Research, Society and Development, v. 9, n. 12, e38291211251, 2020
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v9i12.11251

Previsão de Energia Eólica: Modelo de *Ensemble* Baseado em Modelos Estatísticos e de Aprendizado de Máquina

Wind Power Forecast: Ensemble Model Based in Statistical and Machine Learning Models

Previsión de energía eólica: modelo de Ensemble basado en modelos estadísticos y de aprendizaje automático

Recebido: 16/12/2020 | Revisado: 17/12/2020 | Aceito: 21/12/2020 | Publicado: 27/12/2020

Luís Gustavo Gutierrez Gebin
ORCID: https://orcid.org/0000-0003-3916-4099
Universidade Federal de Alfenas, Brasil
E-mail: lggebin@gmail.com
Ricardo Menezes Salgado
ORCID: https://orcid.org/0000-0002-0989-6259
Universidade Federal de Alfenas, Brasil
E-mail: ricardo@bcc.unifal-mg.edu.br.
Denismar Alves Nogueria
ORCID: https://orcid.org/0000-0003-2285-8764
Universidade Federal de Alfenas, Brasil
E-mail: denisnog@gmail.com

Resumo

O setor de energia é um dos pilares da sociedade moderna, tendo em vista que pode impactar o desenvolvimento de um país em diversos segmentos – econômico, social, ambiental etc. Quanto ao impacto ambiental, sabe-se que sua produção pode ocasionar resíduos e gerar malefícios para o meio ambiente, ocasionando aumento do efeito estufa e contribuindo para o aquecimento global. Neste sentido, ganha destaque o setor de energia eólica, baseado em uma produção derivada do vento e considerado “limpo”. No entanto, existe certa incerteza em relação a sua produção, por não ser possível produzir ou estocar o vento. Por conta disso, tornam-se necessários modelos de predição para que haja segurança na utilização deste tipo de produção. As predições não são tarefas triviais, visto que existem diversas variáveis que impactam sua obtenção, bem como sua estabilidade ou (instabilidade). Diversos modelos já

foram utilizados com este intuito, como os tradicionais (ARIMA), os inteligentes (XGBoost e Random Forest) e os modelos ensemble (junção de modelos), que vêm ganhando muito destaque. O objetivo do trabalho é desenvolver uma estratégia para prever a produção de energia eólica em base horária, em duas estações distintas e com diferentes modelos. Após a realização dos experimentos, pode-se perceber que os modelos individuais (com destaque para o XGBoost) apresentaram bons resultados; no entanto, deve ser notado que, na maioria dos casos de picos/outliers, estes modelos sofrem de superestimação. O modelo ensemble, entretanto, conseguiu suprir tal deficiência em relação aos modelos individuais, corrigindo os casos de superestimação.

Palavras-chave: Energia eólica; Modelos estatísticos; Aprendizado de máquina; Seleção de variáveis.

Abstract

The energy sector is one of the pillars of modern society, considering the fact that it can impact the development of a country in several segments – economic, social, environmental etc. As for the environmental impact, it is known that its production can generate waste and cause harm to the environment, causing an increase in the greenhouse effect and contributing to global warming. In this sense, the wind energy sector stands out, based on a production that is derived from the wind and is considered “clean”. However, there is much uncertainty regarding its production, as it is not possible to produce or store wind. Because of this, prediction models are made necessary, so that there is security in the utilization of this type of production. Predictions are not trivial tasks, since there are several variables that impact their achievement, as well as their stability or (instability). Several models have already been used for this purpose, such as the traditional ones (ARIMA), the intelligent ones (XGBoost and Random Forest) and the ensemble models (joint/combination models), which have been gaining prominence. The goal of the work is to develop a strategy to forecast the production of wind energy on an hourly basis, in two different stations and with different models. After carrying out the experiments, it can be seen that the individual models (especially XGBoost) present good results; however, it should be noted that, in most outlier cases, these models suffer from a problem of overestimation. The ensemble model, however, managed to gap this deficiency in relation to the individual models, correcting overestimation cases.

Keywords: Wind energy; Statistical models; Machine learning; Selection of variables.

Resumen

El sector energético es uno de los pilares de la sociedad moderna, ya que puede afectar el desarrollo de un país en varios segmentos, como el económico, el social y el ambiental. En cuanto al tema ambiental, se sabe que su producción puede causar desperdicios y generar daños al medio ambiente, lo que provoca un aumento del efecto invernadero y contribuye al calentamiento global. En este sentido, se destaca el sector de la energía eólica, en el que es una producción derivada del viento y se considera limpia, sin embargo, existe cierta incertidumbre para su producción porque no es posible producir o almacenar el viento. Como resultado, los modelos de predicción son necesarios para confiar en el uso de este tipo de producción. Sin embargo, estas predicciones no son tareas triviales, ya que hay varias variables que pueden afectar su producción, así como su inestabilidad. Ya se han utilizado varios modelos para este propósito, como el tradicional (ARIMA), inteligente (XGBOOST y Random Forest) y los modelos ensemble (combinación de modelos), en los que han ganado mucha prominencia. Así, el objetivo del trabajo es elaborar una semana de predicción, en escala horaria, de la producción de energía eólica en dos estaciones diferentes con modelos diferentes, además de aplicar un modelo de selección de variables. Usando RMSE y MAPE como medida de evaluación, se puede decir que los modelos individuales, especialmente el XGBoost, mostraron buenos resultados, sin embargo, en la mayoría de los casos se sobreestimaron en los casos de picos / valores atípicos. Sin embargo, el modelo de conjunto logró compensar esta deficiencia en los modelos individuales al corregir estos casos de sobreestimación.

Palabras clave: Energía eólica; Modelos estadísticos; Aprendizaje automático; Selección de variables.

1. Introdução

Uma das maiores preocupações no século XXI é conciliar o desenvolvimento econômico e social com a preservação ambiental. Pode-se identificar, diretamente posicionado nesta intersecção, o setor de energia, visto que, segundo a GEWC (2018) esta produção é responsável por aproximadamente dois terços das emissões globais de gases poluentes no mundo. Vale destacar ainda que, assim como há uma tendência no aumento da população, há uma tendência de aumento da demanda elétrica, o que causa preocupações em relação à preservação ambiental (além de uma preocupação social, tendo em vista a importância da eletricidade para as famílias e empresas). Estes fatores direcionaram um

avanço científico e, por consequência, um interesse maior por fontes renováveis de energia que possam suprir futuros déficits sociais e degradações da natureza. Uma fonte renovável que ganha atenção, por exemplo, é a energia eólica.

Em geral, a energia elétrica pode ser gerada de duas formas: “suja” ou “limpa”. A produção de energia “suja” ocorre quando ela resulta na liberação de resíduos ou gases poluentes que prejudicam o meio ambiente. Na produção “limpa”, por outro lado, isto não ocorre. No que concerne ao Brasil, a maior fonte de energia primária é a produção hidrelétrica, considerada “limpa”, renovável e de baixo custo de operação. Existem, entretanto, alguns pontos que merecem ser ressaltados: é uma forma de produção de energia de alto custo de implementação e que gera impacto social e ambiental devido à desapropriação de terras produtivas e pertencentes à população ribeirinha, além de ter seu potencial de produção extremamente influenciado negativamente pela seca. Devido a este comportamento sazonal, se torna indispensável o uso de alternativas para suprir esta deficiência.

A segunda maior fonte de produção de energia é a termoelétrica (pouco mais de 9% da produção total), que, segundo a ABEEólica (2020), é mais utilizada para complementar a demanda elétrica nestes períodos de baixa produção, ou seja, nos períodos de seca. Porém, sabe-se que a produção termoelétrica acarreta a liberação de CO2, um gás poluente que agrava o aquecimento global. Outro ponto negativo é o custo ambiental de sua implementação.

Já a energia eólica é uma forma de produção de energia “limpa” e renovável, pois depende da energia cinética do vento. No Brasil, segundo a ABEEólica (2020), ela representa cerca de 9% da produção total, e pode ser uma das saídas para a complementação térmica. Ainda conforme a ABEEólica (2020), o Brasil tem uma capacidade instalada de 15,4 GW, mas ainda existe um grande potencial para desenvolvimento e expansão desta produção, com cerca de 2,0 GW de capacidade já contratados e em construção (com data de finalização em um futuro próximo). Ou seja, a energia eólica já é responsável por parte significativa da produção de energia elétrica no Brasil, além de sustentar bom potencial de crescimento.

O que pode causar certa incerteza no que concerne a uma alta dependência deste tipo de produção na matriz energética brasileira é a impossibilidade de produzir ou estocar o vento. Assim, tornam-se indispensáveis modelos de previsão para a produção da energia eólica para auxiliar o sistema elétrico brasileiro, principalmente, no planejamento e na programação energética. Quanto a estes modelos de predição, podemos separá-los, segundo Lei (2009), Foley (2012) e Daraeepour e Echeverri (2014), em 4 metodologias: modelos

meteorológicos, modelos estatísticos tradicionais (Ruppert, 2011; Eldali, 2016) modelos inteligentes (Liu, 2010; Daraeepour; Echeverri, 2014; Yang, 2015) e modelos híbridos (Chang, 2016; Hansen, 1990; Salgado, 2006; Salgado; Machado, 2016; Siqueira; Salgado, 2011).

Os modelos meteorológicos são conhecidos como Numerical weather prediction (NWP) e utilizam dados meteorológicos como a temperatura, a pressão atmosférica etc. para realizar a predição. Quanto aos modelos estatísticos tradicionais, os mais utilizados são os Box-Jenkins, e dentre eles os modelos ARIMA. Porém, segundo Jurado (2013), na predição para dados não-lineares eles não apresentam bons resultados.

Os modelos inteligentes, como o XGBoost e o Random Forest, são métodos e metodologias que vêm ganhando muito destaque no âmbito de big-data e de previsões. Neste intuito, Daraeepour (2014) e Liu (2010) elaboraram modelos inteligentes para a predição de produção de energia eólica e de velocidade do vento e demonstram bons resultados.

Já os modelos híbridos podem ser entendidos como combinações de modelos, também conhecidos como ensemble. O método ensemble consiste na junção de características de modelos individuais com a finalidade da criação de um modelo com melhor desempenho (Hansen; Salamon, 1990). Segundo Salgado (2006), o objetivo principal deste método é aumentar a capacidade de generalização dos modelos individuais em um novo modelo conjunto. Salgado (2006) apresenta em seu trabalho em ensemble que apresenta uma redução de cerca de 25% nos erros em comparação com o melhor modelo individual elaborado para a predição de carga. Os modelos ensemble têm uma gama alta de alternativas e execuções. Salgado e Ohishi (2016), por exemplo, utilizaram um modelo deste tipo para tratamento e identificação de outliers em dados de carga elétrica. Assim, entende-se que a utilização do ensemble possibilita uma melhora significativa em diversos segmentos e objetivos, como os supracitados, para auxílio na predição de produção de energia.

Outro ponto que faz-se necessário destacar é o que se relaciona às diversas covariáveis que podem afetar o modelo de previsão, como o uso de dados climáticos. Como Daraeepour e Echeverri (2014) e Li, Wang e Goel (2015) afirmam, além de ter um modelo bem estruturado, também se torna indispensável a seleção das variáveis que serão inseridas nele. Pois a inserção de muitas covariáveis pode até mesmo atrapalhar a eficácia do modelo no que se refere à predição e deixar o algoritmo lento.

Tendo isso em vista, esta pesquisa objetivou a predição de duas semanas em escala horária da produção de energia eólica (primavera e verão), utilizando apenas o histórico de produção, com os modelos inteligentes (XGBoost e Random Forest) bem como utilizando um

modelo estatístico tradicional (ARIMA), para posteriormente aplicar nos modelos inteligentes um algoritmo de seleção de variáveis, além de propor um ensemble com os modelos individuais.

Este artigo encontra-se estruturado da seguinte forma: na Seção I foram introduzidos os modelos de predição e sua importância para o sistema elétrico. A Seção II discorrerá acerca dos materiais e métodos. A princípio, apresentar-se-á os modelos utilizados e, por fim, o modelo proposto. Quanto à Seção III, os dados serão apresentados e, posteriormente, os resultados (separados por estação), seguidos de uma análise geral da performance dos modelos. A Seção IV apresenta a conclusão acerca da qualidade dos modelos.

2. Fundamentação Teórica

Nesta seção será apresentada a fundamentação teórica dos modelos individuais utilizados para a construção do *ensemble* (*ARIMA*, *XGBoost* e *Random Forest*), além de ser apresentado o modelo de seleção de variáveis utilizado para os modelos inteligentes (*XGBoost* e *Random Forest*).

Modelos de Box-Jenkins (*ARIMA*)

De acordo com Morettin e Toloi (2006), uma série temporal é composta por quatro elementos: 1) **tendência**: verifica o sentido de deslocamento da série ao longo de vários tempos; 2) **ciclo**: movimento ondulatório que ao longo de vários anos tende a ser periódico; 3) **sazonalidade**: movimento ondulatório de curta duração, em geral, inferior a um ano. Está associada, na maioria dos casos, a mudanças climáticas, e; 4) **ruído aleatório ou erro**: compreende a variabilidade intrínseca dos dados.

Quanto às séries temporais, Tubino (2000) afirma que estas previsões são os métodos mais simples e usuais e que partem da premissa de que os valores futuros serão projeções de seus valores passados.

Uma série temporal tem os dados coletados sequencialmente ao longo do tempo, e espera-se que ela apresente uma correlação seriada no tempo. Os modelos de Box-Jenkins, genericamente conhecidos pela sigla *ARIMA* (*Auto Regressive Integrated Moving Averages*) representam um tipo de modelagem para ajustes de séries temporais, e são modelos matemáticos do tipo polinomial que visam captar o comportamento da correlação seriada (ou autocorrelação) entre os valores da série temporal. Com base no comportamento obtido, são

realizadas previsões futuras. Se essa estrutura de correlação for bem modelada, ela fornecerá boas previsões (Wener; Ribeiro, 2003).

O modelo *ARIMA*, desenvolvido por George Box e Gwilym Jenkins, se baseia no ajuste dos valores observados, para então reduzir o erro das predições com os valores observados para próximo de zero. Com este modelo, deve-se trabalhar com séries temporais estacionárias, ou seja, séries temporais em que a média e a variância são constantes e em que o valor da covariância depende da diferença entre os dois períodos do tempo.

Em se tratando dos modelos *ARIMA*, sabe-se que estes possuem uma extensa variedade de possibilidades. Sua forma mais usual não sazonal é o modelo *ARIMA*(p,d,q).

Para a construção deste modelo (p,d,q), conforme Morettin e Toloi (2006), divide-se o processo em três etapas: identificação, estimação e, por fim, verificação. Tais etapas se dão a partir da observação da própria série de dados.

Primeiro, deve verificar-se qual versão dos modelos de Box-Jenkins descreve melhor o comportamento da série estudada, após a verificação das funções de autocorrelação (ACF) e das funções de autocorrelação parcial (PACF). A função de autocorrelação mostra o quão forte o valor observado de hoje está correlacionado com os valores do passado. A função de autocorrelação parcial mostra a correlação entre a variável no instante t e uma de suas defasagens. Posteriormente, estima-se os parâmetros (p,d,q) e, por fim, avalia-se a adequação do modelo e se ele descreve o comportamento dos dados de forma correta (Wener; Ribeiro, 2003).

Assim, a combinação da parte regressiva com o grau de diferença e o grau de parte média móvel expressa o modelo *ARIMA*.

Modelos inteligentes – *Classification and Regression Trees (CART)*

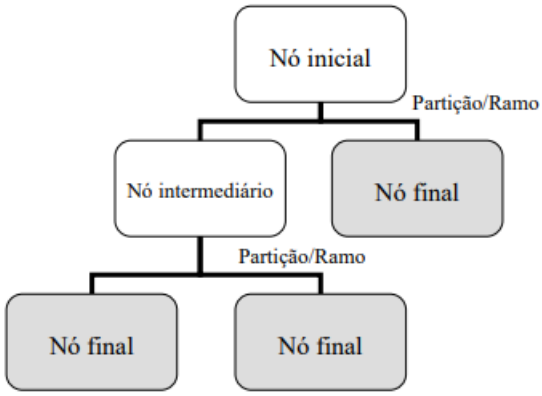
Os modelos *CART* são ótimas alternativas para funções de predição, devido à sua multifuncionalidade e à simplicidade de sua árvore. Esses modelos podem ser utilizados para a explicação de uma variável resposta numérica, como nos casos de regressão, ou categorizada, nos casos de classificação.

Este método, basicamente, se baseia na execução de sucessivas partições binárias de uma amostra, com base nos resultados das covariáveis, buscando a constituição de subamostras internamente homogêneas.

Existem certos termos que são indispensáveis para o entendimento e para a caracterização dos componentes de uma árvore de regressão ou classificação. A lista de termos compreende os seguintes: nó inicial, nós intermediários e nós finais.

O nó inicial se relaciona à amostra original, os nós intermediários às subamostras que originam novas subamostras e os nós finais às subamostras não partidas. As partições executadas são denominadas ramos. Por fim, a árvore é a representação gráfica destes nós e ramos, como visto na Fig. 1.

Figura 1. Ilustração de uma árvore de classificação/regressão.



Fonte: Breiman (2001).

Os modelos inteligentes *Random Forest* e *XGBoost*, presentes nas apresentações da próxima seção, são variações de modelos *CART*.

Modelos inteligentes do tipo *Random Forest*

O *Random Forest* foi desenvolvido por Leo Breiman e, segundo Diaz-Uriarte e Andres (2006) e Breiman (2001), é um algoritmo de aprendizagem de máquina que utiliza o método *ensemble* nas camadas finais da regressão em árvore. É um método que pode ser utilizado para regressão e para classificação. O algoritmo utiliza uma floresta de árvores formada através de entradas aleatórias, com ou sem reposição, e, por fim, a combinação dessas árvores por meio da média (Breiman, 2001).

Pode-se entender então que, apesar de utilizar um processo similar ao dos modelos *CART*, o *Random Forest* gera uma grande quantidade de árvores, para no final combiná-las.

Seleção de variáveis – *Importance* (*Random Forest*)

O método *VarImportance* tem como intuito evidenciar quais foram as variáveis que mais contribuíram para a explicação da variável resposta.

É utilizado o denominado erro *Out-of-bag (OOB)*, que é um método para mensurar os erros de previsão das florestas aleatórias utilizando agregação via *bootstrap* para novas subamostras advindas das amostras de dados utilizadas no treinamento do modelo. Basicamente, o *OOB* se trata do erro médio de previsão em cada amostra de treinamento.

Esta subamostragem permite definir uma estimativa "*Out-of-bag*", ou seja, fora da amostra original, que melhora o desempenho de previsão, avaliando as previsões sobre essas observações que não foram utilizadas para a construção e trazendo-as para o próximo aprendizado do modelo.

Modelos inteligentes – *XGBoost*

O *XGBoost*, abreviação para *Extreme Gradient Boost*, é uma biblioteca de aprendizado de máquina que pode ser utilizada para problemas de regressão e de classificação. Segundo Chen e Guestrin (2016), este método vem ganhando notoriedade, sendo reconhecido e amplamente utilizado por cientistas de dados. O *XGBoost* foi introduzido por Friedman (2001) baseado no método *Gradient Boosted Trees* (*GBT*). Também pode ser utilizado na resolução de problemas de aprendizagem supervisionada de máquina. O *GBT* é um método que pode ser entendido como um *ensemble* que produz resultados baseados em combinações de árvores de decisão.

O *Boosting*, segundo Friedman (2001), que é um processo utilizado no *XGBoost*, tem como objetivo a melhora dos resultados. Assim, o algoritmo baseia-se na ideia de combinar classificadores genéricos e menos específicos, para construir classificadores mais robustos. Além disso, também pode ser verificada a existência da função objetivo, que auxilia o classificador a alcançar o melhor resultado, com auxílio das características e dos parâmetros. A função objetivo é formada por uma função de perda somada a uma função de regularização (Chen; Guestrin; 2016). Ao longo das iterações, o *XGBoost* procura encontrar uma árvore de decisão que minimize a função objetivo.

3. Metodologia Proposta

O trabalho se baseia em métodos quantitativos, que, de acordo com Pereira A.S. et al. (2018), são utilizados quando se coletam e analisam dados quantitativos ou numéricos. A partir desta primeira etapa de coleta é gerado um conjunto de dados que pode ser analisado por meio da estatística ou matemática. Seguindo o pensamento de Pereira A.S. et al. (2018), é comumente utilizado para possibilitar previsões de acontecimentos.

Assim, o enfoque deste trabalho baseia-se na criação de um modelo preditivo (método quantitativo), combinando 3 modelos individuais: *ARIMA*, *XGBoost* e *Random Forest*. O trabalho também utilizará um modelo de seleção de variável para os modelos inteligentes (*XGBoost* e *Random Forest*), bem como apresentará os resultados de cada modelo individual, antes e após a seleção de variáveis.

Ensemble

A elaboração do trabalho proposto pode ser dividida em 4 etapas: elaboração da base de dados, elaboração dos modelos individuais, seleção das variáveis (sendo que esta etapa se aplica apenas aos modelos inteligentes), e, por fim, combinação dos melhores modelos, como visto na Figura 2.

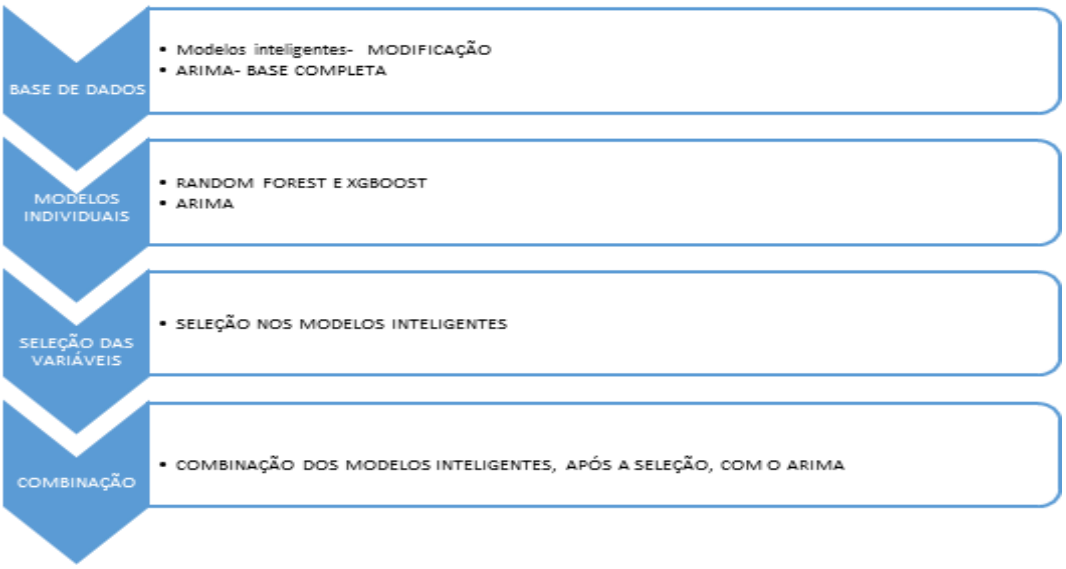
A primeira etapa, *elaboração da base de dados***,** se deu de forma diferente para os modelos inteligentes e para o *ARIMA*. Para o *ARIMA*, o histórico inteiro foi utilizado. Quanto aos modelos inteligentes, o histórico foi transformado em covariáveis. Como se sabe da dependência temporal destes dados para predizer a hora t, foram utilizadas como variáveis explicativas os dados pertencentes ao histórico de duas semanas até a hora imediatamente anterior à hora t (ou seja, de t-14 até t-1). Desta forma, entende-se que houve uma modificação na estrutura da base para se adequar às especificidades dos modelos. Assim, o modelo pode ser entendido como segue na Equação 1.

Equação 1

$$T_0 = \{T_{-1}; T_{-2}; \dots; T_{-x}\} \qquad (1)$$

Sendo T_0 a produção de energia eólica na hora 0.

Figura 2. Construção das etapas.



Fonte: Autores.

Na segunda etapa, *elaboração dos modelos individuais*, se sucedeu da seguinte maneira: 3 modelos foram elaborados: um *ARIMA*, um *Random Forest* e um *XGBoost*, sendo que, no caso dos dois últimos, a elaboração utilizou 2 semanas de histórico como fonte de variáveis explicativas (t-1, ... , t-14). O modelo *ARIMA* foi elaborado seguindo os critérios clássicos de seleção de modelos. Os hiperparâmetros dos modelos inteligentes foram selecionados a partir de diversas simulações.

Para a terceira etapa, *seleção de variáveis*, foi utilizado o *VarImportance*, pacote do *RandomForest* (conforme já apresentado anteriormente neste artigo). A seleção foi aplicada aos modelos *XGBoost* e *Random Forest*, e, após a seleção, foram utilizadas como variáveis explicativas apenas as 5 últimas observações (ou seja, apenas as últimas 5 horas anteriores para predizer a próxima hora).

Até este ponto dos procedimentos, portanto, são elaborados os seguintes modelos: *ARIMA*, *XGBoost* sem seleção de variáveis, *XGBoost* com seleção de variáveis, *Random Forest* sem seleção de variáveis e *Random Forest* com seleção de variáveis.

Para a **última etapa**, *construção do ensemble*, utilizou-se para treino do modelo 700 horas preditas antes de cada semana alvo, além do histórico da produção de energia eólica. Estas predições foram feitas com os modelos *XGBoost*, *Random Forest* e o *ARIMA*. Assim, entende-se que o *ensemble* apresenta 4 variáveis explicativas. O banco de dados utilizado para alimentar o *ensemble*, portanto, será composto por predições dos modelos supracitados e pelo

histórico da produção de energia eólica imediatamente anterior a estas predições e à predição de interesse do modelo *ensemble*, como segue na Equação 2,

Equação 2

$$Ensemble_X = \{ARIMA_X; RF_X; XGB_X; T_{X-1}\} \quad (2)$$

... na qual $Ensemble_X$ se refere à predição do *ensemble* na hora x; $ARIMA_X$, à predição do *ARIMA* na hora x; RF_X, à predição do *Random Forest* na hora x; XGB_X, à predição do *XGboost* na hora x e o T_{X-1} à produção de energia eólica observada até a hora imediatamente anterior à hora x (hora x-1).

Para melhor explicar a metodologia, cabe aqui um exemplo: para predizer a hora 2 do dia 30 de junho, o modelo utiliza como variáveis explicativas as predições dos modelos citados até esta mesma hora e dia, além do histórico real até a hora imediatamente anterior (ou seja, a hora 1 do dia 30 de junho). O banco de dados, então, será composto por 700 predições de cada modelo (em escala horária) e 700 horas do histórico real. Assim, as predições para a construção do banco são iniciadas na hora 21 do dia 31 de maio; ou seja, 700 horas anteriores à semana alvo. Com isso, essas predições se transformam em informações para o *ensemble*. Para combiná-las, foi utilizado o *XGBoost.*

Tendo em vista aprofundar ainda mais a exemplificação da metodologia, apresentar-se-á um pseudocódigo de construção do modelo ensemble do exemplo supracitado:

Construção do pseudocódigo (construção dos modelos Individuais e do Ensemble).

Esta subseção abordará os históricos utilizados para treino e teste. Antes de apresentar o *ensemble*, é importante entender como os modelos individuais foram criados. Entender a construção dos modelos individuais e as nomenclaturas que serão apresentadas facilita o entendimento do *ensemble*.

1) **Construção dos modelos individuais**

- **Passo 1:** Construção dos modelos individuais;
- **Passo 1.a:** Função(Modelo; Período de treino; Período de teste). Esta função escolherá o modelo, o período de treino e o período de teste. A função

retornará a predição do período de teste em escala horária. Caso o período de teste seja maior do que 1, o resultado será um vetor com as predições, denotado por SaídaFunção'Modelo'[].

- **Passo 1.b:** Hist<- Histórico dos dados em escala horária. Para se referir ao histórico dos dados, dar-se-á as seguinte nomenclatura: Hist[1:868] – se refere ao vetor com o histórico dos dados (primeira hora observada até a octingentésima sexagésima oitava hora observada).
- **Passo 1.c:** Separar o histórico em treino e teste. Considerando o histórico supracitado, entende-se que as últimas 168 observações, uma semana em escala horária, referem-se à semana alvo. Com isso, Hist[701:868] será entendido como o período de teste e Hist[1:700] como o período de treino.
- **Passo 1.d:** Criar nomenclaturas para os períodos de treino e teste, conforme segue: o período de treino será Hist[1:700], denominado a partir de agora como Treino[1:700], e o teste será Hist[701:868], denominado a partir de agora como Teste[1:168]; neste contexto, o resultado do modelo será composto pelas predições do período do teste e será denotado por SaídaFunção'Modelo'[1:168] (ou seja, o período que a função irá predizer será Teste[1:168] – as próximas 168 horas).
- **Passo 1.e:** constuir o modelo conforme mostrado em 1.a. mas atualizando o contador para um passo a frente: i indo de 0 até 167 (para realizar uma predição da hora 1 até a 168);Função(RF; Treino[1:700+i];Teste[1+i]). Com isso, o i foi escolhido para fazer a predição 1 hora à frente. Ou seja, após a predição do Teste [1], o modelo utilizará para predizer Teste [2] as informações do Treino atualizadas (Treino[1:700+1]). A saída será um vetor com as predições com a seguinte nomenclatura: SaídaFunçãoRF[1:168].
- **Passo 1.f:** repetir os passos para todos os outros modelos individuais.

Conforme abordado na subseção anterior, cada modelo conta com suas peculiaridades, como, por exemplo, o número de covariáveis utilizadas para o caso dos modelos inteligentes, ou seja, quantas horas anteriores foram utilizadas para predizer a próxima hora (além dos seus hiperparâmetros). Neste pseudocódigo, estas peculiaridades não foram levadas em consideração, mas apenas a lógica para construção do período de treino e de teste.

- **Passo 2:** Após a criação dos modelos individuais, começará a construção do *ensemble*. Para isto, a função para predição será mais elaborada, com mais variáveis. Nos modelos individuais, além dos hiperparâmetros, foram utilizados apenas os históricos. Já no novo modelo *ensemble*, outras 3 variáveis serão utilizadas para alimentar o modelo: O *vetor de predição de cada modelo individual*, ou seja, a *SaídaFunção'RF'[]*, *SaídaFunção'XG'[]* e *SaídaFunção'ARIMA'[]*. Assim, para elaborar o novo *ensemble*, 3 novas base de dados (SaídaFunçãoModelo[]) serão utilizadas para alimentar o modelo e seguirão a mesma lógica. Com isso, continuando o exemplo apresentado pelos modelos individuais, a função para o ensemble seria:
- **Passo 2.a:** i indo de 0 até 167;

FunçãoEnsemble(SaídaFunçãoRF[1+i:700+i];SaídaFunçãoXG[1+i:700+i];
SaídaFunçãoARIMA[1+i:700+i];Treino[1:700+i];Teste[1+i]).

Assim, espera-se que o *ensemble* consiga entender, principalmente, nos casos de maiores variações, os erros mais comuns, e assim diminui-los. Além disso, espera-se que o *ensemble* criado possa captar informações de cada modelo com aproximações diferentes, tornando-o mais robusto e generalizado.

4. Resultados e Discussão

Esta seção será dividida em 6 subseções: A Análise dos dados e matérias apresentará as fontes de extração dos dados e os softwares que foram utilizados. Na subseção Análise descritivas das semanas a serem preditas, análise descritiva dos valores que serão preditos pelos modelos será apresentada. A subseção Parâmetros dos modelos apresentará os parâmetros para os modelos individuais. Quanto à subseção Métricas de avaliação, serão apresentadas as métricas utilizadas para validar os resultados dos modelos. E, por fim, nas subseções Apresentação e discussão das predições em junho e Apresentação e discussão das predições em março, são exibidas as performances de cada modelo nas semanas de junho e março, respectivamente.

Análise dos dados e materiais

Os dados contemplam a produção de energia eólica de 2011 a 2016, anotados a cada hora. Os dados foram disponibilizados a partir do

<https://transmission.bpa.gov/business/operations/wind/>, e contemplam a geração de energia eólica observada nos Estados Unidos. Foram escolhidas para a predição duas semanas em diferentes estações do ano. Os períodos para predição escolhidos foram: do dia 9 de março até o dia 15 de março de 2016 e do dia 9 de julho até o dia 15 de julho de 2016.

Para elaboração do trabalho, utilizou-se o software R para análises e modelagens. Quanto aos modelos Box-Jenkins, utilizou-se os pacotes 'forecast' e 'tseries'. Os modelos de *Random Forest* e o *VarImportance* pelo pacote 'randomForest', e o *XGboost* pelo pacote 'XGboost'.

Análise descritivas das semanas a serem preditas

As duas semanas-alvo foram escolhidas por apresentarem comportamentos substancialmente distintos entre elas no que se refere a distribuição, média, mediana, *outliers* etc. Assim, testou-se o *ensemble* construído, que teve desempenho satisfatório em casos distintos de padrão, verificando assim sua robustez e capacidade de generalização.

A Tabela 1 apresenta as estatísticas descritivas das semanas-alvo. Além disso, pode-se observar os *boxplots* delas na Figura 3.

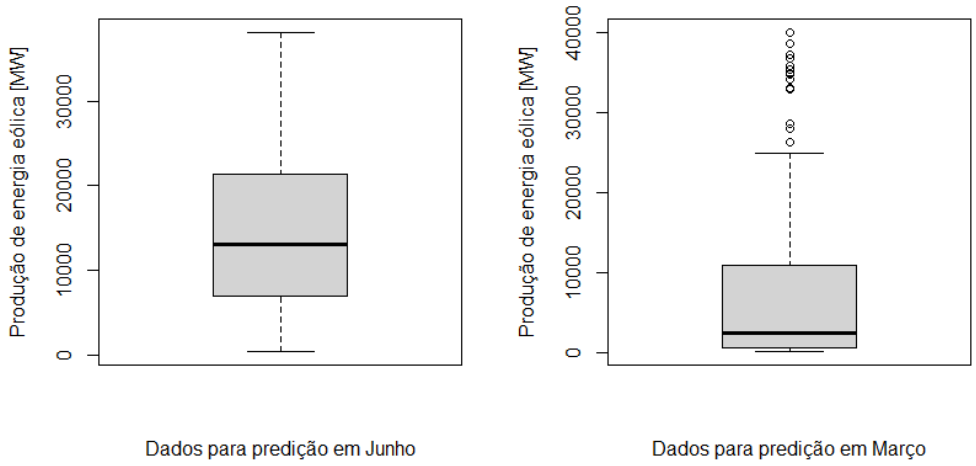
Tabela 1. Medidas descritivas da produção (MW) das duas semanas selecionadas para predição

Semanas	Média	Min	Máx	Maior amplitude*
Março	7.478,40	50	40.104	13.818
Junho	14.700,07	339	38.134	8.378

*Maior amplitude entre duas observações seguidas.
Fonte: Autores.

A partir dos *boxplots*, podem ser pontuadas algumas diferenças entre as duas semanas-alvo. A semana a ser predita em março, conforme a Figura 3, apresentou diversos *outliers*, enquanto a semana referente a junho não apresentou nenhum *outlier*. Além disso, a cauda das duas distribuições é diferente. As semanas apresentam distinções nos quartis, na mediana e na média. Ademais, a semana de março apresenta uma amplitude interquartílica menor, ou seja, quando são retirados os *outliers*, a amplitude da semana é menor do que a de junho.

Figura 3. Boxplot das semanas a serem predita em março e junho.



Fonte: Autores.

Assim, devido aos pontos supracitados, entende-se que os valores a serem preditos em março apresentarão maior dificuldade para os modelos, devido à sua instabilidade e números de *outliers*.

Parâmetros dos modelos

Os parâmetros utilizados na construção do modelo *Random Forest* podem ser observados na Tabela 2, e os parâmetros para o *XGBoost* nas Tabelas 3 e 4. Para o *ARIMA*, os parâmetros escolhidos são mencionados logo após a apresentação das tabelas.

Tabela 2. Parâmetros utilizados para os modelos *Random Forest.*

Ntree [1]	***nodesize*** [2]	***Mtry*** [3]	***nPerm*** [4]	***Corr.bias*** [5]
500	12	9	1	6

1. ***Ntree:*** Número de árvores que serão geradas. Para este parâmetro não deve ser definido um número muito pequeno, para garantir que cada linha de entrada seja prevista pelo menos algumas vezes. É importante notar, no entanto, que um número muito alto pode causar *overfitting*.
2. ***Nodesize*:** Tamanho mínimo dos nós terminais. Definir um número elevado para este parâmetro faz com que árvores menores sejam cultivadas (e que, portanto, levem mais tempo).
3. ***Mtry***: Número de variáveis amostradas aleatoriamente como candidatas em cada divisão.
4. ***NPerm:*** Número de vezes que os dados *Out-of-bag* (*OOB)* são permutados por árvore para avaliar a importância da variável. Um número maior que 1 fornece uma estimativa ligeiramente mais estável, mas não muito eficaz. Mais comumente utilizado para regressão.
5. ***Corr.bias:*** Correção de viés para regressão.
Fonte: Autores.

Tabela 3. Parâmetros utilizados para os modelos *XGBoost 1*.

Nrounds [1]	***Nthread*** [2]	***Max_depth*** [3]	***Eta*** [4]
500	2	10	0,02

1. ***Nrounds***: número máximo de interações.
2. ***Nthread***: número de encadeamentos paralelos usados para executar o *XGBoost*.
3. ***Max_depth***: profundidade máxima de uma árvore de regressão; aumentar esse valor tornará o modelo mais complexo, mas pode causar *overfitting*.
4. ***Eta***: encolhimento do tamanho do passo usado na atualização para evitar *overfitting*. O *Eta* reduz os pesos dos recursos para tornar o processo de reforço mais conservador. Retorna um número entre 0 e 1.
Fonte: Autores.

Tabela 4. Parâmetros utilizados para os modelos *XGBoost 2*.

Min_child_weigh [1]	***Subsample*** [2]	***Objective*** [3]
9	0,2	"reglinear"

1. ***Min_child_weight***: soma mínima do peso da instância (*hessian*) necessária em uma *child*. Se a etapa de partição da árvore resultar em um nó folha com a soma do peso da instância menor que *min_child_weight*, o processo de construção dará mais particionamento. No modo de regressão linear, isso simplesmente corresponde ao número mínimo de instâncias necessárias em cada nó. Quanto maior, mais conservador será o algoritmo.
2. ***Subsample:*** proporção da *subsample* da instância de treinamento. Utilizar um *subsample* < 0,5 pode evitar o *overfitting*.
3. ***Objective***: tipo escolhido para aproximação.
Fonte: Autores.

Para o modelo ARIMA foi ajustado o modelo (1,1,1) entre os modelos candidatos.

Métricas de avaliação

As métricas utilizadas para avaliar a qualidade das predições dos modelos ajustados em relação ao objetivo foram a *Mean Absolut Percentage Error – MAPE* (Equação 3) e o *Root mean squared error – RMSE* (Equação 4).

Equação 3

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \quad (4)$$

Equação 4

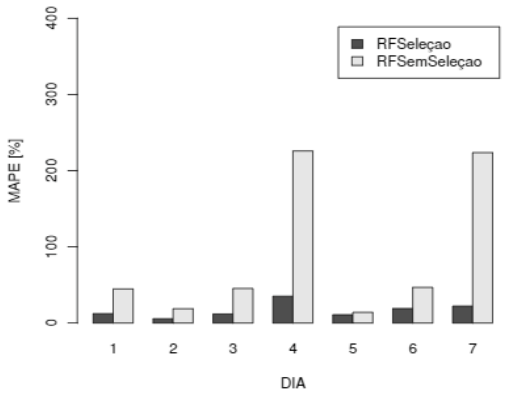
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (5)$$

Onde $\hat{y}_i$ representa o valor previsto e y_i representa o valor observado.

Apresentação e discussão das predições em junho

Em relação à seleção de variáveis, foram escolhidas as últimas 5 horas anteriores como as variáveis explicativas. Para o modelo sem a seleção, as covariáveis selecionadas foram as últimas duas semanas de histórico, ou seja, as 336 horas anteriores. O desempenho desta seleção de variáveis nos modelos *Random Forest* e *XGBoost*, respectivamente, pode ser vistos nas Fig. 4 e 5, que apresentam o *MAPE* separado por dia.

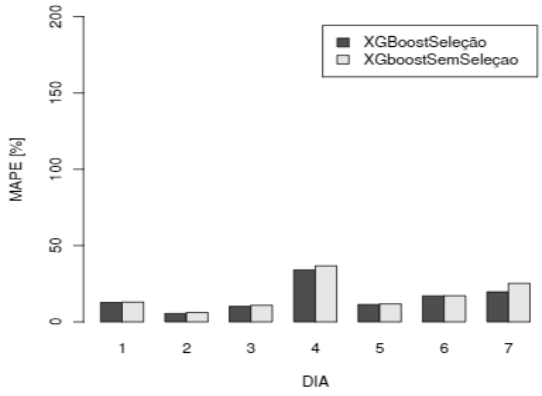
Figura 4. MAPE das dos dias dos modelos *Random Forest* com seleção e sem seleção em junho.



Fonte: Autores.

Conforme observado na Figura 4, o *Random Forest* foi muito favorecido pela seleção, visto que o seu modelo sem a seleção de variáveis não conseguiu apresentar bom desempenho.

Figura 5. MAPE das dos dias dos modelos *XGBoost* com seleção e sem seleção em junho.



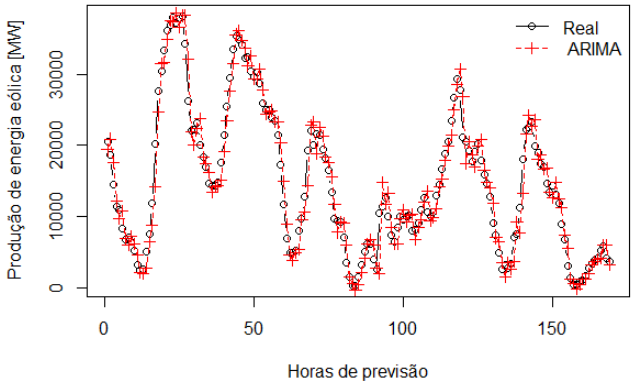
Fonte: Autores.

Como visto na Figura 5, o *XGBoost* apresentou uma melhora mais sutil. Ou seja, podemos entender que o *Random Forest*, devido ao seu método de aproximação, não conseguiu, com as especificações apresentadas, demonstrar bom desempenho. As muitas variáveis atrapalharam o modelo em relação à predição.

Quanto às predições dos modelos individuais ao longo da semana, deve ser ressaltado que nas Fig. 6, 7, 8, 9 e 10 percebe-se um padrão: sempre que existe um pico, os modelos *Random Forest*, *XGBoost* e *ARIMA* supervalorizam suas predições, principalmente quando a amplitude entre as predições é alta.

Como pode-se destacar na Fig. 11, o modelo *ensemble* se diferencia dos outros modelos individuais, principalmente, por conseguir realizar predições melhores nestes casos de extremos.

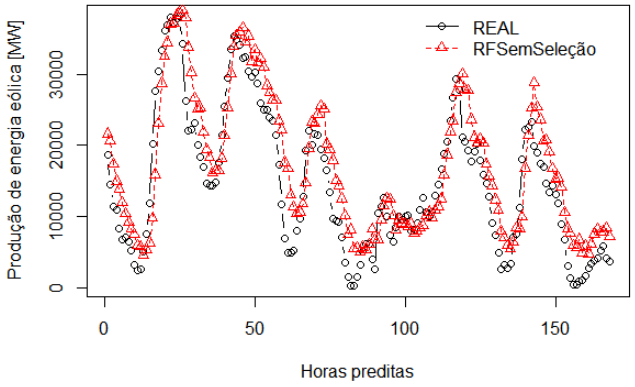
Figura 6. Resultados das predições do modelo *ARIMA*(1,1,1) para junho.



Fonte: Autores.

Conforme Figura 6, entende-se que o ARIMA obteve um resultado satisfatório, porém nos momentos de pico, o mesmo obteve uma superestimação.

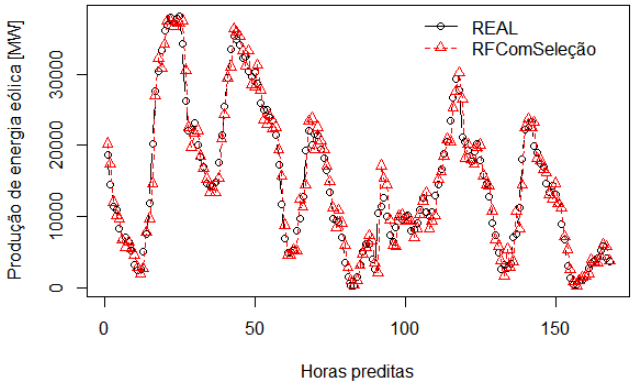
Figura 7. Resultados das predições dos modelos *Random Forest* sem seleção para junho.



Fonte: Autores.

Na Figura 7, observa-se que o *Random Forest* sem a seleção de variáveis obteve um desempenho pior do que os outros modelos.

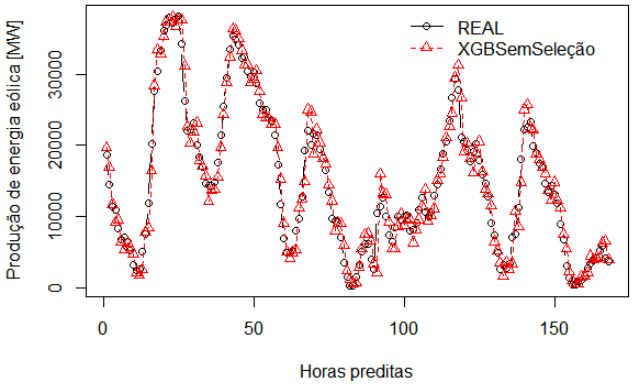
Figura 8. Resultados das predições dos modelos *Random Forest* com seleção para junho.



Fonte: Autores.

Já na Figura 8, é notório a evolução na precisão do modelo *Random Forest* após a seleção de variáveis.

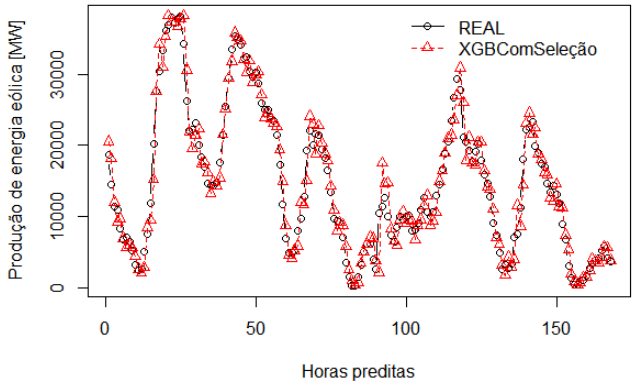
Figura 9. Resultados das predições dos modelos *XGBoost* sem seleção para junho.



Fonte: Autores.

Conforme pode ser observado na Figura 9, o *XGBoost* apresentou um resultado muito mais satisfatório em relação ao *Random Forest* sem seleção.

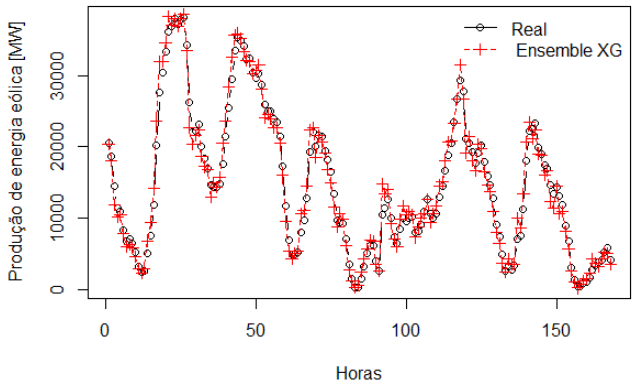
Figura 10. Resultados das predições dos modelos *XGBoost* com seleção para junho.



Fonte: Autores.

Na Figura 10, nota-se que não existiu uma diferença notória após a seleção. Além disso, é percebido que em todos os modelos, ocorreu uma superestimação nos casos de *outliers*.

Figura 11. Resultado das predições do modelo *ensemble* proposto para junho.



Fonte: Autores.

A Figura 11 apresenta o modelo *ensemble* proposto, e é possível observar um alinhamento entre a provisão e os valores observados, além disso, nota-se que, nos casos extremos, o modelo não superestimou, diferente dos modelos individuais.

Na Tabela 5, no que se refere ao *MAPE* das 168 horas de predições dos dados em junho, o *XGBoost* com seleção apresentou desempenho melhor que os outros modelos individuais, mas vale ressaltar que o *Random Forest* obteve um melhor aproveitamento com

seleção de variáveis, sendo que o modelo sem seleção de variáveis não conseguiu propor razoavelmente as predições da semana, apresentando assim o pior resultado entre os modelos estudados.

Outro ponto importante é acerca dos modelos *XGBoost*, visto que, mesmo sem a seleção, são modelos que mantiveram um resultado relevante. Ou seja, mesmo com as diversas variáveis explicativas, o *XGBoost* pôde ponderar bem as variáveis para obter um resultado relevante. O *ARIMA*(1,1,1) conseguiu um resultado (referente ao *MAPE* e o *RMSE*) próximo aos dos modelos inteligentes. Já o *ensemble* obteve um desempenho acima dos modelos individuais em ambas as métricas de avaliação.

Tabela 5. Resultados das métricas de avaliaçao nas 168 horas preditas em junho.

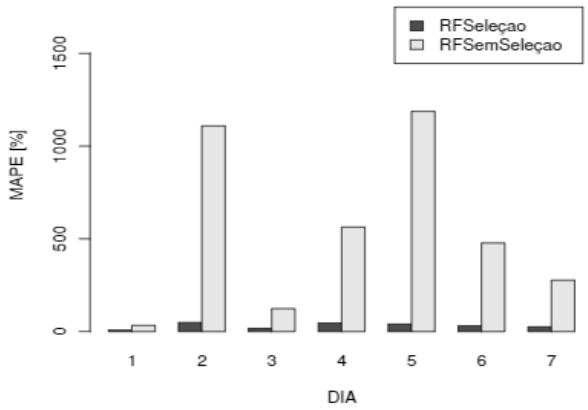
Modelos	***MAPE* [%]**	***RMSE* [MW]**
Random Forest	16	1930,97
***Random Forest* sem seleção**	88	4694,61
XGBoost	15	1881,39
***XGBoost* sem seleção**	17	1900,27
ARIMA(1,1,1)	19	1982,23
Ensemble XG	8	1028,69

Fonte: Autores.

Apresentação e discussão das predições em março

Nas Fig. 12 e 13 apresentam-se os resultados das seleções de variáveis dos modelos *Random Forest* e *XGBoost*, respectivamente.

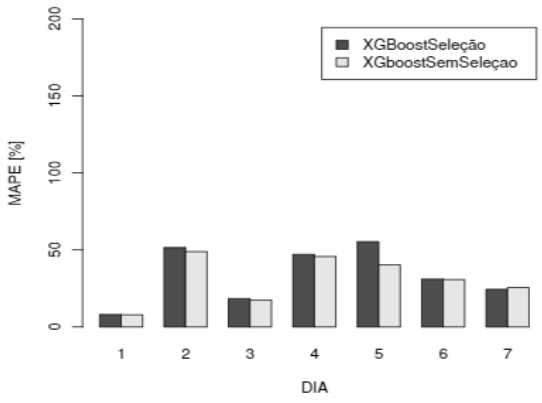
Figura 12. MAPE das dos dias dos modelos *Random Forest* com seleção e sem seleção na estação em março.



Fonte: Autores.

Assim como na semana-alvo de junho, percebe-se que o *Random Forest* foi muito mais beneficiado pela seleção, podendo entender que este modelo, com estas especificações, não conseguiu predizer bem os dados no caso com muitas variáveis.

Figura 13. MAPE das dos dias dos modelos *XGBoost* com seleção e sem seleção na estação em março.

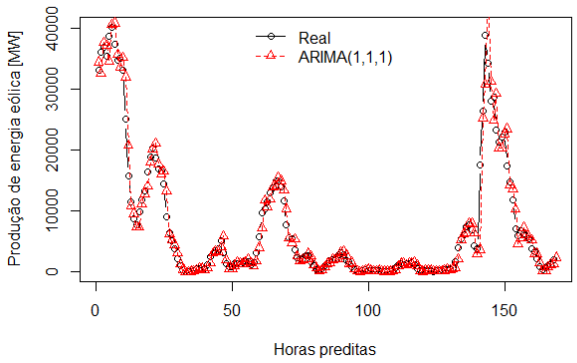


Fonte: Autores.

Percebe-se o modelo *Random Forest* (Figura 12) não tem um desempenho satisfatório quando se trata de diversas variáveis explicativas, enquanto o *XGBoost* (Figura 13) não apresenta tal deficiência. Vale ressaltar que, apesar de ambos serem modelos de regressão em árvore, o método de aproximação de cada um tem suas limitações e especificações.

Assim como nas outras predições, pode ser percebida nas Fig. 14, 15, 16, 17 e 18 a supervalorização dos modelos *ARIMA(1,1,1), Random Forest* e *XGBoost* nos picos. Vale destacar também que o modelo *Random Forest* sem a seleção de variáveis não conseguiu interpretar bem os dados e apresentou pior resultado em comparação com os demais (em relação às métricas de avaliação). Além disso, vale destacar a melhora do desempenho pós-seleção. Quanto ao *ensemble*, como pode ser verificado na Fig. 19, uma melhora foi apresentada em relação aos extremos. Assim, o *ensemble* nas predições das duas semanas obteve desempenho melhor do que os modelos individuais nestes picos, o que se tratava de uma deficiência comum entre eles.

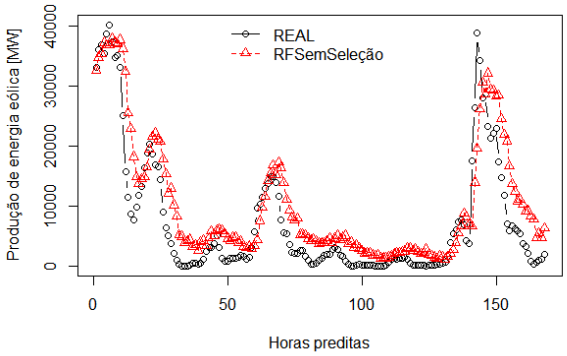
Figura 14. Resultado das predições do modelo *ARIMA (1,1,1)* para a estação em março.



Fonte: Autores.

Conforme citado anteriormente, o modelo ARIMA (1,1,1) obteve resultado satisfatório. Entretando, nos picos, é possível observar uma superestimação da previsão.

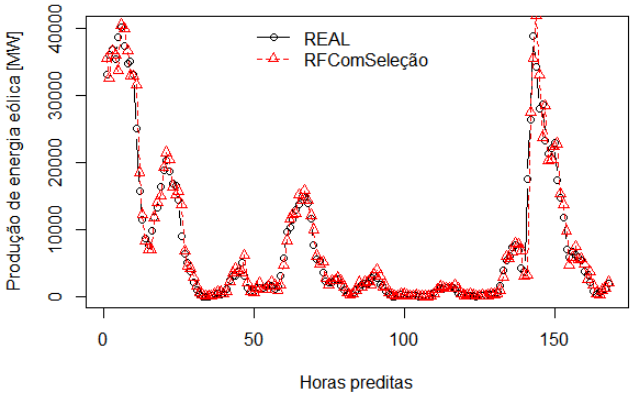
Figura 15. Resultados das predições dos modelos *Random Forest* sem seleção para em março.



Fonte: Autores.

Como visto na Figura 15, o modelo *Random Forest* sem a seleção de variáveis não obteve um resultado satisfatório, assim como na predição anterior.

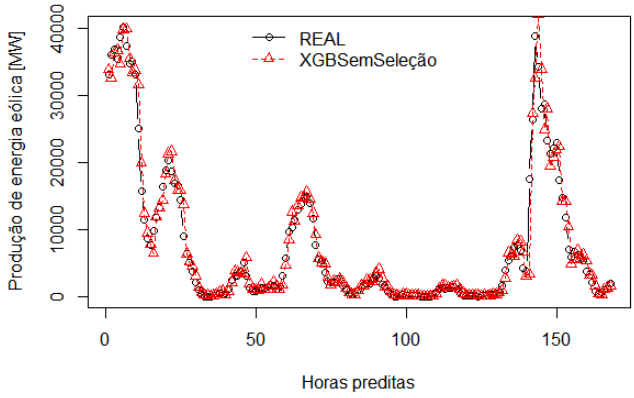
Figura 16. Resultados das predições dos modelos *Random Forest* com seleção para em março.



Fonte: Autores.

Após a seleção de variáveis, o modelo apresentado na Figura 16 obteve um resultado muito próximo aos valores observados.

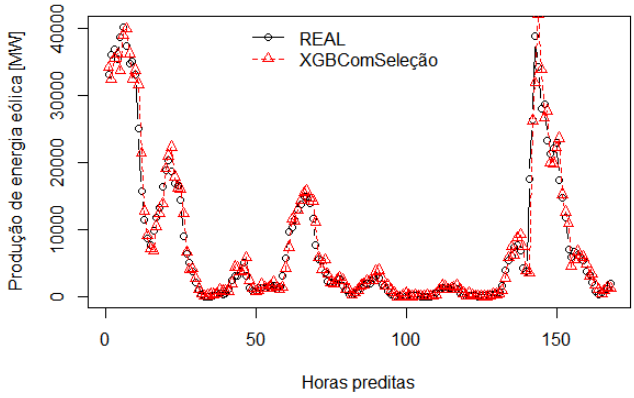
Figura 17. Resultados das predições dos modelos *XGBoost* sem seleção para em março.



Fonte: Autores.

Quanto ao *XGBoost* (Figura 17), o modelo apresentou um resultado próximo aos valores observados mesmo com uma quantidade alta de variáveis independentes.

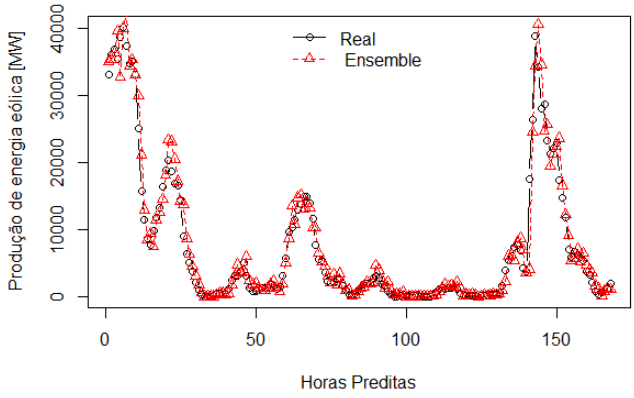
Figura 18. Resultados das predições dos modelos *XGBoost* sem seleção para em março.



Fonte: Autores.

Assim, pode-se entender que os resultados do *XGBoost* não variaram de forma significativa após a seleção de variáveis. O modelo conseguiu desempenhar um bom papel na predição nos dois casos.

Figura 19. Resultado das predições do modelo *ensemble* proposto para em março.



Fonte: Autores.

Quanto ao *ensemble*, nota-se o maior impacto em relação aos modelos individuais: seu desempenho nos momentos de pico. O modelo conseguiu suprir estas deficiências e obteve um bom desempenho nas predições em todo o período analisado.

A Tabela 6, que apresenta os resultados (*MAPE* e *RMSE*) das predições da semana-alvo em março, apresenta o modelo *Random Forest* após a seleção de variáveis como o melhor modelo individual, seguido do *XGBoost* com seleção. Também percebe-se, novamente, que o *Random Forest* apresentou uma melhora significativa com o modelo de seleção. O *ARIMA*(1,1,1) manteve o bom desempenho da primeira estação. Já o *ensemble* obteve um desempenho acima dos modelos individuais apenas no *RMSE*, enquanto em relação ao *MAPE* o desempenho demonstrou-se um tanto inferior aos dos modelos inteligentes, com exceção do *Random Forest* sem aplicação da seleção de variáveis. No entanto, é importante destacar que todos os modelos individuais superestimaram suas predições nos momentos de pico, enquanto o *ensemble* conseguiu suprir essa deficiência.

Tabela 6. Resultados das métricas de avaliaçao nas 168 horas preditas em março.

Classificação	***MAPE* [%]**	***RMSE* [MW]**
Random Forest	22	1911,22
***Random Forest* sem seleção**	538	5244,00
XGBoost	26	1952,00
***XGBoost* sem seleção**	32	2051,00
ARIMA(1,1,1)	43	2764,80
Ensemble XG	33	1497,08

Fonte: Autores.

Análise Geral

Em relação aos resultados dos modelos individuais, pode-se separar a análise em duas etapas: uma análise das métricas e outra dos impactos da seleção de variáveis.

Quanto ao impacto da seleção de variáveis, percebe-se que o modelo *Random Forest* foi o mais beneficiado, visto que o seu modelo com diversas variáveis não conseguiu demonstrar boa performance nas predições (devido ao seu método de aproximação). No entanto, no caso do *XGBoost*, a seleção não teve um impacto significativo. O modelo foi capaz de dar os pesos ideais para as diversas variáveis, mesmo sendo um número excessivo (336 covariáveis).

No que se refere ao *MAPE* e ao *RMSE*, nos modelos individuais, podemos posicionar o *XGBoost* e o *Random Forest* no mesmo patamar, visto que o desempenho destes modelos foram muito similares. O ARIMA, um método mais tradicional, apesar de não ter obtido o melhor desempenho, teve resultado relevante, justificando o fato de ser um método tão usual.

Quanto à combinação dos modelos, o *ensemble* obteve um resultado muito relevante e supriu a principal deficiência dos modelos individuais: a superestimação nos momentos de extremos. Nestes casos, enquanto os modelos individuais apresentavam valores ainda maiores para a estimação, o *ensemble* se aproximava com uma acurácia maior. Este fato direcionou um impacto relevante, principalmente em relação ao *RMSE*, devido ao fato de a equação elaborar o quadrado da diferença, conforme apresentado na Equação 2. Quanto ao *MAPE*, o *ensemble* também apresentou desempenho adequado. Ou seja, entende-se que, apesar de o *ensemble* não ter feito a melhor predição na média, quando errou, errou por menos do que os outros modelos.

5. Considerações Finais

Esta pesquisa teve como objetivo principal propor uma combinação de modelos capaz de elaborar a predição de duas semanas, em escala horária, da produção de energia eólica nos Estados Unidos, utilizando os modelos *Random Forest*, *XGBoost* e *ARIMA*, além de apresentar um modelo de seleção de variáveis. Quanto aos resultados, os modelos individuais apresentaram boa performance, principalmente o *XGBoost* e *Random Forest*, após a seleção de variáveis. O modelo *Random Forest*, utilizando todas as variáveis, foi o único modelo individual que apresentou resultados muito diferentes dos observados, permitindo a conclusão de que o modelo *XGBoost*, para estes casos, teve desempenho superior. Já o *ensemble* proposto atendeu às expectativas, melhorando a acurácia das predições em relação aos modelos individuais, tanto na média quanto nos extremos. Além disso, vale destacar que o modelo apresentou bons resultados para duas estações bem distintas, reforçando que ele possui uma boa capacidade de generalização. Para sugestões de trabalhos futuros, além do histórico da produção, acrescentar outras covaráveis, além de propor novos métodos para o *ensemble*.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Agradeço também ao Laboratório de Inteligência Computacional (LInC) e à UNIFAL pela disponibilização dos recursos computacionais necessários para a execução dos experimentos e elaboração do artigo. Agradeço à Dr. Dalia Patino-Echeverri da Duke University pelo compartilhamento de ideias e ajuda em relação aos métodos aplicados no trabalho.

Referências

ABEE.A.B.E.E. Associação Brasileira de Energia Eólica: Mercado Eólico brasileiro. Recuperado de http://www.abeeolica.org.br /.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Chang, G. W., Lu, H. J., Hsu, L. Y., & Chen, Y. Y. (2016, July). A hybrid model for forecasting wind speed and wind power generation. In 2016 IEEE Power and Energy Society General Meeting (PESGM) (pp. 1-5). IEEE.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Daraeepour, A., & Echeverri, D. P. (2014, February). Day-ahead wind speed prediction by a Neural Network-based model. In ISGT 2014 (pp. 1-5). IEEE.

de Siqueira, T. G., & Salgado, R. M. (2011). An Intelligent Approach for Medium Term Hydropower Scheduling Using Ensemble Model. In Electrical Power Systems and Computers (pp. 353-362). Springer, Berlin, Heidelberg.

Díaz-Uriarte, R., & De Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. BMC bioinformatics, 7(1), 3.

Eldali, F. A., Hansen, T. M., Suryanarayanan, S., & Chong, E. K. (2016, September). Employing ARIMA models to improve wind power forecasts: A case study in ERCOT. In 2016 North American Power Symposium (NAPS) (pp. 1-6). IEEE.

Foley, A. M., Leahy, P. G., Marvuglia, A., & McKeogh, E. J. (2012). Current methods and advances in forecasting of wind power generation. Renewable Energy, 37(1), 1-8.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

GWEC, G. W. E. C. (2018). Global Wind Report: Annual market update 2018. Recuperado de http://gwec. net/.

Hansen, L. K., & Salamon, P. (1990). Neural network ensembles. IEEE transactions on pattern analysis and machine intelligence, 12(10), 993-1001.

Hansen, L. K., & Salamon, P. (1990). Neural network ensembles. IEEE transactions on pattern analysis and machine intelligence, 12(10), 993-1001.

Jurado, S., Peralta, J., Nebot, A., Mugica, F., & Cortez, P. (2013, July). Short-term electric load forecasting using computational intelligence methods. In 2013 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (pp. 1-8). IEEE.

Lei, M., Shiyan, L., Chuanwen, J., Hongling, L., & Yan, Z. (2009). A review on the forecasting of wind speed and generated power. Renewable and Sustainable Energy Reviews, 13(4), 915-920.

Li, S., Wang, P., & Goel, L. (2015). Wind power forecasting using neural network ensembles with feature selection. IEEE Transactions on sustainable energy, 6(4), 1447-1456.

Liu, H., Tian, H. Q., Chen, C., & Li, Y. F. (2010). A hybrid statistical method to predict wind speed and wind power. Renewable energy, 35(8), 1857-1861.

Morettin, P. A., & Toloi, C. M. C. (2006). Análise de Séries Temporais, Segunda Edição.

Ruppert, D. (2011). Statistics and data analysis for financial engineering (Vol. 13). New York: Springer.

Salgado, R. M., Machado, T. C., & Ohishi, T. (2016). Intelligent models to identification and treatment of outliers in electrical load data. IEEE Latin America Transactions, 14(10), 4279-4286.

Salgado, R. M., Pereira, J. J., Ohishi, T., Ballini, R., Lima, C. A. M., & Von Zuben, F. J. (2006, July). A hybrid ensemble model applied to the short-term load forecasting problem. In The 2006 IEEE International Joint Conference on Neural Network Proceedings (pp. 2627-2634). IEEE.

Tubino, D. F. (2000). Administração dos estoques. Manual de planejamento e controle da produção. São Paulo: Atlas, 103-145.

Werner, L., & Ribeiro, J. L. D. (2003). Previsão de demanda: uma aplicação dos modelos Box-Jenkins na área de assistência técnica de computadores pessoais. Gestão & Produção, 10(1), 47-67.

Yang, L., He, M., Zhang, J., & Vittal, V. (2015). Support-vector-machine-enhanced markov model for short-term wind power forecast. IEEE Transactions on Sustainable Energy, 6(3), 791-799.

Porcentagem de contribuição de cada autor no manuscrito

Luís Gustavo Gutierrez Gebin – 50%

Ricardo Menezes Salgado – 33%

Denismar Alves Nogueira – 17%