COBEM 2021 Florianópolis - Brasil
ABCM Associação Brasileira de Engenharia e Ciências Mecânicas
26th ABCM International Congress of Mechanical Engineering
November 22-26, 2021. Florianópolis, SC, Brazil

COB-2021-1902
EVALUATION OF MACHINE LEARNING MODELS FOR VERY SHORT-TERM SOLAR IRRADIANCE FORECASTING: A CASE STUDY FOR PETROLINA / PE, NORTHEASTERN BRAZIL.

Nadja Gomes de Oliveira
Mechanical Engineering Department, Technology Center, Federal University of Ceará, Fortaleza, CE 60020-181, Brazil
nadjagomesde@uoguelph.ca

Paulo Alexandre Costa Rocha
Mechanical Engineering Department, Technology Center, Federal University of Ceará, Fortaleza, CE 60020-181, Brazil
paulo.rocha@ufc.br

Mostafa H. Elsharqawy
School of Engineering, University of Guelph, Guelph, Ontario, N1G 2W1, Canada
melsharq@uoguelph.ca

Abstract. *This work uses the SONDA network irradiance data to forecast global horizontal and direct normal irradiances (GHI and DNI) in very short-term intra-hour for 5 minutes resolution during the period of four years for one solarimetric station in the northeast of Brazil, Petrolina/PE. Four different machine learning models were tested, namely Least Absolute Shrinkage and Selection Operator (LASSO), k-nearest-neighbors (kNN), extreme gradient boosting (XGBoost) and an ensemble combination to form a final forecast (Ensemble with Ridge Regression). Their performance was compared with the RMSE and Forecast Skill relative to the persistence model. Results show that the machine learning models achieve significant forecast improvements over the reference model. In addition, the Ensemble with Ridge Regression and XGBoost models have rarely been used for short-term solar forecasting. This framework can be used to select appropriate machine learning approaches for short-term solar power forecasting and the simulation results can be used as a baseline for comparison. The four methods presented similar behavior for raw and normalized variables, with the RMSE values ranging between 72.85 W/m² and 76.12 W/m² for GHI and values between 104.22 W/m² and 104.66 W/m² for DNI. It is worth to mention that the Ensemble with Ridge model outperformed all the other methods for GHI, obtaining RMSE values between 72.85 W/m² and 74.02 W/m² and for DNI, the XGBoost model outperformed obtaining RMSE values between 104.22 W/m² and 104.66 W/m². Regarding the Forecast Skill (FS), the Ensemble with Ridge model performs 0.60% better than the XGBoost for GHI, although the XGBoost performs 1.58% better than the Ensemble with Ridge model for DNI. Simulation results for FS also showed that the use of the clear-sky index for GHI increased most of the model's performance between 0.70% to 1.15% for GHI, excepted for the LASSO, and between 0.27% and 0.36% for DNI with the LASSO and XGBoost, although a decreased between 0.67% and 0.83% was observed for the kNN and the Ensemble with Ridge models.*

Keywords: *Machine learning, global solar irradiance, direct normal irradiance, intra-hour forecasting.*

1. INTRODUCTION

Technical and financial risks of expanding solar-energy conversion technologies for electricity production can be alleviated through a better comprehension of available solar-resource analysis and forecasting methods applicable to each solar-energy conversion technology. Authentic solar-energy forecasting and resource assessment can reduce the risk in selecting the project location, designing the appropriate solar-energy conversion technology, and operating new sources of solar-power generation integrated into the electricity grid (Kleissl, 2013).

Dependency on meteorological conditions causes renewable energy resources to be inconsistent. Under this constraint, reliable solar irradiance forecast on different time horizons is essential for developing and utilizing solar energy-based systems. Short-term and intra-hour solar forecasts are particularly useful for power plant operations, grid balancing, real-time unit dispatching, automatic generation control (AGC) and trading.

Solar forecasting is therefore an enabling technology for the integration of ever-increasing levels of solar penetration into the grid because it improves the quality of the energy delivered to the grid and reduces the ancillary costs associated with weather dependency. The combination of these two factors (better energy quality through information that is capable of lowering integration and operational costs) has been the driving motivation for the development of a complex field of

research that aims at producing better solar forecasting capabilities for the solar resource at the ground level and for the power output from different solar technologies that depend on the variable irradiance at the ground level. Solar, wind and load forecasting have become integral parts of the so-called smart grid concept (Inman et al., 2013).

Continuous solar irradiance data acquisition is an important component of time series analysis as we need to trace back several steps in time in order to forecast the solar irradiance at the next time step. During night time, solar irradiance is zero, a discontinuity that must be considered when forecasting solar irradiance during the period just after sunrise and before sunset. For locals without solar irradiance measurements, data may be inferred from clear sky models and other meteorological parameters. Weather patterns and their accompanying clouds are the most significant atmospheric phenomena affecting solar irradiance at the earth's surface (Brinsfield et al., 1984).

Targeting these two irradiance components is important: DNI is of particular interest to concentrating solar power (CSP) plants and installations that track the position of the sun; and both DNI and GHI can be used to estimate the plane-of-array irradiance on tilted/tracking PV panels. Intra-hour forecasts are relevant for optimal central plant operations and is an enabling technology for optimal dispatch of ancillary resources and storage systems. Moreover, mitigation measures for large drops in solar irradiance, such as demand response, storage and intra-hour scheduling can only be maximized with accurate and reliable intra-our forecast (Inman et al., 2013).

Solar forecasting techniques can be classified into three main categories: (1) numerical weather prediction (NWP); (2) image-based methods; and (3) statistical and machine learning (ML) methods. NWP studies the weather and generates forecasts of irradiance among hundreds of other meteorological parameters. NWP-based methods are well accepted for 6–48 hours-ahead forecasting. Image-based methods, on the other hand, use sky cameras, shadow cameras, or satellite images to predict solar irradiance, primarily by tracking and adverting clouds. Whereas sky/shadow cameras are used for intra-hour forecasting, satellite-based methods are advantageous for a forecast horizon ranging from 30 min to 6 h. Statistical and ML models use historical data to "train" themselves. The trained model then generates forecasts based on new values of the input variables. Statistical and ML methods are applicable for a wide range of temporal horizons, but mostly appear in hourly forecasting studies (Voyant et al., 2017).

The main objective of our work is to make a local evaluation with different machine learning models for very-short term solar forecasting and analyze the prediction accuracy of each model. We focused on intra-hour global horizontal irradiance (GHI) and direct normal irradiance (DNI) forecasts with forecasting time horizons on a timescale of 5 minutes.

2. DATA AND METHODOLOGY

2.1 Data

All the data were obtained from SONDA project (National Data Organization System, www.sonda.ccst.inpe.br). Although this project is still in progress linked to the area of research in climate and meteorology, there is an aspect of this project aimed at supporting activities in the area of renewable energies. The project aims the development of a complete, integrated, high-quality, and reliable data base that addresses the needs of sectors of society involved with research, development, planning and investment in the energy sector applications (Martins et al., 2006).

The forecasting models are trained for solar irradiance measurements (specifically, GHI and DNI) obtained in Petrolina, PE, Brazil, 09° 04' 08" S and 40° 19' 11" W, northeast of the country. The raw 1-min data was quality controlled to remove physically impossible values, averaged into 5 minutes bins of GHI and DNI directly from the raw data for four consecutive years: January 2013 to December 2016, and divided into three data sets: training, validation and testing.

The first dataset; denoted as training or historical dataset, the radiation itself is used to be predicted using endogenous models. The second dataset; denoted as optimization dataset, is used in the optimization algorithm to determine the several free parameters (explained below) in the forecasting model. The third dataset; the independent testing set, is used to assess the performance of the forecasting model. The three data sets were constructed by grouping disjointed subsets for each month, thus ensuring that all data sets are well representative of the irradiance data over the whole period.

According to Rocha et. al (2021), as a common practice according to the literature, chronologically sequential 70%-30% separation is usually used to include data from all seasons in the training and testing sets. Therefore, for all the applied models in this study, a unique separation of the training set occurred by splitting the four years into the first three years for training and the last year for testing.

Duffie and Beckman (2013) affirms that another common practice in solar energy is work with the clear sky index k_t instead of the original solar irradiance time series. The clear-sky index is defined as:

$$k_t(t) = \frac{I(t)}{I^{clr}(t)} \tag{1}$$

where I is the solar irradiance, GHI or DNI, and I^{clr} is the clear-sky irradiance computed following the algorithm given by Ineichen et al., 2008.

Comparing with the Coimbra et al., (2018), in this study the models selected used only endogenous inputs for generating the forecasts, including the zenith angle as a new auxiliary variable. In other words, the only inputs of the

models are the past solar irradiance data, so we obtained continuous and workable time series of GHI and DNI by applying the following rules to the data:

- Remove all of the data for a solar zenith angle inferior or equal to 85° to avoid the side effects of including the low accuracy of the solar measurements before sunrise and after sunset. Thus, the time series obtained do not contain null night values;
- The backward average for the clear-sky index time series;
- The lagged 5-min average values for the 5-min clear-sky index time series;

2.2 Data Pre-processing

According to (Kuhn and Kjell, 2013), transformations of predictor variables may be needed for several reasons. A few modeling techniques may have strict requirements, such as the predictors having an ordinary scale. In other circumstances, creating a good model can be complex owing specific characteristics of the data (e.g., outliers).

The most straightforward and common data transformation is to center scale the predictor variables. To center a predictor variable, the average predictor value is subtracted from all the values. As a result of centering, the predictor has a zero mean. Similarly, to scale the data, each value of the predictor variable is divided by its standard deviation. Scaling the data coerce the values to have a common standard deviation of one. These manipulations are generally used to improve the numerical stability of some calculations.

To administer this series of transformations to multiple data sets, the *caret* packaged used in R, created by Max Kuhn, has the ability to transform, center, scale, or impute values, as well as apply the spatial sign transformation and feature extraction. For each ML model the pre-processing was tested separately with and without the clear-sky index for GHI and DNI for the 5 minutes time horizon.

For the LASSO method, no predictor pre-processing was available in the library. However, with the XGBoost model, the pre-processing with *center* and *scale* increased the results for the normalized variables: GHI and DNI with clear-sky index, although decreased for the raw variables: GHI and DNI.

The kNN model was the most benefited with *center* and *scale* pre-processing implementations, as with the normalized variables the results increase significantly, and the results decreased for the raw variables. Therefore, we recommend centering and scaling the predictors for normalized variables prior to building XGBoost and kNN models.

2.3 Methods for producing deterministic point forecasts

The following gives a brief overview about the implemented regression models used to calculate the point forecasts for GHI and DNI, and in addition, forecasts are issued every 5 min for solar elevations larger than 5°.

In the first section, it will be shown the application of Lasso in this regression problem. The second and third section discusses how the XGBoost and kNN methods can be applied to solve the proposed regression analysis. Last section introduces an Ensemble with Ridge model which combines the forecasting results of the different statistical methods into more accurate predictions.

Overall, it must be emphasized that the focal point of the current work is the methodology used for developing the forecasting methods and therefore does not address more general questions such as the applicability of the methods to a wide range of solar variability microclimates or the effect of very long (multiple years) training data.

2.3.1 Persistence

The dull and smart persistence model where applied based on the variable and its accuracy is used as the baseline for the point forecasts. The persistence model is often used as a reference for determining the forecast skill. It is useful to know if a forecast model provides better results than any trivial reference model, which is the persistence model. The persistence model considers that the solar radiation at t + 1 is equal to the solar radiation at t. It assumes that the atmospheric conditions are stationary. It is also called the naive predictor.

$$Gt + 1 = Gt \tag{2}$$

Its accuracy decreases with the time horizon and is generally not adequate for more than 1 h. An improved version of this model is the smart persistence model. To take into account the fact that the apparent position of the sun is not identical between t and t + 1, the persistence model is corrected with a clear-sky ratio term and is then called smart persistence.

$$Gt + 1 = Gt \frac{G_{t+1}^{clear-sk}}{G_t^{clear-sky}} \tag{3}$$

2.3.2 The Lasso method

Ridge and Lasso regression are some of the simple techniques to reduce model complexity and prevent over-fitting which may result from simple linear regression. The lasso is a relatively recent alternative to ridge and not only helps in reducing overfitting, but it can help us in feature selection. The cost function for Lasso (least absolute shrinkage and selection operator) regression can be written:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p}|\beta_j| = RSS + \lambda \sum_{j=1}^{p}|\beta_j| \quad (4)$$

The lasso coefficients, $\hat{\beta}_{\lambda}^{L}$, minimize the quantity and as with ridge regression, the lasso shrinks the coefficient estimates towards zero. However, in the case of the lasso, the $\ell 1$ penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large. Hence, much like best subset selection, the lasso performs variable selection. As a result, models generated from the lasso are generally much easier to interpret than those produced by ridge regression (James et al., 2013). For further understanding on the shrinkage and selection procedures of the lasso, we refer the readers to consult Efron et al., 2004.

2.3.3 The kNN method

The kNN model uses the predictors introduced above to forecast GHI and DNI. This model is established on the similarity of the predictors at the forecasting issuing time to the predictors computed with the training data set. The kNN algorithm starts by computing the Euclidean distance for a new data set (i.e. testing or validation) and the features in the training set. This operation yields a distance vector for each feature. These are then combined into a single vector using a weighted sum, denoted as Ds, where the subscript s indicates the set of features used in the calculations.

The algorithm proceeds to extract the *k* instances in the training data with the lowest distance. To each instance there is an associated time stamp $\{\tau_1, \ldots, \tau_2\}$ in the training set. *k* forecasts are then computed using the GHI or DNI training data subsequent to these time stamps:

$$\hat{f}i(t+\Delta t) = \langle kt \rangle_{[t-\Delta t,t]}\, x\, \langle I^{clr} \rangle_{[t,t+\Delta t]}\,,\ \mathrm{i} = 1, \ldots, \mathrm{k} \quad (5)$$

from which the final point forecast is calculated as:

$$\hat{I}(t+\Delta t) = \frac{\sum_{i=1}^{k} \alpha i \hat{f}i\,(t+\Delta t)}{\sum_{i=1}^{k} \alpha i} \quad (6)$$

where the weights α are a function of the distance Ds

$$\alpha \mathrm{i} = \left(\frac{1-Ds,i}{maxDs-minDs}\right)^{n},\ \mathrm{i} = 1, \ldots, \mathrm{k} \quad (7)$$

and n is an adjustable positive integer parameter. The algorithm summarized above depends on several parameters:

1. The number of nearest neighbors, k ∈ {1,2, … , max k}, where maxk = 150 in this case;
2. The set of features S, i.e., which features are used in the search for the nearest neighbors;
3. The weights in the weighted sum Ds denoted as ωi;
4. The exponent n ∈ {1,2, … , 5} for the weights ai in Eq. (7);

The optimal model is determined by minimizing the forecast error for the validation data set:

$$argmin_{k,S,w,n} \sqrt{\frac{1}{n}\sum_{i}^{n}\left(\hat{I}\,(ti+\Delta t, k, S, \omega, n) - I(ti+\Delta t)\right)^2} \quad (8)$$

Further details about the optimization procedure and the respective optimal kNN models for GHI and DNI can be found in Pedro and Coimbra (2018).

2.3.4 The gradient boosting method

According to (Friedman, 2002) and (Hastie et al., 2009), boosting is a general approach that can be applied to many statistical learning methods for regression or classification. For regression problems, given a training data set, the goal is

to find a function f(x) such that a specified loss function is minimized. Boosting approximates f(x) by an additive expansion of the form:

$$\hat{f}(x) \sum_{m=0}^{M} \beta_m \, h(x, \theta_m) \tag{9}$$

where the functions $h(x, \theta_m)$ are simply functions of *x* parameterized by θ_m. $h(x, \theta_m)$ are called "base learners" or "weak learners" (Friedman, 2002). The expansion coefficients β_m and the parameters θ_m are fit to the training data in a forward "stage wise" manner (i.e. without adjusting the previous expansion coefficients and parameters of the base learners that have already been added). Here, we restrict the application of boosting to the context of regression trees (i.e. the base learner $h(x, \theta)$ is a tree $T(\theta)$). For that purpose, boosting builds an ensemble of trees iteratively in order to optimize a loss function ψ: the squared loss function $\psi\big(y, f(x)\big) = (y - f(x))^2$, in this case.

2.3.5 The XGBoost method

XGBoost is an algorithm based on sequential ensemble of decision trees, in which weak learners learn together to build a strong learner. Equation 10 shows the algorithm for the XGBoost method given by Munawar et al., 2019. Since the loss function *l(·)* for calculating residual is hard to optimize, the cost function $L^{(t)}$ is introduce as follows (Chen et al., 2016):

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, {y'}_i^{(t-1)} + f_i(x_i)\,) + \Omega(f_t) \tag{10}$$

where y_i is the index and t is time. y_i is the actual data and ${y'}_i^{(t)}$ is forecasted value. $f_i(x_i)$ is the model being updated itera-tively. $\Omega(f_t)$ is the penalty function. l(·) is the loss function.

2.4 Accuracy measures with error metrics

The error metrics used to evaluate the performance of the applied machine learning models are presented in this section.

2.4.1. Deterministic error metrics

When the goal is to measure the performance of a model for regression problems where we try to predict a numeric value, the residuals are important sources of information. Residuals are computed as the observed value minus the predicted value (i.e., $y_i - \hat{y}_i$).

Mean Absolute Error (MAE) gives the average magnitude of forecast errors and calculates the mean of the absolute differences between the predicted value, $\hat{y}_i$, and the real value, y_i, as indicated in the equation below:

$$MAE = \frac{1}{n} \sum_{j=1}^{n} |y_i - \hat{y}_i| \tag{11}$$

The bias or MBE is the average forecast error representing the systematic error of a forecast model to under or over forecast. As described below, a postprocessing of model output is useful to significantly reduce the bias.

$$MBE = \frac{1}{n} \sum_{j=1}^{n} (y_i - \hat{y}_i) \tag{12}$$

The root mean squared error (RMSE) is commonly used to evaluate models. RMSE is interpreted as how far, on average, the residuals are from zero and it emphasizes the larger errors.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\hat{I}_i - I_i)^2} \tag{13}$$

with N representing the number of samples of the testing set. A second metric, used to evaluate the improvement relative to the baseline model (here the persistence model), is the forecast skill (FS) which is, according to Dazhi Yang (2019), the best parameter to compare forecast models at the moment and is given by:

$$s = \left(1 - \frac{RMSE_m}{RMSE_0}\right) x\ 100[\%] \tag{14}$$

where $RMSE_0$ is the RMSE for the model described by Eq. (12) and $RMSE_m$ is the RMSE for the model m (here the Lasso, kNN, XGBoost or Ensemble with Ridge models).

3. RESULTS

3.1 Forecasting results

Three different machine learning models were applied to the testing set in order to evaluate their performance in an independent data set. The RMSE and the forecast skill for all the models are listed in "Table 1 and 2" for GHI and DNI, respectively. For a better visualization of the numerical data, these results are in Figures 3 and 4.

The results reveal that GHI is much easier to forecast than DNI. The RMSE for 5 minutes forecast horizon ranges between 72.85 and 76.12 W/m² for GHI, whereas for DNI the RMSE range at least 69.90% more than, from 104.22 to 108.33 W/m². The reduction in the RMSE translates into significant forecast skills that range between 19.57% and 23.02%, and between 24.73% and 27.40% for the GHI and DNI testing set, respectively.

Table 1. RMSE and forecast skills for the GHI forecast for the testing set, including the results using clear-sky index. RMSE values are in W/m² and the skill s is in percentage. t+5min is the time horizon for 5 minutes.

t+5min	Error Metrics	Persistence	LASSO	XGBoost	kNN	Ensemble_Ridge
GHI	RMSE	94.75	75.53	74.59	75.58	74.02
	s		20.28%	21.27%	20.23%	21.87%
*kt*GHI	RMSE	94.64	76.12	73.42	74.53	72.85
	s		19.57%	22.42%	21.25%	23.02%

Table 2. RMSE and forecast skills for the DNI forecast for the testing set, including the results using clear-sky index. RMSE values are in W/m² and the skill s is in percentage. t+5min is the time horizon for 5 minutes.

t+5min	Error Metrics	Persistence	LASSO	XGBoost	kNN	Ensemble_Ridge
DNI	RMSE	143.63	108.11	104.66	107.42	105.35
	s		24.73%	27.13%	25.21%	26.65%
*kt*DNI	RMSE	143.56	107.54	104.22	108.33	106.49
	s		25.09%	27.40%	24.54%	25.82%

Moreover, as shown by Figs. 3 and 4, a clear improvement is brought by the XGBoost method that includes the clear-sky index. It was noted that in a recent work (Pedro and Coimbra., 2018), the skill score's result for gradient boost (GB) comparing with the kNN method without images features in a 5-minute horizon had an improvement of 0.4% and 0.8% for GHI and DNI normalized variables, respectively. However, in this study, the XGBoost and Ensemble with Ridge methods increase the skill score obtained by the kNN algorithm by 1.17% and 1.77%, respectively, for GHI normalized variables and by 2.86% and 1.28%, respectively, for DNI normalized variables.

Figures 3 show that the optimized Ensemble with Ridge and XGBoost models reduce the RMSE with respect to the baseline persistence model for comparing with the other models. It is worth to mention that the Ensemble model outperformed all the other methods for GHI, obtaining RMSE values between 72.85 W/m² and 74.02 W/m² and for DNI, the XGBoost model outperformed obtaining RMSE values between 104.22 W/m² and 104.66 W/m². Regarding the Forecast Skill, the Ensemble with Ridge model performs 0.60% better than the XGBoost for GHI, although the XGBoost performs 1.58% better than the Ensemble with Ridge model for DNI.

Simulation results for FS also showed that the use of the clear-sky index for GHI increased most of the model's performance between 0.70% to 1.15% for GHI, excepted for the LASSO, and between 0.27% and 0.36% for DNI with the LASSO and XGBoost, although a decreased between 0.67% and 0.83% was observed for the kNN and the Ensemble with Ridge models when using the clear sky-index.

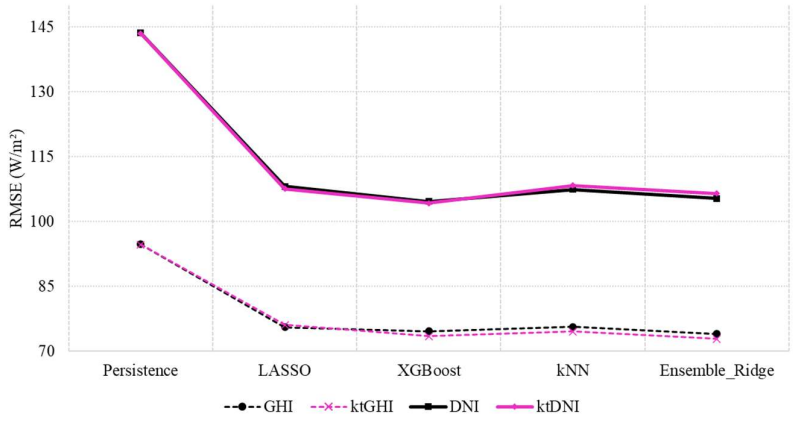

Figure 3. RMSE for GHI and DNI forecasts (testing set).

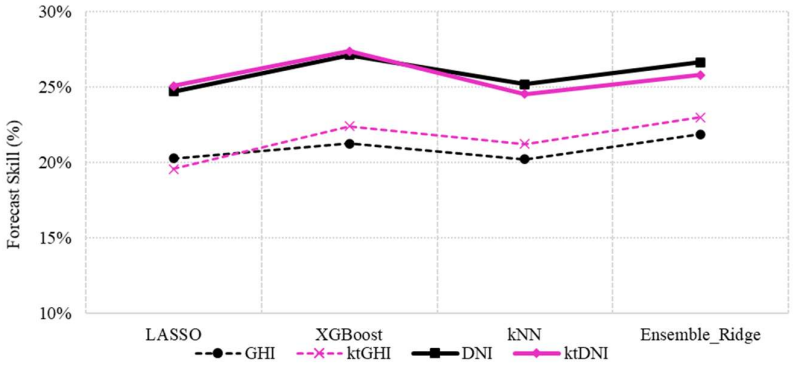

Figure 4. Forecast skill for GHI and DNI forecasts (testing set).

3.2 Overview of Machine Learning models results

All the four techniques (LASSO, XGBoost, kNN and Ensemble with Ridge for raw and normalized variables) clearly outperform the persistence model. This allegation is reinforced by the RMSE decrease results (around 20%) show in Figure 5. Also, the results indicate that, whatever the machine learning technique, the inclusion of clear-sky index does not bring a clear improvement for all the models.

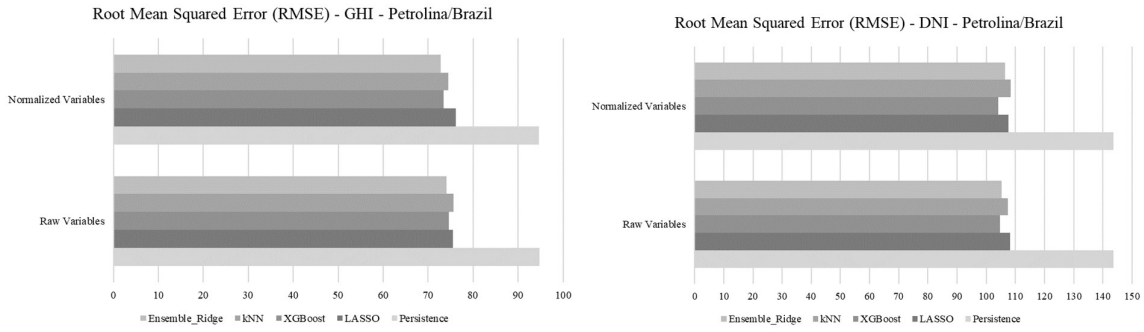

Figure 5. (a) RMSE for GHI and DNI forecasts results comparing with the Persistence model.

Lasso via hierarchical interactions was applied to select appropriate variables, especially because one of the main challenges of building a linear regression model is selecting the independent variables. When working with large datasets, it can be easy to end up with an overwhelming number of independent variables after cleaning the data and generating dummy variables. With LASSO regularization we can identify the most important independent variables and the results for GHI for raw and normalized variables are show in Figure 6 and for DNI normalized variables are show in Figure 7, for the case of 5 min ahead forecasting. It's clearly that the importance of the zenith angle decreases significantly with the use of clear-sky index in GHI and DNI.

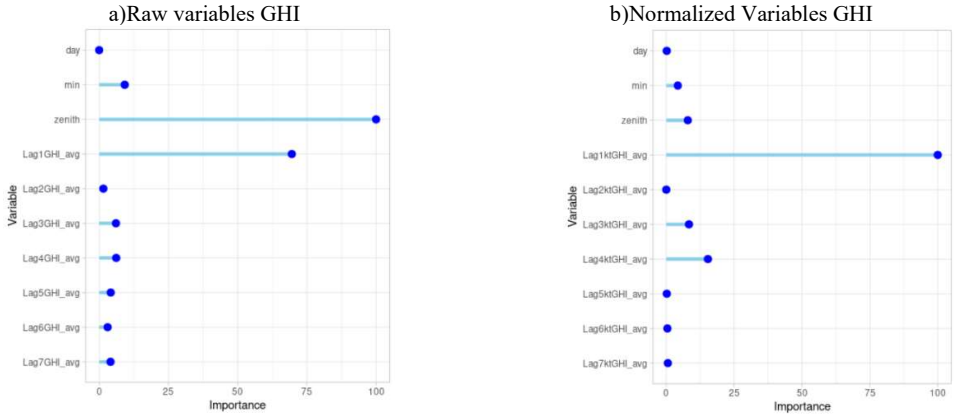

Figure 6. (a), (b) Importance variable with LASSO for GHI for raw and normalized variables respectively.

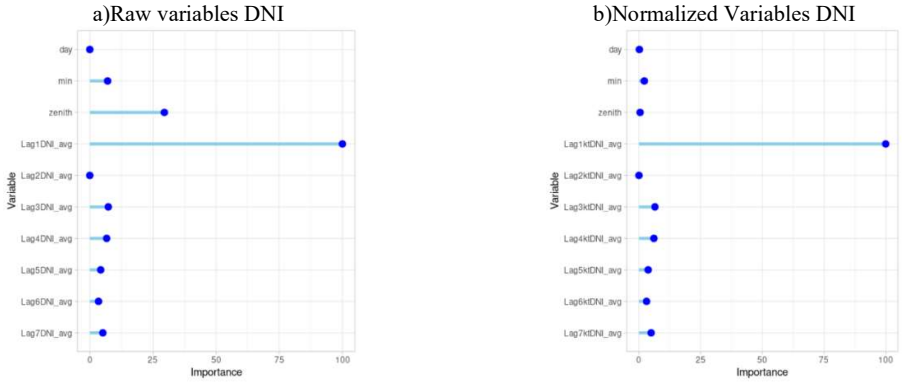

Figure 7. (a), (b) Importance variable with LASSO for DNI for raw and normalized variables respectively.

It can be seen from Figure 6 that the zenith angle account for a high contribution in the model. Through the analysis of model feature contribution, the effectiveness of the feature selection is verified, and it is confirmed that different algorithms have different emphasis on the raw and normalized variables.

In order to verify the predicted and measured forecasting for GHI and DNI for the model with the better results, XGBoost, in Figure 8 the predicted GHI histograms have a different shape as the scatter plot also indicates that the model tends to overestimate when the measure is small and underestimate it when it is larger. Although, we could notice that there isn't a significant difference in the histograms for DNI, but the model tends to underestimate when the measure is larger, as shown in Figure 9. This confirms that the proposed transformation did not statistically distort the response variable.

a)Raw variables GHI b)Normalized Variables GHI

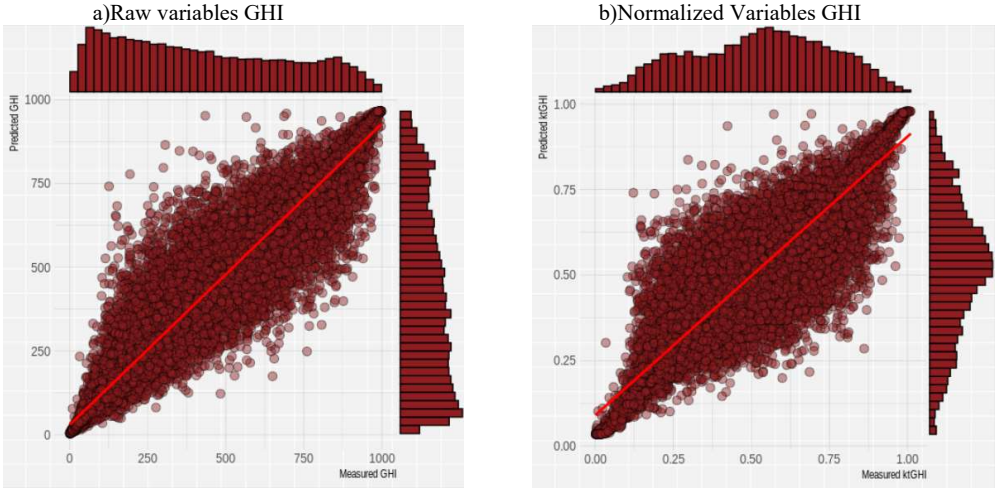

Figure 8. (a), (b) Scatter plot with XGBoost for GHI for raw and normalized variables respectively.

c)Raw variables DNI d)Normalized Variables DNI

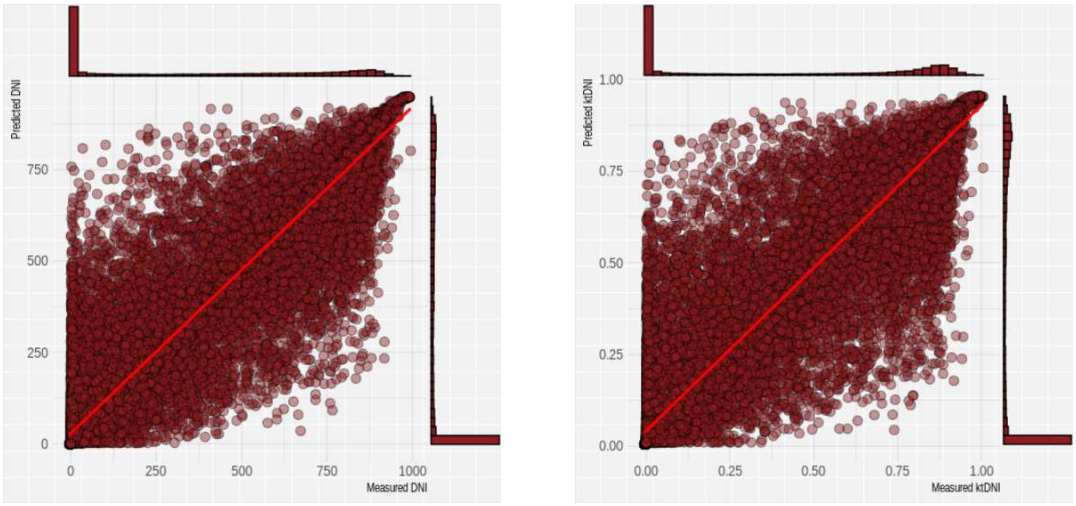

Figure 9. (c), (d) Scatter plot with XGBoost for DNI for raw and normalized variables respectively.

3.3 Conclusion

In this study, we implement a comparison of machine learning techniques for deterministic and probabilistic GHI and DNI forecasts using the National Environmental Data Organization System (SONDA) irradiance data for Petrolina/PE location in Brazil. We carried out a detailed analysis of the forecast performance for the techniques, which are capable of achieving relatively large forecasting skills with open sources and very limited investment in instrumentation. The present work focuses on the methodology used for developing the forecasting methods and therefore does not address more general questions such as the applicability of the methods to a wide range of solar variability microclimates or the effect of long-term training data. Nevertheless, the following conclusions can be drawn from this study:

- The Ensemble with Ridge method resulted in a clear improvement of the skill of the deterministic forecasts for GHI forecasting. Standards for this model range between 21.87% and 23.02% for the 5 min, and ranges between 25.82% and 26.65% for the DNI forecasting, whereas the corresponding values obtained with the XGBoost algorithm ranges between 21.27% and 22.42%, and 27.13% and 27.4%.
- Conversely, regarding the DNI forecasts, the XGBoost algorithm led to better results than the Ensemble with Ridge method especially with an improvement of 0.75%.
- The use of clear-sky index improves the Forecast Skill between 1.15% and 1.02% for GHI with XGBoost and kNN respectively, excepted for LASSO. For the DNI, the use of this index had a positive impact on the

Forecast Skill between 0.27% and 0.36% for XGBoost and LASSO respectively, and a negative impact for kNN and Ensemble with Ridge models.

- An increase, of at least 20%, on the RMSE for all the four techniques (LASSO, XGBoost, kNN and Ensemble with raw and normalized variables) clearly outperform the persistence model.

In conclusion, we can affirm that the simple Ensemble with Ridge method appears to be a viable technique for generating probabilistic forecasts if correctly optimized. Indeed, it has been demonstrated that the XGBoost method is efficient of achieve reliable GHI and DNI forecasts. Conversely, while the inclusion of clear-sky index in the models improves the forecast skill of GHI forecasts, excepted for LASSO, this inclusion (has unfortunately no expressive impacts for DNI forecasts. In this work, all the models were runed using R studio and the *caret* package. In future work we will increase the time horizons until 1 hour ahead in other different locations to see the effect with the 5 macroclimates in Brazil.

4. ACKNOWLEDGEMENTS

Nadja Gomes de Oliveira has received support for her research work from the Ceará Foundation for Scientific and Technological Development Support (FUNCAP), Scientific and Technological Development National Council (CNPq) - Grant n◦. 305456/2019-9, and Global Affairs Canada. P. A. C. Rocha and M. H. Ewsharqawy gratefully acknowledge the partial support provided by the Federal University of Ceara/Brazil and the University of Guelph/Canada.

5. REFERENCES

Brinsfield, R.B., Yaramanoglu,M., Wheaton, F., 1984. *Ground level solar radiation prediction model including cloud cover effects*. Solar Energy 33, 493–499.

Chen T, Guestrin C, 2016. *Xgboost: a scalable tree boosting sys-tem*. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794.

Duffie, J.A. and Beckman, W.A, 2013. Solar Engineering of Thermal Processes. 4th ed. Hoboken, N.J.: John Wiley & Sons.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R., 2004. *Least Angle Regression*. The Annals of Statistics, 32, 407-499.

Friedman, J., 2002. *Stochastic Gradient Boosting*. Computational Statistics & Data Analysis. 38. 367-378. 10.1016/S0167-9473(01)00065-2.

Hastie, T., Tibshirani, R. & Friedman, J., 2009. *Boosting and Additive Trees*. 10.1007/b94608_10.

Ineichen, F.R., Pereira, E.B. and Abreu, S.L., 2007. *Satellite-derived solar resource maps for Brazil under SWERA project*. Journal of Solar Energy.

Ineichen, P., 2008. *A broadband simplified version of the solis clear sky model*, Sol. Energy 82.

Inman, R.H., Pedro, H.T.C., Coimbra, C.F.M., 2013. *Solar forecasting methods for renewable energy integration*. Progress in Energy and Combustion Science, Vol. 39, No. 6, pp. 536-537. Available at: https://www.sciencedirect.com/science/article/pii/S0360128513000294.

James, G., Daniela W., Trevor H., and Tibshirani, R. 2013. *An Introduction to Statistical Learning*. 1st ed. Springer Texts in Statistics. New York, NY: Springer.

Kleissl, J., 2013. *Solar Energy Forecasting and Resource Assessment*. Academic Press, USA.

Kuhn, M. and Kjell, J., 2013. *Applied Predictive Modeling*. New York: Springer. MLA 8th Edition.

Pedro, H.T.C, Coimbra, C.F.M, David, M., and Lauret, P., 2018. *Assessment of machine learning techniques for deterministic and probabilistic intra-hour solar forecasts*. Renewable Energy, Volume 123, Pages 191-203, ISSN 0960-1481.

Rocha, P.A.C., Pedro, H.T.C and Coimbra, C.F.M, 2021. *Nowcasting and short term GHI forecasting using Goes-16 shortwave radiance data: a machine learning case study of Petrolina-PE, Brazil*. In Proceedings of the 22nd International Congress of Mechanical Engineering - COBEM 2021. Florianópolis, Brazil.

Voyant, C., Gilles Notton, G., Kalogirou, S., Nivet, M., Paoli, C., Motte, F. and Fouilloy, A., 2017. Machine Learning methods for solar radiation forecasting: a review. Renewable Energy, Elsevier, 105, pp.569-582.

Yang, D. (2019). A guideline to solar forecasting research practice: Reproducible, operational, probabilistic or physically based, ensemble, and skill (ROPES). Journal of Renewable and Sustainable Energy, 11(2), 022701. doi:10.1063/1.5087462.

Yang, D., 2021. *Validation of the 5-min irradiance from the National Solar Radiation Database (NSRDB).* Journal of Renewable and Sustainable Energy 13, 016101.