STATISTICAL MACHINE LEARNING , 2024

SOLAR POWER PREDICTION

BY: KARTIKEYA SEHGAL &
KRISHNA SHUKLA

Introduction

Problem Statement

Develop machine learning models to forecast solar power output from two plants, using data from solar panels and weather sensors. By modeling factors like irradiation, temperature, and DC power, we aim to improve resource management

BORCELLE

Introduction

Motivation

- India benefits from abundant sunshine and technology.
- However, solar energy is daytime-limited, requiring substantial space and upkeep.
- Solar plant performance depends on sunlight, wind, and temperature, as well as panel efficiency.
- The goal is to understand how these factors interplay to predict solar energy outcomes.

Literary Preview

Subramanian, E., Karthik, M. M., Krishna, G. P., Kumar, V. S., & Prasath, D. V. (2022). Solar Power Prediction Using Machine Learning.

Chakraborty, D., Mondal, J., Barua, H. B., & Bhattacharjee, A. (2023). Computational solar energy – Ensemble learning methods for prediction of solar power generation based on meteorological parameters in Eastern India.

Phan, Q.-T., Wu, Y.-K., & Phan, Q.-D. (2021). Short-term Solar Power Forecasting Using XGBoost with Numerical Weather Prediction.

Barrera, J. M. (2020). Solar Energy Prediction Model Based on Artificial Neural Networks and Open Data. Sustainability, 12(17), 6915.

About Dataset

About Dataset

- The datasets used in the study were sourced from a public contest hosted on Kaggle and focused on solar power plants in India.
- Each dataset was specific to a particular solar power plant and encompassed both solar panel data and weather sensor data.
- Observations were captured at 15-minute intervals over a span of 34 days.
- The data was collected at the inverter level, with each inverter reportedly containing multiple solar panels.

Generation Data

Variable	Meaning
DATE_TIME	Timestamp for each observation, recorded at 15-minute intervals
PLANT_ID	Unique identifier for the specific solar power plant
SOURCE_KEY	Identifier for the specific inverter
AC_POWER	Amount of AC power generated by the inverter (SOURCE KEY) in this 15-minute interval. Measured in kilowatts (kW)
DC_POWER	Amount of DC power generated by the inverter (SOURCE KEY) in this 15-minute interval. Measured in kilowatts (kW)
DAILY_YIELD	Cumulative sum of power generated on that day until that point in time.
TOTAL_YIELD	Total yield for the inverter until that point in time.

Weather Sensor Data

Variable	Meaning
DATE_TIME	Timestamp for each observation, recorded at 15-minute intervals
PLANT_ID	Unique identifier for the specific solar power plant
SOURCE_KEY	Identifier for the specific weather sensor (e.g., temperature sensor, irradiance sensor)
AMBIENT TEMPERATURE	Ambient temperature at the PV plant
MODULE TEMPERATURE	Temperature reading for the solar panel module attached to the sensor panel.
IRRADIATION	Amount of irradiation for the 15-minute interval (Watts per meter-square)

Methodology

01 Exploratory Data Analysis

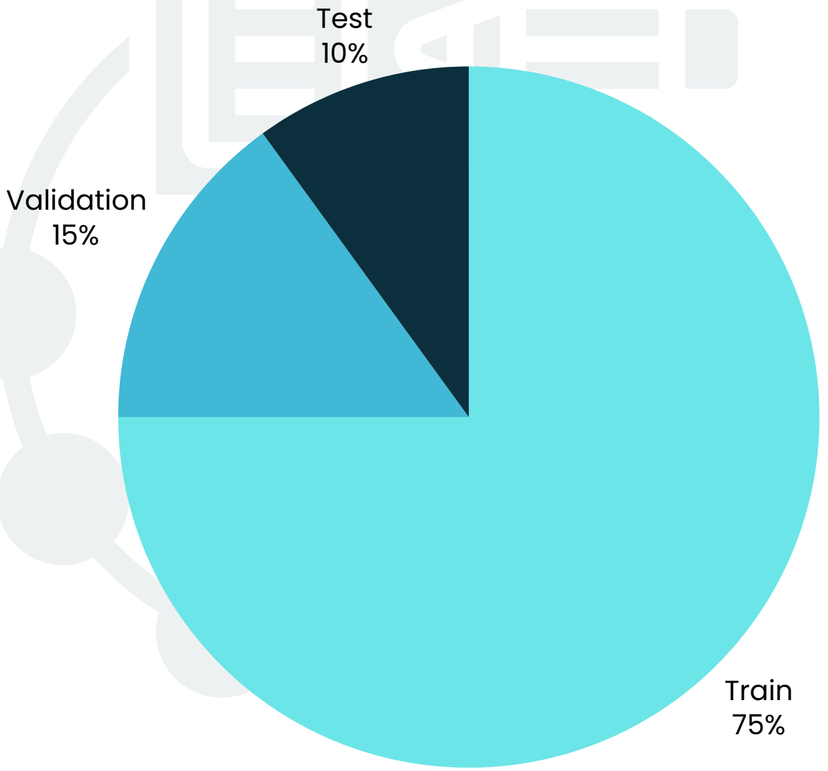
02 Feature Selection

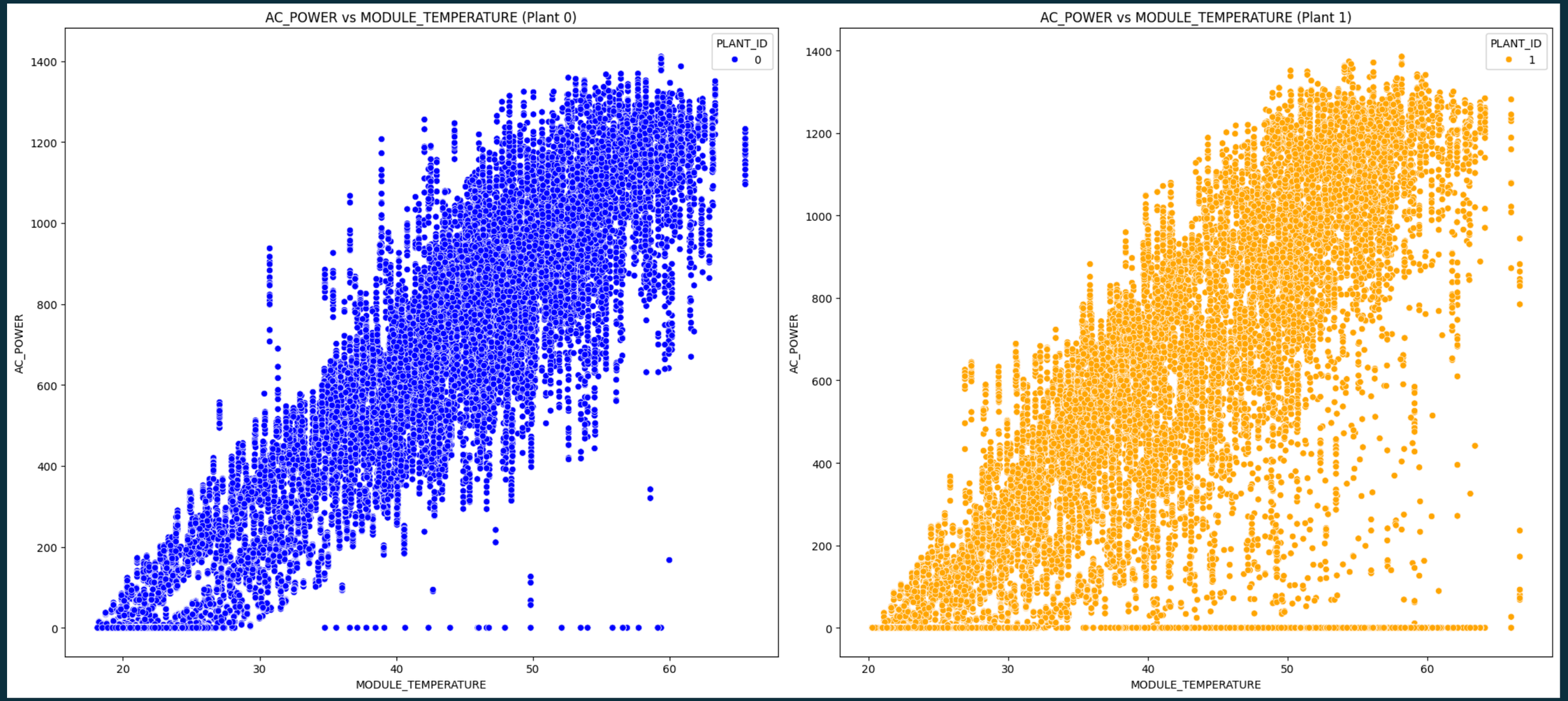
03 Model Training

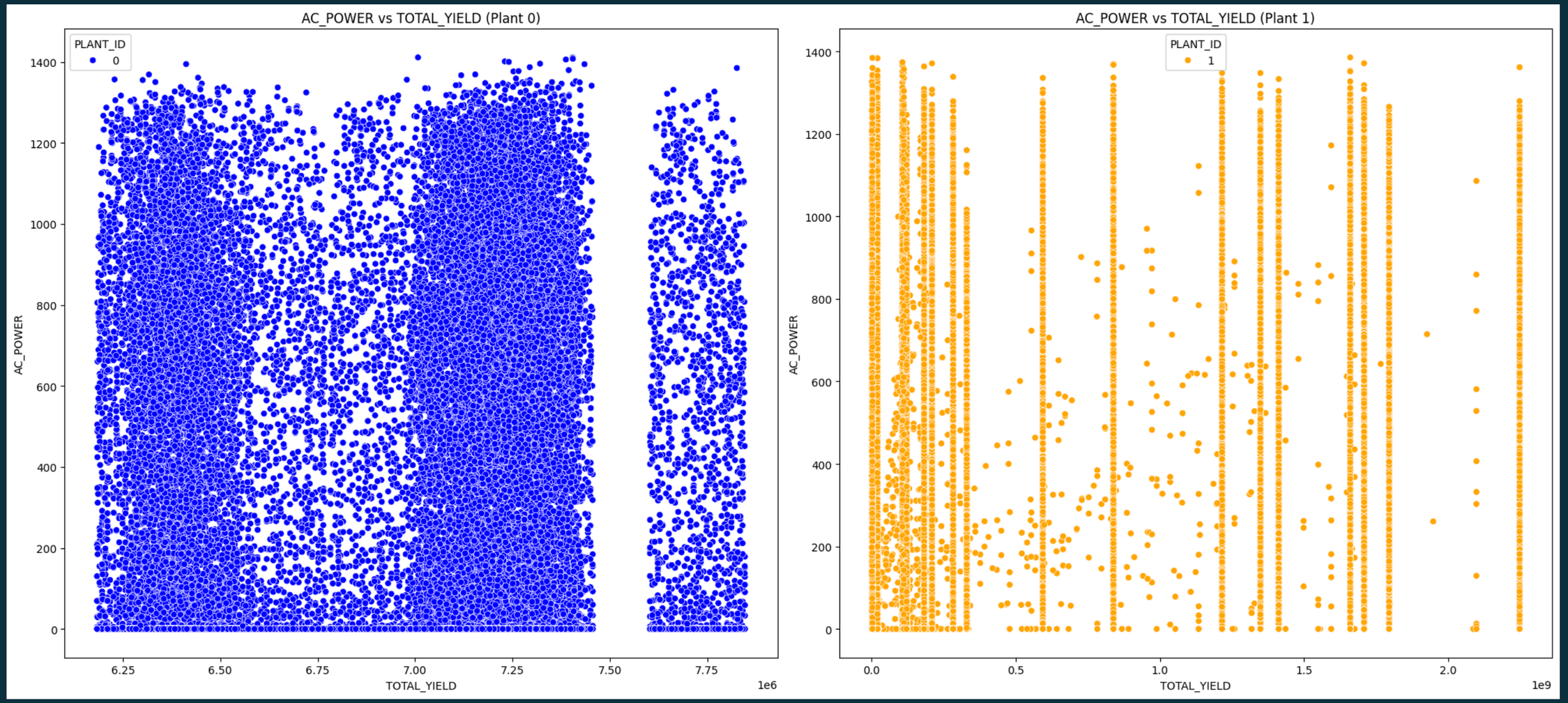
04 Evaluation on validation & test set

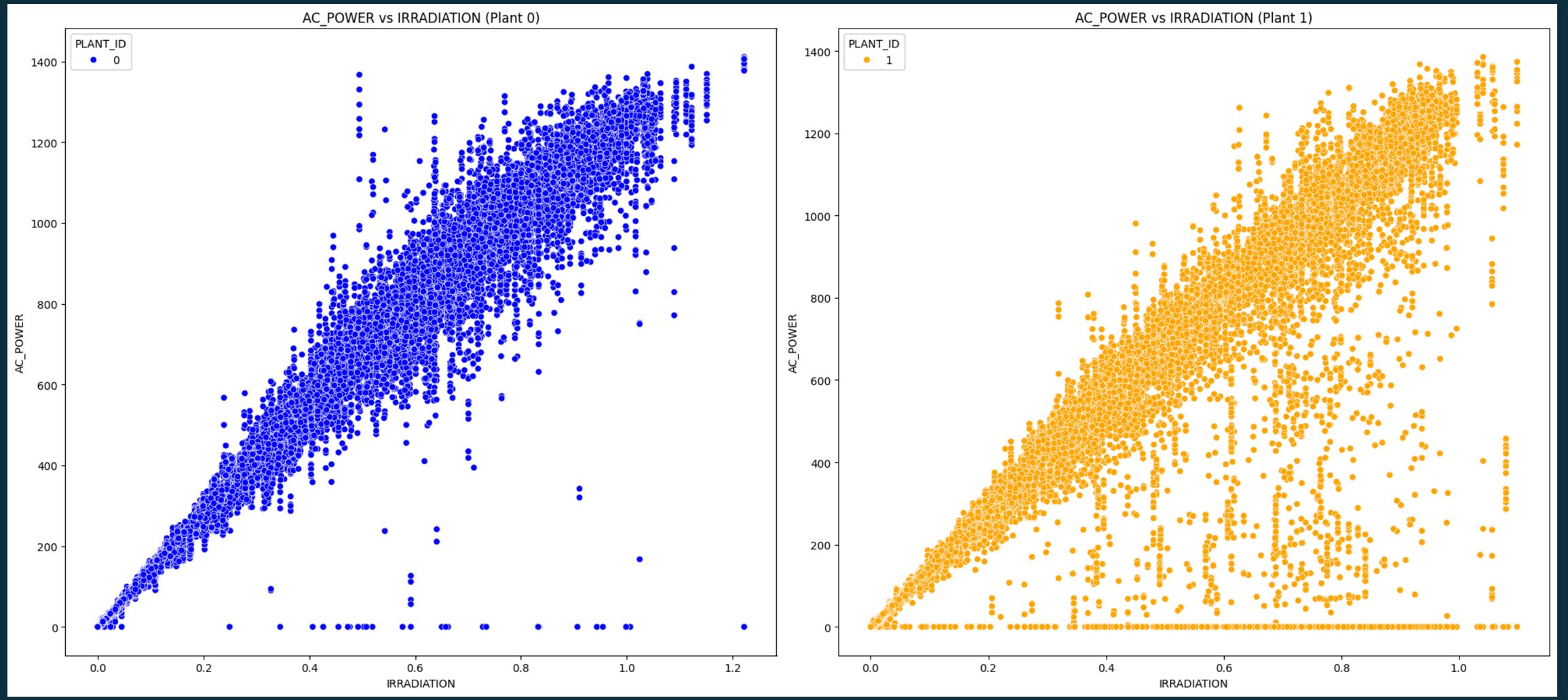
Data Analysis

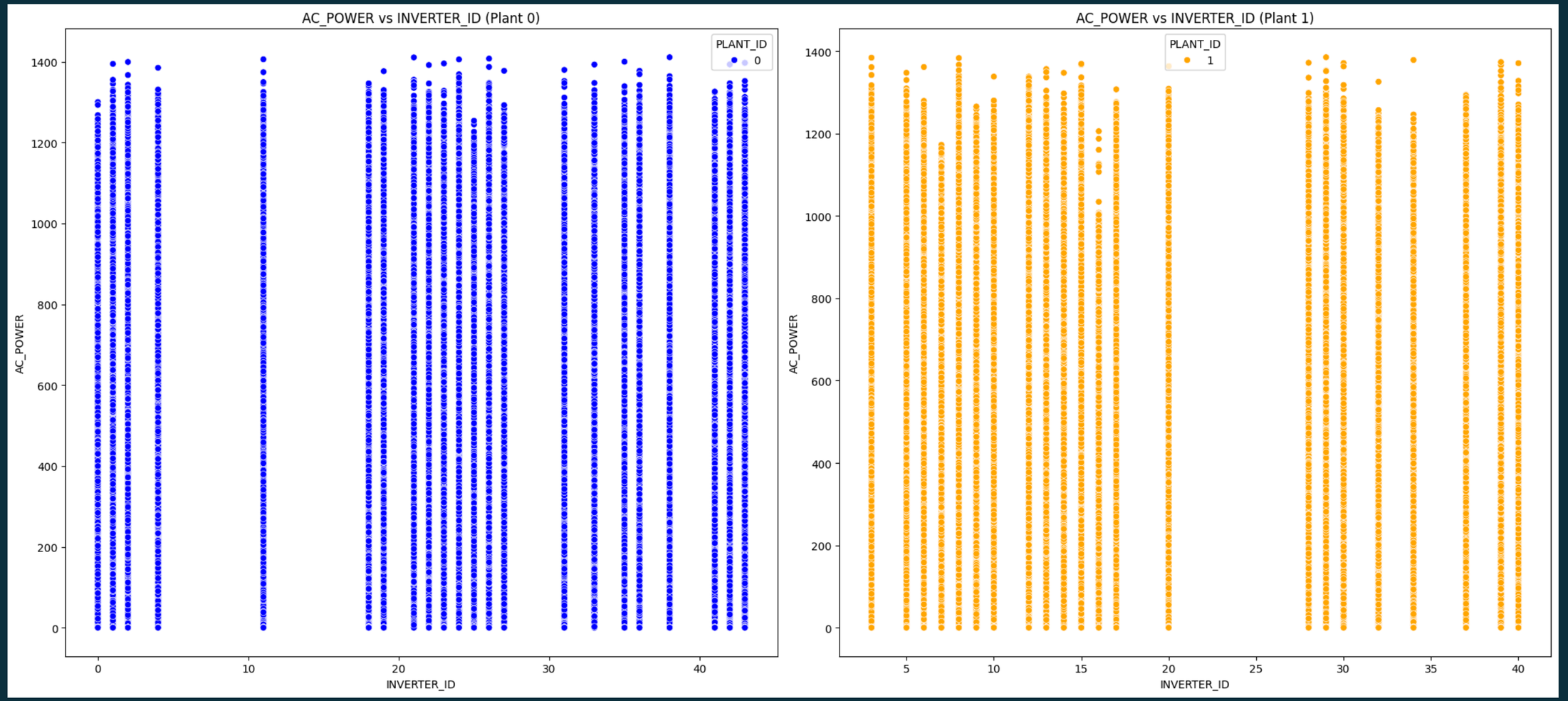
The two datasets were merged on the basis of the common column “DATE_TIME”.
We plot the graphs of all features with the target vector to determine their relationship.
We then plot the correlation matrix and the feature importance of each parameter under consideration.
The reduced dataset is then divided into train, validation and test set as follows.

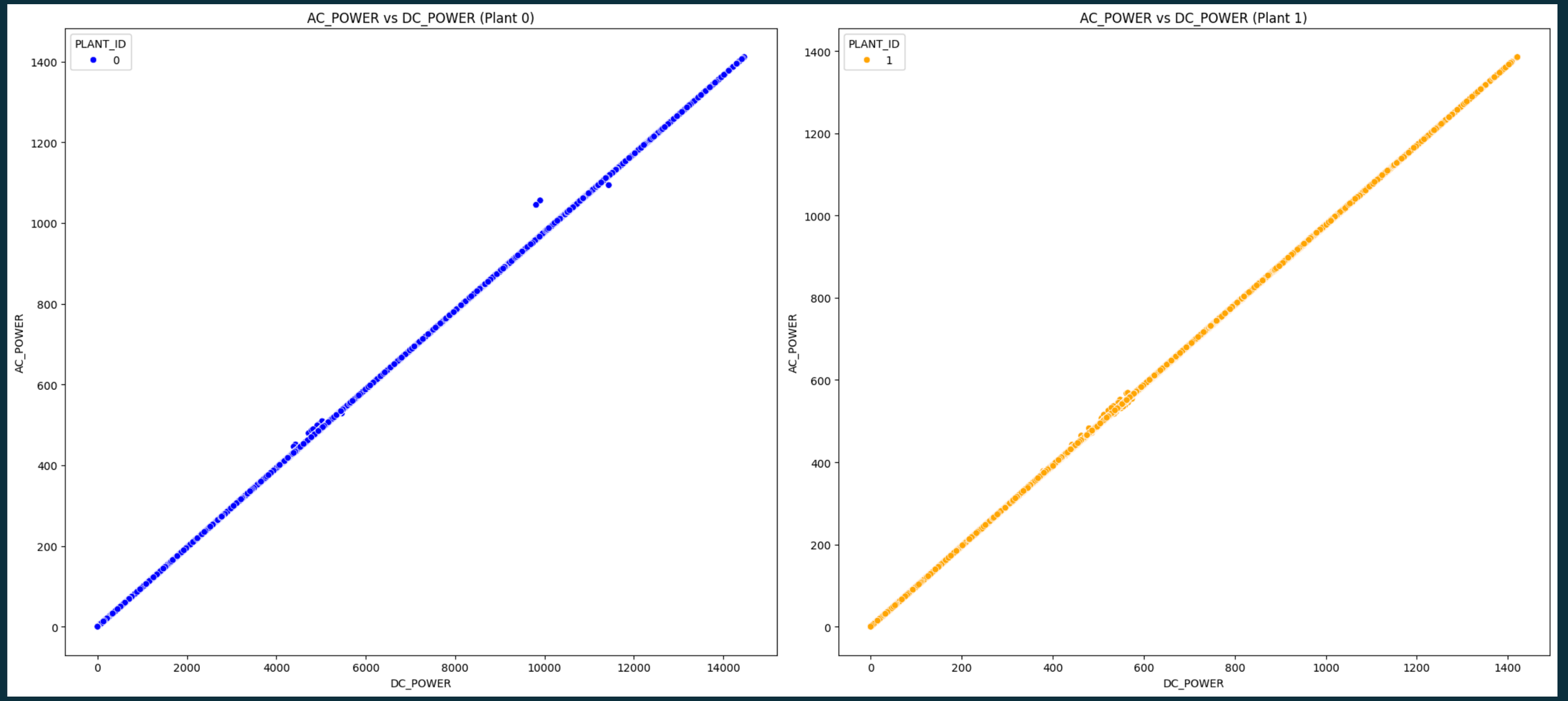


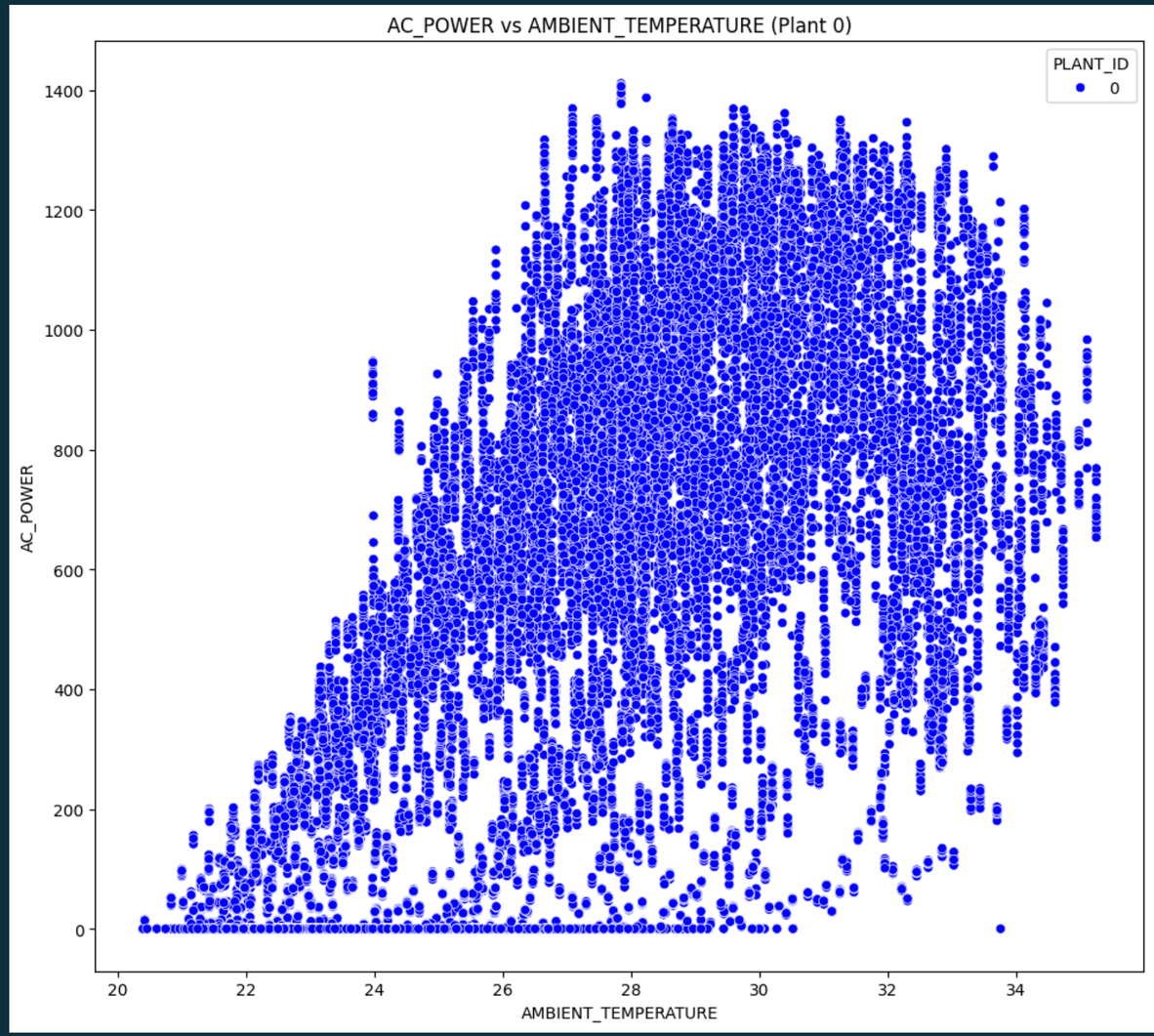

AC_POWER vs MODULE_TEMPERATURE (Plant 0)
PLANT_ID
0
AC_POWER
MODULE_TEMPERATURE
AC_POWER vs MODULE_TEMPERATURE (Plant 1)
PLANT_ID
1
AC_POWER
MODULE_TEMPERATURE

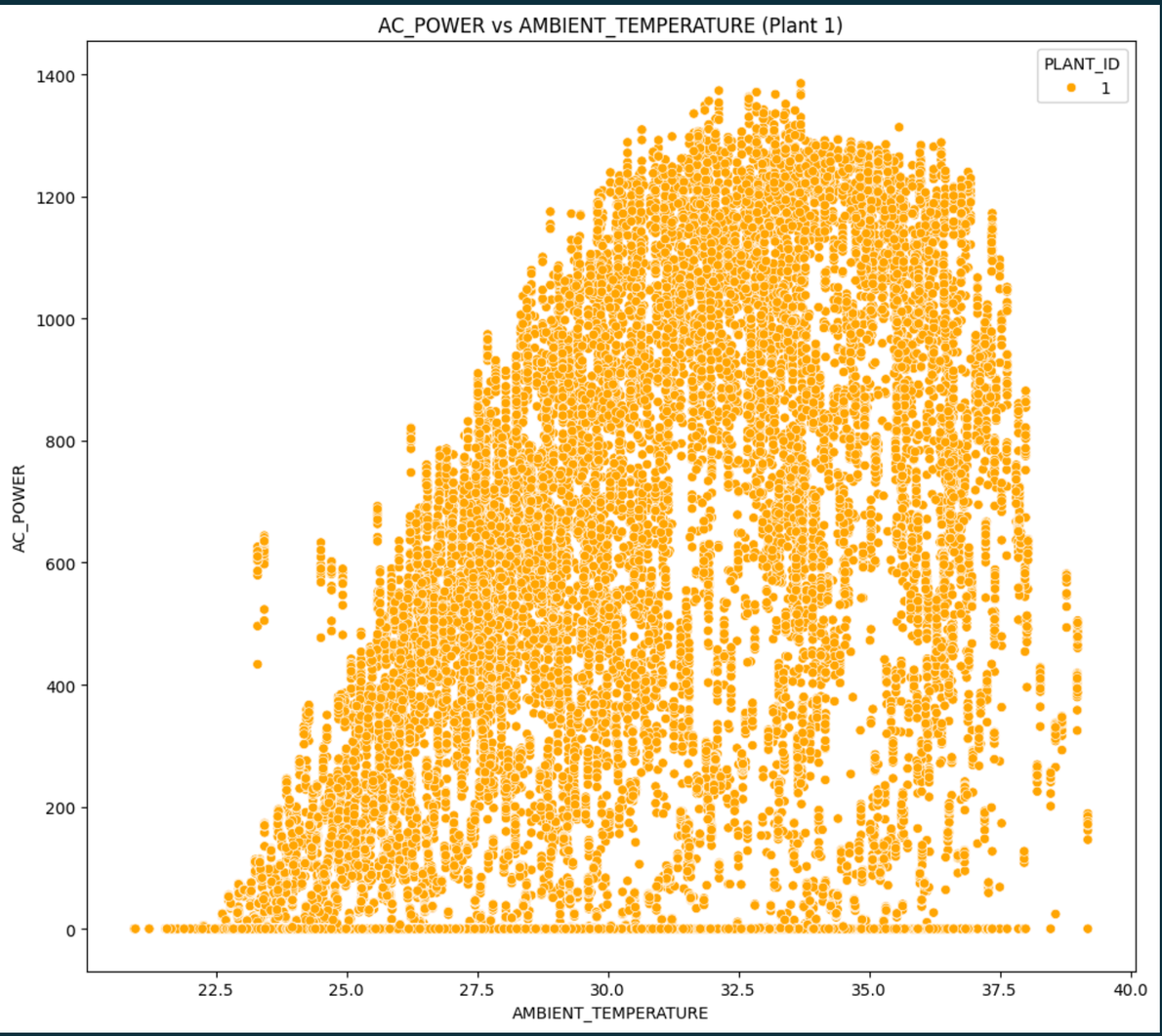

AC_POWER vs TOTAL_YIELD (Plant 0)
PLANT_ID
0
AC_POWER
TOTAL_YIELD
1e6
AC_POWER vs TOTAL_YIELD (Plant 1)
PLANT_ID
1
AC_POWER
TOTAL_YIELD
1e9

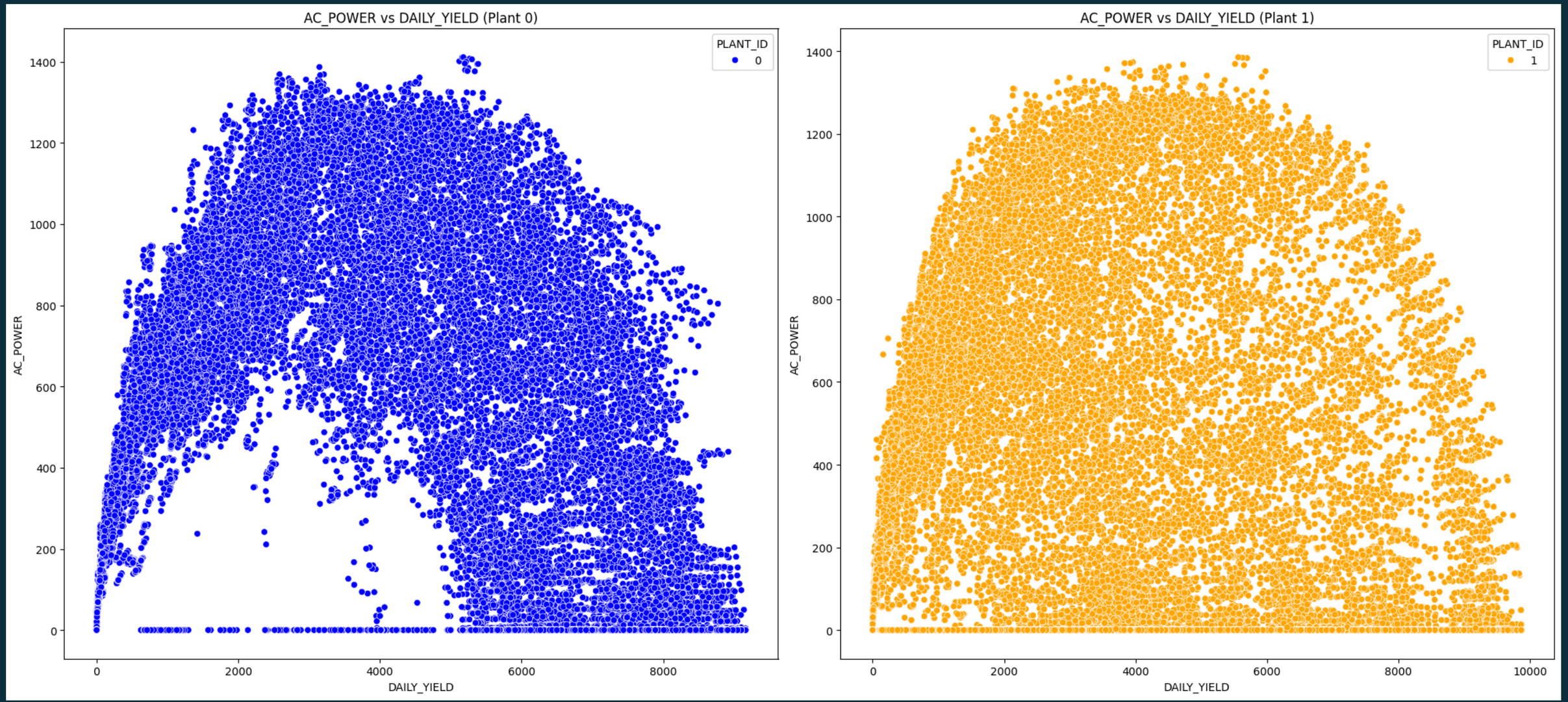

AC_POWER vs IRRADIATION (Plant 0)
PLANT_ID
0
AC_POWER
IRRADIATION
0
200
400
600
800
1000
1200
1400
0.0
0.2
0.4
0.6
0.8
1.0
1.2
AC_POWER vs IRRADIATION (Plant 1)
PLANT_ID
1
AC_POWER
IRRADIATION
0
200
400
600
800
1000
1200
1400
0.0
0.2
0.4
0.6
0.8
1.0

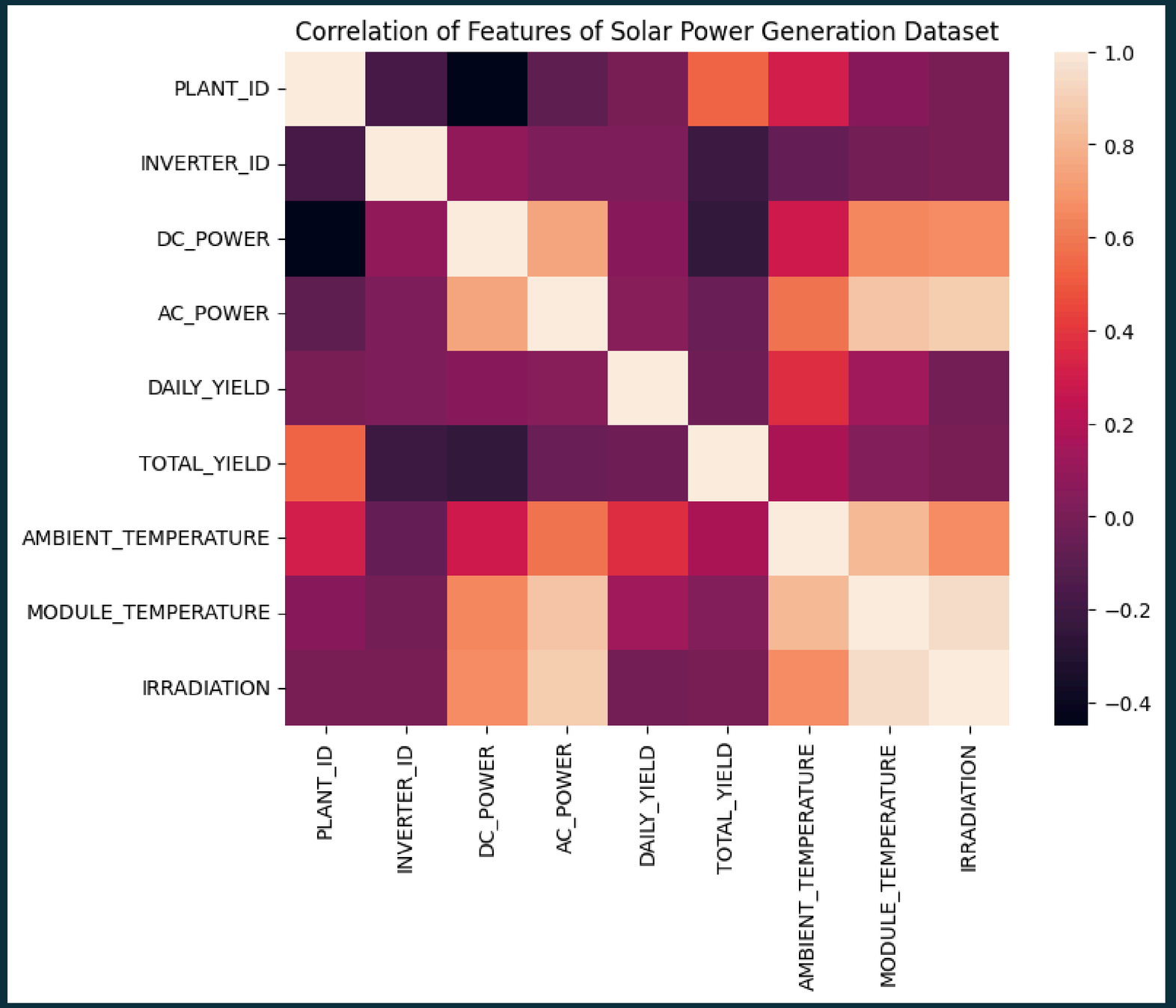

AC_POWER vs INVERTER_ID (Plant 0)
PLANT_ID
0
AC_POWER
INVERTER_ID
0
200
400
600
800
1000
1200
1400
0
10
20
30
40
AC_POWER vs INVERTER_ID (Plant 1)
PLANT_ID
1
AC_POWER
INVERTER_ID
0
200
400
600
800
1000
1200
1400
5
10
15
20
25
30
35
40

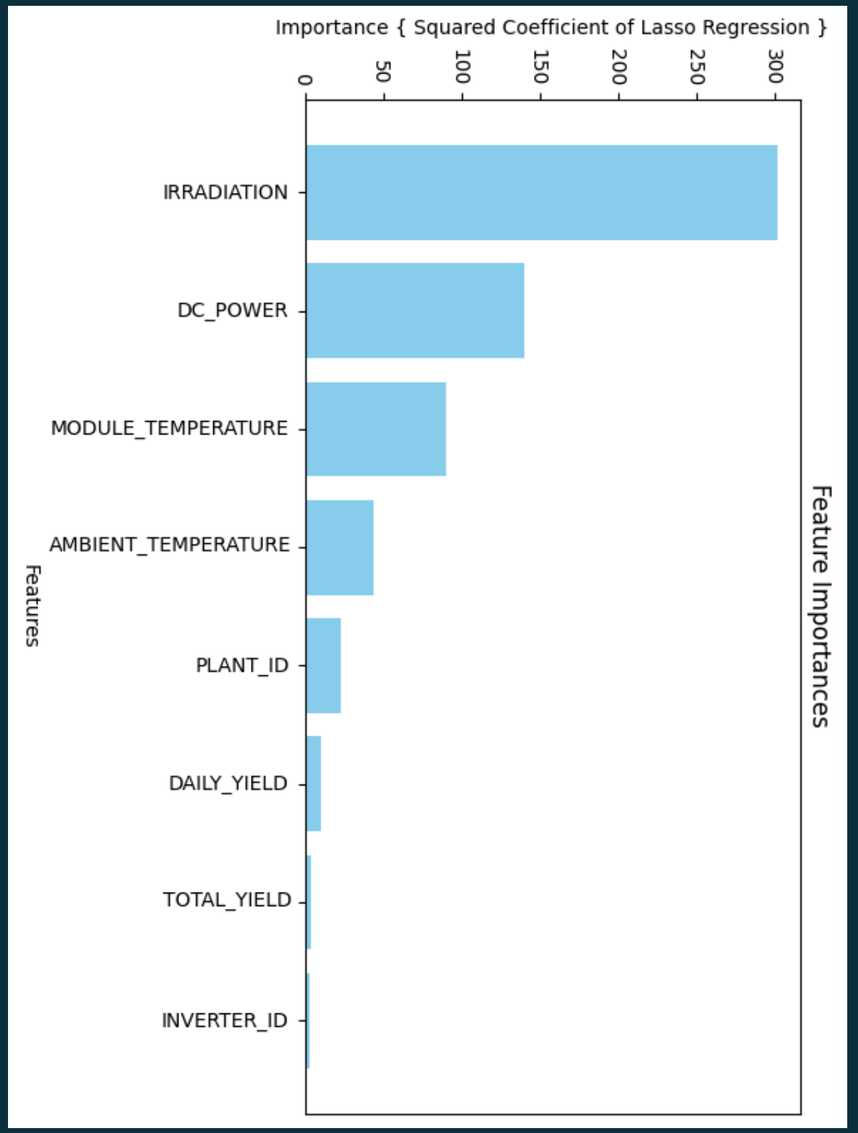

AC_POWER vs DC_POWER (Plant 0)
PLANT_ID
0
AC_POWER
DC_POWER
0
200
400
600
800
1000
1200
1400
0
2000
4000
6000
8000
10000
12000
14000
AC_POWER vs DC_POWER (Plant 1)
PLANT_ID
1
AC_POWER
DC_POWER
0
200
400
600
800
1000
1200
1400


AC_POWER vs AMBIENT_TEMPERATURE (Plant 0)
PLANT_ID
0
AC_POWER
AMBIENT_TEMPERATURE
0
200
400
600
800
1000
1200
1400
20
22
24
26
28
30
32
34


AC_POWER vs AMBIENT_TEMPERATURE (Plant 1)
PLANT_ID
1
AC_POWER
AMBIENT_TEMPERATURE
0
200
400
600
800
1000
1200
1400
22.5
25.0
27.5
30.0
32.5
35.0
37.5
40.0


AC_POWER vs DAILY_YIELD (Plant 0)
PLANT_ID
0
AC_POWER
DAILY_YIELD
0
200
400
600
800
1000
1200
1400
0
2000
4000
6000
8000
AC_POWER vs DAILY_YIELD (Plant 1)
PLANT_ID
1
AC_POWER
DAILY_YIELD
0
200
400
600
800
1000
1200
1400
0
2000
4000
6000
8000
10000


Correlation of Features of Solar Power Generation Dataset
PLANT_ID
INVERTER_ID
DC_POWER
AC_POWER
DAILY_YIELD
TOTAL_YIELD
AMBIENT_TEMPERATURE
MODULE_TEMPERATURE
IRRADIATION
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


Importance { Squared Coefficient of Lasso Regression }
0
50
100
150
200
250
300
IRRADIATION
DC_POWER
MODULE_TEMPERATURE
AMBIENT_TEMPERATURE
PLANT_ID
DAILY_YIELD
TOTAL_YIELD
INVERTER_ID
Features
Feature Importances

Feature Selection

Selected relevant parameters based on visualization and correlation analysis. Prioritized features with high correlation with the target variable. We then reduce the dataset into the following features:

- **Irradiation**
- **Module Temperature**
- **Ambient Temperature**
- **DC Power**

Models

LINEAR REGRESSION

Linear Regression in this scenario is the basic Least Squares Regression. It failed to capture the complexities of the data due to its ’simple’ nature

RANDOM FOREST

Random Forest is a more robust version of Bagging in which we pick √d parameters at random from the input instead of the entire d dimensions (Bagging).

ADABOOST REGRESSION

Adaptive Boosting Regression is an ensemble method in which the ’subsequent regressors focus on the more difficult areas that previous iterations failed to explain.

XGBOOST

- Built-In Cross Validation
- Regularization
- Optimized Tree Pruning

Performed better than the other three.

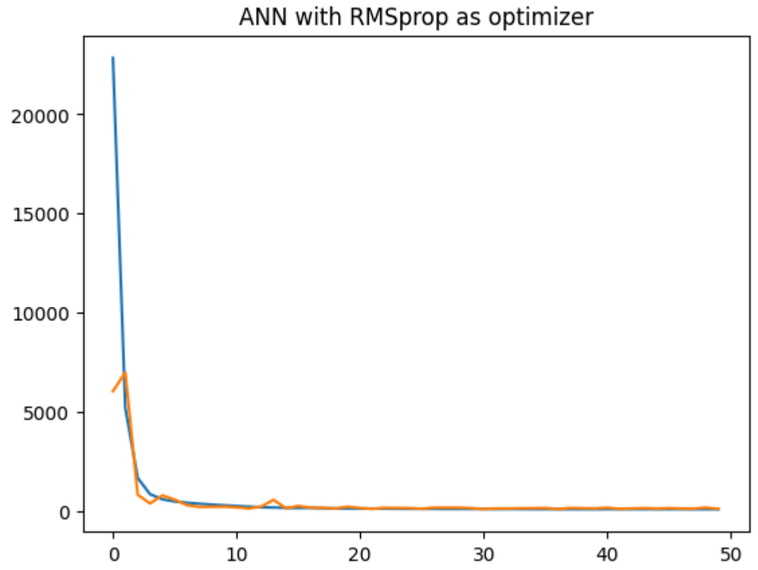
Models

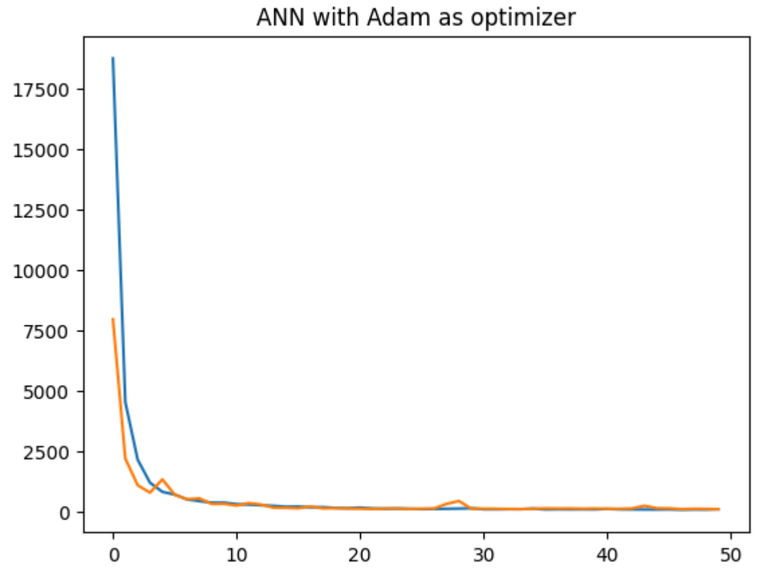
ARTIFICIAL NEURAL NETWORK (ANN)

We developed 2 ANN Models.

- input dimension = 4
- Three ‘Dense’ layers in each
- Loss Function = Mean Squared Error
- ReLU activation on the first two layers
- Linear activation on last layer.

- Model 1 was compiled with Root Mean Squared Back Propagation
- Model 2 was compiled with Adam Optimization
- Both models were trained for 50 epochs





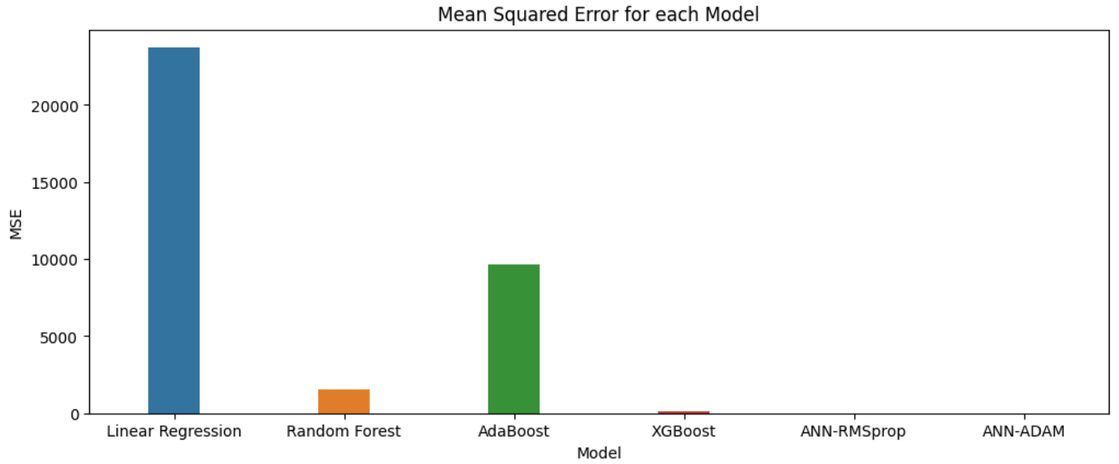
Result

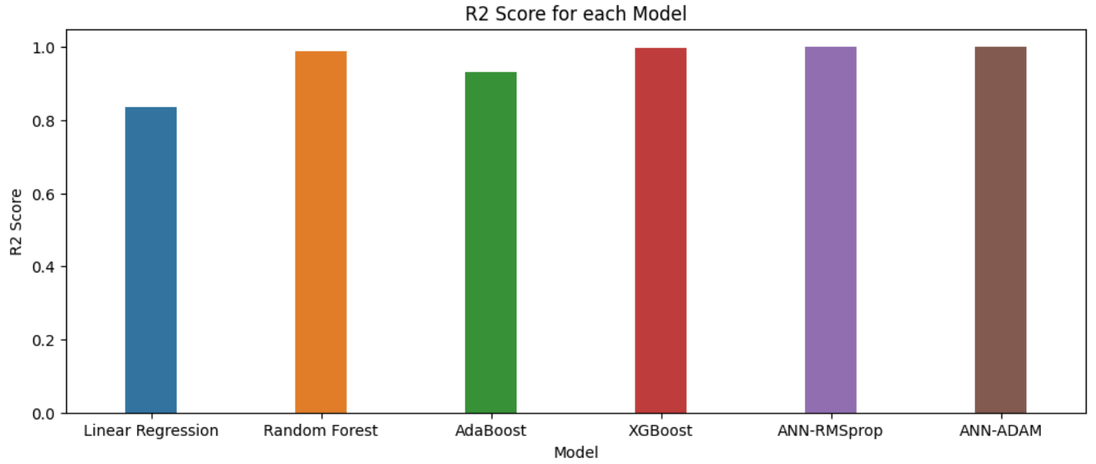
The evaluation metrics chosen were:

1) MSE
2) R2- Score

TABLE I
MODEL PERFORMANCE

Model	MSE	R2-Score
Linear Regression	23702.03	0.8348
AdaBoost	9673.88	0.9326
Random Forest	1567.58	0.9891
XGBoost	97.26	0.9992
ANN-RMSprop	6.39	0.99996
ANN-Adam	3.59	0.99997





Research Result

Conclusion

Evaluation on Test Set

Since the model that performed best on the validation set was the Artificial Neural Network with Adaptive Moment Estimation, we select that model as the model which is evaluated on the TEST Set.

ANN_ADAM Performance on Test Set

Metric	Value
Test MSE	52.41
Test R2-Score	0.9996

BORCELLE

Further Improvement

01

HyperTuning

Applying Cross Validation to determine the optimum number of epochs, the optimum learning rate or penalizing factor.

02

Exploring newer datasets with more features corresponding to WEeather

THANKS

FOR YOUR ATTENTION

Kartikeya Sehgal (2022244)
Krishna Shukla (2022254)