Predição da solubilidade de CO_2 em líquidos iônicos via aprendizado de máquina e otimização de hiperparâmetros

Henrique de Paula Rocha[a], Vinícius Oswaldo M. T. Dias[a], Luciana Y. Akisawa Silva[a], Fernando V. Lima[b],Wilson H. Hirota[a,*]

[a] *Departamento de Engenharia Química, Universidade Federal de São Paulo, Diadema- SP, Brasil*
[b] *Department of Chemical and Biomedical Engineering, West Virginia University, Morgantown, WV, USA*
[*] *wilson.hirota@unifesp.br*

RESUMO

A crescente necessidade de reduzir as emissões de dióxido de carbono (CO_2) tem impulsionado o desenvolvimento de tecnologias de captura mais eficientes. Nesse contexto, os líquidos iônicos (LIs) destacam-se como potenciais substitutos dos solventes convencionais, devido à possibilidade de ajuste de suas propriedades físico-químicas por meio da combinação de diferentes cátions e ânions. Entretanto, o grande número de combinações possíveis, aliado à ampla faixa de condições operacionais, dificulta a determinação experimental do sistema mais adequado. Neste trabalho, algoritmos de aprendizado de máquina foram empregados para predizer a solubilidade de CO_2 em LIs, com ênfase na otimização de hiperparâmetros utilizando ferramentas da framework Optuna. Modelos baseados em Redes Neurais Artificiais e Florestas Aleatórias foram desenvolvidos e avaliados. Os resultados indicam melhorias consistentes no desempenho preditivo, com redução de erros e aumento do coeficiente de determinação, evidenciando o potencial da abordagem proposta.

Palavras-chave: Captura de CO_2; Líquidos Iônicos; Aprendizado de Máquina; Hiperparâmetros.

1 Introdução

A captura de CO_2, consiste em um conjunto de tecnologias destinadas a diminuir as emissões de dióxido de carbono, podendo ser aplicada tanto em correntes de gás de combustão quanto diretamente no ar atmosférico. Dentre os principais métodos de captura de CO_2, destacam-se os processos de pré-combustão, pós-combustão e oxi-combustão, que englobam processos como absorção química e física, adsorção e membranas. Grande parte dos processos de captura de CO_2 envolve predominantemente o uso de solventes, como a monoetanolamina, na absorção química, ou o Selexol (éteres dimetílicos de polietilenoglicol), na absorção física. Entretanto, tais compostos apresentam limitações, incluindo toxicidade, degradação química, alta volatilidade e elevada demanda energética para a regeneração (Mondal et al., 2012).

Diante dessas limitações, diversos estudos têm investigado a substituição dos solventes tradicionais por solventes alternativos que apresentem menor volatilidade, maior estabilidade térmica, menor corrosividade, menor taxa de degradação e menor custo energético associado à regeneração. Neste contexto, os Líquidos Iônicos (LIs) destacam-se como candidatos promissores para a aplicação em processos de captura de CO_2 (Aghaie et al., 2018).

Os LIs são sais compostos por cátions orgânicos e ânions orgânicos ou inorgânicos que, por definição, apresentam pontos de fusão inferiores a 100°C. Esses compostos caracterizam-se pela pressão de vapor desprezível, mesmo em altas temperaturas, elevada estabilidade térmica e baixa inflamabilidade (Patel e Lee, 2012). Estruturalmente, a combinação de cátions volumosos e assimétricos com diferentes ânions confere a esses solventes ampla variabilidade de propriedades físico-químicas. Em razão dessa versatilidade, os LIs são frequentemente descritos na literatura como "solventes sob medida"(Baran e Kloskowski, 2025).

Um dos principais desafios no uso dos LIs é a escolha adequada de combinações cátion-ânion para otimizar o processo em diferentes condições operacionais, uma vez que o número de possibilidades é extremamente elevado. A avaliação experimental de todas as combinações torna-se inviável, tanto pelo alto custo quanto pelo tempo demandado. Nesse contexto, destacam-se duas abordagens principais para a predição da solubilidade de CO2 em LIs: os modelos termodinâmicos e os algoritmos de aprendizado de máquina.

Embora os modelos termodinâmicos apresentem vantagens como consistência e interpretabilidade, sua aplicação pode envolver alta complexidade, principalmente em sistemas não ideais e altas pressões, o que pode limitar

a sua capacidade preditiva em alguns cenários. Já os métodos de aprendizado de máquina podem capturar relações não lineares complexas, embora apresentem limitações como menor interpretabilidade e forte dependência da qualidade e quantidade de dados disponíveis para treinamento. (Tatar et al., 2016).

Para a aplicação dos métodos de aprendizado de máquina, além da escolha do algoritmo, é necessária a definição adequada de hiperparâmetros de cada um. Em Redes Neurais Artificiais (ANN), por exemplo, é necessário determinar o número de camadas, o número de neurônios por camada, e a função de ativação. Já em modelos de Florestas Aleatórias é necessário determinar a profundidade máxima e o número de árvores. A escolha desses hiperparâmetros pode se tornar um empecilho, pois a combinação de diferentes hiperparâmetros influencia diretamente os resultados obtidos, podendo melhorar ou piorar o desempenho. Além disso, há o risco de sobreajuste caso o modelo exceda a complexidade necessária (Samartini, 2023).

Visto que não há regras para determinar esses valores, uma alternativa é o uso de técnicas de otimização de hiperparâmetros. Dentre os métodos disponíveis, existem algumas abordagens diferentes para encontrar os parâmetros, mas a ideia base envolve testar diferentes dados até encontrar a melhor combinação. Por exemplo, há métodos que testam todas as combinações possíveis, enquanto outros exploram conjuntos aleatórios. (Akiba et al., 2019) Portanto, visando a eficiência, uma escolha adequada é a framework Optuna, pois seu sistema de busca envolve testes aleatórios que criam um caminho pelos hiperparâmetros mais promissores, evitando iterações desnecessárias sem negligenciar regiões promissoras do espaço de busca (Shao et al., 2024).

Neste sentido, este trabalho tem como objetivo utilizar ferramentas de otimização de hiperparâmetros para melhorar o desempenho de algoritmos de aprendizado de máquina para prever a solubilidade de CO2 em líquidos iônicos. Além disso, para avaliar os resultados e comparar com a literatura, métricas estatísticas como o Erro Médio Absoluto (MAE), Erro Quadrático Médio (MSE) e o Coeficiente de Determinação (R^2) são aplicadas.

2 Metodologia

Nesse trabalho, toda a implementação computacional foi realizada no ambiente Jupyter Notebook, por meio da distribuição Anaconda e a linguagem Python. Para o desenvolvimento dos modelos de aprendizado de máquina, foram usadas as bibliotecas Scikit-learn e TensorFlow. Adicionalmente, utilizou-se o Pandas para a manipulação e tratamento dos dados, e a biblioteca Optuna para a otimização dos hiperparâmetros.

A base de dados utilizada neste estudo foi disponibilizada por Song et al., 2020 como material suplementar e também utilizada por Dias, 2025. O conjunto de dados é composto por 10.116 registros experimentais, incluindo as variáveis temperatura, pressão, estrutura dos líquidos iônicos e a fração molar de CO_2 dissolvido. A representação estrutural dos líquidos iônicos foi realizada por meio da decomposição dos cátions e ânions em grupos constituintes. Essa abordagem considerou grupos funcionais, núcleos de cátions e estruturas completas de ânions, resultando em um total de 51 descritores utilizados como variáveis de entrada nos modelos. Posteriormente, os registros contendo valores nulos foram removidos, e os dados foram padronizados utilizando o método StandardScaler da biblioteca Scikit-learn.

Tanto os dois trabalhos citados quanto este artigo tiveram abordagens semelhantes na separação dos dados. Os algoritmos usados por Song et al., 2020 e por Dias, 2025 usaram 80% dos dados para o treinamento e 20% para a fase de testes. Já nesse trabalho, a mesma divisão inicial de 80% (8093 dados) para treino e 20% (2023 dados) para teste foi feita, porém houve uma nova divisão nos dados de treinamento: 20% dos 8093 dados foram separados para validação do Optuna. Ou seja, comparando com o total de dados: 64% para treinamento, 20% para teste e 16% para validação. Outra consideração são as variáveis de entrada (*features*) e de saída (*targets*) utilizadas, as quais foram as mesmas nos três trabalhos As *features* estão apresentadas na Tabela 1 e a variável *target* é a fração molar de CO_2 dissolvida no líquido iônico.

Tabela 1: Variáveis de entrada (*features*)

Variável	Faixa	Descrição
T	243,2−453,15	Temperatura (K)
p	0,00798−499,9	Pressão (bar)
Grupos Funcionais	Grupo 1 - Grupo 51	Frequência de ocorrência no LI

A otimização dos hiperparâmetros foi conduzida com o auxílio da biblioteca Optuna. Para cada algoritmo, foi definida uma função objetivo responsável por quantificar o erro entre os valores preditos pelo modelo e os valores experimentais, sendo adotado o erro quadrático médio (MSE-*Mean Squared Error*) como métrica de desempenho. Os hiperparâmetros foram definidos como variáveis de busca e explorados por meio do método *suggest* do Optuna, dentro de intervalos previamente estabelecidos. Os intervalos podem ser classificados de acordo com o tipo de variável desejada, entre eles: categórica e inteiro. A principal diferença entre esses dois tipos é que o inteiro permite o uso de números inteiros em um intervalo com um espaçamento pré-selecionado, e a categórica limita o uso de itens diretamente listados, podendo ser números, palavras (strings), entre outros.

O processo de otimização consistiu na avaliação de diferentes combinações de hiperparâmetros, com o objetivo de minimizar o valor da função objetivo. Ao final, selecionou-se o conjunto de parâmetros que apresentou o menor erro. Com o objetivo de diminuir o número de tentativas na aplicação da ferramenta, uma estratégia para encontrar o número de neurônios no algoritmo ANN foi fazer uma busca com intervalo e espaçamento maiores (busca geral) e afunilar com intervalo e espaçamento menores (busca detalhada). Assim, os intervalos de busca considerados no processo de otimização são apresentados na Tabela 2, para as Redes Neurais, e na Tabela 3 para a Floresta Aleatória.

Tabela 2: Intervalo de testes do Optuna para Redes Neurais Artificiais (ANN)

Hiperparâmetros	Tipo	Espaçamento	Intervalo
Função de ativação	Categórica	-	[swish, tanh, sigmoid, ReLU, linear]
Número de Camadas	Inteiro	1	[1, 9]
Número de neurônios (geral)	Inteiro	10	[1, 251]
Número de neurônios (detalhada)	Inteiro	5	[50, 250]

Tabela 3: Intervalo de testes do Optuna para Floresta Aleatória (RF)

Hiperparâmetros	Tipo	Espaçamento	Intervalo
Profundidade máxima	Categórica	–	[None, 10, 15, 25, 30, 25, 40, 45, 50]
Número de árvores	Inteiro	25	[100, 700]

Após a etapa de otimização, os modelos foram treinados utilizando os hiperparâmetros selecionados, e seu desempenho foi avaliado por meio de métricas estatísticas, como o coeficiente de determinação (R^2), descrito na Equação (1), e o erro absoluto médio (MAE), calculado conforme a Equação (2).

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

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

3 Resultados e discussões

A aplicação da biblioteca Optuna permitiu a determinação de configurações otimizadas de hiperparâmetros para ambos os algoritmos estudados. No caso das Redes Neurais Artificiais, a Tabela 4 apresenta a comparação entre a configuração obtida neste trabalho e aquelas reportadas por Song et al., 2020 e Dias, 2025. Com base nessas configurações, a Tabela 5 apresenta o desempenho preditivo dos modelos, possibilitando uma análise comparativa do impacto da otimização de hiperparâmetros nos resultados obtidos.

É observada uma grande diferença entre os hiperparâmetros usados em cada algoritmo, o que mostra como a tarefa de encontrar uma configuração adequada sem um método sistemático pode gerar diferentes caminhos.

Tabela 4: Hiperparâmetros usados nas Redes Neurais Artificiais (ANN)

	Função de Ativação	Número de camadas	Número de neurônios
Song et al., 2020	tansig e purelin	1	7
Dias, 2025	tanh	9	256; 128; 64; 32; 16; 8; 4; 2; 1
Presente estudo	swish	4	115; 85; 125;150

Tabela 5: Métricas estatísticas do conjunto teste para Rede Neural Artificial (ANN)

	Presente estudo	Song et al., 2020	Dias, 2025
MAE	0,0188	0,0202	0,0194
R^2	0,9857	0,9836	0,9838

Apesar da diferença, todos os modelos analisados apresentaram resultados satisfatórios. Além disso, o modelo com Optuna trouxe uma leve melhora tanto no MAE quanto no R^2, e seu desempenho está exposto na Figura 1 e na Figura 2.

Figura 1: Desempenho do modelo ANN: valores previstos vs. Experimentais

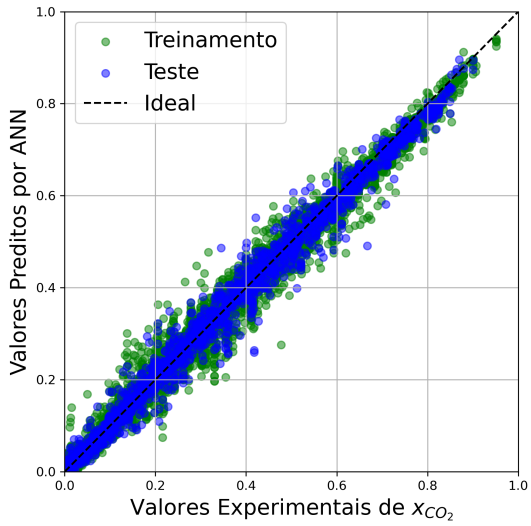
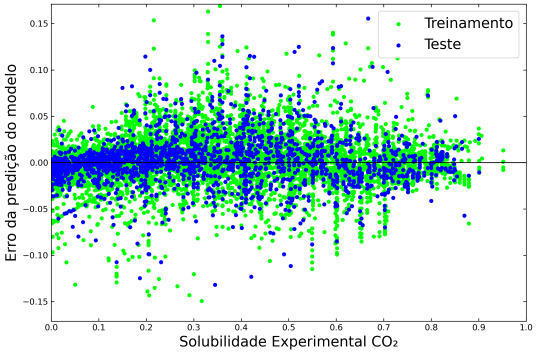


Figura 2: Erro do modelo ANN



Como o trabalho de Song et al., 2020 não utiliza o algoritmo da Floresta Aleatória, as comparações serão feitas apenas com o trabalho de Dias, 2025. Os valores utilizados nos artigos estão representados na Tabela 6 e a comparação dos resultados está presente na Tabela 7.

Tabela 6: Hiperparâmetros usados nas Florestas Aleatórias (RF)

	Número de árvores	Profundidade máxima
Dias, 2025	100	None
Presente estudo	150	25

Tabela 7: Métricas estatísticas do conjunto teste para Floresta Aleatória (RF)

	Presente estudo	Dias, 2025
MAE	0,0219	0,0220
R^2	0,9776	0,9775

Nesse caso, os hiperparâmetros encontrados foram semelhantes aos usados na literatura comparada. Porém, os erros obtidos no modelo com Optuna são ligeiramente menores quando comparados ao trabalho de Dias, 2025, o que mostra como a ferramenta de otimização pode trazer consistência e buscar melhora na procura dos hiperparâmetros. O desempenho do modelo otimizado está apresentado na Figura 3 e na Figura 4.

Figura 3: Desempenho do modelo RF: valores previstos vs. Experimentais

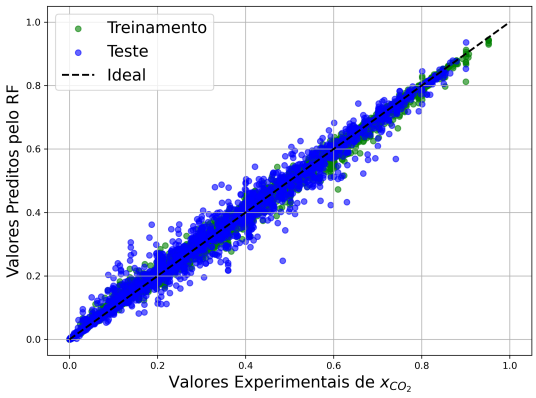
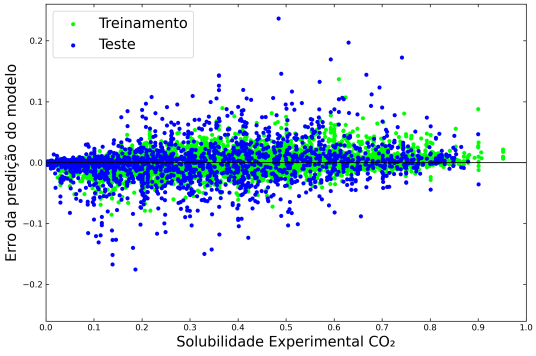


Figura 4: Erro do modelo RF



Com a análise gráfica, é possível observar que tanto no algoritmo de ANN quanto no RF os modelos gerados apresentam valores consistentes para os dados de treinamento e para os dados de teste. Ou seja, outra vantagem do uso de uma ferramenta de otimização é evitar causar sobreajuste nos modelos pelo excesso de complexidade do algoritmo construído.

Outro ponto positivo observado no uso do Optuna em relação ao método de tentativa e erro é a menor demanda de tempo para encontrar os hiperparâmetros ótimos, pois no tempo gasto procurando manualmente é possível executar diversos testes com a ferramenta de otimização.

4 Conclusão

Neste trabalho, investigou–se a aplicação de técnicas de otimização de hiperparâmetros no aprimoramento de modelos de aprendizado de máquina voltados à predição da solubilidade de CO2 em líquidos iônicos. A utilização da framework Optuna possibilitou a exploração eficiente do espaço de busca dos hiperparâmetros, resultando em configurações que proporcionaram melhorias consistentes no desempenho preditivo dos modelos avaliados.

Os resultados obtidos, quando comparados com aqueles reportados na literatura, indicam reduções nas métricas de erros e melhorias no coeficiente de determinação, evidenciando o potencial da abordagem proposta. Além disso, a automatização do processo de ajuste dos hiperparâmetros mostrou-se vantajosa em termos de eficiência computacional e reprodutibilidade, reduzindo a dependência de procedimentos empíricos (tentativa e erro). Embora os ganhos sejam moderados, eles reforçam a importância da etapa de otimização no desenvolvimento de modelos preditivos robustos e confiáveis.

Referências

Aghaie, M., Rezaei, N., & Zendehboudi, S. (2018). A systematic review on CO_2 capture with ionic liquids: Current status and future prospects. *Renewable and Sustainable Energy Reviews*, *96*, 502–525.

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631.

Baran, K., & Kloskowski, A. (2025). Unlocking Structural Insights into CO_2 Absorption with Ionic Liquids: A Comparative Study of Machine Learning, Association Rules, and Meta-Learning for Modeling Henry's Law Constant. *ACS Sustainable Chemistry & Engineering*, *13*(31), 12805–12817.

Dias, V. O. M. T. D. (2025). *Predição da solubilidade de CO_2 em líquidos iônicos usando algoritmos de aprendizado de máquina* [Trabalho de Conclusão de Curso (Graduação em Engenharia)]. Universidade Federal de São Paulo.

Mondal, M. K., Balsora, H. K., & Varshney, P. (2012). Progress and trends in CO_2 capture/separation technologies: A review. *Energy*, *46*(1), 431–441.

Patel, D. D., & Lee, J.-M. (2012). Applications of ionic liquids. *The Chemical Record*, *12*(3), 329–355.

Samartini, N. L., André; Barth. (2023). *Técnicas de Machine Learning* (A. L. Sicsú, Ed.; 1a). Blucher.

Shao, H., Liu, X., Zong, D., & Song, Q. (2024). Optimization of diabetes prediction methods based on combinatorial balancing algorithm. *Nutrition & Diabetes*, *14*(1).

Song, Z., Shi, H., Zhang, X., & Zhou, T. (2020). Prediction of CO_2 solubility in ionic liquids using machine learning methods. *Chemical Engineering Science*, *223*, 115752.

Tatar, A., Naseri, S., Bahadori, M., Hezave, A. Z., Kashiwao, T., Bahadori, A., & Darvish, H. (2016). Prediction of carbon dioxide solubility in ionic liquids using MLP and radial basis function (RBF) neural networks. *Journal of the Taiwan Institute of Chemical Engineers*, *60*, 151–164.