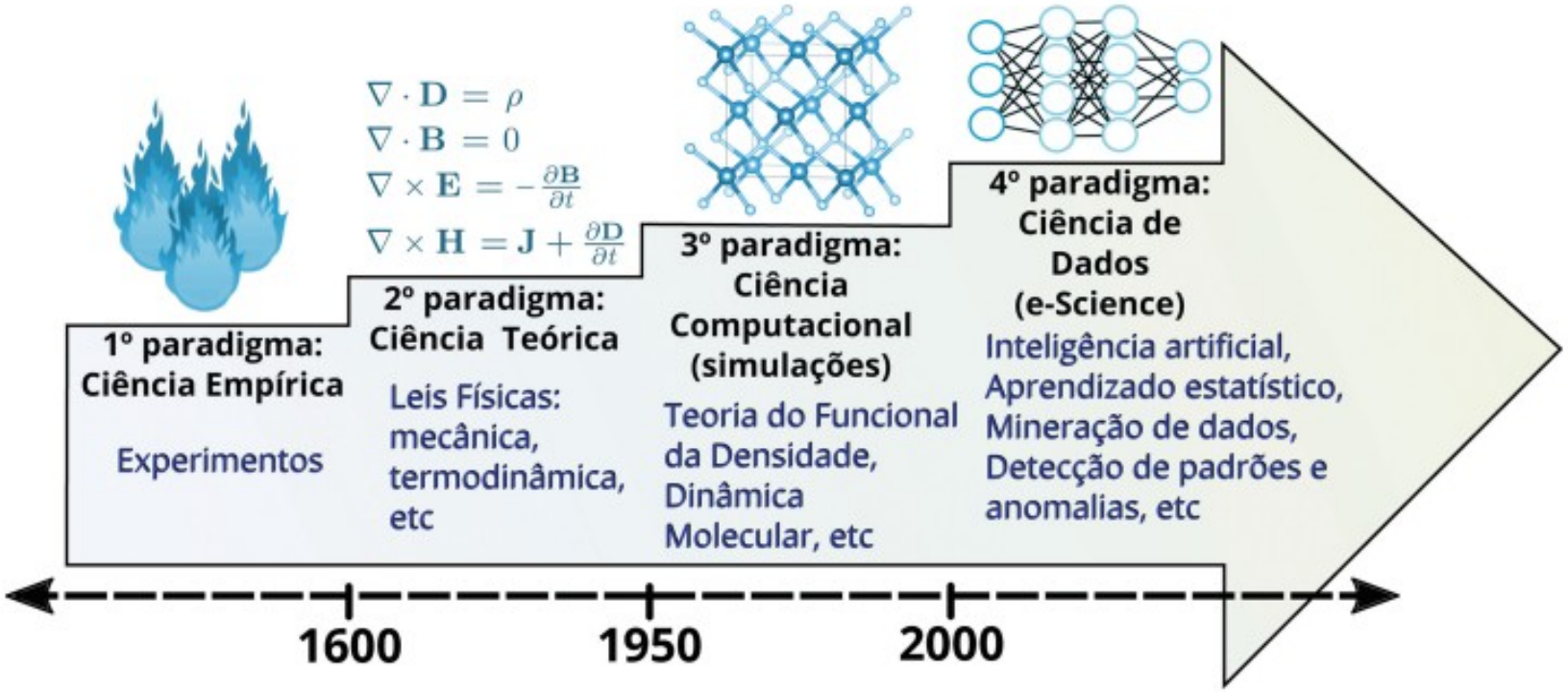
Aprendizado de máquina utilizando árvores de decisão e variantes para a predição da estabilidade de compostos de perovskita

Autor: Erick Fasterra da Silva

Introdução

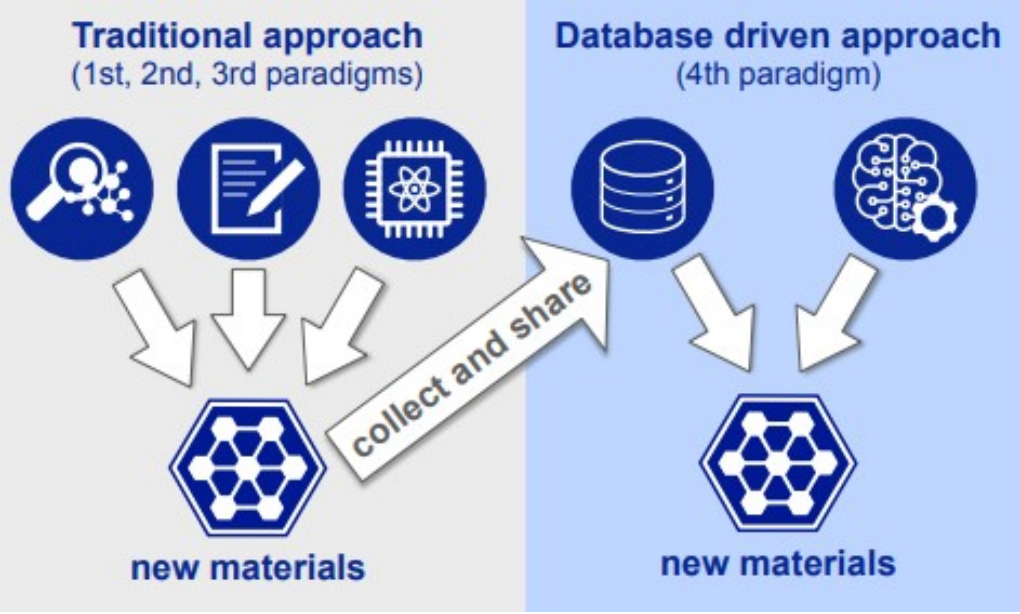
- **Estudo e Avaliação de Novos Materiais**
  - Uso de diferentes abordagens
  - Paradigmas da Ciência aplicados aos Materiais



(BODE et al.,2015)

Introdução

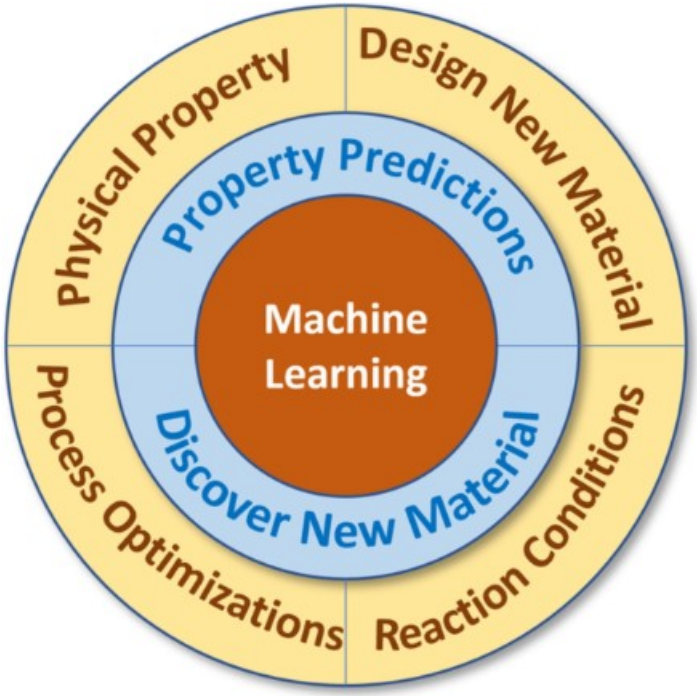
- **4º Paradigma: Orientação à dados**
  - Aprendizado de Máquina e Inteligência Artificial
  - Predição de propriedades baseados em experimentos passados
    - Generalização do comportamento
    - Predição das propriedades
  - Elevada velocidade e precisão



(GIUSTINO et al., 2020)

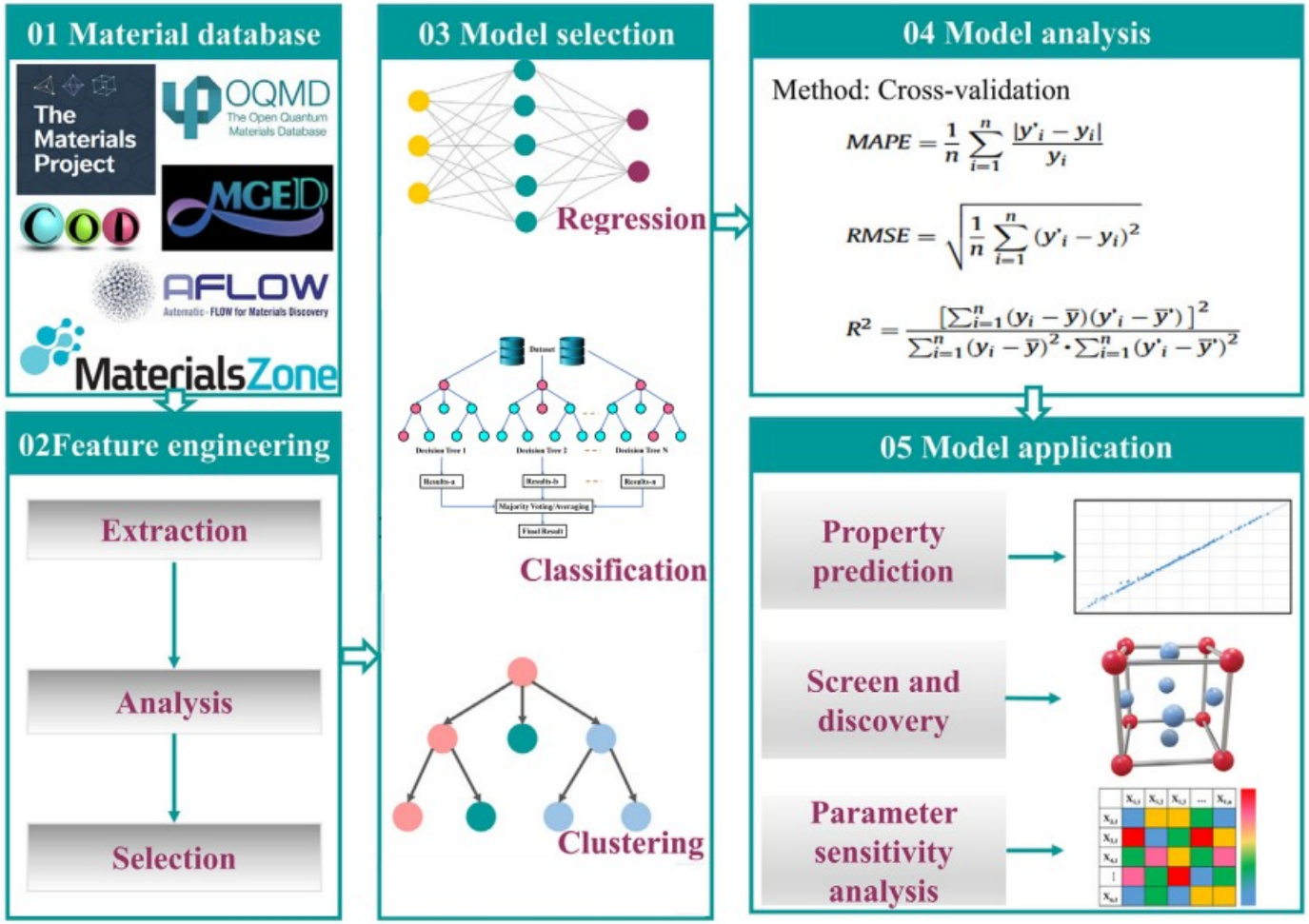
Objetivos

- **Objetivos:**
  - Aplicação de aprendizado de máquina em Ciência de Materiais
    - **Estudo de Caso:** Criação de modelo para a determinação da estabilidade termodinâmica de compostos de perovskita
      - Modelos **Regressores**
      - Modelos **Classificadores**
  - Compreensão de dificuldades na Ciência e Engenharia de Materiais
    - Custo elevado para obtenção de amostras
    - Poucas amostras disponíveis
    - Distribuições não homogêneas



(HINAMEN et al., 2019)

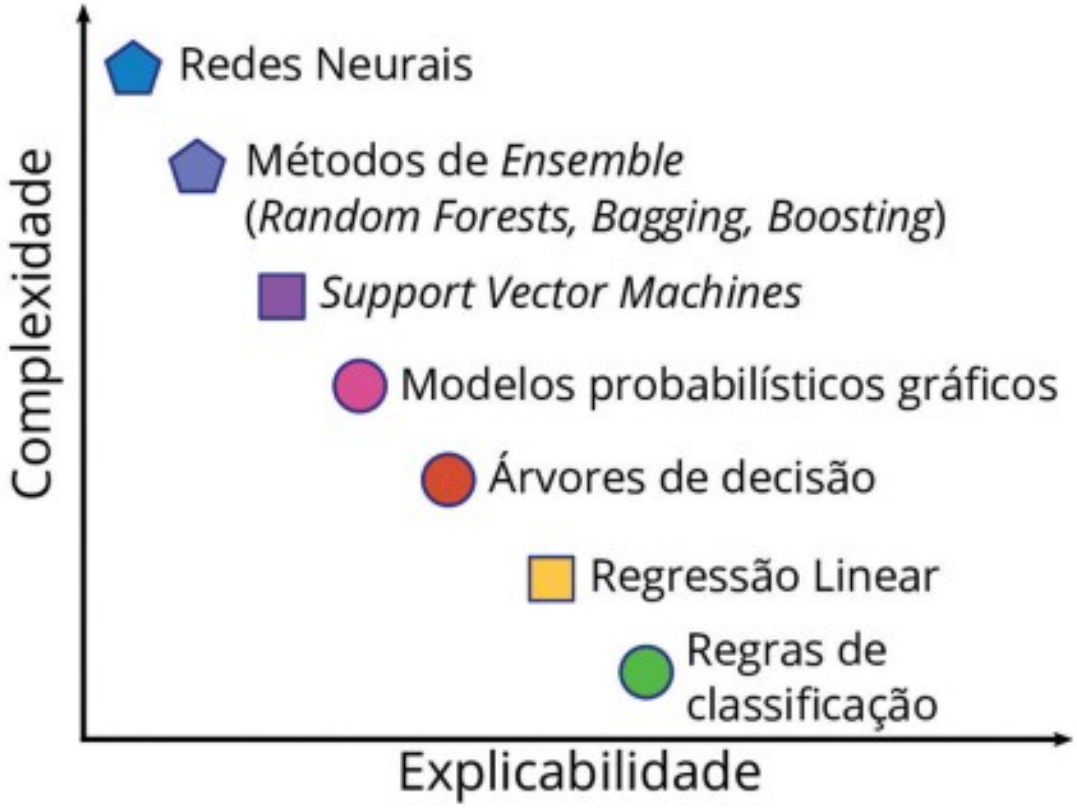
Revisão Bibliográfica – Procedimento de Aplicação



(LIU et al., 2021, adaptado)

Revisão Bibliográfica – Seleção do Modelo

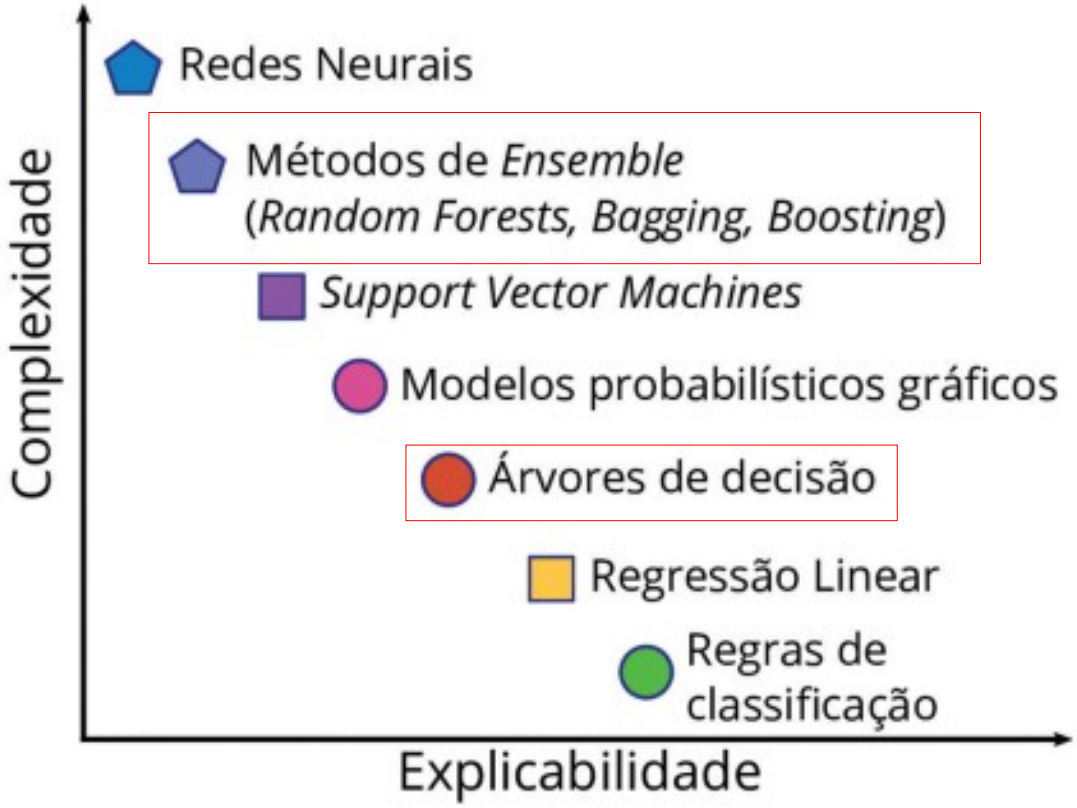
- **Escolha do Modelo**



(SCHLEDER e FAZZIO,2020)

Revisão Bibliográfica – Seleção do Modelo

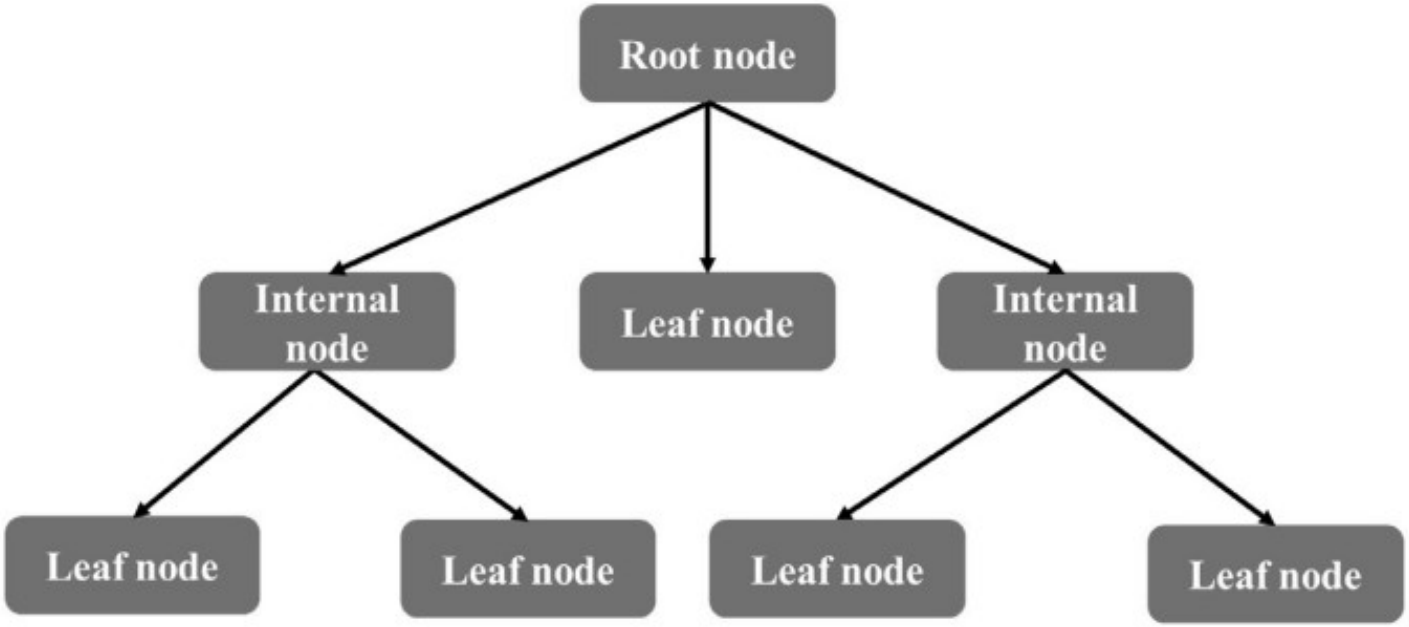
- **Escolha do Modelo**



(SCHLEDER e FAZZIO,2020)

Revisão Bibliográfica – Modelos de Machine Learning

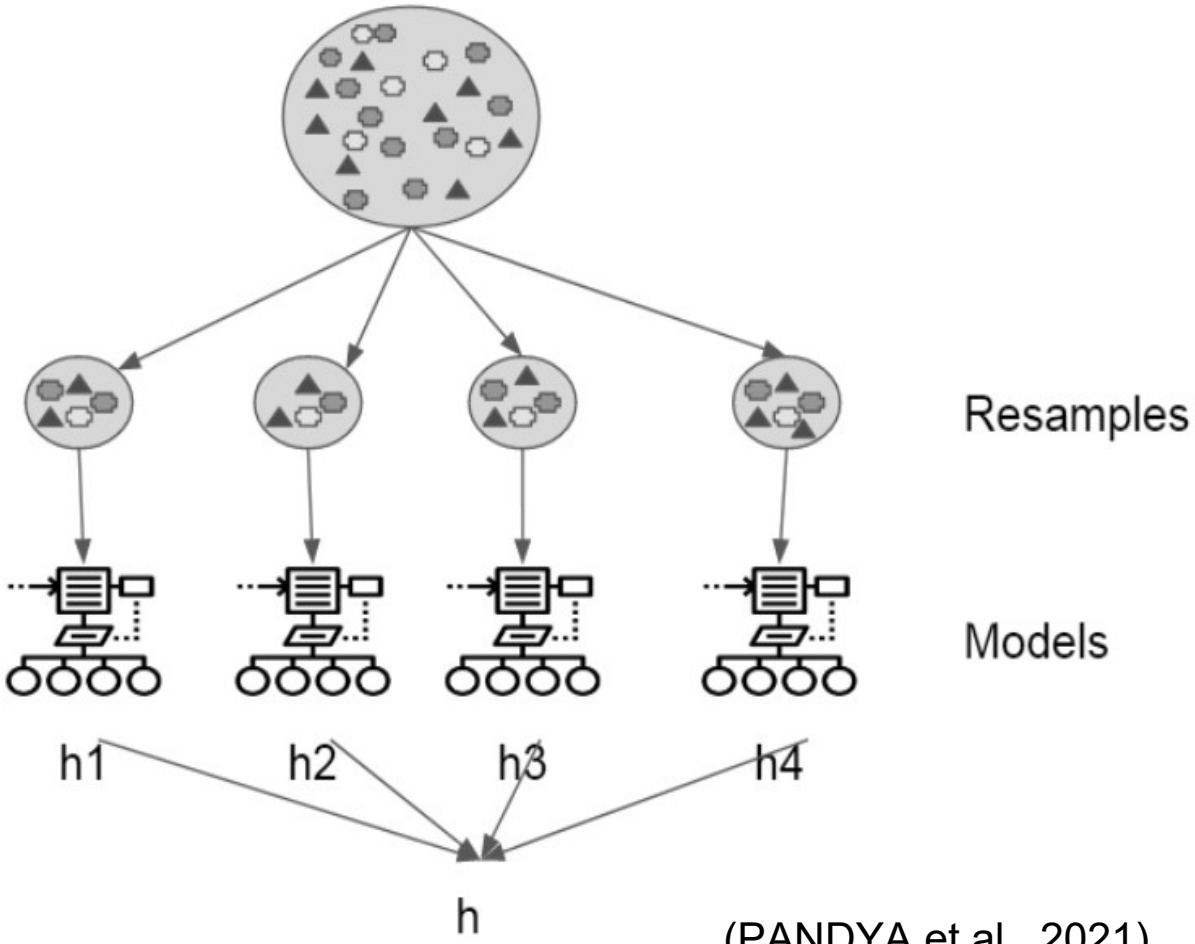
- **Árvores de Decisão (*Decision Trees*)**



(LIU et al., 2021)

Revisão Bibliográfica – Modelos de Machine Learning

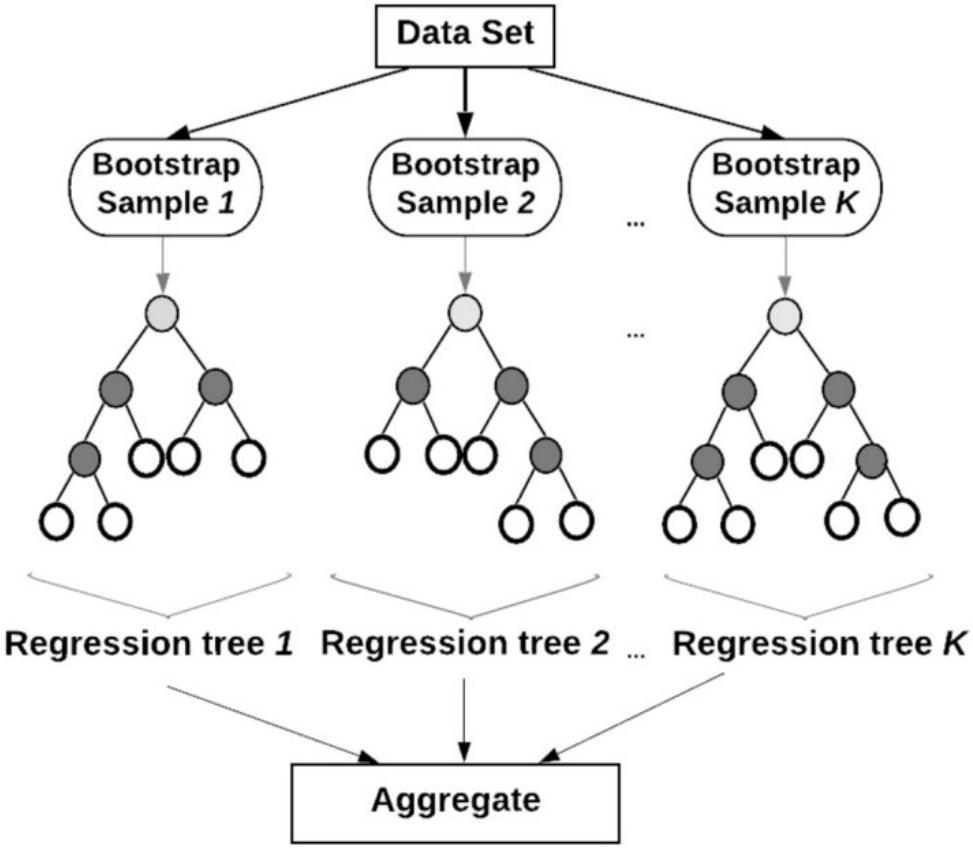
- **Modelos Combinados (*Ensemble*)**



(PANDYA et al., 2021)

Revisão Bibliográfica – Modelos de Machine Learning

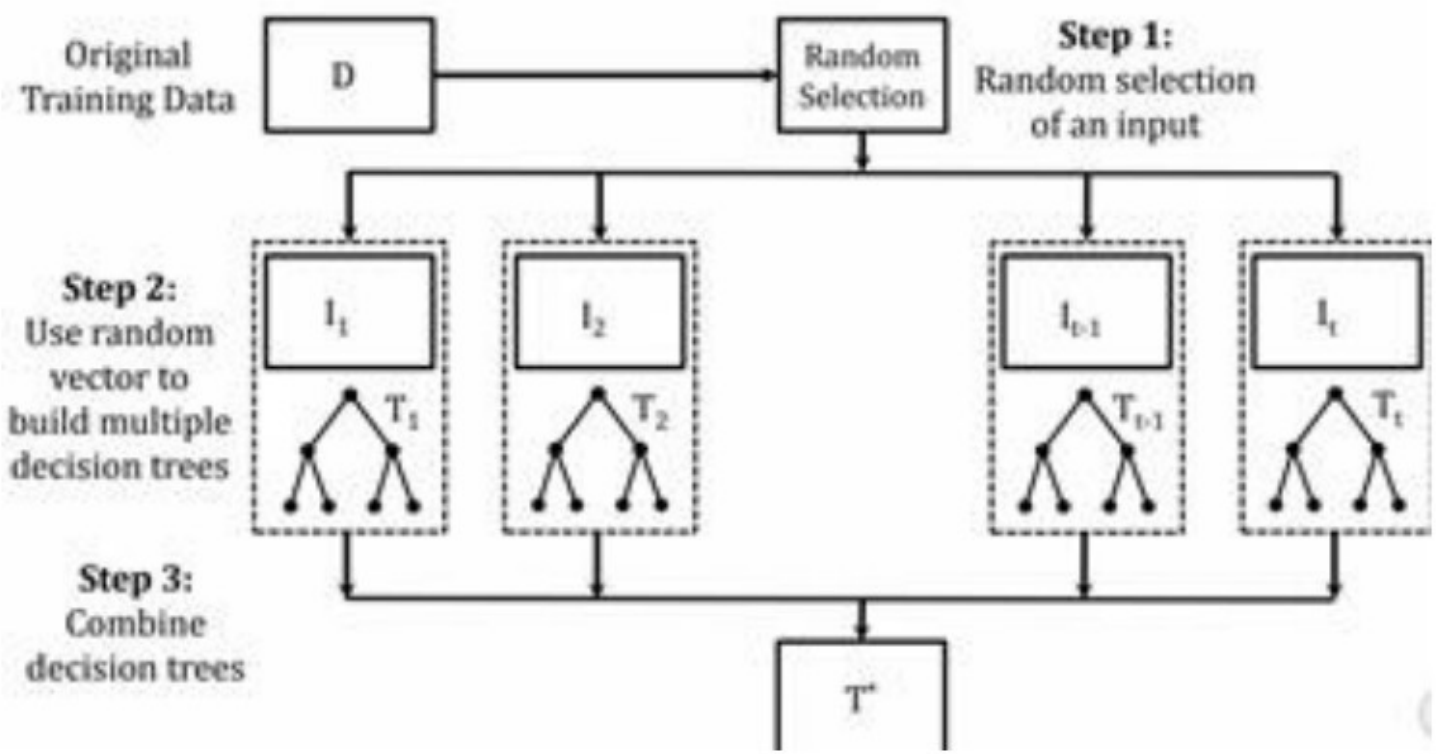
- ***Random Forest***



(MAKARIOU et al., 2021)

Revisão Bibliográfica – Modelos de Machine Learning

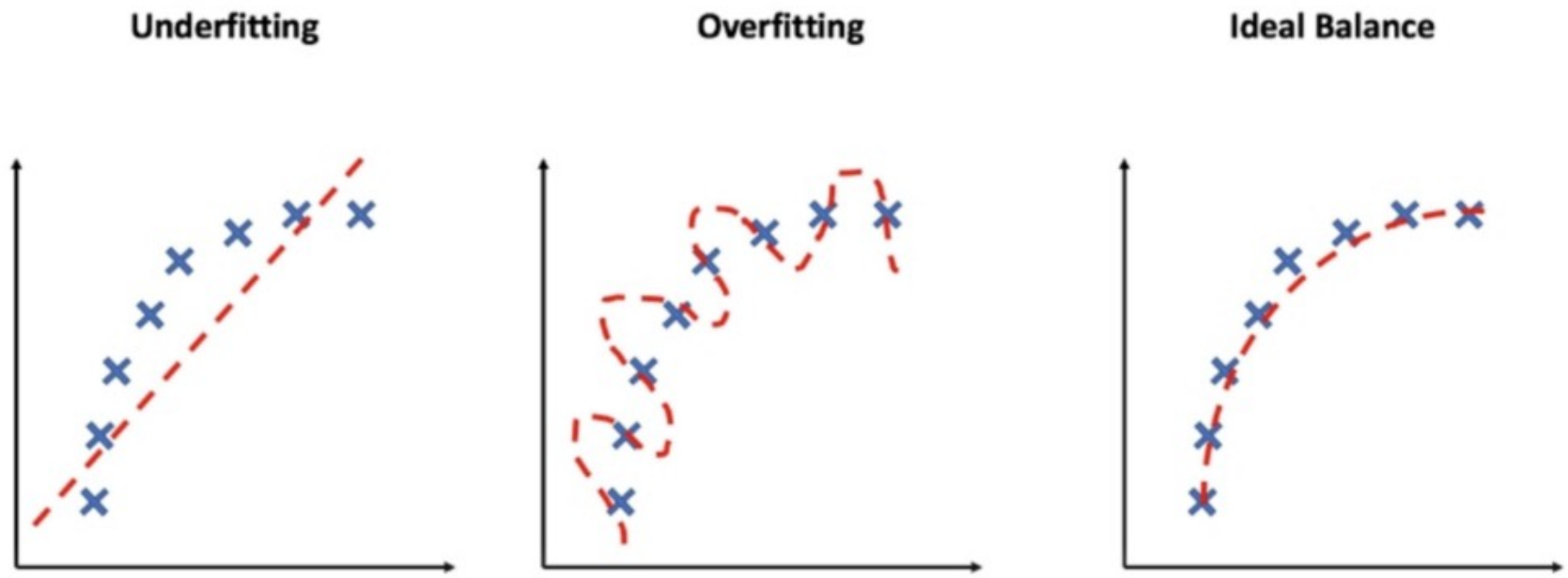
- ***Extra Trees***



(CHAKRABARTY e BISWAS, 2020)

Revisão Bibliográfica – Modelos de Machine Learning

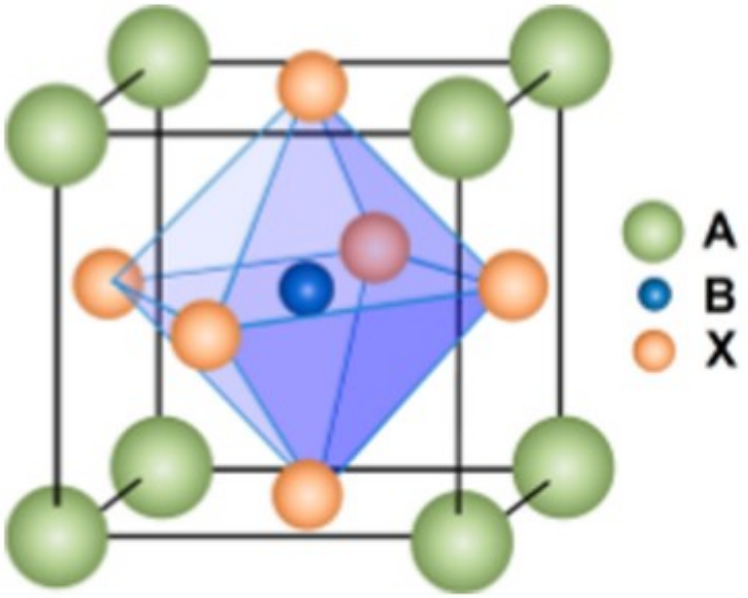
- ***Overfitting* e *Underfitting***



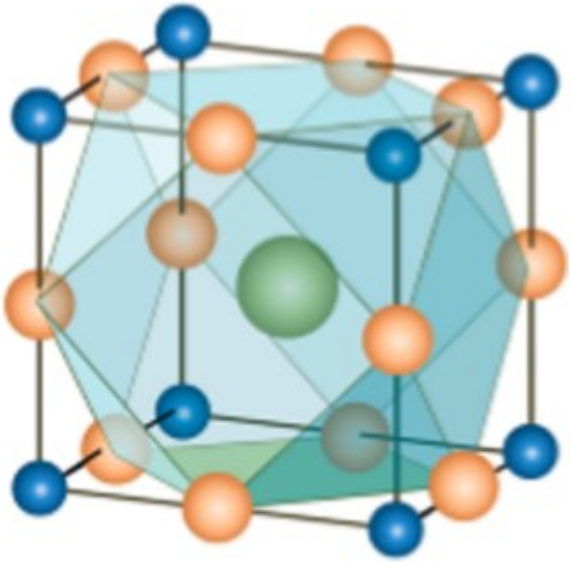
(CHAKRABARTY e BISWAS, 2020)

Metodologia – Estudo de Caso

- **Estudo de Caso:** Determinação da Estabilidade Termodinâmica da Perovskita
  - Fórmula generalizada: ABX_3
  - Estrutura semelhante ao Titanato de Cálcio($CaTiO_3$)



Octaédrico com 6 ânions (BX_6)

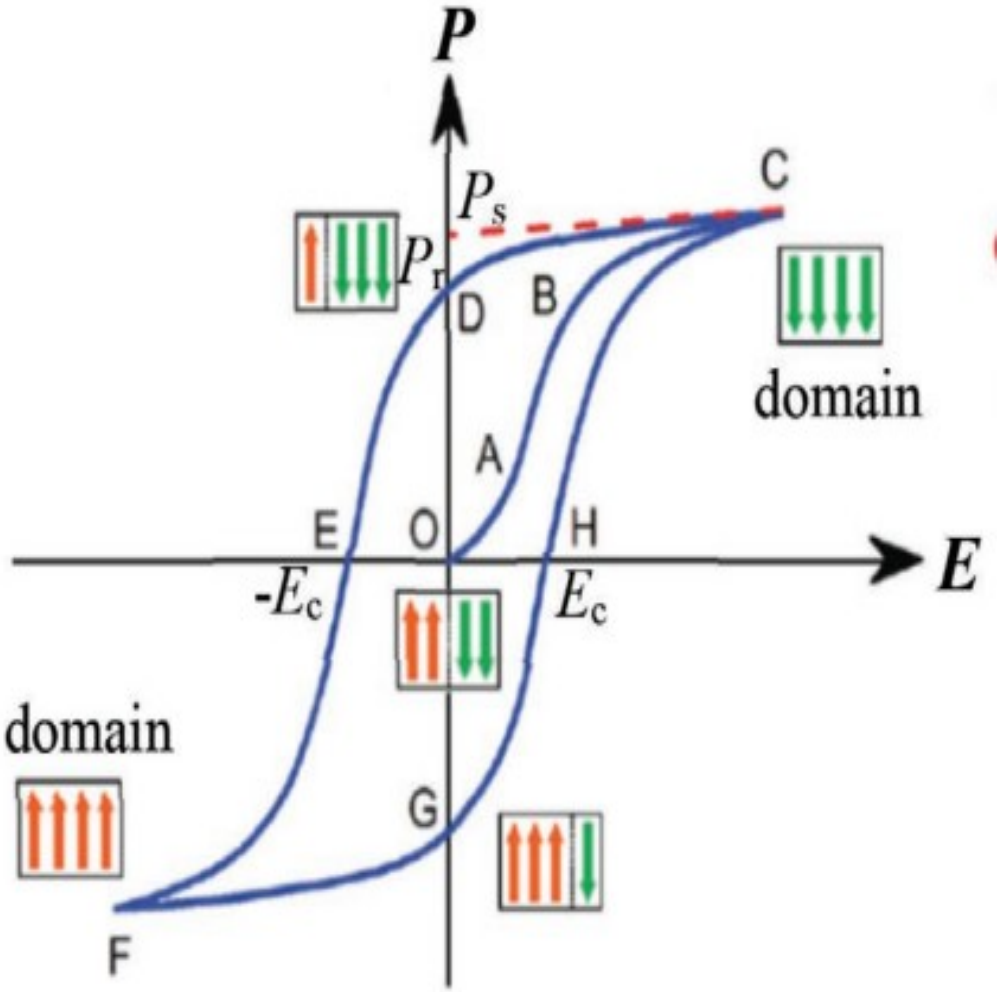


Cuboctaédrico com 12 ânions (AX_{12})

(ROY et al. 2020)

Metodologia – Estudo de Caso

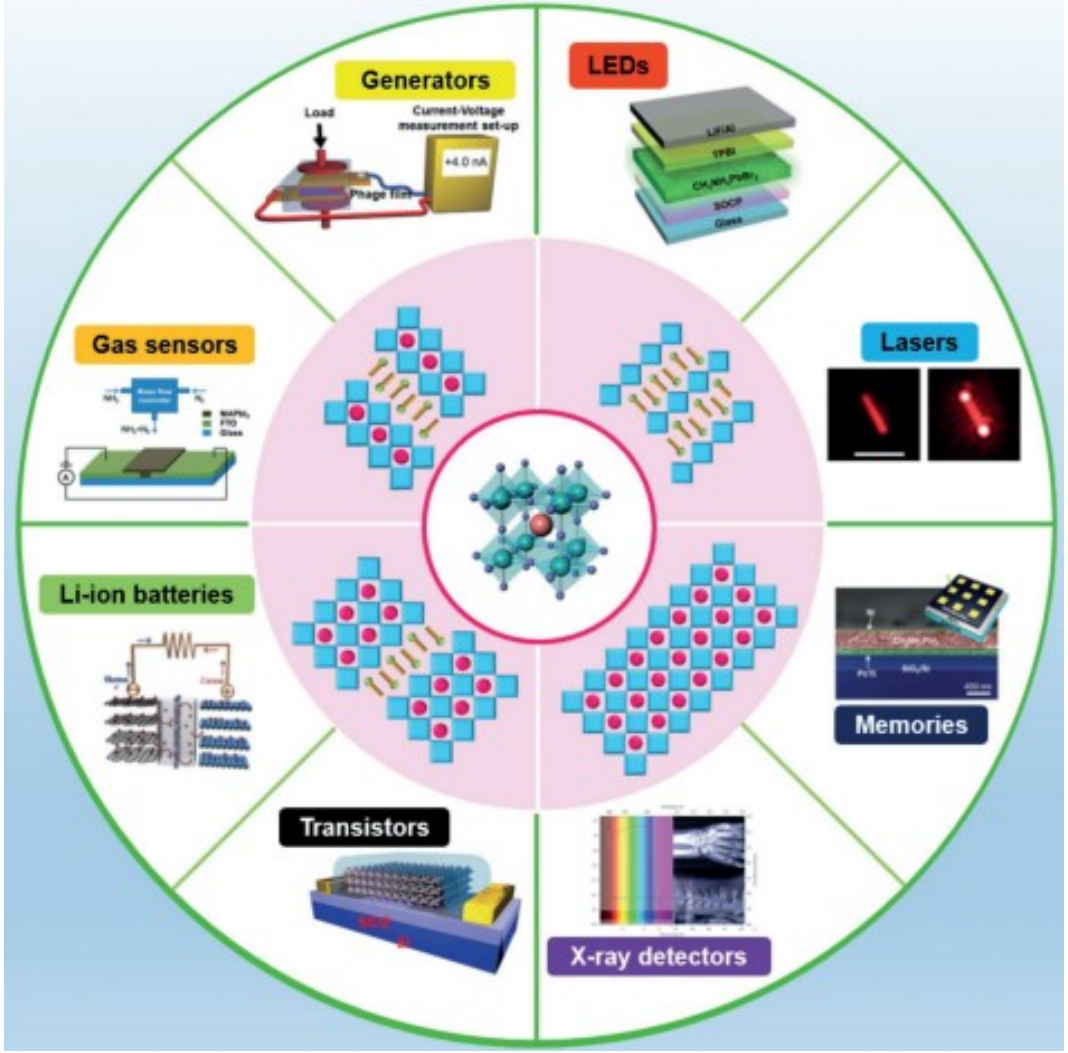
- **Ferroeletricidade**
  - Assimetria em sua estrutura
- **Aplicado estímulo externo:**
  - Orientação dos polos
  - Configuração pode ser mantida



(ROY et al. 2020)

Metodologia – Estudo de Caso

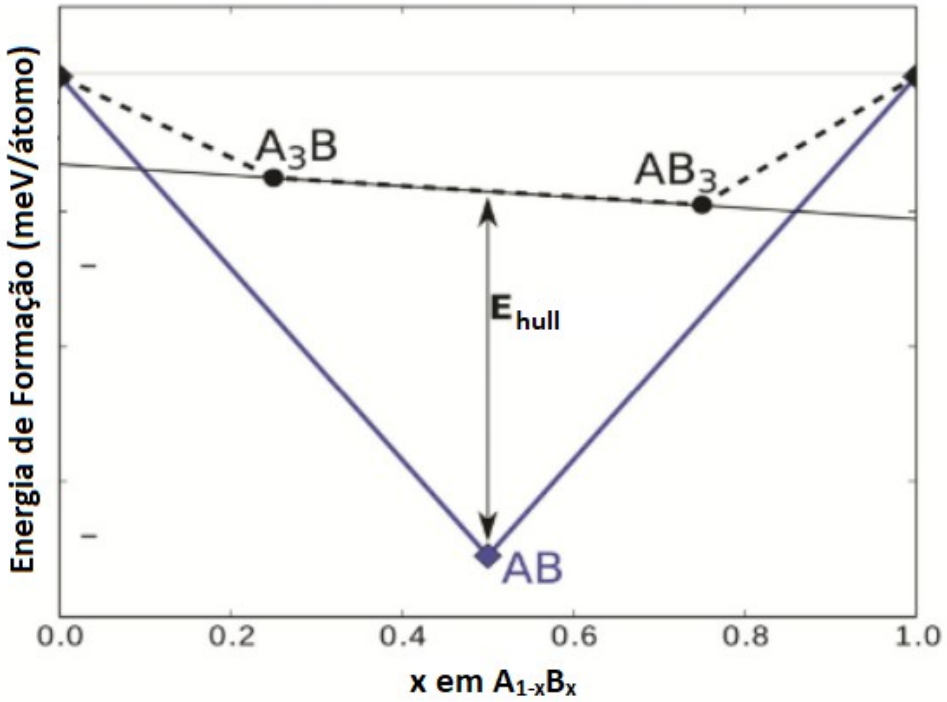
- **Aplicações**



(KIM et al., 2018)

Metodologia – Estudo de Caso

- **Predição da energia acima da envoltória convexa (E_{hull})**
  - Energia de formação em função da composição
  - **Métodos:**
    - Regressão → Análise Quantitativa
    - Classificação → Análise Qualitativa (ponto de corte: **40 meV/átomo)**



(WEINBERGER et al. 2017, adaptado)

Metodologia – Estudo de Caso

- **Li et. al. 2018**
  - Utilizou o mesmo estudo de caso em seu artigo.
    - **Modelos Regressores** → Análise quantitativa
    - **Modelos Classificadores** → Análise qualitativa

- **Extração dos Dados**
  - Base: (JACOBS et al., 2018)
    - Obtidos via simulação computacional DFT (*Density Functional Theory*)
    - 1929 amostras de compostos de perovskita
    - 80 atributos
    - Valores de E_{hull}, entre 0 e 956 meV/átomo.

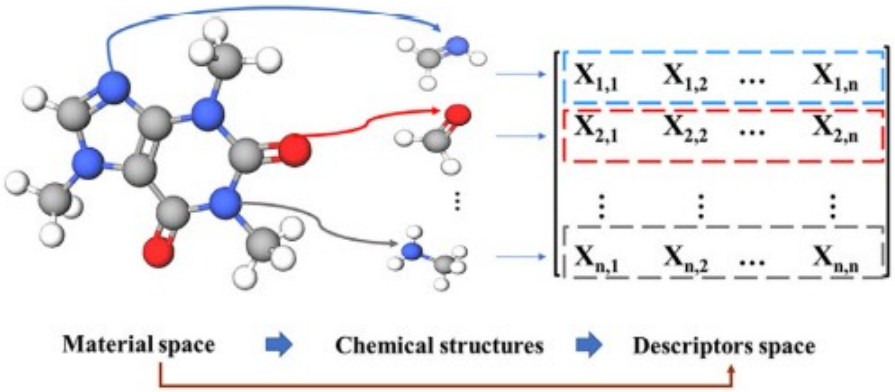
(LIU et al., 2021)

Metodologia – Extração dos Dados

- **Extração dos Dados**
  - Base: (JACOBS et al., 2018)

    - **Atributos (conjunto X de dados)**
      - **Propriedades químicas**
        - Tipos de elementos químicos
        - Número de átomos
        - Raio atômico
      - **Propriedades térmicas**
        - Calor de vaporização
        - Capacidade calorífica específica
        - Condutividade térmica
      - **Propriedades elétricas**
        - Condutividade elétrica

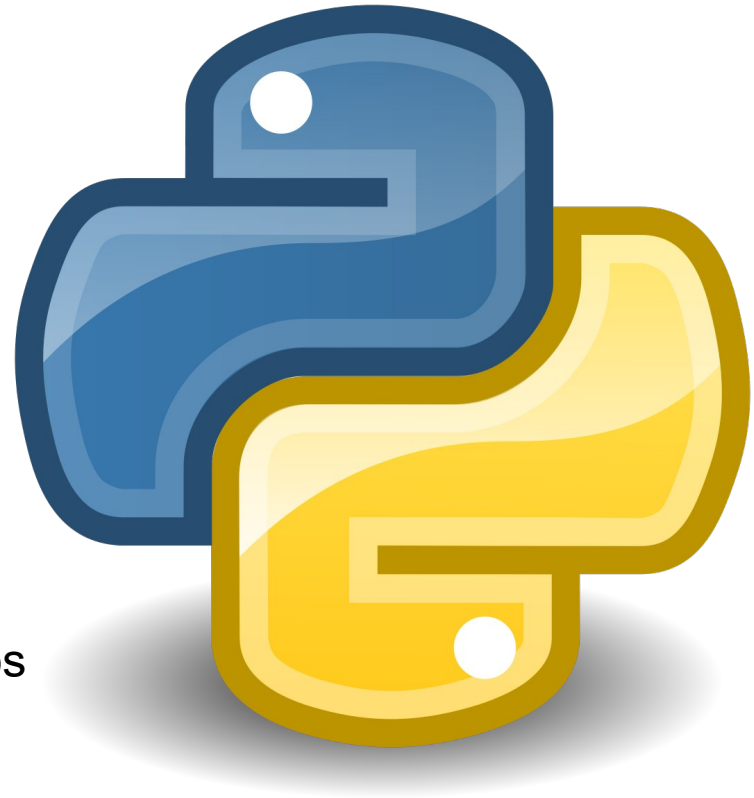
    - **Classe (conjunto y de dados)**
      - Energia acima da envoltória convexa E_{hull}



(LIU et al., 2021)

Metodologia – Ferramentas

- **Ferramentas**
  - Linguagem Utilizada: Python 3.7
  - Bibliotecas
    - Numpy (Algebra Linear)
    - Pandas (Manipulação de dados)
    - Scikit-Learn (Machine Learning)
    - Scikit-Plot (Visualização de ML)
    - Matplotlib (Gráficos)
  - Hardware:
    - Intel Core i5-9300, 2.40GHz
    - 4 núcleos e 8 processadores lógicos

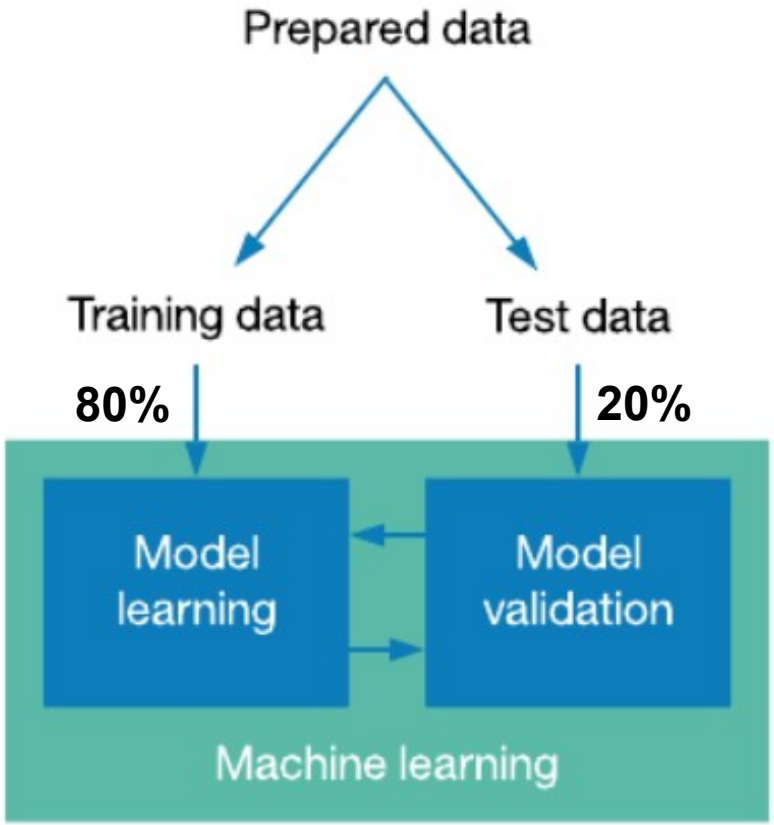


https://upload.wikimedia.org/wikipedia/commons/thumb/0/0a/Python.svg/1200px-Python.svg.png

Metodologia – Preparo dos Dados e Treinamento

- **Tratamento dos Atributos**
  - Eliminação de Outliers (E_{hull} > 400 meV/átomo)
  - Identificação de atributos correlacionados
  - Escalonamento dos atributos

$$X_{escalonado} = \frac{X - X_{mínimo}}{X_{máximo} - X_{mínimo}}$$



- **Tratamento das Classes (E_{hull})**
  - Modelos Regressores:
    - Mantém valores em meV/átomo.
  - Modelos Classificadores:
    - Etiquetação. Corte: 40 meV/átomo

(VOLPI, 2019)

Metodologia – Validação dos modelos Regressores

- **Regressão**

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

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

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

Metodologia – Validação dos modelos Classificadores

- **Classificação**

$$Acurácia = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precisão = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

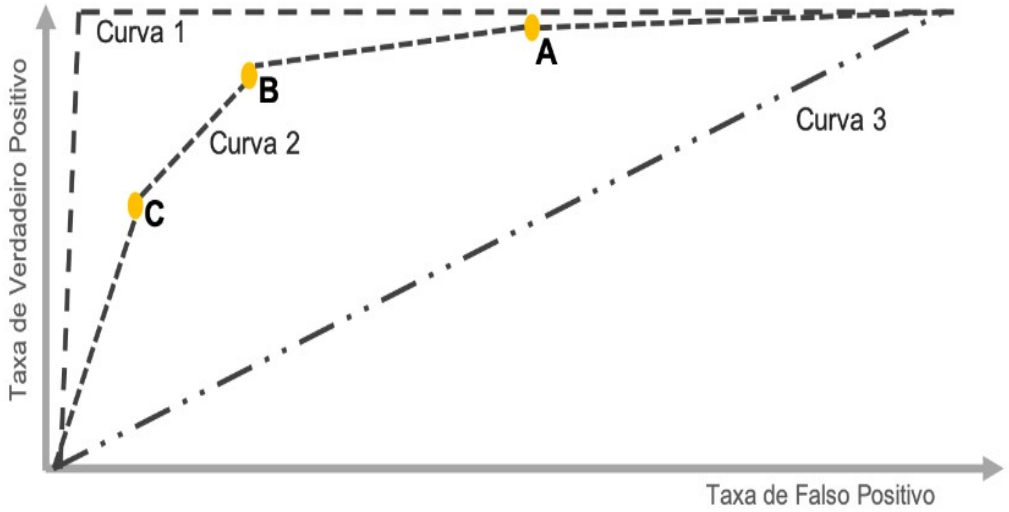
$$F_1 = \frac{2 * Precisão * Recall}{Precisão + Recall}$$

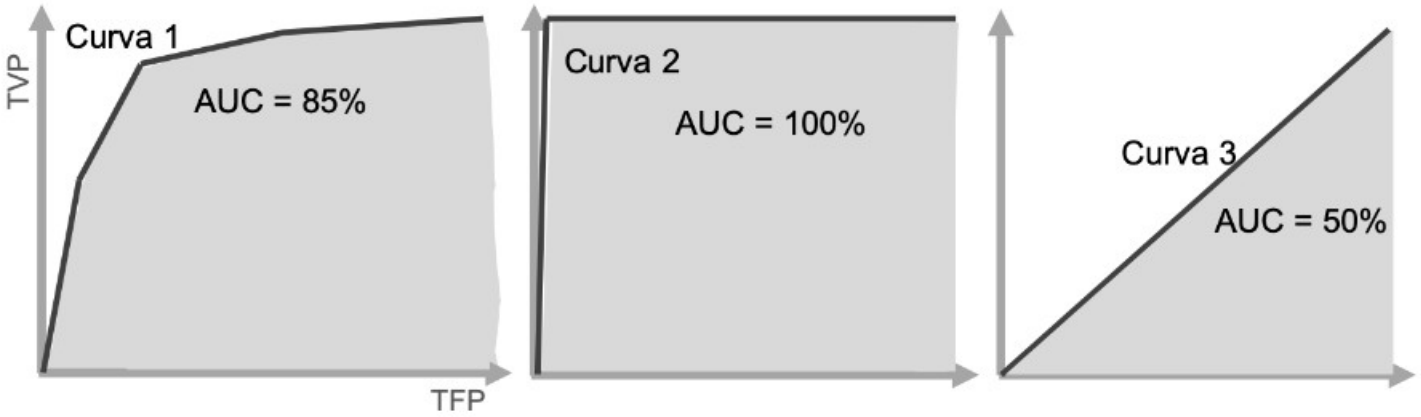
Legenda:

- TP → *True Positive*
- TN → *True Negative*
- FP → *False Positive*
- FN → *False Negative*

Metodologia – Validação dos modelos Classificadores

- **Curvas ROC**

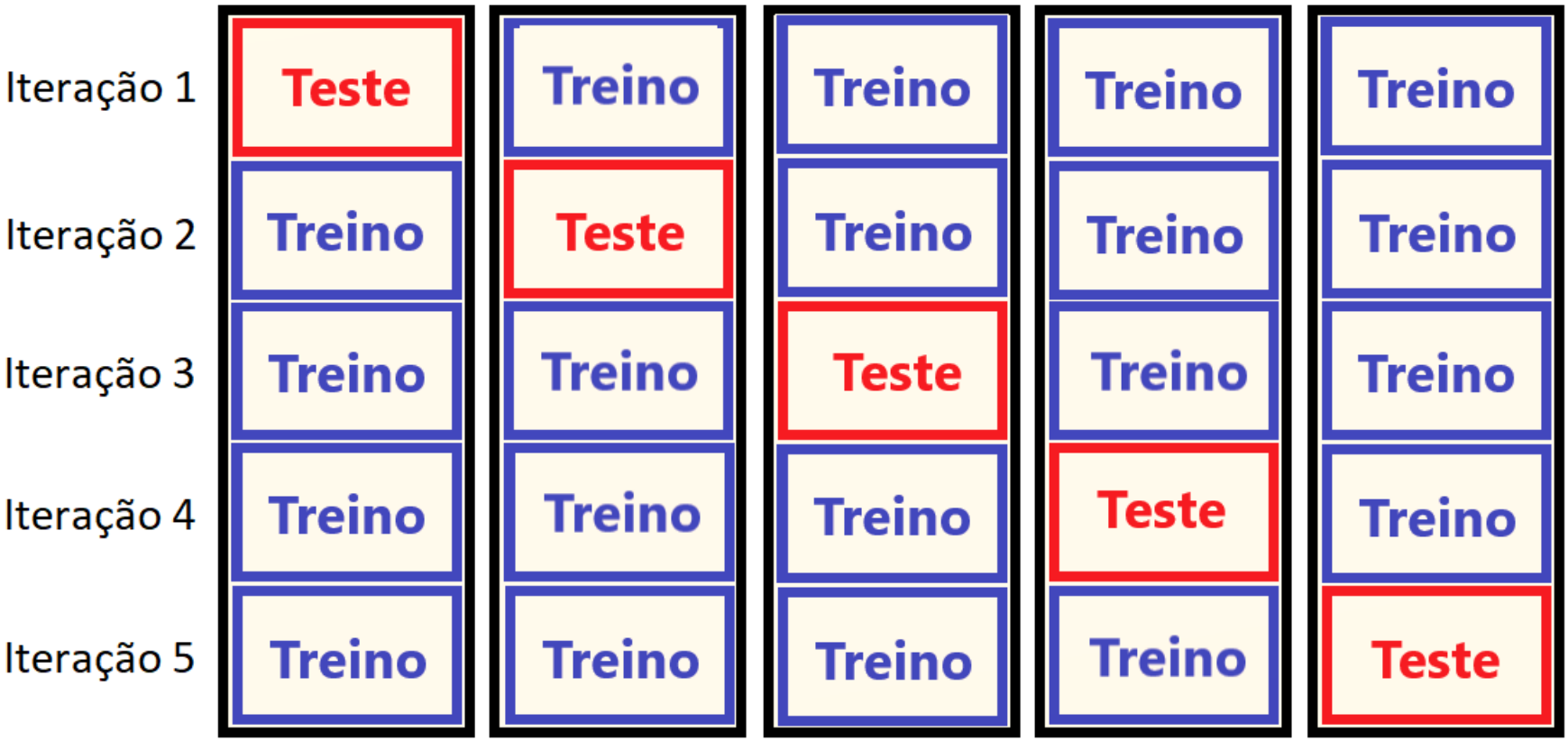




(PYKES, 2020)

Metodologia – Validação dos Modelos

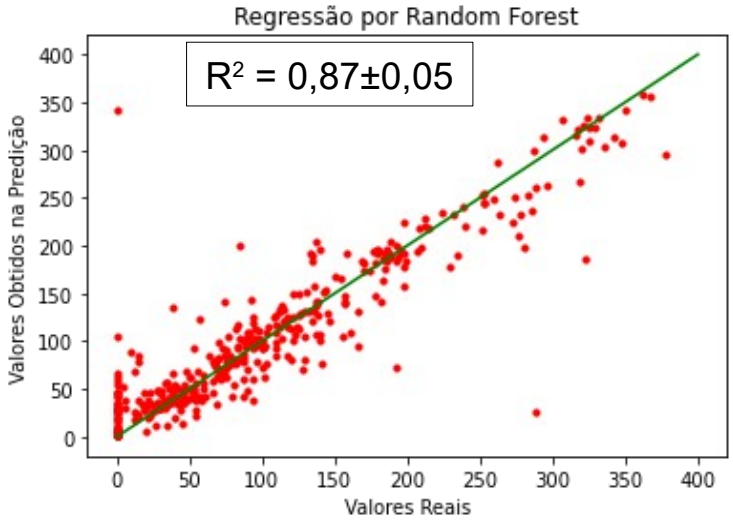
- **Validação Cruzada**

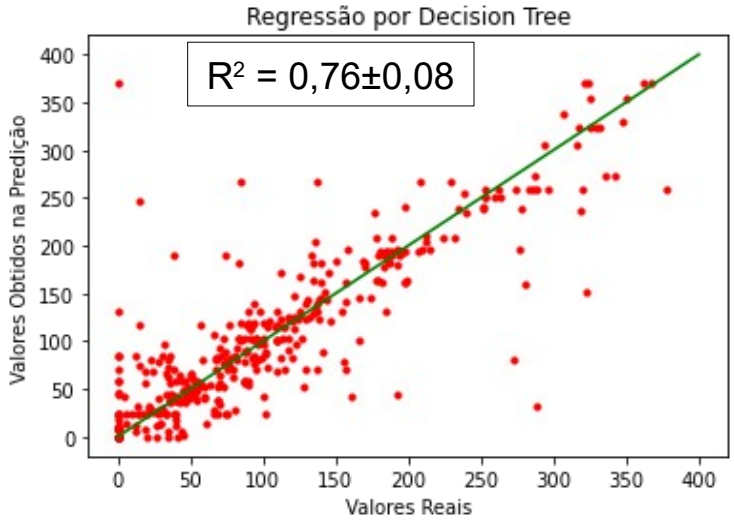


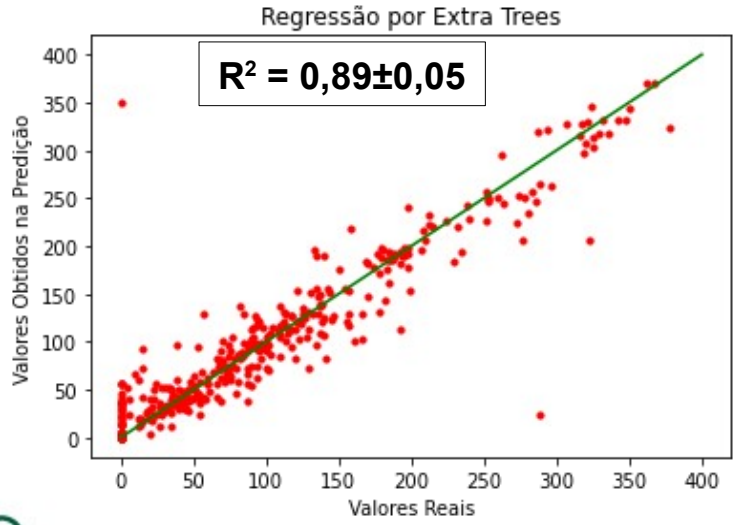
(PEDREGOSA et al. 2011, adaptado)

Resultados – Regressão

- **Modelos Regressores**



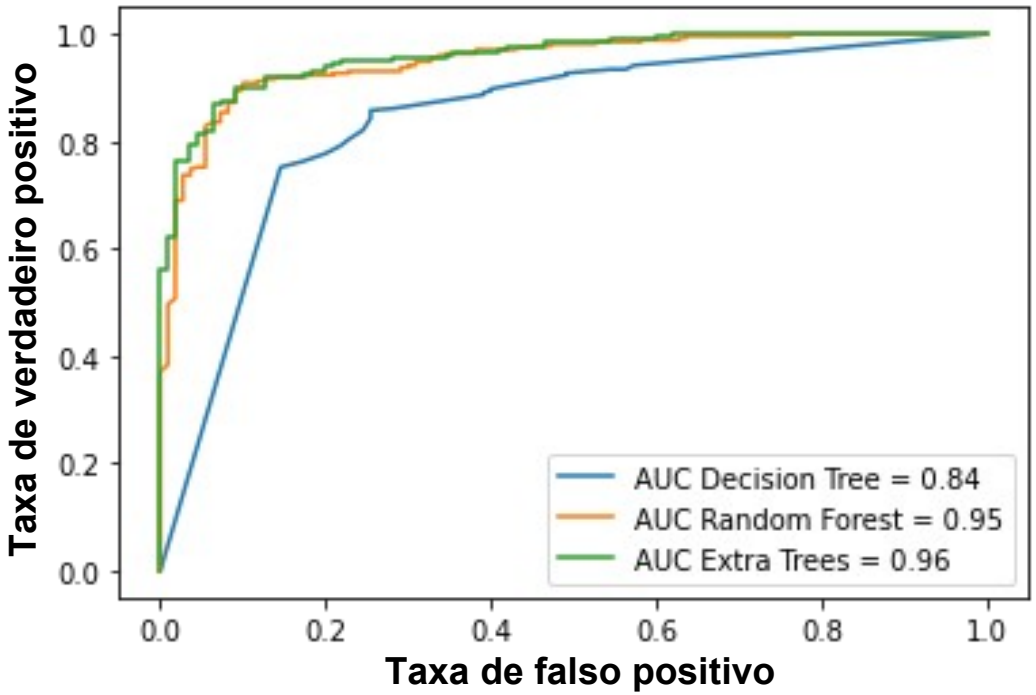




Modelo	MAE (meV/átomo)	RMSE (meV/átomo)	R^2
DT	25,80±3,01	41,37±6,53	0,76±0,08
RF	18,98±2,11	30,53±5,26	0,87±0,05
EX	**16,26±2,20**	**27,64±5,86**	**0,89±0,05**

Resultados – Classificação

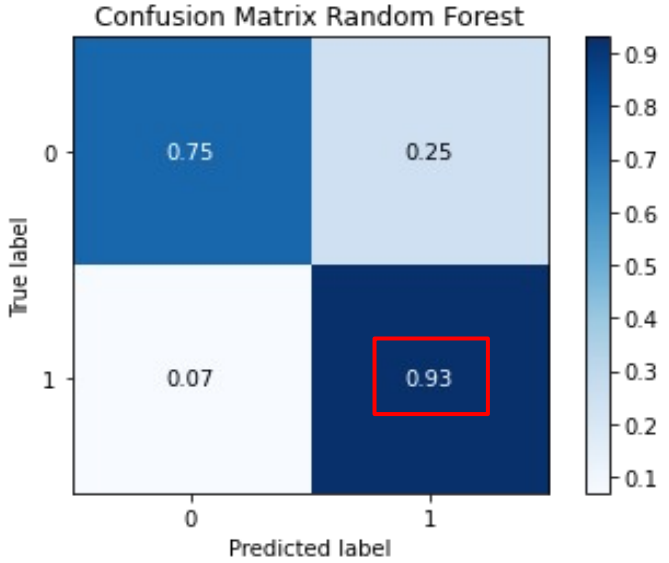
- **Modelos Classificadores**

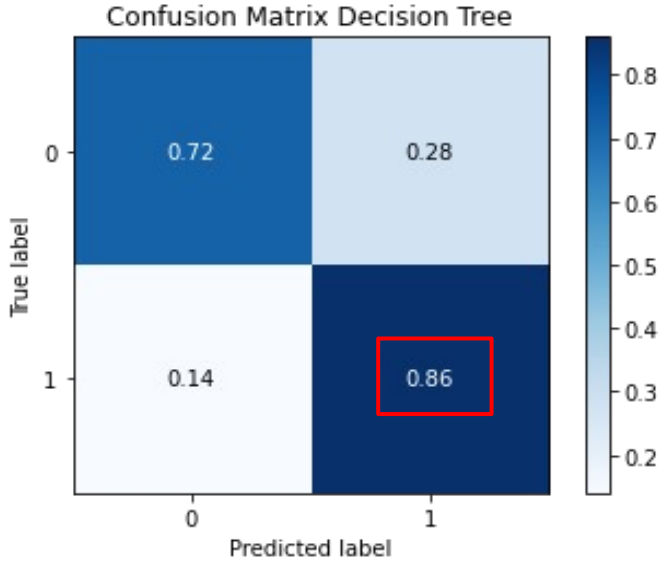


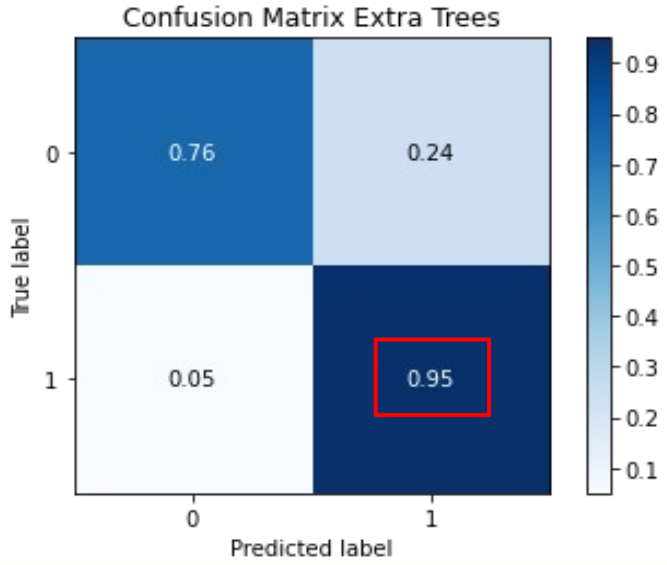
Modelo	Precisão	*Recall*	F_1	Acurácia
DT	0,89±0,04	0,88±0,03	0,88±0,03	0,83±0,04
RF	0,92±0,03	0,94±0,02	0,93±0,02	0,91±0,03
EX	**0,92±0,04**	**0,95±0,02**	**0,94±0,01**	**0,91±0,03**

Resultados – Classificação

- **Modelos Classificadores**







Discussão – Comparação com Li et al. 2018

- **Modelos Regressores**

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

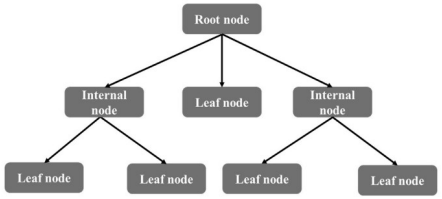
- **Modelos Classificadores**

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

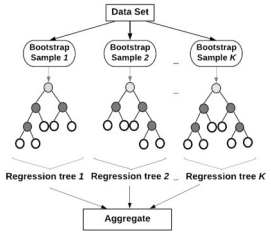
(LI et al, 2018)

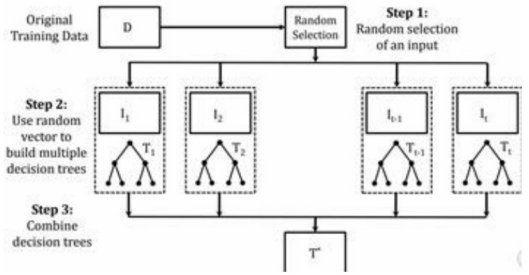
Discussão – Comparação entre modelos

- **Decision Trees**
  - Maior velocidade de predição
  - Explicabilidade elevada
  - Acurácia baixa



- **Modelos Combinados**
  - **Random Forest e Extra Trees**
  - Menor velocidade de predição
  - Explicabilidade baixa
  - Acurácia elevada



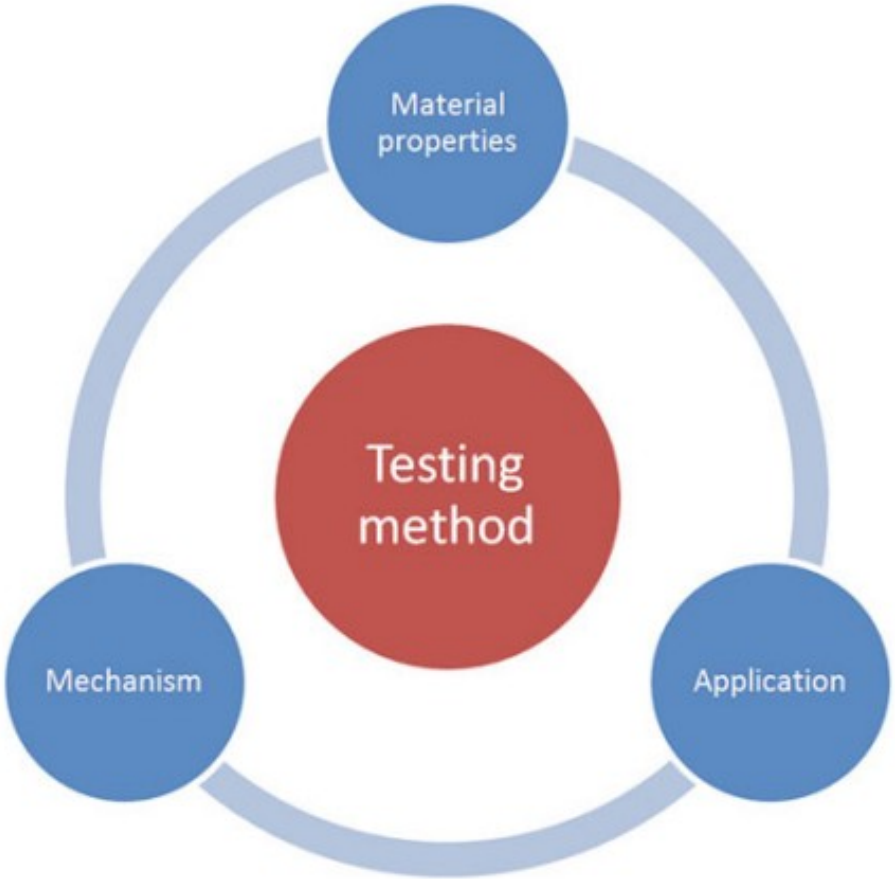


- **Tempo de Execução de uma Predição**

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

Considerações Finais

- **Modelos preditores:**
  - Estudos de novos materiais
  - Predição de propriedades
  - Sistemas de apoio à decisão
- **Limitações:**
  - Poucas amostras
  - Poucos estudos
  - Importância do tratamento de dados
- **Novos estudos:**
  - Maior quantidade de dados
  - Estudo de novos modelos
  - Aplicação de novas técnicas



(BODE et al.,2015)

Referências

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

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

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, abr. 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.

ZHENG, A.; CASARI, A. Feature Engineering for Machine Learning. Principles and Techniques for Data Scientists. **O' Reilly Media Inc**, Sebastopol, 2018

Obrigado!