UNIVERSIDADE ESTADUAL DA PARAÍBA
CAMPUS I – CAMPINA GRANDE
CENTRO CIÊNCIAS E TECNOLOGIA
DEPARTAMENTO QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

MARIA BARBOSA DA SILVA CORDEIRO

ESPECTROSCOPIA NIR E ALGORÍTMOS DE SELEÇÃO DE VARIÁVEIS PARA DETERMINAÇÃO DE TEOR DE ÁGUA E CLASSIFICAÇÃO DE SEMENTES DE ALGODÃO

Campina grande – PB
Julho de 2023

MARIA BARBOSA DA SILVA CORDEIRO

ESPECTROSCOPIA NIR E ALGORÍTMOS DE SELEÇÃO DE VARIÁVEIS PARA DETERMINAÇÃO DE TEOR DE ÁGUA E CLASSIFICAÇÃO DE SEMENTES DE ALGODÃO

Trabalho de Dissertação apresentado ao Programa de Pós-Graduação em Química da Universidade Estadual da Paraíba, como requisito obrigatório à obtenção do título de mestre em Química.

Área de concentração: Metodologia Analítica Avançada.

Orientador: Profª. Drª. Simone da Silva Simões

Coorientador: Dr. Everaldo Paulo de Medeiros

Campina Grande – PB
Julho de 2023

C794e Cordeiro, Maria Barbosa da Silva.

Espectrocospia NIR e algoritmos de seleção de variáveis para determinação de teor de água e classificação de sementes de algodão [manuscrito] / Maria Barbosa da Silva Cordeiro. - 2023.

72 p. : il. colorido.

Digitado.

Dissertação (Mestrado em Química) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2023.

"Orientação : Profa. Dra. Simone da Silva Simões , Departamento de Química - CCT. "

"Coorientação: Prof. Dr. Everaldo Paulo de Medeiros , Embrapa Algodão"

1. Quimiometria. 2. Análise discriminante. 3. Sementes convencionais. 4. Sementes transgênicas . 5. Algodão. I. Título

21. ed. CDD 633.51

Elaborada por Talita M. A. Tavares - CRB - CRB 15/971 BC/UEPB

MARIA BARBOSA DA SILVA CORDEIRO

ESPECTROSCOPIA NIR E ALGORÍTMOS DE SELEÇÃO DE VARIÁVEIS PARA DETERMINAÇÃO DE TEOR DE ÁGUA E CLASSIFICAÇÃO DE SEMENTES DE ALGODÃO

Trabalho de Dissertação apresentado ao Programa de Pós-Graduação em Química da Universidade Estadual da Paraíba, como requisito obrigatório à obtenção do título de mestre em Química

Aprovada em: 25/07/2023.

BANCA EXAMINADORA

gov.br Documento assinado digitalmente
SIMONE DA SILVA SIMOES
Data: 16/08/2023 10:11:13-0300
Verifique em https://validar.iti.gov.br

Profª. Drª. Simone da Silva Simões (Orientadora)
Universidade Estadual da Paraíba (UEPB)

Epmedeiros (Proc SEI 21156.002048/2023-53 Doc 9190019)

Dr. Everaldo Paulo de Medeiros (Co-Orientador)
Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA)

gov.br Documento assinado digitalmente
PAULO HENRIQUE GONCALVES DIAS DINIZ
Data: 09/08/2023 20:47:34-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Paulo Henrique Gonçalves Dias Diniz (Membro interno)
Universidade Federal do Oeste da Bahia (UFOB)

Welma T. Silva Vilar

Profª. Drª. Welma Thaise Silva Vilar (Membra externa)
Faculdade Rebouças de Campina Grande

Aos meus pais, pelo companheirismo e apoio, DEDICO.

AGRADECIMENTOS

Agradeço a Deus por me conceder força e coragem para enfrentar as adversidades da vida em todos os aspectos, e por me dar discernimento para compreender que tudo tem um proposito.

Aos meus pais Manoel e Sebastiana, irmãos, cunhadas e sobrinhos por serem sempre fortalezas em minha caminhada e por me motivarem sempre em busca dos meus objetivos, com sua compreensão e paciência, palavras não define o quanto sou grata a Deus por ter vocês.

As amizades que a graduação e o mestrado me proporcionaram: Agradeço a todos em nome de Alanna, Joabel, Rodrigo, Daniely e Cicero. Vocês foram fundamentais para que a trajetória fosse mais leve. Obrigada por não soltarem a minha mão, unidos até no choro.

A minha orientadora Profª. Drª Simone da Silva Simões por toda orientação e pelos ensinamentos compartilhados. Obrigada por tudo, pelas conversas e contribuições tanto pessoal quanto profissional.

Ao meu co-orientador Dr. Everaldo Paulo de Medeiros por todas suas contribuições e a todos do LATECQ da Embrapa algodão por todo acolhimento e ensinamentos.

A todos os professores do PPGQ UEPB que tiveram contribuição na construção e finalização desta etapa da minha formação intelectual e profissional.

Aos professores que constituíram as bancas de qualificação e dissertação final, por todas as contribuições, correções, sugestões para que fosse realizado o melhor deste trabalho.

Ao gMAQ por tanto aprendizado partilhado durante nossos encontros cheios de quimiometria, prosa e café.

E agradeço a todos que de forma direta ou indireta contribuíram, trilharam e fizeram parte deste ciclo que se encerra, o meu sincero agradecimento.

RESUMO

A diferenciação entre sementes geneticamente modificadas (OGM´s) e convencionais é essencial a viabilidade agroambiental nos diferentes cultivos existentes, pois pode evitar a contaminação de sementes convencionais ou extinção de espécies nativas, eliminação de plantas e insetos, assim como a resistências desses dentro do próprio ambiente produtivo. Os métodos usados para a identificação de OGMs são baseados em análises de DNA, que apesar de serem precisas e confiáveis, geralmente são caras, demoradas e pouco disponíveis. Por outro lado, as técnicas espectroscópicas, bastante utilizadas na área agrícola, para a determinação de requisitos de qualidade em sementes demonstram melhor desempenho para este fim. O uso de algoritmos para a seleção de variáveis gera um subconjunto contendo as variáveis que melhor se relacionam às propriedades de interesse e contribuem para o desenvolvimento de modelos simples e robustos. Nas últimas décadas estudo utilizando algoritmos bionspirados, tem sido explorado em diversas áreas de estudo e em especial na química. A bioinspiração consiste na busca da compreensão dos mecanismos baseados no comportamento de espécies animais ou vegetais. Diante do exposto, neste trabalho estudou-se o desempenho de alguns algoritmos bioinspirados para seleção de varáveis aplicados a dados NIR para a distinção entre sementes convencionais e transgênicas de algodão. Alguns trabalhos já demonstraram o sucesso de modelos de reconhecimento de padrões (RP) construídos utilizando os espectros completos NIR para a solução deste problema analítico. Objetivou-se nesse trabalho, a aplicação de algoritmos para avaliar a capacidade preditiva de modelos de regressão em mínimos quadrados (PLS) construídos com todo o espectro e com intervalos espectrais (iSPA-PLS e i-PLS) para a quantificação de teor de água em sementes de algodão. Além disso, modelos de análise discriminante construídos com as variáveis espectrais NIR selecionadas por algoritmos de seleção de varáveis foram construídos e validados. Para isso foram testados o algoritmo determinístico das projeções sucessivas (SPA) e também, algoritmos bioinspirados como o de colônia de formigas (AOC) e o genético (GA). O conjunto de amostras para a quantificação do teor de umidade foi composto por 30 genótipos de sementes. Já o conjunto de dados destinado aos métodos de RP foi composto por 1195 sementes transgênicas e 1195 sementes convencionais. Todas as amostras foram cedidas pela EMBRAPA algodão. A capacidade preditiva dos modelos de quantificação foi avaliada com base nos valores de figuras de mérito como REP e RMSE. Os modelos de RP consideraram a sensibilidade, seletividade e taxa de classificação correta. Os modelos construídos com as variáveis selecionadas obtiveram resultados comparáveis aos modelos com espectro completo tanto para os modelos de quantificação quanto para os modelos de classificação. No entanto, para a quantificação do teor de água em sementes o iPLS se mostrou um modelo mais robusto apresentando RMSEP = 0,20% e REP = 2,17%. Em relação aos modelos LDA construídos com as variáveis selecionadas, todos alcançaram taxas de classificação corretas superiores a 95%. O algoritmo bioinspirado AOC-LDA forneceu a melhor performance com 97% de classificação correta. Diante do exposto, conclui-se que o uso de algoritmos de seleção de variáveis aliados a métodos multivariados de quantificação ou classificação, quando aplicados a dados NIR, são eficientes. Sendo capazes de quantificar teor de água ou classificar sementes de algodão convencionais e transgênicas de forma rápida e não destrutiva.

Palavras-Chave: Quimiometria; Análise discriminante; Regressão em Mínimos Quadrados Parciais; Seleção de variáveis; Sementes de Algodão.

ABSTRACT

The differentiation between genetically modified seeds (GMOs) and conventional ones is essential for agro-environmental viability in the different existing crops, as it can avoid contamination of conventional seeds or extinction of native species, elimination of plants and insects, as well as their resistance within the production environment itself. The methods used for the identification of GMOs are based on DNA analysis, which despite being accurate and reliable, are generally expensive, time-consuming and not readily available. On the other hand, spectroscopic techniques, widely used in the agricultural area, for determining quality requirements in seeds, demonstrate better performance for this purpose. The use of algorithms for the selection of variables generates a subset containing the variables that best relate to the properties of interest and contribute to the development of simple and robust models. In recent decades, studies using bioinspired algorithms have been explored in several areas of study, especially in chemistry. Bioinspiration consists of seeking to understand the mechanisms based on the behavior of animal or plant species. Given the above, this work studied the performance of some bioinspired algorithms for selection of variables applied to NIR data for the distinction between conventional and transgenic cotton seeds. Some works have already demonstrated the success of pattern recognition (RP) models built using full NIR spectra to solve this analytical problem. The objective of this work was the application of algorithms to evaluate the predictive capacity of least squares regression models (PLS) built with the whole spectrum and with spectral intervals (iSPA-PLS and i-PLS) for the quantification of water content in cotton seeds. Furthermore, discriminant analysis models constructed with NIR spectral variables selected by variable selection algorithms were constructed and validated. For this, the deterministic algorithm of successive projections (SPA) and also bioinspired algorithms such as ant colony (AOC) and genetic (GA) were tested. The set of samples for the quantification of moisture content was composed of 30 seed genotypes. The data set for the PR methods was composed of 1195 transgenic seeds and 1195 conventional seeds. All samples were provided by EMBRAPA cotton. The predictive capacity of the quantification models was evaluated based on the values of figures of merit such as REP and RMSE. PR models considered sensitivity, selectivity and correct classification rate. The models built with the selected variables obtained comparable results to the models with full spectrum both for the quantification models and for the classification models. However, for the quantification of water content in seeds, iPLS proved to be a more robust model, presenting RMSEP = 0.20% and REP = 2.17%. Regarding the LDA models built with the selected variables, all achieved correct classification rates greater than 95%. The AOC-LDA bioinspired algorithm provided the best performance with 97% correct classification. Given the above, it is concluded that the use of variable selection algorithms combined with multivariate methods of quantification or classification, when applied to NIR data, are efficient. Being able to quantify water content or classify conventional and transgenic cotton seeds quickly and non-destructively.

Keywords: Chemometrics; Discriminant analysis; Partial Least Squares Regression; Selection of variables; Cotton Seeds.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABNT	Associação Brasileira de Normas Técnicas
ACO	*Ant Colony Optimization*
ACC	*Acurácia*
BCI	Better Cotton Iniciative
CP	Componentes principais
CTNBio	Comissão Técnica Nacional de Biossegurança
EMBRAPA	Empresa Brasileira de Pesquisa Agropecuária
FAR	Infravermelho distante
GA	*Genetic Algorithm*
GARGS	*Genetic Algorithm – Based Region Selection*
GLM	*Generalized Linear Models*
IUPAC	*International union of purê and applied chemistry*
KS	*Kennard e Stone*
LATECQ	Laboratório Avançado de Tecnologia Química
LASSO	*Least Absolute Selection and Shrinkage Operator*
LDA	Analise discriminante linear
LDL	*Low Density Lipoprotein*
MC-UVE	Monte Carlo – *Uninformative Variable Elimination*
MCS-RPLS	Monte Carlo – *Partial least squares regression*
MSC	*Multiplicative Scatter Correction*
NIR	*Near infrared spectroscopy*
PCA	*Principal Component Analysis*
RP	Reconhecimento de padrão
SNV	*Standard Normal Variate*
SPA	*Successive Projections Algorithm*
SPXY	*Sample set portioning based on joint x-y distance*
SVM	*Suport Vector Machine*

SUMÁRIO

1 INTRODUÇÃO

Os métodos espectroscópicos modernos, como a espectroscopia NIR, são caracterizados por envolver durante uma medida um elevado número de variáveis, que podem ser ruidosas e altamente correlacionadas com redundância de informação. Sendo, possível perceber a importância do uso de técnicas que permitam selecionar as variáveis mais informativas em um conjunto de dados (OZAKI et al. 2021). Este processo é importante para garantir a construção de modelos simples e robustos de classificação e/ou calibração. Além disso, as técnicas de seleção de variáveis contribuem para minimizar os riscos de inferências e melhorar o custo computacional (ANZANELLO et al., 2013). As técnicas de seleção de variáveis em espaços de alta dimensão, como é o caso dos dados espectroscópicos, têm atraído considerável atenção na pesquisa de exploração de dados ao longo dos últimos anos (PASQUINI, 2018; YUN et al., 2019). Ainda mais no tocante aos conjuntos de dados compostos por muitas amostras, como é o caso do desenvolvimento de modelos na área de melhoramento vegetal, em razão de sua natureza abranger a variabilidade biológica e espaço temporal, e também conservação da matriz de trabalho em análises não destrutivas, bem como frequência analítica para aumento de escala.

Os métodos determinísticos de seleção de variáveis apresentam como solução um único subconjunto de variáveis, ou seja, sempre leva à mesma resposta, após diferentes execuções, desde que sejam aplicados sempre às mesmas condições iniciais (GOMES,2012). Já para os métodos estocásticos, várias escolhas são realizadas com base em números aleatórios, selecionados no momento de execução do código. Partindo de um mesmo ponto inicial, cada execução do código seguirá o seu próprio caminho, e possivelmente levará a resultados finais que convergem para as mesmas soluções ótimas com dada precisão. Nas últimas décadas estudo utilizando algoritmos bionspirados para seleção de variáveis, tem sido explorado em diversas áreas de estudo. A bioinspiração consiste na busca da compreensão dos mecanismos baseados no comportamento de espécies animais ou vegetais (PONTES,2020).

Diante do exposto, este trabalho visa avaliar a performance de alguns algoritmos de seleção de variáveis determinísticos e bioinspirados aplicados a dados NIR de sementes de algodão. Neste sentido, foram construídos modelos de quantificação multivariada para predizer teor de água em sementes, bem como, de reconhecimento de padrões para distinguir sementes transgênicas e convencionais. Os modelos construídos utilizando as variáveis selecionadas foram comparados com os modelos de espectro completo.

1.1 OBJETIVOS

Objetivo geral

Explorar a capacidade preditiva de alguns algoritmos de seleção de variáveis para a construção de modelos na predição de teor de umidade em semente de algodão, e de reconhecimento de padrões para a classificação de sementes de algodão transgênicas e convencionais.

Objetivos específicos

- ✓ Desenvolver modelos PLS, iPLS e iSPA-PLS.
- ✓ Comparar o desempenho de modelos de quantificação multivariada construídos a partir dos intervalos espectrais selecionados pelos algoritmos iPLS e iSPA-PLS.
- ✓ Validar os modelos construídos com base nas figuras de mérito REP, RMSE, r, R^2.
- ✓ Comparar o desempenho de modelos de reconhecimento de padrões construídos a partir das variáveis selecionadas por algoritmos estocásticos bio-inspirados, como o ACO e GA, e determinístico, como o SPA-LDA.
- ✓ Validar os modelos construídos com base nas figuras de mérito de sensibilidade, especificidade, acurácia e taxa de classificação correta.
- ✓ Comparar o desempenho dos modelos construídos com todo o espectro e com as variáveis selecionadas em função da taxa de classificação correta, sensibilidade e especificidade obtida para um conjunto externo (teste) de amostras.

2. REVISÃO DE LITERATURA

2.1. Químiometria e procedimentos quimiométricos

A quimiometria, segundo a União Internacional de Química Pura e Aplicada (IUPAC, do inglês *International Union of Pure and Applied Chemistry*), é a ciência que vincula medidas obtidas em sistema ou processo químico por meio da aplicação de métodos estatísticos e matemáticos (HIBBERT, 2016). De uma maneira mais objetiva, a quimiometria utiliza métodos matemáticos e estatísticos para extrair, tratar, interpretar e prever dados químicos complexos de natureza multivariada, envolvendo o máximo de informações relevantes (FERREIRA, 2015). A análise multivariada permite tratar conjuntos de dados que contenham qualquer número de variáveis e extrair informações relevantes na forma de modelos empíricos, que podem ser utilizados para medições quantitativas e qualitativas (KEMSLEY et al., 2019).

A quimiometria é uma ciência de dados químicos multidisciplinar na concepção de ser possível atuar em diversas áreas de conhecimento sobretudo nas ciências analíticas e química, mas também amplamente utilizada nas áreas de ciências dos alimentos, ciências farmacêuticas, engenharia química, agronomia e biologia, dentre outras (VERAS et al 2022).

Uma das características mais relevantes dos métodos quimiométricos é a possibilidade de quantificação ou classificação sem a necessidade de resolução do sinal analítico. Esses métodos demandam um mínimo ou nenhum tipo de pré-tratamento das amostras, sendo ambientalmente desejável por não gerarem resíduos nem consumirem reagentes ou solventes (SENA et. al., 2018). Os métodos quimiométricos podem ser utilizados em dados de natureza multivariados com a finalidade de: processamento de sinais analíticos, planejamento e otimização de experimentos, reconhecimento de padrões e classificação de dados, calibração multivariada, monitoramento e modelagem de processos, métodos de inteligência artificial, entre outros (FERREIRA, 2015). Desse modo, os métodos quimiométricos associados às técnicas espectroscópicas são ferramentas importantes para estratégias de reconhecimento de padrões ou para prever uma determinada propriedade em diferentes amostras (LUNA; GOIS, 2018; ESTEBAN; ARIÑOBLASCO; DÍAZ-CRUZ, 2020).

Os métodos quimiométricos podem ser aplicados para fins quantificação ou classificação. No entanto, antes da construção dos modelos algumas etapas importantes como o pré-processamento dos dados e a seleção dos conjuntos de calibração (ou treinamento) e predição (ou teste) são necessárias.

2.1.1. Pré-processamento dos dados

O pré-processamento dos dados é a etapa inicial do processo de desenvolvimento dos modelos multivariados. Tendo como objetivo geral eliminar ou reduzir a variância aleatória, além de fontes de variação sistemáticas não desejadas. A extração das informações poderá focar-se na variância com maior significância aos objetivos da análise. Esses efeitos podem ser minimizados ou corrigidos empregando métodos de pré-processamento espectral e são atribuídos: mudanças no caminho ótico, sensibilidade do detector, do amplificador e diferenças no tamanho das partículas ou propriedades físicas (morfologia e porosidade) da amostra (FERREIRA, 2015; PASQUINI, 2018).

O passo inicial no pré-tratamento de dados é a sua organização na forma matricial, **X** (I×J). Em que cada uma das linhas *i* refere-se ao conjunto de informações de uma amostra e cada elemento das j colunas refere-se ao registro da intensidade de absorção em dado número de onda. Embora as transformações e pré-processamentos sejam aplicados para a exclusão de variáveis indesejáveis, os mesmos devem ser aplicados com parcimônia afim de evitar distorções nos dados espectrais, e comprometimento dos modelos quimiométricos desenvolvidos (FERREIRA, 2015). Deste modo, é de suma importância que os algoritmos utilizados para pré-processamento espectral podem melhorar a capacidade preditiva ou discriminatória do modelo ao eliminar fontes de variabilidade espectral que mascaram informações de grande relevância para o estudo. No entanto, a sua aplicação deve ser realizada com atenção, pois o excesso ou má aplicação do pré-processamento pode causar perda de informações importantes (PASQUINI, 2018).

Os métodos de pré-processamento podem ser aplicados tanto nas linhas (amostras) quanto nas colunas (variáveis) da matriz de dados (FERREIRA, 2015; SENA et al., 2017). O pré-processamento nas linhas é aplicado em uma amostra de cada vez, correlacionando todas as variáveis. Do mesmo modo, o pré-processamento nas colunas é aplicado a cada variável, correlacionando todas as amostras. As fontes que comumente causam a variância indesejável em espectros NIR são ruídos, espalhamentos e desvios de linha-base (ENGEL et al., 2013).

Os algoritmos de pré-processamento podem ser usados individualmente e em combinação, desde que sejam projetados para superar fontes de variabilidade espectral (FERREIRA, 2015). Como exemplo, a suavização de Savitzky-Golay é aplicada para reduzir o ruído instrumental; a correção de espalhamento multiplicativa - MSC (Multiplicative Scatter Correction) e a Variação Normal Padrão – SNV (Standard Normal Variate) são usadas para

reduzir os efeitos aditivos e multiplicativos causados pelo espalhamento, enquanto a utilização de derivadas corrige o deslocamento e tendência da linha de base (NOLASCO, 2019).

Adicionalmente, as técnicas de suavização, atuam na redução do ruído aleatório existentes nos dados, as quais favorecem o aumento da razão sinal-ruído, com melhor resolução do sinal analítico em relação ao ruído sistemático e aleatório. Essas técnicas atuam como filtros lineares, que incluem a definição de uma janela de pontos, que se desloca por todo o espectro reduzindo todos os pontos da janela, a um ponto central (OLIVEIRA, 2018).

2.1.2 Seleção dos conjuntos de calibração e predição

Após o pré-processamento dos dados instrumentais, é realizado fracionamento do conjunto amostral, em que são definidos os conjuntos de calibração (ou treinamento), validação e predição (ou teste). Essa etapa é indispensável para a construção de modelos sejam eles de classificação ou de calibração, com melhores incrementos por meio de algoritmos, como Kennard-Stone (KS) (KENNARD e STONE, 1969) e SPXY (GOMES et al.2013).

O algoritmo KS (KENNARD e STONE, 1969) utiliza de distância Euclidiana $d_x(p,q)$ para selecionar as amostras mais distantes entre si. Em seguida, uma terceira amostra é rastreada, também distante deste primeiro subconjunto. O fim do ciclo é dado quando o número de amostras selecionadas atinge o número escolhido pelo analista.

De modo análogo, o algoritmo de partição baseado em distâncias SPXY considera a informação simultânea das variáveis independente e dependente (**x**-y). Assim que ocorre a normalização de cada termo, a distância xy atribui importâncias equivalentes na distribuição das amostras em ambos os espaços considerados. O SPXY é geralmente adotado em métodos de regressão utilizados no desenvolvimento de modelos de calibração (GOMES et al.2013).

Após a realização das etapas de pré-processamento e seleção das amostras procede-se a construção dos modelos de calibração multivariada ou reconhecimento de padrões.

2.1.3 Calibração multivariada

A calibração multivariada busca estabelecer um modelo matemático que relaciona um conjunto de medidas instrumentais realizadas em amostras (variáveis independentes, **X**), com determinadas propriedades de interesse, como a concentração de um composto (variáveis dependentes, Y) (SAEYS et al., 2019). O método multivariado associa várias respostas

instrumentais (variáveis), a uma ou mais propriedades de interesse para gerar um modelo. Os parâmetros desse modelo são estimados a partir de um conjunto de amostras de calibração ou treinamento (cerca de 2/3 do universo amostral) para os quais os espectros foram adquiridos e os parâmetros de qualidade de interesse foram quantificados por métodos de referência (DUARTE, 2015).

A calibração multivariada apresenta como vantagens, a realização de determinações diretas na presença de interferentes, desde que presentes no conjunto de calibração e a possibilidade de quantificar simultaneamente diferentes analitos, a partir do mesmo conjunto de espectros, além de apresentar uma redução do erro estimado no modelo devido a utilização de múltiplas variáveis (TIBOLA et al., 2018).

Avanços com as técnicas de calibração multivariada permitiram a determinação de compostos totais como proteína, carboidratos, óleo, ácidos graxos totais e livres. Deste modo, tem sido utilizada progressivamente para determinação de constituintes de alimentos, como a detecção de óleo em milho (MITTELMANN et al., 2006), determinação de organismos geneticamente modificados em alimentos contendo soja (CONCEIÇÃO; MOREIRA; BINSFELD, 2006), composição química de grãos de milho submetidos à secagem e armazenados (GUTKOSKI et al., 2009).

Dos métodos de calibração multivariados, destaca-se a calibração por Mínimos Quadrados Parciais (*Partial Least Squares* – PLS).

2.1.3.1 Calibração por Mínimos Quadrados Parciais (Partial Least Squares – PLS)

A técnica de regressão por Mínimos Quadrados Parciais *(do inglês: Partial Least Squares* – PLS*)* PLS é uma das mais empregadas pelos métodos analíticos baseados em espectroscopia NIR, foi introduzida em 1984, pelo estatístico sueco Herman Wold (WILLIAMS, 2015). O PLS permite identificar fatores (combinações lineares das variáveis **X**) que melhor modelam as variáveis dependentes Y. Este é um modelo que determina correlações quantitativas, servindo por isso para a construção de um modelo de calibração multivariada como uma função linear entre as variáveis dependentes e independentes, em determinadas faixas de resposta da variável dependente (NUNES, 2008).

A regressão PLS tenta maximizar a covariância, capturando assim a variância e correlacionando os dados juntos (FERREIRA, 2015). A qualidade dos modelos obtidos, é avaliada por meio de uma etapa de validação. Inicialmente, o modelo é verificado usando uma validação cruzada, em que as amostras do conjunto de calibração são retiradas, uma de cada

vez, e usadas como amostra de validação. Em seguida, o modelo pode ser revalidado utilizando amostras que não participaram do conjunto de calibração, para uma validação externa, e por meio de ferramentas estatísticas de diagnostico. Assim, o modelo validado pode ser empregado para estimar as propriedades requeridas em amostras desconhecidas, a partir dos seus espectros (DUARTE, 2015). Muitas vezes, dependendo da matriz de dados, os resultados fornecidos pela regressão PLS podem ser otimizados quando aliados a técnicas de seleção de variáveis ou intervalos, como por exemplo o iPLS (do inglês: *Interval Partial Least Squares*).

O iPLS (NORGAARD, et. al, 2000) seleciona um subconjunto de variáveis que fornecerá uma predição superior em comparação ao uso de todas as variáveis em um conjunto de dados. O iPLS realiza uma busca sequencial e exaustiva pela melhor variável ou combinação de variáveis. O *intervalo* para a regressão iPLS pode ser tanto uma única variável quanto uma *janela* de variáveis adjacentes. O número de intervalos a serem testados é escolhido pelo usuário. Inicialmente, os modelos PLS individuais são usados, cada um usando apenas um dos intervalos de variáveis pré-definidos. Se houverem 100 intervalos definidos para um determinado conjunto de dados, a primeira etapa calcula 100 modelos (um para cada intervalo). A validação cruzada é realizada para cada um desses modelos e o intervalo que proporciona o menor erro quadrático médio de validação cruzada (RMSECV) é selecionado. Este é o melhor modelo de único intervalo e o primeiro intervalo selecionado, i1. Se for desejado apenas um intervalo, o algoritmo pode parar nessa etapa. No entanto, se for desejado mais de um intervalo, ciclos adicionais podem ser realizados. No segundo ciclo, o primeiro intervalo selecionado é usado em todos os modelos, mas é combinado com cada um dos outros intervalos restantes, um de cada vez. Isso é repetido para quantos intervalos forem solicitados.

2.1.4 Reconhecimento de padrões

As técnicas de reconhecimento de padrões (RP) se fundamentam na análise de similaridades e dissimilaridades no comportamento espectral das amostras e na tendência de aproximação e separação entre elas no espaço amostral (FERREIRA, 2015; ALVES,2022). Assim as técnicas de reconhecimento de padrões são usadas para identificar as semelhanças e diferenças ente as amostras que foram submetidas a algum tipo de análise, seja por técnicas instrumentais ou pela determinação de variáveis físico-químicos (PONTES, 2020).

As técnicas de reconhecimento de padrões são divididas em técnicas supervisionadas e não supervisionadas. As técnicas de reconhecimento de padrões não supervisionadas buscam encontrar em um conjunto de dados, tendências e/ou agrupamentos sem a necessidade do

conhecimento prévio das classes. Pode-se destacar como principais métodos de RP não supervisionado a análise de componentes principais (PCA, do inglês *Principal Component Analysis*) e análise de agrupamentos hierárquicos (HCA, do inglês *Hierarchical Cluster Analysis*) (HANSEN, 2019).

Em contrapartida, as técnicas de reconhecimento de padrões supervisionadas utilizam uma informação adicional sobre membros das classes. Dessa maneira, a resposta instrumental é relacionada com os índices de classes e modelos multivariados são obtidos e validados com intuito de classificar amostras com identidades desconhecidas. Diversos métodos de reconhecimento de padrões supervisionado são citados na literatura, sendo utilizados em problemas analíticos de classificação, como por exemplo, a Modelagem Independente e Flexível Por Analogia de Classes (SIMCA, do inglês *Soft Independent Modeling of Class Analogy*), a Análise Discriminante Linear (LDA, do inglês *Linear Discriminant Analysis*), a Análise Discriminante (DA, do inglês *Discriminant Analysis)* e a Análise Discriminante com Regressão por Mínimos Quadrados Parciais (PLS-DA, do inglês *Partial Least Square Discriminant Analysis*) (FONTES, 2020).

2.1.4.1 Análise Discriminante Linear

Análise discriminante linear (LDA, do inglês *Linear Discriminant Analysis*) é uma técnica de classificação supervisionada, também conhecida como discriminantes lineares de Fisher, que tem como grande vantagem operar no domínio dos dados originais e permitir uma interpretação mais fácil dos resultados. A LDA é baseada no cálculo de hiperplanos, ou funções discriminantes (FD) lineares (PONTES, 2020). Essas funções maximizam a separação entre as classes e minimizam a variância entre amostras de uma mesma classe. As FD lineares são obtidas por meio de combinações lineares das variáveis originais que melhor discriminem as classes (LIU et al., 2021). Uma fronteira é criada entre as amostras de acordo com as variáveis, estabelecendo o limite entre as duas classes, em que a amostra só pode pertencer a uma única classe. Conforme ocorre a distribuição dos dados das amostras, a LDA gera o discriminante, que atua como classificador, com base na matriz de dados **X** e nos conhecimentos prévios da classe de cada amostra para construção do modelo (FERREIRA, 2015; GOMES et al, 2022).

Assim, o objetivo da LDA é encontrar um espaço de projeção em que as amostras de diferentes classes possam estar com a máxima separação entre si. A LDA tem sido uma das técnicas de reconhecimento de padrão supervisionado mais utilizada em Quimiometria para

resolver problemas de classificação em que pelo menos duas classes estão envolvidas (ALMEIDA et al., 2021).

A LDA, assim como outras técnicas clássicas, tem sua funcionalidade limitada pela colinearidade entre as variáveis utilizadas no modelo. Deste modo, acredita-se que terá um melhor desempenho quando aplicada a conjuntos de dados com menores dimensões (NAES et al., 2001). Além disso, a capacidade de reprodução de modelos também pode ser comprometida por problemas de multicolinearidade. Outra restrição matemática relacionada à colinearidade e número de variáveis de entrada no modelo LDA, é que este deve ser menor que o número de objetos no conjunto de treinamento (FISHER, 1936). Por esse motivo, o uso da LDA para classificação, principalmente com dados espectrométricos, é indicado em matrizes que apresente redução da dimensionalidade inerente (menor número de variáveis do que de objetos) ou quando há aplicação de estratégias de seleção de variáveis (PONTES et al., 2020).

Para superar os problemas de colinearidade muitas vezes é necessário reduzir o número de variáveis do conjunto de dados original. O procedimento utilizado para essa finalidade emprega algoritmos de seleção de variáveis para indicar as variáveis significativas daquelas pouco informativas ou redundantes. As técnicas de seleção de variáveis geralmente contribuem para aumentar o poder discriminatório dos modelos em comparação aos métodos que utilizam todas as variáveis (GOMES et al. 2022).

2.2 Seleção de variáveis

A utilização das técnicas espectroscópicas contemporâneas com interfaceamento digital permite gerar muitas variáveis, dados e de informações sobrepostas. De acordo com Salimi et al. (2018), grandes volumes de dados trazem consigo variáveis irrelevantes ou redundantes, que contribuem apenas para aumentar o tamanho e a complexidade do espaço das variáveis, bem como a dificuldade de compreensão das informações e a interpretação das análises (QU et al., 2019).

A melhoraria do desempenho nos modelos matemáticos que utilizam instrumentação analítica com alta capacidade de gerar dados preconiza estratégias de seleção de variáveis, as quais eliminam comprimentos de onda, no caso de medidas espectrais, que não possuem informações relevantes para as condições avaliadas, bem como interferentes e variáveis responsáveis por produzir não linearidade do modelo, gerando um subconjunto contendo as variáveis que melhor definem a classe da amostra ou melhor se correlacionam com as propriedades de interesse (PASQUINI, 2018).

A aplicação de métodos que possibilitem a redução de dimensionalidade dos dados se torna necessária na calibração multivariada dos espectros, permitindo a identificação de padrões nos dados e a extração de informações relevantes (HUANG; LUO; XIA, 2019). A redução da dimensionalidade pode ser obtida através extração de variáveis, com a transformação das mesmas em novas variáveis em um novo espaço de menor dimensão; ou por meio da identificação do subconjunto das variáveis originais mais informativas para o problema analisado (ZHUO et al., 2021).

A seleção de variáveis auxilia na identificação de um subconjunto de variáveis que podem gerar um modelo mais preciso e exato. A escolha da região espectral é muito importante para a eficiência de um modelo multivariado. Quando escolhida corretamente, permite minimizar os erros de predição de forma a obter modelos mais robustos, de simples interpretação e a aumentar o desempenho de alguns métodos de reconhecimento de padrões e de regressão, pela eliminação de variáveis irrelevantes ou com informações redundantes, ou ainda removendo variáveis ruidosas (WANG et al., 2020). Além disso, reduzir as variáveis de um modelo torna-o menos complexo, facilitando a interpretação e entendimento do sistema que está sendo estudado, além de minimizar a multicolinearidade (BIANCOLILLO et al., 2016). Deste modo, a seleção de variáveis é amplamente utilizada e difundida em diversos campos da quimiometria, como por exemplo, em problemas de classificação (SILVA et al 2016), de calibração multivariada (ABREU et al. 2015), bem como de transferência de calibração (MARTINS et al. 2010; GALVÃO et al. 2015).

Para Xiaobo et al (2010), a seleção de variáveis envolve dois métodos, sendo, a escolha das variáveis que possam ser mais relevantes para os parâmetros a serem determinados e a escolha dos algoritmos que tenham maior capacidade de otimização do desempenho dos modelos. Vários algoritmos com diferentes especificidades têm se consolidado na literatura, e ainda outros têm sido desenvolvidos ou otimizados.

Assim sendo, os quimiometristas dispõem de uma diversidade de ferramentas de seleção de variáveis que podem ser testadas e comparadas. Os métodos de seleção de variáveis, geralmente, fundamentam-se em quatro etapas básicas, geração de um subconjunto de variáveis a partir do conjunto original dos dados, avaliação do subconjunto gerado, definição do critério de parada e validação do resultado (FONTES, 2020).

2.2.1 Geração do subconjunto

A seleção de um subconjunto de variáveis pode ser entendida como um problema de busca (FACELI, et al, 2011). O primeiro passo para se gerar um subconjunto de variáveis é definir o ponto de partida da busca, para o qual uma estratégia de busca deve ser definida. Desse modo, cada ponto no espaço de busca pode ser visto como um possível subconjunto de variáveis (FONTES, 2020).

As estratégias de busca podem ser do tipo completa, sequencial ou aleatória. As estratégias de busca completas asseguram encontrar o subconjunto ótimo, pois estas avaliam todos os possíveis subconjuntos de variáveis. Porém, uma busca exaustiva, em que todos os subconjuntos são testados, pode tornar-se impraticável, uma vez que existem 2^d subconjuntos possíveis para as d variáveis existentes (ARAUZO-AZOFRA et al., 2017; BLUM; LANGLEY, 1997; BOLÓN-CANEDO; SÁNCHEZ-MAROÑO; ALONSO-BETANZOS, 2015).

No caso de estratégias sequenciais, diferentes procedimentos de investigação são aplicados para reduzir o espaço de busca sem comprometer a chances de encontrar o resultado ideal. Estes incluem:

- ✓ Remoção de variáveis do conjunto original de dados, conhecido como abordagem *backward*;
- ✓ Inserção de variáveis em um conjunto inicialmente vazio (*forward*);
- ✓ Adição ou remoção simultânea de variáveis (*stepwise*) (BLUM; LANGLEY, 1997; CILIA et al., 2019). Porém, nem todos os subconjuntos possíveis são testados, assim, estratégias desse tipo podem omitir algumas variáveis relevantes, levando à perda de subconjuntos ótimos (SETIONO, 1997; FONTES, 2020).

Por fim, a estratégia aleatória, a qual busca subconjuntos com algum tipo de aleatoriedade (LIU; YU, 2005), incluem GARGS (*Genetic Algorithm – Based Region Selection*) (HASEGAWA; KIMURA; FUNATSU, 1997), MC-UVE (Monte Carlo – *Uninformative Variable Elimination*) (CAI; LI; SHAO, 2008), e MCS-RPLS (Monte Carlo – *Partial least squares regression*) (ZHANG; ZHANG; IQBAL, 2013).

2.2.2 Avaliação do subconjunto gerado

As abordagens de Seleção de Variáveis visam a seleção de um pequeno subconjunto dos dados originais capaz de reduzir a redundância e maximizar a relevância em

relação a função objetivo (TANG; ALELYANI; LIU, 2014). Cada subconjunto de variáveis gerado precisa ser examinado por um critério de avaliação. As estratégias de avaliação dos subconjuntos gerados têm a possibilidade de ser dependentes ou independentes do algoritmo de indução, sendo divididas em três abordagens principais: *filter* (filtro), *wrapper* (empacotamento) e *embedded* (embutido) ((BOLÓN-CANEDO; SÁNCHEZMAROÑO; ALONSO-BETANZOS, 2015; CILIA et al., 2019; PES, 2019). Há ainda uma tendência em combinar algoritmos de diferentes origens conceituais em um processo sequencial; métodos que fazem uso desta abordagem são denominados híbridos (REMESEIRO; BOLON-CANEDO, 2019).

Nos métodos de abordagem do tipo *filter* ou filtro, as variáveis são avaliadas considerando as características de sua natureza (EBRAHIMPOUR; EFTEKHARI, 2018). Habitualmente, um critério independente é utilizado em modelos do tipo *filter,* em que o processo de seleção de variáveis é guiado por meio da avaliação da qualidade de uma variável ou de um subconjunto de variáveis fazendo uso de uma medida de qualidade independente do algoritmo de indução que será aplicado às variáveis selecionadas (WAN, 2018).

Essas abordagens, normalmente que se apoiam em testes estatísticos de significância, são aplicadas previamente a um algoritmo de classificação ou predição (MURSALIN et al., 2017; JIN et al., 2019). A seleção é utilizada como uma etapa de pré-processamento por variáveis significativas estimando e classificando-as de acordo com sua importância. O subconjunto de atributos selecionados é apresentado como entrada para o algoritmo de classificação. Uma das vantagens da abordagem do tipo *filter, é o* baixo custo computacional e boa capacidade de difusão (BOLÓN-CANEDO; SÁNCHEZ-MAROÑO; ALONSO-BETANZOS, 2015).

Os métodos de abordagem do tipo *wrapper* ou de empacotamento utilizam algoritmos de aprendizado para avaliar os subconjuntos de variáveis e identificar os mais relevantes (BASGALUPP, 2007). Os algoritmos de otimização baseados em metaheurísticas como algoritmo genético (LEARDI, 2000), otimização por enxame de partículas (QASIM; ALGAMAL, 2018) e otimização da colônia de formigas (DORIGO; MANIEZZO; COLORNI, 1996) são característicos dessa abordagem. Do mesmo modo, como as abordagens do tipo *Filter* avaliam subconjuntos de variáveis, as abordagens *wrapper* promove uma busca entre os possíveis subconjuntos a serem avaliados, porém em vez de usar um teste independente como nas abordagens *Filter*, é utilizado o próprio algoritmo de indução para avaliar os subconjuntos de variáveis com base na sua capacidade preditiva (KOHAVI; JOHN, 1997).

As abordagens do tipo *wrapper* utilizam estratégias de busca do tipo sequencial e a importância da variável é baseada em medidas de precisão. Os subconjuntos de variáveis são avaliados utilizando algoritmos de aprendizado, de forma a selecionar o subconjunto que apresentar o melhor desempenho preditivo, embora avaliem as contribuições de cada variável em instrumentos de predição e classificação, caracterizam-se como um processo mais demorado do que o filtro (BASGALUPP, 2007; PARMEZAN et al., 2012). No entanto, a complexidade de tempo na abordagem *wrapper* é relativamente maior que em abordagens *filter* e *embedded*, devido o algoritmo de indução necessitar ser executado diversas vezes (WAN, 2018).

Nos métodos de abordagens do tipo *embedded*, as variáveis são selecionadas durante o processo de aprendizado. Os métodos de abordagem do tipo *embedded* combinam os dois métodos anteriores (*filter* e *wrapper*), porém a seleção de variáveis e o aprendizado não podem ser separados (RODRIGUEZ-GALIANO et al., 2018).

Os métodos de abordagem do tipo *embedded* selecionam o subconjunto de variáveis durante o próprio processo de construção do modelo de classificação (RODRIGUEZ-GALIANO et al., 2018). Alguns exemplos de métodos *embedded* são os algoritmos L1 (*Least Absolute Selection and Shrinkage Operator* ou LASSO) (LEE; CAI, 2018), modelos lineares generalizados (GLM – *Generalized Linear Models*)(TIBSHIRANI, 1996), árvores de decisão (*decision tree*), florestas aleatórias (*random forest*) (GEURTS et al., 2005; WU et al., 2003), máquinas de vetor de suporte (SVM – *Suport Vector Machine*) (WESTON; ELISSEEFF; SCHÖLKOPF, 2003; ZHANG et al., 2006) e redes neurais artificiais (ANN – *Network*)(SABANDO; PONZONI; SOTO, 2019).

Destaca-se que os métodos do tipo *wrapper* e *embedded*, por serem atrelados a algoritmos de aprendizado, apresentam um processo computacionalmente mais demorado do que os de filtro, além de estarem sob o risco de sobreajuste (ZHOU et al., 2021).

Na abordagem híbrida, as variáveis são classificadas com base em sua relevância e, assim, aquelas que apresentam *scores* mais altos são fornecidas ao método *wrapper*, de modo que o número de avaliações necessárias para o método *wrapper* seja menor, reduzindo a complexidade computacional. Além disso, observa-se que os métodos híbridos são computacionalmente mais complexos que os métodos *filter*, uma vez, que os métodos de abordagem do tipo híbridos combinam *wrapper* e *filter* e têm menos generalidade em comparação com os métodos *filter*, uma vez que utilizam o algoritmo de aprendizado supervisionado no processo de seleção de variáveis (FONTES, 2020).

2.2.3 Critério de parada

Em processos de seleção de variáveis é preciso ser adotado um critério de parada, estabelecendo o momento quando se termina a busca pelo melhor subconjunto de variáveis. Tal critério, por exemplo, pode dar-se por um número de variáveis a serem selecionadas ou um número máximo de alternativas testadas de forma que o desempenho do classificador ou do tempo de processamento não seja degradado (FACELI et al., 2011). Uma opção de melhorar a robustez dos algoritmos de seleção de variáveis é utilizar a abordagem *ensemble*, que é uma técnica onde se combinam vários modelos para resolver o mesmo problema (RIBEIRO; DOS SANTOS COELHO, 2020).

Embora a abordagem *ensemble* tenha provado sua eficácia nos últimos anos, a aplicação em outras disciplinas do aprendizado de máquina, como a seleção de variáveis, tem sido pouco explorada. Em geral, este tipo de técnica tem ganhado destaque em aplicações dentro do processo de aprendizado dos algoritmos (predição/classificação) (BOLÓN-CANEDO; ALONSO-BETANZOS, 2019). Apesar do vasto repertório de abordagens disponível na literatura permitir à seleção de variáveis, porém não há um consenso sobre a melhor ou mais eficiente técnica a ser utilizada em cada situação, resultando que a identificação das variáveis mais informativas ainda se mostre como um tema complexo e com abertura para novas abordagens (MUÑOZ-ROMERO et al., 2020).

2.2.4 Algoritmos Determinísticos e Estocásticos para seleção de variáveis

Um critério que deve ser considerado durante a avaliação do método de seleção de variáveis, é a maneira como o algoritmo de seleção de variáveis é executado para um grande número de variáveis. Algoritmos determinísticos, que expressam ao final do cálculo uma solução única apresentam bom desempenho para pequenos problemas, ou seja, tal método sempre leva à mesma resposta, após diferentes execuções, desde que aplicados sempre às mesmas condições iniciais. Porém falham ao passo que o número de variáveis é aumentado. Em contrapartida, algoritmos estocásticos exibem melhores resultados em dimensões mais altas, apesar de não evidenciar exclusivamente um resultado, mas a tendência de encontrar a melhor solução global (RAPHAEL et al., 2003), devido ao uso aleatório em busca de um conjunto de soluções, ou seja, o subconjunto de variáveis selecionado está associado a um certo grau de probabilidade (FISTER et al., 2013; PONTES,2020).

De acordo com o conjunto de dados complexos gerados pelos métodos instrumentais, o desenvolvimento de algoritmos bioinspirados para a seleção de variáveis torna-se essencial para se obter um melhor desempenho nos modelos de classificação (HAIR et al., 2009; SILVA et al 2019). Os algoritmos meta-heurísticos ou bioinspirados são inspirados por comportamentos biológicos de animais, insetos ou aves, visando encontrar a solução ideal através de busca cega ou pesquisa informada mediante função heurística. Deste modo, algoritmos evolutivos realizam busca iterativa com o propósito de determinar a solução para o problema definido pela convergência de respostas ou pelo número de ciclos (DARWISH, 2018). Como exemplos, o algoritmo de otimização por colônia de formigas e o algoritmo genético.

Diversos métodos têm sido propostos para a finalidade de seleção de variáveis. Para fins da construção deste trabalho foram abordados como algoritmos determinísticos, o SPA (PONTES et al., 2005) e como algoritmo estocástico algoritmo de otimização por colônia de formigas (ACO, do inglês *Ant Colony Optimization*) (Pontes *et al* 2020) e algoritmo genético (GA, do inglês *Genetic Algorithm*) (ALLEGRINI; OLIVIERI, 2011; KATOCH; CHAUHAN; KUMAR, 2021).

2.2.4.1 Algoritmo de Projeções Sucessivas (SPA)

O Algoritmo de Projeções Sucessivas (SPA, do inglês: Successive Projections Algorithm) foi originalmente proposto por Araújo e colaboradores em 2001 para minimizar problemas de multicolinearidade em Regressão Linear Múltipla (MLR) para conjuntos de dados espectroscópicos, melhorando assim, a capacidade preditiva desses modelos. O iSPA-PLS é um algoritmo desenvolvido por Araújo et al 2012, que seleciona intervalos otimizados em PLS e emprega o SPA para guiar a seleção de intervalos. O iSPA-PLS é executado em duas fases, onde inicialmente é calculado o número ótimo de fatores para o modelo completo PLS empregando o processo de validação cruzada. Sendo essa uma estratégia do modelo para uma inicial estimativa que indica o número ótimo de fatores na etapa de seleção de intervalos. Na segunda fase, é gerada uma matriz SEL contendo os índices das variáveis representantes que forma cada cadeia de intervalos (FERNANDES, 2016).

Pontes e colaboradores em 2005, posteriormente adaptaram o SPA para modelos de classificação, bem como para resolver problemas de multicolinearidade em análise discriminante linear (LDA) (PONTES, 2020). O SPA é realizado em três fases. Na primeira são geradas as cadeias de variáveis empregando somente a matriz Xcal, geralmente centrada na média das colunas. Na segunda fase avalia-se a correlação das cadeias com o parâmetro de

interesse. Já na terceira etapa elimina-se as variáveis que não trazem melhorias em termos de predição (ARAUJO et al.2001).

Para os métodos de reconhecimento de padrões, a informação da modelagem está incluída nos dados oriundos da resposta instrumental e no índice de classes para cada amostra. Deste modo, a versão do SPA aplicada a LDA, apresenta uma diferença principal, que está na utilização da função de custo. Utiliza-se o risco médio ao invés do RMSEV para guiar a classificação. Com isso, um menor risco será obtido quanto mais distante o objeto estiver do centro da classe incorreta e mais próximo estiver do centro de sua classe verdadeira (SOARES et al., 2014; PONTES et al., 2005).

Soares et al. (2014) propôs uma alternativa para resolver problemas de classificação, que envolvem a utilização de um número reduzido de amostras. Nesse sentido, o conjunto de validação é dispensado e o risco de uma classe incorreta é calculada usando os mesmos dados do conjunto de treinamento que foram empregados para calcular as médias de cada classe e uma matriz de covariância. Dessa forma, evita possíveis problemas de sobreajuste relacionada ao uso repetido do conjunto de treinamento.

2.2.4.2 Algoritmo de otimização por Colônias de Formigas (ACO)

O algoritmo de otimização por colônia de formigas são modelos matemáticos que se baseiam no comportamento das formigas, representando uma classe de meta-heurísticas bioinspiradas baseadas no comportamento das formigas. Desde a proposta do Ant System, o primeiro algoritmo ACO proposto por Doringo et al 1996, muitos resultados significativos de pesquisa foram obtidos.

Nos algoritmos ACO, o sistema é transformado em um vetor memória, e é representado por uma trilha de feromônios. Esta trilha governa a maneira com que os agentes navegam pelo espaço de busca e retomam suas experiências. Essa substância biologicamente ativa faz as formigas encontrarem esse rastro de feromônio, escolherem não saírem mais de forma aleatória em outros caminhos e tendem a seguir o caminho com feromônio. As formigas que encontram o formigueiro primeiro, procuram retornarem mais rapidamente ao ninho pela menor distância alcançada (RANZAN, 2021). O feromônio evapora com o passar do tempo. Quanto maior a quantidade de formigas seguindo o rastro, maior a concentração do rastro dele deixou no caminho do trajeto, levando assim, menor probabilidade das formigas se desviaram no trajeto (ALLEGRINI & OLIVIERI, 2011; PESSOA, 2015). No caso da seleção de variáveis e otimização de modelos, cada variável de entrada inicia a rotina de otimização com uma mesma

quantidade de feromônio. A cada iteração, cada uma das formigas seleciona um subconjunto das variáveis de entrada. Essa seleção se baseia em uns componentes randômicos e um componente relacionado a quantidade de feromônio nas variáveis.

Allegrini e Olivieri (2011) observaram que no primeiro passo de tempo, todas as variáveis têm a mesma probabilidade de serem selecionadas, mas como as quantidades de feromônio são atualizadas em intervalos de tempo sucessivos, essas probabilidades serão diferentes (PONTES, 2011). A formiga entrega para o algoritmo de aprendizado o subconjunto, que ajusta um modelo para predizer as saídas de interesse. A performance do modelo é quantificada, e cada formiga deposita nas variáveis que selecionou uma quantidade de feromônio condizente, quanto melhor a performance, maior a quantidade de feromônio. Por fim, toda a trilha é evaporada, multiplicando a mesma por um valor entre 0 – 1, para penalizar variáveis não selecionadas. Esta operação garante um aumento da probabilidade, nas sucessivas iterações, da seleção de variáveis de entrada que participaram de modelos de alta qualidade, otimizando o preditor (RANZAN, 2021).

O ACO é aplicado para diversas áreas, como por exemplo, na robótica (LINGARA et al., 2013), engenharia (SILVA, 2016), matemática (PIRES, 2019) e na quimiometria (RANZAN et al., 2014; NACHE et al., 2015).

Embora bastante conhecido em problemas de otimização, as aplicações da Técnica de Otimização de Colônia de Formigas, foi introduzido na área da química há pouco mais de uma década. O algoritmo colônia de formigas vem sendo explorado no contexto de calibração multivariada como um método de seleção de variáveis. Em que a maioria dos métodos de seleção de variáveis foram desenvolvidos e empregados nesse contexto e adaptados para classificação multivariada (RANZAN et al., 2014).

Pontes, 2020 propôs o método que consiste numa combinação da seleção de variáveis via colônia de formiga associado a modelos de classificação baseados em analise discriminante linear, nomeado, ACO–LDA.O ACO acoplado ao LDA é um algoritmo interativo que busca uma seleção de um subconjunto de variáveis de para discriminar amostras. Onde são solicitados inicialmente que se forneça alguns valores de entrada, sendo eles: número de formigas na colônia (**ants**), número de colônias (**c**) por interações (**int**), taxa de formigas cegas (**b**) em cada colônia e em cada ciclo, taxa de evaporação do feromônio (γ) e o número máximo de variáveis (**vmax**) que podem ser incluídas por formigas.

Assim, o algoritmo ACO-LDA proposto por Pontes,2020 foi avaliado em três problemas de classificação envolvendo duas diferentes técnicas analíticas (espectrometria NIR e UV-VIS) nos casos: (1) classificação de óleos vegetais comestíveis via espectrometria UV-VIS, (2)

classificação de amostras de chá em relação ao tipo e origem geográfica via espectroscopia NIR e (3) Classificação entre três cultivares de feijões baseada em medidas nano NIR. Nos três casos, o método ACO-LDA selecionou um pequeno subconjunto de variáveis, levando à classificação correta da maioria das amostras e obtendo resultados consideravelmente melhores GA-LDA e tão bons quanto aqueles encontrados usando PLS-DA, ferramentas quimiométricas já bem estabelecidas.

2.2.43 Algoritmo Genético – GA

O algoritmo genético foi proposto por John M. Holland, na década de 60, ele objetiva aperfeiçoar sistemas biológicos complexos (COSTA FILHO, et al., 1999). O primeiro registro da utilização do GA em química foi o trabalho de LUCASIUS e KATEMAN, em 1993, no qual a técnica foi usada para selecionar comprimentos de onda na região ultravioleta aplicado a determinação de nucleotídeos (LUCASIUS et al., 1993).

A seleção de variáveis por GA apresenta basicamente três etapas, sendo a primeira de codificar as variáveis. O processo mais comum é a codificação binária de valor 1 ou 0, que representa, respectivamente, se a amostra está incluída ou não no modelo. A segunda etapa é a avaliação de aptidão, que representa a capacidade de gerar melhores respostas, para seleção de variáveis, quanto maior a aptidão menor o erro de predição obtido. A aptidão é obtida calculando um modelo de regressão para cada amostra e estimando se o Erro Quadrático Médio (RMSE) para um conjunto de amostras externas ou por validação cruzada. Os que apresentarem boa aptidão serão selecionados para a terceira etapa que é a de reprodução (COSTA FILHO, et al., 1999).

Após o desenvolvimento dos modelos, antes que os mesmos possam ser aplicados a predição de amostras as quais não se tem conhecimento prévio da classe, faz-se necessário procedimentos de validação do mesmo, a fim de comprovar sua capacidade preditiva. Para isso, utilizam-se critérios de avaliação de desempenho conhecidos como parâmetros de desempenho.

2.2.5 Avaliação de Desempenho

A validação do modelo gerado pode ser realizada pela comparação da medida do número de acertos (para casos de problemas de classificação) ou de erros (para os casos de problemas de regressão) obtidos pelo modelo com as variáveis selecionadas, em relação à medida obtida com a utilização na totalidade das variáveis disponíveis.

Para as técnicas de calibração multivariada, as medidas mais comumente usadas para a verificação de eficiência ou validação de um modelo de regressão estão relacionadas aos critérios de RMSE, r^2 e REP.

A raiz do erro médio quadrático (do inglês: *Root Mean Square Error – RMSE*) e erro médio quadrático (do inglês: *Mean Square Error – MSE*) são os dois parâmetros estimadores utilizados em modelos de regressão (MOREIRA, 2018). O RMSEC - Raiz quadrada do erro médio quadrado da calibração, expressa a concordância entre os valores estimados e os de referência. Este parâmetro é avaliado utilizando todas as amostras do conjunto de calibração, demostrada na equação 1.

$$\mathbf{RMSEC} = \sqrt{\frac{\sum_{i=1}^{n}(y_{ical} - \widehat{y_{ical}})^2}{\mathbf{n_c - VL - 1}}} \qquad \text{Equação 1.}$$

Sendo, nc o número de amostras do conjunto de calibração, VL é o número de variáveis latentes do modelo, yi,cal, e ŷi,cal são os valores de referência e estimados, respectivamente. Embora o RMSEC seja um parâmetro importante para medir o ajuste nas amostras de calibração, ele não pode ser usado para estimar o número de variáveis latentes do modelo por causar um sobreajuste. Deste modo, as mesmas amostras de calibração são analisadas utilizando um procedimento de validação cruzada, que permite estimar o RMSECV (FERREIRA, 2015).

O RMSECV – Raiz quadrada do erro quadrado da validação cruzada, expressa na equação 2.

$$\mathbf{RMSECV} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{\mathbf{n_c}}} \qquad \text{Equação 2.}$$

O RMSEP – Raiz quadrada do erro médio quadrado da predição, é baseado na concordância entre os valores preditos pelo modelo e os valores de referência de um conjunto de amostras que não estavam presentes no desenvolvimento do modelo. Logo, o RMSEP é uma figura de mérito utilizada para verificar a acurácia do método de regressão na fase de validação, e pode ser expressa pela Equação 3 (FERRÉ et al., 1997; SANTOS JUNIOR et al., 2011).

$$\mathbf{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n}(y_{ival} - \widehat{y_{ival}})^2}{\mathbf{n_v}}} \qquad \text{Equação 3.}$$

Sendo, nv é o número de amostras de validação, yi,val, e ŷi,val são os valores de referência e estimados, respectivamente.

Enquanto, o coeficiente de determinação (**R^2**), indica o nível de ajuste do modelo aos dados, tanto para amostras de calibração como de validação.

Já o REP, erro relativo da predição, permite que a precisão relativa dos conjuntos de calibração e validação possam ser calculados, conforme a Equação 4 (BRAGA et al., 2014)

$$\mathbf{REP} = \sqrt{\frac{1}{n}\sum_{i-1}^{n}\frac{(\hat{y}_i - y_i)^2}{{y_i}^2}}\ \text{x}100 \qquad \text{Equação 4.}$$

Em que n é o número de amostras de calibração (nc) ou validação (nv) dependendo do conjunto de dados considerado, ŷi e yi são, os valores estimados e de referência, respectivamente, da propriedade de interesse, das amostras seja de calibração ou validação.

O Ajuste Linear da reta é a figura de mérito utilizada para estimar a qualidade do modelo no conjunto de validação, e verificar se os valores estimados são estatisticamente equivalentes com aqueles de referência, segundo a proposta de Rius e Rius (1997). O cálculo de regiões elípticas de confiança conjunta (**EJCR**, do inglês - *Elliptical Joint Confidence Regions*) permite a avaliação do ajuste de resposta linear para dados multivariados, bem como a presença de viés, que é um erro sistemático nos resultados. Após a estimativa dos intervalos de confiança para construção de resposta linear de projeção entre as respostas do modelo e de referência. Verifica-se se esses intervalos incluem os valores teóricos de 0 e 1 para a intercepção e inclinação, respectivamente, o que indicará que o método não possui desvio sistemático constante ou proporcional nos resultados. Com a utilização de um gráfico, pode ser visualizado a presença do ponto (1,0), que representa a posição ideal da inclinação e do intercepto, respectivamente, dentro da elipse de confiança (RIU & RIUS, 1997).

A robustez de um modelo de classificação ou de reconhecimento de padrões está associada aos critérios de avaliação de desempenho, como acurácia, sensibilidade e especificidade, também denominadas de parâmetros de desempenho (RODIONOVA; POMERANTSEV,2020). Eles são calculados a partir do número de erros e acertos na predição, ou seja, o número de falsos positivos (FP), verdadeiros positivos (VP), falsos negativos (FN), verdadeiros negativos (VN) (LOPES et al. 2015; FERREIRA, 2015; GHOSH et at., 2019).

A acurácia (exatidão ou erro) ou taxa de classificação correta, correspondente a proporção de classificações adequadas no conjunto de dados, de acordo com a equação 5:

Acurácia (ACC) (%) = $\frac{VP+V}{VP+FN+FP+V}$ X 100 Equação 5.

Os verdadeiros positivos (VP) estão relacionados a quantidade de amostras pertencentes à classe alvo e classificadas como tal e os verdadeiros negativos (VN) estão relacionados a predição correta às classes negativas (ZONTOV *et al.* 2017). Da mesma forma, falsos positivos (FP) dizem respeito a predição incorreta às classes positivas e falsos negativos (FN), a predição incorreta às classes negativas, considerado o número de amostras alvo classificadas erroneamente fora da área de aceitação (GHOSH et at., 2019; NEVES; POPPI, 2020).).

A Sensibilidade é definida como a porcentagem de amostras da classe alvo que são corretamente classificadas como amostras-alvo, baseia-se na capacidade do modelo em estimar corretamente a classe das amostras em razão ao número de amostras total pertencentes a classe modelada (NEVES; POPPI, 2020), como mostrada na Equação 6:

Sensibilidade (%) = $\frac{VP}{VP+F}$ X 100 Equação 6.

A Especificidade é representada como a porcentagem de amostras não-alvo que foram classificadas corretamente como amostras não-alvo. Esse parâmetro determina a capacidade do modelo em estimar corretamente as amostras que não pertencem a classe modelada, demonstrada na Equação 7.

Especificidade (%) = $\frac{VN}{FP+V}$ X 100 Equação 7.

Desse modo, modelos classificatórios que possuem valores ótimos nas figuras de mérito de Acurácia, Sensibilidade e Especificidade com valores limites tendendo para a unidade ou 100% de estimativa, conseguindo assim, descrever ou modelar a maior parte possível das amostras.

2.3 Espectroscopia no infravermelho próximo

A espectroscopia reúne um conjunto de técnicas baseadas na absorção e emissão de luz e outras radiações pela matéria. Essas técnicas podem ser utilizadas para analisar amostras de diferentes naturezas quanto à sua composição química (GATIUS et al., 2017), e são amplamente utilizadas em diferentes áreas de estudo, como médica (GEYIK et al., 2021), alimentícia (WAFULA et al., 2022), bebidas (KAHMANN et al., 2017), farmacêutica

(ANZANELLO et al., 2013), agricultura (XING et al., 2021) e petróleo (ALVES; POPPI, 2013) entre outros.

A região na faixa do infravermelho é abrangente e compreende a radiação eletromagnética com comprimentos de onda de 780 a 1.000.000 nm, sendo subdividida em três sub-regiões (**Tabela 1)** (SKOOG; HOLLER; NIEMAN, 2009).

Tabela 1 – Regiões espectrais do infravermelho

Região Espectral	**Número de onda (cm-1)**	**Comprimento de onda (nm)**	**Frequência (Hz)**
Próximo (NIR)	12.800 a 4.000	780 a 2.500	$3{,}8 \times 10^{14}$ a $1{,}2 \times 10^{14}$
Médio (MID)	4.000 a 200	2.500 a 50.000	$1{,}2 \times 10^{14}$ a $6{,}0 \times 10^{12}$
Distante (FAR)	200 a 10	50.000 a 1.000.000	$6{,}0 \times 10^{12}$ a $3{,}0 \times 10^{11}$

Fonte: SKOOG; HOLLER; NIEMAN (2009)

A técnica de espectroscopia na região do infravermelho é baseada na interação entre a radiação eletromagnética e a amostra. As moléculas da amostra, em seu estado de energia fundamental, absorvem o estímulo da energia e passam para o estado excitado. A radiação absorvida, proveniente de fonte externa e transformada em fenômenos de transição vibracional ou rotacional, é medida em função do comprimento de onda (FERREIRA, 2015).

A espectroscopia na região do infravermelho próximo (do inglês: *Near Infrared Spectroscopy*, NIR) é uma técnica analítica com vantagens de ser rápida e não destrutiva, garantindo a integridade da amostra, permite a obtenção de informações químicas de forma rápida e sem o uso de solvente orgânico, apresenta baixo custo e pode ser aplicado em todas as etapas do processamento, desde a colheita a inspeção do produto (MEDEIROS, 2021) frequentemente, empregada para quantificação e avaliação de informações químicas (RAVAZZI, 2019). Destacando-se pela sua versatilidade, sendo utilizada para análise com pouco ou nenhum pré-tratamento de amostras. Além de poder ser empregada na análise de amostras líquidas, sólidas e gasosas (OZAKI et al., 2021).

Essa técnica é baseada na interação entre a radiação eletromagnética e os constituintes da amostra. Sendo resultantes de sobretons, combinações e ressonâncias de vibrações fundamentais das ligações, envolvendo carbono, nitrogênio, oxigênio e enxofre com um átomo de hidrogênio. Essas interações são detectáveis na região do infravermelho próximo na faixa de comprimento de onda entre de 780 a 2500 nm, de forma que os grupos funcionais presentes

nessas moléculas contêm informações analíticas quantitativas e qualitativas a respeito da amostra relacionadas a sua composição (SU et al., 2017; ZAREEF *et al.,* 2021).

As metodologias que utilizam a tecnologia de espectroscopia NIR têm se mostrado altamente eficientes na substituição de métodos de análises laboratoriais, em campo e em diversos setores, como análises de produtos alimentícios, farmacêuticos e agrícolas, entre outros, viabilizando a avaliação de propriedades e características, em um largo espectro de exceção (PASQUINI, 2018).

A espectroscopia no infravermelho próximo (NIR) tem sido amplamente investigada como um método não destrutivo e demonstrou resultados promissores para a análise de sementes (BADARÓ et al., 2020; BARBIN et al., 2018; CAPORASO et al., 2018a, 2018b; KAUR et al., 2016), e da autenticidade de alimentos (CRUZ-TIRADO et al., 2020; OLIVEIRA et al., 2020; ORRILLO et al., 2019).

As formas de disposição de amostras e obtenção de espectros NIR podem ser por transmitância ou absorbância, reflectância difusa, transflectância e interactância (PASQUINI, 2018). Na medida de sólidos como sementes, o processo mais utilizado é por reflectância difusa para determinações não destrutivas e não invasivas, a técnica de reflectância tem sido utilizada com a finalidade de determinar quantitativamente proteínas, lipídeos e umidade, em produtos agrícolas como grãos e sementes oleaginosas, e em outras amostras sólidas (HOLLER, 2009).

O fenômeno da reflectância difusa é observado quando uma fração da radiação incidente sobre uma superfície de um sólido é refletida e a fração complementar é absorvida pela amostra. Uma função R é obtida para estabelecer a relação entre a radiação incidente e a radiação absorvida (ALMEIDA,2013). A absorbância interna e o espalhamento contribuem com uma variação de intensidade do sinal, implicando em um tratamento rigoroso dos sinais (PASQUINI, 2018).

Cada composto possui um espectro característico (BARBIN et al., 2018), uma vez que os níveis de absorção de radiação em cada comprimento de onda estão associados à sua composição química (RADY & ADEDEJI, 2018). As informações espectrais geradas fornecem uma *impressão digital* do composto e podem ser utilizadas para quantificar compostos específicos primários, como ácidos graxos, açúcares, proteínas e umidade (BADARÓ et al., 2019). E, além disso, para diferenciar e classificar de acordo com fatores relacionados a condições da cultura/processo, origem geográfica, variedades ou genótipos, desde que estes influenciem na sua composição química (COZZOLINO, 2016).

Os modelos desenvolvidos a partir de dados espectroscópicos necessitam de uma etapa de remoção de características espectrais não relacionadas ao problema químico, como por

exemplo ruído espectral. Os espectros brutos, de forma geral, devem ser evitados para criar modelos robustos, pois podem conter fontes de variabilidade espectral que não estão relacionadas com a composição química da amostra, como ruído aleatório, dispersão de luz, variação da inclinação e deslocamento da linha de base. Para que se possa realizar a modelagem os dados instrumentais são submetidos a pré-processamentos (FERREIRA, 2015; PASQUINI, 2018). Apesar da ampla aplicabilidade da espectroscopia NIR, deve-se observar que diferente de outras técnicas espectroscópicas, a informação analítica gerada é condicionada a utilização de ferramentas que viabilizam o processo analítico. Desta forma, a tecnologia requer a combinação da espectroscopia com o uso de ferramentas quimiométricas para o tratamento dos espectros, em que a estatística pode ser agregada aos dados químicos para geração de conhecimento analítico (MAYRINCK, 2018).

2.4 Quimiometria e NIR utilizando algoritmos de seleção de variáveis aplicadas a área agrícola

A produção do algodão está cada vez mais associada a tecnologias de agricultura de precisão, de modo que estão em todas as etapas de produção, a exemplo da utilização de drones para monitoramento e aplicação de produtos químicos e biológicos, uso de softwares de gestão, realização de colheita mecanizada, além das aplicações de biotecnologia, nanotecnologia e do melhoramento genético. Tais inovações tem influência positiva para a agregação de valor na cadeia produtiva de algodão (ARLINDO, 2021).

Com um mercado internacional promissor e o aumento expressivo em área plantada, os produtores demandam cada vez mais por tecnologias e insumos de qualidade que possam garantir a aceitação do produto frente ao mercado. O tratamento em sementes é uma forma eficiente e cada vez mais necessária no manejo de doenças e insetos. A qualidade da semente é um dos mais importantes fatores atrelados ao sistema de produção (LAUXEN et al., 2010). A produção de sementes com qualidade sanitária é um exemplo de tecnologia que pode ser proporcionado pelo alto rendimento dessas culturas (FARIAS, 2017).

Uma das preocupações dos bancos de germoplasma é a análise de qualidade de sementes, uma vez que os estoques de sementes de alguns genótipos podem ser pequenos e a maioria dos testes são destrutivos. Pensando nesse inconveniente, várias técnicas baseadas em espectroscopia óptica têm sido estudadas, cujos benefícios em relação aos métodos tradicionais são a maior velocidade analítica, facilidade de operação e de não destruir a amostra (RAHMAN E CHO, 2016).

A modificação genética tem sido uma alternativa para melhorar a produtividade e a resistência a pregas. O algodão transgênico é uma tecnologia que contribui no combate a pragas no campo. De modo que, antes do desenvolvimento da primeira planta transgênica de algodão, havia a necessidade de grandes quantidades de pulverização de inseticidas para conter os danos causados pelos insetos-pragas do algodoeiro e ervas daninhas, elevando os custos de produção. O desenvolvimento de cultivares geneticamente modificadas representam uma importante estratégia de manejo sustentável (ISAAA, 2019).

No Brasil, a aprovação do primeiro genótipo de algodão transgênico pela Comissão Técnica Nacional de Biossegurança (CTNBio), ocorreu em 2005. Desde então, outros genótipos foram aprovados. Atualmente, cerca de 80% do total da produção de algodão é conduzido adotando sementes transgênicas. Assim como, cerca de 76% de toda lavoura de algodão plantada no mundo emprega esse insumo biotecnológico (CTNbio,2020).

Os métodos atualmente empregados para identificação de produtos transgênicos incluem métodos avançados de biologia molecular associados a marcadores específicos baseados em proteínas, enzimas, sequencias de aminoácidos, microscopia, espectroscopia e cromatografia (WANG et al., 2014). Porém, esses métodos comumente usados, apresentam inconvenientes como destruição da amostra, elevado tempo de análise. Além do que, os métodos que envolvem ferramentas de biologia molecular, possuem alto custo e ainda tem acesso restrito a alguns laboratórios. Desse modo, têm-se aumentado a busca por métodos analíticos rápidos e precisos para aplicações na determinação da qualidade dos produtos e de parâmetros de processamento (KAUFMANN et al., 2019).

A espectroscopia NIR, por suas características intrínsecas como não destrutiva, rápida e com possibilidade de portabilidade e miniaturização, torna-se uma alternativa viável para a identificação de sementes transgênicas de algodão. A base de conhecimento para fenotipagem usando a tecnologia NIR envolve a medida do sinal de absorção relativo às ligações moleculares, como C–H, C–N e C–O, que está relacionada às alterações de composição fenotípicas causadas por alterações genotípicas (SOARES et al., 2016).

Nas pesquisas na área de sementes obtiveram resultados satisfatórios na avaliação da qualidade fisiológica de soja, algodão, café e tomate (BAZONI et al., 2017; GAITÁN et al., 2008; HUANG et al., 2012; GUIMARÃES, 2016; SHRESTHA et al., 2016). Além disso, o uso da espectroscopia no infravermelho associado a análise multivariada foi usado com sucesso para predição do teor de óleo em sementes de genótipos de girassol (GRUNVALD et al., 2014).

Soares (2016), utilizou a espectroscopia NIR associada a quimiometria, para classificação de sementes de algodão de quatro cultivares de alta qualidade genética, para isso

verificou a eficiência de dois métodos de classificação, o SPA-LDA (Algoritmo das Projeções Sucessivas com a Análise Discriminante Linear) e o PLS-DA (Regressão por Mínimos Quadrados Parciais com Análise Discriminante). O autor obteve uma precisão de 96,91% de sementes classificadas corretamente.

Tibola et al. (2018) reuniram as contribuições de uma equipe multidisciplinar e multiinstitucional, abrangendo diferentes áreas de aplicação de espectroscopia no infravermelho próximo para grãos, e produtos derivados na avaliação de indicadores de qualidade e de contaminantes em grãos.

Ferreira (2019) apresentou o desenvolvimento de um método para a identificação de sementes haploides que possuíssem o gene marcador *R1- navajo* em milho utilizando a espectroscopia NIR e a análise discriminante por mínimos quadrados parciais (PLS-DA). Sendo avaliados por parâmetros como especificidade, sensibilidade e eficiência no conjunto de testes de validação do modelo que obtiveram valores iguais a 100% para o conjunto de treinamento e para o conjunto de testes. Os resultados evidenciaram que a espectroscopia NIR combinada PLS-DA é uma tecnologia que apresenta perspectivas favoráveis para a seleção rápida e não destrutiva de sementes de milho haploides.

A combinação da espectroscopia NIR e métodos de classificação na análise de grãos e sementes, é cada vez mais utilizada para desenvolver métodos de triagem capazes de fornecer ferramentas rápidas para a resolução de problemas que envolvam comprovação de autenticidade, atribuição de origem, detecção de contaminações, fraudes e falsificações, dentre outros (VITALE et al., 2013; COZZOLINO, 2014; MARQUETTI et al., 2016).

Mata e colaboradores (2022), utilizou a Análise Discriminante por Mínimos Quadrados Parciais (PLS-DA) em dados NIR e Raman para distinguir genótipos de sementes de algodão convencionais e transgênicos obtendo erros de classificação de predição de 2,23% para NIR e 0% para o Raman.

3. METODOLOGIA

Os procedimentos experimentais deste trabalho foram realizados em colaboração com Laboratório Avançado de Tecnologia Química- (LATECQ) localizado na Embrapa Algodão, em Campina Grande, PB. O conjunto de dados utilizado é composto por espectros de sementes de algodão convencionais e transgênicas, obtidos por Almeida e colaboradores (2013) e Mata e colaboradores (2022). Os referidos conjuntos de dados foram utilizados para o desenvolvimento de modelos de calibração multivariada (i-PLS) e LDA construídos com as variáveis selecionadas pelos algoritmos SPA-LDA, GA-LDA e ACO-LDA.

3.1 Amostras e obtenção de dados instrumentais

3.1.1 Modelos de calibração multivariada para determinação de umidade em sementes de algodão

Foram utilizados os dados espectrais adquiridas por Almeida e colaboradores (2013) das amostras de genótipos de algodão do Banco Ativo de Germoplasma (BAG) da Embrapa Algodão. Foram utilizados os dados espectrais de 30 acessos contendo cerca de 2g de sementes para as medidas e construção dos modelos de calibração. As medidas espectrais de reflectância difusa na região do visível e infravermelho próximo foram adquiridas por Almeida e colaboradores (2013) utilizando um espectrômetro VIS/NIR modelo XDS Analyser (Foss Analytical, Hogans, Sweden) equipado com uma célula de quartzo de 3 cm de diâmetro. Cada espectro foi adquirido como a média de 32 varreduras na faixa de 1101 a 2500 nm, com intervalos de 0,5 nm.

3.1.2 Modelos de reconhecimento da padrões para distinção de sementes transgênicas e convencionais

Foram utilizados os dados espectrais adquiridas por Mata et al 2022 de amostras de sementes convencionais e transgênicas de algodoeiro (*Gossypium* L.), cedidas pela Embrapa Algodão, Campina Grande-PB. A cultivar transgênica utilizada foi a BRS 368 RF, enquanto a cultivar convencional foi a BRS Aroeira. Foram utilizadas 1195 sementes de cada cultivar, totalizando 2390 espectros. As amostras foram acondicionadas por cerca de uma hora antes da aquisição dos espectros em ambiente controlado a uma temperatura média de 20°C e umidade

relativa do ar média de 64%. As medidas espectrais de reflectância difusa na região do visível e infravermelho próximo foram adquiridas no mesmo equipamento e com as mesmas configurações mencionadas no item 3.1.1.

3.2 Determinação do teor de água em sementes pelo método da estufa

O teor de água nas sementes de algodão estudadas foi determinado, de acordo com o método preconizado pelo Ministério da Agricultura, por Almeida e colaboradores (2013). Para isto a temperatura da estufa foi regulada na temperatura 105±3°C. Todos os recipientes foram secos em estufa (a 105°C) por 30 minutos, depois resfriados em dessecador, pesados e identificados. Cerca de 5,0 g de sementes foram pesadas e distribuídas uniformemente nos recipientes. Então, a massa do recipiente com a amostra foi registrada. Em seguida foi colocado na estufa, a 105°C, por 24 horas. Após resfriar em dessecador, a massa foi novamente registrada. O teor de umidade foi calculado pela diferença de massa após a secagem.

3.3 Softwares utilizados e construção dos modelos

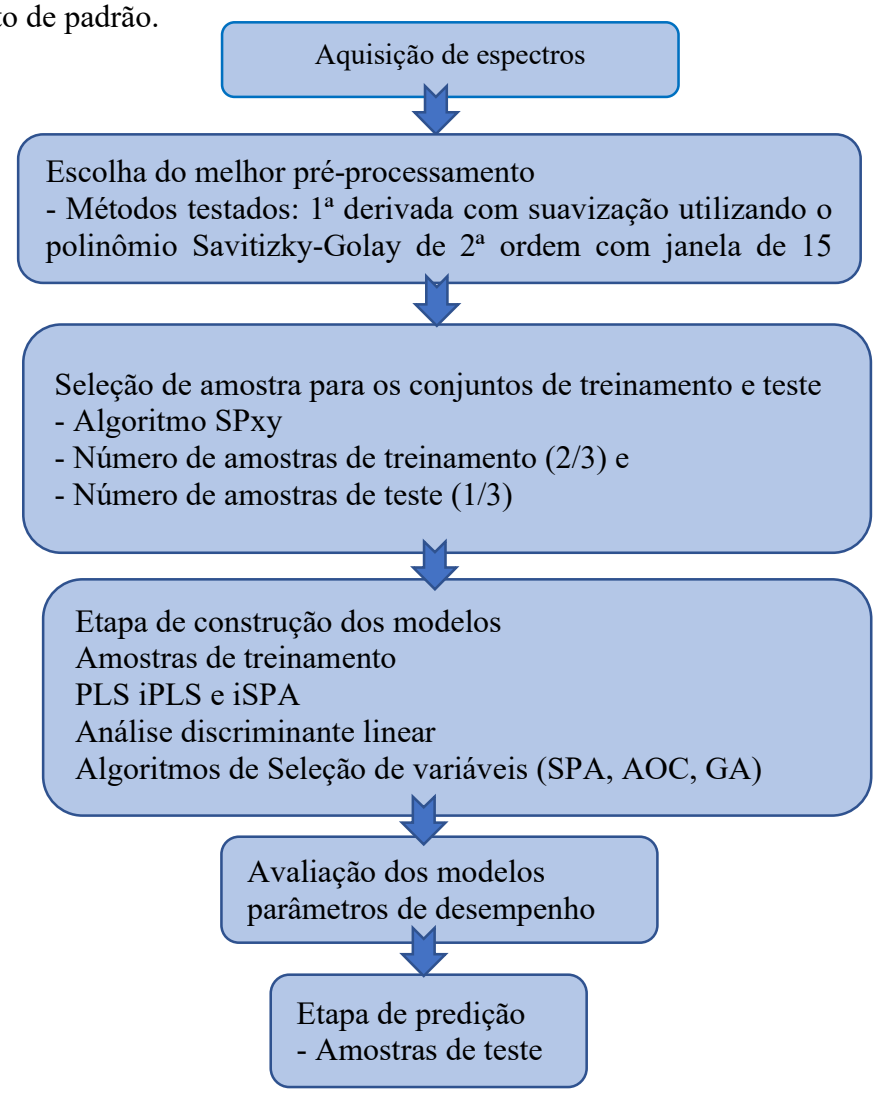
O tratamento dos dados para os modelos de reconhecimento de padrões foi realizado utilizando a interface Linear Discriminant Analysis- Variable Selection toolbox (ldavs_gui) (GOMES et al., 2013) escrito em ambiente Matlab® R2015a (Mathworks Inc., Natick, MA, USA). Os cálculos foram realizados utilizando um computador ACER®, Nitro 5, core i7 sétima geração.

Para os modelos de calibração foi utilizado o pacote PLS Toolbox, versão 6.0.1 da marca Eigenvector, escrito em ambiente MATLAB® R2015a (Mathworks Inc., Natick, MA, USA). Os cálculos foram realizados utilizando um computador ACER®, Nitro 5, core i7 sétima geração.

3.4 Modelos quimiométricos

A construção dos modelos de reconhecimento de padrões supervisionados foi realizada de acordo com o fluxograma da figura 1.

Figura 1. Fluxograma das etapas para a construção de um modelo de calibração e/ou para reconhecimento de padrão.



Fonte: Elaborado pela autora (2023)

Foram realizados dois tipos de pré-processamentos. No primeiro pré-processamento tomando como base o melhor pré-processamento utilizado por Mata e colaboradores (2022), a saber, a 1ª derivada com suavização utilizando o polinômio Savitizky-Golay de 2ª ordem e janela de 15 pontos (Mata e colaboradores, 2022). Como o tamanho da janela pode retirar informação do conjunto de dados testamos como segundo pré-processamento, foi realizado a aplicação da 1ª derivada com suavização utilizando o polinômio Savitizky-Golay de 2ª ordem e janela de 7 pontos.

A separação dos conjuntos de treinamento e teste foi realizada por meio do algoritmo SP-Xy (GOMES et al.2013) tanto para o modelo de calibração quanto o de reconhecimento de

padrão. Para os modelos de calibração multivariada foram selecionadas 20 amostras para calibração e 10 amostras para predição.

Já para os modelos de reconhecimento de padrões foram selecionadas 1593 amostras para o conjunto de treinamento (sendo 891 amostras convencionais e 702 amostras transgênicas) e 797 amostras para o conjunto de predição, constituído por 304 amostras convencionais e 493 amostras transgênicas). Para isto, utilizou-se como entrada para o algoritmo, uma matriz de dados contendo todas as 2390 amostras (convencionais e transgênicas). De modo que a quantidade de amostras convencionais (e transgênicas) que compõe a matriz de calibração e teste foi determinada pelo algoritmo SP-Xy com base nas distâncias entre as mesmas e respeitando a proporção de 70% de amostras para a etapa treinamento e 30% para a etapa de teste.

A matriz X (2390 linhas x 2800 colunas) foi composta pelos espectros de todas as amostras e o vetor Y (2390 linhas x 1 coluna) foi composto dos atributos relativos as classes, em que 1 designava as amostras convencionais e 0 as amostras transgênicas.

Para a construção dos modelos PLS para quantificação do teor água os seguintes parâmetros de entrada foram utilizados de no máximo 10 intervalos selecionados para os modelos iPLS e iSPA.

A verificação da habilidade preditiva dos modelos de calibração multivariada construídos foi avaliada em termos de parâmetros como RMSE, REP, R^2 e elipse de confiança.

Os modelos LDA utilizando os algoritmos de seleção estocásticos AOC e GA foram executados com 10 repetições e o determinístico, SPA, foi executado apenas uma vez.

A seleção de variáveis utilizando o algoritmo AOC e a construção dos modelos LDA foi executada utilizando os seguintes parâmetros: número de formigas na colônia = 50; número de colônias =100; taxa de formigas cegas em cada colônia e em cada ciclo = 0,35; taxa de evaporação do feromônio = 65%; e número máximo de variáveis = 20.

Já para o modelo GA-LDA foi executado utilizando os seguintes parâmetros: tamanho da população (POP) = 50, número de gerações = 100, probabilidade de cruzamento = 60%, máximo de variáveis = 20, probabilidade de mutação = 5%.

O modelo SPA-LDA foi executado com o tamanho da cadeia de variável que será gerada da etapa de projeções sendo definido o menor (*Nmi*n =1) e a maior (*Nmax =20*) quantidade de variáveis na cadeia.

A verificação da habilidade preditiva dos modelos de classificação construídos foi avaliada em termos de parâmetros como a taxa de classificação correta, sensibilidade e especificidade.

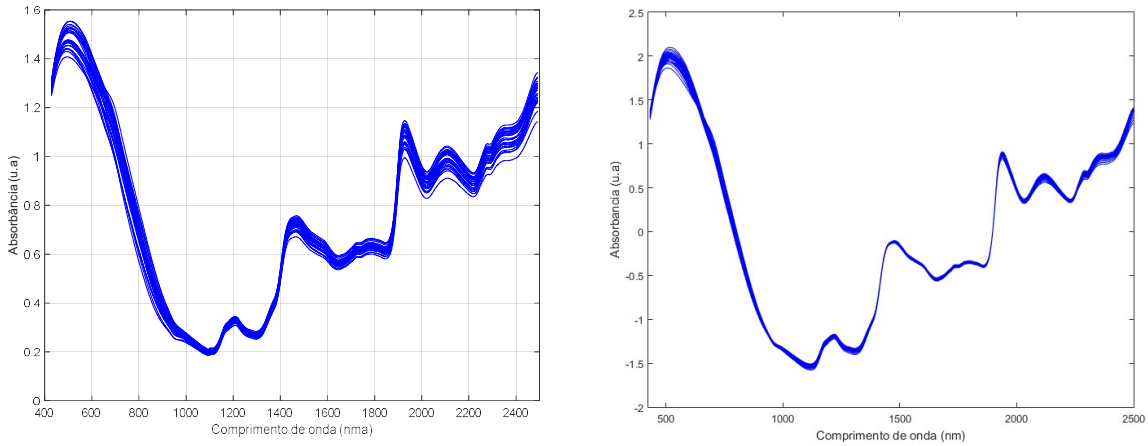
4 RESULTADOS E DISCUSSÕES

4.1. Modelo para a quantificação de teor de água em sementes de algodão

Os espectros são expressos na Figura 2a, que ilustram bandas características. A região espectral em torno de 2270 nm está relacionada com C-H estiramento + deformação CH para grupos de peptídeos; 1450 nm está relacionada com O-H estiramento + deformação do amido; 1483 estiramento N-H (2 m) 1º sobretom de amido, $CONH_2$. Mostram ainda, duas regiões em torno de 1390 e 1890 nm relacionadas ao grupo metil e grupos OH, respectivamente.

A construção do modelo multivariado destinado a determinação de umidade nas sementes, diversos pré-processamentos foram testados e o que apresentou o melhor resultado foi os dados suavizados utilizando o polinômio Savitizky-Golay de 2ª ordem e janela de 15 pontos combinado com SNV (Figura 2b). Após a realização do pré-processamento espectral o conjunto de dados foi dividido em conjunto de calibração e predição. Posteriormente, foram construídos modelos PLS, iSPA-PLS e i-PLS.

Figura 2: Espectros (a) brutos e (b) pre-processados



(a) (b)

O modelo i-PLS foi construído com 4 variáveis latentes que explicam 99,43% da matrix **X** (dados espectrais) e 74,87% do vetor y (valores de umidade). Após a análise do gráfico de Fcaculado/ Fcrítico por amostras (Figura 4), não foi verificada a presença de amostras anômalas.

Na Tabela 2 pode-se notar que tanto o modelo construído com todas as variáveis quanto aqueles construídos com os intervalos selecionados com o i-PLS ou iSPA-PLS mostram valores

de RMSE, REP adequados, representando menos que 10% do menor valor de umidade analisado. Os valores da raiz do erro quadrático médio (RMSE) seja na etapa de calibração, validação cruzada ou predição foram menores que o desvio padrão dos valores de referência (0,93%). Além disso, os valores de REP foram menores que a concentração mínima. No entanto, os modelos iPLS e iSPA-PLS apresentaram os menores valores de RMSEP e REP e os maiores coeficientes de determinação na etapa de predição. Os valores de parâmetro de desempenho obtidos neste trabalho indicam que o modelo i-PLS tem uma melhor performance que o modelo PLS construído com toda faixa espectral desenvolvido.

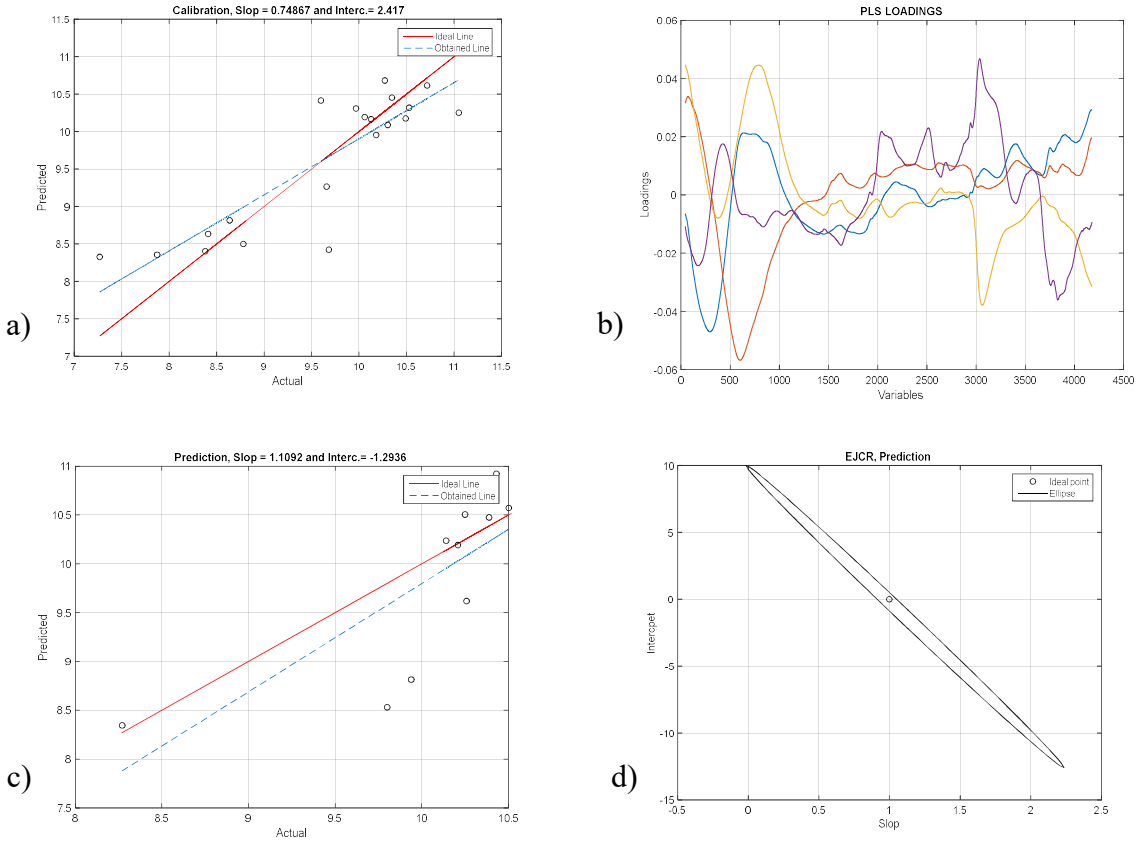
Tabela 2: Valores de parâmetro de desempenho para os modelos de calibração

		Calibração	**Validação cruzada**	**Predição**
PLS (VL=4)	**Número de amostras**	20	20	10
	RMSE(%)	0,5847	0,7340	0,6027
	R^2	0,7487	0,8579	0,5954
	r	0,8653	0,7179	0,7716
	REP(%)			6.2673
iSPA-PLS (VL=4) Nintervalos=7	**Número de amostras**	20	20	10
	RMSE(%)	0,5527	0,6245	0,2615
	R^2	0.77544	0,8243	0,8578
	r	0.88059	0,7942	0,9262
	REP(%)			2,7196
i-PLS (VL=4) Nintervalos=3	**Número de amostras**	20	20	10
	RMSE(%)	0,5325	0,6825	0,2093
	R^2	0,7915	0,8582	0,9137
	r	0,8897	0,7568	0,9559
	REP(%)			2,1763
	Faixa de concentrações (%)	7,27 – 11,05		
	Concentração media (%)	9,75		
	Desvio padrão (%)	0,93		

Nas Figura 3a e 3c observa-se o gráfico de umidade medida vs umidade prevista pelo modelo PLS, para as etapas de calibração e predição, respectivamente. Nele o segmento vermelho representa a regressão ideal e em azul obtida. As amostras de calibração se distribuem ao longo do segmento. No entanto não é observado um bom ajuste para as amostras de predição. Na figura 3b são apresentados o gráfico de loadings para as 3 primeiras VL, pode-se observar que as regiões próximas a 500 nm e entre 3000 e 4000nm são significativas para o modelo PLS. A Figura 3d ilustra a região conjunta da elipse de confiança (EJCR). Nela podemos observar

que o ponto ideal está dentro da elipse, indicando que, a um nível de confiança de 95%, não há bias significativo (tcal = 0.6564 e tcrit 1.8331) no conjunto de dados, com intersecção próximo a zero e inclinação próximo a 1.

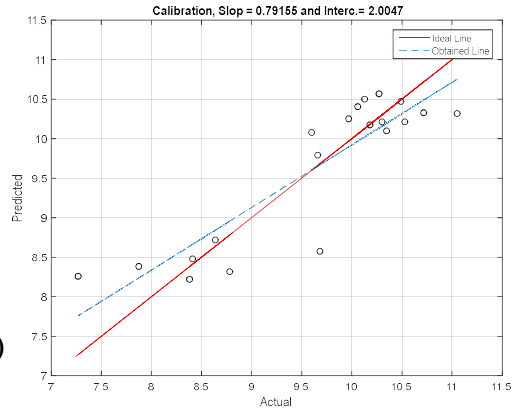
Figura 3: (a e c) Gráfico de umidade medida vs umidade prevista nas etapas de calibração e teste para o modelo PLS, (b) loadings de PC1, PC2 e PC3 (d) EJCR



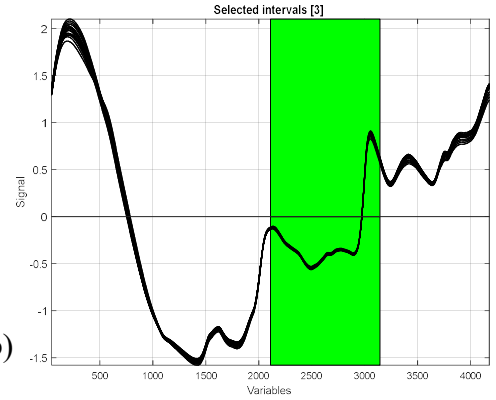
As figuras 4 e 5, mostram os gráficos de valores medidos vs preditos pelos modelos iPLS e iSPA-PLS nas etapas de calibração (4a e 5a) e predição (4c e 5c), bem como os intervalos selecionados (6b e 7b) e região conjunta da elipse de confiança (6d e 7d). Podemos notar um melhor ajuste das amostras de predição a reta de regressão para os modelos construídos com seleção de variáveis. Em relação a curva EJCR, observamos uma elipse estreita, o que evidencia o baixo desvio dos valores de umidade, como o ponto ideal localizado no centro da elipse, indicando a ausência de bias. O que evidencia a qualidade preditiva do modelo, em concordância com as figuras de mérito descritas na Tabela 2.

Figura 4: (a e c) Gráfico de umidade medida vs umidade prevista nas etapas de calibração e teste para o modelo iPLS, (b) intervalo selecionado e (d) EJCR

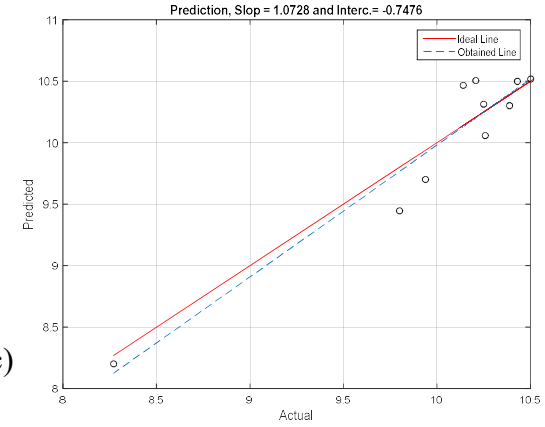
a)



b)



c)



d)

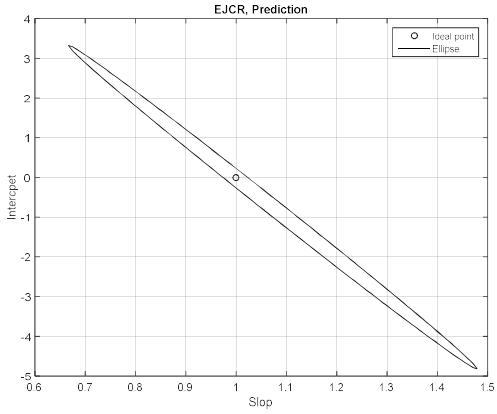
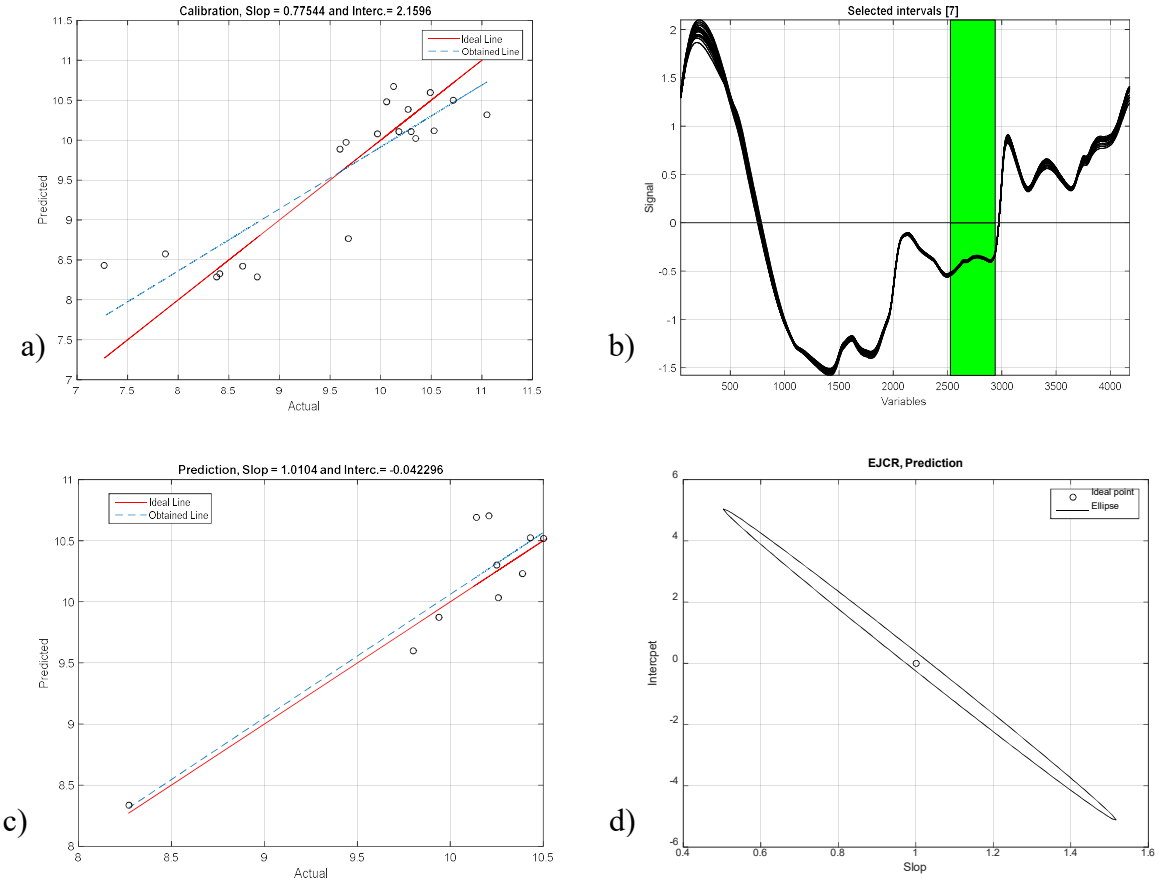


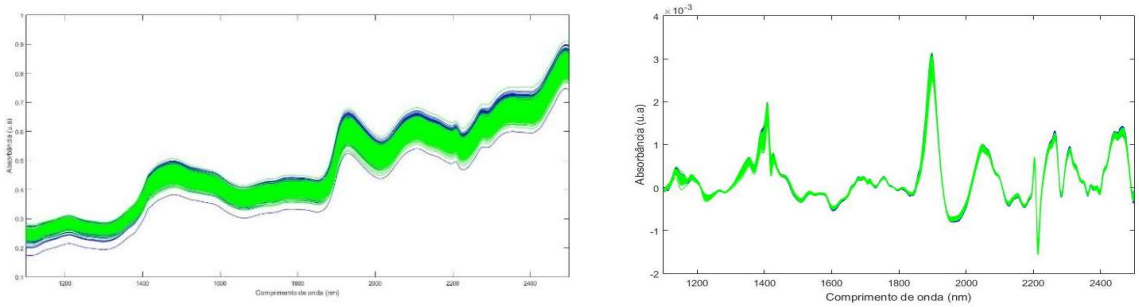
Figura 5: (a e c) Gráfico de umidade medida vs umidade prevista nas etapas de calibração e teste para o modelo iSPA-PLS, (b) intervalo selecionado e (d) EJCR



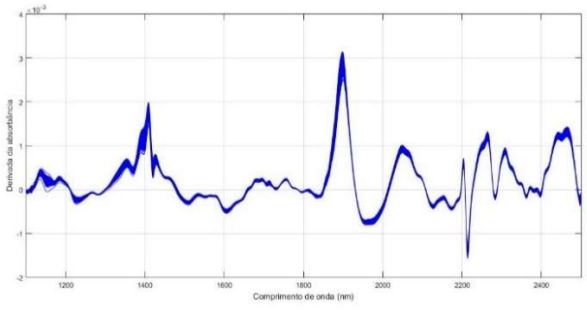
4.2. Modelos de reconhecimento de padrões para distinção de sementes transgênicas e convencionais.

A **Figura 6**, visualiza-se os espectros NIR brutos das 2390 amostras de sementes de algodão convencional e transgênicas, registrados na faixa de 1101 a 2500 nm, com 0,5 nm de intervalo, e também os espectros pré-processados. Nelas estão marcados os principais comprimentos de onda relacionados ao problema de estudo de acordo com Mata e colaboradores (2022). Podemos observar que não há diferenças perceptíveis entre os perfis espectrais de sementes de algodão convencional (verde) e transgênico (azul), o que inviabiliza a distinção entre os genótipos sem o uso de abordagens quimiométricas. No entanto, os espectros fornecem bandas importantes para classificação de sementes de algodão. Os espectros derivados mostram características que são relevantes para a discriminação do genótipo, como descrito na sessão 4.1. Além das bandas já discutidas podemos destacar as bandas em torno de 2051, 2264 e 2466 nm estão relacionadas a combinação amido O-H, estiramento C-O-O, 3º sobretom de proteína, que estão relacionadas as diferenças entre os genótipos.

Figura 6. Espectros **(a)** brutos e com 1ª. derivada com polinômio SG, janelas de **(b)** 7 e **(c)** 15 pontos.



(a) (b)



(c)

As **figuras 6b e 6c**, mostram os espectros derivativos após a aplicação da 1ª derivada com suavização utilizando o polinômio Savitizky-Golay de 2ª ordem e janela de 7 pontos (**Figura 6b**) e de 15 pontos ((**Figura 6c**). O pré-processamento utilizado neste trabalho foi o mesmo utilizado por Mata e colaboradores (2022). Além deste testou-se o uso de uma janela menor (7 pontos).

O pré-processamento dos espectros, os conjuntos de treinamento e teste foram selecionados utilizando o algoritmo SPXY (GOMES et al. 2013). Em seguida os modelos LDA foram construídos e validados utilizando cadeias de variáveis selecionadas baseadas nos parâmetros utilizados por cada algoritmo (SPA, ACO e GA). Tais modelos foram submetidos a classificação do conjunto de teste. A **Tabela 3** fornece os resultados das figuras de mérito ou parâmetros de validação (taxa de classificação correta ou acurácia, sensibilidade, especificidade e taxa de erro de classificação) para os modelos AOC-LDA, GA-LDA e SPA-LDA, bem como, a quantidade de variáveis selecionadas para construção dos modelos.

Os modelos AOC-LDA e GA-LDA por serem estocásticos foram realizados utilizando 10 repetições, deste modo é apresentado o melhor resultado obtido.

Com AOC-LDA e GA-LDA, o tamanho da janela não influenciou de forma significativa a taxa de classificação correta dos modelos. De forma que a diferença na taxa de classificação correta dos modelos construídos com os dados suavizados com janela de 7 ou 15 pontos foi menor que 1%. No entanto, os modelos construídos com a suavização utilizando a menor janela (7pontos) tendeu a selecionar uma variável a menos que o construído com os dados suavizados com a janela de 15 pontos, tanto para o AOC-LDA quanto para o GA-LDA. O que implica que o melhor modelo é o mais parcimonioso, ou seja, o que modifica menos os dados originais e é composto por um menor número de variáveis espectrais.

Em relação aos modelos SPA-LDA tanto o desempenho quanto o número de variáveis selecionadas diferiram em relação ao tamanho da janela de suavização utilizada, de forma que a maior taxa de classificação correta foi alcançada pelo modelo construído com os dados derivativos suavizados com janela de 15 pontos. No entanto, o modelo mencionado utiliza 10 variáveis espectrais enquanto o construído com a janela de 7 pontos, utiliza apenas 4. Considerando o princípio da parcimônia, a seguir serão mostrados detalhes apenas para os modelos construídos com os dados derivativos suavizados com janela de 7 pontos.

TABELA 3. Parâmetro de desempenho para o modelo PLS-DA.

Pré-processamento	Parâmetro	Classe 1: treinamento	Classe 2: treinamento	Classe 1: teste	Classe 2: teste
1ª - D. suavização SG, P-2º, j- 7p. (VL=7)	**ACC***	97,3%		97,7%	
	Erro	2,7%		2,3%	
	Sensibilidade	0,98	0,97	0,97	0,99
	Especificidade	0,97	0,98	0,99	0,97
	No. Variáveis	7			
1ª - D. suavização SG, P-2º, j- 15p (VL=7)	**ACC**	97,3%		97,7%	
	Erro	2,7%		2,3%	
	Sensibilidade	0,98	0,97	0,98	0,96
	Especificidade	0,97	0,98	0,96	0,98
	No. Variáveis	7			

*ACC = Acuraria ou taxa de classificação correta.

TABELA 4. Parâmetro de desempenho para os modelos LDA.

Pré-processamento	Modelos	Parâmetro	Classe 1: treinamento	Classe 2: treinamento	Classe 1: teste	Classe 2: teste
1ª - D. suavização SG, P-2º, j- 7p.	AOC-LDA (VL=7)	**ACC***	91,88%		96,60%	
		Erro	8,11%		3,44%	
		Precisão	0,93	0,91	0,95	0,98
		Sensibilidade	0,93	0,91	0,96	0.97
		Especificidade	0,91	0,93	0.97	0,96
		No. Variáveis	8			
1ª - D. suavização SG, P-2º, j- 15p	AOC-LDA (VL=7)	**ACC**	90,76%		97,36%	
		Erro	9,23%		2,63%	
		Precisão	0,92	0,90	0,96	0,98
		Sensibilidade	0,91	0,90	0,97	0,97
		Especificidade	0,90	0,91	0,97	0,97
		No. Variáveis	9			
1ª - D. suavização SG, P-2º, j- 7p.	GA-LDA (VL=7)	**ACC**	92,85%		96,97%	
		Erro	7,14%		3,02%	
		Precisão	0,94	0,91	0,95	0,98
		Sensibilidade	0,93	0,92	0,97	0,97
		Especificidade	0,92	0,93	0,97	0,97
		No variáveis	8			
1ª - D. suavização SG, P-2º, j- 15p.	GA-LDA (VL=7)	**ACC**	92,04%		96,93%	
		Erro	7,95%		3,06%	
		Precisão	0,94	0,90	0,95	0,98
		Sensibilidade	0,92	0,92	0,97	0,97
		Especificidade	0,92	0,92	0,97	0,97
		No. Variáveis	9			
1ª - D. suavização SG, P-2º, j- 7p.	SPA-LDA (VL=7)	**ACC**	82,81%		91,35%	
		Erro	17,18%		8,64%	
		Precisão	0,86	0,80	0,86	0,95
		Sensibilidade	0,84	0,82	0,91	0,91
		Especificidade	0,82	0,84	0,91	0,91
		No. Variáveis	4			
1ª - D. suavização SG, P-2º, j- 15p.	SPA-LDA (VL=7)	**ACC**	88,76%		95,16%	
		Erro	11,23%		4,83%	
		Precisão	0,91	0,86	0,96	0,96
		Sensibilidade	0,89	0,89	0,93	0,98
		Especificidade	0,89	0,89	0,98	0,93
		No. Variáveis	10			

*ACC = Acuraria ou taxa de classificação correta.

Os valores obtidos são comparáveis aqueles obtidos pelo modelo espectro completo construído por Mata e colaboradores (2022) que alcançou 97,77% de predição correta. No entanto o particionamento dos conjuntos de treinamento e teste utilizado neste trabalho foi diferente do utilizado pelos autores citados. Apesar de ambos os trabalhos utilizarem o SP-xy, no trabalho de Mata (2022) o algoritmo foi aplicado nas amostras convencionais e transgênicas separadamente. Já no presente trabalho as amostras convencionais e transgênicas estavam em uma única matriz. No entanto, para os dois casos o modelo PLS apresentou a mesma taxa de classificação correta no conjunto de teste (97,7%). Os modelos construídos com as variáveis selecionadas pelo GA e AOC obtiveram taxas de classificação correta equivalente as obtidas pelos modelos PLS. O modelo SPA-PLS-LDA obteve uma taxa de classificação correta inferior aos modelos mencionados (95,16%), porém aceitável. Este tipo de modelo tem a vantagem de ser determinístico, em contrapartida aos estocásticos GA e AOC.

Em relação ao tempo computacional, o modelo construído utilizando o SPA para selecionar variáveis consumiu aproximadamente 240 minutos, já aqueles que utilizaram o GA e AOC consumiram 10 e 30 minutos, respectivamente. Apesar de não ser informado no trabalho de Mata e colaboradores (2022), estima-se que o tempo computacional de modelos construídos com todo espectro e utilizando o PLS toobox®, como utilizado pelos autores mencionados, tem custo computacional de aproximadamente 5 minutos, em média.

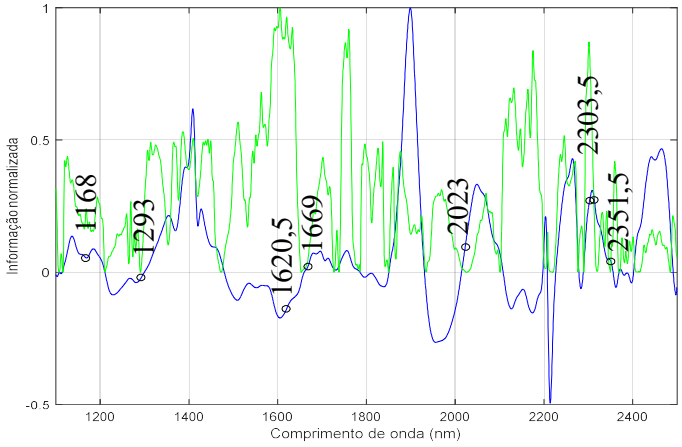
Das variáveis selecionadas listadas na **Tabela 5** e destacadas na **Figura 7**, a maioria das variáveis selecionadas pelo SPA estão relacionadas estiramento O-H e combinação de deformação H-O-H (~1930 nm), provavelmente devido a água e celulose contida nas sementes. Assim, como também foram selecionados comprimentos de onda relativos a $CONH_2$, especialmente devido à ligação de hidrogênio C=O com BH do peptídeo chamado estrutura alfa-hélice relacionada a modificação genética (Mata e colaboradores, 2022). O SPA, apesar de seu desempenho adequado, foi o modelo que obteve a menor taxa de classificação correta. O que deve estar relacionado ao número de variáveis espectrais envolvidas no problema estudado (RAPHAEL et al., 2003). A linha verde mostra o poder discriminatório para cada variável.

Tabela 5. Regiões selecionadas pelos modelos e atribuição de banda.

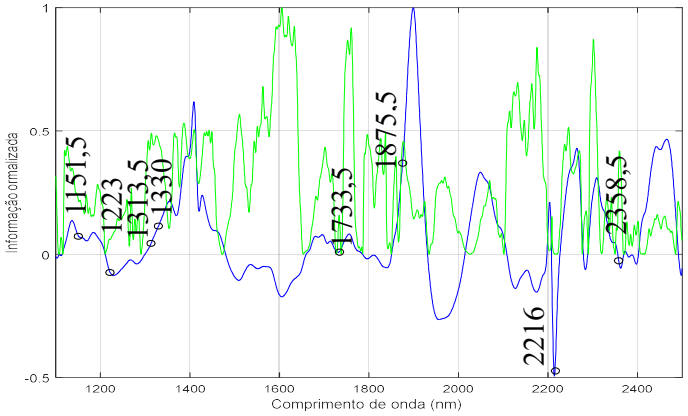
Comprimentos de onda selecionados (nm)			Atribuição das bandas
SPA	AOC	GA	
-	1168 1293	1151,5 1223 1313,5 1330	Segundo sobretom C-H do grupo metil.
-	1620 1669	1733,5	Sobretom de combinação C-H grupo metil, Primeiro sobretom estiramento N-H amido, primeiro sobretom amido O-H Primeiro sobretom C-H+deformação CH para grupos de peptídeos
1876 1890,5 2215	2023	1875,5 2216	Combinações de vibração O-H N-H Celulose e água, estiramento segundo sobretom óleo O-H e deformação de H-O-H, combinação $–CO_2R$, C=O. combinação de estiramento C-O amido. Estiramento C-O-O, terceiro sobretom proteína N-H, segundo sobretom de óleo C-H.
2302	2303,5 2314 2351,5	2310 2358,5	Combinação e estiramento amido C-H, estiramento e deformação CH_2 celulose C-H
2465	-	-	Refere-se a banda da amida I chamado de estrutura de hélice alfa (α-hélice), característica de moléculas de proteína.

Fonte: Baseado em Soares et al 2014; MATA et al., 2022.

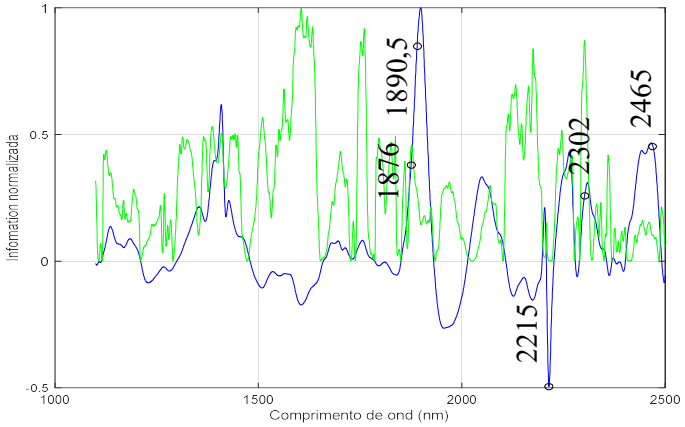
Figura 7. Variavéis selecionadas e poder descriminante para os modelos (a) AOC-LDA, (b) GA-LDA e (c) SPA-LDA.



(a)



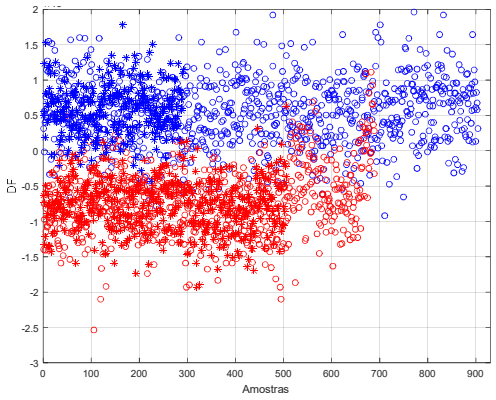
(b)



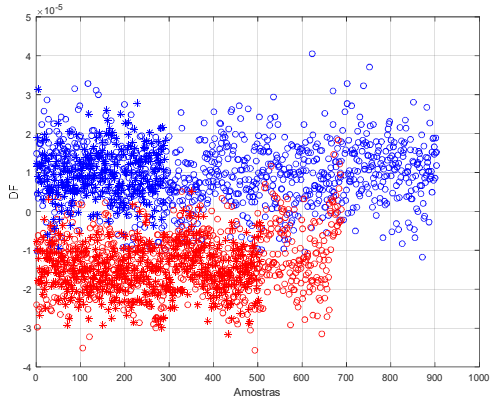
(c)

Na **Figura 8** podemos observar o gráfico de amostras por função discriminante (df), na qual a cor vermelha representa as amostras transgênicas e as azuis as convencionais. Os círculos representam a etapa de treinamento, enquanto os asteriscos representam a etapa de teste. Os modelos AOC-LDA (**figura 8a**) e GA-LDA (**figura 8b**) podemos perceber uma boa separação entre os dois grupos. Já para o modelo SPA-LDA podemos perceber algumas sobreposições, que estão relacionados a porcentagem de classificação incorreta.

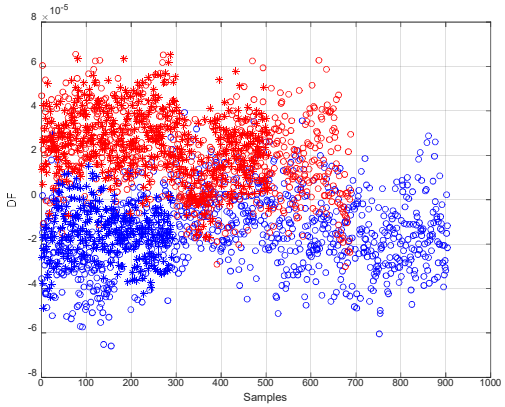
Figura 8. Gráfico das amostras por função descriminante para os modelos (a) AOC, (b) GA e (c) SPA-LDA.



(a)



(b)



(c)

A eficiência dos modelos LDA em distinguir entre as amostras convencionais e transgenicas foram superiores a 95% de classificação correta. Além disso, ressalta-se a eficácia dos algorítmos de seleção de variáveis em selecionar variáveis relacionadas com o problema analítico estudado. Neste sentido os algoritimos bioinspirados (AOC e GA) fornecem uma taxa de classificação correta superiores ao algorítmo determístico (SPA). Isto pode ter sido causado devido a natureza dos dados, nos quais uma foram obtidos de matrizes composta de composição complexa com ampla variabilidade de composição química. No entanto as variáveis seleciondas pelo algorítimo SPA tem mais relação química com o problema em estudo.

5 CONCLUSÃO

A construção de modelos de regressão por mínimos quadrados parciais, mostrou ser eficiente para a predição de teor de água em sementes de algodão. Os modelos construídos com todas as variáveis espectrais ou com os intervalos selecionados são estratégicas analíticas eficientes como prova de conceito. O modelo iPLS deu origem a um modelo mais robusto com valores de RMSE e REP concordantes nas etapas de validação cruzada e de predição. Para este modelo foram obtidos valores de REP de 2,17% na etapa predição.

A técnica de seleção de variáveis aliada a análise discriminante linear, quando aplicada aos dados de espectroscopia NIR, é eficaz para diferenciar sementes de algodão convencionais e transgênicas, com potencial de ampliação em escala de maturidade tecnológica para problemas de classificação, em se tratando das duas cultivares específicas. Todos os modelos da análise discriminante linear construídos com as variáveis selecionadas pelos algoritmos de seleção de variáveis testados, sejam determinísticos ou bioinspirados, forneceram taxas de classificação correta superiores a 95%, com sensibilidade e seletividade próximos a unidade. Sendo o modelo AOC-LDA, construído com os dados pré-processados com 1ª derivada com polinômio Savitzky-Golay de 1ª ordem e janela de 15 pontos, o que obteve a maior taxa de classificação correta (97,36%). A taxa de classificação correta obtida por este modelo foi comparável com a taxa de acertos (97,77%) obtida pelo modelo PLS construído com todo espectro. Apesar disso, o modelo construído com as variáveis selecionadas pode ser considerado mais parcimonioso. Ademais, as variáveis selecionadas podem ser utilizadas para o desenvolvimento de equipamentos portáteis e dedicados, que seriam mais vantajosos economicamente que os equipamentos de bancada.

PROPOSTAS FUTURAS

- Testar a capacidade de modelos de calibração multivariada construídos com as variáveis selecionadas pelos AOC, GA e SPA para a quantificação de teor de óleos, umidade, proteínas e ácidos oléico, linoléico, palmítico, palmitoléico e estearático em sementes de algodão.
- Avaliar a capacidade preditiva de modelos iSPA-PLS para a construção de modelos quantitativo para os parâmetros de qualidade citados acima.
- Aplicar modelos *one class,* como o DD-SIMCA, para distinção entre as sementes de algodão convencional e transgênica.
- Incluir mais cultivares de forma a ampliar o conjunto de dados em relação a novas classes, além das duas cultivares estudadas inicialmente.

REFERÊNCIAS

ABREU, R. E. L.; PAZ, J. E. M.; SILVA, A. C.; PONTES, M. J. C.; LEMOS S. G. Ethanol fuel adulteration with methanol assessed by cyclic voltammetry and Multivariate calibration. Fuel (Guildford) 156 (2015) 20-25.

ALLEGRINI, F., OLIVIERI, A. C. A new and efficient variable selection algorithm based on ant colony optimization. Applications to near infrared spectroscopy/partial least -squares analysis. Analytica Chimica Acta 699: (2011) 18-25.

ALMEIDA, P.B.A. Uso da espectroscopia NIR e calibração multivariada para prospecção de oleaginosas quanto as suas características de óleo e proteína. 2013. Dissertação de mestrado. Ciências agrarias, UEPB.

ALMEIDA, M. R.; CORREA, D. N.; ROCHA, W. F. C.; SCAFI, F. J. O.; POPPI, R. J**.** Discrimination between authentic and counterfeit banknotes using Raman spectroscopy and PLS-DA with uncertainty estimation. ***Microchemical Journal***, v.109, p.170–177, 2013

ALMEIDA, V.E.; FERNANDES, D.D.S.; DINIZ, P.H.G.D.; GOMES, A.A.; VERAS, G.; GALVÃO, R. K. H.; ARAUJO, M.C.U. Scores selection via Fisher's discriminant power in PCA-LDA to improve the classification of food data. **Food Chemistry**, v. 363, 2021.

ALVES, V. D. et al. Método de classificação com seleção meta-heurística de variáveis para identificar adição de leite bovino em caprino. 2022.

ALVES, J. C. L.; POPPI, R. J. Biodiesel content determination in diesel fuel blends using near infrared (NIR) spectroscopy and support vector machines (SVM). **Talanta**, v. 104, p. 155–161, 2013.

ALVES, L. R. A.; BARROS, G. S. C.; BACCHI, M. R. P. Produção e exportação de algodão: efeitos de choques de oferta e de demanda. **Revista Brasileira de Economia**, v. 62, n. 4, p. 381-405, 2008.

ANZANELLO, M. J.; ORTIZ, R. S.; LIMBERGERB, R. P.; MAYORGA, P. A multivariate-based wavenumber selection method for classifying medicines into authentic or counterfeit classes. **Journal of Pharmaceutical and Biomedical Analysis**, v. 83, p. 209–214, 2013.

ARAÚJO, M. C. U., SALDANHA, T. C. B., GALVÃO, R. K. H., YONEYAMA T., CHAME, H. VISANI, C., V. The successive projections algorithm for variable selection in spectroscopy multicomponent analysis. Chemometrics and Intelligent Laboratory Systems 57 (2001) 65-73.

ARAUZO-AZOFRA, A.; MOLINA-BAENA, J.; JIMÉNEZ-VÍLCHEZ, A.; LUQUERODRIGUEZ, M. Using Individual Feature Evaluation to Start Feature Subset Selection Methods for Classification. **ICAART - 9th International Conference on Agents and Artificial Intelligence**, p. 607–614, 2017.

ARLINDO, A. (2021). *Agregação de valor nas cadeias produtivas* https://www.embrapa.br/olhares-para-2030/agregacao-de-valor-nas-cadeias-zrodutivasagricolas/-/asset_publisher/SNN1QE9zUPS2/content/arlindo-de-azevedo-moura?inheritRedirect=true . Acessado em 10/10/2022.

BARBIN, D.F., MACIEL, L.F., BAZONI, C.H.V., RIBEIRO, M.S., CARVALHO, R.D.S., BISPO, E.S., MIRANDA, M.P.S., HIROOKA, E.Y., 2018. Classification and compositional characterization of different varieties of cocoa beans by near infrared spectroscopy and multivariate statistical analyses. J. Food Sci. Technol. 55, 2457–2466. https://doi.org/10.1007/s13197-018-3163-5.

BADARÓ, A.T., GARCIA-MARTIN, J.F., LÓPEZ-BARRERA, M. DEL C., BARBIN, D.F., ALVAREZMATEOS, P., 2020. Determination of pectin content in orange peels by near infrared hyperspectral imaging. Food Chem. 323, 126861. https://doi.org/https://doi.org/10.1016/j.foodchem.2020.126861

BASGALUPP, M. P. **Algoritmos genéticos para seleção de atributos em problemas de classificação de processos de negócio**. [s.l.] PUC-RS, 2007.

BAZONI, C. H., IDA, E. I., BARBIN, D. F., & KUROZAWA, L. E. (2017). Near-infrared spectroscopy as a rapid method for evaluation physicochemical changes of stored soybeans. **Journal of Stored Products Research**, Reino Unido, v. 73, p. 1-6, 2017.

BIANCOLILLO, A.; LILAND, K.; MÅGE, I.; NÆS, T.; BRO, R. Variable selection in multi-block regression. **Chemometrics and Intelligent Laboratory Systems**, Volume 156, 2016

BOLÓN-CANEDO, V.; ALONSO-BETANZOS, A. Ensembles for feature selection: A review and future trends. **Information Fusion**, v. 52, p. 1–12, 2019.

BOLÓN-CANEDO, V.; SÁNCHEZ-MAROÑO, N.; ALONSO-BETANZOS, A. **Feature Selection for High-Dimensional Data.** Springer ed. 2015.

BLUM, A. L.; LANGLEY, P. Selection of relevant features and examples in machine learning. **Artificial Intelligence**, v. 97, p. 245–271, 1997.

BRASIL. **Ministério da Agricultura e do Abastecimento**. Departamento Nacional de Defesa Vegetal. Regras para análise de sementes. Brasília, DF, 1992. 365 p.

BRAGA, J. W. B., SANTOS JUNIOR, A. A. D. MARTINS, I. S., "Determination of viscosity index in lubricant oils by infrared spectroscopy and PLSR," *Fuel,* vol. 120, pp. 171-178, 2014.

BRERETON, R. G.; J. CHEMOM. **2014**, *28*, 749. [Crossref]

CAI, W.; LI, Y.; SHAO, X. A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra. **Chemometrics and Intelligent Laboratory Systems**, v. 90, n. 2, p. 188–194, 2008

CAPORASO, N., WHITWORTH, M.B., FOWLER, M.S., FISK, I.D., 2018a. Hyperspectral imaging for non-destructive prediction of fermentation index, polyphenol content and antioxidant activity in single cocoa beans. Food Chem. 258, 343–351. https://doi.org/10.1016/j.foodchem.2018.03.039
CAPORASO, N., WHITWORTH, M.B., GREBBY, S., FISK, I.D., 2018b. Rapid prediction of single green coffee bean moisture and lipid content by hyperspectral imaging. J. Food Eng. 227, 18–29. https://doi.org/10.1016/j.jfoodeng.2018.01.009

CILIA, N. D.; DE STEFANO, C.; FONTANELLA, F.; SCOTTO DI FRECA, A. A ranking-based feature selection approach for handwritten character recognition. **Pattern Recognition Letters**, v. 121, p. 77–86, 2019
CONCEIÇÃO, F. R.; MOREIRA, A. N.; BINSFELD, P. C. Detecção e quantificação de organismos geneticamente modificados em alimentos e ingredientes alimentares. **Ciência Rural**, Santa Maria, v. 36, n. 1, p. 315-324, 2006.

COSTA, M. W., OLIVEIRA, A. A. M. (2022). Social life cycle assessment of feedstocks for biodiesel production in Brazil. **Renewable and Sustainable Energy Reviews**, 159, 112166, 2022.

COSTA FILHO, C.A., POPPI, R.J. Algoritmo genetico em química. 22: 45, 1999.

CORNEJO-BÁEZ, A. A.; PEÑA-RODRÍGUEZ, L. M.; ALVAREZ-ZAPATA, R.; VAZQUEZ- HERNÁNDEZ, M.; SÁNCHEZ-MEDINA, A. Chemometrics: a complementary tool to guide the isolation of pharmacologically active natural products. **Drug Discovery Today**, v. 25, n. 1, p. 27-37, 2020.

COZZOLINO, D.; ROBERTS, J. Applications and developments on the use of vibrational spectroscopy imaging for the analysis, monitoring and characterization of crops and plants. **Molecules**, v. 21, p.755, 2016

COZZOLINO, D., 2016. Authentication of Cereals and Cereal Products, Advances in Food Authenticity Testing. Elsevier Ltd. https://doi.org/10.1016/b978-0-08-100220-9.00016-3

CRUZ-TIRADO, J.P., FERNÁNDEZ PIERNA, J.A., ROGEZ, H., BARBIN, D.F., BAETEN, V., 2020. Authentication of cocoa (Theobroma cacao) bean hybrids by NIR-hyperspectral imaging and chemometrics. Food Control 118, 107445. https://doi.org/https://doi.org/10.1016/j.foodcont.2020.107445

DARWISH, A. Bio-inspired computing: Algorithms review, deep analysis, and the scope of applications. **Future Computing and Informatics Journal**, v. 3, n. 2, p. 231–246, 2018. doi:10.1016/j.fcij.2018.06.001.

DORIGO, M.; V.; MANIEZZO, A. Colorni. Ant System: Optimization by a Colony of Cooperating Agents. IEEE Transaction of Systems, Man, and Cybernetics-Part B: Cybernetics. 26: (1996) 29-41.

DORIGO, M., STÜTZLE, T. Otimização de colônias de formigas: visão geral e avanços recentes. In: Gendreau M., Potvin JY. (eds) Handbook of Metaheuristics. International Series in Operations Research & Management Science, vol. 272. Springer, Cham. 2019.

DORIGO, M. & BLUM, C. Ant Colony Optimization Theory: A Survey. **Theoretical Computer Science**, v. 344, n. 2-3, p. 243-278, 2005.

DUARTE, L.M., 2015. Modelagem multivariada para determinação de propriedades físicoquímicas de petróleo e para quantificação de edulcorantes artificiais em adoçantes de mesa. Diss. Mestr. em Química. Univ. Fed. Juiz Fora. 120.

EBRAHIMPOUR, M. K.; EFTEKHARI, M. Distributed feature selection: A hesitant fuzzy correlation concept for microarray high-dimensional datasets. **Chemometrics and Intelligent Laboratory Systems**, v. 173, p. 51–64, fev. 2018.

ENGEL, J.; GERRETZEN, J.; SZYMANSKA, E.; JANSEN, J. J.; DOWNEY, G.; BLANCHET, L.; BUYDENS, L. M. C. Breaking with trends in pre-processing? **TrAC – Trends in Analytical Chemistry**, v. 50, p. 96-106, Oct. 2013.

ESTEBAN, M.; ARIÑO-BLASCO, M.C.; DÍAZ-CRUZ, J. M. **Chemometrics in Electrochemistry**. Comprehensive Chemometrics, p. 1-31, 2020. Elsevier

Embrapa. (2017). *Cultura do algodão no cerrado.* (2th ed.). https://www.spo.cnptia.embrapa.br/conteudo?p_p_id=conteudoportlet_WAR_sistemasdeproducaolf6_1ga1ceportlet&p_p_lifecycle=0&p_p_state=normal&p_p_mode=view&p_p_col_id=column-1&p_p_col_count=1&p_r_p_-76293187_sistemaProducaoId=7718&p_r_p_-996514994_topicoId=7985 .

FACELI, KATTI; LORENA, ANA CAROLINA; GAMA, JOÃO; CARVALHO, A. C. P. DE L. F. DE. **Inteligência artificial: uma abordagem de aprendizado de máquina**. LTC ed. Rio de Janeiro, 2011.

FARIAS, O. R. (2017). **Qualidade de sementes de algodoeiro (*Gossypium* spp.) submetidas ao controle biológico**. 74 p. Dissertação (Mestrado em Agronomia) - Universidade Federal da Paraíba, Areia, PB.

FERREIRA, M. M. C. 2015. Quimiometria: Conceitos, Métodos e Aplicações. Campinas, SP: Editora da Unicamp.

FERREIRA. M. M. C. Laboratório de Quimiometria Teórica e Aplicada, Instituto de Química, Universidade Estadual de Campinas, 13083-970 Campinas – SP, Brasil. *Quim. Nova,* Vol. 45, No. 10, 1251-1264, 2022. http://dx.doi.org/10.21577/0100-4042.20170910

FERRÉ, J., BOQUÉ, R., FERNÁNDEZ-BAND, B., LARRECHI, M., RIUS, F. "Figures of merit in multivariate calibration. Determination of four pesticides in water by flow injection analysis and spectrophotometric detection," *Analytica Chimica Acta,* vol. 348, pp. 167-175, 1997.

FISHER, R. A. The use of multiple measurements in taxinomic problems, Annales Eugenics, 7:179, 1936.

FONTES, J. A. Abordagens de seleção de variáveis para classificação e regressão em dados espectrais para controle de qualidade, 2020. Dissertação de mestrado em Engenharia – UFRGS.

GAITÁN-JURADO, A. J., GARCÍA-MOLINA, M., PEÑA-RODRÍGUEZ, F., & ORTIZSOMOVILLA, V. Near infrared applications in the quality control of seed cotton. **Journal of Near Infrared Spectroscopy**, Reino Unido, v. 16, n. 4, p. 421-429, 2008.

GALVÃO, R. K. H.; SOARES, S. F. C.; MARTINS, M. N.; PIMENTEL, M. F.; ARAÚJO, M. C. U. Calibration transfer employing univariate correction and robust regression. Analytica Chimica Acta 864 (2015) 1-8.

GATIUS, F. et al. Comparison of CCA and PLS to explore and model NIR data. **Chemometrics and Intelligent Laboratory Systems**, v. 164, n. March, p. 76–82, 2017.

GOMES, A. A. Algoritmo das Projeções Sucessivas aplicado à seleção de variáveis em regressão PLS. 2012. Dissertação (Mestrado em Química) Curso de Pós-Graduação em Química, Universidade do Federal da Paraíba.

GOMES, A. A.; GALVÃO, R. K. H.; ARAÚJO, M. C. U.; VÉRAS, G.; SILVA, E. C. (2013) The successive projections algorithm for interval selection in PLS. M**icrochemical Journal**, v. 110, p. 202–208,. doi:10.1016/j.microc.2013.03.015

GOMES, A. A., AZCARATE, S. M., DINIZ, P. H. G. D., FERNANDES, D. D. S., VERAS, G. (2022). Variable selection in the chemometric treatment of food data: A tutorial review. Food Chemistry, 370, 131072. doi:10.1016/j.foodchem.2021.131072

GHASEMI, I. NIAZI, A., LEARDI, R. Genetic-algorithm bsed wavelength selection in multicomponent spectrophotometric determination by PLS: application on copper and zinc mixture. Talanta, 50:311, 2003.

GHOSH, J., SHUVO, SB (2019). *Melhorando o Desempenho do Modelo de Classificação Usando Análise Discriminante Linear em Dados Lineares. 2019 10ª Conferência Internacional sobre Tecnologias de Computação, Comunicação e Redes (ICCCNT).* doi:10.1109/icccnt45670.2019.8944632

GRUNVALD, A. K.; CARVALHO, C. G. P. de; LEITE, R. S.; MANDARINO, J. M. G.; ANDRADE, C. A. de B.; SCAPIM, C. A. Predicting the oil contents in sunflower genotype seeds using nearinfrared reflectance (NIR) spectroscopy. **Acta Scientiarum**. Agronomy, Maringá, v. 36, n. 2, p.233-237, 2014.

GUIMARÃES, G. C. Espectroscopia no infravermelho próximo para classificação de sementes de café quanto à qualidade, origem e cultivar. **Tese (doutorado)** – Universidade Federal de Lavras, Lavras. 2016.

GUTKOSKI, L. C.; TEIXEIRA, D. M. F.; DURIGON, A.; GANZER, A. G.; BERTOLIN, T. E.; COLLA, L. M. Influência dos teores de aveia e de gordura nas características tecnológicas e funcionais de bolo. **Ciência e Tecnologia de Alimentos**, Campinas, v. 29, n. 2, p. 254-261, 2009. http://dx.doi.org/10.1590/S0101-20612009000200003

HANSEN, L. Análise de Dados Físico-Químicos de Amostras de Leite Cru do Sul do Brasil Utilizando Métodos Multivariados Exploratórios e Classificatórios. 2019. Dissertação (Mestrado em Química) - Curso de Pós-Graduação em Química, Universidade Federal do Rio Grande do Sul.

HART, J. H.; NORRIS, K. H. Direct spectrophotometric determination of moisture content of grain and seeds. In: WEXLER, A. (Ed.). **Humidity and moisture, principles and methods of measuring moisture in liquid and solids**. New York: Reinhold Publishing Co., 1965. v. 4, p. 19-25.

HAIR, J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E.; TATHAM, R. L. Análise multivariada de dados. Editora Bookman, 6ª ed., Porto Alegre, 682 p., 2009.

HEMMATEENEJAD, B.; SHAMSIPUR M.; ZARE-SHAHABADI V. & AKHOND, M. Building Optimal Regression Tree by Ant Colony System–Genetic Algorithm: Application to Modeling of Melting Points. **Analytica Chimica Acta,** v. 704, n. 1–2, p. 57-62, 2011.

Hibbert, D. B.; *Pure Appl. Chem.* **2016**, *88*, 407.

HOLLER, F. J.; SKOOG, D. A.; CROUCH, S. R. **Princípios de análise instrumental**. 6. ed. Porto Alegre: Bookman, p. 1055, 2009

HUANG, X.; LUO, Y.-P.; XIA, L. An efficient wavelength selection method based on the maximal information coefficient for multivariate spectral calibration. **Chemometrics and Intelligent Laboratory Systems**, v. 194, p. 103872, nov. 2019.

ISAAA: Serviço Internacional para a Aquisição de Aplicações de Agrobiotecnologia.

JIN, L. et al. A ReliefF-SVM-based method for marking dopamine-based disease characteristics: A study on SWEDD and Parkinson's disease. **Behavioural Brain Research**, v. 356, n. September 2018, p. 400–407, jan. 2019.

KAHMANN, A.; ANZANELLO, M. J.; MARCELO, M. C. A.; POZEBON, D. Near infrared spectroscopy and element concentration analysis for assessing yerba mate (Ilex paraguariensis) samples according to the country of origin. **Computers and Electronics in Agriculture**, v. 140, p. 348–360, 2017.

KAUFMANN, K.C., FAVERO, F. DE F., DE VASCONCELOS, M.A.M., GODOY, H.T., SAMPAIO, K.A., BARBIN, D.F., 2019. Portable NIR Spectrometer for Prediction of Palm Oil Acidity. J. Food Sci. 84, 406–411. https://doi.org/10.1111/1750-3841.14467.

KEMSLEY, E.K., DEFERNEZ, M., MARINI, F., 2019. Multivariate statistics: Considerations and confidences in food authenticity problems. Food Control 105, 102–112. https://doi.org/10.1016/j.foodcont.2019.05.021.

KAUR, B., SANGHA, M.K., KAUR, G., 2017. Development of Near-Infrared Reflectance Spectroscopy (NIRS) Calibration Model for Estimation of Oil Content in Brassica juncea and Brassica napus. Food Anal. Methods 227–233. https://doi.org/10.1007/s12161-016-0572-9.

KAUR, B., SANGHA, M.K., KAUR, G., 2016. Calibration of NIRS for the Estimation of Fatty Acids in Brassica Juncea. J. Am. Oil Chem. Soc. 93, 673–680. https://doi.org/10.1007/s11746-016-2802-0

KENNARD, RW E STONE, LA, 1969. Projeto de experimentos auxiliado por computador. Technometrics 11, 137-148

KOHAVI, R.; JOHN, G. H. Wrappers for feature subset selection. **Artificial Intelligence**, v. 97, n. 1–2, p. 273–324, 1997.

LAUXEN, L. R.; VILLELA, F. A.; SOARES, R. C. Desempenho tratadas com tiametoxan. **Revista Brasileira de Sementes**. v. 32, n. 3 p. 061-068, 2010.

LEARDI, R. Application of genetic algorithm-PLS for feature selection in spectral data sets. **Journal of Chemometrics**, v. 14, n. 5–6, p. 643–655, 2000.
LEE, C.-Y.; CAI, J.-Y. LASSO variable selection in data envelopment analysis with small datasets. **Omega**, dez. 2018.

LIU, H.-Y., WADOOD, S. A., XIA, Y., LIU, Y., GUO, H., GUO, B.-L., & GAN, R.-Y. (2021). *Wheat authentication ： An overview on different techniques and chemometric methods. Critical Reviews in Food Science and Nutrition, 1–24.* doi:10.1080/10408398.2021.1942783

LIU, H.; YU, L. Toward integrating feature selection algorithms for classification and clustering. **IEEE Transactions on Knowledge and Data Engineering**, v. 17, n. 4, p. 491–502, 2005.

LINGARA, J. K; SRIKANTHA, N.; MOINUDDIN. K. & LOKESH K.S Swarm Intelligence Based Ant Colony Optimization (ACO) Approach for Maximizing the Lifetime of Heterogeneous Wireless Sensor Networks. **Journal of Engineering Research and Applications,** v. 3, n. 6, p. 167- 172, 2013.

LOPES, E.L.; *Tese: Estratégias para a caracterização de amostras brasileiras de óleo de soja empregando ATR-FTIR e HPLC-ELSD aliadas a ferramentas quimiométricas*. Universidade Federal do Rio Grande do Sul: 2015.

LUCASIUS, C. B., KATEMAN, G. Understanding and using genetic algorithms part 1. Concepts, properties and context. Chemon.int.lab.sist. 19:1, 1993.

LUNA, Aderval S.; GOIS, Jefferson S. de. **Application of Chemometric Methods Coupled with Vibrational Spectroscopy for the Discrimination of Plant Cultivars**

and to Predict Physicochemical Properties Using R. Vibrational Spectroscopy for Plant Varieties And Cultivars Characterization, p. 165-194, 2018. Elsevier.

MAHDAVI, V., MONAJEMI, A. (2014). Optimization of operational conditions for biodiesel production from cottonseed oil on CaO–MgO/Al2O3 solid base catalysts. **Journal of the Taiwan Institute of Chemical Engineers**, 45(5), 2286-2292.

MARCOS FILHO, J. **Fisiologia de sementes de plantas cultivadas**. 2ª edição. Londrina: ABRATES, 659p, 2015.

MAYRINCK, L. G. Uso da espectrometria no infravermelho próximo na avaliação da qualidade fisiológica de sementes de algodão. **Dissertação (mestrado)** – Universidade Federal de Lavras, Lavras. 2018.

MARINI, F. Classification methods in chemometrics. **Current Analytical Chemistry**, v. 6, n. 1, p. 72–79, 2010

MARTINS, M. N.; GALVÃO R. K. H.; PIMENTEL M. F. Multivariate Calibration Transfer Employing Variable Selection and Subagging. Journal of the Brazilian Chemical Society (J. Braz. Chem. Soc.) 21 (2010) 127-134.

MARTENS, H., MARTENS, M. Modied Jack-Knife estenation of parameter uncertaenty in belinear modelling by partial least squares regression (PLSR). Food Quality and preference. 11:5, 2000.

MARQUETTI, I., LINK, J. V., LEMES, A. L. G., SCHOLZ, M. B. D. S., VALDERRAMA, P., BONA, E. **Arabica coffee classification using near infrared spectroscopy and twostage models**. IX Simpósio de Pesquisa dos Cafés do Brasil 24 a 26 de junho de 2015, Curitiba – PR.

MATA, M.M.D.; ROCHA, P.D.; FARIAS, I.K.T.D.; SILVA, J.L.B.D.; MEDEIROS, E.P.; SILVA, C.S.; SIMÕES, S.D.S. Distinguishing cotton seed genotypes by means of vibrational spectroscopic methods (nir and raman) and chemometrics. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2022**, *266*, 120399.

MEDEIROS, M. L. S., 1995-M467a. Aplicação de espectroscopia no infravermelho próximo e imagens hiperespectrais para quantificar o teor de óleo e classificar sementes do gênero *Brassica* / Maria Lucimar da Silva Medeiros. – Campinas, SP: [s.n.], 2021.

MENEZES, I. (2009). Caracterização in situ e diversidade genética de algodoeiros mocós (Gossypiumhirsutum raça marie galante) da região nordeste do Brasil. 2009. 91f. Dissertação (Mestrado em Genética e Biologia Molecular) - Universidade Federal do Rio Grande do Norte, Natal.

MITTELMANN, A. et al. Análise dialélica do teor de óleo em milho. **Revista Brasileira de Agrociência**, Pelotas, v. 12, n. 2, p. 139-143, abr./jun. 2006.

MOREIRA, A. C. O. Espectroscopia NIR, CG-EM e quimiometria para o controle de qualidade do óleo de copaíba (Copaifera spp.). 2018.

MUJELI, M., KEFAS, H. M., SHITU, A., AYUBA, I. (2016). Optimization of Biodiesel Production from Crude Cotton Seed Oil Using Central Composite Design. **American Journal of Chemical and Biochemical Engineering**, 1(1), 8-14.

MURSALIN, M. et al. Automated epileptic seizure detection using improved correlationbased feature selection with random forest classifier. **Neurocomputing**, v. 241, p. 204– 214, jun. 2017

MUÑOZ-ROMERO, S. et al. Informative variable identifier: Expanding interpretability in feature selection. **Pattern Recognition**, v. 98, p. 107077, fev. 2020.

MULLHEN, R. J.; MONEKOSSO, D.; BARMAN, S & REMAGNINO, P. A Review of Ant Algorithms. **Expert SystemsWith Applications**, v. 36, n. 6, p. 9608-9617, 2009.

NACHE, M.; SCHEIER, R.; SCHMIDT, H. & HITZMANN, B. Non-invasive Lactateand pH-monitoring in Porcine Meat Using Raman Spectroscopy and Chemometrics. **Chemometrics and Intelligent Laboratory Systems**, v. 142, p. 197-205, 2015

NAES, T.; MEVIK, B. H., Understandig the collinearity problem in regression and classification, *Journal of Chemometrics*, **15:413, 2001**

NASCIMENTO, C.L.M.M. Valor nutricional e energético do farelo de algodão de alta energia em rações para suínos. 2009. 64f. Dissertação (Mestrado em Zootecnia)- Pós graduação em Zootecnia, Universidade Federal Rural de Pernambuco.

NERGAARD, L., SAUDLAND, A., WAGNER, J., NIELSEN, J.P., MUNCK, L., ENGELSEN, S.B., "Interval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy" **Applied Spectroscopy**, 54 (3), 413-419, 2000

NETO, B. B., SCARMÍNIO, I. S., BRUNS, R. E., "25 anos de quimiometria no Brasil", 124 Quim. Nova, vol. 29, p. 1401–1406, 2006.

NÓBREGA, R. O. Classificação de cafés solúveis usando espectroscopia NIR e quimiometria. Dissertação (Mestrado) - UFPB/CCEN. João Pessoa, 2021

NORGAARD, L. et al. Interval partial least squares regression (iPLS): A comparative chemometric study with na example from Near-infrared spectroscopy. Applied spectroscopy 54: 413,2000.

NOLASCO, I.M.P., 2019. Caracterização e Identificação de carne de aves por métodos não-destrutivos. Universidade Estadual de Campinas.

NUNES, P. G. A. **Uma nova técnica para seleção de variáveis em calibração multivariada aplicada às espectrometrias UV-VIS e NIR**. Tese (Doutorado em Química) - Universidade Federal da Paraíba, João Pessoa, 2008. 106p.

NUNES, R.T.C.; PRADO, T.R.; RIBEIRO, E.B.; VALE, W.S.; MORAIS, O.M. Desempenho fisiológico de sementes de algodão cultivadas em Luís Eduardo Magalhães, Bahia. **Revista Verde**, v. 10. n. 4, p. 69 - 74, 2015.

ORRILLO, I., CRUZ-TIRADO, J.P., CARDENAS, A., ORUNA, M., CARNERO, A., BARBIN, D.F., SICHE, R., 2019. Hyperspectral imaging as a powerful tool for identification of papaya seeds in black pepper. Food Control 101, 45–52. https://doi.org/10.1016/j.foodcont.2019.02.036.

OLIVEIRA, J. S. B.; SCHWAN-ESTRADA, K. R. F.; BONATO, C. M.; CARNEIRO, S. M. T. P. G. Homeopatias de óleos essenciais sobre a germinação de esporos e indução de fitoalexinas. **Revista Ciência Agronômica**, v. 48, n. 1, p. 208, 2017.

OLIVEIRA, M.M., CRUZ-TIRADO, J.P., ROQUE, J. V, TEÓFILO, R.F., BARBIN, D.F., 2020. Portable near-infrared spectroscopy for rapid authentication of adulterated páprica powder. J. Food Compos. Anal. 87, 103403. https://doi.org/https://doi.org/10.1016/j.jfca.2019.103403

OLIVEIRA, D. L. B.; PEREIRA, L. H. S.; SCHNEIDER, M. P.; SILVA, Y. J. A. B.; NASCIMENTO, C.W.A.; van STRAATEN, P.; SILVA, Y. J. A. B.; GOMES, A. A.; VÉRAS, G. Bio-inspired algorithm for variable selection in i-PLSR to determine physical properties, thorium and rare earth elements in soils from Brazilian semiarid region. **Microchemical Journal**, v. 160, Part A, p. 1-7, 2021. doi:10.1016/j.microc.2020.105640

OLIVEIRA, V. S. Análise da autenticidade de cédulas de real utilizando espectroscopia nir portátil e quimiometria. Mestrado (Dissertação) – Universidade Federal de Pernambuco. CCEN. Química Fundamental. Recife, 2018. – 2018. 44 f. : fig.

OZAKI, YUKIHIRO; HUCK, CHRISTIAN; TSUCHIKAWA, SATORU; BALLING NGELSEN, SØREN (EDS.). **Near-infrared spectroscopy: theory, spectral analysis, instrumentation, and applications**. Singapore: Springer Nature, 2021.

PARMEZAN, A. R. S.; LEE, H. D.; SPOLAÔR, N.; CHUNG, W. F. **Avaliação de Métodos para Seleção de Atributos Importantes para Aprendizado de Máquina Supervisionado no Processo de Mineração de Dados**. Foz do Iguaçu, 2012.

PASQUINI, C. Near infrared spectroscopy: fundamentals, practical aspects and analytical applications. **Journal of the Brazilian Chemical Society**, v. 14, n. 2, p. 198-219, 2003.

PASQUINI, C. Near infrared spectroscopy: fundamentals, practical aspects and analytical applications. **Journal of the Brazilian Chemical Society**, v. 14, n. 2, p. 198-219, 2013.

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

PEREIRA, A. F. C., et al.NIR spectrometric determination of qualit parameters in vegetable oils using iPLS and Variable selection. Food Resconch international. 41:341,2008.

PEREIRA, W. A. Calibração multivariada de misturas de óleos vegetais utilizando espectroscopia no infravermelho médio. Dissertação de mestrado em ciências agrárias. Universidade Estadual da Paraíba, 2012. 76 f. : il. Color

PES, B. Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains. **Neural Computing and Applications**, 2019.

PESSOA, C. M. Aperfeiçoamento do Algoritmo Colônia de Formigas para o Desenvolvimento de Modelos Quimiométricos. Dissertação de Mestrado. Universidade Federal do Rio Grande do Sul (UFRGS). Porto alegre. p. 131. 2015.

PONTES, M.J.C., GALVÃO, R.K.H., ARAÚJO, M.C.U., MOREIRA, P.N.T., NETO, O.D.P., JOSÉ, G.E., SALDANHA, T.C.B. The successive projections algorithm for spectral 67ariable selection in classification problems. Chemometrics and Intelligent Laboratory Systems 78: (2005) 11-18.

PONTES, M.J.C., et al. determination the quality of insulating oils using Near infrared spectroscopy and wavelength selection. Michochemical journal 98:254, 2011.

PONTES A. S.; ARAÚJO A.; MARINHO W.; DINIZ, P.H.G.D.; GOMES, A.A.; GOICOECHEA, H.C; SILVA, E.C.; ARAUJO, M. C. Otimização de colônias de formigas para seleção de variáveis em análise linear discriminante. Jornal de Quimiometria. 2020; e3292. https://doi.org/10.1002/cem.3292

PONTES, M.J.C.; PEREIRA, C.F.; PIMENTEL, M.F.; VASCONCELOS, F.V.C.; SILVA, A.G.B. Screening analysis to detect adulteration in diesel/biodiesel blends using near infrared spectrometry and multivariate classification, Talanta, 85 (2011) 2159.

QASIM, O. S.; ALGAMAL, Z. Y. Feature selection using particle swarm optimizationbased logistic regression model. **Chemometrics and Intelligent Laboratory Systems**, v. 182, p. 41–46, nov. 2018.

QU, Y. et al. Non-unique decision differential entropy-based feature selection. **Neurocomputing**, jul. 2019.

RADY, A., ADEDEJI, A., 2018. Assessing different processed meats for adulterants using visible-near-infrared spectroscopy. Meat Sci. 136, 59–67. https://doi.org/10.1016/j.meatsci.2017.10.014

RAHMAN, A.; CHO, B. Assessment of seed quality using non-destructive measurement techniques: a review. **Seed Science and Research**, v.26, n.4, p.285-305, 2016.

RAPHAEL, B.; SMITH, I. F. C. A direct stochastic algorithm for global search. **Applied Mathematics and computation**, v. 146, n. 2-3, p. 729-758, 2003. doi:10.1016/s0096-3003(02)00629-x

RANZAN, L. Metodologias para seleção de variáveis explicativas e detecção de inconformidades de predição aplicados à espectroscopia por fluorescência. Tese de doutorado. UFRGS, 2021.

RANZAN, C.; STROHM, A.; RANZAN, L.; TRIERWEILER, L. F.; HITZMANN, B & TRIERWEILER, J. O. Wheat Flour Characterization Using NIR and Spectral Filter Based on Ant Colony Optimization. **Chemometrics and Intelligent Laboratory Systems**, v. 32, p.133-140, 2014.

RAZZAQ, L., ABBAS, M. M., MIRAN, S., ASGHAR, S., NAWAZ, S., SOUDAGAR, M. E. M., SHAUKAT, N., VEZA, I., KHALIL, S., ABDELRAHMAN, A., KALAM, M. A. (2022). Response Surface Methodology and Artificial Neural Networks-Based Yield Optimization of Biodiesel Sourced from Mixture of Palm and Cotton Seed Oil. **Sustainability**, 14(10), 6130.

RAVAZZI, C.G., 2019. Identificação e quantificação de adulterantes de Whey protein concentrado empregando espectroscopia no infravermelho próximo e resolução multivariada de curvas, in: Dissertção de Mestrado (Química Analítica). Universidade Estadual de Campinas. p. 75

REMESEIRO, B.; BOLON-CANEDO, V. A review of feature selection methods in medical applications. **Computers in Biology and Medicine**, v. 112, p. 103375, 2019.

RIU, J. RIUS, F. X. "Method comparison using regression with uncertainties in both axes," *trends in analytical chemistry,* vol. 16, nº 4, pp. 211-216, 1997.

RIBEIRO, M. H. D. M.; COELHO, L. S. Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. **Applied Soft Computing Journal**, v. 86, p. 105837, 2020

ROCHA, P.D.; MEDEIROS, E. P.; SILVA, C.S.; SIMÕES, S.S. Estratégias quimiométricas para análise de imagens hiperespectrais no infravermelho próximo: classificação de genótipos de sementes de algodão, anal, metanfetamina, 2021.

RODRIGUEZ-GALIANO, V. F. et al. Feature selection approaches for predictive modelling of groundwater nitrate pollution: An evaluation of filters, embedded and wrapper methods. **Science of The Total Environment**, v. 624, p. 661–672, maio 2018.

RODIONOVA, O. Y.; POMERANTSEV, A. L. Chemometric tools for food fraud detection: The role of target class in non-targeted analysis. Food Chemistry, v. 317, p. 126448, 2020.

SABANDO, M. V.; PONZONI, I.; SOTO, A. J. Neural-based approaches to overcome feature selection and applicability domain in drug-related property prediction. **Applied Soft Computing Journal**, v. 85, 2019.

SAEYS, W., NGUYEN DO TRONG, N., VAN BEERS, R., NICOLAÏ, B.M., 2019. Multivariate calibration of spectroscopic sensors for postharvest quality evaluation: A review. Postharvest Biol. Technol. 158, 110981. https://doi.org/10.1016/j.postharvbio.2019.110981

SALIMI, A.; ZIAII, M.; AMIRI, A.; et al. Using a Feature Subset Selection method and Support Vector Machine to address curse of dimensionality and redundancy in Hyperion hyperspectral data classification. **Egyptian Journal of Remote Sensing and Space Science**, v. 21, n. 1, p. 27–36, 2018.

SANTOS JUNIOR, A. A. d. *Determinação de parâmetros de óleos lubrificantes de motores do ciclo Otto e ciclo Diesel através da espectroscopia no infravermelho, métodos multivariados e cartas de controle,* Brasília: Universidade de Brasília - UnB, 2011.

SAVITZKY, A., GOLAY, M. J. E. Smoothing and differentiation of data by simplified least-squares procedures, Analytical Chemistry, 8 (1964)1627.

SENA, M. M.; ALMEIDA, M. R.; BRAGA, J. W. B.; POPPI, R. J. Multivariate statistical analysis and chemometrics. In: FRANCA, A. S.; NOLLET, L. M. L. (Org.). **Spectroscopic methods in food analysis**. Boca Raton: CRC Press, 2017. p. 273-314.

SENA, M. M.; ALMEIDA, M. R. Quimiometria aplicada aos dados espectrais no Infravermelho Próximo. TIBOLA, C. S. et al. (Org.). In: **Espectroscopia no Infravermelho Próximo para Avaliar Indicadores de Qualidade Tecnológica e Contaminantes em Grãos**. Brasília, DF: Embrapa, cap. 2. p. 31–50. 2018.

SEVERINO, L., RODRIGUES, S., CHITARRA, L., LIMA FILHO, J., MOTA, E., MARRA, R., & ARAÚJO, A. (2019). Produto: ALGODÃO - Parte 01: Caracterização e Desafios Tecnológicos. *Embrapa*, 29.

SHRESTHA, S., KNAPIČ, M., ŽIBRAT, U., DELEURAN, L. C., GISLUM, R. Single seednear-infrared hyperspectral imaging in determining tomato (Solanum lycopersicum L.) seed quality in association with multivariate data analysis. **Sensors and Actuators B: Chemical**, Holanda, v. 237, p. 1027-1034, 2016.

SILVA, A. C.; SOARES, S. F.; INSAUSTI, C. M.; GALVÃO, R. K. H.; BAND, B. S.; ARAÚJO, M. C. U. Two-dimensional linear discriminant analysis for classification of three-way chemical data. Analytica Chimica Acta 938 (2016) 53-62.

SILVA, W. R.; PAULA, L. C. M.; SOARES, A. S.; COELHO, C. J. Algoritmo do morcego para seleção de variáveis em calibração multivariada. Revista Arithmós, Escola de Ciências Exatas e da Computação da Pontifícia Universidade Católica de Goiás, Goiânia, v.1, p. 13-17, 2019.

SILVA, S.F. Uso da espectrometria no infravermelho próximo na avaliação de qualidade de sementes de braquiárias. 46f. dissertação de mestrado, UFCE, 2019.

SKOOG, D. A., HOLLER, F. J., & CROUCH, S. R**. Principles of instrumental analysis. Cengage learning,** Seventh edition, Boston, 2017.

SIMEONE, M. L. F., PIMENTEL, M. A. G., NETO, M. M.G., PAES, M.C.D., SILVA, D.D. Uso da espectroscopia no infravermelho próximo e calibração multivariada para avaliar a composição química do milho. Cap.3. EMBRAPA,2018.

SIMEONE, M.L.F., RIBEIRO, M.R., TRINDADE, R.S. Espectroscopia no infravermelho próximo e analise discriminante por mínimos quadrados parciais como método alternativo para a seleção de sementes haploides de milho. EMBRAPA MILHO E SORGO, 2019.

SOCHA, K. & DORIGO, M. Ant Colony Optimization for Continuous Domains. **European Journal of Operational Research**, v. 185, n. 3, p. 1155-1173, 2008

SOARES, S. F. C., GALVÃO, R. K. H., PONTES, M. J. C., ARAÚJO, M. C. U. A new validation criterion for guiding the selection of variables by the successive projections algorithm in classification problems, Journal of the Brazilian Chemocal Society. 25 (2014) 176-181.

SOARES, F. Redução de dimensionalidade para dados espectrais colineares. 2022. Tese de doutorado, UFRGS.

SOARES, S. F. C., MEDEIROS, E. P., PASQUINI, C., DE LELIS MORELLO, C., GALVÃO, R. K. H., & ARAÚJO, M. C. U. Classification of individual cotton seeds withrespect to variety using near-infrared hyperspectral imaging. **Analytical Methods**, Reino Unido, v. 8, n. 48, p. 8498-8505, 2016.

SU, W.H., HE, H.J., SUN, D.W., 2017. Non-Destructive and rapid evaluation of staple foods quality by using spectroscopic techniques: A review. Crit. Rev. Food Sci. Nutr. 57, 1039–1051. https://doi.org/10.1080/10408398.2015.1082966.

TANG, J.; ALELYANI, S.; LIU, H. Feature selection for classification: A review. In: **Data Classification: Algorithms and Applications**. [s.l: s.n.]. p. 37–64, 2014.

TIBOLA, C.S., MEDEIROS, E.P., SIMEONE, M. L. F., OLIVEIRA, M.A., Espectroscopia no infravermelho proximo para avaliar indicadores de qualidade tecnologica e contaminantes em grãos. Brasilia, DF, EMBRAPA, 2018. 200p.

VERAS, G., ALVES, V. D., RAMOS, H. A., GOMES, M., FIGUEIREDO, L., MATIAS, E. V. S**.** Cientometric profile of the Chemometrics in Brazil. *Quim. Nova,* Vol. 45, No. 10, 1315-1321, 2022. http://dx.doi.org/10.21577/0100-4042.20170930

VITALE, R.; BEVILACQUA, M.; BUCCI, R.; MAGRÌ, A. D.; MARINI, F. A rapid and noninvasive method for authenticating the origin of pistachio samples by NIR spectroscopy and chemometric. **Chemometrics and Intelligent Laboratory Systems**, v. 121, p. 90-99, Feb. 2013.

WAFULA, E. N. et al. Antinutrient to mineral molar ratios of raw common beans and their rapid prediction using near-infrared spectroscopy. **Food Chemistry**, v. 368, p. 130773, jan. 2022.

WANG, S. et al. Structured learning for unsupervised feature selection with high-order matrix factorization. **Expert Systems with Applications**, v. 140, p. 112878, fev. 2020.

WAN, C. **Hierarchical Feature Selection for Knowledge Discovery: Application of Data mining to the biology of ageing**. Springer ed. 2018
WESTAD, F.; MARINI. F. Validation of chemometric models – a tutorial. **Analytica Chimica Acta**, v. 893, p. 14-24, Set. 2015.

WILLIAMS, P. How it all began. In: INTERNATIONAL CONFERENCE ON NEAR INFRARED SPECTROSCOPY, 17., 2015, Foz do Iguassu. **Proceedings**… Campinas: International Council for Near Infrared Spectroscopy, 2015. p. 128-132. Disponível em: https://proceedings.galoa.com.br/NIR/papers/how-it-all-began . Acesso em: 20 nov. 2022.

WOLD, S.; *ACS Symposium Series: Years of Chemometrics – From Bruce Kowalski to the Future*, Washington, USA, 2015.

XIA, C., YANG, S., HUANG, M., ZHU, Q., GUO, Y., & QIN, J. (2019). *Maize seed classification using hyperspectral image coupled with multi-linear discriminant analysis. Infrared Physics & Technology, 103077.* doi:10.1016/j.infrared.2019.103077

XIAOBO, Z., JIEWEN, Z., POVEY, MJW, HOLMES, M., & HANPIN, M. (2010). *Métodos de seleção de variáveis em espectroscopia de infravermelho próximo. Analytica Chimica Acta, 667(1-2), 14–32.* doi:10.1016/j.aca.2010.03.048 10.1016/j.aca.2010.03.048

XING, Z. et al. A method combining FTIR-ATR and Raman spectroscopy to determine soil organic matter: Improvement of prediction accuracy using competitive adaptive reweighted sampling (CARS). **Computers and Electronics in Agriculture**, v. 191, p. 106549, dez. 2021.

YESILYURT, K., AYDIN, M. (2020). Experimental investigation on the performance, combustion and exhaust emission characteristics of a compression-ignition engine fueled with cottonseed oil biodiesel/diethyl ether/diesel fuel blends. **Energy Conversion and Management**, 205, 112355.

YUN, Y.-H., LI, H.-D., DENG, B.-C., & CAO, D.-S. (2019). An overview of variable selection methods in multivariate analysis of near-infrared spectra Trends in Analytical Chemistry, 113, 102–115. https://doi.org/10.1016/j.trac.2019.01.018

ZACHET, E., BETIN, F. M. M., SILVA, I. C. A., & RIBEIRO, R. V. **Destoxificação biológica de tortas de sementes de algodão para alimentação animal.** Revista Semana Tecnológica, n.1, 2019**.**

ZAREEF, M., et al., Application of benchtop NIR spectroscopy coupled with multivariate analysis for rapid prediction of antioxidant properties of walnut (Juglans regia), Food Chemistry, Volume 359,2021, 129928, ISSN 0308-8146, https://doi.org/10.1016/j.foodchem.2021.129928.

ZHANG, M.; ZHANG, S.; IQBAL, J. Key wavelengths selection from near infrared spectra using Monte Carlo sampling-recursive partial least squares. **Chemometrics and Intelligent Laboratory Systems**, v. 128, p. 17–24, 2013

ZHOU, R. et al. Supervised Dimensionality Reduction Technology of Generalized Discriminant Component Analysis and Its Kernelization Forms. **Pattern Recognition**, p. 108450, nov. 2021.

ZONTOV, Y., RODIONOVA, O., KUCHERYAVSKIY, S. POMERANTSEV, A., "DD SIMCA – A MATLAB GUI tool for data driven SIMCA approach," *Chemometrics and Intelligent Laboratory Systems,* vol. 167, pp. 23-28, 2017.