Modeling and prediction of a multivariate photovoltaic system, using the multiparametric regression model with Shrinkage regularization and eXtreme Gradient Boosting

Saul Huaquipaco Encinas, Eng[1], Jose Cruz, Phd[1], Norman Jesus Beltran, Phd[2]
Ferdinand Pineda, Msc.[1], Christian Romero, Eng[1], Julio Fredy Chura Acero Msc.[1] Wilson Mamani Machaca, Est[1]
[1]Universidad Nacional del Altiplano, Peru, saul@pizdii.com, josecruz@unap.edu.pe, ferpineda@unap.edu.pe, romeroc24@gmail.com, jchura@unap.edu.pe, wilmamanimac@est.unap.edu.pe
[2]Universidad Nacional de Juliaca, Peru, nbeltran@unaj.edu.pe

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2021.1.1.557
ISBN: 978-958-52071-8-9 ISSN: 2414-6390

Modeling and prediction of a multivariate photovoltaic system, using the multiparametric regression model with Shrinkage regularization and eXtreme Gradient Boosting

Saul Huaquipaco Encinas, Eng[1], Jose Cruz, Phd[1], Norman Jesus Beltran, Phd[2]
Ferdinand Pineda, Msc.[1], Christian Romero, Eng[1], Julio Fredy Chura Acero Msc.[1] Wilson Mamani Machaca, Est[1]
[1]Universidad Nacional del Altiplano, Peru, saul@pizdii.com, josecruz@unap.edu.pe, ferpineda@unap.edu.pe, romeroc24@gmail.com, jchura@unap.edu.pe, wilmamanimac@est.unap.edu.pe
[2]Universidad Nacional de Juliaca, Peru, nbeltran@unaj.edu.pe

Abstract*– Alternative energy systems have more frequently been acquiring a fundamental role in the generation of energy that promotes the development of countries in social, economic, and environmental terms. For the efficient operation of photovoltaic systems (SFV), it is necessary to make predictions about their operation, turning them into intelligent systems. The present work proposes the collection, modeling, and prediction of a multivariate SFV, using a multiparametric regression model, presenting five regression models with machine learning: three that use Shrinkage regularization and two that use eXtreme Gradient Boosting (XGBoost). Results obtained, we note that the five predictions have determination coefficients higher than 99.47%; being XGBoost with n_estimators = 500 which reduces the root mean square error by about 55%. Likewise, in all cases, the test times are less than 1 second. The results were validated so that they not only have mathematical significance, but are also real, showing that XGBoost with n_estimators = 10 does not meet the five validation conditions, so this prediction model should not be considered.

Keywords*-- Modeling, Prediction, Photovoltaic System, Shrinkage Regularization, XGBoost.

I. Introduction

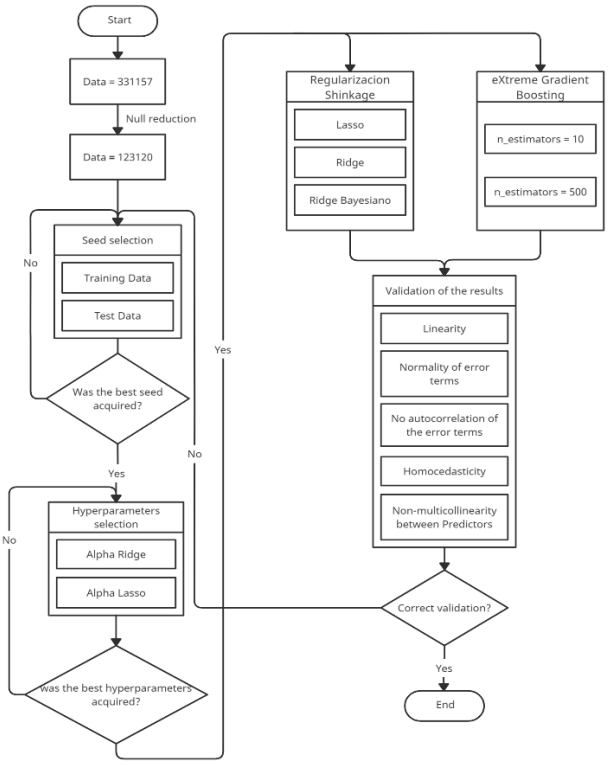
Photovoltaic systems with increasing frequency are acquiring a fundamental role in the generation of energy for the public and private sectors of the countries. For these systems to function efficiently, they need to be turned into smart systems. For this, its operation can be predicted, taking not only characteristics of the system, but also complementary ones such as irradiance and temperature. Therefore, the present work proposes the collection, modeling, and prediction of a multivariate SFV, using a multiparametric regression model. Currently, various investigations are being carried out for both control and forecasting of alternative energy systems such as [1] and [2] propose a system to monitor solar parameters formed by a data acquisition card from a module Arduino, and by a SCADA system developed in Matlab or values delivered in CSV format to reduce implementation costs. [3] uses an application programming interface (APIS) and the Web services of the meteorology system to present the design of a monitoring system for photovoltaic power plants, including not only internal factors but also external factors such as temperature, humidity, wind speed. [4] proposes a renewable energy information management system for energy storage systems taking into account the phases of generation, collection, storage, semantics, and visualization of results using Bigdata. [5] proposes a system based on K-means for the treatment of missing data in data collected from photovoltaic systems in the distribution part, achieving a normalized mean absolute range error (MARNE) of 3%. [6] generated a oneyear database at a sampling frequency of 5KHz of open-circuit voltage, short-circuit current, and maximum power values for SFVs located on the roofs of houses, mentioning that he used regularization and imputation techniques of data. Other authors are beginning to use new methods to make classifications or forecasts, among which those made from XGBoost stand out, such as: [7] which predicts storage from a method built with XGBoost, prediction of electricity prices reducing consumption costs of cloud computing storage data centers, achieving an accuracy of 91%, improving the accuracy of other algorithms such as support vector regression (SVR) that achieve only an accuracy of 88%. Also [8] proposes a method to predict the power in a photovoltaic system, from three submodels: XGBoost, LSTM, and LightGBM, whose results are weighted by means of weights, giving a final result that is better than those obtained with traditional models such as ARIMA or SVM. In the same way [9] proposes a method called dynamic weight set model (DWEM) for the prediction of electricity consumption based on previous values, which in three stages: serialization, four sub-models: MLP, CNN, XGBoost, and RF, combining the results through weightings and different serializations, improves the results comparing them with other processes such as EN3-bestK in 44.47% for the mean absolute error. [10] Also for the efficient prediction of energy consumption, it uses genetic algorithms to

predict the total load consumption, but in the previous stage of variable selection, it uses three methods: XGBoost, SVR, and Knn, achieving a mean absolute percentage error (MAPE) of 3.35%. [11] for combined cooling, heating, and energy systems proposes a two-stage electrical load forecasting system: hourly load forecast using XGBoost and RF, and then combine both results using a multiple linear regression based on sliding windows, achieving a MAPE of 4.49%. [12] [13] implement a solar library and a photovoltaic charging system using 3.0 charging technology in a city at 3800 msnm.

This article seeks to establish a multiparametric regression model for the multivariate photovoltaic system, for which we present five regression models: three that use Shrinkage regularization and two that use eXtreme Gradient Boosting (XGBoost). The contributions of the article are:

- Implementation of a multivariate data acquisition system for an SFV in high altitude conditions.
- Implementation of regularization techniques to reduce or eliminate variables that are not significant for the system, reducing the computational cost.
- With the eXtreme Gradient Boosting method with a small hyperparameter (n_estimator), high precision is obtained that, however, does not comply with the validation of the results, on the other hand, with a high hyperparameter (n_estimator), a precision close to the previous one described is obtained. but that complies with the validations of the results.

II. Proposed Method



A. Data collection

The measurements were obtained according to the IEC 61724-2017 standard, in the city of Juliaca, San Roman province, Puno department in Peru; which is located at a height of 3827 msnm whose coordinates are: 15° 29′27″ S 70° 07′ 37″ W, considering the values of AC voltage, AC current, active power, apparent power, power, reactive, frequency, power factor, total energy, daily energy, DC voltage, DC current, DC power, irradiance, ambient temperature, module temperature [16]. The record started from April to August 2019 in 1 minute intervals 24 hours a day, obtaining a number of 331157 values. The system consists of 12 photovoltaic modules of 60 cells each, a STECAGRID 3010 DE 3kW single-phase power inverter, the instruments used for monitoring were; PT1000 temperature sensor, SENTRON PAC 3200 which is a power quality monitor, a calibrated 15x20 cm photovoltaic cell, all these elements connected through the MODBUS industrial communication protocol.

B. Ridge

In a Ridge-type linear regression, it is expressed by (1).

$$E = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \qquad (1)$$

01

For learning, a Gaussian distribution will have to be made by its mean, this is represented by $N(\mu, \sigma^2), i.e., X \sim N(\mu, \sigma^2)$ where X an input matrix, we have the probability of X_i shown in (2).

$$P(X_i) = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}} \qquad (2)$$

For each occurrence of X_i there is a joint probability that is expressed by (3)

$$p(x_1, x_2, \dots x_N) = \prod_{i=1}^{N} \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}} \qquad (3)$$

The line that will contain the best fit for the regression is shown in (4)

$$P(X|\mu) = p(x_1, x_2, \dots x_N) = \prod_{i=1}^{N} \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}} \qquad (4)$$

To improve the line of fit, the natural logarithm is considered in the probability function, to later make a difference and equal it to 0 as shown in (5).

$$ln\big(P(X|\mu)\big) = ln\big(p(x_1, x_2, \dots x_N)\big) = \qquad (5)$$

$$ln\prod_{i=1}^{N}\frac{1}{2\pi\sigma^2}e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}}=\sum_{i=1}^{N}ln\left(\frac{1}{2\pi\sigma^2}e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}}\right) \quad (6)$$

$$\sum_{i=1}^{N}ln\left(\frac{1}{2\pi\sigma^2}\right)-\sum_{i=1}^{N}\left|\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right. \quad (7)$$

$$\frac{\partial ln\left(P(X|\mu)\right)}{\partial\mu}=\frac{\partial\sum_{i=1}^{N}ln\left(\frac{1}{2\pi\sigma^2}\right)}{\partial\mu}-\frac{\partial\sum_{i=1}^{N}\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}}{\partial\mu} \quad (8)$$

$$=0+\sum_{i=1}^{N}\frac{(x_i-\mu)}{\sigma^2}=\sum_{i=1}^{N}\frac{(x_i-\mu)}{\sigma^2} \quad (9)$$

$$\frac{\partial ln\left(P(X|\mu)\right)}{\partial\mu}=\sum_{i=1}^{N}\frac{(x_i-\mu)}{\sigma^2}=0\Rightarrow\mu=\frac{\sum_{i=1}^{N}x_i}{N} \quad (10)$$

Considering that probability (likelihood) L is equal to the error function and also the Gaussian distribution with mean transposition $(w)*X$ and variance σ^2 is shown in (11).

$$y\sim N(\omega^T X,\sigma^2) \quad o \quad y=\omega^T X+\varepsilon \quad (11)$$

If there will be outliers, a regularization method will be used in the data in order to modify the cost function and also penalize large weights as shown in (12).

$$J_{RIDGE}=\sum_{i=1}^{N}(y_i-\hat{y}_i)^2+\lambda|w|^2 \quad (12)$$

Where: $|w|^2=w^Tw=w_1^2+w_2^2+\cdots+w_D^2$

You get two probabilities: Posterior:

$$P(Y|X,w)=\prod_{i=1}^{N}\frac{1}{2\pi\sigma^2}exp\left(-\frac{1}{2\sigma^2}(y_n-w^Tx_n)^2\right) \quad (13)$$

A priori:

$$P(w)=\frac{\lambda}{\sqrt{2\pi}}exp\left(-\frac{\lambda}{2}w^Tw\right) \quad (14)$$

C. Ridge-Bayesian

Bayesian regression techniques can be used to include regularization parameters in the estimation procedure: the regularization parameter is not set in a hard sense but tuned to the data at hand [18]. BayesianRidge estimates a probabilistic model of the regression problem [14]. Applying Bayes

$$exp(J)=\prod_{n=1}^{N}exp(-(y_n-w^Tx_n)^2)exp(\lambda w^Tw) \quad (15)$$

Bayes application:

$$J=(Y-X_W)(Y-X_W)^T+\lambda w^Tw$$
$$=Y^TT-2Y^TXw+w^TX^TXw+\lambda w^Tw \quad (16)$$

To minimize J, we use $\frac{\partial J}{\partial w}$. Therefore,

$$-2X^T+2X^TXw+2\lambda w=0$$

So $(X^TX+\lambda I)w=X^TY$ or $w=(X^TY)$

Since $P(w)$ is Gaussian and is close to 0 it follows that the weights will be small.

C. Lasso

D So for Lasso, we have

$$J_{LASSO}=\sum_{i=1}^{N}(y_i-\hat{y}_i)^2+\lambda\|w\| \quad (17)$$

Maximizing the likelihood

$$P(Y|X,w)=\prod_{i=1}^{N}\frac{1}{2\pi\sigma^2}exp\left(-\frac{1}{2\sigma^2}(y_n-w^Tx_n)^2\right) \quad (18)$$

This is given by:

$$P(w)=\frac{\lambda}{2}exp(-\lambda|w|) \quad (19)$$

Then $J=(Y-X_W)(Y-X_W)^T+\lambda|w|$

And

$$\frac{\partial J}{\partial w}=-2X^TY+2X^TY+2X^TX_W+\lambda sign(w)=0 \quad (20)$$

Considering that

$$sign(w)=1\text{ If }x>0\text{ and }-1\text{ if }x<0\text{ and }0\text{ if }x=0 \quad (21)$$

C. eXtreme Gradient Boosting

XGBoost is an algorithm that is used in Machine Learning, that is, with tabular data for their prediction or classification [19]. Its operating principle is the decision trees that graphically represent the possible solutions. They then go through a Bagging process that combines various predictions from various decision trees using a majority system. They then go

through a Randon Forest process in which a set of characteristics relevant to the system are randomly selected. Later they go through a Boosting process to minimize the errors of the models obtained previously through a gradient boosting or descending gradient process [17]. So you get an optimized gradient increase algorithm through parallel processing, tree pruning, missing value handling, and regularization to avoid overfitting and biases.

The objective function can be denoted through the two terms of the (22) where the first represents the loss of training and the second the term related to regularization

$$obj(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{k} \Omega(f_k) \quad (22)$$

The and terms represent the current and predicted values and the l term denotes the error between them. The term Ω is the regularization function to avoid overfitting for the k decision trees. In general, the error can be determined through:

$$l(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2 \quad (23)$$

The decision tree can be represented from its complexity:

$$f_i(x) = \omega_{q(x)}, \omega \epsilon R^T, q: R^d \rightarrow \{1,2, \dots T\} \quad (24)$$

where T is the total number of sheets, w is a subvector of sheets, and the function q assigns each data point to the corresponding sheet.

Regularization can be expressed as:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{i=1}^{T} \omega_i^2 \quad (25)$$

where γ and λ are the coefficients associated with the regular terms. The addition model and the forward steps algorithm provide the training and optimization functions for the XGBoost [15] model.

The arithmetic process of the model can be expressed as a function of the predicted values as:

$$\hat{y}_i^t = \hat{y}_i^{t-1} + f_t(x_i) \quad (26)$$

So finally the objective function can be rewritten as:

$$obj(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{t-1} + f_t(x_i)) + \sum_{i=1}^{t} \Omega(f_i) \quad (27)$$

III. Performance Evaluation

A. Data collection

As a first stage, the preprocessing of the data was carried out, reducing its number from 331157 to 123120 because many of the values were obtained at night, considering them null values; as well as values that were not correctly registered by the acquisition system, the study was conducted between the fall and winter seasons. Table I and Table II describes the statistics of the variables to be used: median, standard deviation, values: maximums, minimums, and interquartile ranges.

For all the models proposed in this article, the dependent variable is: 'Active Power' and the independent variables are: 'AC voltage', 'AC current', 'Apparent power', 'Reactive power', 'Frequency', 'Power factor power ','Total energy',' Daily energy ',' DC voltage ',' DC current ',' DC power ',' Irradiance ',' Modulo temp. ',' Ambient temp'.

TABLE I
STATISTICIAN

	AC Coltage	AC Current	Active Power	Apparent Power	Reactive Power	Frequency	Power Factor
count	123120	123120	123120	123120	123120	123120	123120
mean	235,45	6,9652971	1621,9653	1643,2554	219,88262	60,002591	0,9513516
std	2,9435	2,9305708	708,22952	696,3666	66,593441	0,0459694	0,1892854
min	223,9	0,58	0	135	-843,9	59,5	-0,99
25%	233,5	4,639	1071,1	1091	196,2	60	0,983
50%	235,4	7,564	1764	1779,55	228,4	60	0,991
75%	237,6	9,43	2219,3	2232,7	256,2	60	0,994
max	247,9	12,416	2879,2	2898	485,1	60,5	0,998

TABLE II
STATISTICIAN

	Total Energy	Daily Energy	DC Voltage	DC Current	DC Power	Irradiance	Module temp	Ambient temp
count	123120	123120	123120	12312	123120	123120	123120	123120
mean	5233,49	127,68286	334,80517	5,5576	1831,37	669,007	35,109	16,6106
std	1013,12	86,399295	17,333536	2,3894	737,254	291,940	11,2739	3,77327
min	3894,3	0,000209	220,8	0	0	0	2,4	-2
25%	4184,7	56,665777	321,9	3,62	1261,43	432	27,6	14,5
50%	5910,3	113,42258	332,7	5,89	1973,63	706	37	17,4
75%	6175,4	190,36942	346	7,65	2450,60	926	44,2	19,4
max	6427,6	342,90575	420,8	10,78	3142,27	1522	60,3	27,7

C. Results - Multiparametric Regression with Regularizacion Shrinkage

The data were divided into a set of training 98496 (80%) and test 24624 (20%). To randomize the training and test values, the best seed was searched, which is: 8849 as shown in Fig. 1.

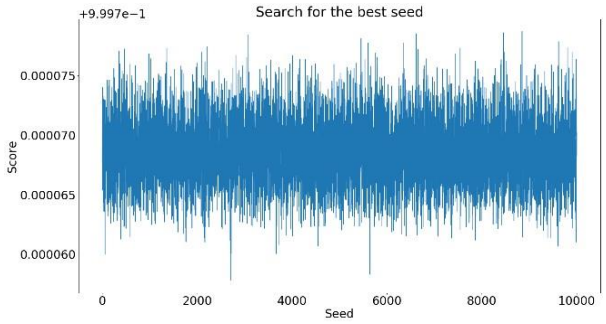


Fig. 1 Seed.

To obtain the hyperparameter alpha, cross-validation was used as shown in Fig. 2, the best value is 0.010.

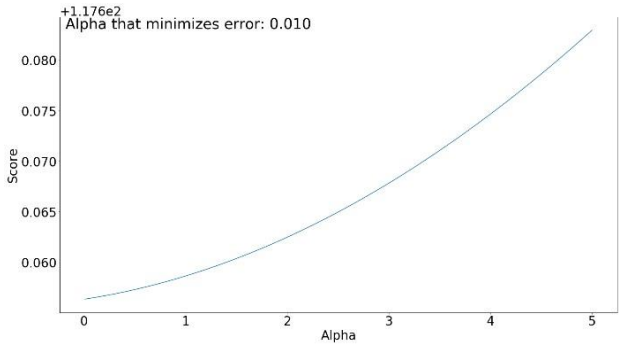


Fig. 2 Alpha.

To optimize the XGBoost model, the hyperparameter called n_estimators (Number of trees with increasing gradient) was modified, which improves the learning of the model and its precision. Fig. 3 and Fig. 4 shows the values of the hyperparameter n_estimators with values of 10 and 500.

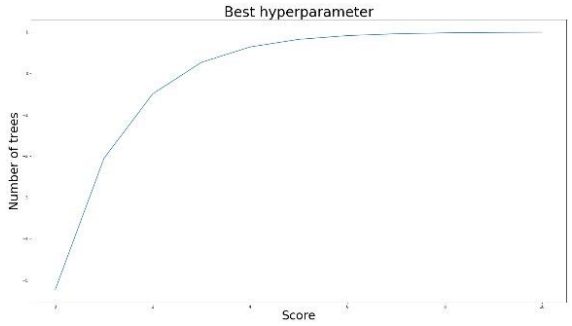


Fig. 3 Value n_estimators=10.

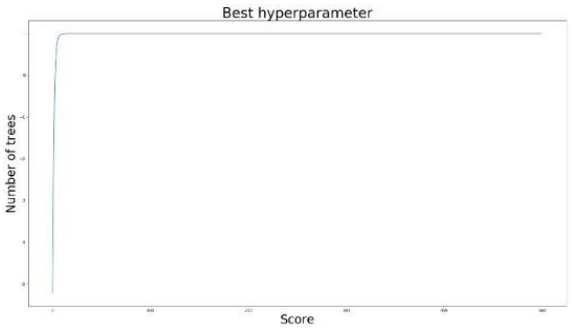


Fig. 4 Value n_estimators=500.

Table III summarizes the results of the 5 proposed models. The first method without the selection of OLS variables, which is a simple regression, was used as a reference point to compare the results obtained with the proposed models.

TABLE III
RESULTS OBTAINED

	No variable selection	Shinkage regularization			XGBoost	
Parameters	OLS	Lasso	Ridge	Ridge Bayesiano	n_estimators= 10	n_estimators= 500
Mean absolute error R	6,038018661	6,077300352	6,037967460	6,037799833	46,421226308	2,769478326
Mean square error R2	11,678861903	11,666872778	11,678845879	11,678793457	51,123567045	4,752143588
Determination coefficient	0,999726436	0,999726997	0,999726436	0,999726439	0,994738850	0,999954541
Adjusted coefficient of determination	0,999726405	0,999726966	0,999726405	0,999726408	0,994738251	0,999954536
Training time	0,0369	15,4285	0,0141	0,0997	0,4595	18,1727
Test time	0,0008	0,0007	0,0007	0,0011	0,0104	0,1759

From Table III it is observed that the mean absolute error for all the regularization methods is reduced by 0.01% while for XGBoost with n = 500, it is reduced by 54%. In the same way, the least-squares error for XGboost with n = 500 is reduced by 60%. The adjusted coefficient of determination for the regularization methods increases by less than 0.00001 while for XGBoost with n_estimators = 500 it increases by 0.02%. As the theory indicates, the training time increases, Ridge being the one that suffers the smallest increase in time with 61%. The test times with the least increase are Lasso with 3.12% while XGBoost with n_estimators = 500 has a real-time of less than 2 tenths of a second. In all cases, the coefficients of determination are greater than 99.47%.

IV. VALIDATION OF THE RESULTS

The results obtained in the five proposed methods give us values of determination coefficients higher than 99.47%, but these values only have mathematical significance. In order for the proposed models to adjust to the values obtained in the real world, the validation of each model was carried out through the fulfillment of the following five conditions:

A. Linearity

The real values vs the predicted values were plotted to find a linear relationship, this is determined through a scatter diagram to demonstrate the linear relationship between the real and predicted data as shown in Fig. 5 and Fig. 6.

A.1. Shrinkage regularization

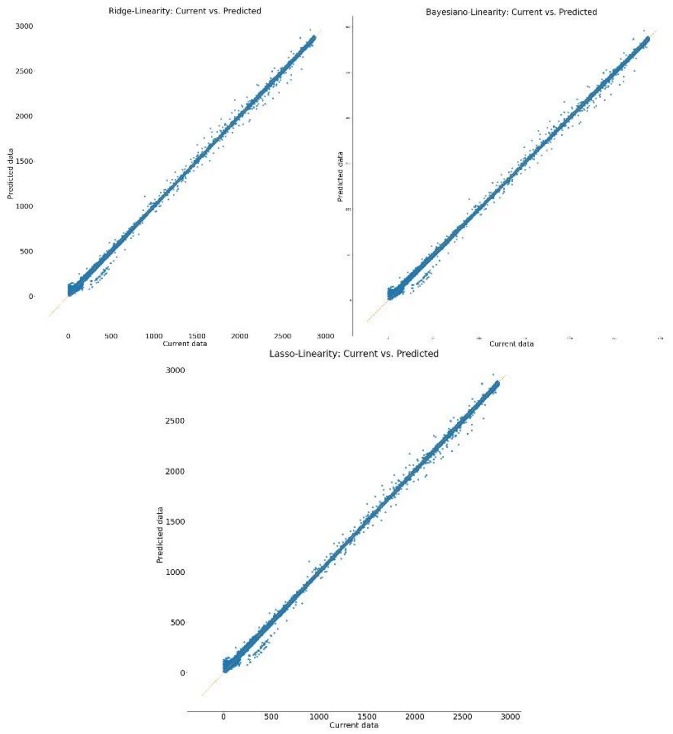


Fig. 5 Linearity: Lasso, Ridge, and Ridge Bayesian.

A.2. Extreme Gradient Boosting regularization

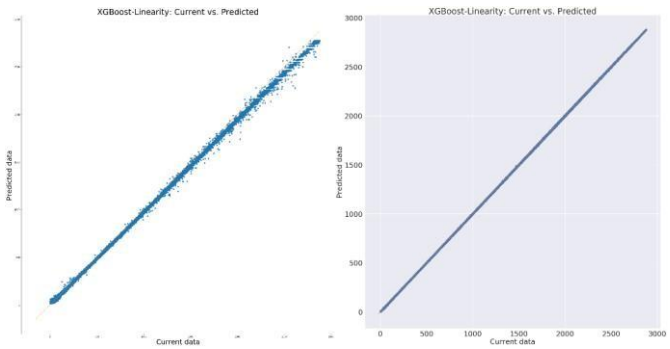


Fig. 6 Linearity: n_estimators=10 and n_estimators=500.

B. Normality of error terms

Through the normality of the error terms, we were able to estimate the confidence intervals for both the regression coefficients and the predicted values, this is shown through the histograms and probability graphs of Fig. 7 and Fig. 8.

B.1. Shrinkage regularization

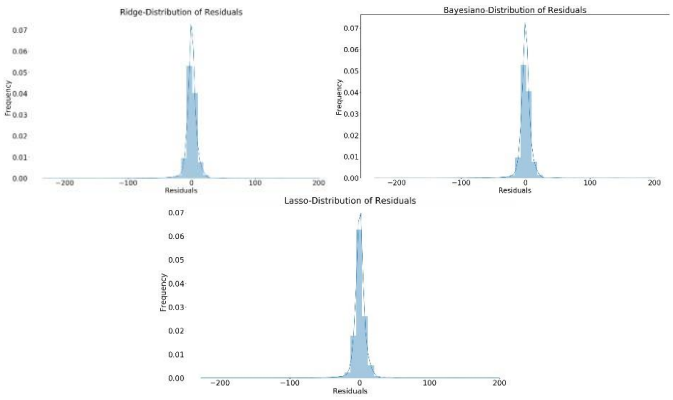


Fig. 7 Normality of error terms: Lasso, Ridge, and Ridge Bayesian.

B.2. Extreme Gradient Boosting regularization

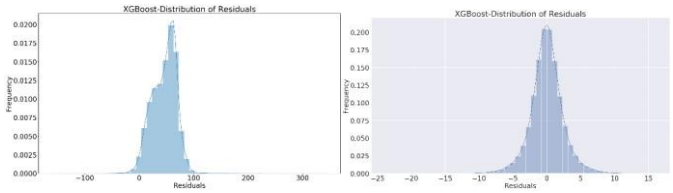


Fig. 8 Normality of error terms: n_estimators=10 and n_estimators=500.

C. No autocorrelation of the error terms

To determine if the model does not capture relevant information, the autocorrelation of error terms is used. For this, the Durbin Watson test was used, which represents a systematic bias below or above the prediction. This value is considered a positive autocorrelation if the result is from 0 to 2 and if the result is from 2 to 4 it is considered a negative correlation.

C.1. Shrinkage regularization

- OLS: 2.003431324284588
- Bayesian Ridge: 2.003429858272002
- Lasso: 2.003505817947447
- Ridge: 2.0034309813174658

C.2. Extreme Gradient Boosting regularization

- n_estimators=10: 0.376042222964736
- n_estimators=500: 2.006383323872818

D. Homocedasticity

To avoid that the model does not assign an excessive weight to a subset of values whose variance of the error is greater than most of the data, the error between the real values and the model must have the same variance. To check this, we use the residual graphs shown in Fig. 9 and Fig. 10.

D.1. Shrinkage regularization

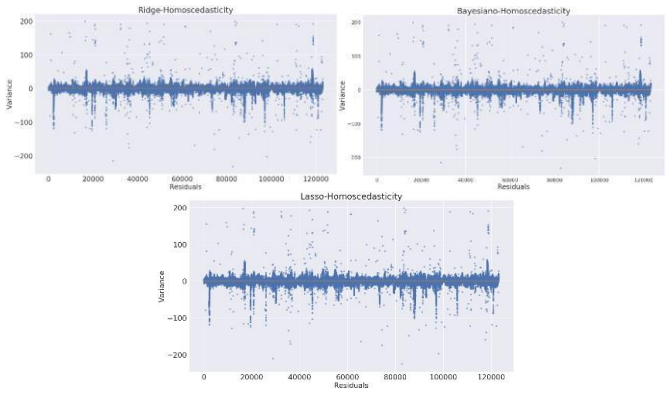


Fig. 9 Homocedasticity: Lasso, Ridge, and Ridge Bayesian.

D.2. Extreme Gradient Boosting regularization

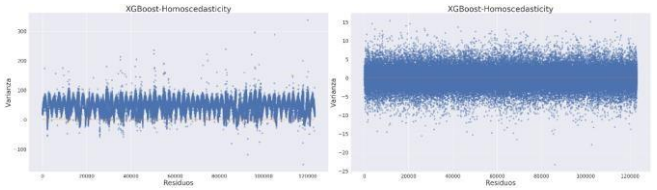


Fig. 10 Homocedasticity: n_estimators=10 and n_estimators=500.

E. Non-multicollinearity between Predictors - CORRELATION

The independent variables (predictors) should not be correlated with each other, since they would cause problems in the interpretation of the coefficients, as well as the error that each one of them provides. To determine Nonmulticollinearity a correlation heat map was used as shown in Fig. 11 and Fig. 12.

E.1. Shrinkage regularization

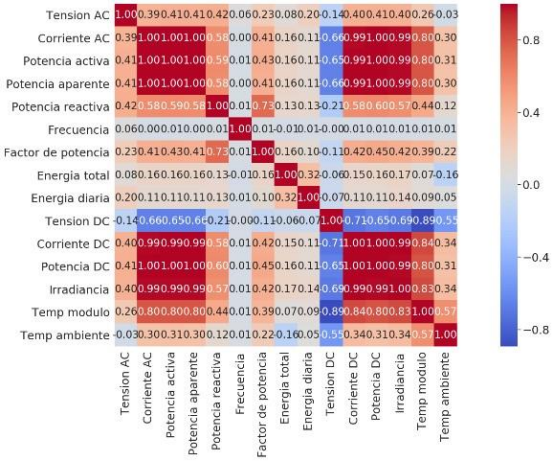


Fig. 11 Correlation Shrinkage.

E.2. Extreme Gradient Boosting regularization

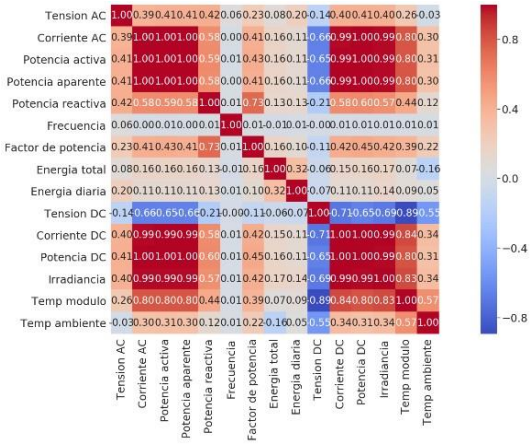


Fig. 12 Correlation XGBoost.

From the conditions previously analyzed for the five proposed models, we note that XGBoost with n_estimator = 10, does not meet any of the validation conditions so that although it is true, the result obtained has mathematical significance, it does not reflect the real-life relationship between variables, so their results are discarded.

V. Conclusions

This article presents the modeling and prediction of a multivariable Photovoltaic System using five multiparametric regression models applying regularization techniques and eXtreme Gradient Boosting to optimize the operation of the photovoltaic system located in a city at 3827 msnm.

The SFV and its data acquisition system were implemented, with industrial components complying with the IEC 61724 -2017 standard. Likewise, three regression methods were implemented using Shrinkage regularization: Lasso, Ridge, and Bayesian Ridge, and two that use XGBoost: n_stimator = 10 and with n_estimator = 500.As result, we have that the mean absolute error for all regularization methods is reduced by 0.01% while for XGBoost with n_estimator = 500, it is reduced by 54%. In the same way, the least-squares error for XGboost with n = _estimator 500 is reduced by 60%. The adjusted coefficient of determination for the regularization methods increases by less than 0.00001% while for XGBoost with n_estimator = 500 it increases by 0.02%. Training time increased in all cases, Ridge being the one with the lowest increase with 61%. Test times are less than 1 second.

Finally, the results were validated so that the proposed models have significance in the real world: linearity, normality of error terms, no autocorrelation of error terms, homoscedasticity, and no multicollinearity between predictors. We note that XGBoost with n_estimator = 10, does not meet any of the validation conditions, so although the result obtained

has mathematical significance, it does not reflect the relationship between the variables in real life, so its results must be discarded. As future work, predictions can be improved using data imputation techniques.

REFERENCES

[1] S. Albahli, M. Shiraz and N. Ayub, "Electricity Price Forecasting for Cloud Computing Using an Enhanced Machine Learning Model," in IEEE Access, vol. 8, pp. 200971-200981, 2020, doi: 10.1109/ACCESS.2020.3035328.

[2] W. Guo, X. Jiang and L. Che, "Short-Term Photovoltaic PowerForecasting based on Machine Learning," 2019 IEEE 3rd International Electrical and Energy Conference (CIEEC), Beijing, China, 2019, pp. 1276-1280, doi: 10.1109/CIEEC47146.2019.CIEEC-2019466.

[3] C. -H. Hu and Y. -L. Chen, "Forecasting Time Series for Electricity Consumption Data Using Dynamic Weight Ensemble Model," 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 2020, pp. 1-9, doi: 10.1109/IJCNN48605.2020.9207108.

[4] P. W. Khan and Y. -C. Byun, "Genetic Algorithm Based Optimized Feature Engineering and Hybrid Machine Learning for Effective Energy Consumption Prediction," in IEEE Access, vol. 8, pp. 196274-196286, 2020, doi: 10.1109/ACCESS.2020.3034101.

[5] S. Park, J. Moon and E. Hwang, "2-Stage Electric Load Forecasting Scheme for Day-Ahead CCHP Scheduling," 2019 IEEE 13th International Conference on Power Electronics and Drive Systems (PEDS), Toulouse, France, 2019, pp. 1-4, doi: 10.1109/PEDS44367.2019.8998960.

[6] P. Lojano, J. Cabrera, A. Lojano, D. Morales and D. Icaza, "Voltage Data Collection using Arduino and Matlab of a Photovoltaic Wind Power System in the Locality of Tarqui the Cuenca Ecuador," 2019 8th International Conference on Renewable Energy Research and Applications (ICRERA), Brasov, Romania, 2019, pp. 582-586, doi: 10.1109/ICRERA47325.2019.8997035.

[7] J. Choi, M. Choi, Y. Shin and I. Lee, "Design of Web-based Monitoring System for Solar Photovoltaic Power Plants," 2020 International Conference on Information Networking (ICOIN), Barcelona, Spain, 2020, pp. 784-786, doi: 10.1109/ICOIN48656.2020.9016482.

[8] T. Ku, W. Park and H. Choi, "Energy Big Data Life Cycle Mechanism for Renewable Energy System," IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Paris, France, 2019, pp. 1067-1068, doi: 10.1109/INFCOMW.2019.8845036.

[9] I. P. Panapakidis, A. S. Bouhouras and G. C. Christoforidis, "A missing data treatment method for photovoltaic installations," 2018 IEEE International Energy Conference (ENERGYCON), Limassol, Cyprus, 2018, pp. 1-6, doi: 10.1109/ENERGYCON.2018.8398780.

[10] Y. Cao, J. A. Magerko, R. Serna, S. Qin, R. C. N. Pilawa-Podgurski and P. T. Krein, "One Year Submillisecond Fast Solar Database: Collection, Investigation, and Application," 2019 IEEE Energy Conversion Congress and Exposition (ECCE), Baltimore, MD, USA, 2019, pp. 2047-2053, doi: 10.1109/ECCE.2019.8913161.

[11] C. C. Onatu and N. E. Mabunda, "Microcontroller Based Data Acquisition System for the Study of Shadowing on a Photo Voltaic Modules," 2019 IEEE AFRICON, Accra, Ghana, 2019, pp. 1-5, doi: 10.1109/AFRICON46755.2019.9133739.

[12] S. Huaquipaco et al., "Solar library," in Proceedings of the ISES Solar World Congress 2019 and IEA SHC International Conference on Solar Heating and Cooling for Buildings and Industry 2019, 2020, pp. 2507–2513, doi: 10.18086/swc.2019.52.01.

[13] S. Huaquipaco et al., "Photovoltaic charger system for mobile devices using quick charge 3.0 technology.," Proc. LACCEI Int. Multi-conference Eng. Educ. Technol., pp. 27–30, 2020, doi: 10.18687/LACCEI2020.1.1.341.

[14] "Bayesian Ridge Regression" scikit. [Online]. Available: https://scikit-learn.org/stable/modules/linear_model.html#bayesian-ridge-regression. [Accessed: 15-Jun-2021].

[15] T. Lumley, "Multiple Imputation Through XGBoost," no. 2012, 2021.

[16] S. Sairam, S. Srinivasan, G. Marafioti, B. Subathra, G. Mathisen, and K. Bekiroglu, "Explainable Incipient Fault Detection Systems for Photovoltaic Panels," pp. 1–8, 2020, [Online]. Available: http://arxiv.org/abs/2011.09843.

[17] E. Zolotareva, "Aiding Long-Term Investment Decisions with XGBoost Machine Learning Model," pp. 1–29.

[18] X. Wu and Q. Cheng, "Top-k Regularization for Supervised Feature Selection," pp. 1–12, 2021, [Online]. Available: http://arxiv.org/abs/2106.02197.

[19] L. de Vito, "LinXGBoost: Extension of XGBoost to Generalized Local Linear Models," 2017, [Online]. Available: http://arxiv.org/abs/1710.03634.