STATISTICAL ANALYSIS AND FORECASTING OF SOLAR ENERGY

Table of Contents

Introduction

We can employ cutting-edge technology, such as thermal power plants, solar power systems, artificial photosynthesis, etc., to harness solar energy, which is the radiant light and heat energy from the sun. It is a type of renewable energy that is regarded as the most trustworthy energy source at the moment because it is abundant, pollution-free, and renewable.

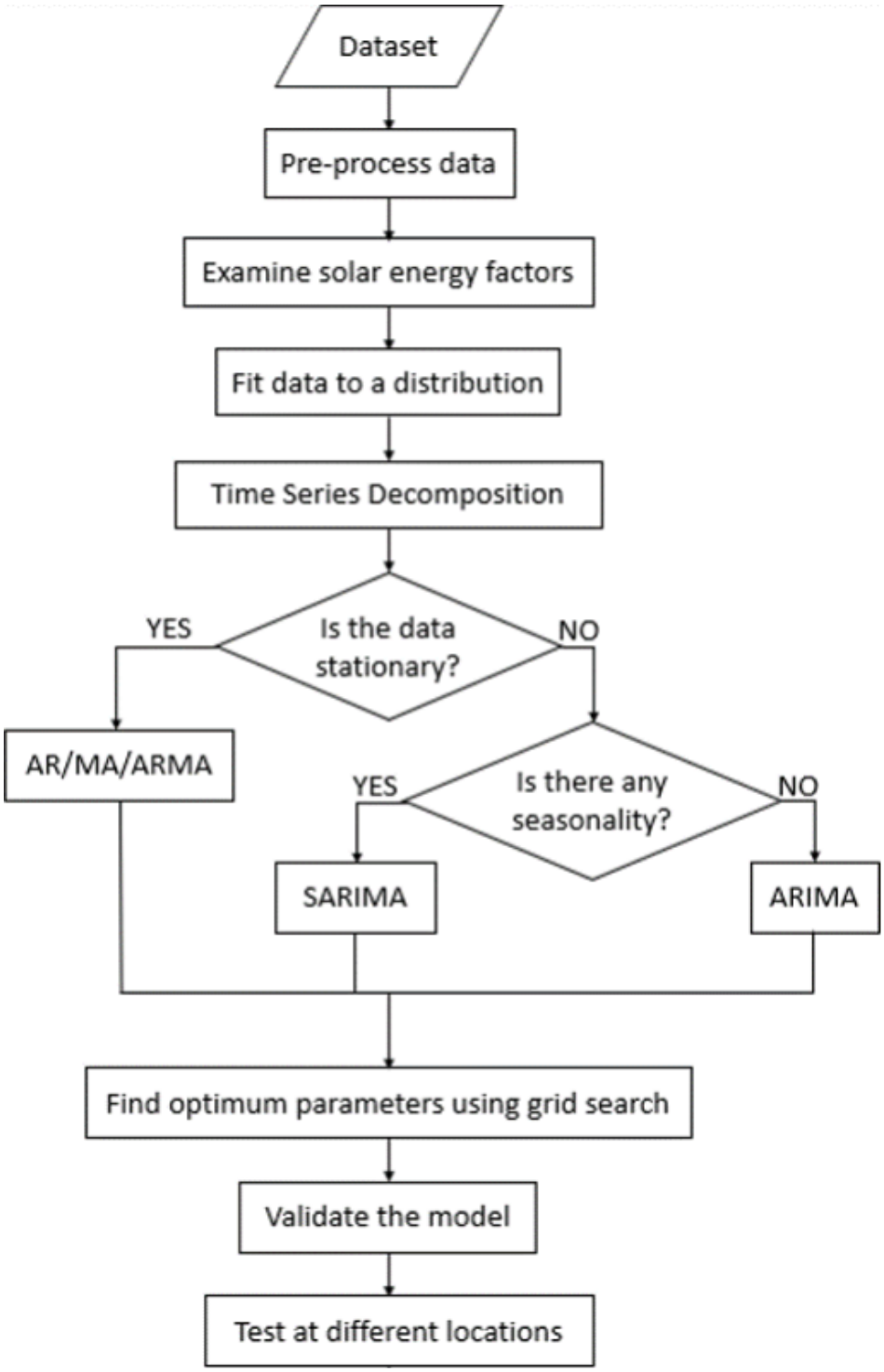
In response to environmental issues including pollution, the adoption of renewable energy sources is rising in popularity. India's National Time series value Change effort now includes solar power as a crucial element, and the National Solar Mission is one of the nation's most significant missions.

For the creation of tools that can harvest solar energy more effectively, forecasting and analysis of solar energy are consequently crucial. The main focus of this report is the intrastate examination of solar energy statistics. Datasets acquired from a range of solar farms in Rajasthan are analysed using a number of statistical techniques.

The report's approach to its methodology is broken down into the following components for the most part:

- Review of Solar Energy Parameters
- Methods for selecting an appropriate distribution fit for the GHI data
- Analysis and breakdown of time series and stationary data analysis
- Forecasting using a variety of different models, including AR, MA, ARMA, ARIMA, and SARIMA
- Conclusion

Methodology

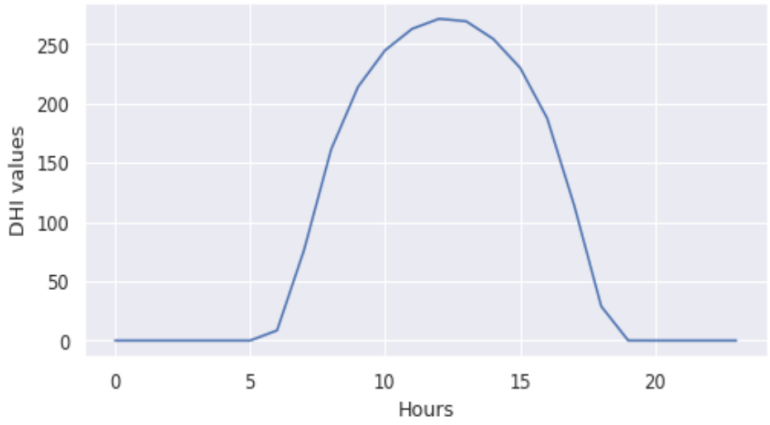


Pre-processing

The findings indicate that a number of variables influence how much solar radiation reaches the Earth's surface. To better understand the situation and to see how the various factors affect the solar radiation value as well as one another (via correlation), let's first define these words.

DHI (Diffuse Horizontal Irradiance)

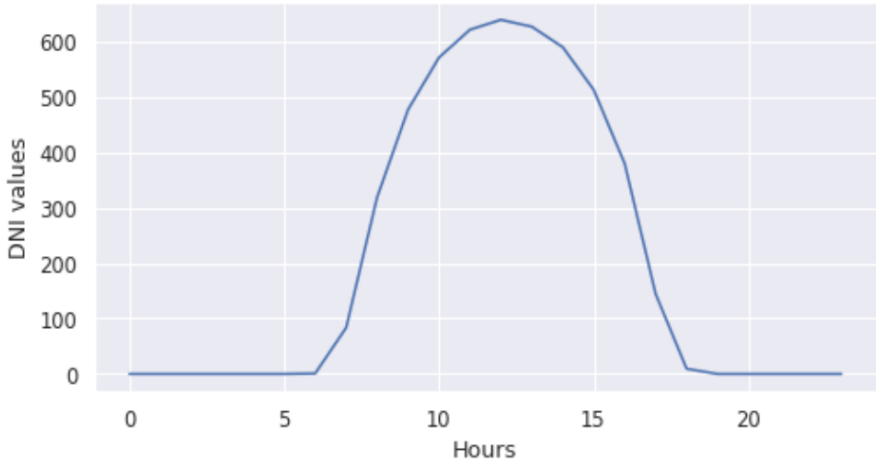
A surface's Diffuse Horizontal Irradiance (DHI) is the amount of radiation it receives per unit area that does not arrive via a straight channel from the sun but is instead scattered by particles and molecules in the atmosphere.
We have charted the DHI daily pattern (fig 2.1) and found that DHI is reaching its peak around



midday, or the 12th hour.

DNI (Direct Normal Irradiance)

The quantity of solar radiation that a surface receives per unit area when it is always held normal (or perpendicular) to the rays that come in a straight line from the direction of the sun at its current position in the sky is known as the Direct Normal Irradiance (DNI).
Daily DNI (Fig. 2.2) followed the same peak-at-12-noon pattern as DHI.

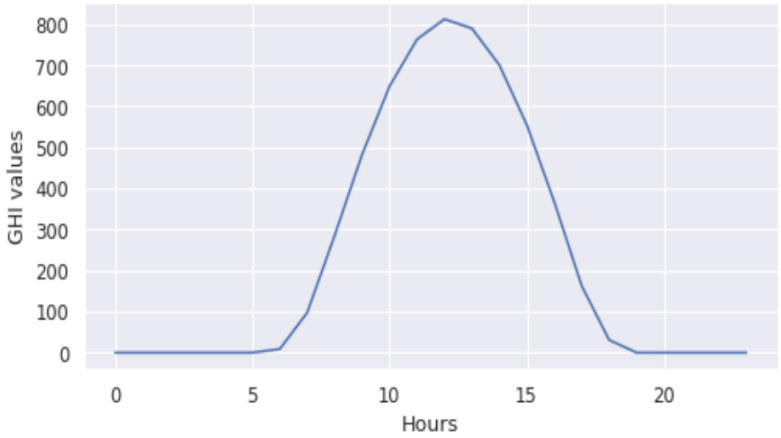


Global Horizontal Irradiance(GHI)

Shortwave radiation from the sun that reaches Earth's surface in a horizontal direction is measured as global horizontal irradiance (GHI). The number is a combination of the Direct Normal Irradiance (DNI) and the Diffuse Horizontal Irradiance (DHI) at the given Solar Zenith Angle(θ)

$$\text{Global Horizontal Irradiance (GHI)} = ((\text{DNI}) \times \cos(\theta)) + (\text{DHI})$$

While DHI and DNI follow roughly the same pattern throughout the day, peaking about noon, GHI's daily trend is slightly less steep.



Dew point

The Dew Point is the temperature at which water droplets begin to condense and dew begins to form in the atmosphere.

Relative humidity

It is the amount of water vapor in the air as a percentage of the amount required for saturation at the same temperature.

Solar zenith angle

It is the angle formed by the vertical and sun’s ray.

Snow depth

It is the vertical height of snow at the ground at the normal observation time.

Wind speed

It is the rate at which wind moves at the observation point

Conclusion

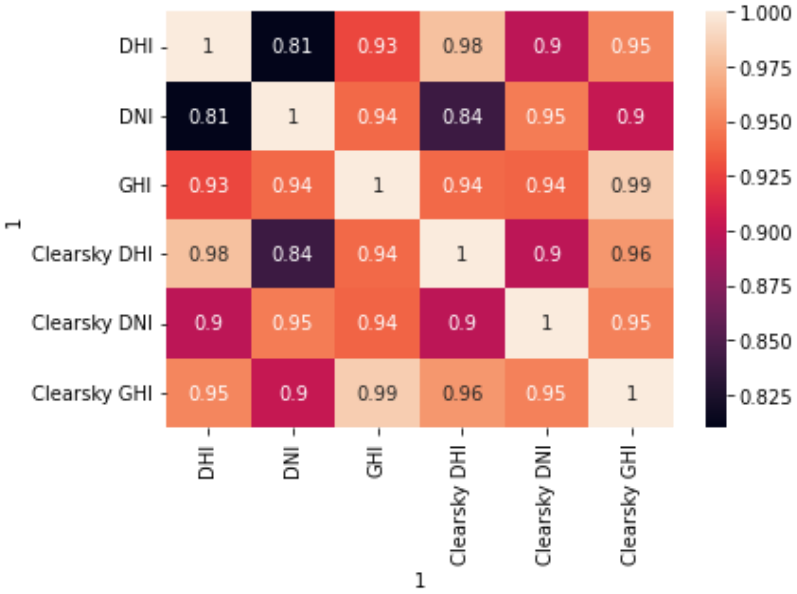
DHI, DNI, GHI, temperature, and relative humidity are relevant to the measuring of solar energy. Dew point is dependent on temperature and relative humidity, making it significant to solar energy as well.

Correlation matrix

	DHI	DNI	GHI	Clearsky DHI	Clearsky DNI	Clearsky GHI
DHI	1.000000	0.810126	0.927168	0.978921	0.897983	0.952093
DNI	0.810126	1.000000	0.940300	0.840339	0.947746	0.902440
GHI	0.927168	0.940300	1.000000	0.940370	0.938249	0.985022
Clearsky DHI	0.978921	0.840339	0.940370	1.000000	0.898374	0.959677
Clearsky DNI	0.897983	0.947746	0.938249	0.898374	1.000000	0.949993
Clearsky GHI	0.952093	0.902440	0.985022	0.959677	0.949993	1.000000

Rajasthan1 - Correlation Matrix

Making a heatmap (Fig. 2.1) of the correlations between all of the variables is the first step in making sense of our data. However, as we have only utilised the Rajasthan 1 dataset from 2000 to 2014 in this paper, it is probable that the data for other solar parks would also exhibit similar pattern for the different parameters.



Rajasthan1 - Correlation Heatmap

Using the heatmap, we can observe that the GHI, DHI, and DNI for Clearsky, respectively, have strong positive relationships. Every feasible pair of these factor-based pairs has a correlation of at least 0.80. We can confidently select one to analyse given the close ties between the aforementioned parameters. Since it considers both the diffuse and direct irradiance from the sun, as well as the solar zenith angle, the Global Horizontal Irradiance (GHI) is utilised.

GHI Data Analysis

We used R software to determine which model was most appropriate for the data, and to check for evidence of a specific distribution pattern in the datasets in question. Specifically, we get the following findings:

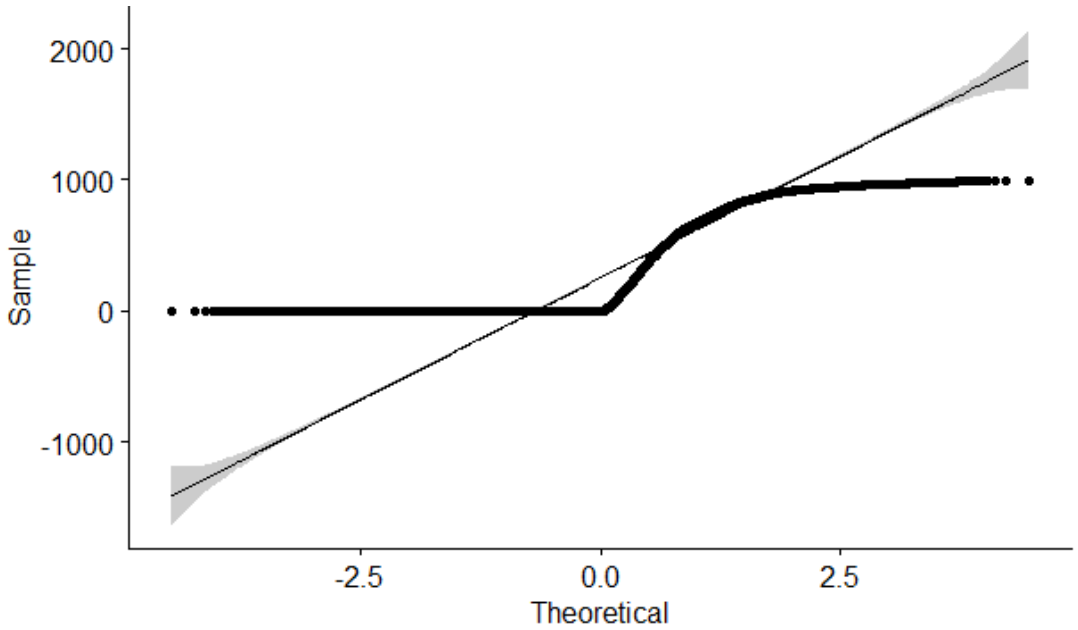


Fig1:- QQ Plots

The same result can also be obtained by carrying out (Kolmogorov–Smirnov) or KS Test. It is a non-parametric test of the equality of continuous one-dimensional probability distributions. Its test is defined as:

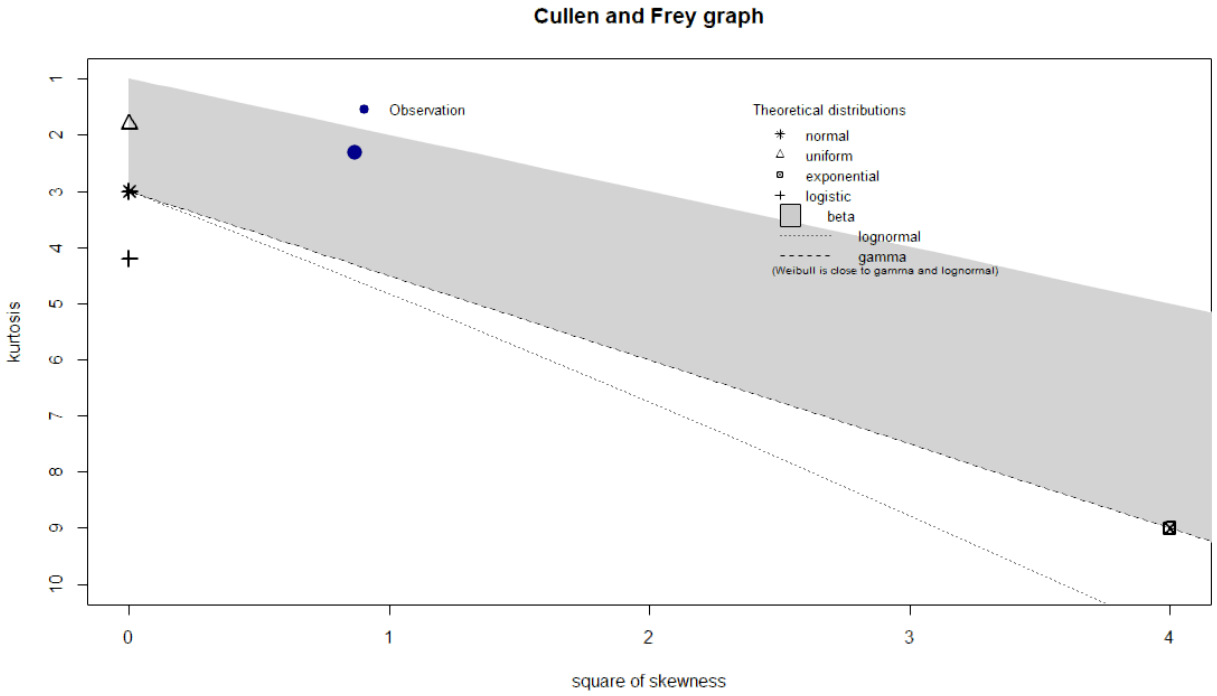
H0: The data follows a specified distribution(Null Hypothesis)
H1: The data doesn't follow a specified distribution(Alternate Hypothesis)

Alpha, representing the significance level, is compared to the KS statistic value. When the fit is strong, the KS statistic value should be close to 1 (Max = 1.0), while when the fit is poor, it should be closer to 0 (Min = 0.0). The resulting information is:

```
> ks.test(Combined$GHI, "pnorm")

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  Combined$GHI
D = 0.5, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Here, p-value< α(level of significance), thus indicating that the given data does not follow normal distribution.

We tried fitting the GHI Data to several probability distributions. For this purpose, we have used the Cullen and Frey Graph which displays the relation between Kurtosis and square of skewness. The graph plotted is as follows:-




```
summary statistics
------
min:  0   max:  995
median:  0
mean:  237.8901
estimated sd:  315.3451
estimated skewness:  0.9314054
estimated kurtosis:  2.311441
```

From the above graph, it can be seen that beta distribution is the distribution which would yield the least sum of squares, although still the goodness of fit tests does not approve of it.

FORECASTING

Autoregressive (AR) Model

The Autoregressive Model, often known as the AR model, uses solely previous period data to forecast future ones. It is a linear model, where current period data are the result of multiplying the total of previous results by a certain number. We write it as AR(p), where "p" stands for the model's order and the number of lag values we wish to take into account.
For instance, if we take X as time-series variable, then an AR(1), also known as a simple autoregressive model, would look something like this:
$X_t = C + \phi_1 X_{t-1} + \epsilon_t$
X_{t-1} represents the value of X during the previous period.
The coefficient ϕ_1 is a numeric constant by which we multiply the lagged variable (X_{t-1}).
ϵ_t = Residual

In order to predict using AR Model, we first fit the AR model to the GHI values of the dataset using the ARIMA function in R. The results obtained are as follows:-

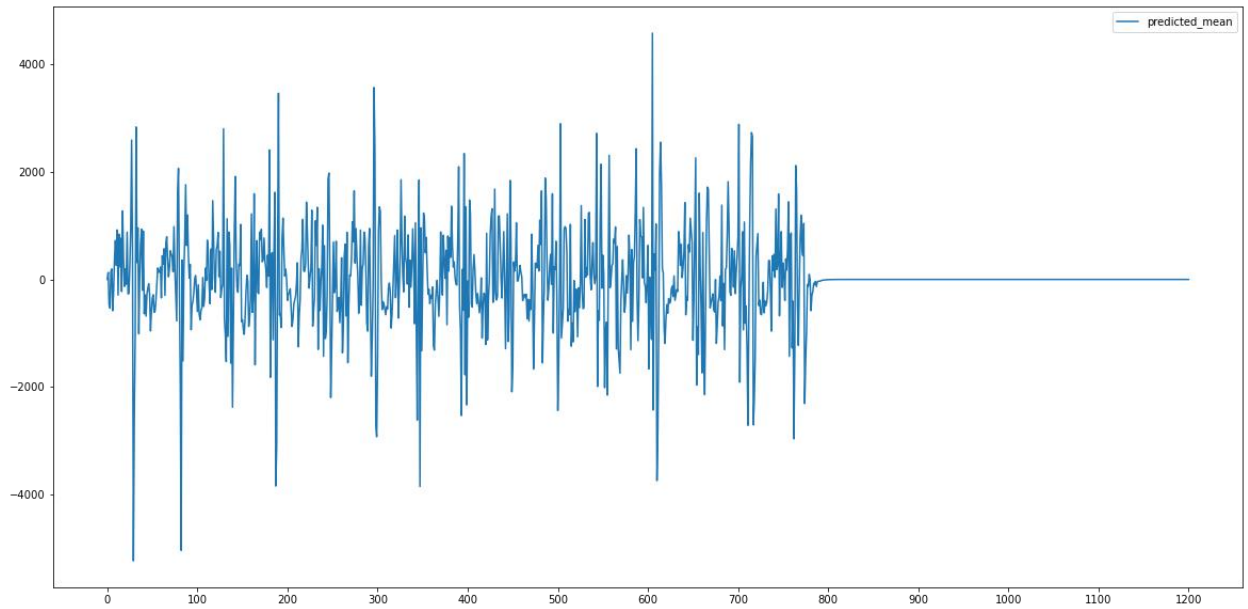
```
> AR<- arima(Combined$GHI, order=c(1,0,0))
> print(AR)

Call:
arima(x = Combined$GHI, order = c(1, 0, 0))

Coefficients:
          ar1  intercept
      0.9358   237.8901
s.e.  0.0010     4.6366

sigma^2 estimated as 12351:  log likelihood = -805441.2,  aic = 1610888
```

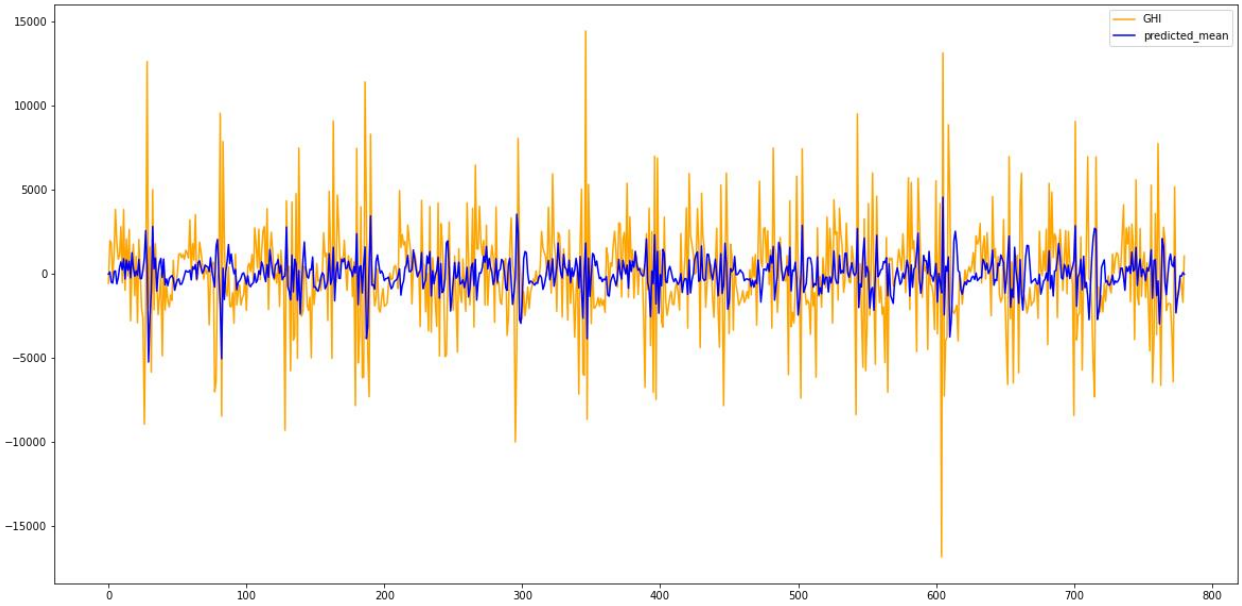
An estimated AR model may be used to predict using the predict() method. The graphs obtained using the AR prediction method is :-

The plot for forecasting using an auto regressive (AR) model is as follows:

Minimum AIC value was obtained to be 14706.814 for the model of order p=6.

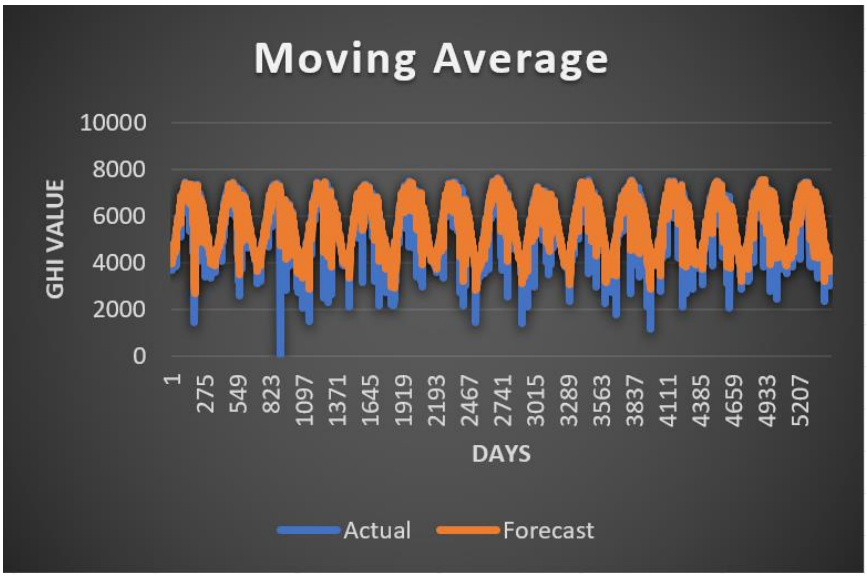
```
MAE = 2137.136
MAPE = 3.572
MSE = 8663585.29
```




GHI
predicted_mean
15000
10000
5000
0
−5000
−10000
−15000
0
100
200
300
400
500
600
700
800

Moving Average Model

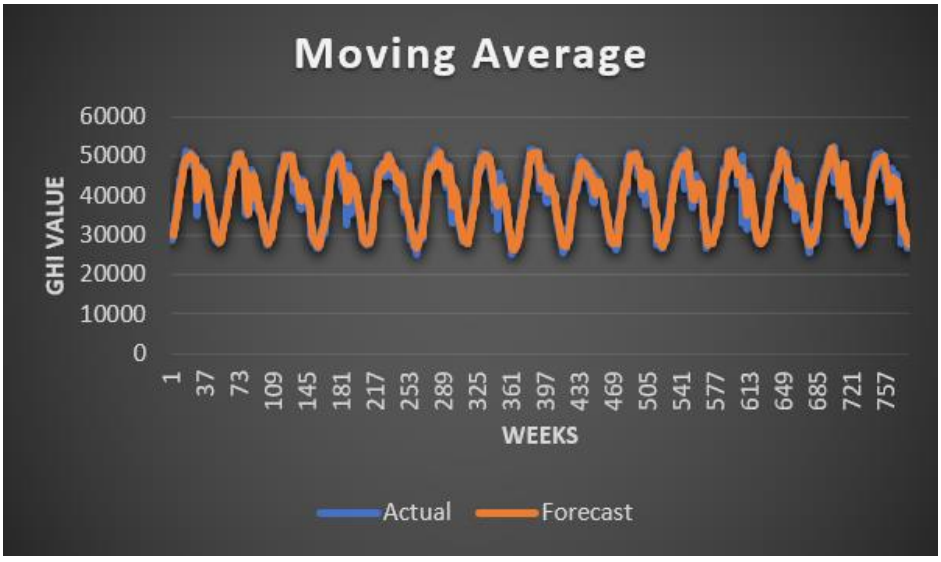
The Moving Average approach allows us to make predictions based on a predetermined window of historical data. By applying the Moving Average model with a window size of 3



, we are able to predict future GHI values with high accuracy(using the daily data)

MAE	MSE	MAPE
221.9051	176540.2	4.524782

For weekly data,



MAE	MSE	MAPE
1584.16	4355122	4.06918

The following is an expression for the MA(q) process:

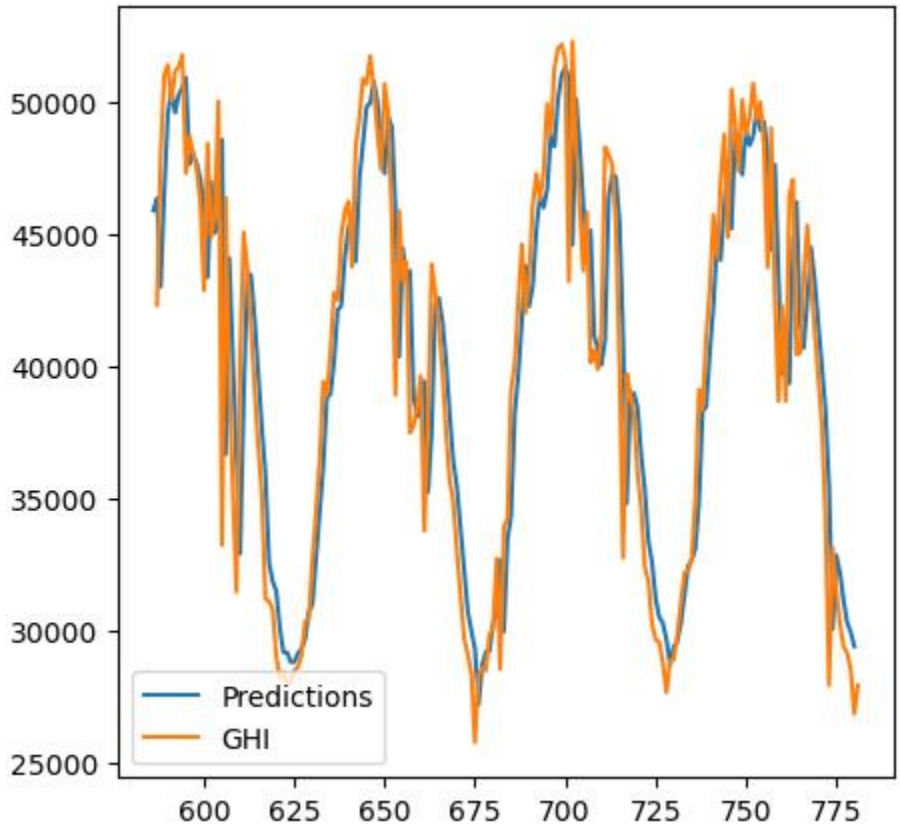
$x_t = (1 - \beta_1 B - \beta_2 B^2 - ... - \beta_p B^p)wt = \phi_q(B)w_t$

where B is a backshift operator and q is a polynomial of degree q. Here, each x_t number represents a weighted moving average of the prior q forecast errors.

Unfortunately, this results in extremely lengthy training times and a high computational cost. The choice of window of moving average also depends on the data. If there is high dependence of new values on previous values, the larger window can be considered for forecasting. As seen from the results of the model, the Moving average model gives quite high forecasting error and hence, is not best suited.Although, with a smaller window gives better results.Following results are obtained when we use a window of interval 3.

Auto Regressive Moving Average (ARMA) Model

ARMA model uses both moving average and auto regression for forecasting future values . The model is defined by two parameters p and q where p is the order of AR model and q is the order of MA model .Here we have taken both p and q to be 1



Where X-axis is the index of the GHI data and Y-axis is the GHI value observed weekly for years till 2014 The AIC value obtained by forecasting using ARMA was 14747.251.

.

If a time series has the form x_t ;t = 0, ±1, ±2, . . then it has the form ARMA(p, q) if it is stationary and if it has the given parameters.

$$X_t = w_t + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{j=1}^{q} \theta_j w_{t-j},$$

$$\text{where } \phi_p \neq 0,\ \theta_q \neq 0,\ \text{and } \sigma_w^2 > 0,\ w_t \sim wn(0, \sigma_w^2).$$

The model may be recast in a more succinct manner with the assistance of the AR operator and the MA operator that we established before as

$$\phi(B)X_t = \theta(B)w_t$$

It's possible that you've noticed that the result of multiplying the same factor by both sides of the equation yields the same result.

$$\eta(B)\phi(B)X_t = \eta(B)\theta(B)w_t$$

If we didn't know about parameter redundancy, we may conclude that the data are associated when, in reality, they aren't at all.

Auto Regressive Integrated Moving Average (ARIMA) Model

The ARIMA model combines autoregression, moving average, and preprocessing "difference." ARMA model cannot work with non-stationary data. The Autoregressive Integrated Moving Average model overcomes this limitation by adding "differencing" to the ARMA model. Each ARIMA model employs three hyperparameters (p,d,q), the meanings of which are comparable to those of the ARMA model's p and q, while d represents the number of times the data must be differenced to produce a stationary output.

ARIMA has the following parameters:

- p: Trend autoregression order
- d: Trend difference order
- q: Trend moving average order

ARIMA model is also written as ARIMA(p,d,q)

P-value from Augmented Dickey Fuller (ADF) Test was obtained as 3.67e-16.
H0: Time series data is non-stationary.
H1: Time Series data is stationary.
As the p-value is very less than significance level 0.05, we can conclude that our data is stationary. Therefore, the ARIMA model will be similar to the ARMA model.

```
ad_fuller_result = adfuller(data['GHI'])
print(f'ADF Statistic: {ad_fuller_result[0]}')
print(f'p-value: {ad_fuller_result[1]}')
```

✓ 0.8s

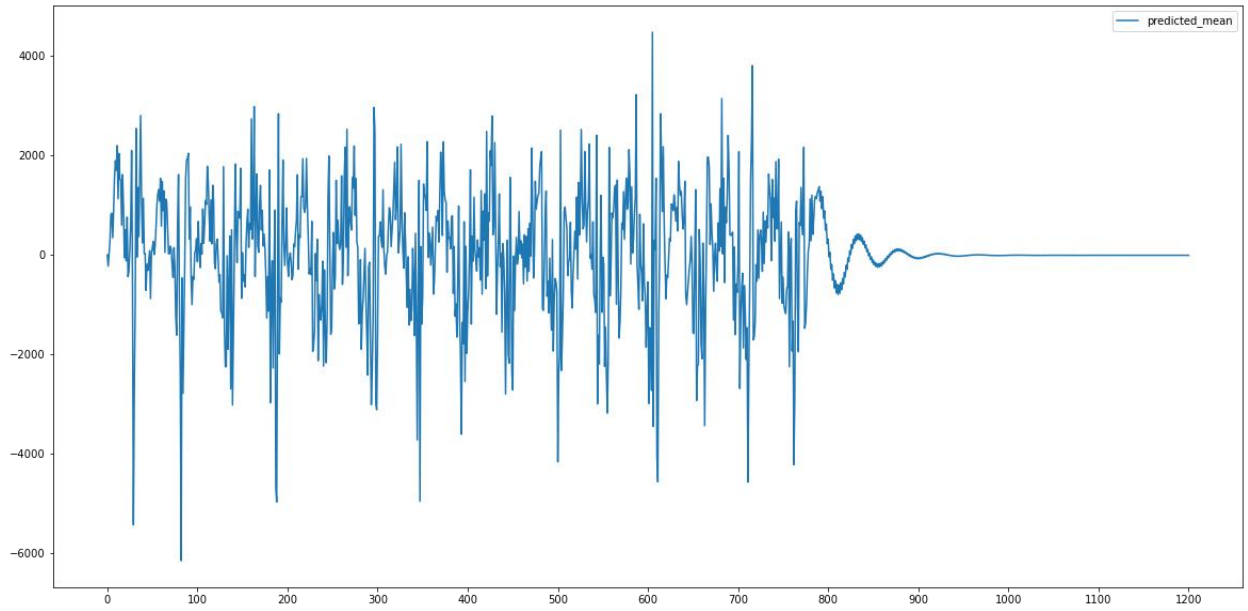
```
ADF Statistic: -9.489838688135503
p-value: 3.6779324671181 82e-16
```

Despite this, a differencing order of 1 was applied on the data prepared by combining the data for all years and converting to weekly form by summing GHI values for a week. We find that the combination (12,1,4) yields the minimum AIC value.

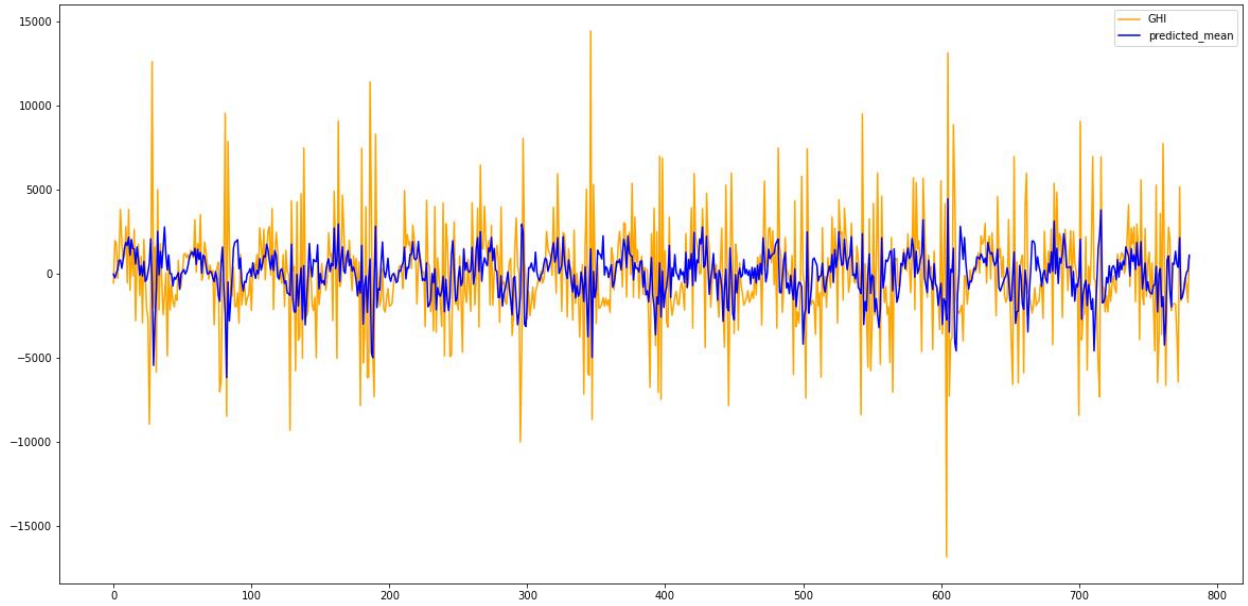
	(p, d, q)	AIC
0	(12, 1, 4)	14664.345554
1	(14, 1, 6)	14667.653330
2	(11, 1, 2)	14672.635652
3	(2, 1, 11)	14683.994771
4	(4, 1, 5)	14684.898595
...	...	...
220	(4, 1, 0)	14782.342682
221	(3, 1, 0)	14835.839571
222	(2, 1, 0)	14955.315523
223	(1, 1, 0)	15120.482980
224	(0, 1, 0)	15475.485213

225 rows × 2 columns

```
MAE = 2041.609
MAPE = 4.449
MSE = 8086807.355
```



The values on the X-axis represent the weeks. The values on the Y-axis represent GHI values predicted by the ARIMA model. The total values in the dataset were 781.



The mathematical form for d-order differencing is: $(1 - B)^d x_t$

for which B stands for the backshift operator. If white noise is produced by differencing by order d, then the series is integrated of order d. The integration by order d is the meaning of the extra I in ARIMA(p, d, q), so the function performs ARMA(p, q) on integrated data. A mathematical expression for this model is:

$$\theta_p(B)(1 - B)^d x_t = \phi_q(B) w_t$$

Seasonal Auto Regressive Integrated Moving Average (SARIMA) Model

- Seasonal Autoregressive Integrated Moving Average, SARIMA or Seasonal ARIMA, is an extension of ARIMA that explicitly supports univariate time series data with a seasonal component.
- It adds three new hyperparameters to specify the autoregression (AR), differencing (I) and moving average (MA) for the seasonal component of the series, as well as an additional parameter for the period of the seasonality.
- A seasonal ARIMA model is formed by including additional seasonal terms in the ARIMA.
- The seasonal part of the model consists of terms that are very similar to the non-seasonal components of the model, but they involve backshifts of the seasonal period.

It adds 3 seasonal hyperparameters (P,D,Q) and a hyperparameter for seasonality m to the three hyperparameters of ARIMA. Thus each SARIMA model is characterized by the hyperparameters: (p,d,q)(P,D,Q,m) where p,d and q have meanings similar to the ARIMA model and the other hyperparameters are used as follows:

• P: seasonal autoregressive order

D: seasonal difference order

• Q: seasonal moving average order

• m: number of time steps in seasonal data

Assumptions for SARIMA -

- Seasonal variation needs to be constant.
- The time series under consideration must be marginally stationary or can be merged to generate a stationary series apart from the seasonal component.
- The error terms are thought to be independent, identically distributed variables that were randomly picked from a normal distribution with a mean of 0.

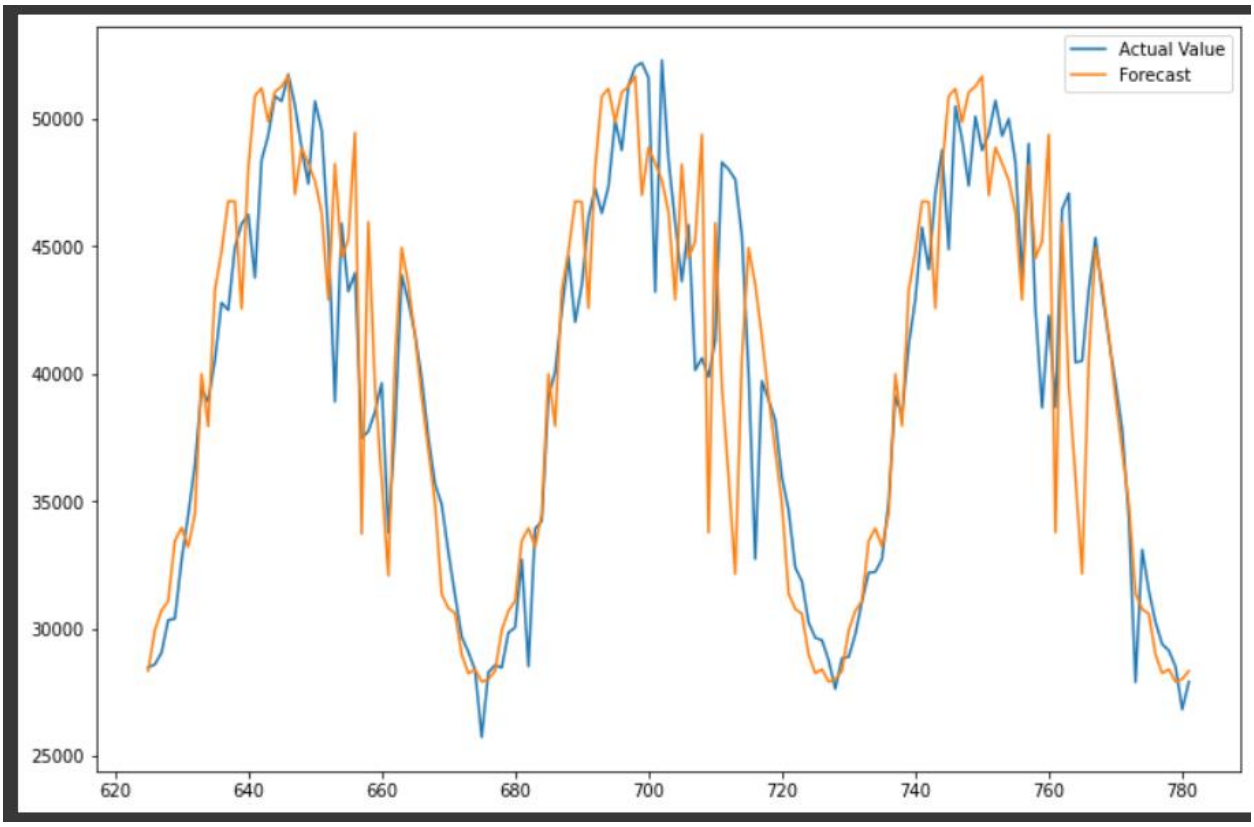


Fig: Rajasthan-1 SARIMA Model Weekly Forecasts

- Parameters - (1,0,1)(1,1,1,52)
- MAPE = 5.927 (weekly)
- MAE =2397.875 (weekly)

```
mape = mean_absolute_percentage_error(test , predictions)
mae = mean_absolute_error(test ,predictions)
print(mape)
print(mae)
```

```
0.059270533562983355
2397.875106934906
```

Comparing the MAPE values for both ARIMA and SARIMA models , we can say that ARIMA is preferred over SARIMA , which infers that there is negligible seasonality in our data.

Machine learning for Time Series Analysis

In recent years, as the availability of data and computer power has increased, Machine Learning has become an integral component of the new generation of Time Series Forecasting models, producing great results.
In traditional models such as AR, MA, etc., feature engineering is conducted manually and needs human parameter optimization. Machine Learning models require just characteristics and dynamics derived directly from the data. This allows them to expedite data preparation and learn complex data patterns more quickly and thoroughly.
In recent years, numerous new architectures have been created as diverse time series problems are investigated in a variety of domains. This has also been facilitated by the increasing availability of open-source frameworks, which has made the development of new bespoke network components easier and quicker.

Some of the well known methods in ML are:

- Recurrent Neural Networks (RNNs): They are the most classical and used architecture for Time Series Forecasting problems.
- Long Short-Term Memory (LSTM): They are an evolution of RNNs developed to overcome the vanishing gradient problem.
- Gated Recurrent Unit (GRU): They are another evolution of RNNs, like LSTM.
- Encoder-Decoder Model: This is a model for RNNs introduced to address the problems where input sequences differ in length from output sequences.
- Attention Mechanism: This is an evolution of the Encoder-Decoder Model, developed in order to avoid forgetting of the earlier parts of the sequence.

The decision is mostly based on whether the target variable has a strong correlation between its past, present, and future values. If strong correlations exist then time series maybe appropriate. Modeling and implementing time series analysis is simpler, but its empirical character is reliant on the assumption of a correlated goal variable. The disadvantage of time series is that prediction modeling does not account for root causes and influences.

This assignment involves univariate forecasting, and research indicates that classical time series models are more accurate for univariate, straightforward time series forecasting. In addition, classical time series methods are easily extensible, but machine learning models are more objectively focused. While machine learning may bring benefit in datasets with complex irregular data, missing observations, excessive noise, complex connection between various variates. But in our situation, for simple univariate forecasting we advise classical approaches because of better accuracy and lesser calculation cost.

Conclusion

- Using the Augmented Dicky Fuller test, we concluded that our data is stationary. So ARMA is best suited for forecasting.
- Although, ARIMA, SARIMA and ARMA work in similar manner for the stationary data, any methodology can be adopted for forecasting.
- We can also see that there is a high correlation between all the GHI,DHI and DNI factors and this is intra-state data for Rajasthan. So we can expect similar results from other solar parks of Rajasthan as well.
- Using the Kolmogorov Smirnoff test, we found that the most probable underlying probability distribution of the data is the beta distribution as it gives the minimum least square error among all the possible theoretical distributions.