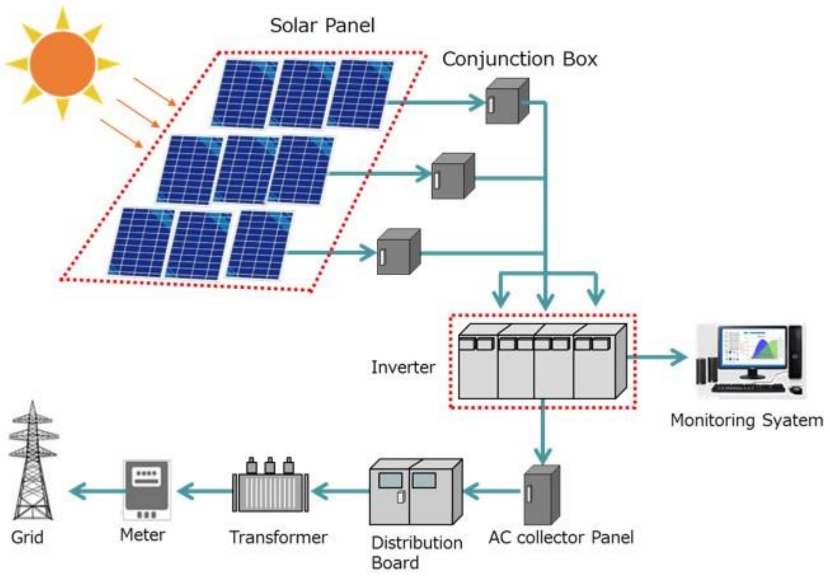
MACHINE LEARNING MODEL DEVELOPMENT FOR PREDICTION AND FORECASTING OF SOLAR POWER GENERATION USING WEATHER DATA

PASINDU LIYANGE

INTRODUCTION

- Solar power is a form of energy harnessed from the power and heat of the sun's rays which makes it one of the most important renewable sources of energy.
- At the present, solar power generation is becoming increasingly popular all around the world due the replenishing nature of non-renewable sources of power generation and hence countries are heavily investing in establishment of solar power plants to meet the energy demands of the future world.



The major challenge with solar power generation is the unpredictable nature of its output. Since the solar power is directly dependent on the power and heat of solar radiation, the production output could vary drastically depending on the weather conditions.

Key Objective of this project is the Development of a Machine Learning Model to accurately forecast the solar power generation with the usage of weather data to support the authorities in effectively planning the supply demand models.

DATASET ANALYSIS

- The dataset used for the model development is taken from an opensource dataset which has been gathered at two solar power plants in India over a 34-day period which consists of two pairs of files - each pair has one power generation dataset and one sensor readings dataset.
- The power generation datasets are gathered at the inverter level where each inverter has multiple lines of solar panels attached to it. The sensor data is gathered at a plant level with a single array of sensors optimally placed at the plant.

Solar Plant 01 Inverter Data Samples

	DATE_TIME	PLANT_ID	SOURCE_KEY	DC_POWER	AC_POWER	DAILY_YIELD	TOTAL_YIELD
0	15-05-2020 00:00	4135001	1BY6WEcLGh8j5v7	0.0	0.0	0.0	6259559.0
1	15-05-2020 00:00	4135001	1IF53ai7Xc0U56Y	0.0	0.0	0.0	6183645.0
2	15-05-2020 00:00	4135001	3PZuoBAID5Wc2HD	0.0	0.0	0.0	6987759.0
3	15-05-2020 00:00	4135001	7JYdWkrLSPkdwr4	0.0	0.0	0.0	7602960.0
4	15-05-2020 00:00	4135001	McdE0feGgRqW7Ca	0.0	0.0	0.0	7158964.0

Solar Plant 01 Weather Data Samples

	DATE_TIME	PLANT_ID	SOURCE_KEY	AMBIENT_TEMPERATURE	MODULE_TEMPERATURE	IRRADIATION
0	2020-05-15 00:00:00	4135001	HmiyD2TTLFNqkNe	25.184316	22.857507	0.0
1	2020-05-15 00:15:00	4135001	HmiyD2TTLFNqkNe	25.084589	22.761668	0.0
2	2020-05-15 00:30:00	4135001	HmiyD2TTLFNqkNe	24.935753	22.592306	0.0
3	2020-05-15 00:45:00	4135001	HmiyD2TTLFNqkNe	24.846130	22.360852	0.0
4	2020-05-15 01:00:00	4135001	HmiyD2TTLFNqkNe	24.621525	22.165423	0.0

Solar Plant 01 Inverter Data Info

#	Column	Non-Null Count	Dtype
0	DATE_TIME	68778 non-null	object
1	PLANT_ID	68778 non-null	int64
2	SOURCE_KEY	68778 non-null	object
3	DC_POWER	68778 non-null	float64
4	AC_POWER	68778 non-null	float64
5	DAILY_YIELD	68778 non-null	float64
6	TOTAL_YIELD	68778 non-null	float64

Solar Plant 01 Weather Data Info

#	Column	Non-Null Count	Dtype
0	DATE_TIME	3182 non-null	object
1	PLANT_ID	3182 non-null	int64
2	SOURCE_KEY	3182 non-null	object
3	AMBIENT_TEMPERATURE	3182 non-null	float64
4	MODULE_TEMPERATURE	3182 non-null	float64
5	IRRADIATION	3182 non-null	float64

METHODOLOGY

DATA PRE-PROCESSING

Data pre-processing is one of the most important parts of the machine learning model development which mainly focuses on preparing the data as per the requirements of the model development.

Handling null occurrences

```
solargen_data01.isnull().sum()
```

```
DATE_TIME      0
SOURCE_KEY     0
DC_POWER       0
AC_POWER       0
DAILY_YIELD    0
TOTAL_YIELD    0
DATE           0
TIME           0
dtype: int64
```

Dropping unnecessary columns

```
solargen_data01.drop(labels='PLANT_ID', axis=1, inplace=True)
```

```
weather_data01.drop(labels='PLANT_ID', axis=1, inplace=True)
```

```
solargen_data02.drop(labels='PLANT_ID', axis=1, inplace=True)
```

```
weather_data02.drop(labels='PLANT_ID', axis=1, inplace=True)
```

Creating separate columns for date and time

```
solargen_data01['DATE'] = solargen_data01['DATE_TIME'].apply(lambda x:x.date())
solargen_data01['TIME'] = solargen_data01['DATE_TIME'].apply(lambda x:x.time())

solargen_data01.sample(5)
```

```
weather_data01['DATE'] = weather_data01['DATE_TIME'].apply(lambda x:x.date())
weather_data01['TIME'] = weather_data01['DATE_TIME'].apply(lambda x:x.time())

weather_data01.sample(5)
```

```
weather_data02['DATE'] = weather_data02['DATE_TIME'].apply(lambda x:x.date())
weather_data02['TIME'] = weather_data02['DATE_TIME'].apply(lambda x:x.time())

weather_data02.sample(5)
```

Formatting timestamp as a datetime object

```
#Formatting DateTime
solargen_data02['DATE_TIME'] = pd.to_datetime(solargen_data02['DATE_TIME'])
solargen_data02.sample(5)
```

Replacing source keys with simple source IDs

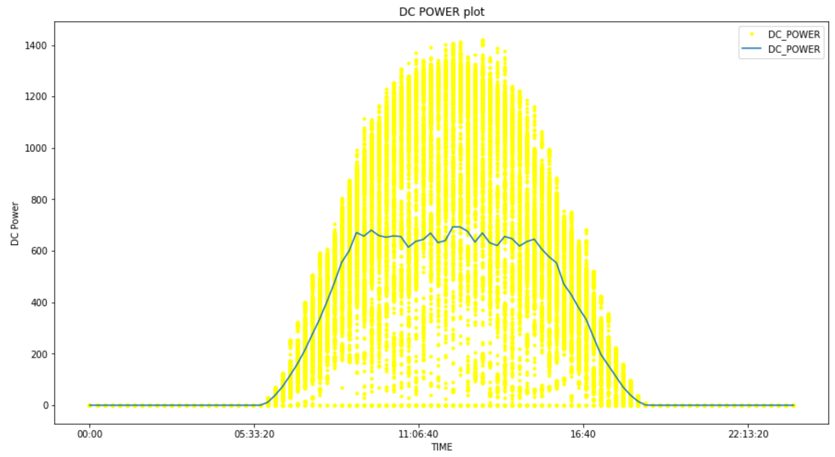
```
solar01_inverter_id = solargen_data01['SOURCE_KEY'].unique()
solargen_data01['SOURCE_KEY'] = solargen_data01['SOURCE_KEY'].apply(lambda x :  int(np.where(solar01_inverter_id == x)[0]))
```

METHODOLOGY

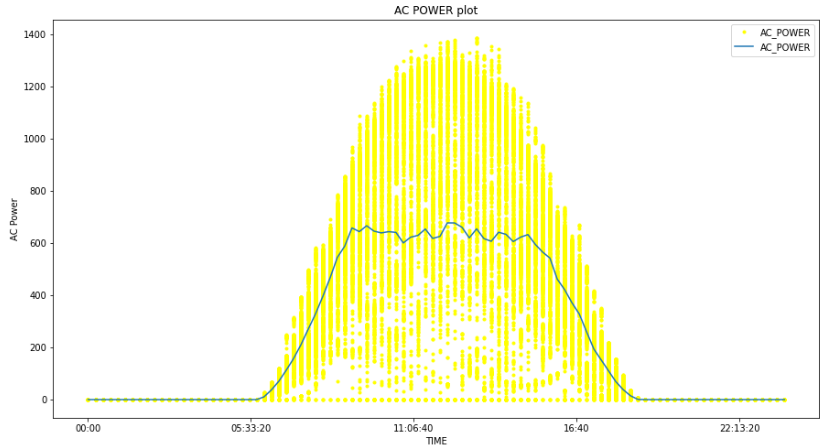
EXPLORATORY DATA ANALYSIS

Exploratory data analysis was carried out on the datasets to understand the patterns and variations in datasets to facilitate the machine learning model development

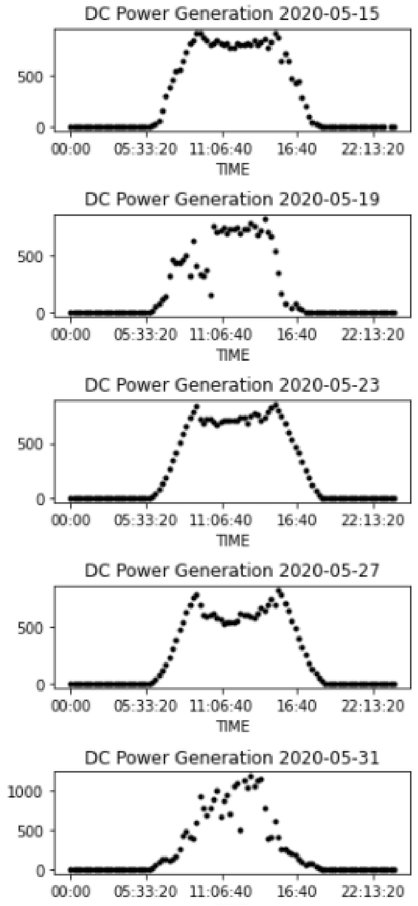
Solar Plant 01 DC Power Variation Analysis



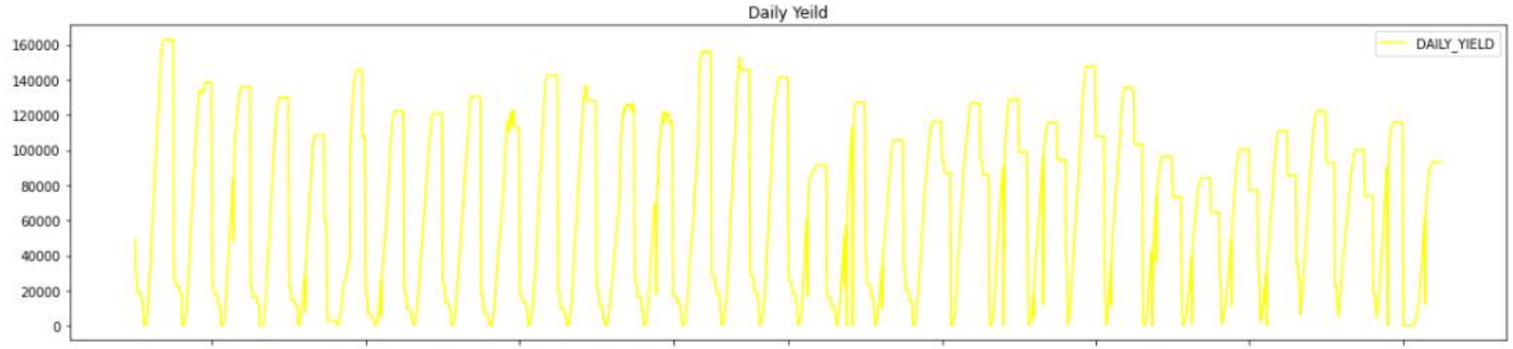
Solar Plant 01 AC Power Variation Analysis



Daily DC Power Generation



Solar Plant 01 DC Power Variation Analysis

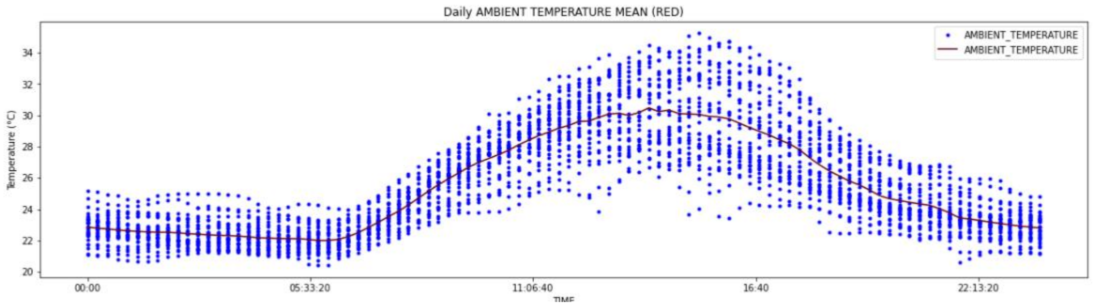


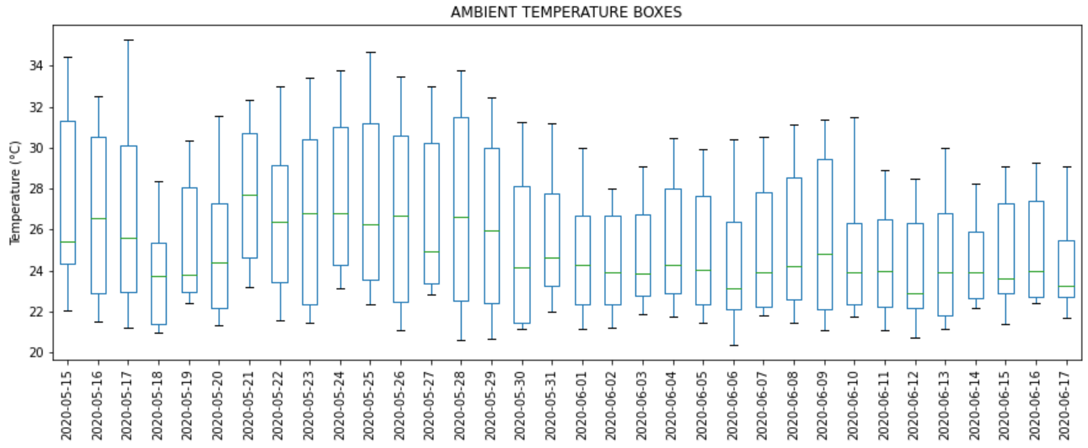
METHODOLOGY

EXPLORATORY DATA ANALYSIS

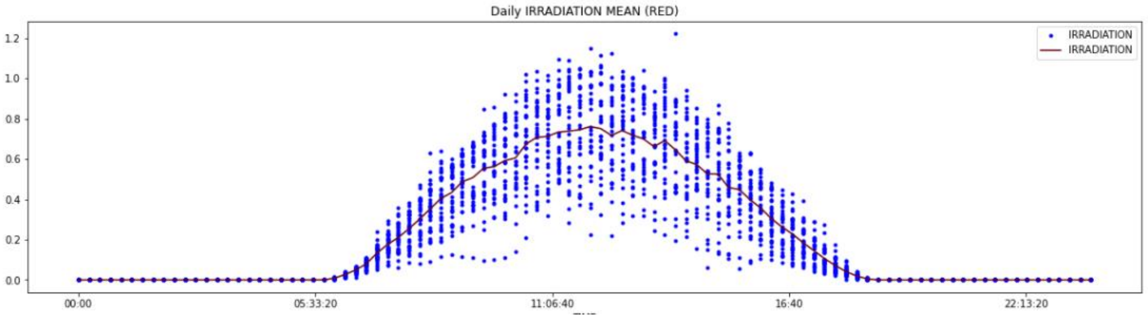
Exploratory data analysis was carried out on the datasets to understand the patterns and variations in datasets to facilitate the machine learning model development

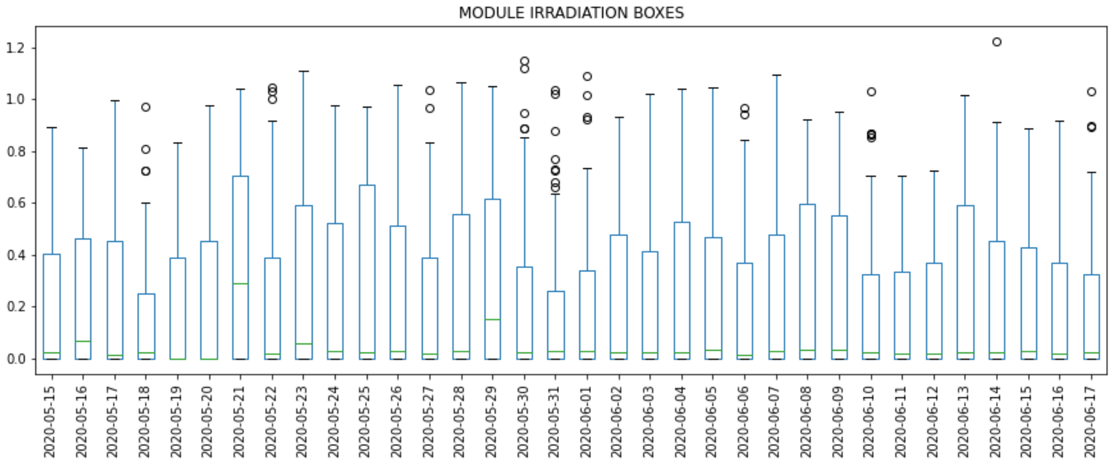
Ambient Temperature Variations





Irradiation Variations



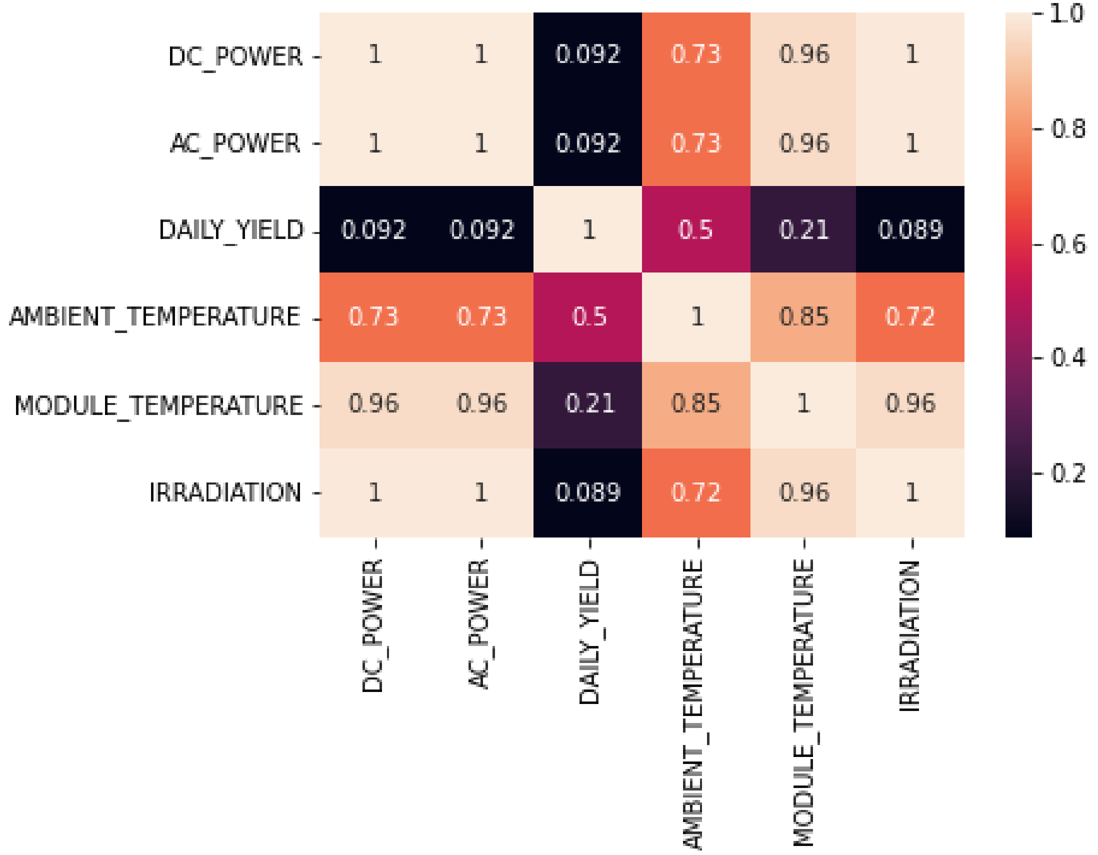


METHODOLOGY

CORRELATION IDENTIFICATION

Correlation matrix was developed for the combined data set of solar power generation and weather data to understand the relationships between solar power generation and weather data

Correlation Matrix



Observations

1. DC Power and AC Power are perfectly linearly related
2. DC Power and AC Power generated has a direct relationship with Irradiation
3. Module temperature is increasing proportional to the ambient temperature and Irradiation

Model Development Strategy

The main objective of the model development is to predict and forecast the solar power generation for the next few days.

As per the observations made during the analysis, it was evident that the AC Power will be directly dependent on irradiation, ambient temperature, and module temperature. However, since the predictions have to be done on weather forecasting module temperature will not be available for the predictions. Hence the final model prediction was done using the irradiation and ambient temperature as input parameters.

MODEL DEVELOPMENT

With the results and observation from exploratory data analysis, it was identified that there is a linear relationship between the AC output yield and some of the weather data. Hence it was decided to use '*Linear Regression*' to carry out the model development.

Type of Model	Linear Regression model from scikit learn
Independent Variables [X]	Irradiation & Ambient Temperature Since these two are parameters have the highest correlation with AC power and these two can be captured from weather forecast
Dependent Variable [Y]	AC Power Since AC Power is the measured output from the solar plant
Train Test Split	Train Data:Test Data – 80:20

```
X = plant01_merged_data[['IRRADIATION','AMBIENT_TEMPERATURE']]
X.head()
```

	IRRADIATION	AMBIENT_TEMPERATURE
0	0.0	25.184316
1	0.0	25.084589
2	0.0	24.935753
3	0.0	24.846130
4	0.0	24.621525

```
y = plant01_merged_data['AC_POWER']
y.head()
```

```
0    0.0
1    0.0
2    0.0
3    0.0
4    0.0
Name: AC_POWER, dtype: float64
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=5)
```

```
print(F"Train sample size = {len(X_train)}")
print(F"Test sample size  = {len(X_test)}")
```

```
Train sample size = 2525
Test sample size  = 632
```

```
model = linear_model.LinearRegression()
```

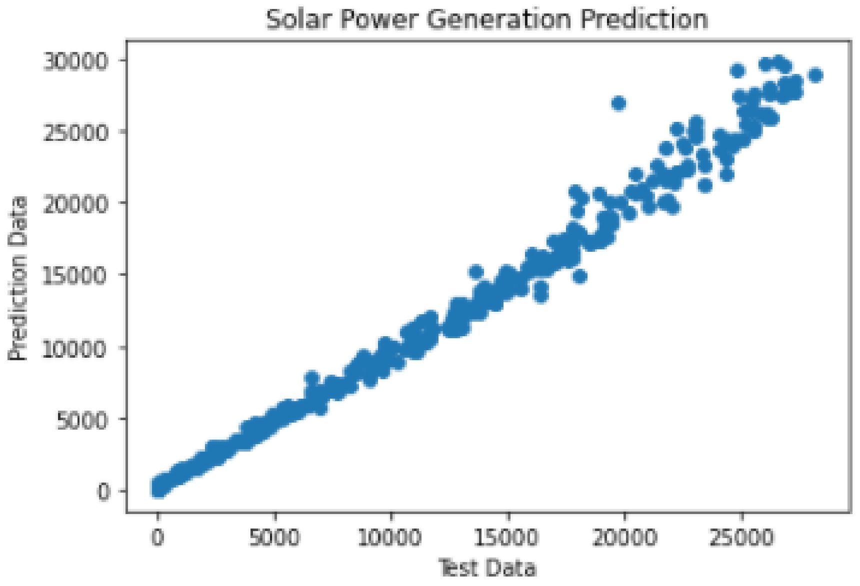
```
model.fit(X_train, y_train)
```

MODEL EVALUATION & RESULTS

Evaluation of the accuracy of the model was done using Mean Squared Error, Root Mean Squared Error and Coefficient of Determination Values

Linear regression model development for solar power generation was successfully completed after several rounds of fine tuning the model.

	Test Data	Predicted Data	Difference	Difference %
2512	318.075000	405.776768	-87.701768	27.572669
3102	14483.887500	13024.903174	1458.984326	10.073154
1218	20299.108929	20810.576090	-511.467161	2.519653
3131	2869.840000	2716.198578	153.641422	5.353658
133	19409.383929	19037.550758	371.833170	1.915739
2189	0.000000	155.190322	-155.190322	inf
1439	0.000000	150.973194	-150.973194	inf
520	3987.000000	4224.694285	-237.694285	5.961733
2485	0.000000	103.137491	-103.137491	inf
2666	0.000000	117.624187	-117.624187	inf




```
MSE: 578069.6881502356
RMSE: 760.3089425688978
R2: 0.9919610845837611
```

CONCLUSION & FUTURE IMPROVEMENTS

After several rounds of testing and fine tuning, a machine learning model was successfully developed using linear regression algorithms for the prediction of solar power generation using weather forecast information.

Model will be tested with Sri Lankan data in the coming days and will be integrated with a solar energy prediction application be used for Sri Lankan energy demand planning operations.