UNIVERSIDADE FEDERAL DE MATO GROSSO DO SUL
CÂMPUS DE CHAPADÃO DO SUL
PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA

REGIMAR GARCIA DOS SANTOS

CLASSIFICAÇÃO DE GENÓTIPOS DE SOJA QUANTO AO DESEMPENHO FISIOLÓGICO UTILIZANDO VARIÁVEIS ESPECTRAIS

CHAPADÃO DO SUL – MS
2022

UNIVERSIDADE FEDERAL DE MATO GROSSO DO SUL
CÂMPUS DE CHAPADÃO DO SUL
PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA

REGIMAR GARCIA DOS SANTOS

Classificação de genótipos de soja quanto ao desempenho fisiológico utilizando variáveis espectrais

Orientadora: Profa. Dra. Larissa Pereira Ribeiro Teodoro

Dissertação apresentada à Universidade Federal de Mato Grosso do Sul, como requisito para obtenção do título deMestre em Agronomia, área de concentração: Produção Vegetal.

CHAPADÃO DO SUL – MS
2022

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

PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA

CERTIFICADO DE APROVAÇÃO

DISCENTE: Regimar Garcia dos Santos

ORIENTADORA: Dra. Larissa Pereira Ribeiro Teodoro

TÍTULO: Classificação de genótipos de soja quanto ao desempenho fisiológico utilizando variáveis espectrais

AVALIADORES:

Profa. Dra. Larissa Pereira Ribeiro Teodoro

Prof. Dr. Paulo Eduardo Teodoro

Prof. Dr. Rafael Ferreira Barreto

Chapadão do Sul, 07 de fevereiro de 2022.

Documento assinado eletronicamente por **Paulo Eduardo Teodoro**, **Professor do Magisterio Superior**, em 07/02/2022, às 10:13, conforme horário oficial de Mato Grosso do Sul, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Rafael Ferreira Barreto**, **Professor do Magisterio Superior - Substituto**, em 07/02/2022, às 10:14, conforme horário oficial de Mato Grosso do Sul, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Larissa Pereira Ribeiro**, **Professora do Magistério Superior**, em 07/02/2022, às 10:14, conforme horário oficial de Mato Grosso do Sul, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

RESUMO

Os caracteres fisiológicos da cultura da soja estão correlacionados com a sua produtividade, o que foi comprovado através de estudos de progresso genético. A mensuração desses caracteres demanda equipamentos complexos e mão de obra especializada, além de ser oneroso e demorado. Problemas que podem ser resolvidos utilizando bandas espectrais e índices de vegetação associados ao sensoriamento remoto. O objetivo deste trabalho foi identificar a melhor técnica de aprendizagem de máquina para a classificação dos clusters de populações F_2 de soja com base em suas características fisiológicas utilizando bandas espectrais e índices de vegetação como variáveis de entrada. O experimento foi realizado no ano agrícola de 2019/2020, no município de Chapadão do Sul, MS. Utilizando um drone foram coletados os valores das bandas espectrais (SB) e índices de vegetação (IV) de 194 populações de soja em F_2. As características fisiológicas avaliadas foram: fotossíntese líquida (A), condutância estomática (gs), concentração interna de CO_2 (Ci), transpiração (E) e eficiência no uso da água (EUA). Utilizando a técnica de k-means as amostras foram divididas em dois clusters, e utilizando Componetes Principais (PCA), as amostras foram particionadas em dois grupos com base em seu desempenho fisiológico. Foram avaliadas como variáveis de entrada as SB+IV, apenas SB e apenas IV. Utilizando SB+IV a técnica com melhor capacidade de classificação foi a de redes neurais artificiais (ANN) com 66,34% de classificações corretas (CC), o algoritmo J48 apresentou o melhor resultado utilizando apenas SB (69,87% CC) e Regressão Logística (RL) obteve melhor resposta quando utilizando apenas IV's como variável de entrada (68,95% CC). Os resultados obtidos demonstram que a melhor forma de se fazer a classificação é utilizando apenas as SB como variáveis de entrada no algoritmo J48, diminuindo o tempo e as chances de erro durante a etapa de cálculo dos IV's.

Palavras-chave: Aprendizagem de Máquina. Sensoriamento remoto. Fisiologia.

ABSTRACT

The physiological characters of the soybean crop are correlated with its productivity, which has been proven through studies of genetic progress. The measurement of these characters requires complex equipment, specialized labor, in addition to being costly and time-consuming. Problems that can be solved using spectral bands and vegetation indices associated with remote sensing. The objective of this work was to identify the best machine learning technique for classifying F2 soybean population clusters based on their physiological characteristics using spectral bands and vegetation indices as input variables. The experiment was carried out in the 2019/2020 agricultural year, in the municipality of Chapadão do Sul, MS. Using a drone, the values of spectral bands (SB) and vegetation indices (IV) were collected from 194 soybean populations in F2. The physiological characteristics evaluated were: net photosynthesis (A), stomatal conductance (gs), internal concentration of CO2 (Ci), transpiration (E) and water use efficiency (USA). Using the k-means technique the samples were divided into two clusters, using Principal Components (PCA), the samples were partitioned into two groups based on their physiological behavior. SB+IV, SB only and IV only were evaluated as input variables. Using SB+IV, the technique with the highest classification capacity was the artificial neural networks (ANN) with 66.34% of correct classifications (CC), the J48 algorithm presented the best result using only SB (69.87% CC) and Logistic Regression (RL) obtained better response when using only IV's as input variable (68.95% CC). The results obtained demonstrate that the best way to make the classification is using only the SB as input variables in the J48 algorithm, reducing the time required and the chances of error during the step of calculating the vegetation indices.

Keywords: Machine Learning. Remote sensing. Physiology.

INTRODUÇÃO

A soja [*Glycina max* (L) Merril] é a principal commoditie produzida no Brasil, que é o maior produtor mundial com uma produção de 135,409 milhões de toneladas na safra 20/21 (CONAB, 2021). A produção aumentou de forma significativa nas últimas décadas. Isso se deve não apenas ao aumento da área plantada, mas principalmente a utilização de novas tecnologias como a transgenia, manejo de pragas e doenças, mecanização, melhoramento genético e técnicas de manejo avançadas (HIRAKURI; LAZZAROTTO, 2014).

As características fisiológicas das plantas de soja tem alta correlação com sua capacidade produtiva, sendo alvo de programas de melhoramento genético que priorizam a seleção indireta por meio de boas taxas fotossintéticas, de respiração, teor de clorofila entre outras que podem incrementar o progresso genético para a produtividade da soja no futuro (TODESCHINI et al., 2019). A correlação entre as características fisiológicas e a produtividade das plantas de soja leva a crer que a busca de melhora das taxas fisiológicas pode servir de incremento em produtividade (AINSWORTH et al., 2012; KARYAWATI et al., 2015).

Para se utilizar utilizar os dados fisiológicos é necessário antes fazer a captação e tratamento desses dados. Mensurar esses caracteres é um desafio, já que os métodos disponíveis para a tarefa são de difícil acesso e demandam mão de obra especializada, por se tratar de aparelhos complexos que exigem diversos passos para calibração e utilização (CAPELIN et al., 2017), além de se tratar de uma operação onerosa e que demanda tempo. Outro fator que pode trazer ainda mais limitações é o clima, para se fazer a aquisição de dados fisiológicos com acurácia é necessário que as avaliações sejam conduzidas em dias com baixa nebulosidade, baixa umidade relativa do ar e em baixa altitude (50-100 m), onde a captação é menos afetada pelos fatores atmosféricos (XUE; SU, 2017). Essas limitações podem ser resolvidas utilizando bandas espectrais e índices de vegetação.

O monitoramento da cultura da soja com base no sensoriamento remoto (SR) por meio de imagens obtidas por Veículos Aéreos Não Tripulados (VANT) ou por satélites constituem um ganho para a agricultura por ser um recurso complementar cada vez mais utilizado no manejo do sistema produtivo da soja (ALVES et al., 2014; GAO et al., 2017). Os IV's têm diversas aplicabilidades na agricultura, com potencial para ser utilizados em diversos estudos e monitoramento da vegetação, como o acompanhamento de ataque de pragas e doenças fúngicas e bacterianas, atividade fotossintética da cultura, monitorar a resposta ao déficit hídrico e a senescência (LEIVAS et al., 2014; ZHANG et al., 2018).

Os sensores captam os valores de refletância de bandas espectrais, atribuindo valores numéricos aos comprimentos de onda, posteriormente essas informações são correlacionadas aos dados obtidos nas avaliações à campo como taxa de transpiração, eficiencia do usa da água e taxa fotossintética por exemplo, gerando um banco de dados contendo os valores das leituras das bandas espectrais e os caracteres fisiológicos observados para a mesma área. As bandas Vermelho, Verde, Borda do Vermelho e Near Infrared (NIR) correspondem a alguns dos intervalos do espectro eletromagnético usualmente utilizados nessas correlações. Os valores atribuídos à bandas espectrais quando combinados podem proporcionar uma melhor avaliação das propriedades de refletância. Utilizando modelos matemáticos é possível obter novosvalores, sendo definidos como Índices de Vegetação (IV) (SHIRATSUCHI et al., 2014).

As técnicas de sensoriamento remoto utilizando as bandas espectrais e índices de vegetação aplicadas ao aprendizado de máquinas tem sido alvo de vários estudos, que comprovaram sua eficiência na tarefa de predição de vários caracteres agronômicos importantes da cultura da soja (DA SILVA et al., 2020; TEODORO et al., 2021). Existem alguns estudos que utilizaram com sucesso os índices de vegetação para fenotipagem de alta precisão de caracteres agronômicos na cultura da soja (SANTANA et al., 2022). Contudo, ainda não existem trabalhos utilizando variáveis espectrais para fenotipagem de caracteres fisiológicos.

A hipótese do estudo é verificar a possibilidade de classificar genótipos de soja em grupos com diferentes características fisiológicas utilizando bandas espectrais e índices de vegetação como variáveis independentes. O objetivo foi identificar a melhor técnica de aprendizagem de máquina para a classificação dos clusters de populações F_2 de soja com base em suas características fisiológicas utilizando bandas espectrais e índices de vegetação como variáveis de entrada nos modelos utilizados.

METODOLOGIA

Condução do experimento

O experimento foi conduzido na safra 2019/2020 na Universidade Federal de Mato Grosso do Sul, localizada no município de Chapadão do Sul (18°41’33’’S, 52°40’45’’W, com 810 m de altitude), Mato Grosso do Sul (Figura 1). O clima da região é caracterizado como Tropical de Savana (Aw) de acordo com a classificação de Koppen e o solo foi classificado como Latossolo Vermelho Distrófico argiloso , com as seguintes características na camada de 0 - 0,20 m: pH (H_2O) = 6,2; Al trocável ($cmol_c\ dm^{-3}$) = 0,0; Ca+Mg ($cmol_c\ dm^{-3}$) = 4,31; P (mg

dm^{-3}) = 41,3; K ($cmol_c\ dm^{-3}$) = 0,2; Matéria orgânica (g dm^{-3}) = 19,74; V (%) = 45; m (%) = 0,0; Soma de bases ($cmol_c\ dm^{-3}$) = 2,3; CTC ($cmol_c\ dm^{-3}$) = 5,1 (TEIXEIRA et al., 2017).

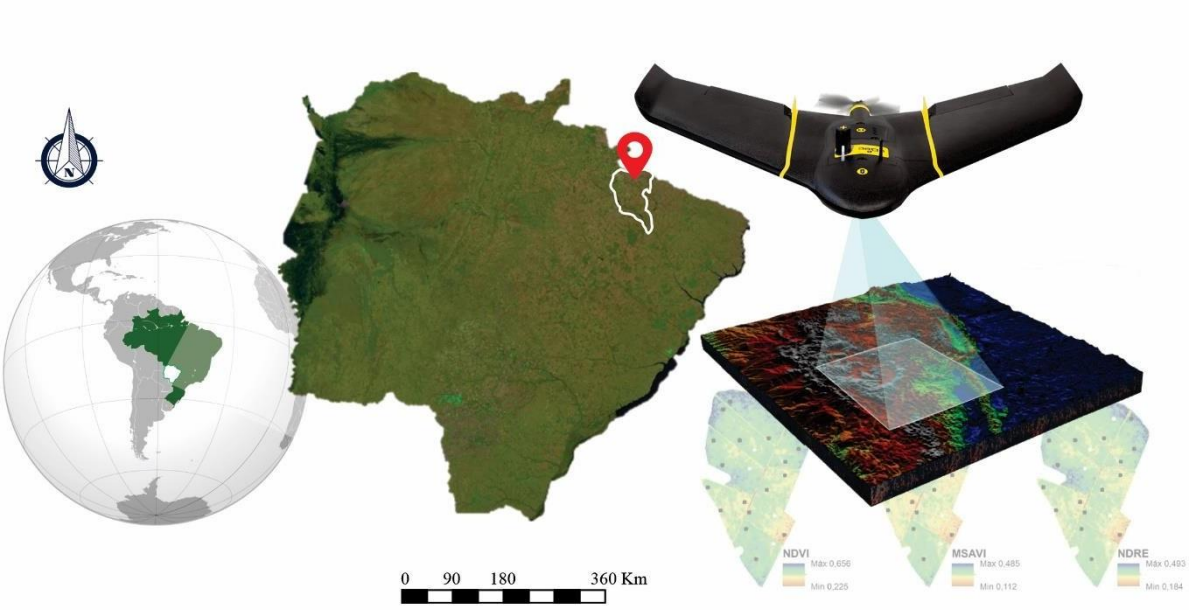

Figura 1. Localização da área de estudo em Chapadão do Sul, Mato Grosso do Sul, Brasil.

Foi utilizado delineamento de blocos casualizados com duas repetições e 194 populações F_2 de soja. As parcelas consistiram de uma linha com três metros de comprimento, com espaçamento de 0,45 m entre fileiras e densidade de 15 plantas por metro. A semeadura ocorreu em outubro de 2019, utilizando preparo convencional do solo (aração e gradagem niveladora). A Figura 2 mostra as condições meteorológicas no decorrer do experimento.

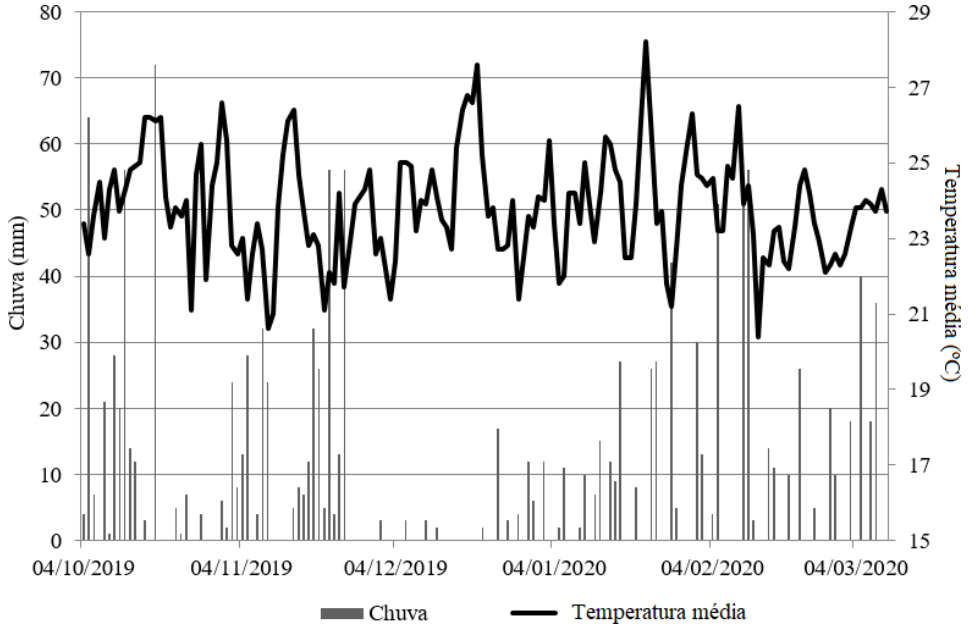

Figura 2. Condições meteorológicas no decorrer da safra 2019/2020.

As sementes foram tratadas com fungicida (Piraclotrobina + 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. Para a fixação biológica de nitrogênio (FBN), as sementes foram inoculadas com bactérias do gênero Bradyrhizobium, 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.

Aquisição dos dados espectrais

Aos 60 dias após a emergência (DAE), foi utilizada a aeronave remotamente pilotada (RPA) de asa fixa Sensefly eBee RTK, com controle autônomo de decolagem, plano de vôo e aterrissagem. O vôo foi realizado às 09:00, em dia sem nebulosidade, a uma altitude de 100 m, permitindo uma resolução espacial de 0,10 m. A calibração radiométrica foi realizada para toda a cena com base em uma superfície refletiva calibrada fornecida pelo frabricante. O eBee foi equipado com o sensor multiespectral Parrot Sequoia. O sensor multiespectral Sequoia faz aquisição da reflectância nos bandas espectrais no comprimento do verde (green, 550 nm ± 40 nm), vermelho (red, 660 nm ± 40 nm), borda do vermelho (rededge, 735 nm ± 40 nm) e infravermelho próximo-(NIR,790 nm ± 40 nm), possuindo um sensor de luminosidade que permite a calibração dos valores adquiridos. As informações obtidas nesses bandas espectrais permitem o cálculo dos diversos mapas dos índices de vegetação a serem utilizados nos algoritmos computacionais.

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. Os índices de vegetação calculados foram: NDVI (Normalized Difference Vegetation Index), NDRE (Normalized Difference Red Edge Index), GNDVI (Green Normalized Difference Vegetation Index), SAVI (Soil Adjusted Vegetation Index), MSAVI (Índice de Vegetação Modificado Ajustado pela Presença de Solo), MCARI (Índice de Refletância de Absorção de Clorofila Modificado) e EVI (Índice de Vegetação Aumentada). Os índices de vegetação estudados estão contidos na Tabela 1.

Tabela 1. Relação dos índices de vegetação calculados pelo uso do sensor multiespectral Sequoia.

Sigla	Índice de Vegetação	Equação
NDVI	Índice de vegetação pela diferença normalizada	$\frac{(R_{NIR} - R_{RED})}{(R_{NIR} + R_{RED})}$
NDRE	Índice de vegetação pela diferença normalizada à transição do vermelho (*Red-edge*)	$\frac{(R_{NIR} - R_{EDGE})}{(R_{NIR} + R_{EDGE})}$
GNDVI	Índice de vegetação pela diferença normalizada ao verde	$\frac{(R_{NIR} - R_{GREEN})}{(R_{NIR} + R_{GREEN})}$
SAVI	Índice de vegetação ajustado pela presença de solo	$\frac{(1 + \mathrm{L})(R_{NIR} - R_{RED})}{(R_{NIR} + R_{RED} + \mathrm{L})}$
MSAVI	Índice de vegetação modificado ajustado pela presença de solo	$\frac{(\ _{NIR} - R\ )}{(\mathrm{L} + \ _{NIR} + R)}(1 + \mathrm{L})$
MCARI	Ìndice de refletância de absorção de clorofila modificado	$[(\mathrm{R700} - \mathrm{R670}) - 0.2(\mathrm{R700} - \mathrm{R550})](\mathrm{R700}/\mathrm{R670})$
EVI	Ìndice de vegetação aumentada	$2{,}5 * \frac{(R_{NIR} - R_{RED})}{((R_{NIR}) + (C1 * R_{NIR}) - (\mathrm{C2} * R_{BLUE}) + \mathrm{L})}$
SCCCI	Ìndice simplificado de conteúdo de clorofila do dossel	$\frac{\mathrm{NDVI}}{\mathrm{NDRE})}$

R_{NIR}: reflectância na faixa do infravermelho próximo; R_{GREEN}: reflectância na faixa do verde; R_{RED}: reflectância na faixa do vermelho; R_{EDGE}: reflectância na faixa de transição do vermelho (Red-edge); L: fator de correção do efeito do solo.

Avaliação dos caracteres fisiológicos

Aos 60 DAE, foram avaliadas as características fisiológicas em cinco plantas por parcela usando um analisador de fotossíntese portátil (Infrared Gas Analyzer - IRGA) modelo Li-6400XT (LiCor Inc., Lincoln, Nebraska, EUA), seguindo os procedimentos descritos em

Teodoro et al. (2019). Foi utilizado fluxo de fótons fotossinteticamente ativos de 1044 μmol m^{-2} s^{-1} e concentrações de CO_2 no ambiente de 372 ± 10 mol m^{-2} s^{-1}. As características fisiológicas medidas foram: fotossíntese líquida (A, μmmol CO_2 m^{-2} s^{-1}), condutância estomática (gs, mol m^{-2} s^{-1}), concentração interna de CO_2 (Ci, μmmol m^{-2} s^{-1}) e transpiração (E, mmol H_2O m^{-2} s^{-1}). A eficiência no uso da água (WUE μmol CO_2/mol H_2O) foi calculado pela razão entre A e E, de acordo com a Equação 1.

$$WUE = \frac{A}{E} \quad (1)$$

As medições foram realizadas entre as 8:00 e 11:00 horas nas plantas amostradas aleatoriamente em cada unidade experimental. As medidas foram realizadas em um dia sem nuvens, com temperaturas entre 26,0 e 26,5ºC e umidade relativa entre 50 e 80%.

Modelos de aprendizagem de máquina

Os modelos testados foram: máquina de vetor suporte (SVM), redes neurais artificiais (ANN), algoritmos de árvores de decisão J48 e RepTree, e random forest (RF). A técnica convencional de regressão logística (LR) foi utilizada como modelo controle. O SVM executa tarefas de predição construindo hiperplanos em um espaço multidimensional para distinguir as classes (RAJVANSHI; CHOWDHARY, 2017). A ANN testada consiste em uma camada oculta única formada por um número de neurônios que é igual ao número de atributos, mais o número de classes, todos divididos por 2 (EGMONT-PETERSEN; DE RIDDER; HANDELS, 2002). O modelo J48 é uma adaptação do classificador C4.5 que pode ser usado em problemas de regressão com uma etapa de poda adicional baseada em uma estratégia de redução de erro (SNOUSY et al., 2011). O REPTree utiliza a lógica da árvore de decisão e cria várias árvores em diferentes iterações. Em seguida, seleciona a melhor árvore utilizando o ganho de informação e realiza a poda de redução do erro como critério de divisão (KALMEGH, 2015). O modelo RF é capaz de produzir várias árvores de predição para o mesmo conjunto de dados e usar um esquema de votação entre todas essas árvores aprendidas para predizer novos valores (BELGIU; DRĂGU, 2016). As análises de aprendizagem de máquina foram realizadas com software Weka 3.9.4. utilizando a configuração padrão para todos os modelos testados (BOUCKAERT et al., 2010), em uma CPU Intel® CoreTM i5 com 8 Gb de RAM.

Análises estatísticas

Para gerar os grupos de genótipos a serem classificados, os dados foram submetidos a análise de componentes principais (PCA). Foi construído um biplot com os dois primeiros componentes principais devido a facilidade de interpretação destes resultados. Neste biplot, foram definidos dois clusters (1 e 2) com base no desempenho dos genótipos para as variáveis fisiológicas com utilização do algoritmo k-means, que agrupa tratamentos cujos centroides estão mais próximos até que não ocorra variação significativa na distância mínima de cada observação a cada um dos centroides. Essas análises foram realizadas com auxílio do pacote "ggfortify" do aplicativo gratuito R (BUNN., 2008). Para as análises de aprendizagem de máquina, os dois clusters formados foram utilizados como variáveis de saída (output) dos modelos, enquanto que as variáveis espectrais red, green e rededge foram utilizadas como entrada dos modelos. A classificação dos clusters foi realizada pelos seis modelos de ML em uma validação cruzada estratificada com k-fold = 10 e dez repetições (100 execuções para cada modelo).

As correlações de Pearson (r) entre os caracteres avaliados foram estimadas utilizando a Equação 2.

$$r = \frac{COV_{(xy)}}{\sqrt{\delta_x^2 * \delta_y^2}} \qquad (2)$$

Em que: $COV_{(xy)}$ é a covariância entre os caracteres X e Y, δ_x^2 é a covariância do caractere X e δ_y^2 é a variância do caractere Y. Um gráfico de correlações e dispersão foi utilizado para expressar as correlações entre as variáveis fisiológicas e os clusters formados, além das correlações entre as variáveis fisiológicas utilizadas. Uma rede de correlações foi utilizada para expressar graficamente a relação funcional entre as variáveis bandas espectrais, índices de vegetação e caracteres fisiológicos, onde a proximidade entre os nós (linhas) são proporcionais ao valor absoluto da correlação entre esses nós. A espessura das arestas foi controlada pela utilização de um valor de corte igual a 0,6, significando que apenas $|r| \geq 0{,}60$ tiveram suas linhas destacadas, sendo as correlações positivas em verde e as correlações negativas representadas em vermelho. As análises foram realizadas com os softwares livres Rbio (BHERING, 2017), e R utilizando o pacote GGEBiplotGUI.

Para avaliar o desempenho de modelos de classificadores, foram utilizadas as seguintes métricas: porcentagem de classificações corretas, estatística Kappa e F- Measure. Essas métricas fazem uso da matriz de confusão, que indica a classificação correta ou

incorreta das classes em uso, agrupando os resultados em quatro classes, sendo elas: Falso Negativo (FN), Falso Positivo (FP), Verdadeiro Positivo (VP) e Verdadeiro Negativo (VN).

Para medir a taxa de acerto foi utilizada F-score e estatística Kappa que é extraída da matriz de confusão gerada na etapa de validação do modelo; o coeficiente de concordância Kappa (k) utiliza todos os elementos da matriz de confusão em seu cálculo (Equação 2). Valores de k menores que 0 significa nenhuma concordância e próximos de 1 significa concordância perfeita, onde n é o número total de amostras e m o número total de classes; xii é o valor da diagonal da matriz de confusão na linha e coluna i; xi+ é a soma da linha i e x+i é a soma da coluna i da matriz.

$$K = \frac{n \sum_{i=1}^{m} x_{ii} - \sum_{i=1}^{m} x_{i+} x_{+i}}{n^2 - \sum_{i=1}^{m} x_{i+} x_{+i}} \quad (3)$$

F-measure (F-score ou Score F1) é uma medida da precisão de um teste, considera tanto a precisão quanto a recordação do teste para calcular a pontuação. O Score F1 pode ser interpretado como uma média harmônica ponderada da precisão e da recordação, em que uma pontuação F1 atinge seu melhor valor em 1 e a pior pontuação em 0. A precisão, também chamada de valor preditivo positivo, é a proporção de resultados positivos que são verdadeiramente positivos. Recall, também chamado de sensibilidade, é a capacidade de um teste para identificar corretamente os resultados positivos para obter a taxa positiva verdadeira (CORNELISSEN, 2019). A Equação (3), demostra como essa métrica é calculada:

$$F1 = 2 * \frac{\text{Precisão} * \text{Recall}}{\text{Precisão} + \text{Recall}} \quad (4)$$

Para avaliação do desempenho dos modelos de aprendizagem de máquina, boxplots para as médias de Porcentagem Classificações Corretas, Estatística Kappa e F-measure foram apresentados juntamente com os resultados do teste Scott-Knott (SCOTT; KNOTT, 1974), a um nível de 5% de significância. Os boxplots foram executados no software R utilizando os pacotes “ggplot2” e “ExpDes.pt”.

RESULTADOS E DISCUSSÃO

Formação e caracterização dos grupos de genótipos quanto ao desempenho fisiológico

Utilizando o algoritmo K-means o conjunto de dados com os genótipos e seus dados fisiológicos foi dividido em dois clusters (C1 e C2), onde os valores foram agrupados com base nas variáveis fisiológicas fotossíntese liquida (A), condutância estomática (gs), concentração interna de CO_2 (Ci), transpiração (E) e eficiência do uso da água (EUA). Simultaneamente, para determinar a qual cluster as amostras pertencem foi aplicada a técnica de componentes

principais (PCA). O número de componentes principais utilizados é determinado pelo acúmulo de 70% ou mais da proporção de variância total (REGAZZI, A.J. 2000). Para o conjunto de dados avaliados foram formados dois grupos de PCA (Figura 3), onde as amostras tem comportamento fisiológico semelhante entre si em um grupo, porém diferente do grupo oposto.

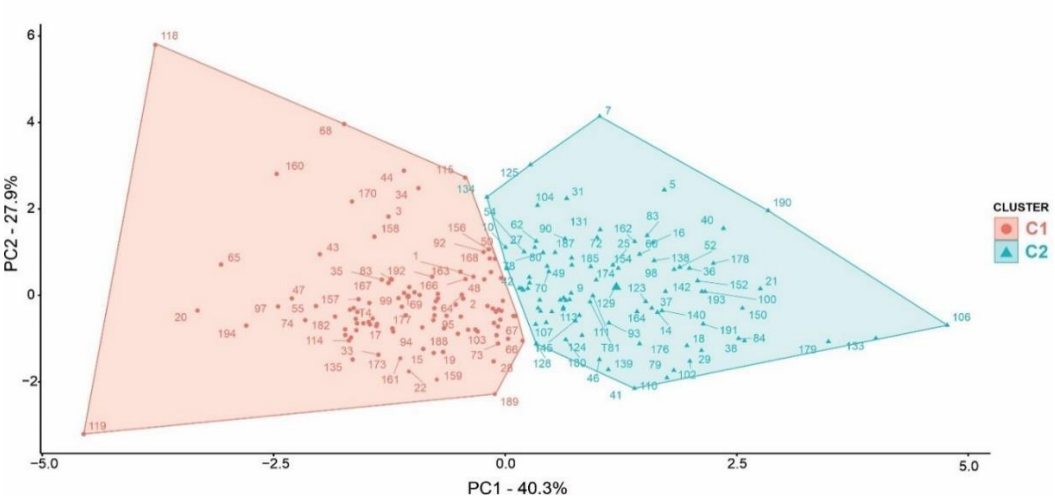

Figura 3.Análise de Componentes Principais (PC) para os clusters formados pelo algoritmo k-means. Cada ponto representa a média de um genótipo avaliado.

A Figura 4 apresenta o gráfico de correlação e dispersão entre as variáveis fisiológicas e os clusters formados pelo k-means. É possível observar que o cluster 1 (C1) apresentou maiores valores de correlação para condutância estomática, concentração interna de CO_2 e transpiração. O cluster 2 (C2) apresentou desempenho superior para as variáveis fotossíntese liquida e eficiência do uso da água.

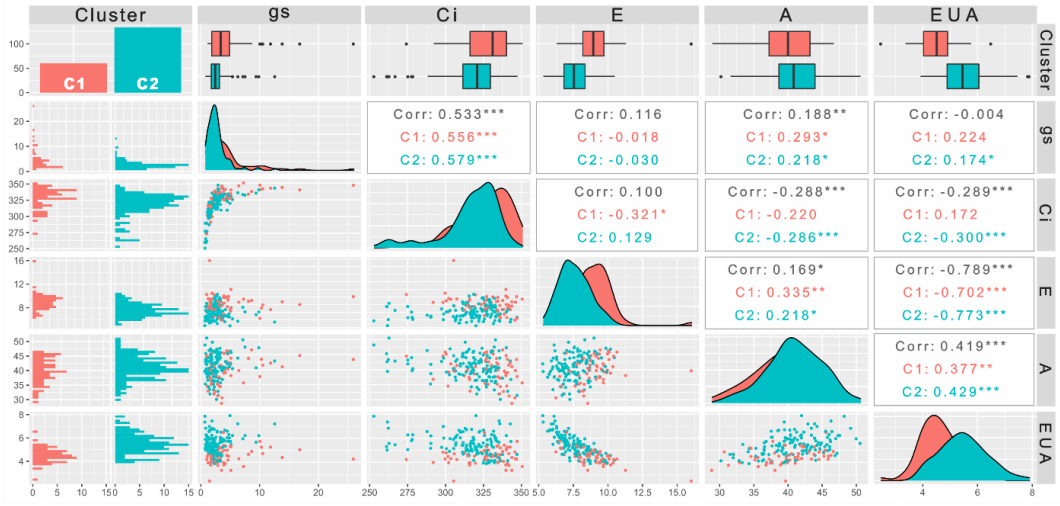

Figura 4. Gráfico de correlações e dispersão entre os clusters e as variáveis Fotossíntese liquida (A), Condutância estomática (gs), Concentração interna de CO_2 (Ci) e Transpiração (E).

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

A Figura 5 apresenta a rede de correlações de Pearson, as linhas verdes representam correlações positivas e as vermelhas apresentam correlações negativas. A espessura das linhas representa a magnitude da correlação entre as variáveis ligadas por elas, as mais espessas representam correlações superiores a 0.6. Houve correlação significativa entre a gs e Ci, resultado que corrobora com Tatagiba; Pezzopane; Reis (2015). Esses autores afirmam que com a redução dos níveis de condutância estomática, o influxo de CO_2 para o interior dos cloroplastos sofre decréscimo, o que contribui ainda para a redução nas taxas fotossintéticas. Esse resultado pode ser observado na Figura 4 pela correlação entre as variáveis condutância estomática e fotossíntese líquida, que apesar do valor de pequena magnitude tem uma interação significativa.

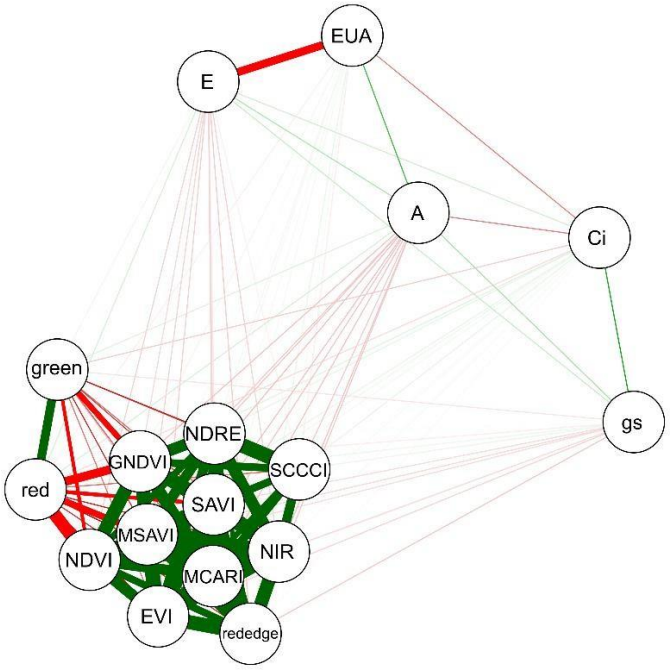

Figura 5. Rede de correlação de Pearson entre as variáveis Fotossíntese líquida (A), Condutância estomática (gs), Concentração interna de CO_2 (Ci), Transpiração (E), as bandas espectrais (green, red, rededge e NIR) e os índices de vegetação NDVI, NDRE, GNDVI, SAVI, MSAVI, MCARI e EVI, avaliados em 194 genótipos de soja na safra 2019/2020.

A variável EUA apresentou interação de baixa magnitude (inferior a 6.0) porém significativa com A, Ci e E, com correlação negativa para concentração de CO_2 e transpiração, inferindo que quanto menor a eficiência no uso da água, menor será o Ci e maior será a taxa de transpiração pela planta. Segundo Taiz e Zeiger (2013), uma baixa eficiência no uso da água leva a redução da fotossíntese por causar o fechamento dos estômatos (diminuição da condutância estomática), diminuindo a eficiência do processo de fixação do carbono pelas plantas, enquanto diminui a perda de água por transpiração, o que pode ser observado pela correlação significativa entre as variáveis EUA, Ci e A.

Acurácia dos modelos de aprendizagem de máquina para classificação dos genótipos de soja

Na Figura 6 estão apresentados o percentual de classificações corretas (CC), estatística Kappa e F-measure para a classificação das diferentes técnicas de aprendizagem de máquina (ML), utilizando como variáveis independentes as bandas espectrais green, red, rededge, NIR e os índices de vegetação avaliados (NDVI, SAVI, GNDVI, NDRE, SCCCI, EVI, MSAVI e MCARI). A técnica de Máquina de Vetor de Suporte (SVM) apresentou a maior média de percentual de classificações corretas com 68,90% de acerto, não diferindo estatisticamente do algoritmo de árvore de decisão J48 (68,17%), Regressão Logística (L) (67,37%) e Redes Neurais Artificiais (ANN) (66,34%). Porém, a estatística Kappa mostra baixa concordância na classificação da SVM, o que também ocorre com o F-measure para esta ML que teve o pior resultado entre as técnicas testadas (0,01), indicando baixa confiabilidade do método na classificação do conjunto de dados.

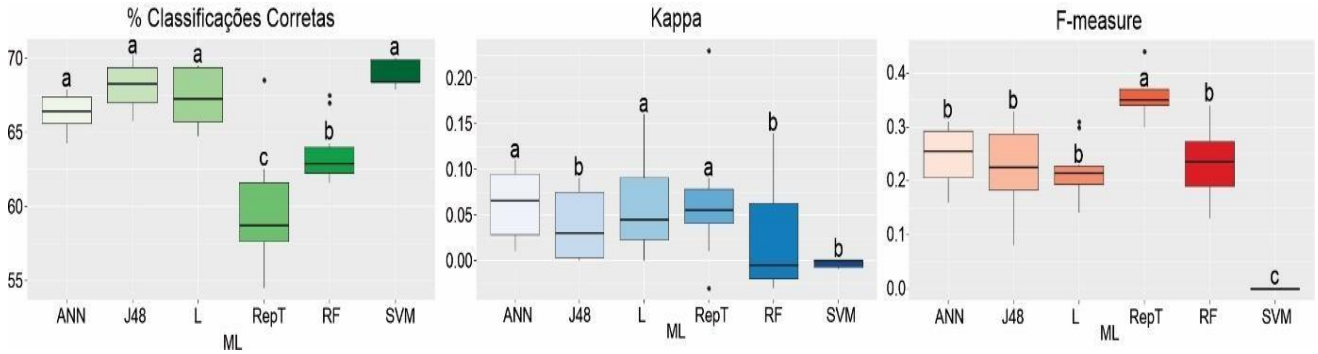

Figura 6. Boxplot para médias de classificações corretas, Kappa e F-measure considerando a análise de classificação logística (L) e técnicas de aprendizagem de máquina: redes neurais (ANN), algoritmos de árvores de decisão J48 e REPTree (RepT), random forest (RF) e máquina de vetor de suporte (SVM). Como variáveis de entrada foram utilizadas as bandas espectrais (green, red, rededge e NIR) e índices de vegetação (NDVI, SAVI, GNDVI, NDRE, SCCCI, EVI, MSAVI e MCARI). Grupos de médias com letras minúsculas iguais não diferem entre si pelo teste de Scott e Knott a 5% de significância.

RF apresentou o segundo pior percentual de classificações corretas além de baixo índice Kappa e F-measure, o que indica baixa capacidade de classificação do modelo para estas variáveis. Outros trabalhos compararam a eficiência dos classificadores RF e SVM utilizando imagens hiperespectrais para classificação (ABDEL-RAHMAN et al., 2014), o estudo destacou que ambos apresentaram resultados semelhantes e satisfatórios. Além disso, RF é considerado um algoritmo adequado para tarefas de classificação, com a vantagem de não apresentar sobre-ajuste dos dados de entrada (BELGIU; DRĂGU, 2016), porém os dois modelos não foram eficientes na classificação das amostras para esse conjunto de dados.

O algoritmo de árvore de decisão RepT apesar de ter apresentado a menor média para CC (59,83%), demonstrou concordância na tarefa de classificação segundo a estatística Kappa e obteve a melhor média para F-measure (0.354), diferindo estatisticamente dos demaismodelos testados para essa métrica, indicando confiabilidade na classificação do modelo. A técnica de redes neurais artificiais mostrou um bom percentual de classificações corretas alémde estatística Kappa não diferindo estatisticamente da melhor média e F-measure,

Na Figura 7 estão apresentados o CC, estatística Kappa e F-measure para a classificação das diferentes técnicas ML e RL, utilizando como variáveis independentes as bandas espectrais green, red, rededge e NIR. O algoritmo de árvore de decisão J48 apresentou a maior porcentagem de classificações corretas (69,30%), não diferindo estatisticamente da técnica SVM com 69,05% de acerto. A menor média foi obtida pelo algoritmo RF com 63,36% de acerto, diferindo assim dos demais modelos a 5% de probabilidade pelo teste de Sckott-Knott.

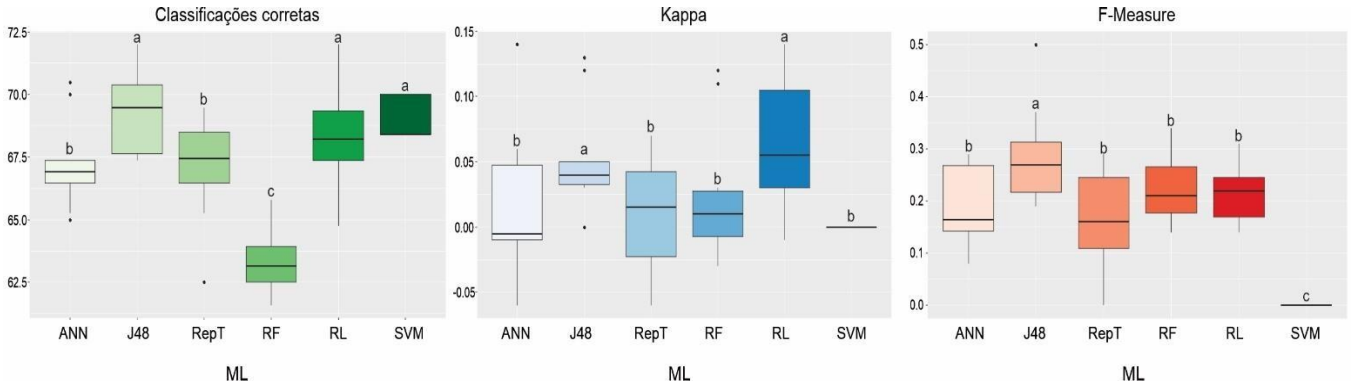

Figura 7. Boxplot para médias de classificações corretas, Kappa e F-measure considerando a análise de classificação regressão logística (RL) e técnicas de aprendizagem de máquina: redes neurais (ANN), algoritmos de árvores de decisão J48 e REPTree (RepT), random forest (RF) e máquina de vetor de suporte (SVM).Como variáveis de entrada foram utilizadas as bandas espectrais green, red e rededge. Grupos de médias com letras minúsculas iguais não diferem entre si pelo teste de Scott e Knott a 5% de significância.

A técnica J48 apresentou ainda o segundo melhor resultado na estatística Kappa com média de 0,053, indicando leve concordância na tarefa de classificação do modelo (LANDIS; KOCH, 1977). Esta análise apresentou ainda o melhor valor para F-measure (0,285), diferindo estatisticamente dos demais modelos. O algoritmo RL obteve resultado estatisticamente semelhante ao obtido pela ANN, RepT e RF para o F-Measure. A SVM apesar de apresentar uma boa CC, obteve o pior resultado entre as técnicas testadas para as métricas estatística Kappa e F-measure, indicando baixa confiabilidade do método na classificação do conjunto de dados.

O algoritmo árvore de decisão J48 foi o modelo que apresentou bons resultados em todas métricas testadas, sendo a técnica com maior poder de classificação utilizando as bandas espectrais como variáveis independentes, ficando à frente de outros algoritmos de árvore de decisão, resultado atípico se tratando de modelos semelhantes entre si. VIEIRA et al., (2012) afirmam que o algoritmo J48 é eficiente na classificação pois elimina a subjetividade, o que diminui a possibilidade de ocorrerem erros humanos na interpretação dos dados. Concluiram ainda que esta técnica é capaz de classificar diferentes usos do solo com um menor número de variáveis independentes.

Foram testados a capacidade de classificação das técnicas de ML utilizando os Índices de Vegetação como variáveis independentes. A Figura 8 mostra boxplots com os valores médios de CC, estatística Kappa e F-measure para a classificação do conjunto de dados. Os maiores valores médios para porcentagem de classificações corretas foram encontrados pelas técnicas SVM (68,95%), RL (68,71%) e J48 (67,93), não havendo diferença estatística entre os três modelos. O pior resultado foi encontrado para o algoritmo de árvore de decisão RF com 65,91% de classificações corretas, não diferindo estatisticamente das técnicas de ANN e RepTree com 67,02 e 67,14% respectivamente.

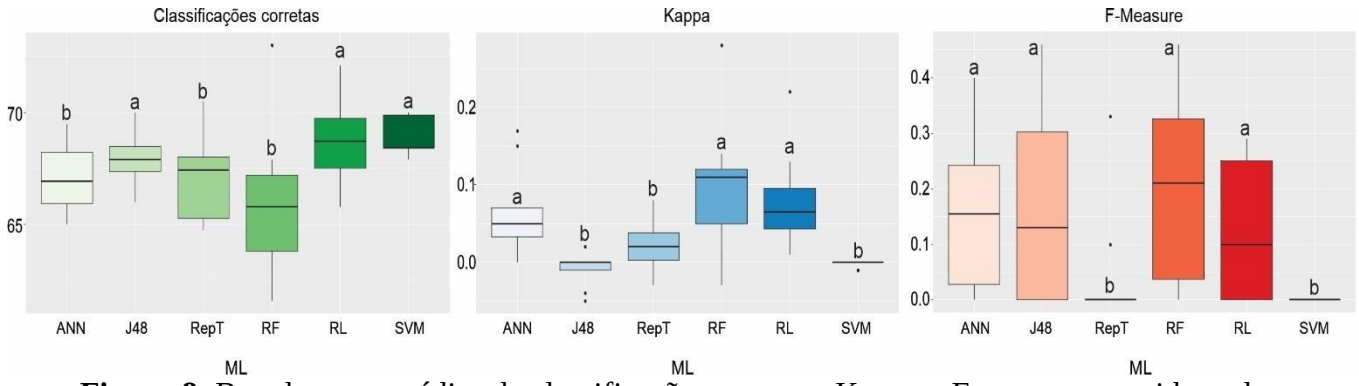

Figura 8. Boxplot para médias de classificações corretas, Kappa e F-measure considerando a análise de classificação regressão logística (RL) e técnicas de aprendizagem de máquina: redes neurais (ANN), algoritmos de árvores de decisão J48 e REPTree (RepT), random forest (RF) e

máquina de vetor de suporte (SVM). Como variáveis de entrada foram utilizadas os índices de vegetação (NIR, NDVI, SAVI, GNDVI, NDRE, SCCCI, EVI, MSAVI e MCARI). Grupos de médias com letras minúsculas iguais não diferem entre si pelo teste de Scott e Knott a 5% de significância.

A análise de classificação regressão logística apresentou a segunda melhor média para classificações corretas e manteve o bom desempenho tanto para o Kappa quanto para F-measure, se mostrando o melhor modelo quando se utiliza os índices de vegetação como variáveis de entrada. A técnica de SVM obteve resultado inferior em todas as as métricas nas análises realizadas quando comparada aos demais modelos testados para esse conjunto de dados. Resultado oposto ao encontrado por Ferreira et al. (2017), onde o modelo obteve resultado superior na tarefa de classificação de cobertura de solo quando comparado aos algoritmos C4.5 e Random Forest.

Foram realizadas três diferentes análises utilizando diferentes arranjos das variáveis independentes, sendo a primeira tendo como variáveis as bandas espectrais associadas aos índices de vegetação, o segundo utilizando apenas as bandas espectrais e o terceiro arranjo utilizando apenas os índices de vegetação. Foram encontrados resultados semelhantes, logo é vantajoso utilizar apenas as bandas espectrais para a análise, diminuindo o tempo gasto com o cálculo dos índices de vegetação, além da redução da quantidade de possíveis erros na fase de captação de dados e da diminuição do banco de dados a ser utilizado.

O avanço dessas tecnologias é um dos principais fatores que contribuem para que o Brasil alavanque sua produção. Elas possibilitaram o monitoramento das culturas por meio de imagens fornecidas por sensores que atribuem valores a radiação eletromagnética e correlacionam esses valores com o potencial produtivo e fenotipagem de cultivares de soja (SANTANA et al., 2022), esses dados ajudam a compreender a relação entre o comportamento da vegetação e as mudanças nas condições ambientais (JOHNSON et al., 2016). O SR permite que sejam coletadas informações de objetos na superfície terrestre sem a necessidade de contato físico com o sensor (SHIRATSUCHI et al., 2014), o que torna a operação mais rápida e com menor custo embutido.

CONCLUSÕES

Foram avaliados três variáveis de entrada e diferentes algoritmos de aprendizagem de máquina para a classificação de genótipos de soja quanto ao desempenho fisiológico. A melhor configuração de entrada depende do algoritmo a ser utilizado.

Para a análise utilizando como variável de entrada as bandas espectrais juntamente com os IV's, a melhor técnica foi a de ANN, enquanto que ao utilizar apenas os bandas espectrais para classificação, o melhor resultado foi obtido pelo modelo J48. Quando foram utilizados apenas os IV's para classificação, o melhor resultado foi encontrado pela regressão logística.

REFERÊNCIAS

AINSWORTH, Elizabeth A. et al. Accelerating yield potential in soybean: Potential targets for biotechnological improvement. **Plant, Cell and Environment**, v. 35, n. 1, p. 38–52, 2012.

ALVES, Claudia Durand et al. MAPPING LAND USE AND LAND COVER CHANGES IN A REGION OF SUGARCANE EXPANSION USING TM AND MODIS DATA Mapeamento de Mudanças de Uso e Cobertura da Terra em uma Região de Expansão de Cana-de-açúcar Utilizando TM e MODIS. v. 2003, n. January, p. 337–347, 2014.

ANDY BUNN, Mikko Korpela. An Introduction to dplR. **Industrial and Commercial Training**, v. 10, n. 1, p. 11–18, 2008.

BELGIU, Mariana; DRĂGU, Lucian. Random forest in remote sensing: A review of applications and future directions. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 114, p. 24–31, 2016.

BHERING, Leonardo Lopes. Rbio: A tool for biometric and statistical analysis using the R platform. **Crop Breeding and Applied Biotechnology**, v. 17, p. 187–190, 2017.

BOUCKAERT, Remco R et al. **WEKA-Experiences with a Java Open-Source Project Eibe Frank Mark A. Hall**. **Journal of Machine Learning Research**. [S.l: s.n.], 2010. Disponível em: <http://www.cs.waikato.ac.nz/ml/weka.>.

CAPELIN, Diogo et al. Utilização do IRGA - Analisador de gases por infravermelho para avaliação de trocas gasosas em plantas: check list de preparação. **O Estresse das Plantas - Teoria & Prática**, n. January, p. 187–194, 2017.

CORNELISSEN, Willy Garabini. Detecção automática do início de notas usando o aprendizado de máquina. p. 1–10, 2019.

DA SILVA, Eder Eujácio et al. UAV-multispectral and vegetation indices in soybean grain yield prediction based on in situ observation. **Remote Sensing Applications: Society and**

Environment, v. 18, n. April, 2020.

DOS SANTOS FERREIRA, Alessandro et al. Weed detection in soybean crops using ConvNets. **Computers and Electronics in Agriculture**, v. 143, n. July, p. 314–324, 2017. Disponível em: <http://dx.doi.org/10.1016/j.compag.2017.10.027>.

EGMONT-PETERSEN, M.; DE RIDDER, D.; HANDELS, H. Image processing with neural networks- A review. **Pattern Recognition**, v. 35, n. 10, p. 2279–2301, 2002.

GAO, Feng et al. Toward mapping crop progress at field scales through fusion of Landsat and MODIS imagery. **Remote Sensing of Environment**, v. 188, p. 9–25, 2017.

KALMEGH, Sushilkumar. Analysis of WEKA Data Mining Algorithm REPTree , Simple Cart and RandomTree for Classification of Indian News. **International Journal of Innovative Science, Engineering & Technology**, v. 2, n. 2, p. 438–446, 2015.

KARYAWATI, Anna Satyana et al. Combining Ability Analysis for Physiological Characters of Soybean (Glycine max L . Merrill). **International Journal of Plant Research**, v. 5, n. 5, p. 113–121, 2015.

LEIVAS, Janice Freitas et al. AVALIAÇÃO DO ÍNDICE DE VEGETAÇÃO PADRONIZADO NO monitoramento indicativa de estiagens em períodos críticos da soja no sul do brasil. **Revista Brasileira de Cartografia**, v. 66, n. 5, p. 1145–1155, 2014.

NITIN RAJVANSHI; K. R. CHOWDHARY,. Comparison of SVM and Naïve Bayes Text Classification Algorithms using WEKA. **International Journal of Engineering Research and**, v. V6, n. 09, 2017.

PAULO CÉSAR TEIXEIRA, GUILHERME KANGUSSU DONAGEMMA, ADEMIR FONTANA, Wenceslau Geraldes Teixeira. **Micromorfologia do solo**. [S.l: s.n.], 2017. Disponível em: <https://ainfo.cnptia.embrapa.br/digital/bitstream/item/194786/1/Pt-5-Cap-1-Micromorfologia-do-solo.pdf>.

REGAZZI, A.J. 2000 - Google Acadêmico. Disponível em: <https://scholar.google.com.br/scholar?start=10&q=REGAZZI,+A.J.+2000&hl=pt-BR&as_sdt=0,5&scioq=REGAZZI,+A.J.+Análise+multivariada,+notas+de+aula+INF+766,+Departamento+de+Informática+da+Universidade+Federal+de+Viçosa,+v.2,+2000.#d=gs_cit&u=%2Fscholar%3Fq%3D>. Acesso em: 16 jan. 2022.

SANTANA, Dthenifer Cordeiro et al. High - throughput phenotyping allows the selection of soybean genotypes for earliness and high grain yield. **Plant Methods**, p. 1–11, 2022. Disponível em: <https://doi.org/10.1186/s13007-022-00848-4>.

SCOTT, Author 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**, v. 30, n. 3, p. 507–512, 1974. Disponível em: <http://www.jstor.org/stable/2529204>.

SHIRATSUCHI, Luciano Shozo et al. Agricultura de precisão - resultados de um novo olhar. **Embrapa Instrumentação**, p. 58–73, 2014.

SNOUSY, Mohmad Badr Al et al. Suite of decision tree-based classification algorithms on cancer gene expression data. **Egyptian Informatics Journal**, v. 12, n. 2, p. 73–82, 2011. Disponível em: <http://dx.doi.org/10.1016/j.eij.2011.04.003>.

TATAGIBA, Sandro Dan; PEZZOPANE, José Eduardo Macedo; REIS, Edvaldo Fialho. FOTOSSÍNTESE EM EUCALYPTUS SOB DIFERENTES CONDIÇÕES EDAFOCLIMÁTICAS - DOI: 10.13083/1414-3984/reveng.v23n4p336-345. **Revista Engenharia na Agricultura - REVENG**, v. 23, n. 4, p. 336–345, 28 ago. 2015. Disponível em: <https://periodicos.ufv.br/reveng/article/view/518>. Acesso em: 26 jan. 2022.

TEODORO, Paulo Eduardo et al. Predicting days to maturity, plant height, and grain yield in soybean: A machine and deep learning approach using multispectral data. **Remote Sensing**, v. 13, n. 22, 2021.

TODESCHINI, Matheus Henrique et al. Soybean genetic progress in South Brazil: physiological, phenological and agronomic traits. **Euphytica**, v. 215, n. 7, 2019.

VIEIRA, Matheus Alves et al. 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**, v. 123, p. 553–562, 2012. Disponível em: <http://dx.doi.org/10.1016/j.rse.2012.04.011>.

XUE, Jinru; SU, Baofeng. Significant remote sensing vegetation indices: A review of developments and applications. **Journal of Sensors**, v. 2017, 2017.

ZHANG, Wenmin et al. Impacts of the seasonal distribution of rainfall on vegetation productivity across the Sahel. **Biogeosciences**, v. 15, n. 1, p. 319–330, 15 jan. 2018. Disponível em: <https://bg.copernicus.org/articles/15/319/2018/>. Acesso em: 22 jun. 2021.