STA-6714 Final Analytical Report

On

“Comprehensive Analysis and Prediction of Solar Irradiance”

By

Jainam Shah (4179144)

Submitted to
University of Central Florida

Table of Contents **Page No.**

Abstract

Solar Radiation is the predominant source of energy on earth and plays a vital role in the process of environmental cycles taking place on earth including hydrological cycles, vegetation photosynthesis and climate change. Precise analysis and accurate prediction of solar radiation is very important in conserving energy and preventing climate change all over the planet. Prediction also helps to ameliorate organization and coordination of solar systems and further helps in diminishing electricity bills and prove to be economically efficient for households. In this project, statistical methods and machine learning techniques have been employed to predict solar irradiance. The methods include building prediction model with random forest regressor and applying cross-validation which prove to be effective in providing accurate prediction whereas rest of the methods fail due to inability of scaling big data and defining long-term dependency. Results indicate that meteorological factors like temperature, time of day and year were important for predictive analytics. The dataset used in this project includes information from HI-SEAS habitat from Hawaii and multiple sources which tracked the formation of solar irradiance. Extreme surface temperature and the extend of solar radiation shows importance of solar irradiance in different places through trend analysis. How much energy is stored in solar panels and the amount of solar energy is required for daily consumption level for mission in HI-SEAS and the factors affecting it have been tracked. Final insights of this project indicate that cross-validation method and Random Forest Regressor performed better than the rest of the methods to predict solar irradiance.

Introduction

Background

In modern 21st century, with the advent of industrialization, innovations of various technologies and the increasing usage of fossil fuels, it is becoming more difficult to maintain the stability of the planet. Electricity is the most important source for mankind in order to sustain and we are most dependent on it. Majority of the electricity is produced from fossil fuels and based on census it is going to increase by 60% more by the end of 2030. Due to lower availability of fossil fuels and because of its major impact on environmental changes, the concept of renewable energy sources and energy conservation came into existence. The renewable energy sources are free and do not harm the living atmosphere or environment of the planet. Among them, Solar energy is the most sustainable source as it free, available in extreme abundance and poses comparatively lower risk on its limiting exposure. Thorough analysis into obtainability of renewable energy sources has spectated evolution of renewable energy sources, majorly focusing on developing nations. As the developing nations strive towards globalization and increasing their status quo on world platform the employment of renewable energy sources is imperative in such regions.

Solar radiation has ample benefits as it determines agricultural production cycle, maintains atmospheric pressure as well as ecological services. Information regarding the parameters in the particular location can prove helpful in predicting solar irradiance in that region and help conserve energy. Solar radiation is the chief parameter for managing solar energy and ensuring its conservation. Through conservation of solar radiation many applications can be fulfilled which include hydrology, ecology, and meteorology. Considering other facilities to predict variables like humidity, temperature, and Air Quality Index (AQI), tools and technology to predict solar radiation are very limited. This is due to complexity observed in its measurement and high maintenance. Through solar radiation maximum amount of energy conserved, if proper techniques are employed at the right time and right place, it can be very fruitful, not only at global level but commercially as well for households and population as it does not deteriorate the environment and maintains stability of the planet.

There are numerous spacecrafts orbiting around the earth, for example International Space Station (ISS), it needs a power source to carry our research in space. For basic needs, ISS works on electrical power to allow astronauts to live and work comfortably in space. In space, sun is a readily available source of energy. NASA has been developing technologies to convert solar energy into power to provide electricity and ensure research fluidity. The technologies developed include solar panels which are charged through sunlight and energy is stored when the spacecraft is not in sun's radar. Sunlight is available to mankind for free and based on that if technologies can be created which can provide sustainable energy on Earth, it would bring revolutionary changes. Currently NASA has funded Hawaii Space Exploration Analog and Simulation (HI-SEAS) which has a habitat like Mars on a remote island at Hawaii. The crew on the island relies on photovoltaic systems i.e. solar panels to conduct research and perform experiments as well as day-to day activities. The main goal is to monitor levels of energy generations from solar panels.

Purpose

Taking into account the conflicting factors hampering the conservation of solar energy in different regions, an accurate model or technique is necessary to predict solar irradiance. Various models can be built for solar radiation prediction based on statistical methods and machine learning techniques. Methods including feature visualization, statistical analysis from cross validation and creating predictive model by fitting necessary parameters in random forest regressor can be employed and correlation can be determined to match with accuracy. More machine learning techniques like regression, extreme gradient lifting (XG Boost), Random Forest Regressor, K-Nearest Neighbor, Decision Tree etc. can be utilized to get more accurate predictions as they solely focus on training the available information and providing insights from that.

With the help of mentioned techniques accurate models can be built which can help in tracking levels of solar radiation over long term and provide economic benefits in the process. The primary purpose of this project is to consider parameters and build a method or machine learning model through which solar irradiance can be predicted as topic of solar energy is palpable and currently capacity of solar batteries and panels to store energy is unknown. The need to develop the technology which can predict solar irradiance is of utmost importance as the solar energy on its current pace is and will be available in extreme abundance. Based on current statistics, considering the meteorological stations across countries like China which has 756 meteorological stations but only 122 of them can measure solar radiation, same goes with Turkey as the country having 1798 meteorological stations but only 129 of them are capable of measuring solar radiation. Therefore, building accurate model and providing thorough analysis of solar radiation is crucial. In numerous literature articles regarding solar energy, models are built using various empirical, statistical and machine learning techniques which predict the solar irradiance. These models focus on different parameters based on available data. This project seeks to build a model which can predict solar irradiance and provide a proper analysis into the availability of solar energy levels based on available data.

Scope

Solar energy is available in abundance throughout the planet, but complexity is observed in measuring proper irradiance and there is high maintenance in handling and monitoring the equipment used to measure solar radiation levels. Due to lack of the mentioned factors, limited data is available is which has accurate information regarding solar irradiance levels. The dataset is obtained from HI-SEAS habitat in Hawaii in which multiple factors are considered for measuring solar irradiance including temperature, pressure, data & time, radiation level, humidity, wind, barometer pressure etc. Based on these factors the prediction of solar irradiance has been performed. The data is limited but the goal of the project is to build a sustainable model for accurate prediction and analysis. Conserving solar energy through specific technologies can revolutionize the way of life and also ensure stability of planet, multiple machine learning techniques can be applied to get the desired results. This project focuses on providing the best solutions to predict solar irradiance based on available data and provide comprehensive analysis.

Literature Review

Taking an overview of some of the literature articles based on prediction of solar irradiance, various methods and techniques are utilized in these research studies in order to gain accurate analysis of the radiation levels of solar energy and ensure its conservation. The following articles have been taken into account and a brief summary has been provided to get a general idea of solar irradiance. The reference of the summarized articles is mentioned in the bibliography section.

Article 1 – "Comprehensive assessment, review, and comparison of AI models for solar irradiance prediction based on different time/ estimation intervals**"**

In the mentioned research article, Ineptitude of various solar energy-based technologies to measure radiation levels of solar energy has been discussed, it is stated that even though such technologies have been invented recently they are still unable to provide accurate results of solar irradiance. To counter that eight different AI models have been developed including Artificial Neural Network (ANN), Convolutional Neural Network (CNN), Extreme Gradient Lifting (XG Boost), Long Short-Term Memory Recurrent Model (LSTM), Multiple Linear Regression (MLR), Decision Tree, Polynomial Regression (PLR) and Random Forest Regression (RFR). In addition, further two more neural networks have been designed for the same purpose. The study in the mentioned research article focuses on developing AI models which evaluates different timestamps including hourly, per minute and daily mean radiation occurring. It also evaluates varied solar irradiance from six different African countries measured with specific technologies. It focuses on developing a global AI model which estimates solar irradiance in all countries for which it focuses on training and testing it with different algorithms mentioned previously.

Based on the inferences made after testing all AI models, XG Boost had comparatively higher performance than the rest of the models in majority of case studies undertakes. It also revealed that hourly prediction of solar irradiance was more accurate than daily mean and per minute timestamp. The research article showcases results of each AI model. As all AI models are tested in the research study, every model shows different level of sustainability based on specific parameters. XG Boost emerges as the best model because it surpasses most test cases with maximum accuracy. Rest of the models in the study do not provide accurate results based on tested cases which indicates more AI models employing different techniques should be developed. For future purposes, more precise research should be conducted which focuses on measuring solar radiation per minute time gap as it would be groundbreaking in terms of technological innovation and would help in energy conservation.

Article 2 – "Solar Irradiance Forecasting Using Deep Neural Networks**"**

In the mentioned research article, specific neural networks have been developed to forecast solar irradiance. Solar energy conservation is the chief factor in renewable energy generation. Through predictive analytics organization and maintenance of photovoltaic systems is possible and is helpful in limiting utility bills in households. Various statistical methods had been employed in the mentioned research study including Support Vector Machine (SVM), Artificial Neural Networks (ANN), and Autoregressive Moving Average (ARMA) to predict solar irradiance. However most statistical methods do not provide accurate results due to poor scalability. Hence, to predict solar irradiance, deep neural networks have been utilized, one of the techniques mentioned called Deep Recurrent Neural Networks (DRNNs) is employed, even though neural networks add complexity to network, it focuses on extraction of multiple features affecting solar irradiance. The data used for the research study in the article is obtained from resources in Canada. Deep Learning methods prove more accurate than the rest of statistical methods in predicting solar irradiance.

Based on the insights gained from the research article. DRNN designed and was applied to real data obtained from Canadian solar farm. Data was prepared by discarding outliers and proper cleaning and further trained to build neural network. The inferences obtained after building neural network were compared with statistical models. The rest of the predictive models exhibited lower accuracies compared to DRNN as in DRNN, scalability is possible when its big data. Therefore, solar radiation forecasting, and predictive analytics is achievable. Time series and trend analysis of solar irradiance data can be performed if data volume is high and higher sampling rate.

Article 3 – "Solar Radiation Prediction Using Different Machine Learning Algorithms and Implications for Extreme Climate Events**"**

Solar radiation is one of the most predominant sources of energy and is responsible for various environmental cycles taking place on planet as mentioned in this research article. Therefore, accurate analysis and prediction of solar radiation is crucial to ensure its sustainability and control climate change. For that twelve machine learning models have been tested in this study for comparing daily and monthly estimates of solar radiation. The insights found in the study portray that sunshine duration, surface temperature and weather visibility are important factors in predicting solar radiation. Through trend analysis between surface temperature and total solar radiation revealed role of solar radiation in extreme different natural phenomenon.

Data cleaning and preprocessing was performed on data collected from Ganzhou station in China dating from 1980 to 2016. Based on that twelve different machine learning algorithms were applied in layers using algorithms like R^2, Root Mean Square (RMSE), Mean Absolute Error (MAE), Extreme Gradient Lifting (XG Boost), Gaussian Process Regression (GPR), Gaussian Boosted Regression Trees (GBRT) and Random Forest in the first layer, followed by multiple linear regression in second layer to create a stacking model and predict radiation levels of solar energy. The fact that solar radiation affects extreme climatic conditions was confirmed when using random forest algorithm to select different parameters and sunlight duration was found to most important because through time series analysis of average ground temperature with respect to solar radiation levels with data available from 1980 to 2016, the ground temperature increased with exposure of

solar radiation levels directly affecting climatic conditions. To conclude, the research study suggested that when GBRT, XG Boost, GPR and Random Forest Regression were performed separately, they did not provide good results, but when the stacking model was created with layers of GBRT, XG Boost, GPR and Random Forest Regression layer by layer, better and accurate results were obtained. In addition to that, when XG Boost model was created to predict solar radiation it also provided efficient results. Hence, based on the research study most efficient techniques were stacking model with using different machine learning algorithms in layers and XG Boost are the best ways to predict solar radiation.

Article 4 – "One month-ahead forecasting of mean daily global solar radiation using time series models**"**

In the mentioned research article, how principal it is forecast solar radiation in order to produce power and preserve energy has been discussed. Due to harmful usage of fossil fuels affecting the environmental balance of the planet, proper integration of energy obtained from sun is crucial and imminent. Solar energy is available in abundance but how, when and where to preserve it and forecast its energy levels is a challenge for scientists. The aim of this research study is to build a method to predict solar radiation every month and based on that two different models were created called Auto Regressive Moving Average (ARMA) and Auto Regressive Integrated Moving Average (ARIMA). These models were used to predict value of solar radiation levels on data available from Tetouan city in Morocco through time series analysis. The two models were also compared based on Goodness-of-fit values. When compared it was found out that ARIMA provided better accurate results compared on ARMA, as it was the optimum model and suitable to predict solar radiation on a monthly basis.

The mentioned research articles present varied information about the concept of solar irradiance and how the solar energy has plethora of benefits. Efficient conservation of solar energy will prove fruitful for mankind as well as it will maintain the stability of the planet. Only way to do that is to know when and where amount of radiation of solar energy is available and how to contain it for better use. As stated in the articles, various machine learning models, different statistical and empirical methods have been developed in order to predict solar irradiance. Taking motivation from that, various insights and methods have been developed in this project in order to predict radiation of solar energy levels and measure amount of energy stored in solar panels and batteries. The methodology to predict solar irradiance and construct a systematic way to achieve that has been performed in this project.

Data Preparation & Integration

Gathering Data is the initial and most crucial step to provide comprehensive understanding of a concept. For predicting solar irradiance in this project, multiple datasets are combined from meteorological data available from HI-SEAS program from NASA through period of four months (September - December 2016) over two missions. The project's aim is to help crew monitor energy levels coming from photovoltaic systems and different parameters affecting it.

Data Overview & Preprocessing

The following organization has been followed for data preprocessing on HI-SEAS Hawaii data.

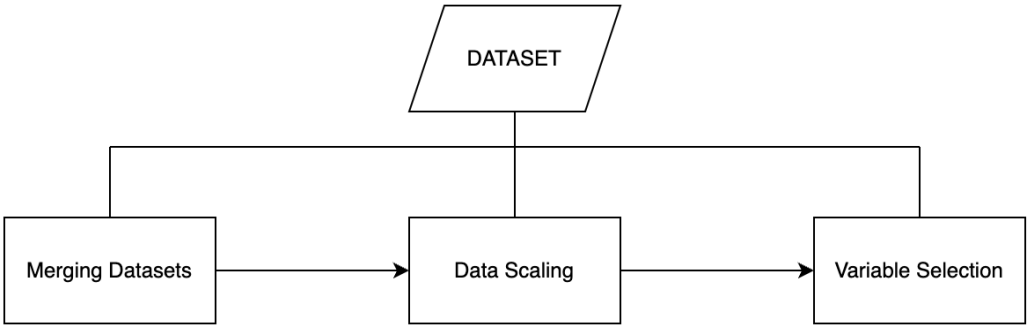


Figure 1 Data preprocessing path

Data Source - https://2017.spaceappschallenge.org/challenges/earth-and-us/you-are-my-sunshine/details
The following parameters are included in each dataset and based on that the datasets are merged.

1. Row number (1-n) for sorting
2. UNIX time_t date (secoonds) for sorting
3. Date in yyyy-mm-dd format
4. Local time in 24-hour format
5. Numeric data
6. Text Data

Numeric data and Text data are unique based on parameters as mentioned below. The merged data frame includes following columns and have been sorted based on UNIXTIME column available in each dataset:

1. UNIXTime – in seconds
2. Date – in yyyy-mm-dd format
3. Time – in 24-hour format
4. Radiation – in watts per meter^2
5. Temperature – in degrees fahrenheit

6. Pressure – in Hg (mercury)
7. Humidity - percent
8. WindDirection (Degrees) – in degrees
9. Speed – in miles per hour
10. TimeSunRise - Hawaii Time
11. TimeSunSet - Hawaii Time

UNIXTime	Data	Time	Radiation	Temperature	Pressure	Humidity	WindDirection(Degrees)	Speed	TimeSunRise	TimeSunSet
1472724008	9/1/2016 12:00:00 AM	00:00:08	2.58	51	30.43	103	77.27	11.25	06:07:00	18:38:00
1472724310	9/1/2016 12:00:00 AM	00:05:10	2.83	51	30.43	103	153.44	9.00	06:07:00	18:38:00
1472725206	9/1/2016 12:00:00 AM	00:20:06	2.16	51	30.43	103	142.04	7.87	06:07:00	18:38:00
1472725505	9/1/2016 12:00:00 AM	00:25:05	2.21	51	30.43	103	144.12	18.00	06:07:00	18:38:00
1472725809	9/1/2016 12:00:00 AM	00:30:09	2.25	51	30.43	103	67.42	11.25	06:07:00	18:38:00

Figure 2 Dataset Overview

Originally "Time" column was available in reverse order. It has been modified and correctly arranged in the data frame by resetting by sorting.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 32686 entries, 7416 to 24522
Data columns (total 11 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   UNIXTime                32686 non-null  int64
 1   Data                    32686 non-null  object
 2   Time                    32686 non-null  object
 3   Radiation               32686 non-null  float64
 4   Temperature             32686 non-null  int64
 5   Pressure                32686 non-null  float64
 6   Humidity                32686 non-null  int64
 7   WindDirection(Degrees)  32686 non-null  float64
 8   Speed                   32686 non-null  float64
 9   TimeSunRise             32686 non-null  object
 10  TimeSunSet              32686 non-null  object
dtypes: float64(4), int64(3), object(4)
memory usage: 3.0+ MB
```

Figure 3 Dataset Info

	UNIXTime	Radiation	Temperature	Pressure	Humidity	WindDirection(Degrees)	Speed
count	3.268600e+04	32686.000000	32686.000000	32686.000000	32686.000000	32686.000000	32686.000000
mean	1.478047e+09	207.124697	51.103255	30.422879	75.016307	143.489821	6.243869
std	3.005037e+06	315.916387	6.201157	0.054673	25.990219	83.167500	3.490474
min	1.472724e+09	1.110000	34.000000	30.190000	8.000000	0.090000	0.000000
25%	1.475546e+09	1.230000	46.000000	30.400000	56.000000	82.227500	3.370000
50%	1.478026e+09	2.660000	50.000000	30.430000	85.000000	147.700000	5.620000
75%	1.480480e+09	354.235000	55.000000	30.460000	97.000000	179.310000	7.870000
max	1.483265e+09	1601.260000	71.000000	30.560000	103.000000	359.950000	40.500000

Figure 4 Dataset Statistics

Feature Engineering

Through feature engineering proper data can be obtained which can be used to train the model accurately and efficiently in order to provide better results. In feature engineering, manipulation and transformation of raw data takes place and based on that solar irradiance prediction become fluid and achievable.

Initial step on importing the dataset was to transform date and time variables into a trackable format and add some parameters that can be used in proper analysis, modelling, and visualization to get a general idea about the data available.

In feature engineering the datetime function and "pytz" library has been utilized to correlate with Hawaii time zone as the data is solely focused on atmospheric entities and weather condition on Hawaii. In the modified dataset new columns are added resetting the index based on datetime and separating "date" and "time" columns based on Month of the Year ("MonthOfYear"), Day of the year ("DayOfYear"), Week of the Year ("WeekOfYear"), Time of the Day in Hours ("TimeOfDay(h)"), Time of the Day in Minutes ("TimeOfDay(m)"), Time of the Day in Seconds ("TimeOfDay(s)") and calculating Day Length ("DayOfLength(s)"). The modified dataset obtained is shown below:

UNIXTime	UNIXTime	Radiation	Temperature	Pressure	Humidity	WindDirection(Degrees)	Speed	MonthOfYear	DayOfYear	WeekOfYear	TimeOfDay(h)	TimeOfDay(m)	TimeOfDay(s)	DayLength(s)
2016-09-01 00:00:08-10:00	1472724008	2.58	51	30.43	103	77.27	11.25	9	245	35	0	0	8	45060
2016-09-01 00:05:10-10:00	1472724310	2.83	51	30.43	103	153.44	9.00	9	245	35	0	5	310	45060
2016-09-01 00:20:06-10:00	1472725206	2.16	51	30.43	103	142.04	7.87	9	245	35	0	20	1206	45060
2016-09-01 00:25:05-10:00	1472725505	2.21	51	30.43	103	144.12	18.00	9	245	35	0	25	1505	45060
2016-09-01 00:30:09-10:00	1472725809	2.25	51	30.43	103	67.42	11.25	9	245	35	0	30	1809	45060

Figure 5 Modified Data

To understand the dataset properly and to achieve the goal of efficient solar irradiance prediction, the data should be modelled and visualized in an elementary way. Through feature visualization and investigating anomalies in data more insights can be obtained. As the final dataset is ready and cleaned for proper analysis, visualizing and applying machine learning algorithms through proper methodology will help dive deeper into overcoming challenges faced in conserving solar energy, monitoring energy levels consumption in photovoltaic systems in HI-SEAS program at Hawaii and predicting solar irradiance in the process.

Methodology

In methodology, step by step through visualization and modelling proper understanding of the data and the purpose of the research study can be understood. Methodology includes feature visualization and machine learning model creation to predict solar irradiance and get more insights into the information available.

Feature Visualization

To get overall understanding of the data, parameters are visualized with bar plots using hourly and monthly means.

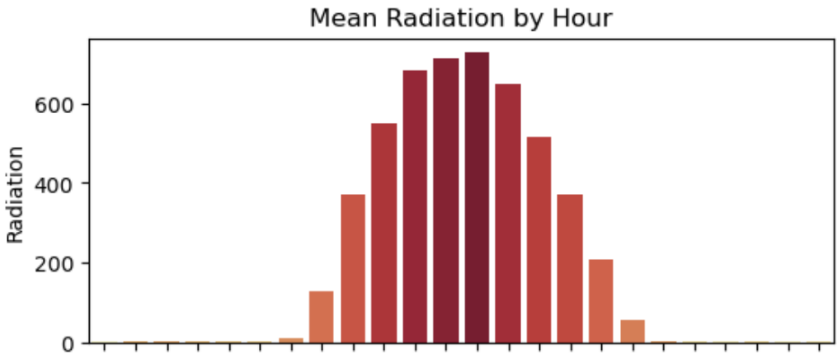


Figure 6 Mean radiation by Hour

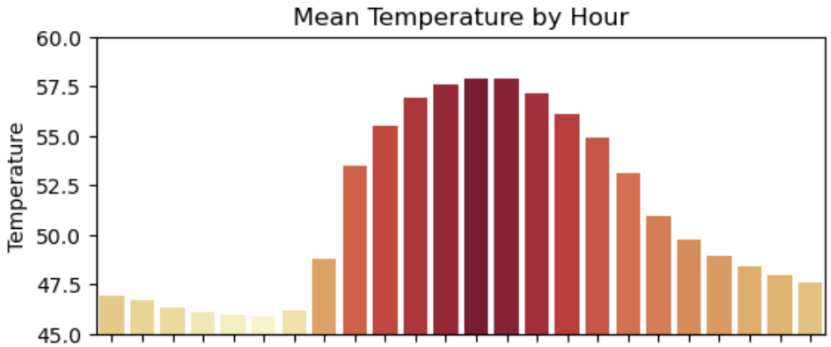


Figure 7 Mean Temperature by Hour

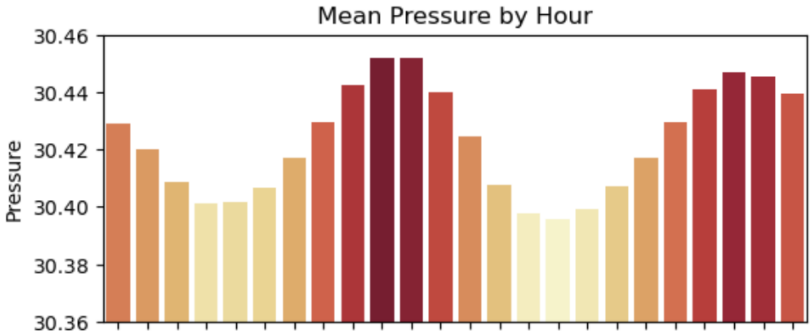


Figure 8 Mean Pressure by Hour

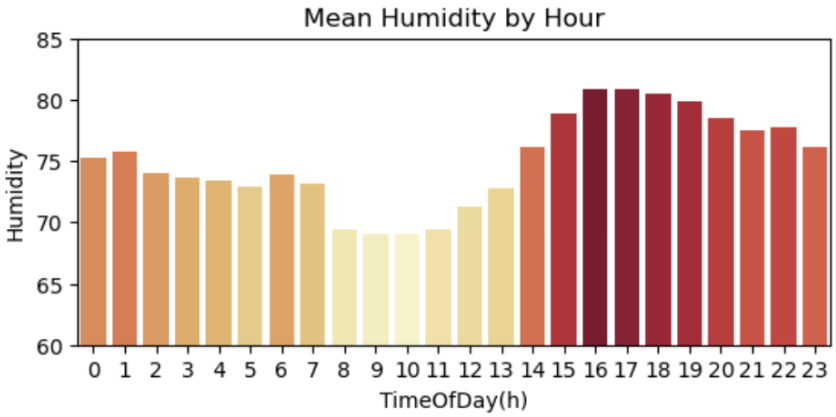


Figure 9 Mean Humidity by Hour

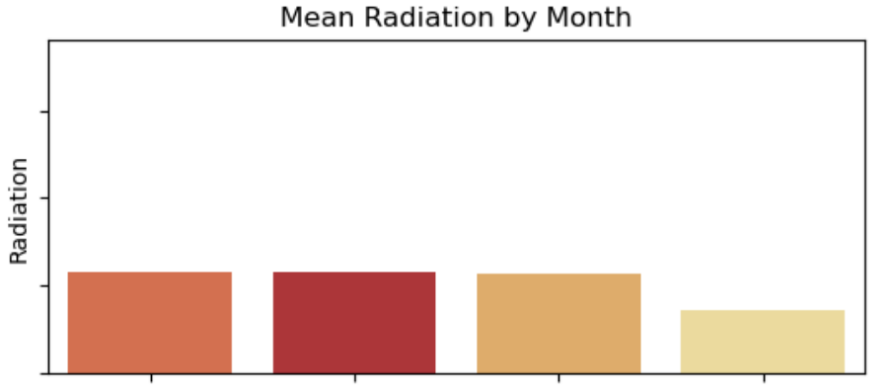


Figure 10 Mean Radiation by Month

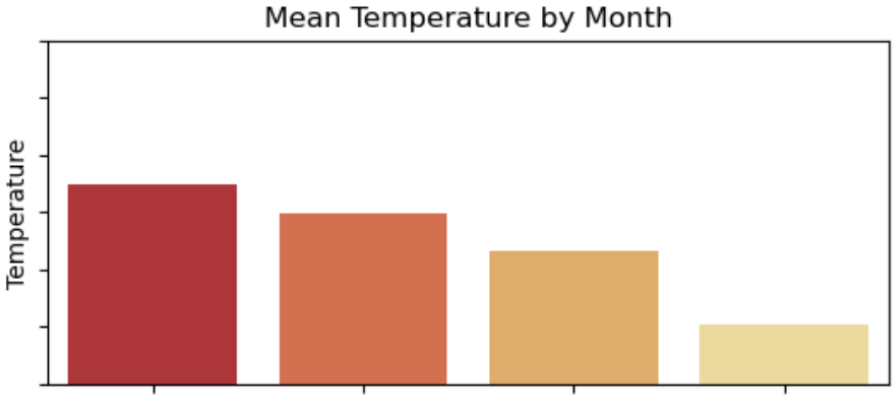


Figure 11 Mean Temperature by Month

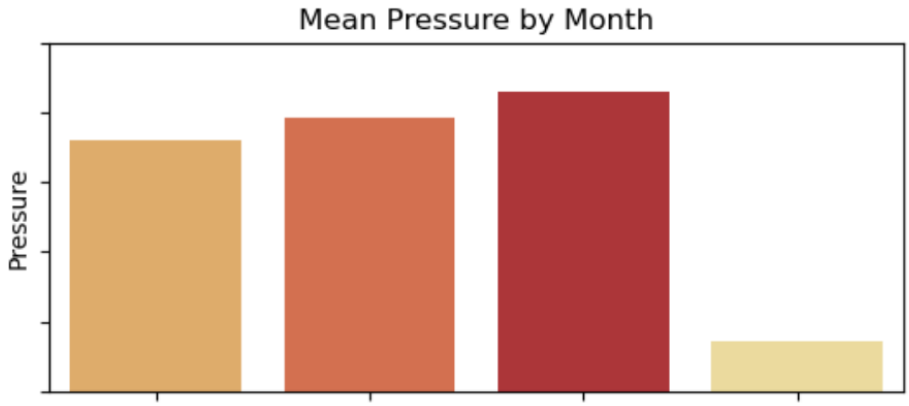


Figure 12 Mean Pressure by Month

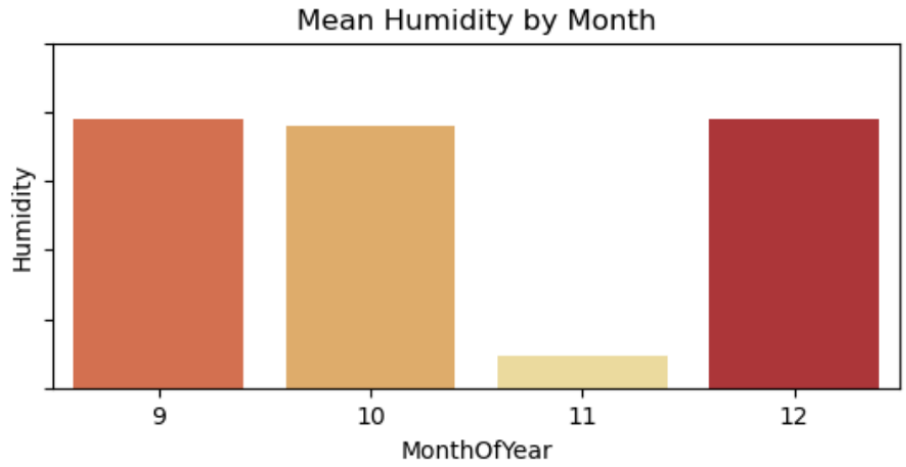


Figure 13 Mean Humidity by Month

From the bar plots obtained from feature visualizations, following insights have been procured: -

- Temperature possesses strong correlation with solar irradiance.
- Relationship between pressure with solar irradiance and humidity with solar irradiance are comparatively less as there is no significant drop measured in values of pressure and humidity hourly. When compared on monthly basis only month of December for pressure and month of November for humidity shows a significant drop.
- Humidity possesses negative correlation with solar irradiance and also with temperature and pressure.
- Solar irradiance and temperature are the highest at approximately between 12:00 to 13:00 hours respectively.
- As the data comprises of information from 4 months i.e from September to December, means of solar irradiance and temperature decreases with winter approaching. Only exception is from month September to October where there is a limited increase in temperature.

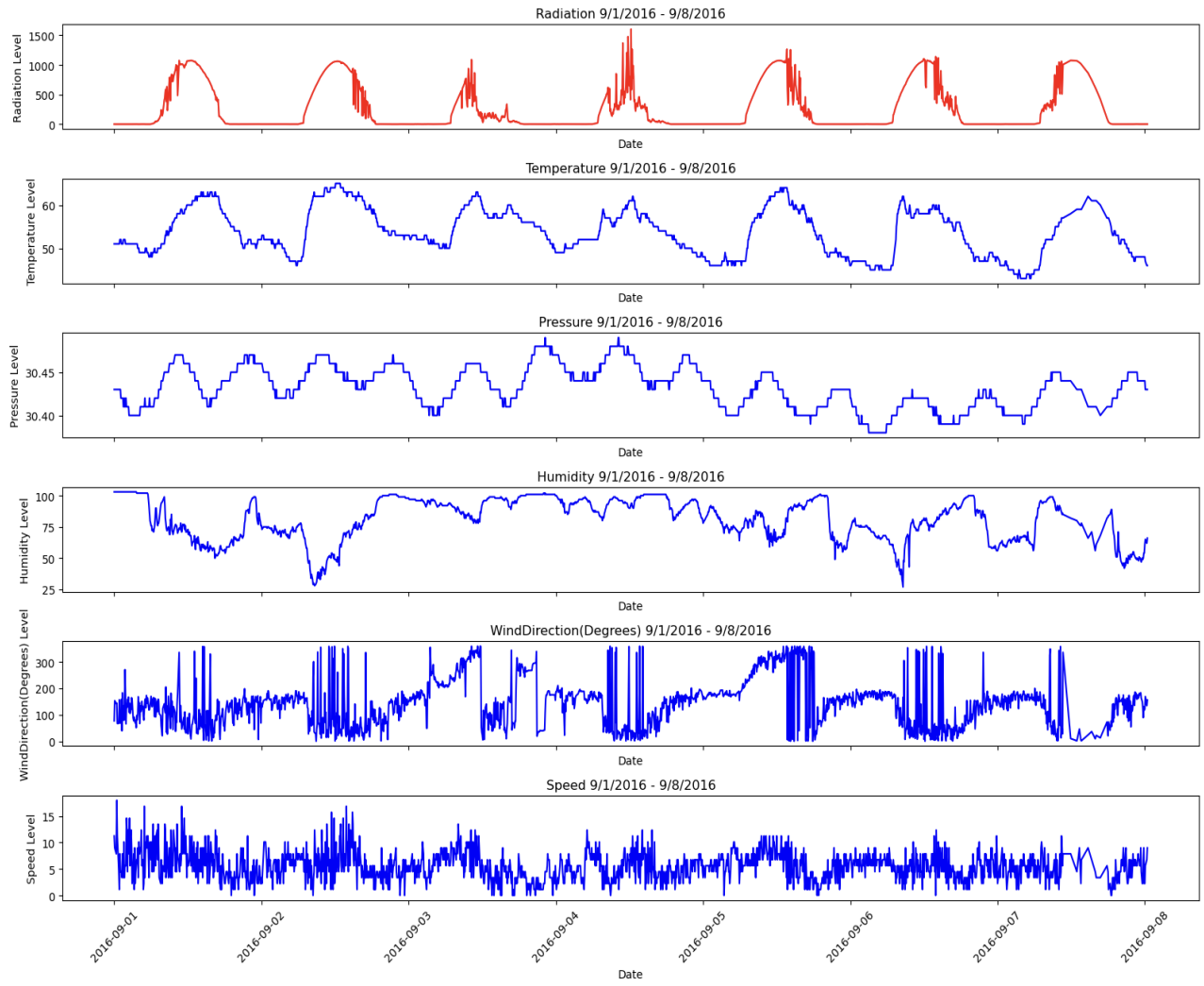


Figure 14 Line Graphs of dataset variable for finding correlation based on weekly division

The following insights can be procured from the above graphs:

- Temperature and radiation have close correlation as oscillations are quite similar.
- Temperamental wind direction might have a relation with radiation occurring.
- Pressure is cyclic in nature but follows different path than radiation.
- Presence of speed and humidity do not create an impact or cause spikes in solar radiation as they are totally wayward.

Correlation

To get a deeper understanding of the relationship of all parameters in the dataset, a Pearson Correlation Heatmap has been plotted. From previous visualizations it is evident that solar irradiance does not possess linear correlation with time of day. In that case, despite any positive correlation between both of them,'TimeofDay' related columns were not used in heatmap. 'MonthofYear','WeekofYear' and 'UNIXTime' were also discarded as combination of 'TimeofDay' and 'DayofYear' combination are likely to be more useful in training and prediction.

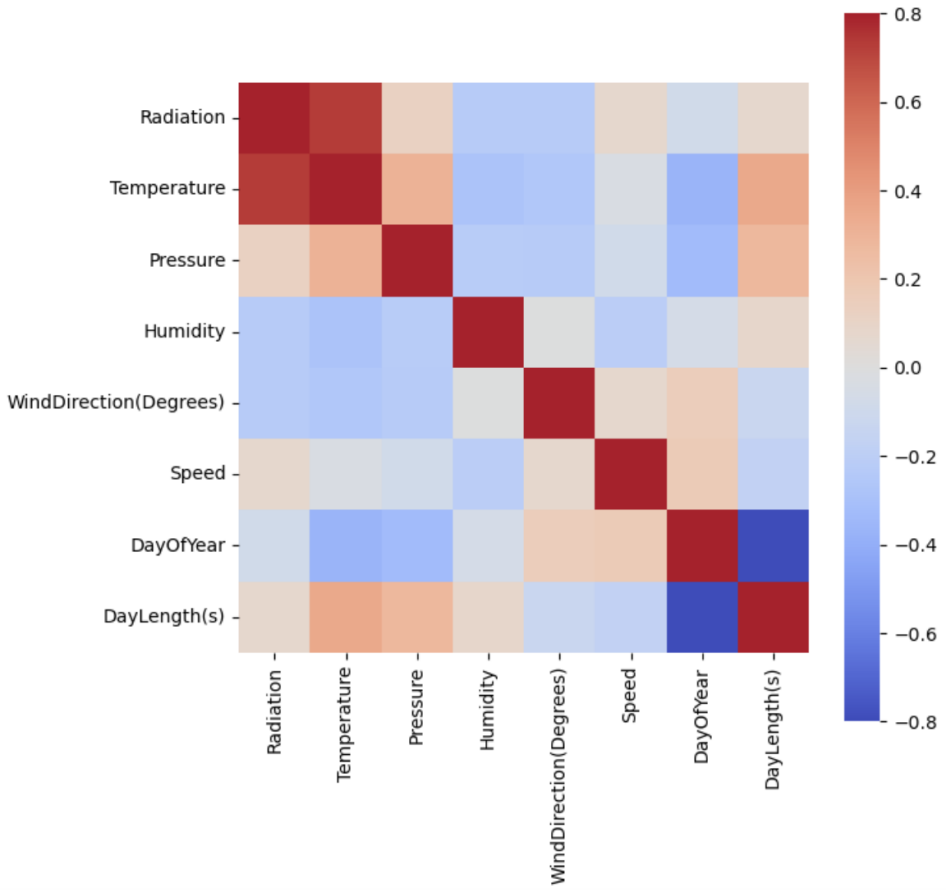


Figure 15 Pearson Correlation Heatmap

Pearson's Correlation Heatmap is the evidence of relationships between variables mentioned in obtained insights and shows that day of the year has weaker relationship with solar irradiance compared to temperature.

Model Creation

Building a machine learning model by employing statistical and machine learning methods is the most efficient way to reach towards project's objective of predicting solar irradiance. In model creation, a structured way is created by recognizing the patterns found in data and based on that inferences can be made of possible outcomes. Model creation in this project follows the path shown in the flowchart below.

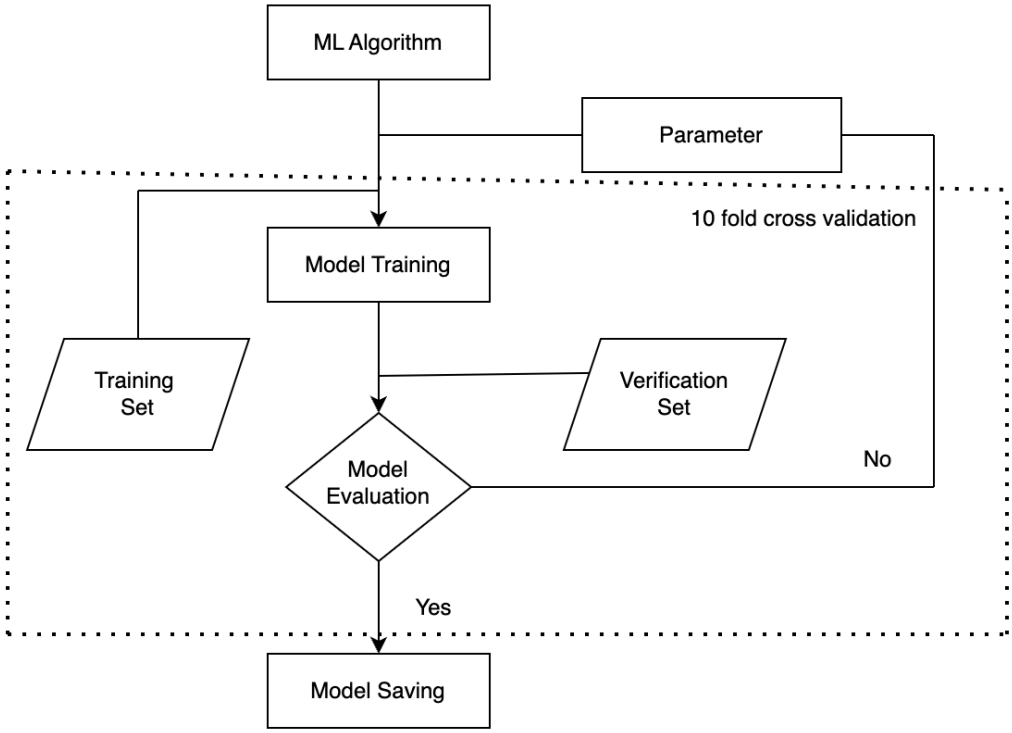


Figure 16 Model Creation

Splitting Independent and Dependent Variables

All parameters excluding solar irradiance were included in independent variables set, 'TimeofDay(s)' and 'DayOfYear' were used to denote date and time. Solar Irradiance (radiation) was set as the dependent variable.

```
X = dataset[['Temperature', 'Pressure', 'Humidity', 'WindDirection(Degrees)',
             'Speed', 'DayOfYear', 'TimeOfDay(s)']]
y = dataset['Radiation']
```

Figure 17 Splitting variables

After that, the dataset was split into testing and training sets with 20% and 80% ratio respectively.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2,
                                                    random_state = 0)
```

Figure 18 Splitting Dataset

As majority of data is nonlinear, employment of linear or multiple regression would not have been suitable for accurate results. Evidently, SciKit Learn's library in python which contains decision tree-based regression algorithm focuses on feature important parameters. With its help, backward elimination procedure was performed where least important parameters of the data were removed and r^2 scores from cross validation (resampling method to test and train data on differentparameters) were recorded for specific models.

	Features	r2 Score
0	Temperature, Pressure, Humidity, WindDirection...	0.932683
1	Temperature, Humidity, WindDirection(Degrees),...	0.931575
2	Temperature, Humidity, DayOfYear, TimeOfDay(s)	0.933883
3	Temperature, DayOfYear, TimeOfDay(s)	0.933187
4	Temperature, TimeOfDay(s)	0.800596

Figure 19 R^2 Scores

From the output, it can inferred that performance of model stays relatively constant until 'DayOfYear' column is discarded, left with only 'Temperature' and 'TimeOfDay(s)'. Without modifying any variables, through random forest regressor (an estimator which fits classifying decision trees based on samples available), model is fit with 'Temperature', 'TimeOfDay(s)' and 'DayOfYear' and r^2 score of approximately 0.93 is achieved.

Based on this obtained result, key regressors are fitted to key features and random forest regressor is trained using "Temperature", "DayOfYear" and "TimeOfDay(s)" as shown below.

```
X_train_best = X_train[['Temperature', 'DayOfYear', 'TimeOfDay(s)']]
X_test_best = X_test[['Temperature', 'DayOfYear', 'TimeOfDay(s)']]
regressor.fit(X_train_best, y_train)
```

Figure 20 Fitting Key Regressors to Key Features

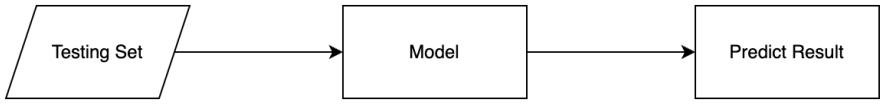
Now performing cross-validation with number of folds=10, produced the following results.

```
accuracies = cross_val_score(estimator = regressor, X = X_train_best,
                             y = y_train, cv = 10, scoring = 'r2')
accuracy = accuracies.mean()
print('r2 = {}'.format(accuracy))
```

```
r2 = 0.934070913136147
```

Figure 21 Cross Validation with folds = 10

Predictive Modelling



Regressors trained in the model creation will be used to predict and test dataset which was not yet used in training dataset. With that results will be obtained for solar irradiance prediction.

```
explained variance = 0.9393423423626016
mse = 6258.650927530407
r2 = 0.939283566667588
```

Figure 22 Solar Irradiance Prediction Accuracy

The obtained variance, r^2 and mean squared error are results depicting accuracy of the created model. The results obtained are almost similar to the one obtained in cross-validation, and it can be inferred that created model is not overfit. Furthermore, the model is visualized to check how accurate the predictions are with actual observations.

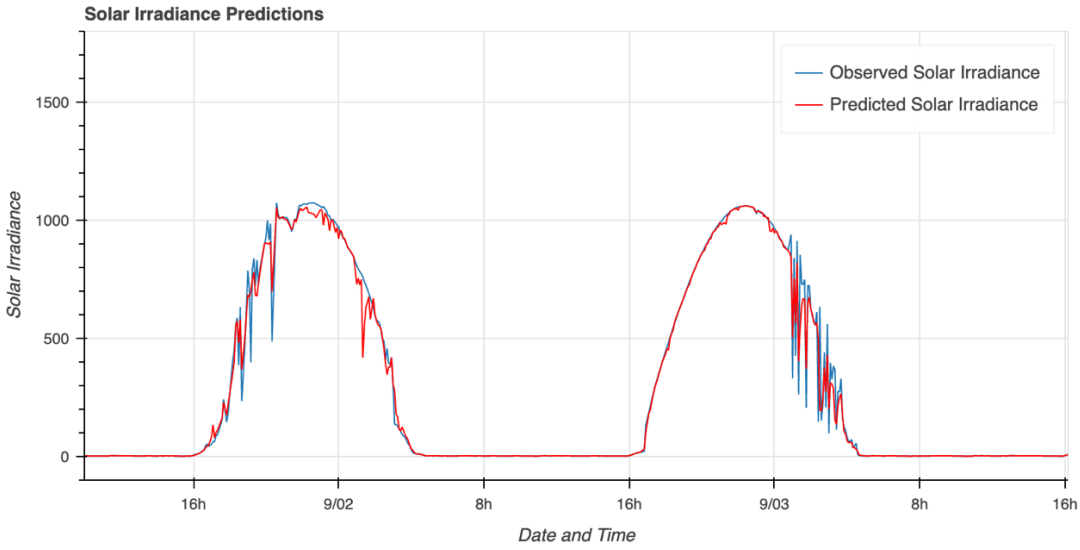


Figure 23 Solar Irradiance Predictions

From the figure, it can be inferred that predictions are very close to the calculated observations.

Analytical Results

Analytical Results depict the key takeaways obtained from research study and outlines the need for the analysis for which the project was undertaken. It provides insights and assumptions based on predictive modelling and feature visualizations obtained from methodology. The following results are obtained in the project.

Based on Feature Visualization **(Referenced from methodology)**: -

- Temperature possesses strong correlation with solar irradiance.
- Relationship between pressure with solar irradiance and humidity with solar irradiance are comparatively less as there is no significant drop measured in values of pressure and humidity hourly. When compared on monthly basis only month of December for pressure and month of November for humidity shows a significant drop.
- Humidity possesses negative correlation with solar irradiance and also with temperature and pressure.
- Solar irradiance and temperature are the highest at approximately between 12:00 to 13:00 hours respectively.
- As the data comprises of information from 4 months i.e., from September to December, means of solar irradiance and temperature decreases with winter approaching. Only exception is from month September to October where there is a limited increase in temperature.
- Temperature and radiation have close correlation as oscillations are quite similar as shown in Figure 14.
- Temperamental wind direction might have a relation with radiation occurring.
- Pressure is cyclic in nature but follows different path than radiation.
- Presence of speed and humidity do not create an impact or cause spikes in solar radiation as they are totally wayward.
- Day of the year has weaker relationship with solar irradiance as tracked through correlation heatmap.

Based on Model Creation and Predictive Modelling

- Most important parameters for prediction of solar irradiance are temperature, time of day and day of the year.
- Through random forest regressor training of the mentioned three variables, a model was created which had a mean r^2 score of 0.93 when passed through cross-validation.
- On comparison of r^2 score with a test set, same r^2 score was obtained.
- Tuning in the random forest regressor to test with more parameters didn't prove as effective as it did with the mentioned parameters. With the mentioned parameters the created model achieved highest accuracy and was able to predict almost correct solar radiation levels when compared to actual observations.

Conclusion

Solar Irradiance is the fundamental process of measuring solar energy available from the sun. As it is energy, it contains some levels of radiation. Proper analysis of when, where and in how much amount energy should be measured and conserved is of utmost importance. Based on the results obtained in this project by using the dataset available from HI-SEAS habitat on Hawaii, it can be observed that solar radiation and temperature have the strongest relationship with each other, whereas humidity has the least effect on solar radiation. The information available in HI-SEAS dataset of Hawaii contains data from the span of 4 months. Based on insights found from predictive modelling, temperature, time of day and day of the year are most important in predicting solar irradiance. After testing multiple machine learning algorithms on parameters available in the data, it was found that, training random forest regressor and applying cross-validation helped achieve the highest accuracy for solar irradiance prediction. Hence, it can be concluded that for predicting solar irradiance for the available size of dataset of HI-SEAS, random forest regressor and cross-validation are the most efficient techniques to achieve the desired goal.

Future Work

Tuning in the random forest regressor for achieving higher r^2 score could have produced better accuracy if more parameters or the amount of data available was larger. Based on the dataset available from HI-SEAS habitat of Hawaii, model would have been unlikely to produce better accuracy based on available parameters and its limited size. In order to create a more efficient model, a regressor could be trained on larger dataset, or data recorded for more years.

Different types of regressors might perform better than random forest regressor when trained on larger dataset, model and feature selection plays an important role when it comes to fitting in such cases. If such amount of radiation levels in photovoltaic systems could be measured on minute time index, it would revolutionize solar technology to higher levels and improve energy sector.

Bibliography

Ahmad Alzahrani, Pourya Shamsi, Cihan Dagli, Mehdi Ferdowsi, "Solar Irradiance Forecasting Using Deep Neural Networks", Procedia Computer Science, Volume 114, 2017, Pages 304-313, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2017.09.045.

Bamisile, O., Cai, D., Oluwasanmi, A. *et al.* "Comprehensive assessment, review, and comparison of AI models for solar irradiance prediction based on different time/estimation intervals". *Sci Rep* **12**, 9644 (2022). https://doi.org/10.1038/s41598-022-13652-w

Brahim Belmahdi, Mohamed Louzazni, Abdelmajid El Bouardi, "One month-ahead forecasting of mean daily global solar radiation using time series models", Optik, Volume 219, 2020, 165207, ISSN 0030-4026, https://doi.org/10.1016/j.ijleo.2020.165207.

Huang, L., Kang, J., Wan, M., Fang, L., Zhang, C., and Zeng, Z., "Solar Radiation Prediction Using Different Machine Learning Algorithms and Implications for Extreme Climate Events", Frontiers in Earth Science, vol. 9, 2021. doi:10.3389/feart.2021.596860.