15º Congresso de Inovação, Ciência e Tecnologia do IFSP - 2024

Modelagem Preditiva do Consumo Energético na Indústria Utilizando Random Forest

VITOR R. MENDES[1], ELCIO R. ARANHA[2]

[1] Graduando em Engenharia de Controle e Automação, Bolsista PIBIFSP, IFSP, Campus Cubatão, vitor.rabello@aluno.ifsp.edu.br.
[2] Docente na área de Automação, IFSP, Campus Cubatão, aranha@ifsp.edu.br.

Área de conhecimento (Tabela CNPq): 3.04.05.03-3 Controle de processos eletrônicos, retroalimentação

RESUMO: A otimização do consumo energético é crucial na indústria siderúrgica devido aos altos custos e emissões associados Modelos de aprendizado de máquina, como o *Random Forest Regressor*, têm se mostrado promissores na previsão e gestão de consumo energético, aproveitando grandes volumes de dados históricos para identificar padrões e propor melhorias. A aplicação desses modelos em ambientes industriais complexos, como a siderurgia, é pouco explorada, por tanto há uma lacuna na compreensão de como essas variáveis interagem e impactam o consumo energético em tempo real, limitando a capacidade de predição precisa e a aplicação de intervenções operacionais imediatas. Este estudo visa preencher essa lacuna desenvolvendo um modelo preditivo com *Random Forest*, utilizando dados operacionais de uma indústria siderúrgica para prever o consumo de energia. O modelo alcançou um coeficiente de determinação (R^2) de 0.92 e um erro médio quadrático (RMSE) de 9.17 kWh, demonstrando alta precisão. Esses resultados indicam que o modelo pode ser uma ferramenta valiosa para reduzir custos e emissões, contribuindo para a sustentabilidade e competitividade da indústria.

PALAVRAS-CHAVE: random forest regressor; eficiência energética; modelagem preditiva.

Predictive Modeling of Energy Consumption in Industry Using Random Forest

ABSTRACT: The optimization of energy consumption is crucial in the steel industry due to the high costs and associated emissions. Machine learning models, such as the Random Forest Regressor, have shown promise in predicting and managing energy consumption by leveraging large volumes of historical data to identify patterns and propose improvements. The application of these models in complex industrial environments, such as steelmaking, is underexplored, creating a gap in understanding how these variables interact and impact real-time energy consumption, thus limiting the ability to make precise predictions and apply immediate operational interventions. This study aims to fill this gap by developing a predictive model using Random Forest, employing operational data from a steel industry to forecast energy consumption. The model achieved a coefficient of determination (R^2) of 0.92 and a root mean square error (RMSE) of 9.17 kWh, demonstrating high accuracy. These results suggest that the model can be a valuable tool for reducing costs and emissions, contributing to the industry's sustainability.

KEYWORDS: random forest regressor; energy efficiency; predictive modeling.

INTRODUÇÃO

A eficiência energética é uma preocupação crescente na indústria, especialmente na siderurgia, devido ao aumento da demanda por recursos energéticos e à necessidade de reduzir custos e impactos ambientais (Tres et al., 2021). A busca por métodos eficientes de monitoramento e controle de processos

industriais tem destacado o uso de dados históricos para aprender padrões de consumo e identificar oportunidades de melhoria na eficiência (Camioto; Rebelatto, 2014).

Nesse contexto, as árvores de decisão surgem como uma abordagem potencialmente eficaz para modelar e prever o desempenho de processos industriais, utilizando dados históricos de eficiência energética (Veloso; Hora, 2019). O uso do algoritmo *Random Forest*, que combina múltiplas árvores de decisão para criar modelos robustos e precisos (BREIMAN, 2001), têm se mostrado promissor para a modelagem preditiva do consumo energético. Ao contrário das árvores de decisão individuais, o Random Forest reduz o *overfitting*, ou seja, é quando o modelo se ajusta excessivamente aos dados de treinamento, capturando ruídos e variações específicas. Essa técnica melhora a capacidade de generalização ao combinar previsões de várias árvores independentes (Nicola, 2021).

Este estudo propõe o uso de *Random Forest* para prever o consumo energético em uma indústria siderúrgica, utilizando dados históricos operacionais fornecidos pela *Korea Electric Power Corporation* (KEPC), obtidos através da plataforma *Kaggle*. A base de dados inclui variáveis como consumo de energia, emissão de CO2 e fatores operacionais, coletadas ao longo de um ano. O objetivo é desenvolver um modelo preditivo capaz de antecipar as necessidades energéticas possibilitando otimizar o consumo.

MATERIAL E MÉTODOS

O desenvolvimento deste estudo foi conduzido utilizando a plataforma *Google Colab*, que proporciona um ambiente baseado em nuvem para execução de códigos em *Python*, facilitando a colaboração e a utilização de bibliotecas especializadas para aprendizado de máquina (NAIK; NAIK; PATIL, 2022). Os dados empregados na pesquisa foram obtidos da plataforma *Kaggle*, fornecidos pela *Korea Electric Power Corporation* (KEPC). O *dataset* compreende registros operacionais de uma indústria siderúrgica ao longo de um ano, incluindo variáveis como consumo energético (kWh) entre outros fatores operacionais.

O tratamento das variáveis foi realizado para adequá-las ao modelo preditivo, visto que inicialmente apresentavam-se em formatos incompatíveis. Por exemplo, as datas estavam em um formato que não permitia a análise direta e precisaram ser convertidas para formatos que permitissem a identificação de padrões temporais. Além disso, os tipos de carga (mínima, média e máxima) foram codificados numericamente como 1, 2 e 3, respectivamente, para facilitar a análise pelo modelo. Esse pré-processamento dos dados foi conduzido no *Excel*, assegurando que os dados estivessem prontos para a análise subsequente e a modelagem preditiva do consumo energético.

Para a implementação e análise dos dados, foram utilizadas as seguintes bibliotecas *Python*: Pandas, para manipulação e análise de dados; *Numpy*, para operações numéricas; *Matplotlib* e *Seaborn*, para visualização de dados; *Scikit-learn*, para implementação do modelo *Random Forest* e avaliação de desempenho; e *Plotly*, para criação de gráficos interativos. No ambiente *Colab*, procedeu-se à exploração dos dados por meio de gráficos de dispersão, buscando entender as relações entre as variáveis independentes (“Semana”, “Mês”, “Dia do ano”, “Dia da Semana”, “Hora”, “Fator de potência adiantado”, “Carga_Valor”) e o consumo de energia (“*Usage_kWh*”).

O *dataset* foi dividido em conjuntos de treino (70%) e teste (30%) utilizando o método *train_test_split* com uma semente aleatória para garantir a reprodutibilidade dos resultados. As variáveis selecionadas por seu potencial impacto no consumo energético e relevância para a otimização dos processos industriais incluem “Semana”, “Mês”, “Dia do ano”, “Dia da Semana”, “Hora”, “Fator de potência adiantado”, “Número de segundos desde a meia-noite” e “Carga_Valor” . A variável alvo foi o consumo energético “*Usage_kWh*”, utilizando o algoritmo *Random Forest Regressor*, para garantir uma alta eficácia do modelo foi feito a otimização dos hiper parâmetros, aplicando uma busca em grade (*Grid Search)*, incluindo o número de estimadores (*n_estimators*), a profundidade máxima das árvores (*max_depth*), e os critérios de divisão dos nós (*min_samples_split* e *min_samples_leaf*). O *Grid Search* é uma técnica de busca onde se avaliam todas as combinações possíveis de um conjunto predefinido de valores de hiper parâmetros. Para cada combinação, o modelo é treinado e validado, permitindo identificar a configuração que proporciona o melhor desempenho (CLAESEN *et al*., 2014). Essa abordagem é fundamental para ajustar o modelo, garantindo que ele opere de forma eficiente e atinja a maior precisão possível.

A performance do modelo foi avaliada por meio do coeficiente de determinação (R^2) e do erro médio quadrático (RMSE). Os resultados obtidos foram visualizados através de gráficos de dispersão

que comparam os valores observados e preditos do consumo energético, além de gráficos específicos para cada variável categórica, o que permitiu uma análise detalhada da precisão do modelo em diferentes cenários operacionais.

RESULTADOS E DISCUSSÃO

Após a aplicação do processo de *Grid Search*, os hiper parâmetros selecionados para o modelo *Random Forest Regressor* foram: profundidade máxima das árvores indefinida (*max_depth*: *None*), tamanho mínimo de amostras na folha igual a 1 (*min_samples_leaf*: 1), divisão mínima de amostras nos nós igual a 5 (*min_samples_split*: 5), e número de estimadores igual a 1000 (*n_estimators*: 1000). Esses valores foram escolhidos para maximizar a capacidade preditiva do modelo, equilibrando a complexidade do modelo com a sua precisão.

O modelo otimizado demonstrou uma alta precisão na predição do consumo energético da indústria siderúrgica. O coeficiente de determinação (R^2) alcançado foi de 0,92, indicando que o modelo foi capaz de explicar 92% da variabilidade nos dados de consumo energético. Além disso, o erro médio quadrático (RMSE) foi de 9,17 kWh, um valor relativamente baixo, o que evidencia a capacidade do modelo de fornecer previsões próximas aos valores reais. Na figura 1, foi representado o gráfico de dispersão comparando os valores reais de consumo energético com os valores preditos pelo modelo, para visualizar a qualidade das predições.

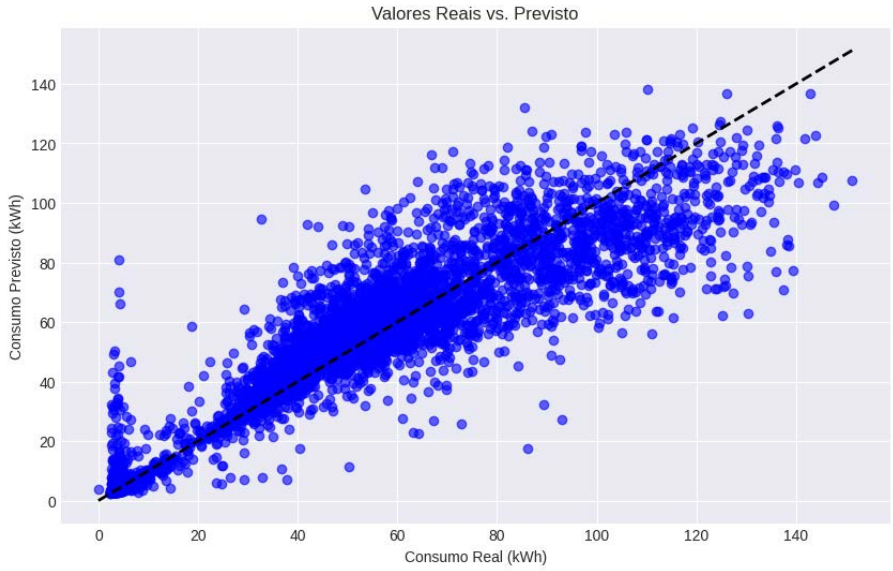

FIGURA 1. Gráfico de dispersão comparando valores reais e valores preditos pelo modelo.

Na Figura 1, a proximidade dos pontos à linha ideal sugere que o modelo preditivo tem uma alta precisão, com uma forte correlação entre os valores reais e previstos. A dispersão dos pontos ao redor da linha, embora presente, é relativamente pequena, o que reforça a capacidade do modelo em capturar os padrões do consumo energético da indústria.

No entanto, é importante considerar as limitações do modelo. Embora o *Grid Search* tenha sido utilizado para otimizar os hiper parâmetros, o modelo ainda depende da qualidade e quantidade dos dados de entrada. Qualquer ruído ou inconsistência nos dados pode impactar negativamente as previsões. Além disso, o modelo pode apresentar dificuldades em capturar variações súbitas no consumo energético, que podem ocorrer devido a fatores externos não contemplados nos dados históricos.

CONCLUSÕES

O presente estudo abordou a aplicação do modelo *Random Forest Regressor* na previsão do consumo energético de uma indústria siderúrgica, utilizando dados operacionais históricos. Os resultados obtidos demonstraram que o modelo foi capaz de prever com alta precisão o consumo de energia, como evidenciado pelo coeficiente de determinação (R^2) de 0.92 e pelo erro médio quadrático

(RMSE) de 9.17 kWh. Esses indicadores confirmam a eficácia do modelo em captar os padrões de consumo e em fornecer previsões eficientes, alinhadas com o objetivo de otimizar o uso energético.

A análise do gráfico de dispersão entre os valores reais e previstos reforçou a capacidade preditiva do modelo, embora algumas discrepâncias apontem para a necessidade de refinamentos adicionais. Esses achados são promissores e indicam que a utilização do *Random Forest* pode contribuir significativamente para a redução de custos e emissões na indústria, atendendo ao propósito de promover a sustentabilidade e a competitividade do setor.

Assim, o modelo desenvolvido não só atende ao objetivo inicial de prever o consumo energético com precisão, mas também se apresenta como uma ferramenta potencial para apoiar a tomada de decisões estratégicas em tempo real, possibilitando intervenções operacionais mais eficazes. Para futuros estudos, sugere-se a integração de técnicas adicionais de pré-processamento de dados e a experimentação com outros algoritmos de aprendizado de máquina, como redes neurais, para avaliar possíveis ganhos em precisão. A inclusão de variáveis externas, como condições climáticas ou variações nos preços da energia, também pode enriquecer o modelo e proporcionar uma predição ainda mais eficaz.

CONTRIBUIÇÕES DOS AUTORES

V.R.M. contribuiu com a curadoria e análise dos dados, desenvolveu o modelo preditivo e foi responsável pela redação do artigo. E.R.A. atuou como orientador, fornecendo orientação metodológica, além de revisar criticamente o manuscrito. Ambos os autores contribuíram para a revisão final do trabalho e aprovaram a versão submetida.

AGRADECIMENTOS

Agradecemos a todos que contribuíram indiretamente a este estudo.

REFERÊNCIAS

BREIMAN, Leo. Random forests. Machine learning, v. 45, p. 5-32, 2001.

CAMIOTO, Flávia de Castro; REBELATTO, Daisy Aparecida do Nascimento. Análise da contribuição ambiental por meio da alteração da matriz energética do setor brasileiro de ferro-gusa e aço. Gestão & Produção, v. 21, p. 732-744, 2014.

CLAESEN, Marc et al. Hyperparameter tuning in python using optunity. In: Proceedings of the international workshop on technical computing for machine learning and mathematical engineering. 2014. p. 3.

NAIK, P.; NAIK, G.; PATIL, M. Conceptualizing Python in Google COLAB. India: Shashwat Publication, 2022.

NICOLA, Márcio José. Adoção de random forest e regressão linear para previsão de falhas em equipamentos agrícolas. 2021. Tese de Doutorado. Universidade de São Paulo.

TRES, N; ZANIN, A.; KRUGER, S. D.; MAGRO, C. B. D. Sustainability practices adopted by industrial companies. Revista de Administração da UFSM, v. 14, n. spe, p. 1140-1159, 2021.

VELLOSO, Higor Medina; HORA, Henrique Rego Monteiro da. Classificação de falhas de um centro de usinagem: um estudo de caso utilizando árvore de decisão. In: SIMPÓSIO DE PESQUISA OPERACIONAL E LOGÍSTICA DA MARINHA, 19., 2019, Rio de Janeiro, RJ. Anais […]. Rio de Janeiro: Centro de Análises de Sistemas Navais, 2019.