



UNIVERSIDADE ESTADUAL PAULISTA  
“JÚLIO DE MESQUITA FILHO”  
Campus de Ilha Solteira

**UNIVERSIDADE ESTADUAL PAULISTA “JÚLIO DE MESQUITA FILHO”  
FACULDADE DE ENGENHARIA  
CAMPUS DE ILHA SOLTEIRA**

**DTHENIFER CORDEIRO SANTANA**

**FENOTIPAGEM DE ALTA PRECISÃO DE CARACTERES INDUSTRIAIS DE SOJA  
USANDO APRENDIZAGEM DE MÁQUINA E DIFERENTES SENSORES  
ESPECTRAIS**

Ilha Solteira  
2023

# **PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA**

**DTHENIFER CORDEIRO SANTANA**

## **FENOTIPAGEM DE ALTA PRECISÃO DE CARACTERES INDUSTRIAIS DE SOJA USANDO APRENDIZAGEM DE MÁQUINA E DIFERENTES SENSORES ESPECTRAIS**

Tese apresentada à Faculdade de Engenharia de  
Ilha Solteira – Unesp como parte dos requisitos  
para obtenção do título de Doutor.<sup>a</sup> Especialidade:  
Produção vegetal.

Paulo Eduardo Teodoro  
**Orientador**

Larissa Pereira Ribeiro Teodoro  
**Coorientadora**

Luciano Shozo Shiratsuchi  
**Coorientador**

Ilha Solteira  
2023

FICHA CATALOGRÁFICA

Desenvolvido pelo Serviço Técnico de Biblioteca e Documentação

S232f Santana, Dthenifer Cordeiro.  
Fenotipagem de alta precisão de caracteres industriais de soja usando  
aprendizagem de máquina e diferentes sensores espectrais / Dthenifer Cordeiro  
Santana. -- Ilha Solteira: [s.n.], 2023  
95 f. : il.

Tese (doutorado) - Universidade Estadual Paulista. Faculdade de Engenharia  
de Ilha Solteira. Área de conhecimento: Sistemas de Produção, 2023

Orientador: Paulo Eduardo Teodoro  
Co-orientador: Larissa Pereira Ribeiro Teodoro  
Co-orientador: Luciano Shozo Shiratsuchi  
Inclui bibliografia

1. Multiespectral. 2. Hiperespectral. 3. Comprimento de onda. 4. Inteligência  
computacional.

  
**Amanda Sertori dos Santos**

Bibliotecária - CRB/8-9061  
Seção Técnica de Referência, Atendimento ao  
Usuário e Documentação  
Diretoria Técnica de Biblioteca e Documentação

## **DEDICATÓRIA**

Dedico enfaticamente esta pesquisa aos meus pais, Areni Cordeiro Batista Santana e Sebastião Luis Santana, cuja dedicação incansável em me incentivar e apoiar foi de valor inestimável para minha formação acadêmica. Expresso também minha dedicação ao meu professor orientador, Paulo Eduardo Teodoro, cuja contribuição foi de suma importância para este trabalho, fornecendo ensinamentos preciosos, paciência inexaurível e constante motivação. Por fim, estendo essa dedicação a todos aqueles que compartilham da crença na importância da educação como ferramenta fundamental para o desenvolvimento pessoal e profissional.

“Se você não consegue explicar alguma coisa de forma simples, é porque ainda não entendeu bem o suficiente.”

Albert Einstein

## AGRADECIMENTOS

Primeiramente, desejo expressar meus agradecimentos de forma profunda reconhecendo o papel divino em meu percurso profissional, apesar da minha falta de assiduidade religiosa e espiritual. Inúmeras circunstâncias e eventos que se desdobraram ao longo do tempo foram impulsionados por forças transcendentais, e sou grata por cada acontecimento, encontro com pessoas e escolhas que me conduziram ao ponto em que me encontro hoje. Tais resultados se devem a uma influência maior, inexplicável por meras causas naturais, que direcionaram os eventos de maneira tão significativa.

Neste contexto, expresso minha sincera gratidão a meus pais, Areni Cordeiro Batista Santana e Sebastião Luis Santana, cujo apoio e crença inabaláveis em mim têm sido fundamentais em minha trajetória de pesquisa. Apesar de não compreenderem completamente minha atividade na área, eles sempre estiveram ao meu lado, oferecendo suporte emocional, encorajamento e segurança ao longo de toda a jornada acadêmica. Minha mãe tem sido meu refúgio afetuoso e acolhedor, enquanto meu pai tem contribuído com conselhos sábios e genuíno interesse pelo meu trabalho, ambos incentivando-me a buscar constantemente o aprimoramento por meio da educação. Meus agradecimentos se estendem igualmente a amigos queridos, como Isadora Luciano de Andrade, Ana Caroline Mello Arevalo e Lucas Hebert Durães dos Santos cujo apoio afetuoso, mesmo à distância, tem sido inestimável nos momentos mais importantes. Além disso, sou profundamente grata ao círculo de amigos eternos da graduação, Douglas Schlosser Machado, Lucas Areco Loureiro, Leandro Lima do Amaral, Luiz Felipe Busanello, Yuri Leite Cáceres e Vinicius Gabriel Beltramim de Oliveira, cuja amizade e companheirismo têm sido valiosos.

Destaco também minha amizade com Gustavo Casadei, cujos abraços reconfortantes e celebrações ao longo do doutorado foram de grande significado para mim. Uma menção especial vai para minha psicóloga, Beatrice Verdelho, cujo auxílio tem sido vital para meu contínuo aprimoramento pessoal, tornando o processo acadêmico mais leve e produtivo.

Sinto-me extremamente grata por ter encontrado anjos em minha jornada acadêmica, cuja amizade tem sido uma verdadeira bênção para toda a vida. Entre eles, Tatiane Deoti Pelissari, minha primeira amiga no doutorado, cuja companhia nos tempos de pandemia foi compartilhada com todas as alegrias e angústias. Izabela Cristina de Oliveira, além de uma querida amiga, tornou-se uma parceira de vida e pesquisa, contribuindo imensamente para enriquecer meu doutorado com risadas, reflexões, momentos de descontração e ideias valiosas. João Lucas Gouveia de Oliveira, meu confidente, fonte de paz, alegria e serenidade, com quem

compartilhei parte da minha jornada pessoal e acadêmica, tornando-a mais leve e feliz. A todos vocês, expresso meu imenso agradecimento por tudo o que têm feito e representado em minha vida.

Sou grata aos integrantes do grupo GEMP, cuja contribuição foi essencial para o desenvolvimento do meu trabalho. Especial reconhecimento aos colegas de pesquisa e amigos queridos Elber Vinicius Martins Silva, Paulo Henrique Menezes das Chagas, Paulo Ademir Dutra Rodrigues Júnior, Luis Gustavo Martins Santos, Yann Victor Muniz de Araújo Portugal, Guilherme Alves Bertasso, Murilo Fortes de Melo, Gustavo Casalino Amorim da Silva e Felipe Carrijo de Araújo.

Além de Deus, meus pais e meus queridos amigos, reservo uma profunda gratidão ao meu orientador, professor Dr. Paulo Eduardo Teodoro. Nenhuma palavra ou gesto será capaz de expressar todo o meu apreço por tê-lo como meu orientador. Sua dedicação, confiança, parceria, ensinamentos e motivação foram fundamentais para o meu crescimento acadêmico e profissional. Acreditar em mim e no meu trabalho, além de abrir portas e construir pontes, possibilitou minha rápida evolução na carreira acadêmica, culminando na conclusão do mestrado em um ano, do doutorado em dois anos e meio e ingresso em um pós-doutorado, assim como na publicação de diversos artigos e aquisição de valiosos conhecimentos.

Desejo também expressar minha gratidão à Profa. Dra. Larissa Pereira Ribeiro Teodoro, que tem sido meu maior exemplo de pesquisadora e me acolhido com doçura e amizade. Agradeço aos professores Carlos Antonio da Silva Junior, Fábio Henrique Rojo Baio, Cid Naudi Silva Campos, Cassiano Garcia Roque e Ana Carina da Silva Cândido Seron pela parceria e confiança em meu trabalho, sendo uma honra trabalhar ao lado de profissionais tão competentes.

Por fim, este trabalho só foi possível graças ao apoio concedido pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Agradeço, portanto, à CAPES pelo suporte financeiro que tornou possível a realização deste projeto acadêmico.

## RESUMO

A utilização de fenotipagem de alta precisão (HTP) no melhoramento de plantas pode auxiliar pesquisadores na seleção mais precisa e rápida de genótipos a partir de seu fenótipo. Aliar técnicas de aprendizagem de máquinas pode facilitar o processamento de dados advindos de distintos sensores que podem ser utilizados na fenotipagem. No capítulo 1 foi utilizado um sensor multiespectral, coletando reflectâncias na faixa do visível obtidos nas seguintes bandas espectrais (SBs): vermelho (660 nm), verde (550 nm), Rededge (735 nm) e NIR (790 nm) e realizados cálculos de índices de vegetação (IVs), objetivando encontrar algoritmos de aprendizagem de máquina e inputs com maior precisão para classificar genótipos de soja com base em características industriais. No capítulo 2 foi utilizado um sensor hiperespectral, que permitiu capturar a reflectância entre os comprimentos de onda entre 450 e 824 nm objetivando: I) classificar genótipos de soja de acordo com as características industriais do grão; II) identificar o(s) algoritmo(s) com melhor acurácia para a classificação dos genótipos com a reflectância das folhas; III) identificação do melhor dado de entrada (input) dos algoritmos que melhoram sua performance. O capítulo 3 utilizou-se sensor hiperespectral com capacidade de medir o espectro na faixa de 350 a 2500 nm objetivando avaliar a acurácia de técnicas de aprendizagem de máquinas em distinguir genótipos de soja de acordo com caracteres industriais dos grãos. Os experimentos foram conduzidos na Universidade Federal de Mato Grosso do Sul, campus Chapadão do Sul em diferentes safras com distintos genótipos. As avaliações espectrais foram realizadas aos 60 dias após a emergência. Ao fim do ciclo foi avaliada a produtividade de grãos e realizada as análises laboratoriais dos caracteres industriais. Os dados foram submetidos a algoritmos de aprendizagem de máquina (ML): rede neural artificial (ANN), algoritmos de árvore de decisão J48 (J48), REPTree (DT) e RandomTree (Rt), floresta aleatória (RF) e regressão logística (LR), usada como controle. No capítulo 1, utilizando o sensor multiespectral os algoritmos de ML que alcançaram maiores precisões foram ANN, DT e SVM utilizando apenas bandas espectrais. No capítulo 2 os algoritmos SVM e J48 foram os que apresentaram melhor performance também utilizando bandas espectrais como input. E no capítulo 3 a regressão logística (RL) apresentou superioridade na classificação dos genótipos de soja com a utilização dos comprimentos de onda como input.

**Palavras-chave:** multiespectral; hiperespectral; comprimento de onda; inteligência computacional.

## ABSTRACT

The use of high-precision phenotyping (HTP) in plant breeding can help researchers in the more accurate and faster selection of genotypes based on their phenotype. Combining machine learning techniques can facilitate the processing of data from different sensors that can be used in phenotyping. In Chapter 1, a multispectral sensor was used, collecting reflectances in the visible range obtained in the following spectral bands (SBs): red (660 nm), green (550 nm), Rededge (735 nm) and NIR (790 nm) and calculations of vegetation indices (IVs) were performed, aiming to find machine learning algorithms and inputs with greater precision to classify soybean genotypes based on industrial characteristics. In chapter 2, a hyperspectral sensor was used, which allowed capturing the reflectance between wavelengths between 450 and 824 nm, aiming to: I) classify soybean genotypes according to the industrial characteristics of the grain; II) identify the algorithm(s) with the best accuracy for classifying genotypes with leaf reflectance; III) identification of the best input data of the algorithms that improve their performance. Chapter 3 used a hyperspectral sensor capable of measuring the spectrum in the range of 350 to 2500 nm, aiming to evaluate the accuracy of machine learning techniques in distinguishing soybean genotypes according to grain industrial traits. The experiments were conducted at the Federal University of Mato Grosso do Sul, Chapadão do Sul campus in different crops with different genotypes. Spectral evaluations were performed 60 days after emergence. At the end of the cycle, grain yield was evaluated and laboratory analyzes of industrial traits were carried out. Data were subjected to machine learning (ML) algorithms: artificial neural network (ANN), J48 decision tree (J48), REPTree (DT) and RandomTree (Rt) algorithms, random forest (RF) and logistic regression (LR), used as control. In Chapter 1, using the multispectral sensor, the ML algorithms that achieved greater precision were ANN, DT and SVM using only spectral bands. In chapter 2, the SVM and J48 algorithms were the ones that presented the best performance also using spectral bands as input. And in chapter 3, logistic regression (RL) showed superiority in the classification of soybean genotypes using wavelengths as input.

**Keywords:** multispectral; hyperspectral; wave-length; computational intelligence.

## SUMÁRIO

|     |                                                                                                                                                                                   |           |
|-----|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------|
| 1   | INTRODUÇÃO GERAL .....                                                                                                                                                            | 12        |
| 2   | REVISÃO BIBLIOGRÁFICA .....                                                                                                                                                       | 13        |
| 2.1 | Espectro Eletromagnético .....                                                                                                                                                    | 13        |
| 2.2 | Índices de Vegetação .....                                                                                                                                                        | 14        |
| 2.3 | Fenotipagem de Alta Precisão no Melhoramento Genético de Plantas .....                                                                                                            | 16        |
| 2.4 | Caracteres Industriais em Soja .....                                                                                                                                              | 17        |
| 2.5 | Aprendizagem de Máquina .....                                                                                                                                                     | 18        |
|     | REFERÊNCIAS.....                                                                                                                                                                  | 20        |
|     | <b>CAPÍTULO 1 – CLASSIFICAÇÃO DE GENÓTIPOS DE SOJA QUANTO A CARACTERES TECNOLÓGICOS INDUSTRIAIS USANDO UAV-MULTIESPECTRAL E APRENDIZAGEM DE MÁQUINA .....</b>                     | <b>26</b> |
| 1   | INTRODUÇÃO .....                                                                                                                                                                  | 27        |
| 2   | MATERIAL E MÉTODOS .....                                                                                                                                                          | 29        |
| 2.1 | Condução do experimento .....                                                                                                                                                     | 29        |
| 2.2 | Variáveis industriais.....                                                                                                                                                        | 30        |
| 2.3 | Aquisição e processamento das imagens multiespectrais .....                                                                                                                       | 31        |
| 2.4 | Modelos de Aprendizagem de Máquina e Análises Estatísticas.....                                                                                                                   | 32        |
| 3   | RESULTADOS .....                                                                                                                                                                  | 34        |
| 4   | DISCUSSÃO .....                                                                                                                                                                   | 40        |
| 5   | CONCLUSÃO.....                                                                                                                                                                    | 44        |
|     | REFERÊNCIAS.....                                                                                                                                                                  | 45        |
|     | <b>CAPÍTULO 2 – FENOTIPAGEM DE ALTO RENDIMENTO COM ESPECTROSCOPIA VIS/NIR NA CLASSIFICAÇÃO DE GENÓTIPOS DE SOJA QUANTO A PRODUTIVIDADE DE GRÃO E CARACTERES INDUSTRIAIS .....</b> | <b>51</b> |
| 1   | INTRODUÇÃO .....                                                                                                                                                                  | 52        |
| 2   | MATERIAL E MÉTODOS .....                                                                                                                                                          | 54        |
| 3   | RESULTADOS .....                                                                                                                                                                  | 58        |
| 4   | DISCUSSÃO .....                                                                                                                                                                   | 66        |
| 5   | CONCLUSÃO.....                                                                                                                                                                    | 68        |
|     | REFERÊNCIAS.....                                                                                                                                                                  | 69        |

|          |                                                                                                                                                                  |           |
|----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------|
|          | <b>CAPÍTULO 3 – CLASSIFICAÇÃO DE GENÓTIPOS DE SOJA QUANTO A<br/>PRODUTIVIDADE DE GRÃOS E CARACTERES INDUSTRIAIS<br/>UTILIZANDO ESPECTROSCOPIA VNIR-SWIR.....</b> | <b>73</b> |
| <b>1</b> | <b>INTRODUÇÃO .....</b>                                                                                                                                          | <b>74</b> |
| <b>2</b> | <b>MATERIAL E MÉTODOS .....</b>                                                                                                                                  | <b>76</b> |
| <b>3</b> | <b>RESULTADOS .....</b>                                                                                                                                          | <b>80</b> |
| <b>4</b> | <b>DISCUSSÃO .....</b>                                                                                                                                           | <b>88</b> |
| <b>5</b> | <b>CONCLUSÃO.....</b>                                                                                                                                            | <b>91</b> |
|          | <b>REFERENCIAS.....</b>                                                                                                                                          | <b>92</b> |
|          | <b>CONSIDERAÇÕES FINAIS .....</b>                                                                                                                                | <b>96</b> |
|          | <b>APÊNDICE A - Acessando o melhor modelo de cada capítulo .....</b>                                                                                             | <b>97</b> |

## 1 INTRODUÇÃO GERAL

A interação do genótipo da planta com o ambiente resulta no fenótipo que são avaliados por técnicas de fenotipagem. A fenotipagem tradicional demanda tempo, mão de obra que demanda investimento alto e equipamentos específicos dependendo do que será avaliado (SOUSA, 2014). Buscar técnicas que possibilitem maior exatidão e agilidade na determinação de características fenotípicas relacionando tais informações com respostas biológicas das plantas em diferentes condições de ambientes são alvo de pesquisa nos últimos anos (SANTOS & YASSITEPE, 2014).

Ao longo dos anos a genotipagem teve grandes avanços devido ao aprimoramento da biologia e engenharia molecular (MORAES, 2002), como as técnicas de sequenciamento genômico que se tornou um método eficiente e barato que geram uma grande quantidade de dados favorecendo a pesquisa em diversos âmbitos. Esse avanço não ocorreu da mesma forma com avaliações fenotípicas (SANTOS; YASSITEPE, 2014).

Nos últimos anos, novas técnicas de fenotipagem de plantas foram desenvolvidas utilizando sensores que captam informações emitidas pelas plantas. A coleta de informação por meio de imagens permite quantificar variáveis da planta de forma não destrutiva, com precisão e com regularidade ao longo do ciclo da planta (CHOUDHURY *et al.*, 2019). A utilização conjunta de sensores, computadores, robótica, aeronaves e inteligência computacional permitiu o desenvolvimento de plataformas e de modelagens para a fenotipagem de alto rendimento buscando a acurácia de informações sobre diversas culturas (SONG *et al.*, 2021). Dentre as plataformas está a utilização, principalmente, de satélites ou de aeronaves remotamente pilotadas (ARP) (JANG *et al.*, 2020). Esse tipo de plataforma é mais utilizado em melhoramento de culturas comerciais de grande escala (SONG *et al.*, 2021).

As plataformas que realizam fenotipagem de alta precisão pode ser divididas em duas categorias: fenotipagem morfológica; ou fisiológica (Li *et al.*, 2021). A reflectância emitida pelo dossel das plantas por meio de comprimentos de onda permite relacionar com características de caráter fisiológico, como teor de nitrogênio foliar, clorofila e estado hídrico da planta (BURUD *et al.*, 2017). Também é possível a mensuração de características morfológicas (VAN EEUWIJK *et al.*, 2019) como altura de plantas ou produtividade de grãos. Dessa forma, a fenotipagem utilizando sensores com auxílio de veículos aéreos não tripulados ou robôs permitem levantamento de grande número de informações sobre as plantas de forma não destrutiva e de maneira econômica (BURUD *et al.*, 2017).

## 2 REVISÃO BIBLIOGRÁFICA

### 2.1 Espectro Eletromagnético

Qualquer corpo com temperatura acima do zero absoluto emite energia eletromagnética, e pode ser considerado como uma fonte de energia. Essa energia pode ser composta de maneira contínua em função de seu comprimento de onda ou de sua frequência, sendo denominada de espectro eletromagnético (MORAES, 2002). A energia eletromagnética de maior quantidade da Terra é oriunda do Sol, que se propaga por meio de onda eletromagnéticas (NAUE *et al.*, 2011). A energia luminosa interage de forma distinta com os diversos alvos da Terra, o qual cada um possui sua própria “assinatura espectral”. Essa assinatura espectral diz respeito a quantidade de energia refletida ou absorvida de maneiras diferentes ao longo das faixas do espectro da luz incidente (FIGUEIREDO, 2005).

A energia eletromagnética possui subdivisões de acordo com a interação que ocorre entre o objeto e a radiação incidente, a transparência da atmosfera e o processo físico originário da energia eletromagnética. O espectro eletromagnético varia de ondas curtas até ondas de comprimentos maiores (MORAES, 2002). Os sensores que captam a radiação eletromagnética obtêm dados sobre os alvos de interesse da superfície terrestre que são transformados em dados passíveis de interpretação (NAUE *et al.*, 2011).

A energia luminosa, além de contribuir para a assinatura espectral dos objetos, contribui para a fotossíntese das plantas afetando seus processos fisiológicos. Além de que a porcentagem de solo coberta por plantas é significativa no planeta afetando a interceptação da radiação e na quantidade que é refletida ou absorvida por essas plantas (BABOLIM *et al.*, 2014). E esse comportamento espectral das plantas envolve desde análise de folhas, plantas isoladas ou por cobertura de vegetação formada com um conjunto de planta, chamado dossel (EDWIRGES *et al.*, 2017). Em culturas de interesse comercial pesquisadores e profissionais realizam todo o acompanhamento do ciclo das safras por meio dos comprimentos de onda refletidos pelas plantas avaliando estádios fenológicos desde a emergência até o enchimento dos grãos (KERN *et al.*, 2018).

Dentro do sensoriamento remoto há a classificação de sensores como passivos e ativos em que os sensores passivos necessitam da radiação do sol para poder capturar os comprimentos de onda do material em avaliação. Já os sensores ativos produzem sua própria radiação, emitindo energia eletromagnética para os objetos capturando parte dessa

energia refletida (MORAES, 2002). O sensoriamento remoto passivo, de maneira geral é separado em dois segundo a resolução espectral: multiespectral e hiperespectral

As câmeras chamadas de RGB são as mais utilizadas pelas aeronaves remotamente pilotadas em análises de fenotipagem em culturas. Esse tipo de sensor possui baixo custo, são leves e o processamento dos dados é relativamente simples. Entretanto, as informações fornecidas pelo RGB são limitadas quanto as características fisiológicas devido aos comprimentos de onda captados, ou seja, bandas espectrais além do RGB fornecem informações adicionais sobre a assinatura espectral da vegetação (TAIT *et al.*, 2019). Sensores multiespectrais são capazes de mensurar a reflectância emitida pelas plantas de partes visíveis e invisíveis do espectro eletromagnético. Já as imagens obtidas por meio de sensores hiperespectrais obtém bandas estreitas e de espectro contínuo (YANG *et al.*, 2017), ou seja, fornece informações muito maiores ao longo do espectro possibilitando respostas mais precisas sobre a fisiologia da planta.

## 2.2 Índices de Vegetação

A superfície das plantas quando saudáveis são visualmente verdes, porque as folhas refletem verde e absorve a luz azul e vermelha (DIN *et al.*, 2017) No comprimento de onda de 475 nm, na região do espectro azul, existe uma forte associação com a presença de xantofilas, carotenos e pigmentos de clorofilas  $\alpha$  e  $\beta$  (SILVA JUNIOR *et al.*, 2018; MOREIRA, 2005). Essa faixa de 475 nm, que pertence à região azul do espectro, está intimamente ligada à absorção de luz pelos pigmentos que conferem cores verdes a amarelas às folhas e são essenciais para a fotossíntese. Por outro lado, a faixa de comprimento de onda de 660 nm, que se encontra na região do vermelho, é altamente absorvida pelas clorofilas durante o processo de fotossíntese e pelos fitocromos para regular os processos de fotoperiodismo. Tanto a luz azul quanto a luz vermelha desempenham papéis cruciais em uma variedade de processos bioquímicos das plantas, especialmente relacionados à fotossíntese. Além dessas regiões do espectro eletromagnético, as faixas conhecidas como Rededge e infravermelho próximo (NIR) também desempenham um papel fundamental em processos relacionados ao fotoperiodismo e ao crescimento de tecidos (BERGAMASCHI & Bergonci, 2017).

A faixa do espectro eletromagnético que abrange a refletância da vegetação é influenciada pelas características da planta quanto a substâncias químicas e morfologia dos seus órgãos. Basicamente a fenotipagem de plantas é baseado na região do ultravioleta (UV) (de 10 a 380 nm), os espectros visíveis, que são compostos das regiões

de comprimento de onda azul (450–495 nm), verde (495–570 nm) e vermelho (620–735 nm) e a banda do infravermelho próximo e médio (850–1700 nm) (CRUDEN *et al.*, 2012; RAHIM *et al.*, 2016; XUE & SU, 2017; X. ZHANG *et al.*, 2014).

A radiação que chega aos sensores implantados nos ARP é medida por meio dos comprimentos de onda citados anteriormente (SHAW; BURKE, 2003). Por meio desses comprimentos de onda é possível realizar equações com um ou mais bandas do espectro eletromagnético, permitindo calcular as relações entre as bandas. Essas equações são chamadas de Índices de vegetação (IVs). Os IVs permitem adquirir informações sobre o fenótipo das culturas em meio a um grande número de indivíduos em qualquer estágio do ciclo (MONTESINOS-LÓPEZ *et al.*, 2017). Eles foram desenvolvidos para diversas finalidades, e na agricultura são usados para otimizar diversos processos avaliativos de interesse (BRANDAO, 2009).

A análise de severidade de doenças em diversas culturas geralmente é realizada por meio de avaliação visual de profissionais, que demanda tempo, mão de obra especializada e que pode gerar dúvidas devido a variação de notas que podem ser dadas a severidade da doença, de acordo com o profissional. Utilizar os IVs nesse processo, sobretudo aqueles que utilizam bandas do infravermelho e do infravermelho próximo, podem fornecer informações sobre a sanidade e nutrição da planta de forma mais rápida de precisa (LI *et al.*, 2014).

O Índice de Vegetação da Diferença Normalizada (NDVI) é um dos índices de vegetação mais conhecidos que dentre sua utilização está a predição de rendimento de grãos, análise da cobertura vegetal e informações quanto a senescência da cultura. Essas informações são alcançadas por meio da utilização de tonalidades de verdes, que são emitidos e medidos em distintos estágios de desenvolvimento da cultura (LI *et al.*, 2014). Todavia o NDVI apresenta uma limitação, em presença de quantidades elevadas de massa vegetal, e tende a saturar e não fornecer informações reais sobre algumas mudanças da cobertura vegetal. Isso ocorre devido a resposta espectral da faixa do infravermelho próximo, que pode saturar rapidamente (GITELSON *et al.*, 2002).

Outro índice bastante conhecido é o SAVI. Com ele não ocorre o problema de saturação que acontece com o NDVI. No SAVI ocorre a suavização dos efeitos do solo, especialmente em estádios iniciais de crescimento da cultura, quando não há fechamento da entrelinha, tornando os dados mais confiáveis (BRAZ *et al.*, 2015). Pode ter diversas utilizações, como na estimativa de índice de nutrição na cultura do milho (ZHAO *et al.*, 2018) e na predição de produção de grãos (VENANCIO *et al.*, 2020). O NDVI é útil para

monitorar o crescimento e a fenologia de diversas culturas (RYU *et al.*, 2021), por exemplo, o NDVI mostra uma correlação elevada com altura das plantas de milho, assim como o SAVI. E o NDRE possui uma associação maior com o teor de nitrogênio foliar das plantas de milho (FLORES *et al.*, 2020) e juntamente com o NDVI fornecem informações precisas quando ao rendimento de grãos da cultura do milho (SANTANA *et al.*, 2021).

### **2.3 Fenotipagem de Alta Precisão no Melhoramento Genético de Plantas**

A fenotipagem de alta precisão ainda é um gargalo, visto que seus avanços são recentes e assim desenvolvimento de protocolos para registros precisos ainda são escassos. Os avanços na fenotipagem de alta precisão são necessários para melhorar a eficiência de seleção de genótipos em programas de melhoramento, acelerando ganhos genéticos e monitorando de forma remota e automatizada a sanidade da cultura diminuindo assim perdas quantitativas e qualitativas. Outro ponto a ser levado em consideração é a eficiência econômica, para que torne mais atrativo para criadores e produtores rurais (CHAWADE *et al.*, 2019).

Para as análises de fenotipagem há a utilização de plataformas de ARP com asa fixa ou asa rotativa, equipados com sensores de imagens. Tais sensores podem obter comprimentos de onda no espectro visível (RGB), sensores infravermelhos, câmeras multiespectrais e sensores hiperespectrais (SONG *et al.*, 2021). Pesquisas recentes no ramo da fenotipagem de alto rendimento usando imagens adquiridas com auxílio das ARP possuem alto potencial na determinação de respostas das culturas a diversos tipos de estresses bióticos e abióticos (ZHOU *et al.*, 2020).

O foco do melhoramento de plantas é o desenvolvimento de novos genótipos de sejam de alguma forma melhores que os existentes. Tal melhora alcançada num determinado período de tempo é realizada por meio de seleção artificial, denominada ganho genético (XU *et al.*, 2017). O melhoramento busca maximizar esses ganhos genéticos gerenciando componentes como variância genética aditiva, intensidade de seleção, controle parental de machos e fêmeas e o tempo para gerar esse ganho. O melhoramento sugere que quanto menor o diferencial de seleção menor é a resposta a ele. E para alterar a frequência genética é preciso haver uma fenotipagem e uma seleção precisa, buscando sempre aumentar a frequência de alelos favoráveis (CHAWADE *et al.*, 2019).

Ensaio a campo para o melhoramento de plantas são compostos por diversas parcelas que avaliam diversos genótipos quanto as suas características agrônômicas. A replicação desses testes é bem comum, testando-os em diversos ambientes, avaliando as alterações entre os locais, épocas do ano ou entre anos, levando em conta a interação genótipo, ambiente e a influência de distintos manejos. Dessa forma se torna um desafio para o melhoramento de plantas quantificar diversas características dos genótipos em distintos ambientes, contornando o tempo necessário para essas mensurações, a precisão das informações colhidas e a análise desses dados. A fenotipagem de alta precisão auxilia na diminuição do tempo, na velocidade da seleção, melhora na precisão e dá suporte para a decisão dos melhores genótipos (ARAUS *et al.*, 2018).

Utilizar a fenotipagem de alta precisão para avaliar caracteres fisiológicos permite que os pesquisadores possam avaliar a cultura de forma não destrutiva e avaliar seu desempenho ao longo do ciclo da cultura (DHONDT *et al.*, 2013; HAIRMANSIS *et al.*, 2014; MUNNS *et al.*, 2010). As plataformas de fenotipagem permitem obter informações não destrutivas das plantas, com alto rendimento para estudos e características complexas como tolerâncias, resistências, fisiologia, arquitetura e crescimento (CHEN *et al.*, 2014).

## 2.4 Caracteres Industriais em Soja

A soja é uma das culturas mais importantes mundialmente devido ao conteúdo de seus grãos, ricos em proteínas, óleo, açúcares, ácidos graxos e aminoácidos. Devido a sua constituição, é mundialmente utilizada na alimentação de animais e humanos (RAHMAN *et al.*, 2023). O melhoramento e os manejos culturais de soja ao longo dos anos foram todos focados para que a cultura tivesse melhor desenvoltura produtiva ao maior número de condições adversas ambientais possíveis (MWIINGA *et al.*, 2020).

É indiscutível o esforço realizado por melhoristas em aumentar a produtividade de soja e melhorar sua performance quanto a resistência de pragas e doenças (RAHMAN *et al.*, 2023), todavia há uma lacuna a ser preenchida pelo melhoramento genético para atender às necessidades de cada segmento industrial (RANI & KUMAR, 2022). Dentre os constituintes da soja de maior importância está a proteína que chega a ser duas vezes maior do que a proteína presente na carne e quatro vezes maior do que a do leite (OGOKE *et al.*, 2003). O teor de proteína de uma cultivar comercial deve atender ao requisito de mercado de ser superior a 47,5% de proteína e de pelo menos 22% de teor de óleo (WILLIS, 2003). Caso o teor de óleo não atenda a esse requisito a quantidade de proteína

deve ser maior que o citado (PATIL *et al.*, 2017). Além de óleo e proteína, a soja possui fibras e carboidratos em torno de 12%, em que as fibras são definidas como fibras dietética, importantes para auxiliar na redução do nível de colesterol (RAHMAN *et al.*, 2023).

O melhoramento da soja tanto para produtividade quanto para qualidade industrial dos grãos é parcialmente dificultado pelos longos ciclos de geração das culturas, ou seja, a determinação desses fatores só será possível ao fim da colheita e com as análises laboratoriais dos grãos. Dessa forma a fenotipagem de alta precisão possibilita determinar genótipos com maior potencial para cada característica antes que o ciclo se encerre (YUAN *et al.*, 2019).

## 2.5 Aprendizagem de Máquina

Os resultados demonstrados pelas bandas espectrais ou pelos índices de vegetação podem gerar dados confusos que podem ser contornados por modelos estatísticos. Essa dificuldade se deve a alta correlação que as bandas espectrais possuem e que geram problemas estatísticos que afetam a precisão dos resultados e suas interpretações (FEILHAUER *et al.*, 2015). Dessa forma a utilização da aprendizagem de máquina surgiu como uma tecnologia que demonstra alto desempenho para desenvolver soluções multidisciplinares tecnológicas na agricultura (LIAKOS *et al.*, 2018). Essa é uma abordagem promissora devido à habilidade dessa técnica em extrair detalhes das imagens e vem sendo aplicada com sucesso na determinação de características em folhas de plantas (ZHOU *et al.*, 2019). A quantidade elevada de dados adquirida por meio da fenotipagem é uma dificuldade contornada com certa facilidade por técnicas de aprendizagem de máquinas, que consegue lidar com dados não lineares (CHLINGARYAN *et al.*, 2018).

Técnicas de aprendizagem de máquinas podem ser divididas em métodos supervisionados, semi-supervisionados e não supervisionados. Os métodos supervisionados são treinados com algoritmos de supervisão. Métodos não supervisionados usam padrões encontrados em conjuntos de dados sem o uso de rótulos. E os métodos semi supervisionados combinam as duas abordagens, que podem melhorar o poder de previsão (LIBBRECHT; NOBLE, 2015).

Zhou *et al.*, 2021, encontraram um modelo de previsão de produtividade de grãos usando aprendizagem de máquinas que resultou em uma previsão de 78% de exatidão. Tal resultado foi encontrado por meio de mensuração de características com

imagens adquiridas com ARP nos períodos iniciais e finais do crescimento reprodutivo da cultura da soja. Outro exemplo da aplicação dessa modelagem é na análise de imagens para classificação de tomates (ZABOROWICZ *et al.*, 2013) e na classificação de maturação de frutos de pêssgo (ALIPASANDI *et al.*, 2013).

Outro algoritmo utilizado para análise de dados é a Árvore de decisão, que possui poder preditivo com precisão, gerando resultados de fácil interpretação (CHASMER *et al.*, 2014). Almeida *et al.*, 2020 utilizaram árvore de decisão para observar a acurácia dos IVs na estimativa de produtividade na cultura soja, observaram que SAVI é preciso na predição de produtividades até 4,300 kg ha<sup>-1</sup> e que, maior que esse valor, o NDVI possui maior precisão. A seleção de variáveis preditivas por meio dessa técnica é realizada na construção do modelo, bem como a estrutura da árvore de decisão (FERNANDES *et al.*, 2014).

## REFERÊNCIAS

- ALIPASANDI, A.; GHAFARI, H.; ALIBEYGLU, S. Z. Classification of three varieties of peach fruit using artificial neural network assisted with image processing techniques. **International Journal of Agronomy and Plant Production**, [s. l.], v. 4, n. 9, p. 2179–2186, 2013.
- ARAUS, J. L.; KEFAUVER, S. C.; ZAMAN-ALLAH, M.; OLSEN, M. S.; CAIRNS, J. E. Translating high-throughput phenotyping into genetic gain. **Trends in plant science**, [s. l.], v. 23, n. 5, p. 451–466, 2018.
- BABOLIM, R. C. G.; SIMIONATO, U. R.; FERREIRA, A. S.; WERNER, F.; IWASAKI, G. S.; NEUMAIER, N.; DEBIASI, H.; FRANCHINI, J. C.; BALBINOT JUNIOR, A. A. Evolução do Índice de Vegetação por Diferença Normalizada (NDVI) da soja em semeadura cruzada e não cruzada. In: JORNADA ACADÊMICA DA EMBRAPA SOJA, 9., 2014, Londrina. **Resumos expandidos [...]**. Londrina: Embrapa Soja, 2014.
- BRANDAO, Z. N. Estimativa da produtividade e estado nutricional da cultura do algodão irrigado via técnicas de sensoriamento remoto. [s. l.], 2009.
- BRAZ, A. M.; DE ANDRADE ÁGUAS, T.; GARCIA, P. H. M. Análise de índices de vegetação NDVI e SAVI e Índice de Área Foliar (IAF) para a comparação da cobertura vegetal na bacia hidrográfica do córrego Ribeirãozinho, município de Selvíria–MS. **Revista Percursos**, [s. l.], v. 7, n. 2, p. 5–22, 2015.
- BURUD, I.; LANGE, G.; LILLEMO, M.; BLEKEN, E.; GRIMSTAD, L.; JOHAN FROM, P. Exploring Robots and UAVs as Phenotyping Tools in Plant Breeding. **IFAC-PapersOnLine**, [s. l.], v. 50, n. 1, p. 11479–11484, 2017.
- CHASMER, L.; HOPKINSON, C.; VENESS, T.; QUINTON, W.; BALTZER, J. A decision-tree classification for low-lying complex land cover types within the zone of discontinuous permafrost. **Remote Sensing of Environment**, [s. l.], v. 143, p. 73–84, 2014.
- CHAWADE, A.; VAN HAM, J.; BLOMQUIST, H.; BAGGE, O.; ALEXANDERSSON, E.; ORTIZ, R. High-throughput field-phenotyping tools for plant breeding and precision agriculture. **Agronomy**, [s. l.], v. 9, n. 5, p. 258, 2019.
- CHEN, D.; NEUMANN, K.; FRIEDEL, S.; KILIAN, B.; CHEN, M.; ALTMANN, T.; KLUKAS, C. Dissecting the phenotypic components of crop plant growth and drought responses based on high-throughput image analysis. **The plant cell**, [s. l.], v. 26, n. 12, p. 4636–4655, 2014.
- CHLINGARYAN, A.; SUKKARIEH, S.; WHELAN, B. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. **Computers and Electronics in Agriculture**, [s. l.], v. 151, p. 61–69, 2018.

- CRUDEN, B. A.; PRABHU, D.; MARTINEZ, R. Absolute radiation measurement in venus and mars entry conditions. **Journal of Spacecraft and Rockets**, [s. l.], v. 49, n. 6, p. 1069–1079, 2012.
- CHOUDHURY, S.; SAMAL, A.; AWADA, T. Leveraging image analysis for high-throughput plant phenotyping. **Frontiers in plant science**, [s. l.], v. 10, p. 508, 2019.
- DE ALMEIDA, M. R. D.; SILVA, J. N. M.; DE BARROS, P. L. C.; ALMEIDA, E. D. S.; DA SILVA, D. A. S.; DE SOUSA, C. S. C. Ajuste e seleção de modelos volumétricos de espécies comerciais no município de Ipixuna (PA). **Revista em Agronegócio e Meio Ambiente**, [s. l.], v. 13, n. 1, p. 259–278, 2020.
- DE MORAES, E. C. Capítulo 1 fundamentos de sensoriamento remoto. [s. l.], 2002.
- DE SOUSA, C. A. F. Fenotipagem de plantas: uma nova abordagem para um velho problema. **Embrapa Agroenergia-Comunicado Técnico (INFOTECA-E)**, [s. l.], 2014.
- DHONDT, S.; WUYTS, N.; INZÉ, D. Cell to whole-plant phenotyping: the best is yet to come. **Trends in plant science**, [s. l.], v. 18, n. 8, p. 428–439, 2013.
- DIN, M.; ZHENG, W.; RASHID, M.; WANG, S.; SHI, Z. Evaluating hyperspectral vegetation indices for leaf area index estimation of *Oryza sativa* L. at diverse phenological stages. **Frontiers in plant science**, [s. l.], v. 8, p. 820, 2017.
- EDWIRGES, T. R.; QUARTAROLI, C. F.; TOSTO, S. G.; AQUINO, E. P. S. D. E.; CAMARGO JUNIOR, A. A. Comportamento espectral de áreas cultivadas com seringueiras em imagens do sensor OLI/Landsat 8. In: CONGRESSO INTERINSTITUCIONAL DE INICIAÇÃO CIENTÍFICA, 11., Campinas, 2017. **Anais [...]**. Campinas: EMPRABA Soja 2017.
- FEILHAUER, H.; ASNER, G. P.; MARTIN, R. E. Multi-method ensemble selection of spectral bands related to leaf biochemistry. **Remote Sensing of Environment**, [s. l.], v. 164, p. 57–65, 2015. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0034425715001339>.
- FERNANDES, M. R.; AGUIAR, F. C.; SILVA, J. M. N.; FERREIRA, M. T.; PEREIRA, J. M. C. Optimal attributes for the object based detection of giant reed in riparian habitats: A comparative study between Airborne High Spatial Resolution and WorldView-2 imagery. **International Journal of Applied Earth Observation and Geoinformation**, [s. l.], v. 32, p. 79–91, 2014.
- FIGUEIREDO, D. Conceitos básicos de sensoriamento remoto. **São Paulo**, [s. l.], 2005.
- FLORES, M. da S.; PASCHOALETE, W. M.; BAILO, F. H. R.; CAMPOS, C. N. S.; PANTALEÃO, A. de A.; TEODORO, L. P. R.; SILVA JÚNIOR, C. A. Da; TEODORO, P. E. Relationship between vegetation indices and agronomic performance of maize varieties under different nitrogen rates. **Biosci. j.(Online)**, [s. l.], p. 1638–1644, 2020.

- GITELSON, A. A.; KAUFMAN, Y. J.; STARK, R.; RUNDQUIST, D. Novel algorithms for remote estimation of vegetation fraction. **Remote sensing of Environment**, [s. l.], v. 80, n. 1, p. 76–87, 2002.
- HAIRMANSIS, A.; BERGER, B.; TESTER, M.; ROY, S. J. Image-based phenotyping for non-destructive screening of different salinity tolerance traits in rice. **Rice**, [s. l.], v. 7, n. 1, p. 1–10, 2014.
- JANG, G.; KIM, J.; YU, J.-K.; KIM, H.-J.; KIM, Y.; KIM, D.-W.; KIM, K.-H.; LEE, C. W.; CHUNG, Y. S. Cost-Effective Unmanned Aerial Vehicle (UAV) Platform for Field Plant Breeding Application. **Remote Sensing**, [s. l.], v. 12, n. 6, p. 998, 2020.
- KERN, A.; BARCZA, Z.; MARJANOVIĆ, H.; ÁRENDÁS, T.; FODOR, N.; BÓNIS, P.; BOGNÁR, P.; LICHTENBERGER, J. Statistical modelling of crop yield in Central Europe using climate data and remote sensing vegetation indices. **Agricultural and forest meteorology**, [s. l.], v. 260, p. 300–320, 2018.
- LI, F.; LI, D.; ELSAYED, S.; HU, Y.; SCHMIDHALTER, U. Using optimized three-band spectral indices to assess canopy N uptake in corn and wheat. **European Journal of Agronomy**, [s. l.], v. 127, p. 126286, 2021.
- LI, L.; ZHANG, Q.; HUANG, D. A review of imaging techniques for plant phenotyping. **Sensors**, [s. l.], v. 14, n. 11, p. 20078–20111, 2014.
- LIAKOS, K. G.; BUSATO, P.; MOSHOU, D.; PEARSON, S.; BOCHTIS, D. Machine learning in agriculture: A review. **Sensors**, [s. l.], v. 18, n. 8, p. 2674, 2018.
- LIBBRECHT, M. W.; NOBLE, W. S. Machine learning applications in genetics and genomics. **Nature Reviews Genetics**, [s. l.], v. 16, n. 6, p. 321–332, 2015.
- MONTESINOS-LÓPEZ, O. A.; MONTESINOS-LÓPEZ, A.; CROSSA, J.; DE LOS CAMPOS, G.; ALVARADO, G.; SUCHISMITA, M.; RUTKOSKI, J.; GONZÁLEZ-PÉREZ, L.; BURGUEÑO, J. Predicting grain yield using canopy hyperspectral reflectance in wheat breeding data. **Plant methods**, [s. l.], v. 13, n. 1, p. 1–23, 2017.
- MORAES, E. C. De. Fundamentos de sensoriamento remoto. **São José dos campos: INPE**, [s. l.], p. 1–7, 2002.
- MUNNS, R.; JAMES, R. A.; SIRAULT, X. R. R.; FURBANK, R. T.; JONES, H. G. New phenotyping methods for screening wheat and barley for beneficial responses to water deficit. **Journal of experimental botany**, [s. l.], v. 61, n. 13, p. 3499–3507, 2010.
- MWIINGA, B.; SIBIYA, J.; KONDWAKWENDA, A.; MUSVOSVI, C.; CHIGEZA, G. Genotype x environment interaction analysis of soybean (*Glycine max* (L.) Merrill) grain yield across production environments in Southern Africa. **Field Crops Research**, [s. l.], v. 256, p. 107922, 2020.
- NAUE, C. R.; MARQUES, M. W.; LIMA, N. B.; GALVÍNCIO, J. D. Sensoriamento remoto como ferramenta aos estudos de doenças de plantas agrícolas: uma revisão (Remote Sensing as a Toll for the Study of Plant Diseases on Agriculture: a Revision). **Revista Brasileira de Geografia Física**, [s. l.], v. 3, n. 3, p. 190–195, 2011.

OGOKE, I. J.; CARSKY, R. J.; TOGUN, A. O.; DASHIELL, K. E. Maturity class and P effects on soya bean grain yield in the moist savanna of West Africa. **Journal of Agronomy and Crop Science**, [s. l.], v. 189, n. 6, p. 422–427, 2003.

PATIL, G.; MIAN, R.; VUONG, T.; PANTALONE, V.; SONG, Q.; CHEN, P.; SHANNON, G. J.; CARTER, T. C.; NGUYEN, H. T. Molecular mapping and genomics of soybean seed protein: a review and perspective for the future. **Theoretical and Applied Genetics**, [s. l.], v. 130, n. 10, p. 1975–1991, 2017.

RAHIM, H. R. B. A.; LOKMAN, M. Q. Bin; HARUN, S. W.; HORNYAK, G. L.; STERCKX, K.; MOHAMMED, W. S.; DUTTA, J. Applied light-side coupling with optimized spiral-patterned zinc oxide nanorod coatings for multiple optical channel alcohol vapor sensing. **Journal of Nanophotonics**, [s. l.], v. 10, n. 3, p. 036009, 2016.

RAHMAN, S. U.; MCCOY, E.; RAZA, G.; ALI, Z.; MANSOOR, S.; AMIN, I. Improvement of Soybean; A Way Forward Transition from Genetic Engineering to New Plant Breeding Technologies. **Molecular Biotechnology**, [s. l.], v. 65, n. 2, p. 162–180, 2023.

RANI, A.; KUMAR, V. Soybean breeding. Em: **Fundamentals of Field Crop Breeding**. [s.l.] : Springer, 2022. p. 907–944.

RYU, J.-H.; OH, D.; CHO, J. Simple method for extracting the seasonal signals of photochemical reflectance index and normalized difference vegetation index measured using a spectral reflectance sensor. **Journal of Integrative Agriculture**, [s. l.], v. 20, n. 7, p. 1969–1986, 2021.

SANTANA, D. C.; COTRIM, M. F.; FLORES, M. S.; ROJO BAIO, F. H.; SHIRATSUCHI, L. S.; SILVA JUNIOR, C. A. Da; TEODORO, L. P. R.; TEODORO, P. E. UAV-based multispectral sensor to measure variations in corn as a function of nitrogen topdressing. **Remote Sensing Applications: Society and Environment**, [s. l.], v. 23, p. 100534, 2021.

SANTOS, T. T.; YASSITEPE, J. E. de C. T. Fenotipagem de plantas em larga escala: um novo campo de aplicação para a visão computacional na agricultura. **Embrapa Informática Agropecuária-Capítulo em livro científico (ALICE)**, [s. l.], 2014.

SONG, H.; AHMAD, A.; FAROOQ, F.; OSTROWSKI, K. A.; MAŚLAK, M.; CZARNECKI, S.; ASLAM, F. Predicting the compressive strength of concrete with fly ash admixture using machine learning algorithms. **Construction and Building Materials**, [s. l.], v. 308, p. 125021, 2021.

TAIT, L.; BIND, J.; CHARAN-DIXON, H.; HAWES, I.; PIRKER, J.; SCHIEL, D. Unmanned aerial vehicles (UAVs) for monitoring macroalgal biodiversity: Comparison of RGB and multispectral imaging sensors for biodiversity assessments. **Remote Sensing**, [s. l.], v. 11, n. 19, p. 2332, 2019.

VAN EEUWIJK, F. A.; BUSTOS-KORTS, D.; MILLET, E. J.; BOER, M. P.; KRUIJER, W.; THOMPSON, A.; MALOSETTI, M.; IWATA, H.; QUIROZ, R.; KUPPE, C.; MULLER, O.; BLAZAKIS, K. N.; YU, K.; TARDIEU, F.; CHAPMAN, S.

C. Modelling strategies for assessing and increasing the effectiveness of new phenotyping techniques in plant breeding. **Plant Science**, [s. l.], v. 282, p. 23–39, 2019.

VENANCIO, L. P.; MANTOVANI, E. C.; DO AMARAL, C. H.; NEALE, C. M. U.; GONÇALVES, I. Z.; FILGUEIRAS, R.; EUGENIO, F. C. Potential of using spectral vegetation indices for corn green biomass estimation based on their relationship with the photosynthetic vegetation sub-pixel fraction. **Agricultural Water Management**, [s. l.], v. 236, p. 106155, 2020.

WILLIS, S. The use of soybean meal and full fat soybean meal by the animal feed industry. *In*: AUSTRALIAN SOYBEAN CONFERENCE. SOY AUSTRALIA, BUNDABERG, 12., 2003, [S. l.]. **Anais...** [s.l: s.n.], 2003.

XU, Y.; LI, P.; ZOU, C.; LU, Y.; XIE, C.; ZHANG, X.; PRASANNA, B. M.; OLSEN, M. S. Enhancing genetic gain in the era of molecular breeding. **Journal of Experimental Botany**, [s. l.], v. 68, n. 11, p. 2641–2666, 2017.

XUE, J.; SU, B. Significant remote sensing vegetation indices: A review of developments and applications. **Journal of sensors**, [s. l.], v. 2017, 2017.

YANG, G.; LIU, J.; ZHAO, C.; LI, Z.; HUANG, Y.; YU, H.; XU, B.; YANG, X.; ZHU, D.; ZHANG, X. Unmanned aerial vehicle remote sensing for field-based crop phenotyping: current status and perspectives. **Frontiers in plant science**, [s. l.], v. 8, p. 1111, 2017.

YUAN, W.; WIJEWARDANE, N. K.; JENKINS, S.; BAI, G.; GE, Y.; GRAEF, G. L. Early Prediction of Soybean Traits through Color and Texture Features of Canopy RGB Imagery. **Scientific Reports**, [s. l.], v. 9, n. 1, p. 14089, 2019.

ZABOROWICZ, M.; BONIECKI, P.; KOSZELA, K.; PRZYBYŁ, J.; MAZUR, R.; KUJAWA, S.; PILARSKI, K. Use of artificial neural networks in the identification and classification of tomatoes. *In*: FIFTH INTERNATIONAL CONFERENCE ON DIGITAL IMAGE PROCESSING (ICDIP 2013), 5., 2013, [S. l.]. **Anais [...]**. [S. l.]: International Society for Optics and Photonics, 2013.

ZHANG, X.; DUAN, A.; ZHANG, J.; XIANG, C. Estimating Tree Height-Diameter Models with the Bayesian Method. **The Scientific World Journal**, [s. l.], v. 2014, p. 683691, 2014.

ZHAO, B.; DUAN, A.; ATA-UL-KARIM, S. T.; LIU, Z.; CHEN, Z.; GONG, Z.; ZHANG, J.; XIAO, J.; LIU, Z.; QIN, A. Exploring new spectral bands and vegetation indices for estimating nitrogen nutrition index of summer maize. **European Journal of Agronomy**, [s. l.], v. 93, p. 113–125, 2018.

ZHOU, J.; FU, X.; ZHOU, S.; ZHOU, J.; YE, H.; NGUYEN, H. T. Automated segmentation of soybean plants from 3D point cloud using machine learning. **Computers and Electronics in Agriculture**, [s. l.], v. 162, p. 143–153, 2019.

ZHOU, J.; ZHOU, J.; YE, H.; ALI, M. L.; NGUYEN, H. T.; CHEN, P. Classification of soybean leaf wilting due to drought stress using UAV-based imagery. **Computers and Electronics in Agriculture**, [s. l.], v. 175, p. 105576, 2020.

ZHOU, S.; MOU, H.; ZHOU, J.; ZHOU, J.; YE, H.; NGUYEN, H. T. Development of an automated plant phenotyping system for evaluation of salt tolerance in soybean. **Computers and Electronics in Agriculture**, [s. l.], v. 182, p. 106001, 2021.

## **CAPÍTULO 1 – CLASSIFICAÇÃO DE GENÓTIPOS DE SOJA QUANTO A CARACTERES TECNOLÓGICOS INDUSTRIAIS USANDO UAV-MULTIESPECTRAL E APRENDIZAGEM DE MÁQUINA**

### **RESUMO**

Os genótipos de soja possuem características físico-químicas distintas, principalmente na quantidade de óleo e proteína. O uso de tecnologias de fenotipagem de alta precisão aliado ao processamento de dados por meio de algoritmos de aprendizagem de máquina (ML) facilita e pode tornar mais rápida e precisa a obtenção de informações das características dos grãos. Dessa forma, o objetivo do trabalho foi encontrar algoritmos de aprendizagem de máquina e inputs com melhor desempenho para classificação de genótipos de soja com base em características industriais. O experimento foi implementado a campo em blocos ao acaso com duas repetições. Foram avaliadas 103 populações F<sub>2</sub> de soja. Foram obtidos dados de reflectância nos comprimentos de onda red, green, Rededge e NIR e calculado os índices de vegetação NDVI, NDRE, GNDVI, SAVI, MSAVI, MCARI, EVI e SCCC. As características industriais avaliadas foram: teores de proteína, óleo, cinzas e fibras. Os dados foram submetidos a análise de correlação expresso por uma rede de correlações. Um agrupamento de genótipos baseado em características industriais foi realizado usando PCA e algoritmo k-means na separação dos grupos segundo os caracteres avaliados, e então os agrupamentos formados foram usados como variáveis de saída dos modelos ML, enquanto três configurações de entrada foram testadas: apenas bandas espectrais (B), apenas índices de vegetação (VIs) e B+VIs. Os algoritmos de ML testados foram: rede neural artificial (RNA), algoritmos de árvore de decisão J48 (J48), REPTree (DT) e RandomTree (Rt), floresta aleatória (RF) e regressão logística (LR), usada como controle. As métricas estatísticas utilizadas para avaliar a acurácia dos modelos foram, porcentagem de classificação correta (CC) e F-score. Os algoritmos de ML que alcançaram as maiores acurácias de classificação foram ANN, DT e SVM. Quanto aos inputs testados, os melhores resultados foram obtidos utilizando apenas bandas espectrais.

**Palavras-chave:** inteligência computacional; fenotipagem de alto rendimento; agricultura de precisão; bandas espectrais; índices de vegetação

## 1 INTRODUÇÃO

A soja é a principal fonte mundial de óleo vegetal e de farelo que é rico em proteína utilizado em rações animais. Além disso, tem sido difundida na indústria alimentícia para humanos devido aos seus benefícios nutritivos (JIANG *et al.*, 2018). De maneira geral os grãos de soja possuem 40% de proteína, 20% de óleo, 35% carboidratos e 5% de cinzas (KARR-LILIENTHAL *et al.*, 2005; Liu, 2012). A quantidade desses atributos nos grãos determina seu uso, ou seja, cultivares ricas em óleos são indicadas para indústria de óleo vegetal ou para a produção de biocombustíveis. Quando aos grãos, são destinados a alimentação humana, aqueles com teores elevados de proteína e açúcar (JIANG *et al.*, 2018).

Características intrínsecas aos grãos como teor de óleo, proteína e fibras possuem arquitetura genética complexa que dificulta a seleção e a modificação dos compostos dos grãos durante os programas de melhoramento (ZHANG *et al.*, 2018). Estimar parâmetros das culturas agrícolas com alta precisão e de baixo custo é imprescindível para o avanço tecnológico e produtivo, sobretudo para o avanço da fenotipagem de alto rendimento (MAIMAITIJIANG *et al.*, 2017).

A fenotipagem de alto rendimento (HTP) possibilita mensurar uma grande quantidade de fenótipos ao longo do tempo e espaço, com custos menores e maior agilidade de trabalho. A HTP aumenta de forma significativa a precisão e auxilia na intensidade de seleção de genótipos em programas de melhoramento genético de plantas (RUTKOSKI *et al.*, 2016). Diversas plataformas semi ou autônomas, equipadas com vários sensores, possibilitam a coleta de dados espaciais e temporais, que permite a obtenção de diversas informações sobre a cultura ao longo do ciclo (SINGH *et al.*, 2016). Algoritmos de inteligência computacional aceleram e automatizam análises de imagens melhorando o processamento de informações de dados advindos de sensores que são informações complexas (MOCHIDA *et al.*, 2019).

A utilização de inteligência computacional como a aprendizagem de máquinas (ML) permite obter diversas respostas sobre as plantas utilizando variáveis espectrais como seleção de genótipos por meio de características ligada a produtividade ou o ciclo da soja (SANTANA *et al.*, 2022), previsão do teor de nitrogênio foliar e altura de plantas (Osco *et al.*, 2020), e associação entre produtividade e fatores relacionados (SANTANA *et al.*, 2021b). Todas essas informações melhoram o processo de tomada de decisão na seleção de genótipos, contribuindo para o avanço do melhoramento de plantas de forma mais tecnológica e confiável (GAO *et al.*, 2022). Dentre as funcionalidades dos

algoritmos de aprendizagem de máquinas está a classificação. No entanto, cada algoritmo tem seu próprio desempenho sobre um conjunto de dados, exigindo o teste de diferentes algoritmos para encontrar o melhor classificador para uma determinada situação (KARAKATIČ & PODGORELEC, 2016).

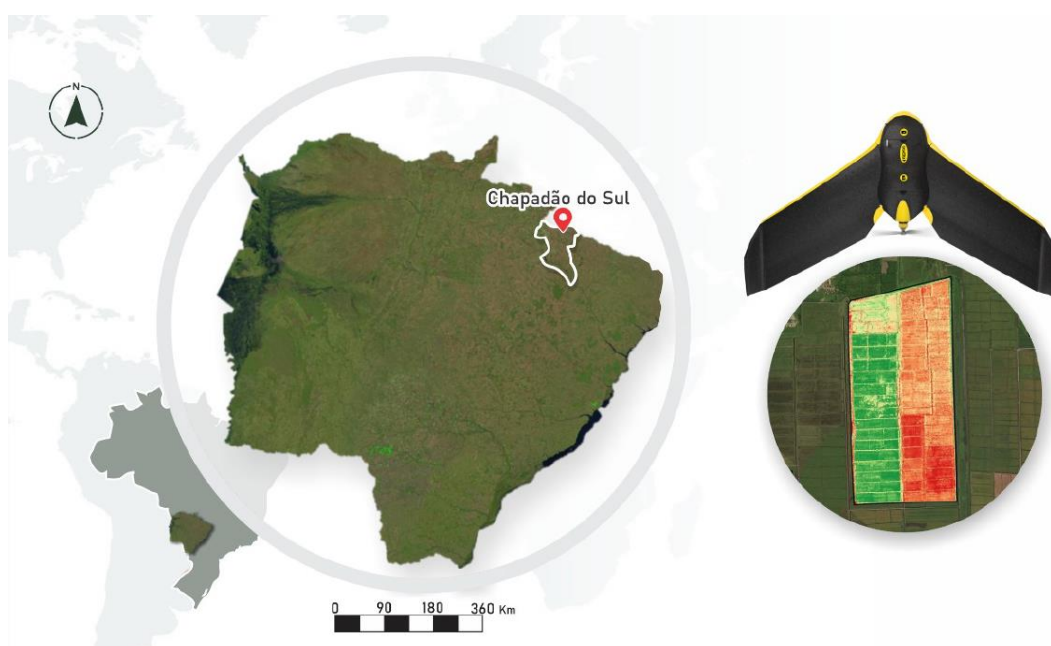
Na literatura são escassas as informações sobre fatores intrínsecos aos grãos como óleo, proteína, fibras, cinzas, etc., principalmente ao que diz respeito a dados oriundos de fenotipagem de alta precisão e utilizando algoritmos de inteligência computacional. Estudos voltados à mensuração e diferenciação de genótipos com base em seus atributos industriais por meio de HTP e processamento de dados por ML são de suma importância para o melhoramento das culturas, pois podem contribuir para a seleção de genótipos de forma mais confiável e rápida, com a plantas ainda no campo. Além de proporcionar maior eficiência dos programas, essa abordagem também pode contribuir para todo o processo produtivo da cultura, onde produtores e empresas podem ter informações sobre a qualidade nutricional e rendimento de óleo das sementes antes da colheita. Assim, a hipótese de pesquisa é que é possível classificar genótipos de soja quanto ao seu desempenho industrial utilizando informações espectrais e processamento de dados por ML. O objetivo foi encontrar algoritmos de ML e inputs com maior acurácia para classificar genótipos de soja com base em características industriais.

## 2 MATERIAL E MÉTODOS

### 2.1 Condução do experimento

O experimento foi conduzido no município de Chapadão do Sul – MS (Figura 1), safra 2019/20, na área experimental da Universidade Federal de Mato Grosso do Sul (18°41'33''S, 52°40'45''W, com 810 m de altitude). O clima da região, segundo a classificação de Koppen, é Tropical de Savana (Aw).

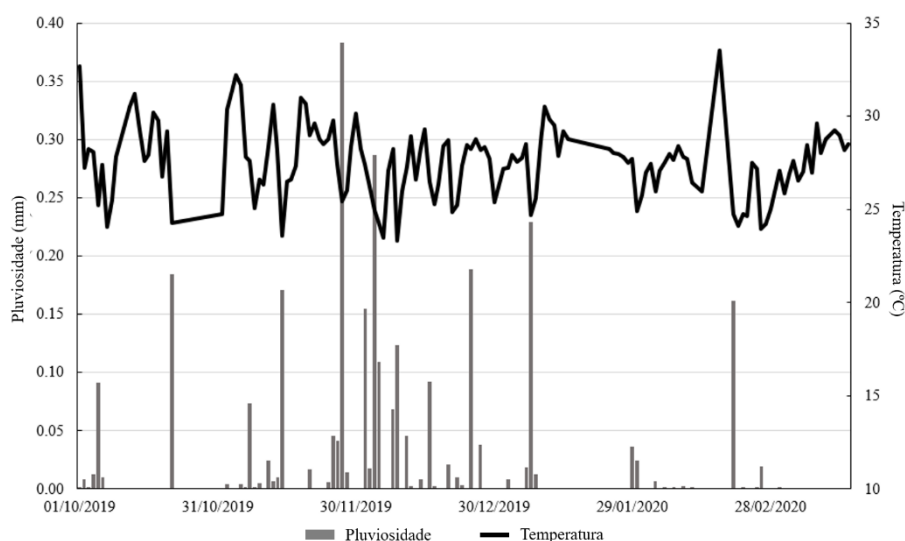
Figura 1 – Localização da área experimental na Universidade Federal de Mato Grosso do Sul, Chapadão do Sul -MS, figura ilustrativa do ARP utilizado e a reflectância das parcelas avaliadas



Fonte: elaboração própria.

A Figura 2 demonstra as condições meteorológicas durante a condução do experimento. O solo da região é classificado como Latossolo Vermelho Distrófico argiloso (Teixeira *et al.*, 2017), tendo características na camada de 0 - 0,20 m: pH (H<sub>2</sub>O) = 6,2; Al trocável (cmol<sub>c</sub> dm<sup>-3</sup>) = 0,0; Ca+Mg (cmol<sub>c</sub> dm<sup>-3</sup>) = 4,31; P (mg dm<sup>-3</sup>) = 41,3; K (cmol<sub>c</sub> dm<sup>-3</sup>) = 0,2; Matéria orgânica (g dm<sup>-3</sup>) = 19,74; V (%) = 45; m (%) = 0,0; Soma de bases (cmol<sub>c</sub> dm<sup>-3</sup>) = 2,3; CTC (cmol<sub>c</sub> dm<sup>-3</sup>) = 5,1.

Figura 2 – Condições meteorológicas no decorrer da safra 2019/2020



Fonte: elaboração própria.

A semeadura foi realizada em outubro de 2019, adotando o sistema de preparo do solo convencional (aração e gradagem). O experimento foi implementado a campo em blocos ao acaso com duas repetições. Foram avaliados 103 genótipos F<sub>2</sub> de soja, as quais foram avaliadas por meio da linha central de três metros de comprimento de cada parcela. O espaçamento entre fileiras foi de 0.45 m e densidade de semeadura de 15 sementes por metro.

Antes da semeadura as sementes foram tratadas com fungicida (Piraclostrobina + Metil Tiofanato) e inseticida (Fipronil), na dose de 200 mL do produto comercial para cada 100 kg de sementes para proteção contra o ataque de pragas e fungos de solo. as sementes foram inoculadas com bactérias do gênero *Bradyrhizobium* Para a fixação biológica de nitrogênio (FBN), sendo utilizada a dose de 200 mL de inoculante líquido concentrado para cada 100 kg de sementes. Os tratos culturais foram realizados de acordo com a necessidade da cultura.

## 2.2 Variáveis industriais

espectroscopia de infravermelho próximo (NIRS) (Modelo DS2500, FOSS Analytical, Dinamarca) com alta precisão óptica foi usado para determinar a proteína bruta (Proteína), rendimento de óleo (Óleo), fibra bruta (Fibra) e teor de cinzas nas

populações de soja F2. As amostras foram homogeneizadas e colocadas em um prato de amostragem. A análise baseou-se em iluminar uma amostra com radiação de comprimento de onda específico na região do infravermelho próximo e, em seguida, medir a diferença entre a quantidade de energia emitida pelo espectroscópio e refletida pela amostra para o detector. O registro dos dados espectrais foi realizado no modo de refletância, dentro da faixa espectral de 400-2500 nm. O resultado obtido foi comparado a um conjunto de resultados já conhecido determinado como padrão, conforme proposto por Horwitz *et al.*, 1970.

### 2.3 Aquisição e processamento das imagens multiespectrais

Aos 60 dias após a emergência (DAE) foram coletadas imagens espectrais com auxílio de uma aeronave remotamente pilotada (RPA) de asa fixa Sensefly eBee RTK, com controle autônomo de decolagem, plano de voo e aterrissagem. O eBee foi equipado com o sensor multiespectral Parrot Sequoia. A aquisição das imagens foi realizada às 09:00 da manhã (horário local), com céu limpo de nuvens a uma altitude de 100 m (altitude local) e resolução espacial de 0,10 m. A calibração radiométrica foi realizada para toda a cena com base em uma superfície refletiva calibrada fornecida pelo fabricante e pelo uso do sensor de luminosidade embutido no próprio equipamento. Os valores de refletância foram obtidos nas seguintes bandas espectrais (SBs): azul (475 nm), vermelho (660 nm), verde (550 nm), Rededge (735 nm) e NIR (790 nm).

As informações adquiridas pelos comprimentos de ondas possibilitaram os cálculos dos índices de vegetação realizadas pelo software ArqGis (Tabela 1). O levantamento aéreo foi realizado com o uso da tecnologia RTK (*Real Time Kinematics*), possibilitando estimar a posição da câmera no instante de coleta da imagem com acurácia de 2,5 cm. As imagens foram mosaicadas e ortorretificadas pelo programa computacional Pix4Dmapper. A acurácia posicional das ortoimagens foi verificada com pontos de controle em campo (GCP – *Ground Control Points*), levantados com o RTK.

Tabela 1 – Relação dos índices de vegetação

| Sigla | Índice de Vegetação                                           | Equação                                                                                | Referência                       |
|-------|---------------------------------------------------------------|----------------------------------------------------------------------------------------|----------------------------------|
| NDVI  | Índice de vegetação pela diferença normalizada                | $\frac{(R_{NIR} - R_{RED})}{(R_{NIR} + R_{RED})}$                                      | (ROUSE <i>et al.</i> , 1974)     |
| NDRE  | Índice de vegetação da diferença normalizada pelo Red-Edge    | $\frac{(R_{NIR} - R_{EDGE})}{(R_{NIR} + R_{EDGE})}$                                    | (GITELS ON <i>et al.</i> , 1996) |
| GNDVI | Índice de vegetação pela diferença normalizada ao verde       | $\frac{(R_{NIR} - R_{GREEN})}{(R_{NIR} + R_{GREEN})}$                                  | (GITELS ON <i>et al.</i> , 1996) |
| SAVI  | Índice de vegetação ajustado pela presença de solo            | $(1 + 0,5) \frac{nir - red}{nir + red + 0,5}$                                          | (HUETE, 1988)                    |
| MSAVI | Índice de vegetação modificado ajustado pela presença de solo | $\frac{2nir + 1 - \sqrt{(2nir + 1)^2 - (8nir - red)}}{2}$                              | (QI <i>et al.</i> , 1994)        |
| MCARI | Índice de refletância de absorção de clorofila modificado     | $R_{700} - R_{670} - 0.2(R_{700} - R_{550}) \frac{R_{700}}{R_{670}}$                   | (DAUGHTRY <i>et al.</i> , 2000)  |
| EVI   | Índice de vegetação aumentada                                 | $2.5 * \frac{(R_{NIR} - R_{RED})}{((R_{NIR}) + (C1 * R_{NIR}) - (C2 * R_{BLUE}) + L)}$ | (HUETE <i>et al.</i> , 1997)     |
| SCCCI | Índice simplificado de conteúdo de clorofila do dossel        | $\frac{NDVI}{NDRE}$                                                                    | (RAPER & VARCO, 2015)            |

R<sub>NIR</sub>: reflectância na faixa do infravermelho próximo; R<sub>GREEN</sub>: reflectância na faixa do verde; R<sub>RED</sub>: reflectância na faixa do vermelho; R<sub>EDGE</sub>: reflectância na faixa de transição do vermelho (Red-edge); L: fator de correção do efeito do solo.

Fonte: elaboração própria.

## 2.4 Modelos de Aprendizagem de Máquina e Análises Estatísticas

Os dados foram submetidos a análise de correlação expressa por uma rede de correlações gerada com o auxílio do software Rbio (Bhering, 2017). Para gerar os grupos de genótipos das populações, os dados foram submetidos a análise de componentes

principais (PCA). Um biplot foi construído com os dois primeiros componentes principais devido à fácil interpretação desses resultados. Neste biplot, dois clusters (C1, C2) foram definidos com base no desempenho dos genótipos para as variáveis tecnológicas industriais para posterior uso do algoritmo k-means, que agrupa tratamentos cujos centroides estão mais próximos até que não haja variação significativa na distância mínima de cada observação para cada um dos centróides. Essas análises foram realizadas com o auxílio do pacote "ggfortify" do software R (TEAM, 2013).

Posteriormente os dados foram submetidos as análises de aprendizagem de máquina (Tabela 2), em que os clusters formados foram utilizados como variáveis de saída (output) dos modelos, enquanto as variáveis espectrais foram utilizadas como entrada dos modelos. Ainda, em relação às variáveis de entrada, foram testadas três diferentes configurações de conjunto de dados: usando apenas IVs, usando apenas SBs e usando IVs+SBs como variáveis de entrada. A classificação dos clusters foi realizada por meio de validação cruzada estratificada com k-fold = 10 e dez repetições (100 execuções para cada modelo). Todos os parâmetros dos modelos foram estabelecidos de acordo com a configuração default do software Weka 3.8.5.

Tabela 2 – Relação dos modelos de aprendizagem de máquinas utilizados na classificação

| Sigla | Modelo de aprendizagem de máquinas | Referência                             |
|-------|------------------------------------|----------------------------------------|
| ANN   | Redes neurais artificiais          | (EGMONT-PETERSEN <i>et al.</i> , 2002) |
| J48   | Árvores de decisão J48             | (QUINLAN, 1993)                        |
| LR    | Regressão logística                | (ŠTEPANOVSKEÝ <i>et al.</i> , 2017)    |
| DT    | REPTree                            | (SNOUSY <i>et al.</i> , 2011)          |
| RF    | Floresta aleatória                 | (BELGIU & DRĂGUȚ, 2016)                |
| Rt    | Random Tree                        | (KALMEGH, 2015)                        |
| SVM   | Máquina de vetor suporte           | (NALEPA & KAWULOK, 2019)               |

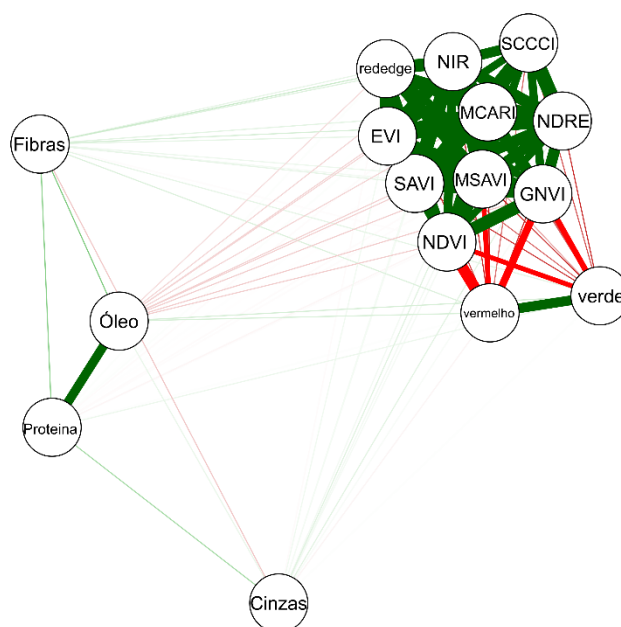
Fonte: elaboração própria.

Nas avaliações quanto ao desempenho dos modelos de classificação utilizados foram usadas as métricas de porcentagem de classificações corretas (CC) e F-score. Para comparar o desempenho dos algoritmos de ML, foram gerados boxplots com as médias CC e F-score agrupados pelo teste de Scott-Knott (Scott; Knott, 1974) ao nível de 5% de significância. O agrupamento das médias foi realizado como o auxílio do software Rbio, enquanto os boxplots foram gerados usando os pacotes ggplot2 e ExpDes.pt do software R.

### 3 RESULTADOS

A partir da análise de correlação linear de Pearson, foi gerada uma rede de correlação entre as variáveis (Figura 3). As linhas em destaque apresentam correlações de magnitude significativa acima de 60%. Linhas em verde representam correlações positivas e as em vermelho indicam correlações negativas.

Figura 3 – Rede de correlações entre as variáveis industriais (Proteína, óleo, Fibras, Cinzas) bandas espectrais (red, green, Rededge, NIR) e índices de vegetação (NDVI, SAVI, GNVDI, NDRE, SCCCI, EVI, MSAVI e MCARI)



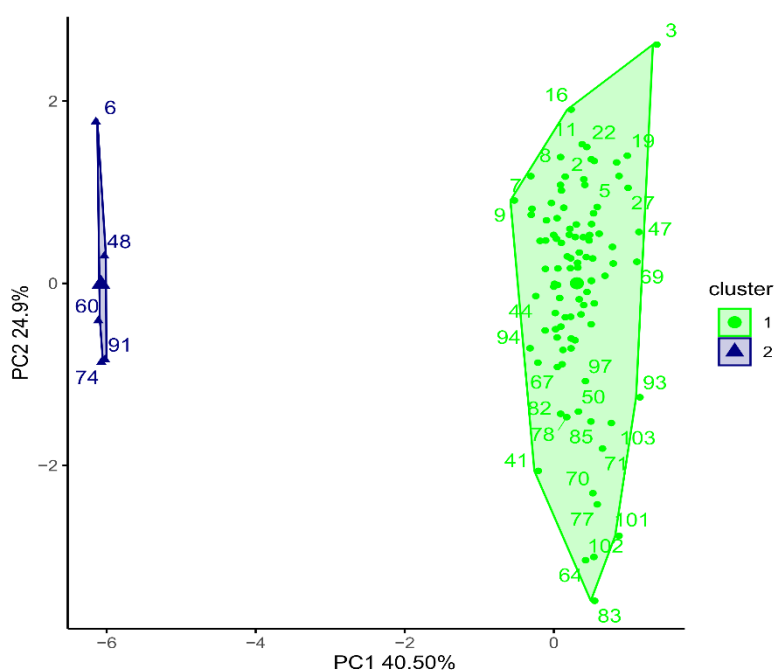
Fonte: elaboração própria.

Proteína e óleo apresentaram correlação de alta magnitude e positiva. As variáveis espectrais demonstraram correlações de alta magnitude entre si e em sua maioria positiva, resultado esperado devido ao uso das bandas red, green e NIR nos cálculos dos índices de vegetação. Entre as variáveis tecnológicas industriais e as espectrais, as correlações foram de baixa magnitude, isso porque não há relação linear entre essas variáveis. Assim, faz-se necessário o uso de outras técnicas que superem a falta de linearidade entre elas.

Os genótipos de soja foram divididos em grupos de acordo com a similaridade entre eles por agrupamento PCA e K-means (Figura 4). Observa-se que os genótipos

foram divididos em dois grupos distintos, nos quais os genótipos contidos em cada grupo compartilham semelhanças entre si quanto às variáveis analisadas. O somatório dos dois primeiros componentes principais explicou 65.4% da variação total dos dados, sendo ligeiramente inferior ao indicado por Regazzi e Cruz, (2020) que recomendam que esse somatório seja acima de 70%. No entanto, o valor foi suficiente e adequado para separar os genótipos para prosseguir com as análises posteriores.

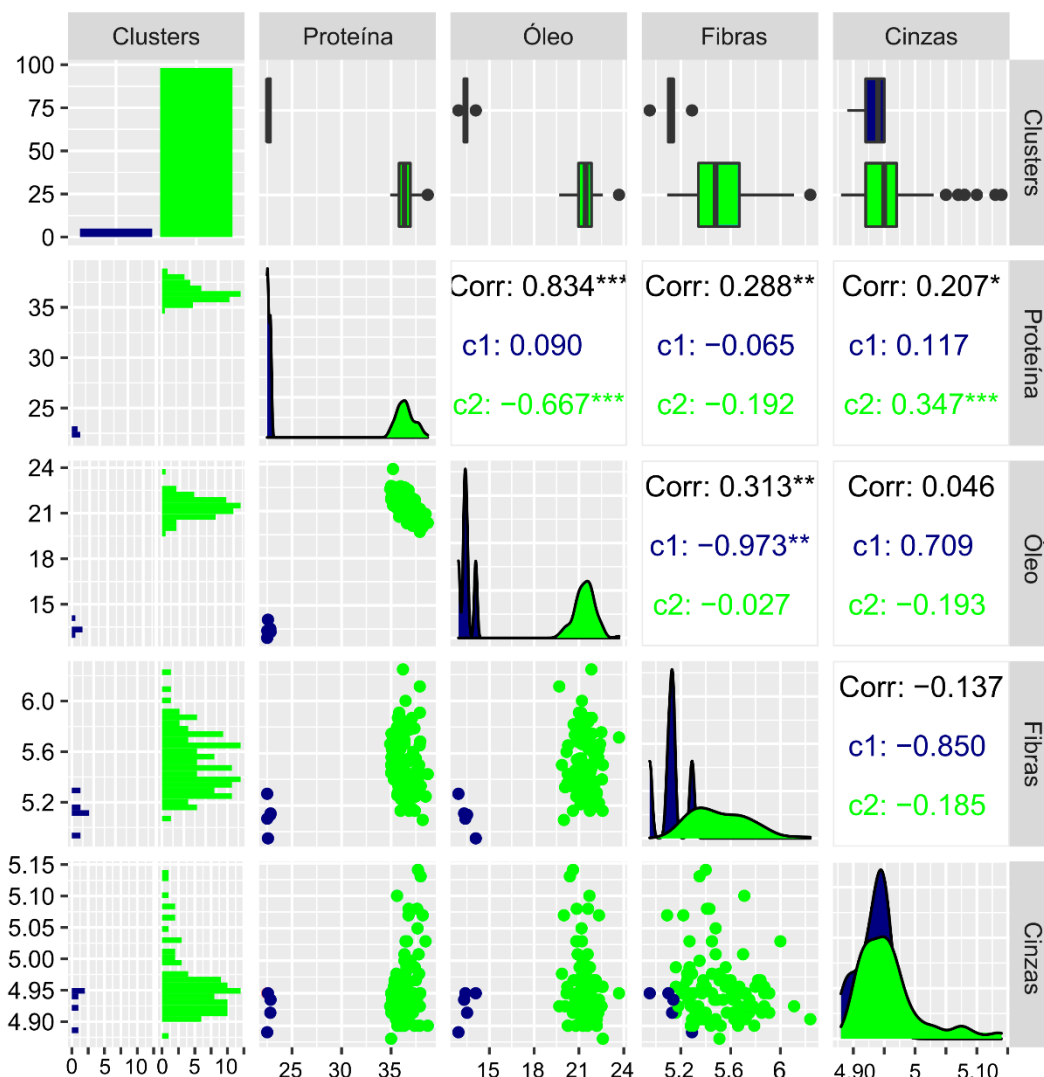
Figura 4 – Análise de Componentes Principais (PC) para os clusters formados pelo algoritmo k-means



Fonte: elaboração própria.

A partir do agrupamento de genótipos, foi realizada uma análise de correlação entre as variáveis industriais dentro de cada agrupamento (Figura 5). Observa-se que a correlação entre óleo e proteína foi alta e negativa (-0,667) entre os genótipos pertencentes ao cluster 2. Entre os genótipos do cluster 1, houve uma alta correlação negativa entre fibras e óleo (-0,973). Correlação alta e positiva entre cinzas e óleo (0,709) e alta e negativa entre cinzas e fibras (0,850) também podem ser observadas.

Figura 5 – Gráfico de correlações e dispersão entre os clusters e as variáveis proteína, óleo, fibras e cinzas



Correlações seguidas por \*\*\*, \*\* e \*, apresentam respectivamente significância de 0,1, 1 e 5% de probabilidade pelo teste F.

Fonte: elaboração própria.

Uma vez que os grupos de genótipos foram divididos pelas análises mencionadas, as análises de classificação foram prosseguidas usando diferentes algoritmos de ML. Essa abordagem busca reconhecer padrões por meio de informações apresentadas que reduzam a subjetividade das variáveis e otimizem o processo de análise, contornando as limitações dos métodos tradicionais (de Medeiros *et al.*, 2020).

Foram utilizadas duas métricas de acurácia foram utilizadas nas análises de classificação: classificação correta (CC) e F-score. Essas métricas foram submetidas a uma análise de variância, onde as entradas foram consideradas como o primeiro fator e

os algoritmos testados como o segundo fator (Tabela 3). Para classificação correta, houve significância para os algoritmos de ML utilizados e sua interação com as entradas, e o F-score foi significativo apenas para os algoritmos testados.

Tabela 3 – Resumo da análise de variância para as variáveis Classificação correta (CC) e F-score

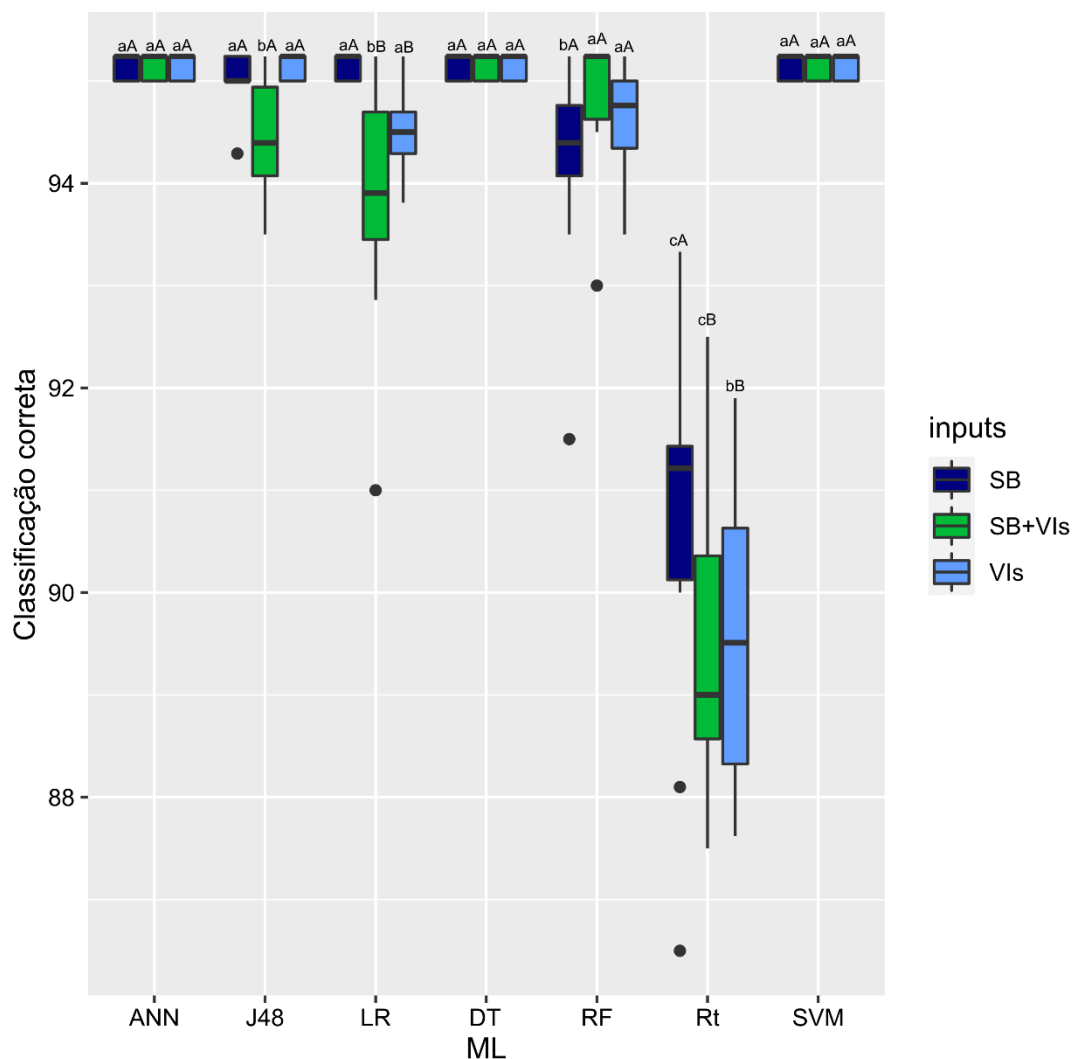
| FV          | GL | CC      | F-score |
|-------------|----|---------|---------|
| Inputs      | 2  | 1,91    | 0,00    |
| ML          | 6  | 109,63* | 0,00**  |
| Inputs x ML | 12 | 1,56**  | 0,00    |

\*significativo a 5% e \*\*significativo a 1% pelo teste F; ML: inteligência computacional.

Fonte: elaboração própria.

A Figura 6 representa a classificação correta em função dos algoritmos de ML utilizando três entradas diferentes: bandas espectrais (SB), índices de vegetação (IVs), bandas espectrais + índices de vegetação (B+IVs). Ao avaliar a entrada utilizando apenas as bandas espectrais, os algoritmos de ML que se destacaram foram: ANN, J48, LR, DT e SVM, não diferenciando estatisticamente entre si. Todos os algoritmos tiveram bom desempenho ao usar os IVs como entrada, exceto Rt. As diferentes entradas testadas não diferiram estatisticamente para os algoritmos ANN, J48, DT, RF e SVM. Para os algoritmos LR e RT, o uso de bandas espectrais como variáveis de entrada obteve melhores resultados do que B+VIs ou apenas IVs.

Figura 6 – Boxplot para porcentagem de classificação correta (CC) para a interação significativa entre os modelos de aprendizagem de máquinas e os inputs testados

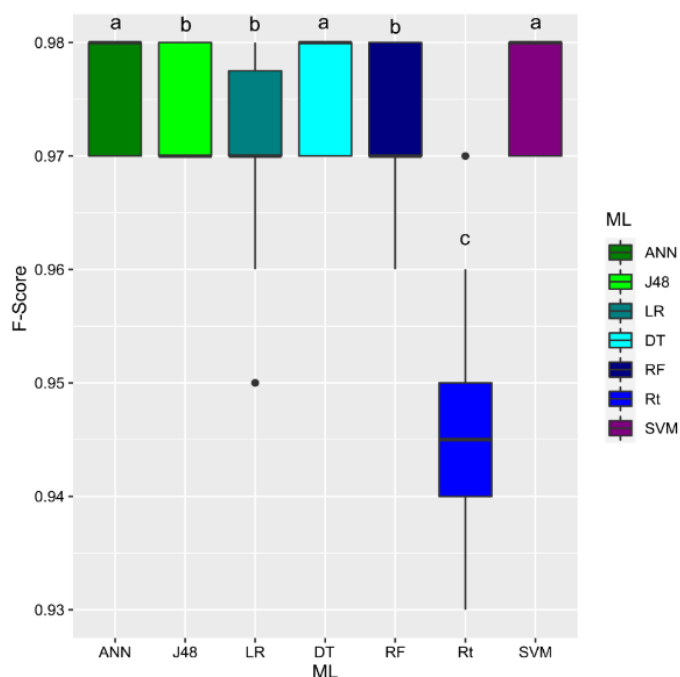


As médias seguidas pelas mesmas letras maiúsculas para as diferentes entradas e as mesmas letras minúsculas para os diferentes algoritmos de ML não diferem pelo teste de Scott-Knott a 5% de probabilidade.

Fonte: elaboração própria.

Na Figura 7, nota-se que para F-score as melhores médias foram observadas para os algoritmos ANN, DT e SVM, enquanto o algoritmo com pior desempenho foi o Rt. O F-score é uma média harmônica que se obtém a partir da precisão e da renovação do banco de dados (Witten *et al.*, 2005).

Figura 7 – Boxplot para F-score para a significância entre os modelos de aprendizagem de máquinas



As médias seguidas pelas mesmas letras não diferem entre si pelo teste de Scott-Knott a 5% de probabilidade.

Fonte: elaboração própria.

ANN, DT e SVM foram os algoritmos que alcançaram bom desempenho em todas as medidas de acurácia testadas. Quanto aos inputs testados a utilização de bandas espectrais alcançaram bons resultados que aliado aos algoritmos ANN, DT e SVM possuem acurácia para a classificação das características industriais da soja.

#### 4 DISCUSSÃO

Proteína e óleo apresentaram correlação de alta magnitude e positiva. Os teores dessas propriedades dos grãos da soja são variáveis e são influenciadas por fatores genéticos e por condições ambientais, principalmente no período de enchimento dos grãos (PÍPOLO & MANDARINO, 2016). As correlações óleo x proteína e rendimento de grãos x proteína são muitas vezes negativas (GESTEIRA *et al.*, 2018; KAMBHAMPATI *et al.*, 2020), tornando a seleção de genótipos de alta proteína e alto rendimento um grande desafio em programas de melhoramento de soja (RODRIGUES *et al.*, 2013). A aquisição de novos genótipos de soja com alto teor de óleo ou proteína e rendimento adequado é um processo demorado devido ao ciclo da cultura, reconhecimento das características do grão de cada genótipo e do trabalho necessário para realizar as avaliações (FINOTO *et al.*, 2017), tornando o processo de melhoria dessas características demorado e caro. A fenotipagem usando refletância do dossel da planta pode ser uma excelente ferramenta para auxiliar na seleção de genótipos (SANTANA *et al.*, 2022), e prever e classificar as características de interesse de forma mais rápida e não destrutiva (YOUSEFZADEH-NAJAFABADI *ET AL.*, 2021).

Para contornar o problema de não linearidade entre variáveis, como foi apresentado entre as variáveis industriais e as espectrais, pode ser utilizado algoritmos de aprendizagem de máquina. Tais algoritmos são eficientes em contornar problemas da não linearidade entre as variáveis lidando de forma mais robusta com esses dados. A integração de técnicas de inteligência computacional com sensoriamento remoto e dados de campo possibilita obter informações mais confiáveis sobre diversos alvos de estudo, como rendimento (SCHWALBERT *et al.*, 2020).

Jiang *et al.* (2018) chegaram ao resultado de altas e negativas correlações fenotípicas e genotípicas entre fibra e óleo em soja, resultado observado entre os genótipos do cluster 1 (Figura 5), sugerindo que alterações no teor de óleo pelo processo de melhoramento de soja causaram um aumento também no teor de fibras. No geral, o teor de fibra na soja tem sido pouco investigado. As correlações observadas entre cinzas e óleo (0,709) e entre cinzas e fibras (-0,850) revelam informações nutricionais importantes sobre os grãos que auxiliam na seleção de cultivares e programas de melhoramento direto para diferentes finalidades (AZAM *et al.*, 2021).

Houve correlação alta e negativa entre óleo e proteína, observado entre os genótipos do cluster 2 (Figura 5). Embora seja possível aumentar simultaneamente o teor de óleo e proteína da soja até certo nível, a correlação negativa entre eles dificulta os

programas de melhoramento (LIN *et al.*, 2015). Isto é especialmente verdadeiro para o teor de óleo dos grãos, que é uma característica quantitativa influenciada tanto por fatores genéticos quanto pelas condições ambientais em que o genótipo está inserido (TURQUETTI-MORAES *et al.*, 2022). Outro fator crítico no melhoramento da soja é que em busca de aumentos produtivos ocorre a diminuição da proteína. Outras variáveis intrínsecas aos grãos não receberam atenção ao longo dos anos de melhoramento da cultura (TAMAGNO *et al.*, 2022) como teor de fibras e cinzas.

Algoritmos de aprendizagem de máquinas são ferramentas importantes no processamento de dados dentro da fenotipagem de alta precisão que geram banco de dados complexos, e a escolha assertiva do melhor algoritmo é crucial para esse fim (A. Singh *et al.*, 2016). Neste estudo, ANN, J48, DT, SVM superaram os outros algoritmos para as três configurações de entrada testadas (Figura 6) com base na porcentagem de classificações corretas.

As ANN são algoritmos eficientes na classificação, chegando a acurácia superior a 80%, a qual a precisão aumenta com a diminuição da complexidade dos dados (MAKSIMENKO *et al.*, 2018). Akbar *et al.*, (2018) obtiveram bons resultados utilizando ANN na predição no rendimento de óleo de cúrcuma, alcançando acurácia na previsão e otimização no processamento dos dados. ANN também pode ser uma ferramenta utilizada na predição do rendimento de biodiesel com alta precisão e reduzindo de maneira efetiva o tempo e o preço do processamento dessas informações (DEVARAJ NAIK & UDAYAKUMAR, 2021).

A utilização do algoritmo J48 pode melhorar a precisão da classificação (ALJAWARNEH *et al.*, 2019), pois elimina a subjetividade e reduz erros nas respostas além de contribuir para a classificação automatizada de imagens (Rossi *et al.*, 2020). Esse algoritmo pode superar as ANN sobretudo quando os dados possuem muito ruídos e são complexos, não seguindo linearidade (BENZER *et al.*, 2022). Massawe *et al.* (2018), concluíram que J48 é mais sensível quanto a quantidade de conjunto de dados do que o RF, alcançando melhores resultados.

O uso de árvores de decisão ou algoritmo similares, auxiliam no processamento de informações obtidas com ARP, proporcionando respostas eficientes sobre o monitoramento constante das culturas (MARIN *et al.*, 2021). Esse algoritmo é conciso na detecção de doenças em plantas de arroz com precisão de 97% (AHMED *et al.*, 2019) e que quando usado em classificação possui um bom desempenho e pode ser mais rápido

no processamento dependendo da comparação com outros ML (PANDEY & PRABHAKAR, 2016).

Quanto aos inputs avaliados a utilização das bandas espectrais apresentaram bons resultados em todos algoritmos de aprendizagem de máquinas (Figura 6), sendo mais prático para o processamento das informações utilizar as bandas espectrais já que não necessitam de cálculos matemáticos como os índices de vegetação. Segundo (ALVES *et al.*, 2019) sensores multiespectrais de bandas largas podem potencializar tarefas de classificação e reduzir custos desses processamentos. As bandas podem ser sensíveis na classificação de genótipos de soja (SILVA JUNIOR *et al.*, 2020), quantificar lesões em folhas causadas por insetos (ALVES *et al.*, 2019), além de ser, não destrutível e de confiança para distinguir variáveis na soja, como óleo de soja transgênica e não transgênica (LUNA *et al.*, 2013).

Os resultados das Figuras 6 e 7, mostra que a acurácia de CC e de F-score atingiram valores acima de 94% para os algoritmos que obtiveram os melhores desempenhos. O algoritmo Rt obteve os piores resultados para todas as métricas. Para F-score os algoritmos ANN, DT e SVM obtiveram os melhores desempenhos. A eficiência e utilização das ANN já foram apresentadas anteriormente, destacando que esse algoritmo de aprendizagem de máquinas possui bons ajustes e adequada capacidade de generalização de dados (EUGENIO *et al.*, 2020). Singh *et al.* (2021), alcançaram alta precisão geral na classificação de sementes de soja doentes usando ANN e DT. Esses algoritmos proporcionaram o melhor desempenho na classificação de genótipos de soja para atributos industriais.

O algoritmo SVM é um método de aprendizagem de máquinas eficiente na classificação de imagens em diversas situações como em casos de incidência de pragas (TETILA *et al.*, 2020) e doença (GUI *et al.*, 2021) na soja. Seu uso também pode melhorar previsões de produtividade de soja (CORRALES *et al.*, 2022). SVM também se mostrou uma boa técnica de aprendizagem de máquinas utilizando imagens RGB na medição de teor de clorofila na soja (HASSANIJALILIAN *et al.*, 2020). SVM também é um método preciso para prever rendimento das culturas com base em IVs (JU *et al.*, 2021).

Utilizar dados espectrais para classificar genótipos de soja com base em características espectrais dos grãos pode otimizar as avaliações de distintos materiais genéticos de soja em programas de melhoramento. Aliar algoritmos de aprendizagem de máquinas a avaliações espectrais torna todo o processo mais rápido e mais preciso. Futuramente, essa tecnologia poderá ser levada a agricultores e empresas para fornecer

informações sobre os atributos industriais dos grãos, principalmente rendimento de óleo e teor de proteína, com a cultura ainda no campo. Dessa forma, empresas e agricultores poderiam definir os preços de produção de forma mais justa. Para estudos futuros, outros genótipos de soja, índices de vegetação e algoritmos de aprendizagem de máquina devem ser avaliados visando maior precisão de classificação.

## 5 CONCLUSÃO

Os algoritmos de ML com maior precisão de classificação foram ANN, algoritmo de árvore de decisão REPTree e Support Vector Machine. Além disso, foram testadas três configurações de entrada para verificar quais informações espectrais (usando apenas bandas, usando apenas IVs ou bandas + IVs) fornecem a maior precisão. Os melhores resultados foram obtidos usando apenas bandas espectrais.

## REFERÊNCIAS

- AHMED, K.; SHAHIDI, T. R.; ALAM, S. M. I.; MOMEN, S. Rice Leaf Disease Detection Using Machine Learning Techniques. *In: INTERNATIONAL CONFERENCE ON SUSTAINABLE TECHNOLOGIES FOR INDUSTRY 4.0 (STI) 2019*, [s. l]. **Anais [...]**. [s.l: s.n.], 2019.
- AKBAR, A.; KUANAR, A.; PATNAIK, J.; MISHRA, A.; NAYAK, S. Application of Artificial Neural Network modeling for optimization and prediction of essential oil yield in turmeric (*Curcuma longa* L.). **Computers and Electronics in Agriculture**, [s. l.], v. 148, p. 160–178, 2018.
- ALJAWARNEH, S.; YASSEIN, M. B.; ALJUNDI, M. An enhanced J48 classification algorithm for the anomaly intrusion detection systems. **Cluster Computing**, [s. l.], v. 22, n. 5, p. 10549–10565, 2019.
- ALVES, T. M.; MOON, R. D.; MACRAE, I. V; KOCH, R. L. Optimizing band selection for spectral detection of *Aphis glycines* Matsumura in soybean. **Pest Management Science**, [s. l.], v. 75, n. 4, p. 942–949, 2019.
- AZAM, M.; ZHANG, S.; QI, J.; ABDELGHANY, A. M.; SHAIBU, A. S.; GHOSH, S.; FENG, Y.; HUAI, Y.; GEBREGZIABHER, B. S.; LI, J.; LI, B.; SUN, J. Profiling and associations of seed nutritional characteristics in Chinese and USA soybean cultivars. **Journal of Food Composition and Analysis**, [s. l.], v. 98, p. 103803, 2021.
- BELGIU, M.; DRĂGUT, L. Random forest in remote sensing: A review of applications and future directions. **ISPRS Journal of Photogrammetry and Remote Sensing**, [s. l.], v. 114, p. 24–31, 2016. Disponível em:
- BENZER, S.; GARABAGHI, F. H.; BENZER, R.; MEHR, H. D. Investigation of some machine learning algorithms in fish age classification. **Fisheries Research**, [s. l.], v. 245, p. 106151, 2022.
- BHERING, L. L. Rbio: A tool for biometric and statistical analysis using the R platform. **Crop Breeding and Applied Biotechnology**, [s. l.], v. 17, p. 187–190, 2017.
- CORRALES, D. C.; SCHOVING, C.; RAYNAL, H.; DEBAEKE, P.; JOURNET, E.-P.; CONSTANTIN, J. A surrogate model based on feature selection techniques and regression learners to improve soybean yield prediction in southern France. **Computers and Electronics in Agriculture**, [s. l.], v. 192, p. 106578, 2022.
- DAUGHTRY, C. S. T.; WALTHALL, C. L.; KIM, M. S.; DE COLSTOUN, E. B.; MCMURTREY III, J. E. Estimating corn leaf chlorophyll concentration from leaf and canopy reflectance. **Remote sensing of Environment**, [s. l.], v. 74, n. 2, p. 229–239, 2000.
- DE MEDEIROS, A. D.; CAPOBIANGO, N. P.; DA SILVA, J. M.; DA SILVA, L. J.; DA SILVA, C. B.; DOS SANTOS DIAS, D. C. F. Interactive machine learning for soybean seed and seedling quality classification. **Scientific Reports**, [s. l.], v. 10, n. 1, p. 11267, 2020.

- DEVARAJ NAIK, B.; UDAYAKUMAR, M. Comparative assessment of response surface methodology and artificial neural networks in forecasting biodiesel yield from waste cooking sunflower oil. **Materials Today: Proceedings**, [s. l.], v. 46, p. 9908–9913, 2021.
- EGMONT-PETERSEN, M.; DE RIDDER, D.; HANDELS, H. Image processing with neural networks—a review. **Pattern Recognition**, [s. l.], v. 35, n. 10, p. 2279–2301, 2002.
- EUGENIO, F. C.; GROHS, M.; VENANCIO, L. P.; SCHUH, M.; BOTTEGA, E. L.; RUOSO, R.; SCHONS, C.; MALLMANN, C. L.; BADIN, T. L.; FERNANDES, P. Estimation of soybean yield from machine learning techniques and multispectral RPAS imagery. **Remote Sensing Applications: Society and Environment**, [s. l.], v. 20, p. 100397, 2020.
- FINOTO, E. L.; SEDIYAMA, T.; ALVES DE ALBUQUERQUE, J. de A.; BERNADES SOARES, M. B.; ALTAFIN GALLI, J.; CORDEIRO JUNIOR, P. S.; SANTOS DE MENEZES, P. H. Antecipação e retardamento de colheita nos teores de óleo e proteína das sementes de soja, cultivar Valiosa RR. **Scientia Agropecuaria**, [s. l.], v. 8, n. 2, p. 99–107, 2017.
- GAO, S.; GUAN, H.; MA, X. A recognition method of multispectral images of soybean canopies based on neural network. **Ecological Informatics**, [s. l.], v. 68, p. 101538, 2022.
- GESTEIRA, G. de S.; BRUZI, A. T.; ZITO, R. K.; FRONZA, V.; ARANTES, N. E. Selection of early soybean inbred lines using multiple indices. **Crop Science**, [s. l.], v. 58, n. 6, p. 2494–2502, 2018.
- GITELSON, A. A.; KAUFMAN, Y. J.; MERZLYAK, M. N. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. **Remote sensing of Environment**, [s. l.], v. 58, n. 3, p. 289–298, 1996.
- GUI, J.; FEI, J.; WU, Z.; FU, X.; DIAKITE, A. Grading method of soybean mosaic disease based on hyperspectral imaging technology. **Information Processing in Agriculture**, [s. l.], v. 8, n. 3, p. 380–385, 2021.
- HASSANIJALILIAN, O.; IGATHINATHANE, C.; DOETKOTT, C.; BAJWA, S.; NOWATZKI, J.; HAJI ESMAEILI, S. A. Chlorophyll estimation in soybean leaves infield with smartphone digital imaging and machine learning. **Computers and Electronics in Agriculture**, [s. l.], v. 174, p. 105433, 2020.
- HORWITZ, W.; CHICHILO, P.; REYNOLDS, H. Official methods of analysis of the Association of Official Analytical Chemists. **Official methods of analysis of the Association of Official Analytical Chemists**, [s. l.], 1970.
- HUETE, A. R. A soil-adjusted vegetation index (SAVI). **Remote sensing of environment**, [s. l.], v. 25, n. 3, p. 295–309, 1988.

HUETE, A. R.; LIU, H. Q.; BATCHILY, K. V.; VAN LEEUWEN, W. A comparison of vegetation indices over a global set of TM images for EOS-MODIS. **Remote sensing of environment**, [s. l.], v. 59, n. 3, p. 440–451, 1997.

JIANG, G. L.; CHEN, P.; ZHANG, J.; FLOREZ-PALACIOS, L.; ZENG, A.; WANG, X.; BOWEN, R. A.; MILLER, A.; BERRY, H. Genetic analysis of sugar composition and its relationship with protein, oil, and fiber in soybean. **Crop Science**, [s. l.], v. 58, n. 6, p. 2413–2421, 2018.

JU, S.; LIM, H.; MA, J. W.; KIM, S.; LEE, K.; ZHAO, S.; HEO, J. Optimal county-level crop yield prediction using MODIS-based variables and weather data: A comparative study on machine learning models. **Agricultural and Forest Meteorology**, [s. l.], v. 307, p. 108530, 2021.

KALMEGH, S. Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News. **IJISSET-International Journal of Innovative Science, Engineering & Technology**, [s.l.], v. 2, n. 2, 2015.

KAMBHAMPATI, S.; AZNAR-MORENO, J. A.; HOSTETLER, C.; CASO, T.; BAILEY, S. R.; HUBBARD, A. H.; DURRETT, T. P.; ALLEN, D. K. On the inverse correlation of protein and oil: Examining the effects of altered central carbon metabolism on seed composition using soybean fast neutron mutants. **Metabolites**, [s. l.], v. 10, n. 1, 2020.

KARAKATIČ, S.; PODGORELEC, V. Improved classification with allocation method and multiple classifiers. **Information Fusion**, [s. l.], v. 31, p. 26–42, 2016.

KARR-LILIENTHAL, L. K.; KADZERE, C. T.; GRIESHOP, C. M.; FAHEY, G. C. Chemical and nutritional properties of soybean carbohydrates as related to nonruminants: A review. **Livestock Production Science**, [s. l.], v. 97, n. 1, p. 1–12, 2005.

LIN, R.; GLAZEBROOK, J.; KATAGIRI, F.; ORF, J. H.; GIBSON, S. I. Identification of differentially expressed genes between developing seeds of different soybean cultivars. **Genomics Data**, [s. l.], v. 6, p. 92–98, 2015.

LIU, K. **Soybeans: chemistry, technology, and utilization**. [s.l.] : Springer, 2012.  
LUNA, A. S.; DA SILVA, A. P.; PINHO, J. S. A.; FERRÉ, J.; BOQUÉ, R. Rapid characterization of transgenic and non-transgenic soybean oils by chemometric methods using NIR spectroscopy. **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**, [s. l.], v. 100, p. 115–119, 2013.

MAIMAITIJANG, M.; GHULAM, A.; SIDIKE, P.; HARTLING, S.;  
MAIMAITIYIMING, M.; PETERSON, K.; SHAVERS, E.; FISHMAN, J.;  
PETERSON, J.; KADAM, S.; BURKEN, J.; FRITSCHI, F. Unmanned Aerial System (UAS)-based phenotyping of soybean using multi-sensor data fusion and extreme learning machine. **ISPRS Journal of Photogrammetry and Remote Sensing**, [s. l.], v. 134, p. 43–58, 2017.

- MAKSIMENKO, V. A.; KURKIN, S. A.; PITSIK, E. N.; MUSATOV, V. Yu.; RUNNOVA, A. E.; EFREMOVA, T. Yu.; HRAMOV, A. E.; PISARCHIK, A. N. Artificial Neural Network Classification of Motor-Related EEG: An Increase in Classification Accuracy by Reducing Signal Complexity. **Complexity**, [s. l.], v. 2018, p. 9385947, 2018.
- MARIN, D. B.; FERRAZ, G. A. e S.; SANTANA, L. S.; BARBOSA, B. D. S.; BARATA, R. A. P.; OSCO, L. P.; RAMOS, A. P. M.; GUIMARÃES, P. H. S. Detecting coffee leaf rust with UAV-based vegetation indices and decision tree machine learning models. **Computers and Electronics in Agriculture**, [s. l.], v. 190, p. 106476, 2021.
- MASSAWE, B. H. J.; SUBBURAYALU, S. K.; KAAYA, A. K.; WINOWIECKI, L.; SLATER, B. K. Mapping numerically classified soil taxa in Kilombero Valley, Tanzania using machine learning. **Geoderma**, [s. l.], v. 311, p. 143–148, 2018.
- MOCHIDA, K.; KODA, S.; INOUE, K.; HIRAYAMA, T.; TANAKA, S.; NISHII, R.; MELGANI, F. Computer vision-based phenotyping for improvement of plant productivity: a machine learning perspective. **GigaScience**, [s. l.], v. 8, n. 1, p. giy153, 2019.
- NALEPA, J.; KAWULOK, M. Selecting training sets for support vector machines: a review. **Artificial Intelligence Review**, [s. l.], v. 52, n. 2, p. 857–900, 2019.
- OSCO, L. P.; JUNIOR, J. M.; RAMOS, A. P.; FURUYA, D. E.; SANTANA, D. C.; TEODORO, L. P.; GONÇALVES, W. N.; BAILO, F. H.; PISTORI, H.; JUNIOR, C. A.; TEODORO, P. E. **Leaf Nitrogen Concentration and Plant Height Prediction for Maize Using UAV-Based Multispectral Imagery and Machine Learning Techniques**, 2020.
- PANDEY, P.; PRABHAKAR, R. An analysis of machine learning techniques (J48 & AdaBoost)-for classification. In: INDIA INTERNATIONAL CONFERENCE ON INFORMATION PROCESSING (IICIP), 1., 2016, [s. l.] **Anais [...]**. [s.l: s.n.], 2016.
- PÍPOLO, A. E.; MANDARINO, J. M. G. Os teores de proteína da soja e a qualidade para a indústria. **Embrapa Soja-Artigo em periódico indexado (ALICE)**, [s. l.], 2016.
- QI, J.; CHEHBOUNI, A.; HUETE, A. R.; KERR, Y. H.; SOROOSHIAN, S. A modified soil adjusted vegetation index. **Remote sensing of environment**, [s. l.], v. 48, n. 2, p. 119–126, 1994.
- QUINLAN, J. R. C4. 5: Programming for machine learning. **Morgan Kauffmann**, [s. l.], v. 38, n. 48, p. 49, 1993.
- RAPER, T. B.; VARCO, J. J. Canopy-scale wavelength and vegetative index sensitivities to cotton growth parameters and nitrogen status. **Precision Agriculture**, [s. l.], v. 16, n. 1, p. 62–76, 2015.

- REGAZZI, A. J.; CRUZ, C. D. Análise multivariada aplicada. **Editora UFV, Viçosa, 401p**, [s. l.], 2020.
- RODRIGUES, J. I. da S.; ARRUDA, K. M. A.; CRUZ, C. D.; PIOVESAN, N. D.; BARROS, E. G. De; MOREIRA, M. A. Associação de marcadores microssatélites com teores de óleo e proteína em soja. **Pesquisa Agropecuária Brasileira**, [s. l.], v. 48, p. 255–262, 2013.
- ROSSI, F. S.; SILVA JUNIOR, C. A. Da; OLIVEIRA-JÚNIOR, J. F. De; TEODORO, P. E.; SHIRATSUCHI, L. S.; LIMA, M.; TEODORO, L. P. R.; TIAGO, A. V.; CAPRISTO-SILVA, G. F. Identification of tillage for soybean crop by spectro-temporal variables, GEOBIA, and decision tree. **Remote Sensing Applications: Society and Environment**, [s. l.], v. 19, p. 100356, 2020.
- ROUSE, J. W.; HAAS, R. H.; SCHELL, J. A.; DEERING, D. W. Monitoring vegetation systems in the Great Plains with ERTS. **NASA special publication**, [s. l.], v. 351, n. 1974, p. 309, 1974.
- RUTKOSKI, J.; POLAND, J.; MONDAL, S.; AUTRIQUE, E.; PÉREZ, L. G.; CROSSA, J.; REYNOLDS, M.; SINGH, R. Canopy Temperature and Vegetation Indices from High-Throughput Phenotyping Improve Accuracy of Pedigree and Genomic Selection for Grain Yield in Wheat. **G3 Genes|Genomes|Genetics**, [s. l.], v. 6, n. 9, p. 2799–2808, 2016.
- SANTANA, D. C.; COTRIM, M. F.; FLORES, M. S.; ROJO BAILO, F. H.; SHIRATSUCHI, L. S.; SILVA JUNIOR, C. A. Da; TEODORO, L. P. R.; TEODORO, P. E. UAV-based multispectral sensor to measure variations in corn as a function of nitrogen topdressing. **Remote Sensing Applications: Society and Environment**, [s. l.], v. 23, p. 100534, 2021.
- SANTANA, D. C.; DE OLIVEIRA CUNHA, M. P.; DOS SANTOS, R. G.; COTRIM, M. F.; TEODORO, L. P. R.; DA SILVA JUNIOR, C. A.; BAILO, F. H. R.; TEODORO, P. E. High-throughput phenotyping allows the selection of soybean genotypes for earliness and high grain yield. **Plant Methods**, [s. l.], v. 18, n. 1, p. 13, 2022.
- SCHWALBERT, R. A.; AMADO, T.; CORASSA, G.; POTT, L. P.; PRASAD, P. V. V.; CIAMPITTI, I. A. Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. **Agricultural and Forest Meteorology**, [s. l.], v. 284, p. 107886, 2020.
- SCOTT, A. A. J.; KNOTT, M. A Cluster Analysis Method for Grouping Means in the Analysis of Variance Published by : International Biometric Society Stable URL : <http://www.jstor.org/stable/2529204>. **Biometrics**, [s. l.], v. 30, n. 3, p. 507–512, 1974.
- SILVA JUNIOR, C. A. Da; TEODORO, L. P. R.; TEODORO, P. E.; BAILO, F. H. R.; PANTALEÃO, A. de A.; CAPRISTO-SILVA, G. F.; FACCO, C. U.; OLIVEIRA-JÚNIOR, J. F. De; SHIRATSUCHI, L. S.; SKRIPACHEV, V.; LIMA, M.; NANNI, M. R. Simulating multispectral MSI bandsets (Sentinel-2) from hyperspectral observations via spectroradiometer for identifying soybean cultivars. **Remote Sensing Applications: Society and Environment**, [s. l.], v. 19, p. 100328, 2020.

SINGH, A.; GANAPATHYSUBRAMANIAN, B.; SINGH, A. K.; SARKAR, S. Machine Learning for High-Throughput Stress Phenotyping in Plants. **Trends in Plant Science**, [s. l.], v. 21, n. 2, p. 110–124, 2016.

SNOUSY, M. B. Al; EL-DEEB, H. M.; BADRAN, K.; KHLIL, I. A. Al. Suite of decision tree-based classification algorithms on cancer gene expression data. **Egyptian Informatics Journal**, [s. l.], v. 12, n. 2, p. 73–82, 2011.

ŠTEPANOVSKÝ, M.; IBROVÁ, A.; BUK, Z.; VELEMÍNSKÁ, J. Novel age estimation model based on development of permanent teeth compared with classical approach and other modern data mining methods. **Forensic Science International**, [s. l.], v. 279, p. 72–82, 2017.

TAMAGNO, S.; SADRAS, V. O.; AZNAR-MORENO, J. A.; DURRETT, T. P.; CIAMPITTI, I. A. Selection for yield shifted the proportion of oil and protein in favor of low-energy seed fractions in soybean. **Field Crops Research**, [s. l.], v. 279, p. 108446, 2022.

TEAM, R. C. R: **A language and environment for statistical computing**. [s. l.], 2013.

TEIXEIRA, P. C.; DONAGEMMA, G. K.; FONTANA, A.; TEIXEIRA, W. G. **Manual de métodos de análise de solo**. Brasília: Embrapa, [s. l.], p. 573, 2017.

TETILA, E. C.; MACHADO, B. B.; ASTOLFI, G.; BELETE, N. A. de S.; AMORIM, W. P.; ROEL, A. R.; PISTORI, H. Detection and classification of soybean pests using deep learning with UAV images. **Computers and Electronics in Agriculture**, [s. l.], v. 179, p. 105836, 2020.

TURQUETTI-MORAES, D. K.; MOHARANA, K. C.; ALMEIDA-SILVA, F.; PEDROSA-SILVA, F.; VENANCIO, T. M. Integrating omics approaches to discover and prioritize candidate genes involved in oil biosynthesis in soybean. **Gene**, [s. l.], v. 808, p. 145976, 2022.

WITTEN, I. H.; FRANK, E.; HALL, M. A.; PAL, C. J.; DATA, M. Practical machine learning tools and techniques. Em: DATA MINING 2005, **Anais...** [s.l: s.n.]

YOOSEFZADEH-NAJAFABADI, M.; EARL, H. J.; TULPAN, D.; SULIK, J.; ESKANDARI, M. Application of Machine Learning Algorithms in Plant Breeding: Predicting Yield From Hyperspectral Reflectance in Soybean. **Frontiers in Plant Science**, [s. l.], v. 11, 2021.

ZHANG, J.; WANG, X.; LU, Y.; BHUSAL, S. J.; SONG, Q.; CREGAN, P. B.; YEN, Y.; BROWN, M.; JIANG, G.-L. Genome-wide Scan for Seed Composition Provides Insights into Soybean Quality Improvement and the Impacts of Domestication and Breeding. **Molecular Plant**, [s. l.], v. 11, n. 3, p. 460–472, 2018. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1674205217303866>.

## **CAPÍTULO 2 – FENOTIPAGEM DE ALTO RENDIMENTO COM ESPECTROSCOPIA VIS/NIR NA CLASSIFICAÇÃO DE GENÓTIPOS DE SOJA QUANTO A PRODUTIVIDADE DE GRÃO E CARACTERES INDUSTRIAIS**

### **RESUMO**

A utilização de sensores VIS/NIR na fenotipagem de alta precisão, pode permitir a relação das características espectrais da folha com o conteúdo do grão, auxiliando o melhoramento genético de soja a direcionar o programa para aprimorar as características do grão de acordo com o interesse dos pesquisadores. A hipótese da pesquisa é que a reflectância foliar de genótipos de soja pode estar diretamente relacionado com caracteres industriais do grão como proteína e fibras. Sendo assim os objetivos do trabalho são: I) classificar genótipos de soja de acordo com as características industriais do grão; II) identificar o(s) algoritmo(s) com melhor acurácia para a classificação dos genótipos com a reflectância das folhas; III) identificação do melhor dado de entrada (input) dos algoritmos que melhoram sua performance. O experimento foi alocado em blocos casualizados com três repetições e 32 genótipos de soja. Aos 60 dias após a emergência (DAE) foi realizada a análise espectral de três amostras foliares de cada parcela. Foi utilizado o sensor hiperespectral que permitiu capturar a reflectância entre os comprimentos de onda entre 450 e 824 nm. Foram selecionadas faixas de intervalos representativos em bandas espectrais agrupados em médias. Após a colheita foi avaliada a produtividade de grãos e realizadas análises laboratoriais dos caracteres industriais. Os dados espectrais, industriais do grão e produtividade foram submetidos a análises estatísticas. Os dados foram analisados usando aprendizagem de máquina, utilizando seis algoritmos: Árvores de decisão J48 (J48) e REPTree (DT), Floresta aleatória (RF), Redes neurais artificiais (RNA), Regressão logística (SL) e Máquina de vetor suporte (SVM). Os clusters formados foram usados como saída (output) dos modelos, enquanto para a entrada dos modelos foram utilizados dois grupos de dados de entrada (input): as variáveis espectrais (WL) obtidas pelo sensor sem o ruído (450-828nm) e as médias espectrais das bandas selecionadas (SB) (450.0-720.6 nm). Os genótipos de soja foram separados em grupos de acordo com suas características industriais e a produtividade dos grãos em que os algoritmos SVM e J48 foram os que apresentaram melhor performance na classificação deles. A utilização das bandas espectrais selecionadas no trabalho proporcionaram melhor acurácia para os algoritmos na classificação.

**Palavras-chave:** sensor hiperespectral; bandas espectrais; proteína bruta; árvore de decisão; máquina de vetor suporte.

## 1 INTRODUÇÃO

A soja [*Glycine max* (L.) Merrill] atualmente no mundo é um dos principais insumos da indústria alimentícia com destaque para a produção e óleo de farelo, destinados a alimentação humana e animal (SANTANA *et al.*, 2023a; ZUFFO *et al.*, 2020). As características intrínsecas ao grão foram melhoradas ao longo dos anos com uma via dupla de aumentar o fornecimento de óleo e proteína ou melhorar uma característica sem impactar negativamente a outra (ZHANG *et al.*, 2020). Segundo Zhang *et al.* (2020), as cultivares de soja lançadas comercialmente nos Estados Unidos na última década possuem teores de proteína reduzido, atribuído ao foco de aumentar o óleo dos grãos e de forma não intencional reduzir a proteína.

No auxílio para direcionar os programas de melhoramento genético de soja para aumentar características do grão de interesse para a indústria pode-se utilizar a fenotipagem de alto rendimento (HTP). A HTP faz uso de tecnologias de sensores para caracterização de fenótipos mensuráveis ao longo do tempo e do espaço de forma não destrutiva, acurada, rápida, a custos menores e requerido pouca mão de obra (ANDRADE *et al.*, 2021; DOBBELS; LORENZ, 2019; SANTANA *et al.*, 2023).

A HTP pode ser utilizada para caracterizar genótipos e populações quanto a morfologia, fisiologia, identificar problemas com estresse abiótico e biótico, relacionar com a composição dos grãos (SINGH *et al.*, 2021) de acordo com a reflectância da planta. A utilização de sensores hiperespectrais para essa caracterização pode fornecer informações relevantes sobre a cultura como na identificação de cultivares de soja de acordo com sua assinatura espectral (SILVA JUNIOR *et al.*, 2018), identificar problemas com estresse hídricos em milho (ASAARI *et al.*, 2019) e tomate (GENANGELI *et al.*, 2023), e salino em quiabo (FENG *et al.*, 2020), dentre outras aplicações.

Dessa forma a utilização da HTP, especialmente na região do VIS/NIR, pode permitir a relação das características espectrais da folha com o conteúdo do grão, auxiliando o melhoramento genético de soja a direcionar o programa para aprimorar as características do grão de acordo com o interesse dos pesquisadores, como é o caso da melhoria de caracteres industriais (SANTANA *et al.*, 2023) direcionando para a indústria de óleo ou de ração. Na implementação da HTP uma série de etapas devem ser consideradas como o tipo de sensor a ser utilizado, a plataforma, a planta a ser estudada,

a problemática a ser resolvida, e por fim as análises de dados que mais irão representar e dar respostas sobre o que se está estudando (KIM, 2020).

Análises precisas para extrair as informações desejadas dos dados espectrais é um gargalo significativo nos programas de melhoramento (TREVISAN *et al.*, 2020) que usam ou venham a usar técnicas de HTP. Isso porque a obtenção de dados espectrais das plantas produz quantidades elevadas de dados em que técnicas tradicionais de análises estatísticas pode impossibilitar ou ser ineficiente em relacionar as características espectrais (GILL *et al.*, 2022) com as características morfológicas, fisiológicas, nutricionais da planta e dos grãos.

A utilização de aprendizagem de máquina é uma alternativa eficiente para a extração de informações úteis dos dados adquiridos por HTP por meio de algoritmos que permitem a extração de características, identificação, classificação e previsão de informações que auxiliam no melhoramento de plantas (GILL *et al.*, 2022). Utilização de dados hiperespectrais processados por meio de técnicas de aprendizagem de máquinas retornam respostas precisas auxiliando na tomada de decisão na seleção de genótipos para características de interesse, e dessa forma contribuir de maneira mais assertiva avanços no melhoramento de plantas de forma mais tecnológica e confiável (GAO *et al.*, 2022).

A hipótese da pesquisa é que a reflectâncias foliar de genótipos de soja pode estar diretamente relacionado com caracteres industriais do grão como proteína e fibras. Sendo assim os objetivos do trabalho são: I) classificar genótipos de soja de acordo com as características industriais do grão; II) identificar o(s) algoritmo(s) com melhor acurácia para a classificação dos genótipos com a reflectância das folhas; III) identificação do melhor dado de entrada (input) dos algoritmos que melhoram sua performance.

## 2 MATERIAL E MÉTODOS

Na Universidade Federal de Mato Grosso do Sul, campus Chapadão do Sul (18°41'33''S, 52°40'45''W, com 810 m de altitude) foi conduzido o experimento com 32 genótipos de soja na safra de 2021/22. Segundo a classificação de Köppen e Geiger O clima da região é Tropical de Savana (Aw). A classificação do solo é Latossolo Vermelho Distrófico Argiloso, com as seguintes características na camada de 0 – 0,20 m: pH (H<sub>2</sub>O) = 6,2; Al trocável (cmol<sub>c</sub> dm<sup>-3</sup>) = 0,0; Ca+Mg (cmol<sub>c</sub> dm<sup>-3</sup>) = 4,31; P (mg dm<sup>-3</sup>) = 41,3; K (cmol<sub>c</sub> dm<sup>-3</sup>) = 0,2; Matéria orgânica (g dm<sup>-3</sup>) = 19,74; V (%) = 45; m (%) = 0,0; Soma de bases (cmol<sub>c</sub> dm<sup>-3</sup>) = 2,3; CTC (cmol<sub>c</sub> dm<sup>-3</sup>) = 5,1.

Os genótipos de soja foram semeados em outubro de 2021, alocados em um experimento em blocos casualizados com três repetições e 32 genótipos de soja. As parcelas foram constituídas por cinco linhas de um metro de comprimento, espaçadas a 0.45 m e densidade de 15 sementes por metro.

O preparo do solo antes da semeadura foi de forma convencional do solo com aração e gradagem niveladora. Antes da semeadura, as sementes de soja receberam tratamento com fungicida (Piraclostrobina + Metil Tiofanato) e inseticida (Fipronil), na dose de 200 mL do produto a cada 100 kg de sementes. Também foi realizado inoculação com *Bradyrhizobium* para a fixação biológica de nitrogênio, seguindo-se a dose de 200 mL de inoculante líquido concentrado a cada 100 kg de sementes. Tratos culturais com fungicidas, inseticidas e herbicidas foram realizados conforme necessidade da cultura.

Aos 60 dias após a emergência (DAE) foi realizada a análise espectral de três amostras foliares escolhidas ao acaso da parte superior do dossel da planta de cada parcela. Foi utilizado o sensor hiperespectral Ocean Optics modelo STS-VIS-L-50-400-SMA, que permitiu capturar a reflectância entre os comprimentos de onda entre 450 e 824 nm. O sensor foi calibrado antes da leitura das amostras de cada bloco, utilizando-se uma placa de sulfato de bário para a calibração do branco e o isolamento do sensor para a calibração no escuro. A captura das reflectâncias foi realizada em laboratório utilizando-se iluminação artificial por duas lâmpadas halógenas de 120 W apontadas para o ponto de leitura.

O sensor estava conectado a um computador para os registros de cada leitura por meio de um software próprio do sensor chamado OceanView, que registra as leituras feitas pelo equipamento com resolução espectral de 1.5 nm. Os comprimentos de onda foram agrupados em médias de intervalos representativos em bandas conforme sugerido

por (SILVA JUNIOR *et al.*, 2018), porém adaptados de acordo com os comprimentos de onda disponibilizado pelo sensor sem o ruído (Tabela 1).

Tabela 1 – Intervalos de Faixa espectral (nm) e comprimento de onda médio (nm) usados no estabelecimento de 16 bandas espectrais

| Bandas (nº) | Faixa espectral (nm) | Comprimento de onda médio (nm) |
|-------------|----------------------|--------------------------------|
| B1          | 450-475              | 462.50                         |
| B2          | 480                  | 480.00                         |
| B3          | 481-500              | 490.47                         |
| B4          | 501-530              | 515.34                         |
| B5          | 531-539              | 534.89                         |
| B6          | 540                  | 540.00                         |
| B7          | 541-649              | 594.80                         |
| B8          | 650                  | 650.00                         |
| B9          | 661-670              | 665.54                         |
| B10         | 675                  | 675.00                         |
| B11         | 676-684              | 680.05                         |
| B12         | 685-689              | 686.83                         |
| B13         | 690-700              | 695.07                         |
| B14         | 701-709              | 705.03                         |
| B15         | 710                  | 710.00                         |
| B16         | 711-730              | 720.60                         |

Fonte: SILVA JUNIOR *et al.*, 2018.

Após o fim do ciclo da cultura, foi realizado a colheita das parcelas experimentais e feita a produtividade de grãos (PROD, kg ha<sup>-1</sup>) foi avaliada pela colheita das fileiras centrais de cada parcela e correção para 13% de umidade e extrapolada para kg ha<sup>-1</sup>. Uma amostra de 50 g foi separada de cada parcela para a realização das análises industriais do grão: a fim de determinar os teores de extrato etéreo (EE) e proteína bruta (PB) conforme metodologia descrita por Silva e Queiroz (2002), fibra em detergente neutro (FDN), fibra em detergente ácido (FDA) e matéria mineral (MM), segundo a metodologia de Van Soest *et al.* (1991)

Os dados espectrais, industriais do grão e produtividade foram submetidos a análises estatísticas. A primeira análise feita foi a separação dos genótipos em grupos por meio do algoritmo k-means. O algoritmo k-means depende do valor de k que sempre precisa ser especificado antes da execução de qualquer análise de agrupamento (M. Ahmed *et al.*, 2020), no trabalho em questão o k que melhor separou os grupos de acordo com as características industriais do grão e a produtividade foi utilizando a separação em

dois grupos. Os genótipos em questão então foram divididos em dois grupos denominados Cluster 1 e Cluster 2.

O k-means realiza comparações entre cada valor das amostras das variáveis utilizadas na análise, calculando o quão “longe” uma está da outra. O algoritmo calcula as distâncias de cada amostra dos centroides dentro de cada classe. Cada amostra fica dentro de seu respectivo grupo de acordo com a distância dos centroides, sendo considerados similares as amostras pertencentes ao mesmo grupo e distinta do outro grupo. Os clusters separados pelo algoritmo k-means foram plotados em um gráfico biplot com o auxílio da análise de componentes principais (PCA) para expressar a separação dos clusters.

Após a separação dos 32 genótipos em dois clusters foram construídos boxplot comparativos entre as médias de cada característica industrial, produtividade e as bandas, buscando observar diferença entre os clusters formados. Também foi realizada análise de correlação de Pearson entre as variáveis e as bandas selecionadas expressas por um gráfico heatmap para cada cluster.

Posteriormente os dados foram analisados usando aprendizagem de máquina, utilizando seis algoritmos (Tabela 2) em que a regressão logística (SL) é utilizado como comparativo por ser um algoritmo tradicional de regressão, em que se busca a superioridade dos outros algoritmos quando comparado a ele. Os clusters formados foram usados como saída (*output*) dos modelos, enquanto para a entrada dos modelos foram utilizados dois grupos de dados de entrada (*input*): as variáveis espectrais (WL) obtidas pelo sensor sem o ruído (450-828nm) e as médias espectrais das bandas selecionadas (SB) (450-720.6 nm) com o intuito de que os algoritmos conseguissem distinguir os clusters de acordo com os dados espectrais dos genótipos pertencentes a cada um deles

Tabela 2 – Relação dos modelos de aprendizagem de máquinas utilizados na classificação

| Sigla | Modelo de aprendizagem de máquinas | Referência                                  |
|-------|------------------------------------|---------------------------------------------|
| DT    | REPTree                            | (SNOUSY <i>et al.</i> , 2011)               |
| J48   | Árvores de decisão J48             | (QUINLAN, 1993)                             |
| RF    | Floresta aleatória                 | (BELGIU; DRĂGUȚ, 2016)                      |
| RNA   | Redes neurais artificiais          | (EGMONT-PETERSEN; DE RIDDER; HANDELS, 2002) |
| SL    | Regressão logística                | (ŠTEPANOVSKÝ <i>et al.</i> , 2017)          |
| SVM   | Máquina de vetor suporte           | (NALEPA; KAWULOK, 2019)                     |

Fonte: elaboração própria.

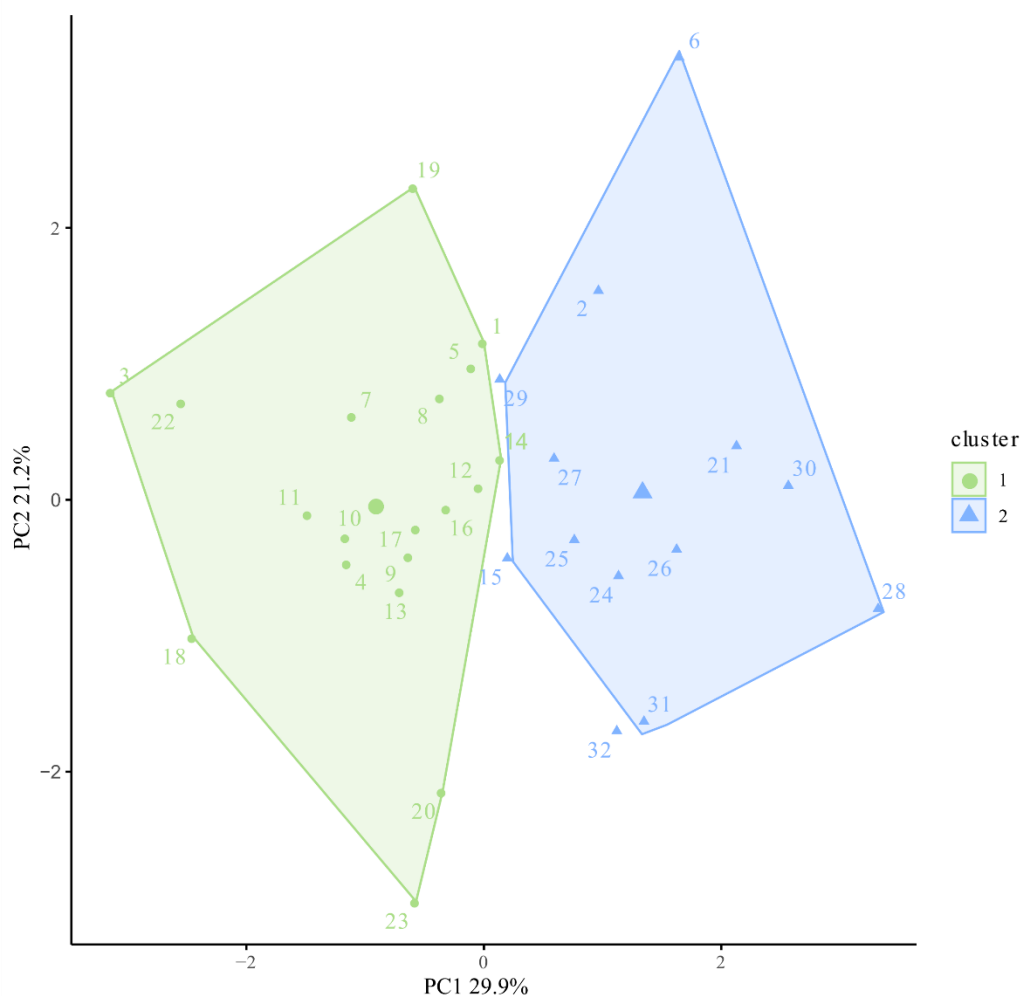
A classificação dos clusters foi feita usando validação cruzada estratificada com k-fold = 10 e dez repetições (100 execuções por modelo). Todos os parâmetros dos modelos seguiram as configurações padrão do software Weka 3.8.5, exceto para as Redes Neurais Artificiais, onde foram utilizadas duas camadas com 10 neurônios cada.

O desempenho de cada algoritmo foi mensurado por meio de duas métricas de acurácia: porcentagem de classificações corretas (CC) e F-score. Foram construídos boxplots para expressar acurácia dos modelos e dos inputs para cada métrica de acurácia. O agrupamento de médias foi realizado por meio do teste de Scott-Knott (Scott; Knott, 1974) a 5% de significância. A análise de PCA-Kmeans, as correlações de Pearson, os gráficos e os testes de médias foram realizados pelo software R (TEAM, 2013) e utilizando os pacotes ggplot2 e ExpDes.pt e Metan.

### 3 RESULTADOS

Os 32 genótipos de soja avaliados foram submetidos ao algoritmo k-means para a separação deles em dois grupos, denominados clusters em que no mesmo grupo ficasse os genótipos similares entre si (cluster 1) de acordo com EE, PB, FDN, FDA, MM e PROD, e distinto dos genótipos pertencentes ao outro grupo (cluster 2). A separação dos grupos foi demonstrada em um PCA para diminuir a dimensionalidade dos dados (Figura 1) após a separação dos grupos.

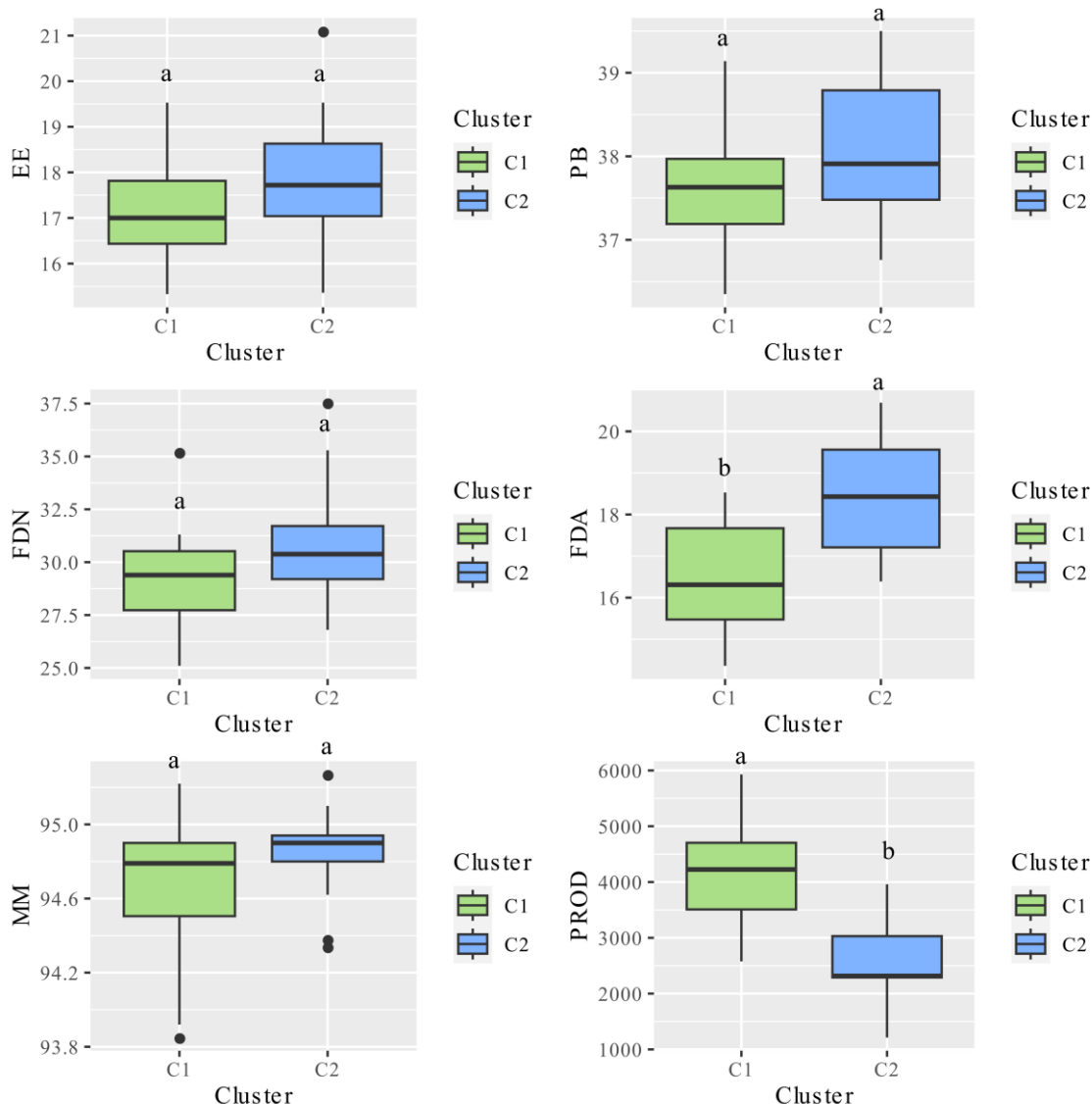
Figura 1 – Clusters formados pelo algoritmo K-means dividindo os 32 genótipos de soja de acordo com a sua similaridade em extrato etéreo (EE); proteína bruta (PB); fibra em detergente neutro (FDN); fibra em detergente ácido (FDA); cinzas ou matéria mineral (MM) e produtividade (PROD)



Fonte: elaboração própria.

Os genótipos separados entre os clusters apresentaram similaridade entre os dois grupos quanto EE, PB, FDN e MM. Entretanto os cluster se diferenciaram quanto a FDA, maior teor entre os genótipos do cluster 2 e PROD que foi maior entre os genótipos do cluster 1 (Figura 2). Podendo atribuir a ambas as características a diferenciação entre os genótipos dos clusters, em que os genótipos do cluster 1 são mais produtivos e do cluster 2 maior teor de FDA.

Figura 2 – Boxplot para as médias de cada cluster segundo os teores de extrato etéreo ou lipídios totais (EE); proteína bruta (PB); fibra em detergente neutro (FDN); fibra em detergente ácido (FDA); cinzas ou matéria mineral (MM) e produtividade (PROD)



EE - Extrato etéreo ou lipídios totais (%); PB - proteína bruta (%); FDN - fibra em detergente neutro (%); FDA - fibra em detergente ácido (%); MM - cinzas ou matéria mineral (%); PROD – produtividade (kg ha<sup>-1</sup>)

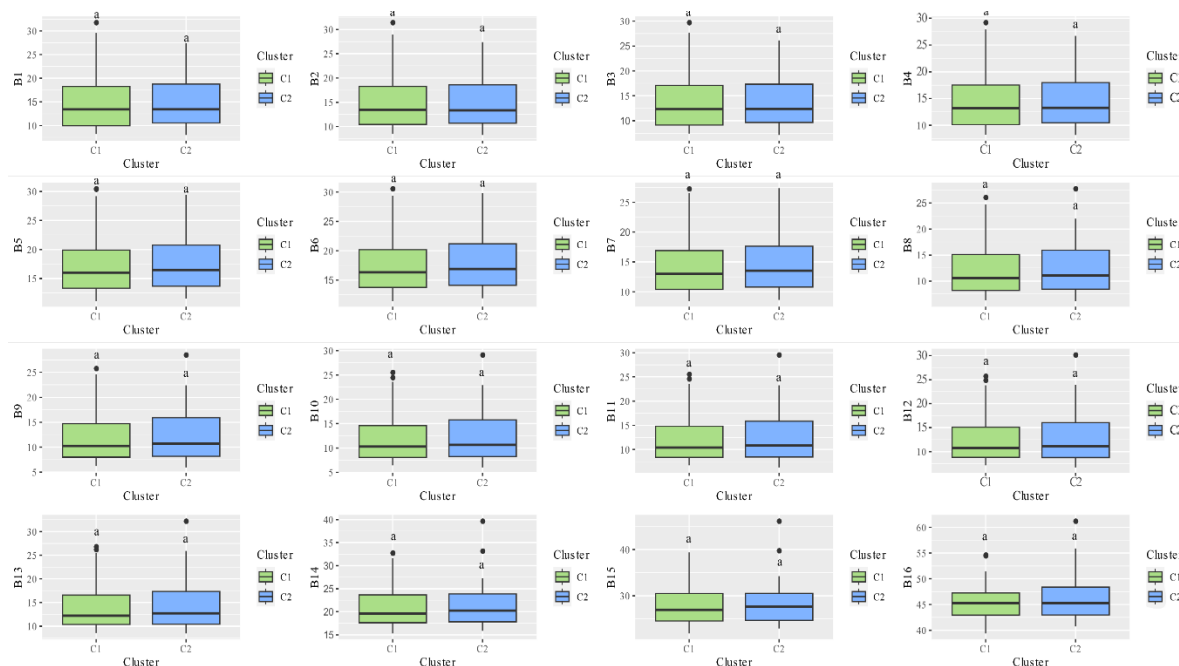
1)

Letras iguais entre os cluster não diferenciam entre si pelo teste T a 5% de probabilidade para cada variável.

Fonte: elaboração própria.

A média dos clusters para as 16 bandas avaliadas apresentaram similaridade entre si (Figura 3). Para cada cluster a média de suas reflectâncias foi muito similar entre elas não apresentando diferença estatística entre suas médias.

Figura 3 – Boxplot para as bandas de cada cluster segundo as bandas selecionadas



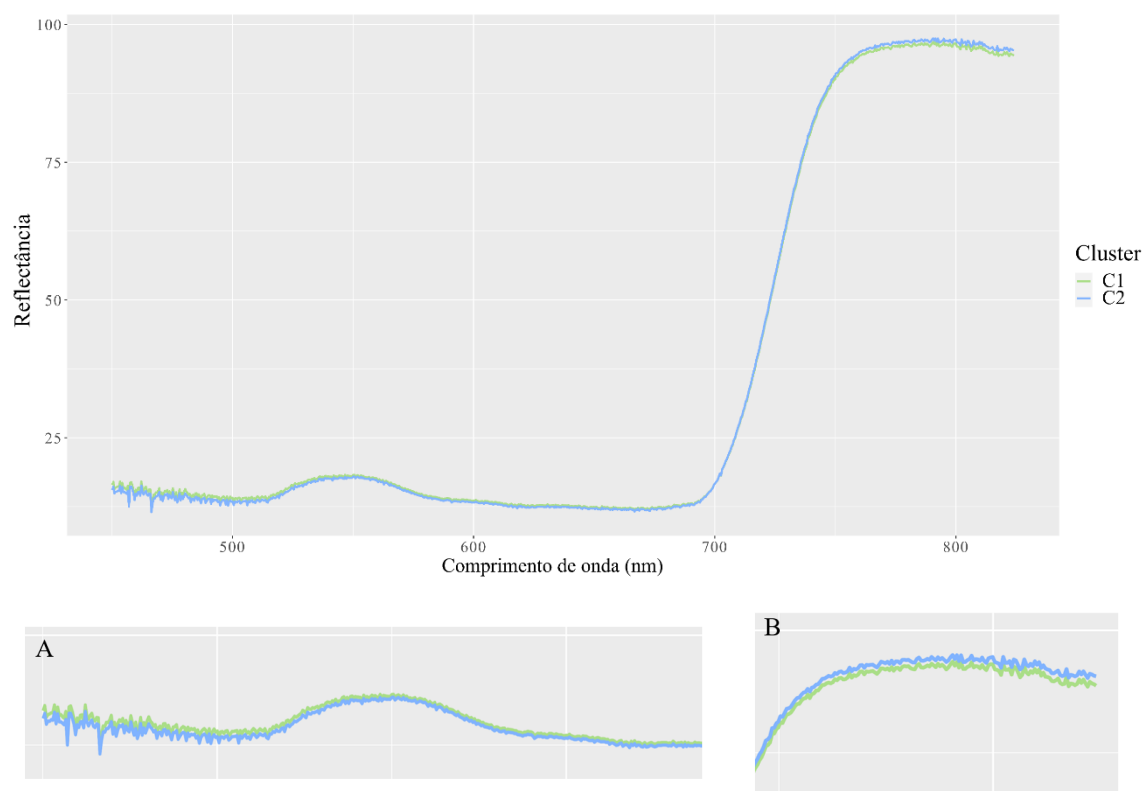
EE - Extrato etéreo ou lipídios totais (%); PB - proteína bruta (%); FDN - fibra em detergente neutro (%); FDA - fibra em detergente ácido (%); MM - cinzas ou matéria mineral (%); PROD – produtividade (kg ha<sup>-1</sup>)

Letras iguais entre os cluster não diferenciam entre si pelo teste T a 5% de probabilidade para cada variável.

Fonte: elaboração própria.

A curva espectral de cada cluster foi construída de acordo com a média dos genótipos pertencentes a cada um e suas respectivas reflectâncias dos comprimentos de onda de 450-824 nm (Figura 4). Nos comprimentos de onda entre 450nm e 550nm é observável uma leve elevação de cluster 1 quando comparado ao cluster 2. Entre 550 nm e 750 nm as duas curvas apresentaram comportamento muito semelhante, diferenciando-se novamente entre 750 nm e 824 nm em que houve uma elevação da curva do cluster 2.

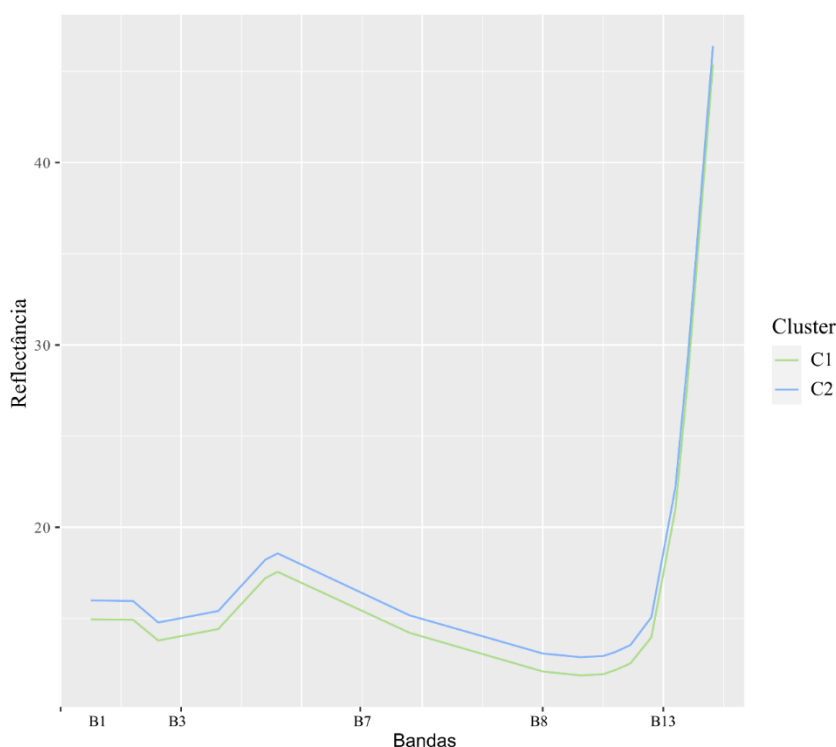
Figura 4 – Assinatura hiperespectral VIS/NIRs de ambos os clusters formados. Destaque para os comprimentos de onda 450nm e 550nm (A) e 750 nm e 824 nm (B) onde houve diferença entre as curvas



Fonte: elaboração própria.

Levando em consideração a média de reflectância dos clusters de acordo com as bandas selecionadas é observado maior diferença entre as curvas (Figura 5), já que é utilizado o comprimento de onda médio entre as faixas espectrais das bandas selecionadas, dessa forma generalizando a informação fornecida pela curva. A curva espectral formada pelo cluster 2 apresentou maior fator de reflectância que a curva formada pelo cluster 1 entre as bandas B1 e B13 abrangendo os comprimentos de onda de 450 - 699 nm, cuja diferença diminuiu a partir de B13, abrangendo os comprimentos de onda entre 700-720 nm.

Figura 5 – Curva dos clusters formados de acordo com as bandas pré-selecionadas

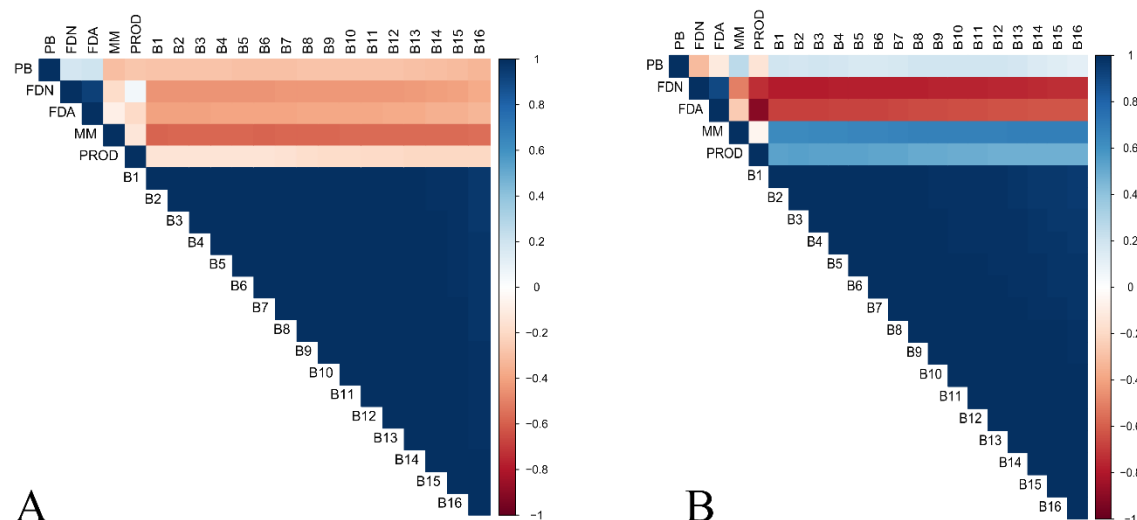


Fonte: elaboração própria.

Foi analisada a relação linear por meio da correlação de Pearson entre as variáveis industriais, produtividade e as bandas selecionadas para cada cluster formado (Figura 6). Para ambos os cluster a correlação entre as bandas foi similar e alta devido à base de obtenção dessas informações serem as mesmas. Diferenças são observadas entre as correlações para a produtividade e as bandas em que no Cluster 1 (Figura 6A) foi quase nula e negativa, o que não ocorreu no Cluster 2 (Figura 6B) em que as correlações por mais que seja pequena e fraca entre a PROD e as bandas possui maior correlação que no cluster 1, indicando uma melhor relação entre as variáveis.

Destaque para as correlações negativas observadas no cluster 2 de forma moderada entre PROD e FDA e fraca entre PROD e FDN. Vale ressaltar também a correlação negativa, mesmo que fraca, entre PROD e PB. FDN teve correlação moderada com EE no cluster 1 e no cluster 2 FDA obteve correlação moderada com EE. De forma geral, com exceção de FDN e FDA que alcançaram alta correlação, as demais relações entre as variáveis químicas do grão, produtividade e as reflectâncias são muito baixas ou inexistentes, abrindo oportunidade para a aplicação de análises estatísticas mais avançadas buscando relacionar tais informações de forma não linear.

Figura 6 – Correlação de Pearson para os teores de extrato etéreo ou lipídios totais (EE); proteína bruta (PB); fibra em detergente neutro (FDN); fibra em detergente ácido (FDA); cinzas ou matéria mineral (MM) e produtividade (PROD) e as bandas para o cluster 1 (A) e o cluster 2 (B)



Fonte: elaboração própria.

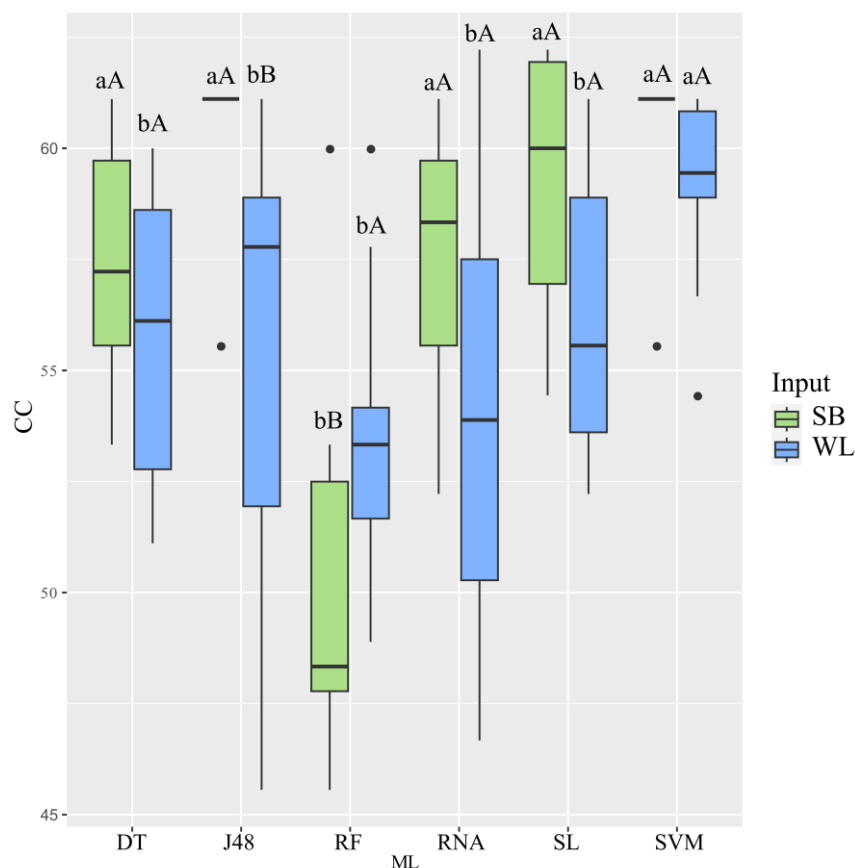
As variáveis dentro de cada cluster comportaram-se de forma diferente, mas de maneira geral as correlações foram baixas ou quase nulas. Assim, torna-se difícil relacionar tais características com uso de análises convencionais estatísticas. Então os dados foram submetidos a análise de aprendizagem de máquinas (ML). Foram utilizados seis algoritmos de ML e para cada um foram utilizados duas formas de entrada dos dados (Input), utilizando todos os comprimentos de onda (WL) e os comprimentos de onda médio para cada faixa espectral das 16 bandas selecionadas.

A interação entre ML e os inputs foram significativas para as duas métricas de acurácia testadas: Classificação Correta (CC) e F-score. Quando utilizado SB como entrada dos algoritmos a melhor performance foi observado por DT, J48, RNA, SL e SVM (Figura 7), destacando que J48 e SVM tiveram menor variabilidade tendendo a ser mais acurados que os demais. Quando WL foi utilizado como input apenas SVM apresentou o melhor desempenho.

Avaliando o comportamento de cada algoritmo com os dois inputs nota-se que DT, RNA, SL e SVM tiveram o mesmo comportamento. J48 apresentou melhor

desempenhou quando utilizado SB como entrada do que WL. RF apresentou melhor desempenho quando utilizado WL do que SB.

Figura 7 – Boxplot para porcentagem de classificação correta (CC) considerando a interação significativa entre os algoritmos de ML e as entradas testadas



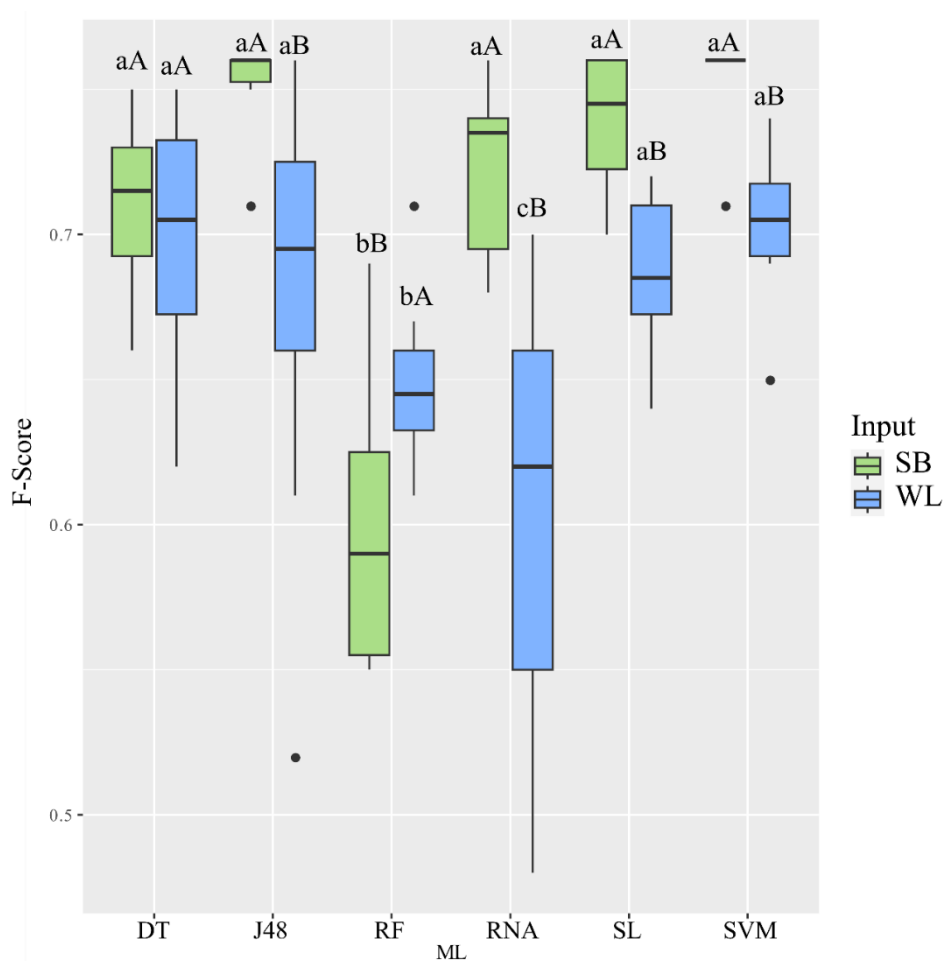
Médias seguidas pelas mesmas letras maiúsculas para as diferentes entradas e mesmas letras minúsculas para os diferentes algoritmos de ML não diferem pelo teste Scott-Knott a 5% de probabilidade.

Fonte: elaboração própria.

Os algoritmos DT, J48, RNA, SL e SVM apresentaram boa performance quando utilizado SB como input (Figura 8), destacando que assim como para CC o algoritmo SVM teve baixa variação dos dados, aumentando sua confiabilidade. Quando WL foi utilizado como input boas performances são observados por DT, J48, SL e SVM.

Comparado a performance de acurácia de cada algoritmo DT apresentou resultados similares usando SB e WL. J48, RNA, SL e SVM apresentou melhores resultados quando utiliza SB do que WL. WL apresentou melhor acurácia que SB para o algoritmo RF.

Figura 8 – Boxplot para F-score considerando a interação significativa entre os algoritmos de ML e as entradas testadas



Médias seguidas pelas mesmas letras maiúsculas para as diferentes entradas e mesmas letras minúsculas para os diferentes algoritmos de ML não diferem pelo teste Scott-Knott a 5% de probabilidade.

Fonte: elaboração própria.

Os resultados concordantes entre as métricas de acurácia foram: SB que garantiu a melhor performance para a maioria dos algoritmos, DT, J48, RNA, SL e SVM; Quando os algoritmos são avaliados utilizando WL como input, apenas SVM apresentou o melhor desempenho para ambas as acurácias.

Os algoritmos quando submetidos a ambos os inputs, DT apresentou o mesmo comportamento para as duas métricas de acurácia. J48 apresentou melhor desempenho quando utilizado SB e RF apresentou melhor desempenho quando utilizado WL.

Sendo assim, de maneira geral SVM e J48 teve destaque quando comparado aos demais algoritmos, especialmente quando utilizado como entrada dos algoritmos as bandas espectrais.

#### 4 DISCUSSÃO

Contribuir para os avanços do melhoramento genético de plantas é essencial para a agricultura moderna para acelerar os ganhos genéticos e monitorar de forma eficiente o ciclo da cultura para diminuir perdas qualitativas e quantitativas durante a produção agrícola (CHAWADE *et al.*, 2019). Ao longo dos anos dentro do melhoramento técnicas de genotipagem avançaram muito em tecnificação o que não ocorreu com a fenotipagem, que ainda continua sendo um gargalo para o desenvolvimento de técnicas precisas, ágeis e facilitada na mensuração de caracteres agronômicos e o monitoramento da cultura (MAHLEIN, 2016; MINERVINI; SCHARR; TSAFTARIS, 2015).

É necessário o desenvolvimento de metodologias para aprimorar o estudo do fenótipo das plantas como é o caso da fenotipagem de alta precisão (HTP) que vem se tornando aliada de pesquisas no melhoramento de plantas. Essa tecnologia vem sendo adotada por programas de melhoramento no âmbito público e privado que buscam melhorar a qualidade e a velocidade do melhoramento de plantas (KIM, 2020). Sendo assim os 32 genótipos de soja foram divididos em dois grupos denominados clusters, em que possuem características similares entre os genótipos pertencentes do mesmo grupo (Figura 2).

Sendo assim se o objetivo do programa de melhoramento for aumentar a produtividade dos genótipos, os pertencentes ao cluster 1 devem continuar no programa. Caso o objetivo seja melhorar características industriais os genótipos pertencentes ao cluster 2 devem ser prosseguidos no programa, devido ao maior teor de FDA dos genótipos do cluster 2 e destaque que PB possui correlação negativa com a PROD, como pode ser observado na correlação de Pearson na Figura 7.

É desafiador para o programa de melhoramento seguir com melhorias no rendimento de grãos e em caracteres do grão simultaneamente, devido as correlações negativas entre as características de interesse (RODRIGUES *et al.*, 2013). Linhagens que darão origem a cultivares comercialmente utilizadas pelos produtores, que possuem alto teor de proteína, possuem como características quantidade menores de sementes por planta, impedindo o alcance de elevados rendimento produtivos (ROTUNDO *et al.*, 2009; WEIH, 2003). Dessa forma, os pesquisadores devem direcionar o programa para o objetivo central do melhoramento que se deseja alcançar.

De forma geral o comportamento espectral dos clusters estudados se diferenciaram de forma mais destacada quando a curva foi construída utilizando a média

espectrais formadas pelas bandas pré-selecionadas (Figura 6). Nota-se que a curva formada pelo cluster 2 apresentou fator de reflectância superior a curva formada pelo cluster 1, podendo concluir que segundo a separação dos clusters, a reflectância foliar dos genótipos que possuíam caracteres industriais melhores é maior do que as dos genótipos que possuem maior produtividade.

A curva espectral dos genótipos é alterada de acordo com diferenças bioquímicas e estruturais entre eles que pode ser atribuída a diferença genética dos mesmos (ZHANG *et al.*, 2023). Observa-se que a maior diferença entre os clusters foi entre as bandas B1 e B13 conferindo de 450-695 nm, compreendendo a faixa espectral do visível. Essa faixa é está diretamente ligada ao conteúdo de  $\beta$ -caroteno (450 nm), Clorofila b (453 nm),  $\alpha$ -caroteno (470nm), Xantofila (475nm),  $\beta$ -caroteno (480 nm) e Síntese de Clorofila (650 nm) (SILVA JUNIOR *et al.*, 2018b). Pode-se inferir que o cluster 1 teve fator de reflectância menor pois estava absorvendo mais os comprimentos de onda citados, indicando maior atividade fotossintética, o que resultou em genótipos mais produtivos.

Havendo separado os genótipos em cluster e observado seu comportamento espectral, os dados foram submetidos a análise de aprendizagem de máquinas em que os algoritmos SVM e J48 alcançaram melhor performance na classificação dos clusters. Santana, *et al.* (2023), apresentaram bons resultados utilizando esses algoritmos para a classificação de soja quanto ao teor de macronutrientes primários em soja. Ambos os algoritmos, possuem diversas vantagens como a eficiência no processamento dos dados em menor tempo (ROSSI NETO *et al.*, 2017; VIEIRA *et al.*, 2012) com boas acurácias (MOUNTRAKIS; IM; OGOLE, 2011). SVM em especial apresenta ser robusto para lidar com classificação utilizando dados oriundos de sensores espectrais (BIGDELI; SAMADZADEGAN; REINARTZ, 2013; OKWUASHI; NDEHEDEHE, 2020).

A utilização das bandas espectrais selecionadas garantiu melhor performance dos algoritmos, especialmente SVM e J48. Pode-se atribuir essas respostas melhores devido a menor dimensão dos dados e a maior distinção espectral dos cluster quando utilizados as bandas que podem ser observadas na Figura 6. Dispor de várias bandas que representam partes importantes fisiologicamente da planta permite que haja melhor exploração das características da planta em questão (RAVIKANTH *et al.*, 2017).

A utilização de aprendizagem de máquinas é uma excelente opção na análise de dados hiperespectrais e variáveis industriais de grãos de soja, contornando problemas estatísticos que podem ser encontrados pela falta de linearidade entre as variáveis em

questão. Em trabalhos futuros seria interessante abranger mais análises inerentes ao grão e comprimentos de onda que vão além dos 800 nm.

## **5 CONCLUSÃO**

Os algoritmos SVM e J48 são os que apresentam melhor acurácia na classificação de características industriais e produtividade dos grãos . A utilização das bandas espectrais selecionadas no trabalho proporcionam melhor acurácia para os algoritmos na classificação.

## REFERÊNCIAS

- AHMED, M.; SERAJ, R.; ISLAM, S. M. S. The k-means algorithm: A comprehensive survey and performance evaluation. **Electronics**, [s. l.], v. 9, n. 8, p. 1295, 2020.
- ANDRADE, S. M.; TEODORO, L. P. R.; BAILO, F. H. R.; CAMPOS, C. N. S.; ROQUE, C. G.; SILVA JUNIOR, C. A. Da; CORADI, P. C.; TEODORO, P. E. High-throughput phenotyping of soybean genotypes under base saturation stress conditions. **Journal of Agronomy and Crop Science**, [s. l.], v. 207, n. 5, p. 814–822, 2021.
- ASAARI, M. S. M.; MERTENS, S.; DHONDT, S.; INZÉ, D.; WUYTS, N.; SCHEUNDERS, P. Analysis of hyperspectral images for detection of drought stress and recovery in maize plants in a high-throughput phenotyping platform. **Computers and Electronics in Agriculture**, [s. l.], v. 162, p. 749–758, 2019.
- BELGIU, M.; DRĂGUT, L. Random forest in remote sensing: A review of applications and future directions. **ISPRS Journal of Photogrammetry and Remote Sensing**, [s. l.], v. 114, p. 24–31, 2016.
- BIGDELI, B.; SAMADZADEGAN, F.; REINARTZ, P. A Multiple SVM System for Classification of Hyperspectral Remote Sensing Data. **Journal of the Indian Society of Remote Sensing**, [s. l.], v. 41, n. 4, p. 763–776, 2013.
- CHAWADE, A.; VAN HAM, J.; BLOMQUIST, H.; BAGGE, O.; ALEXANDERSSON, E.; ORTIZ, R. High-throughput field-phenotyping tools for plant breeding and precision agriculture. **Agronomy**, [s. l.], v. 9, n. 5, p. 258, 2019.
- SILVA JUNIOR, C. A.; NANNI, M. R.; SHAKIR, M.; TEODORO, P. E.; DE OLIVEIRA-JÚNIOR, J. F.; CEZAR, E.; DE GOIS, G.; LIMA, M.; WOJCIECHOWSKI, J. C.; SHIRATSUCHI, L. S. Soybean varieties discrimination using non-imaging hyperspectral sensor. **Infrared Physics & Technology**, [s. l.], v. 89, p. 338–350, 2018.
- DOBBELS, A. A.; LORENZ, A. J. Soybean iron deficiency chlorosis high-throughput phenotyping using an unmanned aircraft system. **Plant methods**, [s. l.], v. 15, n. 1, p. 1–9, 2019.
- EGMONT-PETERSEN, M.; DE RIDDER, D.; HANDELS, H. Image processing with neural networks—a review. **Pattern Recognition**, [s. l.], v. 35, n. 10, p. 2279–2301, 2002.
- FENG, X.; ZHAN, Y.; WANG, Q.; YANG, X.; YU, C.; WANG, H.; TANG, Z.; JIANG, D.; PENG, C.; HE, Y. Hyperspectral imaging combined with machine learning as a tool to obtain high-throughput plant salt-stress phenotyping. **The Plant Journal**, [s. l.], v. 101, n. 6, p. 1448–1461, 2020.
- GAO, S.; GUAN, H.; MA, X. A recognition method of multispectral images of soybean canopies based on neural network. **Ecological Informatics**, [s. l.], v. 68, p. 101538, 2022.

GENANGELI, A.; AVOLA, G.; BINDI, M.; CANTINI, C.; CELLINI, F.; GRILLO, S.; PETROZZA, A.; RIGGI, E.; RUGGIERO, A.; SUMMERER, S. Low-Cost Hyperspectral Imaging to Detect Drought Stress in High-Throughput Phenotyping. **Plants**, [s. l.], v. 12, n. 8, p. 1730, 2023.

GILL, T.; GILL, S. K.; SAINI, D. K.; CHOPRA, Y.; DE KOFF, J. P.; SANDHU, K. S. A Comprehensive Review of High Throughput Phenotyping and Machine Learning for Plant Stress Phenotyping. **Phenomix**, [s. l.], v. 2, n. 3, p. 156–183, 2022.

KIM, J. Y. Roadmap to high throughput phenotyping for plant breeding. **Journal of Biosystems Engineering**, [s. l.], v. 45, p. 43–55, 2020.

MAHLEIN, A.-K. Plant disease detection by imaging sensors—parallels and specific demands for precision agriculture and plant phenotyping. **Plant disease**, [s. l.], v. 100, n. 2, p. 241–251, 2016.

MINERVINI, M.; SCHARR, H.; TSAFTARIS, S. A. Image analysis: the new bottleneck in plant phenotyping [applications corner]. **IEEE signal processing magazine**, [s. l.], v. 32, n. 4, p. 126–131, 2015.

MOUNTRAKIS, G.; IM, J.; OGOLE, C. Support vector machines in remote sensing: A review. **ISPRS Journal of Photogrammetry and Remote Sensing**, [s. l.], v. 66, n. 3, p. 247–259, 2011.

NALEPA, J.; KAWULOK, M. Selecting training sets for support vector machines: a review. **Artificial Intelligence Review**, [s. l.], v. 52, n. 2, p. 857–900, 2019.

OKWUASHI, O.; NDEHEDEHE, C. E. Deep support vector machine for hyperspectral image classification. **Pattern Recognition**, [s. l.], v. 103, p. 107298, 2020.

QUINLAN, J. R. C4. 5: Programming for machine learning. **Morgan Kauffmann**, [s. l.], v. 38, n. 48, p. 49, 1993.

RAVIKANTH, L.; JAYAS, D. S.; WHITE, N. D. G.; FIELDS, P. G.; SUN, D.-W. Extraction of spectral information from hyperspectral data and application of hyperspectral imaging for food and agricultural products. **Food and bioprocess technology**, [s. l.], v. 10, p. 1–33, 2017.

RODRIGUES, J. I. S.; ARRUDA, K. M. A.; CRUZ, C. D.; PIOVESAN, N. D.; DE BARROS, E. G.; MOREIRA, M. A. Associação de marcadores microssatélites com teores de óleo e proteína em soja. **Pesquisa Agropecuária Brasileira**, [s. l.], v. 48, n. 3, p. 255–262, 2013.

ROSSI NETO, J.; DE SOUZA, Z. M.; DE MEDEIROS OLIVEIRA, S. R.; KÖLLN, O. T.; FERREIRA, D. A.; CARVALHO, J. L. N.; BRAUNBECK, O. A.; FRANCO, H. C. J. Use of the Decision Tree Technique to Estimate Sugarcane Productivity Under Edaphoclimatic Conditions. **Sugar Tech**, [s. l.], v. 19, n. 6, p. 662–668, 2017.

ROTUNDO, J. L.; BORRÁS, L.; WESTGATE, M. E.; ORF, J. H. Relationship between assimilate supply per seed during seed filling and soybean seed composition. **Field crops research**, [s. l.], v. 112, n. 1, p. 90–96, 2009.

SANTANA, D. C.; TEIXEIRA FILHO, M. C. M.; DA SILVA, M. R.; CHAGAS, P. H. M. Das; DE OLIVEIRA, J. L. G.; BAIO, F. H. R.; CAMPOS, C. N. S.; TEODORO, L. P. R.; DA SILVA JUNIOR, C. A.; TEODORO, P. E. Machine Learning in the Classification of Soybean Genotypes for Primary Macronutrients' Content Using UAV–Multispectral Sensor. **Remote Sensing**, [s. l.], v. 15, n. 5, p. 1457, 2023.

SCOTT, A. A. J.; KNOTT, M. A Cluster Analysis Method for Grouping Means in the Analysis of Variance Published by : International Biometric Society Stable URL : <http://www.jstor.org/stable/2529204>. **Biometrics**, [s. l.], v. 30, n. 3, p. 507–512, 1974.

SINGH, A. K.; SINGH, A.; SARKAR, S.; GANAPATHYSUBRAMANIAN, B.; SCHAPAUGH, W.; MIGUEZ, F. E.; CARLEY, C. N.; CARROLL, M. E.; CHIOZZA, M. V; CHITERI, K. O.; FALK, K. G.; JONES, S. E.; JUBERY, T. Z.; MIRNEZAMI, S. V; NAGASUBRAMANIAN, K.; PARMLEY, K. A.; RAIRDIN, A. M.; SHOOK, J. M.; VAN DER LAAN, L.; YOUNG, T. J.; ZHANG, J. High-Throughput Phenotyping in Soybean. Em: ZHOU, J.; NGUYEN, H. T. (Eds.). **High-Throughput Crop Phenotyping**. Cham: Springer International Publishing, 2021. p. 129–163.

SNOUSY, M. B. Al; EL-DEEB, H. M.; BADRAN, K.; KHLIL, I. A. Al. Suite of decision tree-based classification algorithms on cancer gene expression data. **Egyptian Informatics Journal**, [s. l.], v. 12, n. 2, p. 73–82, 2011.

ŠTEPANOVSKÝ, M.; IBROVÁ, A.; BUK, Z.; VELEMÍNSKÁ, J. Novel age estimation model based on development of permanent teeth compared with classical approach and other modern data mining methods. **Forensic Science International**, [s. l.], v. 279, p. 72–82, 2017.

TEAM, R. C. R: A language and environment for statistical computing. [s. l.], 2013.

TREVISAN, R.; PÉREZ, O.; SCHMITZ, N.; DIERS, B.; MARTIN, N. High-throughput phenotyping of soybean maturity using time series UAV imagery and convolutional neural networks. **Remote Sensing**, [s. l.], v. 12, n. 21, p. 3617, 2020.

VAN SOEST, P. J. Van; ROBERTSON, J. B.; LEWIS, B. A. Methods for dietary fiber, neutral detergent fiber, and nonstarch polysaccharides in relation to animal nutrition. **Journal of dairy science**, [s. l.], v. 74, n. 10, p. 3583–3597, 1991.

VIEIRA, M. A.; FORMAGGIO, A. R.; RENNO, C. D.; ATZBERGER, C.; AGUIAR, D. A.; MELLO, M. P. Object Based Image Analysis and Data Mining applied to a remotely sensed Landsat time-series to map sugarcane over large areas. **Remote Sensing of Environment**, [s. l.], v. 123, p. 553–562, 2012.

WEIH, M. Trade-offs in plants and the prospects for breeding using modern biotechnology. **New Phytologist**, [s. l.], p. 7–9, 2003.

ZHANG, H.; GOETTEL, W.; SONG, Q.; JIANG, H.; HU, Z.; WANG, M. L.; AN, Y. C. Selection of GmSWEET39 for oil and protein improvement in soybean. **PLoS Genetics**, [s. l.], v. 16, n. 11, p. e1009114, 2020.

ZHANG, J.; HAN, M.; WANG, L.; CHEN, M.; CHEN, C.; SHEN, S.; LIU, J.; ZHANG, C.; SHANG, J.; YAN, X. Study of Genetic Variation in Bermuda Grass along Longitudinal and Latitudinal Gradients Using Spectral Reflectance. **Remote Sensing**, [s. l.], v. 15, n. 4, p. 896, 2023.

ZUFFO, A. M.; STEINER, F.; AGUILERA, J. G.; TEODORO, P. E.; TEODORO, L. P. R.; BUSCH, A. Multi-trait stability index: A tool for simultaneous selection of soya bean genotypes in drought and saline stress. **Journal of Agronomy and Crop Science**, [s. l.], v. 206, n. 6, p. 815–822, 2020.

### **CAPÍTULO 3 – CLASSIFICAÇÃO DE GENÓTIPOS DE SOJA QUANTO A PRODUTIVIDADE DE GRÃOS E CARACTERES INDUSTRIAIS UTILIZANDO ESPECTROSCOPIA VNIR-SWIR**

#### **RESUMO**

O objetivo do trabalho foi avaliar a acurácia de técnicas de aprendizagem de máquinas em distinguir genótipos de soja de acordo com caracteres industriais dos grãos usando reflectância hiperespectral das folhas. A condução do experimento foi realizada na Universidade Federal de Mato Grosso do Sul, campus Chapadão do Sul na safra de 2021/22. No experimento foram avaliados 32 genótipos de soja alocados em blocos casualizados com quatro repetições. Aos 60 dias após a emergência (DAE) foi realizada a análise espectral de três amostras foliares de cada parcela. A análise espectral das folhas foi realizada com sensor hiperespectral disponibilizando as faixas de 350 a 2500 nm. Obtidos os comprimentos de onda, eles foram agrupados em médias de intervalos representativos em bandas. Ao fim do ciclo da cultura foi obtido a produtividade de grãos e posteriormente a determinação de carboidrato, teor de óleo e proteína. Inicialmente, os genótipos foram submetidos à análise de agrupamento por meio do algoritmo k-means e posteriormente, os dados foram submetidos a análise de aprendizagem de máquina, utilizando seis modelos: Árvores de decisão J48 (J48) e REPTree (DT), Floresta aleatória (RF), Redes neurais artificiais (RNA), Regressão logística (SL) e Máquina de vetor suporte (SVM). A regressão logística (SL) foi utilizada como ponto de referência por ser um algoritmo tradicional de regressão. Os clusters formados atuaram como saída (*output*) dos modelos, enquanto para as entradas (*input*) dos modelos, dois grupos de dados foram utilizados: as variáveis espectrais (WL) obtidas pelo sensor (350-2500 nm) e as médias espectrais das bandas selecionadas (SB) (350-2200 nm). A utilização de técnicas de aprendizagem de máquinas apresentou respostas inferiores a técnica padrão utilizada no trabalho, ou seja, a regressão logística (SL) que apresentou superioridade na classificação dos genótipos de soja quanto a variáveis industriais do grão. A utilização dos comprimentos de onda proporcionou melhor desempenho dos algoritmos na classificação.

**Palavras-chave:** regressão logística; comprimento de onda; NIR, SWIR

## 1 INTRODUÇÃO

A soja, *Glycine max* (L.) Merrill, é uma das plantas mais cultivadas comercialmente no mundo devido ser importante fonte de óleo e proteína. Ao longo dos anos de sua domesticação, diversas melhorias genéticas e sistemas de cultivo foram melhorados fazendo com que atualmente a soja seja a oleaginosa mais cultivada e utilizada globalmente (ANDERSON *et al.*, 2019). Devido a crescente demanda mundial por alimentos, e o destaque dado a soja, os programas de melhoramento da cultura devem dar atenção a composição dos grãos. Teor de proteína, óleo e carboidratos são algumas das características importantes que precisam ser melhoradas buscando aumentar o valor nutricional da soja como alimento para humanos e animais (GUPTA; MANJAYA, 2022).

Todas as características que configuram um grão com boas características industriais possuem relações complexas entre si e são fortemente influenciadas pelo ambiente, sendo assim mesmo utilizando linhagens elite, os melhoristas dificilmente conseguiriam melhorar múltiplas características de forma desejável (ZHOU *et al.*, 2019). Diversos avanços em tecnologias no sequenciamento genético, utilização de marcadores revolucionaram o melhoramento genético de diversas culturas e são tecnologias que cada vez mais se tecnificam, tornando todo o processo mais prático e rápido (SINGH *et al.*, 2019). Todavia a avaliação do fenótipo não acompanhou essa evolução, técnicas de fenotipagem rápida e acuradas em larga escala é um grande desafio para melhoristas. A utilização da fenotipagem de alto rendimento (HTP) com resolução espacial e temporal pode auxiliar programas de melhoramento quanto a análises fenotípicas (FURBANK; TESTER, 2011; WHITE *et al.*, 2012).

A utilização de sensores pode fornecer resultados rápidos, acurados em larga escala e de forma não destrutiva ao longo do ciclo da cultura, permitindo aliar informações do genótipo ao fenótipo (XIE; YANG, 2020). O fenótipo da cultura expõe diversas características estruturais, morfológicas, fisiológicas e outros desempenhos, que são resultantes da interação entre os genes das culturas e os fatores ambientais (HAGHIGHATTALAB *et al.*, 2016). Silva Junior *et al.* (2018), utilizando espectroscopia VNIR-SWIR conseguiu discriminar variedades de soja com eficiência, devido a distinção genéticas e fisiológicas entre elas. Há a possibilidade de distinguir genótipos de soja de acordo com teor nutricional de macronutrientes primários (SANTANA *et al.*, 2023a), relacionar caracteres como herdabilidade de caracteres agrônômicos e ganhos de seleção utilizando variáveis multiespectrais (OLIVEIRA *et al.*, 2023) e quantificar caracteres dos

grãos por meio de uso da reflectância das folhas utilizando sensor multiespectral (SANTANA *et al.*, 2023b).

Em análises de fenotipagem de alta precisão são gerados uma grande quantidade de dados que podem ser facilmente analisados com o uso de técnicas de aprendizagem de máquinas. Koc *et al.* (2022), recomendam o uso de sensores espectrais combinados com aprendizagem de máquina na identificação rápida de aparecimento de ferrugem amarela no final da estação de cultivo de trigo com uma precisão de previsão moderada. Dessa forma a utilização de aprendizagem de máquinas como procedimento de análise de dados tem muito a agregar a técnicas de HTP.

Avaliar a cultura a campo no período do florescimento e conseguir distinguir genótipos com diferentes caracteres de interesse para o programa de melhoramento de soja, possibilita que a seleção desses indivíduos seja mais acurada e rápida. Sendo assim o objetivo do trabalho foi avaliar a acurácia de técnicas de aprendizagem de máquinas em distinguir genótipos de soja de acordo com caracteres industriais dos grãos usando reflectância hiperespectral das folhas.

## 2 MATERIAL E MÉTODOS

A condução do experimento foi realizada na Universidade Federal de Mato Grosso do Sul, campus Chapadão do Sul (18°41'33''S, 52°40'45''W, com 810 m de altitude) na safra de 2021/22. O clima da região é Tropical de Savana (Aw), segundo a classificação de Köppen e Geiger. O solo é classificado como Latossolo Vermelho Distrófico Argiloso, com as seguintes características na camada de 0 – 0,20 m: pH (H<sub>2</sub>O) = 6,2; Al trocável (cmol<sub>c</sub> dm<sup>-3</sup>) = 0,0; Ca+Mg (cmol<sub>c</sub> dm<sup>-3</sup>) = 4,31; P (mg dm<sup>-3</sup>) = 41,3; K (cmol<sub>c</sub> dm<sup>-3</sup>) = 0,2; Matéria orgânica (g dm<sup>-3</sup>) = 19,74; V (%) = 45; m (%) = 0,0; Soma de bases (cmol<sub>c</sub> dm<sup>-3</sup>) = 2,3; CTC (cmol<sub>c</sub> dm<sup>-3</sup>) = 5,1.

No experimento foram avaliados 32 genótipos de soja, semeados em outubro de 2022, alocados em blocos casualizados com quatro repetições. As parcelas foram constituídas por quatro linhas de um metro de comprimento com espaçamento de 0,45 m e densidade de 15 sementes por metro.

Antes da semeadura, o solo foi preparado de maneira convencional, utilizando aração e gradagem niveladora. As sementes de soja receberam tratamento com uma combinação de fungicida (Piraclostrobina + Metil Tiofanato) e inseticida (Fipronil), aplicados na proporção de 200 mL de produto para cada 100 kg de sementes. Adicionalmente, foi realizada inoculação com *Bradyrhizobium* via sulco de semeadura para promover a fixação biológica de nitrogênio, na dose recomendada pelo fabricante. Durante o desenvolvimento da cultura, foram efetuados tratos culturais utilizando fungicidas, inseticidas e herbicidas conforme a necessidade.

Aos 60 dias após a emergência (DAE) foi realizada a análise espectral de três amostras foliares de cada parcela. A análise espectral das folhas foi realizada com espectrorradiômetro (FieldSpec 4 Jr, Analytical Spectral Devices, Boulder, EUA). O equipamento tem a capacidade de medir o espectro na faixa de 350 a 2500 nm. foi utilizado o equipamento ASD Plant Probe Para a leitura das folhas, que é projetado para realizar medições espectrais de contato em materiais sólidos. Uma das principais vantagens desse método é que ele não é destrutivo e não sofre interferência da luminosidade do ambiente, permitindo reduzir erros ocasionado pela luz difusa. Na calibração do equipamento foi utilizado uma placa branca que reflete 100% da luz, feita com sulfato de bário. O equipamento foi calibrado após o término de leituras das amostras de cada bloco, totalizando quatro calibrações.

O sensor estava conectado a um computador para os registros de cada leitura por meio de um software próprio do sensor chamado RS<sup>3</sup>, que registra as leituras feitas pelo

equipamento. Posteriormente os arquivos são importados para outro software, ViewSpectroPro para extração dos dados em extensão .txt, facilitando análises estatísticas posteriores.

Obtido os comprimentos de onda, eles foram agrupados em médias de intervalos representativos em bandas (Tabela 1), conforme sugerido por Silva Junior *et al.* (2018). Tais faixas estão relacionadas com processos fisiológicos importantes para a planta, características fisiológicas e suas respostas espectrais também foi descrito por Silva Junior *et al.* (2018)

Tabela 1 – Intervalos de faixa espectral (nm) e comprimento de onda médio (nm) usados no estabelecimento das 28 bandas espectrais.

| <b>Banda (n°)</b> | <b>Faixa espectral (nm)</b> | <b>Comprimento de onda médio (nm)</b> |
|-------------------|-----------------------------|---------------------------------------|
| 1                 | 350–369                     | 359.5                                 |
| 2                 | 370                         | 370.0                                 |
| 3                 | 371–419                     | 395.0                                 |
| 4                 | 420                         | 420.0                                 |
| 5                 | 421–424                     | 422.5                                 |
| 6                 | 425                         | 425.0                                 |
| 7                 | 426–444                     | 435.0                                 |
| 8                 | 445–475                     | 460.0                                 |
| 9                 | 480                         | 480.0                                 |
| 10                | 481–500                     | 490.5                                 |
| 11                | 501–530                     | 515.5                                 |
| 12                | 531–539                     | 535.0                                 |
| 13                | 540                         | 540.0                                 |
| 14                | 541–649                     | 595.0                                 |
| 15                | 650                         | 650.0                                 |
| 16                | 661–670                     | 665.5                                 |
| 17                | 675                         | 675.0                                 |
| 18                | 676–684                     | 680.0                                 |
| 19                | 685–689                     | 687.0                                 |
| 20                | 690–700                     | 695.0                                 |

| <b>Banda (n°)</b> | <b>Faixa espectral (nm)</b> | <b>Comprimento de onda médio (nm)</b> |
|-------------------|-----------------------------|---------------------------------------|
| 21                | 701–709                     | 705.0                                 |
| 22                | 710                         | 710.0                                 |
| 23                | 711–730                     | 720.5                                 |
| 24                | 960                         | 960.0                                 |
| 25                | 1100                        | 1100.0                                |
| 26                | 1400                        | 1400.0                                |
| 27                | 1930                        | 1930.0                                |
| 28                | 2200                        | 2200.0                                |

Fonte: Silva Junior *et al.* (2018).

Ao fim do ciclo da cultura foi realizado a colheita total da parcela, pesada e corrigida a umidade para 13% para fins de cálculo de produtividade de grãos (PROD, kg ha<sup>-1</sup>). Uma amostra de 50 g foi separada de cada parcela e moída para a realização das análises industriais do grão. A proteína bruta (PB) foi determinada conforme metodologia descrita por (SILVA; QUEIROZ, 2002)

A determinação das concentrações de carboidratos solúveis totais nas sementes de soja foi realizada pelo método de DuBois *et al.* (1956). Para as análises 50 mg de sementes secas e moída foram colocadas em tubos de ensaios de 15 mL, onde foi adicionado 5 mL de água destilada e colocados em banho-maria por 30 minutos a 100° C. Posteriormente, os tubos foram levados a centrifugação (3000 rpm) durante 10 minutos. Em seguida, 100 µL do sobrenadante foi retirado e colocados em tubos de ensaio contendo 400 µL de água destilada, sendo submetidos a agitação através de vórtex. Depois foram adicionados 0,5 mL de fenol 5% e 2,5 mL de ácido sulfúrico concentrado. Os tubos de ensaio foram deixados em repouso por 20 minutos e sua leitura foi realizada no espectrofotômetro a 490 nm, tendo como branco 0,5 mL de água destilada, 0,5 mL de fenol e 2,5 mL de ácido sulfúrico. Para o cálculo das concentrações de carboidratos solúveis totais utilizou-se uma curva-padrão de glicose e os resultados foram expressos em g de carboidratos / g MS.

Para a extração de óleo foram separadas 5 g da amostra seca e foram moídas e então, acondicionadas em cartucho de papel filtro. Os cartuchos foram colocados no sistema e 150 mL de hexano para análise (PA), ou seja, de alta pureza, colocados nos balões e conectado a um aparelho de extração de cápsulas. Para este efeito, o gotejamento de extração de solvente e óleo foi monitorado para 4 h. Após a extração, a válvula de

fluxo foi fechada e o solvente foi recuperado. O balão contendo o óleo foi desacoplado do sistema e levado a uma estufa com renovação de ar e circulação para evaporar do solvente até que o par (béquer + óleo) manteve um peso constante. O óleo conteúdo foi determinado gravimetricamente. De acordo com a metodologia descrita pela Association of Químicos Analíticos Oficiais (AOAC, 1990). Com a obtenção das variáveis industriais citadas e a produtividade, foram realizadas análises estatísticas de aprendizagem de máquina. Inicialmente, os genótipos foram submetidos à análise de agrupamento por meio do algoritmo k-means. Primeiramente, o algoritmo k-means é empregado para calcular as distâncias entre cada amostra e os centroides dentro de cada classe. Com base nessas distâncias, as amostras são atribuídas a seus respectivos grupos, considerando-as semelhantes às amostras pertencentes ao mesmo grupo e distintas das do outro grupo. Os clusters resultantes do k-means são então representados em um gráfico biplot por meio da análise de componentes principais (PCA), ilustrando a separação entre eles. Esse método requer a definição prévia do valor de k, que representa o número de grupos desejado antes da execução da análise (AHMED *et al.*, 2020). No presente estudo, o valor de k que melhor discriminou os grupos de acordo com as características industriais do grão e a produtividade foi estabelecido em 2, resultando na separação dos genótipos em dois grupos distintos: Cluster 1 e Cluster 2.

Após a segregação dos 32 genótipos em dois clusters distintos, foram criados boxplots comparativos para analisar as médias de cada característica industrial e produtividade, com o objetivo de observar as diferenças entre os clusters formados. Além disso, foi conduzida uma análise de correlação de Pearson entre as variáveis e as bandas selecionadas, apresentadas em um gráfico heatmap para cada cluster.

Posteriormente, os dados foram submetidos a algoritmos de aprendizagem de máquina, utilizando seis modelos distintos (Tabela 2). A regressão logística (SL) foi utilizada como ponto de referência por ser um algoritmo tradicional de regressão, buscando-se a superioridade dos outros algoritmos em relação a este. Os clusters formados atuaram como saída (output) dos modelos, enquanto para as entradas (input) dos modelos, dois grupos de dados foram utilizados: as variáveis espectrais (WL) obtidas pelo sensor (350-2500 nm) e as médias espectrais das bandas selecionadas (SB) (350-2200 nm). O objetivo era permitir que os algoritmos distinguíssem os clusters com base nos dados espectrais dos genótipos associados a cada um deles.

Tabela 2 – Relação dos modelos de aprendizagem de máquinas utilizados na classificação

| Sigla | Modelo de aprendizagem de máquinas | Referência                                  |
|-------|------------------------------------|---------------------------------------------|
| DT    | REPTree                            | (SNOUSY <i>et al.</i> , 2011)               |
| J48   | Árvores de decisão J48             | (QUINLAN, 1993)                             |
| RF    | Floresta aleatória                 | (BELGIU; DRĂGUȚ, 2016)                      |
| ANN   | Redes neurais artificiais          | (EGMONT-PETERSEN; DE RIDDER; HANDELS, 2002) |
| SL    | Regressão logística                | (ŠTEPANOVSKÝ <i>et al.</i> , 2017)          |
| SVM   | Máquina de vetor suporte           | (NALEPA; KAWULOK, 2019)                     |

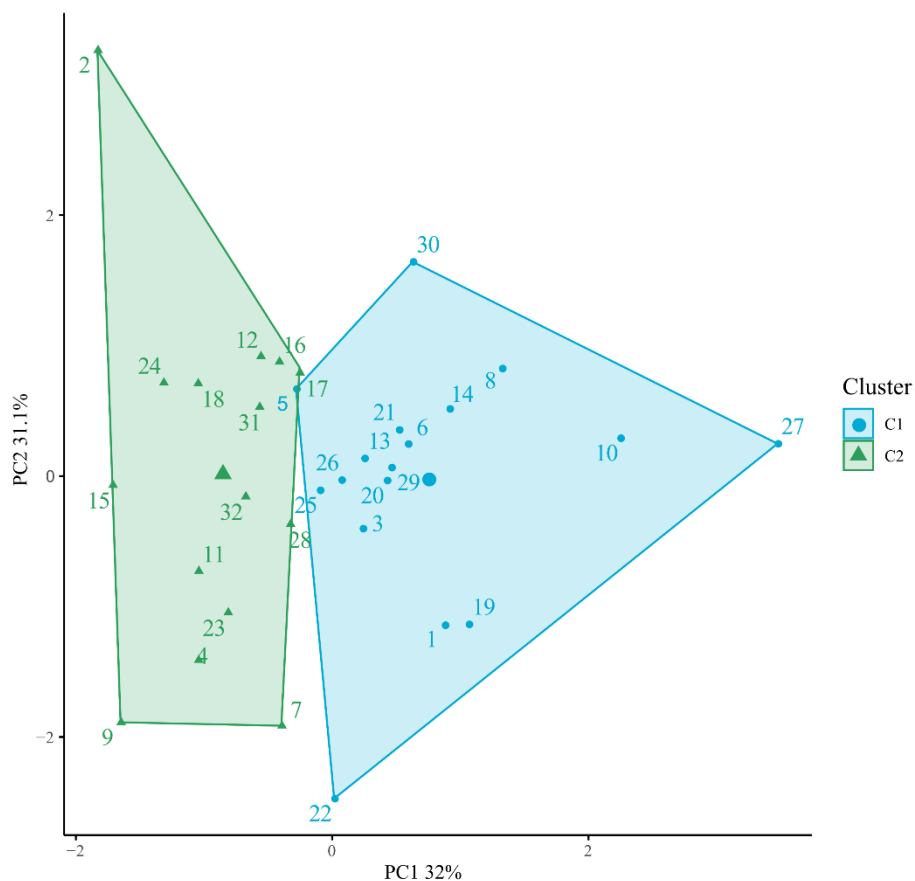
Fonte: elaboração própria.

A avaliação dos clusters pelos algoritmos de aprendizagem de máquina foi conduzida utilizando validação cruzada estratificada com  $k\text{-fold} = 10$  e repetida por dez vezes (totalizando 100 execuções por modelo). Todos os parâmetros dos modelos foram configurados com as opções padrão do software Weka 3.8.5, exceto para as Redes Neurais Artificiais (ANN), que foram configuradas com duas camadas compostas por 10 neurônios cada. Para medir o desempenho de cada algoritmo, foram utilizadas duas métricas de acurácia: a porcentagem de classificações corretas (CC) e o F-score. Os resultados foram representados por meio de boxplots, exibindo a acurácia dos modelos e das entradas para cada métrica. As médias dos resultados foram agrupadas através do teste de Scott-Knott (Scott; Knott, 1974) com nível de significância de 5%. Para a análise de PCA-Kmeans, correlações de Pearson, criação de gráficos e realização dos testes de médias, foram empregados o software R e os pacotes ggplot2, ExpDes.pt e Metan.

### 3 RESULTADOS

Os 32 genótipos de soja foram separados em 2 grupos distintos de acordo com os teores de carboidratos, óleo, proteína e produtividade de grãos. Os genótipos foram separados utilizando os algoritmos k-means determinado para dividir os genótipos em dois grupos, denominado cluster, em que os indivíduos dentro do mesmo grupo possuem características similar entre si, mas distinta do outro grupo. Essa divisão foi plotada em um gráfico de componentes principais buscando diminuir a dimensionalidade dos dados (Figura 1).

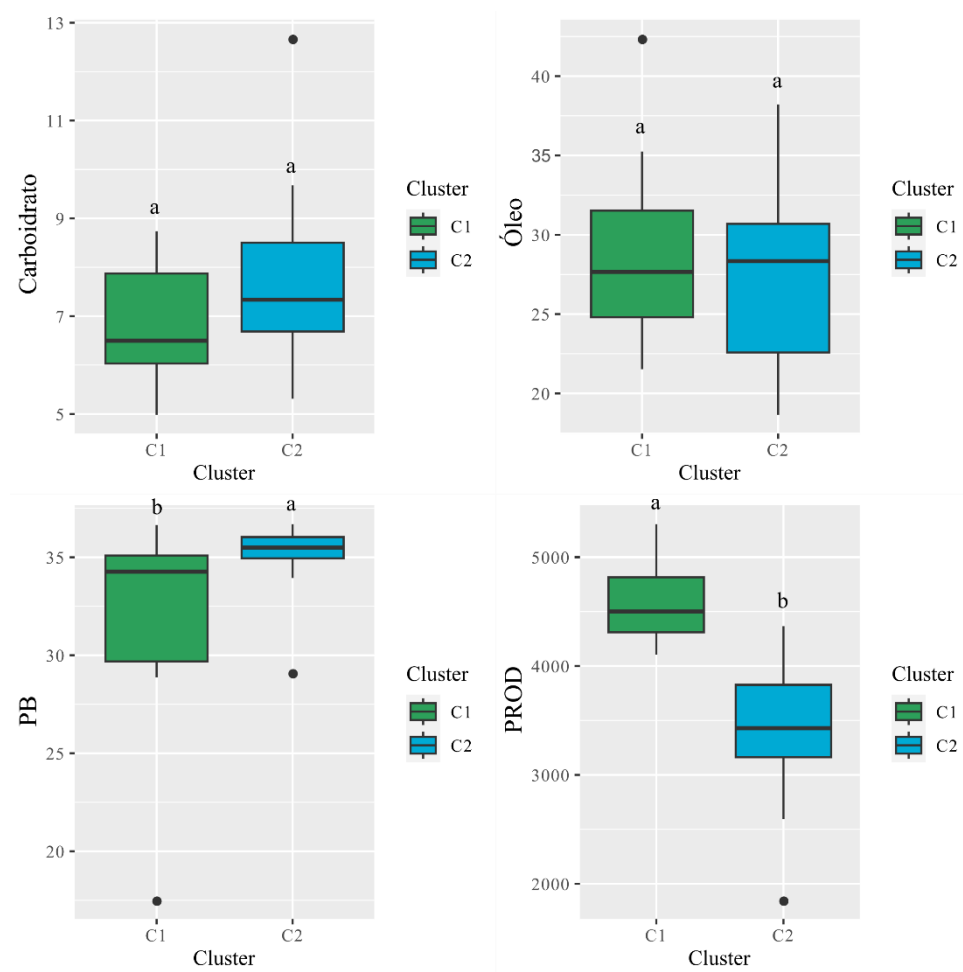
Figura 1 – Divisão dos 32 genótipos de soja em dois Clusters formados pelo algoritmo K-means de acordo com a sua similaridade em: Carboidratos ( $\text{g } 100\text{g}^{-1}$ ); óleo (%), proteína bruta (PB, %) e produtividade (PROD,  $\text{kg ha}^{-1}$ ).



Fonte: elaboração própria.

Buscando identificar as variáveis que fizeram os grupos se distinguirem, foi construído um boxplot com as médias das variáveis industriais de cada cluster (Figura 2). Nota-se que o teor de proteína é maior no cluster 2 e a PROD é a maior para o cluster 1. Dessa forma, pode-se afirmar que os genótipos que constituem o cluster 1 possuem maior PROD que os genótipos do cluster 2, que por sua vez possuem maiores teores de PB.

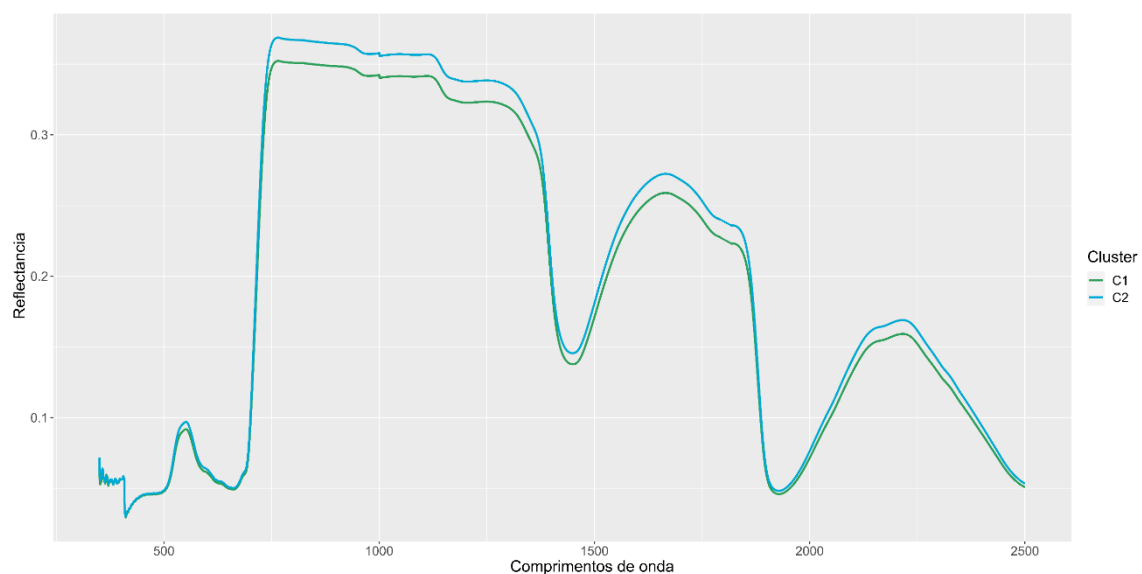
Figura 2 – Boxplot para as médias de carboidratos (g); óleo (%), proteína bruta (PB, %) e produtividade (PROD, kg ha<sup>-1</sup>) para cada cluster.



Letras iguais entre os cluster não diferenciam entre si pelo teste T a 5% de probabilidade para cada variável.  
Fonte: elaboração própria.

A Figura 3 representa a curva espectral das folhas dos genótipos pertencentes a cada cluster formado, demonstrando uma leve diferença no fator de reflectância entre eles, em que o cluster 1 possui menor fator de reflectância que o cluster 2. Como visto anteriormente, o que diferencia os dois clusters é o teor de PB maior para o cluster 2 e a PROD maior para o cluster 1. Pode-se inferir que os genótipos que constituem o cluster 2 possuem maior teor de PB e maior fator de reflectância, e, os genótipos do cluster 1 são mais produtivos e possuem menor fator de reflectância.

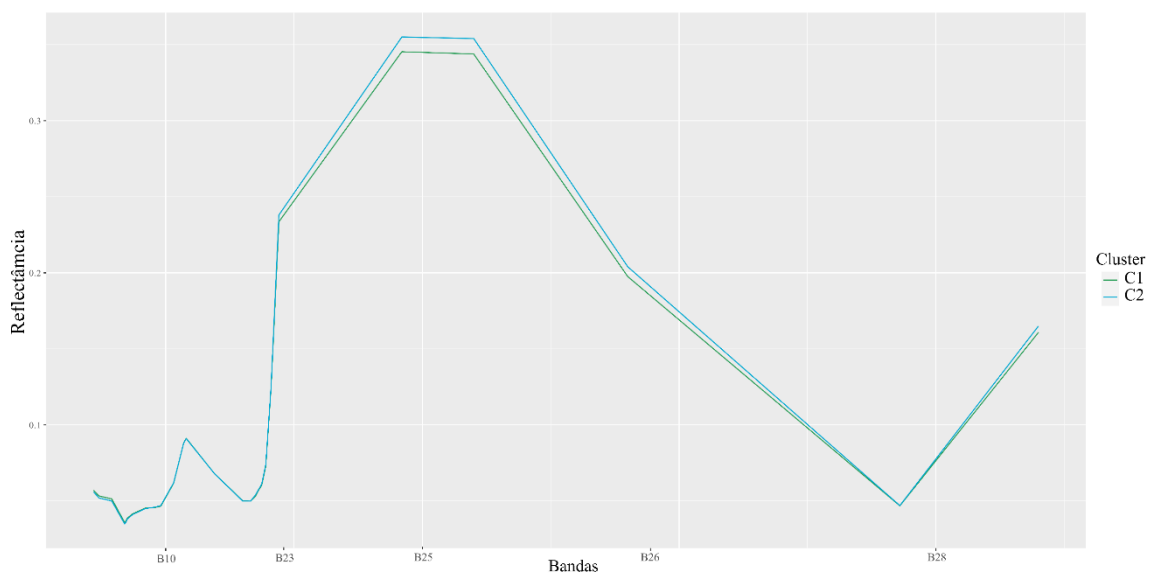
Figura 3 – Assinatura hiperespectral dos clusters formados.



Fonte: elaboração própria.

Na Figura 4 é representada a curva hiperespectral dos clusters de acordo com as bandas selecionadas, em que apresentaram comportamento muito similar a curva espectral usando os comprimentos de onda. A curva do cluster 2 apresentou maior fator de refletância que a do cluster 1.

Figura 4 – Curva hiperespectral dos clusters formados de acordo com as bandas pré-selecionadas.



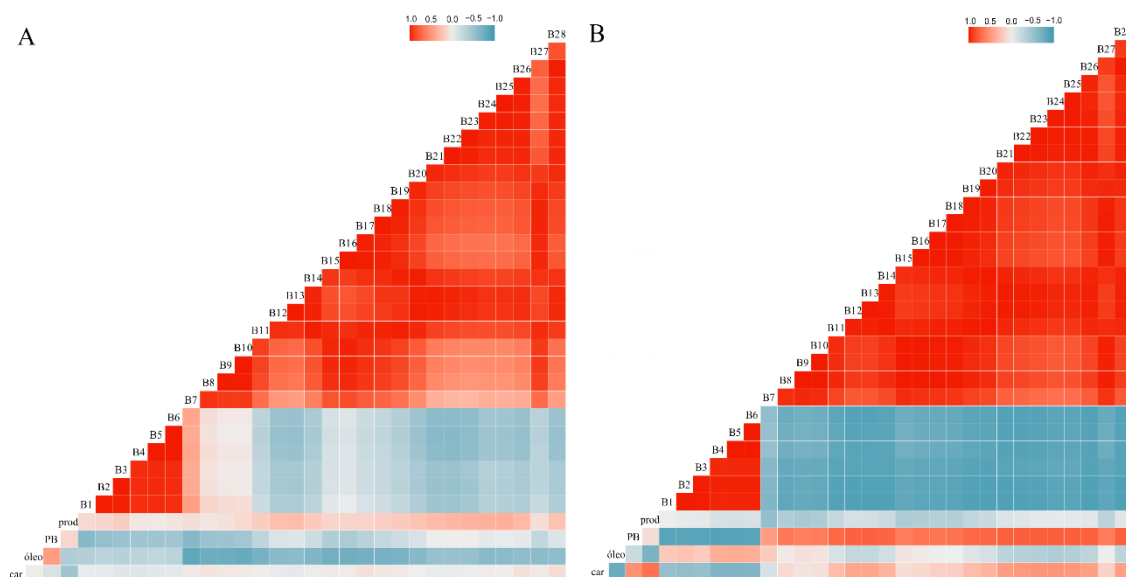
Fonte: elaboração própria.

Além dos teores industriais e a produtividade apresentarem ser diferente entre os cluster a correlação entre essas variáveis dentro de cada cluster apresentaram diferença como pode ser observado na Figura 5. Carboidrato teve alta correlação negativa com óleo entre os genótipos do cluster 2 e correlação muito fraca com a mesma variável entre os genótipos do cluster 1. Dessa forma os genótipos pertencentes ao cluster 2 se melhorados para óleo terá diminuição em carboidrato e vice-versa, o que não ocorrerá entre os genótipos do cluster 1 que podem ainda ser melhorados para ambas as características.

Teor de PB teve alta correlação positiva com óleo entre os genótipos do cluster 1 e uma correlação fraca e negativa entre os genótipos do cluster 2. No cluster 2, PB teve alta correlação positiva com carboidrato. No cluster 1, a PROD obteve alta correlação negativa com carboidrato, moderada e negativa com óleo. No cluster 2, a PROD teve alta correlação positiva com carboidrato e alta correlação negativa com óleo.

No cluster 2, carboidrato e proteína tiveram alta correlação negativa com as bandas 1, 2, 3, 4, 5 e 6. As mesmas bandas tiveram alta correlação negativa com proteína bruta no entre os genótipos do cluster 1. No cluster 1 óleo obteve alta correlação negativa entre as bandas 7 a 27 e correlações positivas de fraca a moderada magnitude pode ser observada entre PROD e as bandas também de 7 a 28. No cluster 2, carboidrato teve correlações positivas de fraca a moderada magnitude entre as bandas de 10 a 28. Correlações de alta magnitude positiva foram observadas entre PB e as bandas B7 a B28. Correlações negativas de fracas a moderadas foram observadas entre as bandas B7 a B28.

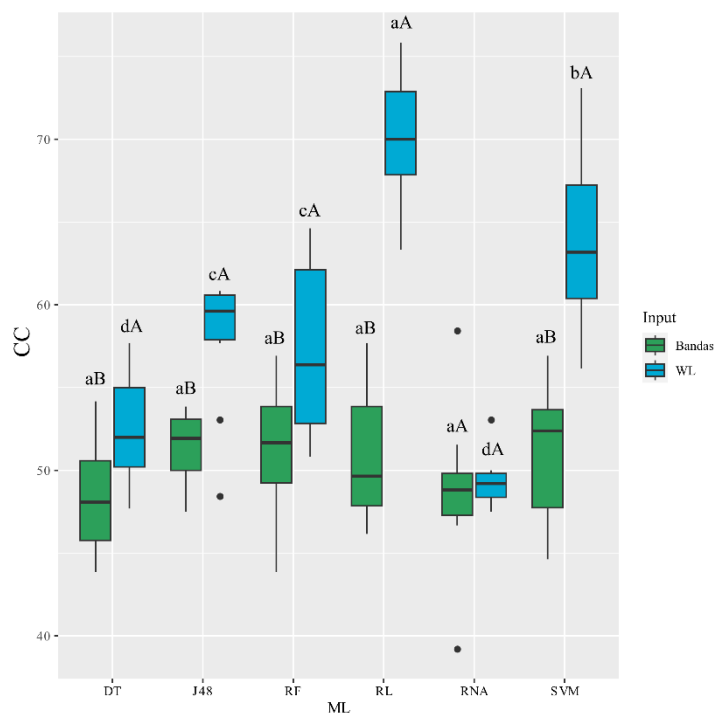
Figura 5 – Correlação de Pearson entre carboidratos ( $\text{g } 100\text{g}^{-1}$ ); óleo (%), proteína bruta (PB, %), produtividade (PROD,  $\text{kg ha}^{-1}$ ) e as 28 bandas selecionadas para o cluster 1 (A) e cluster 2 (B)



Fonte: elaboração própria.

Os dados foram submetidos a análise de aprendizagem de máquina em que a interação entre os modelos testados e os inputs foram significativos para as duas métricas de acurácia testadas (classificação correta e F-score). Quando o input utilizado foi bandas, não houve diferença na acurácia de classificação para CC (Figura 6). Quando o input utilizado foi WL, o melhor algoritmo foi RL. Comparando a performance de cada algoritmo com os dois inputs, para todos, utilizando WL proporcionou melhor acurácia aos algoritmos do que as bandas.

Figura 6 – Médias de classificação correta (CC) para a interação significativa entre os algoritmos de ML e os inputs testados

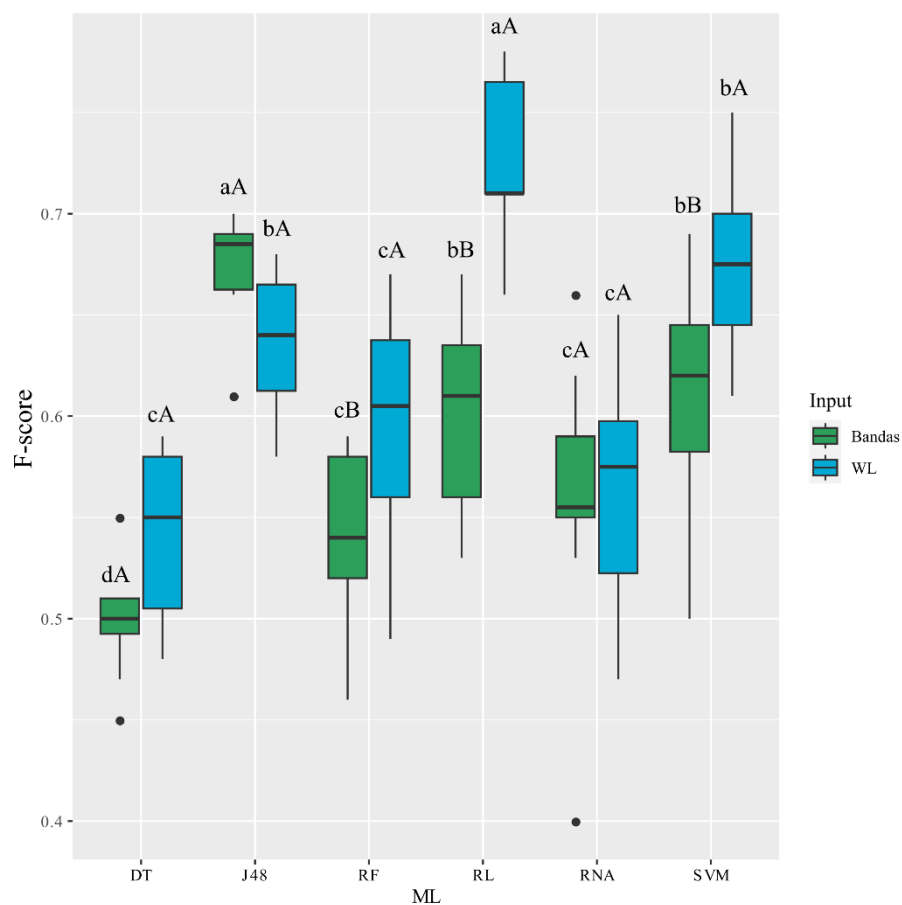


Médias seguidas pelas mesmas letras maiúsculas para os diferentes inputs e mesmas letras minúsculas para os diferentes algoritmos de ML não diferem pelo teste Scott-Knott a 5% de probabilidade.

Fonte: elaboração própria.

A métrica de acurácia F-score apresentou melhor acurácia para RL quando o input utilizado foi WL. Quando input utilizado foi as bandas, o melhor algoritmo foi o J48 (Figura 7). Comparando cada algoritmo com os dois inputs testados, a melhor performance foi observada quando WL foi utilizado.

Figura 7- Médias de F-score para a interação significativa entre os algoritmos de ML e os inputs testados.



Médias seguidas pelas mesmas letras maiúsculas para os diferentes inputs e mesmas letras minúsculas para os diferentes algoritmos de ML não diferem pelo teste Scott-Knott a 5% de probabilidade.

Fonte: elaboração própria.

Na análise de aprendizagem de máquinas pode-se observar que RL utilizado como padrão apresentou melhores resultados que os demais algoritmos para as duas métricas de acurácia. A utilização dos comprimentos de onda como input favoreceu bom desempenho aos algoritmos.

#### 4 DISCUSSÃO

Em programas de melhoramento de soja são avaliados diversos genótipos com o intuito de gerar populações com características melhores, dentre essas características a de maior destaque é a produtividade (HARTMAN; WEST; HERMAN, 2011). Selecionar genótipos de soja com melhores características de grão também são variáveis que merecem atenção no processo de melhoramento da soja (SANTANA *et al.*, 2023). Dessa forma os 32 genótipos de soja foram separados em dois grupos, denominados cluster de acordo com similaridade e diferença pertence entre os genótipos (Figura 1). Os genótipos foram separados de acordo com o teor de proteína e a produtividade. Genótipos mais produtivos foram separados e ficaram no cluster 1 e genótipos com maior teor de proteína bruta ficaram no cluster 2. O teor de proteína da soja é uma variável negativamente correlacionada com o rendimento e o óleo, ou seja, quando melhorado para um o teor do outro diminui, o que dificulta a melhoria simultânea dessas variáveis (PRENGER *et al.*, 2019).

Dessa forma, é importante o programa de melhoramento direcionar os genótipos de acordo com o objetivo final do melhorista (GUO *et al.*, 2022), sendo assim a separação dos genótipos em grupos de acordo com as características de interesse diminui o tamanho dos indivíduos avaliados. Para facilitar ainda mais esse processo, os clusters formados tiveram curvas espectrais diferentes, tanto utilizando os comprimentos de onda (Figura 3) quanto utilizando as médias das bandas selecionadas (Figura 4). Na região do visível de 400 – 700 nm o comportamento de ambos os clusters foi similar. Essa região é associada ao comportamento dos pigmentos fotossintetizantes da folha (SILVA JUNIOR *et al.*, 2018; Moreira, 2001). A primeira diferenciação entre os clusters é observada na região do infravermelho próximo (NIR) que confere a faixa de 700 a 1300 nm, é caracterizado por altos valores de refletância em comparação com a luz visível (JUNGES *et al.*, 2020) e está relacionado com a forma das células e quantidade de espaços intercelulares das folhas (Silva Junior *et al.*, 2018), essa faixa também pode fornecer informações importantes quanto ao teor de nutrientes foliares (PRANANTO; MINASNY; WEAVER, 2020) e por estar relacionado a parte bioquímica da planta pode dar informações importantes sobre a alteração ou perda de alguma funcionalidade do tecido foliar (JUNGES *et al.*, 2020).

A partir de 1300 até 2500 nm é denominado a região do Infravermelho de ondas curtas (SWIR), há uma diminuição gradativa da refletância e está relacionada ao teor de água e atividade química das folhas (JUNGES *et al.*, 2020; SILVA JUNIOR *et al.*, 2018) podendo inferir a respeito do conteúdo de água foliar ou condutância estomática

(HOLZMAN; RIVAS; BAYALA, 2021). De acordo com a assinatura espectral dos cluster formados, o cluster 1 teve menor refletâncias que o cluster 2, assim como avaliando comportamento da curva utilizando as bandas também foi caracterizado pela maior reflectância do cluster 2, especialmente na B25 (1100 nm) pertencente a região do SWIR. Pode-se inferir através dos resultados que os genótipos mais produtivos possuem menos fator de reflectância do que os indivíduos com maior teor de PB. A resposta espectral dos indivíduos se deve a comportamentos bioquímicos, estruturais e genéticos diferentes entre eles (ZHANG *et al.*, 2023) tendo maiores diferenças na região do NIR-SWIR entre os clusters.

Além das curvas espectrais se distinguirem entre os clusters a correlação entre as variáveis avaliadas também se comportaram de forma diferente entre os clusters. No cluster 1, constituído pelos genótipos de maior PROD, houve alta correlação negativa entre PROD e carboidrato e moderada e negativa entre PROD e óleo, ou seja, conforme esses genótipos forem melhorados para aumentar a produtividade, os teores de carboidrato e óleo irá diminuir. Em contrapartida para os genótipos desse cluster, houve alta correlação positiva entre PB e óleo. Óleo e proteína são caracteres correlacionados negativamente, todavia até que esses teores sejam estabilizados no genótipo ele pode ser mesmo sendo aumentado simultâneo até certo ponto (LIN *et al.*, 2015).

No cluster 2 houve alta correlação negativa entre carboidrato e óleo, PROD e óleo e alta correlação positiva entre PB e carboidrato, entre PROD e carboidrato. Ou seja, os genótipos pertencentes a esse grupo foram os que apresentaram maior teor de PB que podem ser melhorados para aumentar mais esse teor e conseqüentemente terá aumento de carboidrato. Caso haja aumento do teor de óleo entre os genótipos desse grupo, haverá declínio de carboidratos e produtividade. Houve também uma correlação positiva entre PROD e carboidrato, todavia o aumento de PROD pode acarretar a diminuição de PB, dessa forma ficando a escolha do melhorista qual das características se alinha aos seus objetivos.

Foi possível observar altas correlações entre as bandas selecionadas e as variáveis industriais dos grãos. Tais características só são possíveis graças ao grande número de informações disponibilizadas pelo sensor hiperespectral, que permite melhor exploração das características com mais detalhes (RAVIKANTH *et al.*, 2017). A esse fato também se atribui o melhor desempenho do algoritmo RL e do input WL. A utilização de algoritmos de aprendizagem de máquina (ML) vem se popularizando devido ao seu poder de resolver problemas com predição e classificação, e seus resultados dependem das

informações fornecidas aos algoritmos (CHRISTODOULOU *et al.*, 2019). Dessa forma em alguns casos o desempenho das técnicas de ML pode ser semelhante a técnicas tradicionais como RL. Tal método estatístico pode ter boa performance como classificador de diversos enfoques agrícolas como os teores de proteína em soja (CAMARGO *et al.*, 2023).

A utilização de todos os comprimentos de onda proporcionou melhores desempenhos para todos os algoritmos testados. Alguns trabalhos comprovam que a utilização de todas as informações espectrais possíveis melhoram o desempenho dos algoritmos quanto a teores industriais do grão utilizando comprimentos de onda da região do visível (SANTANA *et al.*, 2023), teores de macronutrientes primários na folha (SANTANA *et al.*, 2023) ou na identificação de variedades de soja (GAVA *et al.*, 2022). Utilizar os comprimentos de onda é mais prático do ponto de vista processual da análise de dados, diminuído processos nas análises estatísticas (SANTANA *et al.*, 2023) e como pode ser observado no trabalho melhora os resultados, podendo atribuir a maior fornecimento de informações aos algoritmos.

Mais trabalhos com outros genótipos, avaliando mais características industriais de grãos de soja devem ser explorados buscando avaliar as respostas que serão descobertas. A caracterização multi e hiperespectral de genótipos e as respostas industriais de cada genótipo também devem ser exploradas.

## 5 CONCLUSÃO

A utilização de técnicas de aprendizagem de máquinas apresenta respostas inferiores a técnica padrão utilizada no trabalho, ou seja, a regressão logística (RL) apresentou superioridade na classificação dos genótipos de soja quanto a variáveis industriais do grão. A utilização dos comprimentos de onda proporcionou melhor desempenho dos algoritmos na classificação.

## REFERENCIAS

- ANDERSON, E. J.; ALI, M. L.; BEAVIS, W. D.; CHEN, P.; CLEMENTE, T. E.; DIERS, B. W.; GRAEF, G. L.; GRASSINI, P.; HYTEN, D. L.; MCHALE, L. K. Soybean [*Glycine max* (L.) Merr.] breeding: History, improvement, production and future opportunities. In: Al-Khayri, J., Jain, S., Johnson, D. (eds). **Advances in plant breeding strategies: legumes**. [s. l.]: Springer; Cham, 2019. p. 431–516.
- BELGIU, M.; DRĂGUT, L. Random forest in remote sensing: A review of applications and future directions. **ISPRS Journal of Photogrammetry and Remote Sensing**, [s. l.], v. 114, p. 24–31, 2016.
- CHRISTODOULOU, E.; MA, J.; COLLINS, G. S.; STEYERBERG, E. W.; VERBAKEL, J. Y.; VAN CALSTER, B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. **Journal of Clinical Epidemiology**, [s. l.], v. 110, p. 12–22, 2019.
- DA SILVA JUNIOR, C. A.; NANNI, M. R.; SHAKIR, M.; TEODORO, P. E.; DE OLIVEIRA-JÚNIOR, J. F.; CEZAR, E.; DE GOIS, G.; LIMA, M.; WOJCIECHOWSKI, J. C.; SHIRATSUCHI, L. S. Soybean varieties discrimination using non-imaging hyperspectral sensor. **Infrared Physics & Technology**, [s. l.], v. 89, p. 338–350, 2018.
- DE CAMARGO, R. F.; TAVARES, T. R.; DA CRUZ DA SILVA, N. G.; DE ALMEIDA, E.; DE CARVALHO, H. W. P. Soybean sorting based on protein content using X-ray fluorescence spectrometry. **Food Chemistry**, [s. l.], v. 412, p. 135548, 2023.
- DUBOIS, M.; GILLES, K. A.; HAMILTON, J. K.; REBERS, P. A. t; SMITH, F. Colorimetric method for determination of sugars and related substances. **Analytical chemistry**, [s. l.], v. 28, n. 3, p. 350–356, 1956.
- EGMONT-PETERSEN, M.; DE RIDDER, D.; HANDELS, H. Image processing with neural networks—a review. **Pattern Recognition**, [s. l.], v. 35, n. 10, p. 2279–2301, 2002.
- FURBANK, R. T.; TESTER, M. Phenomics—technologies to relieve the phenotyping bottleneck. **Trends in plant science**, [s. l.], v. 16, n. 12, p. 635–644, 2011.
- GAVA, R.; SANTANA, D. C.; COTRIM, M. F.; ROSSI, F. S.; TEODORO, L. P. R.; DA SILVA JUNIOR, C. A.; TEODORO, P. E. Soybean Cultivars Identification Using Remotely Sensed Image and Machine Learning Models. **Sustainability**, [s. l.], v. 14, n. 12, p. 7125, 2022.
- GUO, B.; SUN, L.; JIANG, S.; REN, H.; SUN, R.; WEI, Z.; HONG, H.; LUAN, X.; WANG, J.; WANG, X. Soybean genetic resources contributing to sustainable protein production. **Theoretical and Applied Genetics**, [s. l.], v. 135, n. 11, p. 4095–4121, 2022.

GUPTA, S. K.; MANJAYA, J. G. Advances in improvement of soybean seed composition traits using genetic, genomic and biotechnological approaches. **Euphytica**, [s. l.], v. 218, n. 7, p. 99, 2022.

HAGHIGHATTALAB, A.; GONZÁLEZ PÉREZ, L.; MONDAL, S.; SINGH, D.; SCHINSTOCK, D.; RUTKOSKI, J.; ORTIZ-MONASTERIO, I.; SINGH, R. P.; GOODIN, D.; POLAND, J. Application of unmanned aerial systems for high throughput phenotyping of large wheat breeding nurseries. **Plant Methods**, [s. l.], v. 12, p. 1–15, 2016.

HARTMAN, G. L.; WEST, E. D.; HERMAN, T. K. Crops that feed the World 2. Soybean—worldwide production, use, and constraints caused by pathogens and pests. **Food Security**, [s. l.], v. 3, p. 5–17, 2011.

HOLZMAN, M. E.; RIVAS, R. E.; BAYALA, M. I. Relationship between TIR and NIR-SWIR as indicator of vegetation water availability. **Remote Sensing**, [s. l.], v. 13, n. 17, p. 3371, 2021.

JUNGES, A. H.; ALMANÇA, M. A. K.; FAJARDO, T. V. M.; DUCATI, J. R. Leaf hyperspectral reflectance as a potential tool to detect diseases associated with vineyard decline. **Tropical Plant Pathology**, [s. l.], v. 45, p. 522–533, 2020.

KOC, A.; ODILBEKOV, F.; ALAMRANI, M.; HENRIKSSON, T.; CHAWADE, A. Predicting yellow rust in wheat breeding trials by proximal phenotyping and machine learning. **Plant Methods**, [s. l.], v. 18, n. 1, p. 30, 2022.

LIN, R.; GLAZEBROOK, J.; KATAGIRI, F.; ORF, J. H.; GIBSON, S. I. Identification of differentially expressed genes between developing seeds of different soybean cultivars. **Genomics Data**, [s. l.], v. 6, p. 92–98, 2015.

MOREIRA, M. A. Fundamentos do sensoriamento remoto e metodologia de aplicação. [s. l.], 2001.

NALEPA, J.; KAWULOK, M. Selecting training sets for support vector machines: a review. **Artificial Intelligence Review**, [s. l.], v. 52, n. 2, p. 857–900, 2019.

OLIVEIRA, J. F. De; ALCÂNTARA, J. F. De; SANTANA, D. C.; TEODORO, L. P. R.; BAIO, F. H. R.; CORADI, P. C.; SILVA JUNIOR, C. A. Da; TEODORO, P. E. Spectral variables as criteria for selection of soybean genotypes at different vegetative stages. **Remote Sensing Applications: Society and Environment**, [s. l.], v. 32, p. 101026, 2023.

PRANANTO, J. A.; MINASNY, B.; WEAVER, T. Near infrared (NIR) spectroscopy as a rapid and cost-effective method for nutrient analysis of plant leaf tissues. **Advances in agronomy**, [s. l.], v. 164, p. 1–49, 2020.

PRENGER, E. M.; YATES, J.; MIAN, M. A. R.; BUCKLEY, B.; BOERMA, H. R.; LI, Z. Introgression of a high protein allele into an elite soybean cultivar results in a high-protein near-isogenic line with yield parity. **Crop Science**, [s. l.], v. 59, n. 6, p. 2498–2508, 2019.

QUINLAN, J. R. C4. 5: Programming for machine learning. **Morgan Kauffmann**, [s. l.], v. 38, n. 48, p. 49, 1993.

RAVIKANTH, L.; JAYAS, D. S.; WHITE, N. D. G.; FIELDS, P. G.; SUN, D.-W. Extraction of spectral information from hyperspectral data and application of hyperspectral imaging for food and agricultural products. **Food and bioprocess technology**, [s. l.], v. 10, p. 1–33, 2017.

SANTANA, D. C.; TEODORO, L. P. R.; BAIO, F. H. R.; DOS SANTOS, R. G.; CORADI, P. C.; BIDUSKI, B.; DA SILVA JUNIOR, C. A.; TEODORO, P. E.; SHIRATSUCHI, L. S. Classification of soybean genotypes for industrial traits using UAV multispectral imagery and machine learning. **Remote Sensing Applications: Society and Environment**, [s. l.], v. 29, p. 100919, 2023.

SCOTT, A. J.; KNOTT, M. A cluster analysis method for grouping means in the analysis of variance. **Biometrics**, [s. l.], p. 507–512, 1974.

SILVA, D. J.; QUEIROZ, A. C. De. **Análise de alimentos: métodos químicos e biológicos** UFV. Viçosa: Universidade Federal de Viçosa, 2002.

SINGH, D.; WANG, X.; KUMAR, U.; GAO, L.; NOOR, M.; IMTIAZ, M.; SINGH, R. P.; POLAND, J. High-throughput phenotyping enabled genetic dissection of crop lodging in wheat. **Frontiers in plant science**, [s. l.], v. 10, p. 394, 2019.

SNOUSY, M. B. Al; EL-DEEB, H. M.; BADRAN, K.; KHLIL, I. A. Al. Suite of decision tree-based classification algorithms on cancer gene expression data. **Egyptian Informatics Journal**, [s. l.], v. 12, n. 2, p. 73–82, 2011.

ŠTEPANOVSKEÝ, M.; IBROVÁ, A.; BUK, Z.; VELEMÍNSKÁ, J. Novel age estimation model based on development of permanent teeth compared with classical approach and other modern data mining methods. **Forensic Science International**, [s. l.], v. 279, p. 72–82, 2017.

WHITE, J. W.; ANDRADE-SANCHEZ, P.; GORE, M. A.; BRONSON, K. F.; COFFELT, T. A.; CONLEY, M. M.; FELDMANN, K. A.; FRENCH, A. N.; HEUN, J. T.; HUNSAKER, D. J. Field-based phenomics for plant genetics research. **Field Crops Research**, [s. l.], v. 133, p. 101–112, 2012.

XIE, C.; YANG, C. A review on plant high-throughput phenotyping traits using UAV-based sensors. **Computers and Electronics in Agriculture**, [s. l.], v. 178, p. 105731, 2020.

ZHANG, J.; HAN, M.; WANG, L.; CHEN, M.; CHEN, C.; SHEN, S.; LIU, J.; ZHANG, C.; SHANG, J.; YAN, X. Study of Genetic Variation in Bermuda Grass along Longitudinal and Latitudinal Gradients Using Spectral Reflectance. **Remote Sensing**, [s. l.], v. 15, n. 4, p. 896, 2023.

ZHOU, Z.; LAKHSSASSI, N.; CULLEN, M. A.; EL BAZ, A.; VUONG, T. D.; NGUYEN, H. T.; MEKSEM, K. Assessment of phenotypic variations and correlation

among seed composition traits in mutagenized soybean populations. **Genes**, [s. l.], v. 10, n. 12, p. 975, 2019.

## CONSIDERAÇÕES FINAIS

No Capítulo 1 os algoritmos de Aprendizado de Máquina que apresentaram uma precisão superior na classificação foram as Redes Neurais Artificiais (ANN), o algoritmo de árvore de decisão REPTree e a Máquina de Vetor Suporte (SVM). Adicionalmente, foram realizados testes com três configurações de entrada distintas, a fim de avaliar quais informações espectrais proporcionam o maior nível de precisão. Os resultados mais promissores foram observados quando se utilizaram exclusivamente as bandas espectrais.

Os algoritmos Máquina de Vetor Suporte (SVM) e árvore de decisão J48 no capítulo 2 demonstram uma notável superioridade em termos de acurácia na classificação das características industriais e na produtividade dos grãos. É importante ressaltar que a utilização das bandas espectrais especificamente selecionadas no âmbito deste estudo contribui significativamente para o aumento da acurácia desses algoritmos na classificação.

No capítulo 3 a aplicação de abordagens de aprendizado de máquina demonstrou desempenho inferior em comparação com a técnica convencional adotada neste estudo, que foi a regressão logística (RL), revelando sua superioridade na classificação dos genótipos de soja em relação às variáveis industriais do grão. Vale destacar que a utilização dos comprimentos de onda como atributos de entrada resultou em um desempenho aprimorado dos algoritmos na tarefa de classificação.

## **APÊNDICE A - Acessando o melhor modelo de cada capítulo**

1. No Weka explorer é necessário carregar o arquivo. arff contendo os valores espectrais e as separações dos cluster respectivamente.
2. Abra a guia “Classify” e clique e selecione Conjunto de teste fornecido na opção Opções de teste. Em seguida, clique com o botão direito do mouse na parte em branco da Lista de Resultados. Entre as opções, clique em Carregar Modelo e escolha o arquivo disponível no link: [https://drive.google.com/drive/folders/1tV5WZul0i8WgkEKdFm2bRwPNttSbzp3W?usp=drive\\_link](https://drive.google.com/drive/folders/1tV5WZul0i8WgkEKdFm2bRwPNttSbzp3W?usp=drive_link).
3. Clique com o botão direito do mouse no modelo carregado na Lista de Resultados e escolha a opção "Reavaliar modelo no conjunto de testes atual".