MACHINE LEARNING PARA A PREDIÇÃO DE PROPRIEDADES DE POLIESTIRENO OBTIDO VIA POLIMERIZAÇÃO ARGET-ATPR

Palavras-Chave: polimerização, reação química, aprendizado de máquina.

Autores(as):

Kauê Scaranari Alcantara, UNICAMP – FEQ

Profª. Drª. Liliane Maria Ferrareso Lona (orientadora), UNICAMP – FEQ

Prof. Dr. Nicolas Spogis (coorientador), UNICAMP – FEQ

INTRODUÇÃO

A produção de dados de forma massiva pela indústria torna seu armazenamento, manipulação e tratamento custoso, provocando uma situação desafiadora no sentido de que informações valiosas podem estar sendo descartadas cotidianamente. Portanto, é estabelecido um desafio: avaliar de forma adequada e racional o destino de toda essa informação, de maneira que seja utilizada adequadamente e possa beneficiar a engenharia, a indústria e consequentemente a população de forma geral por meio de soluções e conhecimentos desenvolvidos [1].

Técnicas de aprendizado de máquina (*Machine Learning*) e formas de lidar com dados têm sido exploradas nos últimos 30 anos, mas só recentemente o poder computacional alcançou meios de tornar acessível esse tipo de pesquisa com base em grandes quantidades de dados [2]. Dados, quando isolados de conhecimento, não representam informação útil e, por isso, faz-se necessário o seu tratamento para um subsequente uso coerente. Dessa forma, procedimentos que compreendem a manipulação de equações paramétricas, atualizações de pesos e ajustes de parâmetros/hiperparâmetros têm evoluído e ganhado espaço tanto no mercado quanto na academia como meio de observar padrões e encontrar respostas antes ocultas [3][4].

O anseio pelo desenvolvimento de uma estrutura matemática com comportamento inspirado no cérebro humano resultou no surgimento das redes neurais artificiais (RNAs), que são baseadas em neurônios biológicos. Essas redes recebem dados de entrada e retornam valores de saída, ao passo que passam pelo crivo de uma função matemática conhecida como perda (ou *loss function*), baseando-se na minimização de erros e otimização de parâmetros [2].

No contexto de integração entre tecnologia e indústria, surgem os materiais poliméricos como produtos expoentes no ramo da engenharia, estando presentes em novos utensílios, novas ferramentas, diversas aplicações industriais e recentes alternativas sustentáveis. Assim, a não linearidade das reações de polimerização favorece o uso de RNAs, como é o caso da reação de interesse ARGET-ATRP (*Activators Regenerated by Electron Transfer Atom Transfer Radical Polymerization*) ou Polimerização Radicalar por Transferência de Átomo com Regeneração de Ativadores por Transferência de Elétrons [5]. Portanto, de posse de dados obtidos experimentalmente, torna-se possível obter predições de propriedades do polímero poliestireno e da conversão de monômero (estireno) em um reator do tipo batelada. Ainda, é possível obter as condições operacionais de concentração de reagente, agente redutor e catalisador empregados como condição inicial para o processo de reação química em conjunto com variáveis de operação, como o tempo de reação, a concentração de reagente e a massa molar do monômero [5].

METODOLOGIA

O projeto visa aplicar conceitos e técnicas de Aprendizado de Máquina (*Machine Learning* ou ML) à conhecimentos de engenharia de processos químicos por meio de um estudo de caso. Para tanto, utiliza-se a linguagem de programação Python™ para a criação de um aplicativo com base no *framework* de código aberto Dash, pertencente à empresa Plotly, que

permite o desenvolvimento de aplicativos *web* para realizar visualização, análise e tratamento de dados. O Dash tem como característica agrupar e abstrair ferramentas da linguagem de marcação conhecida como HTML, além do CSS, que é uma linguagem de estilo, permitindo com que o desenvolvedor trabalhe com códigos Python™, focando na aplicação e na análise de dados. Com isso, busca-se estudar as condições operacionais de um reator do tipo batelada utilizado em um processo de reação de polimerização de poliestireno a partir de monômero de estireno, agente redutor e catalisador.

Primeiramente, resolve-se um sistema de dez equações diferenciais ordinárias por meio da função *solve_ivp* presente na biblioteca SciPy. Para isso, é utilizado o método de Runge-Kutta implícito denominado como Radau, em que são empregadas, como condições iniciais, relações de concentração de componentes presentes no reator para prosseguimento do processo de polimerização (concentração de iniciador, concentração de catalisador e concentração de agente redutor). Ainda são utilizadas condições padrão de processo, tais como tempo de reação (40 h), concentração inicial de monômero (5,82 mol/L) e o peso molecular do estireno (104,15 g/mol).

O conjunto de dados estudado foi obtido experimentalmente por Banin, Vieira e Lona (2021), que utilizaram RNAs a para predição de propriedades médias da polimerização ARGET-ATRP a partir do estireno. Como o conjunto original compreendia 50 valores disponíveis para cada variável, sendo três de entrada e três de saída, utiliza-se uma técnica conhecida como amostragem por hipercubo latino (*Latin Hypercube Sampling* ou LHS) para superar essa limitação [6]. Para isso, aplica-se o conceito de planejamento de simulações conhecido por DOE (*Design of Experiments)*, definindo-se variáveis de entrada, intervalos limitantes para cada variável, tipos de variáveis (contínua/discreta) e a quantidade de simulações. O sistema de EDOs que descreve a reação de polimerização é resolvido para cada conjunto de condições iniciais criado pelo método LHS, sendo valores resultantes referentes à conversão de monômero, massa molar e índice de polidispersividade de polímero [6]. Dessa forma, expande-se o conjunto de dados disponíveis de 50 para 5000 valores atribuídos a cada variável, limitando-se entre dados para treinamento e teste das redes neurais artificiais. O objetivo de uso da técnica de LHS compreende obter uma amostragem uniformemente distribuída de forma que haja a menor correlação possível entre as variáveis de entrada [6].

De posse dos dados obtidos como resultado do processo de simulação, renderizam-se gráficos e tabelas com a finalidade de realizar uma análise exploratória de dados. Para evidenciar a baixa correlação entre as variáveis de entrada, empregam-se quatro matrizes que permitem avaliar a dependência entre pares de variáveis, que podem ser classificados como diretamente proporcionais, inversamente proporcionais ou ausentes de proporcionalidade. Para tanto, são aplicadas as correlações de Pearson, Spearman, Kendall e ϕ-k, cada uma com características de interesse. Apesar da presença de quatro métodos distintos e complexos, o enfoque é dado aos dois primeiros: Pearson e Spearman, que permitem avaliar a dependência linear e a monotonicidade da relação entre as variáveis, respectivamente [7]. A monotonicidade diz respeito à característica de uma função em não ter seu valor diminuído, se crescente, nem aumentado, se decrescente ao longo de todo o domínio estudado.

Sequencialmente, produz-se um gráfico de coordenadas paralelas, que apresenta as variáveis em eixos paralelos verticais lado a lado, possibilitando avaliar a relação entre diversas variáveis simultaneamente por meio de suas conexões. As linhas presentes nesse tipo de gráfico passam uma única vez por cada um dos eixos, sendo que uma única linha que passa por todas as variáveis representa uma única simulação, havendo, portanto, 5000 linhas distintas. Ainda, pode-se empregar filtros interativos que permitem a seleção de intervalos de valores específicos para cada variável, tornando possível analisar distintos cenários e relações entre os dados.

O conjunto de dados obtido é conhecido como *dataset* e demanda um tratamento adequado, compreendendo-se a limpeza, a normalização/padronização (se necessário) e a separação entre valores destinados para treinamento e teste das RNAs. Assim, é possível garantir a qualidade dos dados utilizados de modo a evitar problemas como *overfitting*, por exemplo, que é a situação em que um modelo é sobreajustado aos dados de treinamento, tendo seu uso limitado ao modelo utilizado, de forma que a generalização não possa ser realizada. Ainda, emprega-se a RNAs diretas (RNADs) e inversas (RNAIs) para modelagem do processo de polimerização estudado, onde o desempenho das RNAs é avaliado através do erro quadrático médio (*mean squared error* ou MSE) como apresentado pela Equação (1).

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

O projeto utiliza RNAs do tipo multicamada (*Multi-layer Perceptron* ou MLP), que têm como característica apresentarem múltiplas camadas ocultas em sua estrutura [5]. A quantidade de neurônios presentes em cada camada pode variar, assim como o número de camadas ocultas. Nesse sentido, o modelo recebe três variáveis distintas classificadas como preditoras (X), retornando, ao fim, três variáveis alvo ou resultado (Y), sendo X e Y vetores. Avalia-se, a princípio, a RNAD, tendo como variáveis de entrada as relações entre concentração $[P_0X]/[C]$, $[P_0X]/[M]$ e $[C]/[A]$, enquanto as propriedades intrínsecas ao

polímero como o índice de polidispersividade (Đ), o peso molecular médio ($\overline{M}_n$) e a conversão de monômero X são os resultados. Para o caso de RNAI, invertem-se as variáveis de entrada e de saída. É calculado o valor do coeficiente de correlação (R^2) para cada valor de saída obtido.

Ressalte-se que o aplicativo desenvolvido permite a seleção das variáveis de entrada e saída, podendo-se alterar entre o uso das RNADs e RNAIs. Nesse momento, é necessário observar se de fato as RNAs estão interpolando dados, e não extrapolando, sendo que o treinamento deve ser realizado com os valores no interior do intervalo previamente definido. O algoritmo desenvolvido permite realizar a otimização e predição utilizando distintas redes, sendo um processo consideravelmente rápido e baseado na procura de hiperparâmetros. Esses critérios são: a quantidade ideal de neurônios em cada camada, a quantidade de camadas, a função de ativação, a taxa de aprendizagem, o inicializador de pesos, o otimizador, a função perda, o número de amostras processadas (*batch size*) e o número de épocas (*epochs*). Portanto, escolha adequada de cada uma dessas variáveis de interesse é feita por meio da atualização de pesos da função matemática utilizada, visando minimização de erros através do processo conhecido por retropropagação ou *backpropagation* [8].

RESULTADOS E DISCUSSÃO

A Fig. 1 apresenta a interface do aplicativo desenvolvido, que é separado em 13 abas, cada qual com sua funcionalidade. A primeira é responsável por permitir a inserção das condições iniciais e de processo, resultando nos valores finais encontrados para conversão de monômero, polidispersividade e massa molecular do polímero por meio de uma única solução das EDOs. Em seguida, a segunda aba é destinada ao DOE, onde são definidas as características do conjunto de dados para realizar as simulações. A fim de resolver o sistema de dez equações diferenciais, dadas as condições iniciais obtidas para cada uma das 5000 simulações, utiliza-se a terceira aba. A quarta aba é utilizada caso o usuário possua previamente um conjunto experimental de dados com quantidade adequada para uso das RNAs. Ainda, a quinta e a sexta abam são responsáveis pela análise exploratória de dados, onde é produzido o gráfico de coordenadas paralelas presente na Fig. 2., que por meio da seleção de intervalos em seu interior, permite a filtragem de valores de interesse para as variáveis presentes nos eixos paralelos, como na Fig. 3. Esses valores escolhidos são posteriormente utilizados para o treinamento e validação das RNADs e RNAIs, porém. O filtro possibilita avaliar em que local do intervalo para cada variável existe o interesse de trabalhar sobre os dados.

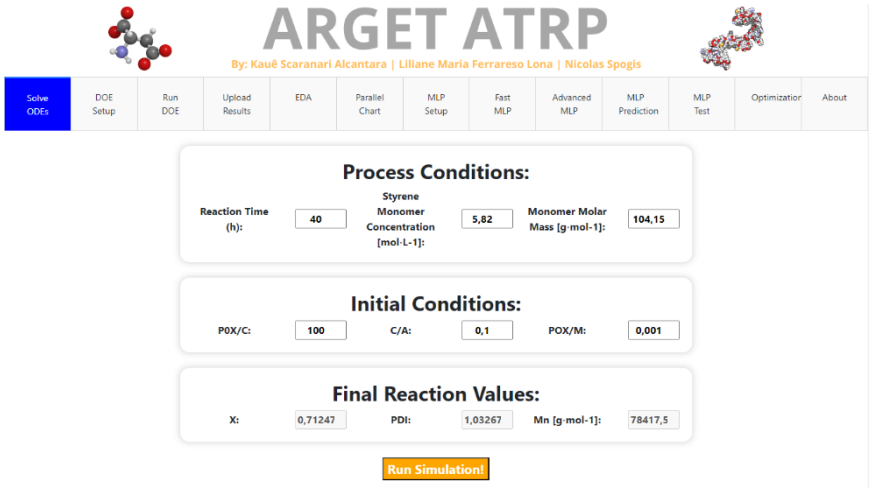

Fig. 1. Interface do aplicativo ARGET-ATRP desenvolvido em Python. Fonte: autoria própria.

Para o proceder correto das análises no aplicativo, deve-se primeiro abrir o gráfico de coordenadas paralelas a partir, que é produzido a partir dos resultados das simulações e, em seguida, rodar a rede neural "*Fast* MLP", dado que os dados utilizados para treinamento e validação são provenientes do filtro definido. Posteriormente, é necessário se atentar aos dados produzidos pela rede neural, porque não é de interesse que ela extrapole dados. Efetivamente, as RNAs são interpoladoras de dados, não sendo esperado que ela retorne resultados fora da faixa previamente estabelecida.

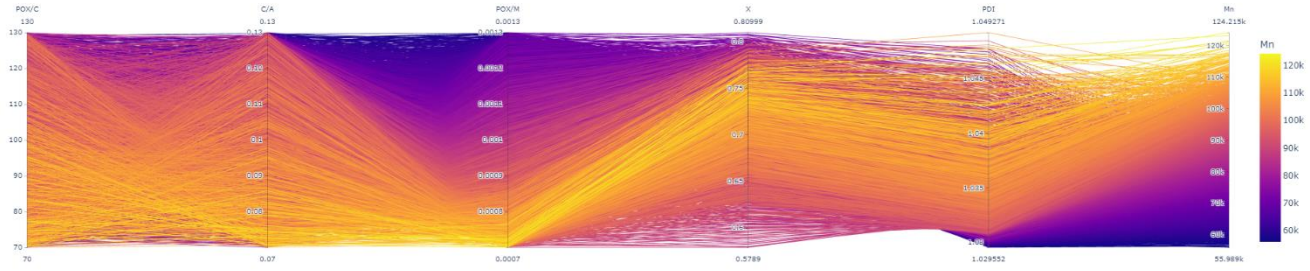

Fig. 2. Gráfico de coordenadas paralelas. Fonte: autoria própria.

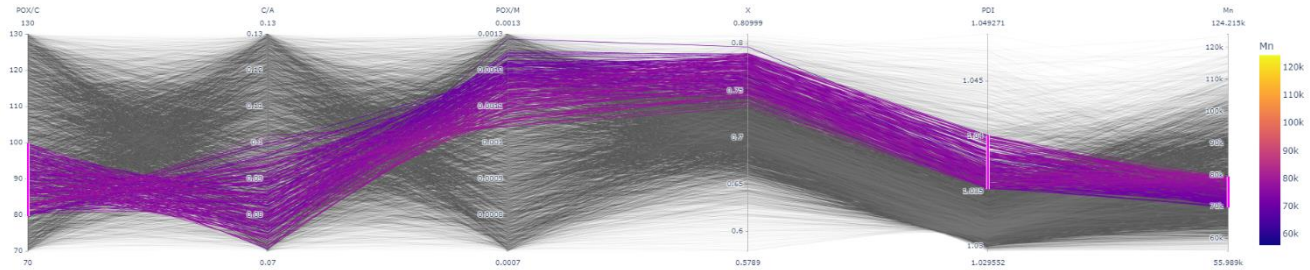

Fig. 3. Gráfico de coordenadas paralelas com filtro aplicado nas variáveis. Fonte: autoria própria.

Em relação às redes neurais, a sétima aba ("MLP *Setup*") permite a seleção das variáveis utilizadas para o caso direto ou indireto. Em seguida, na oitava aba ("*Run* MLP"), pode-se executar a rede neural rápida (*Fast MLP*), que possibilita a regressão a partir do conjunto de dados em poucos segundos. Para tanto, utiliza-se a técnica de busca em grade ou *Grid Search*, capaz de selecionar a melhor combinação de hiperparâmetros por meio de testes sequenciais. Com isso, produz-se para o caso direto o conjunto de gráficos presentes na Fig. 4, onde é possível observar a minimização da perda tanto para os dados de treinamento, quanto para os de validação em Fig. 4a. Essa minimização da perda aparente no eixo *y* conforme a época avança no eixo *x* indica que a RNAD tem o seu modelo aprendeu com os dados fornecidos adequadamente, de forma que apresente aumento em no desempenho, ou seja, na capacidade de predição. Além disso, as Fig. 4b, 4c e 4d apresentam uma comparação entre os dados obtidos por meio de simulação (no eixo *x*) e os resultados preditos pela RNAD (no eixo *y*), explicitando-se nas figuras os respectivos coeficientes de correlação (R^2). Toma-se como referência a reta $y = x$, de modo que os valores que estejam no entorno dessa linha indiquem a previsão de dados realizada pela RNA, sendo o treinamento eficiente e sem sobreajuste de dados. Igualmente, a Fig. 5 apresenta os gráficos obtidos a partir do uso das RNAIs para o processo de ajuste do modelo de regressão aos dados disponíveis. Nesse sentido, as Fig. 5b, 5c e 5d apresentam a comparação entre os resultados obtidos pela simulação e os preditos pela RNAI, onde é possível identificar os coeficientes de correlação com maiores desvios da idealidade se comparados ao caso direto. Isso se dá pela complexidade das situações que a análise inversa provoca, já que podem haver múltiplas possibilidades de combinações entre valores de entrada que produzem um resultado de interesse.

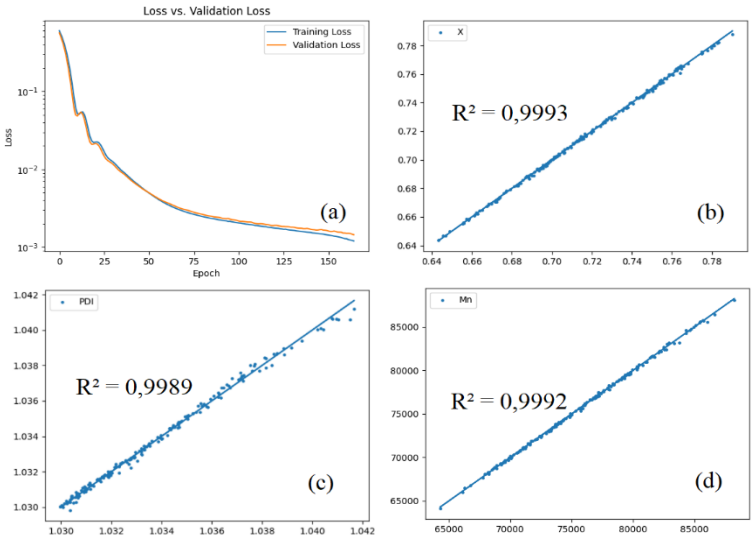

Fig. 4. Resultados para a rede neural direta (RNADs). Fonte: autoria própria.

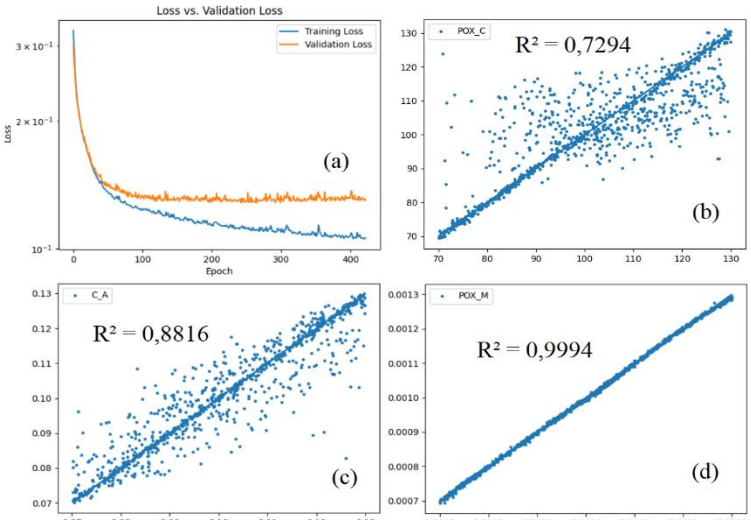

Fig. 5. Resultados para a rede neural indireta (RNAIs). Fonte: autoria própria.

Apesar dos menores valores observados nos coeficientes de correlação da RNAI em relação à RNAD, não são obtidas diferenças significativas entre os resultados da simulação e os da predição, como é possível verificar na comparação de resultados presentes na Tabela 1, que apresenta valores de erro percentual médio absoluto (*mean absolute percentagem error* ou MAPE) para cada variável de no máximo 5%, aproximadamente. Foram utilizados 100 casos teste, sendo que o aplicativo possibilita inserir a quantidade de valores desejados para realizar o cálculo de média do erro. Independentemente do desvio considerável da reta de referência nas Fig. 5b e 5c, a previsão realizada pela regressão é aplicável, pois se encontra em um intervalo de valores aceitáveis no que tange aplicações práticas.

Tabela 1. Erros percentuais médios absolutos para as predições da rede neural direta e inversa.

RNAD	X (MAPE, %)	PDI (MAPE, %)	Mn (MAPE, %)
	0,11	0,01	0,30
RNAI	$[P_0X]/[C]$ (MAPE, %)	[C]/[A] (MAPE, %)	$[P_0X]/[M]$ (MAPE, %)
	5,24	3,65	0,47

Fonte: autoria própria.

CONCLUSÕES:

Desenvolveu-se um aplicativo em Python™ para realizar predições de propriedades do polímero e conversão de monômero utilizando redes neurais artificias diretas (RNADs), de modo que as entradas da rede fossem as relações de concentração. Além disso, utilizando-se de redes neurais artificiais indiretas (RNAIs), pode-se predizer a condição operacional de concentração de agente redutor, catalisador e reagente (monômero de estireno) no início da reação química, dado um processo síntese do poliestireno em um reator batelada operando a temperatura constante [5]. Essa abordagem permite integrar a análise e o tratamento de grandes volumes de dados com simulações e predições em um único meio, sendo relevante para a otimização de processos industriais. Posteriormente, pode-se manipular a temperatura como condição de interesse, já que as constantes da taxa são funções de Arrhenius e essas são funções da temperatura [5].

As redes neurais multicamadas (MLP) criadas puderam confirmar a predição adequada das propriedades do polímero, bem como das relações de concentração, obtendo-se erros (MAPE) de no máximo 5%, aproximadamente, para o caso inverso, sendo considerados resultados satisfatórios. Reforça-se que apesar de o treinamento e validação da RNA inversa possuir um tempo superior de execução por conta das possíveis entradas que podem gerar as mesmas saídas, os erros obtidos não são grosseiros e são aplicáveis, independentemente do menor valor de R^2. Ainda, é possível realizar uma filtragem estatística dos dados por meio do gráfico de coordenadas paralelas, aumentando a restrição dos dados utilizados para treinamento das RNAs.

Nesse sentido, a ferramenta desenvolvida torna eficiente e prática a integração entre engenheiros/cientistas e dados experimentais ou de simulação, independentemente do tamanho da amostra, já que foi utilizada a técnica estatística de LHS para obtenção de valores confiáveis com mínima correlação entre as variáveis de entrada na rede [6]. Essa abordagem de desenvolvimento permitiu a generalização dos procedimentos e, por isso, não é aplicável somente às reações de polimerização. O *software*, portanto, automatiza o processo de simulação, gera RNAs com hiperparâmetros otimizados, determina a melhor condição operacional e fornece resultados estatístico em gráficos/tabelas sem depender de complexa interação com o usuário. Por meio desse procedimento eficaz e inovador, facilita-se o processo de tomada de decisão ao realizar um estudo de caso.

BIBLIOGRAFIA

[1] GANTZ, J.; REINSEL, D. The digital universe in 2020: big data, bigger digital shadows, and biggest growth in the Far East. [s.l.]: [s.n.], 2020.

[2] HIMMELBLAU, D. M. Applications of artificial neural networks in chemical engineering. **Korean Journal of Chemical Engineering**, [s.l.], v. 17, n. 4, p. 373–392, 2000.

[3] CUKIER, K. The rise of big data: how it's changing the way we think about the world. **Foreign Affairs**, [s.l.], v. 92, n. 3, p. 28–40, 2013.

[4] GUARDIEIRO, A. Tudo que você já deveria saber sobre otimização de hiperparâmetros em redes neurais — Parte I. Datarisk.io, 8 abr. 2021. Disponível em: https://medium.com/datarisk-io/tudo-que-voc%C3%AA-j%C3%A1-deveria-saber-sobre-otimiza%C3%A7%C3%A3o-de-hiperpar%C3%A2metros-em-redes-neurais-parte-i-f1d8975f0177. Acesso em: 3 ago. de 2024.

[5] BANIN, G.; VIEIRA, R. P.; LONA, L. M. F. Artificial neural networks towards average properties targets in styrene ARGET-ATRP. **Chemical Engineering Journal**, [s.l.], v. 407, p. 126999, 2021.

[6] SHEIKHOLESLAMI, R.; RAZAVI, S. Progressive Latin Hypercube Sampling: an efficient approach for robust sampling-based analysis of environmental models. Environmental Modelling & Software, v. 93, p. 109-126, 2017. ISSN 1364-8152. Disponível em: https://doi.org/10.1016/j.envsoft.2017.03.010. Acesso em: 3 ago. 2024.

[7] DE WINTER, J. C. F.; GOSLING, S. D.; POTTER, J. Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. **Psychological Methods**, v. 21, n. 3, p. 273-290, set. 2016.

[8] MEHRALIZADEH, Amir; DERAKHSHANFARD, Fahimeh; GHAZI TABATABEI, Zohreh. Applications of multi-layer perceptron artificial neural networks for polymerization of expandable polystyrene by multi-stage dosing Initiator. **Iranian Journal of Chemistry and Chemical Engineering**, v. 41, n. 3, p. 890-901, 2022.