Machine learning aplicado à espectroscopia molecular

Aluno: Alexandre Esposte Santos

Motivação

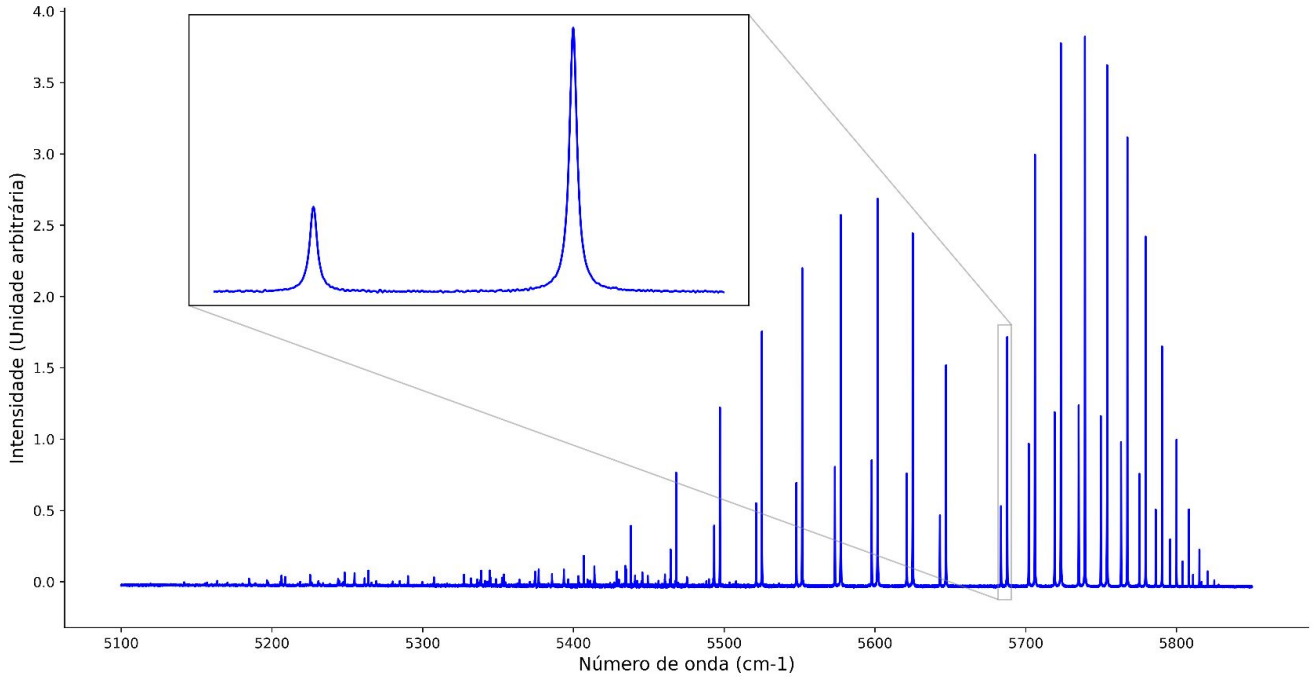
- Escassez de trabalhos utilizando I.A em análise de espectros moleculares de alta resolução por transformada de Fourier
- Utilizar métodos de Inteligência artificial para auxiliar e/ou automatizar análises de espectros moleculares
- Obter informações relevantes através de algoritmos de Machine Learning

Dados

Ao todo foram 13 espectros da molécula de HCl em diferentes temperaturas e pressões.

Temp	-	Pressão (mbar)
20°C	-	78, 145, 200, 398, 790
40°C	-	27, 76.9, 141, 211
42°C	-	181, 218, 301, 439

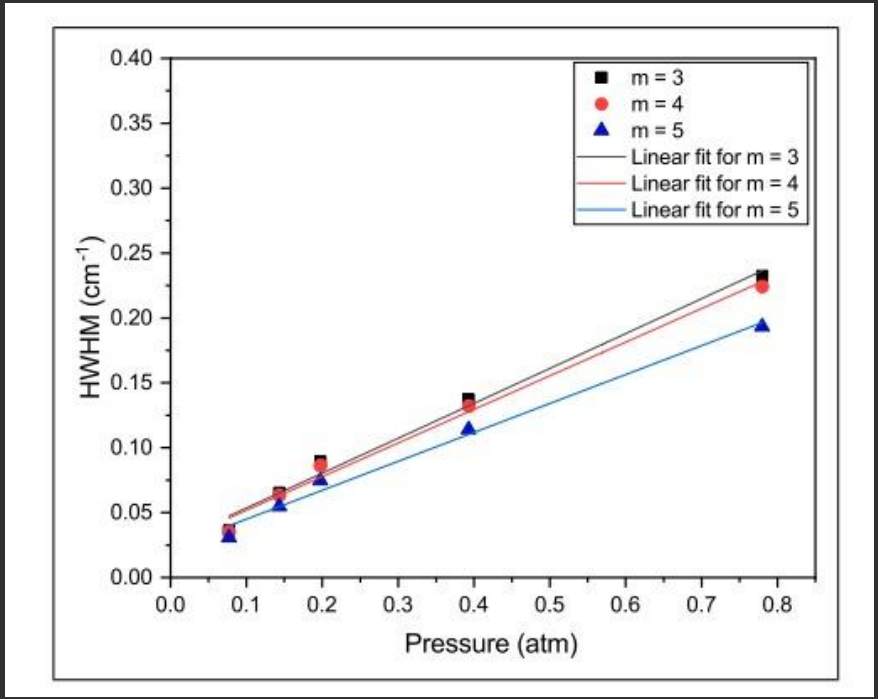
O espectro


Intensidade (Unidade arbitrária)
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
5100
5200
5300
5400
5500
5600
5700
5800
Número de onda (cm-1)

Objetivo

- Estimar as larguras das linhas espectrais

O que obtemos com as larguras



Etapas do processo

Features

Estatísticas obtidas com relação a intensidade e ao número de onda

- média
- mediana
- desvio padrão
- máximo valor
- mínimo valor
- curtose
- assimetria

Dataset

- 421 linhas espectrais
- 23 colunas/features/características

Dataset

	wavenumber	intensity	j	branch	pressure	temperature	fwhm	gamma	sigma	mean_wv	std_wv	skew_wv	kurtosis_wv	max_wv	min_wv	median_wv
0	5683.56703	0.87873	0	R37	145	293	0.068169	0.033105	0.004739	5683.566085	0.029765	0.000005	-1.199996	5683.61693	5683.51524	5683.56608
1	5701.98259	1.23987	1	R37	145	293	0.077525	0.037694	0.005280	5701.981648	0.034114	-0.000005	-1.199997	5702.04003	5701.92327	5701.98165
2	5719.16744	1.36729	2	R37	145	293	0.080923	0.039601	0.004839	5719.166023	0.035473	0.000010	-1.200007	5719.22676	5719.10529	5719.16602
3	5735.10744	1.37998	3	R37	145	293	0.079175	0.038495	0.005394	5735.106026	0.034386	0.000023	-1.199989	5735.16488	5735.04717	5735.10603
4	5749.79130	1.32506	4	R37	145	293	0.069696	0.034073	0.004263	5749.790830	0.030580	0.000007	-1.199999	5749.84309	5749.73857	5749.79083

mean_int	std_int	skew_int	kurtosis_int	max_int	min_int	median_int
0.567632	0.204984	0.105487	-1.423795	0.87873	0.26557	0.549450
0.805899	0.290988	0.084993	-1.431570	1.23987	0.37273	0.784530
0.893575	0.319793	0.065195	-1.431605	1.36729	0.41227	0.874895
0.903325	0.320822	0.067481	-1.428374	1.37998	0.41933	0.883840
0.863854	0.309285	0.075582	-1.441434	1.32506	0.40452	0.846285

Split

Somente as variáveis estatísticas foram utilizadas no split

- Proporção 80/20
- 336 linhas para treino
- 85 linhas para teste/validação

Modelos utilizados

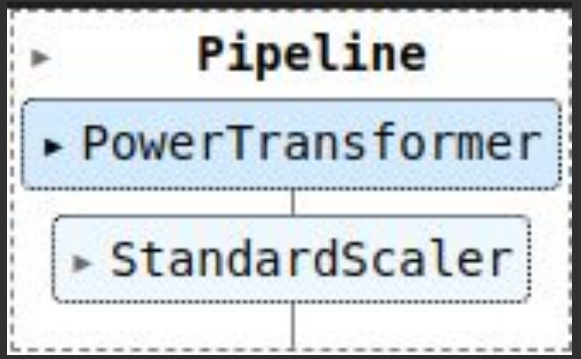
Ao todo foram 10 modelos

- Regressão linear
- Regressão Lasso
- Regressão Ridge
- K-Vizinhos mais próximos
- AdaBoost
- Gradient Boosting
- Light Gradient Boosting Machine
- Random Forest
- Extreme Gradient Boosting
- Support Vector Machine

Métricas de avaliação

- Erro absoluto médio (MAE)
- Raiz quadrada do erro quadrático médio (RMSE)
- Erro percentual absoluto médio (MAPE)

Pipeline



Resultados

Treinamento efetuado com validação cruzada em 5 folds

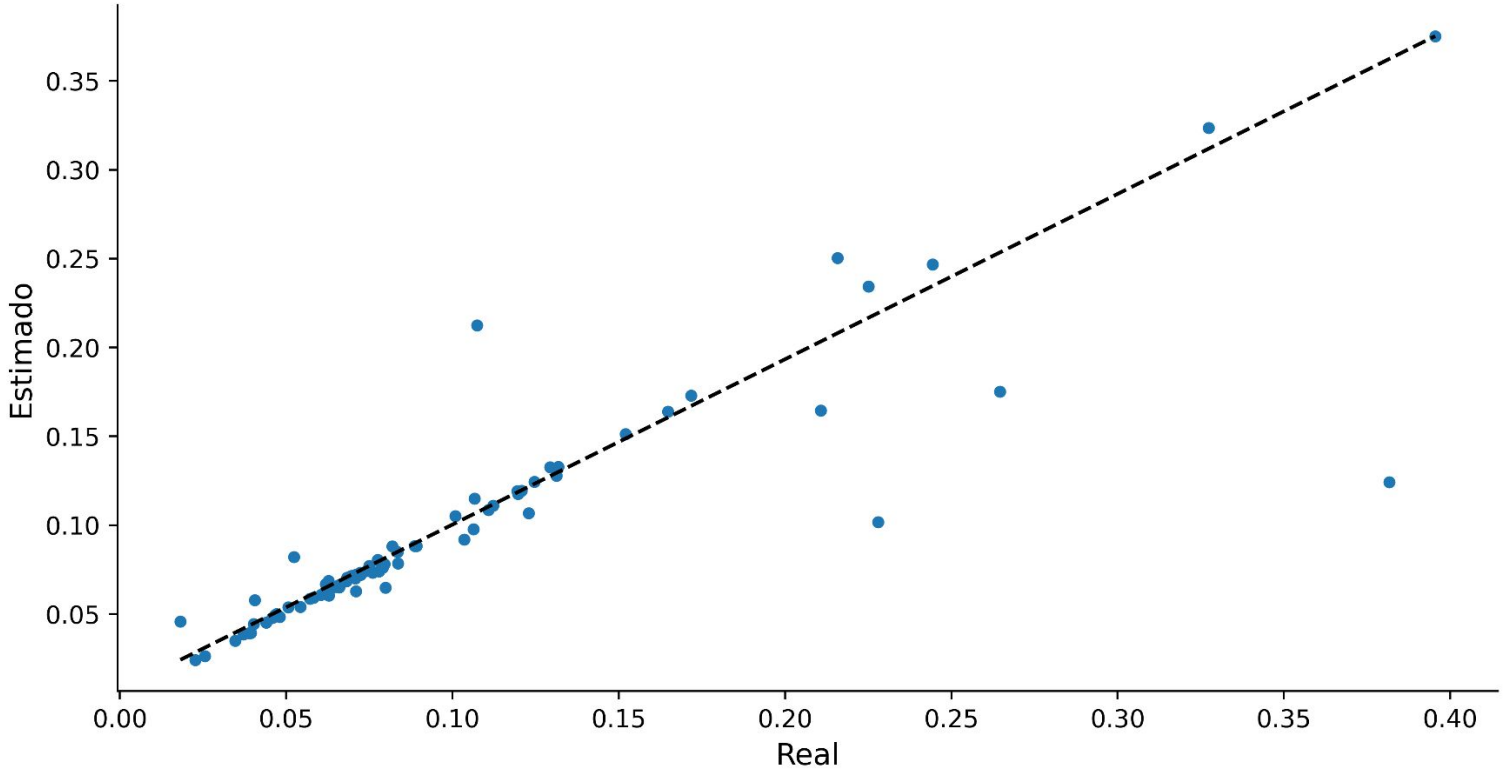
	mae	mape	rms
floresta	0.008218	0.096629	0.023028
gbr	0.008487	0.101332	0.023133
xgb	0.009221	0.112183	0.024020
lgbm	0.011823	0.130744	0.025592
regressao linear	0.013537	0.168398	0.022654
knn	0.020078	0.229507	0.033430
ridge	0.021580	0.255828	0.034647
ada	0.019214	0.261364	0.027993
lasso	0.051288	0.654170	0.073769
svm	0.059708	0.853559	0.067751

O melhor modelo

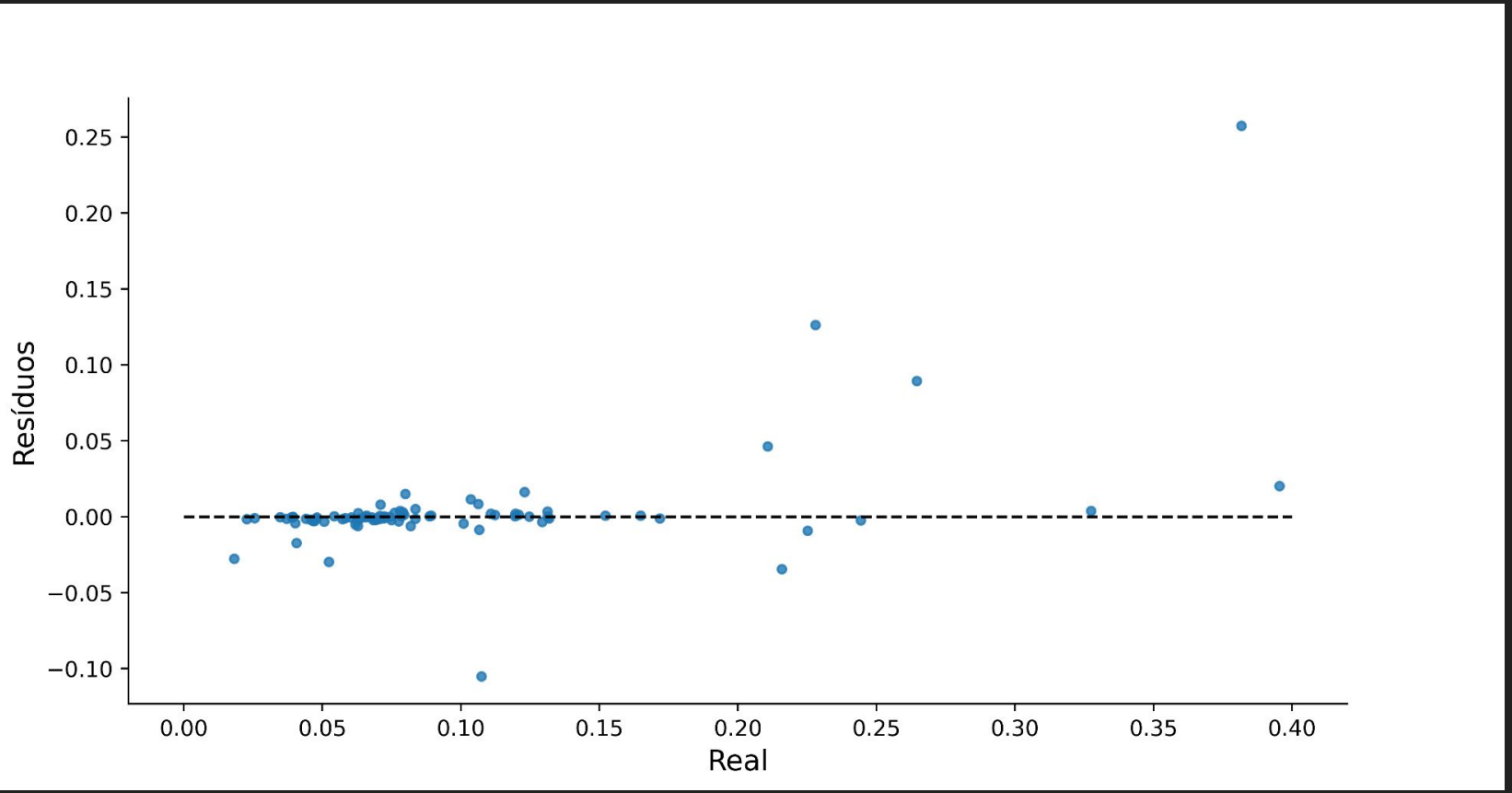
Treinamento efetuado com validação cruzada em 5 folds

- Random Forest Regressor
- Mape Treino = 9,66%
- Mape validação = 9,36%

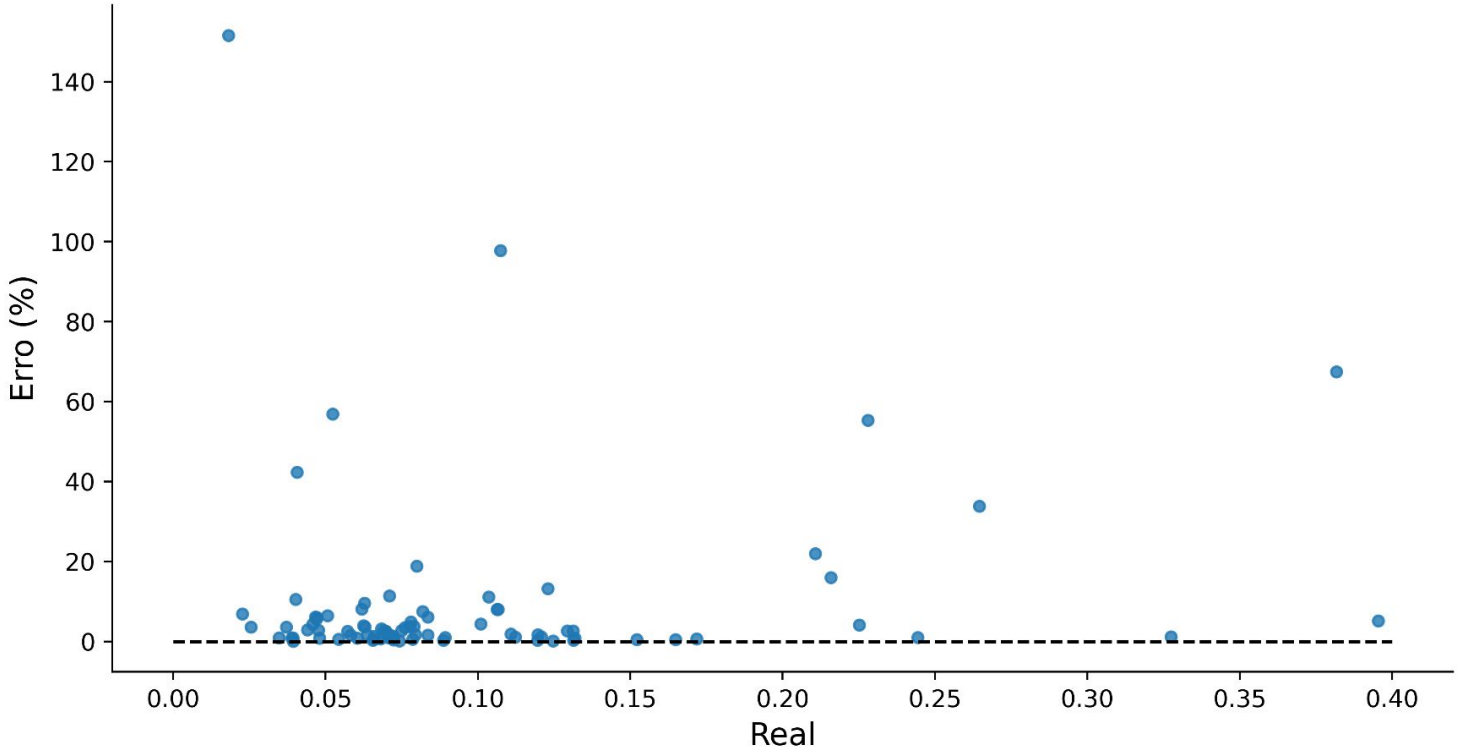
Curva estimado x real



Resíduos x Real



Erro x Real


Erro (%)
140
120
100
80
60
40
20
0
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Real

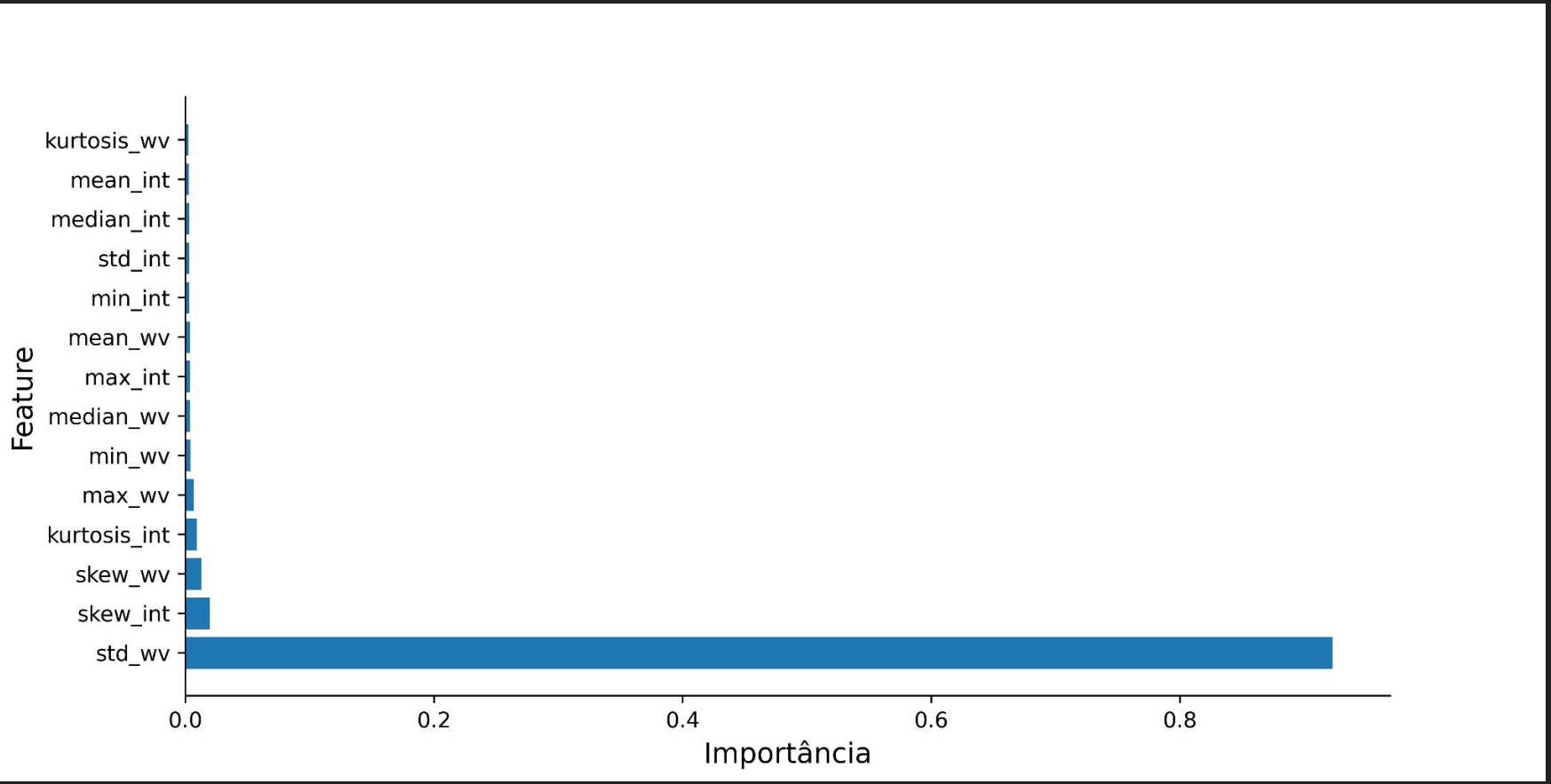
Alguns percentis

50% das estimativas tem erro menor igual a 2.63%

60% das estimativas tem erro menor igual a 3.66%

80% das estimativas tem erro menor igual a 7.98%

Feature importance



Conclusão

- Sucesso na aplicação de Machine learning em espectroscopia molecular
- Obtenção das larguras com estatísticas básicas da linha espectral
- Entendimento sobre qual a variável considerada mais importante
- Erro médio percentual de 9.36%

Referências

1. "Scikit-learn: Machine Learning in Python," Scikit-learn, 2022. [Online]. Available: https://scikit-learn.org.

2. "pandas: Powerful data structures for data analysis, time series, and statistics," pandas, 2021. [Online]. Available: https://pandas.pydata.org.

3. "NumPy: The fundamental package for scientific computing with Python," NumPy, 2020. [Online]. Available: https://numpy.org.

4. "Matplotlib: Visualization with Python," Matplotlib, 2021. [Online]. Available: https://matplotlib.org.

Obrigado pela atenção

Email: alexandreesposte@id.uff.br / alexandreesposte@gmail.com