UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE ENGENHARIA QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA QUÍMICA

LHUCAS TENÓRIO DE MELO DE SOUSA

APRENDIZADO DE MÁQUINA APLICADO À MODELAGEM DA PERMEAÇÃO DE GASES EM MEMBRANAS VISANDO A PURIFICAÇÃO DE METANO

Uberlândia - MG

2023

LHUCAS TENÓRIO DE MELO DE SOUSA

APRENDIZADO DE MÁQUINA APLICADO À MODELAGEM DA PERMEAÇÃO DE GASES EM MEMBRANAS VISANDO A PURIFICAÇÃO DE METANO

Dissertação apresentada ao Programa de Pós-graduação em Engenharia Química da Universidade Federal de Uberlândia como requisito parcial para obtenção do título de Mestre em Engenharia Química.

Área de concentração: Desenvolvimento de Processos Químicos

Orientador: Profª. Drª. Sarah Arvelos Altino

Coorientador: Prof. Drª Ubirajara Coutinho Filho

Uberlândia - MG

2023

Ficha Catalográfica Online do Sistema de Bibliotecas da UFU
com dados informados pelo(a) próprio(a) autor(a).

S725 Sousa, Lhucas Tenório de Melo de, 1998-
2023 Aprendizado de Máquina Aplicado à Modelagem da Permeação de Gases em Membranas Visando a Purificação de Metano [recurso eletrônico] / Lhucas Tenório de Melo de Sousa. - 2023.

Orientadora: Sarah Arvelos Altino.
Coorientador: Ubirajara Coutinho Filho.
Dissertação (Mestrado) - Universidade Federal de Uberlândia, Pós-graduação em Engenharia Química.
Modo de acesso: Internet.
Disponível em: http://doi.org/10.14393/ufu.di.2023.416
Inclui bibliografia.
Inclui ilustrações.

1. Engenharia química. I. Altino, Sarah Arvelos,1986-, (Orient.). II. Coutinho Filho, Ubirajara ,1970-, (Coorient.). III. Universidade Federal de Uberlândia. Pós-graduação em Engenharia Química. IV. Título.

CDU: 66.0

Bibliotecários responsáveis pela estrutura de acordo com o AACR2:
Gizele Cristine Nunes do Couto - CRB6/2091
Nelson Marcos Ferreira - CRB6/3074

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Coordenação do Programa de Pós-Graduação em Engenharia Química
Av. João Naves de Ávila, 2121, Bloco 1K, Sala 206 - Bairro Santa Mônica, Uberlândia-MG, CEP 38400-902
Telefone: (34)3239-4249 - www.ppgeq.feq.ufu.br - secppgeq@feq.ufu.br

ATA DE DEFESA - PÓS-GRADUAÇÃO

Programa de Pós-graduação em:	Engenharia Química				
Defesa de:	Mestrado Acadêmico, 14/2023, PPGEQ				
Data:	28 de julho de 2023	Hora de início:	14:00	Hora de encerramento:	15:45
Matrícula do Discente:	12122EQU006				
Nome do Discente:	Lhucas Tenório de Melo de Sousa				
Título do Trabalho:	Aprendizado de Máquina Aplicado à Modelagem da Permeação de Gases em Membranas Visando a Purificação de Metano				
Área de concentração:	Desenvolvimento de Processos Químicos				
Linha de pesquisa:	Engenharia Bioquímica				
Projeto de Pesquisa de vinculação:	Processos de separação utilizando membranas				
ODS-ONU:	ODS 9 – Indústria, Inovação e Infraestrutura				

Reuniu-se por meio de webconferência, a Banca Examinadora, designada pelo Colegiado do Programa de Pós-graduação em Engenharia Química, assim composta: Professores Doutores: Antonio José Gonçalves da Cruz - DEQ/UFSCar, Rubens Gedraite - PPGEQ/UFU, Ubirajara Coutinho Filho - PPGEQ/UFU, coorientador, e Sarah Arvelos Altino - PPGEQ/UFU, orientadora do candidato.

Iniciando os trabalhos a presidente da mesa, Profª Drª Sarah Arvelos Altino, apresentou a Comissão Examinadora e o candidato, agradeceu a presença do público, e concedeu ao discente a palavra para a exposição do seu trabalho. A duração da apresentação do discente e o tempo de arguição e resposta foram conforme as normas do Programa.

A seguir, a presidente concedeu a palavra, pela ordem sucessivamente aos examinadores, que passaram a arguir o candidato. Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando o candidato:

Aprovado

Esta defesa faz parte dos requisitos necessários à obtenção do título de Mestre.

O competente diploma será expedido após cumprimento dos demais requisitos, conforme as normas do Programa, a legislação pertinente e a regulamentação interna da UFU.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme foi assinada pela Banca Examinadora.

Documento assinado eletronicamente por **Sarah Arvelos Altino**, **Professor(a) do Magistério Superior**, em 28/07/2023, às 15:48, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Rubens Gedraite**, **Professor(a) do Magistério Superior**, em 28/07/2023, às 15:48, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Ubirajara Coutinho Filho**, **Professor(a) do Magistério Superior**, em 28/07/2023, às 15:48, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Antonio José Gonçalves da Cruz**, **Usuário Externo**, em 28/07/2023, às 15:49, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4673944** e o código CRC **37A0ED28**.

Referência: Processo nº 23117.049991/2023-11 SEI nº 4673944

A todos os meus bons professores.

AGRADECIMENTOS

A minha mãe, pelos carinhos, base e inspiração eterna.

A meu pai, pelo apoio incondicional.

À Elizabeth e Jannine, minhas companheiras de vida.

Aos poucos amigos que quero levar para a vida.

À Rafael, pelo calor da sua companhia, liberdade e perspectiva.

À Sarah, pela confiança e infinita paciência.

À extensa comunidade científica, em especial aos experimentalistas. Não se faz nada sozinho.

Às políticas públicas de democratização do acesso ao ensino superior de qualidade, pela oportunidade.

“Existe uma teoria que diz que, se um dia alguém descobrir exatamente para que serve o universo e por que ele está aqui, ele desaparecerá instantaneamente e será substituído por algo ainda mais estranho e inexplicável.

***

Existe uma segunda teoria que diz que isso já aconteceu.”

(Douglas Adams)

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 de três algoritmos de aprendizado de máquina (Floresta Aleatória, Regressão de Vetores de Suporte e Redes Neurais Artificiais) 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 repositório de dados (denominado Banco de Dados Original) foi construído, com 1672 registros referentes a trabalhos experimentais, provenientes de 42 referências bibliográficas diferentes, com foco naqueles que tenham estudado a capacidade de permeação do metano. Para a modelagem, foi realizada uma filtragem, de forma que foi utilizado um banco de dados constituído por 692 registros de 19 referências bibliográficas diferentes. Esses registros se mostraram bem distribuídos quanto ao tipo de membrana e à permeabilidade dos gases. Além disso, foram considerados 3 tipos de membranas diferentes (Polimérica, Zeolítica e de MOF) e 9 diferentes gases na corrente de alimentação (He, H_2, CO_2, O_2, N_2, CH_4, eteno, etano e SF_6) (alguns presentes em mistura binária). Inicialmente, foram considerados 16 atributos de entrada e a permeabilidade foi considerada como o atributo alvo. Todos os modelos de aprendizado de máquina testados se mostraram adequados para prever o desempenho de permeação pelas diferentes membranas em diferentes circunstâncias de utilização do banco de dados. Diferentes funções objetivo foram investigadas no processo de otimização dos hiperparâmetros do modelo. Os principais atributos que influenciaram na predição foram o Tamanho Médio de Poro, o Diâmetro Cinético do Gás de Alimentação e a Área Superficial Específica. A maioria dos outros atributos (Espessura Média, Volume Total de Poro, Idade, Temperatura, Diferença de Pressão, Diâmetro Cinético do Gás de Arraste, Massa Molar do Gás de Alimentação, Polarizabilidade do Gás de Alimentação e Fração do Gás) foram considerados pouco relevantes. Para simplificar a modelagem e retirar as variáveis ruído, alguns atributos foram retirados utilizando o modelo Florestas Aleatórias (FA). Nesse processo, seis atributos foram retirados e o desempenho de previsão permaneceu praticamente inalterado com Erro Percentual Absoluto Médio de 4,9 e 8,3 % e um R^2 de 0,98 e 0,96 para o conjunto de treino e teste, respectivamente. Por fim, as seletividades de quatro sistemas binários (He/CH_4, H_2/CH_4, CO_2/CH_4 e N_2/CH_4) foram estimadas utilizando o modelo FA treinado previamente, que também se mostrou adequado com a obtenção de um R^2 acima de 0,8 e um RMSE abaixo de 0,18 em todos os casos.

Palavras-chave: banco de dados; membranas; separação de gases; aprendizado de máquina; permeabilidade; seletividade.

ABSTRACT

Machine learning models have gained prominence over the years for discovering patterns in data without the need for rule-based programming. The present work aimed to study the application of three machine learning algorithms (Random Forest, Support Vector Regression, and Artificial Neural Networks) to predict the performance of different membranes used in the permeation of pure gases. Additionally, the work aimed to investigate the main factors influencing this process. A data repository (called the Original Database) was built, comprising 1672 records referring to experimental work, from 42 different bibliographic references, focusing on those who have studied the permeation capacity of methane. For modeling, filtering was performed, resulting in a database consisting of 692 records from 19 different bibliographic references. These records were well-distributed in terms of membrane type and gas permeability. Furthermore, 3 different membrane types (Polymeric, Zeolite, and MOF) and 9 different feed stream gases (He, H_2, CO_2, O_2, N_2, CH_4, ethylene, ethane, and SF_6) were considered. Initially, 16 input attributes were used, with permeability as the target attribute. All machine learning models evaluated were adequate to predict the permeation performance of different membranes under various circumstances of the database. Different objective functions were investigated in the process of optimizing the model's hyperparameters. The main attributes that influenced the prediction were the Average Pore Size, the Feed Gas Kinetic Diameter, and the Specific Surface Area. Most other attributes (Average Thickness, Total Pore Volume, Age, Temperature, Pressure Difference, Sweep Gas Kinetic Diameter, Feed Gas Molar Mass, Feed Gas Polarizability, and Gas Fraction) can be considered of low relevance to model prediction. To simplify the modeling and remove the noise variables, some attributes were removed using the Random Forests (FA) model. In this process, six attributes were removed, and the prediction performance remained unchanged with a Mean Absolute Percent Error of 4.9 and 8.3% and an R^2 of 0.98 and 0.96 for the training and test sets, respectively. Finally, the selectivities of four binary systems (He/CH_4, H_2/CH_4, CO_2/CH_4, and N_2/CH_4) were estimated using the previously trained FA model, which also proved to be adequate with an R^2 above 0.8 and an RMSE below 0.18 in all cases.

Keywords: database; membranes; gas separation; machine learning; permeability; selectivity.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE SÍMBOLOS

$\overline{MAPE}_{CV\_teste}$	Média dos MAPEs dos conjuntos de teste da Validação Cruzada
$\overline{MAPE}_{CV\_treino}$	Média dos MAPEs dos conjuntos de treino da Validação Cruzada
$\bar{R}^2_{CV\_teste}$	Média dos coeficientes de determinação dos conjuntos de teste da Validação Cruzada
$\bar{R}^2_{CV\_treino}$	Média dos coeficientes de determinação dos conjuntos de treino da Validação Cruzada
$\hat{y}_i$	Valor predito para o registro *i*
$\mathrm{F_i}$	Função objetivo *i*
$\mathrm{F_{obj}}$	Função objetivo referente ao processo de otimização dos modelos
R^2	Coeficiente de determinação
R^2_{treino}	Coeficiente de determinação do conjunto de treino
V_{ads}	Volume de gás adsorvido a uma determinada pressão e temperatura
V_m	Volume molar de um determinado gás no estado líquido
V_{poro}	Volume total de poro
e^{EQ}	Erro Quadrático
e^{ER}	Erro Relativo
e_l	Função de Custo do modelo Redes Neurais Artificiais
$e^{logcosh}$	Erro do Logaritmo do Cosseno Hiperbólico
f_{RBF}	Função de base radial
f_a	Função de ativação de um neurônio artificial do modelo Redes Neurais Artificiais
f_a^{LReLU}	Função de ativação *Leaky ReLU*
f_a^{ReLU}	Função de ativação ReLU
f_a^{sig}	Função de ativação sigmoide
$n_{entradas}$	Número de conexões de entrada de uma determinada camada do modelo Redes Neurais Artificiais
$n_{saídas}$	Número de conexões de saída de uma determinada camada do modelo Redes Neurais Artificiais
s_i	Valor (sinal) de saída do neurônio artificial *i* do modelo Redes Neurais Artificiais
y_i	Valor real do registro *i*

σ^2	Variância
Δp	Diferença de pressão entre os lados da membrana
C	Coeficiente de penalidade do modelo Regressão de Vetores de Suporte
J	Fluxo de gás no estado estacionário
P	Permeabilidade
Q	Permeância
R	Constante universal dos gases ideais
T	Temperatura
X	Atributo do banco de dados escalonado
cov	Covariância
p	Pressão
r	Valor limite da faixa de inicialização dos pesos do método Uniforme *Glorot* do modelo Redes Neurais Artificiais
t	Espessura do filme (membrana)
w	Peso referente a um determinado neurônio artificial do modelo Redes Neurais Artificiais
x	Atributo do banco de dados
$\mathcal{D}$	Coeficiente de difusão
$\mathcal{S}$	Coeficiente de solubilidade
α	Seletividade ideal
γ	Coeficiente de regularização das funções de *kernel* do modelo Regressão de Vetores de Suporte
ε	Espessura da margem do modelo Regressão de Vetores de Suporte
η	Taxa de Aprendizado do modelo Redes Neurais Artificiais
μ	Média aritmética simples
ρ	Coeficiente de Correlação de Pearson
σ	Desvio padrão

LISTA DE ABREVIATURAS E SIGLAS

AD	Árvores de Decisão
AE	*Absolute Error*
AM	Aprendizado de Máquina
API	*Application Programming Interface*
AR	Árvores de Regressão
BET	*Brunauer*, *Emmett* e *Teller*
CCP	Coeficiente de Correlação de Pearson
CMS	*Carbon Molecular Sieves*
CNTP	Condições Normais de Temperatura e Pressão
DTP	Distribuição do Tamanho dos Poros
EQ	Erro Quadrático
ER	Erro Relativo
FA	Florestas Aleatórias
FFV	*Fractional Free Volume*
GA	Gás de Alimentação
GDP	Gráfico de Dependência Parcial
GLP	Gás Liquefeito de Petróleo
HOF	*Hydrogen-bonded Organic Framework*
IA	Inteligência Artificial
ICE	*Individual Conditional Expectation*
IP	Importâncias de Permutação
Leaky ReLU	*Leaky Rectified Linear Unit*
Log-Cosh	Logaritmo do Cosseno Hiperbólico
MAPE	*Mean Absolute Percentage Error*
MICE	*Multivariate Imputation by Chained Equations*
MMM	Membranas de Matriz Mista
MOF	*Metal Organic Framework*
MSA	*Membrane Society of Australasia*
MVS	Máquina de Vetores de Suporte
PDP	*Partial Dependence Plot*
PGR	Processo Gaussiano de Regressão

PMC	Peneiras Moleculares de Carbono
PSA	*Pressure Swing Adsorption*
PSGM	Processos de Separação de Gases por Membrana
PSM	Processos de Separação por Membranas
RBF	*Radial Basis Function*
ReLU	*Rectified Linear Unit*
REQM	Raiz do Erro Quadrático Médio
RLM	Regressão Linear Múltipla
RMSE	*Root Mean Square Error*
RNA	Redes Neurais Artificiais
RT	*Room Temperature*
RVS	Regressão de Vetores de Suporte
SE	*Squared Error*
SG	Separação de Gases
TAE	Teoria do Aprendizado Estatístico
TPOT	*Tree-based Pipeline Optimization Tool*
VC	Validação Cruzada

SUMÁRIO

1 INTRODUÇÃO

Processos de Separação por Membranas (PSM) têm despertado o interesse de vários pesquisadores ao longo dos anos em função das suas diversas vantagens envolvendo o baixo custo energético, *design* simples, facilidade de operação e alta seletividade quando comparado com processos convencionais (KANCHERLA *et al.*, 2021). A eficiência dos PSM está fundamentalmente relacionada às características das membranas. Elas são o coração do processo e diferentes materiais (orgânicos ou inorgânicos) podem ser utilizados para a sua fabricação. Segundo Habert et al. (2006), 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 é fabricada (HABERT; BORGES; NOBREGA, 2006; VALAPPIL; GHASEM; AL-MARZOUQI, 2021). Suas aplicações abrangem diversas atividades econômicas incluindo o tratamento de água e esgoto; na indústria química, farmacêutica, alimentícia; e, inclusive, na área médica (HABERT; BORGES; NOBREGA, 2006).

O uso de membranas nas atividades relacionadas à Separação de Gases (SG) 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 (IULIANELLI; DRIOLI, 2020). Atualmente, existem membranas sendo aplicadas em várias atividades industriais relacionadas a separação de gases, em específico: na separação de O_2/N_2 para o enriquecimento de oxigênio; na separação de CO_2/CH_4 para a purificação de biogás; na separação de H_2/CO na produção de gás de síntese; e na separação de CO_2/N_2 para diminuir as emissões de carbono na atmosfera (KAMBLE; PATEL; MURTHY, 2021). Com a tecnologia de membranas, a separação de gases pode ser realizada em condições ambiente de baixa pressão e temperatura, o que traz uma vantagem sobre outros processos já bem estabelecidos como o *Pressure Swing Adsorption* (PSA) e a destilação criogênica (SAINI; AWASTHI, 2022). Diferentes desses outros processos, as membranas para SG não requerem uma separação de fases (BERNARDO, P.; DRIOLI; GOLEMME, 2009). É estimado que os processos de SG consumam cerca de 10 a 15% da energia produzida no mundo e em comparação com esses processos de separação tradicionais, as membranas podem economizar até 90% da energia consumida em um processo. No entanto, apesar do notável potencial (mais econômico e sustentável), suas aplicações para sistemas de engenharia ainda não são tão

numerosas, visto que que os outros processos são mais amadurecidos no mercado. O uso de membranas com uma alta seletividade e permeabilidade são a chave para a redução dos custos operacionais dos Processos de Separação de Gases por Membrana (PSGM) e paralelamente para a sua aplicação industrial. Na dessalinização de águas os PSM já são bem estabelecidos e competitivos no mercado, sendo nesse caso específico, 10 vezes mais eficientes energicamente quando comparadas com processos térmicos (BERNARDO, P.; DRIOLI; GOLEMME, 2009; WANG, Jing *et al.*, 2023).

Em específico, o CH_4 se mostra como uma importante fonte de energia e matéria-prima de vários processos industriais, cuja separação por membranas vem sendo estudada amplamente pela comunidade científica. A energia gerada pela combustão do metano está em torno de 9510 $kcal \cdot N^{-1} \cdot m^{-3}$ e esse gás pode ser extraído de diversas fontes, sendo o gás natural (encontrado em reservatórios no subsolo, associados ou não ao petróleo) a mais utilizada para o uso humano (WANG, Qi *et al.*, 2022). O uso de membranas para a purificação de gás natural é relatado desde 1994 e sua utilização está associada à produção de um produto de baixa pureza e à retirada de gases ácidos como CO_2 e H_2S (CARDOSO *et al.*, 2022).

Entretanto, a emissão de gases de efeito estufa proveniente de combustíveis fósseis e outras atividades humanas vem aumentando no último século. O aumento da concentração desses gases desde o século 19 está bastante correlacionado com o aquecimento global e as mudanças climáticas (ISMAIL; KHULBE; MATSUURA, 2015). A diminuição das emissões de CO_2 (que em sua maioria é de origem antropogênica) e o enfrentamento do aquecimento global podem ser considerados um dos desafios mais importantes da atualidade. Segundo o último relatório do IPCC (2018), é previsto que a temperatura média do planeta aumente 1,5 °C até 2052, com relação aos níveis da era pré-industrial (1850-1900). Além disso, é estimado que 1,0 °C desse aquecimento é proveniente de causas antropogênicas. Esses fatos alertam para a necessidade de novas tecnologias que utilizam energias limpas, que impeçam ou desacelerem as mudanças climáticas de forma que assegurem uma segurança energética global baseada em fontes de energia sustentáveis e renováveis (BERNARDO, Gabriel *et al.*, 2020).

Mesmo sendo o segundo gás que mais contribui para o aquecimento global, o metano é visto como uma fonte de energia necessária no processo de transição das matrizes energéticas rumo a uma econômica sustentável. Em comparação com o petróleo e carvão que são as fontes fósseis mais utilizadas atualmente, ele é o que possui a maior razão H/C e menor emissão de poluentes. Posto isso, o biogás é uma fonte de metano que tende a crescer bastante no decorrer dos anos, visto que é proveniente da fermentação de biomassa, uma fonte de energia abundante e renovável. Além disso, as membranas despontam como uma tecnologia atrativa por terem

aplicabilidade em qualquer fase no processo de fabricação do biogás, serem modulares, mais ecológicas, exigirem menor consumo de energia e possuírem alta eficiência e performance (CARDOSO *et al.*, 2022).

Diante desses desafios, as simulações/predições computacionais emergem como grandes aliados no auxílio da resolução desses problemas. As ferramentas matemáticas provaram ser extremamente bem-sucedidas no projeto e na otimização de processos físicos, químicos e biológicos. O uso da modelagem de processos e a simulação computacional vem crescendo no desenvolvimento industrial para projetar, analisar e otimizar processos integrados de modo a fornecer uma solução abrangente (CHIDAMBARAM, 2018).

Predições podem ser obtidas utilizando modelos teóricos, no entanto, muitos deles não podem ser calculados analiticamente e necessitam de um considerável poder computacional para serem obtidos numericamente. O longo tempo necessário para essa tarefa acarreta a utilização de modelos mais simples e, como consequência, uma perda de precisão dos resultados. Diversos modelos empíricos, semiempíricos ou teóricos foram elaborados na última década, porém eles enfrentam a dificuldade de terem uma baixa universalidade e precisão limitada. Além disso, a maioria das membranas se encontram numa fase bem distante do teste em escala piloto e sua performance de separação acaba sendo mais baixa do que o predito pelos modelos teóricos (DOBBELAERE *et al.*, 2021; WANG, Jing *et al.*, 2023).

O Aprendizado de Máquina (AM) pode ser compreendido como "a máquina aprendendo com o exemplo" e possui o potencial de realizar predições de forma precisa e rápida. Os algoritmos de AM possuem base em modelos estatísticos e matemáticos que, por uma análise de um conjunto de dados existentes, consegue identificar padrões que são imperceptíveis aos olhos humanos sem a necessidade de programação baseada em regras (DOBBELAERE *et al.*, 2021). Como campo de estudo, o AM é um subconjunto da Inteligência Artificial (IA). São exemplos de algoritmos aplicáveis à problemas de engenharia química: os “k-vizinhos mais próximos”, as Árvores de Decisão (AD), as Florestas Aleatórias (FA), as máquinas de vetores de suporte e as redes neurais artificiais (THEBELT *et al.*, 2022).

Décadas de modelagem, simulações e experimentos forneceram à comunidade da engenharia química uma enorme quantidade de dados, que adicionam esses algoritmos de AM no leque de opções de ferramentas para a triagem de novos materiais. A crescente digitalização dos processos industriais advinda do fenômeno conhecido como Indústria 4.0, fez crescer o número de dados disponíveis aos engenheiros químicos, os quais estão cada vez mais propensos a lidar no meio industrial com o que se conhece como Ciência de Dados (PICCIONE, 2019).

Nesse contexto, precisa-se compreender que a aplicação de algoritmos de aprendizado de máquina é essencial para tomar decisões rápidas.

O termo Ciência de Dados vem se tornando cada vez mais popular na indústria e nas disciplinas acadêmicas para se referir à combinação de estratégias e ferramentas para lidar com a enorme quantidade de dados disponíveis. O termo cientista de dados é um descritor comum de um engenheiro ou cientista de qualquer formação disciplinar que está equipado para processar, analisar e se comunicar nesse contexto de uso intensivo de dados. As principais áreas da Ciência de Dados são muitas vezes identificadas como: gerenciamento de dados, aprendizado estatístico e de máquina e visualização (BECK *et al.*, 2016).

Por outro lado, o grande crescimento de novos materiais em vista dos recursos experimentais limitados é claramente um grande obstáculo para otimização das estruturas de membranas. Em vez de adotar uma abordagem de tentativa e erro para o seu desenvolvimento, guiada apenas pela experiência e intuição do pesquisador, é muito mais eficiente buscar uma compreensão real dos fenômenos de separação para orientar os novos projetos. O estabelecimento de relações estrutura-propriedade para guiar um planejamento racional futuro de novos materiais e processos economiza tempo e dinheiro investido em testes de bancada. Métodos como Monte Carlo, Dinâmica Molecular e outras técnicas computacionais melhoraram a compreensão das relações entre as características da membrana e as propriedades de separação (THORNTON; HILL; HILL, 2010; WANG, Jing *et al.*, 2023) Dessa maneira, o AM também surge como uma ferramenta com potencial para ajudar a estabelecer essas associações para os PSM auxiliando no projeto de membranas mais adequadas para o futuro uso industrial e eximindo trabalhos desnecessários.

Nesse contexto, a capacidade de realização das associações estrutura-propriedade dependerá da “Interpretabilidade” do modelo. Interpretabilidade é o grau em que um ser humano pode entender a causa de uma decisão e apesar de serem modelos “*black box*”, existem muitas formas de se interpretar os modelos de AM. São exemplos os gráficos de dependência parcial (PDP, em inglês, *Partial Dependence Plot*), valores *Shapley*, substituto local ou global (local *surrogate* e global *surrogate*), gráficos de efeitos locais acumulados (*accumulated local effects*), expectativas individuais condicionais (ICE, em inglês, *Individual Conditional Expectation*) e a Importâncias de Permutação (IP) (MOLNAR, 2021).

1.1 Lacunas do campo de estudo e objetivos da dissertação

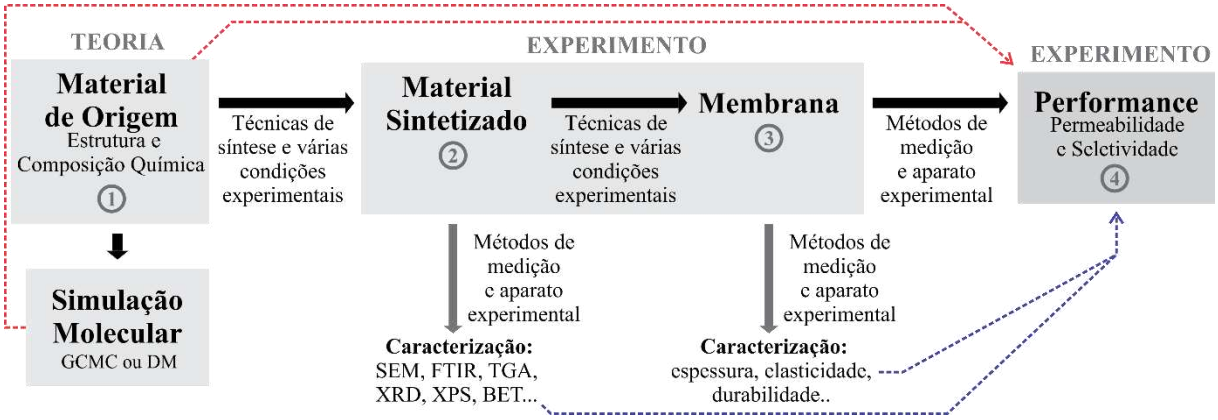
O aprendizado de máquina é um tema recente para processos de separação e ainda há muitas questões a serem exploradas. Mesmo com o gradativo crescimento da Indústria 4.0 e da pesquisa na área dos PSM, poucos estudos exploram a utilização de descritores (variáveis de entrada) experimentais no projeto das membranas. Esse fato revela que ainda existe uma escassez relativa à disponibilidade de informações detalhadas. "O principal problema dos dados na ciência de membranas atualmente é que eles carregam *informações pobres*, isto é, são caracterizados por terem um baixo volume e alta variabilidade" (WANG, Jing *et al.*, 2023). Desse fato surgiram estratégias interessantes e que se mostraram bem-sucedidas como, por exemplo, a utilização de descritores químicos para os polímeros no trabalho de Hasnaoui, Krea e Roizard (2017) ou de métodos de impressão digital no trabalho de Barnett et al. (2020). No contexto das membranas de Peneiras Moleculares de Carbono (PMC), Pan et al. (2022) utilizam como variáveis de entrada características do polímero precursor devido à disponibilização limitada das características desse tipo de membrana na literatura. A maioria dos trabalhos referentes às MOFs (do inglês, *Metal Organic Framework*) e zeólitas também associam as suas propriedades físicas teóricas com a performance das suas membranas.

Entretanto, apesar da escassez, os dados experimentais são mais realistas para a modelagem visto que vários fatores já são levados intrinsicamente em consideração. Os dados relativos à caracterização experimental de um material carregam consigo a história de uma série de procedimentos (etapas reacionais e de tratamentos físico-químicos) que foram realizados para a sua síntese. Por exemplo, no trabalho de Bernardo et al. (2017), uma série de membranas feitas do mesmo material (polímero PIM-1) apresentaram diferenças drásticas de permeação devido à diferença em seus respectivos pós-tratamentos com etanol a diferentes temperaturas. A espessura, a fragilidade ou a porosidade de uma membrana podem mudar em decorrência da técnica ou das condições ambientais em que elas são fabricadas. Dessa forma, a utilização de variáveis de entrada experimentais sensíveis a essa "história de procedimentos" irá descrever melhor o material e quanto mais genéricas elas forem, maior será a capacidade de universalização do modelo de AM. Entretanto, além da escassez, outro problema se mostra no fato de que a precisão e exatidão das medições experimentais estão sujeitas às qualidades do aparato experimental ao qual será feita a medida. Essa é uma questão de difícil resolução que envolve inclusive fatores socioeconômicos de forma que resta ter a consciência de que a precisão do modelo seguirá a dos trabalhos aos quais ele foi treinado.

Vale salientar que um modelo de AM é treinado para predizer o valor experimental da permeabilidade a partir de descritores teóricos. Nesse caso, o modelo retornará um resultado esperado dada uma condição experimental e de testagem mais comum (ZHU *et al.*, 2020). Como a permeabilidade é uma variável "carregada historicamente", em tese, ela seria uma ótima variável de entrada para preditores. Como exemplo, no trabalho de Yuan et al. (2021), observa-se que as permeabilidades de determinados gases foram utilizadas para prever a permeabilidade desconhecida de outros gases, visto que elas possuem uma alta correlação; e no trabalho de Guan et al. (2022) a permeabilidade da matriz polimérica foi utilizada para prever a permeabilidade da MMM (Membranas de Matriz Mista), sendo uma das variáveis mais importantes.

Na Figura 1 é exibido em forma de diagrama o processo de estudo (projeto) de uma membrana dividida em 4 etapas que sofrerão influência de várias fatores: (1) Modelo ideal do material de origem; (2) Síntese desse material; (3) Síntese da membrana; e (4) Testes de desempenho. Guan et al. (2022) e Pan et al. (2022) utilizaram descritores experimentais para modelar e analisar os fatores de influência da permeação de membranas. Pode-se afirmar, portanto, que esses autores utilizaram informações da 2° e 3° etapa para predizer a 4° etapa. Além disso, até onde foi apurado, não há registros na literatura de trabalhos que propuseram modelos (sejam teóricos ou não) que abrangessem mais de um tipo de membrana. Como já foi explanado, os modelos de AM estão sempre restritos a um tipo específico e essa capacidade de generalização nunca foi explorada. Os efeitos de mistura na corrente de alimentação também não foram relatados nesse campo de estudo. Guan et al. (2022) incluem os dados de permeação de mistura no banco de dados utilizado, porém não consideram ou analisam esse aspecto na modelagem e tratam todos os dados da mesma forma. Por outro lado, alguns estudos se propuseram a estudar a abrangência de vários gases em um único modelo de AM. Zhu et al. (2020) atribuíram um vetor para cada gás analisado, isto é, utilizam o "*one-hot-encoding*", e Pan et al. (2022) representaram os gases permeantes por três propriedades físicas (diâmetro cinético, massa molar e potencial de van der Waals).

Figura 1 – Diagrama das etapas do projeto de uma membrana e suas variáveis de influência.
LEGENDA: ■ Utilizando descritores teóricos; ■ Utilizando descritores experimentais.



Fonte: Elaborada pelo autor.

Em vista dessas lacunas, este estudo teve como objetivo geral realizar uma análise detalhada da complexa relação entre os parâmetros potenciais e o desempenho de permeação de gases em membranas de diferentes tipos, visando-se obter *insights* para o desenvolvimento desses materiais porosos. Para isso, focando trabalhos que tenham estudado a capacidade de permeação do metano, descritores experimentais e dados de permeação em membranas de diversos gases leves foram minerados e avaliados fazendo uso de análise exploratória de dados, modelagem por AM e análise de importâncias das variáveis de entrada. Objetivou-se também o treinamento de um modelo mais generalista a partir da utilização de um banco de dados que contenha diferentes tipos de membranas e da utilização de propriedades físicas para representar os gases permeantes na sua forma pura ou em mistura. Trabalhos como este ajudam os PSM a se estabelecerem no mercado e surgem para contribuir com a consolidação do AM como ferramenta de decisão em engenharia e de desenvolvimento de processo.

1.1.1 Objetivos secundários

Esta dissertação tem os seguintes objetivos secundários:

- Construir um banco de dados que abranja vários tipos de membranas e gases permeantes, tanto na forma pura quanto de mistura, de forma que ele seja adequado e seguro para a livre utilização da comunidade científica;
- Investigar a viabilidade do uso de três diferentes modelos de Aprendizado de Máquina (Florestas Aleatórias, Máquinas de Vetores de Suporte e Redes Neurais Artificiais) para previsão da permeação de diferentes gases em diferentes membranas;

- Investigar a utilização de diferentes funções objetivo no processo de otimização dos modelos com o foco em atenuar o problema do sobreajuste com mínima interferência do usuário;
- Investigar os atributos mais importantes para a previsão da permeação de gases através das membranas;
- Analisar a previsão da seletividade ideal dos sistemas He/CH_4, H_2/CH_4, CO_2/CH_4 e N_2/CH_4, utilizando os modelos treinados;
- Investigar os atributos mais importantes para a seletividade das membranas nos quatro sistemas binários.

1.2 Estrutura da dissertação

Após ser abordada uma introdução sobre os temas e os objetivos do presente trabalho, no capítulo 2 será apresentada uma breve revisão de literatura abordando os tópicos sobre os Processos de Separação de Membrana. Nele, há uma discussão sobre a importância histórica das membranas, assim como algumas de suas aplicações e conceitos gerais sobre o tema (classificação e fenômenos físicos existentes). No capítulo 3 será abordado uma breve discussão sobre os algoritmos de Aprendizado de Máquina, explicitando sobre os conceitos gerais e seus vieses indutivos. São discutidos também os principais estudos da literatura que foram referência para o presente trabalho. No capítulo 4, será abordada a metodologia empregada na dissertação, bem como também o detalhamento dos métodos e ferramentas computacionais utilizados. Nele, inicialmente é apresentado um fluxograma referente as etapas do trabalho. No capítulo 5 são apresentados os principais resultados alcançados durante o mestrado, desde a construção do banco de dados esclarecendo a definição de cada um de seus atributos até a aplicação dos modelos de Aprendizado de Máquina para a previsão da permeabilidade e da seletividade das membranas. Por fim, o capítulo 6 apresenta as principais conclusões obtidas no decorrer do trabalho. O referencial bibliográfico e os apêndices, onde aparecem alguns resultados que complementam a discussão do texto, são apresentados no final deste documento.

2 REVISÃO BIBLIOGRÁFICA: MEMBRANAS

2.1 Marcos históricos

A história do desenvolvimento dos PSM começa há mais de dois séculos e, em particular, está intimamente relacionada com o avanço da Ciência de Materiais. Os primeiros registros de estudos sobre membranas apareceram em 1748 por Nollet (1748), um abade francês. Ele observou que uma membrana de origem animal (bexiga) possuiu uma capacidade de separar o etanol de um destilado de vinho (HABERT; BORGES; NOBREGA, 2006). Por um certo tempo, esses materiais despertaram interesse somente de praticantes da área médica e biológica em que membranas naturais (de origem vegetal ou animal) eram utilizadas. Em 1831, Mitchell notou que gases eram capazes de permear filmes elásticos não porosos em diferentes intensidades e traz as explicações fenomenológicas primitivas sobre o fato, que esse processo estava relacionado com a solução e posterior difusão das moléculas ao longo do polímero. Ele conseguiu medir o vazamento de dez gases diferentes dentro de balões naturais (MATTEUCCI *et al.*, 2006; MITCHELL, John Kearsley, 1995; TABE-MOHAMMADI, 1999).

Em 1867, Traube sintetizou a primeira membrana inorgânica constituída de ferrocianeto de cobre apoiada em argila porosa e, contextualmente, apresentou uma ótima seletividade para a retirada de eletrólitos em variadas soluções (FANE; WANG; JIA, 2011; GLATER, 1998). Em 1907, há o primeiro relato da síntese de uma membrana com capacidade ultra filtrante, feita de nitrocelulose, por Bechhold (1907). Esse trabalho também se mostra como pioneiro no teste de membranas submetidas a altas pressões (GLATER, 1998; VALAPPIL; GHASEM; AL-MARZOUQI, 2021). Desde então, protótipos para osmose reversa começaram a ser estudados por alguns pesquisadores. Na década de 1930, as bases da eletrodiálise em membranas foram estabelecidas por Teorell e Meyer e, em 1945, a primeira hemodiálise bem-sucedida clinicamente foi realizada por Willem Johan (FANE; WANG; JIA, 2011).

Ao longo desses anos, esses materiais atraíram a atenção de diversos pesquisadores ajudando-os a elaborar equações e conceitos teóricos clássicos que são largamente utilizadas até hoje. O conceito de transporte de massa entre camadas com diferentes concentrações de um soluto proposto por Fick em 1855 e o mecanismo de solução-difusão proposto por Thomas Graham em 1866 foram desenvolvidos sob os conceitos de separação por membrana (VALAPPIL; GHASEM; AL-MARZOUQI, 2021). Em 1877, Pfeffer utilizou a membrana sintética de Traube para medir a pressão osmótica de uma solução, o que posteriormente levou à clássica teoria de soluções proposta por van’t Hoff’s em 1887, ganhador do primeiro Prêmio

Nobel de Química (1901). O efeito Gibbs–Donnan foi observado experimentalmente em 1911 utilizando esses materiais (FANE; WANG; JIA, 2011; GLATER, 1998). Em 1920, Daynes observou que existia uma relação entre o atraso de tempo ("*time lag*") e o coeficiente de difusão no transporte não estacionário de gases através de uma membrana (TABE-MOHAMMADI, 1999).

Até então, o uso primário das membranas estava associado com a eliminação de micróbios e macromoléculas em líquidos ou gases. Aliado ao investimento do governo dos Estados Unidos a partir do final da 2° guerra mundial, as membranas começaram a ser utilizadas para a sua primeira grande aplicação: a dessalinização de águas (VALAPPIL; GHASEM; AL-MARZOUQI, 2021). A primeira planta comercial para essa aplicação foi posta em operação em 1954 utilizando eletrodiálise e membranas de troca iônica (FANE; WANG; JIA, 2011). Anos depois, após vastas publicações e estudos com membranas poliméricas entre 1957-59, (REID; BRETON, 1959) observaram que o acetato de celulose se mostrava como um material promissor, exibindo altos níveis de rejeição salina, porém com um fluxo de água insatisfatório. Membranas ultrafinas (com espessura abaixo de 10 μm) submetidas a uma pressão acima de 85 bar apresentavam uma rejeição salina de aproximadamente 98% para uma solução de alimentação com 0,01 $mol \cdot L^{-1}$ de NaCl, porém apresentavam um fluxo de água na dimensão de $\mu L \cdot cm^{-2} \cdot h^{-1}$ (GLATER, 1998).

Vale salientar que as membranas poliméricas eram majoritariamente constituídas por derivados de celulose e aquelas mais modernas só foram possíveis após 1937, quando o nylon, a primeira poliamida sintética, foi fabricada por Wallace Carothers. Esse baixo fluxo presente nas membranas da época impedia que processos por osmose reversa possuíssem aplicações industriais até que em 1960-63 Loeb e Sourirajan, motivados a resolver esse problema, aperfeiçoaram a técnica de preparo dessas membranas (futuramente chamada "técnica de inversão de fase por imersão-precipitação") de forma que elas apresentaram um alto fluxo de permeado, livre de defeitos e mantendo a alta retenção de sais (FANE; WANG; JIA, 2011; HABERT; BORGES; NOBREGA, 2006). As membranas desenvolvidas por essa técnica apresentavam uma morfologia assimétrica: em um lado possuíam uma fina camada também chamada de "pele" (com poros muito pequenos, menores que 0,05 μm e com espessura em torno de 0,2 μm), responsável pela seletividade; e ao longo do material os poros aumentavam gradualmente servindo como sustentação mecânica com uma resistência mínima à permeação (HABERT; BORGES; NOBREGA, 2006; TABE-MOHAMMADI, 1999; VALAPPIL; GHASEM; AL-MARZOUQI, 2021). A grande diferença desse método se deu pela adição de

perclorato de magnésio na solução com o polímero dissolvido, que atuou como um agente formador de poros (GLATER, 1998).

Dessa forma, a primeira planta industrial que utilizava osmose reversa foi construída na cidade de Coalinga (California) em 1967. A fábrica produzia quase 19 mil litros de água potável por dia, quantidade adicional suficiente para suprir as necessidades de uma pequena comunidade (STEVENS; LOEB, 1967). Paralelamente, outras inovações sugiram como o projeto de uma planta de configuração placa-quadro com multiestágios e o desenvolvimento de métodos para a fabricação de membranas tubulares (GLATER, 1998). A descoberta de Loeb e Sourirajan se tornou um grande marco na área não apenas no âmbito da dessalinização das águas, mas também propiciou que a utilização das membranas fosse considerada para a comercialização de várias outras aplicações no decorrer dos anos como na separação de gases.

2.1.1 Separação de gases

As primeiras explicações sobre os fenômenos que ocorrem nos PSM foram realizadas utilizando gases. Como já citado, em 1831 Mitchell inicialmente percebeu que diferentes gases permeiam balões naturais de diferentes formas, mas foi Thomas Graham que propôs o conceito de solução-difusão explicando o mecanismo de transporte dos gases através de uma membrana em 1866. Ele também realizou experimentos em que conseguiu separar oxigênio do ar, estudou a permeação do hidrogênio e tentou sintetizar as primeiras membranas compostas (GRAHAM, 1866). R. M. Barrer, que possui uma unidade de medida em sua homenagem, também publicou importantes trabalhos relacionados à permeação de gases em 1939 e 1943 (BARRER, 1943; BARRER; RIDEAL, 1939). A primeira grande aplicação desse tipo de membrana se mostrou em 1945, na separação de Urânio 235 (de 0.17 para 3%) para propósitos militares e aplicações nucleares (FANE; WANG; JIA, 2011; TABE-MOHAMMADI, 1999). Mais de um século se passou desde o trabalho de Mitchell (em 1831) até a primeira grande aplicação das membranas para a separação de gases (em 1945), fato que ilustra a lentidão com a qual a área progride, apesar dos vários trabalhos/estudos experimentais.

As membranas assimétricas descobertas por Loeb e Sourirajan (em 1960-63) se mostraram bastante eficientes para a dessalinizadas de águas, porém não eram adequadas para a separação de gases, pois perdiam sua capacidade seletiva quando secas. Vos e Burris (1969), adicionando um surfactante a água ao qual eram produzidas, resolveram esse problema propiciando assim a introdução das membranas à indústria da separação de gases. O próximo grande progresso na preparação dessas membranas foi o desenvolvimento bem-sucedido de

membranas compostas por Ward, Browall e Salemme (1976). Em seguida, a técnica de revestimento criada por Henis e Tripodi (1980), em que borracha de silicone foi utilizada para obstruir os grandes poros e defeitos da camada seletiva, aprimorou esses materiais aumentando a seletividade H_2/CO_2 em um módulo de baixo custo (FANE; WANG; JIA, 2011; TABE-MOHAMMADI, 1999). No mesmo ano, o desenvolvimento dessas técnicas levou posteriormente à comercialização das membranas "PRISM" feitas de fibra oca de polissulfona pela empresa Monsanto. Elas eram usadas para a recuperação de hidrogênio em um gás de purga para fábricas de amônia e conseguiam retirar H_2 com uma pureza de 95% (HASHEMIFARD *et al.*, 2020).

Outras membranas começaram a ser comercializadas posteriormente na década de 80: a empresa Permea começou a produzir membranas de fibra oca utilizando complexos de ácido-base de Lewis; a Dow começou a construir sistemas para a separação de nitrogênio do ar; e as empresas Cynara, Separex e Grace a construir sistemas para a separação de CO_2 do metano a partir de gás natural. Outras patentes interessantes surgiram na década de 90: Pinnau e Koros (1990) produziram uma membrana ultrafina pelo método de inversão de fase em que a permeabilidade aumentou drasticamente; e Chung, Kafchinski e Vora (1995) produziram membranas assimétricas de fibra oca de alta performance com 6FDA-poliamida. Atualmente, as membranas são utilizadas nos mais diversos processos: na recuperação de gases específicos (hidrogênio, nitrogênio, dióxido de carbono, oxigênio, metano etc.); na desidratação de gás natural; e na remoção de hidrocarbonetos ou de gases ácidos (CO_2 e H_2S) em gás natural (FANE; WANG; JIA, 2011; LI, Guoqiang *et al.*, 2021).

2.2 Aplicações recentes em separação de gases

Considerando que, como apresentado, as membranas e os fenômenos associados a elas vêm sendo estudados há mais de um século, as suas aplicações industriais podem ser consideradas modernas (HABERT; BORGES; NOBREGA, 2006). No contexto das membranas poliméricas, visto que são as mais usadas comercialmente, apenas oito ou nove polímeros são utilizados atualmente para compor 90% das instalações de SG por membranas (ISMAIL; KHULBE; MATSUURA, 2015). No entanto, a pesquisa científica e o uso comercial dos PSM têm crescido no contexto de vários processos industriais em virtude das suas características que, em contraste com os processos convencionais de SG, são melhores. Em específico, ocupam menos espaço, consomem menos energia, não possuem partes móveis,

possuem uma lógica de construção simples e são mais ambientalmente limpos (HASHEMIFARD *et al.*, 2020).

Uma das principais aplicações dos PSM está na separação do ar em que oxigênio e o nitrogênio concentrados são produzidos visto que são os dois componentes majoritários (21 e 79%, respectivamente). A baixa seletividade dos sistemas com membranas pode ser considerada como o principal problema presente (HASHEMIFARD *et al.*, 2020). Nos sistemas de membranas de um único estágio existem duas correntes de saída, o oxigênio e outros gases mais leves permeiam mais fácil a membrana formando a corrente do permeado (rica em oxigênio) e a corrente rica em nitrogênio é composta pela parte retida da membrana (BERNARDO, P.; DRIOLI; GOLEMME, 2009). O nitrogênio é muito utilizado na indústria química (produção de fertilizantes, polímeros, recuperação de solventes, produção de explosivos etc.), alimentícia, metalúrgica, em tratamentos térmicos, na manufatura de eletrônicos e até no processo de tratamento de petróleo e gás natural (HARDENBURGER; ENNIS, 2005).

No que concerne o oxigênio, a sua demanda se tornou considerável da década de 50 em que novos processos para a produção de ligas metálicas começaram a ser utilizados (processo *Linz-Donawitz* ao invés de fornos *Bessemer*). Atualmente, o oxigênio é utilizado como suprimento para indústrias químicas e largamente aplicados na área hospitalar. O uso das membranas é mais adequado para aplicações que demandam baixa pureza de oxigênio. Para a obtenção de um produto com fração molar de 50% de O_2, um módulo otimizado com membranas poliméricas com uma seletividade maior que 8 compete facilmente com os outros processos de separação. Em outro contexto, quando se deseja uma fração molar de aproximadamente 92%, a seletividade da membrana polimérica precisa ser de no mínimo 100. Acima desse limite de pureza, o consumo de energia dos processos que utilizam membranas se torna relativamente maior que os processos convencionais e, portanto, não são adequados (ISMAIL; KHULBE; MATSUURA, 2015).

O objetivo inicial das membranas de SG foi direcionado para a separação do hidrogênio de outros gases. As membranas “PRISM” comercializadas pela Monsanto em 1980 tinham a finalidade de recuperar H_2 de um gás de purga numa indústria de amônia ((FANE; WANG; JIA, 2011; TABE-MOHAMMADI, 1999). O H_2 pode ser produzido em refinarias a partir da reação de reforma a vapor, em que a água reage com metano formando gás de síntese (hidrogênio e monóxido de carbono); pode ser recuperado de diversos processos; e possuem inúmeras aplicações industriais (HASHEMIFARD *et al.*, 2020). A demanda por esse gás tende a aumentar cada vez mais ao longo dos anos por se constituir como matéria-prima de células de

combustível e pela sua necessidade em setores importantes como no refino de petróleo. Existem diversos tipos de membranas que são altamente seletivas ao hidrogênio atualmente. Nesse contexto em sua maioria são membranas poliméricas, porém o interesse pelas membranas inorgânicas tem crescido devido a sua estabilidade térmica e química (ISMAIL; KHULBE; MATSUURA, 2015). Mivechian e Pakizeh (2013) simularam e realizaram uma análise econômica do processo de recuperação de hidrogênio em gases de refinaria pelos processos PSA, por absorção e por membranas. Os resultados mostraram que os PSM levam vantagem em relação aos outros processos contanto que um alto nível de pureza não seja requerido.

As membranas também são aplicadas para a remoção/captura de gases ácidos (como CO_2 e H_2S) em certas correntes. O gás natural bruto possui uma certa quantidade de sulfeto de hidrogênio e dióxido de carbono que precisam ser retirados, pois além de danificarem equipamentos, faz-se necessário padronizar o gás natural para sua comercialização. Além disso, o CO_2 precisa ser separado do H_2 no gás de síntese e capturado de gases de combustão provenientes de fontes fósseis. A maioria das membranas eficientes para essa aplicação são poliméricas devido à sua facilidade de fabricação e a sua melhor adequação a baixas temperaturas (ISMAIL; KHULBE; MATSUURA, 2015). As membranas de SG também são aplicadas em várias outras aplicações como na desidratação de correntes gasosas, separação de componentes voláteis de N_2, na recuperação de GLP (Gás Liquefeito de Petróleo), entre outros. Uma breve descrição dos sucessos comerciais dos PSM pode ser vista na Tabela 1.

Tabela 1 – Sucessos comerciais dos PSM.

n.°	Aplicação
1	Enriquecimento de nitrogênio em correntes gasosas
2	Recuperação de hidrogênio do gás de purga de amônia
3	Retirada de CO_2 de biogás e gás natural
4	Recuperação de monômeros de vasos de armazenamento e reciclagem
5	Separação de gás nitrogênio

Fonte: Adaptado de Ismail, Khulbe e Matsuura (2015, p. 261)

2.2.1 Separação de CH_4

O gás metano pode ser liberado de diversas fontes como em pântanos, plantações de arroz, fermentações entéricas, em estações de tratamento de esgoto, aterros sanitários, mineração de carvão, sistemas de gás natural, indústria do petróleo e em processos de combustão. Ele é o segundo gás que mais contribui para o efeito estufa e o aquecimento global

devido a sua alta retenção de calor, podendo ter um impacto 25 a 32 vezes maior que o do CO_2. Em torno de 25% das suas emissões possuem origens antropogênicas advindas da exploração de combustíveis fósseis (carvão, petróleo e gás natural). Em contrapartida, em comparação com a outras fontes fósseis, o CH_4 se mostra como uma opção de fonte de energia mais limpa (fonte de baixo carbono), pois a combustão de 1 mol de metano libera apenas 1 mol de CO_2 (KALAKECH *et al.*, 2022; WANG, Qi *et al.*, 2022).

O metano corresponde ao principal componente do gás natural e em geral está presente com mais de 85% em volume. A composição do gás natural depende muito da fonte ao qual é extraído, mas é formado também por outros diversos gases como o C_2H_6, CO_2, H_2S, H_2 e N_2. Para a sua comercialização faz-se necessário o seu tratamento e a separação de CO_2/CH_4 se mostra como uma importante etapa (HASHEMIFARD *et al.*, 2020). Além disso, como mencionado, gases ácidos (como o CO_2 e o H_2S) compõem o gás natural e eles podem gerar compostos que corroem e danificam equipamentos. Dessa forma, a retirada desses componentes se mostra essencial para diminuir a frequência de manutenção de equipamentos e para a obtenção de um produto com mais qualidade. O uso de membranas para a purificação de gás natural é relatado desde 1994 e vários avanços em relação a resistência desses materiais frente às condições hostis do tratamento desse recurso foram alcançados. As membranas poliméricas, as mais comumente utilizadas, sofrem de "plastificação" e de redução de seletividade quando submetidas a altas pressões. Atualmente, a maioria dos sistemas de membranas utilizados para a purificação de gás natural são compostos por apenas um estágio de separação, produzindo assim um produto de baixa pureza e ocasionando em perdas de CH_4. A adição de outro estágio de separação por membrana combinado com outro tipo de processo (absorção, por exemplo), formando um processo híbrido, aumentaria a eficiência de remoção do CO_2 (CARDOSO *et al.*, 2022).

Com objetivo de atenuar os efeitos do aquecimento global e garantir uma cadeia de suprimento de energia mais híbrida, o biogás vem se mostrando como uma fonte energia que tende a crescer bastante no decorrer dos anos. O biogás pode ser produzido em reatores de fermentação (anaeróbica ou aeróbica) podendo ter diversas matérias-primas (biomassas) como substrato. A biomassa é uma fonte de energia renovável, abundante e disponível que tem o potencial de substituir os combustíveis fósseis no contexto de uma economia circular e sustentável (HOSSEINI *et al.*, 2023; KALAKECH *et al.*, 2022). Resíduos agrícolas (palha, bagaço de cana, folhas de plantas etc.), lodo, lixos orgânicos, resíduos de matadouro e esterco são exemplos de matérias orgânicas que podem ser utilizadas nesse processo (WANG, Quanliang *et al.*, 2023). O biogás bruto possui 50-70% em volume de biometano e 30-40% em

volume de CO_2. Assim como no gás natural, o CO_2 se encontra como o segundo componente majoritário e precisa ser retirado para prevenir corrosão em dutos e aumentar a sua qualidade (LI, Guoqiang *et al.*, 2021).

Vários outros processos podem ser utilizados para purificar o biogás incluindo absorção (aminas) e adsorção (PSA, *Temperature Swing Adsorption*, *Vacuum Pressure Swing Adsorptio*n), porém, quando comparadas com os PSM, essas tecnologias exigem mais energia, maior custo de capital e um contínuo fornecimento de matéria-prima. As membranas despontam como uma tecnologia atrativa por terem aplicabilidade em qualquer fase no processo de fabricação do biogás, serem modulares, mais ecológicos, exigirem pouco consumo de energia e possuírem alta eficiência e performance. Na perspectiva de materiais, a maioria das pesquisas relacionadas ao tema estudam a aplicação de membranas poliméricas e MMMs para a purificação de biogás bruto focando na retirada de CO_2 e enxofre. (HOSSEINI *et al.*, 2023). Devido à alta eficiência para essa aplicação utilizando membranas poliméricas, atualmente há certa de 300 sistemas estabelecidos no mundo para a fabricação de biogás com alta pureza (VALAPPIL; GHASEM; AL-MARZOUQI, 2021). Ainda assim, há muito o que se estudar não apenas no âmbito da fabricação de materiais mais eficientes, mas também no desenvolvimento de módulos, configurações de processo e na análise de condições de operação mais otimizadas para melhorar a aplicação das membranas de SG nesse contexto.

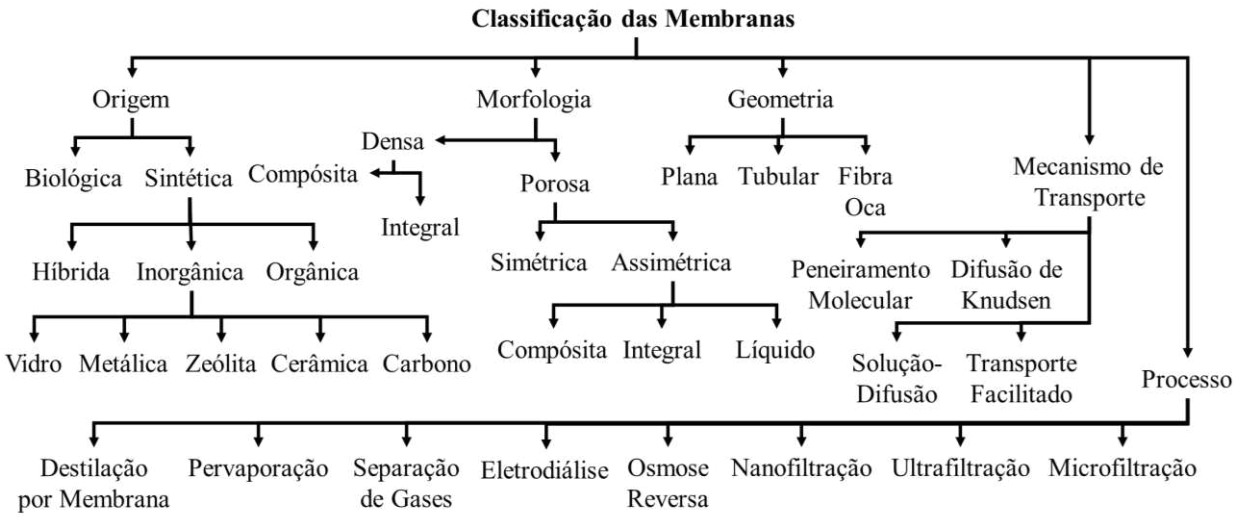
2.3 Classificação das membranas

Uma membrana pode ser definida como uma barreira que possui uma permeabilidade seletiva, que permite a passagem de algumas espécies enquanto retém a de outras. Como é mostrado na Figura 2, ela pode ser classificada quanto a sua origem, morfologia, geometria, mecanismo de transporte predominante e seu processo de aplicação. As membranas podem ser constituídas por inúmeros materiais, de origens naturais (biológicas) ou sintéticas. Os materiais sintéticos são os únicos de interesse comercial e podem ser classificados entre orgânicos e inorgânicos, sendo que algumas membranas são híbridas, isto é, quando esses dois tipos de materiais estão presentes (HABERT; BORGES; NOBREGA, 2006; SADRZADEH; MOHAMMADI, 2019).

Os polímeros são os materiais mais comuns para a fabricação das membranas e são os mais aplicados pois, atualmente, apresentam o menor custo de produção, melhor estabilidade mecânica, variabilidade de módulos/configurações e ótima performance. Entretanto, elas possuem baixa estabilidade química e maiores limitações quanto a relação de *trade-off* entre a

permeabilidade e seletividade. Além disso, essas membranas possuem um difícil ajuste de tamanho de poro quando comparadas com aquelas fabricadas por materiais inorgânicos em que essa variável é facilmente controlada. As membranas inorgânicas em geral possuem melhor estabilidade térmica e química, podem ser operadas em condições mais hostis e tendem a superar o limite superior de Robeson (2008); porém possuem baixa seletividade, são mais caras e mais frágeis, dificultando o seu escalonamento industrial (HASHEMIFARD *et al.*, 2020; SADRZADEH; MOHAMMADI, 2019).

Figura 2 – Diagrama de classificação das membranas.



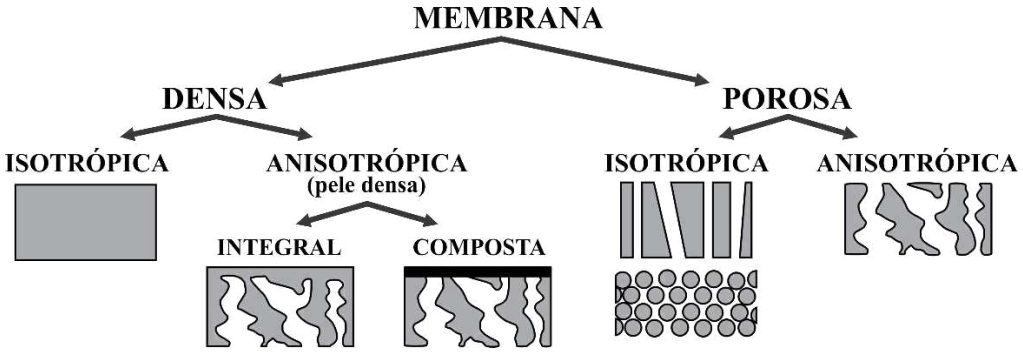
Fonte: Adaptado de Sadrzadeh e Mohammadi (2019).

As membranas compósitas (também chamadas de Membranas de Matriz Mista) são fabricadas de forma que possuem características mistas, tanto dos materiais inorgânicos quanto dos poliméricos. Em geral, possuem boa estabilidade térmica e mecânica; conseguem superar os limites superiores de Robeson (2008); e tendem a ser mais baratas que as membranas inorgânicas. Essas características melhoradas estão diretamente interligadas com interação entre os diferentes tipos de materiais presentes (entre os *fillers* e a matriz). A liberdade no projeto dessas membranas vem atraindo a atenção de vários pesquisadores de forma que elas tendem a substituir as membranas poliméricas no protagonismo comercial futuramente (ISMAIL; KHULBE; MATSUURA, 2015; SADRZADEH; REZAKAZEMI; MOHAMMADI, 2018).

Quanto a morfologia, as membranas podem ser classificadas como densas e porosas, característica que dependerá do mecanismo de transporte predominante. Quando o transporte dos componentes através da membrana envolver a sua dissolução e posterior difusão

(mecanismo solução-difusão), ela é considerada densa. Além disso, ela ainda pode ser dividida em duas subcategorias: isotrópica (simétrica) ou anisotrópica (assimétrica). Se houver uma variação de densidade ao longo da seção transversal da membrana, ela é considerada anisotrópica. Além disso, membranas anisotrópicas podem ser híbridas e são as mais utilizadas em processos de SG (HABERT; BORGES; NOBREGA, 2006; ISMAIL; KHULBE; MATSUURA, 2015). Na Figura 3 é ilustrada a classificação no âmbito morfológico das membranas.

Figura 3 – Diagrama de classificação das membranas acordo com a morfologia e representação das suas respectivas seções transversais.

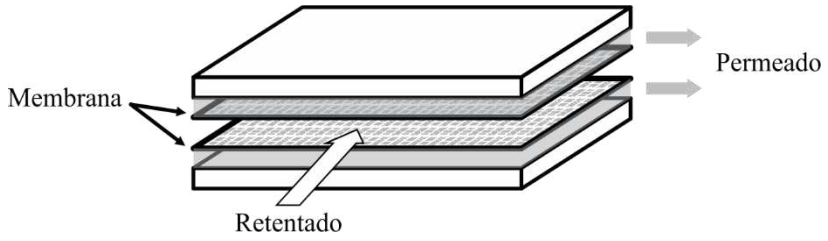


Fonte: Adaptado de Habert, Borges e Nobrega (2006)

Nos sistemas industriais as membranas são organizadas em módulos que podem possuir diferentes configurações dependendo da sua geometria e da posição que elas se encontram em relação aos fluxos de alimentação e do permeado (BERK, 2009). O formato (geometria) ao qual as membranas são "empacotadas" podem ser divididas em três categorias: plana, tubular e fibra oca. A escolha de um formato em específico dependerá de fatores como a natureza da membrana, sua estabilidade mecânica, reprodutibilidade de uma determinada estrutura, natureza do gás permeante e fatores econômicos. Membranas planas podem ser encontradas em módulos de Placa-e-Quadro e de Espiral. O projeto da configuração de Placa-e-Quadro, que pode ser representado pela Figura 4, é semelhante à de um filtro prensa industrial e pode ser circular ou quadrada, arranjada verticalmente ou horizontalmente. Esses módulos não suportam altas pressões e possuem uma baixa relação entre a área de permeação e o volume do módulo. Nas configurações Espiral, que podem ser representadas pela Figura 5, as membranas se encontram "envelopadas" e enroladas em torno de um tubo central onde se coleta o permeado. Em cada envelope a membrana é colocada entre dois espaçadores em que um escoará o fluxo de alimentação e em outro o fluxo de permeado (vedado em três lados para direcionar o

escoamento). A razão entre a área superficial e o volume é alta (BERK, 2009; HABERT; BORGES; NOBREGA, 2006; HASHEMIFARD *et al.*, 2020).

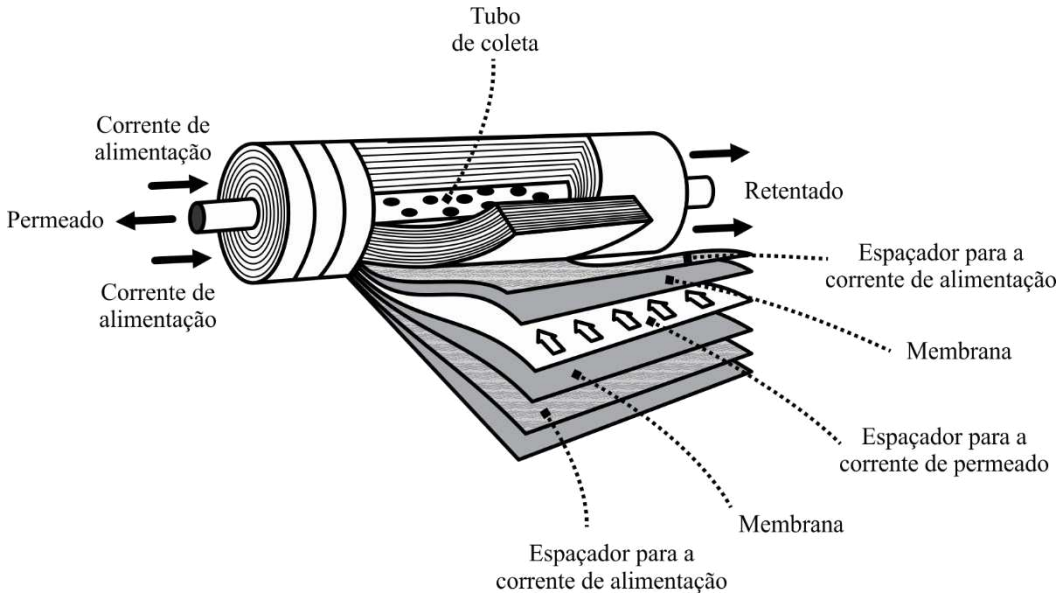
Figura 4 – Representação de uma configuração plana de uma membrana no módulo Placa-e-Quadro.



Fonte: Adaptado de Berk (2009).

Os módulos de configuração tubulares aparentam um trocador de calor casco e tubo. Geralmente a fina camada seletiva é depositada no interior de tubos porosos (suportes). Como pode ser visto na Figura 6, esses tubos são organizados paralelamente uns com os outros e o fluxo do permeado segue o sentido de dentro para fora. Devido ao seu diâmetro relativamente maior, essa configuração permite altas vazões de alimentação e são adequadas quando há sólidos em suspensão ou muito concentrados. A razão entre a área superficial e o volume é baixa (BERK, 2009; CUI; JIANG; FIELD, 2010).

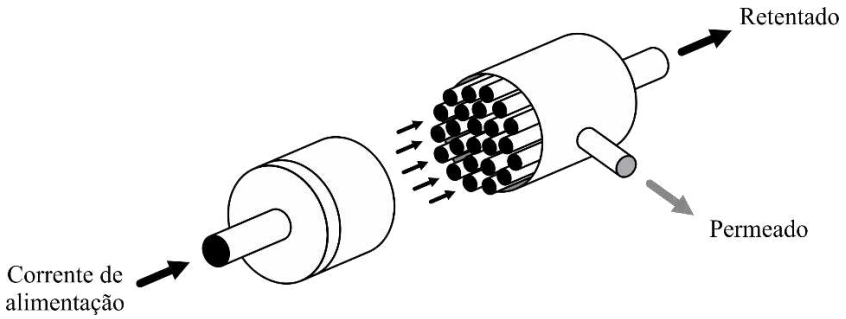
Figura 5 – Representação de uma configuração plana de uma membrana no módulo Espiral.



Fonte: Adaptado de Ismail, Khulbe e Matsuura (2015).

As configurações de fibra oca são similares com as tubulares, exceto pelo fato de que os diâmetros dos "tubos" são bem mais pequenos, em torno de 0,2 e 3 mm. As fibras ocas possuem uma boa estabilidade mecânica e, portanto, não necessitam de suporte. São resistentes a altas pressões e são um dos módulos mais econômicos energeticamente. A direção do fluxo de permeação pode variar de dentro para fora ou o oposto. Essa configuração possui a maior razão entre a área superficial e o volume comparado com as outras. Em contrapartida, são suscetíveis a incrustações e entupimentos (principalmente quando o operadas com o fluxo de dentro para fora), limitando seu uso para fluidos "limpos" e com baixa viscosidade (BERK, 2009; CUI; JIANG; FIELD, 2010).

Figura 6 – Representação de uma configuração tubular ou de fibra oca de uma membrana.



Fonte: Adaptado de Ismail, Khulbe e Matsuura (2015).

2.3.1 Mecanismos de transporte dos PSM para SG

Os PSM podem utilizar o gradiente de potencial químico e o potencial elétrico como força motriz da permeação. Nos processos de separação de gases a diferença de pressão entre os lados da membrana (uma forma de gradiente de potencial químico) é utilizada como força motriz. Dessa forma, a diferença de concentração causada entre as duas faces da membrana induz a difusão dos gases (HABERT; BORGES; NOBREGA, 2006; ISMAIL; KHULBE; MATSUURA, 2015). A taxa de transporte de gás pode ser medida pela permeabilidade (P) ou pela permeância (Q). No contexto dos gases, a permeabilidade é normalmente expressa em *barrer* (1 barrer $=10^{-10}\cdot cm^3_{(STP)}\cdot cm\cdot cm^{-2}\cdot s^{-1}\cdot cmHg^{-1}$) e pode ser definida pela Equação (2.1):

$$P = \frac{J.t}{\Delta p} \tag{2.1}$$

Em que J corresponde ao fluxo de gás (em $cm^3_{(STP)} \cdot cm^{-2} \cdot s^{-1}$) no estado estacionário, t à espessura do filme (em cm) e Δp à diferença de pressão gerada entre os lados da membrana (em cmHg). Pressupondo que a membrana seja isotrópica, a Equação (2.1) pode ser obtida tendo como ponto de partida a primeira e segunda Lei de Fick. Além disso, é considerado a Lei de Henry para descrever o fluxo em função da pressão ao invés da concentração dos componentes. Dessa forma, a permeabilidade também pode ser definida pela Equação (2.2) como o produto entre o coeficiente de difusão ($\mathcal{D}$) e o coeficiente de solubilidade ($\mathcal{S}$) (ISMAIL; KHULBE; MATSUURA, 2015).

$$P = \mathcal{D}.\mathcal{S} \tag{2.2}$$

Vale salientar que essas Equações – (2.1) e (2.2) – foram obtidas partindo do pressuposto que a difusividade independe da concentração das espécies (contextualmente medida em pressão parcial), da espessura do filme e do tempo; circunstâncias adequadas para representar um filme denso cujas interações gás-material são fracas. O coeficiente de permeabilidade, como também pode ser chamado, é uma característica do material e que na prática depende da espessura e da "história" da membrana (ISMAIL; KHULBE; MATSUURA, 2015). A permeância (Q) é ser definida pela Equação (2.3):

$$Q = \frac{P}{t} \tag{2.3}$$

Normalmente, Q é expressa em $mol \cdot m^{-2} \cdot s^{-1} \cdot Pa^{-1}$ ou em GPU (1 GPU = $10^{-6} \cdot cm^3_{(STP)} \cdot cm^{-2} \cdot s^{-1} \cdot cmHg^{-1}$) e indica a quantidade de permeado que atravessa uma determinada área de membrana, a um dado tempo e diferença de pressão. A permeância é preferivelmente calculada no contexto em que as membranas são muito finas ou são consideradas assimétricas. No caso de membranas simétricas (homogêneas) ou quando a espessura é bem definida (normalmente em filmes densos), a permeabilidade é preferivelmente calculada (DU *et al.*, 2012; ISMAIL; KHULBE; MATSUURA, 2015). Em procedimentos experimentais, devido a diversidade morfológica e de mecanismos de transporte das membranas, outras relações (empíricas ou não) diferentes da Equação (2.1) são utilizadas para medir a taxa transporte dos gases. A Tabela 2 exibe os fatores de conversão entre algumas unidades dessas variáveis.

A seletividade ideal (α) ou fator de separação é outra característica chave que determinará a eficiência de separação. Ela retrata a capacidade seletiva da membrana e pode ser definida pela Equação (2.4) como a razão entre as taxas de permeação de dois gases puros.

$$\alpha_{AB} = \frac{P_A}{P_B} = \frac{Q_A}{Q_B} \tag{2.4}$$

Em que α_{AB} é denominado como a seletividade do gás permeante A em relação ao gás permeante B. Por convenção, esse valor precisa ser sempre maior que 1, isto é, a espécie mais veloz sempre estará no numerador da fração (KAMBLE; PATEL; MURTHY, 2021).

Tabela 2 – Fatores de conversão entre as unidades de permeabilidade e de permeância.

Dado um valor de P ou Q com a seguinte unidade		Multiplicar pelos fatores abaixo para converter às variáveis com as unidades correspondentes			
		P		Q	
		barrer	$\frac{mol \cdot m}{m^2 \cdot s \cdot Pa}$	GPU	$\frac{mol}{m^2 \cdot s \cdot Pa}$
P	*barrer*	1	3.34641×10^{-16}	$10^{-4}/t$	$(3.34641\times10^{-14})/t$
	$\frac{mol \cdot m}{m^2 \cdot s \cdot Pa}$	2.98828×10^{15}	1	$(2.98828\times10^{11})/t$	$10^{2}/t$
Q	GPU	$10^{4}\times t$	$3.34641\times10^{-12} \times t$	1	3.34641×10^{-10}
	$\frac{mol}{m^2 \cdot s \cdot Pa}$	$2.98828\times10^{13} \times t$	$10^{-2}\times t$	2.98828×10^{9}	1

P=Permeabilidade. Q=Permeância. t = Valor da espessura da membrana em cm. Fonte: Elaborada pelo autor.

Dessa forma, ocorrem diferentes mecanismos de transporte que vão reger a taxa de permeação das espécies ao longo das membranas. A característica morfológica desses materiais, que pode ser dividida entre densa ou porosa, é considerada como o principal fator para a determinação do tipo de fenômeno predominante. No contexto dos filmes porosos, a difusão das moléculas se mostra como a etapa limitante no processo de permeação. Existem 4 tipos de mecanismos de transporte propostos: difusão de Knudsen, difusão superficial, condensação capilar e peneiramento molecular.

A difusão de Knudsen irá ocorrer quando o percurso livre médio das moléculas é maior que o raio do poro da membrana. Diferente do fluxo convectivo normal (ou fluxo de *Poiseuille*)

em que não há separação, nesse mecanismo as moléculas de gás possuem mais interação com as paredes do poro do que entre elas (SADRZADEH; REZAKAZEMI; MOHAMMADI, 2018). Nesse caso a seletividade ideal de separação pode ser estimada como a raiz quadrada da razão entre as massas molares dos gases (ISMAIL; KHULBE; MATSUURA, 2015).

A difusão superficial ocorre quando as moléculas de gás são adsorvidas nas paredes dos poros e se deslocam ao longo da superfície. Esse fenômeno intensifica a permeação daqueles gases que possuem boa interação com o material e consequentemente aumenta a seletividade de separação. Além disso, o filme criado pela adsorção diminui o tamanho efetivo dos poros aumentando assim a predominância de fenômenos paralelos como a difusão de *Knudsen* (ISMAIL; KHULBE; MATSUURA, 2015; SATO; NAGAI, 2020).

A condensação capilar ocorre quando os poros são completamente preenchidos por uma fase condensada das moléculas permeantes. Dessa forma, apenas moléculas solúveis a essa fase condensada são capazes de permear o poro. Assim como a difusão superficial, a condensação capilar irá depender fortemente da capacidade adsortiva do material e, portanto, seus efeitos positivos relativos à seletividade e ao fluxo vão depender de variáveis como a temperatura, tamanho dos poros e composição de alimentação. Por fim, a permeação dos gases no peneiramento molecular dependerá apenas dos seus respectivos tamanhos (diâmetro cinético) e da dimensão dos poros do material. Esse fenômeno ocorre quando os poros da membrana são tão pequenos quanto os gases permeantes de forma que eles se comportam literalmente como uma peneira. (ISMAIL; KHULBE; MATSUURA, 2015)

O mecanismo de permeação de gases ao longo de membranas densas (majoritariamente poliméricas) conhecido como solução-difusão foi proposto por Thomas Graham em 1866 (GRAHAM, 1866) e pode ser dividida em 3 etapas: (i) sorção da espécie permeante na superfície, (ii) Difusão ao longo da membrana e (iii) dessorção da espécie permeante no outro lado da membrana. O fluxo de permeação pode ser representado pela Equação (2.1). (SADRZADEH; REZAKAZEMI; MOHAMMADI, 2018). A difusividade é fortemente dependente do volume livre entre as cadeias poliméricas e inversamente proporcional ao tamanho das moléculas permeantes. De outra forma, a sorção/dessorção sofrerá influência das condições de operação, da condensabilidade do gás (que é diretamente proporcional ao tamanho do gás) e principalmente das interações intermoleculares entre o polímero e os gases. Existe uma relação de troca entre os coeficientes de difusão e sorção dos gases cuja separação dependerá de suas respectivas magnitudes (LI, Guoqiang *et al.*, 2021; SADRZADEH; MOHAMMADI, 2019).

No âmbito microscópico, devido aos pequenos espaços gerados pelas cadeias poliméricas, existe uma semelhança entre o fenômeno de solução-difusão e o de peneiramento molecular que ocorre nos materiais inorgânicos. Apesar de induzir um peneiramento molecular, as cadeias poliméricas nos materiais densos formam poros mais flexíveis e termicamente ajustáveis de maneira que suas posições se tornam dinâmicas e não são fixas como no peneiramento molecular dos materiais porosos (ISMAIL; KHULBE; MATSUURA, 2015).

Do ponto de vista físico-químico, há apenas uma fase nos sistemas de permeação das membranas densas, enquanto nas membranas porosas existe no mínimo duas fases (HABERT; BORGES; NOBREGA, 2006). Vale salientar que em alguns casos as moléculas vão se mover através das membranas por mais de um mecanismo de transporte e que não necessariamente apenas um será predominante. Na prática, o fenômeno de solução-difusão tem sido utilizado como uma base genérica para modelar tanto membranas porosas quanto densas (ISMAIL; KHULBE; MATSUURA, 2015). A maioria das membranas poliméricas possuem uma camada seletiva densa, em que o mecanismo de transporte solução-difusão será predominante (HAYEK *et al.*, 2020; WANG, Ming *et al.*, 2017). Entretanto, o mecanismo de transporte nas membranas inorgânicas, que possuem uma estrutura porosa, será dependente da distribuição de poros do material. De forma genérica, para tamanhos de poro menores que 0,6 nm, ocorrerá peneiramento molecular; quando os poros forem menores que 2 nm, ocorrerá difusão de superfície; e quando os poros estiverem entre 2 e 50 nm, ocorrerá difusão de Knudsen (CARDOSO *et al.*, 2022).

3 REVISÃO BIBLIOGRÁFICA: APRENDIZADO DE MÁQUINA

3.1 Fundamentos de AM

Originado de um algoritmo que ensinava um computador a jogar xadrez em 1959, o Aprendizado de Máquina é um subcampo da Inteligência Artificial que se desenvolveu para uma técnica especializada na previsão a partir de vários exemplos observados (WANG, Jing *et al.*, 2023). Precisamente, segundo Tom Mitchell (1997, p.2), "pode-se dizer que um programa aprende com a experiência E, a respeito de alguma classe de tarefas T medida pelo desempenho P, se seu desempenho nessas tarefas T, medido por P, melhora com a experiência". De outro modo, pode-se afirmar que o AM é a capacidade que um programa possui de melhorar o desempenho na realização de alguma tarefa por meio da experiência. Assim, esses algoritmos são construídos de forma a induzir uma função ou hipótese para esclarecer alguma resposta utilizando dados que representam registros sobre o problema a ser resolvido (FACELI *et al.*, 2011). No presente contexto, a hipótese criada pode ser definida como o modelo gerado pelo algoritmo de AM. Dessa forma, aprender corresponde a adquirir conceitos gerais considerando apenas alguns exemplos sobre o problema (MITCHELL, Tom, 1997).

Esses exemplos podem ser chamados de registros, objetos, padrões ou eventos e são armazenados em um Banco de Dados. Eles são constituídos por atributos que serão utilizados para a indução de hipóteses. Para algumas tarefas (modelos supervisionados), um desses atributos (ou mais) é chamado de atributo alvo (ou variável dependente), cujo valor precisa ser estimado utilizando os valores dos outros atributos que serão definidos como atributos de entrada (descritores, variáveis independentes ou variáveis preditoras). Esses atributos podem ser classificados como quantitativos e qualitativos. Os atributos quantitativos carregam valores inteiros (escala discreta) ou reais (escala contínua). Quando são reais, podem assumir um número infinito de valores e normalmente possuem uma unidade de medida associada, algo de suma importância que deve ser conhecida para interpretação dos dados adquiridos. Os atributos qualitativos (também chamados de categóricos) possuem valores associados às classes ou categorias. Eles podem ser ordinais ou nominais, quando existe ou não existe, respectivamente, uma relação de ordem entre as classes (FACELI *et al.*, 2011).

Vale ressaltar que o AM desfruta da teoria estatística ao construir modelos matemáticos, visto que a seu papel principal é fazer inferências a partir de uma amostra adequada. Dessa forma, no aprendizado indutivo, é fundamental que a hipótese construída seja a melhor para representar as instâncias ausentes no conjunto de treinamento (MOHRI; ROSTAMIZADEH;

TALWALKAR, 2018). Mais precisamente, é assumido essencialmente que "qualquer hipótese encontrada para aproximar bem uma função alvo em um conjunto suficientemente grande de exemplos de treinamento também aproximará bem a função alvo em relação a outros exemplos não observados" (MITCHELL, 1997, p.23).

Esse processo de inferência é quantificado ao medir a capacidade preditiva do modelo para um novo conjunto de dados. Assim, é comum a divisão dos registros disponíveis em duas partes: treino e teste. O conjunto de treino será utilizado para a construção do modelo e o conjunto de teste para analisar a sua capacidade de generalização. É dito que o modelo está sobreajustado (ocorre *overfitting*) quando sua performance no conjunto de treino é muito superior do que a vista no conjunto de teste, indicando uma capacidade limitada de extrapolação. É normal que isso ocorra quando se utiliza um modelo muito complexo em relação a quantidade de registros disponíveis (que podem ser pouco representativos) ou ao fenômeno que está sendo tratado, de forma que até os ruídos presentes são levados em consideração. É essencial que os modelos funcionem bem mesmo com dados imperfeitos, isto é, com ruídos, inconsistências, valores ausentes ou redundantes (FACELI *et al.*, 2011; IZBICKI; SANTOS, 2020).

Por outro lado, quando o algoritmo de AM induz hipóteses que não são adequadas para explicar bem os padrões presentes em nenhum conjunto de dados (treino ou teste), pode-se afirmar que o modelo está sub-ajustado (ocorre *underfitting*). Normalmente isso ocorre pois o modelo empregado é muito simples ou está mal ajustado. Além disso, os atributos descritores podem não ser tão relevantes para o fenômeno estudado (IZBICKI; SANTOS, 2020). Quando o conjunto de dados é relativamente pequeno, ao dividi-lo, resta uma quantidade insuficiente de registros para ser utilizada no treinamento. Dessa forma, recomenda-se a utilização do método de Validação Cruzada (que será explicado posteriormente na Seção 4.6) para medir a capacidade de generalização de um modelo.

Vale salientar que existem diversos algoritmos de AM e, segundo Faceli et al. (2011), cada um deles terá dois vieses indutivos que determinará a representação e a busca da hipótese induzida. O viés de representação ou a forma como o modelo é apresentado pode restringir o conjunto de hipóteses que podem ser inferidas pelo algoritmo. O modelo Florestas Aleatórias é constituído por uma série de árvores de decisão que possuem um conjunto de regras em forma de árvore para tomar decisões. O modelo Regressão de Vetores de Suporte (RVS) possui vetores de suporte como critério no processo de previsão e as Redes Neurais Artificiais representam uma hipótese por um conjunto de pesos associados a cada neurônio artificial. Por outro lado, viés de busca é a forma como o algoritmo busca a hipótese que melhor se ajusta aos

dados de treinamento. Em árvore de decisão, busca-se árvores com poucos nós; no RVS, procura-se ajustar uma função de regressão com uma determinada margem de erro; e em Redes Neurais, deseja-se minimizar uma função de custo no algoritmo de retropropagação (FACELI *et al.*, 2011).

No entanto, mesmo que o objetivo primordial da hipótese formada (modelo construído) pelo algoritmo seja obter um bom poder preditivo, a sua interpretabilidade também é importante, pois isso gera segurança na sua utilização e pode trazer percepções sobre o fenômeno ao qual ocorre determinada tarefa (IZBICKI; SANTOS, 2020). A interpretabilidade pode ser definida como o grau em que um ser humano é capaz de compreender as razões subjacentes a uma decisão tomada por um modelo ou de prever consistentemente os seus resultados. Quanto maior a Interpretabilidade de um modelo de AM, mais facilmente alguém irá entender por que certas decisões ou previsões foram feitas. Um modelo é melhor interpretável do que outro quando suas decisões são mais facilmente compreendidas por um ser humano (MILLER, 2019).

Nesse contexto, existem diferentes tipos de tarefas de aprendizado (ou cenários) que serão utilizadas a depender do objetivo do usuário, do tipo dos dados disponíveis, do conjunto de teste empregado para avaliar o algoritmo e, por fim, da ordem e método em que os dados são obtidos (MOHRI; ROSTAMIZADEH; TALWALKAR, 2018). De acordo com esses critérios, normalmente essas tarefas são divididas em preditivas ou descritivas. Em tarefas de previsão, os algoritmos constroem modelos utilizando um banco de dados em que os valores do atributo alvo são conhecidos e são utilizados para treinar e avaliar a capacidade preditiva. Eles são denominados como modelos de aprendizado supervisionado e podem classificar ou realizar regressão, caso o atributo alvo possua rótulos discretos ou contínuos, respectivamente. Em tarefas de descrição, os algoritmos não possuem um atributo alvo e o objetivo é explorar ou descrever os registros de um conjunto de dados. Os modelos gerados podem agrupar os registros com base em suas respectivas similaridades; sumarizá-los para descrevê-los de forma simples e compacta; ou identificar relações ou associações frequentes entre eles (FACELI *et al.*, 2011).

3.2 Aplicações

O desenvolvimento de algoritmos cada vez mais eficazes e eficientes, a capacidade computacional cada vez maior e a digitalização dos processos em geral são fatores que favorecem a expansão do AM a diversos campos de estudo atualmente. Além disso, essa área

é inerentemente multidisciplinar, de modo que suas bases são formadas com conceitos de probabilidade e estatística, teoria da complexidade computacional, teoria do controle, teoria da informação, filosofia, psicologia, neurobiologia e outros campos (MITCHELL, Tom, 1997).

Atualmente, esses algoritmos de aprendizado já são parte integrante do cotidiano das pessoas e são largamente aplicados para a classificação de documentos (detecção de *spam*); no processamento de linguagem natural (análise sintática, reconhecimento de fala e escrita); para o reconhecimento óptico de caracteres; em tarefas de reconhecimento de imagem; na biologia computacional (estruturação de proteínas); em detecções de fraudes (bancos e segurança computacional); em diagnósticos médicos; na orientação de navegação de veículos; na publicidade (recomendação de produtos ou serviços); e em diferentes ramos do entretenimento (MOHRI; ROSTAMIZADEH; TALWALKAR, 2018).

Entretanto, especificamente no campo da engenharia, as técnicas de AM ganharam popularidade por serem capazes de reconhecer padrões que são muito difíceis para as pessoas identificarem. Foi principalmente após a popularização das técnicas de aprendizado profundo (Redes Neurais Artificiais Profundas) por meio de bibliotecas computacionais como o *Keras*, *TensorFlow* ou o *scikit-learn*, que o interesse de pesquisadores sobre o tema cresceu bastante após 10 anos de adormecimento, desde a popularização das Redes Neurais Simples e de técnicas de clusterização (DOBBELAERE *et al.*, 2021).

Existem vários trabalhos referentes a previsão de propriedades químicas quânticas usando AM com o intuito de diminuir o custo computacional associado aos métodos como o *ab initio* (DOBBELAERE *et al.*, 2021). Além disso, modelos de AM (principalmente Redes Neurais) também são utilizadas para auxiliar na modelagem de colunas de destilação; de processos de adsorção e de absorção; no projeto e reatores químicos; e, inclusive, no planejamento de plantas industriais inteiras (no controle, operação e gerenciamento de energia) (YAN; BORHANI; CLOUGH, 2020). Essas técnicas também vêm sendo exploradas de uma forma muito otimista como uma nova ferramenta para desenvolvimento racional de materiais para diversas aplicações de engenharia. A complexidade inerente da procura em um enorme gama de opções cria um cenário propício para a utilização das ferramentas disponibilizadas pelo AM. (MOOSAVI; JABLONKA; SMIT, 2020)

3.2.1 Principais autores e citações do campo de estudo

Alguns pesquisadores já se propuseram a explorar algoritmos de AM ao estudo de Processos de Separação de Gases por Membranas (PSGM) utilizando variáveis de entrada

experimentais ou simuladas. Wessling et al. (1994) utilizaram dados do espectro de infravermelho para ganhar informações da estrutura química de 50 polímeros diferentes e tentar prever a permeabilidade do CO_2. Os autores do estudo usaram redes neurais e, em comparação com os trabalhos atuais, não obtiveram resultados satisfatórios. O tamanho do banco de dados foi pequeno e, devido ao custo computacional, a rede neural construída tinha 24 atributos de entrada e apenas uma camada oculta com 16 neurônios artificiais. Segundo Zhu et al. (2020), essa foi a primeira tentativa de desenvolver um modelo baseado em AM para prever o desempenho de membranas. Hasnaoui, Krea e Roizard (2017) utilizaram Redes Neurais Artificiais (RNA) para prever a permeabilidade do O_2, N_2, CO_2 e do CH_4. Os autores coletaram dados de permeação de 149 polímeros diferentes (122 sobre polímeros vítreos e 27 sobre polímeros plásticos) e definiram como variáveis de entrada 21 descritores do modelo de contribuição de grupo de Yampolskii et al. (1998). Um modelo foi treinado para cada gás e três camadas de neurônios foram suficientes para obtenção de bons resultados com REQM (Raiz do Erro Quadrático Médio) abaixo de 2,84 para o N_2, O_2 e o CH_4. A predição não foi tão precisa apenas para os polímeros que possuíam baixa permeabilidade, fato que segundo os autores foi consequência de imprecisões experimentais. Segundo Wang et al. (2023a), esse foi o primeiro trabalho a apresentar resultados favoráveis a modelos de AM, mesmo utilizando um banco de dados pequeno e apresentando uma baixa generalização para outros polímeros.

Barnett et al. (2020) treinaram um algoritmo de AM (Processo Gaussiano de Regressão - PGR) para projetar membranas poliméricas com propriedades acima do limite superior do gráfico de Robeson (2008). Os autores predisseram a performance de separação (permeabilidade e seletividade) de mais de 11 mil homopolímeros. Eles treinaram seis modelos com conjuntos de treino diferentes (variando de 282 a 523 dados) para cada gás (N_2, O_2, H_2, He, CH_4 e CO_2). A estrutura química dos polímeros foi definida como atributo de entrada para prever o logaritmo da permeabilidade experimental. A seletividade foi calculada a partir dos resultados desses modelos. Uma técnica de impressão digital foi utilizada para representar a estrutura química de cada polímero tratado como variável de entrada. Essa técnica de impressão digital analisa os fragmentos da molécula e suas ligações para, em seguida, aplicar um algoritmo *hash* e construir um código binário que represente computacionalmente a molécula. Desse modo, elabora-se uma representação mais dinâmica, pois permite a extrapolação para novas moléculas sem a necessidade de um retreino e retém informações sobre as conectividades entre as unidades das moléculas. Entre as membranas mais promissoras preditas, duas foram sintetizadas com a finalidade de validar a precisão do modelo e bons resultados foram encontrados. Os dados experimentais e previstos apresentaram uma boa concordância entre si

(dentro da faixa de erros do modelo) e, além disso, as duas membranas excederam o limite superior do gráfico de Robeson.

De maneira parecida, Zhu et al. (2020) também usaram o modelo PGR e um método de impressão digital para os polímeros. Utilizando um banco de dados com 1501 registros e considerando 315 polímeros diferentes, os autores treinaram um modelo de AM abrangendo os gases He, H_2, CO_2, O_2, N_2 e CH_4 [diferente do trabalho de (BARNETT *et al.*, 2020) que treinou um modelo para cada gás]. Para esse único modelo treinado, foi obtido um coeficiente de determinação de 0,99 para o conjunto de treino e de 0,93 para o conjunto de teste.

Yuan et al. (2021) utilizaram alguns modelos de AM (Regressão Linear Bayesiana e Árvores Extremamente Aleatórias) em um algoritmo MICE (em inglês, *Multivariate Imputation by Chained Equations*) para predizer a permeabilidade desconhecida de determinados gases com base em dados de permeabilidade conhecidos de outros gases. Os autores trabalharam com o banco de dados da Sociedade de Membranas da Australásia (em inglês, *Membrane Society of Australasia* – MSA) abrangendo os gases He, H_2, O_2, N_2, CO_2 e CH_4. Juntamente com o algoritmo MICE, ambos os modelos testados demonstraram uma predição aceitável, de modo que possam ser utilizados nos estágios iniciais de medidas experimentais, concedendo um discernimento sobre as membranas que valem a pena investigar em detalhes e, em vista disso, diminuindo a quantidade de testes experimentais a serem realizados.

De forma semelhante ao trabalho de Zhu et al. (2020) e Barnett et al. (2020), Yang et al. (2022) analisaram dois modelos de AM (Florestas Aleatórias e Redes Neurais Profundas) para o projeto de novas membranas poliméricas. Eles empregaram um método de descrição química e um método de impressão digital para descrever 353 polímeros diferentes. Além disso, foi utilizado um conjunto de treino com 778 registros coletados do banco de dados PoLyInfo e MSA, com cada um contendo a permeabilidade de um dos seguintes gases: He, H_2, O_2, N_2, CO_2 e CH_4. Os modelos foram treinados com um conjunto separado para cada gás. Os autores conseguiram prever a estrutura química de mais de nove milhões de polímeros hipotéticos (separados em três banco de dados) e identificaram inúmeros candidatos com performance superior aos limites do gráfico de Roberson. A validação dessas predições foi realizada via simulação por Dinâmica Molecular.

Daglar e Keskin (2022) realizam várias simulações moleculares (*Grand Canonical Monte Carlo* e Dinâmica Molecular) para treinar algoritmos de Aprendizado de Máquina e estimar diferentes propriedades de MOFs. Utilizando dados simulados de 5249 MOFs diferentes para alimentar os modelos (8, no total), os autores previram a adsorção e a difusão

de quatro gases diferentes (He, H_2, CH_4 e N_2) para depois calcular a respectiva permeabilidade e a seletividade. Utilizando 20 atributos de entrada (de propriedades químicas e físicas), eles aplicaram a técnica de Aprendizado de Máquina automatizada TPOT (*Tree-based Pipeline Optimization Tool*) (OLSON *et al.*, 2016) para selecionar o melhor modelo de AM e otimizar os hiperparâmetros. Nessa abordagem, procura-se eficientemente com o mínimo de intervenção humana tratar os atributos de entrada, selecionar e otimizar o melhor modelo de AM entre vários disponíveis, de forma que o usuário seja responsável apenas pela coleta e curadoria do banco de dados. Boas métricas foram alcançadas em todos os modelos e nas análises de importâncias realizadas, as propriedades físicas foram mais relevantes para a predição de ambos os atributos alvo. Bons resultados também foram obtidos após os autores utilizarem os dados estimados no modelo de *Maxwell* (MAXWELL, 1954) para prever a permeabilidade de membranas MMM de MOFs. Além disso, os modelos construídos foram utilizados para prever a adsorção e a difusão de 500 MOFs hipotéticas (que ainda não foram sintetizadas) e seus resultados foram comparados com os obtidos via simulação molecular, mostrando uma precisão adequada.

Existem publicações que fazem uso de algoritmos de AM para tratar sobre temas relacionado às membranas, porém poucos estudos exploram dados experimentais para descrever os materiais em foco. Guan et al. (2022) utilizaram Florestas Aleatórias e um método de transferência de aprendizado para guiar o projeto de MMMs sintetizadas com MOFs (em inglês, Metal Organic Frameworks). O AM foi aplicado com o objetivo de fornecer associações entre a estrutura das MOFs e a performance das membranas (permeabilidade e seletividade) para a captura de CO_2. Um banco de dados com 648 registros (incluindo 36 tipos de MOFs e 41 tipos de polímeros) foi construído contendo informações de permeação tanto de gases puros, quanto de gases em mistura. Dois modelos foram treinados no contexto da separação CO_2/CH_4 utilizando 12 atributos de entrada divididos em quatro classes: propriedades das MOFs, propriedades da matriz polimérica, características da MMM e condições de teste. A permeabilidade do CO_2 e a seletividade CO_2/CH_4 foram definidas como variável alvo para cada um respectivamente. Pela interpretação de gráficos PDP, os autores concluíram que MOFs com um diâmetro de poro maior que 1 nm e uma área superficial BET (*Brunauer*, *Emmett* e *Teller*) com aproximadamente 800 $m^2 \cdot g^{-1}$ seriam mais adequadas para a síntese de uma MMM com boa performance para a captura de CO_2. Para validar, foram sintetizadas duas novas MOFs, ambas com um diâmetro de poro recomendado (~1,2 nm), mas com áreas superficiais diferentes (124 e 718 $m^2 \cdot g^{-1}$) para serem utilizadas na fabricação de MMM. As predições do modelo foram adequadas quando comparadas com os testes experimentais, de modo que a membrana fabricada com a melhor MOFs superou o limite superior de Robeson (2008). O método de

transferência de aprendizado foi aplicado para elaborar um outro modelo focado em predizer a separação no contexto CO_2/N_2, que demonstrou ser também adequado.

Pan et al. (2022) investigaram os modelos Máquina de Vetores de Suporte (MVS) e Regressão Linear Múltipla (RLM) para analisar os múltiplos fatores que podem influenciar a performance de membranas compostas por PMC. Um banco de dados com 399 registros foi construído contendo dados de permeação de 5 gases diferentes (CO_2, CH_4, N_2, O_2 e H_2). Os autores testaram diferentes *kernels* e métodos de regressão utilizando 11 atributos de entrada que foram divididos em 4 classes: estrutura do precursor, condição de carbonatação, estrutura do microcristal de carbono e propriedades do gás permeante. Além disso, definiu-se um interessante índice quantitativo, nomeado de distância característica, para servir como indicativo de performance. Em um dado gráfico de Robeson para um par de gás específico, a distância característica foi definida como a distância perpendicular entre o ponto referente a membrana e a reta do limite superior de Robeson (2008). Desse modo, foi realizada uma análise quantitativa da importância das variáveis de entrada para a predição da permeabilidade e da distância característica. O estudo constatou a partir das análises que os principais fatores que afetam a performance da separação estão mais associados às características microestruturais (que podem ser alterados durante a síntese) do que associados às propriedades do gás permeante. Esses resultados podem auxiliar uma otimização estrutural e no aprimoramento da performance de separação de membranas de PMC.

4 METODOLOGIA

4.1 Panorama geral do estudo

O presente trabalho pode ser dividido em 5 etapas:

i. Construção do banco de dados, em que será feita a coleta dos mais variados registros na literatura e seleção daqueles que serão utilizados na modelagem;
ii. Análise Exploratória do banco de dados, em que será investigado a distribuição dos atributos de entrada e a correlação entre eles;
iii. Modelagem do atributo alvo utilizando os três modelos de AM e investigando as diferentes funções objetivo;
iv. Seleção do melhor modelo e a retirada das variáveis ruído;
v. Cálculo da seletividade utilizando o melhor modelo.

Sendo assim, os procedimentos realizados na etapa *iii*, onde os modelos de AM são treinados, dada uma função objetivo e um conjunto de atributos de entrada, podem ser resumidos em forma de pseudocódigo da seguinte forma:

Algoritmo 1 – Procedimentos realizados para treinar e avaliar os modelos de AM.

Entrada: Banco de dados referente a permeação de gases em membranas.
Saída: Modelo de AM treinado e suas respectivas métricas.

1 Carregar o banco de dados a partir de um arquivo do formato ".csv".
2 Separar os atributos de entrada e o atributo alvo do banco de dados.
3 Codificar os atributos categóricos, se houver.
4 Escalonar todos os atributos via *z-score*.
5 Separar o conjunto de dados em treino (80 %) e teste (20 %).
6 Otimizar o modelo de AM utilizando a biblioteca computacional Optuna:
    i. Definir uma função objetivo para a otimização dos hiperparâmetros;
    ii. Definir os hiperparâmetros a serem otimizados e suas faixas de procura;
    iii. Executar a busca no espaço dos hiperparâmetros e selecionar os valores ótimos.
7 Treinar o modelo de AM com os parâmetros otimizados.
8 Calcular as métricas de desempenho para o conjunto de treino e teste.
9 Interpretar os resultados obtidos via outros métodos computacionais (Importâncias por permutação ou Gráficos de Dependência Parcial).

4.2 Banco de dados

Primeiramente, uma análise exploratória foi realizada na literatura de forma a elucidar os descritores que: i) carregam mais informações sobre o processo de permeação nas membranas; ii) são mais frequentemente fornecidas em publicações acadêmicas; e iii) que caracterizem de forma mais genérica os tipos de membranas as quais o presente trabalho se propôs a investigar. A pesquisa teve como foco trabalhos que tenham estudado a capacidade de permeação do metano, porém, porventura trabalhos relacionados a outros temas (que focam em outros gases) também foram adicionados em menor quantidade (por exemplo, que estudaram a separação de H_2 ou a captura de CO_2). A busca foi realizada pelas plataformas *ScienceDirect* (https://www.sciencedirect.com/) e *Google Scholar* (https://scholar.google.com/). Ela foi concentrada entre os anos de 2009 e 2022 e foi sistemática (não teve uma metodologia probabilística) (BOLFARINE; BUSSAB, 2005), de forma que foram focados os estudos de membranas com os seguintes tipos, na ordem: (i) MMM, (ii) zeólitas, (iii) poliméricas e (iv) MOFs. Ao longo da procura, os trabalhos selecionados foram os que forneceram muitas características sobre as membranas e seus materiais de origem.

Desta forma, um banco de dados, constituído por 1672 registros (ou linhas) referentes a trabalhos experimentais de membranas aplicadas na separação de gases, foi construído com base em 42 referências bibliográficas diferentes. A ferramenta *WebPlotDigitizer v4.5* (ROHATDI, 2021) foi utilizada para a prospecção de dados apresentados em forma gráfica. Como pode ser observado na Tabela 3, para cada registro foram extraídos 29 atributos (que originaram 67 colunas no arquivo do banco de dados) que foram divididos em 4 categorias: características morfológicas da membrana, características de processo, de performance e bibliográficas. Assim, cada registro se refere a uma membrana, caracterizada pelos atributos morfológicos, sendo aplicada em uma determinada circunstância de utilização, caracterizada pelos atributos de processo. Dessa forma, uma mesma membrana pode ser submetida a vários ensaios de permeação diferentes, originando diferentes registros. Os atributos de performance são os únicos que são exclusivos para cada linha e não se repetem. Esse banco de dados foi denominado “Banco de Dados Original”, pois a partir dele, outros foram construídos possuindo

diferentes formatações e informações complementares. O arquivo[1] que contém esse banco de dados está disponível no formato "*xlsx*" para que os interessados possam utilizá-lo livremente.

Tabela 3 – Definição dos atributos do arquivo que contém o Banco de Dados Original.

(continua)

Categoria	Atributo[1]	Descrição do atributo
Morfologia	"*type*" (Tipo)	Classificação da membrana pelo material ao qual ela foi sintetizada.
	"*description*" (Descrição)	Descrição nominal da membrana utilizada pelos autores do artigo de referência.
	"*support_material*" (Material de Suporte)	Material do suporte da membrana, caso ela não for autossuportada.
	"*configuration*" (Configuração)	Configuração da membrana nos ensaios experimentais de permeação.
	"*subtype*" (Subtipo)	Nome específico do material ao qual a membrana é composta.
	"*filler_loading*" (Carga de *Filler*)	Carga de *filler* (normalmente zeólitas) adicionada no polímero. Vale apenas para MMM (em %m/m).
	"*min_thickness*" e "*max_thickness*" (Espessura Mínima e Máxima)	Espessura mínima e máxima da membrana (sem considerar o suporte). Nos casos em que os autores não forneceram uma faixa mínima e máxima, as duas variáveis possuem o mesmo valor (em μm).
	"*min_pore_size*" e "*max_pore_size*" (Tamanho Mínimo e Máximo de Poro)	Tamanho mínimo e máximo de poro do material da membrana (sem considerar o suporte). Nos casos em que os autores não forneceram uma faixa mínima e máxima, as duas variáveis possuem o mesmo valor (em nm).
	"*pore_size_type*" (Tipo do Tamanho de Poro)	Método utilizado para a descrição do tamanho de poro.

[1]O nome entre aspas se refere a como são chamadas os atributos/colunas no banco de dados e o nome em parênteses se refere a como o atributo será chamado no presente trabalho. Fonte: Elaborada pelo autor.

[1] https://github.com/tenoriolms/databank_membranes/blob/main/databank_Original_Databank.xlsx

Tabela 3 – Definição dos atributos do arquivo que contém o Banco de Dados Original.

(continuação)

Categoria	Atributo[1]	Descrição do atributo
Morfologia	"*total_pore_volume*" (Volume Total de Poro)	Volume total de poro do material da membrana (em $cm^3 \cdot g^{-1}$).
	"*micropore_volume*" (Volume de Microporos)	Volume de microporos do material da membrana (em $cm^3 \cdot g^{-1}$).
	"*specific_surface_area*" (Área Superficial Específica)	Área superficial específica BET do material da membrana (em $m^2 \cdot g^{-1}$).
	"*aging*" (Idade)	Idade da membrana (em dia).
Processo	"*sufarce_area*" (Área Superficial)	Área superficial da membrana em contato com a corrente de alimentação utilizada nas medições experimentais (cm^2).
	"*temperature*" (Temperatura)	Temperatura da membrana no momento da medição experimental (em K).
	"*feed_pressure*" (Pressão de Alimentação)	Pressão na corrente de alimentação (em Pa).
	"*permeate_pressure*" (Pressão do Permeado)	Pressão na corrente do permeado (em Pa).
	"*delta_pressure*" (Diferença de Pressão)	Diferença entre a pressão na corrente de alimentação e no permeado (em Pa).
	"*feed_flow_rate*" (Vazão de Alimentação)	Vazão da corrente de alimentação ($mL_{(STP)} \cdot min^{-1}$).
	"*sweep_gas*" (Gás de Arraste)	Fórmula química do gás de arraste presente na corrente de permeado.
	"*sweep_gas_flow*" (Vazão do Gás de Arraste)	Vazão do gás de arraste presente na corrente de permeado.
	"*stage_cut*" (*Stage-cut*)	*Stage-cut* máximo permitido durante os procedimentos experimentais. Essa variável é definida como a razão entre a vazão de permeado e a vazão de alimentação (em %).
	"*x_*" (Fração do Gás)	Prefixo para fração molar/volumétrica do gás na corrente de alimentação.

[1]O nome entre aspas se refere a como são chamados os atributos/colunas no banco de dados e o nome em parênteses se refere a como o atributo será chamado no presente trabalho. Fonte: Elaborada pelo autor.

Tabela 3 – Definição dos atributos do arquivo que contém o Banco de Dados Original.

(conclusão)

Categoria	Atributo[1]	Descrição do atributo
Performance	"*Py_*"	Prefixo (no arquivo) para permeabilidade do gás de alimentação (em barrer).
	"*Pe_*"	Prefixo (no arquivo) para permeância do gás de alimentação (em $mol \cdot sqm^{-1} \cdot s^{-1} \cdot Pa^{-1}$).
Bibliografia	"*provided_data_type*" (Tipo de Performance Fornecida)	Especificação de qual variável de performance da membrana é originalmente apresentada no artigo: Permeabilidade ou Permeância.
	"*in_reference_data_location*"	Figura ou tabela da referência onde está localizado o registro de performance da membrana.
	"*reference*"	Abreviação da referência no formato: Abreviação do(s) nome(s) do(s) Autor(es) e ano de publicação. Caso existirem apenas 2 autores, suas duas respectivas abreviações são exibidas. Caso existirem mais de 2 autores, a abreviação do primeiro autor é seguida por "et al".
	"*url*"	Endereço virtual da página onde se encontra a referência.

[1]O nome entre aspas se refere a como são chamados os atributos/colunas no banco de dados e o nome em parênteses se refere a como o atributo será chamado no presente trabalho. Fonte: Elaborada pelo autor.

A classificação dos registros do atributo Tipo foi baseada no trabalho de Ismail, Khulbe e Matsuura (2015). Os autores classificam os materiais das membranas em: poliméricos, cerâmicos, de sílica, zeolíticos, MOFs, MMM (compósitos), metálicos e baseados em carbono (CMS, em inglês, *Carbon Molecular Sieves*). Vale salientar que não existe uma padronização no fornecimento desses atributos e, portanto, alguns deles estarão ausentes em determinadas referências e apresentadas de diferentes formas em outras. 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 considerações foram realizadas no momento da mineração dos dados. A Pressão do Permeado foi considerada nula (vácuo) para aquelas referências que não a especificaram, mas forneceram informações sobre a Pressão de Alimentação. As frações molar

e volumétrica (a depender do que foi fornecido) foram representadas por um mesmo atributo chamado Fração do Gás como pode ser visto na Tabela 3. Essa aproximação é adequada visto que, principalmente para os gases leves, essas frações possuem valores aproximados e as medições sobre a composição das misturas são realizadas em condições de temperatura e pressão amenas. Além disso, algumas referências não especificaram a base (seja molar, mássica ou volumétrica) que utilizaram para a representação da composição da corrente de alimentação.

Vale destacar ainda que existem registros de membranas autossuportadas no banco de dados e nesses casos, o atributo Material de Suporte tem o valor nulo ("NaN"). Além disso, por definição, não foram considerados os suportes para descrição da Espessura Mínima e Máxima das membranas, uma vez que a seletividade do processo é mantida pelo filme denso (ou "pele") e o suporte poroso não oferece resistência significativa ao transporte de massa. Sendo assim, apenas a espessura do filme deve ser considerada como variável de resistência a permeação (HABERT; BORGES; NOBREGA, 2006).

O Tamanho Mínimo e Máximo de Poro das membranas foram descritores de difícil definição visto que sua real distribuição irá depender das técnicas/condições de síntese e de medição empregadas pelos autores. As membranas poliméricas não possuem um tamanho de poro tão determinado quanto o das membranas inorgânicas. Dessa forma, foi priorizada a mineração de trabalhos que contivessem análises sobre a Distribuição do Tamanho dos Poros (DTP, em inglês *Pore Size Distribution*). Nesses casos os valores dessas variáveis foram calculados por uma média aritmética ponderada utilizando os valores dos respectivos gráficos. O diferencial do volume de poro (dV) foi definido como o peso no cálculo dessa média. Nos trabalhos que trataram de membranas inorgânicas (zeólitas e MOFs) e não analisaram a DTP, o valor teórico fornecido pelos autores foi considerado. O atributo Tipo do Tamanho de Poro tem a finalidade de indicar a forma que foi utilizada para o preenchimento dos respectivos atributos do tamanho de poro, portanto, um desses atributos não estará ausente se o outro não estiver.

Além disso, quando não fornecidos, alguns valores do Volume Total de Poro também foram calculados. Nesse caso, assumindo que os poros são totalmente preenchidos com o adsorbato líquido, o volume de poro (V_{poro}) foi calculado utilizando sempre os dados da isoterma de adsorção de N_2 (P/P_o=0,98) com a Equação (4.1) (LOWELL et al., 2004, p. 111).

$$V_{poro} = \frac{W_{ads}}{d_l} \tag{4.1}$$

Em que W_{ads} corresponde à massa de gás adsorvido e d_l à densidade do adsorbato no estado líquido. Considerando o N_2 como um gás ideal, pode-se representar a Equação (4.1) pela Equação (4.2).

$$V_{poro} = \frac{pV_{ads}V_m}{RT} \tag{4.2}$$

Em que T, p e R correspondem à temperatura, à pressão e à constante universal dos gases ideais, respectivamente. V_{ads} corresponde ao volume de gás adsorvido e V_m ao volume molar do gás adsorvido no estado líquido. Como as isotermas foram retratadas à Condições Normais de Temperatura e Pressão (CNTP), a temperatura e a pressão foram definidas como 273,15 K e 1 atm (0,101325 MPa), respectivamente. O volume molar (V_m) considerado para o nitrogênio no estado líquido foi de 34,6 $cm^3 \cdot mol^{-1}$ (@77K) e o R foi considerado como sendo 8,314462 $cm^3 \cdot MPa \cdot K^{-1} \cdot mol^{-1}$ (PERRY; GREEN; MALONEY, 1997).

Nos casos em que a referência forneceu a espessura da membrana, foram registradas as duas variáveis de performance (permeabilidade e permeância), pois uma delas foi calculada utilizando essa espessura pela Equação (2.3). O atributo Tipo de Performance Fornecida tem a utilidade de indicar qual delas é originalmente fornecida pelos autores do artigo e que, portanto, por motivos de precisão possui preferência de uso. Além disso, para a análise exploratória do Banco de Dados Original e para dar seguimento ao seu pré-tratamento antes da modelagem, as seguintes considerações sobre as informações especificadas pelas referências foram realizadas:

i. Quando a referência não especificou o Material de Suporte ("*NaN*"), essa variável foi considerada como inexistente ("*None*"). Isto é, não existe suporte. Todas as referências relativas a zeólitas e MOFs forneceram essa variável;
ii. Quando a referência não especificou a Configuração ("*NaN*"), essa variável foi considerada como "*flat*", pois é a configuração tradicional dos dispositivos de bancada;
iii. A carga de *filler* foi considerada como 0 em membranas que não sejam do tipo "MMM" (Membranas de Matriz Mista);
iv. Os atributos "*min_thickness*" e "*max_thickness*" foram substituídos pelo atributo "*mean_thickness*" (Espessura Média). O valor dessa nova variável corresponde à média aritmética simples das variáveis antecessoras;
v. Os atributos "*min_pore_size*" e "*max_pore_size*" foram substituídas pelo atributo "*mean_pore_size*" (Tamanho Médio de Poro). O valor dessa nova variável corresponde à média aritmética simples das variáveis antecessoras;

vi. Quando a referência não especificou a Idade, as membranas foram consideradas como novas (Idade=0);

vii. Temperatura ambiente (escrita como "RT" [do inglês, *Room Temperature*] no banco de dados) foi considerada como sendo 298 K;

viii. Quando a referência não especificou o Gás de Arraste ("*NaN*"), essa variável foi considerada como inexistente ("*None*"). Isto é, não existe Gás de Arraste.

4.3 Análise exploratória e pré-processamento do banco de dados

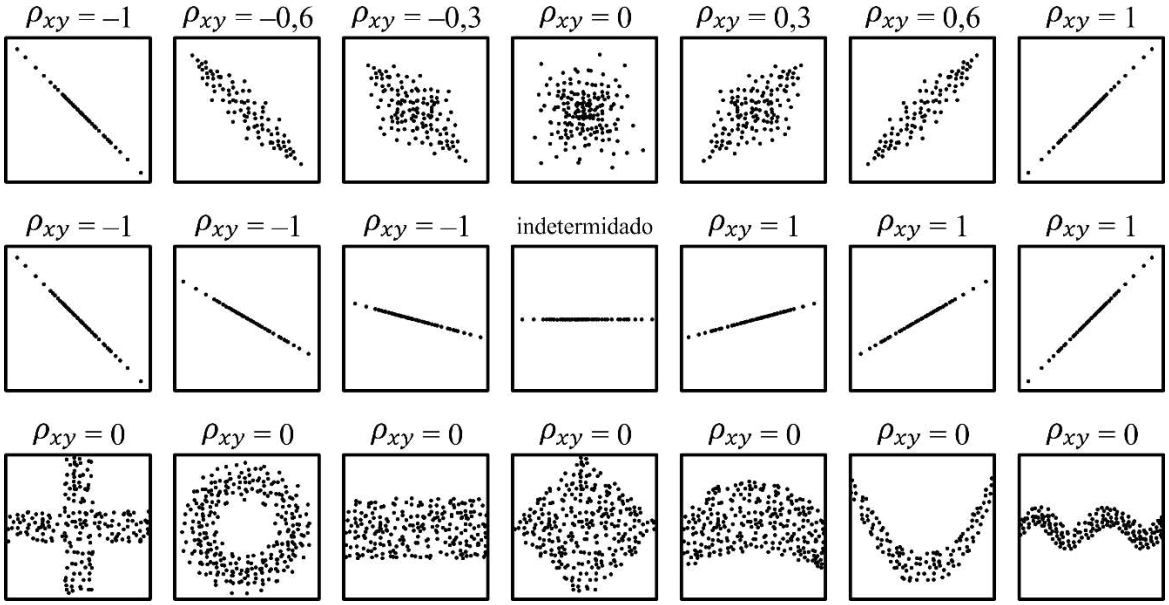
Algumas ferramentas da estatística descritiva foram utilizadas nos dados minerados com o intuito de entender a distribuição dos atributos e as relações entre eles. A faixa dos valores de cada atributo foi analisada por gráfico de barras, se forem categóricos, e por histogramas, quando forem numéricos. Esses gráficos mostram o número de instâncias (frequência) no eixo vertical que possuem um determinado valor ou intervalo de valores no eixo horizontal. Eles são muito úteis para reconhecer o tipo de dados ao qual se está trabalhando (GÉRON, 2019). As relações entre os atributos dos bancos de dados foram analisadas pela Matriz de Gráficos de Dispersão e pelo Coeficiente de Correlação de Pearson (CCP). O CPP (ρ) é um indicativo estatístico que corresponde a uma medida quantitativa da dependência linear entre dois descritores numéricos e pode ser calculada pela Equação (4.3).

$$\rho_{ij} = \frac{cov(i,j)}{\sqrt{\sigma^2(i) \cdot \sigma^2(j)}} \tag{4.3}$$

Em que as funções cov e σ^2 correspondem à covariância e à variância, respectivamente, dos atributos independentes *i* e *j*. O valor de ρ pode variar entre 1 à –1. Quanto mais próximo dos extremos for o seu valor, mais forte será a correlação, seja positiva ou negativa. Coeficientes nulos indicam que não há correlação linear. Na Figura 7 é mostrado o valor do CCP para diferentes conjuntos de dados. Pode-se perceber que nas duas primeiras linhas, ocorre correlação linear, porém na terceira linha, ocorre correlação não linear. Como pode ser vista na segunda linha da Figura 7, esse coeficiente não está relacionado com a inclinação (intensidade da "subida" ou "descida") (GÉRON, 2019). Neste trabalho, apenas os atributos numéricos foram analisados e, antes de avaliar as correlações, os valores de cada coluna/atributo foram padronizados via *z-score*. Além disso, ao selecionar os atributos *i* e *j*, os conjuntos de valores

repetidos foram retirados de forma que cada um seja único. A manutenção dos valores repetidos pode gerar uma análise subestimada da correlação entre os atributos (*i* e *j*), visto que não existe variância entre conjuntos que compartilham um único ponto e, dessa forma, o valor do CCP tenderá a ser mais nulo.

Figura 7 – Coeficiente de Correlação de Pearson para vários conjuntos de dados ilustrados em gráficos de dispersão.



Fonte: Elaborada pelo autor.

Os valores categóricos dos atributos de entrada (incluídos do banco de dados como *strings*) foram convertidos, de forma ordinária, a números inteiros característicos (converteram-se em atributos do tipo quantitativo discreto). Isto é, para um dado atributo com *n* classes diferentes, cada classe foi substituída por um número inteiro exclusivo que pode variar de *0* a *(n-1)*. Essa codificação se fez necessária visto que os modelos que foram estudados não permitem categorias de texto[2]. Além disso, a fim de manter todos os atributos na mesma escala e não gerar enviesamento do modelo para aquele cujos valores de escala são maiores, todos os

[2] Segundo Faceli et al. (2011), a técnica utilizada na conversão Simbólico-Numérico depende de o atributo ser nominal ou ordinal e essa transformação deve manter a relação de ordem (ou a inexistência dela) nos valores numéricos gerados. Como nenhum atributo categórico possui uma escala ordinal em seus valores, a utilização do codificador *OneHotEncoder* seria ideal (GÉRON, 2019). Porém, o mesmo foi testado e pouquíssimas diferenças foram observadas na qualidade dos modelos. Em prol de uma interpretação mais simples das importâncias desses atributos, um codificador ordinal foi implementado.

atributos foram padronizados via *z-score*. Dessa forma, o atributo escalonado (X) pode ser definido pela Equação (4.4).

$$X = \frac{x - \mu}{\sigma} \tag{4.4}$$

Em que μ e σ correspondem à média aritmética simples e ao desvio padrão do atributo x, respectivamente. Assim, a média do grupo de valores em que foi realizada a padronização possui uma média igual a 0 e um desvio padrão igual a 1. Esse escalonamento é muito menos afetado por *outliers*, característica favorável para o contexto do presente trabalho, visto que há uma grande variabilidade nos atributos e não foi realizada nenhuma poda nos valores muito altos ou baixos do banco de dados tendo em vista que os dados não obedecem a nenhuma distribuição estatística em específico.

4.4 Modelos

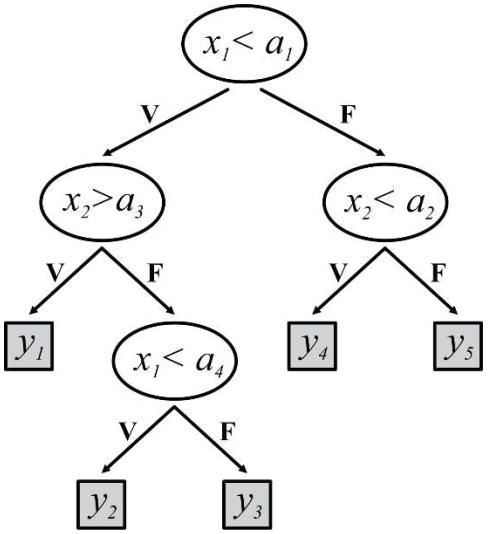
4.4.1 Florestas Aleatórias

A Floresta Aleatória (do inglês, *Random Forest*) é um algoritmo de AM introduzido pela primeira vez por Breiman (2001) e corresponde a uma coleção de várias Árvores de Decisão (AD) treinadas em subconjuntos aleatórios de características para fazer a predição de um mesmo problema (GÉRON, 2019; IZBICKI; SANTOS, 2020). As ADs são modelos simples que usam divisões binárias em variáveis preditoras para determinar previsões de resultados. Elas são práticas, oferecendo um método intuitivo para prever o resultado que divide os valores “alto” versus “baixo” de um atributo de entrada relacionado à variável alvo (SPEISER *et al.*, 2019). Mais precisamente, quando utilizadas em problemas de regressão, elas podem ser denominadas como Árvores de Regressão (AR). As ADs são adequadas para encontrar regras de previsão não lineares e são bem interpretáveis, embora sua instabilidade seja um motivo de preocupação.

Mais precisamente, como pode ser visto na Figura 8, uma árvore de decisão pode ser definida como “um grafo acíclico direcionado em que cada nó ou é um nó de divisão, com dois ou mais sucessores, ou um nó folha” (FACELI et al., 2011, p. 83). Cada nó de divisão será representado por um teste condicional baseado nos valores de um determinado atributo de entrada. No processo de decisão, percorre-se a árvore a partir da raiz (primeiro nó de decisão),

passando pelos testes condicionais, até o nó folha, que corresponde a decisão final (relativa ao atributo alvo) daquela árvore.

Figura 8 – Representação de uma árvore de decisão, em que *x* corresponde ao valor do atributo de entrada de um determinado registro. Em cada nó de decisão, representado pelas elipses, existe um teste condicional referente a um valor *a* para o atributo *x*. Ao percorrer a árvore, cada teste pode ser verdadeiro (V) ou falso (F) até chegar a uma resposta final referente ao atributo alvo *y*, denominada de nó folha e representada pelos quadrados.



Fonte: Adaptado de Faceli et al. (2011, p. 84)

Dessa forma, pode-se definir a profundidade da árvore como o número máximo de testes condicionais (nós de decisão) que é possível passar para se ter uma decisão. Esses algoritmos são muito flexíveis, pois não assumem nenhuma distribuição do dados; são robustos, pois são praticamente invariantes a escala dos atributos de entrada; tendem a não selecionar atributos irrelevantes ou redundantes; são eficientes, pois possuem uma complexidade de tempo linear ao número de registros de treinamento; e, por fim, são interpretáveis, pois decisões complexas são aproximadas por uma série de decisões mais simples baseadas nos valores dos atributos de entrada (FACELI et al. 2011).

Em contrapartida, algoritmos de ordenação são necessários ao se trabalhar com atributos contínuos, algo que pode diminuir a sua eficiência. Além disso, existem questões inerentes relativas à replicação de teste condicionais e, como já dito, à instabilidade. Muitos pesquisadores apontam que pequenas variações no conjunto de treino podem provocar grandes mudanças na árvore final (FACELI et al. 2011).

Curiosamente, a instabilidade das ADs as tornam ótimas candidatas para a aplicação de métodos de aprendizado em conjunto. Nesses procedimentos, vários modelos do mesmo tipo

são combinados para gerar uma previsão final e assim, é essencial que haja uma a alta variância entre os modelos que formam o conjunto. Dessa forma, pode-se afirmar que são combinadas as decisões de diferentes especialistas para se obter uma resposta final. Existem diferentes formas para criar diversidade entre os modelos do conjunto: manipulando os registros de treinamento, fazendo uma amostragem dos atributos de entrada, embaralhando os registros teste ou inserindo aleatoriedade na inicialização dos parâmetros (ALPAYDIN, 2020; FACELI *et al.*, 2011).

Assim, o algoritmo FA pode ser considerado um método de conjunto e foi projetado para superar os problemas referentes às AD. Por causa disso, se tornou muito popular ao combinar a interpretabilidade com o desempenho de algoritmos de aprendizagem modernos, como redes neurais e as máquinas de vetores de suporte (ALTMANN *et al.*, 2010). Uma breve descrição dos hiperparâmetros desse modelo pode ser vista na Tabela 4.

Tabela 4 – Descrição dos hiperparâmetros de treinamento do modelo Florestas Aleatórias.

Hiperparâmetro[1]	**Descrição**
Número de Árvores (*n_estimators*)	Quantidade de árvores de decisão (estimadores) na Floresta.
Critério (*criterion*)	Função para medir a qualidade da divisão de um nó.
Profundidade Máxima (*max_depth*)	Profundidade máxima das árvores de decisão.
Mínimo de Folhas (*min_samples_leaf*)	O mínimo de amostras requeridas em um nó de folha.
Número de Atributos (*max_features*)	Número de atributos de entrada a serem considerados na procura da melhor divisão.
Tamanho da Amostra (*max_samples*)	Quantidade de registros que será utilizado para treinar o modelo.
K-fold (*cv*)	Número de divisões para a Validação Cruzada.

[1] O nome em parênteses se refere ao nome do hiperparâmetro na biblioteca computacional. Fonte: Elaborada pelo autor.

4.4.2 Máquinas de Vetores de Suporte

Máquinas de Vetores de Suporte (MVS) são sistemas de aprendizado que podem ser usados tanto para classificação quanto para regressão e que utilizam um espaço de hipóteses de funções lineares em um espaço de atributos de alta dimensão. A ideia básica desse algoritmo se mostra no cálculo de instâncias, que também podem ser chamados de hiperplanos ou vetores

de suporte, que separem da melhor forma duas classes existentes ou que represente melhor a relação entre variáveis. Essas instâncias são definidas (ou "suportadas") pelas suas respectivas margens, que servem como "limites de decisão" e cujas características são regidas pelos hiperparâmetros do modelo (GÉRON, 2019). Os MVSs são treinados por algoritmos embasados pela Teoria do Aprendizado Estatístico (TAE), um método poderoso desenvolvido por Vapnik (2000) que, devido a sua criação para um contexto industrial, é utilizado em uma ampla variedade de aplicações mesmo após poucos anos desde sua introdução. A TAE estabelece as condições matemáticas para a escolha de um classificador particular (ou hiperplano) a partir de um conjunto de dados de treinamento (CRISTIANINI; SHAWE-TAYLOR, 2000; FACELI *et al.*, 2011).

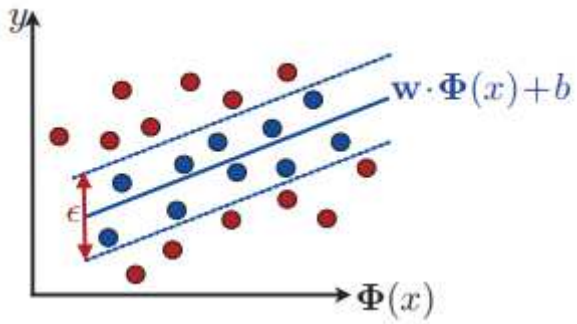
Tal sistema aplicado em problemas de regressão é denominado como Regressão de Vetores de Suporte (RVS), uma extensão do MVS, e podem trabalhar com dados linearmente ou não-linearmente separáveis. Nesse contexto, utiliza-se o algoritmo chamado ε-SVR (do inglês, *Support Vector Regression*) que tem como objetivo encontrar uma função *f(x)* que retorne valores com margem de erro de no máximo ε com relação aos valores reais e que, ao mesmo tempo, seja o mais plano/linear o possível. (FACELI *et al.*, 2011; SMOLA; SCHÖLKOPF, 2004). De outra forma, como pode ser visto na Figura 9, o RVS tenta ajustar um "tubo" ou uma "via" com largura ε aos dados, de forma que os dados de treinamento dentro do "tubo *epsilon*" (pontos azuis) não são contabilizados (MOHRI; ROSTAMIZADEH; TALWALKAR, 2018).

A maioria das aplicações do mundo real são complexas de forma que se torna inevitável lidar com dados cujas relações não são lineares. Nesse contexto, a aplicação de *kernels* se mostra como uma ótima solução para esse caso sem aumentar tanto o custo computacional. Nessa abordagem, o espaço original do conjunto de treinamento, denominado espaço de entradas, é transformado para um novo espaço de maior dimensão, denominado espaço de características. Isto é, transforma-se os objetos de $\mathbb{R}^d$ para $\mathbb{R}^{d+n}$ utilizando um transformador não linear. A partir desse novo espaço, em tese, uma relação linear entre os conjuntos de dados pode ser obtida e o algoritmo RVS linear pode ser implementado (FACELI *et al.*, 2011). Vários *kernels* podem ser empregados para essa transformação e os mais comuns são os polinomiais, os de função de base radial (RBF – *Radial Basis Function*) e os sigmoidais. A função do *kernel* RBF pode ser representada pela Equação (4.5).

$$f_{RBF} = e^{-\gamma \|x-x'\|^2} \tag{4.5}$$

Em que x e x' correspondem às variáveis independentes (ou aos atributos do banco de dados) e γ (*gamma*) é um parâmetro arbitrário que precisa ser maior que 0. No processo de otimização do hiperplano utilizando margens suaves[3], um coeficiente de penalidade (C), que representa uma forma de tolerância ao erro, é aplicado. Esse hiperparâmetro serve como um termo de regularização de forma que quanto maior for o seu valor, menor será as margens do hiperplano e menos generalista tenderá a ser o modelo. Além disso, recomenda-se diminuir o valor desse parâmetro para diminuir um possível sobreajuste.

Figura 9 – Representação de um vetor de suporte de largura ε para o modelo Regressão de Vetores de Suporte.



Fonte: (MOHRI; ROSTAMIZADEH; TALWALKAR, 2018)

De forma semelhante, γ é um parâmetro que aparece nas definições dos *kernels* mais comuns de modo que também pode ser visto como um hiperparâmetro de regularização. Como o C, caso o modelo estiver sobreajustado, recomenda-se reduzi-lo de forma a aumentar a capacidade de generalização. Especificamente no *kernel* RBF, aumentar o valor de γ estreita a curva em forma de sino da função e diminui o raio de influência dos vetores de suporte de modo que sua forma se torna mais irregular (com mais curvas). Em contrapartida, diminuir o seu valor torna a curva em forma de sino mais espaçosa e as instâncias de decisão ficam mais suaves (GÉRON, 2019). Uma breve descrição dos hiperparâmetros desse modelo pode ser vista na Tabela 5.

[3] MVSs de margens suaves é o termo utilizado quando se permite que alguns dados do conjunto de treinamento violem as restrições da margem de tamanho ε, de forma que o modelo consiga trabalhar com dados mais gerais. Essa violação de margem é regulada pelo coeficiente C. Quando não se permite violações tem-se um MVS de margem rígida (FACELI *et al.*, 2011).

Tabela 5 – Descrição dos hiperparâmetros de treinamento do modelo Máquinas de Vetores de Suporte.

Hiperparâmetro[1]	**Descrição**
Gama (*gamma*)	Coeficiente de regularização associado à função definição do kernel utilizado.
C (*C*)	Termo de regularização que impõe um peso no processo de minimização dos erros marginais.
Epsilon (*épsilon*)	Espessura da margem. É um parâmetro insensível ao conjunto de dados.

[1] O nome em parênteses se refere ao nome do hiperparâmetro na biblioteca computacional. Fonte: Elaborada pelo autor.

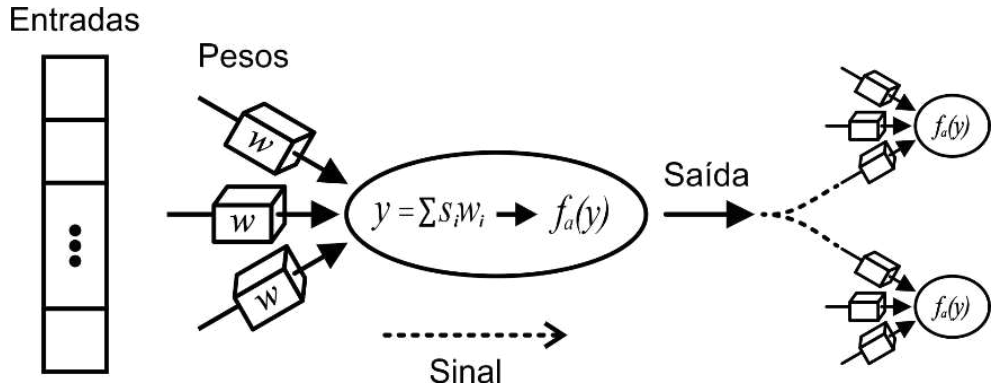
4.4.3 Redes Neurais Artificiais

As Redes Neurais Artificiais (RNA) são algoritmos de AM cuja teoria foi inspirada no sistema nervoso humano, em específico, pela forma como os neurônios biológicos recebem e transmitem informação (FACELI *et al.*, 2011). Mcculloch e Pitts (1943) foram os pioneiros a estudar esses modelos que, inicialmente, foram criados com a finalidade de simular e estudar o cérebro humano. Entretanto, os atuais sistemas computacionais diferem bastante desse contexto tanto em finalidade quanto em funcionamento. O *perceptron* é um modelo de neurônio artificial criado por Rosenblatt (1958) (também chamado de *Linear Threshold Unit*) e corresponde à unidade básica de processamento cuja junção, caracterizada pela forma com a qual serão conectadas, formará uma rede neural. Na Figura 10 é mostrada a representação de um *perceptron* cujas entradas (s_i) podem ser originadas de um banco de dados ou da saída de outros neurônios. Cada entrada possui um peso associado (w_i) e para "transmitir o sinal" a unidade primeiramente fará uma soma ponderada delas e, em seguida, aplicará uma função de ativação, que determinará o valor limite para a informação ou "sinal" ser passado adiante.

Treinar uma RNA equivale a ajustar seus pesos para minimizar uma função de custo, que é uma representação do erro da previsão. Esses pesos podem possuir valores positivos ou negativos, o que caracterizará uma conectividade excitatória ou inibitória, respectivamente (ALPAYDIN, 2020; FACELI *et al.*, 2011; GÉRON, 2019). Pode-se dividir as características de uma RNA em dois grupos: arquitetura e aprendizado. A arquitetura está relacionada à forma com que os neurônios estão conectados, à quantidade de unidades e aos respectivos tipos

(definido pela função de ativação). O aprendizado está relacionado às regras utilizadas para o treinamento e à informação que será utilizada para isso (FACELI *et al.*, 2011).

Figura 10 – Representação de um neurônio artificial em que *y* corresponde ao "sinal" recebido pelo neurônio. Os símbolos s_i e w_i correspondem ao sinal de saída emitido e o respectivo peso do neurônio *i* da camada anterior. O sinal emitido por cada neurônio passa por uma função de ativação f_a.



Fonte: Adaptado de FACELI et al. (2011).

No presente trabalho, uma rede neural multicamada *feedforward* (que não permitem conexões de retropropagação) foi utilizada para investigar a performance das membranas. A rede neural que será implementada pode ser visualizada na Figura 11 e possui uma camada de entrada, apenas uma camada oculta um neurônio na camada de saída. Os quadrados representam os atributos de entrada do banco de dados, os círculos representam os neurônios artificiais e as linhas suas conectividades associadas a um respectivo peso. Um neurônio de viés foi adicionado na camada de entrada e na camada oculta. Ele pode ser definido como uma unidade que transmite apenas o valor 1, que será ajustado pelo respectivo peso para a próxima unidade e cuja adição ajuda a tornar o modelo mais generalista. Dessa forma, para um determinado número de neurônios na camada anterior ou valores de entrada (*d*), cada neurônio receberá um valor *y* calculado com base na Equação (4.6).

$$y = \sum_{i=1}^{d} (s_i w_i) + w_0 \tag{4.6}$$

Em que s_i corresponde ao sinal de saída do neurônio *i* da camada anterior (ou um atributo de entrada) e w_i corresponde ao respectivo peso. w_0 corresponde ao peso do neurônio de viés da camada anterior. Foram investigadas três funções de ativação para compor os neurônios da

camada oculta: ReLU (do inglês, *Rectified Linear Unit)*, sigmoide e *Leaky ReLU* (do inglês, *Leaky Rectified Linear Unit*). Para um dado valor de entrada *i*, as funções ReLU (f_a^{ReLU}), sigmoide (f_a^{sig}) e *Leaky ReLU* (f_a^{LReLU}) foram definidas respectivamente como apresentado nas Equações (4.7) a (4.9).

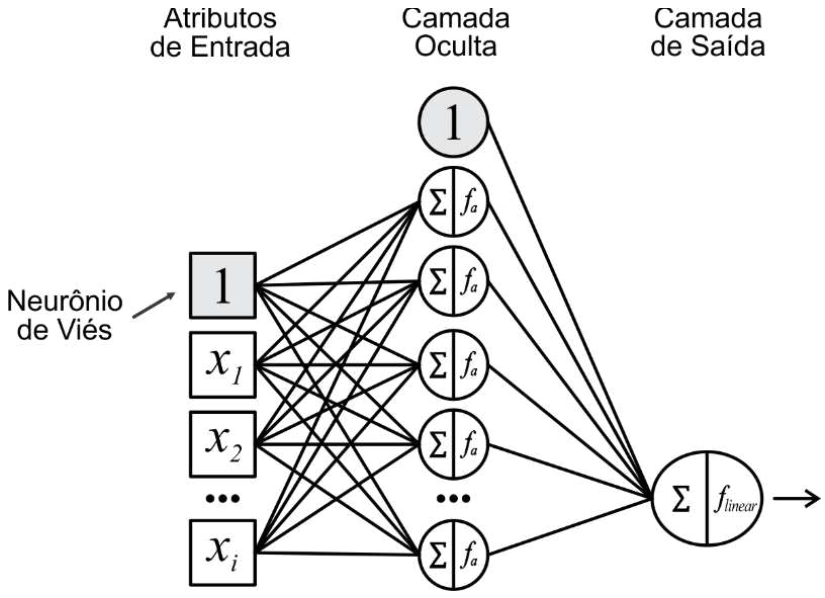
$$f_a^{ReLU}(i) = \max(0, i) \tag{4.7}$$

$$f_a^{sig}(i) = \frac{1}{1 + e^{-i}} \tag{4.8}$$

$$f_a^{LReLU}(i) = \begin{cases} 0{,}3i & se\ i < 0 \\ i & se\ i \geq 0 \end{cases} \tag{4.9}$$

Em que *max* retorna o maior valor entre 0 e a variável *i*. A função de ativação do neurônio da camada de saída foi a identidade, também chamada de função linear, caracterizada por não alterar o sinal de entrada.

Figura 11 – Diagrama da RNA utilizada no presente trabalho.



Fonte: Elaborada pelo autor.

Para estimar os pesos das conecções, deve-se especificar uma função de custo (ou objetivo) a ser minimizada. O algoritmo utilizado para isso, denominado retropropagação (*back-propagation*), foi proposto inicialmente por (RUMELHART; HINTON; WILLIAMS, 1986) e vem sendo continuamente aperfeiçoado desde então. Ele é baseado numa regra chamada "delta generalizada" e é semelhante ao que se chama atualmente de gradiente

descendente estocástico. Nele, há a iteração de duas etapas, uma para frente (*forward*), no sentido padrão ao qual se faz a predição, e a outra para trás (*backward*), quando os pesos são ajustados com base no erro de cada neurônio. Esse ajuste na etapa *backward* prossegue até a camada com os primeiros neurônios e, para cada iteração *m*, é regido pela Equação (4.10) (ALPAYDIN, 2020; FACELI *et al.*, 2011; GÉRON, 2019).

$$w_{jl}(m+1) = w_{jl}(m) + \eta \frac{\delta e_l(\widehat{y_l}, y_i)}{\delta w_{jl}} \tag{4.10}$$

Em que w_{jl} representa o peso entre o neurônio *l* e o *j*-ésimo atributo de entrada ou o j-ésimo neurônio da camada anterior, e_l representa a função de custo, que mede o erro do *l*-ésimo neurônio entre o valor predito $\widehat{y_l}$ e o valor real y_i (IZBICKI; SANTOS, 2020). O valor predito $\widehat{y_l}$ pode ser representado pelos pesos a partir Equação (4.6). No presente trabalho foram investigadas três funções de custo no processo de otimização do modelo: o Erro Quadrático (EQ), o Erro Relativo (ER) e o Logaritmo do Cosseno Hiperbólico (*Log-Cosh*) que foram definidas respectivamente como apresentado nas Equações (4.11) a (4.13).

$$e^{EQ}(\widehat{y_l}, y_i) = (y_i - \widehat{y_l})^2 \tag{4.11}$$

$$e^{ER}(\widehat{y_l}, y_i) = \left|\frac{y_i - \widehat{y_l}}{y_i}\right| \times 100 \tag{4.12}$$

$$e^{logcosh}(\widehat{y_l}, y_i) = \ln(\cosh(\widehat{y_l} - y_i)) \tag{4.13}$$

A variável η da Equação (4.10) é um hiperparâmetro denominado taxa de aprendizado. Ele tem forte influência no tempo de convergência da rede e indiretamente na qualidade da previsão. Valores altos demais levam a grandes atualizações nos pesos e divergem o treinamento. Por outro lado, valores pequenos aumentam o tempo de treinamento e exigem mais custo computacional (FACELI *et al.*, 2011).

Antes de tudo, a escolha da metodologia mais adequada para a inicialização dos pesos é de suma importância principalmente em redes neurais profundas para evitar problemas dos gradientes (*vanishing* ou *exploding*) em que o algoritmo de treinamento converge para um mínimo local ou apresenta problemas numéricos (quando as camadas recebem atualizações de peso muito grandes fazendo divergir o algoritmo). A inicialização dos pesos foi realizada pelo método Uniforme *Glorot* (GLOROT; BENGIO, 2010), em que os pesos de cada camada são inicializados aleatoriamente e podem ter valores entre a faixa [*r*, *–r*], com distribuição de

probabilidade uniforme. O valor limite *r* para a faixa é definido pela Equação (4.14) (GÉRON, 2019).

$$r = \sqrt{\frac{6}{n_{entradas} + n_{saídas}}} \tag{4.14}$$

Em que $n_{entradas}$ e $n_{saídas}$ representam o número de conexões de entrada e de saída da camada, respectivamente. Os pesos dos neurônios de viés foram inicializados em 0. Além disso, a fim de acelerar o treinamento, o otimizador *Adam* foi implementado. Basicamente, nesse algoritmo a taxa de aprendizado é estimada adaptativamente e seus valores dependem dos seus antecessores (KINGMA; BA, 2014). No processo de treinamento, a versão padrão do algoritmo de retropropagação ajusta os pesos para cada registro apresentado, entretanto, existe uma variação denominada *Mini-Batch Gradient Descent* ou Aprendizado em Batelada (do inglês, *Batch Learning*) que foi utilizada nesse trabalho, em que a atualização dos pesos é feita apenas após a apresentação de *n* de registros denominados lotes ou *batchs* (HAYKIN, 2009). Nesse contexto, a média do erro desses *n* registros são contabilizados para o ajuste dos pesos de acordo com a Equação (4.10). Além disso, denomina-se o hiperparâmetro época (ou *epochs*) a quantidade de vezes que o algoritmo passa pelo conjunto de treino que é dividido em lotes e a cada etapa embaralhado aleatoriamente (FACELI *et al.*, 2011; GÉRON, 2019; GOODFELLOW; BENGIO; COURVILLE, 2016).

Devido a característica estocástica das redes neurais, o modelo foi treinado várias vezes e a média do seu desempenho foi utilizado no processo de otimização dos hiperparâmetros. Segundo Faceli et al. (2011), apesar de não ser um procedimento comum devido ao elevado custo computacional, é recomendado que isso seja feito utilizando diferentes valores iniciais de peso e configurações de dados. Essa metodologia pode ser classificada como um método de conjunto (*ensemble method*), uma técnica primeiramente implementada por Breiman (1996) com a finalidade de diminuir o erro de generalização combinando o resultado de vários modelos com características aleatórias e diferentes. Algumas implementações não determinísticas de redes neurais são suficientes para fazer com que diferentes modelos do conjunto cometam erros parcialmente independentes, com baixo viés e alta variância, validando assim o uso desse método. Nesse caso, no início de cada época o conjunto de treino é embaralhado aleatoriamente e os pesos são inicializados aleatoriamente (FACELI *et al.*, 2011; GOODFELLOW; BENGIO;

COURVILLE, 2016). Uma breve descrição dos hiperparâmetros do modelo RNA pode ser vista na Tabela 6.

Tabela 6 – Descrição dos hiperparâmetros de treinamento do modelo Redes Neurais Artificiais.

Hiperparâmetro[1]	**Descrição**
Número de Neurônios (*units*)	Quantidade de neurônios na camada oculta.
Função de Ativação (*activation*)	Função de ativação dos neurônios da camada oculta.
Taxa de Aprendizado (*learning_rate*)	Taxa de convergência do treinamento do modelo. Valor do parâmetro η da Equação (4.12).
Função de custo (*loss*)	Função de erro a ser minimizada no treinamento do modelo.
Lotes (*batchs*)	Define o número de registros para o cálculo da média do erro antes de atualizar/ajustar os pesos.
Épocas (*epochs*)	Quantidade de vezes que o algoritmo de treinamento utilizará o banco de dados inteiro.

[1] O nome em parênteses se refere ao nome do hiperparâmetro na biblioteca computacional. Fonte: Elaborada pelo autor.

4.5 Métricas de desempenho

A fim de quantificar a qualidade da predição dos modelos, foram utilizados três indicadores estatísticos: a Raiz do Erro Quadrático Médio (em inglês, *Root Mean Square Error* - RMSE), o Erro Percentual Absoluto Médio (em inglês, *Mean Absolute Percentage Error* – MAPE) e o Coeficiente de Determinação (R^2). O RMSE para um conjunto de dados com tamanho n pode ser definido pela Equação (4.15) (PAN *et al.*, 2022).

$$RMSE(y, \hat{y}) = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(y_i - \hat{y_i})^2} \tag{4.15}$$

Em que $\hat{y_i}$ e y_i correspondem ao valor predito e real do registro i, respectivamente. Essa métrica apresentará uma média do erro absoluto dos valores preditos, na mesma unidade de medida do atributo alvo. O MAPE para um conjunto de dados com tamanho n pode ser definido pela Equação (4.16) (KIM; KIM, 2016).

$$MAPE(y, \hat{y}) = \frac{1}{n}\sum_{i=0}^{n-1}\frac{|y_i - \hat{y}_i|}{\max(\epsilon, |y_i|)} \tag{4.16}$$

Em que $\hat{y}_i$ e y_i correspondem ao valor predito e real do registro i, respectivamente. Para evitar resultados indefinidos (divisão por 0), a função max retorna o maior valor entre ϵ, um valor arbitrário bem pequeno, e $|y_i|$. Pode-se interpretar o MAPE como uma média dos módulos dos erros relativos dos valores preditos. Essa métrica se mostra útil nos casos em que há grande variabilidade do atributo alvo e permite uma melhor comparação entre os registros de diferentes contextos (e.g. comparar o erro de previsão da permeabilidade do H_2, um gás leve, com o do SF_6, um gás mais pesado). Por fim, o Coeficiente de Determinação (R^2) para um conjunto de dados com tamanho n foi definido pela Equação (4.17) (PAN *et al.*, 2022).

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n-1}(y_i - \bar{y})^2} \tag{4.17}$$

Em que $\hat{y}_i$ e y_i correspondem ao valor predito e real do registro i, respectivamente. A variável $\bar{y}$ corresponde a média dos valores reais. Essa métrica representa a capacidade que um modelo possui de acompanhar a variabilidade dos valores reais, dado um conjunto de variáveis de entrada independentes. Os valores de R^2 podem variar de 1 a $-\infty$ e quanto mais positivo for, mais ajustado estará o modelo. Um modelo que sempre prever o valor esperado (média) terá um R^2 igual a 0. No decorrer deste trabalho, o MAPE e o R^2 foram definidos como os principais indicadores para medir a adequação dos modelos e, portanto, irão reger as escolhas metodológicas nos respectivos processos de otimização/ajuste.

4.6 Otimização e modelagem

Para o uso do modelo de Aprendizado de Máquina supervisionado, 20% do banco de dados (139 registros) foi separado para o conjunto de teste e 80% (553 registros) para o conjunto de treino. Esta proporção é amplamente aceita pela comunidade científica (ALIBAKSHI, 2018) e foi calibrada via testes preliminares realizados pela construção de curvas de aprendizado.

Para busca dos hiperparâmetros ótimos dos modelos, foi utilizada a biblioteca computacional Optuna (AKIBA *et al.*, 2019). Tendo como referência uma função objetivo

(F_{obj}), tal biblioteca permite a pesquisa automatizada de hiperparâmetros ao implementar o algoritmo Bayesiano como metodologia de otimização (PRAVIN *et al.*, 2022). Neste trabalho, foram estudadas a aplicação de sete funções objetivo para a otimização dos hiperparâmetros que podem ser vistas na Tabela 7. Dessa forma, objetivou-se investigar as suas respectivas influências nos hiperparâmetros ótimos e nas métricas dos modelos, com o foco em solucionar o problema do sobreajuste sem interferência do usuário.

Tabela 7 – Funções objetivo (F_{obj}) utilizadas para a otimização dos hiperparâmetros dos modelos.

Função objetivo (F_{obj})		Descrição das variáveis
$F_0 = 1 - R^2_{treino}$	(4.18)	R^2_{treino} = Coeficiente de determinação do conjunto de treino. $\bar{R}^2_{CV\_treino}$ = Média dos coeficientes de determinação dos conjuntos de treino da VC. $\bar{R}^2_{CV\_teste}$ = Média dos coeficientes de determinação dos conjuntos de teste da VC. $\overline{MAPE}_{CV\_treino}$ = Média dos MAPEs dos conjuntos de treino da VC. $\overline{MAPE}_{CV\_teste}$ = Média dos MAPEs dos conjuntos de teste da VC.
$F_1 = 1 - \bar{R}^2_{CV\_treino}$	(4.19)	
$F_2 = 1 - \bar{R}^2_{CV\_teste}$	(4.20)	
$F_3 = \left(1 - \bar{R}^2_{cv\_teste}\right) + \left\lvert 1 - \frac{\bar{R}^2_{CV\_treino}}{\bar{R}^2_{CV\_teste}} \right\rvert$	(4.21)	
$F_4 = \overline{MAPE}_{CV\_treino}$	(4.22)	
$F_5 = \overline{MAPE}_{CV\_teste}$	(4.23)	
$F_6 = \overline{MAPE}_{CV\_teste} + \left\lvert \overline{MAPE}_{CV\_treino} - \overline{MAPE}_{CV\_teste} \right\rvert$	(4.24)	

Fonte: Elaborada pelo autor.

A função objetivo F_0 é a comumente utilizada para a otimização dos parâmetros nos trabalhos da literatura. As outras funções objetivo foram testadas como alternativas e fazem uso das métricas do método de Validação Cruzada (VC). Ao todo, quatro funções objetivo fazem uso do coeficiente de determinação (R^2) e três fazem uso do MAPE em sua definição. Na Validação Cruzada, divide-se os dados de treino aleatoriamente em *k* lotes de tamanho aproximadamente igual. Os lotes de partições são treinados por um preditor e a partição restante é então testada/validada. Esse processo é repetido *k* vezes de forma que cada lote foi usado como teste (FACELI *et al.*, 2011; IZBICKI; SANTOS, 2020). Assim, pode-se afirmar que o

processo de VC também divide o conjunto de treino original em duas partições: treino e teste. No processo de minimização das funções objetivo, denominou-se o parâmetro "*k*-fold" como o número de divisões da Validação Cruzada. Vale salientar que ele é relevante apenas nessa etapa e que após a obtenção dos hiperparâmetros ótimos, os modelos não serão submetidos à VC para as suas respectivas avaliações. As funções objetivo F_3 e F_6 possuem termos que medem a diferença de desempenho entre o conjunto de treino e teste da VC e foram denominados "termos de aproximação". Pretende-se investigar se a presença deles consegue diminuir um possível sobreajuste dos modelos analisados.

4.7 Interpretabilidade

4.7.1 Importância de Permutação

Para avaliar as importâncias dos atributos de entrada na previsão da permeabilidade foi utilizado o recurso de permutação. Esse método de medição de importância foi introduzido por Breiman (2001) para Florestas Aleatórias cujo algoritmo é descrito na sequência (MOLNAR, 2021).

Algoritmo 2 – Importância de Permutação

Entrada: Modelo treinado *m* e um banco de dados tabular *D*.
Saída: Importância para cada atributo *j*.

1 Estimar o erro de *m* para o banco de dados *D*, definido como S_D.
2 Para cada atributo *j* (coluna de *D*):
3 Para cada repetição *k* entre [1, ..., *K*]:
4 Embaralhar aleatoriamente a coluna de *j* e gerar uma versão corrompida de *D*, definida como $D_{j,k}$.
5 Estimar o erro de *m* para $D_{j,k}$, definido como $S_{Dj,k}$.
6 Calcular a importância (I_j) do atributo *j*, definido como:

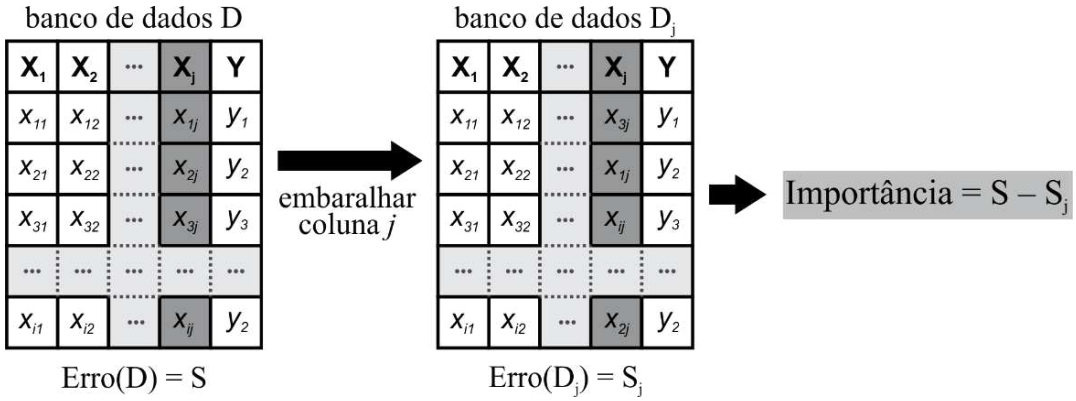
$$I_j = S_D - \frac{1}{K}\sum_{k=1}^{K} S_{Dj,k}$$

7 Retornar a importância I_j de cada atributo *j*.

Nessa análise chamada Importância de Permutação, a relevância de um atributo de entrada é proporcional à diminuição na qualidade de um modelo (que pode ser medida por algum indicador estatístico como o R^2 ou o RMSE) quando os seus valores são embaralhados, como pode ser visto na Figura 12. Desse modo, quebra-se a relação entre a variável de entrada

e o alvo, de modo que a queda na qualidade serve como indicativo quantitativo de dependência do modelo (BREIMAN, 2001). Um recurso é "sem importância" caso o erro do modelo se mantiver inalterado ao embaralhar os valores do atributo, visto que nesse caso o modelo ignorou o recurso para a previsão. Nessa análise, o R^2 foi a métrica escolhida para quantificar a qualidade de previsão do modelo e, para cada atributo, 30 embaralhamentos (repetições) foram realizados. Como pode ser observado no algoritmo acima, a importância é definida como a média dos resultados de cada embaralhamento.

Figura 12 – Etapas para o cálculo da importância de um atributo X_j em um banco de dados D. D_j corresponde ao banco de dados com a coluna X_j embaralhada. O erro da previsão utilizando D equivale a S e utilizando D_j equivale a S_j.



Fonte: Elaborada pelo autor.

4.7.2 Gráficos de Dependência Parcial

Para entender como cada atributo afeta a variável alvo, Gráficos de Dependência Parcial (GDP) de apenas uma dimensão foram construídos. Essa análise mostra o efeito marginal médio de um ou dois atributos na variável alvo de um modelo de AM. Esse método pode ser considerado global, isto é, pode ser aplicado em qualquer modelo de AM, e irá mostrar o perfil da relação entre o atributo de entrada e alvo (linear, monotônica ou mais complexa) (GUAN *et al.*, 2022; MOLNAR, 2021). A função de dependência parcial (f_{GDP}) para um determinado atributo de entrada x_S pode ser definida pela Equação (4.25) (FRIEDMAN, 2001).

$$f_{GDP}(x_S) = \int \hat{y}(x_S, x_C)\mathbb{P}(x_C)dx_C \tag{4.25}$$

Em que x_C representa todos os outros atributos de entrada (atributos complementares) utilizados na modelagem e $\mathbb{P}(x_C)$ corresponde à suas respectivas densidades de probabilidade

marginal. Essa função pode ser estimada calculando uma média de um conjunto de dados, também chamado de método Monte Carlo como apresentado na Equação (4.26).

$$f_{GDP}(x_S) = \frac{1}{n}\sum_{i=1}^{n} \hat{y}\left(x_S, x_C^{(i)}\right) \tag{4.26}$$

Em que $x_C^{(i)}$ corresponde aos atributos complementares do *i*-ésimo registro e *n* corresponde à quantidade de registros analisados. Dessa forma, o valor do GDP para um ponto existente de x_S (no conjunto de dados) é definido como a média das previsões do modelo utilizando todos os cenários possíveis para as variáveis complementares x_C, mantendo constante x_S. Nesse trabalho, todos os registros do conjunto de teste foram considerados para o cálculo.

4.8 Ambiente de programação

Além da modelagem, toda a análise exploratória e processamento dos dados foram realizados no ambiente online *Google Colaboratory*[4] utilizando a linguagem de programação de alto nível *Python* (versão 3.7.13). Essa ferramenta é um recurso da *Google* especialmente adequado para aplicações de AM, que oferece um ambiente de desenvolvimento integrado na nuvem (máquina virtual vinculada a conta do usuário), semelhante ao *Jupyter Notebook*. Foram utilizadas diversas bibliotecas computacionais para o tratamento dos dados, execução da modelagem e visualização dos resultados. Dentre as principais estão a *scikit-learn*, *pandas*, *NumPy* e *Matplotlib*.

4.8.1 Scikit-learn e Keras

O *scikit-learn* (versão 1.3) é uma biblioteca de AM de código livre que suporta tarefas de aprendizado supervisionado e não supervisionado. Além de possuir diversos algoritmos de regressão, classificação e clusterização, fornece ferramentas para o pré-processamento dos

[4] https://research.google.com/colaboratory/intl/pt-BR/faq.html

dados, avaliação e interpretabilidade dos modelos. A separação dos conjuntos (treino e teste)[5] e a implementação dos modelos (Florestas Aleatórias[6] e Regressão de Vetores de Suporte[7]) foram realizadas por essa biblioteca.

A Rede Neural Artificial foi implementada utilizando a biblioteca computacional *Keras*[8] (versão 2.13.1). Ela também possui uma API escrita em *Python* e é executada sob a plataforma *TensorFlow*. Nela, é fornecido uma interface de alto nível para a criação de redes neurais, que incluem a criação de modelos sequenciais ou mais complexos com várias camadas (incluindo redes convolucionais, redes recorrentes e diversas combinações). O *Keras* é bem reconhecido no campo de estudo e permite treinar os modelos de Redes Neurais utilizando programação paralela (GPU e TSU) de uma maneira simples, flexível e rápida.

4.8.2 Pandas e NumPy

O *pandas*[9] (versão 2.0.3) é uma das bibliotecas de código aberto mais populares para a análise de dados. Ela fornece estruturas e diversas ferramentas para a sua manipulação de dados com um alto desempenho e uma API construída com base na linguagem *Python*. O *DataFrame* é a principal estrutura de dados da biblioteca que possui uma organização tabular e foi amplamente utilizada no presente trabalho. Nele, cada coluna corresponde a um atributo e cada linha representa uma observação (registro). Ele é construído sobre a biblioteca *NumPy* (abreviação para *Numerical Python*), que oferece suporte a operações vetorizadas e eficientes em termos de computação numérica.

O *NumPy*[10] (versão 1.25.0) é uma biblioteca de código aberto que é utilizada em quase todos os campos da ciência e engenharia. Ela trabalha com uma estrutura de dados chamada *ndarray* (um objeto de matriz *n*-dimensional homogêneo) e métodos eficientes de alto nível que permitem manipulá-los. Além disso, é considerada o padrão universal para se trabalhar com dados numéricos em *Python*. Sua API é utilizada extensivamente em outras bibliotecas como a

---

[5] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

[6] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

[7] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html#sklearn.svm.SVR

[8] https://keras.io/

[9] https://pandas.pydata.org/

[10] https://numpy.org/

scikit-learn e *Matplotlib* (NUMPY, 2023). Todas as manipulações nos dados, que incluem transformações, filtragem, ordenação, união de conjuntos de dados, foram realizadas com o auxílio dessas duas bibliotecas.

4.8.3 Matplotlib

Matplotlib[11] (versão 3.7.1) é uma biblioteca computacional bem abrangente utilizada para criar gráficos estáticos, animados e visualizações interativas utilizando *Python* em uma ampla variedade de formatos e estilos. Ela é altamente integrada com outras bibliotecas de análise de dados como as citadas anteriormente (*scikit-learn*, *keras* e *pandas*). Nela, há uma ampla gama de opções para a personalização dos gráficos, permitindo alterar cada elemento como cores, títulos, rótulos, tamanho da fonte, entre outros. Com poucas exceções, os gráficos do presente trabalho foram construídos utilizando a API dessa biblioteca.

[11] https://matplotlib.org/

5 RESULTADOS E DISCUSSÃO

5.1 Banco de dados

Das 42 referências diferentes presentes no Banco de Dados Original, a maioria delas (23) foi publicada entre os anos de 2020 e 2022 como pode ser visto na Figura 13. Esse fato ilustra que a maioria dos dados contidos no banco de dados construído são atuais. As referências mais antigas foram publicadas em 2009.

Figura 13 – Frequência do ano de publicação dos trabalhos de referência do banco de dados original.

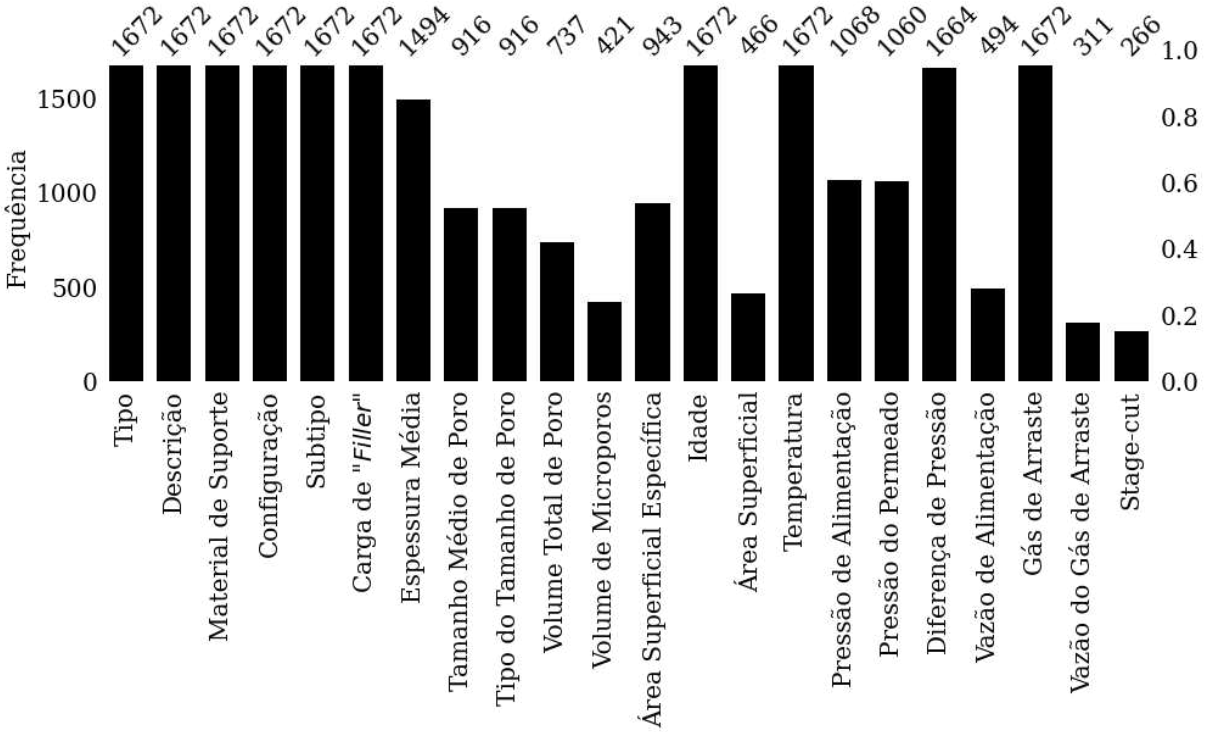
Fonte: Elaborada pelo autor.

Na Figura 14 é mostrada a quantidade de registros não nulos para as variáveis morfológicas e de processo no Banco de Dados Original após as considerações. Os atributos restantes que não aparecem na Figura 14 não possuem valores ausentes, isto é, possuem os 1672 registros de permeação. Vale salientar que cada registro equivale a um ensaio de permeação em uma determinada membrana. Alguns atributos foram dificilmente encontrados como o Volume de Microporos, Área Superficial, Vazão de Alimentação, Vazão do Gás de Arraste e *Stage-cut*. O Material de Suporte, Configuração, Carga de *Filler*, Idade e Gás de Arraste possuíram todos os registros devido as considerações anteriormente relatadas. A Temperatura foi o único atributo de processo que todas as referências analisadas forneceram, mesmo indicando que era uma temperatura ambiente (*Room Temperature*).

Foram coletados dados referentes a 11 gases: He, H_2, CO_2, O_2, H_2S, CO, N_2, CH_4, C_2H_4, C_2H_6 e SF_6. O banco de dados é constituído por 1140 registros referentes a gases puros e 532 registros referentes a gases em misturas (511 misturas binárias e 21 misturas ternárias). Na Figura 15 é mostrado os detalhes sobre a frequência desses registros no Banco de Dados Original. Registros sobre o CO_2 e, como esperado, sobre o CH_4 foram encontrados em maior

quantidade em ambos os contextos, na forma pura e em mistura. Vale salientar que a mineração dos dados teve foco em trabalhos que estudaram a capacidade de permeação do CH_4. Mesmo em menor quantidade, o H_2, N_2 e He também estiveram muito presentes com 203, 198 e 108 registros na sua forma pura e 67, 88 e 5 registros em mistura, respectivamente. Os outros gases somam um total de 130 registros em ambos os contextos e o H_2S só foi encontrado em mistura. Nenhum hidrocarboneto com mais de um carbono foi encontrado em mistura, fato que colabora com a aproximação realizada para a definição da variável Fração do Gás.

Figura 14 – Frequência dos valores não nulos para os atributos morfológicos e de processo (com exceção da Fração do Gás) do Banco de Dados Original.

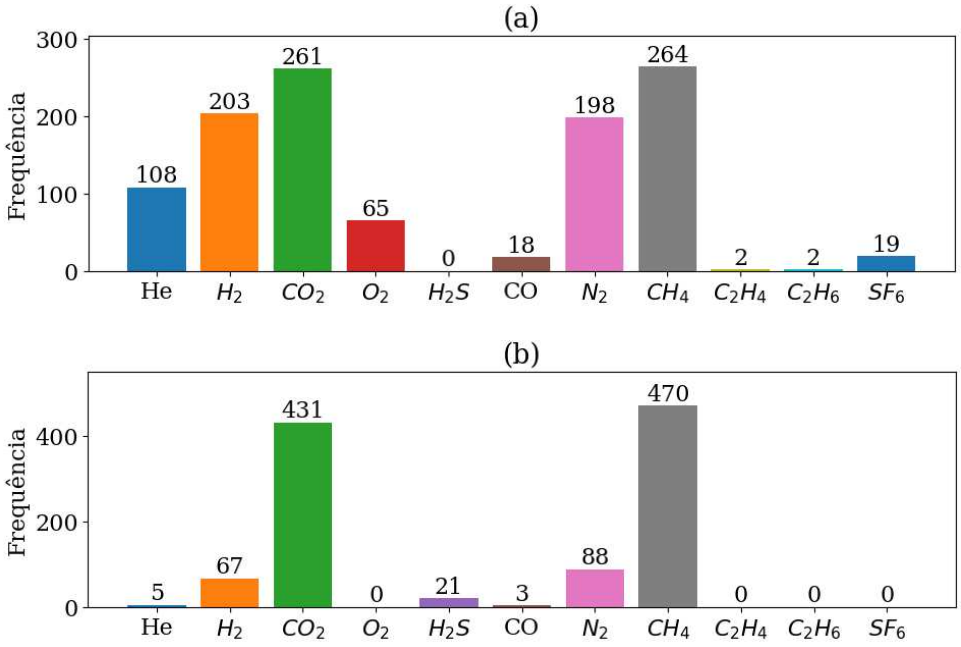


Fonte: Elaborada pelo autor.

Ao todo, o atributo Tipo foi dividido em 5 classes de materiais diferentes: polímero, zeólita, MMM, *Carbon Molecular Sieves* (CMS), MOF e cerâmica. Essa classificação foi baseada no trabalho de Ismail, Khulbe e Matsuura (2015). Como pode ser visto na Figura 16, membranas poliméricas foram mineradas com mais frequência, com 603 registros, seguida de membranas zeolíticas, compósitas (MMM) e de MOFs, com 506, 375 e 143 registros respectivamente. Os registros de tipo “CMS” e “cerâmica” são provenientes de uma referência bibliográfica, cada uma, e em razão disso elas estão em menor quantidade no banco de dados. Os histogramas e gráficos de barras para os atributos morfológicos e de processo do Banco de

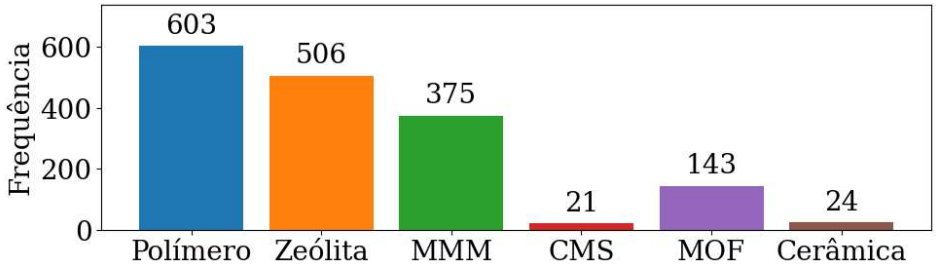
Dados Original se encontram no Apêndice A. Devido à grande variabilidade dos valores/colunas, os atributos “Descrição”, “Subtipo” e “Fração do Gás” não foram analisados na Figura A1 (Apêndice A).

Figura 15 – Quantidade de registros para cada gás (a) na forma pura e (b) em mistura presente no Banco de Dados Original. Os gases estão em ordem crescente ao respectivo diâmetro cinético.



Fonte: Elaborada pelo autor.

Figura 16 – Quantidade de registros para cada do atributo Tipo no Banco de Dados Original.



Fonte: Elaborada pelo autor.

Para a modelagem, foi construído um banco de dados[12] diferente, baseado nos atributos do Banco de Dados Original, porém com um novo formato. Neste novo banco de dados,

[12] https://github.com/tenoriolms/databank_membranes/blob/main/databank_used_in_modeling.xlsx

somente os atributos de performance e o atributo de processo Fração do Gás (Tabela 3) não foram aproveitados da maneira como estavam. Cada registro do novo banco de dados referiu-se obrigatoriamente a apenas um Gás de Alimentação (GA) e sua respectiva Fração do Gás. No Banco de Dados Original um único registro pode conter informações de mais de um GA, contanto que ele esteja se referindo a uma mistura. Sendo assim, as 39 colunas anteriormente referentes a cada GA (no banco de dados: "*x_*", "*Py_*" e "*Pe_*") foram simplificadas, sendo substituídas por seis novas colunas para representar:

i. A fórmula química do gás na corrente de alimentação;
ii. O atributo Fração do Gás do respectivo GA do registro;
iii. A performance, que será representada pela Permeabilidade ou pela Permeância;
iv. As propriedades físicas do respectivo GA: diâmetro cinético, massa molar e polarizabilidade (3 colunas/atributos).

Além disso, mais uma coluna foi adicionada para representar o diâmetro cinético do Gás de Arraste. O diâmetro cinético está relacionado com a esfera de influência para colisões moleculares e, de certa forma, com a resistência de permeação. A massa molar está relacionada com a difusão livre desses gases (JOOS; 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). Na Tabela 8 é apresentada a definição dos atributos adicionados. Em suma, essas modificações permitiram a criação de um banco de dados mais simplificado que será utilizado para o treino e teste dos modelos de AM.

Com esse novo formato, tendo como critério a disponibilidade dos dados e a capacidade de generalização, somente alguns atributos foram selecionados para compor a lista de atributos de entrada dos modelos. Os atributos Volume de Microporos, Área Superficial, Vazão de Alimentação, Vazão do Gás de Arraste e *Stage-cut* não foram escolhidos devido à baixa quantidade de registros fornecidos. Além disso, assim como os atributos informativos (Tabela 3), a Descrição e o Subtipo não foram escolhidos pois são atributos morfológicos muito específicos. Os atributos Pressão de Alimentação e Pressão do Permeado não foram escolhidos a fim de evitar redundância conceitual com o atributo Diferença de Pressão e por possuírem menos registros.

Tabela 8 – Definição dos atributos adicionados ao novo banco de dados. Os atributos morfológicos, de processo (com exceção da Fração do Gás) e informativos do Banco de Dados Original foram mantidos.

Categoria	Atributo[1]	Descrição do atributo
Processo	"*gas*" (Gás de Alimentação)	Fórmula química do gás na corrente de alimentação ao qual o registro se refere.
	"*x*" (Fração do Gás)	Fração molar/volumétrica do Gás de Alimentação.
	"gases_kinetic_diameter" (Diâmetro Cinético do GA)	Diâmetro cinético do Gás de Alimentação (em Å).
	"gases_molar_mass" (Massa Molar do GA)	Massa molar do Gás de Alimentação ($g \cdot mol^{-1}$).
	"gases_polarizability" (Polarizabilidade do GA)	Polarizabilidade do Gás de Alimentação (em $cm^{-3} \cdot 10^{-25}$).
	"sweep_gas_KD" (Diâmetro Cinético do Gás de Arraste)	Diâmetro cinético do Gás de Arraste (em Å).
Performance	"*Py*" (Logaritmo da Permeabilidade)	Logaritmo da permeabilidade do Gás de Alimentação.

[1]O nome entre aspas se refere a como são chamados os atributos/colunas no banco de dados e o nome em parênteses se refere a como o atributo será chamado no presente trabalho. GA = Gás de Alimentação. Fonte: Elaborada pelo autor.

Após a retirada desses atributos, foram eliminados os registros que portavam algum valor ausente em qualquer atributo ou que fossem repetidos (linhas duplicadas). Após esse processo, não se manteve nenhum registro relativo ao Tipo MMM. Por esse motivo, o atributo Carga de *Filler* também foi retirado por ser instrutiva apenas para esse tipo de membrana e, portanto, evitar um "ruído" evidente nas fases de treinamento. Dessa forma, os atributos de entrada selecionado foram:

- Tipo;
- Material de Suporte;
- Configuração;
- Espessura Média;
- Tamanho Médio de Poro;
- Tipo do Tamanho de Poro;

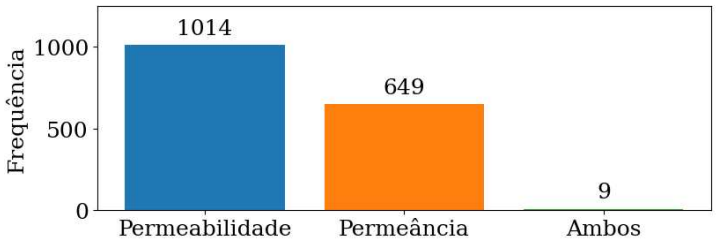
- Volume Total de Poro;
- Área Superficial Específica;
- Idade;
- Temperatura;
- Diferença de Pressão;
- Diâmetro Cinético do Gás de Arraste;
- Diâmetro Cinético do Gás de Alimentação;
- Massa Molar do Gás de Alimentação;
- Polarizabilidade do Gás de Alimentação;
- Fração do Gás.

Empiricamente, sabe-se que alguns atributos como a configuração, o Material de Suporte ou o Diâmetro Cinético do Gás de Arraste são irrelevantes frente a Espessura Média ou o Diâmetro Cinético do Gás de Alimentação. A manutenção delas como atributos de entrada também tem a finalidade de investigar se os modelos conseguirão reconhecer os padrões corretamente de forma que as importâncias dessas variáveis correspondam com a realidade.

A performance das membranas (atributo alvo) foi representada pelo logaritmo da permeabilidade (em barrer), assim como faz Barnett et al. (2020). A permeabilidade foi escolhida devido a sua maior frequência de disponibilização nos trabalhos de referência, como pode ser observado na Figura 17, a qual mostra que 1023 registros do Banco de Dados Original (aproximadamente 61%) fornecem originalmente a permeabilidade como variável de performance de permeação em comparação com a permeância. É importante destacar que, como mencionado anteriormente, as permeabilidades foram calculadas pela Equação (2.3) nas referências que não a forneceram, mas disponibilizam a Espessura Média. Além disso, as permeabilidades foram logaritmizadas da base 10. Essa transformação é uma das mais aplicadas em análises de regressão e é particularmente útil quando os dados utilizados possuem um desvio padrão muito alto em relação à média. Nesse contexto, a logaritmização amortece a variabilidade dos dados, pode reduzir a assimetria da sua distribuição e diminuir a heterocedasticidade (quando os resíduos possuem um perfil não constante ao longo das observações) (CHATTERJEE; HADI, 2012; KUTNER *et al.*, 2005). Como pode ser visto na Figura 18, os valores da permeabilidade possuem uma variabilidade muito ampla, com um mínimo e máximo de $1{,}7\times10^{-2}$ e $1{,}8\times10^{6}$ barrer, respectivamente. Percebe-se também que essa distribuição se mantém quando se considera apenas um subconjunto de dados pertencente a um nicho específico (polimérica, zeólitas, MMM). Essa amplitude de valores é comum no campo

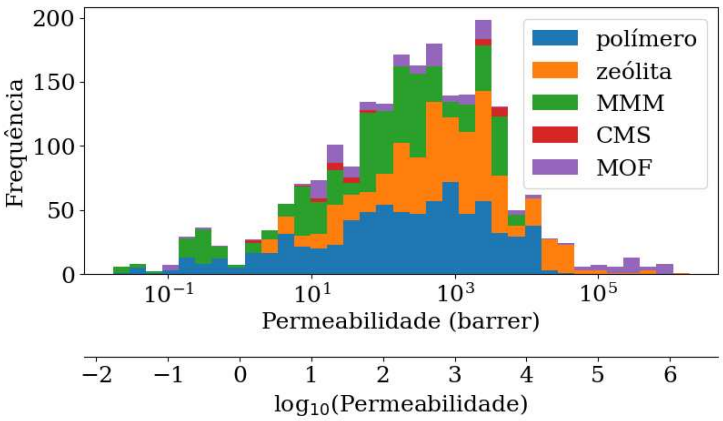
de estudo dos PSGM e essa característica é ampliada devido a heterogeneidade do banco de dados, que além de possuir 6 classes para o atributo Tipo, considera 13 Gases de Alimentação diferentes. Por essa razão, o atributo de performance escolhido (permeabilidade) para representar o atributo alvo sofreu uma transformação logarítmica (BARNETT *et al.*, 2020), resultando em uma distribuição aproximadamente gaussiana no seu histograma (Figura 18). Antes de calcular as métricas de desempenho do modelo, o atributo alvo foi "despadronizado", isto é, a transformação inversa do *z-score* – Equação (4.4) – foi aplicada e foram analisadas as métricas referentes ao logaritmo da permeabilidade. Um histograma dos valores da permeabilidade do Banco de Dados Original de acordo com o gás de alimentação pode ser visto na Figura A2 (Apêndice A).

Figura 17 – Frequência de cada classe do atributo Tipo de Performance Fornecida no Banco de Dados Original.



Fonte: Elaborada pelo autor.

Figura 18 – Histograma dos valores da permeabilidade do Banco de Dados Original de acordo com o atributo Tipo. Os registros referentes a classe "cerâmica" não forneceram a permeabilidade nem a Espessura Média para o seu cálculo.



Fonte: Elaborada pelo autor.

5.2 Análise exploratória do banco de dados

Todo o processo realizado nas etapas descritas anteriormente objetivou a construção do banco de dados que foi utilizado para treinamento dos algoritmos de AM. O mesmo possui um total de 692 registros e 17 atributos (incluindo o atributo alvo), sem registros duplicados ou valores ausentes. Como pode ser observado na Figura 19, dos três valores únicos do atributo Tipo presentes no banco de dados, 48% correspondem às membranas poliméricas, 34% às membranas zeolíticas e 18% às membranas de MOF. Membranas do Tipo "MMM", "CMS" e "Cerâmica" estão presentes em vários registros (420) do Banco de Dados Original, porém não foram selecionadas devido ao processo de poda dos atributos. Por questão de definição, nenhum registro da classe das MMMs apresentou valores para os atributos Tamanho Médio de Poro, Tipo do Tamanho de Poro, Volume Total de Poro e Área Superficial Específica. Nessas membranas o filme é fabricado por um material híbrido ("mistura" de dois ou mais materiais) e as características fornecidas pela literatura no campo de estudo (e.g. área BET, tamanho de poro, volume de poro) raramente se referem a esse compósito, mas geralmente aos materiais precedentes. Por essa razão, as MMM não foram analisadas no presente trabalho. Na Tabela B1 (Apêndice B) são apresentadas as referências utilizadas para a construção do banco de dados, além dos seus respectivos materiais e gases testados. Ao todo, 31 materiais diferentes de 19 referências bibliográficas estão presentes no banco de dados. Alguns trabalhos estudam mais de uma membrana, pois, ao mudar a metodologia de síntese, um material pode originar membranas com diferentes características (e.g. diferentes espessuras ou áreas superficiais).

Foram encontrados 4 tipos de suporte de membranas apenas nas membranas zeolíticas e de MOF que foram classificados em: "alumina", "α-alumina", "silicalite-1" e "óxido de alumínio anodizado" (Figura 19). A alumina foi o material mais comum a ser utilizado como suporte – 90% das membranas que utilizaram suporte, a utilizaram – e, como ficou evidente, a alumina acabou sendo dividida em duas categorias no momento de mineração dos dados ("alumina" e "α-alumina"), em que os registros cujas referencias não especificaram a fase da alumina tiveram o seu atributo Material de Suporte considerado como "alumina". As membranas poliméricas não possuíram material de suporte (todas enquadradas na classificação "D"), o que é justificável, uma vez que nesse contexto não há necessidade de tais materiais. A mulita foi encontrada em vários registros do Banco de Dados Original (Figura A1 – Apêndice A), porém devido ao processo de poda dos atributos eles não foram selecionados.

As configurações presentes no banco de dados foram a "tubular" e a "plana". Todas as membranas zeolíticas foram encontradas em uma configuração tubular enquanto as membranas

dos outros tipos foram encontradas (todas) em uma configuração plana. A configuração de fibra oca apareceu em alguns registros do Banco de Dados Original referentes a membranas de MOF (Figura A1 – Apêndice A), porém não apareceu devido ao processo de poda dos atributos.

Os registros das espessuras médias presentes no banco de dados variaram de 0,01 a 248 μm (Figura 19). As espessuras das membranas poliméricas apresentaram as maiores dimensões em comparação com a dos outros dois tipos, que de fato, devido aos métodos de síntese relacionados, normalmente possuem menores espessuras. Durante a prospecção dos dados, ficou evidente que esse atributo se mostrou como uma das principais características das membranas. Consultando a Equação (2.1), entende-se que o desempenho de separação desses filmes está correlacionado com a sua respectiva espessura e portanto, mesmo que a tendência entre permeabilidade e a espessura tenham relações de intensidades diferentes a depender do material de síntese, 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; AWASTHI, 2022).

Os registros do Tamanho Médio de Poro presentes no banco de dados variaram de 0,35 a 6,79 nm (Figura 19). Pela natureza do fenômeno, sabemos que o tamanho de poro também se mostra como um fator determinante para a permeação dos gases nesses materiais. No contexto das membranas em que o peneiramento molecular é predominante (o que irá ocorrer quando os poros das membranas estiverem na faixa de 0,5–2 nm) moléculas maiores que o tamanho do poro da membrana serão rejeitadas, enquanto moléculas menores irão passar (SADRZADEH; MOHAMMADI, 2019; SADRZADEH; REZAKAZEMI; MOHAMMADI, 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; MOHAMMADI, 2019). Com base no atributo Tipo de Tamanho de Poro, 63% dos registros de tamanho de poro foram obtidos pelos autores por DTP (os gráficos de distribuição de poros), dos quais se encontram todas as membranas poliméricas. Para as membranas zeolíticas essa variável foi obtida de forma majoritariamente teórica (valores também fornecidos pelos autores), enquanto para as membranas de MOF o tamanho de poro foi obtido majoritariamente por DTP.

Os registros relativos ao volume de poro variaram de 0,11 a 1,6 $cm^3 \cdot g^{-1}$ (Figura 19). Os volumes de poro das membranas poliméricas tenderam a ser maiores do que o das membranas de zeólitas, enquanto se observa um perfil disperso para as membranas de MOF. O volume de

poro é comumente estimado como um derivado da quantidade de moléculas de adsorbato que um determinado material consegue adsorver, assumindo que os poros estão totalmente preenchidos com o adsorbato líquido que é geralmente N_2 ou H_2 através de uma técnica conhecida como BET. Considerando a dimensão teórica dos átomos destes adsorvatos, também é possível estimar a área superficial do sólido. A área superficial BET das membranas avaliadas variaram de 165 a 1718 $m^2 \cdot g^{-1}$ (Figura 19). Assim como o tamanho e volume de poro, as membranas de MOF apresentaram as maiores áreas superficiais. Sabe-se que essa propriedade tem relação positiva com a capacidade adsortiva do material, o que pode indicar resistência à permeação dependendo das suas respectivas propriedades químicas.

Os registros da Idade no banco de dados variaram de 0 a 2088 dias (Figura 19). Não houve registros de membranas zeolíticas ou de MOF envelhecidas. 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. O envelhecimento dificulta na comercialização dessas membranas para a separação de gases e sua intensidade irá depender de múltiplos fatores que incluem a temperatura, as condições de armazenamento, espessura e o pré-tratamento da membrana (ALBERTO *et al.*, 2018; AMEEN *et al.*, 2021; BERNARDO, P. *et al.*, 2017). Desta forma, a idade da membrana é um fator importante a ser considerado na previsão da performance.

As temperaturas registradas no banco de dados variaram entre 273 e 453 K (Figura 19). Provavelmente por motivos de estabilidade térmica, membranas poliméricas não foram testadas a temperaturas acima de 328 K. Os registros do atributo Diferença de Pressão variaram entre 0 e 30 bar (Figura 19). Apenas membranas zeolíticas foram submetidas a diferenças de pressão acima de 9 bar. Somente dois gases de arraste foram utilizados na corrente de permeado, o Ar e o He. A maioria dos registros (71%) não utilizaram gás de arraste ao passo que, no contexto daqueles que utilizaram, 79% optaram pelo gás Ar. O N_2 foi utilizado em alguns registros do Banco de Dados Original (Figura A1 – Apêndice A), mas não apareceu no banco de dados devido a poda dos atributos realizada anteriormente.

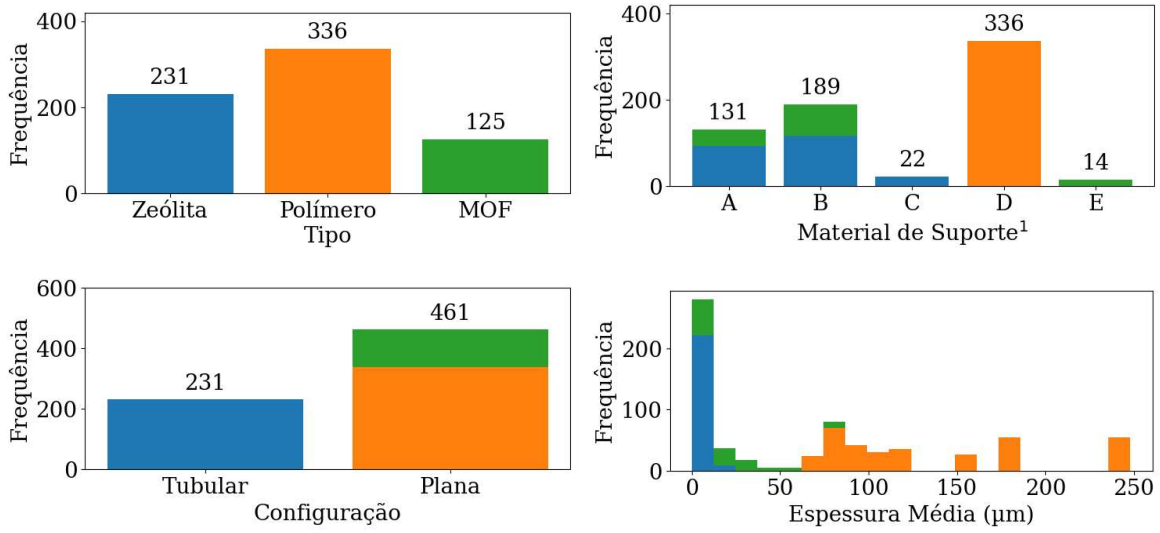
As frações dos gases no banco de dados variaram de 0,14 a 1, sendo que a maioria dos gases esteve presente na sua forma pura. Aqueles presentes em misturas ($x<1$) são provenientes de misturas binárias. Na Figura 19, observa-se apenas frações equimolares nos registros referentes às membranas zeolíticas e de MOF, enquanto outras diferentes condições podem ser vistas sendo testadas nas membranas poliméricas. Como pode ser observado na Figura 20, 66%

(460) dos registros do banco de dados foram referentes aos gases puros. Ao todo, 9 gases de alimentação estão presentes no banco de dados após a poda: He, H_2, CO_2, O_2, N_2, CH_4, eteno, etano e SF_6. Assim como no Banco de Dados Original, os gases CH_4 e CO_2 estão presentes majoritariamente com um total de 163 e 211 registros, respectivamente. Em menor quantidade, porém ainda assim de forma considerável, estão os gases H_2, N_2, He e O_2 com um total de 119, 106, 42 e 41 registros, respectivamente. Apenas os gases CO_2, CH_4, H_2 e He foram encontrados em situação de mistura com 106, 77, 26 e 23 registros, respectivamente. Os restantes dos gases (eteno, etano e SF_6) foram encontrados em poucas referências, porém foram mantidos com o objetivo de fornecer ao modelo de AM um caráter mais generalista. Nenhum registro referente ao H_2S e o CO foi selecionado após a poda dos atributos de entrada e dos valores ausentes.

Os registros do logaritmo da permeabilidade variaram entre –0,88 e 6,02 (equivalente a 0,13 e $1{,}05 \cdot 10^6$ barrer, respectivamente), com os três tipos de membrana apresentando um perfil de distribuição semelhante para essa variável. As membranas de MOF se mostraram mais dispersas, apresentando as maiores e as menores permeabilidades.

Figura 19 – Distribuição de registros dos atributos do banco de dados utilizado nas modelagens de acordo com seus respectivos valores categóricos ou numéricos. As cores estão classificadas de acordo com o atributo Tipo de cada registro. Legenda: ■ Zeólita; ■ Polímero; ■ MOF. [1] A = “alumina”; B = “α-alumina”; C = “silicalite-1”; D = “Nenhum”; E = “Óxido de alumínio anodizado”.

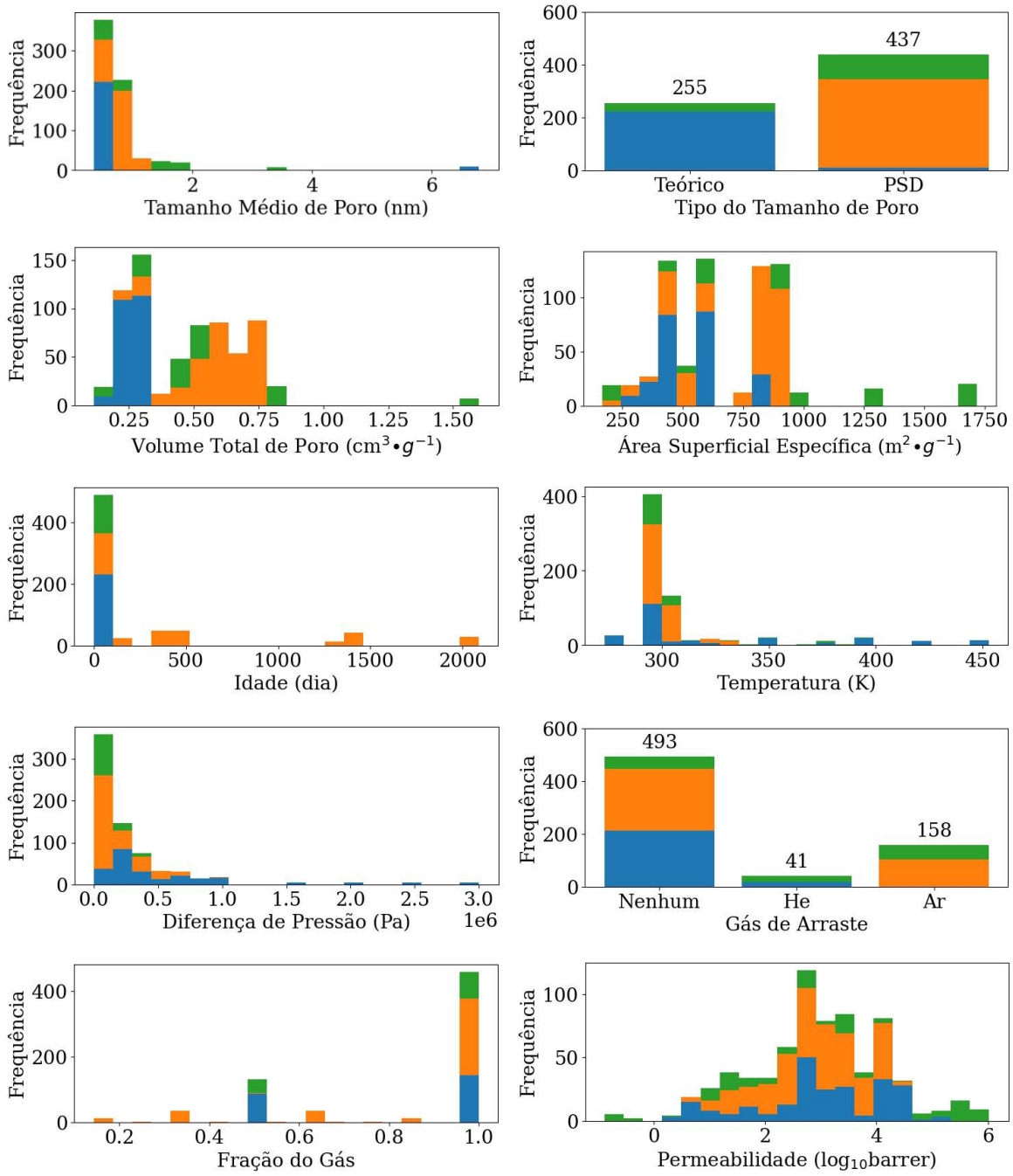
(continua)



Fonte: Elaborada pelo autor.

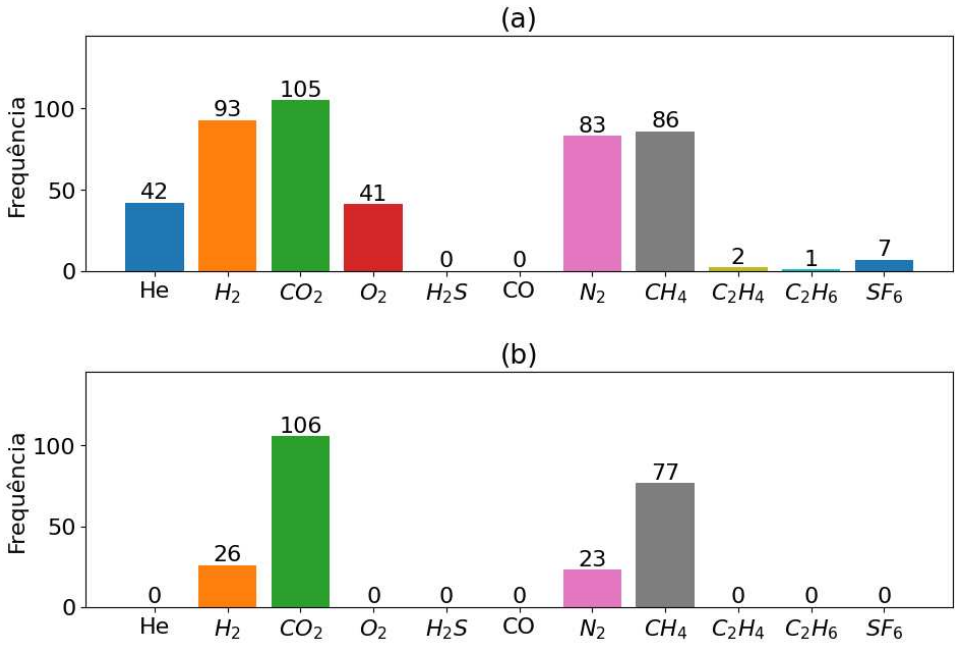
Figura 19 – Distribuição de registros dos atributos do banco de dados utilizado nas modelagens de acordo com seus respectivos valores categóricos ou numéricos. As cores estão classificadas de acordo com o atributo Tipo de cada registro. Legenda: ■ Zeólita; ■ Polímero; ■ MOF. [1] A = “alumina”; B = “α-alumina”; C = “silicalite-1”; D = “Nenhum”; E = “Óxido de alumínio anodizado”.

(conclusão)



Fonte: Elaborada pelo autor.

Figura 20 – Quantidade de registros para cada gás (a) na forma pura e (b) em mistura presente no banco de dados. Os gases estão em ordem crescente ao respectivo diâmetro cinético.



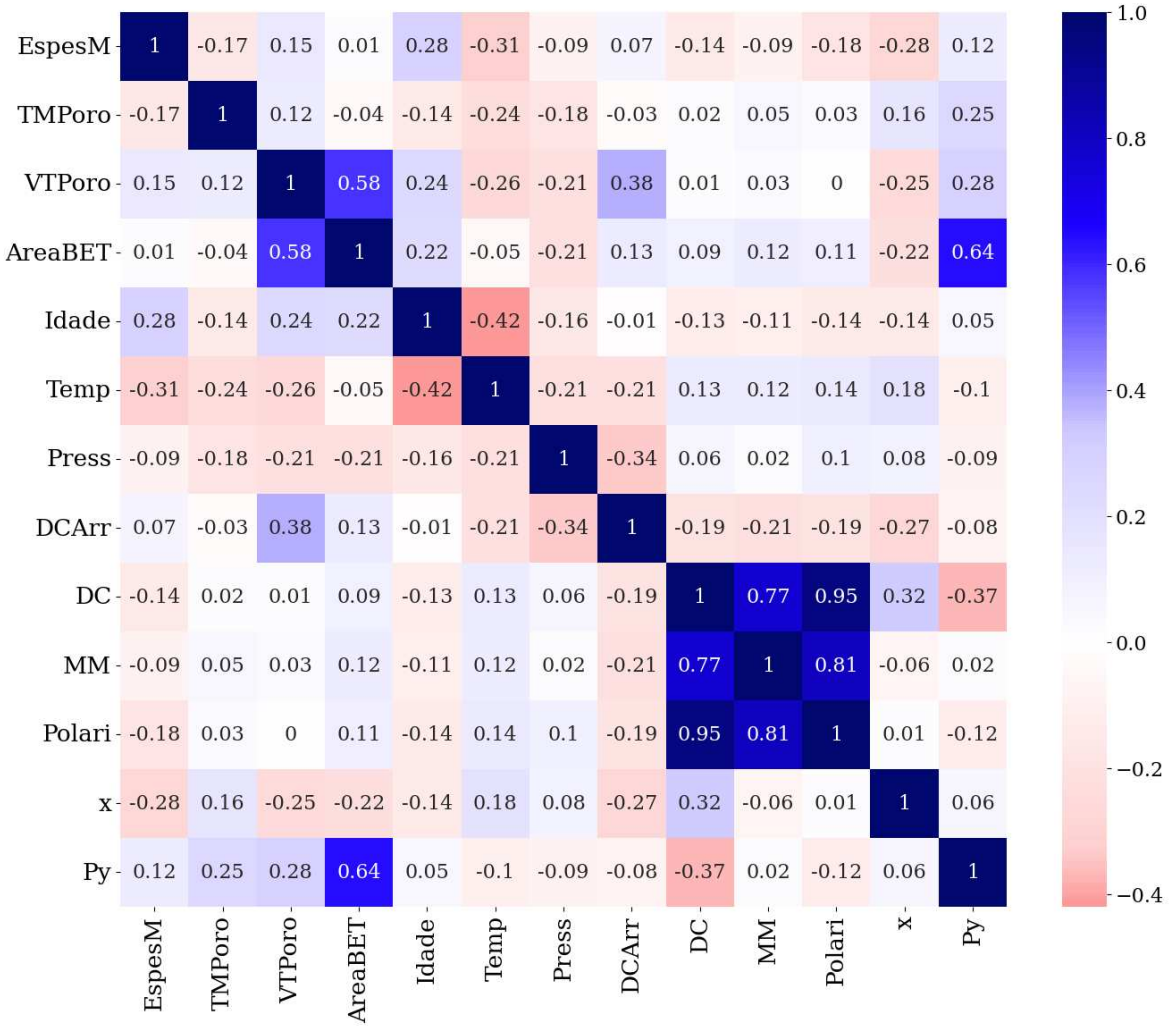
Fonte: Elaborada pelo autor.

A Figura 21 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, Shansheng *et al.*, 2005). Se ρ for positivo as variáveis relacionadas são diretamente proporcionais e se for negativo, as variáveis são inversamente proporcionais. Caso o coeficiente de Pearson for muito alto e ultrapassar ±0,8 ou ±0,9 as variáveis podem ser consideradas fortemente correlacionadas entre si e, geralmente, alguma delas é retirada da modelagem (MUKAKA, 2012; PAN *et al.*, 2022; WERE *et al.*, 2015).

Sendo assim, como pode ser observado na Figura 21, com exceção do caso entre a polarizabilidade e o diâmetro cinético, todos os atributos possuíram uma correlação linear abaixo de 0,9. A maioria das variáveis não possui correlação linear entre si, indicando uma relação complexa entre elas. Uma correlação moderada positiva pode ser vista entre a área superficial e a permeabilidade ($\rho_{xy} = 0,64$) e entre a área superficial e o volume de poro ($\rho_{xy} = 0,58$). Em contrapartida, as variáveis referentes às propriedades físicas do gás estão bastante positivamente correlacionadas entre si ($\rho_{xy} > 0,77$), principalmente a polarizabilidade e o diâmetro cinético com um $\rho_{xy} = 0,95$. Apesar disso, a priori, decidiu-se manter essas duas variáveis juntas como atributos de entrada com o intuito de investigar qual delas será mais

relevante no processo de predição, visto que ambas carregam informações diferentes para o modelo.

Figura 21 – Matriz de correlação de Pearson entre os atributos numéricos do banco de dados. LEGENDA: “EspesM” = Espessura Média; “TMPoro” = Tamanho Médio de Poro; “VTPoro” = Volume Total de Poro; “AreaBET” = Área Superficial Específica; “Idade” = Idade; “Temp” = Temperatura; “Press” = Diferença de Pressão; “DCArr” = Diâmetro Cinético do Gás de Arraste; “DC” = Diâmetro Cinético do Gás de Alimentação; “MM” = Massa Molar do Gás de Alimentação; “Polari” = Polarizabilidade do Gás de Alimentação; “x” = Fração do Gás; “Py” = Permeabilidade.

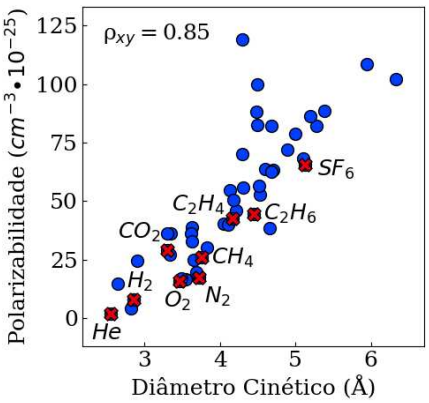


Fonte: Elaborada pelo autor.

A polarizabilidade, dentre todos os atributos, é a única variável que possui relação direta com o aspecto químico do processo de permeação, isto é, com a intensidade das interações intermoleculares do gás, enquanto o diâmetro cinético retrata uma dimensão mais associada ao tamanho. Além disso, como pode ser visualizado na Figura 22, moléculas com maior polarizabilidade tendem a ter maiores diâmetros cinéticos, uma vez que uma nuvem eletrônica

mais deformável resulta em uma maior extensão no espaço. Essa relação se torna mais complexa e menos linear ao se analisar grandes moléculas, uma vez que as diferentes possibilidades de conformação e composição química tornam-se mais relevantes. O fato de o CCP ser igual a 0,85 (Figura 22) para o conjunto de dados mais amplo reafirma essa não linearidade. Em razão disso, após a seleção de modelo e metodologias ótimos, uma avaliação individual sobre a influência dessas variáveis na predição será realizada. A matriz dos gráficos de dispersão para todos os atributos pode ser visualizada na Figura B1, no apêndice B. Na Figura B2 é mostrada a matriz dos coeficientes de correlação de Spearman entre as variáveis do banco de dados. Esse coeficiente é uma medida não paramétrica da monotonicidade entre dois conjuntos de dados. Assim como o Coeficiente de Pearson, valores nulos implicam a inexistência de correlação monotônica e valores extremos (1 ou –1) implicam uma correlação monotônica exata. A relação é diretamente proporcional quando o valor for positivo e inversamente proporcional quando for negativo (TRIOLA, 2010). Pode-se observar que as relações monotônicas entre os atributos são similares às suas relações lineares.

Figura 22 – Gráfico de dispersão entre a polarizabilidade e o diâmetro cinético de vários gases (LI, Jian-Rong; KUPPLER; ZHOU, 2009). Os pontos marcados em vermelho são referentes aos gases presentes no banco de dados.



Fonte: Elaborada pelo autor.

5.3 Modelagem

5.3.1 Florestas Aleatórias

O modelo de AM Florestas Aleatórias (FA) foi o primeiro a ser analisado para a predição da performance. Primeiramente, os hiperparâmetros do modelo foram otimizados utilizando as sete funções objetivo diferentes ($F_{0–6}$ na Tabela 7) e 500 passos iterativos (tentativas) foram

realizados para cada uma. Vale salientar que, com exceção de F_0, as funções objetivo a serem minimizadas utilizam métricas do conjunto de teste e treino da validação cruzada. Ao todo, foram otimizados sete hiperparâmetros cuja descrição, faixa de procura e valores ótimos são apresentados na Tabela 9. O tempo de execução e o melhor valor encontrado para cada caso pode ser visto na Tabela C1 (Apêndice C).

A escolha da faixa utilizada para a busca dos parâmetros foi inicialmente guiada por curvas de validação. Em seguida, em todos os contextos testados, a faixa foi ajustada de forma que o valor ótimo dos parâmetros não se encontre nos limites mínimos ou máximos delimitados. Isto é, a faixa de procura teve a finalidade de "orientar" o algoritmo de busca e não de limitar os valores que os parâmetros podem ter. Dessa forma, percebeu-se que os valores das funções objetivo eram minimizadas ao máximo e que a influência de cada uma delas nos resultados do modelo foi mais perceptível.

Tabela 9 – Hiperparâmetros avaliados durante a otimização do modelo Florestas Aleatórias para a predição da performance das membranas.

Hiperparâmetro	**Faixa de procura** [1]	**Valores ótimos**						
		F_0	**F_1**	**F_2**	**F_3**	**F_4**	**F_5**	**F_6**
Número de Árvores	1 – 400 (1)	322	397	204	198	245	167	19
Critério	SE e AE[2]	SE	SE	SE	SE	SE	SE	SE
Profundidade Máxima	2 – 50 (1)	44	21	21	46	27	19	17
Mínimo de Folhas	1 – 4 (1)	1	1	1	1	1	1	1
Número de Atributos	1 – 16 (1)	2	1	1	2	5	5	5
Tamanho da Amostra	50 – 100 % do tamanho do conjunto de treino (1%)	100%	100%	98%	99%	100%	98%	99%
K-fold	2 – 25 (1)	-	17	19	6	24	15	19

[1] O valor entre parênteses corresponde ao passo (*step*) definido para cada faixa. [2] SE = Erro quadrático (em inglês, *Squared Error*); AE = Erro Absoluto (em inglês, *Absolute Error*). Fonte: Elaborada pelo autor.

Na Tabela 10 são apresentadas as métricas de desempenho do modelo FA ao utilizar os parâmetros otimizados de cada função objetivo. Procura-se obter sempre um modelo com R^2 alto e um MAPE baixo. Em geral, pode-se perceber que o coeficiente de determinação dos conjuntos de treino e teste se encontram em torno de 0,99 e 0,96, respectivamente, e o MAPE, que pode ser interpretado como um erro relativo médio, teve valores em torno de 2 e 5 % para

o conjunto de treino e teste, respectivamente. Em todos os casos, pode-se observar um sobreajuste do modelo, caracterizado pelo bom desempenho no conjunto de treino que não é passado para o conjunto de teste. Além disso, nesse caso, a adição de um termo de aproximação nas funções F_3 e F_6 não foi suficiente nem para evitá-lo, nem para diminuir a diferença entre as métricas quando se compara com as outras funções (sem termo de aproximação). Apesar disso, o modelo se mostrou adequado e com ótimos resultados para prever a performance de permeação de diferentes gases em variados tipos de membranas.

Tabela 10 – Métricas de desempenho (MAPE e R^2) do modelo FA otimizado por diferentes funções objetivo utilizando o conjunto de treino e teste.

Métrica[1]	**F_0**	**F_1**	**F_2**	**F_3**	**F_4**	**F_5**	**F_6**
R^2 treino	0,9966	0,9964	0,9964	0,9965	0,9960	0,9957	0,9950
R^2 teste	0,9671	0,9590	0,9621	0,9668	0,9723	0,9722	0,9642
ΔR^2 ($\cdot 10^{-2}$)	2,95	3,74	3,43	2,97	2,37	2,35	3,08
MAPE treino (%)	1,9893	2,2627	2,4622	2,0956	1,8930	1,8887	1,8683
MAPE teste (%)	5,2096	5,9598	6,0820	5,3727	4,4018	4,4432	4,4427
Δ MAPE (%)	3,2203	3,6971	3,6198	3,2771	2,5088	2,5545	2,5744

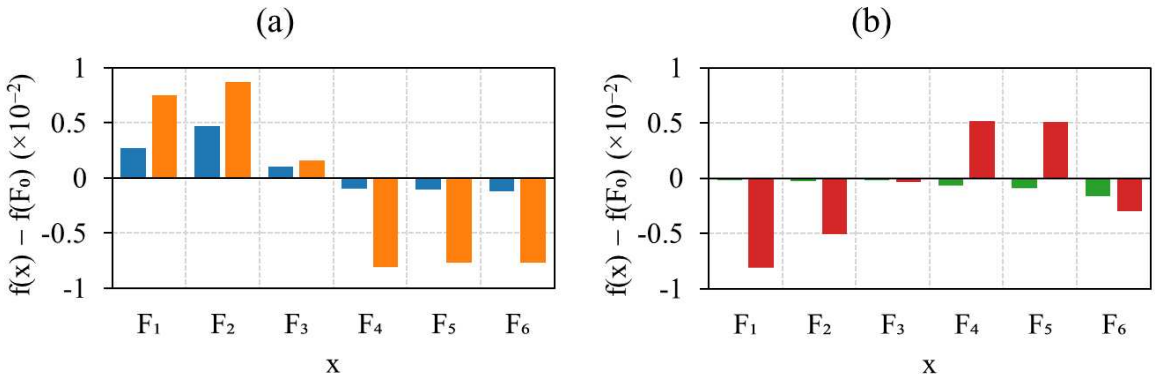
[1] Δ = |teste – treino|. Fonte: Elaborada pelo autor.

As diferenças nas métricas foram pequenas, porém a seleção dos hiperparâmetros ótimos foi bastante determinada pela função objetivo como pode-se observar na Tabela 9. O Número de Árvores, a Profundidade Máxima e o Número de Atributos foram os parâmetros que mais variaram. As funções F_0, F_1 e F_4, que fazem uso de métricas exclusivamente de conjunto de treino, possuem os maiores Números de Árvores, com 322, 397 e 245, respectivamente. As funções que fazem uso do MAPE apresentaram os maiores valores para o Número de Atributos, com o valor de 5. As Profundidades Máximas das árvores variaram entre 21 e 46 níveis. Os *k*-folds ótimos, que são relevantes apenas no processo de otimização, variaram entre 6 e 24. Em todos os casos, o Tamanho da Amostra foi alto (acima de 98%) e os valores dos parâmetros Critério e o Mínimo de Folhas foram os mesmos para todos os casos (SE e 1, respectivamente).

A fim de identificar as diferenças entre esses 7 leques de parâmetros, a Figura 23 exibe as métricas do modelo FA utilizando os parâmetros otimizados pelas diferentes funções objetivo ($F_{1–6}$) em relação a função F_0, visto é a mais usualmente utilizada. Entre as funções que utilizaram o coeficiente de determinação, F_0 foi a que gerou um modelo com menores erros relativos e os maiores R^2 (visto que F_1, F_2 e F_3 possuem MAPEs maiores e R^2 menores). Além

disso, em F_0 houve uma menor diferença entre as métricas dos conjuntos, indicando menos sobreajuste, e F_3 gerou um modelo com um MAPE bem próximo de F_0. Não obstante, as funções objetivo que possuem o MAPE em sua definição (F_{4-6}) geraram modelos com os menores erros relativos que ficaram bem próximos entre si. Entre elas, F_4 e F_5 possuíram os melhores coeficientes de determinação.

Figura 23 – (a) MAPE e (b) R^2 do modelo FA utilizando os parâmetros otimizados pelas diferentes funções objetivo *x*. As métricas relativas a cada função objetivo *x* estão sendo exibidas em relação a respectiva métrica de F_0. LEGENDA: (a) ■ MAPE do conjunto de treino; ■ MAPE do conjunto de teste. (b) ■ R^2 do conjunto de treino; ■ R^2 do conjunto de teste.



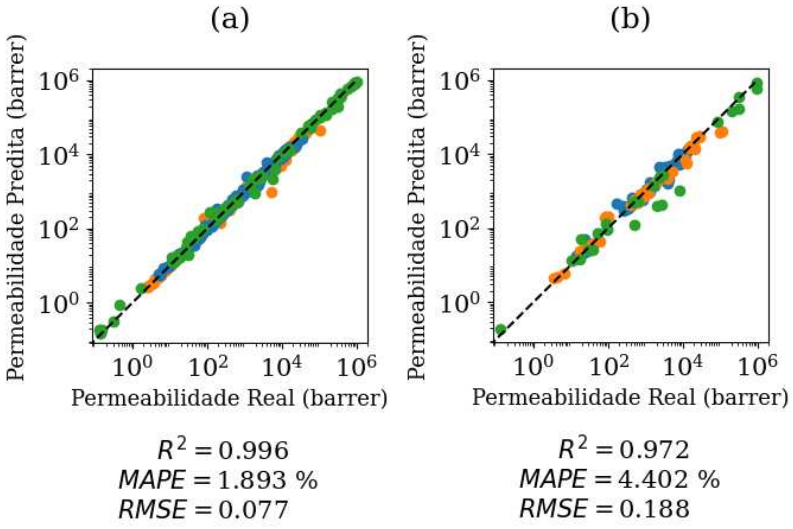
Fonte: Elaborada pelo autor.

Como pode ser visto na Figura 23, as métricas do conjunto de teste variaram mais conforme a função objetivo do que as métricas do conjunto de treino. Pode-se afirmar que o desempenho de generalização do modelo se mostrou mais sensível à escolha dos hiperparâmetros. Observa-se na Tabela 9 que aparentemente não há um padrão simples para a melhor parametrização do modelo e que as relações entre os hiperparâmetros são complexas. As variações numéricas do Número de Árvores (19 – 397) e da Profundidade Máxima (17 – 44), assim como as diferentes proporções entre elas ilustram esse fato. Em contrapartida, nesse contexto, percebe-se que há uma correlação entre o Número de Atributos e a qualidade de previsão do modelo, visto que os modelos que utilizaram mais atributos de entrada obtiveram, em média, os menores MAPEs.

As funções F_4 e F_5 geraram os modelos com as melhores métricas, entretanto, o leque de parâmetros da função F_4 foi escolhido para realizar a análise importância das variáveis, visto que suas métricas possuem valores ligeiramente melhores. Na Figura 24 são exibidos gráficos de dispersão entre as permeabilidades preditas e reais do conjunto de treino e teste.

Visualmente, pode-se perceber que as membranas de MOF apresentaram um desempenho de previsão um pouco pior em comparação com os outros dois tipos, visto que no conjunto de teste há alguns pontos subestimados (valores preditos abaixo do real). Entretanto, mesmo com essa diferença, as métricas do modelo podem ser consideradas adequadas para aplicações de engenharia (triagem e guia para o desenvolvimento de membranas).

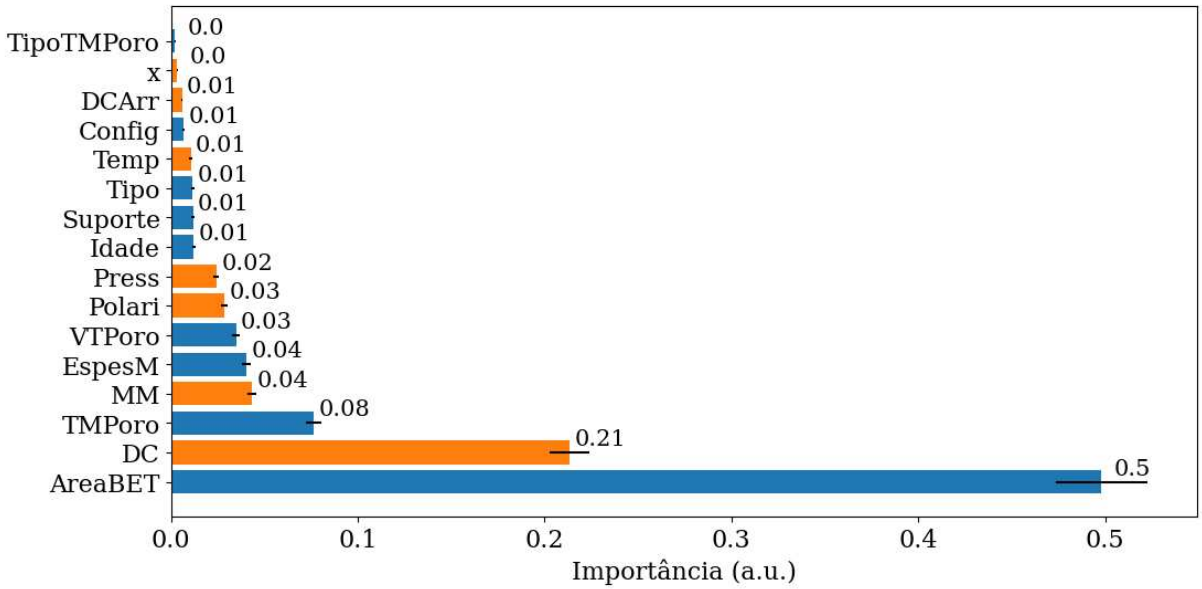
Figura 24 – Métricas e gráficos de dispersão das permeabilidades reais e preditas do modelo FA utilizando o (a) conjunto de treino e (b) conjunto de teste. LEGENDA: ■ Polímero; ■ Zeólita; ■ MOF.



Fonte: Elaborada pelo autor.

A análise de Importâncias de Permutação dos atributos de entrada do modelo FA foi realizada utilizando o conjunto de treino e seus resultados podem ser visualizados na Figura 25. A Área Superficial Específica BET foi o atributo de entrada mais relevante entre todas as incluídas na regressão, seguido do Diâmetro Cinético do Gás de Alimentação e do Tamanho Médio de Poro. O diâmetro cinético foi a propriedade física do gás de alimentação mais importante, seguida da massa molar e da polarizabilidade. Além disso, o tipo das membranas (Tipo) foi uma variável pouco relevante, estando no mesmo nível que o Tipo do Tamanho de Poro e o Diâmetro Cinético do Gás de Arraste, fato que indica certa capacidade de generalização do modelo.

Figura 25 – Importâncias de cada atributo de entrada do modelo FA para a predição do logaritmo da permeabilidade. LEGENDA: ■ Atributos morfológicos; ■ Atributos de processo. “AreaBET” = Área Superficial Específica; “DC” = Diâmetro Cinético do Gás de Alimentação; “TMPoro” = Tamanho Médio de Poro; “MM” = Massa Molar do Gás de Alimentação; “EspesM” = Espessura Média; “VTPoro” = Volume Total de Poro; “Polari” = Polarizabilidade do Gás de Alimentação; “Press” = Diferença de Pressão; “Idade” = Idade; “Suporte” = Material de Suporte; “Tipo” = Tipo; “Temp” = Temperatura; “Config” = Configuração; “DCArr” = Diâmetro Cinético do Gás de Arraste; “x” = Fração do Gás; “TipoTMPoro” = Tipo do Tamanho de Poro.



Fonte: Elaborada pelo autor.

5.3.2 Máquinas de Vetores de Suporte

O modelo de regressão “Regressão de Vetores de Suporte” (RVS) foi o próximo a ser analisado. Nesse caso os hiperparâmetros do modelo foram otimizados utilizando apenas três funções objetivo diferentes; a F_0 que é mais comumente utilizada; a F_4 e a F_5 que geraram os melhores modelos no caso anterior. Além disso, o *kernel* RBF (função de base radial) cuja definição pode ser vista na Equação (4.5) foi utilizado. Conforme o procedimento padrão, 500 passos iterativos (tentativas) foram realizados para a otimização de cada caso. Os valores ótimos, e a faixa de procura dos hiperparâmetros podem ser visualizados na Tabela 11. A escolha da faixa utilizada para a busca dos parâmetros foi inicialmente guiada por curvas de validação e posteriormente ajustada de forma que o valor ótimo dos parâmetros não se encontre nos limites mínimos ou máximos delimitados. Nesse modelo, a única exceção esteve na faixa do parâmetro *k-fold* que por motivos de performance computacional teve sua faixa limitada a

um valor máximo de 3. O tempo de execução e o melhor valor encontrado para cada caso pode ser visto na Tabela C2 (Apêndice C).

Tabela 11 – Hiperparâmetros avaliados durante a otimização do modelo Máquinas de Vetores de Suporte para a predição da performance das membranas.

Hiperparâmetro	**Faixa de procura** [1]	**RBF**		
		F_0	F_4	F_5
Gama (γ)	0.1 – 7 ($1 \cdot 10^{-4}$)	0,1401	6,0410	$1{,}75 \cdot 10^{-2}$
C	1000 – 150000 (1)	36708	62559	66082
Epsilon (ε)	0 – 1 ($1 \cdot 10^{-6}$)	$1{,}3 \cdot 10^{-3}$	$4 \cdot 10^{-6}$	$3{,}7 \cdot 10^{-5}$
k-fold	2 – 3 (1)	-	2	3

[1] O valor entre parênteses corresponde ao passo (*step*) definido para cada faixa. Fonte: Elaborada pelo autor.

Na Tabela 12 são apresentadas as métricas de desempenho do modelo RVS ao utilizar os parâmetros otimizados de cada função objetivo. Primeiramente, observa-se um alto sobreajuste no modelo gerado pela função objetivo F_4, visto que apesar de possuir o maior R^2 e menor MAPE de treino dentre todas as funções objetivo (com 0,9998 e 0,07 %, respectivamente), possui as piores métricas para o conjunto de teste (com R^2 e MAPE de 0,37 e 27 %, respectivamente). Diferente do ocorrido no modelo Florestas Aleatórias, o sobreajuste e o alto erro associado ao conjunto de teste torna o modelo gerado por F_4 inadequado para previsão da permeabilidade. No âmbito dos hiperparâmetros, ele possui o maior γ, maior C e menor ε, características que explicam o sobreajuste impedindo que o modelo tenha um caráter generalista (GÉRON, 2019).

Tabela 12 – Métricas de desempenho (MAPE e R^2) do modelo RVS otimizado por diferentes funções objetivo utilizando o conjunto de treino e teste.

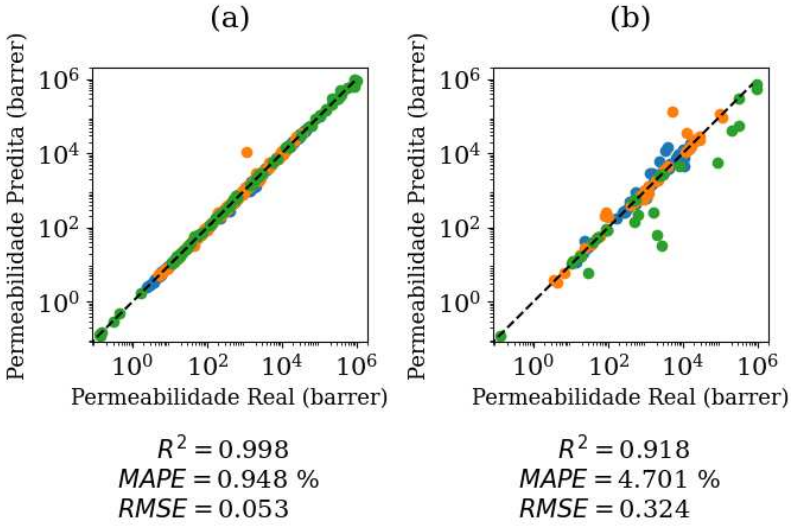
Métrica[1]	**F_0**	**F_4**	**F_5**
R^2 treino	0,9986	0,9998	0,9981
R^2 teste	0,9489	0,3712	0,9179
Δ R^2 ($\cdot 10^{-2}$)	4,97	62,86	8,02
MAPE treino (%)	0,2256%	0,0701%	0,9481%
MAPE teste (%)	3,8866%	27,6501%	4,7006%
Δ MAPE (%)	3,6610%	27,5800%	3,7525%

[1] Δ = |teste – treino|. Fonte: Elaborada pelo autor.

Por outro lado, embora possuam hiperparâmetros distintos, os modelos gerados pelas funções objetivo F_0 e F_5 apresentaram métricas melhores. Em ambos os casos, os valores ótimos do hiperparâmetro γ são pequenos e percebe-se que há uma ação de compensação dele em relação ao hiperparâmetro C. Considerando apenas essas duas funções, enquanto o valor de γ é oito vezes maior para F_0, o valor de C é duas vezes maior para a F_5. Por outro lado, a espessura ótima das margens dos vetores de suporte (hiperparâmetro ε) acompanha a mesma tendencia do hiperparâmetro γ, em que F_0 possuiu o maior valor, com $1{,}3 \cdot 10^{-3}$. Assim como anteriormente, a seleção dos hiperparâmetros ótimos foi bastante determinada pela função objetivo. Vale salientar que o sobreajuste desses modelos foram maiores em comparação daqueles gerados pelo método Florestas Aleatórias, entretanto os valores podem ser considerados aceitáveis para ambos os conjuntos.

De toda forma, o melhor modelo foi o gerado pela função F_0, pois apresentou as melhores métricas nos conjuntos de treino/teste e apresentou um desempenho contextualmente mediano no conjunto de extrapolação. Sendo assim, os parâmetros de F_0 foram escolhidos para a realização das análises posteriores. Na Figura 26 são exibidas as métricas de desempenho e os gráficos de dispersão entre as permeabilidades reais e preditas utilizando os diferentes conjuntos de dados. Pode-se visualizar graficamente o sobreajuste presente ao utilizar o modelo RVS, pois ao mesmo tempo que ajusta muito bem o conjunto de treino, sua performance decai no conjunto de teste (principalmente para a membranas de MOF).

Figura 26 – Métricas e gráficos de dispersão das permeabilidades reais e preditas do modelo RVS utilizando o (a) conjunto de treino e (b) conjunto de teste. LEGENDA: ■ Polímero; ■ Zeólita; ■ MOF.

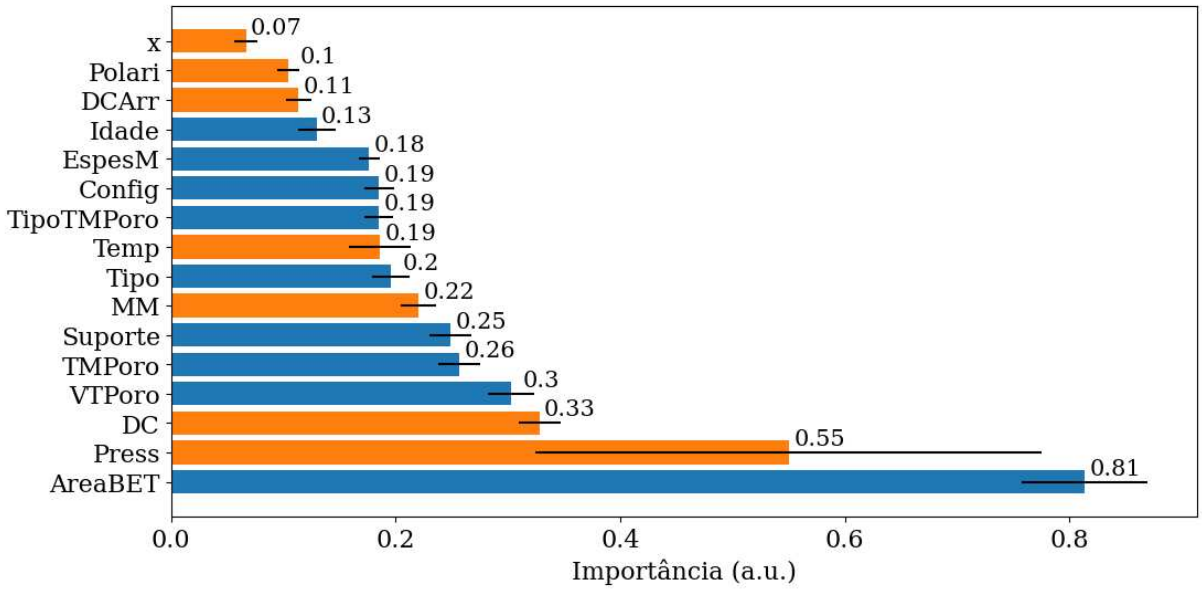


Fonte: Elaborada pelo autor.

A análise de Importâncias de Permutação dos atributos de entrada do modelo RVS foi realizada utilizando o conjunto de treino e seus resultados podem ser visualizados na Figura 27. Percebe-se, em comparação com o modelo Florestas Aleatórias, que nesse caso as importâncias estão distribuídas mais uniformemente. A Área Superficial Específica BET foi o atributo de entrada mais relevante de todos e o Diâmetro Cinético foi a propriedade física do GA mais importante, ficando nesse caso em 3° lugar no total. A polarizabilidade foi um dos atributos menos relevantes para o modelo juntamente com a Fração do Gás, que ficou em último. A Diferença de Pressão foi o segundo atributo mais importante, fato que é estranho visto que, para o modelo, ele é mais importante que as propriedades físicas do GA ou atributos morfológicos como o tamanho de poro. A alta incerteza da análise indicada pelo alto desvio padrão da sua importância ilustra esse fato.

O mesmo ocorre com o Material de Suporte, cuja importância é superestimada e aparece com o mesmo nível de relevância que o Tamanho Médio de Poro e o Volume Total de Poro. Sabe-se teoricamente que essas variáveis influenciam na permeação de gases, porém não na intensidade das outras que a acompanham nessa análise na Figura 27. Por outro lado, a Espessura Média é subestimada pelo modelo, pois aparece como menos relevante que a Configuração ou o Tipo do Tamanho de Poro. Esses fatos podem explicar o sobreajuste do modelo, entretanto, da maneira que foi treinado, ele ainda se mostra adequado para a triagem e desenvolvimento de membranas.

Figura 27 – Importâncias de cada atributo de entrada do modelo RVS para a predição do logaritmo da permeabilidade. LEGENDA: ■ Atributos morfológicos; ■ Atributos de processo. “AreaBET” = Área Superficial Específica; “Press” = Diferença de Pressão; “DC” = Diâmetro Cinético do Gás de Alimentação; “VTPoro” = Volume Total de Poro; “TMPoro” = Tamanho Médio de Poro; “Suporte” = Material de Suporte; “MM” = Massa Molar do Gás de Alimentação; “Tipo” = Tipo; “Temp” = Temperatura; “TipoTMPoro” = Tipo do Tamanho de Poro; “Config” = Configuração; “EspesM” = Espessura Média; “Idade” = Idade; “DCArr” = Diâmetro Cinético do Gás de Arraste; “Polari” = Polarizabilidade do Gás de Alimentação; “x” = Fração do Gás.



Fonte: Elaborada pelo autor.

5.3.3 Redes Neurais

Finalmente, as Redes Neurais Artificiais (RNA) foram estudadas para a previsão da permeabilidade de gases nas membranas. Para isso a biblioteca computacional *Keras* foi utilizada, entretanto ela não possui uma API (do inglês, *Application Programming Interface*) própria para a aplicação do método de Validação Cruzada e por esse motivo apenas a função objetivo F_0 foi utilizada no processo de otimização. A rede neural mostrada na Figura 11, com apenas uma camada oculta, foi utilizada e a otimização dos hiperparâmetros foi dividida em duas etapas: (1) arquitetura e (2) aprendizado. A primeira etapa teve 200 tentativas e englobou a análise da quantidade de neurônios na camada oculta e suas respectivas funções de ativação. Os outros hiperparâmetros possuíram o valor padrão recomendados pela biblioteca computacional (0,001 para a Taxa de Aprendizado, EQ para a Função de Custo, 32 Lotes e 1 Época). Na segunda etapa, com 500 tentativas e utilizando os valores ótimos da primeira etapa, determinou-se a melhor Função de Custo e os melhores valores para a Taxa de Aprendizado,

Lotes e Épocas. Apenas no processo de otimização as métricas foram obtidas a partir da média do treino de 3 redes neurais.

Os valores ótimos e a faixa de procura dos hiperparâmetros podem ser vistos na Tabela 13. Conforme o procedimento padrão, faixa escolhida para a busca dos parâmetros foi inicialmente guiada por curvas de validação e posteriormente ajustada para que o valor ótimo não se encontre nos limites delimitados. Vale destacar que, ao utilizar a função objetivo F_0, percebeu-se que o melhor valor para o Número de Neurônios e para as Épocas tenderam a se aproximar do máximo definido. Em função disso e devido a questões computacionais, o número máximo para os neurônio e épocas foram definidos como sendo 500 e 700, respectivamente. O tempo de execução e o melhor valor encontrado para F_0 pode ser visto na Tabela C3 (Apêndice C).

Tabela 13 – Hiperparâmetros avaliados durante a otimização do modelo Redes Neurais Artificiais para a predição da performance das membranas.

Hiperparâmetro	**Faixa de procura** [1]	**F_0**
Número de Neurônios	1 – 500 (1)	487
Função de Ativação	ReLU, sigmoide e Leaky ReLU	"ReLU"
Taxa de Aprendizado	$1{\cdot}10^{-5}$ – 0,1 (*log*) [2]	$1{,}981{\cdot}10^{-2}$
Função de Custo	EQ, ER e Log-Cosh [3]	Log-Cosh
Lotes	28 – 524 (2)	284
Épocas	1 – 700 (1)	687

[1] O valor entre parênteses corresponde ao passo (*step*) definido para cada faixa. [2] Para a biblioteca, o passo *log* indica que o valores da faixa serão escolhidos arbitrariamente com base no domínio logarítimo e não linear como os outros. [3] EQ = Erro Quadrático; ER = Erro Relativo; e Log-Cosh = Erro do Logarítmo do Cosseno Hiperbólico. Fonte: Elaborada pelo autor.

Na Tabela 14 são apresentadas as métricas de desempenho do modelo RNA utilizando os parâmetros otimizados e a Figura 28 exibe os respectivos gráficos de dispersão para os diferentes conjuntos de dados. O modelo apresentou um bom desempenho nos conjuntos de treino e teste e percebe-se que foi obtido um sobreajuste semelhante ao do modelo FA com uma diferença de $1{,}77{\cdot}10^{-2}$ para o R^2 e 1,93 % para o MAPE. Além disso, visualmente, o desempenho foi similar para todos os tipos de membrana.

A ordem geral das importâncias dos atributos, que pode ser vista na Figura 29, é diferente da obtida pelos outros dois modelos, porém percebe-se que permanecem alguns padrões. A Área Superficial Específica BET foi novamente o atributo mais importante e, nesse

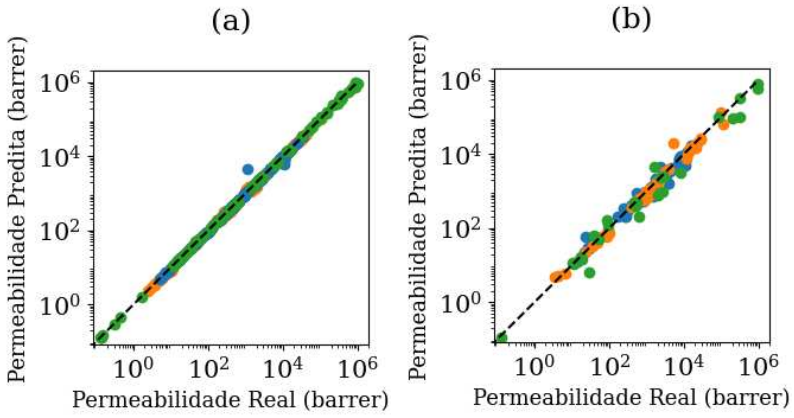
caso, foi seguido do Volume Total de Poro e do Diâmetro Cinético do Gás de Alimentação. As importâncias das propriedades físicas do gás de alimentação seguiram a ordem que também foi encontrada nos outros modelos: diâmetro cinético > massa molar > polarizabilidade. A Fração do gás foi o atributo menos importante e o Tipo das membranas, assim como em FA, foi o 6° atributo menos relevante. Por outro lado, é estranho o fato de que o Material de Suporte se encontra como o 6° atributo mais importante, em contraste com atributos como a Espessura Média, a Temperatura e a Pressão, que ficaram nas últimas posições da lista de importâncias.

Tabela 14 – Métricas de desempenho (MAPE e R^2) do modelo RNA para o conjunto de treino e teste.

Métrica[1]	**F_0**
R^2 treino	0,9989
R^2 teste	0,9788
ΔR^2 ($\cdot 10^{-2}$)	2,01
MAPE treino (%)	1,0117
MAPE teste (%)	3,8870
Δ MAPE (%)	2,8753

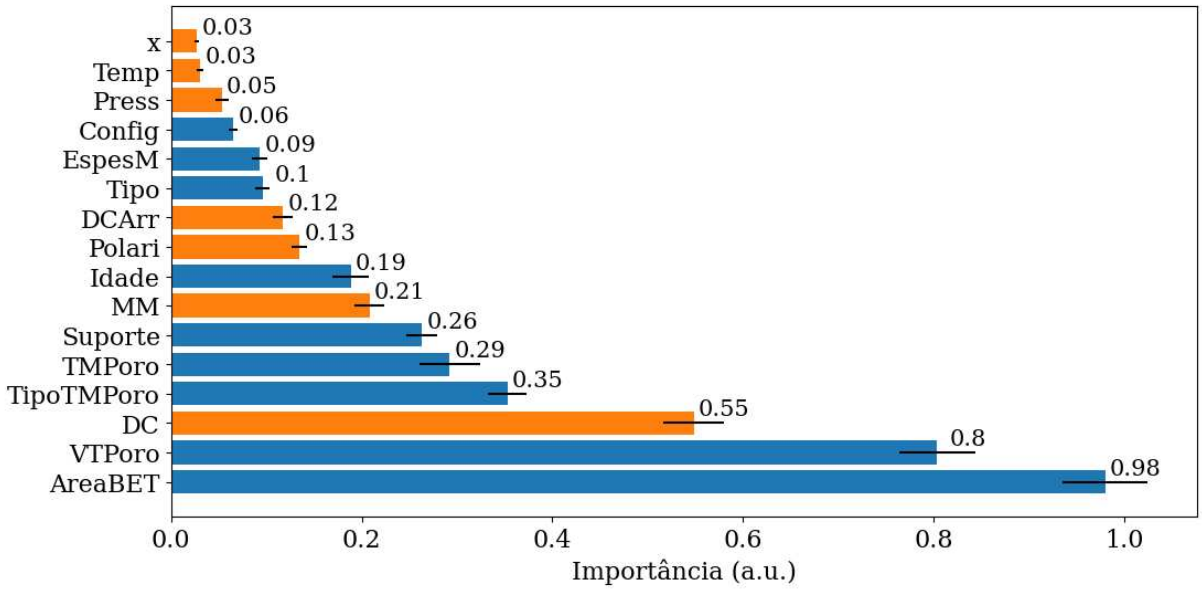
[1] Δ = |teste – treino|. Fonte: Elaborada pelo autor.

Figura 28 – Métricas e gráficos de dispersão das permeabilidades reais e preditas do modelo RNA utilizando o (a) conjunto de treino e (b) conjunto de teste. LEGENDA: ■ Polímero; ■ Zeólita; ■ MOF.



Fonte: Elaborada pelo autor.

Figura 29 – Importâncias de cada atributo de entrada do modelo RVS para a predição do logaritmo da permeabilidade. LEGENDA: ■ Atributos morfológicos; ■ Atributos de processo. “AreaBET” = Área Superficial Específica; “VTPoro” = Volume Total de Poro; “DC” = Diâmetro Cinético do Gás de Alimentação; “TipoTMPoro” = Tipo do Tamanho de Poro; “TMPoro” = Tamanho Médio de Poro; “Suporte” = Material de Suporte; “MM” = Massa Molar do Gás de Alimentação; “Idade” = Idade; “Polari” = Polarizabilidade do Gás de Alimentação; “DCArr” = Diâmetro Cinético do Gás de Arraste; “Tipo” = Tipo; “EspesM” = Espessura Média; “Config” = Configuração; “Press” = Diferença de Pressão; “Temp” = Temperatura; “x” = Fração do Gás.



Fonte: Elaborada pelo autor.

5.4 Melhor modelo, F_{obj} e poda dos atributos de entrada

Todos os modelos (FA, RVS e RNA) obtiveram métricas aceitáveis para o conjunto de treino e teste, entretanto o modelo Florestas Aleatórias foi selecionado para a análise posteriores devido a sua simplicidade e rapidez de otimização. Como pode ser visto nas Tabela C1, Tabela C2 e Tabela C3 (no Apêndice C), o tempo de otimização foi menor para FA e, além disso, sua análise de importâncias se mostrou mais fisicamente coerente. No modelo RVS (Figura 27), a importância da Diferença de Pressão e do Material foi contextualmente alta enquanto a Espessura Média foi subestimada. No modelo RNA (Figura 29), a Temperatura e a Diferença de Pressão foram menos relevantes que a Configuração, o Diâmetro Cinético do Gás de Arraste ou o Material de Suporte. Essas incoerências podem diminuir a capacidade de extrapolação dos modelos.

Vale salientar que os desempenhos relatados dependem da metodologia que foi aplicada e é possível que resultados melhores possam ser encontrados após a aplicação de, por exemplo,

outras arquiteturas de redes neurais, de outros *kernels* no modelo RVS ou de um banco de dados maior. No âmbito das funções objetivo, F_4 apresentou-se como a melhor para o conjunto de teste no modelo FA. Pode-se afirmar também que os hiperparâmetros encontrados pela função F_0, que é a mais usual, foram satisfatórios em todos os casos, apresentando um desempenho mediano em FA e se mostrando como o melhor leque no modelo RVS.

Ao todo, o modelo Florestas Aleatórias com os hiperparâmetros obtidos a partir da função objetivo F_4 dispôs de um R^2 de 0,996 e 0,972 para o conjunto de treino e teste, respectivamente; e um MAPE de 1,9 e 4,4 % para o conjunto de treino e teste, respectivamente. Analisando a suas importâncias (Figura 25), percebe-se que a maioria dos atributos podem ser considerados pouco relevantes para a predição do modelo, visto que 13 deles possuem importâncias pequenas (abaixo de 0,05) frente aos três primeiros atributos. Variáveis pouco importantes, também chamadas de variáveis ruído (em inglês, *noisy features*), possuem menor influência sobre a variável alvo e, em geral, podem ser retiradas da modelagem, pois isso permite que, virtualmente, a predição seja mais precisa, simples e eficiente (LI, Xiao *et al.*, 2019; PHUNG *et al.*, 2022).

Como existem atributos com algum nível de correlação (não são de fato independentes), a análise de IP pode não ser capaz de reconhecer as variáveis ruído adequadamente. A interpretabilidade da importância de cada variável de entrada para a previsão depende implicitamente da suposição de que elas são ortogonais (ou pouco correlacionadas). Foi notado em outros trabalhos que na análise de IP há contribuições conjuntas quando existem atributos de entrada correlacionados, de forma que as importâncias de variáveis irrelevantes são superestimadas. Dessa forma, a colinearidade (ou multicolinearidade) pode não interferir no desempenho de um modelo, mas impacta negativamente na interpretação dos atributos usados para construi-lo (BREIMAN, 2001; CHATTERJEE; HADI, 2012; WEI; LU; SONG, 2015).

Para identificar essas variáveis, cada atributo de entrada foi retirado de forma cumulativa na ordem de importância da Figura 25 e o desempenho de predição do modelo FA foi sucessivamente avaliado. Como procedimento padrão, a cada remoção o modelo foi otimizado utilizando a função objetivo F_4 e 500 tentativas. Os hiperparâmetros ótimos para cada caso podem ser vistos na Tabela C4 (Apêndice C). Na Figura 30 são mostrados os MAPEs dos modelos FA para o conjunto de treino e teste a cada atributo retirado cumulativamente. Pode-se observar que os erros de ambos os conjuntos possuem perfis semelhantes e variam muito pouco com a retirada da maioria das variáveis.

Como visto pela análise de IP, o Tamanho Médio de Poro, o Diâmetro Cinético do GA e a Área Superficial Específica são, de forma destacada, as variáveis que mais carregam

informação sobre a permeação dos gases. Isso pode ser ilustrado na Figura 30, visto que mesmo com apenas esses três atributos, o modelo ainda é razoavelmente adequado apresentando um MAPE em torno de 4,9 e 8,3 % no conjunto de treino e teste, respectivamente. Algumas variáveis que foram relativamente relevantes na análise de IP, tiveram pouco impacto depois que foram retiradas como a Massa Molar do GA, a Espessura Média, o Volume Total de Poro e a Polarizabilidade do GA. Isso ocorre devido aos níveis de correlação suficientemente altos que esses atributos possuem com as variáveis que não foram retiradas. Os coeficientes de Pearson e Spearman (Figura 21 e Figura B2 - Apêndice B) são medidas estatísticas que podem fornecer percepções sobre essa relação. Além disso, a Espessura Média especificamente é uma característica aproximada que é medida visualmente por meio de microscopia eletrônica (HABERT; BORGES; NOBREGA, 2006). Assim, pode não fornecer tanta informação sobre a resistência de permeação dos gases.

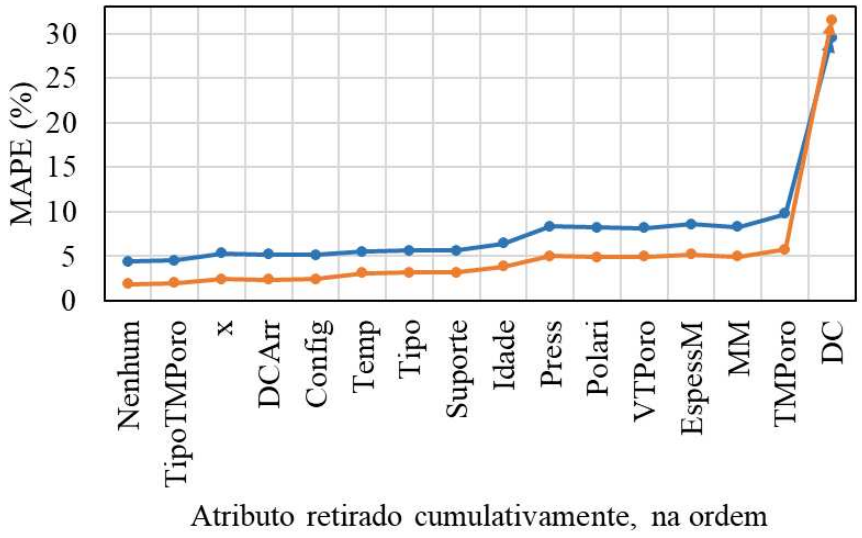
Por outro lado, alguns atributos foram subestimados. Isso pode ser ilustrado na Figura 30 pela presença de dois níveis ligeiramente distintos até o atributo Tamanho Médio de Poro, sendo ocasionado pela remoção dos atributos Idade e Diferença de Pressão. De fato, pela natureza do fenômeno, sabe-se que temperatura e pressão são fatores que afetam diretamente a difusividade e adsorção de gases em meios porosos. Mesmo com a relativa baixa variabilidade dos dados para as características citadas (aproximadamente 75% dos registros referentes a temperatura e pressão se encontram abaixo de 308 K e 0,6 MPa, respectivamente), elas carregam informações importantes para a previsão.

De toda forma, vale salientar que o atributo Tipo e Fração do Gás continuam sendo pouco relevantes, o que pode indicar uma certa capacidade generalista do modelo. Sabe-se que o comportamento de mistura de um determinado gás não irá depender apenas da sua composição (fração molar, volumétrica e mássica) mas também do(s) outro(s) gás(es) presentes. A depender do diâmetro cinético e das características químicas das moléculas do meio, o perfil de permeação de um gás pode não ser ideal (HABERT; BORGES; NOBREGA, 2006). Assim, é possível que somente a Fração do Gás não contenha informações suficientes para descrever os comportamentos de mistura e resulte em ser pouco relevante. Além disso, a Espessura Média e a Idade podem servir como indicativo ao tipo das membranas. Como pode ser visto nos gráficos da Figura B1 (Apêndice B), as membranas velhas ou com maiores espessuras são poliméricas. Dessa forma, tem-se atributos correlacionados que podem mascarar a real importância do atributo Tipo.

Vale ressaltar que as métricas de desempenho são adequadas e essas considerações não alteram o fato de que, frente a Área Superficial e ao Diâmetro Cinético do GA, os outros

atributos não desempenharam um papel tão relevante no processo de previsão. Na Figura C1 (Apêndice C) são mostrados os valores de R^2 dos modelos para cada atributo retirado cumulativamente e essa métrica também se mantém adequada e relativamente estável após a retirada da maioria dos atributos.

Figura 30 – MAPE dos modelos FAs otimizados a cada retirada sucessiva de um atributo de entrada. LEGENDA: ■ Conjunto de treino; ■ Conjunto de teste. “TipoTMPoro” = Tipo do Tamanho de Poro; “x” = Fração do Gás; “DCArr” = Diâmetro Cinético do Gás de Arraste; “Config” = Configuração; “Temp” = Temperatura; “Tipo” = Tipo; “Suporte” = Material de Suporte; “Idade” = Idade; “Press” = Diferença de Pressão; “Polari” = Polarizabilidade do Gás de Alimentação; “VTPoro” = Volume Total de Poro; “EspesM” = Espessura Média; “MM” = Massa Molar do Gás de Alimentação; “TMPoro” = Tamanho Médio de Poro; “DC” = Diâmetro Cinético do Gás de Alimentação.

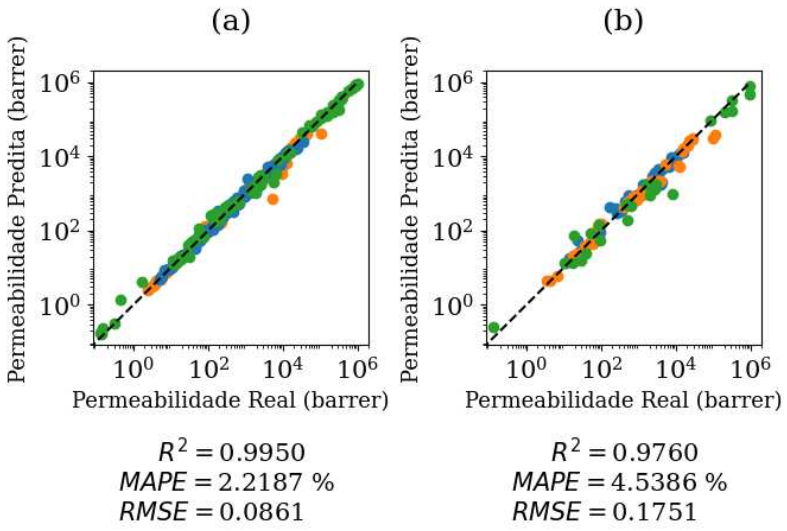


Fonte: Elaborada pelo autor.

Dessa forma, o conjunto de teste foi considerado como referência para a retirada das variáveis ruído e simplificar o modelo. A Figura C2 (Apêndice C) mostra um gráfico de cascata para ilustrar o efeito cumulativo da remoção de cada atributo no MAPE do conjunto de teste. Assim, foram retirados seis atributos que não se mostraram tão impactantes para a previsão: Massa Molar do GA, Diâmetro Cinético do Gás de Arraste, Polarizabilidade do GA, Material de Suporte, Configuração e Tipo do Tamanho de Poro. Devido a conhecimentos prévios sobre o processo, é esperado que atributos como o referente ao gás de arraste, o Material de Suporte e a Configuração não possuam tanta importância para a previsão. Além disso, mesmo sendo menos importantes, a outras variáveis (como o Tipo, o Volume Total de Poro ou a Temperatura) foram mantidas para auxiliar minimamente na previsão.

Na Figura 31 são apresentadas as métricas de desempenho do modelo FA após a poda dos atributos de entrada e a otimização dos hiperparâmetros. Os valores ótimos para cada hiperparâmetro pode ser visto na Tabela C5 (Apêndice C). Em suma, pequenas alterações podem ser vistas após a retirada desses atributos. O R^2 do conjunto de treino diminuiu $8{\cdot}10^{-4}$ enquanto o seu MAPE aumentou 0,33 %. Para o conjunto teste, o R^2 aumentou $3{,}7{\cdot}10^{-3}$ e o MAPE aumentou 0,09 %. Além disso, o sobreajuste também teve uma pequena redução.

Figura 31 – Métricas e gráficos de dispersão das permeabilidades reais e preditas (variável alvo deslogaritmizada) do modelo FA otimizado após a poda dos atributos de entrada utilizando o (a) conjunto de treino e (b) conjunto de teste. Δ R^2 ($\cdot 10^{-2}$) = 1,90. Δ MAPE = 2,32 %. Em que Δ = |teste – treino|. LEGENDA: ■ Polímero; ■ Zeólita; ■ MOF.



Fonte: Elaborada pelo autor.

Esses resultados estão adequados quando se compara com aqueles obtidos em outros trabalhos similares no campo de estudo. Como pode ser visto na Tabela 15, o presente trabalho possui os maiores R^2 e menores RMSEs para os conjuntos, apesar de utilizar um banco de dados relativamente heterogêneo com diferentes tipos de materiais e gases. Vale ressaltar que a comparação entre os trabalhos precisa ser realizada com ponderação, visto que existem diferenças metodológicas e bancos de dados de treinamento distintos são utilizados entre os trabalhos. Por exemplo; Hasnaoui, Krea e Roizard (2017); Zhu et al. (2020); Barnett et al. (2020); e Yang et al. (2022); utilizam de descritores teóricos para representar a estrutura química de diferentes polímeros e, portanto, diferentes materiais. Esses trabalhos consideram respectivamente 149, 315, 698 e 778 polímeros/materiais diferentes no banco de dados. Por outro lado, naqueles trabalhos que utilizam descritores experimentais, Pan et al. (2022) consideram 16 polímeros precursores diferentes e Guan et al. (2022) consideram 36 e 41 tipos

de MOFS e polímeros diferentes, respectivamente. O presente trabalho utilizou um banco de dados com 31 materiais diferentes.

Tabela 15 – Comparação das métricas de desempenho entre o presente trabalho e os similares encontrados na literatura.

Referência Bibliográfica	Tipo de material das membranas	Gás(es)	Tamanho do Banco de Dados	R^2 treino	R^2 teste	RMSE treino	RMSE teste
Este Trabalho	**Polímero, zeólita e MOF**	**He, H_2, CO_2, O_2, N_2, CH_4, eteno, etano, e SF_6**	**692**	**0,99**	**0,98**	**0,09**	**0,18**
Hasnaoui, Krea e Roizard (2017)[1 2]	Polímero	N_2, O_2 e CO_2	144	0,96	-	2,05	4,33
Zhu et al. (2020)	Polímero	CH_4, CO_2, H_2, He, N_2 e O_2	1501	0,99	0,93	-	-
Barnett et al. (2020)[1]	Polímero	N_2, O_2, H_2, He, CH_4 e CO_2	698	0,99	0,86	-	0,42
Pan et al. (2022)	PMC	CO_2, CH_4, N_2, O_2 e H_2	399	0,84	-	0,41	-
Guan et al. (2022)	MMM (MOFs)	CO_2	648	0,90	0,89	0,56	0,68
Yang et al. (2022)[1]	Polímero	He, H_2, O_2, N_2, CO_2 e CH_4	778	0,89	0,90	0,37	0,36

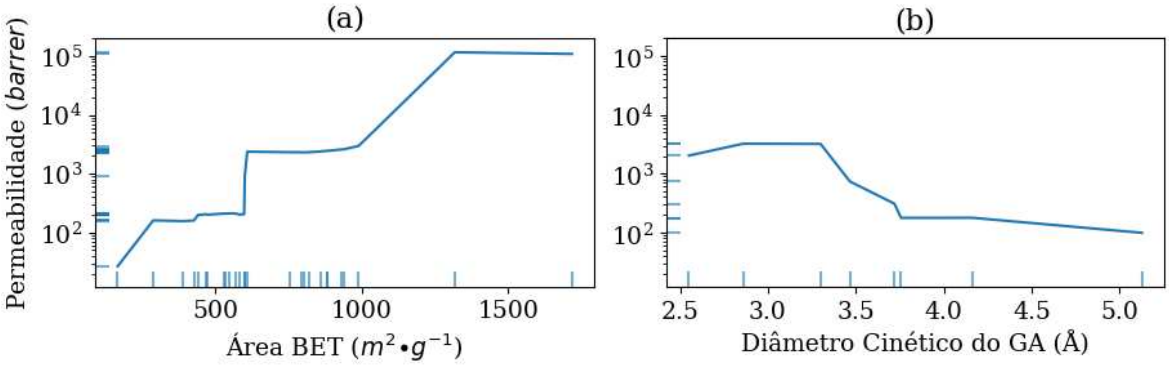
[1] Nesses trabalhos, os autores treinaram um modelo de AM para cada gás. As métricas exibidas nesta tabela são uma média dos resultados desses modelos. O tamanho do banco de dados corresponde ao maior utilizado dentre todos. [2] Por apresentar resultados inadequados, não foram considerados os dados referentes ao CH_4. Fonte: Elaborada pelo autor.

Após essa retirada, a ordem de importâncias dos atributos de entrada mudou pouco (Figura C3 – Apêndice C). O Volume Total de Poro foi o único atributo que ficou relativamente mais importante. Os valores das importâncias da Área Superficial Específica e do Diâmetro Cinético do Gás de Alimentação ficaram significantemente maiores, tornando-os mais destacados. Na Figura 32 são mostrados os Gráficos de Dependência Parcial desses dois atributos de entrada, que foram os mais importantes. Pode-se perceber que existe uma relação positiva entre a permeabilidade e Área Superficial Específica. No que diz respeito ao Diâmetro Cinético, observa-se um perfil muito comum do observado em vários trabalhos experimentais do banco de dados (HAYAKAWA; HIMENO, 2020; JIANG *et al.*, 2020, 2021; KIDA; MAETA; YOGO, 2017; WANG, Bin; GAO; *et al.*, 2019; WANG, Bin; ZHENG; *et al.*, 2019).

Neles, a permeação cresce até o diâmetro do gás CO_2 para, em seguida, decrescer de uma maneira exponencial, atingindo um patamar quase constante.

Em outros trabalhos, provavelmente devido às baixas capacidades adsortivas do material da membrana, não se observa esse comportamento do gás CO_2 e a permeação decresce em função do diâmetro cinético, indicando predominância de peneiramento molecular (FENG *et al.*, 2020; LEE *et al.*, 2012; LIAN *et al.*, 2022; PENG *et al.*, 2017). Dessa forma, ao comparar com o perfil do modelo (Figura 32-a), é possível que parte do resíduo encontrado seja consequência dessa diferença.

Figura 32 – Gráficos de Dependência Parcial da Permeabilidade em função da (a) Área Superficial Específica e (b) do Diâmetro Cinético do GA.



Fonte: Elaborada pelo autor.

Para trabalhos futuros, recomenda-se acrescentar atributos referentes a capacidade adsortiva dos materiais de forma que a modelagem seja mais refinada. Além disso, Liu et al. (2009) encontram uma relação quase linear entre a permeância e raiz do inverso massa molar do GA. Assim como a transformação logarítmica realizada na permeabilidade, vale como sugestão investigar futuramente transformações nas propriedades físicas dos GA como a realizada por esses autores. Os GDP dos outros atributos podem ser visualizados na Figura C4 (Apêndice C) e, pela mudança de escala dos gráficos, pode-se reiterar que eles auxiliam bem menos na previsão. Pode-se observar pela Figura C4-c (Apêndice C) que o tamanho de poro possui uma relação positiva com a permeabilidade.

Pan et al. (2022) analisam os fatores que influenciam na permeação de gases em membranas de PMC e os atributos estruturais, que incluem o FFV (do inglês, *Fractional Free Volume*) do precursor e o espaçamento médio das camadas cristalinas de carbono, são os mais importantes. No trabalho de Yang et al. (2022), dos 146 descritores químicos utilizados na modelagem, o mais importante foi aquele relacionado à área superficial de Van der Waals dos

grupos funcionais. Daglar e Keskin (2022) treinaram modelos de AM para prever a difusão e a capacidade adsortiva (duas propriedades que podem descrever o coeficiente de permeabilidade) de MMM de MOFs e as características estruturais (especialmente a porosidade dos materiais) foram aquelas mais relevantes.

Vale destacar que embora o modelo possua uma certa capacidade de generalização, o banco de dados foi construído com registros de membranas que foram planejadas com uma lógica para "serem eficientes" na separação de gases. Portanto, evidentemente, precisa-se ter compreensão das limitações do modelo ao analisar seus resultados e extrapolar os dados. No caso do aprendizado de máquina, essas restrições são regidas principalmente pelo banco de dados ao qual ele foi treinado e, portanto, a distribuição dos valores dos atributos de entrada (Figura 19) podem servir como critério para a defini-las (DOBBELAERE *et al.*, 2021).

Dessa forma, é notório que o tamanho de poro e a espessura da membrana são características que possuem uma importância significativa no processo de permeação. Membranas muito grossas ou com poros muito largos terão uma baixa permeabilidade e seletividade, respectivamente (HABERT; BORGES; NOBREGA, 2006). Portanto, é mais preciso afirmar que, dado a faixa em que foram analisadas, elas não possuem tanta importância quanto os outros atributos. De outro modo, deve-se manter a metodologia utilizada para fabricação das membranas e, a partir de então, focar em conseguir uma área superficial maior para alcançar maiores permeabilidades.

5.5 Seletividade

Conhecer a seletividade de uma membrana para um sistema de mistura binária é essencial para determinar a sua eficiência de separação e planejar suas futuras aplicações (KAMBLE; PATEL; MURTHY, 2021). Nesta sessão, foi investigada a capacidade do modelo Florestas Aleatórias que foi otimizado e treinado anteriormente (Seção 5.4) em predizer a seletividade ideal (α) – Equação (2.4) – das membranas presentes no banco de dados. Vale ressaltar que nele cada registro se refere a uma membrana (caracterizada pelos atributos morfológicos) sendo aplicada em uma determinada circunstância de utilização (caracterizada pelos atributos de processo). Para estimar a seletividade ideal de um sistema binário de uma determinada membrana, basta calcular a permeabilidade dos dois gases de interesse (na sua forma pura) utilizando o modelo e, logo após, aplicar a Equação (2.4).

Ao todo, 4 sistemas foram investigados: He/CH_4, H_2/CH_4, CO_2/CH_4 e N_2/CH_4. Para analisar a qualidade da previsão, a seletividade real precisa ser obtida do banco de dados. Assim,

para cada sistema, apenas as membranas que foram testadas, concomitantemente, com os dois gases de interesse na sua forma pura foram selecionadas e, em seguida, foi calculada a seletividade ideal real. Dessa forma, após essa filtragem, a quantidade de registros separados para cada sistema binário para avaliar a previsão do modelo pode ser visto na Tabela 16. Esses dados também estão disponíveis com mais detalhes no formato "*xlsx*"[13].

Por padronização, todas as métricas de desempenho do modelo foram relativas ao logaritmo da seletividade. Como pode ser visto graficamente na Tabela 16, o modelo se mostrou adequado para a previsão da seletividade para todos os sistemas com um R^2 acima de 0,8 e um RMSE abaixo de 0,18 em todos os casos. Na Figura D1 são exibidos os gráficos de dispersão entre as seletividades reais e preditas dos sistemas analisados. Guan et al. (2022) também treinaram um modelo FA, porém nesse caso para prever a seletividade do sistema CO_2/CH_4 de membranas MMM de MOFs. Os autores obtiveram um R^2 de 0,88 e 0,86 para o conjunto de treino e teste, respectivamente; e um RMSE de 0,22 e 0,2 para o conjunto de treino e teste, respectivamente.

Tabela 16 – Métricas de desempenho (MAPE e R^2) do modelo FA para a previsão do logaritmo da seletividade.

Sistema	Quantidade de registros	R^2	MAPE (%)	RMSE
He/CH_4	39	0,9726	158,8082	0,0794
H_2/CH_4	61	0.8197	15,7561	0,1788
CO_2/CH_4	83	0,9475	12,1240	0,1528
N_2/CH_4	68	0.9503	43,9820	0.0852

Fonte: Elaborada pelo autor.

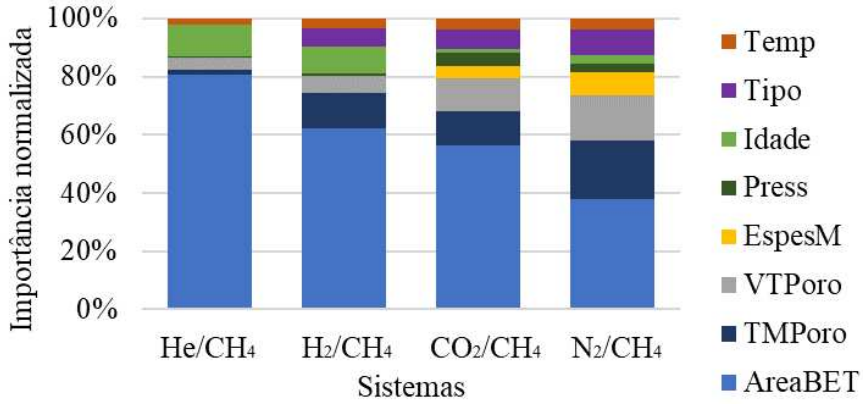
Na Figura 33 são apresentadas as importâncias normalizadas dos atributos de entrada para cada sistema binário. Pode-se perceber que a Área Superficial Específica foi novamente o atributo mais relevante e, ao passo que os gases ficam mais similares (o que pode ser medido pelo tamanho), a importância desse atributo diminui. Quanto mais parecidos em tamanho forem os gases, mais a características químicas (relativas à solubilidade/adsorção) serão relevantes no processo de separação. Além disso, a capacidade de adsorção do material exerce uma influência

[13] https://github.com/tenoriolms/databank_membranes/blob/main/databank_pure_data_for_selectivity.xlsx

significativamente maior na permeabilidade do N_2 do que o habitual (DAGLAR; KESKIN, 2022).

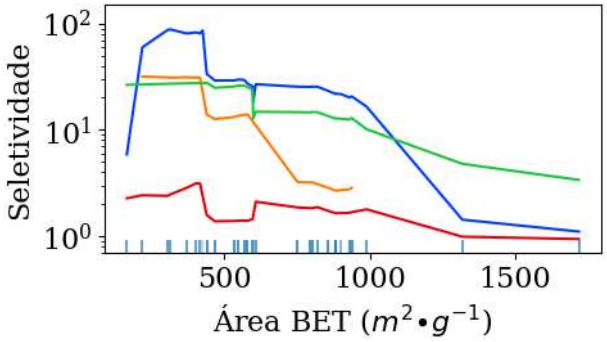
Na Figura 34 são mostrados os Gráficos de Dependência Parcial da Área Superficial Específica e percebe-se uma relação inversamente proporcional com a seletividade, diferente do que ocorre com a permeabilidade. Esse fato ilustra o *trade-off* entre a seletividade e a permeabilidade discutido por Robeson (2008), porém em uma perspectiva diferente. Isto é, analisando os resultados do modelo, membranas com uma área superficial grande tendem a ter uma alta permeabilidade, porém perdem seletividade. É interessante salientar que, ao analisar apenas as características estruturais da membrana, foi possível identificar esse comportamento. Os GDP dos outros atributos podem ser visualizados na Figura D2 (Apêndice D). Vale observar que, assim como a Área Superficial, porém com menos intensidade, o Tamanho Médio de Poro possui uma relação negativa com a seletividade.

Figura 33 – Importâncias normalizadas dos atributos de entrada do modelo FA para a predição do logaritmo da seletividade para cada sistema binário. LEGENDA: “AreaBET” = Área Superficial Específica; “TipoTMPoro” = Tipo do Tamanho de Poro; “VTPoro” = Volume Total de Poro; “EspesM” = Espessura Média; “Press” = Diferença de Pressão; “Idade” = Idade; “Tipo” = Tipo; “Temp” = Temperatura.



Fonte: Elaborada pelo autor.

Figura 34 – Gráfico de Dependência Parcial da seletividade de cada sistema binário em função da Área Superficial Específica. LEGENDA: ■ He/CH_4; ■ H_2/CH_4; ■ CO_2/CH_4; ■ N_2/CH_4.



Fonte: Elaborada pelo autor.

6 CONCLUSÃO

Um banco de dados referente a membranas para separação de gases foi construído com dados experimentais de 42 referências diferentes. Para a modelagem, ele foi filtrado de forma que foram utilizados 692 registros de 19 referências bibliográficas diferentes. Os registros se mostraram bem distribuídos quanto ao tipo de membrana e à permeabilidade dos gases. Além disso, foram considerados 3 tipos de membranas diferentes (Polimérica, Zeolítica e de MOF) e 9 diferentes gases na corrente de alimentação (He, H_2, CO_2, O_2, N_2, CH_4, eteno, etano e SF_6). Três modelos de AM (FA, RVS e RNA) foram testados e todos se mostraram adequados para prever o desempenho de permeação das diferentes membranas em diferentes circunstâncias de utilização do banco de dados. No modelo FA, foram investigadas 7 funções objetivo para a otimização dos hiperparâmetros. Nesse contexto a função objetivo F_4 se mostrou a mais adequada, gerando um modelo com o menor MAPE e maior R^2. No modelo RVS foram investigadas três funções objetivo para a otimização dos hiperparâmetros de forma que a F_0, que a mais comumente utilizada, se mostrou mais adequada.

Devido a simplicidade, rapidez de otimização e coerência física na análise de IP, o modelo FA foi utilizado para a análise de poda dos atributos de entrada com o intuito de simplificar a modelagem e retirar as variáveis correlacionadas. Os principais atributos que influenciaram na predição foram o Tamanho Médio de Poro, o Diâmetro Cinético do GA e a Área Superficial Específica. O modelo continua relativamente adequado ao considerar apenas esses 3 atributos de entrada, de forma que o desempenho da previsão se encontra com um MAPE de 4,9 e 8,3 % e um R^2 de 0,98 e 0,96 para o conjunto de treino e teste, respectivamente. Vale salientar que atributos como o Tipo e a Fração do Gás foram poucos relevantes para o modelo.

Foram retirados 6 atributos de entrada (incluindo a massa molar e polarizabilidade do GA) de forma que, após otimizado, o desempenho do modelo permaneceu praticamente inalterado com um MAPE de 2,2 e 4,5 % e um R^2 de 0,99 e 0,97 para o conjunto de treino e teste, respectivamente. Esses resultados estão adequados quando se compara com resultados de outros trabalhos do campo de estudo. A seletividade de quatro sistemas binários (He/CH_4, H_2/CH_4, CO_2/CH_4 e N_2/CH_4) foram estimadas utilizando o modelo FA treinado previamente. Dessa forma, o modelo também se mostrou adequado para a previsão dessa variável com a obtenção de um R^2 acima de 0,8 e um RMSE abaixo de 0,18 em todos os casos.

Para o modelo FA, a Área Superficial Específica foi o atributo mais relevante tanto para a previsão da permeabilidade, quanto da seletividade. Entretanto, esse atributo influencia de

maneira distinta nessas variáveis de performance, de modo que membranas com grande área superficial apresentarão uma alta permeabilidade, porém uma baixa seletividade. Pode haver um valor ideal para esse atributo, de forma que haja um equilíbrio entre essas duas características.

REFERÊNCIAS BIBLIOGRÁFICAS

AKHTAR, Farid; SJÖBERG, Erik; KORELSKIY, Danil; RAYSON, Mark; HEDLUND, Jonas; BERGSTRÖM, Lennart. Preparation of graded silicalite-1 substrates for all-zeolite membranes with excellent CO 2 /H 2 separation performance. **Journal of Membrane Science**, vol. 493, p. 206–211, nov. 2015. https://doi.org/10.1016/j.memsci.2015.06.020.

AKIBA, Takuya; SANO, Shotaro; YANASE, Toshihiko; OHTA, Takeru; KOYAMA, Masanori. Optuna: A Next-generation Hyperparameter Optimization Framework. 2019. https://doi.org/10.48550/ARXIV.1907.10902.

ALBERTO, Monica; BHAVSAR, Rupesh; LUQUE-ALLED, Jose Miguel; VIJAYARAGHAVAN, Aravind; BUDD, Peter M.; GORGOJO, Patricia. Impeded physical aging in PIM-1 membranes containing graphene-like fillers. **Journal of Membrane Science**, vol. 563, p. 513–520, out. 2018. https://doi.org/10.1016/j.memsci.2018.06.026.

ALIBAKSHI, Amin. Strategies to develop robust neural network models: Prediction of flash point as a case study. **Analytica Chimica Acta**, vol. 1026, p. 69–76, out. 2018. 10.1016/j.aca.2018.05.015.

AL-MAYTHALONY, Bassem A.; SHEKHAH, Osama; SWAIDAN, Raja; BELMABKHOUT, Youssef; PINNAU, Ingo; EDDAOUDI, Mohamed. Quest for Anionic MOF Membranes: Continuous **sod** -ZMOF Membrane with CO_2 Adsorption-Driven Selectivity. **Journal of the American Chemical Society**, vol. 137, nº 5, p. 1754–1757, 11 fev. 2015. https://doi.org/10.1021/ja511495j.

ALPAYDIN, Ethem. **Introduction to machine learning**. 4°. Cambridge, Massachusetts: The MIT Press, 2020.

ALTMANN, André; TOLOŞI, Laura; SANDER, Oliver; LENGAUER, Thomas. Permutation importance: a corrected feature importance measure. **Bioinformatics**, vol. 26, nº 10, p. 1340–1347, 2010. https://doi.org/10.1093/bioinformatics/btq134.

AMEEN, Ahmed W.; JI, Jing; TAMADDONDAR, Marzieh; MOSHENPOUR, Sajjad; FOSTER, Andrew B.; FAN, Xiaolei; BUDD, Peter M.; MATTIA, Davide; GORGOJO, Patricia. 2D boron nitride nanosheets in PIM-1 membranes for CO2/CH4 separation. **Journal of Membrane Science**, vol. 636, p. 119527, out. 2021. https://doi.org/10.1016/j.memsci.2021.119527.

BARNETT, J. Wesley; BILCHAK, Connor R.; WANG, Yiwen; BENICEWICZ, Brian C.; MURDOCK, Laura A.; BEREAU, Tristan; KUMAR, Sanat K. Designing exceptional gas-separation polymer membranes using machine learning. **Science Advances**, vol. 6, nº 20, 15 maio 2020. https://doi.org/10.1126/sciadv.aaz4301.

BARRER, R. M. The zone of activation in rate processes. **Transactions of the Faraday Society**, vol. 39, p. 237, 1943. https://doi.org/10.1039/tf9433900237.

BARRER, R. M.; RIDEAL, Eric K. Permeation, diffusion and solution of gases in organic polymers. **Transactions of the Faraday Society**, vol. 35, p. 628, 1939. https://doi.org/10.1039/tf9393500628.

BECHHOLD, H. Kolloidstudien mit der Filtrationsmethode. **Zeitschrift für Physikalische Chemie**, vol. 60U, nº 1, p. 257–318, 1 jul. 1907. https://doi.org/10.1515/zpch-1907-6013.

BECK, David A C; CAROTHERS, James M; SUBRAMANIAN, Venkat R; PFAENDTNER, Jim. Data science: Accelerating innovation and discovery in chemical engineering. **AIChE Journal**, vol. 62, nº 5, p. 1402–1416, 2016. https://doi.org/10.1002/aic.15192.

BERK, Zeki. Membrane processes. **Food Process Engineering and Technology**. [*S. l.*]: Elsevier, 2009. p. 233–257. https://doi.org/10.1016/B978-0-12-373660-4.00010-7.

BERNARDO, Gabriel; ARAÚJO, Tiago; DA SILVA LOPES, Telmo; SOUSA, José; MENDES, Adélio. Recent advances in membrane technologies for hydrogen purification. **International Journal of Hydrogen Energy**, vol. 45, nº 12, p. 7313–7338, mar. 2020. https://doi.org/10.1016/j.ijhydene.2019.06.162.

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, Yu.P.; SHANTAROVICH, V.; JANSEN, J.C. Effect of physical aging on the gas transport and sorption in PIM-1 membranes. **Polymer**, vol. 113, p. 283–294, mar. 2017. https://doi.org/10.1016/j.polymer.2016.10.040.

BERNARDO, P.; DRIOLI, E.; GOLEMME, G. Membrane Gas Separation: A Review/State of the Art. **Industrial & Engineering Chemistry Research**, vol. 48, nº 10, p. 4638–4663, 20 maio 2009. https://doi.org/10.1021/ie8019032.

BEZZU, C. Grazia; CARTA, Mariolino; FERRARI, Maria-Chiara; JANSEN, Johannes C.; MONTELEONE, Marcello; ESPOSITO, Elisa; FUOCO, Alessio; HART, Kyle; LIYANA-ARACHCHI, T. P.; COLINA, Coray M.; MCKEOWN, Neil B. The synthesis, chain-packing simulation and long-term gas permeability of highly selective spirobifluorene-based polymers of intrinsic microporosity. **Journal of Materials Chemistry A**, vol. 6, nº 22, p. 10507–10514, 2018. https://doi.org/10.1039/C8TA02601G.

BEZZU, C. Grazia; FUOCO, Alessio; ESPOSITO, Elisa; MONTELEONE, Marcello; LONGO, Mariagiulia; JANSEN, Johannes Carolus; NICHOL, Gary S.; MCKEOWN, Neil B. Ultrapermeable Polymers of Intrinsic Microporosity Containing Spirocyclic Units with Fused Triptycenes. **Advanced Functional Materials**, vol. 31, nº 37, p. 2104474, 25 set. 2021. https://doi.org/10.1002/adfm.202104474.

BOLFARINE, Heleno; BUSSAB, Wilton O. **Elementos de Amostragem**. São Paulo: Edgar Blucher, 2005.

BREIMAN, Leo. Bagging predictors. **Machine Learning**, vol. 24, nº 2, p. 123–140, ago. 1996. https://doi.org/10.1007/BF00058655.

BREIMAN, Leo. Random Forests. **Machine Learning**, vol. 45, nº 1, p. 5–32, 2001. DOI 10.1023/A:1010933404324. Disponível em: https://doi.org/10.1023/A:1010933404324.

CARDOSO, Anne Raquel Teixeira; AMBROSI, Alan; DI LUCCIO, Marco; HOTZA, Dachamir. Membranes for separation of CO2/CH4 at harsh conditions. **Journal of Natural Gas Science and Engineering**, vol. 98, p. 104388, fev. 2022. https://doi.org/10.1016/j.jngse.2021.104388.

CHATTERJEE, Samprit; HADI, Ali S. **Regression analysis by example**. 5°. New York: John Wiley & Sons, 2012.

CHIDAMBARAM, M. **Mathematical Modelling and Simulation in Chemical Engineering**. 1°. [*S. l.*]: Cambridge University Press, 2018.

CHUNG, Tai-Shung; KAFCHINSKI, Edward R; VORA, Rohitkumar H. **SIXEF^{TM}-durene polyimide hollow fibers**. US, US: 5413852, 9 maio 1995.

CRISTIANINI, Nello; SHAWE-TAYLOR, John. **An Introduction to Support Vector Machines and Other Kernel-based Learning Methods**. [*S. l.*]: Cambridge University Press, 2000. https://doi.org/10.1017/CBO9780511801389.

CUI, Z.F.; JIANG, Y.; FIELD, R.W. Fundamentals of Pressure-Driven Membrane Separation Processes. **Membrane Technology**. [*S. l.*]: Elsevier, 2010. p. 1–18. https://doi.org/10.1016/B978-1-85617-632-3.00001-X.

DAGLAR, Hilal; KESKIN, Seda. Combining Machine Learning and Molecular Simulations to Unlock Gas Separation Potentials of MOF Membranes and MOF/Polymer MMMs. **ACS Applied Materials & Interfaces**, vol. 14, nº 28, p. 32134–32148, 20 jul. 2022. https://doi.org/10.1021/acsami.2c08977.

DOBBELAERE, Maarten R.; PLEHIERS, Pieter P.; VAN DE VIJVER, Ruben; STEVENS, Christian V.; VAN GEEM, Kevin M. Machine Learning in Chemical Engineering: Strengths, Weaknesses, Opportunities, and Threats. **Engineering**, vol. 7, nº 9, p. 1201–1211, set. 2021. https://doi.org/10.1016/j.eng.2021.03.019.

DU, Naiying; PARK, Ho Bum; DAL-CIN, Mauro M.; GUIVER, Michael D. Advances in high permeability polymeric membrane materials for CO 2 separations. **Energy Environ. Sci.**, vol. 5, nº 6, p. 7306–7322, 2012. https://doi.org/10.1039/C1EE02668B.

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

FANE, A. G.; WANG, Rong; JIA, Yue. Membrane Technology: Past, Present and Future. **Membrane and Desalination Technologies**. Totowa, NJ: Humana Press, 2011. p. 1–45. https://doi.org/10.1007/978-1-59745-278-6_1.

FENG, Yang; WANG, Zhikun; FAN, Weidong; KANG, Zixi; FENG, Shou; FAN, Lili; HU, Songqing; SUN, Daofeng. Engineering the pore environment of metal–organic framework membranes *via* modification of the secondary building unit for improved gas separation. **Journal of Materials Chemistry A**, vol. 8, nº 26, p. 13132–13141, 2020. https://doi.org/10.1039/C9TA13547B.

FRIEDMAN, Jerome H. Greedy function approximation: A gradient boosting machine. **The Annals of Statistics**, vol. 29, nº 5, 1 out. 2001. https://doi.org/10.1214/aos/1013203451.

GÉRON, Aurélien. **Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow**. 1°. Rio de Janeiro: Alta Books, 2019.

GHANEM, Bader; ALASLAI, Nasser; MIAO, Xiaohe; PINNAU, Ingo. Novel 6FDA-based polyimides derived from sterically hindered Tröger's base diamines: Synthesis and gas permeation properties. **Polymer**, vol. 96, p. 13–19, jul. 2016. https://doi.org/10.1016/j.polymer.2016.04.068.

GLATER, Julius. The early history of reverse osmosis membrane development. **Desalination**, vol. 117, nº 1–3, p. 297–309, set. 1998. https://doi.org/10.1016/S0011-9164(98)00122-2.

GLOROT, Xavier; BENGIO, Yoshua. Understanding the difficulty of training deep feedforward neural networks. **Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings**, vol. 9, p. 249–256, 2010.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep Learning**. [*S. l.*]: MIT Press, 2016.

GRAHAM, Thomas. On the Absorption and Dialytic Separation of Gases by Colloid Septa. **Philosophical Transactions of the Royal Society of London**, vol. 156, p. 399–439, 1866.

GUAN, Jian; HUANG, Tan; LIU, Wei; FENG, Fan; JAPIP, Susilo; LI, Jiali; WANG, Xiaonan; ZHANG, Sui. Design and prediction of metal organic framework-based mixed matrix membranes for CO2 capture via machine learning. **Cell Reports Physical Science**, vol. 3, nº 5, p. 100864, 2022. https://doi.org/10.1016/j.xcrp.2022.100864.

HABERT, Alberto Cláudio; BORGES, Cristiano Piacsek; NOBREGA, Ronaldo. **Processos de separação por membranas**. [*S. l.*]: Editora e-papers, 2006. vol. 3.

HARDENBURGER, Thomas L.; ENNIS, MattheW. Nitrogen. **Kirk-Othmer Encyclopedia of Chemical Technology**. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2005. https://doi.org/10.1002/0471238961.1409201808011804.a01.pub2.

HASHEMIFARD, Seyed Abdollatif; KHOSRAVI, Arash; ABDOLLAHI, Farideh; ALIHEMATI, Zahra; REZAEE, Mohsen. Synthetic polymeric membranes for gas and vapor separations. **Synthetic Polymeric Membranes for Advanced Water Treatment, Gas Separation, and Energy Sustainability**. [*S. l.*]: Elsevier, 2020. p. 217–272. https://doi.org/10.1016/B978-0-12-818485-1.00011-3.

HASNAOUI, Hanaa; KREA, Mohamed; ROIZARD, Denis. Neural networks for the prediction of polymer permeability to gases. **Journal of Membrane Science**, vol. 541, p. 541–549, nov. 2017. https://doi.org/10.1016/j.memsci.2017.07.031.

HAYAKAWA, Eiji; HIMENO, Shuji. Synthesis of all-silica ZSM-58 zeolite membranes for separation of CO2/CH4 and CO2/N2 gas mixtures. **Microporous and Mesoporous Materials**, vol. 291, p. 109695, jan. 2020. https://doi.org/10.1016/j.micromeso.2019.109695.

HAYEK, Ali; YAHAYA, Garba O.; ALSAMAH, Abdulkarim; PANDA, Saroj K. Fluorinated copolyimide membranes for sour mixed-gas upgrading. **Journal of Applied Polymer Science**, vol. 137, nº 5, p. 48336, 5 fev. 2020. https://doi.org/10.1002/app.48336.

HAYKIN, Simon. **Neural networks and learning machines**. 3°. New Jersey: Prentice Hall, 2009.

HENIS, Jay M. S.; TRIPODI, Mary K. A Novel Approach to Gas Separations Using Composite Hollow Fiber Membranes. **Separation Science and Technology**, vol. 15, nº 4, p. 1059–1068, 19 maio 1980. https://doi.org/10.1080/01496398008076287.

HOSSEINI, Seyed Saeid; AZADI TABAR, Mohammad; VANKELECOM, Ivo F.J.; DENAYER, Joeri F.M. Progress in high performance membrane materials and processes for biogas production, upgrading and conversion. **Separation and Purification Technology**, vol. 310, p. 123139, abr. 2023. https://doi.org/10.1016/j.seppur.2023.123139.

HU, Xiaofan; HE, Yabing; WANG, Zhen; YAN, Jingling. Intrinsically microporous co-polyimides derived from ortho-substituted Tröger's Base diamine with a pendant tert-butyl-phenyl group and their gas separation performance. **Polymer**, vol. 153, p. 173–182, set. 2018. https://doi.org/10.1016/j.polymer.2018.08.013.

HU, Xiaofan; LEE, Won Hee; BAE, Joon Yong; ZHAO, Jiayi; KIM, Ju Sung; WANG, Zhen; YAN, Jingling; LEE, Young Moo. Highly permeable polyimides incorporating Tröger's base (TB)

units for gas separation membranes. **Journal of Membrane Science**, vol. 615, p. 118533, dez. 2020. https://doi.org/10.1016/j.memsci.2020.118533.

HU, Xiaofan; LEE, Won Hee; ZHAO, Jiayi; BAE, Joon Yong; KIM, Ju Sung; WANG, Zhen; YAN, Jingling; ZHUANG, Yongbing; LEE, Young Moo. Tröger's Base (TB)-containing polyimide membranes derived from bio-based dianhydrides for gas separations. **Journal of Membrane Science**, vol. 610, p. 118255, set. 2020. https://doi.org/10.1016/j.memsci.2020.118255.

IPCC. Summary for policymakers: Global warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the. 2018. Disponível em: https://archive.ipcc.ch/pdf/special-reports/sr15/sr15_spm_final.pdf.

ISMAIL, Ahmad Fauzi; KHULBE, Kailash Chandra; MATSUURA, Takeshi. Application of Gas Separation Membranes. **Gas Separation Membranes**. Cham: Springer International Publishing, 2015. p. 241–287. https://doi.org/10.1007/978-3-319-01095-3_6.

IULIANELLI, Adolfo; DRIOLI, Enrico. Membrane engineering: Latest advancements in gas separation and pre-treatment processes, petrochemical industry and refinery, and future perspectives in emerging applications. **Fuel Processing Technology**, vol. 206, p. 106464, set. 2020. https://doi.org/10.1016/j.fuproc.2020.106464.

IZBICKI, Rafael; SANTOS, Tiago Mendonça dos. **Aprendizado de máquina: uma abordagem estatística**. 1°. São Carlos, SP: Rafael Izbicki, 2020.

JIANG, Shuangshuang; SHI, Xinli; SUN, Fuxing; ZHU, Guangshan. Fabrication of Crystalline Microporous Membrane from 2D MOF Nanosheets for Gas Separation. **Chemistry – An Asian Journal**, vol. 15, nº 15, p. 2371–2378, 3 ago. 2020. https://doi.org/10.1002/asia.202000143.

JIANG, Shuangshuang; SHI, Xinli; ZU, Yucong; SUN, Fuxing; ZHU, Guangshan. Interfacial growth of 2D MOF membranes *via* contra-diffusion for CO_2 separation. **Materials Chemistry Frontiers**, vol. 5, nº 13, p. 5150–5157, 2021. https://doi.org/10.1039/D1QM00154J.

JOOS, Georg; FREEMAN, Ira M. **Theoretical physics**. 3°. Glasgow, London: Blackie and Son, 1958.

KALAKECH, Carla; SOHAIB, Qazi; LESAGE, Geoffroy; MERICQ, Jean-Pierre. Progress and challenges in recovering dissolved methane from anaerobic bioreactor permeate using membrane contactors: A comprehensive review. **Journal of Water Process Engineering**, vol. 50, p. 103218, dez. 2022. https://doi.org/10.1016/j.jwpe.2022.103218.

KAMBLE, Ashwin R.; PATEL, Chetan M.; MURTHY, Z.V.P. A review on the recent advances in mixed matrix membranes for gas separation processes. **Renewable and Sustainable Energy Reviews**, vol. 145, p. 111062, jul. 2021. https://doi.org/10.1016/j.rser.2021.111062.

KANCHERLA, Ravichand; NAZIA, Shaik; KALYANI, Swayampakula; SRIDHAR, Sundergopal. Modeling and simulation for design and analysis of membrane-based separation processes. **Computers & Chemical Engineering**, vol. 148, p. 107258, maio 2021. https://doi.org/10.1016/j.compchemeng.2021.107258.

KIDA, Koji; MAETA, Yasushi; YOGO, Katsunori. Preparation and gas permeation properties on pure silica CHA-type zeolite membranes. **Journal of Membrane Science**, vol. 522, p. 363–370, jan. 2017. https://doi.org/10.1016/j.memsci.2016.09.002.

KIM, Sungil; KIM, Heeyoung. A new metric of absolute percentage error for intermittent demand forecasts. **International Journal of Forecasting**, vol. 32, nº 3, p. 669–679, jul. 2016. https://doi.org/10.1016/j.ijforecast.2015.12.003.

KINGMA, Diederik P.; BA, Jimmy. Adam: A Method for Stochastic Optimization. 22 dez. 2014.

KUTNER, Michael H.; NACHTSHEIM, Christopher J.; NETER, John; LI, William. **Applied Linear Statistical Models**. 5º. New York: McGraw-Hill, 2005.

LEE, Dong-Joo; LI, Qiming; KIM, Hern; LEE, Kisay. Preparation of Ni-MOF-74 membrane for CO2 separation by layer-by-layer seeding technique. **Microporous and Mesoporous Materials**, vol. 163, p. 169–177, nov. 2012. https://doi.org/10.1016/j.micromeso.2012.07.008.

LI, Guoqiang; KUJAWSKI, Wojciech; VÁLEK, Robert; KOTER, Stanisław. A review - The development of hollow fibre membranes for gas separation processes. **International Journal of Greenhouse Gas Control**, vol. 104, p. 103195, jan. 2021. https://doi.org/10.1016/j.ijggc.2020.103195.

LI, Jian-Rong; KUPPLER, Ryan J; ZHOU, Hong-Cai. Selective gas adsorption and separation in metal–organic frameworks. **Chem. Soc. Rev.**, vol. 38, nº 5, p. 1477–1504, 2009. https://doi.org/10.1039/B802426J.

LI, Xiao; WANG, Yu; BASU, Sumanta; KUMBIER, Karl; YU, Bin. A Debiased MDI Feature Importance Measure for Random Forests. 2019. https://doi.org/10.48550/ARXIV.1906.10845.

LIAN, Haiqian; SONG, Eryue; BAO, Bin; YANG, Wenhe; YANG, Yu; PAN, Yichang; JU, Shengui. Highly steam-stable CHA-type zeolite imidazole framework ZIF-302 membrane for hydrogen separation. **Separation and Purification Technology**, vol. 281, p. 119875, jan. 2022. https://doi.org/10.1016/j.seppur.2021.119875.

LIU, Yunyang; NG, Zhenfu; KHAN, Easir A.; JEONG, Hae-Kwon; CHING, Chi-bun; LAI, Zhiping. Synthesis of continuous MOF-5 membranes on porous α-alumina substrates. **Microporous and Mesoporous Materials**, vol. 118, nº 1–3, p. 296–301, fev. 2009. https://doi.org/10.1016/j.micromeso.2008.08.054.

LOWELL, S.; SHIELDS, Joan E.; THOMAS, Martin A.; THOMMES, Matthias. **Characterization of Porous Solids and Powders: Surface Area, Pore Size and Density**. Dordrecht: Springer Netherlands, 2004. vol. 16. https://doi.org/10.1007/978-1-4020-2303-3.

MATTEUCCI, Scott; YAMPOLSKII, Yuri; FREEMAN, Benny D; PINNAU, Ingo. Transport of Gases and Vapors in Glassy and Rubbery Polymers. **Materials Science of Membranes for Gas and Vapor Separation**. [*S. l.*]: John Wiley & Sons, Ltd, 2006. p. 1–47. https://doi.org/https://doi.org/10.1002/047002903X.ch1.

MAXWELL, J. Clerk. A treatise on electricity and magnetism Dover Publications. **Dover Publications: New York**, vol. 2, 1954.

MCCULLOCH, Warren S.; PITTS, Walter. A logical calculus of the ideas immanent in nervous activity. **The Bulletin of Mathematical Biophysics**, vol. 5, nº 4, p. 115–133, dez. 1943. https://doi.org/10.1007/BF02478259.

MEEK, Scott T; TEICH-MCGOLDRICK, Stephanie L; PERRY, John J; GREATHOUSE, Jeffery A; ALLENDORF, Mark D. Effects of Polarizability on the Adsorption of Noble Gases at Low Pressures in Monohalogenated Isoreticular Metal–Organic Frameworks. **The Journal of**

Physical Chemistry C, vol. 116, nº 37, p. 19765–19772, 20 set. 2012. https://doi.org/10.1021/jp303274m.

MILLER, Tim. Explanation in artificial intelligence: Insights from the social sciences. **Artificial Intelligence**, vol. 267, p. 1–38, 2019. https://doi.org/10.1016/j.artint.2018.07.007.

MITCHELL, John Kearsley. On the penetrativeness of fluids. **Journal of Membrane Science**, vol. 100, nº 1, p. 11–16, 1995. https://doi.org/10.1016/0376-7388(94)00227-P.

MITCHELL, Tom. **Machine Learning**. 1º ed. [*S. l.*]: McGraw-Hill , 1997.

MIVECHIAN, Ali; PAKIZEH, Majid. Hydrogen recovery from Tehran refinery off-gas using pressure swing adsorption, gas absorption and membrane separation technologies: Simulation and economic evaluation. **Korean Journal of Chemical Engineering**, vol. 30, nº 4, p. 937–948, 18 abr. 2013. https://doi.org/10.1007/s11814-012-0221-y.

MOHRI, Mehryar; ROSTAMIZADEH, Afshin; TALWALKAR, Ameet. **Foundations of machine learning**. Cambridge, Massachusetts: The MIT Press, 2018.

MOLNAR, Christoph. **Interpretable Machine Learning**. Morrisville: Lulu Press, 2021.

MOOSAVI, Seyed Mohamad; JABLONKA, Kevin Maik; SMIT, Berend. The Role of Machine Learning in the Understanding and Design of Materials. **Journal of the American Chemical Society**, vol. 142, nº 48, p. 20273–20287, 2 dez. 2020. https://doi.org/10.1021/jacs.0c09105.

MUKAKA, M M. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. **Malawi medical journal : the journal of Medical Association of Malawi**, vol. 24, nº 3, p. 69–71, set. 2012.

NOLLET, J.A. **Leçons de physique expérimentale**. Paris: Chez Hippolyte-Louis Guérin and Louis-François Delatour, 1748.

NUMPY. User Guide. NumPy: the absolute basics for beginners, 3 jul. 2023. Disponível em: https://numpy.org/doc/stable/user/absolute_beginners.html. Acessado em: 2 jul. 2023.

OLSON, Randal S; URBANOWICZ, Ryan J; ANDREWS, Peter C; LAVENDER, Nicole A; KIDD, La Creis; MOORE, Jason H. Automating biomedical data science through tree-based pipeline optimization. 2016.

PAN, Yanqiu; HE, Liu; REN, Yisu; WANG, Wei; WANG, Tonghua. Analysis of Influencing Factors on the Gas Separation Performance of Carbon Molecular Sieve Membrane Using Machine Learning Technique. **Membranes**, vol. 12, nº 1, p. 100, 17 jan. 2022. https://doi.org/10.3390/membranes12010100.

PENG, Yuan; LI, Yanshuo; BAN, Yujie; YANG, Weishen. Two-Dimensional Metal–Organic Framework Nanosheets for Membrane-Based Gas Separation. **Angewandte Chemie International Edition**, vol. 56, nº 33, p. 9757–9761, 7 ago. 2017. https://doi.org/10.1002/anie.201703959.

PERRY, Robert H.; GREEN, Don W.; MALONEY, James O. **Perry's Chemical Engineers' Handbook**. 7th ed. New York: Mc Graw-Hill Book Company, 1997.

PHUNG, Vera Ling Hui; OKA, Kazutaka; HIJIOKA, Yasuaki; UEDA, Kayo; SAHANI, Mazrura; WAN MAHIYUDDIN, Wan Rozita. Environmental variable importance for under-five

mortality in Malaysia: A random forest approach. **Science of The Total Environment**, vol. 845, p. 157312, 2022. https://doi.org/10.1016/j.scitotenv.2022.157312.

PICCIONE, Patrick M. Realistic interplays between data science and chemical engineering in the first quarter of the 21st century: Facts and a vision. **Chemical Engineering Research and Design**, vol. 147, p. 668–675, 2019. https://doi.org/10.1016/j.cherd.2019.05.046.

PINNAU, Ingo; KOROS, William J. **Defect-free ultrahigh flux asymmetric membranes**. US: 4902422, 20 fev. 1990.

PRAVIN, P.S; TAN, Jaswin Zhi Ming; YAP, Ken Shaun; WU, Zhe. Hyperparameter optimization strategies for machine learning-based stochastic energy efficient scheduling in cyber-physical production systems. **Digital Chemical Engineering**, vol. 4, p. 100047, set. 2022. https://doi.org/10.1016/j.dche.2022.100047.

REID, C. E.; BRETON, E. J. Water and ion flow across cellulosic membranes. **Journal of Applied Polymer Science**, vol. 1, nº 2, p. 133–143, mar. 1959. https://doi.org/10.1002/app.1959.070010202.

ROBESON, Lloyd M. The upper bound revisited. **Journal of Membrane Science**, vol. 320, nº 1–2, p. 390–400, jul. 2008. https://doi.org/10.1016/j.memsci.2008.04.030.

ROHATDI, A. **WebPlotDigitizer**, v4.5. Pacifica, California, 2021. Disponível em: https://automeris.io/WebPlotDigitizer

ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. **Psychological Review**, vol. 65, nº 6, p. 386–408, 1958. https://doi.org/10.1037/h0042519.

RUMELHART, David E.; HINTON, Geoffrey E.; WILLIAMS, Ronald J. Learning representations by back-propagating errors. **Nature**, vol. 323, nº 6088, p. 533–536, out. 1986. https://doi.org/10.1038/323533a0.

SADRZADEH, Mohtada; MOHAMMADI, Toraj. **Nanocomposite membranes for water and gas separation**. [*S. l.*]: Elsevier, 2019.

SADRZADEH, Mohtada; REZAKAZEMI, Mashallah; MOHAMMADI, Toraj. Fundamentals and Measurement Techniques for Gas Transport in Polymers. **Transport Properties of Polymeric Membranes**. [*S. l.*]: Elsevier, 2018. p. 391–423. https://doi.org/10.1016/B978-0-12-809884-4.00019-7.

SAINI, Nishel; AWASTHI, Kamlendra. Insights into the progress of polymeric nano-composite membranes for hydrogen separation and purification in the direction of sustainable energy resources. **Separation and Purification Technology**, vol. 282, p. 120029, fev. 2022. https://doi.org/10.1016/j.seppur.2021.120029.

SATO, Shuichi; NAGAI, Kazukiyo. Synthetic polymer-based membranes for acidic gas removal. **Synthetic Polymeric Membranes for Advanced Water Treatment, Gas Separation, and Energy Sustainability**. [*S. l.*]: Elsevier, 2020. p. 173–190. https://doi.org/10.1016/B978-0-12-818485-1.00009-5.

SCHÄF, Oliver; TORTET, Laurence; SIMON-MASSERON, Angélique; PATARIN, Joël; DEFOUR, Stephanie; BLANC, Rosine; COSTE, Christophe; ZEREGA, Yves. 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**, vol. 259, p. 127457, 2020. https://doi.org/10.1016/j.chemosphere.2020.127457.

SEN, Mitali; DANA, Kausik; DAS, Nandini. Development of LTA zeolite membrane from clay by sonication assisted method at room temperature for H2-CO2 and CO2-CH4 separation. **Ultrasonics Sonochemistry**, vol. 48, p. 299–310, nov. 2018. https://doi.org/10.1016/j.ultsonch.2018.06.007.

SMOLA, Alex J.; SCHÖLKOPF, Bernhard. A tutorial on support vector regression. **Statistics and Computing**, vol. 14, nº 3, p. 199–222, ago. 2004. https://doi.org/10.1023/B:STCO.0000035301.49549.88.

SPEISER, Jaime Lynn; MILLER, Michael E; TOOZE, Janet; IP, Edward. A comparison of random forest variable selection methods for classification prediction modeling. **Expert Systems with Applications**, vol. 134, p. 93–101, 2019. https://doi.org/10.1016/j.eswa.2019.05.028.

STEVENS, Douglas; LOEB, Sidney. Reverse osmosis desalination costs derived from the Coalinga pilot plant operation. **Desalination**, vol. 2, nº 1, p. 56–74, jan. 1967. https://doi.org/10.1016/S0011-9164(00)80146-0.

TABE-MOHAMMADI, Abdulreza. A Review of the Applications of Membrane Separation Technology in Natural Gas Treatment. **Separation Science and Technology**, vol. 34, nº 10, p. 2095–2111, 7 dez. 1999. https://doi.org/10.1081/SS-100100758.

THEBELT, Alexander; WIEBE, Johannes; KRONQVIST, Jan; TSAY, Calvin; MISENER, Ruth. Maximizing information from chemical engineering data sets: Applications to machine learning. **Chemical Engineering Science**, vol. 252, p. 117469, 2022. https://doi.org/10.1016/j.ces.2022.117469.

THORNTON, Aaron W.; HILL, James M.; HILL, Anita J. Modelling Gas Separation in Porous Membranes. **Membrane Gas Separation**. Chichester, UK: John Wiley & Sons, Ltd, 2010. p. 85–109. https://doi.org/10.1002/9780470665626.ch5.

TRIOLA, Mario F. **Elementary statistics**. 11°. Boston: Pearson/Addison-Wesley, 2010.

VALAPPIL, Riya Sidhikku Kandath; GHASEM, Nayef; AL-MARZOUQI, Mohamed. Current and future trends in polymer membrane-based gas separation technology: A comprehensive review. **Journal of Industrial and Engineering Chemistry**, vol. 98, p. 103–129, jun. 2021. https://doi.org/10.1016/j.jiec.2021.03.030.

VAPNIK, Vladimir N. **The Nature of Statistical Learning Theory**. New York, NY: Springer New York, 2000. https://doi.org/10.1007/978-1-4757-3264-1.

VOS, Kenneth D.; BURRIS, F. O. Drying Cellulose Acetate Reverse Osmosis Membranes. **Product R&D**, vol. 8, nº 1, p. 84–89, 1 mar. 1969. https://doi.org/10.1021/i360029a016.

WANG, Bin; GAO, Feng; ZHANG, Feng; XING, Weihong; ZHOU, Rongfei. Highly permeable and oriented AlPO-18 membranes prepared using directly synthesized nanosheets for CO_2/CH_4 separation. **Journal of Materials Chemistry A**, vol. 7, nº 21, p. 13164–13172, 2019. https://doi.org/10.1039/C9TA01233H.

WANG, Bin; ZHENG, Yihong; ZHANG, Jinfeng; ZHANG, Wenjuan; ZHANG, Feng; XING, Weihong; ZHOU, Rongfei. Separation of light gas mixtures using zeolite SSZ-13 membranes. **Microporous and Mesoporous Materials**, vol. 275, p. 191–199, fev. 2019. https://doi.org/10.1016/j.micromeso.2018.08.032.

WANG, Jing; TIAN, Kai; LI, Dongyang; CHEN, Muning; FENG, Xiaoquan; ZHANG, Yatao; WANG, Yong; VAN DER BRUGGEN, Bart. Machine learning in gas separation membrane developing: Ready for prime time. **Separation and Purification Technology**, vol. 313, p. 123493, maio 2023. https://doi.org/10.1016/j.seppur.2023.123493.

WANG, Ming; WANG, Zhi; ZHAO, Song; WANG, Jixiao; WANG, Shichang. Recent advances on mixed matrix membranes for CO2 separation. **Chinese Journal of Chemical Engineering**, vol. 25, nº 11, p. 1581–1597, nov. 2017. https://doi.org/10.1016/j.cjche.2017.07.006.

WANG, Qi; YU, Yixuan; LI, Yunhe; MIN, Xiubo; ZHANG, Jin; SUN, Tianjun. Methane separation and capture from nitrogen rich gases by selective adsorption in microporous Materials: A review. **Separation and Purification Technology**, vol. 283, p. 120206, jan. 2022. https://doi.org/10.1016/j.seppur.2021.120206.

WANG, Quanliang; XIA, Changlei; ALAGUMALAI, Krishnapandi; THANH NHI LE, Thi; YUAN, Yan; KHADEMI, Tayebeh; BERKANI, Mohammed; LU, Haiying. Biogas generation from biomass as a cleaner alternative towards a circular bioeconomy: Artificial intelligence, challenges, and future insights. **Fuel**, vol. 333, p. 126456, fev. 2023. https://doi.org/10.1016/j.fuel.2022.126456.

WARD, W.J.; BROWALL, W.R.; SALEMME, R.M. Ultrathin silicone/polycarbonate membranes for gas separation processes. **Journal of Membrane Science**, vol. 1, p. 99–108, jan. 1976. https://doi.org/10.1016/S0376-7388(00)82259-0.

WEI, Pengfei; LU, Zhenzhou; SONG, Jingwen. Variable importance analysis: A comprehensive review. **Reliability Engineering & System Safety**, vol. 142, p. 399–432, out. 2015. https://doi.org/10.1016/j.ress.2015.05.018.

WERE, Kennedy; BUI, Dieu Tien; DICK, Øystein B; SINGH, Bal Ram. 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. **Ecological Indicators**, vol. 52, p. 394–403, 2015. https://doi.org/10.1016/j.ecolind.2014.12.028.

WESSLING, M.; MULDER, M.H.V.; BOS, A.; VAN DER LINDEN, M.; BOS, M.; VAN DER LINDEN, W.E. Modelling the permeability of polymers: a neural network approach. **Journal of Membrane Science**, vol. 86, nº 1–2, p. 193–198, jan. 1994. https://doi.org/10.1016/0376-7388(93)E0168-J.

YAMPOLSKII, Yu.; SHISHATSKII, S.; ALENTIEV, A.; LOZA, K. Group contribution method for transport property predictions of glassy polymers: focus on polyimides and polynorbornenes. **Journal of Membrane Science**, vol. 149, nº 2, p. 203–220, out. 1998. https://doi.org/10.1016/S0376-7388(98)00152-5.

YAN, Y.; BORHANI, T. N.; CLOUGH, P. T. Chapter 14: Machine Learning Applications in Chemical Engineering. **Machine Learning in Chemistry: The impact of Artificial Inteligence**. 1°. London: The Royal Society of Chemistry, 2020.

YANG, Jason; TAO, Lei; HE, Jinlong; MCCUTCHEON, Jeffrey R.; LI, Ying. Machine learning enables interpretable discovery of innovative polymers for gas separation membranes. **Science Advances**, vol. 8, nº 29, 22 jul. 2022. https://doi.org/10.1126/sciadv.abn9545.

YANG, Shansheng; LU, Wencong; CHEN, Nianyi; HU, Qiannan. Support vector regression based QSPR for the prediction of some physicochemical properties of alkyl benzenes. **Journal of**

Molecular Structure: THEOCHEM, vol. 719, nº 1, p. 119–127, 2005. https://doi.org/10.1016/j.theochem.2004.10.060.

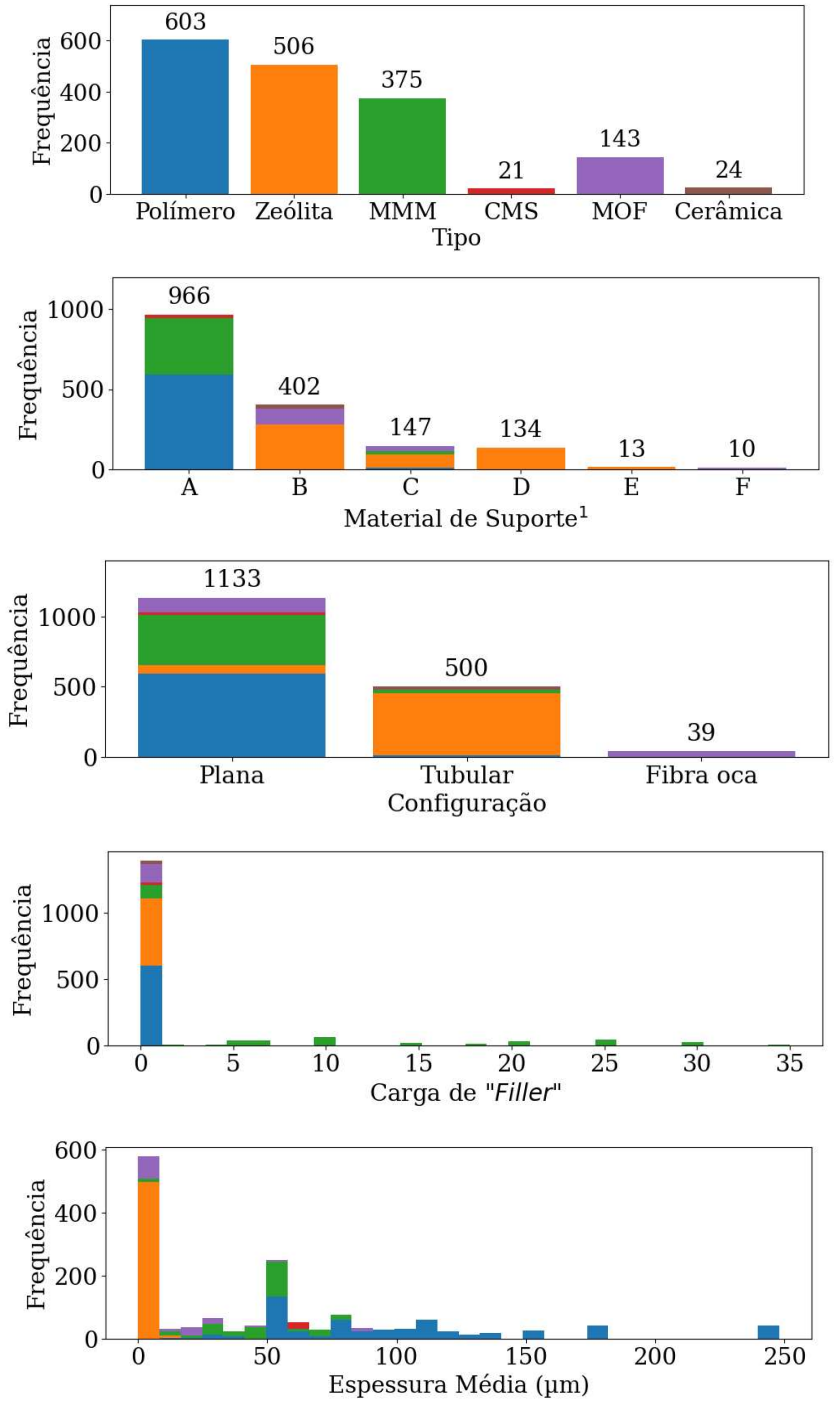
YUAN, Qi; LONGO, Mariagiulia; THORNTON, Aaron W.; MCKEOWN, Neil B.; COMESAÑA-GÁNDARA, Bibiana; JANSEN, Johannes C.; JELFS, Kim E. Imputation of missing gas permeability data for polymer membranes using machine learning. **Journal of Membrane Science**, vol. 627, p. 119207, jun. 2021. https://doi.org/10.1016/j.memsci.2021.119207.

ZHU, Guanghui; KIM, Chiho; CHANDRASEKARN, Anand; EVERETT, Joshua D.; RAMPRASAD, Rampi; LIVELY, Ryan P. Polymer genome–based prediction of gas permeabilities in polymers. **Journal of Polymer Engineering**, vol. 40, nº 6, p. 451–457, 28 jul. 2020. https://doi.org/10.1515/polyeng-2019-0329.

APÊNDICE A - BANCOS DE DADOS ORIGINAIS

Figura A1 – Distribuição de registros dos atributos morfológicos e de processo do Banco de Dados Original de acordo com seus respectivos valores categóricos ou numéricos. Devido à grande variabilidade dos valores/colunas, os atributos “Descrição”, “Subtipo” e “Fração do Gás” não foram analisados. As cores estão classificadas de acordo com o Tipo de cada registro. LEGENDA: ■ Polímero; ■ Zeólita; ■ MMM; ■ CMS; ■ MOF; ■ Cerâmica. [1] A = “Nenhum”; B = “α-alumina”; C = “alumina”; D = “mulita”; E = “silicalite-1”; F = “Óxido de alumínio anodizado”.

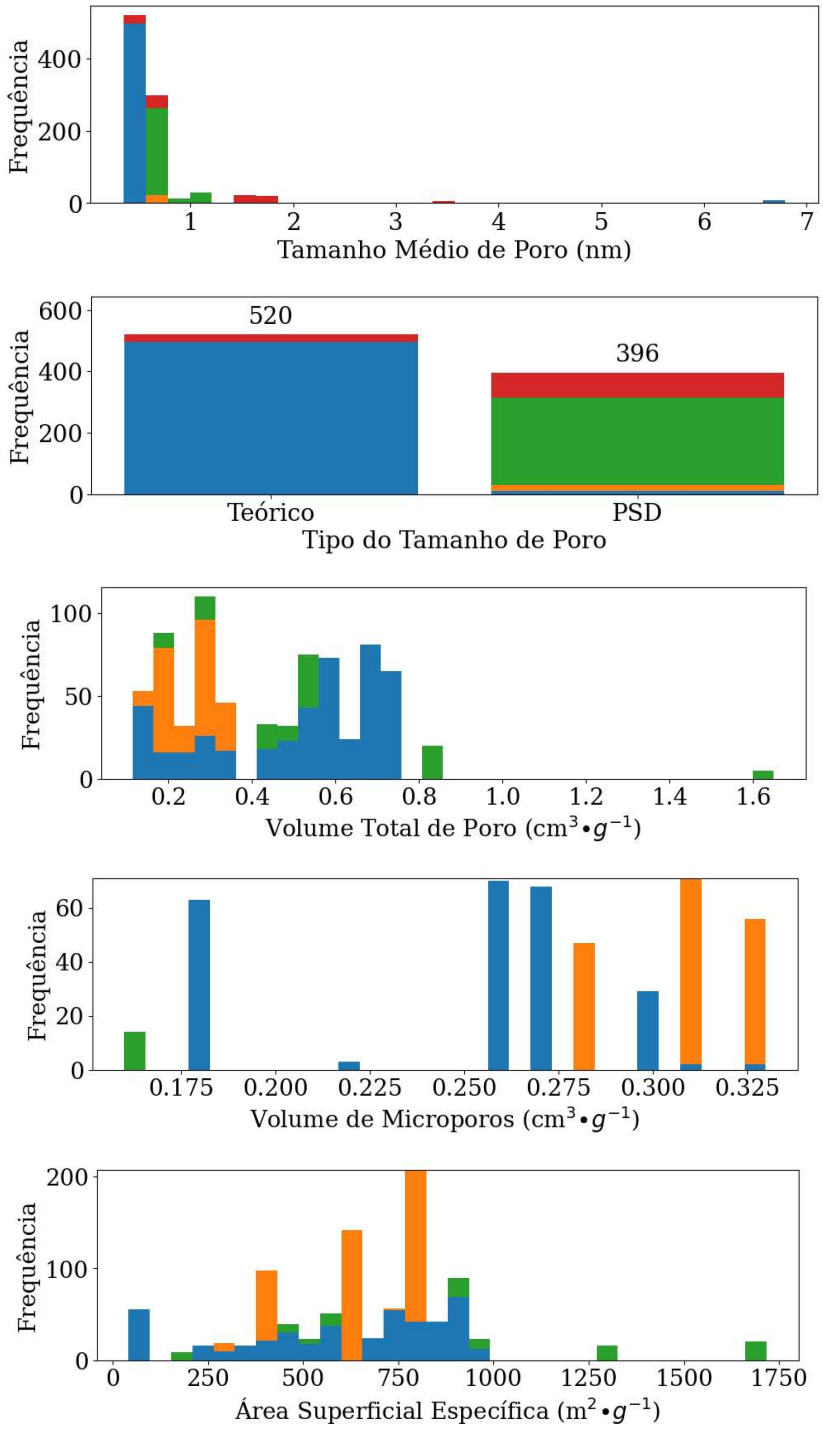
(continua)



Fonte: Elaborada pelo autor.

Figura A1 – Distribuição de registros dos atributos morfológicos e de processo do Banco de Dados Original de acordo com seus respectivos valores categóricos ou numéricos. Devido à grande variabilidade dos valores/colunas, os atributos “Descrição”, “Subtipo” e “Fração do Gás” não foram analisados. As cores estão classificadas de acordo com o Tipo de cada registro. LEGENDA: ■ Polímero; ■ Zeólita; ■ MMM; ■ CMS; ■ MOF; ■ Cerâmica. [1] A = “Nenhum”; B = “α-alumina”; C = “alumina”; D = “mulita”; E = “silicalite-1”; F = “Óxido de alumínio anodizado”.

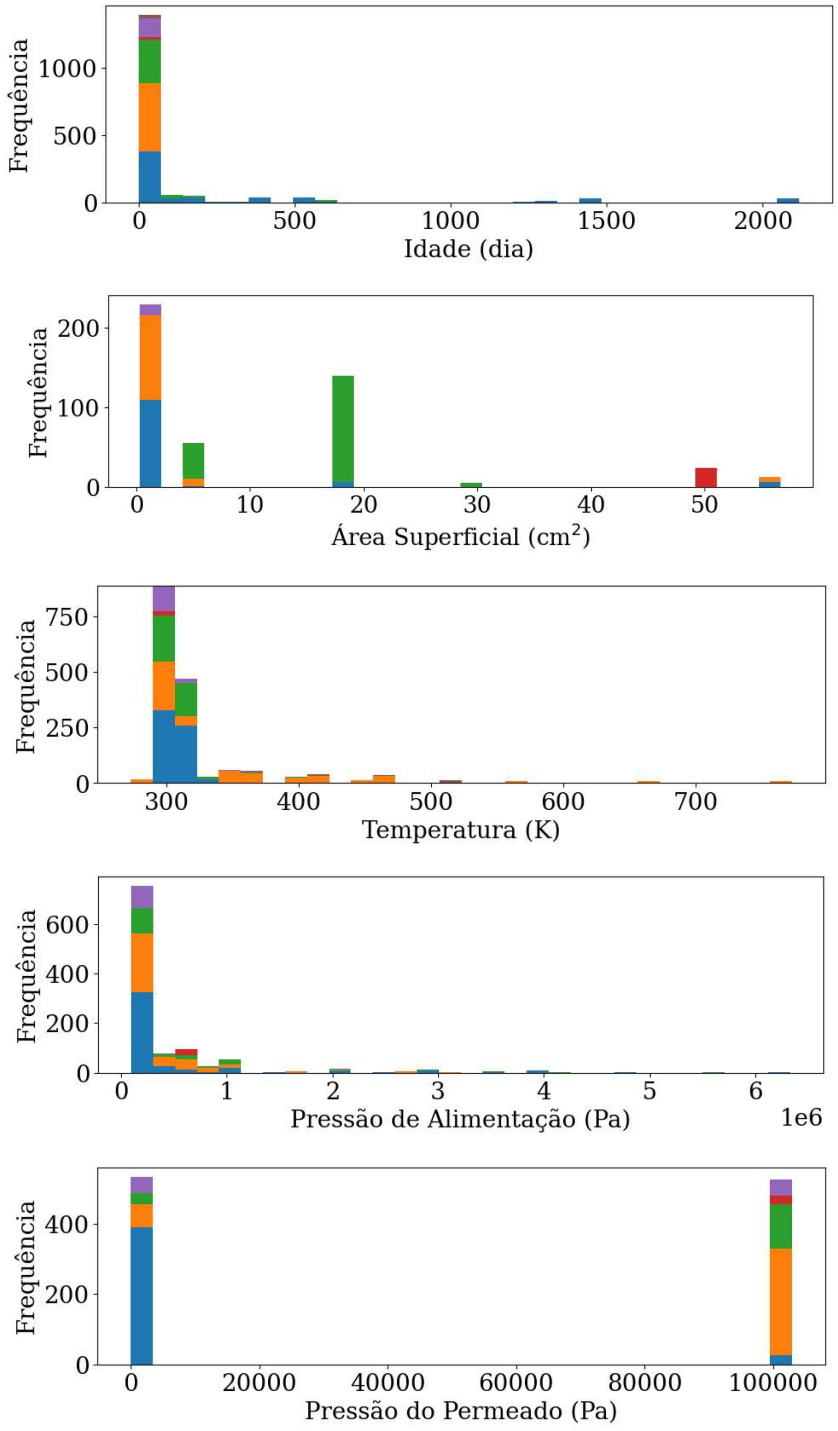
(continuação)



Fonte: Elaborada pelo autor.

Figura A1 – Distribuição de registros dos atributos morfológicos e de processo do Banco de Dados Original de acordo com seus respectivos valores categóricos ou numéricos. Devido à grande variabilidade dos valores/colunas, os atributos “Descrição”, “Subtipo” e “Fração do Gás” não foram analisados. As cores estão classificadas de acordo com o Tipo de cada registro. LEGENDA: ■ Polímero; ■ Zeólita; ■ MMM; ■ CMS; ■ MOF; ■ Cerâmica. [1] A = “Nenhum”; B = “α-alumina”; C = “alumina”; D = “mulita”; E = “silicalite-1”; F = “Óxido de alumínio anodizado”.

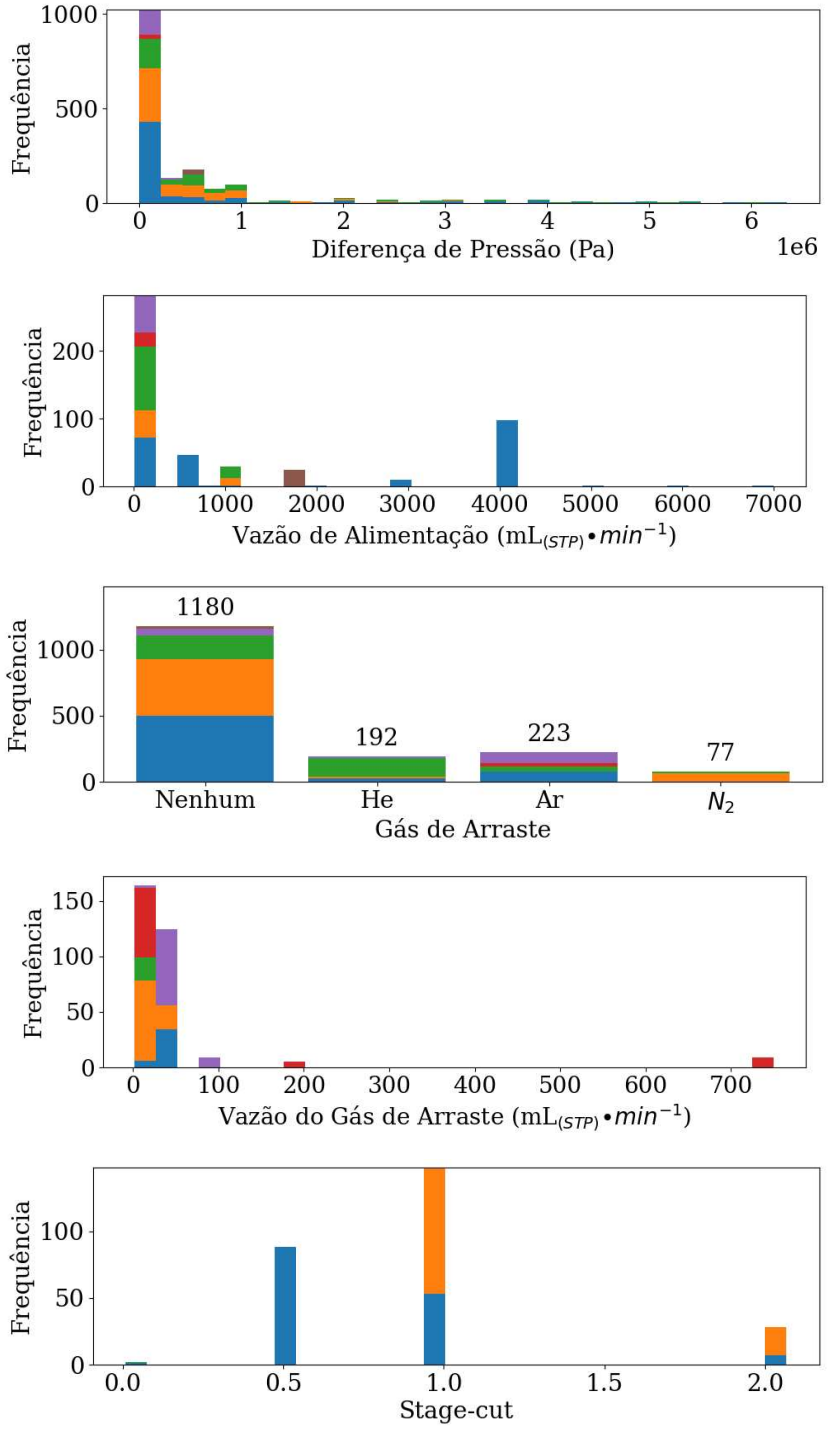
(continuação)



Fonte: Elaborada pelo autor.

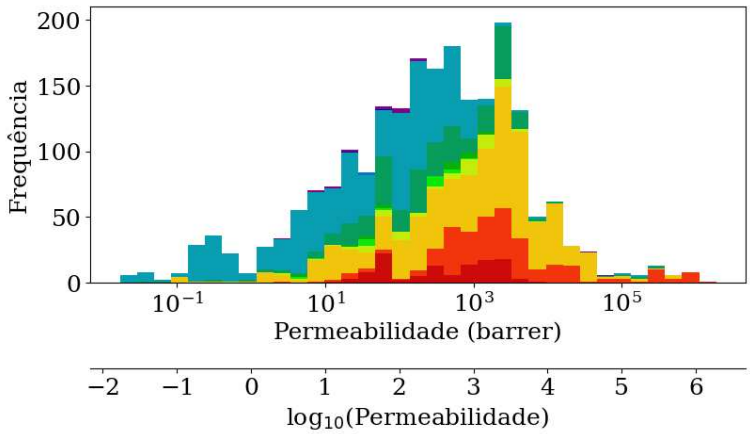
Figura A1 – Distribuição de registros dos atributos morfológicos e de processo do Banco de Dados Original de acordo com seus respectivos valores categóricos ou numéricos. Devido à grande variabilidade dos valores/colunas, os atributos “Descrição”, “Subtipo” e “Fração do Gás” não foram analisados. As cores estão classificadas de acordo com o Tipo de cada registro. LEGENDA: ■ Polímero; ■ Zeólita; ■ MMM; ■ CMS; ■ MOF; ■ Cerâmica. [1] A = “Nenhum”; B = “α-alumina”; C = “alumina”; D = “mulita”; E = “silicalite-1”; F = “Óxido de alumínio anodizado”.

(conclusão)



Fonte: Elaborada pelo autor.

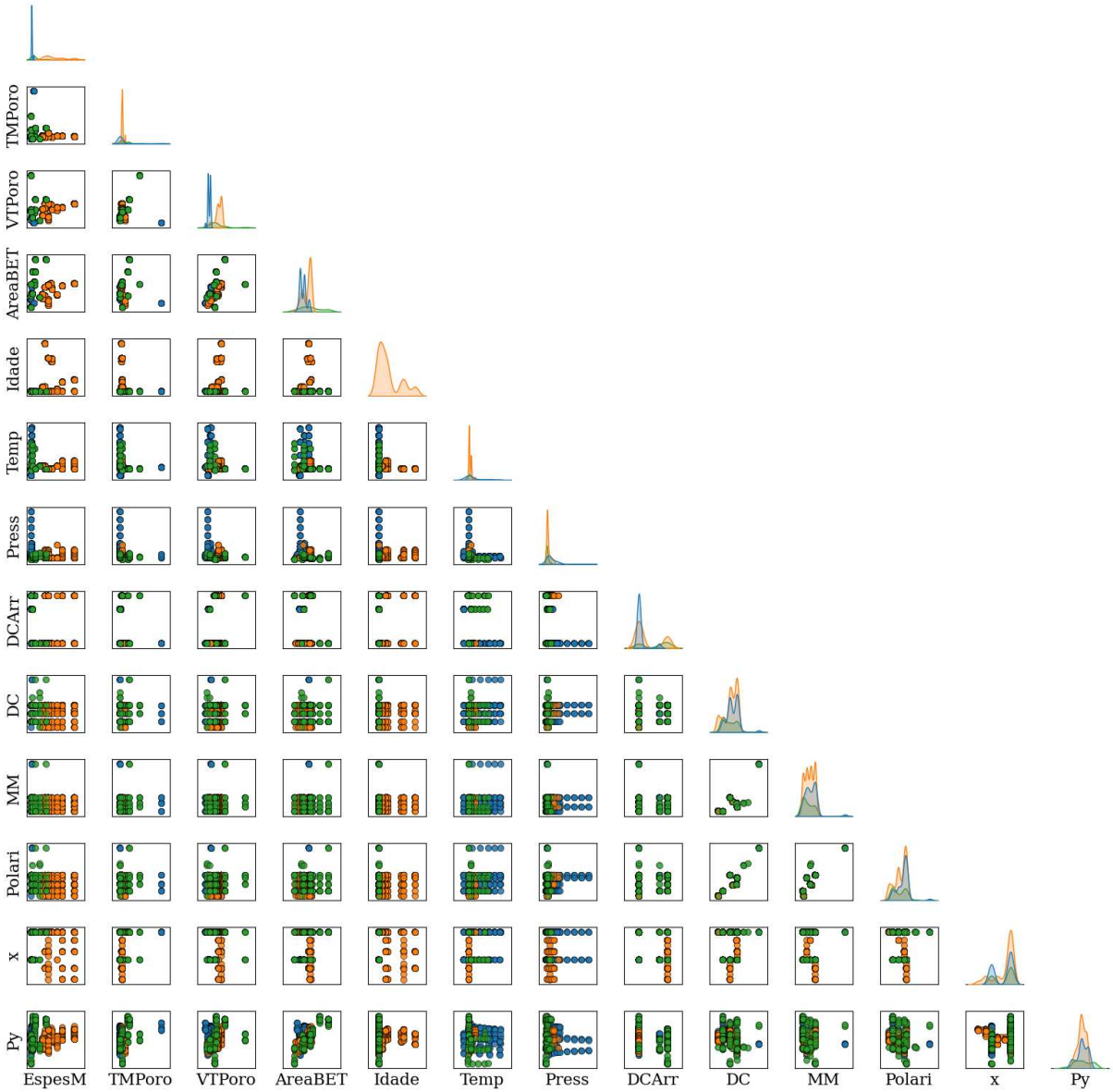
Figura A2 – Histograma dos valores da permeabilidade do Banco de Dados Original de acordo com o gás de alimentação. LEGENDA: ■ He; ■ H2; ■ CO2; ■ O2; ■ H2S; ■ CO; ■ N2; ■ CH4; ■ C2H4; ■ C2H6; ■ SF6.



Fonte: Elaborada pelo autor.

APÊNDICE B – BANCO DE DADOS UTILIZADO NA MODELAGEM

Figura B1 – Matriz dos gráficos de dispersão para os atributos do banco de dados. LEGENDA das Cores: ■ Zeólita; ■ Polímero; ■ MOF. “EspesM” = Espessura Média; “TMPoro” = Tamanho Médio de Poro; “VTPoro” = Volume Total de Poro; “AreaBET” = Área Superficial Específica; “Idade” = Idade; “Temp” = Temperatura; “Press” = Diferença de Pressão; “DCArr” = Diâmetro Cinético do Gás de Arraste; “DC” = Diâmetro Cinético do Gás de Alimentação; “MM” = Massa Molar do Gás de Alimentação; “Polari” = Polarizabilidade do Gás de Alimentação; “x” = Fração do Gás; “Py” = Permeabilidade.



Fonte: Elaborada pelo autor.

Tabela B1 – Detalhes sobre as referências utilizadas para a construção do banco de dados.

(continua)

Número	Referência	Quant. de registros	Material(is)[1]	Gases
Atributo Tipo = Zeólita				
1	(WANG, Bin; GAO; *et al.*, 2019)	84	Nanofolhas de AlPO-18 (AEI)	CO_2, H_2, N_2 e CH_4
2	(KIDA; MAETA; YOGO, 2017)	32	Si-CHA e SSZ-13	H_2, CO_2, N_2, SF_6 e CH_4
3	(AKHTAR *et al.*, 2015)	22	Silicalita-1 (MFI)	He, H_2, CO_2 e N_2
4	(HAYAKAWA; HIMENO, 2020)	84	ZSM-58 (DDR)	CO_2, CH_4 e N_2
5	(SEN; DANA; DAS, 2018)	9	De topologia LTA originada de argila bentonita (montmorilonita)	H_2, CO_2 e N_2
Atributo Tipo = Polímero				
6	(BEZZU *et al.*, 2018)	112	Cinco polímeros diferentes, todos PIM denominados PIM-SBI. A unidade SBI (espirobisindano) do PIM-1 foi substituída por SBF (espirobifluoreno)	He, H_2, O_2, N_2, CO_2 e CH_4
7	(BEZZU *et al.*, 2021)	108	Dois polímeros, o PIM-SBI-Trip (com a unidade de espirobisindano e tripticeno) e o copolímero PIM-1/PIM-SBI-Trip	N_2, O_2, CO_2, CH_4, H_2 e He
8	(HU; LEE; BAE; *et al.*, 2020)	26	Dois polímeros PITB (do inglês, *Polyimide Troger's Base*) preparados a partir de bioprodutos de liginina diferentes (MMDA e FDDA)	He, H_2, CO_2, O_2, N_2 e CH_4
9	(HU; LEE; ZHAO; *et al.*, 2020)	36		
10	(GHANEM *et al.*, 2016)	24	Dois polímeros sintetizados com 6FDA e dois novos monômeros de diamina baseados em TB impedidos estericamente (denominados TBDA1 e TBDA2)	He, H_2, N_2, O_2, CH_4 e CO_2
11	(HU *et al.*, 2018)	30	Seis copolímeros constituídos por diferentes poliimidas contendo TB (TB-PIMPIs)	He, CO_2, O_2, N_2 e CH_4

[1] PIM = *Polymers of Intrinsic Microporosity*; TB = *Troger's Base*; PI = Poliimida; SBU = *Secondary Building Unit*. Fonte: Elaborada pelo autor.

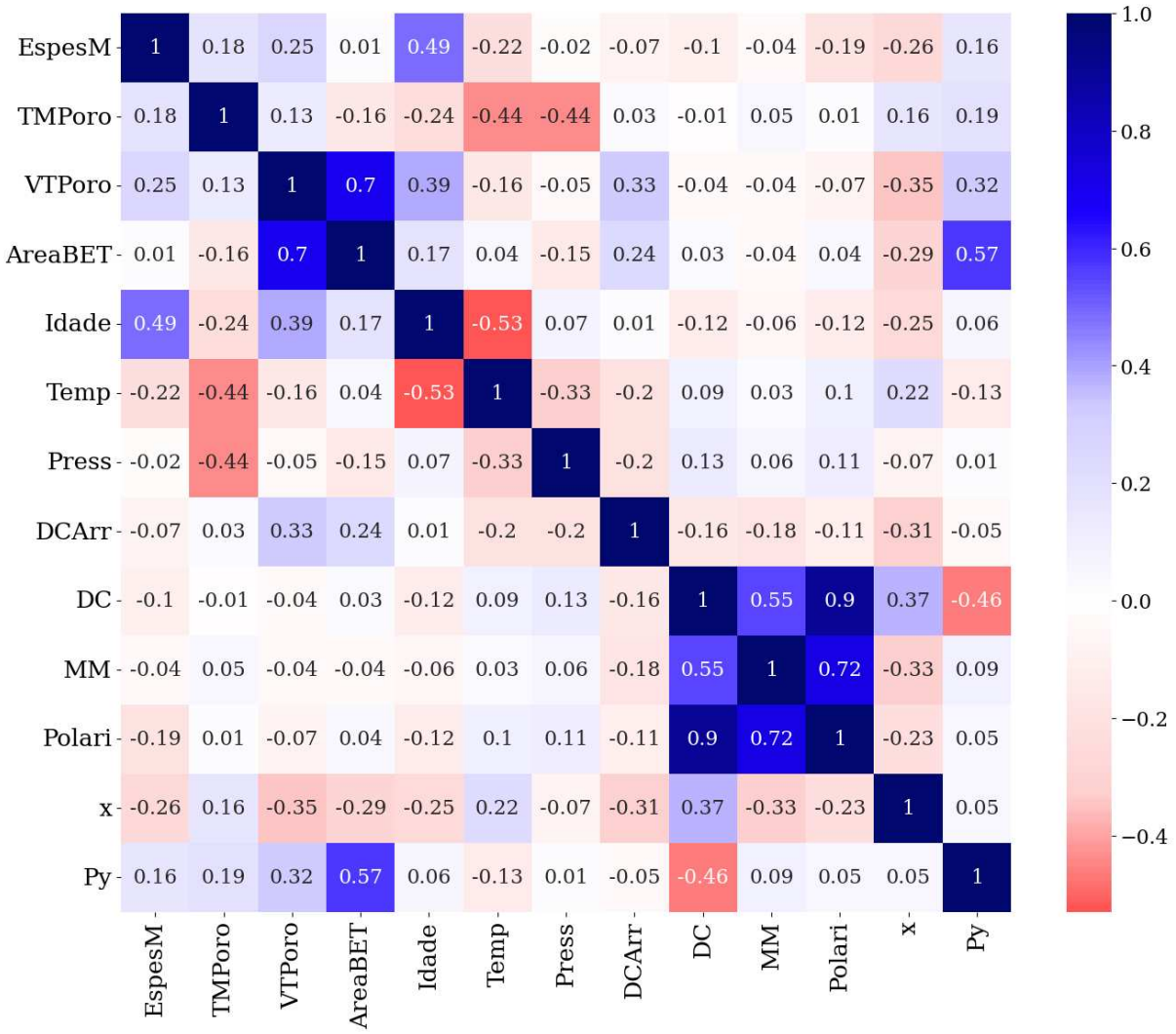
Tabela B1 – Detalhes sobre as referências utilizadas para a construção do banco de dados.

(conclusão)

Número	Referência	Quant. de registros	Material(is)[1]	Gases
Atributo Tipo = MOF				
12	(FENG *et al.*, 2020)	28	Dois cristais denominados soc-MOF (PCN-250) e soc-MOF-IM (em que uma SBU foi modificado com imidazol)	H_2, CO_2, N_2 e CH_4
13	(LEE *et al.*, 2012)	16	Ni-MOF-74	H_2, CO_2, N_2 e CH_4
14	(AL-MAYTHALONY *et al.*, 2015)	10	ZMOF-1 com topologia de sodalita (sod-ZMOF)	He, H_2, N_2, O_2, CH_4, CO_2, C_2H_4 e C_2H_6
15	(JIANG *et al.*, 2021)	7	Nanofolhas de $Ni_3(HITP)_2$ (HITP = 2,3,6,7,10,11-hexaaminotrifenilene)	CO_2, N_2 e CH_4
16	(JIANG *et al.*, 2020)	7		
17	(PENG *et al.*, 2017)	14	Nanofolhas de $[Zn_2(benzimidazole)_3(OH)(H_2O)]_n$	H_2, CO_2, N_2 e CH_4
18	(LIAN *et al.*, 2022)	23	ZIF-302 (CHA)	H_2, CO_2, N_2, CH_4 e C_2H_4
19	(LIU *et al.*, 2009)	20	MOF-5	H_2, CO_2, N_2, CH_4 e SF_6

[1] PIM = *Polymers of Intrinsic Microporosity*; TB = *Troger's Base*; PI = Poliimida; SBU = *Secondary Building Unit*. Fonte: Elaborada pelo autor.

Figura B2 – Matriz de correlação de postos de Spearman entre os atributos numéricos do banco de dados. LEGENDA: “EspesM” = Espessura Média; “TMPoro” = Tamanho Médio de Poro; “VTPoro” = Volume Total de Poro; “AreaBET” = Área Superficial Específica; “Idade” = Idade; “Temp” = Temperatura; “Press” = Diferença de Pressão; “DCArr” = Diâmetro Cinético do Gás de Arraste; “DC” = Diâmetro Cinético do Gás de Alimentação; “MM” = Massa Molar do Gás de Alimentação; “Polari” = Polarizabilidade do Gás de Alimentação; “x” = Fração do Gás; “Py” = Permeabilidade.



Fonte: Elaborada pelo autor.

APÊNDICE C – INFORMAÇÕES ADICIONAIS SOBRE A MODELAGEM

Tabela C1 – Tempo de execução e respectivos melhores valores encontrados na otimização dos hiperparâmetros do modelo Florestas Aleatórias utilizando as diferentes funções objetivo.

	F_0	**F_1**	**F_2**	**F_3**	**F_4**	**F_5**	**F_6**
Tempo de execução	18min	1h20min	1h14min	1h16min	1h18min	1h22min	1h30min
Melhor valor para F_{obj} ($\cdot 10^{-3}$)	3,3594	3,7979	25,1426	47,2176	18,9551	48,6296	77,5635

Fonte: Elaborada pelo autor.

Tabela C2 – Tempo de execução e respectivos melhores valores encontrados na otimização dos hiperparâmetros do modelo Máquinas de Vetores de Suporte utilizando as diferentes funções objetivo.

	F_0	**F_4**	**F_5**
Tempo de execução	1h41min	6h52min	19h32min
Melhor valor para F_{obj} ($\cdot 10^{-3}$)	1,3451	0,4383	49,9687

Fonte: Elaborada pelo autor.

Tabela C3 – Tempo de execução e respectivo melhor valor encontrado na otimização dos hiperparâmetros do modelo Redes Neurais Artificiais.

	F_0
Tempo de execução	3h15min
Melhor valor para F_{obj} ($\cdot 10^{-3}$)	139,37

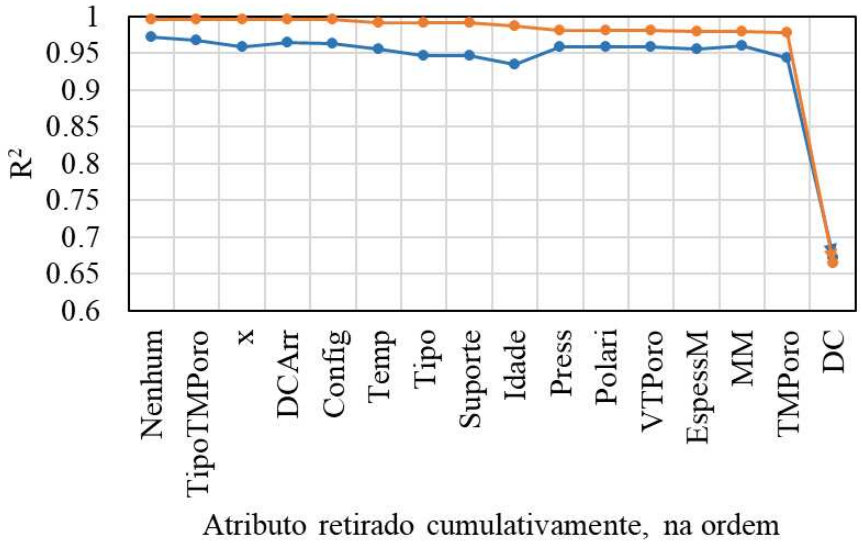
Fonte: Elaborada pelo autor.

Tabela C4 – Hiperparâmetros avaliados durante a otimização do modelo Florestas Aleatórias ao retirar os atributos de entrada de forma cumulativa. A função objetivo F_4 foi utilizada. O hiperparâmetro *min_samples_leaf* teve o valor 1 em todos os casos. FAIXA DE PROCURA: *estimators* = [1, 600]; *criterion* = [SE, AE]; *max_depth* = [2, 70]; *min_samples_leaf* = [1,4]; *max_features* = [1, máximo de atributos]; *max_samples* = [0,5, máximo de amostras]; *cv* = [2, 10].

Atributos retirados, na ordem:	***n_estimators***	***criterion*[1]**	***max_depth***	***max_features***	***max_samples***	***cv***
Nenhum	245	SE	27	5	1,00	24
Tipo do Tamanho de Poro	140	SE	32	7	1,00	10
Fração do Gás	73	SE	39	4	1,00	10
Diâmetro Cinético do Gás de Arraste	371	SE	20	3	1,00	10
Configuração	371	SE	50	3	1,00	10
Temperatura	330	SE	17	2	1,00	9
Tipo	277	SE	58	3	1,00	10
Material de Suporte	518	SE	16	3	1,00	10
Idade	467	SE	68	3	1,00	4
Diferença de Pressão	129	SE	51	3	1,00	10
Polarizabilidade do Gás de Alimentação	141	SE	36	2	1,00	10
Volume Total de Poro	145	SE	70	2	1,00	6
Espessura Média	129	SE	42	2	1,00	10
Massa Molar do Gás de Alimentação	75	AE	47	2	0,99	7
Tamanho Médio de Poro	223	AE	22	2	1,00	9
Diâmetro Cinético do Gás de Alimentação	1	SE	62	1	0,92	2

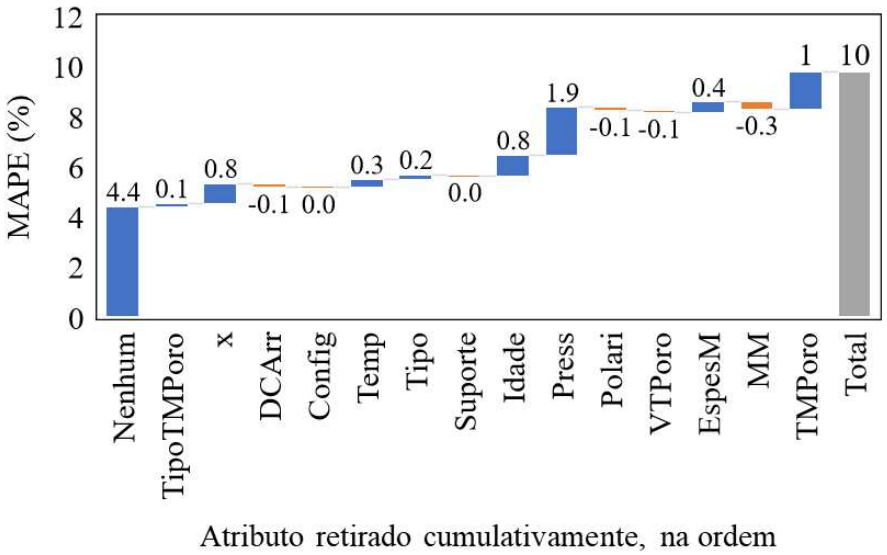
SE = Erro quadrático (em inglês, *Squared Error*); AE = Erro Absoluto (em inglês, *Absolute Error*). Fonte: Elaborada pelo autor.

Figura C1 – R^2 dos modelos FAs otimizados a cada retirada sucessiva de um atributo de entrada. LEGENDA: ■ Conjunto de treino; ■ Conjunto de teste. “TipoTMPoro” = Tipo do Tamanho de Poro; “x” = Fração do Gás; “DCArr” = Diâmetro Cinético do Gás de Arraste; “Config” = Configuração; “Temp” = Temperatura; “Tipo” = Tipo; “Suporte” = Material de Suporte; “Idade” = Idade; “Press” = Diferença de Pressão; “Polari” = Polarizabilidade do Gás de Alimentação; “VTPoro” = Volume Total de Poro; “EspesM” = Espessura Média; “MM” = Massa Molar do Gás de Alimentação; “TMPoro” = Tamanho Médio de Poro; “DC” = Diâmetro Cinético do Gás de Alimentação.



Fonte: Elaborada pelo autor.

Figura C2 – Gráfico de cascata do MAPE do conjunto teste para cada remoção sucessiva de atributo. LEGENDA: ■ Aumento; ■ Diminuição; ■ Total. “TipoTMPoro” = Tipo do Tamanho de Poro; “x” = Fração do Gás; “DCArr” = Diâmetro Cinético do Gás de Arraste; “Config” = Configuração; “Temp” = Temperatura; “Tipo” = Tipo; “Suporte” = Material de Suporte; “Idade” = Idade; “Press” = Diferença de Pressão; “Polari” = Polarizabilidade do Gás de Alimentação; “VTPoro” = Volume Total de Poro; “EspesM” = Espessura Média; “MM” = Massa Molar do Gás de Alimentação; “TMPoro” = Tamanho Médio de Poro; “DC” = Diâmetro Cinético do Gás de Alimentação.



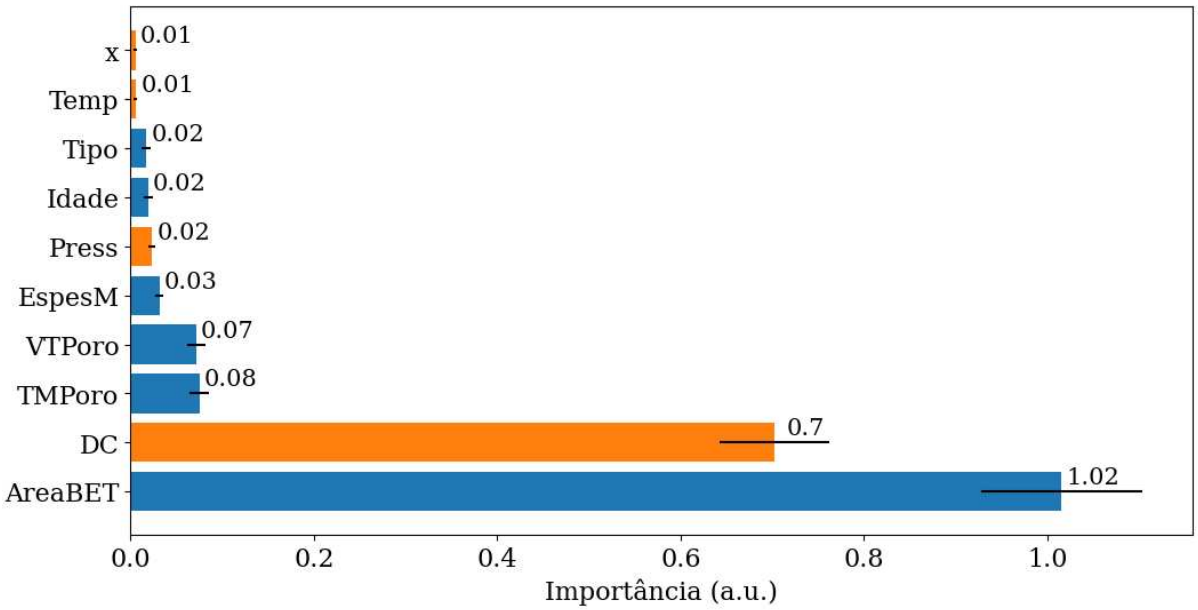
Fonte: Elaborada pelo autor.

Tabela C5 – Hiperparâmetros avaliados durante a otimização do modelo Florestas Aleatórias utilizando a função objetivo F_4, após a poda dos atributos de entrada.

Hiperparâmetro [1]	**Faixa de procura** [2]	**Valores ótimos**
Número de Árvores (*n_estimators*)	1 – 600 (1)	355
Critério (*criterion*)	SE e AE[3]	SE
Profundidade Máxima (*max_depth*)	2 – 70 (1)	23
Mínimo de Folhas (*min_samples_leaf*)	1 – 4 (1)	1
Número de Atributos (*max_features*)	1 – 10 (1)	8
Tamanho da Amostra (*max_samples*)	50 – 100 % do tamanho do conjunto de treino (1%)	100 %
K-fold (*cv*)	2 – 10 (1)	9

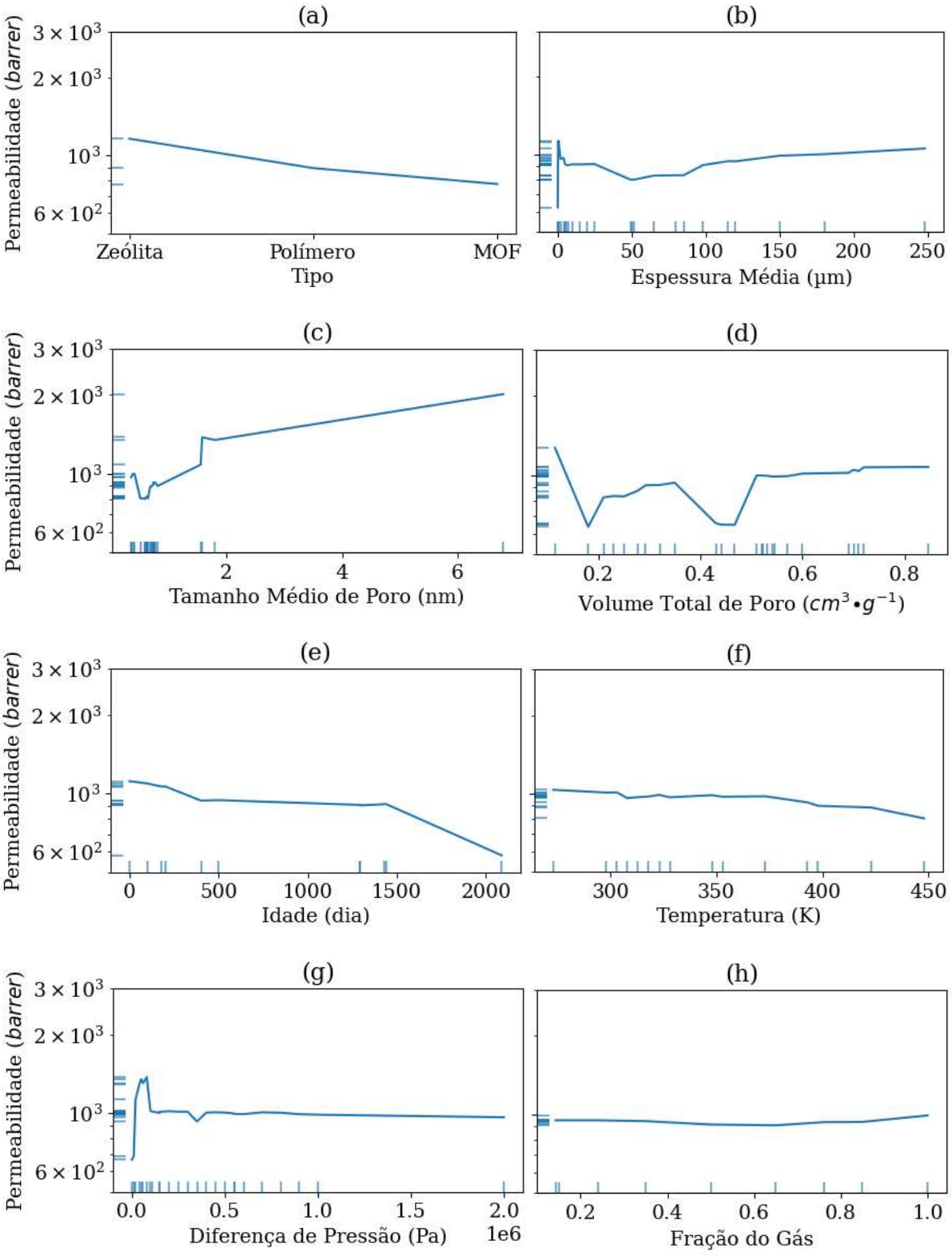
[1] O nome em parênteses se refere ao nome do hiperparâmetro na biblioteca computacional. [2] O valor entre parênteses corresponde ao passo (*step*) definido para cada faixa. [3] SE = Erro quadrático (em inglês, *Squared Error*); AE = Erro Absoluto (em inglês, *Absolute Error*). Fonte: Elaborada pelo autor.

Figura C3 – Importâncias de cada atributo de entrada do modelo FA (após a retirada das variáveis ruído) para a predição da permeabilidade. LEGENDA: ■ Atributos morfológicos; ■ Atributos de processo. “AreaBET” = Área Superficial Específica; “DC” = Diâmetro Cinético do Gás de Alimentação; “TMPoro” = Tamanho Médio de Poro; “VTPoro” = Volume Total de Poro; “EspesM” = Espessura Média; “Press” = Diferença de Pressão; “Idade” = Idade; “Tipo” = Tipo; “Temp” = Temperatura; “x” = Fração do Gás.



Fonte: Elaborada pelo autor.

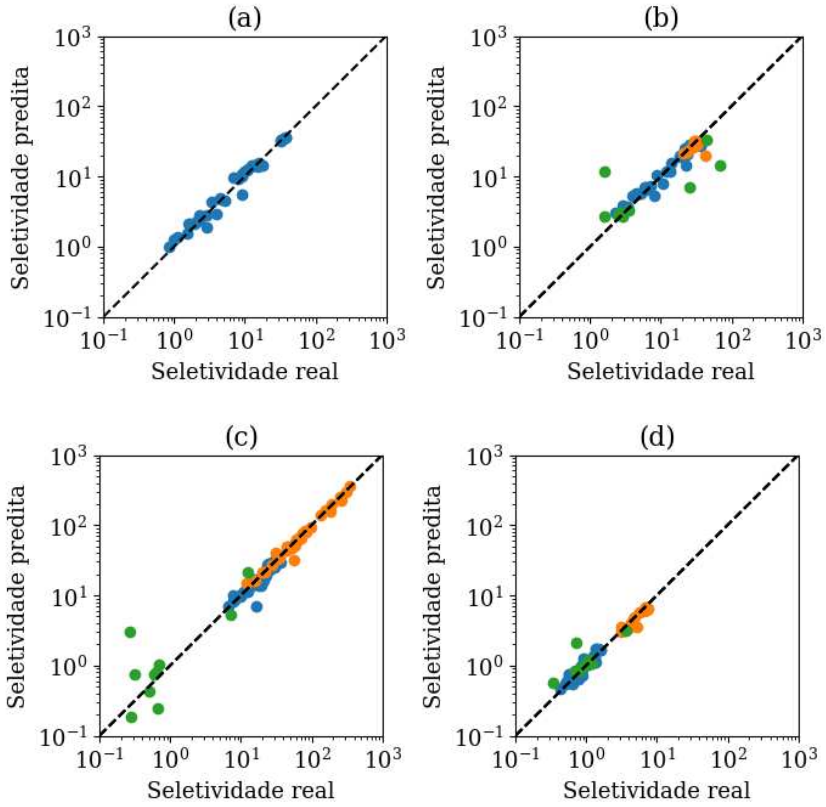
Figura C4 – Gráficos de Dependência Parcial da Permeabilidade em função do (a) Tipo, (b) Espessura Média, (c) Tamanho Médio de Poro, (d) Volume Total de Poro, (e) Idade, (f) Temperatura, (g) Diferença de Pressão e (h) Fração do Gás.



Fonte: Elaborada pelo autor.

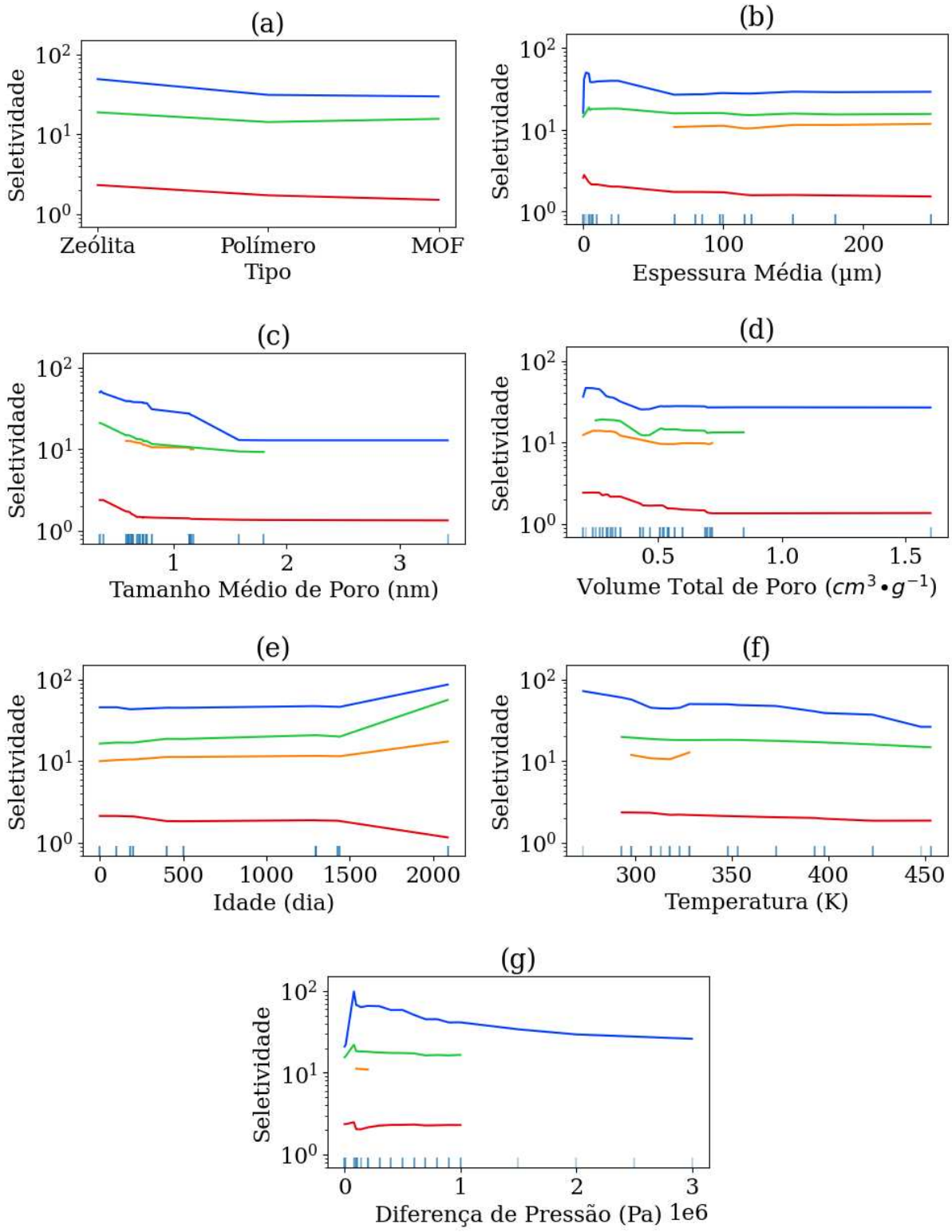
APÊNDICE D – SELETIVIDADE

Figura D1 – Gráficos de dispersão entre as seletividades reais e preditas dos sistemas (a) He/CH_4, (b) H_2/CH_4, (c) CO_2/CH_4 e (d) N_2/CH_4 utilizando o modelo FA treinado anteriormente. LEGENDA: ■ Polímero; ■ Zeólita; ■ MOF.



Fonte: Elaborada pelo autor.

Figura D2 – Gráficos de Dependência Parcial da Seletividade de cada sistema binário em função do (a) Tipo, (b) Espessura Média, (c) Tamanho Médio de Poro, (d) Volume Total de Poro, (e) Idade, (f) Temperatura e (g) Diferença de Pressão. LEGENDA: ■ He/CH_4; ■ H_2/CH_4; ■ CO_2/CH_4; ■ N_2/CH_4.



Fonte: Elaborada pelo autor.