Solar Power Forecasting Based on Ensemble Learning Methods

Naylene Fraccanabbia*, Ramon Gomes da Silva†, Matheus Henrique Dal Molin Ribeiro†‡, Sinvaldo Rodrigues Moreno§, Leandro dos Santos Coelho†§¶ and Viviana Cocco Mariani*¶

*Mechanical Engineering Graduate Program (PPGEM), Pontifical Catholic University of Parana (PUCPR), Curitiba, Parana, Brazil. 80215–901

†Industrial & Systems Engineering Graduate Program (PPGEPS), Pontifical Catholic University of Parana (PUCPR), Curitiba, Parana, Brazil. 80215–901

‡Department of Mathematics, Federal Technological University of Parana (UTFPR), Pato Branco, Parana, Brazil. 85503–390

§Electrical Engineering Graduate Program (PPGEE), Federal University of Parana (UFPR), Curitiba, Parana, Brazil. 81530–000

¶Department of Electrical Engineering, Federal University of Parana (UFPR), Curitiba, Parana, Brazil. 81530–000

Emails: nayfraccanabbia@outlook.com, gomes.ramon@pucpr.edu.br, matheus.dalmolinribeiro@gmail.com, sinvaldo.moreno@gmail.com, leandro.coelho@pucpr.br, viviana.mariani@pucpr.br

***Abstract*—Alternative energy sources are becoming more and more common around the world. In order to reduce environmental pollution and $\mathrm{CO_2}$ emissions, in addition to being an ideal solution to overcome the energy crisis. In this context, power energy stands out, as it is the most abundant and most widely available natural resource on the entire planet. Due to the high level of uncertainty of the factors that directly interfere in the generation of solar power, such as temperature and solar radiation, make predictions of solar power with high precision is a challenge. Thus, the objective of this article is to develop a forecasting model, through time series, that makes it possible to predict the production of power energy, using a database collected in a photovoltaic plant in Uruguay. For the development of the proposal, models (base-learners), pre-processing techniques and models (meta-learners) used in the Stacking-Ensemble Learnig (STACK) method were used, which were compared using the measurements of performance Relative Root Mean Square Error (RRMSE), Symmetric Mean Absolute Percentage Error (sMAPE) and Determination Coefficient (R^2) in addition to statistical tests. In the end, it can be concluded that the combination Correlation Matrix (CORR) and Language Model (LM), from Layer-0 obtained the best results, in the three performance measures and the combination of models (base-learners) and pre-processing techniques (Layer-0) presented the best results when compared to Layer-1, obtaining satisfactory values in all performance measures.**

***Index Terms*—Solar power, time series, forecasting, stacking-ensemble learning, machine learning.**

I. Introduction

Currently, energy consumption is one of the criteria used in the analysis of countries economic development, which is in sharp expansion due to economic progress and development and increased demand, especially in the renewable energy market [1] [2]. Non-renewable sources of energy, such as coal, oil and natural gas, are used on a large scale by the world population. However, it must be considered that such fossil fuels may end up running out, due to the increase in the world population and economic development [3] In addition, fossil fuels harm the environment, increasing global warming, impacting climatic conditions and the government's economic policy [1]. Due to factors like these presented, the whole world started the development and use of alternative sources of clean and sustainable energy [4], in order to reduce environmental pollution and $\mathrm{CO_2}$ emissions, in addition to being an ideal solution to overcome the energy crisis [5].

Analyzing [6], it can be noted that world oil production has been decreasing since the mid-1970s, in relation to Global GDP. Such an achievement can be justified due to the prominence of renewable energy sources, which are growing on a large scale, since it has become common to find homes, industries, among others, implementing other sources of energy generation, clean and sustainable [2].

The most used renewable energy sources today are hydroelectric energy, from river waters; wind energy, from the strength of the winds; solar power, from sunlight; biomass energy, from organic materials; tidal energy, wave strength; and geothermal energy, from the internal heat of the earth, all of which are abundantly available to us, being theoretically unlimited. Conventional energy sources, such as coal, oil and natural gas, formed over millions of years, are limited and not homogeneously distributed around the world [3]. In the present work, the study was directed to solar power, since it is the most abundant and most widely available natural resource in the entire planet [2].

The authors would like to thank the National Council of Scientific and Technologic Development of Brazil – CNPq (Grants number: 307958/2019-1-PQ, 307966/2019-4-PQ, 404659/2016-0-Univ, 405101/2016-3-Univ), PRONEX '*Fundação Araucária*' 042/2018, and *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* - Brasil (CAPES) - Finance Code 001 for financial support of this work.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

Based on the growing availability of historical data, together with the need to make accurate forecasts, mainly in the area of energy, a sector in constant growth, it is possible to predict future values of a given sequence, being made through the forecasting of time series, which has attracted the attention of researchers in the field of machine learning, who seek to address the limitations of traditional forecasting methods [7]. Due to the high level of uncertainty of the factors that directly interfere in the generation of solar power, such as temperature and solar radiation, making predictions of solar power with high precision is a challenge. Nowadays, time series forecasting plays a fundamental role in several real-world problems, such as weather forecasting, financial markets, network traffic, the oil industry, among others.

Machine learning is a field in the domain of computer science that aims to develop methods and algorithms in order to learn and predict based on data. It is a data analysis method that automates the construction of analytical models, in addition to being a branch of artificial intelligence that is based on the idea that systems can learn from data, identify patterns and also make decisions with the least possible human intervention [8].

There are several methods of machine learning, for forecasting purposes for example, the classification and regression algorithms are widely used. To predict the components, the work combines methods and models of machine learning. Originally, most of these algorithms are initially classified and end up being modified to produce real values according to the input data set [8]. In this study the following algorithms will be used for analysis: Support Vector Regression with kernel Linear (SVRL), Multivariate Adaptive Regression Splines (MARS), Bayesian Regularization Neural Networks (BRNN), Language Model (LM) as the base-learner (weak models) and CUBIST and Support Vector Regression with kernel Radial (SVRR) as the meta-learners (strong models). In layer-0 of Stacking-ensemble learning (STACK), a combination of each pre-processing technique was made, namely principal component analysis (PCA) and Correlation matrix analysis (CORR), with each base-learner. Thus, eight forecasts were generated: PCA–SVRL, PCA–MARS, PCA–BRNN, PCA–LM, CORR–SVRL, CORR–MARS, CORR–BRNN and CORR–LM. In layer-1, the PCA and CORR pre-processing techniques were combined with the SVRR and CUBIST methods as meta-learners. Thus, eight final predictions were generated: PCA–PCA–SVRR–STACK, PCA–PCA–CUBIST–STACK, PCA–CORR–SVRR–STACK, PCA–CORR–CUBIST–STACK, CORR–PCA–SVRR–STACK, CORR–PCA–CUBIST–STACK, CORR–CORR–SVRR–STACK, CORR–CORR–CUBIST–STACK.

This work aims to predict, through time series, the production of solar power. To this end, we seek to find the most effective machine learning method, since in recent times, several approaches have been proposed with the objective of providing forecasts and offering support systems for decision making [7]. Thus, it is intended to analyze which method presents better results for studies of this profile, in addition to making use of the joint learning method, STACK, with the objective of improving forecasting accuracy, integrating several sub-models and operating by layers.

The contributions of this article aim at the development of a forecasting model, through time series, enabling the knowledge of future values of solar power production, a sector that is constantly growing, due to the worldwide need to use renewable energies and for being abundantly, freely and cleanly available to all [9]. In recent times, a large number of different techniques have been applied to problems that aim to estimate solar power. Most of these approaches are algorithms based on Machine Learning, which have a high capacity to obtain robust results in the estimation of solar energy, using different input variables, such as temperature and solar radiation [10].

The rest of this document is structured as follows: section II presents the materials used, addressing the data and its characteristics, in addition to describing the methods used in this study. In section III, the methodology is presented, describing the modeling steps and their performances. Section IV addresses the results and discussions and finally, section V presents the conclusion of this work and the intentions of future research.

II. Methodology

This section describes the steps adopted as a methodology, which was applied in this document. The structure is illustrated in Figure 1.

A. Description

- Through the autocorrelation (ACF - 2) and partial autocorrelation (PACF - 3) functions, significant delays for precision of solar power were verified. In this article lag 1 was adopted, using the following forecasting structure was adopted:

$$y_{(t+1)} = f\left\{y_t, \mathbf{x}_{it}\right\} + \epsilon \tag{1}$$

  where f is a function related to the model adopted in the training process, $y_{(t+1)}$ is the value predicted for a step forward, y_t is the value of the output observed in time t, $\mathbf{x}_{it}$ is a matrix of i entries in time t, i the number of system entries, and ϵ is the random error that follows a normal distribution of mean 0 and variance σ^2.
- The data set was divided into training and testing, with training 302 observations and testing 129. In addition, 10 features were generated as inputs, for y and the lag of y, which are: meanstra, sdtra, skewtra, diftra, expo2, expo3, ftanh, flog, min, and max.
- In layer-0 of STACK, a combination of each pre-processing technique PCA and CORR was made, with each base-learner, being: SVRL, MARS, BRNN and LM. Thus, eight forecasts were generated: PCA–SVRL, PCA–MARS, PCA–BRNN, PCA–LM, CORR–SVRL, CORR–MARS, CORR–BRNN and CORR–LM;
- In layer-1, the PCA and CORR pre-processing techniques were combined with the SVRR and CUBIST

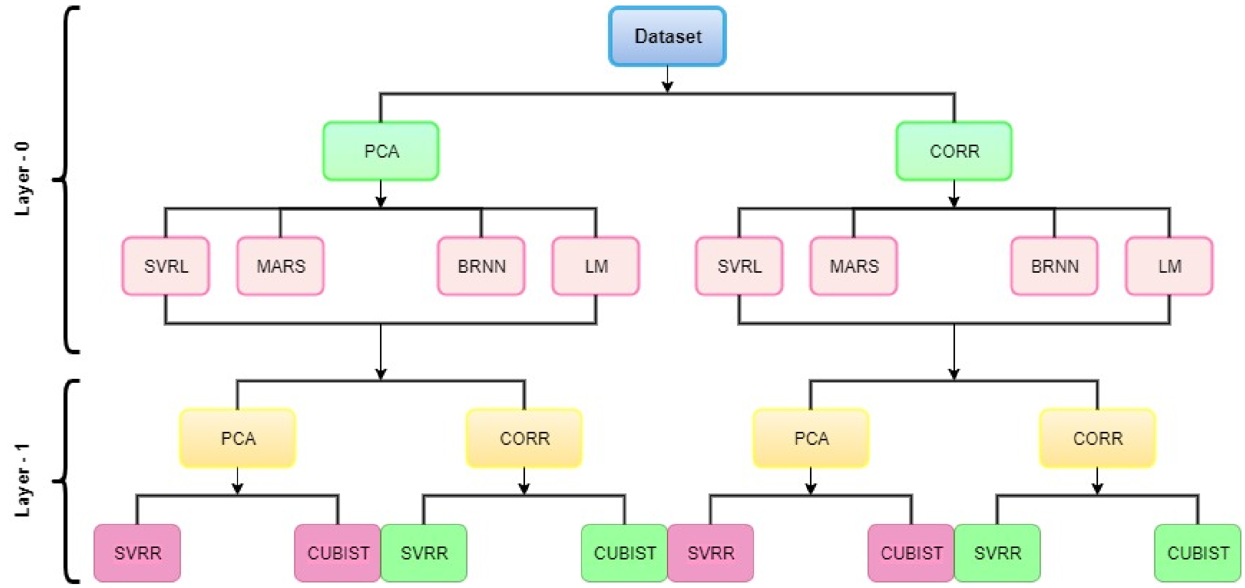

Fig. 1. Steps for developing the proposed model.

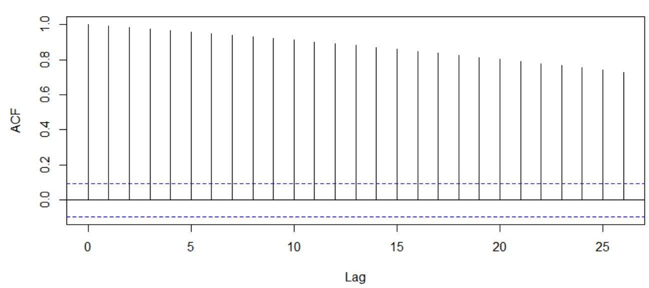

Fig. 2. Autocorrelation function.

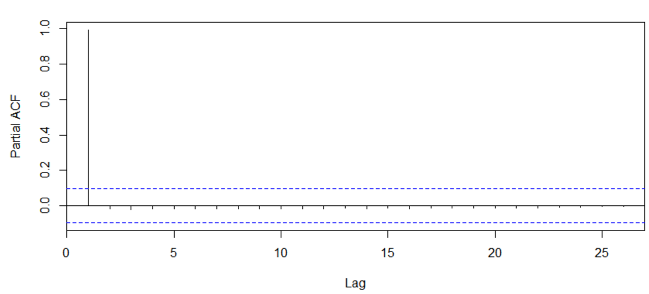

Fig. 3. Partial autocorrelation function.

methods as meta-learners. Thus, eight final predictions were generated: PCA–PCA–SVRR–STACK, PCA–PCA–CUBIST–STACK, PCA–CORR–SVRR–STACK, PCA–CORR–CUBIST–STACK, CORR–PCA–SVRR–STACK, CORR–PCA–CUBIST–STACK, CORR–CORR–SVRR–STACK, CORR–CORR–CUBIST–STACK.

- Table I shows the hyperparameters of the models used in this article. To measure acting, performance measures and statistical tests were obtained.

B. *Performance Measures*

- *Relative Root Mean Square Error*: The RRMSE is an indicator that is calculated by dividing the RMSE with the average value of the measured data [11], being given by the formula:

$$\text{RRMSE} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}}{\frac{1}{n}\sum_{i=1}^{n} y_i}, \tag{2}$$

  where n represents the number of observations of the training and test sets, y_i represents the i-th value observed, and $\hat{y}_i$ o i-th predicted value.

- *Symmetric mean absolute percentage error*: The SMAPE or sMAPE is a precision measure based on percentage or relative errors [12], given by the formula:

$$\text{sMAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{(|y_i| + |\hat{y}_i|/2)}\right| \tag{3}$$

  where n represents the number of observations of the training and test sets, y_i represents the i-th value observed, and $\hat{y}_i$ the i-th predicted value.

- *Determination coefficient*: R^2 can be considered a multiple correlation coefficient, that is, the correlation between the dependent variable and the set of independent variables [13], being given by the formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}[y_i(t) - \hat{y}_i(t)]^2}{\sum_{i=1}^{n}[y_i(t) - \overline{y}_i(t)]^2} \tag{4}$$

TABLE I

CONTROL HYPERPARAMETERS FOR THE MODELS

Model	**Control Hyperparameters**		**Model**	**Control Hyperparameters**		
PCA-SVRL	Cost 1	Kernel Linear	**PCA-PCA-SVRR-STACK**	Sigma 3.671366	Cost 2	Kernel Radial
PCA-MARS	Nprune 5	Degree 1	**PCA-PCA-CUBIST-STACK**	Committees 20	Neighbors 9	- -
PCA-BRNN	Neurons 1	- -	**PCA-CORR-SVRR-STACK**	Sigma 3518.273	Cost 0.25	Kernel Radial
PCA-LM	Intercept TRUE	- -	**PCA-CORR-CUBIST-STACK**	Committees 20	Neighbors 5	- -
CORR-SVRL	Cost 0.25	Kernel Linear	**CORR-PCA-SVRR-STACK**	Sigma 19.48925	Cost 0.25	Kernel Radial
CORR-MARS	Nprune 4	Degree 1	**CORR-PCA-CUBIST-STACK**	Committees 1	Neighbors 5	- -
CORR-BRNN	Neurons 5	- -	**CORR-CORR-SVRR-STACK**	Sigma 7031063	Cost 0.25	Kernel Radial
CORR-LM	Intercept TRUE	- -	**CORR-CORR-CUBIST-STACK**	Committees 20	Neighbors 9	- -

where n represents the number of observations of the training and test sets, y_i represents the i-th value observed, and $\hat{y}_i$ the i-th predicted value.

C. Diebold-Mariano test

In this article, a Diebold-Mariano (DM) test is performed, since with it it is possible to compare the forecasting errors of the models. The DM test checks whether one model's forecasting errors are lower or higher than another model [14] [15]. There is the hypothesis test, with the null hypothesis saying that there is no difference between the forecasting errors of the models that have been compared, and the alternative hypothesis says that the forecasting error of the proposed model is less than that of the compared model. With this hypothesis test, it is possible to know if the errors of the proposed model are lower than that of the compared model. The hypothesis test (5) and statistic of DM test (6) can be defined as follow:

$$H : \begin{cases} H_0 : \epsilon_i^P = \epsilon_i^C \\ H_1 : \epsilon_i^P < \epsilon_i^C, \end{cases} \tag{5}$$

$$\mathrm{DM} = \frac{\frac{\sum_{n}^{i=1} [d_i]}{n}}{\sqrt{\frac{\mathrm{var}(d_i)}{n-1}}}, \tag{6}$$

where $d_i = L(\epsilon_i^P) - L(\epsilon_i^C)$, L is a loss function that estimates the accuracy of each model, ϵ_i^P and ϵ_i^C are the error of the proposed model and the compared model, respectively, $\mathrm{var}(d_i)$ is an estimate for the variance of d_i.

III. MATERIAL & METHODS

The Materials and Methods section presents the description of the analyzed material (Section III-A), in addition to describing the methods used in this article (Section III-B).

A. Material

The set of data collected is from a 26.35MWp photovoltaic plant located in Artigas, Uruguay. The period of the sample used, from the data set, starts on April 14, 2018 at 00:00 and ends on April 16, 2018 at 23:50h, with data collected every 10 minutes. Therefore, the number of observations in this case is 431, as can be seen in Figure 4. The sample consists of three variables, as can be seen in Table II, where power is the system's output, and temperature and radiation are the system's inputs.

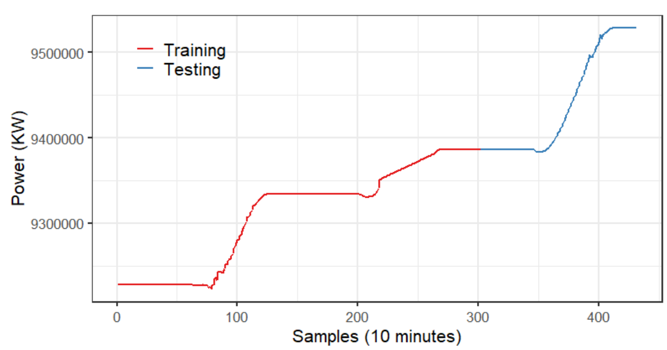

Fig. 4. Number of observations (sample) divided into training and testing.

The sample was divided into training (70%) and test (30%) sets. It is important to make this division before applying the machine learning algorithms, as it is possible to train the algorithm with a large volume of training data, it is possible to validate the sequence or result of this algorithm with test data. This gives confidence that the algorithm can actually visualize real data. Table III presents a summary of the statistical indicators of the data set, which are Maximum (Max), Minimum (Min), Mean, Median and Standard Deviation (Std).

TABLE II
INPUTS AND OUTPUT OF THE SYSTEM

Type	Description	Unit Measure
Input	Temperature	C°
Input	Radiation	W/m^2
Output	Power	kW

B. Methods

- Pre-processing techniques
  - *Principal Component Analysis*: PCA has as main idea to reduce the dimensionality of a data set that consists of many variables correlated with each other. This happens with the transformation of variables, in a new set of variables, which are the main components (PC) [16] [17].
  - *Correlation Matrix*: CORR is a correlation matrix, that is, the predictors are pre-processed using a correlation matrix, which removes the predictors that are cor7related higher than a limit. Thus, the function corr calculates correlations between predictor variables [18].

    In this case, pre-processing techniques PCA and CORR were used to reduce the amount of inputs generated with the features presented in Section III-A.
- Stacking-Ensemble Learnig
  - *Stacking-Ensemble Learnig*: STACK is a method that aims to improve the forecasting accuracy, integrating several sub-models and operating in layers [19].

    Therefore, STACK is an ensemble learning technique, which provides the predictions of a group of individual learners (base learners) as inputs to a second level learning algorithm (meta-learner). Thus, it combines the model prediction in an ideal way to form a final set of predictions [20].
- Models used in the STACK method
  - *CUBIST*: Cubist is a rules-based algorithm, used to build predictive models based on the analysis of input data. In a regression tree, a rule is built for each leaf, associated with the information contained in it. Final predictions are based on the linear combination of rules, which occurs when all rules are built. In order to try to improve the accuracy of the models, this model generates a set of rules, aiming to correct the predictions of previous members [21] [22].
  - *Multivariate Adaptive Regression Splines*: The non-linear and non-parametric method known as MARS, also known as Multivariate Adaptive Regression Splines, can be defined as an implementation of techniques popularized by [23], which are intended to solve regression problems, with the main objective of predicting the values of a dependent variable or continuous variable of a set of independent or predictor variables. The present method makes no assumptions about the underlying functional relationship between dependent and independent variables, instead it builds the relationship through a set of coefficients and base functions entirely "oriented" through regression data. can be considered suitable for problems with larger input dimensions [24].

    The process of building models takes place in an indented way. In the forward pass, the base functions are added to the model, in pairs, until the modification in the residual error is too small to continue, or else the maximum number of terms has been reached [25]. From this, an overfit model is usually obtained, that is, the step back becomes necessary in order to improve the generalization capacity, a process in which the least effective term is removed from the model until the best one is found submodel. In addition, there is the function spline, which divides the data into segments separated by different slope values, the MARS base function being dependent on the same [25].
  - *Language Model*: LM can be characterized as a probability distribution in sequences of words. These models have an important contribution when used as a base model, for various natural language processing tasks, including classification, information retrieval and other applications. Estimating the relative probability of different phrases is useful in many natural language processing applications, especially those that generate text as an output [26].
  - *Bayesian Regularized Neural Network*: One of the biggest difficulties encountered when designing a model of ANN (Artificial Neural Network), is to determine the number of hidden neurons, since many of them can lead to over-adjustment. It is known that an ANN model characterized by being too complex or too simple will have poor predictive performances. Thus, the BRNN (Bayesian Regularized Neural Network), which can be defined as a type of neural network, aims to prevent excessive adjustments from being made. BRNN consists of an entrance and a hidden layer that uses Bayesian methods that seek to estimate parameters, preventing such adjustments from being made [27] [28].
  - *Support Vector Regression with kernel Linear and Support Vector Regression with kernel Radial*: Support Vector Regression (SVR) [29] [30] works on principles similar to the Support Vector Machine (SVM) classification. It can be said that SVR is the adapted form of SVM when the dependent variable is numeric and not categorical. The Support Vector Machines (SVMs) can be characterized as a set of related supervised learning methods, which are popular in performing classification and regression analysis. Each specific method varies according to the structure and attributes of the classifier, the most well-known SVM being a linear classifier. In more detail, SVM creates a hyperplane or a set of hyperplanes to classify all entries in a high-dimensional space [31].

SVR technique depends on kernel functions, and the Kernel is a function used to map smaller data points to larger data points. There are several types of kernel: linear, radial, polynomial, Gaussian, among others.

TABLE III
SUMMARY OF THE STATISTICAL INDICATORS OF THE INPUTS AND OUTPUT OF THE DATASET

Variable	Samples	Percentage	# of Samples	Statistical indicator				
				Max	Min	Mean	Median	Std
Temperature	All set	100%	431	29.0	15.3	21.3	20.5	7.3977
	Training set	70%	302	28.4	16.3	20.9	20.1	5.5108
	Testing set	30%	129	29.0	15.3	22.2	22.3	5.1875
Radiation	All set	100%	431	963.2	0.0	254.8	0.0	341.2318
	Training set	70%	302	963.2	0.0	232.9	0.0	358.5303
	Testing set	30%	129	958.7	0.0	307.8	25.2	295.0030
Power	All set	100%	431	9528535	9222822	9349565	9337349	3052534
	Training set	70%	302	9386772	9222822	9310893	9334026	3207673
	Testing set	30%	129	9528535	9383581	9440742	9408307	2647387

IV. RESULTS

Table IV shows the performance measures of the models. The best results for training and testing are highlighted in bold. The applied performance measures were: sMAPE, RRMSE and R^2, first for the predictions formed by the combination of the models (base-learners) and the pre-processing techniques (Layer-0) and finally for the predictions formed by the combination of the models (meta-learners) used in the STACK method with pre-processing techniques (Layer-1). When analyzing the Table, it is noted that the Layer-0 CORR-LM combination obtained the best results, in the three performance measures, as highlighted in bold. Already analyzing Layer-1 separately, who obtained the best results was the combination CORR-CORR-CUBIST-STACK.

Therefore, after carrying out the experiments, the combination of the models (base-learners) and the pre-processing techniques (Layer-0) showed better results when compared to Layer-1, obtaining the most satisfactory results in all measurements performance. The hyperparameters, shown in Table III, were defined by grid search.

In addition, Diebold-Mariano (DM) tests were performed to compare the proposed models [(A) PCA-PCA-SVRR-STACK, (B) PCA-PCA-CUBIST-STACK, (C) PCA-CORR-SVRR-STACK, (D)PCA-CORR-CUBIST-STACK, (E) CORR-PCA-SVRR-STACK, (F) CORR-PCA-CUBIST-STACK, (G) CORR-CORR-SVRR-STACK, (H) CORR-CORR-CUBIST-STACK] with others models for each forecast horizon. Analyzing Table V, which presents the results of the tests, it can be seen that the model (A) is statistically equal to the model (E), the model (B) is statistically equal to the model (D) and the model (N) is statistically equal to the model (P).

V. CONCLUSION

This article was developed with the objective of predicting the production of solar energy, using a database collected in a photovoltaic plant of 26.35MWp, located in Artigas in Uruguay, in the period from April 14, 2018 at 00:00 until 16:00 April 2018 at 11:50 pm, with data collected every 10 minutes. The work covered data referring to power energy due to the growing demand for the use of renewable energy and because it is a natural resource available all over the planet.

For the development of the proposal, models (base-learners), pre-processing techniques and models (meta-learners) used in the STACK method were used, which were compared using the performance measures sMAPE, RRMSE and R^2 and with statistical tests.

At the end, it can be concluded that the combination CORR-LM, from Layer-0 obtained the best results, in the three performance measures and the combination of models (base-learners) and pre-processing techniques (Layer-0) presented the best results when compared to Layer-1, obtaining satisfactory values in all performance measures.

As a proposal for future research, (i) it is proposed to use different algorithms both in the models used in the STACK method, as in the other combinations; (ii) Increase the number of steps ahead for forecasting; and (iii) using other renewable energy sources - such as wind and biomass - for comparative studies and the use of other techniques.

REFERENCES

[1] M. K. Behera, I. Majumder, and N. Nayak, "Solar photovoltaic power forecasting using optimized modified extreme learning machine technique," *Engineering Science and Technology, an International Journal*, vol. 21, no. 3, pp. 428–438, 2018.

[2] Y. Feng, W. Hao, H. Li, N. Cui, D. Gong, and L. Gao, "Machine learning models to quantify and map daily global solar radiation and photovoltaic power," *Renewable and Sustainable Energy Reviews*, vol. 118, p. 109393, 2020.

[3] F. Zeren and H. T. Akkuş, "The relationship between renewable energy consumption and trade openness: New evidence from emerging economies," *Renewable Energy*, vol. 147, pp. 322–329, 2020.

[4] B. K. Sahu, "A study on global solar pv energy developments and policies with special focus on the top ten solar pv power producing countries," *Renewable and Sustainable Energy Reviews*, vol. 43, pp. 621–634, 2015.

[5] 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, 2019.

[6] O. A. Pedrosa Jr and A. C. de França Corrêa, "A crise do petróleo e os desafios do pré-sal," *Boletim de Conjuntura*, no. 2, pp. 4–14, 2016.

[7] A. Sagheer and M. Kotb, "Time series forecasting of petroleum production using deep lstm recurrent networks," *Neurocomputing*, vol. 323, pp. 203–213, 2019.

[8] H. Demolli, A. S. Dokuz, A. Ecemis, and M. Gokcek, "Wind power forecasting based on daily wind speed data using machine learning algorithms," *Energy Conversion and Management*, vol. 198, p. 111823, 2019.

[9] Y. Zhang, J. Ren, Y. Pu, and P. Wang, "Solar energy potential assessment: A framework to integrate geographic, technological, and economic indices for a potential analysis," *Renewable Energy*, vol. 149, pp. 577–586, 2020.

[10] L. Cornejo-Bueno, C. Casanova-Mateo, J. Sanz-Justo, and S. Salcedo-Sanz, "Machine learning regressors for solar radiation estimation from satellite data," *Solar Energy*, vol. 183, pp. 768–775, 2019.

[11] M. Despotovic, V. Nedic, D. Despotovic, and S. Cvetanovic, "Evaluation of empirical models for predicting monthly mean horizontal diffuse solar radiation," *Renewable and Sustainable Energy Reviews*, vol. 56, pp. 246–260, 2016.

[12] D. Konstantin and Z. Dominique, "Variational mode decomposition," *IEEE Transactions on Signal Processing*, vol. 62, no. 3, pp. 531–544, 2014.

[13] R. Riffenburgh, "Chapter 28 - methods you might meet, but not every day," in *Statistics in Medicine (Third Edition)*, third edition ed., R. Riffenburgh, Ed. San Diego: Academic Press, 2012, pp. 581 – 591. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780123848642000287

TABLE IV
PERFORMANCE MEASURES RESULTS OF THE MODELS

METRIC METHODS \Layer-0				**METRIC METHODS STACK \Layer-1**			
	sMAPE	RRMSE	R^2		sMAPE	RRMSE	R^2
PCA-SVRL	9.449549e-04	9.779133e-04	0.9997	PCA-PCA-SVRR-STACK	0.0108	0.0145	0.7025
PCA-MARS	9.473985e-04	1.018069e-03	0.9984	PCA-PCA-CUBIST-STACK	0.0056	0.0083	0.3338
PCA-BRNN	1.742184e-03	2.488815e-03	0.9946	PCA-CORR-SVRR-STACK	0.0103	0.0131	0.5773
PCA-LM	1.851099e-04	2.111932e-04	0.9998	PCA-CORR-CUBIST-STACK	0.0057	0.0084	0.4026
CORR-SVRL	1.074919e-03	1.171680e-03	0.9999	CORR-PCA-SVRR-STACK	0.0117	0.0147	0.6002
CORR-MARS	2.798447e-05	6.844400e-05	0.9999	CORR-PCA-CUBIST-STACK	0.0039	0.0059	0.9489
CORR-BRNN	1.969842e-03	3.000576e-03	0.9959	CORR-CORR-SVRR-STACK	0.0101	0.0132	0.4232
CORR-LM	**1.100414e-05**	**2.372946e-05**	**0.9999**	CORR-CORR-CUBIST-STACK	0.0014	0.0024	0.9874

TABLE V
DIEBOLD-MARIANO TEST RESULTS

Model	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)	(K)	(L)	(M)	(N)	(O)	(P)
(A) PCA-PCA-SVRR-STACK	-	8.8299*	8.9003*	8.7990*	1.2679	8.6965*	8.8302*	8.6453*	8.6432*	8.6432*	8.6451*	8.6432*	8.6432*	8.6432*	8.6467*	8.6432*
(B) PCA-PCA-CUBIST-STACK	-	-	-8.7708*	-1.757726	-8.9679*	7.2834*	-8.8186*	7.2895*	7.2863*	7.2860*	7.2895*	7.2866*	7.2861*	7.2866*	7.2870*	7.2866*
(C) PCA-CORR-SVRR-STACK	-	-	-	8.7187*	-9.2565*	8.5678*	-9.9921*	8.4939*	8.4909*	8.4908*	8.4937*	8.4909*	8.4909*	8.4909*	8.4958*	8.4909*
(D) PCA-CORR-CUBIST-STACK	-	-	-	-	-8.9366*	7.8333*	-8.7683*	7.6795*	7.6726*	7.6724*	7.6789*	7.6729*	7.6725*	7.6729*	7.6813*	7.6729*
(E) CORR-PCA-SVRR-STACK	-	-	-	-	-	8.8166*	9.2023*	8.7595*	8.7572*	8.7572*	8.7593*	8.7572*	8.7572*	8.7572*	8.7611*	8.7572*
(F) CORR-PCA-CUBIST-STACK	-	-	-	-	-	-	-8.6131*	7.2270*	7.2159*	7.2150*	7.2258*	7.2168*	7.2150*	7.2168*	7.2181*	7.2168*
(G) CORR-CORR-SVRR-STACK	-	-	-	-	-	-	-	8.5384*	8.5353*	8.5353*	8.5382*	8.5354*	8.5353*	8.5354*	8.5403*	8.5354*
(H) CORR-CORR-CUBIST-STACK	-	-	-	-	-	-	-	-	6.8960*	6.8658*	4.8139*	6.9226*	6.8642*	6.9228*	-7.4720*	6.9229*
(I) PCA-SVRL	-	-	-	-	-	-	-	-	-	-2.1311**	-6.8965*	10.2051*	-8.3503*	10.2451*	-7.1830*	10.2507*
(J) PCA-MARS	-	-	-	-	-	-	-	-	-	-	-6.8604*	6.6480*	-4.4925*	6.6627*	-7.1694*	6.6631*
(K) PCA-BRNN	-	-	-	-	-	-	-	-	-	-	-	6.9237*	6.8613*	6.9240*	-7.4023*	6.9240*
(L) PCA-LM	-	-	-	-	-	-	-	-	-	-	-	-	-9.5986*	4.4538*	-7.1964*	8.5665*
(M) CORR-SVRL	-	-	-	-	-	-	-	-	-	-	-	-	-	9.6156*	-7.1692*	9.6163*
(N) CORR-MARS	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-7.1965*	1.3622
(O) CORR-BRNN	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	7.1965*
(P) CORR-LM	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Note: *1% significance level; **5% significance level.

[14] M. H. D. M. Ribeiro, V. H. A. Ribeiro, G. Reynoso-Meza, and L. S. Coelho, "Multi-objective ensemble model for short-term price forecasting in corn price time series," in *International Joint Conference on Neural Networks (IJCNN)*, Budapest, Hungary, July 2019, pp. 1–8.

[15] F. X. Diebold and R. S. Mariano, "Comparing predictive accuracy," *Journal of Business and Economic Statistics*, vol. 13, no. 3, pp. 253–263, 1995.

[16] X. Du and F. Zhu, "A novel principal components analysis (pca) method for energy absorbing structural design enhanced by data mining," *Advances in Engineering Software*, vol. 127, pp. 17–27, 2019.

[17] Y. Gündoğdu, P. Karabağlı, H. Alptekin, M. Şahin, and H. Ş. Kılıç, "Comparison of performances of principal component analysis (pca) and factor analysis (fa) methods on the identification of cancerous and healthy colon tissues," *International Journal of Mass Spectrometry*, vol. 445, p. 116204, 2019.

[18] M. Kuhn and K. Johnson, *Applied predictive modeling*. Springer, 2013, vol. 26.

[19] D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.

[20] M. H. D. M. Ribeiro and L. S. Coelho, "Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series," *Applied Soft Computing*, vol. 86, 2020.

[21] S. C. Worland, W. H. Farmer, and J. E. Kiang, "Improving predictions of hydrological low-flow indices in ungaged basins using machine learning," *Environmental modelling & software*, vol. 101, pp. 169–182, 2018.

[22] R. Houborg and M. F. McCabe, "A hybrid training approach for leaf area index estimation via cubist and random forests machine-learning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 135, pp. 173–188, 2018.

[23] J. H. Friedman *et al.*, "Multivariate adaptive regression splines," *The annals of statistics*, vol. 19, no. 1, pp. 1–67, 1991.

[24] S. Mehdizadeh, F. Fathian, and J. F. Adamowski, "Hybrid artificial intelligence-time series models for monthly streamflow modeling," *Applied Soft Computing*, vol. 80, pp. 873–887, 2019.

[25] S. S. Roy, R. Roy, and V. E. Balas, "Estimating heating load in buildings using multivariate adaptive regression splines, extreme learning machine, a hybrid model of mars and elm," *Renewable and Sustainable Energy Reviews*, vol. 82, pp. 4256–4268, 2018.

[26] M. Sundermeyer, R. Schlüter, and H. Ney, "Lstm neural networks for language modeling," in *Thirteenth annual conference of the international speech communication association*, 2012.

[27] D. T. Bui, B. Pradhan, O. Lofman, I. Revhaug, and O. B. Dick, "Landslide susceptibility assessment in the hoa binh province of vietnam: A comparison of the levenberg–marquardt and bayesian regularized neural networks," *Geomorphology*, vol. 171-172, pp. 12 – 29, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169555X12002061

[28] D. J. MacKay, "Bayesian interpolation," *Neural computation*, vol. 4, no. 3, pp. 415–447, 1992.

[29] S. Shataee, S. Kalbi, A. Fallah, and D. R. Pelz, "Forest attribute imputation using machine-learning methods and aster data: comparison of k-nn, svr and random forest regression algorithms," 2012.

[30] M. H. D. M. Ribeiro, R. G. da Silva, V. C. Mariani, and L. d S Coelho, "Short-term forecasting covid-19 cumulative confirmed cases: Perspectives for brazil," *Chaos, Solitons & Fractals*, p. 109853, 2020.

[31] G. Junior, "Máquina de vetores suporte: estudo e análise de parâmetros para otimização de resultado," *Ciência da Computação, Universidade Federal de Pernambuco*, 2010.