Serviço Público Federal
Ministério da Educação
Fundação Universidade Federal de Mato Grosso do Sul

Curso de Física Bacharelado

Trabalho de Conclusão de curso

RANDOM FOREST PARA A CLASSIFICAÇÃO DE ESPECTROS: TÉCNICAS DE NORMALIZAÇÃO E SELEÇÃO DE VARIÁVEIS

ALAN GABRIEL VARELA TARGINO

CAMPO GRANDE, MS

2024

RANDOM FOREST PARA A CLASSIFICAÇÃO DE ESPECTROS LIBS: TÉCNICAS DE NORMALIZAÇÃO E SELEÇÃO DE VARIÁVEIS

Alan Gabriel Varela Targino

Orientador: Bruno Spolon Marangoni

Trabalho de Conclusão de Curso apresentado como parte das atividades para obtenção do título de Físico, do curso de Física Bacharelado da Fundação UFMS

Campo Grande, MS
2024

“Irmão, você não percebeu
Que você é o único representante
Do seu sonho na face da terra
Se isso não fizer você correr, chapa
Eu não sei o que vai.”
(Leandro Roque de Oliveira, Emicida)

AGRADECIMENTOS

Agradeço inicialmente aos meus pais, Danielle e Alan, que sempre me apoiaram na minha jornada e provieram todas as condições para que eu pudesse me dedicar aos estudos. Aos meus irmãos, Luís Fernando e Ana Beatriz, que me fizeram companhia durante toda a minha graduação, proporcionando momentos de descontração durante as sessões de resolução de exercícios em casa. Aos meus avós, Hélio e Jovita por terem um papel fundamental na minha criação, bem como meus tios, Ricardo e Luciane. À minha avó Izilda que me deu a oportunidade de desenvolver mais habilidades comunicativas e de conciliar o trabalho com a graduação.

Agradeço também a meu amor, Adrielli, por estar comigo durante todo o processo de escrita do TCC e final da graduação, pelos momentos de descanso e respiro aos fins de semana, pelo apoio incondicional à minhas aspirações, pelas conversas e parceria.

Agradeço também aos amigos e colegas que fiz antes e durante a graduação: Ao Estevão, pelas ligações diárias e intermináveis durante a pandemia, para resolver listas e conversar sobre a vida; A meu camarada Cássio, pelas conversas descontraídas, parceria, ajuda, debates e indignações diárias; Ao João Miguel pela ajuda, risadas e parceria; Ao Carlos Henrique, pela companhia, pelos momentos descontraídos e pelas discordâncias futebolísticas; Ao Wender, Mateus, João Augusto e Giovanni, pela companhia e parceria durante o fim da graduação; Ao Murilo pela ajuda, momentos descontraídos e parceria; Ao Nicolas Andrei e Luis Spengler pelas conversas na sala do PET e companhia nessa reta final de graduação.

Agradeço ao PET-Física e a todos os seus integrantes pelos inúmeros aprendizados que obtive durante a graduação, especialmente àqueles presentes na Comissão Organizadora das VI e VII Semana da Física.

A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) (CAPES/PROCAD nº 88881.516322/2020-01).

Por fim agradeço aos professores que me conduziram durante a graduação, em especial ao meu orientador Bruno S. Marangoni, por me auxiliar com as dúvidas, me guiar no desenvolvimento das análises e no processo de escrita deste trabalho, pela sua serenidade, calma e responsabilidade em realizar isso de forma leve e tranquila.

RESUMO

Pesquisas recentes evidenciam o uso da técnica LIBS (*laser-induced breakdown spectroscopy*) para a classificação de materiais nas mais diversas áreas. Na maioria dos casos, essa classificação é realizada por meio de algoritmos de aprendizado de máquina. Nesse sentido, outros trabalhos mostram a eficiência na associação de métodos de normalização, seleção e redução de variáveis com os classificadores que utilizam aprendizado de máquina, dentre eles o *Random Forest*. O presente trabalho apresenta um protocolo que utiliza técnicas de normalização, seleção e redução de variáveis em associação com o *Random Forest* na classificação de resíduos de armas de fogo provenientes de munições não tóxicas. Os espectros foram normalizados utilizando *Standard Normal Variate*, posteriormente foram aplicados a Análise das Componentes Principais e a técnica Boruta, de forma permutada, para reduzir e selecionar variáveis com a finalidade de reduzir a dimensionalidade dos dados e treinar o modelo de classificação *Random Forest*. Enquanto treinados, os dados foram testados utilizando a validação interna, por meio do método *leave-one-out cross-validation* e os hiperparâmetros do *Random Forest* foram otimizados por meio do uso do *Grid Search* de modo a fornecer a melhor acurácia possível para o modelo. Esse melhor resultado demonstrou uma acurácia de 100% para ambas as formas de validação (interna e externa), evidenciando a eficiência do protocolo desenvolvido.

Palavras-chave: Random Forest, Boruta, Espectroscopia LIBS, Classificação.

ABSTRACT

Recent research highlights the use of the LIBS (Laser-Induced Breakdown Spectroscopy) technique for material classification across various fields. In most cases, this classification is carried out using machine learning algorithms. In this regard, other studies show the effectiveness of combining normalization, feature selection, and dimensionality reduction methods with machine learning classifiers, including Random Forest. This study presents a protocol that employs normalization, feature selection, and dimensionality reduction techniques in conjunction with Random Forest for classifying firearm residue from non-toxic ammunition. The spectra were normalized using Standard Normal Variate (SNV), and then Principal Component Analysis (PCA) and the Boruta technique were applied in a permuted fashion to reduce and select features, aiming to reduce data dimensionality and train the Random Forest classification model. While training, the data were tested using internal validation through leave-one-out cross-validation, and the Random Forest hyperparameters were optimized using Grid Search to achieve the best possible accuracy for the model. This optimal result demonstrated 100% accuracy for both internal and external validation, highlighting the effectiveness of the developed protocol.

Keywords: Random Forest, Boruta, LIBS Spectroscopy, Classification.

LISTA DE FIGURAS E TABELAS

SUMÁRIO

1. INTRODUÇÃO

A técnica LIBS (*laser induced breakdown spectroscopy*) combina emissão e ablação a laser em uma amostra, obtendo um espectro de emissão/absorção dessa amostra. O espectro de emissão/absorção contém a assinatura espectral dos elementos que compõem a amostra e possibilita a identificação dela. Além disso, a técnica LIBS, aliada ao *Random Forest,* é amplamente utilizada para análise e classificação de diferentes tipos de amostras (Tang, 2015; Liang, 2020; Chen, 2021). Ademais, a combinação do *Random Forest* com a técnica de seleção de variáveis *Boruta* (S. S. Kumar and T. Shaikh, 2017) e, de forma independente, com a técnica de redução de dimensionalidade PCA (*Principal Componentes Analysis)* (Jia, 2016) se mostraram bastante eficazes na obtenção de uma melhor acurácia na classificação de dados.

Em 2021, o Brasil registrou 47.847 homicídios no total, sendo 33.039, 69,1% do total, ocasionados por disparos de arma de fogo. Além disso, de 2011 a 2021, houveram em média 4.492 homicídios ocultos ao ano, nos quais não foi possível identificar a causa (Atlas 2023). Nesse sentido, a detecção e caracterização de resíduos de armas de fogo (GSR - *gunshot residues*) têm grande importância na solução de casos de crimes envolvendo esses disparos. Essa identificação é feita analisando os resíduos orgânicos e inorgânicos que são espalhados no momento do disparo e se depositam na pele e roupas do atirador, em pessoas próximas e no ambiente ao redor. Essa análise é realizada utilizando testes de identificação de metais pesados (Pb, Sb e Ba) característicos desse processo, como exemplo: os testes colorimétricos, a Microscopia Eletrônica de Varredura (MEV) e a técnica LIBS (*laser-induced breakdown spectroscopy*). Ocorre que os testes colorimétricos não são suficientemente eficazes na caracterização de resíduos advindos de munições não tóxicas (NTA - *non-toxic ammunition*), pois os GSR não são compostos por metais pesados. Além disso, há também uma alteração na morfologia das partículas de GSR-NTA, em relação aos GSR convencionais, dificultando a identificação via MEV. Nesse âmbito, faz-se necessário o desenvolvimento de alternativas para a caracterização de munições NTA (Carneiro, 2022).

Desse modo, analisando a literatura, nota-se que a técnica LIBS, já é utilizada para caracterizar GSR de munições usuais (Dona-Ferñandez, 2018), portanto é razoável testar sua aplicabilidade à caracterização de munições NTA. Tendo isso em vista, o intuito deste trabalho é

utilizar a técnica LIBS (*laser induced breakdown spectroscopy*) juntamente à técnicas de normalização (SNV), seleção de variáveis (Boruta), Análise de Componentes Principais (PCA) e classificação utilizando aprendizado de máquina (*Random Forest*).

2. REVISÃO BIBLIOGRÁFICA

Os resíduos de arma de fogo (GSRs - *Gunshot Residues*) são conjuntos de compostos inflamáveis ou não, que reagem com o propelente e o *primer* da munição, bem como seus componentes metálicos e suas munições. A composição do *primer* faz com que os GSRs sejam identificados pela presença simultânea de Chumbo (Pb), Bário (Ba) e Antimônio (Sb), além das partículas de GSR possuírem diâmetro de 0,5~10µm com morfologia esferoidal. Dessa forma, quando o gatilho é apertado, as partículas são vaporizadas pela alta temperatura e pressão e então expelidas através das frestas da arma, são espalhadas no local do disparo e depositadas nas mãos do atirador e ao seu redor. Nesse sentido, esses resíduos são utilizados para investigar incidentes envolvendo disparos de arma de fogo, comprovando suspeitas e auxiliando na reconstrução da cena do crime (CARNEIRO, 2022).

Uma das formas de analisar os resíduos de arma de fogo é o exame das amostras utilizando a microscopia eletrônica de varredura com detector de energia de dispersão de raios X (MEV/EDX). Esse método é utilizado para analisar materiais em uma escala nanoscópica a microscópica. A análise é feita através da aplicação de um feixe de elétrons em um determinado *range* de energia (normalmente 100 - 30.000 eV). Esse feixe é focalizado e os elétrons entram na amostra a uma determinada profundidade, isso gera os sinais captados pelos detectores e então é produzida uma imagem da amostra (MOHAMMED, 2018). Nesse âmbito, o MEV/EDX tem se mostrado bastante eficaz, uma vez que possibilita examinar objetos sem danificá-los, e é capaz de extrair informações sobre sua morfologia e composição, permitindo assim identificar partículas características de GSR (Pb, Ba e Sb) (BROZEK-MUCHA, 2001).

Uma alternativa ao MEV/EDX é o teste colorimétrico, como por exemplo os testes de Rodizonato de Sódio e de Griess. O teste de rodizonato de sódio é feito para detectar o Chumbo nas roupas ou mãos do atirador. O rodizonato de sódio reage com íons da camada bivalente do metal, causando uma alteração na cor da amostra, em um tom azul-violeta. Já o teste de Griess, utiliza um reagente descrito por Peter Griess em 1858 para identificar nitrito presente no resíduo de arma de fogo. Nesse teste, é realizado um procedimento com papel de Bromida tratado com 2-naftilamina 4:8 ácido dissulfônico, o papel é então colocado em cima da amostra, coberto com

20% de CH_3COOH e então pressionado com ferro quente que então faz surgir uma cor vermelho escuro que indica a presença de GSR na amostra (Shivastava, 2021).

Recentemente, o uso de munições não tóxicas (NTA), sem o uso de metais no *primer,* se tornaram um problema na caracterização de GSRs, uma vez que os *primes* passaram a possuir elementos como Alumínio (Al), Silício (Si), Potássio (K), Titânio (Ti), Zinco (Zn) e Enxofre (S). Assim, as partículas de GSR-NTA são mais difíceis de distinguir de partículas do ambiente do que as de GSR convencionais. Apesar disso, o MEV/EDX ainda se mostra eficaz na caracterização de GSR-NTA, com ressalvas, uma vez que é necessário se atentar a marcadores diferentes do habitual (Pb, Sb e Ba), como Ti-Zn-K-Cu-Zn e Al-Si-K-S-Cu-Zn (Romano, 2019).

Além disso, as desvantagens do uso do MEV, como seu alto custo, a demora na análise e a dificuldade em preparar as amostras, fazem com que o uso do teste colorimétrico seja mais utilizado (Romolo, 2001). Nesse sentido, o uso de testes colorimétricos na caracterização de munições NTA se faz ineficaz, uma vez que eles se baseiam na identificação de metais, como Chumbo, que não se fazem presentes nos resíduos provenientes de munições não tóxicas (Costa, 2016).

A técnica LIBS, é amplamente utilizada na análise e caracterização de marcadores químicos em amostras em diversas áreas, como biomedicina (Dixon e Hahn, 2005), arqueologia (Melessanaki et al., 2005) e perícia forense (Dockery, 2003) onde se mostrou rápida, de simples manuseio e versátil na análise de resíduos presentes nas mãos de atiradores (Almirall, 2015). A versatilidade do LIBS permite a análise de inúmeros materiais, superfícies e inclusive seu uso se mostrou ser viável na reconstrução de incidentes envolvendo armas de fogo (Rejos, 2020). Em complemento, a análise utilizando LIBS se mostrou eficaz na caracterização de munições sem chumbo, sendo significativamente mais eficaz que o MEV/EDX (Fambro, 2017). O uso da técnica LIBS é frequentemente associada ao uso de classificadores de aprendizado de máquina, como *Support Vector Machine*, redes neurais, *Random Forest*, dentre outros e a técnicas de seleção de variáveis e redução de dimensionalidade, como é o caso do Boruta e da Análise das Componentes Principais, respectivamente.

A técnica LIBS foi aplicada juntamente ao *Random Forest* nas mais diversas áreas e obteve sucesso, como exemplo: sua aplicação na análise quantitativa do nível de pH no solo, onde obteve uma acurácia de 99,95% (Chen, 2021); sua aplicação na classificação e caracterização de componentes de salvia miltiorrhiza, um importante e tradicional medicamento

chinês, onde obteve uma acurácia de 96,19% (Liang, 2020); sua aplicação na classificação de amostras de *slag*, pedaços rochosos, subprodutos da produção de aço, onde obteve uma acurácia de 97,78% (Tang, 2015).

A técnica Boruta é um método de seleção de variáveis, baseado no *Random Forest*, que funciona como um filtro, removendo as variáveis que não possuem muita relevância no conjunto de dados. Seu uso funcionou bem em algoritmos de regressão e classificação, bem como aliado ao *Random Forest* ocasionou um aumento na acurácia de 80,9% a 98,88% no conjunto de dados de Cleveland, um repositório de *machine learning* utilizado para aprendizado e testes de modelos de classificação (S.S. Kumar e T.. Shaikh, 2017). Além disso, o Boruta também possui aplicação na exploração da variabilidade espacial de metais pesados na superfície e interior do solo (Shaheen, 2018) e, mais recentemente, aplicado juntamente ao LIBS, onde selecionou as variáveis que foram posteriormente classificadas por um algoritmo de Aprendizado de Máquina Extremo (EML), onde o uso do Boruta aumentou o coeficiente de correlação de predição (R_p^2) de 0,8417 para 0,9962 (Ding, 2024).

Além do Boruta, a Análise das Componentes Principais (PCA - *Principal Component Analysis*), uma ferramenta utilizada na redução da dimensionalidade dos dados, também é utilizada em conjunto ao *Random Forest* e ao LIBS. Seu uso em associação com o LIBS é amplamente difundido, seja na classificação de tipos de plástico (Unnikrishnan, 2013), na classificação de amostras do solo em áreas geotermais (Chatterjee, 2018) e na avaliação de espécimes de arroz (Ribeiro, 2020). Em associação com o *Random Forest*, seu uso é conhecido em aplicações como: método de reconhecimento facial, onde obteve resultados satisfatórios, com acurácias em torno de 70%~80%; em um sistema de reconhecimento facial, onde obteve uma acurácia de 96,78% (Waskle, 2020).

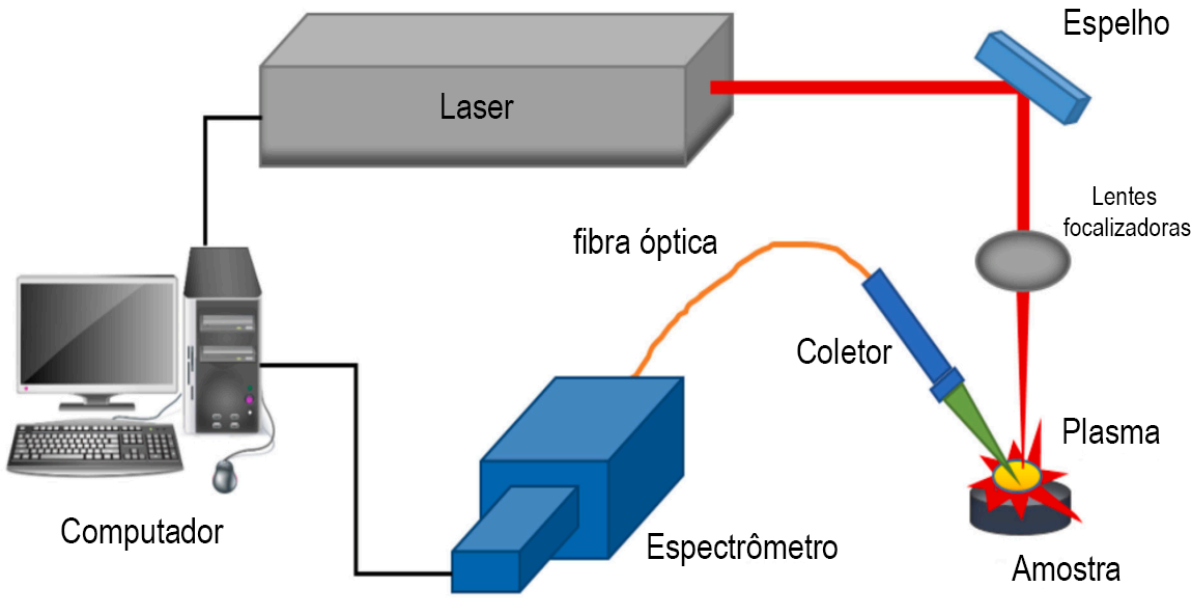
Tendo isso em vista, a aplicação da técnica LIBS, em conjunto com o PCA, Boruta, e *Random Forest* - como ferramentas de redução de dimensionalidade, seleção de variáveis e classificação dos dados - se mostra uma alternativa promissora em relação aos métodos já existentes de classificação de resíduos de arma de fogo provenientes de munições não tóxicas, dada a velocidade em que a análise é feita e a acessibilidade do uso do LIBS.

3. METODOLOGIA

3.1. LIBS - Laser induced breakdown spectroscopy

A técnica LIBS utiliza o conceito de espectroscopia de emissão atômica para realizar inferências de uma determinada amostra. A instrumentação (Figura 1) é composta por um laser de alta energia, um sistema óptico que focaliza o *laser*, um espectrômetro óptico que realiza a leitura dos sinais e um computador que os processa.

Figura 1 - Esquema de Instrumentação LIBS

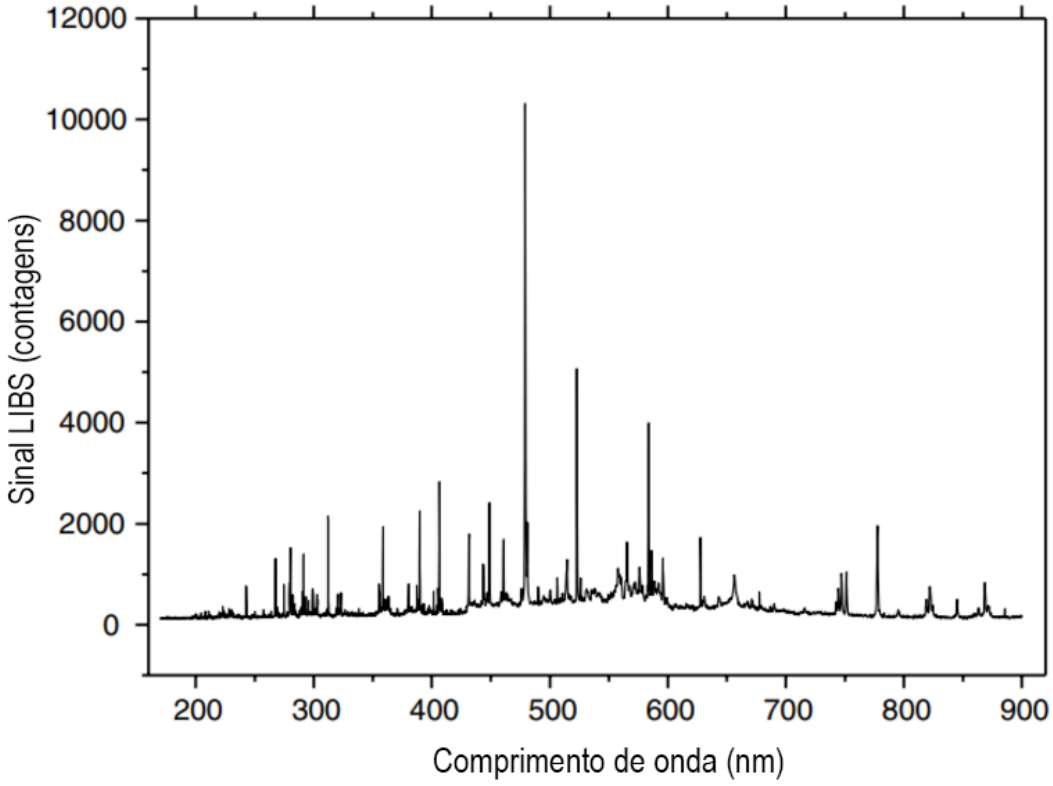


Fonte: Referência 45, adaptada pelo autor.

O processo é realizado da seguinte maneira: o laser é focalizado na amostra, removendo parte de sua superfície e excitando seus átomos, elétrons e íons, que fazem formar um plasma a uma temperatura da ordem de $10^5 K$, ocasionando uma rápida expansão (na ordem da velocidade do som) e em seguida um rápido resfriamento. No processo de resfriamento, os átomos de maior energia, presentes na amostra, transitam para um estado de menor energia emitindo radiação em comprimento de onda e intensidade característicos dos elementos que compõem a amostra. Essas linhas de emissão são coletadas pelo espectrômetro e analisadas pelo computador (Zhang, 2021),

formando o espectro LIBS, que contém os picos de emissão característicos da amostra de cada amostra, um exemplo é mostrado na Figura 2.

Figura 2 - Exemplo de espectro LIBS



Fonte: Referência 31, adaptada pelo autor.

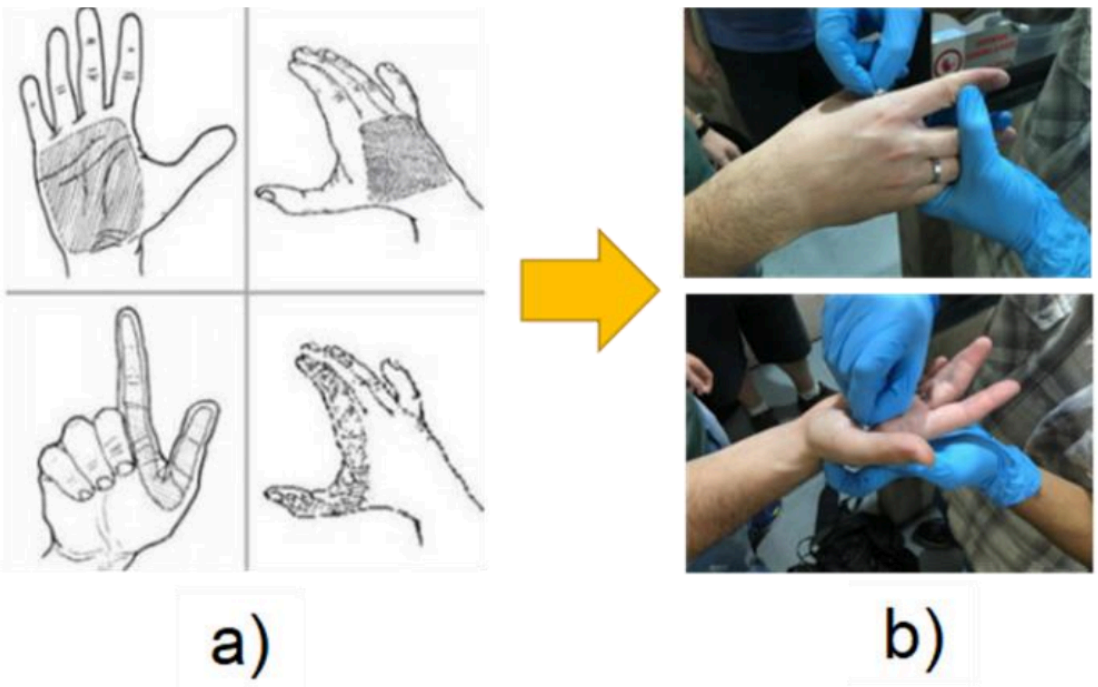
3.2. Coleta de amostras

As amostras dos resíduos de arma de fogo foram recolhidas conforme o protocolo padrão para a coleta de resíduos de disparos de armas de fogo, descrito no Procedimento Operacional Padrão (POP) nº1.4, publicado em setembro de 2013 pela Secretaria Nacional de Segurança Pública (SENASP) do Brasil. Nesse protocolo, foi utilizada uma fita adesiva transparente de dupla face da marca 3M, empregada na coleta em campo pela polícia brasileira e por laboratórios forenses, para obter amostras das mãos de quatro voluntários que haviam sido previamente lavadas (com água e detergente), tanto antes do disparo quanto após (Figura 3). O processo de coleta foi realizado por peritos forenses designados ao Centro Regional de Criminalística

(NRC/CR), à Unidade Regional de Perícias e Identificação de Costa Rica/MS (URPI/MS) e à Coordenação Geral de Perícias de MS (CGP).

Posteriormente, essas amostras foram então enviadas para análises espectroscópicas na Universidade Federal de Mato Grosso do Sul (UFMS). Na UFMS, as amostras foram classificadas em dois grupos: "Atirador" e "Não Atirador", e analisadas por LIBS na faixa espectral (190-1050 nm). A luz do plasma foi capturada por um espectrômetro Stellarnet, operando na faixa mencionada, com resolução óptica de 0,2 nm. O equipamento usou um laser primário Nd de 1064 nm com duração de pulso de 8 ns (marca Quanta-Ray®, Spectra-Physics). A sincronização do sistema foi realizada com um gerador de atraso comercial, com precisão de sub-nanosegundo. Um único pulso foi disparado para cada medição LIBS, a fim de detectar as partículas de GSR depositadas na superfície. O atraso entre o pulso e a captura pelo espectrômetro foi otimizado para 500 ns, visando a melhor relação sinal-ruído. Os espectros foram registrados em intervalos regulares na superfície da fita, com uma sequência de 15 leituras espectrais por faixa de coleta (Cioccia, 2023).

Figura 3 - a) área das mãos na qual foi realizada a coleta, b) procedimento da coleta de dados



Fonte: Referência 28.

3.3. SNV - *Standard Normal Variate*

A técnica LIBS, como as demais técnicas de espectroscopia, é influenciada por flutuações nos sinais captados, devido a propriedades do plasma (Han, 2010). Sabendo disso, é necessário reduzir essas flutuações e, nesse tarefa, o SNV (*Standard Normal Variate*) se mostra uma técnica de normalização eficaz (Guezenoc, 2019).

O processo de cálculo do SNV leva em conta cada espectro individualmente, centralizando-o em seu valor médio e dividindo pelo seu desvio padrão (Equação 1).

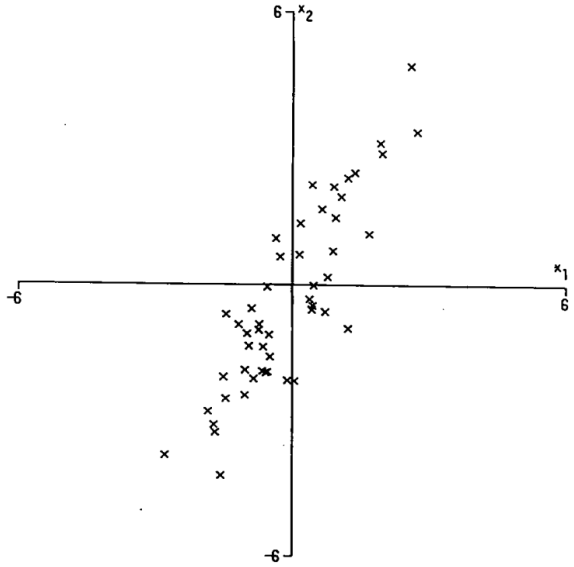
$$I_l^{SNV} = \frac{I_l - I_{Média}}{\sigma} \quad (1)$$

Onde: I_l^{SNV} = l-ésima intensidade normalizada; I_l = l-ésima intensidade do espectro original; $I_{Média}$ = média das intensidades do espectro original; e σ = desvio padrão do espectro original.

3.4. PCA - *Principal Component Analysis*

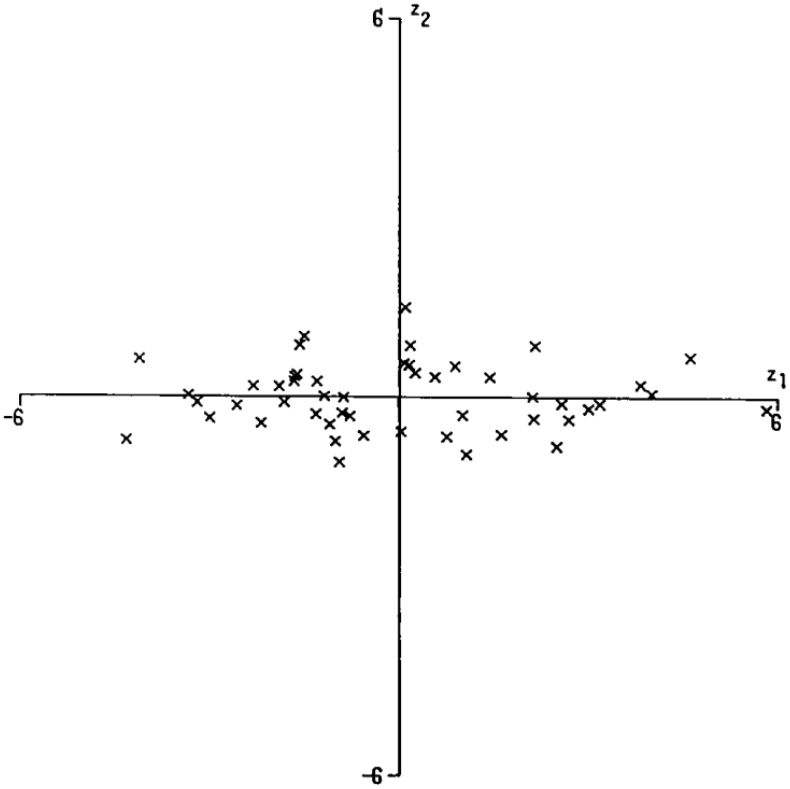
A Análise de Componentes Principais (PCA - *Principal Component Analysis)* é utilizada para reduzir a dimensionalidade de um conjunto de dados, utilizando como base a correlação entre suas variáveis. Dadas p variáveis ($x_1, x_2, \dots, x_p$) com n medidas cada, é construída uma função na forma $z = \alpha_1 x_1 + \alpha_2 x_2 + \dots + \alpha_p x_p$, onde $\alpha_1, \alpha_2, \dots \alpha_p$ são constantes. A partir disso, é calculada a variância dessa função e aquela que tiver maior variância é chamada 1º Componente Principal, a 2º maior variância a 2º Componente Principal e assim por diante (Jolliffe, 1990). As Figura 4 e 5 exemplificam o processo de transformação de duas variáveis (x_1 e x_2) em duas componentes principais (z_1 e z_2).

Figura 4 - 50 pontos com respeito às variáveis x_1 e x_2



Fonte: Referência 22.

Figura 5 - 50 pontos, com respeito às componentes principais z_1 e z_2



Fonte: Referência 22.

Também é possível escrever a função z em notação matricial (Equação 2):

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\alpha} \tag{2}$$

Onde **Z** é a matriz que contém os valores da função z, **α** é a matriz das constantes α e **X** é a matriz que contém as variáveis (x_1, x_2,..., x_p) para cada uma das n medidas. Como **Z** é uma combinação linear, a sua variância é dada por var(**Z**) = var(**Xα**) = **α'Sα**, onde **α'** é a transposta de **α** e **S** é a matriz de covariância associada à matriz **X**. Para resolver o problema da maximização da variância, devemos impor que **α'α =** 1 e então podemos utilizar o multiplicador de Lagrange (λ) e reescrever var(**Xα**) na forma da Equação 3 (Jolliffe, 2016):

$$\text{var}(\mathbf{X}\boldsymbol{\alpha}) = \boldsymbol{\alpha}'\mathbf{S}\boldsymbol{\alpha} - \lambda(\boldsymbol{\alpha}'\boldsymbol{\alpha} - 1) \tag{3}$$

Derivando a Equação 3 em relação a **α** e igualando a **0** (vetor nulo), obtemos a Equação 4, que maximiza a função var(**Z**):

$$\mathbf{S}\boldsymbol{\alpha} - \lambda\boldsymbol{\alpha} = \mathbf{0} \rightarrow \mathbf{S}\boldsymbol{\alpha} = \lambda\boldsymbol{\alpha} \tag{4}$$

Queremos também que λ seja máximo e como os autovalores e autovetores são definidos pela Equação 4, substituindo ela na Equação 3, obtemos que var(**Xα**) = **α'**λ**α** = λ, mostrando que λ indica justamente a variância, logo também é máximo.

Sendo assim, as combinações lineares $\mathbf{X}\boldsymbol{\alpha}_k$ são chamadas componentes principais do conjunto de dados, as matrizes $\boldsymbol{\alpha}_k$ são chamadas de *loadings* e os elementos das componentes principais são chamados de *scores.*

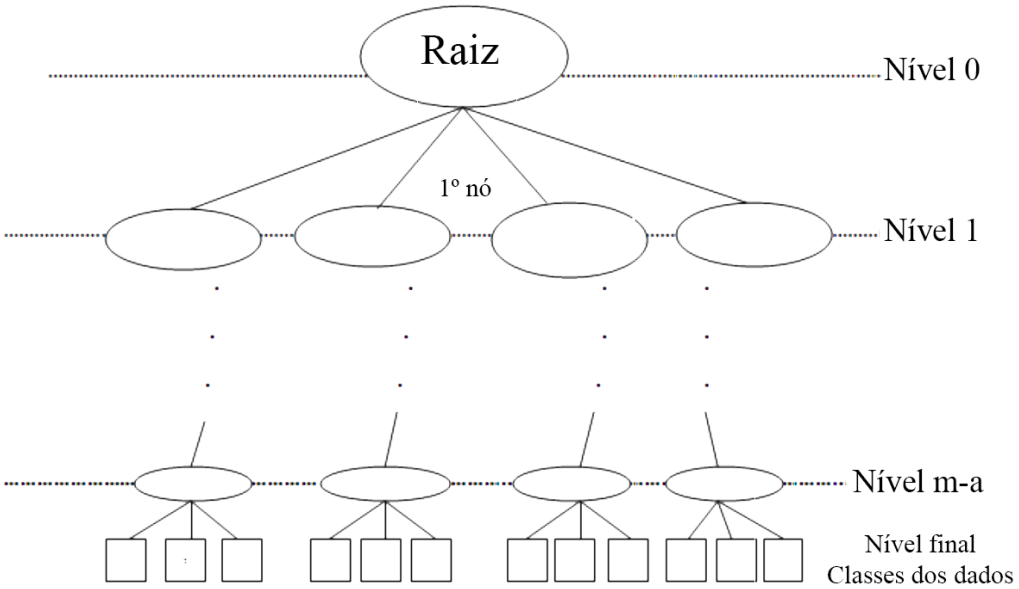
3.5.Random Forest Classifier

O *Random Forest Classifier* se baseia no modelo de árvores de decisão (*Decision Trees*). Os *Decision Trees* (DTs), são modelos preditivos que analisam os dados utilizando uma cadeia de decisões com uma estrutura semelhante a uma árvore (Figura 6).

Nesse modelo, cada nó representa um teste em uma determinada variável, cada ramo representa um caminho a ser percorrido, de acordo com a resposta obtida no nó anterior e cada folha representa a classe atribuída às variáveis (**X**). O nó raiz não possui ramificações anteriores e todas as variáveis (**X**) a serem classificadas são inseridas nele. A árvore realiza a classificação através das ramificações presentes em cada nó. Esse processo é realizado até que as variáveis alcancem o nível final, onde os dados são enfim classificados (Guerts, 2009).

Existem diversas formas de construir uma árvore de decisão, porém algumas limitações são recorrentes nesses modelos: A complexidade da árvore tem papel crucial na acurácia do modelo e o seu tamanho cresce com o tamanho dos dados, de modo que para conjunto de dados volumosos, a árvore torna-se cada vez mais complexa, usar árvores compactas para analisar grandes conjuntos de dados causa uma perda grande na acurácia; A mudança na ordem das variáveis altera a construção da árvore de decisão, de modo que caso ocorra, é necessário refazer a árvore de decisão; As DTs também possui problemas em classificar dados que possuem ruídos, frequentemente atribuindo a eles uma importância na classificação das variáveis, o que torna o modelo viciado, com pouca capacidade de generalização. Neste âmbito, existem modelos que visam contornar tais limitações, como é o caso de *Ensembles*, *Bagging*, *Boosting*, e o *Random Forest* (Priyanka, 2020).

Figura 6 - Esquema exemplificando um modelo de Árvore de Decisão



Fonte: Referência 34, adaptado pelo autor.

Na classificação utilizando *Ensembles*, diversas árvores são criadas para analisar determinada parte do conjunto de dados e então é realizada uma votação com base na classificação de cada árvore e a classe mais votada é a escolhida pelo modelo como resposta. Esse método possui uma acurácia maior que qualquer uma das árvores de decisão presentes nos *Ensembles*. No modelo chamado *Bagging*, são feitas cópias das árvores de decisão, nas quais são inseridos o conjunto de dados, juntamente a algumas cópias, que então são classificadas. A classificação do modelo também é realizada por votação, a classe mais votada pelas árvores é a definida pelo modelo. No modelo *Boosting*, são criados classificadores que, iterativamente, têm o peso do seu voto definido com base na taxa de erro de sua classificação. A classificação do modelo é proveniente de uma votação que considera o peso dado ao voto de cada classificador (Priyanka, 2020).

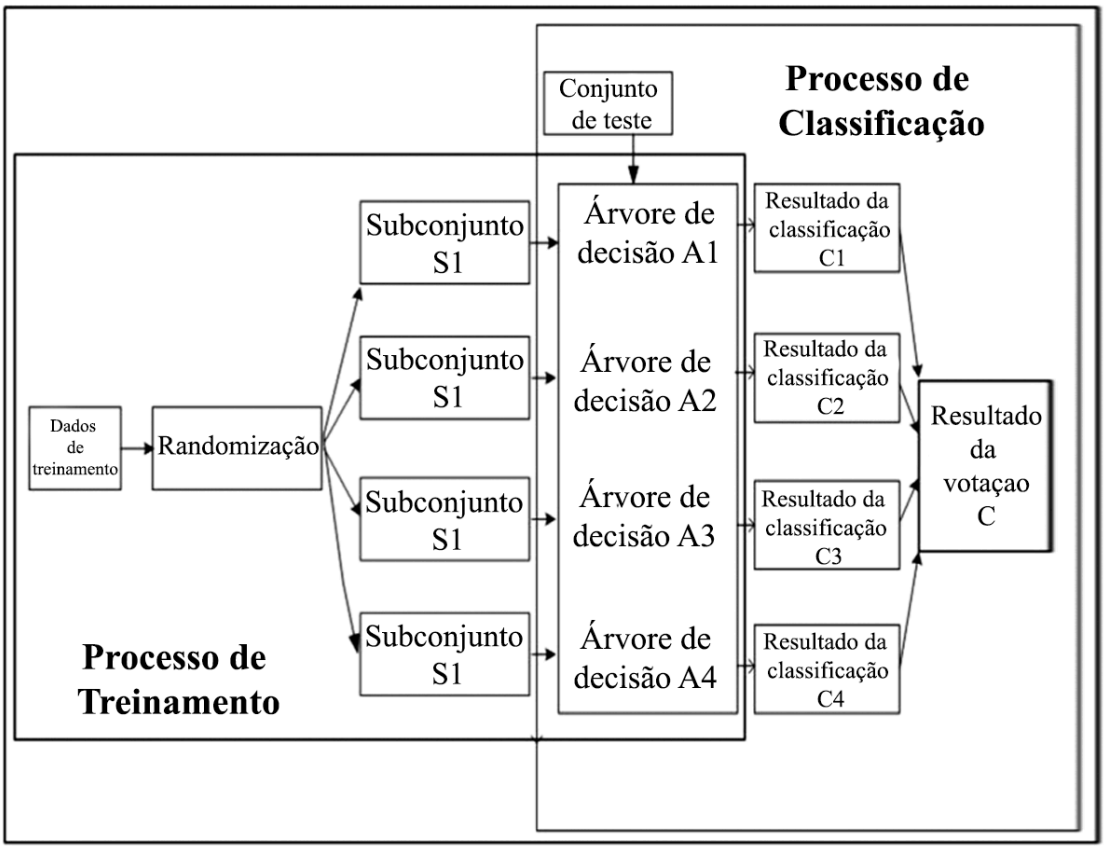
O *Random Forest* utiliza conceitos dos modelos dos tipos *Ensembles* e *Bagging* e adiciona aleatoriedade a esses métodos, o que o faz se destacar perante eles. Nesse sentido, o *Random Forest Classifier* consiste de um conjunto de classificadores com N árvores de decisão, de modo que para a k-ésima árvore, é gerado um vetor aleatório θ_k independente dos vetores anteriores $\theta_1 \ldots \theta_{k-1}$, mas com a mesma distribuição. A partir disso, uma árvore de decisão é gerada utilizando os dados de treino e θ_k, resultando em um classificador $h(\mathbf{x}, \theta_k)$, onde $\mathbf{x}$ é o vetor composto pelas variáveis de entrada. Dessa forma, o *Random Forest* é o conjunto desses classificadores $\{h(\mathbf{x}, \theta_k), k = 1, \ldots, N\}$ pelos quais os dados são passados, suas classes são definidas por cada classificador, e a classe que obtiver mais votos, ou seja, for classificada individualmente mais vezes, é a classificação obtida pelo *Random Forest (Breiman, 2001)*.

Um resumo do funcionamento do modelo de classificação é apresentado na Figura 7, nele é exemplificado seu uso com apenas 3 árvores de decisão e um passo a passo, de modo a tornar a explicação mais clara, é apresentada a seguir (Parmar, 2018):

1. É definido um valor para "M", que define a quantidade de cada subconjunto de variáveis (*shadow features*).
2. É selecionado um novo subconjunto θ_k dos dados, de forma randômica e de acordo com M. Cada subconjunto θ_k é independente dos demais subconjuntos $\theta_1, \ldots, \theta_N$.
3. Os subconjuntos são utilizados para criar árvores de decisão para cada grupo do conjunto de treinamento.

4. É escolhido um novo θ_k e o processo é repetido até que todos os subconjuntos sejam validados.
5. São inseridos os dados do subconjunto de teste. Cada árvore decide a classificação da amostra e então é fornecido o resultado do modelo, levando em conta a votação das árvores de decisão.

Figura 7 - Esquema da estrutura de funcionamento do Random Forest Classifier utilizando 3 árvores de decisão



Fonte: Referência 32, adaptada pelo autor.

3.5.Boruta

O *Random Forest*, mesmo com sua aleatoriedade intrínseca, ainda sofre influência de variáveis com baixa correlação em relação ao resultado. Nesse âmbito, o Boruta atua selecionando as variáveis que mais influenciam nos dados, adicionando ainda mais aleatoriedade ao modelo.

De forma resumida, o Boruta faz uma cópia de parte dos dados (*shadow features*), os embaralha e adiciona aos dados, comparando seus impactos com as demais variáveis, de modo a selecionar aquelas que têm mais importância que as *shadow features*. O cálculo da importância é feito utilizando o *z-score*, descrito pela Equação 5.

$$Z = \frac{x - x_{med}}{\sigma} \tag{5}$$

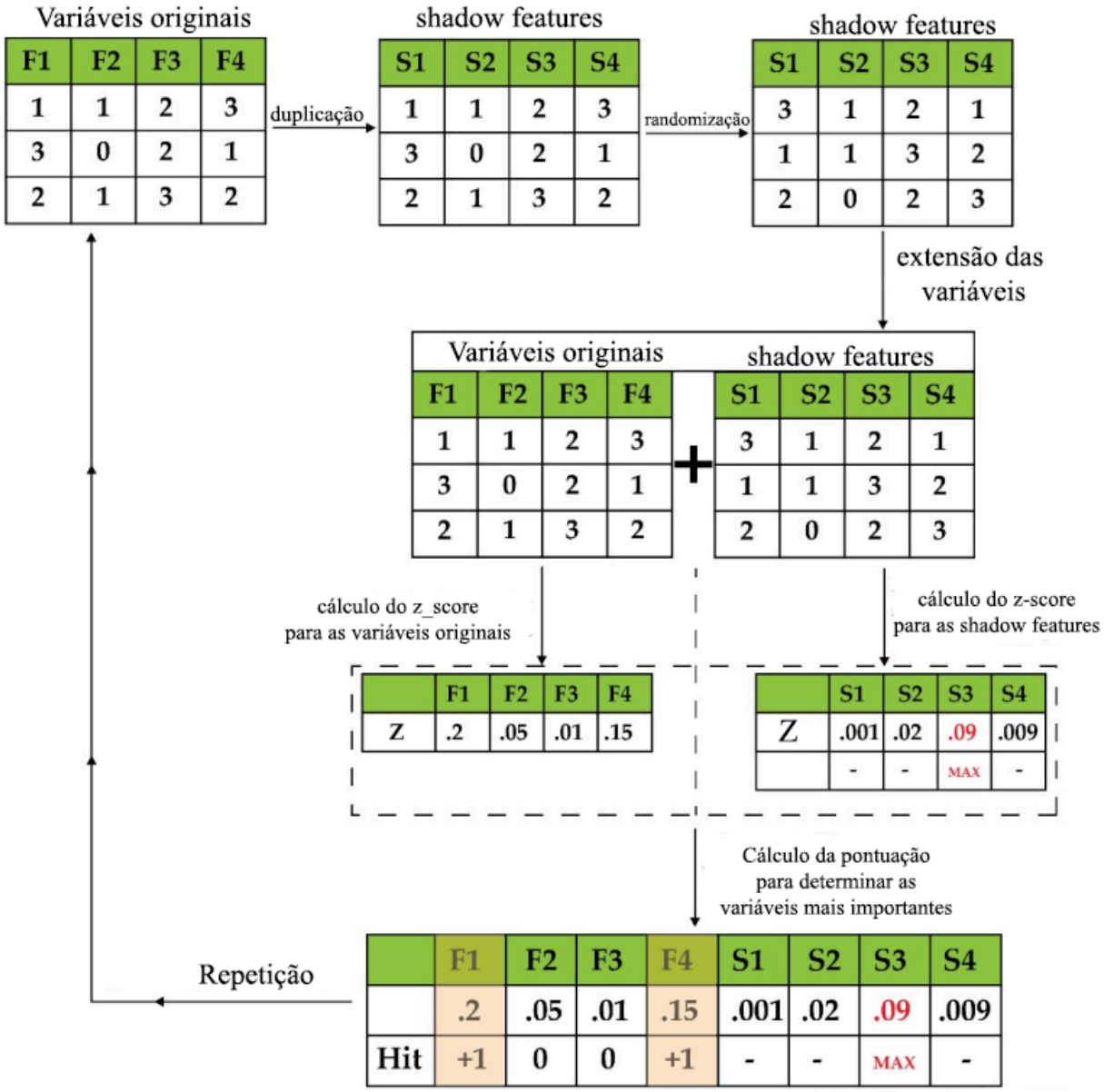
Onde, Z é o chamado *z-score*, x a variável, x_{med}, a média da variável e σ o desvio padrão da variável.

O processo de seleção de variáveis é descrito pelos seguintes passos (Kursa, 2010) e exemplificado na Figura 8 (Hasan, 2020):

1. São criadas as *shadow features* randomicamente, de modo que a relação entre a variável e o resultado seja aleatória. Ou seja, esses dados são uma espécie de ruído, onde cada classificação possui 50% de chance de ser correta ou não para casos de duas classes.
2. É rodado o *Random Forest* múltiplas vezes, refazendo as *shadow features* para cada iteração.
3. É computada a importância de cada variável. Isso é, se a variável original possui um *z-score* maior que o *z-score* máximo das *shadow features*, se sim, a variável original pontua.
4. A variável é considerada importante se a sua importância é maior que a importância máxima das shadow features. Ou seja, quando seu *z-score* é maior que o *z-score* máximo das *shadow features* em várias iterações.
5. É feito um teste estatístico para cada variável, onde é mensurado quantas vezes a importância é maior ou menor que a importância máxima das shadow features. O número médio de acertos para N iterações é $E(N) = 0,5N$ e a variável é considerada importante quando o número de acertos é significativamente maior que E e desimportante quando o número de acertos é significativamente menor. As variáveis que não são classificadas nesse sistema binário (importante/desimportante) são classificadas como indefinidas.
6. O procedimento é feito até N vezes ou até todas as variáveis serem classificadas em importante/desimportante, o que vier primeiro.

7. Por fim, as *shadow features* são removidas do conjunto de dados e são filtradas as variáveis.

Figura 8 - Passos para seleção de variáveis no algoritmo Boruta



Fonte: Referência 20, adaptada pelo autor.

No caso do presente trabalho, foram realizadas 100 iterações (N = 100) e consideradas apenas as variáveis classificadas como importantes, sendo descartadas as variáveis desimportantes e indefinidas.

*3.6.*Validação Interna: *Leave One Out Cross Validation*

Muitos estudos utilizam o *Leave One Out Cross Validation* para avaliar a performance de um algoritmo de classificação de dados, ele se sobressai quando são poucas classes a serem determinadas. Nesse método, os dados são divididos em treino e teste, sendo 1 conjunto de variáveis para teste e as demais para treino. Esse ciclo se repete para as n medidas presentes no conjunto de dados (Wong, 2015). A partir disso, é calculada a acurácia do modelo, dada pela Equação 5:

$$A_C = \frac{n_{Acertos}}{n} \tag{5}$$

No presente trabalho, os dados utilizados para treinar o algoritmo (68% - 17 amostras de cada classe) são avaliados utilizando esse método, enquanto os demais são apenas utilizados para aferir a acurácia externa do modelo.

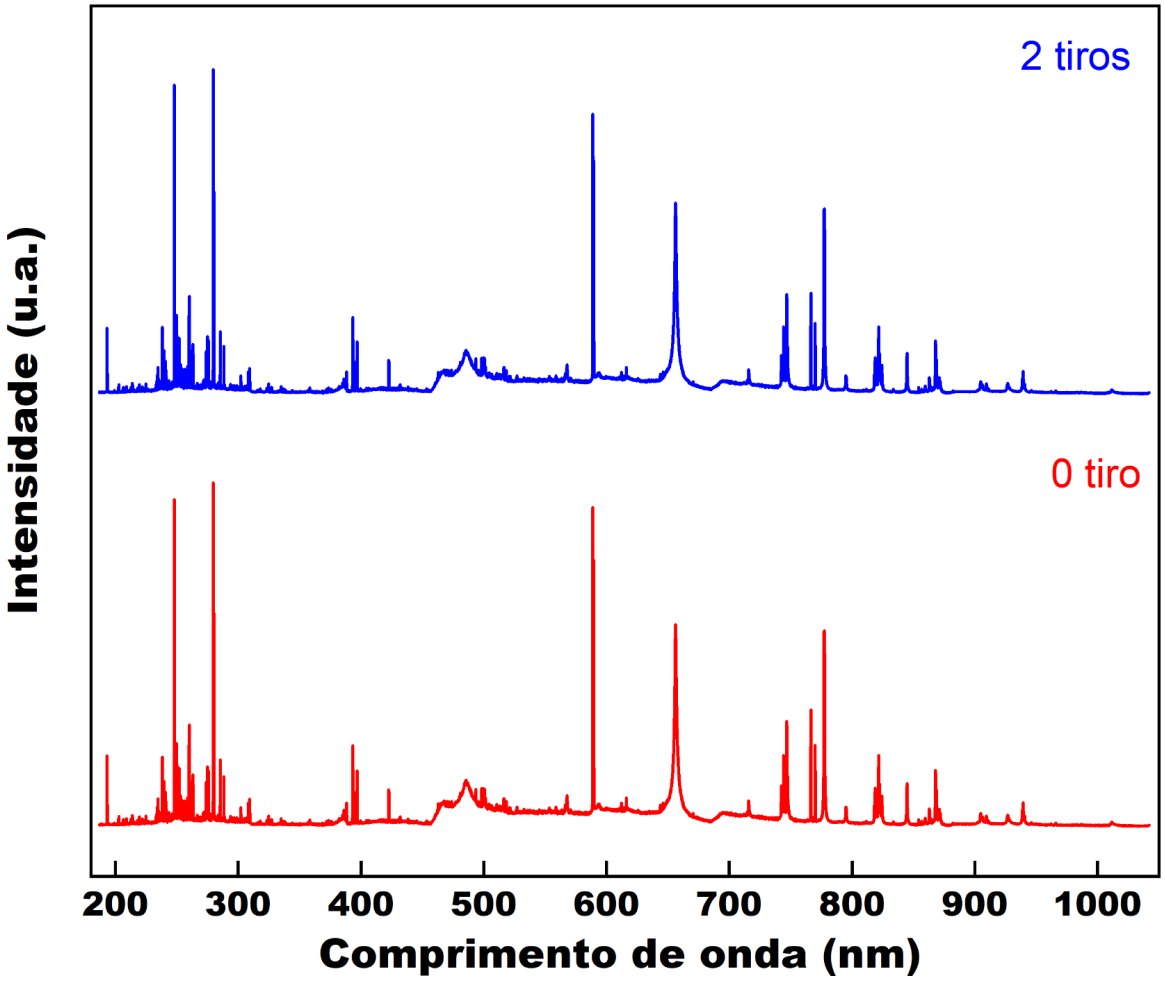
3.7.Validação Externa

A validação externa utiliza os dados inicialmente separados para teste (32% - 8 amostras de cada classe), para obter a acurácia externa do modelo treinado com os dados utilizados na etapa anterior. Vale ressaltar que esses dados, em nenhum momento, são utilizados para treinar o modelo. Um bom modelo apresenta valores de acurácia interna e externa próximos, demonstrando que o modelo treinado é coerente em prever os dados usados para treino e teste, ou seja, possui uma boa capacidade de generalização.

4. RESULTADOS E DISCUSSÕES

Inicialmente foi realizada uma análise exploratória dos espectros brutos, onde foi calculada a média espectral de cada classe (0 tiro, 2 tiros, 25 amostras de cada classe), e realizada a normalização utilizando o SNV, como é mostrado na Figura 9.

Figura 9 - Média espectral de cada classe - 0 tiro (em vermelho) e 2 tiros (em azul)

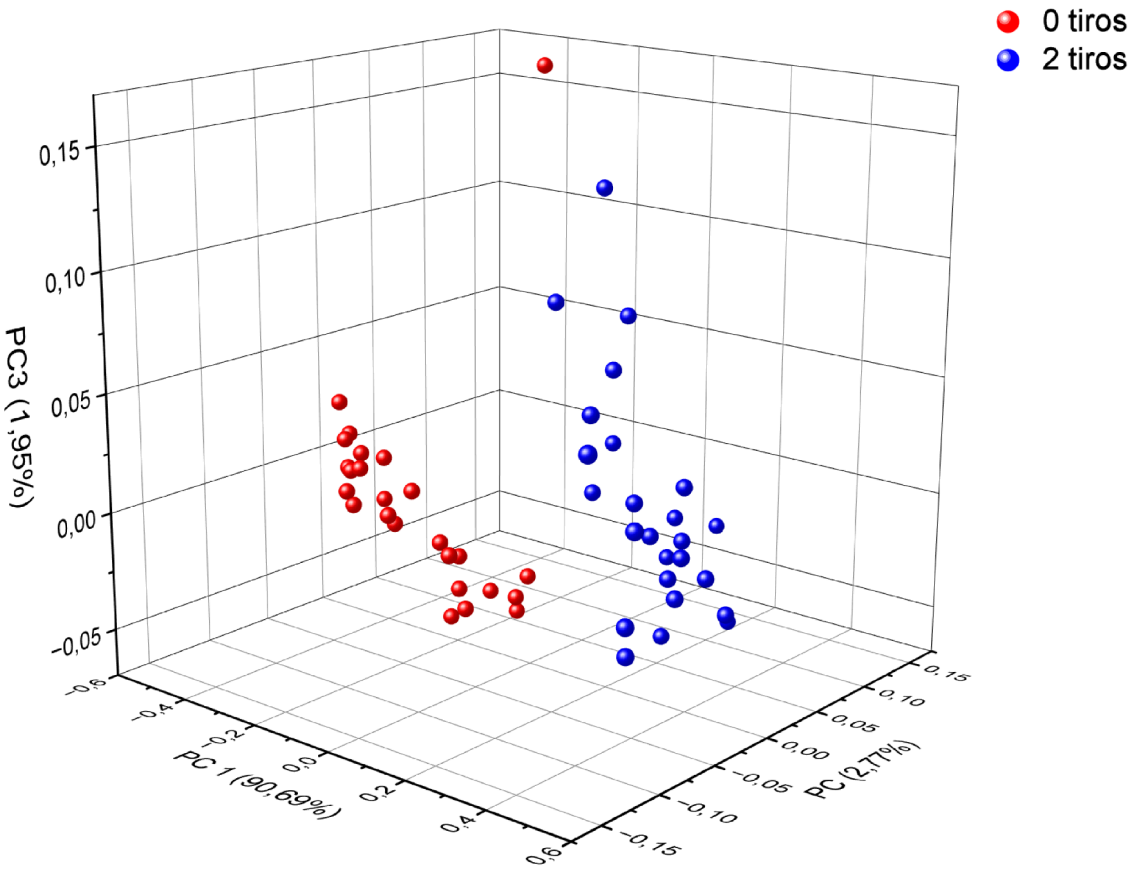


Fonte: Autor.

Na Figura 9 é possível notar a semelhança entre os espectros, o que impossibilita realizar a classificação deles utilizando apenas a visualização do gráfico. Dessa forma, é necessário utilizar técnicas mais robustas para realizar a análise desses dados.

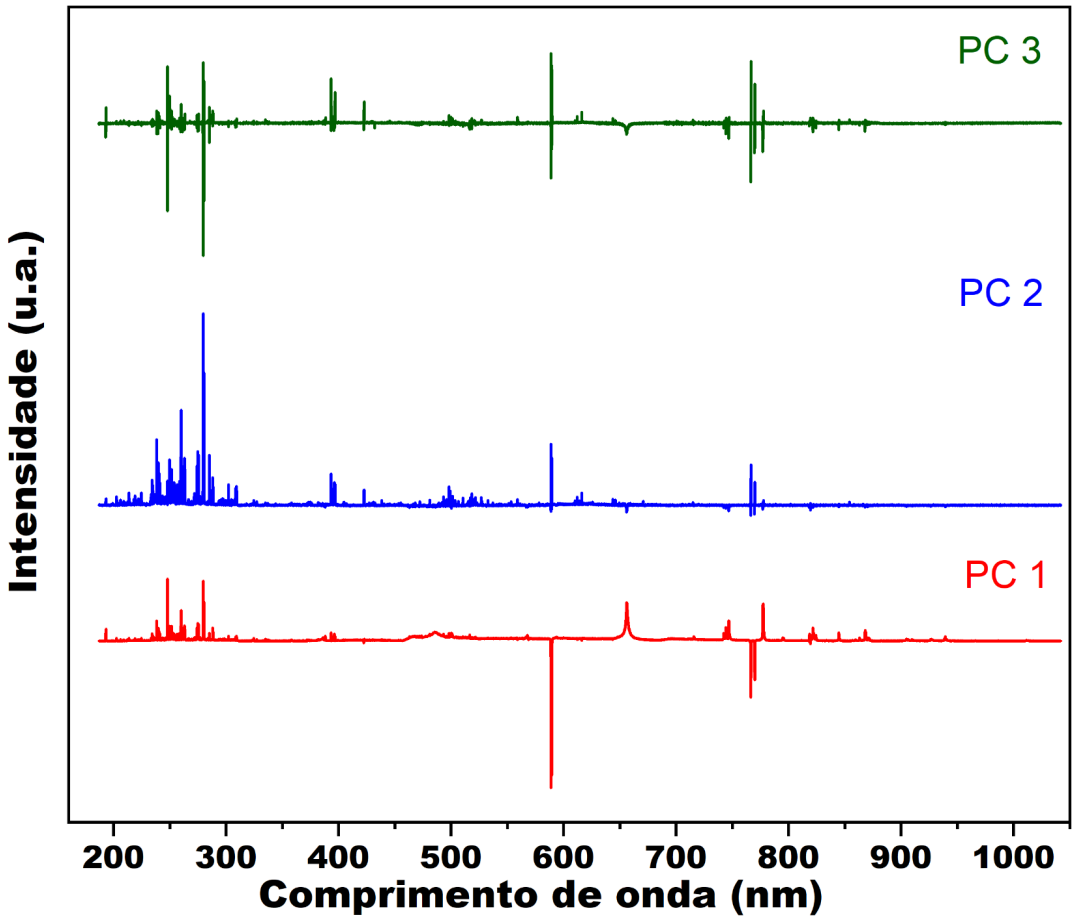
Nesse âmbito, após a normalização via SNV, foi aplicado o PCA com o intuito de realizar a análise não supervisionada e, para melhor visualização dos dados, o *range* espectral foi reduzido a 3 Componentes Principais (PCs). As Figuras 10 e 11 mostram, respectivamente, o *score plot* e os *loadings* das PCs.

Figura 10 - Score plot das 3 primeiras PCs



Fonte: Autor.

Figura 11 - Loadings das três primeiras PCs para o intervalo espectral completo



Fonte: Autor.

Na Figura 11 é mostrada a relação entre as PCs e as variáveis originais. Já na Figura 10 é possível notar uma tendência de separação das amostras, havendo poucas amostras de uma classe invadindo o *cluster* de outra, o que demonstra a potencialidade dos dados para a aplicação do aprendizado de máquina na classificação de futuras amostras.

Após a análise exploratória e aplicação do PCA, foi realizado o treinamento de algoritmos de aprendizado de máquina supervisionados, visando buscar métodos eficazes de classificação das amostras.

Foram utilizadas as bibliotecas *numpy, pandas, sklearn, BorutaPy e matplotlib* da linguagem de programação Python para manejo dos dados, processamento, treinamento de

classificadores, cálculo de métricas, seleção de recursos e visualização de dados, de modo a construir um protocolo de aprendizado de máquina e análise de dados.

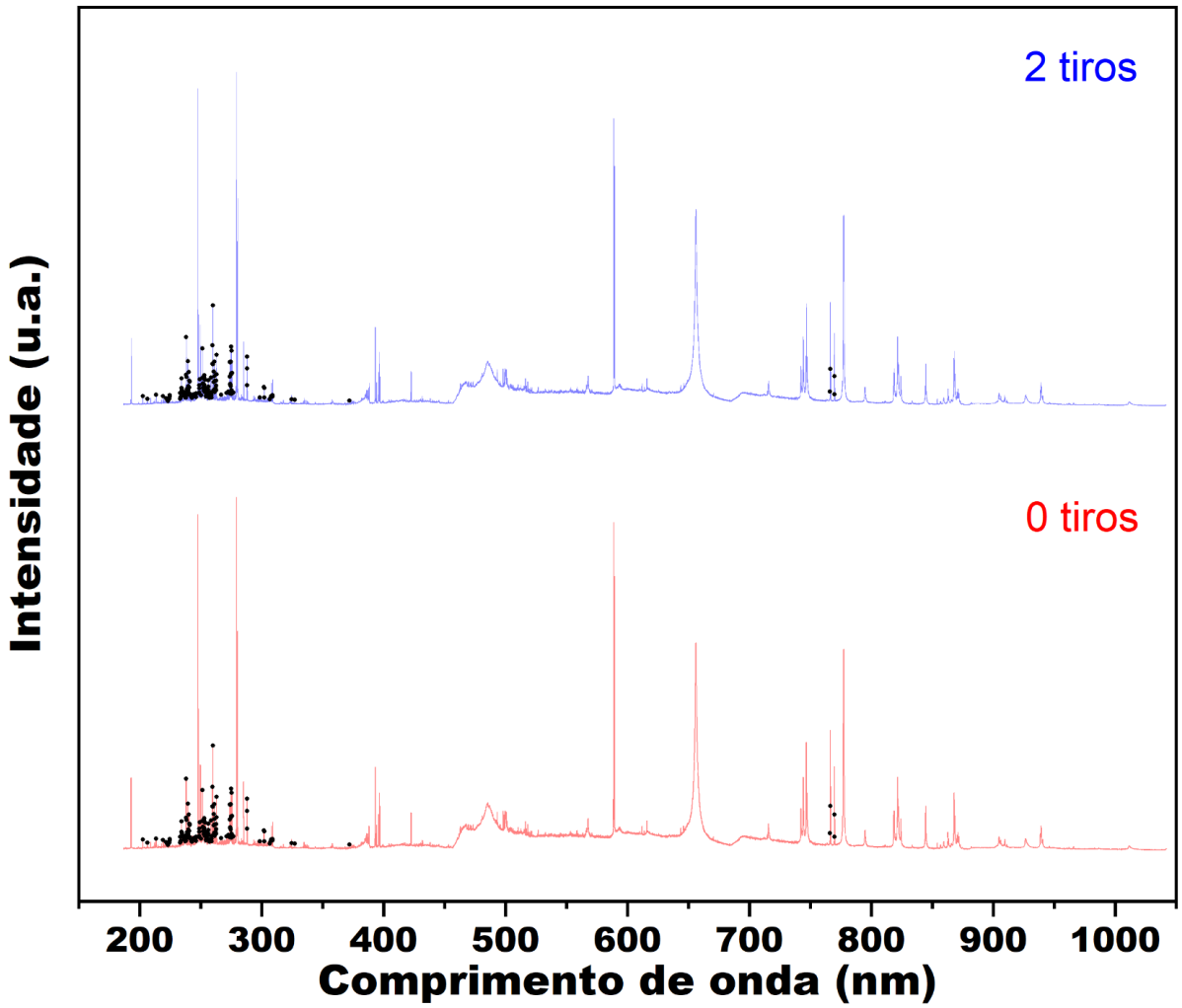
Foram utilizadas para o treinamento do modelo e validação interna, 34 amostras, 17 de cada classe. As demais amostras, 16 sendo 8 de cada classe, foram utilizadas apenas para a validação externa.

Foi realizada a construção de um modelo de classificação *Random Forest*, utilizando 100 árvores de decisão (parâmetro padrão); *random_state* = 42, parâmetro que define como deve ser realizada a randomização na criação das árvores de decisão, é importante manter ele igual dentre as análises para que essa randomização não afete a variação dos resultados; *n_jobs* = -1, parâmetro que indica que todas as CPUs estão sendo utilizadas simultaneamente para realizar o treinamento e classificação do modelo; e demais parâmetros nas configurações padrão que foi aplicado a cinco tratamentos de dados diferentes:

a) Variáveis sem tratamento: Espectros brutos utilizados diretamente na análise, no qual a intensidade de cada comprimento de onda é uma variável.
b) Variáveis selecionadas pelo Boruta (158/12288): Espectros brutos selecionados pelo Boruta, utilizando 100 iterações, foram selecionados 158 comprimentos de onda, onde a intensidade foi julgada, pelo modelo, como relevante para a análise.
c) Variáveis reduzidas a 8 PCs: Espectros brutos reduzidos a 8 Componentes Principais. O número de Componentes Principais utilizado foi o que obteve melhor acurácia na classificação das amostras. Esse processo foi realizado via Grid Search, variando o número de PCs de 1 a 50.
d) Variáveis reduzidas a 8 PCs e posteriormente selecionadas pelo Boruta (3 PCs): Das 8 Componentes Principais obtidas anteriormente, foram selecionadas apenas 3, consideradas mais relevantes pelo método Boruta, novamente se utilizando de 100 iterações.
e) Variáveis selecionadas (158/12288) pelo Boruta e então reduzidas a 8 PCs: Foi aplicado o PCA, reduzindo as variáveis obtidas em b) (158 comprimentos de ondas, e suas respectivas intensidades) para 8 Componentes Principais.

A Figura 11 mostra os 158 comprimentos de onda selecionados pelo Boruta, selecionados na Figura 8. Nota-se que as variáveis selecionadas se encontram majoritariamente na faixa dos 200-350 nm com alguns outros pontos próximos a 770 nm.

Figura 12 - Médias espectrais para cada classe e pontos selecionados pelo Boruta como relevantes (preto)

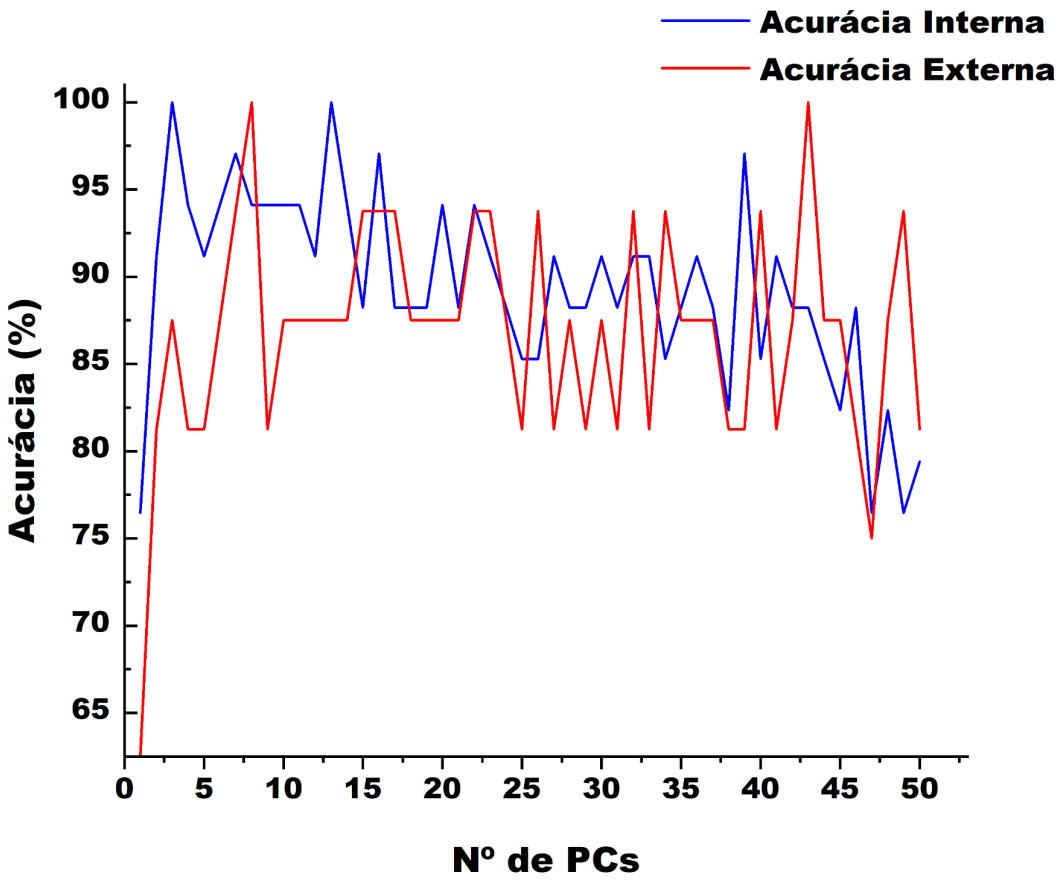


Fonte: Autor.

Esse resultado nos mostra a potencialidade do uso do Boruta para a seleção de variáveis, seu uso foi capaz de reduzir a complexidade dos dados em 98,71% e demonstrou que a faixa que contém as informações relevantes se concentra de 200 a 350 nm, indicando que apenas esse conjunto de dados é necessário para a análise de GSR advinda de munições não tóxicas.

As Figuras 13 e 14 contém o *Grid Search* utilizado para encontrar o número de PCs que produzem maior acurácia no modelo.

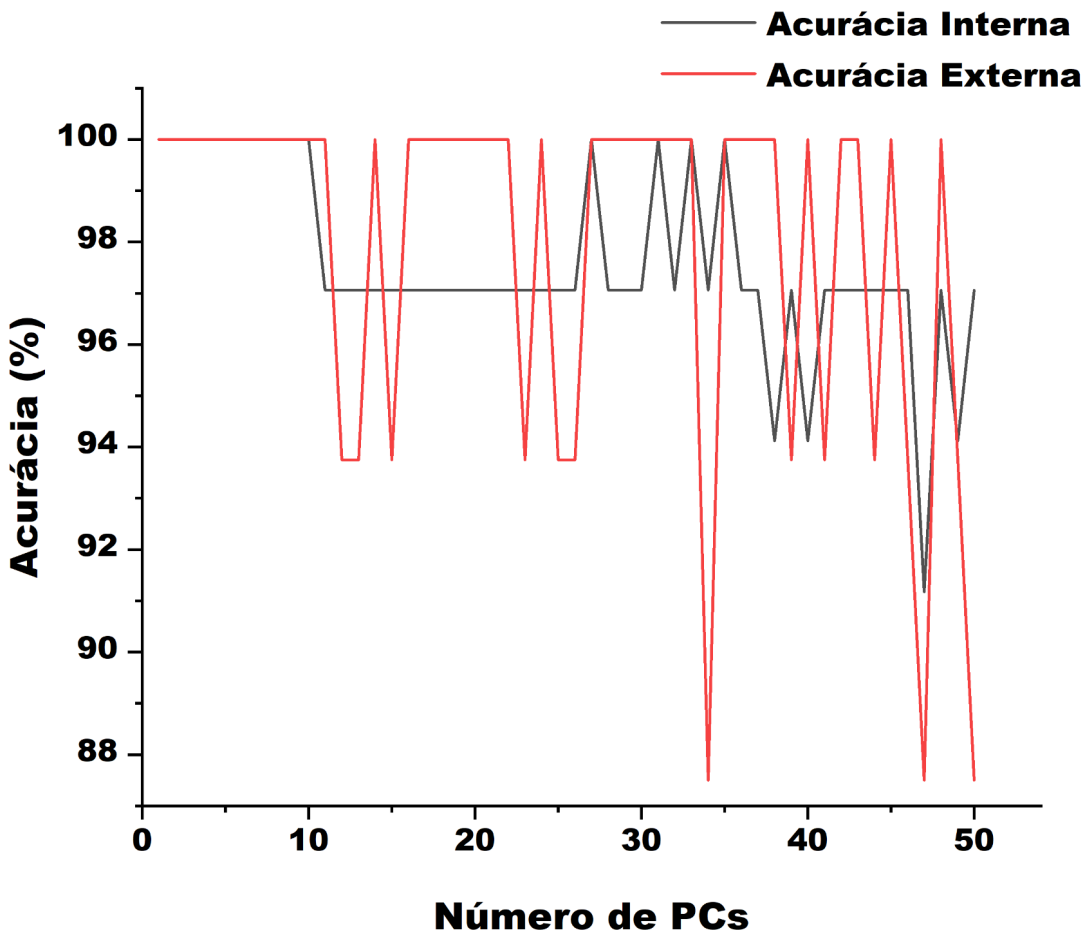
Figura 13 - Grid Search variando o número de PCs de 1 a 50, PCA aplicado nos dados antes do Boruta



Fonte: Autor.

Na Figura 13 é possível notar a variação das acurácias conforme o número de Componentes Principais aumenta. A acurácia externa (em vermelho) atinge seu valor máximo em duas ocasiões, para 8 e 43 PCs, porém a acurácia interna (em azul) é maior para 8 PCs, logo esse número de PCs foi o escolhido para as análises.

Figura 14 - Grid Search variando o número de PCs de 1 a 50, PCA aplicado nos dados selecionados pelo Boruta

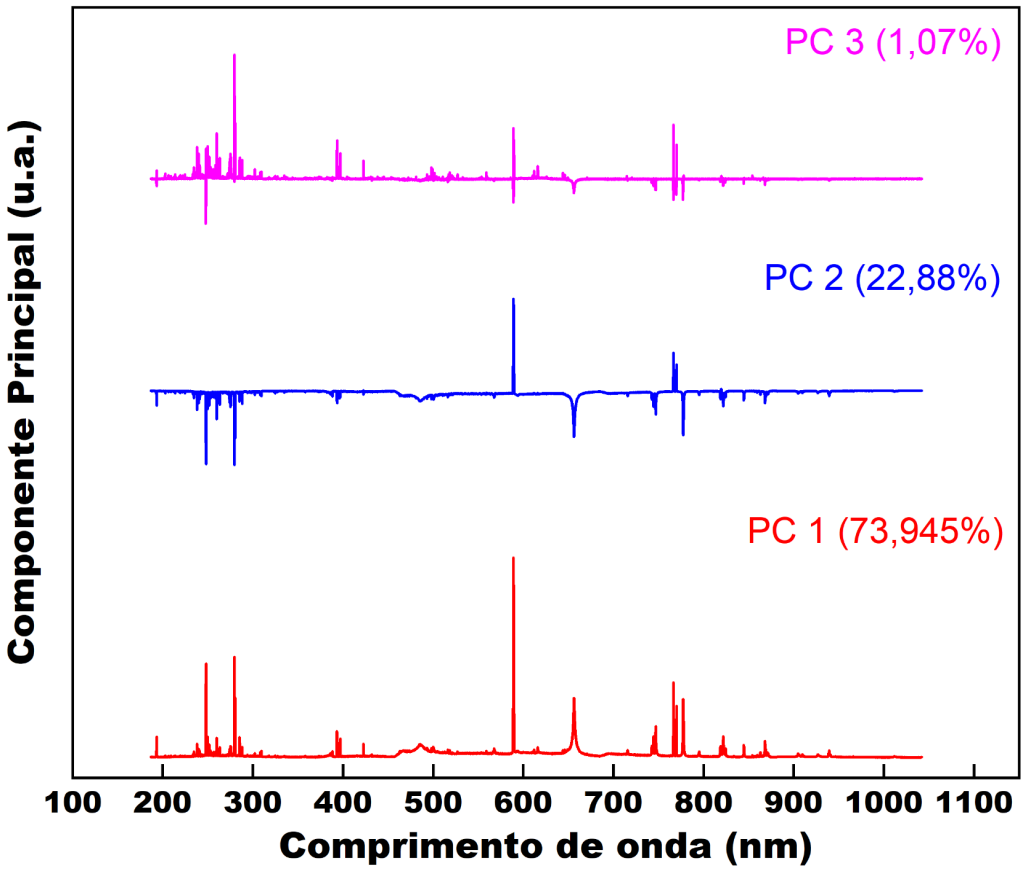


Fonte: Autor.

Na Figura 14 é possível observar que as acurácias são máximas até o número de PCs ser igual a 9, após isso, tanto a acurácia interna (em preto), quanto a externa (em vermelho), oscilam e não estabilizam para um valor específico. Nesse sentido, foram escolhidas 8 PCs para as análises, já que para esse número, ambas as acurácias são máximas e também para fator de comparação com o conjunto de dados c).

A Figura 15 mostra os loadings das 3 PCs selecionadas pelo algoritmo Boruta no tratamento d).

Figura 15 - Loading das 3 PCs selecionadas pelo Boruta



Fonte: Autor.

É possível notar que os espectros das PCs 1, 2 e 3 selecionadas pelo algoritmo Boruta se assemelham com as Componentes Principais obtidas na Figura 10. Nesse contexto, é possível notar a concentração de transições nos intervalos selecionados pelo Boruta (200-350 nm e ~ 770 nm) nas 3 PCs selecionadas. Também é observada uma transição na faixa dos 600 nm, que foi não foi considerada relevante pelo Boruta na seleção de variáveis do tratamento b).

A Tabela 1 apresenta os resultados da acurácia de cada um desses tratamentos.

Tabela 1 - Acurácias (interna e externa) do modelo de classificação Random Forest aplicado a diferentes tipos de tratamentos ((a), b), c), d) e e)) do mesmo conjunto de dados

Tipo de tratamento	**Acurácia Interna**	**Acurácia Externa**
a) Sem tratamento	94,12%	93,75%
b) Apenas Boruta (158 variáveis)	94,12%	93,75%
c) Apenas PCA (8 PCs)	100%	87,50%
d) PCA + Boruta (3/8 PCs)	100%	100%
e) Boruta + PCA (158 variáveis - 8 PCs)	100%	100%

Fonte: Autor.

Comparando as acurácias dos tratamentos a) e b), é possível notar que utilizar apenas o Boruta não foi eficaz para aumentar o desempenho do modelo, mas ainda sim foi eficaz na seleção de variáveis, reduzindo a dimensionalidade dos dados e, por consequência, reduzindo o poderio computacional e tempo necessários para treinar e classificar os dados.

Os tratamentos c), d) e e) alcançaram uma acurácia interna de 100%, indicando que o uso do PCA potencializa a eficiência do modelo em realizar a classificação dos dados utilizados para treino. O tratamento c), que utiliza apenas o PCA para reduzir a dimensionalidade dos dados, teve a sua acurácia externa reduzida se comparado aos tratamentos a) e b), indicando que seu uso reduziu a capacidade de generalização do modelo.

Os tratamentos d) e e), que combinam a seleção de variáveis do Boruta e a redução de dimensionalidade do PCA, obtiveram 100% em ambas as acurácias, indicando que a combinação dos dois métodos é bastante eficaz em selecionar os dados que mais interferem na classificação das amostras e descartando aqueles que não são relevantes, de modo a potencializar a eficiência do *Random Forest*.

5. CONCLUSÃO

A utilização da técnica LIBS, combinada com métodos de aprendizado de máquina, procedimentos de normalização e seleção de variáveis, demonstrou-se extremamente eficaz na classificação e identificação de resíduos de armas de fogo gerados por munições não tóxicas. A implementação da normalização utilizando o método SNV foi eficiente em mitigar ruídos e interferências indesejadas nas medições, resultando em dados mais limpos e precisos. As abordagens de Análise de Componentes Principais (PCA) e Boruta foram eficazes na redução e seleção de amostras. Além disso, foi evidenciada uma vantagem no uso do Boruta, uma vez que através da seleção de variáveis, foi possível identificar as transições que mais contribuíram para a separação, demonstrando ser uma ferramenta poderosa para entender esse processo como um todo. Essas técnicas permitiram a eliminação de variáveis que não apresentavam relevância significativa, o que não apenas simplificou o conjunto de dados, mas também potencializou a eficácia do modelo, propiciando uma classificação mais confiável e robusta.

Os resultados obtidos, tanto nas validações internas quanto externas, alcançando uma acurácia de 100%, ressaltam o potencial dessas técnicas em fornecer resultados consistentes e confiáveis. Isso não apenas evidencia a eficácia do protocolo desenvolvido, mas também demonstra seu potencial para aprimorar a classificação de resíduos de armas de fogo derivados de munições não tóxicas. Ademais, o protocolo pode ser estendido para a classificação de uma variedade de outros tipos de amostras, abrindo novas possibilidades de aplicação nas áreas forense e de análise de materiais.

6. REFERÊNCIAS BIBLIOGRÁFICAS

1. ABDULLAH, A.; MOHAMMED, A. **Scanning Electron Microscopy (SEM): A Review**. [s.l: s.n.].
2. BREIMAN, L. **Machine learning**, v. 45, n. 1, p. 5–32, 2001.
3. BROZEK-MUCHA, Z.; JANKOWICZ, A. Evaluation of the possibility of differentiation between various types of ammunition by means of GSR examination with SEM-EDX method. **Forensic science international**, v. 123, n. 1, p. 39–47, 2001.
4. BURGIO, L. et al. Pigment identification in paintings employing laser induced breakdown spectroscopy and Raman microscopy. **Spectrochimica acta. Part B: Atomic spectroscopy**, v. 56, n. 6, p. 905–913, 2001.
5. CHATTERJEE, S. et al. Application of laser-induced breakdown spectroscopy (LIBS) coupled with PCA for rapid classification of soil samples in geothermal areas. **Analytical and bioanalytical chemistry**, v. 411, n. 13, p. 2855–2866, 2019.
6. CHEN, T. et al. The spectral fusion of laser-induced breakdown spectroscopy (LIBS) and mid-infrared spectroscopy (MIR) coupled with random forest (RF) for the quantitative analysis of soil pH. **Journal of analytical atomic spectrometry**, v. 36, n. 5, p. 1084–1092, 2021.
7. CICERO RIBEIRO, M. et al. Discrimination of maize transgenic and non-transgenic varieties by laser induced spectroscopy (LIBS) and machine learning algorithms. **Microchemical journal, devoted to the application of microtechniques in all branches of science**, v. 203, n. 110898, p. 110898, 2024.
8. CIOCCIA, G. et al. Laser-induced breakdown spectroscopy associated with the design of experiments and machine learning for discrimination of Brachiaria brizantha seed vigor. **Sensors (Basel, Switzerland)**, v. 22, n. 14, p. 5067, 2022.
9. CIOCCIA, G. et al. Probabilistic-based identification of gunshot residues (GSR) using Laser-Induced Breakdown Spectroscopy (LIBS) and Support Vector Machine (SVM) algorithm. **Microchemical journal, devoted to the application of microtechniques in all branches of science**, v. 207, n. 112142, p. 112142, 2024.
10. COSTA, R. A. et al. Gunshot residues (GSR) analysis of clean range ammunition using

SEM/EDX, colorimetric test and ICP-MS: A comparative approach between the analytical techniques. **Microchemical journal, devoted to the application of microtechniques in all branches of science**, v. 129, p. 339–347, 2016.

11. DING, Y. et al. Substrate-assisted laser-induced breakdown spectroscopy combined with variable selection and extreme learning machine for quantitative determination of fenthion in soybean oil. **Photonics**, v. 11, n. 2, p. 129, 2024.
12. DIXON, P. B.; HAHN, D. W. Feasibility of detection and identification of individual bioaerosols using laser-induced breakdown spectroscopy. **Analytical chemistry**, v. 77, n. 2, p. 631–638, 2005.
13. DOCKERY, C. R.; GOODE, S. R. Laser-induced breakdown spectroscopy for the detection of gunshot residues on the hands of a shooter. **Applied optics**, v. 42, n. 30, p. 6153–6158, 2003.
14. DOÑA-FERNÁNDEZ, A. et al. Real-time detection of GSR particles from crime scene: A comparative study of SEM/EDX and portable LIBS system. **Forensic science international**, v. 292, p. 167–175, 2018.
15. FAMBRO, L. A. et al. Laser-induced breakdown spectroscopy for the rapid characterization of lead-free gunshot residues. **Applied spectroscopy**, v. 71, n. 4, p. 699–708, 2017.
16. GEURTS, P.; IRRTHUM, A.; WEHENKEL, L. Supervised learning with decision tree-based methods in computational and systems biology. **Molecular bioSystems**, v. 5, n. 12, p. 1593–1605, 2009.
17. GUEZENOC, J.; GALLET-BUDYNEK, A.; BOUSQUET, B. Critical review and advices on spectral-based normalization methods for LIBS quantitative analysis. **Spectrochimica acta. Part B: Atomic spectroscopy**, v. 160, n. 105688, p. 105688, 2019.
18. GUPTA, I. et al. **PCA-RF: An efficient Parkinson's disease prediction model based on random forest classification**. 2022. Disponível em: <http://arxiv.org/abs/2203.11287>.
19. HAHN, D. W.; OMENETTO, N. **Laser-induced breakdown spectroscopy (LIBS), part I: review of basic diagnostics and plasma-particle interactions: still-challenging issues within the analytical plasma community. Applied spectroscopy**, v. 64, n. 12, p. 335–366, 2010.

20. HASAN, M. J. et al. **Health state classification of a spherical tank using a hybrid bag of features and K-nearest neighbor. Applied sciences** (Basel, Switzerland), v. 10, n. 7, p. 2525, 2020.
21. JIA, J. et al. **The facial expression recognition method of random forest based on improved PCA extracting feature**. 2016 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC). **Anais**...IEEE, 2016.
22. JOLLIFFE, I. T. PRINCIPAL COMPONENT ANALYSIS: A BEGINNER'S GUIDE - I. introduction and application. **Weather**, v. 45, n. 10, p. 375–382, 1990.
23. JOLLIFFE, I. T.; CADIMA, J. Principal component analysis: a review and recent developments. **Philosophical transactions. Series A, Mathematical, physical, and engineering sciences**, v. 374, n. 2065, p. 20150202, 2016.
24. KUMAR, S. S.; SHAIKH, T. **Empirical evaluation of the performance of feature selection approaches on random forest**. 2017 International Conference on Computer and Applications (ICCA). **Anais**...IEEE, 2017.
25. KURSA, M. B.; JANKOWSKI, A.; RUDNICKI, W. R. Boruta – A System for Feature Selection. **Fundamenta informaticae**, v. 101, n. 4, p. 271–285, 2010.
26. LIANG, J. et al. Data fusion of laser induced breakdown spectroscopy (LIBS) and infrared spectroscopy (IR) coupled with random forest (RF) for the classification and discrimination of compound salvia miltiorrhiza. **Chemometrics and intelligent laboratory systems: an international journal sponsored by the Chemometrics Society**, v. 207, n. 104179, p. 104179, 2020.
27. MARANGONI, B. S. et al. Multi-elemental analysis of landfill leachates by single and double pulse laser-induced breakdown spectroscopy. **Microchemical journal, devoted to the application of microtechniques in all branches of science**, v. 165, n. 106125, p. 106125, 2021.
28. NEVES, G. C. **ELABORAÇÃO DE PROTOCOLO PARA IDENTIFICAÇÃO DE RESÍDUOS DE DISPARO POR ARMA DE FOGO UTILIZANDO ESPECTROSCOPIA LIBS E APRENDIZADO DE MÁQUINA**. Campo Grande, MS: UFMS, 2023.
29. NICOLODELLI, G. et al. Differentiation of latex biomembrane with collagen and non-collagen using laser induced breakdown spectroscopy. **Materials today.**

Communications, v. 30, n. 103099, p. 103099, 2022.

30. **Oxford Lead Symposium Lead Ammunition: understanding and minimising the risks to human and environmental health 10th**. UK: [s.n.].
31. PALLESCHI, V. Laser-induced breakdown spectroscopy: principles of the technique and future trends. **ChemTexts**, v. 6, n. 2, 2020.
32. PARMAR, A.; KATARIYA, R.; PATEL, V. A review on random forest: An ensemble classifier. In: **International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018**. Cham: Springer International Publishing, 2019. p. 758–763.
33. PATHAK, A. K. et al. Assessment of LIBS for spectrochemical analysis: A review. **Applied spectroscopy reviews**, v. 47, n. 1, p. 14–40, 2012.
34. PRIYANKA, N. A.; KUMAR, D. Decision tree classifier: a detailed survey. **International journal of information and decision sciences**, v. 12, n. 3, p. 246, 2020.
35. ROMANÒ, S. et al. Characterisation of gunshot residues from non-toxic ammunition and their persistence on the shooter's hands. **International journal of legal medicine**, v. 134, n. 3, p. 1083–1094, 2020.
36. SAVERIO ROMOLO, F.; MARGOT, P. Identification of gunshot residue: a critical review. **Forensic science international**, v. 119, n. 2, p. 195–211, 2001.
37. SENESI, G. S. et al. Geochemical identification and classification of cherts using handheld laser induced breakdown spectroscopy (LIBS) supported by supervised machine learning algorithms. **Applied geochemistry: journal of the International Association of Geochemistry and Cosmochemistry**, v. 151, n. 105625, p. 105625, 2023.
38. SHAHEEN, A.; IQBAL, J. Spatial distribution and mobility assessment of carcinogenic heavy metals in soil profiles using geostatistics and Random Forest, Boruta Algorithm. **Sustainability**, v. 10, n. 3, p. 799, 2018.
39. SHRIVASTAVA, P.; JAIN, V. K.; NAGPAL, S. Gunshot residue detection technologies—a review. **Egyptian journal of forensic sciences**, v. 11, n. 1, 2021.
40. TANG, H. et al. Classification of different types of slag samples by laser-induced breakdown spectroscopy (LIBS) coupled with random forest based on variable importance (VIRF). **Analytical methods: advancing methods and applications**, v. 7, n.

21, p. 9171–9176, 2015.

41. TARIFA, A.; ALMIRALL, J. R. Fast detection and characterization of organic and inorganic gunshot residues on the hands of suspects by CMV-GC-MS and LIBS. **Science & justice: journal of the Forensic Science Society**, v. 55, n. 3, p. 168–175, 2015.
42. UNNIKRISHNAN, V. K. et al. Analytical predictive capabilities of Laser Induced Breakdown Spectroscopy (LIBS) with Principal Component Analysis (PCA) for plastic classification. **RSC advances**, v. 3, n. 48, p. 25872, 2013.
43. WASKLE, S.; PARASHAR, L.; SINGH, U. **Intrusion detection system using PCA with random forest approach**. 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC). **Anais**...IEEE, 2020.
44. WONG, T.-T. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. **Pattern recognition**, v. 48, n. 9, p. 2839–2846, 2015.
45. ZHANG, Y.; ZHANG, T.; LI, H. Application of laser-induced breakdown spectroscopy (LIBS) in environmental monitoring. **Spectrochimica acta. Part B: Atomic spectroscopy**, v. 181, n. 106218, p. 106218, 2021.