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

ABORDAGENS DE APRENDIZADO DE MÁQUINA PARA PREVER O EQUILÍBRIO DE ADSORÇÃO DE GASES LEVES EM ZEÓLITAS, CARVÕES ATIVADOS E REDES METALORGÂNICAS

THAYLANE DA ROCHA BEZERRA

Uberlândia – MG
2023

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

ABORDAGENS DE APRENDIZADO DE MÁQUINA PARA PREVER O EQUILÍBRIO DE ADSORÇÃO DE GASES LEVES EM ZEÓLITAS, CARVÕES ATIVADOS E REDES METALORGÂNICAS

Dissertação submetida ao Programa de Pós-Graduação em Engenharia Química da Universidade Federal de Uberlândia como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Química.

Mestranda: Thaylane da Rocha Bezerra
Orientadora: Profa. Dra. Sarah Arvelos Altino
Coorientador: Prof. Ubirajara Coutinho Filho

Uberlândia – MG
2023

Ficha Catalográfica Online do Sistema de Bibliotecas da UFU

com dados informados pelo(a) próprio(a) autor(a).

B574 2023

Bezerra, Thaylane da Rocha, 1993-

ABORDAGENS DE APRENDIZADO DE MÁQUINA PARA PREVER O EQUILÍBRIO DE ADSORÇÃO DE GASES LEVES EM ZEÓLITAS, CARVÕES ATIVADOS E REDES METALORGÂNICAS [recurso eletrônico] / Thaylane da Rocha Bezerra. - 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.2023.617
Inclui bibliografia.
Inclui ilustrações.

1. Engenharia química. I. Altino, Sarah Arvelos ,1986-, (Orient.). II. Filho, Ubirajara Coutinho ,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, 15/2023, PPGEQ				
Data:	27 de novembro de 2023	Hora de início:	15:00	Hora de encerramento:	17:00
Matrícula do Discente:	12122EQU011				
Nome do Discente:	Thaylane da Rocha Bezerra				
Título do Trabalho:	ABORDAGENS DE APRENDIZADO DE MÁQUINA PARA PREVER O EQUILÍBRIO DE ADSORÇÃO DE GASES LEVES EM ZEÓLITAS, CARVÕES ATIVADOS E REDES METALORGÂNICAS				
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 da candidata.

Iniciando os trabalhos a presidente da mesa, Profª Drª Sarah Arvelos Altino, apresentou a Comissão Examinadora e a candidata, agradeceu a presença do público, e concedeu à discente a palavra para a exposição do seu trabalho. A duração da apresentação da 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 a candidata. Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando a candidata:

Aprovada

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 27/11/2023, às 17:00, 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 27/11/2023, às 17:01, 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 27/11/2023, às 17:01, 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 27/11/2023, às 17:01, 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 **4993877** e o código CRC **A2F04825**.

Dedico aos meus pais Cícero e Tânia, meu irmão Thaylan e meu esposo Vandeyberg. Pessoas que muito amo.

AGRADECIMENTOS

Agradeço em primeiro lugar a Deus, por ter me dado forças e persistência para vencer mais este desafio. Obrigada por seu amor incondicional, sua presença em minha vida e por cada graça recebida. Que eu nunca me esqueça de tudo o que fizeste por mim. Teu amor me envolve, minha devoção é Tua!

Aos meus pais, José Cícero Bezerra e Maria Tânia da Rocha Bezerra, expresso minha profunda gratidão por todo incentivo e amor. Agradeço por estarem ao meu lado em todas as etapas da minha vida. Cada conquista em minha vida é fruto do esforço e dedicação de vocês a mim. Sou grata ainda ao meu irmão Thaylan Benício Rocha Bezerra por toda alegria que você trouxe à minha vida, por ter em você não somente um irmão, mas também um amigo. Eu amo vocês!

Ao meu amado esposo, Vandeyberg Nogueira de Souza, agradeço por seu apoio incondicional durante essa jornada. Seu suporte e dedicação em me encorajar foram fundamentais para alcançar esse marco. Sou grata por ter um companheiro tão extraordinário ao meu lado. Te amo, bem!

À Universidade Federal de Uberlândia, em especial à Faculdade de Engenharia Química, por ter aberto suas portas para mim e me fornecido oportunidades únicas. Obrigada a todos os professores, servidores e colegas de turma.

À minha orientadora Sarah Arvelos Altino, expresso minha sincera gratidão pela sua orientação, seu comprometimento e dedicação. Ter você como orientadora foi realmente uma bênção, sou extremamente grata por sua disponibilidade e por todos os ensinamentos que recebi ao seu lado.

Ao meu líder Raul Santana Filho, quero agradecer o incentivo, suporte e flexibilidade que me foram concedidos durante esse período. Sua compreensão foi fundamental para que eu conciliasse as demandas profissionais e acadêmicas.

A cada um de vocês, sou profundamente grata por fazerem parte dessa realização significativa em minha vida. Sem o carinho, apoio e motivação de vocês, esse marco não teria sido possível. Mais uma vez, meu sincero agradecimento a todos vocês!

“A vida é muito curta para ser pequena”

Benjamin Disraeli

RESUMO

Este estudo teve como objetivo extrair conhecimento sobre a adsorção de gases leves por materiais microporosos (zeólitas, MOFs e carvões de ativação) através de análise de dados e algoritmos de aprendizado de máquina: K-vizinhos mais próximos (KNN), Árvores de Decisão (AD) e Regressão por Vetores de Suporte (RVS) dos dados reportados em 22 artigos publicados entre os anos 1974 e 2022. Um banco de dados contendo 3.352 pontos de dados exibindo os efeitos de 8 variáveis de entrada (volume de poros do sólido; área superficial do sólido; temperatura e pressão do experimento; técnica de medida da capacidade de adsorção, polarizabilidade, diâmetro cinético e massa molecular dos gases) sobre a capacidade de adsorção foi construído. Diagramas de caixa, histogramas, gráficos de barra e dispersão foram aplicados, como parte da análise exploratória de dados, para determinar como várias variáveis de entrada se relacionam entre si e com a variável de desempenho. Além disso, os modelos KNN, AD e RVS foram utilizados para regressão dos dados de capacidade adsorvida. O estudo paramétrico destes modelos permitiu determinar a importância relativa das variáveis de entrada e a dependência parcial entre elas visando explorar a interpretabilidade da modelagem (para deduzir heurísticas para alta ou baixa capacidade de adsorção). Constatou-se a partir da análise exploratória dos dados que a pressão, temperatura, polarizabilidade e massa molar do gás foram as variáveis mais significativas que afetaram a capacidade de adsorção. Além disso, combinações de variáveis de entrada que levam a um alto desempenho de adsorção foram reveladas por meio da análise dos modelos, as quais podem ser usadas como diretrizes para estudos futuros nesta área.

Palavras-chave: adsorção. gases leves. aprendizado de máquina.

ABSTRACT

This study aimed to extract knowledge about the adsorption of light gases by microporous materials (zeolites, MOFs, and activated carbons) through data analysis and machine learning algorithms: K-nearest neighbors (KNN), Decision Trees (DT), and Support Vector Regression (SVR) of data reported in 22 articles published between 1974 and 2022. A database containing 3352 data points displaying the effects of 8 input variables (solid pore volume; solid surface area; experimental temperature and pressure; adsorption capacity measurement technique; gas polarizability, kinetic diameter, and molecular mass) on adsorption capacity was constructed. Box plots, histograms, bar charts, and scatter plots were applied (as part of exploratory data analysis) to determine how various input variables relate to each other and the performance variable. Additionally, KNN, DT, and SVR models were used for the regression of the adsorbed capacity data. Furthermore, the parametric study of these models allowed determining the relative importance of input variables and partial dependence among them to explore model interpretability (to deduce heuristics for high or low adsorption capacity). The exploratory data analysis found that pressure, temperature, gas polarizability, and molecular mass were the most significant variables affecting adsorption capacity. Additionally, combinations of input variables leading to high adsorption performance were revealed through the model analysis, which can be used as guidelines for future studies in this area.

Keywords: adsorption. light gases. machine learning.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AC	Carvão ativado
AD	Árvore de Decisão
BET	Relativa à teoria clássica de Brunauer-Emmett-Teller
BJH	Relativa à teoria clássica de Barrett-Joyner-Halenda
CA	Carvão Ativado
CART	Árvores de classificação e regressão, do inglês *Classification and Regression Trees*
COF	Redes orgânicas covalentes, do inglês *Covalent Organic Frameworks*
CPU	Unidade de Processamento Central, do inglês *Central Processing Unit*
DI	Distribuição Interquartilar
GEP	Programação de Expressão Genética, do inglês *Gene Expression Programming*
GMDH	Método de Grupo para Manipulação de Dados, do inglês *Group Method of Data Handling*
IUPAC	União Internacional de Química Pura e Aplicada, do inglês *International Union of Pure and Applied Chemistry*
KNN	K-vizinhos mais próximos, do inglês *k-nearest neighbors*
MAE	Erro Médio Absoluto, do inglês *Mean Absoluto Error*
ML	Aprendizado de máquina, *do inglês Machine learning*
MLP	Perceptron multicamadas, do inglês *Multilayer Perceptron*
MOF	Redes metalorgânicas, do inglês *Metal Organic Frameworks*
MVS/SVM	Máquinas de Vertores de Suporte, do inglês *Support Vector Machine*
PDP	Gráfico de dependência parcial, do inglês *Parcial Dependence Plot*
PSA	Adsorção por Variação de Pressão, do inglês *Pressure Swing Adsorptio*n
RBF	*Radial Basis Function*
RF	Floresta aleatória, do inglês *Random Forest*
RMSE	Erro Quadrático Médio, do inglês *Root Mean Squared Error*
RNA	Redes Neurais Artificiais
RVS/RVS	Regressão por Vetores de Suporte, do inglês *Support Vector Regression*
SHAP	Relativo ao método de explicabilidade *Shapley Additive exPlanations*
TSA	Adsorção por Variação de temperatura, do inglês *Temperature Swing Adsorption*

UCS/SBU	Unidades de Construção Secundária, do inglês *Secondary Building Units*
VPSA	Adsorção por Variação de Pressão a Vácuo, do inglês *Vacuum Swing Adsorption*

LISTA DE SÍMBOLOS

C	Penalidade do erro no modelo MVS/RVS
Cov	Covariância
d	Grau da função de kernel polinomial
$d_{\text{Euclidiana}}(P, Q)$	Distância Euclidiana entre os pontos P e Q
$d_{Cosseno}(P, Q)$	Distância Cosseno entre os pontos P e Q
$d_{\text{Manhattan}}(P, Q)$	Distância de Manhattan entre os pontos P e Q
$d_{\text{Minkowski}}(P, Q)$	Distância de Minkowski entre os pontos P e Q
$h(x)$	Função contínua
K	Dados vizinhos mais próximos, referente ao modelo KNN (*K-Nearest Neighbors*)
m	Parâmetro da distância de Minkowski
M	Ponto M em um espaço bidimensional
n	Espaço de dimensão n
N	Ponto N em um espaço bidimensional
$N_k(\boldsymbol{x}')$	Vizinhança de $\boldsymbol{x}'$ formada pelos padrões de treinamento $\boldsymbol{x_i}$ que corresponde aos K vizinhos mais próximos a $\boldsymbol{x}'$
P	Pressão
P_o	Pressão de saturação
$P = (p_1, p_2, \dots, p_n)$	Ponto P em um espaço de dimensão n
p_i	Elemento i do ponto $P = (p_1, p_2, \dots, p_n)$
$Q = (q_1, q_2, \dots, q_n)$	Ponto Q em um espaço de dimensão n
q_i	Elemento i do ponto $Q = (q_1, q_2, \dots, q_n)$
$\boldsymbol{r}$	Termo de viés
R^2	Coeficiente de determinação
Tol	Tolerância para o critério de convergência
w_i	Peso associado ao i-ésimo vizinho de $\boldsymbol{x}'$
$\boldsymbol{x}'$	Novo dado de entrada
$\boldsymbol{x_1}$	Vetor de características (atributos do banco de dados)
$\boldsymbol{x_2}$	Vetor de características (atributos do banco de dados)
$\boldsymbol{x_i}$	Padrões de treinamento
y	Vetor de valores reais da variável alvo
$\hat{y}$	Vetor de valores preditos para a variável alvo

y_i	Valor real do registro i de y
$\hat{y}_i$	Valor predito do registro i de $\hat{y}$
$\bar{y}$	Média dos valores reais
$y(\boldsymbol{x_i} \in N_k(\boldsymbol{x'}))$	Valor alvo para o dado de entrada $\boldsymbol{x_i}$
$\hat{y}(\boldsymbol{x'})$	Valor predito para o novo dado de entrada $\boldsymbol{x'}$
z	Variável padronizada
α	Polarizabilidade
γ	Termo para controle da influência de cada amostra de treino individual na fronteira de decisão
ε	Desvio máximo dos dados de treinamento à função *h(x)*
θ	Ângulo compreendido entre dois vetores
μ	Média aritmética
ρ	Coeficiente de Correlação de Pearson
σ	Desvio padrão
σ_A	Desvio padrão da variável A
σ_B	Desvio padrão da variável B

SUMÁRIO

1. INTRODUÇÃO

As indústrias químicas e petroquímicas estão cada vez mais conscientes de que o seu desenvolvimento sustentável depende fortemente da aplicação de processos inovadores que utilizem materiais e energia de forma mais eficiente. Os processos de separação e purificação representam em torno de 40 a 60% do capital e dos custos operacionais industriais, respectivamente. A melhoria nesses processos pode reduzir expressivamente os custos, o uso de energia e a geração de resíduo (PULLUMBI; BRANDANI; BRANDANI, 2019).

Entre os processos de separação e purificação, encontra-se a adsorção que é uma operação de transferência de massa que ocorre a partir da propriedade que alguns sólidos possuem de aderir em sua superfície determinadas substâncias presentes em fluidos líquidos ou gasosos. Em relação a vários outros métodos de separação e purificação nas indústrias, a adsorção em colunas tem se destacado devido ao seu baixo custo, a ser um processo descomplicado e ecologicamente correto. Processos baseados em adsorção podem ser uma alternativa atrativa em relação aos processos atuais baseados em destilação criogênica que requerem um alto consumo de energia (BORDONHOS *et al.*, 2021; MAJD *et al.*, 2022).

Devido à atratividade oferecida pela possível redução no consumo de energia, o desenvolvimento de um método eficaz de separação de gases, como é o caso da adsorção, tem sido incentivado e estudado. Pesquisadores vêm investigando diferentes tipos de adsorventes como carvão ativado, zeólitas, titanossilicatos e materiais semelhantes à zeólitas para a separação de hidrocarbonetos e gases leves (GOLIPOUR *et al.*, 2020). E, a todo tempo, vários sólidos e materiais estão sendo descobertos. Logo, é preciso compreender como as características de um sólido e as condições de processo afetam o fenômeno de adsorção para identificar o melhor candidato para uma aplicação específica. Entretanto, a análise experimental das propriedades de transporte de novos materiais pode ser demorada e estudos precisos requerem equipamentos especializados, além de ser considerado economicamente inviável o teste experimental de muitos materiais. Desta forma, a triagem rápida de materiais via técnicas computacionais seria vantajosa para avaliar o potencial de adsorção de um gás puro, ajudando a concentrar o esforço experimental nos sistemas mais promissores (TANG *et al.*, 2021; WANG, Z. *et al.*, 2022; YUAN *et al.*, 2021).

As ferramentas da análise de dados em conjunto com os modelos de *Machine Learning* (ML), em português Aprendizado de Máquina, têm se mostrando como guias poderosos em diversos campos científicos e tecnológicos para avaliação e descrição de fenômenos e padrões.

Essas ferramentas e modelos vêm ganhando destaque em diversos estudos na área de adsorção para triagem rápida de materiais promissores para determinadas tarefas (CAO, 2022; LI *et al.*, 2022; ZHANG, XUAN; ZHENG; HE, 2022).

Diante do que foi apresentado, este estudo teve como objetivo realizar uma análise detalhada da complexa relação entre os parâmetros potenciais e os desempenhos de adsorção de zeólitas, estruturas metálicas orgânicas – conhecidas também como MOF, do inglês *Metal-Organic Frameworks* – e carvões ativados (CA), visando-se obter *insights* para o desenvolvimento de adsorventes de alta capacidade. Para isso, criou-se um banco de dados experimentais de adsorção de gases leves – CO_2, N_2, CH_4, C_2H_4 e C_2H_6 – envolvendo os materiais adsorventes mencionados. Esses dados foram avaliados fazendo uso de análise exploratória de dados e algoritmos de Aprendizagem de Máquina (ML, do inglês *Machine learning*) – k-vizinhos mais próximos (KNN), Regressão por Vetores de Suporte (RVS) e Árvores de Decisão (AD) – e análise de importâncias das variáveis de entrada.

1.1. Objetivos

1.1.1. Objetivo Geral

Desenvolver um banco de dados relativo a dados experimentais de equilíbrio de adsorção de gases leves em sólidos porosos e ajustar modelos de aprendizado de máquina clássicos para regressão das capacidades adsorvidas, visando contribuir para a redução de custos e tempo associados a análises laboratoriais e processos industriais.

1.1.2. Objetivos Específicos

i. Coleta de dados: construir e disponibilizar para a comunidade científica um banco de dados robusto relativo a dados experimentais de equilíbrio de adsorção de gases leves em diversos sólidos porosos.
ii. Modelagem preditiva: avaliar três modelos de aprendizado de máquina (KNN, AD e SVR) na previsão da capacidade de adsorção, considerando as características presentes no banco de dados construído.
iii. Desempenho dos modelos: avaliar a performance dos modelos por meio de métricas de avaliação para validar a eficácia das previsões em relação aos dados reais. Comparar a capacidade de predição dos algoritmos de aprendizado considerados no estudo.

iv. Interpretabilidade dos modelos: Utilizar técnica de análise de importância das variáveis por pemrutação, para compreender a contribuição relativa das diferentes características na capacidade de adsorção, identificando aquelas que mais influenciam na precisão do modelo. Além disso, aplicar o Gráfico de Dependência Parcial (PDP, do inglês *Partial Dependence Plot*) para visualizar e interpretar as relações individuais das variáveis de entrada com a saída do modelo, oderecendo *insights* sobre padrões e tendências.

1.2. Estrutura e Organização do Trabalho

A maneira pela qual este projeto foi conduzido ancorou-se em uma programação de atividades organizadas em forma cooperativa para alcançar os objetivos específicos e consequentemente o geral. O trabalho foi desenvolvido de acordo com as seguintes etapas:

i. Revisão bibliográfica do assunto proposto a fim de se obter um melhor entendimento do problema, fortalecer os conhecimentos sobre o problema abordado e seu entendimento físico; Inicialmente foi realizada uma revisão da literatura sobre adsorção e os diferentes tipos de adsorventes, englobando sólidos de diferentes classes tais como redes metalorgânicas (MOFs, do inglês *Metal Organic Frameworks*) , (COFs, do inglês *Covalent Organic Frameworks*), carvões ativados, sílicas e estruturas zeolíticas (aluminossilicatos, silicoaluminofosfatos e titanossilicatos), utilizados para o processo de separação das misturas gasosas em diferentes proporções e condições experimentais;

ii. Mediante extensiva revisão e consulta na literatura, foram coletados dados experimentais de adsorção (temperatura, pressão, concentração inicial dos gases), características do adsorvente (área superficial, composição mássica e diâmetro médio de poros) e capacidade de adsorção no equilíbrio para adsorvatos gasosos.

iii. Como a linguagem Python foi aplicada nesse trabalho para aplicação de *Machine Learning*, foi realizado um estudo dos fundamentos da linguagem Python, visando tornar-se ciente sobre a criação de variáveis e os diversos tipos de operadores (aritméticos, relacionais, de atribuição, lógicos, unários e ternários). Também foram estudadas as bibliotecas que foram utilizadas nesse trabalho, como, por exemplo, Scikit-learn.

iv. Na sequência, os algoritmos de KNN, AD e SVR foram utilizados para regressão das informações de capacidade de adsorção.

v. A interpretabilidade dos modelos foi estudada usando o recurso de análise de importância por permutação e o Gráfico de Dependência Parcial (PDP).

O presente trabalho está organizado em 6 capítulos, além das seções de Referências Bibliográficas e Apêndice A. O capítulo inicial e aqui descrito introduz a motivação pela qual este trabalho foi desenvolvido.

O segundo capítulo abrange os fundamentos teóricos essenciais para o desenvolvimento deste trabalho. São abordados temas relacionados ao processo de adsorção; princípios fundamentais de aprendizado de máquina e dos algoritmos utilizados para prever a capacidade de adsorção de gases leves (KNN, AD e SVR); conceitos fundamentais sobre interpretabilidade de modelos de aprendizado de máquina, incluindo análise de importância por permutação e Gráfico de Dependência Parcial (PDP); além de aspectos relevantes da linguagem de programação Python e suas bibliotecas.

No terceiro capítulo, aborda-se a metodologia, detalhando todas as fases do desenvolvimento do trabalho, desde a criação do banco de dados até a aplicação dos algoritmos escolhidos para prever a capacidade de adsorção de gases leves com base nas características contidas no banco de dados. Além disso, exploramos o estudo da interpretabilidade dos modelos utilizados.

O quarto capítulo relata os resultados e discussões, onde são apresentados os resultados alcançados, englobando inicialmente a análise básica e exploratória dos dados e finalizando com o estudo de interpretabilidade do modelo selecionado.

O quinto capítulo expõe a conclusão sobre os resultados alcançados no estudo e o capítulo 6 as oportunidades para estudos futuros. Em seguida, há as seções de Referências Bibliográficas que apresenta todo o referencial bibliográfico empregado neste trabalho e a seção Apêndice A que inclui informaçõe adicionais relevantes deste trabalho.

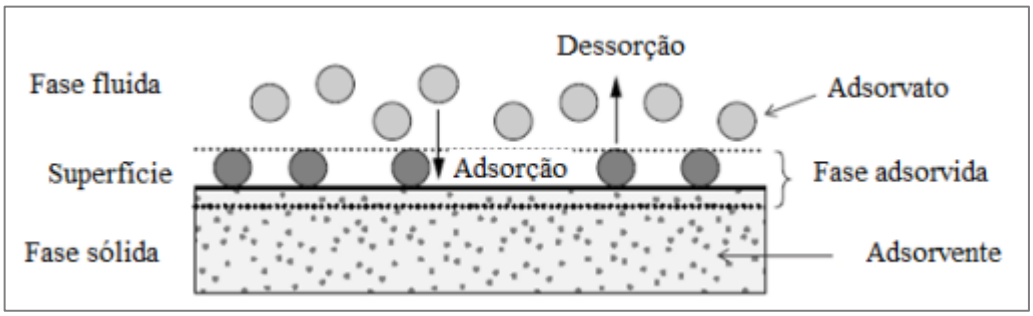
2. FUNDAMENTAÇÃO TEÓRICA

2.1. Adsorção

Ao longo de extensos períodos, a humanidade tem detido conhecimento acerca dos fenômenos de adsorção, os quais têm crescentemente sido empregados na consecução da separação ou purificação de distintas misturas. O cerne de um procedimento de adsorção, em regra, reside em um substrato sólido dotado de porosidade. A adoção de uma matriz sólida porosa decorre primordialmente da sua capacidade em fornecer uma vasta área superficial ou um considerável volume de microporos, sendo essa elevada superfície ou volume de microporos o alicerce subjacente à capacidade adsorvente almejada. Contudo, é notável que a estrutura porosa frequentemente se correlaciona com poros de dimensões diminutas, o que por sua vez instiga as moléculas adsorvatas a trilharem um trajeto rumo ao interior da área superficial ou volume de microporos. A emergência deste processo de "desbravamento do percurso" suscita o surgimento do que se conhece como resistência difusional ao fluxo molecular. A compreensão da capacidade do adsorvente insere-se no âmbito do equilíbrio, ao passo que a compreensão da resistência difusional está dentro no escopo da cinética (DO, DUONG D., 1998).

O processo de adsorção é classificado como um fenômeno de superfície em que ocorre a concentração de uma espécie química, conhecida como adsorvato, podendo essa ser um fluido líquido ou gasoso, sobre a superfície de um sólido, sendo esse o adsorvente (ALAQARBEH, 2021; CHIOU, 2002; NASCIMENTO *et al.*, 2014). Este processo normalmente é reversível, ou seja, as espécies adsorvidas podem ser liberadas da superfície e transferidas de volta para a fase fluida ao alterar as propriedades dessa fase (por exemplo, temperatura, concentração e pH), sendo esse fenômeno inverso denominado dessorção (ALAQARBEH, 2021; WORCH, 2021). A Figura 1 apresenta, de forma esquemática, esses dois processos, assim como seus termos básicos.

Figura 1 - Processo de adsorção, dessorção e seus termos básicos



Fonte: Worch (2021)

O processo de adsorção pode ser classificado em dois tipos a depender da natureza das forças envolvidas no processo:

Quimissorção: refere-se à interação química entre uma substância adsorvida e uma superfície sólida. Nesse processo, ocorre uma reação química entre os átomos ou moléculas adsorvidas e os sítios ativos presentes na superfície do material adsorvente. Essa interação química resulta na formação de ligações químicas fortes entre a substância adsorvida e a superfície sólida;

Fisissorção: refere-se à adsorção física, onde a interação entre a substância adsorvida e a superfície sólida é principalmente de natureza física, como forças de Van der Waals. Nesse processo, não ocorre uma reação química entre a substância adsorvida e a superfície, e a adsorção é geralmente mais fraca e reversível (CONDON, 2006; CYCHOSZ; THOMMES, 2018; VAREDA, 2023).

De acordo com Ruthven (2008), pode-se sumarizar as principais diferenças e entre estes tipos de acordo com a Tabela 1.

Tabela 1 - Comparação entre os processos de adsorção física e química.

Adsorção física	Adsorção química
Baixo calor de adsorção (1,0 a 1,5 vezes do calor latente de evaporação)	Alto calor de adsorção (maior que 1,5 vezes do calor latente de evaporação)
Não específica	Altamente específica
Monocamada ou multicamada	Monocamada apenas
Sem dissociação de espécies adsorvidas	Pode envolver dissociação
Significante apenas em temperaturas relativamente baixas	Possível em uma ampla faixa de temperaturas
Rápida, não ativada e reversível	Ativada, pode ser lenta e irreversível
Sem transferência de elétrons, embora a polarização do adsorvato possa ocorrer	Transferência de elétrons levando à formação de ligação entre o adsorvato e a superfície

Fonte: Ruthven (2008).

2.1.1. Fatores que afetam o processo adsortivo

Os fatores mais importantes a afetar o processo adsortivo são:

- A área superficial do adsorvente: A característica fundamental do adsorvente reside na extensão de sua superfície. Comumente, os materiais adsorventes exibem numerosos poros de configuração esférica ou irregular. Contudo, a dimensão reduzida das partículas do adsorvente amplifica a superfície exterior, propiciando um rápido engajamento da porosidade com a fase fluida. Como decorrência desse engajamento de forma mais eficiente, o processo ocorre em uma taxa mais elevada, culminando na acumulação aumentada de espécies adsorvidas (POURHAKKAK *et al.*, 2021);
- O tamanho da partícula do adsorvente: Conforme o tamanho dos materiais adsorventes diminui, as moléculas dos compostos a serem adsorvidos podem se deslocar com maior facilidade, transpondo as barreiras à transferência de massa. Em outras palavras, isso significa que o sistema tende a atingir o equilíbrio mais rapidamente, o que, por sua vez, permite a realização mais eficaz da capacidade total de adsorção. Contudo, é válido ressaltar que a utilização de adsorventes excessivamente pequenos não é viável em todos os contextos. Por exemplo, empregar esses materiais em colunas empacotadas resultaria em quedas

significativas de pressão, inviabilizando sua aplicação nesse tipo específico de processo (CHAHBANI; TONDEUR, 2001);

- O tempo de contato: Este item se refere ao período durante o qual a fase fluida (líquida ou gasosa) está em contato com a superfície do adsorvente. Este fator desempenha um papel crucial no processo de adsorção, influenciando a quantidade de substância adsorvida e a velocidade com que esse processo ocorre. Um tempo de contato mais longo geralmente resulta em uma maior quantidade de substância adsorvida na superfície do adsorvente. Isso ocorre porque as moléculas adsorvatas têm mais tempo para interagir e aderir à superfície, levando a uma acumulação maior. Além disso, à medida que o tempo de contato aumenta, o sistema tende a se aproximar do equilíbrio de adsorção (INGLEZAKIS; POULOPOULOS, 2006):
- O tamanho da molécula do adsorvato em relação ao tamanho dos poros: A forma geométrica da molécula também é um fator que afeta a taxa de adsorção, e consequentemente, moléculas com tamanhos grandes podem ser grandes demais para se difundir nos canais do adsorvente para atingir os sítios de adsorção, e isso diminuirá a capacidade de adsorção dessa molécula (DO, D. D. *et al.*, 2010).
- A polarizabilidade do adsorvato: A polarizabilidade refere-se à facilidade com que a distribuição de cargas em uma molécula pode ser distorcida por um campo elétrico externo, levando à formação de um momento dipolar induzido. A polarizabilidade da molécula gasosa afeta a adsorção de várias maneiras. Moléculas polarizáveis são mais propensas a participar de interações de forças de London (também conhecidas como forças de dispersão ou forças de Van der Waals). Essas interações ocorrem devido à flutuação momentânea das densidades de carga nas moléculas, criando dipolos instantâneos que induzem dipolos em outras moléculas próximas. Além disso, a polarizabilidade das moléculas gasosas permite uma atração mais forte com a superfície do adsorvente. Outro fator importante é que moléculas grandes e complexas frequentemente possuem maior polarizabilidade devido à presença de elétrons mais distantes do núcleo. Isso pode influenciar a adsorção, especialmente em sistemas onde as forças de dispersão são dominantes (GOLDEN; SIRCAR, 1994; KISELEV; LOPATKIN; SHULGA, 1985).

- A Temperatura: A adsorção geralmente é um processo exotérmico devido às interações atrativas entre as moléculas adsorvatas (aquelas que estão sendo adsorvidas) e a superfície do adsorvente (a substância sólida à qual as moléculas estão aderindo). Essas interações, no caso de gases, geralmente são de natureza van der Waals, dipolo-dipolo e ligações de hidrogênio. Além disso, o aumento no movimento molecular leva a um aumento da solubilidade na fase fluida e consequentemente a uma diminuição da adsorção (KOOPAL; TAN; AVENA, 2020).
- A pressão: Exceto em casos raros nos quais a estrutura microscópica de uma superfície sólida é quase uniforme, a maioria dos sólidos possui superfícies heterogêneas, o que resulta em variações nas energias de adsorção. Os sítios de adsorção são ocupados de forma sequencial, começando pelos sítios de maior energia e avançando em direção aos sítios de menor energia à medida que a pressão parcial ou concentração do soluto aumentam até atingir a saturação (CHIOU, 2002). A relação entre a quantidade de moléculas adsorvidas e a pressão em temperatura constante é chamada de isoterma de adsorção.

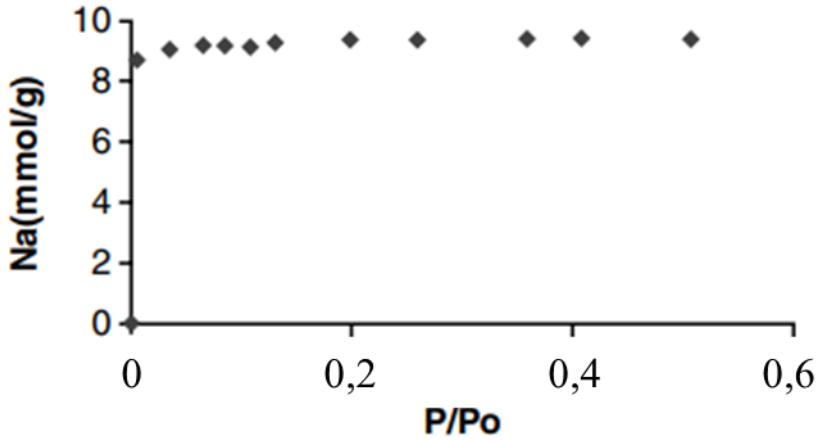
2.1.2. Equilíbrio de Adsorção

O equilíbrio de adsorção refere-se a um estado no qual as taxas de adsorção (ligação de moléculas de uma substância em fase gasosa ou líquida a uma superfície sólida) e dessorção (liberação das moléculas adsorvidas da superfície) atingem um ponto de estabilidade. Nesse estado, a quantidade de moléculas adsorvidas na superfície sólida permanece constante ao longo do tempo, uma vez que a taxa de adsorção é igual à taxa de dessorção. Em outras palavras, é um ponto no qual as interações entre as moléculas adsorventes e a superfície sólida alcançam um equilíbrio dinâmico (AL-GHOUTI; DA'ANA, 2020; MAJD *et al.*, 2022).

O equilíbrio de adsorção é governado por diversos fatores, incluindo a natureza das moléculas adsorventes e da superfície sólida, a temperatura, a pressão e a concentração das moléculas na fase gasosa ou líquida. Para entender esse equilíbrio, é comum utilizar a isoterma de adsorção, que é um gráfico que expressa a quantidade de adsorvato (substância adsorvida) em função da concentração na fase líquida ou da pressão na fase gasosa (MAJD *et al.*, 2022; WANG, J.; GUO, 2020). A Figura 2 ilustra o exemplo de uma isoterma de adsorção de nitrogênio a 77 K em zeólita sintética do tipo Na-Y, sendo P a pressão e P_o a pressão de saturação a 77K.

A isoterma de adsorção é essencial para a compreensão das relações entre a quantidade de adsorvato e as variáveis do sistema, permitindo determinar a capacidade máxima de adsorção, o coeficiente de distribuição e outros parâmetros relevantes. A partir desses dados, é possível otimizar processos de adsorção, como em operações de purificação, separação de componentes e remediação ambiental, levando em consideração as condições de equilíbrio que afetam a eficiência e a aplicabilidade desses processos (MAJD *et al.*, 2022).

Figura 2 - Isoterma de adsorção de nitrogênio a 77 K para diferentes pressões (P) em zeólita sintética do tipo Na-Y. Po indica a pressão de saturação a 77 K.



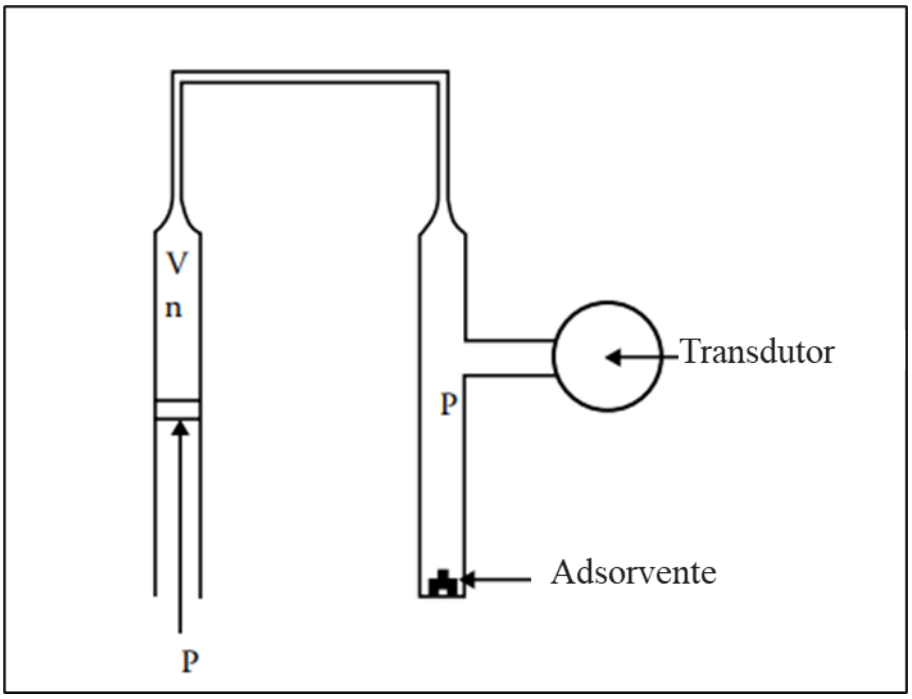
Fonte: Roque-Malherbe (2007).

2.1.3. Métodos experimentais para obtenção de dados de equilíbrio de adsorção de gases

Os métodos volumétrico e gravimétrico são dois métodos comumente usados para obter dados de equilíbrio de adsorção em estudos de adsorção de substâncias em superfícies sólidas.

No volumétrico, uma quantidade conhecida da substância alvo é colocada em contato com o material adsorvente em um sistema fechado. O equilíbrio de adsorção é alcançado à medida que a substância é adsorvida na superfície sólida. Para determinar a quantidade adsorvida, a concentração da substância no sistema é medida antes e depois do equilíbrio. Isso pode ser feito por meio de análise volumétrica, como titulação, espectrofotometria ou cromatografia. A diferença nas concentrações antes e depois do equilíbrio fornece uma medida da quantidade adsorvida (ROQUE-MALHERBE, 2007). A Figura 3 apresenta uma representação esquemática do experimento por aparato volumétrico.

Figura 3 - Representação esquemática do experimento de adsorção com aparato volumétrico.



Fonte: Adaptado de Roque-Malherbe (2007).

O início do desenvolvimento da técnica volumétrica para medição do equilíbrio de adsorção de gases data do início do século XX, por Sieverts, que considerava um aparato volumétrico de vidro para absorção e difusão de gases. Este aparato volumétrico inicial foi superado tecnologicamente e foi aprimorado com a medição de pressão, controle de temperatura e havendo o aumento da precisão do dispositivo para diferentes processos de adsorção de gás. As unidades volumétricas convencionais geralmente trabalham com quantidades em escala de grama de adsorventes, que considerando o tempo necessário para síntese, custo e capacidade de carregamento de novas amostras é uma das principais desvantagens dos aparelhos desenvolvidos. Geralmente, este tipo de unidade não tem muita flexibilidade em relação a estudos em diferentes sistemas de adsorção incluindo: 1) faixas de pressão e temperatura (por exemplo, para estudos criogênicos), 2) tipo de adsorventes (pó, grânulos, pastilhas, carbonos impressos em 3D, monólitos, etc.), 3) o tamanho e a proporção da célula de adsorção para a célula de referência. Ao mesmo tempo, a compra de unidades comerciais requer um orçamento alto em relação às unidades “artesanais” (KARIMI, M.; RODRIGUES; SILVA, 2021).

Já no método gravimétrico, a massa do material adsorvente é medida antes e depois do equilíbrio de adsorção. A massa inicial do adsorvente é registrada e, em seguida, o adsorvente é colocado em contato com a substância a ser adsorvida até atingir o equilíbrio. Em seguida, a

massa final do adsorvente é medida. A diferença entre as massas inicial e final fornece a quantidade adsorvida. Geralmente, uma balança de suspensão magnética é empregada para realização destas medidas (DREISBACH; SEIF; LÖSCH, 2003).

2.1.4. Sólidos Adsorventes

2.1.4.1. Zeólitas

A história das zeólitas teve início quando o mineralogista sueco Crönstedt descobriu a estilbita em 1756. Ao submeter a zeólita ao aquecimento, foi observada a liberação de água ocluída, que deu origem ao nome geral desses materiais, zeólita, derivado das palavras gregas "ξειν" (zeo), que significa ferver, e "λιθος" (lithos), que significa pedra. Esses materiais são minerais microporosos pertencentes ao grupo dos aluminossilicatos e são caracterizadas por uma estrutura cristalina tridimensional com poros e canais de dimensões moleculares, sendo composta por tetraedros de silício, alumínio e oxigênio, cuja composição química é representada pela Equação 1 (IBRAHIM, 2007; LASSINANTTI, 2001; YANG, R. T., 2003).

$$\mathrm{M}_{x/n}.\left[(\mathrm{AlO}_2)_x(\mathrm{SiO}_2)_y\right].\mathrm{zH}_2\mathrm{O} \qquad (1)$$

onde x e y são inteiros com y/x igual ou maior que 1, n é a valência do cátion M, e z é o número de moléculas de água em cada célula unitária.

De acordo com Kianfar (2020), as propriedades mais importantes desses materiais são estrutura bem definida, alta área superficial, sorção seletiva de pequenas moléculas (peneira molecular) e troca iônica que tornam as zeólitas extremamente versáteis e amplamente utilizadas em diversas áreas, englobando as áreas de catálise, adsorção, separação molecular e troca iônica. Dentre os diversos campos de suas aplicações industriais, encontra-se seu uso como catalisadores em processos industriais, como na produção de petróleo, química fina, indústria petroquímica e tratamento de gases de exaustão; no tratamento de água, controle de poluição ambiental, purificação de gases, remoção de compostos indesejáveis e separação de misturas complexas; em aplicações como amaciamento de água, remoção de metais pesados em tratamento de efluentes e purificação de líquidos (IBRAHIM, 2007; RHODES, 2010).

Ibrahim (2007) afirma que existem cerca de 40 zeólitas naturais que foram identificadas durante nos últimos 200 anos e mais de 150 zeólitas foram sintetizados. Dentre os tipos de zeólitas mais relevantes, muitas das quais empregadas em diversas aplicações comerciais, destacam-se as zeólitas naturais analcima, chabazita, clinoptilolita, erionita, mordenita e

phillipsita; e no grupo das sintéticas os tipos A, X, Y, L, "zeolon" mordenita, ZSM - 5, beta e MCM – 22, e as zeólitas F e W (IBRAHIM, 2007; KULPRATHIPANJA, 2010). De acordo com as zeólitas comerciais são encontradas em diversas formas, como: esferas, pellets e extrudados.

2.1.4.2. Carvões Ativados

O uso do Carvão Ativado (AC) remete ao Egito Antigo (1500 a.C), sendo usado pelos Egípcios para purificação da água como também para fins medicinais, devido às suas características adsorventes. Sua primeira aplicação no setor industrial ocorreu na Inglaterra em 1794, quando foi utilizado como agente descolorante na indústria açucareira. Por volta de 1900, ocorreu a primeira produção industrial desse ativo para a aplicação em indústria de refino de açúcar. Entretanto, foi durante a Primeira Guerra Mundial que o uso desse material cresceu de forma potencial e capitalizada com a finalidade de proteger as vidas humanas contra produtos químicos, sendo utilizado em máscaras de gás contra gases e vapores tóxicos (BANSAL, R. C.; GOYAL, 2005; MARSH; RODRÍGUEZ-REINOSO, 2006; TADDA *et al.*, 2016).

Os carvões ativados são sorventes amplamente utilizados, disponíveis em três formas principais, sendo estas pó, granulado e pellet, entretanto as mais utilizadas são essas duas primeiras (TADDA *et al.*, 2016). A Figura 4 apresenta uma pilha de AC granular. A produção comercial do AC, considerados materiais econômicos e ecológicos, é proveniente de diversos precursores como madeira, turfa, carvão, coque de petróleo, ossos, casca de coco e nozes de frutas (MARSH, 2001; YANG, 2003).

Figura 4 - Carvão ativado na forma granular

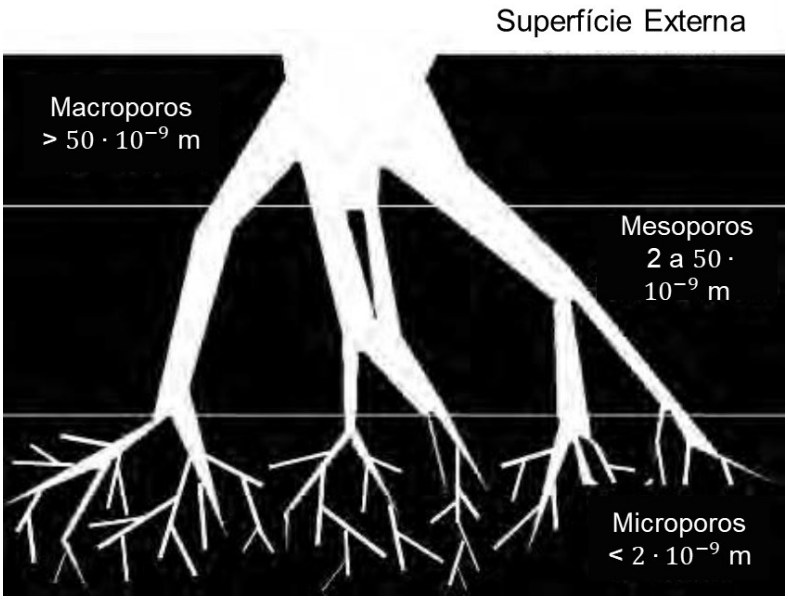
Fonte: Tadda et al. (2016).

Suas aplicações são diversas, sendo esse material utilizado para purificação de ar; remoção de odor, cor, sabores indesejáveis e outros compostos orgânicos e inorgânicos; remoção de impurezas de águas residuais domésticas e industriais; recuperação de solventes;

controle de escapamentos industriais e automotivos; na purificação de muitos produtos químicos, produtos farmacêuticos e alimentícios; e até mesmo em aplicações avançadas em plantas nucelares (BANSAL, R. C.; GOYAL, 2005; MUTTIL *et al.*, 2022; TADDA *et al.*, 2016).

De acordo com Bandosz (2016), os adsorventes de carvões ativados apresentam uma estrutura de carbono extremamente porosa com pequenas quantidades de heteroátomos – átomos diferentes dos átomos de carbono - como o oxigênio e o hidrogênio, e é formada por poros de diferentes tamanhos que são classificadas de acordo com a IUPAC (União Internacional de Química Pura e Aplicada, do inglês *International Union of Pure and Applied Chemistry*) em três grupos principais: microporos com poros menores que $2 \cdot 10^{-9}$ m; mesoporos com tamanhos de 2 a $50 \cdot 10^{-9}$ m e macroporos com tamanhos maiores que $50 \cdot 10^{-9}$ m, conforme esquemático apresentado na Figura 5. Destaca-se ainda que devido ao seu alto grau de porosidade, esse adsorvente também é caracterizado por uma alta área superficial que varia entre 300 e aproximadamente 4.000 m²/g (YANG, R. T., 2003).

Figura 5 - Representação esquemática da rede de poros de um adsorvente de carbono



Fonte: Adaptado de Bandosz (2016)

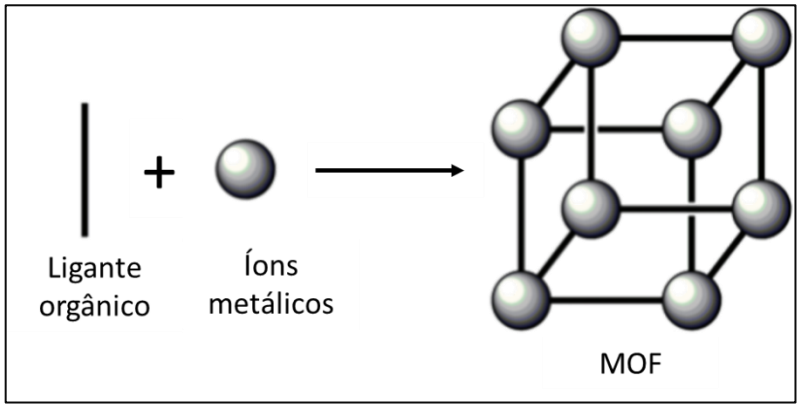
Conforme mencionado por Yang (2003), o carvão ativado se distingue da maioria dos outros sorventes devido à sua superfície apolar ou ligeiramente polar, resultado da presença dos grupos de óxido de superfície e impurezas inorgânicas. Essa propriedade dos carvões ativados proporciona algumas vantagens como:

- É o único produto sorvente disponível no mercado que permite realizar processos de separação e purificação sem exigir a remoção rigorosa de umidade prévia.
- Devido à sua grande superfície interna acessível e seu grande volume de poros, o carvão ativado adsorve mais moléculas orgânicas apolares e fracamente polares do que outros sorventes.
- O calor de adsorção, ou força de ligação, é geralmente menor no carvão ativado em comparação a outros sorventes. Isso ocorre porque apenas froças de van der Waals estão disponíveis como as principais forças de adsorção. Como resultado, a remoção das moléculas adsorvidas é relativamente mais fácil e resulta em requisitos de energia relativamente mais baixos para a regeneração de o sorvente.

2.1.4.3. Redes Metalorgânicas

As redes metalorgânicas (MOFs, do inglês *Metal-Organic Frameworks*) são materiais híbridos com uma estrutura cristalina formada por íons metálicos conectados por ligantes orgânicos (SAFAEI *et al.*, 2019; SHARMIN; ZAFAR, 2016; SONI; BAJPAI; ARORA, 2018), conforme esquemático apresentado na Figura 6. Por se tratar de compostos de coordenação, a ligação química metal-ligante envolvida na formação de MOFs é de natureza predominantemente covalente do tipo ácido/base de Lewis (BATTEN *et al.*, 2012).

Figura 6 - Esquemático de estrutura do MOF

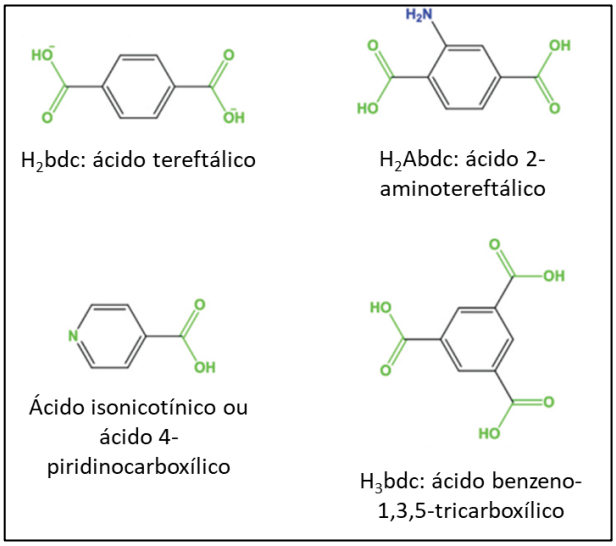


Fonte: Adaptado de Sharmin e Zafar (2016).

De acordo com Sharmin e Zafar (2016), as unidades orgânicas, também conhecidas como ligantes orgânicos, consistem em carboxilatos, como ilustrado na Figura 7, ou ânions, como fosfonato, sulfonato e compostos heterocíclicos, conforme exemplificado na Figura 8. As

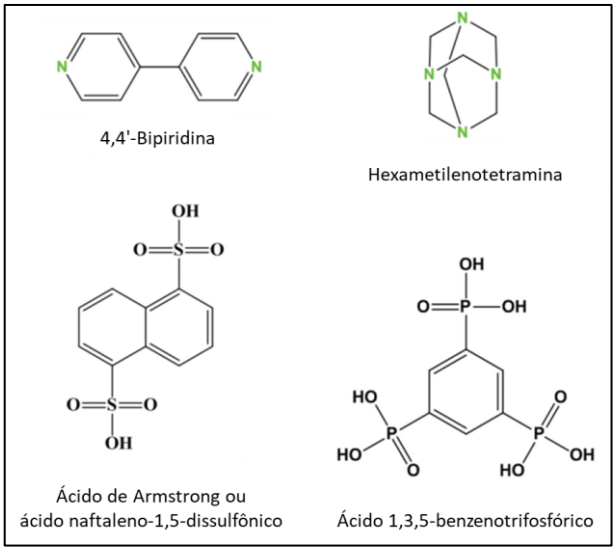
unidades inorgânicas são os íons metálicos, denominadas Unidades de Construção Secundária (UCSs ou SBUs, do inglês *Secondary Building Units*), e apresentam uma ampla variedade de geometrias com diferentes números de pontos de extensão, incluindo o octaedro (seis pontos), o prisma trigonal (seis pontos), a roda de pás quadrada (quatro pontos) e o triângulo (três pontos), conforme ilustrado na Figura 9.

Figura 7 - Alguns exemplos de ligantes orgânicos com funcionalidade carboxílica utilizados para a preparação de MOFs



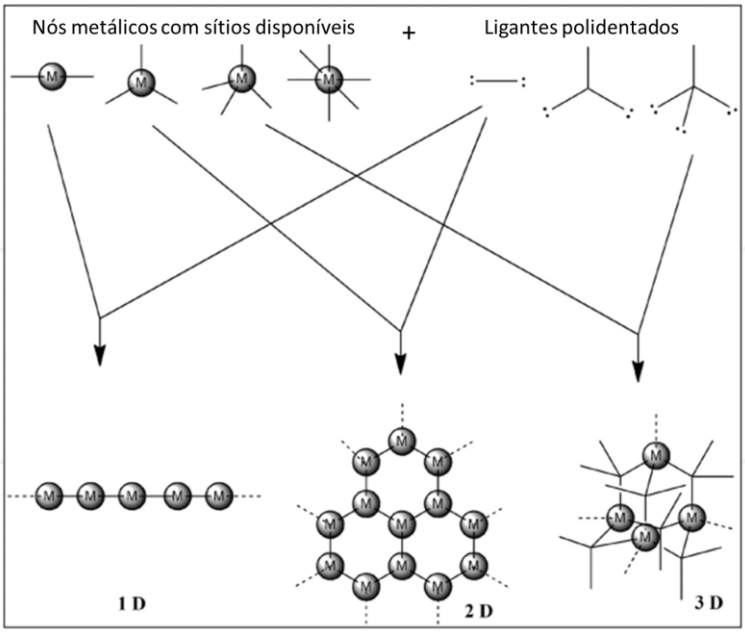
Fonte: Adaptado de Sharmin e Zafar (2016).

Figura 8 - Alguns exemplos de ligantes contendo nitrogênio, enxofre, fósforo e heterocíclicos utilizados para a preparação



Fonte: Adaptado de Sharmin e Zafar (2016).

Figura 9 - MOFs resultantes de diferentes nós metálicos e ligantes.



Fonte: Adaptado de Sharmin e Zafar (2016).

A formação de MOFs por meio da coordenação de unidades contendo metal (UCSs), resulta na criação de estruturas que exibem características notáveis. Estas incluem porosidade permanente, estabilidade estrutural, extensa área superficial e um amplo volume de poros, conforme destacado por Sharmin e Zafar (2016). Vale ressaltar que os valores de área superficial das MOFs geralmente se situam na faixa de 1.000 a 10.000 m^2/g, superando significativamente os valores associados a materiais porosos tradicionais, como zeólitas e carbonos, conforme mencionado por Furukawa (2013).

A diversificada gama de íons metálicos e ligantes orgânicos disponíveis possibilita a síntese de MOFs com uma variedade de propriedades, despertando considerável interesse nesses materiais. Isso resulta na preparação e estudo de milhares de compostos anualmente, como mencionado por Furukawa (2013). As características singulares dos MOFs os tornam altamente atrativos em diversas aplicações, tornando uma alternativa aos materiais porosos inorgânicos tradicionais, como zeólitas e sílica, para a remediação ambiental devido às suas características fascinantes. Estudos recentes têm demonstrado que os MOFs são um dos adsorventes ou catalisadores mais eficientes em aplicações de separação de gases, conversão de energia solar e aplicações fotocatalíticas.

A ampla variedade de íons metálicos e ligantes orgânicos disponíveis possibilita a síntese de MOFs com diversas propriedades, gerando um notável interesse nesses materiais. Isso conduz à preparação e ao estudo de milhares de compostos anualmente, conforme mencionado por Furukawa (2013). As características únicas dos MOFs os tornam altamente

atraentes em diversas aplicações, oferecendo uma alternativa aos materiais porosos inorgânicos tradicionais, como zeólitas e sílica, para a remediação ambiental devido às suas características fascinantes (WEN et al., 2019). Wen et al. (2019) expressam ainda que estudos recentes têm demonstrado que os MOFs são altamente eficientes como adsorventes ou catalisadores em aplicações de separação de gases, conversão de energia solar e processos fotocatalíticos.

2.2. Aprendizado de máquina

Sah (2020) define o termo *Machine Learning* (ML), também conhecido como aprendizagem de máquina, como um subcampo da inteligência artificial que faz o estudo de algoritmos de computador que fornece aos sistemas a capacidade de aprender e melhorar automaticamente com a experiência. Esses algoritmos são classificados como supervisionados, não supervisionados ou ainda como de reforço (BERRY; AZLINAH MOHAMED; YAP, 2020; SAH, 2020).

Os algoritmos supervisionados são aqueles em que o aprendizado envolve atributos de saída predeterminados, além dos de entrada, ou seja, o algoritmo aprende a função de mapeamento da entrada para a saída. Por sua vez, no não supervisionado, as instâncias não são rotuladas e o aprendizado ocorre pelo reconhecimento de padrões sem o envolvimento de um atributo alvo (BERRY; AZLINAH MOHAMED; YAP, 2020; KOTSIANTIS; ZAHARAKIS; PINTELAS, 2006; SAH, 2020).

Já o algoritmo de reforço visa aprender a controlar um sistema de forma a maximizar algum valor numérico que representa um objetivo de longo prazo. Nessa operação, um agente artificial obtém recompensas ou penalidades pelas ações que realiza e seu objetivo é maximizar a recompensa total (FRANÇOIS-LAVET *et al.*, 2018). O uso desse algoritmo inclui o aprendizado de como jogar jogos de computador, por exemplo.

No aprendizado de máquina, tanto a regressão quanto a classificação são tarefas importantes, mas diferem em seus objetivos e abordagens. A regressão é usada quando o objetivo é prever um valor contínuo ou quantitativo. Em outras palavras, a regressão busca estimar uma relação funcional entre as variáveis de entrada e a variável de saída, permitindo prever um valor numérico. Por exemplo, a regressão pode ser usada para prever o preço de uma casa com base em suas características, como tamanho, número de quartos, localização, etc. Por outro lado, a classificação é usada quando o objetivo é atribuir uma observação a uma categoria ou classe específica. É uma tarefa de aprendizado supervisionado em que o modelo é treinado com exemplos rotulados para aprender a classificar novas amostras em categorias pré-definidas. Por exemplo, a classificação pode ser usada para determinar se um e-mail é spam ou não spam

com base em várias características do e-mail (DIETTERICH, 1995; DING *et al.*, 2021; DOBBELAERE *et al.*, 2021; JORDAN; MITCHELL, 2015; KITCHIN, 2018; YANG, L.; SHAMI, 2020).

Neste trabalho foram aplicados três métodos de aprendizado de máquina supervisionados para regressão dos valores de capacidade adsorvida no equilíbrio: K-vizinhos mais próximos (KNN), Árvores de Decisão (AD) e Regressão por Vetores de Suporte (RVS).

2.2.1. K-vizinhos mais próximos

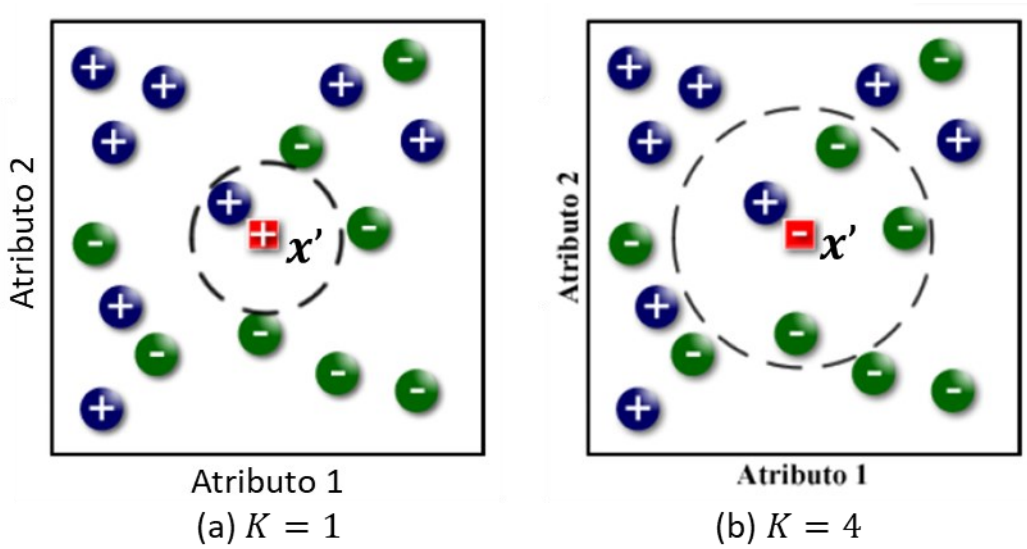
O modelo KNN (*K-Nearest Neighbors*) é um algoritmo de aprendizado de máquina que originalmente foi utilizado para tarefas de classificação, contudo nas últimas décadas também tem sido usado para regressão (AL-DOSARY; AL-HAMED; ABOUKARIMA, 2019). Ele se baseia em uma abordagem simples, mas eficaz, conhecida como "aprendizado baseado em instância", ou seja, o modelo toma decisões com base na similaridade entre novos exemplos de dados e instâncias previamente observadas no conjunto de treinamento. Em vez de construir um modelo explícito durante a fase de treinamento, o aprendizado baseado em instâncias armazena e memoriza os exemplos de treinamento e por esse motivo é conhecido como um método de aprendizado *lazy* ou preguiçoso (FERRERO, 2009; AL-DOSARY; AL-HAMED; ABOUKARIMA, 2019; TAUNK *et al.*, 2019).

Métodos de aprendizagem baseados em instâncias, como o KNN e regressão ponderada localmente, são abordagens conceitualmente simples para aproximar valores reais ou funções alvo de valor discreto. Uma diferença fundamental entre essas abordagens e os outros métodos é que as abordagens baseadas em instâncias podem construir uma aproximação diferente para a função de destino para cada instância de consulta distinta que deve ser classificada ou prevista. Na verdade, muitas técnicas constroem apenas uma aproximação local para a função alvo que se aplica na vizinhança da nova consulta por exemplo, e nunca constroem uma aproximação projetada para ter um bom desempenho ao longo de todo o espaço da instância. Isto tem vantagens significativas quando a função alvo é muita complexa (MITCHELL, 1997).

A ideia central do KNN é classificar ou prever um ponto de dados com base na maioria dos K pontos de dados vizinhos mais próximos, utilizando uma medida de similaridade também conhecida como métrica para cálculo de distância entre os pontos, a ele no espaço de características (TAUNK et al., 2019; AL-DOSARY; AL-HAMED; ABOUKARIMA, 2019). A Figura 10 ilustra exemplos de uso do algoritmo KNN, apresentado por Ferrero (2009), para um problema de classificação considerando Valor K igual a um (K=1) e quatro (K=4) vizinhos mais próximo, com um conjunto de exemplos de treinamento descrito por dois atributos, no

qual, exemplos com rótulo positivo (+) referem-se a pacientes doentes e exemplos com rótulo negativo (-) a não doentes. No primeiro caso, K=1, o exemplo $\boldsymbol{x'}$ é classificado como positivo, já no segundo caso, K=4, a maioria dos quatro exemplos mais próximos é negativo e E_i será classificado como negativo.

Figura 10 - Exemplos de aplicação do algoritmo KNN, com (a) $K = 1$ e (b) $K = 4$



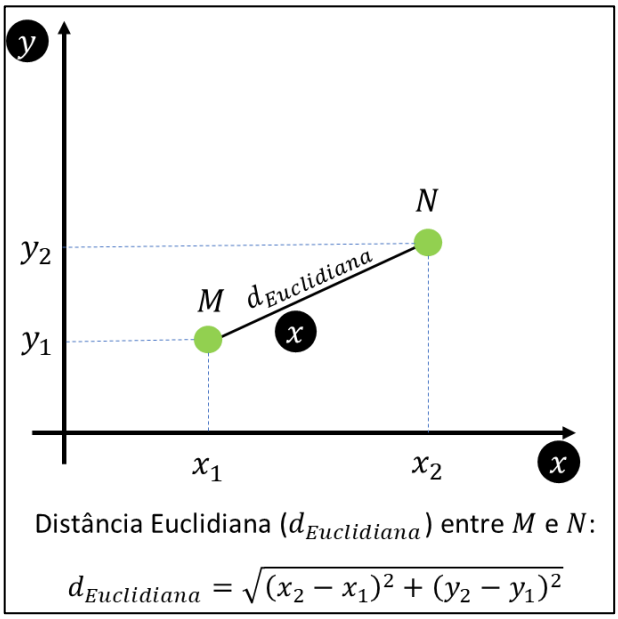
Fonte: Adaptado de Ferrero (2009)

Diante do exemplo apresentado, fica claro que o número de vizinhos mais próximos influencia fortemente a previsibilidade do modelo (FERRERO, 2009). Os principais parâmetros do modelo KNN são o número de vizinhos (K), a função de distância e a função de ponderação de vizinhos. Como mencionado anteriormente, K representa o número de vizinhos mais próximos a serem considerados ao fazer uma previsão. Escolher um valor adequado para K é crucial, pois valores muito baixos podem tornar o modelo sensível a ruído, causando sobreajuste denominado também como *overfitting*, enquanto valores muito altos podem suavizar demais a decisão, levando a um ajuste insuficiente, ou seja, *underfitting*. Já a escolha da função de distância afeta diretamente como os pontos são medidos em relação ao ponto de dados de interesse. Em alguns casos, pode-se ponderar os vizinhos com base em sua proximidade. Vizinhos mais próximos podem ter um peso maior na votação do que vizinhos mais distantes (ALI; NEAGU; TRUNDLE, 2019; BANSAL, M.; GOYAL; CHOUDHARY, 2022; TAUNK *et al.*, 2019).

A métrica mais comum para o cálculo de distância entre dois pontos é a Euclidiana (Equação 2), sendo sua representação para um sistema bidimensional apresentada na Figura 11. Contudo outras medidas para cálculo de distâncias são comuns: Manhattan, Minkowski,

Cosseno (ABU ALFEILAT *et al.*, 2019; EHSANI; DRABLØS, 2020; HECHENBICHLER; SCHLIEP, 2004). As Equações de 2 a 5 apresentam as métricas de distâncias mencionadas entre dois pontos em um espaço de dimensão n, sendo os pontos $P = (p_1, p_2, \dots, p_n)$ e $Q = (q_1, q_2, \dots, q_n)$.

Figura 11 - Cálculo da distância euclidiana entre dois pontos M e N para um sistema bidimensional.



Fonte: Bansal; Goyal e Choudhary (2022).

- Euclidiana: representa a raiz da soma do quadrado das diferenças entre os opostos valores em vetores (ABU ALFEILAT *et al.*, 2019; COST; SALZBERG, 1993; FAN *et al.*, 2019; HECHENBICHLER; SCHLIEP, 2004).

$$d_{\text{Euclidiana}}(P, Q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \tag{2}$$

- Manhattan: esta distância representa a soma das diferenças absolutas entre as coordenadas dos pontos (HECHENBICHLER; SCHLIEP, 2004).

$$d_{\text{Manhattan}}(P, Q) = \sum_{i=1}^{n} |q_i - p_i| \tag{3}$$

- Minkowski: essa diferença é uma generalização que abrange outras distâncias comuns, como a Euclidiana (Equação 2) quando $m = 2$ e a de Manhattan

(Equação 3) quando $m = 1$ (ABU ALFEILAT *et al.*, 2019; COST; SALZBERG, 1993; HECHENBICHLER; SCHLIEP, 2004).

$$d_{\text{Minkowski}}(P, Q) = \sqrt[m]{\sum_{i=1}^{n} |q_i - p_i|^m} \tag{4}$$

- Cosseno: também conhecida como distância angular. Essa distância é definida como a diferença entre 1 e a similaridade por cosseno que avalia o valor do cosseno do ângulo θ compreendido entre dois vetores (DEZA; DEZA, 2009; HECHENBICHLER; SCHLIEP, 2004).

$$d_{Cosseno}(P, Q) = 1 - \frac{\sum_{i=1}^{n} q_i p_i}{\sum_{i=1}^{n} q_i^2 \sum_{i=1}^{n} p_i^2} = 1 - \cos\theta \tag{5}$$

O uso do KNN para regressão é análogo ao seu emprego em casos de classificação. Contudo em vez de considerar as classes dos vizinhos mais próximos, o KNN para prevê um valor contínuo em uma abordagem simples computa a média dos valores alvo dos K vizinhos como a previsão para um novo dado de entrada $\boldsymbol{x'}$, conforme Equação 6 (MITCHELL, 1997). Onde $N_k(\boldsymbol{x'})$ denota a vizinhança de $\boldsymbol{x'}$ formada pelos padrões de treinamento $\boldsymbol{x_i}$ que corresponde aos K vizinhos mais próximos a $\boldsymbol{x'}$, $y(\boldsymbol{x_i} \in N_k(\boldsymbol{x'}))$ o valor alvo para o dado de entrada $\boldsymbol{x_i}$ e $\hat{y}(\boldsymbol{x'})$ o valor predito.

$$\hat{y}(\boldsymbol{x'}) = \frac{1}{K} \sum_{i=1}^{K} y(\boldsymbol{x_i} \in N_k(\boldsymbol{x'})) \tag{6}$$

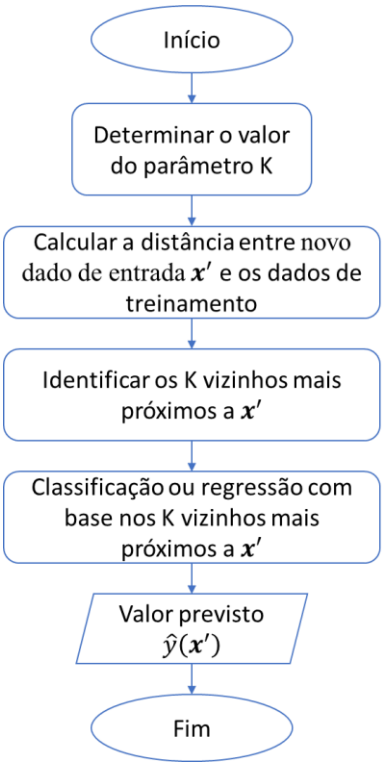
De acordo com Fan et al. (2019) e Mitchell (1997), outra abordagem é atribuir pesos diferentes à contribuição de cada vizinho à decisão final, utilizando uma média ponderada pelo inverso da distância dos K vizinhos mais próximos, ou seja, os vizinhos mais próximos possuem maior influência na previsão (Equação 7), sendo $w_i, i = 1, 2, \ldots, K$ o peso associado ao i-ésimo vizinho de $\boldsymbol{x'}$.

$$\hat{y}(\boldsymbol{x'}) = \frac{\sum_{i=1}^{K} w_i y(\boldsymbol{x_i} \in N_k(\boldsymbol{x'}))}{\sum_{i=1}^{K} w_i} \tag{7}$$

Mitchell (1997) expressa que ao adicionar ponderação à distância não há nenhum problema em permitir que todos os exemplos de treinamento influenciem na previsão de uma nova instância, pois exemplos muito distantes terão efeito mínimo. Contudo o mesmo autor afirma que a desvantagem principal reside na eficiência computacional, já que considerar todos os exemplos pode tornar a previsão mais lenta. Quando todos os exemplos de treinamento são considerados, o algoritmo é chamado de método global; se apenas os mais próximos são utilizados, é denominado método local.

Uma descrição geral do funcionamento do KNN é apresentada no fluxograma disponível na Figura 12. Cada etapa delineada no fluxograma é detalhada a seguir.

Figura 12 - Descrição geral do funcionamento do KNN



Fonte: Autoria própria (2023).

1. Determinar um valor para K: O primeiro passo é determinar o valor de K, que representa o número de vizinhos mais próximos a serem considerados ao fazer uma previsão. Um valor adequado para K é geralmente escolhido com base em tentativa e erro ou técnicas de validação cruzada.

2. Cálculo da distância: Em seguida, o KNN calcula a distância entre o ponto de dados que se deseja classificar ou prever e todos os outros pontos de dados no

conjunto de treinamento. Como mencionado, a distância mais comum é a distância euclidiana, mas outras medidas de distância também podem ser usadas, dependendo do problema.

3. Identificação dos K vizinhos mais próximos: O algoritmo seleciona os K pontos de dados com as menores distâncias em relação ao ponto de dados de interesse.

4. Previsão (classificação ou regressão): Para tarefas de classificação, o KNN realiza uma votação entre os K vizinhos mais próximos e atribui a classe mais frequente ao ponto de dados de interesse. Em outras palavras, a classe que ocorre com mais frequência entre os K vizinhos é a classe prevista. Para tarefas de regressão, o KNN calcula a média (ou outra medida, como a média ponderada pelos pesos dos vizinhos) dos valores alvo dos K vizinhos mais próximos e atribui esse valor ao ponto de dados de interesse (BANSAL, M.; GOYAL; CHOUDHARY, 2022).

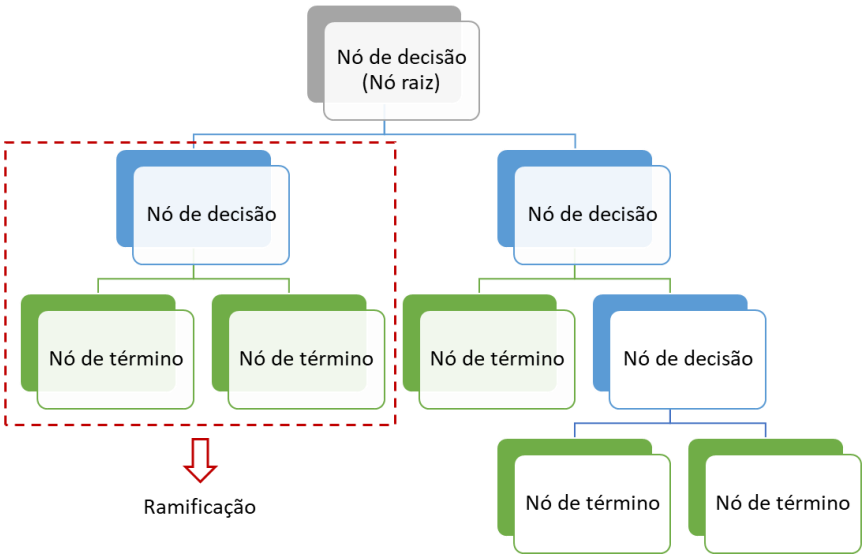
2.2.2. .Árvores de Decisão

Árvore de Decisão (AD) é o modelo de aprendizado de máquina mais conhecido e desenvolvido para tarefas de previsão e diagnósticos. A sua popularidade decorre da sua fácil interpretabilidade, eficiência e flexibilidade. Além disso, o algoritmo de Árvore de Decisão faz parte dos modelos de aprendizado supervisionado e pode ser usado para resolver problemas de classificação e regressão. Como um modelo de aprendizado supervisionado, a classe de cada objeto é conhecida no conjunto de dados. O objetivo é construir um modelo a partir de um conjunto de exemplos para encontrar uma descrição para cada uma das classes a partir de propriedades comuns entre os exemplos (KOTSIANTIS, 2013).

O aprendizado das regras em uma Árvore de Decisão consiste em separar os objetos em subamostras (as quais não possuem elementos em comum) onde a maioria dos objetos, idealmente, possuem o mesmo valor para a variável de saída, isto é, a mesma classe para o problema de classificação. A análise consiste em nós internos que representam as estruturas dos ramos (avaliação das características do conjunto de dados), com o objetivo de representar o veredito dado pelo algoritmo. Cada nó folha representa um resultado. Existem dois tipos de nós: o primeiro é o nó de decisão, usado para tomar uma decisão e que possui vários ramos; o segundo é o nó folha, que é a saída dos nós de decisão e não possui mais ramos. O nó raiz é um

ponto de partida que se expande em vários ramos, tornando-se uma estrutura semelhante a uma árvore, conforme mostra Figura 13 (BANSAL, M.; GOYAL; CHOUDHARY, 2022).

Figura 13 - Diagrama esquemático de uma árvore de decisão



Fonte: Adaptado de Bansal; Goyal e Choudhary (2022).

Ao construir uma Árvore de Decisão com muitos ramos o aprendizado pode ser muito específico para os dados do conjunto de treinamento, não sendo possível generalizar para o conjunto de teste (*overfitting*). Assim, para melhorar o modelo, são usados métodos de poda (*pruning*), com o objetivo de melhorar a taxa de acerto do modelo para novas amostras que não foram usadas no treinamento. Os métodos de poda podem ser separados em pré-poda e pós-poda. A pré-poda é realizada durante a construção da árvore e, caso o ganho de informação de um nó for menor que um valor pré-estabelecido, o nó vira uma folha. Enquanto isso, a pós-poda é realizada após a construção da árvore. Para cada nó interno da árvore é calculada a taxa de erro caso esse nó vire folha e caso não haja poda, se a diferença entre as duas taxas de erro for menor que um valor pré-estabelecido, a árvore é podada, caso contrário, não ocorre a poda (COSTA, V. G.; PEDREIRA, 2023).

Atualmente há diversos softwares disponíveis para construção de modelos de Árvore de Decisão. Entre os mais usados, encontra-se a biblioteca scikit-learn da linguagem Python. O modelo possui diversos parâmetros que precisam ser otimizados para obter melhores resultados de previsão. Alguns dos principais parâmetros a serem calibrados estão listados abaixo (a lista completa pode ser encontrada na documentação da biblioteca).

- Divisor (*Splitter*): valor *defaut* = *best*. Define a estratégia escolhida para dividir cada nó, pode receber os valores *best* (escolhe a melhor divisão) e *random* (escolhe a melhor divisão aleatória).

- Profundidade máxima da árvore (*max_depth*): possui valor *default* = *None*¸ caso em que os nós são expandidos até todas as folhas conterem menos objetos do que o especificado em *min_samples_split*. Um valor alto desse parâmetro causa o *overfitting* do modelo, enquanto um valor muito baixo causa o *underfitting*.

- Número mínimo para divisão (*min_samples_split*): default = 2. Número mínimo de amostras necessárias para dividir um nó interno.

- Número mínimo de amostras necessárias em uma folha (*min_samples_leaf*): possui valor *default* = 1. Uma divisão do nó em qualquer profundidade só será considerada se sobrar o número mínimo de amostras em cada folha, o que pode causar uma suavidade do modelo, especialmente no caso de regressão.

- Número máximo de nós folha na árvore (*max_leaf_nodes*): *default* = *None*. Cresce a árvore até o número máximo de nós folha definido. Se o valor for *none*, cresce até um número ilimitado de folhas.

- Número máximo de atributos considerados para cada divisão (*max_features*): *default* = *None*. Define o número de atributos a serem considerados nas divisões dos nós.

2.2.3. Regressão por Vetores de Suporte

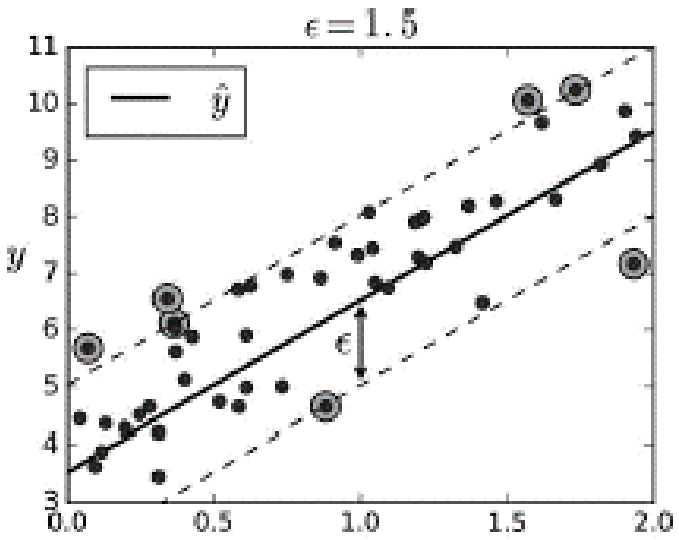
Os modelos baseados em vetores de suporte são amplamente discutidos e utilizados na literatura, com aplicação em diversas áreas. Esses modelos podem ser utilizados para solução de problemas de classificação, sendo, nesse caso, conhecidos como Máquinas de Vetores de Suporte (MVS ou SVM, do inglês *Support Vector Machine*) e, para regressão, onde são chamados por Regressão por Vetores de Suporte (RVS ou SVR, do inglês *Support Vector Regression*) (FACELI *et al.*, 2011).

Apesar do uso desses modelos ser mais difundido para tarefas de classificação, os resultados obtidos em tarefas de regressão são expressivos em diversas áreas, tais como: engenharia, medicina e química. O RVS pode ser entendido como um algoritmo que constrói

hiperplanos em um espaço n-dimensional para regredir dados. O RVS é um modelo muito poderoso e versátil e, por esta razão, é um dos modelos mais populares na área do conhecimento de aprendizado de máquina. O RVS é particularmente adequado para a regressão de conjuntos de dados complexos, porém de pequeno ou médio porte (Géron, 2019).

O algoritmo RVS objetiva encontrar uma função *h(x)* que produza saídas contínuas para os dados de treinamento que desviem no máximo de um ε (que representa o desvio máximo dos dados de treinamento à função *h(x)*) de seu rótulo desejado. Em sua formulação mais simples, a função *h(x)* possui forma linear em que os dados de treinamento ficam dentro de um limite ε da reta de regressão, conforme apresentado na Figura 14.

Figura 14 - SVR com função linear



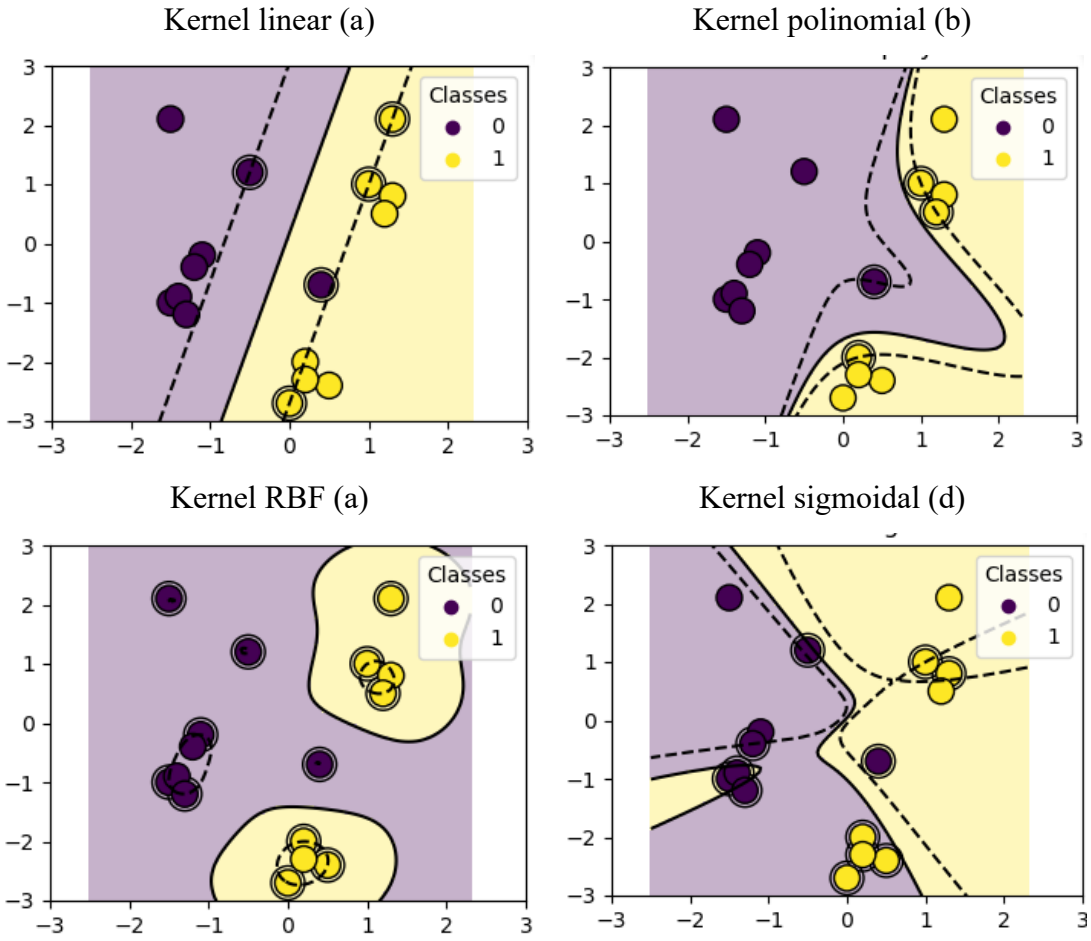
Fonte: Géron (2019)

Para os casos de regressões não lineares são utilizados *kernels* para mapear os objetos para um espaço de características, onde a função linear mais regular e com baixo erro de treinamento é encontrada (FACELI *et al.*, 2011). Diante disto, a função de *kernel* é um componente chave desse algoritmo para resolver problemas não lineares (CHIH-CHENG YANG; WAN-JUI LEE; SHIE-JUE LEE, 2006). As funções típicas de kernel incluem função polinomial, *Radial Basis Function* (RBF), também conhecida como gaussiana, e sigmoide (RAGHAVENDRA. N; DEKA, 2014). A Equação 8 mostra a formulação de cada kernel, onde $\boldsymbol{x_1}$ e $\boldsymbol{x_2}$ são vetores de características (pontos) nos quais o kernel está sendo aplicado, d é o grau da função polinomial, γ é um termo para controle da influência de cada amostra de treino individual na fronteira de decisão e r é o termo de viés ($coef0$). A Figura 15 mostra a separação de um conjunto de dados usando o SVM com os kernels linear, polinomial, RBF e sigmoidal.

Para esse trabalho serão analisados os kernels RBF e sigmoide, os quais apresentaram resultados satisfatórios em estudos anteriores de adsorção (KOOH et al., 2022).

$$K(\boldsymbol{x_1}, \boldsymbol{x_2}) = \begin{cases} (\boldsymbol{x_1^T}.\boldsymbol{x_2}) & Linear \\ (\boldsymbol{x_1^T}.\boldsymbol{x_2} + \boldsymbol{r})^d & Polinomial \\ e^{\left(-\gamma||x_1 - x_2||^2\right)} & RBF \\ \tanh(\gamma \boldsymbol{x_1}.\boldsymbol{x_2} + \boldsymbol{r}) & Sigmoide \end{cases} \tag{8}$$

Figura 15 - Separação de um conjunto de dados usando o SVM com os kernels linear (a), polinomial (b), RBF (c) e sigmoidal (d) para um caso de classificação



Fonte: Scikit-Learn (2007)

2.3. Interpretabilidade de modelos de aprendizado de máquina

A interpretabilidade de modelos de aprendizado de máquina refere-se à capacidade de compreender e explicar como um modelo toma decisões ou faz previsões com base nos dados de entrada. É a capacidade de entender o raciocínio interno do modelo e os principais fatores que influenciam suas decisões. A interpretabilidade é uma propriedade desejável em muitos

cenários, especialmente quando as decisões do modelo afetam diretamente os seres humanos (RIBEIRO; SINGH; GUESTRIN, 2016).

A interpretabilidade é valorizada uma vez que os usuários podem entender e confiar nas decisões tomadas pelo modelo. Além disso, modelos do tipo "caixa preta" podem gerar decisões viesadas. A interpretabilidade permite identificar e corrigir este possível problema. Além disso, a interpretabilidade permite fornecer explicações claras e compreensíveis sobre porque um modelo tomou uma determinada decisão. Por fim, analisar a interpretabilidade permite entender o funcionamento interno do modelo e assim os desenvolvedores podem identificar possíveis erros ou problemas de desempenho e ajustar o modelo de acordo (MOLNAR, 2020).

Existem várias abordagens para aumentar a interpretabilidade de modelos de aprendizado de máquina, como o uso da análise de importância das variáveis por permutação e do Gráfico de Dependência Parcial (PDP, do inglês *Partial Dependence Plot*) que foram empregados nesse trabalho e são detalhados a seguir.

2.3.1. Análise de importância das variáveis por permutação

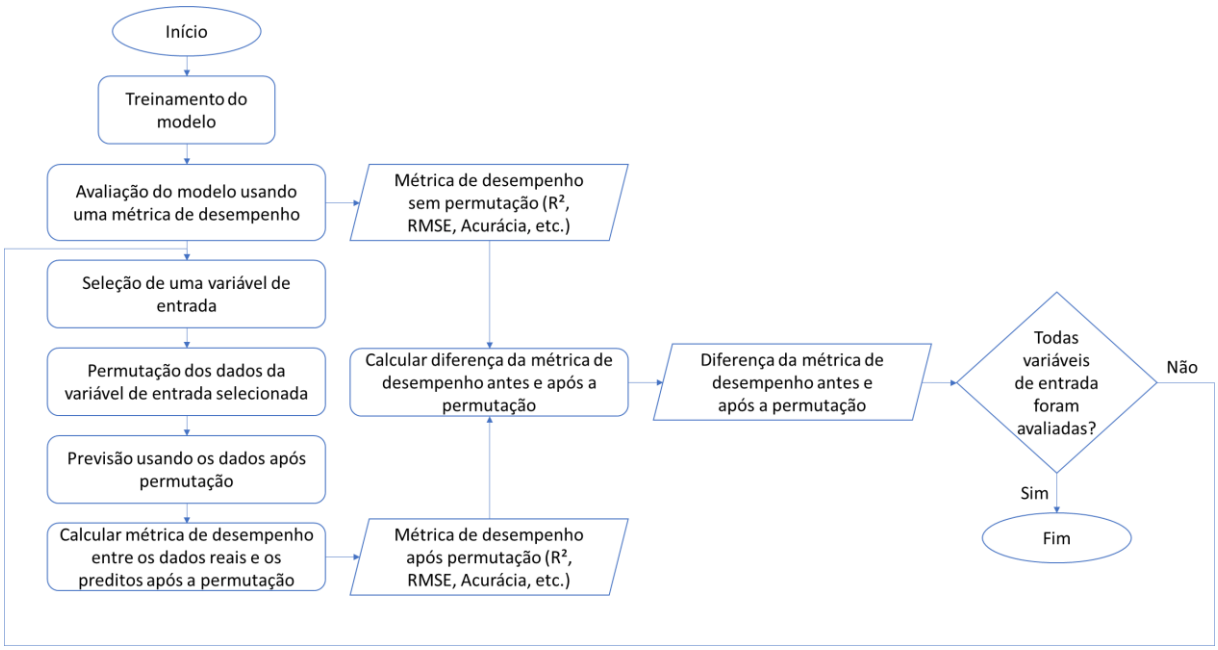
A importância das variáveis de um modelo pode ser avaliada pelo recurso de permutação que é um método de inspeção de modelo que pode ser usada quando os dados são tabulares. A importância de uma variável por permutação é definida como a diminuição na performance de um modelo quando os valores dessa variável são embaralhados aleatoriamente. Este procedimento quebra a relação entre a variável e o alvo. Portanto, a queda na pontuação do modelo, medida por algum indicador de desempenho do modelo (como por exemplo R^2, RMSE ou acurácia), é indicativa de quanto o modelo depende da variável (BREIMAN, 2001). Em outras palavras, um atributo é "importante" se ao embaralhar seus valores aumenta o erro do modelo, porque neste caso o modelo dependia do atributo para a previsão e um atributo é "sem importância" se embaralhar seus valores deixa o erro do modelo inalterado, porque neste caso o modelo ignorou o atributo para a previsão (MOLNAR, 2022).

O processo de análise de importância por permutação pode ser descrito pelas etapas a seguir e é representado pelo fluxograma apresentado na Figura 16.

1. Construção do modelo para previsão da variável alvo.
2. Avaliação da performance do modelo por alguma métrica de desempenho como R^2 (coeficiente de determinação), RMSE (Erro Quadrático Médio, do inglês *Root Mean Squared Error*), MAE (Erro Médio Absoluto, do inglês *Mean Absoluto Error*), acurária, entre outras.

3. Permutação dos valores em uma única variável de entrada e previsões usando o conjunto de dados resultante. A partir dessas previsões e os valores reais da variável alvo, avaliar a diferença entre a métrica de desempenho sem e após a permutação. Essa diferença, como mencionado, mede a importância da variável que foi embaralhada.
4. A partir dos dados originais, sem permutação, repetir a etapa 3 para a próxima variável de entrada até que todas sejam avaliadas.

Figura 16 - Processo da análise de importância por permutação



Fonte: Autoria própria (2023).

Para exemplificar o recurso de importância de um atributo por permutação, vamos considerar um caso em que foi treinado um modelo para prever a nota dos alunos com base nas seguintes variáveis de entrada: horas de estudo por dia, horas de sono por noite e horas em redes sociais por dia. Após treinar o modelo, é realizada a avaliação de desempenho deste usando uma métrica de desempenho, como por exemplo o R^2. Para avaliar a importância das variáveis por permutação é realizado o embaralhando aleatório dos valores de uma variável de entrada por vez e medida a diferença entre a métrica original (sem permutação) e a métrica após permutação, indicando a importância da variável. Quanto maior a diferença, maior será a importância desta na previsão do modelo.

Para o exemplo mencionado, suponha que o R^2 no conjunto original tenha sido de 0,956 e os resultados de R^2 obtidos após a permutação das variáveis horas de estudo por dia, horas de sono por noite e horas em redes sociais por dia foram respectivamente de 0,800, 0,888 e 0,940.

Desta forma, a diferença entre o R^2 sem permutação e os R^2 após a permutação das variáveis horas de estudo por dia, horas de sono por noite e horas em redes sociais por dia são, nesta ordem, de 0,156, 0,068 e 0,016. Neste caso, a análise de importância por permutação neste exemplo fictício sugere que a quantidade de horas de estudo por dia teve um impacto mais significativo no R^2 do modelo, indicando que esta variável desempenha um papel mais importante na previsão quando comparada as outras duas. A Figura 17 exemplifica o funcionamento de permutação para a análise de importância considerando a etapa de embaralhamento da variável horas de sono por noite.

Figura 17 - Permutação da coluna "Horas de sono por noite" para a avaliação de análise importância desta variável na predição da nota do aluno

Horas de estudo por dia	Horas de sono por noite	Horas em redes socias por dia	Nota do aluno
5	7	1	8,0
3	6	2	7,5
6	8	3	9,2
2	5	4	6,0
...	...	...	...
4	4	5	7,8

Permutação da variável "Horas de sono por noite"

Horas de estudo por dia	Horas de sono por noite	Horas em redes socias por dia	Nota do aluno
5	4	1	8,0
3	5	2	7,5
6	6	3	9,2
2	8	4	6,0
...	...	...	...
4	7	5	7,8

Fonte: Autoria própria (2023).

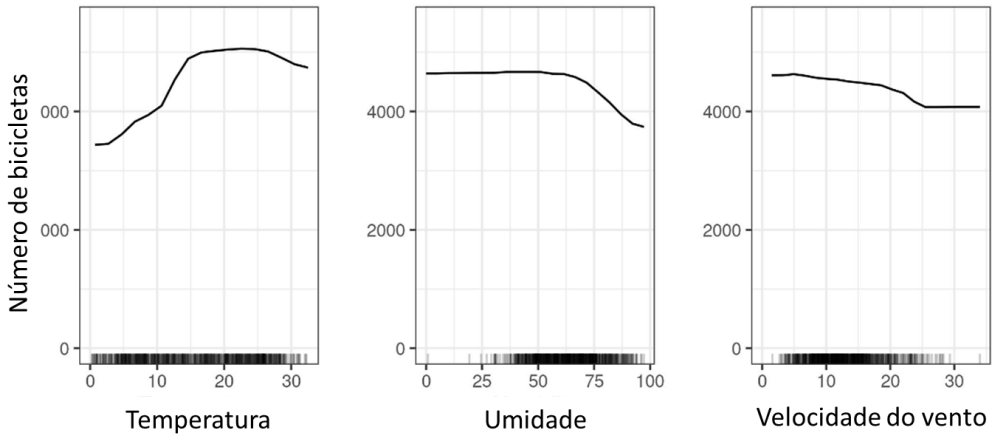
2.3.2. Gráfico de Dependência Parcial

Outra técnica de interpretabilidade é o gráfico de dependência parcial (PDP) foi aplicada. Os PDPs possuem uma função regressora que marginaliza o impacto de todos os atributos de entrada no modelo de previsão, exceto para um (ou dois, se necessário) de interesse, apresentando assim análise de sensibilidade local (JOHNSON *et al.*, 2022). Normalmente, é analisado o efeito conjunto de no máximo dois atributos no valor alvo, uma vez que a análise

de um atributo produz um gráfico 2D e dois atributos um gráfico 3D. Análises com 3 os mais atributos tornaria a interpretação dos resultados bem mais complexa.

A Figura 18 mostra um exemplo do uso do PDP para analisar a interpretabilidade de um modelo de Floresta Aleatória (*Random Forest*) para previsão do número de bicicletas que serão alugadas em um dado dia a partir da informação de temperatura, umidade e ventania nesse dia (MOLNAR, 2022). Os dados usados no modelo podem ser obtidos do repositório de aprendizado de máquina da UC Irvine. Para esse caso é possível perceber que para um clima quente, mas não muito quente, o modelo prevê, em média, um elevado número de bicicletas alugadas. Além disso, os potenciais ciclistas ficam cada vez mais inibidos em alugar uma bicicleta quando a umidade excede 60%. Por fim, quanto mais vento, menos pessoas gostam de pedalar. Como é possível perceber no exemplo apresentado, os resultados fornecidos pelo PDP mostram de forma simples os efeitos dos atributos no valor alvo (aluguel de bicicletas) para um modelo (*Random Forest*) de difícil interpretação.

Figura 18 - Exemplo do uso do PDP para analisar a interpretabilidade de um modelo de Floresta Aleatória para previsão do número de bicicletas que serão alugadas em um dado dia a partir da informação de temperatura, umidade e ventania nesse dia



Fonte: Adaptado de Molnar (2022)

2.4. Linguagem Python

Python é uma linguagem de programação criada pelo matemático e programador Guido van Rossum em 1991. Muitas linguagens são utilizadas para aprendizado de máquina como R, C/C++, Fortran, Go e Python. Entretanto, de acordo com Fenner (2020), a linguagem Python provou ser notavelmente popular para essa aplicação principalmente devido à disponibilização da biblioteca de ML Scikit-learn que torna fácil não apenas treinar uma série de modelos diferentes, mas também projetar recursos, avaliar a qualidade do modelo e pontuar novos dados.

Na versão final desta dissertação, serão apresentadas informações gerais sobre as principais bibliotecas empregadas na modelagem.

2.4.1. Pandas

O "pandas" é o pacote mais utilizado para trabalhar com conjuntos de dados grandes em Python. Ele é projetado para trabalhar com conjuntos de dados geralmente abaixo ou ao redor de 1 GB de tamanho, mas esse limite pode variar dependendo das restrições de memória do dispositivo em que está sendo executado. Uma boa regra geral é ter pelo menos cinco a dez vezes a quantidade de memória do dispositivo em relação ao tamanho do conjunto de dados. Quando o conjunto de dados começa a ultrapassar a faixa de alguns gigabytes, geralmente é recomendado usar uma biblioteca diferente. O nome "pandas" veio do termo "*panel data*" que se refere a dados tabulares. A ideia é que você pode criar painéis a partir de um painel maior de dados, como mostrado na Tabela 2 (STEPANEK, 2020).

Tabela 2. Exemplo de dados dispostos em forma de painel

Bairro de localização do restaurante	Média da avaliação dos clientes (*score*)	Mês da pesquisa
Centro	90	Junho
Santa Mônica	100	Julho
Segismundo Bairro	80	Agosto
Taiamã	70	Junho

Fonte: Autoria própria (2023).

Quando o pandas foi desenvolvido inicialmente, ele estava fortemente associado ao NumPy. A biblioteca NumPy é uma das mais importantes e populares bibliotecas em Python para computação numérica. Ela fornece suporte para arrays multidimensionais (matrizes) e funções matemáticas de alto desempenho que permitem realizar operações eficientes em grandes conjuntos de dados numéricos. Mesmo na versão moderna do pandas usada atualmente, ainda podemos notar essa estreita relação, como evidenciado pela exposição do tipo "Not a Number" (NaN) e por parâmetros da API, como o "dtype". Para desenvolver o pandas rapidamente e torná-lo mais acessível para sua base de usuários e colaboradores existentes, o DataFrame, estrutura de dados central do pandas, foi construído utilizando a funcionalidade do NumPy (BRESSERT, 2012; HARRISON, 2016; IDRIS, 2014).

Um DataFrame do pandas é uma estrutura de dados bidimensional, semelhante a uma tabela, que é amplamente utilizada para análise e manipulação de dados em Python. Ele é uma das principais estruturas fornecidas pela biblioteca pandas e permite organizar dados em linhas e colunas, tornando mais fácil e intuitiva a realização de operações de limpeza, manipulação, transformação e análise de dados. As principais características de um Dataframe são: estrutura tabula, rótulos de linha e coluna, flexibilidade, funcionalidades avançadas (para lidar com filtragem, seleção, agregação, ordenação, pivoteamento e junção de dados), integração com outras bibliotecas (como a matplotlib, scipy e scikit-learn) e leitura e escrita de dados (suportando diversos formatos para importação e exportação como CSV, Excel, SQL e JSON) (MORGAN, 2018).

O NumPy é construído com extensões em linguagem C, e apesar de fornecer uma interface em Python, a maior parte das operações ocorre quase inteiramente em C, o que torna sua execução extremamente eficiente. A linguagem C é consideravelmente mais rápida que o Python porque é uma linguagem de baixo nível, o que significa que não requer a mesma quantidade de memória e recursos da CPU (Unidade de Processamento Central, do inglês Central Processing Unit) que o Python necessita para fornecer facilidades de alto nível, como gerenciamento de memória. A diferença de desempenho entre Python e C geralmente não é muito significativa para a maioria dos desenvolvedores. Em muitos casos, o Python é suficientemente rápido, e as vantagens da linguagem de alto nível do Python (como o gerenciamento automático de memória e a sintaxe semelhante a pseudocódigo, para citar algumas) geralmente superam as complicações de gerenciar a memória manualmente (STEPANEK, 2020; SUMMERFIELD, 2010).

2.4.2. Scikit-Learn

O scikit-learn é uma biblioteca de aprendizado de máquina em Python que foi criada com o objetivo de fornecer uma ferramenta simples e eficiente para análise de dados e modelagem preditiva. Seu nome "scikit" vem da junção das palavras "SciPy" (outra biblioteca Python para computação científica) e "toolkit" (kit de ferramentas), refletindo sua associação com a comunidade científica e sua natureza de ser uma caixa de ferramentas para aprendizado de máquina (LEEKHA, 2022).

O scikit-learn foi iniciado como um projeto de código aberto em 2007, como parte do projeto Google Summer of Code. Desde então, muitos outros desenvolvedores contribuíram para o seu crescimento e aprimoramento. Originalmente, o scikit-learn foi desenvolvido como uma ferramenta simples para regressão e classificação, mas ao longo do tempo, evoluiu para

incluir diversos algoritmos e técnicas de aprendizado de máquina, abrangendo também agrupamento, redução de dimensionalidade, pré-processamento de dados e outras técnicas (GERON, 2019; PEDREGOSA *et al.*, 2011).

O scikit-learn se baseia em outras bibliotecas Python importantes, como NumPy, SciPy e matplotlib, aproveitando suas funcionalidades e integrando-se perfeitamente com o ecossistema científico de Python. Uma das principais metas do scikit-learn é ser fácil de usar, com uma interface consistente e uma documentação abrangente. Ele foi projetado para ser acessível tanto para iniciantes em aprendizado de máquina quanto para profissionais experientes, tornando-se uma escolha popular para a comunidade científica e de ciência de dados. A biblioteca ganhou reconhecimento e adoção generalizada por suas características como alta performance, eficiência computacional, suporte à validação cruzada e seleção automática de hiperparâmetros (DEO; SANJU, 2023; PEDREGOSA *et al.*, 2011; SANTOS-PEREIRA; GRUENWALD; BERNARDINO, 2022).

2.5. Estado da Arte: Algoritmos de aprendizagem de máquina aplicados ao estudo do equilíbrio de adsorção

As técnicas de Aprendizado de Máquina têm sido utilizadas para modelar e aprimorar processos de adsorção em diversos campos de aplicação como tratamento de águas residuais para remoção de fármacos (PRASAD *et al.*, 2023) , poluentes em geral (ZHANG, W. *et al.*, 2023) e coloração (AHMADI AZQHANDI *et al.*, 2017; KOOH *et al.*, 2022); qualidade do ar, incluindo a captura de dióxido de carbono (CO_2) que é um tema bastante atual (ASHAYERI *et al.*, 2021; SITU *et al.*, 2022; XIE, C. *et al.*, 2023); catálise heterogênea (ESTERHUIZEN; GOLDSMITH; LINIC, 2022; GHANEKAR; DESHPANDE; GREELEY, 2022; JÄGER *et al.*, 2018; SCHLEXER LAMOUREUX *et al.*, 2019) e demanda energética (MENG; ZHONG; WEI, 2020; AMAR et al., 2022).

Apesar da ampla aplicação da técnica de *Machine Learning* voltada ao estudo de adsorção, é essencial destacar que, até o momento, não há registros de estudos abordando a temática tratada nesta dissertação com bancos de dados coletados na literatura visando estimar a capacidade de adsorção de diversos gases puros em uma ampla variedade de adsorventes. Estudos recentes tem se concentrado na avaliação da adsorção de gases individuais em um adsorvente específico por modelo (AMAR et al., 2022; KOOH et al., 2022; MENG; ZHONG; WEI, 2020; WANG, S. et al., 2019; XIE, C. et al., 2023).

A seletividade do CO_2/N_2 em carvões ativados é investigada por Wang et al. (2019), utilizando técnicas de *Deep Learning*, também conhecido como aprendizado profundo. O

trabalho se baseia em um conjunto de dados experimentais, composto por 1.138 dados de adsorção de CO_2 e 314 dados para a adsorção de N_2 em diversas amostras de carbonos porosos, abrangendo diferentes estruturas e características. Para realizar a análise, os pesquisadores empregaram a arquitetura de Rede Neural Profunda (*Deep Neural Networks*) com duas camadas ocultas, sendo um modelo para o conjunto de dados de adsorção de CO_2 e outro para o de N_2, com as seguintes variáveis de entrada para prever a seletividade do CO_2/N_2 em carbonos porosos: área superficial, volume de micro e mesoporos, temperatura e pressão.

Os resultados obtidos por Wang et al. (2019) a partir do modelo de aprendizado profundo revelaram *insights* importantes sobre como a estrutura dos carvões ativados influencia a seletividade de CO_2 em relação ao N_2, demonstrando que as melhores características para maior captura de CO_2 não eram necessariamente as mais adequadas para a seletividade do CO_2/N_2. Em vez disso, focar nas regiões de menor captura de N_2 com volumes moderados de microporos e mesoporos se mostrou uma estratégia mais eficaz para alcançar a maior seletividade de CO_2/N_2. Essa abordagem proporcionou uma compreensão mais abrangente do processo seletivo em carvões ativados e oferece novas perspectivas para o desenvolvimento de materiais com maior seletividade para a captura de CO_2 em aplicações de sequestro e mitigação de gases de efeito estufa.

A adsorção do metano em xisto foi estudada por Meng, Zhong e Wei (2020). Os autores empregam modelos clássicos e modelos baseados em aprendizado de máquina para prever a adsorção do metano *in-situ*. Uma base de dados abrangente com 352 pontos de dados foi coletada da literatura, e as variáveis de pressão, temperatura, umidade e Carbono Orgânico Total (COT) foram usadas como entradas para as correlações implementadas. Os modelos baseados em aprendizado de máquina empregados foram quatros: RNA (Redes Neurais Artificiais), RF (Floresta aleatória, do inglês *Random Forest*), RVS e *Extreme Gradient Boosting* (XGBoost), sendo este último o que mostrou ser o mais eficiente para prever a adsorção de metano em formações de gás de xisto, fornecendo os melhores resultados comparados com os demais modelos clássicos e baseados em aprendizado de máquina. Os autores destacam ainda que o XGBoost não se limita às condições isotérmicas e a um único tipo de xisto, além de ser extrapolável além do intervalo de teste. Essa capacidade de generalização do XGBoost apresenta uma vantagem significativa em relação aos modelos clássicos, permitindo uma maior flexibilidade e aplicabilidade a uma variedade maior de cenários e formações geológicas.

Amar et al. (2022) abordaram a modelagem da capacidade de adsorção de metano (CH_4) em formações de gás de xisto usando técnicas de aprendizado de máquina supervisionadas do

tipo "white-box", que são modelos com interpretabilidade. Neste estudo duas técnicas rigorosas baseadas em dados - Programação de Expressão Genética (GEP, do inglês *Gene Expression Programming*) e o Método de Grupo para Manipulação de Dados (GMDH, do inglês *Group Method of Data Handling*), foram utilizadas para fornecem expressões matemáticas explícitas precisas e confiáveis para prever a adsorção de metano. Para realizar as previsões, os autores utilizaram a mesma base de dados experimentais e variáveis de entrada empregada no trabalho de Meng, Zhong e Wei (2020).

Os resultados do trabalho de Amar et al. (2022) mostraram que ambas as correlações – GEP e GMDH - podem fornecer previsões precisas para a capacidade de adsorção de CH_4 em formações de xisto. No entanto, a correlação baseada em GEP apresentou previsões mais confiáveis para a adsorção de metano, superando a correlação baseada em GMDH com um Coeficiente de Correlação (R^2) de 0,9837 e Erro Quadrático Médio (RMSE) de 0,0625. Além disso, foi demonstrado que a correlação baseada em GEP pode prever com precisão a variação da capacidade de gás de xisto para diferentes valores de entrada. Constatou-se também que a adsorção de metano depende significativamente do valor de umidade, enquanto a temperatura, o COT e a pressão são as variáveis mais influentes após a umidade. Essa abordagem é relevante para a indústria de petróleo e gás, pois permite melhorar a compreensão e previsão da capacidade de armazenamento e produção de gás em reservatórios de xisto, oferecendo, por meio da utilização de técnicas de aprendizado de máquina interpretáveis, percepções adicionais sobre os fatores que afetam a adsorção de metano, facilitando tomadas de decisões para a exploração e produção de gás natural.

Xie et al. (2023) exploraram o uso de técnicas de Aprendizado de Máquina - Florestas Aleatórias (RF, do inglês, *Random Forest*), Regressão Logística, Regressão por Vetores de Suporte (RVS) e Redes Neurais Artificiais (RNA) de múltiplas camadas, também conhecidas como MLP (do inglês, *Multilayer Perceptron*) - para compreender e prever a capacidade de adsorção de CO_2 em carbonos porosos, também conhecidos como carvões ativados. O banco de dados do estudo é compreendido por 4.589 dados, referentes a 239 carbonos porosos, que foram extraídos de 31 trabalhos. As variáveis de entrada escolhidas incluem as características físicas e estruturais do carbono poroso - área superficial, volume de micro e mesoporos -, sua composição, em gramas, de Nitrogênio (N), Carbono (C), Hidrogênio (H) e Oxigênio (O), temperatura e pressão. E, a quantidade de dióxido de carbono adsorvido pelo material poroso foi escolhida como variável de saída.

Dentre os modelos avaliados no trabalho de XIE et al. (2023), o RF foi o que apresentou uma melhor acurácia com o maior valor de coeficiente de determinação (R^2 de 91%) e menor

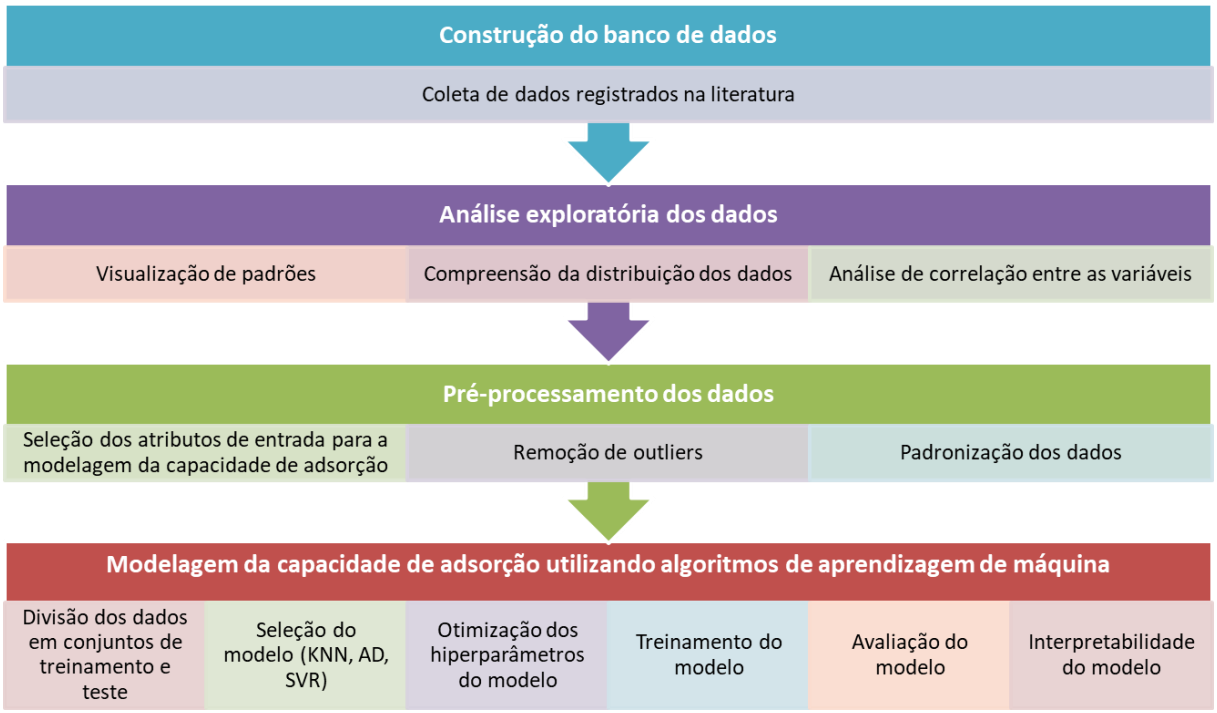
RMSE, de 0,31 para o conjunto de teste. Os autores empregaram ainda ao modelo que apresentou melhor precisão, RF, o método *Shapley Additive exPlanations* (SHAP) de explicabilidade no campo de aprendizado de máquina para quantificar e desvendar a influência de cada propriedade na capacidade de adsorção de CO_2 em carvões ativados. Além da influência de cada propriedade físico-química do carbono poroso, os resultados mostraram a faixa de valores desejados para alcançar alta capacidade de adsorção de CO_2 em baixas pressões e temperatura de 298 K, especificamente: área superficial entre 1.100 e 1.300 m^2/g, volume de microporos entre 0,4 e 0,6 cm^3/g, volume de mesoporos inferior a 0,1 cm^3/g, composição em peso de C entre 72 e 86%, H entre 1 e 4%, O inferior a 16% e N menor que 1%. Desta forma, os autores apresentam resultados promissores na aplicação de modelos de aprendizado de máquina interpretáveis para melhorar a compreensão desse processo crítico na captura de dióxido de carbono, contribuindo assim para o avanço de soluções sustentáveis de mitigação de gases de efeito estufa.

Os trabalhos apresentados destacam a importância significativa das técnicas de aprendizado de máquina no campo da adsorção devido à sua capacidade de analisar e compreender dados complexos relacionados aos processos de adsorção. Com a crescente necessidade de desenvolver materiais adsorventes altamente seletivos e eficientes para aplicações como captura de gases, separação de misturas, purificação de fluidos, tratamento de efluentes, entre outras, o aprendizado de máquina oferece uma abordagem poderosa para otimizar as propriedades dos materiais (TANG *et al.*, 2021; WANG, Z. *et al.*, 2022). Essas técnicas permitem o processamento rápido e preciso de grandes conjuntos de dados experimentais, identificando padrões sutis e relacionamentos ocultos que podem levar a avanços significativos no projeto e caracterização de materiais adsorventes (GHANEKAR; DESHPANDE; GREELEY, 2022; YUAN *et al.*, 2021). Além disso, o uso de aprendizado de máquina na adsorção possibilita a previsão da capacidade de adsorção, permitindo uma seleção mais eficiente de materiais para atender a requisitos específicos de aplicação (ZHANG *et al.*, 2022). Combinando a expertise científica com as vantagens das técnicas de *Machine Learning*, os avanços na adsorção têm o potencial de contribuir significativamente para a sustentabilidade ambiental e a eficiência de processos industriais.

3. METODOLOGIA

O presente trabalho foi constituído das etapas apresentadas na Figura 19. Cada etapa é detalhada a seguir.

Figura 19 - Metodologia aplicada



Fonte: Autoria própria.

3.1. Construção de Banco de Dados

A base de dados foi construída extraindo os dados de adsorção sobre diversos materiais publicados entre os anos 1974 e 2022. Primeiramente, mais de 50 artigos foram estudados e as variáveis potenciais que poderiam ser críticas no desempenho dos adsorventes foram coletadas. Na segunda etapa, todas as publicações foram reavaliadas a fim de remover aquelas que não continham informações sobre as propriedades selecionadas (por exemplo, propriedades estruturais, etc.) juntamente com os dados de adsorção.

Ao fazer isso, foram eleitas apenas zeólitas, MOFs e carvões ativados como as estruturas mais usualmente investigadas. Devido à natureza das metodologias de mineração de dados, não deve haver nenhum valor exclusivo ou alguns valores para uma variável de entrada no banco de dados usado para modelagem. Isso acontece porque algumas técnicas utilizadas na modelagem são indutivas. Um ou poucos exemplos impossibilitam a previsão ou generalização

dos padrões. Portanto, materiais raramente utilizados não puderam ser admitidos no banco de dados, embora pudessem apresentar resultados animadores. Pelo mesmo motivo, a análise de dados ao longo deste trabalho é estabelecida nas faixas de variáveis mais frequentemente acessíveis na literatura.

Por fim, a base de dados criada foi composta por 3.352 dados extraídos de 22 publicações para 39 diferentes materiais, as quais estão listadas na Tabela 3. As unidades adotadas são as usualmente empregadas pelos experimentalistas.

O banco de dados foi construído com base em 9 variáveis de entrada agrupadas em quatro: variáveis de operação, técnica de medida da capacidade de adsorção, propriedades adsorventes e propriedades do gás. A lista das variáveis de entrada juntamente com suas faixas é apresentada na Tabela 4. Como a maioria das avaliações de desempenho de adsorção de gases puros são baseadas na capacidade de adsorção (em mmol/g adsorvente), esta variável foi tomada como variável alvo na modelagem. Este alvo variou de 0 a 17,04 mmol/g.

É importante salientar que, a princípio, buscou-se incluir no banco de dados informações sobre densidade aparente do sólido, densidade real, porosidade e diâmetro médio de poros. Contudo, não foi possível coletar 30% de exemplos dotados destas características, razão pela qual tais variáveis foram abandonadas.

Tabela 3 - Detalhes sobre os artigos utilizados para a criação do banco de dados.

Número	Referência	Gases	Sólido
1	Nakahara (1986)	CH_4, C_2H_4, C_2H_6, C_3H_6	Carvão ativado MSC 5A
2	Reich (1974)	C_2H_4, C_2H_6, CH_4 e CO_2	Carvão ativado BPL
3	Danner; Choi (1979)	C_2H_4 e C_2H_6	Zeólita 13X
4	Costa et al. (1989)	C_2H_4, C_2H_6 e CH_4	Carvão ativado AC-40
5	Choi et al. (2003)	CH_4, C_2H_4, C_2H_6, N_2	Carvão ativado PCB Calgon Co.
6	Al-Muhtaseb; Al-Rub; Zarooni (2007)	CH_4, C_2H_4, C_2H_6, N_2	Carvão ativado BDH
7	Al-Muhtaseb (2010)	CH_4, C_2H_4, C_2H_6, N_2	Carvão atibado Date-Pit
8	(Mishra; Mekala et al. (2012)	N_2, CH_4 e CO_2	MOF ZnDABC
9	Mishara; Uppara et al. (2012)	C_2H_6 e C_3H_8	MOF ZnDABC
10	Mofarahi; Salehi, (2013)	C_2H_4 e C_2H_6	Zeólita 5A
11	Mcewen; Hayman; Ozgur Yazaydin (2013)	N_2, CH_4 e CO_2	MOF ZIF-8 e Zeólita 13 X
12	Pham et al. (2014)	N_2 e CO_2	Zeólitas Beta, CHA, FER, MFI e STT
13	Zhang et al. (2015)	N_2, CH_4 e CO_2	MOF MIL-101
14	Martins et al. (2015)	C_2H_4 e C_2H_6	MOF Cu-BTC
15	Birkmann et al. (2017)	C_2H_6, C_3H_8	Zeólita 13X e Carvão ativado Norit RX 1.5 Extra
16	Charalambous et al. (2018)	N_2 e CO_2	Zeólita AQSOA-FAM-Z02
17	Kloutse et al. (2018)	N_2, CH_4 e CO_2	MOF ZIF-8
18	Moreira et al. (2019)	N_2, CH_4 e CO_2	MOF UiO-66
19	Khoramzadeh; Mofarahi; Lee (2019)	N_2 e CO_2	Zeólita 13X, 5A, 4A e Beta
20	Golipour et al. (2020)	C_2H_4 e C_2H_6	Zeólita 13X, CuX, 5A, CuA
21	Ursueguía; Díaz; Ordóñez (2020)	N_2 e CH_4	MOFs Basolite C300, F300 e A100
22	Boer et al. (2022)	CO_2 e CH_4	Zeólitas do tipo LTA

Fonte: Autoria própria.

Para algumas referências, em especial, os carvões ativados, o método de estimativa de volume de poros (método clássico BJH - Barrett-Joyner-Halenda) permitiu a estimativa dos volumes de microporos e mesoporos, separadamente. Para a confecção do banco de dados, a variável “Volume de poro”, disponível na Tabela 4, indica a soma destas duas grandezas.

Ainda sobre as variáveis apresentadas na Tabela 4, adotou-se a nomenclatura "Área superficial BET média" tendo em vista que alguns estudos reportam a Área superficial BET (relativa à teoria clássica de Brunauer-Emmett-Teller) em forma de intervalo numérico. Neste caso, no banco de dados, adotou-se a média do intervalo.

Tabela 4 - Detalhes sobre as variáveis contidas no banco de dados.

Variável de entrada	**Intervalos para as variáveis contínuas ou identidades para as variáveis discretas**
Classe do adsorvente [-]	Zeólita, MOF e carvão ativado
Volume de poro [cm^3/g]	0,096-1,31
Área superficial BET média [m^2/g]	24 - 3054
Temperatura [K]	212,70 – 499,44
Pressão [KPa]	0,00 – 9032,26
Método de medida [-]	Gravimétrico e volumétrico
Diâmetro cinético do adsorvato [Å]	3,300 - 4,71
Peso molecular do adsorvato [g/mol]	14,01 - 44,10
Polarizabilidade do adsorvato [cm^3]	$1{,}7403E^{-24}$ - $6{,}33E^{-24}$

Fonte: Autoria própria.

É importante salientar que, até onde se sabe, não existem estudos desta natureza, envolvendo aprendizado de máquina, com bancos de dados coletados na literatura, visando estimar a capacidade adsortiva de gases puros em mais de um tipo de sólido. Trabalhos recentes envolvem bancos de dados gerados por simulação, em geral, calculados por métodos de química teórica. Além disso, os estudos realizados focam na avaliação da adsorção de uma única espécie gasosa (CAO, 2022; HUANG *et al.*, 2022; MENG; ZHONG; WEI, 2020; NAIT AMAR *et al.*, 2022; RAJI *et al.*, 2022; XIE, J.; ZHANG, 2022; YAN *et al.*, 2021, 2022; ZHANG, XUAN; ZHENG; HE, 2022).

Através de dados gerados por simulação, o pesquisador pode coletar uma gama de informações sobre a topografia do sólido. Fato que não está acessível para um banco de dados dotado de informações experimentais. Por esta razão, no banco de dados analisado, trabalhou-se com um artifício para expressão dos efeitos de peneiramento molecular e avaliação da interação energética dos sítios ativos: foram computadas variáveis (ou *features*) que indicam

características físico-químicas dos gases adsorvidos: diâmetro cinético, peso molecular e polarizabilidade (vide Tabela 4). Esta metodologia ainda não foi reportada em estudos visando modelar o processo de adsorção por ML.

Com base nas características de exclusão de tamanho molecular do adsorvente, poros estreitos não estão disponíveis para adsorvatos com diâmetros maiores que a largura do poro, enquanto este mesmo poro pode ser acessível para um adsorvato menor (MASOUD JAHANDAR *et al.*, 2012). Como resultado, entende-se que o diâmetro cinético é uma variável que está diretamente relacionada ao volume de poros real do adsorvente. Seguindo este mesmo raciocínio, o peso molecular expressa-se como mais uma variável relacionada ao volume acessível do material.

Por outro lado, as interações eletrônicas entre a superfície dos sólidos podem ser explicadas, em parte, com base na polarizabilidade das moléculas. A análise teórica do fenômeno de adsorção mostra que diversos grupos funcionais disponíveis na superfície do sólido são polarizáveis e estão diretamente associadas ao aumento da adsorção de certos gases, em especial, os com maior capacidade de formar dipolos instantâneos (MEEK *et al.*, 2012; SCHÄF *et al.*, 2020).

3.2. Análise básica e exploratória dos dados

A análise básica de dados em ciência de dados refere-se a um conjunto de técnicas e métodos utilizados para explorar e compreender os dados em um estágio inicial do processo de análise. É uma etapa fundamental que envolve a organização, visualização e resumo dos dados, a fim de obter insights iniciais e identificar padrões, tendências ou problemas potenciais nos dados.

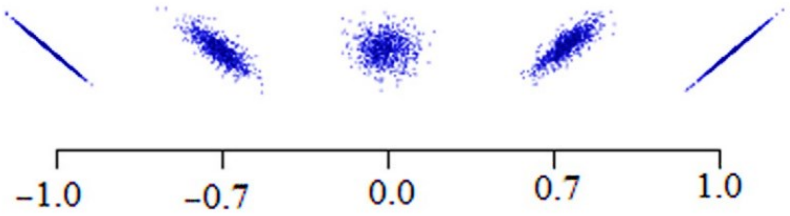
Geralmente, a análise básica envolve atividades de exploração, visualização e sumarização (isso envolve examinar as características dos dados, como sua estrutura, tamanho, tipos de variáveis, distribuições, entre outros). É importante entender a natureza dos dados antes de prosseguir com análises mais avançadas como a aplicação de modelos de aprendizado de máquina (MUKHIYA; AHMED, 2020).

No caso da presente dissertação, foram incluídos diagramas de caixa (*box plots*), histogramas, gráficos por pares (*pairplots*) e análise de correlação de Pearson que fornece um coeficiente conhecido como Coeficiente de Correlação de Pearson que indica uma medida quantitativa de dependência linear entre dois descritores numéricos e é definido, de acordo com Devore (2015), pela Equação 9. Na equação, $Cov(A, B)$ é a covariância entre as variáveis A e B e $\sigma_A \; e \; \sigma_B$ os desvios padrão de A e B, respectivamente.

$$\rho = \frac{Cov(A,B)}{\sigma_A \sigma_B} \tag{9}$$

O valor de ρ pode variar de -1 a 1, onde 0 indica não correlação, 1 correlação positiva total e -1 correlação negativa total. Uma correlação positiva entre a variável A e B significa que se a variável A subir, então B também aumentará, enquanto se o valor da correlação for negativo, se A aumentar, B diminui (NETTLETON, 2014). A Figura 20 ilustra pares de variáveis numéricas plotadas em relação entre si, com o valor de correlação, mostrado no eixo x, correspondente entre as duas variáveis.

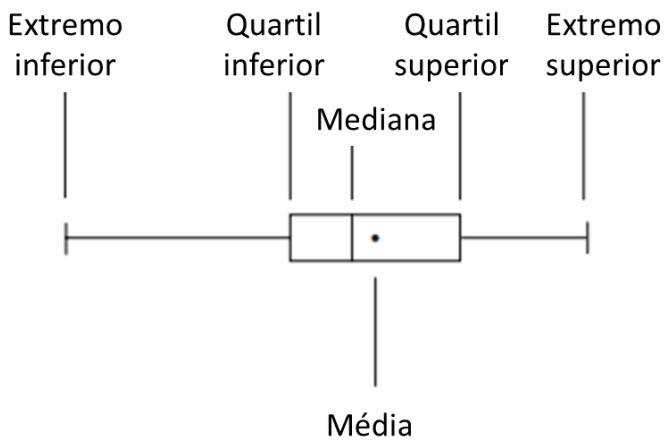
Figura 20 - Correlações entre duas variáveis numéricas



Fonte: Nettleton (2014)

Diagramas de caixa, também conhecido como boxplot ou *box plot*, são úteis para visualizar os quartis de uma determinada variável. Além disso, é um método efetivo para a representação de grandezas estatísticas como a mediana na análise de uma única variável. Além disso, é uma ferramenta útil para detecção de *outliers*, os quais são pontos que caem em uma região muito afastada do centro da distribuição (muito afastados da média e da mediana). Não há um consenso sobre a definição de um *outlier*. Porém, no caso do *box plot* em geral, existe uma definição formal. A maior parte das definições considera que pontos acima do valor do 3º quartil somado a 1,5 vezes a distribuição interquartilar (DI) ou os pontos abaixo do valor do 1º quartil diminuído de 1,5 vezes a DI são considerados outliers. A Figura 21 apresenta o formato de um *blox plot*. São considerados *outliers* valores menores que os pontos extremos (MYATT; JOHNSON, 2014).

Figura 21 - Formato de um diagrama de caixa.



Fonte: Adaptado de Myatt e Johnson (2014).

3.3. Detalhes sobre a modelagem e uso de ferramentas computacionais

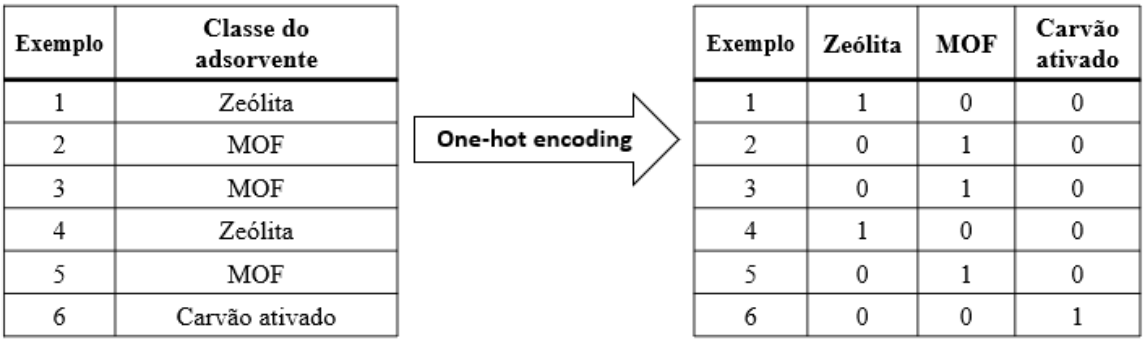
Para realização da análise exploratória de dados e regressão com modelo de aprendizado de máquina foi empregado o *software* online *Google Colaboratory* em ambiente de programação Python (versão número 3.7.13). Para análise de dados, as principais bibliotecas de uso foram numpy[1], matplotlib[2], seaborn[3] e pandas[4].

Antes da regressão dos modelos, as variáveis Classe do adsorvente (rótulos Zeólita, MOF, carvão ativado) e Método de medida (rótulos gravimétrico e volumétrico) foram transformadas através de *one-hot encoding*, a qual é uma técnica utilizada em ML para representar variáveis categóricas como vetores binários. Ela é usada quando queremos converter variáveis categóricas em um formato numérico adequado para algoritmos de aprendizado de máquina, que normalmente operam em cima de valores numéricos (RASCHKA; MIRJALILI, 2019).

O processo de *one-hot encoding*[5] envolve a criação de uma nova coluna para cada categoria distinta na variável categórica original. Cada coluna representa uma categoria específica, onde o valor binário 1 é atribuído para indicar a presença da categoria e 0 é atribuído para indicar a ausência dela. Portanto, para cada exemplo de entrada, apenas uma coluna terá o valor 1, enquanto todas as outras colunas terão o valor 0, conforme exemplifica a Figura 22.

[1] https://numpy.org/
[2] https://matplotlib.org/
[3] https://seaborn.pydata.org/
[4] https://pandas.pydata.org/
[5] https://pandas.pydata.org/docs/reference/api/pandas.get_dummies.html

Figura 22 - Representação esquemática da transformação de variáveis conhecida como *one-hot encoding*.



Fonte: Autoria própria.

Neste trabalho, a biblioteca scikit-learn[6] foi empregada para implementação dos algoritmos KNN, AD e RVS.

O crescimento das Árvores de Decisão foi realizado pelo algoritmo padrão (CART: *Classification and Regression Trees*) com o índice de diversidade de Gini como critério de impureza (BREIMAN, 2001).

Para o sucesso do algoritmo RVS aplicadas aos dados as funções de *kernel*, as quais distribuem os pontos não linearmente separáveis. Ou seja, essas funções realizam transformações nos pontos, tornando possível sua separação por um hiperplano. As funções de kernel podem ser bastante simples (THARWAT, 2019). No caso do presente trabalho, foram estudados os *kernels* implementados e disponíveis na biblioteca em uso: *Radial Basis Function* (RBF) e sigmoide.

Para o uso do modelo de aprendizado de máquina supervisionado, 25% do banco de dados foi separado para o conjunto de teste e 75% 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 [7].

Para a modelagem, a variável alvo foi logaritmizada (via ln) por duas razões. Uma estatística e outra proporcional. Supondo que a distribuição da variável em questão possui um viés, ou seja, uma das extremidades elevadas e uma cauda longa, medidas como correlação ou regressão podem ser bastante influenciadas pelo pico da distribuição, outliers, dentre outros. A aplicação da transformação pode reduzir o efeito do viés. Por outro lado, os dados de adsorção

[6] https://scikit-learn.org/stable/
[7] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.learning_curve.html

apresentam valores com escalas de grandezas diferentes. A transformação logarítmica auxilia a suavizar a diferença entre os valores extremos (LANE, 2022; WANG, Z. F., 2022).

Para a modelagem utilizando os modelos KNN e SVR, as variáveis foram padronizadas (escritas em termos de *z-scores*), conforme Equação 10, sendo z a variável padronizada e μ e σ a média aritmética simples e o desvio padrão do atributo x, respectivamente. A normalização e a padronização são técnicas frequentemente aplicadas na etapa de preparação dos dados, com o objetivo de colocá-los em um intervalo de valores comuns. Como KNN e SVR lidam com cálculos de distâncias, a padronização auxiliou a colocação de todos as características modeladas na média dos dados em 0 e o desvio padrão em 1.

$$z = \frac{x - \mu}{\sigma} \tag{10}$$

Para busca dos (hiper)parâmetros ótimos dos modelos foi utilizada a biblioteca computacional Optuna[8]. Tal biblioteca permite a pesquisa automatizada de (hiper)parâmetros ao implementar o algoritmo Bayesiano como metodologia de otimização (PRAVIN *et al.*, 2022). Para busca extensiva da região de ótimo em todo o espaço de parâmetros, foi empregado o buscador GridSearchCV[9]. Os (hiper)parâmetros que foram avaliados durante a otimização, bem como sua breve descrição, são apresentados na Tabela 5 para o KNN, Tabela 6 para o AD e na Tabela 7 para o RVS.

A função objetivo de otimização pela rotina implementada com o auxílio da biblioteca Optuna trabalhou com a maximização da média dos coeficientes de determinação (R^2) obtidos pelo treinamento do modelo. Como a variável alvo apresentou valores em diferentes escalas de grandeza (de 10^{-4} a 10^{1}), seus valores foram transformados pelo operador logaritmo de base 10 antes da regressão.

Tabela 5 - Hiperparâmetros para o modelo KNN, sua descrição breve e intervalo de busca.

(Hiper)parâmetros	**Descrição**	**Intervalores para variáveis contínuas ou identidades para variáveis discretas**
K	Número de vizinhos	2-20
metric	Métrica para avaliação de distâncias entre os pontos	Minkowski, euclidiana, Chebyshev, cossenos e Manhattan
weights	Função de peso	Uniforme e distância

Fonte: Autoria própria (2023).

[8] https:// https://optuna.org/
[9] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

Tabela 6 - Hiperparâmetros para o modelo AD, sua descrição breve e intervalo de busca.

(Hiper)parâmetros	**Descrição**	**Intervalores para variáveis contínuas ou identidades para variáveis discretas**
splitter	Estratégia utilizada para escolher o *split* em casa nós	Best, random
max_features	Número de características (*features*) consideradas quando for avaliado o melhor *split*	2 – número de *features*
max_depth	Máxima profundidade da(s) árvore(s)	5-20
min_samples_leaf	Número mínimo de amostras requerido para estar em um nó-folha	1 – número de amostras
max_leaf_nodes	Número máximo de nós-folhas	15-500

[a] – somente para FA

Fonte: Autoria própria (2023).

Tabela 7 - (Hiper)parâmetros para o modelo RVS, sua descrição breve e intervalo de busca.

(Hiper)parâmetros	**Descrição**	**Intervalo**
$\boldsymbol{C}$	Penalidade do erro	1- 100
ε	Largura da via	10^{-2} - 2
γ	Sensibilidade às diferenças nos vetores de erro das características	10^{-3} - 2
Coef0	Termo de viés na função do kernel	0 - 4
Tol	Tolerância para o critério de parada	10^{-6} - 0,1

Fonte: Autoria própria (2023).

Inicialmente, realizou-se uma avaliação da influência do número de *k-folds* sobre o coeficiente de determinação e verificou-se não existir influência significativa para valores até 21 (vide Apêndice A, Figura A.1). Desta forma, durante o treinamento, foi realizado o procedimento de validação cruzada[10] com *k-folds* fixado em 7 (ZHANG; LIU, 2022). Para cada procedimento de busca, 300 passos iterativos foram realizados para a busca do ponto de máximo da função objetivo.

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

A otimização buscou maximizar valores de R^2. Contudo, na análise dos resultados, outras métricas de ajuste foram utilizadas para comparar a performance dos modelos que foram o Erro Quadrático Médio (MAE - do inglês, *Mean Absolute Error*) e a Raiz do Erro Quadrático Médio (RMSE - do inglês, *Root Mean Squared Error*). As definições destas grandezas são apresentadas nas Equações 11 a 13 (AMAR et al., 2022; PAN et al., 2022).

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n-1}(y_i, \hat{y}_i)^2}{\sum_{i=0}^{n-1}(y_i, \bar{y})^2} \tag{11}$$

$$RMSE(y, \hat{y}) = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2} \tag{12}$$

$$MAE(y, \hat{y}) = \frac{1}{n}\sum_{i=0}^{n-1}|y_i - \hat{y}_i| \tag{13}$$

Em que y_i e $\hat{y}_i$ correspondem ao valor real e predito do registro i, respectivamente, e $\bar{y}$ refere-se a média dos valores reais.

3.4. Interpretabilidade do modelo

A seguir, é delineada a aplicação da interpretabilidade do modelo, mediante a utilização das técnicas de análise de importância por permutação e do Gráfico de Dependência Parcial (PDP), no contexto deste estudo.

3.4.1. Importância por permutação

A análise de importância por permutação foi aplicada para avaliar a relevância de cada atributo na previsão do modelo, utilizando o coeficiente de determinação (R^2) como métrica de avaliação. Este procedimento foi realizado por meio de trinta repetições de embaralhamento dos dados para cada atributo, proporcionando uma avaliação robusta e abrangente da influência de cada variável no processo de adsorção. Os resultados foram consolidados pela obtenção da média dos R^2 obtidos nas trinta repetições, sendo então avaliada a diferença entre essa média e o valor inicial de R^2 sem permutação dos dados.

Os atributos que mais impactaram a previsão no modelo de capacidade de adsorção foram identificados como aqueles que apresentaram uma significativa variação no R^2 após a

permutação dos dados. Essa variação expressiva reflete a sensibilidade desses atributos nas predições, destacando sua relevância para o desempenho do modelo. A aplicação da métrica R^2 em conjunto com a análise repetitiva reforça a confiabilidade estatística dos resultados, conferindo solidez às conclusões alcançadas neste estudo.

3.4.2. Gráfico de Dependência Parcial

No decorrer da investigação sobre a capacidade de adsorção, além da análise de importância por permutação, empregou-se o *Partial Dependence Plot*[11] (PDP), conhecido também como Gráfico de Dependência Parcial, como uma ferramenta adicional para a interpretabilidade do modelo, visando proporcionar uma visualização mais clara e intuitiva das relações entre variáveis, permitindo uma compreensão aprofundada do impacto individual de cada atributo na capacidade de adsorção de gases leves.

Essa abordagem combinada, envolvendo análise de importância por permutação e PDP, visa não apenas avaliar a influência relativa de variáveis, mas também proporcionar *insights* visuais sobre como essas variáveis interagem e afetam a capacidade de adsorção. Essa estratégia robusta de interpretabilidade fortalece a fundamentação do estudo, contribuindo para uma compreensão mais abrangente e confiável do modelo utilizado.

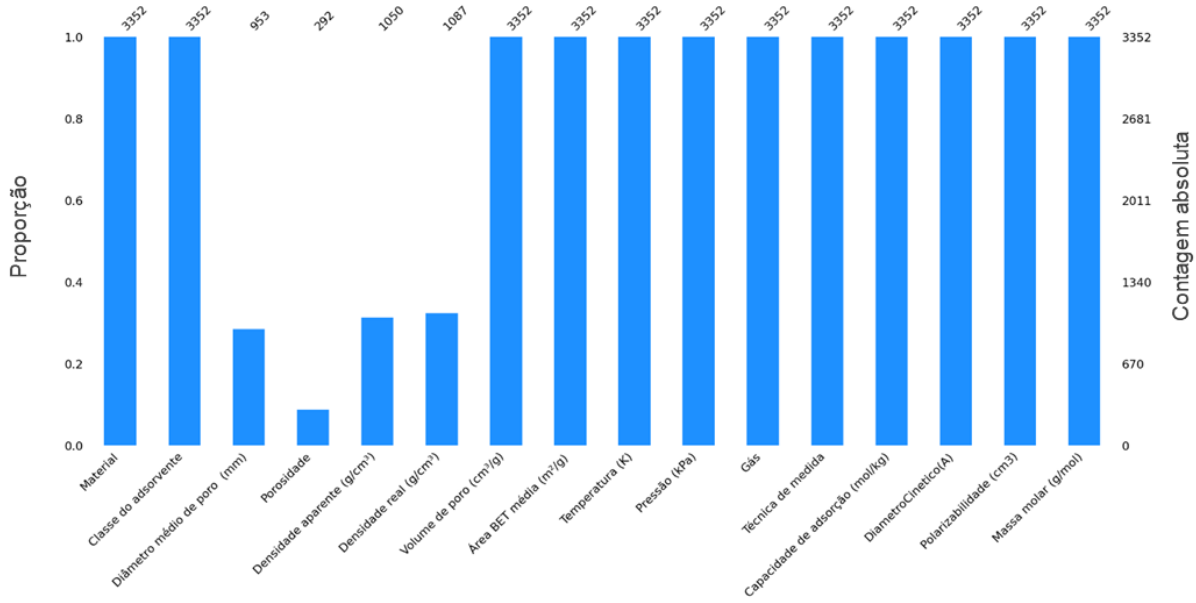
[11] https://scikit-learn.org/stable/modules/generated/sklearn.inspection.plot_partial_dependence.html

4. RESULTADOS E DISCUSSÕES

4.1. Análise básica e exploratória dos dados

A Figura 23 apresenta um gráfico indicando o número de dados por variável coletada (e/ou adicionada). Surpreende a falta de dados sobre o diâmetro médio de poros uma vez que é uma informação crucial para discussão de fenômenos de superfície. Informações sobre porosidade, densidade real e aparente são informações cruciais quando o experimentalista trabalha com experimentos em pellets, visando descrever a permeação do gás pelo leito fixo. Contudo, a maioria de dados reportam adsorção em pós, devido à escala dos equipamentos disponíveis em escala laboratorial. Assim, estes dados também se mostraram escassos. Devido as lacunas citadas, estas variáveis (diâmetro médio de poros, densidades e porosidade) não foram utilizadas na modelagem por aprendizado de máquina.

Figura 23 - Número de informações por característica do banco de dados.

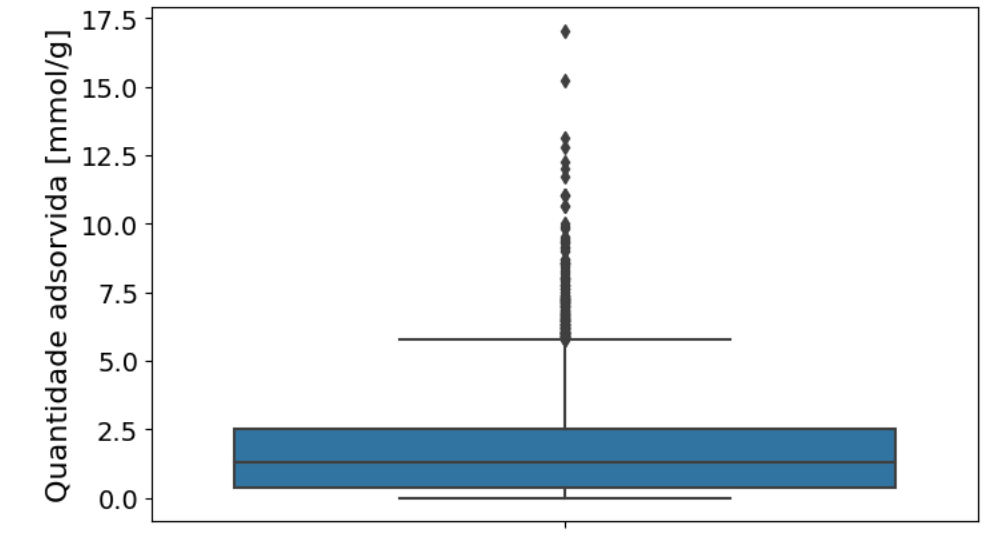


Fonte: Autoria própria (2023).

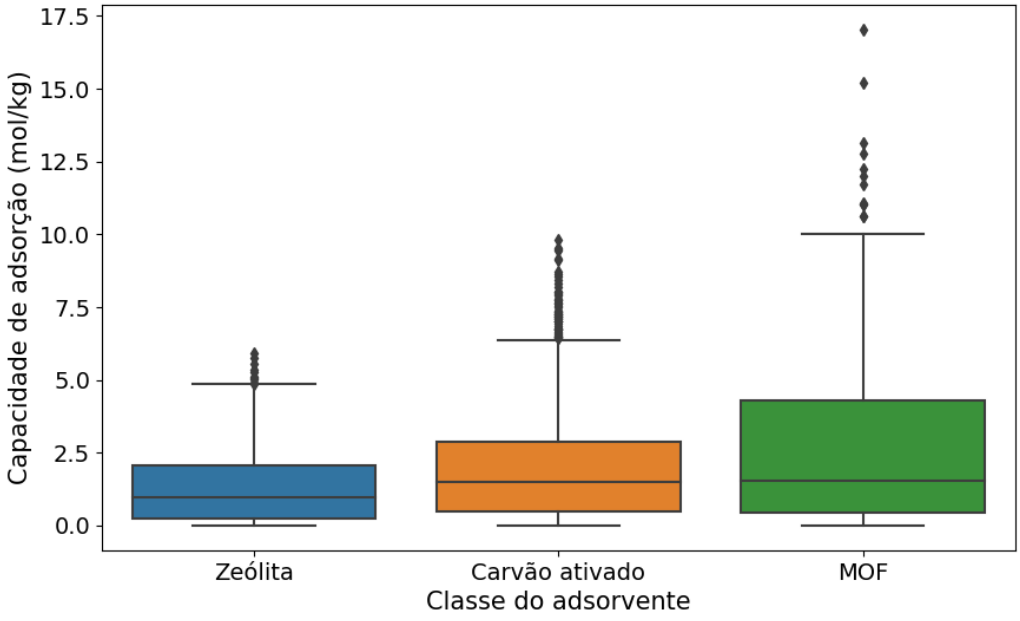
A Figura 24(a) apresenta o *boxplot* da quantidade adsorvida tendo por base o todo os dados do banco de dados (considerando as diferentes classes de materiais). Nota-se que os valores mais altos se distanciam da média global. A Figura 24(b), por outro lado, apresenta os *boxplots* tendo por base os tipos de materiais. Devido à heterogeneidade dos dados coletados no que se refere à capacidade adsorvida, fato que dificultou em etapas futuras o ajuste de modelos de aprendizado de máquina, a remoção de dados do banco de dados inicial precisou

ser efetuada. Com base nos resultados apresentados na Figura 24(b) os dados acima de 10 mmol/g foram removidos, ficando o banco de dados com 3.339 exemplos (foram removidos 13 itens).

Figura 24 - Diagramas de caixa considerando a quantidade adsorvida em cada classe de material disponível no banco de dados: Zeólita, Carvão Ativado e MOFs



(a)



(b)

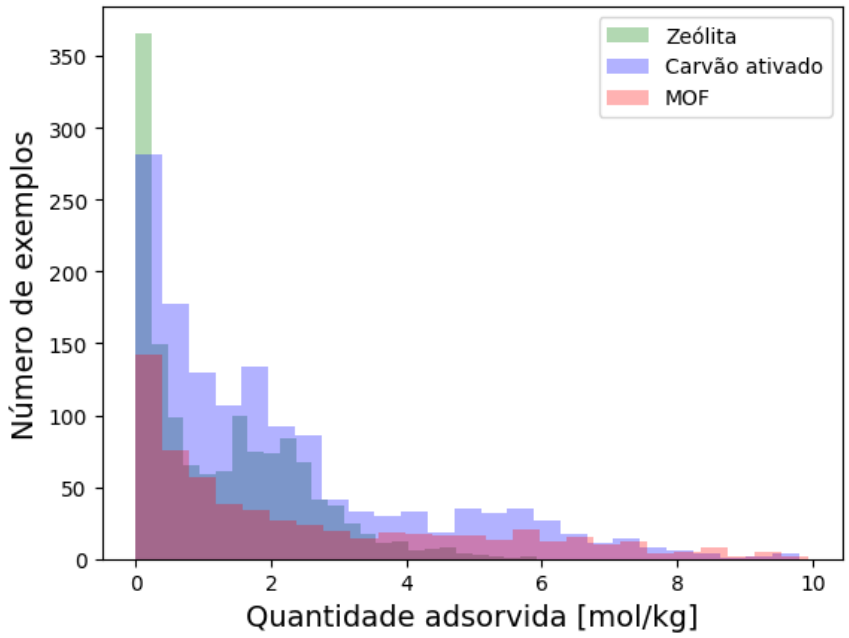
Fonte: Autoria própria (2023).

Ao final da remoção, seguiu-se com 1359 exemplos em carvão ativado, 606 em MOFs e 1374 em zeólitas. Também é interessante comentar que os MOFs, apesar de serem materiais relativamente novos e proeminentes dentro de diversas áreas dentro da ciência de materiais,

ainda têm poucos dados experimentais reportados em relação aos materiais classicamente empregados nos processos industriais que são os carvões e as zeólitas.

As Figura 25 apresenta o histograma com os valores de frequência de capacidade adsorvida, tendo-se por alicerce a classe do adsorvente. A análise da Figura 25 indica que o banco de dados apresenta classes de sólidos com desempenhos variados, o que é fundamental quando se deseja aplicar um algoritmo de aprendizado indutivo.

Figura 25 - Frequência dos valores de capacidade adsorvida indicados em termos das classes de adsorventes.

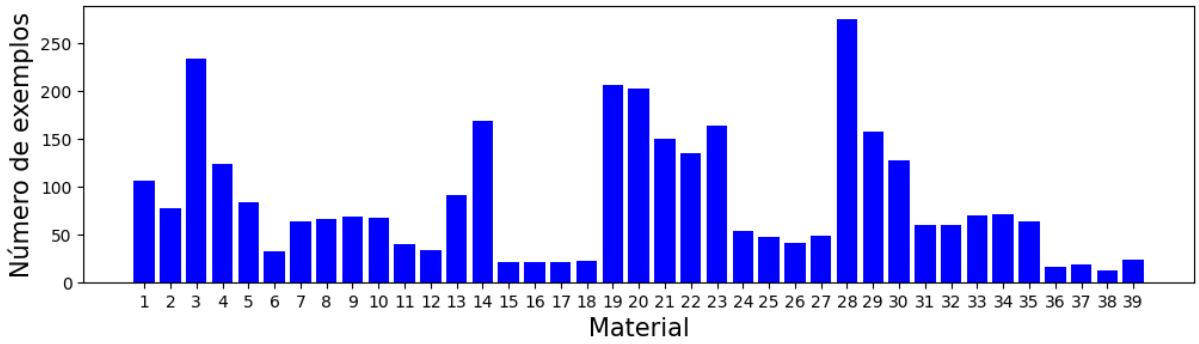


Fonte: Autoria própria (2023)

A Figura 26 apresenta o número de dados tendo por referência os 39 materiais incluídos no banco de dados. O material 28, o qual se trata de um carvão ativado é o que computa maior número de exemplos. Al-Muhtaseb; Al-Rub e Zarooni (2007) obtiveram isotermas para quatro tipos de gases trabalhando em 4 temperaturas fixas para alcançar esse patamar de informações.

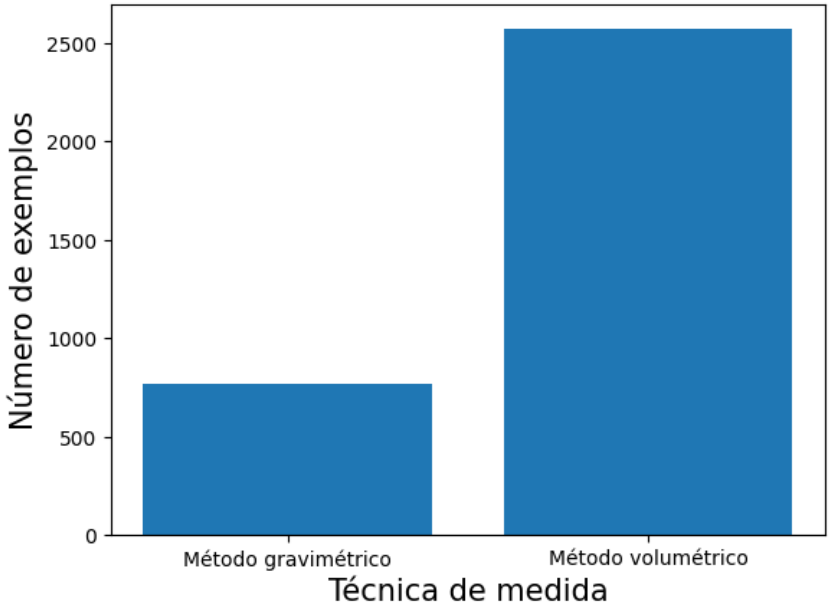
Em se tratando da técnica de medida (Figura 27) utilizada para a obtenção dos dados experimentais, nota-se que o método volumétrico é majoritariamente empregado. Conforme revisão de literatura, o método gravimétrico é mais direto e não requer análise química adicional.

Figura 26 - Número de exemplos por material estudado no banco de dados.



Fonte: Autoria própria (2023)

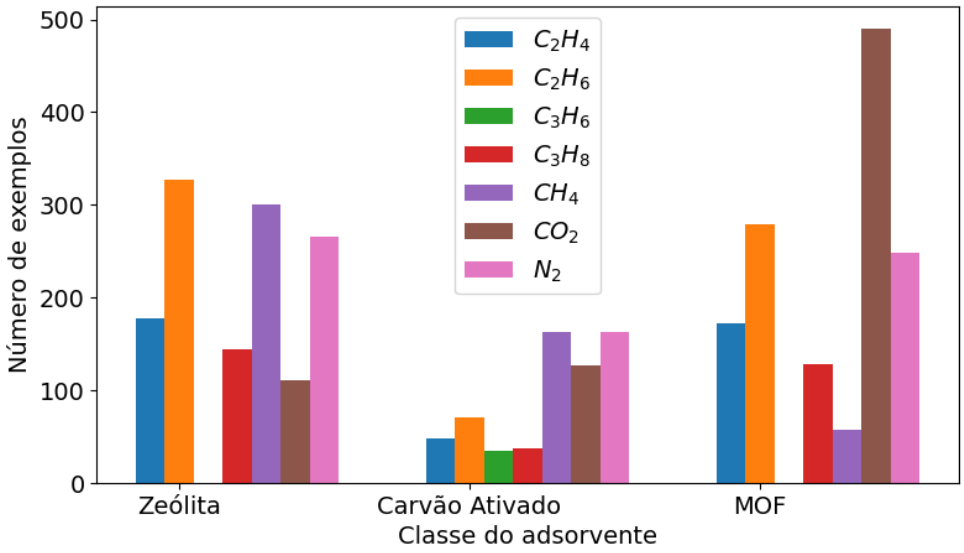
Figura 27 - Número de exemplos categorizados em termos da técnica de medida dos dados de equilíbrio de adsorção.



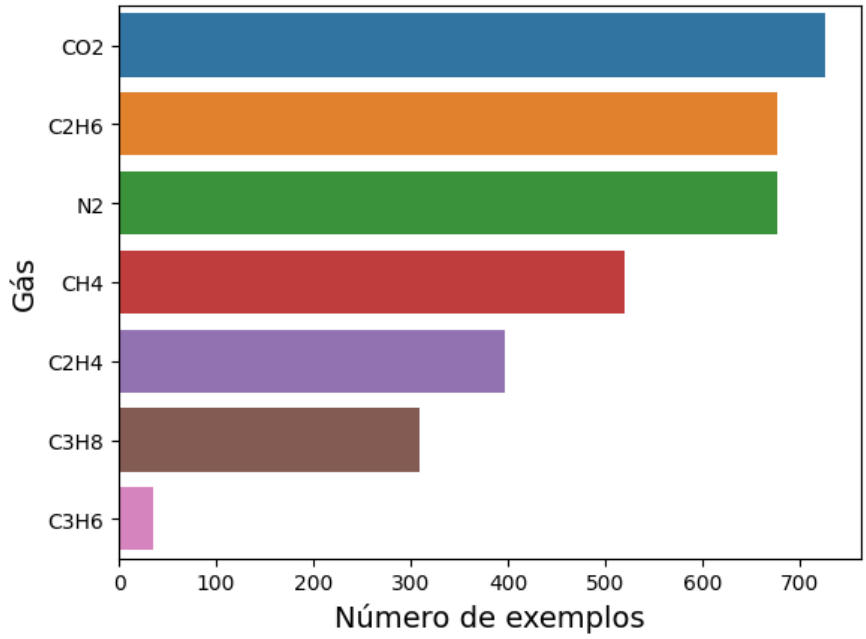
Fonte: Autoria própria (2023)

Foram coletados dados para todos os gases sendo adsorvidos em todas as classes de adsorventes, exceto C_3H_6 (Figura 28 (a)). De forma geral, nota-se que o banco de dados é bastante heterogêneo em termos das espécies selecionadas, apresentando 726 dados para CO_2 e 35 para o C_3H_6 (Figura 28 (b)). Grande parte dos dados referem-se à adsorção de CO_2 e N_2 em zeólitas (Figura 28(b)), dados que estão amplamente disponíveis na literatura devido ao apelo tecnológico dos processos dos processos de captura de CO_2, purificação dos gases de chaminé e, ao mesmo tempo, da ampla disponibilidade comercial de materiais zeolíticos (FLANIGEN, 2001; KHORAMZADEH; MOFARAHI; LEE, 2019).

Figura 28 - Número de exemplos considerando cada gás do banco de dados (a) por classe de adsorvente; (b) pelo banco de dados total.



(a)



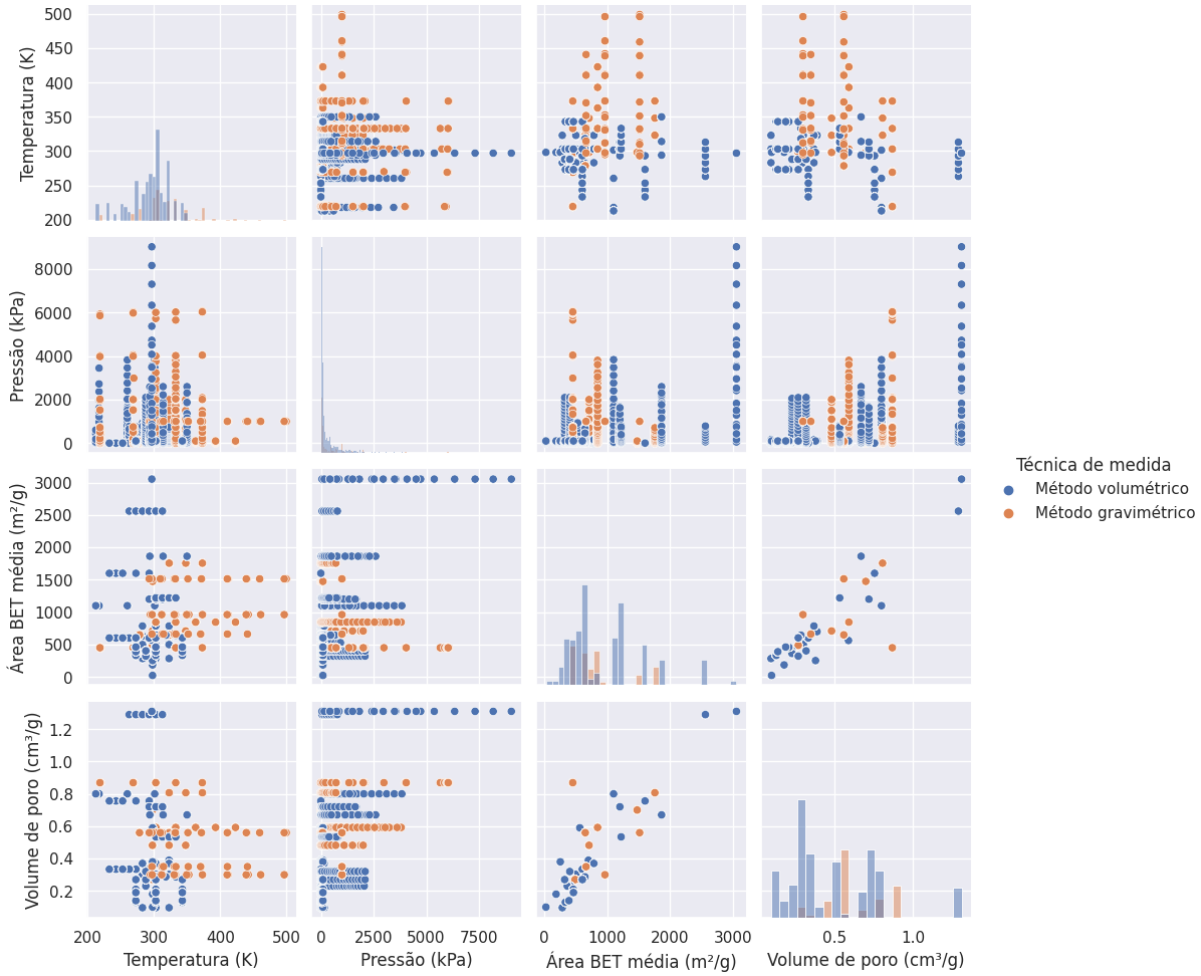
(b)

Fonte: Autoria própria (2023).

A Figura 29 apresenta uma sequência de gráficos de dispersão e histogramas para as variáveis Temperatura (K), Pressão (kPa) e Área superficial BET média (m²/g), tendo por base o método empregado para obtenção das medidas de adsorção. Nota-se que o banco de dados idealizado apresenta, em sua maioria, dados coletados a pressões baixas e temperaturas próximas à ambiente. Tal constatação pode estar relacionada ao fato de baixas pressões e temperaturas ambientes serem comumente adotadas na etapa de adsorção ou dessorção de

diversos sistemas de colunas como PSA (Adsorção por Variação de Pressão, do inglês *Pressure Swing Adsorption*), VPSA (Adsorção por Variação de Pressão a Vácuo, do inglês *Vacuum Swing Adsorption*) e TSA (Adsorção por Variação de temperatura, do inglês *Temperature Swing Adsorption*) (DANACI; WEBLEY; PETIT, 2021; SPEIGHT, 2019). Prevalentemente, também são observados dados contendo áreas BET médias inferiores à 1000 m²/g, característica esta referente a maioria dos adsorventes sintéticos (KARIMI, S.; TAVAKKOLI YARAKI; KARRI, 2019). Não foi possível observar qualquer tendência entre as variáveis ilustradas e os métodos empregados nas medidas experimentais.

Figura 29 - Gráfico de pares para as variáveis Temperatura (K), Pressão (KPa), Área BET média (m²/g) e Volume de poros (cm³/g) contidas no banco de dados analisado tendo por discriminante a técnica de medida empregada no experimento de adsorção.



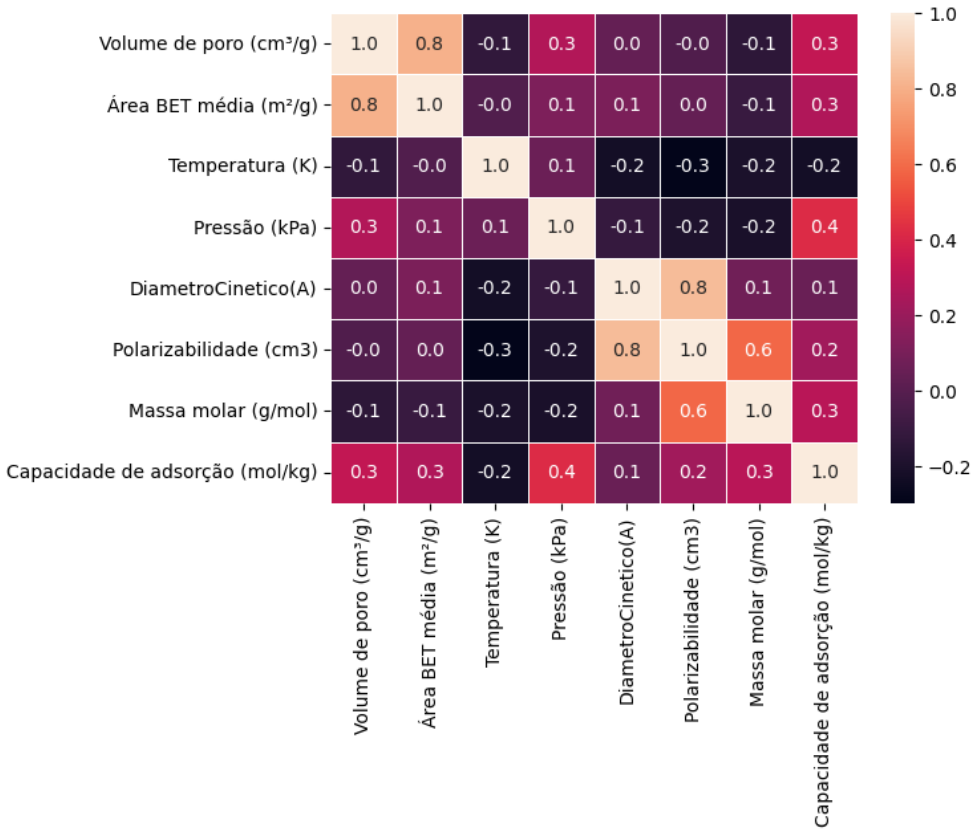
Fonte: Autoria própria (2023)

Durante o pré-tratamento dos dados, observou-se que as variáveis: área superficial e volume médio de poros estavam linearmente correlacionadas de forma positiva (Figura 30). Tal fato já era esperado, tendo em vista a natureza dos métodos BET (Brunauer, Emmett e Teller)

(FREEMAN; MCLEOD, 1983) e BJH (Barrett-Joyner-Halenda), os quais foram aplicados nas referências citadas. O uso de ambas as variáveis ou não foi explorado na etapa de regressão pois elas codificam efetivamente as mesmas informações.

O diâmetro cinético é uma medida aplicada a átomos e moléculas que expressa a probabilidade de uma molécula em um gás colidir com outra molécula. É uma indicação do tamanho da molécula como alvo. Este diâmetro, no caso dos gases presentes no banco de dados, se mostrou fortemente correlacionado com a polarizabilidade. Segundo Loukhovitski e Sharipov (2021), a polarizabilidade (α) é uma propriedade física fundamental das partículas em colisão uma vez que o choque é essencialmente responsável pela interação intermolecular do tipo de van der Waals. Assim, é esperado que a densidade eletrônica de um átomo (e, portanto, grandezas relativas ao seu tamanho em escala linear), pelo menos dentro da aproximação quase clássica, tenham alguma proporcionalidade.

Figura 30 - Coeficiente de correlação de Pearson entre as variáveis do banco de dados.



Fonte: Autoria própria (2023).

Por fim, a mostra a grandezas da estatística descritiva do banco de dados, incluindo aquelas que resumem a tendência central, dispersão e forma da distribuição de um conjunto de

dados. Como indicado na Tabela 8, as distribuições de Área BET média e Pressão são bastante assimétricas, o que pode ser observado pelo afastamento dos valores de média e mediana.

Tabela 8 - Grandezas estatísticas básicas do banco de dados utilizado para regressão do modelo de aprendizado de máquina

Grandeza	Média	Desvio Padrão	Valor Mínimo	Quartil 25%	Mediana	Quartil 75%	Valor máximo
Volume de poro (cm^3/g)	0,52	0,28	0,10	0,30	0,53	0,72	1,31
Área BET média (m^2/g)	924,31	577,50	24,00	499,74	650,00	1220,00	3054,00
Temperatura (K)	300,02	33,77	212,70	283,00	303,00	323,00	499,44
Pressão (kPa)	370,60	747,73	0,00	19,22	86,50	411,48	9032,26
Diâmetro Cinético (Å)	3,92	0,48	3,30	3,64	3,76	4,44	4,71
Polarizabilidade (cm^3)	$3,45.10^{-24}$	$1,38.10^{-24}$	$1,74.10^{-24}$	$2,59.10^{-24}$	$2,91.10^{-24}$	$4,46.10^{-24}$	$6,33.10^{-24}$
Massa molar (g/mol)	28,85	12,00	14,01	16,04	30,07	44,01	44,10
Capacidade de adsorção (mol/kg)	1,81	1,89	0,00	0,36	1,27	2,50	9,94

Fonte: Autoria própria (2023)

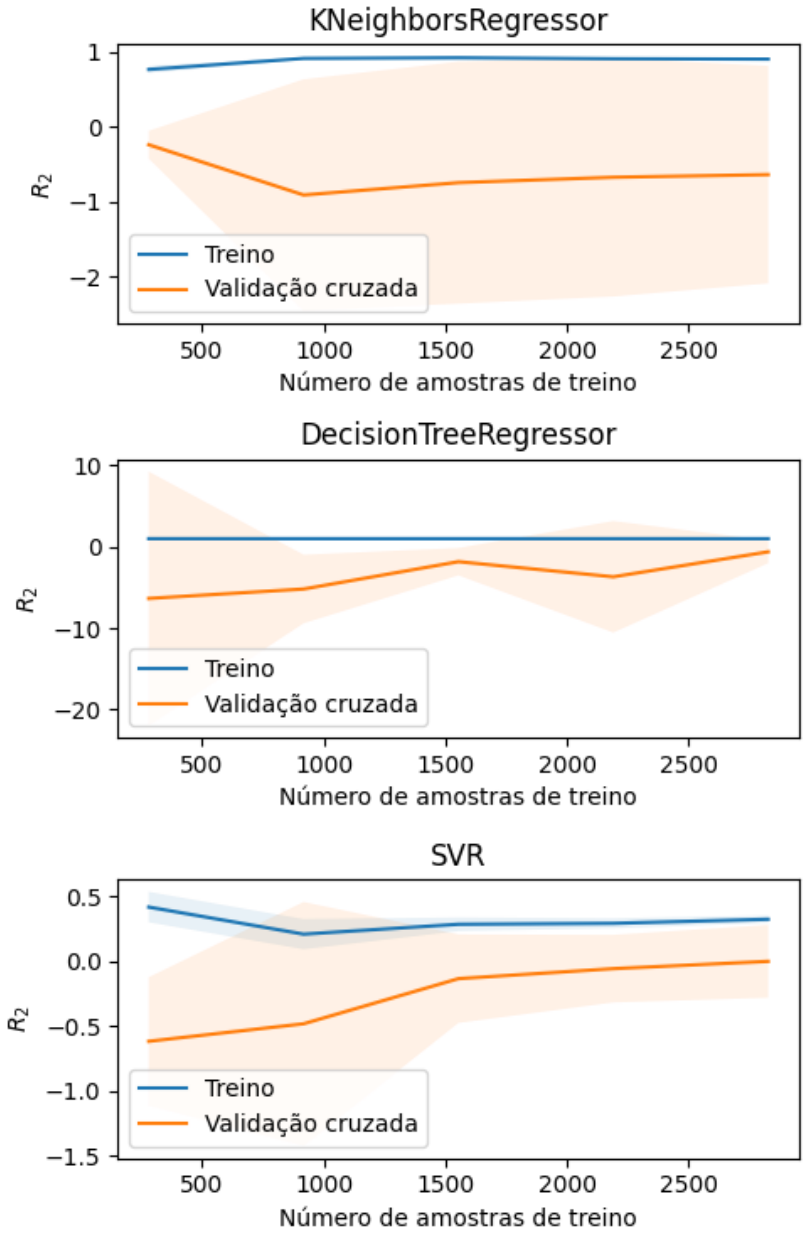
4.2. Curvas de aprendizado

Anteriormente ao ajuste dos modelos, foram construídas as curvas de aprendizado considerando os parâmetros *default* das bibliotecas KNeighborsRegressor(), DecisionTreeRegressor() e SVR(). Para construção das curvas, foram considerados os valores de *k-folds* iguais a 3, 7, 11 e 15. A Figura 31 apresenta os resultados para *k-folds=7,* valor que foi fixado nas etapas posteriores do trabalho. Demais resultados foram incluídos no Apêndice A desta dissertação.

A curva de aprendizado possibilita que verifiquemos os graus de sobreajuste dos modelos. Por exemplo, conforme Figura 31, o modelo SVR apresenta um menor efeito de sobreajuste para uma quantidade de amostras de treino maior que 1.500. Já o modelo KNN apresenta alto sobreajuste para qualquer tamanho de amostra.

É importante salientar que quando há mais de um hiperparâmetro a ser ajustado, a análise do efeito de apenas um conjunto deles pode ser enganosa, já que a relação expressa em uma curva de aprendizado pode ser completamente diferente para outras combinações dos demais hiperparâmetros. De qualquer forma, trata-se de uma ferramenta útil para verificar faixas hiperparamétricas em que possa ocorrer subajuste e sobreajuste. A partir dos resultados obtidos, trabalhou-se com uma amostra de treinamento igual a 75% da amostra total.

Figura 31 - Variação dos valores do coeficiente de determinação (R^2) para a regressão dos valores de capacidade adsorvida para os modelos KNN, AD e SVR como função do tamanho das amostras de treinamento considerando o procedimento de validação cruzada com *k-folds=7*



Fonte: Autoria própria (2023)

4.3. Regressão com KNN

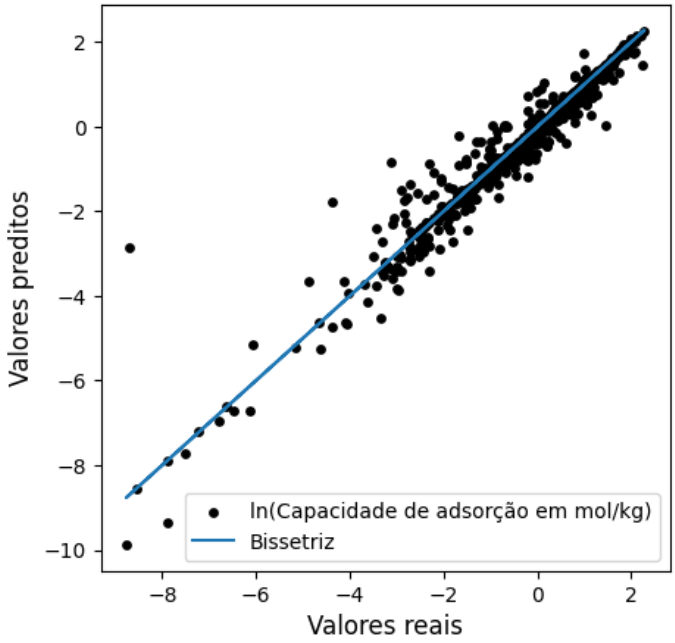
As métricas de ajuste para os conjuntos de treino e teste após o processo de otimização são apresentadas na Tabela 9. Nota-se que o modelo explicou em torno de 94% da variância dos dados. Além disso, há pouco sobreajuste, tendo em vista as métricas de treino e teste próximas. A Figura 32 atesta o sucesso do modelo, apresentado a comparação entre os valores reais e os preditos da variável alvo para o conjunto de treino.

Tabela 9 - Coeficiente de determinação (R^2), erro quadrático médio (MAE) e a raiz do erro quadrático médio (RMSE) para os conjuntos de treino e teste e parametrização pelo KNN

Métrica	**Treino – Validação cruzada**	**Teste**
R^2 [-]	0,940	0,945
MAE [kg/mol]	0,223	0,183
RMSE [kg/mol]	0,406	0,386

Fonte: Autoria própria (2023)

Figura 32 - Variável alvo (ln da capacidade de adsorção em mol/kg): valores reais *versus* valores preditos pelo algoritmo KNN para o conjunto de teste



Fonte: Autoria própria (2023)

Os hiperparâmetros ideais para o modelo KNN foram determinados por meio de um processo de ajuste, que incluiu a exploração de diversas combinações de hiperparâmetros utilizando a biblioteca Optuna. O objetivo desse processo foi encontrar, por meio de 300 iterações, a configuração mais adequada que maximizasse o desempenho do modelo por meio da métrica R^2. Os parâmetros ótimos foram: métrica tipo "cosseno" (Equação 5), $K = 3$ e pesos baseados no inverso das distâncias (Tabela 10). A Figura 33(a) mostra como 300 passos computacionais são suficientes para a convergência da busca. A Figura 33(b) mostra como a função de cálculo de distâncias do tipo "cosseno" apresentou, em geral, desempenho superior às demais para a maioria dos valores de K e com pesos baseados nas distâncias.

Tabela 10 - Melhores hiperparâmetros para o modelo KNN obtidos pela ferramenta Optuna

(Hiper)parâmetros	**Faixa de busca**	**Valor ótimo**
Número de vizinhos *(n neighbors)*	2-20	3
Métrica para avaliação de distâncias entre os pontos (*metric*)	Minkowski, euclidiana, Chebyshev, cosseno e Manhattan	Cosseno
Função de peso (*weights*)	Uniforme e Distância	Distância

Fonte: Autoria própria (2023)

A distância calculada pela função cosseno entre dois vetores de atributos é computada de acordo com a equação 5, sendo θ o cosseno entre os vetores. O cosseno de θ é um tipo de indicativo de similaridade entre os vetores. O cosseno de 0 é 1 e é menor que 1 para qualquer outro ângulo. Logo, o mesmo mede uma orientação e não uma magnitude: dois vetores com a mesma orientação têm cosseno 1 (similaridade de máxima ou "distância" nula); dois vetores à 90 graus têm similaridade de 0 e distância de 1 e dois vetores opostos têm similaridade de -1 e distância de 2. Este tipo de métrica de similaridade tem sido aplicada na resolução de diferentes problemas de mineração de texto, como classificação de texto, sumarização de texto, recuperação de informações, resposta a perguntas, entre outros (MANNING; RAGHAVAN; SCHÜTZE, 2009; ZOBEL *et al.*, 1995).

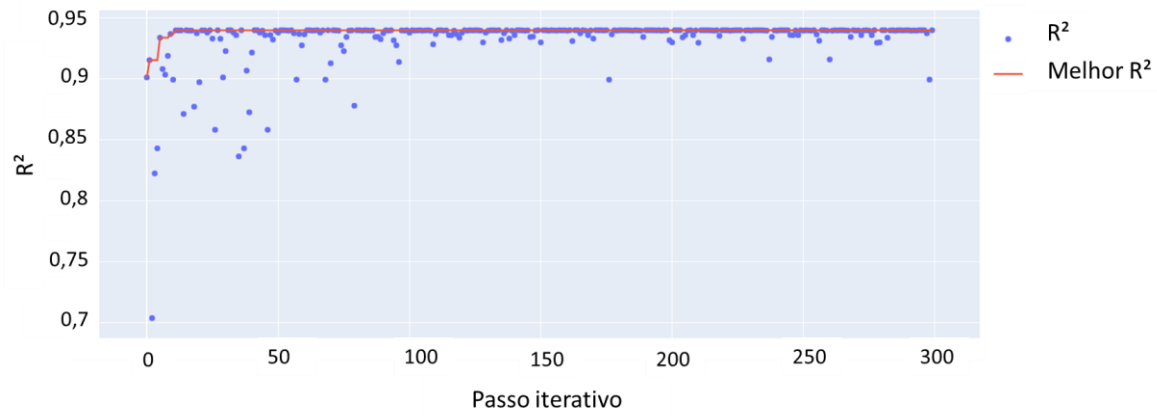
A biblioteca Optuna retorna após a otimização as importâncias dos parâmetros como um dicionário onde as chaves consistem nos nomes dos parâmetros e seus valores de importância. As importâncias são representadas por números decimais não negativos, onde valores mais altos indicam que os parâmetros são mais importantes. O dicionário retornado está ordenado por seus valores em ordem decrescente. Por padrão, a soma dos valores de importância é normalizada para 1,0. No caso da otimização com o KNN, conforme Tabela 11, as importâncias foram: 0,698 (para *weights*), 0,274 (para K) e 0,028 (para *metric*). Logo, o parâmetro mais sensível para o ajuste da modelagem se refere aos pesos dados na contabilização das distâncias. Quando *weights=Uniforme* o algoritmo retorna como valor predito na regressão a média dos K vizinhos (independente da distância). Já quando *weights=Distância,* a média é ponderada pelo inverso da sua distância entre os K vizinhos. Neste caso, os vizinhos mais próximos de um ponto de consulta terão uma maior influência do que os vizinhos que estão mais distantes. Este tipo de estratégia foi apresentado por Dudani (1976) e auxilia na minimização de erros advindos de classificações e regressões quando os valores de K são pequenos.

Tabela 11 - Análise de importância dos hiperparâmetros do modelo KNN

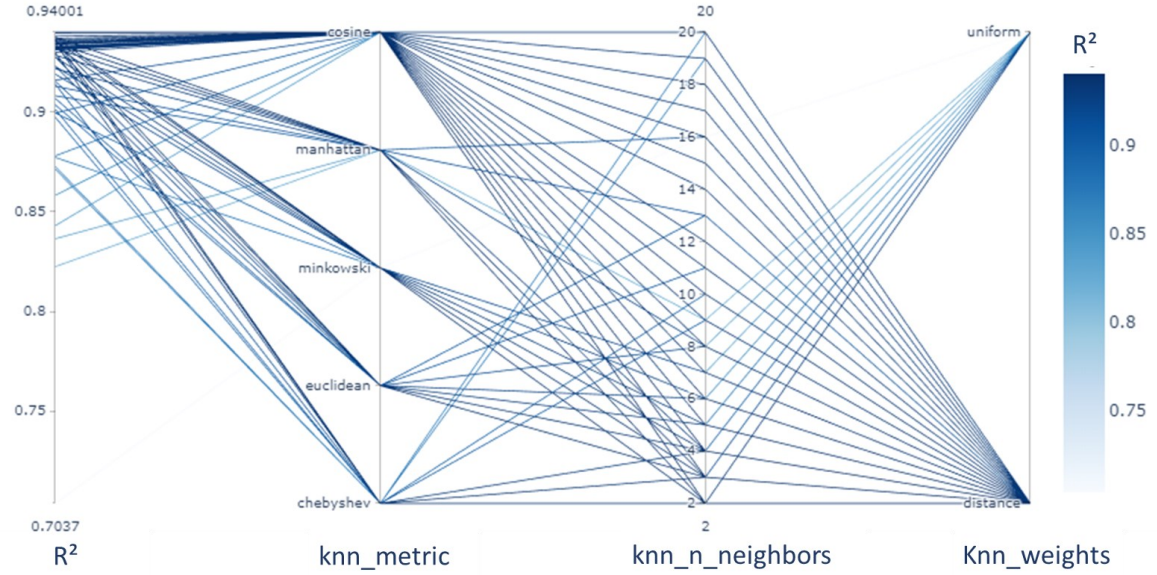
Hiperparâmetro	**Importância relativa**
Número de vizinhos *(n_neighbors)*	0,274
Métrica para avaliação de distâncias entre os pontos (*metric*)	0,028
Função de peso (*weights*)	0,698

Fonte: Autoria própria (2023)

Figura 33 - Processo de otimização dos hiperparâmetros do modelo KNN usando R^2 como métrica da função objetivo: (a) Gráfico do histórico de otimização; (b) Gráfico de coordenadas paralelas



(a)



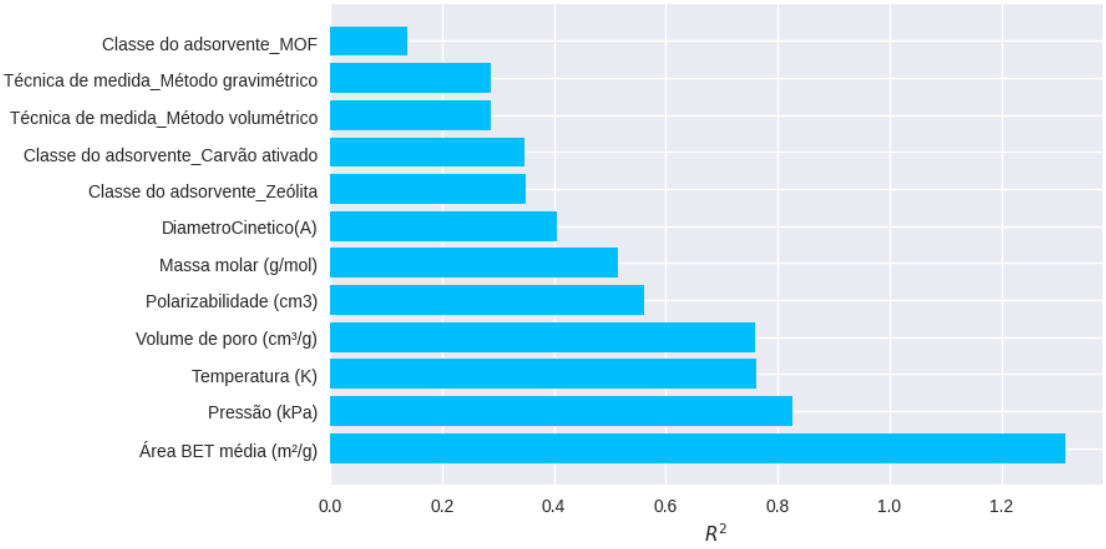
(b)

Fonte: Autoria própria (2023)

As importâncias de permutação das variáveis para o conjunto de teste são apresentadas na Figura 34. Nota-se que as variáveis operacionais (temperatura – 0,83 - e pressão – 0,76) e

estruturais (volume de poro – 0,76 - e área BET média – 1,31) figuram entre as mais importantes para a previsão do modelo. Já as classes e técnicas de medidas entre as variáveis que menos influenciam. Sendo a variável Classe do adsorvente - MOF a de menor valor com 0,14. Teste adicionais mostraram que não é possível remover variáveis pouco importantes para obter-se melhores ajustes. As variáveis volume de poro e área BET média que se mostraram correlacionadas na análise de dados, também não puderam ser removidas para este fim.

Figura 34 - Importâncias de permutação das variáveis para o modelo KNN otimizado



Fonte: Autoria própria (2023)

4.4. Regressão com AD

Conforme as métricas apresentadas na Tabela 9, o modelo de Árvore de Decisão (AD) para prever a capacidade de adsorção de gases leves demonstrou ser altamente promissor. Isso é evidenciado pelo valor do R^2, tanto no conjunto de treinamento (0,931) quanto no conjunto de teste (0,933), indicando que o modelo se ajustou bem aos dados de treinamento e é capaz de generalizar para novos dados. Além disso, é notável a obtenção de baixos valores de MAE (Erro Absoluto Médio) e RMSE (Erro Quadrático Médio) para ambos os conjuntos de dados - treino e teste -, inferindo desta forma que o modelo possui uma precisão considerável nas previsões. A Figura 35, que exibe a comparação entre os valores reais e as previsões da capacidade de adsorção obtidas pelo algoritmo de Árvore de Decisão no conjunto de teste, reforça o excelente desempenho observado pelas métricas do modelo. Isso significa que as previsões do modelo estão muito próximas dos valores reais, corroborando sua eficácia na tarefa de previsão. Desta forma, com base nas métricas e na análise visual dos resultados, o modelo de Árvore de Decisão

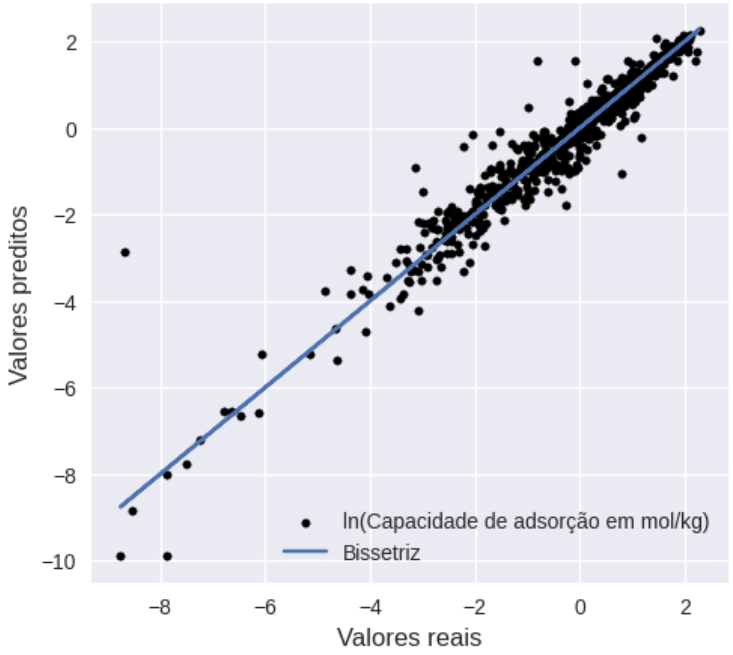
demonstrou ser uma escolha sólida e confiável para a previsão da capacidade de adsorção de gases leves, com capacidade de generalização e precisão notáveis.

Tabela 12 - Coeficiente de determinação (R^2), erro quadrático médio (MAE) e a raiz do erro quadrático médio (RMSE) para os conjuntos de treino e teste e parametrização por AD

Métrica	**Treino – Validação cruzada**	**Teste**
R^2 [-]	0,931	0,933
MAE [kg/mol]	0,270	0,238
RMSE [kg/mol]	0,435	0,426

Fonte: Autoria própria (2023)

Figura 35 - Variável alvo (capacidade de adsorção em mol/kg): valores reais versus valores preditos pelo algoritmo AD para o conjunto de teste

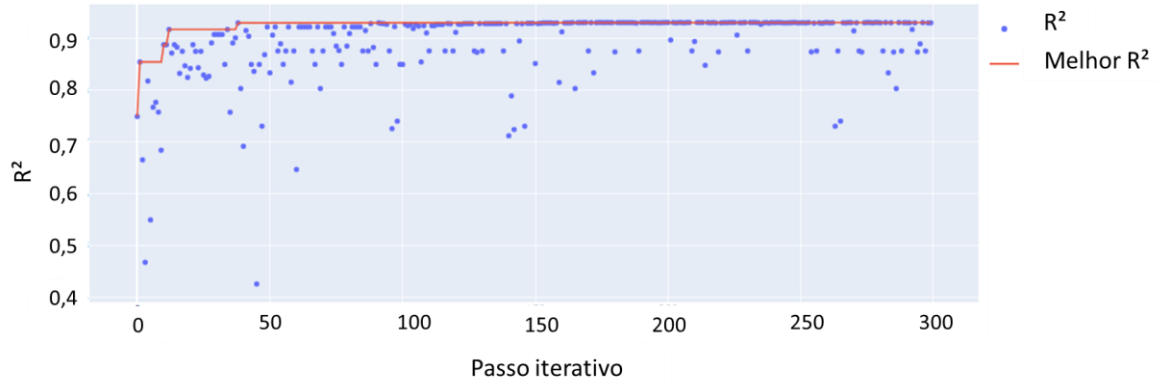


Fonte: Autoria própria (2023)

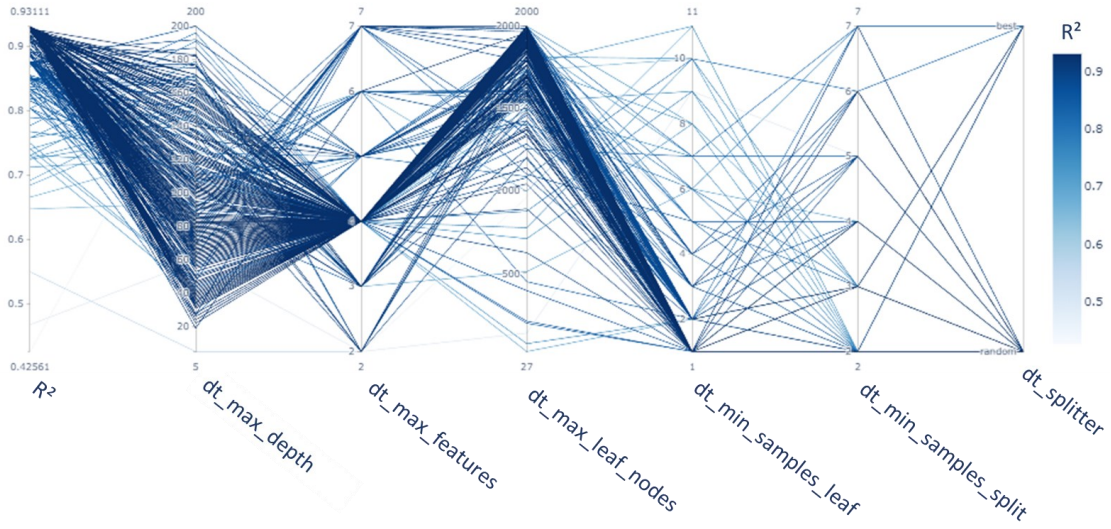
Assim como para o modelo KNN, o modelo de Árvore de Decisão (AD) foi obtido por meio de um processo de ajuste que envolveu a exploração de várias combinações de hiperparâmetros, tendo como objetivo encontrar a configuração ideal que maximizasse o desempenho do modelo. Nesse processo, a biblioteca Optuna desempenhou um papel importante ao buscar, em 300 passos iterativos, a otimização do desempenho do modelo de AD com base na métrica R^2. A Figura 36(a) mostra que 300 passos computacionais foram suficientes para a convergência da busca. A Figura 36(b) demonstra que há diversas

combinações de hiperparâmetros em que a função objetivo resulta em um valor de R^2 acima de 0,90.

Figura 36 - Processo de otimização dos hiperparâmetros do modelo AD usando R^2 como métrica da função objetivo: (a) Gráfico do histórico de otimização; (b) Gráfico de coordenadas paralelas



(a)



(b)

Fonte: Autoria própria (2023)

As faixas de busca dos hiperparâmetros, assim como os valores resultantes otimizados obtidos pela ferramenta Optuna são apresentados na Tabela 13. Os resultados da análise de importância dos hiperparâmetros (Tabela 14), obtidos por meio do Optuna, revelaram insights valiosos sobre a otimização do modelo de AD. Os hiperparâmetro 'dt_min_samples_leaf' e 'dt_max_features' se destacaram como os mais influentes na busca pelos melhores hiperparâmetros, indicando que a definição adequada do número mínimo de amostras por folha e o limite do número de atributos considerados para cada divisão desempenharam um papel fundamental no aprimoramento do desempenho do modelo, com uma importância relativa de

aproximadamente 47,9% e 34,2%, respectivamente. Além disso, ‘dt_max_leaf_nodes’ também apresentou importância significativa, com valor de cerca de 11,9%, enquanto ‘dt_splitter’, ‘dt_max_depth’ e ‘dt_min_samples_split’ demonstraram menor influência, com 1,5%, 2,2% e 2,3%, respectivamente. Esses resultados informam estrategicamente futuras iterações de otimização, direcionando a atenção para os hiperparâmetros mais relevantes e oferecendo insights valiosos sobre como ajustar o modelo de árvore de decisão para alcançar um desempenho otimizado.

Tabela 13 - Melhores hiperparâmetros para o modelo de AD obtidos pela ferramenta Optuna

Hiperparâmetro	**Faixa de busca**	**Valor ótimo**
Estratégia usada para escolher os atributos ao dividir um nó (dt_splitter)	*Best e random*	*random*
Profundidade máxima da árvore (dt_max_depth)	5 - 200	84
Número mínimo de amostras necessárias para dividir um nó (dt_min_samples_split)	2 - 7	2
Número mínimo de amostras necessárias em uma folha (dt_min_samples_leaf)	1 - 11	1
Limitação do número de atributos considerados para cada divisão (dt_max_features)	2 - Quantidade de variáveis de entrada selecionadas	4
Número máximo de nós folha na árvore (dt_max_leaf_nodes)	10 - 2000	2000

Fonte: Autoria própria (2023).

Tabela 14 - Análise de importância dos hiperparâmetros do modelo de árvore de decisão

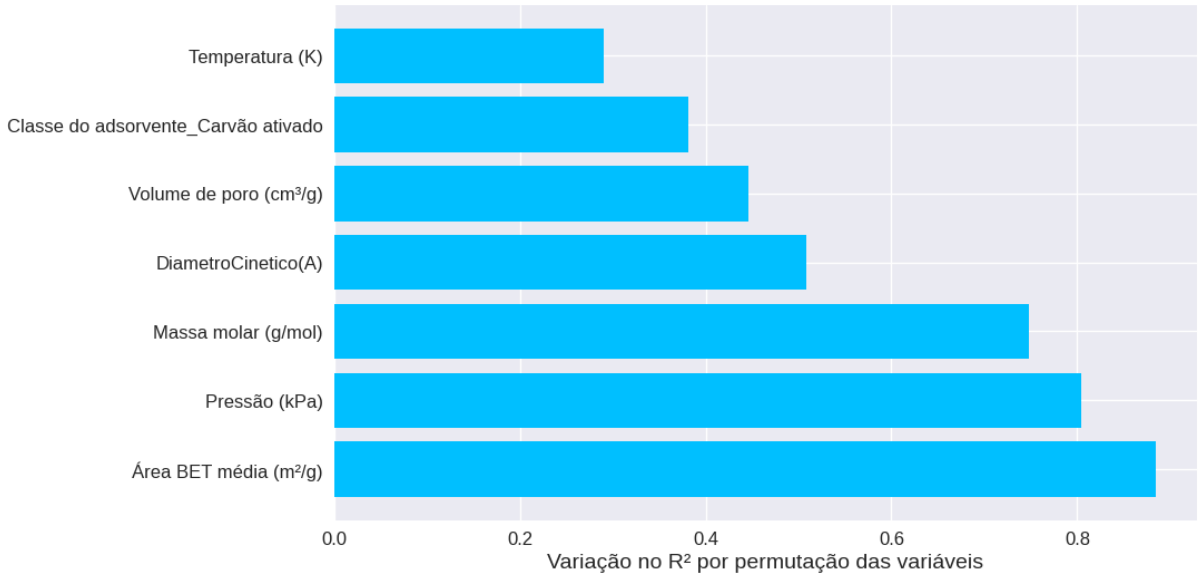
Hiperparâmetro	**Importância relativa**
Estratégia usada para escolher os atributos ao dividir um nó (*dt_splitter*)	0,015
Profundidade máxima da árvore (*dt_max_depth*)	0,022
Número mínimo de amostras necessárias para dividir um nó (*dt_min_samples_split*)	0,023
Número mínimo de amostras necessárias em uma folha (*dt_min_samples_leaf*)	0,479
Limitação do número de atributos considerados para cada divisão (*dt_max_features*)	0,342
Número máximo de nós folha na árvore (*dt_max_leaf_nodes*)	0,119

Fonte: Autoria própria (2023).

Visando aprimorar e simplificar o modelo de AD, as variáveis de entrada foram selecionadas a partir da análise de importância por permutação, considerando o conjunto de teste, removendo aquelas que apresentaram uma variação no R^2 inferior a 5% em relação ao R^2 obtido sem a permutação das variáveis. A Figura 37 apresenta o impacto das variáveis no desempenho do modelo construído após a remoção das variáveis que demonstraram possuir pouca importância, sendo elas: Técnica de medida - Método gravimétrico, Técnica de medida - Método volumétrico, Classe do adsorvente - MOF, Classe do adsorvente – Zeólita, e Polarizabilidade (cm^3).

Conforme pode-se observar pela Figura 37, que apresenta a variação no R^2 pela permutação das variáveis de entrada do modelo, semelhantemente ao resultado obtido pelo modelo KNN, as variáveis Área BET (0,88) e Pressão (0,80) estão entre as variáveis mais importantes para a previsão do modelo, além disso verifica-se que neste modelo de AD as propriedades do gás, Massa molar (0,75) e Diâmetro cinético (0,51), também apresentam impacto significativo.

Figura 37 - Importâncias de permutação das variáveis para o modelo AD otimizado após remoção das seguintes variáveis: Técnica de medida - Método gravimétrico, Técnica de medida - Método volumétrico, Classe do adsorvente - MOF, Classe do adsorvente – Zeólita, Classe do adsorvente – Carvão ativado e Polarizabilidade (cm³).



Fonte: Autoria própria (2023).

4.5. Regressão com RVS

Conforme métricas apresentadas na Tabela 12, os resultados obtidos a partir do modelo *Support Vector Regressor* (SVR) com kernel RBF demonstraram um desempenho bastante sólido na tarefa de previsão da capacidade de adsorção de gases leves, apesar de ter apresentado um desempenho inferior aos modelos pelos algoritmos KNN e AD. O valor de R^2, que se aproxima de 0,86 tanto no conjunto de treinamento quanto no conjunto de teste, sugere que o modelo é capaz de explicar uma parcela significativa da variação nos dados de teste. Além disso, os baixos valores de MAE (0,356) e RMSE (0,624) no conjunto de teste indicam que as previsões do modelo estão, em média, próximas dos valores reais com uma margem de erro reduzida. A estabilidade entre os resultados de treinamento e teste sugere que o modelo não está sofrendo de superajuste, o que é uma indicação positiva de sua capacidade de generalização. A Figura 38 atesta o sucesso do modelo ao utilizar o kernel RBF, apresentado a comparação entre os valores reais e os preditos da variável alvo para o conjunto de treino. Contudo, em relação aos modelos KNN e AD, verificam-se erros de predição na região de dados intermediária.

Ainda de acordo com a Tabela 15, pode-se observar que o kernel Sigmoide não apresentou desempenho satisfatório na previsão da capacidade de adsorção, sendo este expressivamente inferior nesta tarefa quando comparado com o kernel RBF que de acordo com

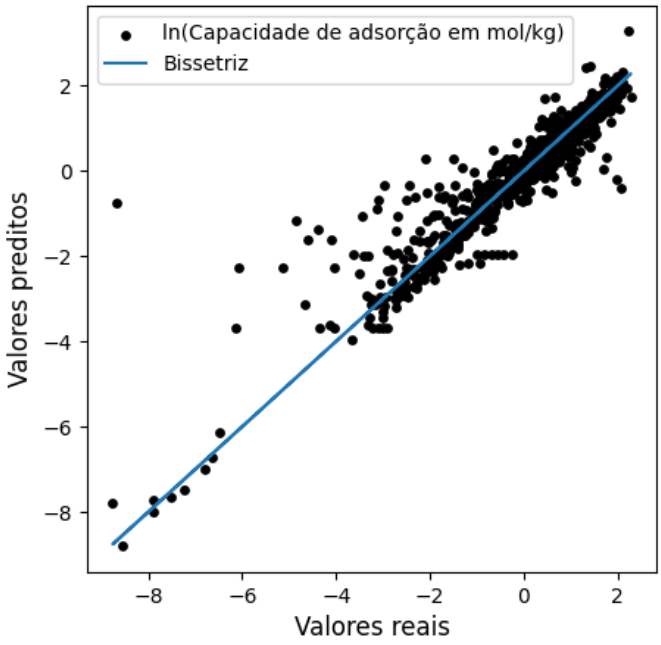
Géron (2019) e (PENG; LING, 2015) funciona bem na maioria dos casos, sendo desta forma amplamente utilizado. Destaca-se ainda que a função de kernel RBF tem demonstrado eficácia em estudos de adsorção (FATHALIAN *et al.*, 2022; KOOH *et al.*, 2022; PARVEEN; ZAIDI; DANISH, 2019).

Tabela 15 - Coeficiente de determinação (R^2), erro quadrático médio (MAE) e a raiz do erro quadrático médio (RMSE) para os conjuntos de treino e teste e parametrização pelo SVR com kernel RBF e sigmoide.

Métrica	**Treino – Validação cruzada**		**Teste**	
	RBF	**Sigmoide**	**RBF**	**Sigmoide**
R^2 [-]	0,861	0,296	0,857	0,250
MAE [kg/mol]	0,366	0,945	0,356	0,969
RMSE [kg/mol]	0,616	1,403	0,624	1,428

Fonte: Autoria própria (2023).

Figura 38 - Variável alvo (capacidade de adsorção em mol/kg): valores reais versus valores preditos pelo algoritmo SVR com kernel RBF para o conjunto de teste.

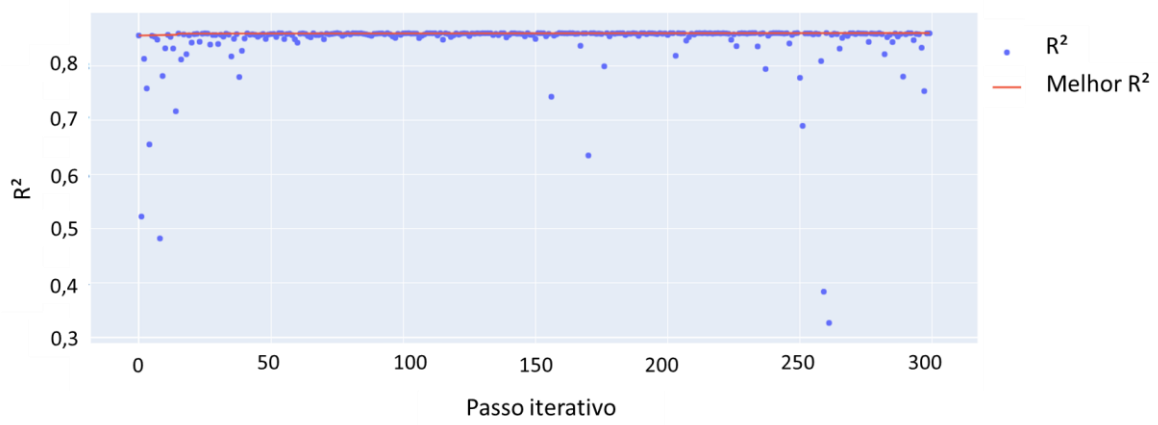


Fonte: Autoria própria (2023).

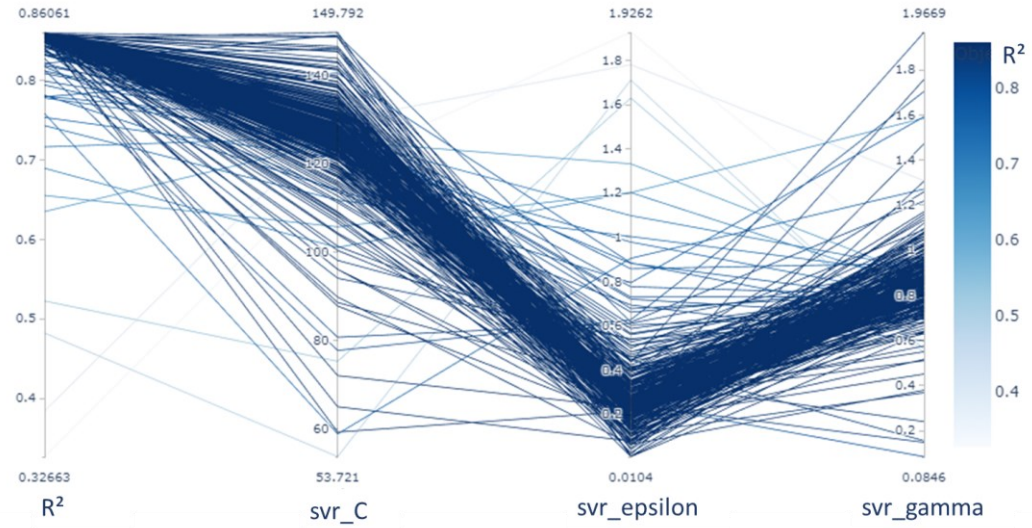
Assim como para os outros modelos, o modelo SVR com função kernel RBF foi obtido a partir da otimização de seus hiperparâmetros utilizando a biblioteca Optuna com 300 passos iterativos, sendo esse valor suficiente para convergência de busca pelos hiperparâmetros ótimos

conforme apresentado na Figura 39(a). A Figura 39(b) demonstra que existem diversas combinações que alcançam valores de R^2 superiores a 0,80 pela função objetivo.

Figura 39 - Processo de otimização dos hiperparâmetros do modelo SVR com kernel RBF usando R^2 como métrica da função objetivo: (a) Gráfico do histórico de otimização; (b) Gráfico de coordenadas paralelas



(a)



(b)

Fonte: Autoria própria (2023).

As faixas de busca dos hiperparâmetros, bem como os valores otimizados resultantes da busca pela ferramenta Optuna, são detalhados na Tabela 16. Uma análise de importância dos hiperparâmetros, conforme indicado na Tabela 17, revelou que o parâmetro de Largura da Faixa de Apoio (*svr__epsilon*), responsável por determinar a extensão na qual os erros são desconsiderados, exerce uma influência notável na busca pelos melhores hiperparâmetros, com uma significativa importância relativa de 98,5%. Isso enfatiza que mesmo pequenas variações

nesse parâmetro têm o potencial de causar um impacto substancial nas previsões geradas pelo modelo.

Tabela 16 - Melhores hiperparâmetros para o modelo de SVR com função kernel RBF obtidos pela ferramenta Optuna

Hiperparâmetro	**Faixa de busca**	**Valor ótimo**
Parâmetro de regularização (*svr__C*)	50 – 150	125,7
Largura da faixa de apoio (*svr__epsilon*)	0,01 – 2	0,250
Parâmetro de suavidade (*svr__gamma*)	0,001 - 2	0,840

Fonte: Autoria própria (2023).

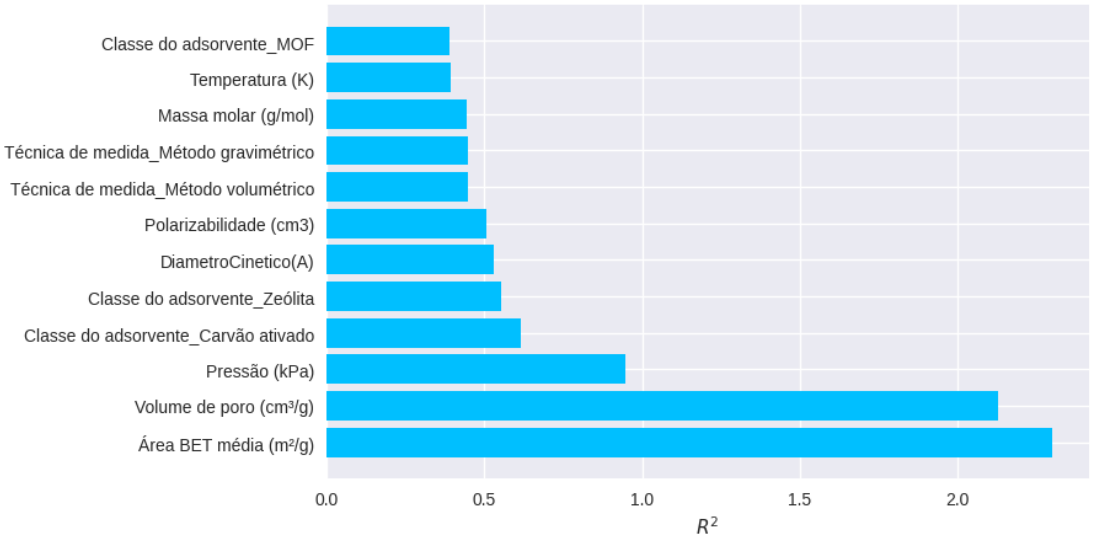
Tabela 17 - Análise de importância dos hiperparâmetros do modelo SVR com kernel RBF

Hiperparâmetro	**Importância relativa**
Parâmetro de regularização (*svr__C*)	0,006
Largura da faixa de apoio (*svr__epsilon*)	0,985
Parâmetro de suavidade (*svr__gamma*)	0,009

Fonte: Autoria própria (2023).

A análise de importância das variáveis utilizando o método de permutação apresentada na Figura 40 demonstrou os fatores que mais influenciam a previsão da capacidade de adsorção de gases leves no modelo SVR com kernel RBF. Os resultados destacam que as características Área BET média (m^2/g) e Volume de poro (cm^3/g) desempenham papéis cruciais, com variações de 2,30 e 2,13 respectivamente tendo como base a métrica R^2, indicando que variações nessas propriedades têm um impacto significativo nas previsões do modelo. Além disso, a variável Pressão (kPa) – 0,95 – também desempenha papel relevante.

Figura 40 - Importâncias de permutação das variáveis para o modelo SVR com kernel RBF otimizado.



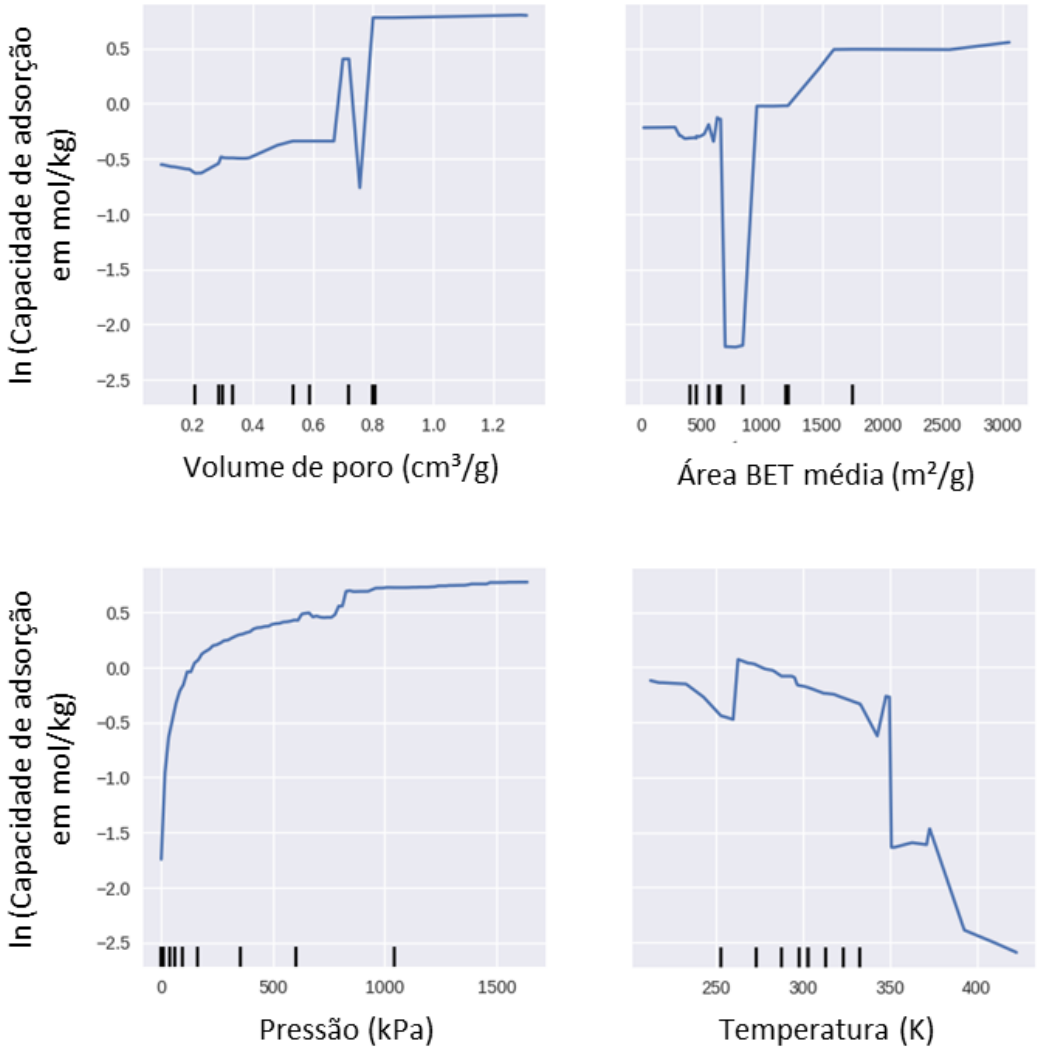
Fonte: Autoria própria (2023).

4.6. Interpretabilidade do modelo de AD usando Gráfico de Dependência Parcial

Todos os modelos - KNN, AD e SVR com kernel RBF - apresentaram métricas satisfatórias tanto para o conjunto de treino quanto para o de teste. Entretanto, notou-se que os modelos KNN e AD demonstraram um desempenho superior em comparação com o SVR. Dado o desempenho semelhante dos modelos KNN e AD, optou-se por selecionar apenas o modelo AD para a análise de interpretabilidade por meio da ferramenta PDP (*Partial Dependence Plot*).

A relação entre a variável alvo, que é o logaritmo natural da capacidade de adsorção em mol/kg, e as característica “Volume de Poro (cm³/g)”, “Área BET média (m²/g)”, “Temperatura (K)” e “Pressão (kPa)”, foram avaliadas usando a ferramenta PDP, conforme mostrado na Figura 41.

Figura 41 - Gráficos de Dependência Parcial para o modelo AD.

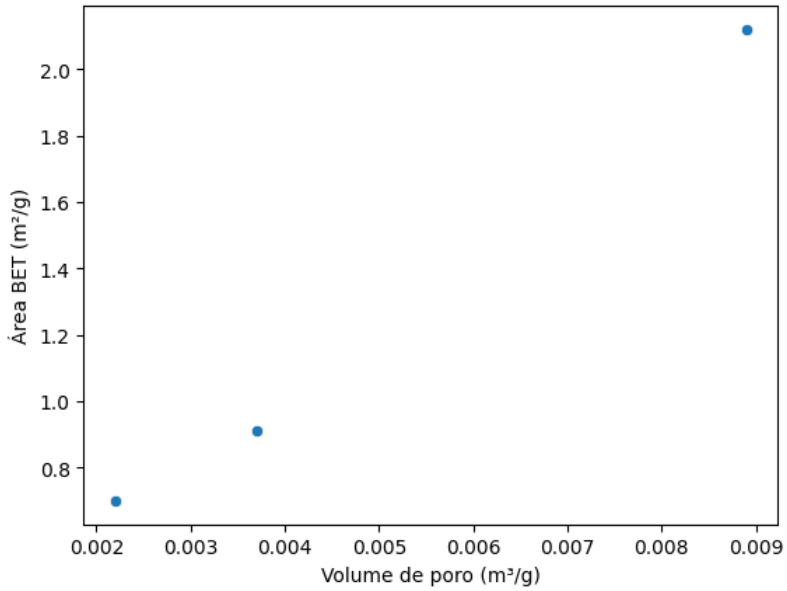


Fonte: Autoria própria (2023).

De maneira geral, de acordo com a Figura 41, observa-se aumento na variável alvo à medida que o Volume de Poro e a Área BET média aumentam. Domingues (2005) destaca que as propriedades dos adsorventes – área superficial específica, porosidade, volume de poros, distribuição do tamanho de poros, grupos funcionais presentes na superfície do adsorvente e natureza do material precursor – são fatores determinantes para a capacidade e a taxa de adsorção. Um maior volume de poro proporciona mais sítios de adsorção disponíveis para as moléculas do adsorvato, resultando, em geral, em uma maior quantidade total de adsorção (DO, D. D. et al., 2010). O estudo de Seifi et al. (2023) demonstra um aumento da área superficial à medida que o volume de poro também aumenta ao avaliar três tipos de materiais (Figura 42). Como a adsorção é o fenômeno de superfície, a capacidade de adsorção é proporcional à área

superficial específica, que representa à superfície total ativa por unidade de massa de adsorvente (DOMINGUES, 2005).

Figura 42 - Relação entre volume de poro e área superficial específica pelo método BET.



Fonte: Adaptado de Seifi et al. (2023).

Ainda de acordo com a Figura 41, destaca-se que, para pressões mais baixas, o aumento dessa variável contribui significativamente para o aumento da capacidade de adsorção, atingindo um ponto em que a influência da pressão se torna menos expressiva. Este comportamento foi evidenciado no estudo de Kang et al. (2020) em que os resultados demonstram que a capacidade de adsorção do CO_2 aumentou com o aumento da pressão de equilíbrio de adsorção e aumentou rapidamente a baixa pressão. Esse resultado é explicado pelo fato da maioria dos sólidos adsorventes possuir superfícies heterogêneas, o que resulta em variações nas energias de adsorção. Desta forma, os sítios de adsorção são ocupados de forma sequencial, começando pelos sítios de maior energia e avançando em direção aos sítios de menor energia à medida que a pressão parcial atinge a saturação (CHIOU, 2003).

A análise da Figura 41 revela que a capacidade de adsorção apresenta uma tendência decrescente com o aumento da temperatura. Este comportamento está em consonância com as observações de Chen; Jin e Chen (2011) e Horsfall Jnr e Spiff (2005), que afirmam que, de acordo com a teoria da adsorção, o aumento da temperatura geralmente resulta na diminuição da adsorção e as moléculas previamente adsorvidas em uma superfície tendem a dessorver da superfície a temperaturas elevadas. No entanto, Horsfall Jnr e Spiff (2005) destacam que em alguns cenários ocorre o fenômeno oposto, onde temperaturas mais altas induzem à diminuição

da viscosidade, facilitando a adsorção de moléculas nos poros e resultando em aumento da capacidade de adsorção.

Vale ressaltar que a relação identificada entre a capacidade de adsorção, representada no presente trabalho pelo logaritmo natural dessa variável, e a temperatura foi corroborada pelo estudo de Kang et al. (2020). Este estudo constatou que a capacidade de adsorção de CO_2 em caulinita diminuiu com o aumento da temperatura. Resultados semelhantes foram observados por Chen, Jin e Chen (2011) em seu estudo sobre a capacidade de adsorção de CH_4 em carvão, conduzido na faixa de temperatura de 20 a 60°C e pressão inferior a 6 MPa.

5. CONCLUSÃO

Neste estudo, buscou-se prever a capacidade de adsorção de gases leves em uma ampla variedade de sólidos, conduzindo uma investigação abrangente fundamentada em técnicas avançadas de aprendizado de máquina e análise estatística. A construção de um banco de dados consistente, contendo 3.339 pontos e contemplando a adsorção de diversos gases leves em uma variedade de materiais adsorventes a partir de 22 fontes de dados abrangendo o período de 1974 a 2022, foi um passo crucial, revelando a riqueza da amostragem, uma vez que foi observada heterogeneidade na distribuição dos dados entre diferentes tipos de sólidos e gases.

A aplicação de três modelos distintos de aprendizado de máquina - KNN, Árvore de Decisão (AD) e Regressão por Vetores de Suporte (RVS) com kernel RBF -proporcionou uma predição satisfatória da capacidade de adsorção de gases leves em diferentes tipos de adsorventes. Contudo, o modelo RVS com kernel sigmoide não se mostrou adequado para esta tarefa. Entre esses modelos, notamos que os modelos AD e KNN emergiram como os mais eficazes na previsão da capacidade de adsorção, apresentando desempenhos notáveis.

Além da predição, explorou-se a contribuição de cada variável na predição da capacidade de adsorção por meio da análise de importância das variáveis por permutação. Nessa análise, para todos os modelos – KNN, AD e SVR com kernel RBF –, um padrão consistente se destacou: as variáveis Área BET média e Pressão foram consistentemente identificadas como importantes na predição da capacidade de adsorção, apontando para sua influência significativa no processo.

Uma análise mais profunda foi conduzida no modelo AD, empregando o Gráfico de Dependência Parcial (PDP), revelando comportamentos específicos: observou-se, conforme esperado pela literatura, que a capacidade de adsorção tende a diminuir com o aumento da temperatura, mas aumenta em resposta ao incremento da pressão, volume de poro e Área BET média.

O alinhamento entre os resultados empíricos e as expectativas teóricas fortalece a validade do modelo, sugerindo uma compreensão robusta do fenômeno estudado. Além da predição acurada, a interpretabilidade dos modelos propostos enriquece a análise, fornecendo *insights* valiosos sobre os fatores críticos na capacidade de adsorção. Estes resultados destacam o potencial promissor da técnica de *Machine Learning* para conduzir triagens rápidas na escolha

de adsorventes promissores em aplicações específicas, contribuindo para a redução de custos e tempo associados a análises laboratoriais. Este estudo não apenas contribui para o avanço do conhecimento nessa área, mas também oferece ferramentas práticas para a previsão e compreensão mais profunda da capacidade de adsorção em contextos diversos.

6. SUGESTÕES PARA TRABALHOS FUTUROS

Para direcionar pesquisas subsequentes, propõem-se abordagens capazes de aprofundar e enriquecer os conhecimentos adquiridos neste trabalho:

- Inclusão de variáveis adicionais: Avaliar a inclusão de variáveis adicionais, como propriedades específicas dos adsorventes e dos gases, que possam impactar a capacidade de adsorção. Essa análise visa ampliar o escopo do estudo, explorando potenciais impactos e verificando se contribuirá para elevar a precisão dos modelos.
- Comparação dos modelos desenvolvidos nesse trabalho para predição da capacidade de adsorção com métodos tradicionais: realizar comparações com métodos tradicionais de previsão de capacidade de adsorção pode enriquecer a discussão e fornecer uma visão mais holística. Essa comparação permitirá contextualizar a eficácia dos modelos de aprendizagem de máquina, destacando suas vantagens e limitações em relação às práticas convencionais.
- Testar a robustez dos modelos em condições não exploradas: A exposição a condições não previamente consideradas pode revelar limitações do modelo, possibilitando ajustes e melhorias para torná-lo mais abrangente e preciso. Ao assegurar que o modelo é seguro em condições diversas, aumenta-se a confiança em sua aplicabilidade prática, possibilitando a utilização em uma variedade de contextos industriais e ambientais. Desta forma, avaliar se o modelo pode ser extrapolável possibilita uma compreensão mais abrangente e realista da eficácia do modelo, fortalecendo sua relevância e aplicabilidade em diversas situações.
- Explorar a aplicabilidade dos modelos em processos de separação de gases: utilizar as previsões dos modelos para estimar a capacidade de adsorção de diferentes gases e aplicar a equação de seletividade para avaliar a eficiência de separação, contribuindo para estratégias mais eficientes em processos de separação de gases.

REFERÊNCIAS BIBLIOGRÁFICAS

ABU ALFEILAT, H. A. *et al.* Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review. *Big Data*, v. 7, n. 4, p. 221–248, 1 dez. 2019. https://doi.org/10.1089/big.2018.0175.

AHMADI AZQHANDI, M. H. *et al.* Application of random forest, radial basis function neural networks and central composite design for modeling and/or optimization of the ultrasonic assisted adsorption of brilliant green on ZnS-NP-AC. *Journal of Colloid and Interface Science*, v. 505, p. 278–292, nov. 2017. https://doi.org/10.1016/j.jcis.2017.05.098.

ALAQARBEH, M. ADSORPTION PHENOMENA: DEFINITION, MECHANISMS, AND ADSORPTION TYPES: SHORT REVIEW. *Green and Applied Chemistry*, v. 13, p. 43–51, 2021.

AL-DOSARY, N. M. N.; AL-HAMED, S. A.; ABOUKARIMA, A. M. K-NEAREST NEIGHBORS METHOD FOR PREDICTION OF FUEL CONSUMPTION IN TRACTOR-CHISEL PLOW SYSTEMS. *Engenharia Agrícola*, v. 39, n. 6, p. 729–736, dez. 2019. https://doi.org/10.1590/1809-4430-eng.agric.v39n6p729-736/2019.

AL-GHOUTI, M. A.; DA'ANA, D. A. Guidelines for the use and interpretation of adsorption isotherm models: A review. *Journal of Hazardous Materials*, v. 393, p. 122383, jul. 2020. https://doi.org/10.1016/j.jhazmat.2020.122383.

ALI, N.; NEAGU, D.; TRUNDLE, P. Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets. *SN Applied Sciences*, v. 1, n. 12, p. 1559, 6 dez. 2019. https://doi.org/10.1007/s42452-019-1356-9.

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

AL-MUHTASEB, S. A. Adsorption and Desorption Equilibria of Nitrogen, Methane, Ethane, and Ethylene on Date-Pit Activated Carbon. *Journal of Chemical & Engineering Data*, v. 55, n. 1, p. 313–319, 14 jan. 2010. https://doi.org/10.1021/je900350k.

AL-MUHTASEB, S. A.; AL-RUB, F. A. A.; ZAROONI, M. AL. Adsorption Equilibria of Nitrogen, Methane, and Ethane on BDH-Activated Carbon. *J. Chem. Eng. Data*, v. 52, n. 1, p. 60–65, 2007. https://doi.org/10.1021/je060215+.

AMAR, M. N. *et al.* Modeling of methane adsorption capacity in shale gas formations using white-box supervised machine learning techniques. *Journal of Petroleum Science and Engineering*, v. 208, 1 jan. 2022. https://doi.org/10.1016/j.petrol.2021.109226.

ASHAYERI, M. *et al.* Predicting intraurban PM2.5 concentrations using enhanced machine learning approaches and incorporating human activity patterns. *Environmental Research*, v. 196, p. 110423, maio 2021. https://doi.org/10.1016/j.envres.2020.110423.

AURÉLIEN GÉRON. *Mãos à obra: aprendizado de máquina com Scikit-Learn & TensorFlow*. 1. ed. São Paulo: Alta Books, 2019.

BANDOSZ, T. J. *Activated Carbon Surfaces in Environmental Remediation*. [S.l.]: Elsevier Science, 2016. v. 7.

BANSAL, M.; GOYAL, A.; CHOUDHARY, A. A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning. *Decision Analytics Journal*, v. 3, p. 100071, jun. 2022. https://doi.org/10.1016/j.dajour.2022.100071.

BANSAL, R. C.; GOYAL, M. *Activated Carbon Adsorption*. [S.l.]: CRC Press, 2005. https://doi.org/10.1201/9781420028812.

BATTEN, S. R. *et al.* Coordination polymers, metal–organic frameworks and the need for terminology guidelines. *CrystEngComm*, v. 14, n. 9, p. 3001, 2012. https://doi.org/10.1039/c2ce06488j.

BERRY, M. W.; AZLINAH MOHAMED, M.; YAP, B. W. *Supervised and Unsupervised Learning for Data Science*. Cham: Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-22475-2.

BIRKMANN, F. *et al.* Trace Adsorption of Ethane, Propane, and *n* -Butane on Microporous Activated Carbon and Zeolite 13X at Low Temperatures. *Journal of Chemical & Engineering Data*, v. 62, n. 7, p. 1973–1982, 13 jul. 2017. https://doi.org/10.1021/acs.jced.6b01068.

BOER, D. G. *et al.* Binderless zeolite LTA beads with hierarchical porosity for selective CO2 adsorption in biogas upgrading. *Microporous and Mesoporous Materials*, v. 344, p. 112208, out. 2022. https://doi.org/10.1016/j.micromeso.2022.112208.

BORDONHOS, M. *et al.* Exploring periodic mesoporous organosilicas for ethane–ethylene adsorption–separation. *Microporous and Mesoporous Materials*, v. 317, p. 110975, abr. 2021.

BREIMAN, L. Random Forests. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001. https://doi.org/10.1016/j.micromeso.2021.110975.

BRESSERT, E. *SciPy and NumPy*. 1th. ed. Sebastopol (CA-US): O'Reilly Media, 2012.

CAO, L. Recent advances in the application of machine-learning algorithms to predict adsorption energies. *Trends in Chemistry*, v. 4, n. 4, p. 347–360, abr. 2022. https://doi.org/10.1016/j.trechm.2022.01.012.

CHAHBANI, M. H.; TONDEUR, D. Pressure drop in fixed-bed adsorbers. *Chemical Engineering Journal*, v. 81, n. 1–3, p. 23–34, jan. 2001. https://doi.org/10.1016/S1385-8947(00)00215-1.

CHARALAMBOUS, C. *et al.* Pure and Binary Adsorption of Carbon Dioxide and Nitrogen on AQSOA FAM Z02. *Journal of Chemical & Engineering Data*, v. 63, n. 3, p. 661–670, 8 mar. 2018. https://doi.org/10.1021/acs.jced.7b00864.

CHEN, S.; JIN, L.; CHEN, X. The effect and prediction of temperature on adsorption capability of coal/CH4. *Procedia Engineering*, v. 26, p. 126–131, 2011. https://doi.org/10.1016/j.proeng.2011.11.2149.

CHIH-CHENG YANG; WAN-JUI LEE; SHIE-JUE LEE. Learning of Kernel Functions in Support Vector Machines. 2006, [S.l.]: IEEE, 2006. p. 1150–1155. https://doi.org/10.1109/IJCNN.2006.246820.

CHIOU, C. T. Fundamentals of the Adsorption Theory. *Partition and Adsorption of Organic Contaminants in Environmental Systems*. [S.l.]: Wiley, 2002. p. 39–52. https://doi.org/10.1002/0471264326.ch4.

CHOI, B.-U. *et al.* Adsorption Equilibria of Methane, Ethane, Ethylene, Nitrogen, and Hydrogen onto Activated Carbon. *Journal of Chemical & Engineering Data*, v. 48, n. 3, p. 603–607, 18 mar. 2003. https://doi.org/10.1021/je020161d.

CONDON, J. B. *Surface Area and Porosity Determinations by Physisorption*. [S.l.]: Elsevier, 2006.

COST, S.; SALZBERG, S. A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features. *Machine Learning*, v. 10, n. 1, p. 57–78, 1993. https://doi.org/10.1007/BF00993481.

COSTA, E. *et al.* Equilibrium adsorption of methane, ethane, ethylene, and propylene and their mixtures on activated carbon. *Journal of Chemical & Engineering Data*, v. 34, n. 2, p. 156–160, 1 abr. 1989. https://doi.org/10.1021/je00056a003.

COSTA, V. G.; PEDREIRA, C. E. Recent advances in decision trees: an updated survey. *Artificial Intelligence Review*, v. 56, n. 5, p. 4765–4800, 10 maio 2023. https://doi.org/10.1007/s10462-022-10275-5.

CYCHOSZ, K. A.; THOMMES, M. Progress in the Physisorption Characterization of Nanoporous Gas Storage Materials. *Engineering*, v. 4, n. 4, p. 559–566, ago. 2018. https://doi.org/10.1016/j.eng.2018.06.001.

DANACI, D.; WEBLEY, P. A.; PETIT, C. Guidelines for Techno-Economic Analysis of Adsorption Processes. *Frontiers in Chemical Engineering*, v. 2, 20 jan. 2021. https://doi.org/10.3389/fceng.2020.602430.

DANNER, R.; CHOI, E. C. CORRECTION- Mixture Adsorption Equilibria of Ethane and Ethylene on 13X Molecular Sieves. *Industrial & Engineering Chemistry Fundamentals*, v. 18, n. 3, p. 300–300, 27 ago. 1979. https://doi.org/10.1021/i160071a600.

DEO, T. Y.; SANJU, A. Data imputation and comparison of custom ensemble models with existing libraries like XGBoost, CATBoost, AdaBoost and Scikit learn for predictive equipment failure. *Materials Today: Proceedings*, v. 72, p. 1596–1604, 2023. https://doi.org/10.1016/j.matpr.2022.09.410.

DEVORE, J. L. *Probability and Statistics for Engineering and the Sciences*. 8. ed. Boston: Cengage Learning, 2015.

DEZA, E.; DEZA, M. M. *Encyclopedia of Distances*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. https://doi.org/10.1007/978-3-642-00234-2.

DIETTERICH, T. Overfitting and undercomputing in machine learning. *ACM Computing Surveys*, v. 27, n. 3, p. 326–327, set. 1995. https://doi.org/10.1145/212094.212114.

DING, J. *et al.* Machine learning for molecular thermodynamics. *Chinese Journal of Chemical Engineering*, jan. 2021. https://doi.org/10.1016/j.cjche.2020.10.044.

DO, D. D. *et al.* The role of accessibility in the characterization of porous solids and their adsorption properties. *Adsorption*, v. 16, n. 1–2, p. 3–15, 1 jun. 2010. https://doi.org/10.1007/s10450-009-9203-8.

DO, DUONG D. *Adsorption analysis: equilibria and kinetics*. London: Imperial College Press, 1998. https://doi.org/10.1142/9781860943829.

DOBBELAERE, M. R. *et al.* Machine Learning in Chemical Engineering: Strengths, Weaknesses, Opportunities, and Threats. *Engineering*, jul. 2021. https://doi.org/10.1016/j.eng.2021.03.019.

DOMINGUES, V. M. F. *Utilização de um produto natural (cortiça) como adsorvente de pesticidas piretróides em águas*. 2005. 1–224 f. Faculdade de Engenharia da Universidade do Porto, Porto, 2005.

DREISBACH, F.; SEIF, R.; LÖSCH, H. W. Adsorption equilibria of CO/H2 with a magnetic suspension balance: Purely gravimetric measurement. *Journal of Thermal Analysis and Calorimetry*, v. 71, n. 1, p. 73–82, 2003. https://doi.org/10.1023/A:1022206031461.

DUDANI, S. A. The Distance-Weighted k-Nearest-Neighbor Rule. *IEEE Transactions on Systems, Man, and Cybernetics*, v. SMC-6, n. 4, p. 325–327, abr. 1976. https://doi.org/10.1109/TSMC.1976.5408784.

EHSANI, R.; DRABLØS, F. Robust Distance Measures for KNN Classification of Cancer Data. *Cancer Informatics*, v. 19, p. 117693512096554, 13 jan. 2020. https://doi.org/10.1177/1176935120965542.

ESTERHUIZEN, J. A.; GOLDSMITH, B. R.; LINIC, S. Interpretable machine learning for knowledge generation in heterogeneous catalysis. *Nature Catalysis*, v. 5, n. 3, p. 175–184, 17 mar. 2022. https://doi.org/10.1038/s41929-022-00744-z.

FACELI, K. *et al. Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina*. 1th. ed. São Paulo: LTC, 2011.

FAN, G.-F. *et al.* Application of the Weighted K-Nearest Neighbor Algorithm for Short-Term Load Forecasting. *Energies*, v. 12, n. 5, 9 mar. 2019. https://doi.org/10.3390/en12050916.

FATHALIAN, F. *et al.* Intelligent prediction models based on machine learning for CO2 capture performance by graphene oxide-based adsorbents. *Scientific Reports*, v. 12, n. 1, p. 21507, 13 dez. 2022. https://doi.org/10.1038/s41598-022-26138-6.

FERRERO, C. A. *Algoritmo kNN para previsão de dados temporais: funções de previsão e critérios de seleção de vizinhos próximos aplicados a variáveis ambientais em limnologia*. 2009. Universidade de São Paulo, São Carlos, 2009.

FLANIGEN, E. M. Zeolites and molecular sieves: An historical perspective. [S.l: s.n.], 2001. p. 11–35. https://doi.org/10.1016/S0167-2991(01)80243-3.

FRANÇOIS-LAVET, V. *et al.* An Introduction to Deep Reinforcement Learning. *Foundations and Trends® in Machine Learning*, v. 11, n. 3–4, p. 219–354, 2018. https://doi.org/10.1561/2200000071.

FREEMAN, J. J.; MCLEOD, A. I. Nitrogen BET surface area measurement as a fingerprint method for the estimation of pore volume in active carbons. *Fuel*, v. 62, n. 9, p. 1090–1091, set. 1983. https://doi.org/10.1016/0016-2361(83)90147-3.

GERON, A. *Mãos à obra: aprendizado de máquina com Scikit-Learn & TensorFlow*. 1ª ed. São Paulo: Alta Books, 2019.

GHANEKAR, P. G.; DESHPANDE, S.; GREELEY, J. Adsorbate chemical environment-based machine learning framework for heterogeneous catalysis. *Nature Communications*, v. 13, n. 1, p. 5788, 2 out. 2022. https://doi.org/10.1038/s41467-022-33256-2.

GOLDEN, T. C.; SIRCAR, S. Gas Adsorption on Silicalite. *Journal of Colloid and Interface Science*, v. 162, n. 1, p. 182–188, jan. 1994. https://doi.org/10.1006/jcis.1994.1023.

GOLIPOUR, H. *et al.* Experimental Measurement for Adsorption of Ethylene and Ethane Gases on Copper-Exchanged Zeolites 13X and 5A. *Journal of Chemical & Engineering Data*, v. 65, n. 8, p. 3920–3932, 13 ago. 2020. https://doi.org/10.1021/acs.jced.0c00251.

HARRISON, M. *Learning the pandas library*. 1th. ed. Scotts Valley (CA-US): CreateSpace Publishing, 2016.

HECHENBICHLER, K.; SCHLIEP, K. *Weighted k-Nearest-Neighbor Techniques and Ordinal Classification*, 2004.

HORSFALL JNR, M.; SPIFF, A. I. Effects of temperature on the sorption of Pb2+ and Cd2+ from aqueous solution by Caladium bicolor (Wild Cocoyam) biomass. *Electronic Journal of Biotechnology*, v. 8, n. 2, p. 162–169, 15 ago. 2005. https://doi.org/10.2225/vol8-issue2-fulltext-4.

HUANG, M. *et al.* Fast prediction of methane adsorption in shale nanopores using kinetic theory and machine learning algorithm. *Chemical Engineering Journal*, v. 446, p. 137221, out. 2022. https://doi.org/10.1016/j.cej.2022.137221.

IBRAHIM, S. A. B. Synthesis and characterization of zeolites from sodium aluminosilicate solution. 2007. 1–42 f. Universiti Sains Malaysia, Penang, 2007. Disponível em: <https://core.ac.uk/download/pdf/11933197.pdf>. Acesso em: 3 set. 2023.

IDRIS, I. *Learning NumPy Array*. 1th. ed. BIRMINGHAM: Packt Publishing, 2014.

INGLEZAKIS, V.; POULOPOULOS, S. *Adsorption, Ion Exchange and Catalysis*. New York: Elsevier, 2006. https://doi.org/10.1016/B978-044452783-7/50002-1.

JÄGER, M. O. J. *et al.* Machine learning hydrogen adsorption on nanoclusters through structural descriptors. *npj Computational Materials*, v. 4, n. 1, p. 37, 19 jul. 2018. https://doi.org/10.1038/s41524-018-0096-5.

JOHNSON, P. M. *et al.* Using machine learning to examine freight network spatial vulnerabilities to disasters: A new take on partial dependence plots. *Transportation Research Interdisciplinary Perspectives*, v. 14, p. 100617, jun. 2022. https://doi.org/10.1016/j.trip.2022.100617.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. *Science*, v. 349, n. 6245, p. 255–260, 17 jul. 2015. https://doi.org/10.1126/science.aaa8415.

KANG, G. *et al.* Effect of Pressure and Temperature on CO2/CH4 Competitive Adsorption on Kaolinite by Monte Carlo Simulations. *Materials*, v. 13, n. 12, p. 2851, 25 jun. 2020. https://doi.org/10.3390/ma13122851.

KARIMI, M.; RODRIGUES, A. E.; SILVA, J. A. C. Designing a simple volumetric apparatus for measuring gas adsorption equilibria and kinetics of sorption. Application and validation for CO2, CH4 and N2 adsorption in binder-free beads of 4A zeolite. *Chemical Engineering Journal*, v. 425, p. 130538, dez. 2021. https://doi.org/10.1016/j.cej.2021.130538.

KARIMI, S.; TAVAKKOLI YARAKI, M.; KARRI, R. R. *A comprehensive review of the adsorption mechanisms and factors influencing the adsorption process from the perspective of bioethanol dehydration*. *Renewable and Sustainable Energy Reviews*. [S.l.]: Elsevier Ltd. , 1 jun. 2019. https://doi.org/10.1016/j.rser.2019.03.025.

KHORAMZADEH, E.; MOFARAHI, M.; LEE, C.-H. Equilibrium Adsorption Study of CO2 and N2 on Synthesized Zeolites 13X, 4A, 5A, and Beta. *Journal of Chemical & Engineering Data*, v. 64, n. 12, p. 5648–5664, 12 dez. 2019. https://doi.org/10.1021/acs.jced.9b00690.

KIANFAR, E. Zeolites: Properties, Applications, Modification and Selectivity. In: MAHLER, A. (Org.). *Zeolites: Advances in Research and Applications*. New York: Nova Science Publishers, 2020. p. 1–22.

KISELEV, A. V.; LOPATKIN, A. A.; SHULGA, A. A. Molecular statistical calculation of gas adsorption by silicalite. *Zeolites*, v. 5, n. 4, p. 261–267, jul. 1985. https://doi.org/10.1016/0144-2449(85)90098-3.

KITCHIN, J. R. Machine learning in catalysis. *Nature Catalysis*, v. 1, n. 4, p. 230–232, 16 abr. 2018. Disponível em: <http://www.nature.com/articles/s41929-018-0056-y>. https://doi.org/10.1038/s41929-018-0056-y.

KLOUTSE, F. A. *et al.* Experimental benchmark data of CH4, CO2 and N2 binary and ternary mixtures adsorption on MOF-5. *Separation and Purification Technology*, v. 197, p. 228–236, maio 2018. https://doi.org/10.1016/j.seppur.2018.01.013.

KOOH, M. R. R. *et al.* Machine learning approaches to predict adsorption capacity of Azolla pinnata in the removal of methylene blue. *Journal of the Taiwan Institute of Chemical Engineers*, v. 132, p. 104134, mar. 2022. https://doi.org/10.1016/j.jtice.2021.11.001.

KOOPAL, L.; TAN, W.; AVENA, M. Equilibrium mono- and multicomponent adsorption models: From homogeneous ideal to heterogeneous non-ideal binding. *Advances in Colloid and Interface Science*, v. 280, p. 102138, jun. 2020. https://doi.org/10.1016/j.cis.2020.102138.

KOTSIANTIS, S. B. Decision trees: a recent overview. *Artificial Intelligence Review*, v. 39, n. 4, p. 261–283, 29 abr. 2013. https://doi.org/10.1007/s10462-011-9272-4.

KOTSIANTIS, S. B.; ZAHARAKIS, I. D.; PINTELAS, P. E. Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, v. 26, n. 3, p. 159–190, 10 nov. 2006. https://doi.org/10.1007/s10462-007-9052-3.

KULPRATHIPANJA, S. *Zeolites in Industrial Separation and Catalysis*. [S.l.]: Wiley, 2010.

LANE, D. M. *Log Transformations*. Disponível em: <https://onlinestatbook.com/2/transformations/log.html>. Acesso em: 12 set. 2023. https://doi.org/10.1002/9783527629565.

LASSINANTTI, M. *Synthesis, Characterization and properties of zeolites filmes and membranes*. 2001. 1–32 f. PhD – Luleå University of Technology, Luleå, 2001. Disponível em: <https://www.diva-portal.org/smash/get/diva2:991559/FULLTEXT01.pdf>. Acesso em: 3 set. 2023.

LEEKHA, G. *Learn AI with Python*. 1th. ed. Noida: BPB Publications, 2022.

LI, W. *et al.* Implementation of AdaBoost and genetic algorithm machine learning models in prediction of adsorption capacity of nanocomposite materials. *Journal of Molecular Liquids*, v. 350, p. 118527, mar. 2022. https://doi.org/10.1016/j.molliq.2022.118527.

LOUKHOVITSKI, B. I.; SHARIPOV, A. S. Molecular Collision Diameters and Electronic Polarizabilities: Inherent Relationship and Fast Evaluation. *The Journal of Physical Chemistry A*, v. 125, n. 23, p. 5117–5123, 17 jun. 2021. https://doi.org/10.1021/acs.jpca.1c02201.

MAJD, M. M. *et al.* Adsorption isotherm models: A comprehensive and systematic review (2010−2020). *Science of The Total Environment*, v. 812, p. 151334, mar. 2022. https://doi.org/10.1016/j.scitotenv.2021.151334.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2009. https://doi.org/10.1017/CBO9780511809071.

MARSH, H. *Activated carbon compendium*. London: Elsevier, 2001.

MARSH, H.; RODRÍGUEZ-REINOSO, F. *Activated Carbon*. 1. ed. [S.l.]: Elsevier Science & Technology Books, 2006. https://doi.org/10.1016/B978-008044463-5/50015-7.

MARTINS, V. F. D. *et al.* Ethane/ethylene separation on a copper benzene-1,3,5-tricarboxylate MOF. *Separation and Purification Technology*, v. 149, p. 445–456, jul. 2015. https://doi.org/10.1016/j.seppur.2015.06.012.

MASOUD JAHANDAR, L. *et al.* Effect of the adsorbate kinetic diameter on the accuracy of the Dubinin–Radushkevich equation for modeling adsorption of organic vapors on activated carbon. *Journal of Hazardous Materials*, v. 241–242, p. 154–163, nov. 2012. https://doi.org/10.1016/j.jhazmat.2012.09.024.

MCEWEN, J.; HAYMAN, J.-D.; OZGUR YAZAYDIN, A. A comparative study of CO2, CH4 and N2 adsorption in ZIF-8, Zeolite-13X and BPL activated carbon. *Chemical Physics*, v. 412, p. 72–76, fev. 2013. https://doi.org/10.1016/j.chemphys.2012.12.012.

MEEK, S. T. *et al.* Effects of Polarizability on the Adsorption of Noble Gases at Low Pressures in Monohalogenated Isoreticular Metal–Organic Frameworks. *The Journal of Physical Chemistry C*, v. 116, n. 37, p. 19765–19772, 20 set. 2012. https://doi.org/10.1021/jp303274m.

MENG, M.; ZHONG, R.; WEI, Z. Prediction of methane adsorption in shale: Classical models and machine learning based models. *Fuel*, v. 278, p. 118358, out. 2020. https://doi.org/10.1016/j.fuel.2020.118358.

MISHRA, P.; MEKALA, S.; *et al.* Adsorption of CO2, CO, CH4 and N2 on a zinc based metal organic framework. *Separation and Purification Technology*, v. 94, p. 124–130, jun. 2012. https://doi.org/10.1016/j.seppur.2011.09.041.

MISHRA, P.; UPPARA, H. P.; *et al.* Adsorption of Lower Alkanes on a Zinc Based Metal Organic Framework. *Journal of Chemical & Engineering Data*, v. 57, n. 9, p. 2610–2613, 13 set. 2012. https://doi.org/10.1021/je3007265.

MITCHELL, T. M. *Machine Learning*. 1th. ed. London: McGraw-Hill, 1997.

MOLNAR, C. *Interpretable Machine Learning*. 1th. ed. Morrisville: Lulu.com, 2020.

MOLNAR, C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2. ed. [S.l: s.n.], 2022.

MOREIRA, M. A. *et al.* Adsorption Equilibrium of Carbon Dioxide, Methane, Nitrogen, Carbon Monoxide, and Hydrogen on UiO-66(Zr)_$(COOH)_2$. *Journal of Chemical & Engineering Data*, v. 64, n. 11, p. 4724–4732, 14 nov. 2019. https://doi.org/10.1021/acs.jced.9b00053.

MORGAN, P. *Data Analysis From Scratch With Python*. [S.l.]: AI Sciences LLC, 2018.

MUKHIYA, S. K.; AHMED, U. *Hands-On Exploratory Data Analysis with Python*. 1th. ed. Mumbai: Packt Publishing, 2020.

MUTTIL, N. *et al.* Production, Types, and Applications of Activated Carbon Derived from Waste Tyres: An Overview. *Applied Sciences*, v. 13, n. 1, p. 257, 25 dez. 2022. https://doi.org/10.3390/app13010257.

MYATT, G. J.; JOHNSON, W. P. *Making Sense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining*. 2nd. ed. Hoboken: Wiley, 2014. https://doi.org/10.1002/9781118422007.

NAIT AMAR, M. *et al.* Modeling of methane adsorption capacity in shale gas formations using white-box supervised machine learning techniques. *Journal of Petroleum Science and Engineering*, v. 208, p. 109226, jan. 2022. https://doi.org/10.1016/j.petrol.2021.109226.

NAKAHARA, T. Calculation of adsorption equilibria for the binary gaseous mixtures on heterogeneous surface. *Chemical Engineering Science*, v. 41, n. 1, p. 2093–2098, 1986. https://doi.org/10.1016/0009-2509(86)87126-3.

NASCIMENTO, R. F. DO *et al. Adsorção: Aspectos Teóricos e Aplicações Ambientais*. Fortaleza: Imprensa Universitária, 2014. Disponível em: <https://repositorio.ufc.br/bitstream/riufc/10267/1/2014_liv_rfdnascimento.pdf>. Acesso em: 3 set. 2023.

NETTLETON, D. Selection of Variables and Factor Derivation. *Commercial Data Mining*. [S.l.]: Elsevier, 2014. p. 79–104. https://doi.org/10.1016/B978-0-12-416602-8.00006-6.

PAN, Y. *et al.* Analysis of Influencing Factors on the Gas Separation Performance of Carbon Molecular Sieve Membrane Using Machine Learning Technique. *Membranes*, v. 12, n. 1, 17 jan. 2022. https://doi.org/10.3390/membranes12010100.

PARVEEN, N.; ZAIDI, S.; DANISH, M. Support vector regression (SVR)-based adsorption model for Ni (II) ions removal. *Groundwater for Sustainable Development*, v. 9, out. 2019. https://doi.org/10.1016/j.gsd.2019.100232.

PEDREGOSA, F. *et al.* Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

PENG, H.; LING, X. Predicting thermal–hydraulic performances in compact heat exchangers by support vector regression. *International Journal of Heat and Mass Transfer*, v. 84, p. 203–213, maio 2015. https://doi.org/10.1016/j.ijheatmasstransfer.2015.01.017.

PHAM, T. D. *et al.* Experimental and computational studies on the adsorption of CO2 and N2 on pure silica zeolites. *Microporous and Mesoporous Materials*, v. 185, p. 157–166, fev. 2014. https://doi.org/10.1016/j.micromeso.2013.10.030.

POURHAKKAK, P. *et al.* Fundamentals of adsorption technology. [S.l.]: Elsevier, 2021. p. 1–70. https://doi.org/10.1016/B978-0-12-818805-7.00001-1.

PRASAD, K. *et al.* Predicting the Adsorption Efficiency Using Machine Learning Framework on a Carbon-Activated Nanomaterial. *Adsorption Science & Technology*, v. 2023, p. 1–11, 2 jun. 2023. https://doi.org/10.1155/2023/4048676.

PRAVIN, P. S. *et al.* Hyperparameter optimization strategies for machine learning-based stochastic energy efficient scheduling in cyber-physical production systems. *Digital Chemical Engineering*, p. 100047, jul. 2022. https://doi.org/10.1016/j.dche.2022.100047.

PULLUMBI, P.; BRANDANI, F.; BRANDANI, S. Gas separation by adsorption: technological drivers and opportunities for improvement. *Current Opinion in Chemical Engineering*, v. 24, p. 131–142, jun. 2019. https://doi.org/10.1016/j.coche.2019.04.008.

RAGHAVENDRA. N, S.; DEKA, P. C. Support vector machine applications in the field of hydrology: A review. *Applied Soft Computing*, v. 19, p. 372–386, jun. 2014. https://doi.org/10.1016/j.asoc.2014.02.002.

RAJI, M. *et al.* Novel prosperous computational estimations for greenhouse gas adsorptive control by zeolites using machine learning methods. *Journal of Environmental Management*, v. 307, p. 114478, abr. 2022. https://doi.org/10.1016/j.jenvman.2022.114478.

RASCHKA, S.; MIRJALILI, V. *Python Machine Learning*. 3th. ed. [S.l.]: Packt, 2019.

REICH, R. *Adsorption on activated carbon of methane, ethane, and ethylene gases and their mixtures and carbon dioxide at 212 K, 260 K, and 301 K and up to thirty-five atmospheres*. 1974. 1–182 f. Thesis – Georgia Institute of Technology , Atlanta, GA, 1974.

RHODES, C. J. Properties and applications of Zeolites. *Science Progress*, v. 93, n. 3, p. 223–284, 1 ago. 2010. https://doi.org/10.3184/003685010X12800828155007.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "Why Should I Trust You?" 13 ago. 2016, New York, NY, USA: ACM, 13 ago. 2016. p. 1135–1144.

ROQUE-MALHERBE, R. M. A. *Adsorption and Diffusion in Nanoporous Materials*. 1th. ed. Boca Raton: CRC Press, 2007.

RUTHVEN, D. M. Fundamentals of Adsorption Equilibrium and Kinetics in Microporous Solids. In: KARGE, H. G.; WEITKAMP, J. (Org.). *Adsorption and Diffusion*. 1th. ed. New York: Springer, 2008. v. 7. p. 1–43. https://doi.org/10.1007/3829_007.

SAFAEI, M. *et al.* A review on metal-organic frameworks: Synthesis and applications. *TrAC Trends in Analytical Chemistry*, v. 118, p. 401–425, set. 2019. https://doi.org/10.1016/j.trac.2019.06.007.

SAH, S. Machine Learning: A Review of Learning Types. *Preprints.org*, 10 jul. 2020. https://doi.org/10.20944/preprints202007.0230.v1.

SANTOS-PEREIRA, J.; GRUENWALD, L.; BERNARDINO, J. Top data mining tools for the healthcare industry. *Journal of King Saud University - Computer and Information Sciences*, v. 34, n. 8, p. 4968–4982, set. 2022. https://doi.org/10.1016/j.jksuci.2021.06.002.

SCHÄF, O. *et al.* 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*, v. 259, p. 127457, nov. 2020. https://doi.org/10.1016/j.chemosphere.2020.127457.

SCHLEXER LAMOUREUX, P. *et al.* Machine Learning for Computational Heterogeneous Catalysis. *ChemCatChem*, v. 11, n. 16, p. 3581–3601, 21 ago. 2019. https://doi.org/10.1002/cctc.201900595.

SEIFI, S. *et al.* Microstructure of Dry Mortars without Cement: Specific Surface Area, Pore Size and Volume Distribution Analysis. *Applied Sciences*, v. 13, n. 9, p. 5616, 2 maio 2023. https://doi.org/10.3390/app13095616.

SHARMIN, E.; ZAFAR, F. Introductory Chapter: Metal Organic Frameworks (MOFs). *Metal-Organic Frameworks*. [S.l.]: InTech, 2016. https://doi.org/10.5772/64797.

SITU, Y. *et al.* Large-Scale Screening and Machine Learning for Metal–Organic Framework Membranes to Capture CO2 from Flue Gas. *Membranes*, v. 12, n. 7, p. 700, 11 jul. 2022. https://doi.org/10.3390/membranes12070700.

SONI, S.; BAJPAI, P. K.; ARORA, C. A review on metal-organic framework: synthesis, properties and application. *Characterization and Application of Nanomaterials*, v. 2, n. 2, 5 set. 2018.

SPEIGHT, J. G. Hydrogen Production. *Heavy Oil Recovery and Upgrading*. [S.l.]: Elsevier, 2019. p. 657–697. https://doi.org/10.1016/B978-0-12-813025-4.00015-5.

STEPANEK, H. Introduction. *Thinking in Pandas*. Berkeley, CA: Apress, 2020. p. 1–7. https://doi.org/10.1007/978-1-4842-5839-2_1.

SUMMERFIELD, M. *Programming in Python 3*. Boston: Pearson Education, 2010.

TADDA, M. A. *et al.* A Review on Activated Carbon: Process , Application and Prospects. *Journal of Advanced Civil Engineering Practice and Research*, v. 2, p. 7–13, 2016.

TANG, H. *et al.* Rapid Screening of Metal–Organic Frameworks for Propane/Propylene Separation by Synergizing Molecular Simulation and Machine Learning. *ACS Applied Materials & Interfaces*, v. 13, n. 45, p. 53454–53467, 17 nov. 2021. https://doi.org/10.1021/acsami.1c13786.

TAUNK, K. *et al.* A Brief Review of Nearest Neighbor Algorithm for Learning and Classification. maio 2019, [S.l.]: IEEE, maio 2019. p. 1255–1260. https://doi.org/10.1109/ICCS45141.2019.9065747.

THARWAT, A. Parameter investigation of support vector machine classifier with kernel functions. *Knowledge and Information Systems*, v. 61, n. 3, p. 1269–1302, 1 dez. 2019. https://doi.org/10.1007/s10115-019-01335-4.

URSUEGUÍA, D.; DÍAZ, E.; ORDÓÑEZ, S. Adsorption of methane and nitrogen on Basolite MOFs: Equilibrium and kinetic studies. *Microporous and Mesoporous Materials*, v. 298, p. 110048, maio 2020. https://doi.org/10.1016/j.micromeso.2020.110048.

VAREDA, J. P. On validity, physical meaning, mechanism insights and regression of adsorption kinetic models. *Journal of Molecular Liquids*, v. 376, p. 121416, abr. 2023. https://doi.org/10.1016/j.molliq.2023.121416.

WANG, J.; GUO, X. Adsorption isotherm models: Classification, physical meaning, application and solving method. *Chemosphere*, v. 258, p. 127279, nov. 2020. https://doi.org/10.1016/j.chemosphere.2020.127279.

WANG, S. *et al.* Insights into CO2 /N2 Selectivity in Porous Carbons from Deep Learning. *ACS Materials Letters*, v. 1, n. 5, p. 558–563, 4 nov. 2019. https://doi.org/10.1021/acsmaterialslett.9b00374.

WANG, Z. *et al.* Identification of optimal metal-organic frameworks by machine learning: Structure decomposition, feature integration, and predictive modeling. *Computers & Chemical Engineering*, v. 160, p. 107739, abr. 2022. https://doi.org/10.1016/j.compchemeng.2022.107739.

WANG, Z. F. *Data Analysis with R: Diamonds & Price Predictions*. Disponível em: <https://rstudio-pubs-static.s3.amazonaws.com/94067_d1fdfafd20b14725a2578647031760c2.html>. Acesso em: 12 set. 2023.

WORCH, E. *Adsorption Technology in Water Treatment: Fundamentals, Processes, and Modeling*. 2nd. ed. Berlim: de Gruyter, 2021. https://doi.org/10.1515/9783110715507.

XIE, C. *et al.* Explainable machine learning for carbon dioxide adsorption on porous carbon. *Journal of Environmental Chemical Engineering*, v. 11, n. 1, p. 109053, fev. 2023. https://doi.org/10.1016/j.jece.2022.109053.

XIE, J.; ZHANG, L. Machine learning and symbolic regression for adsorption of atmospheric molecules on low-dimensional TiO2. *Applied Surface Science*, v. 597, p. 153728, set. 2022. https://doi.org/10.1016/j.apsusc.2022.153728.

YAN, Y. *et al.* Adsorption behavior of metal-organic frameworks: From single simulation, high-throughput computational screening to machine learning. *Computational Materials Science*, v. 193, p. 110383, jun. 2021. https://doi.org/10.1016/j.commatsci.2021.110383.

YAN, Y. *et al.* Machine learning and in-silico screening of metal–organic frameworks for O2/N2 dynamic adsorption and separation. *Chemical Engineering Journal*, v. 427, p. 131604, jan. 2022. https://doi.org/10.1016/j.cej.2021.131604.

YANG, L.; SHAMI, A. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, v. 415, p. 295–316, nov. 2020. https://doi.org/10.1016/j.neucom.2020.07.061.

YANG, R. T. *Adsorbents: Fundamentals and Applications*. [S.l.]: Wiley, 2003. https://doi.org/10.1002/047144409X.

YUAN, Q. *et al.* Imputation of missing gas permeability data for polymer membranes using machine learning. *Journal of Membrane Science*, v. 627, p. 119207, jun. 2021. https://doi.org/10.1016/j.memsci.2021.119207.

ZHANG, W. *et al.* Modeling, optimization and understanding of adsorption process for pollutant removal via machine learning: Recent progress and future perspectives. *Chemosphere*, v. 311, p. 137044, jan. 2023. https://doi.org/10.1016/j.chemosphere.2022.137044.

ZHANG, XINYU; LIU, C.-A. Model averaging prediction by K-fold cross-validation. *Journal of Econometrics*, maio 2022. https://doi.org/10.2139/ssrn.4032249.

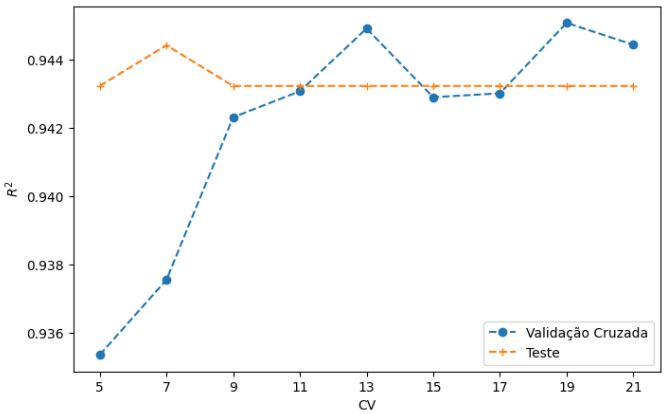
ZHANG, XUAN; ZHENG, Q.; HE, H. Machine-learning-based prediction of hydrogen adsorption capacity at varied temperatures and pressures for MOFs adsorbents. *Journal of the Taiwan Institute of Chemical Engineers*, v. 138, p. 104479, set. 2022. https://doi.org/10.1016/j.jtice.2022.104479.

ZHANG, Y. *et al.* Adsorption Equilibrium of N2, CH4, and CO2 on MIL-101. *Journal of Chemical & Engineering Data*, v. 60, n. 10, p. 2951–2957, 8 out. 2015. https://doi.org/10.1021/acs.jced.5b00327.

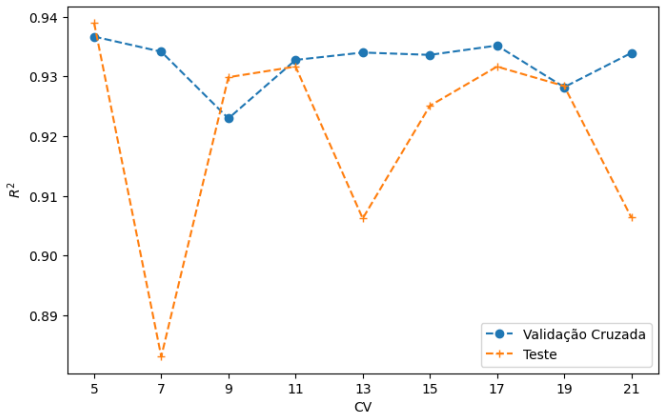
ZOBEL, J. *et al.* Efficient retrieval of partial documents. *Information Processing & Management*, v. 31, n. 3, p. 361–377, maio 1995. https://doi.org/10.1016/0306-4573(94)00052-5.

APÊNDICE A

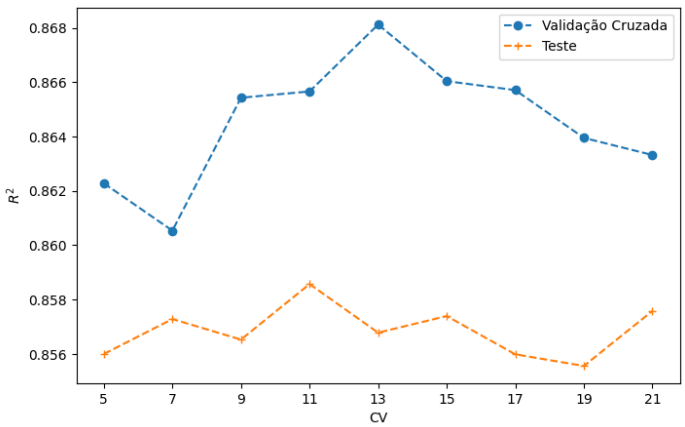
Figura A. 1. - Valores de coeficiente de determinação (R^2) dos conjuntos de treino e teste da validação cruzada para os modelos KNN (a), AD (b) e SVR (c) considerando diferentes valores para os *k-folds* (CV). Resultados obtidos após a otimização de demais parâmetros pela biblioteca Optuna.



(a)



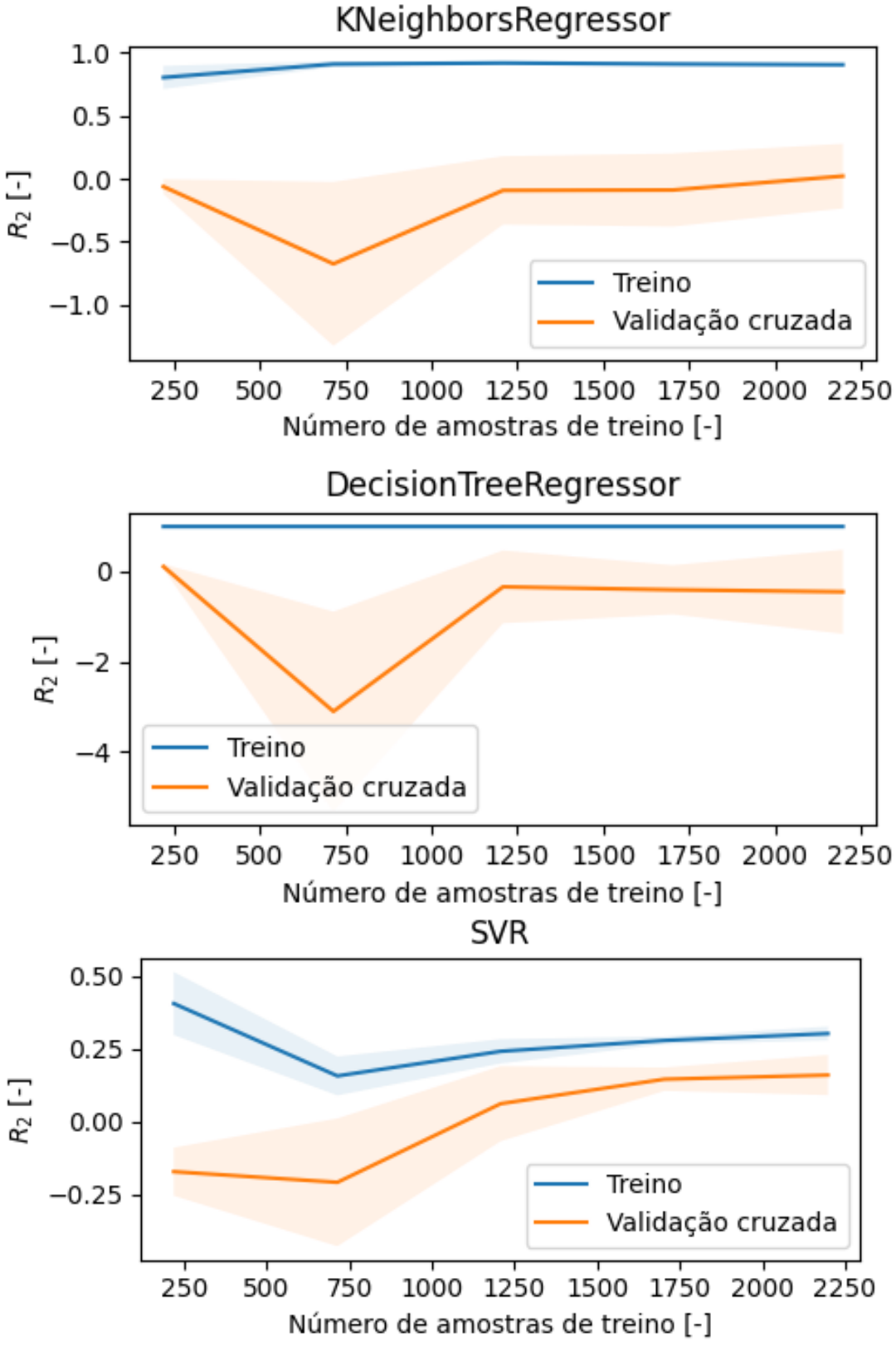
(b)



(c)

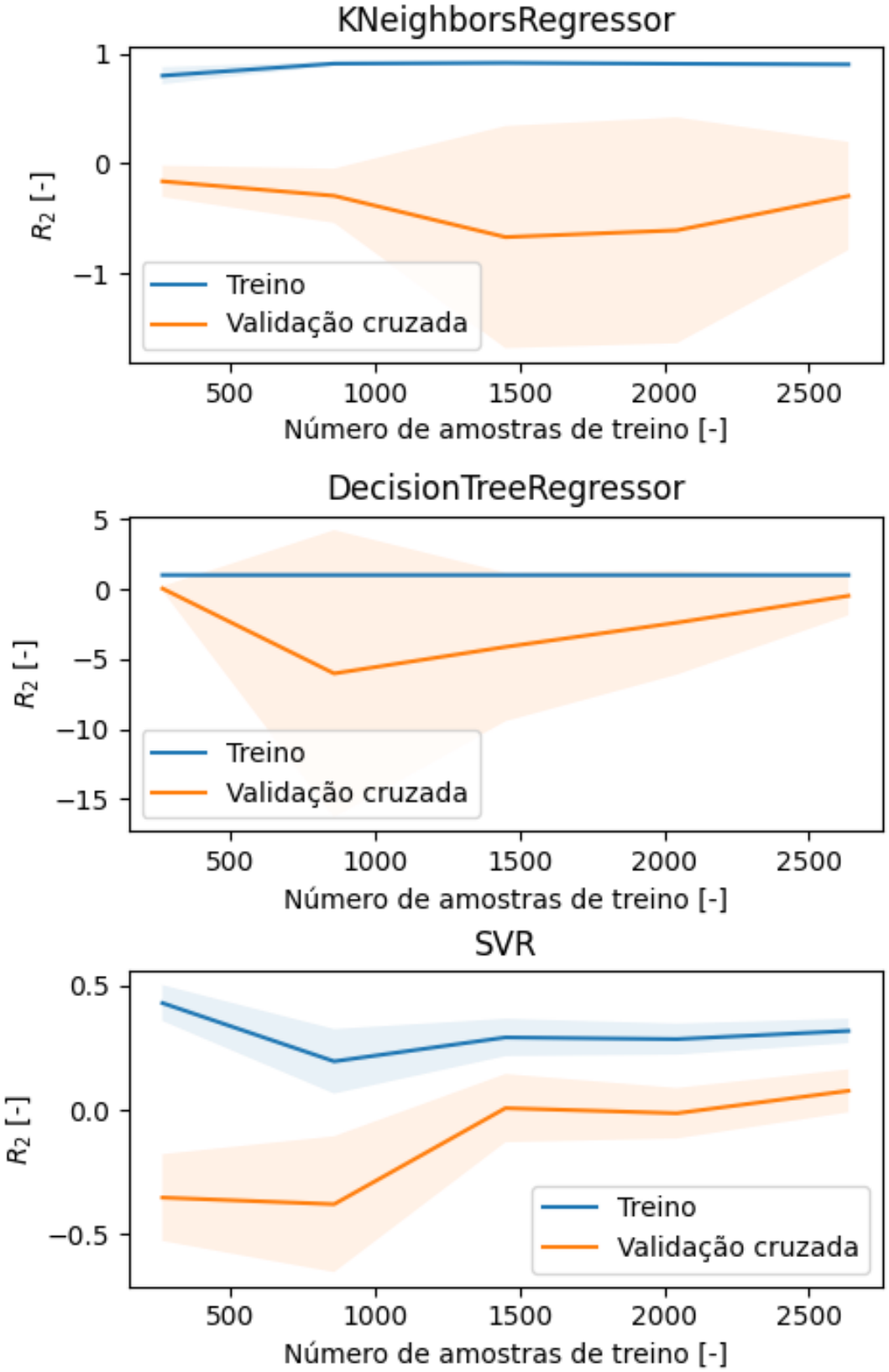
Fonte: Autoria própria (2023)

Figura A. 2. - Variação dos valores do coeficiente de determinação (R^2) para a regressão dos valores de capacidade adsorvida para os modelos KNN, AD e SVR como função do tamanho das amostras de treinamento considerando o procedimento de validação cruzada com k-folds=3.



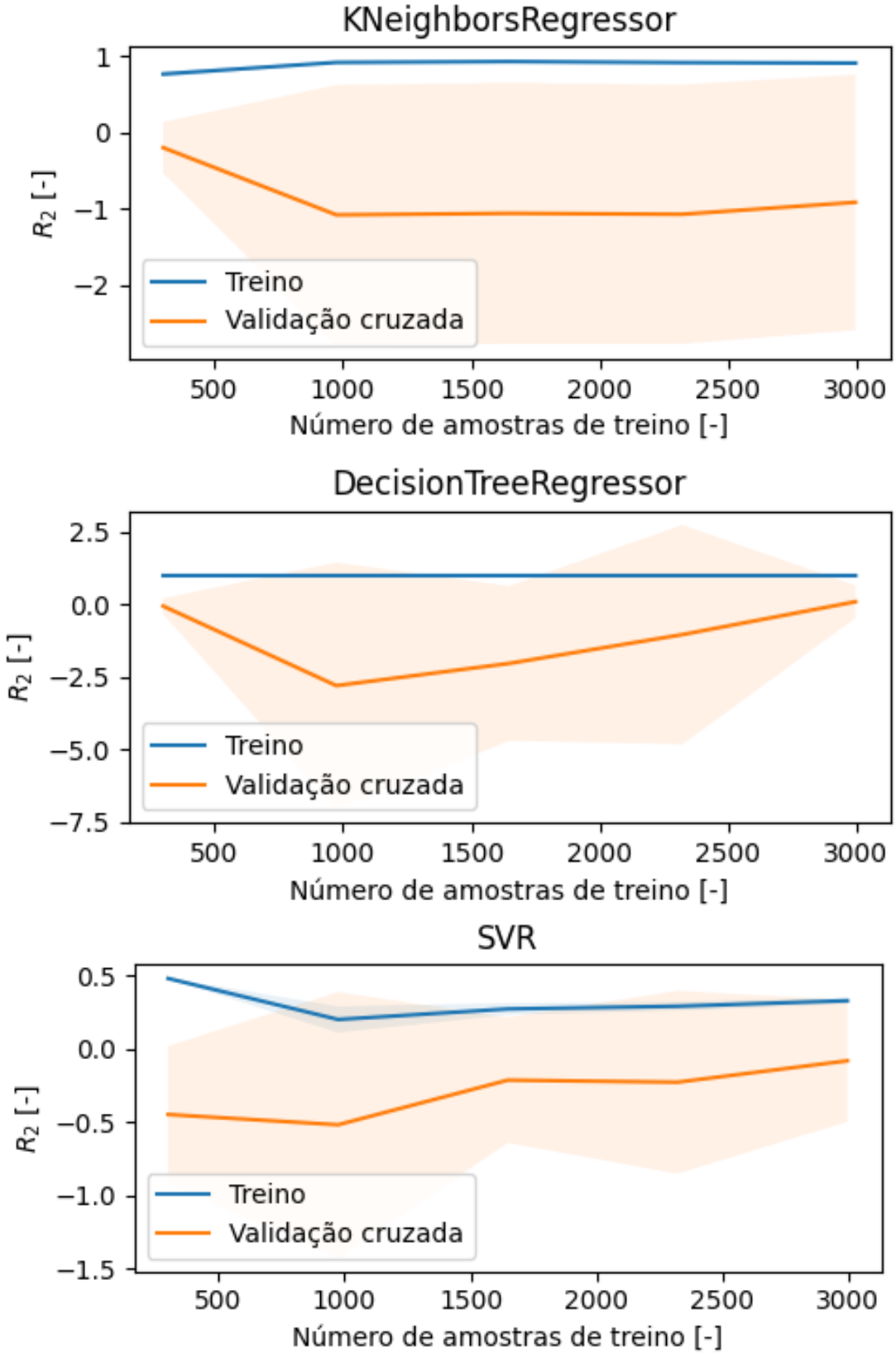
Fonte: Autoria própria (2023)

Figura A. 3 - Variação dos valores do coeficiente de determinação (R^2) para a regressão dos valores de capacidade adsorvida para os modelos KNN, AD, FA e SVR como função do tamanho das amostras de treinamento considerando o procedimento de validação cruzada com k-folds=5.



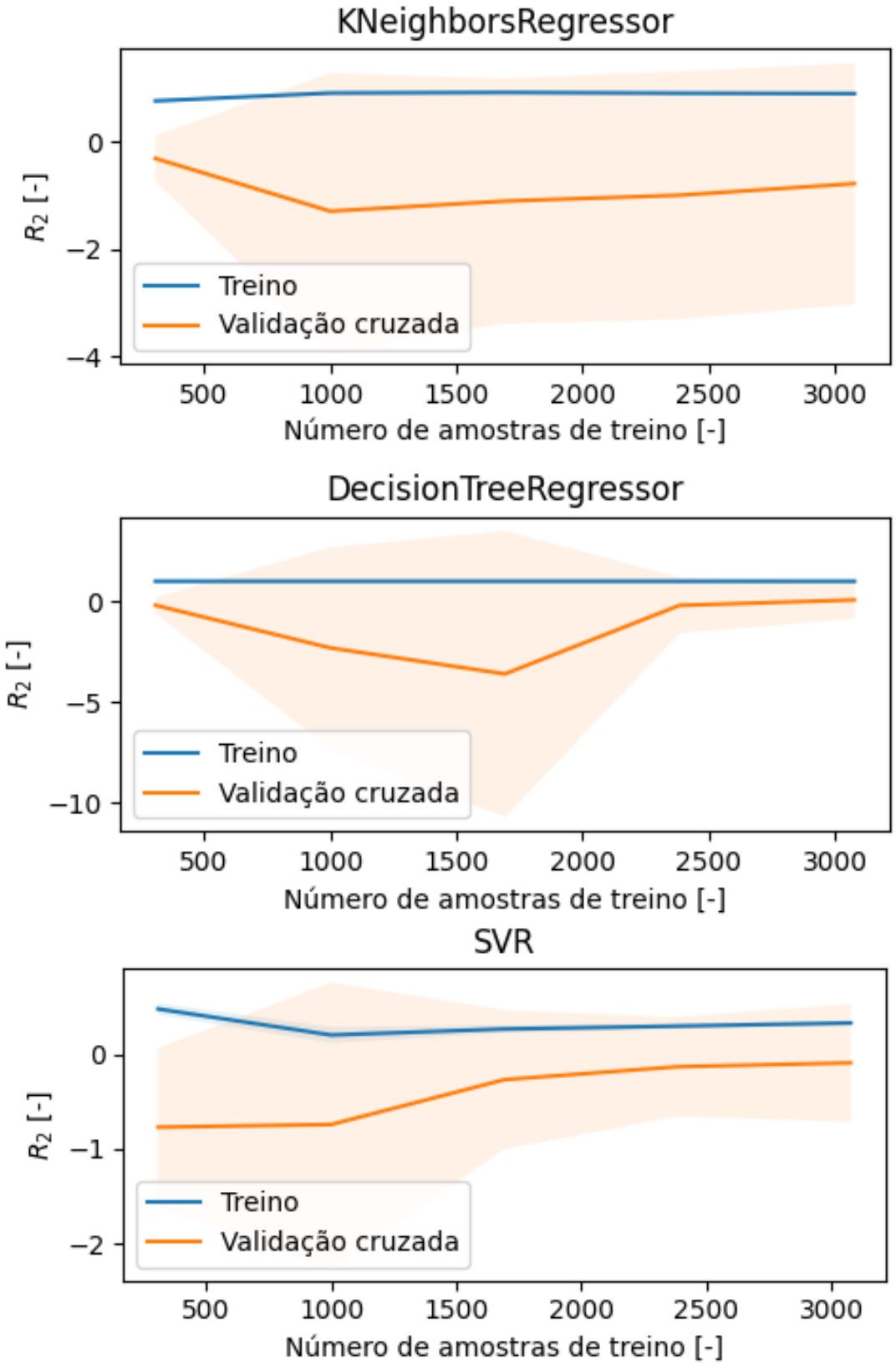
Fonte: Autoria própria (2023)

Figura A. 4 - Variação dos valores do coeficiente de determinação (R^2) para a regressão dos valores de capacidade adsorvida para os modelos KNN, AD, FA e MVS como função do tamanho das amostras de treinamento considerando o procedimento de validação cruzada com k-folds=11.



Fonte: Autoria própria (2023)

Figura A. 5 - Variação dos valores do coeficiente de determinação (R^2) para a regressão dos valores de capacidade adsorvida para os modelos KNN, AD, FA e MVS como função do tamanho das amostras de treinamento considerando o procedimento de validação cruzada com k-folds=15.



Fonte: Autoria própria (2023)