Journal of Renewable Energies

Revue des Energies Renouvelables

journal home page : https://revue.cder.dz/index.php/rer

Cubist Regression, Random Forest and Support Vector Regression for Solar Power Prediction

Souhaila Chahboun [a,*], Mohamed Maaroufi [a]

[a] *Department of Electrical Engineering, Mohammadia School of Engineers, Mohammed V University in Rabat, Rabat, Morocco*

** Corresponding author, E-mail address: souhaila.chahboun93@gmail.com*

Tel.: + 212 639507337

Abstract

At a time when the energy transition is inescapable and artificial intelligence is rapidly advancing in all directions, solar renewable energy output forecasting is becoming a popular concept, especially with the availability of large data sets and the critical requirement to forecast these energies, known to have a random nature. Therefore, the main goal of this study is to investigate and exploit artificial intelligence's revolutionary potential for the prediction of the electricity generated by solar photovoltaic panels. The main algorithms that will be studied in this article are cubist regression, random forest and support vector regression. This forecast is beneficial to both providers and consumers, since it will enable for more efficient use of solar renewable energy supplies, which intermittency makes their integration into the existing electrical networks a challenging task.

Keywords: Photovoltaic, Machine learning, Artificial intelligence, Solar energy, Prediction

1. Introduction

Climate change, energy resource shortages, cost growth, and environmental pollution are all driving demand for green energy. Solar energy, which is the oldest form of energy, is being promoted in particular [1]. However, due to the insecurity of renewable energies [2], production sources have diversified and the network has become increasingly hard to maintain. As a result, anticipating the electricity generated by renewable energies has become critical.

The evaluation and forecasting of energy demands is one of the key problems of facility managers. From the perspective of power system grid operation, short-term load forecasting is critical. The short-term time frame might range from half-hourly forecasting to monthly forecasting. Accurate forecasting would aid the utility in terms of grid dependability and stability, ensuring that sufficient supply is available to fulfil demand.

In this regard, artificial intelligence's machine learning appears to be one of the tools to achieve this goal, as it makes renewable energies more predictable and hence more valuable [3]. Machine learning approaches can provide insights into improving distribution, balancing energy consumption loads, and managing oscillations in renewable energy output by evaluating and examining vast amounts of data from renewable energy producers or connected items. The advantages of artificial intelligence prediction are not just financial for energy producers, but they also increase the supply reliability of the entire electrical system and lend legitimacy to new energy sources, making their integration into the energy mix easier. In this study, machine learning is used to forecast the hourly PV power output. With irradiance and climatic data as inputs and PV power as output, the entire system can be viewed as a black box. Most traditional models believe that there are only linear relationships between the model's input features and the PV power output, whereas in this work the investigated methods, namely cubist regression, random forest and, support vector regression, have the ability to learn non-linear relationships. Finally, using residual analysis, the chosen regression models are graphically evaluated. The main purpose of this study is to compare the three machine learning algorithms listed above to.

2. Materials and Methods

2.1 Data source

The PV output power is obtained hourly from a PV installation with a total capacity of 6 KW. The solar irradiation and meteorological datasets used in this study, on the other hand, were obtained from SoDa, a free data source. In this study, five parameters are employed as inputs, as shown in Table 1.:

Table 1. Input Parameters

Parameter	**Unit**
Global Horizontal irradiation	Wh⁄m²
Hour	Hour
Ambient Temperature	°C
Cell Temperature	°C
PV panels efficiency	%

2.2 Machine learning algorithms

2.2.1 Support vector regression

Support vector regression (SVR) predicts future data using a learning method generated from past data [4]. [4]. The basic idea behind this approach is to find support vectors that maximize

the difference between two-point classes determined from the difference between the target value and a threshold. The kernel notion is added to the SVR technique because the majority of real problems have nonlinear features [5].

2.2.2 Random forest

Random forest (RF) extracts bootstrap data samples to form trees independently. Furthermore, each node in the tree is generated using a random collection of variables. The forecasting results obtained indicate the average prediction of the individual trees [6]. This method allows to make a forest with a number of decision trees. The forecast becomes stronger as more trees are added, resulting in increased accuracy.

2.2.3 Cubist regression

Cubist regression (CB) is a rule-based regression technique that was created using a mixture of Quinlan's principles. Unlike RF, CB retrieves a set of models rather than a single final model. Cubist models are a type of decision tree modeling in which the data is subset using rules. There are two steps in the main algorithm. The first step is to create a set of criteria for segmenting the data into smaller subsets. The second component of the procedure uses a regression model to arrive at a forecast for these smaller groupings[7].

2.3 Optimization of the Hyperparameters

To avoid overfitting, machine learning algorithms must be properly tuned. Users can control the complexity of these algorithms by adjusting the hyper-parameters. mtry is employed in the case of RF approach. At each split, it shows the number of indicators that are randomly sampled as candidates. In addition, there are two hyper-parameters to tune in the CB algorithm: neighbors (#Instances) and committees (#Committees). These two parameters are the most likely to have the most impact on the Cubist model's final performance. Finally, for a nonlinear SVR with a Gaussian radial basis function kernel, (#Cost) parameters is used.

3. Results and discussion

3.1 Final models

The final hyperparameters of the investigated methods are presented as follows (see Fig. 1, Fig. 2 and Fig. 3):

3.1.1 Support vector regression

The final value used for the model is Cost = 128.

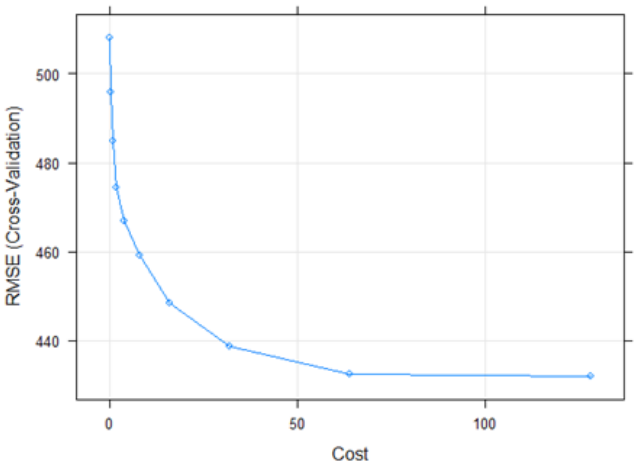


Fig 1. SVR model plot

3.1.2 Random forest

The final value used for the model is mtry = 2.

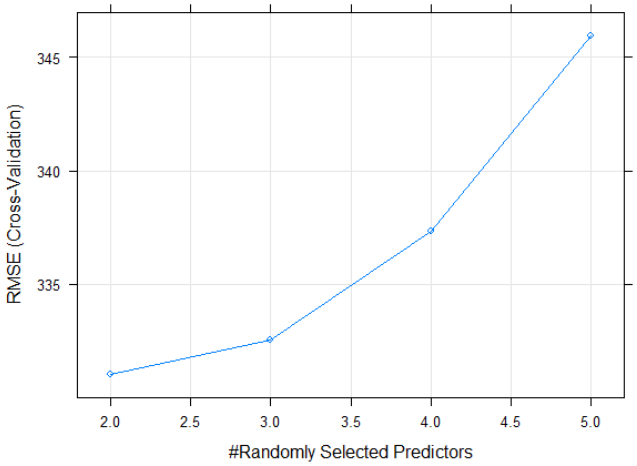


Fig 2. RF model plot

3.1.2 Cubist regression

The final values used for the model are committees = 20 and neighbors = 9

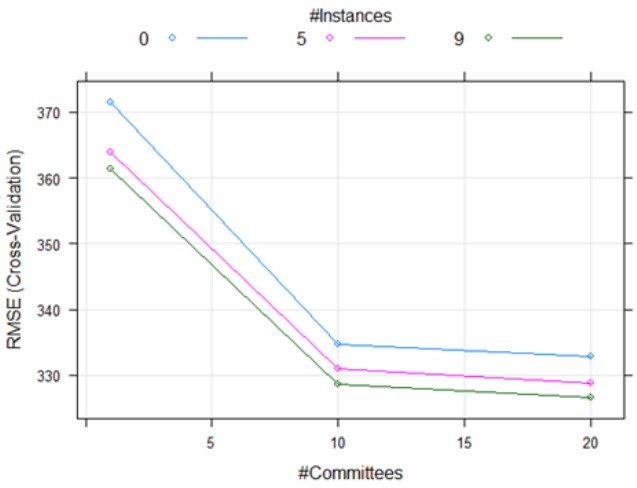


Fig 3. CB model plot

3.2 Performance metrics

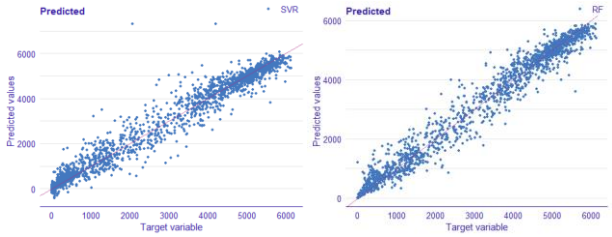
For the training and testing phases, the accuracy of regression models is assessed using the key performance metrics R^2, RMSE, and MAE, as seen in Table 2 and Table 3:

Table 2. Performance metrics in the training phase (80%)

Machine learning	R^2	RMSE (Kw)	MAE (Kw)
SVR	0.9672	0.3831	0.2502
RF	0.9950	0.1496	0.9654
CB	0.9858	0.2516	0.1667

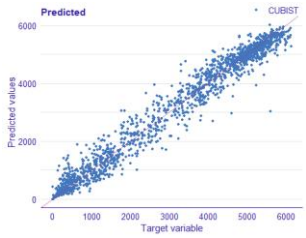
Table 3. Performance metrics in the testing phase (20%)

Machine learning	R^2	RMSE (Kw)	MAE (Kw)
SVR	0.9606	0.4191	0.2722
RF	0.9759	0.3274	0.2171
CB	0.9765	0.3232	0.2153



SVR

RF



CB

Fig.4 Predicted versus observed values plots

3.3 Residual analysis

The trained model's residuals are examined using residual analysis. As shown in Fig. 5, the first plot is residual boxplot, which illustrates the distribution of absolute residual values.

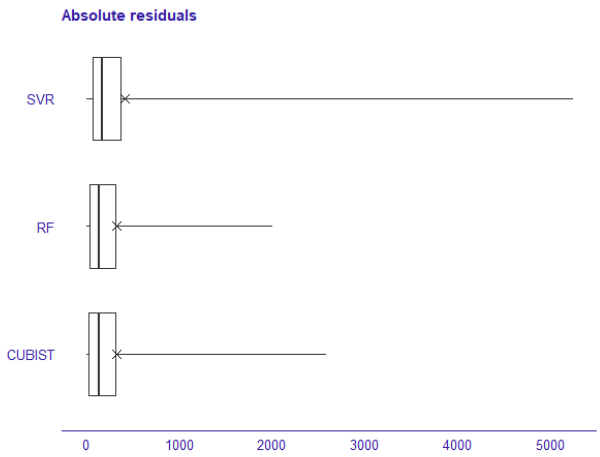


Fig.5 Residual Boxplot

REC (Regression Error Characteristic) curves (see Fig. 6) are a more advanced variant of the widely used Receiver Operating Characteristic (ROC) curves. [8]. In comparison to alternatives such as performance metric tables, the use of REC curves allows for a better visual comparison of regression models as well as a more persuasive representation of regression results.

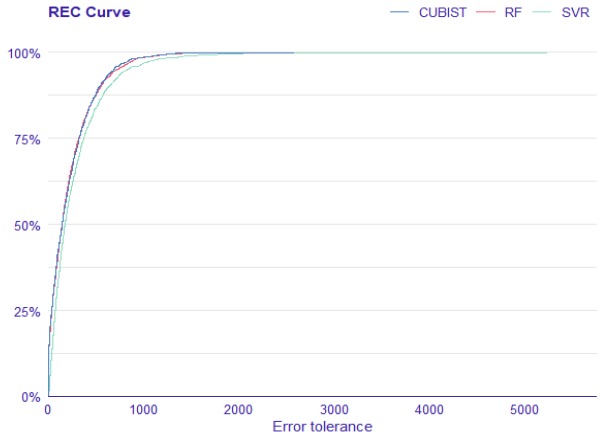


Fig.6 REC Curve

3.4 Discussion

Based on the results of performance metrics obtained in Tables 2 and 3, it can be seen that CB and RF achieved the best balance between the expected and observed values with an R^2=97% in the testing phase, followed by SVR algorithm with an R^2=96%. The results obtained for the three approaches studied in this work are similar, this is mainly due to the fact that these algorithms are more promising than conventional regressions because they better incorporate the dynamics of data relationships and capture nonlinear correlations between input and output variables. Moreover, residual analysis carried out in our study, confirms the results obtained. When we look at residual boxplots (see Fig. 4), we can observe that CB regression has the smallest residuals, followed by RF, and SVR. Moreover, as presented above, REC curve (see

Fig. 5) is frequently used to assess and analyze the quality of models at various tolerance levels. The steady improvement in accuracy indicates that there are no issues with the models.

4. Conclusions

Two significant contributions are made by this paper. To begin, a comparison study is conducted to discover which machine learning algorithms produce the most accurate photovoltaic output power prediction. Second, despite the fact that multiple studies have proposed various approaches such as neural network methods, this study has demonstrated that easy to implement algorithms such CB, RF and SVR may be highly effective in forecasting.
Finally, it is worth noting, too, that most machine-learning research in renewable-energy forecasting has centred on solar or wind energy forecasting. Instead of solar and wind energy projections, other forms of renewable energy predictions, such as biomass energy, wave energy, and hydraulic power, could be potential areas for future research. In addition, hybrid models may be promising techniques to predicting renewable energies.

5. References

[1] S. K. H. Chow, E. W. M. Lee, and D. H. W. Li, "Short-term prediction of photovoltaic energy generation by intelligent approach," Energy Build., vol. 55, pp. 660–667, 2012, doi: 10.1016/j.enbuild.2012.08.011.

[2] B. Wang, J. Che, B. Wang, and S. Feng, "A Solar Power Prediction Using Support Vector Machines Based on Multi-source Data Fusion," 2018 Int. Conf. Power Syst. Technol. POWERCON 2018 - Proc., no. 201805280000160, pp. 4573–4577, 2019, doi: 10.1109/POWERCON.2018.8601672.

[3] S. Jogunuri and F. T. Josh, "Artificial intelligence methods for solar forecasting for optimum sizing of PV systems: A review," Res. J. Chem. Environ., vol. 24, no. I, pp. 174–180, 2020.

[4] K. A. Baharin, H. Abd Rahman, M. Y. Hassan, and C. K. Gan, "Hourly Photovoltaics Power Output Prediction for Malaysia Using Support Vector Regression," Appl. Mech. Mater., vol. 785, pp. 591–595, 2015, doi: 10.4028/www.scientific.net/amm.785.591.

[5] M. H. D. M. Ribeiro and L. dos Santos Coelho, "Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series," Appl. Soft Comput. J., vol. 86, p. 105837, 2020, doi: 10.1016/j.asoc.2019.105837.

[6] L. Visser, T. Alskaif, and W. Van Sark, "Benchmark analysis of day-ahead solar power forecasting techniques using weather predictions," Conf. Rec. IEEE Photovolt. Spec. Conf., pp. 2111–2116, 2019, doi: 10.1109/PVSC40753.2019.8980899.

[7] K. John, N. M. Kebonye, P. C. Agyeman, and S. K. Ahado, “Comparison of Cubist models for soil organic carbon prediction via portable XRF measured data,” Environ. Monit. Assess., vol. 193, no. 4, pp. 1–15, 2021, doi: 10.1007/s10661-021-08946-x.

[8] J. Hernández-Orallo, “ROC curves for regression,” Pattern Recognit., vol. 46, no. 12, pp. 3395–3411, 2013, doi: 10.1016/j.patcog.2013.06.014.