May-June 2020
ISSN: 0193-4120 Page No. 8883 - 8891

Predicting Best Regression Model for Prediction of Global Irradiance on Solar Dataset

Raunak Kumar Jha, Ria Sinha, Rishavh Saxena, Mr Amit Rai
Department of Electrical and Electronics Engineering, Galgotias College of Engineering and Technology, Dr. A.P.J Abdul Kalam Technical University

Article Info
Volume 83
Page Number: 8883 - 8891
Publication Issue:
May - June 2020

Article History
Article Received*: 19 November 2019*
Revised*: 27 January 2020*
Accepted*: 24 February 2020*
Publication*: 18 May 2020*

Abstract:
Solar energy refers to the process of capturing the energy from the sun and then converting it into the electricity. After then we can use that electricity to light up our houses, streets, and power our machines as well. It is very important because it is a safe alternative which can replace current fossil fuels like coal and gases for generation of electricity. [7]
Solar Irradiance is defined as the power per unit area, received when energy of the sun is captured in the form of the electromagnetic radiations. When all the energy is measured it is known as Total Solar Irradiance. When it is measured as a measure of wavelength it is known as spectral Irradiance.
The S.I Unit of Irradiance is watt per meter square. [9]
Predicting solar Irradiance has been an important topic in renewable energy generation nowadays. Prediction improves the planning and operation of the photovoltaic systems and yields many economic advantages for the electric utilities. They are also used for planning the development of the solar power systems. Different solar power technologies use different components of the total solar Irradiance. Accurate forecasting of solar energy also plays an important role in the integration into the grid. [11]

Keywords*: Solar Irradiance, Irradiance,* fossil fuels.

INTRODUCTION

Production of electrical energy in most of the countries is done by the fossils fuels resources but it is a type of renewable sources of energy which can get extinguished. Therefore, it has been replaced by non- renewable sources like solar energy so that they can reduce the strain in the environment.
There is a wide increment in the growing demand for environment friendly energy. So the generation and production of renewable energy makes it a perfect profit future material. Solar energy is obtained from the sun and is also one of the renewable energy sources.
Everyday earth receives sunlight above 1366w approx. This is an unlimited source of energy which is fully free of cost. The other advantages of this solar energy as compared to the conventional sources are that it can be directly converted from sun rays to solar energy with the help of photovoltaic (PV) cells. This energy is distributed among wide range of geographical range and has negligible maintenance cost. [12]
Solar radiation is all radiant energy emitted by the sun. Solar irradiance is the power per unit area received from the sun in form of electromagnetic radiation.
The goal of our project is to use the machine learning techniques to figure out the best model for the prediction of global irradiance. The machine learning techniques can be divided into 2 types: - Supervised and Unsupervised Learning. Further supervised learning can be classified into two types 'Classification and Regression'.

May-June 2020
ISSN: 0193-4120 Page No. 8883 - 8891

Regression models that we will be using for the prediction are:-

1. LINEAR REGRESSION
2. SUPPORT VECTOR REGRESSOR MODEL
3. RANDOM FOREST REGRESSION MODEL
4. POLYNOMIAL REGRESSION

By using these models, we would find out the best model that gives the best accuracy for the prediction of Global Irradiance. [13]

INTRODUCTION OF OUR DATASET:

The dataset on which we are working is a time series dataset which consists of 17,520 entries. It consists of following columns: Global irradiance, H_sun (height of sun), T2m (dry bulb temperature at 2m), Ws10m (Wind speed at 10m). The sample of our dataset is given below.

	time	G(i)	H_sun	T2m	WS10
0	20131231:2349	0	0.00	9.12	2.84
1	20140101:0049	0	0.00	9.61	2.57
2	20140101:0149	0	0.00	10.09	2.30
3	20140101:0249	264.87	11.08	10.58	2.03
4	20140101:0349	509.45	21.48	11.45	1.89

METHODOLOGY

Machine learning is an application of the artificial intelligence that enables the machines to access the different types of data and then tells how to perform the specific task with the help of the various algorithms. It also allows the machines to find insight and at time fix anomalies. There are many advantages of machine learning such as it helps people to work more creatively and efficiently. It also performs complex problems very easily.

It can be divided into two types: Supervised and unsupervised learning.

Supervised Learning is the branch of Machines where we deal with Labelled data. A labelled data supervises the learning of a model by telling the model what answers to give under what circumstances, hence the name- Supervised Learning.

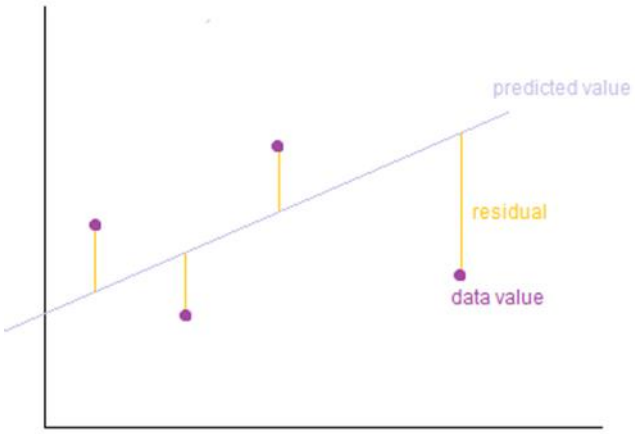
Unsupervised learning is the branch of machine learning where we deal with Unlabelled data. Since there is no dependent/target variable, there is no way to evaluate the correctness of the model. Unsupervised learning is used to cluster entities with similar behaviour. [14]

Supervised Learning can be classified as Classification and Regression. The type of supervised learning which we are using to make our model learn for prediction is Regression.

Regression is a type of supervised learning in which dependent data is in continuous form. The regression plot is a scatter plot and each point shows\representation of row of our data. We find the relationship between dependent and independent variables in regression through graph. The line for which the error between the predicted values and the observed values is minimum is called the best fit line

May-June 2020
ISSN: 0193-4120 Page No. 8883 - 8891

or the **Regression Line.** The residuals(error) can be visualized by the vertical lines from the observed data value to the regression line.



To define and measure the error of our model we define the cost function as the sum of the squares of the residuals The cost function is denoted by:

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h(x^i) - y^i)^2$$

Where the hypothesis function h(x) is denoted by:

$$h(x) = \theta_0 + \theta_1 x_1 + ... + \theta_n x_n$$

and m is the total number of training examples in our dataset.

Our objective is to find the model parameters so that the **cost function is minimum**. We will be using Gradient Descent to find this. [15]

Regression is divided into various other types: Simple Linear regression, Multilinear Regression, Polynomial Regression, Decision Trees regression, Random Forest Regression, Support vector Regression. [4]

Here are the descriptions of the models that we are using for our predictions:

Simple Linear Regression Model: It is a univariate regression technique. In simple words linear regression is predicting the value of a dependent variable and independent variable provided that there is a linear relationship. Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous variables. This linear relationship between the two variables can be represented by a straight line (called Regression line). The linear regression model can be represented by the following equation:

$$Y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_n x_n$$

where,

1. Y is the predicted value
2. 2. θ_0 is the bias term
3. $\theta_1,\ldots,\theta_n$ are the model parameters
4. 4. $x_1, x_2,\ldots,x_n$ are the features values.

Support Vector Regressor model: The method of support vector Classification can be extended to solve regression problems. This method is called Support Vector Regression. In case of Support Vector Classification, the model produced by support vector depends only on a subsets of the training data, because the cost function for building the model does not care about training points that lie beyond the margin.

Random Forest Regressor is a linear/ non-linear, multivariate, non-parametric, ensemble regression technique. Random Forest Regressor (RFR) is an example of Ensemble learning. It predicts the answer by taking the mean of the prediction of the decisions trees contained within the forest. It can be applied on linear as well as nonlinear relationship. It can handle multiple independent variables.

Polynomial Regression Model: Polynomial Regression is a form of linear regression in which the relationship between the independent variable and the dependent variable is modelled as a nth degree polynomial. We use polynomial regression because of:

Hypothesis: There are some relationships that a researcher will hypothesize is curvilinear. Clearly, if this is the case, include polynomial terms.

May-June 2020
ISSN: 0193-4120 Page No. 8883 - 8891

Visual inspection of your variables: This is one of those reasons for always doing univariate and bivariate inspection of your data before you begin your regression analysis. A simple scatter plot reveals a curvilinear relationship.

Inspection of residuals: If you try to fit a linear model to curved data, scatter plot of residual (Y axis) on the predictor (X axis) will have patches of many positive residuals in the middle. This is good sign that a linear model is not appropriate, and a polynomial may do better. [16]

RESIDUAL PLOTS: The difference between the value observed of the dependent variable and the predicted value is known as error / residual.
Residual= Observed value- Predicted value
Both the sum and mean of the residuals are equal to zero.
The type of graph which shows the residuals / errors on its vertical axis and independent variable on its horizontal axis is known as Residual plots. A regression model is fit for the data can be detected by seeing whether the residual plots are randomly dispersed around its horizontal axis. In case if it doesn't then the non-linear model is more appropriate. [17]

IMPLEMENTATION:

The software which we will be working for coding are Spyder and Jupyter. The platform to get these software is Anaconda. It is a free and open source distribution of Python programming language for scientific calculation such as machine learning applications, large scale data processing, predictive analytics etc.
The software which we have preferred for the coding of our project is Jupyter. There are some libraries like Numpy, Pandas, Seaborn, yellowbricks which we have used for pre-processing of our data and plotting.
Numpy and Pandas are the open sources that are used for the scientific computing. Out of Numpy and Pandas, Pandas are most widely used because it provides high performance, easier data analysis tools. These libraries are used for the pre-processing of the data. The other two libraries Seaborn and yellowbricks are used for the data visualizations. [3]
Now let's take a look at the steps that we have applied for the coding of our project:

1. Firstly, we need to import all the required libraries that we would be using while coding our problem statement.
2. The Next step would be reading the dataset on which we would be working in Jupyter Notebook. We took the help of Pandas Library for this step.
3. After reading out the dataset, we have done the datapreprocessing under which:
   - We have dropped the unnecessary columns after finding the correlation matrix.
   - The necessary steps require in data preprocessing are:
     Handling missing values: under which we find out the total numbers of missing values in each row of the datasets.
     Encoding Categorical Variables: Sometimes there are columns which contains categorical values and which we cannot use in our code for future purposes. Therefore, it is compulsory to convert those categorical values into integers values. This can be done by following two methods One Hot Encoding and Label Encoding.
     Feature Scaling: In Python's language, Feature is a synonym used in place of column. So basically under this step, we need to scale all those columns of dependent variables which are having a very huge difference. This will help us in getting better results from the model. [6]
4. Now, after finding the relationship between the columns by using the correlation matrix, we have seen the following relations with the help of visualisations.

May-June 2020
ISSN: 0193-4120 Page No. 8883 - 8891

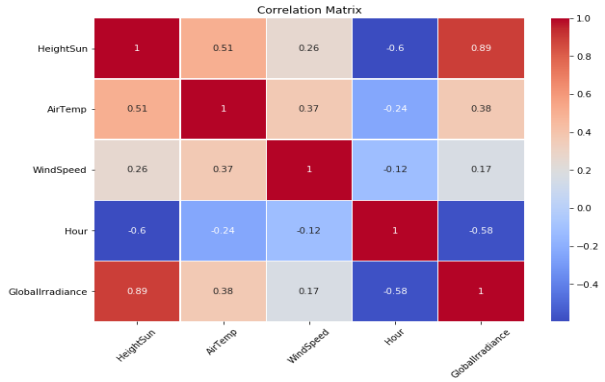


Figure 1 (Corr matrix vis.)

Fig.1 is a correlation matrix which depicts the values of the dependency between the the independent and dependent variables.

GLOBAL IRRADIANCE VS AIR TEMPERATURE

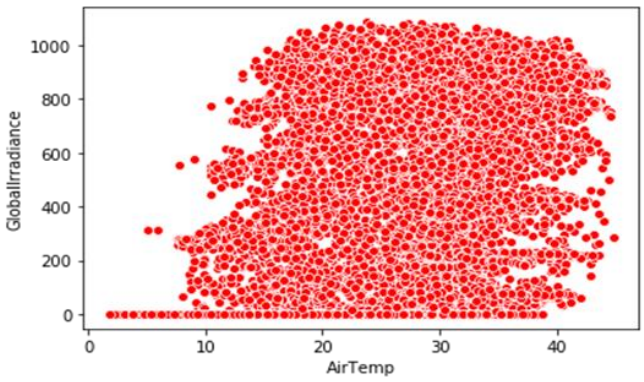


Figure 2

This graph shows that there is around 17% affect of air temperature on global irradiance which can be taken into consideration but does not affect very much.

GLOBAL IRRADIANCE VS HEIGHT OF THE SUN

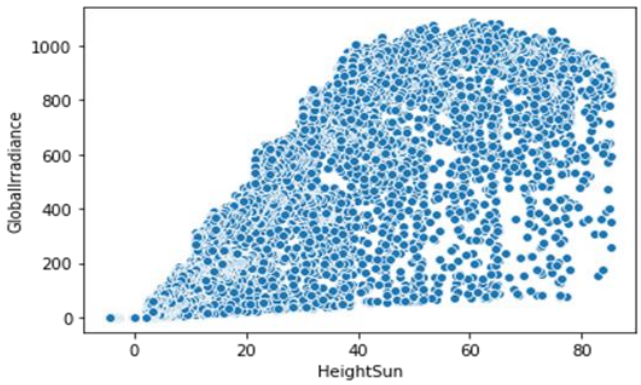


Figure 3

According to correlation matrix,Global irradiance has the highest dependancy value on Height of the sun with around 89%.

GLOBAL IRRADIANCE VS WIND SPEED

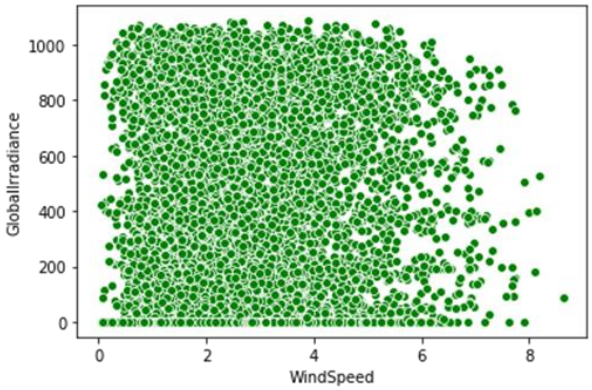


Figure 4

There is around 28% dependency of global irradiance on the speed of the wind which we can see in the correlation matrix in fig.1.

5. After data preprocessing is done, we have divided our data into dependent and independent variables. This is followed by giving the data into training and testing dataset using the scikit learn library under which 75% of the data is given to training dataset and 25% is given to the testing dataset. We do so to train our model and once the model is being trained, we can do the predictions on the basis of this training dataset.

6. Once this is done, then we have reached the last step of the code which is to use different types of regression model and then find the best model with high accuracy rate and low root mean square error for the predictions of our problem statement.

7. The accuracy rate of any model depends upon the hyper parameters. Machine learning models are parameterized so that their behaviour can be tuned differently for different problems. These models can have many parameters and finding the best combination of parameters is known as **Grid**

May-June 2020
ISSN: 0193-4120 Page No. 8883 - 8891

Search process which is used for model tuning.

8. There is another way of improving our model accuracy which is **K-Fold Cross Validation.** In K-Folds cross validation we split our data into k different subsets or folds. We use k-1 subsets to train our data and leave the last subset as test data. We then average the model against each of the folds and then finalize our model. After that we test it against the test set.

So, these are the steps that we have done to make our model train for the predictions and find out the best model to solve our problem. [8]

RESULT AND DISCUSSION:

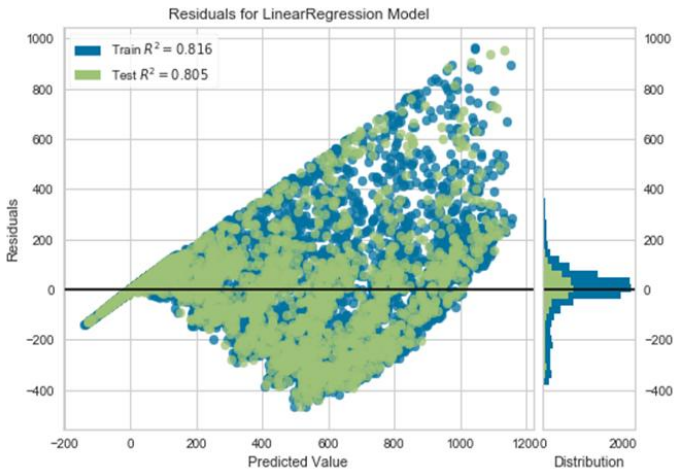
The variables most relevant to the prediction of solar irradiance were found to be wind speed, height of sun, dry bulb temperature. Training the various type of regressor models such as random forest regressor etc. with these three variables produced a model that achieves different root mean squared error and accuracy score. The below mentioned table is the list of all the models that has been used along with their accuracy score and the respective Root mean squared error values.

S.NO	NAME OF THE MODEL	ACCURACY SCORE(%)	ROOT MEAN SQUARE ERROR
1.	LINEAR REGRESSION	81.20	148.962
2.	SUPPORT VECTOR REGRESSOR	84.50	135.233
3.	RANDOM FOREST REGRESSION	92.51	93.997
4.	POLYNOMIAL REGREESION	93.32	88.798

There are possibilities of tuning all the models to obtain an even higher accuracy score by using **GridSearch CV** but realistically the model produced is unlikely to be any better at solar irradiance prediction, it will just fit the observations of one particular year extremely well. To create a more useful model, the regresssor should be trained on data recorded over several years. [6]

Residuals Plot using Yellow Brick Library for Each Regression Model

1. **LINEAR REGRESSION**



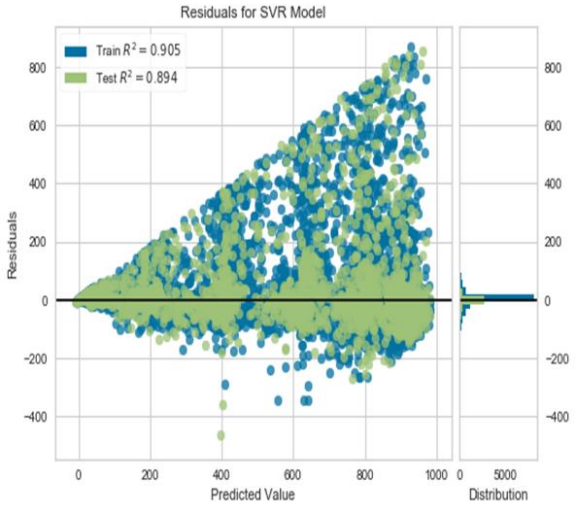
The residuals are unevenly distributed. This shows that a non-linear regressor might perform better on the dataset.

Majority of the residuals seem to be positive, indicating that in most of the cases, our predictions are higher than the actual answers

For smaller predictions (leftmost predictions), we have obvious errors as Global Irradiance can't be negative. Residuals are widely spread in both directions; this may indicate that our model is Under fitted to the dataset [7]

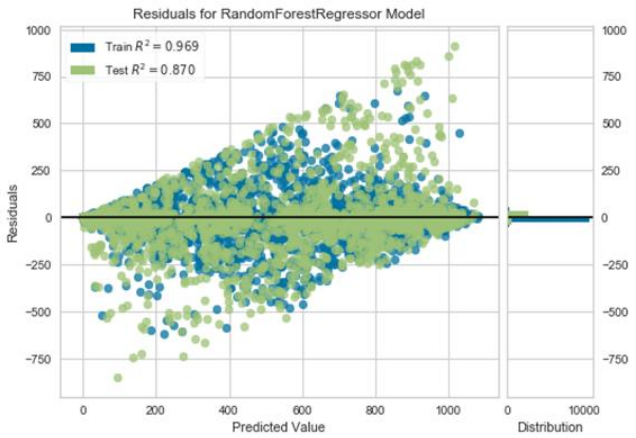
2. **SUPPORT VECTOR MODEL**

May-June 2020
ISSN: 0193-4120 Page No. 8883 - 8891



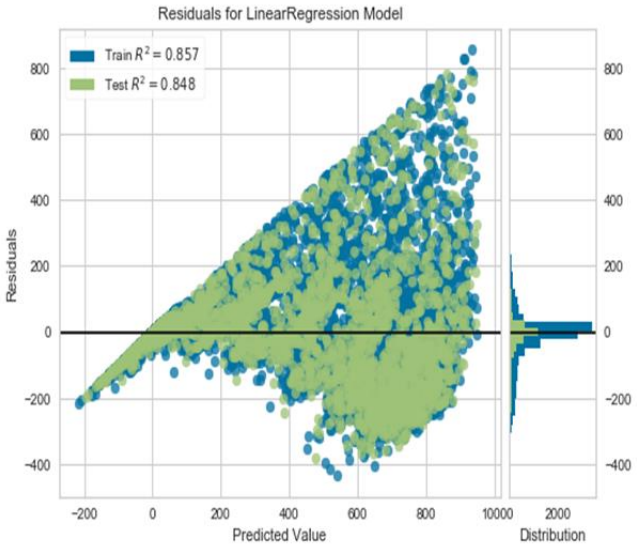
The residuals aren't distributed at random. This shows that a non-linear regressor like this is a good choice for this dataset. Majority of residuals are along the 0 axis, suggesting that the model makes lesser errors in prediction The residuals on the test dataset are negative in the mid-range of predictions, but higher if the expected value is higher. There is not much difference in the training and testing residuals, indicate that our model is very nicely fitting the dataset, no Overfitting or Under fitting.

3. RANDOM FOREST REGRESSOR MODEL



On the Train dataset the model is working very well The residuals are not distributed randomly. This shows that a regressor like this is a good choice for this dataset Majority of residuals are along the 0 axis, suggesting that the model makes lesser predictive errors.The residuals on the test dataset are negative in the mid-range of predictions, but higher if the expected value is higher as there is a lot of variance in the monitoring and testing residuals, this could mean that our model is slightly overfitting on the train dataset.

4. POLYNOMIAL REGRESSION MODEL



The residuals aren't distributed at random. This shows that a non-linear regressor on the dataset will do better. Most residuals tend to be negative, suggesting that in most cases our estimates are lower than the actual answers Residuals are widely distributed in both directions; this may mean that our model is Underfitted.

Findings and Conclusion

Solar power is a vast, free and renewable resources that can be used to produce electricity. Solar generated electricity produces no greenhouse gases or emissions of any kind. Solar energy is a commercially proven, rapidly growing form of electricity generation. Machine learning algorithms were used to predict the solar irradiance. Parameters such as global irradiance, wind speed, height of sun

May-June 2020
ISSN: 0193-4120 Page No. 8883 - 8891

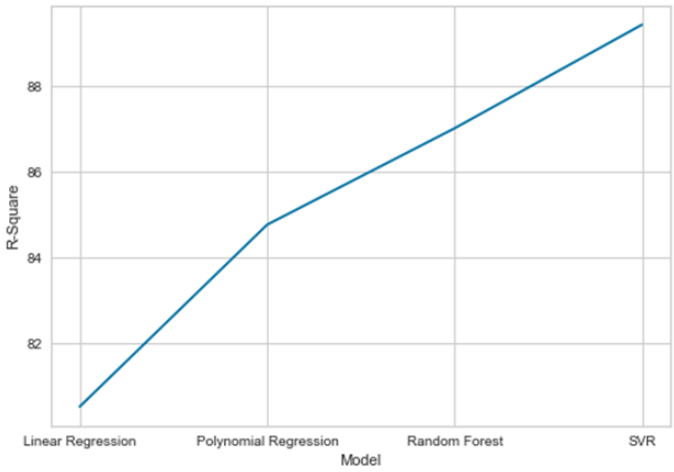
and lastly air temperature was given to train the given data.

Machine learning algorithms such as Linear Regression Model, Random Forest Model, Polynomial Model and state vector model were used. But out of these the best accuracy score was of the Polynomial Regression Model. The accuracy shown for the models are listed below:

S.NO	NAME OF THE MODEL	ACCURACY SCORE(%)	ROOT MEAN SQUARE ERROR
1.	LINEAR REGRESSION	80.51	151.057
2.	SUPPORT VECTOR REGRESSOR	89.44	111.243
3.	RANDOM FOREST REGRESSION	87.02	123.333
4.	POLYNOMIAL REGREESION	84.77	133.579

It may be possible to tune the random forest regressor to achieve an even higher r_2 score, but realistically the model created is unlikely to be any better on the prediction of solar irradiance, it will only match extremely well with the observations of a given year. The regressor should be trained on data collected over a period of several years to construct a more reliable model.

Likewise, other types of regressors that perform better than the random forest regressor when trained on the same set of features, was put into model selection after achieving such a high r_2 score with the random forest regressor.



So, at the end by looking at the graph we can say that SUPPORT VECTOR REGRESSION is the best fitted model for our problem i.e. to predict the Global Irradiance and leads the Linear Regression model **by 11.1 %** and from Random Forest **by 2.5 %** , Polynomial regression model **by 7 %.**

References

1. https://pandas.pydata.org/pandas-docs/stable/getting_started/10min.html
2. https://www.wikipedia.org/
3. https://scikit-learn.org/stable/
4. https://www.sciencedirect.com/science/article/pii/S2210832719302108
5. https://link.springer.com/chapter/10.1007/978-3-319-10422-5_29
6. https://www.google.com/amp/s/www.igs.com/amp/machine-learningsolar7technology/RGY1aVFoUFVhU0FHb0RzeVhHUys3OXgrR1JNPQ2
7. https://www.researchgate.net/publication/335880904_Applying_Data_Science_to_Improve_Solar_Power_Production_and_Reliability
8. ht https://www.saveonenergy.com/learning-center/post/3-challenges-facing-the-solar-energy-industry/ https://towardsdatascience.com/tagged/solar-energy
9. https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861
10. https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/

May-June 2020
ISSN: 0193-4120 Page No. 8883 - 8891

11. https://towardsdatascience.com/linear-regression-using-python-b136c91bf0a2
12. https://www.saurenergy.com/solar-energy-articles/the-role-of-ai-and-ml-in-solar-energy
13. https://www.google.com/amp/s/www.igs.com/amp/machine-learning-solar-technology/RGY1aVFoUFVhU0FHb0RzeVhHUys3OXgrR1JNPQ2
14. https://www.google.com/amp/www.cleanfuture.co.in/2019/03/07/ai-ml-are-being-used-extensively-in-renewable-energy-space/amp/
15. https://github.com/shashanksira/Solar-Radiation-Prediction
16. https://analyticsindiamag.com/ai-has-become-a-key-contributing-factor-in-the-renewable-energy-sector/
17. https://pexapark.com/blog/technology/renewable-energy-machine-learning-artificial-intelligence/