Solar Energy Prediction using Machine Learning with Support Vector Regression Algorithm

Idamakanti Kasireddy(✉), K. Padmini, R. V. D. Ramarao, B. Seshagiri, and B. Venkata Naga Rani

Vishnu Institute of Technology, Bhimavaram, India
kaasireddy.i@vishnu.edu.in

Abstract. Machine Learning is almost applied in every field such as engineering, science, medical etc. In this work, the concept of machine learning has been adopted for predicting solar energy. The solar Energy is widely known renewable energy due to its massive advantages. Solar energy prediction can help to determine the energy consumption beforehand and plays a major role in future planning. The grid operators are facing hardships because of unreliable weather conditions, which lead to the reduction in solar energy output. So, they are unable to satisfy the needs of consumers. Our proposed solution intends to make prediction models by using machine learning algorithms such as Linear Regression, Lasso Regression, Ridge Regression and Support Vector Regression (SVR). These algorithms use past weather data including temperature, dew, wind, cloud and visibility. Based on these data, analysis has been carried out in Jupyter Notebook. From the analysis, it has found that, SVR algorithm performed well when compared with other algorithms.

Keywords: Solar energy · Machine learning · Support Vector Regression (SVR)

1 Introduction

Solar energy has many advantages but it also has its own downside which is its production is highly irregular. Grid operators are facing problems because of the irregular output of solar energy as they are unable to meet the demand of consumers. But, by predicting solar energy grid operators can satisfy the consumers. If the production of solar energy at particular time is known then it is easy to plan according to the needs of users. There are many factors that affect the solar energy prediction. To predict solar energy one must collect data regarding these factors. The factors that are effecting the solar energy production involve wind, humidity, temperature and dew etc. Solar energy production can be predicted by analyzing the data which includes various factors that affect the solar energy with respect to the time and some other essential data. The data required can be collected from meteorological department or it is easily available in online. Machine

N. Gupta et al. (Eds.): IC4S 2022, LNICST 472, pp. 17–25, 2023.
https://doi.org/10.1007/978-3-031-28975-0_2

learning is the most effective approach to analyze data. It is classified into three types; they are supervised learning, unsupervised learning and reinforcement learning [1–6].

In supervised learning the model is provided with input data to get desired output. In unsupervised learning only input data is given to the model, leaving it on its own to find structure and in reinforcement learning a computer program interacts with a dynamic environment in which it must perform a certain goal. As it navigates its problem space, the program is provided feedback that's similar to rewards. In this study supervised learning has been used. Supervised learning is further classified into two types which include regression and classification. The most used regression algorithms are such as linear regression, lasso regression, ridge regression and support vector regression. Regression is basically predicting the value of dependent variables using independent variables.

Machine learning follows specific steps during its implementation, which includes data collecting, data analyzing, data wrangling, train & test and accuracy check. Firstly, the data is collected and collected data is analysed in order to check for duplicate values, Null values and wrong format. If null values are present in any rows of data those rows can be dropped or should be filled with appropriate values. The rows with duplicate values should be removed and wrong format is altered into correct format. This process is known as data cleaning. Then the model should be trained with 80% of data and remaining 20% is used to test the model. Finally the model undergoes accuracy check. Accuracy can be measured by using various methods such as R-Squared method. The algorithm is most efficient if it is more accurate.

Python simplifies the Machine Learning Algorithms by providing some libraries. Libraries that are used for machine learning are numpy, pandas, matplotlib, pyplot, sklearn and seaborn. These algorithms make machine learning algorithms easy to implement. Python is most used to implement machine learning algorithms due to its advantages such as python is easiest language and it provides many libraries. The platforms that are used to run machine learning algorithms include Anaconda, Google Colab and Jupyter. Among them Jupyter is simpler and it is very easy to share files using Jupyter Notebook. Jupyter is an open source which allows users to do mathematical computation such as trigonometry and Fourier transforms.

2 Methodology

2.1 Linear Regression

In this study we have used various regressions. Linear Regressions is one of the basic regression [7]. It helps to find the relation between independent and dependent variables. This regression is based on fitting best line into the graph and uses various methods to reduce error between best fit line and data points. Best fit line is also known as regression line. Linear Regression uses least-squares method to fit a line to the data and it uses R-squared value is the statistical measure of how close the data to the regression line. R-squared value is considered as accuracy and if its value is more than 0.5 than model is considered as good. It is used in Trend Forecasting, Evaluating trends and sales estimates.

Slope for the estimated regression equation is given by (1) and (2)

$$b_1 = \frac{\sum (x_1 - \overline{x})(y_1 - \overline{y})}{\sum (x_1 - \overline{x})^2} \tag{1}$$

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = \frac{\sum_i (y_i - \widehat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2} \tag{2}$$

2.2 Lasso Regression

Lasso Regression [8] is used for Regularization. It is the technique to prevent model from getting over fitting. Sometimes the machine learning model works well with training data set but, when it is tested with testing data set it produces high cost function when compared to training data set hence it leads to over fitting. In case of Lasso regression the lambda is multiplying with the weights.

$$\sum_{i=1}^{n} (y_i - \sum_j x_{ij}\beta_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3}$$

2.3 Ridge Regression

Ridge Regression [8] is also used for Regularization. It is the technique to prevent model from getting over fitting. Sometimes the machine learning model works well with training data set but, when it is tested with testing data set it produces high cost function when compared to training data set hence it leads to over fitting. In case of Ridge Regression the cost function is changed by adding the penalty term to it. The amount of penalty added to the model is known as Ridge Regression Penalty. We can calculate it by multiplying with the lambda to squared weights

$$\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = RSS + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{4}$$

2.4 Support Vector Regression

Support Vector Regression algorithms [9] are used to predict discrete values. Support Vector Regression uses the same principle as SVMs. In SVR we find the best fit line. Best fit line is the hyper plane that has the maximum no. of points. The SVR tries to fit the best line with a threshold value. The threshold value is the distance between hyper plane and boundary line. The advantages of SVR when compared to other algorithms are it is very easy to implement and it has high prediction capability.

2.5 Purpose

These four algorithms are analysed in this study to find out most accurate method for solar energy prediction. SVR turns out to be most accurate model among the other models.

3 Results

Firstly, in this analysis it is important to know about dependent and independent variables. In this data solar energy is the dependent variable and various factors affecting the solar energy are independent variables. Now, we will see the implementation of machine learning models.

Various regression algorithms are implemented and analysed for predicting solar energy in Jupyter Notebook. Initially, the python libraries such as numpy, pandas, matplotlib.pyplot and seaborn have been imported. The following commands are used to import libraries.

$$\textit{import pandas as pd}$$
$$\textit{import numpy as np}$$
$$\textit{import matplotlib.pyplot}$$
$$\textit{import seaborn as sn}$$

Then the solar weather data has been imported into the Jupyter notebook using the following command.

$$\textit{pd.read\_excel('datainfo.xls')}$$

Here the command varies with file type. If the file is csv type then the command changes to read_csv and name of the file should be written inside the brackets.

The imported data has been analysed using various commands such as info(), describe(), isnull().sum(). These methods are used to analyze data, data must be analysed in order to know about empty values, duplicate values and wrong formatted values. info() method is used to get information related to data such as type of value, is value null or non-null value, describe() is used to get information related to no of rows, max value, min value, standard deviation, 25%, 50% and 75% of values. The difference between values must be minimized in order to get more accuracy.

If any null values are present in data then the rows with null values must be removed or replaced with a suitable value. The rows with duplicate values must be dropped and rows with wrong format should be altered into correct format. In our data this type of values are absent.

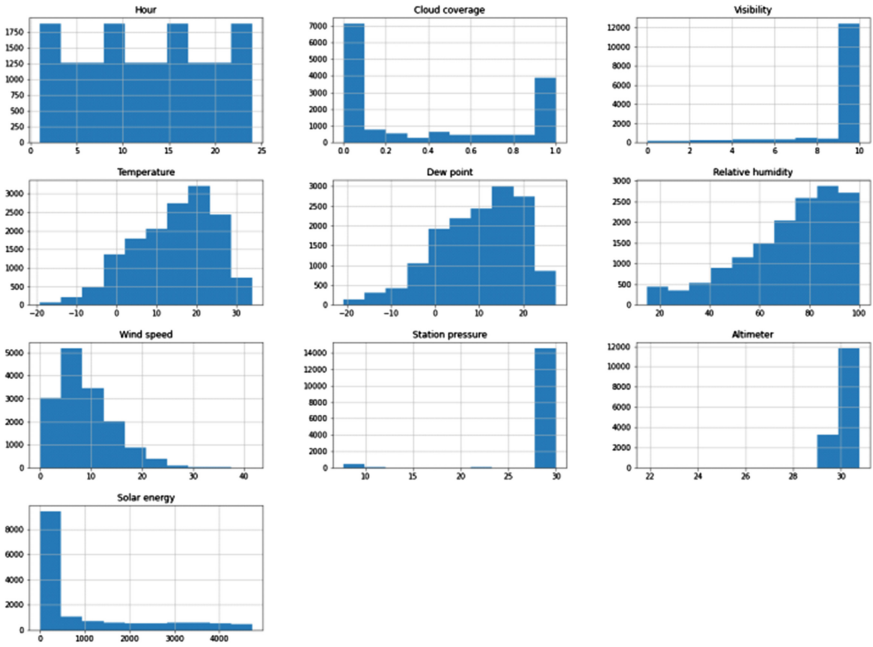

Fig. 1. Solar data histogram

The above graph shows the values present in columns of the data. The x-axis shows values in columns and y-axis shows no. of rows are present in data with same value. The above histogram can be drawn by using various methods present in matplotlib. pyplot library. Not only histograms but we can draw various graphs using the methods present in this library (Fig. 1).

Figure 2 shows the correlation between columns. If correlation between any columns is greater than 0.5 or −0.5 than they are highly correlated and also it varies from −1 to +1. Negative values stand for negative correlation and it shows inverse proportionality. Positive value indicates direct proportionality. The value of 1 is one to one relationship. Corr() method is used to find correlation and this method ignores the non-numeric values. In data non numeric values should be converted into numeric values for accurate analysis.

The training and testing of data can be done by importing train_test_split from sklearn.model selection library. It can be done by using following command.

$$\textit{from sklearn.model\_selectionimporttrain\_test\_split}$$

The various algorithms can be directly implemented by using sklearn library. The following command is used to implement algorithms.

$$\textit{from sklearn.linear\_modelimportLinearRegression, Lasso, Ridge}$$

These algorithms should be trained by using data.

$$\textit{lr=LinearRegression()}$$
$$\textit{lr.fit(x\_train, y\_train)}$$

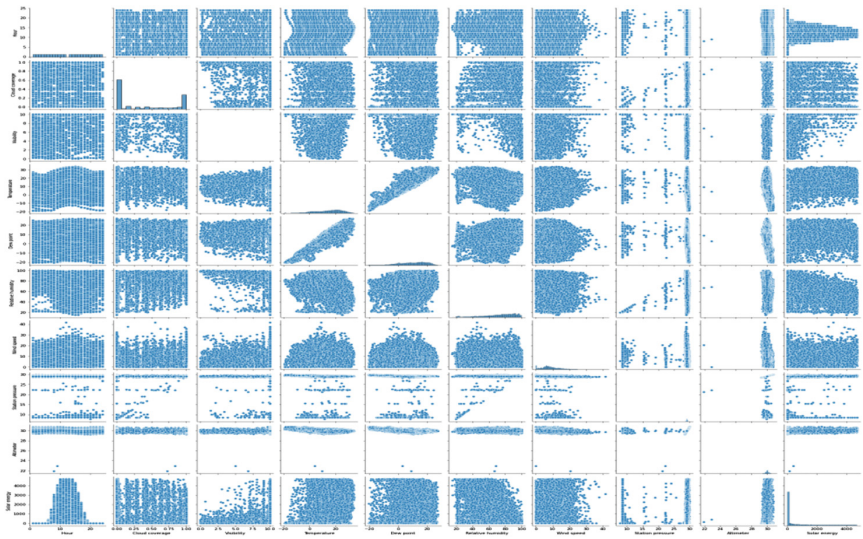

Fig. 2. Correlation pair plot

In the above command the linear regression is taken as lr and it is trained by using fit method. The same process follows for both Ridge and Lasso regressions.

$$las = Lasso(alpha = 0.1)$$
$$las.fit(x\_train, y\_train)$$
$$rid = Ridge()$$
$$rid.fit(x\_train, y\_train)$$

las represents lasso regression and rid represents ridge regression. As the training of model completed, next the testing of data should be done. For testing also various python provides various libraries.

$$las.predict(x\_train)$$
$$las.score(x\_test, y\_test)$$

In the above command the models is tested by using testing data and the accuracy of the model can be known by using above commands. The above commands are implemented for both Ridge and Linear Regression. SVR has to be imported from SVM as SVR is a part of SVM. For this also sklearn library is used and SVR also tested and trained in the same way as mentioned above.

Finally, the accuracies of various models are obtained as follows (Table 1).

Table 1. Comparison of various algorithms

Algorithm	Accuracy
Linear regression	51.3%
Lasso regression	51.2%
Ridge regression	51%
Support vector regression	88.4%

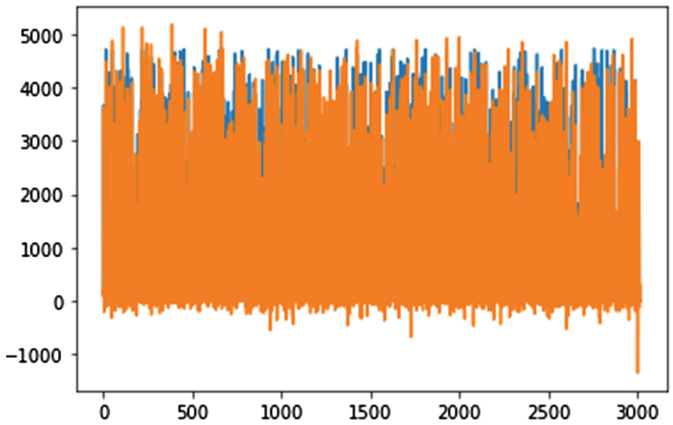

Fig. 3. Comparison of testing output (blue color) and predicted output (yellow color) for SVR

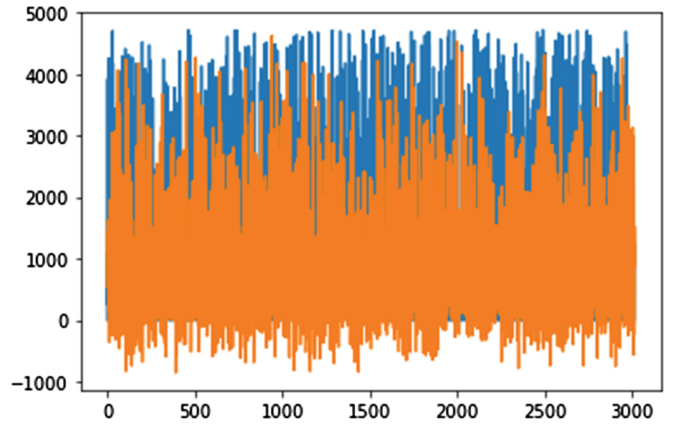

Fig. 4. Comparison of testing output (blue color) and predicted output (yellow color) for Linear Regression

Figures 3, 4, 5, and 6 depicts solar energy predictions for various algorithms. From Fig. 3, 4, 5 and 6, it is noticed that SVR algorithm outperforming when compared to

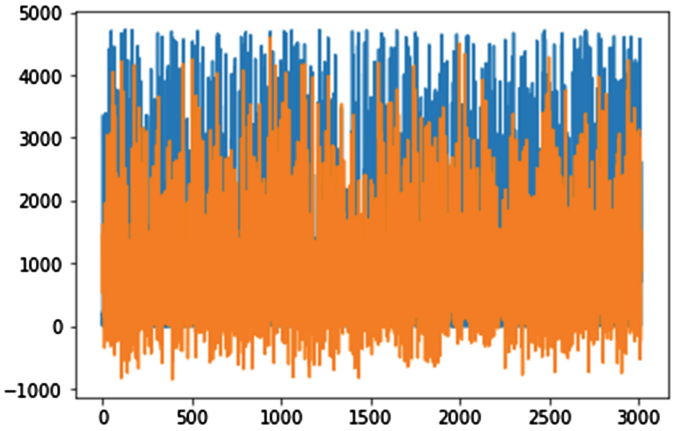

Fig. 5. Comparison of testing output (blue color) and predicted output (yellow color) for Lasso Regression

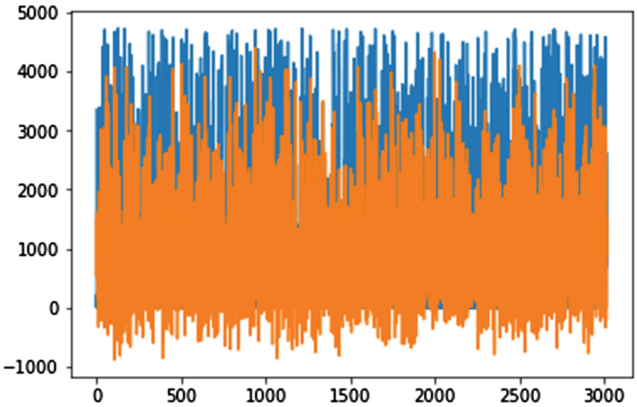

Fig. 6. Comparison of testing output (blue color) and predicted output (yellow color) for Ridge Regression

remaining algorithms. Hence it can be conclude that Support Vector Regression is most accurate model when compared to others.

4 Conclusion

In this work machine Learning models such as Lasso Regression, Ridge Regression, Linear Regression and Support Vector Regression are implemented and analyzed in Jupyter Notebook for solar energy prediction. From the analysis, accuracies of Lasso, Ridge and Linear regressions are found in between 51%–52%. Whereas the accuracy of Support Vector Regression is 88.4%. Hence it is clear that Support vector regression outperformed the remaining regressions models for predicting solar energy.

References

1. Javed, A., et al.: Predicting solar irradiance using machine learning techniques. In: 15th International Wireless Communications and Mobile Computing Conference (IWCMC), pp. 1458–1462 (2019)
2. Ak, R., et al.: Two machine learning approaches for short-term wind speed time-series prediction. IEEE Trans. Neural Netw. Learn. Syst. **27**(8), 1734–1747 (2016)
3. Nasir, A.W., Kasireddy, I., Rahul Tiwari, B.K., Ahmed, I., Furquan, A.: Data-based tuning of PI controller for first-order system. In: Bhaumik, S., Chattopadhyay, S., Chattopadhyay, T., Bhattacharya, S. (eds.) Proceedings of International Conference on Industrial Instrumentation and Control: ICI2C 2021, pp. 547–555. Springer Nature Singapore, Singapore (2022). https://doi.org/10.1007/978-981-16-7011-4_52
4. Sharma, A., Kakkar, A.: Forecasting daily global solar irradiance generation using machine learning. Renew. Sustain. Energy Rev. **82**(3), 2254–2269 (2018)
5. Kasireddy, I., et al.: Application of FOPID-FOF controller based on IMC theory for automatic generation control of power system. IETE J. Res. **68**(3), 2204–2219 (2022). https://doi.org/10.1080/03772063.2019.1694452
6. Kasireddy, I., et al.: Determination of stable zones of LFC for a power system considering communication delay. AIP Conf. Proc. **2418**, 040014 (2022). https://doi.org/10.1063/5.0081986
7. Maulud, D., Abdulazeez, A.M.: A review on linear regression comprehensive in machine learning. JASTT **1**(4), 140–147 (2020)
8. Yang, X., et al.: Lasso regression models for cross-version defect prediction. IEEE Trans. Reliab. **67**(3), 885–896 (2018)
9. Crone, S.F., Guajardo, J., Weber, R.: A study on the ability of support vector regression and neural networks to forecast basic time series patterns. In: Bramer, M. (ed.) IFIP AI 2006. IIFIP, vol. 217, pp. 149–158. Springer, Boston, MA (2006). https://doi.org/10.1007/978-0-387-34747-9_16