Solar Power Prediction, Management and Predictive Maintenance

Six months Training Report

SUBMITTED IN PARTIAL FULFILLMENT REQUIREMENT FOR

THE AWARD OF DEGREE OF

Bachelor of Technology

(Computer Science & Engineering)

SUBMITTED BY

Anshul Aggarwal and Yogesh Mani

(UNIVERSITY ROLL No. 1800192)

December, 2021

SABUDH

D-228, Third Floor, Sector 74,

Sahibzada Ajit Singh Nagar,

Punjab 140307

CANDIDATE'S DECLARATION

I hereby certify that I have under gone six months industrial training at SABUDH FOUNDATION and worked on project entitled, "**Solar Power Prediction, Management and Predictive Maintenance**", in partial fulfillment of requirements for the award of Degree of **Bachelor of Technology** in Computer Science at **Amritsar Group of Colleges**, having University Roll No.1800192, is an authentic record of my own work carried out during a period from July, 2021 to December, 2021 under the supervision of Er. Vijay Garg and Er. Lakshmi Narayanan.

(**Anshul Aggarwal and Yogesh Mani)**

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

(**Er. Vijay Garg, Er. Lakshmi Narayanan)**

Person from Organisation with designation

ABSTRACT

The project aims at Solar Power prediction, management and predictive maintenance. There is an increasing adoption of renewable energy sources globally. In particular, in India, solar energy plays an increasingly important part of the country's growth and sustainability. The project will perform data mining and knowledge discovery of power quality events associated with the solar grids and estimate the solar PV output loss (in kWh). The production of power from Solar farms depends on the weather and the state of maintenance of the Solar panels. Consumption and Production often don't match each other, hence strategies for pushing production into the grid or storage for later use need to be managed. Further, predictive maintenance of panels can help ensure full efficiency to be maintained. Sustainability being the key interest of all stakeholders in the system our partner Institution Akal Academy, Academic Institution in Punjab which remains operational in twelve sites run on fully functional hybrid solar power plant integrated to the grid. Here, we handle with Time-Series data to optimize the power consumption and build alert systems for power backups in the time of maintenance.

ACKNOWLEDGEMENT

I am highly grateful to **Er. Vijay Garg, Data Scientist**, Sabudh Foundation, Mohali, for providing opportunity to carry out six months training at Sabudh Founadtion from July-December 2021.

(Er. Vijay Garg) has provided great help in carrying out my work and is acknowledged with reverential thanks. Without the wise counsel and able guidance, it would have been impossible to complete the training in this manner.

I would like to express thanks profusely to thank **Er. Vijay Garg**, for stimulating me time to time. I would also like to thank entire team of Sabudh Foundation. I would also thank my friends who devoted their valuable time and helped me in all possible ways towards successful completion of this training.

(Anshul Aggarwal and Yogesh Mani)

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Full Form
RE	Renewable Energy
DelREMO	Delta Remort Monitoring Solution
API	Application programming interface
CSV	Comma Seperated Value
IQR	Interquartile Range
GRU	Gated Recurrent Unit
RNN	Interquartile Range
LSTM	Long Short - Term Memory
CEC	Constant Error Carrousels
ROI	Return on Investment
AR	Auto Regression
MA	Moving Average
AIC	Akaike Information Criterion
ACF	Auto Correlation Function
ADF	Augmented Dickey-Fuller

TABLE OF CONTENTS

Chapter 1

Introduction to Organisation

Sabudh Foundation - An NGO that applies data science for social good. Sabudh Foundation is formed by the leading data scientists in the industry with the objective to bring together data and young data scientists to work on focused, collaborative projects for social benefit. Sabudh foundation is working on solving the real and high impact problems in areas such as education, governance, healthcare, and agriculture using Artificial Intelligence and Machine Learning techniques.

Data science can be used across a number of industries in order to be beneficial for the society. For example in agriculture, there are now Agrobots and drones being used to gauge the health of the harvest that can help farmers improve their crop yield and reduce costs. With the help of advanced technologies, we're able to save 90
The foundation has taken steps to involve Colleges, Universities, and Industry from the region for the social cause. Particularly, the foundation has signed academic and research-based MoUs with Panjab University, Chandigarh, GNDEC, Ludhiana, BML Munjal University, Punjab Government (Punjab Police), Punjabi University, Patiala, and Punjab Engineering College, Chandigarh.

Chapter 2

Introduction to Project

2.1 Overview

There is an increasing adoption of renewable energy sources globally. In particular, in India, solar energy plays an increasingly important part of the country's growth and sustainability.

Akal Academy is a group of academic institutions in Punjab India, that helps students mostly from the deprived sections of society gain access to quality education. There are twelve (12) sites in remote areas of Punjab. The sites were electrified through hybrid solar plants integrated with the grid.

The aim of this study was twofold, (1) to perform a data mining and knowledge discovery of power quality events associated with the grid that takes place at Akal Academy Balbhera site, and (2) to estimate the solar PV output (in kWh) at the Balbhera Site.

The scope of the study included daily time-series data between February 2019 and May 2021 for four the Akal Academy hybrid solar plant sites in Punjab, namely, Balbhera. The study also includes the events data associated with the plant which includes, Grid Events, Maintenance Events and Weather Events

2.2 Solar Energy as Renewable resource

Most of our electricity comes from coal, nuclear, and other non-renewable power plants. Producing energy from these resources takes a severe toll on our environment, polluting our air, land, and water. Renewable energy sources can be used to produce electricity

with fewer environmental impacts. It is possible to make electricity from renewable energy sources without producing carbon dioxide (CO2), the leading cause of global climate change. Thus, Renewable energy like energy from wind /sun is getting traction all over the globe

Photovoltaics (PV), also called solar cells, are electronic devices that convert sunlight directly into electricity and is the fastest-growing renewable energy technologies. Solar PV installations can be combined to provide electricity on a commercial scale, or arranged in smaller configurations for mini-grids or personal use. Prediction of these energies around the year across selected locations helps to make successful energy investment plans. This also helps in operating and managing power systems more efficiently.

2.3 Stages and Scope to Solar Power Generation

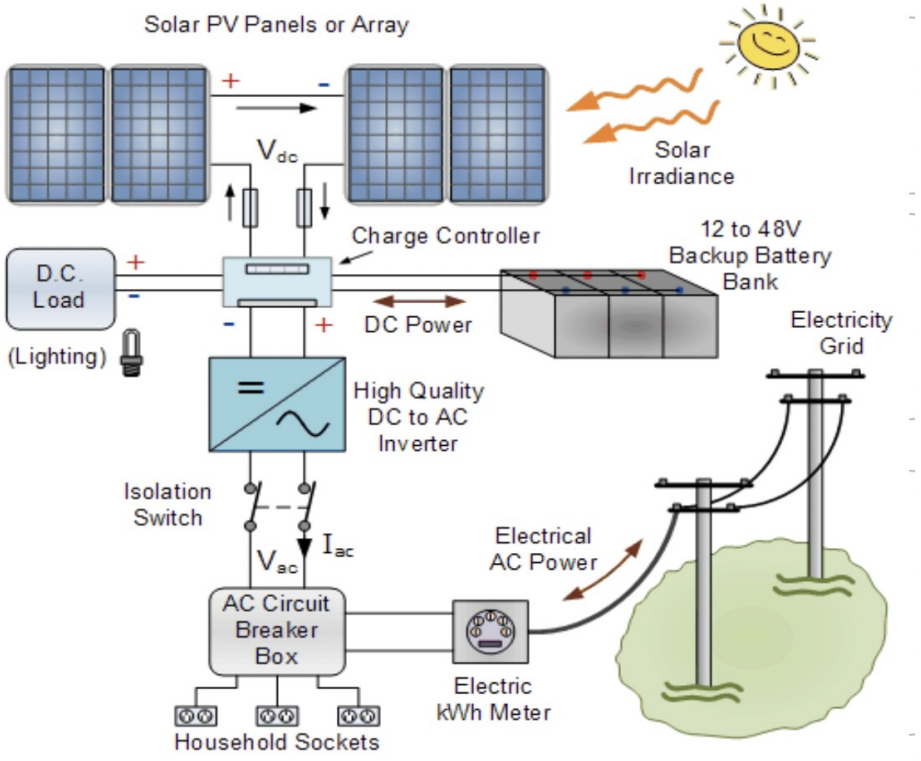

FIGURE 2.1: Solar Connected System

- Stage 1 : Here, the power is being generated.

Scope:

- Predicting amount of solar irradiance
- And how much solar power is / can be produced

- Stage 2: Next stage is transformation of power.

  Scope:

  - Monitoring output and working towards optimization

- Stage 3: Next, the power is distributed accordingly

  Scope:

  - Monitoring how much power is consumed daily and needed
  - And how much is exported to the grid

Chapter 3

Literature Review

Can renewable energy power the future?: Fossil fuels face resource depletion, supply security, and climate change problems; renewable energy (RE) may offer the best prospects for their long-term replacement.The most important RE sources, wind and solar energy, are intermittent, which will necessitate major energy storage if these sources are to dominate total energy supply in future. The need to maintain ecosystem services will reduce global RE potential. [1].

Adoption of residential solar power under uncertainty, Implications for renewable energy incentives: Many incentives at the state and federal level exist for household adoption of renewable energy like solar photovoltaic (PV) panels. Despite generous financial incentives the adoption rate is low. One of solar energy's primary benefits is its low carbon emissions. Greenhouse gas emissions from electricity produced using a solar photovoltaic (PV) system is over 90 percent lower compared to electricity from substitutes such as oil, coal and natural gas, making it an ideal renewable energy source (Weisser 2007). In addition, the adoption of solar power contributes to the local economy by providing green jobs. Uncertainty in benefits and costs leads to delay in investment timing. Rebates and other financial incentives decrease adoption time, but their effect is attenuated if households apply the option value decision rule to solar PV investments. [2].

Time Series Forecasting on Solar Irradiation using Deep Learning: Time series forecasting is currently used in various areas. Energy management is also one of the most prevalent application areas. Thus in case of solar ower prediction, prediction of the

solar irradiance plays an important part. Solar Irradiation is a type of property which is used to measure solar radiation. There are different kinds of measurement: Total Solar Irradiance, Spectral Solar Irradiance, Global Horizontal Irradiance, etc. Measurements are done primarily through satellite or surface sensors. Generated energy is measured from Solar power generator in days, weeks, or even months. This is useful in managing power distribution from different kinds of sources. [3].

Assessing Evidence for Weather Regimes Governing Solar Power Generation in Kuwait: With electricity representing around 20 percent of the global energy demand, and increasing support for renewable sources of electricity, there is also an escalating need to improve solar forecasts to support power management. While considerable research has been directed to statistical methods to improve solar power forecasting, few have employed finite mixture distributions. A statistically-objective classification of the overall sky condition may lead to improved forecasts. Identify prevalent sky conditions (clear, semi- cloudy, and cloudy) and explore associated weather pattern. [4].

Prediction of solar irradiation and performance evaluation of grid connected solar 80KWp PV plant in Bangladesh: Energy security is the triggering factor for a developing country. Thus, for ensuring energy security renewable energy such as PV plant could be a best alternative. This study deals with the performance analysis of 80 KWp grid-connected solar power plant in Dhaka. This study presents a solar irradiation predict model based on fuzzy logic and artificial neural networks which aims to achieve a good accuracy at different weather conditions. The accuracy of predicted solar irradiation will affect the power output forecast of grid-connected photovoltaic systems which is important for power system operation and planning. Real time power generation should be investigated precisely for grid performance, because a high penetration of PV production could create instability in the grid . Also, the uncertainty of the photovoltaic performance models needs to be reviewed. [5]

Chapter 4

Exploratory Data Analysis

4.1 Dataset

4.1.1 Collection of Data

Akal Academy hybrid solar plant sites data are collected in real-time and reported via a web-based platform, DelREMO.

The data sets stored in DelREMO can be broadly classified into two categories, (1) critical alarms data, and (2) plant output data. The critical alarms data represent a different type of events or alarms a site may encounter daily. The plant output data represent the amount of solar PV energy generated (in kWh) daily.

For this study, the data for the Balbhera site was sourced through DelREMO. In particular, the daily critical alarms data and daily plant output data for each month between February 2019 and May 2021 were collected from DelREMO. For the similar time period, the weather data for whole of Patiala was also collected with the help of an online weather data API available at visual crossing weather site.

4.1.2 Size of datasets

- The critical alarms data set for the Balbhera site had 2649 observations with five (5) variables in total.
- The plant output data set for four the Balbhera site had 910 observations with two (2) variables in total.
- The weather data had 912 obvservations with (17) variables

4.1.3 Description of data

The data attributes collected for critical alarms data were:

- Date - start date and time of event in DD/MM/YYYY HH:MM:SS format
- Device - categorical variable, with Plant or inverter device ID e.g., 1
- Details - categorical variable, with description of events
- Deactive date - end date and time of event in DD/MM/YYYY HH:MM:SS format
- Duration - total duration of event in D.HH:MM:SS format

The data attributes collected for plant output data were:

- Date - date in DD/MM/YYYY format
- Total yield [kWh] - numeric value of float data type, eg, 9.9

The data attributes collected for weather data were:

- Name - Categorical variable with the name of city i.e., Patiala
- Date time - Date of the weather in DD/MM/YYYY format
- Maximum temperature - numeric value of float data type, with maximum temperature reached on a particular day
- Minimum temperature - numeric value of float data type, with minimum temperature reached on a particular day
- Temperature - numeric value of float data type, with the average temperature that remained on a particular day

- Wind Chill - numeric value of float data type, with how cold the wind was on the given day
- Heat Index - numeric value of float data type, with the heat index on the given day
- Precipitation - numeric value of float data type, with the amount of rain fell on the given day
- Snow - empty column
- Snow depth - empty column
- Wind Speed - numeric value of float data type, with the speed at which the wind was flowing at the given day
- Wind direction - numeric value of float data type, with the direction in which the wind was flowing
- Wind Gust - empty column
- Visibility - numeric value of float data type, with the amount of visiblity
- Cloud Cover - numeric value of float data type, with the amount of cloud coverage on the given day
- Relative Humidity - numeric value of float data type
- Conditions - Categorical variable with the weather conditions, whether the weather was clear, rainy or partially cloudy

4.1.4 Cleaning of data

The data cleaning and transformation tasks were systematically performed through the Python programming language. The tasks were decomposed and grouped into logical modules through custom-built functions in Python.

The critical alarms data cleaning tasks involve the following:

- Input: Critical Alarm Dataset files in Excel (XLSX) format.
- Output(s): processed_alarm_Dataset.csv, pivoted_alarm_Dataset.csv
  - Merging all the files containing yearly data into one dataframe
  - Categorized the grid and non-grid events in a seperate column
  - Determined the period in which the event took place
  - Calculated the hours for which the event was active
  - Pivoted the event data on the basis of the details of the event to get the Grid Duration Hours, Maintenance Duration hours and the hours for which DelREMO was not communicating, as seperate columns
  - Finally, saved the datasets as CSV files

The weather data cleaning tasks involve the following:

- Input: Weather Dataset in CSV format.
- Output(s): weather_data.csv
  - Dropping the unnecessary columns

The plant output data cleaning tasks involve the following:

- Input: Plant Output Dataset files in Excel (XLSX) format.
- Output(s): Combined_Solar_Dataset.csv
  - Transforming the plant output data sets from Excel (XLSX) format to CSV (commaseparated value) data sets respectively

4.1.5 Formation of final dataset

- Input(s): Plant Output Data, Critical Alarm Data and Weather Dataset in CSV format.
- Output(s): final_dataset.csv
  - Merging all the input files to form a final dataset containing all the necessary columns and excluding all erroneus data.

The data contains (7) independent variables and (1) dependent variable which are used for the analysis and model selection as follows,

- Total Yield [kWh] - Output generated from the plant **(Dependent Column)**
- GridDurationHours - represents the number of hours the grid events took place
- MaintDurationHours - represents the number of hours the Maintenance events took place
- DelREMO - represents the number of hours for which DelREMO site stopped communicating
- Temp - Temperature on a particular day
- Prcp - represents the amount of precipitation recorded per day
- Cloud Cover - represents the coverage of the clouds per day
- Conditions - Categorical variable captures the weather conditions, whether the weather was clear, rainy or partially cloudy

4.2 Exploratory Data Analysis and Visualisations

	Date	Total Yield [kWh]	GridDurationHours	MaintDurationHours	DelREMO	Temp	Prcp	Cloud Cover	Conditions
0	2019-02-01	190.3	0.000000	0.0	0.000000	14.3	1.75	67.1	Rain, Partially cloudy
1	2019-02-02	246.5	0.004167	0.0	0.000000	12.4	0.00	12.0	Clear
2	2019-02-03	227.3	0.000000	0.0	0.000000	12.3	0.00	0.0	Clear
3	2019-02-04	205.9	0.001944	0.0	0.000000	12.1	0.00	40.7	Partially cloudy
4	2019-02-05	191.5	0.003889	0.0	4.345833	16.2	0.00	74.2	Partially cloudy

FIGURE 4.1: Final Dataset

Out of all the the datasets data with continuous time series was chosen, i.e., 1st February 2019 to 31st May, 2021

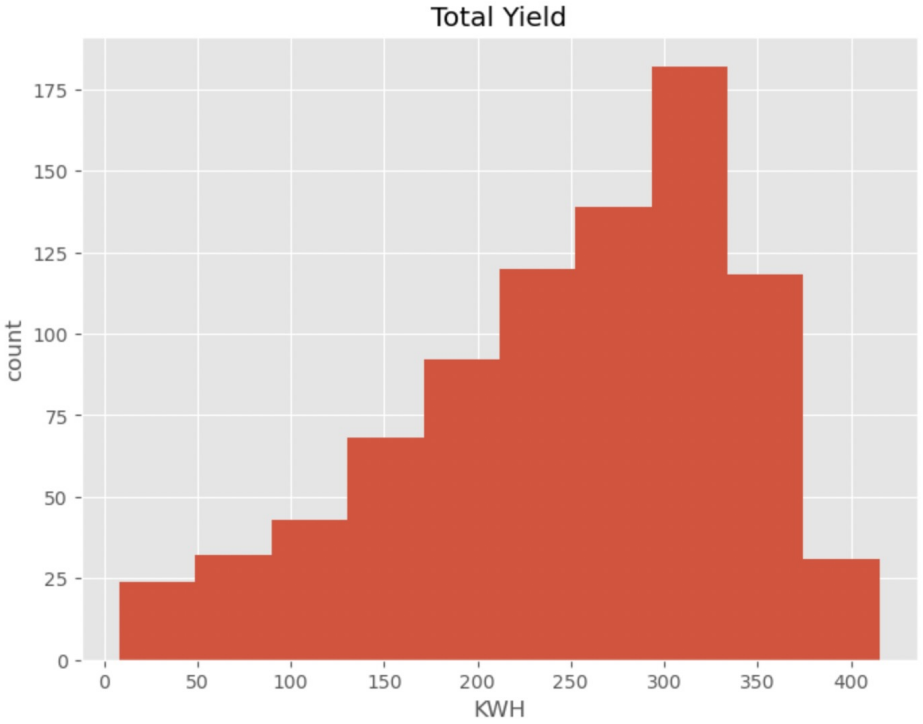

FIGURE 4.2: Total Yield Count

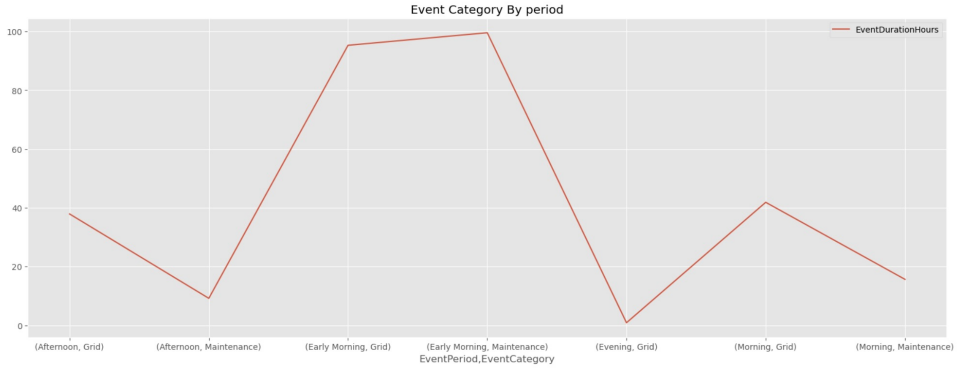

FIGURE 4.3: Event Category by Period

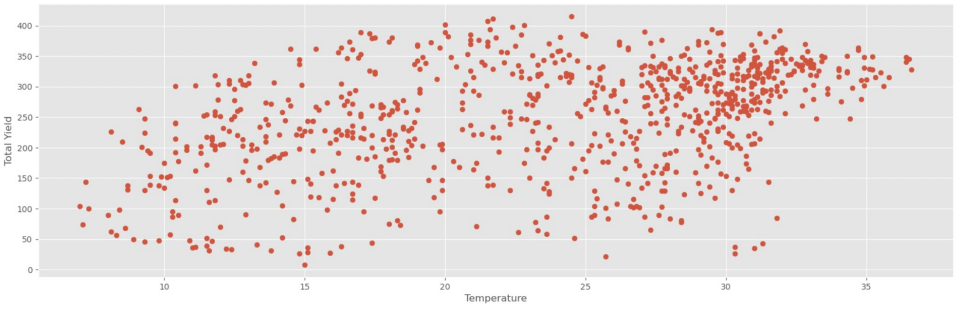

FIGURE 4.4: Temperature vs Yield

FIGURE 4.5: Yield when either events happened and none happened

- The minimum and maximum yield throughout ranges between 0 to 415.3 kWh
- And, mostly the yield ranged between 300-350 throughout the given time series
- The minimum and maximum temperature throughout for the site ranges between 7 degrees to 36.6 degrees

- With increase in temperature, Yield also increases
- The weather remains mostly clear during the year
- The grid and maintenance events mostly happened in early mornings
- More the event hours, less was the yield recorded
- The Precipitation column indicates towards presence of a lot of outliers due to the large difference between max value and upper quartile value

4.2.1 Verification of Skewness and Kurtosis

Skewness: Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.

Kurtosis: Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. Data sets with low kurtosis tend to have light tails, or lack of outliers.

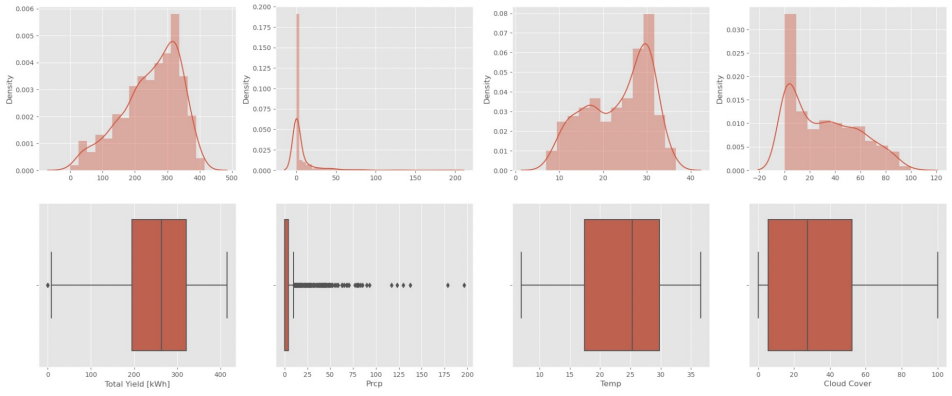

FIGURE 4.6: Distribution and Box Plots

```
Range of Skewness, S < |0.5|


Skewness of Total Yield:  -0.6236119855557846
Skewness of Prcp:  4.646326223976378
Skewness of Temp:  -0.4163484195121324
Skewness of Cloud Cover:  0.4835140983110726


Kurtosis of Total Yield:  -0.3282618349013515
Kurtosis of Prcp:  29.347855781171127
Kurtosis of Temp:  -1.0284953539592268
Kurtosis of Cloud Cover:  -0.9347320157450669
```

FIGURE 4.7: Values of Skewness and Kurtosis

- The presence of outliers in prcp column is verified
- Prcp column is positively skewed and Total yield column is partially negatively skewed
- Prcp column has a high value of Kurtosis which indicates towards presence of reasonable number of outliers and needs to be considered
- IQR test and Log transformation was performed on the prcp column to remove the outliers and handle skewness

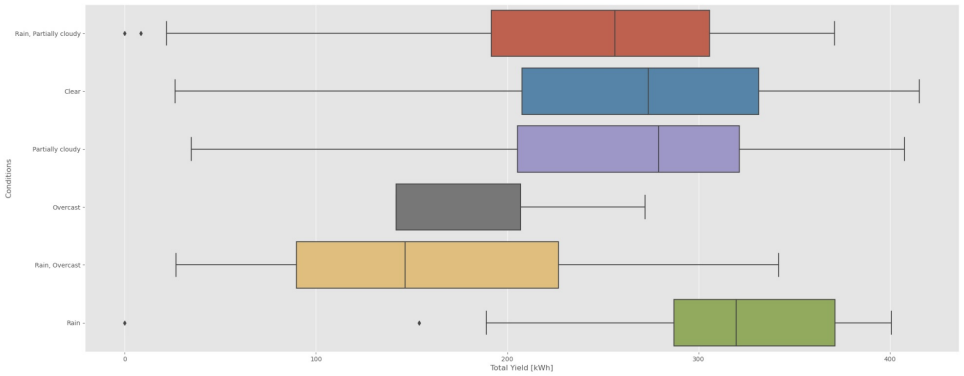

FIGURE 4.8: Conditions Boxplot

- The median yield is maximum when the condition is either Rain or Clear

- The yield is highest when condition is clear
- With the conditions being 'Rain,Overcast' or Overcast, yield comes out be low

4.2.2 Correlation

The performance of some algorithms can deteriorate if two or more variables are tightly related, called multicollinearity. A correlation could be positive, meaning both variables move in the same direction, or negative, meaning that when one variable's value increases, the other variables' values decrease. Correlation can also be neutral or zero, meaning that the variables are unrelated. The performance of some algorithms can deteriorate if two or more variables are tightly related, called multicollinearity.

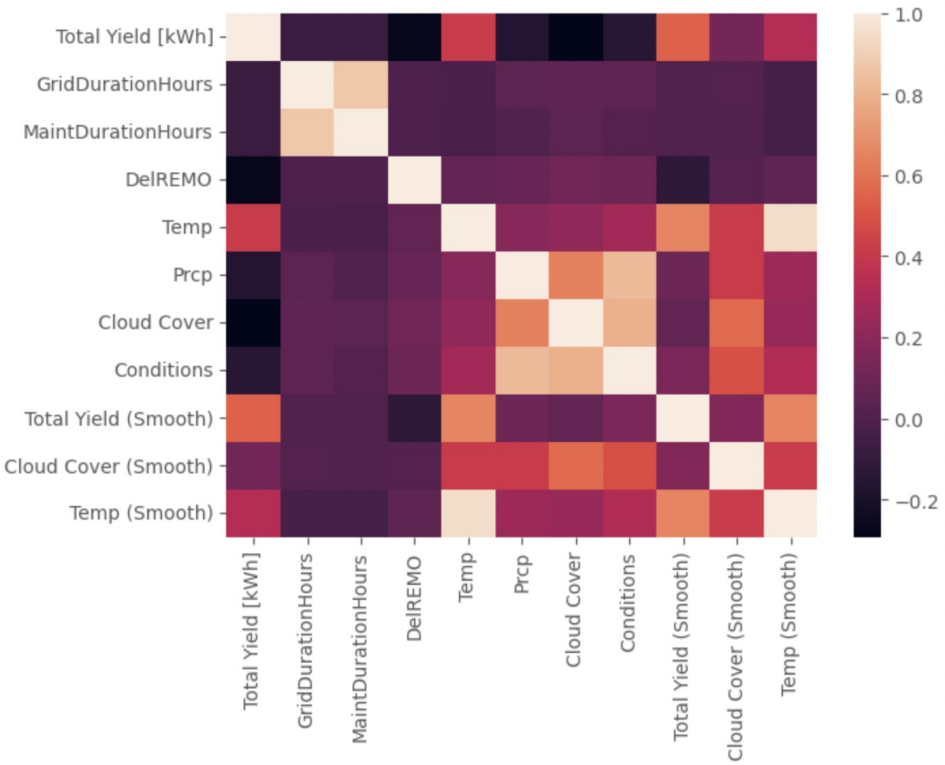

FIGURE 4.9: Correlation Plot

The correlation plot shows,

- Prcp, Cloud Cover and Conditions are highly correlated to each other

- Temperature and Cloud Cover are correlated to Total Yield

4.2.3 Noise removal (Smoothing of data columns)

In time series forecasting, the presence of dirty and messy data can hurt the final predictions. This is true, especially in this domain, because the temporal dependency plays a crucial role when dealing with temporal sequences.

The scope of representing time series models in the state-space form is the availability of a set of general algorithms, for the computation of the Gaussian likelihood, which can be numerically maximized to obtain the maximum likelihood estimation of the model's parameters.

The resulting smoothed time series holds the same temporal pattern present in the raw data but with a consistent and rational noise reduction.

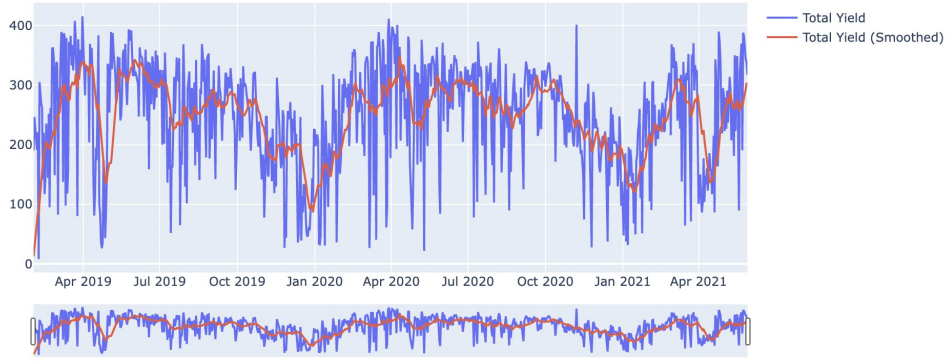

FIGURE 4.10: Total Yield Plot

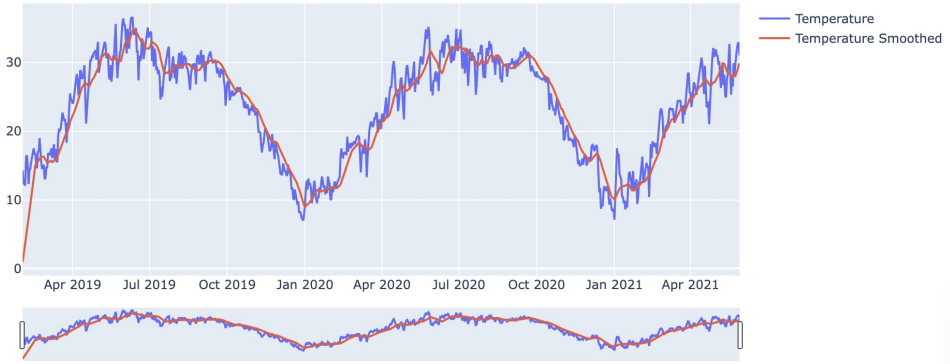

FIGURE 4.11: Temperature Plot

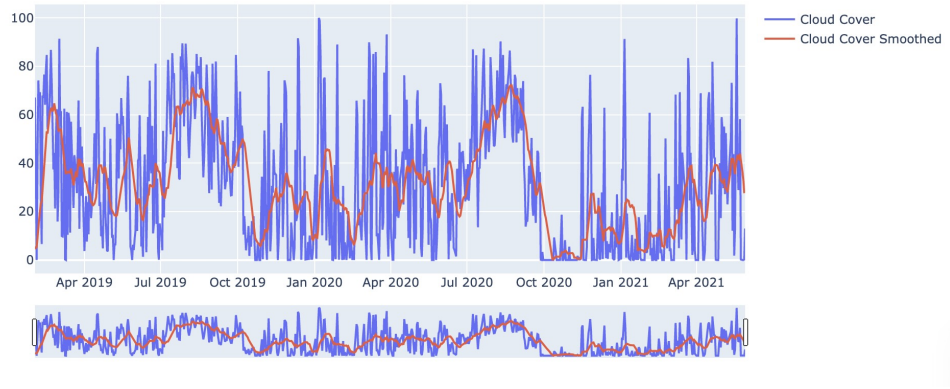

FIGURE 4.12: Cloud Cover Plot

Chapter 5

Methodology

This section discusses the methodology to develop the ML model.

5.1 Introduction to Languages (Front End and Back End)

5.1.1 PYTHON

Python is a computer programming language often used to build websites and software, automate tasks, and conduct data analysis. Python is a general purpose language, meaning it can be used to create a variety of different programs and isn't specialized for any specific problems. This versatility, along with its beginner-friendliness, has made it one of the most-used programming languages today. It is commonly used for developing websites and software, task automation, data analysis, and data visualization.

Python in Data Analysis and Machine Learning. Python has become a staple in data science, allowing data analysts and other professionals to use the language to conduct complex statistical calculations, create data visualizations, build machine learning algorithms, manipulate and analyze data, and complete other data-related tasks.

Python can build a wide range of different data visualizations, like line and bar graphs, pie charts, histograms, and 3D plots. Python also has a number of libraries that enable coders to write programs for data analysis and machine learning more quickly and efficiently, like TensorFlow and Keras.

5.1.2 HTML

The HyperText Markup Language, or HTML is the standard markup language for documents designed to be displayed in a web browser. Web browsers receive HTML documents from a web server or from local storage and render the documents into multimedia web pages. HTML de- scribes the structure of a web page semantically and originally included cues for the appearance of the document.

5.1.3 CSS

Cascading Style Sheets (CSS) is a style sheet language used for describing the presentation of a document written in a markup language such as HTML. CSS is designed to enable the separation of presentation and content, including layout, colors, and fonts. CSS describes how HTML elements are to be displayed on screen, paper, or in other media.

5.1.4 JAVASCRIPT

JavaScript is a text-based programming language used both on the client-side and server-side that allows you to make web pages interactive. Where HTML and CSS are languages that give structure and style to web pages, JavaScript gives web pages interactive elements that engage a user. Incorporating JavaScript improves the user experience of the web page by converting it from a static page into an interactive one.

5.2 Any other Supporting Languages/ packages

5.2.1 TENSORFLOW

TensorFlow is an end-to-end open source platform for machine learning. It has a comprehensive, flexible ecosystem of tools, libraries and community resources that lets researchers push the state-of-the-art in ML and developers easily build and deploy ML powered applications.

5.2.2 MATPLOTLIB

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible. It is used for forming quality plots and for customizing visual plots and layouts.

5.2.3 PANDAS

It is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language. It is a Python package that offers various data structures and operations for manipulating numerical data and time series. It is mainly popular for importing and analyzing data much easier. Pandas is fast and it has high-performance and productivity for users.

5.2.4 NUMPY

NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more. At the core of the NumPy package, is the ndarray object. This encapsulates n-dimensional arrays of homogeneous data types, with many operations being performed in compiled code for performance.

5.2.5 FLASK

Flask is a Python web framework built with a small core and easy-to-extend philosophy.Flask is an API of Python that allows us to build up web-applications. It was developed by Armin Ronacher. A Web-Application Framework or Web Framework is the

collection of modules and libraries that helps the developer to write applications without writing the low-level codes such as protocols, thread management, etc. Flask is based on WSGI(Web Server Gateway Interface) toolkit and Jinja2 template engine.

5.3 ML algorithm discussion

5.3.1 SARIMA Model

If a time series, has seasonal patterns, then you need to add seasonal terms and it becomes SARIMA, short for 'Seasonal ARIMA'. In many time series data, frequent seasonal effects come into play. Take for example the average temperature measured in a location with four seasons. There will be a seasonal effect on a yearly basis, and the temperature in this particular season will definitely have a strong correlation with the temperature measured last year in the same season.

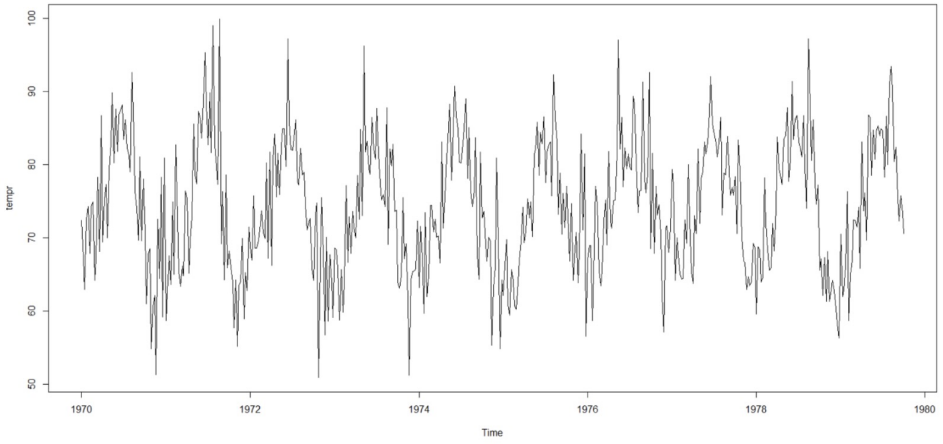

FIGURE 5.1: Example: Sarima Model

The plot above shows the yearly cyclical rise and fall in temperatures, and there is a strong basis for assuming temperatures will fall to 60 deg F near the end of the year, while temperatures will surge past 80 deg F near the middle of the year.

Configuring a SARIMA requires selecting hyperparameters for both the trend and seasonal elements of the series. There are three trend elements that require configuration. They are the same as the ARIMA model; specifically:

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

There are four seasonal elements that are not part of ARIMA that must be configured; specifically:

- P: Seasonal autoregressive order.
- D: Seasonal difference order.
- Q: Seasonal moving average order.
- m: The number of time steps for a single seasonal period.

5.3.2 GRU

GRU is a RNN (Recurrent Neural Network) based model. RNN are a powerful and robust type of neural network. because of their internal memory, RNN's can remember important things about the input they received, which allows them to be very precise in predicting what's coming next. This is why they're the preferred algorithm for sequential data like time series, speech, text, financial data, audio, video, weather and much more. Recurrent neural networks can form a much deeper understanding of a sequence and its context compared to other algorithms.

RNN contain cycles that feed the network activations from a previous time step as inputs to the network to influence predictions at the current time step. These activations are

stored in the internal states of the network which can in principle hold long-term temporal contextual information. This mechanism allows RNNs to exploit a dynamically changing contextual window over the input sequence history

5.3.3 LSTM

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems. standard RNNs fail to learn in the presence of time lags greater than 5 – 10 discrete time steps between relevant input events and target signals. The vanishing error problem casts doubt on whether standard RNNs can indeed exhibit significant practical advantages over time window-based feedforward networks. LSTM can learn to bridge minimal time lags in excess of 1000 discrete time steps by enforcing constant error flow through "constant error carrousels" (CECs) within special units, called cells.

All recurrent neural networks have the form of a chain of repeating modules of neural network. In standard RNNs, this repeating module will have a very simple structure, such as a single tanh layer.

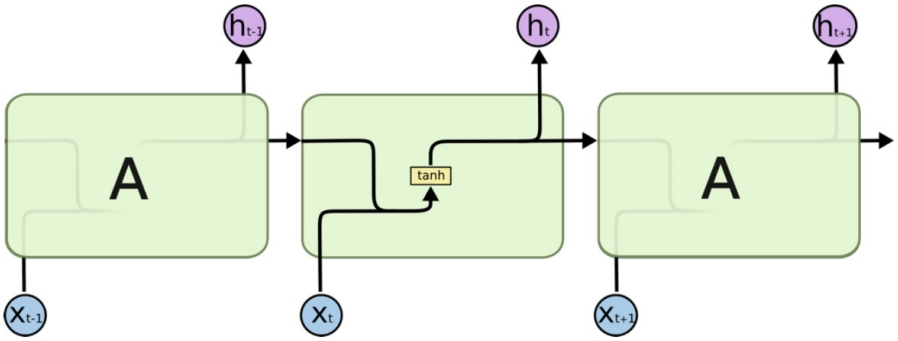

FIGURE 5.2: Image 1: RNN

LSTMs also have this chain like structure, but the repeating module has a different structure. Instead of having a single neural network layer, there are four, interacting in a very special way.

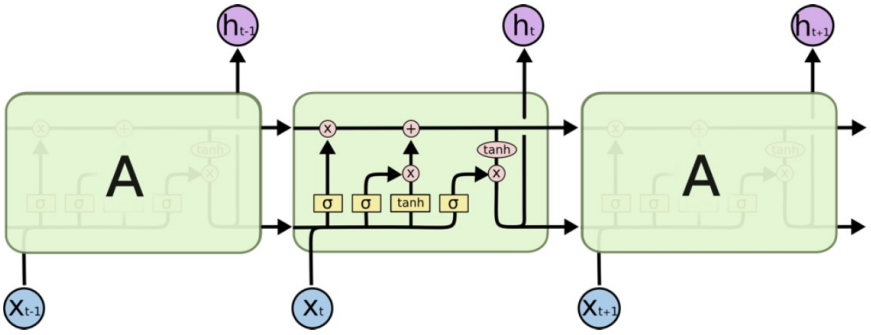

FIGURE 5.3: Image 2: LSTM

5.4 Implementation of Algorithm

5.4.1 Checking Stationarity

Stationarity means that the statistical properties of a process generating a time series do not change over time. stationarity's importance is its ubiquity in time series analysis, making the ability to understand, detect and model it necessary for the application of many prominent tools and procedures in time series analysis.

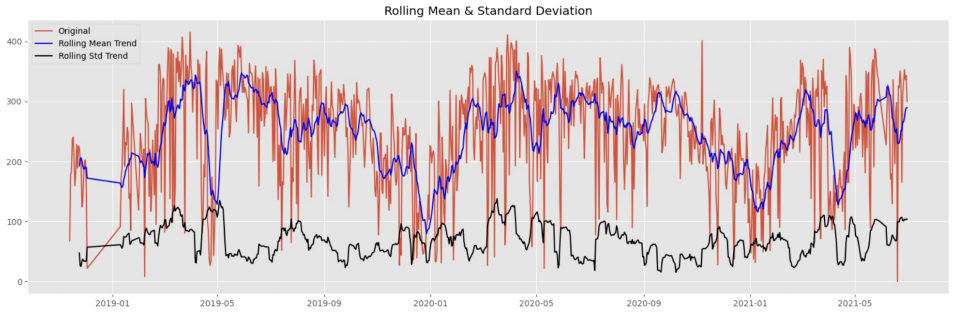

FIGURE 5.4: Rolling Mean and Standard Deviation

```
Test Statistic                  -6.097004e+00
p-value                          1.004455e-07
#Lags Used                       6.000000e+00
Number of Observations Used      9.030000e+02
Critical Value (1%)             -3.437612e+00
Critical Value (5%)             -2.864746e+00
Critical Value (10%)            -2.568477e+00
dtype: float64
```

FIGURE 5.5: Statistics for Stationarity

Here, the p-value is less than 0.05, so our data is taken to be stationary

5.4.2 Decomposition plot

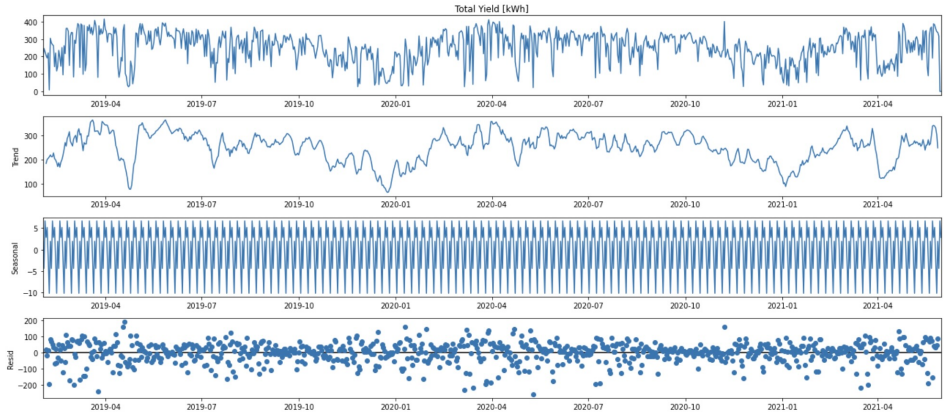

FIGURE 5.6: Decomposition Plot

This plot helps to visualise the trend and seasonality present in the dataset.

- **Trend:** a general systematic linear or (most often) nonlinear component that changes over time and does not repeat
- **Seasonality:** the repeating short-term cycle in the series
- **Residual / Noise:** a non-systematic component that is nor Trend/Seasonality within the data

5.4.3 Univariate analysis using SARIMA model

Three stages in SARIMA Forecasts:

1. Identification: Check for stationarity and find the combinations of AR and MA order;
2. Estimation: the parameters and AIC are computed for each models;
3. Validation: Select the model wherein the residuals show presence of white noise.

5.4.3.1 Identification

Stationary series have constant mean and variance, and an auto-correlation that only depends on the lag between two periods. We check the series is stationary by looking at the ACF plot, and performing unit root tests like ADF test.

```
Results of Dickey-Fuller Test for column: Total Yield [kWh]
Test Statistic                  -4.388458
p-value                          0.000311
No Lags Used                    11.000000
Number of Observations Used    839.000000
Critical Value (1%)             -3.438168
Critical Value (5%)             -2.864991
Critical Value (10%)            -2.568608
dtype: float64
Conclusion:====>
Reject the null hypothesis
Data is stationary
```

FIGURE 5.7: ADF Test Results

The ADF test (Said and Dickey, 1984) evolved from the Dickey Fuller Test (DF), in which the null-hypothesis is that $\alpha < 1$ for the model $x_t = \alpha x_{t-1} + u_t$ in which u_t is white noise. The ADF allows the differenced series u_t to be any stationary process rather than only white noise. If $\alpha > 1$ the series explodes and does not return to the mean

(i.e. non-stationary). The test resulted in a Dickey-Fuller value of -4.388458 with a p-value equal to 0.000311. The p-value is statistically significant, therefore we reject the null-hypothesis.

5.4.3.2 Estimation

In this stage we account, the parameters $\varphi, \phi, \vartheta, \Theta$ for each model are estimated by maximizing the likely-hood function i.e the probability of obtaining the data given the model. Model is selected based on the Akaike Information Criteria (AIC) sincee the AIC penalizes models with too many parameters:

$$\text{AIC} = 2k - 2\ln(P)$$

In which k equals the log-likelihood and P the number of parameters. The lower the value of AIC the better the fit of the model to the series.

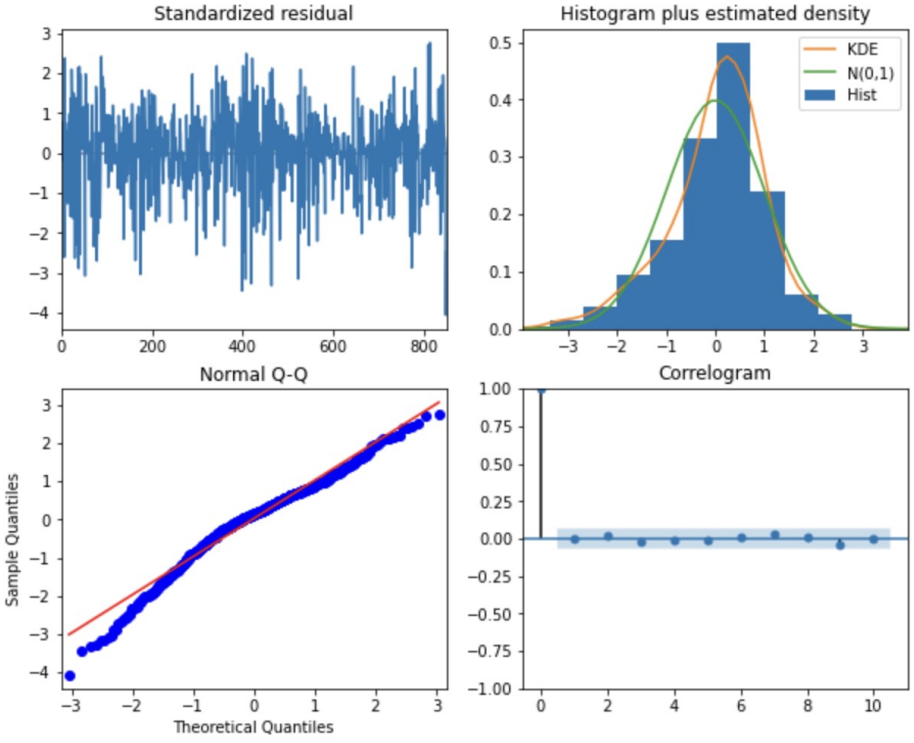

FIGURE 5.8: SARIMA Model (Diagnostics)

5.4.3.3 Validation (SARIMA(1,1,2)(0,1,1))

The model has only one significant value of auto-correlation. The residuals resemble a normal distribution. The values are all bellow the significance level and thus the residuals of this model fit are purely white noise. This makes it a good candidate for forecasting.

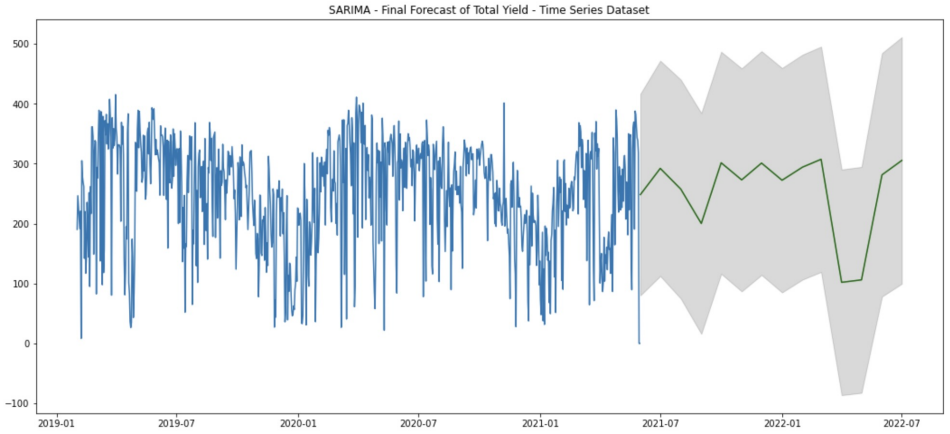

FIGURE 5.9: SARIMA Model (Final Forecasting)

Test RMSE (SARIMA): 43.2

5.4.4 Multivariate analysis using LSTM model

	var1(t-1)	var2(t-1)	var3(t-1)	var4(t-1)	var5(t-1)	var6(t-1)	var7(t-1)	var8(t-1)	var1(t)
1	0.458223	0.0	0.0	0.000000	0.246622	0.008917	0.671	1.0	0.593547
2	0.593547	0.0	0.0	0.000000	0.182432	0.000000	0.120	0.0	0.547315
3	0.547315	0.0	0.0	0.000000	0.179054	0.000000	0.000	0.0	0.495786
4	0.495786	0.0	0.0	0.000000	0.172297	0.000000	0.407	0.4	0.461112
5	0.461112	0.0	0.0	0.183235	0.310811	0.000000	0.742	0.4	0.531423
...	...	...	...	...	...	...	...	...	...
846	0.868312	0.0	0.0	0.000000	0.847973	0.000000	0.000	0.0	0.826126
847	0.826126	0.0	0.0	0.000000	0.871622	0.000000	0.000	0.0	0.814857
848	0.814857	0.0	0.0	0.021483	0.875000	0.000000	0.002	0.0	0.764243
849	0.764243	0.0	0.0	0.000000	0.810811	0.065525	0.130	0.6	0.000000
850	0.000000	0.0	1.0	0.014322	0.739865	0.050189	0.026	0.6	0.000000

850 rows × 9 columns

FIGURE 5.10: Multivariate Model (dataset)

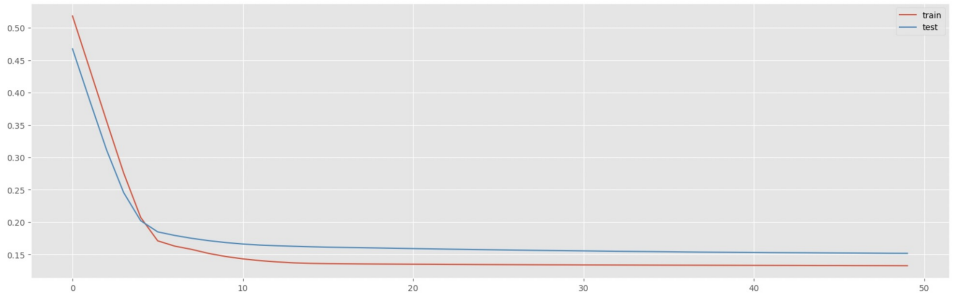

FIGURE 5.11: LSTM Model (Loss Plot)

Test RMSE (LSTM): 10.885

5.4.5 Multivariate analysis using GRU model

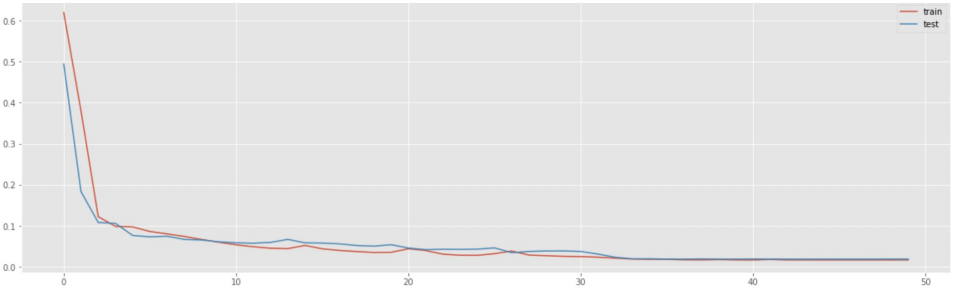

FIGURE 5.12: GRU Model (Loss Plot)

Test RMSE (GRU): 8.876

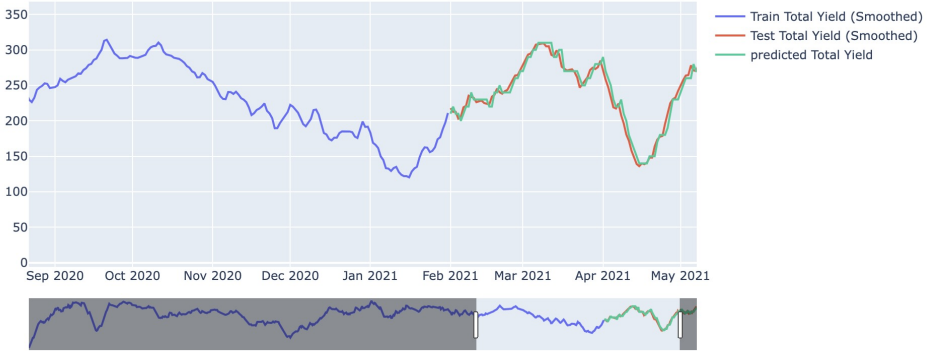

FIGURE 5.13: GRU Model (Prediction Plot)

5.5 Deployment

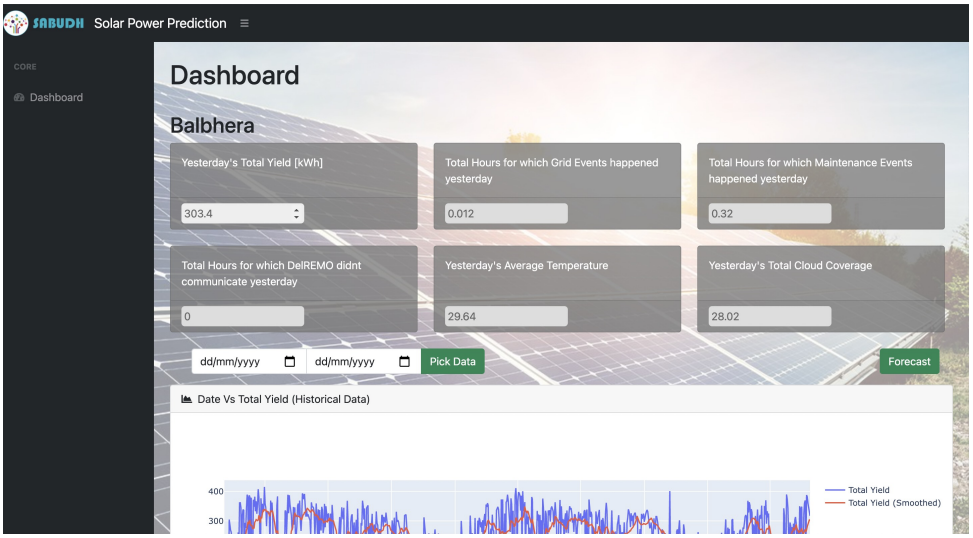

FIGURE 5.14: Deployment (1)

FIGURE 5.15: Deployment (2)

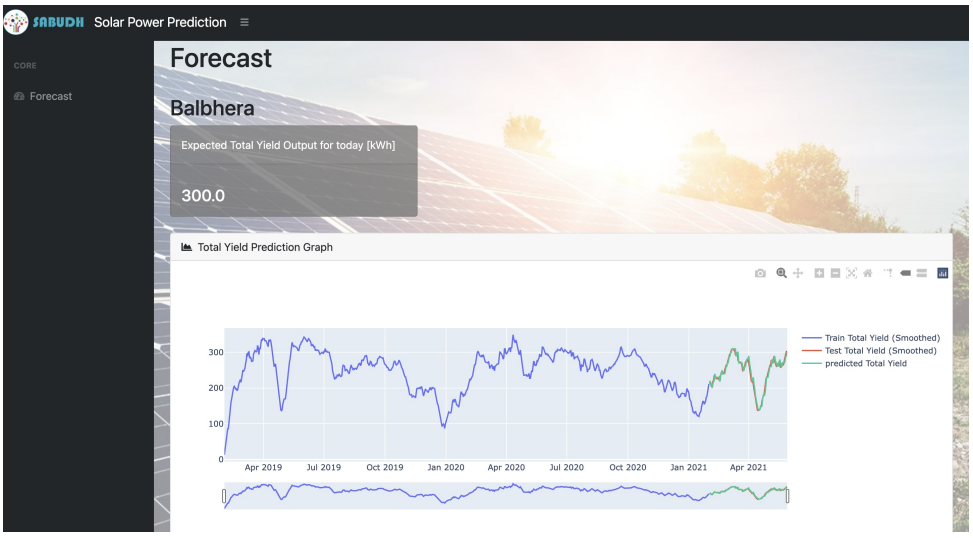

FIGURE 5.16: Deployment (3)

In the deployment part, the Index page includes 6 input fields namely,

- **Yesterday's Total Yield [kWh]:** expects the user to input the recorded total yield [kWh] the previous day.
- **Total Hours for which Grid Events happened yesterday:** expects the user to input the hours for which the grid events happened.
- **Total Hours for which Maintenance Events happened yesterday:** expects the user to input the hours for which the maintenance events happened.
- **Total Hours for which DelREMO didn't communicate yesterday:** expects the user to input the hours for which DelREMO site didn't communicate.
- **Yesterday's Average Temperature:** expects the user to input the recorded average temperature the previous day.
- **Yesterday's Total Cloud Coverage:** expects the user to input the recorded total cloud coverage the previous day.

Also the Index page displays the Date vs Total Yield plot and Date vs Temperature plot formed with historical data provided. Also these plots can be tuned based on the date range provided to view the historical plots for particular time period.

The next coded page is the Forecast page which displays the **Expected Total Yield Output for today** [**kWh**] based on the inputs recorded from the Index page. It also displays the Total Yield Forecast plot based on the testing dataset provided to the used model.

Chapter 6

Conclusion and Future Scope

6.1 Conclusion

This study aimed to discover power quality events associated with the grid and estimate the solar PV output loss (in kWh).

6.2 Future Scope

- The study's findings can be leveraged to improve the smoothness of the grid and minimise PV output loss.
- Setting up Investment plans and ROI to have a clear picture of the costs and returns associated with solar installments in rural areas.
- The estimation of PV output loss can be further extended to estimation of carbon dioxide (CO2) emissions and financial loss to determine the impact and severity.

Bibliography

[1] P. Moriarty and D. Honnery, "Can renewable energy power the future?" *Energy Policy*, vol. 93, pp. 3–7, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S030142151630088X

[2] C. Bauner and C. L. Crago, "Adoption of residential solar power under uncertainty: Implications for renewable energy incentives," *Energy Policy*, vol. 86, pp. 27–35, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S030142151500227X

[3] M. Sorkun, C. Paoli, and O. Incel, "Time series forecasting on solar irradiation using deep learning," 11 2017.

[4] M. R. Tye, S. E. Haupt, E. Gilleland, C. Kalb, and T. Jensen, "Assessing evidence for weather regimes governing solar power generation in kuwait," *Energies*, vol. 12, no. 23, 2019. [Online]. Available: https://www.mdpi.com/1996-1073/12/23/4409

[5] M. B. A. Shuvho, M. A. Chowdhury, S. Ahmed, and M. A. Kashem, "Prediction of solar irradiation and performance evaluation of grid connected solar 80kwp pv plant in bangladesh," *Energy Reports*, vol. 5, pp. 714–722, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352484719302057