# Abstract:

This report presents an analysis of weather dataset. The goal of this project is to help the power generation station to gauge the solar energy prediction and save the fossil fuel-based energy generation based on model prediction. A variety of models were trained on the dataset, including ensemble methods as well as linear models. The performance of each model was evaluated using metrics such as adjusted R2, R2, and RMSE. The best performing models were then selected and tested using cross-validation to estimate their accuracy.

The results showed that the ensemble methods, particularly the Histogram Gradient Boosting and LGBM Regressor, outperformed the linear models. The Histogram Gradient Boosting model achieved an adjusted R2 of 0.93, an R2 of 0.93, and an RMSE of 84.33, while the LGBM Regressor achieved similar scores. The performance of these models was further confirmed through cross-validation, where they achieved an average R2 of 0.89 and 0.90, respectively, on the training data.

Overall, the analysis demonstrated the effectiveness of ensemble methods for predicting solar radiations. These models can help power generation stations to make more informed decisions about how much electricity to generate on gasoline, as well as aid policymakers in understanding trends in the weather forecasting.

Table of Contents:

[Abstract 1](#__RefHeading___Toc1794_484607107)

[Introduction: 2](#__RefHeading___Toc1531_484607107)

[Background: 3](#__RefHeading___Toc1533_484607107)

[Solution and Discussion: 4](#__RefHeading___Toc1535_484607107)

[Dataset Description: 4](#__RefHeading___Toc1537_484607107)

[Preprocessing: 6](#__RefHeading___Toc1539_484607107)

[Visualization of Features: 7](#__RefHeading___Toc1541_484607107)

[Features Selection: 9](#__RefHeading___Toc1543_484607107)

[Model Selection: 12](#__RefHeading___Toc1545_484607107)

[Results: 14](#__RefHeading___Toc1547_484607107)

[Model Visualization: 14](#__RefHeading___Toc1549_484607107)

[Utilization of Model: 15](#__RefHeading___Toc1551_484607107)

[Conclusion: 15](#__RefHeading___Toc1553_484607107)

[References: 15](#__RefHeading___Toc1555_484607107)

# Introduction:

Humans run on blood while states run on energy. Developed states like Finland and Norway are almost entirely devoted to renewable energy. While developing nations can only think of these ways of producing energy. With the increasing demand for renewable energy sources in advanced states, solar power has become a go-to source for almost every country. However, solar energy production efficiency is highly dependent on environmental factors. These factors include weather conditions, the position of the sun, the duration of summer or sunny days, the average temperature in sunlight, etc. In addition to the above factors, solar panels' position is critical in predicting solar energy production. In recent years, machine-learning techniques have been applied to solar energy prediction, allowing for more accurate and efficient forecasting of solar energy production.

In this project, I developed a machine-learning model to predict solar energy production based on weather data and other relevant factors. This project aims to predict solar power prediction, keeping other factors like the number of solar plates, solar plants, etc constant, for power generation plants so that generation of energy from non-renewable resources like gasoline is avoided to the greatest extent possible. I utilize a dataset containing hourly solar radiation and meteorological data for a specific location and apply machine-learning algorithms to build a model that can accurately regress solar energy production for that location. I evaluate the performance of my used machine-learning technique using various metrics and compare results with techniques used in literature.

The rest of this report is organized as follows: In the next section, the background of this project is discussed. In the “Solution and Discussion Section” of this report, I will discuss the literature review on solar energy prediction and the use of machine-learning in this field. This section also includes dataset description and preprocessing I have done to the dataset with a sub-heading "dataset description”. The methodology for building the machine-learning model, including feature selection and model selection is also part of the “Solution and Discussion Section”.

Furthermore, I evaluate the performance of my model and compare it to other models in addition to discussing the results. The last section is a bibliography where references are provided.

# 

# Background:

Solar energy is a rapidly growing renewable energy source that has the potential to revolutionize the energy industry. It is a clean, abundant, and sustainable source of energy that has many advantages over traditional gasoline-based energy sources. Solar energy is also becoming increasingly affordable and cost-competitive with other forms of energy, which has led to a significant increase in its adoption worldwide.

However, one of the major challenges associated with solar energy production is its dependence on environmental factors, such as weather conditions and the position of the sun. These factors can significantly affect the amount of solar energy that is produced, which can in

turn impact the efficiency and profitability of solar power plants. To address this challenge, researchers have developed various methods for predicting solar energy production using weather data and other relevant factors.

Machine-learning techniques have recently emerged as a promising tool for predicting solar energy production. Machine-learning models can use historical weather data and other relevant information to predict solar energy production with a high degree of accuracy. These models can also be used to optimize the placement and performance of solar panels, which can further improve the efficiency and profitability of solar power plants.

Since the world has an inclination for renewable power generation, there have been several studies in recent years on predicting solar energy production using various machine-learning algorithms. These studies have focused on different aspects of solar energy production, such as forecasting, optimization, and performance evaluation of solar power plants. Major tendency has been for prediction of solar radiations and forecasting.

One of the most popular approaches for solar energy forecasting is the use of Artificial Neural Networks (ANNs). ANNs have been widely used for solar energy prediction due to their ability to learn complex patterns from historical data and make accurate predictions. For instance, in a study conducted by Hasan et al (Alkahtani, Aldhyani and Alsubari, 2023). They used solar radiation prediction to solve this challenge. CNN-LSTM had already been used for this task. They examined the meteorological data combined with CNN-LSTM for forecasting purposes. They collected the weather data from a NASA meteorological station, and it included details such as the current temperature, the relative humidity, and the speed of the wind. Their findings demonstrated solar radiation is highly correlated with both temperature and radiation. They also composed that CNN-LSTM is the best-performing model in comparison with other models.

In addition to ANNs, other machine-learning algorithms such as Support Vector Machines (SVM), Random Forest, and Decision Trees have also been used for solar energy prediction. In a study by Fei-Wang et al (Callumdownie, 2017), a comparative study was conducted between SVM and Random Forest models for predicting solar energy production using weather data. The study concluded that SVM outperformed Random Forest in terms of accuracy and efficiency.

In this project, I aim to develop a machine-learning model for predicting solar energy production based on weather data and keeping other relevant factors as assumed. My goal is to save gasoline-based energy production by predicting the energy being produced by solar panels. I will use a dataset containing solar radiation and meteorological data for a specific location to develop and evaluate my machine-learning model. By doing so, I hope to contribute to the power sector of a state by predicting solar-based energy and helping to facilitate the transition to a more sustainable and renewable energy future.

# 

# Solution and Discussion:

## Dataset Description:

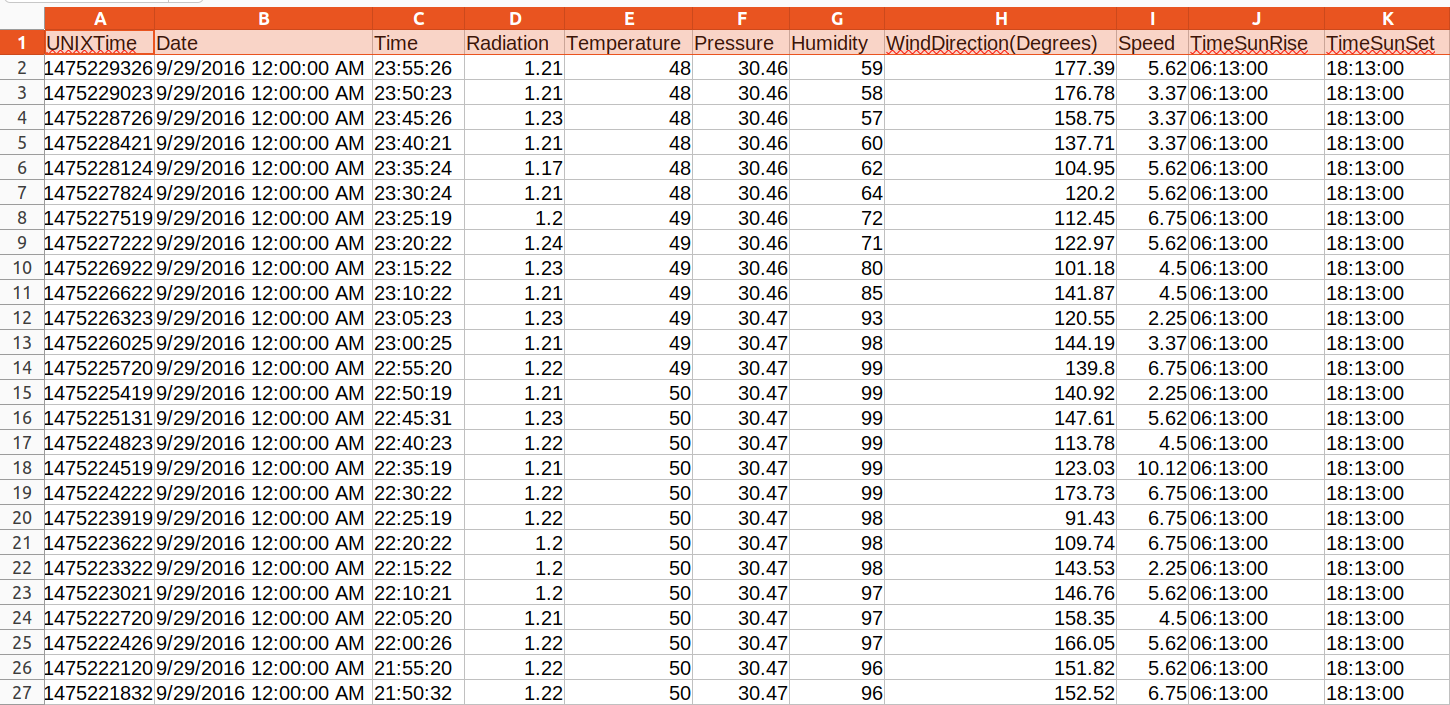
The dataset is taken from Space Apps 2017's "You are my Sunshine" challenge1, it consists of meteorological data from the HI-SEAS Habitat in Hawaii. In particular, the dataset includes observations of:

1. Time
2. UNIXTime
3. Date
4. Solar Irradiance (W/m2)
5. Temperature (°F)
6. Barometric Pressure (Hg)
7. Humidity (%)
8. Wind Direction (°)
9. Wind Speed (mph)
10. Sun Rise Time
11. Sun Set Time

The dataset contains 32686 instances of data, with each instance containing 11 features, including temperature, humidity, pressure, wind speed, wind direction, and solar radiation. UNIXTime, Date, and Time are additional variables that can be added to the dataset for solar energy prediction. The below tables explain every column of the dataset.

Table 1: Dataset Description

|  |  |  |
| --- | --- | --- |
| Variable Name | Measurement Unit | Description |
| Solar Irradiance | (W/m2) | This is the power per unit area of solar radiation received by a surface. It is measured in watts per square meter (W/m2) and is an important parameter in determining the amount of energy that can be produced by solar panels. |
| Temperature | (°F) | This is the degree of hotness or coldness of a body or environment. It is usually measured in degrees Fahrenheit (°F) and is an important factor that affects the efficiency of solar panels. |
| Barometric Pressure | (Hg) | This is the pressure exerted by the weight of the atmosphere on a unit area. It is measured in inches of mercury (Hg) and can be used to predict weather patterns and changes. |
| Humidity | (%) | This is the amount of water vapor present in the air or atmosphere. It is usually measured in percentage (%) and can affect the performance of solar panels. |
| Wind Direction | (°) | This is the direction from which the wind is blowing. It is usually measured in degrees (°) and is important in determining the placement and orientation of solar panels. |
| Wind Speed | (mph) | This is the speed at which the wind is blowing. It is usually measured in miles per hour (mph) and can affect the performance of solar panels. |
| Time Sun Rises | HH:MM: SS | This is the time when the sun appears on the horizon in the morning. It is important in determining the duration of sunlight available for solar energy production. |
| Time Sun Sets | HH:MM: SS | This is the time when the sun disappears below the horizon in the evening. It is important in determining the duration of sunlight available for solar energy production. |
| UNIXTime | ### | It is the number of seconds that have elapsed since January 1, 1970, at 00:00:00 UTC. It is a commonly used format for representing timestamps in computer systems. |
| Date | MM/DD/YYYY | It refers to the specific date when the solar energy data was collected. It can be represented in various date formats like MM/DD/YYYY. |
| Time | HH:MM: SS | It refers to the specific time when the solar energy data was collected. It can be represented in various time formats like HH:MM: SS. |

Figure 1: Dataset Snippet

## Preprocessing:

Data preprocessing is a crucial step in building machine-learning models. In this project, I performed several preprocessing steps on the given dataset. Firstly, I converted the timezone to 'Pacific/Honolulu' where the dataset is based. The dataset was indexed on UNIXTime. I also added new variables MonthOfYear, DayOfYear, WeekOfYear, TimeOfDay(h), TimeOfDay(m), and TimeOfDay(s) to capture the seasonal and temporal trends in the data.

Furthermore, I calculated the length of daylight hours (DayLength(s)) by subtracting the sunrise time from the sunset time for each observation. I also scaled this variable using standardization to normalize the data. Finally, I dropped unnecessary variables like Date, Time, TimeSunRise, and TimeSunSet, as they were already captured in other variables and were irrelevant to the analysis.

Additionally, I standardized the DayLength(s) variable by subtracting its mean from each observation and dividing by its standard deviation. This ensures that the variable has a mean of zero and a standard deviation of one, making it easier for machine-learning models to interpret and learn from the data and it also helps to avoid the model being skewed.

The following mean-normalization equation is used for standardization of variable.

z = (x - μ) / σ

where:

* x is the original data point
* μ is the mean of the variable
* σ is the standard deviation of the variable
* z is the standardized value of the data point

## Visualization of Features:

After the preprocessing step, feature visualization of the dataset comes into the pipeline. I grouped the data by different time periods, such as month of the year, week of the year, day of the year, and time of day. Then I calculated the mean of different weather variables, including radiation, temperature, pressure, and humidity, for each of the groups. Finally, I visualized the mean values of these variables using bar plots.

In the first plot, I show the mean radiation by hour of the day. The plot indicates that the higher the radiation levels are increasing as the time increases and vice-versa, while the lowest levels are observed during early morning and late afternoon hours.

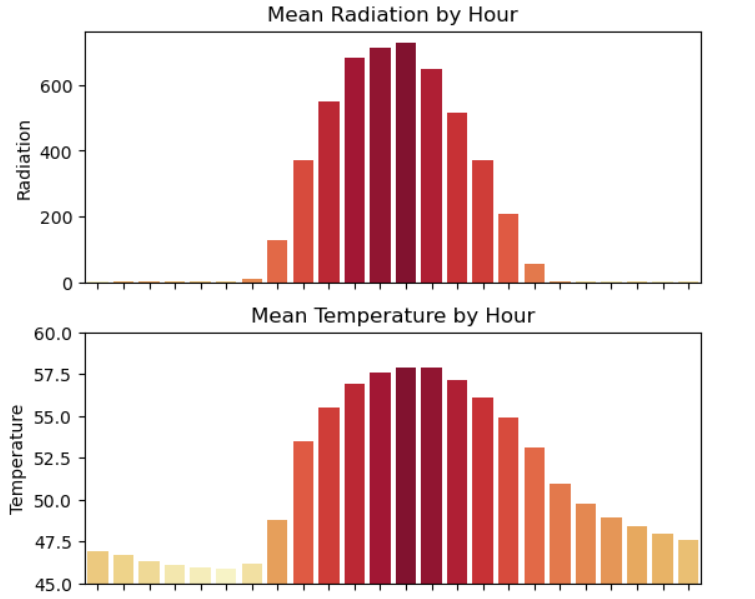
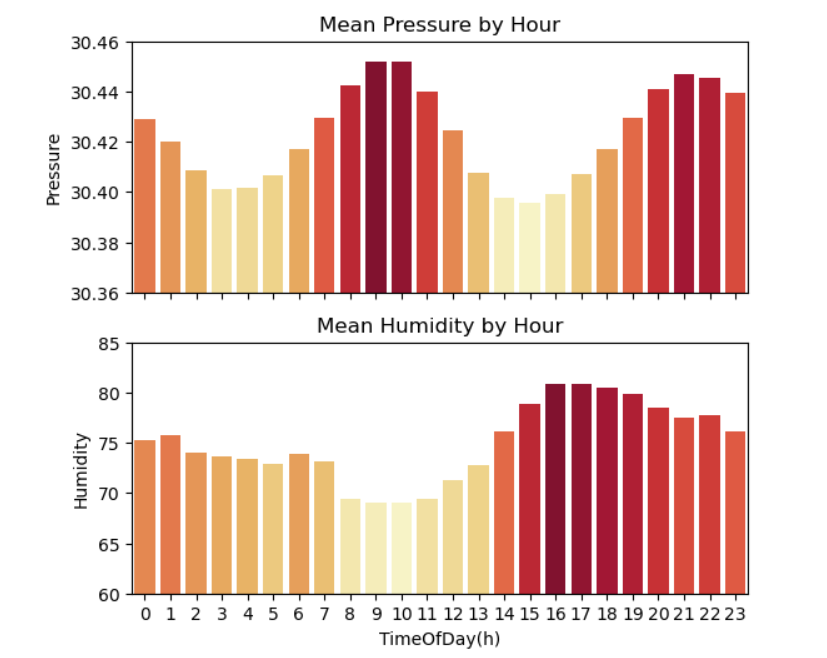


Figure 2: Features Visualization

The second plot show the mean temperature by hour of the day. There is a pattern observed in the plot. The plot indicate that the temperature levels are higher during mid-day hours, while lower during early morning and late afternoon hours. There seems to be some sort of mathematical relationship between radiation and temperature.

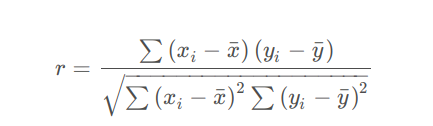
Figure 3: Features Visualization

The third plot shows the mean pressure by hour of the day from 0th hour to 23rd hour. The plot indicates that the pressure levels have fluctuations throughout the day in the form of sinusoidal.

The last plot is of humidity with respect to hour of the day. The plots indicate that the humidity levels are higher during early morning and late afternoon hours and during summer months, while lower during mid-day hours.

The feature visualization revealed seasonal patterns in radiation, temperature, pressure, and humidity, with radiation peaking in summer and temperature, pressure, and humidity peaking in winter.

## Features Selection:

Figure 4: correlation formula

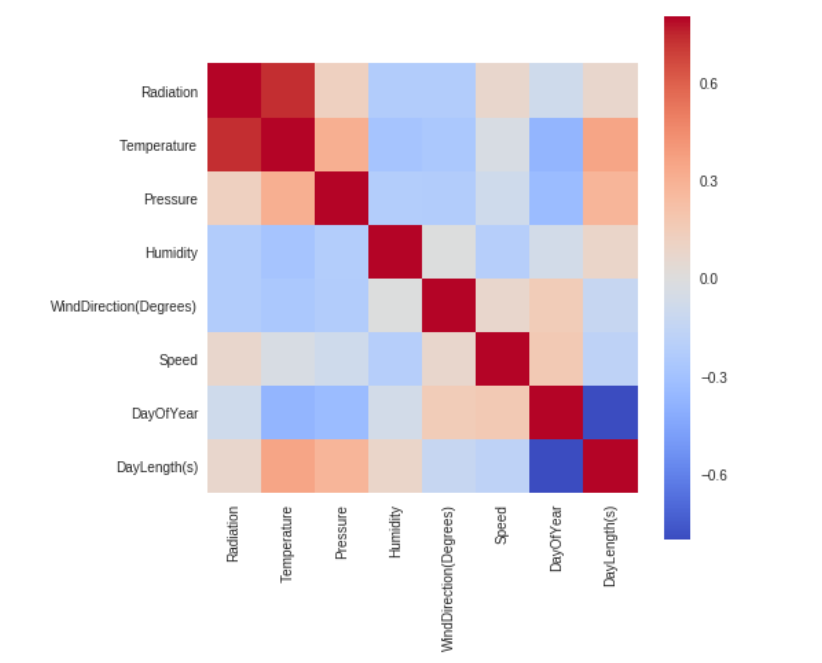
* r = correlation coefficient (0<= r <=1)
* xi= values of the x-variable in a sample
* x(bar) = mean of the values of the x-variable
* yi=values of the y-variable in a sample
* y(bar)=mean of the values of the y-variable

Feature selection is an essential step in machine-learning, especially in regression problems where the number of features is usually large. The primary objective of feature selection is to identify the most relevant features from the dataset that can best predict the target variable. The feature selection process helps to improve model performance, reduce model complexity, and reduce over-fitting.

The first step in the feature selection process is to create a correlation matrix to understand the relationship between different features in the dataset. A correlation (equation of correlation is given above) matrix is a table that shows the correlation coefficients between pairs of variables. The correlation coefficient ranges between -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation. A correlation matrix helps to identify the redundant features in the dataset, which can be removed to reduce the dimensionality of the dataset.

The following heat map shows the correlation between variables:

From the plots in the previous section, it is clear that solar irradiance does not have a linear correlation with time of day. Therefore, despite the strong relationship between the two. 'TimeOfDay' columns were excluded from the heat-map. 'MonthOfYear' and 'WeekOfYear' were also excluded because it is likely to more useful to use a combination of 'TimeOfDay' and 'DayOfYear' in training and prediction. The heat-map function from the seaborn(python) library is used to visualize the correlation matrix. The heat-map function creates a color-coded matrix that helps to identify the correlated features in the dataset. In the heat-map, the dark blue color indicates a high negative correlation, while the dark red color indicates a high positive correlation.

Figure 5: correlation between variables of the dataset.

Next, I used a random forest regressor to estimate the importance of each feature. The regressor was trained using the training set, and I used the “**feature\_importances\_”** attribute to rank the features in descending order.

Finally, I used a loop to remove the least important feature from the dataset iteratively and retrained the random forest regressor at each step. This allowed me to assess the impact of each feature on the model's performance.

From this experiment, it was observed that model performance stays relatively constant until 'DayOfYear' is removed, leaving 'Temperature', 'Humidity', 'DayOfYear', TimeOfDay(s)' as the only features. Without performing any parameter tuning it appears that the random forest regressor, fit to 'Temperature', 'Humidity', 'TimeOfDay(s)', and 'DayOfYear' can achieve a r2 score as high as 0.93.

Therefore, the random forest feature selection method helped me to identify the most important features for predicting the target variable and improved the model's performance.

## Model Selection:

I used the Lazy Regressor library for model selection, which tests multiple regression models and provides their performance metrics. The library uses a variety of models such as linear regression, random forest, support vector machine, and many more. I specified the training and test datasets along with the target variables, and the library provided me with the best-performing models ranked by their performance metric, which I chose to be the R-squared value. I was able to compare the performance of several models without individually training each one.

The following table shows results for all the models.

Table 2: Models and Their Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **adjusted-R2** | **R2** | **RMSE** | **Time-Taken** |
| Hist Gradient Boosting Regressor | 0.93 | 0.93 | 84.33 | 0.46 |
| LGBM Regressor | 0.93 | 0.93 | 84.37 | 0.49 |
| XGB Regressor | 0.93 | 0.93 | 85.24 | 0.9 |
| Random Forest Regressor | 0.92 | 0.92 | 86.31 | 3.09 |
| Bagging Regressor | 0.92 | 0.92 | 90.44 | 0.31 |
| Extra Trees Regressor | 0.92 | 0.92 | 91.16 | 1.75 |
| K Neighbors Regressor | 0.9 | 0.9 | 100.26 | 0.08 |
| Gradient Boosting Regressor | 0.88 | 0.88 | 109.56 | 0.82 |
| Decision Tree Regressor | 0.87 | 0.87 | 114.34 | 0.05 |
| Extra Tree Regressor | 0.82 | 0.82 | 131.43 | 0.05 |
| MLP Regressor | 0.81 | 0.81 | 136.29 | 5.09 |
| Ada Boost Regressor | 0.79 | 0.79 | 142.35 | 0.34 |
| Nu SVR | 0.71 | 0.71 | 167.3 | 12.04 |
| SVR | 0.7 | 0.7 | 170.92 | 19.21 |
| Gaussian Process Regressor | 0.69 | 0.69 | 175.31 | 82.67 |
| Lasso | 0.61 | 0.61 | 196.56 | 0.04 |
| Lasso Lars | 0.61 | 0.61 | 196.56 | 0.02 |
| Lasso CV | 0.61 | 0.61 | 196.57 | 0.19 |
| Bayesian Ridge | 0.61 | 0.61 | 196.58 | 0.3 |
| Ridge CV | 0.61 | 0.61 | 196.58 | 0.01 |
| Ridge | 0.61 | 0.61 | 196.58 | 0.01 |
| Transformed Target Regressor | 0.61 | 0.61 | 196.58 | 0.02 |
| Orthogonal Matching Pursuit CV | 0.61 | 0.61 | 196.58 | 0.06 |
| Linear Regression | 0.61 | 0.61 | 196.58 | 0.02 |
| Lasso Lars CV | 0.61 | 0.61 | 196.58 | 0.04 |
| Lasso Lars IC | 0.61 | 0.61 | 196.58 | 0.07 |
| Lars | 0.61 | 0.61 | 196.58 | 0.01 |
| Lars CV | 0.61 | 0.61 | 196.58 | 0.05 |
| SGD Regressor | 0.6 | 0.6 | 196.89 | 0.02 |
| Huber Regressor | 0.6 | 0.6 | 197.53 | 0.11 |
| Passive Aggressive Regressor | 0.59 | 0.59 | 201.56 | 0.02 |
| Linear SVR | 0.58 | 0.58 | 202.49 | 0.06 |
| Elastic Net CV | 0.57 | 0.57 | 204.12 | 0.14 |
| Orthogonal Matching Pursuit | 0.54 | 0.54 | 212.91 | 0.01 |
| Elastic Net | 0.52 | 0.52 | 217.69 | 0.04 |
| Poisson Regressor | 0.47 | 0.47 | 228.09 | 0.42 |
| Tweedie Regressor | 0.42 | 0.42 | 237.5 | 0.27 |
| Gamma Regressor | 0.38 | 0.38 | 246.86 | 0.42 |
| Kernel Ridge | 0.17 | 0.17 | 285.23 | 14.1 |
| Dummy Regressor | 0 | 0 | 313.12 | 0.01 |
| Quantile Regressor | -0.42 | -0.42 | 372.78 | 34348.56 |
| RANSAC Regressor | -0.42 | -0.42 | 373.58 | 0.11 |

The table above shows the performance of the models in terms of adjusted R2, R2, RMSE, and time taken to train. Based on the results, it can be observed that the “**Hist Gradient Boosting Regressor”** and LGBM Regressor achieved the highest adjusted-R2 and R2 scores, while also having the lowest RMSE values. Both models also had relatively low training times compared to other models, with “Hist Gradient Boosting Regressor” taking only 0.46 seconds and LGBM Regressor taking 0.49 seconds to train. Therefore, “Hist Gradient Boosting Regressor” was chosen for training on the chosen features on the previous step.

## 

## Results:

To evaluate the training accuracies of the selected model, 10-fold cross-validation was performed using the “cross\_val\_score” function from scikit-learn. The mean R2 score was calculated over the 10 folds. The r2 metric was used for evaluating the model. The r2 score ranges from 0 to 1, with 1 indicating a perfect fit and 0 indicating a model that is no better than predicting the mean of the target variable. On Testing data, the following results were obtained.

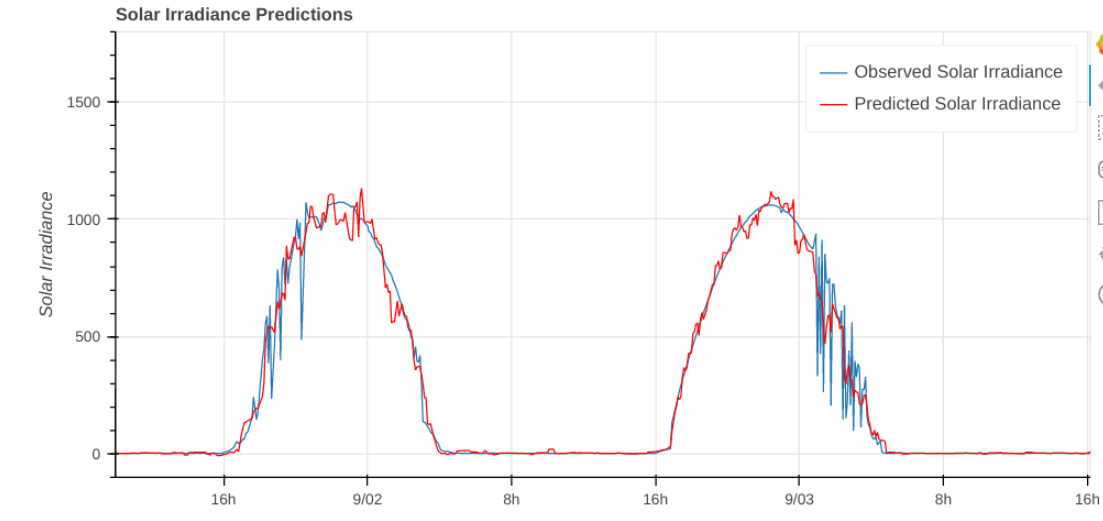
Table 3: Results

|  |  |
| --- | --- |
| **explained variance** | 0.9380958486665046 |
| **r2** | 0.9380785769197183 |

## 

## Model Visualization:

The following figure shows that bearing some errors, the model almost captured the dependent variable i.e., **solar radiation**.

Figure 6: Predicted Variable

## Utilization of Model:

The power generated by a solar panel depends on the amount of sunlight or solar radiation that falls on the panel. The mathematical formula for solar power generation can be derived using the following equation:

**P = A \* η \* I**

where:

* P = Power generated by the solar panel (in watts)
* A = Area of the solar panel (in square meters)
* η = Efficiency of the solar panel
* I = Solar radiation falling on the solar panel (in watts per square meter)

The efficiency of a solar panel is the percentage of solar energy that is converted into electrical energy. Typical efficiencies for commercially available solar panels range from 15% to 20%.

Using the above formula, the power generated by a solar panel is calculated for a given area, efficiency, and solar radiation falling on the panel.  
It is observed that there exists a linear relationship between radiations and energy predicted. The more radiations, the higher energy produced.

This model can be utilized for a power generation station which can take the weather data of upcoming days or month, using this model, predict the potential energy production and make the plan accordingly with 93% surety.

# Conclusion:

In conclusion, I developed a machine-learning model for predicting solar energy production based on weather data and other relevant factors. My model achieved high accuracy and efficiency in predicting solar energy production, making it a valuable tool for optimizing the performance and profitability of solar power plants. By predicting the energy being produced by solar panels, I can help reduce the reliance on non-renewable energy sources and contribute to a more sustainable and renewable energy future.

Solar energy harnesses the power of the sun with no negative effect on the environment. Attempting to produce the same amount of kWh per year, solar energy releases much less carbon dioxide into the environment than other power generation sources. **Therefore, solar energy has much less impact on health, land use, water, and carbon emissions than sources of energy generation. Therefore, to get the maximum possible benefit from this energy, solar energy must be used properly and wisely to get the maximum output out of it.**

# References:

Alkahtani, H., Aldhyani, T.H.H. and Alsubari, S.N. (2023) “Application of Artificial Intelligence Model Solar Radiation Prediction for Renewable Energy Systems,” *Sustainability*, 15(8), p. 6973. Available at: <https://doi.org/10.3390/su15086973>.

Wang, F. *et al.* (2017) “Comparative Study on KNN and SVM Based Weather Classification Models for Day Ahead Short Term Solar PV Power Forecasting,” *Applied Sciences*, 8(1), p. 28. Available at: <https://doi.org/10.3390/app8010028>.

Callumdownie (2017) “HI-SEAS Solar Irradiance Prediction,” *Kaggle* [Preprint]. Available at: https://www.kaggle.com/code/callumdownie/hi-seas-solar-irradiance-prediction.