

**Elements of AIML**

**ASSIGNMENT 1**

**Name:** Somya Vats

**Batch:** 11

**SAP:** 500119012

**Roll no:** R2142230349

# **Predicting Solar Plant Yield Using Machine Learning on Weather and Power Data**

* [*GOOGLE DRIVE FOLDER*](https://drive.google.com/drive/folders/1AR4G3YjDdZSGfRpAtZKXJQ8EXupjzynV?usp=sharing)
* [*GITHUB LINK*](https://github.com/vatssomya/Assignment_1)

**Introduction**

As climate change concerns continue to grow, global reliance on renewable sources has been increasingly sought out in order to not only follow a greener path, but to specifically address Sustainable Development Goals – SDG 7: Affordable and Clean Energy, and SDG 13: Climate Action. Solar energy provides clean energy, but it remains a hard area to maximise due to the uncontrollable factors of nature, which ultimately leads to variability in solar generation. All these factors make solar power forecasting an indispensable task when it comes to efficient energy management systems, plannings, and resource utilizations.

In this project, we focus on estimating how much energy a solar plant can produce by looking at past generation of the plant and weather conditions. Through the use of python ML models, the objective is to build a model that manages to forecast solar generation for a given weather scenario. Such models are a useful piece of machinery for solar plant operators as they can enhance the decision making by improving resource distribution and assisting in sustainable energy consumption as well.

This task is dependent on using Machine Learning given its sophistication in recognizing intricate structures within extensive amounts of data, self adjusting in real time, and providing forecast abilities which would be impossible otherwise. This project shows the other side of ML: how it changes the game for the renewable business.

**Problem Statement**

The production of solar energy is subject to the whims of sunlight, temperature, and cloud cover, making it an unreliable source that is difficult for grid operators to effectively budget for. This variability can threaten grid stability, make energy storage planning difficult, and create energy distribution inefficiencies, especially with the growing worldwide need for renewable energy.

The International Renewable Energy Agency (IRENA) says that solar energy will need to increase dramatically in order to meet climate action goals, possibly increasing by a factor of ten by the year 2050. In order to satisfy these requirements and to increase reliability, precise solar yield forecasts are necessary. This project will use machine learning and historical weather and power generation data to predict how much a solar plant will yield, so that the operators can manage the flow of energy, maximize the stability of the grid, and reach energy supply goals.

**Solution Overview**

Based on the examination of past weather statistics and related energy generation data, this project proposes a machine learning technique for estimating the total energy output for a solar plant. To achieve this goal the model inputs four environmental variables; sunlight intensity, cloud cover, temperature and humidity. This enables solar power plant managers to optimize their choices for energy storage, integration into the grid and investment level for deployment resources.

The time-series data contains information about the solar power output and the changes in the atmosphere including different weather conditions such as day and night and the seasons is what the model is trained upon. Then this model is applied to predict energy generation during the next hours while attempting to balance energy demand and supply in the grid. This makes it possible for the operators to appropriately prepare for high production periods and low production periods.

The flexibility of the model means that it can be transferred across various locations and applied to other similar solar plants with appropriate geographical data. Also, since the system assimilates additional information, its precision and flexibility enhance over time, making it an effective tool for the renewable energy sector to meet optimized productivity and efficiency.

**Tech Stack**

This application was developed with the help of HTML/CSS and Flask as a framework for process-oriented model training and interaction with users, along with the combination of several libraries of the Python programming language.  
  
 **Python Libraries:**  
  
I. Pandas and NumPy were used for data handling and storage as well as for statistical data analysis, that required dealing with big amounts of data.  
II. Scikit-learn was utilized for performing most of the machine learning algorithms, model training and hyperparameters tuning.  
III. Matplotlib and seaborn libraries aimed to illustrate data and models results in graphical representations.  
  
**Flask (Backend):** As the backend framework application, Flask offered the simplest integration tool to the developers as minimalistic API routes existed in order to make predictions and query data.  
  
**HTML/CSS (Frontend):** Used to construct model visualizations and model forecast outputs allowing users to perform prediction tasks and other engaged procedures actively and simply and look at descriptive parts of the data such as images.

# **CODE DESCRIPTION**

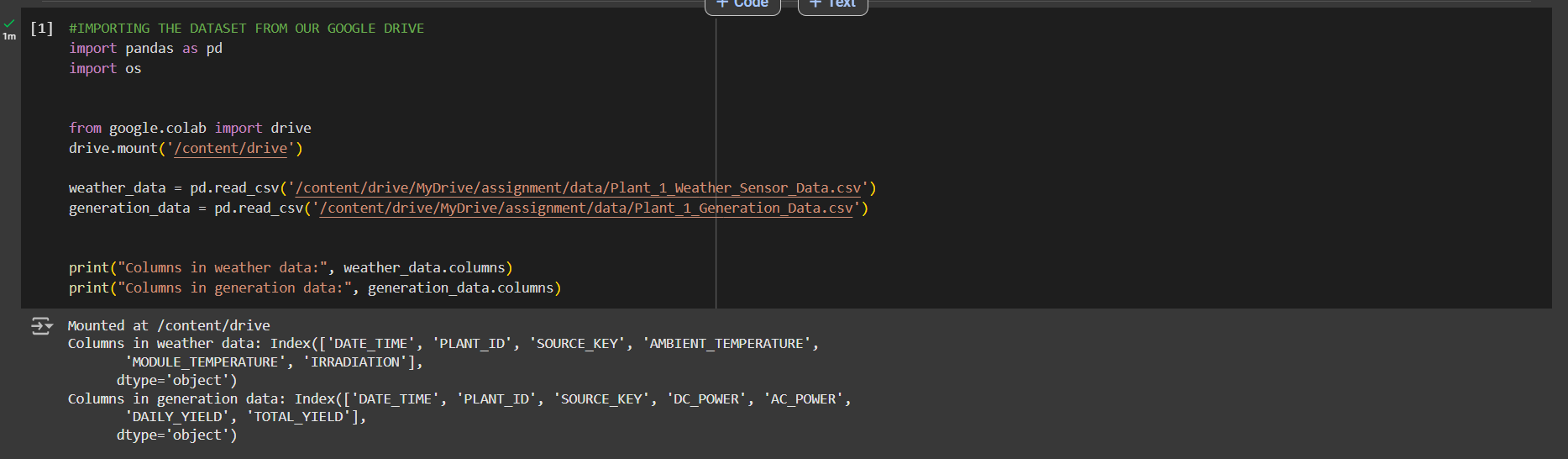
* [*GOOGLE DRIVE FOLDER*](https://drive.google.com/drive/folders/1AR4G3YjDdZSGfRpAtZKXJQ8EXupjzynV?usp=sharing)
* [*GITHUB LINK*](https://github.com/vatssomya/Assignment_1)

**STEP 1 DATA ACQUIRING**

*1.   Dataset has been taken from kaggle and uploaded into google drive as a ".csv" file*

*2.   Drive is imported and so is the data.*

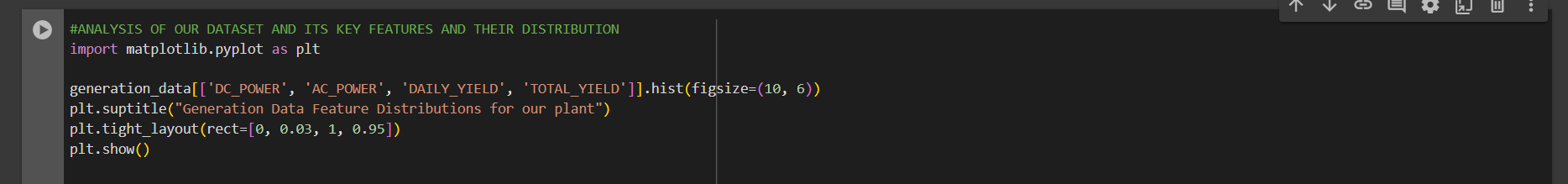
*3.   Then, just to check the success of these steps, few content is printed.*

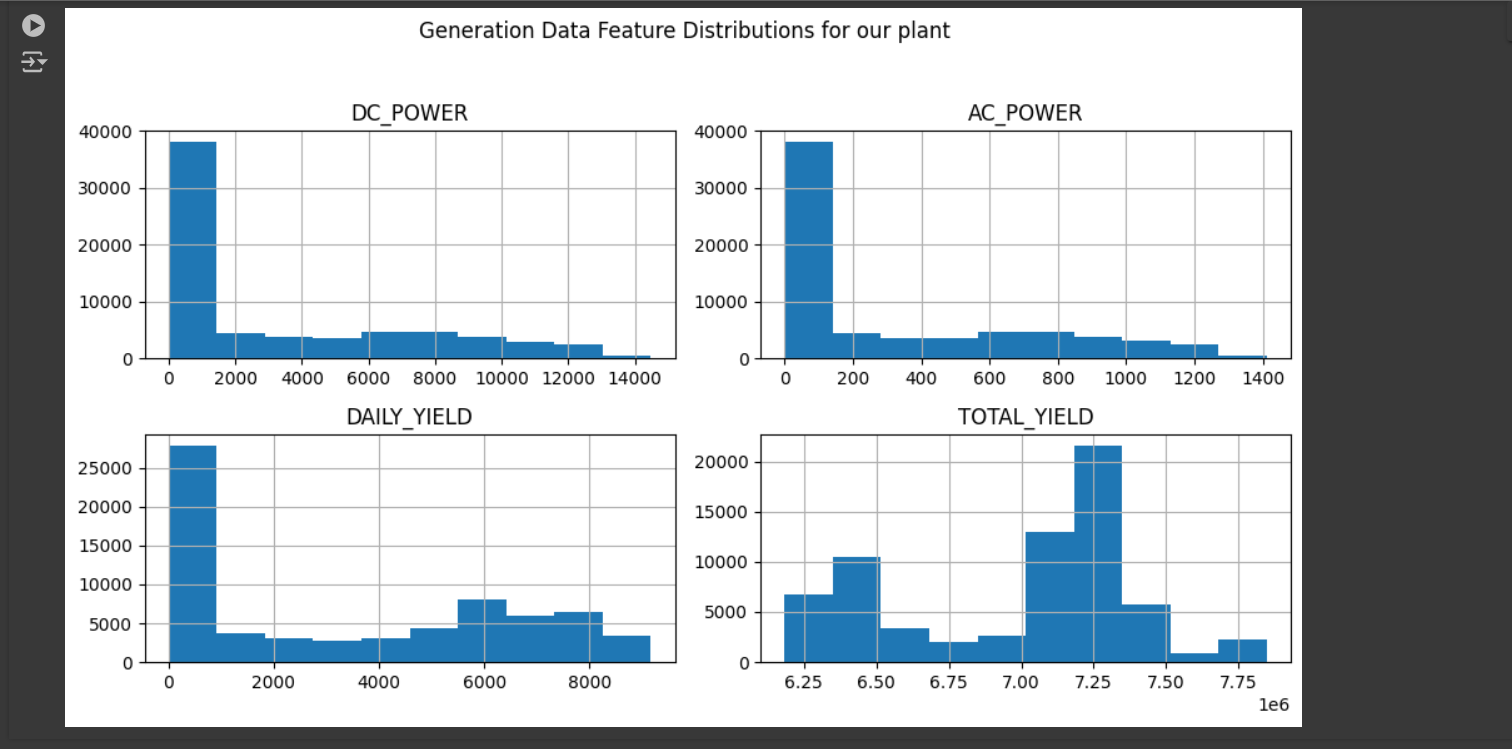


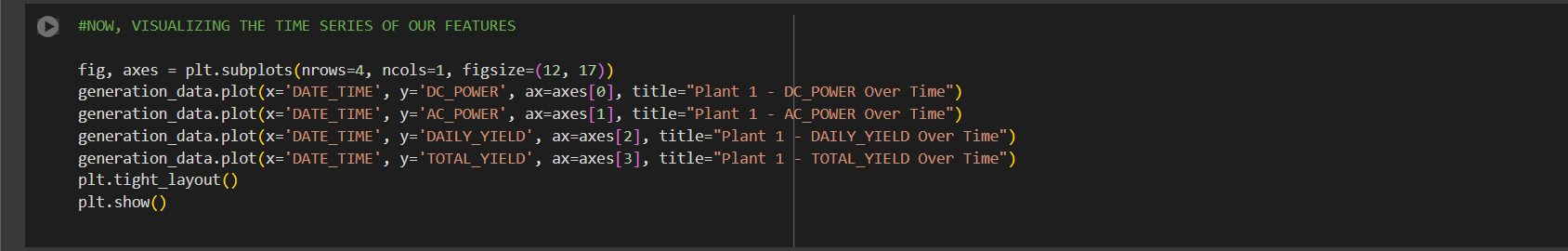
**STEP 2 VISUALIZATION OF CURRENT DATA'S TREND AND FEATURES**

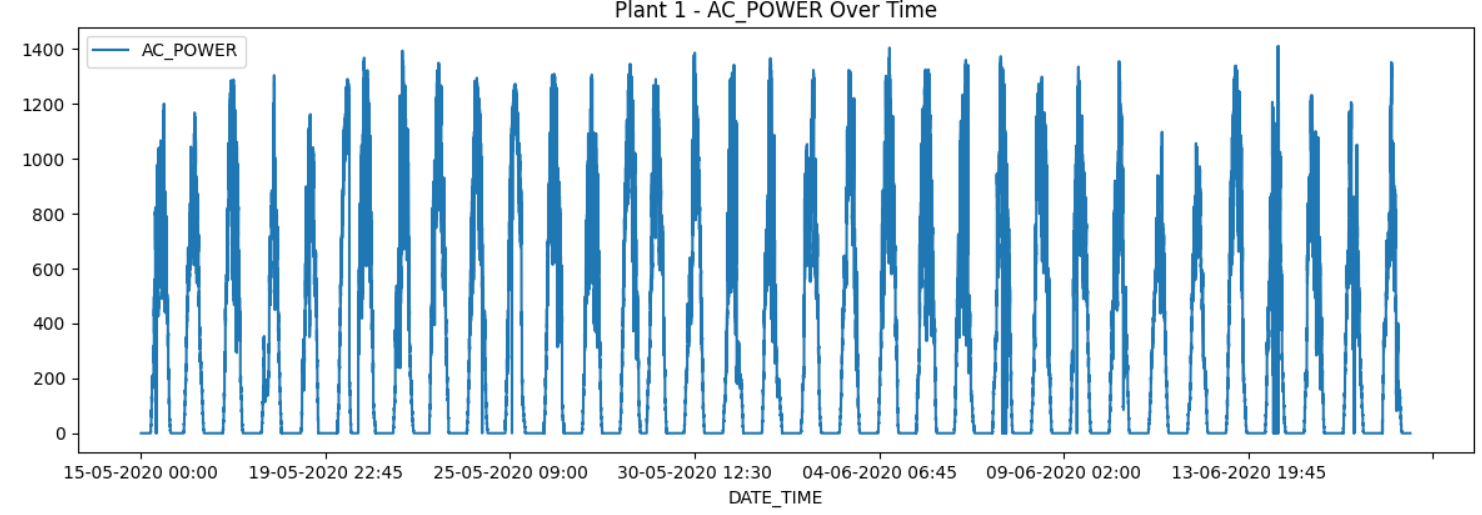
*1.   Checking the distribution of the features of the data*

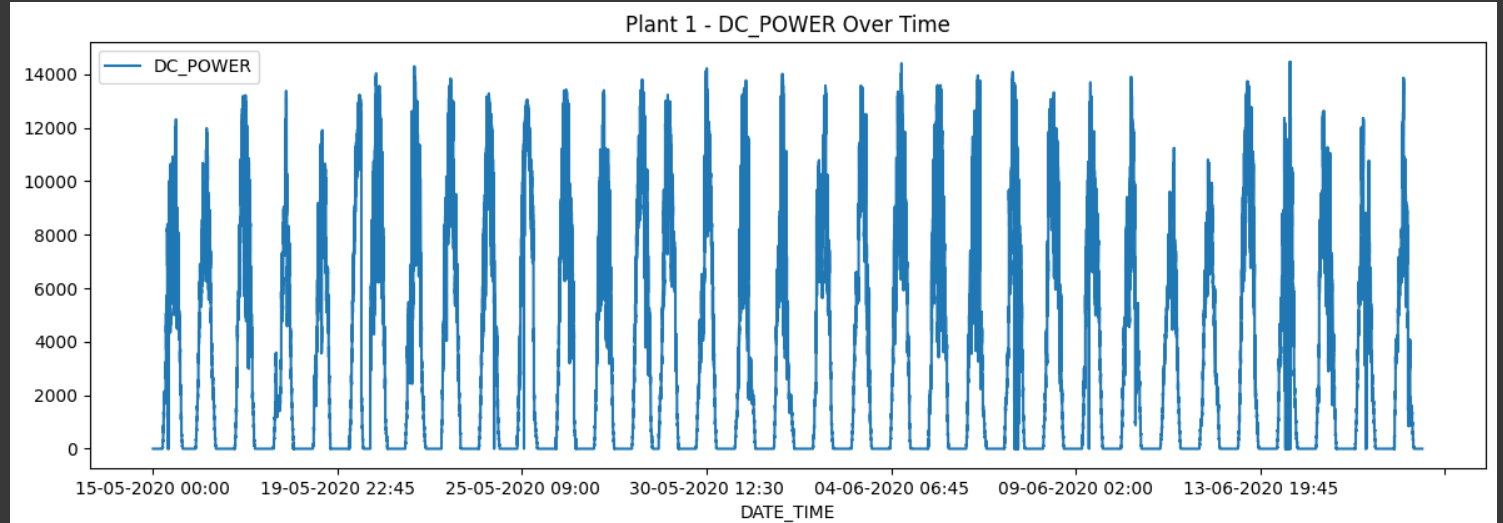
*2.   Then, checking the time series trends, as later, we merge both our dataset based off the "Date\_time" column*

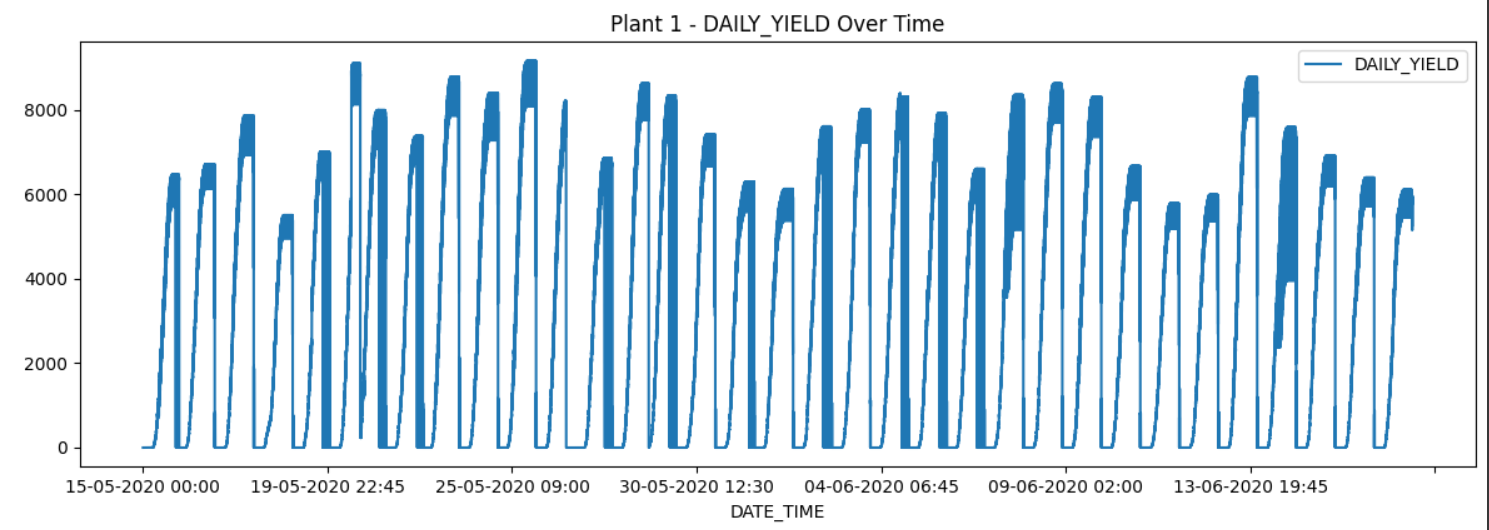


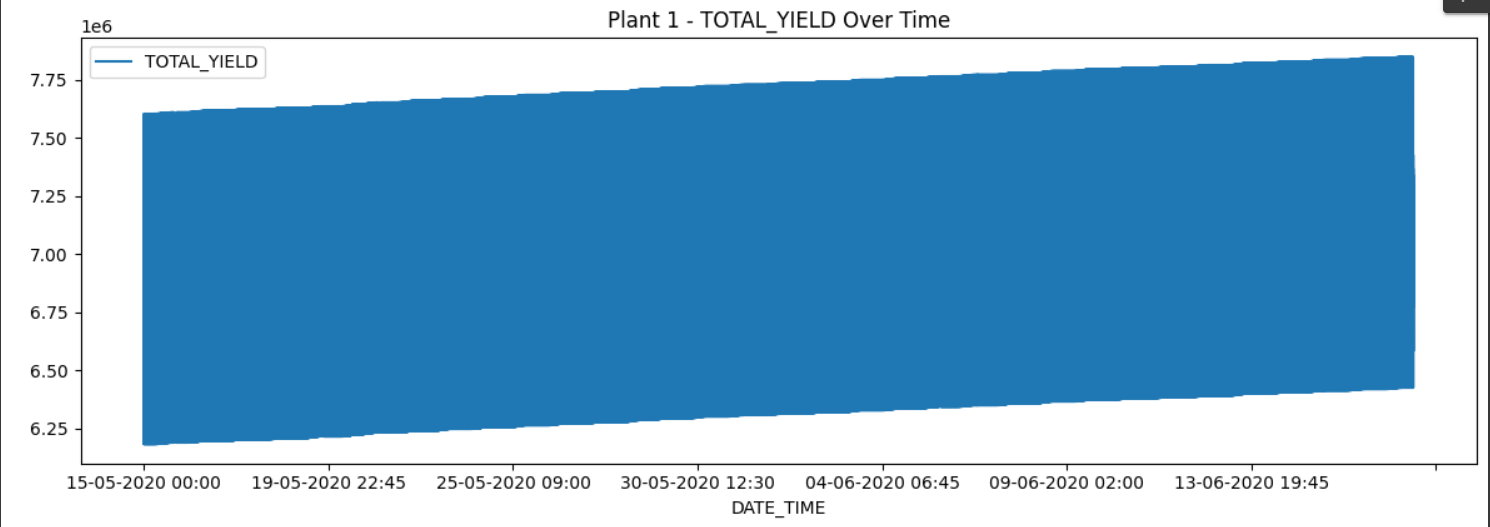












**STEP 3. DATA PROCESSING**

*1.   We change the format of the "DATE\_TIME" column into the datetime data format in python*

*2.   Then, we check the data types and missing values for both the datasets*

*3.   Since generation data is huge, we condense it*

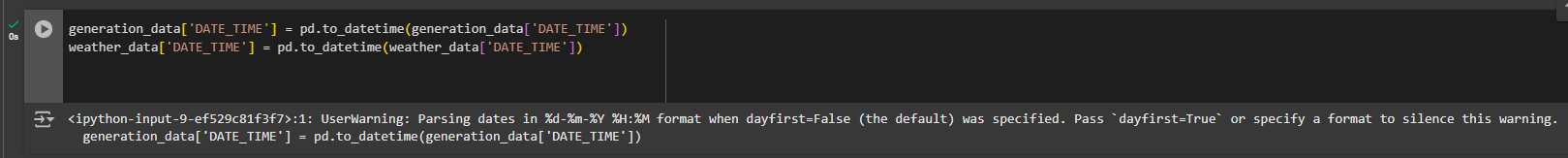
*4.   Then, on the basis of "DATE\_TIME" and "PLANT\_ID", merge both the datasets as the prediction will be made on basis of combined features*

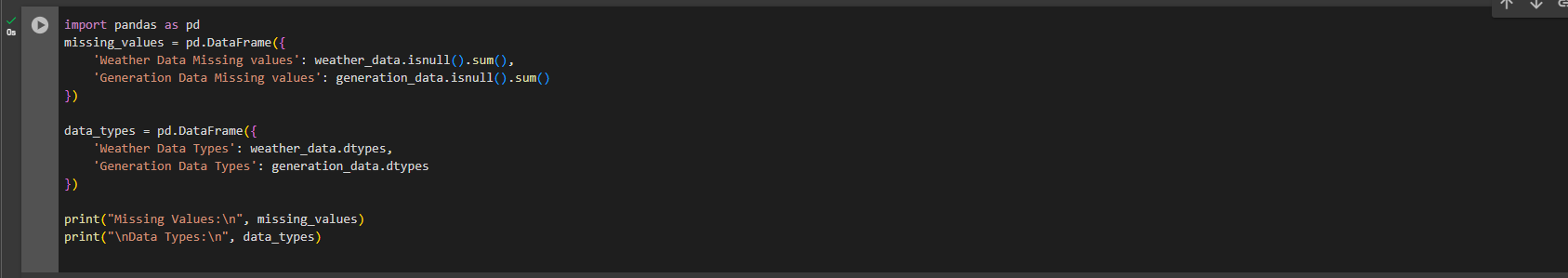
*4.   We condense "merged\_data" into a sub-dataset as the initial one was very large size*

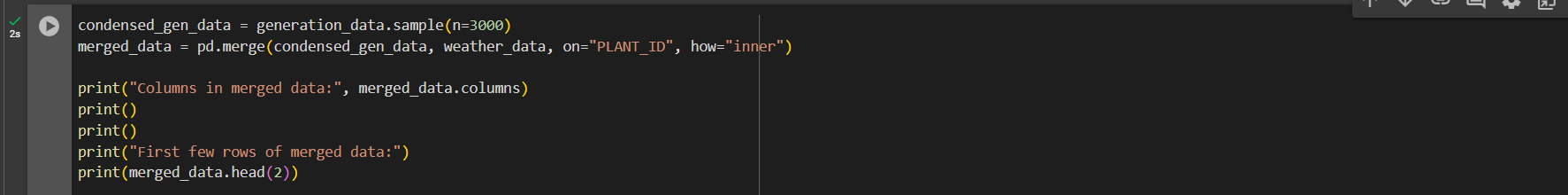
*5.   We distribute the train and validation splits over the data*

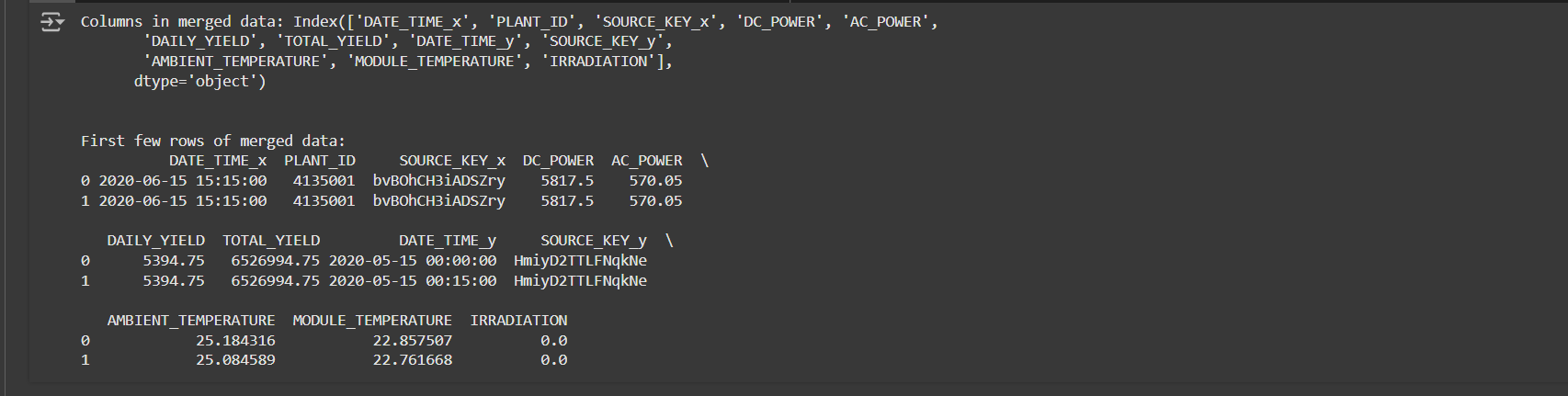
*6.   We check if we require SMOTE application for our data*

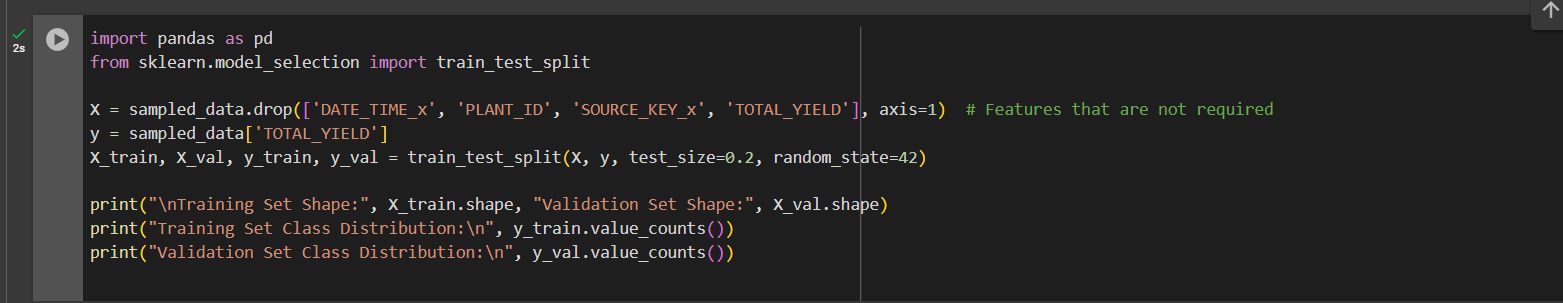
*7.   We further check the class distribution for imbalances, so to further check the applicabilty of SMOGN, etc*





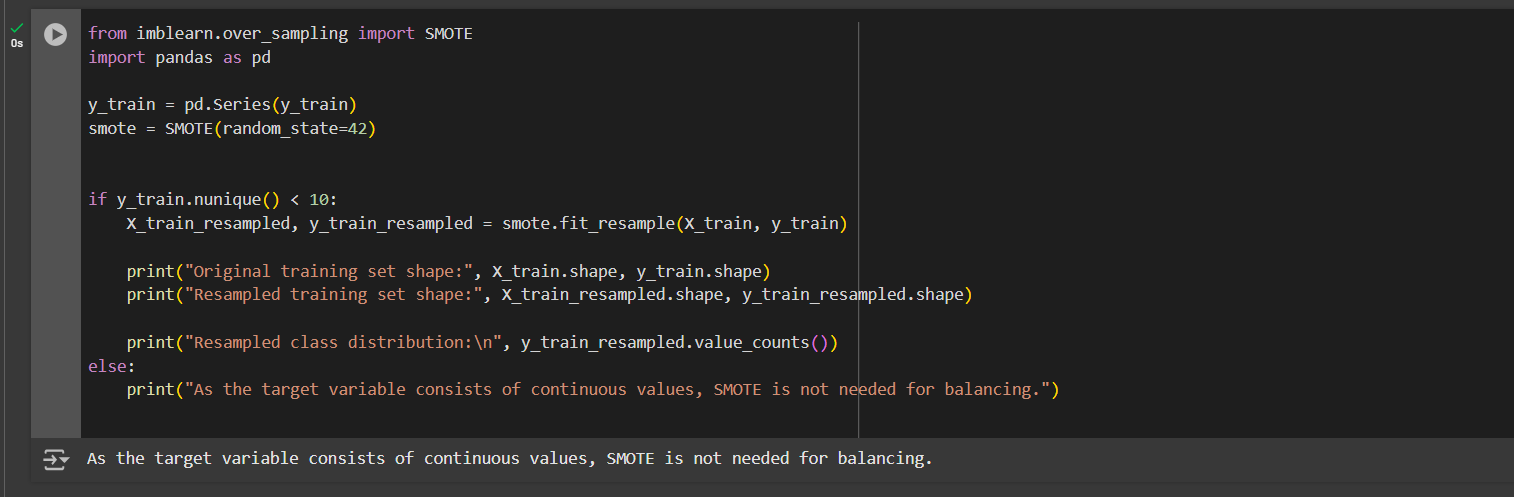








***SMOTE*** *is not applied to the continuous data*

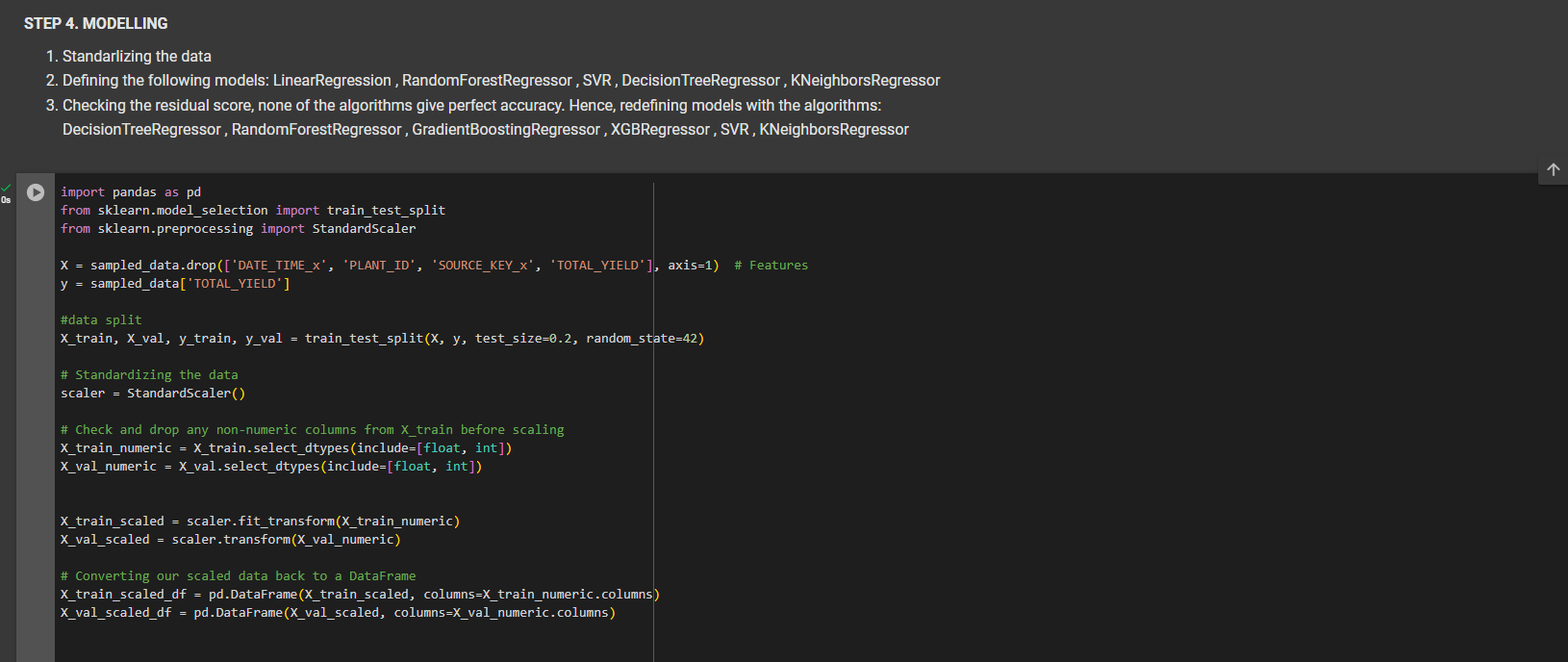


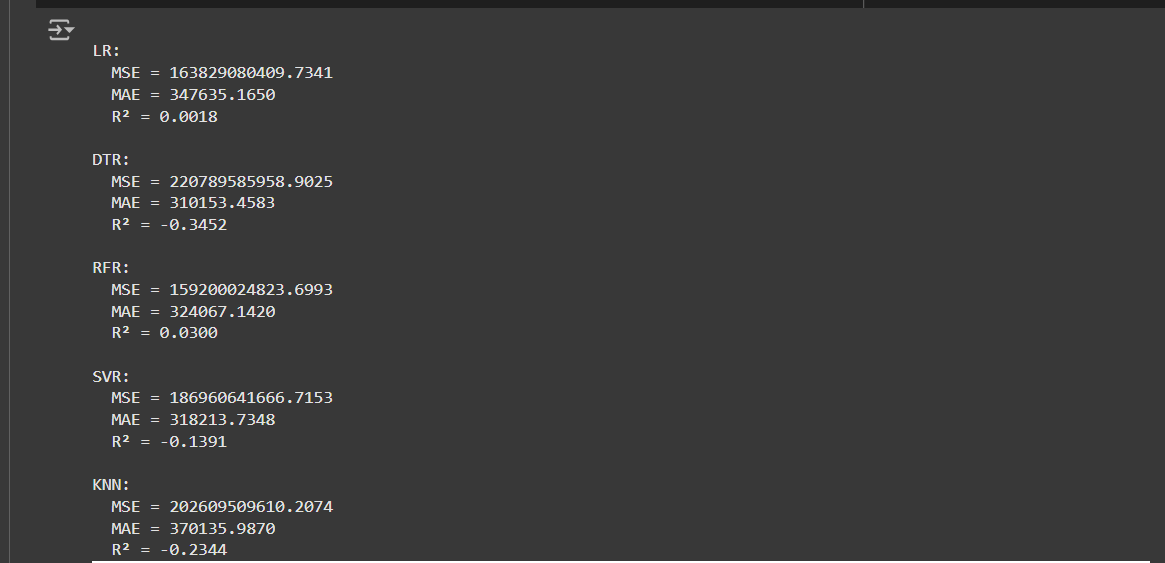
**STEP 4. MODELLING**

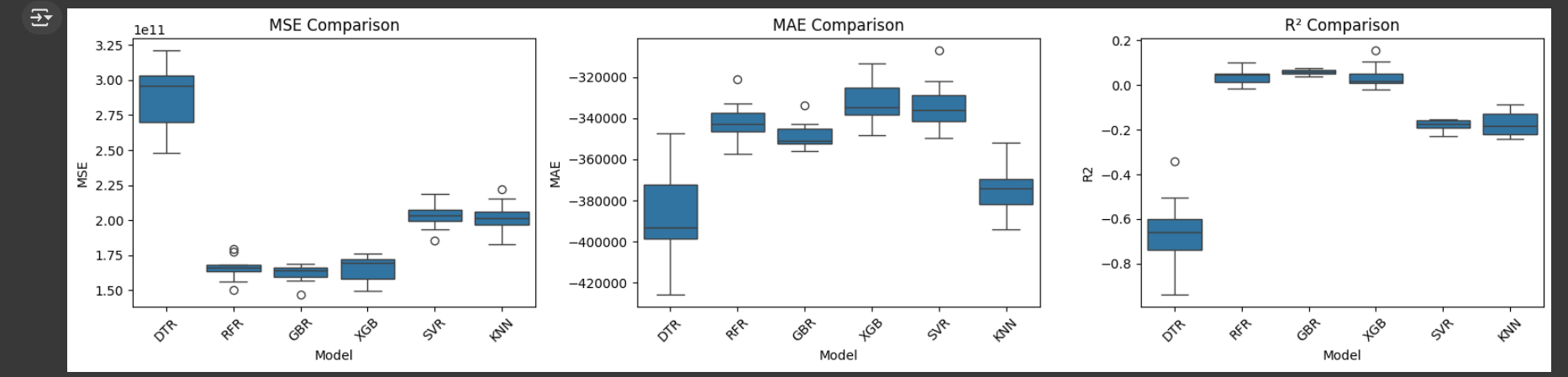
*1.   Standarlizing the data*

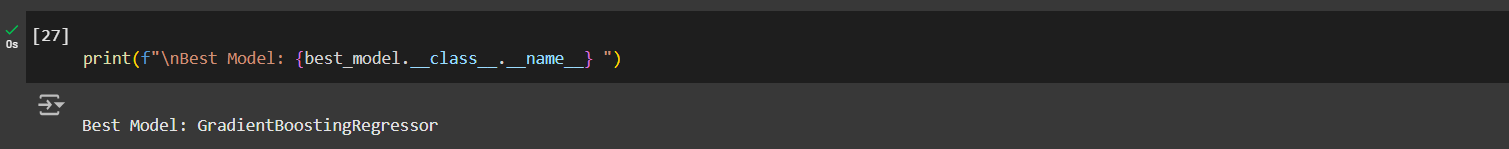
*2.   Defining the following models:*

*3.    Checking the residual score*





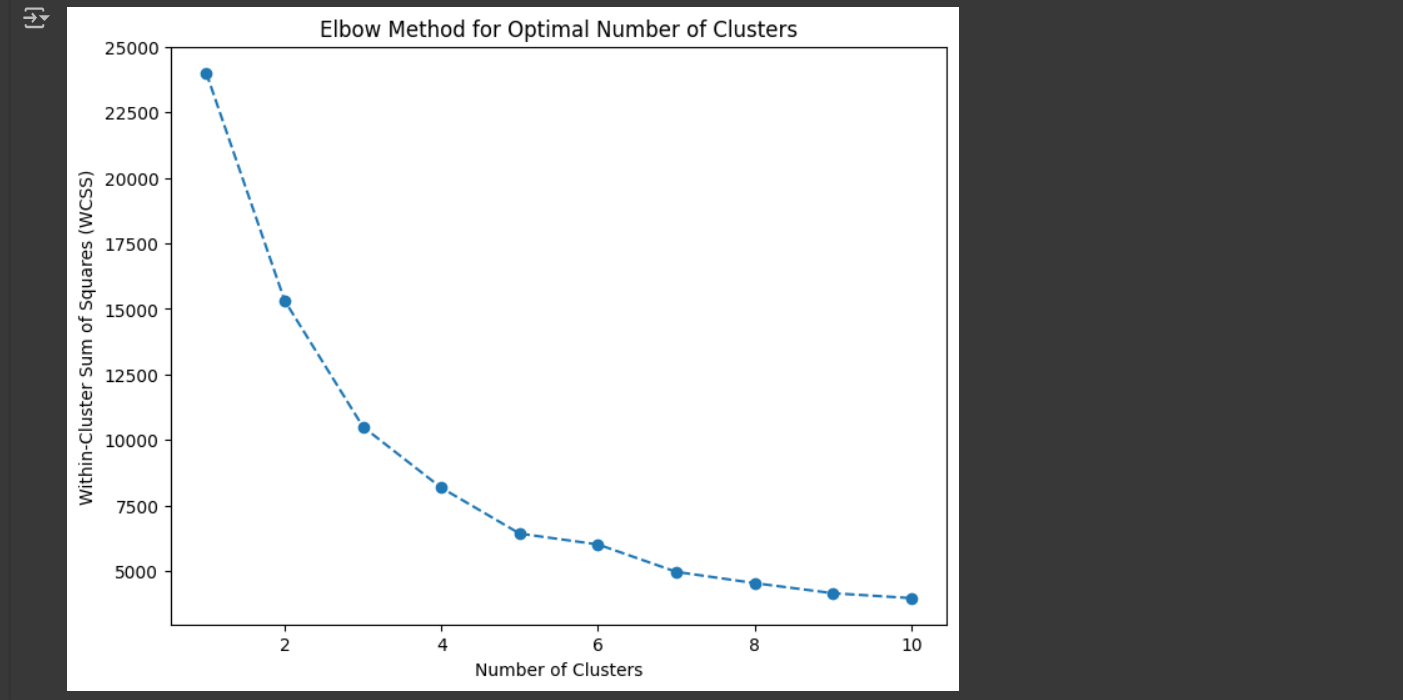


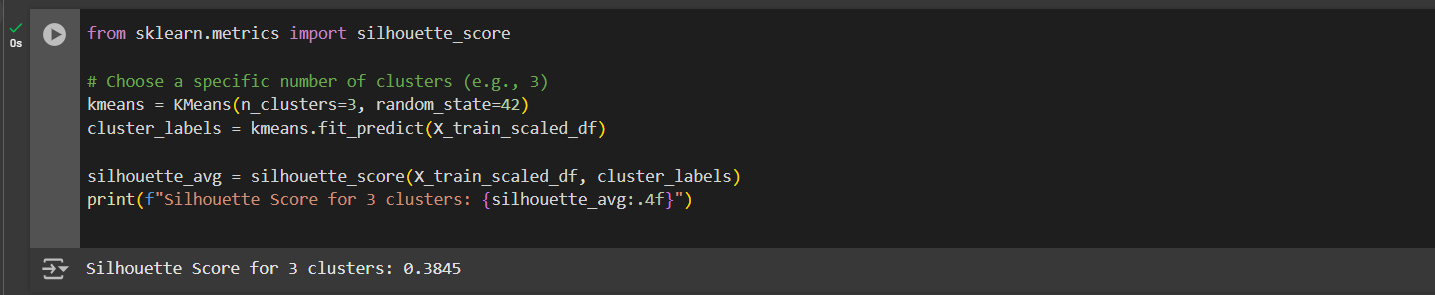


**STEP 6. CLUSTER MEANS AND SILHOUTTE SCORE**

*1.   Optimal number of clusters*

*2.   Silhoutte score*



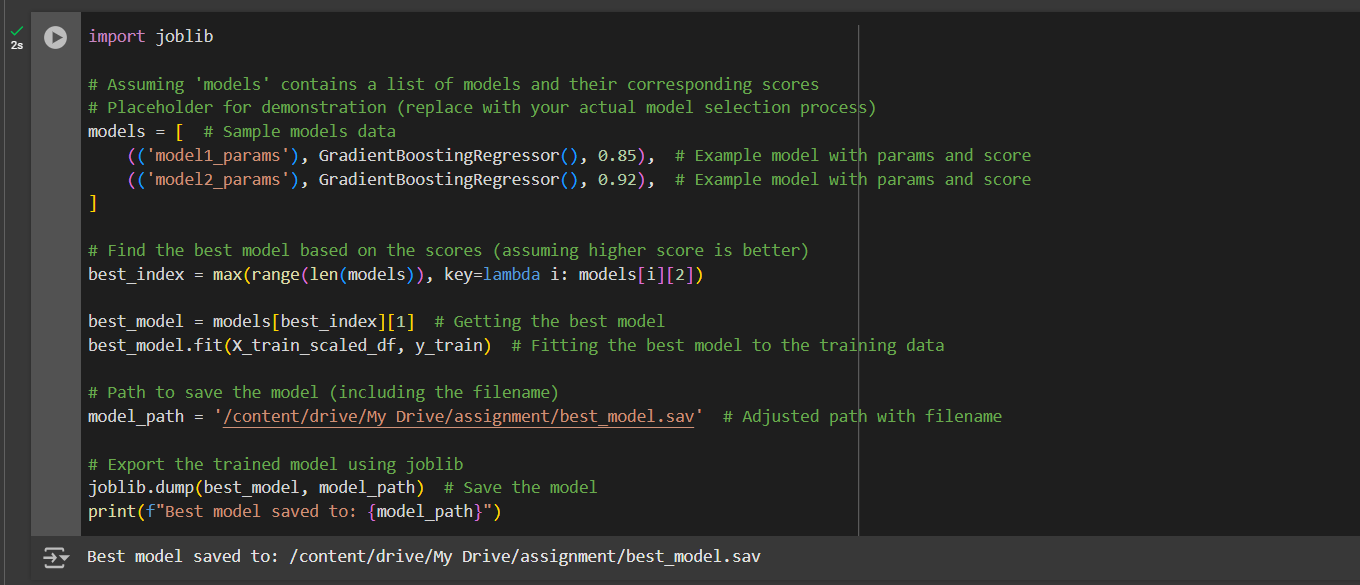


**STEP 6. BEST MODEL FOR PREDICTION**

*1.   Checking the maximum and minimum range for our input*

*2.   Submitting a test data (input) and predicting the total yield based off our best model*

*3.   Saving our best model into the google drive folder*

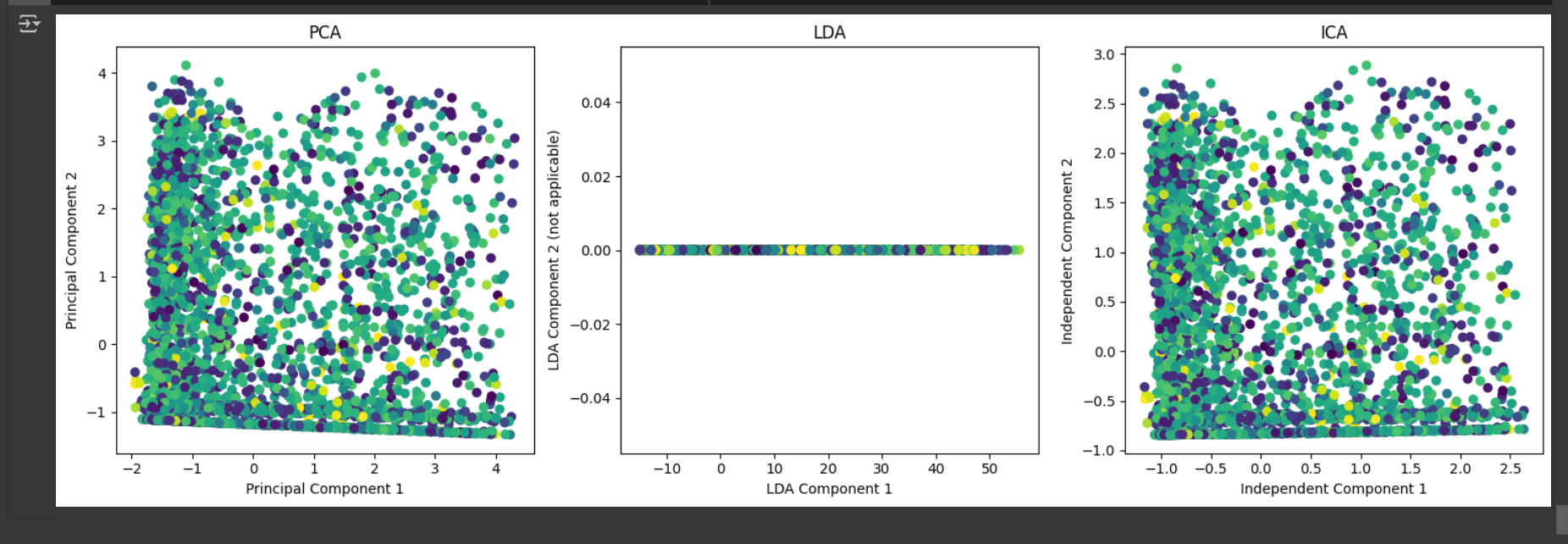


**STEP 7. VISUALIZATION FOR DIMENSIONAL ANALYSIS**

*1.   Applying PDA algorithm*

*2.   Applying LDA algorithm*

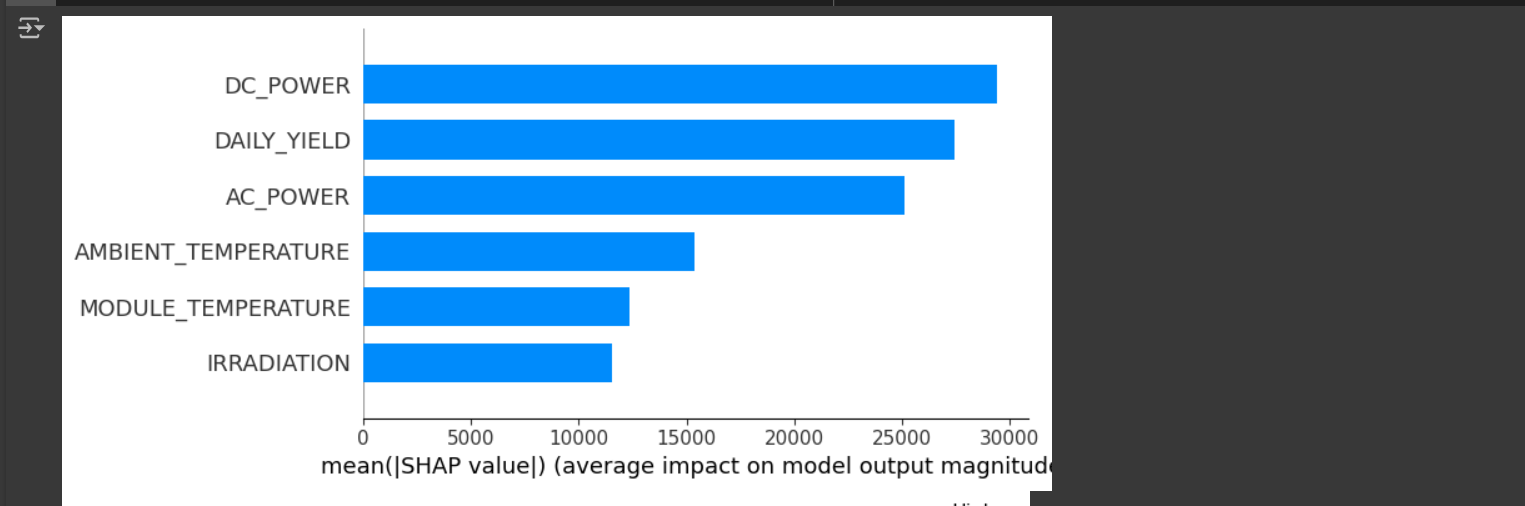
*3.   Applying ICA algorithm*

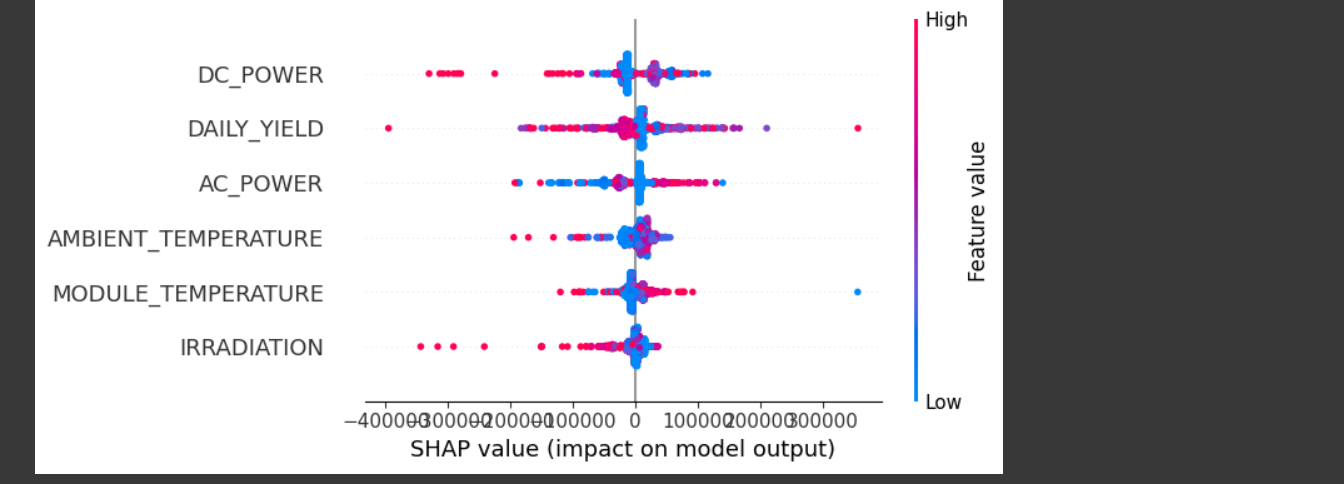


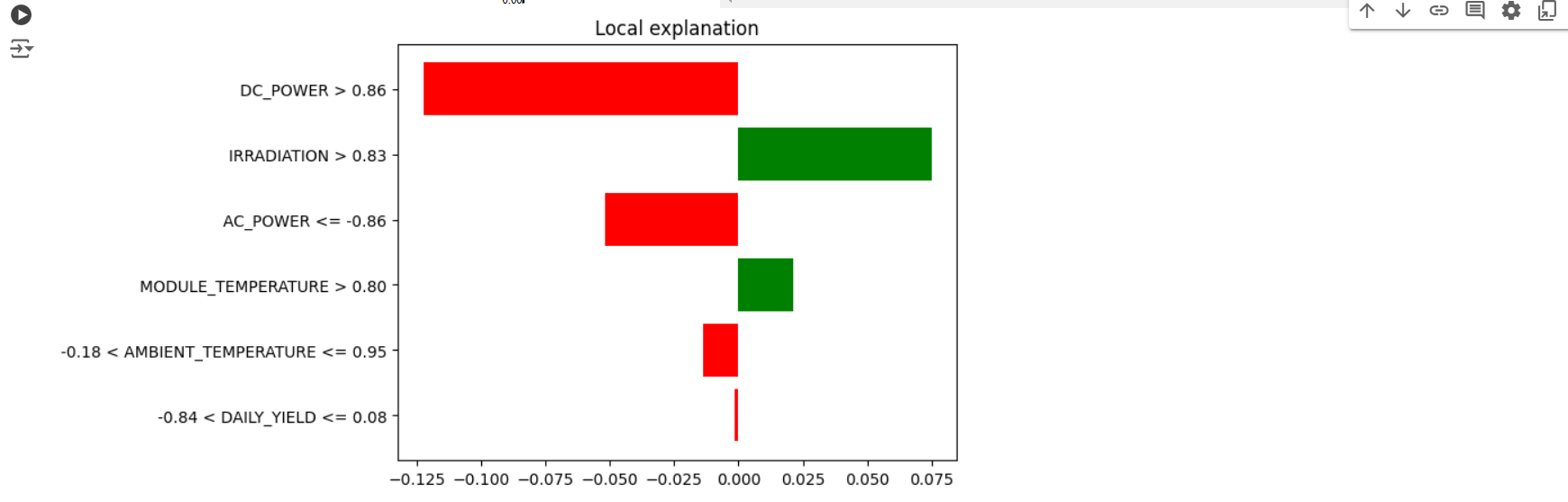
**Step 8. Understanding our model with the help of Explainable AI models**

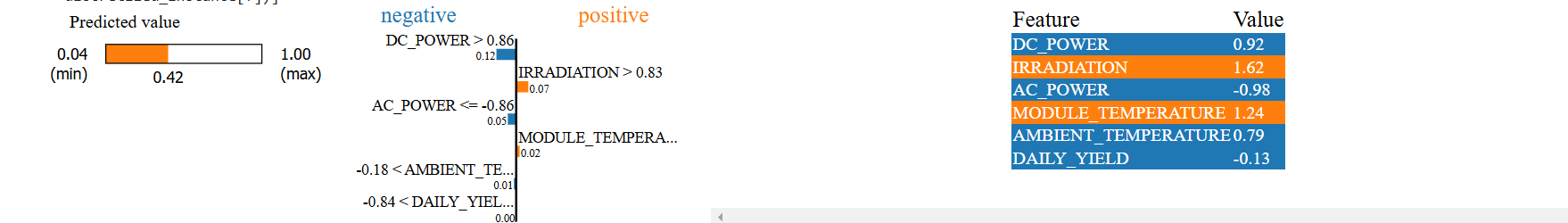
*1.   SHAP*

*2.   LIME*









**Practical Applications**

The model for predicting solar energy output has practical uses in the renewable energy industry as it supplies the operators, energy companies and policymakers involved in energy production with the necessary foresight. Such predictions ensure effective and efficient energy production. The main advantages are the following:

1. **Higher Efficiency:** When the operators are able to determine the amount of energy to be generated accurately, they can change the production plans in such a way as to optimize efficiency and reduce wastage of energy resources.
2. **Cost Saving:** Factors such as overproduction or underproduction have great operational cost implications due to the need to install energy storage systems or balance the power grid.
3. **Improved Grid Orchestration:** Accurate analyses also help install solar power plants scattered all over the grid as their supporting roles, enhancing grid stability and reducing fossil energy uses.
4. **Sustainability Orientation**: The constructs of the solution endorse energy companies to attain the sustainability agenda as they adopt the low carbon emission options which make solar energy competitive in the energy mix. Such a predictive solution also has a significant impact to other sectors that are not directly associated with renewable energy.

For instance, hospitals that mostly rely on renewable energy sources for emergency needs can be assured of reliability in supply due to the accurate estimation of yield. This model also has applicability such as in government agencies by which its use can determine the performance of solar PV capacity of a nation and thus provide guidance on renewable energy’s policies and investment opportunities.

**Key Takeaways and Challenges**

Through this project, I gained insights into:

* **Technical Skills**: Enhanced understanding of data preprocessing, feature selection, and working with regression models to address real-world problems.
* **Conceptual Insights**: The project emphasized the connection between weather patterns and energy production, illustrating how ML models can leverage these insights.

Key challenges included:

1. **Data Quality**: Cleaning and preprocessing data to handle missing values and inconsistencies.
2. **Model Accuracy**: Initial predictions were suboptimal, requiring model experimentation and hyperparameter tuning to improve accuracy.

**Acknowledgements**

I am grateful for my mentor’s guidance throughout this project, helping me understand core concepts in machine learning and solar energy forecasting. I also appreciate my peers for reviewing my work and providing valuable feedback.

**References**

1. Anikannal, "Solar Power Generation Data," Kaggle, [Kaggle link](https://www.kaggle.com/datasets/anikannal/solar-power-generation-data).
2. J. Brownlee, "Machine Learning Mastery," [ML Mastery](https://machinelearningmastery.com/)
3. Scikit-learn Documentation, [Scikit-learn link](https://scikit-learn.org/stable/documentation.html)
4. VanderPlas, "Data Science Handbook," O’Reilly Media.
5. Springer, "Renewable Energy Forecasting," Springer Publications, 2016.

* [*GOOGLE DRIVE FOLDER*](https://drive.google.com/drive/folders/1AR4G3YjDdZSGfRpAtZKXJQ8EXupjzynV?usp=sharing)
* [*GITHUB LINK*](https://github.com/vatssomya/Assignment_1)

**Name:** Somya Vats

**Batch:** 11

**SAP:** 500119012

**Roll no:** R2142230349