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

ENIO ANTONIO MANFROI FILHO

APRENDIZAGEM DE MÁQUINA PARA PREDIÇÃO DE VARIÁVEIS DENDROMÉTRICAS EM ESPÉCIES NATIVAS POR VARIÁVEIS HIPERESPECTRAIS

CHAPADÃO DO SUL – MS
2023

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

ENIO ANTONIO MANFROI FILHO

APRENDIZAGEM DE MÁQUINA PARA PREDIÇÃO DE VARIÁVEIS DENDROMÉTRICAS EM ESPÉCIES NATIVAS POR VARIÁVEIS HIPERESPECTRAIS

Orientador: Prof. Dr. Gileno Brito de Azevedo

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

CHAPADÃO DO SUL – MS
2023

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

PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA

CERTIFICADO DE APROVAÇÃO

DISCENTE: Enio Antonio Manfroi Filho

ORIENTADOR: Dr. Gileno Brito de Azevedo

TÍTULO: Modelos de aprendizagem de máquina para predição de variáveis dendrométricas em espécies nativas usando variáveis hiperespectrais.

AVALIADORES:

Prof. Dr. Gileno Brito de Azevedo

Prof. Dr. Fabio Henrique Rojo Baio

Profa. Dra. Glauce Tais de Oliveira Sousa Azevedo

Chapadão do Sul, 27 de setembro de 2023.

Documento assinado eletronicamente por **Gileno Brito de Azevedo**, **Professor do Magisterio Superior**, em 24/10/2023, às 09:24, conforme horário oficial de Mato Grosso do Sul, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Fabio Henrique Rojo Baio**, **Professor do Magisterio Superior**, em 24/10/2023, às 09:27, conforme horário oficial de Mato Grosso do Sul, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Glauce Tais de Oliveira Sousa Azevedo**, **Professora do Magistério Superior**, em 24/10/2023, às 10:13, conforme horário oficial de Mato Grosso do Sul, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

AGRADECIMENTOS

Inicialmente agradeço a meu bom e amado Deus - fonte de todo conhecimento. Ele dá sabedoria e entendimento àqueles que o buscam com amor.

A minha amada e resiliente esposa. Pela oportunidade que tenho em ombrear corajosamente a vida a seu lado, na busca constante da verdade e da justiça.

A meu Orientador, Prof. Dr. Gileno Brito de Azevedo, pela oportunidade de trabalharmos juntos, pelos ensinamentos, e, sobretudo, pela disponibilidade e paciência em todos os momentos.

Aos docentes do Programa de Pós-Graduação em Agronomia, pelo apoio e pela disponibilidade que sempre demonstraram durante o curso.

Aos membros da banca examinadora, Profa. Dra. Larissa Pereira Ribeiro Teodoro e Profa. Dra. Glauce Tais de Oliveira Sousa Azevedo e Prof. Dr. Fábio Henrique Rojo Baio pela disponibilidade e contribuição na melhoria do trabalho.

Agradeço também ao Prof. Dr. Paulo Eduardo Teodoro por ter me guiado na conclusão desta etapa em minha caminhada como servidor público. Exemplo profissional de dedicação à pesquisa e ensino.

À Universidade Federal de Mato Grosso do Sul (UFMS) e ao Programa de Pós-Graduação em Agronomia pela oportunidade

.

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

O uso de sensoriamento remoto combinado com técnicas de aprendizagem de máquina (AM) é uma abordagem promissora para estimar o crescimento e a produtividade das árvores. Muitos estudos mostram melhorias na precisão das estimativas quando os modelos de AM são implementados em comparação com os métodos tradicionais. O objetivo deste estudo foi investigar o desempenho de Técnicas de AM e leituras espectrais para prever o diâmetro à altura do peito (DAP) e altura (Ht) de espécies florestais nativas por meio de variáveis hiperespectrais. A área de estudo compreendeu um povoamento florestal misto, no qual 195 árvores foram aleatoriamente amostradas. Para a aquisição dos dados hiperpectrais, foram realizadas leituras das reflectâncias foliares com o equipamento ASD FieldSpec® 4. Os comprimentos de onda obtidos pelas leituras espectrais, cuja faixa variou de 350 a 2500 nm, foram utilizados como variáveis de entrada dos modelos. As técnicas de AM testadas foram, Redes neurais artificiais (ANN), REPTree (DT), Árvore de decisão (M5P), Zero R (R), Floresta aleatória (RF) e Máquina de vetor suporte (SVM). Foram testadas duas configurações de inputs: 1) utilizando apenas os comprimentos de onda na predição das variáveis dendrométricas, sem inclusão da variável qualitativa espécie (SE) e, 2) utilizando os comprimentos de onda em conjunto com a variável qualitativa espécie (CE). O desempenho de cada algoritmo foi verificado usando coeficiente de correlação (r), o erro absoluto médio MAE e raiz quadrada do erro médio RMSE. Houve interação significativa entre inputs e algoritmos de AM para as variáveis r, MAE e RMSE. O algoritmo melhor ranqueado foi DT, tanto para DAP como para Ht em todos os testes. Houve diferença significativa quando utilizada a variável de entrada CE demonstrando melhora nos resultados de predição. Portanto, o estudo demonstra que é possível predizer DAP e Ht com relativa precisão utilizando-se de bandas espectrais como entrada nos modelos de AM testados. Quando incluída a variável qualitativa espécie, os algoritmos DT, M5P e SVM apresentaram melhor desempenho. Quando não incluída essa informação como entrada, verificou-se que o algoritmo RF obteve os melhores resultados devido à sua capacidade de predição e alta estabilidade.

Palavras-chave: Inteligência computacional. Bandas espectrais. Mensuração Florestal. Sensoriamento hiperespectral.

ABSTRACT

The use of remote sensing combined with machine learning (ML) techniques is a promising approach for estimating tree growth and productivity. Many studies show improvements in estimation accuracy when ML models are implemented compared to traditional methods. The objective of this study was to investigate the performance of AM techniques and spectral readings to predict diameter at breast height (DBH) and height (Ht) of native forest species using hyperspectral variables. The study area comprised a mixed forest stand, in which 195 trees were randomly sampled. To acquire hyperspectral data, leaf reflectance readings were carried out with the ASD FieldSpec® 4 equipment. The wavelengths obtained by spectral readings, whose range varied from 350 to 2500 nm, were used as input variables for the models. The ML techniques tested were, Artificial Neural Networks (ANN), REPTree (DT), Decision Tree (M5P), Zero R (R), Random Forest (RF) and Support Vector Machine (SVM). Two input configurations were tested: 1) using only wavelengths in the prediction of dendrometric variables, without including the qualitative variable species (SE) and, 2) using wavelengths together with the qualitative variable species (CE). The performance of each algorithm was checked using correlation coefficient (r), the mean absolute error MAE and root mean square error RMSE. There was a significant interaction between inputs and AM algorithms for the variables r, MAE and RMSE. The best ranked algorithm was DT, both for DAP and Ht in all tests. There was a significant difference when using the input variable CE, demonstrating improvement in prediction results. Therefore, the study demonstrates that it is possible to predict DAP and Ht with relative accuracy using spectral bands as input in the AM models tested. When the qualitative variable species was included, the DT, M5P and SVM algorithms performed better. When this information was not included as input, it was found that the RF algorithm obtained the best results due to its prediction capacity and high stability.

Keywords: Computational intelligence. Spectral bands. Forest Measurement. Hyperspectral sensing.

SUMÁRIO

1. INTRODUÇÃO

A mensuração precisa dos atributos das árvores, como diâmetro de fuste, altura e volume, é essencial para o planejamento florestal, estimativa da produção de madeira e da avaliação do estoque florestal. Além disso, a mensuração florestal fornece informações essenciais para o monitoramento e a avaliação dos impactos ambientais e das mudanças climáticas nas florestas (BREDE et al. 2019)

Ao longo das décadas, a mensuração florestal tem sido reconhecida como uma prática fundamental para compreender e gerenciar efetivamente os recursos florestais. Dentre as várias técnicas utilizadas, a regressão (linear e não linear) tem sido tradicionalmente empregada para modelar a relação entre as variáveis e estimar atributos das árvores. Essa abordagem permite ajustar modelos que podem predizer com precisão características como peso, altura, volume e produção por unidade de área das florestas, tanto no presente quanto no futuro (BREDE et al. 2019; MARQUES RAMOS et al. 2020). Essa análise estatística das relações entre variáveis mensuradas é essencial para embasar decisões de manejo florestal e garantir a sustentabilidade desses ecossistemas (BREDE et al. 2019). Nos métodos tradicionais, os conjuntos de dados são originados de medições realizadas diretamente e anualmente em campo, resultando em um inventário florestal contínuo para estimativa de produtividade. Ainda assim, esta é uma tarefa demorada, cara e altamente exigente que pode ser suportada por novas abordagens como o uso de dados de sensoriamento remoto (PONZONI; HIMABUKURO, 2010)**.**

Nos modelos de regressão, as estimativas dos parâmetros dos modelos lineares normalmente são obtidas pela solução do sistema de equações normais, considerando a minimização da soma do quadrado dos erros (Mínimos Quadrados Ordinários) (CAMPOS; LEITE, 2017). No caso de modelos não lineares, as estimativas são obtidas por algum método de aproximações sucessivas ou técnica de otimização numérica. No entanto, nos últimos anos, métodos mais eficientes relacionados às técnicas de aprendizagem de máquina (AM) ganharam destaque para realização dessa tarefa (SILVA et al. 2019). Ultimamente as análises de regressão, onde muitos dos problemas agrícolas e ambientais estão incluídos, tem recebido grande atenção, tornando-se um ponto de destaque neste campo de AM (HUANG et al. 2020).

Muitos estudos mostram melhorias significativas na precisão das estimativas quando os modelos de AM são implementados em comparação com os métodos tradicionais (ÖZÇELIK et al. 2010). Há estudos que comprovam que o uso desses algoritmos tem sido empregados com sucesso em diversas atividades aplicadas às ciências florestais, como por exemplo na modelagem da produção volumétrica de árvores de eucalipto (SILVA et al. 2009) e da relação

hipsométrica em plantações de eucalipto e pinus (COSTA FILHO et al. 2019; ROCHA et al. 2021), na predição do diâmetro e altura de árvores de espécies de eucalipto a partir de dados hiperespectrais (SILVA et al. 2009; Da SILVA et al. 2021); mudanças espaço-temporais da produtividade primária líquida (NPP) em florestas subtropicais (LI et al. 2022); e na classificação do uso da terra (TSAI et al. 2023).

Entre as técnicas de AM utilizadas, destacam-se o Random Forest (RF), algoritmos de árvore de decisão: Árvore de Decisão (DT), árvore de decisão (M5P); Redes Neurais Artificiais (ANN - Artificial Neural Network), algoritmo Zero R (R), Máquina de Vetor de Suporte (SVM). O algoritmo RF tem a capacidade de gerar múltiplas árvores de classificação para o mesmo conjunto de dados, sendo considerado um dos modelos mais práticos e amplamente utilizados em inferência indutiva (BELGIU; DRĂGU, 2016). O algoritmo REPTree (DT) é uma variação do classificador C4. Ambos os algoritmos são usados para a construção de modelos de classificação baseados em árvores de decisão, que pode ser empregada em problemas de regressão. Ele foi desenvolvido para lidar com algumas limitações do C4.5, especialmente em termos de simplificação e redução de árvores complexas, com a adição de uma etapa de poda baseada em uma estratégia de redução de erros, essa simplificação evita o ajuste excessivo e tornar o modelo fácil de interpretar. A ideia principal é que, durante a construção da árvore de decisão, o modelo pode ficar muito complexo e acabar se ajustando aos dados de treinamento, prejudicando o desempenho em dados (BOUCKAERT, 2010).

Por sua vez, o M5P é um algoritmo que substitui globalmente todos os valores ausentes pela média/moda do atributo e utiliza técnicas de regressão para estimar os valores faltantes. Esses algoritmos baseados em árvores possuem capacidade de lidar com problemas de regressão e apresentou bom desempenho em estudos anteriores (SILVA et al. 2019).

O algoritmo Zero R (R) é considerado um algoritmo de aprendizagem de máquina extremamente simples e básico. Sua principal função é realizar uma predição com base em uma única variável independente, geralmente a variável de saída, ignorando completamente as demais variáveis do conjunto de dados. Mesmo que o Zero R seja extremamente simples e não forneça insights profundos sobre os dados, ele pode ser útil como uma linha de base de comparação para outros algoritmos de AM mais sofisticados.

O algoritmo ANN é um modelo de AM que trabalha de forma similar ao funcionamento do cérebro humano. Ele foi idealizado por constituir de um conjunto interconectado de unidades chamadas de neurônios artificiais ou perceptron, que são unidades de processamento utilizadas no processo de aprendizagem de relações lineares e não-lineares. É uma área de pesquisa ampla e em constante evolução, com contribuições de diversos cientistas ao longo do tempo

(EGMONT-PETERSEN et al. 2002). O algoritmo SVM é um método de AM supervisionado utilizado também para tarefas de predição e classificação, com boa capacidade de lidar com problemas de classificação mais complexos (NALEPA e KAWULOK, 2019).

Estas técnicas de AM, associadas à utilização de sensores hiperespectrais, representam uma ferramenta com grande potencial para uma ampla gama de estudos nesta área. Pesquisas recentes têm demonstrado que essas técnicas proporcionam resultados altamente precisos, em linha com os achados deste estudo. Da Silva et al. (2021) realizaram predições do diâmetro à altura do peito (DAP) e altura total (Ht) de seis espécies de eucalipto utilizando de técnicas de AM associadas ao processamento de índices de vegetação extraídos de imagens multiespectrais baseadas em drone. Já Lim et al. (2020) utilizaram aprendizagem de máquina para classificação de cinco espécies de árvores usando informações espectrais do Sentinel-2 na região Norte Coreana com sucesso. Outro estudo similar é o de Deur et al. (2020), que avaliaram a acurácia de classificação de espécies de árvores em uma floresta decídua mista de várzea em Jastrebarski Lugovi no centro da Croácia associou características espectrais com dois algoritmos de aprendizagem de máquina. Com base nas investigações mencionadas anteriormente, torna-se evidente o potencial de aplicação dessas técnicas. No entanto, é importante destacar que, na revisão da literatura, não foram identificados estudos que tenham empregado dados hiperespectrais para a predição de variáveis dendrométricas em povoamentos florestais mistos compostos por espécies nativas de ambientes tropicais. Assim, a aplicação estratégica e integrada de técnicas de aprendizagem por máquina e sensoriamento remoto podem otimizar o processo para a obtenção de estimativas precisas para diversas variáveis. Por meio do sensoriamento remoto é possível realizar coleta de informações sobre um objeto, área ou fenômeno sem a necessidade de contato direto com o mesmo. Isso é feito por meio de sensores, que podem estar em satélites, aeronaves, drones, balões ou equipamentos terrestres. Esses sensores captam dados a partir da radiação eletromagnética refletida ou emitida pelo objeto ou área de interesse, permitindo a obtenção de informações sobre características físicas, químicas, biológicas e geográficas.

O sensoriamento remoto desempenha uma função de destaque no campo na aprendizagem de máquina ao fornecer informações que podem ser utilizadas como entradas em algoritmos (MAXWELL et al. 2018). Estes dados captados por sensores remotos são digitalizados, podendo ser processados e analisados pelos algoritmos.

Com isso, técnicas de aprendizagem de máquina têm ganhado atenção nas práticas de agricultura de precisão, uma vez que abordam com eficiência múltiplas aplicações, como estimar o crescimento e a produtividade de árvores em plantações florestais (SILVA et al.

2021), no trabalho de classificação de espécies de árvores (MARRS et al. 2019) e identificação de espécies arbóreas urbanas (CETIN et al. 2022). O uso correto dessas técnicas apresentadas, ou seja, frente as tradicionais já utilizadas, podem resultar em uma série de vantagens e auxiliar nos manejos florestais. Os modelos de AM permitem processar grandes conjuntos de dados altamente complexo, contornando a falta de linearidade que existe entre os dados (ROELL et al. 2020).

Dessa forma, o objetivo deste estudo foi avaliar a acurácia e predição de estimativas das variáveis DAP e Ht utilizando dados hiperespectrais por meio de modelos de aprendizagem de máquina. Os objetivos específicos foram: 1) avaliar quais os melhores modelos de aprendizagem de máquina para a predição das variáveis DAP e Ht a partir de dados hiperespectrais; 2) verificar se a inclusão da variável de entrada espécie contribui para a melhoria da acurácia das estimativas de DAP e Ht a partir de dados hiperespectrais.

2. MATERIAL E MÉTODOS

A área de estudo é compreendida por um povoamento florestal misto, que se caracteriza por uma área de floresta ou floresta plantada composta por diferentes espécies de árvores ou plantas em vez de uma monocultura, esta misturada pode ocorrer intencionalmente ou naturalmente, neste caso, a implantação foi realizada por meio de plantio. A área possui 4,8 ha (área de reposição florestal), com 9,7 anos de idade e caracterizado pelo plantio misto de espécies florestais nativas da flora brasileira, distribuídos de forma aleatória e com variações na densidade de indivíduos de cada espécie. O povoamento foi implantado em março de 2013 e está localizado no município de Chapadão do Sul-MS a 8°46′32"S e 52°36′59"W, com altitude de aproximadamente 793 m. O clima local de acordo com a classificação de Köppen é do tipo tropical úmido (Aw) com duas estações bem definidas, uma chuvosa no verão e outra seca no inverno. A precipitação média varia de 750 a 1.800 mm ano^{-1} e a temperatura média anual varia de 20 a 25ºC. A vegetação predominante na região é tipicamente savana (PAGOTTO e SOUZA, 2006). O solo predominante na área é o Latossolo Distrófico (EMBRAPA, 2018).

Para a coleta dos dados, realizou-se um caminhamento aleatório na área, onde foram amostrados aleatoriamente 195 indivíduos pertencentes a 19 espécies florestais (Tabela 1). A quantidade de indivíduos amostrados é diferente entre as espécies, sendo selecionados os indivíduos de forma aleatória, conforme a sua distribuição e frequência na área de estudo. A

grafia dos nomes científicos foi verificada de acordo a base de dados Flora do Brasil (FLORA DO BRASIL, 2022).

Tabela 1. Número de indivíduos amostrados em cada uma das espécies encontradas na área de estudo.

Nome Científico	ESP	Família Botânica	n
Cedrela fissilis Vell.	Ced	Meliaceae	33
Guazuma ulmifolia Lam.	Gua	Malvaceae	29
Enterolobium contortisiliquum (Vell.) Morong	Ent	Fabaceae	18
Peltophorum dubium (Spreng.) Taub.	Pel	Fabaceae	16
Anadenanthera colubrina (Brenan.) Brenan	Anc	Fabaceae	15
Senegalia polyphylla (DC.) Britton	Sen	Fabaceae	15
Luehea divaricata Mart.	Lue	Malvaceae	13
Ceiba speciosa (A.St.-Hil., A.Juss. & Cambess.) Ravenna	Cei	Malvaceae	11
Schinus molle L.	Scm	Anacardiaceae	8
Parapiptadenia rigida (Benth.) Brenan	Par	Fabaceae	8
Schinus terebinthifolia Raddi	Sct	Anacardiaceae	7
Gallesia integrifolia (Spreng.) Harms	Gal	Petiveriaceae	6
Bauhinia forficata Link	Bau	Fabaceae	5
Bixa orellana L.	Bix	Bixaceae	3
Croton floribundus Spreng.	Cro	Euphorbiaceae	2
Clitoria fairchildiana R.A.Howard	Cli	Fabaceae	2
Handroanthus impetiginosus (Mart. ex DC.) Mattos	Han	Bignoniaceae	2
Anadenanthera peregrina (L.) Speg.	Anp	Fabaceae	1
Piptadenia gonoacantha (Mart.) J.F.Macbr.	Pip	Fabaceae	1
Total			195

Em que: ESP = código de identificação das espécies; n = nº de indivíduos amostrados para cada espécie.

Para cada indivíduo foram mensuradas as variáveis dendrométricas: circunferência do fuste à altura do peito, 1,3 m do nível do solo (CAP, em centímetros) e altura total (Ht, em metros). A CAP foi obtida com auxílio de uma fita métrica e posteriormente foi convertida para diâmetro (DAP) a partir da divisão de seu valor por π. Como critério de inclusão foi adotado

árvores com CAP ≥ 15 cm. No caso das árvores como multifustes, a medida de cada fuste foi obtida separadamente e, posteriormente, foi obtido o DAP equivalente (expressão 1), que representa o diâmetro correspondente a uma área basimétrica proporcional ao conjunto de fustes da árvore. A Ht foi obtida com auxílio de um clinômetro digital, e no caso das árvores multifustes, foi realizada a leitura apenas para o fuste de maior altura. A Figura 1 apresenta a variação de DAP, Ht e número de fustes dos indivíduos amostrados para cada uma das espécies. Ao analisar a figura, torna-se evidente a notável variabilidade dos dados apresentados, além da considerável quantidade de faixas espectrais por indivíduo, ou seja, 2.141. Com isso, observou-se uma ampla diversidade entre os indivíduos e as espécies em relação à Ht, diâmetro e quantidade de fustes. É importante ressaltar que as condições do ambiente foram preservadas, e não se buscou coletar indivíduos com características que fossem mais facilmente convertidas em dados, com o propósito de reproduzir fielmente as condições naturais de povoamento em florestas nativas.

$$DAPe = \sqrt[3]{\sum_{i=1}^{n} DAP_i^2} \quad (1)$$

Em que: DAP*e* = diâmetro equivalente para cada indivíduo amostrado (cm); DAP*i* = diâmetro do fuste a 1,3 m do nível do solo.

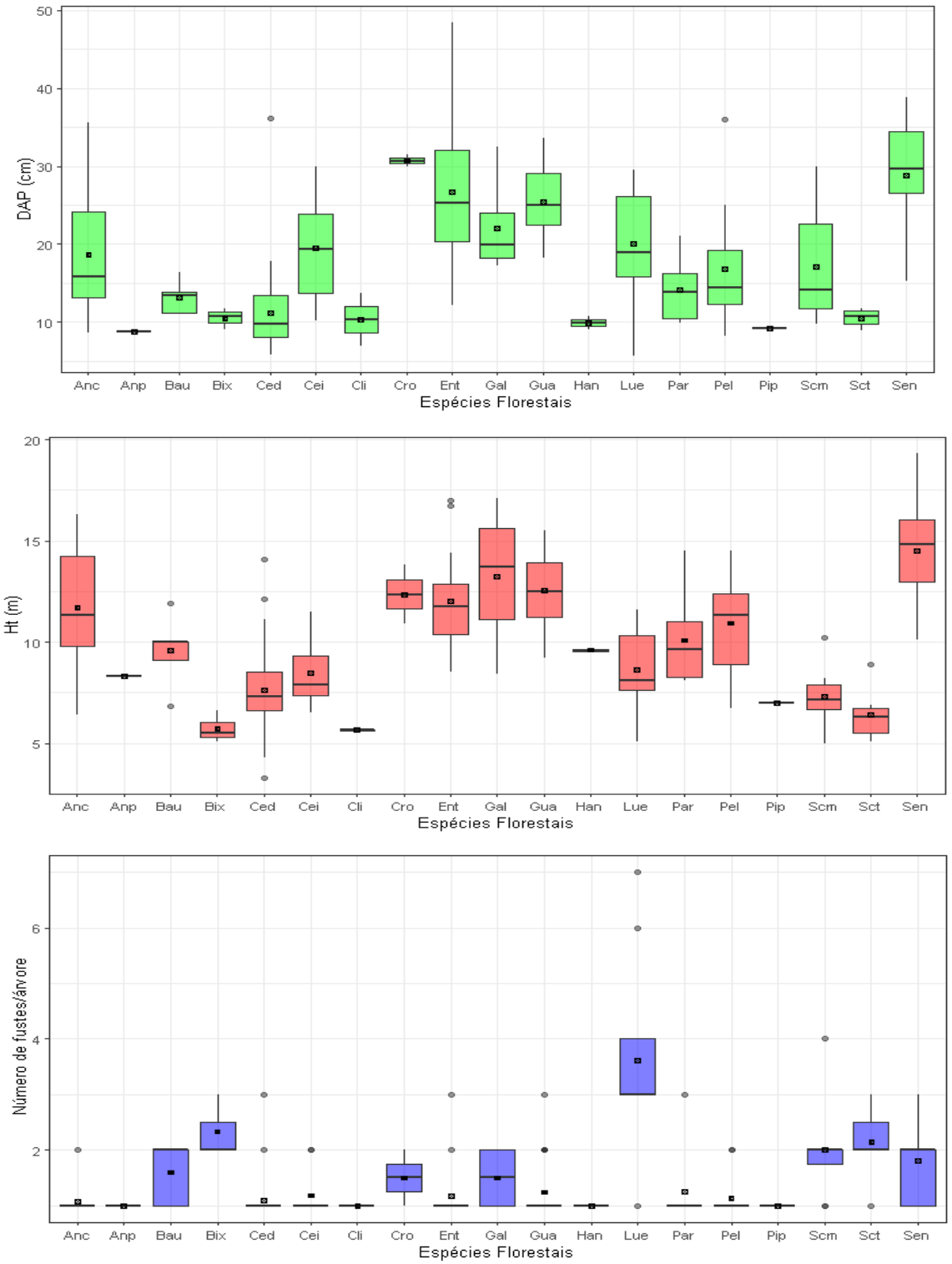

Figura 1. Boxplot para o diâmetro à 1,3 m do nível do solo (DAP), altura total (Ht) e número de fustes dos indivíduos de cada uma das espécies amostradas.

A leitura dos dados hiperespectrais foi realizada em três folhas aleatórias de cada indivíduo amostrado, totalizando 585 folhas analisadas. Os dados foram obtidos com o equipamento ASD FieldSpec® 4 que oferece desempenho espectral de leitura no espectro de

reflectância solar de amplo alcance (350 - 2500 nm). A resolução espectral aprimorada entre (350 a 2500 nm) são largamente utilizadas no sensoriamento remoto na agricultura, pois cobrem uma gama de comprimentos de onda que capturam informações importantes sobre as plantas e o solo. As informações que podem ser coletadas com ampla utilidade na agricultura se dividem por faixas. Entre estes cumprimentos de ondas, os que mais se destacam são o Ultravioleta (UV, 350-400 nm) que apesar de não estar entre as mais utilizadas pode fornecer informações sobre o estado de saúde das plantas e detectar doenças e estresses iniciais. Visível (400-700 nm) nesta faixa encontramos o espectro que é visível a olho nu. A clorofila nas plantas absorve a luz na faixa visível, e a reflexão nessa faixa pode indicar a saúde e a atividade fotossintética das plantas. Infravermelho Próximo (NIR, 700-1400 nm), usado para estimar a quantidade de clorofila nas folhas e a quantidade de água nas plantas. Faixa que pode determinar o estresse hídrico e a saúde geral das plantas. Infravermelho Médio (MIR, 1400-3000 nm), nesta faixa pode-se determinar as características químicas das plantas e do solo, composição de nutrientes e estruturas moleculares. Acima da faixa do infravermelho médio, entra em cena o Infravermelho Térmico (3000 nm a 14.000 nm), desempenhando um papel crucial na avaliação das temperaturas da superfície das plantas e do solo. Embora seja uma ferramenta valiosa para monitorar o estresse térmico, vale destacar que sua eficácia em relação às previsões e estimativas de crescimento, como abordadas neste estudo, seria ser limitada.

O banco de dados hiperespectrais foi composto por 2.151 variáveis espectrais, abrangendo comprimentos de onda que variam de 350 nm a 2500 nm, com incremento de 1 nm para cada amostra foliar. Essas amostras foram obtidas pela média aritmética das três leituras realizadas em cada indivíduo. Foram utilizados para a modelagem das variáveis dendrométricas em função das hiperespectrais a partir de seis algoritmos de aprendizagem de máquina (Tabela 2). Como variável de saída (output) foram utilizadas as variáveis DAP e Ht. Para as variáveis de entrada (input), foram testadas duas configurações: 1) utilizando apenas os comprimentos de onda na predição das variáveis dendrométricas, sem inclusão da variável qualitativa espécie (SE) e, 2) utilizando os comprimentos de onda em conjunto com a variável qualitativa espécie (CE). A predição foi realizada por meio de validação cruzada estratificada com k-fold = 10 e dez repetições (100 execuções para cada modelo). Foi utilizado o algoritmo ZeroR como controle para avaliar o desempenho de outros algoritmos mais sofisticados (FRANK, et al. 2016). 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 predição das variáveis dendrométricas em um povoamento florestal misto.

Sigla	Modelo de aprendizagem de máquinas	Referência
ANN	Redes neurais artificiais	Egmont-Petersen et al. (2002)
DT	REPTree	Snousy et al. (2011)
M5P	Árvore de decisão M5P	Blaifi et al. (2018)
R	Zero R	Quinlan (1993)
RF	Floresta aleatória	Belgiu e Drăguț (2016)
SVM	Máquina de vetor suporte	Nalepa e Kawulok (2019)

Nas avaliações do desempenho dos modelos de predição testados foram utilizadas as métricas: coeficiente de correlação (r) de Pearson; erro absoluto médio (MAE) e; raiz quadrada do erro médio (RMSE) entre valores observados e preditos. Para verificar a significância dos inputs, dos algoritmos testados e a interação entre ambos, foi realizada uma análise de variância. Havendo a presença de significância, foram gerados boxplots com as médias de r, MAE e RMSE, agrupados pelo teste de Scott-Knott ao nível de 5% de probabilidade. O agrupamento das médias e os boxplots foram gerados usando os pacotes ggplot2 e ExpDes.pt do software R.

3. RESULTADOS E DISCUSSÃO

As medidas de precisão das estimativas das variáveis dendrométricas DAP e Ht foram influenciadas pela interação entre os algoritmos de aprendizagem de máquina e as configurações de entrada comprimento de onda, com ou sem a inclusão da variável qualitativa espécie. Está variável representa atributos categóricos ou qualidades que não podem ser quantificadas numericamente. Em contraste com variáveis numéricas, as variáveis qualitativas são caracterizadas por categorias ou rótulos, tornando a tarefa de classificação mais acessível para o algoritmo. Neste estudo, as variáveis qualitativas foram identificadas pelo nome da espécie (Tabela 3).

Tabela 3. Resumo da análise de variância, com o quadrado médio do resíduo para as métricas utilizadas na avaliação da precisão dos algoritmos de aprendizagem de máquinas para predição de variáveis diâmetro à 1,3 m do nível do solo (DAP) e altura de árvores (Ht) em função de dados hiperespectrais em um povoamento florestal misto.

DAP				
FV	GL	r	MAE	RMSE
Algoritmo (A)	5	0,616*	8,943*	11,672*
Input (I)	1	1,091*	25,687*	25,743*
A x I	5	0,193*	4,137*	3,790*
Erro	108	0,004	0,147	0,309
CV (%)		17,84	5,590	6,580
Ht				
FV	GL	r	MAE	RMSE
Algoritmo (A)	5	0,779*	1,193*	1,719*
Input (I)	1	0,879*	2,846*	3,591*
A x I	5	0,142*	0,408*	0,523*
Erro	108	0,003	0,012	0,021
CV (%)		13,5	4,530	4,790

*significativo a 5% de probabilidade pelo teste F; G.L. graus de liberdades; C.V. coeficiente de variação. Em que: r = coeficiente de correlação entre valores observados e estimados; MAE = erro absoluto médio; RMSE = raiz quadrada do erro médio.

O desempenho dos algoritmos de aprendizagem de máquina na predição das variáveis dendrométricas DAP e Ht foi influenciado pelas variáveis de entrada utilizadas em seu treinamento (Figura 2). Não houve um algoritmo superior em todas as situações. De maneira geral, a inclusão da espécie como variável de entrada (CE), em conjunto com os dados hiperespectrais, proporcionou um melhor desempenho nas estimativas das variáveis dendrométricas DAP e Ht. Isso foi observado através dos maiores valores de r e dos menores valores de MAE e RMSE obtidos pelos algoritmos DT, M5P e SVM (Figura 2). Por outro lado, quando a variável espécie não foi incluída entre as variáveis de entrada (SE), as estimativas de DAP e Ht com maior acurácia foram obtidas a partir do algoritmo RF. Entre os algoritmos de melhor desempenho, o RF proporcionou a maior estabilidade quando comparadas as estimativas com e sem a inclusão a variável de entrada espécie (Figura 2). Tanto para o DAP

quanto para Ht, as métricas de precisão avaliadas apresentaram comportamentos semelhantes para esse algoritmo.

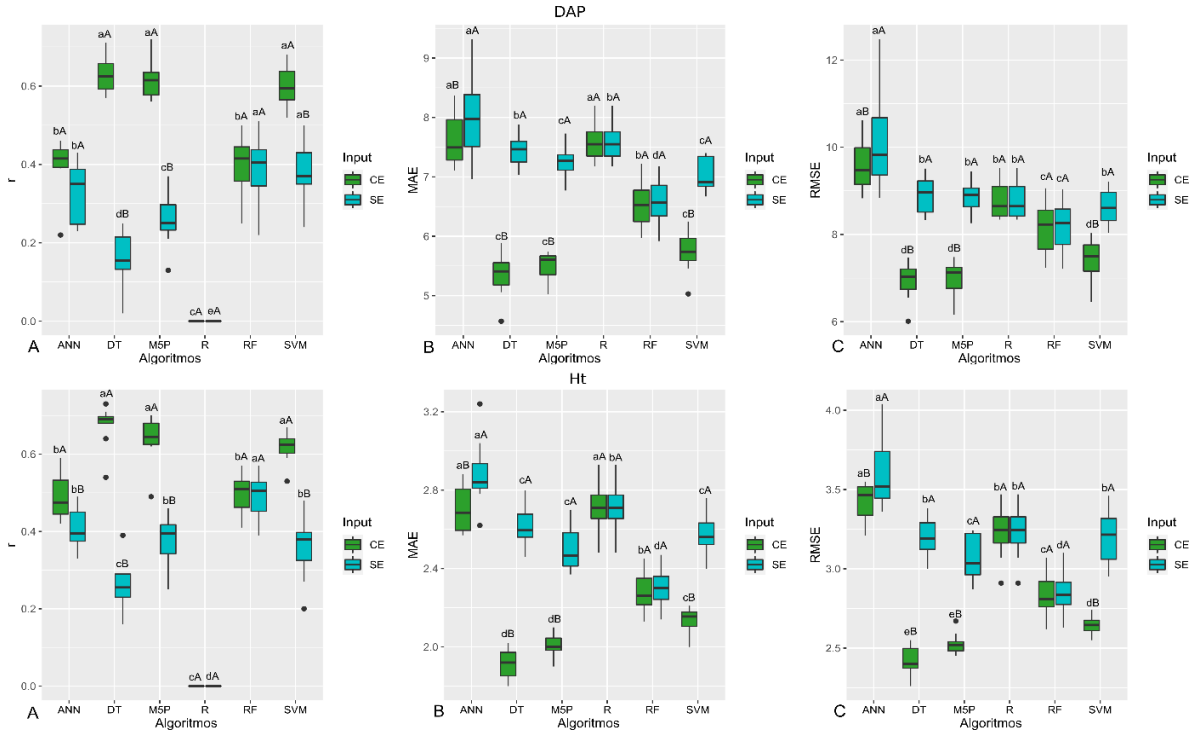

Figura 2. Boxplot para comparação de médias das medidas de precisão para a predição das variáveis diâmetro à altura do peito (DAP) e altura de árvores (Ht) em um povoamento florestal misto. A) coeficiente de correlação (r); B) erro absoluto médio (MAE); C) raiz quadrada do erro médio (RMSE). Em que: ANN (Redes neurais artificiais), DT (REPTree), M5P (Árvore de decisão M5P), R Zero R, RF (Floresta aleatória), SVM (Máquina de vetor suporte). As médias seguidas pelas mesmas letras maiúsculas não diferem entre si para as entradas testadas pelo teste de Scott-Knott a 5% de probabilidade; as médias seguidas pelas mesmas letras minúsculas não diferem entre si para os algoritmos testados pelo teste de Scott-Knott a 5% de probabilidade**.**

Na literatura não foram encontrados estudos que utilizaram dados hiperespectrais para a predição de variáveis dendrométricas em povoamentos florestais mistos compostos por espécies nativas de ambientes tropicais. Contudo, os resultados do presente estudo corroboram com os obtidos por Da Silva et al. (2021). Estes autores realizaram a predição do DAP e Ht de árvores de seis espécies de eucalipto a partir de índices espectrais baseados em veículos aéreos não tripulados (VANTs) e aprendizagem de máquina, encontrando interação entre os modelos de aprendizagem e os inputs avaliados. Estes autores também verificaram que a inclusão das

espécies como variável preditora contribuiu para melhoria da precisão das estimativas obtidas a partir dos modelos de aprendizagem de máquina.

As métricas de acurácia r e MAE obtidas por Da Silva et al. (2021) apresentaram desempenho ligeiramente superior às do presente estudo. Essa diferença pode ser considerada aceitável, pois o estudo anterior utilizou dados de apenas dois gêneros e uma família botânica, enquanto neste estudo foram amostrados indivíduos pertencentes a 19 gêneros e oito famílias botânicas. Essa ampliação da diversidade botânica pode ter contribuído para a maior variabilidade dos dados e, consequentemente, para uma precisão ligeiramente menor nos modelos. Quanto maior a variabilidade natural dos dados, menor tende a ser a precisão obtida nos modelos.

Dessa forma, esses resultados demonstram que a inclusão da espécie como variável de entrada nos modelos de aprendizagem de máquina pode contribuir significativamente para a melhoria da precisão das estimativas das variáveis dendrométricas a partir das variáveis hiperespectrais. Cada espécie de árvore possui características únicas, como crescimento, densidade e forma do tronco, e consequentemente, tem influência significativa nas variáveis dendrométricas, como diâmetro, altura, volume e massa. Assim, ao incluir essa informação como variável de entrada nos algoritmos de predição, é possível capturar padrões específicos de cada espécie, resultando em estimativas mais precisas (Da SILVA et al. 2021). Além disso, a inclusão da variável espécie pode auxiliar no planejamento e manejo florestal, pois com estimativas mais precisas, é possível tomar decisões mais assertivas, como determinar as áreas de colheita, estimar a produtividade por espécie e otimizar a alocação de recursos.

Por outro lado, a utilização da informação de espécie requer dados completos e precisos que identifiquem corretamente cada árvore em relação à sua espécie. Isso pode exigir um esforço adicional na coleta e registro dos dados, especialmente considerando a vasta diversidade de árvores, com mais de 60.000 espécies em todo o mundo e cerca de 10.000 apenas no Brasil (GAISBERGER, 2020). A inclusão da variável espécie também pode aumentar a complexidade da análise, principalmente se houver um grande número de espécies presente na amostra. Portanto, é necessário encontrar um equilíbrio entre essas vantagens e desvantagens, levando em consideração o contexto específico do estudo e as metas do manejo florestal. Com base nos resultados das análises dos algoritmos DT, M5P e SVM, o algoritmo RF apresentou uma abordagem oposta. RF mostrou boa estabilidade e desempenho superior mesmo sem a utilização da informação de espécie. Essa descoberta sugere que o RF pode ser uma escolha adequada em situações em que a inclusão dessa variável não é viável, além de exigir menos esforço adicional.

Os algoritmos de melhor desempenho no presente estudo também se destacaram em outros estudos com aplicações florestais. Da Silva et al. (2021) verificaram que o algoritmo RF proporcionou uma estimativa globalmente superior para todas as configurações testadas nas seis espécies de eucalipto no Brasil. Os autores também verificaram que o algoritmo RBF (Radial Basis Functions) também proporcionou desempenho superior na predição do DAP, superando numericamente o RF tanto em r quanto em MAE, em alguns casos. Já para Ht, os autores observaram que a técnica que obteve o menor MAE foi SVM. Zou et al. (2019) verificaram que o algoritmo RF, seguido de SVM, proporcionou maior precisão nas estimativas do DAP e Ht de oito cultivares de *Cunninghamia lanceolata* (Lamb.) na China, apesar dos autores não terem utilizado variáveis categóricas como variáveis de entrada. Li et al. (2022), ao realizar a estimativa e análise de mudanças espaço-temporais da produtividade primária líquida (NPP) em florestas subtropicais na China, também encontraram melhor desempenho para o algoritmo RF.

No presente trabalho o algoritmo RF destacou-se por apresentar a maior estabilidade nas estimativas de DAP e Ht, com e sem inclusão a variável de entrada espécie. A manutenção de boa capacidade preditiva, mesmo com a retirada de algumas variáveis de entrada, é uma característica importante, pois mantem a capacidade preditiva dos modelos com demanda de menor esforço para a obtenção dos dados a serem utilizados no treinamento dos modelos. A estabilidade apresentada para o algoritmo RF se deve a sua característica de utilizar o método *bootstrap* para extrair várias amostras das amostras originais, conduzindo a modelagem da árvore de decisão para cada amostra *bootstrap* e, em seguida, combina as árvores de decisão para obter o resultado final da previsão por meio de votação. RF prova ser um algoritmo altamente preciso especialmente em problemas de predição utilizando variáveis hiperespectrais como input dos modelos agrícolas (ZOU et al. 2019; RAMOS et al. 2020; SILVA et al. 2020). Um grande número de estudos teóricos e práticos provou que a precisão da previsão de RF é alta e tem boa tolerância para outliers e ruídos (BREIMAN, 2001).

Dessa forma, os resultados do presente estudo evidenciam que a combinação de dados hiperespectrais e algoritmos de aprendizagem de máquina apresenta um potencial significativo para gerar estimativas precisas das variáveis dendrométricas mais comumente mensuradas em inventários florestais, DAP e a Ht. Essas variáveis desempenham um papel fundamental na avaliação do crescimento e produtividade das florestas, sendo essenciais para o planejamento e tomada de decisões em manejo florestal. No entanto, a obtenção desses dados geralmente requer uma grande quantidade de esforço e mão de obra, pois envolve a medição em unidades de amostra distribuídas em toda a área de interesse, seguindo um esquema de amostragem.

Portanto, o uso de abordagens baseadas em dados hiperespectrais e algoritmos de aprendizagem de máquina pode proporcionar uma alternativa eficiente e econômica para a obtenção dessas informações dendrométricas essenciais.

Além disso, uma das descobertas significativas deste estudo é o papel da variável espécie no treinamento de algoritmos de aprendizagem de máquina para a predição das variáveis dendrométricas. No entanto, devido à vasta diversidade de espécies florestais, a identificação precisa das espécies ainda representa um desafio que requer profissionais experientes nesse reconhecimento. Nesse contexto, a coleta de dados *in loco* desempenha um papel crucial na construção de bibliotecas espectrais de espécies, possibilitando uma caracterização mais precisa de suas assinaturas espectrais. Essas informações podem ser posteriormente utilizadas como variáveis de entrada no treinamento de algoritmos de aprendizagem de máquina, visando o reconhecimento das espécies. Esses algoritmos podem ser integrados a sistemas que envolvem o uso de sensoriamento remoto, ampliando as aplicações e benefícios dessa abordagem. Essa abordagem integrada pode otimizar a coleta de dados e a análise remota, permitindo obter estimativas precisas das variáveis dendrométricas a um menor custo e, consequentemente, impulsionar avanços no manejo florestal, fornecendo uma base sólida para a tomada de decisões embasadas em dados precisos e contribuindo para a conservação e sustentabilidade das florestas.

Em geral, todos os algoritmos de AM obtiveram acurácia satisfatória e observou-se que a inclusão da variável espécie como input foi determinante para obtenção de maiores valores de r e menores valores MAE e RMS para os algoritmos DT, M5P e SVM.

Esses resultados representam uma confirmação científica significativa para os programas de inventário florestal, demonstrando a viabilidade da predição precisa das espécies de florestas nativas utilizando variáveis hiperespectrais e técnicas de AM.

4. CONCLUSÕES

O estudo demonstra que é possível predizer DAP e Ht com relativa precisão usando bandas espectrais como entrada nos modelos de AM testados.

Ao comparar os algoritmos sem incluir a informação de espécie como entrada, verificou-se que o algoritmo RF obteve os melhores resultados devido à sua capacidade de predição e alta estabilidade.

REFERÊNCIAS

BELGIU, M.; DRĂGUȚ, L. 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. DOI: 10.1016/j.isprsjprs.2016.01.011.

BOUCKAERT, R.; FRANK, E.; HALL, M.; KIRKBY, R.; REUTEMANN, P.; SEEWALD, A. S. **WEKA Manual for Version 3-7-1.** University of Waikato, 2010.

BREDE, B. et al. Non-destructive tree volume estimation through quantitative structure modeling: Comparing UAV laser scanning with terrestrial LIDAR. **Remote Sensing of Environment**, v. 233, p. 111355, 2019. DOI: 10.1016/j.rse.2019.111355.

BREIMAN, Leo. Random forests. **Machine Learning**, v. 45, p. 5-32, 2001.

CAMPOS, J. C.; LEITE, H. G. **Mensuração Florestal: Perguntas e Respostas.** 5ª ed. Viçosa: UFV, 2017. 407 p.

CETIN, Z.; YASTIKLI, N. The use of machine learning algorithms in urban tree species classification. **ISPRS International Journal of Geo-Information**, v. 11, n. 4, p. 226, 2022.

COSTA FILHO, S. V. S. et al. Configuração de algoritmos de aprendizado de máquina na modelagem florestal: um estudo de caso na modelagem da relação hipsométrica. **Ciência Florestal**, v. 29, n. 4, p. 1501-1515, 2019. DOI: 10.5902/1980509828392.

DA SILVA, A. K. V. et al. Predicting Eucalyptus Diameter at Breast Height and Total Height with UAV-Based Spectral Indices and Machine Learning. **Forests**, v. 12, n. 582, p. 2-13, 2021. DOI: 10.3390/f12050582.

DEUR, M. et al. Tree Species Classification in Mixed Deciduous Forests Using Very High Spatial Resolution Satellite Imagery and Machine Learning Methods. **Remote Sensing**, v. 12, p. 3926, 2020. DOI: 10.3390/rs12233926.

EGMONT-PETERSEN, M.; DE RIDDER, D.; HANDELS, H. Image processing with neural networks—A review. **Pattern Recognition**, v. 35, p. 2279-2301, 2002. DOI: 10.1016/S0031-3203(01)00178-9.

EMBRAPA. Sistema Brasileiro de Classificação de Solos. **Embrapa Solos**, 2018.

GAISBERGER, H.; VINCETI, B. As**sessing threats to genetic resources of food-tree species in Burkina Faso**. Case study for FAO's SOFO 2020. 2020. DOI: 10.4060/ca8642en.

FLORA DO BRASIL. **Flora e Funga do Brasil. Disponível em :** https://floradobrasil.jbrj.gov.br/reflora/listaBrasil/ConsultaPublicaUC/ConsultaPublicaUC.do#CondicaoTaxonCP. . Acesso em: 23 de novembro 2022.

HUANG, Jui-Chan et al. Application and comparison of several machine learning algorithms

and their integration models in regression problems. **Neural Computing and Applications**, v. 32, p. 5461-5469, 2020.

Li, T.; Li, M.; Ren, F.; Tian, L. Estimation and Spatio-Temporal Change Analysis of NPP in Subtropical Forests: A Case Study of Shaoguan, Guangdong, China. **Remote Sensing**. V. 14, p. 2541, 2022, https://doi.org/10.3390/rs14112541

LIM, J. et al. Machine Learning for Tree Species Classification Using Sentinel-2 Spectral Information, Crown Texture, and Environmental Variables. **Remote Sensing**, v. 12, p. 2049, 2020. DOI: 10.3390/rs12122049.

MARRS, J. et al. Machine Learning Techniques for Tree Species Classification Using Co-Registered LiDAR and Hyperspectral Data. **Remote Sensing**, v. 11, p. 819, 2019. DOI: 10.3390/rs11070819.

MAXWELL, A. E.; WARNER, T. A.; FANG, F. Implementation of machine-learning classification in remote sensing: an applied review. **International Journal of Remote Sensing**, v. 39, n. 9, p. 2784-2817, 2018. DOI: 10.1080/01431161.2018.1433343.

NALEPA, J.; KAWULOK, M. Selecting training sets for support vector machines: A review. **Artificial Intelligence Review**, v. 52, p. 857-900, 2019. DOI: 10.1007/s10462-017-9611-1.

ÖZÇELIK, R. et al. Estimating tree trunk volume using artificial neural network models for four species in Turkey. **Journal Environment. Manage**, v. 91, p. 742-753, 2010. DOI: 10.1016/j.jenvman.2009.10.002.

PAGOTTO E SOUZA, T. C. S.; SOUZA, P. R. Biodiversidade do complexo aporé Sucuriú: acessórios à conservação e ao manejo do Cerrado. **Edição UFMS**, Campo Grande, 2006.

Ponzoni, F. J., Shimabukuro, Y. E., & Kuplich, T. M. (2007). **Sensoriamento remoto no estudo da vegetação**, p. 127, São José dos Campos: Parêntese.

RAMOS, Ana Paula Marques et al. A random forest ranking approach to predict yield in maize with uav-based vegetation spectral indices. **Computers and Electronics in Agriculture**, v. 178, p. 105791, 2020.

ROCHA, J. E. C. et al. Configuração de redes neurais artificiais para relação hipsométrica de árvores de Eucalyptus spp. **Scientia Forestalis**, v. 49, n. 132, p. e3706, 2021. DOI: 10.18671/scifor.v49n132.08.

ROELL, Y. E., Beucher, A., Møller, P. G., Greve, M. B., & Greve, M. H. Comparing a random forest based prediction of winter wheat yield to historical yield potential. **Agronomy**, v. 10, n. 3, p. 395, 2020.

SILVA, M. L. M. D., Binoti, D. H. B., Gleriani, J. M., & Leite, H. G. Ajuste do modelo de Schumacher e Hall e aplicação de redes neurais artificiais para estimar volume de árvores de

eucalipto. **Revista Árvore**, v. 33, p. 1133-1139, 2009.

SNOUSY, M. B. A. et al. Suite of decision tree-based classification algorithms on cancer gene expression data. **Egyptian Informatics Journal.**, v. 12, p. 73–82, 2011. DOI: 10.1016/j.eij.2011.04.003.

TSAI, M.-D. et al. Exploring Airborne LiDAR and Aerial Photographs Using Machine Learning for Land Cover Classification. **Remote Sensing.** v. 15, p. 2280, 2023. DOI: 10.3390/rs15092280.

FRANK, Eibe; HALL, Mark A., **Data mining: Practical machine learning tools and techniques**. 2016.