Forecasting of Wind Power Generation Cristalândia wind farm using Tree-Based Machine Learning Approaches

Camila Albuquerque Fitipaldi
Departamento de Engenharia Elétrica
Universidade Federal de Pernambuco
Recife, Brazil
camila.fitipaldi@ufpe.br,

Rodrigo de Paula Monteiro
Unicap-Icam International School
Universidade Católica de Pernambuco
Recife, Brazil
rodrigo.paula@unicap.br

Diego Pinheiro
Universidade de Pernambuco
Recife, Brazil
dmpfs@ecomp.poli.br

Waldemar Ferreira
Unidade Acadêmica de Belo Jardim
Universidade Federal Rural de Pernambuco
Recife, Brazil
waldemar.ferreira@ufrpe.br

Liliane Sheyla da Silva Fonseca
Unicap-Icam International School
Universidade Católica de Pernambuco
Recife, Brazil
liliane.fonseca@unicap.br

Andrea Maria Nogueira Cavalcanti Ribeiro
Departamento de Engenharia Elétrica
Universidade Federal de Pernambuco
Recife, Brazil
andrea.marianogueira@ufpe.br

***Abstract*—Wind power has gained increasing attention as a rapidly growing source of sustainable electricity generation. As variable renewable energy, however, its reliability, stability, and efficiency depend on factors such as wind speeds, air density, and turbine characteristics. As a result, an effective energy management strategy requires the accurate forecasting of wind power generation. Machine learning approaches have been applied to forecasting wind power generation, but their proper fine-tuning is still not fully understood. In this work, we trained using 5-fold cross-validation and fine-tuned using a GridSearch tree-based machine learning models, namely, Extreme Gradient Boosting (XGBoost) and Random Forest, for the forecasting of wind power generation. We evaluated XGBoost and Random Forest using data from the Cristalândia wind farm in Brumado-BA. The results suggest that tree-based models can accurately forecast wind power generation. Since they are relatively simple and easy to train when compared to machine learning models based on neural architectures, tree-based models are competitive approaches to forecast wind power generation.**

***Index Terms*—Machine learning. Prediction. Linear Regression. Random Forest. XGBoost.**

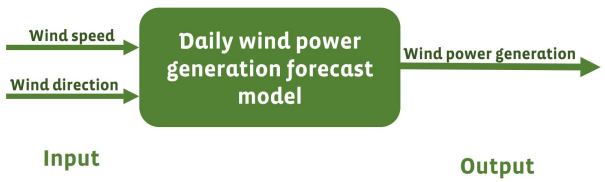
I. Introduction

2Renewable and sustainable electricity generation has gained increasing attention worldwide [1]. In Brazil, hydroelectric plants play a crucial role in electricity generation, corresponding to 65.2% of its electricity matrix in 2020 [2]. However, the lack of regular rains has led to a water crisis that increased the search for other renewable energy sources such as wind, biomass, and solar. Among those sources, Wind power rapidly emerged as an outstanding alternative.

According to the Brazilian Energy Operation Plan 2020, wind power generation will grow 11% in 2024 [3]. Besides, the Brazilian Wind Energy Association highlights that wind power is the second-largest contribution to the electricity matrix in Brazil, which has 795 wind farms with 21.5 GW of installed capacity [4]. However, this power generation modality depends on ever-changing wind availability over the year. Such fluctuations in the wind patterns affect the power generation efficiency and require power substitution from other sources that might not be available in the short term, *e.g.*, coal plants. Accurately forecasting wind power generation arises as a challenging and relevant task.

In this work, we use artificial intelligence to build a forecasting model of wind power generation for the Cristalândia wind farm in Brumado-Brazil. The machine learning model is represented in Figure 1. With this model, we aim to reduce the economic losses regarding the power generation deficit in the wind farm, which has an installed capacity of 90 MW and consists of 45 wind turbines with a nominal power of 2 kW. The motivation for choosing this wind farm regards its importance to the local economy, *e.g.*, its power generation can serve around 170,000 families [5]. Also, Cristalândia has open data on daily wind generation, in addition to the anemometric tower in Brumado.

Fig. 1. Forecasting wind power generation model.



By developing a forecasting model for the daily wind power generation, we expect to provide the following benefits to the company that manages the wind farm:

- Create more effective predictive maintenance plans;
- Define energy efficiency projects;

- Avoid paying fines to the concessionaire due to the power generation deficit;
- Monitor whether the enterprise is supplying the generation value foreseen in the project;
- Through energy efficiency projects, increase the availability of transmission lines.

This work is organized as follows: in Section I (Introduction), we present the problem, motivation, proposal, and contributions of this work. Section II (Related Work) lists the published works related to the forecasting of wind power generation. Section III (Materials and Methods) presents the dataset and methodology used in this work. In Section IV (Results), we present and discuss the results achieved, and in Section V (Conclusion), we make the final considerations about the study developed.

II. Related work

In the scientific community, studies about the forecasting of wind power generation gained prominence as the countries started investing in renewable energy sources.

Demir and Taşcı [6] used machine learning-based approaches to forecast the power generated by a wind turbine. They used the following algorithms to perform the power forecasting: linear regression, polynomial regression, decision tree, *AdaBoost*, and random forest. Those algorithms were fed with wind direction and speed data, collected for 12 months in 2018 at a 10-minute sampling rate. The metric R^2 was used to evaluate the forecasting models, and linear regression achieved the best results.

Ahmed *et al.* [7] also used linear regression models to forecast wind generation. However, the wind speed data was the only input available. Also, the study focused on using Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) metrics to evaluate the forecasting.

Chen *et al.* [8] combined deep learning and genetic algorithms to perform wind power forecasting. They trained convolutional neural networks using different algorithms, *e.g.*, backpropagation, and genetic algorithm. They used the metrics Mean Absolute Percentage Error (MAPE), RMSE, and R^2 to compare different training algorithms for wind power forecasting. The forecasting models were fed with wind direction and speed data, collected for 65 hours in July 2017 at a 10-minute sampling rate. The genetic algorithm provided the best training performance.

Vaitheeswaran and Ventrapragada [9] and Niu *et al.* [10] forecast the wind generation with gated recurrent units. The former evaluated the forecasting models with MAE and used the following input variables: wind speed, wind direction, wind cube height, time of day, seasonal index, and horizon index. On the other hand, Niu *et al.* used wind speed, direction, temperature, pressure, air density, and seasonality as input variables. The resulting models were evaluated by metrics such as Normalized Root Mean Square Error (NRMSE), MAPE, and Coefficient of the Variation of the Root Mean Square Error (CV-RMSE).

Dong *et al.* [11] used deep and traditional machine learning-based models to forecast wind power. The deep learning models were recurrent neural networks with long short-term memory cells, which are suitable structures for tasks regarding time series. The recurrent neural networks were compared with other machine learning models: linear regression, random forest, and gradient boosting. In this study, only the wind speed fed the forecasting models. The metric used to evaluate the model was the MSE, and the deep learning-based models achieved the best results. Chandran *et al.* [12] also used deep learning for this task. Their forecasting models were based on Long Short-Term Memories (LSTMs), Gated Recurrent Units (GRUs), and Recurrent Neural Networks (RNNs), using input variables such as temperature, wind speed, and wind direction. The MSE was the evaluation metric as well.

Demolli *et al.* [13] also used machine learning-based approaches to perform wind power forecasting. The authors performed a Least Absolute Shrinkage and Selection Operator (LASSO), k-nearest neighbors, Extreme Gradient Boosting (XGBoost), random forest, and support vector machines. Those models performed the forecasting by using wind speed and turbine data as input information. They trained the forecasting models with data collected for four years at a 1-hour sampling rate and used R^2, MAE, and RMSE as evaluation metrics. In Fahim *et al.* [14], wind forecasting is performed by artificial neural networks and XGBoost algorithms, with wind speed and power as inputs and RMSE as an evaluation metric.

These paper have as contributing use only weather data to do wind power generation forecast with RMSE, MAE and R^2 how metrics and using Linear regression, XGBoost and Random Forest like the machine learning models. These algorithms had as input wind speed data and wind direction how long 2018-2021 as wind power forecasting daily average. The results obtained was wind power forecasting average one day forward.

III. Materials and methods

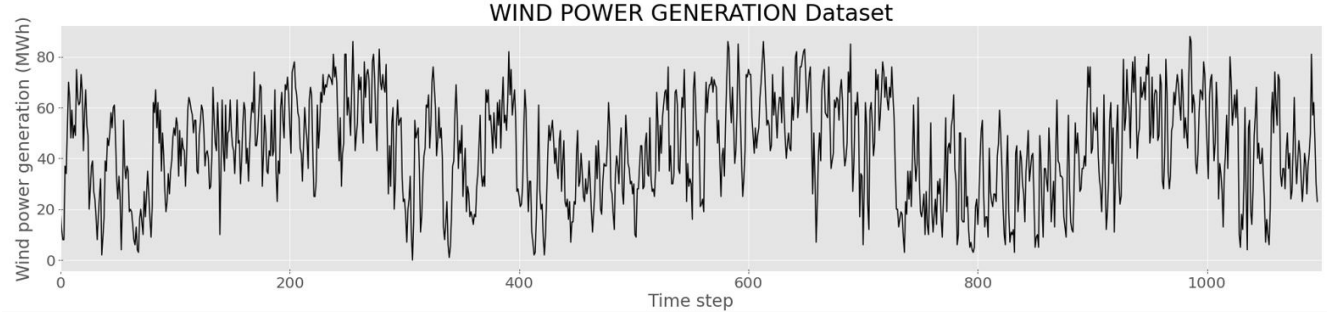
We used weather and wind power generation data to train and evaluate the artificial intelligence models. A total of 1,097 daily samples were collected from the Cristalândia wind farm, in Brumado-BA, from January 1st, 2018 to January 1st, 2021. The weather data source was the Brazilian National Institute of Meteorology [15], and the wind power generation data were obtained from the Brazil's National Grid Operator [16]. The wind power generation data contained the wind power generation (Figure 2.a), and the weather data consisted of the wind speed (Figure 2.b) and wind direction (Figure 2.c).

A. Data preprocessing

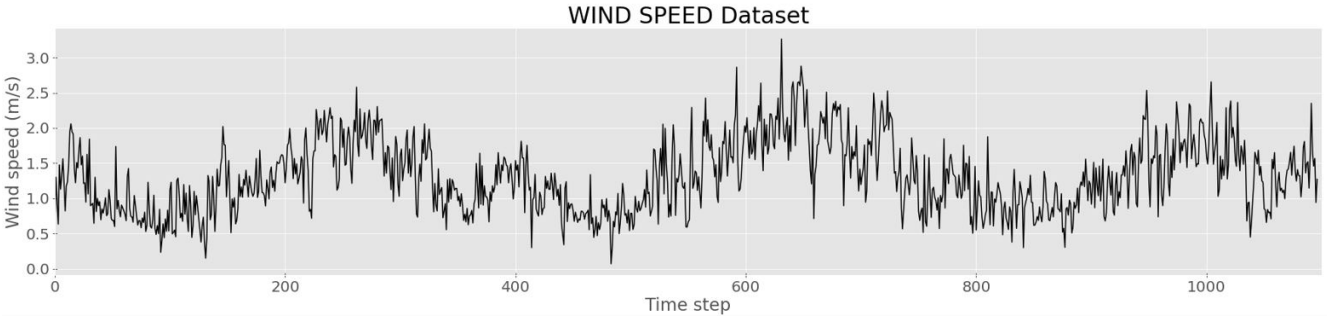
Both weather and wind power generation data presented missing data. Thus, we performed data imputation using the moving average method where the solution of missing value is the average of five numbers prior to them.

We normalized the data to the range [0, 1] through MinMaxScaler function contained in sklearn library based on

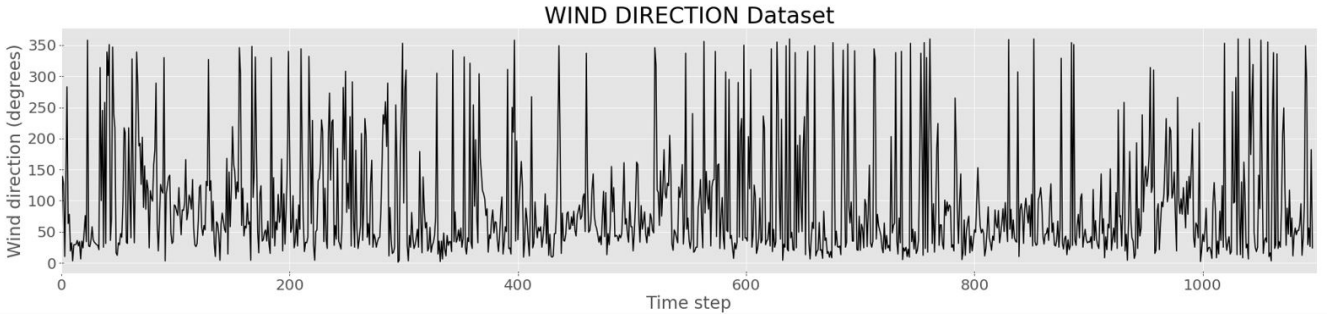
Fig. 2. Daily wind (a) power generation dataset, (b) speed dataset and (c) direction dataset.



(a)



(b)



(c)

Equation (1).

$$X'_i = \frac{X_i - X_{min}}{X_{max} - X_{min}} \tag{1}$$

Where X'_i is the rescaled value, X_i is the original value, X_{min} the minimum value in feature and X_{max} is the maximum value in feature.

It provides the time series a consistent scale and can improve the performance of machine learning-based algorithms.

B. Forecasting models

Section II presented some algorithms commonly employed in wind power forecasting. Algorithms based on deep learning tend to perform better, especially when dealing with a complex and large amount of data. However, they impose a high computational burden on the process. As seen in Subsection III-A, our dataset consists of 1,097 low-dimensional samples. Thus, by using regular machine learning models, *e.g.*, linear regression, random forest, and XGBoost, we achieve satisfactory results with less computational burden. Those models are presented in the remainder of this Subsection.

1) Linear regression: Linear regression is a supervised machine learning algorithm used for regression tasks. It maps a set o input values x to output values y by a linear function whose parameters are learned from data [7]. Equation (2) illustrates the linear regression model.

$$y(t) = a_0 + \sum_{i=1}^{n} a_i \cdot x_i(t) + r(t) \tag{2}$$

in which $y(t)$ is the predicted value, $x_i(t)$ are the input values, and $r(t)$ is the residual value at time t. a_i are the linear regression parameters.

2) Random forest: Random forest is a supervised machine learning algorithm suitable for classification and regression tasks. It is a set of decision trees trained with random samples extracted from the complete dataset. Concerning regression

tasks, the Random Forest returns the average forecasting of the individual trees [17], [18].

This algorithm presents some advantages: it optimizes the use of computational resources by allowing parallelization, and deals with noisy data better than other machine learning algorithms. However, since the random forest consists of a large combination of trees, it compiles slowly and requires a lot of computational memory [19], [20].

3) EXtreme Gradient Boosting: XGBoost is a supervised machine learning algorithm used for classification and regression tasks. It is a decision tree-based ensemble algorithm that uses a gradient boosting framework.

In this algorithm, decision trees are created in sequential form. Then, weights are assigned to all independent variables, which feed the decision tree for forecasting. The weight of variables wrongly predicted by the tree increases, and then these variables will feed a second decision tree, repeating the process.

XGBoost outperforms other machine learning-based algorithms in problems with small-to-medium structured data. Also, it is a resilient and robust method that prevents and cubs over-fitting quite easily. However, this algorithm does not perform well on sparse and unstructured data. Besides, it is sensitive to outliers since every classifier is forced to fix the errors in the predecessor learners [21].

C. Hyper-parameterization

We split the data into training and testing sets to perform the hyper-parameterization. The proportion was 80% of samples for training and 20% for testing. To determine the best hyperparameters for each forecasting model, we used the grid search algorithm with 5-fold cross-validation [22], and for each technique we performed the grid search 30 times to achieve more reliable results. The hyperparameters assessed in this process were: tree depth and number of trees. Table I lists the hyperparameters regarded in this process and their respective range of values.

TABLE I
MACHINE LEARNING PARAMETERS AND LEVELS.

Technique	**Parameters**	**Levels**
Random forest XGBoost	Tree	From 10 to 130, step 30
Random forest XGBoost	Maximum depth	From 2 to 10, step 2

In order to carry out the comparative analysis of different configurations of the studied models, we selected the root mean squared error (RMSE) and mean absolute error (MAE) metrics to assist the choice of the forecasting model with the best performance, that is, the most accurate model. The choice of metrics was justified, as they were present in most of the related works [7]–[10], [13].

The RMSE is obtained when the difference between the true value and the predicted is squared to then takes the square root of the average of these values like represented on Equation (3).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

y_i is the true value, $\hat{y}_i$ the predicted value and n corresponding the sample size.

The MAE is the average of module difference between true value and the predicted as represented on Equation (4) [23].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4}$$

The R^2 is the coefficient of determination regression score function. How much more the R^2 approaches of 1.0 the better the forecasting model. R^2 can be represented in Equation (5) [23].

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5}$$

A baseline was also selected for comparison purposes, the linear regression technique.

IV. RESULTS

Our first step was calculating the baseline using the linear regression technique. We adopted the grid search to choose the best fitting regression technique parameters [24]. Table II lists the average grid search results for the linear regression baseline. We observe that setting the intersection parameter to *True* provided the lowest forecasting errors regarding both metrics. The R^2 presented the better value 0.465252 what it means that the variance of generation is few explained by X.

TABLE II
MAE, RMSE AND R^2 OF *linear regression* FOR DIFFERENT INTERSECTION VALUES.

	True	False
MAE	**0.134552**	0.137738
RMSE	**0.169004**	0.174679
R^2	**0.465252**	0.423481

Figure 3 presents the grid search results for XGBoost and random forest. According to both RMSE and MAE metrics, the XGBoost achieved the best performance for a maximum depth equal to 2 (Figure 3.a and Figure 3.c). On the other hand, the number of trees was different regarding each metric. While the RMSE, which penalizes the most significant individual errors, suggested that 40 trees provided the best results, the MAE suggested 70 trees. This way, we can assume that by using 40 trees, the absolute error slightly increased regarding the configuration with 70 trees, but the most significant forecast errors were reduced.

For parallel runtimes models machine learning we represented the Table III. How hoped the linear regression have

Fig. 3. Grid search result for (a) XGBoost - RMSE, (b) random forest - RMSE, (c) XGBoost - MAE and (d) random forest - MAE.

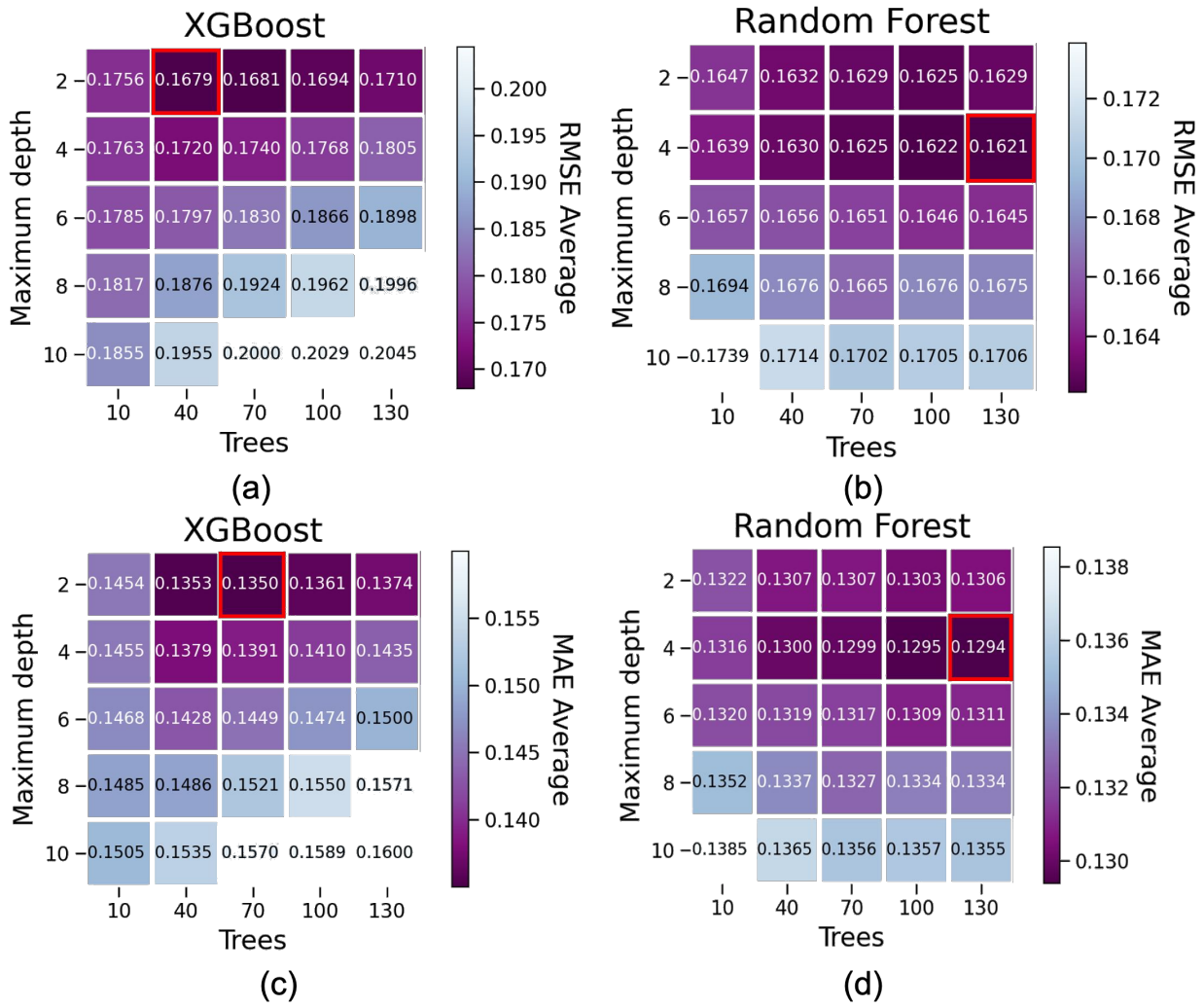
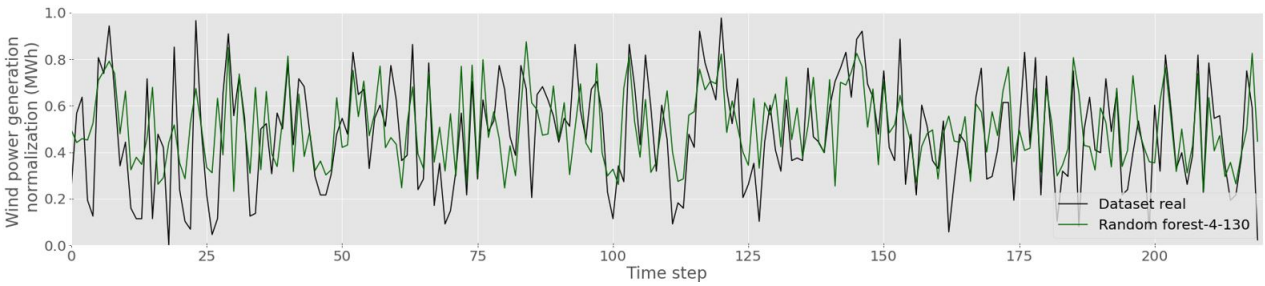


Fig. 4. Forecasting daily wind power generation using the random forest-4-130.



the smaller runtime because is a simple model where relates linearly independent variables with dependent variables doing then your runtime be little [25].

TABLE III
RUNTIMES IN SECONDS OF MODELS MACHINE LEARNING

Models	Mean fit time	Mean score time
Linear Regression	0.0042	0.0024
Random Forest 4 - 130	1.1425	0.2844
XGBoost 2 - 40	0.0608	0.0032
XGBoost 2 - 70	0.1833	0.0194

TABLE IV
RMSE AND MAE RESULTS FOR THE SELECTION OF ANALYZED MODELS.

Models	RMSE	MAE
Linear regression	0.169004	0.134552
Random forest 4-130	**0.162119**	**0.129373**
XGBoost 2-40	0.167933	0.135295
XGBoost 2-70	0.168127	0.135000

Table IV summarizes the results achieved by the best configuration of each forecast algorithm regarding the RMSE and MAE. We observe that the random forest presented the best performance for both metrics. The RMSE of random forest was about 3.5% lower than the second-best result, achieved by the XGBoost with 70 trees and a maximum depth of 2. The MAE, on the other hand, was about 3.9% lower than the second-best result, achieved by the linear regression baseline with intersection parameter *True*.

Concerning the random forest, both metrics suggested that the best configuration of hyperparameters was the maximum depth four with 130 trees (Figure 3.b and Figure 3.d). The performance of random forest is depicted in Figure 4, which shows the daily wind power forecasts for the algorithm with a maximum depth of four and 130 trees. They suggested that this configuration was the best one to reduce the absolute forecasting error and the variation among the individual errors. We choice these model too because random forest allow optimize parameters besides is a algorithm of high perform precision and for the lenght sample was needed little storage of computacional memory. The random forest describe well phenomenous that have missing data how the case study where sometimes anemometric tower can't measure winds speed.

When compared with random forest the linear regression is most fast but the linear regression have limitations in your model like impaired precision due outliers and the daily wind energy generation have big variance due winds speed oscillation.

Then when we go select a model for describe one problem first is need understand the nature of phenom and your behavior to that the machine learning model be consistent with case study.

V. CONCLUSION AND FUTURE WORK

The wind is an attractive alternative energy source since it is natural, renewable, relatively cheap, and carbon-free. In principle, it is possible to produce energy from wind turbines every day. Even systems that require the energy to be continuously supplied can exploit wind energy. However, using wind energy is challenging because it demands a high initial investment in analysis before establishing a wind plant. Some challenges are the varying wind period, the distance of wind-efficient areas to the national grids, and its environmentally disruptive effects.

Regarding these challenges, this study used machine learning algorithms to perform wind power forecasting based on daily wind speed and direction data. In particular, we used data from the Cristalândia wind farm in Brumado-Brazil. The classification algorithms used to forecast wind power were: linear regression, random forest, and XGBoost. The results showed that the best algorithm to fit wind power is the random forest using the number of trees equal to 130 and depth equal to four.

An essential outcome of this study is that machine learning algorithms can be successfully used to forecast wind power production based on wind direction and speed. So, before establishing wind plants in an unknown geographical location, it is reasonable to use our approach to forecast wind power based on local data.

This study suggests that machine learning models may be used to forecast daily average wind power production. Further we can a good wind power forecasting only using weather as input and understanding how the o behavior of phenomenon have relationship for use the better machine learning model these choice is very important in construction of a artificial intelligence project. However, ensemble solutions combining machine learning and deep learning may provide better results and are worthy of exploration. Another relevant topic to be studied is adopting others time windows for forecasting. Our result only forecasts wind production to the next day; however, we can evaluate our model to predict the next three days or one week, for instance. Finally, we can apply our method to evaluate data from other farm plants worldwide.

REFERENCES

[1] R. Banos, F. Manzano-Agugliaro, F. Montoya, C. Gil, A. Alcayde, and J. Gómez, "Optimization methods applied to renewable and sustainable energy: A review," *Renewable and sustainable energy reviews*, vol. 15, no. 4, pp. 1753–1766, 2011.

[2] EPE, "Matriz energética e elétrica," urlhttps://www.epe.gov.br/pt/abcdenergia/matriz-energetica-e-eletrica, 2021.

[3] ONS, "Pen," 2020. [Online]. Available: http://www.ons.org.br/

[4] ABEEólica, "O tamanho da industria no brasil," 2022. [Online]. Available: https://abeeolica.org.br/

[5] Engexpor, "Parque eólico de cristalândia," urlhttps://engexpor.com/projeto/parque-eolico-de-cristalandia/, 2020.

[6] F. Demir and B. Taşcı, "Predicting the power of a wind turbine with machine learning-based approaches from wind direction and speed data," in *2021 International Conference on Technology and Policy in Energy and Electric Power (ICT-PEP)*, 2021, pp. 37–40.

[7] M. I. Ahmed, P. Pan, R. Kumar, and R. K. Mandal, "Wind generation forecasting using python," in *2020 International Conference on Emerging Frontiers in Electrical and Electronic Technologies (ICEFEET)*, 2020, pp. 1–5.

[8] G. Chen, J. Shan, D. Y. Li, C. Wang, C. Li, Z. Zhou, X. Wang, Z. Li, and J. J. Hao, "Research on wind power prediction method based on convolutional neural network and genetic algorithm," in *2019 IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia)*, 2019, pp. 3573–3578.

[9] S. S. Vaitheeswaran and V. R. Ventrapragada, "Wind power pattern prediction in time series measuremnt data for wind energy prediction modelling using lstm-ga networks," in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2019, pp. 1–5.

[10] Z. Niu, Z. Yu, W. Tang, Q. Wu, and M. Reformat, "Wind power forecasting using attention-based gated recurrent unit network," *Energy*, vol. 196, p. 117081, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360544220301882

[11] D. Dong, Z. Sheng, and T. Yang, "Wind power prediction based on recurrent neural network with long short-term memory units," in *2018 International Conference on Renewable Energy and Power Engineering (REPE)*, 2018, pp. 34–38.

[12] V. Chandran, C. K. Patil, A. Merline Manoharan, A. Ghosh, M. Sumithra, A. Karthick, R. Rahim, and K. Arun, "Wind power forecasting based on time series model using deep machine learning algorithms," *Materials Today: Proceedings*, vol. 47, pp. 115–126, 2021, nCRABE.

[13] A. E. Halil Demolli, Ahmet Sakir Dokuz and M. Gokcek, "Wind power forecasting based on daily wind speed data using machine learning algorithms," *Energy Conversion and Management*, vol. 198, 2019.

[14] M. Fahim, V. Sharma, T.-V. Cao, B. Canberk, and T. Q. Duong, "Machine learning-based digital twin for predictive modeling in wind turbines," *IEEE Access*, vol. 10, pp. 14 184–14 194, 2022.

[15] INMET, "banco de dados meteorológicos de brumado," 2022. [Online]. Available: https://bdmep.inmet.gov.br/

[16] ONS, "Geração de energia eólica," 2022. [Online]. Available: http://www.ons.org.br/

[17] M. W. Ahmad, J. Reynolds, and Y. Rezgui, "Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees," *Journal of Cleaner Production*, vol. 203, pp. 810–821, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0959652618325551

[18] D. Assouline, N. Mohajeri, and J.-L. Scartezzini, "Large-scale rooftop solar photovoltaic technical potential estimation using random forests," *Applied Energy*, vol. 217, pp. 189–211, 05 2018.

[19] D. Liu and K. Sun, "Random forest solar power forecast based on classification optimization," *Energy*, vol. 187, p. 115940, 08 2019.

[20] Z.-C. Chen, F. Han, L. Wu, J. Yu, S. Cheng, P. Lin, and H. Chen, "Random forest based intelligent fault diagnosis for pv arrays using array voltage and string currents," *Energy Conversion and Management*, vol. 178, pp. 250–264, 12 2018.

[21] S.-H. Wu and Y.-K. Wu, "Probabilistic wind power forecasts considering different nwp models," in *2020 International Symposium on Computer, Consumer and Control (IS3C)*, 2020, pp. 428–431.

[22] J. Wainer and P. Fonseca, "How to tune the rbf svm hyperparameters? an empirical evaluation of 18 search algorithms," *Artificial Intelligence Review*, vol. 54, no. 6, pp. 4771–4797, 2021.

[23] V. Plevris, G. Solorzano, N. P. Bakas, and M. E. A. Ben Seghier, "Investigation of performance metrics in regression analysis and machine learning-based prediction models," in *8th European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS Congress 2022)*. European Community on Computational Methods in Applied Sciences, 2022.

[24] P. Lerman, "Fitting segmented regression models by grid search," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 29, no. 1, pp. 77–84, 1980.

[25] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4, no. 4.