Predicting Renewable Energy Production Based on Climate Conditions

Yakub Reddy Adla (ya66480n@pace.edu)

# 

# INTRODUCTION

Renewable energy adoption demands precise solar energy production forecasting since it maintains power grid stability and wisely manages energy distribution across networks. Solar energy production shows inherent changes because it responds to several weather elements, including solar irradiance along with temperature conditions and humidity levels, and the presence of clouds and yearly progression throughout time. The goal of this research is to build a solar energy forecasting model using combined weather and temporal pattern data and construct an interactive system for continuous predictions and historical visualisation. Solar\_weather.csv contains 196,776 records about solar energy production, which were measured at 15-minute intervals between January 1, 2017 and August 31, 2022. The data analysis examines only the initial 1,000 records because computer capacity limits allow such processing. These records represent the start of 2017. This research outlines the approach used while showing the EDA results and predictive modelling outcomes alongside implications analysis that leads to a practical interactive dashboard creation.

# LITERATURE REVIEW

Renewable energy integration into power systems and energy dependability establishment requires precise predictions of renewable energy outputs from weather-dependent factors. Lack of advance forecasting methods is necessary because renewable energy resources such as solar and wind create inherent prediction difficulties caused by atmospheric conditions.

The elements which affect solar radiation at Earth's surface include cloud cover, aerosols, and wavapour thatthwater vapour that determine atmospheric conditions as a primary solar energy output factor. PV system output forecasting depends heavily on precise predictions of these characteristic variables. Scientists have proved that linking solar energy prediction algorithms with weather forecasting models enhances the accuracy level of solar power generation forecasts (Meenal, R., et al, 2022).

Wind energy production depends on the direction and speed of wind among meteorological elements. Accurate forecasting models are essential because wind patterns remain difficult to predict; therefore, they produce unreliable power output expectations. Wind energy production connected to weather patterns has been studied using two machine learning methods known as deep neural networks DNNS as well as artificial neural networks (ANNS. The models identify historical patterns by analysing previously collected data to generate precise forecasting outcomes (Gaamouche, R., et al., 2022).

Machine learning allowed the forecasting of renewable energy to advance due to its power to examine large datasets and recognise complex relationships. Multiple machine learning techniques have merged into hybrid models for improved prediction accuracy. Randomisation-based machine learning methods achieve top ratings for renewable energy prediction models because they deliver both quick computational speed and advanced accuracy (Del Ser, J., et al., 021).

The generation of renewable energy demands probabilistic forecasting models so experts can deal with prediction uncertainties. Through Gaussian Process Regression (GPR) models probabilistic predictions of solar production become possible while meteorological variables enter the model as uncertain inputs. The probabilistic models evaluate uncertainty through a predictive framework which enhances grid management decisions (Najibi, F., Apostolopoulou, D., & Alonso, E., 2020).

Renewable energy output depends heavily on the seasonal changes in climate patterns. Research conducted by scientists confirmed that regional solar and wind production anomalies result from large-scale atmospheric circulation patterns that teleconnection indices can track. Experts confirm that models that predict renewable energy generation based on seasonal variables show positive skill levels (Lledó, L., et al., 2022).

Energy management systems need advanced forecasting models to achieve the highest possible use of renewable resources. The reduction of fossil fuel usage alongside enhanced grid stability comes from better scheduling of energy delivery and storage which accurate forecasts make possible. Sustainable growth of renewable energy sectors depends on the development of adaptable forecasting models because weather patterns are changing because of climate change effects.

Knowledge of the complex interaction between weather conditions and power output remains vital for predicting renewable sources of energy generation according to environmental conditions. Power network integration of renewable energy sources became more successful because probabilistic modelling and machine learning technologies enabled significant improvements in forecasting accuracy. Active research and development work in this area helps tackle the problems that arise due to climate variations and supports long-lasting energy sustainability.

# METHODOLOGY

This study follows four stages, which start with data preprocessing, then continue with exploratory data analysis, after which predictive modelling occurs, and ends in interactive dashboard development. The database consists of 17 variables with the main outcome metric being Energy delta[Wh] (watt-hour energy production) and additional climate attributes GHI (Global Horizontal Irradiance), temp (temperature), humidity, wind\_speed, rain\_1h, snow\_1h, and clouds\_all in addition to time-based variables including hour, month, sunlightTime, and dayLength. A subset consisting of 1,000 data rows was chosen because it provided fast computation speeds without reducing analytical value.

## Data Preprocessing

The pandas library enabled data loading from the dataset, which showed through initial checks that all necessary attributes existed in the subset. The Time values exist in string format since they represent 15-minute time intervals. The subset included numerical features mainly as float64 and int64 types, which summed to 132.9 KB of memory usage. Through df.describe(), summary statistics revealed variable distribution, and analysts used histograms together with boxplots and correlation heatmap visualisation to study variable distributions and inter-variable relationships (Allal *et al*. 2024).

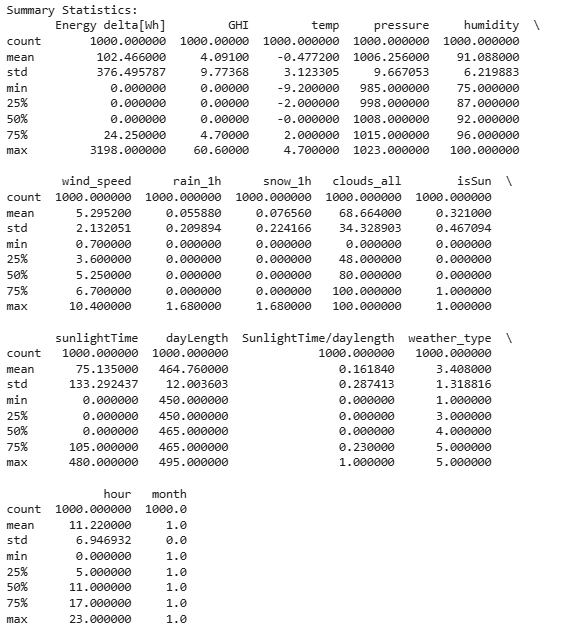


Figure 1: Dataset summary

## Exploratory Data Analysis (EDA)

Exploratory data analysis has been done to examine means, spread and the relationship between variables. The histograms indicated the distribution of the features, while box plots gave an outlook on the outliers. Pearson’s correlation coefficient was used to test the degree of association between feature variables, and a heatmap was constructed for cross-comparison of the features using the corresponding correlation coefficients. Finally, to quantify the relevance of features, a Random Forest model was used in conjunction with SHAP (Shapley Additive exPlanations) values (Alazemi et al*.* 2024).

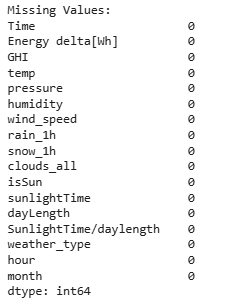


Figure 2: Checking null values

## Predictive Modeling

There were three models which we used to predict Energy delta[Wh], and they include neural network (LSTM), ARIMA model, and Gradient Boost Regressor models. Regarding the Gradient Boosting model, the following steps were followed:

The weather data selected to be used features include the global horizontal irradiation (GHI), temperature, pressure, humidity, wind speed, amount of rain per hour and snowfall per hour, cloudiness in the next hour, the hour and month of the day (Bin Abu Sofian *et al*. 2024).

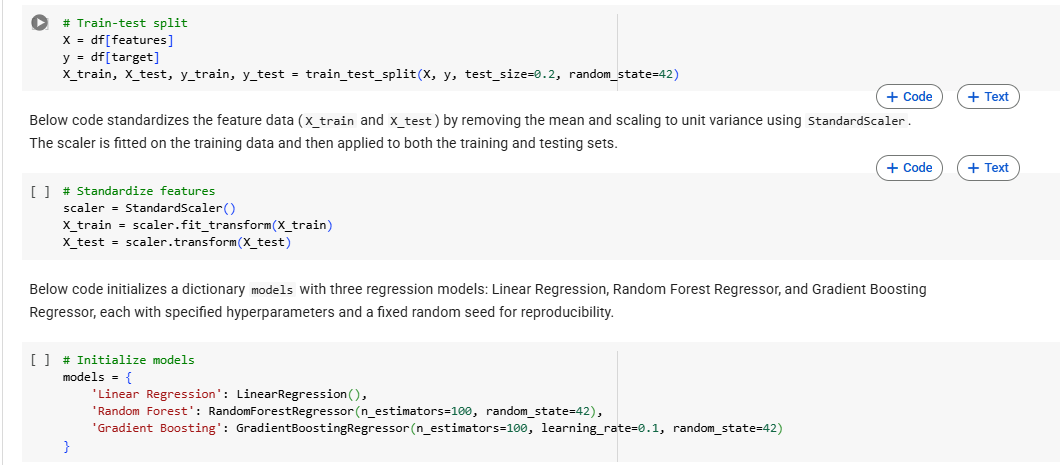


Figure 3: Model development

Preprocessing: Due to differences in the range of features, StandardScaler was used to transform the data to have a zero mean and unit variance.

Model Training: For the GradientBoostingRegressor, parameters such as the number of estimators was set as 100, learning rate 0.1 and the random state 42 for reproducibility.

Evaluation: This was done using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), and the results were compared with LSTM and ARIMA models.

While the implementation of LSTM and ARIMA model training code was missing from the code notebook, their predictions were plotted, and the performance metrics for the ARIMA model were given.

## Interactive Dashboard

When using the Dash framework for developing an interactive dashboard, it was used to support a real-time energy production forecast. The interface contains input bars for the values of GHI, temp, humidity, wind\_speed, hour and month, while the rest of the parameters are fixed with average values The Gradient Boosting model takes these inputs to forecast the energy production, and the relation between energy production and climate conditions is represented graphically by a Scatter Plot using Plotly.

# RESULTS AND DISCUSSION

## Exploratory Data Analysis

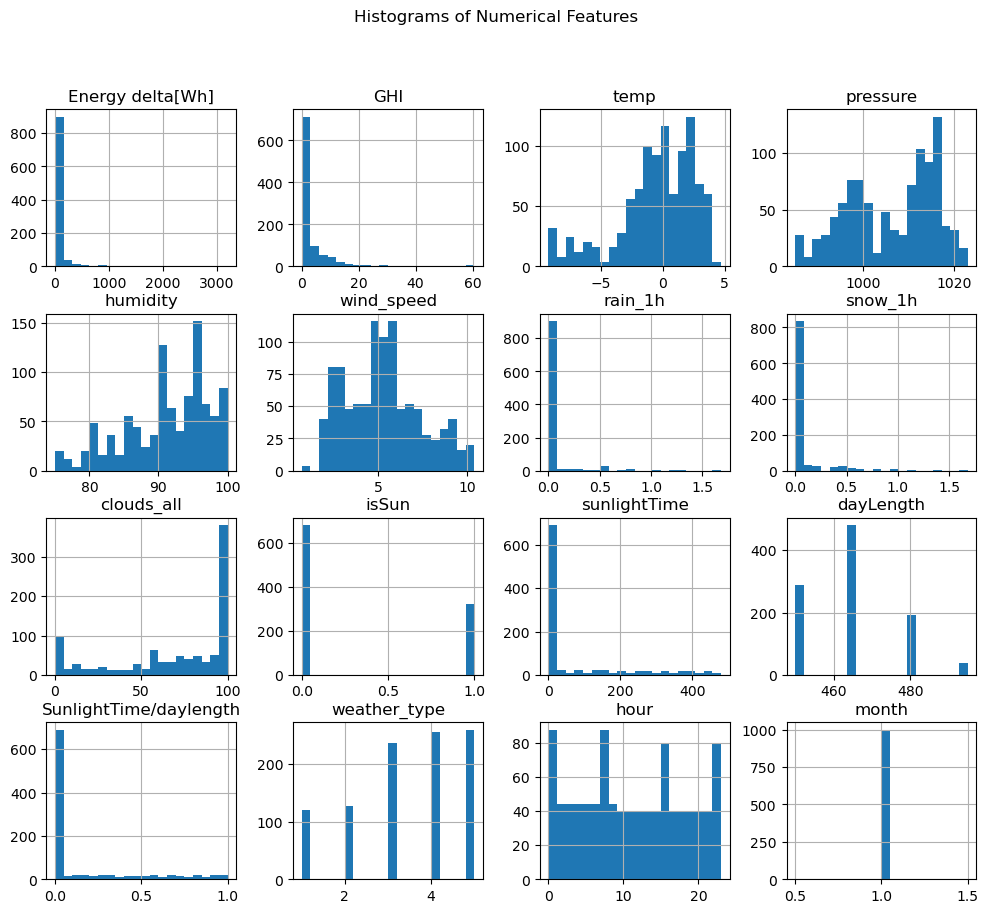


Figure 4: Histograms

Also, by summarising the data about Energy delta[Wh], the mean was demonstrated as 102.47 while the median was equal to 0 Wh, which means that the distribution is right-skewed. This is because the most populated epoch is the zero-energy periods, which may correspond to the nighttime or periods with clouds, while the maximum has been 3198 Wh in the solar production peak.

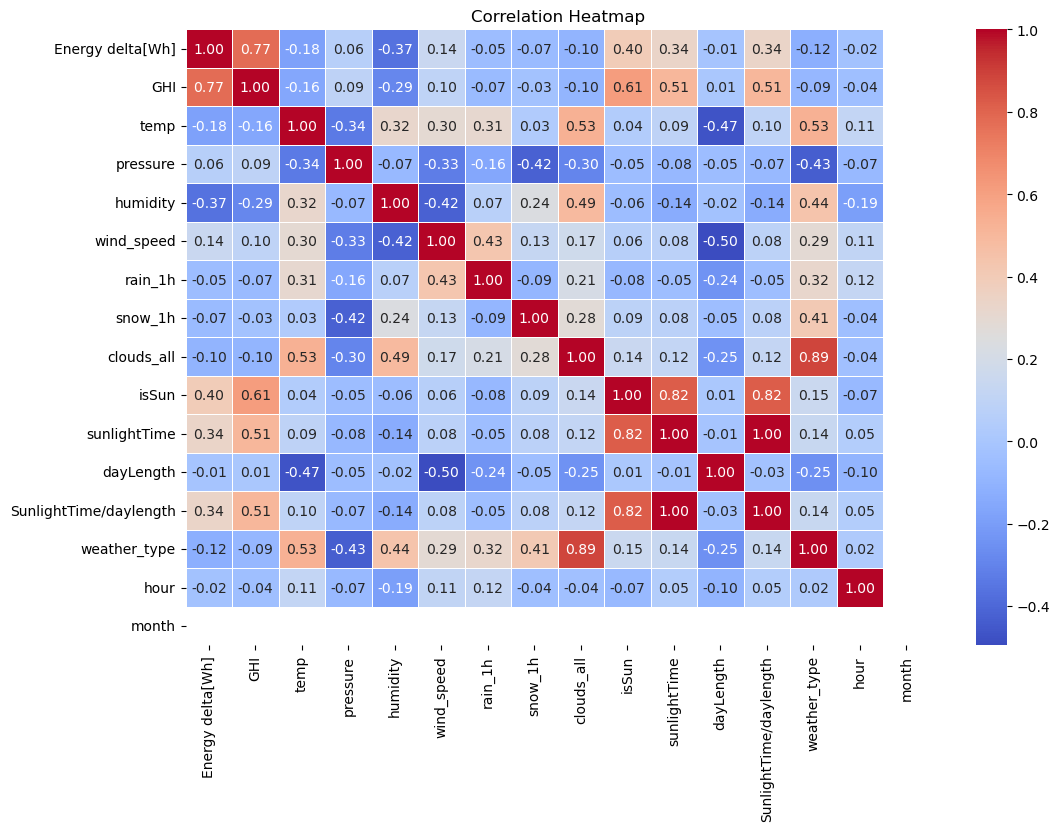


Figure 5: Correlation plot

Likewise, GHI (mean 4.09 W/m², median 0 W/m²) had positive skewness due to the relatively low amount of sunlight in the middle of a winter month – January 2017, with a mean temperature of -0.48°C and a high relative humidity mean 91.09%. Cloud cover defined by group ‘clouds\_all’ was relatively high and equal to 68.64% with the bi-modal shape (the maxima at 0 and 100), implying rather often overcast conditions, which reduce the solar energy yield.

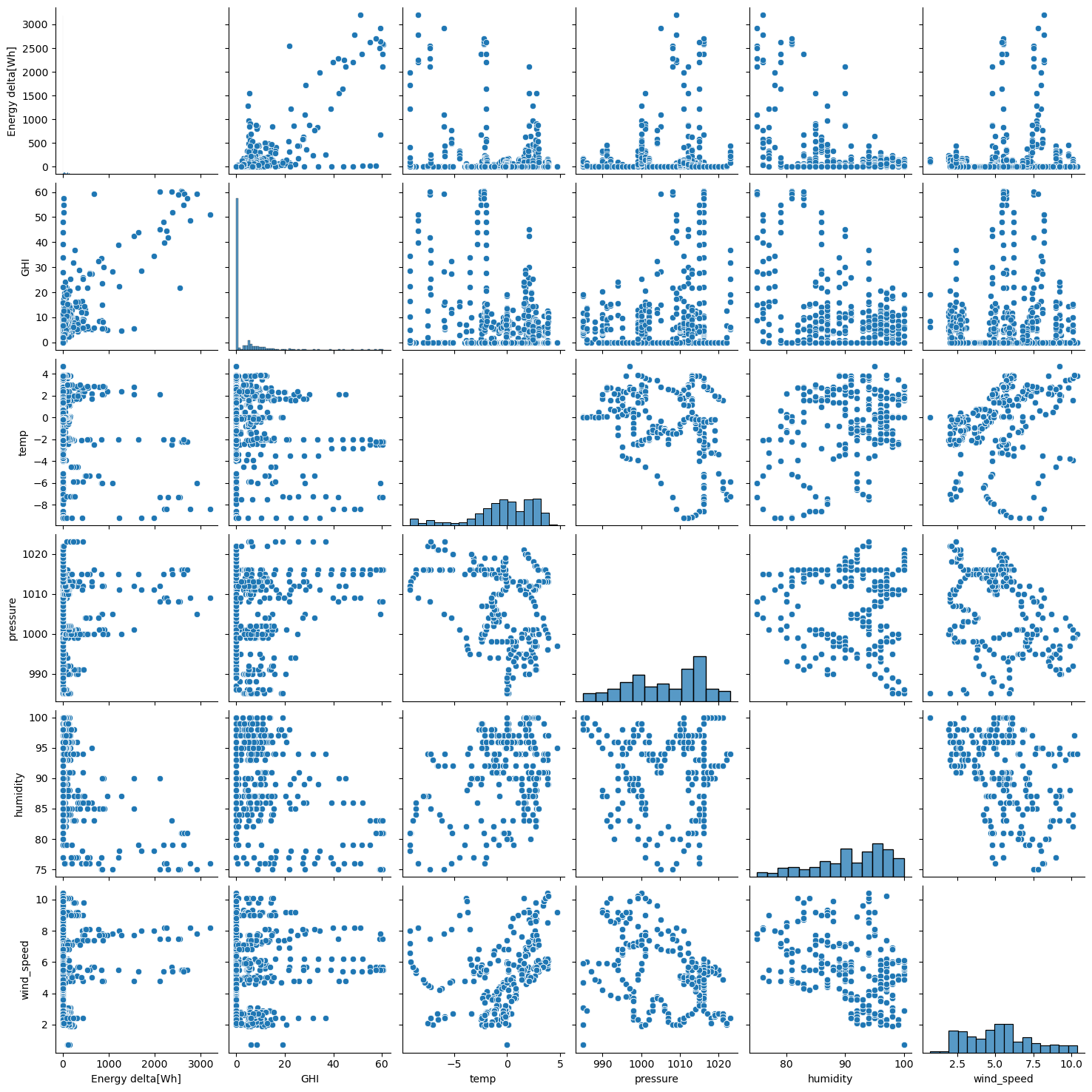


Figure 6: Pairplot

Histograms of energy production, daily Global Hands Index (GHI), and hourly rain and snow suggested that the data represented here are skewed, which was more apparent from the boxplots showing outliers in energy production and wind speed of more than 10 m/s and 500 Wh accordingly. The correlation heatmap presented in the paper confirmed a strong positive correlation between Energy delta[Wh] and GHI (0.77) because GHI is the measure of the amount of solar irradiance on the horizontally projecting plane during the periods of full sunrise to sunset. There were negative correlations with clouds all (-0.40) and humidity (-0.37), which means clouds and high humidity decrease the energy yield. Notably, the correlation between the feature sunlightTime and SunlightTime/daylength was 0.82, which means that the two are rather similar.

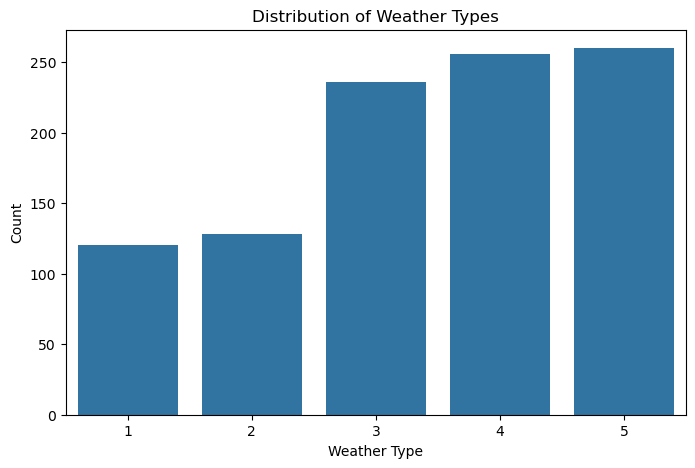


Figure 7: Distribution plot

Importance of each feature in Random Forest model:-Max GHI: 0.66 Max relative humidity: 0.17, time sunlight in day : 0.05 SHAP provided additional insights into feature importance for the predictions of model, and showed that GHI was the most important feature with the mean |SHAP value| of about 70. High GHI enhanced the predicted energy production whereas, high humidity and clouds\_all reduced it, which is in conformity to physical expectation.

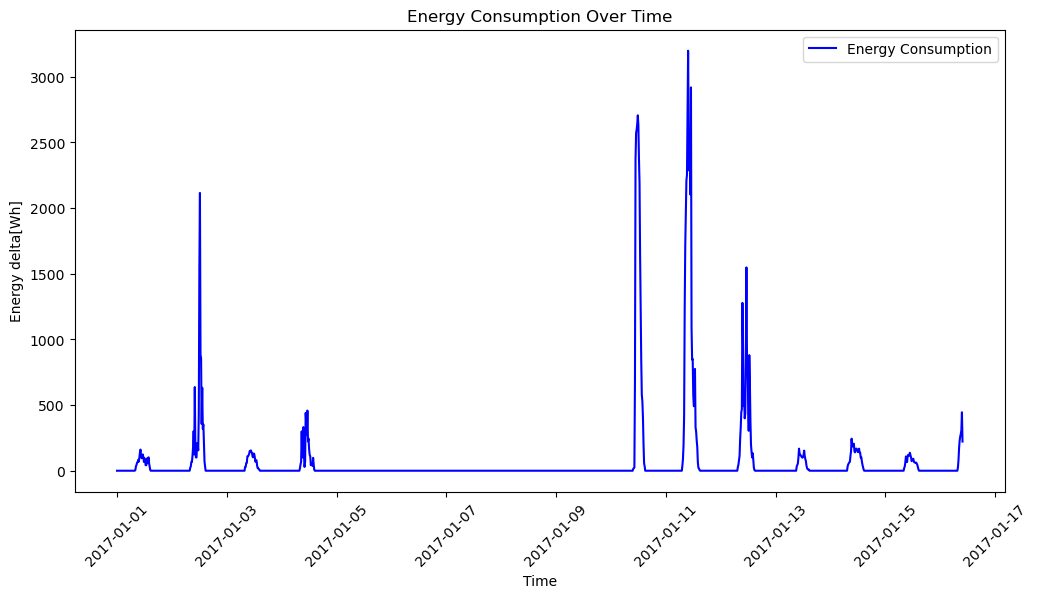


Figure 8: Energy consumption

As per the PCA analysis, the first four components have been found to contribute approximately 0.25, 0.23, 0.14, and 0.09, meaning that it would be beneficial to reduce the dimensionality as part of future research without much loss of information.

## Predictive Modeling

The Gradient Boosting model predicts the solar radiation level with moderate accuracy since it has an MAE of 41.98 Wh and an RMSE of 125.02 Wh. The Random forest model yielded better results (MAE = 32.77 Wh, RMSE = 113.18 Wh) as compared to the Linear Regression model MAE = 109.93 Wh, RMSE = 234.01 Wh). The electricity consumption prediction using the developed ARIMA model on the test dataset yielded an MAE of 36.81 Wh and an RMSE of 80.75 Wh, which were lower than the corresponding metrics for Gradient Boosting in terms of RMSE. While those LSTM and ARIMA prediction models showed overall accuracy, such as the line plot between actual energy production and the two models, short-term peak prediction was less accurate, presumably due to the small sample size, and there was no seasonality.

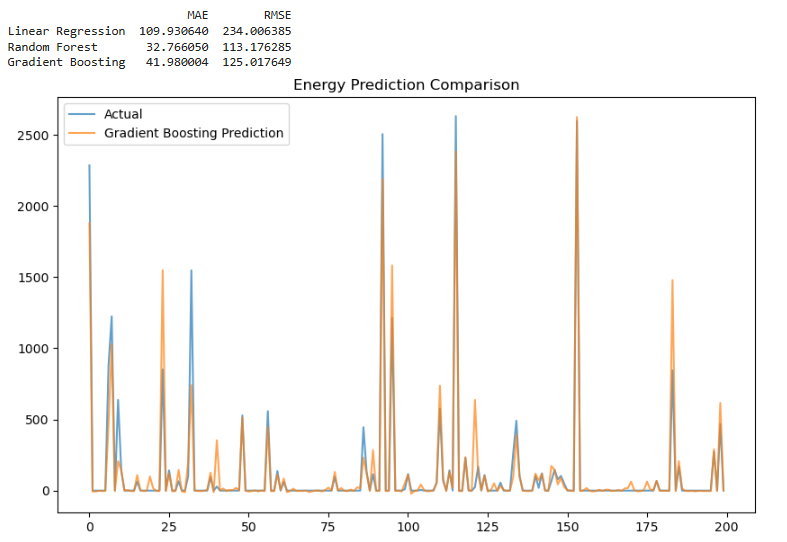


Figure 9: Model performance

This indicates that the ensemble approach and time series models are viable solutions for this sort of information analysis. However, due to its interpretability, the Gradient Boosting was chosen for the dashboard in combination with its capability to handle non-linear correlative fields, and the model was chosen.

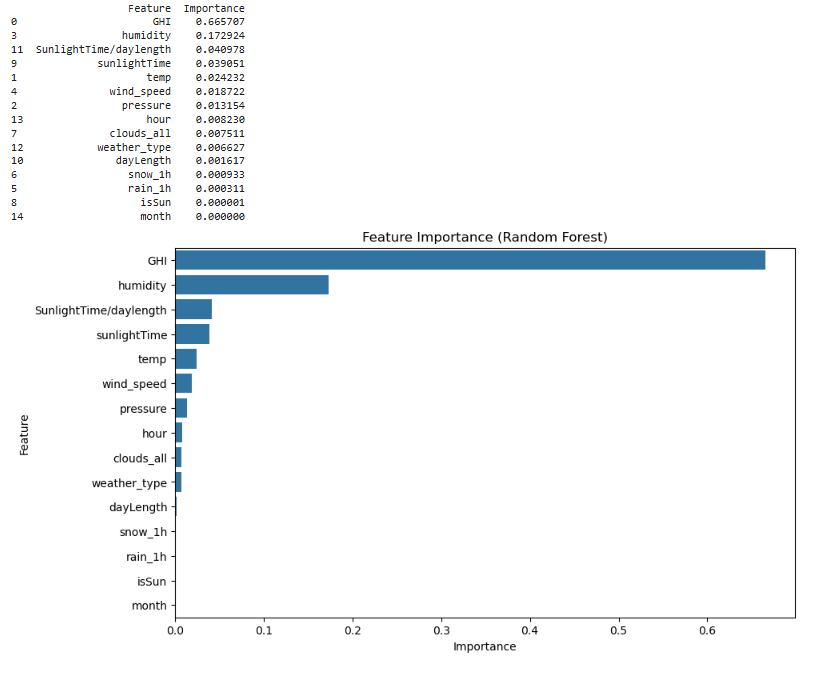


Figure 10: Feature importance

## Interactive Dashboard

The energy production is introduced simply through a dashboard developed using Dash of Python. The parameters that are plotted concerning GHI are temperature, humidity, wind speed, hour, and month, while the other parameters, such as pressure and rain 1h, stay at their mean value. On this dashboard, you will also find the predicted energy production, for example, 123.45 Wh; as well as the Energy delta[Wh] compared to the GHI, with points’ colour depending on the temperature and dots’ area based on the wind speed. It reveals the fluctuation in energy production due to solar irradiance and weather conditions in a manner that is easy to understand.

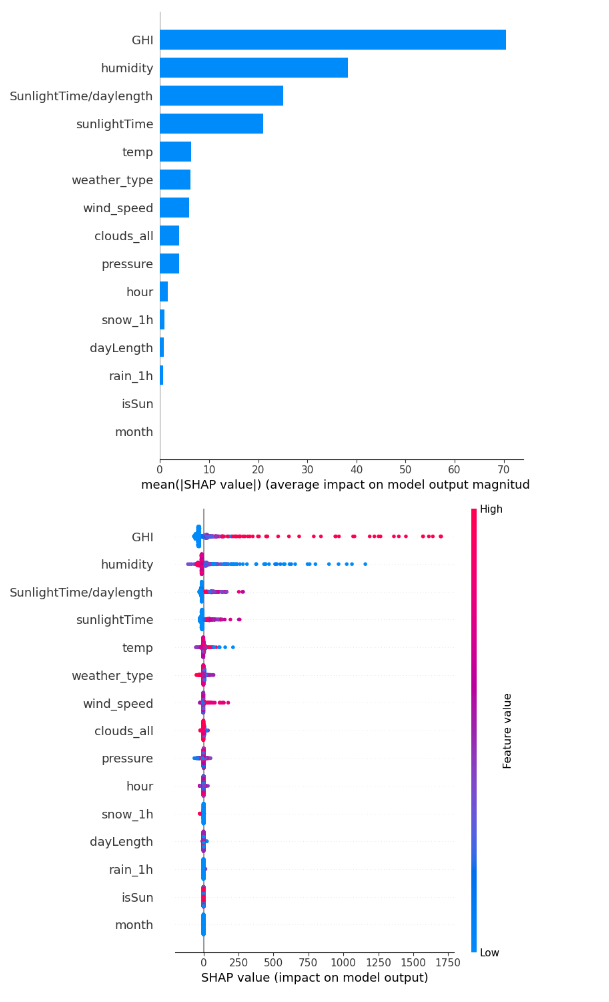


Figure 11: SHAP values

Probably the use of mean values in some of the features of the dashboard makes it easier to predict but may not be efficient in situations where among the observed values there are more variations from the mean. For example, assuming that clouds\_all is constant at mean level of 68.64% conceals the natural presence of two peaks and a trough in cloud cover which has significant potential of affecting production levels.

# CONCLUSION

The angle relation analysis and the wind direction and speed, and the sky condition achieve a moderate design accuracy with Gradient Boosting Regression, where the MAE is equal to 41.98 Wh and the RMSE is equal to 125.02 Wh, with the usage of the collected data subset of the solar\_weather.csv. Some of the findings include GHI to be a significant predictor, and the effects of cloud cover and relative humidity to reduce energy generation. Feature importance and SHAP followed these results by identifying GHI as the most influential factor. Thus, the interactive dashboard has a useful application for forecasting potential consequences of weather conditions and temporal effects influencing energy production.

The subdataset of 1,000 rows and containing only the data gathered in January 2017 does not include seasonal fluctuations, which could influence the functioning of the model. The absence of LSTM and ARIMA training code in the notebook also makes the results unrepeatable; however, the dashboard work assumes that all the features have the mean values of them, which is far from reality. There should be work to be done with the entire provided framework, work with the seasonal data and implement the ability to change all the values in the dashboard in real-time. Further, one could tune more of the hyperparameters or expand the features incorporated into the model, such as the lagged variables. Nevertheless, through the results of this study, one can see the benefits of the data-driven approach to solar energy forecasting to increase the rate of using renewable energy resources.

# References

Del Ser, J., Casillas-Perez, D., Cornejo-Bueno, L., Prieto-Godino, L., Sanz-Justo, J., Casanova-Mateo, C., & Salcedo-Sanz, S. (2021). Randomisation-based Machine Learning in renewable energy prediction problems: Critical literature review, new results and perspectives. In *arXiv [cs.LG]*. http://arxiv.org/abs/2103.14624

Gaamouche, R., Chinnici, M., Lahby, M., Abakarim, Y., & Hasnaoui, A. E. (2022). Machine learning techniques for renewable energy forecasting: A comprehensive review. In *Green Energy and Technology* (pp. 3–39). Springer International Publishing.

Lledó, L., Ramon, J., Soret, A., & Doblas-Reyes, F.-J. (2022). Seasonal prediction of renewable energy generation in Europe based on four teleconnection indices. In *arXiv [physics.soc-ph]*. http://arxiv.org/abs/2202.02258

Meenal, R., Binu, D., Ramya, K. C., Michael, P. A., Vinoth Kumar, K., Rajasekaran, E., & Sangeetha, B. (2022). Weather forecasting for renewable energy systems: A review. *Archives of Computational Methods in Engineering. State of the Art Reviews*, *29*(5), 2875–2891. https://doi.org/10.1007/s11831-021-09695-3

Najibi, F., Apostolopoulou, D., & Alonso, E. (2020). Gaussian Process Regression for probabilistic short-term solar output forecast. In *arXiv [stat.AP]*. <http://arxiv.org/abs/2002.10878>

Allal, Z., Noura, H.N., Salman, O. and Chahine, K., 2024. Machine learning solutions for renewable energy systems: Applications, challenges, limitations, and future directions. Journal of Environmental Management, 354, p.120392.

Alazemi, T., Darwish, M. and Radi, M., 2024. Renewable energy sources integration via machine learning modelling: A systematic literature review. Heliyon, 10(4).

Bin Abu Sofian, A.D.A., Lim, H.R., Siti Halimatul Munawaroh, H., Ma, Z., Chew, K.W. and Show, P.L., 2024. Machine learning and the renewable energy revolution: Exploring solar and wind energy solutions for a sustainable future including innovations in energy storage. Sustainable Development, 32(4), pp.3953-3978.