UNIVERSIDADE FEDERAL DO ABC

CENTRO DE ENGENHARIA, MODELAGEM E CIÊNCIAS SOCIAIS APLICADAS

Erick Fasterra da Silva

Aprendizado de máquina utilizando árvores de decisão e variantes para a predição da estabilidade de compostos de perovskita

Santo André

2022

Erick Fasterra da Silva

Aprendizado de máquina utilizando árvores de decisão e variantes para a predição da estabilidade de compostos de perovskita

Trabalho de Graduação apresentado à Universidade Federal do ABC como parte dos requisitos para a obtenção do título de Engenheiro de Materiais.

Orientador: Prof. Dr. Wallace Gusmão Ferreira

Santo André

2022

Sistema de Bibliotecas da Universidade Federal do ABC
Elaborada pelo Sistema de Geração de Ficha Catalográfica da UFABC
com os dados fornecidos pelo(a) autor(a).

Fasterra da Silva, Erick
Aprendizado de máquina utilizando árvores de decisão e variantes para a predição da estabilidade de compostos de perovskita / Erick Fasterra da Silva. — 2022.

88 fls. : il.

Orientador: Wallace Gusmão Ferreira

Trabalho de Conclusão de Curso — Universidade Federal do ABC, Bacharelado em Engenharia de Materiais, Santo André, 2022.

1. Aprendizado de Máquina. 2. Perovskita. 3. Random Forest. 4. Extra Trees. 5. Cerâmicos. I. Gusmão Ferreira, Wallace. II. Bacharelado em Engenharia de Materiais, 2022. III. Título.

FOLHA DE ASSINATURAS

Santo André, 18 de Agosto de 2022.

BANCA EXAMINADORA

_____

Orientador: Prof. Dr. Wallace Gusmão Ferreira

_____

Prof. Dr. Jeverson Teodoro Arantes Junior

_____

Prof. Dr. Roberto Gomes de Aguiar Veiga

Dedico este trabalho a todas e todos os que contribuíram diretamente e indiretamente para a sua elaboração

Agradecimentos

Agradeço ao Prof. Dr. Wallace pela disponibilidade, dedicação e paciência na orientação do meu trabalho de graduação, o que foi possível a sua execução.

Agradeço também aos familiares e amigos que me deram o apoio e o incentivo nesta jornada.

Absque sudore et labore nullum opus perfectum est.
“Sem suor e sem trabalho nenhuma obra é terminada”.
- Schrevellius 1176

RESUMO

O paradigma da ciência orientada a dados permite que as características de novos materiais sejam estudadas com velocidade superior às análises de laboratório e simulação computacional, sem perda ou mesmo com melhora em sua performance. Contudo, a área da Engenharia de Materiais deve lidar com o conjunto de dados geralmente escasso, o que gera limitações nas predições ao se aplicar métodos de aprendizado de máquina para a predição de características de materiais. Para avaliar a aplicação de tais métodos e estudar as suas limitações, foram aplicadas técnicas de regressão e classificação para predições, tais como *decision trees*, *random forest* e *extra trees*, para estimar a estabilidade termodinâmica de materiais de estrutura perovskita, utilizando a energia de formação acima da envoltória convexa como parâmetro. Os resultados demonstraram predições bastante precisas, mesmo com a limitação do conjunto de dados, sendo o melhor modelo o *extra tree*s para os métodos de classificação e regressão em termos de acurácia, e as *decision trees* o melhor em termos de velocidade de predição.

Palavras-Chave: Aprendizado de Máquina, Perovskita, Random Forest, Extra Trees, Cerâmicos

ABSTRACT

Data-oriented science paradigm allows the characteristics of new materials to be studied faster than laboratory analysis and computer simulation, without loss or having even improvement in their performance. However, Materials Engineering's area must deal with the generally scarce dataset, which generates limitations in predictions when applying Machine Learning methods to predict material characteristics. To evaluate the application of such methods and study their limitations, regression and classification techniques were applied for predictions, such as decision trees, random forest and extra trees, to estimate the thermodynamic stability of materials with perovskite structure, using the energy of formation above the convex envelope as a parameter. The results showed very accurate predictions, even with the limitation of the dataset, the best model being extra trees for classification and regression methods in terms of accuracy, and decision trees being the best in terms of prediction speed.

Keywords: Machine Learning, Perovskite, Random Forest, Extra Trees, Ceramics

Sumário

1 INTRODUÇÃO....11
1.1 Paradigmas da Ciência em Engenharia de Materiais....12
1.2 Objetivos....21
1.3 Estrutura de Capítulos....21
2 FUNDAMENTAÇÃO TEÓRICA....23
2.1 Técnicas de Aprendizado de Máquina....23
2.2 Extração dos Dados....27
2.3 Engenharia de Atributos....28
2.4 Seleção do modelo....32
2.5 *Decision Tree*....33
2.6 Modelos Combinados....36
2.6.1 Random Forest....38
2.6.2 Extra Trees....40
2.7 Validação dos Modelos....41
2.8 Métricas Estatísticas....43
2.9 Otimização dos Modelos....49
2.10 Exemplo de Aplicação na Literatura....50
3 METODOLOGIA....51
3.1 Problema de Estudo....51
3.2 Ferramentas Computacionais Utilizadas....54
3.3 Descrição da base de dados....55
3.4 Pré-processamento dos dados....59
3.5 Separação do Conjunto de Dados....61
3.6 Modelos de Aprendizado de Máquina....62
3.7 Validação do Treinamento....63
4 RESULTADOS....64
4.1 Pré-processamento dos Dados....64
4.2 Modelos Regressores....66
4.3 Modelos Classificadores....68
4.4 Tempo de Processamento....70
5 DISCUSSÃO....71
5.1 Implicações na Ciência e Engenharia de Materiais....71
5.2 Regressão....72
5.3 Classificação....74
5.4 Resumo....78
6 CONSIDERAÇÕES FINAIS....79
REFERÊNCIAS....82
ANEXOS....88

1 INTRODUÇÃO

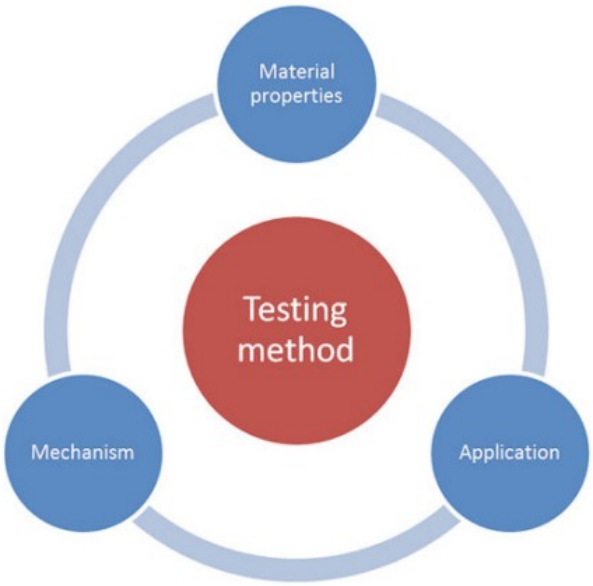
O desenvolvimento tecnológico industrial que utiliza a Engenharia de Materiais demanda cada vez mais o estudo e desenvolvimento de novos materiais, muitas vezes com composições mais complexas que visam atender às exigências e as especificações para as funções as quais serão aplicados (PICKLUM; BEETZ, 2019).

Com isso, as suas propriedades devem ser bem avaliadas, tais como o seu comportamento em ambientes sob influência da temperatura e pressão, composição química, densidade, etc. Além disso, suas características atômicas também devem ser estudadas, para compreender os fenômenos presentes no material que podem influenciar em suas características (GALAN et al. 2020).

A ciência dos materiais visa atender este objetivo, com a descrição, explicação e predição dos fenômenos físicos e químicos dos materiais, desde a extração da matéria-prima, síntese, processamento e utilização, até o seu descarte ou possível reutilização. Contudo, a natureza destes fenômenos é demasiadamente complexa, sendo necessário a realização de aproximações ou simplificações a fim de descrever o seu comportamento, obtendo assim variáveis tangíveis que possibilitam a sua compreensão e consequentemente a sua aplicação (SCHLEDER; FAZZIO,2020).

Testes envolvendo caracterização de materiais podem ser aplicados para confirmar a natureza dos fenômenos presentes nos materiais estudados. A Figura 1 mostra uma representação simplificada deste conceito, indicando a dependência dos métodos de caracterização dos materiais com as suas propriedades, mecanismos de funcionamento e aplicação (BODE et al. 2015).

Figura 1 - Relação dos métodos de teste e caracterização de materiais com as suas propriedades, mecanismos de funcionamento e aplicação



BODE et al. 2015

No entanto, os métodos mais tradicionais para a avaliação de materiais, tais como a experimentação laboratorial e a simulação computacional, tem se mostrado demasiadamente custosa, além de consumir tempo considerável para cada análise (LU, 2021).

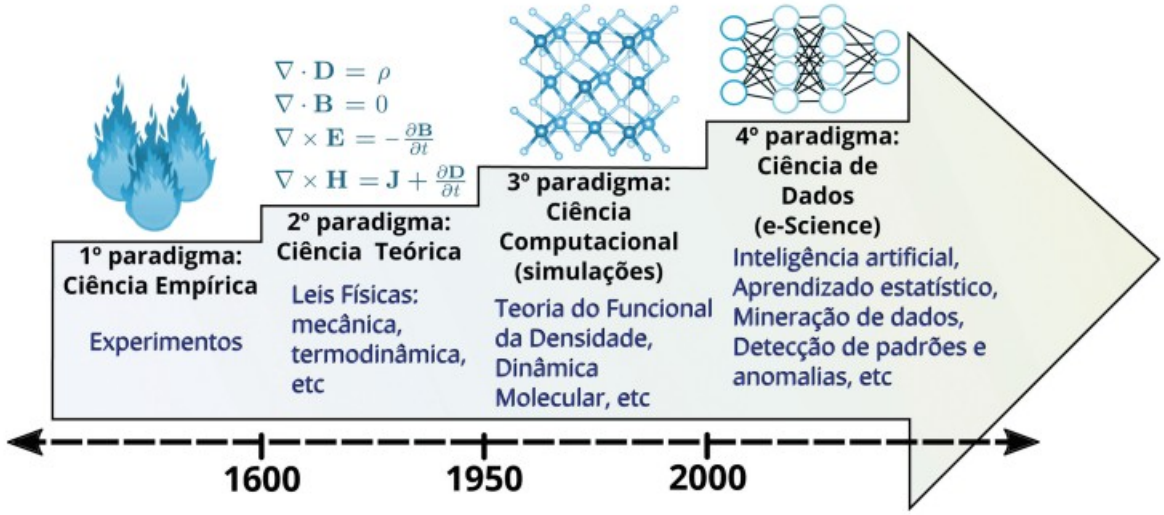
Desta forma, novos métodos foram desenvolvidos para que essas desvantagens sejam minimizadas sem perder a precisão e o rigor das análises dos materiais. Uma das abordagens inclui o uso de aprendizado de máquina, que, no contexto da ciência e engenharia de materiais, visa predizer diversas propriedades de diferentes tipos de materiais, dados como entrada atributos relevantes para a predição desejada (LU, 2021).

1.1 Paradigmas da Ciência em Engenharia de Materiais

Ao longo da história, paradigmas foram estabelecidos para possibilitar a compreensão dos fenômenos envolvidos na ciência de materiais para a obtenção do

melhor conjunto de propriedades que atendam a uma exigência específica. A Figura 2 apresenta os quatro paradigmas presentes no desenvolvimento, compreensão e análise de novos materiais, e que serão detalhados em seguida (SCHLEDER; FAZZIO, 2020).

Figura 2 - Os quatro paradigmas da ciência



SCHLEDER; FAZZIO, 2020

O primeiro paradigma se baseia na ciência empírica, com a experimentação em laboratório por tentativa e erro, utilizando reagentes e equipamentos. O processamento de diversos materiais sob determinadas condições levam a um conjunto de propriedades específicas e podem inferir em vários resultados finais diferentes, registrando todo o processo, seja com resultados favoráveis ou desfavoráveis à aplicação estudada (LU, 2021).

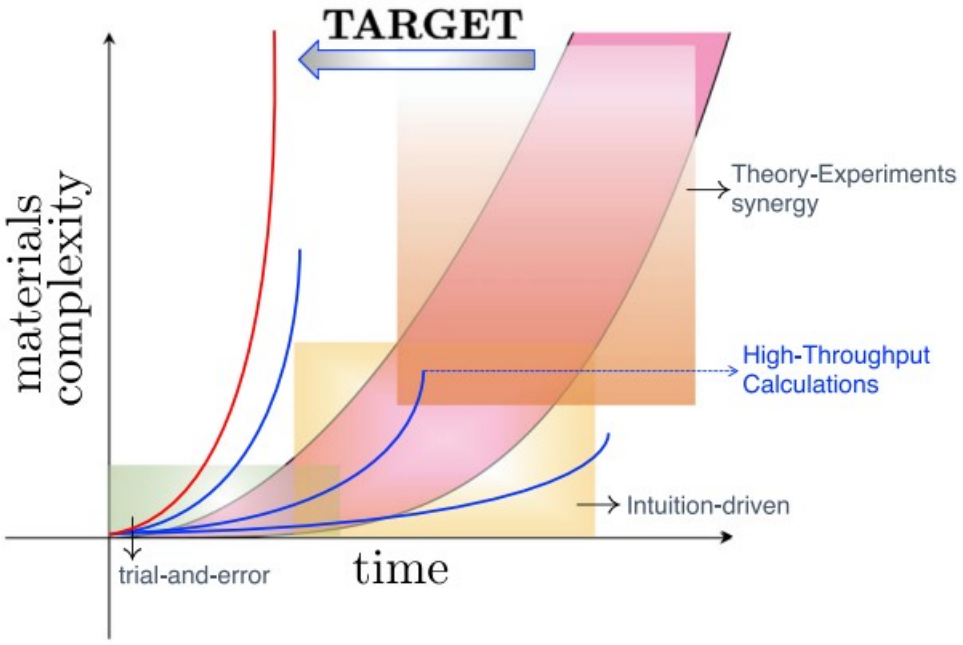
No entanto, a realização de testes físicos utilizando equipamentos de laboratório é um processo que demanda bastante tempo e gera bastante custo, além de bastante esforço laboratorial para suprir as demandas industriais (SCHLEDER; FAZZIO, 2020).

Isso leva ao segundo paradigma, baseado na ciência teórica, no qual as leis físicas relacionadas à mecânica, termodinâmica, etc., são estudadas e modeladas

para a resolução de um problema. Equações matemáticas que visam simplificar o problema em questão são aplicadas para a predição do comportamento do material, para que assim as experimentações em laboratório possam ser direcionadas (SCHLEDER; FAZZIO, 2020).

Porém, à medida que surgem novas revoluções na ciência, ocasionadas pela criação de novas ferramentas de análise, o conceito teórico se tornou cada vez mais complexo, com modelagens extremamente específicas para cada natureza de um problema. Além disso, a demanda por materiais cada vez mais complexos continua consumindo bastante tempo. A Figura 3 mostra a trajetória de descoberta de novos materiais e sua complexidade em função do tempo requerido para a obtenção e análise das propriedades desejadas com o uso de métodos presentes no primeiro e segundo paradigma (LOOKMAN et al. 2019).

Figura 3 - Aumento na complexidade de materiais em função do tempo para a obtenção e análise das propriedades desejadas. Idealmente, o tempo não deve variar após um determinado nível de complexidade, idealidade representada pela linha vermelha.



LOOKMAN et al. 2019

Desta forma, foi-se necessário descrever os fenômenos de novas maneiras de forma a generalizar o problema estudado sem perder a sua eficácia de análise. Muitas análises requerem um elevado grau de precisão, e comumente necessitam atingir níveis microscópicos inacessíveis às análises com equipamentos mecânicos tradicionais, sendo necessário equipamentos modernos que requerem que as informações obtidas no processo de caracterização do material seja transformado em uma medida ou representação que possa ser compreendida pelo analista (WEI et al. 2019).

Assim, surge o terceiro paradigma, baseado em ciência computacional e simulação. Um exemplo de simulação é a aplicação computacional da teoria do funcional da densidade (*Density Functional Theory*, ou DFT), a qual descreve propriedades eletrônicas na ciência dos materiais e física do estado sólido através de funcionais, que são funções que recebem outra função como variável através da densidade eletrônica. Simulações que envolvem teoria do funcional da densidade, dinâmica molecular, Monte Carlo, etc., se tornaram extremamente úteis para o estudo dos materiais, simulando o seu comportamento através de simplificações e simulações realizadas computacionalmente (SHOLL et al. 2011).

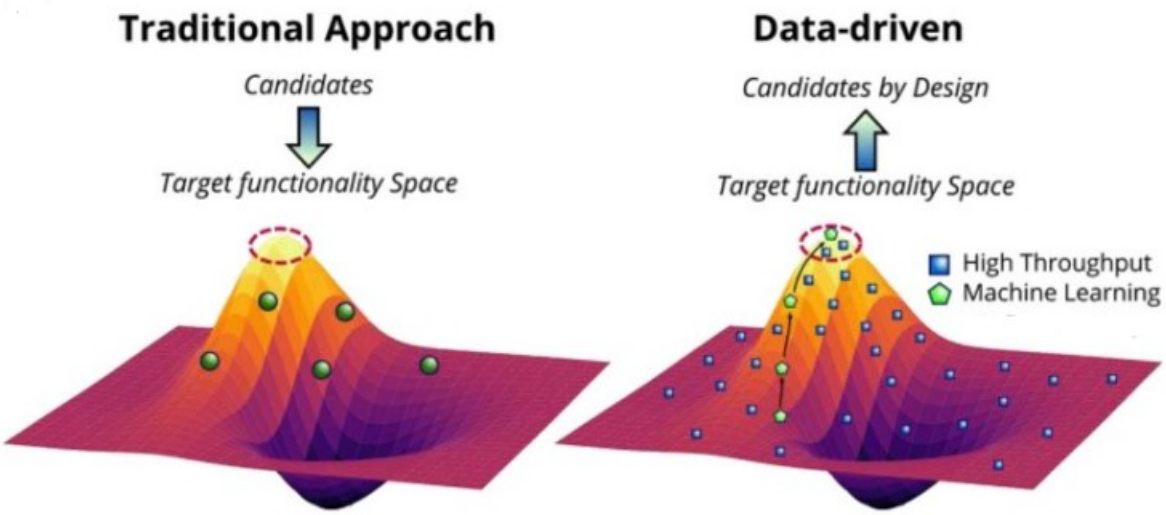
A computação se tornou um elemento crucial para o armazenamento de informações sobre os materiais aplicados na indústria, bem como o processamento desta informação através do processamento digital de imagens e simulação, e a sua representação sob a forma de gráficos, imagens bidimensionais e tridimensionais, animações que representam o seu comportamento em nível atômico e molecular, etc. Isso permite o estudo analítico das características do material a ser empregado ou desenvolvido para uma determinada aplicação (PICKLUM; BEETZ, 2019).

Para se atingir o conjunto de características desejadas para um determinado material, faz-se necessárias diversas etapas para o seu processamento, obtendo assim um conjunto de atributos que podem ser influenciadores de um outro determinado conjunto. Desta forma, pode-se ter diversas combinações de atributos dentro deste mesmo conjunto, elevando exponencialmente a quantidade de valores combinados para um mesmo material para serem buscados, visando a característica desejada (WEI et al. 2019).

Assim, mesmo com o emprego da computação para auxiliar na rotina de um Engenheiro de Materiais neste processo, este pode se tornar extremamente penoso ou mesmo computacionalmente inviável, graças à “maldição da dimensionalidade”, que ocorre quando existe a dificuldade de se trabalhar com muitos elementos que possuem muitos atributos para cada um. A operação com equipamentos e processos de laboratório se tornam neste cenário praticamente impossíveis, e as simulações se tornam lentas e exaustivas (PICKLUM; BEETZ, 2019).

Para contornar esta limitação, pode ser aplicado o quarto e último paradigma de análise, que é baseado em ciência de dados e aprendizado de máquina. Este paradigma permite que experimentos realizados no passado possam ser aplicados para predizer o comportamento de um novo experimento. Isso possibilita estimar as características requeridas para uma determinada aplicação sem a necessidade de realizar diversos experimentos fisicamente, além de se tornar computacionalmente viável e de dar espaço as novas possibilidades de combinações de atributos e de possibilitar a descoberta de novos materiais com as propriedades desejáveis e otimizadas para uma determinada aplicação (LU, 2021). A Figura 4 mostra as matrizes de propriedades do material, representadas pelos pontos verdes, comparando a sua obtenção através do método tradicional por tentativa e erro (primeiro paradigma) com o método baseado em ciência de dados (GIUSTINO et al. 2020).

Figura 4 - Comparação entre o método tradicional por tentativa e erro e o método atráves de ciência de dados na busca por novos materiais que atendam aos critérios definidos

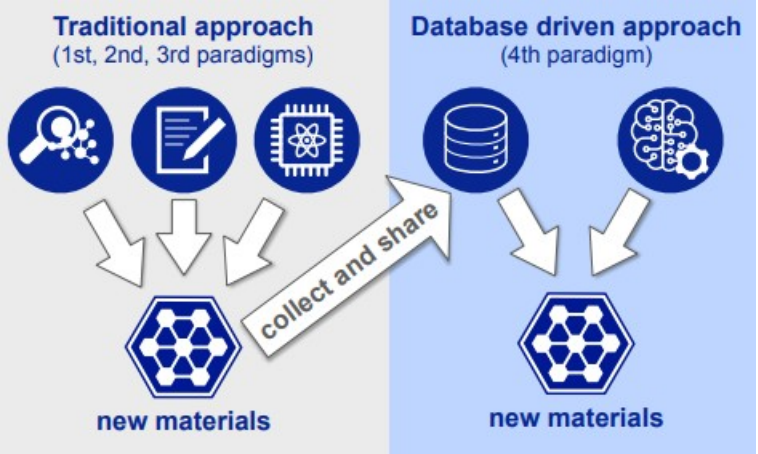


GIUSTINO et al. 2020

Além disso, este paradigma permite o estudo e simulação de um grande número de compostos, visando a busca de características otimizadas ou específicas para uma determinada aplicação. No paradigma baseado em simulação, essa catalogação não era possível, sendo necessário o estudo caso a caso de cada composto investigado (SCHLEDER; FAZZIO, 2020).

É importante notar que cada paradigma utiliza e complementa conceitos de paradigmas anteriores. O quarto paradigma, baseado em ciência de dados, sua dependência das experimentações, teorias e simulações realizadas anteriormente visa obter o conjunto de conhecimento necessário para a sua execução. Assim, neste estudo apesar do foco e da aplicação direta do quarto paradigma, os três paradigmas anteriores estarão implicitamente presentes, como mostra a Figura 5 (SCHLEDER; FAZZIO, 2020).

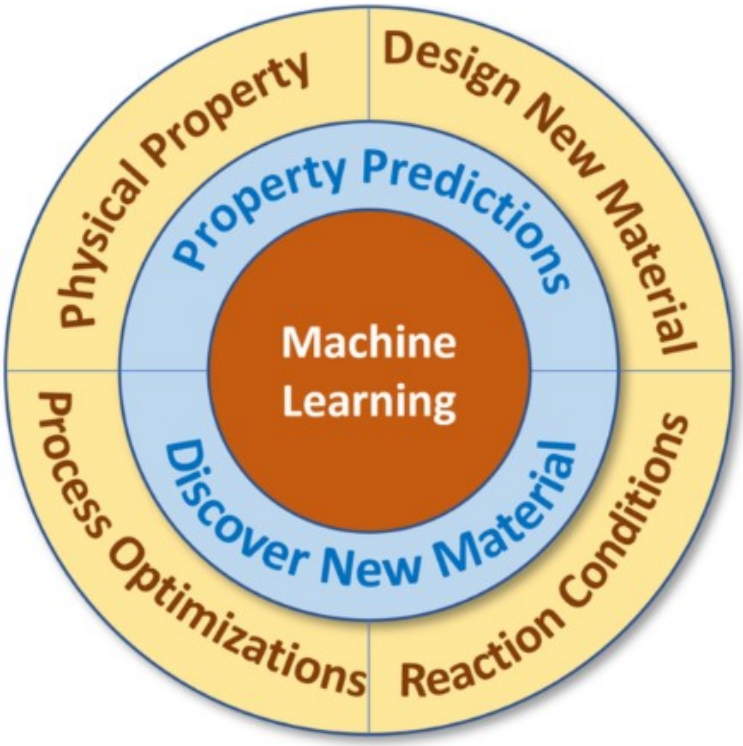
Figura 5 - Dependência dos três primeiros paradigmas para a execução do paradigma baseado em ciência de dados



HIMANEN et al. 2019

O quarto paradigma permite a aplicação de aprendizado de máquina para realizar predições no âmbito da ciência e engenharia de materiais. As possibilidades de predição abrangem praticamente todas as propriedades de materiais estudadas pela ciência dos materiais, como as propriedades mecânicas, térmicas, elétricas, magnéticas, ópticas, etc, além de avaliar comportamentos funcionais dos materiais, tais como a piezoeletricidade, a ferroeletricidade e a supercondutividade. Mudanças no ambiente, reações a estímulos externos e efeitos decorrentes do tempo e da temperatura são considerações que também podem estar aplicadas nas predições das propriedades dos materiais (CHAN et al. 2022). A Figura 6 elucida as possibilidades do aprendizado de máquina na engenharia de materiais.

Figura 6 - Possibilidades abrangentes do aprendizado de máquina na área de ciência e engenharia de materiais



CHAN et al. 2022

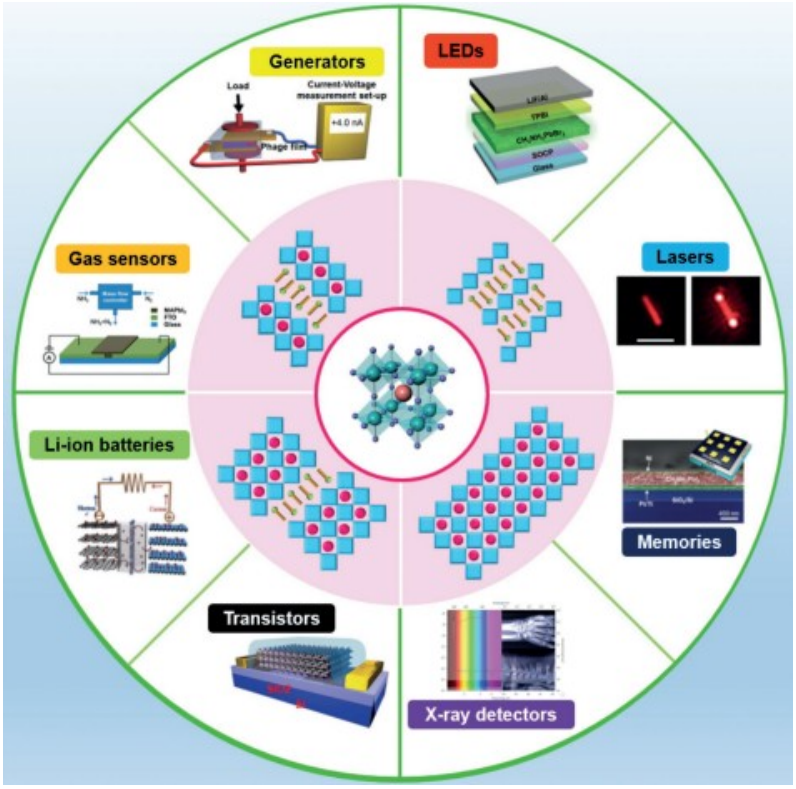
Além da ampla possibilidade de predição, os modelos de aprendizagem de máquina consomem pouco tempo e são menos custosos em relação aos métodos de simulação computacional tradicionais. Isso possibilita maior eficiência na avaliação dos materiais, reduz custos significativos e ganho significativo de tempo (LIANG et al. 2022).

Apesar do aprendizado de máquina ser um método que está em constante desenvolvimento e cada vez mais aplicado em diversas áreas da ciência, a ciência e engenharia de materiais ainda apresenta muitas dificuldades para a sua aplicação. Um dos motivos principais é a falta de conjunto de dados de experimentação realizadas e consolidadas. Existem poucos destes conjuntos disponíveis, e quando existem, a quantidade de dados é pequena ou insuficiente para alimentar um algoritmo de aprendizado de máquina (SCHLEDER; FAZZIO, 2020).

Alguns materiais são particularmente mais difíceis de serem abordados por tais algoritmos, como por exemplo os materiais de estrutura perovskita. Por possuírem diversas possibilidades de combinações em sua composição estrutural, o número de compostos únicos pode ser maior que um milhão, tornando assim inviável a experimentação ou a simulação de cada um destes compostos (LI et al. 2018).

A ampla possibilidade de combinações nos compostos de perovskita permite que novos materiais possam ser desenvolvidos, a fim de otimizar as propriedades requeridas para diversas aplicações, como por exemplo, no emprego destes materiais na fabricação de células solares, LEDs, capacitores, transistores, detectores de raios-X, dentre outros. A Figura 7 mostra algumas aplicações de materiais de estrutura perovskita (KIM et al. 2018).

Figura 7 - Aplicações de materiais de estrutura perovskita além de seu uso em células solares.



KIM et al. 2018

1.2 Objetivos

Assim, neste trabalho, busca-se estudar os principais conceitos de aprendizado de máquina e sua aplicação na engenharia de materiais utilizando métodos e ferramentas que abrangem desde a extração à predição de propriedades de materiais, além de compreender as dificuldades presentes no quarto paradigma da ciência e engenharia de materiais, como a quantidade limitada de amostras presentes para a predição de propriedades.

Para este estudo, três modelos de aprendizado de máquina foram aplicados para a determinação da estabilidade de compostos de perovskita, sendo um deles um modelo base e os outros dois modelos combinados. Para cada um dos modelos, métodos de regressão e classificação foram utilizados a fim de verificar a melhor abordagem utilizando métodos estatísticos de avaliação de modelos de aprendizado de máquina e comparando com a literatura.

1.3 Estrutura de Capítulos

No capítulo 2, será dado uma visão geral sobre o funcionamento do aprendizado de máquina aplicado à análise de novos materiais, desde a extração dos dados até a seleção dos modelos e avaliação final, destacando as métricas estatísticas utilizadas no estudo.

No capítulo 3, será explicado o estudo de caso e o procedimento para o tratamento dos dados e para o desenvolvimento do aprendizado de máquina, bem como a metodologia aplicada para a validação da eficiência do algoritmo gerado.

No capítulo 4, os resultados serão explanados, constando o comportamento e descrição dos dados avaliados neste estudo, os parâmetros que foram encontrados para se obter o modelo mais eficiente, os valores estatísticos obtidos ao serem introduzidos amostras de novas composições e análises gráficas obtidas no decorrer dos experimentos.

No capítulo 5, será realizada a discussão dos resultados obtidos, avaliando o processamento dos dados, a seleção dos modelos, a comparação entre os métodos, as métricas estatísticas observadas e a viabilidade dos modelos para aplicações

reais de estudo de novos compostos de perovskita. Assim, os resultados foram comparados com a literatura.

Por fim, no capítulo 6, será realizado uma reflexão sobre o experimento, destacando as dificuldades encontradas e a possibilidade de estudos posteriores.

2 FUNDAMENTAÇÃO TEÓRICA

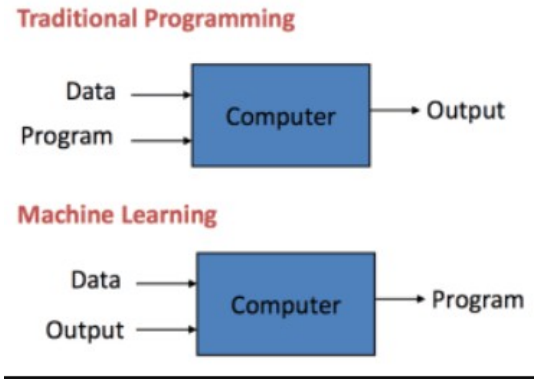
2.1 Técnicas de Aprendizado de Máquina

O aprendizado de máquina (ou *Machine Learning*, ML em forma abreviada) trata-se da detecção de padrões presentes nos dados fornecidos a um algoritmo de análise automatizada, permitindo assim que algoritmos aprendam a encontrar a solução para um determinado problema a partir dos dados que lhe são fornecidos, sem possuírem programação explicita para isso (BELGIU; DRÂGUT, 2016).

O algoritmo também permite a previsão de comportamentos desconhecidos dado uma sequência de atributos, como as propriedades de um material, que podem servir para a geração de *insights* e em processos de tomada de decisão. Uma definição mais rudimentar deste conceito para descrever o aprendizado de máquina é qualquer classe de métodos para que computadores imitem o comportamento da inteligência humana na geração de conclusões a partir de experiências previamente conhecidas, relacionando as informações disponíveis para que deduções entre elas possam ser realizadas ou adquirindo informação adicional que complementem as informações já existentes. (SCHLEDER; FAZZIO, 2020; RAKHRA et al. 2021).

A programação tradicional possui um conjunto de instruções próprias com inputs esperados, gerando um *output* com a sua solução. O aprendizado de máquina, por outro lado, busca encontrar o melhor conjunto de regras para a solução do problema a partir de um conjunto de dados em sua entrada, e pode ser composto por uma ou mais estratégias de predição, como o uso de regras do tipo “se-então”, lógica computacional, regressão linear, dentre outros. Muitas vezes, este mesmo algoritmo pode aprender a resolver diversos tipos de problemas dado um novo conjunto de dados (RAKHRA et al. 2021). Essa diferença entre a programação tradicional e o *Machine Learning* pode ser visualizado na Figura 8.

Figura 8 - Comparação das entradas e saídas presentes na programação tradicional e no aprendizado de máquina



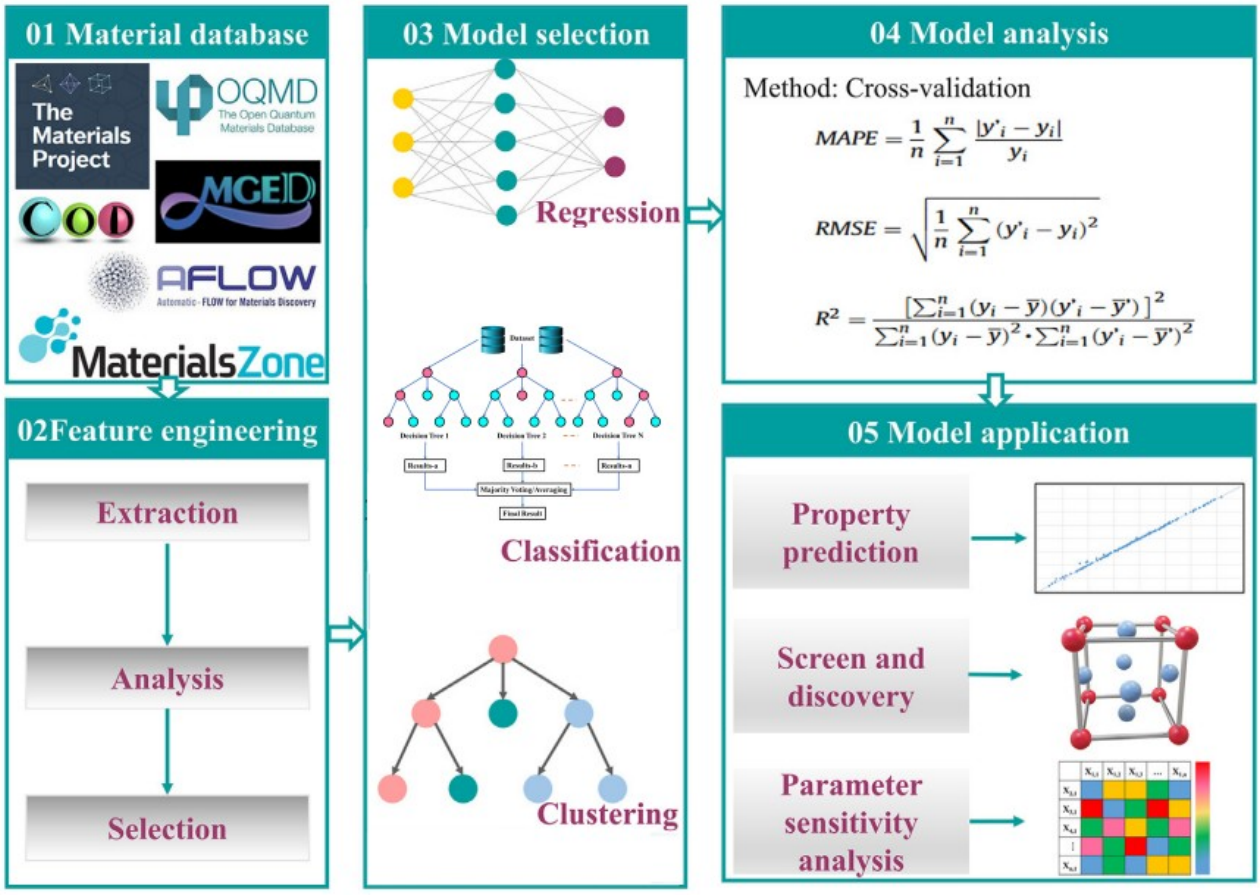
BROWNLEE, 2020

Desta forma, cálculos de elevada complexidade computacional podem ser substituídos por modelos mais simples e eficientes, como na predição de diversas propriedades inerentes aos materiais, como a energia livre, de formação e total de uma amostra, presença e possíveis ocorrências de defeitos, propriedades térmicas, mecânicas, elétricas, magnéticas, ópticas, etc.(SCHLEDER; FAZZIO, 2020).

Um dos tipos de aprendizado de máquina é o aprendizado supervisionado. Este permite que o algoritmo encontre a solução de um problema através de um conjunto de soluções já conhecidas para um conjunto de dados de entrada. Seu processo se constitui basicamente em treinamento e validação do modelo de aprendizado de máquina, e seu nome se dá por conta da necessidade do operador humano de verificar se as soluções dos dados de entrada são condizentes com os valores reais esperados. (LIU et al. 2021).

A Figura 9 resume as etapas presentes no aprendizado de máquina. Em seguida, será explicada cada uma dessas etapas, com ênfase no aprendizado de máquina supervisionado aplicado à ciência e engenharia de materiais para o estudo de novos compostos (LIU et al. 2021).

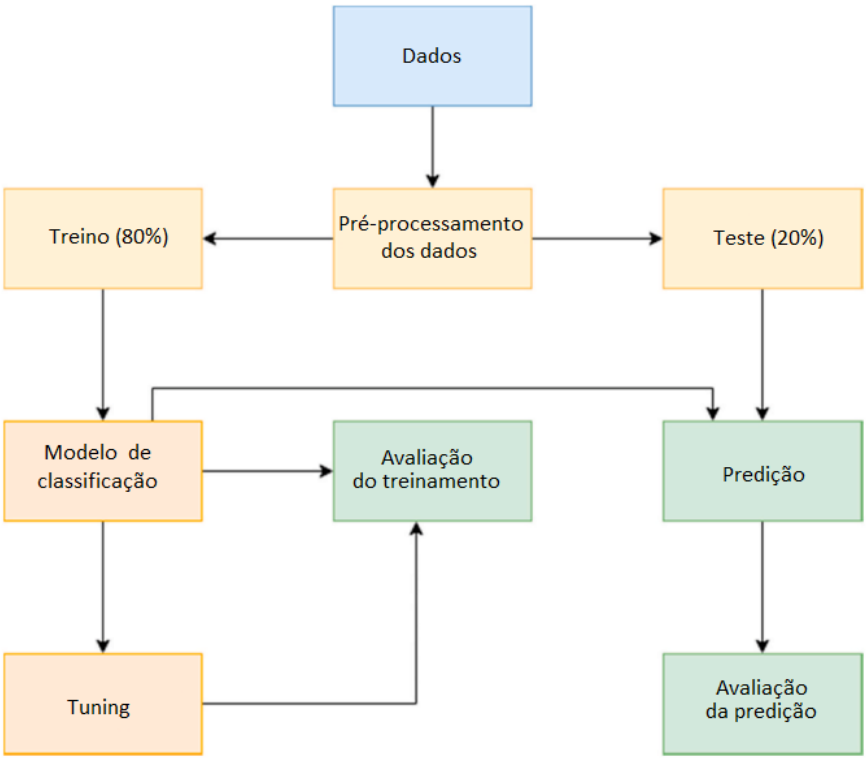
Figura 9 - Esquematização generalizada do aprendizado de máquina aplicado à diferentes tipos de predições em materiais



LIU et al. 2021, adaptado

Uma simplificação do fluxo de aprendizado de máquina pode ser visualizado na Figura 10. Essas etapas foram utilizadas no presente estudo e serão detalhadas nas seções seguintes.

Figura 10 - Processo simplificado do treinamento e predição de um aprendizado de máquina supervisionado



AL-AZZAM; SHATNAWI, 2021, adaptado

Todo o processo depende do método escolhido para a modelagem, desde a extração até a análise dos dados. Estes métodos podem ser separados em métodos de regressão ou métodos de classificação (LIU et al. 2021).

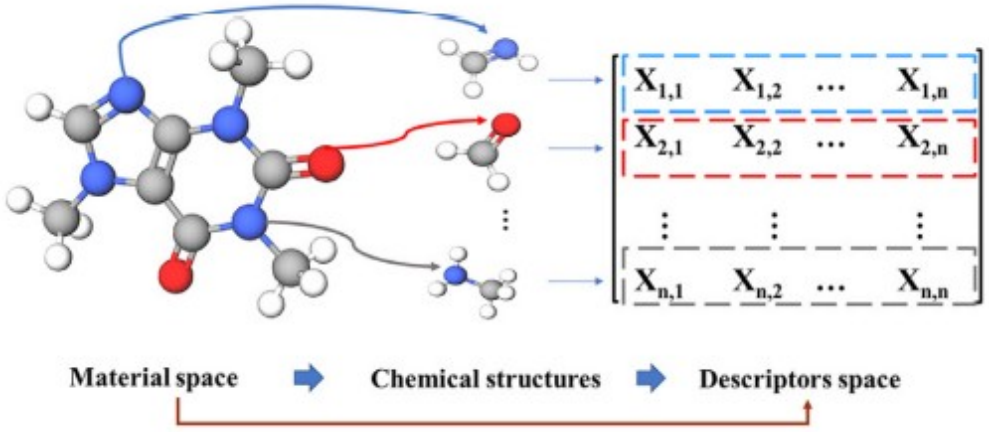
Para a regressão, espera-se estimar como dados de saída valores numéricos, como por exemplo, valores em escala decimal de energia de formação de um determinado material. A classificação, por outro lado, visa obter valores bem definidos, como por exemplo se um material é termodinamicamente estável ou não (LIU et al. 2021).

2.2 Extração dos Dados

A base de dados para compor o modelo de aprendizagem de máquina supervisionado poderá ser extraída de artigos científicos ou de bases confiáveis na internet que contém informações extremamente úteis dos materiais a serem analisados, como suas propriedades físico-químicas, processamento, aplicações, etc.

Desta forma, a coleta de dados relevantes ao problema, bem como o seu devido tratamento, são cruciais para garantir o desempenho do algoritmo de aprendizado de máquina supervisionado. A coleta deve providenciar uma série de valores legíveis para o modelo, normalmente disposto na forma de uma matriz composta por diversos vetores de característica, como mostrado na Figura 11 (LIU et al. 2021).

Figura 11 - Extração dos dados em uma matriz interpretável por um modelo de aprendizagem de máquina

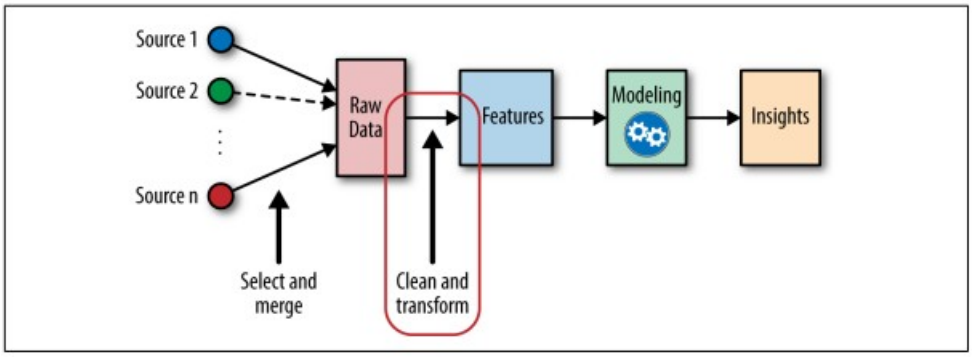


LIU et al. 2021

2.3 Engenharia de Atributos

O tratamento dos dados envolvem a remoção de valores vazios ou faltantes no conjunto de dados trabalhado, remoção de *outliers* (valores atípicos), verificação de tipagem, normalização dos dados, verificação de dados correlacionados, etc. Essa etapa se trata do pré-processamento dos dados, e é uma etapa bastante importante para garantir o desempenho do modelo. Comumente é a etapa que mais demanda tempo por parte do seu projetista, incluso também a área de ciência dos materiais pela dificuldade e alto custo da extração dos dados e tratamento dos mesmos (WARD, 2016). A Figura 12 indica a etapa onde a engenharia de atributos (ou atributos) atua no processo de aprendizagem de máquina.

Figura 12 - Processo simplificado de aprendizagem de máquina, com destaque na etapa de limpeza e transformação de dados, onde a engenharia de atributos atua



ZHANG; CASARI, 2018

Para o treinamento de um modelo baseado em aprendizado supervisionado, as amostras devem ser balanceadas, de forma a obter o conjunto de valores o mais distribuído possível, tanto para o conjunto de treinamento quanto para o conjunto de teste. No caso da regressão, dado o exemplo de resultados entre 0 e 10, o modelo se comportaria com uma performance muito baixa, caso haja muitas amostras cujos resultados se aproximam de 10 e poucas amostras com resultados próximos de 0, considerando que os valores reais são distribuídos nesta faixa de valores. Neste

caso, os resultados do modelo tendem a favor dos valores mais presentes na distribuição (WARD, 2016).

O mesmo ocorre com os modelos baseados em classificação. Quando existem muito mais amostras de uma determinada classe e escassez de amostras de outra classe, pode ocorrer o desbalanceamento (WARD, 2016).

Assim, caso haja desbalanceamento, técnicas de balanceamento podem ser empregadas com o decréscimo ou acrescimo artificial de amostras até que ambas as classes apresentem o mesmo número de amostras (WARD, 2016).

No entanto, a escolha ao se utilizar essa abordagem deve ser realizada com muita cautela, e deve ser evitada quando possível. Uma das razões é a inserção de vieses no modelo, uma vez que se dados importantes são eliminados do modelo, pode haver perda de generalização do mesmo, gerando dificuldades para se considerar os casos que foram eliminados. Elevando a quantidade de amostras das classes minoritárias, pode ocorrer em acrescimo de influência de amostras que apresentam anomalias. Além disso, amostras adicionadas artificialmente podem não refletir na realidade dos experimentos (VERZINO, 2021).

Sendo assim, a não homogeneidade do modelo é um parâmetro do mundo real que deve ser considerado, sendo a verificação de sua influência de suma importância para a performance dos modelos de aprendizado de máquina. Se as amostras das classes minoritárias apresentarem características intrínsecas, com dependências relevantes dos atributos das amostras, estas podem ser suficientes para que o modelo consiga identificar essas características (VERZINO, 2021).

Para isso, deve haver uma quantidade considerável de amostras, o que se apresenta como um desafio para a aplicação de aprendizado supervisionado na área de Engenharia de Materiais, por seus experimentos e obtenção de dados serem usualmente caros. Poucas amostras no modelo pode dificultar a sua aprendizagem, diminuindo a acurácia no processo de validação do modelo, como observa-se na Figura 13. (BELGIU; DRÂGUT, 2016)

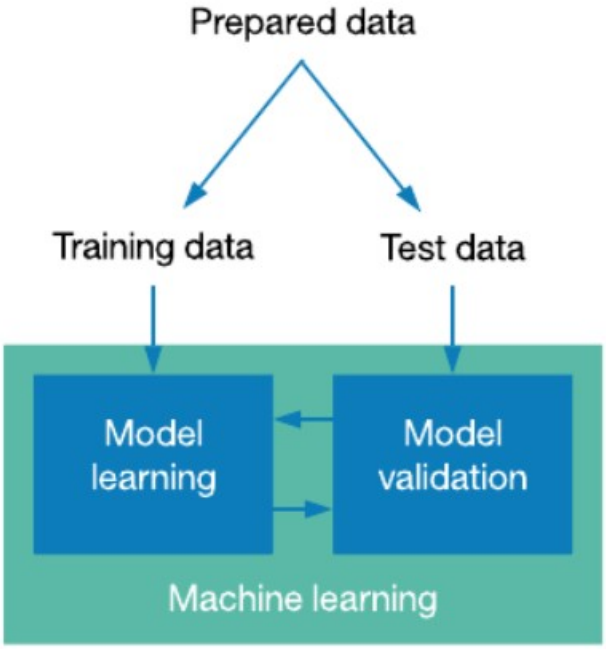
Figura 13 - Relação entre a quantidade de amostras e a acurácia de um modelo de aprendizagem supervisionada

SCHLEDER; FAZZIO,2020

Após a limpeza e a transformação dos dados, estes devem ser separados em: conjunto de treinamento, onde contém os dados e soluções conhecidos pelo modelo, servindo para a aprendizagem pelo algoritmo; e o conjunto de teste, o qual contém dados cujas soluções são desconhecidas pelo modelo, sendo utilizados para avaliar a sua performance, ou seja, se ao serem introduzidos novos dados ao modelo, as predições são coerentes com o resultado esperado (AL-AZZAM; SHATNAWI, 2021).

A separação do conjunto de dados normalmente obedece uma distribuição de 80% dos dados para o conjunto de treinamento e 20% para o conjunto de teste (AL-AZZAM; SHATNAWI, 2021). Essa divisão pode ser visualizada na Figura 14.

Figura 14 - Divisão dos dados em conjuntos de treinamento e de teste



VOLPI, 2019

Os dados de treinamento e de teste devem ser estatisticamente independentes, ou seja, a partir dos resultados obtidos no conjunto de treinamento, não deve ser possível inferir qualquer conclusão sobre os resultados do conjunto de teste, e vice-versa. Isso significa que quando o modelo é treinado, deve-se necessariamente introduzir os dados de teste, sem os quais não é possível obter qualquer conclusão. Os modelos que utilizam aprendizagem supervisionada costumam ser bastante sensíveis às correlações existentes entre as classes das amostras de treinamento, e devem assim ser evitados (BELGIU; DRÂGUT, 2016).

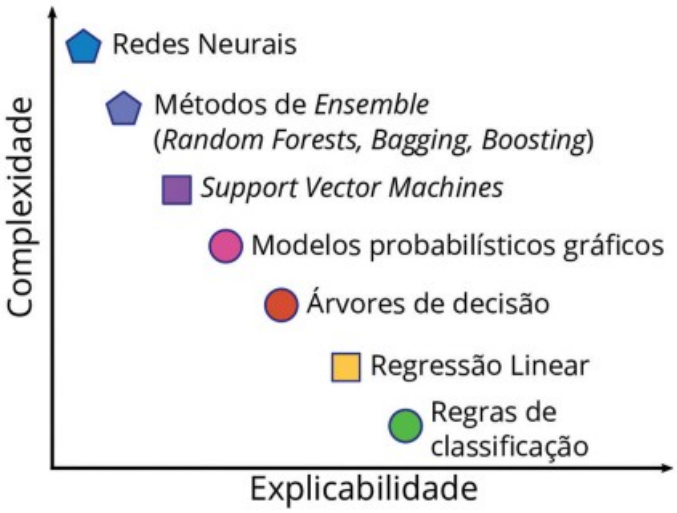
As amostras de treinamento devem refletir o comportamento das amostras de teste. Isso significa que mesmo seguindo a regra de independência estatística, as amostras de teste devem apresentar a mesma natureza de eventos das amostras de treinamento, ou seja, se é realizada uma análise no conjunto de treinamento, como a estimativa de estabilidade termodinâmica de diversos compostos, a mesma análise também deve ser realizada no conjunto de teste (ZHANG; CASARI, 2018).

2.4 Seleção do modelo

A escolha do modelo a ser aplicado no problema é crucial para se obter eficiência nas predições com a maximização de sua acurácia. Seu êxito deve considerar a qualidade dos dados inseridos após o pré-processamento dos dados e sua complexidade, e deve haver equilíbrio entre a explicabilidade e a complexidade dos algoritmos de aprendizado de máquina, de acordo com a aplicação e o tipo de estudo que está sendo realizado (BELGIU; DRÂGUT, 2016).

Modelos muito complexos são denominados modelos de caixa-preta, uma vez que a determinação dos atributos mais importantes também é complexa, enquanto que modelos menos complexos tendem a simplificar o problema, não sendo apropriado para a resolução de problemas de elevada complexidade ou cujos dados não satisfazem totalmente um ou mais critérios mencionados na seção anterior. A relação da complexidade de diversos modelos presentes na literatura com sua explicabilidade pode ser observada na Figura 15 (SCHLEDER; FAZZIO, 2020).

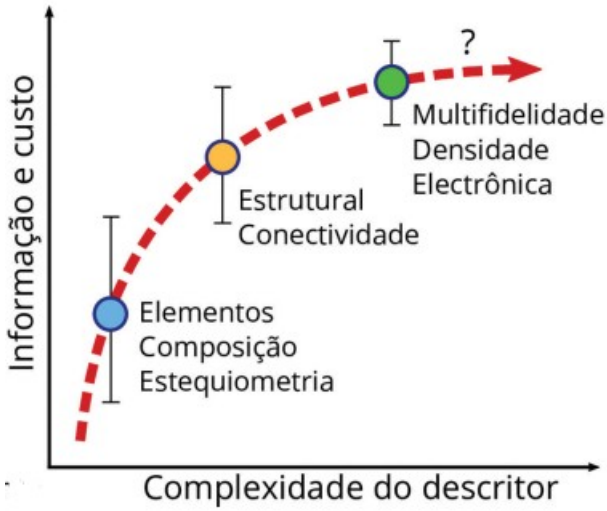
Figura 15 – Relação entre complexidade e explicabilidade de diversos modelos de aprendizagem supervisionada



SCHLEDER; FAZZIO, 2020

No geral, objetiva-se a escolha do modelo que seja o mais simples possível. Assim, a escolha do modelo é diretamente afetada pelos dados que são inseridos. Muitos atributos no âmbito de materiais possuem muita informação atribuída, tornando a análise mais custosa. Quanto mais informações inseridas nos dados de entrada, maior a complexidade requerida pelo modelo, como pode ser observado na Figura 16 (SCHLEDER; FAZZIO, 2020).

Figura 16 - Informação e custo para diversas entradas comuns em ciência dos materiais



SCHLEDER; FAZZIO, 2020

2.5 *Decision Tree*

A *decision Tree* (árvore de decisão em português) é um modelo comumente utilizado em aprendizado de máquina supervisionado, com o potencial para transformar um problema maior e complexo em diversos problemas menores e mais simples, sendo fácil de ser interpretado. Este modelo pode ser aplicado para problemas que envolvem classificação ou regressão (AHMAD, 2018).

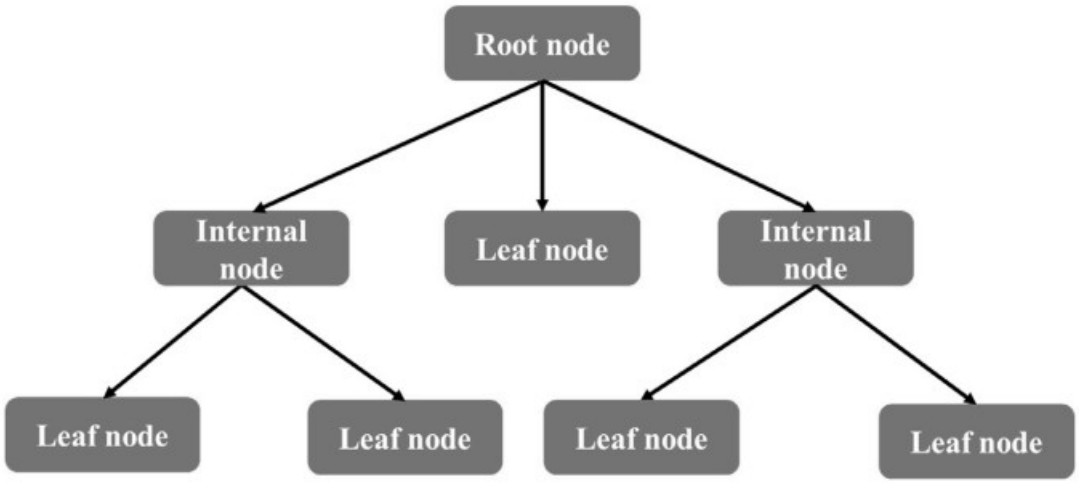
Uma *decision tree* regressiva se inicia com um nodo de uma árvore de partição recursiva, que cresce à medida que regressões múltiplas são realizadas a partir do conjunto de treinamento (AHMAD, 2018).

O conjunto de possíveis atributos inseridos no modelo são divididos em subconjuntos que não possuem união entre si, e são particionados de forma binária e recursiva até que seja obtida a menor soma dos quadrados dos erros residuais (*residual sum of squares*, RSS) possível para cada nodo gerado. Quanto maior o RSS, maior a variação atribuída ao erro (JAMES et al. 2021). A equação do RSS está representada na Equação 1.

$$RSS = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 \quad (1)$$

Com o crescimento da árvore, o modelo gera regressões mais específicas para um determinado problema, com cada nodo interno representando um conjunto de condições para se percorrer o nodo, e cada folha, como mostra a Figura 17, contendo os resultados estimados para o conjunto de entrada aplicado (JAMES et al. 2021).

Figura 17 - Esquema de *decision tree* simples



LIU et al. 2021, adaptado

Na esquematização da Figura 17, observa-se que a estrutura se assemelha a uma árvore de cabeça para baixo, com suas folhas na parte interior da estrutura e sua raiz na parte superior (JAMES et al. 2021)

Como o problema de se construír uma *decision tree* ótima se trataria de um problema computacionalmente inviável, a otimização é realizada de forma local para cada partição. Desta forma, o menor valor de RSS encontrado para cada nodo representa o melhor conjunto de condições (JAMES et al. 2021).

O processo recursivo se repete até que um critério de parada seja atingido. Esse critério é configurado utilizando hiperparâmetros inerentes às *decision trees*, que são parâmetros pré-definidos que caracterizam o modelo, como o tamanho máximo das árvores formadas. Limitantes também podem ser configurados como hiperparâmetros, como o número mínimo de folhas geradas e o minimo de partições realizadas para cada iteração (AHMAD, 2018).

Para garantir a generalização desejada no modelo de aprendizagem supervisionado, o processo de poda de folhas pode ser aplicado, além de limitar o crescimento da árvore até um determinado limite, diminuindo a complexidade da árvore e consequentemente evitando vieses de dados (AHMAD, 2018).

Para a *decision tree* aplicada para classificação, o seu funcionamento é semelhante a de uma *decision tree* para regressão, mas que utiliza a classificação que apresenta maior ocorrência nos subconjuntos gerados. Neste caso, como se trata de uma abordagem qualitativa, o RSS não pode ser utilizado, podendo assim ser substituído pelo índice Gini, que verifica a variância total de cada classe nas partições formadas (JAMES et al. 2021).

Valores menores de Gini indicam predominância de uma determinada classe no modelo, enquanto que valores maiores indicam uma maior frequência de um elemento aleatório ser identificado de forma incorreta até um máximo de 0,5. O índice Gini pode ser visualizado na Equação 2, onde $\hat{p}$ denota a probabilidade de um elemento ser classificado em uma determinada classe (JAMES et al. 2021).

$$G = \sum_{i=1}^{K} \hat{p}_i (1 - \hat{p}_i) \tag{2}$$

A *decision tree* foi aplicada neste trabalho como um dos modelos preditivos que foram treinados e validados. Além disso, a *decision tree* também serve como

base para a compreensão e composição dos modelos *Random Forest* e *Extra Trees*, que serão explanados a seguir.

2.6 Modelos Combinados

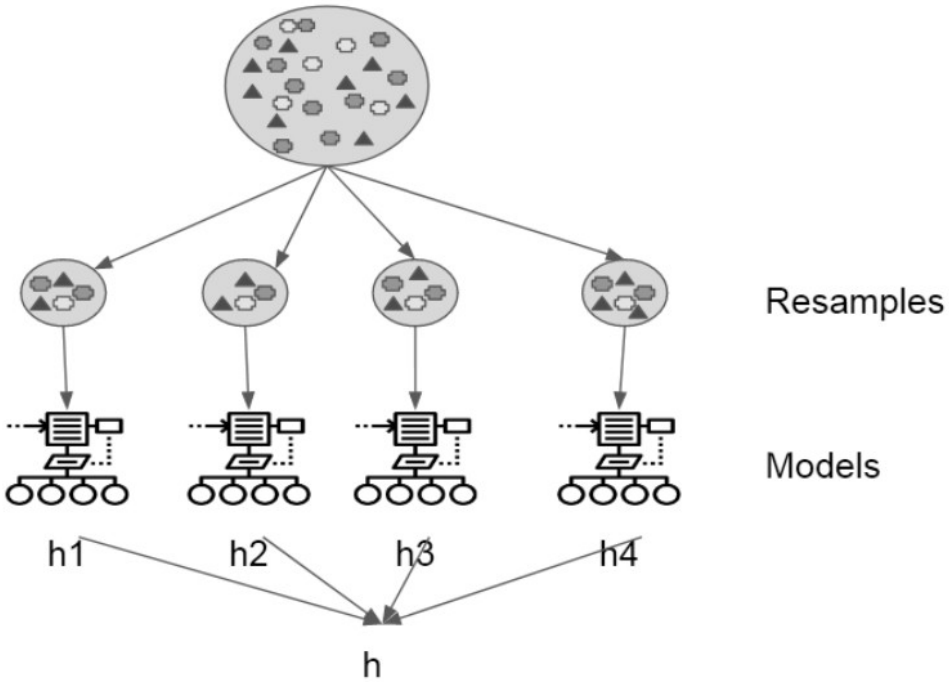
O aprendizado de máquina utilizando modelos combinados (*ensemble learning*, em inglês), baseia-se na combinação de modelos de predição mais simples para a modelagem de um sistema de aprendizagem mais complexo, constituindo na soma de suas partes e treinados para atingir o mesmo objetivo. Essa abordagem eleva a performance da predição, diminui a quantidade de viés e reduz a variância, tornando-os assim modelos mais robustos do que suas partes mais simples (MOHANA et al. 2021).

Qualquer algoritmo de aprendizado de máquina pode ser utilizado para compor um modelo combinado, como os modelos de regressão linear, redes neurais ou *decision tree*, sendo geralmente escolhido somente um modelo base para a sua composição, classificado como preditor homogêneo (ARIA et al. 2021).

Um dos métodos empregados nos modelos combinados é o *Bagging*, acrônimo para *Bootstrap Aggregating*. Este método constitui-se na agregação dos resultados de diversos modelos-base utilizando múltiplos conjuntos de amostras, e então uma média simples é aplicada (MAKARIOU et al. 2021).

O resultado é a diminuição dos vieses em comparação aos modelos-base isolados. Com a separação dos subconjuntos de dados, alguns subconjuntos podem ser utilizados diversas vezes em um mesmo processo de treinamento, enquanto que outros subconjuntos podem não ser necessariamente selecionados, tornando assim o modelo mais generalizado, com resultados de maior confiabilidade (BELGIU; DRÂGUT, 2016). O seu esquema pode ser visualizado na Figura 18.

Figura 18 - Funcionamento de um modelo combinado utilizando o método *Bagging*



PANDYA et al. 2021

Assim como os modelos que fazem parte de sua composição, os modelos combinados devem ser calibrados, com os seus hiperparâmetros ajustados e otimizados, podendo assim obter uma melhoria considerável na performance do modelo (ALI KHAN et al. 2021).

Este tipo de modelo é amplamente utilizado em diversos sistemas de aprendizagem de máquina supervisionado em várias áreas do conhecimento, desde as ciências exatas até as ciências humanas. Com estes modelos, é possível a obtenção de uma boa performance utilizando poucos dados de treinamento e de teste, mesmo com uma dimensionalidade relativamente elevada dos dados e consequentemente diversos atributos a serem explorados pelo modelo. Assim, também é possível o seu uso em engenharia de materiais, e será o tipo de modelo que será empregado na predição proposta neste trabalho (BELGIU; DRÂGUT, 2016).

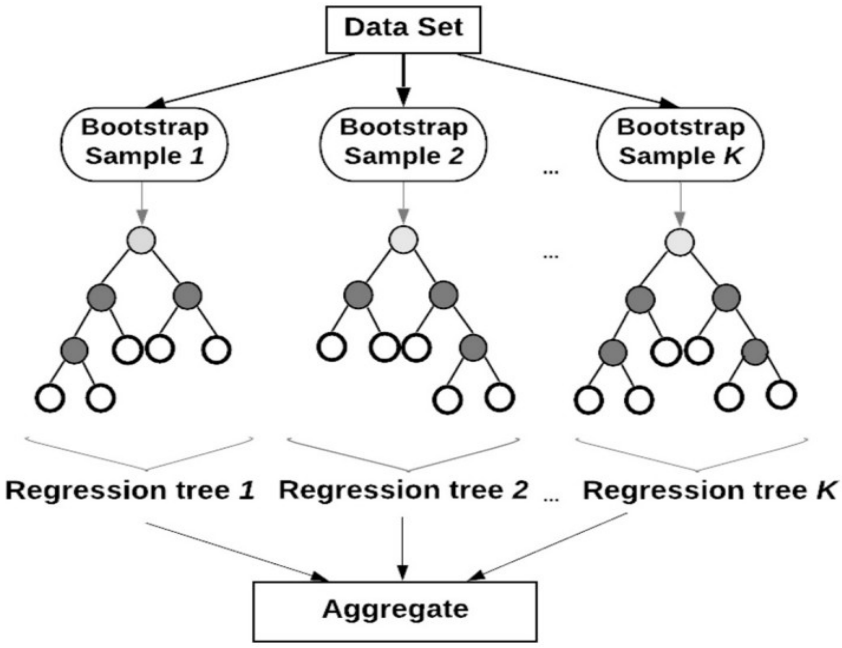
2.6.1 Random Forest

Um dos modelos combinados utilizado neste trabalho é o *Random Forest* (RF, ou florestas aleatórias, em português,). A metodologia utilizada neste modelo possui o formato “dividir para conquistar”, possuindo como modelo base as *decision trees*, que podem ser expandidas para compor um modelo combinado (BREIMAN, 2001).

Ao agregar diversas *decision trees*, a predição pode ser aprimorada significativamente, pois elementos de aleatoriedade são adicionados aos resultados, tornando-os menos correlacionados entre si (MAKARIOU et al. 2021).

O seu funcionamento se baseia nos conceitos de *Bagging*, já explicado anteriormente neste trabalho, com duas etapas adicionais: a primeira se trata da divisão do conjunto de dados de treinamento em grupos menores (*bootstrapping*) para cada *decision tree*, e o segundo é o uso de subconjuntos randômicos de variáveis para cada nó interno da árvore, que cresce no formato de partição recursiva. O seu crescimento é então encerrado quando um mínimo de observações em um determinado nodo é atingido, gerando K árvores e consequentemente K estimadores, como mostrado na Figura 19 (MAKARIOU et al. 2021).

Figura 19 - Formação do *Random Forest* e suas árvores de regressão, com os nós finais representados em círculos brancos, nós intermediários em cinza escuro e nó raiz em cinza claro



MAKARIOU et al. 2021

Assim, a redução da variância ocorre com o calculo da média das predições geradas pelas *decision trees*. Funções h são geradas na forma h_1, h_2, h_3, … h_k. A solução final está representada pela Equação 3 (BREIMAN, 2001):

$$h_{en} = \frac{1}{K} \sum_{k=1}^{K} h_k(x_n) \tag{3}$$

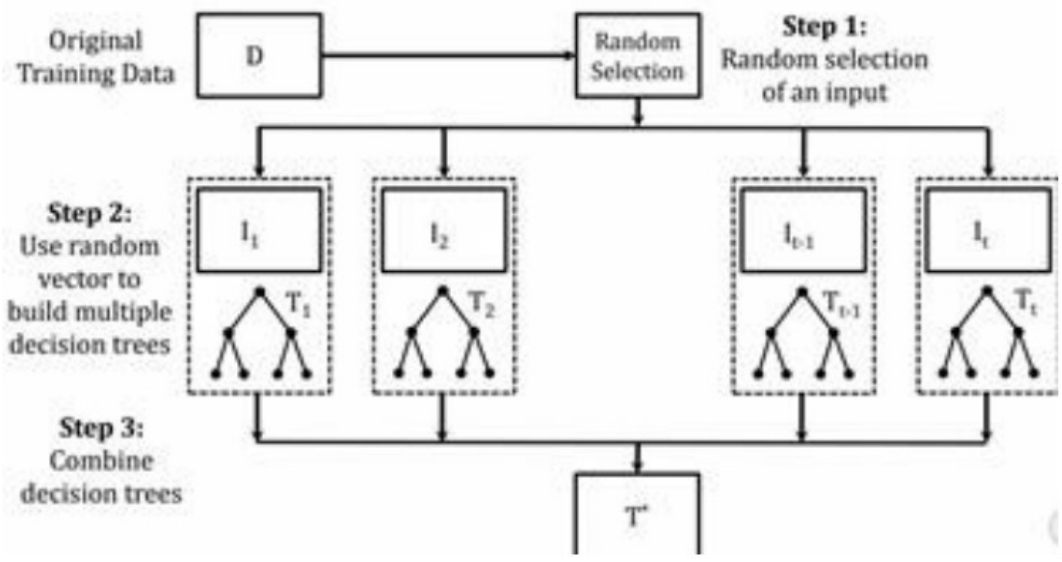
Os hiperparâmetros ajustáveis em um modelo de *Random Forest* incluem o número de árvores presentes na floresta no modelo final, profundidade máxima das árvores geradas partindo do nodo raiz, número mínimo de amostras para formação dos nodos internos, número mínimo de amostras para formação de folhas, número máximo de atributos ao se considerar a melhor partição e uso de *Bootstrapping* para a agregação dos resultados de múltiplos conjuntos de amostras (MAKARIOU et al. 2021).

2.6.2 Extra Trees

As *Extra Trees* (Extremely Randomized Trees, EX) se trata de uma versão modificada do *Random Forest*, tratando-se também de um modelo combinado. Utilizando os seus mesmos princípios. No entanto, este modelo separa os nodos das árvores de maneira aleatória, utilizando todo o conjunto de dados de seu treinamento para o treinamento de seus modelos-base. As árvores são construídas independentemente dos valores gerados através dos dados de entrada, e seus hiperparâmetros são os mesmos de um *Random Forest* (OKORO, 2021).

Assim, os atributos mais relevantes para a construção das *decision trees* são selecionados aleatoriamente. Isso eleva o grau de aleatoriedade do modelo, devendo reduzir ainda mais a variância dos dados com a diminuição da correlação entre as amostras (AHMAD, 2018). O esquema de um modelo *Extra Trees* pode ser visualizado na Figura 20.

Figura 20 - Esquema simplificado do funcionamento do modelo *Extra Trees*

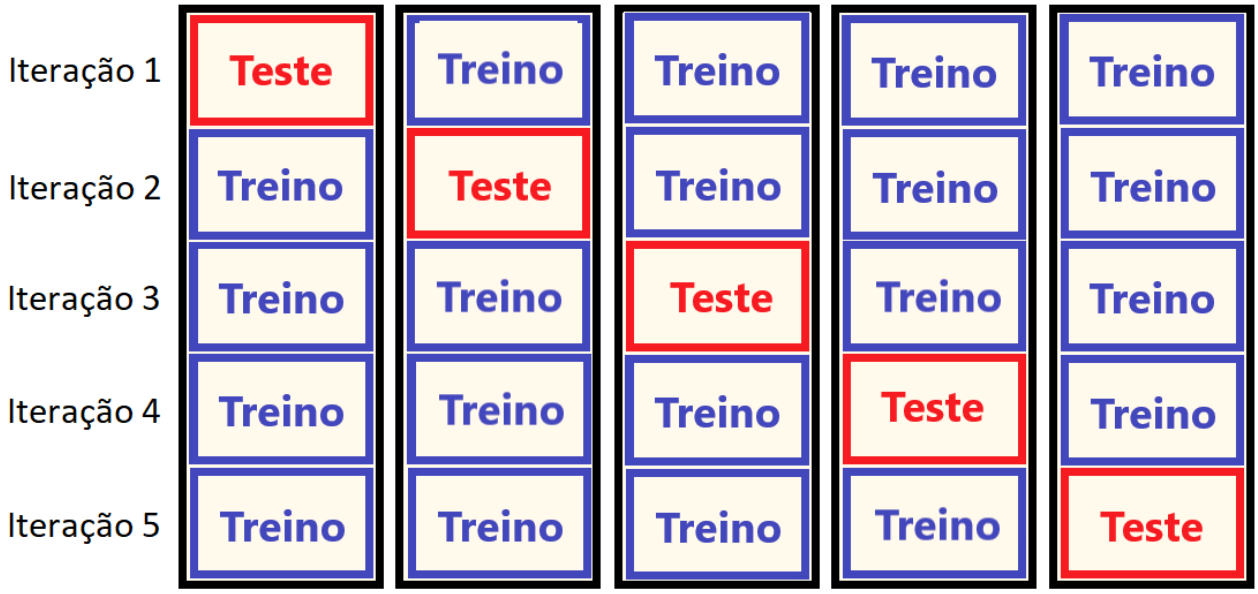


CHAKRABARTY; BISWAS, 2020

2.7 Validação dos Modelos

Para a validação dos modelos, métricas estatísticas são aplicados nos resultados do conjunto de teste para que a performance dos modelos sejam avaliadas e comparadas com os resultados reais. Além disso, as amostras devem ser particionadas em diversas amostras menores, respeitando a quantidade de dados de treinamento e de teste. Esse tratamento é chamado de validação cruzada (BELGIU; DRÂGUT, 2016), e pode ser verificada na Figura 21.

Figura 21: Esquematização da validação cruzada para 5 partições



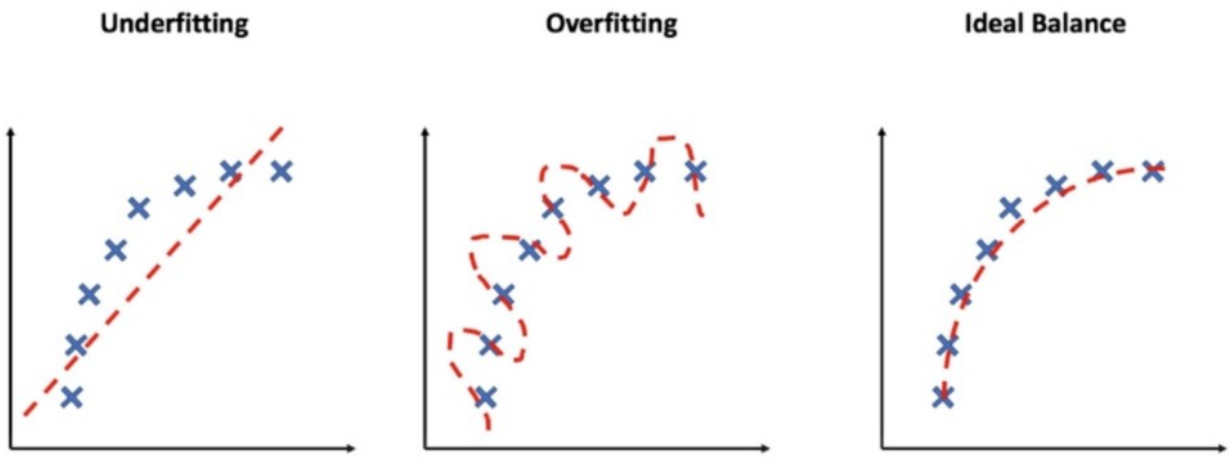
PEDREGOSA et al. 2011, adaptado

Com as amostras particionadas, o treinamento é realizado considerando o novo conjunto de amostras de treinamento e de teste. Para a avaliação final, seus resultados estatísticos são calculados com a média e desvio-padrão das métricas utilizadas em cada modelo (PEDREGOSA et al. 2011).

Essa etapa é necessária, pois ao avaliarmos os modelos de aprendizado de máquina, deve-se considerar a possibilidade de ocorrência de *underfitting* (subajuste, em português) ou *overfitting* (sobreajuste, em português) no modelo. O

underfitting ocorre quando o modelo não consegue realizar predições mesmo com dados já conhecidos, no caso os dados de treinamento, enquanto que no *overfitting*, o modelo possui desempenho ruim ao predizer o conjunto de teste, mesmo performando de forma excelente no conjunto de treinamento, impactando negativamente na capacidade de generalização do modelo (BRUCE; BRUCE, 2019). A representação do *underfitting* e *overfitting* presentes nos dados de treinamento pode ser verificado na Figura 22.

Figura 22 - Representação do u*nderfitting* e *overfitting* no aprendizado de máquina



ADDAGATIA, 2021

A importância aumenta quando considera-se trabalhar com uma quantidade limitada de dados, como no caso de análises na área de ciência e engenharia de materiais. Além disso, a validação cruzada se torna essencial para verificar a ocorrência de *overfitting*, pois durante o treinamento, os modelos considerados tendem a apresentar desempenho de 100% de acerto dado que mais da metade das árvores formadas no modelo combinado conhecem o resultado esperado para uma determinada ramificação (GAVRILOV et al., 2018).

Desta forma, a comparação entre as predições utilizando dados de treinamento e teste não pode ser utilizada para avaliar a presença de *underfitting* ou *overfitting* no caso dos modelos combinados. A validação cruzada permite que essa avaliação seja realizada, pois com a partição dos dados, se em uma partição ocorre

uma elevada taxa de acerto quando existe *overfitting* no modelo, em outra partição essa taxa tende a ser reduzida (BRUCE; BRUCE, 2019).

2.8 Métricas Estatísticas

No método de regressão, são avaliados o RMSE (*Root Mean Squared Error*), o MAE (*Mean Absolut Error*) e o R^2 (CHUGH, 2020).

O MAE indica a média da diferença absoluta entre os valores reais e os preditos pelos modelos de regressão. Esta métrica não aplica penalidades para distâncias muito elevadas dos erros em relação aos valores reais, e por isso deve ser avaliada juntamente com outras métricas estatísticas. Quando maior o MAE, maior o erro presente entre os valores, e seu cálculo está representado na Equação 4 (CHUGH, 2020).

$$MAE = \frac{1}{n}\sum_{j=1}^{n} |y_j - \hat{y}_j| \tag{4}$$

O RMSE representa a raiz quadrada do erro quadrático médio, e possibilita a medida dos desvios padrão. Por elevar o erro ao quadrado, erros de maior magnitudes presentes nas predições, como a presença de *outliers*, penalizam esta métrica, elevando ainda mais o seu valor. Assim, busca-se o menor valor possível desta métrica para se determinar a acurácia do modelo de regressão. A Equação 5 representa o RMSE (CHUGH, 2020).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{5}$$

O R^2, ou coeficiente de determinação, apresentado valores entre 0 e 1, determina a proporção da variância presente em uma variável dependente em um modelo de regressão. Quando maior o R^2, maior é a ajustabilidade do modelo linear

à amostra, ou seja, maior a dependência da variância em relação ao modelo e menor a variância residual. A Equação 6 representa o R^2 (CHUGH, 2020).

$$R^2 = 1 - \frac{Variância_{Residual}}{Variância_{Total}} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} \quad (6)$$

Além das métricas utilizadas, gráficos que comparam os resultados reais com os preditos podem ser plotados para uma melhor visualização de seu comportamento (CHUGH, 2020).

Os modelos de classificação permitem a comparação dos resultados reais dos preditos como verdadeiro ou falso. Para classificações binárias, como por exemplo, a predição da estabilidade de materiais como positiva (estável) e negativa (instável), pode-se ter quatro tipos de ocorrências: Verdadeiro positivo (*true positive*, TP), quando a classe positiva foi prevista corretamente, como a estabilidade do material; Verdadeiro negativo (*false negative*, FN), quando a classe negativa foi prevista corretamente, como a instabilidade do material; Falso positivo (*false positive*, FP), quando a classe positiva foi predita incorretamente, como um material instável classificado como um material estável; e o falso negativo (*false negative*, FN), quando a classe negativa foi predita incorretamente, como um material estável classificado como um material instável (PEDREGOSA et al. 2011).

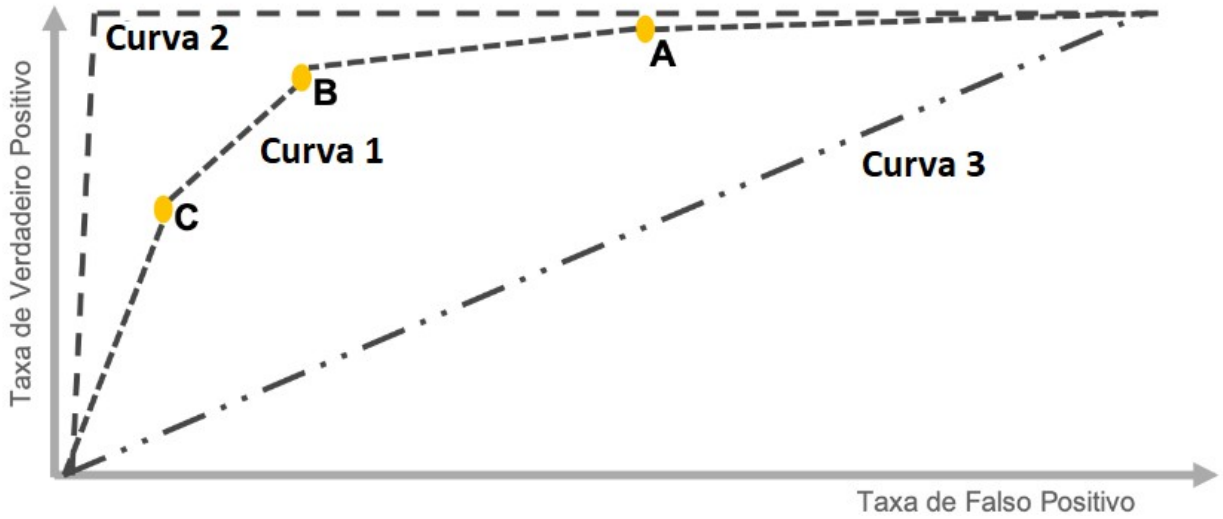
Para a determinação de cada uma destas ocorrências em um modelo de classificação, uma matriz de confusão poderá ser plotada, com cada quadrante representando cada ocorrência, seja em valores absolutos ou percentuais (PEDREGOSA et al. 2011), e está representada na Figura 23.

Figura 23. Representação da raiz de confusão, destacando cada um de seus quadrantes

PEDREGOSA et al. 2011

Curvas ROC (ou *Receiver Operating Characteristic*, em inglês) relacionam a taxa de verdadeiros positivos, sendo esta a sensibilidade do modelo, com a taxa de falsos positivos, ou seja, a sua especificidade. Com isso, é gerado uma probabilidade estimada para o verdadeiro positivo dado um ponto de corte na probabilidade estimada (BRUCE; BRUCE, 2019). Um escopo desta curva pode ser visualizado na Figura 24.

Figura 24 - Representação de uma curva ROC, com exemplos de pontos de corte denotados pelos pontos A, B e C, 30%, 50% e 70% de pontos de corte respectivamente



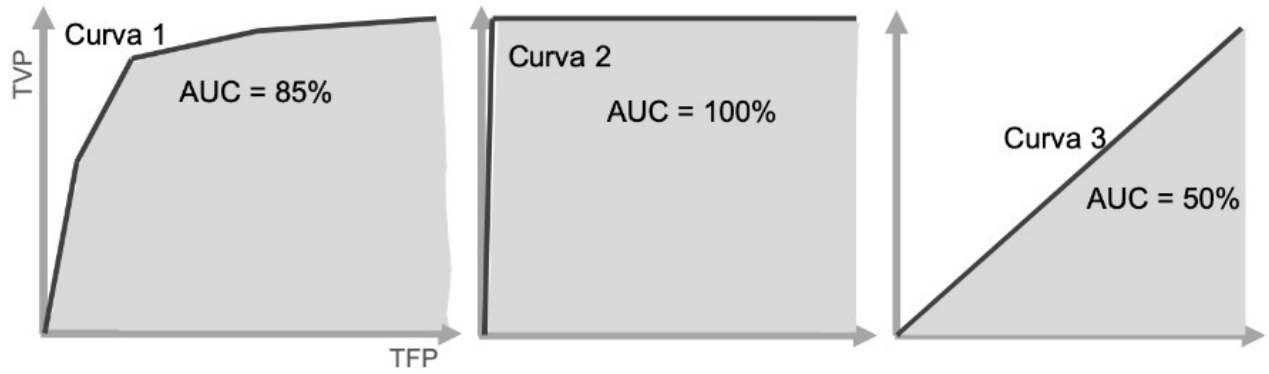
PYKES, 2020, adaptado

O ponto A representado na Figura 24 indica um ponto de corte onde aproximadamente 30% das amostras são classificadas como verdadeiro positivo, enquanto que no ponto C, aproximadamente 70% são classificados como verdadeiro positivo, apresentando neste ponto maior ocorrência de falsos positivos do que no ponto A (PYKES, 2020).

Idealmente, deseja-se que a curva seja a mais distante possível da reta indicada na curva 3 da Figura 24 e mais próxima do canto superior esquerdo, representado pela curva 1 do gráfico. Curvas próximas da reta indica um modelo que não agrega na classificação, estando próximo a um modelo que realiza chutes aleatórios, enquanto que quanto mais próximo do comportamento da curva 1, menor a ocorrência de falsos positivos a medida em que o ponto de corte aumenta (BRUCE; BRUCE, 2019).

Para uma estimativa mais precisa, é verificado a área abaixo da curva (*Area Under the Curve*, AUC), que é a área obtida pela curva ROC em porcentagem. A Figura 25 mostra o comportamento destas curvas para diferentes porcentagens de AUC.

Figura 25 - Exemplos de curvas ROC para 3 diferentes valores de AUC



PYKES, 2020

Na Figura 25, a Curva 1 indica um modelo comum, com a ocorrência da taxa de falsos positivos aumentando de acordo com o aumento da taxa de verdadeiros positivos, enquanto que a Curva 2 indica um modelo perfeito com nenhuma ocorrência de falsos positivos, e a Curva 3 indica um modelo que somente realiza chutes aleatórios na classificação (PYKES, 2020).

A acurácia determina a quantidade de acertos do modelo. Apesar de apresentar um resultado direto quanto á performance do modelo, muitas vezes esta pode ser equivocada devido à não homogeneidade das classes presentes na amostra. Se por exemplo busca-se determinar a estabilidade de um conjunto de amostras de materiais, sendo 99% materiais estáveis e 1% instável, se o modelo indica que todas as amostras são estáveis, é obtido uma acurácia de 99%, mesmo com o modelo falhando em determinar o material instável. Desta forma, é importante a utilização deste métrica com outras métricas estatísticas que auxiliam na determinação da performance dos modelos de classificação, e serão descritos adiante. A Equação 7 representa o cálculo da acurácia dos modelos (RODRIGUES, 2019).

$$Acurácia = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

A Precisão indica todas as predições positivas corretas realizadas pelo modelo. Isso evidencia a presença de falsos positivos presentes no modelo, sendo útil quando este parâmetro possui maior importância do que a presença de falsos negativos (RODRIGUES, 2019). Sua equação está descrita na Equação 8.

$$Precisão = \frac{TP}{TP + FP} \quad (8)$$

O *Recall* representa todas as predições positivas das situações da classe como valor esperado, permitindo avaliar a presença de falsos negativos no modelo, útil quando a identificação dos falsos negativos é mais importante do que os falsos positivos (RODRIGUES, 2019). A Equação 9 mostra o cálculo do *Recall*.

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

Por fim, o F_1 é a média harmônica da precisão e do *recall*, permitindo que somente uma métrica seja avaliada no lugar das duas. Com um F_1 baixo, a precisão ou o *recall* se apresenta com um valor baixo. Essa métrica auxilia a interpretação da métrica de acurácia, uma vez que se as amostras não são homogêneas e o modelo apresentar erros, mesmo que a acurácia apresente um valor elevado, o F_1 apresentará um valor reduzido mesmo com a menor ocorrência de qualquer uma das classes (RODRIGUES, 2019). Sua equação está representada na Equação 10.

$$F_1 = \frac{2 * Precisão * Recall}{Precisão + Recall} \quad (10)$$

Desta forma, podemos utilizar as métricas estatísticas de acurácia, precisão, *recall* e F_1 para compor a validação de modelos classificadores (RODRIGUES, 2019).

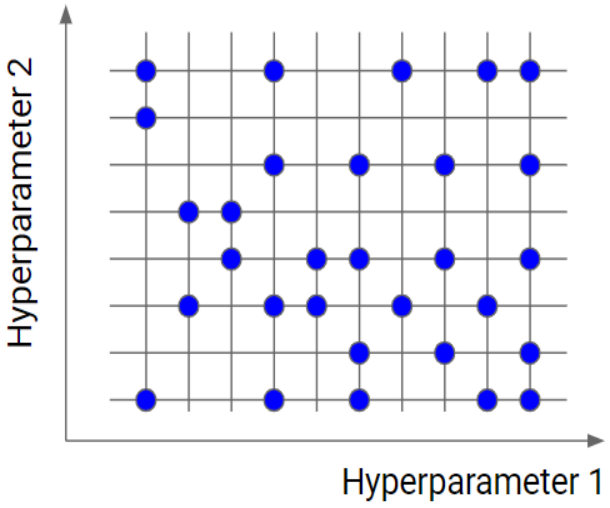
2.9 Otimização dos Modelos

A otimização dos hiperparâmetros do modelo pode ser realizada para se obter melhores resultados. Para esta otimização, um espaço de valores de hiperparâmetros é configurado no modelo, sendo assim treinado, avaliado e comparado com o treinamento de outros modelos sob diferentes configurações de hiperparâmetros, sendo assim utilizado a configuração que permite o modelo a obter os melhores resultados (SHARMA et al. 2021).

Os hiperparâmetros abrangem a configuração dos modelos de aprendizado de máquina treinados. Para os modelos combinados, estes envolvem o tamanho das árvores geradas, a extensão dos nodos, o particionamento das amostras, o limite de folhas geradas, entre outros (AL-AZZAM; SHATNAWI, 2021).

O modelo de otimização mais conhecido é o *Grid Search.* Neste modelo, uma busca em grade é realizada para diversos valores diferentes de hiperparâmetros A sua esquematização em duas dimensões pode ser visualizada na Figura 26, sendo estendida para n dimensões para um conjunto n de hiperparâmetros (PEDREGOSA et al. 2011).

Figura 26. Esquematização da otimização dos hiperparâmetros



PEDREGOSA et al. 2011

Com o modelo treinado e com os hiperparâmetros otimizados, este passa a interpretar e a compreender estes dados, produzindo uma função generalizada que poderá prever a solução de novos dados de entrada desconhecidos pelo modelo (AL-AZZAM; SHATNAWI, 2021).

2.10 Exemplo de Aplicação na Literatura

A aplicação do aprendizado de máquina na ciência e engenharia de materiais, mesmo com suas limitações, está presente na literatura, a exemplo do estudo realizado por LI et al., 2018. No artigo, os autores selecionaram um estudo de caso que visa a determinação da estabilidade termodinâmica de compostos de perovskita utilizando diversos modelos de aprendizado de máquina (LI et al. 2018).

Neste artigo, os autores consideram diversos atributos numéricos e categóricos de diferentes compostos de perovskita, e lidam com a limitação da quantidade reduzida de amostras e da quantidade não homogênea das classes de predição, no caso a energia acima da envoltória convexa (LI et al. 2018).

Os autores apresentam resultados bastante satisfatórios que permitem que os modelos realizem a predição da estabilidade do material. Contudo, diferentes abordagens e uso de diferentes tecnologias podem inferir em novos resultados, às vezes apresentando melhor desempenho (LI et al. 2018).

Desta forma, utilizando o mesmo conjunto de dados, novos modelos podem ser desenvolvidos, caracterizando o escopo deste trabalho, e assim comparados com os resultados obtidos por LI et al. 2018.

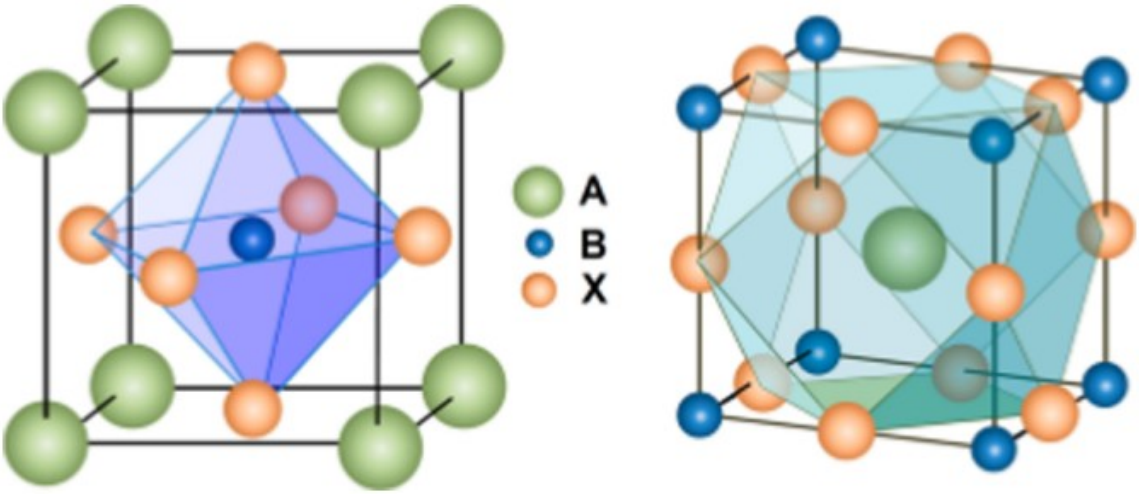
3 METODOLOGIA

3.1 Problema de Estudo

A estrutura perovskita é uma classe de cristais cuja estrutura é semelhante à do titanato de cálcio ($CaTiO_3$), com a generalização de sua fórmula para ABX_3. A e B são cátions de diferentes tamanhos, sendo o cátion A geralmente composto por metais alcalinos terrosos ou por lantanídeos, enquanto que o cátion B é geralmente composto por metais de transição 3d, 4d ou 5d. Na maioria das estruturas perovskitas, o ânion X é composto por oxigênio, mas também pode ser formado por nitrogênio ou halogênios (ROY et al. 2020).

Os cátions na posição A formam geometrias cuboctaédricas com 12 ânions X, enquanto que os cátions na posição B formam geometrias octaédricas com 6 ânions em sua estrutura. A sua representação pode ser visualizada na Figura 27 (ROY et al. 2020).

Figura 27. Estrutura perovskita ABX_3. Na esquerda, os ânions estão dispostos em geometria octaédrica, interagindo com o cátion B (BX_6), e na direita, formação da estrutura cuboctaédrica pelos ânions, que interagem com o cátion A (AX_{12})

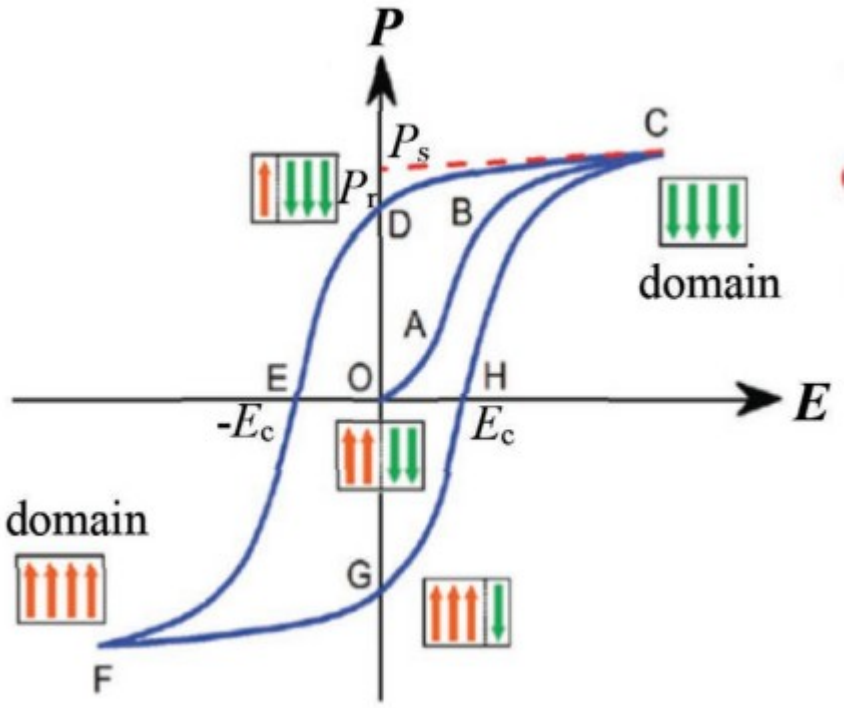


ROY et al. 2020

A perovskita apresenta assim características favoráveis que permitem diversos tipos de aplicações, entre elas a produção de células solares, capacitores, sensores e dispositivos de som, devido às suas propriedades elétricas, magnéticas e ópticas e seu caráter ferroelétrico. (ROY et al. 2020).

A ferroeletricidade presente na estrutura perovskita permite que polarizações espontâneas sejam formadas no material devido à assimetria presente em sua estrutura. A orientação dos pólos pode ser modificada ao se induzir estímulos elétricos externos no material e sua configuração pode ser mantida mesmo após a retirada deste estímulo. (SHAHROKHI et al. 2020). A curva de histerese apresentada na Figura 28 caracteriza o comportamento ferroelétrico da perovskita.

Figura 28 - Curva de Histerese P-E padrão para o comportamento ferroelétrico. Quando um estímulo E é aplicado e ultrapassa E_c, a polarização tende a mudar a sua orientação para um domínio positivo. O inverso ocorre para E abaixo de $-E_c$, cuja orientação tende a ser negativa



SHAHROKHI et al. 2020

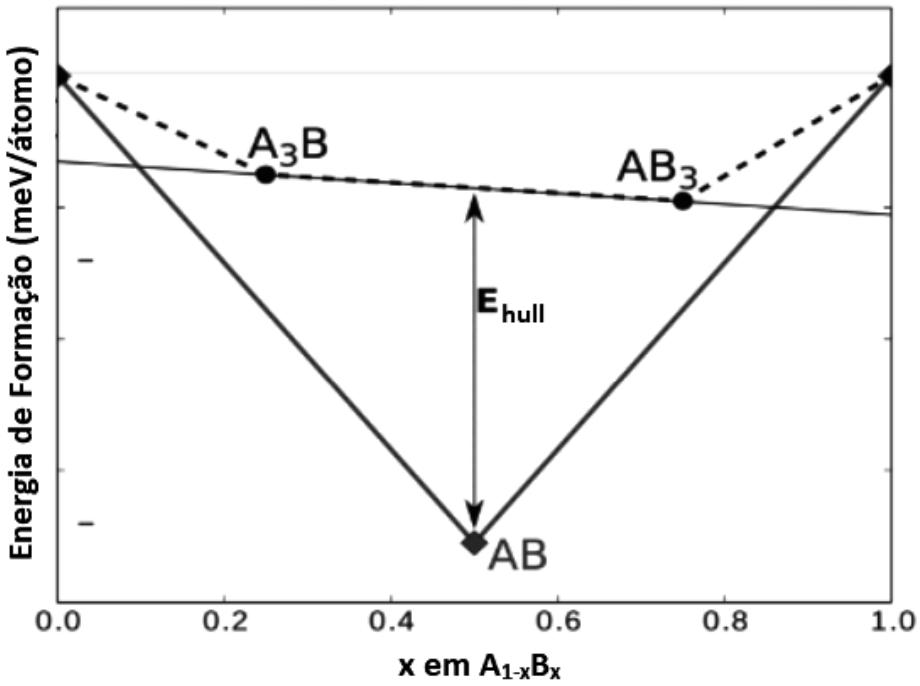
Os elementos que compõem a estrutura perovskita são um dos fatores que influenciam as suas propriedades, bem como na angulação da deformação da estrutura, afetando a geometria BX_6 formada na perovskita que determina a maior parte de suas propriedades, como a sua transição de fase, estabilidade ou capacidade de transportar elétrons (ROY et al. 2020).

A estrutura perovskita possui uma flexibilidade bastante ampla para a formação de novos materiais ao permitir diversas composições moleculares e a adição de elementos. Assim, o estudo de novas composições que formam a estrutura perovskita permite que estas propriedades sejam ajustadas de acordo com a aplicação desejada, além de abrir espaço para o estudo e desenvolvimento de novos materiais com diversas propriedades funcionais além da ferroeletricidade, como piezoeletricidade, supercondutividade, dentre outras (LI et al. 2018).

Para a formação da estrutura com as propriedades desejadas, é de suma importância a verificação da estabilidade termodinâmica da perovskita formada. Um dos parâmetros a ser considerado é a energia acima da envoltória convexa (*Energy above the convex hull*, E_{hull}). Esse parâmetro determina a energia de decomposição do composto com fases que possuem energia menor do que qualquer outra fase em combinação linear com o composto estudado (LI et al. 2018).

Fases presentes na envoltória convexa possuem estabilidade termodinâmica na presença de tais fases de menor energia, não havendo assim decomposição sob temperatura 0K. Nestas condições, o material possui maior estabilidade e maior facilidade de síntese. Valores acima da envoltória convexa, ou seja, $E_{hull} > 0$, podem se apresentar como instáveis ou metaestáveis (LI et al. 2018). Sua esquematização pode ser visualizada na Figura 29.

Figura 29 - Esquematização da Energia acima da envoltória convexa para um composto AB genérico



WEINBERGER et al. 2017, adaptado

O experimento realizado neste trabalho visa o estudo e a aplicação do aprendizado de máquina no contexto de Engenharia de Materiais, com a obtenção de um modelo que prediz a estabilidade termodinâmica de diversas estruturas perovskitas de diferentes composições através da obtenção da E_{hull} de cada material, reduzindo assim o consumo de tempo experimental ou computacional necessário para a realização deste tipo de análise.

3.2 Ferramentas Computacionais Utilizadas

A linguagem *Python* na versão 3.7 e as bibliotecas *pandas* e *scikit-learn* foram utilizadas no experimento para o pré-processamento do conjunto de dados, construção e treinamento de um modelo de *Machine Learning*, a fim de predizer a

E_{hull} de variadas amostras de compostos de perovskita, determinando assim a sua estabilidade termodinâmica.

A linguagem utilizada também permite a avaliação dos modelos construídos. A visualização gráfica dos resultados foi obtida utilizando a biblioteca *matplotlib* e foi complementada com a biblioteca *scikit-plot* para melhor apresentação dos resultados.

O notebook com a codificação e os resultados estão em anexo neste trabalho para fins de reprodução e melhor visualização de gráficos mais extensos.

Todo o processamento foi realizado utilizando um computador com processador Intel Core i5-9300, 2.40GHz com 4 núcleos e 8 processadores lógicos.

3.3 Descrição da base de dados

O conjunto de dados utilizado neste experimento foi extraído de JACOBS et al. 2018, sendo disponibilizado em repositório aberto de resultados de artigos científicos na plataforma Figshare, disponibilizada no anexo deste trabalho. Sua viabilidade como aplicação para métodos de *Machine Learning* foi estudado por LI et at. 2018, cujo trabalho foi comparado com os resultados deste estudo para fins de validação.

A base possui valores de E_{hull} dados em meV/átomo para 1929 compostos de perovskitas, variando entre 0 e 956 meV/átomo. Os valores de E_{hull} foram simulados computacionalmente através da teoria do funcional da densidade (DFT) e do uso do *Pymatgen toolkit* pelos autores do conjunto de dados (JACOBS et al. 2018). A simulação tratou os materiais em ambiente sob a temperatura de 1073K e pressão de 0,2atm, em presença de H_2 e umidade de 30%. Através das composições obtidas na simulação, a base de dados MAGPIE (*Materials Agnostic Platform for Informatics and Exploration*) e a base RTC (R*esources for Teaching Science*) foram utilizadas para obter as características dos compostos estudados (LI et at. 2018).

Além dos valores de E_{hull}, o conjunto conta com 80 atributos, sendo 71 atributos numéricos para serem utilizados no treinamento e predição, 7 atributos

categóricos, 1 classe adicional obtidas por simulação e uma coluna de catalogação, contendo a composição do material.

A classe adicional trata-se da energia de formação, ou seja, a energia liberada ou absorvida na formação do composto. Esta classe foi eliminada da análise para evitar o enviesamento e correlação errônea dos atributos com a classe a ser analisada, dado que a experimentação realizada pelo elaborador da base de dados para a obtenção desta classe foi conduzida de forma independente, não podendo assim ser utilizado como um atributo (LI et al. 2018). Essa eliminação também foi realizada por LI et at. 2018 em seu estudo, a fim de evitar vieses de dados.

Os atributos numéricos do conjunto de dados se constituem das propriedades dos dois elementos de maior composição presente no material, denotados por A-site e B-site. Alguns atributos receberam modificadores específicos, como a diferença entre os componentes dos sítios A e B, propriedade do elemento no sítio A ou no sítio B, média aritmética, média ponderada por fração atômica, o valor mínimo e o valor máximo. A Tabela 1 lista os atributos utilizados neste experimento.

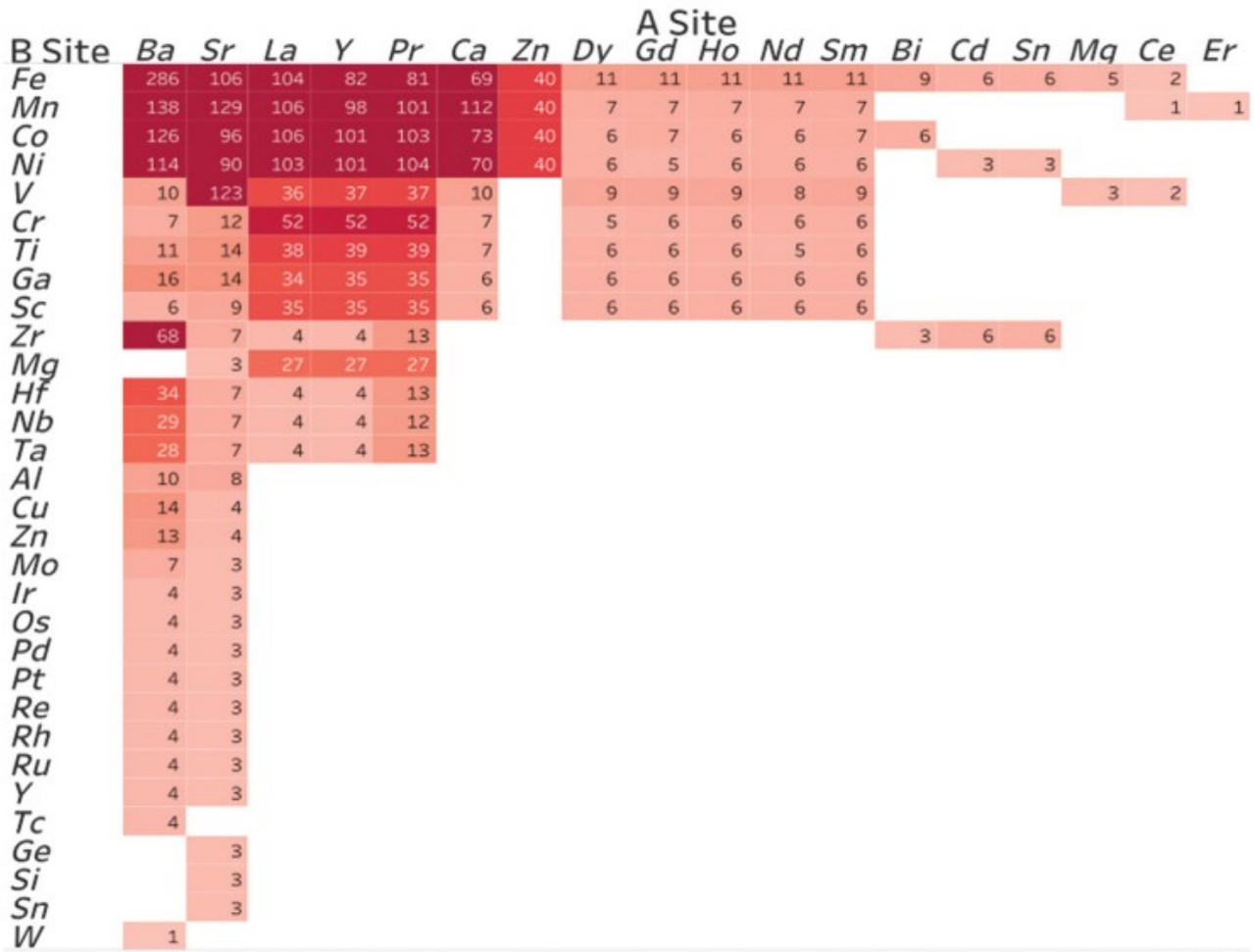
Tabela 1 - Atributos numéricos utilizados no conjunto de dados

Nome no conjunto de dados	**Definição**
NfValence_weighted	Número de valências não preenchidas na camada d.
Heat of Vaporization	Calor de Vaporização do elemento
First Ionization Potential (V)	Primeiro potencial de ionização
Second Ionization Potential (V)	Segundo potencial de ionização
Third Ionization Potential (V)	Terceiro potencial de ionização
BCCefflatcnt	BCC Fermi – Nível de Fermi do elemento calculado via DFT
MendeleevNumber	Número de Mendeleev, indicando a posição do elemento na tabela periódica
BCCenergydiff	Energia BBC (energia do elemento no OQMD – Open Quantum Materials Database, menos o BCC DFT à 0K)
ICSDVolume	Volume de acordo com o ICSD – Inorganic Crystal Structure Database
num_of_atoms	Número de Átomos
Atomic Volume (cm³/mol)	Volume Atômico
at. #	Número Atômico
thermal condutivity	Condutividade Térmica
density	Densidade
shannon_radii	Raio iônico de Shannon
covalent_radius	Raio de covalência
at. wt.	Peso Atômico
Ionization Energy (kJ/mol)	Energia de ionização em kJ/mol
Electron Affinity (kJ/mol)	Afinidade eletrônica
Atomic Radius (Å)	Raio atômico (em Armstrongs)
Specific heat capacity	Capacidade calorífica específica
Electrical conductivity	Condutividade elétrica
IsNoblegas, IsBoron, IsHalogen, IsPnictide, IsAlkali, IsMetal, IsCubic	Tipo de Elemento (gás nobre, boro, halogênio, Lantanídeo, alcalino terroso, metal, cúbico)

Os atributos categóricos contém os elementos presentes em cada um dos espaços da estrutura perovskita. Todas as composições apresentam no máximo 3 elementos para A-site e 3 elementos para B-site, com o máximo de 7 elementos em sua composição incluindo o oxigênio. A posição X-site abriga o elemento O, posição reservada para este elemento.

Assim, no conjunto existem 18 diferentes elementos presentes nos sítios A nos compostos e 31 elementos nos sítios B, como pode ser visualizado na Figura 30.

Figura 30 – Elementos presentes nas estruturas perovskita nas amostras do conjunto de dados. Cada valor representa o número de componentes com seus correspondentes nas posições A e B do conjunto



LI et al. 2018

A base também contém uma coluna contendo informações sobre a energia de formação de cada elemento, sendo obtido separadamente através de experimentos de simulação por DFT.

Por fim, a coluna com as informações da composição dos elementos constitui-se do sumário dos elementos contidos nos atributos numéricos em forma de composição química, sendo utilizado para a catalogação das análises a serem realizadas.

3.4 Pré-processamento dos dados

Com a extração dos dados, faz-se necessário a avaliação da dimensionalidade dos mesmos, a fim de eliminar atributos que apresentam irrelevância ou redundância, avaliando atributos que possuem elevada correlação entre si. Neste caso, estes atributos são eliminados da análise, reduzindo assim a dimensionalidade do conjunto de dados e garantindo a estabilidade do modelo a ser construído.

A avaliação é realizada através da verificação do coeficiente de correlação de Pearson para cada par de atributos, que variam entre 1 e -1, e determinam o quanto que ambos os atributos são dependentes linearmente entre si. Coeficientes mais próximos de 1 e -1 indicam atributos altamente correlacionadas entre si. Um par de atributos com valor 1 indica uma dependência diretamente proporcional entre si, e um par com valor -1 indica dependência inversamente proporcional entre si. O valor 0 indica independência linear entre o par de atributos considerado.

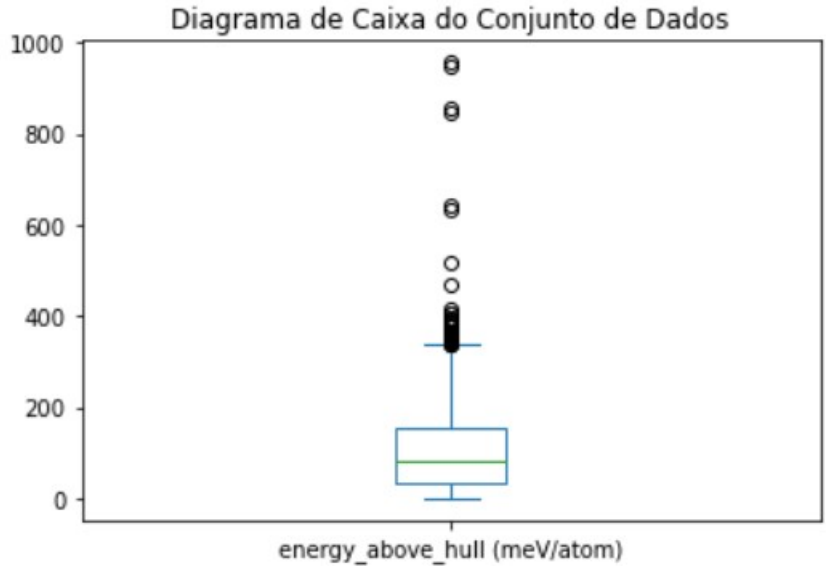
Os atributos categóricos possuem relação intrínseca com os atributos numéricos, pois a partir deles foram gerados as características das perovskitas. Além disso, a sua categorização deveria ser realizada por meio da criação de variáveis *dummy*, que definem um atributo para cada elemento presente no conjunto, mas elevariam consideravelmente a quantidade de colunas do conjunto.

Como a base possui diversos elementos químicos diferentes na composição das perovskitas, a formação das variáveis *dummy* (ou artificiais, em português) pode resultar na maldição da dimensionalidade, a qual torna a base um conjunto muito

grande de atributos, prejudicando a performance do modelo, tanto para as predições quanto para o seu tempo de execução. Por conta destes fatores, os atributos categóricos foram removidos do experimento.

Das 1929 amostras do conjunto de dados, foi realizada a eliminação dos *outliers*, com o estabelecimento de um limite de E_{hull} de 400 meV/átomo. Esse limite foi estabelecido seguindo o limite configurado por JACOBS et al., 2018. Este limite é justo, pois valores acima de 400 apresentam instabilidade muito elevada, dez vezes maior do que o limite estabelecido neste trabalho para a classificação de estabilidade do material, estudo que será explicado adiante. Além disso, pode-se observar na Figura 31 o diagrama de caixa do conjunto de dados, que mostra o espaço interquartil das quantidades de E_{hull} presentes nas amostras e a presença dos *outliers* próxima e acima de 400 meV/átomo.

Figura 31 - Diagrama de Caixa contendo o espaço interquartil de E_{hull} presente nas amostras, com o seu valor máximo e mínimo, seus quartis de 25% e 75% e mediana. Os pontos representam os *outliers* presentes no conjunto de dados.



Para a classificação, foram utilizados os mesmos procedimentos de pré-processamento do conjunto de dados, com a inclusão de uma etapa adicional de etiquetamento dos valores E_{hull}, sendo E_{hull} > 40 etiquetado com o valor 1,

representando compostos instáveis, e valores iguais ou abaixo de 40 com o valor 0, representando os compostos estáveis. Este critério é justificável, pois segundo LI et al., 2018, os compostos com menos de 36 meV/átomo apresentam estabilidade significativa, verificado por WU et al, 2013, em experimentos via DFT e validados sinteticamente em laboratório, sendo o valor arredondado para 40 meV/átomo para fins de classificação.

Assim como nos modelos treinados por LI et al., 2018, as classes do conjunto de dados não foram balanceadas, para que amostras importantes do conjunto de dados não sejam perdidas, anomalias não sejam evidenciadas no modelo e não haja distorção dos dados reais introduzidos no modelo com a inserção de dados artificiais, minimizando assim os vieses dos modelos.

3.5 Separação do Conjunto de Dados

Com o pré-processamento dos dados finalizado, é realizada a separação do conjunto de dados em dois subconjuntos: um conjunto de treinamento, utilizado para alimentar o modelo de aprendizado de máquina, e um conjunto de teste, que será utilizado para avaliar a performance do modelo.

Os dados de treinamento são conhecidos pelo modelo, e são divididos em atributos denominados como previsores, que serão utilizados na predição, e a classe, valor que será predito, no caso a E_{hull}.

Os dados de teste, possuem a mesma estrutura dos dados de treinamento, mas em contrapartida, são desconhecidos pelo modelo, que deve predizer as classes para cada liga baseando-se nas informações contidas nos previsores. Todo o conjunto é então dividido randomicamente, 80% para o conjunto de treinamento e 20% para o conjunto de teste.

Após a separação, foi realizado o escalonamento dos atributos no conjunto de treinamento, deixando a classe intacta. Neste escalonamento, os valores de cada atributo foram convertidos para uma escala de 0 a 1, mantendo a proporção entre os valores originais e permitindo a sua tradução. O cálculo de escalonamento está elucidado na Equação 11.

$$X_{escalonado} = \frac{X - X_{mínimo}}{X_{máximo} - X_{mínimo}} \quad (11)$$

Essa etapa foi realizada para que não haja atributos sob diferentes escalas, evitando a concentração ou diminuição da importância entre os atributos estudados. Feito o escalonamento do conjunto de treinamento, os parâmetros de valores mínimos e máximos foram utilizados para realizar o escalonamento do conjunto de teste, traduzindo a tratativa feita sobre os dados do conjunto de treinamento.

Assim, a separação de escalonamento garante que não haja vazamento de informações do conjunto de treinamento para o conjunto de teste, uma vez que os valores de treinamento não devem influenciar nos valores de teste.

3.6 Modelos de Aprendizado de Máquina

O estudo foi tratado com a aplicação da aprendizagem supervisionada, e foram empregados três modelos de aprendizagem de máquina: *Decision Tree*(DT), *Random Forest* (RF) e *Extra Trees* (EX), sendo os dois últimos modelos combinados. Para cada modelo, foram aplicados os métodos de regressão e classificação, visando a obtenção do algoritmo de maior explicabilidade e que apresente a acurácia mais otimizada possível.

Assim, os modelos foram treinados utilizando o conjunto de treinamento preprocessado e otimizados por pesquisa em grade, iterando diversos valores de hiperparâmetros para cada modelo e extraindo a configuração que resulta na melhor acurácia.

Para garantir a performance do modelo, foi realizada a validação cruzada dos dados, os quais os modelos foram treinados e avaliados utilizando novas porções de todo o conjunto de dados na proporção 80/20%, da forma que ao fim das iterações, todos os dados tenham sido empregados temporariamente como conjunto de teste. Os resultados das iterações foram extraídos e sua média foi determinada.

3.7 Validação do Treinamento

Após o treinamento dos modelos, a predição dos valores de E_{hull} do conjunto de teste foi realizado através da inserção do conjunto de teste no modelo treinado, que são dados desconhecidos pelo modelo e que, a partir dos valores de seus atributos, deve predizer a E_{hull} de cada amostra. Os hiperparâmetros de cada modelo foram otimizadas utilizando a busca em grade, a fim de se obter modelos que sejam adequados ao problema estudado.

Para o método de regressão, o desempenho de cada modelo foi avaliado comparando os valores de E_{hull} preditos, com os valores das classes de teste, sendo estes os valores reais de E_{hull}, para que sejam geradas as métricas MAE, RMSE e R^2. Estas métricas permitem a comparação entre os modelos de aprendizado de máquina empregados, a fim de determinar o modelo que apresente os valores de E_{hull} preditos mais próximos possíveis dos valores de E_{hull} reais. Essa relação será evidenciada com a plotagem de um gráfico de valores preditos versus valores reais.

Para a classificação, a avaliação foi realizada verificando a acurácia, precisão, *recall* e F_1 dos modelos treinados. Tais métricas foram consideradas na avaliação pela não homogeneidade das classes das amostras que compõem os modelos. Foi também gerada uma matriz de confusão para cada modelo, a fim de estudar a presença de falsos positivos e falsos negativos presentes nas predições, e a suas curvas ROC foram plotadas a fim de se determinar a AUC para cada modelo e determinar a probabilidade de ocorrência de falsos positivos no modelo.

Para ambas as regressões e classificações, a validação cruzada foi utilizada para elevar a generalização do modelo e garantir que possíveis vieses decorrentes de amostras desbalanceadas sejam identificadas.

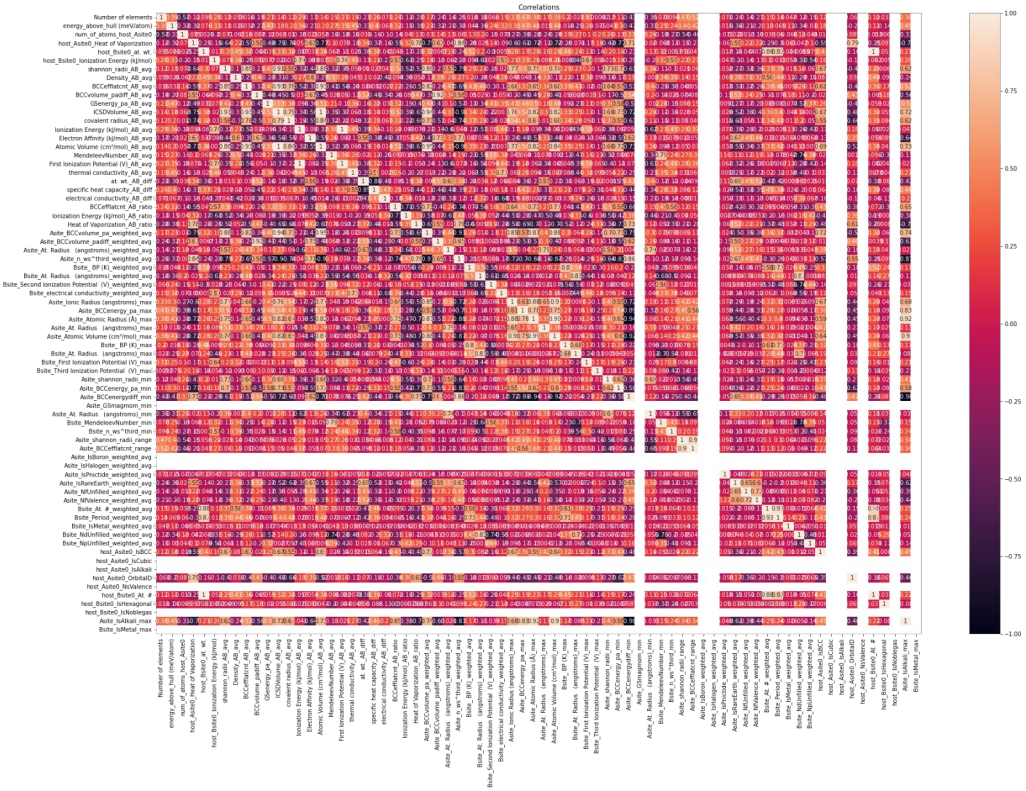
4 RESULTADOS

4.1 Pré-processamento dos Dados

Para a seleção dos atributos que compõem o modelo, foram utilizadas somente os atributos numéricos, tanto para o procedimento de regressão quanto para a classificação.

A fim de excluir os atributos numéricos altamente correlacionados, um mapa de calor contendo valores de coeficientes de Pearson para diversos pares de atributos foi gerado para este fim, sendo este visualizado na Figura 32.

Figura 32 – Mapa de calor de cada atributo analisada no experimento para a avaliação de sua correlação. Uma melhor visualização poderá ser obtida no *notebook* do experimento, disponibilizado em anexo ao trabalho.



Os coeficientes de Pearson de E_{hull} também foram analisados, a fim de verificar a relação da classe com seus atributos. Assim, o valor máximo absoluto de coeficiente de Pearson foi de 0,38 e o valor mínimo absoluto foi de 0,01. Isso indica que a classe não apresenta forte dependência linear com nenhum dos atributos considerados, e ao mesmo tempo existem atributos os quais contribuem para a variação do valor de E_{hull}, e que podem ser considerados como atributos relevantes para a sua determinação.

Assim, para o restante dos atributos, no experimento foram desconsideradas todos os atributos cujo coeficiente de Pearson absoluto seja superior à 0,90, eliminando assim um total de 12 atributos. Estes atributos podem ser visualizados na Tabela 2, com os seus valores máximos absolutos de coeficiente de Pearson.

Tabela 2. Valores máximos absolutos do coeficiente de Pearson para os atributos que foram removidos do modelo.

Atributo	**Coeficiente de Pearson**
Asite_Atomic Volume (cm³/mol)_max	0,99
Atomic Volume (cm³/mol)_AB_avg	0,99
First Ionization Potential (V)_AB_avg	0,99
host_Bsite0_At. #	0,99
Asite_IsAlkali_max	0,98
Asite_BCCvolume_pa_weighted_avg	0,95
Asite_BCCenergydiff_min	0,94
ICSDVolume_AB_avg	0,93
Asite_BCCvolume_padiff_weighted_avg	0,92
Bsite_Period_weighted_avg	0,92
Asite_n_ws^third_weighted_avg	0,91
Asite_BCCefflatcnt_range	0,90

No processo de remoção de *outliers*, foram removidas 11 amostras do conjunto por apresentarem valores de E_{hull} acima de 400 meV/átomo, restando 1918 amostras para compor o modelo. Destas 1918 amostras, 80% foram separadas

aleatóriamente para serem utilizadas como conjunto de treinamento dos modelos, totalizando 1534 amostras.

Os 20% restantes foram separados para compor o conjunto de teste, totalizando 384 amostras para teste. Os valores de E_{hull} foram atribuídos para cada conjunto de atributos, sendo esta variável a classe do conjunto.

4.2 Modelos Regressores

Os hiperparâmetros dos modelos regressores foram otimizados utilizando o método de busca por grade, testando diversos valores de hiperparâmetros até que seja obtido a melhor eficácia possível dos modelos.

A Tabela 3 mostra os hiperparâmetros que foram aplicados para cada modelo após a otimização.

Tabela 3 – Hiperparâmetros dos modelos após otimização por busca em grade

Hiperparâmetros	**DT**	**RF**	**EX**
Número de árvores	-	2400	1200
Profundidade máxima	20	20	20
Nº mínimo de amostras para formação dos nodos internos	10	2	2
Nº mínimo de amostras para formação de folhas	4	1	1
Bootstrap	-	True	False

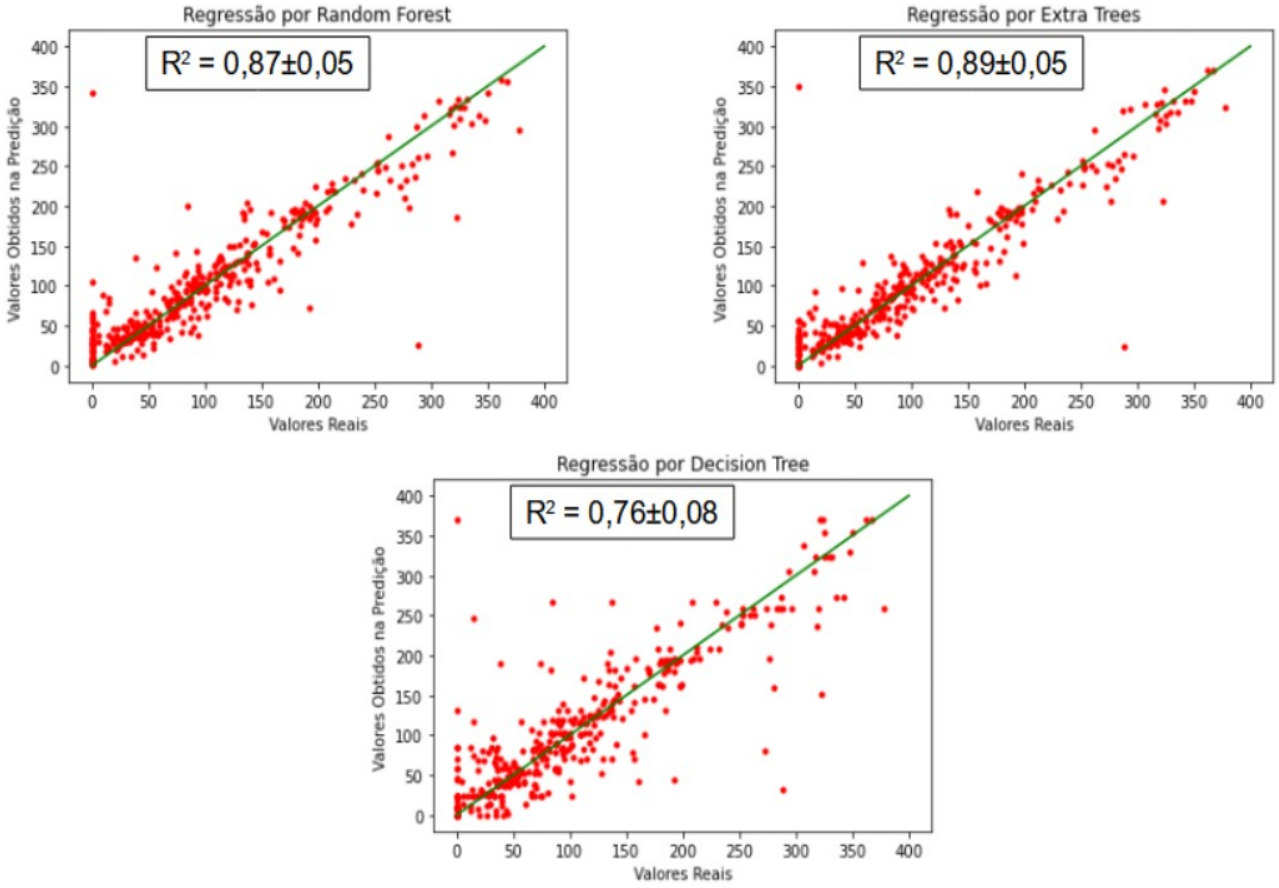
Para garantir que os modelos não estão enviesados para um conjunto de treinamento fixo, foi aplicada a validação cruzada com 5 cortes no conjunto para cada modelo regressor. Os resultados são apresentados na Tabela 4.

Tabela 4 – Resultados estatísticos dos modelos regressores após validação cruzada.

Modelo	**MAE (meV/átomo)**	**RMSE (meV/átomo)**	**R^2**
DT	25,80 ± 3,01	41,37 ± 6,53	0,76 ± 0,08
RF	18,98 ± 2,11	30,53 ± 5,26	0,87 ± 0,05
EX	16,26 ± 2,20	27,64 ± 5,86	0,89 ± 0,05

Através dos hiperparâmetros otimizados, foi possível a comparação dos valores de E_{hull} reais do conjunto de teste com os valores preditos pelos modelos regressores. A Figura 33 apresenta essa comparação para cada modelo estudado.

Figura 33 – Relação de valores reais com os valores preditos pelos modelos no conjunto de teste



Valores próximos da reta indicam valores preditos próximos dos valores reais, enquanto que os valores apresentados acima da reta mostram resultados preditos acima do valor real, e abaixo da reta indicam valores preditos abaixo do valor real.

4.3 Modelos Classificadores

Com a catalogação dos valores de E_{hull}, foi obtido assim 1351 amostras de compostos instáveis e 569 compostos estáveis, destacando assim a não homogeneidade das classes presentes nas amostras no conjunto de dados.

Como os valores numéricos de E_{hull} após catalogação não foram necessários para o treinamento dos modelos, essa coluna foi descartada, sendo substituída pelos valores binários e compondo as classes dos conjuntos de treinamento e teste.

Os modelos utilizados foram os mesmos utilizados na regressão, mas com a abordagem de classificação, sendo assim realizado um novo *tuning* dos hiperparâmetros via busca por grade. Estes hiperparâmetros estão listados na Tabela 5.

Tabela 5 - Hiperparâmetros dos modelos após otimização por busca em grade

Hiperparâmetros	**DT**	**RF**	**EX**
Número de árvores	-	600	1200
Profundidade máxima	20	20	20
Nº mínimo de amostras para formação dos nodos internos	10	2	2
Nº mínimo de amostras para formação de folhas	4	1	1
Bootstrap	-	True	False

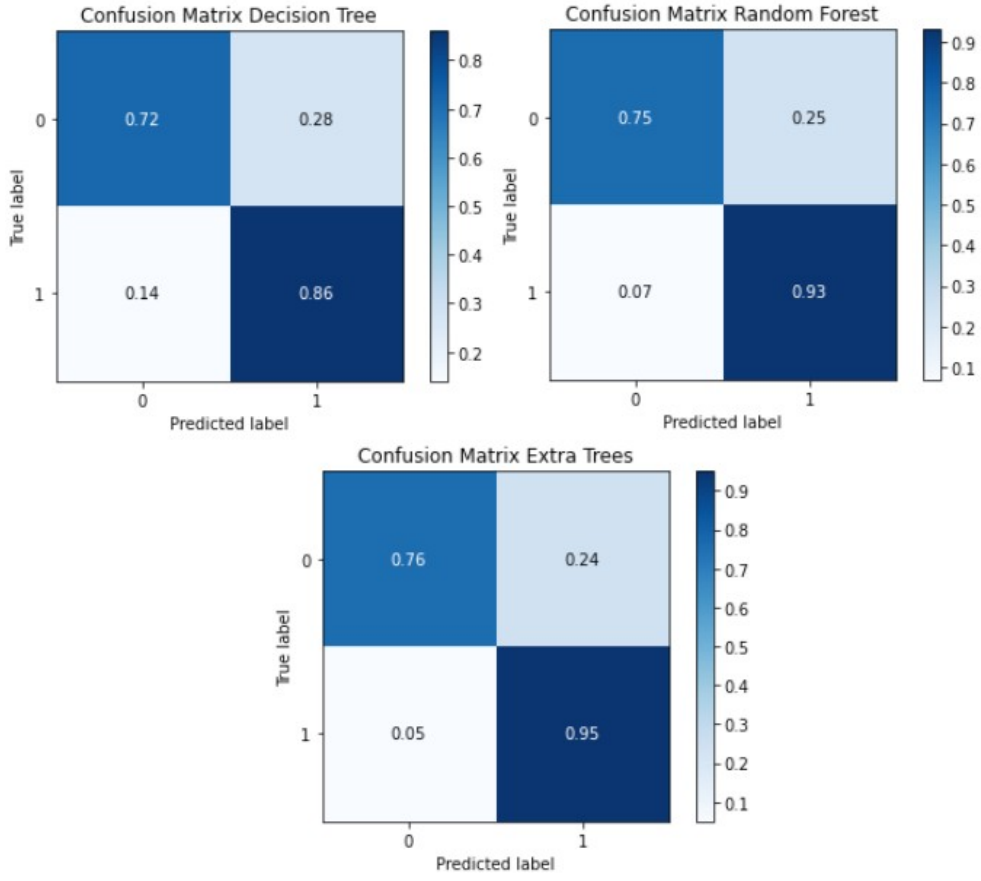
Os valores estatísticos obtidos para a classificação após a validação cruzada podem ser visualizados na Tabela 6.

Tabela 6 - Resultados estatísticos dos modelos classificadores após validação cruzada

Modelo	**Precisão**	***Recall***	**F_1**	**Acurácia**
DT	0,89 ± 0,04	0,88 ± 0,03	0,88 ± 0,03	0,83 ± 0,04
RF	0,92 ± 0,03	0,94 ± 0,02	0,93 ± 0,02	0,91 ± 0,03
EX	0,92 ± 0,04	0,95 ± 0,02	0,94 ± 0,01	0,91 ± 0,03

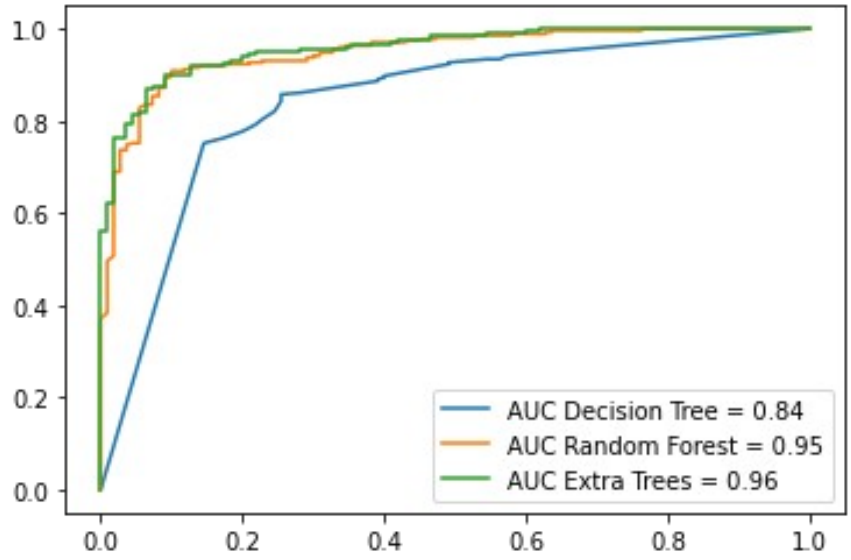
A Figura 34 mostra as matrizes de confusão geradas pelos modelos de classificação, normalizados em valores entre 0 e 1 para melhor visualização do comportamento da ocorrência dos falsos positivos e falsos negativos.

Figura 34 - Matrizes de confusão para os modelos estudados. O label 0 indica materiais classificados como estáveis, enquanto que o label 1 indica materiais instáveis.



Por fim, as curvas ROC também foram plotadas para que seja avaliada a performance do modelo comparada, considerando a taxa de verdadeiros positivos com a taxa de falsos positivos. A Figura 35 mostra as curvas geradas, bem como os valores AUC.

Figura 35 - Curvas ROC geradas por modelo de classificação estudado utilizando o conjunto de teste, com os valores de AUC em porcentagem.



4.4 Tempo de Processamento

O tempo de processamento de cada um dos modelos estudados pode ser observado na Tabela 7. O valor foi calculado realizando 20 execuções em uma amostra de teste aleatória e obtendo a sua média.

Tabela 7 - Tempo médio para a execução das predições pelos modelos treinados.

	Decision Tree	***Random Forest***	***Extra Trees***
Regressão	0,0009s	0,1540s	0,0740s
Classificação	0,0010s	0,0560s	0,0960s

5 DISCUSSÃO

5.1 Implicações na Ciência e Engenharia de Materiais

A aplicação do aprendizado de máquina utilizando *decision tree* e modelos combinados na ciência e engenharia de materiais neste estudo de caso se mostrou possível devido às simulações computacionais realizadas por JACOBS et al., 2018, e que por sua vez podem ser validadas e comparadas por modelagem matemática e experimentações em laboratório.

Com isso, verifica-se a dependência do quarto paradigma da cîencia com os seus antecessores. Como já discutido neste trabalho, o estudo de materiais orientado a dados exige que os dados históricos obtidos sejam confiáveis e com informações suficientes que permitam correlacionar os atributos com as classes as quais se deseja predizer.

Neste estudo, limita-se a 1929 amostras para compor todo o processo de aprendizado de máquina. Destas amostras, ainda foi necessária a remoção de *outliers* e de atributos linearmente dependentes entre si, além da separação de 20% do conjunto para compor o conjunto de teste, diminuindo ainda mais a quantidade de informação no conjunto de treinamento do modelo.

O custo elevado para a obtenção destas amostras impactou diretamente no desenvolvimento dos modelos, com a consequente limitação em sua disponibilidade. A não homogeneidade das classes, reflexo da dificuldade de disponibilização dos dados, também influenciou no tratamento e escolha do modelo.

Assim, a presença de *underfitting* ou *overfitting* tendem a ser mais evidentes nestas condições, e os tratamentos descritos na metodologia deste trabalho visaram minimizar a ocorrência tanto de *underfitting* como de *overfitting*, como a remoção dos atributos de elevada correlação, a análise de *outliers*, a validação cruzada e a otimização por busca em grade.

Estes tratamentos cumprem o objetivo de elevar a generalização do modelo e de avaliar corretamente a sua presença. O uso dos modelos combinados também

contribuiu para evitar o *overfitting*, pela adição de uma quantidade maior de eventos aleatórios no treinamento dos modelos.

5.2 Regressão

O estudo do comportamento dos modelos envolvendo *decision tree*, *random forest* e *extra trees* permitiu que a estabilidade de diversos compostos de perovskita, mesmo que não conhecidos pelos modelos, sejam preditos através da regressão ou da classificação.

Na regressão, para a *decision tree*, ao verificar o gráfico que relaciona os valores preditos com os valores reais, nota-se uma maior dispersão dos pontos em relação aos outros modelos, indicando a dificuldade encontrada pelo modelo em predizer os valores de E_{hull} de forma precisa. Este resultado é esperado, uma vez que esse tipo de modelo é mais simplificado e rudimentar do que os modelos combinados, que são uma versão mais robusta da utilização de árvores para a predição de estabilidade de compostos de perovskita.

A regressão por *Random Forest* apresentou dispersão dos pontos semelhantes com o *Extra Trees*. As pequenas diferenças encontradas podem ser justificadas com o comportamento diferente no tráfego dos nodos das árvores no processo de predição com melhores resultados, verificado tanto pela análise gráfica quanto pela análise dos valores estatísticos obtidos na validação do modelo. Para ambos os modelos, os pontos apresentara-se menos dispersos em relação à *decision tree*.

Com a visualização do gráfico, justifica-se a aplicação dos modelos combinados para a predição da E_{hull} dos compostos de perovskita. Pode-se confirmar com a análise de seus valores estatísticos, no qual os modelos combinados apresentam melhores resultados em comparação com a *decision tree*, com menores valores de MAE e RMSE, e maior valor de R^2.

Para os três modelos, o RMSE possui valores maiores do que o MAE, devido ao fator quadrático do RMSE que penaliza a ocorrência de *outliers,* indicando que a presença destes possui forte influência no treinamento do modelo, ainda que tenha

sido realizado o tratamento dos *outliers* no pré-processamento dos dados. A presença destes *outliers* é reflexo das limitações na quantidade de amostras no âmbito da ciência de materiais, discutidas anteriormente. Assim, o aumento na quantidade de amostras pode reduzir significativamente o RMSE, de forma a se assemelhar com o MAE.

O R^2 sumariza a dependência e proximidade dos resultados preditos com os reais, com o *Extra Trees* apresentando o melhor resultado dentre os modelos estudados.

Pode-se observar da experimentação o aumento da generalização do modelo e a redução do *overfitting* ao utilizar modelos combinados, tendo como base a *decision tree*. Essa observação é justificada verificando o desvio-padrão das métricas, obtidas através da validação cruzada. A *decision tree* apresentou um desvio padrão maior do que os modelos combinados para as três métricas estudadas. Isso significa que ocorre maior aprendizagem para determinadas partições de treinamento em detrimento a outras partições, indicando a perda de generalização neste modelo.

Pode-se assim comparar os resultados obtidos por estes modelos com os modelos de regressão gerados por LI et al. 2018, o qual utilizou procedimentos semelhantes para o treinamento de diversos modelos de aprendizado de máquina, incluindo o *Extra Trees*. A relação do *Extra Trees* utilizado em comparação com o *Extra Trees* obtido por LI et al. 2018 pode ser observado na Tabela 8.

Tabela 8 - Resultados estatísticos dos melhores modelos obtidos por LI et al. 2018 comparados com o modelo de *Extra Trees* obtido neste trabalho.

Métrica	**Extra Trees (LI et al, 2018)**	**Extra Trees (este trabalho)**
R^2	0,888 ± 0,054	0,893 ± 0,050
RMSE (meV/átomo)	29,4 ± 7,3	27,6 ± 5,9
MAE (meV/átomo)	16,0 ± 2,2	16,2 ± 2,2

Dentre os modelos de regressão estudados pelo artigo, o *Extra Trees* foi o que melhor performou em relação aos modelos treinados pelos autores do artigo. Destes resultados, pode-se notar um valor menor de RMSE do Extra Tree obtido neste trabalho em comparação com o menor valor de RMSE obtido por LI et al, 2018, indicando a obtenção de um modelo um pouco mais preciso, dado a semelhança de R^2 e MAE utilizando o mesmo conjunto de dados de compostos de perovskita.

A diferença pode ser justificada pela escolha dos hiperparâmetros envolvidos em cada modelo, e pela decisão de não utilizar os dados categóricos neste experimento, uma vez que estes contribuem para a maldição da dimensionalidade do modelo e contribuem com a redundância entre os atributos, possuindo relações intrínsecas com os dados numéricos.

Com estes resultados, para aplicações práticas, os modelos regressores podem ser utilizados como um sistema de apoio à decisão quando faz-se necessário uma avaliação precisa dos valores de E_{hull} para a determinação da estabilidade termodinâmica da perovskita, como no estudo de novos materiais ou para a seleção de seus aditivos.

O erro obtido nestes modelos deve ser considerado no projeto, estimando uma tolerância de erro baseado no MAE e RMSE obtidos em sua avaliação estatística. Além disso, deve ser verificada a possibilidade de valores divergentes quando materiais que se enquadram como *outliers* são inseridos no modelo para predição de sua estabilidade. Mais uma vez, para amenizar estes casos, novas amostras que englobam estes *outliers* são necessários para compor o treinamento do modelo.

5.3 Classificação

Nos modelos de classificação, ao trabalhar com valores categóricos binários, o detalhe dos valores de E_{hull} são perdidos, tornando a análise quantitativa em uma análise qualitativa ao categorizar as classes dos compostos de perovskita como

estáveis ou instáveis. Essa análise também permite a verificação simplificada da ocorrência de falsos positivos e falsos negativos.

Para o modelo *Extra Trees*, verifica-se através da matriz de confusão, que a taxa de verdadeiros negativos, isto é, os materiais instáveis que realmente foram classificados como instáveis, é maior do que a taxa de verdadeiros positivos, materiais classificados como estáveis. Isso indica que o modelo possui maior facilidade em predizer materiais instáveis do que os estáveis.

Essa facilidade implica em redução de custos e tempo na tentativa de síntese em laboratório. Por outro lado, a maior dificuldade em predizer materiais estáveis presume perdas de compostos em potencial que poderiam ser estudados, mas que seriam descartados. Isso reforça a necessidade de tomar o modelo como um sistema de apoio à decisão e não deve considerar a predição como conclusão final para fins de classificação da estabilidade dos materiais.

Através da matriz de confusão também é possível verificar que este modelo apresenta maior taxa de falsos negativos, ou seja, quando o material é estável mas foi classificado como instável pelo modelo, e menor taxa de falsos positivos, quando o material é instável mas foi classificado como estável.

Essa relação deve ser considerada na aplicação prática do modelo, uma vez que o estudo deve definir o impacto que um falso positivo ou um falso negativo acarreta em sua aplicação.

Esse comportamento ocorre nos três modelos treinados, sendo a *decision tree* a que possui maior taxa de falsos positivos e negativos, e o *Extra Trees* sendo o modelo com menor taxa de falsos positivos e negativos.

A acurácia medida para cada um dos modelos permite analisar a performance de cada um dos modelos. No entanto, como o conjunto de dados utilizado não possui distribuição de classes homogênea, faz-se necessário avaliar as outras métricas analizadas, a fim de determinar o modelo que mais se adequa ao comportamento esperado.

As métricas de precisão e *recall* obtidos nos modelos de classificação indicam que os modelos combinados obtiveram maior confiabilidade em evitar a ocorrência

de falsos negativos do que falsos positivos, devido ao seu maior valor de *recall*. Em contra-partida, a *decision tree* apresenta maior confiança para evitar falsos positivos.

Ainda assim, a *decision tree* possui performance menor em relação aos modelos combinados, o que é verificado na relação entre a precisão e o *recall*, no caso o F_1. Nesta análise, verifica-se que o F_1 da *decision tree* é menor do que a dos modelos combinados, com o maior valor atribuído ao *Extra Trees*, sendo este o que apresenta maior performance entre os demais modelos.

A ocorrência de falsos positivos em relação ao aumento dos verdadeiros positivos foi verificado através das curvas ROC dos três modelos, evidenciando a melhor performance do *Extra Trees* dado a maior proximidade da curva no canto superior esquerdo, indicado numericamente pelo seu valor de AUC elevado. Nesta análise, verifica-se que o seu comportamento não se assemelha a um modelo que realiza chutes aleatórios no processo de classificação, com menor probabilidade de ocorrência de falsos positivos para cada verdadeiro positivo estimado. Em contrapartida, a *decision tree* apresentou uma curva mais próxima da reta, com menor valor de AUC em relação aos modelos combinados, possuindo assim maior chance de ocorrência falsos positivos.

A presença de *overfitting* também é verificada nos modelos de classificação. Os modelos combinados apresentaram no geral desvios-padrão menores do que a *decision tree*, mantendo uma taxa elevada de acurácia, indicando a redução do *overfitting* ao se utilizar os modelos combinados.

Assim, é possível comparar o melhor modelo de classificação obtido neste trabalho com o melhor modelo de classificação obtido por LI et al., 2018. O resultado pode ser visualizado na Tabela 9.

Tabela 9 - Comparação do *Extra Trees* deste trabalho em relação ao *Extra Trees* obtido por LI et al, 2018

Métrica	***Extra Trees* (Li et al. 2018)**	***Extra Trees* (este trabalho)**
Acurácia	0,93 ± 0,02	0,91 ± 0,03
Precisão	0,89 ± 0,07	0,92 ± 0,04
Recall	0,87 ± 0,05	0,95 ± 0,02
F_1	0,88 ± 0,03	0,94 ± 0,03

Observa-se que os modelos de classificação deste trabalho performaram de forma menos eficaz do que os modelos obtidos por LI et al. 2018 em termos de acurácia, mas se destacando nos outros resultados estatísticos, como na precisão, *recall* e F_1. Essa diferença pode ser justificada pelos mesmos motivos descritos pelo método de regressão, no caso a diferença do tratamento dos dados e dos hiperparâmetros utilizados.

A acurácia do *Extra Trees* deste trabalho permanece dentro dos limites dos valores dos desvios-padrão apresentados pelos modelos de LI et al. 2018. Assim, para se obter um valor mais preciso de E_{hull} através dos modelos deste trabalho, o *Extra Trees* abordado como um problema de regressão é o mais apropriado em comparação com os outros modelos estudados.

Essa abordagem pode ser útil para os casos em que os detalhes numéricos de E_{hull} não sejam estritamente necessários, sendo a catalogação da estabilidade suficiente para a análise do novo composto. Além disso, deve-se considerar a ocorrência de materiais meta-estáveis dentro do conjunto dos materiais classificados como estáveis.

5.4 Resumo

Em síntese, considerando a quantidade limitada de dados presentes neste experimento que reflete a problemática do paradigma da ciência orientada a dados presente na engenharia de materiais, os três modelos estudados apresentaram vantagens e desvantagens, as quais devem ser avaliadas de acordo com as necessidades de estudo e aplicação.

A *decision tree*, embora apresente resultados de acurácia inferiores aos outros modelos, tanto no método de regressão quanto por classificação, apresentou maior velocidade de predição, sendo assim aplicável quando a velocidade é um parâmetro chave maior do que a acurácia. A explicabilidade da *decision tree* também é favorecida, uma vez que dentre os modelos estudados, trata-se do mais simples entre eles, sendo mais fácil a compreensão dos fatores que levaram o modelo a realizar uma determinada predição.

O *Random Forest*, por ser um dos tipos de modelos combinados mais tradicionais e por apresentar resultados bastante satisfatórios, pode ser aplicável para a maioria dos problemas que envolvem aprendizado de máquina. A sua desvantagem se dá no maior tempo de predição comparado com os outros modelos estudados, como verificado nos experimentos realizados.

O *Extra Trees*, por sua vez, apresentou uma performance igual ou superior aos outros modelos estudados, sem prejuízo em seu tempo de execução, sendo aplicável em estudos que permitem maior confiança ao modelo ou maior rigor técnico para a compreensão do comportamento do modelo.

Para todos os modelos, é notório o baixo tempo de execução das predições, verificado pelo tempo de execução dos modelos para a predição de uma única amostra. Comparados com as simulações computacionais via DFT, que levam em torno de 8 a 9 horas para processar cada composto de perovskita (LI et al. 2018), o tempo de execução na ordem de milissegundos para a predição dos modelos estudados tornam o método extremamente eficiente para ser aplicado na ciência e engenharia de materiais, com a consequente economia em processamento e energia e redução de custos em comparação com a simulação.

6 CONSIDERAÇÕES FINAIS

As experimentações de predição da E_{hull} utilizando modelos de regressão e classificação foram possíveis de serem realizadas graças ao conjunto de dados gerados e validados por etapas anteriores ao quarto paradigma da ciência. A experimentação em laboratório dos fenômenos presentes nos compostos de perovskita, a modelagem através de fórmulas matemáticas que refletem o comportamento das variáveis envolvidas na atribuição da E_{hull} e a simulação computacional tornaram possível a generalização a fim de predizer o comportamento dos compostos estudados de acordo com a variação de cada atributo e de atribuir esse comportamento para novos compostos.

Assim, o estudo cumpre o seu objetivo de avaliar a estabilidade termodinâmica da perovskita como estudo de caso para a aplicação do paradigma da ciência orientado a dados utilizando aprendizado de máquina, bem como de gerar modelos preditivos baseados em *decision tree* e suas variantes, considerando as dificuldades comumente encontradas na ciência dos materiais neste tipo de abordagem.

No entanto, uma vez que para que a ciência através dos dados seja realizada ainda é necessário um custo bastante elevado para a geração do conjunto de amostras, esse processo ainda está longe de ser ideal para todo tipo de estudo de fenômenos para qualquer material. O presente estudo só foi possível de ser realizado por apresentar um conjunto de dados com quantidade e homogeneidade o suficiente para que o seu comportamento possa ser predito.

Com o aumento no número de amostras e mantendo a homogeneidade das classes, tanto na abordagem por regressão quanto na classificação, pode-se obter resultados ainda mais precisos com o consequente aumento na acurácia dos modelos, o que pode servir como base para estudos posteriores envolvendo predições em compostos de estrutura perovskita.

Vale destacar a importância da já mencionada homogeneidade do conjunto de dados, uma vez que, para exemplificar, se tivermos presente no conjunto de dados

muitas amostras com elevada E_{hull} e poucas amostras com E_{hull} próximas de zero, pode-se desta forma induzir os modelos ao erro, uma vez que haverá muito mais informações de compostos instáveis e poucos para compostos estáveis, gerando assim enviesamento dos dados ou mesmo a geração de valores estatísticos que não refletiriam a eficácia do modelo.

Neste estudo de caso, contudo, a influência da não homogeneidade das classes foi contornada com o devido pré-tratamento dos dados e com a escolha do modelo, além da validação do seu comportamento através da validação cruzada, das métricas estatísticas e das curvas ROC.

As etapas realizadas no pré-processamento também garantiram a redução da ocorrência de *overfitting* com a eliminação dos atributos que apresentaram alta correlação entre si, remoção de *outliers* e uso de modelos combinados. No estudo de conjunto de dados envolvendo a ciência de materiais, essa etapa se mostrou crucial.

Contudo, é importante destacar que embora um modelo regressor com R^2 próximo de 0.89, como obtido neste experimento, consiga obter resultados bastante condizentes com o observado por experimentos reais e simulações computacionais, estes são mais indicados para sistemas de apoio a tomada de decisão, ou seja, além da predição encontrada pelo modelo escolhido para um determinado composto de perovskita, outros critérios devem ser analisados para definir a utilização prática de um material, como as técnicas tradicionais de seleção de materiais, avaliação e caracterização dos mesmos, estudos de experimentos já realizados e consolidados cientificamente, entre outros.

O modelo utilizando ciência orientado a dados avalia compostos promissores a serem estudados e descarta outros que apresentam valores preditos não satisfatórios. Assim, as predições validadas, simuladas e obtidas experimentalmente podem ser utilizadas para alimentar o modelo preditivo, podendo assim elevar a sua eficiência.

Por fim, com as validações e resultados apresentados, o modelo *Extra Trees* poderá ser devidamente aplicado para fins de predição da estabilidade de novos compostos de perovskita, o que poderá ser aplicado em estudos para diminuir uma

das principais desvantagens do uso da perovskita, que são o uso de metais que, se descartados indevidamente podem agredir o meio ambiente.

Como sugestões de novos estudos, novas amostras podem ser obtidas através do DFT ou de novas técnicas de obtenção das propriedades dos compostos. Além disso, novos modelos de aprendizado de máquina podem ser verificados, bem como variações no pré-processamento e na configuração dos hiperparâmetros dos modelos.

REFERÊNCIAS

ADDAGATIA, A. Investigating Underfitting and Overfitting. **Medium**. Disponível em: <https://medium.com/geekculture/investigating-underfitting-and-overfitting-70382835e45c>. Acesso em: 01 jul. 2022.

AHMAD, M. W.; REYNOLDS, J.; REZGUI, Y. Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. **Journal of Cleaner Production**, v. 203, p. 810–821, dez. 2018.

AL-AZZAM, N.; SHATNAWI, I. Comparing supervised and semi-supervised Machine Learning Models on Diagnosing Breast Cancer. **Annals of Medicine and Surgery**, v. 62, p. 53–64, fev. 2021.

ALI KHAN, M. et al. Application of random forest for modelling of surface water salinity. **Ain Shams Engineering Journal**, p. S2090447921004007, nov. 2021.

ARIA, M.; CUCCURULLO, C.; GNASSO, A. A comparison among interpretative proposals for Random Forests. **Machine Learning with Applications**, v. 6, p. 100094, dez. 2021.

BELGIU, M.; DRĂGUŢ, L. Random forest in remote sensing: A review of applications and future directions. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 114, p. 24–31, abr. 2016.

BODE, S. et al. Characterization of Self-Healing Polymers: From Macroscopic Healing Tests to the Molecular Mechanism. Em: HAGER, M. D.; VAN DER ZWAAG, S.; SCHUBERT, U. S. (Eds.). **Self-healing Materials**. Advances in Polymer Science. Cham: Springer International Publishing, 2015. v. 273, p. 113–142.

BRUCE, P.; BRUCE, A. Estatística Prática para Cientistas de Dados: 50 Conceitos Essenciais. **O' Reilly Media Inc, Alta Books Editora**, 1. ed, 2019.

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, p.5-32, 2001

BROWNLEE, J. Basic Concepts in Machine Learning. **Machine Learning Mastery**, 2020. Disponível em: <https://machinelearningmastery.com/basic-concepts-in-machine-learning/> (Acesso em 07/03/2022)

CHAKRABARTY, N.; BISWAS, S. Navo Minority Over-sampling Technique (NMOTe): A Consistent Performance Booster on Imbalanced Datasets. **Journal of Electronics and Informatics**, v. 2, n. 2, p. 96–136, 4 jun. 2020.

CHAN, C. H.; SUN, M.; HUANG, B. Application of machine learning for advanced material prediction and design. **EcoMat**, 7 mar. 2022.

CHUGH, A. MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared – Which Metric is Better? **Medium**, 2020. Disponível em: <https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e> Acesso em: 01 jul. 2022.

GALAN, E. A. et al. Intelligent Microfluidics: The Convergence of Machine Learning and Microfluidics in Materials Science and Biomedicine. **Matter**, v. 3, n. 6, p. 1893–1922, dez. 2020.

GAVRILOV, A. D. et al. Preventing Model Overfitting and Underfitting in Convolutional Neural Networks: **International Journal of Software Science and Computational Intelligence**, v. 10, n. 4, p. 19–28, out. 2018.

GIUSTINO, F. et al. The 2021 quantum materials roadmap. **Journal of Physics: Materials**, v. 3, n. 4, p. 042006, 1 out. 2020.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The Elements of Statistical Learning. Data Mining, Inference and Prediction. **Springer**, v. 2, 2009.

HIMANEN, L. et al. Data-driven materials science: status, challenges and perspectives. **Advanced Science**, v. 6, n. 21, p. 1900808, nov. 2019.

JACOBS, R. et al. Material Discovery and Design Principles for Stable, High Activity Perovskite Cathodes for Solid Oxide Fuel Cells. **Advanced Energy Materials**, v. 8, n. 11, p. 1702708, abr. 2018.

JAMES, G. et al. An Introduction to Statistical Learning. **Springer**. Nova Iorque, 2 ed. 2021

KIM, H. et al. Halide Perovskites for Applications beyond Photovoltaics. **Small Methods**, v. 2, n. 3, p. 1700310, mar. 2018.

LI, W.; JACOBS, R.; MORGAN, D. Predicting the thermodynamic stability of perovskite oxides using machine learning models. **Computational Materials Science**, v. 150, p. 454–463, jul. 2018.

LIANG, Z. et al. A Machine Learning Method for Material Property Prediction: Example Polymer Compatibility. p. 11, [s.d.].

LIU, Y. et al. Machine learning for advanced energy materials. **Energy and AI**, v. 3, p. 100049, mar. 2021.

LOOKMAN, T. et al. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. **npj Computational Materials**, v. 5, n. 1, p. 21, dez. 2019.

LU, Z. Computational discovery of energy materials in the era of big data and machine learning: A critical review. **Materials Reports: Energy**, v. 1, n. 3, p. 100047, ago. 2021.

MAKARIOU, D.; BARRIEU, P.; CHEN, Y. A random forest based approach for predicting spreads in the primary catastrophe bond market. **Insurance: Mathematics and Economics**, v. 101, p. 140–162, nov. 2021.

MOHANA, R. M. et al. Random forest algorithms for the classification of tree-based ensemble. **Materials Today: Proceedings**, p. S2214785321008853, fev. 2021.

MORGAN, D. **Machine Learning Materials Datasets. Figshare**, 2018. Disponível em: <https://doi.org/10.6084/m9.figshare.7017254.v5>. Acesso em: 9 set. 2021

OKORO, E. E. et al. Application of artificial intelligence in predicting the dynamics of bottom hole pressure for under-balanced drilling: Extra tree compared with feed forward neural network model. **Petroleum**, p. S240565612100016X, mar. 2021.

PANDYA, Y. Ensemble Methods in Machine Learning. **Medium**, 2021. Disponível em <https://medium.com/analytics-vidhya/ensemble-methods-in-machine-learning-31084c3740be>. Acesso em 04 mar. 2022

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

PICKLUM, M.; BEETZ, M. MatCALO: Knowledge-enabled machine learning in materials science. **Computational Materials Science**, v. 163, p. 50–62, jun. 2019.

PYKES, K. Comprehension of the AUC-ROC curve. **Towards Data Science**, 2020. Disponível em <https://towardsdatascience.com/comprehension-of-the-auc-roc-curve-e876191280f9>. Acesso em 28 jul. 2022.

RODRIGUES, V. Métricas de Avaliação: acurácia, precisão, recall… quais as diferenças?. **Medium,** 2019. Disponível em: <https://vitorborbarodrigues.medium.com/m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-acur%C3%A1cia-precis%C3%A3o-recall-quais-as-diferen%C3%A7as-c8f05e0a513c>. Acesso em: 01 jul. 2022.

RAKHRA, M. et al. Crop Price Prediction Using Random Forest and Decision Tree Regression:-A Review. **Materials Today: Proceedings**, p. S2214785321022902, abr. 2021.

ROY, P. et al. A review on perovskite solar cells: Evolution of architecture, fabrication techniques, commercialization issues and status. **Solar Energy**, v. 198, p. 665–688, mar. 2020.

SCHLEDER, G. R.; FAZZIO, A. Machine Learning na Física, Química, e Ciência de Materiais: Descoberta e Design de Materiais. **Revista Brasileira de Ensino de Física**, v. 43, n. suppl 1, p. e20200407, 2021.

SHAHROKHI, S. et al. Emergence of Ferroelectricity in Halide Perovskites. **Small Methods**, v. 4, n. 8, p. 2000149, ago. 2020.

SHARMA, D. K. et al. Classification of COVID-19 by using supervised optimized machine learning technique. **Materials Today: Proceedings**, p. S2214785321074101, nov. 2021.

SHOLL, D.; STECKEL, J. Density Functional Theory: A Practical Introduction. **John Wiley & Sons**, ed. 1, 2011.

VERZINO, G. Why Balancing Classes is Over-Hyped. **Towards Data Science**, 2021. Disponível em; <https://towardsdatascience.com/why-balancing-classes-is-over-hyped-e382a8a410f7>. Acesso em 29 jul. 2022

VOLPI, G. F. 6 amateur mistakes I’ve made working with train-test splits. **Towards Data Science**, 2019. Disponível em: <https://towardsdatascience.com/6-amateur-mistakes-ive-made-working-with-train-test-splits-916fabb421bb>. Acesso em 04 mar. 2022.

WARD, L. A general-purpose machine learning framework for predicting. **npj Computational Materials**, p. 7, 2016.

WEI, J. et al. Machine learning in materials science. **InfoMat**, v. 1, n. 3, p. 338–358, set. 2019.

WEINBERGER, C. R. et al. Ab initio investigations of the phase stability in group IVB and VB transition metal nitrides. **Computational Materials Science**, v. 138, p. 333–345, out. 2017.

WU, Y. et al. First principles high throughput screening of oxynitrides for water-splitting photocatalysts. **Energy Environ. Sci.**, v. 6, n. 1, p. 157–168, 2013.

ZHANG, Y. X. et al. A two-step fused machine learning approach for the prediction of glass-forming ability of metallic glasses. **Journal of Alloys and Compounds**, v. 875, p. 160040, 2021.

ANEXOS

O dataset utilizado neste trabalho pode ser encontrado no site FigShare, disponibilizado através do link: https://figshare.com/articles/dataset/MAST-ML_Education_Datasets/7017254?file=12978425 (Último acesso em 24/06/2022)

O repositório contendo o código utilizado nas análises está disponibilizado no GitHub do autor, no seguinte link: https://github.com/Erick-Faster/ml-perovskite-tg (Última atualização em 05/08/2022)