International Journal of Computer Applications (0975 – 8887)
Volume 184 – No.5, March 2022

Implementation of an Intelligent Model to Predict Solar Energy in North Morocco

Chaker El Amrani
Department of Computer Engineering,
FST de Tanger,
AbdelmalekEssaadi University
FST de Tanger, PO 416, Tangier, Morocco

Kawtar Chmichi
Department of Computer Engineering,
FST de Tanger,
AbdelmalekEssaadi University
FST de Tanger, PO 416, Tangier, Morocco

ABSTRACT
Air pollution is mainly due to the use of fossil energy in industrial and transport activities [1,2]. A solution to this problem is to replace fossil fuels by solar energy.

This study is about the prediction of solar energy production, in order to decide when and where the switch between the two sources can be made.

In this work, different prediction techniques weretested. They were developed with different Machine Learning models, namely, Decision Tree, Random Forest and Neural Networks. The best proposedalgorithmwas implemented in a Web application that shows prediction results, based on environmental variables values.

General Terms
Solar energy prediction, Artificial Intelligence algorithms, data acquisition and processing

Keywords
Solar energy, Artificial Intelligence

1. INTRODUCTION
Solar energy produced by photovoltaic panels is the most widely used renewable energy resource. Photovoltaic systems can be small, implemented in houses, or giant installed in large areas [3]. However, the intermittent nature of solar energy makes it difficult to integrate it into the electricity grid.

This study aims to predict the solar energy production, so that decision makers can schedule the switch between fossil and solar energy [4].

Predictions are based on archived data of a location in North-Morocco. Processing was made with Artificial Intelligence algorithms.

2. LOCATION AND DATA ACQUISITION

2.1 Location
This study concerns a location with latitude: 35.767846 and longitude: -5.917985, in North Morocco ((see Figure 1), particularly in an area called Mediouna. This zone is candidate to host a system of solar panels(see Figure 2).

This location is close to bothMediterranean Sea and Atlantic Ocean, and located near the forest, and it is not far from the city downtown of Tangier and from rural areas. That is why this location can be consideredas an ideal site andas the pilot location in this study. Furthermore, it is very representative of the whole North of Morocco, in terms of geographic, meteorological and environmental diversities.

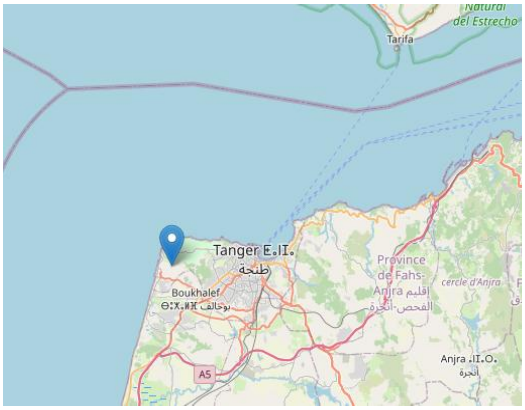

Fig 1: The study location

Fig 2: Solar panels installation design

2.2 Data acquisition and archives
Several sensors are installed in the studied region to measure 21 environmental variables including shortwave radiation, temperature, relative humidity, wind speed, the produced solar energy (KWh)w. The archive, available in csv format, covers daily data of 12 years, from 2007 to 2019.

The objective of this study is to predict the production of solar energy based on selected environmental variables, then to compare and validate the predictions with “generated power kw” variable which values are provided by the sensor.

3. PROCESSING DATA AND PREDICTIONS

3.1 Selection of representative variables
To build the proposed model, we first imported the data using Pandas libraries [5], an open source data analysis tool,to create a dataframe that include the environmental variables datasets.

The data pre-processing consists in assigning "artificial values" to missing values. In the present case, there are no missing values in the imported dataset, as shown in the following table:

Table 1. Number of missing values for each variable

date	0
temperature_2_m_above_gnd	0
relative_humidity_2_m_above_gnd	0
mean_sea_level_pressure_MSL	0
total_precipitation_sfc	0
snowfall_amount_sfc	0
total_cloud_cover_sfc	0
high_cloud_cover_high_cld_lay	0
medium_cloud_cover_mid_cld_lay	0
low_cloud_cover_low_cld_lay	0
shortwave_radiation_backwards_sfc	0
wind_speed_10_m_above_gnd	0
wind_direction_10_m_above_gnd	0
wind_speed_80_m_above_gnd	0
wind_direction_80_m_above_gnd	0
wind_speed_900_mb	0
wind_direction_900_mb	0
wind_gust_10_m_above_gnd	0
angle_of_incidence	0
zenith	0
azimuth	0
generated_power_kw	0

In statistics, correlation or dependence is a measure of how two variables are linearly related. The function "corr()" in Python, evaluates the correlation between all the variables. The result can be represented with colors (see Figure 3).

When score is zero or close to zero, this means that no correlation is found between the two variables. The dark blue indicates a perfect positive correlation, whereas the very light blue designates a perfect negative correlation.

We used "Lasso Regression" technique to find out the most relevant variables to the artificial intelligence model. If thevariable is not pertinent, Lasso assigns a value equal to zero (see Figure 4).

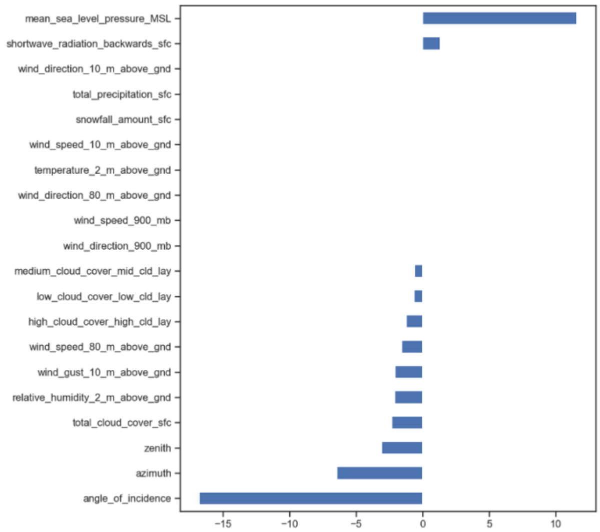

Fig 4: Feature importance using Lasso Model

Lasso model selects 12 variables and eliminates the others.We also used Backward Elimination, a feature selection technique, with "p-value" performance metric. When "p-value" is greater than 0.05, the candidate variable is eliminated. Otherwise, it is maintained. This method selects the same 12 variables than the Lasso Model.Another two models: Random Forest and Decision Trees were used to calculate the importance of the variables (See Figures 5 and 6).

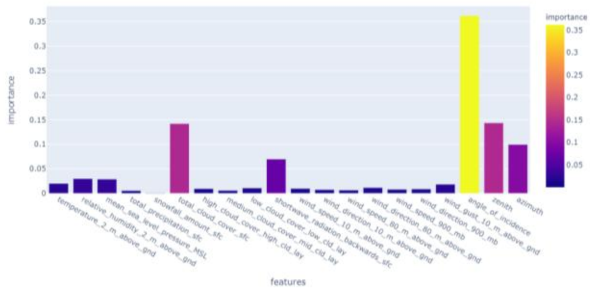

Fig 5: Importance of the variables with "Random Forest"

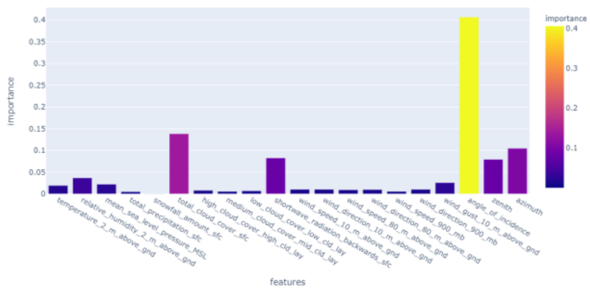

Fig 6: Importance of the variables with "Decision Trees"

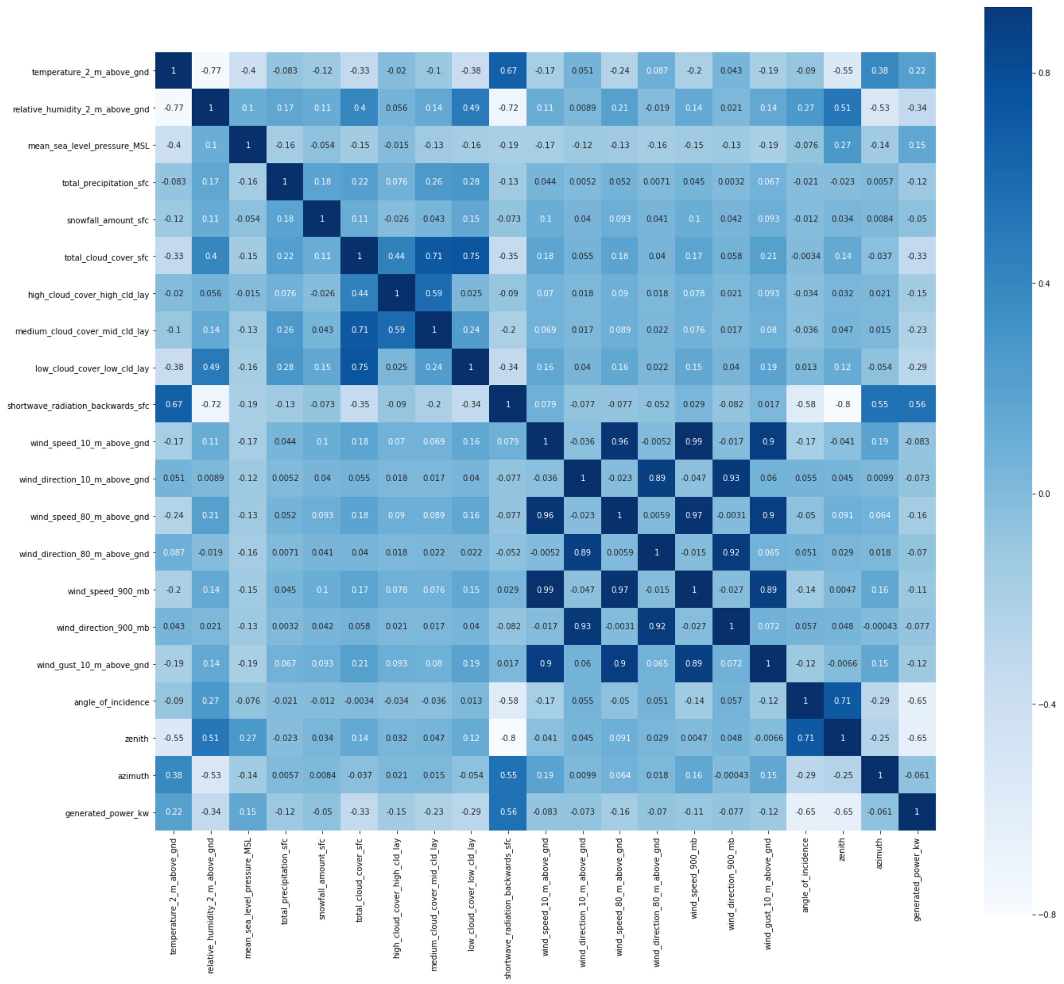

Fig 3 : Correlation between variables

We can see that each variable has an importance score, the higher this score is, the more pertinent are the variable.

3.2 Data decomposition

X and Y attributes are defined as follows:

- X: are variables found on the basis of the abovemodels' results and correlation.Accordingly, X is a set of 8 variables:
  - temperature_2_m_above_gnd
  - relative_humidity_2_m_above_gnd
  - mean_sea_level_pressure_MSL
  - total_cloud_cover_sfc
  - low_cloud_cover_low_cld_lay
  - shortwave_radiation_backwards_sfc
  - angle_of_incidence
  - zenith
- Y: is the target variable, which is assigned to the attribute: "generated_power_kw"

Python provides a function called: train_test_split ( ), which allows splitting the dataset into two groups, one for the training and the other for the testing. The last group serves to test and validate the prediction models developed with data from the training group.In this study, 75% of full dataset are the training set and the 25% remaining amount are the test set.

3.3 Predictive models

There are many models for prediction studies [6,7]. Among them, the most used by Data Miners are NeuralNetworks, Decision Trees and Random Forests.

In this work, we implemented, tested and compared these three algorithms. The best of them, called champion model,was integrated into the solar energy prediction application.

Brief descriptions of the three algorithms are as follows.

3.3.1 Neural Networks

Is a network of interconnected nodes inspired of brain neuron. An artificial Neural Network is composed (ANN) is composed of an input layer, hidden layers and an output layer. It uses learning algorithms that can operate adjustments or learn during the acquisition of new inputs. This makes ANN a strong tool for non-linear statistical data modeling.

3.3.2 Decision Trees

The objective of this algorithm is to develop a model that predicts the value of a target variable by learning decision rules provided by the data features.

3.3.3 Random Forests

The random forest is a classification algorithm including many decisions trees. It utilizes ensemble learning that combines several classifiers in order to provide solutions to complex problems.

3.4 Building models in Python

Variables are already divided into X and Y, and data are separated between "train" and "test". In order to build the model in Python, we used the Sickit-learn library, a free software machine learning library for Python, and the class: RandomForestRegressor to create the Random Forest model:

```
from sklearn.ensemble import RandomForestRegressor
rf = RandomForestRegressor()
```

and the learning is made as follows:

```
rf.fit(train_features, train_labels)
```

We used the "DecisionTreeRegressor" class to build the decision trees model, and we trained it as follows:

```
from sklearn.tree import DecisionTreeRegressor
tr = DecisionTreeRegressor()
```

```
rf.fit(train_features, train_labels)
```

The ANN modelis based on the Keras regression architecture (see Figure7). The network input is composed of the eight variables shown in paragraph 3.2. The model output is a neuron with a linear activation function « ReLuLeaky »to predict solar energy.

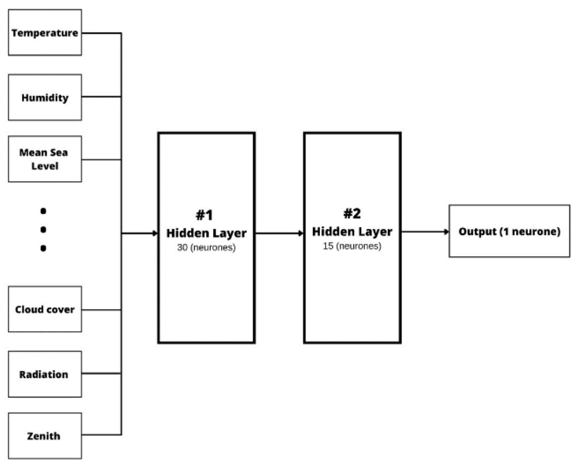

Fig7 : Regression with Keras

We defined a multilayer perceptron (MLP) with a function called: “create _mlp”. After the import of the necessary libraries, and the definition of MLP architecture, the model training started.

```
from keras.optimizers import Adam
from keras import optimizers


model = create_mlp(train_features.shape[1], regress=True)
model.compile(loss="mean_absolute_error", optimizer='Adam')

# train the model
print("[INFO] training model...")
hist=model.fit(train_features, train_labels, validation_data=(test_features, test_labels),epochs=300, batch_size=10000)
```

Grid search is an optimization technique that computes the optimum values of hyperparameters. The application of Grid search can save time and effort.

For Random Forests, we used Grid search with “Cross validation” (CV) technique. Usually we do not know the best hyperparameters. The optimal approach to find them is to assess a large number of values corresponding to each hyperparameter. To do so, we used “RandomizedSearchCV” function from Scikit-Learn,and we adjusted the following hyperparameters: n_estimators, max_features, max_depth, min_sample_split, min_samples_leaf and bootstrap.

For Decision Trees, we used GridSearchCV function, which is a useful tool to find the optimal hyperparameters to increase the model performance. The study will focus on these hyperparameters: max_depth, min_samples_split, min_samples_leaf, max_features.

The main difference between the two methods is that the GridSearchCV defines the combinations and do the training of the model,mainwhile the RandomizedSearchCV selects the combinations randomly.

3.5 Models assessment

The assessment of the three models was made with R^2Score, which is a statistical measure used to evaluate the performance of linear regression models. The following table shows the scores:

Table 2. Assessment of the models

	Without Grid Search	With Grid Search
Random Forest	0,79	0,82
Decision Trees	0.59	0.67
Neural Network	0.72	

We can see that "Random Forest" model provides the best result without "Grid Search". This score is improved with the use of "Grid Search". This is the model that will be employed in the following part of the work.

4. DEVELOPMENT OF A WEBSITE

A Website: http://solar.watching-environment.com/was designed and developed to predict the solar energy produced, based on the above eight environmental variables. We used Streamlit, a framework for Machine Learning and Data Science [8], and included libraries of Pandas.The Website is composed of three sub-menus (see Figure 8):

- Data collection

- Data Predictions

- Data Visualization

The Website provides one and fifteen days ahead forecasting of solar energy production. The firstoption serves to optimize the market bids and decide when and whether to switch temporarily to fossil fuels. The second option allows makingaverage term statisticsfor decision makers, mainly to avoid a decrease of solar energy production.

In predictions sub-menu,a user can choose the values of variableswith "Sliders", developed with "slicer()" function of Streamlit, and click on "predict" button which activate the "predict()" function.Then results are shown thanks to the "success()" function of Streamlit.

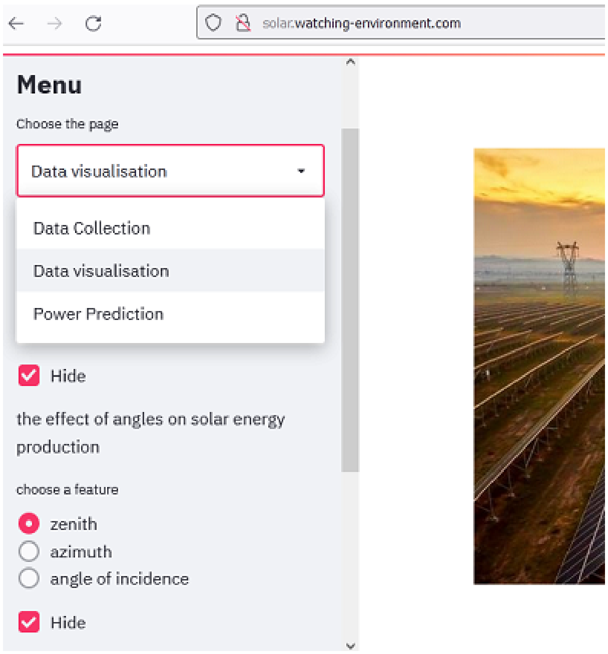

Fig8 : A Websitescreenshot

5. CONCLUSION

In this study, we proposed a methodology to predict solar energy of a given region. We chose eight variables as input and solar energy production as output of the prediction models. We split data to train and then to test the model. We compared three intelligent algorithms:Neural Networks, Decision Trees and Random Forests.We found that Random Forest with Grid Search has the best performance. On this basis, a Website was developed to allow users choosethe values of the eight variables and get as output one-ahead or fifteen-ahead days forecasting of solar energy production.The proposed future work consists on combining Eumetsat satellites and ground sensors datasets, to provide solar energy predictions in different regionsof Morocco.

6. ACKNOWLEDGMENTS

The authors would like to express their gratitude to the funding institution, AbdelmalekEssaadi University, for supporting this investigation.

7. REFERENCES

[1] Chaker El Amrani, Mohamed Akram Zaytar, Gilbert L. Rochon, Tarek El-Ghazawi, "Processing EUMETSAT Big Datasets to Monitor Air Pollution", The European Geosciences Union General Assembly 2018 (EGU), Vienna, Austria, 8–13 April 2018.

[2] Meryeme Boumahdi, Chaker El Amrani and Siegfried Denys, "An Innovative Air Purification Method and Neural Network Algorithm Applied to Urban Streets", International Journal of Embedded and Real-Time Communication Systems (IJERTCS), ISSN: 1947-3176, Volume 10, Issue 4, pp. 1-19, September 2019.

[3] K.N. Nwaigwe, P. Mutabilwa, E. Dintwa, "An overview of solar power (PV systems) integration into electricity grids", Materials Science for Energy Technologies, Volume 2, Issue 3, Pages 629-633, December 2019.

[4] Christil Pasion, Torrey Wagner, Clay Koschnick, Steven Schuldt, Jada Williams and Kevin Hallinan, "Machine Learning Modeling of Horizontal Photovoltaics Using Weather and Location Data", Energies 13, 2570; doi:10.3390/en13102570, May 2020.

[5] Pandas Website: https://pandas.pydata.org/

[6] Abdoul Kader Sidibe, Mounir Ghogho, Chaker El Amrani, Mustapha Faquir, "A comparative study of simple clear sky irradiance models", 2nd International Conference on Electrical and Information Technologies, Tangier, Morocco, May 4-7, 2016. (http://ieeexplore.ieee.org/document/7519572)

[7] Cyril Voyant, Gilles Notton, SoterisKalogirou, Marie-Laure Nivet, Christophe Paoli, Fabrice Motte, Alexis Fouilloy, "Machine learning methods for solar radiation forecasting: A review", Renewable Energy, Volume 105, Pages 569-582, May 2017.

[8] Streamlit Website: https://streamlit.io/