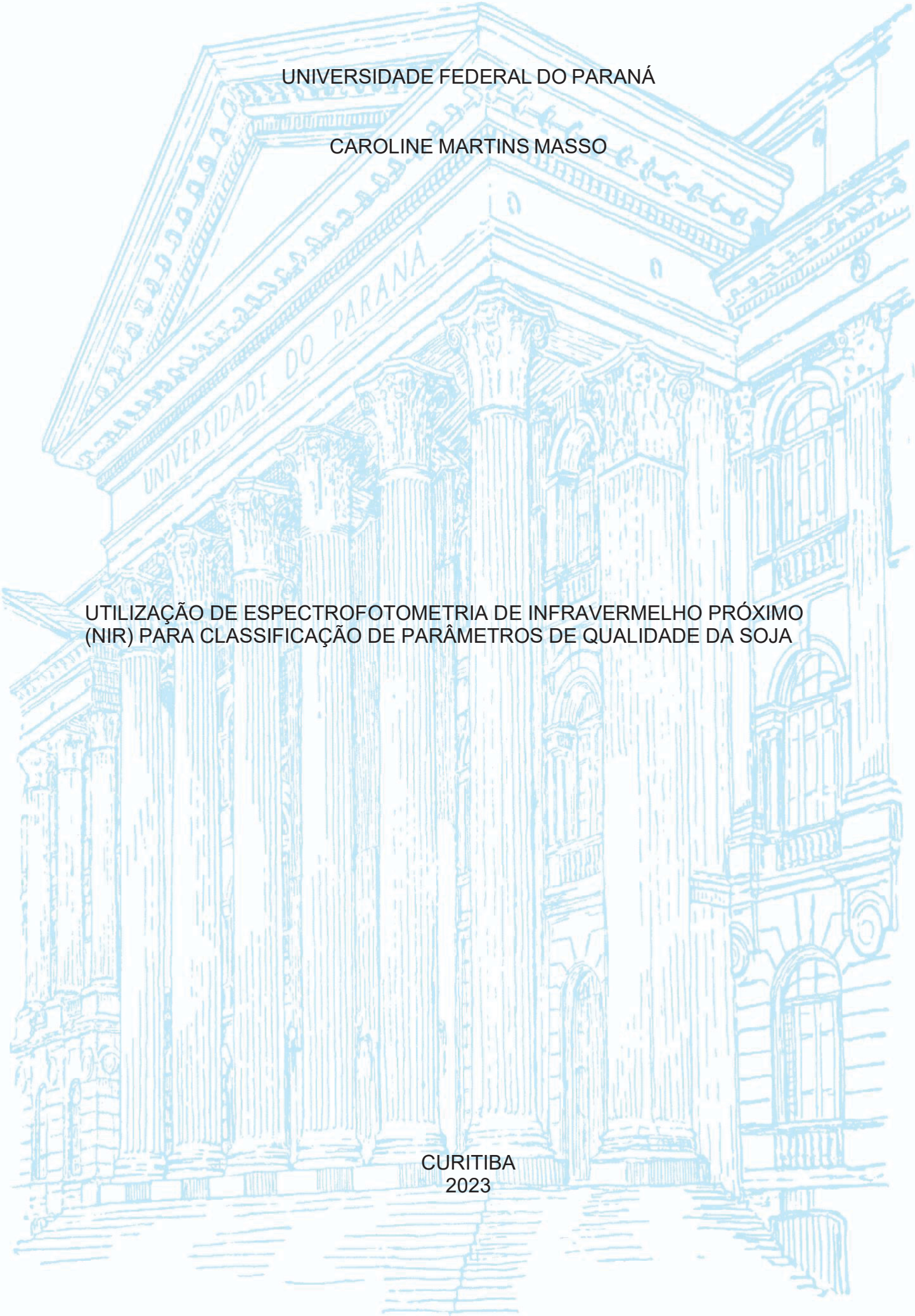
UNIVERSIDADE FEDERAL DO PARANÁ

CAROLINE MARTINS MASSO




# UTILIZAÇÃO DE ESPECTROFOTOMETRIA DE INFRAVERMELHO PRÓXIMO (NIR) PARA CLASSIFICAÇÃO DE PARÂMETROS DE QUALIDADE DA SOJA

CURITIBA
2023

CAROLINE MARTINS MASSO

# UTILIZAÇÃO DE ESPECTROFOTOMETRIA DE INFRAVERMELHO PRÓXIMO (NIR) PARA CLASSIFICAÇÃO DE PARÂMETROS DE QUALIDADE DA SOJA

Dissertação apresentada ao Curso de Pós-Graduação em Engenharia Química, Setor de Tecnologia, da Universidade Federal do Paraná, como requisito parcial à obtenção do título de Mestre em Engenharia Química.

Orientador(a): Prof. Dr. Alexandre Ferreira Santos

CURITIBA
2023

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
UNIVERSIDADE FEDERAL DO PARANÁ
SISTEMA DE BIBLIOTECAS – BIBLIOTECA DE CIÊNCIA E TECNOLOGIA

Masso, Caroline Martins
Utilização de espectrofotometria de infravermelho próximo (NIR) para classificação de parâmetros de qualidade da soja / Caroline Martins Masso. – Curitiba, 2023.
1 recurso on-line : PDF.

Dissertação (Mestrado) - Universidade Federal do Paraná, Setor de Tecnologia, Programa de Pós-Graduação em Engenharia Química.

Orientador: Alexandre Ferreira Santos

1. Soja. 2. Sementes – Qualidade. 3. Espectrofotometria. 4. Espectro infravermelho. I. Universidade Federal do Paraná. II. Programa de Pós-Graduação em Engenharia Química. III. Santos, Alexandre Ferreira. IV. Título.

Bibliotecário: Elias Barbosa da Silva CRB-9/1894

MINISTÉRIO DA EDUCAÇÃO
SETOR DE TECNOLOGIA
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO ENGENHARIA QUÍMICA
- 40001016056P9

# TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação ENGENHARIA QUÍMICA da Universidade Federal do Paraná foram convocados para realizar a arguição da dissertação de Mestrado de **CAROLINE MARTINS MASSO** intitulada: **Utilização de espectrofotometria de infravermelho próximo (NIR) para classificação de parâmetros de qualidade da soja**, sob orientação do Prof. Dr. ALEXANDRE FERREIRA SANTOS, que após terem inquirido a aluna e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

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

CURITIBA, 04 de Agosto de 2023.

Assinatura Eletrônica
04/08/2023 11:25:15.0
ALEXANDRE FERREIRA SANTOS
Presidente da Banca Examinadora

Assinatura Eletrônica
04/08/2023 11:45:44.0
MARCELO KAMINSKI LENZI
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
04/08/2023 11:39:00.0
EMERSON MARTIM
Avaliador Externo (PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ)

Dedico este trabalho à minha família,
que sempre me incentivou, me apoiou a realizá-lo e sempre
acreditou em mim.

## AGRADECIMENTOS

Agradeço primeiramente a Deus, por sempre ter me dado sabedoria e força para minhas conquistas e enfrentar os desafios.

Agradeço ao meu professor orientador, Professor Dr. Alexandre Ferreira Santos por confiar no meu trabalho e me ajudar a torná-lo possível.

Agradeço aos meus pais, Terezinha e Valdir, por toda educação, incentivo e carinho. Por nunca medirem esforços em me ajudar a conquistar meus sonhos, eles que são meus maiores exemplos. Agradeço também a minha querida irmã, Laura, por estar ao meu lado em todos os momentos, pelos conselhos e momentos de sabedoria, que sempre me motivou a ir além. A qual, mesmo a um oceano de distância sempre se fez presente com todo seu carinho. Agradeço ao meu Amado Renato, que sempre esteve ao meu lado me incentivando e me apoiando a buscar sempre o melhor. O qual em momentos que pareciam difíceis me fez enxergar as possibilidades existentes.

Agradeço ao Laboratório de Engenharia Química, Usina Piloto, o qual representado pela Dra. Montserrat Fortuny, que não mediu esforços na procura de alternativas e realização de análises, sempre buscando melhores caminhos para resultados.

Agradeço também meus amigos da Rumo Logística, os quais sempre me motivaram no dia a dia a dar um passo a mais e buscar sempre ser uma melhor profissional. Ao Mauro e Êmili, por sempre ter uma palavra de incentivo e por toda ajuda nas etapas de programação.

Por último, mas não menos importante, agradeço aos meus professores da graduação por todos os ensinamentos, sem eles não seria possível chegar até aqui.

"Algumas pessoas querem que aconteça,
outras desejam que aconteça, outras fazem acontecer."
(Michael Jordan)

## RESUMO

A semente de soja tem uma grande representatividade na agricultura do Brasil, sendo o país um dos maiores produtores deste grão no mundo. Para garantir seu valor comercial e o atendimento dos requisitos normativos dos órgãos de controle, os grãos de soja devem atender aos critérios e parâmetros de qualidade estabelecidos na legislação. Estes são de extrema importância na classificação dos grãos na entrada de sua unidade de armazenamento, e para controle de qualidade na exportação dos mesmos. Existem diversos danos que podem ser causados ao grão tanto durante seu transporte quanto durante a armazenagem, e isso pode afetar diretamente o seu valor comercial. Além disso, fatores biológicos, químicos e físicos também podem causar danos ao grão de soja. Uma tecnologia que vem sendo aplicada com relativo sucesso na análise de propriedades de grãos é espectrofotometria de infravermelho próximo (NIR). Dentre as propriedades analisadas, destacam-se umidade, teores de proteínas, gorduras e outros fatores químicos de sementes e grãos. A utilização do NIR se destaca por sua facilidade de manuseio, rapidez na obtenção do resultado e assertividade. Quando se observa o processo de análise dos grãos para obtenção de parâmetros de qualidade nas unidades armazenadoras, conclui-se que o processo de classificação é bastante manual. Para uma maior agilidade para o processo, estuda-se a aplicação de um espectrofotômetro NIR portátil (MicroNIR) nas análises de tais parâmetros de qualidade. Diferentes técnicas de análise multivariável e de processamento de espectros NIR foram investigadas para fins de classificação dos grãos de soja entre esverdeados, ardidos, queimados e bons. Neste trabalho, demonstra-se a excelente capacidade do MicroNIR na classificação dos grãos de soja quando se efetua o uso combinado de espectros brutos e Análise de Componente Principal com Análise Discriminatória (PCA–DA), permitindo predições com elevada exatidão e erros médios de 0,8%.



Palavras-chave: Grão de Soja. Infravermelho próximo. Parâmetros de Qualidade. MicroNIR.

## ABSTRACT

Soybean seed has a great representativeness in Brazilian agriculture, being the country one of the largest producers of this grain in the world. To ensure its commercial value and to meet the regulatory requirements of the control bodies, soybeans must meet the quality criteria and parameters established in the legislation. These are of extreme importance in the classification of grains at the entrance of their storage unit, and for quality control in their export. There are several damages that can be caused to the grain both during its transportation and storage, and this can directly affect its commercial value. In addition, biological, chemical and physical factors can also cause damage to soybeans. A technology that has been applied with relative success in the analysis of grain properties is near infrared spectrophotometry (NIR). Among the properties analyzed are moisture, protein, fat and other chemical factors of seeds and grains. The use of NIR stands out for its ease of handling, speed in obtaining the result and assertiveness. When observing the process of grain analysis to obtain quality parameters in storage units, it is concluded that the classification process is quite manual. For a greater agility for the process, the application of a portable NIR spectrophotometer (MicroNIR) in the analysis of such quality parameters is studied. Different techniques of multivariable analysis and NIR spectra processing were investigated for the purpose of classification of soybeans between greenish, burnt, burnt and good. In this work, the excellent ability of MicroNIR in the classification of soybeans is demonstrated when the combined use of raw spectra and Principal Component Analysis with Discriminant Analysis (PCA-DA) is performed, allowing predictions with high accuracy and average errors of 0.8%.



Keywords: Soybean. Near Infrared. Quality Parameters. MicroNIR.

## LISTA DE FIGURAS

## LISTA DE GRÁFICOS

## LISTA DE TABELAS

## LISTA DE ABREVIATURAS OU SIGLAS

| | |
|---|---|
| NIR | - *Near infrared* |
| PLS | - *Partial Least Square* |
| RL | - Regressão Logística |
| SVM | - Suporte Vetorial de Máquinas |
| PCA | - Análise de Principais Componentes |

**SUMÁRIO**

# 1 INTRODUÇÃO

A soja possui um crescimento considerável em sua produção ao redor do mundo, número que ultrapassa a casa das 300 milhões de toneladas métricas anualmente. Este grão pertence à família das leguminosas alimentares, sendo que apresenta em relação mássica a maior quantidade de proteína e maior teor de óleo, sendo respectivamente, 40% e 20%. Dentre os vários motivos que podem ser considerados incentivadores para aumento de plantio da soja, um deles é o quão versátil é o uso do grão em seu modo final, sendo possível a aplicação desde alimentação humana, ração animal e matéria prima industrial. Além disso, pode ser utilizado também para a produção de biocombustíveis (Liu, 2016).

Sabe-se que até que o grão vá para o processamento industrial, há uma etapa de logística, armazenamento e comercialização do mesmo, sendo muito importante sempre manter os parâmetros de qualidade, pois este fator afeta diretamente no valor do produto. Apesar da agricultura brasileira possuir bastante tecnologia, ainda ocorrem perdas tanto qualitativas quanto quantitativas. Estas podem ser oriundas do processo de pós-colheita, e, durante o armazenamento, a massa de grãos é constantemente submetida a fatores externos, os quais podem ser físicos, como temperatura e umidade; químicos, oxigenação, fermentação e entre outros fatores. Estes pontos de qualidade do grão possuem alta relevância devido ao tempo de armazenagem, que pode durar até mesmo cerca de um ano. Para um armazenamento seguro proporciona-se uma condição desfavorável ao desenvolvimento de fatores biológicos que possam afetar a qualidade do grão. Em particular, em regiões tropicais os fatores climáticos afetam fortemente a qualidade do grão, exigindo uma maior atenção e controle das condições de temperatura e umidade. A aparência é um outro aspecto considerado de bastante relevância na comercialização do grão, sendo que na etapa de armazenagem, podem ocorrer alterações na coloração do grão, assim como alterações biológicas (Alencar et al., 2009). Desta maneira, torna-se essencial todo monitoramento e controle de qualidade aplicado para uma garantia da conformidade do grão a ser comercializado.

O espectrofotômetro de infravermelho próximo (*NIR – Near Infrared*) é um equipamento utilizado para realizar leitura de parâmetros de amostras, sendo uma

característica importante sua metodologia não destrutiva, rápida, além de que não há um pré-requisito de pré-tratamento da amostra para realizar leituras. Esta técnica espectrofotométrica vem sendo utilizada amplamente em produtos agrícolas, e principalmente em pesquisas que buscam análise de proteínas, gorduras, umidade e coloração da soja, por exemplo. Para tanto, são utilizadas técnicas estatísticas para correlação dos dados espectrais com as propriedades químicas e físicas das amostras. A precisão dos resultados desse método e sua reprodutibilidade são influenciadas pelo tamanho das partículas e a uniformidade das amostras. (Zhu et al., 2018). Com o devido controle destes fatores, torna-se possível empregar a espectrofotometria NIR para determinação de parâmetros de qualidade do grão.

Apesar da literatura apresentar trabalhos empregando com sucesso a espectroscopia NIR na inferência de alguns parâmetros químicos e físicos dos grãos de soja, ainda inexistem relatos sobre o uso de ferramentas para monitoramento de uma série de parâmetros requisitados por órgãos públicos e que são fundamentais para a aceitação dos grãos provenientes de produtores e cooperativas para fins de estocagem em terminais e exportação.

Considerando os problemas expostos, as preocupações relacionadas aos parâmetros de qualidade dos grãos na etapa de armazenamento e comercialização dos grãos, e a grande importância que a soja possui no cenário da agricultura no Brasil, o objetivo deste trabalho é avaliar a utilização da espectroscopia NIR combinada com modelos matemáticos para análise simultânea de características do grão de soja, incluindo umidade, quantidade de grãos ardidos e verdes em uma amostra, para fins de controle de parâmetros de qualidade tanto físicos quanto químicos dos grãos de soja.

## 1.1 JUSTIFICATIVA

Considerando a grande volumetria de plantio de soja no Brasil e sua grande representatividade no cenário agrícola mundial, torna-se evidente a motivação dos estudos visando a garantia de qualidade do grão de soja durante as etapas de armazenagem e comercialização.

Os requisitos de qualidade dos grãos de soja especificados na legislação brasileira incluem teores de umidade, grão ardido, grão verde, grão quebrado,

impurezas e temperatura em uma mesma análise da amostragem retirada na chegada do produto na unidade armazenadora.

Apesar de terem sido realizados diversos estudos que abordam os parâmetros de qualidade que envolvem teores de umidade, proteínas e óleos, e também danos físicos ao grão, não há estudos que abordem parâmetros físicos e químicos simultaneamente, além de todas as variáveis consideradas para qualidade nos armazéns brasileiros. Desta maneira, o desenvolvimento de uma ferramenta capaz de inferir simultaneamente diferentes parâmetros de qualidade do grão de soja é de importância estratégica para produtores, cooperativas, terminais de exportação, unidades armazenadoras e órgãos de controle, podendo trazer agilidade na tomada de decisão, redução de custos de transporte e armazenamento.

## 1.2 OBJETIVOS

### 1.2.1 Objetivo geral

O objetivo geral desta dissertação é desenvolver uma metodologia baseada na espectrofotometria de infravermelho próximo para monitoramento e caracterização simultânea de propriedades dos grãos de soja, incluindo grãos queimados, grão ardido e grãos esverdeados.

### 1.2.2 Objetivos específicos

- Seleção de um conjunto de amostras de grãos de soja com diferentes morfologias;
- Determinação experimental de diferenciação de alguns parâmetros de qualidade dos grãos de soja com base em metodologias recomendadas pelos órgãos de controle.
- Investigação de métodos quimiométricos e de pré-processamento de dados espectrais.
- Desenvolvimento e validação de modelos espectrais para avaliação qualitativa da separação de algumas características referente a qualidade relacionadas as amostras analisadas.

# 2 REVISÃO DE LITERATURA

## 2.1 A SOJA NA AGRICULTURA

Há registros de que a origem da soja ocorreu na região da China em um período entre os anos de 2883 e 2838 AC, sendo que este produto era considerado um grão sagrado, assim como outros grãos da cultura, como arroz e trigo. As plantações de soja de cinco milênios atrás apresentam muita diferença das plantações que conhecemos, pois eram plantas rasteiras que se desenvolviam ao longo de rios e lagos. O processo de mudança na plantação da soja ocorreu no século XI a.C., a partir de cruzamentos naturais de espécies, que foram realizados na ciência chinesa. A partir dessa evolução e pesquisa do grão de soja seu cultivo começa a ser introduzido no Sul da China, indo para a Coréia, o Japão e outros países do atual Sudeste da Ásia. Na parte ocidental, o grão surge no final do século XV e início do século XVI, época conhecida pelas grandes navegações europeias (APROSOJA, 2021). No Brasil, o plantio de soja começa a ser visto com potencial comercial a partir da década de 1960, sendo que até o momento, o principal plantio era de trigo e a soja aparecia somente como uma safra para o verão. Assim, utilizando uma visão estratégica comercial, a soja começa a ser considerada uma necessidade em 1966. Mas somente em 1970 os agricultores brasileiros despertaram um grande interesse na plantação, enxergando a lucratividade da safra do Brasil ocorrendo na entressafra estadunidense. Assim, cada vez mais o país vem investindo em tecnologia e estudos para o plantio deste grão (EMBRAPA, 2016).

Na safra de 2019/2020, o Brasil recuperou o primeiro lugar em produção mundial do grão, sendo que seu plantio representa cerca de 37% em relação às plantações no mundo. Na Tabela 1 apresentam-se os dados referentes à safra de soja 2019/2020. Observa-se também na safra de 2019/2020 que o Brasil foi o maior exportador mundial de soja.

TABELA 1 - DADOS PRODUÇÃO DE SOJA SAFRA 2019/2020.

| Local | Produção | Exportação | Estoque final |
|---|---|---|---|
| Mundo | 337,7 | 149,15 | 96,67 |
| Brasil | 123 | 76 | 31,17 |
| Estados Unidos | 96,84 | 48,31 | 12,92 |
| Argentina | 53 | 8,2 | 25,89 |
| China | 18,1 | 0,13 | 19,13 |
| Paraguai | 10,2 | 6,2 | 0,11 |
| Europa | 2,6 | 0,25 | 1,07 |
| Sudeste Asiático | 0,64 | 0,04 | 1,25 |
| México | 0,24 | 0 | 0,2 |

FONTE: Adaptado de USDA (2020)

LEGENDA: Milhões de toneladas

Pelo extenso território brasileiro, há algumas regiões que se destacaram por sua grande produtividade de soja na safra 2019/2020. A Tabela 2 apresenta a relação dos quatro principais estados brasileiros.

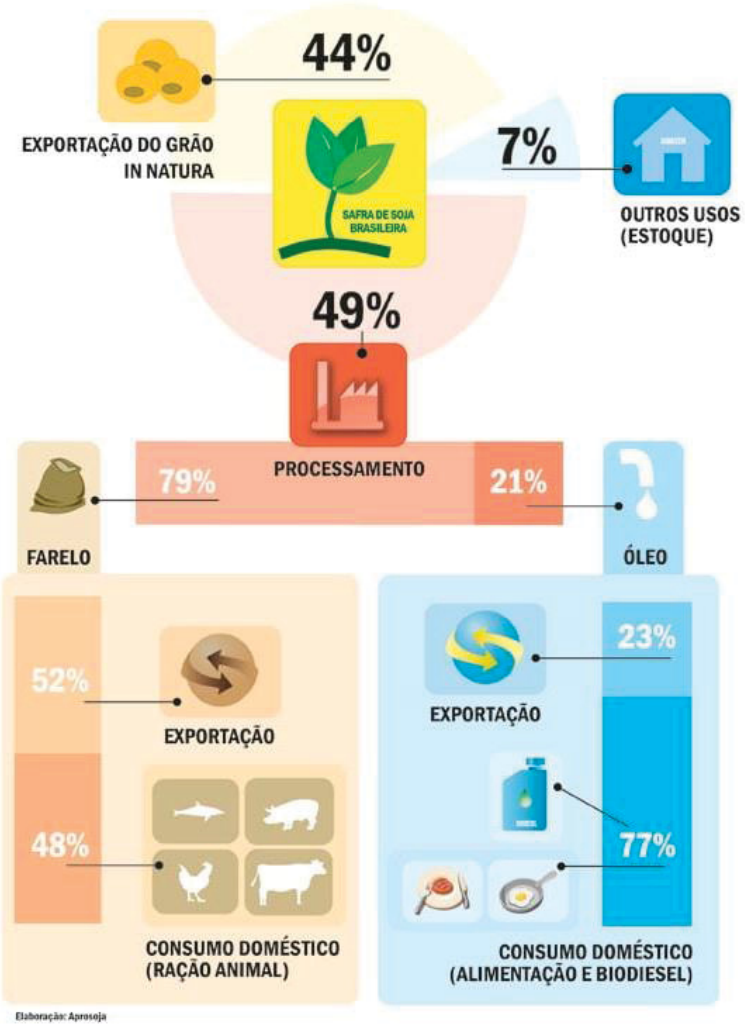
TABELA 2 - DADOS PRODUÇÃO DE SOJA POR REGIÃO SAFRA 2019/2020.

| Região | Produção (milhões toneladas) | Área plantada (milhões hectares) | Produtividade (Kg/hectare) |
|---|---|---|---|
| Mato Grosso | 35,88 | 10 | 3,58 |
| Paraná | 21,59 | 5,5 | 3,92 |
| Rio Grande do Sul | 11,44 | 5,90 | 1,93 |
| Goiás | 13,15 | 3,54 | 3,71 |

FONTE: Adaptado de EMBRAPA (2020)

De toda produção brasileira de soja, sua aplicação e utilização é dividida dentre os setores representados na Figura 1.

FIGURA 1 - DESTINAÇÃO DA SOJA NO BRASIL




FONTE: APROSOJA, 2018

Desta maneira, é possível observar a grande abrangência e ramificação para a soja no Brasil, além de considerar a grande extensão do território brasileiro. Além disso, pode-se observar a representação em porcentagem da utilização de cada item em diferentes setores. E assim, surge a necessidade de um investimento na logística e armazenagem deste produto, para que assim tenha a garantia de qualidade para a exportação e uso interno.

## 2.2 PARÂMETROS DE QUALIDADE DA SOJA

Os parâmetros de qualidade do grão de soja possuem bastante importância para comercialização e processamento, podendo influenciar diretamente no valor do

produto. Durante a etapa de armazenamento e transporte o grão sofre bastante influência de fatores como temperatura, umidade, além de influências biológicas, como bactérias, fungos, insetos. Outro fator a ser levado em consideração é a característica física do grão, já que um grão ardido, quebrado e entre outros fatores, também são levados em consideração ao se analisar os atributos de qualidade de um grão. Com toda tecnologia empregada na colheita e pós-colheitas, os fatores de influência podem ser minimizados, mas não anulados (Alencar et al., 2009).

No Brasil, segundo a Lei 9972 / 2000 instrução IN11 para soja de 16/05/2007, do Ministério da Agricultura, Pecuária e Abastecimento, os grãos possuem parâmetros de qualidade que devem ser medidos e comprovados pela unidade armazenadora e/ou produtora. Para cada parâmetro, há a definição dos mesmos e a especificação de porcentagem máxima tolerada em uma amostragem. Além disso, há a classificação primária do grão em dois grupos:

- Grupo I: Soja destinada ao consumo in natura;
- Grupo II: Soja destinada a outros usos. (Instrução Normativa 11/2007, 2007)

Das definições dos parâmetros de acordo com a Lei 9972 / 2000 instrução IN11 de 16/05/2007, do Ministério da Agricultura, Pecuária e Abastecimento:

Umidade: quantidade percentual de água encontrada na amostra em seu estado *in natura*. (SENAR, 2017)

Matérias Estranhas e/ou Impurezas: materiais que vazem nas peneiras ou fiquem retidos nelas, sendo estas com as seguintes especificações (SENAR, 2017):

- Espessura de chapa: 0,8 mm
- Quantidade de furos: 400/100 cm²
- Diâmetro dos furos: 3,0 mm

Avariados: são os grãos que se encontram com as seguintes características: ardidos, brotados, imaturos, chochos, mofados ou danificados. Grãos que possuem casca enrugada ou possuem alteração na coloração. (SENAR, 2017)

Ardidos: grãos apresentam, pela ação do calor e umidade, visivelmente fermentados com coloração marrom ou escura na casca e afetando a polpa. (SENAR, 2017)

FIGURA 2 - GRÃOS ARDIDOS

FONTE: SENAR, 2017

- Brotados: Grãos com indícios de germinação ou germinados, ou seja, apresentam a emissão de radícula. (SENAR, 2017)

FIGURA 3 - GRÃOS BROTADOS

FONTE: SENAR, 2017

- Imaturos: Grãos que não tiveram seu completo desenvolvimento fisiológico, por isso apresentam formato irregular. (SENAR, 2017)

FIGURA 4 - GRÃOS IMATUROS

FONTE: SENAR, 2017

- Chochos: Grãos que em seu desenvolvimento possuem característica de estarem enrugados e atrofiados. (SENAR, 2017)

FIGURA 5 - GRÃOS CHOCHOS

FONTE: SENAR, 2017

- Mofados: Grãos que apresentam estarem afetados por fungos, sendo que estes podem ser mofos ou bolor, que podem ser vistos sem utilizar equipamentos de visualização. (SENAR, 2017)

FIGURA 6 - GRÃOS MOFADOS

FONTE: SENAR, 2017

- Danificados: Grãos que apresentam características (manchas na polpa, alterados, deformados) de estarem afetados por pragas ou doenças, que podem ser causados em qualquer fase de desenvolvimento do grão. (SENAR, 2017)

FIGURA 7 - GRÃOS DANIFICADOS

FONTE: SENAR, 2017

- Quebrados: Pedaços de grãos que estão em bom estado e ficam retidos no processo de peneiramento de impurezas. (SENAR, 2017)

FIGURA 8 - GRÃOS QUEBRADOS

FONTE: SENAR, 2017

- Esverdeados: Grãos que apresentam coloração esverdeada tanto na casca quanto na polpa, apesar de estarem completamente desenvolvidos. (SENAR, 2017)

FIGURA 9 - GRÃOS ESVERDEADOS

FONTE: SENAR, 2017

Desses parâmetros de qualidade, há também a definição da porcentagem tolerada dos principais parâmetros que são medidos para controle da qualidade, assim, na Tabela 3 é possível visualizar os parâmetros de qualidade mínimos exigidos IN11 para soja de 16/05/2007, do Ministério da Agricultura, Pecuária e Abastecimento, e a porcentagem permitida para o Grupo I.

TABELA 3 - PARÂMETROS DE QUALIDADE E PORCENTUAL TOLERADO GRUPO I

| Tipo | Avariados (ardidos e queimados) | Máximo de queimados | Mofados | Total (1) | Esverdeados | Partidos, Quebrados e Amassados | Matérias estranhas e impurezas |
|---|---|---|---|---|---|---|---|
| 1 | 1% | 0,3% | 0,5% | 4,0% | 2,0% | 8,0% | 1,0% |
| 2 | 2% | 1% | 1,5% | 6,0% | 4,0% | 15,0% | 1,0% |

FONTE: IN11 para soja de 16/05/2007, do Ministério da Agricultura, Pecuária e Abastecimento.

(1) Representa a soma de queimados, ardidos, mofados, fermentados, germinados, danificados, imaturos e chochos.

A partir das porcentagens coletadas na amostra, há então a classificação em tipo 1 ou tipo 2, do grupo I. Os parâmetros de qualidade definidos para o grupo 2 se encontram na Tabela 4.

TABELA 4 - PARÂMETROS DE QUALIDADE E PORCENTUAL TOLERADO GRUPO II

| Tipo | Avariados (ardidos e queimados) | Máximo de queimados | Mofados | Total (1) | Esverdeados | Partidos, Quebrados e Amassados | Matérias estranhas e impurezas |
|---|---|---|---|---|---|---|---|
| Padrão básico | 4,0% | 1,0% | 6,0% | 8,0% | 8,0% | 30,0% | 1,0% |

FONTE: IN11 para soja de 16/05/2007, do Ministério da Agricultura, Pecuária e Abastecimento.

(1) Representa a soma de queimados, ardidos, mofados, fermentados, germinados, danificados, imaturos e chochos.

Diante dos parâmetros de qualidade e tolerâncias, são realizadas as análises, para garantia para armazenagem e transporte do grão. Entretanto, estas análises atualmente são essencialmente manuais, demoradas, e muito dependentes da acuidade visual do operador. Após a coleta do grão, realiza-se a medição do teor de umidade presente na amostra e o registro da temperatura. Então inicia-se a análise

dos parâmetros de qualidade especificados na Tabela 3. Para tanto, o operador retira visualmente os itens associados a cada um dos parâmetros e realiza a pesagem individualmente. Assim, obtendo esse valor de massa, bem como a massa total da amostra, é possível calcular a porcentagem que cada um dos itens de parâmetros da qualidade representa na amostragem.

## 2.3 ESPECTROSCOPIA DE INFRAVERMELHO PRÓXIMO (NIR)

A utilização da tecnologia NIR está ligada à análise de grãos e produtos derivados, desde a década de 1960, sendo Karl Norris o primeiro envolvido no desenvolvimento de metodologias NIR, na época utilizada no lugar da titulação (Hart et al., 1962). Sua equipe realizou um trabalho visando efetuar a determinação do teor de umidade da semente. Somente no ano de 1980, o espectrofotômetro NIR era utilizado pelo Serviço Federal de Inspeção de Grãos dos Estados Unidos. Hoje, há vários trabalhos que mostram medidas de umidade, proteína, medidas de características físicas utilizando o NIR (Ferreira et al., 2014).

A técnica de espectroscopia de infravermelho próximo (NIR) possui como característica ser uma tecnologia não invasiva, as quais apresentam boas medições quantitativas, com uma boa velocidade e facilidade de manuseio, todos esses fatores fazem com que o espectrofotômetro NIR seja utilizado em vários setores, sendo o alimentício um deles. Os equipamentos NIR oferecem a possibilidade de serem calibrados de acordo com sua aplicação, podendo ser desde constituintes da agricultura, como soja, trigo, cevada e semente madura e também de alimentos como lácteos, carne vermelha, peixes e entre outros (Baianu et al., 2011). A faixa coberta pela radiação no infravermelho próximo (NIR) é de 780 e 2500 nanômetros, sendo que o produto analisado é irradiado com radiação NIR, e a radiação medida é a refletida ou transmitida. (Nicolaï et al., 2007). Sabe-se que as bandas de absorção que podem ser observadas na região de espectroscopia NIR são em sua maioria sobretons, as quais possuem a tendência de possuírem a intensidade mais fraca. Entretanto, essa característica pode ser avaliada como vantajosa, pelo fato de que a regia de banda da espectroscopia NIR são suficientes para observar grupos funcionais que possuem átomo de hidrogênio ligado a carbono, nitrogênio ou oxigênio. Sendo que, esses grupos funcionais são constituintes principais de alimentos, como

proteínas, lipídios, água e carboidratos. Portanto, dependendo do tipo da amostra a ser avaliada, é possível realizar medições de refletância e transmitância pela espectroscopia NIR (Baianu et al., 2011).

Pode-se dizer que as principais fontes de variação no espectro incluem as seguintes características: tamanho, forma e curvatura do grão, assim como seu posicionamento durante a leitura do espectro (Kosmowski & Worku, 2018). Desta maneira, com a utilização da metodologia NIR podem-se inferir características tanto físicas quanto químicas de grãos, tornando possível extrair informações sobre os parâmetros de qualidade dos grãos a partir dos espectros.

## 2.4 APLICAÇÃO DAS METODOLOGIAS DO NIR EM MODELOS MATEMÁTICOS

Para processamento dos dados de infravermelho de proximidade nas análises dos grãos é necessário a utilização de métodos de modelagem que possuem eficácia de acordo com a característica do trabalho a ser realizada, para produção de resultados mais tratados. Existem vários métodos que podem ser utilizados para modelos de classificação, como *K-means*, *Random Forest* (RF), Regressão Logística (LR) e o Mínimos Quadrados Parciais (PLS). (Wang et al., 2021)

Uma outra metodologia de análise a ser considerada é a de componentes principais (PCA), o qual é um método estatístico multivariado que pode trabalhar com grandes conjuntos de dados, revelando padrões e estrutura interna. Sua análise pode ser realizada através da transformação ortogonal, um grupo de dados de variáveis que podem ter correlação é transformado em um grupo de variáveis linearmente não correlacionadas, e o grupo de variáveis transformadas é chamado de componentes principais (CPs). De modo geral, os primeiros componentes principais podem ser usados para maximizar a variação espectral dos conjuntos de dados, interpretar agrupamentos de amostras e semelhanças. Desta maneira, para utilizar de métodos de pré-processamento e analisar se há algum tipo de padrão comportamental e de correlação é interessante a utilização da metodologia de PCA. (Wang et al., 2021)

### 2.4.1 Linguagem de Programação – Python

Atualmente, existem muitas linguagens de programação, sendo o Python uma dessas linguagens. Desta maneira, podemos dizer que Python é uma linguagem de programação de uso geral frequentemente aplicada em funções que podem ser chamadas de script, ou programa, para assim descrever um código. Por ser uma linguagem relativamente nova, possuem alguns pontos que justificam sua grande usabilidade:

- Qualidade do software: O código Python é projetado para ser legível. Tornando-o desta maneira reutilizável e sustentável. Além de que, a uniformidade do código Python facilita a compreensão, mesmo que não seja você o autor.
- Produtividade do desenvolvedor: Normalmente, o código Python possui um terço a um quinto do tamanho do código C++ ou Java. Os programas Python também possuem uma execução imediata sem as etapas demoradas de compilação e vinculação exigidas por algumas outras ferramentas, aumentando ainda mais a velocidade do programador.
- Portabilidade do programa: a grande maioria dos programas em Python podem ser executados sem quaisquer alterações nas principais plataformas do computador. Além de oferecer diversas opções para codificar interfaces gráficas de usuário portáteis, programas de acesso a banco de dados, sistemas baseados na Web, por exemplo.
- Bibliotecas de suporte: possui uma grande coleção de funcionalidades pré-construídas e portáteis, que pode ser nomeada de biblioteca padrão. Essa biblioteca oferece suporte a uma variedade de tarefas de programação em nível de aplicativo, que são desde a correspondência de padrões de texto até scripts de rede. Outro ponto, é que o Python pode ser estendido com bibliotecas próprias e uma vasta coleção de software de suporte a aplicativos de terceiros.
- Integração de componentes: existe uma variedade de integração que permitem que o Python se comunique com facilidade em outras partes de aplicativos. Sendo assim, o mesmo pode ser usado como ferramenta de extensão ou customização de produtos.

- Prazer: Devido a sua facilidade de uso e ao conjunto de ferramentas integrado, o Python permite que programar seja algo divertido.

Estes são alguns fatores positivos que podem explicar a grande utilização da linguagem de programação Python em diversas aplicações. (Lutz, 2007)

### 2.4.2 Análise por componentes principais (PCA)

O método de análise por componentes principais (PCA) é muito comum, sendo empregado na análise de informações, principalmente utilizada pela sua capacidade de agrupar os dados, isso devido a correlação existente nas mais diversas variáveis mensuradas. Na aplicação de um algoritmo de PCA em um conjunto de variáveis, sendo elas por exemplo, leituras de espectros no infravermelho de proximidade, o as variáveis originais dessa amostra é substituído por um novo conjunto de variáveis, o qual é denominado de Componentes Principais (CPs). Pode ser considerado como a principal característica deste novo conjunto a ortogonalidade, no entanto, o mesmo é facilmente reconstruído por meio de uma combinação linear das variáveis originais (espectros da leitura). A utilização do novo conjunto de variáveis (CPs) possui como uma vantagem o fato de concentrar a maior parte das informações em poucas variáveis. Assim, diminui a dimensionalidade dos dados, sem que ocorra a perda significativa da informação química dos mesmos. (Goelzer Sabin et al., 2004) Uma grande parte de softwares que fazem a análise de PCA disponíveis no mercado utilizam a técnica de decomposição do valor singular (SDV) para obtenção dos CPs. Desta maneira, a primeira componente principal (CP1) é definida no eixo do gráfico de maior variância do conjunto de variáveis originais. Um exemplo desta aplicação quando há cerca de 3 CPs, o sistema se assemelha ao sistema cartesiano de coordenadas, em que todos os eixos são linearmente independentes, isto é, ortogonais entre si. Para o exemplo de conjuntos de espectros, podemos considerar a matriz de dados X (m x n), sendo que m corresponde a quantidade de amostras e n o número de variáveis que pode ser decomposta em 3 outras matrizes, U, S e V. Assim, a equação pode ser demonstrada da seguinte maneira:

$$X\ (m * n) = U * S * V^t$$

Equação 1

Onde as colunas de U e V são ortogonais. A matriz V corresponde aos pesos, em que a primeira coluna contém os pesos dos PCs, isso depende da quantidade de componente principal que possui a amostra analisada. O produto de U x S corresponde à matriz T dos escores. O S representa a matriz diagonal, no qual os elementos possuem informações da quantidade de variância que cada componente principal apresenta. A matriz S representa, a partir das análises anteriores a aplicação do PCA, a quantidade correspondente de componentes principais. Sendo assim, em sua composição os autovalores que possuem valores pequenos serão excluídos, destacando as informações importantes para as análises, eliminando os ruídos experimentais. (Goelzer Sabin et al., 2004)

### 2.4.3 Análise Discriminante Linear (LDA)

Uma grande parte das abordagens tradicionais de classificação na ciência são chamadas de análise discriminante e são chamadas de formas de “modelagem rígida”. Diversas medições são necessárias para determinar o grupo ao qual uma amostra pertence, levando em considerações algumas similaridades para realizar o agrupamento. Em áreas de trabalho como a espectroscopia ocorre de que alguns comprimentos de onda ou regiões do espectro possuam uma maior utilidade do que outros para análise discriminante. Principalmente nas leituras da espectroscopia de infravermelho próximo (NIR). Um outro fator é que diferentes partes de um espectro podem ter intensidades muito diferentes, tornando algumas classes são mais difusas do que outras. Existem algumas ramificações para análise discriminante, podendo ser a Linear, a Mahalanobis e a Quadrática. A de Mahalanobis usa uma distância euclidiana, cada medição assume igual significância, portanto, variáveis correlacionadas, que podem representar uma característica irrelevante, podem ter uma influência desproporcional na análise. Quando deseja se obter a análise de padrões supervisionado, um dos principais objetivos é definir a distância de um objeto

do centro de uma classe. Existem dois usos principais de distâncias estatísticas. A primeira é obter uma medida análoga a uma pontuação, muitas vezes chamada de função discriminante linear, proposta pela primeira vez pelo estatístico Ronald Fisher. Isso difere da distância acima porque é um único número se houver apenas duas classes. Na função quadrática assume-se que a discriminação entre a variância de classes é uma função quadrática.

Observa-se nas equações 2 e 3 o cálculo da distância do objeto ao centro da classe.

$$fi = \left(\overline{Xa} - \overline{Xb}\right) * C_{ab} * Xi^T$$

Equação 2

$$C_{ab} = \frac{(Na - 1) * Ca + (Nb - 1) * Cb}{(Na + Nb - 2)}$$

Equação 3

Onde $Na$ representa o número de objetos no grupo A, e $Ca$ a matriz variância-covariância para este grupo, em que elementos diagonais correspondem à variância de cada variável e os elementos fora da diagonal à covariância, com $Xa$ representando o centroide correspondente para classe A, e o mesmo equivale para o $Xb$. Assume-se que a matemática se torna mais complexa se houver mais de dois grupos. Esta função pode assumir valores negativos. A partir da equação representada é possível aplicar o cálculo da distância para cada um dos modelos. O modelo linear de Fischer apresenta-se como um problema de autovalores e autovetores. (Brereton, 2003) No qual busca a maximizar a variância na razão entre a matriz interclasse das amostras e a matriz intraclasse. A diferença de uma discriminação linear para uma quadrática é que se assume que a separação entre as classes é não-linear. Para determinar a distância na metodologia de Mahalanobis ao centroide de qualquer grupo dado dependendo da quantidade de classes. Haverá uma distância separada para o centro de cada grupo definido para cada classe:

$$d_{in} = \sqrt{(xi - \overline{Xn})} * C_a^{-1} * (xi - \overline{Xn})'$$

Equação 4

,

onde $xi$ é um vetor linha para a amostra i e $\overline{Xn}$ é a medida média para a classe $n$. (Brereton, 2003)

### 2.4.4 Modelagem Independente Suave de Analogia de Classe (SIMCA)

O método de análise SIMCA, foi defendido pela primeira vez por S. Wold, no início dos anos 1970. Esta metodologia é considerada como uma forma de modelagem suave usada no reconhecimento de padrões químicos. Isso embora tenha a existência de algumas diferenças com a análise discriminante linear empregada na estatística tradicional, essa distinção não é tão radical quanto muitos poderiam acreditar. A ideia de modelagem suave pode ser explicada pelo fato de que duas classes podem se sobrepor (e, portanto, são 'suave'), e não há problema com um objeto pertencer a mais de uma classe simultaneamente ou, então, a nenhuma. Na maioria das outras áreas da estatística, há uma insistência para que um objeto pertença a uma classe discreta, e assim vem dessa definição o conceito de modelagem rígida. (Brereton, 2003)Desta maneira, é possível calcular as distâncias de classe a partir da análise discriminante próximas a dois grupos ou mais. Em contrapartida, usando a análise discriminante clássica, todo o procedimento de modelagem deve ser repetido se um número extra de grupos for adicionado, uma vez que a matriz de variância-covariância agrupada deve ser calculada novamente.

$$X_i = T_i * P_i^T + E_i$$

Equação 5

Onde o valor de $X_i$é a amostra original, $T_i$ é a matriz de escores, $P_i$ representa as variáveis que serão consideradas no modelo e $E_i$ a matriz residual. Assim, a variância residual para uma classe do modelo pode ser calculada conforme a equação 6.

$$S_0^2 = \sum_{i=1}^{n}\sum_{j=1}^{p} \frac{e_{ij}^2}{(n - A - 1) * (p - A)}$$

Equação 6

Desta maneira, o símbolo $e_{ij}$ é definido como os resíduos de cada classe, $p$ é o número de variáveis, $n$ é o número de amostras, $A$ número de componentes principais de cada classe. Assim, é possível realizar o cálculo do modelo para classificação de classes. Pode-se definir que então o modelo SIMCA calcula um modelo de PCA para cada classe que é reconhecida pela metodologia, sendo a distância entre as classes representada por uma matriz de valores numéricos. (Brereton, 2003)

### 2.4.5 Mínimos Quadrados Parciais (PLS)

O PLS (*Partial Least Square Regression*) é uma técnica bastante interessante para análise de dados que possuam uma quantidade significativa de ruídos experimentais, colinearidades e que possam apresentar também não linearidades. Todas as variáveis que possuem relevância devem ser inseridas nos modelos via PLS, o que implica que a calibração pode ser realizada eficientemente mesmo na presença de interferentes, não havendo necessidade do conhecimento do número e natureza dos mesmos. O método PLS tem se tornado uma ferramenta extremamente útil e importante em muitos campos da química, como a físico-química, ainda no controle de inúmeros processos industriais. (Malegori et al., 2017) Algoritmos de decomposição cruzada encontram as relações fundamentais entre duas matrizes (X e Y). As quais tentarão encontrar a direção multidimensional no espaço X que explica a direção da variância multidimensional máxima no espaço Y. De uma maneira mais simplificada, o PLS projeta X e Y em um subespaço de dimensão inferior, de modo que a covariância entre transformado(X) e transformado(Y) seja máxima. A equação do PLS pode ser descrita da seguinte maneira:

$$Y = \sum_{i \in \Omega_{NIR}} (A_i * X_i + B_i)$$

Equação 7

Onde:

$\Omega_{NIR}$é o conjunto de números de onda medidos pelo espectrômetro;

$X$ é o vetor que representa a absorbância da amostra para cada número de onda pertencente a $\Omega_{NIR}$;

$Y$ é o vetor que representa as classes preditas de cada amostra. Para a regressão desenvolvida, numericamente foi estabelecido que:

1 = grãos esverdeados

2 = grãos bons

3 = grãos ardidos

4 = grãos queimados

$A$ e $B$ são, respectivamente, os vetores de coeficientes angulares e lineares, a serem calculados pelo algoritmo do PLS de forma a minimizar o erro médio quadrático (MSE) da predição em relação aos valores reais.

O PLS é uma forma de regressão linear regularizada onde o número de componentes controla a força da regularização. Desta maneira, o modelo computa um valor de y real (y_true) e y predito (y_predicted), e a saída do modelo pode ser entendido da seguinte maneira:

Para um melhor entendimento da equação do $R^2$ sua representação de uma maneira de matemática global é a seguinte:

$$R^2 = 1 - \frac{\sum_{n=1}^{N}(y_{true} - y_{pred})^2}{\sum_{n=1}^{N}(y_{true} - y_{true,mean})^2}$$

Equação 8

Os valores considerados para N é a quantidade de amostras a serem analisadas, ou seja, inseridas no modelo. A melhor pontuação possível para o $R^2$ é 1,0 e a mesma também pode ser negativa, demonstrando que o modelo é pior ou o não indicado para a aplicação dos dados a serem tratados. Um modelo com caracterizado como constante é aquele que sempre prevê o valor esperado de y, desconsiderando os recursos de

entrada, obteria uma pontuação de 0,0. (Scikit-learn developers, 2023b; Yang et al., 2022)

Uma outra forma de equação matemática importante para análise do desempenho do PLS é a raíz quadrada do erro, nomeada RMSE. Representada na equação 11.

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N}(y_{true} - y_{pred})^2}{N}}$$

Equação 9

O RMSE se comporta de maneira inversa ao $R^2$ onde para um bom resultado, o RMSE deve ser menor a medida que o $R^2$. (Yang et al., 2022)

### 2.4.6 Regressão Logística

A Regressão Logística é uma técnica estatística aplicada para análise e mineração de dados, o qual busca a elaboração de um modelo que resulte na predição de dados, seja eles utilizando dados iniciais aplicados com variáveis binárias, ou variáveis com multicomponentes. Este modelo é utilizado também para aplicações em modelos de aprendizagem de máquinas (*Machine Learning*), pois ajuda a criar predições mais precisas para suas diversas aplicações. A grande diferença da regressão logística para uma regressão linear é que sua predição é categórica, ou seja, é capaz de separar as variáveis por uma categoria a qual representa (Gonzalez Azevedo, 2018). A equação capaz de traduzir a regressão logística de uma maneira geral, considerando um conjunto de variáveis pode ser expressa conforme a equação 12.

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n$$

Equação 10

Onde:

$x$ representa o conjunto de variáveis independentes, podendo variar de 1 até n variáveis, dependendo do sistema a ser modelado.

Quando se iguala o $g(x)$ a equação logarítmica, é descrito na equação:

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n$$

Equação 11

Onde:

$p$ é o que se estima na equação, que representa a probabilidade ou a predição do evento a ser equacionado.

Para isolar o valor de $p$ é necessário aplicar o antilogaritmo, expressa na equação 14.

$$\hat{p} = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\cdots+\beta_n x_n}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\cdots+\beta_n x_n}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\cdots+\beta_n x_n)}}$$

Equação 12

De uma maneira geral, esta é uma representação da equação logística com multicomponentes. (Gonzalez Azevedo, 2018)

Utilizando linguagem de programação Python é possível implementar a regressão logística aplicando a biblioteca *sckit-learn*, que contém *LogisticRegression.* Este modelo também pode ser usado em função binária e multicomponentes, depende dos dados a serem analisados e o resultado que se deseja obter. Dentro dessa função, há também os solvers que são possíveis de serem aplicados, sendo a escolha de acordo com a quantidade de dados e tipos dos dados a serem tratatados, sendo o solver nomeado ‘lbfgs’ o mais robusto, e que é utilizado em grande maioria das vezes. No entanto, além deste citado, existem outros, como ‘liblinear’,’newton-cg’, ‘newton-cholesky, ‘sag’ e ’saga’. Os solvers que possuem maior compatibilidade com o tipo (binário ou multicomponentes) e quantidade da amostragem de dados a serem tratados de predição das características do grão são: ‘lbfgs’e ‘liblinear’. Testando-os é possível identificar o que melhor se adapta e fornece melhores predições. (scikit-learn developers, 2023) Para tratamentos em grãos de soja, especificamente análise de características físico-químicas há poucos trabalhos utilizando essa modelagem

matemática para classificação das características dos grãos, no entanto, há trabalhos com utilização de regressão logística em outras aplicações. (Côrtes, 2022)

### 2.4.7 Máquina vetores de suporte (SVM)

Máquinas vetores de suporte (SVMs) podem ser definidos como um conjunto de métodos de aprendizado supervisionados que possuem aplicação para classificação, regressão e detecção de *outliers*. Algumas vantagens de utilizar esse método é sua eficácia em espaços dimensionais elevados, e nos casos em que o número de dimensões são maiores que o número de amostragem. Para realizar a classificação, o SVM minimiza o erro de classificação empírica e, de maneira simultânea maximiza a margem geométrica interclasse resultando em uma solução única. Um outro ponto positivo dos modelos SVM é que eles podem operar em um espaço de recursos induzido pelo *kernel*, permitindo modelagem não linear. Além disso, podem apresentar um resultado satisfatório com uma base de dados com uma quantidade limitada de dados. No entanto, as maiores dificuldades que podem ser apontadas para este modelo são: a otimização dos parâmetros e a dificuldade de interpretar o mesmo.(Devos et al., 2009; Scikit-learn developers, 2023c)

O modelo apresenta parâmetros configuráveis para classificações binárias lineares, e não lineares, além da classificação multicomponentes. Uma forma mais ampla de representar matematicamente o modelo não linear, pode ser interpretado com a equação 15.

$$f(x) = f\big(\emptyset(x)\big) = \sum_{i=1}^{n} y_i \propto_i \emptyset(x_i)\emptyset(x) + b \quad com\ 0 \leq \propto_i \geq C$$

Equação 13

O produto escalar $(x_i)\emptyset(x)$ é substituído pela função Kernel. Existem algumas funções que podem ser utilizadas dependendo do modelo e problema a ser utilizado e resolvido, dentre os mais utilizados e o que se encaixa no trabalho aqui apresentado é o kernel RBF, também pode ser descrito como base radial, isso porque p qualquer formato de contorno dos clusters da base de dados pode ser obtido com o RBF. (Devos et al., 2009; Scikit-learn developers, 2023c)

# 3 MATERIAIS E MÉTODOS

## 3.1 GRÃOS DE SOJA

Para realização deste trabalho foram utilizadas 6 amostras com 500 g de grãos de soja doadas por uma cooperativa e um armazém, ambos localizados no Estado do Paraná. As amostras recebidas eram constituídas por ampla variedade de grãos cobrindo todas as classes discutidas anteriormente (seção 2.2). Algumas das amostras são padrão exportação e outras, com qualidade inferior, são tipicamente encaminhadas para fabricação de ração animal.

## 3.2 METODOLOGIA PARA ANÁLISE DE DADOS

Os ensaios analíticos deste trabalho seguiram as seguintes etapas:

- Recebimento de uma amostra de grãos de soja de um terminal de recebimento e armazenagem do grão;
- Preparação da amostra dos grãos em tamanhos diferentes, para leitura de espectro;
- Análises de espectrofotometria utilizando um espectrofotômetro NIR;
- Pré-análise de dados com diferentes amostras verificando variações significativas de características físicas dos grãos;
- Avaliação estatística através de ferramentas baseadas em PCA, PLS, Regressão Logística e Máquina Vetores de Suporte de utilizando o software Unscrambler e o ambiente de programação em Python, com objetivo de identificar a melhor técnica de classificação dos grãos e utilização do software unscrambler para aplicação de modelos lineares.

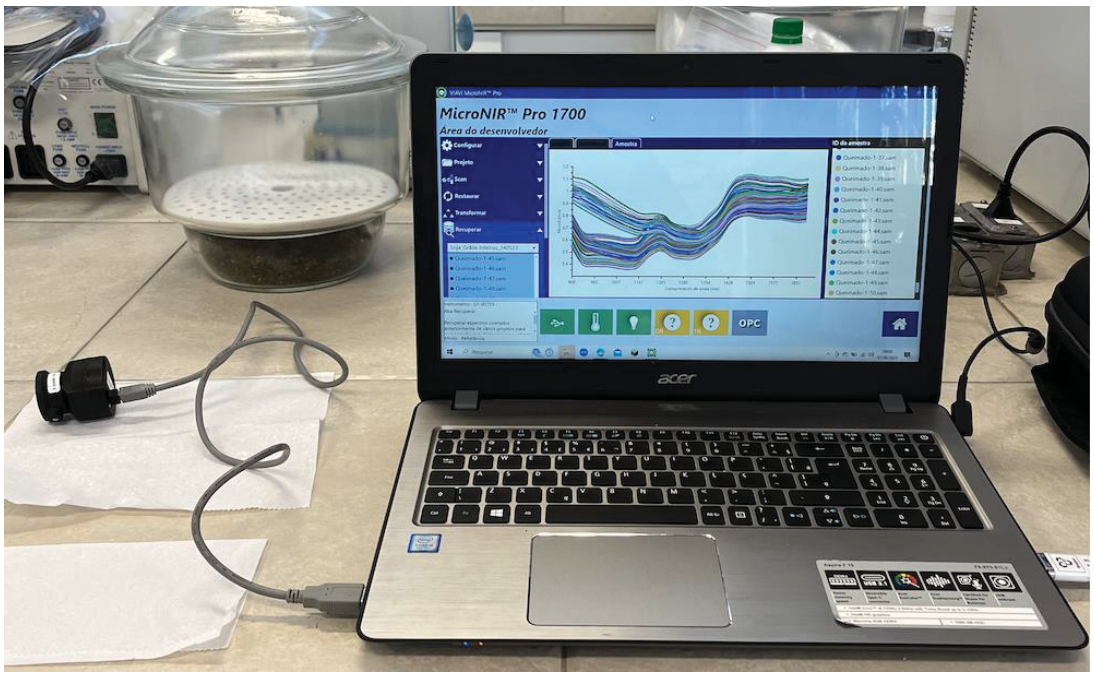
No Laboratório de Pesquisa Experimental I (LPE I/UFPR), foram realizadas análises de espectrofotometria utilizando um equipamento portátil, micro-NIR (Viavi *Solutions*). Os grãos de soja analisados correspondiam às 4 classificações

- Ardidos;
- Queimados;
- Bom;

- Esverdeados.

Os parâmetros foram medidos para as amostras de grãos de soja, sendo dessa maneira possível realizar a medição das amostras classificadas pela qualidade e notando a diferença das leituras, as quais são relacionadas as características dos grãos. (SENAR, 2017). As análises seguiram os procedimentos de que primeiramente, foi realizada a leitura de calibração do equipamento, onde é realizada uma leitura com o feixe de luz do MicroNIR em preto e depois no branco. Após o equipamento calibrado, as amostras foram colocadas em um cadinho, cobrindo cerca de metade com as amostras, a cada 5 leituras, mudando o direcional do equipamento. Esse procedimento de calibração foi repetido para cada características da amostra, resultando 50 leituras por amostras e 200 leituras totais. Posteriormente, foi realizada mais um total 40 leituras para as classes (10 para cada classe), o qual foi usado para validação dos dados, e mais um total de 19 leituras para grãos ardidos e queimados que apresentavam algumas características de grãos fungados. Foram utilizados um total de 59 leituras para predição do modelo.

FIGURA 10 - EQUIPAMENTO MICRO-NIR COM SOFTWARE




FONTE: O autor (2023).

FIGURA 11 - MICRO NIR


FONTE: O autor (2023).

Para leitura das amostras os grãos foram categorizados em 4 classes, para posterior classificação pelos modelos. As figuras das amostras dos grãos separados em classes representadas pelas Figuras 12 a 15.

FIGURA 12 - AMOSTRA DE GRÃOS ESVERDEADO

FONTE: O autor (2023).

FIGURA 13 - AMOSTRA DE GRÃOS BOM

FONTE: O autor (2023).

FIGURA 14 - AMOSTRA DE GRÃOS QUEIMADO

FONTE: O autor (2023).

FIGURA 15 - AMOSTRA DE GRÃOS ARDIDOS

FONTE: O autor (2023).

Para realizar as 200 leituras explicadas acima, foram utilizadas as categorias das 4 amostras ilustradas nas Figuras 12 a 15. E nas leituras das 59 amostras, que foram usadas para a predição do modelo, foi utilizado as amostras categorizadas nas Figuras 12 a 15, e mais as amostras nas Figuras 16 e 17. Um ponto é que, para realizar os grãos foram categorizados em 4 amostras, como no treinamento, e os grãos que se apresentavam com a característica de ardidos, mas com outras características foram categorizados como grãos ardidos, e o mesmo foi realizado para os grãos com características a mais que queimados, para separados das amostras de grãos queimado.

FIGURA 16 - AMOSTRA GRÃOS ARDIDOS COM OUTRAS CARACTERÍSTICAS

FONTE: O autor (2023).

FIGURA 17 - AMOSTRA GRÃOS QUEIMADOS COM OUTRAS CARACTERÍSTICAS

FONTE: O autor (2023).

## 3.3 MÉTODO DE ANÁLISE DE DADOS

Foram utilizadas amostras de grãos de soja, as quais foram submetidas a uma classificação de qualidade, realizada da maneira manual por um operador responsável pela classificação ao receber os grãos nas empresas antes do armazenamento, sendo os padrões de qualidade os descritos na portaria nº 282 do ministério da agricultura,

pecuária e abastecimento. Dentre as amostras utilizadas, foram realizadas análises iniciais antes da leitura do NIR, e posteriormente, foram analisados via ferramentas de classificação estatística descritos na seção 2.4 e detalhados a seguir para verificação das análises e plotagem dos gráficos.

### 3.3.1 Análises iniciais das amostras

Primeiramente, foram selecionados grãos de 6 amostras diferentes sem serem classificadas por qualidade segundo classificação enviada junto as amostras, as mesmas foram colocadas em um vidro de relógio, e então foram tiradas fotografias dessas amostras separadamente. Para esta etapa foram consideradas somente 6 amostras que apresentavam maiores diversidades de grãos. Posteriormente, as fotografias das amostras foram impressas para contabilização de grãos por diâmetro, classificação de qualidade e determinação da curva de distribuição de diâmetro de grão geral e por classe de qualidade. Para fins de visualização comparativa das distribuições adotou-se o seguinte procedimento:

Distribuição dos diâmetros de amostragem de acordo com as frequências de cada um;

Foram calculadas com equações do Excel: o mínimo, máximo, média, desvio padrão e incremento. Além disso foram utilizados:

$$valor\ calculado = \ média\ \pm 4 * \delta$$

Equação 14

Sendo $\delta$ o desvio padrão, e os valores calculados separadamente somando, apara obter um valor a e subtraindo para obtenção de um valor b.

O número de ponto é estimado de acordo com a amostragem de dados.

Os valores calculados de função de probabilidade de massa (FPM) são provenientes da equação do Excel de Distribuição Normal, onde considera-se a média, desvio padrão, o valor de x, sendo que estes são os valores calculados no item 2, e a seleção falso para distribuição cumulativa, pois assim é possível a construção da curva de distribuição.

Além do cálculo para a distribuição Gaussiana, foi também utilizada a relação direta dos valores obtidos pelas imagens dos grãos de soja para obtenção de curva de distribuição normal por amostra, utilizando-se dos diâmetros obtidos e fazendo a relação direta com a frequência, a qual é calculada de acordo com a Equação 15.

$$F = \left(\frac{Pd}{Pt}\right) * 100$$

Equação 15

Onde:
F = Frequência
Pd = Soma de grãos por diâmetro
Pt = Quantidade total de grãos por amostra

Uma terceira forma de obtenção de distribuição normal que pode ser empregada como comparativo com as demais se dá pelo software Minitab, o qual a partir de inserção de dados da distribuição dos diâmetros de amostragem de acordo com as frequências de cada, e utilizando-se da opção de histograma é possível obtenção da curva de distribuição, além de parâmetros de desvio padrão, média e quantidade de grãos por amostra (N). Os cálculos descritos foram utilizados com objetivo de pré análise para entendimento do comportamento dos dados da amostra, além de ser possível comparar e observar o comportamento das amostragens de acordo com os parâmetros de qualidade do grão que foi observado. Vale ressaltar que os parâmetros foram classificados de acordo com conhecimento adquirido na literatura, comparando com a imagem impressa.

## 3.4 METODOLOGIAS E ANÁLISES UTILIZANDO LINGUAGEM DE PROGRAMAÇAO

Com as leituras obtidas pelo espectrofotômetro, foram então plotadas as curvas utilizando linguagem de programação Python, sendo o ambiente de desenvolvimento utilizado para as aplicações deste trabalho foi o Jupyter Notebook para esta plotagem em gráfico dos espectros foram utilizadas as bibliotecas pandas, numpy e matplotlib. Para os resultados de PCA, PLS e SIMCA foi utilizado o software Unscrambler e para

os modelos SVM e RL foi utilizado a linguagem de programação Python. Sendo que, as plotagens foram realizadas tanto para as 200 leituras das 4 amostras, quanto para as 59 leituras das 4 amostras das características separadas.

Os modelos foram aplicados tanto para os dados dos espectros sem pré-tratamento, quanto para os espectros com pré-tratamento, o qual foi utilizada a primeira derivada. Neste caso, a aplicação da primeira derivada nos espectros NIR realiza a remoção do deslocamento e do termo linear dominante dos dados espectrais sendo esta metodologia uma das primeiras tentativas de corrigir as variações de tamanho de partícula. Desta maneira, aplicando a derivada aos dados busca-se compensar parcialmente o deslocamento da linha de base entre as amostras e reduzir os efeitos que são provenientes do instrumento de leitura do espectro. Frequentemente, o uso mais importante de derivadas é obter informações adicionais do espectro, potencializando a capacidade de análise de um modelo matemático a ser aplicado nos dados. (Burns & Ciurczak, 2008)

### 3.4.1 PCA (Análise de Componente Principal)

Para aplicar a metodologia de PCA foi utilizado o software Unscrambler. O PCA foi aplicado aos dados espectrais dos grãos já separados por característica, com o intuito de avaliar a sensibilidade da técnica às diferentes classes de grãos de soja e uma possível correlação dos dados. Para classificação o modelo de PCA foi aplicado juntamente com o LDA, visando a partir de uma análise discriminatória aumentar o desempenho do modelo. Neste caso, visando enriquecer os comparativos de análises e entender o comportamento obtido pela leitura foi utilizada a análise discriminatória linear quadrática de Mahalanobis. Assim, o método foi aplicado para as 200 amostras considerando tanto os espectros brutos, quanto os espectros tratados com primeira derivada, com o objetivo de verificar a influência do pré-tratamento nas variações dos dados espectrais.

### 3.4.2 Mínimos Quadrados Parciais (PLS - *Partial Least Square*)

Na aplicação do PLS foi utilizado o software Unscrambler. O modelo foi aplicado tanto para as 200 leituras para treinamento, e 59 para predição, considerando os dados espectrais brutos e tratados com primeira derivada. Após a aplicação, foi utilizada uma matriz de confusão para análise dos resultados. A partir da mesma foi possível entender o comportamento do modelo. Desta maneira é possível analisar tanto as características que o modelo teve uma maior assertividade, quanto a menor. O modelo de PLS foi aplicado juntamente com o modelo de LDA, visando aumentar a assertividade do modelo utilizando uma análise discriminatória.

### 3.4.3 SIMCA

Para aplicação do modelo SIMCA foi utilizado o software Unscrambler. O modelo foi treinado com as 200 leituras separadas em 4 classes (esverdeado, bom, ardido, queimado) e posteriormente foi realizada a predição com as 59 leituras. Para as amostras de treinamento e predição foi elaborada uma matriz de confusão com avaliação de resultados por classe para melhor entendimento do desempenho do modelo para os dados usados na entrada.

### 3.4.4 Regressão Logística (RL)

O modelo de regressão logística foi investigado com base em uma sub-rotina elaborada via linguagem de programação Python, capaz de tratar os dados espectrais obtidos através do micro-NIR. O modelo foi aplicado para as 200 leituras (brutas e primeira derivada) do conjunto de treinamento, e para as 59 leituras (brutas e primeira derivada) do conjunto de predição. A partir do resultado utilizou-se o *classification_report*, para analisar o desempenho e a matriz de confusão.

### 3.4.5 SVM

Para aplicação do modelo de SVM foram utilizadas as bibliotecas pandas, numpy, sklearn.svm e posteriormente, foi utilizado o *classification_report* (resumo da classificação), para análise do desempenho. O modelo foi aplicado para as 200 leituras do conjunto de treinamento e as 59 leituras do conjunto de predição, considerando dados espectrais brutos e tratados com primeira derivada. Conforme reconhecido na literatura, a dimensão do conjunto de dados influencia o comportamento dos modelos de classificação. Esse modelo já possui como saída a matriz de confusão para análise do resultado.

# 4 RESULTADOS

Para se avaliar o desempenho da espectrofotometria NIR na classificação dos grãos de soja, foram selecionadas amostras de diferentes classes, as quais foram submetidas a ensaios de determinação de distribuição de tamanho (DTG) e leitura espectral. Os diferentes algoritmos de calibração e processamento de espectros relacionados nas seções anteriores desta dissertação foram aplicados. Além disso, As métricas de classificação utilizadas na avaliação da performance dos modelos investigados via linguagem Python (PLS, Regressão Logística e SVM) foram a precisão, *recall e F1score*. As definições para cada uma dessas são as seguintes:

- Precisão: é a quantidade de predições positivas que o modelo acertou, em relação às predições positivas totais feitas pelo modelo.
- *Recall:* é a quantidade de predições positivas corretas em relação aos positivos reais.
- *F1score:* representa a média harmônica ponderada dos valores da precisão e do *recall.* Do valor resultante, quanto mais próximo de 1 significa que melhor será o modelo. (Scikit-learn developers, 2023a).

Para analisar o desempenho dos modelos neste trabalho, um dos tópicos utilizados para modelos não lineares será o *F1score* principalmente, pois é a quantidade de acertos do modelo.

Diferentes métricas para avaliação de desempenho de métodos lineares foram utilizadas com base no software Unscrambler, incluindo a sensibilidade, especificidade e exatidão do modelo. A partir dos valores obtidos da exatidão, pode ser então calculado o erro do modelo. Desta maneira é possível definir:

- Sensibilidade: capacidade de um modelo de identificar corretamente os casos positivos, ou seja, proporção de casos positivos no modelo em relação a todos os dados positivos;
- Especificidade: habilidade do modelo na detecção dos casos negativos. Proporção de casos em que o resultado é negativo, pelo total de valores negativos do modelo.

- Exatidão: é a fração de resultados corretos do modelo em relação ao total de dados inseridos como entrada. Ou seja, são os resultados corretos em relação ao total de casos inseridos.

Para os cálculos das métricas dos resultados utilizam-se as seguintes equações.

$$Exatidão = \frac{VP + VN}{VP + FP + FN + VN} * 100$$

Equação 16

$$Sensibilidade = \frac{VP}{VP + FN} * 100$$

Equação 17

$$Especificidade = \frac{VN}{VN + FP}$$

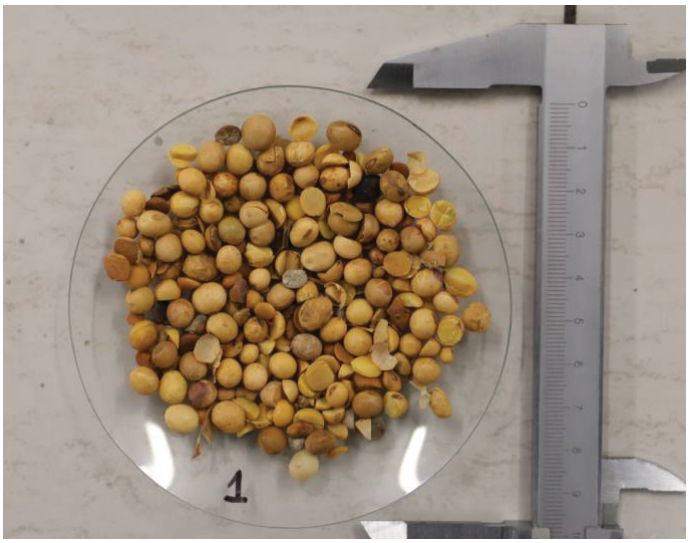
Equação 18

Onde o $VP$ representa os valores encontrados de verdadeiros positivos, $VN$ são os valores encontrados de verdadeiros negativos, $FN$ são os valores encontrados de falsos negativos e $FP$ são os valores encontrados de falsos positivos. (Kaufmann et al., 2022)

## 4.1 DISTRIBUIÇÃO DE TAMANHO DE GRÃO

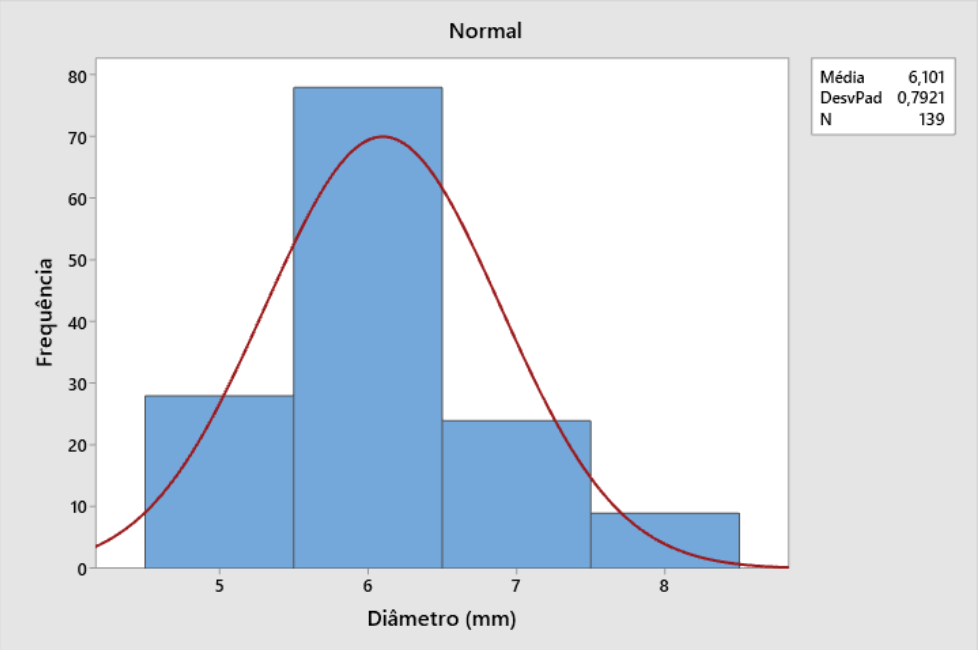
As amostras de grãos de soja foram analisadas através de fotografias quanto à distribuição de tamanho de grãos (DTG), através de contagem manual considerando entre 100-200 grãos. Através de uma planilha eletrônica, os dados experimentais de DTG descritos na forma de histograma foram comparados ao modelo de distribuição normal (curva Gaussiana). Os resultados para a Amostra 1 são exibidos nas Figuras 18 e 19.

FIGURA 18 - GRÃOS DE SOJA: AMOSTRA 1




FONTE: O autor (2023).

FIGURA 19 - DISTRIBUIÇÃO NORMAL AMOSTRA 1 - MINITAB



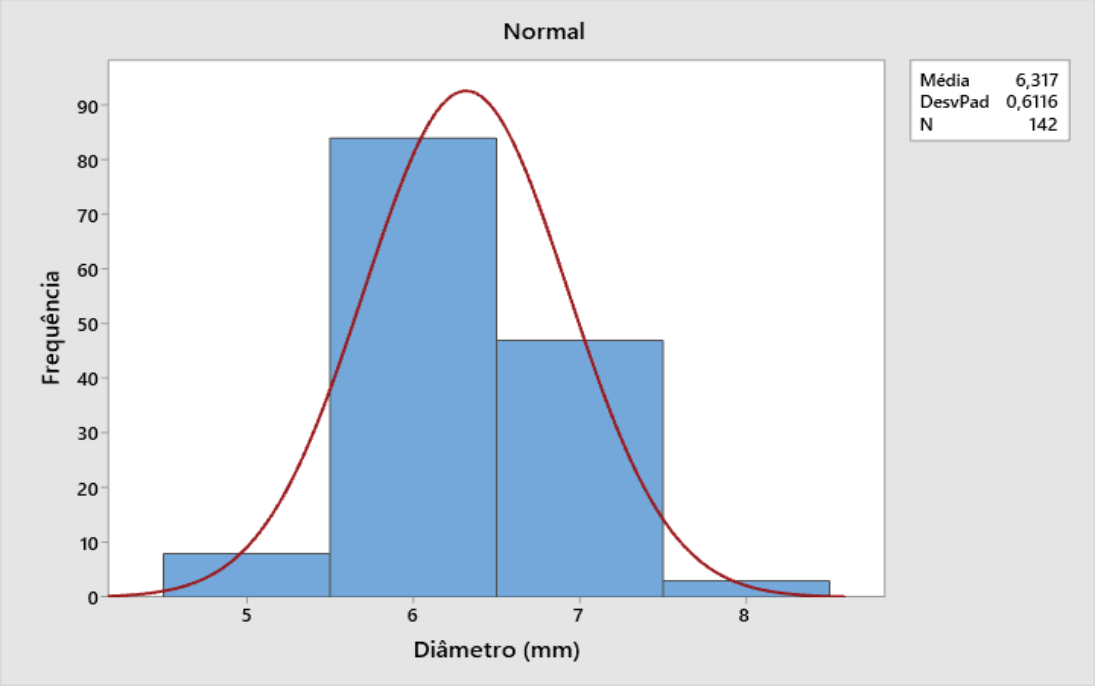

FONTE: O autor (2023).

Percebe-se que, a maior concentração de diâmetro dos grãos na amostra 1 foi de 6 mm.

FIGURA 20 - GRÃOS DE SOJA: AMOSTRA 2


FONTE: O autor (2023).

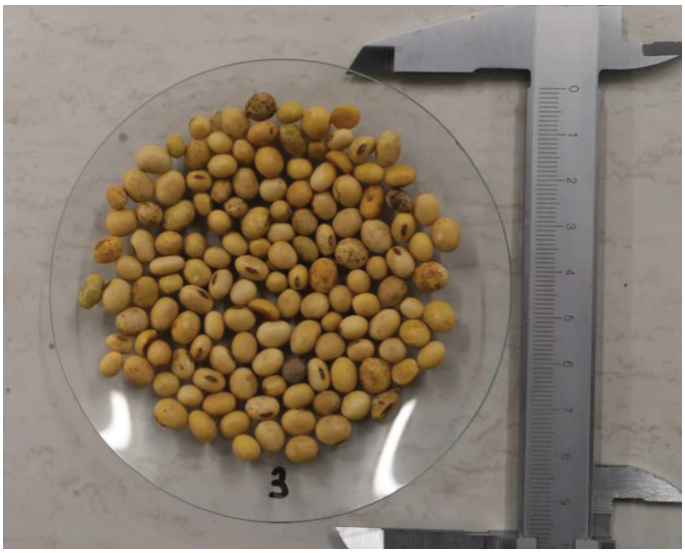
FIGURA 21 - DISTRIBUIÇÃO NORMAL AMOSTRA 2 - MINITAB




FONTE: O autor (2023).

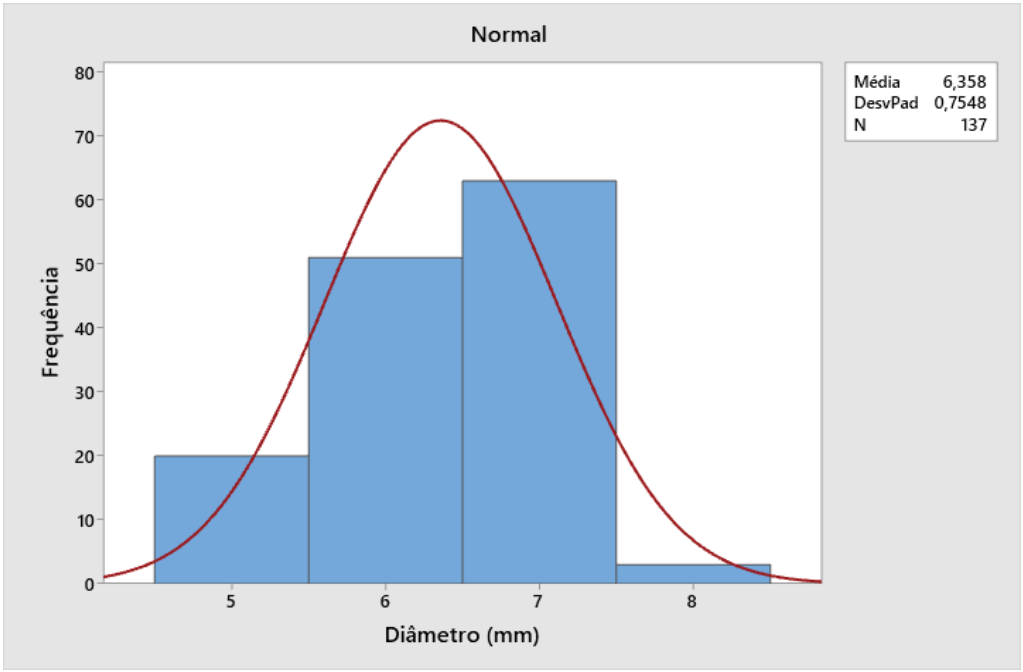
Para a amostra 2 (Figuras 20 e 21) há uma distribuição em relação ao diâmetro dos grãos parecida com a amostra 1, na qual, a maior frequência se encontra entre os diâmetros 6 e 7.

FIGURA 22 - GRÃOS DE SOJA: AMOSTRA 3




FONTE: O autor (2023).

FIGURA 23 - DISTRIBUIÇÃO NORMAL AMOSTRA 3 - MINITAB



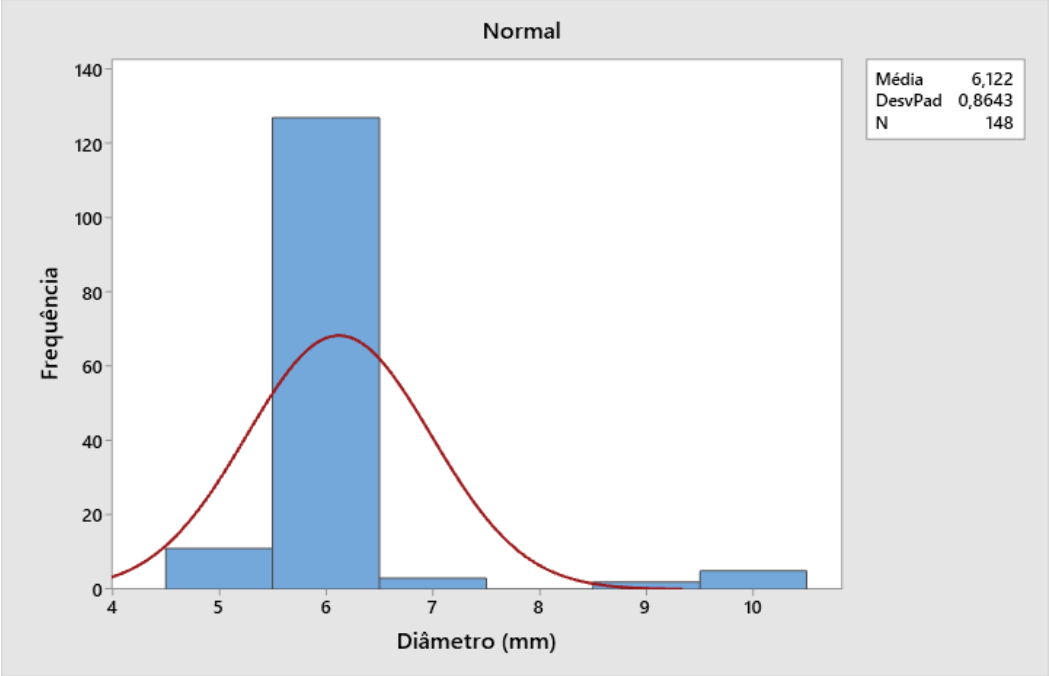

FONTE: O autor (2023).

Na amostra 3 (Figuras 22 e 23) os grãos apresentam uma frequência maior no diâmetro de 7mm. Além disso, quando analisadas as frequências os comportamentos são bem similares.

FIGURA 24 - GRÃOS DE SOJA: AMOSTRA 4


FONTE: O autor (2023).

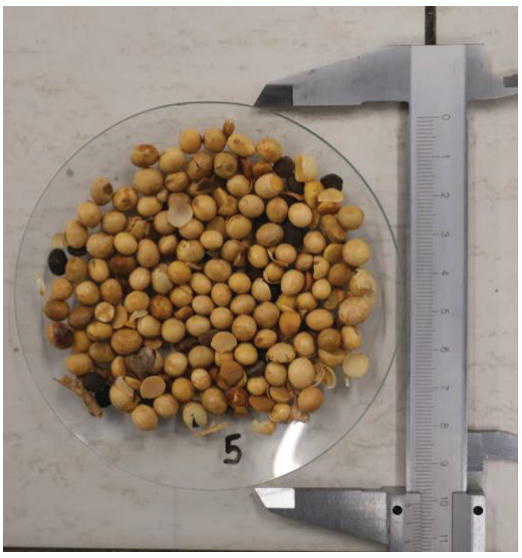
FIGURA 25 - DISTRIBUIÇÃO NORMAL AMOSTRA 4 - MINITAB




FONTE: O autor (2023).

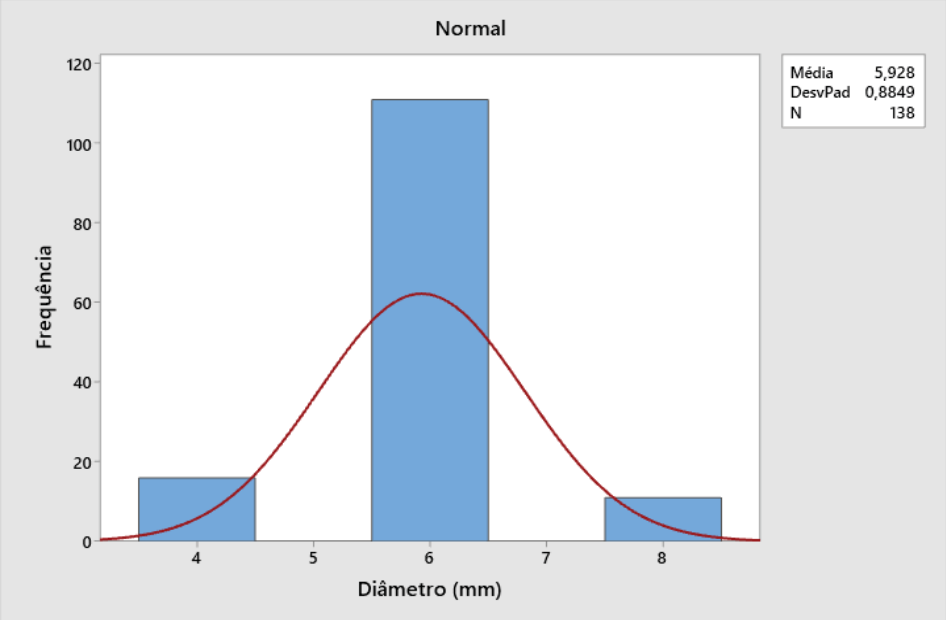
Para a amostra 4 (Figuras 24 e 25), o diâmetro de 6 mm se destaca com maior frequência, principalmente quando comparamos com a Figura 16. Tal amostra ainda apresenta grãos com diâmetros maiores, chegando até a 10 mm, resultando em uma curva de DTG larga, com desvio-padrão superior a 0,8 mm.

FIGURA 26 - GRÃOS DE SOJA: AMOSTRA 5




FONTE: O autor (2023).

FIGURA 27 - DISTRIBUIÇÃO NORMAL AMOSTRA 5 - MINITAB



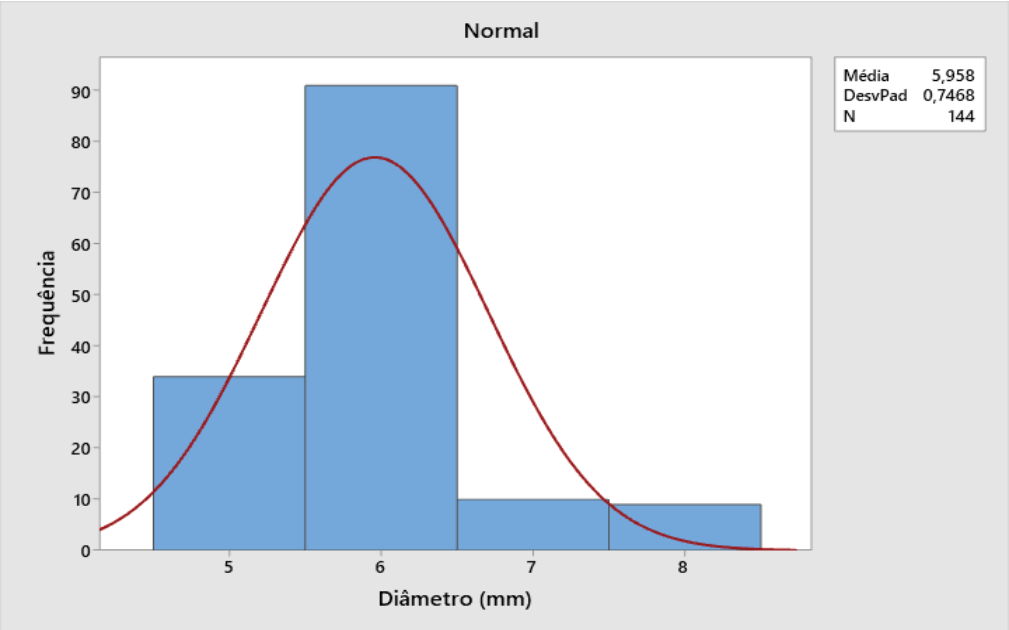

FONTE: O autor (2023).

A amostra 5 (Figuras 26 e 27) apresenta a maior frequência de amostragem no diâmetro 6 mm, mas também inclui grãos pequenos e grandes, resultando em ampla curva de DTG e desvio-padrão superior a 0,8mm. Trata-se da amostra com maior desvio-padrão.

FIGURA 28 - GRÃOS DE SOJA: AMOSTRA 6

FONTE: O autor (2023).

FIGURA 29 - DISTRIBUIÇÃO NORMAL AMOSTRA 6 - MINITAB



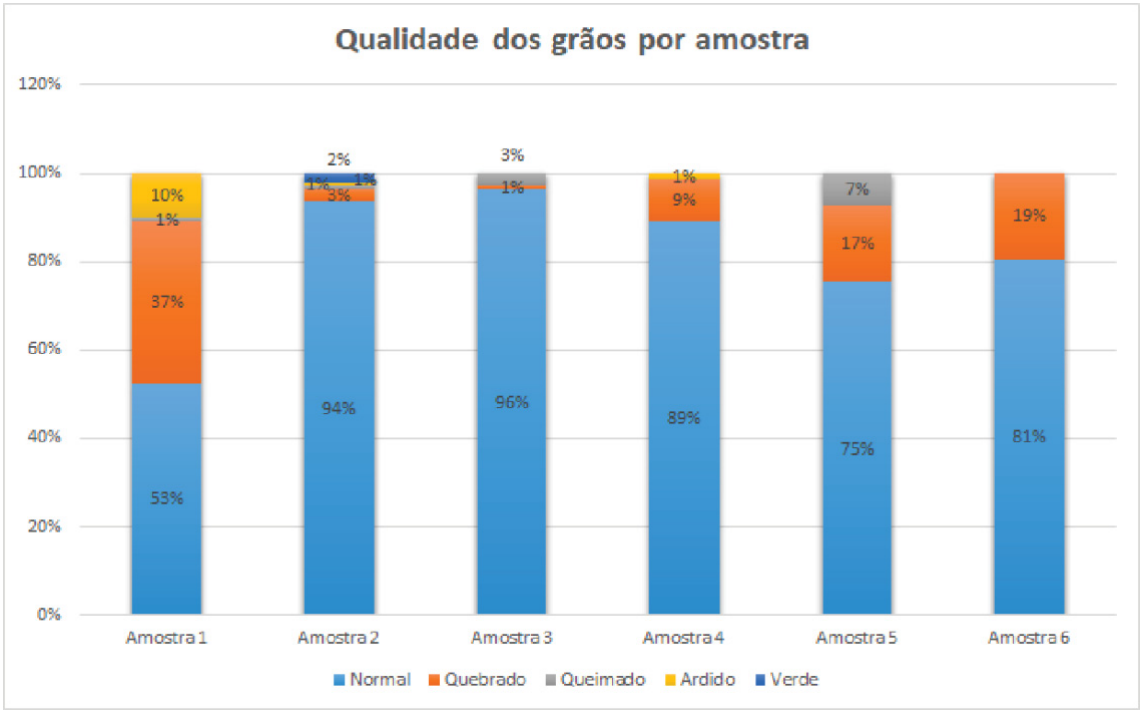

FONTE: O autor (2023).

A amostra 6 (Figuras 28 e 29) apresenta uma grande concentração no diâmetro 6 mm, assim como pode-se observar nas amostras anteriores, no entanto, a frequência é maior do que nas demais, não apresentando uma distribuição tão suave para outros diâmetros. O pico de frequência para o diâmetro de 6 mm é observado em todas as características mapeadas dos grãos dessa amostra.

Desta maneira, pode-se observar que em todas as amostras o diâmetro de 6 mm apresenta uma frequência alta. Em algumas amostras, há uma maior quantidade de variações referente à qualidade do grão, apresentando grãos ardidos, queimados e quebrados do que em outras. Esses pontos de qualidade podem ocorrer em grãos com qualquer diâmetro, contudo como nessas amostras há uma maior quantidade de grãos com 6 mm, a chance de obtê-los é maior. Já a calibração do sistema de imagem é feita baseada nos parâmetros e características disponibilizados pela portaria do Ministério da Agricultura.

De modo geral, as variações de diâmetro de grão observadas não são específicas por classe. Entretanto, os grãos com a característica normal, apresentam como maioria diâmetro variando entre 5mm e 6mm, o que equivale aos valores encontrados em safras de 2015/2016 e 2016/2017 de grãos colhidos na região médio-norte do Mato Grosso (Onetta & Ruffato, 2018).

FIGURA 30 - QUALIDADE DO GRÃO POR AMOSTRA




FONTE: O autor (2023).

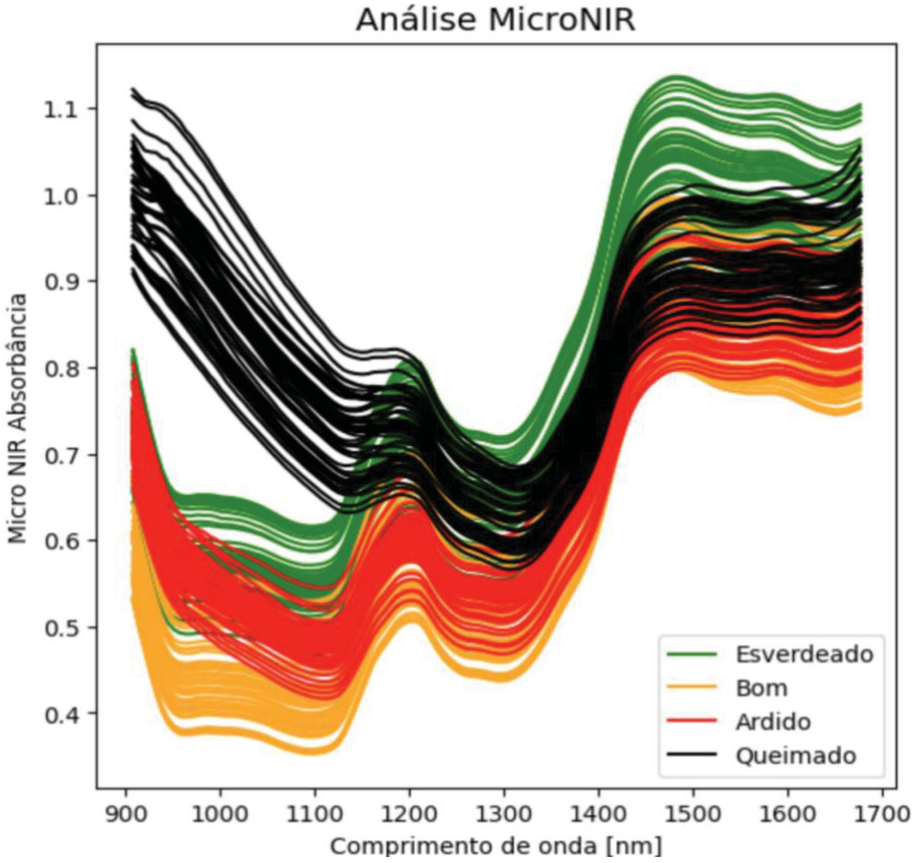
As quantidades de grãos caracterizados como normais são a grande maioria das amostras, o que de fato representa algo esperado. No entanto, na amostragem 1, as quantidades de grãos quebrados se destacam, além de possuir uma certa quantidade de grãos com característica de ardidos. As amostras 4, 5 e 6 apresentam uma quantidade razoável de grãos caracterizados como quebrados, sendo respectivamente uma frequência de 9,45%, 17,39% e 19,44%. Mesmo com frequências significativas, as amostras se enquadram dentro do limite tolerado, apresentado na Tabela 3. Outra relação para essa variação de características pode ser notada pela variação do tamanho diâmetro da semente, que ao possuírem um tamanho maior em conjunção ao choque que ocorre durante o transporte, pode gerar uma maior quebra.

## 4.2 ESPECTROS OBTIDOS VIA MICRO-NIR

A partir das 6 amostras estudadas anteriormente, foram separados visualmente cerca de 100 grãos para cada uma das 4 classes investigadas neste trabalho: Bom, Ardido, Queimado e Esverdeados. Foram coletados espectros para cada categoria. A

partir dessas leituras, realizadas 50 vezes para cada classe foi possível obter curvas espectrais representadas na Figura 31.

FIGURA 31 - ESPECTRO DOS GRÃOS DE SOJA CLASSIFICADOS POR QUALIDADE

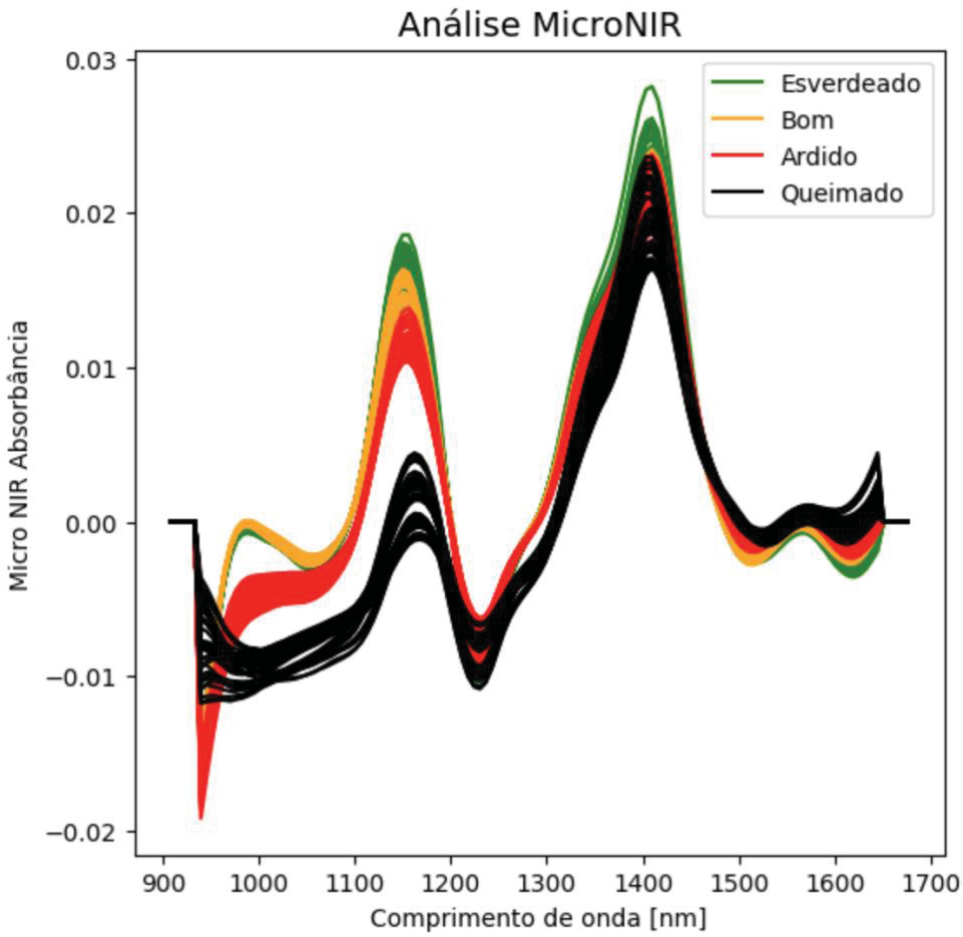



FONTE: O autor (2023).

Analisando a Figura 31 é possível observar que em alguns comprimentos de onda ocorre sobreposição espectral das classes, como por exemplo no comprimento de onda de 900 a 1100 nm, onde esverdeados, bons e ardidos coincidem. Cabe destacar que se reconhece na literatura que no espectro NIR é comum haver interferência de umidade, o que poderia então explicar a sobreposição gráfica em alguns pontos (dos Santos et al., 2018). Ainda assim, é possível identificar faixas de onda com padrão espectral específico por classe. Esta identificação torna-se mais acurada com o uso de algoritmos de classificação discutidos nas seções anteriores.

Com o objetivo de tratar os dados brutos apresentados na Figura 31 foi aplicada a primeira derivada nos dados espectrais, conforme demonstrado na Figura 32:

FIGURA 32 - ESPECTRO DA PRIMEIRA DERIVADA DOS GRÃOS DE SOJA CLASSIFICADOS POR QUALIDADE



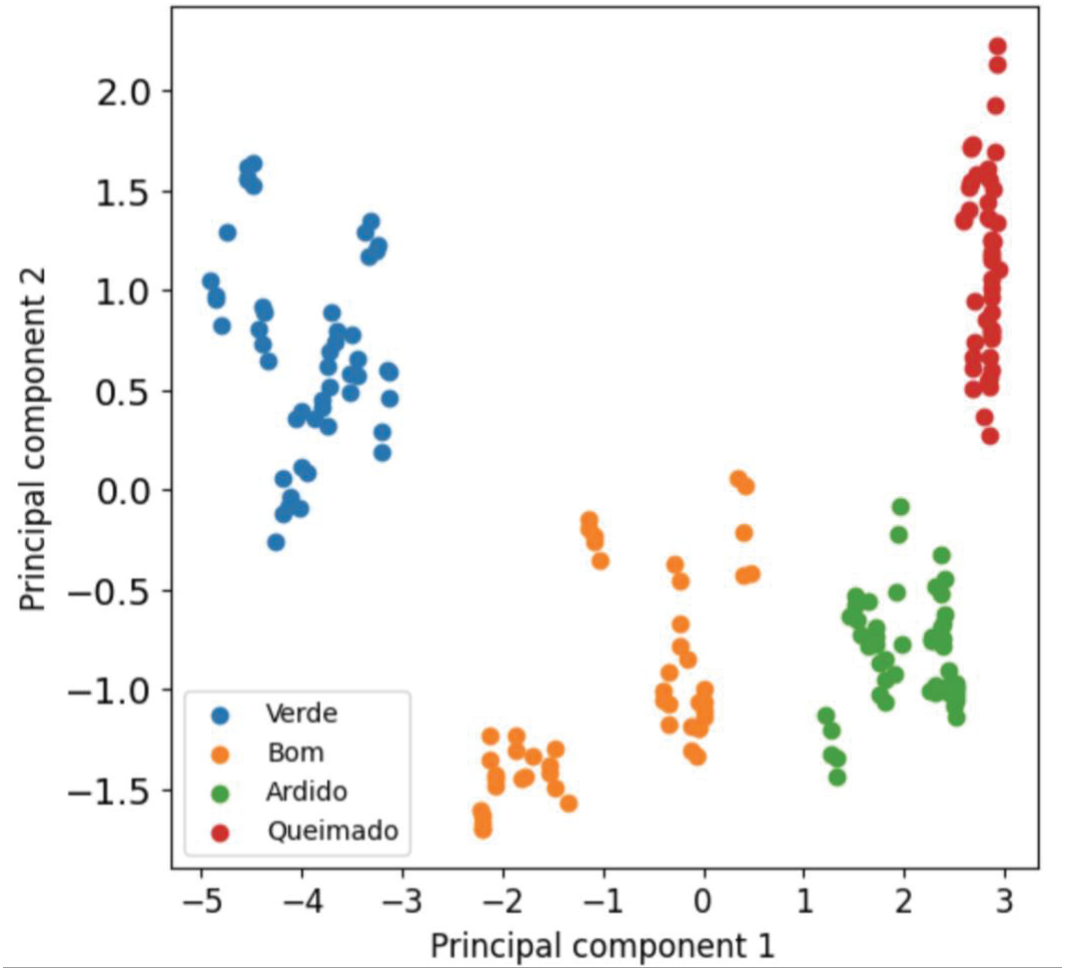

FONTE: O autor (2023).

Com aplicação do pré-tratamento, o qual é a primeira derivada dos espectros, observam-se maiores diferenciações nas ondas das quatro categorias principalmente no pico em que se encontra no intervalo do comprimento de onda de 1100 a 1200 nm. A primeira derivada retira os principais efeitos das características físicas dos grãos, enfatizando assim as informações químicas do material analisado. Por esse motivo as curvas se tornam mais concentradas e com picos bem proeminentes, ressaltando melhor as diferenças entre as amostras, conforme discussão teórica no capítulo 2.2 deste trabalho.

### 4.2.1 Análise de Componente Principal (PCA)

Para obtenção dos dados de PCA dos espectros do micro-NIR foi utilizado o software Unscrambler, o qual permite a geração de modelos estatísticos na forma de matrizes de pesos e escores. Neste software foi necessário adicionar o valor de entrada de 5 principais componentes, no qual, o modelo convergiu para 2 PCs utilizando os dados dos grãos sem tratamento prévio, ou seja, os dados espectrais brutos.

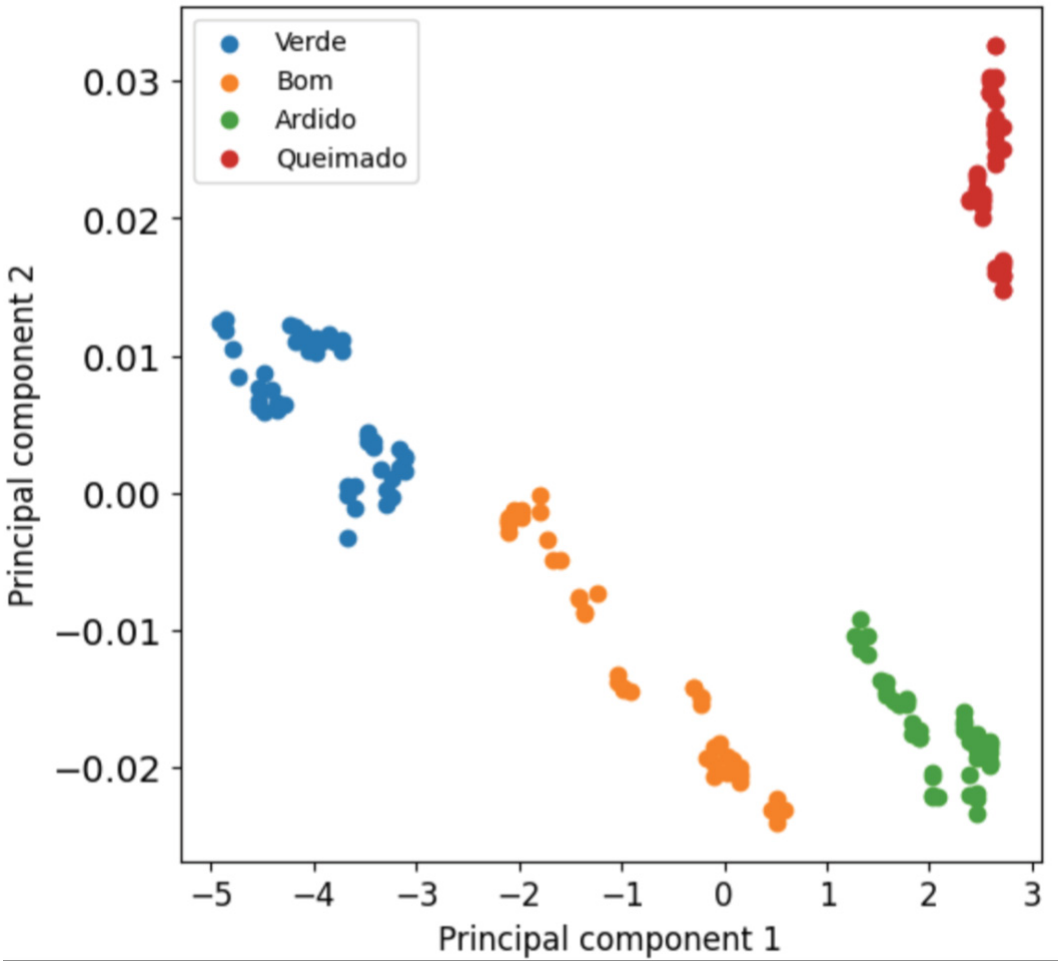
FIGURA 33 - GRÁFICO PCA PARA DADOS BRUTOS DA LEITURA DOS GRÃOS




FONTE: O autor (2023).

Realizando um pré-tratamento dos dados, o qual neste caso foi a aplicação da primeira derivada, obteve-se o resultado representado na Figura 34.

FIGURA 34 - GRÁFICO PCA PARA DADOS DA DERIVADA DA LEITURA DOS GRÃOS



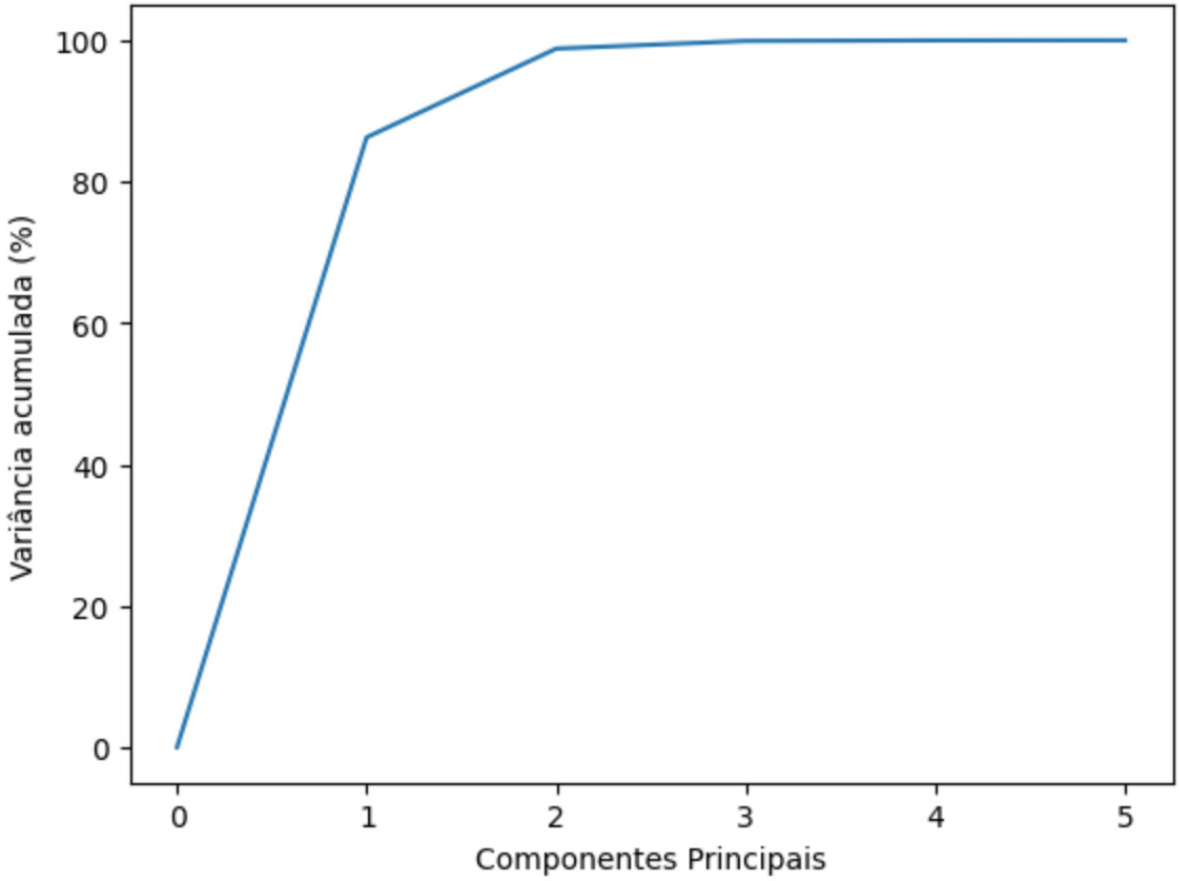

FONTE: O autor (2023).

Analisando o gráfico dos componentes principais (PCA) das amostras para espectros brutos é possível analisar que os clusters das amostras ficam bem isolados. Os grãos ardidos possuem maior variância com o PC1 e os grãos verdes com o PC2. Os grãos bons ficam bem divididos com o PC1 e PC2, já os grãos ardidos se mostram com uma boa disposição tanto para PC1 quanto para PC2, apesar de que os valores de variação são um pouco maiores no PC1.

Para a FIGURA 34, onde é apresentado o PCA a partir dos dados da primeira derivada, percebe-se que os valores ficam melhor classificados, os agrupamentos possuem maiores correlações com o PC1. Isso se deve ao fato que esse tratamento remove efeitos tipicamente físicos, deixando mais evidente as características químicas dos grãos, enfatizando, portanto, uma maior distinção química entre os mesmos.

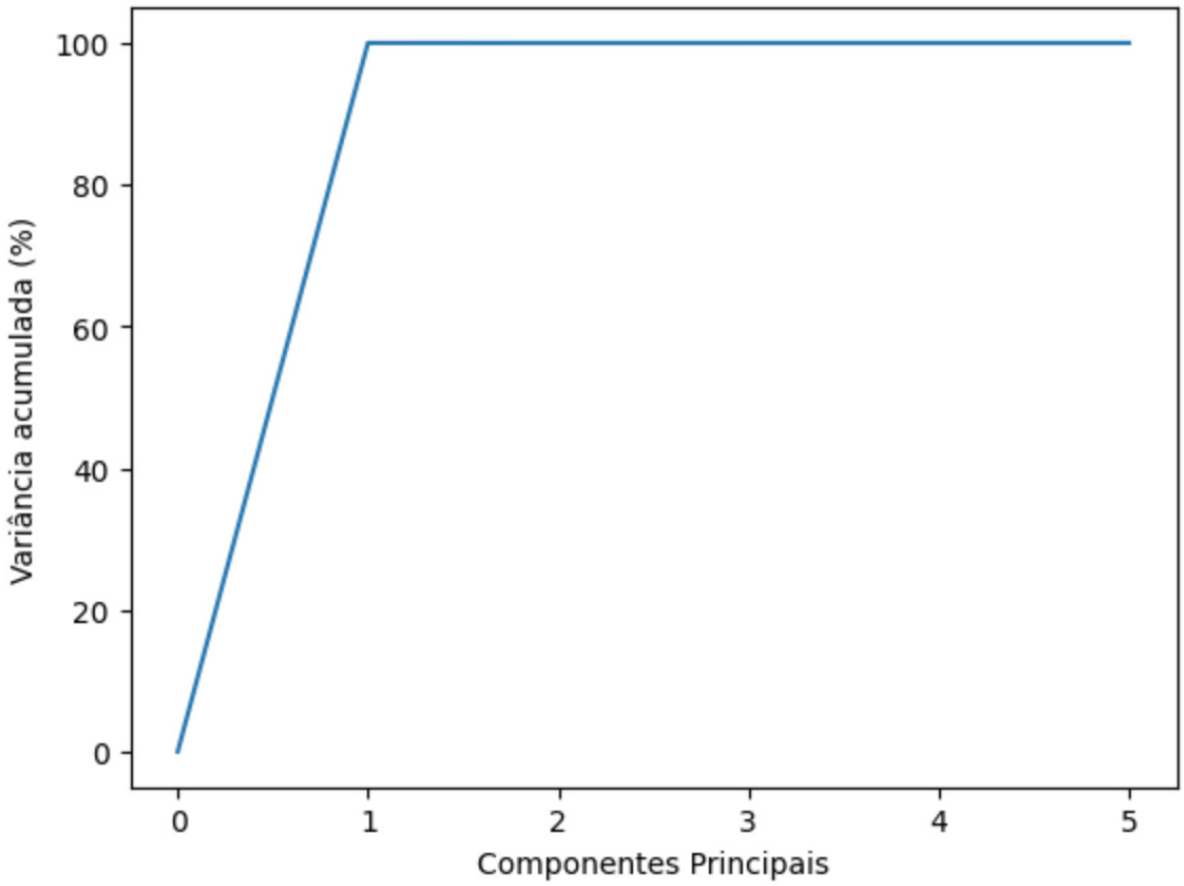
A FIGURA 35 ilustra a capacidade de explicação da modelagem PCA com a adição de componentes principais para o caso de dados espectrais brutos. Observa-se neste caso que 100% da variância dos dados é explicada com 2 componentes principais. Ou seja, para encontrar as direções de máxima variância para os dados brutos, foram necessários 2 componentes principais.

FIGURA 35 – CONVERSÃO DE COMPONENTES DOS DADOS BRUTOS




FONTE: O autor (2023).

FIGURA 36 - CONVERSÃO DE COMPONENTES DOS DADOS DA DERIVADA

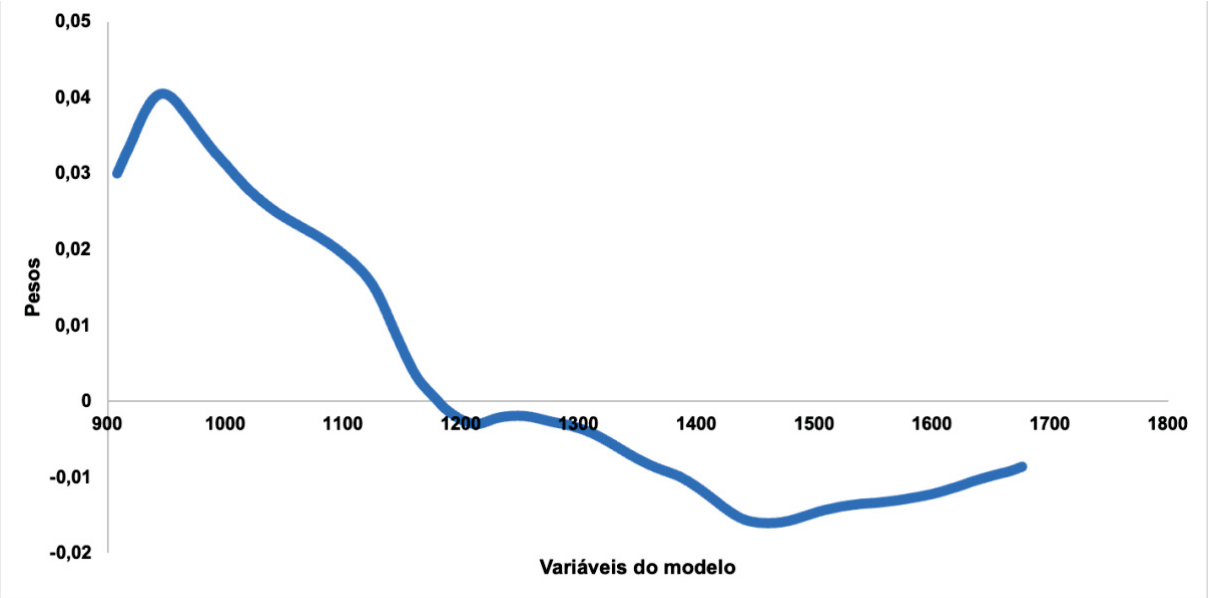



FONTE: O autor (2023).

Já a FIGURA 36, a qual representa a variância acumulada para o caso de espectros tratados com primeira derivada, mostra que apenas 1 componente principal é suficiente para completa descrição da variância dos dados. Como os dados estão tratados, sem maiores influências de outras características, foi possível encontrar a máxima variância em um componente único, representando que os clusters estão bem definidos e não há sobreposição.

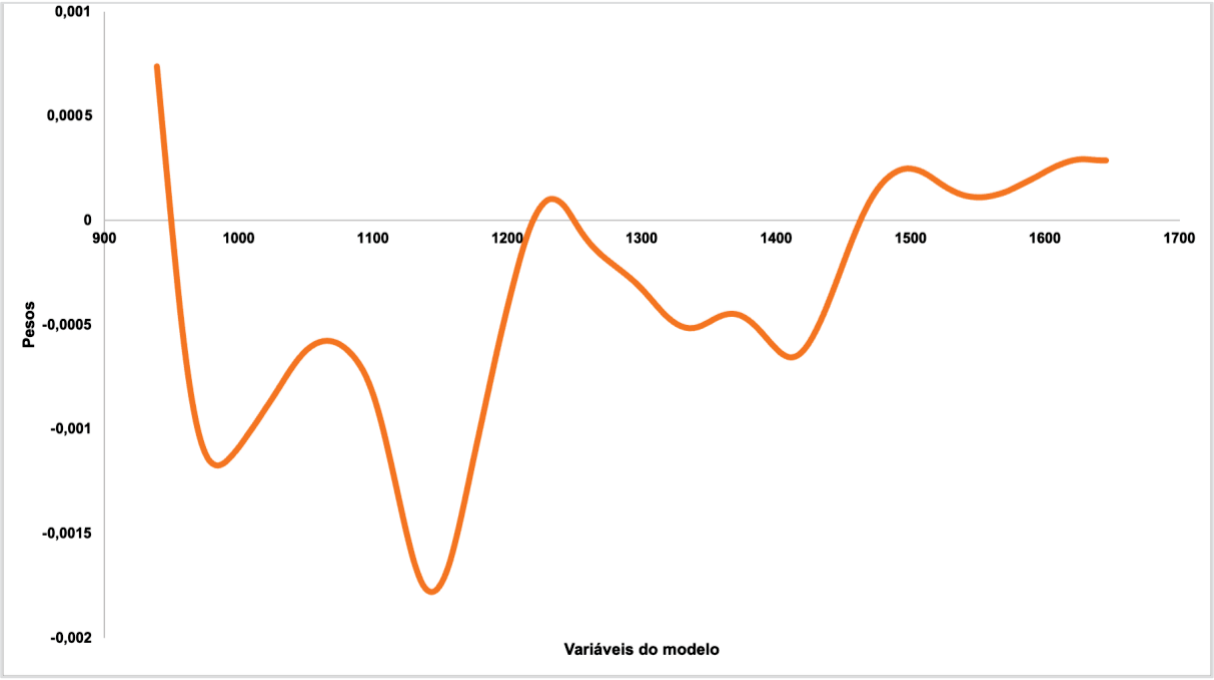
Cabe destacar que em outros trabalhos observam-se maiores quantidades de PCs para a plena explicação da variância dos dados indicando a maior variabilidade das amostras para os estudos em questão (Wang et al., 2021) Para este trabalho, os espectros tratados com primeira derivada apresentaram bons resultados para o modelo não supervisionado, estabelecendo assim boas perspectivas de aplicação para fins de classificação de grãos.

FIGURA 37 - PESOS DO MODELO POR VARIÁVEL PARA PC1




FONTE: O autor (2023).

FIGURA 38 - PESOS DO MODELO POR VARIÁVEL - DERIVADA PARA PC1



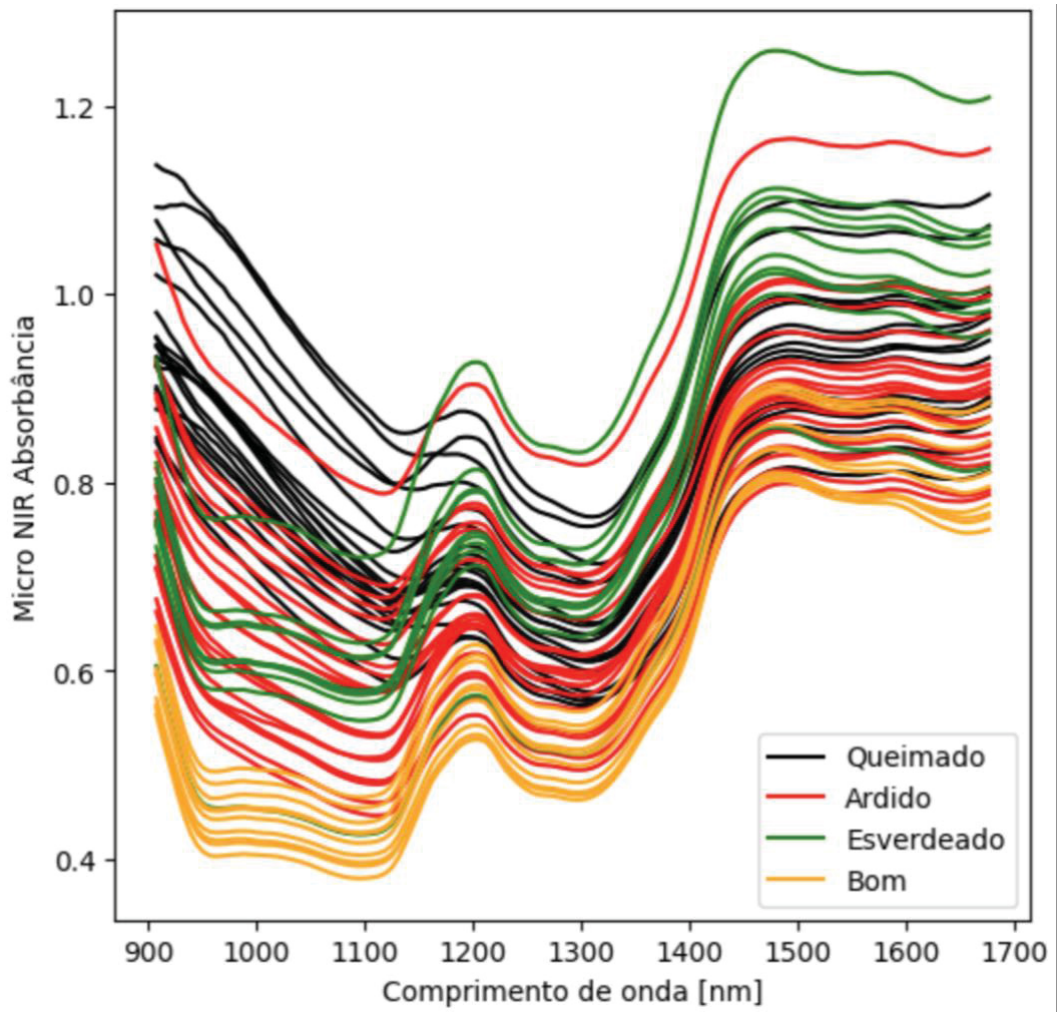

FONTE: O autor (2023).

Na FIGURA 37 pode-se analisar os pesos das variáveis do modelo utilizadas na análise de componentes principais. Neste caso, as variáveis são os comprimentos de ondas dos espectros do micro-NIR, sendo o maior peso obtido no comprimento de

onda entre 900 e 1000 nm. A partir do valor de 1000 nm até próximo a 1200 o peso das variáveis diminui até atingirem pesos muito pequenos para o modelo. É possível interpretar que os parâmetros de leituras do intervalo de comprimento de onda de 900 nm a 1100 nm são bastante importantes e apresentam algumas características próprias que possuem significância para máxima variância. Já na FIGURA 38 percebe-se que para os dados da derivada, o maior peso se encontra entre o comprimento de onda de 900 nm até cerca de 950 nm, e posteriormente a partir do comprimento de onda de 1500 nm. Além disso, pelos dados estarem mais distintos entre si, pode ser uma possibilidade de que os pesos para o PC1 tenham sido bem menores, já que as diferenciações para o modelo se apresentavam mais evidentes.

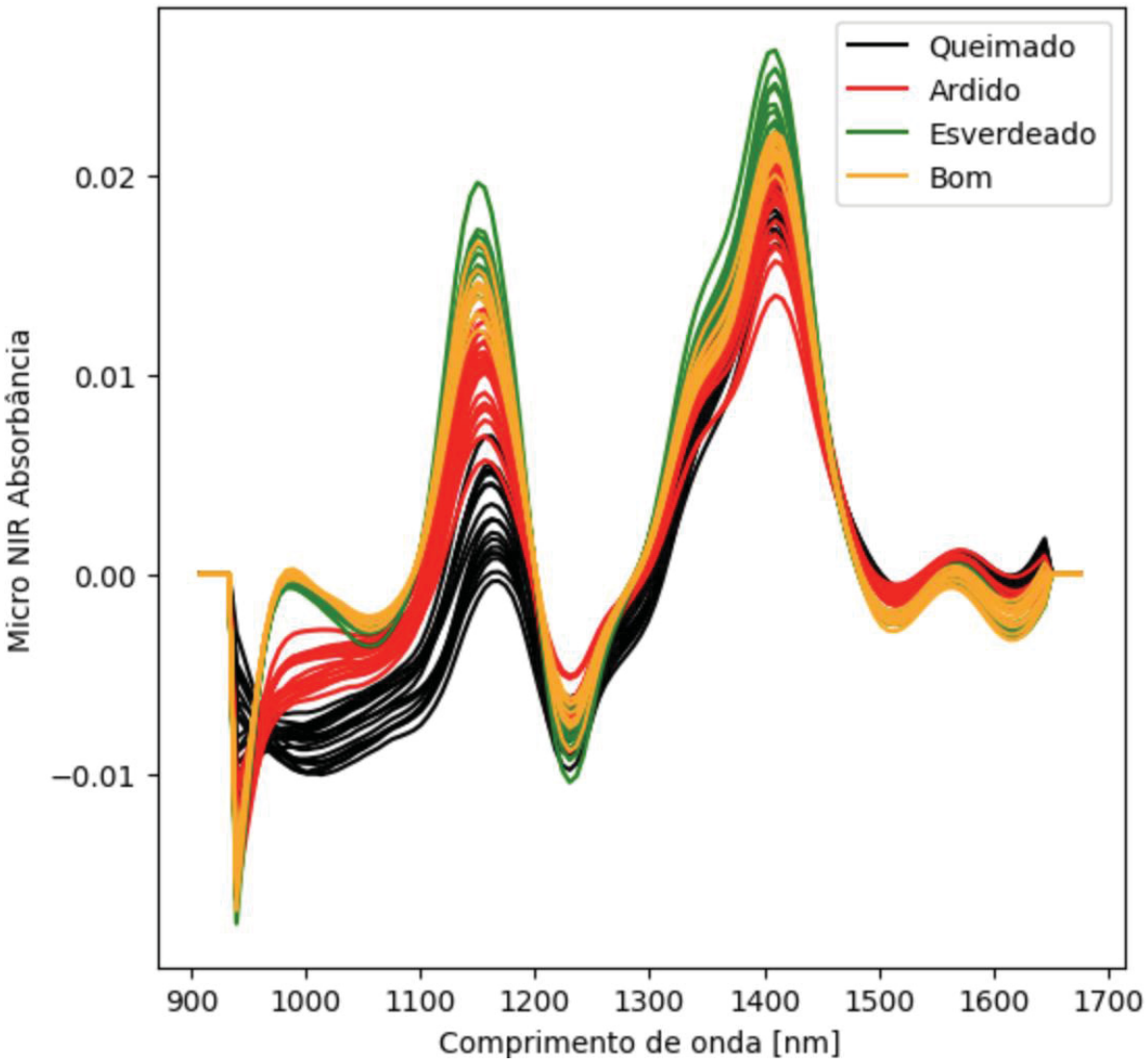
Para aplicação na predição dos dados nos modelos foram realizadas 59 leituras das classes de Ardidos, Queimados, Bom e Esverdeados. Os espectros dos dados sem pré-tratamento e com pré-tratamento (1ª derivada) podem ser observados na Figura 39 e 40.

FIGURA 39 - ESPECTROS BRUTOS PARA PREDIÇAO DOS MODELOS MATEMÁTICOS




FONTE: O autor (2023).

FIGURA 40 - ESPECTROS DERIVADA PARA PREDIÇAO DOS MODELOS MATEMÁTICOS

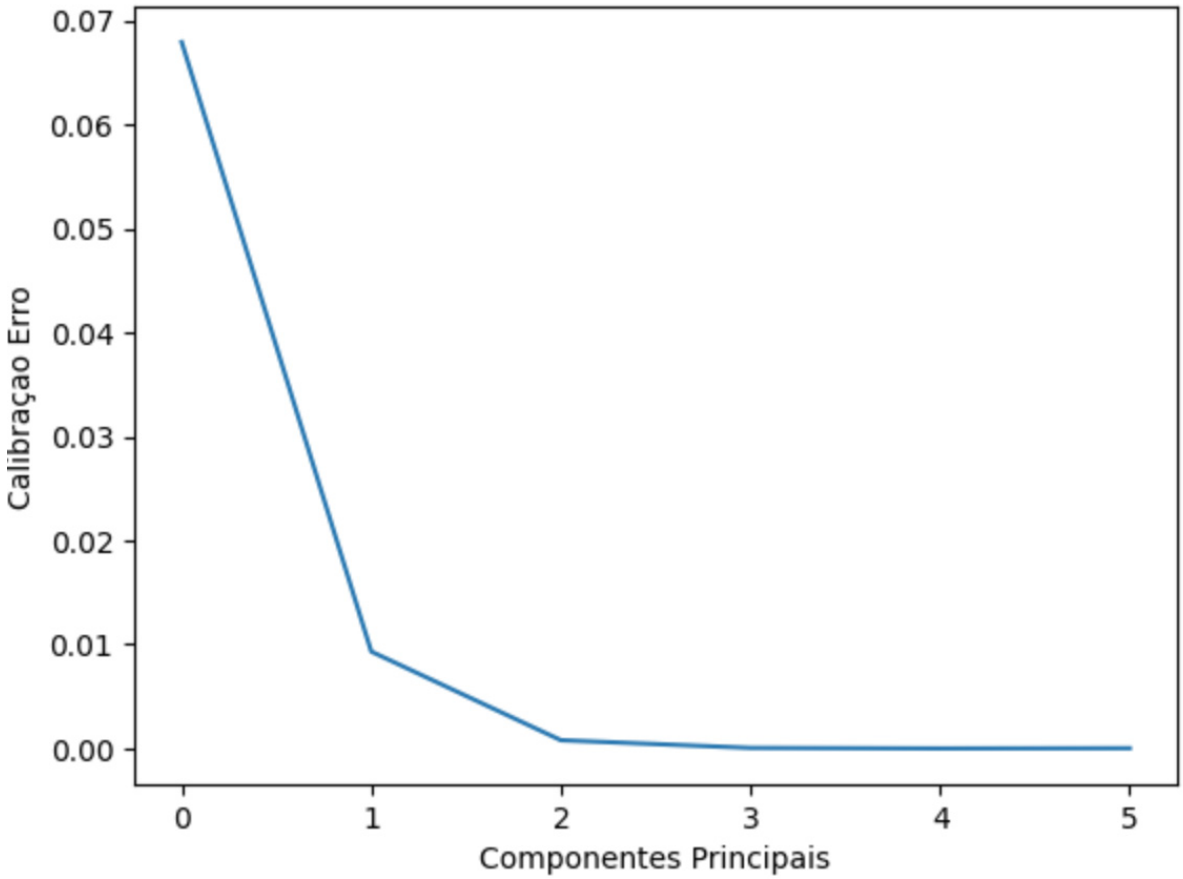



FONTE: O autor (2023).

### 4.2.2 Análise de Componente Principal com Análise Discriminatória (PCA – DA)

Para as análises de componente principal combinada com análise discriminatória foi utilizado o software Unscrambler. Para esta análise foram utilizadas as 200 leituras para o treinamento do modelo e as 59 leituras realizadas para predição do modelo. Para os dados de leitura bruto o número ótimo de componentes principais (PC) foi 2, como demonstrado na seção anterior. E assim, o erro residual pode ser observado graficamente na Figura 41.

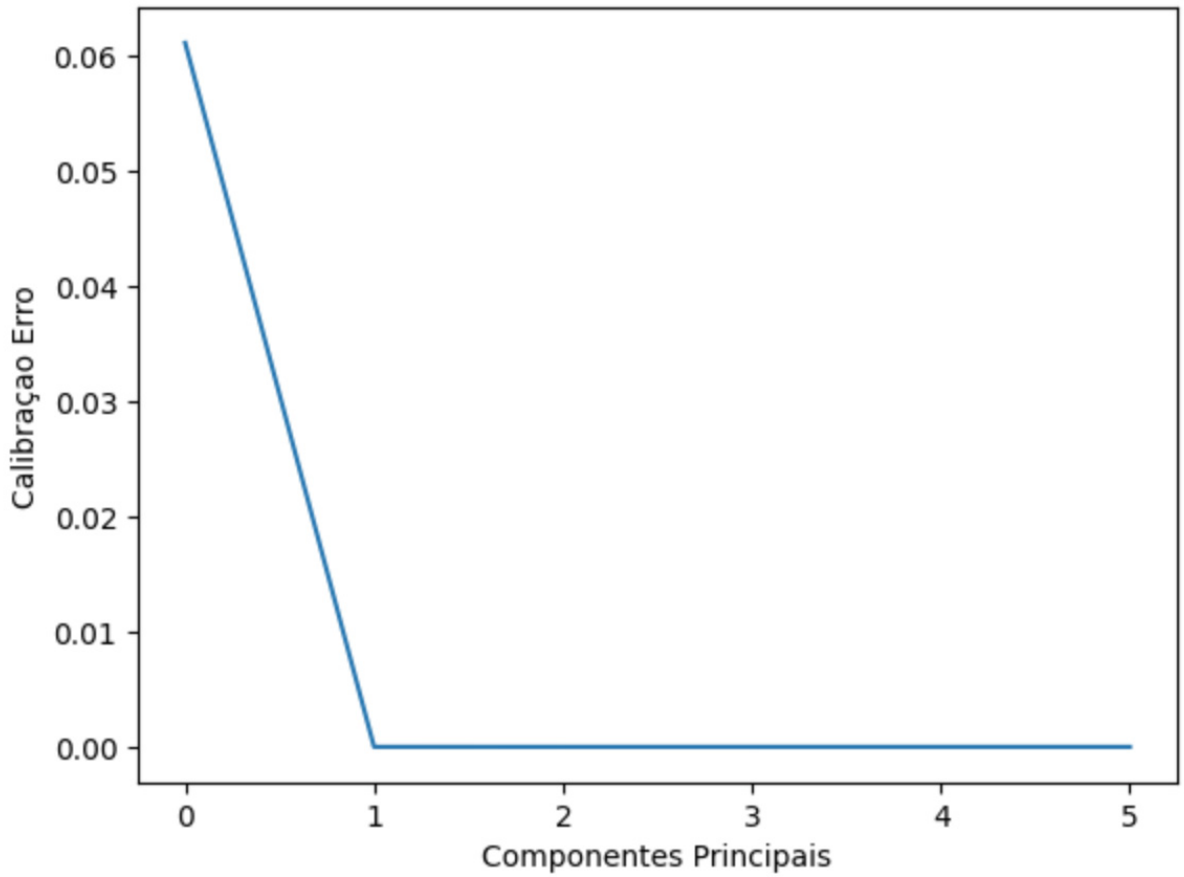
FIGURA 41 - ERRO RESIDUAL PARA LEITURA DE DADOS BRUTOS




FONTE: O autor (2023).

Desta maneira, é possível perceber que o erro diminui intensamente até PC=2. Outrossim, nota-se que as variâncias explicadas pelas componentes PC1 e PC2 são de 86% e 13%, respectivamente.

FIGURA 42 - ERRO RESIDUAL PARA LEITURA DE DADOS DERIVADA




FONTE: O autor (2023).

Para o conjunto tratado com primeira derivada, novamente foi suficiente adotar PC=1 para plena explicação da variância. Desta maneira, seu erro residual diminui subitamente até PC=1.

Na aplicação do modelo nas amostras usadas para treinamento do mesmo foram utilizados os dados brutos e os dados da primeira derivada. Foram aplicados os modelos linear, quadrático e Mahalanobis. (vide Tabela 5 a 7)

TABELA 5 – RESULTADOS TREINAMENTO PCA-DA LINEAR

| **PCA Bruto e Derivada – Linear** | | | | |
|---|---|---|---|---|
| | Sensibilidade | Especificidade | Exatidão | Erro |
| Esverdeado - Bruto | 90,0% | 94,7% | 93,5% | 6,5% |
| Bom - Bruto | 84,0% | 96,7% | 93,5% | 6,5% |
| Ardido-Bruto | 100,0% | 100,0% | 100,0% | 0,0% |
| Queimado-Bruto | 100,0% | 100,0% | 100,0% | 0,0% |
| Esverdeado - Derivada | 64,0% | 92,7% | 85,5% | 14,5% |
| Bom - Derivada | 78,0% | 88,0% | 85,5% | 14,5% |

| | | | | |
|---|---|---|---|---|
| Ardido - Derivada | 100,0% | 100,0% | 100,0% | 0,0% |
| Queimado - Derivada | 100,0% | 100,0% | 100,0% | 0,0% |

FONTE: O autor (2023).

TABELA 6 - RESULTADOS TREINAMENTO PCA-DA QUADRÁTICO

| **PCA Bruto e Derivada – Quadrática** | | | | |
|---|---|---|---|---|
| | Sensibilidade | Especificidade | Exatidão | Erro |
| Esverdeado - Bruto | 100% | 100% | 100% | 0% |
| Bom - Bruto | 100% | 100% | 100% | 0% |
| Ardido-Bruto | 100% | 100% | 100% | 0% |
| Queimado-Bruto | 100% | 100% | 100% | 0% |
| Esverdeado - Derivada | 76,0% | 98,0% | 92,5% | 7,5% |
| Bom - Derivada | 94,0% | 92,0% | 92,5% | 7,5% |
| Ardido - Derivada | 100% | 100% | 100% | 0% |
| Queimado - Derivada | 100% | 100% | 100% | 0% |

FONTE: O autor (2023).

TABELA 7 - RESULTADOS TREINAMENTO PCA-DA MAHALANOBIS

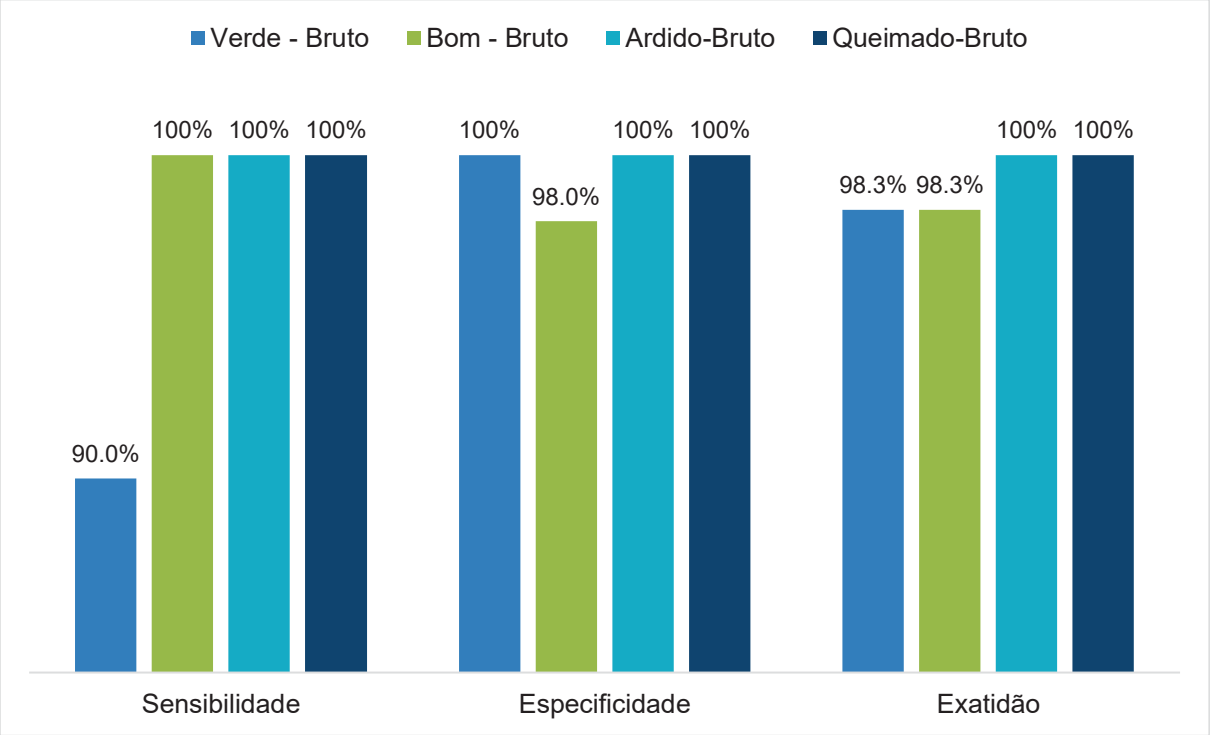
| **PCA Bruto e Derivada – Mahalanobis** | | | | |
|---|---|---|---|---|
| | Sensibilidade | Especificidade | Exatidão | Erro |
| Esverdeado - Bruto | 100% | 100% | 100% | 0% |
| Bom - Bruto | 100% | 100% | 100% | 0% |
| Ardido-Bruto | 100% | 100% | 100% | 0% |
| Queimado-Bruto | 100% | 100% | 100% | 0% |
| Esverdeado - Derivada | 68,0% | 89,3% | 84,0% | 16% |
| Bom - Derivada | 68,0% | 89,3% | 84,0% | 16% |
| Ardido - Derivada | 100% | 100% | 100% | 0% |
| Queimado - Derivada | 100% | 100% | 100% | 0% |

FONTE: O autor (2023).

Analisando os resultados obtidos no treinamento do modelo utilizando as 200 leituras pode-se observar que para os dados brutos o modelo quadrático e Mahalanobis apresenta uma maior exatidão e em consequência um menor erro, neste caso respectivamente os valores são 100% e 0%. Tais resultados demonstram que o modelo desempenhou bem para as leituras sem pré-tratamentos. Já para os dados espectrais tratados com 1ª derivada, o melhor modelo no treinamento foi o quadrático, com um menor erro e uma maior exatidão.

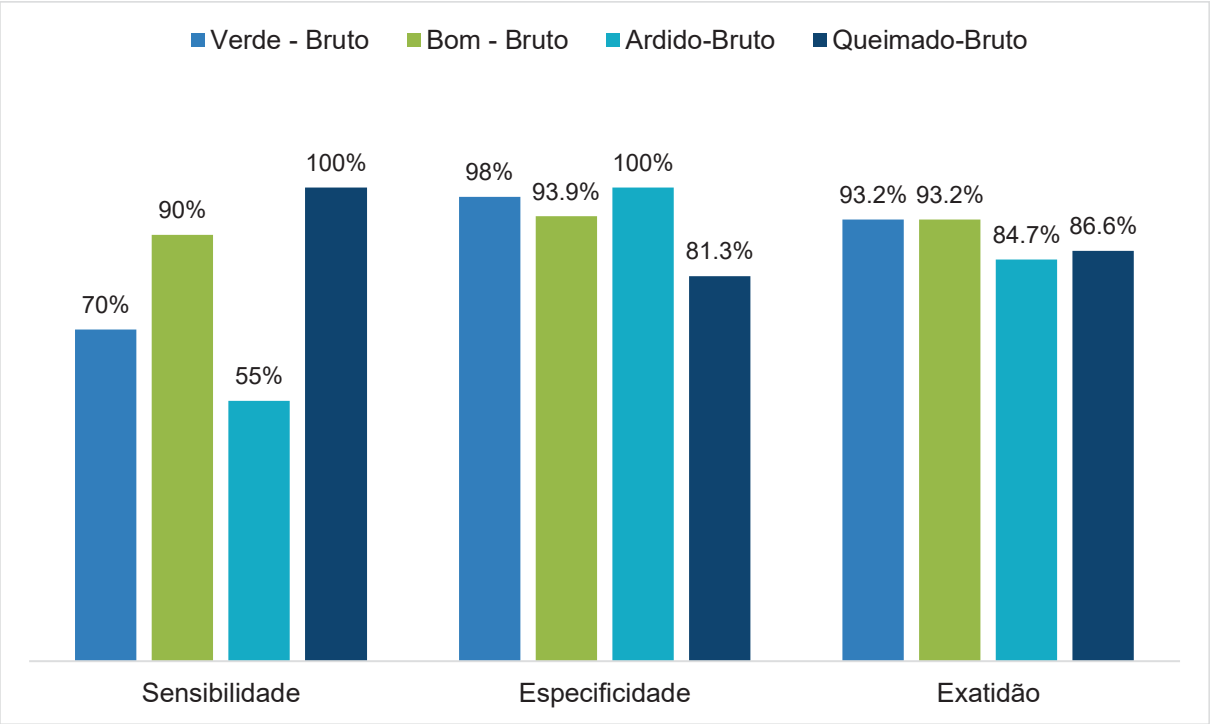
Para a avaliação do desempenho do modelo foram utilizadas 59 leituras para que o modelo fizesse a predição das amostras por classe, bem como o cálculo das métricas de desempenho.

GRÁFICO 1 - RESULTADO PREDIÇAO DADOS BRUTOS PCA-DA



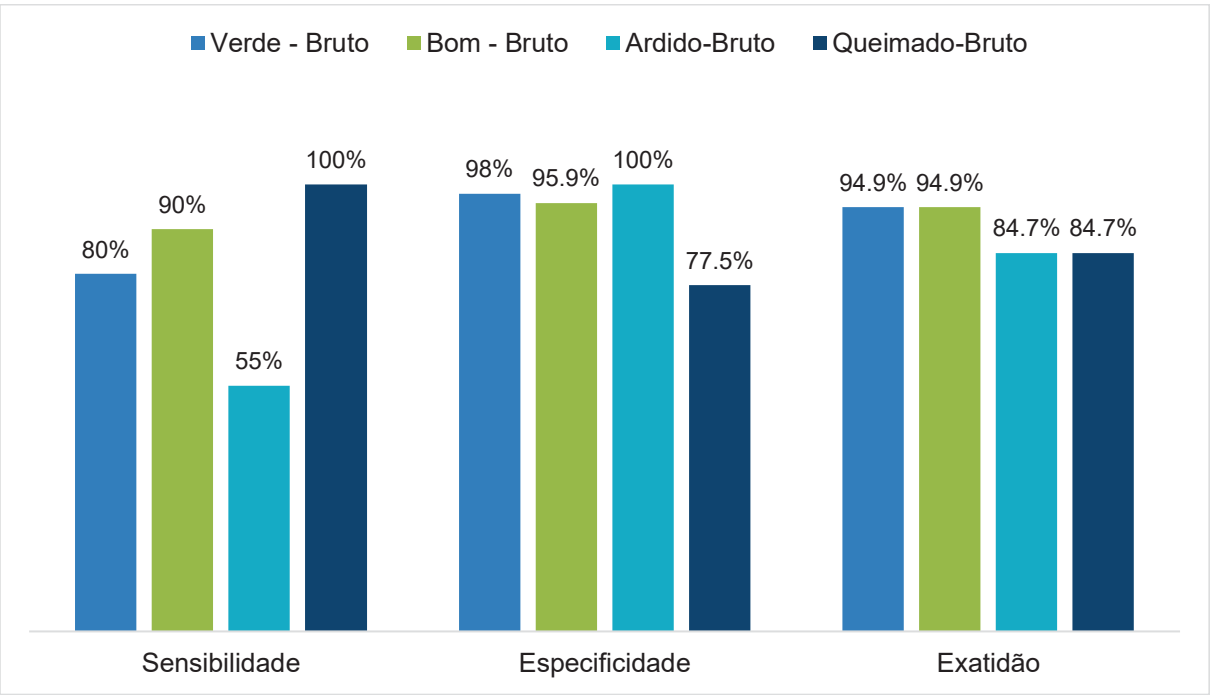

FONTE: O autor (2023).

GRÁFICO 2 - RESULTADO PREDIÇAO DADOS BRUTOS PCA-QA



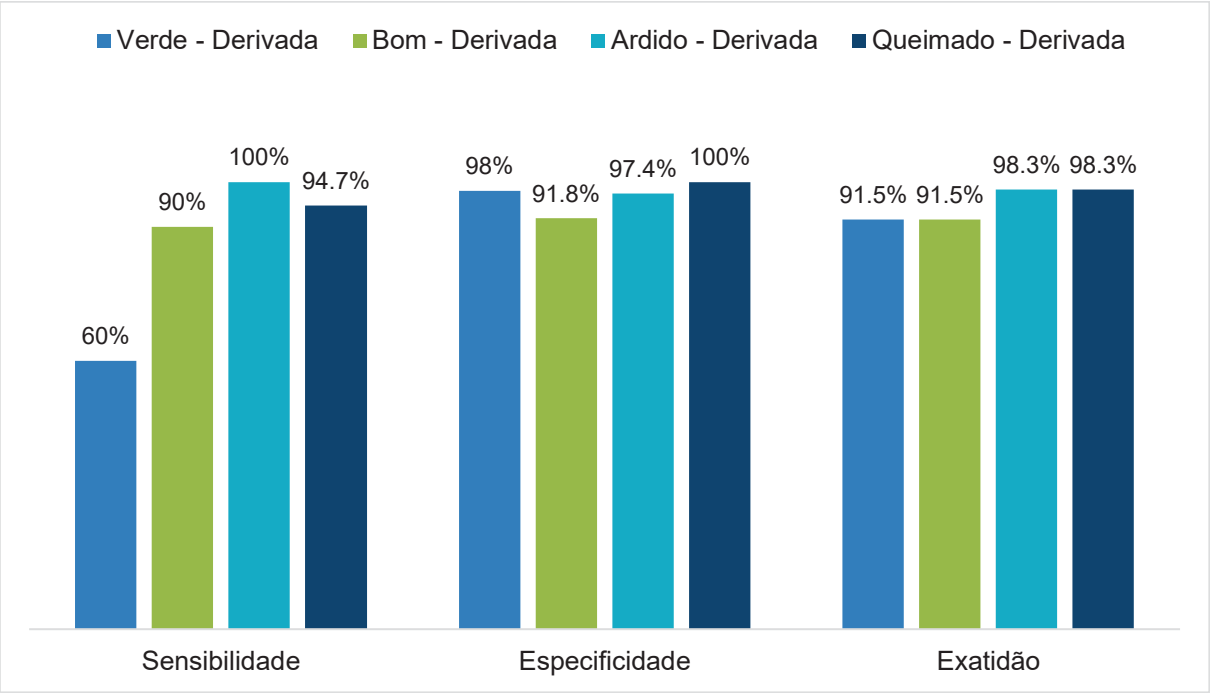

FONTE: O autor (2023).

GRÁFICO 3 - RESULTADO PREDIÇAO DADOS BRUTOS PCA-MA



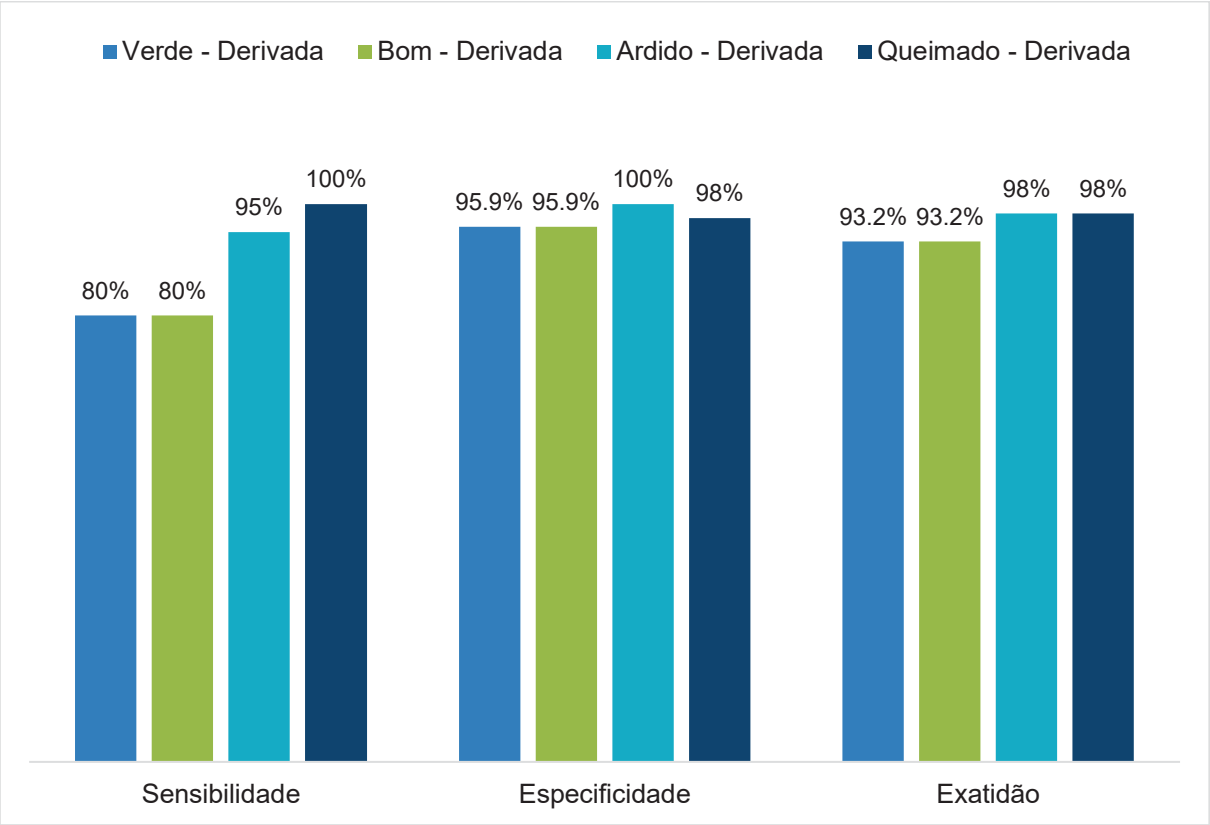

FONTE: O autor (2023).

GRÁFICO 4 - RESULTADO PREDIÇAO DADOS DERIVADA PCA-DA



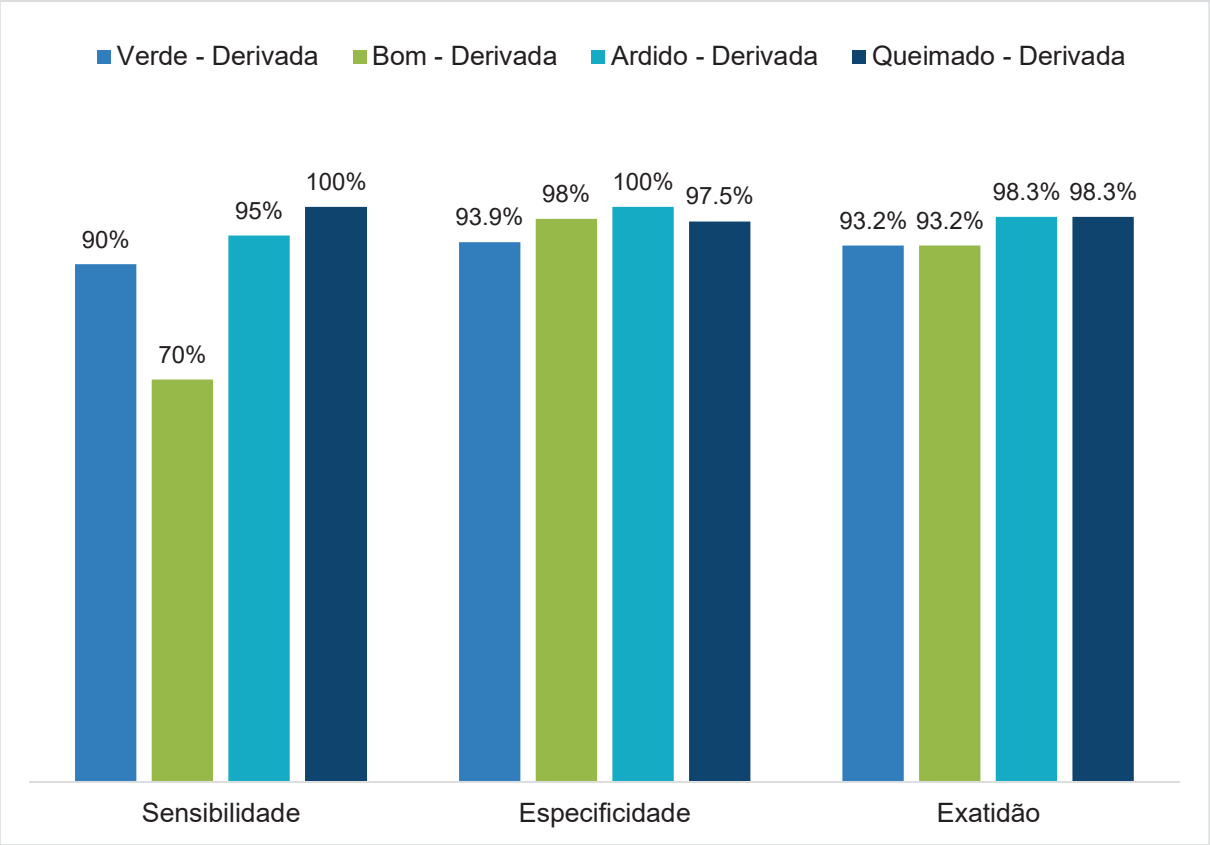

FONTE: O autor (2023).

GRÁFICO 5 - RESULTADO PREDIÇAO DADOS DERIVADA PCA-QA




FONTE: O autor (2023).

GRÁFICO 6 - RESULTADOS PREDIÇAO DADOS DERIVADA PCA-MA




FONTE: O autor (2023).

Com os resultados obtidos na etapa de predição dos modelos, foi possível analisar que o modelo com maior exatidão e por consequência um menor erro foi o PCA-DA linear para os dados brutos, seguido por Mahalanobis e posteriormente o quadrático. Para os dados da derivada, os modelos Mahalanobis e quadrático apresentam valores muito próximos, e o linear apresenta um maior valor de erro. Entretanto, para análise dos valores totais, o baixo erro e uma exatidão mais alta é encontrada no modelo PCA-DA nos espectros brutos. Nos espectros brutos se sobressaem características físicas dos grãos, o que pode ser considerado pelo modelo para a classificação das classes das amostras.

Analisando o trabalho de Kaufmann et al. (2022), em que foi aplicado a metodologia de classificação LDA para óleos, a partir da leitura realizada por um micro-NIR, os valores encontrados de exatidão apresentada chegam a 99% para a classe de óleo de soja, sendo a menor exatidão a de 90%, para óleo de canola. Esses resultados de validação e comparando com os resultados encontrados neste trabalho,

demonstram que o uso combinado de espectros de MicroNIR e a análise discriminatória linear (LDA) resulta em modelo capaz de distinguir as características das amostras e então realizar a categorização das classes com uma eficiência maior para alguns casos e menores para outros. Isso porque, há algumas características que se apresentam mais distante de outras analisando as categorias das amostras, ajudando o modelo a identificar a característica. Esses pontos de reconhecimento de uma classe com maior eficiência e outras com menor ocorre tanto no trabalho citado, como neste trabalho.(Kaufmann et al., 2022).

### 4.2.3 SIMCA

Na aplicação do SIMCA foi utilizada a mesma metodologia do tópico anterior, em que 200 leituras das classes foram utilizadas para treinamento do modelo e 40 para a predição do mesmo. Para os espectros brutos, pode-se observar que a maior distância calculada pelo modelo foram as dos grãos queimados em relação aos demais. Isso se deve pelo fato dos grãos queimados, por terem uma característica bem particular, em que se apresentam carbonizados. A distância do grão queimado para o ardido é a mais próxima observada, mas ainda se apresenta com uma relativa distância para separação das classes. As classes que se apresentam mais próximas são das de grãos verdes e bons, pois os dois apresentam características consideradas não tão diferentes para o modelo. Nas Tabelas 8 a 9 apresenta com as distâncias numéricas observadas de cada classe, considerando dados espectrais brutos (Tabela 8) e tratados com 1ª derivada (Tabela 9).

TABELA 8 - DISTÂNCIA ENTRE OS MODELOS COM DADOS BRUTOS

| | **PCA Esverdeados** | **PCA Bom** | **PCA Ardido** | **PCA Queimado** |
|---|---|---|---|---|
| PCA Esverdeados | 1 | 4,225 | 31,758 | 558,041 |
| PCA Bom | 4,225 | 1 | 73,543 | 879,448 |
| PCA Ardido | 31,758 | 73,543 | 1 | 225,675 |
| PCA Queimado | 558,041 | 879,448 | 225,675 | 1 |

FONTE: O autor (2023).

TABELA 9 - DISTÂNCIA ENTRE OS MODELOS COM DADOS DA DERIVADA

| | **PCA Esverdeados** | **PCA Bom** | **PCA Ardido** | **PCA Queimado** |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| PCA Esverdeados | 1 | 7,522 | 303,527 | 1811,272 |
| PCA Bom | 7,522 | 1 | 451,160 | 1853,706 |
| PCA Ardido | 303,527 | 451,160 | 1 | 843,262 |
| PCA Queimado | 1811,272 | 1853,706 | 843,262 | 1 |

FONTE: O autor (2023).

É possível perceber que nos dados pré-tratados com 1ª derivada as distâncias entre os modelos são maiores. Conforme discussão anterior, o pré-tratamento aplicado reduz as influências físicas das amostras sobre os espectros, evidenciando assim o comportamento químico, o que acentua as diferenças de classes para este modelo. Por essa razão é possível verificar que as distâncias são maiores quando comparada com os dados brutos.

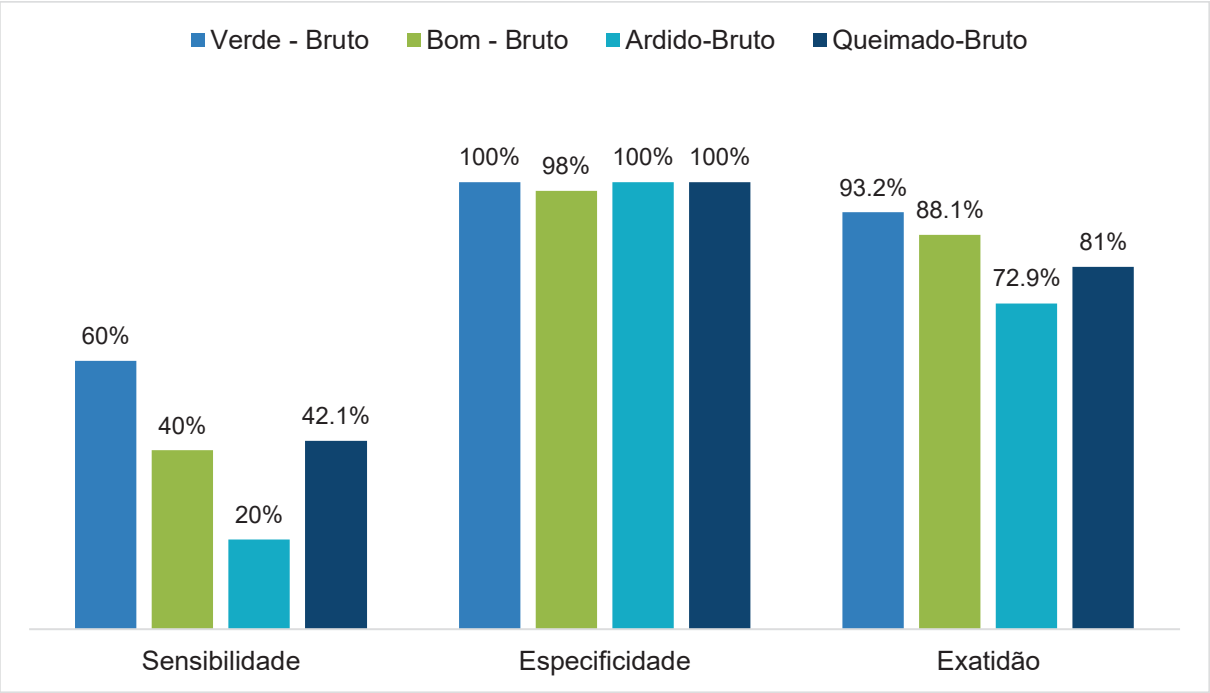
Então, como consequência, na Tabela 10 é possível verificar o resultado dos dados de treinamento do modelo, tanto para os dados brutos, quanto para os dados com pré-tratamento (1ª derivada).

TABELA 10 - RESULTADO TREINAMENTO SIMCA

| **SIMCA Bruto e Derivada** | | | | |
|---|---|---|---|---|
| | Sensibilidade | Especificidade | Exatidão | Erro |
| Esverdeado - Bruto | 64,0% | 100,0% | 91,0% | 9,0% |
| Bom - Bruto | 70,0% | 100,0% | 92,5% | 7,5% |
| Ardido-Bruto | 98,0% | 100,0% | 99,5% | 0,5% |
| Queimado-Bruto | 100,0% | 100,0% | 100,0% | 0,0% |
| Esverdeado - Derivada | 92,0% | 100,0% | 98,0% | 2,0% |
| Bom - Derivada | 70,0% | 100,0% | 92,5% | 7,5% |
| Ardido - Derivada | 96,0% | 100,0% | 99,0% | 1,0% |
| Queimado - Derivada | 100,0% | 100,0% | 100,0% | 0,0% |

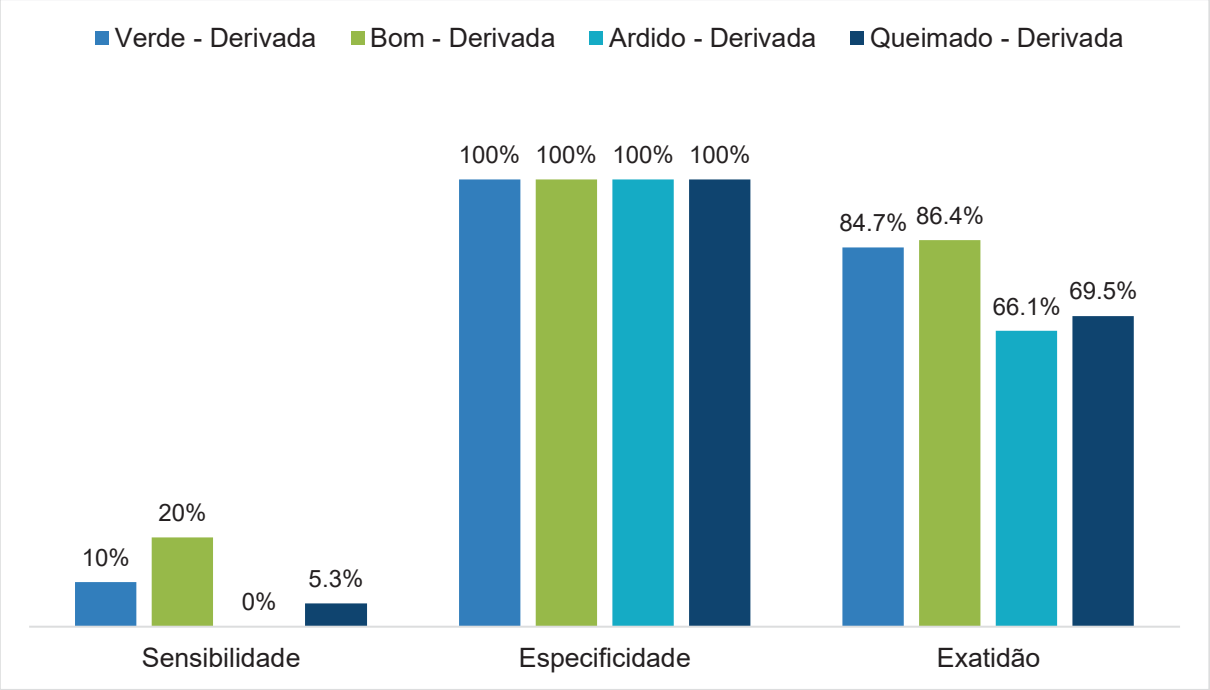
FONTE: O autor (2023).

GRÁFICO 7 - RESULTADOS PREDIÇÃO DADOS BRUTOS




FONTE: O autor (2023).

GRÁFICO 8 - RESULTADOS PREDIÇÃO DADOS DERIVADA




FONTE: O autor (2023).

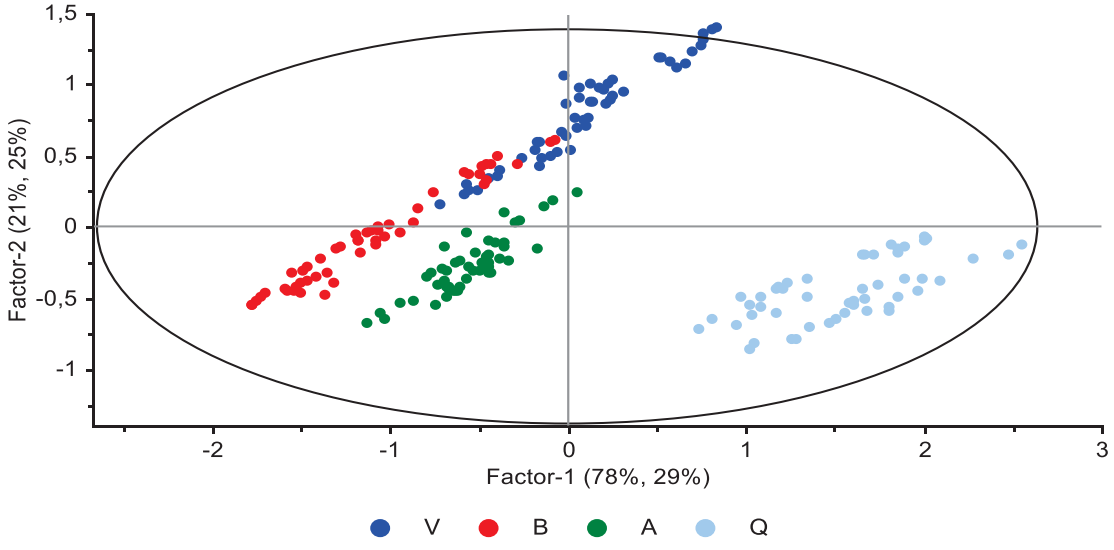
A partir da Tabela 10 é possível observar que o modelo apresentou um melhor resultado de exatidão para os dados de 1ª derivada, o que possui relação com as distâncias entre classes, ou seja, com esses dados pré-tratados o modelo conseguiu separar as classes com uma maior eficiência. O ponto de maior erro observado fica para os grãos bons, por sua proximidade característica com os grãos esverdeados. Cabe destacar que estes resultados se referem ao conjunto de dados de treinamento.

Os gráficos 9 e 10 ilustram as métricas de desempenho para o conjunto de dados de predição, considerando espectros brutos e tratados, respectivamente. Observa-se uma intensa perda de sensibilidade e de exatidão da modelagem SIMCA no teste de predição, sobretudo para os espectros tratados com 1ª derivada, indicando a reduzida acurácia do método para classificação de dados novos.

### 4.2.4 Mínimos Quadrados Parciais combinado com Análise Discriminante (PLS-DA)

Na aplicação do modelo PLS-DA foi feito primeiramente o treinamento do modelo, utilizando 200 espectros. Primeiramente, realizou-se uma análise da capacidade de separação de classes do modelo através de gráficos do 1º e 2º fatores, que constituem as principais componentes que descrevem as variações dos dados espectrais. As Figuras 43 e 44 ilustram a separação das classes de grãos com base em dados espectrais brutos e tratados, respectivamente.

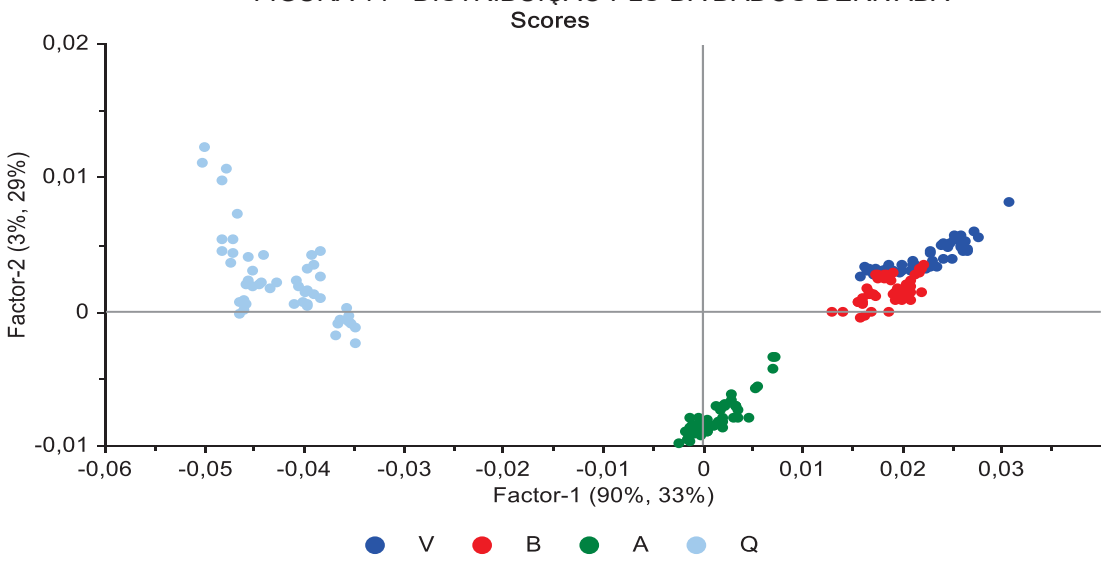
FIGURA 43 - DISTRIBUIÇÃO PLS-DA DADOS BRUTOS



FONTE: O autor (2023).

LEGENDA: V = Esverdeados; B = Bons; A = Ardidos; Q = Queimados.

FIGURA 44 - DISTRIBUIÇÃO PLS-DA DADOS DERIVADA




FONTE: O autor (2023).

LEGENDA: V = Esverdeados; B = Bons; A = Ardidos; Q = Queimados.

Nas Figuras 43 e 44, é possível observar que os grãos queimados se distanciam dos demais, e que as classes de esverdeados e bons se sobrepõem em alguns pontos. Entretanto, há ainda pontos das duas classes mencionadas suficientemente distantes passíveis de serem utilizados pelo modelo para distinção entre tais classes. Assim, a partir da Tabela 11 é possível verificar os resultados obtidos de treinamento do modelo.

TABELA 11 – RESULTADO TREINAMENTO PLS-DA DADOS BRUTOS E DERIVADA

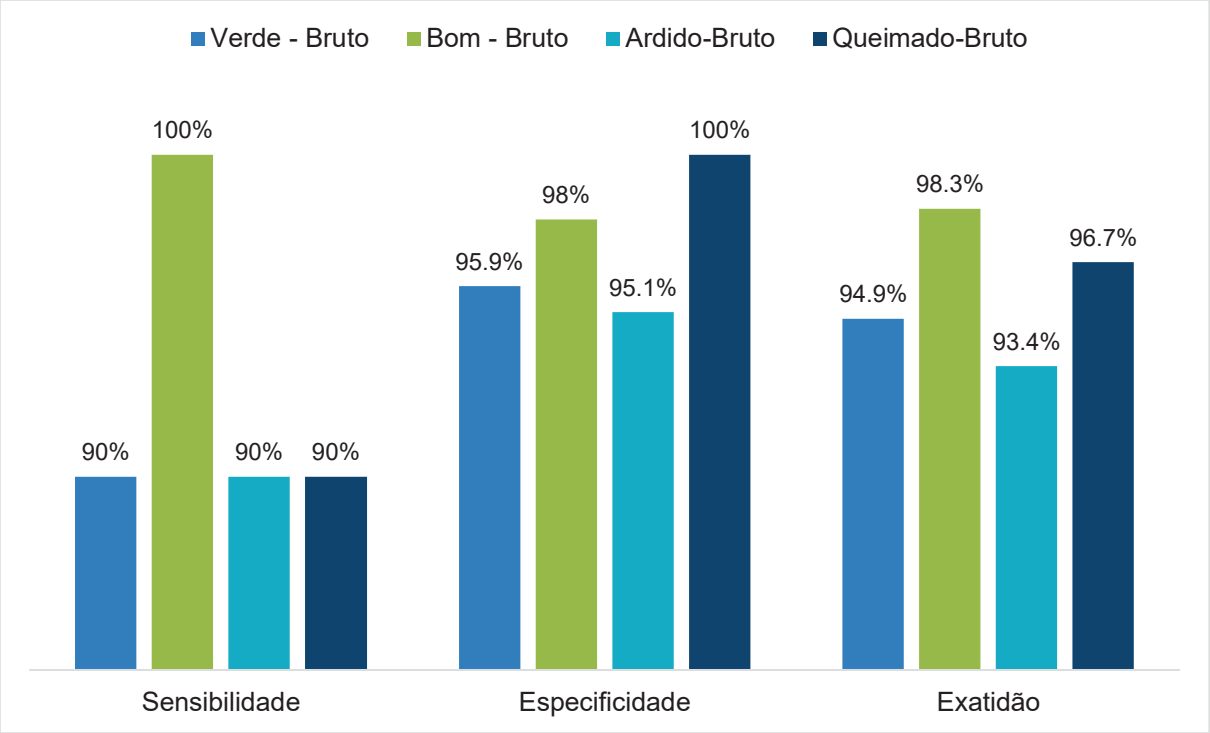
| **PLS-DA Bruto e Derivada** | | | | |
|---|---|---|---|---|
| | Sensibilidade | Especificidade | Exatidão | Erro |
| Esverdeado - Bruto | 100,0% | 100,0% | 100,0% | 0,0% |
| Bom - Bruto | 100,0% | 100,0% | 100,0% | 0,0% |
| Ardido-Bruto | 100,0% | 100,0% | 100,0% | 0,0% |
| Queimado-Bruto | 100,0% | 100,0% | 100,0% | 0,0% |
| Esverdeado - Derivada | 100,0% | 100,0% | 100,0% | 0,0% |
| Bom - Derivada | 100,0% | 100,0% | 100,0% | 0,0% |
| Ardido - Derivada | 100,0% | 100,0% | 100,0% | 0,0% |
| Queimado - Derivada | 100,0% | 100,0% | 100,0% | 0,0% |

FONTE: O autor (2023).

O resultado do treinamento do modelo foi obtido uma exatidão de 100% e consequente erro de 0% para todas as classes, mostrando que o modelo teve uma boa adequação aos dados apresentados, sem distinção se os dados de entradas foram brutos ou de 1ª derivada.

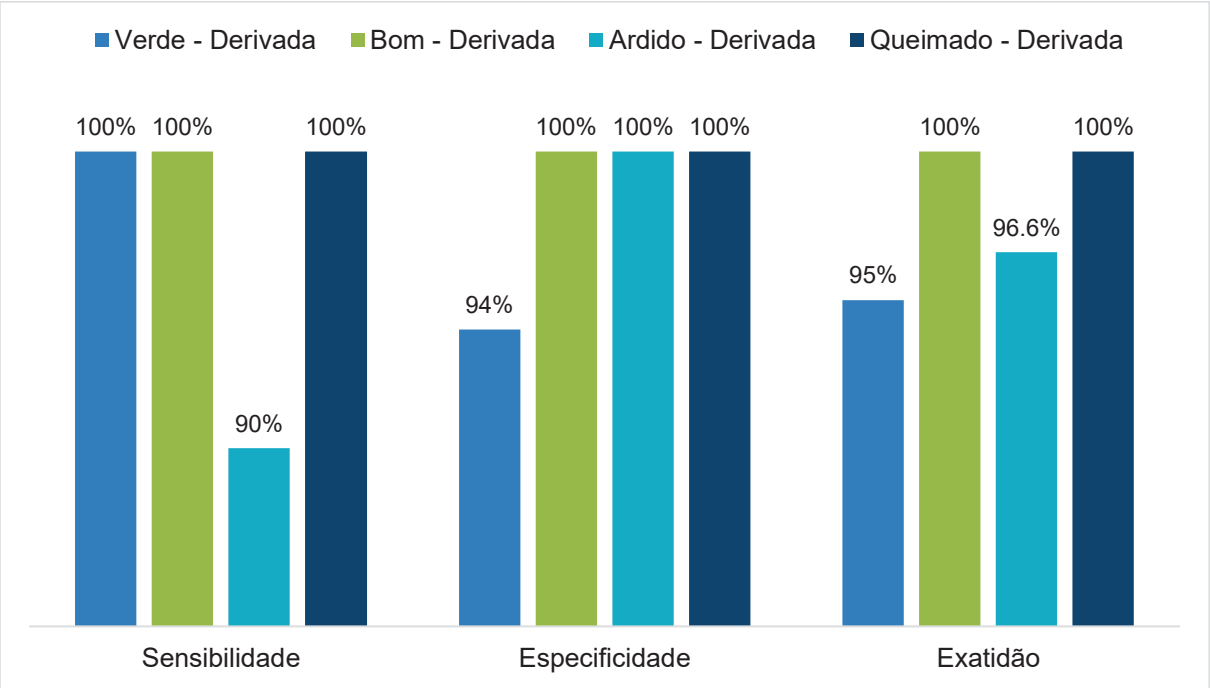
A fim de avaliar a capacidade preditiva da modelagem PLS-DA, o modelo foi testado no conjunto de predição, conforme ilustram os Gráficos 11 e 12. Observa-se que de modo geral a modelagem PLS-DA possui boa capacidade preditiva para as diferentes classes, uma vez que seu desempenho em sensibilidade, especificidade e exatidão encontra-se sempre acima de 90%, independentemente do tipo de tratamento ou de classe. Nota-se também o desempenho superior do conjunto tratado com 1ª derivada, com performance de 100% em todas as métricas para grãos bons e queimados.

GRÁFICO 9 - RESULTADOS PREDIÇAO PLS-DA DADOS BRUTOS




FONTE: O autor (2023).

GRÁFICO 10 - RESULTADOS PREDIÇAO PLS-DA DADOS DERIVADA




FONTE: O autor (2023).

O desempenho do modelo PLS-DA desenvolvido nesta dissertação pode ser confrontado com a modelagem PLS descrita por Wang et al (2002), que combinou modelos PLS e espectrofotometria NIR para classificação de sementes de soja danificadas e íntegras. A modelagem apresentada nesta dissertação, ainda que envolvendo mais classes de grãos, mostrou-se de maior assertividade na classificação em comparação com o referido trabalho.

### 4.2.5 Regressão Logística (RL)

Para aplicação do modelo de classificação de regressão logística foi realizada a divisão dos dados de treinamento e teste em 80% e 20%. O número total de dados de leituras do modelo foram de 200, e o modelo foi aplicado tanto para as leituras dos dados brutos, sem pré-tratamento, quanto para os dados da derivada, ou seja, com pré-tratamento. Esta primeira etapa foi interessante analisar o desempenho do modelo para somente as 200 leituras, para posterior aplicação das 59 leituras para predição do modelo, sendo que esta configuração dos dados de treinamento das 200 leituras e teste (predição) das 59 é realizada na própria programação do modelo de regressão

logística. Nas Tabelas 12 e 13 são exibidos os resultados do modelo aplicado para os dois tipos de dados mencionados acima.

TABELA 12 - RESULTADOS DE MAIOR E MENOR ACURÁCIA DE RL PARA DADOS DO ESPECTRO BRUTO

| Amostra | Precisão | Recall | F1-score | Acurácia |
|---|---|---|---|---|
| Esverdeado - Bruto | 0,96 | 1,00 | 0,98 | 97,5% |
| Bom - Bruto | 1,00 | 0,96 | 0,98 | |
| Ardido - Bruto | 1,00 | 1,00 | 1,00 | |
| Queimado - Bruto | 1,00 | 1,00 | 1,00 | |
| Esverdeado - Derivada | 1,00 | 1,00 | 1,00 | 100% |
| Bom - Derivada | 1,00 | 1,00 | 1,00 | |
| Ardido - Derivada | 1,00 | 1,00 | 1,00 | |
| Queimado - Derivada | 1,00 | 1,00 | 1,00 | |

FONTE: O autor (2023).

Para a Tabela 12, os valores de acurácia do modelo em instâncias do treinamento apresentaram um valor médio que variaram de 97,5% a 100%, sendo o menor valor o de 97,5%, representando, portanto, um bom desempenho do modelo na classificação. Além desse fato, os valores observados do *F1score* se apresentam muito próximos a 1, ou efetivamente 1, refletindo o bom desempenho do modelo em relação aos dados utilizados como entrada.

A Tabela 13 exibe o desempenho da modelagem RL utilizando os dados espectrais tratados com 1ª derivada. Neste caso, os valores de treinamento do modelo apresentaram uma variação média de 72,5% a 100%, representando um delta de 27,5%. O menor valor de acurácia obtido foi o de 72,5%.

TABELA 13 – RESULTADOS DE MAIOR E MENOR ACURÁCIA DE RL PARA DADOS DO ESPECTRO DA DERIVADA

| Amostra | Precisão | Recall | F1-score | Acurácia |
|---|---|---|---|---|
| Esverdeado - Bruto | 1,00 | 1,00 | 1,00 | 72,5% |
| Bom - Bruto | 1,00 | 1,00 | 1,00 | |
| Ardido - Bruto | 0,00 | 0,00 | 0,00 | |
| Queimado - Bruto | 0,50 | 1,00 | 0,67 | |
| Esverdeado - Derivada | 1,00 | 1,00 | 1,00 | 100% |
| Bom - Derivada | 1,00 | 1,00 | 1,00 | |
| Ardido - Derivada | 1,00 | 1,00 | 1,00 | |
| Queimado - Derivada | 1,00 | 1,00 | 1,00 | |

FONTE: O autor (2023).

Para o menor valor da acurácia obtida modelo se equivocou em todas as amostras da classe ardido. Plotando a matriz de confusão do modelo para este

resultado foi possível analisar que as amostras que deveriam ser classificadas como 3, foram classificadas como 4. A maior variação de delta obtida nos dados de entradas da derivada pode ser devido ao fato que o pré-tratamento processa e retira algumas informações relevantes para o modelo utilizar em sua classificação. Admite-se neste caso que a suavização produzida pela 1ª derivada tenha sido prejudicial à capacidade de classificação do modelo.

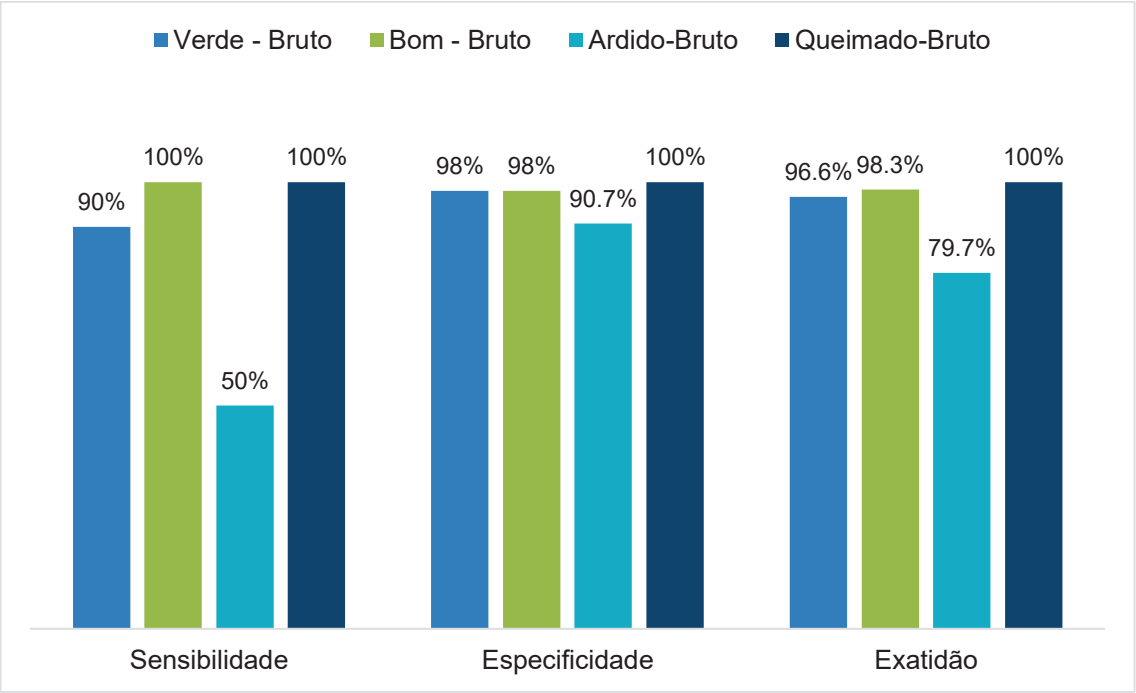
Procedeu-se também a avaliação de desempenho da modelagem RL com base nas métricas de Sensibilidade, Especificidade e Exatidão discutidas anteriormente. A Tabela 14 resume a aplicação de tais métricas para o conjunto de treinamento, observa-se a facilidade do modelo RL no treinamento com dados brutos e sua dificuldade no treinamento a partir de espectros tratados para grãos ardidos e queimados.

TABELA 14 - RESULTADOS APLICAÇÃO DE MÉTRICAS MODELO DE REGRESSÃO LOGÍSTICA

| **RL Bruto e Derivada** | | | | |
|---|---|---|---|---|
| | Sensibilidade | Especificidade | Exatidão | Erro |
| Esverdeado – Bruto | 100% | 100% | 100% | 0,0% |
| Bom – Bruto | 100% | 100% | 100% | 0,0% |
| Ardido – Bruto | 100% | 100% | 100% | 0,0% |
| Queimado – Bruto | 100% | 100% | 100% | 0,0% |
| Esverdeado – Derivada | 100% | 100% | 100% | 0,0% |
| Bom – Derivada | 100% | 100% | 100% | 0,0% |
| Ardido – Derivada | 0% | 78% | 78% | 21,6% |
| Queimado – Derivada | 100% | 78% | 81% | 19,3% |

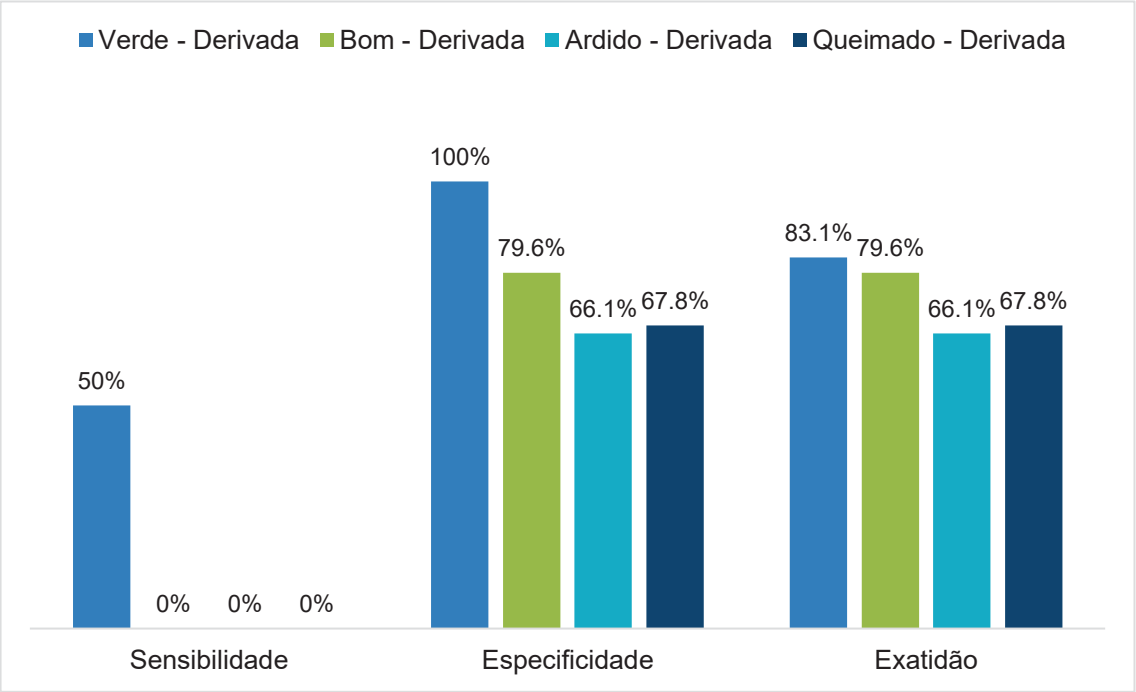
FONTE: O autor (2023).

GRÁFICO 11 - RESULTADOS PREDIÇÃO RL DADOS BRUTOS




FONTE: O autor (2023).

GRÁFICO 12 - RESULTADOS PREDIÇÃO RL DADOS DERIVADA




FONTE: O autor (2023).

Os gráficos 13 e 14 ilustram a capacidade de predição do modelo RL, percebendo-se baixos desempenhos sobretudo para os espectros tratados. A classe

para a qual o modelo apresenta uma maior dificuldade é a de ardidos, assim como nos dados brutos. Os valores da derivada apresentam resultados de exatidão menores e maiores erros em relação aos dados brutos, mostrando que o pré-tratamento espectral resultou em perda de informação relevante para a modelagem RL.

Comparando as duas métricas para o modelo para os dados da derivada, analisando o *F1score* no desempenho de acurácia de 72,5%, o qual apresenta um valor de 0,67 e para o ardido apresenta um resultado de 0 acertos. Para os ardidos, olhando a sensibilidade, percebe-se a dificuldade do modelo já no treinamento. Além disso, os valores de *F1score* demonstram a baixa assertividade do modelo quanto as duas classes (ardido e queimado). Na predição, os valores são diferentes, e mostram a dificuldade encontrada do modelo na classificação dos dados.

### 4.2.6 Máquinas de Vetores Suporte (SVM)

No desenvolvimento do modelo de SVM foram utilizadas as métricas de treinamento e teste, sendo os 200 dados de leitura usados para treinamento e os 59 para o teste (predição) do modelo. Os dados de entrada para o modelo de classificação foram os dados brutos e os dados com pré-tratamento, que corresponde neste trabalho a derivada. Os resultados estão na Tabela 15.

TABELA 15 - RESULTADO SVM PARA DADOS BRUTOS E DERIVADA

| Amostra | Precisão | Recall | F1-score | Acurácia |
|---|---|---|---|---|
| Esverdeados (1) | 1,00 | 1,00 | 1,00 | 100% |
| Bom (2) | 1,00 | 1,00 | 1,00 | |
| Ardido (3) | 1,00 | 1,00 | 1,00 | |
| Queimado (4) | 1,00 | 1,00 | 1,00 | |

FONTE: O autor (2023).

Os resultados obtidos para o modelo de SVM foi de 100% de acurácia, o mesmo resultado foi obtido por um intervalo de repetições realizadas para as duas entradas de dados. Certamente, neste modelo há uma boa adequação de classificação, sendo que, tanto pela entrada dos dados brutos quanto pela entrada dos dados pré-tratados, não causam diferenças no desempenho do modelo. Em outros tipos de classificação de grãos, o SVM mostrou um bom desempenho, sendo

seu resultado otimizado após um pré-tratamento. (Sampaio et al., 2020) Um fato bastante importante a ser destacado, pois para este trabalho os dois dados de entrada obtiveram um bom resultado de classificação.

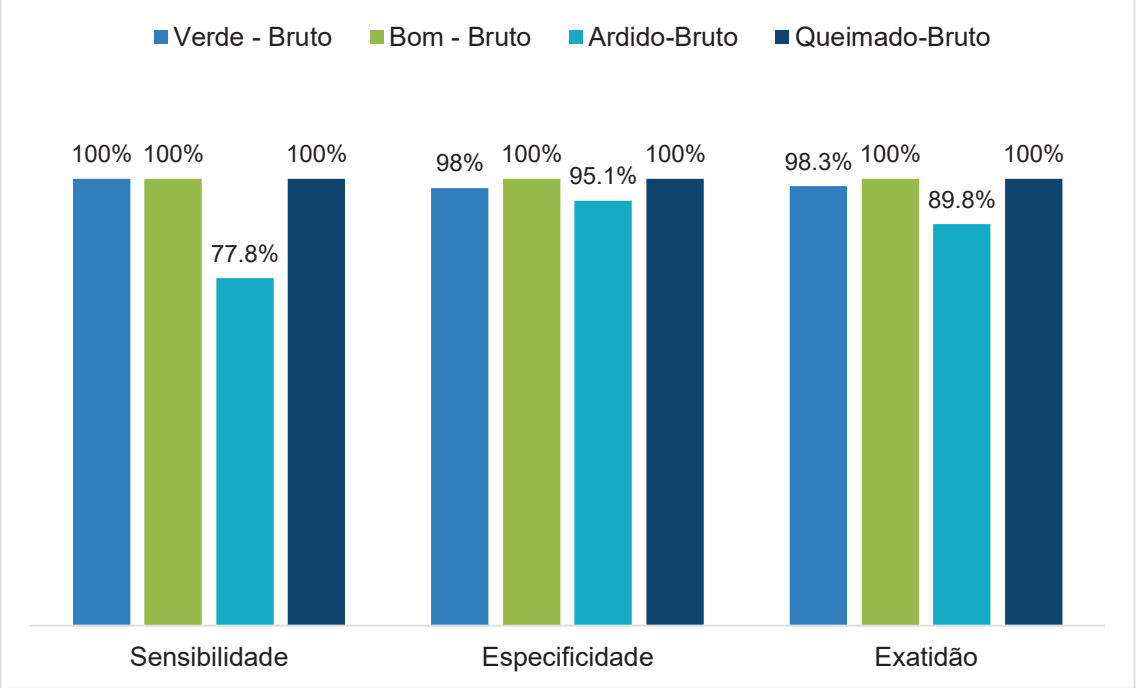
Para analisar os dados de predição do modelo, visando observar seu desempenho nas mesmas métricas que os anteriores, aplicou-se o modelo SVM ao conjunto de predição, conforme resultados demonstrados nos Gráficos 15 e 16.

TABELA 16 - RESULTADOS SVM PARA DADOS DE TREINAMENTO

| **SVM Bruto e Derivada** | | | | |
|---|---|---|---|---|
| | Sensibilidade | Especificidade | Exatidão | Erro |
| Esverdeado - Bruto | 100,0% | 100,0% | 100,0% | 0,0% |
| Bom - Bruto | 100,0% | 100,0% | 100,0% | 0,0% |
| Ardido-Bruto | 100,0% | 100,0% | 100,0% | 0,0% |
| Queimado-Bruto | 100,0% | 100,0% | 100,0% | 0,0% |
| Esverdeado - Derivada | 100,0% | 100,0% | 100,0% | 0,0% |
| Bom - Derivada | 100,0% | 100,0% | 100,0% | 0,0% |
| Ardido - Derivada | 100,0% | 100,0% | 100,0% | 0,0% |
| Queimado - Derivada | 100,0% | 100,0% | 100,0% | 0,0% |

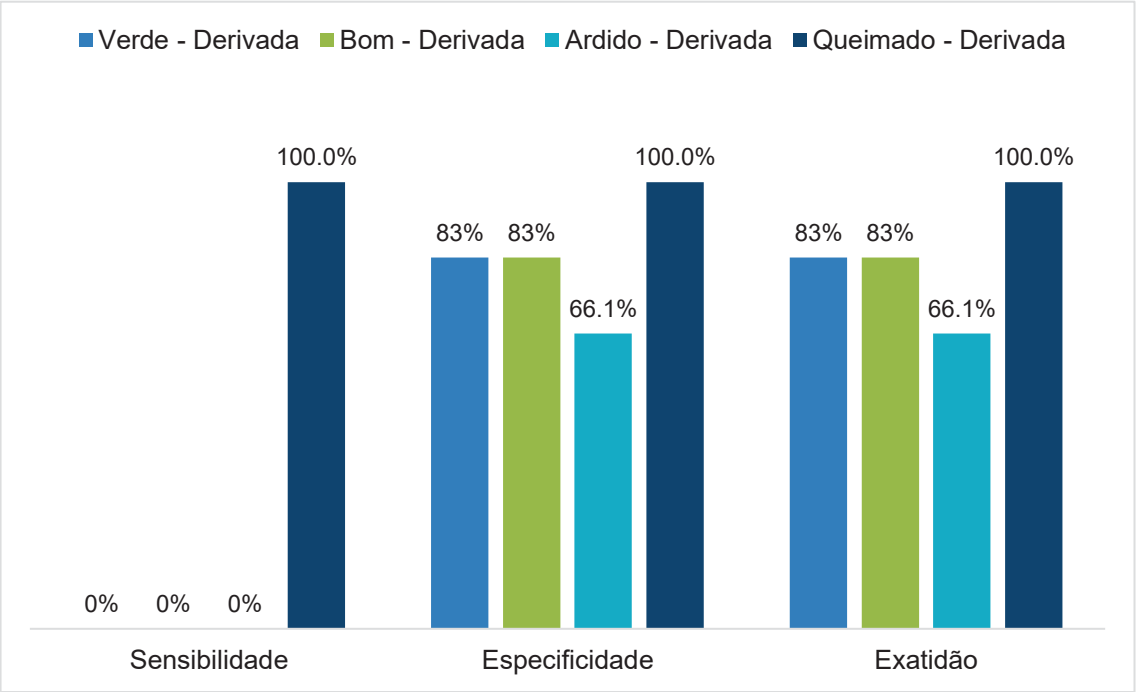
FONTE: O autor (2023).

GRÁFICO 13 - RESULTADOS PREDIÇÃO SVM DADOS BRUTOS




FONTE: O autor (2023).

GRÁFICO 14 - RESULTADOS PREDIÇÃO SVM DADOS DERIVADA




FONTE: O autor (2023).

Os resultados obtidos na predição demonstram que os dados de entradas sem pré-tratamento se adaptaram melhor ao modelo, apresentando uma exatidão de 100% para duas classes, que são os grãos bom e categoria queimado. Isso pode demonstrar que o modelo de SVM se adaptou melhor aos dados brutos, pois os dados de derivada apresentam somente exatidão de 100% para os grãos queimados. Assim como a sensibilidade do modelo se mostra baixa nos dados de entrada pré-tratados, no caso, a 1ª derivada.

### 4.2.7 Comparativos dos resultados dos modelos

Analisando os resultados dos modelos, separando por: treinamento dados brutos, treinamento dados de 1ª derivada, predição dados brutos e predição dados de 1ª derivada é possível verificar qual o modelo que apresentou o melhor desempenho. A Tabela 17 demonstra em ordem de desempenho dos dados brutos usados no treinamento do modelo, seguida pela Tabela 18, os quais traz os resultados dos dados das derivadas para treinamento do modelo.

TABELA 17 – RANKING COMPARATIVO DOS RESULTADOS DO TREINAMENTO DOS MODELOS COM DADOS BRUTOS

| RANKING MODELOS - TREINAMENTO - DADOS BRUTOS | | | | | |
|---|---|---|---|---|---|
| | | Sensibilidade | Especificidade | Exatidão | Erro |
| PLS-DA | Esverdeado | 100,0% | 100% | 100,0% | 0,0% |
| | Bom | 100,0% | 100,0% | 100,0% | 0,0% |
| | Ardido | 100,0% | 100,0% | 100,0% | 0,0% |
| | Queimado | 100,0% | 100,0% | 100,0% | 0,0% |
| | **Média do modelo** | **100,0%** | **100,0%** | **100,0%** | **0,0%** |
| | | Sensibilidade | Especificidade | Exatidão | Erro |
| SVM | Esverdeado | 100,0% | 100% | 100,0% | 0,0% |
| | Bom | 100% | 100,0% | 100,0% | 0,0% |
| | Ardido | 100,0% | 100,0% | 100,0% | 0,0% |
| | Queimado | 100,0% | 100,0% | 100,0% | 0,0% |
| | **Média do modelo** | **100,0%** | **100,0%** | **100,0%** | **0,0%** |
| | | Sensibilidade | Especificidade | Exatidão | Erro |
| PCA-QA | Esverdeado | 100,0% | 100% | 100,0% | 0,0% |
| | Bom | 100% | 100,0% | 100,0% | 0,0% |
| | Ardido | 100,0% | 100,0% | 100,0% | 0,0% |
| | Queimado | 100,0% | 100,0% | 100,0% | 0,0% |
| | **Média do modelo** | **100,0%** | **100,0%** | **100,0%** | **0,0%** |
| | | Sensibilidade | Especificidade | Exatidão | Erro |
| RL | Esverdeado | 100,0% | 100% | 100,0% | 0,0% |
| | Bom | 100,0% | 100,0% | 100,0% | 0,0% |
| | Ardido | 100,0% | 100,0% | 100,0% | 0,0% |
| | Queimado | 100,0% | 100,0% | 100,0% | 0,0% |
| | **Média do modelo** | **100,0%** | **100,0%** | **100,0%** | **0,0%** |
| | | Sensibilidade | Especificidade | Exatidão | Erro |
| PCA-MA | Esverdeado | 100,0% | 100% | 100,0% | 0,0% |
| | Bom | 100% | 100,0% | 100,0% | 0,0% |
| | Ardido | 100,0% | 100,0% | 100,0% | 0,0% |
| | Queimado | 100,0% | 100,0% | 100,0% | 0,0% |
| | **Média do modelo** | **100,0%** | **100,0%** | **100,0%** | **0,0%** |
| | | Sensibilidade | Especificidade | Exatidão | Erro |
| PCA - DA | Esverdeado | 90,0% | 95% | 93,5% | 6,5% |
| | Bom | 84% | 96,7% | 93,5% | 6,5% |
| | Ardido | 100,0% | 100,0% | 100,0% | 0,0% |
| | Queimado | 100,0% | 100,0% | 100,0% | 0,0% |
| | **Média do modelo** | **93,5%** | **97,8%** | **96,8%** | **3,3%** |
| | | Sensibilidade | Especificidade | Exatidão | Erro |
| SIMCA | Esverdeado | 64,0% | 100% | 91,0% | 9,0% |
| | Bom | 70% | 100,0% | 92,5% | 7,5% |
| | Ardido | 98,0% | 100,0% | 99,5% | 0,5% |
| | Queimado | 100,0% | 100,0% | 100,0% | 0,0% |
| | **Média do modelo** | **83,0%** | **100,0%** | **95,8%** | **4,3%** |

FONTE: O autor (2023).

Analisando os resultados de desempenho do modelo para os dados brutos, nota-se um bom resultado da maioria dos modelos, com 100% em todas as métricas e 0% de erro. Entretanto, nos modelos PCA-DA e SIMCA o mesmo resultado não se repete, apresentando um menor desempenho para grãos esverdeados e bons nos dois modelos, e para o SIMCA um menor desempenho também para os grãos ardidos.

Para os resultados obtidos no treinamento dos modelos utilizando a derivada o desempenho se mostrou diferente de quando foram usados os dados brutos na entrada do modelo. Na Tabela 18 é possível visualizar o desempenho dos modelos, sendo o destaque principalmente para exatidão obtida em cada um e consequentemente o erro calculado.

TABELA 18 – RANKING COMPARATIVO DOS RESULTADOS DO TREINAMENTO DOS MODELOS COM DADOS DERIVADA

| **RANKING MODELOS - TREINAMENTO - DADOS DERIVADA** | | | | | |
|---|---|---|---|---|---|
| | | Sensibilidade | Especificidade | Exatidão | Erro |
| PLS-DA | Esverdeado | 100,0% | 100,0% | 100,0% | 0,0% |
| | Bom | 100,0% | 100,0% | 100,0% | 0,0% |
| | Ardido | 100,0% | 100,0% | 100,0% | 0,0% |
| | Queimado | 100,0% | 100,0% | 100,0% | 0,0% |
| | **Média do modelo** | **100,0%** | **100,0%** | **100,0%** | **0,0%** |
| | | Sensibilidade | Especificidade | Exatidão | Erro |
| SVM | Esverdeado | 100,0% | 100,0% | 100,0% | 0,0% |
| | Bom | 100,0% | 100,0% | 100,0% | 0,0% |
| | Ardido | 100,0% | 100,0% | 100,0% | 0,0% |
| | Queimado | 100,0% | 100,0% | 100,0% | 0,0% |
| | **Média do modelo** | **100,0%** | **100,0%** | **100,0%** | **0,0%** |
| | | Sensibilidade | Especificidade | Exatidão | Erro |
| SIMCA | Esverdeado | 92,0% | 100% | 98,0% | 2,0% |
| | Bom | 70% | 100,0% | 92,5% | 7,5% |
| | Ardido | 96,0% | 100,0% | 99,0% | 1,0% |
| | Queimado | 100,0% | 100,0% | 100,0% | 0,0% |
| | **Média do modelo** | **89,5%** | **100,0%** | **97,4%** | **2,6%** |
| | | Sensibilidade | Especificidade | Exatidão | Erro |
| PCA-QA | Esverdeado | 76,0% | 98% | 92,5% | 7,5% |
| | Bom | 94% | 92,0% | 92,5% | 7,5% |
| | Ardido | 100,0% | 100,0% | 100,0% | 0,0% |
| | Queimado | 100,0% | 100,0% | 100,0% | 0,0% |
| | **Média do modelo** | **92,5%** | **97,5%** | **96,3%** | **3,8%** |
| RL | | Sensibilidade | Especificidade | Exatidão | Erro |

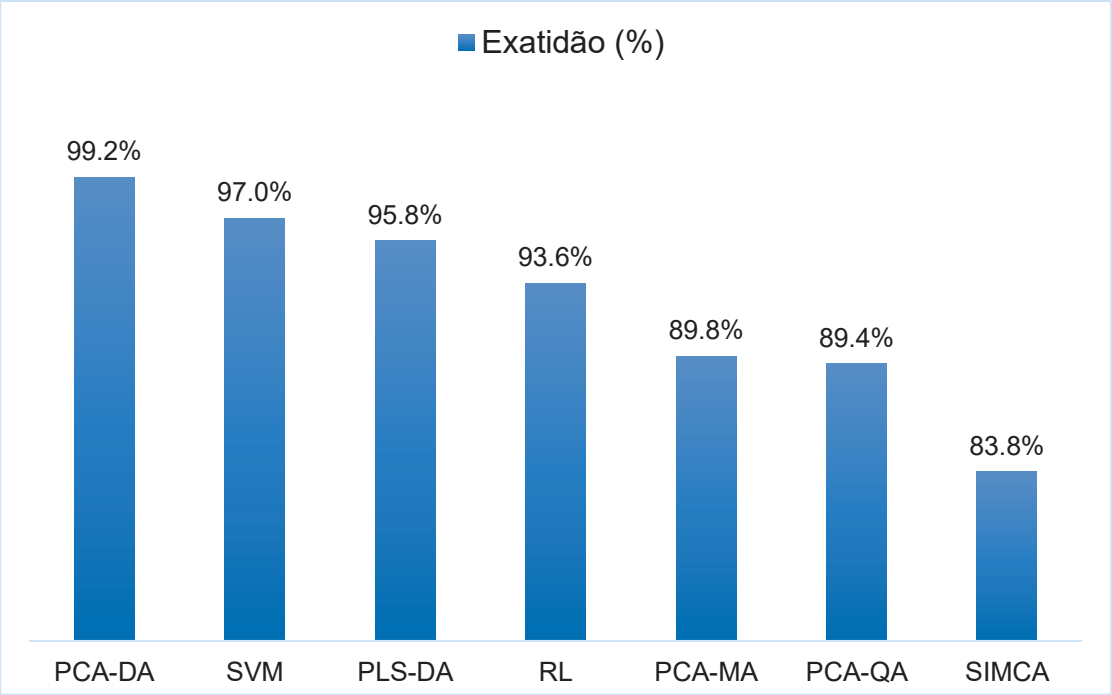
| | | | | | |
|---|---|---|---|---|---|
| | Esverdeado | 100,0% | 100% | 100,0% | 0,0% |
| | Bom | 100% | 100,0% | 100,0% | 0,0% |
| | Ardido | 0,0% | 78,4% | 78,4% | 21,6% |
| | Queimado | 100,0% | 100,0% | 100,0% | 0,0% |
| | **Média do modelo** | **75,0%** | **94,6%** | **94,6%** | **5,4%** |
| | | Sensibilidade | Especificidade | Exatidão | Erro |
| PCA-DA | Esverdeado | 64,0% | 93% | 85,5% | 14,5% |
| | Bom | 78% | 88,0% | 85,5% | 14,5% |
| | Ardido | 100,0% | 100,0% | 100,0% | 0,0% |
| | Queimado | 100,0% | 100,0% | 100,0% | 0,0% |
| | **Média do modelo** | **85,5%** | **95,2%** | **92,8%** | **7,3%** |
| | | Sensibilidade | Especificidade | Exatidão | Erro |
| PCA-MA | Esverdeado | 68,0% | 89% | 84,0% | 16,0% |
| | Bom | 68% | 89,3% | 84,0% | 16,0% |
| | Ardido | 100,0% | 100,0% | 100,0% | 0,0% |
| | Queimado | 100,0% | 100,0% | 100,0% | 0,0% |
| | **Média do modelo** | **84,0%** | **94,7%** | **92,0%** | **8,0%** |

FONTE: O autor (2023).

Para os dados de entrada da derivada, os modelos de PLS e SVM apresentam 100% nas métricas e um erro de 0%. No entanto, para os outros 5 modelos, todos apresentaram dificuldades no treinamento para algumas das classes, apresentando um desempenho total menor que os 100%, e consequente erro. Isso mostra que, para os dados da derivada, apesar do erro ser no máximo de 8% para o modelo de PCA-MA(Mahalanobis), os modelos desempenharam melhor no treinamento dos dados brutos. Isto é, dos 7 modelos, 5 apresentaram um desempenho de 100% nas classes, enquanto nos dados utilizando a derivada como entrada, somente 2 modelos apresentaram esse resultado. Isso representa que para os modelos, os dados brutos, que consideram aspectos físicos também, possuem uma importância na distinção das classes.

Quando analisamos desempenho dos modelos, é importante para uma avaliação precisa inferir a validade do modelo para um conjunto de amostras desconhecidas. Neste trabalho, foram preparados conjuntos de espectros de grãos não utilizados na etapa de treinamento (conjunto de predição). Desta maneira, nos Gráficos 15 e 16 é possível inferir o desempenho dos modelos treinados anteriormente com base no conjunto de predição.

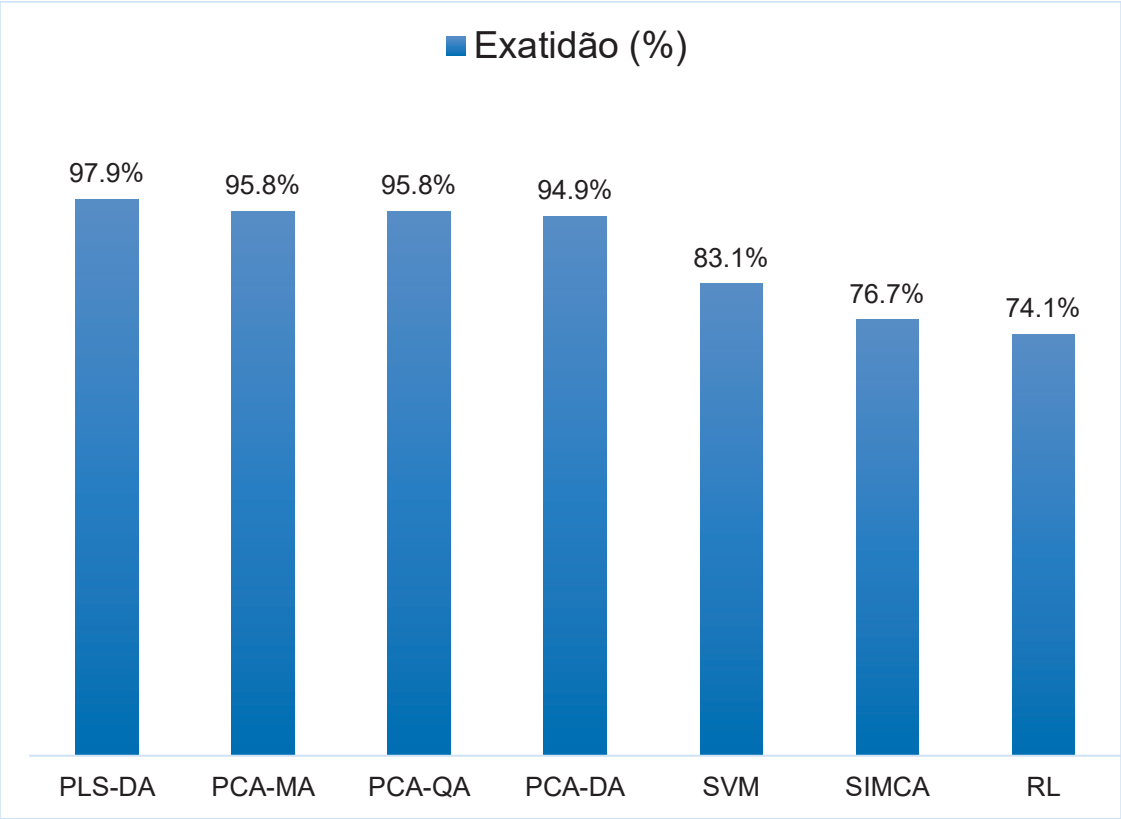
GRÁFICO 15 - RESULTADOS DA MÉDIA DA EXATIDÃO DOS MODELOS COM DADOS BRUTOS




FONTE: O autor (2023).

Nos resultados da predição dos modelos utilizando como entrada os dados brutos observam-se um bom desempenho do modelo PCA-DA, o qual apresenta uma exatidão média de 99,2% e um erro menor que 1%. Comparando aos demais modelos, notamos que em seguida vem o SVM, que é um modelo não linear, com uma precisão de 97% e depois o PLS-DA. A diferença entre os modelos que estão na 2ª e 3ª posição é de 1,2% no erro e exatidão, mas o SVM (2ª posição) se distancia do PCA-DA em 2,2%. Isso demonstra que o modelo PCA-DA se destaca pelo bom desempenho na predição dos dados de entrada bruto.

GRÁFICO 16 - RESULTADOS DA MÉDIA DA EXATIDÃO DOS MODELOS COM DADOS DERIVADA




FONTE: O autor (2023).

Os resultados encontrados com os dados de entrada da 1ª derivada, o modelo PLS-DA apresenta um desempenho que se destaca dos demais modelos, sendo uma exatidão média de 97,9% e um erro de 2,1%. Observando os modelos que apresentam um desempenho que ficaria na segunda posição o PCA-MA e PCA-QA, onde ambos possuem uma diferença de 2,1% abaixo do primeiro modelo em desempenho. O PCA-DA fica em uma terceira posição, com uma exatidão de 94,9%. Os demais modelos apresentam um desempenho menor que os demais. Para os valores de predição, os modelos não lineares SVM e RL apresentam desempenhos bem mais baixos que as posições acima.

Comparando os dados de predição dos dados brutos e derivada os valores de exatidão são melhores no primeiro. Mesmo os modelos que apresentam os valores de exatidão menores, quando comparado aos resultados da derivada, os valores encontrados dos dados brutos são melhores. Os modelos apresentam um melhor desempenho para os dados sem pré-tratamento, isso representa que os modelos se

adequam mais satisfatoriamente a essas leituras como entrada, sem suavização de características físicas dos grãos.

Embora inexistam na literatura trabalhos envolvendo o uso de espectrofotometria NIR para classificação de grãos de soja nos moldes desta dissertação, o trabalho de Chelladurai et al. (2014) se aproxima. Neste trabalho, os autores visavam a classificação de danos causados em parte por características biológica dos grãos (danos causados por infestações de insetos no grão), aplicando LDA e QDA aos dados após processamento via PCA. Os modelos obtidos resultaram em acurácias de 62% a 77% para conjuntos envolvendo 5 classes. Os melhores resultados para PCA-DA e para PCA-QA respectivamente foram de 40% e 80%. As classes analisadas no trabalho (Chelladurai et al., 2014), como explanado, foram bem diferentes das demonstradas no presente trabalho.

Nesta dissertação, verificou-se que a estratégia de modelagem PCA-DA mostrou-se bastante sensível às quatro classes aqui analisadas (esverdeados, bom, ardido e queimado). As métricas de desempenho de todos os modelos foram relacionadas na forma de tabelas no APÊNDICE I. A estratégia PLS-DA também resultou em boas classificações e predições, fornecendo elevada exatidão sobretudo para dados tratados com 1ª derivada, ficando bem próximos da exatidão obtida para o PCA-DA com dados brutos.

# 5 CONCLUSÕES

A utilização da espectrofotometria NIR para a classificação de produtos tem sido amplamente investigada. Entretanto, poucos trabalhos têm contemplado o uso de espectrofotômetro NIR portátil e de baixo custo na avaliação de processos e produtos. Em particular, o uso do espectrofotômetro MicroNIR para classificação de grãos de soja in-natura é ainda um tema inovador. De acordo com a legislação brasileira, as classificações dos grãos de soja e milho são estabelecidas de modo visual por operadores, uma prática passível de falhas tendo em vista os grandes volumes de produção do agronegócio nacional. Neste cenário, torna-se essencial o desenvolvimento de tecnologias que permitam o monitoramento em tempo real da qualidade dos grãos. Neste trabalho, mostrou-se que o uso combinado de técnicas de calibração multivariável e espectros obtidos via MicroNIR permite a classificação de grãos de soja in-natura. Foram testados espectros brutos e pré-tratados com 1ª derivada, além de diferentes modelos de classificação.

O modelo PCA-DA para os dados brutos foi o que apresentou o melhor desempenho na etapa de predição, devendo ser então recomendado para a classificação de grãos de soja. Este modelo linear se destaca pelo baixo valor do erro (1,7% para grãos esverdeados e bons) e elevada exatidão, resultando em um erro médio de 0,8% na etapa de predição. Em seguida, o modelo não linear SVM, que apresentou valores abaixo do PCA-DA, mas melhores que o desempenho dos demais modelos. A estratégia PLS-DA mostrou-se bastante efetiva sobretudo quando baseada em dados pré-tratados com a 1ª derivada, desempenhando de modo similar (porém abaixo) ao PCA-DA. É importante ressaltar que a estratégia de modelagem PCA-DA que apresentou melhor desempenho geral foi baseada em dados espectrais brutos, o que torna a estratégia ainda mais interessante pela simplicidade. Além disso, demonstra a importância da utilização de aspectos tanto químicos quanto físicos das amostras na classificação dos grãos de soja.

Desta maneira, é possível concluir que as leituras utilizando micro-NIR foram satisfatórias, além de ser um equipamento portátil e de fácil manuseio, suas leituras apresentaram bons resultados quando inseridas em modelos para classificação, atendendo aos objetivos pontuados pelo trabalho.

## 5.1 RECOMENDAÇÕES PARA TRABALHOS FUTUROS

Como mostrado neste trabalho há diversos parâmetros de qualidade no recebimento e transporte dos grãos de soja. O presente trabalho teve como objetivo a classificação via espectrofotometria NIR de grãos de soja em termos de parâmetros de qualidade tais como grãos esverdeados, ardidos, bons e queimados.

Para futuros trabalhos recomenda-se a aplicação do micro-NIR para uma maior quantidade de amostras de grão de soja com classes diferentes, a fim de buscar a classificação de demais características que são importantes no recebimento e transporte dos grãos. Além disso, estima-se que outros modelos de aprendizagem de máquina também possam apresentar bom desempenho analisando mais tipos de classes.

Por fim, sugere-se também o estudo aplicado para amostras de grãos de milho, que possuem classificações específicas, com a mesma metodologia de classificação manual. Espera-se que o uso combinado das técnicas multivariáveis aqui estudadas e o espectrofotômetro MicroNIR permita classificar corretamente amostras de grãos de milho.

# REFERÊNCIAS

Alencar, E. R. De, Faroni, L. R. D., Filho, A. F. L., Peternelli, L. A., & Costa, A. R. (2009). **Qualidade dos grãos de soja armazenados em diferentes condições** Quality of soy bean grains stored under different conditions. 31, 606–613.

APROSOJA. (2018). ECONOMIA. https://aprosojabrasil.com.br/a-soja/economia/. Acesso em: 16 Set. 2021.

APROSOJA. (2021). **A história da soja.** Aprosoja Mato Grosso. http://www.aprosoja.com.br/soja-e-milho/a-historia-da-soja. Acesso em: 20 Out. 2021.

Baianu, I., Guo, J., Nelson, R., You, T., & Costescu, D. (2011). NIR Calibrations for Soybean Seeds and Soy Food Composition Analysis: Total Carbohydrates, Oil, Proteins and Water Contents [v.2]. **Nature Precedings**. https://doi.org/10.1038/npre.2011.6611.2. Acesso em: Janeiro 2023.

Brereton, R. G. (2003). **Chemometrics: data analysis for the laboratory and chemical plant.**

Burns, D. A., & Ciurczak, E. W. (2008). **Handbook of Near-Infrared Analysis** (Third Edition, Vol. 35).

Casiraghi, E. (2017). Comparing the analytical performances of Micro-NIR and FT-NIR spectrometers in the evaluation of acerola fruit quality, using PLS and SVM regression algorithms. **Talanta**, 165, 112–116. https://doi.org/10.1016/j.talanta.2016.12.035. Acesso em: Janeiro 2023.

Côrtes, F. da S. (2022). **Modelo de predição da ocorrência de ferrugem asiática na cultura da soja a partir de variáveis climáticas e clusterização**. Universidade Federal de Goiás.

Chelladurai, V., Karuppiah, K., Jayas, D. S., Fields, P. G., & White, N. D. G. (2014). Detection of Callosobruchus maculatus (F.) infestation in soybean using soft X-ray and

NIR hyperspectral imaging techniques. **Journal of Stored Products Research**, 57, 43–48. https://doi.org/10.1016/j.jspr.2013.12.005. Acesso em: Janeiro 2023.

Devos, O., Ruckebusch, C., Durand, A., Duponchel, L., & Huvenne, J. P. (2009). Support vector machines (SVM) in near infrared (NIR) spectroscopy: Focus on parameters optimization and model interpretation. **Chemometrics and Intelligent Laboratory Systems**, 96(1), 27–33. https://doi.org/10.1016/j.chemolab.2008.11.005. Acesso em: Março 2023.

EMBRAPA. (2016). **História**. https://www.embrapa.br/soja/cultivos/soja1/historia. Acesso em: Outubro 2021.

Ferreira, D. S., Galão, O. F., Pallone, J. A. L., & Poppi, R. J. (2014). Comparison and application of near-infrared (NIR) and mid-infrared (MIR) spectroscopy for determination of quality parameters in soybean samples. **Food Control**, 35(1), 227–232. https://doi.org/10.1016/j.foodcont.2013.07.010. Acesso em: Fevereiro 2023.

Goelzer Sabin, J., Flôres Ferrão, M., & Carlos Furtado, J. (2004). Análise multivariada aplicada na identificação de fármacos antidepressivos. Parte II: Análise por componentes principais (PCA) e o método de classificação SIMCA. **Revista Brasileira de Ciências Farmacêuticas Brazilian Journal of Pharmaceutical Sciences**, 40(3).

Gonzalez Azevedo, L. (2018). **Regressão Logística e suas Aplicações**.
Malegori, C., Nascimento Marques, E. J., de Freitas, S. T., Pimentel, M. F., Pasquini, C., &

Hart, J. R., Norris, K. H., & Golumbic, G. (1962). Determination of the moisture content of seeds by near-infrared spectrophotometry of their methanol extracts. **In Cereal Chemistry** (Vol. 39, pp. 94–99).

Instrução Normativa 11/2007, (2007).

Kaufmann, K. C., Sampaio, K. A., García-Martín, J. F., & Barbin, D. F. (2022). Identification of coriander oil adulteration using a portable NIR spectrometer. **Food Control**, 132. https://doi.org/10.1016/j.foodcont.2021.108536. Acesso em: Fevereiro 2023.

Kosmowski, F., & Worku, T. (2018). Evaluation of a miniaturized NIR spectrometer for cultivar identification: The case of barley, chickpea and sorghum in Ethiopia. **PLoS ONE**, 13(3), 1–17. https://doi.org/10.1371/journal.pone.0193620. Acesso em: Janeiro 2023.

Liu, K. (2016). Soybean: Overview. **In Encyclopedia of Food Grains: Second Edition** (2nd ed., Vols. 1–4). Elsevier Ltd. https://doi.org/10.1016/B978-0-12-394437-5.00028-0. Acesso em: Janeiro 2023.

Lutz, M. (2007). ***Learning Python*** (3rd. Edition). California: O 'Reilly, 2008.

Malegori, C., Nascimento Marques, E. J., de Freitas, S. T., Pimentel, M. F., Pasquini, C., & Casiraghi, E. (2017). Comparing the analytical performances of Micro-NIR and FT-NIR spectrometers in the evaluation of acerola fruit quality, using PLS and SVM regression algorithms. **Talanta,** 165, 112–116. https://doi.org/10.1016/j.talanta.2016.12.035. Acesso em: Janeiro 2023.

Nicolaï, B. M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K. I., & Lammertyn, J. (2007). Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. **Postharvest Biology and Technology**, 46(2), 99–118. https://doi.org/10.1016/j.postharvbio.2007.06.024. Acesso em: Fevereiro 2023.

Onetta, J. de S., & Ruffato, S. (2018). **Propriedades Físicas de Grãos de Soja em Diferentes Safras**. VII Conferência Brasileira de Pós-Colheita, 593–599. https://eventos.abrapos.org.br/anais/paperfile/910_20181103_02-45-08_837.pdf

Sampaio, P. S., Castanho, A., Almeida, A. S., Oliveira, J., & Brites, C. (2020). Identification of rice flour types with near-infrared spectroscopy associated with PLS-DA and SVM methods. ***European Food Research and Technology***, *246*(3), 527–537. https://doi.org/10.1007/s00217-019-03419-5. Acesso em: Fevereiro 2023.

Scikit-learn developers. (2023). **Scikit-Learn Cross Decomposition**. Https://Scikit-Learn.Org/Stable/Modules/Cross_decomposition.Html#cross-Decomposition. Acesso em: Março 2023.

scikit-learn developers. (2023). **Linear Models**. Scikit-Learn.Org. Acesso em: 20 Abril 2023.

Scikit-learn developers. (2023). **Scikit-Learn Cross Decomposition**. https://Scikit-Learn.Org/Stable/Modules/Cross_decomposition.Html#cross-Decomposition. Acesso em: Abril 2023.

SENAR. (2017). **Grãos: classificação de soja e milho** (178th ed.).

Wang, L., Huang, Z., & Wang, R. (2021). Discrimination of cracked soybean seeds by near-infrared spectroscopy and random forest variable selection. **Infrared Physics and Technology**, 115. https://doi.org/10.1016/j.infrared.2021.103731. Acesso em: Fevereiro 2023.

Wang, D., Ram, M. S., & Dowell, F. E. (2002.). Classification of damaged soybean seeds using near-infrared spectroscopy. Transactions of the **ASAE,** 45(6), 1943–1948.

Zhu, Z., Chen, S., Wu, X., Xing, C., & Yuan, J. (2018). Determination of soybean routine quality parameters using near-infrared spectroscopy. **Food Science and Nutrition**, 6(4), 1109–1118. https://doi.org/10.1002/fsn3.652. Acesso em: Fevereiro 2023.

Yang, J., Li, J., Hu, J., Yang, W., Zhang, X., Xu, J., Zhang, Y., Luo, X., Ting, K. C., Lin, T., & Ying, Y. (2022). An interpretable deep learning approach for calibration transfer

among multiple near-infrared instruments. **Computers and Electronics in Agriculture**, 192**.** https://doi.org/10.1016/j.compag.2021.106584. Acesso em: Março 2023.

# APÊNDICE I – TABELAS PREDIÇÃO

RESULTADOS PREDIÇÃO PCA-DA LINEAR

**PCA Bruto e Derivada – Linear**

| | Sensibilidade | Especificidade | Exatidão | Erro |
|---|---|---|---|---|
| Esverdeado - Bruto | 90,0% | 100% | 98,3% | 1,7% |
| Bom - Bruto | 100% | 98,0% | 98,3% | 1,7% |
| Ardido-Bruto | 100,0% | 100,0% | 100,0% | 0,0% |
| Queimado-Bruto | 100,0% | 100,0% | 100,0% | 0,0% |
| Esverdeado - Derivada | 60,0% | 98,0% | 91,5% | 8,5% |
| Bom - Derivada | 90,0% | 91,8% | 91,5% | 8,5% |
| Ardido - Derivada | 100% | 97,4% | 98,3% | 1,7% |
| Queimado - Derivada | 94,7% | 100% | 98,3% | 1,7% |

FONTE: O autor (2023).

RESULTADOS PREDIÇÃO PCA-DA QUADRÁTICO

**PCA Bruto e Derivada – Quadrática**

| | Sensibilidade | Especificidade | Exatidão | Erro |
|---|---|---|---|---|
| Esverdeado - Bruto | 70,0% | 98,0% | 93,2% | 6,8% |
| Bom - Bruto | 90,0% | 93,9% | 93,2% | 6,8% |
| Ardido-Bruto | 55,0% | 100% | 84,7% | 15,3% |
| Queimado-Bruto | 100% | 81,3% | 86,6% | 13,4% |
| Esverdeado - Derivada | 80,0% | 95,9% | 93,2% | 6,8% |
| Bom - Derivada | 80,0% | 95,9% | 93,2% | 6,8% |
| Ardido - Derivada | 95% | 100% | 98% | 1,7% |
| Queimado - Derivada | 100% | 98% | 98% | 1,7% |

FONTE: O autor (2023).

RESULTADOS PREDIÇÃO PCA-DA MAHALANOBIS

**PCA Bruto e Derivada – Mahalanobis**

| | Sensibilidade | Especificidade | Exatidão | Erro |
|---|---|---|---|---|
| Esverdeado - Bruto | 80,0% | 98,0% | 94,9% | 5% |
| Bom - Bruto | 90,0% | 95,9% | 94,9% | 5% |
| Ardido-Bruto | 55,0% | 100% | 84,7% | 15% |
| Queimado-Bruto | 100% | 77,5% | 84,7% | 15% |
| Esverdeado - Derivada | 90,0% | 93,9% | 93,2% | 7% |
| Bom - Derivada | 70,0% | 98,0% | 93,2% | 7% |
| Ardido - Derivada | 95,0% | 100,0% | 98,3% | 2% |
| Queimado - Derivada | 100,0% | 97,5% | 98,3% | 2% |

FONTE: O autor (2023).

RESULTADOS PREDIÇÃO SIMCA

**SIMCA Bruto e Derivada**

| | Sensibilidade | Especificidade | Exatidão | Erro |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| Esverdeado - Bruto | 60,0% | 100,0% | 93,2% | 6,8% |
| Bom - Bruto | 40,0% | 98,0% | 88,1% | 11,9% |
| Ardido-Bruto | 20,0% | 100,0% | 72,9% | 27,1% |
| Queimado-Bruto | 42,1% | 100,0% | 81,0% | 19,0% |
| Esverdeado - Derivada | 10,0% | 100,0% | 84,7% | 15,3% |
| Bom - Derivada | 20,0% | 100,0% | 86,4% | 13,6% |
| Ardido - Derivada | 0,0% | 100,0% | 66,1% | 33,9% |
| Queimado - Derivada | 5,3% | 100,0% | 69,5% | 30,5% |

FONTE: O autor (2023).

RESULTADO PREDIÇÃO REGRESSÃO LOGÍSTICA

| **RL Bruto e Derivada** | | | | |
|---|---|---|---|---|
| | Sensibilidade | Especificidade | Exatidão | Erro |
| Esverdeado - Bruto | 90,0% | 98,0% | 96,6% | 3,4% |
| Bom - Bruto | 100,0% | 98,0% | 98,3% | 1,7% |
| Ardido-Bruto | 50,0% | 90,7% | 79,7% | 20,3% |
| Queimado-Bruto | 100,0% | 100,0% | 100,0% | 0,0% |
| Esverdeado - Derivada | 50,0% | 100,0% | 83,1% | 17,0% |
| Bom - Derivada | 0,0% | 79,6% | 79,6% | 20,4% |
| Ardido - Derivada | 0,0% | 66,1% | 66,1% | 33,9% |
| Queimado - Derivada | 0,0% | 67,8% | 67,8% | 32,2% |

FONTE: O autor (2023).

RESULTADO PREDIÇÃO PLS-DA DADOS BRUTOS

| **PLS-DA Bruto e Derivada** | | | | |
|---|---|---|---|---|
| | Sensibilidade | Especificidade | Exatidão | Erro |
| Esverdeado – Bruto | 90,0% | 95,9% | 94,9% | 5,1% |
| Bom – Bruto | 100,0% | 98,0% | 98,3% | 1,7% |
| Ardido-Bruto | 90,0% | 95,1% | 93,4% | 6,6% |
| Queimado-Bruto | 90,0% | 100,0% | 96,7% | 3,3% |
| Esverdeado – Derivada | 100,0% | 94,0% | 95,0% | 5,0% |
| Bom – Derivada | 100,0% | 100,0% | 100,0% | 0,0% |
| Ardido – Derivada | 90,0% | 100,0% | 96,6% | 3,4% |
| Queimado – Derivada | 100,0% | 100,0% | 100,0% | 0,0% |

FONTE: O autor (2023).

RESULTADOS SVM PARA DADOS DE PREDIÇÃO

| **SVM Bruto e Derivada** | | | | |
|---|---|---|---|---|
| | Sensibilidade | Especificidade | Exatidão | Erro |
| Esverdeado – Bruto | 100,0% | 98,0% | 98,3% | 1,7% |
| Bom – Bruto | 100,0% | 100,0% | 100,0% | 0,0% |
| Ardido-Bruto | 77,8% | 95,1% | 89,8% | 10,2% |
| Queimado-Bruto | 100,0% | 100,0% | 100,0% | 0,0% |
| Esverdeado – Derivada | 0% | 83% | 83% | 17,0% |
| Bom – Derivada | 0% | 83% | 83% | 17,0% |

| | | | | |
|---|---|---|---|---|
| Ardido – Derivada | 0% | 66,1% | 66,1% | 33,9% |
| Queimado – Derivada | 100,0% | 100,0% | 100,0% | 0,0% |

FONTE: O autor (2023).

RANKING COMPARATIVO DOS RESULTADOS DA PREDIÇÃO DOS MODELOS COM DADOS BRUTOS

| **RANKING MODELOS - PREDIÇÃO - DADOS BRUTOS** | | | | | |
|---|---|---|---|---|---|
| | | Sensibilidade | Especificidade | Exatidão | Erro |
| PCA-DA | Esverdeado | 90,0% | 100,0% | 98,3% | 1,7% |
| | Bom | 100,0% | 98,0% | 98,3% | 1,7% |
| | Ardido | 100,0% | 100,0% | 100,0% | 0,0% |
| | Queimado | 100,0% | 100,0% | 100,0% | 0,0% |
| | **Média do modelo** | **97,5%** | **99,5%** | **99,2%** | **0,8%** |
| | | Sensibilidade | Especificidade | Exatidão | Erro |
| SVM | Esverdeado | 100,0% | 98,0% | 98,3% | 1,7% |
| | Bom | 100,0% | 100,0% | 100,0% | 0,0% |
| | Ardido | 77,8% | 95,1% | 89,8% | 10,2% |
| | Queimado | 100,0% | 100,0% | 100,0% | 0,0% |
| | **Média do modelo** | **94,4%** | **98,3%** | **97,0%** | **3,0%** |
| | | Sensibilidade | Especificidade | Exatidão | Erro |
| PLS-DA | Esverdeado | 90,0% | 96% | 94,9% | 5,1% |
| | Bom | 100% | 98,0% | 98,3% | 1,7% |
| | Ardido | 90,0% | 95,1% | 93,4% | 6,6% |
| | Queimado | 90,0% | 100,0% | 96,7% | 3,3% |
| | **Média do modelo** | **92,5%** | **97,3%** | **95,8%** | **4,2%** |
| | | Sensibilidade | Especificidade | Exatidão | Erro |
| RL | Esverdeado | 90,0% | 98% | 96,6% | 3,4% |
| | Bom | 100% | 98,0% | 98,3% | 1,7% |
| | Ardido | 50,0% | 90,7% | 79,7% | 20,3% |
| | Queimado | 100,0% | 100,0% | 100,0% | 0,0% |
| | **Média do modelo** | **85,0%** | **96,7%** | **93,6%** | **6,4%** |
| | | Sensibilidade | Especificidade | Exatidão | Erro |
| PCA-MA | Esverdeado | 80,0% | 98% | 94,9% | 5,1% |
| | Bom | 90% | 95,9% | 94,9% | 5,1% |
| | Ardido | 55,0% | 100,0% | 84,7% | 15,3% |
| | Queimado | 100,0% | 77,5% | 84,7% | 15,3% |
| | **Média do modelo** | **81,3%** | **92,8%** | **89,8%** | **10,2%** |
| | | Sensibilidade | Especificidade | Exatidão | Erro |
| PCA-QA | Esverdeado | 70,0% | 98% | 93,2% | 6,8% |
| | Bom | 90% | 93,9% | 93,2% | 6,8% |
| | Ardido | 55,0% | 100,0% | 84,7% | 15,3% |
| | Queimado | 100,0% | 81,3% | 86,6% | 13,4% |
| | **Média do modelo** | **78,8%** | **93,3%** | **89,4%** | **10,6%** |
| SIMCA | | Sensibilidade | Especificidade | Exatidão | Erro |

| | | | | | |
|---|---|---|---|---|---|
| | Esverdeado | 60,0% | 100% | 93,2% | 6,8% |
| | Bom | 40% | 98,0% | 88,1% | 11,9% |
| | Ardido | 20,0% | 100,0% | 72,9% | 27,1% |
| | Queimado | 42,1% | 100,0% | 81,0% | 19,0% |
| | **Média do modelo** | **40,5%** | **99,5%** | **83,8%** | **16,2%** |

FONTE: O autor (2023).

RANKING COMPARATIVO DOS RESULTADOS DA PREDIÇÃO DOS MODELOS COM DADOS DERIVADA

| **RANKING MODELOS - PREDIÇÃO - DADOS DERIVADA** | | | | | |
|---|---|---|---|---|---|
| | | Sensibilidade | Especificidade | Exatidão | Erro |
| PLS-DA | Esverdeado | 100,0% | 94,0% | 95,0% | 5,0% |
| | Bom | 100,0% | 100,0% | 100,0% | 0,0% |
| | Ardido | 90,0% | 100,0% | 96,6% | 3,4% |
| | Queimado | 100,0% | 100,0% | 100,0% | 0,0% |
| | **Média do modelo** | **97,5%** | **98,5%** | **97,9%** | **2,1%** |
| | | Sensibilidade | Especificidade | Exatidão | Erro |
| PCA-MA | Esverdeado | 90,0% | 93,9% | 93,2% | 6,8% |
| | Bom | 70,0% | 98,0% | 93,2% | 6,8% |
| | Ardido | 95,0% | 100,0% | 98,3% | 1,7% |
| | Queimado | 100,0% | 97,5% | 98,3% | 1,7% |
| | **Média do modelo** | **88,8%** | **97,3%** | **95,8%** | **4,2%** |
| | | Sensibilidade | Especificidade | Exatidão | Erro |
| PCA-QA | Esverdeado | 80,0% | 96% | 93,2% | 6,8% |
| | Bom | 80% | 95,9% | 93,2% | 6,8% |
| | Ardido | 95,0% | 100,0% | 98,3% | 1,7% |
| | Queimado | 100,0% | 97,5% | 98,3% | 1,7% |
| | **Média do modelo** | **88,8%** | **97,3%** | **95,8%** | **4,2%** |
| | | Sensibilidade | Especificidade | Exatidão | Erro |
| PCA-DA | Esverdeado | 60,0% | 98% | 91,5% | 8,5% |
| | Bom | 90% | 91,8% | 91,5% | 8,5% |
| | Ardido | 100,0% | 97,4% | 98,3% | 1,7% |
| | Queimado | 94,7% | 100,0% | 98,3% | 1,7% |
| | **Média do modelo** | **86,2%** | **96,8%** | **94,9%** | **5,1%** |
| | | Sensibilidade | Especificidade | Exatidão | Erro |
| SVM | Esverdeado | 0,0% | 83% | 83,1% | 17,0% |
| | Bom | 0% | 83,1% | 83,1% | 17,0% |
| | Ardido | 0,0% | 66,1% | 66,1% | 33,9% |
| | Queimado | 100,0% | 100,0% | 100,0% | 0,0% |
| | **Média do modelo** | **25,0%** | **83,1%** | **83,1%** | **17,0%** |
| | | Sensibilidade | Especificidade | Exatidão | Erro |
| SIMCA | Esverdeado | 10,0% | 100% | 84,7% | 15,3% |
| | Bom | 20% | 100,0% | 86,4% | 13,6% |

| | | | | | |
|---|---|---|---|---|---|
| | Ardido | 0,0% | 100,0% | 66,1% | 33,9% |
| | Queimado | 5,3% | 100,0% | 69,5% | 30,5% |
| | **Média do modelo** | **8,8%** | **100,0%** | **76,7%** | **23,3%** |
| | | Sensibilidade | Especificidade | Exatidão | Erro |
| RL | Esverdeado | 50,0% | 100% | 83,1% | 17,0% |
| | Bom | 0% | 79,6% | 79,6% | 20,4% |
| | Ardido | 0,0% | 66,1% | 66,1% | 33,9% |
| | Queimado | 0,0% | 67,8% | 67,8% | 32,2% |
| | **Média do modelo** | **12,5%** | **78,4%** | **74,1%** | **25,9%** |

FONTE: O autor (2023).