# Solar Radiation Prediction Using Decision Tree and Random Forest Models in Open-Source Software

*Lisbeth* Tucumbi [1], *Jefferson* Guano [1], *Roberto* Salazar-Achig[1*], and *Diego L.* Jiménez J.[1]

[1]Facultad de Ciencias de la Ingeniería y Aplicadas, Universidad Técnica de Cotopaxi, Ecuador

**Abstract.** The present research focuses on solar radiation prediction, which is important for energy production in thermal and solar systems. For this purpose, open-source software (Python) and a methodology involving the creation, implementation, and testing of specific machine learning models random forest (RF) and decision tree (DT) were used. The metrics used to identify the effectiveness of the models in predicting solar radiation were the coefficient ($R^2$), the mean square error (MSE), and the mean absolute error (MAE). The evaluation of the two methods is presented in three cases: for one, two, and seven days. The results show that the RF model has better results than the DT, with MAE and MSE values of 36.96 and 4238.77, respectively, and a determination coefficient of 0.96. The study emphasizes the importance of selecting the appropriate model based on the prediction horizon to estimate solar availability and improve solar and thermal energy system planning.

**Keywords:** Decision tree, machine learning, random forest, solar radiation, prediction, Python.

## 1 Introduction

Meteorological variables have been of immense importance since ancient times, leading to the creation of measurement methods aimed at forecasting these parameters. In this context, mathematical, statistical, and software methods are applied, which are commonly non-linear [1]. Currently, interest in sustainability has grown, and energy efficiency has increased. An accurate forecast of meteorological variables is essential for electricity production using renewable energy sources. Effective solar radiation forecasting is essential to optimizing the performance of solar and thermal energy systems.

However, accurately forecasting solar radiation is challenging due to the complexity of atmospheric phenomena and temporal and spatial variability. Although robust models and traditional statistical methods such as linear regression and time series exist, their precision and generalizability are often limited. On the contrary, machine learning has established itself as an effective solution to improve the accuracy of solar radiation forecasting.

* Corresponding author: edgar.salazar7619@utc.edu.ec

Machine learning algorithms can identify complex non-linear relationships in large volumes of data, making them suitable for modeling solar radiation under various conditions. Unlike traditional models, machine learning methods can continually learn and adapt to new data.

In this work, we use machine learning models to forecast solar radiation, specifically focusing on the decision tree and random forest models, which we implement in the Python programming environment. Since decision trees offer a clear and easy-to-interpret structure, random forests combine multiple decision trees, improving the model's robustness and precision in such a way as to reduce the risk of overfitting.

## 1.1 Relevant research

The study [2] discusses the advantages and disadvantages of various prediction methods applied to the early warning of solar radiation. This includes the application of various regression techniques, including line regression, polynomial, decision tree, and random forest, as well as their complementarity with the gradient. Where the comparison of forecast accuracy is carried out by splitting data for implementation and training, defining the accuracy and finding the discrepancy between the estimated values and the actual values, it can be measured using different evaluation criteria, applying evaluation criteria such as the Mean Square Error (MSE) and the coefficient of determination ($R^2$) to calculate the precision of the model.

The work carried out by [3], carries out the measurement in percentage of the (MSE and NRMSE) between several linear regression models such as decision tree and random forest. It obtained values less than 12% error in all cases, resulting in a lower random forest error value, defining the latter as the precise algorithm without taking into account the computational weight. The model uses an incomplete database at certain times on some days. The absence of data leads to an unacceptable percentage, necessitating the use of techniques such as Monte Carlo to fill the information gap.

Finally, [4] presents the application of the Monte Carlo method for decision tree and random forest optimization. The implementation of these methods reduces the percentage of mean squared error and normalized mean squared error based on completing non-existent data in the database.

In many cases, decision tree and random forest models are preferred due to their simplicity, efficiency, and robustness. Decision Tree is easier to interpret and works well with structured data, providing competitive results without requiring large amounts of data [5]. Random Forest in conjunction with Decision Tree improves stability and reduces the risk of overfitting, making them more resistant to noise and unbalanced data. Both models require fewer computing resources and are faster to train [6].

Neural networks, on the other hand, require larger data sets and complexity to learn meaningful patterns and are less interpreted [7]. In addition, gradient boosting is considered a more efficient model, but it is more prone to overfitting; it requires careful selection of hyperparameters, as it requires careful tuning to avoid overfitting or misfitting; and it is slow to train, especially for large amounts of data [8].

Table 1 offers a bibliographic summary of the past four years, examining several research studies that showcase the potential for solar radiation predictions through decision trees and random forests.

**Table 1.** Literature Review.

| Year of publication | | Authors | Comments | Reference |
|---|---|---|---|---|
| 1 | 2021 | Shanmugasundar, G., Vanitha, M., Čep, R., Kumar, V., Kalita, K., Ramachandran, M. | The article develops a solar radiation prediction model using collected data. The study demonstrates that, despite the widespread use of neural networks, other techniques like random forest and AdaBoost also demonstrate efficiency and accuracy in solar radiation prediction. | [9] |
| 2 | 2022 | Ordoñez Palacios, L.E., Bucheli Guerrero, V., Ordoñez, H. | This article presents a study that utilizes the random forest algorithm to predict solar radiation from satellite images. This technique produced a determination index ($R^2$) of 0.82 and an RMSE of 107.05. The study's goal is to investigate the discrepancy between the prediction with random forest and the NASA M1 model. | [10] |
| 3 | 2023 | Ehteram, M., Shabanian, H. | This article presents a solar radiation prediction model using the random forest algorithm. They used real data, measured by a pyranometer. | [11] |

### 1.2 Nomenclature

**Random Forest**

- Hj(x): Estimated class for instance x.
- argmaxy: Class selection mechanism with the highest number of votes.
- I(Hj(x)=Y): Indicator function responsible for determining whether the jth model assigns class y to input x.

**Gini gain**

- Aij: Features used to divide the tree.
- j: Class number.
- C: Shows the total set of data or a specific category of it.
- Mi: Different values that belong to the characteristic Ai.
- p(Aij): Contribution of probability that Aij takes its jth value.
- p(Ck/Aij): Demonstrates the probability that an example belongs to Ck for each attribute Aij to take a jth value.
- G(C/Aij): Represents the impurity measure of the data set C by dividing it according to the attribute A in its partition Ai.

**Mean Absolute Error (MAE) and Mean Square Error (MSE)**

- n: Sample size.
- xi: Prediction value.
- yi: Real value.
- ỹi: Prediction value of observation i.
- $R^2$: Represents the coefficient of determination

## 2 Methods

Accurate solar radiation forecasting has various applications but is especially important in the electrical field. In this context, it enables the optimal planning of solar and thermal energy generation, adjusting production to the expected demand under the forecasted climatic conditions.

Figure 1 illustrates the methodology for solving the proposed solar radiation forecasting problem. First, we need an input database to establish the historical solar radiation behavior. Next, we need to use machine learning models to train these data and validate the forecast results. The precision of models is evaluated with error metrics in order to determine which model has greater precision.

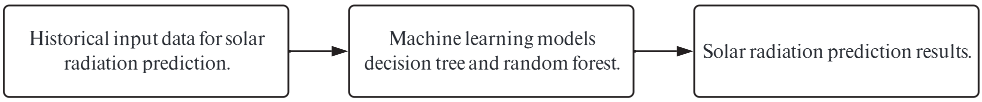



**Fig. 1.** Processes for determining the solar radiation forecast.

## 2.1 Data processing

The Meteorological Station historical database of the National University of Piura was used. From this database, six specific meteorological variables were selected for the study, with solar radiation being the main variable. Furthermore, five additional variables were considered to adjust the prediction model: outdoor temperature, relative humidity, wind speed, solar energy, and ultraviolet radiation intensity.

Data from the period from April 25, 2023 to April 25, 2024 were used, with hourly measurements (24 per day). One year of historical data was taken, which is important for forecasting. There were some missing data in the database, with 421 missing data corresponding to 18 days of the 8760 annual measurements. To solve this problem, the Monte Carlo method was applied, which allows completing the time series in a coherent and robust manner, improving the reliability of the forecasts [12]. Figure 2 shows how this method completed the nonexistent data in the repository.

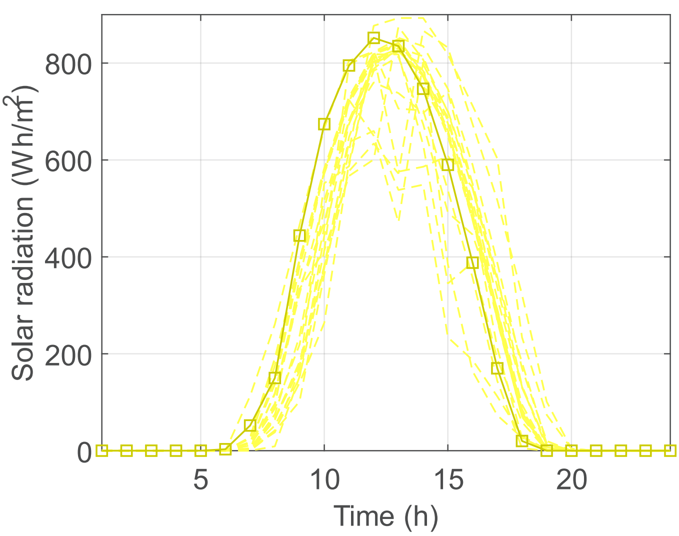



**Fig. 2.** Synthetic series using the Monte Carlo method of Solar Radiation.

Figure 3's block diagram shows the input data used for the training and validation of the decision tree and random forest learning models, with a focus on solar forecasting.

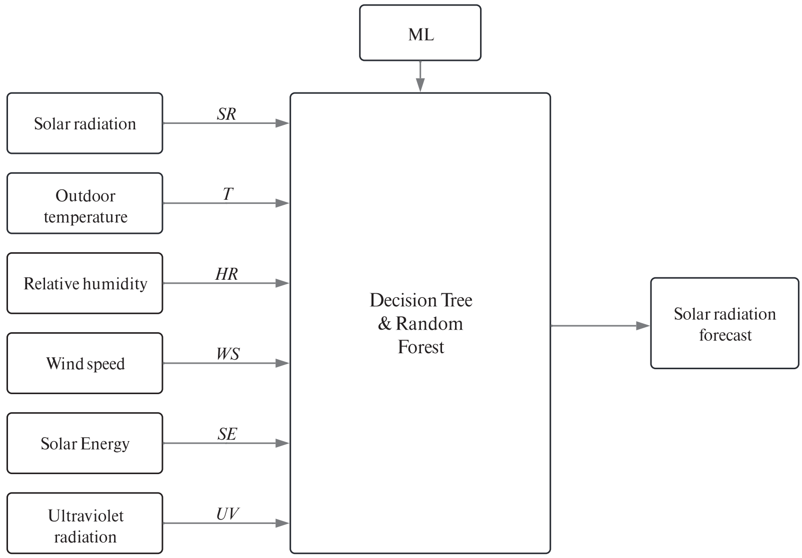



**Fig. 3.** Block diagram used for solar radiation forecasting.

## 2.2 Algorithms

The decision tree and random forest machine learning algorithms were applied using the scikit-learn Python library, which includes the necessary algorithms for making forecasts. The algorithms used to implement these models are presented below.

### *2.2.1 Decision Tree*

Equation 1 represents the algorithm that utilizes the decision tree learning model.

$$Entropy(p) = -\sum_{j=1}^{n} \frac{|pj|}{|p|} log\left(\frac{|pj|}{|p|}\right) \quad (1)$$

The variable n signifies the total number of classes in the database p. The variable n represents the variety of categories under evaluation; |pj| determines the number of elements of class j found in the data set, while p represents the frequency measure of class j. |p| indicates the data set's total size. It gives the total number of elements found in the set; |pj|/|p| performs the calculation of the proportion of elements in the set of class j. After that, log(|pj|/|p|) measures the amount of information belonging to class j; finally, the negative sign in the algorithm guarantees a positive entropy [13]. Equation 2 can be used to calculate the information gain.

$$Gain(p, j) = Entropy(p - entropy(jlp)) \quad (2)$$

The weighted entropy of the subsets obtained after breaking down the set p according to the characteristic is represented by the information gain entropy $(jlp)$. Then, using Equation 3, the impurity measurement is carried out:

$$G(A_i) = \sum_{j=1}^{M_i} p(A_{ij}) G\left(\frac{C}{A_{ij}}\right) \quad (3)$$

Where $M_i$ represents the total number of possible values of the characteristic $Ai_i$, $p(Ai_i)$ expresses the probability of an element in the data set that is part of $A_{ij}$, $G(C/A_{ij})$ indicates the Gini impurity of data set C [13].

Where, $G(C/Aij)$ is equal to the expression of Equation 4:

$$G\left(\frac{C}{A_{ij}}\right) = -\sum_{j=1}^{M_i} p\left(\frac{C_k}{A_{ij}}\right)\left(1 - p\left(\frac{C_k}{A_{ij}}\right)\right) \tag{4}$$

$p(C_k/A_{ij})$ shows the probability of an element in a subset when $A_{ij}$, belongs to $C_k$.

### 2.2.2 Random Forest

The Random Forest algorithm utilizes decision tree algorithms and divides data based on the gain of attributes, as determined by Equation 5 [14]:

$$Importance(f) = \frac{1}{N}\sum_{r=1}^{Ta}\left(Impurity_{befores} - Impurity_{after}\right) \tag{5}$$

Where $Importance(f)$ indicates the importance of feature f in the model. It measures the contribution of the attribute f to the decrease in impurity in the model trees. N indicates the total number of nodes where the feature f is useful for splitting, Ta represents the total number of trees in the model. Impuritybefores shows the impurity of the node before the breakdown using the feature $f$. It can also be calculated with the Gini impurity equation, entropy or any other method for measuring impurity metrics. $Impurity_{after}$ presents the impurity measure of the child nodes after the split using the characteristic $f$, this impurity is the combination of the resulting nodes after the split [14]. After this, random forest uses the following algorithm to define the final prediction using Equation 6:

$$H(x){:}\, argmax_y \sum_{j=1}^{k}\left(I\left(H_j(x) = Y\right)\right) \tag{6}$$

Where $argmax_y$ denotes the operation that identifies the value of y maximizes the following expression, in this case it refers to the class y that demonstrates the greatest number of upvotes, k is the total number of trees in the model, and finally $I\left(H_j(x) = Y\right)$ refers to the function in charge of indicating that it is 1 in case the prediction of the $jth\ tree$ $(Hj(x))$ is equal to the class y. This equation is able to count the quantity of the number of trees that the class y predicts [14].

## 2.3 Evaluation metrics

### 2.3.1 Mean Absolute Error (MAE)

This algorithm measures the average error difference between the predicted values and the actual observed values and can be calculated with Equation 7:

$$MAE = \frac{\sum_{i=0}^{n}|y_i - x_i|}{n} \tag{7}$$

The algorithm makes a sum of the absolute error of each prediction from to, where i represents a variable that operates with all the elements of a set, n is the total number of observations indicating the number of predicted values, and real in are present in a data set [15].

### 2.3.2 Mean square error (MSE)

In the case of regression, the MSE seeks to measure the quality of a prediction model by defining the difference between the predicted values with respect to the values observed in the database.

In order to carry out its evaluation, Equation 8 is shown below.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \tag{8}$$

Where $n$ indicates the total number of observations, $1/n$ the normalized factor, the real observation value is $yi$, and the prediction value with respect to the observation as $\tilde{y}i$, the sum of the difference of these two types of values raised to the square [15].

### *2.3.3 $R^2$ square*

The $R^2$ squared, also known as the coefficient of determination; This algorithm can measure the number of dependent variables that can be applied to an independent variable. It defines your ability to fit a model to observed data; Equation 9 can denote the similarity level of the predicted data and the actual data [15]:

$$R^2 = 1 - \frac{\sum(y_i - x_i)^2}{\sum(y_i - \mu_y)} \tag{9}$$

## 2.4 Programming

For programming, Python software was used, which offers a wide variety of useful libraries and tools for data analysis and processing, as well as for applying machine learning algorithms. Among the libraries used for programming are Pandas, Numpy, Matplotlib, Seaborn, Sklearn.

The decision tree and random forest models were used to forecast solar radiation at 1 day and 1 week. The data set was divided, using 80% to train the models and 20% to validate them. These models are effective for short-term forecasts, as seen in the programming process in Figure 4.

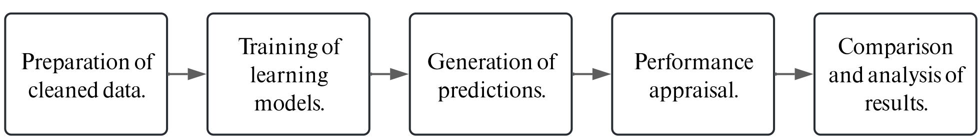



**Fig. 4.** Process of using machine learning models.

## 2.5 Responses from learning models

The purpose of the analysis of the response in question is to examine and contrast the capacity of the two machine learning models decision tree and random forest to predict the magnitude of solar radiation. To predict solar radiation, the procedure illustrated in Figure 5 was applied.

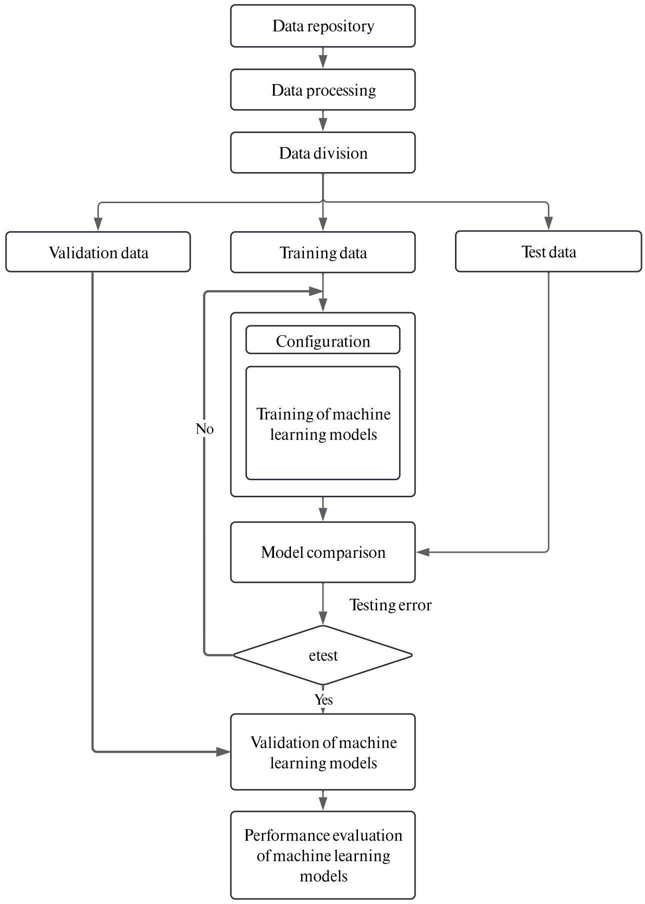



**Fig. 5.** Flowchart of the general methodology used for the pre-drying process.

Where, first of all, research was carried out on the existing literature to understand the state of the art in solar radiation prediction. Then, the historical database of meteorological variables from the University of Piura (UNP) was compiled, which underwent a preprocessing and selection process to ensure its quality. The decision tree and random forest machine learning models were implemented in Python, adjusting hyperparameters and additional meteorological variables to optimize their performance. The database was divided into training, testing and validation sets. To determine the prediction capacity of the models, evaluation metrics such as the mean square error (MSE), the mean absolute error (MAE) and the mean percentage error (MAPE) were used, which serve to measure the accuracy of the forecasts.

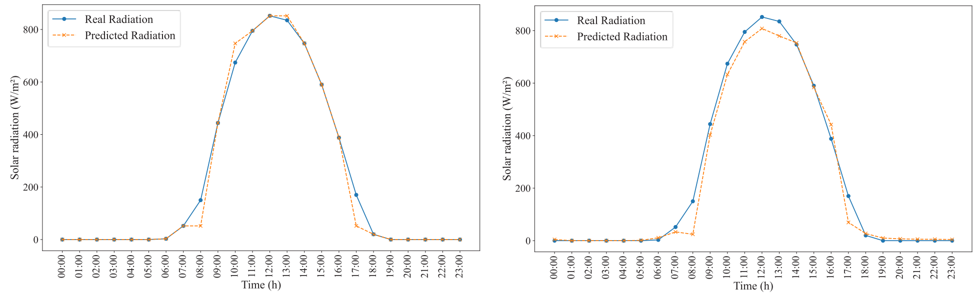
# 3 Results

This section shows the benefits derived from machine learning models (random forest and decision tree), in order to determine their ability to predict using error metrics and in this way, choose the most effective machine learning model and more suitable to be used in prediction.Machine learning models were used to forecast solar radiation. Open-source software (Python) was used in the study. The information was obtained from the UNP Conventional Meteorological Station and covered a year from April 2023 to April 2024, with evidence of hourly data that resulted in 8760 solar radiation records.

To confirm the machine learning models, a fraction of 20% of the data set was used, while 80% was used for training. Finally, evaluation metrics were used to compare the accuracy between these models, and as a result.

### 3.1 Prediction for one day

Next, Figure 6 displays the solar radiation levels on April 14, 2024, along with the predictions made by the decision tree model and the random forest model. The blue line displays how the database actually behaved, while the orange line displays the predicted solar radiation levels.

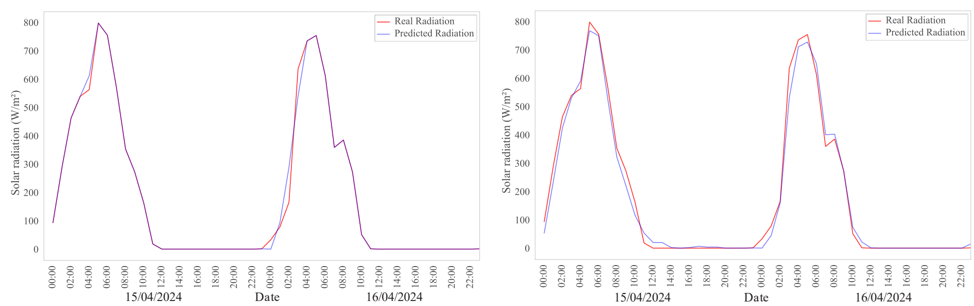



a) b)

**Fig. 6.** Solar prediction models for one day (14/04/2024): a) decision tree, b) random forest.

### 3.2 Two-day forecast

Figure 7 shows the behavior of solar radiation, both real and predicted by the two models, in a range of two days from April 15 to April 16, 2024. The blue contour represents the real behavior of the data, while the red outline reflects the solar radiation forecast for two days.

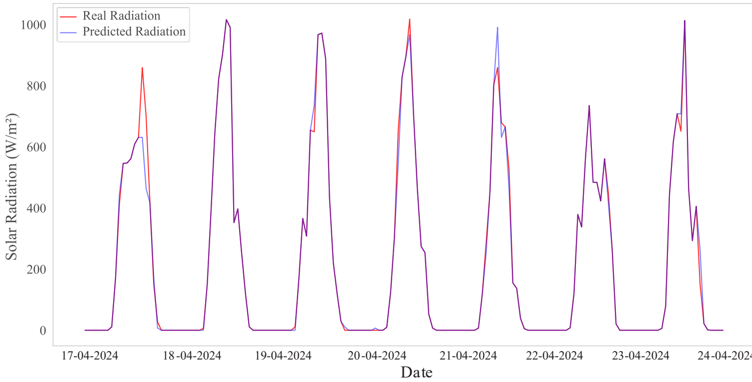



a) b)

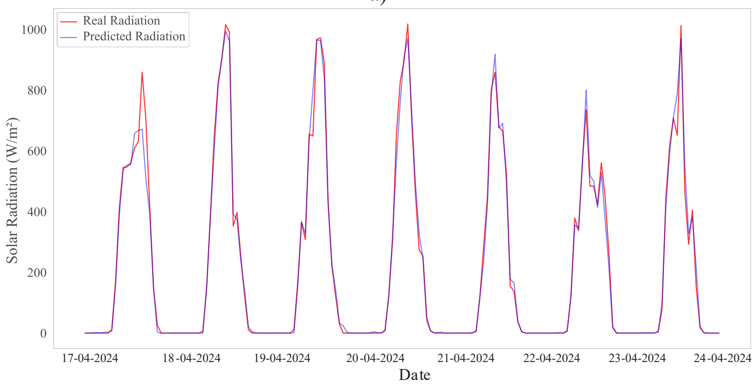
**Fig. 7.** Two-day solar prediction models: a) decision tree, b) random forest.

### 3.3 Seven-day forecast

Figure 8 shows how the real and predicted solar radiation acts within a week included in the date range from 17/04/2024 to 24/04/2024. The blue contour depicts the graph of the real data, while the red contour represents the forecast from the two learning models, decision trees and random forests, respectively.




a)




b)

**Fig. 8.** Seven-day solar prediction models: a) decision tree, b) random forest.

## 3.4 Analysis of machine learning model metrics

Table 2 presents the results of the error metrics for the comparison between the machine learning models for one, two and seven days, which are the mean absolute error (MAE), the mean square error (MSE) and the determination coefficient ($R^2$). These metrics are essential to evaluate the accuracy and effectiveness of the models.

**Table 2.** Comparison with error metrics between models.

| Model | Días | MAE | MSE | $R^2$ |
|---|---|---|---|---|
| Decision Tree | 1 | 12.85 | 1214.52 | 0.99 |
| Random Forest | | 24.28 | 1639.18 | 0.98 |
| Decision Tree | 2 | 31.70 | 2930.70 | 0.95 |
| Random Forest | | 22.67 | 1429.19 | 0.97 |
| Decision Tree | 7 | 38.21 | 5106.50 | 0.95 |
| Random Forest | | 36.96 | 4238.77 | 0.96 |

The results in Table 2 indicate that the random forest and decision tree models have appropriate performance for solar radiation prediction. In the one-day prediction, the decision tree model obtained the best performance, with a MAE of 12.85, an MSE of 1214.52 and an R2 of 0.99. On the other hand, for the two- and seven-day predictions, the random forest model outperformed the decision tree, with a MAE of 22.67 and 36.96, an MSE of 1429.19 and 4238.77, and an R2 of 0.97 and 0.96, respectively.

This indicates that the choice of machine learning model must consider the desired prediction horizon. For short-term predictions (one day), the decision tree is the appropriate option, while for medium- and long-term predictions (two and seven days), the random forest presents better performance. These results have important implications for the planning and management of solar and thermal energy systems by allowing better estimation of the availability of solar resources in the future.

## 4 Conclusions

The results of the study show that the use of techniques such as Monte Carlo simulation and the proper handling of incomplete data contribute to improving the accuracy of machine learning models.

In conclusion, the random forest machine learning model is presented as the most reliable option for predicting solar radiation in the medium and long term (two and seven days). Although for short-term forecasts (one day) the decision tree obtained slightly better results, the random forest demonstrated a precise fit to the data observed at distant horizons. This research highlights the importance of selecting the appropriate machine learning model based on the required prediction horizon, in order to optimize the estimation of the future availability of the solar resource, for the planning and management of solar and thermal energy systems.

## References

1. Segovia, J.A., Toaquiza, J.F., Llanos, J.R., Rivas, D.R.: Meteorological Variables Forecasting System Using Machine Learning and Open-Source Software. Electronics. 12, 1007 (2023). https://doi.org/10.3390/electronics12041007
2. Allal, Z., Noura, H.N., Chahine, K.: Machine Learning Algorithms for Solar Irradiance Prediction: A Recent Comparative Study. e-Prime - Advances in Electrical Engineering, Electronics and Energy. 7, 100453 (2024). https://doi.org/10.1016/j.prime.2024.100453
3. Ağbulut, Ü., Gürel, A.E., Biçen, Y.: Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison. Renewable and Sustainable Energy Reviews. 135, 110114 (2021). https://doi.org/10.1016/j.rser.2020.110114
4. Villegas-Mier, C.G., Rodriguez-Resendiz, J., Álvarez-Alvarado, J.M., Jiménez-Hernández, H., Odry, Á.: Optimized Random Forest for Solar Radiation Prediction Using Sunshine Hours. Micromachines. 13, 1406 (2022). https://doi.org/10.3390/mi13091406
5. Sagi, O., Rokach, L.: Explainable decision forest: Transforming a decision forest into an interpretable tree. Information Fusion. 61, 124–138 (2020). https://doi.org/10.1016/j.inffus.2020.03.013
6. Banik, R., Biswas, A.: Improving Solar PV Prediction Performance with RF-CatBoost Ensemble: A Robust and Complementary Approach. Renewable Energy Focus. 46, 207–221 (2023). https://doi.org/10.1016/j.ref.2023.06.009

7. Gao, Y., Miyata, S., Akashi, Y.: Interpretable deep learning models for hourly solar radiation prediction based on graph neural network and attention. Applied Energy. 321, 119288 (2022). https://doi.org/10.1016/j.apenergy.2022.119288
8. Alam, M.S., Al-Ismail, F.S., Hossain, M.S., Rahman, S.M.: Ensemble Machine-Learning Models for Accurate Prediction of Solar Irradiation in Bangladesh. Processes. 11, 908 (2023). https://doi.org/10.3390/pr11030908
9. Shanmugasundar, G., Vanitha, M., Čep, R., Kumar, V., Kalita, K., Ramachandran, M.: A Comparative Study of Linear, Random Forest and AdaBoost Regressions for Modeling Non-Traditional Machining. Processes. 9, 2015 (2021). https://doi.org/10.3390/pr9112015
10. Ordoñez Palacios, L.E., Bucheli Guerrero, V., Ordoñez, H.: Machine Learning for Solar Resource Assessment Using Satellite Images. Energies. 15, 3985 (2022). https://doi.org/10.3390/en15113985
11. Ehteram, M., Shabanian, H.: Unveiling the SALSTM-M5T model and its python implementation for precise solar radiation prediction. Energy Reports. 10, 3402–3417 (2023). https://doi.org/10.1016/j.egyr.2023.10.029
12. Jiménez J, D.L., Toapanta, J., Muñoz, P., Achig, R.S.: State-of-Health Assessment of Lithium-Ion Batteries in Two Scenarios: Microgrids and Electric Vehicles. In: 2023 IEEE Seventh Ecuador Technical Chapters Meeting (ECTM). pp. 1–6 (2023). https://doi.org/10.1109/ETCM58927.2023.10309053
13. Singh, N., Jena, S., Panigrahi, C.K.: A novel application of Decision Tree classifier in solar irradiance prediction. Materials Today: Proceedings. 58, 316–323 (2022). https://doi.org/10.1016/j.matpr.2022.02.198
14. Huang, H., Zhu, Q., Zhu, X., Zhang, J.: An Adaptive, Data-Driven Stacking Ensemble Learning Framework for the Short-Term Forecasting of Renewable Energy Generation. Energies. 16, 1963 (2023). https://doi.org/10.3390/en16041963
15. Achig, R.S., Gonzales, J.A., Hidalgo, C.A.: Flow Prediction for Hydropower Generation using LMST Neural Networks. J. Phys.: Conf. Ser. 2609, 012006 (2023). https://doi.org/10.1088/1742-6596/2609/1/012006