

Automatização do Aprendizado de Máquina com PyCaret para Tratamento de Dados Químicos

*Márcia Helena Cassago Nascimento¹ (PQ), Lara Fábila F. Gerhardt¹ (IC), Felipe R. Nascimento¹ (IC), Madson P. Zanoni¹ (PG), Lais S. Gomes¹ (PG), Paulo Roberto Filgueiras¹ (PQ)

*marcia.cassago@gmail.com

¹Departamento de Química, Universidade Federal do Espírito Santo, Av. Fernando Ferrari, 514 - Goiabeiras, Vitória - ES, 29075-910, Brasil,

Palavras Chave: RMN ¹H, NIR, quimiometria, ensemble

Introdução

Os avanços da inteligência artificial e da capacidade computacional têm viabilizado a geração de grandes volumes de dados químicos por técnicas analíticas, assim como o uso de aprendizado de máquina no tratamento desses dados. No entanto, a aplicação desses métodos geralmente demanda tempo e programação extensiva na otimização e seleção de modelos capazes de extrair informações relevantes e identificar padrões nos dados. Nesse cenário, o PyCaret, uma biblioteca de código aberto, desenvolvida em Python, surge como uma alternativa promissora por automatizar fluxos completos de modelagem com poucas linhas de código, incluindo comparação, otimização e combinação de algoritmos. Entretanto, na diversidade de modelos envolvidos, informações relacionadas às variáveis mais relevantes podem ser diluídas ou perdidas na etapa de combinação dos modelos. Este trabalho propõe o uso de PyCaret na automação de modelos aplicados aos problemas químicos, com foco na construção de *ensembles* interpretáveis. O objetivo é explorar o potencial de automação oferecido pela biblioteca, criar *ensembles* por meio da combinação de diferentes abordagens e extrair, a partir deles, de forma unificada, as variáveis-chave associadas aos problemas químicos analisados.

Resultados e Discussão

Neste trabalho, utilizou-se quatro conjuntos de dados, descritos na Tabela 1.

Tabela 1. Conjuntos de dados utilizados neste estudo

Conjunto	Técnica analítica	Problema abordado	Dimensão (linhas variáveis)	Fonte
01	NIR	Distinção entre azeite extravirgem e outros óleos vegetais ou compostos	40 125	Sintéticos adaptados de Borghi et al. (2020)
02	NIR	Distinção de petróleo e óleo de reuso	150 125	Sintéticos adaptados de Santos et al.(2021)
03	ICP-MS	Distinção entre cafés puros e adulterados	68 22	Reproduzidos de Costa et al. (2024)
04	RMH ¹ H	Distinção de petróleos pelo API	115 7242	Projeto Quimiometria-Labpetro-UFES

NIR: espectroscopia na região do infravermelho próximo; ICP-MS: espectrometria de massas com plasma indutivamente acoplado; RMN

¹H: Ressonância magnética nuclear de hidrogênio

Os dados foram pré-processados com correção de linha de base, derivada, normalização, autoescalamento e/ou centralização na média. As análises foram conduzidas em Python (versão 3.10.18), utilizando a interface *Jupyter Notebook* e bibliotecas auxiliares. O módulo de classificação da biblioteca PyCaret (versão 3.2.0) foi empregado para automatizar, otimizar e comparar modelos pela validação cruzada, utilizando as métricas de sensibilidade, exatidão e coeficiente de correlação de Matthews (MCC). *Ensembles* formados pelos cinco modelos com melhor desempenho apresentaram sensibilidade entre 80 e 98,6%, exatidão entre 90 e 100% e MCC de 0,800 a 0,972. As variáveis mais relevantes foram identificadas por meio da combinação dos classificadores baseados em árvores com maior desempenho, utilizando o índice de Gini como critério. A ponderação pela performance de cada modelo assegurou a interpretabilidade química dos resultados.

Conclusões

A abordagem proposta neste estudo pela estratégia de *ensembles* com Pycaret promove a otimização de modelos de aprendizado de máquina de alta performance sem que se perca a interpretabilidade dos resultados, o que é essencial para resolução de problemas químicos.

Agradecimentos

LABPETRO, UFES, FEST, FAPES, CAPES, CNPq

BORGHI, Flávia T. et al. *Microchemical J.*, v. 159, p. 105544-105555, 2020.

SANTOS, Francine D. et al. *Fuel*, v.283, p.118854, 2021.

COSTA, Tainá de O. et al. *Food Chemistry*, v.442, p.138268, 2024.