

XGBoost na Previsão da Geração de Energia Elétrica em Parques Eólicos

André Luis Andrade Machado
Curso de Engenharia Elétrica
Universidade Federal do
Rio Grande do Norte
Natal/RN, Brasil
andre.luis.e3@gmail.com

Luiz Gonzaga de Q. Silveira Junior
Departamento de Engenharia de Comunicações
Universidade Federal do
Rio Grande do Norte
Natal/RN, Brasil
junior@ct.ufrn.br

Marcus Alexandre Nunes
Departamento de Estatística
Universidade Federal do
Rio Grande do Norte
Natal/RN, Brasil
marcus@marcusnunes.me

Resumo—Neste trabalho, propõe-se um modelo baseado em *XGBoost* para a previsão da geração de energia elétrica em parques eólicos. O desenvolvimento e a validação do modelo proposto foram realizados através da participação de um dos autores numa competição internacional de Ciência de Dados, que é promovida pela *Compagnie Nationale du Rhône*, uma empresa francesa que atua no segmento de produção de energia elétrica. A partir de comparações de desempenho realizadas no final da competição com outros métodos concorrentes, foi possível verificar que o modelo proposto previu com boa precisão a dinâmica da produção de energia elétrica gerada em seis parques eólicos, de localização anônima, durante um período de aproximadamente oito meses. Além disso, o modelo proposto tem baixa complexidade. Portanto, possui uma boa relação de custo-benefício, o que favorece seu emprego em condições reais, geralmente menos desafiadoras do que aquelas enfrentadas na competição de Ciência de Dados.

Palavras-chave—Previsão, Geração de Energia Elétrica, Parques Eólicos, Aprendizado de Máquina, *XGBoost*.

I. INTRODUÇÃO

Nos diversos meios de informação atuais, o cenário econômico global é geralmente caracterizado pelo seu desenvolvimento acelerado, marcado por uma revolução tecnológica de velocidade frenética onde a inovação se supera a cada dia, trazendo à tona novos produtos, serviços e metodologias que contribuem cada vez mais para o aumento da qualidade de vida da sociedade.

No centro desta revolução se encontra o processo de industrialização, cujo desenvolvimento, além de ter contribuído para o impulsionamento da inovação tecnológica, acompanhou nitidamente a evolução do processo de extração de energia da natureza e sua devida conversão para formas convenientes para o seu aproveitamento pelo ser humano. Este processo evolutivo, entretanto, se deu de forma concentrada em algumas fontes energéticas específicas, o que veio se evidenciando como a origem de alguns problemas ao longo do crescimento do setor energético, levantando principalmente questões relacionadas ao impacto ambiental e à sustentabilidade.

A partir desse contexto, o setor energético passou a investir cada vez mais no desenvolvimento de novas tecnologias para extração de energia de fontes renováveis que pudessem complementar, e a longo prazo substituir, as fontes tradicionais

de energia. Dentre as diversas fontes alternativas levantadas neste processo, a energia eólica vem ganhando um destaque considerável, principalmente pelo seu caráter ilimitado, baixo impacto ambiental e preços cada vez mais competitivos. Sua adoção, entretanto, traz alguns desafios. Apesar de possuir grande abundância e estabilidade como recurso, o vento ainda possui um comportamento bastante aleatório, o que gera um desafio para a operação do sistema elétrico, visando manter a sua estabilidade principalmente a partir do correto equilíbrio entre geração e consumo da energia [1] [2].

Para otimizar o planejamento das atividades de operação de um sistema de geração eólico em um cenário de incertezas, diversos modelos de previsão da produção de energia elétrica têm sido propostos na literatura ao longo dos últimos anos, e o seu desenvolvimento vem sendo impulsionado tanto pelo amadurecimento do campo de Ciência de Dados, quanto pelo advento de técnicas de aprendizado de máquina cada vez mais eficazes, o que permite alcançar horizontes de previsão cada vez maiores, mantendo-se um nível de precisão adequado ao planejamento da operação.

Este trabalho propõe um modelo de previsão da produção de energia elétrica por um parque eólico baseado em *Árvores de Decisão*, o *XGBoost*, que apresenta um menor custo computacional quando comparado com outras abordagens de aprendizado de máquina como, por exemplo, redes neurais.

Para a obtenção desse modelo, foram consideradas variáveis meteorológicas, com destaque para a velocidade e a direção do vento. Desta forma, a previsão da geração de energia elétrica gerada pelo parque eólico através do modelo baseado em *XGBoost*, proposto neste trabalho, considera o comportamento estocástico das variáveis envolvidas e o período de oito meses.

Foram realizadas comparações de desempenho na precisão da previsão com modelos concorrentes em uma competição internacional de Ciência de Dados, promovida pela *Compagnie Nationale du Rhône*, uma empresa francesa que atua no segmento de produção de energia elétrica.

A partir dos resultados obtidos, é possível constatar que o modelo proposto possui uma boa relação de custo-benefício, o que favorece seu emprego em condições reais.

II. REVISÃO DA LITERATURA

Com relação ao tema específico deste trabalho, é possível encontrar uma gama de trabalhos com uma certa variedade de metodologias utilizadas, havendo entretanto uma predominância de trabalhos baseados no uso de métodos estatísticos, métodos físicos e, mais recentemente, como já citado, baseados na utilização de modelos de aprendizado de máquina.

Neste último contexto, alguns trabalhos podem ser encontrados como em [3], onde o autor propõe um modelo para previsões de energia eólica gerada a curto prazo a partir da utilização de Redes LSTM, em [4], onde os autores propõem um modelo híbrido, baseado na utilização de uma Rede Neural Convolucional em conjunto com uma Rede Neural de Funções de Base Radial, e em [5], onde os autores utilizam uma Rede Neural de Wavelets associada métodos específicos de otimização e decomposição de séries temporais para modelar a produção de energia eólica.

Entretanto, com exceção de alguns outros trabalhos, como em [6], onde é proposto um modelo composto por uma Rede Neural Convolucional associado ao LightGBM, que consiste em um modelo semelhante ao XGBoost utilizado neste trabalho, é possível perceber que grande parte dos modelos propostos na literatura se concentram no uso exclusivo de Redes Neurais. Este trabalho, assim como o último citado, consistem em um esforço em contribuir à gama de trabalhos referentes a este segmento a partir da exploração de modelos distintos da família de Aprendizado Profundo, mas que podem apresentar um igual potencial de geração de resultados de alta precisão.

III. METODOLOGIA

Nesta seção, a Metodologia utilizada neste trabalho é detalhada. A mesma envolve a aplicação de alguns passos, listados a seguir.

- Obtenção dos Dados;
- Pré-Processamento dos Dados;
- Criação de Variáveis (*Feature Engineering*);
- Ajuste de Hiper Parâmetros e Validação;
- Treinamento e Aplicação do Modelo.

A. Obtenção dos Dados

Neste trabalho foi utilizado um conjunto de dados fornecidos pela Compagnie Nationale du Rhône em uma competição de ciência de dados hospedada pelo portal www.challengedata.ens.fr, cujo objetivo dos participantes consistia em realizar a previsão de geração de energia de seis parques eólicos na França em intervalos de 1 hora durante um período de 8 meses e 15 dias.

Para isso são fornecidos dois conjunto de dados (treinamento e teste) contendo previsões em passos de 1 hora de algumas variáveis meteorológicas, fornecidas por 4 NWP (Numerical Weather Prediction) diferentes, geradas em dias e horários específicos. NWP consistem em modelos meteorológicos de alta precisão mantidos por diversos centros de meteorologia no mundo, e seu uso é muito importante para

diversas áreas, dentre as quais a tratada neste trabalho. As variáveis meteorológicas fornecidas no conjunto de dados em questão são:

- Velocidade do Vento Zonal e Meridional a 10 metros e 100 metros de Altura;
- Temperatura;
- Percentual de Cobertura de Nuvens.

A previsão de cada uma dessas variáveis é apresentada de acordo com o NWP que a gerou, o dia em que a previsão foi obtida, podendo ser 1 ou 2 dias antes ou no mesmo dia em que a previsão ocorre de fato e o horário em que a previsão foi feita, onde é possível encontrar 00h, 06h, 12h e 18h.

Além disso, para permitir a aplicação de um modelo supervisionado, no conjunto de treinamento é fornecido os valores reais de produção de energia de cada parque eólico em faixas de 1 hora no período considerado, variável esta que desejamos obter como resultado do modelo no conjunto de teste.

B. Pré-Processamento dos Dados

Devido à diversidade de NWP, horários e dias de previsões, o conjunto de dados original apresenta uma alta dimensionalidade, o que pode se tornar um obstáculo nos passos seguintes. Além disso, a existência de NWP diferentes pode trazer uma variância indesejável para as suas previsões, o que pode ser atenuado pela agregação das diferentes previsões disponíveis de forma a alcançar resultados mais precisos [7].

Portanto, nesta etapa os dados são submetidos a uma redução de dimensionalidade feita a partir da agregação de previsões relacionadas à mesma grandeza meteorológica utilizando a mediana, escolhida pela sua robustez a possíveis valores extremos. Dessa forma, o conjunto de dados resultante passa não só a apresentar uma dimensionalidade inferior, mas também apresenta uma variância de suas previsões menos intensa.

C. Criação de Variáveis

A partir do conjunto de dados obtido na etapa de pré-processamento, algumas variáveis foram criadas a partir de conhecimento geral do problema, que possuem um potencial de auxiliar o modelo em seu processo de aprendizado e consequentemente contribuir para o aumento da sua precisão. Na literatura, este processo é comumente conhecido como *Feature Engineering*.

Para este trabalho, as seguintes variáveis foram criadas:

- Módulo e Direção (Em Graus) do Vento a 10 metros e 100 metros de Altura;
- Versões Atrasadas das Variáveis Originais em 7, 14 e 21 horas;
- Média e Variância de Meses Anteriores para cada Variável Original;
- Dia, Hora, Minuto, Dia da Semana e Dia do Ano em Formato Senoidal;
- Média e Variância Móvel com Janelas de 7 e 14 Horas de cada Variável Original;
- Máximo em Janela Expansiva para cada Variável Original.

O Módulo e Direção do Vento são obtidas de forma semelhante a um formato vetorial a partir das duas componentes do vento fornecidas pelo conjunto de dados original, e consistem numa representação diferente da velocidade do vento que pode facilitar o aprendizado modelo pela evidenciação de padrões nos dados. O módulo e a direção podem ser equacionados por:

$$|S| = \sqrt{U^2 + V^2} \quad (1)$$

$$\angle S = \arctan(V/U) \quad (2)$$

onde U corresponde à velocidade do vento zonal, V corresponde à velocidade do vento meridional e S corresponde ao vetor resultante da velocidade do vento a partir dessas duas componentes. Para os valores de ângulos negativos, seu valor é somado a 360 para possuir um valor equivalente positivo.

As Versões Atrasadas das Variáveis Originais basicamente consiste em apresentar, no tempo t , a magnitude da mesma variável em um tempo $t - \delta t$, onde δ corresponde ao atraso desejado. Sua presença no treinamento do modelo, pode ajudar a identificar tendências de crescimento ou decrescimento da variável no período de atraso considerado. Dessa forma, esta variável pode ser equacionada por:

$$L[t] = F[t - \delta t] \quad (3)$$

onde L corresponde à versão atrasada da variável F .

A Média e Variância de Meses Anteriores corresponde à aplicação destas operações de agregação nas amostras do mês anterior à amostra analisada, agregando informações gerais sobre o período analisado. Dessa forma, essas variáveis podem ser agregadas como:

$$\begin{cases} \bar{F}_i = \text{Média}([x_0, x_1, \dots, x_n]), \text{ se Mês}(x_n) = i - 1 \\ \delta(F)_i = \text{Variância}([x_0, x_1, \dots, x_n]), \text{ se Mês}(x_n) = i - 1 \end{cases} \quad (4)$$

Os Valores Temporais em Formato Senoidal foram obtidos a partir da aplicação das operações de seno e cosseno nos valores originais de forma a obter variáveis que reflitam o caráter cíclico do tempo, desde minutos até dias no ano. Matematicamente, essas variáveis podem ser representadas por:

$$\begin{cases} [\cos]\text{Dia do Mês} = \cos(2\pi(\text{Dia} - 1)/\text{Dias no Mês}) \\ [\sen]\text{Dia do Mês} = \sen(2\pi(\text{Dia} - 1)/\text{Dias no Mês}) \\ [\cos]\text{Dia da Semana} = \cos(2\pi\text{Dia}/7) \\ [\sen]\text{Dia da Semana} = \sen(2\pi\text{Dia}/7) \\ [\cos]\text{Dia do Ano} = \cos(2\pi(\text{Dia} - 1)/365) \\ [\sen]\text{Dia do Ano} = \sen(2\pi(\text{Dia} - 1)/365) \\ [\cos]\text{Hora} = \cos(2\pi\text{Hora}/24) \\ [\sen]\text{Hora} = \sen(2\pi\text{Hora}/24) \\ [\cos]\text{Minuto} = \cos(2\pi\text{Minuto}/60) \\ [\sen]\text{Minuto} = \sen(2\pi\text{Minuto}/60) \end{cases} \quad (5)$$

Para as variáveis envolvendo dia cujo valor inicial é 1, foi aplicada uma subtração unitária para que o primeiro dia possua valor nulo. Além disso, para a variável de dia no mês, o denominador interno varia de acordo com a quantidade de dias do mês considerado.

As Variáveis de Janela Móvel são aplicadas visando obter informações de um sinal ignorando suas componentes frequenciais elevadas e mantendo apenas componentes mais baixas, onde se acredita que a informação principal do sinal esteja contida. Neste trabalho, foram calculadas, em formato de janela móvel, a média e a variância das variáveis meteorológicas originais, que podem ser equacionadas como:

$$\begin{cases} MA(F[t]) = \text{Média}([x_t - n, \dots, x_t]) \\ M\delta(F[t]) = \text{Variância}([x_t - n, \dots, x_t]) \end{cases} \quad (6)$$

onde MA corresponde à operação de média móvel, $M\delta$ corresponde à operação de variância móvel, F corresponde à variável em questão e n corresponde ao tamanho da janela móvel.

Por fim, foi calculada a Variável de Máximo em Janela Expansiva, que basicamente consiste no valor máximo de uma variável até o momento sendo analisado. Sua presença pode ser útil para agregar ao aprendizado do modelo a distância do valor sendo analisado ao máximo tido até o momento em questão. Seu equacionamento pode ser dado por:

$$\text{Max Móvel} = \max([x_0, \dots, x_t]) \quad (7)$$

D. Ajuste de Hiper Parâmetros e Validação

Com a execução da etapa anterior, o conjunto de dados a ser utilizado já se encontra em um formato otimizado para a aplicação da solução proposta neste trabalho. A partir disso, esta etapa se preocupa com a otimização do modelo a ser utilizado e aplicado no conjunto de dados obtidos.

Como já citado anteriormente, este trabalho se baseia na utilização do XGBoost (*Extreme Gradient Boosting*), um modelo de aprendizado de máquina cujo funcionamento se baseia na utilização de múltiplas árvores de decisão geradas a partir de uma técnica denominada de *Boosting* [8]. Esta técnica, por sua vez, consiste em um processo iterativo de treinamento, onde cada árvore de decisão é treinada a partir dos residuais de árvores anteriores, de forma a complementar o aprendizado geral do modelo, direcionando-se para os comportamentos ainda não modelados pelos modelos anteriores. Esse processo é esquematizado de maneira simplificada na Fig. 1.

Para a otimização deste modelo no contexto da aplicação envolvida, é aplicado o processo de ajuste de hiper parâmetros atrelado a um processo de validação do modelo.

Os hiper parâmetros consistem em variáveis internas ao modelo que regulam seu processo de aprendizagem, sendo o seu ajuste importante para adaptar a execução do modelo ao problema abordado da melhor forma possível, e assim contribuir para uma melhor precisão das predições geradas.

O processo de validação por sua vez, consiste em uma técnica onde o modelo é aplicado em contexto controlado de

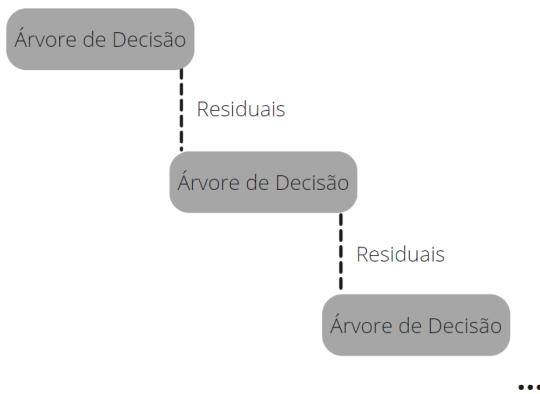


Fig. 1. Esquema de Treinamento do XGBoost.

forma a obter-se uma estimativa de performance geral. Neste trabalho, é utilizado como técnica de validação o *Nested Cross Validation*, cujo funcionamento consiste em promover várias iterações onde o conjunto de validação corresponde a uma determinada amostra do conjunto de treinamento, enquanto que todos os dados anteriores são utilizados para treinamento, assim como demonstrado na Fig. 2. Dessa forma, a ordenação dos dados é respeitada, o que torna esta técnica apropriada para aplicação em dados de séries temporais, que é o caso deste trabalho.

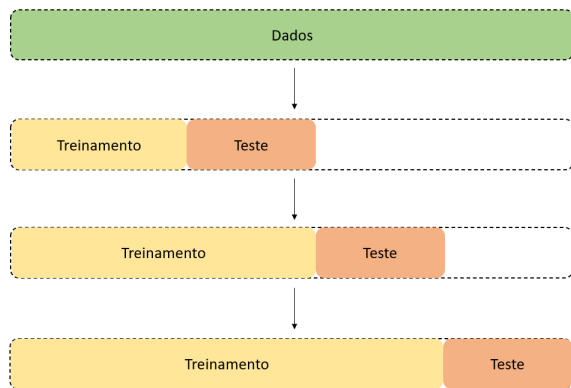


Fig. 2. Esquema de Nested Cross Validation

O objetivo desta etapa da Metodologia é acoplar os dois conceitos apresentados nesta seção a partir da ideia de que os hiper parâmetros podem ser selecionados visando obter um modelo que apresente a melhor performance retornada pelo processo de validação. Em outras palavras, o ajuste de hiper parâmetros pode ser feito de forma a otimizar a resposta dada pelo processo de validação.

Após a seleção dos hiper parâmetros a partir do processo de validação, um último teste é feito com o modelo resultante, avaliando seu desempenho em um outro subconjunto de dados separado previamente do subconjunto utilizado para a validação, comumente referenciado na literatura como *holdout*.

Dessa forma, o modelo selecionado é treinado em todo o subconjunto utilizado no ajuste de hiper parâmetros, e realiza previsões no subconjunto de *holdout*, que neste trabalho correspondeu a aproximadamente 1 mês de previsões de NWP, cujos valores reais da variável-alvo são utilizadas para avaliar o desempenho do modelo. Uma análise gráfica obtida a partir da aplicação do modelo nesse subconjunto pode ser vista na seção de Apresentação de Resultados.

E. Treinamento e Aplicação do Modelo

A partir do modelo selecionado na etapa anterior, pode-se treiná-lo em todo o conjunto de treinamento obtido após o pré-processamento e criação de variáveis e enfim aplicá-lo no conjunto de testes de forma a gerar as previsões requisitadas pela competição e necessárias para a resolução do problema. Este processo, entretanto, deve ser feito de forma a manter, o mais próximo possível, as mesmas condições nas quais o modelo utilizado foi gerado. Como o processo de validação envolveu a avaliação de modelos que operam gerando previsões em subconjuntos de validação pequenos, é de se esperar que o modelo selecionado seja otimizado para operar nestas condições, e portanto as mesmas devem ser mantidas ao realizar a aplicação do modelo no conjunto de testes.

Para isso, o processo de geração de previsão adotado neste trabalho segue um esquema iterativo onde o modelo selecionado é treinado iterativamente para realizar previsões sequenciais de períodos de 1 mês, até cobrir todo o conjunto de testes proposto, de 8 meses e 15 dias. Entretanto, para garantir que cada mês seja previsto a partir de um modelo que captura informações desde as amostras mais antigas até o período mais recente possível, cada mês é previsto por um modelo treinado a partir do conjunto de treinamento juntamente com os meses já previstos no conjunto de teste, utilizando para isso, as próprias previsões do modelo para a energia produzida para treinar o modelo nos meses anteriores do conjunto de teste. Dessa forma, a medida que a geração de previsões avança ao longo do conjunto de testes, o conjunto utilizado para treinamento não se restringe apenas ao delimitado originalmente pela competição, mas se estende a partir das previsões já estabelecidas pelo modelo. Um esquema desse processo iterativo pode ser visto na Fig. 3, onde os blocos em verde constituem os dados utilizados para treinamento do modelo, enquanto que em amarelo se encontram os conjuntos de dados previstos pelo modelo em cada iteração.

IV. RESULTADOS E DISCUSSÕES

Nesta seção, os resultados obtidos através da aplicação da metodologia abordada são apresentados e em seguida discutidos.

A. Resultados Obtidos

Os resultados obtidos pela aplicação do método discutido na seção anterior são avaliados seguindo a mesma medida de erro proposta pela competição, denominada de Erro Percentual Absoluto Acumulado (do inglês *Cumulated Absolute Percentage Error*, ou CAPE), cujo equacionamento é dado por:

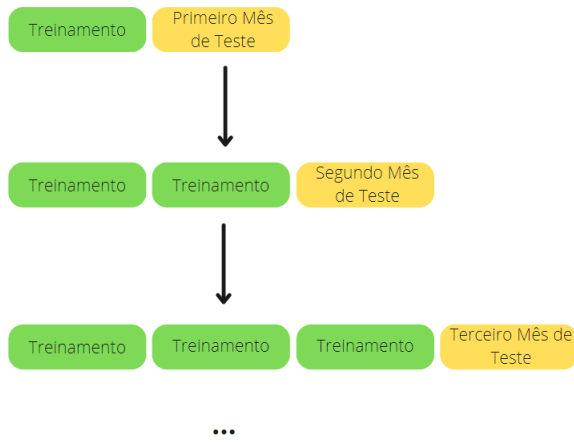


Fig. 3. Esquema de Processo de Geração de Previsões.

$$CAPE_k(\hat{Y}_k, Y_k) = 100 * \frac{\sum_{i=1}^{N_k} |Y_{i,k} - \hat{Y}_{i,k}|}{\sum_{i=1}^{N_k} Y_{i,k}} \quad (8)$$

onde $CAPE_k(\hat{Y}_k, Y_k)$ corresponde ao Erro CAPE para o parque eólico k calculado a partir dos vetores de valores observados (Y_k) e valores previstos pelo modelo (\hat{Y}_k), N_k corresponde à quantidade de amostras no conjunto de teste para o parque eólico k , e $Y_{i,k}$ e $\hat{Y}_{i,k}$ correspondem aos valores observados e previstos pelo modelo para o parque eólico k e horário i .

Nesse contexto, a aplicação do modelo selecionado na etapa de ajuste de hiper parâmetros e validação gerou um erro no conjunto de *holdout* de 35.88, de acordo com o erro CAPE. Esse resultado pode ser analisado visualmente na Fig. 4, que realiza um comparativo entre as previsões geradas pelo modelo e os valores de produção de energia reais do período. A aplicação do modelo no conjunto de teste, por sua vez, gerou um resultado de CAPE 33.45.

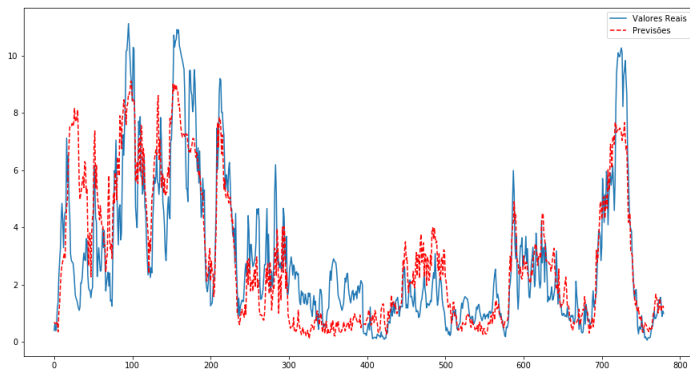


Fig. 4. Gráfico Comparativo de Previsões do Modelo no Conjunto de *Holdout*.

B. Discussão

O resultado obtido no conjunto de testes, que consiste de fato no resultado de interesse por parte dos organizadores

da competição, coloca o modelo construído neste trabalho na posição 48 do ranking final da competição. Comparativamente, o vencedor da competição apresentou um modelo com erro CAPE de 29.6168, além da própria competição apresentar um modelo de *benchmark*, que consiste no próprio modelo utilizado comercialmente pela organização promotora da competição, que apresentou um erro CAPE de 31.7269, ocupando a posição de número 30 no ranking final. Nesse quesito, pode-se perceber que os resultados obtidos não foram tão satisfatórios frente a outros modelos propostos pelos demais participantes.

Visualmente entretanto, o modelo apresentou uma performance razoável, acompanhando a tendência dos valores reais de produção, bem como algumas oscilações de frequências mais elevadas. O modelo deixa a desejar majoritariamente em regiões de baixa geração de energia, onde alguns comportamentos não foram corretamente previstos, gerando erros visualmente significativos.

V. CONCLUSÕES

Neste trabalho, um modelo de previsão de produção de energia elétrica por um parque eólico a partir de variáveis meteorológicas foi apresentado utilizando como base o XG-Boost, com o objetivo de facilitar o planejamento de operações de um sistema eólico frente à natureza aleatória do vento.

O modelo obtido foi capaz de prever com boa precisão o comportamento geral da produção da energia gerada a um custo computacional menor, se comparado aquele de outras abordagens da literatura como, por exemplo, redes neurais. Portanto, possui boa relação de custo-benefício.

Para trabalhos futuros, a metodologia adotada pode ser aplicada paralelamente a outros modelos de aprendizado de máquina, como as redes neurais, e os resultados obtidos a partir dos múltiplos modelos a serem desenvolvidos podem ser combinados adequadamente através de técnicas como, por exemplo, *stacking*, gerando previsões ainda mais precisas para a geração de energia eólica.

REFERÊNCIAS

- [1] V. S. Lopes and C. L. T. Borges, "Impact of the combined integration of wind generation and small hydropower plants on the system reliability," *IEEE Transactions on Sustainable Energy*, vol. 6, no. 3, pp. 1169–1177, 2015.
- [2] A. Tsikalakis, N. Hatzigiorgiou, Y. Katsigiannis, and P. Georgilakis, "Impact of wind power forecasting error bias on the economic operation of autonomous power systems," *Wind Energy*, vol. 12, pp. 315 – 331, 05 2009.
- [3] U. Cali and V. Sharma, "Short-term wind power forecasting using long-short term memory based recurrent neural network model and variable selection," *International Journal of Smart Grid and Clean Energy*, vol. 8, no. 2, pp. 103–110, 2019.
- [4] Y.-Y. Hong and C. L. P. P. Rioflorida, "A hybrid deep learning-based neural network for 24-h ahead wind power forecasting," *Applied Energy*, vol. 250, pp. 530–539, 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S030626191930889X>
- [5] P. Du, J. Wang, W. Yang, and T. Niu, "A novel hybrid model for short-term wind power forecasting," *Applied Soft Computing*, vol. 80, pp. 93–106, 2019.
- [6] Y. Ju, G. Sun, Q. Chen, M. Zhang, H. Zhu, and M. U. Rehman, "A model combining convolutional neural network and lightgbm algorithm for ultra-short-term wind power forecasting," *IEEE Access*, vol. 7, pp. 28 309–28 318, 2019.

- [7] P. Zhao, Q. J. Wang, W. Wu, and Q. Yang, "Which precipitation forecasts to use? deterministic versus coarser-resolution ensemble nwp models," *Quarterly Journal of the Royal Meteorological Society*, 2020.
- [8] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.