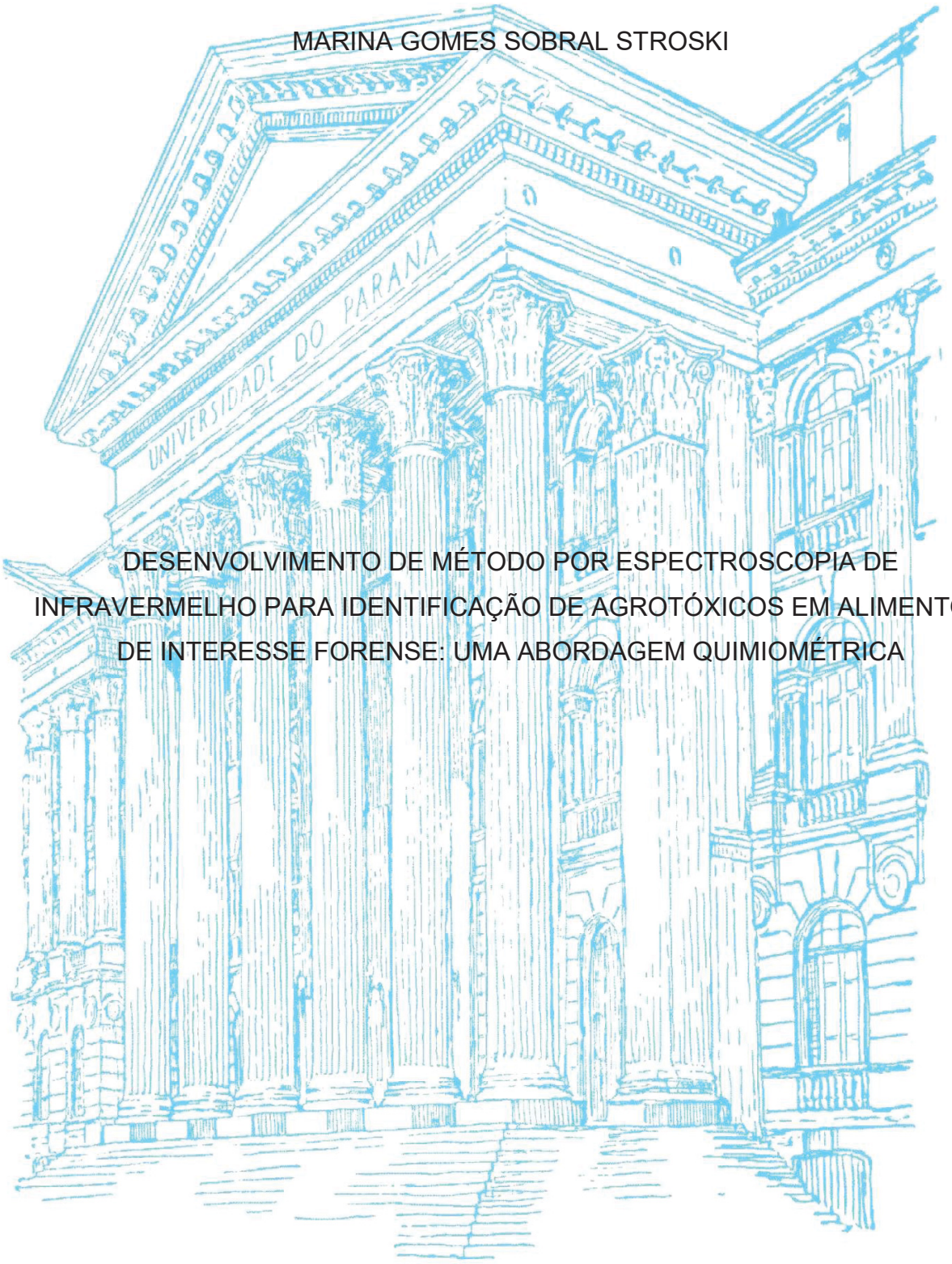
UNIVERSIDADE FEDERAL DO PARANÁ

MARINA GOMES SOBRAL STROSKI



DESENVOLVIMENTO DE MÉTODO POR ESPECTROSCOPIA DE INFRAVERMELHO PARA IDENTIFICAÇÃO DE AGROTÓXICOS EM ALIMENTOS DE INTERESSE FORENSE: UMA ABORDAGEM QUIMIOMÉTRICA

CURITIBA

2025

MARINA GOMES SOBRAL STROSKI

DESENVOLVIMENTO DE MÉTODO POR ESPECTROSCOPIA DE INFRAVERMELHO PARA IDENTIFICAÇÃO DE AGROTÓXICOS EM ALIMENTOS DE INTERESSE FORENSE: UMA ABORDAGEM QUIMIOMÉTRICA

Dissertação apresentada Programa de Pós-Graduação em Ciências Farmacêuticas da Universidade Federal do Paraná, área de Medicamentos, Insumos e Correlatos, linha de Pesquisa de Produção e Qualidade, Setor de Ciências da Saúde, como requisito parcial à obtenção do título de Mestre em Ciências Farmacêuticas.

Orientador: Prof. Dr. Roberto Pontarolo

CURITIBA
2025

Stroski, Marina Gomes Sobral
Desenvolvimento de método por espectroscopia de infravermelho para identificação de agrotóxicos em alimentos de interesse forense [recurso eletrônico]: uma abordagem quimiométrica / Marina Gomes Sobral Stroski. – Curitiba, 2025.
1 recurso online : PDF

Dissertação (mestrado) – Programa de Pós-Graduação em Ciências Farmacêuticas. Setor de Ciências da Saúde, Universidade Federal do Paraná, 2025.

Orientador: Prof. Dr. Roberto Pontarolo

1. Agroquímicos. 2. Espectroscopia de infravermelho com transformada de Fourier. 3. Quimiometria. 4. Aprendizado de máquina. I. Pontarolo, Roberto. II. Universidade Federal do Paraná. III. Título.

CDD 615.9

Maria da Conceição Kury da Silva CRB 9/1275

MINISTÉRIO DA EDUCAÇÃO
SETOR DE CIÊNCIAS DA SAÚDE
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO CIÊNCIAS FARMACÊUTICAS - 40001016042P8

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação CIÊNCIAS FARMACÊUTICAS da Universidade Federal do Paraná foram convocados para realizar a arguição da dissertação de Mestrado de **MARINA GOMES SOBRAL STROSKI**, intitulada: **Desenvolvimento de método por espectroscopia de infravermelho para identificação de agrotóxicos em alimentos de interesse forense: uma abordagem quimiométrica.**, sob orientação do Prof. Dr. ROBERTO PONTAROLO, que após terem inquirido a aluna e realizada a avaliação do trabalho, são de parecer pela sua aprovação no rito de defesa.

A outorga do título de mestra está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 31 de Julho de 2025.

ROBERTO PONTAROLO
Presidente da Banca Examinadora

RAUL EDISON LUNA LAZO
Avaliador Externo (UNIVERSIDADE FEDERAL DO PARANÁ - UFPR)

RAQUEL DE OLIVEIRA VILHENA
Avaliador Externo (UNIVERSIDADE FEDERAL DO PARANÁ)

Dedico este trabalho aos meus pais:

Josafá Ferreira Sobral e Sônia Gomes de Sousa,

que com muito amor e carinho, sempre me incentivaram a estudar.

AGRADECIMENTOS

Agradeço primeiramente à Deus, pela presença constante em minha vida. Foi d'Ele que veio o fôlego necessário para que eu conseguisse trilhar e concluir mais esta etapa. Aos meus pais, Sônia e Josafá, e minha irmã Marília, minha gratidão eterna. Desde pequena lembro da minha mãe falando que podem tirar tudo de mim, menos o conhecimento. Por isso agradeço cada incentivo, cada palavra de apoio e por serem o alicerce sobre o qual construí os meus sonhos. Agradeço também ao meu amor Fernando e ao meu amado filho Francisco, que são meu porto seguro, minha paz, meu descanso e o lugar onde meu coração sempre quer estar.

Agradeço aos meus amigos espirituais que me acompanharam silenciosamente ao longo deste percurso. Mesmo invisíveis aos olhos, senti sua presença em momentos decisivos — nas intuições certeiras, no alívio em meio ao cansaço, no consolo sereno das horas difíceis. À vocês, minha prece sincera e o reconhecimento por terem me sustentado com luz, força e coragem quando precisei.

Agradeço ao professor Roberto Pontarolo pela confiança depositada em mim ao aceitar orientar este trabalho, sempre com respeito e paciência, da forma mais elegante possível. Seu apoio foi fundamental para que eu chegasse até aqui.

Meu sincero reconhecimento à Polícia Científica do Paraná e à Academia de Ciências Forenses, pelo suporte institucional e pela aprovação deste projeto. Aos amigos da Seção de Química Forense, minha segunda casa, obrigada por cada aprendizado e companheirismo. Agradeço nominalmente à Fabia, Gabriela, Amanda e Danilo pelo conhecimento compartilhado, por me apoiarem nos meus momentos de preocupação e por fazerem parte da minha caminhada. Agradeço à Maria Eduarda pelos conselhos sinceros - de vida e de trabalho - que me acompanharam em momentos importantes. Agradeço ao Luís pelas valiosas trocas de ideias que contribuíram significativamente para o aprimoramento técnico deste trabalho. Também não posso de deixar de agradecer ao amigo Eduardo, minha referência na Química Forense, pelas conversas que ultrapassavam os limites do trabalho e tantas vezes me ofereceram inspiração, apoio e descanso quando precisei, mesmo sem ele perceber. Meus agradecimentos também à Isabella, que estava como chefe da Seção de Química Forense no início deste mestrado, por todo o apoio para reorganizar minha carga horária e viabilizar minha formação. Obrigada pelo

conhecimento compartilhado e pelas infinitas vezes em que me acolheu nos meus períodos de desânimo. Com especial carinho também agradeço à Raquel que, com sua sensibilidade e leveza, foi, aos poucos, conquistando um lugar em meu coração, tornando-se hoje uma referência técnica pra mim e, principalmente, uma referência de pessoa em quem sei que posso confiar.

Ao professor doutor Frederico, excelente quimiometrista do Departamento de Química da Universidade Federal do Paraná, minha sincera gratidão pelas inúmeras mensagens trocadas ao longo desta jornada, sempre respondidas com prontidão, paciência e generosidade, esclarecendo dúvidas e contribuindo de forma valiosa para o desenvolvimento deste trabalho.

Ao Centro de Estudos em Biofarmácia e ao Grupo de Estudos em Avaliação de Tecnologias em Saúde da Universidade Federal do Paraná, meu carinho e gratidão por me acolherem com tanta paciência desde o início, quase literalmente me guiando pela mão para que eu pudesse compreender o básico e crescer a partir dele. Ao doutor Raul, que foi o primeiro a me receber, oferecendo todo o apoio que precisei. À querida Aline, ao doce Ahmad, ao anjo João, à toda a luz que o Moisés me proporcionou, aos estagiários que me auxiliaram na coleta dos espectros, à querida Sindy pelo reencontro providencial e pelas conversas que sempre me faziam bem, ao Alexandre, ao Dalton, à Patrícia, ao Eric, à Carol e a todos que me ajudaram: saibam que este trabalho também carrega um pedaço de cada um de vocês. E aos que, de alguma forma, fizeram parte desta trajetória: meu mais profundo e sincero muito obrigada!

Por fim, meus agradecimentos à Universidade Federal do Paraná por mais uma vez me proporcionar uma formação pública, gratuita e de excelência. Sinto-me privilegiada por integrar novamente esta instituição.

Eu posso ir muito além de onde estou
Vou nas asas do Senhor
O Teu amor é o que me conduz
Posso voar e subir sem me cansar
Ir pra frente sem me fatigar
Vou com asas, como águia
Pois confio no Senhor!

Eros Biondini

RESUMO

Em laboratórios de química forense, é comum a análise em materiais contaminados por agrotóxicos relacionados a homicídios ou crimes ambientais. Entre os compostos envolvidos, destacam-se os raticidas anticoagulantes e os inseticidas organofosforados e carbamatos. As técnicas clássicas de identificação, como a cromatografia gasosa e líquida acoplada à espectrometria de massas, embora sensíveis, são dispendiosas, exigem extrações laboriosas e fazem uso intensivo de solventes. Nesse contexto, este trabalho propõe o desenvolvimento de uma metodologia baseada em espectroscopia de infravermelho com Transformada de Fourier (FTIR), associada a técnicas quimiométricas, para a detecção de agrotóxicos em alimentos. A proposta busca contribuir para a área da química forense por meio da avaliação do potencial de um método de triagem rápido, não destrutivo, de baixo custo e alinhado aos princípios da química verde, devido à redução do uso de solventes. As amostras incluíram cinco grupos de alimentos: pão, mortadela, arroz, carne moída e uma mistura contendo diferentes alimentos. Cada matriz foi contaminada com diferentes agrotóxicos (aldicarbe, brodifacoum, bromadiolona, clorpirifós, metomil e terbufós) em concentrações variadas (entre 2 e 50 ppm). As análises espectrais foram conduzidas no modo de transmitância, com resolução de 4 cm^{-1} e 24 varreduras por espectro, na faixa de 1800 a 400 cm^{-1}. Constatou-se que, entre os cinco grupos alimentares, o pão e a mortadela apresentaram melhor desempenho, sendo, portanto, selecionados para as principais análises quimiométricas. As médias espectrais de cada grupo foram utilizadas na análise de componentes principais (PCA), o que permitiu identificar claramente os padrões de agrupamento entre os alimentos contaminados. Em seguida, a análise discriminante por mínimos quadrados parciais (PLS-DA) foi aplicada para construir um modelo de classificação, que atingiu acurácia de 86%. A aplicabilidade do modelo foi avaliada com duas amostras reais da Polícia Científica do Paraná. A amostra de pão contaminada com brodifacoum foi corretamente classificada; entretanto, a amostra de mortadela contaminada com terbufós foi identificada como contendo clorpirifós. Considerando as limitações do modelo treinado com o algoritmo PLS-DA, foi realizada uma etapa complementar utilizando a biblioteca PyCaret no Google Colab para comparar o desempenho de 15 algoritmos de classificação. Os dois melhores modelos foram os treinados com os algoritmos *Linear Discriminant Analysis* (LDA), com acurácia de 81%, e *Light Gradient Boosting Machine* (LightGBM), com acurácia de 67%. O LDA classificou corretamente a amostra de pão contaminada com brodifacoum, mas errou na classificação da mortadela contaminada com terbufós. Por outro lado, o LightGBM identificou corretamente a contaminação por terbufós na mortadela e classificou o pão como bromadiolona, um erro no composto, porém coerente com a classe química, devido à semelhança estrutural entre brodifacoum e bromadiolona. Os resultados obtidos devem ser compreendidos como parte de uma prova de conceito, servindo de base para futuras etapas de refinamento metodológico. Entre as perspectivas futuras, destacam-se o ajuste dos hiperparâmetros, a diversificação da base espectral, a definição de limites de detecção, a avaliação da robustez frente à variabilidade experimental e a integração a plataformas automatizadas amigáveis.

Palavras-chave: FTIR; química forense; agrotóxicos; quimiometria; aprendizado de máquina.

ABSTRACT

In forensic chemistry laboratories, it is common to analyze materials contaminated with pesticides related to homicides or environmental crimes. Among the compounds involved, anticoagulant rodenticides and organophosphate and carbamate insecticides stand out. Classic identification techniques, such as gas and liquid chromatography coupled with mass spectrometry, although sensitive, are expensive, require laborious extractions, and make intensive use of solvents. In this context, this work proposes the development of a methodology based on Fourier Transform Infrared Spectroscopy (FTIR), associated with chemometric techniques, for the detection of pesticides in food. The proposal seeks to contribute to the field of forensic chemistry by evaluating the potential of a rapid, non-destructive, low-cost screening method that is aligned with the principles of green chemistry due to the reduction in the use of solvents. The samples included five food groups: bread, bologna, rice, ground beef, and a mixture containing different foods. Each matrix was contaminated with different pesticides (aldicarb, brodifacoum, bromadiolone, chlorpyrifos, methomyl, and terbufos) in varying concentrations (between 2 and 50 ppm). Spectral analyses were conducted in transmittance mode, with a resolution of 4 cm^{-1} and 24 scans per spectrum, in the range of 1800 to 400 cm^{-1}. Among the five food groups, bread and mortadella performed best and were therefore selected for the main chemometric analyses. The spectral averages of each group were used in principal component analysis (PCA), which allowed clear identification of the grouping patterns among the contaminated foods. Next, partial least squares discriminant analysis (PLS-DA) was applied to construct a classification model, which achieved 86% accuracy. The applicability of the model was evaluated with two real samples from the Paraná Forensic Police. The bread sample contaminated with brodifacoum was correctly classified; however, the mortadella sample contaminated with terbufos was identified as containing chlorpyrifos. Considering the limitations of the model trained with the PLS-DA algorithm, a complementary step was performed using the PyCaret library in Google Colab to compare the performance of 15 classification algorithms. The two best models were those trained with the Linear Discriminant Analysis (LDA) algorithm, with 81% accuracy, and Light Gradient Boosting Machine (LightGBM), with 67% accuracy. LDA correctly classified the bread sample contaminated with brodifacoum, but misclassified the mortadella contaminated with terbufos. On the other hand, LightGBM correctly identified the terbufos contamination in the mortadella and classified the bread as bromadiolone, an error in the compound, but consistent with the chemical class, due to the structural similarity between brodifacoum and bromadiolone. The results obtained should be understood as part of a proof of concept, serving as a basis for future stages of methodological refinement. Future prospects include hyperparameter adjustment, spectral base diversification, detection limit definition, robustness assessment against experimental variability, and integration with user-friendly automated platforms.

Keywords: FTIR; forensic chemistry; pesticides; chemometrics; machine learning.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

a.C	-	Antes de Cristo
Anvisa	-	Agência Nacional de Vigilância Sanitária
ATR	-	*Attenuated Total Reflectance*
AUC	-	*Area Under the Curve*
AutoML	-	*Automated Machine Learning*
CG-EM	-	Cromatografia Gasosa acoplada à Espectrometria de Maasas
DDT	-	1,1,1-tricloro-2,2-di (ρ-clorofenil) etano
DL_{50}	-	Dose letal 50%
ESR	-	Ressonância Eletrônica de Spin
EVD	-	*Eigenvalue Decomposition*
FT	-	*Fourier Transform*
FTIR	-	*Fourier Transform Infrared Spectroscopy*
FTIR-ATR	-	*Fourier Transform Infrared Spectroscopy with Attenuated Total Reflectance*
GLS-W	-	*Generalized Least Squares Weighting*
GC-MS	-	*Gas Chromatography – Mass Spectrometry*
HCA	-	*Hierarchical Cluster Analysis*
HPLC	-	*High-Performance Liquid Chromatography*
IA	-	Inteligência Artificial
ICP-MS	-	*Inductively Coupled Plasma Mass Spectrometry*
KNN	-	*k-Nearest Neighbour*
SIMCA	-	*Soft Independent Modelling of Class Analogy*
kg	-	quilograma
LC-MS	-	*Liquid Chromatography – Mass Spectrometry*
LC-MS/MS	-	Liquid Chromatography – Tandem Mass Spectrometry
MIR	-	*Mid Infrared*
MIR-FTIR	-	*Mid-Infrared Fourier Transform Infrared Spectroscopy*
ML	-	*Machine Learning*
NIR	-	*Near Infrared*
ROC	-	*Receiver Operating Characteristic*

PC	-	*Principal Component*
PCA	-	*Principal Component Analysis*
PCR	-	*Principal Component Regression*
PLS	-	*Partial Least Squares*
PLS-DA	-	*Partial Least Squares Discriminant Analysis*
QuERChERS	-	*Quick, Easy, Cheap, Effective, Rugged, and Safe*
RMN	-	Ressonância Magnética Nuclear
RNA	-	Redes Neurais Artificiais
SVM	-	*Support Vector Machine*
UV/VIS	-	Espectroscopia na Região do Ultravioleta e Visível

SUMÁRIO

NOTA DA AUTORA

Com o intuito de preservar a fidelidade conceitual e visual de determinados conteúdos, algumas figuras foram mantidas em seu idioma original. Da mesma forma, optou-se por conservar certas siglas em inglês ao longo do texto, considerando que esses termos estão amplamente consolidados na literatura científica internacional e são comumente utilizados em sua forma original, mesmo em publicações em língua portuguesa.

1 INTRODUÇÃO

Os agrotóxicos são substâncias químicas empregadas no controle de organismos considerados pragas, abrangendo animais, vegetais, fungos e microrganismos. Seu uso se estende a diversos setores, como a agricultura, a pecuária, a indústria, a medicina veterinária, a saúde pública e as campanhas sanitárias. Desenvolvidos para interferir em processos biológicos naturais, esses compostos possuem, portanto, propriedades tóxicas que podem representar sérios riscos à saúde humana e ao meio ambiente (PARANÁ, 2018).

A Lei Federal nº 15.070 de 23/12/2024, no seu Artigo 40º, traz a seguinte definição de agrotóxico:

> produtos e agentes de processos físicos ou químicos isolados ou em mistura com biológicos destinados ao uso nos setores de produção, no armazenamento e no beneficiamento de produtos agrícolas, nas pastagens ou na proteção de florestas plantadas, cuja finalidade seja alterar a composição da flora ou da fauna, a fim de preservá-las da ação danosa de seres vivos considerados nocivos (BRASIL, 2024).

As intoxicações por agrotóxicos em seres humanos podem ocorrer de forma intencional (tentativa de suicídio, homicídio ou abortamento), acidental (reutilização de embalagens, fácil acesso das crianças), ocupacional (no exercício da atividade laboral) e ambiental (contaminação da água, do ar e do solo) (PARANÁ, 2018).

Dados extraídos do Sistema de Gerenciamento de Documentos e Laudos da Polícia Científica do Paraná revelaram que, no período de 2022 a 2024, o laboratório de Química Forense foi requisitado por dezenas de ofícios para conduzir análises periciais em materiais potencialmente contaminados por agrotóxicos, abarcando casos relacionados a eventos de violência ou suicídio. Dentre essas amostras, 39% estavam inseridas em uma matriz alimentar, com predomínio notório de produtos cárneos (bovinos ou suínos) em 46% dos casos relacionados a alimentos sólidos. Em relação aos alimentos líquidos, o maior contingente de análises solicitadas (60%) estava associado à água, havendo também a inclusão, no rol de amostras passíveis de análise, do café pronto pra consumo e sucos (PCI-PR, 2025).

Entre os diferentes agentes agrotóxicos implicados nessas ocorrências, destacam-se os raticidas anticoagulantes, tendo como exemplos o brodifacoum e a bromadiolona, justificando, assim, uma das classes escolhidas para esse trabalho.

A inclusão dos compostos organofosforados e carbamatos neste estudo justifica-se, por sua vez, pela relevância que esses agentes químicos mantêm em casos de intoxicação intencional. O aldicarbe, por exemplo, foi proibido no Brasil em 2012 devido ao seu uso irregular e indiscriminado em tentativas de homicídio e suicídio, continuando, até os dias atuais, a causar vítimas. Em sua "Nota Técnica da Reavaliação do Ingrediente Ativo Aldicarbe", a Agência Nacional de Vigilância Sanitária (Anvisa) classificou o referido composto como um "grave problema de saúde pública, com alcance nacional, em virtude da facilidade de acesso, especialmente em áreas urbanas". O documento ainda relata que, somente no estado do Rio de Janeiro, aproximadamente cem pessoas morriam anualmente em decorrência de envenenamentos por aldicarbe. Entre os raticidas clandestinos, popularmente conhecidos como “chumbinho”, o aldicarbe - pertencente à classe dos carbamatos - figura como um dos princípios ativos mais frequentemente identificados (ANVISA, 2006). Ainda importa destacar que, em análises periciais realizadas na Polícia Científica do Paraná, também têm sido detectados outros compostos de relevância, como o metomil - igualmente classificado como carbamato -, o clorpirifós e o terbufós, ambos representantes da classe dos organofosforados. Essas substâncias, incluídas no escopo do presente estudo, demandam a mesma atenção e aprofundamento investigativo.

A análise direcionada a uma determinada classe de agrotóxicos torna-se necessária quando há indícios visuais evidentes no material analisado, como a presença de grânulos pretos ou acinzentados - típicos de compostos organofosforados e carbamatos - ou iscas coloridas, geralmente em tons de rosa, azul, roxo ou verde, comumente associadas aos raticidas anticoagulantes. Diversos métodos específicos têm sido descritos na literatura para a análise desses compostos em casos suspeitos envolvendo alimentos contaminados ou produtos agrícolas (ARMENTA *et al.,* 2005; FABUNMI, 2019). Entretanto, tais métodos não possibilitam a triagem simultânea das três principais classes de agrotóxicos. A adaptação metodológica a um analito específico pode acarretar impactos significativos na produtividade, na eficiência operacional e nos custos laboratoriais, uma vez que cada abordagem requer treinamento técnico específico, aquisição de

reagentes e insumos próprios, além da execução independente para cada tipo de substância (DI LORENZO; BUTT, 2022).

A identificação inequívoca dessas substâncias em matrizes alimentares, muitas vezes complexas e em avançado estado de decomposição, representa um desafio analítico significativo (SOUZA; RIBEIRO, 2020). Além do escopo forense, diversas técnicas analíticas têm sido descritas recentemente na literatura com o objetivo de detectar adulterações em bebidas à base de frutas, destacando-se a cromatografia líquida de alta eficiência (na sigla em inglês, HPLC - *High-Performance Liquid Chromatography*) (HAN *et al.*, 2012; ABAD-GARCÍA *et al.*, 2014; ASADPOOR; ANSARIN; NEMATI, 2014) e a cromatografia gasosa acoplada à espectrometria de massas (CG-EM) (NUNCIO-JÁUREGUI *et al.*, 2014). Apesar da elevada sensibilidade e seletividade, esses métodos são, em geral, caros, trabalhosos, demorados e dispendiosos, além de demandarem grande volume de solventes e gerarem uma quantidade considerável de resíduos químicos (NIESSEN, 2001; NIU *et al.*, 2021).

Neste cenário, a espectroscopiadeo infravermelho médio com transformada de Fourier (MIR-FTIR, na sigla em inglês), especialmente quando associada ao uso de acessórios de refletância total atenuada (ATR), surge como uma alternativa promissora para a detecção de agrotóxicos em alimentos contaminados. Trata-se de uma técnica rápida, não destrutiva, com baixo consumo de reagentes e capaz de fornecer informações químicas relevantes por meio da interação da radiação infravermelha com as moléculas das amostras analisadas (COZZOLINO, 2015). A simplicidade operacional e a possibilidade de obtenção direta de espectros de alimentos sólidos ou semissólidos tornam a espectroscopia MIR-FTIR particularmente atrativa em áreas forenses, nas quais a celeridade e a rastreabilidade dos dados são requisitos essenciais (BARTH, 2007).

A espectroscopia MIR-FTIR tem se consolidado como uma ferramenta analítica versátil na área forense, atendendo de forma cada vez mais abrangente às demandas investigativas. Sua aplicação tem sido demonstrada em uma ampla variedade de contextos periciais, incluindo a análise de amostras de cabelo humano masculino e feminino (THAKUR *et al.*, 2024), a identificação de substâncias em comprimidos de *ecstasy* (SOUZA; POPPI, 2012), a detecção de manchas de sangue sobre diferentes substratos (CAVALCANTE *et al.*, 2024), estudos entomotoxicológicos (WEI *et al.*, 2024), a identificação de cocaína (SILVA *et al.*,

2023), a análise de manchas de fluidos biológicos em superfícies diversas (JOHN *et al.*, 2023), a caracterização de produtos derivados de drogas ilícitas (CANO-TRUJILLO *et al.*, 2023), a identificação de drogas sintéticas em papéis apreendidos (DECONINCK *et al.*, 2022), a análise forense de batons (CUSTÓDIO et al., 2021) e a investigação de recortes de unhas com fins de classificação e predição de sexo (CHOPHI *et al.*, 2020).

Uma das características mais marcantes desses instrumentos é a capacidade de obter, a partir de uma única amostra, um grande volume de variáveis analíticas representadas pelos valores de transmitância (ou absorbância) registrados em cada número de onda. Atualmente, é possível registrar rotineiramente a intensidade de transmitância em centenas ou mesmo milhares de números de onda em um único espectro, o que proporciona uma base de dados extremamente rica para análises posteriores (SNYDER *et al.*, 2014; SHAH *et al.*, 2010). As principais vantagens das técnicas espectroscópicas residem justamente nessa capacidade de avaliação multivariada da composição de diferentes amostras, utilizando quantidades mínimas de material, de forma não destrutiva, simples, rápida e ambientalmente sustentável (FERNÁNDEZ-GONZÁLEZ *et al.*, 2014; KIRTIL *et al.*, 2017). Além disso, a possibilidade de detectar múltiplas substâncias por meio de um único procedimento analítico, como se propõe neste trabalho, representa um avanço na otimização da eficiência, da produtividade e da economia de recursos no contexto laboratorial. No âmbito da Polícia Científica do Paraná, essa abordagem tem o potencial de tornar o processo de triagem pericial mais ágil e econômico do ponto de vista técnico.

Com o desenvolvimento das tecnologias de aquisição de dados na química analítica, especialmente com a integração de instrumentos a sistemas computacionais, passou a ser possível gerar grandes volumes de informações complexas (HE *et al.*, 2007). No entanto, a complexidade espectral das amostras alimentares contaminadas, somada à frequente sobreposição de bandas características, impõe desafios significativos à análise direta dos dados, o que torna-se imprescindível o uso de abordagens quimiométricas avançadas para a correta interpretação dos espectros obtidos. Neste contexto, a quimiometria ocupa papel central ao viabilizar a construção de modelos preditivos e classificatórios, especialmente por meio da aplicação de algoritmos de aprendizado de máquina. Técnicas como análise de componentes principais (PCA), regressão por mínimos

quadrados parciais (PLS) e seus derivados, como a Análise Discriminante por Mínimos Quadrados Parciais (PLS-DA), têm sido amplamente utilizadas para discriminar grupos, detectar padrões ocultos e correlacionar espectros com identidade química (WOLD; SJÖSTRÖM; ERIKSSON, 2001; ESBENSEN; SANCHEZ, 2010).

A união entre espectroscopia e quimiometria representa uma abordagem estratégica para o desenvolvimento de métodos analíticos forenses. Ao combinar a capacidade de detecção não invasiva da espectroscopia com o potencial preditivo da inteligência artificial, é possível criar um método de triagem de agrotóxicos em alimentos capaz de responder com eficiência às demandas da perícia criminal.

Diante da ocorrência de eventos violentos envolvendo agrotóxicos no Estado do Paraná e das limitações inerentes às técnicas analíticas atualmente empregadas, que, além de onerosas, demandam procedimentos de extração laboriosos e utilizam grandes volumes de solventes, contrariando princípios da Química Verde, evidencia-se a necessidade de aprimoramento nos métodos de detecção e identificação dessas substâncias. Na rotina da Polícia Científica do Paraná, as amostras são submetidas a um protocolo de extração genérico, visando abranger o maior número possível de compostos. No entanto, essa abordagem ampla compromete a sensibilidade e a seletividade para determinados analitos de interesse forense, uma vez que não permite foco em grupos químicos específicos. Ademais, até o momento, não se dispõe de um método sistematizado de triagem capaz de fornecer respostas preliminares de forma ágil e direcionada.

1.1 OBJETIVOS

1.1.1 Objetivo Geral

Desenvolver e validar uma prova de conceito, com caráter exploratório, para demonstrar a viabilidade de um método analítico baseado em espectroscopia de infravermelho médio com Transformada de Fourier (MIR-FTIR), associada a técnicas quimiométricas, visando à triagem simultânea de seis agrotóxicos de interesse forense - aldicarbe, metomil, clorpirifós, terbufós, brodifacoum e bromadiolona - avaliados em diferentes concentrações (2, 5, 10, 30, 40 e 50 ppm) em cinco matrizes alimentares distintas: arroz branco, carne moída, pão francês, mortadela e uma mistura composta por arroz branco, feijão, carne suína, maionese e doce de leite.

1.1.2 Objetivos Específicos

1. Realizar análise exploratória dos dados espectrais para identificar padrões, tendências e possíveis agrupamentos entre as amostras.

2. Desenvolver e validar modelos preditivos, com base em algoritmos de aprendizado de máquina, para identificar os seis agrotóxicos selecionados.

3. Aplicar os modelos desenvolvidos na identificação de resíduos de agrotóxicos em amostras alimentares contaminadas, provenientes de casos reais analisados pela Seção de Química Forense da Polícia Científica do Paraná.

2 REVISÃO DE LITERATURA

2.1 ASPECTOS GERAIS DO USO DE AGROTÓXICOS

No decurso da civilização, tem sido comum o uso de substâncias químicas para o controle de pragas. Um exemplo disso é o enxofre, que os sumérios utilizavam como inseticidas desde 2500 a.C. No ano 400 a.C., o piretro, uma substância extraída da planta *Chrysanthemum cinerariaefolium*, passou a ser empregado no controle de piolhos. Na China, durante o século XIV, compostos à base de arsênio e mercúrio eram empregados para controlar diversos tipos de insetos e pragas (BRAIBANTE; ZAPPE, 2012). Com o avanço da agricultura, especialmente a partir do século XVI, substâncias orgânicas como a nicotina, extraída do tabaco, e a rotenona, proveniente de raízes de plantas como o timbó (*Derris* sp.), foram amplamente empregadas na Europa e nos Estados Unidos para o controle de pragas (BULL; HATHAWAY, 1986).

Um marco importante para o controle de pragas foi a descoberta, em 1939, da atividade inseticida do 1,1,1-tricloro-2,2-di (p-clorofenil) etano, conhecido como DDT. Esse inseticida foi utilizado pela primeira vez durante a Segunda Guerra Mundial para combater piolhos que infestavam as tropas norte-americanas na Europa. Entretanto, devido às características dos compostos organoclorados (alta solubilidade em líquidos apolares e, consequentemente, em óleos e gorduras, o que ocasiona o acúmulo do DDT no tecido adiposo dos organismos vivos, e alta estabilidade, pois demora muitos anos para ser degradado na natureza devido à baixa reatividade das ligações químicas presentes no composto em condições normais), houve a necessidade de serem desenvolvidos novos compostos com eficiência no controle de pragas. Assim, foram sintetizados os organofosforados e os carbamatos (BRAIBANTE; ZAPPE, 2012).

A disseminação generalizada do uso de agrotóxicos na agricultura teve início nos Estados Unidos na década de 1950, marcando o advento da chamada “Revolução Verde”, cujo objetivo era modernizar as práticas agrícolas e elevar os níveis de produtividade (SILVA *et al.*, 2005). No Brasil, esse movimento se consolidou na década de 1960 e ganhou impulso significativo ao longo dos anos

1970, com a implementação do Programa Nacional de Defensivos Agrícolas (PNDA). Tal programa vinculava a utilização dessas substâncias à concessão de créditos agrícolas, sendo o Estado um dos principais incentivadores dessa política (SIQUEIRA *et al*., 2013). Nesse contexto, os agroquímicos foram incorporados à agricultura brasileira com o propósito de prevenir e/ou erradicar pragas que pudessem comprometer o rendimento das lavouras (VEIGA, 2007).

Na primeira década do século XXI, o Brasil expandiu em 190% o mercado de agrotóxicos, o que colocou o país em primeiro lugar no *ranking* mundial de consumo desde 2008 (PAZ; REZENDE; GAMEIRO, 2023). Dentro desse cenário, a região Sul responde por aproximadamente 30% do total consumido no país (CREMONESE *et al.*, 2012).

Existem diversas categorias de agrotóxicos, delineadas com base nos padrões de aplicação e no tipo de praga que se almeja controlar ou erradicar. As principais classes compreendem inseticidas, herbicidas, fungicidas e raticidas. Ademais, a Agência de Proteção Ambiental dos Estados Unidos (EPA - *Environmental Protection Agency*) efetua uma categorização adicional dos pesticidas, classificando-os em compostos químicos (como organofosforados, carbamatos, organoclorados e piretroides), componentes biológicos (como microrganismos, protetores incorporados nas plantas por Engenharia Genética e feromônios) e instrumentos e dispositivos destinados ao controle de pragas (OGA; CAMARGO, BATISTUZZO, 2008).

Conforme estabelecido na Resolução nº 296, de 29 de julho de 2019, emitida pela Agência Nacional de Vigilância Sanitária (Anvisa), vinculada ao Ministério da Saúde (BRASIL, 2019), os agrotóxicos podem ser classificados com base na dose necessária para causar a morte de 50% de uma população de animais de laboratório submetida a testes (DL_{50}). As diferentes classes toxicológicas estão indicadas na TABELA 1, conforme diretrizes dessa resolução, e estão organizadas em seis categorias, variando de “extremamente tóxico” (classe 1) a “não classificado” (risco mínimo ou ausência de toxicidade aguda mensurável). Adicionalmente, todos os produtos devem apresentar, em seus rótulos, uma faixa colorida que identifique sua respectiva classe toxicológica. Essa sinalização tem como objetivo facilitar a identificação do grau de periculosidade por parte dos usuários, funcionando como uma importante ferramenta de comunicação de risco.

TABELA 1 - CLASSIFICAÇÃO TOXICOLÓGICA DOS AGROTÓXICOS DE ACORDO COM A TOXICIDADE AGUDA

Categoria	Toxicidade	Concentração Limite (dose oral em mg/kg)	Faixa
1	Extremamente Tóxico	≤5	Vermelha
2	Altamente Tóxico	>5-50	Vermelha
3	Moderadamente Tóxico	>50-300	Amarela
4	Pouco Tóxico	>300-2000	Azul
5	Improvável de Causar Dano Agudo	>2000-5000	Azul
	Não classificado	>5000	Verde

FONTE: A autora (2025).

Embora a classificação toxicológica seja definida com base em valores de DL_{50} obtidos em modelos animais, é fundamental compreender o significado prático desses números no contexto da exposição humana. Para isso, torna-se útil estabelecer aproximações que permitam relacionar os dados experimentais à realidade de um indivíduo adulto. A TABELA 2 a seguir apresenta uma correlação entre os valores de DL_{50} determinados em animais de laboratório e a quantidade estimada de produto necessária para causar a morte de um ser humano com peso médio de 70 kg. Tal estimativa, ainda que aproximada, auxilia na visualização da gravidade potencial associada à ingestão de diferentes substâncias (PARANÁ, 2018).

TABELA 2 - CORRELAÇÃO ENTRE DOSE LETAL EM ANIMAIS DE LABORATÓRIO E DOSE LETAL EM HUMANOS

DL_{50} Oral Animais de Laboratório	Dose Letal Provável em Humanos
1 mg/kg	Algumas Gotas
1 - 50 mg/kg	1 Colher de Chá
50 - 500 mg/kg	30 g ou 30 mL
0,5 g - 5 g/kg	500 g ou 500 mL
5 - 15 g/kg	1 kg ou 1 Litro
> 15 g/kg	> de 1 kg ou 1 Litro

FONTE: PARANÁ (2018).

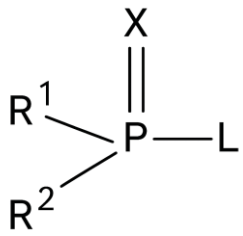
2.2 AGROTÓXICOS SELECIONADOS NESTE ESTUDO

2.2.1 Inseticidas Inibidores da Colinesterase

Em contextos forenses**,** a identificação de inseticidas inibidores da colinestease (como os da classe dos organofosforados e carbamatos) em amostras alimentares é frequentemente associada a intoxicações deliberadas (COUTO *et al.*, 2024). Essas substâncias, pelo seu potencial de causar morte rápida e sintomas neurológicos agudos, são frequentemente investigadas em casos de envenenamento suspeito, principalmente por ingestão de alimentos contaminados.

Os compostos organofosforados e os carbamatos são largamente utilizados no controle e no combate a pragas, principalmente como inseticidas (agrícola, doméstico e veterinário), desempenhando um papel no controle de parasitas em diversas áreas, tais como fruticultura, horticultura, cereais, tratamento de sementes e em paisagismo (OGA; CAMARGO; BATISTUZO, 2008). Os organofosforados são estruturalmente definidos como ésteres amido ou tiol-derivados dos ácidos fosfórico, fosfônico, fosforotioico e fosfonotioico, podendo ser representados pela fórmula geral apresentada na FIGURA 1.

FIGURA 1 - ESTRUTURA GERAL DOS INSETICIDAS ORGANOFOSFORADOS



FONTE: A autora (2025).

LEGENDA: **X** – Oxigênio (O) ou Enxofre (S)
R1 e R2 – Substituintes químicos menos reativos
L – Substituinte químico mais reativo

Essa classe de compostos também se destaca por suas características ambientais, sobretudo no que diz respeito à degradação e à persistência no solo. Os organofosforados são compostos biodegradáveis e, por essa razão, apresentam curta persistência no ambiente. A hidrólise em condições alcalinas constitui o principal mecanismo de degradação desses compostos no solo (JOKANOVIC, 2001). Eles são amplamente utilizados devido ao baixo custo, disponibilidade, ampla eficácia no controle de pragas e à capacidade de atuar contra diversas espécies (TANKIEWICZ; FENIK; BIZIUK, 2010).

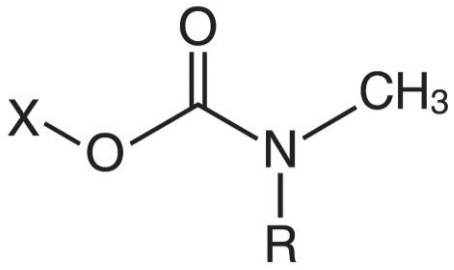
Do ponto de vista bioquímico, os organofosforados são suscetíveis à ação de diversas enzimas, que atuam em diferentes regiões da molécula. Dentre os processos metabólicos relevantes, destaca-se a dessulfuração oxidativa - isto é, a oxidação do grupamento fosforotioato (P=S) para fosfato (P=O), convertendo as formas tions em oxons -, reação que gera metabólitos com toxicidade acentuada para insetos e mamíferos. No entanto, esses análogos oxidados podem ser rapidamente hidrolisados por hidrolases presentes nos tecidos de mamíferos, o que contribui para sua desintoxicação. Os insetos, por outro lado, apresentam frequentemente deficiência dessas enzimas, o que os torna mais suscetíveis à ação tóxica desses compostos (OGA; CAMARGO; BATISTUZO, 2008).

De forma semelhante, os carbamatos constituem outra classe de compostos amplamente utilizada como inseticida e que também atua na inibição da acetilcolinesterase, embora com características químicas e históricas distintas. Seu desenvolvimento está associado ao uso tradicional da planta *Physostigma venenosum*, originária do oeste da Ásia e popularmente conhecida como feijão-de-calabar. Seu extrato aquoso era empregado em julgamentos de feitiçaria: a ingestão por indivíduos acusados de determinado crime resultaria em uma "prova" de sua culpabilidade se ele morresse, ou de sua inocência, caso sobrevivesse. Na metade do século XIX, foi isolado o composto responsável pelos efeitos medicinais e tóxicos dessa planta, o qual continha o grupo funcional carbamato (BRANCO, 2003 apud BRAIBANTE; ZAPPE, 2012).

Do ponto de vista estrutural, o grupo dos carbamatos é formado por derivados do ácido N-metil-carbâmico (FIGURA 2), bem como por tiocarbamatos e ditiocarbamatos, que diferem quanto ao mecanismo de ação, aplicação e toxicidade. Estes últimos, por não compartilharem o mesmo modo de ação inibitória sobre a acetilcolinesterase, apresentam usos distintos e devem ser considerados

separadamente. De modo geral, os carbamatos são compostos instáveis, cuja degradação pode ser influenciada por diversos fatores ambientais, como umidade, temperatura, luminosidade e volatilidade (BARBOSA, 2004 *apud* DONATO, 2012).

FIGURA 2 - ESTRUTURA GERAL DOS INSETICIDAS CARBAMATOS



FONTE: A autora (2025).

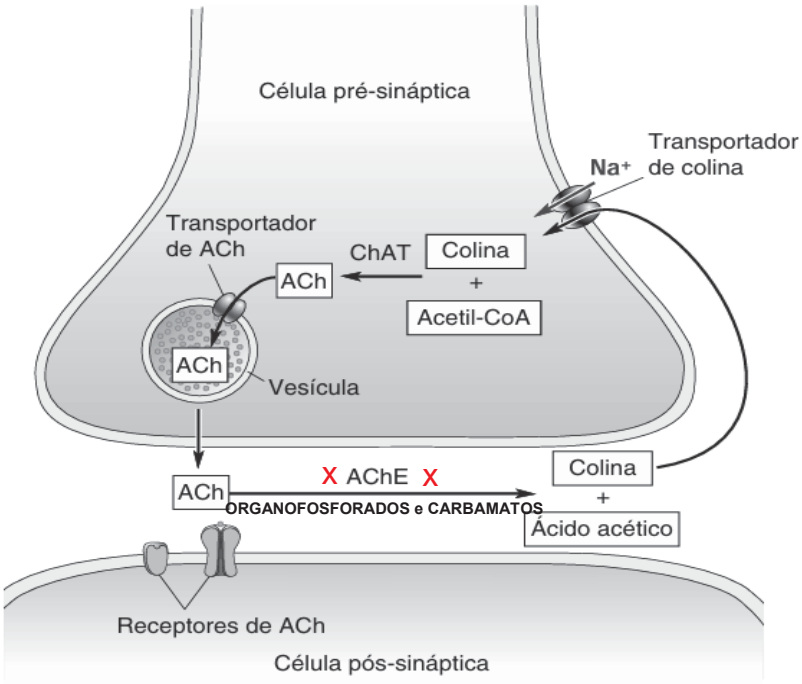
LEGENDA: **X** – Substituinte químico mais reativo (oxima, grupamento aromático)
R – Substituinte químicos menos reativo (H, CH_3)

Tanto os organofosforados quanto os carbamatos, a partir de suas estruturas químicas específicas, exercem sua toxicidade principalmente por meio da inibição das colinesterases, em especial da enzima acetilcolinesterase (AChE), responsável pela hidrólise da acetilcolina nas sinapses. Como consequência, ocorre o acúmulo desse neurotransmissor no espaço sináptico, intensificando os estímulos colinérgicos (OGA; CAMARGO; BATISTUZO, 2008).

A AChE possui dois sítios ativos - um sítio aniônico e um sítio esterásico - e a interação com os inibidores ocorre de maneira distinta dependendo da classe do composto. Os organofosforados se ligam exclusivamente ao sítio esterásico, onde o átomo de fósforo forma uma ligação covalente estável com a enzima, resultando na formação da enzima fosforilada. A hidrólise dessa ligação é lenta, podendo levar dias ou semanas, o que prolonga a inibição enzimática. Por outro lado, os carbamatos também inibem a AChE, mas o fazem de forma transitória. Inicialmente, ocorre a formação de um complexo reversível entre o carbamato e a enzima, seguido de uma carbamilação da AChE. No entanto, diferentemente dos organofosforados, a descarbamilação ocorre de forma relativamente rápida por

hidrólise, resultando na liberação da enzima funcional e de fragmentos inativos do carbamato (OGA; CAMARGO; BATISTUZO, 2008). A FIGURA 3 representa um desenho esquemático da inibição da AChE pelos inseticidas organofosforados e carbamatos.

FIGURA 3 - MECANISMO DE AÇÃO DOS ORGANOFOSFORADOS E CARBAMATOS



FONTE: Adaptado de BEAR; CONNORS; PARADISO (2017).

LEGENDA: Representação esquemática do mecanismo de ação dos inseticidas organofosforados e carbamatos na fenda sináptica. Esses compostos inibem a enzima acetilcoinesterase (AChE), impedindo a degradação de acetilcolina (ACh) em colina e ácido acético. Como resultado, ocorre acúmulo de ACh na fenda sináptica, com estimulação contínua dos receptores colinérgicos.

A consequência fisiológica dessa inibição prolongada ou transitória da acetilcolinesterase é o acúmulo de acetilcolina nas sinapses, levando a uma hiperestimulação colinérgica. Em mamíferos, o quadro clínico decorrente da ingestão de inibidores da AChE é caracterizado por manifestações muscarínicas, nicotínicas e centrais, com sintomas como lacrimejamento, salivação excessiva, sudorese, diarreia, tremores e alterações cardiorrespiratórias. Estas últimas incluem broncoconstrição, aumento das secreções brônquicas e bradicardia, além de depressão do sistema nervoso central, sendo essas complicações as principais

responsáveis pelos casos de morbidade e mortalidade associadas à exposição a tais compostos (ECOBICHON; JOY, 1991 *apud* CAVALIERI et al., 1996).

Sob a perspectiva analítica, a detecção de organofosforados e carbamatos em matrizes alimentares pode ser realizada por meio de técnicas de alta sensibilidade e seletividade, como cromatografia líquida ou gasosa acoplada à espectrometria de massas (LC-MS/MS, GC-MS). Estudos demonstram uso dessas metodologias na identificação simultânea de múltiplos resíduos de pesticidas em diferentes tipos de amostras, como frutas, vegetais, produtos de origem animal, água e fluidos biológicos (LIU *et al.*, 2005; KIM; YANG; CHOI, 2024; CHEN *et al.*, 2009). Além disso, abordagens com o método QuEChERS (*Quick, Easy, Cheap, Effective, Rugged, and Safe*) aliado à LC-MS/MS têm sido amplamente adotadas para análise de agrotóxicos em alimentos (MDENI *et al.*, 2022; TSAGKARIS *et al.*, 2020).

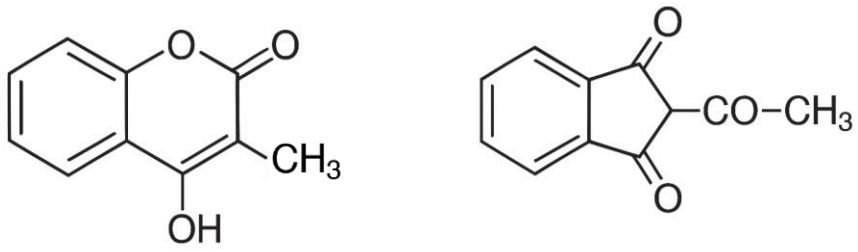
2.2.2 Raticidas Anticoagulantes

Os raticidas anticoagulantes constituem o principal grupo de agrotóxicos empregados no controle de roedores. Dentre eles, destacam-se os compostos hidroxicumarínicos, cuja descoberta remonta à década de 1940 e que, desde então, vêm sendo amplamente utilizados. Atualmente, esses agentes figuram entre os raticidas mais difundidos no mundo, sendo frequentemente associados a inúmeros casos de intoxicação, tanto em seres humanos quanto em animais (VALCHEV; YORDANOVA; NIKOLOV, 2008).

Nesse contexto, os raticidas têm aplicação diversificada, abrangendo os setores agropecuário, industrial, doméstico e campanhas de saúde pública. Usualmente, são disponibilizados sob as formas de grãos, iscas, *pellets* ou blocos, apresentando coloração variada, que pode incluir tonalidades como lilás, vermelho, laranja e verde-azulado (OGA; CAMARGO; BATISTUZO, 2008). No Brasil, a concentração máxima permitida de brodifacoum e bromadiolona é de 0,005%, em embalagens de até 20 g, conforme os índices monográficos B10 (brodifacoum) e B27 (bromadiolona), estabelecidos pela Anvisa (ANVISA, 2025).

Do ponto de vista químico, esses compostos são classificados em dois grupos principais: as hidroxicumarinas e as indandionas, sendo estas últimas representadas por substâncias como clorofacinona, difacinona, pindona e valona (FIGURA 4). Os derivados hidroxicumarínicos, por sua vez, subdividem-se em compostos de primeira geração - como coumacloro, coumafuril, coumatetralil e varfarina - e de segunda geração, que incluem brodifacoum, bromadiolona, difenacoum, difetialona e flocumafen (VALCHEV; YORDANOVA; NIKOLOV, 2008).

FIGURA 4 - ESTRUTURA QUÍMICA DOS RATICIDAS ANTICOAGULANTES



Raticidas hidroxicumarínicos Raticidas indandionas

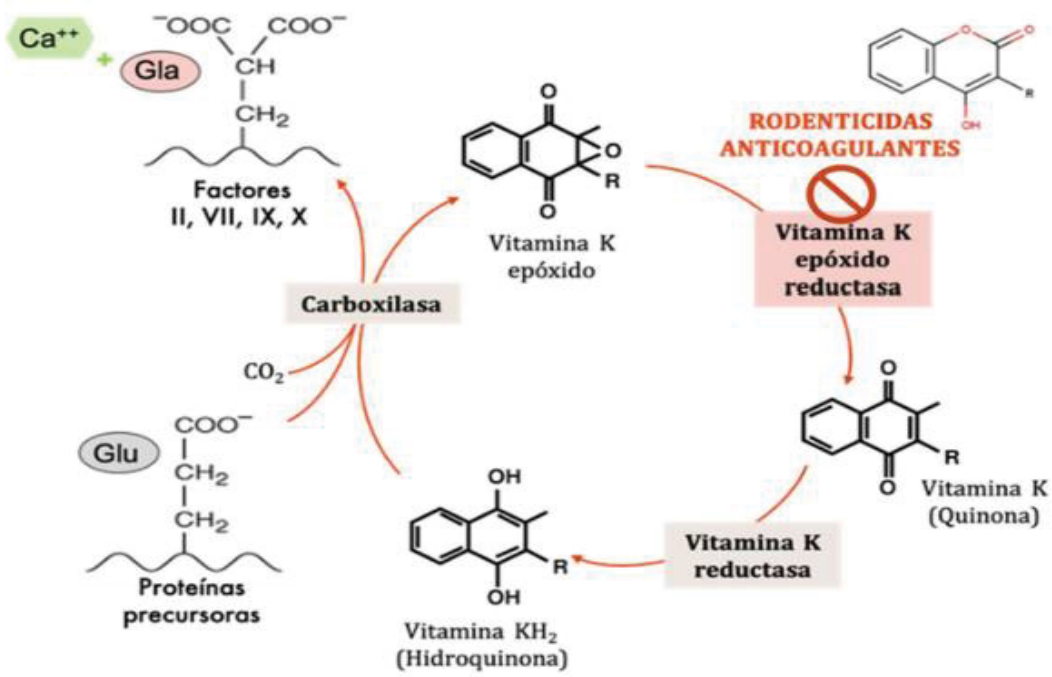
FONTE: A autora (2025).

Dentre os derivados hidroxicumarínicos de segunda geração, destaca-se o brodifacoum, amplamente empregado como raticida anticoagulante devido à sua elevada potência. Desenvolvido em 1977 pela empresa britânica Sorex e posteriormente comercializado pela *Imperial Chemicals Incorporated – Plant Protection Division*, o composto tornou-se notável por sua eficácia: à época de seu lançamento, era o único raticida anticoagulante capaz de provocar 100% de mortalidade na maioria das espécies de roedores com apenas uma única dose administrada em 24 horas (CHALERMCHAIKIT; FELICE; MURPHY, 1993).

As hidroxicumarinas e as indandionas exercem sua ação tóxica por meio da inibição das enzimas vitamina K epóxido-redutase e vitamina K redutase, fundamentais para o ciclo de regeneração da vitamina K no fígado (FIGURA 5). Como consequência, ocorre a depleção da forma ativa dessa vitamina, essencial para a ativação dos fatores de coagulação II, VII, IX e X. Esse bloqueio leva à redução progressiva da atividade desses fatores e, por fim, ao prolongamento do

tempo de protrombina. A manifestação do distúrbio hemorrágico, no entanto, só se inicia após a depuração dos fatores de coagulação previamente sintetizados e já circulantes no organismo, uma vez que a deficiência de vitamina K compromete apenas a ativação de novos fatores. O prolongamento do tempo de protrombina se torna clinicamente evidente quando a concentração plasmática desses fatores atinge cerca de 25% de seus níveis basais. As respectivas meias-vidas são: fator VII – de 4 a 7 horas, fator IX – 24 horas, fator X – de 36 a 48 horas e fator II – aproximadamente 50 horas (OGA; CAMARGO; BATISTUZO, 2008).

FIGURA 5 - MECANISMO DE AÇÃO DOS RATICIDAS ANTICOAGULANTES



FONTE: MODREGO (2022).

LEGENDA: Representação esquemática do ciclo da vitamina K no fígado, essencial para a ativação dos fatores de coagulação II, VII, IX e X. A enzima carboxilase catalisa a conversão de resíduos de ácido glutâmico (Glu) em ácido γ-carboxiglutâmico (Gla), processo dependente da vitamina K na forma reduzida (hidroquinona). Após participar da carboxilação, a vitamina K é convertida em sua forma oxidada (epóxido) e regenerada pelas enzimas vitamina K epóxido-redutase e vitamina K redutase. Os raticidas anticoagulantes atuam inibindo a epóxido-redutase, bloqueando a regeneração da vitamina K e, consequentemente, interrompendo a ativação dos fatores de coagulação.

Em seres humanos, as manifestações clínicas decorrentes da intoxicação por raticidas anticoagulantes podem variar consideravelmente, refletindo a

complexidade da resposta individual ao comprometimento do ciclo da vitamina K. Enquanto alguns pacientes desenvolvem coagulopatia com sangramentos ativos, outros apresentam apenas alterações laboratoriais assintomáticas ou sequer manifestam sinais de desordem hemostática. Fatores fisiológicos intrínsecos, como hipoalbuminemia e níveis elevados de alanina aminotransferase, mostram-se significativamente associados ao risco de coagulopatia após a ingestão desses compostos. Além disso, elementos externos, como o consumo concomitante de álcool ou paracetamol, podem potencializar os efeitos tóxicos dos raticidas. A duração da atividade anticoagulante também varia de acordo com o princípio ativo envolvido: estima-se que, após uma única exposição, os efeitos perdurem por aproximadamente 21 dias no caso da bromadiolona e até 30 dias para o brodifacoum (STEENSMA *et al.*, 1994; TAM; CHAN; LIU, 2021).

Os sinais clínicos costumam surgir, em geral, entre 24 e 48 horas após a ingestão, período necessário para que ocorra a depleção dos fatores de coagulação circulantes. Em casos de ingestão maciça, entretanto, os sintomas podem manifestar-se em até 12 horas. O quadro clínico é marcado por sinais de sangramentos espontâneos, como gengivorragia, equimoses e hematomas — especialmente em áreas de apoio como joelhos, cotovelos e nádegas —, além de hemorragia subconjuntival, hematúria acompanhada de dor lombar, epistaxe, sangramentos vaginais e digestivos. Em paralelo, surgem sintomas de anemia secundária, como fadiga, palidez e dispneia. Nos casos mais graves, podem ocorrer hemorragias cavitárias, incluindo sangramento intracraniano ou intra-abdominal, com risco de evolução para choque hipovolêmico e óbito (OGA; CAMARGO; BATISTUZO, 2008).

A gravidade do quadro clínico está diretamente relacionada à quantidade de princípio ativo ingerido. Em humanos, a menor dose fatal de brodifacoum documentada na literatura varia entre 0,12 e 0,172 mg/kg. Considerando um indivíduo adulto com peso médio de 70 kg, essa faixa corresponde a uma dose total aproximada de 8,4 mg (PATOCKA; PETROIANU; KUCA, 2013; TAM; CHAN; LIU, 2021). No Brasil, os raticidas à base de brodifacoum ou bromadiolona são usualmente comercializados em embalagens de 20 gramas, com concentração de 0,005% do princípio ativo, o que equivale a cerca de 1,0 mg por embalagem. Assim, para que um adulto saudável atinja a dose tóxica mínima de brodifacoum, seria necessário ingerir o conteúdo de, aproximadamente, sete dessas embalagens.

2.3 ESPECTROSCOPIA DE INFRAVERMELHO

A espectroscopia é a ciência que estuda a interação da radiação eletromagnética com a matéria, possibilitando uma análise detalhada da estrutura e da composição de compostos químicos. Dependendo da energia da radiação incidente, essa interação pode resultar em diversos fenômenos, como reflexão, espalhamento, fluorescência, reações fotoquímicas ou absorção (AMAYA, 1999). Por sua versatilidade e precisão na caracterização de substâncias, essa técnica é amplamente utilizada em áreas como a química de alimentos e as ciências forenses. Entre suas principais vantagens, destacam-se a possibilidade de analisar pequenas quantidades de amostra ou de seus extratos de forma não destrutiva, rápida, direta e com mínima ou nenhuma preparação. Além disso, permite a identificação simultânea de muitos compostos em uma única análise (ZHANG *et al.*, 2011; ESSLINGER; RIEDL; FAUHL-HASSEK, 2013; FERNÁNDEZ-GONZÁLEZ *et al.*, 2014; KIRTIL *et al.*, 2017).

Entre as diversas regiões do espectro eletromagnético, destaca-se a região do infravermelho, que compreende comprimentos de onda entre aproximadamente 2,5 μm e 25 μm. A maioria dos compostos orgânicos e inorgânicos que apresentam ligações covalentes absorvem radiação nessa faixa, o que está relacionado a vibrações moleculares específicas de seus grupos funcionais. A compreensão dessas interações baseia-se em princípios físicos fundamentais, como a relação inversa entre a frequência (ν) e o comprimento de onda (λ), expressa pela equação ν = c/λ, onde c representa a velocidade da luz. Além disso, a energia da radiação eletromagnética está diretamente associada à sua frequência, de acordo com a equação 1 abaixo.

$$E = h\upsilon \tag{1}$$

Em que:
h é a constante de Planck.

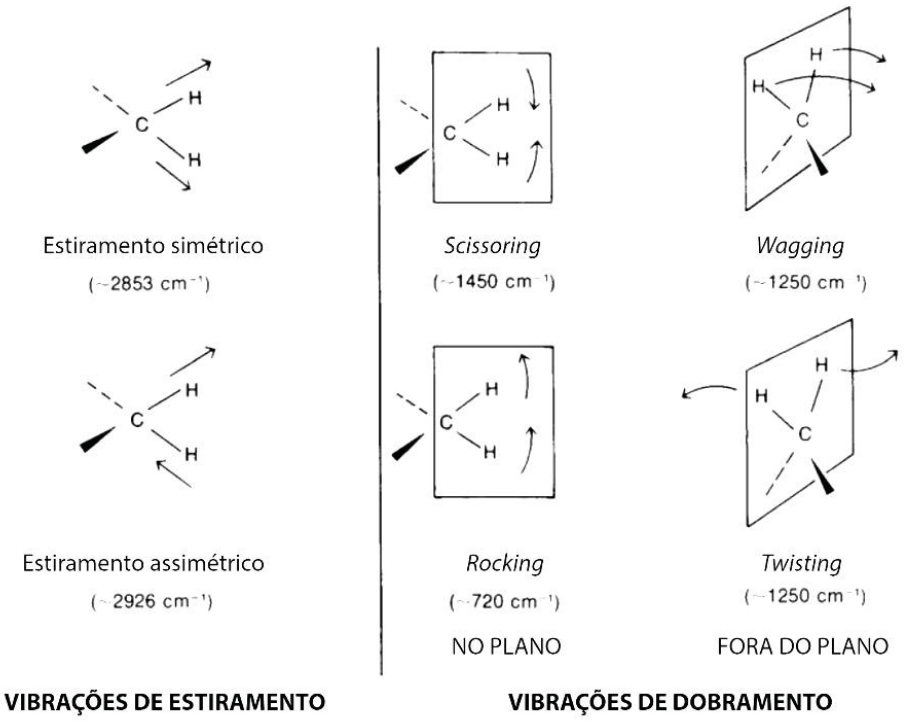
Assim, quanto maior a frequência, maior a energia da radiação. Essa relação explica por que radiações de alta energia, como os raios X, são capazes de romper ligações químicas, enquanto radiações de menor energia, como as de radiofrequência, são utilizadas para promover transições de spin em técnicas como Ressonância Magnética Nuclear (RMN) e Ressonância Eletrônica de Spin (ESR) (PAVIA *et al.*, 2015).

Dentre as diversas regiões do espectro eletromagnético, a região do infravermelho ocupa papel central na caracterização molecular, especialmente pela sua capacidade de revelar informações estruturais por meio das vibrações atômicas. Essa região pode ser subdividida em três faixas distintas: o infravermelho distante, de 400 a 10 cm^{-1}, o infravermelho próximo (*Near Infrared* – NIR), que abrange radiações com números de onda entre 15.000 e 4.000 cm^{-1}, e o infravermelho médio (*Mid Infrared* – MIR), que se estende de 4.000 a 400 cm^{-1}. No infravermelho médio ocorrem as chamadas transições fundamentais, nas quais a molécula passa do estado vibracional fundamental para o primeiro estado excitado. Essa característica confere à espectroscopia no infravermelho médio uma elevada sensibilidade na detecção de grupos funcionais, uma vez que cada tipo de ligação química exibe uma banda de absorção característica em frequência específica (AMAYA, 1999). Por essa razão, essa técnica é amplamente empregada na identificação e elucidação estrutural de compostos orgânicos.

Quando uma molécula absorve radiação na região do infravermelho, ocorrem transições vibracionais e, em alguns casos, também rotacionais. Esse processo só se verifica para determinadas frequências da radiação incidente, as quais devem coincidir com as frequências vibracionais naturais da molécula. A energia absorvida resulta no aumento da amplitude dos movimentos vibracionais das ligações covalentes, como os modos de estiramento *(stretching*) e de flexão (*bending*). No entanto, nem todas as ligações presentes em uma molécula são capazes de absorver radiação infravermelha, mesmo quando a frequência da radiação coincide exatamente com a frequência vibracional do grupo funcional. Para que haja absorção, é necessário que ocorra uma variação do momento de dipolo durante a vibração. Apenas as ligações que apresentam um dipolo elétrico oscilante ao longo do tempo podem interagir com a radiação infravermelha. Ligações simétricas, como a ligação Cl–Cl, por exemplo, não exibem variação de dipolo e, por isso, são transparentes à radiação nessa região do espectro (PAVIA *et al.*, 2015).

Os modos de estiramento envolvem variações no comprimento da ligação entre dois átomos, podendo ocorrer de forma simétrica, quando ambos os átomos se afastam ou se aproximam simultaneamente do centro molecular, ou de forma assimétrica, quando um átomo se aproxima enquanto o outro se afasta. Por outro lado, os modos de dobramento referem-se a alterações no ângulo entre ligações químicas, sendo geralmente observados em frequências mais baixas que as dos estiramentos. Esses movimentos incluem subtipos realizados no plano molecular, como os modos de tesoura e balanço, e fora do plano, como os movimentos de abano e torção (FIGURA 6). Cada tipo de vibração fornece informações específicas sobre a geometria e o ambiente químico dos grupos funcionais presentes na molécula. A análise combinada dos modos de estiramento e de dobramento é essencial para a identificação de grupos funcionais e para a caracterização estrutural de compostos por espectroscopia no infravermelho, pois permite reconhecer padrões vibracionais característicos associados a diferentes tipos de ligações e geometrias moleculares (COATES, 2006; PAVIA *et al.*, 2015).

FIGURA 6 - MODOS VIBRACIONAIS MOLECULARES

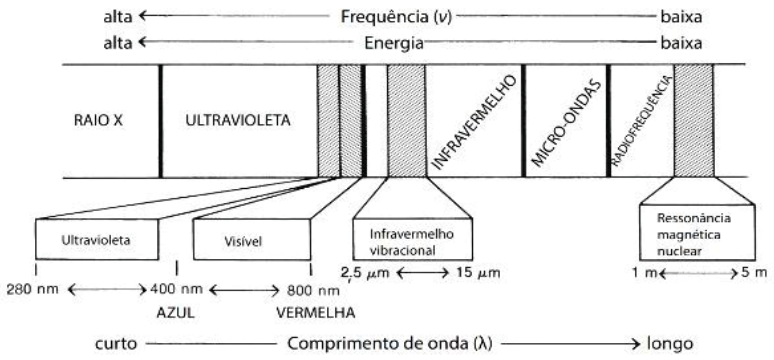


FONTE: PAVIA *et al.* (2015).

2.3.1 Espectrômetro de Infravermelho com Transformada de Fourier

O espectro de absorção no infravermelho de um composto é obtido por meio de um espectrômetro de infravermelho. Existem dois tipos principais de equipamentos: os dispersivos e os de transformada de Fourier (FT – *Fourier Transform*). Ambos fornecem espectros na faixa de 4000 a 400 cm^{-1}, mas os espectrômetros FT destacam-se pela maior rapidez na aquisição dos dados e melhor relação sinal-ruído em comparação aos sistemas dispersivos. Esses equipamentos identificam as posições e intensidades das bandas de absorção e as registram graficamente, originando o espectro infravermelho do composto. A absorção de radiação infravermelha é uma propriedade comum à maioria dos compostos que apresentam ligações covalentes, ocorrendo em frequências específicas relacionadas às vibrações moleculares típicas de cada ligação química. Por isso, a região vibracional do infravermelho é amplamente empregada na caracterização química, dada sua sensibilidade às variações estruturais e funcionais das moléculas. Essa faixa do espectro situa-se entre a luz visível (400 - 800 nm) e as micro-ondas (>1 mm) e sua descrição baseia-se na relação inversa entre o comprimento de onda (λ) e a frequência (ν), dada por ν = c/λ, onde c representa a velocidade da luz (FIGURA 7). Como a energia da radiação é diretamente proporcional à frequência, segundo a Equação 1 já apresentada nesta dissertação, números de onda mais altos correspondem a radiações de maior energia (PAVIA *et al.*, 2015).

FIGURA 7 - ESPECTRO ELETROMAGNÉTICO: RELAÇÃO ENTRE FREQUÊNCIA, ENERGIA E COMPRIMENTO DE ONDA



FONTE: PAVIA *et al.* (2015).

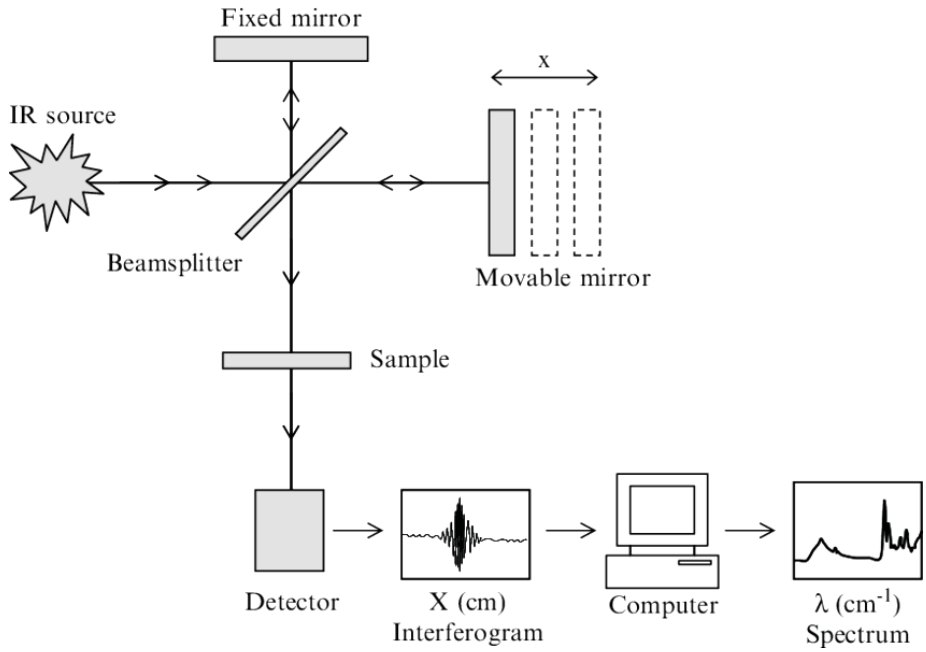
Os espectrômetros de infravermelho modernos, como os de Transformada de Fourier, geram, após o caminho óptico, um interferograma. Esse interferograma apresenta um padrão em ondas que inclui todas as frequências que compõem o espectro infravermelho. O interferograma é um gráfico que relaciona a intensidade com o tempo (um espectro no domínio temporal). No entanto, os químicos preferem visualizar o espectro como um gráfico de intensidade *versus* frequência (um espectro no domínio da frequência). A transformada de Fourier é a operação matemática utilizada para separar as frequências das absorções presentes no interferograma. O principal benefício do espectrômetro FTIR é sua capacidade de produzir um interferograma em menos de um segundo, permitindo a coleta de vários interferogramas da mesma amostra, que são armazenados no computador. Ao aplicar a transformada de Fourier sobre esses interferogramas acumulados, é possível obter um espectro com uma melhor razão sinal/ruído (PAVIA *et al*, 2015).

O componente óptico central em um espectrômetro FTIR é o interferômetro de Michelson, responsável pela modulação da radiação infravermelha antes que ela alcance a amostra. Esse sistema permite que todas as frequências do espectro sejam processadas simultaneamente, por meio da criação de padrões de interferência gerados pelo deslocamento controlado de um dos espelhos. A partir dessas interferências, forma-se o interferograma, que contém informações sobre todas as frequências absorvidas pela amostra. O uso do interferômetro de Michelson é o que diferencia fundamentalmente os espectrômetros de transformada de Fourier dos modelos dispersivos tradicionais, sendo crucial para a eficiência, sensibilidade e rapidez na obtenção dos espectros (GRIFFITHS, HASETH, 2007).

O FTIR, como ilustrado na FIGURA 8, usa um interferômetro para manipular a energia enviada à amostra. No interferômetro, a energia da fonte atravessa um divisor de feixes, um espelho posicionado em um ângulo de 45º em relação à radiação que entra, separando-a em dois feixes perpendiculares: um segue na direção original e o outro é desviado por um ângulo de 90º. Um feixe, desviado por 90º, vai para um espelho estacionário, ou "fixo", e é refletido de volta para o divisor de feixes. O feixe que não sofreu desvio vai para um espelho que se move e também é refletido para o divisor de feixes. O movimento do espelho faz variar a trajetória do segundo feixe. Quando os dois feixes se encontram no divisor de feixes, eles se recombinam, mas as diferenças de caminhos dos dois feixes causam interferências. O feixe combinado contém esses padrões de interferência e dá

origem ao interferograma, o qual contém toda a energia da radiação que veio da fonte (PAVIA *et al.*, 2015).

FIGURA 8 - DIAGRAMA ESQUEMÁTICO DE ESPECTRÔMETRO DE INFRAVERMELHO COM TRANSFORMADA DE FOURIER



FONTE: PAVIA *et al.* (2015).

LEGENDA: A radiação infravermelha é emitida por uma fonte (IR *source*) e direcionada a um divisor de feixe (*beamsplitter*), que separa a luz em dois caminhos: um segue para um espelho fixo (*fixed mirror*) e outro para um espelho móvel (*movable mirror*). O espelho móvel desloca-se para frente e para trás, gerando diferenças no caminho óptico dos feixes, o que resulta em padrões de interferência. Após serem refletidos, os feixes são recombinados no *beamsplitter* e direcionados à amostra (*sample*). Parte da radiação é absorvida pela amostra, de acordo com os grupos funcionais presentes. A radiação transmitida é então detectada (*detector*) e registrada como um interferograma (*interferogram*), que representa as variações de intensidade da luz em função do tempo. Esse sinal bruto é processado por um computador (*computer*), que aplica a Transformada de Fourier para convertê-lo em um espectro de absorção (*spectrum*), o qual revela as faixas de número de onda (cm^{-1}) em que ocorrem absorções características das ligações químicas presentes na amostra. NOTA DO AUTOR: Para preservar a fidelidade conceitual e visual do diagrama original, optou-se por mater os termos técnicos em inglês.

Resumidamente, antes de se obter o espectro de uma amostra por espectroscopia no infravermelho com transformada de Fourier, é necessário que o analista registre inicialmente um interferograma de “fundo”. Esse fundo corresponde à absorção de componentes atmosféricos presentes no ambiente, como dióxido de carbono e vapor d’água, que exibem bandas na região do infravermelho. O

interferograma de fundo é então submetido à transformada de Fourier, gerando um espectro que representa essas absorções indesejadas. Em seguida, a amostra é posicionada no caminho óptico, e um novo interferograma é registrado. A aplicação da transformada de Fourier sobre esse novo sinal resulta em um espectro que contém tanto as bandas do fundo quanto as da amostra. A subtração do espectro de fundo permite, então, isolar as informações espectrais da amostra propriamente dita (PAVIA *et al.*, 2015).

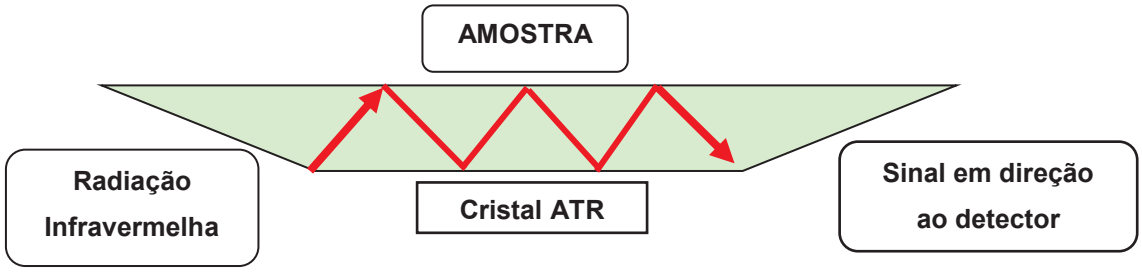
Nesse contexto, o uso de espectrômetros baseados em Transformada de Fourier, aliados a acessórios como a refletância total atenuada (FTIR-ATR – *Fourier Transform Infrared Spectroscopy with Attenuated Total Reflectance*), revolucionou a análise de amostras sólidas e líquidas. Essa abordagem moderna simplifica significativamente o preparo das amostras e melhora a reprodutibilidade dos espectros, superando as limitações dos métodos tradicionais de transmissão e absorção, que são geralmente mais dispendiosos e menos eficientes (SOUZA, 2009).

Para compreender melhor os fatores que influenciam a qualidade do espectro obtido, é essencial entender o princípio físico por trás da técnica ATR. Na espectroscopia ATR, a amostra sólida ou líquida é posicionada sobre um cristal opticamente denso, com alto índice de refração, como o diamante, por exemplo. A radiação infravermelha que incide sobre esse cristal é refletida internamente em sua superfície de contato com a amostra. Nessa reflexão interna total, o feixe penetra levemente na amostra por meio de uma onda evanescente, cuja profundidade típica varia entre 0,5 μm e 5,0 μm. Essa onda perde energia nos comprimentos de onda em que a amostra apresenta absorção, produzindo assim um espectro infravermelho de superfície (PAVIA *et al.*, 2015). A intensidade da radiação transmitida é atenuada em função das múltiplas reflexões ao longo do cristal, caracterizando o fenômeno de refletância total atenuada (FIGURA 9).

Apesar das vantagens, o uso do elemento ATR pode apresentar limitações práticas, especialmente no que diz respeito à reprodutibilidade das medições. Um dos problemas mais comuns é a variação na intensidade das bandas em função da pressão aplicada sobre a amostra. Quanto maior a pressão, melhor tende a ser o contato entre a amostra e o cristal, o que resulta em bandas mais intensas no espectro. Além disso, a área de contato entre a superfície da amostra e o cristal é outro fator crítico: para que os resultados sejam consistentes, é fundamental que

toda a superfície do cristal esteja em contato com a amostra. No entanto, irregularidades da amostra podem dificultar esse contato ideal, comprometendo a uniformidade e a qualidade espectral dos dados obtidos (COLEMAN, 1993; MIRABELA, 1985).

FIGURA 9 - REPRESENTAÇÃO ESQUEMÁTICA DO ACESSÓRIO DE REFLETÂNCIA TOTAL ATENUADA (ATR)



FONTE: A autora (2024).

A incorporação de acessórios ATR ampliou ainda mais as possibilidades analíticas da técnica, especialmente quando associada a métodos quimiométricos, permitindo o desenvolvimento de modelos quantitativos e classificatórios robustos, sem a necessidade de separações físicas ou químicas dos componentes interferentes (KAROUI; PIERNA; DUFOUR, 2008). Nesse sentido, o uso da quimiometria tem se mostrado essencial para a extração eficiente das informações contidas nos espectros gerados.

Com base nesse princípio, a espectroscopia de infravermelho torna-se uma ferramenta poderosa para a caracterização molecular, uma vez que cada grupo funcional exibe bandas de absorção em regiões específicas do espectro, de acordo com seus modos vibracionais. Dessa forma, o espectro obtido funciona como uma verdadeira impressão digital da substância analisada (FERNÁNDEZ-GONZÁLEZ *et al.*, 2014).

2.3.2 Aplicações da espectroscopia de infravermelho médio na detecção de agrotóxicos em alimentos e outras análises forenses

A espectroscopia FTIR tem se consolidado como uma técnica analítica poderosa em diferentes contextos forenses, destacando-se por sua capacidade de fornecer informações moleculares detalhadas de forma rápida e não destrutiva. Entre suas principais vantagens, ressalta-se a agilidade na obtenção dos resultados, especialmente quando comparada a métodos tradicionais como cromatografia gasosa e líquida acopladas à espectrometria de massas (GC-MS e LC-MS, nas siglas em inglês) ou espectrometria de massas com plasma indutivamente acoplado (ICP-MS, na sigla em inglês). Essa técnica apresenta-se, assim, como uma alternativa mais econômica e eficiente para análises de triagem, com o benefício adicional de preservar as amostras para exames confirmatórios posteriores.

Entre as diversas aplicações dessa técnica, destaca-se a sua utilização na detecção de resíduos de agrotóxicos em alimentos, aspecto amplamente explorado na literatura científica. Estudos demonstram que, quando associada a métodos quimiométricos multivariados - como Análise de Componentes Principais (PCA, na sigla em inglês), Análise Discriminante por Mínimos Quadrados Parciais (PLS-DA, na sigla em inglês) ou Máquinas de Vetores de Suporte (SVM, na sigla em inglês) - a espectroscopia FTIR tem apresentado excelente desempenho na diferenciação entre amostras contaminadas e não contaminadas, bem como na classificação de acordo com os níveis de concentração dos agrotóxicos (SANTOS *et al*., 2020).

Diversos estudos têm demonstrado a eficácia da espectroscopia MIR-FTIR na detecção de traços de agrotóxicos em frutas, hortaliças, cereais, óleos e produtos processados. Em um estudo conduzido por Steidle Neto *et al.* (2024), a aplicação da espectroscopia MIR-FTIR em conjunto com modelo utilizando algoritmo PLS-DA possibilitou a detecção de fungicidas em alfaces. Resultados semelhantes foram obtidos por Xiao *et al.* (2015), que evidenciaram a viabilidade da técnica na identificação de resíduos de clorpirifós em amostras de maçãs.

Complementando essas abordagens voltadas à matriz alimentar, Armenta *et al.* (2005) desenvolveram uma metodologia validada de FTIR para a determinação rápida de pesticidas, como cipermetrina e clorpirifós, em formulações comerciais. Utilizando extração com clorofórmio e medição direta por FTIR, o método reduziu

significativamente o uso de solventes e aumentou a eficiência analítica. Em um estudo subsequente, os mesmos autores ampliaram essa aplicação para a determinação de Diuron.

Além das aplicações convencionais voltadas ao controle de qualidade e à análise de formulações, a espectroscopia FTIR também tem despertado interesse no campo da ciência forense - foco central deste trabalho. Em situações de envenenamento intencional por agrotóxicos, como as que motivam investigações periciais, a técnica pode ser empregada na análise de amostras alimentares suspeitas, com o objetivo de identificar padrões espectrais compatíveis com a presença de compostos tóxicos. Essa abordagem se destaca pela rapidez e praticidade, sendo particularmente útil como etapa de triagem preliminar antes da aplicação de métodos confirmatórios, como a cromatografia acoplada à espectrometria de massas (GC-MS ou LC-MS) (FERNÁNDEZ-GONZÁLEZ *et al.*, 2014). Portanto, a espectroscopia MIR-FTIR tem se consolidado como uma ferramenta versátil, aplicável tanto em contextos de segurança alimentar quanto em análises forenses. Sua natureza acessível, limpa e alinhada aos princípios da química verde, somada à capacidade de fornecer resultados rápidos e reprodutíveis, reforça seu potencial como método de triagem e como suporte analítico em investigações envolvendo contaminações intencionais.

Nesse mesmo contexto forense, a espectroscopia FTIR tem sido utilizada em investigações de adulteração de produtos de consumo, como bebidas alcoólicas. Um caso notório ocorreu em 2021, quando a contaminação de cervejas por monoetilenoglicol levantou suspeitas sobre envenenamento acidental, levando a internações e mortes na cidade de Belo Horizonte - MG. O estudo de Fulgêncio *et al.* (2022) demonstrou que a técnica FTIR-ATR, aliada à quimiometria, permitiu discriminar com eficácia amostras adulteradas com monoetilenoglicol. Assim como nos casos de contaminação alimentar deliberada, essa abordagem mostrou-se eficaz como método de triagem rápida.

Além disso, técnicas quimiométricas, como a PCA e a PLS-DA têm sido amplamente utilizadas para identificar substâncias tóxicas em amostras biológicas. Estudos recentes, como o de Silva *et al.* (2023), evidenciaram o uso eficaz dessas técnicas para detectar pesticidas organofosforados em larvas de moscas em cadáveres, auxiliando na determinação da causa da morte em casos de envenenamento. Em paralelo, Wei *et al.* (2024) aplicaram com sucesso o ATR-FTIR

em conjunto com algoritmos de aprendizado de máquina para diferenciar manchas de sangue humano de animal, uma aplicação inovadora que avança as investigações forenses ao proporcionar uma metodologia rápida e não destrutiva para a identificação de fluidos corporais em diferentes substratos.

Além da toxicologia forense, a espectroscopia FTIR-ATR tem se mostrado uma ferramenta promissora também na documentoscopia, especialmente na diferenciação de tintas utilizadas em canetas esferográficas e impressoras, bem como na determinação da ordem cronológica das escritas. No estudo conduzido por Farid *et al.* (2021), a combinação do FTIR-ATR com técnicas de análise de imagem e quimiometria demonstrou ser eficaz na identificação de adulterações em documentos suspeitos.

A versatilidade da espectroscopia FTIR também se estende à análise de vestígios automotivos, outra área de grande relevância na ciência forense. A técnica tem sido aplicada também na identificação de fragmentos de tinta em batidas de carro em cenas de crime. Em um estudo conduzido por Duarte *et al.* (2022), a espectroscopia FTIR foi utilizada para determinar a composição de amostras de tinta automotiva recolhidas diretamente em cenas de acidentes de trânsito, com o objetivo de discriminar os revestimentos conforme a marca do veículo e a cor.

Além dos vestígios veiculares, a aplicação do FTIR na ciência forense tem se expandido para outras matrizes de interesse investigativo. O estudo de Cox *et al.* (2000) demonstrou sua eficácia na discriminação de amostras de solo com base em suas composições orgânicas, sendo particularmente útil em investigações realizadas em ambientes externos. De forma semelhante, a análise de drogas sintéticas impregnadas em papéis foi conduzida com sucesso por Custodio (2021), utilizando FTIR combinado com PLS-DA, resultando em um método rápido, limpo e não destrutivo para o controle forense de entorpecentes. Complementando essa abordagem, Materazzi *et al.* (2017) empregaram FTIR-ATR para a criação de perfis preditivos de amostras de cocaína adulterada, estabelecendo um modelo eficiente para fins de classificação e comparação em contextos periciais.

Esses achados reforçam que a espectroscopia FTIR apresenta-se como uma ferramenta poderosa e versátil na ciência forense. Sua aplicação abrange desde a identificação e triagem de substâncias até a diferenciação de amostras em diversas matrizes, destacando-se por ser uma abordagem econômica, eficiente e minimamente invasiva.

Diante desse panorama promissor, é eviden mais relevante para a química forense moderna é a integração da espectroscopia FTIR com abordagens baseadas em inteligência artificial e aprendizado de máquina. Essa combinação tem potencializado a capacidade analítica da técnica, especialmente no tratamento de grandes volumes de dados espectrais e na realização de classificações complexas de forma automatizada. A junção de espectros ricos em informação com algoritmos computacionais tem ampliado a confiabilidade das interpretações periciais, reduzindo a subjetividade na análise e permitindo a construção de modelos preditivos capazes de revelar padrões ocultos em extensas bases de dados (BASAK *et al.*, 2022). Esse será, portanto, o próximo tema abordado na contextualização desta dissertação.

2.4 QUIMIOMETRIA

De acordo com Ferreira (2015a), autora do livro “Quimiometria: Conceitos, Métodos e Aplicações”, existem várias definições de quimiometria, mas há um consenso de que essa ciência se situa na interseção de três áreas principais: Química, Matemática e Estatística. A autora também destaca tanto a natureza interdisciplinar da quimiometria, que tem aplicações importantes na química medicinal e computacional, como no estudo de novos fármacos, além de ser amplamente utilizada nas indústrias farmacêutica, química e de alimentos para controle de qualidade e tratamento de imagens.

Complementando essa perspectiva, Kowalski (1975) define a quimiometria como a disciplina química que recorre a métodos matemáticos, estatísticos e a outras abordagens baseadas em lógica formal para planejar ou selecionar procedimentos de medição ideais, além de extrair o máximo de informação química relevante a partir da análise de dados experimentais. Em síntese, trata-se do conjunto de métodos apropriados para transformar grandes volumes de dados em informações significativas do ponto de vista químico. Essa integração entre a química e ferramentas quantitativas torna a quimiometria uma aliada indispensável na interpretação eficiente de dados complexos, sobretudo em contextos analíticos, industriais e de pesquisa científica.

A moderna instrumentação de análises químicas é capaz de gerar uma quantidade considerável de dados sobre uma única amostra, em um curto espaço de tempo: um espectrômetro pode registrar sinais provenientes de mais de mil números de onda ou um único cromatograma pode apresentar mais de cem picos (ROBINSON, 2001; VOET; VOET; PRATT, 2000). Assim, para que informação útil seja obtida deste grande volume de dados, é necessário que se utilizem técnicas matemáticas adequadas, sendo a quimiometria um dos campos de estudo da química que fornece tais ferramentas (ROBINSON, 2001; LAJOLO; NUTTI, 2003).

A quimiometria pode ser dividida em quatro principais vertentes: análise exploratória de dados químicos, utilizando técnicas como HCA (*Hierarchical Cluster Analysis*) e PCA (*Principal Component Analysis*); calibração multivariada, com PCR (*Principal Component Regression*) e PLS (*Partial Least Squares*); modelos de classificação, aplicando ferramentas como KNN (*k-Nearest Neighbour*), SIMCA (*Soft Independent Modelling of Class Analogy*) e PLS-DA (*Partial Least Squares for Discriminant Analysis*); e planejamento fatorial para otimizar condições experimentais (FERREIRA *et al.*, 1999).

As vertentes quimiométricas mencionadas são amplamente empregadas na identificação de padrões entre grupos de amostras com o objetivo de classificá-las, e podem ser agrupadas em dois tipos principais: métodos supervisionados e não supervisionados. Nos métodos supervisionados - como PLS - parte-se do pressuposto de que os dados já pertencem a classes previamente definidas, sendo essa informação utilizada na construção e validação dos modelos de classificação. Por outro lado, os métodos não supervisionados - como PCA - não requerem conhecimento prévio sobre as classes dos dados, promovendo o agrupamento com base apenas nas características intrínsecas extraídas dos próprios dados experimentais (FERREIRA, 2015a).

Dentre as técnicas quimiométricas citadas, as de calibração multivariada possibilitam a realização de determinações quantitativas de compostos presentes na ordem de porcentagem ou até de micro-constituintes, em matrizes complexas. Outras vantagens de sua aplicação são a redução considerável da necessidade de preparo das amostras, bem como da utilização de reagentes químicos por vezes nocivos ao homem ou ao meio ambiente, e a possibilidade da utilização de métodos não destrutivos (ROBINSON, 2001; VOET; VOET; PRATT, 2000).

Inicialmente, a maioria dos trabalhos da literatura utilizando quimiometria tratava dados obtidos em química analítica instrumental. A maioria das aplicações envolvia técnicas como espectroscopia de absorção no UV/VIS, espectroscopia no infravermelho e cromatografia. A partir dos anos 1990, a quimiometria foi se expandindo para outras áreas de aplicação, como biologia e medicina, juntamente com aplicações de outras técnicas analíticas, como espectrometria de massas e ressonância magnética nuclear. O uso da quimiometria em análises de alimentos para detecção de adulterações e fraudes e também em aplicações forenses utilizando informações químicas e espectroscópicas para determinar, por exemplo, a origem de amostras, também vem crescendo significativamente, como já introduzido na seção anterior. A interface da quimiometria com a bioinformática é mais recente, estando relacionada a dados biológicos, havendo, também, aplicações na área de metabolômica (BRERETON, 2007).

Na química de alimentos, a quimiometria tem se consolidado como uma ferramenta indispensável para a análise de dados complexos, sobretudo na interpretação de espectros obtidos a partir de amostras alimentícias. O emprego dessas técnicas está alinhado à crescente demanda por métodos analíticos sustentáveis - os chamados métodos verdes -, que fornecem respostas rápidas sem a necessidade de reagentes químicos ou solventes e baixa geração de resíduos, reduzindo significativamente o impacto ambiental. Um produto alimentar exposto à radiação do infravermelho apresentará um espectro característico que será, essencialmente, o resultado da absorção por vários componentes químicos. Por conseguinte, haverá um perfil típico de espectros para cada alimento. Embora simples em termos de conceito, esta comparação não é fácil, sendo necessárias técnicas quimiométricas para a sua realização (MANLEY; DOWNEY; BAETEN, 2008).

Dando continuidade a essa perspectiva, serão detalhadas a seguir as abordagens que fundamentam a quimiometria clássica adotada neste trabalho: a PCA, uma técnica não supervisionada voltada à redução da dimensionalidade e à exploração de padrões em dados complexos; e a PLS-DA, uma técnica supervisionada com foco na construção de modelos preditivos baseados em classes previamente definidas. Ambas as metodologias se mostram particularmente eficazes na análise de dados espectrais obtidos por técnicas de infravermelho. Complementarmente, este estudo também contempla algoritmos modernos de

classificação, igualmente supervisionados, que integram os avanços do aprendizado de máquina (*machine learning*) à espectroscopia, ampliando o potencial preditivo e interpretativo da análise quimiométrica.

2.4.1 Análise de Componentes Principais (PCA)

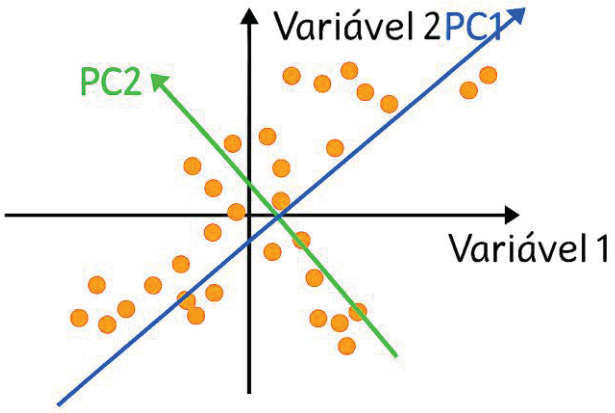
A PCA é um método matemático de análise exploratória aplicado a dados multivariados, com o objetivo de reduzir a dimensionalidade do conjunto original, preservando o máximo possível da variabilidade presente nos dados e mantendo inalteradas as relações entre as amostras (FERREIRA, 2015b). Essa redução permite simplificar a estrutura dos dados, facilitando sua interpretação, ao mesmo tempo em que evidencia amostras com comportamento atípico. Com a projeção em um novo espaço de menor dimensão, essas amostras tendem a se destacar visualmente.

Tal projeção permite revelar, de maneira mais clara, padrões ocultos, tendências e agrupamentos presentes nos dados. Além disso, os gráficos resultantes - especialmente os gráficos de escores - constituem uma ferramenta poderosa para a visualização das similaridades, diferenças e para a identificação de amostras anômalas no conjunto analisado (JOLLIFFE; CADIMA, 2016).

Um método de projeção como a PCA reduz o número de variáveis envolvidas - assim como uma sombra reduz a forma tridimensional de um objeto para um plano bidimensional. Essa compressão de dados permite representar as informações essenciais com menor complexidade, sem perda significativa de conteúdo. A PCA realiza essa compressão com base na correlação entre variáveis, identificando componentes que carregam, de forma concentrada, a maior parte da variabilidade dos dados. Isso é especialmente útil em conjuntos provenientes de espectroscopia ou cromatografia, com múltiplas variáveis altamente correlacionadas. Nesse contexto, é possível reduzir o número de variáveis reais mantendo grande porcentagem da informação relevante, por meio de combinações lineares que agrupam variáveis com conteúdo análogo (WOLD; ESBENSEN; GELADI, 1987; BRO; SMILD, 2014).

Na PCA, essas novas variáveis são denominadas componentes principais (PCs – *Principal Components*). Elas são matematicamente definidas como combinações lineares das variáveis originais e possuem duas propriedades fundamentais: são ortogonais (isto é, não correlacionadas entre si) e são ordenadas de acordo com a variância explicada. A primeira componente (PC1) carrega a maior variância dos dados; a segunda (PC2) carrega a segunda maior, mas em um subespaço ortogonal à primeira; e assim sucessivamente (FIGURA 10). Esse ordenamento garante que a informação redundante seja eliminada, mantendo-se apenas as dimensões mais informativas (WOLD; ESBENSEN; GELADI, 1987; BRO; SMILD, 2014).

FIGURA 10 - REPRESENTAÇÃO GRÁFICA DA ANÁLISE DE COMPONENTES PRINCIPAIS (PCA)



FONTE: A autora (2025).

LEGENDA: Distribuição bidimensional de dados simulados ilustrando a aplicação da PCA. As setas em azul (PC1) e verde (PC2) representam os eixos das componentes principais, obtidas por meio de combinações lineares das variáveis originais (Variável 1 e Variável 2). A PC1, orientada na direção de maior variância, capta a maior parte da informação dos dados, enquanto a PC2, ortogonal à primeira, representa a variância remanescente.

Matematicamente, a PCA corresponde à decomposição da matriz **X**, de dimensão *n* (amostras) × *m* (variáveis), em três matrizes menores: **T**, $\mathbf{P}^t$ e **E**, sendo **T** a matriz de *scores*, $\mathbf{P}^t$ a matriz de *loadings* (com “t” representando a transposta), e **E** a matriz de resíduos não explicados pelo modelo (FERREIRA, 2015).

Os escores representam as coordenadas das amostras no novo sistema de eixos definidos pelas componentes principais. Cada componente principal é constituída por uma combinação linear das variáveis originais, cujos coeficientes são chamados de pesos (*loadings*). Esses coeficientes podem ser interpretados como os cossenos dos ângulos entre as variáveis originais e os novos eixos (PCs), indicando o grau de contribuição de cada variável para a formação da respectiva componente. Assim, a PC1 é construída no sentido da maior variação nos dados, a PC2 é ortogonal à PC1 e capta a maior parte da variação não explicada pela primeira, e assim por diante. A análise simultânea de escores e pesos permite entender quais variáveis são responsáveis pelas diferenças observadas entre as amostras e os agrupamentos formados (WOLD; ESBENSEN; GELADI, 1987). Dessa forma, a PCA configura-se como uma ferramenta poderosa para a investigação exploratória de dados multivariados, promovendo a redução dimensional com preservação da variância e revelando padrões significativos ocultos nos conjuntos de dados.

2.4.2 Análise Discriminante por Mínimos Quadrados Parciais (PLS-DA)

A PLS-DA é uma variação da técnica de regressão multivariada PLS adaptada para tarefas de classificação supervisionada. Seu princípio consiste em estabelecer uma relação linear entre os dados espectrais (matriz **X**) e a classe à qual a amostra pertence (variável ou matriz **Y**), permitindo a discriminação entre grupos de amostras com base em seus perfis químicos. Essa técnica é amplamente utilizada na identificação de resíduos de pesticidas e outras substâncias químicas em matrizes alimentares e em análises forenses (BARKER; RAYENS, 2003).

O objetivo do desenvolvimento de modelos utilizando algoritmos PLS-DA é semelhante ao da Análise Discriminante Linear (LDA). Segundo Barker e Rayens (2003), a PLS-DA representa, essencialmente, o inverso da abordagem de mínimos quadrados aplicada ao LDA, produzindo resultados equivalentes, mas com maior robustez frente a colinearidades e ruídos nos dados (WISE *et al.*, 2006). Essa característica o torna especialmente útil em aplicações com dados altamente multivariados e complexos, como os provenientes de espectroscopia ou espectrometria de massa.

Na PLS-DA, a variável dependente (y) assume valores binários ou categóricos, em que o valor 1 indica que a amostra pertence à classe, e 0 indica que não pertence (BARKER; RAYENS, 2003). O modelo retorna valores contínuos entre 0 e 1, que representam a probabilidade de a amostra pertencer à classe considerada. Como os valores previstos raramente são exatamente 0 ou 1, utiliza-se um limite de decisão (*threshold*) calculado com base em estatística bayesiana, assumindo que os valores de y seguem uma distribuição normal. Esse limite é definido de modo a minimizar os erros de classificação — falsos positivos e falsos negativos (BYLESJÖ *et al.*, 2006; PULIDO *et al.*, 2003).

Em situações que envolvem mais de duas classes, essa estratégia binária não é suficiente, pois atribuir valores inteiros (como 1, 2, 3, 4...) às classes introduziria uma falsa relação de continuidade entre elas. A solução consiste em aplicar uma codificação *one-hot*, na qual um vetor coluna é criado para cada classe. Nesse caso, pode-se empregar o algoritmo PLS1 (ajustando um modelo para cada classe individualmente) ou PLS2 (modelando todas as classes simultaneamente em uma matriz de respostas). O PLS2-DA fornece, para cada amostra, uma predição vetorial do tipo *Nx1*, com valores entre 0 e 1 para cada classe. A classificação final é feita com base na maior dessas probabilidades, de forma análoga ao caso binário (BARKER; RAYENS, 2003).

Resumidamente, existem duas variantes da técnica PLS-DA: a PLS1-DA e a PLS2-DA. Na abordagem PLS1-DA, cada coluna da matriz de resposta **Y** é modelada individualmente. Assim, no caso de três classes — “a”, “b” e “c” — são construídos três modelos distintos. No primeiro, os valores de Y são codificados como 1 para a classe “a” e 0 para as demais; no segundo, 1 para a classe “b” e 0 para as outras; e o mesmo procedimento é repetido para a classe “c”. Já na PLS2-DA, constrói-se um único modelo multivariado em que os escores e pesos latentes são calculados de forma conjunta para todas as colunas de **Y**, implicando na necessidade de utilizar o mesmo número de variáveis latentes para modelar simultaneamente todas as classes (SANTANA *et al.*, 2020).

Do ponto de vista metodológico, o algoritmo PLS-DA é considerado uma das ferramentas padrão para análise de dados quimiométricos provenientes de técnicas como infravermelho, espectrometria de massas e RMN (GROMSKI *et al.*, 2015; MENDEZ; REINKE; BROADHURST, 2019). Sua popularidade deve-se à capacidade de lidar com grandes conjuntos de dados com muitas variáveis correlacionadas,

projetando os dados originais em um subespaço de menor dimensão composto por variáveis latentes. Essas variáveis são ortogonais entre si, o que resolve problemas de colinearidade e ainda contribui para a identificação das variáveis mais relevantes para a separação entre grupos (GROMSKI *et al.*, 2015).

Entretanto, algumas advertências são importantes. A validação do modelo é essencial e frequentemente negligenciada, o que pode levar a interpretações enganosas. Além disso, há risco de sobreajuste, especialmente quando o número de componentes utilizados é excessivo (GROMSKI *et al.*, 2015). Brereton e Lloyd (2014) também destacam que, embora o PLS-DA seja amplamente utilizado, sua aplicação como método de classificação exige cautela metodológica. O modelo pode produzir resultados convincentes mesmo em situações em que a separação entre grupos não é real, principalmente quando interpretado sem garantia de sua capacidade preditiva. Por outro lado, seu potencial exploratório é significativo, especialmente para identificar e visualizar variáveis discriminantes, como metabólitos ou bandas espectrais relevantes, auxiliando na geração de hipóteses e na compreensão de processos químicos ou biológicos. Dessa forma, o modelo PLS-DA configura-se como uma ferramenta poderosa e interpretável, especialmente útil em análises espectroscópicas, desde que seja utilizada com critérios estatísticos rigorosos, validações apropriadas e consciência de suas limitações.

A próxima seção abordará outros algoritmos classificatórios supervisionados, comumente utilizados em análises multivariadas e aprendizado de máquina, permitindo a comparação de desempenho e robustez em relação ao PLS-DA.

2.5 INTELIGÊNCIA ARTIFICAL: FUNDAMENTOS E OUTROS ALGORITMOS DE APRENDIZADO DE MÁQUINA

Estamos vivenciando uma nova revolução industrial, impulsionada pelo desenvolvimento de tecnologias avançadas como a Inteligência Artificial (IA). As máquinas não se limitam mais a realizar trabalhos manuais, mas também desempenham tarefas racionais, tradicionalmente associadas à inteligência humana (LUDERMIR, 2021). A IA, campo da ciência da computação, concentra-se no desenvolvimento de sistemas capazes de aprender, raciocinar, perceber, tomar

decisões e resolver problemas de maneira semelhante aos seres humanos (LINARDATOS *et al.*, 2021).

Algumas técnicas de inteligência artificial (IA) foram propostas ainda na década de 1950. No entanto, a limitação de dados disponíveis e do poder computacional impedia que essas técnicas fossem aplicadas em larga escala. Atualmente, com o advento de unidades de processamento gráfico (GPU), maior capacidade de processamento e a disponibilidade massiva de dados, tornou-se possível treinar algoritmos de aprendizado de máquina em problemas muito mais complexos. Assim, técnicas já conhecidas ganharam protagonismo, beneficiando-se do aumento da capacidade computacional e do acesso a grandes bases de dados (LUDERMIR, 2021).

É importante destacar que o desejo de fazer com que máquinas aprendam não é recente. Alan Turing, considerado o pai da computação, já em 1950 propôs o conhecido Teste de Turing, cujo objetivo era avaliar se um computador poderia demonstrar comportamento inteligente indistinguível do humano (TURING, 1950). Embora atualmente muitos sistemas de IA sejam capazes de superar esse teste em tarefas específicas, sua capacidade de aprendizado ainda não se equipara à dos seres humanos. Isso se deve, em parte, ao fato de que ainda não compreendemos integralmente como ocorre o processo de aprendizado humano. Mesmo assim, algoritmos atuais já conseguem “aprender” determinadas tarefas com eficiência, fornecendo resultados significativos em contextos específicos (LUDERMIR, 2021).

As técnicas de *machine larning* (ML), ou aprendizado de máquina, permitem que o computador aprenda por exemplos, ou seja, aprenda por meio dos dados. Dentro desse universo, o ML tem ganhado protagonismo por sua habilidade de identificar padrões complexos e fazer predições a partir de grandes volumes de dados, sendo amplamente aplicado em áreas como saúde, segurança, agricultura e ciências forenses. Esse crescimento é impulsionado tanto pelo aprimoramento dos algoritmos quanto pela crescente disponibilidade de bases de dados acessíveis e pelos avanços na computação de alto desempenho e baixo custo (RUSSELL; NORVIG, 2010; ALPAYDIN, 2016; BEAM; KOHANE, 2018; JORDAN; MITCHELL, 2015; LECUN; BENGIO; HINTON, 2015).

O ML pode ser dividido em três categorias principais: supervisionado, não supervisionado e por reforço. No aprendizado supervisionado, atualmente a abordagem mais utilizada, cada exemplo apresentado ao algoritmo deve estar

associado à resposta correta - ou seja, a um rótulo que identifique a classe à qual o exemplo pertence. Em um problema de classificação de imagens, por exemplo, seria necessário indicar se cada imagem corresponde a um gato ou a um cachorro. Cada amostra é descrita por um vetor de atributos e pelo rótulo associado, permitindo que o algoritmo construa um modelo capaz de classificar corretamente novos exemplos ainda não rotulados. Quando os rótulos correspondem a categorias discretas, trata-se de um problema de classificação; quando são valores contínuos, caracteriza-se como um problema de regressão (LUDEMIR, 2021). Em geral, para esses casos, os dados são divididos em treino e teste, comumente na proporção de 70% para treinamento e 30% para teste, embora essa divisão possa variar conforme o tamanho do conjunto e a complexidade do problema (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Entre os principais algoritmos de aprendizado supervisionado estão *Support Vector Machines* (SVM), *K-Nearest Neighbors* (KNN), Árvores de Decisão, Redes Neurais Artificiais (RNA), *Logistic Regression*, entre outros. Esses algoritmos atuam sobre dados rotulados para inferir padrões e classificar novas instâncias. O processo pode ser comparado ao de um aluno guiado por um professor: o algoritmo recebe exemplos com respostas conhecidas, ajusta-se com base nos erros cometidos e passa a classificar corretamente novas amostras (FERNEDA, 2006).

Para além da divisão inicial dos dados, é fundamental adotar estratégias que assegurem uma avaliação mais confiável do modelo supervisionado. Nesse contexto, a validação cruzada (*cross-validation*) é uma técnica amplamente empregada para estimar o desempenho do modelo. A validação cruzada *k-fold* consiste em dividir o conjunto de dados em k partes (valores típicos incluem 5, 10 ou até mesmo 20). Em cada uma das k iterações, uma dobra é usada como teste e as demais como treinamento. O modelo é treinado e avaliado em cada rodada, e ao final, calcula-se a média das métricas obtidas. Esse processo fornece uma estimativa mais confiável do desempenho do modelo, reduzindo o risco de sobreajuste (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

No aprendizado não supervisionado, os exemplos são apresentados ao algoritmo sem rótulos previamente definidos. Nessa abordagem, o sistema busca identificar padrões ocultos nos dados, agrupando-os de acordo com suas similaridades. Como resultado, são formados agrupamentos, ou *clusters*, que posteriormente precisam ser interpretados para que se compreenda o que cada

grupo representa dentro do contexto analisado (LUDERMIR, 2021). O objetivo, neste caso, é identificar padrões, estruturas e relações ocultas entre os dados (BARTNECK *et al.*, 2021). Exemplos incluem algoritmos de agrupamento (*clustering*), detecção de anomalias e redução de dimensionalidade.

Por fim, o aprendizado por reforço baseia-se na interação contínua entre um agente e seu ambiente, no qual as ações tomadas geram recompensas ou penalidades, guiando o aprendizado pela maximização de um objetivo cumulativo (GÉRON, 2022). O algoritmo não recebe a resposta correta para cada situação, mas sim um sinal de avaliação, que pode assumir a forma de recompensa ou punição. A partir dessa interação, o sistema formula hipóteses, testa suas ações e ajusta seu comportamento conforme os resultados obtidos. Esse tipo de aprendizado é amplamente empregado em áreas como jogos e robótica (LUDERMIR, 2021).

É importante destacar que a aplicação de aprendizado de máquina à resolução de problemas envolve uma série de pré-requisitos. O primeiro deles é a disponibilidade de um conjunto de exemplos representativo e de boa qualidade, o que muitas vezes exige a construção e a atualização constante das bases de dados. Além disso, como os dados coletados nem sempre são ideais, torna-se necessário aplicar técnicas de pré-processamento que melhorem sua qualidade. Outro aspecto crucial é a escolha dos algoritmos adequados ao tipo de problema a ser resolvido, já que não existe uma solução única para todos os casos. Uma vez selecionados os algoritmos, é preciso ajustar seus parâmetros - como o número de camadas em uma rede neural - e avaliar, após o treinamento, se o modelo está efetivamente solucionando o problema e com que nível de precisão. Por fim, sistemas de aprendizado de máquina devem ser continuamente atualizados, uma vez que mudanças nos dados podem comprometer seu desempenho ao longo do tempo (LUDERMIR, 2021).

Apesar de suas vantagens, o aprendizado de máquina apresenta desafios, especialmente quanto à capacidade de generalização. Os principais problemas envolvem tanto o *overfitting* (ajuste excessivo aos dados de treino) quanto o *underfitting* (incapacidade de capturar os padrões). Outros fatores incluem amostras desbalanceadas, dados de teste distintos dos dados de treino, seleção inadequada de variáveis, presença de *outliers* e escolha inadequada de hiperparâmetros ou algoritmos. Por isso, mesmo com bom desempenho inicial, é fundamental monitorar

continuamente os modelos com novos dados, garantindo a manutenção da performance preditiva (GOODFELLOW; BENGIO; COURVILLE, 2016).

A crescente demanda por modelos inteligentes levou ao desenvolvimento de ferramentas automatizadas, como o *AutoML* (*Automated Machine Learning*). Essa abordagem permite testar diversos algoritmos e ajustar hiperparâmetros automaticamente, economizando tempo e conhecimento técnico. Dentre as ferramentas mais populares destaca-se a PyCaret, biblioteca que possibilita, com poucas linhas de código aberto em *Python*, treinar múltiplos modelos e compará-los por meio de métricas de desempenho e matrizes de confusão (VELOSO, 2023; WHIG *et al.*, 2023).

PyCaret compara o desempenho de diferentes algoritmos de classificação supervisionada, cujas siglas e respectivas denominações completas são descritas a seguir: ET (*Extra Trees*), Light GBM (*Light Gradient Boosting Machine*), RF (*Random Forest*), GBC (*Gradient Boosting Classifier*), DT (*Decision Trees*), QDA (*Quadratic Discriminant Analysis*), KNN (*K-Nearest Neighbors*), LR (*Logistic Regression*), LDA (*Linear Discriminant Analysis*), NB (*Naive Bayes*) e SVM (*Support Vector Machine*), Ada (*AdaBoos*t), Dummy (classificador base de referência) e Ridge (*Ridge Classifier*).

O *design* e a simplicidade da PyCaret são influenciados pelo surgimento dos "cientistas de dados cidadãos". Esses são usuários avançados capazes de realizar tarefas analíticas que, anteriormente, exigiriam um conhecimento técnico mais aprofundado. Em essência, PyCaret busca democratizar o acesso ao *machine learning*, facilitando a vida tanto de especialistas quanto de usuários menos técnicos (ARAÚJO, 2022).

3 MATERIAL E MÉTODOS

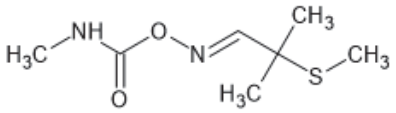
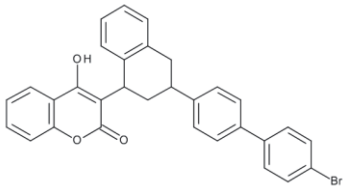
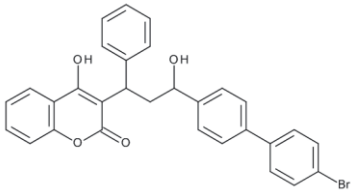
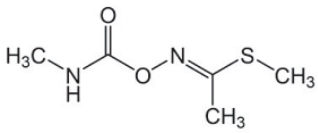
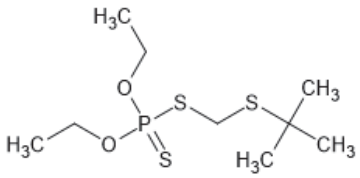
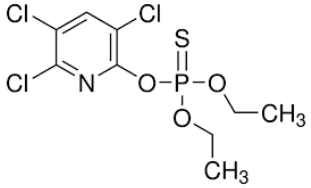
3.1 PREPARO DAS SOLUÇÕES ESTOQUE DE PADRÕES ANALÍTICOS

A primeira etapa deste estudo consistiu na seleção, separação e preparo dos padrões analíticos a serem utilizados nas análises. Todos os padrões foram obtidos a partir da coleção de substâncias de referência da Seção de Química Forense e da Seção de Toxicologia Forense da Polícia Científica do Paraná, devidamente catalogados e mantidos sob condições controladas de temperatura, conforme protocolos institucionais de conservação.

Os padrões analíticos utilizados compreenderam seis substâncias de referência de agrotóxicos: aldicarbe, brodifacoum, bromadiolona, clorpirifós, metomil e terbufós. Aldicarbe (99%) e metomil (99%) foram fornecidos pela ChemService; brodifacoum (99,3%), bromadiolona (99,8%) e terbufós (98,6%) foram adquiridos da Pestanal®; e o clorpirifós (98,4%) foi comprado da Dr. Ehrenstorfer™. Um resumo das principais informações comerciais de cada padrão analítico encontra-se listado no QUADRO 1.

A preparação das soluções estoque foi realizada com base na concentração de 1 mg/mL para cada substância de referência. Para isso, foram pesados exatamente 10 mg de cada composto utilizando uma balança analítica Mettler Toledo, modelo Excellence Plus XP 205, com precisão de 0,01 mg (Columbus, EUA). Em seguida, os padrões foram dissolvidos em 10 mL de acetonitrila grau HPLC (Tedia, Fairfield, EUA), em tubos cônicos de centrífuga de 15 mL com tampa de rosca da Techno Plastic Products AG (Trasadingen, Suíça). As soluções preparadas foram armazenadas em freezer a -40 °C, em frascos devidamente identificados, visando à preservação de sua estabilidade química e à prevenção de processos de degradação. Todos os equipamentos e materiais empregados nesta etapa fazem parte da infraestrutura do Centro de Estudos em Biofarmácia da Universidade Federal do Paraná (UFPR), onde foi realizada a parte experimental deste estudo.

QUADRO 1 - CARACTERÍSTICAS COMERCIAIS DOS PADRÕES ANALÍTICOS SELECIONADOS

ANALITO	MASSA MOLECULAR (g/mol)	TEOR	MARCA	FÓRMULA ESTRUTURAL
Aldicarbe	190,27	99%	Chem Service	
Brodifacoum	523,43	99,3%	PESTANAL ®	
Bromadiolona	527,41	98,8%	PESTANAL ®	
Metomil	126,21	99%	Chem Service	
Terbufós	288,43	98,6%	PESTANAL ®	
Clorpririfós	350,6	98,4%	Dr. Ehrenstorfer™	

FONTE: A autora (2025).

3.2 PREPARO DA MATRIZ ALIMENTAR

Foram analisados cinco tipos de alimentos cozidos, adquiridos em um restaurante localizado na cidade de Curitiba – PR. Os alimentos selecionados foram: arroz branco, carne moída, pão, mortadela (esta adquirida em supermercado) e uma mistura composta por diferentes itens alimentícios, preparada com 100 gramas de arroz branco, 82 gramas de macarrão, 158 gramas de feijão, 96 gramas de carne de porco, 66 gramas de maionese e 62 gramas de doce de leite condensado. Cada um dos cinco grupos alimentares foi homogeneizado em liquidificador até a obtenção de uma mistura uniforme e pastosa. As amostras resultantes foram então armazenadas em freezer a -40 °C, onde permaneceram até o momento da fortificação.

A escolha das matrizes utilizadas neste estudo foi alinhada à realidade observada no âmbito das análises periciais. Optou-se, portanto, por empregar cinco diferentes matrizes alimentares selecionadas com base no histórico das amostras mais frequentemente submetidas à perícia, nos últimos anos, na Seção de Química Forense da Polícia Científica do Paraná. Tal decisão teve como objetivo assegurar que o método desenvolvido apresentasse elevada aplicabilidade prática e relevância no contexto forense, refletindo, de forma fidedigna, a diversidade e a complexidade dos materiais comumente analisados.

3.3 FORTIFICAÇÃO DO ALIMENTO COM OS PADRÕES ANALÍTICOS

Para a etapa de fortificação, as amostras alimentares previamente homogeneizadas foram submetidas à adição controlada de padrões analíticos de agrotóxicos, com o objetivo de simular situações de contaminação deliberada.

Foram pesados 2 g de cada matriz alimentar e acondicionados em tubos de centrífuga, aos quais foram adicionadas alíquotas das soluções dos padrões analíticos, de modo a se obter concentrações finais de 2, 5, 10, 30, 40 e 50 ppm para cada agrotóxico em cada grupo alimentar. Essa estratégia visou representar diferentes níveis de contaminação e permitir a avaliação do desempenho analítico do método proposto em distintas faixas de concentração. Considerando que foram

avaliados cinco tipos de alimentos, e que cada um foi contaminado individualmente com seis agrotóxicos em seis concentrações diferentes, foram preparados, ao todo, 180 tubos contendo amostras fortificadas.

A definição das concentrações estudadas foi fundamentada em uma estratégia que também leva em consideração o contexto forense. Diferente dos cenários toxicológicos ambientais ou ocupacionais, nos quais os níveis de exposição são geralmente baixos e estão sujeitos a limites máximos de resíduos, em casos de contaminação intencional, presume-se o uso de concentrações significativamente superiores às normalmente toleradas. As concentrações de 2 a 50 ppm estabelecidas podem parecer baixas se comparadas a amostras forenses reais, onde a intencionalidade do ato criminoso tende a envolver doses elevadas de substâncias tóxicas, com o objetivo de causar danos deliberados à vida. No entanto, essa faixa de concentrações visa não apenas refletir a variabilidade esperada em situações reais de contaminação proposital, mas também avaliar a robustez e a sensibilidade dos modelos.

Essa escolha também pode ser ilustrada por situações práticas observadas em perícias. Por exemplo, considere a manipulação criminosa de uma quantidade moderada de alimento, como 200 g de arroz, com a adição de um pacote inteiro de raticida à base de bromadiolona - substância cuja comercialização no Brasil é permitida apenas em embalagens de 20 g, com concentração de 0,005%. Se a mistura for devidamente homogenizada, o cenário resultaria em uma concentração média de cerca de 5 ppm no alimento. No entanto, considerando que é comum a dispersão não homogênea do contaminante nesses eventos, é razoável supor que porções do alimento possam apresentar concentrações ainda mais elevadas. Dessa forma, a faixa de 2 a 50 ppm adotada neste estudo revela-se adequada para simular condições realistas de contaminação intencional no contexto forense, abrangendo tanto os níveis potencialmente encontrados em vestígios periciais quanto aqueles mais elevados, típicos de atos deliberados de envenenamento.

Após a adição dos padrões, o conteúdo de cada tubo foi homogeneizado por meio de agitação em vórtex (Genie 2, Bohemia, EUA) por 5 minutos, seguida de um banho ultrassônico em cuba Branson 2510 (Danbury, EUA) por 10 minutos. Em seguida, as amostras foram submetidas à centrifugação refrigerada utilizando a centrífuga de marca Eppendorf, modelo 5810-R (Hamburgo, Alemanha), utilizando

4.000 rpm por 5 minutos, a 8 °C. Por fim, cada tubo foi novamente agitado em vórtex por 2 minutos adicionais a fim de garantir a completa homogeneização.

Concluído o processo de preparo, as amostras foram armazenadas em freezer a -40 °C, permanecendo sob essas condições até o momento da coleta dos espectros na região do infravermelho.

3.4 COLETA DOS ESPECTROS FTIR

As amostras foram analisadas em um espectrômetro de infravermelho médio com Transformada de Fourier da marca Bruker (Bruker OPTIK GmbH, Ettlingen, BW, Alemanha), modelo Alpha P, equipado com acessório de Refletância Total Atenuada (ATR) com cristal de diamante. As análises foram realizadas sob condições controladas de temperatura (25,0 ± 0,2 ºC) e umidade (45–55%).

As amostras, previamente armazenadas a -40 ºC, foram retiradas do freezer e deixadas em repouso até atingirem a temperatura ambiente. Em seguida, foram homogeneizadas em vórtex por 2 minutos e, então, cuidadosamente aplicadas sobre o cristal de diamante do acessório ATR, assegurando contato adequado com a superfície do cristal. Para garantir essa interface, foi utilizada uma pressão moderada exercida pelo sistema de fixação automática acoplado ao equipamento.

Com o objetivo de reduzir a variabilidade entre porções, foram analisadas 20 alíquotas individuais de cada tubo, totalizando 3700 espectros ao final do experimento. Entre cada aquisição espectral, o cristal foi higienizado com álcool isopropílico e papel macio, a fim de evitar contaminações cruzadas.

A decisão de registrar 20 espectros para cada grupo experimental fundamenta-se em princípios consolidados da quimiometria, segundo os quais a qualidade e a utilidade dos dados analíticos dependem diretamente da forma como são adquiridos. Essa premissa reforça a importância de um protocolo experimental bem delineado e de um volume amostral capaz de refletir adequadamente as variações naturais do sistema investigado.

Visando reduzir a interferência do ruído estocástico - flutuações aleatórias decorrentes de erros experimentais, de amostragem ou de instrumentação -, adotou-se a estratégia de realizar 20 replicações espectrais por grupo. Essa abordagem

apoia-se no princípio de que o ruído, por apresentar natureza tanto positiva quanto negativa, tende a ser atenuado por meio da média de múltiplas medições independentes. Ao consolidar os espectros dessa forma, busca-se representar com maior fidedignidade o sinal real de cada grupo, fortalecendo a qualidade dos modelos quimiométricos e garantindo que estes operem sobre uma base de dados representativa e confiável (FERREIRA, 2015a).

Os espectros foram adquiridos por meio do *software* OPUS (versão 6.0 para Windows), da Bruker Optics (Billerica, MA, EUA), na faixa espectral de 4000–400 cm^{-1}, com 24 varreduras por amostra e resolução de 4 cm^{-1}, resultando em um total de 2541 variáveis por espectro. Para minimizar interferências causadas por CO_2 e H_2O presente no ambiente, foi registrado um espectro de fundo das condições atmosféricas imediatamente antes da leitura a cada 10 amostras analisadas.

3.5 ORGANIZAÇÃO DOS DADOS ESPECTRAIS

Os espectros obtidos foram exportados em formato de texto (.dpt), contendo as intensidades dos picos para cada variável (comprimento de onda), e organizados em uma planilha por meio do software Microsoft Excel 2019. A estruturação dos dados resultou em uma matriz **X**, composta por 3700 linhas (amostras) e 2541 colunas (variáveis espectrais). Para fins de modelagem, todas as amostras pertencentes a um mesmo analito foram agrupadas em uma única classe, independentemente da concentração utilizada na fortificação do alimento.

Em seguida, os dados foram importados para o software Origin (versão 9.6.5.169, 2019b) com o objetivo de adequar o formato ao ambiente MATLAB (versão 7.5.0.342; R2007b, 2007, Natick, MA, EUA), onde foram processados utilizando o pacote PLS toolbox 3.0.16 (Eigenvector Research Inc., Wenatchee, WA, EUA). O MATLAB, amplamente utilizado em estudos quimiométricos, permitiu a interpretação e o multiprocessamento dos dados por meio dos algoritmos de PCA e PLS-DA.

Além disso, a planilha inicial organizada no Microsoft Excel também foi utilizada em etapas posteriores de classificação supervisionada, por meio da

aplicação de algoritmos disponíveis na biblioteca PyCaret, executada de forma automatizada em ambiente Google Colab.

3.6 ANÁLISE QUIMIOMÉTRICA

3.6.1 Análise de Componentes Principais (PCA)

O método escolhido para a análise exploratória dos dados foi a PCA, uma técnica amplamente utilizada para investigar estruturas em conjuntos de dados complexos. Nesse estudo, a PCA foi aplicada com o objetivo de explorar a estrutura multivariada dos dados espectrais, identificar padrões de agrupamento entre as amostras e, simultaneamente, reduzir a dimensionalidade do conjunto de dados. A condução da análise seguiu as recomendações metodológicas de Souza e Poppi (2012).

Antes da importação dos dados para o ambiente do MATLAB, os espectros foram organizados em uma matriz (matriz **X**), na qual as linhas representavam as amostras e as colunas correspondiam aos valores de transmitância para cada número de onda. Foram criados sete vetores de classes (vetor C) para identificar o grupo ao qual cada amostra pertencia: alimento não contaminado, alimento contaminado com aldicarbe, alimento contaminado com brodifacoum, alimento contaminado com bromadiolona, alimento contaminado com clorpirifós, alimento contaminado com metomil e alimento contaminado com terbufós. Além disso, foi elaborado um vetor de variáveis (vetor V), contendo os números de onda associados a cada coluna da matriz **X**.

Com o intuito de minimizar efeitos indesejados, como ruído instrumental, variações de linha de base e dispersão de luz, os dados espectrais foram submetidos a etapas de pré-processamento. Essas etapas foram divididas em dois tipos: transformações aplicadas às amostras (linhas da matriz **X**) e pré-tratamentos aplicados às variáveis (colunas da matriz **X**). Antes da aplicação desses procedimentos, realizou-se uma inspeção visual dos espectros brutos para identificar

possíveis erros grosseiros nos dados, garantindo a qualidade inicial da matriz espectral.

A matriz de dados, após o pré-tratamento, foi submetida à PCA utilizando o método de decomposição por autovalores (*Eigenvalue Decomposition* – EVD). Matematicamente, a PCA decompõe a matriz X no produto de duas matrizes — escores (T) e pesos (P) — mais uma matriz de resíduos (E), conforme representado na seguinte Equação 2 abaixo:

$$X = TP^T + E \tag{2}$$

Em que:

T representa os escores (projeções das amostras nos novos eixos);

P representa os pesos (*loadings*), que indicam a contribuição de cada variável para os componentes principais;

E representa os resíduos (parte da variação não explicada pelo modelo).

A definição do número ótimo de componentes principais considerou a variância explicada acumulada. Os gráficos de escores foram utilizados para investigar possíveis agrupamentos naturais entre as amostras, revelando diferenças nos perfis espectrais. Paralelamente, os gráficos de pesos foram analisados para identificar as regiões espectrais mais relevantes para a separação entre os grupos, permitindo inferências sobre os grupos funcionais responsáveis pelas diferenças observadas.

A qualidade da análise e a presença de amostras anômalas foram avaliadas por meio de critérios estatísticos baseados nas distâncias de *Hotelling* (T^2) e nas distâncias ortogonais (Q *residual*), conforme proposto por Rodionova (2021). A distância de *Hotelling* (T^2) mede a posição relativa das amostras no espaço modelado pelas componentes principais, sendo valores elevados indicativos de variações sistemáticas. Já a distância ortogonal (Q residual) quantifica a parte da variabilidade não explicada pelo modelo, ou seja, o quanto da estrutura original dos dados permanece fora do espaço definido pelos componentes principais. Valores elevados de Q indicam que a amostra não está bem representada.

Para facilitar a interpretação conjunta desses parâmetros, foi construído um gráfico bidimensional (Q *versus* T^2), no qual as amostras foram classificadas em três categorias (SANTANA *et al.*, 2020):

- Regulares, quando situadas dentro dos limites de confiança (95%) para T^2 e Q;
- Extremas, quando apresentam T^2 elevado, mas Q dentro do limite;
- *Outliers*, quando os valores de T^2 e Q ultrapassam simultaneamente os limites críticos.

Os limites críticos para T^2 e Q foram definidos com base em distribuições teóricas para um nível de significância de 5% ($p = 0,05$). Essa abordagem permitiu uma avaliação objetiva e estatisticamente fundamentada da adequação do modelo e da presença de amostras potencialmente problemáticas.

Por fim, a interpretação conjunta dos gráficos de escores e pesos seguiu as diretrizes de Oliveri *et al.* (2019), permitindo a associação entre padrões de agrupamento e variáveis espectrais relevantes. A interpretação química das regiões espectrais mais influentes foi realizada com base na abordagem sistemática proposta por Coates (2006), que organiza a análise segundo as frequências vibracionais características dos principais grupos funcionais presentes nos espectros de infravermelho.

3.6.2 Análise discriminante de mínimos quadrados parciais (PLS-DA)

A Análise Discriminante por Mínimos Quadrados Parciais (PLS-DA) foi empregada neste estudo como abordagem supervisionada para a classificação das amostras com base em seus espectros de infravermelho. O método foi escolhido por sua capacidade de lidar com conjuntos de dados altamente colineares e com número elevado de variáveis, característica comum em matrizes espectrais. Diferentemente da PCA, que é uma técnica exploratória não supervisionada, o algoritmo PLS-DA utiliza informações a priori sobre as categorias das amostras (BARKER, RAYENS, 3003).

Os espectros foram inicialmente organizados em uma matriz **X**, composta pelos espectros, a qual serviu como entrada para o método de classificação PLS-DA. Nessa matriz, cada linha representou uma amostra, enquanto as colunas corresponderam aos valores de transmitância obtidos para cada número de onda.

Neste estudo, foi adotado o método supervisionado PLS2-DA para classificar as amostras em sete categorias distintas: alimento não contaminado, e alimento contaminado com aldicarbe, brodifacoum, bromadiolona, clorpirifós, metomil e terbufós. A escolha pelo método PLS2-DA se justifica pela sua capacidade de modelar simultaneamente todas as classes por meio de um único conjunto de variáveis latentes, o que simplifica o processo de modelagem. Caso fosse empregada a abordagem do PLS1-DA, seria necessário desenvolver sete modelos individuais, um para cada classe, o que tornaria o procedimento mais trabalhoso e computacionalmente mais oneroso. Assim, sempre que houver referência ao modelo PLS-DA ao longo deste trabalho, estará implícita a utilização da variante PLS2-DA.

Uma das etapas principais na construção do modelo PLS-DA foi a escolha adequada do número de variáveis latentes. Quando se escolhe um número insuficiente de variáveis latentes, ocorre o subajuste do modelo, ou seja, nem toda a informação útil disponível é aproveitada em sua construção. Por outro lado, a seleção de um número excessivo de variáveis latentes pode levar ao sobreajuste, situação em que são incorporadas ao modelo informações irrelevantes à propriedade de interesse, como, por exemplo, ruídos espectrais (HAALAND; THOMAS, 1988).

A escolha do número de variáveis latentes foi realizada por meio de validação cruzada do tipo *venetian blinds* com 10 divisões (*splits*), na qual os dados foram separados em dois subconjuntos: um de treinamento (ou calibração), contendo 70% das amostras de cada classe, e outro de teste (ou validação), com os 30% restantes. A divisão 70/30 é fundamental para garantir a robustez do modelo e sua capacidade de generalização, evitando sobreajustes e assegurando a confiabilidade das previsões futuras. A escolha de utilizar a mesma matriz empregada na análise exploratória, por sua vez, reforça o caráter complementar do PLS-DA em relação à PCA, agregando à análise a capacidade de realizar classificações orientadas pelas informações de classe (FERREIRA, 2015a; PERIS-DÍAZ; KRE¸ZEL, 2021; WESTAD; MARINI, 2015).

Em seguida, diferentes modelos foram construídos com variados números de variáveis latentes (de 1 a 15), e as amostras do conjunto de validação foram previstas por cada um desses modelos. O número ótimo de variáveis latentes foi então definido com base na análise conjunta das figuras de mérito sensibilidade e especificidade ao longo da validação cruzada, considerando-se a estabilidade desses parâmetros em função da quantidade de variáveis latentes empregadas.

Antes do início da construção do modelo de classificação propriamente dito, foi fundamental realizar a identificação de possíveis amostras anômalas (*outliers*). Para isso, empregaram-se as estatísticas Q (resíduos) e T^2 de *Hotelling* nos conjuntos de treinamento e teste, adotando-se um nível de significância de 5% (ou seja, um nível de confiança de 95%). Amostras que apresentaram, simultaneamente, valores de Q e T^2 acima dos limites críticos foram consideradas anômalas e, consequentemente, excluídas do modelo (PASQUINI, 2018).

Após a etapa de avaliação dos *outliers*, procedeu-se à análise do gráfico de valores preditos de Y ($\hat{Y}$), gerado pelo modelo PLS-DA para as amostras dos conjuntos de treinamento e teste. Como a variável de resposta categórica é modelada de forma contínua - ou seja, seus valores variam ao longo de um intervalo, geralmente entre 0 e 1 -, e não como categorias discretas, tornou-se necessário definir um valor de corte (*threshold*) que permitisse converter essas previsões contínuas em classificações categóricas, facilitando a interpretação por parte do usuário. Esse limiar foi automaticamente estimado pelo software MATLAB, com base na análise da curva ROC (*Receiver Operating Characteristic*), sendo determinado o ponto ótimo que proporciona o melhor equilíbrio entre sensibilidade e precisão.

A partir da representação visual dos gráficos do Y predito, foi possível identificar os falsos positivos, falsos negativos, verdadeiros positivos e verdadeiros negativos, o que permitiu o cálculo das métricas de desempenho do modelo, como acurácia, sensibilidade e especificidade.

Em modelos de classificação, a sensibilidade representa a capacidade do modelo em identificar corretamente as amostras positivas, ou seja, aquelas que de fato pertencem à classe positiva (IUPAC, 1997). Esse parâmetro é calculado por meio da Equação 3, apresentada a seguir:

$$Sensibilidade = \frac{VP}{VP + FN} \times 100 \quad (3)$$

A especificidade corresponde à capacidade do modelo em identificar corretamente as amostras negativas, ou seja, aquelas que de fato não pertencem à classe positiva (IUPAC, 1997). Esse parâmetro é calculado por meio da Equação 4, apresentada a seguir:

$$Especificidade = \frac{VN}{VN + FP} \times 100 \quad (4)$$

A acurácia, também denominada eficiência, é um parâmetro estatístico que fornece uma medida global do desempenho do modelo de classificação. Seu valor é obtido a partir da razão entre o número total de amostras corretamente classificadas, independentemente da classe, e o número total de amostras avaliadas (IUPAC, 1997), conforme expressa a Equação 5 a seguir:

$$Acurácia = \frac{VP + VN}{VP + VN + FN + FP} \times 100 \quad (5)$$

Além da avaliação do desempenho global do modelo, foi também analisada a contribuição individual de cada variável espectral para a classificação. Para isso, utilizou-se o gráfico de variáveis importantes na projeção (VIP), por meio do qual foi possível identificar as regiões do espectro mais relevantes para a diferenciação entre os grupos.

Por fim, a aplicabilidade prática do modelo desenvolvido foi avaliada utilizando duas amostras provenientes de casos reais já periciados e laudados pela Seção de Química Forense da Polícia Científica do Paraná. Essas amostras foram submetidas ao mesmo tratamento espectral e classificadas pelo modelo PLS-DA, com o objetivo de verificar sua capacidade de generalização e acurácia frente a amostras externas, reforçando sua potencial utilização em contextos forenses reais.

3.6.3 Outros algoritmos de aprendizado de máquina

A metodologia empregada para a construção do modelo de aprendizado de máquina foi implementada na plataforma Google Colab, utilizando a linguagem Python e seguindo uma sequência bem definida de etapas. Inicialmente, a matriz **X** - contendo os dados espectrais e originalmente estruturada em uma planilha do Excel (.xls) - foi importada para o ambiente de desenvolvimento com o objetivo principal de avaliar o desempenho de 12 algoritmos de classificação supervisionada. Em seguida, procedeu-se à importação da biblioteca Pandas, amplamente reconhecida por sua eficiência na manipulação e organização de dados em formato tabular.

Na etapa subsequente, foi realizada a instalação da biblioteca PyCaret, ferramenta que automatiza e simplifica processos relacionados ao aprendizado de máquina, permitindo a comparação direta entre diferentes algoritmos classificatórios. Após sua instalação, a biblioteca foi devidamente importada e utilizada nas etapas seguintes do fluxo de modelagem.

Com o ambiente devidamente configurado, iniciou-se o pré-processamento dos dados, etapa essencial para assegurar a qualidade e integridade das informações antes do treinamento dos modelos. Durante a configuração da modelagem com a função `setup()` da PyCaret, diversas operações foram executadas automaticamente, como a imputação de valores ausentes utilizando a média para variáveis numéricas, se necessário.

Outro aspecto relevante definido nessa etapa foi a validação cruzada, implementada por meio do método *StratifiedKFold*, que assegura a manutenção da proporção entre as classes em cada subdivisão dos dados. Foram utilizados 10 *folds*, conforme configuração padrão da biblioteca, o que implicou na divisão do conjunto total em dez subconjuntos alternados entre treino e teste a cada iteração.

Finalizado o pré-processamento, foi empregada a função `compare_models()`, responsável pela construção e comparação automatizada dos modelos de classificação. Nessa etapa, até 15 algoritmos supervisionados foram avaliados e ranqueados com base em um conjunto abrangente de métricas, incluindo: acurácia, curva ROC (AUC), *recall*, precisão, F1-score, índice Kappa, coeficiente de correlação de Matthews (MCC) e tempo de treinamento (TT). O modelo com melhor desempenho foi selecionado para análise aprofundada e validação externa.

A acurácia representa a proporção de classificações corretas em relação ao total de predições realizadas. A AUC expressa a capacidade discriminativa do modelo em diferenciar as classes. O *recall* (ou sensibilidade) indica a fração de verdadeiros positivos corretamente identificados, enquanto a precisão refere-se à proporção de verdadeiros positivos entre todas as amostras classificadas como positivas. O F1-score é a média harmônica entre precisão e *recall*, refletindo o equilíbrio entre essas duas métricas. O índice Kappa avalia o grau de concordância entre as predições do modelo e os valores reais, ajustando os acertos esperados ao acaso. O MCC (coeficiente de Matthews) é uma métrica robusta que considera todos os elementos da matriz de confusão, sendo especialmente útil em cenários de classes desbalanceadas. Já o tempo de treinamento (TT), expresso em segundos, permite avaliar a eficiência computacional de cada abordagem (JUNIOR; CARPINETTI, 2019; DOMINGUES; PEDROSA; BERNARDINO, 2020; CASTRO; BRAGA, 2011).

Na sequência, o algoritmo de melhor desempenho foi selecionado com o auxílio da função `create_model()`, responsável por treinar o modelo sobre os dados pré-processados, empregando validação cruzada estratificada. A divisão dos dados foi automatizada, alocando 70% das amostras para o conjunto de calibração e 30% para o conjunto de validação.

Concluído o treinamento, diversos gráficos foram gerados automaticamente por meio da função `plot_model()` da PyCaret, entre os quais se destacam: a matriz de confusão, o gráfico de confiabilidade da previsão, o *classification report* (relatório de classificação) e o gráfico de importância das variáveis espectrais — este último, crucial para identificar os atributos os números de onda relevantes para o processo de classificação.

Por fim, os modelos desenvolvidos foram devidamente salvos, assegurando sua reprodutibilidade e possibilitando sua integração com sistemas externos para futuras aplicações em amostras reais. Essa etapa final reforça o potencial prático da metodologia desenvolvida, evidenciando sua aplicabilidade em contextos reais da química forense.

O código em linguagem Python empregado neste trabalho, incluindo as versões utilizadas e as parametrizações dos modelos classificatórios, encontra-se descrito de forma detalhada no Apêndice 8 desta dissertação.

3.7 PREPARO PARA VALIDAÇÃO PRÁTICA DOS MODELOS CLASSIFICATÓRIOS COM AMOSTRAS REAIS PERICIADAS PELA POLÍCIA CIENTÍFICA DO PARANÁ

Duas amostras distintas, oriundas de casos reais analisados pelo Laboratório de Química Forense da Polícia Científica do Paraná, foram utilizadas para testar a aplicabilidade prática dos modelos desenvolvidos. Esses materiais foram apreendidos no ano de 2024 e analisados por técnicas instrumentais de referência. Para o material ilustrado à esquerda na FIGURA 11, empregou-se a cromatografia líquida de alta eficiência com detector de arranjo de diodos (HPLC-DAD), a qual permitiu a identificação de brodifacoum. Já o material apresentado à direita na mesma figura foi analisado por cromatografia gasosa acoplada à espectrometria de massas (GC-MS), resultando na identificação de terbufós. Ambas as amostras foram submetidas aos mesmos procedimentos de preparo aplicados às amostras utilizadas na calibração dos modelos classificatórios, conforme descrito nas Seções 3.3 e 3.4 deste trabalho.

FIGURA 11 - AMOSTRAS REAIS ORIUNDAS DA SEÇÃO DE QUÍMICA FORENSE DA POLÍCIA CIENTÍFICA DO PARANÁ

FONTE: A autora (2025).

LEGENDA: À esquerda, pão em avançado estado de decomposição, com presença de mofo e fragmentos vegetais aderidos à superfície, contaminado com brodifacoum. À direita, dois corpos sólidos de formato circular, identificados como mortadela, contaminados com terbufós. Ambas as amostras foram analisadas pelo Laboratório de Química Forense da Polícia Científica do Paraná em 2024.

Cabe destacar que o presente projeto foi aprovado pela Academia de Ciências Forenses da Polícia Científica do Paraná, sob o número de protocolo 23.333.944-0, o que autorizou formalmente o uso tanto dos padrões analíticos pertencentes à coleção oficial das Seções de Química e Toxicologia Forense quanto das amostras remanescentes de exames periciais. Ressalta-se que essas amostras já haviam sido devidamente periciadas, com laudos conclusivos emitidos, e que contraprovas foram devidamente armazenadas sempre que possível, em conformidade com os procedimentos técnicos e legais vigentes.

4 RESULTADOS E DISCUSSÕES

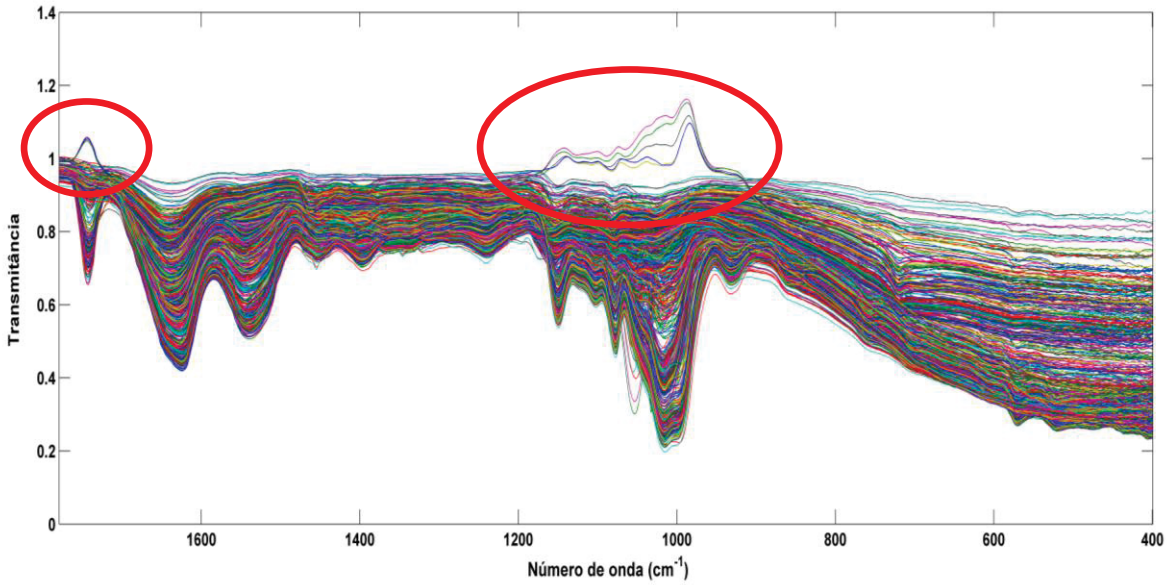
4.1 PERFIL ESPECTRAL

A etapa inicial do estudo quimiométrico consistiu na obtenção e organização dos espectros de infravermelho referentes aos diferentes agrotóxicos incluídos neste estudo. Como descrito na seção “Material e Métodos”, o delineamento experimental previa a obtenção de 600 espectros para cada agrotóxico e, para o grupo dos alimentos isentos de contaminação, estabeleceu-se a aquisição de 100 espectros, totalizando 3700 espectros de infravermelhos.

Cabe destacar que, em alguns casos específicos - como nas classes referentes ao metomil, brodifacoum e bromadiolona - o número final de espectros utilizados foi inferior ao previsto no delineamento experimental, em razão da exclusão de 20, 1 e 2 espectros, respectivamente. Essa redução decorreu de uma etapa inicial fundamental, na qual todos os espectros foram sobrepostos e inspecionados visualmente. Essa análise exploratória inicial, conduzida por inspeção visual (a olho nu), teve como propósito identificar possíveis falhas grosseiras na aquisição dos espectros, resultando na detecção de perfis espectrais nitidamente discrepantes em relação aos demais pertencentes à mesma classe. Essa inconsistência, observável em região específica ilustrada na FIGURA 12, justificou a exclusão dos respectivos espectros, assegurando maior homogeneidade e confiabilidade ao conjunto de dados utilizado nas etapas subsequentes.

Outro ponto a se destacar é referente à região do infravermelho médio analisado. Embora os espectros tenham sido adquiridos no intervalo de 400 a 4000 cm^{-1}, a avaliação quimiométrica concentrou-se no trecho de 400 a 1800 cm^{-1} (FIGURA 12). Essa escolha teve por objetivo minimizar a interferência da banda intensa e característica do estiramento O–H da molécula de água - presente em todas as matrizes alimentares - que, por sua elevada intensidade, poderia mascarar regiões espectrais mais relevantes para a discriminação entre os grupos de amostras (COATES, 2006).

FIGURA 12 - ESPECTROS MIR-FTIR SOBREPOSTOS



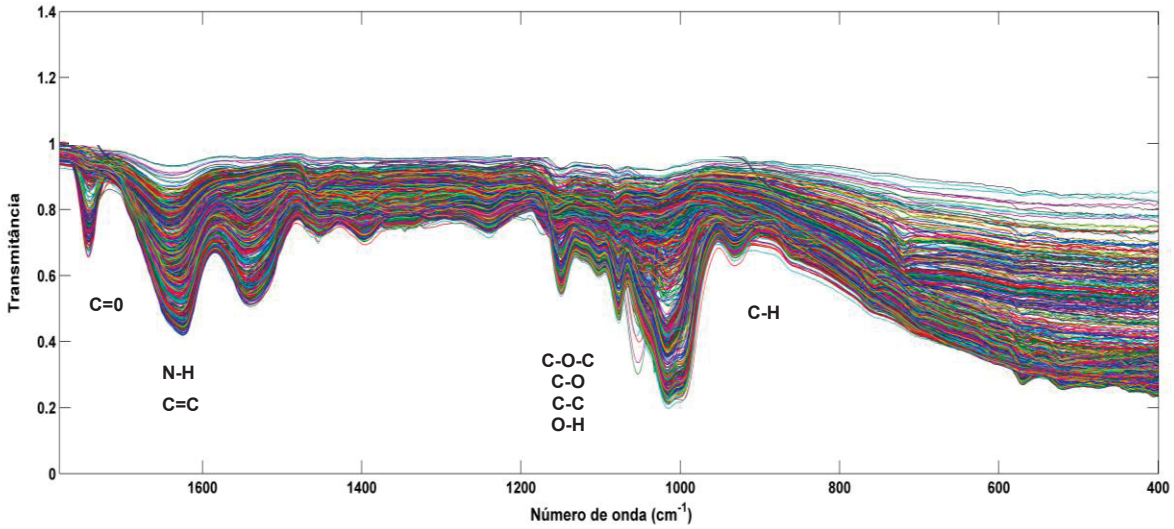
FONTE: A autora (2025).

LEGENDA: Representação dos espectros coletados na faixa de 400 a 1800 cm^{-1}, sobrepostos, com o objetivo de ilustrar o comportamento e a variabilidade espectral das amostras analisadas. Foram adquiridas 20 replicatas para cada matriz alimentar (mistura de alimentos, pão, mortadela, carne e arroz), contaminadas individualmente com os agrotóxicos aldicarbe, brodifacoum, bromadiolona, clorpirifós, metomil e terbufós, em diferentes níveis de concentração. O conjunto total de dados é composto por 3.700 espectros. Destacam-se alguns espectros anômalos, que foram posteriormente excluídos para a análise quimiométrica.

De modo geral, os espectros obtidos e sobrepostos revelaram perfis semelhantes entre si, não sendo possível identificar visualmente, de forma clara, a presença dos diferentes agrotóxicos. Essa limitação pode ser atribuída tanto à baixa concentração dos contaminantes quanto à própria composição das matrizes alimentares analisadas. Tais alimentos, compostos majoritariamente por macronutrientes como proteínas, lipídeos e carboidratos, apresentam concentrações muito superiores às dos agrotóxicos investigados. Como consequência, os sinais espectrais dos constituintes majoritários dominam o espectro global, ofuscando, a olho nu, as contribuições espectrais associadas aos contaminantes presentes.

Ainda assim, mesmo diante da elevada complexidade das amostras e da ausência de informações precisas sobre a composição exata de cada matriz, é possível realizar uma interpretação simplificada das regiões espectrais observadas. Para isso, os espectros foram divididos em quatro faixas principais de absorção, conforme ilustrado na FIGURA 13, permitindo uma análise preliminar dos grupos funcionais predominantes

FIGURA 13 - EVIDÊNCIA DAS PRINCIPAIS BANDAS DE ABSORÇÃO DOS ESPECTROS MIR-FTIR ANALISADOS



FONTE: A autora (2025).

LEGENDA: Espectros MIR-FTIR sobrepostos de todas as amostras analisadas. Os picos observados correspondem a absorções características de diferentes grupos funcionais presentes nas amostras, fornecendo informações sobre a sua composição química.

A primeira região espectral abrange o intervalo de aproximadamente 1850 cm^{-1} a 1610 cm^{-1}, onde se destacam bandas características associadas ao grupo carbonila (C=O), típicas de compostos como proteínas e ésteres derivados de lipídeos (COATES, 2006; PAVIA *et al.*, 2015).

Em continuidade, a segunda região, situada entre 1650 cm^{-1} e 1500 cm^{-1}, apresenta bandas que podem estar associadas ao estiramento de ligações duplas C=C, possivelmente atribuídas a lipídeos — como os ácidos graxos insaturados — ou a compostos que contenham anéis aromáticos. Nessa mesma faixa, também se observam absorções que, muito provavelmente, correspondem às amidas presentes nas proteínas, especialmente à banda conhecida como Amida II, relacionada à deformação angular da ligação N–H (COATES, 2006; PAVIA *et al.*, 2015).

A terceira região relevante localiza-se entre 1300 cm^{-1} e 1050 cm^{-1}, englobando bandas que provavelmente estão associadas tanto aos ésteres presentes em lipídeos quanto aos carboidratos. Nos ésteres, destacam-se os estiramentos assimétrico e simétrico da ligação C–O–C, geralmente observados nas faixas de 1300–1250 cm^{-1} e 1180–1050 cm^{-1}, respectivamente. Essa faixa também se mostra particularmente importante para a análise dos carboidratos, com bandas atribuídas aos estiramentos das ligações C–O e C–C, além das deformações

angulares das ligações O–H. Entre 1000 e 880 cm^{-1}, podem ainda ser observadas bandas que possivelmente correspondem à deformação angular fora do plano dos grupos C–H. Ressalta-se que a segunda e a terceira regiões compõem a chamada "região de impressão digital", considerada extremamente útil para a confirmação da identidade de compostos específicos (BARBOSA, 2007; COATES, 2006).

Por fim, a quarta e última região analisada estende-se de 800 cm^{-1} a 400 cm^{-1}. Trata-se de uma faixa mais difusa do espectro, provavelmente por múltiplas absorções de bandas. Nessa região, também se encontram vibrações características de compostos aromáticos, heteroaromáticos, alquenos e cadeias alifáticas longas (COATES, 2006).

4.2 ANÁLISE DE COMPONENTES PRINCIPAIS (PCA)

A PCA deste estudo foi conduzida com o objetivo de explorar a estrutura dos dados espectrais obtidos a partir das diferentes matrizes alimentares previamente mencionadas na seção “3.6.1 Análise de Componentes Principais (PCA)” dos “Materiais e Métodos”. Embora espectros tenham sido adquiridos para todas as matrizes alimentares investigadas, os melhores resultados - no sentido de proporcionar uma separação mais nítida entre as classes e permitir a interpretação adequada da variabilidade dos dados - foram obtidos ao restringir a análise às amostras de pão e mortadela. Diferentes combinações de pré-tratamentos espectrais foram avaliadas; no entanto, mesmo com essas abordagens, não se observou uma separação clara entre os grupos nas demais matrizes.

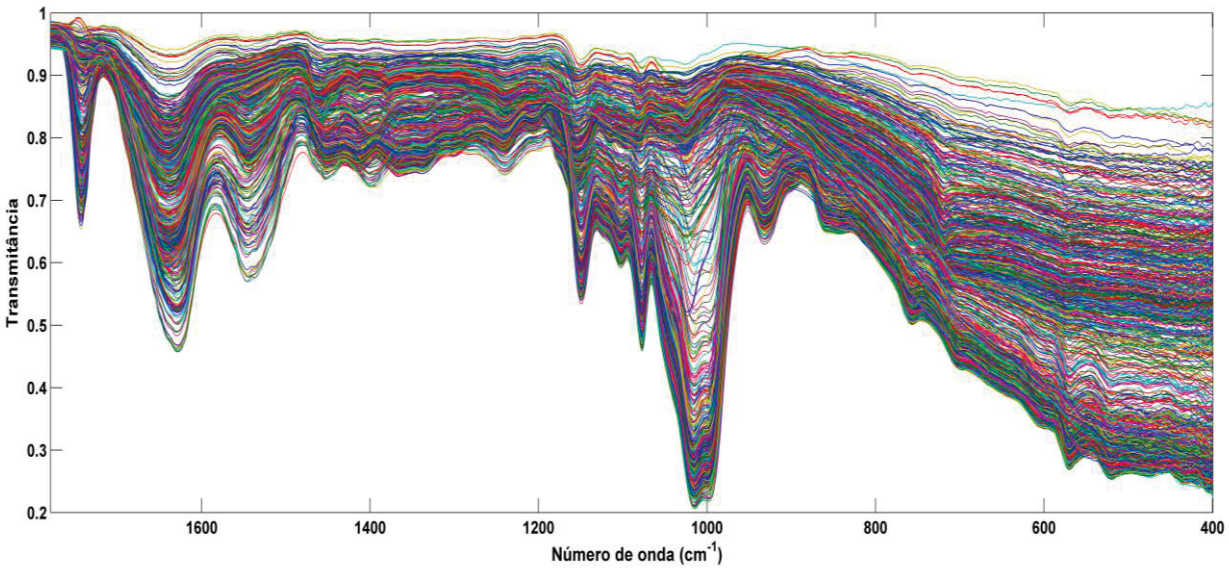
Essa escolha se mostrou estratégica para o prosseguimento tanto da PCA quanto da modelagem subsequente por PLS-DA, cujos detalhes serão aprofundados na seção seguinte. Sendo assim, importante que o leitor saiba que, a partir deste ponto, todas as referências às análises de PCA e PLS-DA dizem respeito exclusivamente às matrizes pão e mortadela.

Com o intuito de apresentar uma visão geral do comportamento do conjunto completo, o APÊNDICE 1 deste trabalho ilustra a distribuição dos grupos considerando todas as matrizes alimentares avaliadas e embora não tenham exibido a mesma eficiência discriminativa nesta etapa inicial, os resultados referentes ao

conjunto total de matrizes serão oportunamente apresentados e discutidos ao longo da dissertação.

Dando continuidade à PCA, a matriz de dados foi segmentada em sete classes distintas: alimento sem contaminante (pão e mortadela), alimento contaminado com aldicarbe, alimento contaminado com brodifacoum, alimento contaminado com bromadiolona, alimento contaminado com clorpirifós, alimento contaminado com metomil e alimento contaminado com terbufós. A FIGURA 14 exibe os espectros de transmitância obtidos para todas as classes, dispostos de forma sobreposta. As médias espectrais de cada grupo foram calculadas e empregadas como perfis representativos, constituindo a base para a aplicação da PCA.

FIGURA 14 - ESPECTROS MIR-FTIR OBTIDOS DAS MATRIZES PÃO E MORTADELA CONTAMINADOS COM AGROTÓXICOS



FONTE: A autora (2025).

Antes da aplicação das abordagens quimiométricas, foi imprescindível submeter os dados espectrais a procedimentos de pré-tratamento adequados. Isso porque dados experimentais organizados na forma matricial podem apresentar variações indesejadas que não foram eliminadas durante a aquisição e que, caso não sejam devidamente tratadas, podem comprometer significativamente a

interpretação dos modelos. O objetivo central do pré-tratamento, portanto, consiste em reduzir tais variações, preservando as informações pertinentes ao sinal verdadeiro, ao passo que minimiza as contribuições estocásticas (flutuações aleatórias introduzidas por fatores imprevisíveis, como ruído eletrônico, erros experimentais e de amostragem, variações ambientais transitórias) e sistemáticas (variações previsíveis e reprodutível associadas a fatores externos e à instrumentação). Neste contexto, é importante salientar que os procedimentos de pré-tratamento podem ser classificados, de maneira geral, em duas categorias: transformação, quando aplicados às amostras (linhas da matriz **X**), e pré-processamento, quando direcionados às variáveis (colunas da matriz **X**) (FERREIRA, 2015a).

Diversas abordagens de pré-tratamento foram avaliadas com o objetivo de identificar aquela que oferecesse a melhor representação do conjunto de dados espectrais. Dentre os métodos testados, a combinação que se revelou mais eficaz consistiu na aplicação do *Generalized Least Squares Weighting* (GLS-W), com o parâmetro α ajustado para 0,0002, associada à normalização Euclidiana - ambos empregados como transformações - e à centragem dos dados na média, esta adotada como pré-processamento.

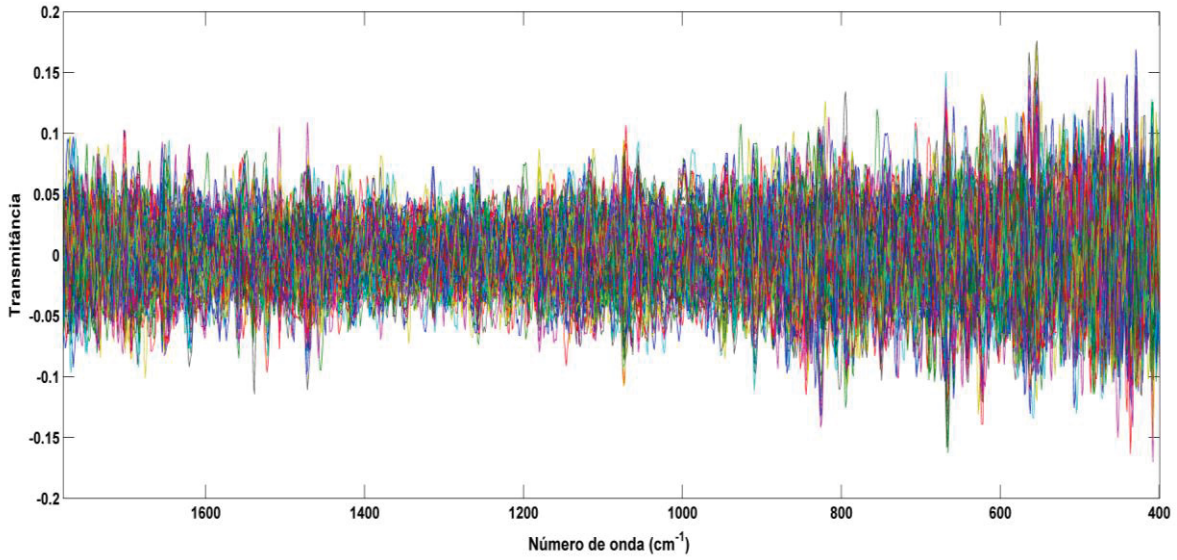
O primeiro pré-tratamento realizado foi o uso do GLS-W, que atua como um filtro que atribui menor peso a determinadas faixas espectrais, com o objetivo de reduzir as diferenças entre amostras semelhantes, por exemplo, aquelas pertencentes à mesma classe (WISE *et al.*, 2019; ZORZETTI, SHAVER e HARYNUK, 2011).

Após a aplicação do filtro inicial, os espectros foram submetidos à normalização Euclidiana, procedimento no qual os valores de cada variável de uma amostra são divididos por um fator de normalização — neste caso, a norma Euclidiana, calculada como a raiz quadrada da soma dos quadrados de suas componentes (FERREIRA, 2015a). Essa etapa revelou-se particularmente relevante para este estudo, uma vez que é indicada quando o objetivo é preservar apenas as informações qualitativas capazes de diferenciar as amostras, eliminando, assim, variações relacionadas a diferenças meramente quantitativas de concentração, o que se adequa perfeitamente ao escopo da presente análise.

Na sequência, com vistas a ajustar adequadamente os dados para a modelagem multivariada, aplicou-se a centragem na média. Nesse procedimento,

calcula-se inicialmente o valor médio de cada coluna da matriz de dados e, em seguida, subtrai-se esse valor de cada elemento da respectiva coluna (SOUZA e POPPI, 2012). Trata-se de uma operação que implica apenas uma translação dos dados em relação ao seu eixo, sem modificar a estrutura intrínseca do conjunto. *Os e*spectros resultantes dessas etapas encontram-se apresentados na FIGURA 15.

FIGURA 15 - ESPECTROS MIR-FTIR PRÉ-TRATADOS OBTIDOS DAS MATRIZES PÃO E MORTADELA CONTAMINADAS COM AGROTÓXICOS

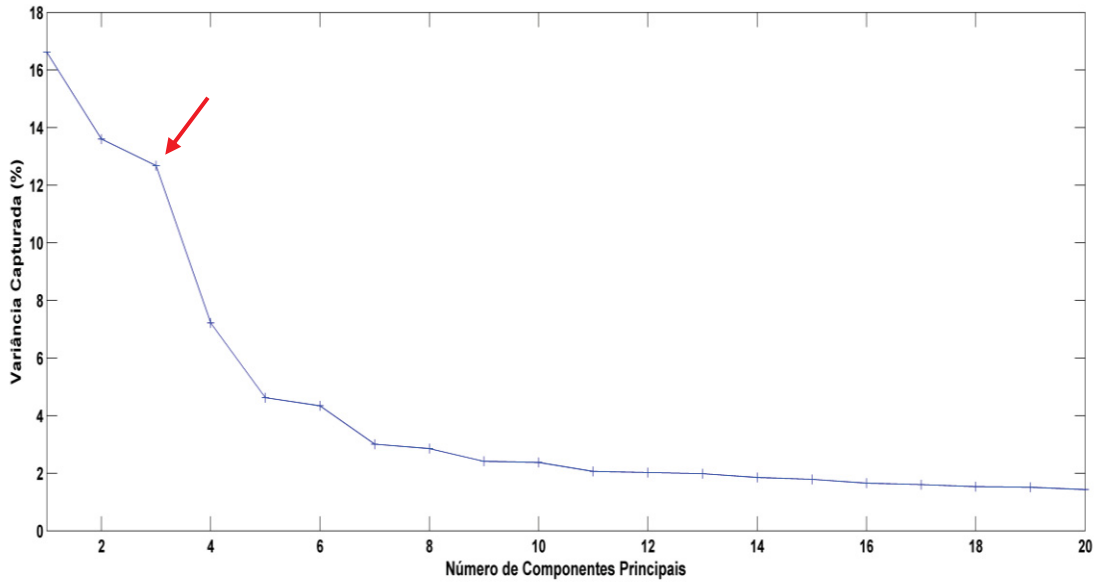


FONTE: A autora (2025).

Em seguida, procedeu-se à definição do número adequado de componentes principais (PCs) a serem utilizados, de modo a assegurar uma representação eficiente dos dados, minimizando resíduos e preservando informações relevantes. Para tanto, foi analisado o gráfico que relaciona a variância explicada ao número de componentes (FIGURA 16), a partir do qual se concluiu que a utilização de três componentes principais seria a mais apropriada. Após a terceira componente, a variância adicional capturada pelas demais PCs torna-se progressivamente reduzida, o que sugere que essas componentes posteriores podem estar associadas majoritariamente a ruído experimental, em vez de informações úteis para a modelagem.

A escolha por três componentes principais se justifica também pela sua contribuição para a explicação da variabilidade total dos dados. Juntas, essas três PCs foram responsáveis por descrever 42,90% da variância total presente no conjunto original. Especificamente, a contribuição individual de cada componente foi de 16,62% para a primeira, 13,60% para a segunda e 12,68% para a terceira.

FIGURA 16 - ANÁLISE GRÁFICA DA ESCOLHA DO NÚMERO DE COMPONENTES PRINCIPAIS



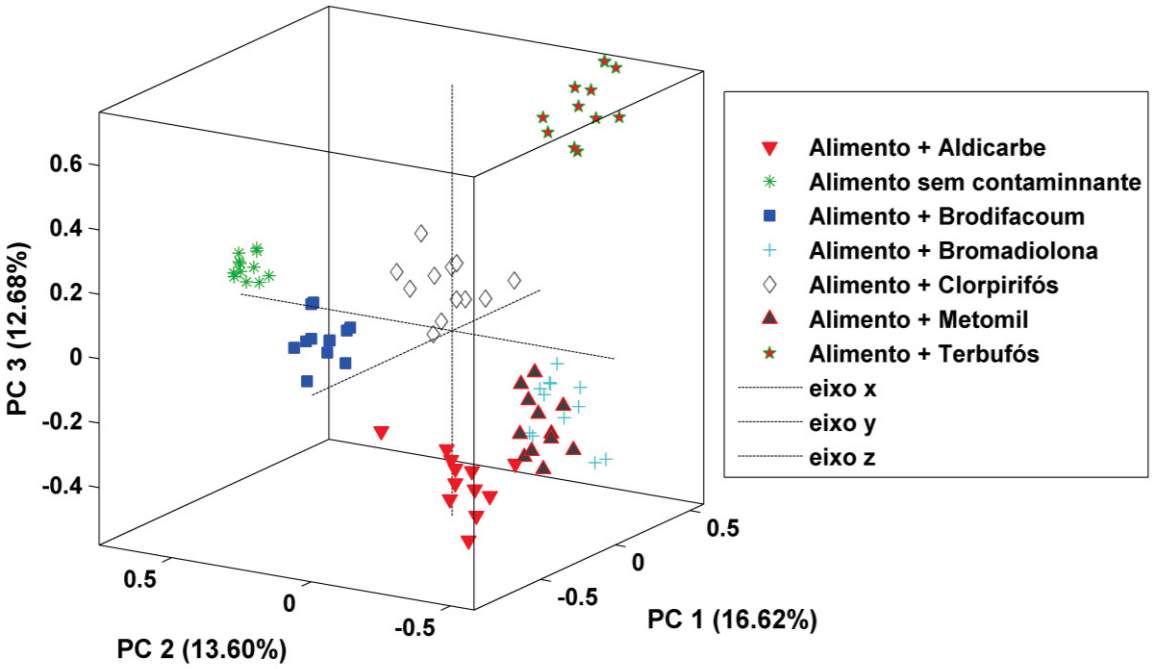
FONTE: A autora (2025).

LEGENDA: A seta indicando o número de componentes principais em que a curva começa a descer assintoticamente para zero, indicando a quantidade ideal.

Dando continuidade à aplicação das três componentes principais selecionadas, a FIGURA 17 apresenta o gráfico de escores, no qual é possível observar a projeção das amostras no espaço definido por essas três PCs. Cada ponto no gráfico representa um espectro individual. Observa-se uma distribuição organizada das amostras, com formação clara de sete agrupamentos. Esse comportamento é indicativo de que a PCA foi eficaz em capturar a variabilidade relevante dos dados, conseguindo resumir em poucas dimensões as informações essenciais para a diferenciação entre os grupos. A proximidade entre os pontos

reflete similaridade espectral, enquanto distâncias maiores sugerem dissimilaridades significativas entre as amostras.

FIGURA 17 - GRÁFICO DOS ESCORES DA PC1 X PC2 X PC3 REFERENTE AOS ESPECTROS PRÉ-TRATADOS DAS AMOSTRAS DE PÃO E MORTADELA CONTAMINADAS COM DIFERENTES AGROTÓXICOS



FONTE: A autora (2025).

Observa-se, nesta etapa da análise, que a PC1 apresentou papel determinante na segregação dos alimentos contaminados com brodifacoum, clorpirifós e terbufós em relação às demais classes avaliadas. A PC2, por sua vez, destacou-se por diferenciar as amostras de alimento isento de contaminantes, bem como aquelas contendo brodifacoum e clorpirifós, das demais categorias. Por fim, a PC3 contribuiu para a separação das amostras de alimento sem contaminantes, alimentos contaminados com terbufós e clorpirifós, em relação às demais classes presentes no conjunto de dados.

Complementando essa análise, uma observação mais criteriosa do gráfico de escores revelou que, embora metomil e bromadiolona não compartilhem

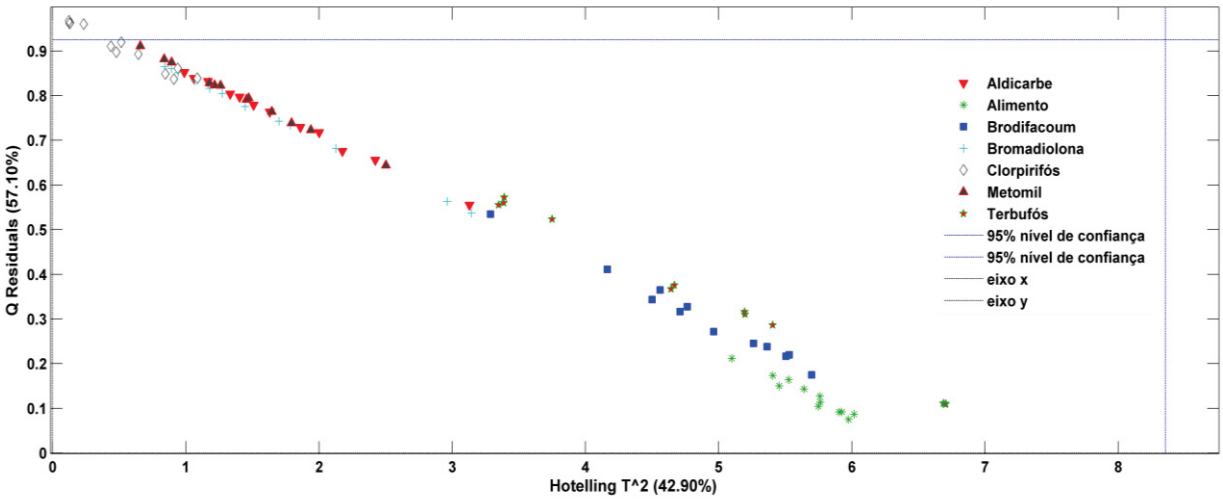
similaridade química evidente, suas respectivas amostras foram projetadas em regiões próximas no espaço das três PCs. Ainda mais intrigante foi a constatação de que as amostras contaminadas com brodifacoum e bromadiolona - compostos estruturalmente relacionados, pertencentes à classe dos raticidas anticoagulantes cumarínicos - exibiram comportamentos distintos, localizando-se em áreas afastadas no gráfico. À primeira vista, tal fenômeno pode parecer paradoxal. No entanto, conforme salientado por Ferreira (2015a), as componentes principais são definidas com o objetivo de maximizar a variância com base nas informações contidas nas variáveis, o que implica que sua orientação não está, necessariamente, vinculada à diferenciação entre as classes. Essa característica, por sua natureza exploratória, constitui uma limitação intrínseca à técnica de PCA e justifica a adoção de métodos supervisionados, como o PLS-DA, os quais são fundamentais para promover uma discriminação orientada pelas informações de classe.

Complementarmente, a qualidade da análise exploratória com a identificação de possíveis amostras anômalas foi avaliada com base em critérios estatísticos fundamentados nas distâncias de *Hotelling* (T^2) e nas distâncias ortogonais (resíduo Q). A distância de *Hotelling* (T^2) expressa a posição relativa das amostras no espaço formado pelos componentes principais, sendo que valores elevados de T^2 indicam que a amostra se encontra mais afastada do centro do modelo, podendo refletir variações sistemáticas dentro do espaço modelado. Por sua vez, a distância ortogonal, ou resíduo Q, mede a variabilidade que não foi explicada pela PCA - ou seja, a porção da estrutura original dos dados que permanece fora do espaço definido pelos componentes principais. Altos valores de Q sugerem que a amostra não está adequadamente representada pelo modelo (RODIONOVA; 2021).

Para uma interpretação conjunta dessas métricas, foi construído o gráfico de resíduos Q *versus* T^2 (FIGURA 18), no qual os valores de T^2 e Q foram dispostos em um plano bidimensional, permitindo a classificação visual das amostras em três categorias distintas. As amostras consideradas regulares situam-se dentro dos limites de confiança estatísticos, estabelecidos em 95%, tanto para T^2 quanto para Q. Amostras classificadas como extremas apresentam valores elevados de T^2, ainda que permaneçam dentro da análise, denotando uma variabilidade sistemática acentuada. Já os *outliers*, por exibirem simultaneamente altos valores de T^2 e de Q, localizam-se fora da região esperada para a variabilidade natural do sistema,

configurando desvios que podem comprometer a robustez do modelo e exigir análise específica (SANTANA *et al.*, 2020).

FIGURA 18 - GRÁFICO DE T^2 *HOTELLING VERSUS* RESÍDUOS Q DA PCA COM 3 COMPONENTES PRINCIPAIS



FONTE: A autora (2025).

Observa-se que, em sua maioria, as amostras encontram-se bem distribuídas ao longo da tendência principal do gráfico, permanecendo dentro dos limites estabelecidos e demonstrando boa representatividade no espaço das três componentes principais selecionadas, sem evidência de amostras anômalas. Contudo, algumas observações atribuídas à classe Clorpirifós apresentaram valores elevados de resíduos Q. Tais amostras correspondem a espectros de pão contaminado com clorpirifós nas concentrações de 30, 40 e 50 ppm e, por refletirem variações inerentes ao sistema real analisado, foram mantidas na base de dados, uma vez que não se trata de valores espúrios, mas sim de amostras de fato representativas.

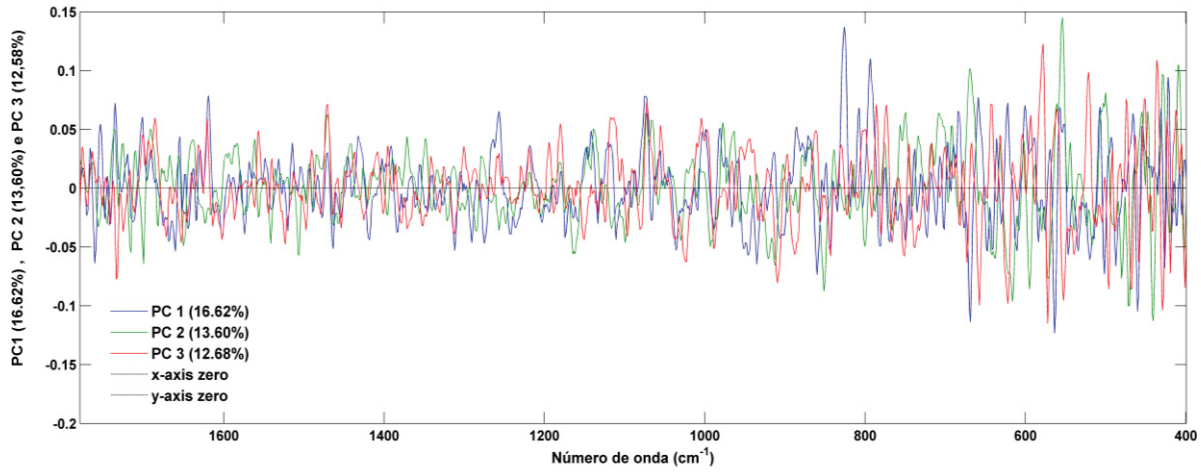
Dando prosseguimento à análise exploratória, foi examinado o gráfico dos pesos, com o intuito de identificar as variáveis espectrais que mais contribuíram para a formação dos agrupamentos observados no gráfico de escores. Esse gráfico evidenciou que determinadas regiões ao longo do espectro exerceram maior influência na separação entre as classes, indicando os comprimentos de onda que mais impactaram a estrutura dos dados no espaço das componentes principais. A

análise detalhada dos pesos das três primeiras componentes principais (PC1, PC2 e PC3), apresentada na FIGURA 19, revelou que, embora todo o espectro tenha contribuído para a modelagem, algumas regiões mostraram-se especialmente relevantes.

No caso da PC1, responsável por 16,62% da variância explicada, destacaram-se as regiões próximas de 950, 800 e 450 cm^{-1} (porção positiva), além das faixas em torno de 680 e 580 cm^{-1} (porção negativa). A PC2, que explica 13,60% da variância, demonstrou forte influência das regiões de 750, 700 e 680 cm^{-1} (porção positiva), bem como de 600 e 400 cm^{-1} (porção negativa). Já a PC3, com contribuição de 12,68%, apresentou influência mais localizada, com destaque para as variáveis situadas nas proximidades de 530 e 580 cm^{-1}, ambas na porção positiva do gráfico.

No gráfico dos pesos, a porção positiva ou negativa indica a contribuição de cada variável para a posição de uma amostra no espaço das componentes principais. Por exemplo, quando um número de onda associado à PC1 está localizado na porção positiva do gráfico de pesos, isso sugere que as amostras com escores positivos na direção da PC1, no gráfico de escores, apresentam maior intensidade naquele número de onda. Da mesma forma, variáveis situadas na porção negativa do gráfico de pesos indicam que amostras com escores negativos na PC1 tendem a ser mais influenciadas por essas regiões espectrais.

FIGURA 19 - GRÁFICO DOS PESOS DA PC1, PC2 E PC3



FONTE: A autora (2025).

De modo geral, observou-se que todo o espectro contribuiu para a separação das classes, ainda que de maneira distribuída. No entanto, determinadas regiões se destacaram por seu papel preponderante na discriminação entre as amostras, especialmente aquelas localizadas na porção final do espectro (entre aproximadamente 900 e 400 cm^{-1}), onde se concentram bandas vibracionais altamente específicas e sensíveis às variações estruturais dos compostos presentes nas amostras analisadas. As bandas nas regiões de 950 e 800 cm^{-1} podem ser atribuídas à deformação angular fora do plano da ligação C–H de compostos aromáticos, presentes em clorpirifós, brodifacoum e bromadiolona. Entre 700 e 600 cm^{-1}, observa-se uma banda que pode ser característica do estiramento da ligação C–S do grupo tiometílico, encontrado em aldicarbe e metomil. Já a região próxima a 450 cm^{-1} pode estar relacionada ao estiramento da ligação C–Cl, característica estrutural do clorpirifós. A banda em 580 cm^{-1}, por sua vez, pode corresponder ao estiramento da ligação C–Br, típica de compostos organobromados como brodifacoum e bromadiolona. Além disso, na faixa entre 550 e 650 cm^{-1}, identificam-se bandas compatíveis com o estiramento da ligação P–S, típico de organofosforados do tipo fosforotioato, como o clorpirifós e o terbufós (PAVIA *et al.*, 2015; COATES, 2006).

As interpretações apresentadas no parágrafo acima foram fundamentadas na literatura científica especializada, bem como na análise dos espectros individuais dos agrotóxicos avaliados, os quais se encontram disponíveis ao leitor nos apêndices desta dissertação.

Em síntese, a interpretação das bandas que mais contribuíram para a PCA demonstra que, embora a separação entre os grupos tenha resultado da atuação conjunta de várias regiões espectrais, determinadas bandas associadas a grupos funcionais específicos - notadamente haletos orgânicos e sistemas aromáticos - exerceram papel particularmente relevante na diferenciação observada no espaço das componentes principais. Ressalta-se, contudo, que essa interpretação representa uma tentativa didática e simplificada de associar bandas espectrais a estruturas moleculares. Na prática, as amostras analisadas constituem matrizes alimentares extremamente complexas, contaminadas com diferentes agrotóxicos, cujas interações químicas são múltiplas e nem sempre plenamente compreendidas. Assim, embora tais atribuições auxiliem na compreensão dos resultados, devem ser

interpretadas com cautela, considerando-se as limitações do conhecimento atual diante da complexidade dos sistemas reais analisados.

Encerrada a análise exploratória, a seção seguinte apresenta a aplicação do método supervisionado PLS-DA sobre os mesmos dados, com o objetivo de aprofundar a discriminação entre as classes.

4.3 ANÁLISE DISCRIMINATE POR MÍNIMOS QUADRADOS MÍNIMOS PARCIAIS (PLS-DA)

Dando sequência à abordagem analítica, a presente seção apresenta a aplicação do método algoritmo supervisionado PLS-DA sobre os mesmos dados utilizados na modelagem exploratória por PCA. Diferentemente da PCA, que busca identificar padrões de variância sem considerar informações de classe, o algoritmo PLS-DA incorpora o conhecimento prévio das categorias investigadas, promovendo uma discriminação direcionada entre os grupos com base em suas características espectrais (SANTANA *et al.*, 2020).

Dando continuidade às análises multivariadas, a construção do modelo PLS-DA seguiu as diretrizes descritas na seção “Materiais e Métodos”, sendo a matriz de dados dividida em duas porções: uma de calibração, correspondente a 70% das amostras, e outra de validação, contendo os 30% restantes. A etapa de calibração teve por objetivo treinar o modelo a partir de um subconjunto representativo dos dados, permitindo que ele aprendesse a associar os espectros às respectivas classes previamente definidas. Por sua vez, a etapa de validação buscou testar a capacidade preditiva do modelo gerado, avaliando seu desempenho frente a dados independentes, não utilizados durante o treinamento (FERREIRA, 2015a).

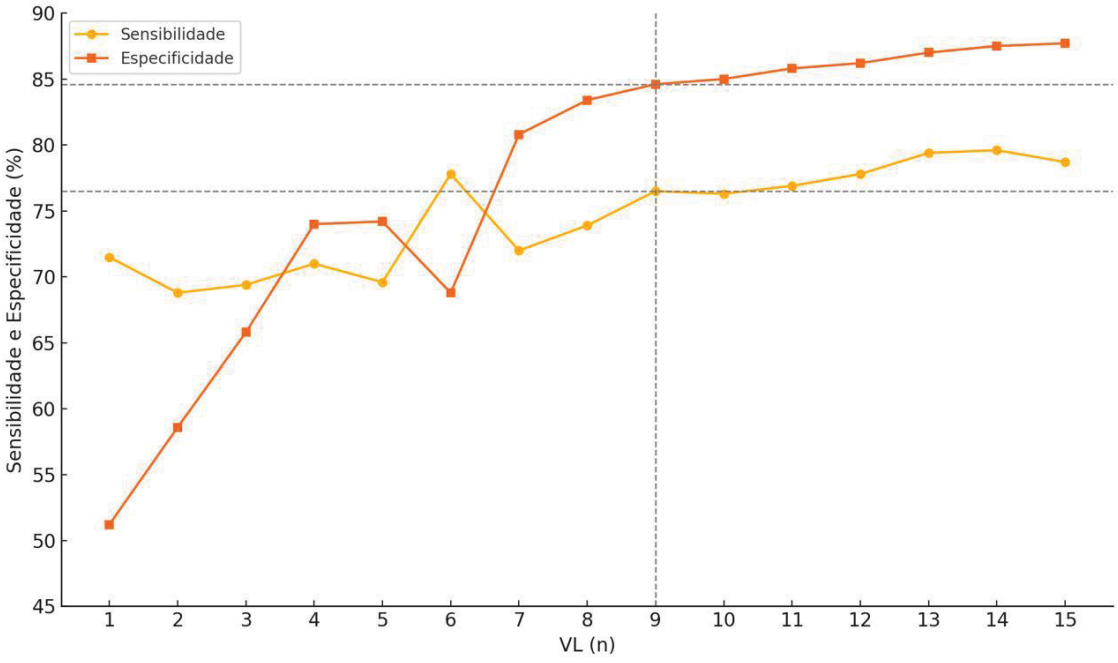
Importa destacar que os mesmos pré-tratamentos espectrais aplicados na análise exploratória por PCA foram igualmente empregados na construção do modelo PLS-DA. Essa uniformidade metodológica garante consistência analítica ao processo, além de possibilitar uma comparação mais direta entre os métodos exploratório e supervisionado adotados neste estudo.

À semelhança da PCA, a construção do modelo PLS-DA também requer a definição do número ideal de componentes a serem utilizados. No entanto, nesse

contexto, tais componentes são denominados variáveis latentes, pois representam combinações lineares das variáveis originais que não apenas explicam a variância presente no bloco de dados X, mas também se correlacionam diretamente com a variável de resposta Y. Essas variáveis latentes são extraídas de forma a maximizar a covariância entre X e Y, desempenhando, assim, um papel central na capacidade de discriminação entre as classes (SANTANA *et al.*, 2020).

Em outras palavras, enquanto a PCA busca identificar componentes que capturam a maior variabilidade possível do conjunto X, independentemente de qualquer informação de classe, a PLS-DA é orientada por essa informação e procura variáveis latentes que representem, da melhor forma possível, a relação entre os dados espectrais e as categorias de interesse. Dessa forma, as variáveis latentes funcionam como uma representação reduzida e orientada por classe do conjunto de dados, permitindo ao modelo reconhecer e explorar os padrões mais relevantes para a tarefa de classificação (FERREIRA, 2020).

FIGURA 20 - FIGURAS DE MÉRITO COM OS DADOS DE TREINAMENTO NA VALIDAÇÃO CRUZADA, DE ACORDO COM O NÚMERO DE VARIÁVEIS LATENTES



FONTE: A autora (2025).

LEGENDA: A linha tracejada vertical destaca o ponto correspondente a 9 variáveis latentes. A linha tracejada horizontal preta marca o valor de 76% para a sensibilidade e 84% para a especificidade.

A definição do número ótimo de variáveis latentes foi realizada com base na análise conjunta das figuras de mérito sensibilidade e especificidade, calculadas durante a validação cruzada do modelo. Essa avaliação foi conduzida em função da quantidade de variáveis latentes utilizadas, permitindo identificar o ponto em que ambos os parâmetros atingiram estabilização. Tal estratégia possibilitou uma escolha criteriosa, que favorecesse o equilíbrio entre a capacidade discriminativa e o desempenho preditivo do modelo. Ao todo, foram testados quinze modelos, e os valores de sensibilidade e especificidade obtidos encontram-se resumidos na FIGURA 20 na página anterior.

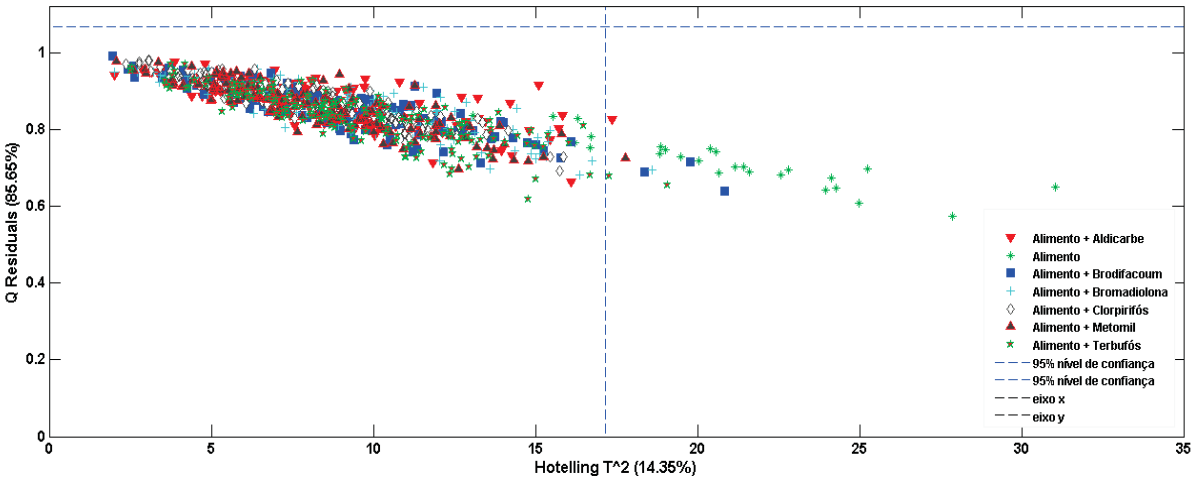
De acordo com os resultados obtidos, o número ótimo de variáveis latentes para o modelo de classificação PLS-DA foi igual a nove, ponto em que tanto a sensibilidade quanto a especificidade demonstraram estabilização. Nessa configuração, o modelo foi capaz de explicar 60,11% da variância presente nos dados originais. Os valores obtidos para as métricas de desempenho do modelo final foram de 84% de especificidade e 76% de sensibilidade.

Dando continuidade à análise, procedeu-se à avaliação de possíveis amostras anômalas, com o objetivo de verificar a presença de observações que pudessem comprometer o desempenho do modelo, também conhecidas como *outliers*. Tais desvios podem surgir por diversas razões, como erros de digitação, falhas experimentais ou, ainda, pela presença de espectros de baixa qualidade, frequentemente associados a ruídos ou interferências durante a aquisição. Uma vez identificadas, essas anomalias devem ser analisadas criticamente, sendo recomendada, sempre que possível, a repetição da coleta e análise da amostra. Nos casos em que tal repetição não seja viável, recomenda-se a exclusão da amostra, sobretudo se sua permanência afetar negativamente a estrutura do modelo. Ainda assim, é importante ressaltar que nem toda amostra atípica deve ser automaticamente descartada, uma vez que algumas delas podem refletir variabilidades reais e relevantes do sistema em estudo (FERREIRA, 2020).

Com o número de variáveis latentes previamente definido, foi construído um gráfico de T^2 de *Hotelling versus* Q *residual*, utilizando nove variáveis latentes (FIGURA 21), com o intuito de detectar possíveis *outliers*. Como já citada durante os resultados da PCA, essa abordagem é amplamente empregada na quimiometria para a identificação de amostras anômalas em contextos multivariados, pois permite avaliar simultaneamente tanto a posição relativa das amostras no espaço latente do

modelo quanto o grau em que são adequadamente explicadas por ele (FERREIRA, 2020; SANTANA *et al.*, 2020).

FIGURA 21 - GRÁFICO DE T^2 *HOTELLING VS.* Q RESIDUAL DO MODELO PLS-DA COM 9 VARIÁVEIS LATENTES



FONTE: A autora (2025).

LEGENDA: As linhas verticais e horizontais indicam os limites críticos (95%) para T^2 e Q residual, respectivamente.

No gráfico acima, o eixo X representa os valores de T^2 de *Hotelling*, a qual, neste caso, responde por 14,35% da variância modelada. Essa métrica reflete a variabilidade interna de cada amostra no espaço das variáveis latentes, sendo que valores mais elevados indicam posições mais extremas em relação ao centro da distribuição multivariada. Por sua vez, o eixo Y corresponde aos resíduos Q, que medem a fração da variabilidade não explicada pelo modelo. Valores elevados de Q indicam que a amostra apresenta um comportamento atípico em relação ao padrão descrito pelas variáveis latentes, sugerindo que parte de sua estrutura espectral não foi adequadamente representada pelo modelo.

A análise revelou que a maior parte das amostras se concentra na região inferior esquerda do gráfico, o que indica que foram bem modeladas e não exercem influência excessiva sobre a estrutura latente. No entanto, observou-se um agrupamento de amostras na região inferior direita, ou seja, com valores elevados de T^2, mas com resíduos Q dentro da faixa de aceitação. Verificou-se posteriormente que esse grupo corresponde, majoritariamente, às amostras de alimentos sem

contaminação por agrotóxicos. Tal distribuição pode ser atribuída à elevada homogeneidade espectral dessas amostras, que, por apresentarem um perfil químico uniforme e distinto das classes contaminadas, tendem a formar um agrupamento coeso, afastado do centróide da estrutura latente global.

Na sequência, a capacidade discriminatória do modelo PLS-DA foi avaliada individualmente para cada uma das classes de amostras, por meio do gráfico de Y predito, o qual ilustra a precisão das predições realizadas para cada categoria. Nas FIGURAS 22 a 28, observa-se como o modelo distingue as diferentes classes com base nos valores preditos, evidenciando sua aptidão para capturar as variações relevantes entre os grupos. Esse tipo de representação é fundamental para verificar se o modelo está realizando predições consistentes e confiáveis para cada classe de interesse (SANTANA *et al.*, 2020).

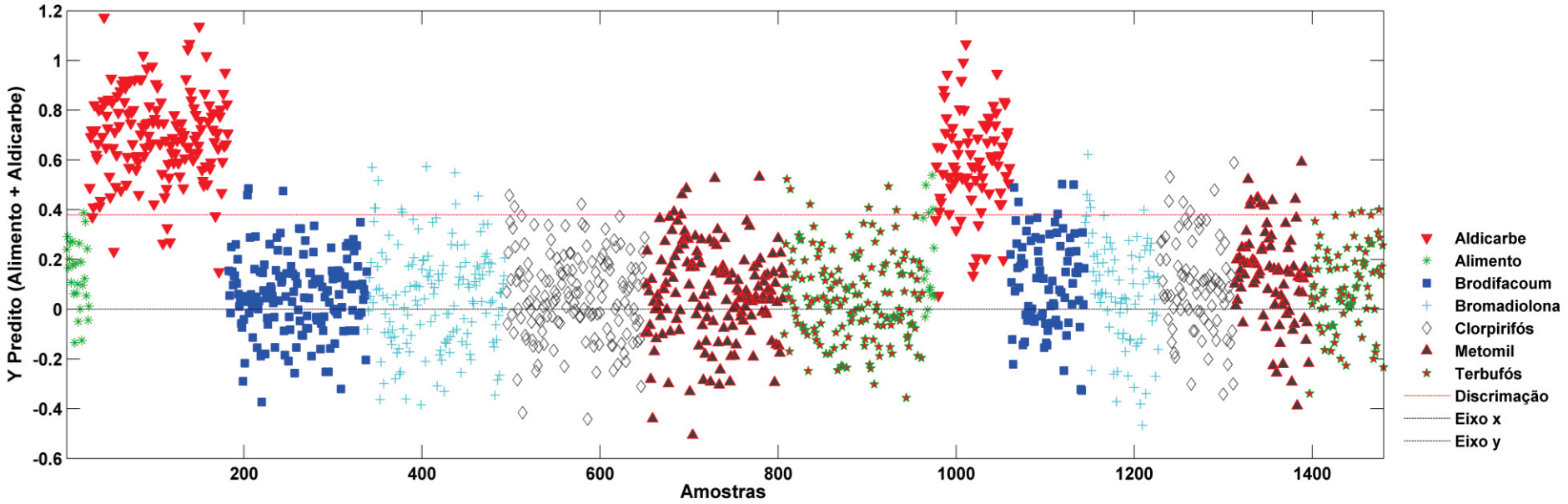
Com base na análise dos gráficos de Y predito, foi possível identificar e quantificar os casos de falsos positivos, falsos negativos, verdadeiros positivos e verdadeiros negativos para cada classe. A partir dessas informações, estimou-se a acurácia do modelo, métrica que fornece uma visão geral do seu desempenho global na tarefa de classificação. Os resultados obtidos estão organizados na TABELA 3 a seguir.

TABELA 3 - VALORES DE ACURÁCIA PARA CADA CLASSE DO MODELO UTILIZANDO O ALGORITMO PLS-DA

Classe	Acurácia
Alimento	98%
Alim + Aldicarbe	88%
Alim + Brodifacoum	86%
Alim + Bromadiolona	77%
Alim + Clorpirifós	86%
Alim + Metomil	86%
Alim + Terbufós	86%
MÉDIA	86%

FONTE: A autora (2025).

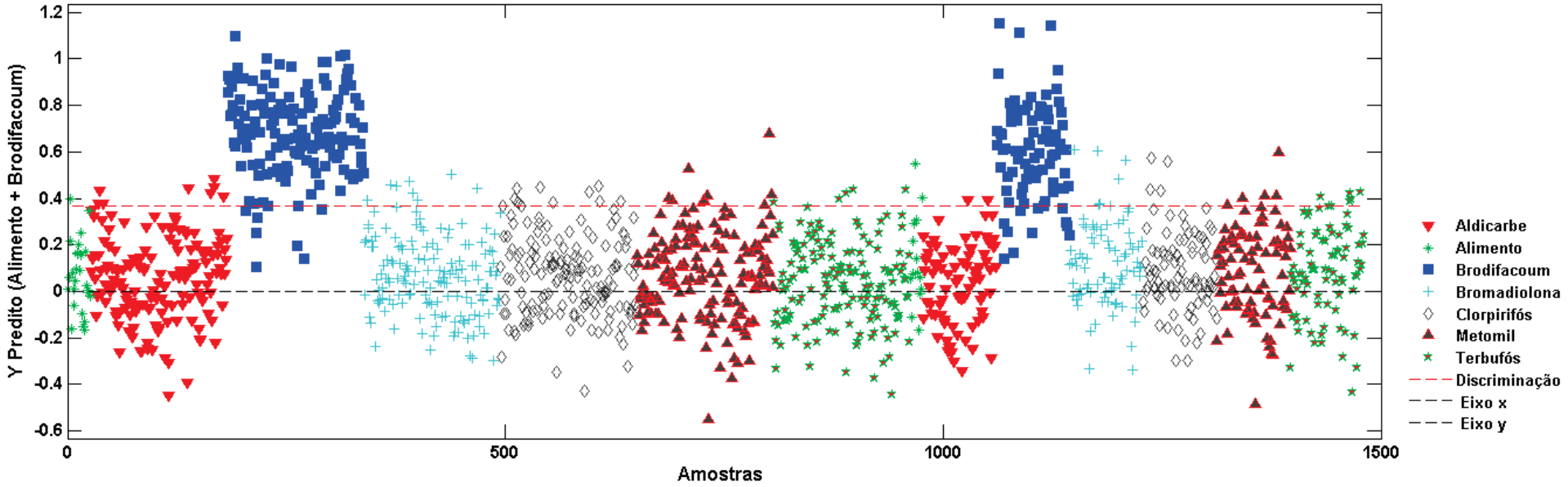
FIGURA 22 - GRÁFICO Y PREDITO PARA AVALIAÇÃO DA CAPACIDADE DISCRIMINATÓRIA DO MODELO PLS-DA COM NOVE VARIÁVEIS LATENTES PARA ALIMENTOS CONTAMINADOS COM ALDICARBE



FONTE: A autora (2025).

LEGENDA: Ao longo do eixo das abcissas, estão dispostas todas as varreduras espectrais obtidas no experimento, sendo cada grupo de alimento contaminado com agrotóxico representado por um símbolo distinto, de modo a facilitar a diferenciação visual entre as classes. As varreduras destinadas à etapa de calibração do modelo encontram-se posicionadas à esquerda do gráfico, enquanto aquelas reservadas à validação estão localizadas à direita. A linha vermelha horizontal representa o limiar de decisão do modelo, utilizado para distinguir as amostras classificadas como pertencente à classe “Alimento + aldicarbe”.

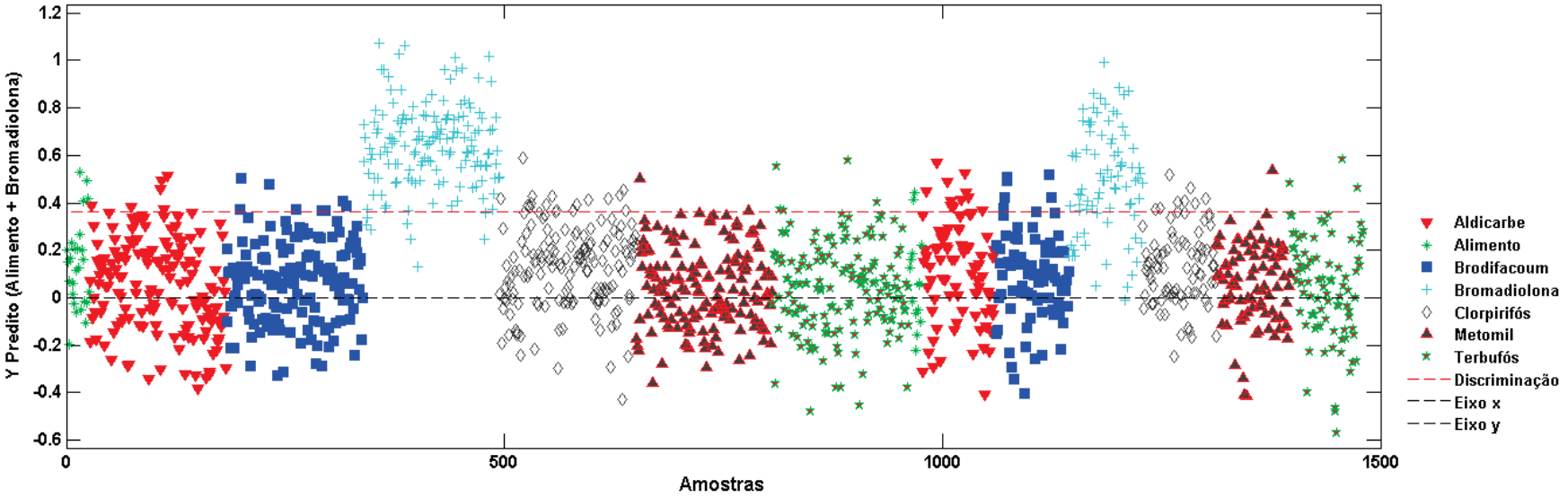
FIGURA 23 - GRÁFICO Y PREDITO PARA AVALIAÇÃO DA CAPACIDADE DISCRIMINATÓRIA DO MODELO PLS-DA COM NOVE VARIÁVEIS LATENTES PARA ALIMENTOS CONTAMINADOS COM BRODIFACOUM



FONTE: A autora (2025).

LEGENDA: Ao longo do eixo das abcissas, estão dispostas todas as varreduras espectrais obtidas no experimento, sendo cada grupo de alimento contaminado com agrotóxico representado por um símbolo distinto, de modo a facilitar a diferenciação visual entre as classes. As varreduras destinadas à etapa de calibração do modelo encontram-se posicionadas à esquerda do gráfico, enquanto aquelas reservadas à validação estão localizadas à direita. A linha vermelha horizontal representa o limiar de decisão do modelo, utilizado para distinguir as amostras classificadas como pertencente à classe “Alimento + brodifacoum”.

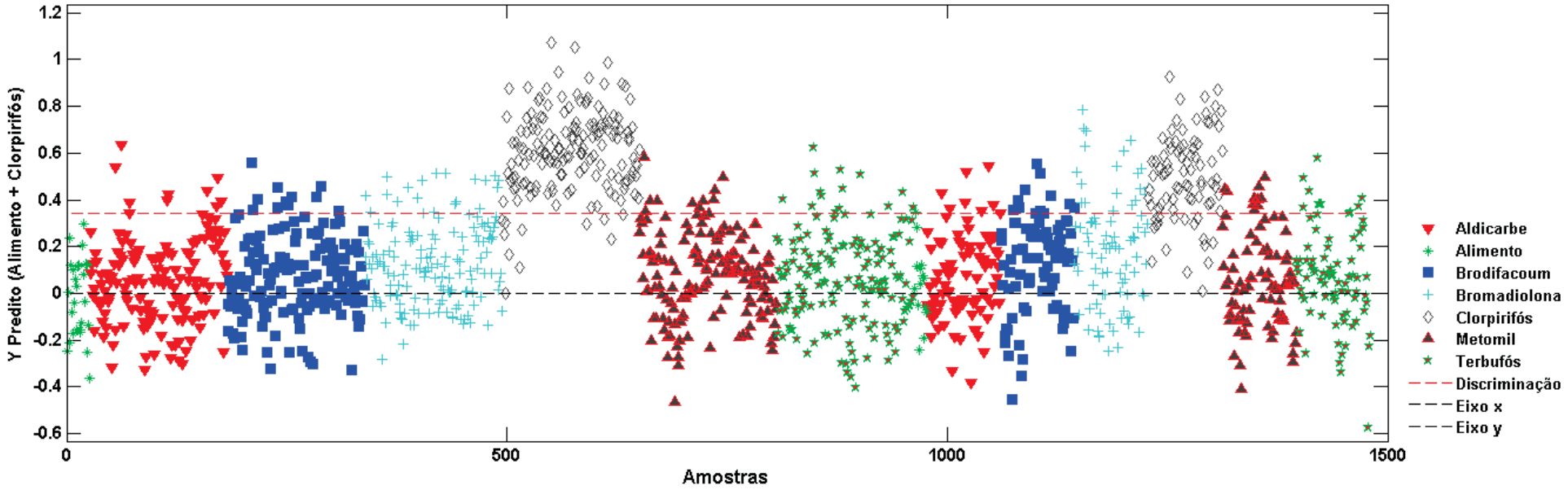
FIGURA 24 - GRÁFICO Y PREDITO PARA AVALIAÇÃO DA CAPACIDADE DISCRIMINATÓRIA DO MODELO PLS-DA COM NOVE VARIÁVEIS LATENTES PARA ALIMENTOS CONTAMINADOS COM BROMADIOLONA



FONTE: A autora (2025).

LEGENDA: Ao longo do eixo das abcissas, estão dispostas todas as varreduras espectrais obtidas no experimento, sendo cada grupo de alimento contaminado com agrotóxico representado por um símbolo distinto, de modo a facilitar a diferenciação visual entre as classes. As varreduras destinadas à etapa de calibração do modelo encontram-se posicionadas à esquerda do gráfico, enquanto aquelas reservadas à validação estão localizadas à direita. A linha vermelha horizontal representa o limiar de decisão do modelo, utilizado para distinguir as amostras classificadas como pertencente à classe “Alimento + bromadiolona”.

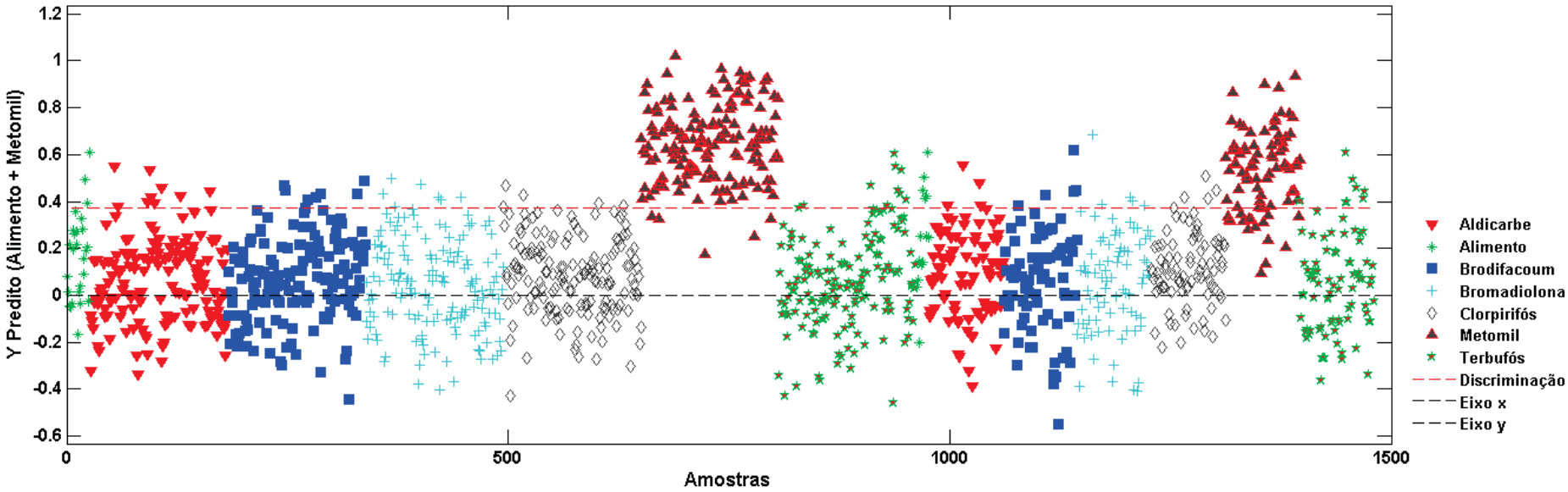
FIGURA 25 - GRÁFICO Y PREDITO PARA AVALIAÇÃO DA CAPACIDADE DISCRIMINATÓRIA DO MODELO PLS-DA COM NOVE VARIÁVEIS LATENTES PARA ALIMENTOS CONTAMINADOS COM CLORPIRIFÓS



FONTE: A autora (2025).

LEGENDA: Ao longo do eixo das abcissas, estão dispostas todas as varreduras espectrais obtidas no experimento, sendo cada grupo de alimento contaminado com agrotóxico representado por um símbolo distinto, de modo a facilitar a diferenciação visual entre as classes. As varreduras destinadas à etapa de calibração do modelo encontram-se posicionadas à esquerda do gráfico, enquanto aquelas reservadas à validação estão localizadas à direita. A linha vermelha horizontal representa o limiar de decisão do modelo, utilizado para distinguir as amostras classificadas como pertencente à classe “Alimento + clorpirifós”.

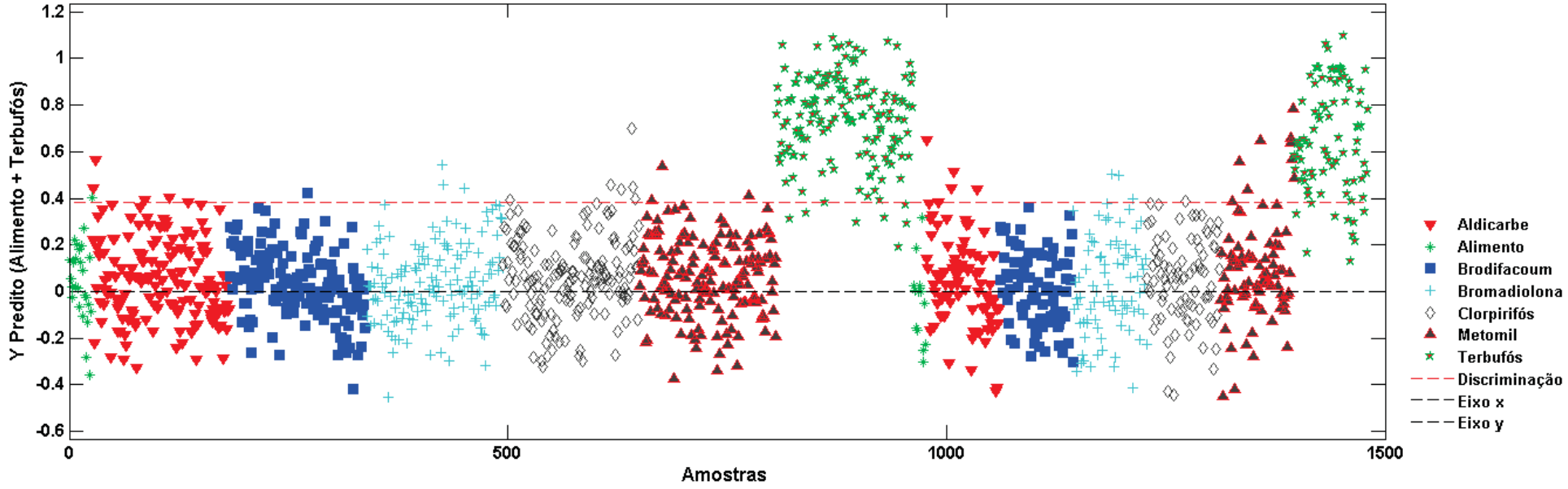
FIGURA 26 - GRÁFICO Y PREDITO PARA AVALIAÇÃO DA CAPACIDADE DISCRIMINATÓRIA DO MODELO PLS-DA COM NOVE VARIÁVEIS LATENTES PARA ALIMENTOS CONTAMINADOS COM METOMIL



FONTE: A autora (2025).

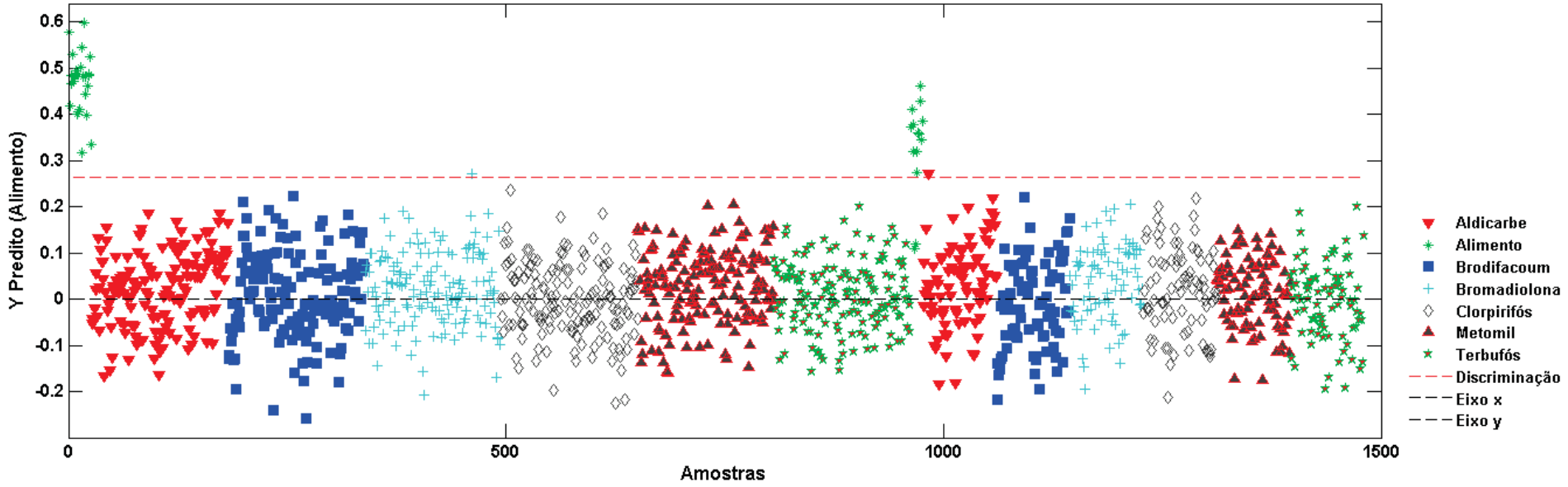
LEGENDA: Ao longo do eixo das abcissas, estão dispostas todas as varreduras espectrais obtidas no experimento, sendo cada grupo de alimento contaminado com agrotóxico representado por um símbolo distinto, de modo a facilitar a diferenciação visual entre as classes. As varreduras destinadas à etapa de calibração do modelo encontram-se posicionadas à esquerda do gráfico, enquanto aquelas reservadas à validação estão localizadas à direita. A linha vermelha horizontal representa o limiar de decisão do modelo, utilizado para distinguir as amostras classificadas como pertencente à classe “Alimento + metomil”.

FIGURA 27 - GRÁFICO Y PREDITO PARA AVALIAÇÃO DA CAPACIDADE DISCRIMINATÓRIA DO MODELO PLS-DA COM NOVE VARIÁVEIS LATENTES PARA ALIMENTOS CONTAMINADOS COM TERBUFÓS



FONTE: A autora (2025).
LEGENDA: Ao longo do eixo das abcissas, estão dispostas todas as varreduras espectrais obtidas no experimento, sendo cada grupo de alimento contaminado com agrotóxico representado por um símbolo distinto, de modo a facilitar a diferenciação visual entre as classes. As varreduras destinadas à etapa de calibração do modelo encontram-se posicionadas à esquerda do gráfico, enquanto aquelas reservadas à validação estão localizadas à direita. A linha vermelha horizontal representa o limiar de decisão do modelo, utilizado para distinguir as amostras classificadas como pertencente à classe “Alimento + terbufós”.

FIGURA 28 - GRÁFICO Y PREDITO PARA AVALIAÇÃO DA CAPACIDADE DISCRIMINATÓRIA DO MODELO PLS-DA COM NOVE VARIÁVEIS LATENTES PARA ALIMENTOS SEM CONTAMINAÇÃO



FONTE: A autora (2025).

LEGENDA: Ao longo do eixo das abcissas, estão dispostas todas as varreduras espectrais obtidas no experimento, sendo cada grupo de alimento contaminado com agrotóxico representado por um símbolo distinto, de modo a facilitar a diferenciação visual entre as classes. As varreduras destinadas à etapa de calibração do modelo encontram-se posicionadas à esquerda do gráfico, enquanto aquelas reservadas à validação estão localizadas à direita. A linha vermelha horizontal representa o limiar de decisão do modelo, utilizado para distinguir as amostras classificadas como pertencente à classe “Alimento sem contaminação”.

A análise comparativa entre o desempenho do modelo PLS-DA desenvolvido neste trabalho e os resultados reportados na literatura científica, evidencia a eficácia da abordagem adotada — sobretudo diante da elevada complexidade da matriz alimentar analisada, composta por pão, mortadela e diferentes analitos. Com acurácia de 86%, sensibilidade de 76% e especificidade de 84%, o modelo demonstrou desempenho satisfatório e consistente na discriminação de amostras contaminadas com distintos agrotóxicos, reforçando seu potencial como ferramenta.

Diversos estudos têm demonstrado o potencial do modelo PLS-DA em aplicações forenses, ainda que com desempenhos variados a depender da complexidade do sistema analisado. Qureshi *et al.* (2022) aplicaram FTIR associada ao PLS-DA para classificar tintas de canetas esferográficas, gel e *oil*-gel, conseguindo distinguir não apenas os diferentes tipos de tinta, mas também suas origens — nacionais ou importadas. O modelo apresentou sensibilidade de 60% para canetas do tipo gel no conjunto de validação, refletindo os desafios associados à semelhança espectral entre determinadas classes. Em outro exemplo, Pereira *et al.* (2016) utilizaram espectrometria de massa com ionização por spray de papel, combinada ao PLS-DA, para diferenciar amostras de cerveja. Após seleção de variáveis, o modelo atingiu uma taxa de acerto de 100%, evidenciando a acurácia da abordagem. Esses exemplos reforçam a versatilidade do PLS-DA em diferentes contextos forenses.

Na etapa seguinte deste estudo, foi conduzida uma análise dos gráficos de Variáveis Importantes na Projeção (VIP), obtidos a partir do modelo PLS-DA, com o objetivo de identificar as regiões espectrais mais relevantes para a discriminação dos diferentes grupos contaminados com agrotóxicos. Ainda que essa abordagem tenha caráter necessariamente simplificado e didático, ela se mostra valiosa por reforçar a consistência entre os dados espectrais e a classificação observada.

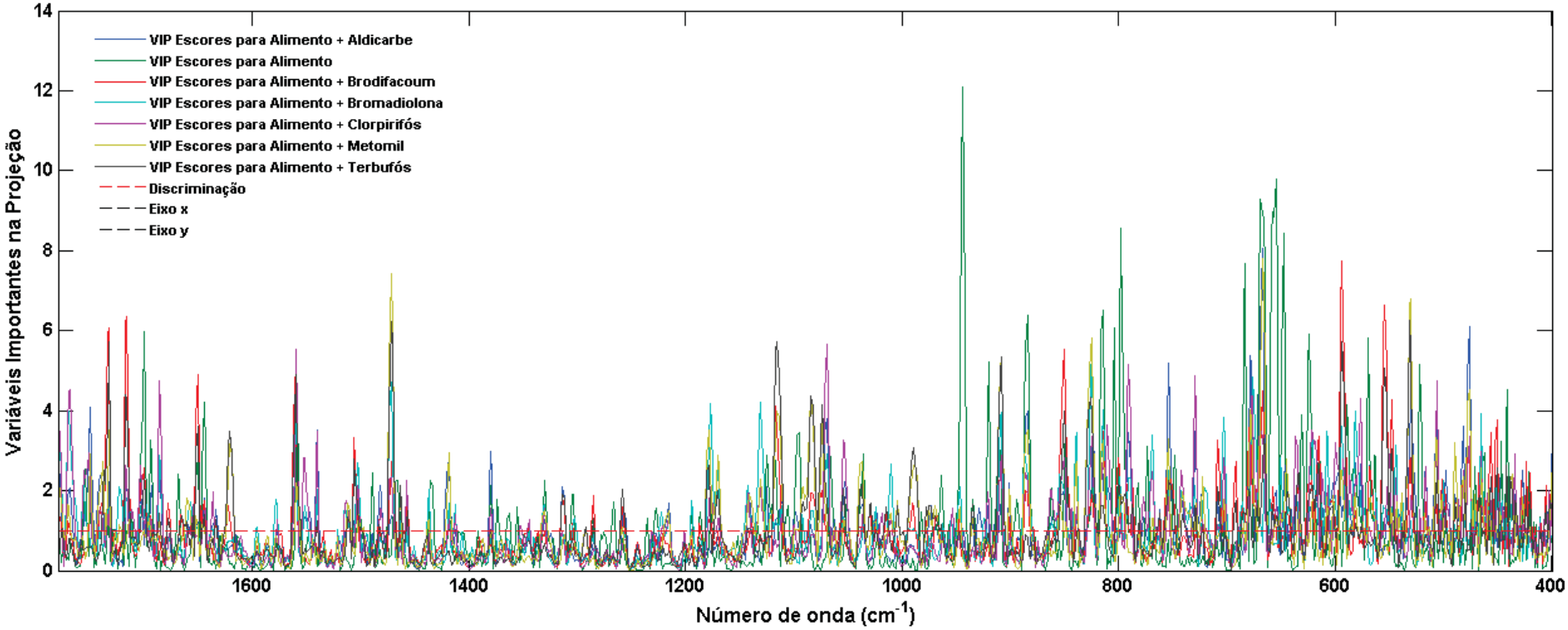
Deve-se ressaltar, entretanto, que o material em análise se trata de uma matriz alimentar altamente complexa, composta por uma mistura de pão, mortadela e contaminantes em diferentes proporções. Como tal, a interpretação espectral precisa ser entendida como limitada pela sobreposição de sinais, interferências da matriz e presença de múltiplas substâncias com grupos funcionais redundantes. Uma análise estrutural definitiva demandaria o uso de técnicas complementares, como a cromatografia gasosa ou líquida acoplada à espectrometria de massas.

Conforme evidenciado na FIGURA 29, na página seguinte, observa-se que toda a faixa espectral analisada se encontra acima do limiar de significância, o que indica que praticamente toda a região contribuiu para a separação entre as classes. Ainda assim, algumas bandas se destacam por sua expressiva relevância discriminante. A atribuição química das vibrações observadas nos picos mais proeminentes foi realizada com base nas funções espectroscópicas características dos grupos funcionais presentes em biomoléculas, conforme descrito na literatura especializada (SHURVELL, 2006; BARBOSA, 2007; COATES, 2006; PAVIA *et al.*, 2015), aliada à análise visual dos espectros de referência dos agrotóxicos empregados neste estudo, os quais se encontram disponíveis ao leitor nos apêndices desta dissertação.

No que se refere ao aldicarbe, observou-se a predominância de variáveis importantes nas faixas entre 900 e 700 cm^{-1}, correspondentes a deformações fora do plano de grupos metila e C–H, características da estrutura alifática simples desse composto. Além disso, em torno de 1400 cm^{-1}, destacam-se bandas associadas aos estiramentos das ligações C–N e C–O do grupo funcional carbamato.

Para o brodifacoum, um rodenticida cumarínico halogenado, os picos de maior relevância discriminante concentram-se nas regiões associadas às deformações dos anéis aromáticos substituídos, principalmente entre 1600 e 1450 cm^{-1}, bem como nas vibrações fora do plano localizadas entre 850 e 650 cm^{-1}, relacionadas ao padrão de substituição dos sistemas aromáticos policíclicos. A presença de dois átomos de bromo (Br) ainda confere um padrão espectral característico na faixa entre 650 e 500 cm^{-1}, onde se localizam os estiramentos da ligação C–Br. De forma semelhante, a bromadiolona, estruturalmente próximo ao brodifacoum, apresentou valores expressivos de VIP em faixas compatíveis com vibrações de carbonila (C=O) e com estiramentos C=C aromáticos, entre 1700 e 1600 cm^{-1}. Também se destacam as bandas entre 850 e 600 cm^{-1}, atribuídas a deformações fora do plano dos anéis aromáticos. Os sinais característicos das ligações C–Br voltam a se manifestar entre 600 e 500 cm^{-1}, servindo como marcadores vibracionais relevantes da estrutura halogenada da molécula.

FIGURA 29 - GRÁFICO DAS VARIÁVEIS IMPORTANTES NA PROJEÇÃO OBTIDOS PELO MODELO PLS-DA NA DISCRIMINAÇÃO ENTRE ALIMENTOS CONTAMINADOS COM DIFERENTES AGROTÓXICOS



FONTE: A autora (2025).

LEGENDA: Cada cor representa o perfil de importância de variáveis para uma classe específica, sendo os valores de número de onda plotados no eixo xe os valores de VIP no eixo y. Valores de VIP superiores a 1 (linha tracejada em vermelho) indicam regiões espectrais com maior contribuição para a discriminação das classes.

Para o clorpirifós, um inseticida organofosforado clorado, o gráfico de VIP destacou picos intensos na faixa de 1100 a 1000 cm^{-1}, correspondentes às vibrações das ligações P=O, P–O–C e P–O–Ar — elementos estruturais centrais da porção fosforada do composto. A região entre 850 e 600 cm^{-1} também apresentou elevada importância, englobando deformações fora do plano dos anéis aromáticos substituídos e, especialmente, os estiramentos da ligação C–Cl, cujas absorções típicas se concentram entre 400 e 600 cm^{-1}. Além disso, foi observada atividade vibracional em torno de 1550 cm^{-1}, compatível com os estiramentos das ligações C=C do anel aromático, indicando a contribuição da porção arílica do clorpirifós para sua diferenciação espectral.

No caso do metomil, pertencente à classe dos carbamatos, os picos mais expressivos se concentraram nas regiões próximas de 1500 e 1400 cm^{-1}, refletindo as vibrações do grupo funcional N–C=O, além das ligações C–N e C–O. Também foram identificados picos significativos abaixo de 900 cm^{-1}, atribuídos a deformações fora do plano de grupos metila e cadeias curtas, características da estrutura alifática do composto.

Por fim, o gráfico de VIP referente ao terbufós, um organofosforado do tipo fosforoditioato, evidenciou picos expressivos em regiões espectrais diretamente associadas aos seus grupos funcionais característicos. As bandas entre 1250 e 1000 cm^{-1} foram atribuídas aos estiramentos das ligações P=O e P–O–C. Já a região entre 600 e 500 cm^{-1} apresentou sinais relacionados às vibrações envolvendo fósforo e enxofre, reforçando seu papel como elemento discriminante adicional no modelo.

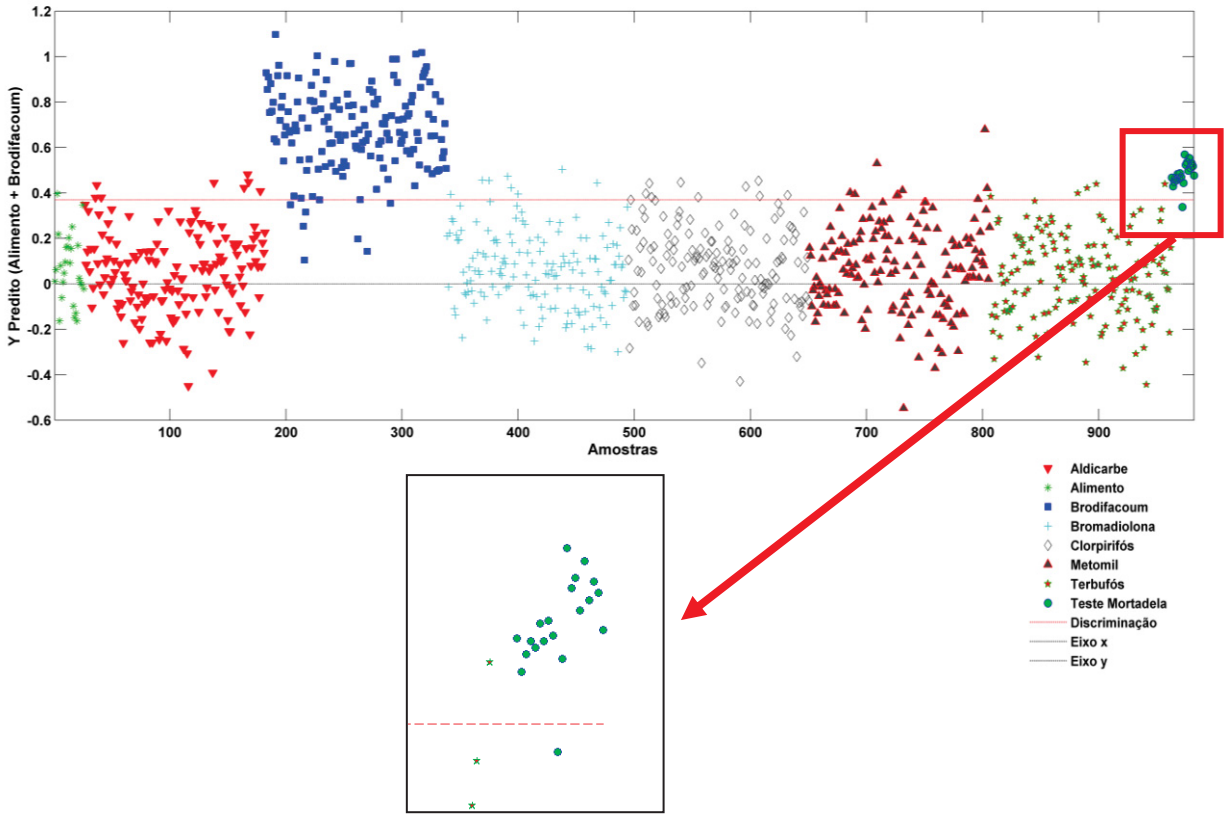
Finalizada a etapa de construção e validação, a aplicabilidade do modelo PLS-DA foi comprovada por meio de sua capacidade de prever corretamente a classe de amostras desconhecidas, provenientes de matriz semelhante, mas não idêntica àquelas utilizadas na calibração. Todo o esforço investido nos procedimentos de pré-tratamento e na otimização do modelo teve como objetivo assegurar uma boa resposta em contextos práticos.

Dessa forma, apresentam-se a seguir os resultados obtidos a partir da aplicação da metodologia aos espectros coletados de duas amostras reais, oriundas de casos periciais encaminhados à Seção de Química Forense da Polícia Científica do Paraná. A fase de previsão exigiu o cumprimento rigoroso de condições experimentais semelhantes às adotadas na calibração. As amostras de teste foram

preparadas utilizando os mesmos protocolos e instrumentos empregados na fase de construção do modelo, e seus espectros foram submetidos ao mesmo tratamento aplicado aos dados de calibração, assegurando coerência na projeção dos resultados.

A primeira análise de teste foi conduzida com uma amostra de pão contaminado com brodifacoum, cuja presença havia sido previamente confirmada. A FIGURA 30 apresenta os resultados da projeção dos valores preditos de Y para a classe “Alimento + Brodifacoum”, evidenciando o desempenho do modelo na correta classificação dessa amostra de teste.

FIGURA 30 - PROJEÇÃO DO Y PREDITO DA CLASSE “ALIMENTO + BRODIFACOUM” COM O RESULTADO DA AMOSTRA TESTE DE PÃO CONTAMINADO COM BRODIFACOUM



FONTE: A autora (2025).

LEGENDA: Representação gráfica da predição do modelo, destacando a amostra teste (quadrado vermelho) e sua respectiva projeção no espaço do modelo de classificação. No gráfico superior, observa-se a distribuição dos espectros por classe, codificados por cores distintas, correspondentes aos diferentes agrotóxicos e ao grupo controle. A região ampliada evidencia o posicionamento da amostra de teste, cuja classificação foi atribuída pelo modelo à classe “Alimento + Brodifacoum”.

Observa-se que 19 dos 20 espectros analisados foram corretamente classificados como pertencentes à referida classe, evidenciando a elevada capacidade discriminatória do modelo para esse analito.

A amostra de pão proveniente do material periciado, por se tratar de uma matriz complexa e distinta daquela utilizada na etapa de calibração, poderia, em princípio, dificultar a diferenciação espectral. No entanto, a correta alocação da maioria dos espectros no espaço da classe “Brodifacoum” demonstra que o modelo foi eficaz em capturar padrões espectrais característicos do rodenticida.

O único espectro classificado de forma incorreta pode ser atribuído a flutuações pontuais, ruído instrumental ou heterogeneidade local da amostra. Ainda assim, esse desvio isolado não compromete o desempenho geral do modelo, que se mostrou altamente sensível e assertivo.

Um ponto relevante a ser considerado é que, com o passar do tempo, alterações nos equipamentos, nas condições ambientais ou nos procedimentos de amostragem podem comprometer a capacidade preditiva do modelo. Assim, recomenda-se o uso regular de amostras, portanto, de referência como forma de monitoramento da estabilidade do modelo. Os resultados obtidos com essas amostras devem ser comparados com valores futuros, permitindo a avaliação contínua da confiabilidade do sistema e a identificação da necessidade de revalidação, quando necessário. Desse modo, compreende-se que a aplicação de um modelo PLS-DA não se limita à sua validação inicial, mas integra um processo contínuo de verificação da qualidade preditiva (FERREIRA, 2020). Essa abordagem assegura que o modelo se mantenha robusto, sensível e confiável frente às variações inerentes ao ambiente analítico e às características das matrizes alimentares reais.

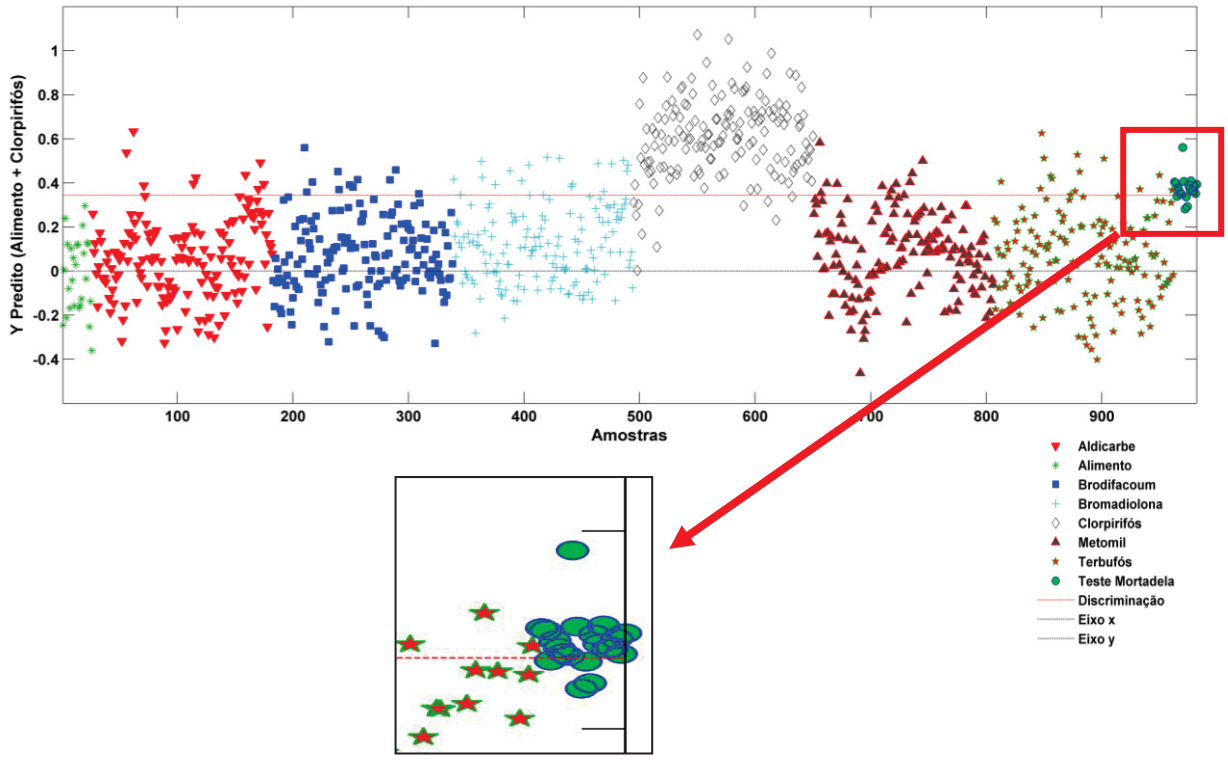
A segunda análise de teste envolveu a previsão de uma amostra de mortadela contaminada com terbufós. A FIGURA 30 apresenta os resultados obtidos a partir da aplicação do modelo PLS-DA a essa amostra, cuja contaminação foi previamente confirmada por cromatografia gasosa acoplada à espectrometria de massas.

A análise dessa amostra revela um aspecto relevante da capacidade discriminatória do modelo. Embora os espectros tenham sido corretamente agrupados quanto ao grupo químico ao qual pertencem, observou-se uma limitação na identificação precisa do analito. Esperava-se, inicialmente, que os espectros

fossem alocados na classe “Alimento + Terbufós”; no entanto, ao serem projetados no espaço do modelo, constatou-se que todas as amostras se posicionaram abaixo da linha de discriminação estabelecida para essa classe, indicando que elas não atendem plenamente aos critérios espectrais característicos do terbufós.

Por outro lado, quando essas mesmas amostras são analisadas sob a perspectiva da classe “Alimento + Clorpirifós”, observa-se que 16 dos 20 espectros foram classificados como pertencentes a essa categoria. Tal comportamento sugere uma confusão na definição específica do analito de interesse, embora o modelo tenha sido capaz de reconhecer corretamente o grupo químico ao qual ele pertence.

FIGURA 31 - PROJEÇÃO DO Y PREDITO DA CLASSE “ALIMENTO + TERBUFÓS” COM O RESULTADO DA AMOSTRA TESTE MORTADELA CONTAMINADA COM CLORPIRIFÓS



FONTE: A autora (2025).

LEGENDA: Representação gráfica da predição do modelo, destacando a amostra de teste (quadrado vermelho) e sua respectiva projeção no espaço do modelo de classificação. No gráfico superior, observa-se a distribuição dos espectros por classe, codificados por cores distintas, correspondentes aos diferentes agrotóxicos e ao grupo controle (Alimento). A região ampliada evidencia o posicionamento da amostra de teste, cuja classificação foi atribuída pelo modelo à classe “Alimento + Clorpirifós”.

Diante desse cenário, recomenda-se a adoção de estratégias complementares visando ao aprimoramento do desempenho do modelo, como, por exemplo, a aplicação de técnicas de seleção de variáveis, capazes de identificar e isolar as regiões espectrais com maior poder discriminativo. Além disso, a inclusão de matrizes alimentares em diferentes estágios de conservação poderia representar um avanço importante, uma vez que simula de forma mais realista as condições em que os alimentos periciados são frequentemente encontrados, muitas vezes em estado de degradação. Essa abordagem ampliaria a robustez do modelo frente às variações inerentes à rotina pericial.

Adicionalmente, serão exploradas abordagens com o uso de outros algoritmos classificatórios, visando superar limitações observadas nas predições com o modelo PLS-DA - em especial, as falhas na identificação correta das amostras de mortadela contaminadas com terbufós. Ressalta-se, ainda, que a metodologia desenvolvida até este ponto foi aplicada apenas a duas matrizes específicas (pão e mortadela), mas o conjunto de dados abrange cinco grupos distintos de alimentos, cuja complexidade e variabilidade ainda não foram plenamente exploradas. Assim, a aplicação de novos algoritmos permitirá uma avaliação mais ampla de todo o conjunto amostral, ampliando o potencial discriminatório da abordagem proposta. Essa estratégia será detalhada na seção seguinte deste trabalho.

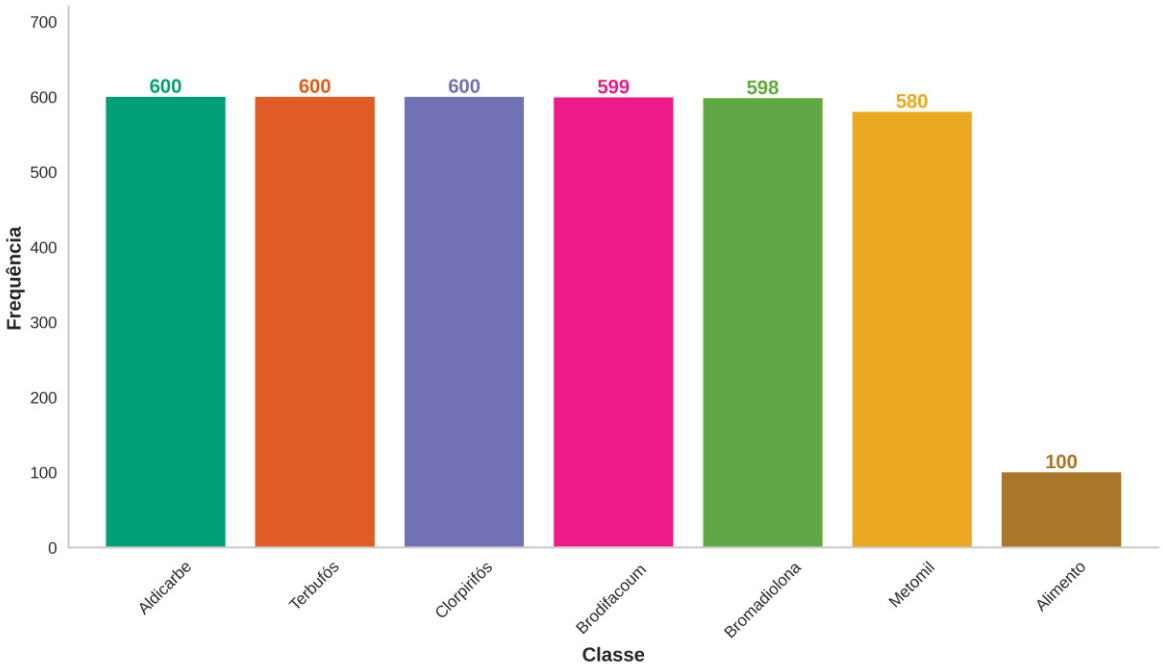
4.4 OUTROS ALGORITMOS DE APRENDIZADO DE MÁQUINA

Dando continuidade à investigação analítica, foram explorados algoritmos modernos de aprendizado de máquina disponibilizados pela biblioteca PyCaret, utilizando a linguagem Python executada na plataforma Google Colab, com o objetivo de desenvolver modelos classificatórios preditivos mais robustos e generalizáveis. Para essa etapa, diferentemente do procedimento adotado na modelagem com PLS-DA, optou-se por utilizar todas as matrizes alimentares preparadas — incluindo mistura de alimentos, pão, mortadela, arroz e carne moída — no processo de treinamento dos modelos. A inclusão conjunta dessas matrizes elevou consideravelmente a complexidade da base de dados, impondo um desafio

adicional aos algoritmos, especialmente se considerado que, nas etapas anteriores de análise exploratória por PCA e modelagem supervisionada por PLS-DA, os melhores desempenhos foram alcançados apenas quando se utilizaram exclusivamente as matrizes de pão e mortadela.

A primeira etapa dessa nova abordagem envolveu a visualização dos espectros sobrepostos de todas as classes, com o objetivo de realizar uma inspeção visual preliminar e excluir eventuais amostras anômalas. Conforme já verificado na análise exploratória por PCA, identificaram-se espectros atípicos — sobretudo na classe referente à mistura de alimentos contaminada com metomil —, os quais foram removidos por apresentarem comportamento discrepante. Adicionalmente, foram excluídos dois espectros da classe “alimento contaminado com bromadiolona” e um da classe “alimento contaminado com brodifacoum”, em virtude de variações significativas em relação ao padrão da respectiva classe. O total final de espectros por classe, após as exclusões realizadas, encontra-se representado na FIGURA 32.

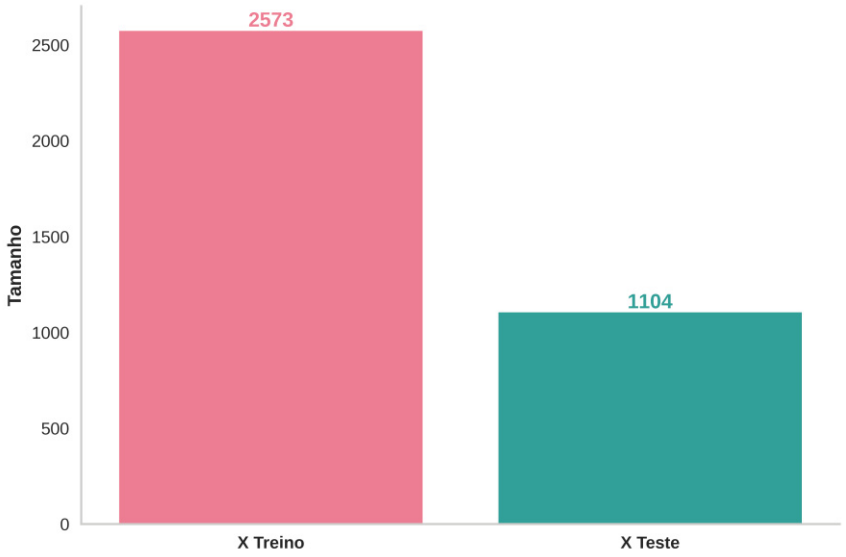
FIGURA 32 - DISTRIBUIÇÃO DO NÚMERO DE ESPECTROS POR GRUPO UTILIZADOS NA CONSTRUÇÃO DOS MODELOS CLASSIFICATÓRIOS DA BIBLIOTECA PYCARET



FONTE: A autora (2025).

Com a matriz de dados devidamente ajustada, procedeu-se à divisão do conjunto em amostras de treino e teste de forma automatizada, adotando-se a proporção de 70% para o treinamento e 30% para a validação do modelo. A distribuição das amostras em cada grupo pode ser visualizada na FIGURA 33, que apresenta a quantidade de espectros destinados a cada fase do processo de modelagem.

FIGURA 33 - DISTRIBUIÇÃO DO NÚMERO DE AMOSTRAS NOS CONJUNTOS DE TREINAMENTO E TESTE UTILIZADOS NA CONSTRUÇÃO DOS MODELOS CLASSIFICATÓRIOS DA BIBLIOTECA PYCARET



FONTE: A autora (2025).

Após o pré-processamento automatizado realizado pela biblioteca PyCaret, foram testados 15 algoritmos de classificação supervisionada disponíveis na plataforma, os quais foram automaticamente avaliados e ranqueados com base na acurácia. Os resultados dessa etapa estão sintetizados na TABELA 4 da página seguinte. Dentre os modelos testados, aqueles que apresentaram o melhor desempenho preditivo foram o LDA (*Linear Discriminant Analysis*) e o LightGBM (*Light Gradient Boosting Machine*).

Observa-se que o modelo LDA apresentou o desempenho global mais expressivo entre os 15 algoritmos avaliados, alcançando acurácia de 82,07%, AUC

de 0,9711 e métricas igualmente elevadas para recall, precisão, F1-score, Kappa e MCC, todas situadas na faixa entre 0,78 e 0,82. Apesar de sua simplicidade e natureza linear, o LDA demonstrou excelente desempenho, com tempo de treinamento reduzido (0,42 segundos), evidenciando sua baixa complexidade computacional. Tal resultado pode ser atribuído ao fato de que, em conjuntos de dados estruturados e estatisticamente bem comportados, modelos lineares ainda se mostram altamente competitivos (HEINTZ *et al.*, 2023).

TABELA 4 - DESEMPENHO DOS MODELOS DE CLASSIFICAÇÃO SUPERVISIONA AVALIADOS PELA BIBLIOTECA PYCARET

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lda	Linear Discriminant Analysis	0.8207	0.9711	0.8207	0.8204	0.8204	0.7868	0.7868	0.4200
lightgbm	Light Gradient Boosting Machine	0.6540	0.9183	0.6540	0.6567	0.6502	0.5874	0.5878	12.6800
xgboost	Extreme Gradient Boosting	0.6522	0.9214	0.6522	0.6539	0.6485	0.5854	0.5857	32.1700
rf	Random Forest Classifier	0.6313	0.9111	0.6313	0.6372	0.6264	0.5604	0.5612	1.0000
et	Extra Trees Classifier	0.6286	0.9064	0.6286	0.6295	0.6245	0.5573	0.5577	0.3500
gbc	Gradient Boosting Classifier	0.5888	0.8917	0.5888	0.5852	0.5840	0.5100	0.5105	534.1300
dt	Decision Tree Classifier	0.5118	0.7111	0.5118	0.5126	0.5118	0.4201	0.4202	3.0100
knn	K Neighbors Classifier	0.4873	0.8178	0.4873	0.4974	0.4867	0.3891	0.3911	0.0900
ridge	Ridge Classifier	0.3922	0.0000	0.3922	0.3707	0.3745	0.2743	0.2766	0.1300
qda	Quadratic Discriminant Analysis	0.3732	0.6469	0.3732	0.1937	0.2473	0.2531	0.2907	0.6500
lr	Logistic Regression	0.3424	0.7234	0.3424	0.3279	0.3229	0.2145	0.2172	3.7800
ada	Ada Boost Classifier	0.3415	0.6984	0.3415	0.3404	0.3384	0.2157	0.2164	14.2300
svm	SVM - Linear Kernel	0.2636	0.0000	0.2636	0.3348	0.1467	0.1203	0.1690	1.0100
nb	Naive Bayes	0.2364	0.5911	0.2364	0.1777	0.1772	0.0877	0.0981	0.1000
dummy	Dummy Classifier	0.1630	0.0000	0.1630	0.0266	0.0457	0.0000	0.0000	0.0700

FONTE: A autora (2025).

LEGENDA: Os modelos estão listados na primeira coluna, seguidos pelas métricas de avaliação. A coluna **Accuracy** representa a proporção de classificações corretas em relação ao total de predições realizadas. A métrica **AUC** (Área sob a Curva ROC) expressa a capacidade do modelo em distinguir entre as classes. O **Recall**, também conhecido como sensibilidade, corresponde à fração de verdadeiros positivos corretamente identificados. A Precisão (**Prec.**) indica a proporção de verdadeiros positivos entre todas as predições classificadas como positivas. O valor de **F1** refere-se à média harmônica entre precisão e recall. O **Kappa** mede o grau de concordância entre as classificações do modelo e os valores reais, ajustando as previsões esperadas pelo acaso. Já o **MCC** (Coeficiente de Correlação de Matthews) é uma métrica abrangente que considera todos os elementos da matriz de confusão. Por fim, a coluna **TT (Sec)** indica o tempo total, em segundos, necessário para o treinamento e validação de cada modelo. Todas essas métricas foram destacadas com sombreamento amarelo para o melhor modelo.

Em contraste, o modelo LightGBM, embora tenha alcançado uma AUC elevada (0,9183), apresentou acurácia inferior (65,40%) e métricas de concordância consideravelmente mais baixas em comparação ao LDA (Kappa: 0,5874; MCC: 0,5878), além de demandar um tempo de treinamento quase 30 vezes superior.

Modelos mais complexos, como XGBoost, *Random Forest* e *Gradient Boosting Classifier*, também não superaram o desempenho do LDA, ainda que tenham apresentado resultados razoáveis em algumas métricas. Esse cenário reforça a ideia de que, em determinadas bases de dados, modelos mais simples podem superar algoritmos sofisticados (GHOJOGH; CROWLEY, 2019).

Diante disso, os dois modelos com melhor desempenho — LDA e LightGBM — foram selecionados para uma nova etapa de calibração, desta vez com validação cruzada explícita. Ambos foram treinados com o conjunto de treino e avaliados com base em métricas de desempenho padrão, permitindo uma comparação mais robusta da capacidade preditiva dos modelos após o ajuste. As figuras de mérito obtidas após essa etapa estão apresentadas na TABELA 5 a seguir.

TABELA 5 - DESEMPENHO DOS DOIS MELHORES MODELOS APÓS A VALIDAÇÃO CRUZADA

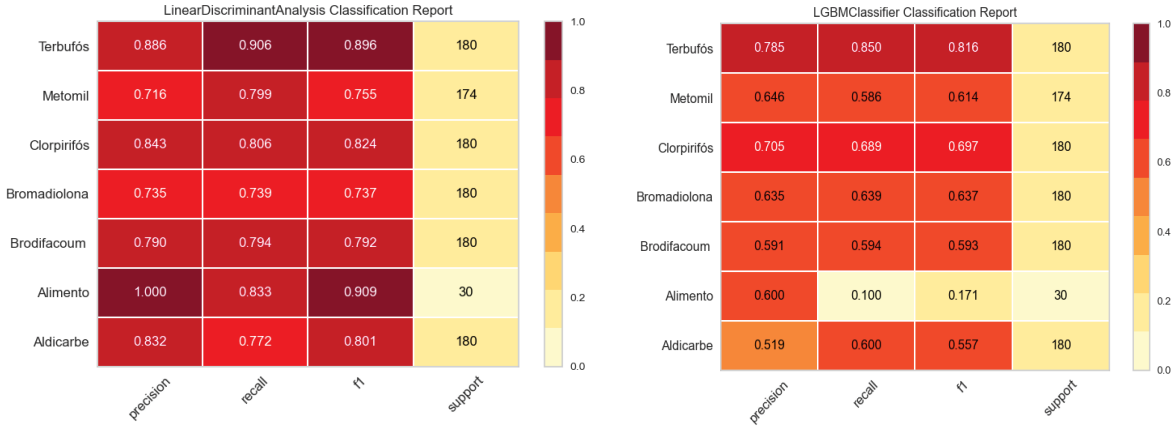
Modelo	Acurácia Treino	Acurácia Teste	*F1 Score*	*Precision*	*Recall*
Linear Discriminant Analysis	0.99	0.81	0.81	0.81	0.81
LightGBM Classifier	1.00	0.67	0.67	0.67	0.67

FONTE: A autora (2025).

Após a identificação dos modelos com melhor desempenho geral, prosseguiu-se com uma análise aprofundada desses classificadores por meio de representações gráficas. Tais visualizações complementam as informações numéricas previamente apresentadas, permitindo avaliar de forma mais detalhada o comportamento dos modelos em relação à classificação das diferentes classes de amostras. A utilização de relatório de classificação, curva ROC, matrizes de confusão, confiabilidade da previsão e gráficos de importância das variáveis possibilitou uma compreensão mais refinada sobre a capacidade discriminativa, os padrões de erro e os atributos espectrais mais relevantes para a tomada de decisão do modelo.

Como verificado nos relatórios de classificação da FIGURA 34, o modelo LDA apresentou desempenho superior em relação ao LightGBM na classificação das diferentes classes de amostras. As métricas obtidas pelo LDA evidenciam sua capacidade de identificar corretamente tanto as amostras contaminadas com diferentes agrotóxicos quanto aquelas não contaminadas. Especificamente, a classe “Alimento” obteve desempenho perfeito em termos de precisão, com valor igual a 1.000, indicando que o modelo foi capaz de reconhecer corretamente todas as amostras não contaminadas, ou seja, sem ocorrência de falsos positivos para essa classe.

FIGURA 34 - RELATÓRIOS DE CLASSIFICAÇÃO DOS MODELOS LDA E LIGHTGBM APLICADOS ÀS AMOSTRAS ESPECTRAIS



FONTE: A autora (2025).

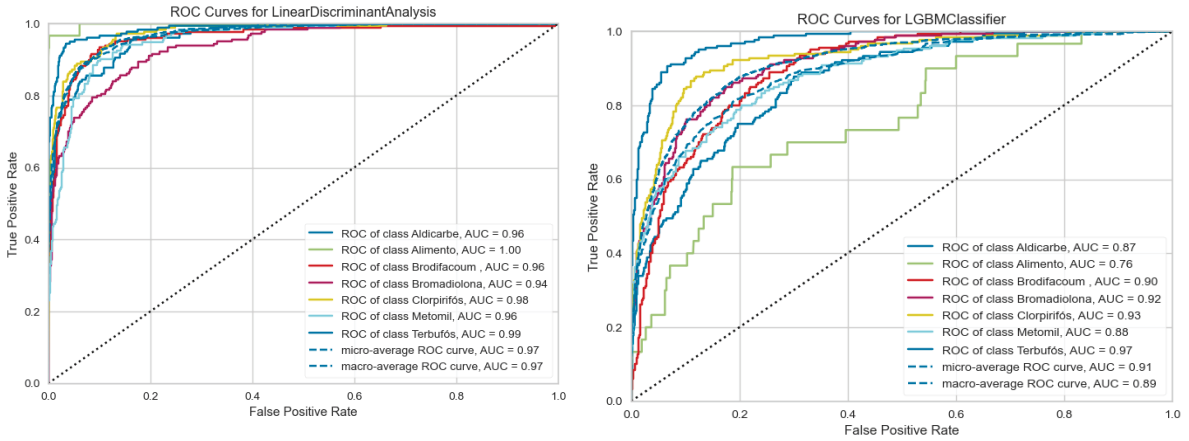
LEGENDA: Relatórios de desempenho dos modelos Linear Discriminant Analysis (LDA), na parte superior, e LightGBM, na parte inferior, representados por mapas de calor. Cada linha corresponde a uma das sete classes de amostras analisada. As colunas indicam as métricas de avaliação: precisão (*precision*), sensibilidade (*recall*), f1-score e número de amostras por classe (*support*). Os valores estão representados em escala de cor, variando de tonalidades claras (baixa performance) a escuras (alta performance).

Em contraste, o modelo LightGBM apresentou desempenho mais irregular. Embora tenha alcançado resultados razoáveis para algumas classes, como Terbufós e Clorpirifós, demonstrou dificuldade considerável na classificação de outras categorias. O caso mais crítico foi o da classe “Alimento”, cuja taxa de acerto foi extremamente baixa. O *recall* de apenas 0.100 indica que apenas 10% das amostras

não contaminadas foram corretamente identificadas como tal. Esses resultados sugerem que, apesar da sofisticação do algoritmo, o LightGBM foi mais sensível a variações ou ruídos nos dados, além de possivelmente ter sido afetado pelo desbalanceamento entre as classes.

Complementando essa análise de desempenho, a FIGURA 35 apresenta as curvas ROC (*Receiver Operating Characteristic*) geradas para os dois modelos de classificação, representando a relação entre a taxa de verdadeiros positivos (sensibilidade) e a taxa de falsos positivos (especificidade). Quanto maior a área sob a curva (AUC – *Area Under the Curve*), melhor o desempenho do modelo na distinção entre as diferentes classes.

FIGURA 35 - CURVAS ROC DOS MODELOS LDA E LIGHTGBM APLICADOS ÀS AMOSTRAS ESPECTRAIS



FONTE: A autora (2025).

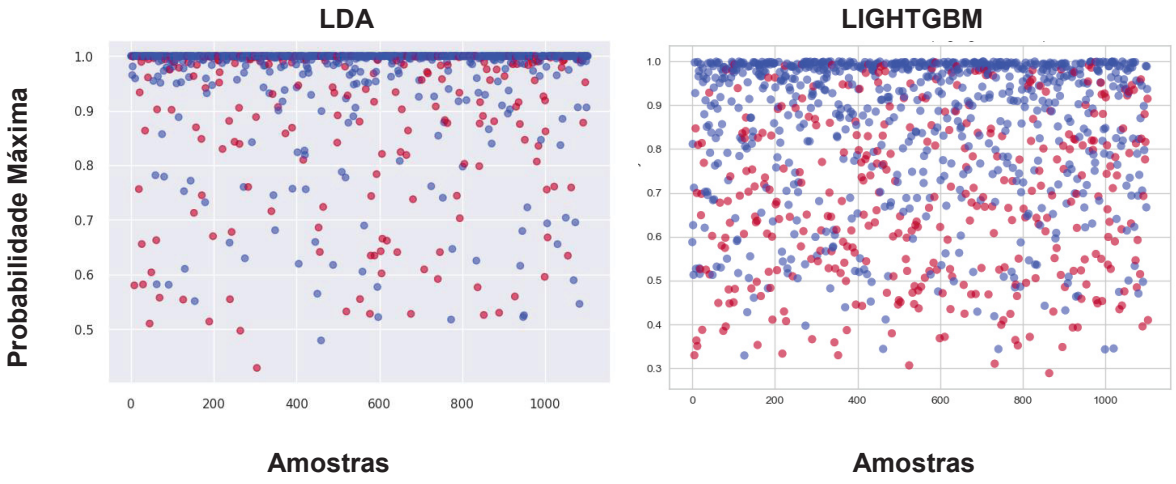
LEGENDA: Relatórios de desempenho dos modelos LDA, na parte superior, e LightGBM, na parte inferior, representados por mapas de calor. Cada linha corresponde a uma das sete classes de amostras analisada. As colunas indicam as métricas de avaliação: precisão (*precision*), sensibilidade (*recall*), f1-score e número de amostras por classe (*support*). Os valores estão representados em escala de cor, variando de tonalidades claras (baixa performance) a escuras (alta performance).

O modelo LDA demonstrou desempenho superior e mais homogêneo nas curvas ROC, com valores de AUC variando entre 0,96 e 1,00 para todas as classes. A classe “Alimento” obteve AUC de 1,00, sugerindo excelente separabilidade em relação às demais, em contraste com o desempenho observado para o LightGBM.

Neste último, os valores de AUC oscilaram entre 0,76, para a classe “Alimento”, e 0,92, para a classe “Bromadiolona”, com desempenhos intermediários nas demais categorias. Já com relação ao LightGBM, apesar de se tratar de um algoritmo mais moderno, apresentou desempenho mais modesto na distinção da classe do alimento sem contaminação, evidenciando maior dificuldade na diferenciação dessa categoria em relação às demais, conforme já apontado no relatório de classificação.

Além da métrica AUC, os dois modelos também foram comparados com base nos gráficos de “Confiabilidade da Previsão” (FIGURA 36). Esses gráficos revelam diferenças marcantes na distribuição das probabilidades máximas atribuídas às previsões corretas, indicando distintos níveis de confiança na classificação realizada por cada modelo.

FIGURA 36 - CONFIABILIDADE DA PREVISÃO DOS MODELOS LDA E LIGHTGBM



FONTE: A autora (2025).

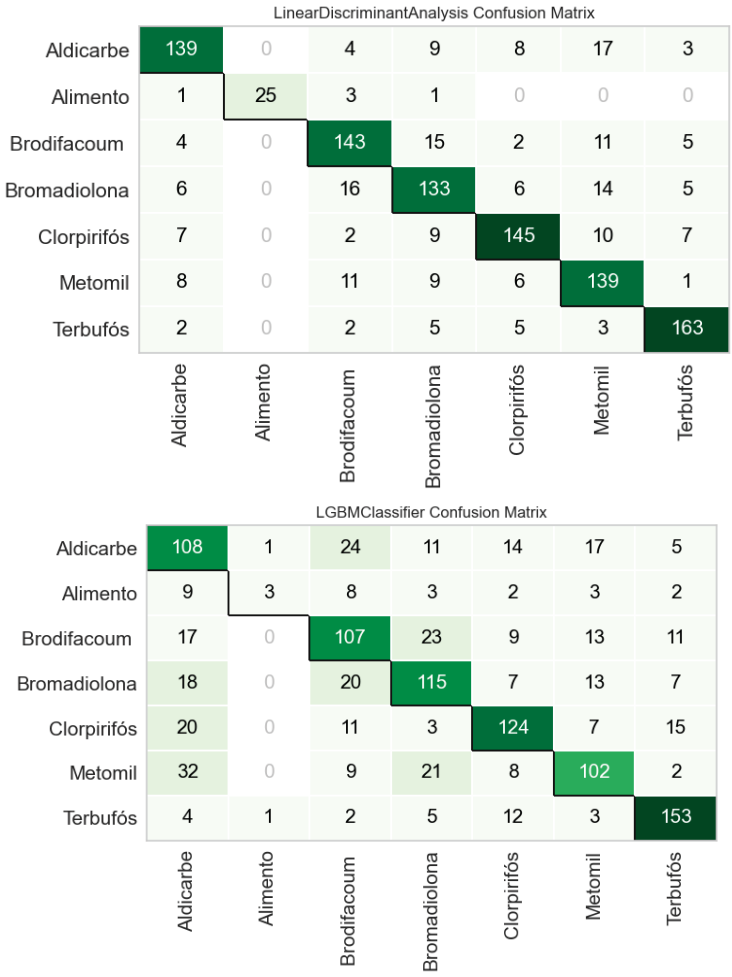
LEGENDA: No eixo X está representado o número sequencial das amostras do conjunto de teste. No eixo Y, tem-se a probabilidade máxima de previsão, ou seja, o quanto o modelo “acreditou” que a amostra pertence à classe atribuída. Valores próximos de 1,0 no eixo Y indicam que o modelo fez a predição com alta confiança, atribuindo probabilidade próxima de 100% para a classe predita. Em contraste, valores mais próximos de 0,5 indicam que o modelo estava incerto, pois atribuiu a classe predita com apenas cerca de 50% de certeza.

No caso do LDA, observa-se que a grande maioria das amostras foi classificada com alta probabilidade preditiva, concentrando-se em valores superiores a 0,85. Essa uniformidade na distribuição da confiabilidade, ainda que com algumas variações pontuais, demonstra que o modelo realiza suas classificações com

elevado grau de certeza. Tal comportamento é coerente com o bom desempenho obtido pelo LDA nas demais métricas avaliadas, como acurácia e AUC.

A análise das matrizes de confusão (FIGURA 37) também permitiu uma avaliação comparativa do desempenho dos dois modelos.

FIGURA 37 - MATRIZ DE CONFUSÃO PARA OS MODELOS LDA E LIGHTGBM



FONTE: A autora (2025).

LEGENDA: 0 – Aldicarbe, 1 – Alimento controle, 2 – Brodifacoum, 3 – Bromadiolona, 4 – Clorpirifós, 5 – Metomil e 6 – Terbufós. Cada célula contém um número absoluto que indica a quantidade de amostras da classe real (horizontal) que foram classificadas como pertencentes a determinada classe prevista (vertical). A diagonal principal da matriz, onde as classes reais e previstas coincidem, representa os acertos do modelo. Já os valores fora da diagonal indicam erros de classificação — também chamados de predições incorretas ou confusões entre classes.

De modo geral, o LDA apresentou menor número de falsos positivos na atribuição das classes em comparação ao LightGBM. Um exemplo representativo é

a classe Metomil: enquanto o LDA obteve 139 classificações corretas, o LightGBM alcançou apenas 102 acertos, além de apresentar uma considerável dispersão de erros, com classificações equivocadas direcionadas principalmente às classes Aldicarbe e Bromadiolona.

Outro ponto de destaque refere-se à classe correspondente aos alimentos sem contaminação. O LDA foi capaz de distinguir essa classe com maior clareza, contabilizando 25 acertos, ao passo que o LightGBM apresentou confusão com todas as demais classes do estudo. Situação análoga foi observada nas classes Brodifacoum e Bromadiolona, dois raticidas cumarínicos estruturalmente similares. Apesar de ambos os modelos apresentarem confusão cruzada entre essas classes, o LDA ainda assim superou o LightGBM em termos de precisão individual, obtendo 143 acertos para Brodifacoum e 133 para Bromadiolona, frente a 107 e 115 acertos, respectivamente, registrados pelo LightGBM.

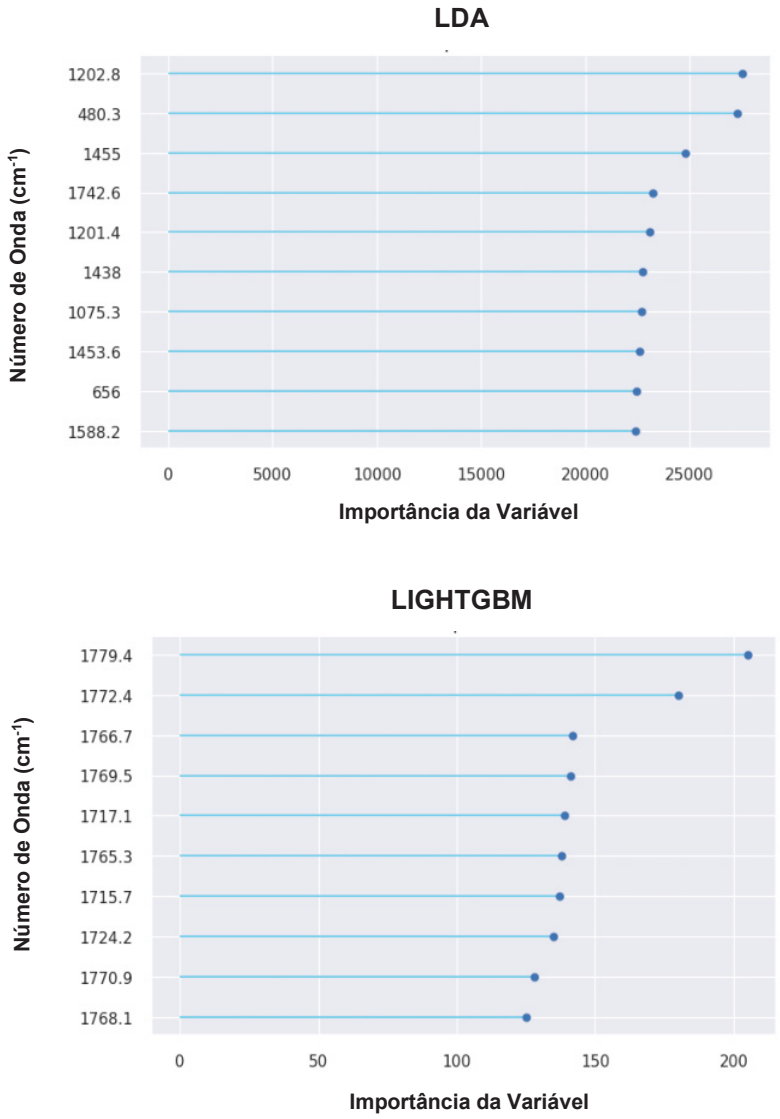
Após a análise detalhada das matrizes de confusão, que permitiu identificar os padrões de acertos e erros específicos de cada classe, procede-se agora à investigação da importância das variáveis espectrais utilizadas pelos algoritmos na etapa de classificação. Essa nova abordagem tem como objetivo esclarecer quais regiões do espectro de infravermelho médio exerceram maior influência na distinção entre as classes analisadas, oferecendo, assim, uma compreensão mais aprofundada dos fatores que sustentam o desempenho discriminativo observado nos modelos avaliados.

A análise do gráfico de importância das variáveis gerado para o modelo LDA (FIGURA 38) evidenciou que, de modo geral, praticamente toda a faixa espectral avaliada contribuiu para a discriminação entre as classes. Esse resultado indica que o modelo não se baseou apenas em regiões pontuais do espectro, mas considerou informações distribuídas ao longo de diferentes comprimentos de onda no processo de classificação, reforçando a complexidade espectral envolvida na distinção entre os grupos analisados.

Em contraste, a análise da importância das variáveis espectrais para o modelo LightGBM revelou um comportamento marcadamente distinto. As variáveis de maior relevância encontravam-se concentradas em uma faixa espectral bastante estreita, entre 1765 e 1779 cm^{-1}, correspondendo a uma região classicamente atribuída aos estiramentos da ligação C=O - ou seja, às bandas de carbonila (SHURVELL, 2006; BARBOSA, 2007; COATES, 2006; PAVIA et al., 2015). Essa

concentração sugere que o LightGBM atribuiu peso decisivo a uma região espectral química muito específica, o que pode explicar, em parte, a maior confusão observada entre classes com perfis espectrais semelhantes fora dessa faixa.

FIGURA 38 - IMPORTÂNCIA DOS NÚMEROS DE ONDA PARA OS MODELOS LDA E LIGHTGB



FONTE: A autora (2025).

LEGENDA: O eixo x representa a importância relativa da variável para cada modelo, ou seja, o quanto cada número de onda contribuiu para a tomada de decisão do algoritmo. O eixo y estão indicados os números de onda (em cm^{-1}). As escalas do eixo x são diferentes entre os dois gráficos devido às metodologias distintas utilizadas por cada algoritmo para calcular a importância das variáveis. Assim, os valores absolutos de importância não são diretamente comparáveis entre os modelos.

Com o intuito de avaliar o desempenho prático dos modelos em condições reais, foi conduzida uma etapa de predição utilizando amostras experimentais previamente analisadas pelo modelo PLS-DA: a mortadela contaminada com terbufós e a amostra de pão contaminado com brodifacoum. Essa aplicação permitiu verificar como os modelos LDA e LightGBM se comportam frente a novas amostras, possibilitando uma comparação mais direta de sua capacidade de generalização. Para facilitar a interpretação dos resultados obtidos, foi elaborado o QUADRO 2, que apresenta um resumo das predições realizadas por cada modelo.

QUADRO 2 - SÍNTESE COMPARATIVA FINAL DA PREVISÃO DE AMOSTRAS REAIS NOS DOIS MELHORES MODELOS CLASSIFICATÓRIOS DA BIBLIOTECA PYCARET

Critério	LDA	LightGBM
Acerto no pão com brodifacoum	**Sim (correto)**	Não (respondeu bromadiolona)
Acerto na mortadela com terbufós	Bromadiolona	**Sim (correto)**

FONTE: A autora (2025).

O modelo LDA classificou corretamente a amostra de pão contaminada com brodifacoum, demonstrando bom desempenho na identificação de anticoagulantes cumarínicos. O modelo LightGBM, por sua vez, obteve resultado oposto: identificou corretamente a amostra de mortadela com terbufós, mas errou na classificação do pão com brodifacoum, atribuindo-o à classe bromadiolona. Esse equívoco, no entanto, não pode ser considerado totalmente desfavorável, já que a predição recaiu sobre uma substância estruturalmente semelhante. Dada a proximidade química entre brodifacoum e bromadiolona (ambos anticoagulantes cumarínicos halogenados), é plausível que o LightGBM, ao concentrar sua atenção quase exclusivamente nas regiões associadas às bandas de carbonila, tenha deixado de captar as sutis diferenças presentes na região de impressão digital do espectro, especialmente entre 400 e 600 cm^{-1}.

Esses achados ilustram os pontos fortes e as limitações complementares de cada abordagem: o LDA, por explorar múltiplas regiões espectrais quimicamente relevantes, evidenciou maior capacidade de generalização; já o LightGBM, embora tenha mostrado boa resposta frente a compostos distintos, apresentou menor acurácia na distinção entre estruturas similares. Modelos baseados em árvores de decisão e aprendizado por gradiente, como o LightGBM, vêm demonstrando desempenho competitivo em diversas tarefas, mas frequentemente requerem ajustes refinados de seus hiperparâmetros. Essa característica foi ressaltada pelos próprios desenvolvedores do LightGBM, Ke *et al.* (2017), que destacam que, embora o algoritmo seja altamente eficiente e preciso em muitos contextos, sua flexibilidade o torna mais sensível a ruídos e à complexidade dos dados — especialmente em domínios caracterizados por alta colinearidade espectral.

Em síntese, os resultados obtidos neste estudo demonstram o potencial dos métodos quimiométricos aliados à espectroscopia de infravermelho na triagem forense de alimentos contaminados. A correta classificação de amostras reais por modelos supervisionados treinados reforça a aplicabilidade prática da estratégia proposta. Ainda que limitações pontuais tenham sido identificadas, os resultados alcançados evidenciam a robustez da abordagem adotada e apontam caminhos promissores para aprimoramentos futuros.

5 PERSPECTIVAS FUTURAS

Ao final deste trabalho, surgem naturalmente questionamentos sobre os próximos passos necessários para transformar essa abordagem promissora em uma ferramenta aplicável na prática forense. O que pode ser aprimorado? Quais barreiras ainda precisam ser superadas? E, sobretudo, o que deve ser feito para que, no futuro, a combinação entre espectroscopia e inteligência artificial se consolide como uma estratégia confiável de triagem pericial? Com base nessas reflexões, esta seção propõe caminhos e diretrizes que poderão orientar o aprimoramento da metodologia desenvolvida, visando sua validação, robustez operacional e integração efetiva ao contexto real de laboratórios forenses. Embora o caminho até a aplicação institucional plena seja desafiador, alguns pontos se destacam como fundamentais nesse processo.

Em primeiro lugar, recomenda-se o aprimoramento do desempenho dos modelos. Uma estratégia consiste em testar alterar hiperparâmetros a serem ajustados no caso do modelo LDA e LightGBM. Outra sugestão é o aumento do número de espectros coletados para todas as classes, principalmente para a classe de alimentos sem contaminantes, incluindo a incorporação de novas matrizes alimentares, ou até mesmo alimentos líquidos, o que contribuiria para ampliar a variabilidade da matriz de calibração. Adicionalmente, sugere-se realizar o refinamento na seleção das variáveis espectrais, priorizando os comprimentos de onda mais relevantes para a tarefa classificatória, o que pode resultar em melhorias tanto na acurácia quanto na interpretabilidade dos modelos.

Outra etapa essencial consiste na determinação dos limites de detecção de cada analito, bem como na avaliação da robustez do modelo frente às variabilidades inerentes ao processo experimental. Para isso, é imprescindível considerar diferentes lotes de preparo da matriz contaminada, contemplando variações como a marca dos alimentos utilizados, os modos de preparação, os dias analíticos, as condições ambientais e, idealmente, a reprodutibilidade em ambientes interlaboratoriais. Essa abordagem visa garantir que o modelo mantenha desempenho consistente e confiável mesmo diante das flutuações naturais do processo analítico.

Ainda no que se refere à matriz, recomenda-se a verificação da estabilidade dos analitos ao longo do tempo nas condições específicas de armazenamento e manuseio. É fundamental avaliar se o intervalo entre a preparação da amostra e sua análise interfere no perfil espectral, bem como os possíveis efeitos causados por ciclos sucessivos de congelamento e descongelamento. Essas informações são cruciais para assegurar a confiabilidade dos resultados em cenários reais, nos quais nem sempre a análise ocorre imediatamente após a coleta.

Outro aspecto crucial refere-se à documentação, rastreabilidade e revalidação do modelo. Recomenda-se a elaboração de um manual técnico que contenha informações detalhadas sobre os critérios de aceitação, as faixas de concentração válidas e os tipos de amostra permitidos para análise. Para assegurar a confiabilidade ao longo do tempo, é fundamental estabelecer protocolos de revalidação periódica, especialmente em contextos nos quais o modelo poderá ser aplicado de forma contínua.

Pensando em uma perspectiva de longo prazo, é pertinente considerar a possibilidade de integrar a metodologia desenvolvida a um *software* automatizado amigável. Tal integração permitiria o processamento ágil dos espectros e a emissão de resultados interpretáveis de forma direta, como “positivo” ou “negativo” para a presença de contaminantes, viabilizando a aplicação da técnica em cenários operacionais com demanda por triagem rápida

Por fim, é imprescindível considerar que a reprodutibilidade dos resultados precisa estar devidamente documentada, de modo a garantir sua admissibilidade como prova técnica em contextos judiciais. Ademais, recomenda-se que o modelo seja validado conforme os requisitos estabelecidos por normas reconhecidas internacionalmente.

6 CONCLUSÕES

O presente estudo demonstrou a viabilidade da utilização da espectroscopia MIR-FTIR, aliada a técnicas de quimiometria clássica e moderna, como uma ferramenta de triagem para a detecção de resíduos de agrotóxicos em matrizes alimentares com relevância forense. A análise exploratória, conduzida por meio da PCA, possibilitou reduzir a dimensionalidade dos dados e identificar padrões de agrupamento.

Dentre os modelos de classificação supervisionada avaliados, a análise por PLS-DA apresentou desempenho satisfatório, demonstrando boa capacidade discriminativa entre as classes, com dados de acurácia, sensibilidade e especificidade de 86%, 76% e 84%, respectivamente. Contudo, essa abordagem apresentou uma limitação importante: sua performance adequada foi restrita ao cenário em que apenas as matrizes de pão e mortadela foram consideradas.

Nesse contexto, os melhores resultados preditivos foram obtidos por meio da aplicação de algoritmos modernos de classificação supervisionada, com destaque para o *Linear Discriminant Analysis* (LDA), que obteve 81% de acurácia, e o *Light Gradient Boosting Machine* (LightGBM), com 67%. Esses modelos mostraram maior robustez diante da variabilidade de matrizes, superando a limitação observada na PLS-DA.

A aplicação dos modelos em amostras reais oriundas da Polícia Científica do Paraná reforçou esses achados. O modelo PLS-DA foi capaz de classificar corretamente a amostra de pão contaminado com brodifacoum, mas falhou na previsão da amostra de mortadela, embora tenha indicado corretamente a classe do contaminante. Por sua vez, o modelo LDA acertou a previsão da amostra contaminada com brodifacoum, enquanto o LightGBM teve êxito ao identificar corretamente a amostra contaminada com terbufós. Esses resultados evidenciam que cada algoritmo apresenta pontos fortes distintos, sendo sensíveis a diferentes padrões espectrais e à variabilidade das matrizes.

Em síntese, este trabalho mostrou, por meio de uma abordagem de prova de conceito, que a combinação entre técnicas espectroscópicas e algoritmos de aprendizado de máquina representa uma alternativa promissora, rápida e não destrutiva para a triagem de agrotóxicos em alimentos.

REFERÊNCIAS

ABAD-GARCÍA, B. *et al*. Comprehensive characterization of phenolic and other polar compounds in fruit juices by high-performance liquid chromatography with diode array detection coupled to electrospray ionization and triple quadrupole mass spectrometry. **Journal of Chromatography A**, v. 1216, n. 43, p. 7245-7254, 2014.

Agência Nacional de Vigilância Sanitária (ANVISA). Gerência Geral de Toxicologia. Gerência de Análise Toxicologica. **NOTA TÉCNICA DA REAVALIAÇÃO DO INGREDIENTE ATIVO ALDICARBE**. Brasília, 2006.

ALPAYDIN, E. **Machine learning: the new AI**. MIT Press, 2016.

AMAYA, D. A. **Espectroscopia no infravermelho médio e seus fundamentos**. 1999.

ANVISA – AGÊNCIA NACIONAL DE VIGILÂNCIA SANITÁRIA. ***Monografias de Ingredientes Ativos de Produtos Saneantes – Brodifacoum (B10) e Bromadiolona (B27)***. Brasília: Anvisa, 2025.

ARAUJO, D. **Estudo comparativo entre algoritmos de aprendizagem de máquina aplicados à detecção de fraudes de cartão de crédito**. 2022. Universidade Federal de Pernambuco, Recife

ARMENTA, S.; QUINTÁS, G.; GARRIGUES, S.; DE LA GUARDIA, M. A validated and fast procedure for FTIR determination of Cypermethrin and Chlorpyrifos. **Talanta**, v. 67, p. 634-639, 2005.

ASADPOOR, M.; ANSARIN, M.; NEMATI, M. Simultaneous determination of preservatives (benzoate, sorbate and sulfite) in orange juice. **Food Chemistry**, v. 145, p. 531-535, 2014.

BARBOSA, L. C. A. **OS PESTICIDAS, O HOMEM E O MEIO AMBIENTE**. Viçosa: UFV. 215 p, 2004.

BARBOSA, L. C. A. **Espectroscopia no infravermelho na caracterização de compostos orgânicos**. 2007. UFV. ISBN 978-85-7269-280-9.

BARKER, M.; RAYENS, W. Partial least squares for discrimination. **Journal of Chemometrics**, 2003.

BARTNECK, C.; LÜTGE, C.; WAGNER, A.; WELSH, S. **An introduction to ethics in robotics and AI.** Springer Nature, 2021.

BEAM, A. L.; KOHANE, I. S. Big data and machine learning in health care. **JAMA**, v. 319, n. 13, p. 1317–1318, 2018.

BEAR, M. F.; CONNORS, B. W.; PARADISO, M. A. **Neurociência: desvendando o sistema nervoso**. 4. ed. Porto Alegre: Artmed, 2017.

BRAIBANTE, M. E. F.; ZAPPE, J. A. A Química dos Agrotóxicos. **Química Nova na Escola**, v. 34, n. 1, p. 10-15, 2012.

BRASIL. **Agência Nacional de Vigilância Sanitária (ANVISA)**. Resolução nº 296, de 29 de julho de 2019. Dispõe sobre as informações toxicológicas para rótulos e bulas de agrotóxicos, afins e preservativos de madeira. Diário Oficial da União, Brasília, DF, 31 jul. 2019.

BRERETON, R. G.; LLOYD, G. R. Partial least squares discriminant analysis: Taking the magic away. **Journal of Chemometrics**, v. 28, n. 4, p. 213-225, 2014.

BRERETON, R. G. **Applied Chemometrics for Scientists**. England: John Wiley & Sons, 2007. 379 p.

BULL, D.; HATHAWAY, D. **Pragas e Venenos:** Agrotóxicos no Brasil e no Terceiro Mundo. Tradução e Ampliação de David Hathaway. Petrópolis, Rio de Janeiro: Vozes, 1986.

BYLESJÖ, M. *et al.* OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. **Journal of Chemometrics**, v. 20, p. 341-351, 2006.

CANO-TRUJILLO, C.; GARCÍA-RUIZ, C.; ORTEGA-OJEDA, F. E.; ROMOLO, F.; MONTALVO, G. Forensic analysis of biological fluid stains on substrates by spectroscopic approaches and chemometrics: A review. **Anal Chim Acta**, v, 1282, n. 341841, 2023.

CASTRO, C. L.; BRAGA, A. P. Aprendizado supervisionado com conjuntos de dados desbalanceados. **Sba: Controle & Automação Sociedade Brasileira de Automatica**, v. 22, p. 441-466, 2011.

CAVALCANTE, J. A.; SOUZA, J. C.; ROHWEDDER, J.J.R.; MALDANER, A.O.; PASQUINI C.; HESPANHOL, M,C. A compact Fourier-transform near-infrared spectrophotometer and chemometrics for characterizing a comprehensive set of seized ecstasy samples. **Spectrochim Acta A Mol Biomol Spectrosc**., v. 314, n. 124163, 2024.

CHALERMCHAIKIT, T., FELICE, L.J., MURPHY, M.J. Simultaneous determination of eight anticoagulant rodenticides in blood serum and liver. **Journal of Analytical Toxicology**, v. 17, n. 1, p. 56–61, 1993.

CHEN, H. *et al.* Simultaneous analysis of carbamate and organophosphorus pesticides in water by single-drop microextraction coupled with GC–MS. **Chromatographia**, v. 70, p. 165–172, 2009.

CHOPHI, R.; SHARMA, S.; SINGH, R. Forensic analysis of red lipsticks using ATR-FTIR spectroscopy and chemometrics. **Forensic Chemistry**, v. 17, n. 100209, 2020.

COATES, J. Interpretation of Infrared Spectra, A Practical Approach. *In*: **Encyclopedia of Analytical Chemistry**, 2006.

COLEMAN, R. **Spectroscopy in Infrared Analysis**. 1993.

COUTO, T. J. G.; BORDINI, P. H. C.; SILVA, L. S.; SALES, S. B. L. Detection of pesticides in forensic expertise: profile of cases and deaths. **Brazilian Journal of Forensic Science Medical Law and Bioethics**, São Paulo, v. 12, n. 2, p. 167–192, ago. 2024.

COX, M. *et al*. The forensic application of Fourier Transform Infrared Spectroscopy. **Journal of Analytical Toxicology**, 2000.

COZZOLINO, D. Infrared spectroscopy as a versatile analytical tool for the quantitative determination of antioxidants in agricultural products, foods and plants. **Antioxidants**, v. 4, n. 3, p. 482–497, 2015.

CREMONESE, C.; FREIRE, C.; MEYER, A.; KOIFMAN, S. Exposição a agrotóxicos e eventos adversos na gravidez no Sul do Brasil, 1996-2000. **Cadernos de Saúde Pública**, Rio de Janeiro, v. 28, n. 7, p. 1263-1272, jul. 2012.

CUSTÓDIO, M. F.; MAGALHÃES, L. O.; ARANTES, L. C.; BRAGA, J. W. B. Identification of synthetic drugs on seized blotter papers using ATR-FTIR and PLS-DA: routine application in a forensic laboratory. **Journal of the Brazilian Chemical Society**, v. 32, n. 3, p. 513-522, 2021.

DECONINCK, E.; DUCHATEAU, C.; BALCAEN, M.; GREMEAUX, L.; COURSELLE, P. Chemometrics and infrared spectroscopy – A winning team for the analysis of illicit drug products. **Reviews in Analytical Chemistry**, 41, p. 228–255, 2022.

DI LORENZO, R. A.; BUTT, C. M. **Sensibilidade como vantagem para simplificar a preparação de amostra e melhorar a produtividade de laboratórios de análise de contaminantes.** Sciex. 2022. Portfólio de diagnóstico clínico da SCIEX.

DOMINGUES, R.; PEDROSA, I.; BERNARDINO, J. Indicadores chave de desempenho em marketing. **Revista Ibérica de Sistemas e Tecnologias de Informação (RISTI)**, n. esp. 35, p. 128–140, 2020.

DONATO, F. F. **Resíduos de agrotóxicos em água potável usando SPE e determinação rápida por LC-MS/MS e GC-MS/MS.** 166 f**.** Dissertação (Mestrado em Química) - Universidade Federal de Santa Maria, Santa Maria, 2012.

DUARTE, J. M.; DE CARVALHO, A. M. F.; POPPI, R. J.; PACHECO, M. T. T. Discrimination of white automotive paint samples using ATR-FTIR and PLS-DA for forensic purposes. **Talanta**, v. 246, n. 123492, 2022.

ECOBICHON; D. J.; JOY, R. M. **Pesticides and neurological diseases**. In: Casarett, L. J. & Doull, J. Toxicology the basic science of poisons. 4th ed. Boca Raton, CRC Press, 1991. p. 565-622.

ESBENSEN, K. H.; SANCHEZ, M. T. **Multivariate Data Analysis — in Practice**. 5. ed. Oslo: CAMO Software, 2010.

ESSLINGER, S.; RIEDL, J.; FAUHL-HASSEK, C. Potential and limitations of nontargeted fingerprinting for authentication of food in official control. **Food Research International**, v. 60, p. 189-204, 2013.

FABUNMI, O. "**A More Efficient Method for Extracting and Analyzing Pesticides in Baby foods**" (2019). Master of Science in Chemical Sciences Theses. 24.

FARID, S.; KASEM, M. A.; ZEDAN, A. F.; MOHAMED, G. G.; EL-HUSSEIN, A. Exploring ATR Fourier transform IR spectroscopy with chemometric analysis and laser scanning microscopy in the investigation of forensic documents fraud. **Optics & Laser Technology**, v. 135, 2021.

FERNÁNDEZ-GONZÁLEZ, M. *et al.* Application of visible-near infrared spectroscopy and multivariate data analysis for rapid authentication of orange juice. **Food Control**, v. 46, p. 1-7, 2014.

FERNEDA, E. Redes neurais e sua aplicação em sistemas de recuperação de informação. **Ciência da Informação**, Brasília, v. 35, n. 1, p. 25-30, jan./abr. 2006.

FERRÃO, M.F. *et al*. Determinação simultânea dos teores de cinza e proteína em farinha de trigo empregando NIRR-PLS e DRIFT-PLS. **Ciência e Tecnologia dos Alimentos**, v. 24, n. 30, p. 333-340, 2004.

FERREIRA, M. M. C. **Quimiometria: Conceitos, Métodos e Aplicações**. 1. ed. Campinas: Editora UNICAMP, 2015a.

FERREIRA, M. M. C. **Aplicações de quimiometria em análises químicas**. Química Nova, 2015b.

FERREIRA, M. M. C. **Quimiometria: Conceitos, Métodos e Aplicações**. Campinas: Editora UNICAMP, 2020.

FERREIRA, M. M. C.; ANTUNES, A. M.; MELGO, M. S.; VOLPE, P. L. O. Quimiometria I: Calibração multivariada, um tutorial. **Química Nova**, v. 22, n. 5, p. 724–731, 1999.

FULGÊNCIO, A.C.C.; RESENDE, G.A.P.; TEIXEIRA, M.C.F.; BOTELHO, B.G.; SENA, M.M. Screening method for the rapid detection of diethylene glycol in beer based on chemometrics and portable near-infrared spectroscopy. **Food Chemistry**, v. 391, p. 133258, 2022.

GÉRON, A. **Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow**. O'Reilly Media, Inc., 2022.

GHOJOGH, B.; CROWLEY, M. The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial. **arXiv preprint**, arXiv:1905.12787, 2019.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. MIT Press, 2016.

GRIFFITHS, P. R.; DE HASETH, J. A. **Fourier Transform Infrared Spectrometry**. 2. ed. Hoboken: Wiley-Interscience, 2007.

GROMSKI, P. S. *et al.* A tutorial review: Metabolomics and partial least squaresdiscriminant analysis - a marriage of convenience or a shotgun wedding. **Analytica Chimica Acta,** v. 879, p. 10–23, 2015.

HAALAND, D. M.; THOMAS, E. V. Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. **Analytical Chemistry**, v. 60, n. 11, p. 1193–1202, 1 jun. 1988.

HAN, J. *et al*. PCR and DHPLC methods used to detect juice ingredient from 7 fruits. **Food Control**, v. 25, p. 696-703, 2012.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. H. **The elements of statistical learning: data mining, inference, and prediction**. 2. ed. Springer, 2009.

HE, Y.; RODRIGUEZ-SAONA, L. E.; GIUSTI, M. M. Midinfrared spectroscopy for juice authenticity determination. **Journal of Agricultural and Food Chemistry**, v. 55, n. 11, p. 4443-4450, 2007.

HEINTZ, N.; FRANCART, T.; BERTRAND, A. Minimally informed linear discriminant analysis: training an LDA model with unlabelled data. **arXiv**, 2310.11110, 2023.

IUPAC. **Compendium of Chemical Terminology: IUPAC Recommendations.** 2. ed. Compiled by Alan D. McNaught and Andrew Wilkinson. Oxford: Blackwell Science, 1997. 450 p.

JOCANOVIC, M. Biotransformation of organophosphorus compunds. **Toxicology**, v. 166, n. 3, p. 139-160,. 2001.

JOHN, D. K.; DOS SANTOS SOUZA, K.; FERRÃO, M. F. Overview of cocaine identification by vibrational spectroscopy and chemometrics. **Forensic Sci Int**., v. 342 n. 111540, 2023.

JOLLIFFE, I. T.; CADIMA, J. Principal component analysis: a review and recent developments. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, v. 374, 2016.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: trends, perspectives, and prospects. **Science**, v. 349, n. 6245, p. 255–260, 2015.

JUNIOR, F. R. L.; CARPINETTI, L. C. R. Modelos de decisão para avaliação de desempenho de cadeias de suprimentos baseados no Scor®: uma revisão da literatura. **Brazilian Journal of Development**, v. 5, n. 7, p. 10301-10318, 2019.

KAROUI, R.; PIERNA, J. F.; DUFOUR, E. Spectroscopic technique: mid-infrared (MIR) and Fourier transform mid-infrared (FT-MIR) spectroscopies. In: SUN, D.W. **Modern Techniques for Food Authentication**, Hardbound: Academic Press, 2008, p. 27-64.

KE, G.; MENG, Q.; FINLEY, T.; WANG, T.; CHEN, W.; MA, W.; YE, Q.; LIU, T. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. **In: Advances in Neural Information Processing Systems**, v. 30, p. 3146-3154, 2017.

KIM, B.; YANG, SEUNG-HYUN; CHOI, H. Organophosphate detection in animal-derived foods using a modified Quick, Easy, Cheap, Effective, Rugged, and Safe method with liquid chromatography–mass spectrometry. **Foods**, v. 13, n. 16, p. 2642, 2024.

KIRTIL, E. *et al*. Recent advances in time domain NMR & MRI sensors and their food applications. **Current Opinion in Food Science**, v. 17, p. 9-15, 2017.

KOWALSKI. Chemometrics: Views and Propositions. **Journal of Chemical Information and Computer Sciences**, 15, n. 4, p. 201-203, 1975.

LAJOLO, F.M.; NUTTI, M.R. **Transgênicos: bases científicas da sua segurança**. São Paulo: SBAN, 2003.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, v. 521, n. 7553, p. 436–444, 2015.

LIU, M.; HASHI, Y.; SONG, Y.; LIN, J. Simultaneous determination of carbamate and organophosphorus pesticides in fruits and vegetables by liquid chromatography–mass spectrometry. **Journal of Chromatography A**, v. 1097, n. 1-2, p. 183–187, 2005.

LINARDATOS, P.; PAPASTEFANOPOULOS, V.; KOTSIANTIS, S. Explainable ai: A review of machine learning interpretability methods. **Entropy**, v. 23, n. 1, 2021.

LUDERMIR, T. B. Inteligência Artificial e Aprendizado de Máquina: estado atual e tendências. **Estudos Avançados**, v. 35, n. 101, p. 85–94, 2021.

MANLEY, M.; DOWNEY, G.; BAETEN, V. Spectroscopic technique: near-infrared (NIR) spectroscopy. In: SUN, D.W. Modern techniques for food authentication, Hardbound. **Academic Press**, 2008, p. 65-116.

MATERAZZI, S.; GREGORI, A.; RIPANI, L.; APRICENO, A.; RISOLUTI, R. Cocaine profiling: implementation of a predictive model by ATR-FTIR coupled with chemometrics in forensic chemistry. **Talanta**, v. 172, p. 227-234, 2017.

MDENI, N. L. *et al.* Analytical evaluation of carbamate and organophosphate pesticides in human and environmental matrices: a review. **Molecules**, v. 27, n. 3, p. 618, 2022.

MENDEZ, K. M.; REINKE, S. N.; BROADHURST, D. I. A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification. **Metabolomics**, v. 15, 2019.

MIRABELA, F.M. International reflection spectroscopy. **Applied Spectroscopy**, v. 21, p. 45-178, 1985.

MODREGO, I. N. **Intoxicación por rodenticidas anticoagulantes en veterinaria: estado de la cuestión.** Trabajo fin de grado en veterinaria. 2022.

OGA, S.; CAMARGO, M. M. A. C.; BATISTUZZO, J. A. O**. FUNDAMENTOS DE TOXICOLOGIA**. 3. ed. São Paulo: Atheneu Editora, 2008.

OLIVERI, P.; MALEGORI, C.; SIMONETTI, R.; CASALE, M. The impact of signal pre-processing on the final interpretation of analytical outcomes - A tutorial. **Anal Chim Acta**, 1058, p. 9-17, Jun 13 2019.

PARANÁ. Secretaria da Saúde. **Material técnico - Intoxicações Agudas por Agrotóxicos – Atendimento Inicial do Paciente Intoxicado**. Paraná, 2018. 120p.

PASQUINI, C. Near infrared spectroscopy: A mature analytical technique with new perspectives - A review. **Anal Chim Acta**, v. 1026, p. 8-36, 2018.

PATOCKA, J.; PETROIANU, G.; KUCA, K. Toxic potential of superwarfarin: brodifacoum. **Military Medical Science Letters**, v. 82, n. 1, p. 32–38, 2013.

PAVIA, D.; LAMPMAN, G.; KRIZ, G.; VYVYAN, J. **Introdução à espectroscopia**. Tradução do inglês. Edição Português. São Paulo: Cengage Learning, 2015.

PAZ, J. V.; REZENDE, V. T.; GAMEIRO, A. **Agrotóxicos no Brasil: entre a produção e a segurança alimentar**. Jornal da USP, São Paulo, 1 ago. 2023. Disponível em: https://jornal.usp.br/artigos/agrotoxicos-no-brasil-entre-a-producao-e-a-seguranca-alimentar/. Acesso em: 31 maio 2025.

PEREIRA, H. V. *et al.* Paper spray mass spectrometry and PLS-DA improved by variable selection for the forensic discrimination of beers. **Analytica Chimica Acta**, Amsterdam, v. 940, p. 104–112, 12 out. 2016.

PERIS-DÍAZ, M. D.; KREŻEL, A. **A guide to good practice in chemometric methods for vibrational spectroscopy, electrochemistry, and hyphenated mass spectrometry**. **TrAC - Trends in Analytical Chemistry.** Elsevier. 1 fev. 2021.

POLÍCIA CIENTÍFICA DO PARANÁ (PCI-PR). Gestor de Documentos e Laudos (GDL). **Pesquisa Geral**. Paraná, 2023.

PULIDO, A. *et al.* Uncertainty of results in routine qualitative analysis. **TrAC - Trends in Analytical Chemistry**, v. 22, p. 647-654, 2003.

ROBINSON, C. **Genetic modification technology and food**. Brussels: ILSI Europe, 2001.

RODIONOVA. Efficient tools for principal component analysis of complex data— A tutorial. **Chemometrics and Intelligent Laboratory Systems**, 213, 2021.

RUSSELL, S.; NORVIG, P. **Artificial Intelligence: A Modern Approach**. 2010.

SANTANA, F. B. *et al*. Experimento didático de quimiometria para classificação de óleos vegetais comestíveis por espectroscopia no infravermelho médio combinado com análise discriminante por mínimos quadrados parciais: um tutorial, parte V. **Química Nova**, v. 43, n. 3, p. 371–381, 2020.

SHAH, N. N. *et al.* A study of fruit juices authentication using MIR spectroscopy. **Food Analytical Methods**, v. 3, n. 1, p. 91-98, 2010.
SHURVELL, H. F. Spectra-Structure Correlations in the Mid- and Far- Infrared. **Handbook of Vibrational Spectroscopy**. 2006.

SILVA, H.K.T.A. *et al.* Detection of terbufos in cases of intoxication by means of entomotoxicological analysis using ATR-FTIR spectroscopy combined with chemometrics. **Acta Tropica**, v. 238, p. 106779, 2023.

SILVA, J. M.; NOVATO-SILVA, E.; FARIA, H. P.; PINHEIRO, T. M. M. Agrotóxico e trabalho: uma combinação perigosa para a saúde do trabalhador rural. **Ciência & Saúde Coletiva**, v. 10, n. 4, p. 891-903, 2005.

SIQUEIRA, D. F.; MOURA, R. M.; LAURENTINO, G. E. C.; ARAÚJO, A. J.; CRUZ, S. L. ANALISE DA EXPOSIÇÃO DE TRABALHADORES RURAIS A AGROTÓXICOS. **Revista Brasileira em Promoção da Saúde**, Fortaleza, v. 26, n. 2, p. 182-191, 2013.

SNYDER, A. B. *et al*. Mid-infrared spectroscopy coupled with chemometrics for identification of fruit juices. **Journal of Agricultural and Food Chemistry**, v. 62, p. 4088-4094, 2014.

SOUZA, R. M. ATR: AVANÇO NA ESPECTROSCOPIA DE INFRAVERMELHO NA ANÁLISE DE MATERIAIS PLÁSTICOS. **Instituto de Tecnologia de Alimentos**, v. 21, n. 3, 2009.

SOUZA, A. M.; POPPI, R. J. Teaching experiment of chemometrics for exploratory analysis of edible vegetable oils by mid infrared spectroscopy and principal component analysis: a tutorial, part I. **Química Nova**, v, 35, n.1, 2012.

SOUZA, R.; RIBEIRO, L. N. Agrotóxicos em alimentos: desafios periciais e perspectivas metodológicas. **Cadernos de Perícia Criminal**, v. 7, p. 55-70, 2020.

STEENSMA, A., BEAMAND, J. A., WALTERS, D. G., PRICE, R. J., LAKE, B. G. Metabolism of coumarin and 7-ethoxycoumarin by rat, mouse, guinea pig, cynomolgus monkey and human precision-cut liver slices. **Xenobiotica**, V. 24, n. 9, p. 893-907, 1994.

STEIDLE NETO, A. J.; DE LIMA, J. L. M. P.; JARDIM, A. M. D. R. F.; LOPES, D. D. C.; SILVA, T. G. F. D. Discrimination of Fungicide-Contaminated Lettuces Based on Maximum Residue Limits Using Spectroscopy and Chemometrics. **Horticulturae**, v. 10, n. 8, 2024.

TAM, K. W.; CHAN, C. K.; LIU, S. Anticoagulant rodenticide ingestion: Who will develop coagulopathy? **Hong Kong Journal of Emergency Medicine**, v. 30, n. 2, p. 1-9, 2021.

TANKIEWICZ, M; FENIK, J.; BIZIUK, M. Determination of organophosphorus and organonitrogen pesticides in water samples. **Trends in Analytical Chemistry**, v. 29, n. 9, p. 1050-63, 2010.

THAKUR, S.; SHARMA, A.; CIEŚLA, R.; KUMAR MISHRA, P. K.; SHARMA, V. A novel approach using ATR-FTIR spectroscopy and chemometric analysis to distinguish male and female human hair samples. **The Science of Nature**, v. 111, n. 9, 2024.

TSAGKARIS, A. *et al*. Screening of carbamate and organophosphate pesticides in food matrices using an affordable and simple spectrophotometric acetylcholinesterase assay. **Applied Sciences**, v. 10, n. 2, p. 565, 2020.

TURING, A. M. Computing Machinery and Intelligence. **Mind**, n. 236, p.433-460, 1950.

VALCHEV, I; BINEV, R.; YORDANOVA, V.; NIKOLOV, Y. Anticoagulant Rodenticide Intoxication in Animals - A Review. **Turkish Journal of Veterinary & Animal Sciences**, v. 32, n. 4, p. 237-243, 2008.

VEIGA, M. M. Agrotóxicos: eficiência econômica e injustiça socioambiental. **Ciência & Saúde Coletiva**, v. 12, n. 1, p. 145-152, 2007.

VELOSO, B. C. **Análise de algoritmos de machine learning para deteção de violência em áudio.** 2023. Universidade do Minho, Braga.

VOET, D.; VOET, J.; PRATT, C. **Fundamentos de bioquímica**. 2000.

WEI, C.T.; YOU, J. L.; WENG, S. K.; JIAN, S. Y.; LEE, J. C.; CHIANG, T. L. Enhancing forensic investigations: Identifying bloodstains on various substrates through ATR-FTIR spectroscopy combined with machine learning algorithms. **Spectrochim Acta A Mol Biomol Spectrosc**., v. 308, n. 123755, 2024.

WESTAD, F.; MARINI, F. Validation of chemometric models - A tutorial. **Analytica Chimica Acta,** v. 893, p. 14-24, 2015.

WHIG, P. *et al*. A novel method for diabetes classification and prediction with Pycaret. **Microsystem Technologies**, p. 1-9, 2023.

WISE, B. M. *et al.* PLS_Toolbox Version 4.0 for use with MATLAB™. **Eigenvector Research**, Inc, v. 3905, 2006.

WISE, B.M. *et al. Chemometrics Tutorial for PLS_Toolbox and Solo*. Wenatchee, WA: **Eigenvector Research**, 2019.

WOLD, S.; ESBENSEN, K.; GELADI, P. Principal Component Analysis. **Chemometrics and Intelligent Laboratory Systems**, v. 2, p. 37-52, 1987.

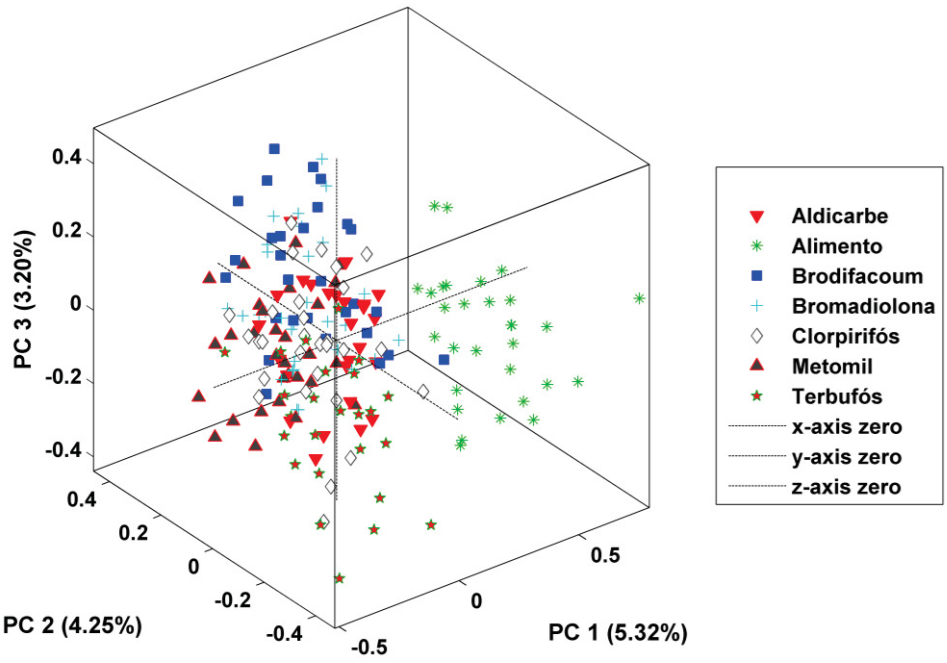
WOLD, S.; SJÖSTRÖM, M.; ERIKSSON, L. PLS-regression: a basic tool of chemometrics. **Chemometrics and Intelligent Laboratory Systems**, v. 58, p. 109–130, 2001.

XIAO, G.; DONG, D.; LIAO, T.; LI, Y.; ZHENG, L.; ZHANG, D.; ZHAO, C. Detection of Pesticide (Chlorpyrifos) Residues on Fruit Peels Through Spectra of Volatiles by FTIR. **Food Analytical Methods**, v. 8, n. 5, 2015.

ZHANG, J. *et al.* Review of the current application of fingerprinting allowing detection of food adulteration and fraud in China. **Food Control**, v. 22, p. 1126-1135, 2011.

ZORZETTI, B. M.; SHAVER, J. M.; HARYNUK, J. J. Estimation of the age of a weathered mixture of volatile organic compounds. **Analytica Chimica Acta**, v. 694, p. 31–37, 2011.

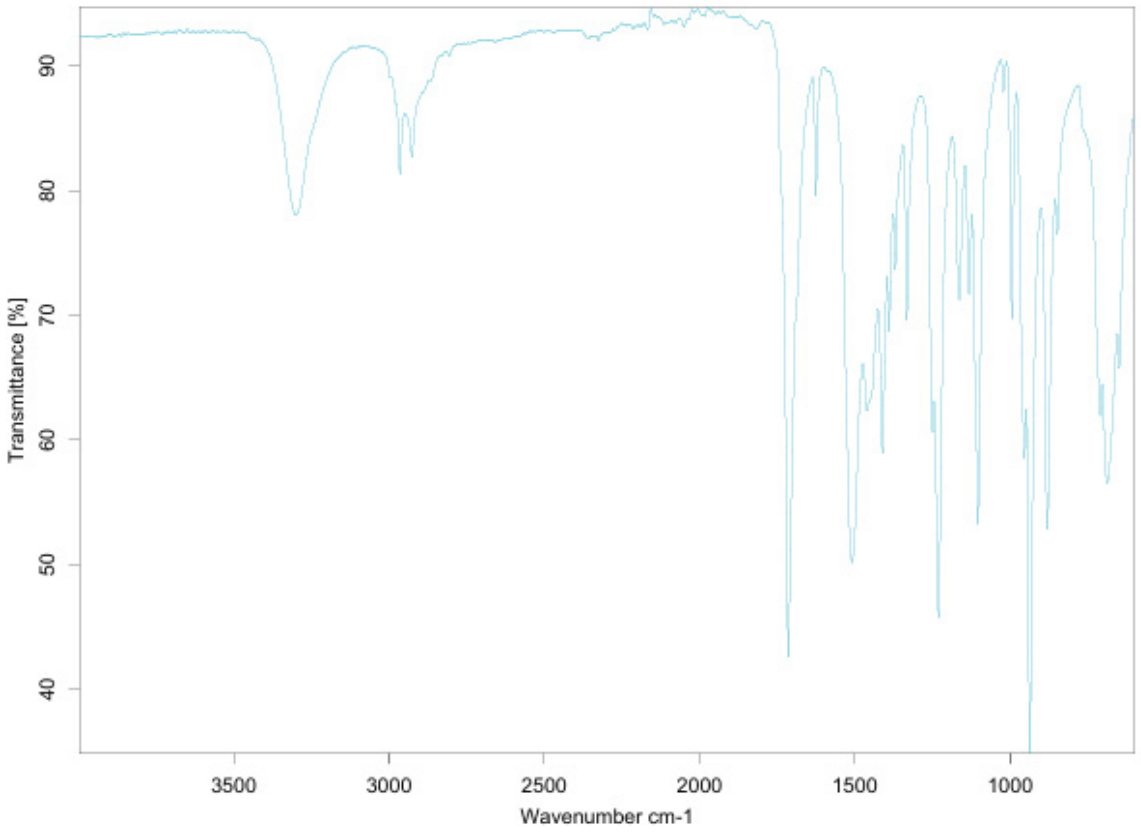
APÊNDICE 1 - GRÁFICO DOS ESCORES DA PC1 X PC2 X PC3 OBTIDOS POR ANÁLISE DE COMPONENTES PRINCIPAIS DOS ESPECTROS PRÉ-TRATADOS DE AMOSTRAS CONTAMINADAS COM DIFERENTES AGROTÓXICOS



FONTE: A autora (2025).

LEGENDA: Representação tridimensional dos escores das três primeiras componentes principais, responsáveis por 5,32%, 4,35% e 3,20% da variância explicada. Observa-se agrupamento parcial entre as classes, prejudicando a verificação de padrões associados à composição química dos contaminantes.

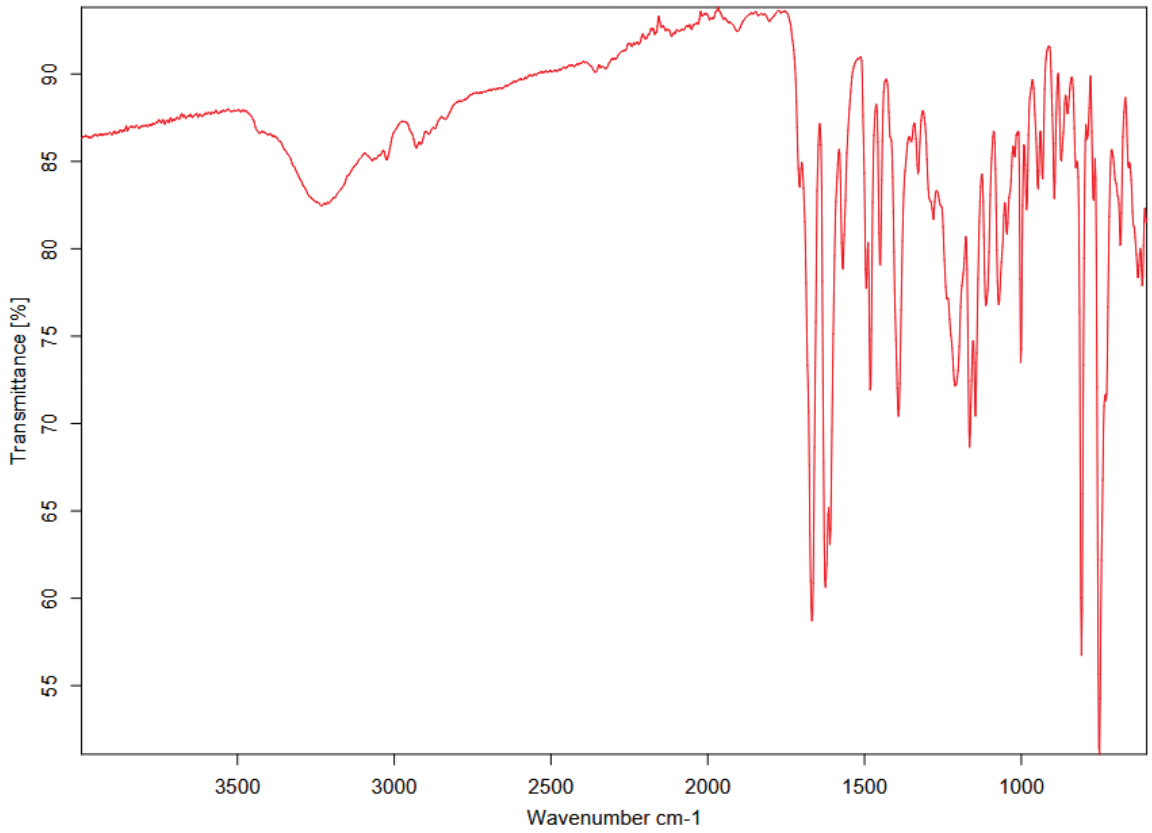
APÊNDICE 2 - ESPECTRO DE ESPECTROSCOPIA FTIR-ATR DO ALDICARBE



FONTE: A autora (2025).

LEGENDA: Espectro obtido por espectroscopia no infravermelho com transformada de Fourier (FTIR), utilizando acessório de reflectância total atenuada (ATR), na faixa de 4000 a 600 cm^{-1}.

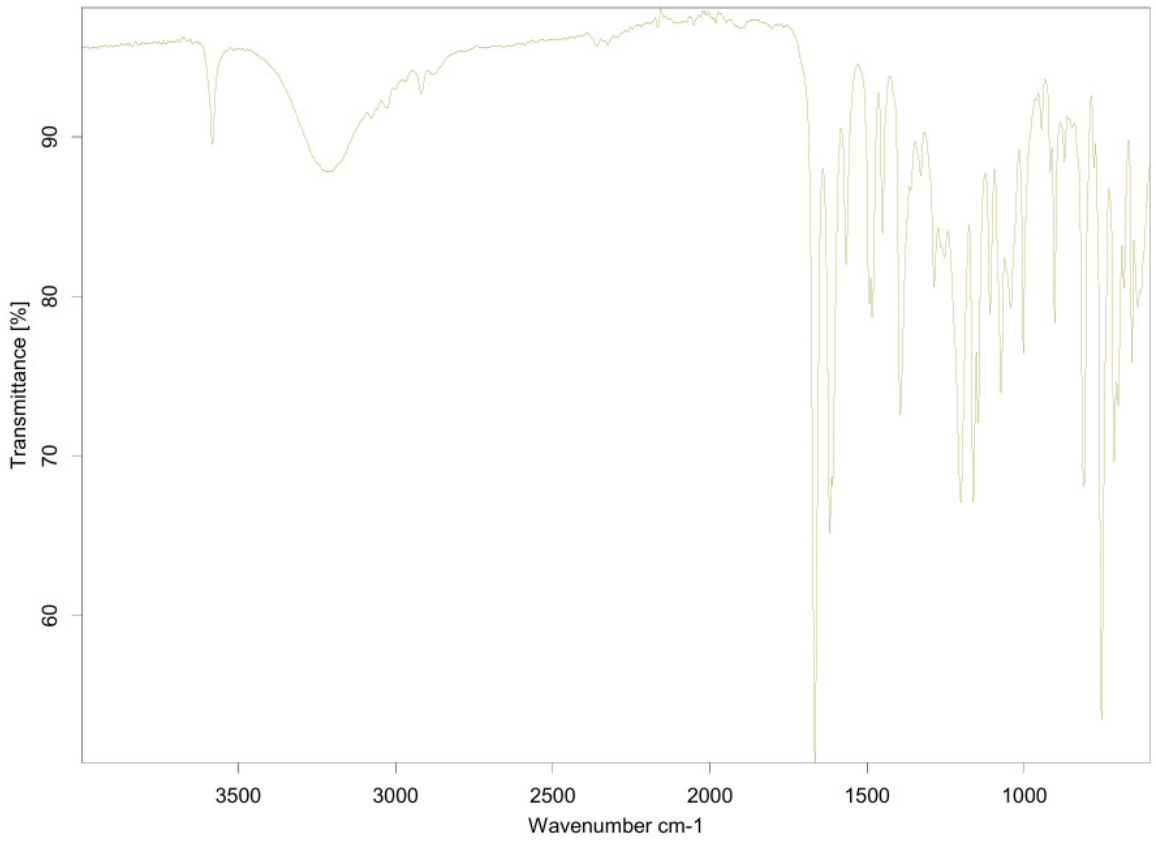
APÊNDICE 3 - ESPECTRO DE ESPECTROCOPIA FTIR-ATR DO BRODIFACOUM



FONTE: A autora (2025).

LEGENDA: Espectro obtido por espectroscopia no infravermelho com transformada de Fourier (FTIR), utilizando acessório de reflectância total atenuada (ATR), na faixa de 4000 a 600 cm^{-1}.

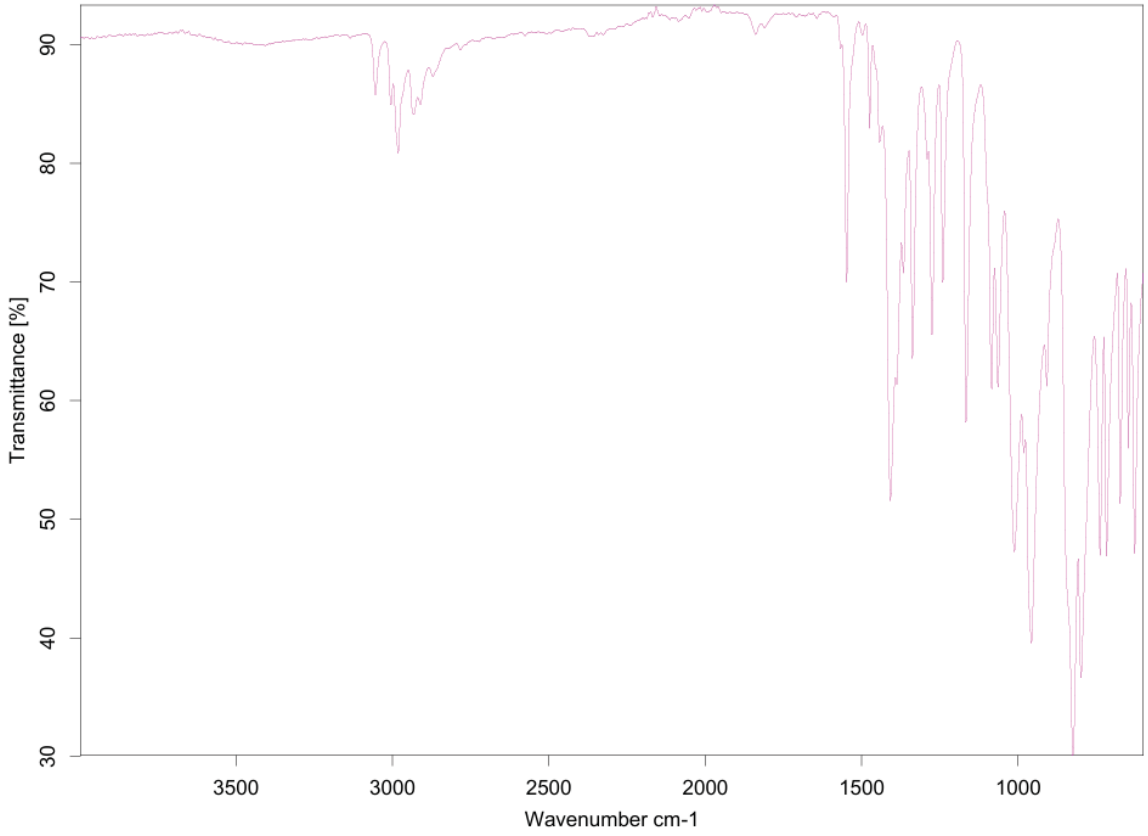
APÊNDICE 4 - ESPECTRO DE ESPECTOSCOPIA FTIR-ATR DA BROMADIOLONA



FONTE: A autora (2025).

LEGENDA: Espectro obtido por espectroscopia no infravermelho com transformada de Fourier (FTIR), utilizando acessório de reflectância total atenuada (ATR), na faixa de 4000 a 600 cm^{-1}.

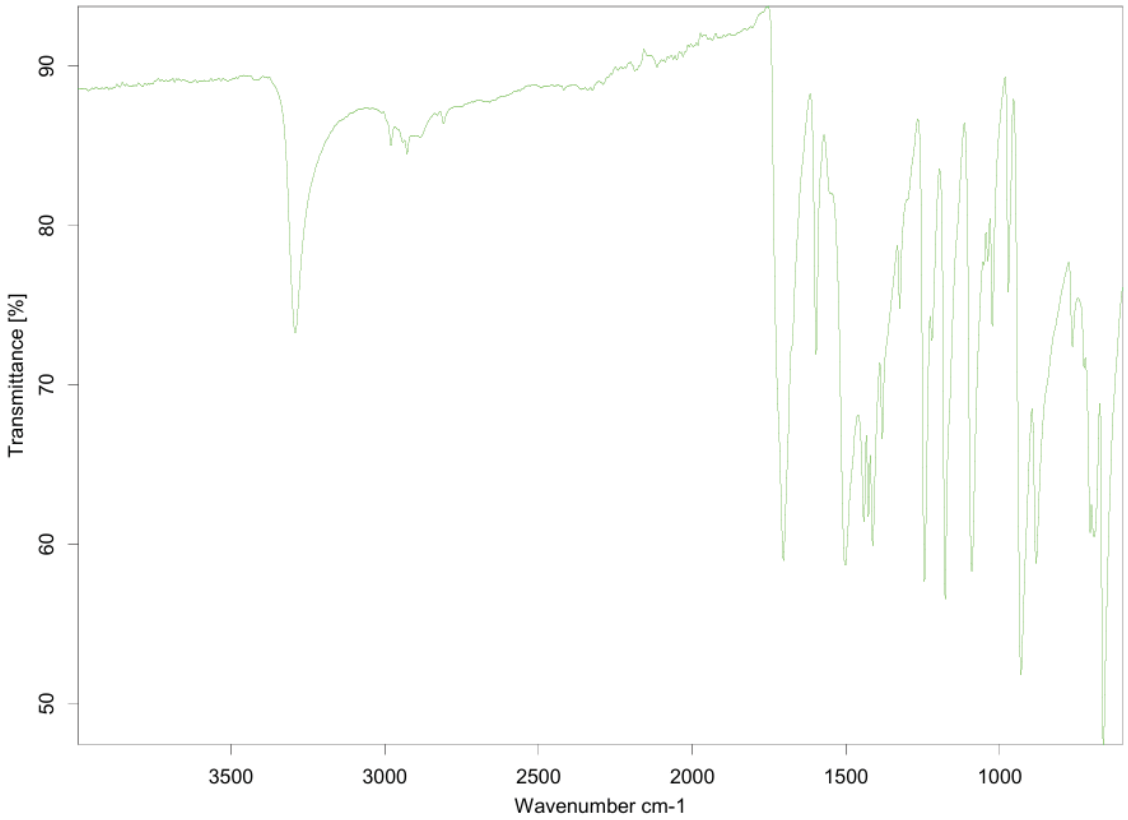
APÊNDICE 5 - ESPECTRO DE ESPCTROSCOPIA FTIR-ATR DO CLORPIRIFÓS



FONTE: A autora (2025).

LEGENDA: Espectro obtido por espectroscopia no infravermelho com transformada de Fourier (FTIR), utilizando acessório de reflectância total atenuada (ATR), na faixa de 4000 a 600 cm^{-1}.

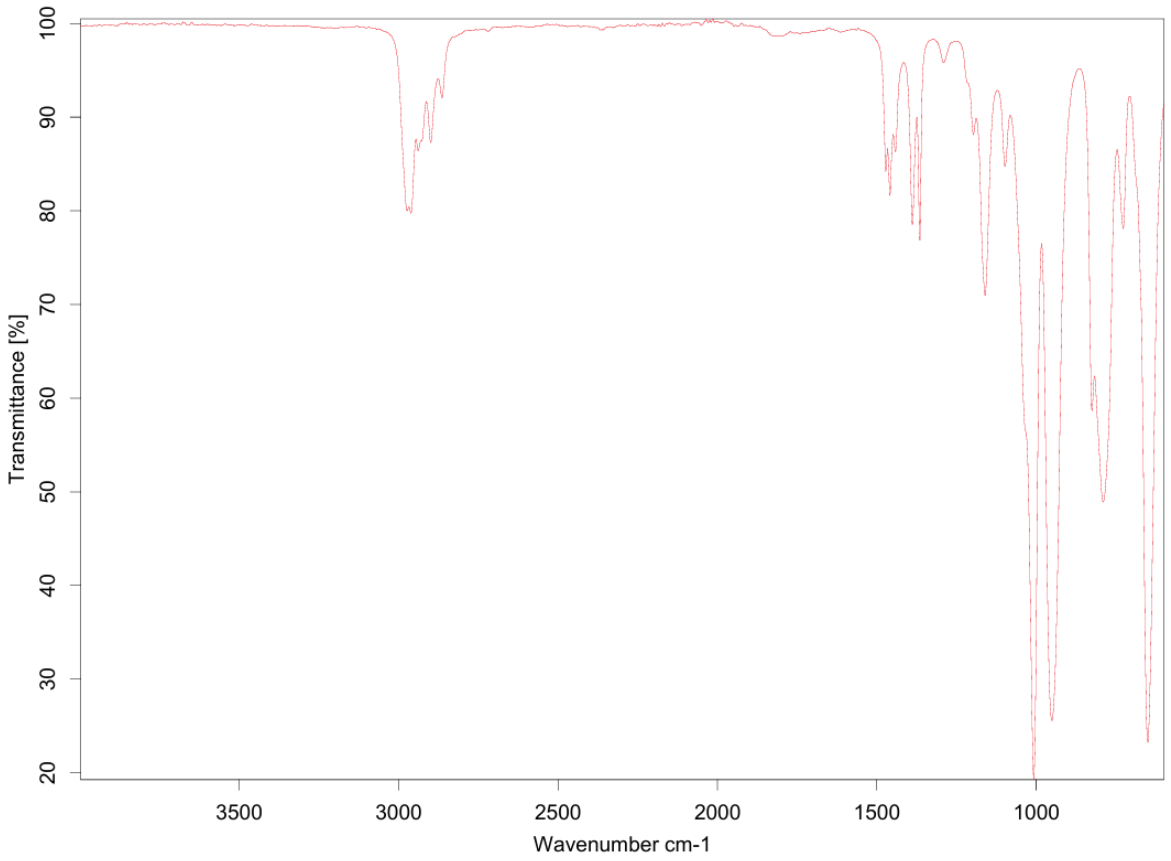
APÊNDICE 6 - ESPECTRO DE ESPECTROSCOPIA FTIR-ATR DO METOMIL



FONTE: A autora (2025).

LEGENDA: Espectro obtido por espectroscopia no infravermelho com transformada de Fourier (FTIR), utilizando acessório de reflectância total atenuada (ATR), na faixa de 4000 a 600 cm^{-1}.

APÊNDICE 7 - ESPECTRO DE ESPECTROSCOPIA FTIR-ATR DO TERBUFÓS



FONTE: A autora (2025).

LEGENDA: Espectro obtido por espectroscopia no infravermelho com transformada de Fourier (FTIR), utilizando acessório de reflectância total atenuada (ATR), na faixa de 4000 a 600 cm^{-1}.

APÊNDICE 8 - DESCRIÇÃO DO CÓDIGO EM PYTHON PARA ANÁLISE CLASSIFICATÓRIA COM PYCARET

PARTE 1 - REQUISITOS, INSTALAÇÕES, IMPORTAÇÕES E VARIÁVEIS GLOBAIS:
Visando evitar que as instalações de programas e importações de bibliotecas sejam distribuídas repetidas vezes ao longo do código, todas elas estão condensadas nesta parte. Do mesmo modo, endereços de pastas e nomes de variáveis que mudariam de acordo com o equipamento, analista e termos da pesquisa são declarados aqui para evitar que tenham que ser editados em todo o código.
Esta parte deve então ser executada sempre que o código for reaberto.

REQUISITOS PARA EXECUÇÃO DESTE CÓDIGO:

Instale o anaconda 3 ou miniconda e crie um ambiente (enviroment) onde deve ser instalado o Python 3 na versão 3.10 ou superior.
Este será usado como kernel base para execução desse projeto.
Execute o código com extensão .ipynb do júpiter notebook no Google Colab, Anaconda navigator ou Usando uma IDE como o VSCODE com a extensão Python e Jupyter notebook support. Caso utilize o Google Colab e deseje salvar ou ler arquivos diretamente no Google Drive observe as opções comentadas com '#' e retire estes comentários.

1.1. INSTALAR PACOTES:

```
# Upgrade pip to the latest version
%pip install --upgrade pip

# Install packages one by one
%pip install fastapi
%pip install kaleido
%pip install python-multipart
%pip install uvicorn
%pip install chembl_webresource_client
%pip install rdkit
%pip install lazypredict
%pip install pycaret
%pip install wget
%pip install unzip
%pip install matplotlib
%pip install scikit-learn
%pip install xgboost
%pip install pandas
%pip install seaborn
%pip install padelpy
%pip install tensorflow
%pip install keras
%pip install jinja2
%pip install umap-learn
%pip install openpyxl
%pip install pycaret
%pip install lightgbm
```

```
# sudo apt-get install python3-tk #instalar no terminal
# %pip install cuml


## **1.2.  IMPORTAR BIBLIOTECAS:**

# Manipulação de expressões regulares:
import re

# Manipulação de pastas e arquivos:
import glob
import os
import shutil

#Gerador de data e hora:
from datetime import datetime

# Manipulação de subprocessos:
import subprocess

#Gerador de interfacees gráficas:
import tkinter as tk
from tkinter import messagebox, filedialog

# Armazenamento de modelos treinados:
import joblib

# Manipulação de dados em tabelas:
import pandas as pd

# Manipulação de gráficos:
import seaborn as sns
import matplotlib.pyplot as plt

# Manipulação de arrays e cálculos matemáticos:
import numpy as np

# Geração de descritores:
from padelpy import padeldescriptor

# Acesso a base de dados do ChEMBL:
from chembl_webresource_client.new_client import new_client

# Aplicação de métodos estatísticos:
import scipy.stats as stats
from scipy.stats import shapiro
from scipy.stats import zscore
from scipy.stats import mannwhitneyu
from scipy.stats import ttest_ind
from scipy.stats import norm
from scipy.stats import pearsonr
```

```
from scipy.stats import gaussian_kde
from scipy.stats import spearmanr
from scipy.stats import kendalltau
from scipy.stats import pearsonr

# Manipula estruturas químicas e propriedades moleculares:
import rdkit
from rdkit import Chem
from rdkit.Chem import Descriptors, Lipinski

# # Realiza um secreening de modelos de Machine Learning:
# import lazypredict
# from lazypredict.Supervised import LazyRegressor
# from lazypredict.Supervised import LazyClassifier

from pycaret.classification import *
from pycaret import classification
from pycaret.regression import *

# Base de dados de modelos de Machine Learning:
import sklearn

#Modelos de regressão:
from sklearn.ensemble import RandomForestRegressor, AdaBoostRegressor, BaggingRegressor, GradientBoostingRegressor,
ExtraTreesRegressor
from sklearn.linear_model import Ridge, Lasso, ElasticNet, BayesianRidge
from sklearn.neighbors import KNeighborsRegressor
from sklearn.svm import NuSVR, SVR
from sklearn.tree import DecisionTreeRegressor
from sklearn.experimental import enable_hist_gradient_boosting  # necessário para HistGradientBoostingRegressor
from sklearn.ensemble import HistGradientBoostingRegressor
import lightgbm as lgb
import xgboost as xgb

#Modelos de classificação:
from sklearn.ensemble import (RandomForestClassifier, BaggingClassifier,
                              GradientBoostingClassifier, ExtraTreesClassifier,
                              AdaBoostClassifier)
from sklearn.tree import DecisionTreeClassifier
from sklearn.calibration import CalibratedClassifierCV
from sklearn.svm import SVC, NuSVC
from sklearn.neighbors import KNeighborsClassifier
from sklearn.linear_model import LogisticRegression, RidgeClassifier, SGDClassifier, RidgeClassifierCV
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.naive_bayes import GaussianNB
from sklearn.experimental import enable_hist_gradient_boosting  # Necessário para HistGradientBoostingClassifier
from sklearn.ensemble import HistGradientBoostingClassifier
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
from sklearn.metrics import roc_auc_score, roc_curve, auc
from sklearn.metrics import precision_score, recall_score, f1_score
from sklearn.model_selection import cross_val_score, cross_val_predict, KFold
```

```
from sklearn.preprocessing import label_binarize

#Modelos de redes neurais:
from sklearn.neural_network import MLPRegressor
from sklearn.neural_network import MLPClassifier

#Outros recursos do sklearn:
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import cross_val_score, cross_val_predict
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.feature_selection import VarianceThreshold
from sklearn.decomposition import PCA
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score

# Trabalha com redes neurais e uso de GPUs:
# import tensorflow as tf
# from tensorflow.keras.models import Sequential
# from tensorflow.keras.layers import Dense
# from tensorflow.keras.optimizers import Adam

## **1.3. ATRIBUIR VARIÁVEIS GLOBAIS, ENDEREÇOS E CRIAR AS PASTAS:**

**ATENÇÃO:** É importante uma análise exploratória prévia dos dados no website do chembl para avaliar melhor os parâmetros necessários.

# Definindo os endereços para as pastas onde os dataframes, imagens e resultados serão salvos:
# data_hora = datetime.now().strftime('%Y%m%d_%H%M%S')
caminho = os.getcwd()
caminho_dataset_interno = f'{caminho}/DATASET_INTERNO'
caminho_dataset_externo = f'{caminho}/DATASET_EXTERNO'
caminho_resultados = f'{caminho}/RESULTADOS'
caminho_resultados_predicoes = f'{caminho}/RESULTADOS/PREDIÇÕES'
caminho_resultados_figuras = f'{caminho}/RESULTADOS/FIGURAS'
caminho_resultados_modelos = f'{caminho}/RESULTADOS/MODELOS'
caminho_modelos = f'{caminho}/MODELOS'
caminho_modelos_classificacao = f'{caminho}/MODELOS/CLASSIFICAÇÃO'
caminho_modelos_redes_neurais = f'{caminho}/MODELOS/REDES_NEURAIS'

# Create all directories at once
directories = [
    f'{caminho}/DATASET_INTERNO',
    f'{caminho}/DATASET_EXTERNO',
    f'{caminho}/RESULTADOS',
    f'{caminho}/RESULTADOS/PREDIÇÕES',
    f'{caminho}/RESULTADOS/FIGURAS',
    f'{caminho}/RESULTADOS/MODELOS',
    f'{caminho}/MODELOS',
    f'{caminho}/MODELOS/CLASSIFICAÇÃO',
    f'{caminho}/MODELOS/REDES_NEURAIS'
]
```

```
for dir_path in directories:
    os.makedirs(dir_path, exist_ok=True)

# **PARTE 2 - SCREENING DOS MODELOS DE MACHINE LEARNING:**

Nesta parte são aplicadas as bibliotecas **lazypredict e Pycaret** com a finalidade de realizarmos um screening de diferentes modelos de machine learning de modo a selecionar os melhores para nosso conjunto de dados. Dessa forma, teremos um direcionamento inicial que nos permitirá, na sequência, otimizar o modelo selecionado e avaliar suas métricas de performance.

## 2.1. Importação e visualização do dataframe pré-tratado:

Eliminando valores infinitos ou fora do intervalo float64:

from google.colab import files
import pandas as pd
import io

# Etapa 1: Fazer upload do arquivo Excel
uploaded = files.upload()

# Etapa 2: Ler o arquivo Excel corretamente
for nome_arquivo in uploaded.keys():
    df = pd.read_excel(io.BytesIO(uploaded[nome_arquivo]))  # para .xlsx ou .xls

# Etapa 3: Exibir o conteúdo
df

# Função para selecionar o arquivo (CSV ou XLSX):
def selecionar_arquivo():
    root = tk.Tk()
    root.withdraw()  # Esconde a janela principal
    caminho = filedialog.askopenfilename(filetypes=[("Excel files", "*.xlsx")])
    # caminho = filedialog.askopenfilename(filetypes=[("CSV files", "*.csv")])
    root.destroy()  # Destroi a janela principal após a seleção do arquivo
    return caminho

# Selecionar o arquivo CSV
caminho = selecionar_arquivo()

# Carregar os dados do arquivo CSV
df = pd.read_excel(caminho)
# df = pd.read_csv(caminho)
display(df.head())

## 2.2. Definindo os valores de **X e Y**:

x = df.iloc[:, 1:]  # Todas as colunas, exceto a primeira (features)
y = df.iloc[:, 0]     # Primeira coluna (label)
bioatv_name = df.columns[-1] # Nome da coluna de bioatividade
print("Features:")
```

```
display(x)
print("Label:")
display(y)

## 2.3. Dividindo os conjuntos de **TREINO e TESTE**:

# Número para a semente do gerador de números aleatórios.
# Caso queira testar um conjunto de dados diferente, basta alterar o número aleatório.
# Algumas instâncias que estavam no conjunto treino passarão para o conjunto de teste e vice-versa.
rand = 123

# Dividir o conjunto de dados em treino e teste:
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.3, random_state=rand)

# Verificando o tamanho dos grupos:
tamanhos = [len(x_train), len(x_test)]
grupos = ['X Train', 'X Test']

# Criando um gráfico de barras
ax = sns.barplot(x=grupos, y=tamanhos, palette="husl")
plt.xticks(fontweight='bold')
# Remover as bordas superiores e direitas:
ax.spines[['top', 'right']].set_visible(False)
# plt.title("Tamanhos dos Grupos")
plt.ylabel("Size", fontweight='bold')

# # Adicionando valores numéricos acima de cada barra
for bar in ax.patches:
    plt.text(bar.get_x() + bar.get_width() / 2, bar.get_height() + 0.1, f'{int(bar.get_height())}',
             va='bottom', ha='center', color=bar.get_facecolor(), weight='bold')

plt.savefig(f'{caminho_resultados_figuras}/Treino_Teste.png', bbox_inches='tight', dpi=500)

plt.show()

### Novo Processamento (Opicional):

## Normalizar os dados:
# scaler = StandardScaler()
# x = scaler.fit_transform(x)
# y = scaler.fit_transform(y.values.reshape(-1, 1)).flatten()

## Redução de dimensionalidade com PCA:
# Redução adicional de dimensionalidade com PCA (opcional)
# pca = PCA(n_components=100)  # Ajustar o número de componentes principais
# x_reduced = pca.fit_transform(x)

## 2.4. LazyPredict:

from lazypredict.Supervised import LazyClassifier
from sklearn.model_selection import train_test_split
```

```
# Supondo que x e y são os DataFrames de features e labels, respectivamente
# Dividir os dados em conjuntos de treinamento e teste
# x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)

# Instanciar o LazyClassifier
clf = LazyClassifier(verbose=0, ignore_warnings=True, custom_metric=None, predictions=True)

# Treinamento e avaliação dos modelos
modelos, predições = clf.fit(x_train, x_test, y_train, y_test)

# Exibir os resultados
display(modelos)
display(predições)
```

Visualizando o resultado do Screening de Modelos:

```
# Visualizando os resultados de R2:
plt.figure(figsize=(5, 10))
sns.set_theme(style="darkgrid")
# Criando um gráfico de barras
ax = sns.barplot(y="Model", x="Accuracy", data=modelos, palette="husl")
ax.set(xlim=(0, 1))

# Ajusta o layout para evitar que elementos fiquem cortados; se necessário, aumente a margem esquerda
plt.tight_layout()
plt.subplots_adjust(left=0.3)  # Ajuste o valor se necessário para centralizar

# Salvando a imgem:
plt.savefig(f'{caminho_resultados_figuras}/Screening_Lazy_Accuracy.png', bbox_inches='tight', dpi=500)

# Visualizando os resultados de RMSE:
plt.figure(figsize=(5, 10))
sns.set_theme(style="darkgrid")
ax = sns.barplot(y="Model", x="F1 Score", data=modelos, palette="husl")
ax.set(xlim=(0, 10))

# Ajusta o layout para evitar que elementos fiquem cortados; se necessário, aumente a margem esquerda
plt.tight_layout()
plt.subplots_adjust(left=0.3)  # Ajuste o valor se necessário para centralizar

plt.savefig(f'{caminho_resultados_figuras}/Screening_Lazy_F1.png', bbox_inches='tight', dpi=500)
```

2.5. Pycaret:

```
# Pre-processing the data
classification_setup = classification.setup(data = df, target = "Class", fold = 10)

# Building and comparing different models
best = classification.compare_models(sort = "Accuracy", cross_validation=False)
```

```
classes_list = ["Aldicarbe", "Alimento", "Brodifacoum ", "Bromadiolona", "Clorpirifós", "Metomil", "Terbufós"]

model_name = 'xgboost' # Ajuste o nome do modelo que deseja realizar validação cruzada e gerar os gráficos de performance
best_model_2 = classification.create_model(model_name)

# Gerar a lista de classes a partir da coluna 'Class' do DataFrame
# classes_list = df['Class'].unique().tolist()

# Criar lista com os tipos de gráficos usados pelo pycaret
plot_types = ['error', 'feature', 'auc', 'confusion_matrix', 'class_report']

# Salva como PNG todos esses gráficos utilizando a lista de classes gerada dinamicamente.
for plot in plot_types:
    classification_setup.plot_model(
        best_model_2,
        plot=plot,
        plot_kwargs={'classes': classes_list},
        save=True
    )

# Diretório atual onde os gráficos são salvos
current_directory = os.getcwd()
data_hora = datetime.now().strftime("%Y%m%d_%H%M%S")

# Renomeia e Transfere os gráficos gerados pelo pycaret para uma pasta específica.
for arquivo in glob.glob(os.path.join(current_directory, "*.png")):
    # Separa o nome base e a extensão
    base, ext = os.path.splitext(os.path.basename(arquivo))
    # Cria o novo nome: data_hora, model_name e o nome original
    novo_nome = f"{data_hora}_{model_name}_{base}{ext}"
    destino = os.path.join(caminho_resultados_figuras, novo_nome)
    shutil.move(arquivo, destino)

joblib.dump(best_model_2, f'{caminho_modelos_classificacao}/{model_name}_pycaret.joblib')

# **PARTE 3: TREINAMENTO, AVALIAÇÃO E OTIMIZAÇÃO DE PARÂMETROS DO MELHOR MODELO:**

## **MODELOS DE CLASSIFICAÇÃO:**

### Importando o dataframe:

from google.colab import files
import pandas as pd
import io

# Etapa 1: Fazer upload do arquivo Excel
uploaded = files.upload()  # Isso abrirá uma janela para você selecionar o arquivo .xlsx do seu PC

# Etapa 2: Ler o conteúdo do arquivo Excel
for nome_arquivo in uploaded.keys():
    df = pd.read_excel(io.BytesIO(uploaded[nome_arquivo]))  # Lê o Excel (primeira aba por padrão)
```

```
# Etapa 3: Exibir as primeiras linhas do DataFrame
df.head()

# Função para selecionar o arquivo (CSV ou XLSX):
def selecionar_arquivo():
    root = tk.Tk()
    root.withdraw()  # Esconde a janela principal
    caminho = filedialog.askopenfilename(filetypes=[("Excel files", "*.xlsx")])
    # caminho = filedialog.askopenfilename(filetypes=[("CSV files", "*.csv")])
    root.destroy()  # Destroi a janela principal após a seleção do arquivo
    return caminho

# Selecionar o arquivo CSV
caminho = selecionar_arquivo()

# Carregar os dados do arquivo CSV
df = pd.read_excel(caminho)
# df = pd.read_csv(caminho)
df

### Atribuindo um código a cada classe:

# Atribui a cada elemento de classe um código numérico:
# target_mapping = {i: classe for i, classe in enumerate(df['Class'].unique())}
# Dicionário de mapeamento
class_map = {
    'Aldicarbe': 0,
    'Alimento': 1,
    'Brodifacoum ': 2,
    'Bromadiolona': 3,
    'Clorpirifós': 4,
    'Metomil': 5,
    'Terbufós': 6
}

# Aplicando o mapeamento à coluna 'Class'
df['Class'] = df['Class'].replace(class_map)
# Inverte o dicionário para obter: classe -> código
# inverted_mapping = {v: k for k, v in target_mapping.items()}

# # Substitui os valores de texto na coluna 'Class' pelos números correspondentes
# df['Class'] = df['Class'].map(target_mapping)

#exibe os resultados:
print(class_map)
display(df)

### Definindo os valores de x e y e os conjuntos de treino e teste:

x = df.iloc[:, 1:]  # Todas as colunas, exceto a primeira (features)
```

```
y = df.iloc[:, 0]     # Primeira coluna (label)
bioatv_name = df.columns[-1] # Nome da coluna de bioatividade
rand=123
# Dividir o conjunto de dados em treino e teste:
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.3, random_state=rand)
print("Features:")
display(x)
print("Label:")
display(y)

### Selecione o modelo de classificação que deseja implementar e ajuste parâmetros:

# Retire o comentário dos modelos nas listas a sequir:
# Atenção! O último modelo não comentado deve estar sem linha ao final.

# Criar um classificador base
# base_clf = RandomForestClassifier(
#     n_estimators=100,           # Número de árvores na floresta
#     max_depth=None,             # Profundidade máxima da árvore
#     min_samples_split=2,        # Número mínimo de amostras necessárias para dividir um nó interno
#     min_samples_leaf=1,         # Número mínimo de amostras necessárias para estar em um nó folha
#     random_state=42             # Semente para o gerador de números aleatórios
# )

base_clf = lgb.LGBMClassifier(
        force_col_wise=True,
        num_leaves=31,           # Número de folhas na árvore
        learning_rate=0.1,       # Taxa de aprendizado
        n_estimators=100,        # Número de árvores
        max_depth=-1,            # Profundidade máxima (-1 = sem limite)
        feature_fraction=1.0,    # Proporção de features a serem utilizadas em cada árvore
        random_state=rand
    )

modelos_class = [
    # ('RandomForestClassifier', RandomForestClassifier(
    #     n_estimators=100,        # Número de árvores na floresta
    #     max_depth=None,          # Profundidade máxima das árvores (None = sem limite)
    #     min_samples_split=2,     # Número mínimo de amostras necessárias para dividir um nó
    #     min_samples_leaf=1,      # Número mínimo de amostras que um nó folha deve ter
    #     max_features=None,     # Número de recursos a serem considerados para a melhor divisão
    #     random_state=rand
    # )),

    ('XGBClassifier', xgb.XGBClassifier(
        n_estimators=100,        # Número de árvores
        learning_rate=0.1,       # Taxa de aprendizado
        max_depth=None,          # Profundidade máxima para cada árvore
        tree_method='hist',   # Método de construção da árvore (gpu_hist, hist, auto)
        subsample=1.0,           # Subamostragem dos dados
        colsample_bytree=1.0,    # Subamostragem das features por árvore
```

```
    gamma=0,                # Mínima redução de perda necessária para realizar uma divisão adicional
    random_state=rand
)),

# ('HistGradientBoostingClassifier', HistGradientBoostingClassifier(
#     max_iter=100,             # Número máximo de iterações
#     learning_rate=0.1,        # Taxa de aprendizado
#     max_depth=None,           # Profundidade máxima das árvores (None = sem limite)
#     random_state=rand
# )),

# ('LGBMClassifier', lgb.LGBMClassifier(
#     force_col_wise=True,
#     num_leaves=31,            # Número de folhas na árvore
#     learning_rate=0.1,        # Taxa de aprendizado
#     n_estimators=100,         # Número de árvores
#     max_depth=-1,             # Profundidade máxima (-1 = sem limite)
#     feature_fraction=1.0,     # Proporção de features a serem utilizadas em cada árvore
#     random_state=rand
# )),

# # ('BaggingClassifier', BaggingClassifier(
# #     estimator=DecisionTreeClassifier(
# #         max_depth=None,       # Profundidade máxima das árvores (None = sem limite)
# #         min_samples_split=2,  # Número mínimo de amostras necessárias para dividir um nó
# #         min_samples_leaf=1,   # Número mínimo de amostras que um nó folha deve ter
# #         random_state=rand
# #     ),
# #     n_estimators=100,         # Número de estimadores (árvores)
# #     n_jobs=-1,                # Usa todos os núcleos disponíveis
# #     random_state=rand
# # )),

# # ('NuSVC', NuSVC(
# #     kernel='rbf',             # Tipo de kernel (núcleo)
# #     nu=0.5,                   # Parâmetro nu que controla a fração de suportes e margens de erro
# #     gamma='scale'             # Coeficiente do kernel: 'scale', 'auto' ou um valor float
# # )),

# # ('SVC', SVC(
# #     kernel='rbf',             # Tipo de kernel (núcleo)
# #     C=1.0,                    # Parâmetro de penalidade
# #     probability=True,         # Habilita a estimação de probabilidades
# #     gamma='scale',            # Coeficiente do kernel: 'scale', 'auto' ou um valor float
# #     random_state=rand
# # )),

# # ('GradientBoostingClassifier', GradientBoostingClassifier(
# #     n_estimators=100,         # Número de árvores
# #     learning_rate=0.1,        # Taxa de aprendizado
# #     max_depth=3,              # Profundidade máxima das árvores
```

```
# #     subsample=1.0,           # Fração dos dados amostrados para cada árvore
# #     max_features=None,       # Número de features a considerar para a melhor divisão
# #     random_state=rand
# # )),

# # ('ExtraTreesClassifier', ExtraTreesClassifier(
# #     n_estimators=100,        # Número de árvores na floresta
# #     max_depth=None,          # Profundidade máxima das árvores (None = sem limite)
# #     min_samples_split=2,     # Número mínimo de amostras para dividir um nó
# #     min_samples_leaf=1,      # Número mínimo de amostras que um nó folha deve ter
# #     max_features=None,     # Número de features a serem consideradas nas divisões
# #     random_state=rand
# # )),

# # ('AdaBoostClassifier', AdaBoostClassifier(
# #     n_estimators=100,        # Número de estimadores
# #     learning_rate=0.1,       # Taxa de aprendizado
# #     random_state=rand
# # )),

# # ('KNeighborsClassifier', KNeighborsClassifier(
# #     n_neighbors=5,           # Número de vizinhos a serem considerados
# #     weights='uniform',       # Pesos uniformes para todos os vizinhos: 'uniform', 'distance'
# #     algorithm='auto'     # Algoritmo para computar os vizinhos: 'auto', 'ball_tree', 'kd_tree', 'brute'
# # )),

# # ('LogisticRegression', LogisticRegression(
# #     C=1.0,                   # Parâmetro de regularização
# #     solver='lbfgs',          # Algoritmo de otimização
# #     max_iter=1000,           # Número máximo de iterações
# #     random_state=rand
# # )),

# ('CalibratedClassifierCV', CalibratedClassifierCV(
# estimator=base_clf,          # Classificador base
# method='isotonic',           # Método de calibração: 'sigmoid', 'isotonic'
# cv=5,                      # Número de dobras na validação cruzada
# n_jobs=-1                  # Número de trabalhos a serem executados em paralelo (-1 usa todos os processadores)
# )),

# ('RidgeClassifier', RidgeClassifier(
#     alpha=1.0,               # Parâmetro de regularização
#     solver='auto'            # Algoritmo para resolver o sistema: 'auto', 'svd', 'cholesky', 'lsqr', 'sparse_cg', 'sag', 'saga'
# )),

# ('RidgeClassifierCV', RidgeClassifierCV(
#     alphas=[1.0],            # Valores de alpha para busca
#     cv=5                     # Número de dobras na validação cruzada
# )),

('LinearDiscriminantAnalysis', LinearDiscriminantAnalysis(
```

```
        solver='svd'              # Algoritmo para solver: 'svd', 'lsqr', 'eigen'
    ))
]


### **TREINAMENTO E TESTE DOS MODELOS:**

# Criando um dataframe para as métricas de todos os modelos de classificação e limpando registros anteriores:
df_models_class = pd.DataFrame(columns=['Modelo', 'Accuracy Treino', 'Accuracy Teste', 'F1 Score', 'Precision', 'Recall'])
df_models_class.drop(df_models_class.index, inplace=True)

# Dicionário para armazenar os modelos treinados.
trained_models = {}

for model_name, model in modelos_class:
    try:
        # Treinamento do modelo:
        model.fit(x_train, y_train)

        # Armazena o modelo treinado para uso posterior.
        trained_models[model_name] = model

        # Previsões para treino e teste:
        y_train_pred = model.predict(x_train)
        y_test_pred = model.predict(x_test)

        # Cálculo das métricas:
        acc_train = accuracy_score(y_train, y_train_pred)
        acc_test = accuracy_score(y_test, y_test_pred)
        f1 = f1_score(y_test, y_test_pred, average='weighted')
        precision = precision_score(y_test, y_test_pred, average='weighted')
        recall = recall_score(y_test, y_test_pred, average='weighted')

        # Adicionando os resultados à tabela consolidada:
        df_models_class.loc[len(df_models_class)] = [model_name, acc_train, acc_test, f1, precision, recall]
        print(f'Modelo {model_name} avaliado com sucesso.')
    except Exception as e:
        print(f'Erro ao avaliar o modelo {model_name}: {e}')

# Encontra o modelo com a maior Accuracy de teste e, em caso de diferença menor ou igual a 0.01, seleciona o que apresentar maior F1 Score:
best_model = df_models_class[df_models_class['Accuracy Teste'] >= df_models_class['Accuracy Teste'].max() - 0.01] \
            .sort_values('F1 Score', kind='mergesort', ascending=False).iloc[0]['Modelo']

# Exibe o DataFrame consolidado:
df_models_class = df_models_class.sort_values(by=df_models_class.columns[2], ascending=False)
display(df_models_class.style.format({col: "{:.3f}" for col in df_models_class.columns if col != "Modelo"}).hide(axis="index"))

# Salva o DataFrame consolidado em um arquivo CSV:
data_hora = datetime.now().strftime('%Y%m%d_%H%M%S')
file_path = os.path.join(caminho_resultados_modelos, f'{data_hora}_resultado_modelos.csv')
```

```
df_models_class.to_csv(file_path, index=False)

# Salva o modelo com melhor performance em um arquivo .joblib:
best_model_instance = trained_models[best_model]
joblib.dump(best_model_instance, f'{caminho_modelos_classificacao}/{data_hora}_{best_model}_classifier.joblib')


### Salvando outro modelo a sua escolha:

# Salvando o segundo modelo com melhor performance em um arquivo .joblib:
data_hora = datetime.now().strftime('%Y%m%d_%H%M%S')
second_best_model = df_models_class.iloc[1]['Modelo']
second_best_model_instance = trained_models[second_best_model]
joblib.dump(second_best_model_instance,
f'{caminho_modelos_classificacao}/{data_hora}_{second_best_model}_classifier.joblib')

### Gerando gráficos de performance de um modelo selecionado:

# Function to select the model file (JOBLIB):
def selecionar_arquivo():
    root = tk.Tk()
    root.withdraw()  # Hides the main window
    caminho = filedialog.askopenfilename(filetypes=[("JOBLIB files", "*.joblib")])
    root.destroy()  # Destroys the main window after file selection
    return caminho

# Select the JOBLIB file
caminho = selecionar_arquivo()

# Load the model from the JOBLIB file
best_model = joblib.load(caminho)

# Extract the model name
best_model_name = type(best_model).__name__

# List of classes generated dynamically from the 'Class' column of df
classes_list = df['Class'].unique().tolist()

# Generate predictions on the test set
y_pred = best_model.predict(x_test)

data_hora = datetime.now().strftime('%Y%m%d_%H%M%S')

# --- Plot 1: Prediction Confidence (Error Plot) ---
# If the model has predict_proba, we can visualize the confidence of the predictions.
if hasattr(best_model, "predict_proba"):
    y_proba = best_model.predict_proba(x_test)
    max_proba = np.max(y_proba, axis=1)
    errors = (y_pred != y_test)
    plt.figure()
    plt.scatter(range(len(max_proba)), max_proba, c=errors, cmap='coolwarm', alpha=0.6)
```

```
    plt.xlabel('Samples')
    plt.ylabel('Maximum Probability')
    plt.title('Prediction Confidence (Highlighted Errors)')
    plt.savefig(os.path.join(caminho_resultados_figuras, f'{data_hora}_{best_model_name}_error_plot.png'))
    plt.close()

# --- Plot 2: Feature Importance ---
# If the model has the attribute feature_importances_
if hasattr(best_model, "feature_importances_"):
    importances = best_model.feature_importances_
    try:
        feature_names
    except NameError:
        feature_names = [f'Feature {i}' for i in range(len(importances))]

    # Get the indices of the 10 most important features
    indices = np.argsort(importances)[::-1][:10]

    # Get the importances and names of the 10 most important features
    top_importances = importances[indices]
    top_feature_names = [feature_names[i] for i in indices]

    # Create the horizontal bar chart
    plt.figure(figsize=(10, 6))
    plt.title("10 Most Important Features")
    plt.barh(range(len(top_importances)), top_importances, align='center')
    plt.yticks(range(len(top_importances)), top_feature_names)
    plt.xlabel('Importance')
    plt.gca().invert_yaxis()  # Invert y-axis so that the most important feature is at the top
    plt.tight_layout()
    plt.savefig(os.path.join(caminho_resultados_figuras, f'{data_hora}_{best_model_name}_feature_importance.png'))
    plt.close()

# --- Plot 3: ROC Curve / AUC (Multi-class, micro-average) ---
# Requires the model to have predict_proba
if hasattr(best_model, "predict_proba"):
    # Binarize the labels
    y_test_bin = label_binarize(y_test, classes=classes_list)
    y_score = best_model.predict_proba(x_test)
    # Compute the micro-average ROC curve
    fpr, tpr, _ = roc_curve(y_test_bin.ravel(), np.array(y_score).ravel())
    roc_auc = auc(fpr, tpr)
    plt.figure()
    plt.plot(fpr, tpr, label='ROC (AUC = %0.2f)' % roc_auc)
    plt.plot([0, 1], [0, 1], 'k--')
    plt.xlim([0.0, 1.0])
    plt.ylim([0.0, 1.05])
    plt.xlabel('False Positive Rate')
    plt.ylabel('True Positive Rate')
    plt.title('ROC Curve - Micro-average')
    plt.legend(loc="lower right")
```

```
    plt.savefig(os.path.join(caminho_resultados_figuras, f'{data_hora}_{best_model_name}_roc_curve.png'))
    plt.close()

# --- Plot 4: Confusion Matrix ---
unique_labels = np.unique(np.concatenate((y_test, y_pred)))
cm = confusion_matrix(y_test, y_pred, labels=unique_labels)
plt.figure()
plt.imshow(cm, interpolation='nearest', cmap=plt.cm.Blues)
plt.title('Confusion Matrix')
plt.colorbar()
tick_marks = np.arange(len(unique_labels))
plt.xticks(tick_marks, unique_labels, rotation=45)
plt.yticks(tick_marks, unique_labels)
plt.ylabel('True Class')
plt.xlabel('Predicted Class')
plt.grid(False)
thresh = cm.max() / 2.
for i in range(cm.shape[0]):
    for j in range(cm.shape[1]):
        plt.text(j, i, format(cm[i, j], 'd'), ha="center",
                 color="white" if cm[i, j] > thresh else "black")
plt.tight_layout()
plt.savefig(os.path.join(caminho_resultados_figuras, f'{data_hora}_{best_model_name}_confusion_matrix.png'))
plt.close()

# --- Plot 5: Classification Report ---
report = classification_report(y_test, y_pred, target_names=classes_list, output_dict=True)
report_df = pd.DataFrame(report).transpose()

# Ensure all classes are included in the report DataFrame
for class_name in classes_list:
    if class_name not in report_df.index:
        report_df.loc[class_name] = [0.0, 0.0, 0.0, 0.0]

plt.figure(figsize=(8, 4))
plt.axis('off')
tbl = plt.table(cellText=report_df.round(2).values,
                rowLabels=report_df.index,
                colLabels=report_df.columns,
                cellLoc='center',
                loc='center')
tbl.auto_set_font_size(False)
tbl.set_fontsize(10)
plt.title("Classification Report")
plt.savefig(os.path.join(caminho_resultados_figuras, f'{data_hora}_{best_model_name}_classification_report.png'))
plt.close()
```

REDES NEURAIS ARTIFICIAIS (RNAs):

Para valores de **Y EM VARIÁVEIS CATEGÓRICAS** usando os mesmos valores dos modelos de classificação.

MLPClassifier:

Multi-layer Perceptron Classifier

```
# Criar e treinar o modelo MLPClassifier
modelo_mlpc = MLPClassifier(hidden_layer_sizes=(100, 50), max_iter=500, random_state=rand)
modelo_mlpc.fit(x_train, y_train)

# Acurácia no treinamento e teste
acc_train = modelo_mlpc.score(x_train, y_train)
acc_test = modelo_mlpc.score(x_test, y_test)

print("Acurácia Treino: " + str(acc_train))
print("Acurácia Teste: " + str(acc_test))

# Calculando as previsões para o conjunto de teste
y_mlpc = modelo_mlpc.predict(x_test)

# Exibindo o relatório de classificação e a matriz de confusão
print("Relatório de Classificação:")
print(classification_report(y_test, y_mlpc))

print("Matriz de Confusão:")
print(confusion_matrix(y_test, y_mlpc))

# Salva o modelo em um arquivo .joblib:
modelo_mlpc_instance = modelo_mlpc
joblib.dump(modelo_mlpc_instance, f'{caminho_modelos_redes_neurais}/MLPC_RNA.joblib')
```

PARTE 7: APLICANDO OS MODELOS DE ML PARA PREDIÇÃO EM BASES DE DADOS EXTERNAS:

Nesta parte utilizaremos os modelos de ML com melhor desempenho para realizar a predição das classes em bases de dados externas.

Para um único Dataset Externo escolhido:

```
# Imports necessários
from google.colab import files
import pandas as pd
import numpy as np
import joblib
import io

# === Upload do DATASET DE TREINAMENTO (.csv ou .xlsx) ===
print("▲ Faça o upload do DATASET DE TREINAMENTO (.csv ou .xlsx)")
uploaded_int = files.upload()

# Carregar o arquivo (testando codificações se for CSV)
df_int = None
```

```
for filename in uploaded_int.keys():
    if filename.endswith('.csv'):
        encodings_testadas = ['utf-8', 'latin1', 'windows-1252', 'utf-16']
        for enc in encodings_testadas:
            try:
                df_int = pd.read_csv(io.BytesIO(uploaded_int[filename]), encoding=enc)
                print(f"✔ CSV de treinamento lido com sucesso com encoding: {enc}")
                break
            except Exception as e:
                print(f"✖ Falha com encoding {enc}: {e}")
        if df_int is None:
            raise ValueError("⚠ Não foi possível ler o CSV de treinamento com nenhuma codificação.")
    else:
        df_int = pd.read_excel(io.BytesIO(uploaded_int[filename]))
        print("✔ XLSX de treinamento lido com sucesso.")

# Separar x_int e extrair features
x_int = df_int.iloc[:, 1:]
features_train = x_int.columns.tolist()

# Criar mapeamento de índices para rótulos da variável alvo
target_mapping = {i: classe for i, classe in enumerate(df_int['Class'].unique())}

# === Upload do MODELO (.joblib) ===
print("🔼 Faça o upload do MODELO (.joblib)")
uploaded_model = files.upload()

# Carregar o modelo
for filename in uploaded_model.keys():
    best_model = joblib.load(io.BytesIO(uploaded_model[filename]))
    best_model_name = type(best_model).__name__
    print(f"✔ Modelo carregado: {best_model_name}")

# === Upload do ARQUIVO DE DADOS EXTERNOS (.csv ou .xlsx) ===
print("🔼 Faça o upload do ARQUIVO DE DADOS EXTERNOS (.csv ou .xlsx)")
uploaded_ext = files.upload()

# Carregar dados externos
for filename in uploaded_ext.keys():
    if filename.endswith('.csv'):
        df_ext = pd.read_csv(io.BytesIO(uploaded_ext[filename]))
    else:
        df_ext = pd.read_excel(io.BytesIO(uploaded_ext[filename]))

# Separar x_ext (todas as colunas exceto a primeira)
x_ext = df_ext.iloc[:, 1:]

# Verificar e alinhar features
features_to_use = [col for col in features_train if col in x_ext.columns]
```

```
missing_features = [col for col in features_train if col not in x_ext.columns]
if missing_features:
    print("⚠ As seguintes features estão ausentes no novo dataset:", missing_features)

x_ext = x_ext[features_to_use].reset_index(drop=True)

# Fazer predições
predicoes = best_model.predict(x_ext)

# Criar DataFrame com os resultados
df_predict = pd.DataFrame(predicoes, columns=['predictions'])
df_result = pd.concat([df_predict, df_ext.reset_index(drop=True)], axis=1)
df_result['Class'] = df_result['predictions'].map(target_mapping)

# Exibir resultados
print(f"\n✅ PREDICTIONS FOR THE SELECTED DATASET USING MODEL: {best_model_name}")
pd.set_option('display.max_rows', None)
display(df_result)
```

De forma automatizada para todas as bases presentes na pasta **DATASET_EXTERNO**:

```
# Defina o mapeamento de códigos para classes
target_mapping = {
    0: 'Aldicarbe',
    1: 'Alimento',
    2: 'Brodifacoum',
    3: 'Bromadiolona',
    4: 'Clorpirifós',
    5: 'Glifosato',
    6: 'Metomil',
    7: 'Terbufós'
}

# Iterar sobre os arquivos no diretório externo
for file_name in os.listdir(caminho_dataset_externo):
    if file_name.endswith(".csv"):
        # Leia o arquivo .csv considerando \t e espaço como separadores
        csv_path = os.path.join(caminho_dataset_externo, file_name)
        df_ext = pd.read_csv(csv_path)

        # Separando x :
        x_ext = df_ext.iloc[:, 1:]

        # Obtém a lista de features utilizadas no treinamento
        file_path = os.path.join(caminho_dataset_interno, 'df_Z.csv') # Alterar para o dataframe usado no treinamento e teste do
modelo
        df_train = pd.read_csv(file_path)
        rand = 123
        x_train, x_test, y_train, y_test = train_test_split(df_train.iloc[:, 1:], df_train.iloc[:, [0]], test_size=0.3, random_state=rand)
        for linha_index in range(len(x_int)):
```

```
    # Selecionar a linha desejada
    linha = x_int.iloc[linha_index]

    # Extrair os valores de x (nomes das colunas) e y (valores da linha)
    features_train = linha.index.astype(float)

# Garante que as features do novo dataset sejam apenas aquelas presentes no treinamento, mantendo a ordem original
features_to_use = [col for col in features_train if col in x_ext.columns]

# Opcional: avisa se alguma feature utilizada no treinamento não estiver presente no novo dataset
missing_features = [col for col in features_train if col not in x_ext.columns]
if missing_features:
    print("Atenção: as seguintes features estão ausentes no novo dataset:", missing_features)

# Filtra o novo dataframe para manter apenas as colunas de interesse
x_ext = x_ext[features_to_use].reset_index(drop=True)

# Carrega o modelo salvo em .joblib
loaded_model = joblib.load(f'{caminho_resultados_modelos}/best_model.joblib') # Alterar para o modelo que será aplicado nas predições

# Aplica o modelo carregado para fazer as predições usando somente as features alinhadas
predicoes = loaded_model.predict(x_ext)

# Converte o resultado para um dataframe em pandas:
df_predict = pd.DataFrame(predicoes, columns=['predictions'])
df_result = pd.concat([df_predict, df_ext], axis=1)

# Interpreta os códigos gerando o valor da classe:
df_result['Class'] = df_result['predictions'].map(target_mapping)

# Salvar o DataFrame resultante em um arquivo CSV
data_hora = datetime.now().strftime('%Y%m%d_%H%M%S')
file_path = os.path.join(caminho_resultados_predicoes, f'{data_hora}_df_result.csv')
df_result.to_csv(file_path, index=False)

print(f'PREDICTIONS FOR THE EXTERNAL DATASET USING THE MODEL {loaded_model}')
display(df_result)
```