QUIMIOMETRIA COM *SOFTWARE NO-CODE*

Felipe Rodrigues Nascimento[1], Lara Fábia Ferreira Gerhardt[1], Márcia Helena Cassago Nascimento[1], Matthews Silva Martins[1], Valério Garrone Barauna, Paulo Roberto Filgueiras[1]

[1]Universidade Federal do Espírito Santo, Av. Fernando Ferrari, 514 - Goiabeiras, Vitória - ES, 29075-910, Brasil, felipe.nascimento.09@ufes.edu.br, paulo.filgueiras@ufes.br

Resumo
A quimiometria surgiu nos anos 70, beneficiando-se do avanço computacional que crescia em paralelo à química analítica e permitia a introdução de computadores nos centros de química. Após anos de aperfeiçoamento tecnológico e computacional, a vertente *no-code* surgiu a fim de contrapor-se à necessidade de extensas linhas de código que geralmente estavam presentes no trabalhos. Este trabalho centraliza-se na aplicação *no-code,* dentro da quimiometria, usando o *software Orange Data Mining,* o qual além de utilizar código aberto tem uma interface *user-friendly*. Para demonstrar a praticidade desta evolução *no-code* na quimiometria, foram treinados modelos de classificação. Com o auxilo da validação cruzada, os modelos SVM e *Gradient Boosting* foram escolhidos, pois obtiveram as melhores métricas de performance com valores de F-Scores de 1,0 e de 0,95 para ambos, nos conjuntos de treinamento e teste, respectivamente. Dessa forma, conhecendo as ferramentas adequadas é possível realizar quimiometria a partir de um software de código aberto e acessível pelo formato *no-code*, o que pode ajudar a desmistificar e disseminar a quimiometria.

Palavras-chave: Quimiometria. *No-code.* Classificação. Adulteração de alimentos
Área do Conhecimento: Ciências exatas e da terra

Introdução

A evolução da ciência é impulsionada pela necessidade de explorar e compreender o que existe nas lacunas do conhecimento. Neste contexto, a química analítica surgiu sob o viés de analisar e quantificar substâncias presentes no cotidiano (Andrade, Alvim, 2018). Nos anos 70, em consequência da evolução gradativa da área, juntamente com o advento e popularização da computação, a qual permitiu que programação computacional fosse aplicada, surgiu a quimiometria, um campo de estudo que visava a aplicação de métodos estatísticos e matemáticos, juntamente com o aprendizado de máquina (do inglês *machine learning)*, para efetuar-se a análise de dados químicos. No Brasil, o marco inicial da quimiometria pode ser associado à vinda do Professor Dr. Bruce Kowalski para ministrar um curso intensivo de quimiometria em novembro de 1980 (Veras et al., 2022). Ao longo de mais de 4 décadas a quimiometria estabeleceu-se como uma das áreas mais relevantes da química.

É muito comum ver trabalhos quimiométricos que utilizam softwares como Matlab®, *RStudios* e *Octave* para a realização de suas análises. Tais softwares apresentam eficácia e possuem uma gama de bibliotecas e *toolboxes* capazes de tornar uma análise de dados mais prática e dinâmica. Todavia, a necessidade de um alto conhecimento em programação computacional, acaba por afastar o interesse de indivíduos com mais dificuldade em trabalhos computacionais. A fim de contornar esse fator, nos últimos anos diversos autores como Cahil e colaboradores (2024), Bravenec e Wark (2023) e Ellick e colaboradores (2022) têm realizado estudos quimiométricos com *Python,* que é uma linguagem de programação mais simples que outras. Além disso, também surgiram algumas interfaces com as propostas mais centradas na interação do usuário com o *software*. Como exemplo pode-se citar o aplicativo GAMMA-GUI, que se trata de uma interface gráfica para Matlab®, a qual visa facilitar a interação entre novos usuários e o *software*, trazendo consigo uma interface intuitiva e visualmente mais amigável que promete facilitar o planejamento de experimentos e outros métodos de análise multivariada de dados (Galvan; Bona, 2024). A interface do GAMMA-GUI é um passo relevante para a se contornar as longas linhas de código presentes em trabalhos de quimiometria. Entretanto, ainda se utiliza linhas de comando juntamente com ferramentas sem código, ou seja, ainda exige um conhecimento em linguagem de programação que não é abrangido por todos.

Sob esse viés surgiram vertentes computacionais que se baseiam nas aplicações *low-code* e *no-code* dentro da análise de dados. Os termos vêm de uma nova vertente da área computacional que amplia a cadeia de possibilidades, estendendo-se a profissionais de outras áreas que não possuem um conhecimento técnico avançado. O termo *low-code* refere-se à tática de construção de *softwares* que utilizem uma menor quantidade de linhas de código e permitam que a plataforma seja mais intuitiva e didática (RAMOS ALVES, GOMES SOARES ACALÁ, 2022). O termo *no-code* possui pontos em comum ao termo anterior com o diferencial de sua metodologia remover por completo a aplicação de linhas de código, substituindo-as por interfaces gráficas e intuitivas que facilitam o processo. Plataformas como *LUMIOS, Orange Data Mining* e *CODA,* vem sendo citadas por diversos autores (VIEIRA, 2024; KIM, 2024; MULLIE, 2024) como exemplos de plataformas que se adequam dentro dessa vertente computacional.

Este artigo centra sua proposta na aplicação de um software dessa abordagem dentro da quimiometria, a fim de apresentar uma alternativa de emprego de métodos quimiométricos a dados espectroscópicos, por meio da reprodução de estudos desenvolvidos e publicados pelo nosso grupo de pesquisa, a partir de uma estratégia *no-code* e de código aberto.

Metodologia

Para este trabalho, o software *Orange Data Minig* foi selecionado por possuir uma interface mais intuitiva e por possuir um pacote de espectroscopia que permitiria um melhor tratamento dos dados. O *software* pode ser baixado gratuitamente e está disponível, até o momento da escrita deste artigo, na versão 3.37.0 (a versão utilizada neste estudo foi a 3.36.0). O *Orange* trata-se de uma plataforma que possui uma interface *no-code* e de código aberto, que foi construída a partir do *python* e a construção do trabalho é feita por meio de *widgets* que dão corpo a um *workflow* (“fluxograma”) que permite ao usuário construir algo de maneira intuitiva.

O conjunto de dados utilizado foi um conjunto contendo dados de espectroscopia na região do infravermelho médio com acessório de reflectância total atenuada (ATR-FTIR) de sessenta e seis espectros adquiridos a partir de amostras de três marcas de *whey protein,* as quais foram divididas em seis misturas, contaminadas com diferentes quantidades de farinha de trigo, variando de 0 a 100% (em termos de 10%). O desenho experimental e maiores detalhes a respeito do preparo das amostras e aquisição espectral podem ser encontrados em (MARTINS et al.,2023).

Para classificação das amostras utilizou-se como parâmetro de atribuição de classes a concentração de adulterante (farinha de trigo): para concentração de adulterante ≤ 20% seriam consideradas classe 1 e para concentração de adulterante > 20% seriam consideradas Classe 2, baseando-se no limite permitido pela legislação brasileira para suplementos (até 20%) (BRASIL, 2020).

Os espectros foram cortados na faixa de 1800 a 800 cm^{-1} seguido de uma correção de linha de base pelo método *rubber band* e média das triplicatas, para uma melhor visualização dos espectros e uma diminuição no número de variáveis de saída. Além disso, os espectros foram pré-processados pelo método de padronização Variação Normal Padrão (do inglês, *standard normal variate – SNV).*

A seguir, foi realizada a construção do modelo iniciando-se pela divisão dos dados em subconjuntos de treinamento (70%) e teste externo (30%). Diferentes métodos de classificação foram avaliados, utilizando-se a técnica de validação cruzada (do inglês *cross-validation*) seguida da escolha pelo(s) modelo(s) de melhores métricas de performance nas análises de validação cruzada e de teste externo.

Resultados

A Tabela 1 traz uma relação entre determinadas regiões de espectros de infravermelho médio e os respectivos compostos associados a essas regiões, também chamadas de bandas.

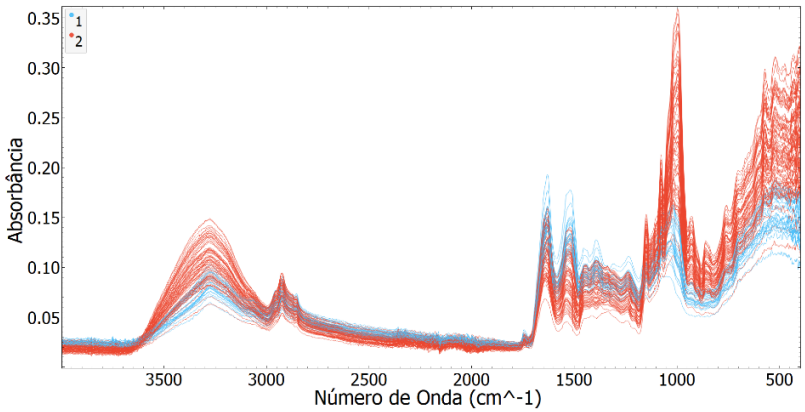
Tabela 1- Atribuição espectral de bandas características

Banda	Número de onda (cm^{-1})	Referência
Amida I (vC =O, vC-N)	~1640	(X. Wang et al., 2018)
Amida II (δN-H, vC-N)	~1550	(X. Wang et al., 2018
Vibrações de polissacarídeos	~1200-900	(Moros et al., 2006)
Estiramento de –OH de carbohidratos	~ 1080	(Moros et al., 2006)

v: vibração de estiramento de ligação química; δ: vibração de deformação angular de ligação química
Fonte: Autoria Própria

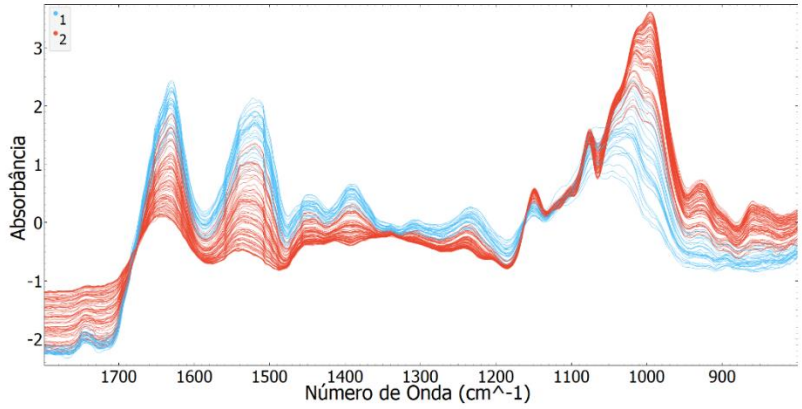
Ao se observar o espectro bruto (Figura 1), onde as amostras da classe 1 (≤ 20%) são as linhas cor azul e classe 2 (> 20%) são as linhas vermelhas, é possível notar que há um ruído de fundo ao longo dos espectros. A fim de remover o impacto desse ruído, foi utilizado o método *Rubber Band's* para correção de linha de base. Após o recorte da região de interesse, abrangendo bandas de absorção de proteínas e de carboidratos (1800 – 800 cm^{-1}) e correção da linhas de base, é possível observar na Figura 2, que à medida que a quantidade de farinha de trigo vai aumentando nas amostras é possível notar o aumento da banda de espectro de carboidratos (1080 cm^{-1}) junto ao decréscimo da banda relacionada a proteína (1640-1550 cm^{-1}), isso demonstra uma relação direta entre a alteração composicional, provocada pela adulteração, e a informação espectral.

Figura 1- Espectro Bruto



Fonte: Autoria própria

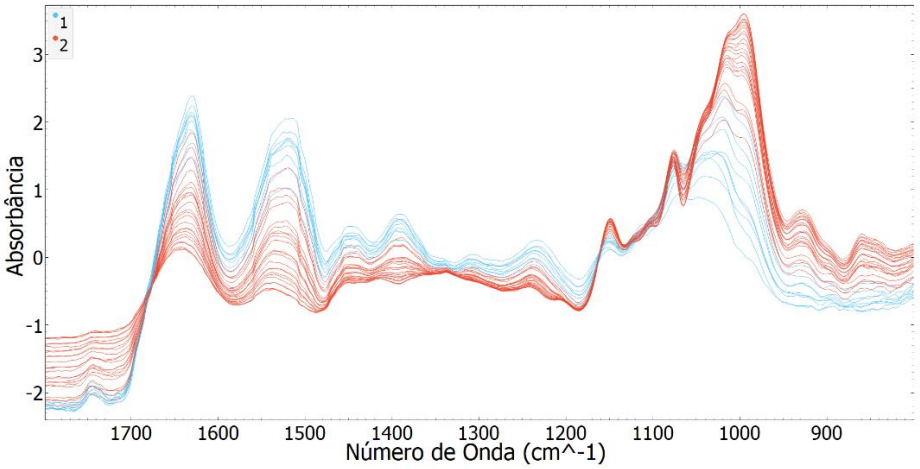
Figura 2 – Espectro pré-processado



Fonte: Autoria Própria

Além do pré-processamento e análise dos espectros, realizou-se a média das triplicatas para promover uma menor quantidade de saídas para as próximas etapas, uma vez que acima de 200 variáveis de saída o *software* alerta para uma queda de eficiência. O espectro médio também auxilia para caso haja uma amostra com algum valor muito discrepante em alguma análise, pois há uma normalização das linhas espectrais.

Figura 3- Espectro das médias



Fonte: Autoria Própria

Após os pré-processamentos, realizou-se a separação dos dados, utilizando um *widget data sampler,* que os dividiu, aleatoriamente, em duas porções: uma de 70% voltada para treino e outra de 30% voltada para teste, respeitando a porcentagem dentro de cada classe. O método de validação cruzada utilizando o *widget test&score*, foi eficiente para demostrar quais modelos de classificação entregariam os melhores resultados. As métricas resultantes de cada modelo são demonstradas na Tabela 2, na qual é possível verificar que o *software* entregrou modelos eficientes para os dados utilizados.

Tabela 2- *Cross-Validation*

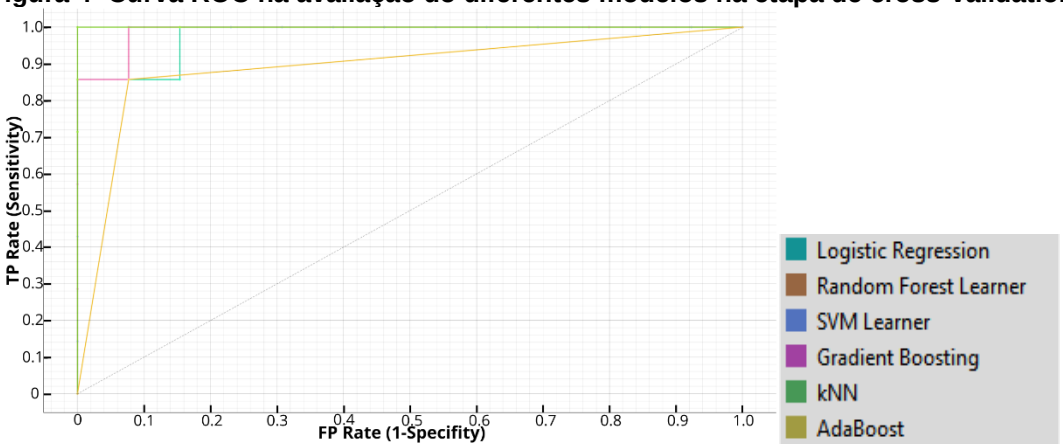
Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.978	0.900	0.900	0.900	0.900	0.780
Random Forest Learner	0.989	0.950	0.949	0.954	0.950	0.892
SVM Learner	0.989	0.850	0.852	0.858	0.850	0.685
Gradient Boosting	0.989	0.900	0.900	0.900	0.900	0.780
kNN	1.000	0.950	0.949	0.954	0.950	0.892
AdaBoost	0.890	0.900	0.900	0.900	0.900	0.780

Legenda: AUC: Área abaixo da curva de curva ROC; CA: Exatidão de classificação (do inglês, Classification accuracy); F1: F1 Score, uma média harmônica ponderada de precisão e sensibilidade; Prec: Precisão da classificação; Recall: Sensibilidade; MCC: Coeficiente de correlação de Matthews.
Fonte: Autoria Própria

Todos os modelos apresentaram resultados satisfatórios, mas para realizar o treinamento e a validação do modelo pelo teste externo era necessário selecionar os melhores dentre estes, a fim de não sobrecarregar a plataforma. Utilizou-se a curva ROC como critério de escolha (Figura 4), pois com essa análise foi possível verificar graficamente a qualidade dos modelos. A curva ROC pode ser interpretada como “a confiabilidade do modelo”, uma vez que seu gráfico possui uma linha central

pontilhada que indica o valor de 0,5 ou 50%, associada a resultados aleatórios. Portanto, os modelos em que a curva ROC estiver mais próxima dessa linha central são aqueles que possuem uma maior aleatoriedade na classificação das amostras, ou seja, um modelo dado à sorte, o que não condiz com uma classificação dada pela informação química existente nos dados.

Figura 4- Curva ROC na avaliação do diferentes modelos na etapa de cross-validation



Fonte: Autoria própria

Nota-se na curva ROC que o método *k-Nearest Neighbors (k-NN)* possui o melhor resultado dentre todos, com AUC de 1,00, seguido de *Random Forest*, *Support vector machine* (SVM) e *Gradient Boosting*, cujas linhas se sobrepõem no gráfico, ambos com 0,989 de AUC. Os de menor desempenho foram *Logistic Regression* (AUC de 0,978) e *AdaBoost* (AUC de 0,890). Todavia, as métricas de treinamento do modelo KNN obtiveram um valor inferior ao esperado, substituindo-o por: SVM e *Gradient Boosting,* os quais obtiveram os melhores resultados dentro dos parâmetros de treino e teste. A realização do treinamento entregou modelos 100% precisos, todos os parâmetros obtiveram valores de resposta igual a 1 (Tabela 3), demonstrando modelos eficazes. O teste externo obteve valores de resposta inferiores, o que já se esperava, e suas métricas continuam excelentes (Tabela 4).

Tabela 3-Métricas de performance modelos classificação - conjunto Treinamento

Modelo	AUC	CA	F1	Prec	Recall	MCC
SVM	1,00	1,00	1,00	1,00	1,00	1,00
Gradient Boosting	1,00	1,00	1,00	1,00	1,00	1,00

Legenda: AUC: Área abaixo da curva de curva ROC; CA: Exatidão de classificação (do inglês, Classification accuracy); F1: F1 Score, uma média harmônica ponderada de precisão e sensibilidade; Prec: Precisão da classificação; Recall: Sensibilidade; MCC: Coeficiente de correlação de Matthews.

Fonte: Autoria Própria

Tabela 4-Métricas de performance de modelos de classificação - conjunto Teste externo

Modelo	AUC	CA	F1	Prec	Recall	MCC
SVM	1,00	0,923	0,952	1,00	0,909	0,778
Gadient Boosting	1,00	0,923	0,952	1,00	0,909	0,778

Legenda: AUC: Área abaixo da curva de curva ROC; CA: Exatidão de classificação (do inglês, Classification accuracy); F1: F1 Score, uma média harmônica ponderada de precisão e sensibilidade; Prec: Precisão da classificação; Recall: Sensibilidade; MCC: Coeficiente de correlação de Matthews.

Fonte: Autoria Própria

Conclusão

O *software Orange Data Mining* se mostrou eficiente ao se trabalhar com modelos de classificação. A plataforma possui certas limitações que podem ser associadas ao *no-code,* o qual não permite que sejam feitas alterações internas nos *widgets* nem em sua interface, isso acaba acarretando análises mais superficiais. Outro problema interno da plataforma é a queda de eficiência quando os

dados aplicados são em grande volume, como geralmente são os dados de espectroscopia. Isso pode acabar gerando diferenças e/ou erros ao longo do processo de análise. O *software* também traz consigo alguns travamentos a depender da capacidade do computador em que se está trabalhando. Porém, apesar desses pequenos casos isolados, o *software* é uma ótima alternativa, pois torna mais acessível a aplicação de quimiometria mantendo-se as performances dos modelos.

Referências

BRASIL. Agência Nacional de Vigilância Sanitária (ANVISA). **Resolução da Diretoria Colegiada - RDC nº 429, de 8 de outubro de 2020**. Diário Oficial da União, Brasília, DF, 9 out. 2020.

ANDRADE, J. C. de; ALVIM, T. R. **Química analítica básica: uma visão histórica da análise qualitativa clássica**. Revista Chemkeys, n. 9, p. 1–8, 2018.

BRAVENEC, A. D.; WARD, K. D. **Interactive python notebooks for physical chemistry**. Journal of Chemical Education, v. 100, n. 2, p. 933–940, 2023.

CAHILL, S. T. et al. **Assignment of regioisomers using infrared spectroscopy: A python coding exercise in data processing and machine learning**. Journal of Chemical Education, v. 101, n. 7, p. 2925–2932, 2024. DOI: doi.org/10.1021/acs.jchemed.4c00295

ELLICK, S. et al. **Chemometric approaches to resolving base oil mixtures**. Rapid Communications in Mass Spectrometry: RCM, v. 36, n. 1, 2022. DOI: 10.1002/rcm.9214

GALVAN, D.; BONA, E. **Aplicativo GAMMA-GUI: uma interface gráfica amigável para planejamento de experimentos no MATLAB**. Química Nova, 2024.

KIM, S.-Y.; JEON, I.; KANG, S.-J. **Integrating data science and machine learning to chemistry education: Predicting classification and boiling point of compounds**. Journal of Chemical Education, v. 101, n. 4, p. 1771–1776, 2024. DOI: 10.1021/acs.jchemed.3c01040

MARTINS, MATTHEWS S., et al. **Detection and Quantification Using ATR-FTIR Spectroscopy of Whey Protein Concentrate Adulteration with Wheat Flour.** *LWT*, vol. 172, Dec. 2022, p. 114161, https://doi.org/10.1016/j.lwt.2022.114161.

MULLIE, L. et al. **CODA: an open-source platform for federated analysis and machine learning on distributed healthcare data**. Journal of the American Medical Informatics Association: JAMIA, v. 31, n. 3, p. 651–665, 2024. DOI : https://doi.org/10.1093/jamia/ocad235

VERAS, G. et al. **Perfil cientométrico da quimiometria no Brasil**. Química Nova, 2022.

VIEIRA, R.; DE SOUSA, K. A.; CASTRO-GAMBOA, I. **LUMIOS – Label using machine in organic samples – A software for dereplication, molecular docking, and combined machine and deep learning**. Expert Systems with Applications, v. 248, n. 123447, 2024.

Agradecimentos

O presente trabalho foi realizado com apoio da Fundação de Amparo à Pesquisa e Inovação do Espírito Santo (FAPES) [processos #032/2023, #691/2022, #1036/2022, #442/2021]; a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) [processo nº 88887.487966/2020-00]; e Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [processo #409700/2022-3, #310349/2021-4].