


Integration of satellite imagery and meteorological data to estimate solar radiation using machine learning models


Luis Eduardo Ordoñez Palacios

(Universidad del Valle, Santiago de Cali, Colombia)

 <https://orcid.org/0000-0001-5154-9472>, luis.ordonez.palacios@correounivalle.edu.co


Víctor Bucheli Guerrero

(Universidad del Valle, Santiago de Cali, Colombia)

 <https://orcid.org/0000-0002-0885-8699>, victor.bucheli@correounivalle.edu.co

Hugo Ordoñez

(Universidad del Cauca, Popayán, Colombia)

 <https://orcid.org/0000-0002-3465-5617>, hugoordonez@unicauca.edu.co

Abstract: Knowing the behavior of solar energy is imperative for its use in photovoltaic systems; moreover, the number of weather stations is insufficient. This study presents a method for the integration of solar resource data: images and datasets. For this purpose, variables are extracted from images obtained from the GOES-13 satellite and integrated with variables obtained from meteorological stations. Subsequently, this data integration was used to train solar radiation prediction models in three different scenarios with data from 2012 and 2017. The predictive ability of five regression methods was evaluated, of which, neural networks had the highest performance in the scenario that integrates the meteorological variables and features obtained from the images. The analysis was performed using four evaluation metrics in each year. In the 2012 dataset, an R^2 of 0.88 and an RMSE of 90.99 were obtained. On the other hand, in the 2017 dataset, an R^2 of 0.92 and an RMSE of 40.97 were achieved. The model integrating data improves performance by up to 4% in R^2 and up to 10 points less in the level of dispersion according to RMSE, with respect to models using separate data.

Keywords: GOES-13, Meteorological stations, Solar Radiation, Sunshine, Predictive model

Categories: H.3.1, H.3.2, H.3.3, H.3.7, H.5.1

DOI: 10.3897/jucs.98648

1 Introduction

The sun provides us with a source of energy that can be harnessed to produce electricity [Acciona, 21]; sunlight can be transformed into electric power using photovoltaic systems and the heat of the Sun; however, the Sun radiation that arrives at the surface of Earth undergoes a weakening process due to several dispersion, reflection and absorption factors, as can be observed in [IDEAM, 21]. Therefore, it is important to understand the solar power levels at a location to determine the size of systems that allow its utilization.

There are various types of instruments that measure solar radiation [Kipp & Zonen, 21] and some organizations interested in monitoring the meteorological conditions of a region have at their disposal a limited number of measurement stations to make

observations regarding the behavior of the characteristics that determine the weather. In Colombia, IDEAM is a government entity dependent on the Minister of Environment and Sustainable Development and oversees handling the scientific information related to the environment; likewise, there are other entities [Alcaldía de Santiago de Cali, 21], [Agroclima, 21] in the public and private spheres that also make these kinds of measurements.

Although there are various organizations that monitor weather conditions, the total number of weather stations [IDEAM, 20], is not sufficient to cover the Colombian territory, since they imply additional costs for maintenance and surveillance [Stel et al., 19]. Additionally, many stations do not capture radiation values due to the adverse conditions they are exposed to [Rodríguez Gómez, 19]. Therefore, several researchers have developed different physical, statistical, and artificial intelligence models to estimate solar radiation supported by terrestrial and satellite measurement instruments. Likewise, integrating data and models can enhance algorithm performance and yield more dependable estimation outcomes.

According to the research of [Suárez Vargas, 13], physical and statistical models are based on an energy balance between the radiation reaching the top of the atmosphere and the radiation reflected by the satellite. On the other hand, physical models use parameters of absorption, spreading, cloud albedo and superficial albedo, although the difficulty of these models lies in knowing these atmospheric values at a local level, according to the research of [Zarzalejo et al., 06]. In the case of statistical models, regressions are used between the radiometric measurements on the surface and the information recorded by the satellite, as observed in the work of [Poveda Matallana, 20].

Many studies have used satellite images to calculate solar radiation; the work of [Doncel Ballén, 11] used the Heliostat 1 algorithm to estimate solar irradiance using images from the GOES satellite over the Cundiboyacense region in Colombia; in a similar way, the research of [Albarelo et al., 15] used a modification of the Heliosat-II method, developed to process images from the Meteostat satellite, for its use with images from the GOES satellite in solar radiation estimation over the French Guyana; in the same way, the work of [Pagola et al., 14] implemented a combination between the Heliostat 1 and Heliostat 2 methods to estimate solar radiation in Spain, using images coming from the second generation Meteostat satellite. The Heliostat method has also been used to estimate solar radiation in research developed by [Hammer et al., 01], [Hammer et al., 03], [Kallio-Myers et al., 20], [Lorenz et al., 12], [Lorenz et al., 04], [Rigollier et al., 04] and [Zarzalejo et al., 06].

Other researchers have used the Angström-Prescott method to calculate solar power using radiometric stations. According to the [Prescott, 40], the model relates the monthly average with the solar radiation on a clear day using the hours of daily sunlight. Research by [Poveda Matallana, 20] validated the solar radiation on the surface over Orinoquia from images obtained from the GOES satellite. The research indicated that the Angström-Prescott coefficients depend on geographic and climatic parameters and on dynamic, spatial and physical properties of the atmosphere, which explains the need for radiation data and sunshine obtained by measurement stations on earth. In the same way, the work of [Guzman M. et al., 21] utilized the Angström-Prescott coefficients to estimate global solar radiation using sunshine data in the coffee zone in Colombia.

In the research of [Nwokolo et al., 22], predictions of global solar radiation potential were made using probabilistic methods. Twenty-nine Angström-Prescott (AP)

empirical models were analyzed. The M1-M13 models were fitted using generalized datasets by altering the original model. Models M14-M20 were taken from the literature. Models M21-M29 used datasets from meteorological stations in Nigeria. Model M13 achieved the highest performance with an RMSE of 0.0001, an rRMSE of 0.0176%, a maximum R^2 value of 0.990 and a global performance indicator GPI of 0.9321.

[Geetha et al., 22], performed solar radiation estimates in India using neural network methods. The best neural network model was adjusted with 12 hidden layers and showed an R^2 of 0.9340. [Oyewola et al., 22] demonstrated that air temperature and humidity improve solar power predictions. Twenty (20) models were adopted based on historical data captured over 35 years (1984-2018) at 6 monitoring stations in the Fiji Islands. The MD17 model achieved the one with the highest performance with an R^2 0.988.

In this review, we also found studies that used artificial intelligence techniques to evaluate solar resources using historical climatic data or images provided by geostationary satellites. The studies of [Eissa et al., 13], [Hammer et al., 01], [Linares-Rodriguez et al., 13], [Linares-Rodriguez et al., 15], [Martín Pomares et al., 06], [Mazorra Aguiar et al., 15], [Ordoñez-Palacios et al., 20], [Gürel et al., 20], [Jumin et al., 21] and [Ağbulut et al., 21], used machine learning algorithms. On the other hand, the studies of [Alzahran et al., 17], [Jiang et al., 19], [Jiang et al., 20], [Kaba et al., 18], and [Chandola et al., 20] utilized deep learning techniques.

The work of [Ordoñez Palacios et al., 22] uses machine learning algorithms to evaluate the solar resource from satellite images. It compares the results obtained from models trained with datasets from different altitudes in Colombia. The best result was achieved with Random Forest in the M1-M5 models using 100% of the data. An R^2 of 0.82 and an RMSE of 107.05 were obtained. In contrast, this research achieves superior performance in solar radiation prediction because it integrates two dimensions of information. The neural networks achieved a performance of 0.92 in R^2 and 40.97 in RMSE.

This research uses a mixed model. It uses a mathematical method to extract features from satellite imagery, then integrates them with meteorological variables and implements a machine learning model to estimate solar radiation. 1447x1636 pixel images taken from the GOES-13 meteorological satellite from 2012 and 2017 were processed [US Department of Commerce, 21] using Python libraries (Rasterio, Pyproj) and mathematical equations. It is important to highlight that (i) a model was built based on the empirical Ångström-PreScott method for calculating solar radiation from historical solar brightness data; (ii) a solar radiation prediction scenario was proposed based on variables extracted from the images; (iii) a solar radiation prediction scenario was constructed using meteorological variables; (iv) a scenario was implemented to estimate solar radiation incorporating two dimensions of information: variables extracted from the images and meteorological variables.

This paper contains the sources of information and methodology used, the results of the research, discussion of the results, conclusions, and future work.

2 Materials and methods

The following are the questions that guided the research, the sources of information, the data processing and the models used to evaluate the integration of data for solar radiation prediction.

2.1 Questions of interest

Solar power is a renewable resource that can be transformed into electric power using photovoltaic systems; therefore, it is fundamental to understand its behavior regarding the radiation levels reaching the Earth, either by measurement instruments, satellite images or artificial intelligence predictive models.

Considering that the number of measurement stations in Colombia is insufficient and that they are exposed to constant risks, such as to crime issues, geographic location, and adverse climatic conditions, it is necessary to evaluate existing methods for estimation; therefore, questions such as the following are important: what is the process to calculate solar radiation from images and what other variables are obtained? What techniques of machine learning can be utilized to predict solar radiation, and which have a better performance? What are the results obtained by utilizing meteorological variables extracted from the images or the integration of both? These questions will be answered throughout this paper.

2.2 Information sources

For this research, historical images of the visible channel were utilized, obtained from the GOES-13 meteorological satellite during 2012 and 2017; additionally, two sets of data from air quality stations (Republic of Argentina School –ERA– and Compartir) of the Administrative Department of Environment Management from the Mayorality of Cali and a dataset of daily sunshine from the Univalle station provided by IDEAM were obtained.

The images were captured by the satellite every hour and a half, and for this research, the pictures obtained from 6 am to 6 pm were utilized; however, it is important to highlight that only a few days throughout the year account for all the images; on the other hand, both sets of air quality data (ERA and Compartir) include historical observations of wind speed and direction, temperature, rain, humidity and solar radiation recorded every hour. Tables 1 and 2 show the metadata of the images and datasets utilized.

ID	Year	Images	Size	Total
1	2012	1991 out of 4758	9.6 MB	19.11 GB
2	2017	2135 out of 4745	9.6 MB	20.50 GB

Table 1: GOES-13 Satellite images

Station	ERA	Univalle	Compartir
Latitude	3.44779	3.3780	3.4282312578
Longitude	-76.51918	-76.53388889	-76.4665448467
Year	2012	2012	2017

Registers	8634 out of 8784	329 out of 366	8166 out of 8760
Interval	Hourly	Daily	Hourly
Porcentaje	98.3%	89.9%	93.2%

Table 2: Datasets from the air quality stations

The images were taken by the government agency for Oceanic and Atmospheric Administration (NOAA) [NOAA Class, 21]. In this research, a web tracker was developed for the automatic download of satellite images, files with a goal extension were filtered, and images from channels other than the visible channel. The images obtained from the satellite were processed with the software Weather and Climate Tools from the NOAA (WTC) [NOAA, 21] to visualize and export the data to the NetCDF format, which was used for the model developed in the programming language Python to calculate solar irradiance using the Angström-PreScott method.

Estimation of solar radiation (H) of a geographic location using the Angström-PreScott model requires calculation of the coefficients a and b , the cloudiness index (n_c) and solar radiation at the boundary of the earth's atmosphere (H_{ext}), according to research by [Poveda Matallana, 20]. The coefficients are obtained from the relationship between the number of hours of sunlight and global solar radiation, with both parameters captured on the surface, according to [Guzman M., 21]. In this sense, to obtain the cloudiness index, first, it is necessary to transform the images to the TIFF format using the libraries Rasterio and PyProj from Python. The research by [Ordoñez Palacios et al., 22] presents the process to obtaining the variables involved in the calculation of the cloudiness index (n_c).

Estimating solar radiation from satellite imagery requires solar radiation and brightness data captured by ground-based measurement stations; however, the air quality datasets from the DAGMA (ERA and Compartir) contain only solar radiation. Consequently, sunshine data was requested from a nearby IDEAM station. In this case, they provided data from the Univalle station, located 9.3 km from the Compartir station, and 7.8 km from the ERA station. Later, the sunshine data were integrated into the ERA station, although it was not possible to do the same for the data from the Compartir station, since IDEAM does not have data on sunshine for 2017.

2.3 Methodology

This work is based on the CRISP-DM methodology. Some manuscripts such as that of [Solano et al., 22], [Plotnikova et al., 22] and [Huber et al., 19], are also based on the same methodology for building data mining models. The process for extracting the satellite image features, data processing and hyperparameter fitting can be seen in [Ordoñez Palacios et al., 22]. The empirical Angström-PreScott method was used to obtain solar radiation from solar brightness data. Then, three scenarios were built to predict irradiance based on meteorological data, data extracted from satellite images and the integration of the variables from both cases.

Five regression methods were developed and applied to each of the proposed scenarios. The regression models were developed in the Python programming language and their performance was evaluated using four metrics described below. The analysis of the results obtained in the research can be seen in the discussion and conclusions.

The MinMaxScaler method was used to regulate the data between 1 and 0. Regression algorithms such as multiple linear regression, artificial neural networks, and the ensemble methods XGBoost for regression, gradient boosting regressor and random forest regressor were used.

2.3.1 Linear Regression

This algorithm fits a linear model with coefficients to minimize the residual sum of squares between the targets observed in the data set and the targets predicted by the linear approximation [Scikit-learn, 07]).

2.3.2 Gradient Boosting for regression

Is a estimator builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage a regression tree is fit on the negative gradient of the given loss function [Scikit-learn, 07]).

2.3.3 Random Forest regressor

Is a meta estimator that fits a number of classifying decision trees on various subsamples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting [Scikit-learn, 07]).

2.3.4 Neural networks

In terms of neural networks, the Multi-layer Perceptron (MLP) is a supervised learning algorithm that learns a function $f(\cdot): R^m \rightarrow R^o$ by training on a dataset, where m is the number of dimensions for input and o is the number of dimensions for output. Given a set of features $X = x_1, x_2, \dots, x_m$ and a target y , it can learn a non-linear function approximator for either classification or regression [Scikit-learn, 07]).

2.3.5 XGBoost for regression

According to the documentation, XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way [XGBoost, 22].

2.3.6 Hyperparameter fitting

In the matter of neural networks, the Multilayer Perceptron was used and tested with the 2012 data, about 4000 different configurations in 7 hyperparameters, establishing a search space of 3 to 7 different values in each hyperparameter as shown in table 3. In this work we have used the hyperparameter fitting method based on the random search of the scikit-learn library. The hyperparameter search space, the description taken form [Scikit-learn, 07]), and best estimator values used in the artificial neural networks are presented in Table 3.

ID	Hyper-parameter	Description	Search space	Found value
1	Optimization algorithm	Is used to fit a neural network model to a training data set.	["lbfgs", "adam", "sgd"]	Adam
2	Number of hidden layers	Refers to the number of hidden layers that represent higher level characteristics or attributes of the data.	[[125, 100, 75, 50, 10], [100, 75, 50, 10], [100, 75, 50], [100, 75]]	[125, 100, 75, 50, 10]
3	L2 regularization term	Allows you to reduce the value of parameters to make them small.	loguniform(1e-5, 1e-3)	0.000146232
4	Activation function	It transmits the information generated by the linear combination of weights and inputs through the output connections.	["identity", "logistic", "tanh", "relu"]	Relu
5	Learning rate	It is a value que affects the speed at which the algorithm reaches (converges to) the optimal weights.	["constant", "adaptive", "invscaling"]	Invscaling
6	Maximum number of iterations	Number of iterations of the optimization algorithm up to convergence.	[200, 1000, 5000, 10000]	1000
7	Maximum number of epochs	For stochastic optimization algorithms, determine the number of times each data point will be used, not the number of gradient steps.	[5, 10, 15, 20, 25, 30, 40]	30

Table 3: Search space and values of the best model

2.3.7 Description of variables

Three regression models were built to evaluate the integration of data for solar radiation prediction: the first (M1) used meteorological variables, the second (M2) only included the features extracted from the satellite images and the third (M3) integrated both groups of variables. Table 4 describes each of the variables used by the different models.

ID	Source	Variable	Description
1	Measurement station	Wind speed	It refers to the displacement of air at a point and at a given moment; it is measured in meters per second (m/s).
2		Wind direction	Indicates the direction in degrees (0-360) from where the wind is coming.
3		Temperature	It is related to the notion of heat in the atmosphere and is measured in Celsius degrees.
4		Rain	It is defined as the amount of water that falls per unit of time in each place, measured in millimeters (mm).
5		Humidity	It refers to the vapor present in the atmosphere.
6		Solar Radiation (Target Variable)	Energy received from the sun by electromagnetic waves, measured in W/m ² .
7	Satellite images	Reflectance	It corresponds to the value of the solar radiation that is reflected by the clouds.
8		Cloudiness index	It is a value related to the cloudiness conditions of clear, partly cloudy, and cloudy skies (0-1).
9		Solar radiation at the edge of the atmosphere	It is the value of the electromagnetic radiation emitted by the Sun, before entering the atmosphere, measured in W/m ² .
10		Number of hours of solar brightness	It is the time in hours during which the sun has an effective solar brightness.

Table 4: Description of variables

2.3.8 Evaluation Metrics

The predictive ability of the regression methods was calculated using the metrics MBE, R², RMSE and rRMSE.

The MBE metric (1) provides insight into the long-term performance of models [Manju, 19]. The closer the MBE value to zero, the better the estimation result [Fan et al., 18].

$$MBE = \frac{1}{n} \sum_{i=1}^n (y_i - x_i) \quad (1)$$

The R² metric (2) is used to determine how well the regression line approximates the actual data points [Gouda et al., 19]. R² value changes between 0 and 1, and the closer this value is to 1, the better the performance of the model.

$$R^2 = 1 - \frac{\sum(y_i - x_i)^2}{\sum(x_i - \bar{x}_i)^2} \quad (2)$$

The RMSE metric (3) represents the difference between the estimated and observed values [Fan et al., 19]. RMSE takes positive values and the closer the RMSE value to zero, the better the estimation result [Bakay and Ağbulut, 21]

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - x_i)^2} \quad (3)$$

The rRMSE metric (4) provides a percentage value, the result of dividing the RMSE by the mean of the real values. An rRMSE value close to zero explains that the models perform better. The success of algorithms according to this metric are ranked according to the research of [Ağbulut et al., 21], [Fan et al., 19] and [Bakay and Ağbulut, 21].

$$rRMSE = \frac{\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - x_i)^2}}{\bar{x}_i} \times 100 \quad (4)$$

2.4 Model architecture

Figure 1 explains the data flow from satellite and ground-based data sources to solar radiation predictions using regression and hyperparameter fitting techniques.

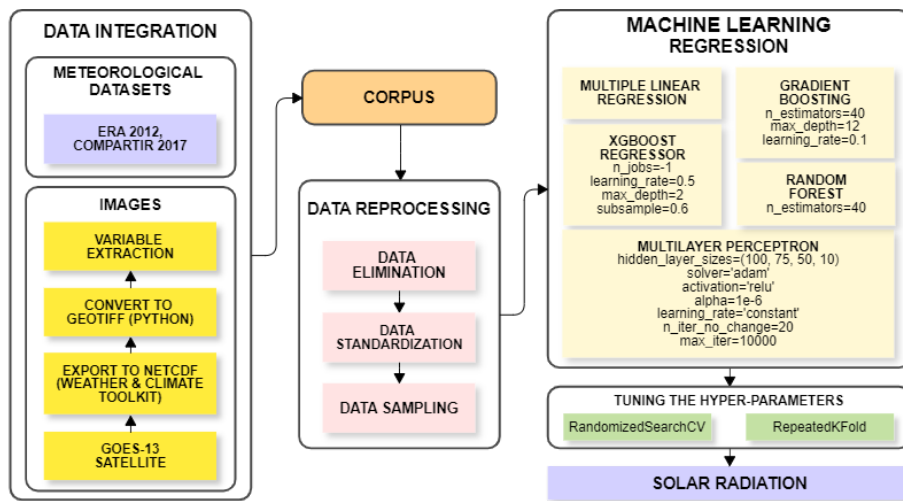


Figure 1: Model architecture. Source: Own elaboration

3 Results

The empirical Angström-Prescott method uses monthly coefficients a and b obtained from the 2012 dataset to calculate solar radiation using regression and considering the relationship between solar radiation and solar brightness captured on the ground

(Univalle dataset). For 2017, it was not possible to estimate solar radiation because of the lack of historical solar brightness data. Table 5 presents the values of the coefficients a and b and the performance of the regression to calculate the monthly solar radiation according to the determination coefficient R^2 .

Month	a	b	R^2
1	0.207	0.419	0.758
2	0.222	0.393	0.648
3	0.236	0.410	0.740
4	0.245	0.375	0.603
5	0.274	0.319	0.611
6	0.314	0.297	0.540
7	0.282	0.366	0.647
8	0.300	0.335	0.712
9	0.323	0.335	0.795
10	0.226	0.416	0.550
11	0.263	0.339	0.521
12	0.262	0.311	0.623

Table 5: Results obtained by the empirical Angström-Prescott method for the 2012 dataset

Figure 1 shows the scatter diagram obtained from daily solar radiation captured in 2012 by the measurement station and estimated by the empirical Angström-Prescott method with an R^2 of 0.552 and an RMSE of 1062.71. The experiment was also conducted with hourly solar radiation data; however, the results showed a much lower performance.

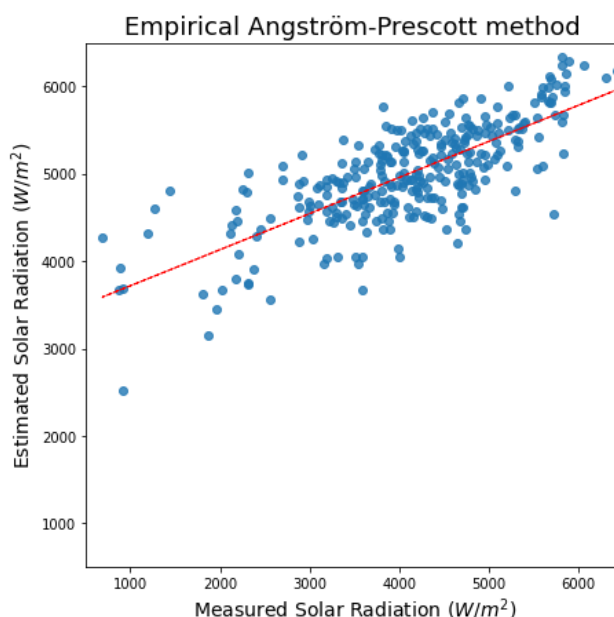


Figure 2: Daily solar radiation observed by the station and calculated from satellite images. Source: Own elaboration

The dataset from the ERA station for 2012 was integrated into the dataset generated by processing the images of the GOES-13 satellite taken in 2012; in the same way, the dataset of the Compartir station for 2017 was integrated into the dataset of the images of the GOES-13 satellite taken in 2017. In both datasets, records were eliminated due to the nonexistence of meteorological data of the dates and times in which data from the images did exist. Table 6 presents the characteristics of the datasets used in the regression models.

ID	Datasets	Registers	Eliminated	Total
1	2012	1991	28	1963
2	2017	2135	110	2025

Table 6: Datasets used in the regression

Tables 7 and 8 show the performance of the regression algorithms applied to the 2012 and 2017 datasets. In both cases, neural networks had the highest performance; likewise, the model that integrates the meteorological variables and the variables obtained from the images had a higher performance than the other two models. It is important to point out that the machine learning algorithms performed significantly better than the empirical model.

The ensemble methods (gradient boosting, XGBoost regressor and random forest) with the default hyperparameters, showed results very close to 1 in training, according to the determination coefficient R^2 , which indicated an overfitting of the algorithms.

Therefore, cross-validation was used as a preventive measure against overtraining, dividing the data into 5 subsets and training the model iteratively. The results presented are the average of the values obtained in each data subset; the regularization technique was also used to artificially force the algorithm to be simpler.

The adjustment of the hyperparameters of each model was performed by random search of the scikit-learn library. A search space of between 6 and 10 different values was defined in the hyperparameters: `n_estimators`, `max_depth`, `learning_rate`, `min_samples_leaf`. In the case of the `learning_rate` parameter, a range of values between: $1e-3$ and $1e-1$ was established using the loguniform function of the python `scipy` library.

ID	SCENARIO	METRICS	MULTIPLE LINEAR REGRESSION	GRADIENT BOOSTING	XGBOOST REGRESSOR	RANDOM FOREST	NEURAL NETWORKS
M1	METEOROLOGICAL VARIABLES	MBE	9.467	-0.269	0.171	0.215	-0.808
		R ² ENT	0.564	0.868	0.863	0.857	0.858
		R ² PRU	0.559	0.824	0.817	0.820	0.849
		RMSE	169.49	116.10	106.91	103.44	98.68
		rRMSE	51.14	33.12	31.43	30.97	29.78
M2	VARIABLES OBTAINED FROM THE IMAGES	MBE	-0.422	-0.249	-1.541	-0.886	1.813
		R ² ENT	0.374	0.852	0.853	0.842	0.836
		R ² PRU	0.432	0.810	0.800	0.807	0.836
		RMSE	192.26	110.97	106.74	108.47	103.27
		rRMSE	58.01	32.56	31.79	31.58	31.16
M3	METEOROLOGICAL VARIABLES + VARIABLES OBTAINED FROM THE IMAGES	MBE	-0.248	0.993	-2.968	-2.909	-0.051
		R ² ENT	0.669	0.904	0.899	0.884	0.909
		R ² PRU	0.685	0.860	0.848	0.846	0.880
		RMSE	143.19	96.79	94.87	98.10	90.99
		rRMSE	43.21	28.36	27.23	26.78	26.70

Table 7: Results obtained from the 2012 dataset

According to Table 7, the neural network algorithm had the best performance in most evaluation metrics. In the case of MBE, the M3 model had the value closest to zero (-0.051) of all models used in the research, followed by the M1 model (0.171). Positive MBE values indicated that the average of the results predicted by the models was greater than the average of the actual observations.

Moreover, the M3 model had a higher performance than the M1 model. In regard to training, the model was better by 5%, and in tests, it was also better by 3% according to the determination coefficient R^2 ; in addition, the model had 8% a smaller number of errors according to the RMSE. The M2 model had the lowest performance of the 3 models, although it did not require surface-observed data. Considering the relative mean square error (rRMSE), the neural networks also had the highest performance; however, the estimated results in the M2 model were higher than 30%, which indicated a poor performance, and in the case of the M1 and M3 models, the forecast results were

between 20% and 30%, which represented a regular performance of the regression method.

Figure 2 shows the scatter diagrams obtained with the multilayer perceptron technique for each model, using the actual solar power captured by the station and the estimated solar power by the models for 2012. At first glance, the differences between the three images are not very evident; however, the M3 model fits the data better and represents less variability around the mean.

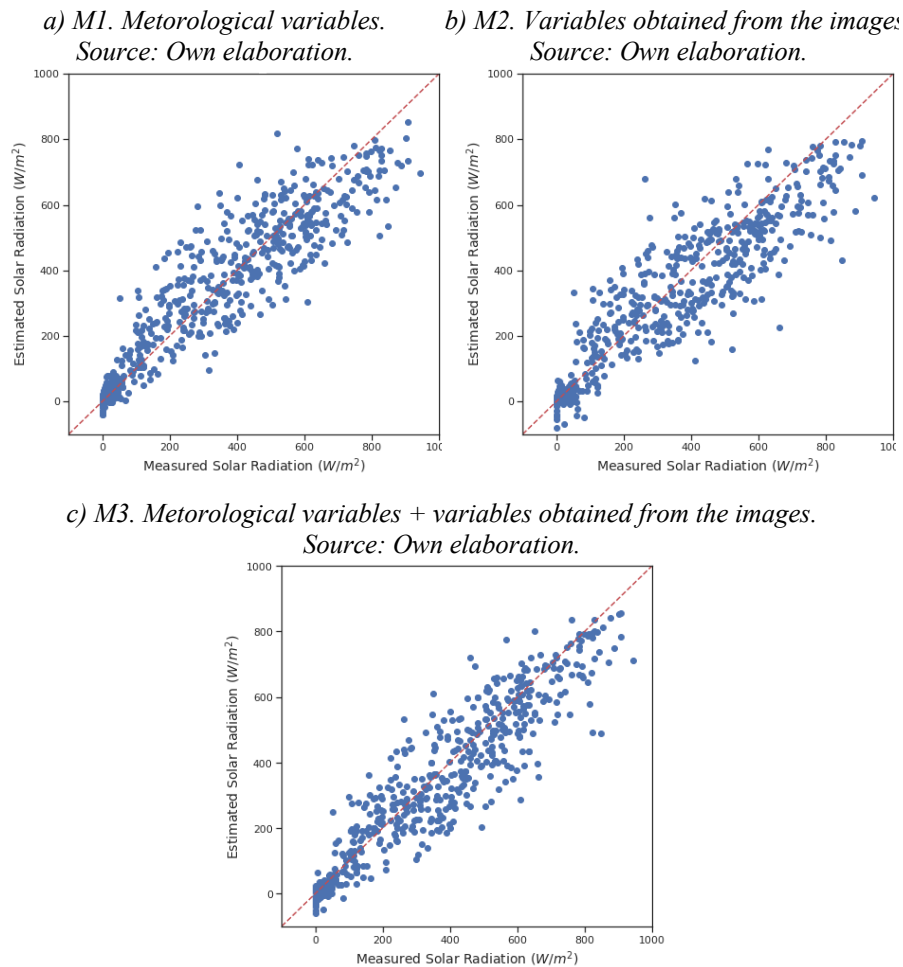


Figure 3: Real and estimated solar radiation in 2012 by using the multilayer perceptron

According to the results in Table 8, the neural network algorithm had the best performance in most evaluation metrics. In the case of MBE, the M3 model had the closest value to zero (-0.146) of all models used in the research, followed by the M1 model (-0.149). In this case, the MBE value was negative in both cases, which indicated

that the average of the results estimated by the models was less than the average of the actual observations.

Moreover, the M3 model also had a higher performance than the M1 model; in training, the model was better by 4%; and in tests, it was also better by 3% according to the determination coefficient R^2 , which indicated that the model increased the level in the explanation of the variability of the data around the mean and that it fit the data better. In addition, the model had 6% less error between the real and estimated datasets, according to the RMSE. The M2 model had the lowest performance of the 3 models, although it did not require surface-observed data.

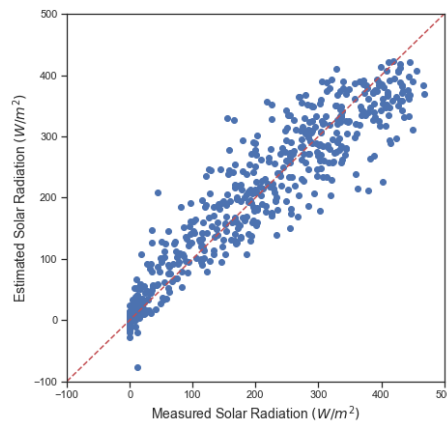
Analyzing the rRMSE metric, the neural networks also had the highest performance, and the results estimated by the three models were between 20% and 30%, which represented a regular performance of the algorithm. The M3 model had the best performance in predictions.

ID	SCENARIO	METRICS	MULTIPLE LINEAR REGRESSION	GRADIENT BOOSTING	XGBOOST REGRESSOR	RANDOM FOREST	NEURAL NETWORKS
M1	METEOROLOGICAL VARIABLES	MBE	-8.345	-0.452	1.506	-0.149	1.687
		R ² ENT	0.581	0.901	0.900	0.893	0.882
		R ² PRU	0.611	0.870	0.867	0.866	0.890
		RMSE	88.48	49.03	48.19	47.20	46.97
		rRMSE	48.17	28.11	26.48	27.01	25.57
M2	VARIABLES OBTAINED FROM THE IMAGES	MBE	-11.715	1.856	-3.739	0.992	-0.748
		R ² ENT	0.349	0.900	0.901	0.889	0.895
		R ² PRU	0.325	0.872	0.868	0.863	0.881
		RMSE	116.58	46.73	46.28	46.36	49.04
		rRMSE	63.46	26.67	25.44	26.22	26.70
M3	METEOROLOGICAL VARIABLES + VARIABLES OBTAINED FROM THE IMAGES	MBE	-8.994	-1.076	-0.298	0.580	-0.146
		R ² ENT	0.657	0.933	0.925	0.916	0.927
		R ² PRU	0.691	0.902	0.891	0.887	0.917
		RMSE	78.85	44.10	47.95	42.59	40.97
		rRMSE	42.92	24.73	25.74	24.92	22.30

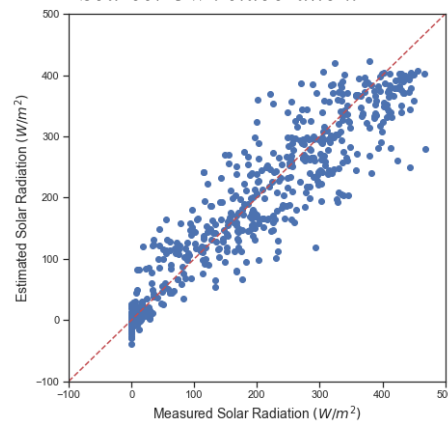
Table 8: Results obtained from the 2017 dataset

Figure 3 represents the scatter diagrams obtained with the multilayer perceptron technique for each model, using the real solar radiation captured by the station and the estimated solar radiation captured by the models for 2017. Although visually, the differences between the images that represent each model are not very evident, the M3 model fits the data better and represents less variability around the average.

a) M1. Meteorological variables.
Source: Own elaboration.



b) M2. Variables obtained from the images.
Source: Own elaboration.



c) M3. Meteorological variables + variables obtained from the images.
Source: Own elaboration.

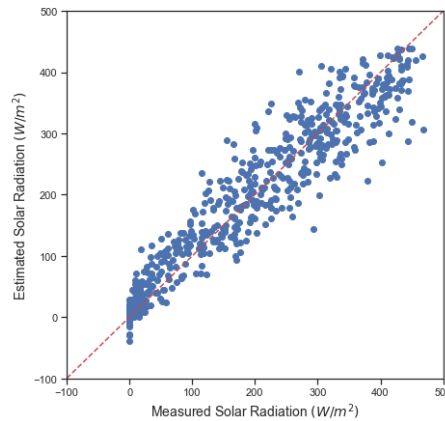


Figure 4: Real and estimated solar radiation in 2017 using the multilayer perceptron

4 Conclusions

In this manuscript, we evaluate the predictive ability of five regression algorithms for predicting solar radiation in Colombia. The research considers the characteristics obtained from the images taken by the GOES-13 satellite in 2012 and 2017, two datasets provided by air quality measurement stations (Escuela República de Argentina -ERA- and Compartir) from the Mayoralty of Cali and a dataset of daily solar brightness from the Univalle station, supplied by IDEAM.

The time required for the request, download, georeferencing and processing of the images to obtain the reflectance of the pixels depends on the internet connection, the availability of the NOAA Class server and the characteristics of the computer

equipment. On the other hand, it was not possible to evaluate the solar power in the 2017 empirical Angström-Prescott model due to the lack of solar brightness data at the Univalle station.

Considering the correlation values in the regression models, the temperature variable, followed by the reflectance variable, had a greater correspondence with solar radiation; similarly, the artificial neural network called the multilayer perceptron, had the best performance in solar radiation predictions compared with the other regression algorithms based on R^2 and RMSE. Furthermore, the model that integrated the meteorological variables and the variables obtained from the satellite images produced the best results in solar radiation estimations in comparison with the empirical model and the other learning models.

The M2 model had a lower performance than the M1 and M3 prediction models, however, it provided better results than the empirical model. It also allows the prediction of solar power at any geographic location on the planet using only data obtained from the processing of the images recovered from the GOES-13 satellite. Considering the results achieved in this work, we propose acquiring satellite images and recent solar radiation data for at least three consecutive years to use time series to make solar radiation predictions in the future.

5 Future Work

Researches on solar radiation predictions occurs in many cases due to an insufficient number of monitoring stations; additionally, because some of them do not provide instruments to measure it, some methods that estimate solar power from satellite images require radiation and solar brightness data observed on the ground and depend on different geographic and climatic parameters and some atmospheric properties, as observed in the work by [Poveda Matallana, 20].

This work contributes to the solution of the global problem of electricity production associated with the emission of greenhouse gasses generated by the burning of fossil fuels, according to the study by [Bakay and Ağbulut, 21]. Its importance lies in supporting decision makers in the installation of solar farms in remote places. It provides data on the behavior of solar radiation using artificial intelligence techniques based on characteristics obtained from satellite images. This work uses the empirical Angström-Prescott method to extract features from satellite images and integrate them with meteorological data to evaluate solar resources at a given location.

The data acquired from satellite images allow the calculation of variables that directly affect the solar power that reaches the Earth's surface; among others, the cloudiness index is a parameter that depends on the minimum and maximum reflectance of each hour of the day. Reflectance values can be above 1, which is why, according to the research by [Laguarda et al, 18], only 80% of the maximum reflectance should be used without compromising the performance of the model. Even so, the values of this index must be between 0 and 1. Thus, they must be adjusted in the case of exceeding the limits of the domain.

The low performance of the 2012 empirical Angström-Prescott method, in part, was due to the lack of images provided by the satellite. Only 41.8% of the total images were obtained (1991 out of 4758 possible). In addition, between 6 am and 6 pm, only 7.9% of the days of the year had 13 images; in contrast, 66.4% of the days of the year

had 4 images or less, and finally, the parameters observed by the DAGMA air quality stations (ERA and Compartir) did not include the solar brightness variable. Therefore, data from a nearby IDEAM station, the Univalle station was used, which provided only 89.9% of the daily solar brightness data.

The optimized Angström-Prescott model OPM3, built in the work of [Nwokolo et al., 22] outperforms with an R^2 of 0.998 the best result of our Angström-Prescott model which achieved an R^2 of 0.795 as shown in Table 5. However, our model performs better than 20 of the basic models (M1-M20) and 3 of the improved models (OPM14, OPM17 and OPM19) developed in the study. Consequently, the neural networks with the 2017 data, improve the performance of our Angström-Prescott method with an R^2 of 0.917 being surpassed only by the result of the optimized models OPM3 and OPM20 from the study in comparison.

In this research, the neural networks performed best in the scenario that integrates the variables extracted from the images and the meteorological variables. An R^2 of 0.880 was obtained with the 2012 data, while an R^2 of 0.917 was obtained with the 2017 data. This is because for 2017 there were 7.23% more images than in 2012. It is important to highlight that missing image at one hour of the day leads to the exclusion of meteorological records for that same hour.

Although the best models from the research of [Geetha et al., 22] with an R^2 of 0.9340 and [O. M. Oyewola et al, 22] with an R^2 of 0.988, outperform our best-case scenario M3, which achieved an R^2 of 0.917. It is critical to recognize that the error difference of approximately 7% does not represent significant costs in terms of what might occur if an estimate made by our model is wrong. If this error is translated to the context of a microgrid, the worst that could happen is that a backup battery or diesel generator could be used to support the average consumption load.

This research generated the work of [Ordoñez Palacios et al., 22] as an extension of the M2 model to evaluate the solar resource. The study tested seven models that include data from different altitudes above sea level. The results showed a slight trend in the evaluation of the solar resource where the altitude and the performance of the models is inversely proportional. Unlike the present study, the Random Forest algorithm stood out for achieving the best results with a value of 0.82 in R^2 , compared to a value of 0.81 obtained with the same technique in the M2 model of this work.

References

- [Acciona, 21] Acciona, «¿Qué beneficios tiene la energía solar?», 2021. <https://www.acciona.com/es/energias-renovables/energia-solar/>
- [Ağbulut, 21] Ü. Ağbulut, A. E. Gürel, and Y. Biçen, «Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison», *Renewable and Sustainable Energy Reviews*, vol. 135, p. 110114, Jan. 2021, doi: 10.1016/j.rser.2020.110114.
- [Agroclima, 21] Agroclima, «Servicio Meteorológico de la Federación Nacional de Cafeteros de Colombia-FNC», 2021. <https://agroclima.cenicafe.org/>
- [Albarelo, 15] T. Albarelo, I. Marie-Joseph, A. Primerose, F. Seyler, L. Wald, and L. Linguet, «Optimizing the Heliosat-II Method for Surface Solar Irradiation Estimation with GOES Images», jul. 2015.

- [Alcaldía de Santiago de Cali, 21] Alcaldía de Santiago de Cali, «Departamento Administrativo de Gestión del Medio Ambiente», 2021. www.cali.gov.co/dagma
- [Alzahrán, 17] A. Alzahrán, P. Shamsia, C. Daglib, and M. Ferdowsia, «Solar Irradiance Forecasting Using Deep Neural Networks», *Procedia Computer Science*, vol. 114, pp. 304-313, Jan. 2017, doi: 10.1016/j.procs.2017.09.045.
- [Bakay, 21] M. S. Bakay and Ü. Ağbulut, «Electricity production based forecasting of greenhouse gas emissions in Turkey with deep learning, support vector machine and artificial neural network algorithms», *Journal of Cleaner Production*, vol. 285, p. 125324, feb. 2021, doi: 10.1016/j.jclepro.2020.125324.
- [Chandola, 20] D. Chandola, H. Gupta, V. A. Tikkiwal, and M. K. Bohra, «Multi-step ahead forecasting of global solar radiation for arid zones using deep learning», *Procedia Computer Science*, vol. 167, pp. 626-635, Jan. 2020, doi: 10.1016/j.procs.2020.03.329.
- [Doncel Ballén, 11] D. Doncel Ballén, «Estimación de irradiancia solar aplicando el algoritmo HELIOSAT 1 con imágenes satelitales GOES en la Región Cundiboyacense para el Año 2011», Universidad Distrital Francisco José de Caldas, 2018. Accessed: January 15, 2021. [Online]. Available at: <http://repository.udistrital.edu.co/handle/11349/13314>
- [Eissa, 13] Y. Eissa, P. R. Marpu, I. Gherboudj, H. Ghedira, T. B. M. J. Ouarda, and M. Chiesa, «Artificial neural network based model for retrieval of the direct normal, diffuse horizontal and global horizontal irradiances using SEVIRI images», *Solar Energy*, vol. 89, pp. 1-16, mar. 2013, doi: 10.1016/j.solener.2012.12.008.
- [Fan, 18] J. Fan, X. Wang, L. Wu, F. Zhang, H. Bai, Lu X., and Y. Xiang, «New combined models for estimating daily global solar radiation based on sunshine duration in humid regions: A case study in South China», *Energy Conversion and Management*, vol. 156, pp. 618-625, ene. 2018, doi: 10.1016/j.enconman.2017.11.085.
- [Fan, 19] J. Fan, L. Wu, F. Zhang, H. Cai, X. Ma, and H. Bai, «Evaluation and development of empirical models for estimating daily and monthly mean daily diffuse horizontal solar radiation for different climatic regions of China», *Renewable and Sustainable Energy Reviews*, vol. 105, pp. 168-186, may 2019, doi: 10.1016/j.rser.2019.01.040.
- [Geetha, 22] A. Geetha, J. Santhakumar, K. M. Sundaram, S. Usha, T. M. T. Thentral, C. S. Boopathi, R. Ramya, and R. Sathyamurthy, «Prediction of hourly solar radiation in Tamil Nadu using ANN model with different learning algorithms», *Energy Reports*, vol. 8, pp. 664-671, abr. 2022, doi: 10.1016/j.egy.2021.11.190.
- [Gouda, 19] S. G. Gouda, Z. Hussein, S. Luo, and Q. Yuan, «Model selection for accurate daily global solar radiation prediction in China», *Journal of Cleaner Production*, vol. 221, pp. 132-144, jun. 2019, doi: 10.1016/j.jclepro.2019.02.211.
- [Gürel, 20] A. E. Gürel, Ü. Ağbulut, and Y. Biçen, «Assessment of machine learning, time series, response surface methodology and empirical models in prediction of global solar radiation», *Journal of Cleaner Production*, vol. 277, p. 122353, dic. 2020, doi: 10.1016/j.jclepro.2020.122353.
- [Guzman M., 21] O. Guzman M., J. V. Baldion R., O. Simbaqueva F., H. J. Zapata, and C. Chacon C., «Coeficientes para estimar la radiación solar global a partir del brillo solar en la zona cafetera colombiana», 2015, [Online]. Available at: <https://biblioteca.cenicafe.org/handle/10778/526>
- [Hammer, 01] A. Hammer, D. Heinemann, C. Hoyer-Klick, and E. Lorenz, «Satellite based short-term forecasting of solar irradiance - Comparison of methods and error analysis», *The 2001 EUMETSAT Meteorological Satellite Data User's Conference*, Jan. 2001.

- [Hammer, 03] A. Hammer, D. Heinemann, C. Hoyer-Klick, R. Kuhlemann, E. Lorenz, R. Müller, and H. G. Beyer, «Solar energy assessment using remote sensing technologies», *Remote Sensing of Environment*, vol. 86, pp. 423-432, ago. 2003, doi: 10.1016/S0034-4257(03)00083-X.
- [Huber, 19] S. Huber, H. Wiemer, D. Schneider, and S. Ihlenfeldt, «DMME: Data mining methodology for engineering applications – a holistic extension to the CRISP-DM model», *Procedia CIRP*, vol. 79, pp. 403-408, Jan. 2019, doi: 10.1016/j.procir.2019.02.106.
- [IDEAM, 20] IDEAM, «Catálogo Nacional de Estaciones del IDEAM | Datos Abiertos Colombia», 18 de mayo de 2020. <https://www.datos.gov.co/Ambiente-y-Desarrollo-Sostenible/Cat-logo-Nacional-de-Estaciones-del-IDEAM/hp9r-jxuu>
- [IDEAM, 21] IDEAM, «La Radiación Solar y su paso por la Atmósfera», 2021. <http://www.ideam.gov.co/web/tiempo-y-clima/la-radiacion-solar-y-su-paso-por-la-atmosfera>
- [Jiang, 19] H. Jiang, N. Lu, J. Qin, W. Tang, and L. Yao, «A deep learning algorithm to estimate hourly global solar radiation from geostationary satellite data», *Renewable and Sustainable Energy Reviews*, vol. 114, p. 109327, oct. 2019, doi: 10.1016/j.rser.2019.109327.
- [Jiang, 20] H. Jiang, N. Lu, G. Huang, L. Yao, J. Qin, and H. Liu, «Spatial scale effects on retrieval accuracy of surface solar radiation using satellite data», *Applied Energy*, vol. 270, p. 115178, jul. 2020, doi: 10.1016/j.apenergy.2020.115178.
- [Jumin, 21] E. Jumin, F. B. Basaruddin, Y. B. M. Yusoff, S. D. Latif, and A. N. Ahmed, «Solar radiation prediction using boosted decision tree regression model: A case study in Malaysia», *Environ Sci Pollut Res Int*, Jan. 2021, doi: 10.1007/s11356-021-12435-6.
- [Kaba, 18] K. Kaba, M. Sarigül, M. Avcı, and H. M. Kandirmaz, «Estimation of daily global solar radiation using deep learning model», *Energy*, vol. 162, pp. 126-135, nov. 2018, doi: 10.1016/j.energy.2018.07.202.
- [Kallio-Myers, 20] V. Kallio-Myers, A. Riihelä, P. Lahtinen, and A. Lindfors, «Global horizontal irradiance forecast for Finland based on geostationary weather satellite data», *Solar Energy*, vol. 198, pp. 68-80, mar. 2020, doi: 10.1016/j.solener.2020.01.008.
- [Kipp, 21] Kipp & Zonen, «Instrumentos solares», *Instrumentos solares - Kipp & Zonen*, 2021. <https://www.kippzonen.es/ProductGroup/85/Instrumentos-Solares>
- [Laguarda, 18] A. Laguarda, R. Alonso-Suárez, and G. Abal, «Modelo semi-empírico simple de irradiación solar global a partir de imágenes satelitales GOES», 2018, [Online]. Available at: <https://www.colibri.udelar.edu.uy/jspui/handle/20.500.12008/21610>
- [Linares-Rodríguez, 13] A. Linares-Rodríguez, J. A. Ruiz-Arias, D. Pozo-Vazquez, and J. Tovar-Pescador, «An artificial neural network ensemble model for estimating global solar radiation from Meteosat satellite images», *Energy*, vol. 61, pp. 636-645, nov. 2013, doi: 10.1016/j.energy.2013.09.008.
- [Linares-Rodríguez, 15] A. Linares-Rodríguez, S. Quesada-Ruiz, D. Pozo-Vazquez, and J. Tovar-Pescador, «An evolutionary artificial neural network ensemble model for estimating hourly direct normal irradiances from meteosat imagery», *Energy*, vol. 91, pp. 264-273, nov. 2015, doi: 10.1016/j.energy.2015.08.043.
- [Lorenz, 04] E. Lorenz, A. Hammer, and D. Heinemann, «Short term forecasting of solar radiation based on satellite data», *EUROSUN2004 (ISES Europe Solar Congress)*, Jan. 2004.
- [Lorenz, 12] E. Lorenz, J. Kühnert, and D. Heinemann, «Short Term Forecasting of Solar Irradiance by Combining Satellite Data and Numerical Weather Predictions», Jan. 2012, doi: 10.4229/27thEUPVSEC2012-6DO.12.1.

- [Manju, 19] S. Manju and M. Sandeep, «Prediction and performance assessment of global solar radiation in Indian cities: A comparison of satellite and surface measured data», *Journal of Cleaner Production*, vol. 230, pp. 116-128, sep. 2019, doi: 10.1016/j.jclepro.2019.05.108.
- [Martín Pomares, 06] L. Martín Pomares, L. Zarzalejo, J. Polo, B. Espinar, and L. Santigosa, «Predicción de la Irradiancia Solar Diaria a partir de Imágenes de Satélite Mediante Técnicas Estadísticas», Jan. 2006.
- [Mazorra Aguiar, 15] L. Mazorra Aguiar, B. Pereira, M. David, F. Díaz, and P. Lauret, «Use of satellite data to improve solar radiation forecasting with Bayesian Artificial Neural Networks», *Solar Energy*, vol. 122, pp. 1309-1324, dic. 2015, doi: 10.1016/j.solener.2015.10.041.
- [NOAA, 21] NOAA, «NOAA's Weather and Climate Toolkit (Viewer and Data Exporter)», 2021. <https://www.ncdc.noaa.gov/wct/>
- [NOAA Class, 21] NOAA Class, «NOAA's Comprehensive Large Array-data Stewardship System», 2021. <https://www.avl.class.noaa.gov/saa/products/welcome>
- [Nwokolo, 22] S. C. Nwokolo, S. O. Amadi, A. U. Obiwulu, J. C. Ogbulezie, and E. E. Eyibio, «Prediction of global solar radiation potential for sustainable and cleaner energy generation using improved Angstrom-Prescott and Gumbel probabilistic models», *Cleaner Engineering and Technology*, vol. 6, p. 100416, feb. 2022, doi: 10.1016/j.clet.2022.100416.
- [Ordoñez-Palacios, 20] L. E. Ordoñez-Palacios, D. A. León-Vargas, V. A. Bucheli-Guerrero, and H. A. Ordoñez-Eraso, «Predicción de radiación solar en sistemas fotovoltaicos utilizando técnicas de aprendizaje automático», *I*, vol. 29, n.º 54, Art. n.º 54, sep. 2020, doi: 10.19053/01211129.v29.n54.2020.11751.
- [Ordoñez Palacios, 22] L. E. Ordoñez Palacios, V. Bucheli Guerrero, and H. Ordoñez, «Machine Learning for Solar Resource Assessment Using Satellite Images», *Energies*, vol. 15, n.º 11, Art. n.º 11, Jan. 2022, doi: 10.3390/en15113985.
- [Oyewola, 22] O. M. Oyewola, T. E. Patchali, O. O. Ajide, S. Singh, and O. J. Matthew, «Global solar radiation predictions in Fiji Islands based on empirical models», *Alexandria Engineering Journal*, feb. 2022, doi: 10.1016/j.aej.2022.01.065.
- [Pagola, 14] I. Pagola, M. Gastón, A. Bernardos, and C. Fernández-Peruchena, «A Combination of Heliosat-1 and Heliosat-2 Methods for Deriving Solar Radiation from Satellite Images», *Energy Procedia*, vol. 57, pp. 1037-1043, Jan. 2014, doi: 10.1016/j.egypro.2014.10.088.
- [Plotnikova, 22] V. Plotnikova, M. Dumas, and F. P. Milani, «Applying the CRISP-DM data mining process in the financial services industry: Elicitation of adaptation requirements», *Data & Knowledge Engineering*, vol. 139, p. 102013, may 2022, doi: 10.1016/j.datak.2022.102013.
- [Poveda Matallana, 20] W. D. Poveda Matallana, «Validación de la radiación solar en superficie para la región Orinoquía a partir de imágenes de satélite», Bogotá - Ciencias - Maestría en Ciencias - Meteorología, 2020. Accessed: January 19, 2021. [Online]. Available at: <https://repositorio.unal.edu.co/handle/unal/77981>
- [Prescott, 40] J. Prescott, «Evaporation from a Water Surface in Relation to Solar Radiation», *Transactions of the Royal Society of South Australia*, vol. 64, pp. 114-118, 1940.
- [Rigollier, 04] C. Rigollier, M. Lefèvre, and L. Wald, «The method Heliosat-2 for deriving shortwave solar radiation from satellite images», *Solar Energy*, vol. 77, n.º 2, pp. 159-169, Jan. 2004, doi: 10.1016/j.solener.2004.04.017.

[Rodríguez Gómez, 19] J. D. Rodríguez Gómez, «¿Qué está pasando con las estaciones que miden la calidad del aire en Bogotá?», *RCN Radio*, 16 de marzo de 2019. <https://www.rcnradio.com/estilo-de-vida/medio-ambiente/que-esta-pasando-con-las-estaciones-que-miden-la-calidad-del-aire-en>

[Scikit-learn, 07] Scikit-learn. (2007). *API Reference*. Scikit-Learn. <https://scikit-learn/stable/modules/classes.html>.

[Solano, 22] J. A. Solano, D. J. Lancheros Cuesta, S. F. Umaña Ibáñez, and J. R. Coronado-Hernández, «Predictive models assessment based on CRISP-DM methodology for students performance in Colombia - Saber 11 Test», *Procedia Computer Science*, vol. 198, pp. 512-517, Jan. 2022, doi: 10.1016/j.procs.2021.12.278.

[Stel, 19] L. Stel, G. Haldbrant, and G. Sanchez, «Sobre mantenimiento e interpretación de fallas de estaciones automáticas.» October 2019. [Online]. Available at: http://repositorio.smn.gob.ar/bitstream/handle/20.500.12160/1191/Nota_Tecnica_SMN_2019-61.pdf?sequence=1&isAllowed=y

[Suárez Vargas, 13] D. A. Suárez Vargas, «Evaluación de la radiación solar en Bogotá a partir de imágenes del satélite Goes», Universidad Nacional de Colombia, Bogotá, 2013. Accessed: January 20, 2021. [Online]. Available at: <https://core.ac.uk/reader/19485169>

[US Department of Commerce, 21] US Department of Commerce, «NOAA's Office of Satellite and Product Operations», 2021. <https://www.ospo.noaa.gov/Operations/GOES/13/index.html>

[XGBoost, 22] XGBoost, developers. (2022). *XGBoost Documentation—Xgboost 1.7.4 documentation*. <https://xgboost.readthedocs.io/en/stable/>

[Zarzalejo, 06] L. Zarzalejo, L. Santigosa, J. Polo, L. Martín Pomares, and B. Espinar, «Estimación de la radiación solar a partir de imágenes de satélite: nuevos mapas de evaluación de la irradiancia solar para la península Ibérica», *Averma*, vol. 10, p. 11.71-11.78, Jan. 2006.