Apresentação TAAD

Gonçalo Moreira Dias
Professor Dr. Ricardo Gonçalo
4/2/22

Machine learning modeling of superconducting critical temperature

Artigo escolhido

Introdução

Introdução

A supercondutividade tem sido o foco de um enorme esforço de pesquisa desde sua descoberta há mais de um século. No entanto, ainda existem vários mistérios em torno deste tema.

A ligação entre a supercondutividade e as propriedades químicas/estruturais dos materiais é o principal.

Foram criados vários esquemas de ML que são desenvolvidos para modelar as temperaturas críticas (T_C) dos mais de 12000 supercondutores conhecidos.

Objetivo

Objetivo

Inicial: Criar modelos separados para prever os valores de T_C para compostos de **cobre aniónico**, à **base de ferro** e de **baixa T_C**.

Final: Com base nos modelos de regressão iniciais, são combinados num *pipeline* integrado no banco de dados de estrutura cristalográfica inorgânica (ICSD) para identificar potenciais novos supercondutores.

Resultados e discussão

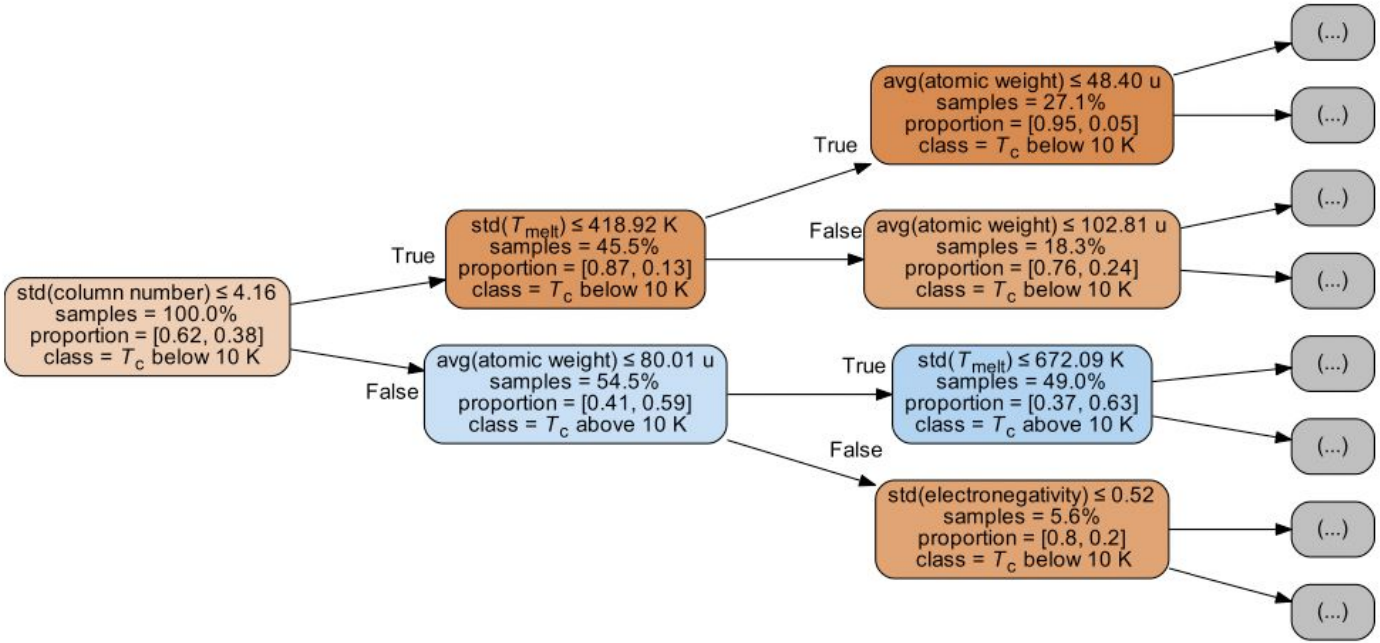
Métodos e resultados

Todos os algoritmos usados são variantes do método **Random Forest**.

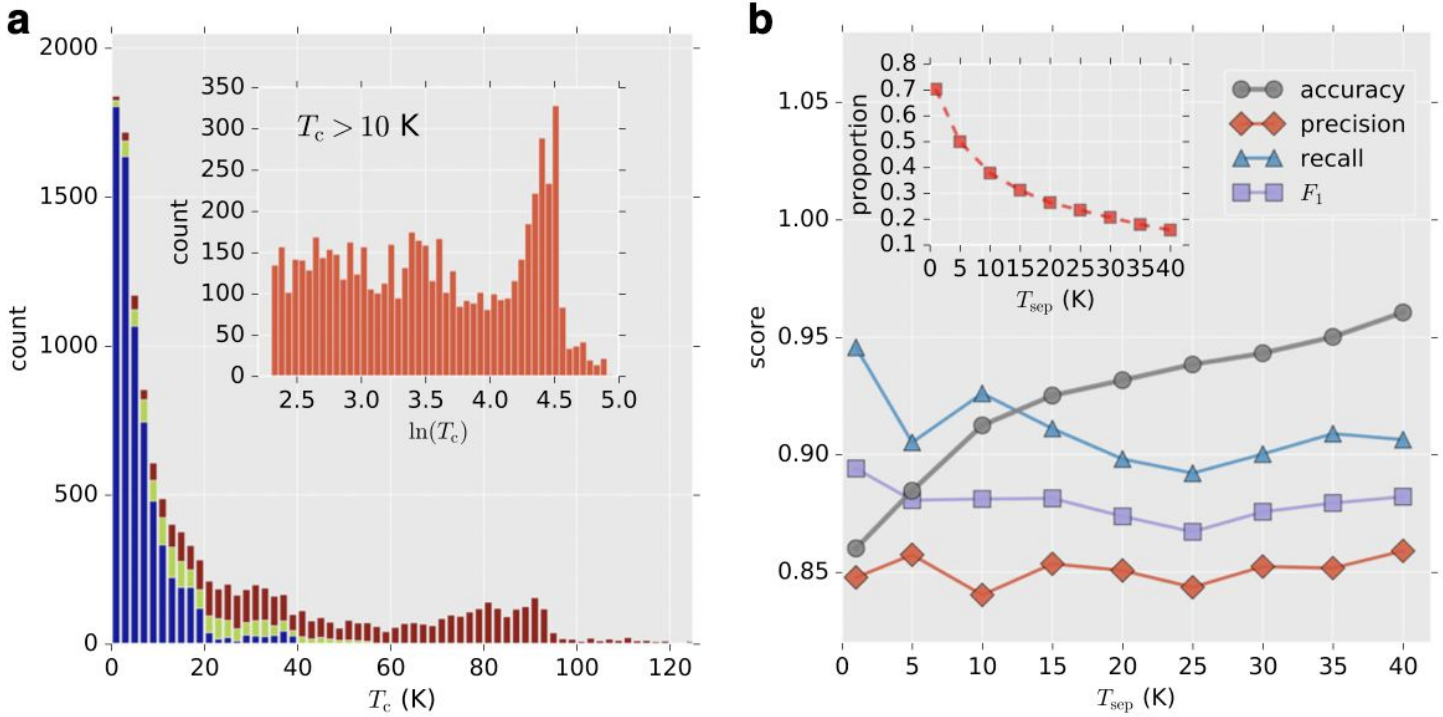
Foi feita uma separação desde início entre valores **acima e abaixo de 10K**. Criara-se assim **2 classes**.

Foi um **parâmetro de ajuste T_{sep}** . Ajudou, à semelhança de outros pormenores, a melhorar a capacidade de previsão do modelo.

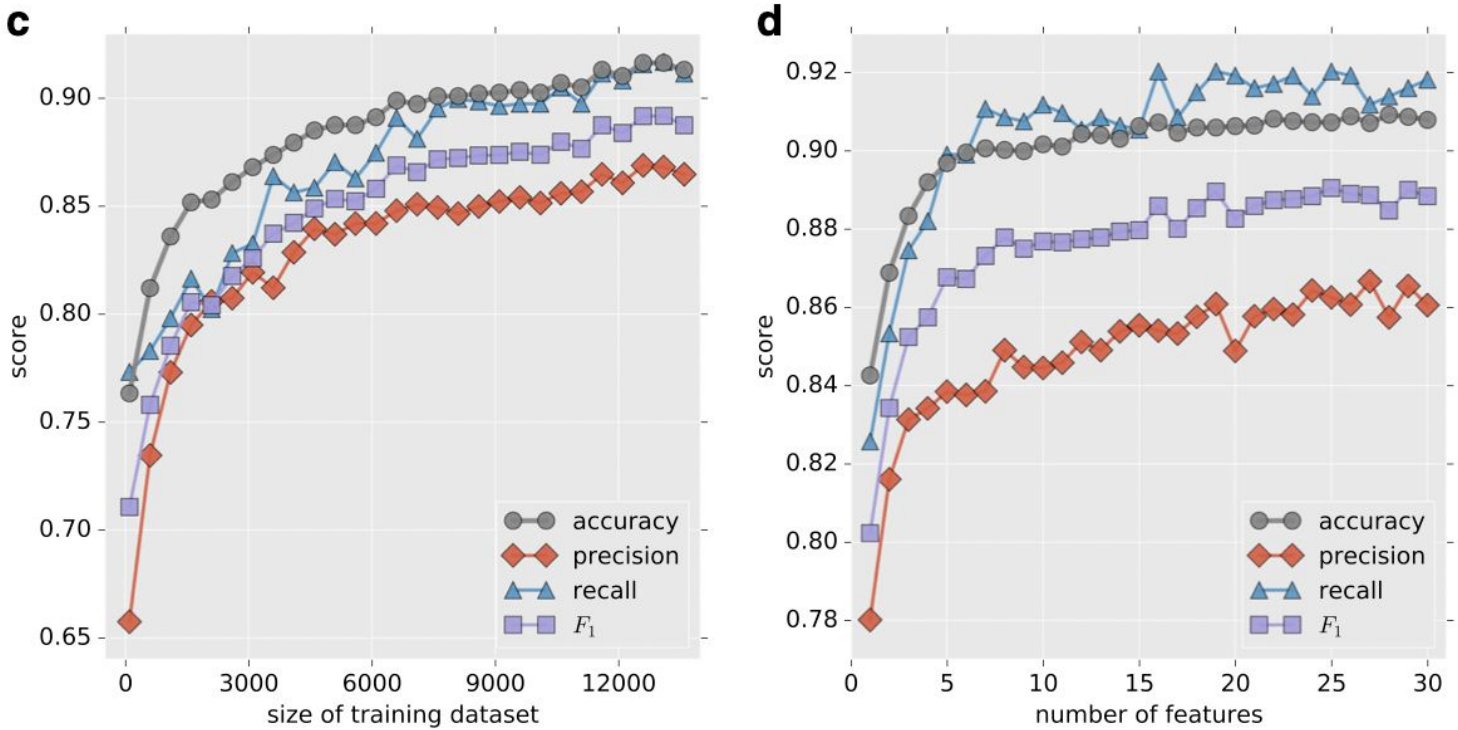
Exemplo de *decision tree*



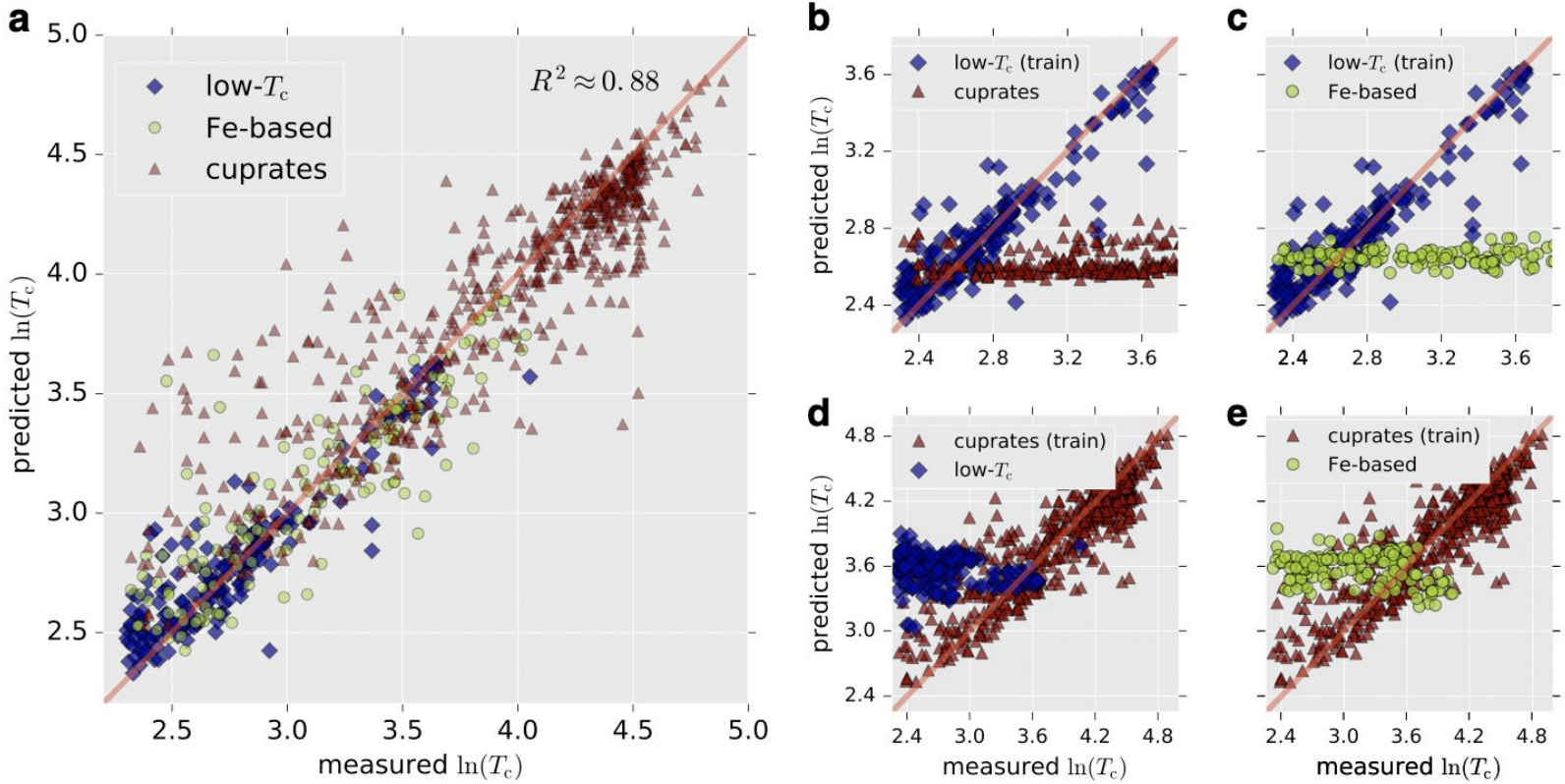
Construção e teste - modelos classificativos



Construção e teste - modelos classificativos



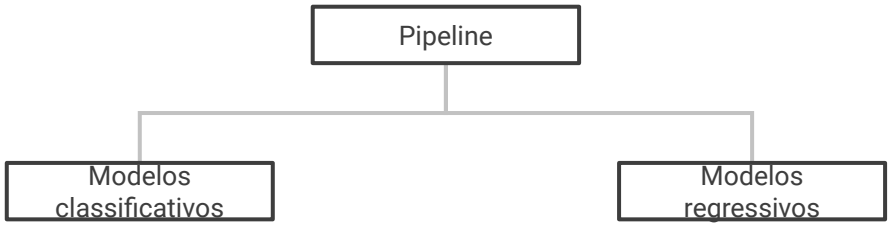
Previsão de T_C - modelos regressivos



Potenciais supercondutores

Foi estudada a correlação entre T_C e propriedades como:

- Massa atómica
- Raio covalente
- Número de eletrões de valência
- Número de orbitais não preenchidas



AFLOW

Table 3. List of potential superconductors identified by the pipeline

Compound	ICSD	SYM
$CsBe(AsO_4)$	074027	Orthorhombic
$RbAsO_2$	413150	Orthorhombic
$KSbO_2$	411214	Monoclinic
$RbSbO_2$	411216	Monoclinic
$CsSbO_2$	059329	Monoclinic
$AgCrO_2$	004149/025624	Hexagonal
$K_{0.8}(Li_{0.2}Sn_{0.76})O_2$	262638	Hexagonal
$Cs(MoZn)(O_3F_3)$	018082	Cubic
$Na_3Cd_2(IrO_6)$	404507	Monoclinic
$Sr_3Cd(PtO_6)$	280518	Hexagonal
$Sr_3Zn(PtO_6)$	280519	Hexagonal
$(Ba_5Br_2)Ru_2O_9$	245668	Hexagonal
$Ba_4(AgO_2)(AuO_4)$	072329	Orthorhombic
$Sr_5(AuO_4)_2$	071965	Orthorhombic
$RbSeO_2F$	078399	Cubic
$CsSeO_2F$	078400	Cubic
$KTeO_2F$	411068	Monoclinic
$Na_2K_4(Tl2O_6)$	074956	Monoclinic
$Na_3Ni_2BiO_6$	237391	Monoclinic
$Na_3Ca_2BiO_6$	240975	Orthorhombic
CsCd(BO3)	189199	Cubic
$K_2Cd(SiO_4)$	083229/086917	Orthorhombic
$Rb_2Cd(SiO_4)$	093879	Orthorhombic
$K_2Zn(SiO_4)$	083227	Orthorhombic
$K_2Zn(Si2O_6)$	079705	Orthorhombic
$K_2Zn(GeO_4)$	069018/085006/085007	Orthorhombic
$(K_{0.6}Na_{1.4})Zn(GeO_4)$	069166	Orthorhombic
$K_2Zn(Ge_2O_6)$	065740	Orthorhombic
$Na_6Ca_3(Ge_2O_6)_3$	067315	Hexagonal
$Cs_3(AlGe_2O_7)$	412140	Monoclinic
$K_4Ba(Ge_3O_9)$	100203	Monoclinic
$K_{16}Sr_4(Ge_3O_9)_4$	100202	Cubic
$K_3Tb[Ge_3O_8(OH)_2]$	193585	Orthorhombic
$K_3Eu[Ge_3O_8(OH)_2]$	262677	Orthorhombic
$KBa_6Zn_4(Ga_7O_{21})$	040856	Trigonal

Also shown are their ICSD numbers and symmetries. Note that for some compounds there are several entries. All of the materials contain oxygen

Conclusão

Conclusão

Demonstrou-se que os modelos de ML são importantes em desempenhar a pesquisa de supercondutividade.

Com base nos registos da SuperCon e outros bancos de dados pode ser possível uma melhor **compreensão da possível conexão entre a química/estrutura dos materiais** e, consequentemente, a supercondutividade.

A aplicação de algoritmos de ML sofisticados tem o potencial de **acelerar drasticamente** a busca por supercondutores de alta temperatura candidatos.

Machine Learning in Material Science

Orientadores do Projeto: Márcio Ferreira e Tiago Cerqueira

Índice

Introdução e enquadramento

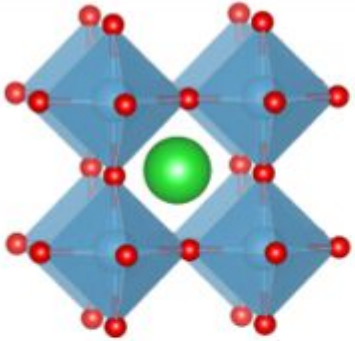
Introdução e enquadramento

Descobrir novos materiais, com propriedades interessantes, ao fazer uma **combinação aleatória de elementos**.

Uma das propriedades mais importantes para a **estabilidade de novos materiais é a Energia de Hull** (Ehull).

	A	B	X	Ehull
0	W	Ir	Pt	2210.302
1	Ga	Tc	Pb	537.228
2	Se	Ba	Cu	2274.020
3	Zr	P	C	1716.121
4	Mn	Te	Hg	826.878
...	...	...	...	...
170832	Pt	Cu	Sc	771.021
170833	Ga	Pt	Sb	767.354
170834	Cl	W	Fe	1856.558
170835	Pd	Se	Hf	944.208
170836	As	P	N	1154.846

170837 rows × 4 columns



ABX_3

	IA	IIA	IIIB	IVB	VB	VIB	VIIB	VIIIB	VIIIB	VIIIB	IB	IIB	IIIA	IVA	VA	VIA	VIIA	VIIIA
1	1 H																	2 He
2	3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
3	11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
4	19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
5	37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
6	55 Cs	56 Ba	57 La	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn

Objetivo

Através de algoritmos de ML queremos prever o valor “Ehull” para uma certa combinação de elementos **ABX**.

Essa previsão é feita com base nas **propriedades intrínsecas de cada elemento**, que são as features associadas a cada elemento.

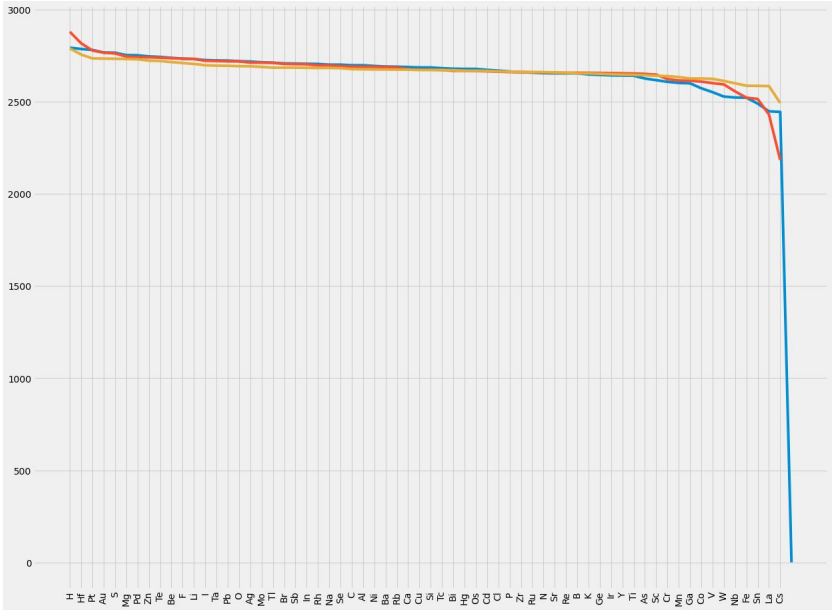
Minimizar o MAE (Mean Absolute Error).

Tratamento inicial dos dados

Eliminar o elemento ‘Lu’

	A	B	X	Ehull
0	W	Ir	Pt	2210.302
1	Ga	Tc	Pb	537.228
2	Se	Ba	Cu	2274.020
3	Zr	P	C	1716.121
4	Mn	Te	Hg	826.878
...	...	...	...	...
170832	Pt	Cu	Sc	771.021
170833	Ga	Pt	Sb	767.354
170834	Cl	W	Fe	1856.558
170835	Pd	Se	Hf	944.208
170836	As	P	N	1154.846

170837 rows × 4 columns



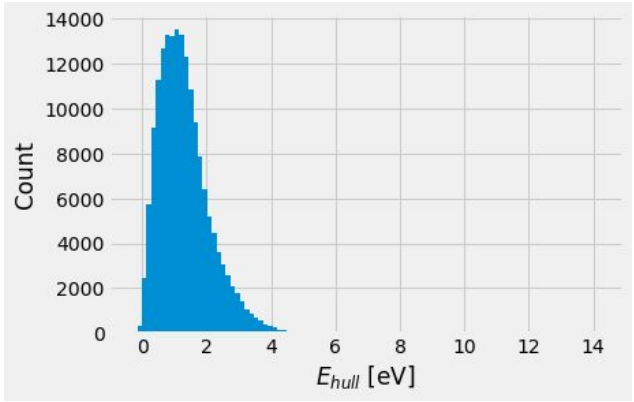

```
data[data.A == "Lu"]
```

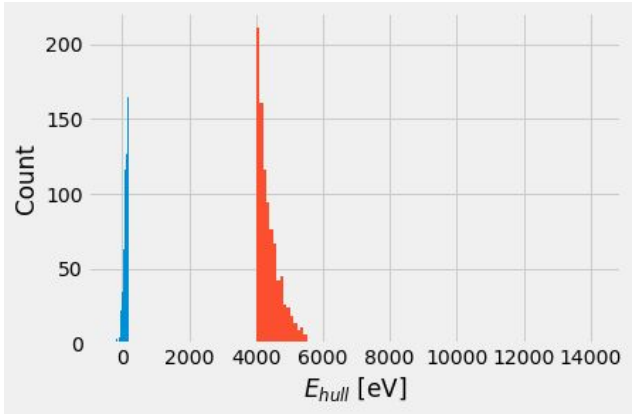
	A	B	X	Ehull
61835	Lu	Y	Te	799.615

Eliminar possíveis *outliers*

	A	B	X	Ehull
0	W	Ir	Pt	2210.302
1	Ga	Tc	Pb	537.228
2	Se	Ba	Cu	2274.020
3	Zr	P	C	1716.121
4	Mn	Te	Hg	826.878
...	...	...	...	...
170832	Pt	Cu	Sc	771.021
170833	Ga	Pt	Sb	767.354
170834	Cl	W	Fe	1856.558
170835	Pd	Se	Hf	944.208
170836	As	P	N	1154.846

170837 rows × 4 columns

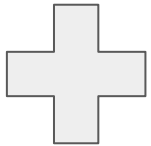




Preparação do dataset

	A	B	X	Ehull
0	W	Ir	Pt	2210.302
1	Ga	Tc	Pb	537.228
2	Se	Ba	Cu	2274.020
3	Zr	P	C	1716.121
4	Mn	Te	Hg	826.878
...	...	...	...	...
170832	Pt	Cu	Sc	771.021
170833	Ga	Pt	Sb	767.354
170834	Cl	W	Fe	1856.558
170835	Pd	Se	Hf	944.208
170836	As	P	N	1154.846

170837 rows × 4 columns



Biblioteca - Element()


```
{'Atomic mass': 1.00794,
 'Atomic no': 1,
 'Atomic orbitals': {'1s': -0.233471},
 'Atomic radius': 0.25,
 'Atomic radius calculated': 0.53,
 'Boiling point': '20.28 K',
 'Brinell hardness': 'no data MN m<sup>-2</sup>',
 'Bulk modulus': 'no data GPa',
 'Coefficient of linear thermal expansion': 'no data x10<sup>-6</sup>K<sup>-1</sup>',
 'Common oxidation states': [-1, 1],
 'Critical temperature': '33 K',
 'Density of solid': 'no data kg m<sup>-3</sup>',
 'Electrical resistivity': 'no data 10<sup>-8</sup> &Omega; m',
 'Electronic structure': '1s<sup>1</sup>',
 'ICSD oxidation states': [1, -1],
 'Liquid range': '6.27 K',
 'Melting point': '14.01 K',
 'Mendeleev no': 103,
 'Mineral hardness': 'no data',
 'Molar volume': '11.42 cm<sup>3</sup>',
 'Name': 'Hydrogen',
 'Oxidation states': [-1, 1],
 'Poissons ratio': 'no data',
 'Reflectivity': 'no data %',
 'Refractive index': '1.000132 (gas; liquid 1.12)(no units)',
 'Rigidity modulus': 'no data GPa',
 'Shannon radii': {'1': {'I': {'': {'crystal_radius': -0.24,
```

Dataset final

Y ⇩ X →

	A	B	X	Ehull	mass_A	ns_A	np_A	nd_A	elecneg_A	radius_A	...	Boiling point_X	Liquid range_X	Melting point_X	Mendeleev no_X	Molar volume_X	Therm conductivity_
0	W	Ir	Pt	2210.302	183.840000	2.0	0.0	4.0	1.470	193.0	...	4098.00	2056.60	2041.40	68.0	9.09	72.0000
1	Ga	Tc	Pb	537.228	69.723000	2.0	1.0	10.0	1.756	136.0	...	2022.00	1421.39	600.61	82.0	18.26	35.0000
2	Se	Ba	Cu	2274.020	78.960000	2.0	4.0	10.0	2.424	103.0	...	3200.00	1842.23	1357.77	72.0	7.11	400.0000
3	Zr	P	C	1716.121	91.224000	2.0	0.0	2.0	1.320	206.0	...	4300.00	500.00	3800.00	95.0	5.29	140.0000
4	Mn	Te	Hg	826.878	54.938045	2.0	0.0	5.0	1.750	161.0	...	629.88	395.56	234.32	74.0	14.09	8.3000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
170832	Pt	Cu	Sc	771.021	195.084000	1.0	0.0	9.0	1.720	177.0	...	3103.00	1289.00	1814.00	19.0	15.00	16.0000
170833	Ga	Pt	Sb	767.354	69.723000	2.0	1.0	10.0	1.756	136.0	...	1860.00	956.22	903.78	88.0	18.19	24.0000
170834	Cl	W	Fe	1856.558	35.453000	2.0	5.0	0.0	2.869	79.0	...	3134.00	1323.00	1811.00	61.0	7.09	80.0000
170835	Pd	Se	Hf	944.208	106.420000	0.0	0.0	10.0	1.580	169.0	...	4876.00	2370.00	2506.00	50.0	13.44	23.0000
170836	As	P	N	1154.846	74.921600	2.0	3.0	10.0	2.211	114.0	...	77.36	14.31	63.05	100.0	13.54	0.0258

170837 rows × 61 columns

Algoritmos de ML

Regressão linear
Regressão ridge
Regressão lasso
Random forest

Regressão linear, ridge e lasso

Por vezes quando é utilizado o modelo de regressão linear pode existir *overfitting* na fase de treino.

Como tal, criaram-se os modelos de ridge, lasso e outros.

Ridge ou regularização L2 :

$$\sum_{i=1}^{n}(y - Xw)^2 + alpha\sum_{j=1}^{p} w_j{}^2$$

Lasso ou regularização L1 :

$$\frac{1}{2m}\sum_{i=1}^{m}(y - Xw)^2 + alpha\sum_{j=1}^{p}\left|w_j\right|$$

Random forest

Conjunto de árvores de decisão.

É um estimador que ajusta vários classificadores de árvore de decisão em várias sub amostras do conjunto de dados e usa a média para melhorar a precisão preditiva e controlar o *overfitting*.

Regressão linear

Treino - 90% Teste - 10% (Random state= 20)	
R squared (%)	64.38
MAE - teste	361.3
MAE - treino	364.3
RMS Error	469.1

Treino - 50% Teste - 50% (Random state= 20)	
R squared (%)	64.37
MAE - teste	364.9
MAE - treino	363.8
RMS Error	474.7

Regressão Ridge

Treino - 70% Teste - 30% (alpha = 1.0)	
R squared (%)	64.37
MAE - teste	364.3
MAE - treino	363.9
RMS Error	476.4

Treino - 50% Teste - 50% (alpha = 25.0)	
R squared (%)	64.36
MAE - teste	364.4
MAE - treino	363.9
RMS Error	474.7

Treino - 90% Teste - 10% (alpha = 0.01)	
R squared (%)	64.38
MAE - teste	361.3
MAE - treino	364.3
RMS Error	469.1

Regressão Lasso

Treino - 70% Teste - 30% (alpha = 1.0)	
R squared (%)	64.22
MAE - teste	365.6
MAE - treino	365.1
RMS Error	477.5

Treino - 50% Teste - 50% (alpha = 25.0)	
R squared (%)	61.14
MAE - teste	381.2
MAE - treino	381.2
RMS Error	495.6

Treino - 90% Teste - 10% (alpha = 0.01)	
R squared (%)	64.36
MAE - teste	361.4
MAE - treino	364.4
RMS Error	469.2

Random forest

Treino - 90% Teste - 10% (Random state= 20)	
R squared (%)	99.42
MAE - teste	69.84
MAE - treino	23.59
RMS Error	105.5

Hiperparâmetros:

bootstrap=False,
max_depth=30,
max_features='sqrt',
min_samples_leaf=2,
min_samples_split=4,
n_estimators=200

Comparação de resultados

MAE	ML model
363.1	Regressão linear
363.3	Regressão ridge
369.4	Regressão lasso
69,84	Random forest

MAE	machine learning model
298.9 ± 0.3	ridge regression
155.5 ± 4.8	neural network
140.0 ± 0.6	random forests
126.6 ± 1.0	AdaBoost/random forests
123.1 ± 0.8	extremely randomized trees
121.3 ± 0.8	AdaBoost/extremely randomized trees

Conclusão

Conclusão

Posso concluir que o modelo que apresentou melhores resultados foi o **Random forest.** Foram bastante positivos.

Os restantes modelos mostraram não ser os mais indicados para este tipo de dados.

Podendo a escolha dos hiperparâmetros ter influenciado os resultados finais.