Solar Irradiation Prediction Based on M5 Model Tree and Feature Importance Evaluation

Lejla Idrizović, Lejla Lulić Skopljak, Faris Haznadarević, Haris Ahmetović
International Burch University
Sarajevo, Bosnia and Herzegovina
lejla.idrizovic@stu.ibu.edu.ba
lejla.lulic.skopljak@stu.ibu.edu.ba
faris.haznadarevic@stu.ibu.edu.ba
haris.ahmetovic@ibu.edu.ba

Orginal research

Abstract: *In the last decade, the usage of renewable energy is on the rise, and that trend will only continue because technology is becoming more developed, so renewable energy sources are going to offer more for the same price. Besides all positive properties, there are also some negatives like direct dependence on the weather conditions. That means energy production is constantly changing, so it must be as precisely as possible predicted to be usable on a large scale. Fifteen attributes were analyzed using M5 regression tree. High positive degree of correlation was found between participle water and dew point temperature, air temperature with dew point, air temperature with precipitation of water, snow depth with Albedo daily, zenith angle with relative humidity, GHI with Air temperature. It was found that the zenith angle, between the normal of the Earth's surface and the Sun, was the most important feature of the dataset for solar irradiation prediction.*

Keywords: machine learning, photovoltaic, renewable energy, solar irradiation, weather forecast.

1. Introduction

Today, energy is increasingly used in all its forms, but conventional methods have had a severe effect on our environment. Therefore, new methods were developed. People are becoming more aware of environmental problems caused by the usage of fossil fuels. Therefore, there is a high demand for the usage of renewable energy. Limited supply, increasing costs, climate change concerns and government mandates are also driving a desire to increase the percentage of electricity generated by renewable energy sources. Besides all positive properties, there are also some negatives like direct dependence on the weather conditions. That means energy production is constantly changing, so it must be as precisely as possible predicted to be usable on a large scale. Currently, the largest and most widespread energy source used by humans in the world is fossil fuel. Their uncontrolled exploitation and use in the last century have caused extensive environmental pollution, caused by enormous production of greenhouse gases during exploitation and use, resulting in climate changes. Another reason for reducing reliance on fossil fuels is that they are limited, so they can be depleted very easily if they continue to be used to the present extent.

On the other hand, clean energy from the sun and wind is present everywhere around the world in unlimited quantities, and we only need to harvest and use that energy. The solution to the problem of environmental pollution and climate change is not and cannot be instantaneous, so the results will be visible over a longer period. Renewable energy sources are a feasible and cost-effective solution that when used for the generation of electric energy also introduces certain additional complications into existing electrical energy systems. To meet the needs of modern society for energy and to achieve sustainability it is necessary to make a planned transition from the use of fossil fuels to renewable energy sources [1]. Renewable electrical energy sources introduce additional complexity to the process of maintaining the power quality, stability, and reliability of electric power systems. The reason for that additional complexity is that renewable energy comes from natural sources (processes), which are intermittent by nature as shown in Figure 1, where we presented examples of Global Horizontal Irradiation (GHI) patterns through one day. Regarding that, we can conclude that renewable energy depends on weather conditions, and there are also important tasks of reducing that unpredictability [2]. That challenge can be solved or at least significantly reduced with the use of modern analyzing techniques of current and historic weather data with the goal of precise prediction of the weather conditions. Furthermore, the development of computer hardware with higher computational power enabled the use of large amounts of data for providing highly useful results using machine learning techniques.

The purpose of this paper is to give a contribution to further development and research of solar irradiation prediction using machine learning techniques. The need for the development of this topic is on the rise because renewable energy is everywhere around the world, and humans only need to harvest it in the right way and use it. If the use of renewable energy continues to grow, it will modernize the world's electricity grid, making it smarter, more secure, and better integrated around the world, also most important is that our environment and our lives will be much healthier and more sustainable.

2. Literature Review

In [3], authors applied multivariate adaptive regression tree, M5, and random forest models for solar irradiation prediction for 1-day to 6- day ahead hourly prediction. Nine variables, minimum temperature, maximum temperature, wind speed, rainfall, dew point, global solar irradiation, atmospheric pressure, and solar azimuth were used as the inputs for model creation. Authors used root means square error to determine which models provided the best results and the result was validated using the t-static error.

Authors in [4] employed different machine learning algorithms for the precise estimation of solar irradiation for two locations. Ten attributes, namely year, month, day, hour, pressure, temperature, humidity, wind speed, hourly solar duration, and solar irradiation were used for the best attribute selection using six different feature selection methods to create data for five selection groups. It was shown that hourly solar duration was the most important feature for both selected data groups.

In [5], authors used maximum temperature, minimum temperature, sunshine hours, wind speed and relative humidity for inputs to build models for estimating solar radiation. Kriging, response surface method, multivariate adaptive regression and M5 model tree were applied.

Meenal and Slevakumar [6] evaluated artificial neural networks (ANN), Support Vector Machine (SVM) and empirical solar radiation models with different combination models. Eight attributes, month, latitude, longitude, bright sunshine hours, day length, relative humidity, maximum and minimum temperature were used as parameters. Best attributes were determined, and the most important feature to reduce the dimensionality of the data was identified to improve the correlation coefficient and the prediction accuracy of solar irradiation models.

3. Data Collection

Data used in this research are from Solcast website, which provides solar forecasting data for free in limited amounts if they are not going to be used commercially. Solcast is providing historical forecasting weather data from the beginning of 2007. year to present days, extracted from the high-resolution satellite imagery using advanced modeling techniques. These images are taken using geostationary meteorological satellites of the newer generation capable of providing high-resolution (1-2km) imagery every 5 to 15 minutes. There are 5 different measurement resolutions from every 60 minutes to every 5 minutes, and all this data are in the Comma Separated Value (CSV) format [7].

Historical weather data for the location of one potential photovoltaic (PV) power plant close to the center of the capital city of Bosnia and Herzegovina, Sarajevo, (longitude of 18,444170° and latitude of 43,891336°) are used. The time period of the used data set is chosen to be the longest possible, which is the period from the beginning of the year 2007. until the time when this data set is downloaded during December of 2020. Within that wide time period, data set resolution is limited to 10 minutes, because of computational limits of hardware used for model calculation and because this data set already gave highly satisfactory results.

In the research, the prediction of Solar irradiation has been performed for the given attributes. The sun radiates over the given area and provides insight into the possible energy production of a PV system. Furthermore, with that information, the number of solar panels and their installed power can be estimated. The solar irradiation value that has been predicted in this study is more precisely the Global Horizontal Irradiation (GHI). GHI is the amount of direct and diffuse solar irradiance received on the horizontal surface, measured with the unit of watts per meter squared (W/m^2) [7]. In Figure 1, a few examples of a daily GHI's from one summer month with different weather conditions are shown. Case from an entirely clear sky sunny day (yellow) to partly sunny day (green), and a completely cloudy day are shown (gray and orange).

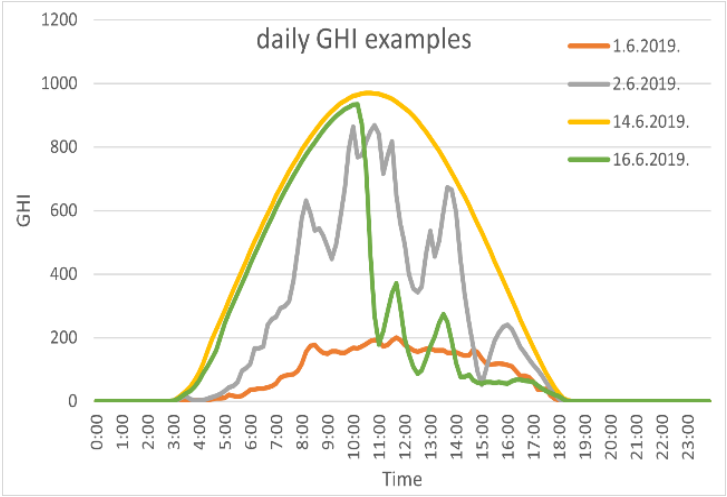

Figure 1. Example days of GHI

Preliminary Analysis of Attributes and the Corellation Between Them

The data set that enabled prediction of GHI contains 15 attributes, which are shown in Table 1, also they are sorted and numbered by relevance to the GHI prediction.

The Zenith Angle is most important for the prediction of GHI, it is the angle between the normal of the Earth's surface and the Sun, as it is shown in Figure 2.

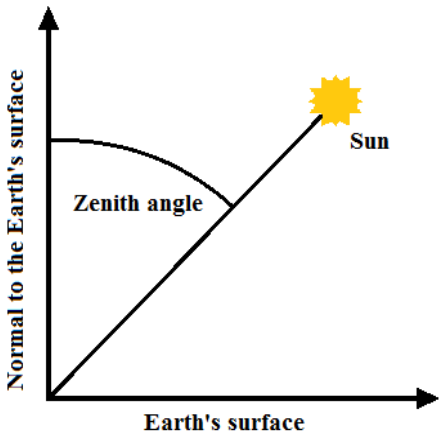

Figure 2. Zenith angle

Relative Humidity is estimated humidity for observed location, and it is presented in percentages (%).

Cloud opacity is the value of how opaque clouds are for sunlight in the observed region, and it is also presented in percentages (%).

Dew Point Temperature is the temperature of air for which air is saturated with water vapor and in contact with colder items it starts to condense on the item's surface [8].
Precipitable Water is the amount of water concentrated in the column of clouds that extends above some surface on the Earth, potentially available for precipitation [9]. Measured with units of kilograms on meter square (kg/m²). Albedo Daily is average daylight surface reflectivity, presented with a value between 0 and 1, where 0 is total absorption and 1 is total reflection.

To analyze the data and their relationship, Pearson correlation coefficient was used, where the value ±1 represents perfect positive (+1) and negative relationship (-1), from ±0.5 to ±1 represents a strong relationship, from ±0.30 to ±0.49 represents a medium relationship, and below ±0.29 represent a small relationship. Pearson correlation coefficient for two sets of values, x and y, is given by the formula:

$$r = \frac{\sum(x - \overline{x})(y - \overline{y})}{\sqrt{\sum(x - \overline{x})^2 \sum(y - \overline{y})^2}}$$

where $\overline{x}$ and $\overline{y}$ are the sample means of the two arrays of values.

Between the attributes in Table 1, high positive degree of correlation was found between the following attributes:

- Participle water and Dew point temperature,
- Air temperature with Dew point,
- Air temperature with precipitation of water,
- Snow depth with Albedo daily,
- Zenith with Relative Humidity,
- GHI with Air temperature.

A high negative correlation degree was found between the following attributes:

- Zenith with GHI,
- Relative Humidity with GHI.

A moderate degree of correlation was between the following attributes:

- Relative Humidity with Cloud Opacity,
- Wind speed with Cloud Opacity.

According to the Pearson correlation coefficient, for GHI, the zenith angle is the most important attribute, because they have a high negative degree of relationship (-0.764), and the attribute Air temperature, having a high degree of relationship of 0.502.
The highest degree of correlation between attributes was found between Participle water and Dew point temperature being 0.892, meaning for an increase in Participle water Dew Point temperature increases accordingly and vice versa.

Also, there is a high degree of correlation between the Air temperature with Dew point temperature being 0.889, and precipitation of water is 0.788. With high temperature, water evaporates into the atmosphere, hence increasing the amount of water in the air, affecting dew point and precipitation of water attributes. Snow depth and Albedo Daily have a high degree of correlation of 0.621. The reason behind this is that the snow being of the white color is more reflective than the darker ground surface [10].

Zenith and Relative humidity correlating 0.563. As the Sun is positioned normal to the Earth's surface, the angle decreases. During that period of the day, the Sun irradiating the Earth's surface is the highest, hence increasing air temperature. As the temperature increases, in turn, the humidity decreases and vice versa. This is also shown by the fact that the correlation between Zenith and air temperature is -0.504, indicating an inverse correlation between the attributes. In other words, when the Zenith angle increases, the air temperature decreases and vice-versa.

Table 1. Attributes used - sorted relevance to GHI predictions

	Attributes
1.	Zenith Angle
2.	Relative Humidity
3.	Air Temperature
4.	Cloud Opacity
5.	Dew Point Temperature
6.	Precipitable of Water
7.	Time
8.	Albedo Daily
9.	Snow height
10.	Surface Pressure
11.	Wind Speed
12.	Wind Direction
13.	Month
14.	Day
15.	Year

4. Modeling and Results

After the extensive analysis and tune up of the data set, it is ready to be fed to the machine learning algorithms. For that experimental setup, a specialized machine learning software WEKA is used in order to predict solar irradiation.

WEKA stands for Waikato Environment for Knowledge Analysis, and it is free software developed at the University of Waikato in New Zealand [11]. WEKA provides usage of many different machine learning tools and techniques. Data filters, data visualization, data classifiers and attribute selection tools were utilized in this study. Classifiers are algorithms that perform the process of predicting the class attribute, in this project GHI, of a given data points. Because GHI is in a range of 0 to 1000, it can be predicted using regression algorithms. Table 2 shows the statistics of the GHI from the used data set, including Minimum value, Mean value, Maximum value, and Standard deviation [8].

Table 2. Summary

Statistics	**Value**
Minimum	0
Maximum	991
Standard deviation	153,74
Mean	240,97

The M5' tree algorithm is a tree-based model algorithm, which constructs trees that can have multivariate linear models. By comparing it to the regression trees algorithms, M5' algorithm learns efficiently and it is good in tasks including big data sets and a high number of attributes [12].

After extensive testing and attribute selection processes, a summary of the results of the M5' algorithm are presented in Table 3, including Correlation coefficient, mean absolute error, Root mean squared errors, Relative absolute error, Root relative squared error and the Total number of instances.

Table 3. Performance of predicting GHI

Cross Validation	
Correlation coefficient	0,9995
Mean absolute error	3,6793
Root mean squared error	7,3679
Relative absolute error	1,9705%
Root relative squared error	3,0576%
Total number of instances	733383

The mean value from Table 2 is an average value of GHI in the data set, and the Mean absolute error from Table 3 is the average value of the predicted GHI error, so from these two values and by comparing them it can be known how good predictions are. Besides numerical interpretation, the precision of the predicted test data can be graphically interpreted, which is shown in the following Figure 4.

The next interesting step is attribute selection, which are collected by repeating model algorithm and every time excluding one attribute in the specific order of the attribute relevance for GHI prediction, like it is shown in Table 1.

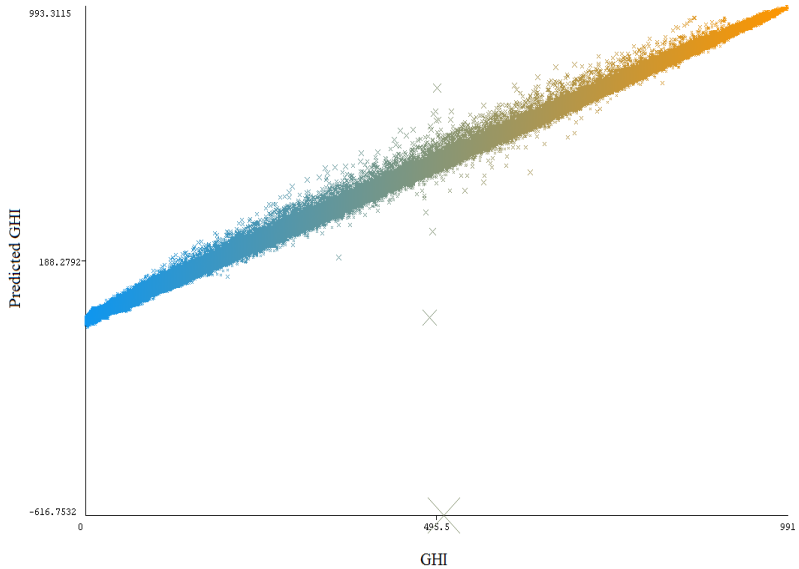

Figure 3. Scatter plot

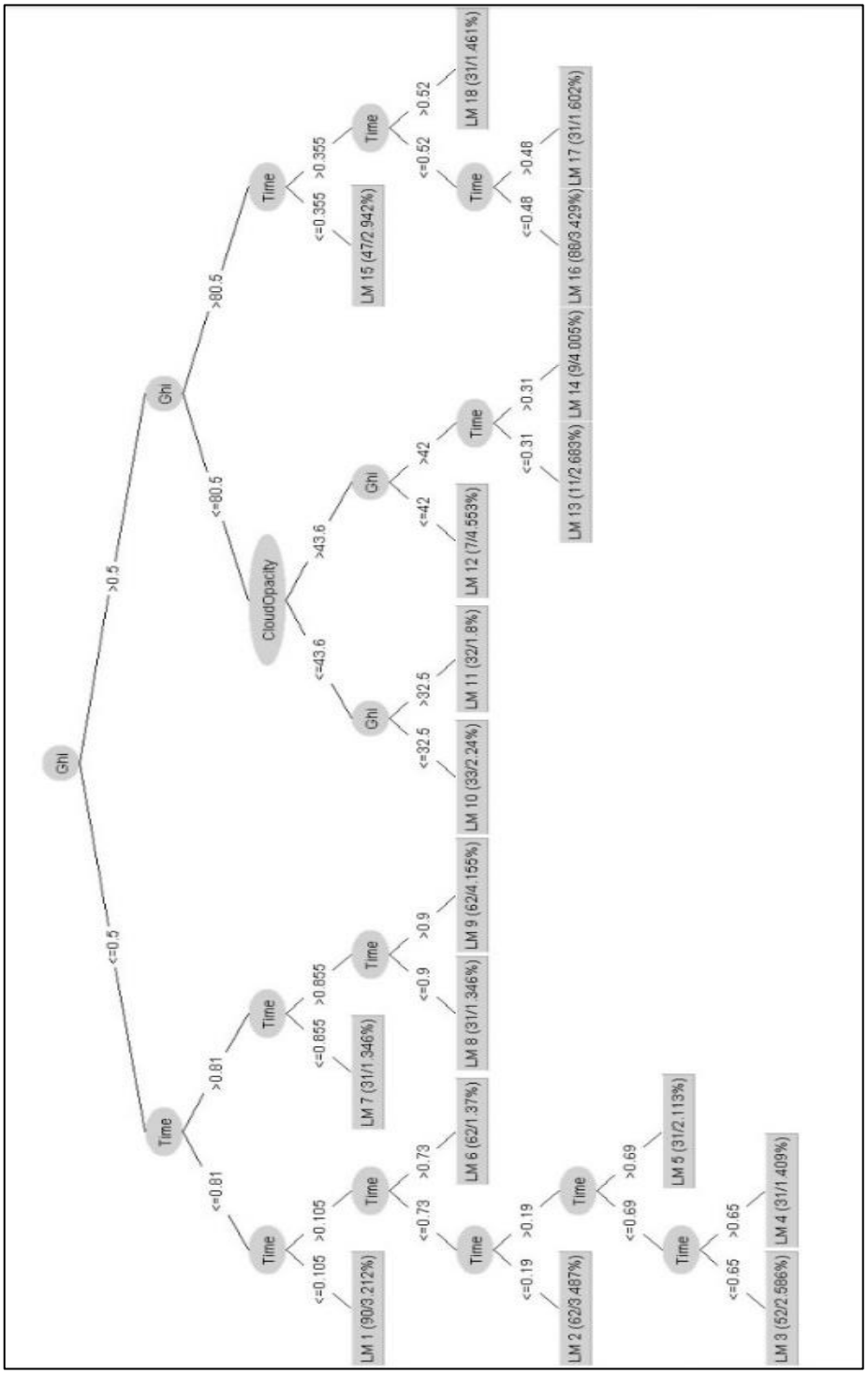

Figure 4. The tree of algorithm

The tree of our M5' with 1000 instances is shown in Figure 4. The results of this attribute selection are shown in Table 4.

Table 4. Attribute selection

Number of attributes	Attribute selection		
	Previously excluded attribute	***Correlation coefficient***	***Mean absolute error***
15		0,9995	3,6793
14	Year	0,9995	3,6904
13	Day	0,9995	3,6928
12	Month	0,9995	3,7509
11	Wind direction	0,9995	3,769
10	Wind speed	0,9995	3,7736
9	Surface pressure	0,9995	3,7776
8	Snow height	0,9995	3,8599
7	Albedo daily	0,9995	3,9416
6	Time	0,9995	4,0076
5	Precipitable of water	0,9993	4,6096
4	Dew point temperature	0,9992	4,7918
3	Cloud opacity	0,9281	44,8639
2	Air temperature	0,9231	47,1382
1	Relative humidity	0,8744	63,5234
0	Zenith	-0,003	186,7254

5. Conclusion

This study utilized the M5 regression tree for solar irradiation prediction. Azimuth angle was determined as the most important feature, year was the least important feature in the dataset. By excluding the least important feature, the correlation coefficient did not change. However, by its exclusion, the mean absolute error increased. Besides that, it also showed how solar energy sources and power grids with them can be easily upgraded in their operation. The proposed study has the potential to be even more precise and developed in the working, highly precise solar irradiance prediction tool, which could be used in some real-life applications.

References

1. A. M. P. O. A. V.-K. V. I. Mehrnoosh Torabi, "A Hybrid Machine Learning Approach For Daily Prediction of Solar Radiation," in Lecture Notes in Networks and Systems 53, 2018.

2. D. S. Alex Kim, "Predicting Solar Power Generation from Weather Data," Stanford University, 2019.
3. Srivastava, R., Tiwari, A. N., & Giri, V. K. (2019). Solar radiation forecasting using MARS, CART, M5, and random forest model: A case study for India. Heliyon, 5(10), e02692.
4. Guher, A. B., Tasdemir, S., & Yaniktepe, B. (2020). Effective Estimation of Hourly Global Solar Radiation Using Machine Learning Algorithms. International Journal of Photoenergy, 2020.
5. Keshtegar, B., Mert, C., & Kisi, O. (2018). Comparison of four heuristic regression techniques in solar radiation modeling: Kriging method vs RSM,
6. Meenal, R., & Selvakumar, A. I. (2018). Assessment of SVM, empirical and ANN based solar radiation prediction models with most influencing input parameters. Renewable Energy, 121, 324-343.
7. "Solcast," Solcast, [Online]. Available: https://solcast.com. [Accessed 2020/2021].
8. Cassel, Encyclopedia of Soils in the Environment, North Carolina: North Carolina State University, Raleigh, NC, USA, 2005..
9. "Department of Atmospheric Sciences (DAS) of the University of Illinois," University of Illinois, 2010. [Online]. Available: http://ww2010.atmos.uiuc.edu/(Gh)/guides/maps/sfcobs/dwp.rxml.
10. Albedo and reflective properties of various types of snow and water | AMAP. (n.d.). AMAP. Retrieved May 10, 2021, from https://www.amap.no/documents/doc/albedo-and-reflective-propertis-of-various-types-of-snow-and-water/971
11. "waikato," University of Waikato, [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/ .
12. J. R. Quinlan, "Learning with continuous classes," Sydney, 2006.