

Uso de aprendizado de máquina com amostras sintéticas para identificação de variáveis preditoras na caracterização química de *Coffea canephora*

Evandro Messias¹(PG)*, Cleidiana V. Guimarães³(PG), José M. R. Luz³(PQ), Elisson H. de Paulo⁴(PG), Marcos V. V. Lyrio¹(PG), Márcia H. C. Nascimento¹(PQ), Victor G. K. Cardoso⁵(PQ), Marco F. Ferrão⁴(PQ), Lucas L. Pereira² (PQ), Paulo R. Filgueiras¹(PQ).

evandroqt@gmail.com

1. Universidade Federal do Espírito Santo, Av. Fernando Ferrari, Vitória, ES, Brasil.

2. Instituto Federal do Espírito Santo, Av. Elizabeth Minete Perim, Venda Nova do Imigrante, ES, Brasil.

3. Universidade Federal de Viçosa, Av. P.H. Rolfs, Viçosa, MG, Brasil.

4. Universidade Federal do Rio Grande do Sul, Av. Bento Gonçalves, Porto Alegre, RS, Brasil.

5. Universidade de Copenhagen, Rolighedsvej 26, Frederiksberg, Dinamarca.

Palavras Chave: Café conilon, FTIR, Random Forest, Amostras sintéticas e Qualidade sensorial.

Introdução

Métodos de aprendizado de máquina como o Random Forest (RF), têm sido destacados por sua alta eficiência na identificação de variáveis preditoras relevantes em análises da composição química do café¹. Em estudos quimiométricos, a distribuição dos dados representa um desafio significativo, especialmente em conjuntos contendo número reduzido de amostras ou desbalanceamento entre as classes analisadas, problemas frequentes nos estudos de qualidade de café devido à complexa variabilidade química influenciada por fatores como o processamento pós-colheita. O método de geração de amostras sintéticas emerge como uma solução promissora para equilibrar a quantidade de informação, preservando a integridade das características espectrais das amostras originais e melhorando a generalização dos modelos preditivos². Neste trabalho, combinamos o método de geração de amostras sintéticas com modelos RF para identificar variáveis preditoras importantes na caracterização química dos diferentes processos de fermentação de *C. canephora*, utilizando dados obtidos por espectroscopia na região do infravermelho médio (FTIR).

Resultados e Discussão

A geração de amostras sintéticas foi realizada com base na análise de componentes principais dos dados originais. Utilizando um conjunto de dados composto por aproximadamente 20% de amostras originais e 80% de amostras sintéticas, foi possível desenvolver o modelo RF com desempenho satisfatório na identificação das variáveis preditoras, alcançando sensibilidade acima de 95% para todas as classes e exatidão em torno de 97%. As bandas em 2940 e 2840 cm^{-1} selecionadas pelo modelo, correspondem a vibrações de estiramento das ligações C-H que podem estar associadas à presença de moléculas de lipídios. Regiões entre 900 e 650 cm^{-1} , relacionadas a vibrações de deformação de ligações N-H e C-N, indicam a presença de amins e amidas, sugerindo aminoácidos e proteínas nas amostras analisadas. Essas observações são significativas, uma vez que a fermentação induzida pode modular a degradação e síntese de compostos nitrogenados, como proteínas e aminoácidos durante o processo fermentativo do café¹.

Conclusões

A integração de geração de amostras sintéticas com modelo de aprendizado de máquina como o Random Forest constitui uma abordagem poderosa para identificar variáveis químicas relevantes em estudos de qualidade de cafés por espectroscopia FTIR, permitindo a detecção de padrões analíticos importantes para a avaliação da qualidade sensorial do *C. canephora*, tais como ácidos orgânicos, açúcares, proteínas e lipídeos.

Agradecimentos

[UFES, UFRGS, IFES, FAPES, CAPES, CNPq]

¹Messias, E.; Guimarães, C. V.; da Luz, J. M. R.; de Paulo, E. H.; Oliveira, E. C. S.; Nascimento, M. H. C.; Ferrão, M. F.; Guarçoni, R. C.; Filgueiras, P. R.; Pereira, L. L. Food Chem. 456, 139879, 2025.

²Folli, G. S.; Nascimento, M. H. C.; Lopes, B. P. O.; Romão, W.; Filgueiras, P. R. Chemolab. J. 251, 2024.