# CLASSIFICAÇÃO DE AMOSTRAS DE ÁGUA E EFLUENTES PELA PRESENÇA DE COMPOSTOS FENÓLICOS USANDO QUIMIOMETRIA ASSOCIADA ÀS ESPECTROSCOPIAS DE FLUORESCÊNCIA E ABSORÇÃO NO UV-VIS

**Ellisson Henrique de Paulo**[(1)]
Mestre em Química pela Universidade Federal do Espírito Santo (UFES). Atualmente doutorando em Química pela Universidade Federal do Rio Grande do Sul (UFRGS) e pesquisador da Tommasi Ambiental e do Tommasi Laboratório.
**Otávio Arruda Heringer**[(2)]
Mestre em Ciências Farmacêuticas pela Universidade Vila Velha (UVV). Atualmente diretor operacional da Tommasi Ambiental e pesquisador do Tommasi Laboratório.
**Paulo Roberto Filgueiras**[(3)]
Doutor em Ciências pela Universidade Estadual de Campinas (Unicamp). Atualmente professor de Química na Universidade Federal do Espírito Santo (UFES).
**Marco Flôres Ferrão**[(4)]
Doutor em Ciências pela Universidade Estadual de Campinas (Unicamp). Atualmente professor de Química na Universidade Federal do Rio Grande do Sul (UFRGS).

**Endereço**[(1)]**:** Rua Arara Azul, 187, Galpão branco - Novo Horizonte - Serra - Espírito Santo - CEP: 29163-306 - Brasil - Tel: +55 (27) 3340-8200 - e-mail: ellisson.hp@gmail.com, ellisson@tommasi.com.br.

## RESUMO

Os compostos fenólicos são monitorados em águas naturais e efluentes devido aos seus potenciais efeitos nocivos. A espectroscopia UV-vis e a fluorescência EEM são técnicas que podem ser empregadas para essa finalidade, sendo a última mais sensível e seletiva. A aplicação de métodos quimiométricos na análise qualitativa dos dados espectrais resulta em modelos de classificação. Neste estudo, modelos PLS-DA e SVM-DA acoplados a geração de amostras sintéticas foram utilizados para classificar amostras de água e efluentes quanto à presença de compostos fenólicos a partir dos dados espectrais. Os resultados indicam que a aplicação do ADASYN favorece a melhoria das métricas dos modelos.

**PALAVRAS-CHAVE:** Fluorescência EEM, fenóis, ADASYN.

## INTRODUÇÃO

Os compostos fenólicos são um grupo de compostos orgânicos que podem ser encontrados em águas naturais e efluentes industriais. Essas substâncias são caracterizadas pela presença de um ou mais grupos hidroxila ($OH^-$) ligados a um anel aromático, conferindo-lhes propriedades únicas e uma ampla gama de aplicações industriais e ambientais (Anku et al., 2017; Ramos et al., 2024).

A presença de compostos fenólicos na água e efluentes pode ter várias fontes, incluindo atividades industriais, agrícolas e domésticas (Anku et al., 2017; Ramos et al., 2024). Esses compostos são frequentemente liberados no ambiente como subprodutos de processos industriais, como fabricação de papel, produção de produtos químicos e tratamento de águas residuais. Além disso, pesticidas, herbicidas e produtos de higiene pessoal também podem contribuir para a presença de compostos fenólicos na água (Anku et al., 2017; Ramos et al., 2024).

A detecção e quantificação de fenóis totais na água e efluentes são de grande importância devido aos seus potenciais efeitos nocivos à saúde humana e ao meio ambiente. Muitos compostos fenólicos são tóxicos e podem causar danos aos organismos aquáticos, além de representarem um risco para a saúde humana se presentes em concentrações elevadas na água potável (Anku et al., 2017; Ramos et al., 2024).

A determinação de fenóis totais é comumente realizada por meio de métodos analíticos, sendo a espectroscopia no ultravioleta-visível (UV-vis) uma das técnicas mais utilizadas (Lipps et al., 2023; Thomas & Burgess, 2017). Nesse método, os fenóis totais na amostra reagem com um reagente específico para formar um complexo colorido, cuja

intensidade de absorção de luz UV está diretamente relacionada à concentração de fenóis presentes. A leitura da absorbância da solução resultante em um comprimento de onda permite quantificar os fenóis totais na amostra (Baird et al., 1990; Thomas & Burgess, 2017).

Além da espectroscopia UV-vis, outras técnicas analíticas, como a espectroscopia de emissão de fluorescência, também podem ser empregadas para a determinação de fenóis totais com alta sensibilidade e seletividade (Lakowicz, 2006; Xiao et al., 2014). Os compostos fenólicos podem exibir fluorescência em diferentes comprimentos de onda na faixa do UV-vis (Xiao et al., 2014). Em geral, os compostos fenólicos absorvem luz em comprimentos de onda UV-vis devido aos grupos cromóforos presentes em sua estrutura molecular, como o anel aromático conjugado e os grupos hidroxila. Quando excitados por luz UV-vis, esses compostos podem emitir fluorescência geralmente na faixa de 275 a 560 nm (Nørgaard, 1995).

Neste sentido, a fluorescência em matriz de excitação-emissão (EEM, do inglês *excitation-emission matrix*) permite a obtenção das intensidades de fluorescência de uma amostra usando uma série de comprimentos de onda de emissão em função de uma série de comprimentos de onda de excitação. O que permite a coleta de muita informação química sobre os fenóis em um único espectro (Tchaikovskaya et al., 2002; Xiao et al., 2014).

De maneira alternativa, métodos matemáticos podem ser empregados na avaliação da presença de compostos fenólicos via fluorescência EEM e absorção no UV-vis (Nørgaard, 1995). A quimiometria aplica técnicas de estatística multivariada aliada a recursos computacionais para tratar dados de origem química (Ferreira, 2015). Neste caso, os espectros de emissão de fluorescência ou absorção no UV-vis podem ser processados por metodologias de reconhecimento de padrões, como a análise discriminante por mínimos quadrados parciais (PLS-DA, do inglês, *partial least squares discriminant analysis*) e a máquina de vetores de suporte com análise discriminante (SVM-DA, do inglês *support vector machines discriminant analysis*) para classificar amostras quanto à presença de fenóis na água (Ballabio & Consonni, 2013; Filisbino et al., 2020; Nørgaard, 1995).

Modelos de classificação associados à fluorescência EEM foram usados por Paulo et al (2024) para classificar amostras de água pela presença bacteriana (Paulo et al., 2024). Os modelos PLS-DA, SVM e floresta aleatória (RF, do inglês *random forest*) se mostraram como as melhores estratégias quando acopladas ao desdobramento multivias (do inglês *unfold-multiway*). Neste trabalho, foram aplicados os métodos PLS-DA e SVM-DA a dados de absorção no UV-vis e fluorescência EEM para a classificação de amostras de água e efluentes usando a presença de compostos fenólicos como critério.

## OBJETIVOS

Classificar amostras de água e efluentes pela presença de fenóis totais usando os métodos PLS-DA e SVM-DA aplicados a absorção no UV-vis e fluorescência EEM.

## MATERIAIS E MÉTODOS

As amostras utilizadas neste estudo foram coletadas nos estados de Minas Gerais e Espírito Santo seguindo protocolos de referência (Brandão et al., 2011). O conjunto de amostras totalizou 148 com exemplos de água doce, água salina, efluentes sanitários e industriais.

Os fenóis totais foram quantificados usando o método 5530 A da *American Public Health Association*, *American Water Works Association, and Water Environment Federation* via reação com aminoantipirina, extração com clorofórmio e detecção por espectrofotometria em 480-500 nm (Lipps et al., 2023). Os espectros de fluorescência EEM e absorção no UV-vis foram adquiridos simultaneamente num espectrômetro Horiba Aqualog na faixa de excitação/absorção em 200 a 800 nm e emissão em 250 a 800 nm (ASTM International, 2022). Os espectros de fluorescência EEM foram corrigidos pelo efeito de filtro interno, espalhamento Raman e Rayleigh antes do processamento de dados (Bahram et al., 2006).

Para a construção dos modelos, as amostras que apresentaram teor de fenóis abaixo do limite de quantificação (0,003 $mg \cdot L^{-1}$) foram colocadas como classe 1, totalizando 48 amostras, e as amostras que quantificaram fenóis foram colocadas como classe 2, totalizando 100 amostras. Dessa maneira, como a quantidade de amostras nas classes 1 e 2 não é equivalente, o algoritmo amostragem sintética adaptativa (ADASYN, do

inglês *adaptive synthetic sampling*) foi testado para o balanceamento de classes com a geração de amostras sintéticas (He et al., 2008).

As amostras foram divididas em conjunto treinamento (tre) (70%) e teste (tes) (30%) pelo algoritmo Kennard-Stone respeitando a quantidade das classes (Kennard & Stone, 1969). Para o desdobramento dos espectros de fluorescência EEM, o *unfold-multiway* e a análise de fatores paralelos (PARAFAC, do inglês *parallel factor analysis*) foram testados antes da construção dos modelos de classificação (Bro, 1997; Murphy et al., 2013; Olivieri et al., 2015). Os espectros UV-vis e fluorescência foram pré-tratados com normalização, variação normal padrão e autoescalamento (Barnes et al., 1989; Rinnan et al., 2009). Foram construídos modelos PLS-DA e SVM-DA para a classificação.

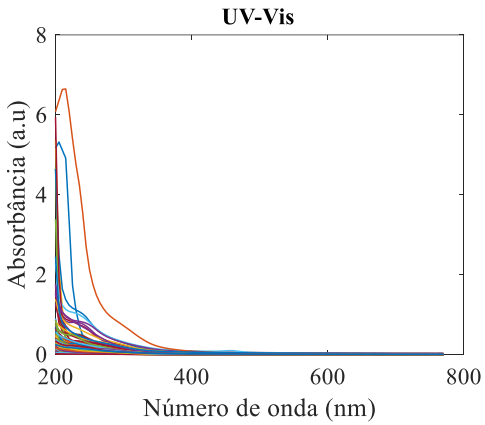
Os modelos de classificação usando UV-vis foram nomeados PLS-DA e SVM-DA sem amostragem sintética e suas versões com amostragem são ADASYN-PLS-DA e ADASYN-SVM-DA. Para os espectros de fluorescência EEM com o *unfold-multiway*, os modelos de classificação são UPLS-DA, USVM-DA, ADASYN-UPLS-DA e ADASYN-USVM-DA. Já espectros de fluorescência EEM usando PARAFAC, os modelos são PARAFAC-PLS-DA, PARAFAC-SVM-DA, ADASYN- PARAFAC-PLS-DA e ADASYN-PARAFAC-SVM-DA.

Os modelos foram avaliados pelos conjuntos de treinamento e teste, e as métricas de desempenho incluem verdadeiro positivo (TP), verdadeiro negativo (TN), falso positivo (FP), falso negativo (FN), sensibilidade (SEN), especificidade (SPE), precisão (PRE) e acurácia (ACC) (Brereton, 2021). Toda modelagem quimiométrica foi realizada no software MATLAB® versão 2013a.

## RESULTADOS E DISCUSSÃO

A Figura 1 apresenta os espectros de absorção no UV-vis das amostras usadas neste estudo. Poucas amostras demonstraram absorção elevada se comparadas às demais (acima de 5 u.a.). A região do UV-254, que é caracterizada por ser marcador de matéria orgânica foi muito pronunciada para grande parte das amostras (Thomas, 2017). Dessa forma, o espectro UV-vis pode conter bastante informação química sobre a parte orgânica dos compostos fenólicos dissolvidos na água.

**Figura 1. Espectros de absorção na região do ultravioleta-visível.**



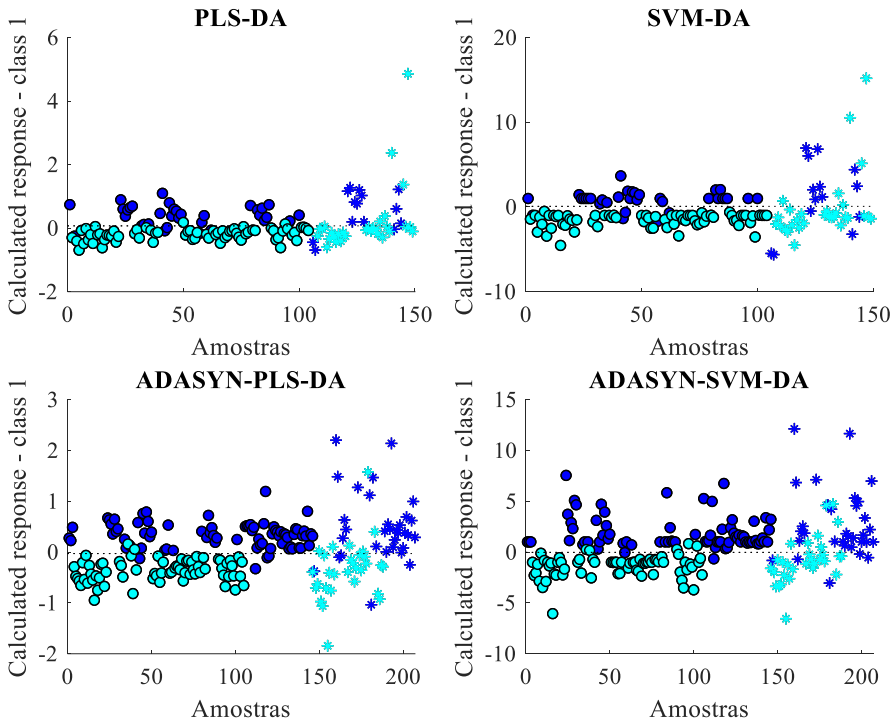

Fonte: Os autores, 2024.

A Figura 2 mostra os resultados dos modelos PLS-DA, SVM-DA, ADASYN-PLS-DA e ADASYN-SVM-DA usando absorção no UV-vis. As amostras da classe 1 (sem fenóis) e da classe 2 (com fenóis) formaram grupos bem distintos. A distribuição das amostras no espaço melhorou após a aplicação do ADASYN, principalmente corrigindo amostras que inicialmente se dispersavam muito do resto do conjunto, uma vez que outros exemplos parecidos foram sintetizados. Poucas amostras classificadas como FN e FP foram identificadas.
Na Tabela 1, são apresentados os resultados da comparação desses modelos. Para o modelo PLS-DA no conjunto de treinamento, os resultados para a Classe 1 mostraram uma sensibilidade de 85,3%, especificidade de 97,1%, precisão de 93,5% e acurácia de 93,3%. Para a Classe 2, a sensibilidade foi de 97,1%, a especificidade de 85,3%, a precisão de 93,2% e a acurácia de 93,3%. No conjunto de teste, a Classe 1 apresentou uma sensibilidade de 78,6%, especificidade de 83,3%, precisão de 68,8% e acurácia de 81,8%. A Classe 2 apresentou sensibilidade de 83,3%, especificidade de 78,6%, precisão de 89,3% e acurácia de 81,8%.

Para o modelo SVM-DA no conjunto de treinamento, a Classe 1 teve uma sensibilidade de 85,3%, especificidade de 100%, precisão de 100% e acurácia de 95,2%. A Classe 2 apresentou sensibilidade de 100%, especificidade de 85,3%, precisão de 93,3% e acurácia de 95,2%. No conjunto de teste, a Classe 1 apresentou sensibilidade de 64,3%, especificidade de 80,0%, precisão de 60,0% e acurácia de 75,0%. A Classe 2 apresentou sensibilidade de 80,0%, especificidade de 64,3%, precisão de 82,8% e acurácia de 75,0%.

Para o modelo ADASYN-PLS-DA no conjunto de treinamento, a Classe 1 teve uma sensibilidade de 94,7%, especificidade de 97,1%, precisão de 97,3% e acurácia de 95,9%. A Classe 2 apresentou sensibilidade de 97,1%, especificidade de 94,7%, precisão de 94,4% e acurácia de 95,9%. No conjunto de teste, a Classe 1 apresentou sensibilidade de 84,4%, especificidade de 80,0%, precisão de 81,8% e acurácia de 82,3%. A Classe 2 apresentou sensibilidade de 80,0%, especificidade de 84,4%, precisão de 82,8% e acurácia de 82,3%.

**Figura 2. Modelos PLS-DA e SVM-DA construídos usando UV-vis. As amostras de treinamento estão representadas pelos círculos e as amostras de testes são as estrelas. A cor azul representa a classe 1 e a cor ciano representa a classe 2.**




Fonte: Os autores, 2024.

Para o modelo ADASYN-SVM-DA no conjunto de treinamento, a Classe 1 teve uma sensibilidade de 97,4%, especificidade de 92,9%, precisão de 93,7% e acurácia de 95,2%. A Classe 2 apresentou sensibilidade de 92,9%, especificidade de 97,4%, precisão de 97,0% e acurácia de 95,2%. No conjunto de teste, a Classe 1 apresentou sensibilidade de 84,4%, especificidade de 76,7%, precisão de 79,4% e acurácia de 80,6%. A Classe 2 apresentou sensibilidade de 76,7%, especificidade de 84,4%, precisão de 82,1% e acurácia de 80,6%.

Esses resultados demonstram as diferenças de desempenho entre os modelos e conjuntos de dados, destacando as variações na sensibilidade, especificidade, precisão e acurácia para cada classe e método de análise. O uso de técnicas como ADASYN para balanceamento de dados parece melhorar o desempenho geral dos modelos em comparação com os métodos tradicionais, especialmente em termos de precisão e acurácia.

**Tabela 1. Resultados dos modelos de classificação usando absorção no UV-vis.**

| Model | Set | Cl | TP | TN | FP | FN | SEN | SPE | PRE | ACC |
|---|---|---|---|---|---|---|---|---|---|---|
| **PLS-DA** | Tre | 1 | 29 | 68 | 2 | 5 | 85,3 | 97,1 | 93,5 | 93,3 |
| | | 2 | 68 | 29 | 5 | 2 | 97,1 | 85,3 | 93,2 | 93,3 |
| | Tes | 1 | 11 | 25 | 5 | 3 | 78,6 | 83,3 | 68,8 | 81,8 |
| | | 2 | 25 | 11 | 3 | 5 | 83,3 | 78,6 | 89,3 | 81,8 |
| **SVM-DA** | Tre | 1 | 29 | 70 | 0 | 5 | 85,3 | 100,0 | 100,0 | 95,2 |
| | | 2 | 70 | 29 | 5 | 0 | 100,0 | 85,3 | 93,3 | 95,2 |
| | Tes | 1 | 9 | 24 | 6 | 5 | 64,3 | 80,0 | 60,0 | 75,0 |
| | | 2 | 24 | 9 | 5 | 6 | 80,0 | 64,3 | 82,8 | 75,0 |
| **ADASYN-PLS-DA** | Tre | 1 | 72 | 68 | 2 | 4 | 94,7 | 97,1 | 97,3 | 95,9 |
| | | 2 | 68 | 72 | 4 | 2 | 97,1 | 94,7 | 94,4 | 95,9 |
| | Tes | 1 | 27 | 24 | 6 | 5 | 84,4 | 80,0 | 81,8 | 82,3 |
| | | 2 | 24 | 27 | 5 | 6 | 80,0 | 84,4 | 82,8 | 82,3 |
| **ADASYN-SVM-DA** | Tre | 1 | 74 | 65 | 5 | 2 | 97,4 | 92,9 | 93,7 | 95,2 |
| | | 2 | 65 | 74 | 2 | 5 | 92,9 | 97,4 | 97,0 | 95,2 |
| | Tes | 1 | 27 | 23 | 7 | 5 | 84,4 | 76,7 | 79,4 | 80,6 |
| | | 2 | 23 | 27 | 5 | 7 | 76,7 | 84,4 | 82,1 | 80,6 |

Fonte: elaboração própria a partir dos resultados dos modelos (2024).

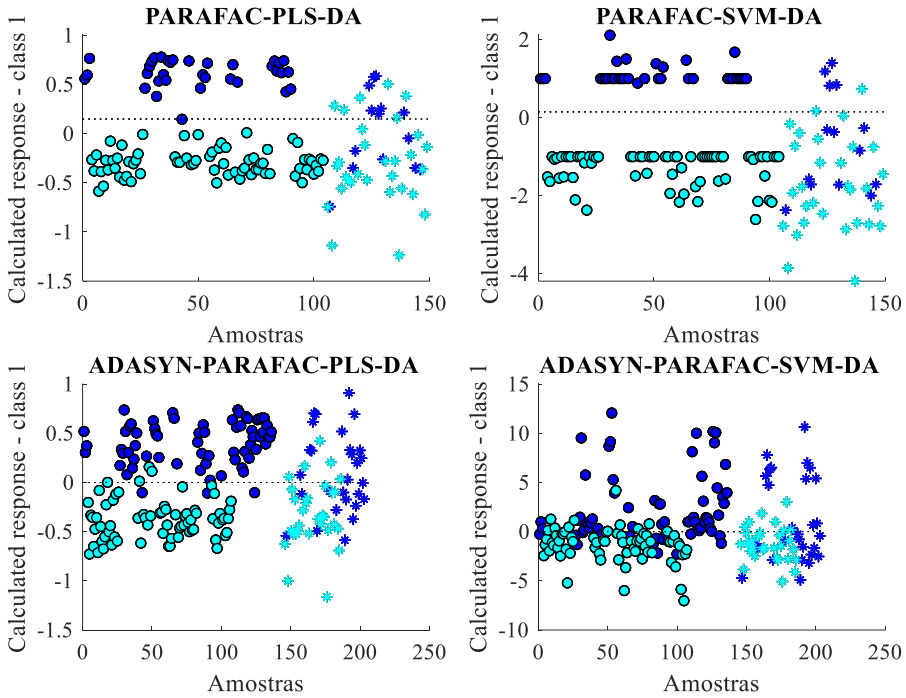
A Figura 3 ilustra a comparação visual dos modelos de classificação aplicados aos dados de fluorescência EEM que foram analisados utilizando o método PARAFAC. Com exceção do modelo ADASYN-PARAFC-SVM-DA, todos os outros modelos demonstraram uma excelente capacidade de separação no conjunto de treinamento. No entanto, ao avaliar os conjuntos de teste, a distinção entre as classes 1 e 2 não foi tão eficaz, resultando em muitas amostras classificadas de forma incorreta.

Essa tendência de desempenho menos satisfatório persistiu nos modelos que empregaram a geração de amostras sintéticas. Isso sugere que os espectros de fluorescência EEM, após o processamento pelo PARAFAC, contribuem pouco para uma separação eficiente entre as classes. Muitas amostras foram erroneamente classificadas como falsos positivos e falsos negativo. Esses resultados indicam que as informações espectrais contidas nos espectros de fluorescência EEM, embora estejam altamente correlacionadas com compostos fenólicos, não foram totalmente aproveitadas pela modelagem utilizando o algoritmo PARAFAC neste contexto específico.

Os resultados avaliados na Tabela 2 são os dos modelos PARAFAC-PLS-DA, PARAFAC-SVM-DA, ADASYN-PARAFAC-PLS-DA e ADASYN-PARAFAC-SVM-DA. Para o modelo PARAFAC-PLS-DA no conjunto de treinamento, os resultados para a Classe 1 mostraram uma sensibilidade de 97,1%, especificidade de 100%, precisão de 100% e acurácia de 99%. Para a Classe 2, a sensibilidade foi de 100%, especificidade de 97,1%, precisão de 98,6% e acurácia de 99%. No conjunto de teste, a Classe 1 apresentou uma sensibilidade de 50%, especificidade de 80%, precisão de 53,8% e acurácia de 70,5%. A Classe 2 apresentou sensibilidade de 80%, especificidade de 50%, precisão de 77,4% e acurácia de 70,5%.

Para o modelo PARAFAC-SVM-DA no conjunto de treinamento, a Classe 1 teve uma sensibilidade de 100%, especificidade de 100%, precisão de 100% e acurácia de 100%. A Classe 2 também apresentou uma sensibilidade de 100%, especificidade de 100%, precisão de 100% e acurácia de 100%. No conjunto de teste, a Classe 1 apresentou uma sensibilidade de 28,6%, especificidade de 90%, precisão de 57,1% e acurácia de 70,5%. A Classe 2 apresentou sensibilidade de 90%, especificidade de 28,6%, precisão de 73% e acurácia de 70,5%.

**Figura 3. Modelos PLS-DA e SVM-DA construídos usando fluorescência EEM com PARAFAC. As amostras de treinamento estão representadas pelos círculos e as amostras de testes são as estrelas. A cor azul representa a classe 1 e a cor ciano representa a classe 2.**




Fonte: Os autores, 2024.

**Tabela 2. Resultados dos modelos de classificação usando fluorescência EEM com PARAFAC.**

| Model | Set | Cl | TP | TN | FP | FN | SEN | SPE | PRE | ACC |
|---|---|---|---|---|---|---|---|---|---|---|
| **PARAFAC-PLS-DA** | Tre | 1 | 33 | 70 | 0 | 1 | 97,1 | 100,0 | 100,0 | 99,0 |
| | | 2 | 70 | 33 | 1 | 0 | 100,0 | 97,1 | 98,6 | 99,0 |
| | Tes | 1 | 7 | 24 | 6 | 7 | 50,0 | 80,0 | 53,8 | 70,5 |
| | | 2 | 24 | 7 | 7 | 6 | 80,0 | 50,0 | 77,4 | 70,5 |
| **PARAFAC-SVM-DA** | Tre | 1 | 34 | 70 | 0 | 0 | 100,0 | 100,0 | 100,0 | 100,0 |
| | | 2 | 70 | 34 | 0 | 0 | 100,0 | 100,0 | 100,0 | 100,0 |
| | Tes | 1 | 4 | 27 | 3 | 10 | 28,6 | 90,0 | 57,1 | 70,5 |
| | | 2 | 27 | 4 | 10 | 3 | 90,0 | 28,6 | 73,0 | 70,5 |
| **ADASYN-PARAFAC-PLS-DA** | Tre | 1 | 63 | 66 | 4 | 3 | 95,5 | 94,3 | 94,0 | 94,9 |
| | | 2 | 66 | 63 | 3 | 4 | 94,3 | 95,5 | 95,7 | 94,9 |
| | Tes | 1 | 15 | 25 | 5 | 13 | 53,6 | 83,3 | 75,0 | 69,0 |
| | | 2 | 25 | 15 | 13 | 5 | 83,3 | 53,6 | 65,8 | 69,0 |
| **ADASYN-PARAFAC-SVM-DA** | Tre | 1 | 51 | 59 | 11 | 15 | 77,3 | 84,3 | 82,3 | 80,9 |
| | | 2 | 59 | 51 | 15 | 11 | 84,3 | 77,3 | 79,7 | 80,9 |
| | Tes | 1 | 13 | 23 | 7 | 15 | 46,4 | 76,7 | 65,0 | 62,1 |
| | | 2 | 23 | 13 | 15 | 7 | 76,7 | 46,4 | 60,5 | 62,1 |

Fonte: elaboração própria a partir dos resultados dos modelos (2024).

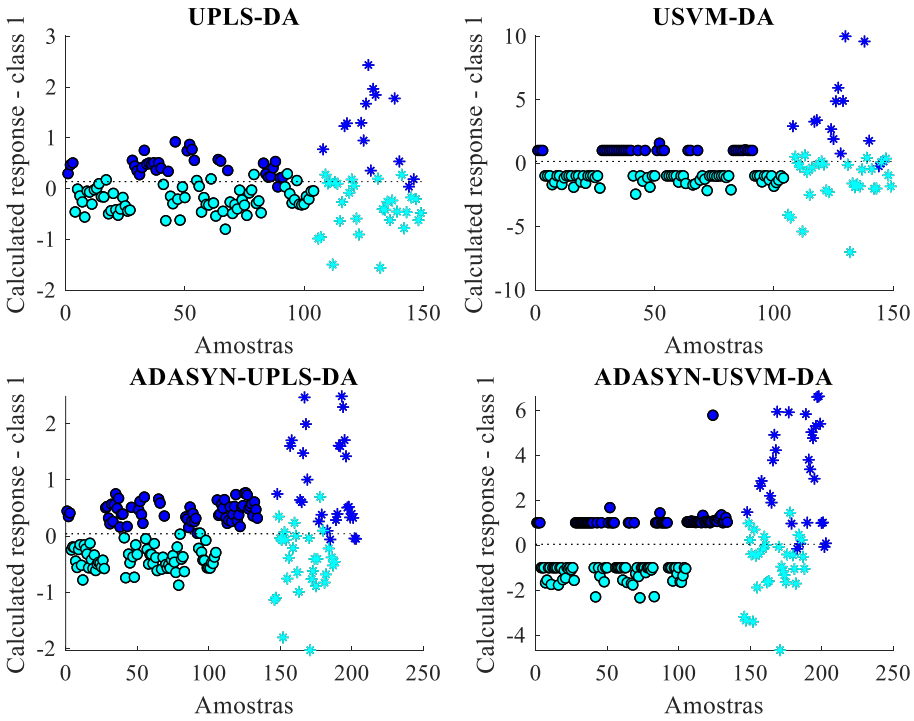
Para o modelo ADASYN-PARAFAC-PLS-DA no conjunto de treinamento, a Classe 1 teve uma sensibilidade de 95,5%, especificidade de 94,3%, precisão de 94% e acurácia de 94,9%. A Classe 2 apresentou sensibilidade de 94,3%, especificidade de 95,5%, precisão de 95,7% e acurácia de 94,9%. No conjunto de teste, a Classe 1 apresentou sensibilidade de 53,6%, especificidade de 83,3%, precisão de 75% e acurácia de 69%. A Classe 2 apresentou sensibilidade de 83,3%, especificidade de 53,6%, precisão de 65,8% e acurácia de 69%.

Para o modelo ADASYN-PARAFAC-SVM-DA no conjunto de treinamento, a Classe 1 teve uma sensibilidade de 77,3%, especificidade de 84,3%, precisão de 82,3% e acurácia de 80,9%. A Classe 2 apresentou sensibilidade de 84,3%, especificidade de 77,3%, precisão de 79,7% e acurácia de 80,9%. No conjunto de teste, a Classe 1 apresentou sensibilidade de 46,4%, especificidade de 76,7%, precisão de 65% e acurácia de 62,1%. A Classe 2 apresentou sensibilidade de 76,7%, especificidade de 46,4%, precisão de 60,5% e acurácia de 62,1%. Esses resultados demonstram que, no conjunto de treinamento, o modelo PARAFAC-SVM-DA obteve desempenho perfeito (100%) para ambas as classes, indicando que foi capaz de classificar corretamente todas as amostras. No entanto, esse desempenho não se manteve no conjunto de teste, onde o modelo apresentou uma sensibilidade de apenas 28,6% para a Classe 1, embora tenha mantido uma alta especificidade de 90%. O modelo PARAFAC-PLS-DA também apresentou um excelente desempenho no conjunto de treinamento, mas teve uma queda significativa no conjunto de teste, especialmente na sensibilidade da Classe 1 (50%).

A aplicação do ADASYN melhorou a distribuição dos dados de treinamento, mas os resultados no conjunto de teste mostraram uma variação considerável. O modelo ADASYN-PARAFAC-PLS-DA apresentou uma acurácia de 69% para ambas as classes no conjunto de teste, enquanto o modelo ADASYN-PARAFAC-SVM-DA apresentou a menor acurácia geral (62,1%). Em resumo, embora os modelos apresentem um excelente desempenho no treinamento, a generalização para novos dados (conjunto de teste) ainda apresenta desafios, destacando a importância de um equilíbrio entre o ajuste do modelo e a capacidade de generalização para novas amostras.

A Figura 4 apresenta os resultados da comparação entre diferentes modelos de classificação aplicados aos dados de fluorescência EEM, que foram analisados utilizando a técnica *unfold-multiway*. Os modelos avaliados incluem UPLS-DA, USVM-DA, ADASYN-UPLS-DA e ADASYN-USVM-DA.

**Figura 4. Modelos PLS-DA e SVM-DA construídos usando fluorescência EEM com *unfold-multiway*. As amostras de treinamento estão representadas pelos círculos e as amostras de testes são as estrelas. A cor azul representa a classe 1 e a cor ciano representa a classe 2.**




Fonte: Os autores, 2024.

Em todos os casos, os modelos de classificação demonstraram uma notável capacidade de separação entre as classes no conjunto de treinamento. Nos conjuntos de teste, observou-se uma separação excelente sem a

necessidade de geração de amostras sintéticas. Apesar da aplicação do método ADASYN, essa qualidade na separação persistiu e a precisão permaneceu elevada. Esses resultados indicam que os modelos de classificação foram eficazes na diferenciação das classes, tanto no conjunto de treinamento quanto no conjunto de teste, mesmo após a utilização do ADASYN para equilibrar as classes.

A Tabela 3 apresenta os resultados estatísticos dos modelos mencionados acima. Para o modelo UPLS-DA no conjunto de treinamento, os resultados para a Classe 1 mostraram uma sensibilidade de 97,1%, especificidade de 91,4%, precisão de 84,6% e acurácia de 93,3%. Para a Classe 2, a sensibilidade foi de 91,4%, especificidade de 97,1%, precisão de 98,5% e acurácia de 93,3%. No conjunto de teste, a Classe 1 apresentou uma sensibilidade de 92,9%, especificidade de 86,7%, precisão de 76,5% e acurácia de 88,6%. A Classe 2 apresentou sensibilidade de 86,7%, especificidade de 92,9%, precisão de 96,3% e acurácia de 88,6%.

Para o modelo USVM-DA no conjunto de treinamento, a Classe 1 teve uma sensibilidade de 100%, especificidade de 100%, precisão de 100% e acurácia de 100%. A Classe 2 também apresentou uma sensibilidade de 100%, especificidade de 100%, precisão de 100% e acurácia de 100%. No conjunto de teste, a

**Tabela 3. Resultados dos modelos de classificação usando fluorescência EEM com *unfold-multiway*.**

| Model | Set | Cl | TP | TN | FP | FN | SEN | SPE | PRE | ACC |
|---|---|---|---|---|---|---|---|---|---|---|
| **UPLS-DA** | Tre | 1 | 33 | 64 | 6 | 1 | 97,1 | 91,4 | 84,6 | 93,3 |
| | | 2 | 64 | 33 | 1 | 6 | 91,4 | 97,1 | 98,5 | 93,3 |
| | Tes | 1 | 13 | 26 | 4 | 1 | 92,9 | 86,7 | 76,5 | 88,6 |
| | | 2 | 26 | 13 | 1 | 4 | 86,7 | 92,9 | 96,3 | 88,6 |
| **USVM-DA** | Tre | 1 | 34 | 70 | 0 | 0 | 100,0 | 100,0 | 100,0 | 100,0 |
| | | 2 | 70 | 34 | 0 | 0 | 100,0 | 100,0 | 100,0 | 100,0 |
| | Tes | 1 | 13 | 24 | 6 | 1 | 92,9 | 80,0 | 68,4 | 84,1 |
| | | 2 | 24 | 13 | 1 | 6 | 80,0 | 92,9 | 96,0 | 84,1 |
| **ADASYN-UPLS-DA** | Tre | 1 | 64 | 69 | 1 | 0 | 100,0 | 98,6 | 98,5 | 99,3 |
| | | 2 | 69 | 64 | 0 | 1 | 98,6 | 100,0 | 100,0 | 99,3 |
| | Tes | 1 | 25 | 25 | 5 | 3 | 89,3 | 83,3 | 83,3 | 86,2 |
| | | 2 | 25 | 25 | 3 | 5 | 83,3 | 89,3 | 89,3 | 86,2 |
| **ADASYN-USVM-DA** | Tre | 1 | 64 | 70 | 0 | 0 | 100,0 | 100,0 | 100,0 | 100,0 |
| | | 2 | 70 | 64 | 0 | 0 | 100,0 | 100,0 | 100,0 | 100,0 |
| | Tes | 1 | 26 | 22 | 8 | 2 | 92,9 | 73,3 | 76,5 | 82,8 |
| | | 2 | 22 | 26 | 2 | 8 | 73,3 | 92,9 | 91,7 | 82,8 |

Fonte: elaboração própria a partir dos resultados dos modelos (2024).

Classe 1 apresentou uma sensibilidade de 92,9%, especificidade de 80%, precisão de 68,4% e acurácia de 84,1%. A Classe 2 apresentou sensibilidade de 80%, especificidade de 92,9%, precisão de 96% e acurácia de 84,1%. Para o modelo ADASYN-UPLS-DA no conjunto de treinamento, a Classe 1 teve uma sensibilidade de 100%, especificidade de 98,6%, precisão de 98,5% e acurácia de 99,3%. A Classe 2 apresentou sensibilidade de 98,6%, especificidade de 100%, precisão de 100% e acurácia de 99,3%. No conjunto de teste, a Classe 1 apresentou sensibilidade de 89,3%, especificidade de 83,3%, precisão de 83,3% e acurácia de 86,2%. A Classe 2 apresentou sensibilidade de 83,3%, especificidade de 89,3%, precisão de 89,3% e acurácia de 86,2%.

Para o modelo ADASYN-USVM-DA no conjunto de treinamento, a Classe 1 teve uma sensibilidade de 100%, especificidade de 100%, precisão de 100% e acurácia de 100%. A Classe 2 também apresentou uma sensibilidade de 100%, especificidade de 100%, precisão de 100% e acurácia de 100%. No conjunto de teste, a

Classe 1 apresentou sensibilidade de 92,9%, especificidade de 73,3%, precisão de 76,5% e acurácia de 82,8%. A Classe 2 apresentou sensibilidade de 73,3%, especificidade de 92,9%, precisão de 91,7% e acurácia de 82,8%.

Os resultados demonstram que no conjunto de treinamento, os modelos USVM-DA e ADASYN-USVM-DA alcançaram um desempenho perfeito (100%) para ambas as classes, indicando que classificaram corretamente todas as amostras. No entanto, esse desempenho não se manteve no conjunto de teste, onde houve uma queda nas métricas, especialmente na especificidade do modelo ADASYN-USVM-DA para a Classe 1 (73,3%). O modelo UPLS-DA também apresentou um desempenho sólido no conjunto de treinamento, mas a acurácia e outras métricas caíram no conjunto de teste, particularmente na precisão da Classe 1 (76,5%).

A aplicação do ADASYN melhorou a distribuição dos dados de treinamento, resultando em uma melhoria geral nas métricas de acurácia e precisão. No entanto, no conjunto de teste, embora as métricas tenham melhorado para algumas classes, os desafios na generalização dos modelos ainda são evidentes. Em resumo, os resultados destacam a necessidade de equilibrar o ajuste do modelo com a capacidade de generalização para novas amostras, evidenciando a importância de validar os modelos em conjuntos de dados independentes.

## CONCLUSÕES

A análise comparativa entre modelos de classificação utilizando espectros UV-vis e fluorescência EEM demonstra variações no desempenho dos modelos em termos de sensibilidade, especificidade, precisão e acurácia. Modelos foram desenvolvidos com e sem a técnica de amostragem sintética ADASYN, resultando em PLS-DA, SVM-DA, ADASYN-PLS-DA, ADASYN-SVM-DA para espectros UV-vis, e UPLS-DA, USVM-DA, ADASYN-UPLS-DA, ADASYN-USVM-DA, PARAFAC-PLS-DA, PARAFAC-SVM-DA, ADASYN-PARAFAC-PLS-DA e ADASYN-PARAFAC-SVM-DA para espectros de fluorescência EEM.

O método ADASYN-PLS-DA obteve a maior média de acurácia no teste (82,3%) para dados UV-vis. Já para dados de fluorescência EEM com PARAFAC, os modelos PARAFAC-PLS-DA e PARAFAC-SVM-DA alcançaram 70,5% de acurácia média no teste. O modelo UPLS-DA obteve a maior média de acurácia no teste (88,6%) para dados de fluorescência EEM com *unfold-multiway*. A aplicação do método ADASYN melhorou a performance dos modelos PLS-DA e SVM-DA para dados UV-vis. Porém, para os modelos EEM com PARAFAC, houve uma redução na acurácia em ambos os conjuntos de treinamento e teste. Nos modelos EEM com *unfold-multiway*, a acurácia aumentou no treinamento, mas houve uma pequena queda no teste em comparação com os modelos sem amostragem sintética, devido ao aumento de amostras classificadas erroneamente. No entanto, o método ADASYN contribuiu para a melhoria das métricas dos modelos em todos os espectros ao gerar amostras sintéticas.

Em termos de aplicabilidade prática, essas metodologias oferecem *insights* para o monitoramento de fenóis em situações reais. A técnica de amostragem sintética ADASYN pode ser particularmente útil em contextos onde a quantidade de amostras é limitada, ajudando a melhorar a robustez dos modelos. Para aplicações no monitoramento ambiental ou industrial, onde a detecção precisa e a classificação de fenóis são cruciais, o uso de espectros UV-vis e fluorescência EEM com modelos aprimorados por ADASYN pode proporcionar um desempenho superior. A integração dessas metodologias em sistemas de monitoramento em tempo real pode levar a melhorias na detecção e na análise de fenóis, facilitando uma resposta mais ágil e precisa a alterações nas condições ambientais ou processos industriais.

## REFERÊNCIAS BIBLIOGRÁFICAS

1. ANKU, W.W., MAMO, M.A., & GOVENDER, P.P. *Phenolic Compounds in Water: Sources, Reactivity, Toxicity and Treatment Methods. Phenolic Compounds - Natural Sources, Importance and Applications.* 2017.

2. ASTM International. *ASTM D8431-22 - Standard Test Method for Detection of Water-soluble Petroleum Oils by A-TEEM Optical Spectroscopy and Multivariate Analysis.* 2022.

3. BAHRAM, M., BRO, R., STEDMON, C., & AFKHAMI, A. *Handling of Rayleigh and Raman scatter for PARAFAC modeling of fluorescence data using interpolation. Journal of Chemometrics*, 20(3–4), 99–105. 2006.

4. BAIRD, R.B., EATON, A.D., & RICE, E.W. *Standard methods: For the examination of water and wastewater. In American Public Health Association; American Water Works Association; Water Environment Federation*, 1990.

5. BALLABIO, D., & CONSONNI, V. *Classification tools in chemistry. Part 1: Linear models. PLS-DA. In Analytical Methods*, 5(16), 3790–3798. 2013.

6. BARNES, R. J., DHANOA, M. S., & LISTER, S. J. *Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra. Applied Spectroscopy*, 43(5), 772–777. 1989.

7. BRANDÃO, C.J., BOTELHO, M.J.C., SATO, M.I.Z., & LAMPARELLI, M.C. *Guia Nacional De Coleta E Preservação De Amostras*, São Paulo: CETESB; Brasília: ANA, 2011. 325 p. 2011.

8. BRERETON, R.G. *Contingency tables, confusion matrices, classifiers and quality of prediction. Journal of Chemometrics*, 35(11). 2021.

9. BRO, R. *PARAFAC: Tutorial and applications. Chemometrics and Intelligent Laboratory Systems.* 38(2), 149–171. 1997.

10. FERREIRA, M.M.C. *Quimiometria: conceitos, métodos e aplicações (1st ed.). Editora da Unicamp.* 2015.

11. FILISBINO, T.A., GIRALDI, G.A., & THOMAZ, C.E. *Support vector machine ensembles for discriminant analysis for ranking principal components. Multimedia Tools and Applications*, 79(35–36), 25277–25313. 2020.

12. HE, H., BAI, Y., GARCIA, E. A., & LI, S. *ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. In IEEE Xplore. International Joint Conference on Neural Networks*, pp. 1322–1328. 2008.

13. KENNARD, R.W., & STONE, L.A. *Computer Aided Design of Experiments. Technometric*, 11(1), 137–148. 1969.

14. RAMOS, R.L., MOREIRA, V.R., & AMARAL, M.C.S. *Phenolic compounds in water: Review of occurrence, risk, and retention by membrane technology. Journal of Environmental Management*, 351, 119772. 2024.

15. LAKOWICZ, J.R. *Principles of fluorescence spectroscopy. Springer.* 2006.

16. LIPPS, W.C., BRAUN-HOWLAND, E.B., & BAXTER, T.E. *Standard Methods for the Examination of Water and Wastewater* (24th ed.). APHA, AWWA, WEF. 2023.

17. MURPHY, K.R., STEDMON, C.A., GRAEBER, D., & BRO, R. *Fluorescence spectroscopy and multi-way techniques. PARAFAC. Analytical Methods*, 5(23), 6557–6566. 2013.

18. NØRGAARD, L. *Direct standardisation in multi wavelength fluorescence spectroscopy. Chemometrics and Intelligent Laboratory Systems*, 29(2), 283–293. 1995.

19. OLIVIERI, AC., ESCANDAR, G.M., GOICOECHEA, H.C., & DE LA PEÑA, A.M. *Unfolded and Multiway Partial Least-Squares with Residual Multilinearization: Fundamentals. Data Handling in Science and Technology*, 29, 365–397. 2015.

20. PAULO, E.H., MAGALHÃES, G.B., MOREIRA, M.P.B., NASCIMENTO, M.H.C., HERINGER, O.A., FILGUEIRAS, P.R., & FERRÃO, M.F. *Classification of water by bacterial presence using chemometrics associated with excitation-emission matrix fluorescence spectroscopy*. *Microchemical Journal*, 197. 2024.

21. RINNAN, Å., BERG, F. van den, & ENGELSEN, S. B. (2009). *Review of the most common pre-processing techniques for near-infrared spectra. TrAC - Trends in Analytical Chemistry*, 28(10), 1201–1222. https://doi.org/10.1016/j.trac.2009.07.007

22. TCHAIKOVSKAYA, O., SOKOLOVA, I., BAZYL, O., SWETLICHNYI, V., KOPYLOVA, T., MAYER, G., & SULTIMOVA, N. *The fluorescence analysis of laser photolysis of phenols in water. International Journal of Photoenergy*, 4(2), 79–83. 2002.

23. THOMAS, O., & BURGESS, C. *UV-Visible Spectrophotometry of Water and Wastewater. In UV-Visible Spectrophotometry of Water and Wastewater.* (2nd ed.). Elsevier Science. 2017.

24. XIAO, Q., CHEN, F., ZHOU, Y., CHEN, L., & LI, J. *Analysis of Spectra and Intensity of 3D Fluorescence of Phenol Dissolved in Water. Proceedings of the 2014 International Conference on Mechatronics, Control and Electronic Engineering*, 113, 575–579. 2014.