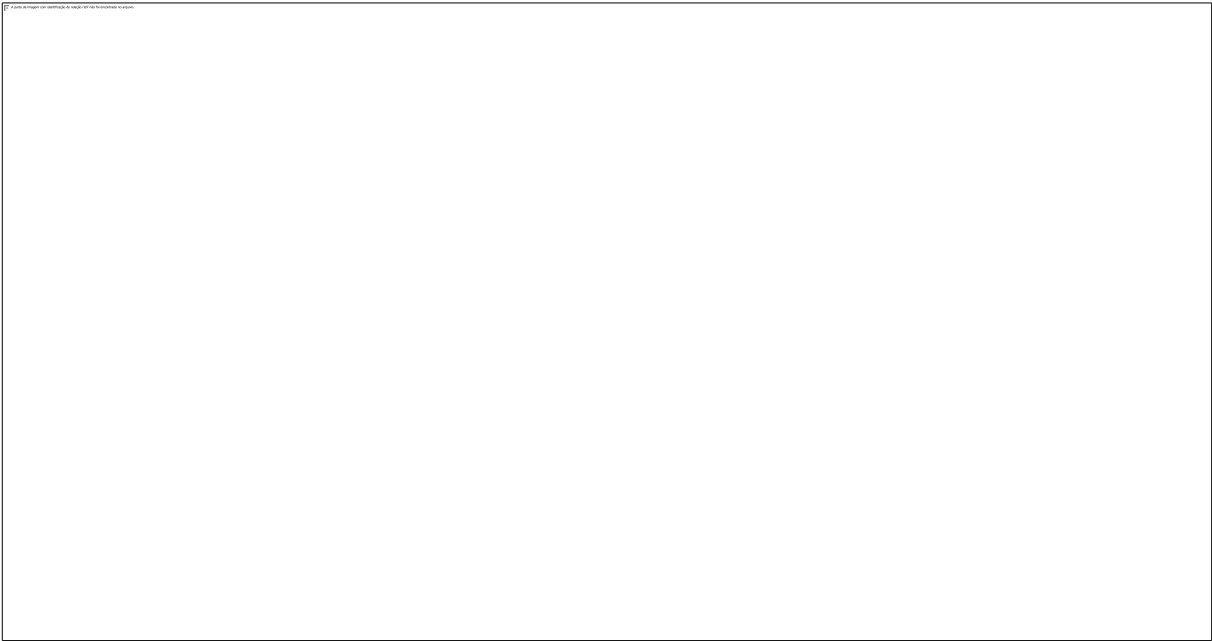


12º CONGRESSO BRASILEIRO DE PESQUISA E DESENVOLVIMENTO EM PETRÓLEO E GÁS



TÍTULO DO TRABALHO:

Modelo estimador da densidade de misturas eletrólíticas: Aplicação de *Machine Learning* e comparação com modelos da literatura.

AUTORES:

Detogni JLH, Monteiro MF, O-Chiavone Filho

INSTITUIÇÃO:

Universidade Federal do Rio Grande do Norte (UFRN)

MODELO ESTIMADOR DA DENSIDADE DE MISTURAS ELETRÓLITICAS: APLICAÇÃO DE *MACHINE LEARNING* E COMPARAÇÃO COM MODELO DA LITERATURA

Resumo

Misturas aquosas contendo eletrólitos fortes possuem vastas aplicações industriais. A densidade é uma propriedade chave para o cálculo do equilíbrio de fases, determinação do regime de escoamento e compreensão das interações estruturais. A modelagem das propriedades físico-químicas dessas misturas é complexa devido as forças eletrostáticas fortes e intensa solvatação dos íons. Portanto, é essencial modelar a densidade dessas misturas em amplas faixas de concentração, temperatura e tipos de eletrólitos. *Machine Learning* é um ramo da simulação computacional capaz de criar algoritmos que aprendem com dados e apresentam alta performance e capacidade de generalizar resultados. Diversas aplicações da Engenharia Química fazem uso desse tipo de modelagem, com destaque para modelos baseados em conjuntos de Árvores de Decisões. O objetivo do trabalho foi treinar um modelo de *Machine Learning* utilizando dados experimentais compilados da literatura para prever a densidade de misturas do tipo água + eletrólito a 1 atm em $\text{kg}\cdot\text{m}^{-3}$ para 72 espécies distintas em uma faixa de temperatura de 283.15 K a 373.15 K e concentração molal de 0.0 a 62 $\text{mol}\cdot\text{kg}^{-1}$. Ademais, o algoritmo treinado foi confrontado com dois modelos semiempíricos de referência. O modelo proposto obteve um desvio médio relativo absoluto inferior a 0.5 % e coeficiente de determinação superior a 0.99 apresentando performance semelhante as referências.

Palavras-chave: densidade, mistura, eletrólito, Machine Learning

Abstract

Aqueous mixtures containing strong electrolytes have wide industrial applications. Mass density is a key property for calculating phase equilibrium, determining the flow regime and understanding structural interactions. Modeling the physicochemical properties of these mixtures is complex due to the strong electrostatic forces and intense solvation of the ions. Therefore, it is essential to model the density of these mixtures over wide ranges of concentration, temperature, and electrolyte types. Machine Learning is a branch of computer simulation capable of creating algorithms that learn from data and present high performance and the ability to generalize results. Several Chemical Engineering applications make use of this type of modeling, with emphasis on models based on sets of Decision Trees. The objective of the work was to train a Machine Learning model using experimental data compiled from the literature to predict the density of water + electrolyte mixtures at 1 atm in $\text{kg}\cdot\text{m}^{-3}$ for 72 different species in a temperature range from 283.15 K to 373.15 K and molal concentration from 0.0 to 62 $\text{mol}\cdot\text{kg}^{-1}$. Furthermore, the trained algorithm was compared with two semi-empirical reference models. The proposed model obtained an absolute relative average deviation of less than 0.5% and a correlation coefficient greater than 0.99, presenting similar performance to the references.

Keywords: density, mixture, electrolyte, Machine Learning

Introdução

Misturas aquosas com eletrólitos fortes são sistemas de interesse em diversos processos industriais.¹ Esses sistemas, constituídos de eletrólitos dissolvidos em água liberando íons, alteram o comportamento volumétrico da solução, causam incrustações, corrosões e influenciam no equilíbrio de fases. A complexidade em modelar as propriedades físico-químicas dessas misturas, como a densidade, viscosidade, capacidade calorífica e outras, reside nas interações de longo e curto alcance provocadas pelas cargas iônicas no solvente.²

A densidade é uma valiosa propriedade que influencia no regime de escoamento, coeficiente de atividade, permite estimar a mudança volumétrica e auxilia na compreensão de interações estruturais do tipo íon-íon, solvente-solvente e íon-solvente.³ Portanto, a modelagem da densidade em diversas faixas de temperatura, concentração e eletrólitos tornou-se uma necessidade científica.⁴ Essa propriedade é influenciada pela temperatura da solução, concentração do soluto e natureza do eletrólito dissolvido.

Machine Learning (ML) é um ramo da computação que desenvolve algoritmos capazes de aprender com dados. Um modelo de ML atua como uma função que relaciona variáveis independentes X com uma variável resposta Y , as regras de associação são obtidas ao captar padrões em um conjunto de dados para treinamento com exemplos de $Y(X)$ em diversas condições. O modelo é validado ao confrontá-lo com um conjunto separado de validação/teste contendo dados não vistos durante o treinamento. ML tem sido amplamente empregado para resolver problemas de engenharia devido a sua alta performance e capacidade de estimar situações não vistas durante o treinamento.⁵

Árvore de Decisão é uma classe de modelo de ML que se ajusta a dados identificando padrões de dependência e proporção entre as variáveis e a resposta. A construção desse algoritmo envolve um processo recursivo de criação de nós até atingir um critério de parada. Um nó é uma regra de comparação para a variável $x \in X$ que divide um subconjunto de Y em dois outros de menor impureza. A impureza reflete a heterogeneidade do conjunto em relação à média e pode ser mensurada pela variância.⁶ Ao fim do processo, diversos subconjuntos de Y com baixa impureza denominados folhas podem ser acessados por caminhos específicos de nós e correspondem as previsões do modelo.⁷ Os critérios de parada são limitações para o desenvolvimento do algoritmo, como por exemplo o número máximo de nós, o número mínimo de amostras para consolidar um nó ou o decréscimo mínimo da Impureza em um nó. As configurações que influenciam no desenvolvimento do modelo são denominadas Hiper Parâmetros.

Modelos de ML podem ser combinados para formar um novo preditor de performance superior, nesse caso, o modelo resultante é um algoritmo de conjunto (*ensemble*).⁸ A técnica de *Gradient Boosting* (Gboost) utiliza um conjunto de Árvores de Decisões em sequência onde cada modelo do grupo é ajustado aos desvios produzidos pelo agente anterior, dessa forma, a previsão do Gboost é igual a estimativa da primeira Árvore mais o somatório das previsões dos desvios de cada modelo subsequente.⁹

Apresentada a necessidade de modelar as propriedades físico-químicas de misturas aquosas contendo eletrólitos fortes e destacado a capacidade promissora de implementação de técnicas de ML para construir algoritmos capazes de aprender com dados e extrapolar resultados para condições não estudadas, o objetivo do trabalho será desenvolver um modelo baseado em Árvores de decisões utilizando a técnica de conjunto de *Gradiente Boosting* para estimar a densidade de soluções contendo eletrólitos utilizando uma base de dados experimentais compilada da literatura. Além disso, o modelo proposto será confrontado com referências da literatura.

Metodologia

Um modelo do tipo Gboost será treinado para prever a densidade, ρ , em $\text{kg}\cdot\text{m}^{-3}$ de uma mistura água + eletrólito forte a 1 atm de pressão em função das variáveis apresentadas na Tabela 1. A separação entre cátion e ânion aumenta a generalidade do modelo pois permite que eletrólitos não vistos durante o treinamento possam ser estimados desde que os íons que o constituem sejam previamente conhecidos.

Tabela 1: Variáveis utilizadas pelo modelo Gboost para calcular a densidade de uma mistura aquosa contendo eletrólito forte.

VARIÁVEL	SIMBOLOGIA	UNIDADE
TEMPERATURA DA SOLUÇÃO	T	K
POTENCIAL IÔNICO DO CÁTION	I_{cat}	$\text{mol}\cdot\text{kg}^{-1}$
POTENCIAL IÔNICO DO ÂNION	I_{ani}	$\text{mol}\cdot\text{kg}^{-1}$
MASSA MOLECULAR DO CÁTION	M_{cat}	g/mol
MASSA MOLECULAR DO ÂNION	M_{ani}	g/mol
COEFICIENTE ESTEQUIOMÉTRICO DO CÁTION	n_{cat}	-
COEFICIENTE ESTEQUIOMÉTRICO DO ÂNION	n_{ani}	-

O potencial iônico, I , foi utilizado ao invés da concentração molal, b , dos íons. A propriedade citada expressa a concentração dos íons juntamente com o efeito de suas cargas, q , na atividade iônica do meio e é calculada por meio da Equação 1.¹⁰

$$I = \frac{b_{\text{ion}} \cdot q_{\text{ion}}^2}{2} \quad (1)$$

Uma base de dados experimentais foi compilada da literatura para o desenvolvimento do modelo. A base contém 17657 pontos de densidade para mistura aquosa contendo eletrólito a 1 atm de pressão em uma faixa de temperatura de 273.15 K até 373.15 K e concentração molal dos solutos variando de 0.0 até 64 $\text{mol}\cdot\text{kg}^{-1}$. A Figura 1 apresenta a distribuição de temperatura e concentração da base de dados experimentais. É possível observar a aglomeração em condições de diluição infinita e temperatura ambiente associados a densidades próximas a da água pura.

Foram estudados 72 eletrólitos distintos listados em ordem decrescente de disponibilidade de dados experimentais: NaCl, KCl, LiCl, NaBr, KBr, LiBr, NaOH, HNO₃, KOH, MgCl₂, NH₄Cl, CaCl₂, H₂SO₄, Na₂SO₄, MgSO₄, NaF, HCl, KI, CsCl, NaI, NaNO₃, ZnCl₂, CuSO₄, K₂SO₄, KF, ZnSO₄, NiCl₂, KNO₃, H₃PO₄, MnCl₂, CoCl₂, LiI, K₂CO₃, (NH₄)₂SO₄, NH₄NO₃, RbCl, Na₂CO₃, Li₂SO₄, CsI, BaCl₂, CuCl₂, MnSO₄, NiSO₄, SrCl₂, FeCl₂, NaHSO₃, CdCl₂, NaHCO₃, NaClO₃, FeSO₄, Al₂(SO₄)₃, Fe₂(SO₄)₃, LiF, Na₂S₂O₃, CsBr, FeCl₃, CdSO₄, Na₂SO₃, RbI, NaH₂PO₄, CoSO₄, Na₂MoO₄, RbBr, AlCl₃, RbF, Cr₂(SO₄)₃, Na₃PO₄, Na₂HPO₄, CsF, NaNO₂, CrCl₃. Os 3 primeiros eletrólitos apresentam abundância de dados na literatura somando juntos 30 % da base de dados, essa mesma porcentagem é atingida juntando os dados disponíveis para os últimos 55 solutos, isto é, todos seguidos ao NaF.

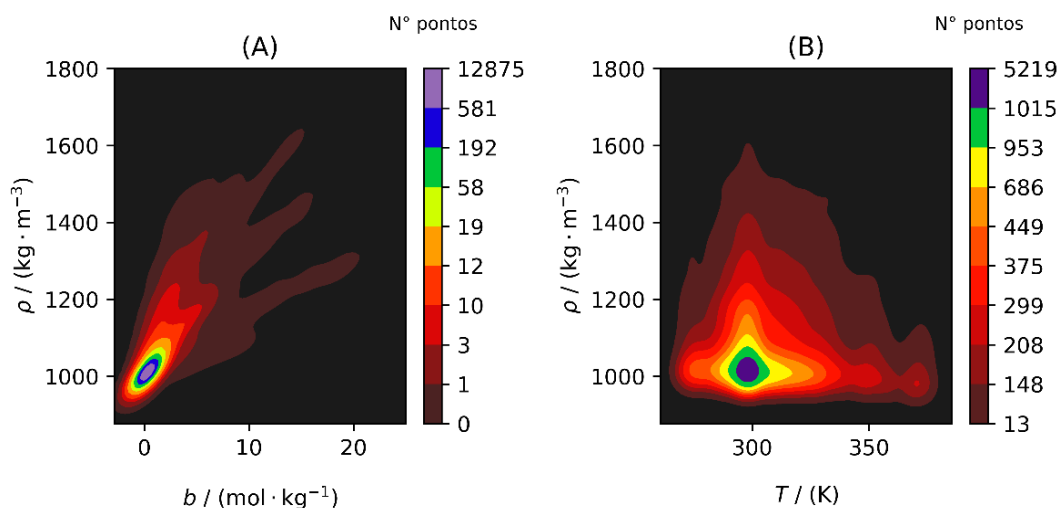


Figura 1: Distribuição de densidade da base de dados experimentais em função da molalidade (A) e temperatura (B).

A construção do modelo Gboost utilizou a linguagem de programação *Python* juntamente com o pacote Pandas manipulação de dados e *Scikit-Learn* para ajuste de modelos de ML.¹¹⁻¹² Foram destinados 20 % dos dados para o conjunto de validação utilizando uma estratificação por eletrólito para garantir que todos os solutos estejam disponíveis para o treinamento e teste. Às variáveis foram reescaladas para valores entre 0 e 1 no intuito de reduzir a disparidade de ordem de grandeza.

O treinamento consistiu em uma varredura de Hiper Parâmetros do algoritmo de Gboost (Tabela 2) para treinar diversos preditores com configurações distintas. Cada modelo realizou o cálculo da base de dados de validação e uma métrica de avaliação mensurou suas performances, para esse trabalho o coeficiente de determinação, R^2 . O melhor Gboost obtido foi confrontado com dois modelos da Literatura. A Equação semiempírica de Laliberté-Cooper (LC) consegue calcular a densidade de 59 eletrólitos distintos utilizando a temperatura da solução e a fração mássica do soluto juntamente com 5 parâmetros.¹³ O software industrial OLI Studio é um modelo semiempírico com pacote especializado para eletrólitos em mistura de solventes.¹⁴

Tabela 2: Hiper Parâmetros do algoritmo Gboost. A função para o cálculo dos resíduos pode assumir o desvio médio quadrático (MSE) ou absoluto (MAE) e o desvio médio quadrático de Friedman está disponível para o cálculo da Impureza.

HIPER PARÂMETRO	DESCRIÇÃO	VALORES
F_{PERDA}	Função para calcular os resíduos das estimativas	MSE ou MAE
Ω	Velocidade de aprendizado	De 0 até 0.8
N	Número máximo de estimadores consecutivos	De 100 até 3000
I	Função para o cálculo da impureza	MSE ou Friedman
N_{MIN} (NÓ)	Número min. de amostras para iniciar nó	De 2 até 100
N_{MIN} (FOLHA)	Número min. de amostras para consolidar folha	De 2 até 100
D_{MAX}	Profundidade máxima de cada estimador	De 100 até 1500
ΔI_{MIN}	Variação min. Da impureza em cada nó	De 0.01 até 0.95

Resultados e Discussão

O modelo de ML da classe *Gradient Boosting* (Gboost) foi obtido de uma varredura de Hiper Parâmetros contendo 10.000 exemplares. A Figura 2 apresenta um gráfico de coordenadas paralelas com as configurações dos 200 melhores modelos e desempenhos refletidos na última coluna. A função para o cálculo da Impureza foi igual ao Erro médio quadrático de Friedman.¹⁵

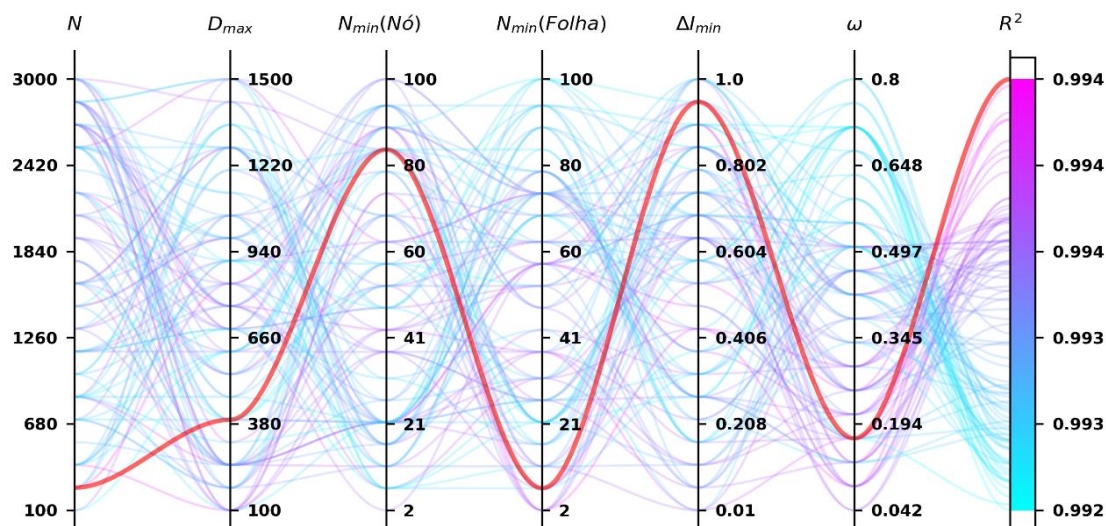


Figura 2: Coordenadas paralelas das configurações de Hiper Parâmetros para 200 modelos da classe Gboost e R^2 ao calcular o conjunto de validação. O melhor modelo está destacado em vermelho.

A Figura 3-A apresenta os desvios obtidos com um gradiente de cores indicando o desvio relativo. Foi constatado que mais de 97 % das densidades calculadas obtiveram desvio relativo igual ou inferior a 2 % em módulo. O Gboost obteve desvio relativo absoluto médio igual a 0.35 % ($4.002 \text{ kg} \cdot \text{m}^{-3}$) e desvio absoluto máximo igual a $226.06 \text{ kg} \cdot \text{m}^{-3}$. A Figura 3-B apresenta a correlação linear entre as densidades calculadas e experimentais, o valor do coeficiente de determinação foi igual a 0.9944, indicando forte correlação.

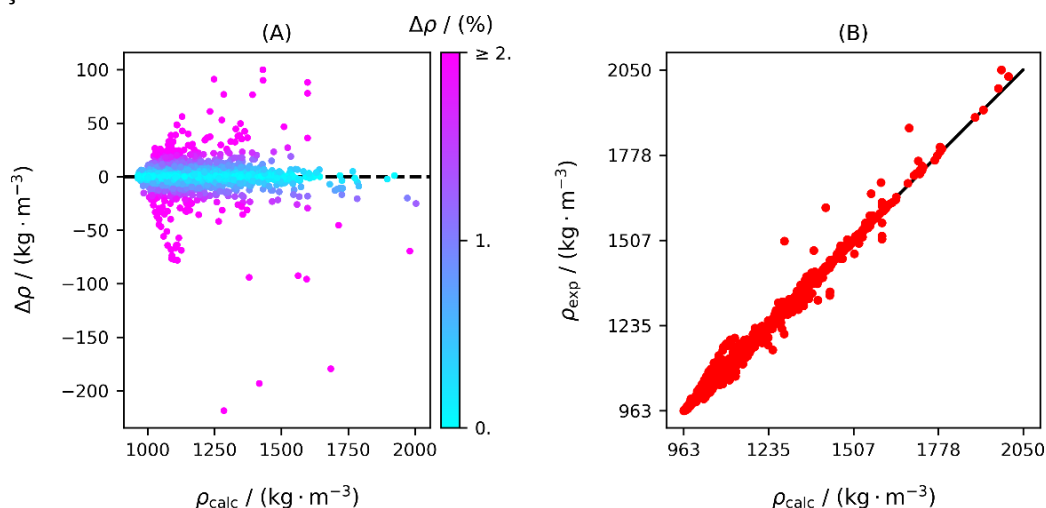


Figura 3: Desvios e gradiente de cores indicando desvio relativo (A) e correlação linear (B) do melhor modelo Gboost ao calcular a base de dados de validação.

A Figura 4-A apresenta um gráfico de radar para comparar os desvios médios absolutos e a Figura 4-B os desvios médios relativos absolutos obtidos entre os modelos para diversos eletrólitos. O desempenho do modelo Gboost foi similar ao de ambas as referências, com o OLI fornecendo os maiores desvios. A Figura 5 apresenta as curvas de desvio relativo absoluto acumulado para os 3 modelos. A equação de LC obteve a melhor performance (converge mais rapidamente). O Gboost possui desempenho superior ao software OLI, que acumula o total de pontos no desvio absoluto relativo acumulado igual a 102 %.

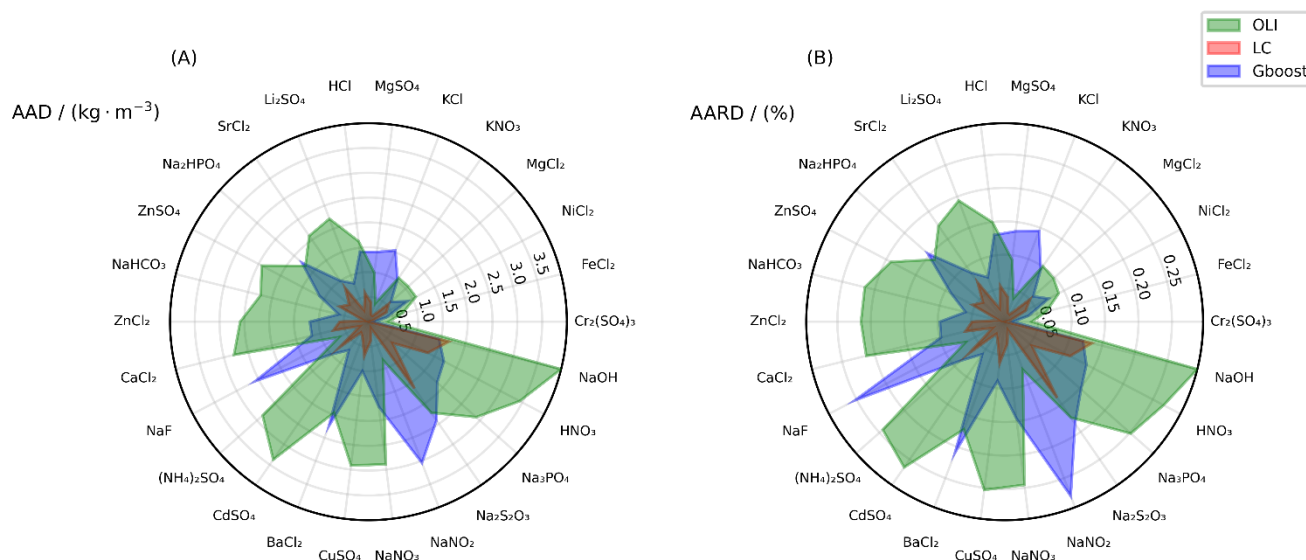


Figura 4: Desvios absolutos (A) e relativos (B) de cada modelo ao prever a densidade de alguns eletrólitos.

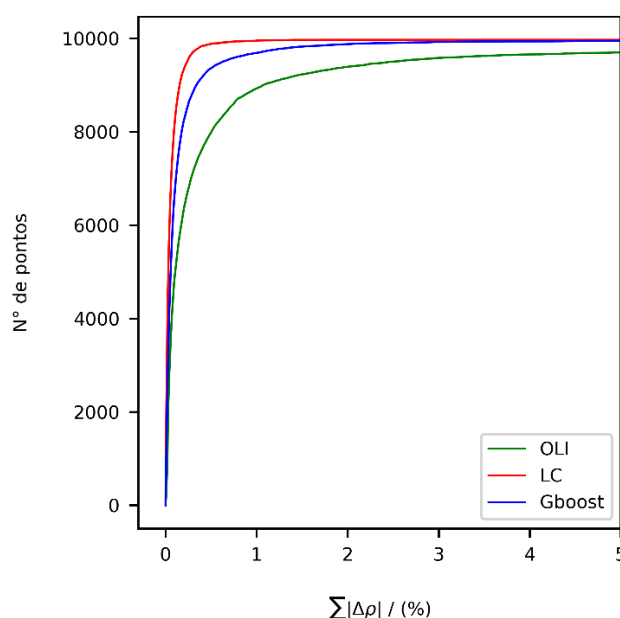


Figura 5: Curva de desvio relativo absoluto acumulado para os modelos OLI, LC e Gboost. As curvas convergem para a quantidade total de pontos calculados.

Conclusões

Um modelo de ML baseado em Árvores de Decisões com uma otimização de conjunto *Gradient Boosting* (Gboost) foi treinado para calcular a densidade de uma solução aquosa com eletrólito a 1 atm em função da temperatura, concentração e características dos íons. O algoritmo obteve desvios inferiores a 2 % em módulo para um volume próximo ao total de pontos da base de dados, além disso, um desvio relativo absoluto médio inferior a 0.5 % e forte correlação linear foram observados. O modelo Gboost foi comparado com duas referências da literatura e obteve desempenho similar a ambos.

Agradecimentos

Ao apoio financeiro do Programa de Recursos Humanos 44.1 da Agencia Nacional de Petróleo, Gás Natural e Biocombustíveis.

Referências Bibliográficas

- [1] MOURA NETO, Mário Hermes de et al. Density and Electrical Conductivity for Aqueous Mixtures of Monoethylene Glycol and Sodium chloride: Experimental Data and data-driven Modeling for Composition Determination. *Journal of Chemical & Engineering Data*, v. 66, p. 1914–1928, 2021.
- [2] BUCKINGHAM, A. D. A theory of ion-solvent interaction. *Discuss. Faraday Soc.*, v. 24, p. 151-157, 1957.
- [3] MILLERO, Frank J. Molal Volumes of Electrolytes. *Chemical Reviews*, v. 71, p. 147-176, 1971.
- [4] KONTOGEORGIS, Georgios M. et al. Conclusions from Round Table Discussion during IUT of ESAT 2021 electrolyte thermodynamics challenges - From industrial needs to academic research. *Fluid Phase Equilibria*, v. 556, p. 113399, 2022.
- [5] FANG, Jian et al. Development of Machine Learning Algorithms for Predicting Internal Corrosion of Crude Oil and Natural Gas Pipelines. *Computers & Chemical Engineering*, v. 177, p. 108358-108358, 2023.
- [6] ZHANG, Guangyi; GIONIS, Aristides. Regularized Impurity reduction: Accurate Decision Trees with Complexity Guarantees. *Data Mining and Knowledge Discovery*, v. 37, p. 434-475, 2022.
- [7] MYLES, Anthony J. et al. An introduction to decision tree modeling. *Journal of Chemometrics*, v. 18, n. 6, p. 275-285, 2004.
- [8] DIETTERICH, Thomas G. Ensemble Methods in Machine Learning. In: *Multiple Classifier Systems*. [S. l.]: Springer, 2000. v. 1857, p. 1-15.
- [9] BENTÉJAC, Candice; CSÖRGŐ, Anna; MARTÍNEZ-MUÑOZ, Gonzalo. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, v. 54, n. 3, p. 1937-1967, 2021.
- [10] SOLOMON, Theodros. The Definition and Unit of Ionic Strength. *Journal of Chemical Education*, v. 78, p. 1691, 2001.
- [11] THE PANDAS DEVELOPMENT TEAM. *pandas-dev/pandas: Pandas*. Zenodo, v. 2.2.2, abr. 2024.
- [12] PEDREGOSA, Fabian et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.
- [13] LALIBERTÉ, Marc; COOPER, W. Edward. Model for Calculating the Density of Aqueous Electrolyte Solutions. *Journal of Chemical & Engineering Data*, v. 49, p. 1141-1151, 2004.
- [14] WANG, Peiming; ANDERKO, Andrzej; YOUNG, Robert D. A speciation-based model for mixed-solvent electrolyte systems. *Fluid Phase Equilibria*, v. 203, n. 1, p. 141-176, 2002.
- [15] FRIEDMAN, Jerome H. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, v. 29, p. 1189–1232, 2001.