# Machine Learning Application in Power Generation Prediction

- ## Introduction and Background

There is a global shift of the majority of people towards renewable energy sources and among such renewable sources of energy, solar power energy is one of the most attractive ones. Solar power is the energy generated by converting the energy from the sunlight into electricity or thermal energy. Solar power is a major form of clean and free source of energy. It is the best alternative to the traditional energy source like fossil fuels, coal, gas, and oil which are non-renewable energy sources and are bound to get depleted if not used sustainably and preserved for future generations. It was predicted by BP that within 53 years, all of the oil reserves will be consumed at the current rate [1] [2]. The demand for global energy in 2018 was supplied with natural gases, oil, and coal, which resulted in a 1.7% rise in carbon emissions [3]. Solar power doesn't contribute to such carbon emissions as it is a source of renewable energy which is clean energy and if utilized properly can be the major alternative to non-renewable sources for the coming generation. So, to sum up, some of the major merits of using solar power that makes it attractive is that it is a clean energy source that doesn't contribute to pollution or global warming, it reduces our dependability on the dwindling supply of non-renewable sources, and it is easily producible. Furthermore, between 2010 and 2019, the average solar panel costs have decreased from $7.34 per watt to $2.53 per watt and cost estimation $2.22 per watt in 2021[4]. According to International Energy Agency (IEA), solar is the cheapest energy source in history[5].

It is very clear from the above merits that solar power is the future of energy source. The sun generates more energy in just one hour than the entire world uses in a year [6]. Hence, there is a lot of potential for solar energy generation. The generation of solar power is predicted to climb to 48% of total renewable energy generation in the U.S by 2050, in contrast to 11% in 2017 [7]. The amount of power that can be generated from a solar power source depends on various factors like Wind direction, Humidity, Pressure, Distance to Solar Noon, Temperature, etc. All these environmental factors affect power generation directly or indirectly to a certain degree.

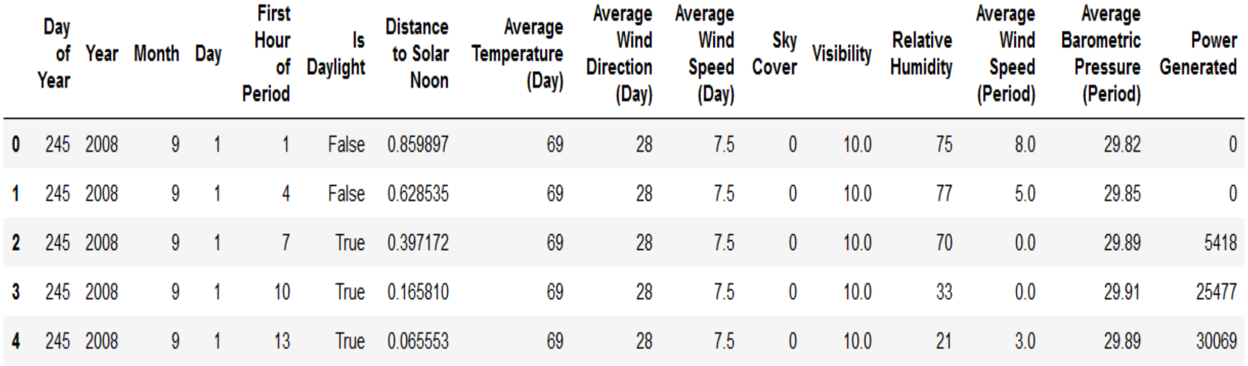
Unfavorable weather conditions can drastically reduce the amount of solar power that a power plant can derive. Hence, the power generation prediction of a power plant is done to extract valuable information. However, the lack of advanced technology of solar forecasting means inaccurate results from the forecasts [8]. It will then result in inadequate power generation [8]. So, to make up for these inaccurate results and lack of enough power, the grid operators in the power plants must use short-term power sources which are expensive [8]. Such short-term power sources supplement will most likely be from power generation companies that use fossil fuels [9]. This not only increases the cost drastically but also supports the burning of limited fossil fuels. Furthermore, it also increases the carbon emission in the environment. So, proper forecasting of power generation in a solar power plant is vital and is a major motivation for power plant owners to invest in accurate forecasting techniques. To maintain a sound relationship with the customers, power companies want to create stability of services by eliminating any disturbances in the power distribution [10]. So, power companies would want to know the exact amount of power they can provide in a certain time frame [10]. Solar power generation forecasting will help us to predict energy generation and maintain the balance between generation and consumption [11]. This information can also help us to minimize disruptions and costly power outages. There is a detailed explanation of the benefits of solar forecasting for energy imbalance markets in the paper [12]**.** Another major motivation of solar power generation prediction is that it provides us the information about where sunlight will strike and thus increasing the photovoltaic panels' efficiency vastly. It helps us to increase the efficiency of the solar power plant.

Hence, solar power generation forecasting has a substantial impact on energy power plants. However, as stated earlier the generation of solar power depends on various environmental conditions like the weather variables. That's why many researchers have applied machine learning concepts to forecast solar power generation. Some solar forecasting techniques have been proposed to date[13] [14] [15].

Considering all the fore-mentioned points, I am proposing to utilize the machine learning concepts to predict solar power generation by taking into account the effect of various environmental factors on solar power generation. I want to implement various machine learning models to accurately predict the power generation from the solar Photovoltaic (PV) cells or predict a certain range of power generation from those power plants.

- **Dataset Discussion**

This dataset was collected and compiled by Ph.D. candidate Alexandra Constantin. The dataset contains the information of the solar Photovoltaic(PV) systems at the University of California, Berkeley. The solar PV is installed at various sites within the campus in the building rooftops and parking areas. The dataset contains information on various environmental factors that could affect the power generation of the power plant. Those features are included as columns in the dataset. The total number of columns in the dataset is 16. It includes ‘Day of Year’, ‘Year’, ‘Month’, ‘Day’, ‘First Hour of Period’, ‘Is Daylight’, ‘Distance to Solar Noon’, ‘Average Temperature (Day)’, ‘Average Wind Direction (Day)’, ‘Average Wind Speed (Day)’, ‘Sky Cover’, ‘Visibility’, ‘Relative Humidity, ‘Average Wind Speed (Period)’, ‘Average Barometric Pressure (Period)’ and ‘Power Generated’. There are 2,921 rows in the dataset filled with integers for 10 columns, Boolean for one column, and float values for 5 columns, with one empty value for the Average Wind Speed (Period) column in the 714th row. The output label ‘Power Generated’ has 1529 unique values. In the project, most of the columns will be used for the machine learning model. The first five rows of the dataset along with the feature names are shown in figure 1.



| | Day of Year | Year | Month | Day | First Hour of Period | Is Daylight | Distance to Solar Noon | Average Temperature (Day) | Average Wind Direction (Day) | Average Wind Speed (Day) | Sky Cover | Visibility | Relative Humidity | Average Wind Speed (Period) | Average Barometric Pressure (Period) | Power Generated |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 245 | 2008 | 9 | 1 | 1 | False | 0.859897 | 69 | 28 | 7.5 | 0 | 10.0 | 75 | 8.0 | 29.82 | 0 |
| 1 | 245 | 2008 | 9 | 1 | 4 | False | 0.628535 | 69 | 28 | 7.5 | 0 | 10.0 | 77 | 5.0 | 29.85 | 0 |
| 2 | 245 | 2008 | 9 | 1 | 7 | True | 0.397172 | 69 | 28 | 7.5 | 0 | 10.0 | 70 | 0.0 | 29.89 | 5418 |
| 3 | 245 | 2008 | 9 | 1 | 10 | True | 0.165810 | 69 | 28 | 7.5 | 0 | 10.0 | 33 | 0.0 | 29.91 | 25477 |
| 4 | 245 | 2008 | 9 | 1 | 13 | True | 0.065553 | 69 | 28 | 7.5 | 0 | 10.0 | 21 | 3.0 | 29.89 | 30069 |

Figure 1: Dataset Sample

- **Procedures**

The main goal of the project is to forecast the power generation of a solar power station by utilizing the information from the dataset described above. The dataset consists of various features that need to be first analyzed before using the machine learning techniques. So, the first step in the process is ‘Data Preprocessing’, which is to clean the data and check for any inconsistency in the datasets like empty feature values for certain rows. ‘Encoding’ will also be done to transform the nominal features. Principal component analysis which is a technique for feature extraction will be done to prepare the data for the model. The label or target feature of the dataset is the power generated feature which is shown in table 1. Most of the columns will be used as the features for the machine learning model.

The last column named ‘Power Generation’ will be the target feature or the label. As the dataset consists of the target feature; supervised machine learning algorithms will be used in the project. Some of those supervised machine learning algorithms are logistic regression, decision trees, and support vector machines. All these models are to be implemented using the Scikit-learn library. The dataset will be split into two sets. One set will be the training set and the other set will be the test set, which will be used for the testing of the model. Hyperparameter tuning will also be performed to optimize the model[16]. ‘Pipeline’, ‘Grid-search’, and validation techniques will be used for optimization. Various supervised learning algorithms as mentioned above will be used with the training and testing data. The accuracy of those models will be noted along with precision, recall, and Receiver Operating Characteristic Area under the curve (ROC AUC) curve. The basic flow chart of the procedure is shown in figure 2.

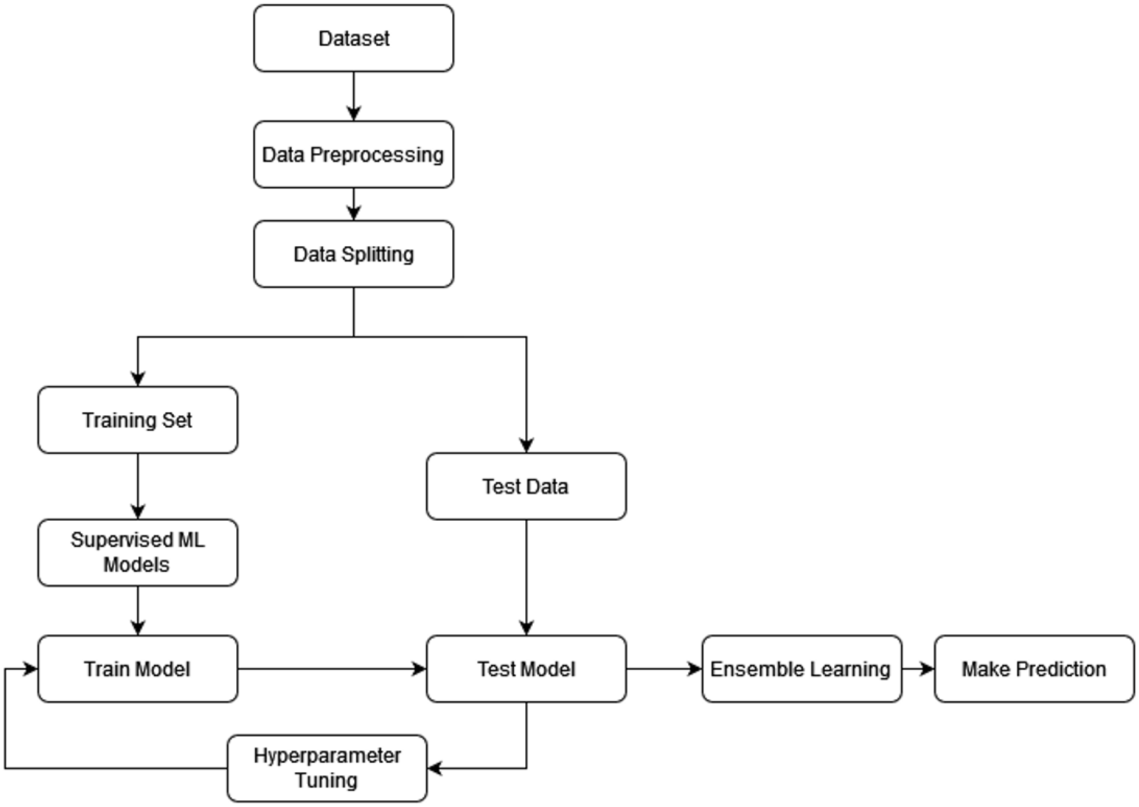


Figure 2: Procedure Flowchart

- **Validation Process**

The output of the machine learning techniques will be validated by using one of the various validation processes. One of the common choices for validation is the cross-validation technique, which can be implemented to determine the optimal values of the various hyperparameters used for various machine learning techniques. One of the cross-validation processes is the k-fold cross-validation technique. The training data is split randomly into k number of folds[16]. For the training of the model, k-1 number of folds are used and for evaluation of the performance, a single fold is used [16]. The process is performed ‘k’ number of times [16]. Accuracy of the training and testing data and precision are some numerical metrics to analyze the performance of supervised learning. Learning curves, ROC (Receiver Operating Characteristics) AUC (Area Under the Curve) score can provide the graphical evaluation of the machine learning model.

The best model will be validated after the ensemble process. The test set, which is the data sample not seen by the model before, will be tested on the model. The accuracy score from the training set, validation set, and testing set will be compared. The accuracy of the model will be validated by the accuracy of the model for the test set.

## References

[1] “When Fossil Fuels Run Out, What Then? - MAHB.” https://mahb.stanford.edu/library-item/fossil-fuels-run/ (accessed Oct. 17, 2021).

[2] “How long will world’s oil reserves last? 53 years, says BP,” *Christian Science Monitor*, Jul. 14, 2014. Accessed: Oct. 17, 2021. [Online]. Available: https://www.csmonitor.com/Environment/Energy-Voices/2014/0714/How-long-will-world-s-oil-reserves-last-53-years-says-BP

[3] J. Howarth, “When will fossil fuels run out?,” *Octopus Energy*. https://octopus.energy/blog/when-will-fossil-fuels-run-out/ (accessed Oct. 17, 2021).

[4] “2021 Solar Panel Costs | Average Installation Cost Calculator,” *HomeGuide*. https://homeguide.com/costs/solar-panel-cost (accessed Oct. 20, 2021).

[5] “Solar is now ‘cheapest electricity in history’, confirms IEA,” *Carbon Brief*, Oct. 13, 2020. https://www.carbonbrief.org/solar-is-now-cheapest-electricity-in-history-confirms-iea (accessed Oct. 20, 2021).

[6] “Solar - City of Berkeley, CA.” https://www.cityofberkeley.info/solar/ (accessed Oct. 17, 2021).

[7] "Renewable Energy," *Center for Climate and Energy Solutions*, Oct. 21, 2017. https://www.c2es.org/content/renewable-energy/ (accessed Oct. 17, 2021).

[8] M. Chou, "Solar Forecasting," Nov. 06, 2017. http://large.stanford.edu/courses/2017/ph240/chou1/ (accessed Oct. 17, 2021).

[9] F. Jawaid and K. Junejo, "Predicting Daily Mean Solar Power Using Machine Learning Regression Techniques," Sep. 2016. DOI: 10.1109/INTECH.2016.7845051.

[10] B. Carrera and K. Kim, "Comparison Analysis of Machine Learning Techniques for Photovoltaic Prediction Using Weather Sensor Data," *Sensors*, vol. 20, no. 11, p. 3129, Jun. 2020, DOI: 10.3390/s20113129.

[11] S. Dise, "What is the value of accurate solar forecasting for utility-scale PV plants?," *SolarAnywhere*, Mar. 08, 2017. https://www.solaranywhere.com/2017/accurate-solar-forecasting-value/ (accessed Oct. 17, 2021).

[12] A. Kaur, L. Nonnenmacher, H. Pedro, and C. Coimbra, "Benefits of solar forecasting for energy imbalance markets," *Renew. Energy*, vol. 86, pp. 819–830, Feb. 2016, DOI: 10.1016/j.renene.2015.09.011.

[13] C. Voyant *et al.*, "Machine learning methods for solar radiation forecasting: A review," *Renew. Energy*, vol. 105, pp. 569–582, May 2017, DOI: 10.1016/j.renene.2016.12.095.

[14] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. J. Martinez-de-Pison, and F. Antonanzas-Torres, "Review of photovoltaic power forecasting," *Sol. Energy*, vol. 136, pp. 78–111, Oct. 2016, doi: 10.1016/j.solener.2016.06.069.

[15] T. Chuluunsaikhan, A. Nasridinov, W. S. Choi, D. B. Choi, S. H. Choi, and Y. M. Kim, "Predicting the Power Output of Solar Panels based on Weather and Air Pollution Features using Machine Learning," *J. Korea Multimed. Soc.*, vol. 24, no. 2, pp. 222–232, Feb. 2021, DOI: 10.9717/KMMS.2021.24.2.222.

[16] S. Raschka and V. Mirjalili, *"Python machine learning : machine learning and deep learning with Python, Scikit-learn, and TensorFlow 2 (Third edition)". Packt Publishing, Limited. 2019.*, Third Edition.