XL CONGRESSO BRASILEIRO DE SISTEMAS PARTICULADOS
ENEMP 2022
23 a 26 de outubro de 2022
Uberlândia/MG

ANÁLISE EXPLORATÓRIA DE BANCO DE DADOS E APLICAÇÃO ALGORITMO DE APRENDIZAGEM DE MÁQUINA NA PREDIÇÃO DO DESEMPENHO DE MEMBRANAS PARA PERMEAÇÃO DE GASES

LHUCAS T. DE M. DE SOUSA[1], THAYLANE DA R. BEZERRA[1], SARAH. A. ALTINO [1*]

[1]Universidade Federal de Uberlândia, Programa de Pós-Graduação em Engenharia Química
Faculdade de Engenharia Química
*e-mail: sarah.arvelos@ufu.br

RESUMO - Os modelos de aprendizado de máquina vêm se destacando ao longo dos anos pelo fato de descobrir padrões em dados sem a necessidade de programação baseada em regras. O presente trabalho teve o objetivo de estudar a aplicação do algoritmo de aprendizado de máquina Floresta Aleatória para a predição da performance de diferentes membranas empregadas na permeação de gases puros. Além disso, o trabalho visou investigar os principais fatores que influenciam nesse processo. Um banco de dados, constituído por 533 registros referentes a trabalhos experimentais, foi construído com base em 16 referências bibliográficas diferentes. Foram consideradas 9 variáveis de entrada: o Tipo, a Espessura Média, o Tamanho de Poro e a Idade da membrana; a Temperatura e a Pressão do Processo; e o Diâmetro cinético, Massa Molar e a Polarizabilidade do gás de alimentação. A Permeabilidade foi considerada como a variável alvo. Após otimizar os parâmetros do modelo, foram obtidos coeficientes de determinação (R^2) de teste de 0,92 com uma Raiz do Erro quadrático Médio de 326,3 *barrer*. Os principais fatores que influenciaram na predição foram o diâmetro cinético do gás de alimentação, o tamanho médio do poro e a espessura média da membrana.

INTRODUÇÃO

Décadas de modelagem, simulações e experimentos forneceram à comunidade de engenharia química uma enorme quantidade de dados, que adicionam a opção de fazer previsões a partir da experiência (Piccione, 2019). Os modelos de aprendizado de máquina (AM) são modelos estatísticos e matemáticos que podem "aprender" com a experiência e descobrir padrões em dados sem a necessidade de programação baseada em regras (Dobbelaere *et al*., 2021; Thebelt *et al*., 2022).

A Floresta Aleatória, FA, (do inglês, *Random Forest*) é um procedimento popular de AM que pode ser usado para desenvolver modelos de previsão. Introduzido pela primeira vez por Breiman (2001), FAs são uma coleção de árvores de decisão (AD) de classificação ou regressão, que são modelos simples que usam divisões binárias em variáveis preditoras para determinar previsões de resultados (Speiser *et al*., 2019). O algoritmo FA se tornou muito popular porque combina a interpretabilidade das ADs com o desempenho de algoritmos de aprendizagem modernos, como redes neurais e as máquinas de vetores de suporte (Altmann *et al*., 2010).

Dentre as formas de se interpretar modelos de AM, destaca-se a técnica de análise de importâncias de permutação (IP). A IP de uma característica é o aumento do erro de previsão do modelo depois de termos permutado os valores do recurso, que quebra a relação entre o recurso e o resultado real (Miller, 2019; Molnar, 2020).

As membranas podem ser definidas como uma barreira que separa duas fases e que restringe total ou parcialmente o transporte de uma ou várias espécies químicas presentes nas fases. As suas propriedades de transporte, em específico, sua permeabilidade e capacidade seletiva, determinarão suas circunstâncias de utilização e estão associadas com o material e a metodologia em que a membrana foi fabricada (Habert *et al*., 2006).

O uso de membranas nas atividades relacionadas à separação de gases vem se destacando ao longo dos anos devido a melhor adequação aos interesses em fortalecer uma economia circular e sustentável, dado suas vantagens associadas ao menor custo e consumo de energia. Atualmente, existem membranas sendo aplicadas em várias atividades industriais relacionadas a separação de gases (Kamble *et al*., 2021). Uma forma bastante utilizada de se avaliar o desempenho de uma membrana para a separação de uma mistura gasosa é a seletividade ideal. No caso de misturas binárias, esta seletividade nada mais é que a razão entre as permeabilidades dos gases puros em uma condição fixa de temperatura e pressão (Saqib *et al*., 2020).

A permeabilidade de um gás (P_A) é a capacidade da membrana de permitir que o gás permeante (A) se difunda através do material da membrana como consequência da diferença de pressão sobre a membrana, e pode ser medida em termos de vazão do permeado, espessura e área da membrana e pressão diferença através da membrana. Por definição,

$$P_A = \frac{V_p t}{A_m(p_h - p_l)} \quad (1)$$

na qual, P_A representa a permeabilidade (Barrer), V_p o fluxo de permeado (cm³/s), A_m a área superficial da membrana (cm²) e p_h e p_l representam a pressão (cmHg) nos lados da alimentação e do permeado, respectivamente.

Assim, este estudo teve como objetivo confeccionar um banco de dados que apresentasse dentre suas variáveis informações potencialmente relacionadas ao desempenho de permeação de gases puros em membranas de diferentes tipos. Além disso, visou obter *insights* para o desenvolvimento de materiais porosos. Dados de permeabilidade de diversos gases leves puros sobre estes materiais publicados entre 2009 e 2022 foram extraídos e avaliados fazendo uso de análise exploratória de dados, regressão pelo algoritmo FA e análise de importâncias de permutação.

METODOLOGIA

Primeiramente, uma análise exploratória foi realizada na literatura de forma a elucidar as variáveis que: i) mais influenciam nos processos de separação por membranas, ii) são mais frequentemente fornecidas em publicações acadêmicas; iii) sejam linearmente independentes e iv) que caracterizem de forma mais genérica os tipos de membranas os quais o presente trabalho se propôs a investigar.

Desta forma, um banco de dados[1], constituído por 533 registros referentes a trabalhos experimentais de membranas aplicadas na separação de gases, foi construído com base em 16 referências bibliográficas diferentes. A ferramenta *WebPlotDigitizer* v4.5 (Rohatdi, 2021) foi utilizada para a prospecção de dados apresentados em forma gráfica. Para cada registro foram extraídos 9 atributos de entrada (Tabela 1) que foram divididas em 2 categorias: características do processo e da membrana. É possível observar na Tabela 1 o intervalo de cada variável (entrada) presente no banco de dados.

A permeância e a permeabilidade dos gases são variáveis habituais utilizadas para representar a performance das membranas. A permeância é preferivelmente calculada no contexto em que as membranas são muito finas ou são consideradas assimétricas. No caso de membranas simétricas ou quando a espessura é bem definida (normalmente em filmes densos), a permeabilidade é preferivelmente calculada (Du *et al*., 2012). A performance das membranas (variável alvo) foi representada pela permeabilidade (em *barrer*) dos gases de alimentação devido a sua maior frequência de disponibilização nos trabalhos de referência. Para alguns poucos registros que forneceram apenas a permeância, a permeabilidade foi calculada a partir da Equação 2.

$$P_A = Q_A \times l \quad (2)$$

Em que Q_A equivale à permeância do gás A em $cm^3(CNTP).cm^{-2}.\ s^{-1}.cmHg^{-1}$ e l à espessura média experimental da membrana em cm.

[1] https://github.com/tenoriolms/databank_ENEMP

As membranas podem ser classificadas por vários aspectos: quanto a origem, morfologia, geometria, mecanismo de transporte e aplicação (Sadrzadeh e Mohammadi, 2019). No banco de dados construído a variável "Tipo" caracteriza a membrana quanto ao material de fabricação (origem), se é inorgânica, orgânica ou uma Membrana de Matriz Mista (MMM) (membranas híbridas). As médias da espessura e do tamanho de poro apresentadas no banco de dados foram calculadas utilizando o valor mínimo e máximo das respectivas variáveis.

Tabela 1: Atributos de entrada coletados e seus respectivos intervalos ou classes no banco de dados.

Categoria	**Variável de entrada**	**Intervalo das variáveis numéricas ou classes das variáveis categóricas**
Variáveis de processo	Temperatura (K):	293 - 773
	Diferença de pressão entre a corrente de alimentação e do permeado (MPa):	0,02 – 6,35
	Diâmetro cinético do gás de alimentação (pm):	260 - 550
	Massa molar do gás de alimentação (g/mol):	2,02 – 146,06
	Polarizabilidade do gás de alimentação ($Å^3$):	0,208 – 4,49
Variáveis morfológicas da membrana	Tipo:	Poliméricas; Membranas de Matriz Mista (MMM); e inorgânicas.
	Espessura média (µm):	0,5 – 132,508
	Tamanho médio do poro (nm):	0,372 – 1,2
	Idade (dia):	0 - 2117

Os gases da corrente de alimentação foram representados por três propriedades físicas: diâmetro cinético, massa molar e a polarizabilidade (Li *et al*., 2009). O diâmetro cinético está relacionado com a esfera de influência para colisões moleculares e, de certa forma, com a resistência. A massa molar está relacionada com a difusão livre desses gases (Joos e Freeman, 1958; Pan *et al*., 2022). As interações eletrônicas entre a superfície dos poros e o gás podem ser explicadas (em parte) com base na polarizabilidade das moléculas (Meek *et al*., 2012; Schäf *et al*., 2020).

Vale salientar que não existe uma padronização no fornecimento desses atributos de entrada e, portanto, algumas variáveis estarão ausentes em determinadas referências. A presença de dados incompletos é um problema comum na ciência de dados e que pode ser corrigido de várias maneiras (Faceli *et al*., 2011). Algumas hipóteses foram realizadas para o preenchimento de alguns valores faltantes: a temperatura ambiente foi considerada como sendo 298 K e foram consideradas novas (Variável Idade = 0) as membranas cujas referências não forneceram informações sobre a envelhecimento. A Espessura Média e o Tamanho Médio do Poro também possuíram valores faltantes. No entanto, devido à natureza complexa dessas variáveis, o preenchimento desses campos por valores fixos (média, moda ou variável *dummy*) provocaria um enviesamento no banco de dados. Sendo assim, decidiu-se empregar um algoritmo indutor para estimar os valores desses atributos (Faceli *et al*., 2011). O modelo de AM K-NN (K Nearest Neighbor) foi aplicado para realizar o preenchimento desses campos (Faisal e Tutz, 2021).

Com o intuito de avaliar a confiabilidade do modelo de FA, dois indicadores estatísticos foram utilizados entre os valores calculados e os reais: o Coeficiente de Determinação (Equação 3) e a Raiz do Erro Quadrático Médio (Equação 4).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (3)$$

$$REQM = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (4)$$

Em que $\hat{y}_i$ equivale ao valor predito e y_i equivale ao valor real do i-ésimo registro. $\bar{y}_i$ corresponde à média dos valores reais e n ao número total de registros. As variáveis do banco de dados foram analisadas estatisticamente pelo Coeficiente de Correlação de Pearson, calculado pela Equação 5.

$$\rho = \frac{cov(x_i,x_j)}{\sqrt{\sigma^2(x_i)\sigma^2(x_j)}} \quad (5)$$

Em que $cov(x_i, x_j)$ corresponde à covariância das variáveis independentes x_i e x_j. $\sigma^2(x_i)$ e $\sigma^2(x_j)$ corresponde às suas variâncias. Esse coeficiente corresponde a uma medida quantitativa da dependência linear entre duas variáreis aleatórias.

Para busca dos (hiper)parâmetros ótimos do modelo FA foi utilizada a biblioteca computacional *Optuna* (Akiba *et al*., 2019). Tal biblioteca permite a pesquisa automatizada de (hiper)parâmetros ao implementar o algoritmo Bayesiano como metodologia de otimização (Pravin *et al*., 2022).

A função objetivo que foi utilizada pelo algoritmo Bayesiano fez uso os coeficientes de determinação do método *k-fold cross-validation*[2]. Nessa abordagem, divide-se os dados de treino aleatoriamente em *k* lotes de tamanho aproximadamente igual. Os lotes de $k-1$ partições são treinadas por um preditor e a partição restante é então testada/validada. Esse processo é repetido *k* vezes de forma que cada lote será usado como teste (Faceli *et al*., 2011; Izbicki e Santos, 2020). Assim, o processo de validação cruzada divide o conjunto de treino original em duas partições: treino e validação. Desse modo, a função objetivo que foi minimizada é apresentada na Equação 6.

$$F_{obj} = (1 - \bar{R}^2_{valid}) + \left|1 - \frac{\bar{R}^2_{treino}}{\bar{R}^2_{valid}}\right| \quad (6)$$

Em que $\bar{R}^2_{treino}$ e $\bar{R}^2_{valid}$ correspondem à média dos coeficientes de determinação obtidos no treinamento e validação, respectivamente, após aplicar o método de validação cruzada no conjunto de treino. No procedimento de busca, 500 passos iterativos (tentativas) foram realizados para a minimização da função objetivo. Os parâmetros que foram avaliados durante a otimização, bem como sua breve descrição, são apresentados na Tabela 2.

Para avaliar as importâncias das variáveis de entrada foi utilizado o recurso de permutação[3]. Nessa análise, a importância de uma variável é proporcional à diminuição na qualidade de um modelo (que pode ser medida por R^2 ou RMSE) quando ela é removida aleatoriamente. Deste modo, quebra-se a relação entre a variável de entrada e o alvo, e portanto, a queda na qualidade serve como indicativo de quanto o modelo depende da variável (Breiman, 2001).

Imputação dos valores ausentes utilizando K-NN: De um total de 533 dados, 86 registros para a espessura média e 207 registros do tamanho médio do poro estão originalmente ausentes nos trabalhos de referência. Esses valores foram preenchidos com o emprego do modelo de AM K-NN (*K Nearest Neighbor*) (Faisal e Tutz, 2021).

Demais variáveis disponíveis no banco de dados, inclusive a Permeabilidade, foram empregadas como variáveis de entrada durante o procedimento de preenchimento. Foram feitas duas regressões para preenchimento de cada variável alvo: uma para a Espessura Média e outra para o Diâmetro Médio do Poro. A mesma função objetivo descrita anteriormente, utilizando a biblioteca *Optuna*, foi empregada para a otimização dos parâmetros do modelo. A quantidade de vizinhos próximos a serem considerados no cálculo foram variados de 1 a 12. As métricas testadas para o cálculo das distâncias foram a euclidiana, *manhattan*[4]e *minkowski*[5].O número de lotes (*k-folds*) da validação cruzada foram variados de 2 a 13. As variáveis de entrada, os valores dos parâmetros otimizados e os coeficientes de determinação

[2] https://scikit-learn.org/stable/modules/cross_validation.html

[3]https://scikit-learn.org/stable/modules/permutation_importance.html#id2

[4]https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.cityblock.html#scipy.spatial.distance.cityblock

[5]https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.minkowski.html#scipy.spatial.distance.minkowski

obtidos para cada modelagem podem ser observados na Tabela 3.

Tabela 2: Parâmetros avaliados durante a otimização do modelo Florestas Aleatórias para a predição da permeabilidade.

Parâmetros	Intervalos de procura para as variáveis contínuas ou classes para as categóricas.
Número de árvores de decisão na Floresta. (*n_estimators*)	1-100
Função para medir a qualidade da divisão de um nó. (*criterion*)	Erro quadrático, absoluto e de Poisson
Número de atributos de entrada a serem considerados na procura da melhor divisão.	4 – 9
Máxima profundidade de uma árvore de decisão. *(max_depth)*	2-14
O mínimo de amostras requeridas em um nó de folha. (*min_samples_leaf*)	1 - 4
Número de divisões para a Validação Cruzada. (*k-fold*)	2-13

RESULTADOS E DISCUSSÃO

Análise Básica do Banco de Dados

Dos três valores únicos da variável Tipo (de membranas) presentes no banco de dados, 38% correspondem às membranas poliméricas, 31% às MMM e 30% às membranas inorgânicas (Figura 1). Durante a prospecção dos dados, ficou evidente que a espessura se mostra como uma das principais características das membranas. Consultando a Equação 2, sabemos que o desempenho de separação de uma membrana está correlacionado com a sua respectiva espessura. A análise do banco de dados mostrou que mesmo que a tendência entre permeabilidade e a espessura tenham relações de intensidades diferentes a depender do material em que a membrana é sintetizada, é evidente que uma menor espessura promove uma alta permeabilidade.

Membranas densas com baixa espessura apresentam uma alta permeabilidade devido a um menor tempo de passagem pelo meio poroso e uma maior frequência de colisão entre as paredes dos poros (Saini e Awasthi, 2022). Os registros de espessuras médias presentes no banco de dados variaram de 0,05 a 132,5 µm (Figura 1). As espessuras das membranas poliméricas apresentaram as maiores dimensões e as das híbridas se encontraram interpostas entre as dos outros dois tipos.

Além disso, pela natureza do fenômeno sabemos que o tamanho de poro também se mostra como um fator determinante para a permeação nesses materiais. No contexto de membranas em que o peneiramento molecular é predominante moléculas maiores que o tamanho do poro da membrana serão rejeitadas, enquanto moléculas menores irão passar (Sadrzadeh *et al*., 2018). Em membranas consideradas densas, em que os poros possuem um tamanho de nível molecular, o transporte de gás é governado pelo mecanismo de "difusão-solução". Moléculas menores que o poro da membrana irão se transportar por difusão enquanto aquelas maiores e condensáveis se transportarão pelo fenômeno de sorção (Sadrzadeh e Mohammadi, 2019). Os registros de tamanho médio de poro presentes no banco de dados variaram de 0,372 a 1,2 nm (Figura 1).

Os efeitos do envelhecimento físico das membranas irão depender do material ao qual elas são sintetizadas e as membranas orgânicas (em específico, as fabricadas de polímeros vítreos) são aquelas que mais sofrem com essa variável. Nesses materiais há uma perda no volume livre e um consequente declínio da permeabilidade devido a um rearranjo segmental para um estado de maior equilíbrio (Alberto *et al*., 2018; Ameen *et al*., 2021; Bernardo *et al*., 2017). Desta forma, a idade da membrana é um fator importante a ser considerado na previsão da performance. Os registros de Idade no banco de dados variaram de 0 a 2117 dias (Figura 1). Não houve registros de membranas inorgânicas envelhecidas.

Vários gases de alimentação estão presentes no banco de dados. São eles: He, H_2, CO_2, O_2, N_2, CO, CH_4, etano e SF_6. Na Figura 2 é possível observar o número de exemplos desses gases no banco de dados. Em sua maioria tem-se CH_4, CO_2, N_2, H_2 e o He. Os

restantes dos gases foram encontrados em poucas referências, porém foram mantidos com o objetivo de fornecer ao modelo de AM um caráter mais generalista.

As temperaturas registradas no banco de dados variaram entre 293 e 773 K (Figura 1). Por motivos de estabilidade térmica, percebeu-se que apenas membranas inorgânicas foram testadas a temperaturas acima de 338 K. Os registros da variável Diferença de Pressão variaram entre 0,02 e 6,35 MPa (Figura 1). Não foram registradas membranas inorgânicas submetidas a altas pressões, acima de 1 MPa. Os registros de permeabilidade variaram entre 0,02 e 8427 *barrer*, com os três tipos de membrana apresentando um mesmo perfil para essa variável.

A Figura 3 mostra uma matriz dos Coeficientes Correlação de Pearson (ρ) entre as variáveis do banco de dados. Caso duas variáveis possuam uma forte correlação linear e forem mantidas, o modelo pode ser afetado negativamente (Yang *et al*., 2005). Valores de $\rho > 0{,}9$ podem ser consideradas fortemente correlacionadas entre si (Mukaka, 2012) e, geralmente, alguma variável é retirada apenas se o Coeficiente de Pearson for muito alto e ultrapassar 0,8 ou 0,9 (Pan *et al*., 2022; Were *et al*., 2015). Sendo assim, neste trabalho as variáveis com uma correlação acima de 0,9 foram retiradas. Se ρ for positiva as variáveis relacionadas são diretamente proporcionais e se for negativa, as variáveis são inversamente proporcionais.

Na Figura 3, é possível perceber que a maioria das variáveis não possuem correlação linear entre si, indicando uma relação complexa entre elas. Além disso, as variáveis referentes às propriedades físicas do gás estão bastante correlacionadas entre si, porém não o suficiente para prejudicar a capacidade preditiva do modelo. Portanto, elas foram mantidas com o intuito de conhecer as suas respectivas relevâncias para a predição da permeabilidade pelo modelo na análise de importâncias.

Predição da permeabilidade e análise de importâncias

Após a otimização, os valores dos parâmetros encontrados para o modelo de FA podem ser observados na Tabela 4. O $\bar{R}^2_{treino}$ foi de 0,835 com um REQM de 563,8 *barrer*. Como podem ser observados na Figura 4, o alto valor de $R^2_{teste} = 0{,}928$ e o baixo valor para o $REQM = 326{,}3\ barrer$ indicam que o modelo de FA é apropriado para prever a permeabilidade de diferentes gases com variados tipos de membranas e que os resultados obtidos estão, de alguma forma, correlacionados com os dados experimentais.

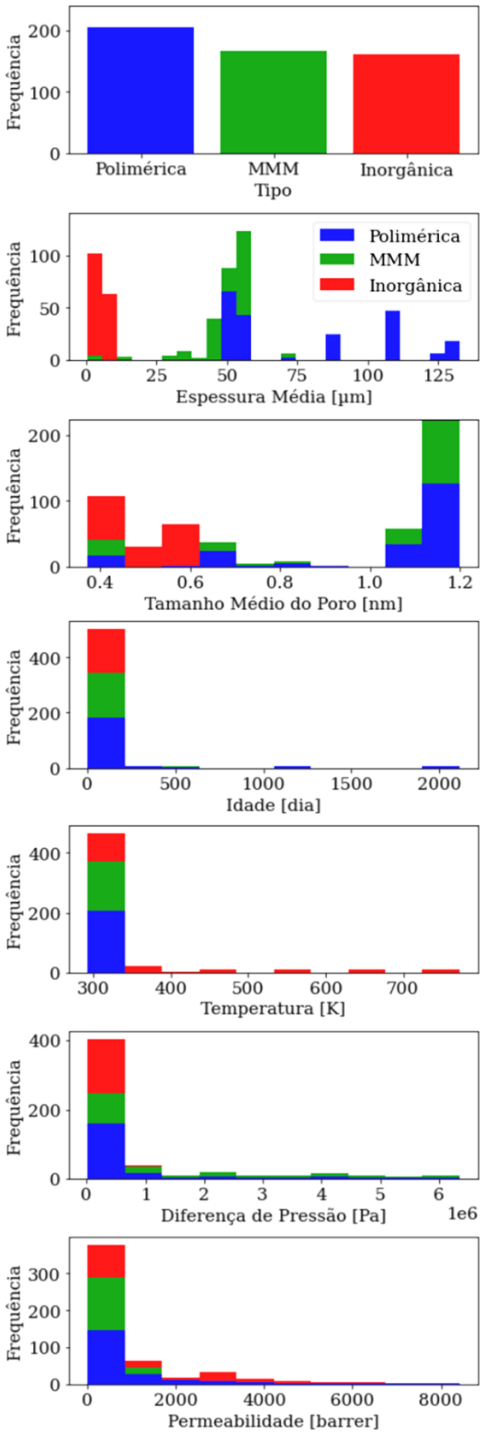

Figura 1: Histogramas das variáveis de entrada referentes à morfologia da membrana, da Temperatura, da Diferença de Pressão e da variável alvo Permeabilidade.

Tabela 3: Variáveis de entrada, parâmetros otimizados e os coeficientes de determinação obtidos para cada modelagem feita para a imputação dos dados ausentes.

Variável imputada (alvo)	**Variáveis de entrada**	**Parâmetros otimizados**	**Coeficientes de determinação***
Espessura média	Tipo; idade; temperatura; diferença de pressão; diâmetro cinético, massa molar e polarizabilidade do gás de alimentação; e permeabilidade.	Número de Vizinhos: 4 Métrica de distância: *Manhattan* *k-fold:* 6	$\bar{R}^2_{valid} = 0.9379$ $R^2_{teste} = 0.9632$
Tamanho médio do poro	Tipo; temperatura; diferença de pressão; diâmetro cinético, massa molar e polarizabilidade do gás de alimentação; e permeabilidade.	Número de Vizinhos: 2 Métrica de distância: *Manhattan* *k-fold*: 12	$\bar{R}^2_{valid} = 0.9544$ $R^2_{teste} = 0.8992$

*$\bar{R}^2_{valid}$ corresponde à média dos coeficientes de determinação do conjunto de validação após aplicar o método de validação cruzada no conjunto de treino. R^2_{teste} corresponde ao coeficiente de determinação do conjunto de teste.

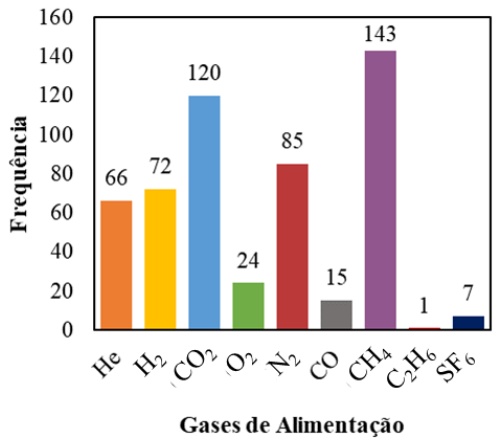

Figura 2: Gases puros de alimentação presentes no banco de dados.

Pan *et al.* (2022) trabalharam com um modelo de AM denominado Máquina de Vetores de Suporte no contexto de separação de gases em membranas de peneira moleculares de carbono, em um banco de dados de 399 itens, contendo dados sobre gases diversos gases puros (CO_2, CH_4, N_2, O_2, e H_2). Neste banco de dados idealizado por estes autores, as propriedades consideradas dos gases permeantes foram: massa, diâmetro cinético e potencial de Van der Waals entre o gás e o carbono. Os autores obtiveram um R^2_{teste} de 0,841 para a predição da permeabilidade. Guan *et al.* (2022) também utilizam o modelo de FA, porém para prever as permeabilidades em membranas do tipo MMM sintetizadas com MOF (em inglês, *Metal Organic Framework*). Os autores computaram 648 exemplos para gases puros (CO_2 e CH_4) e suas misturas. Como variáveis de entrada, foram testadas características intrínsecas das membranas, propriedades da matriz polimérica e condições de operação dos experimentos (temperatura e pressão). Eles encontram um R^2_{teste} de 0,77 para a predição da permeabilidade.

Assim, comparando-se a metodologia apresentada no presente trabalho com os trabalhos recentes consultados na literatura, nota-se que a inclusão de gases diversos no banco de dados bem como também a inclusão da polarizabilidade dentre as variáveis de entrada faz com que seja possível obter-se uma boa representatividade da permeabilidade do gás por um modelo de FA.

Uma análise de importâncias das variáveis do modelo FA por permutação foi realizada utilizando o conjunto de treino. Seus resultados normalizados podem ser visualizados na Figura 5. A princípio, pode-se observar que a Idade, a Temperatura e a Diferença de Pressão são as variáveis menos importantes para a exatidão do modelo, com um valor de 0,02 cada.

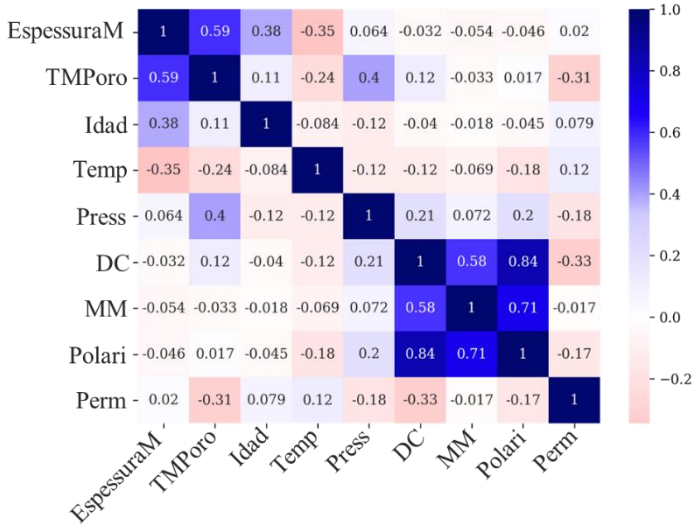

Figura 3: Matriz de correlação de Pearson entre as variáveis de entrada e alvo do banco de dados. LEGENDA: EspessuraM = Espessura Média. TMPoro = Tamanho Médio do Poro. Idad = Idade. Temp = Temperatura. Press = Diferença de Pressão. DC = Diâmetro Cinético do Gás de Alimentação. MM = Massa Molar do Gás de Alimentação. Polari = Polarizabilidade do Gás de Alimentação. Perm = Permeabilidade.

Tabela 4: Valores otimizados para os parâmetros do modelo de FA.

Parâmetros	**Valor otimizado**
n_estimators	30
criterion	Erro quadrático
max_features	6
max_depth	6
min_samples_leaf	1
k-fold	6

Pela natureza do fenômeno, sabe-se que Temperatura e Pressão são fatores que afetam diretamente a difusividade de gases em meios porosos. Logo, especula-se que a baixa influência da Temperatura e da Diferença de pressão ocorreu devido à baixa variabilidade dos dados para as características citadas. 75% dos registros referentes a temperatura e pressão se encontram abaixo de 308K e 0,6MPa respectivamente.

Variáveis pouco importantes, também chamadas de variáveis ruído (em inglês, *noisy features*), são têm menor influência sobre a variável alvo e, em geral, podem ser retiradas da modelagem, pois isso permite, virtualmente, que a predição seja mais precisa e eficiente (Li *et al*., 2019; Phung *et al*., 2022). Contudo, mesmo sendo as menos importantes, a retirada das variáveis em questão acarretou numa pequena queda de exatidão do modelo (otimizado sem essas variáveis). Foi decidido então mantê-las para análises de importâncias posteriores.

O diâmetro cinético (com 0,54) foi a variável de entrada mais relevante entre todas as incluídas na regressão, sendo a mais importante entre todas as propriedades físicas do gás de alimentação seguida da polarizabilidade (com 0,11) e da massa molar (com 0,05), respectivamente. O tamanho do poro (0,46) e a espessura da membrana (0,43) também foram variáveis bastante relevantes para a predição da permeabilidade, sendo os atributos mais importantes referentes à morfologia da membrana. No contexto de membranas de peneiras moleculares de carbono, Pan *et al.* (2022) observaram que as variáveis estruturais (morfológicas) foram mais importantes dos que as referentes às propriedades dos gases.

Para realização de um estudo comparativo com o trabalho de Pan *et al.* (2022), com o modelo já treinado, foram realizadas análises de importâncias considerando em cada uma delas, um subconjunto de treino filtrado conforme uma determinada característica. A Figura 6 mostra as importâncias normalizadas tendo por base a variável Tipo.

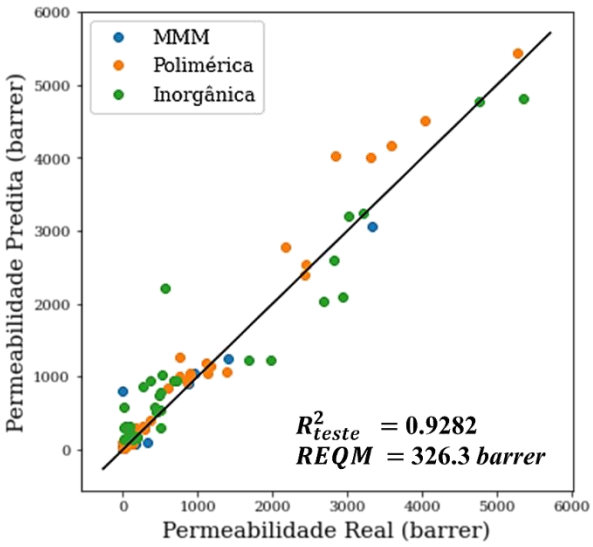

Figura 4: Permeabilidades reais e preditas do conjunto de teste. REQM = Raiz do Erro Quadrático Médio.

Novamente, pode-se perceber que a temperatura, a pressão e a idade praticamente não possuem relevância na predição. Além disso, o diâmetro cinético se mostrou bastante relevante, independentemente do tipo de membrana, acima das outras propriedades físicas do gás de alimentação. O diâmetro cinético está relacionado com o livre percurso médio das moléculas e portanto, é fundamental para a compreensão do comportamento dos gases em meios porosos e para a determinação de suas respectivas permeabilidades (Kunze *et al*., 2022). Li *et al.* (2022) observaram experimentalmente uma relação inversa entre o diâmetro cinético e a permeabilidade de gases puros (H_2, CO_2, N_2 e CH_4) numa MMM.

Dentre as características morfológicas da membrana, é possível perceber na Figura 6 que nas membranas inorgânicas, diferentes das outras, o tamanho de poro é a variável mais importante compreendendo 46,7% de importância. Nas membranas poliméricas e híbridas, a relevância do tamanho de poro foi pequena e a espessura se apresentou como uma das categorias mais importantes, com 39,3 e 42,2% respectivamente.

De fato, o transporte de gases nas membranas orgânicas em questão (que são poliméricas e consideradas densas) é governado pelo mecanismo de "difusão-solução" que depende mais da espessura da membrana do que do tamanho do poro, que são pequenos. (Sadrzadeh e Mohammadi, 2019). É evidente que membranas densas mais finas, apesar de normalmente sofrerem com alta frequência de defeitos e serem mais sensíveis à efeitos de envelhecimento, possuem uma alta permeabilidade devido ao menor tempo de passagem do gás ao longo da membrana e à maior frequência de colisões nos caminhos dos poros (Saini e Awasthi, 2022).

No que concerne às membranas híbridas. A adição de "*fillers"* (porosos ou não) ocasiona na criação de espaços vazios e no acréscimo do volume livre na matriz polimérica. Dessa forma, a permeabilidade da membrana aumenta quando comparada a sua respectiva versão "sem *fillers*" (Kamble *et al*., 2021; Saini e Awasthi, 2022). Ainda assim, sua análise de importância se assemelha mais ao tipo da sua matriz, que é feita de polímeros, pois depende bem mais da sua espessura.

Em comparação com as demais, as membranas inorgânicas normalmente possuem uma permeabilidade muito maior. Realmente, nesses materiais a separação pode ser regida geralmente por diferentes mecanismos de transporte que serão determinados principalmente pelo tamanho de poro da membrana. Dentre eles estão o fluxo de *Poiseuille* (escoamento convectivo), difusão de *Knudsen*, difusão superficial e peneiramento molecular (Sadrzadeh e Mohammadi, 2019).

A Figura 7 mostra as importâncias normalizadas de acordo com os gases de alimentação que foram mais frequentes no banco de dados. Os gases estão em ordem crescente de diâmetro cinético. Primeiramente, pode-se perceber que a espessura da membrana se torna mais relevante conforme o aumento do diâmetro cinético. De fato, é adequado afirmar que a resistência à permeabilidade aumentará conforme a espessura da membrana e que quanto maior for o diâmetro cinético do gás (menor for o livre percurso médio), mais essa forma de resistência se mostrará importante.

Além disso, observa-se que o tamanho do poro possui um comportamento crescente até o CO_2 (valor máximo), logo após diminui até se tornar insignificante no CH_4. As outras variáveis somadas possuíram uma pequena importância de maneira geral, com valores somados sempre abaixo de 9,3%.

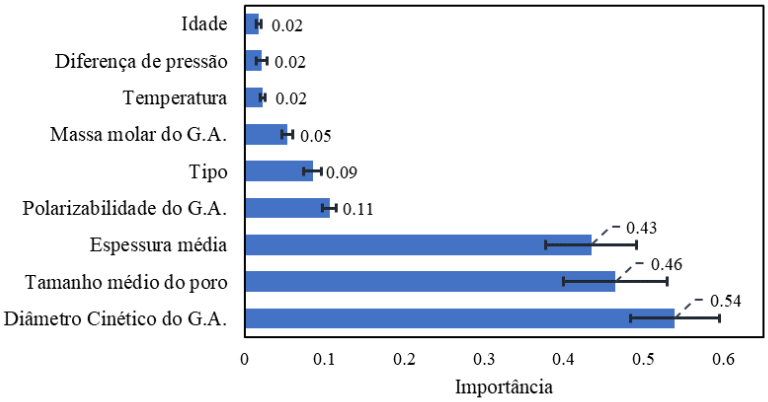

Figura 5: Importâncias de cada variável para a predição da permeabilidade. G.A. = Gás de Alimentação.

A variável Diferença de Pressão foi uma variável pouco relevante para a predição da permeabilidade de todos os gases. Dessa forma, o modelo de FA treinado com um banco de dados heterogêneo foi capaz de segregar adequadamente os atributos mais importantes tanto para cada tipo de membrana, quanto para cada gás de alimentação.

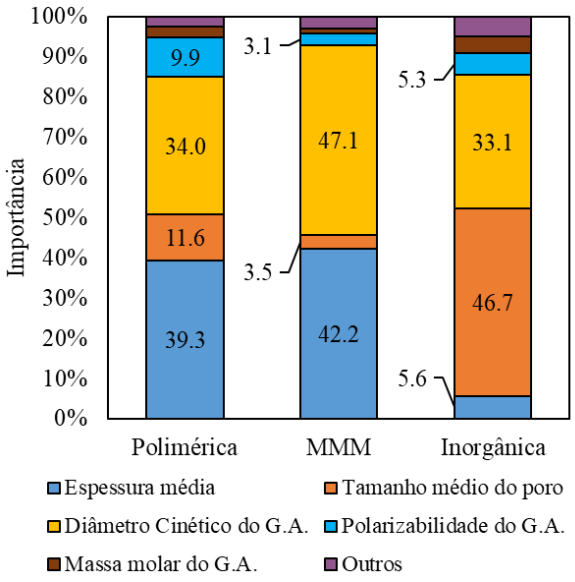

Figura 6: Importâncias relativas dos subconjuntos de treino filtrados conforme o tipo de membrana. G.A. = Gás de Alimentação.

CONCLUSÕES

Um banco de dados referentes a membranas para separação de gases foi construído com dados experimentais de 16 referências diferentes.

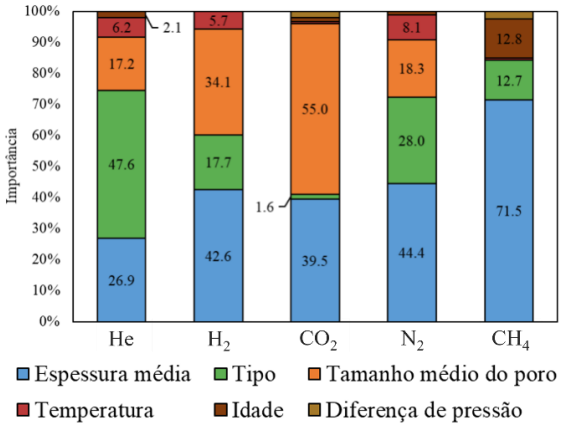

Figura 7: Importâncias relativas dos subconjuntos de treino filtrados conforme os gases de alimentação mais frequentes no banco de dados. Os gases estão em ordem crescente referente ao diâmetro cinético.

Os registros se mostraram bem distribuídos quanto ao tipo de membrana e à permeabilidade dos gases. Além disso, foram considerados 9 diferentes gases na corrente de alimentação (He, H_2, CO_2, O_2, N_2, CO, CH_4, etano e SF_6.). O modelo de aprendizado de máquina FA foi utilizado com o objetivo de prever a permeabilidade desses gases puros. Um alto valor de R^2_{teste} e um baixo valor para o $REQM$ foram obtidos, indicando que o modelo de FA é apropriado para prever a permeabilidade de diferentes gases em variados tipos de membranas. Os principais fatores que influenciaram na predição foram o diâmetro cinético do gás de alimentação, o tamanho médio do poro e a espessura média da membrana. A temperatura, a diferença pressão e a idade da membrana foram os atributos menos relevantes para o modelo.

AGRADECIMENTOS

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001. O presente trabalho foi realizado com apoio da Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG), Edital 001/2021 – Demanda Universal.

NOMENCLATURA

AD	Arvore(s) de Decisão
AM	Aprendizado de Máquina
A_m	Área superficial da membrana (cm^2)
$cov(x_i, x_j)$	covariância das variáveis x_i e x_j
FA	Floresta(s) Aleatória(s)
F_{obj}	Função Objetivo
IP	Importância(s) de Permutação
K-NN	*K Nearest Neighbor*
l	Espessura média experimental da membrana (*cm*)
MMM	Membrana(s) de Matriz Mista
MOF	*Metal Organic Framework*
P_A	Permeabilidade do gás puro A (*barrer*)
p_h	Pressão na corrente de alimentação (*cmHg*)
p_l	Pressão no permeado (*cmHg*)
Q_A	Permeância do gás puro A ($cm^3(CNTP).cm^{-2}.s^{-1}.cmHg^{-1}$)
R^2	Coeficiente de Determinação
$REQM$	Raiz do Erro Quadrático Médio
$\bar{R}^2_{treino}$	Média dos coeficientes de determinação do conjunto de validação da validação cruzada.
R^2_{teste}	Coeficiente de determinação do conjunto de teste
$\bar{R}^2_{valid}$	Média dos coeficientes de determinação do conjunto de treino da validação cruzada.
V_p	Fluxo de permeado (cm^3/s)
y_i	Valor real do i-ésimo registro
$\hat{y}_i$	Valor predito do i-ésimo registro
$\bar{y}_i$	Média dos valores reais
ρ	Coeficiente de Correlação de Pearson
$\sigma^2(x_i)$	Variância da variável x_i

REFERÊNCIAS

AKIBA, T., SANO, S., YANASE, T., OHTA, T., KOYAMA, M. (2019), Optuna: A Next-generation Hyperparameter Optimization Framework.

ALBERTO, M., BHAVSAR, R., LUQUE-ALLED, J.M., VIJAYARAGHAVAN, A., BUDD, P.M., GORGOJO, P (2018), Impeded physical aging in PIM-1 membranes containing graphene-like fillers. J. Memb. Sci. 563, 513–520.

ALIBAKSHI, A. (2018), Strategies to develop robust neural network models: Prediction of flash point as a case study. Anal. Chim. Acta 1026, 69–76.

ALTMANN, A., TOLOŞI, L., SANDER, O., LENGAUER, T. (2010), Permutation importance: a corrected feature importance measure. Bioinformatics 26, 1340–1347.

AMEEN, A.W., JI, J., TAMADDONDAR, M., MOSHENPOUR, S., FOSTER, A.B., FAN, X., BUDD, P.M., MATTIA, D., GORGOJO, P. (2021), 2D boron nitride nanosheets in PIM-1 membranes for CO_2/CH_4 separation. J. Memb. Sci. 636, 119527.

BERNARDO, P., BAZZARELLI, F., TASSELLI, F., CLARIZIA, G., MASON, C.R., MAYNARD-ATEM, L., BUDD, P.M., LANČ, M., PILNÁČEK, K., VOPIČKA, O., FRIESS, K., FRITSCH, D., YAMPOLSKII, Y.P., SHANTAROVICH, V., JANSEN, J.C. (2017), Effect of physical aging on the gas transport and sorption in PIM-1 membranes. Polymer (Guildf). 113, 283–294.

BREIMAN, L. (2001), Random Forests. Mach. Learn. 45, 5–32.

DOBBELAERE, M.R., PLEHIERS, P.P., VAN DE VIJVER, R., STEVENS, C. V., VAN GEEM, K.M. (2021), Machine Learning in Chemical Engineering: Strengths, Weaknesses, Opportunities, and Threats. Engineering 7, 1201–1211.

DU, N., PARK, H.B., DAL-CIN, M.M., GUIVER, M.D. (2012), Advances in high permeability polymeric membrane materials for CO_2 separations. Energy Environ. Sci. 5, 7306–7322.

FACELI, K., LORENA, A.C., GAMA, J., CARVALHO, A.C.P.D.L.F.D. (2011), Inteligência artificial: uma abordagem de aprendizado de máquina, 1°. ed. LTC, Rio de Janeiro.

FAISAL, S., TUTZ, G. (2021), Multiple imputation using nearest neighbor methods. Inf. Sci. (Ny). 570, 500–516.

GÉRON, A. (2019), Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow, 1°. ed. Alta Books, Rio de Janeiro.

GUAN, J., HUANG, T., LIU, W., FENG, F., JAPIP, S., LI, J., WANG, X., ZHANG, S. (2022), Design and prediction of metal organic framework-based mixed matrix membranes for CO2 capture via machine learning. Cell Reports Phys. Sci. 3, 100864.

HABERT, A.C., BORGES, C.P., NOBREGA, R.

(2006), Processos de separação por membranas. Editora e-papers.

IZBICKI, R., SANTOS, T.M. DOS (2020), Aprendizado de máquina: uma abordagem estatística, 1°. ed. Rafael Izbicki, São Carlos, SP.

JOOS, G., FREEMAN, I.M. (1958), Theoretical physics, Third. ed. Blackie and Son, Glasgow, London.

KAMBLE, A.R., PATEL, C.M., MURTHY, Z.V.P. (2021), A review on the recent advances in mixed matrix membranes for gas separation processes. Renew. Sustain. Energy Rev. 145, 111062.

KUNZE, S., GROLL, R., BESSER, B., THÖMING, J. (2022), Molecular diameters of rarefied gases. Sci. Rep. 12, 2057.

LI, J.-R., KUPPLER, R.J., ZHOU, H.-C. (2009), Selective gas adsorption and separation in metal–organic frameworks. Chem. Soc. Rev. 38, 1477–1504.

LI, W., LI, Y., CARO, J., HUANG, A. (2022), Fabrication of a flexible hydrogen-bonded organic framework based mixed matrix membrane for hydrogen separation. J. Memb. Sci. 643, 120021.

LI, X., WANG, Y., BASU, S., KUMBIER, K., YU, B. (2019), A Debiased MDI Feature Importance Measure for Random Forests.

MEEK, S.T., TEICH-MCGOLDRICK, S.L., PERRY, J.J., GREATHOUSE, J.A., ALLENDORF, M.D. (2012), Effects of Polarizability on the Adsorption of Noble Gases at Low Pressures in Monohalogenated Isoreticular Metal–Organic Frameworks. J. Phys. Chem. C 116, 19765–19772.

MILLER, T. (2019), Explanation in artificial intelligence: Insights from the social sciences. Artif. Intell. 267, 1–38.

MOLNAR, C. (2020), Interpretable Machine Learning, 1°. ed. Lulu.com.

MUKAKA, M.M. (2012), Statistics corner: A guide to appropriate use of correlation coefficient in medical research. Malawi Med. J. 24, 69–71.

PAN, Y., HE, L., REN, Y., WANG, W., WANG, T. (2022), Analysis of Influencing Factors on the Gas Separation Performance of Carbon Molecular Sieve Membrane Using Machine Learning Technique. Membranes (Basel). 12, 100.

PHUNG, V.L.H., OKA, K., HIJIOKA, Y., UEDA, K., SAHANI, M., WAN MAHIYUDDIN, W.R. (2022), Environmental variable importance for under-five mortality in Malaysia: A random forest approach. Sci. Total Environ. 845, 157312.

PICCIONE, P.M. (2019), Realistic interplays between data science and chemical engineering in the first quarter of the 21st century: Facts and a vision. Chem. Eng. Res. Des. 147, 668–675.

PRAVIN, P.., TAN, J.Z.M., YAP, K.S., WU, Z. (2022), Hyperparameter optimization strategies for machine learning-based stochastic energy efficient scheduling in cyber-physical production systems. Digit. Chem. Eng. 4, 100047.

ROHATDI, A. (2021), WebPlotDigitizer, v4.5. Pacifica, California, USA. https//automeris.io/WebPlotDigitizer.

SADRZADEH, M., REZAKAZEMI, M., MOHAMMADI, T. (2018), Fundamentals and Measurement Techniques for Gas Transport in Polymers, in: Transport Properties of Polymeric Membranes. Elsevier, p. 391–423.

SAINI, N., AWASTHI, K. (2022), Insights into the progress of polymeric nano-composite membranes for hydrogen separation and purification in the direction of sustainable energy resources. Sep. Purif. Technol. 282, 120029.

SAQIB, S., RAFIQ, S., MUHAMMAD, N., KHAN, A.L., MUKHTAR, A., MELLON, N.B., MAN, Z., NAWAZ, M.H., JAMIL, F., AHMAD, N.M. (2020), Perylene based novel mixed matrix membranes with enhanced selective pure and mixed gases (CO_2, CH_4, and N_2) separation. J. Nat. Gas Sci. Eng. 73, 103072.

SCHÄF, O., TORTET, L., SIMON-MASSERON, A., PATARIN, J., DEFOUR, S., BLANC, R., COSTE, C., ZEREGA, Y. (2020), Importance of PCDD/F molecules' polarizability and steric hindrance on their adsorption onto zeolites in a standard EN1948-1 sampling device for incinerator emission monitoring. Chemosphere 259, 127457.

SPEISER, J.L., MILLER, M.E., TOOZE, J., IP, E. (2019), A comparison of random forest variable selection methods for classification prediction modeling. Expert Syst. Appl. 134, 93–101.

THEBELT, A., WIEBE, J., KRONQVIST, J., TSAY, C., MISENER, R. (2022), Maximizing information from chemical engineering data sets: Applications to machine learning. Chem. Eng. Sci. 252, 117469.

WERE, K., BUI, D.T., DICK, Ø.B., SINGH, B.R. (2015), A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecol. Indic. 52, 394–403.

YANG, S., LU, W., CHEN, N., HU, Q. (2005), Support vector regression based QSPR for the prediction of some physicochemical properties of alkyl benzenes. J. Mol. Struct. THEOCHEM 719, 119–127.