Projeto final - Machine Learning aplicado à espectrocopia molecular

Alexandre Esposte Santos

7 de dezembro de 2023

1 Introdução

Este trabalho surge em resposta à escassez de estudos que empregam inteligência artificial na análise de espectros moleculares de alta resolução por transformada de Fourier. A motivação para este estudo reside na necessidade de utilizar métodos de inteligência artificial para auxiliar e/ou automatizar análises de espectros moleculares, visando obter informações relevantes por meio de algoritmos de machine learning .

A análise de espectros moleculares é fundamental em diversas áreas, incluindo química, física e bioquímica. Nesse contexto, a aplicação de técnicas de machine learning pode proporcionar avanços significativos, permitindo a estimativa de larguras das linhas espectrais e a extração de informações valiosas a partir dos dados espectroscópicos.

Ao longo deste estudo, são abordadas as etapas do processo, os modelos utilizados, as métricas de avaliação e os resultados obtidos, contribuindo para o avanço do conhecimento no campo da espectroscopia molecular e demonstrando o potencial da integração entre inteligência artificial e análise de espectros.

2 Conjunto de Dados

O conjunto de dados utilizado neste trabalho consiste em 421 linhas espectrais e 23 colunas ou características que descrevem as propriedades das linhas espectrais da molécula de ácido clorídrico. A criação desse conjunto envolveu a coleta de estatísticas básicas, tais como média, mediana, desvio padrão, máximo, mínimo, curtose e assimetria das linhas espectrais de 13 espectros distintos, estando cada um deles em diferentes temperaturas e pressões.

3 Objetivo

A partir desse conjunto de dados, o estudo buscou aplicar técnicas de machine learning para estimar as larguras das linhas espectrais e obter insights sobre as variáveis consideradas mais importantes consideradas pelo melhor modelo treinado.

4 Pré-Processamento de Dados

No processo de pré-processamento dos dados, implementamos uma pipeline composta por duas transformações consecutivas: o Power Transformer e o Standard Scaler. A primeira transformação foi aplicada às features, visando modificar suas distribuições originais para torná-las mais semelhantes à distribuição normal. Já a segunda transformação foi responsável por padronizar as features, garantindo que todas elas sejam representadas em uma mesma escala.

5 Split dos Dados

A divisão dos dados em treino e teste foram realizados utilizando uma proporção de 80% para treino e 20% para o teste. Desse modo, 396 linhas foram para treino e 85 para teste.

6 Modelagem

Os modelos utilizados na modelagem incluem a regressão linear, regressão Lasso, regressão Ridge, k-vizinhos mais próximos, AdaBoost, Gradient Boosting, Light Gradient Boosting Machine, Random Forest, Extreme Gradient Boosting e Support Vector Machine, totalizando 10 modelos . Cada modelo foi treinado e avaliado através da aplicação de validação cruzada em 5 folds no conjunto de dados de treinamento. Posteriormente, realizou-se uma comparação dos modelos treinados por meio de métricas de erro, que abrangem o erro absoluto médio, a raiz quadrada do erro quadrático médio e o erro percentual absoluto médio.

7 Resultados

O Random Forest foi selecionado como sendo o melhor modelo, obtendo um erro médio percentual de 9,66% durante a validação cruzada e 9,36% nos dados de teste. Esses números refletem a notável capacidade do modelo em estimar com certa precisão as larguras das linhas espectrais.

Após o treinamento do modelo, realizou-se uma análise dos erros, permitindo a identificação de seus pontos fortes e fracos. Observou-se um número limitado de estimativas com erro elevado, que impactaram negativamente as métricas de erro, uma vez que todas são baseadas em média. Essa questão foi resolvida ao analisar os percentis do erro, revelando que, na mediana, o erro percentual é reduzido para 2,63%.

Além disso, a análise das features proporcionou insights valiosos sobre quais características estatísticas da linha espectral desempenham um papel mais significativo na estimativa das larguras das linhas espectrais. Esse conhecimento não apenas aprimora a compreensão do funcionamento do modelo, mas também contribui para a geração de insights analíticos.

8 Conclusão

As conclusões deste estudo destacaram o sucesso da aplicação de técnicas de machine learning na análise de espectroscopia molecular, evidenciando a capacidade dessas abordagens em lidar com dados espectroscópicos e extrair informações relevantes .

Além disso, a obtenção das larguras das linhas espectrais com estatísticas básicas demonstrou a viabilidade e a eficácia da abordagem proposta, fornecendo insights importantes sobre as características dos espectros moleculares analisados .

A identificação da variável considerada mais importante na análise das linhas também forneceu informações valiosas para pesquisas e aplicações práticas.

Por fim, o erro médio percentual de 9,66% obtido na validação cruzada em 5 folds e de 9,36% em teste reforçou a robustez e a confiabilidade do modelo desenvolvido, demonstrando sua capacidade de estimativa das larguras das linhas espectrais com precisão aceitável.

Em síntese, as conclusões deste estudo ressaltaram o potencial e a relevância da aplicação de machine learning na análise de espectros moleculares, abrindo novas perspectivas para a utilização dessas técnicas em pesquisas e aplicações práticas no campo da espectroscopia molecular.