Modelagens de *Machine Learning* na predição da produção de H_2 por microorganismos

Machine Learning modeling in predicting H_2 production by microorganisms

Arielly Araujo Luiz[1], Elis Regina Duarte[2], Eduardo Bittencourt Sydney[3], Walter Jose Martinez Burgos[4]

RESUMO

Modelagens de Machine Learning tem sido ferramentas muito utilizadas na atualidade, sendo técnicas poderosas implementadas em diversos processos. A aplicação de *Machine Learning* em bioprocessos gera grande interesse por demonstrar ser um recurso com alta capacidade de auxiliar na solução de problemas e dinamizar processos. O presente trabalho buscou avaliar modelagens com 3 diferentes algoritmos (Arvore de Decisão, Floresta Randômica e Rede Neural Artificial), construídos em linguagem Python com utilização da biblioteca Keras, para predição de biohidrogênio por microorganismos. Foram configuradas um total de 48 modelagens diferentes, variando a porcentagem de dados para teste e validação dos modelos, e o desempenho destes foi avaliado tendo como métrica o Erro Médio Quadrático (MSE) e o Erro Médio Absoluto (MAE) para avaliação de dois modelos particulares. Os resultados obtidos neste trabalho mostraram que a utilização de modelagens de *Machine Learning* é muito promissora na implementação desse tipo de bioprocesso e podem ser aprimoradas com melhor treinamento e estudo de variação de parâmetros dos modelos.

Palavras-chave*: Machine Learning*, Redes Neurais Artificiais, Produção de biohidrogênio.

ABSTRACT

Machine Learning Modeling has been a widely used tool nowadays, being powerful techniques implemented in several processes. The application of Machine Learning in bioprocesses generates great interest as it demonstrates that it is a resource with a high capacity to help solve problems and streamline processes. The present work aimed to evaluate models with 3 different algorithms (Decision Tree, Random Forest and Artificial Neural Network), built in Python language using the Keras library, for prediction of biohydrogen by microorganisms.A total of 48 different models were configured, varying the percentage of data for testing and validating the models, and their performance was evaluated using the Mean Square Error (MSE) and the Mean Absolute Error (MAE) as a metric for evaluating two particular models. The results obtained in this work showed that the use of Machine Learning modeling is very promising in the implementation of this type of processes and can be improved with better training.

Keywords: Machine Learning, Neural Artificial Network, Biohidrogênio production.

[1] Engenharia de Bioprocessos e Biotecnologia, Universidade Tecnológica Federal do Paraná, Ponta Grossa, Paraná, Brasil; arielly@alunos.utfpr.edu.br
[2] Universidade Tecnológica Federal do Paraná, Campus Ponta Grossa; erduarte@utfpr.edu.br
[3] Universidade Tecnológica Federal do Paraná, Ponta Grossa, Paraná, Brasil; eduardosydney@utfpr.edu.br
[4] Universidade Federal do Paraná, Curitiba, Paraná, Brasil; wjmartinez@gmail.com

1 INTRODUÇÃO

A Inteligência artificial (IA), propõe por recursos tecnológicos em que as máquinas realizam decisões por meio de algoritmos para resolver problemas, encontrar soluções e adotar decisões no lugar de pessoas para facilitar o cotidiano das mesmas (DA SILVA, et al., 2019). Essas teorias e o desenvolvimento da IA vem sendo realizado a aproximadamente 70 anos por John McCarthy, Alan Turing, Marvin Minsky e John Atkinson, cientistas computacionais. (GOMES, 2018). Machine Learning (ML) e Deep Learning (DL) (subcampo do ML) têm sido usados ultimamente em diversos processos devido ao grande potencial que possuem de fazer análise de dados, alta capacidade de aprendizado, a fim de chegarem a determinadas conclusões de forma eficiente, não sendo apenas algoritmos baseados somente em instruções específicas.

Modelos de Árvore de Decisão (AD), criado por Ross Quinlan nos anos 90 foi implementado na linguagem JAVA e vem sendo utilizado por apresentar um padrão comportamental nos conjuntos de dados, ou seja, apresenta poucas restrições mostrando-se adequado para os procedimentos que envolvem características qualitativas, além disso, não exige uma distribuição de probabilidade específica (NONG, 2014). As principais vantagens estão relacionadas a capacidade de processar valores em falta ou dados insertos (com ruídos), e ainda, assim, terem resultados de alto desempenho com baixo custo computacional (BHARGAVA, 2013). Uma das maiores dificuldades desse método é a adequação de determinadas bases de dados.

De acordo com Breiman (2001), o algoritmo Floresta Randômica (FR) faz uso da técnica Bagging. Foi desenvolvido com a combinação de múltiplas AD e também pode ser utilizada para as tarefas de classificação ou regressão, ajudando a reduzir a variância das previsões combinando os resultados de várias AD, modeladas em diferentes subamostras do mesmo conjunto de dados (BERNARDO, 2020). Dessa forma podemos perceber que trabalhar com o método FR necessita de uma experiência prévia na utilização de AD.

Modelagens de DL como Redes Neurais Artificiais (RNA) vem se consolidando mundo afora para lidar com problemas mais complexos onde é necessário trabalhar com uma grande quantidade de dados em análises multidisciplinares trabalhando em conjunto com a área estatística e computacional (KOVÁCS, 2020).

As RNAs são construídas em camadas de neurônios artificiais que se ligam entre si. De acordo com a informação processada, um neurônio ativa outro. Isto é, o valor de saída definido por pesos e função de ativação de um neurônio, é a entrada de outro neurônio em outra camada. As RNAs podem ter diferentes arquiteturas de construção, e também serem constituídas de uma ou múltiplas camadas. (FURTADO, 2019; HAYKIN, 2001; FINOCCHIO, 2014). RNAs são modelos computacionais que possuem a capacidade de manter a aquisição e manutenção do conhecimento, desta forma, consegue obter resultados de máquina muito expressivos. Ela é a mais recente técnica e mais complexa utilizada para o desenvolvimento deste trabalho.

O biohidrogênio é o único biocombustível conhecido nos dias de hoje por não se utilizar da queima de carbono e ao reagir com o oxigênio, dessa forma ele libera apenas moléculas de água e calor na sua combustão, permitindo o classificar como uma fonte limpa e renovável de energia (TELES, 2020). A produção de hidrogênio a partir de diferentes microrganismos tem sido reconhecida como uma solução eficiente devido ao fato de ser uma fonte inesgotável e por ter baixo custo de produção (SHOW et al., 2011). Esta produção pode vir a partir de vários substratos, como o glicerol, resíduos de efluente e subprodutos do biodiesel, mas o critério predominante para a escolha do substrato sempre será a produção da molécula de hidrogênio.

De modo geral, as modelagens de ML apresentam ser muito convenientes pela capacidade de aprender e organizar dados, tornando-se ferramentas muito úteis e de apoio a processos industriais e biotecnológicos, sendo que podem resultar em modelagens eficientes para classificação de dados, reconhecimento de padrões,

tomada de decisões, otimização, monitoramento e controle de bioprocessos. (AQUINO, 2016; FLECK et al, 2016; MENDES et al, 2011; STAUDT, 2019).

O presente trabalho tem por objetivo estudar os modelos de ML que se adequem a produção do biocombustível, de forma a serem capazes de prever o volume de biohidrogênio produzido por microorganismos, sendo possível dinamizar processos, escolher e avaliar qual o melhor método a ser empregado em alta escala.

2 METODOLOGIA

Um total de 81 amostras de produção do biohidrogênio por microorganismos foram coletadas a partir de 3 consórcios, cada um tendo 27 amostras, com dados extraídos experimentalmente e variando em quatro parâmetros: temperatura, pH, inóculo e razão entre carbono nitrogênio (C:N) do meio.

A partir das análises, estudou-se a criação de modelagens de ML a partir de 3 algoritmos diferentes (AD, FR e RNA) que comportassem as 4 variáveis para predizer a produção do biohidrogênio, a fim de entender como os modelos se desempenham ao trabalhar com esse tipo de dados. Todos os modelos foram construídos em linguagem Python e executados com diferentes divisões do conjunto de dados em cada execução, em 30%, 25% e 20% dos dados para teste e validação do modelo. O parâmetro utilizado para medir a performance das modelagens foi o erro quadrático médio (MSE) calculado conforme a Eq. (1).

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

Nas primeiras modelagens, buscou-se construir modelos de RNA *Multi Layer Perceptron* (MLP), usando a biblioteca Keras com otimizador ADAM e duas camadas ocultas. Os neurônios receberam função de ativação RELU, e variaram de 5 a 30, em intervalos de 5, em cada uma das camadas, gerando um total de 36 configurações diferentes. Cada um dos modelos foi executado 3 vezes, variando entre as execuções a porcentagem de conjuntos separados para teste e validação.

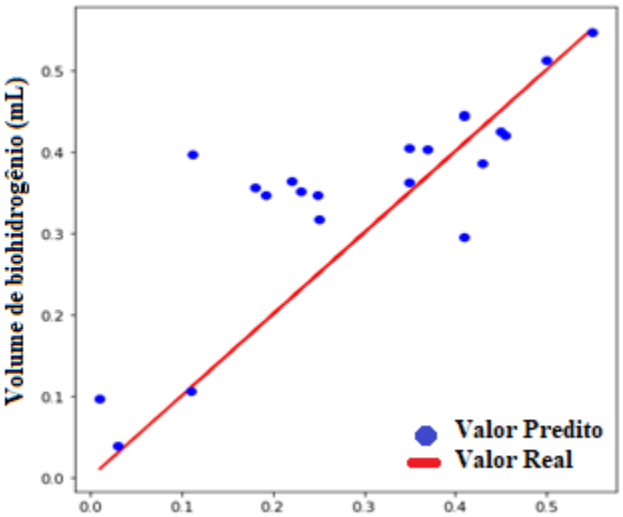
As modelagens seguintes consistiram em algoritmos de AD e FR. Para fins comparativos de desempenho, buscou-se incluir como variável de entrada os consórcios (1, 2 e 3) que pertenciam os dados, resultando em modelos com 4 variáveis (temperatura, pH, inóculo e razão carbono nitrogênio) e 5 variáveis (consórcio, temperatura, pH, inóculo e razão carbono nitrogênio) de entrada, variando a porcentagem de separação de dados conforme estabelecido anteriormente. Dessa maneira, configurou-se 12 modelagens.

3 RESULTADOS

As modelagens foram executadas e o *score* medido através do MSE de cada uma.

Entre as RNAs, a configuração que apresentou o melhor desempenho na predição, representado na Fig.1, em que a linha representa o valor real e os pontos o valor predito, foi o modelo 36, com 25% dos dados separados para teste e validação e 30 neurônios em ambas camadas ocultas, resultando em um MSE de 0.010325965782825125. O modelo foi capaz de prever alguns valores com boa precisão, embora seja possível notar valores dispersos e longe do valor real.

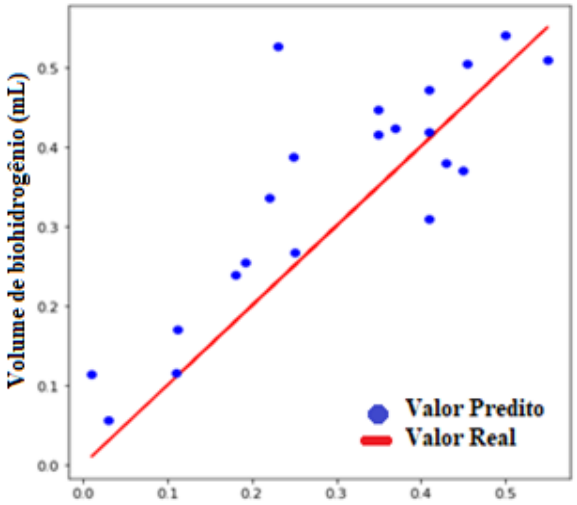
Figura 1 – Predição do modelo 36.



Fonte: Autoria própria (2021).

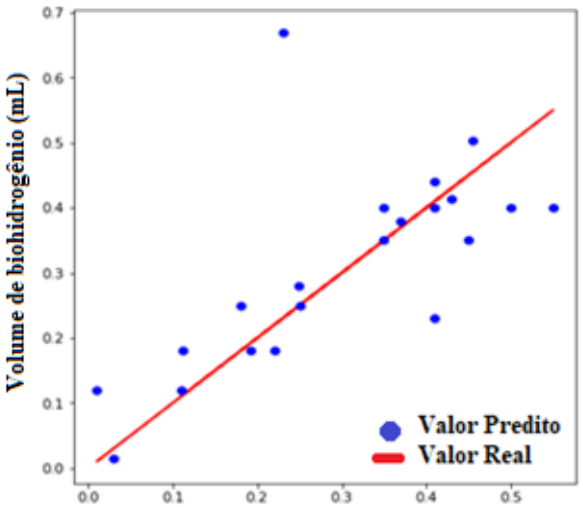
De todos os algoritmos configurados no trabalho, o modelo 47, utilizando algoritmo de FR, com separação de 25% dos dados para teste, obteve o melhor *score* tendo o MSE de 0.0089510659636024 e seu desempenho na predição está representado na Fig. 2. O modelo apresentou um dos melhores índices de predição com todo o conjunto tendo a menor variação do resultado real, mas ainda não sendo a predição ideal. Entretanto, o modelo 41 foi outro a chamar atenção na performance de predição. Este, utilizando algoritmo de AD, também utilizando 25% dos dados para teste, conseguiu predizer com boa precisão e pequena variação a maior quantidade de valores em relação aos outros modelos e seu desempenho na predição está representado na Fig.3. Porém esta modelagem resultou um MSE de 0.014268128095238097, superior ao modelo 47.

Figura 2 – Predição do modelo 47.



Fonte: Autoria própria (2021).

Figura 3 – Predição do modelo 41.



Fonte: Autoria própria (2021).

Nota-se que o MSE do modelo 41 sofreu grande penalidade ao predizer valores muito distantes, mesmo o modelo tendo apresentado boa precisão da predição da maioria valores. A fim de comparar os últimos dois modelos, outra métrica foi interessante ser analisada. O erro absoluto médio (MAE) é menos afetado por *outliers*, além de também ser robusto e preciso, sendo uma boa métrica para analisar e comparar modelos numéricos e sua capacidade de reproduzir a realidade. O MAE pode ser calculado como mostra a Eq. 2.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{2}$$

Como previsto, ao considerar o MAE e diminuindo a sensibilização do *score* por *outliers* muito distantes, o modelo 41 apresentou um MAE de 0.07112, sendo menor em relação ao 47 que resultou em um MAE de 0.07280, indicando ser a melhor modelagem na predição do biohidrogênio configurada neste trabalho.

4 CONCLUSÃO

É possível afirmar que a utilização de modelagens de *Machine Learning* para a predição do biohidrogênio a partir de microorganismos é promissora. As performances das modelagens estudadas neste trabalho têm grande potencial de serem aprimoradas com o aperfeiçoamento dos algoritmos, uma vez que seria possível estudar a variação de parâmetros empregados nos mesmos.

Todavia, a complexidade dos algoritmos não necessariamente poderá chegar aos resultados mais promissores. Os modelos se ajustam de acordo com os tipos de dados fornecidos, dessa forma, neste trabalho um algoritmo mais simples como uma Árvore de Decisão demonstrou expressar melhores resultados que uma Rede Neural Artificial, um algoritmo de maior complexidade que exige um treinamento mais rebuscado e melhor exploração dos parâmetros.

AGRADECIMENTOS

Ao Programa de Bolsas de Iniciação Científica – PIBIC, Fundação Araucária.

REFERÊNCIAS

AQUINO, Pedro Luiz da Mota. **Inteligência computacional aplicada à modelagem e otimização de bioprocessos**. Dissertação (Doutorado em Engenharia Química). 160 p. Departamento de Engenharia Química, Universidade Federal de São Carlos, São Carlos. 2016.

BERNARDO, Felipe et al. **Análise de Algoritmos de Árvores de Decisão e Floresta Randômica**. 2020.

BHARGAVA, Neeraj; BHARGAVA Ritu; SHARMA Girja; MATHURIA Manish. **"Decision Tree Analysis on J48 Algorithm for Data Mining."** 2013.

BREIMAN, Leo. **Random forests. Machine learning**, v. 45, n. 1, p. 5-32, 2001.

DA SILVA, Jennifer Amanda Sobral; MAIRINK, Carlos Henrique Passos. **Inteligência artificial**. Libertas: Revista de Ciências Sociais Aplicadas, v. 9, n. 2, p. 64-85, 2019.

FINOCCHIO, Marco Antonio Ferreira. **Noções de redes neurais artificiais**. Departamento de Engenharia Elétrica, Universidade Tecnológica Federal do Paraná, Cornélio Procópio. 2014.

FLECK, Leandro. et al. **Redes neurais artificiais: Princípios básicos**. Revista Eletrônica Científica Inovação e Tecnologia. Universidade Tecnológica Federal do Paraná. Medianeira, Paraná. v. 1, n. 13, p. 47-57, 2016.

FURTADO, Maria Inês Vasconcellos**. Redes Neurais Artificiais: Uma abordagem para sala de aula**. Editora Atena, 2019.

GOMES, Hermes Oliveira. **Inteligência artificial na saúde pública e privada é possível?**. Revista de Ciências Médicas e Biológicas, v. 17, n. 3, p. 285-286, 2018.

HAYKIN, Simon. **Redes Neurais: Princípios e prática**. 2ª Edição. Editora Bookman, 2001.

KOVÁCS, Zsolt László. **Redes neurais artificiais**. Editora Livraria da Fisica, 2002.

MENDES, Álvaro José Boareto; VALDMAN Belkis; JÚNIOR, Maurício Bezerra de Souza. **Uma revisão de modelagem matemática em bioprocessos. Parte II**: Modelos mecanicistas e redes neuronais artificiais. Revista Militar de Ciência e Tecnologia, Rio de Janeiro, v. 18, p. 40-86, 2011.

NONG, Ye. **Datamining: Theories,algorithms,and examples**, CRC press. 2014.

TELES, Felipe; FLORENTINO, Helenice; CERASUOLO, Marianna. **Um Modelo Matemático para Produção de Biohidrogênio**. Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 7, n. 1, 2020.

STAUDT, Tiago Bauermann. **Predição do volume de biogás produzido em sistemas de biodigestão utilizando redes neurais artificiais.** 120 p. Dissertação (Graduação em Engenharia Elétrica) - Área de Conhecimento de Ciências Exatas e Engenharias, Universidade de Caxias do Sul, Caxias do Sul. 2019.

SHOW, Kuan Yeow et al. **Biohydrogen production: Current perspectives and the way forward. International Journal of Hydrogen Energy**, 37 (20), 15616-15631. 2011.