Aprendizado de máquina e inteligência artificial em física - 2025

Atividade 2

Instruções: Produzir um notebook com a solução da atividade. Salvar o notebook em formato pdf legível. Enviar a atividade via Moodle até o dia 23/out 23:59. O prazo não será prorrogado. Não serão aceitas atividades enviadas por email.

Descrição do conjunto de dados:

Vamos utilizar um conjunto de dados sobre supercondutividade, disponível em: https://archive.ics.uci.edu/ml/datasets/superconductivty+data e em https://github.com/tiagofiorini/MLinPhysics/blob/main/Superconductivity.csv. O conjunto possui dados sobre cerca de 21 mil materiais, com 81 colunas. Nossa variável alvo será a temperatura crítica (*critical_temp*). Os atributos são baseados em propriedades físicas e químicas dos materiais, descritas na Tabela 1 abaixo. Os materiais possuem de 1 a 9 elementos em sua composição. A partir das propriedades listadas na Tabela 1, foram criados atributos como: a massa atômica média dos elementos presentes em cada material (*mean_atomic_mass*), a média ponderada da massa atômica (*wtd_mean_atomic_mass*), a entropia (entropy_atomic_mass), a faixa de valores de massa atômica (*range_atomic_mass*), e assim por diante. Dessa forma, para cada uma das 8 propriedades listadas na Tabela 1, existem 10 atributos calculados de acordo com a Tabela 2, totalizando 80 atributos para cada material.

Variable	Units	Description
Atomic Mass	atomic mass units (AMU)	total proton and neutron rest masses
First Ionization Energy	kilo-Joules per mole (kJ/mol)	energy required to remove a valence electron
Atomic Radius	picometer (pm)	calculated atomic radius
Density	kilograms per meters cubed (kg/m^3)	density at standard temperature and pressure
Electron Affinity	kilo-Joules per mole (kJ/mol)	energy required to add an electron to a neutral atom
Fusion Heat	kilo-Joules per mole (kJ/mol)	energy to change from solid to liquid without temperature change
Thermal Conductivity	watts per meter-Kelvin ($W/(m \times K)$)	thermal conductivity coefficient κ
Valence	no units	typical number of chemical bonds formed by the element

Tabela 1: Propriedades consideradas no conjunto de dados (Hamidieh, 2018).

Feature & Description	Formula	Sample Value
Mean	$= \mu = (t_1 + t_2)/2$	35.5
Weighted mean	$= \nu = (p_1 t_1) + (p_2 t_2)$	44.43
Geometric mean	$= (t_1 t_2)^{1/2}$	33.23
Weighted geometric mean	$= (t_1)^{p_1} (t_2)^{p_2}$	43.21
Entropy	$= -w_1 \ln(w_1) - w_2 \ln(w_2)$	0.63
Weighted entropy	$= -A \ln(A) - B \ln(B)$	0.26
Range	$= t_1 - t_2 \ (t_1 > t_2)$	25
Weighted range	$= p_1 t_1 - p_2 t_2$	37.86
Standard deviation	$= [(1/2)((t_1 - \mu)^2 + (t_2 - \mu)^2)]^{1/2}$	12.5
Weighted standard deviation	$= [p_1 (t_1 - \nu)^2 + p_2 (t_2 - \nu)^2)]^{1/2}$	8.75

Tabela 2: Métricas utilizadas para a criação de atributos, tendo como base as propriedades listadas na tabela 1. As fórmulas exemplificam os cálculos realizados para um material constituído por dois elementos, com propriedades distintas. Os valores numéricos mostram um exemplo para o material Re_7Zr_1 (Hamidieh, 2018).

Atividade: Fazer uma análise de regressão supervisionada para determinar a temperatura crítica de materiais. Utilizar diferentes modelos e subconjuntos de dados, comparando o desempenho e a interpretabilidade.

1) Análise exploratória

2) Preparação dos dados: escalonamento, transformação, particionamento

3) Avaliar a importância dos atributos com base nos coeficientes de um modelo de regressão linear múltipla (com ou sem regularização, à sua escolha).

4) Avaliar a importância dos atributos com base em um regressor Random Forest ou Gradient Boosting.

5) Com base nos resultados dos itens 3 e 4, selecionar os atributos mais importantes. O número de atributos fica à sua escolha. Justificar sua escolha. Discutir brevemente se os atributos escolhidos possuem significado físico, ou seja, se de fato pode existir uma relação com a variável alvo (temperatura crítica).

6) Aplicar uma técnica de redução de dimensionalidade, como PCA (análise de componentes principais), criando novos atributos a partir de uma combinação dos atributos originais. O número de componentes principais a serem utilizadas fica à sua escolha (justificar escolha).

7) Construir um modelo de regressão linear múltipla com:
    a) os atributos mais importantes, escolhidos no item 5;
    b) usando as componentes principais como atributos (item 6).

   Comparar o desempenho desses modelos na predição da temperatura crítica e a sua interpretabilidade (isto é, se é fácil ou não interpretar o significado físico dos coeficientes ajustados). Avalie se o modelo é capaz de predizer diferentes faixas de valores de temperatura crítica.

8) Construir um modelo de regressão baseado em Random Forest ou Gradient Boosting com:
    a) apenas os atributos mais importantes, escolhidos no item 5;
    b) usando as componentes principais como atributos (item 6).

   Lembre-se de otimizar os hiperparâmetros. Comparar o desempenho desses modelos na predição da temperatura crítica e sua interpretabilidade. Avalie se o modelo é capaz de predizer diferentes faixas de valores de temperatura crítica.

9) Fazer uma breve discussão crítica sobre o desempenho, a interpretabilidade e o custo computacional dos modelos lineares e dos modelos baseados em árvores de decisão.