LVI SBPO SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL
DE 04 ATÉ 07 NOVEMBRO 2024
FORTALEZA | CEARÁ

ANÁLISE DE DESEMPENHO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA NA CLASSIFICAÇÃO DE DANOS OCASIONADOS POR MACACOS-PREGO EM PLANTAÇÕES DE PINUS

Flavio Chiarello
Programa de Pós-Graduação em Engenharia de Produção e Sistemas - PPGEPS,
Escola Politécnica – EP, Pontifícia Universidade Católica do Paraná - PUCPR,
Imaculada Conceição, 1155, CEP. 80215-901, Curitiba, PR, Brasil
flavio_chi@hotmail.com

Gilberto Reynoso Meza
Programa de Pós-Graduação em Engenharia de Produção e Sistemas - PPGEPS,
Escola Politécnica – EP, Pontifícia Universidade Católica do Paraná - PUCPR,
Imaculada Conceição, 1155, Prado Velho, CEP. 80215-901, Curitiba, PR, Brasil
g.reynosomeza@pucpr.br

Roberto Zanetti Freire
Programa de Pós-Graduação em Engenharia de Produção e Sistemas - PPGEPS-PB
Universidade Tecnológica Federal do Paraná - UTFPR,
Av. Sete de Setembro, 3165, Rebouças, CEP. 80230-901, Curitiba, PR, Brasil
robertofreire@utfpr.edu.br

RESUMO

A quebra de paradigmas alterou a dinâmica da silvicultura moderna tornando-a multidisciplinar, incluindo processos de melhoria contínua por meio de métodos naturais ou artificiais, como a modificação genética ou pela introdução de ferramentas tecnológicas. Neste sentido, o objetivo deste trabalho é traçar perfis sobre o padrão de danos caudados por *Sapajus nigritus* Goldfuss, 1809, popularmente conhecido como macaco-prego, em plantios de *Pinus taeda* L., com o auxílio de técnicas de aprendizado de máquina, com foco na classificação de árvores que serão ou não atacadas. Os indicadores evidenciaram uma boa convergência dos métodos *XGBoost* e *Random Forest*, fornecendo 73,5% e 72,4% de Acurácia, respectivamente, otimizando seus hiperparâmetros. Ao mesmo tempo, os gráficos expõem a importância das variáveis: data do inventário, confirmando a dependência temporal dos dados, e altura e faixa, o que sugere uma possível relação geográfica da ação dos macacos-prego em áreas distantes da ação humana.

PALAVRAS-CHAVE. Macaco-prego, Aprendizado de Máquina, Pinus.

ABSTRACT

The breaking of paradigms has altered the dynamic of modern silviculture, working on multidisciplinary continuous improvements through natural or artificial methods, such as genetic modification or the introduction of technological tools. In this sense, the objective of this work is to profile the pattern of damage caused by *Sapajus nigritus* Goldfuss, 1809, popularly known as capuchin monkey, in *Pinus taeda* L. plantations using machine learning methods to classify attacked or non-attacked trees. When optimized hyperparameters were assumed, performance measures showed reasonable convergence of the XGBoost and Random Forest methods, with 73.5% and 72.4% accuracy, respectively. At the same time, the graphical analysis exposed the importance of the variable inventory date, confirming the temporal dependence of the data, height, and band, suggesting a possible geographic relationship of capuchin monkey action in areas far from human activity.

KEYWORDS. Capuchin Monkey, Machine Learning, Pinus.

DOI: 10.59254/sbpo-2024-193805
https://proceedings.science/p/193805?lang=pt-br

1. Introdução

A interconectividade da Floresta 4.0 representa a aplicação de inúmeras tecnologias avançadas nas atividades basilares dos sistemas florestais, sobretudo na melhoria dos níveis de gestão, estratégias de manejo, aumento da rotatividade e produtividade, e de intensificadas filosofias de redução de desperdícios. O macroambiente nacional brasileiro, hoje, consolida como uma das grandes forças econômicas as operações de agronegócio, incluindo a indústria madeireira de reflorestamento, com a política de atender as demandas presentes e futuras sem a necessidade de comprometer a sustentabilidade e os ecossistemas nativos.

Com posição de destaque no desenvolvimento da economia, a modalidade Florestal alcançou, em 2019, uma receita bruta total de R$ 97,4 bilhões, além de gerar oportunidades a 3,75 milhões de pessoas em todo o país, segundo relatório disponibilizado pela Indústria Brasileira de Árvores (IBÁ). O alinhamento deste cenário com a concepção de que o setor é considerado um dos principais pilares e fornecedores de insumos para a indústria em geral, evidencia o potencial e a relevância no abastecimento interno da economia, subsidiando positivamente no aumento de 1,2% do Produto Interno Bruto, o PIB nacional [IBÁ 2020].

Apesar do viés econômico apresentar grande rentabilidade, a incidência de pragas e doenças é um dos principais fatores que acometem o crescimento dos povoamentos comerciais, resultando em elevados prejuízos financeiros [Mikich e Liebsch 2014b; Liebsch et al. 2015]. Um dos problema de dimensões significativas tem despertado a atenção de empresas e especialistas. O *Sapajus nigritus* (Goldfuss, 1809), popularmente conhecido como macaco-prego, vem afetando plantios de *Pinus* spp. desde a década de 1980, com maiores intensidades a partir de 1990. As áreas com maior ocorrências pertencem à faixa sul e sudeste do Brasil, envolvendo os estados de São Paulo, Paraná, Santa Catarina e Rio Grande do Sul [Mikich e Liebsch 2009; Liebsch et al. 2015].

Geralmente os macacos-prego estão concentrados, estrategicamente, no estrato médio da floresta, mas podem percorrer distâncias em solo para a captação de alimentos ou deslocamento. Apresentam hábitos onívoros, incluindo insetos, ovos de aves, pequenos vertebrados, folhas, bulbos, frutos e sementes [Galetti e Pedroni 1994], além de desempenhar importante papel na dinâmica das florestas, atuando diretamente como dispersores de sementes [Mikich et al. 2015].

Para [Hill 2000], espécies como *Sapajus nigritus* são frequentemente referenciadas pela alcunha de pragas ou pestes por, simplesmente, condicionarem perdas econômicas e produtivas. [Santos et al. 2007], no entanto, reiteram que o termo adequado seja “populações problema”, visto que os danos são localizados e causados por grupos, e não pela espécie como um todo. Os ataques são classificados pela retirada da casca em tiras longas, geralmente no terço superior da árvore, em dois tipos básicos de padrão, conforme Figura 1. A janela contempla o traço mais comum, na qual apenas uma parte do tronco é danificado, enquanto o descascamento em toda circunferência do caule é tipificado como sendo anelamento [Mikich e Liebsch 2009].

Figura 1: Apresentação dos tipos de danos causados por macacos-prego em árvores de Pinus taeda: (a) práticas de janelamento; (b) anelamento.



(a) (b)

Fonte: Adaptado de [Mikich e Liebsch 2009].

DOI: 10.59254/sbpo-2024-193805
https://proceedings.science/p/193805?lang=pt-br

Diante das características econômicas do assunto, a Empresa Brasileira de Pesquisa Agropecuária (Embrapa) desenvolve continuamente projetos em parceria com instituições públicas e privadas desde 2003 [Embrapa Florestas - Programa Macaco-prego 2024]. O teor destes estudos envolve a construção de metodologias eficientes e diretrizes operacionais no controle e/ou redução dos danos causados aos povoamentos de *Pinus* spp.

Inserido nessa temática, este trabalho faz uso de métodos de Aprendizado de Máquina - AM, (do inglês: *Machine Learning*), os quais vêm se tornando sólidos aliados na concepção de métodos inovadores que podem ser utilizados em exercícios plurivalentes de quantificação de variáveis intrínsecas, detecção de ataques, classificação de coberturas e monitoramento periódico, por exemplo, todos aplicados a área florestal e seu beneficiamento.

Diante disso, o objetivo deste estudo consiste em avaliar o desempenho das técnicas de aprendizado de máquina, *Random Forest* e *XGBoost*, na classificação de árvores que sofreram, ou não, danos por macacos-prego, buscando classificar árvores quanto a expectativa de ataques.

Este artigo foi organizado considerando a estrutura apresentada a seguir. Na Seção 2, apresenta-se uma revisão de literatura que descreve os métodos aplicados à pesquisa, bem como algumas definições importantes. Na sequência, a Seção 3 aborda todo aspecto de coleta e uso de dados, enquanto a Seção 4 trabalha com os resultados encontrados e a Seção 5 traz, além do das conclusões do trabalho, algumas oportunidades de trabalhos futuros.

2. Materiais e Métodos

Esta seção apresenta os métodos de AM adotados ao longo deste estudo e algumas técnicas complementares visando melhorar o desempenho dos modelos.

2.1 *Random Forest*

Resultado de conceitos derivativos, *Random Forest* (RF) é uma metodologia de comitê de máquinas (*ensemble*) que agrega aspectos estruturais e sistemáticos da Árvore de decisão, do inglês *Decision Tree* - DT [Quinlan 1986]. É constituída por um conjunto de árvores sucessivas que crescem em razão de outras, como sendo um relacionamento genético de pai e filho. Cada árvore é determinada de forma independente e aleatória, utilizando-se de reamostragens (*Bootstrap*) da base de dados original. No trabalho apresentado em [Breiman 2001], define-se o RF como sendo uma combinação de preditores orientados como classificadores estruturados em árvores.

Apesar de estarem vinculadas à dinâmicas parecidas, uma das justificativas para se trabalhar com RF, ao invés do DT, pode ser encontrada durante o processo de construção do algoritmo. Para a primeira técnica, são desenvolvidas e analisadas várias árvores ao mesmo tempo, com a seleção de conjuntos aleatórios de dados de treinamento e de subconjuntos para a modelagem do problema, o que permite a previsão agregada de informações e, frequentemente, o oferecimento de melhores resultados do que apenas uma única árvore.

O RF é constantemente aplicado por vários pesquisadores, entre eles, por [Miranda et al. 2022] para a estimativa da altura individual das árvores com inicialização por algoritmos genéticos, e em [Alckmin et al. 2022], no mapeamento da quantidade de matéria seca de biomassa com imagens obtidas por drones.

2.2 *Boosting*

A técnica de *Boosting* representa um método que pode ser utilizado para melhorar o desempenho de qualquer algoritmo de AM [Freund e Schapire 1996]. O método conta com arquitetura flexível e de rápida implementação. Além disso, vem apresentando eficiência interessante em processos de classificação e previsão, especialmente com o *Gradient Boosting Machine* (GBM) e sua evolução, o *Extreme Gradient Boosting* (*XGBoost*) [Chen e Guestrin 2016].

O funcionamento do *Boosting* baseia-se na combinação de vários mecanismos de aprendizado na busca pela construção de um modelo mais forte. A essência está no aprimoramento do poder preditivo das respostas com base na aplicação, de forma sequencial, de um conjunto de preditores executados paralelamente, cada um tentando corrigir a resposta de seu antecessor [Chen e Guestrin 2016].

DOI: 10.59254/sbpo-2024-193805
https://proceedings.science/p/193805?lang=pt-br

As arquiteturas *Boosting* integram diversos hiperparâmetros que devem ser analisados e ajustados, o que acaba tornando possível a adaptação em problemáticas lineares ou não lineares e permitindo a obtenção de melhores níveis de eficiência [Miller et al. 2017]. As metodologias assumem características de Árvores de Decisão e buscam a minimização da função objetivo através da geração de novas árvores de probabilidade ao final de cada iteração anterior. A existência de várias ramificações só atende o critério de parada quando a adição de uma nova ramificação não reduz significativamente a função, o que sugere que o modelo encontrado está em sua fase inteligente (estágio satisfatório) com os parâmetros apresentados.

O conjunto de respostas é então ajustado ao longo de cada iteração, em função da soma de resíduos anteriores, regularizado por uma taxa de aprendizado. Em casos de regressão, são consideradas as diferenças entre os valores reais e simulados, enquanto a classificação é atendida pela seleção de uma categoria dentro de uma coleção pré-definida [Miskell et al. 2019]. Os estudos de [Watt et al. 2021], na determinação de índices de sítio associados ao processo produtivo em povoamentos de *Pinus radiatta*, e de [Huang et al. 2023], na estimativa de biomassa acima do solo, são exemplos de aplicação do método em variáveis presentes no setor florestal.

2.3 *Hiperparametrização*

A hiperparametrização consiste em um processo de adaptação dos aspectos que influenciam diretamente no comportamento do modelo e que não são aprendidos durante a fase de treinamento. A definição dos valores é geralmente baseada em experimentos empíricos em que várias opções de configuração são testadas por meio de um conjunto de validação, de forma a verificar se as modificações nos hiperparâmetros conseguem determinar os melhores índices de desempenho. Nesse trabalho, esta ação foi feita pelo uso da técnica automática de busca em caixa preta chamada de *OpenBox* [Li et al. 2021].

A diversificação de abordagens com especificações e mecanismos distintos, é uma característica do espaço ambientado pelo *tuning*. A popularização de métodos como o *Search Grid* (modelo de pesquisa em grade que faz varreduras em todas as combinações possíveis), *Random Search* (buscas aleatórias), *Bayesian Optimization* (estratégia de projeto sequencial para funções de caixa preta) e os algoritmos evolutivos populacionais, por exemplo, denotam uma literatura em crescimento.

A interoperabilidade do *OpenBox* como otimizador faz com que seja possível avaliar diferentes *frameworks* de busca, incluindo os híbridos, Bayesianos (Processos Gaussianos e *Tree-structured Parzen Estimator*) e a RF probabilística.

3. Descrição da Área e Base de Dados

Os pesquisadores da Embrapa Florestas, no ano de 2007, instalaram um experimento envolvendo um povoamento de *Pinus taeda* plantado em 2001, de propriedade privada. Com espaçamento entre as árvores de 3 m × 3 m, de área fixada no município de Ponte Serrada, em Santa Catarina. O povoamento integrava 1.111 árvores/ha e um tratamento silvicultural de poda até dois metros de altura aos quatro anos de idade. As coordenadas geodésicas estão entre 27°1’27’’S e 51°53’9’’O e a altitude média é de 1.080 metros. O clima da região é caracterizado como subtropical úmido mesotérmico, segundo a classificação de Köppen (1948) [Liebsch et al. 2015], com geadas severas, ausência de estações de secas e verão brando.

Com aproximadamente 50 hectares (1.000 m × 500 m) de extensão, a região recebeu, sistematicamente, a alocação de 169 parcelas homogêneas distribuídas em formato de cruz de malta. Os grupos, distantes 50m um dos outros, eram compostos por 25 árvores ordenadas em quatro linhas, partindo de um ponto de referência central. Mesmo que em condições atípicas, a supressão de árvores, ou simplesmente falhas, implicava na atividade de registro e rotulação dos pontos faltantes, por existir riscos de comprometimento da dinâmica do processo. Os diferentes vetores de informações foram roteirizados durante o período de 2007 a 2019, tendo como recurso principal as aferições manuais recorrentes durante o mês de dezembro (exceto pelo primeiro ano que contou, também, com medições no marco zero do projeto - junho) [Liebsch et al. 2015].

DOI: 10.59254/sbpo-2024-193805
https://proceedings.science/p/193805?lang=pt-br

As recursivas avaliações em campo ao longo das unidades amostrais não só permitiram a coleta de dados anuais da circunferência a altura do peito (CAP) individual, como também a identificação do tipo de ataque ocasionado pelo macaco-prego. É importante destacar que as árvores janeladas, ao sofrerem anelamento, recebem a designação do efeito dominante de maior impacto, isto é, anelamento. A sensibilidade dessas anotações mostra significativa precisão das observações, principalmente na etapa descritiva sobre a quantidade e altura do dano, índices de mortalidade e na incidência de indivíduos com ponta seca ou quebrada.

As intensificadas ações de monitoramento periódico no decorrer dos 13 anos de pesquisa trouxeram informações detalhadas. A potencialização das análises viabilizou a estruturação de um conjunto de 46.813 amostras tabulares. A série de dados mostra desde o número de tratamento e operações silviculturais de manejo, até o estado em que a árvore se encontra durante a linha temporal. O experimento apresentado neste estudo utilizou apenas registrados até 2013, isto é, 23.300 dados (1.829 árvores atacadas e 21.471 não atacadas) [Embrapa Florestas - Programa Macaco-prego). A lista compreende o total de 14 variáveis, sem contar a saída desejada, conforme apresentada pela Tabela 1.

Tabela 1. Descrição dos dados utilizados no estudo.

Parâmetros	Variações assumidas no estudo	Modalidade dos dados
Posição (da Árvore)	Índice central, superior, atrás, lado direito e esquerdo.	Qualitativa
Manejo	Com ou sem desbastes, LT e outros.	Qualitativa
CodTratamento	1 a 11 (Especificação de cada tratamento)	Quantitativa
Faixa	Intervalo de 1 a 12.	Quantitativa
Linha	1 até 22.	Quantitativa
Unidade Amostral	Entre 1 e 169.	Quantitativa
Nº Indivíduo	1 até 6.	Quantitativa
Data do Inventário	Junho/2007 e Dezembro (2007 a 2019).	Quantitativa
Circunferência a altura do peito (CAP)	Medidas que variam conforme a idade (24 cm a 170 cm).	Quantitativa
Diâmetro a altura do peito (DAP)	Medidas que variam conforme a idade (7,63 até 54,11).	Quantitativa
Corte (da árvore)	Falso ou verdadeiro.	Qualitativa
ID_Tab_Árvores	Registro de identificação do indivíduo (1 até 3.413).	Quantitativa
Altura	Valores Positivos.	Quantitativa
Volume	Valores Positivos.	Quantitativa

Fonte: Adaptado do Programa Macaco-prego Embrapa Florestas.

3. Metodologia

Devido ao atual escopo, propôs-se a extensão das informações com o ajuste de duas modelagens: A relação hipsométrica de Stofells e Soest (1953) na determinação da altura relativa individual (em metros) e o uso de um método empírico capaz de calcular o volume unitário de madeira (m³). A finalidade das parametrizações é construir uma base para alcançar os resultados esperados da pesquisa. As expressões que descrevem este processo estão nas Equações (1) e (2):

$$Hr_{a,i} = \alpha_i * DAP_{a,i}{}^2 + \beta_i * DAP_{a,i} + \delta_i, \quad \forall\, i \in S;\ \alpha \in V, \tag{1}$$

sendo Hr_{ai} a altura relativa da árvore a no período i, V é denominado o espaço que compreende o número total de árvores, DAP_{ai} denota o diâmetro do indivíduo a na altura de 1,30 m na idade i, S

DOI: 10.59254/sbpo-2024-193805
https://proceedings.science/p/193805?lang=pt-br

é o espaço que define a quantidade de anos do experimento e, por fim, α, β e δ retornam um valor constante (pré-definido) diferente para cada ano i.

$$V_{a,i} = \frac{\left(FF * \pi * DAP_{ai}^2 * Hr_{ai}\right)}{40.000}, \qquad \forall\, i \in S, \tag{2}$$

onde FF configura o valor referente ao fator de forma das árvores (por convenção foi usado o *default* do *Software* SisPinus – 0.4915) [Embrapa Florestas, 2024].

A etapa seguinte consiste no balanceamento amostral do conjunto de dados. Como a proporção de árvores atacadas é significativamente menor do que as não atacadas, houve a necessidade de aplicação da técnica SMOTE (*Synthetic Minority Oversampling Technique*) [Chawla et al. 2002], uma abordagem de aumento de dados (*oversampling*) que utiliza todas as informações repassadas para criar amostras fictícias da classe minoritária (Tabela 2).

Tabela 2. Detalhamento da base de dados original e expansão com SMOTE.

Banco de Dados sem reincidência de danos				
Etapa de Treinamento e Validação dos modelos				
Conjunto:	Real		Expandido (Sintético)	
	N° Atacadas	N° Não Atacadas	N° Atacadas	N° Não Atacadas
N° Instâncias	1.829	21.471	21.471	21.471
Total:	23.300		42.942	

Fonte: Os autores [2024].

Partindo do contexto de que o número de árvores atacadas (2.033 – considerando reincidência de dano, isto é, a mesma árvore pode receber de 1 a 8 ataques [Liebsch et al. 2015]), é muito inferior às não atacadas (21.471), decidiu-se selecionar 406 padrões igualmente distribuídos (203 para cada possibilidade) na intenção de formar um conjunto de teste. A escolha sucedeu de maneira aleatória com base em 10% do número total de danos por ano. Por exemplo: junho de 2007 terminou com 8 registros e apenas 1 padrão selecionado. O ano de 2013, por sua vez, contabilizou 868 lesões e, portanto, 87 amostras. Já para a classe dos indivíduos com ausência de ataque, o processo é mais simples: exceto 2013 que ficou com 28, os demais anos tiveram um valor fixado de 25 exemplares (Tabela 3).

Tabela 3. Descrição dos novos dados (fase de teste).

Ano	Total de Instâncias	N° Atacadas	N° Não Atacadas
Jun / 2007	26	1	25
Dez / 2007	27	2	25
Dez / 2008	29	4	25
Dez / 2009	39	14	25
Dez / 2010	45	20	25
Dez / 2011	48	23	25
Dez / 2012	77	52	25
Dez / 2013	115	87	28
Total:	406	203	203

Fonte: Os autores [2024].

A fim de simplificar o conteúdo, as árvores denominadas cortadas foram retiradas no respectivo ano de desbaste, enquanto as células nulas passaram a ser preenchidas com zeros. A transformação dos dados para um formato padronizado facilita a entrada das instâncias nos métodos de AM. Para o presente estudo, foi utilizada a normalização no intervalo fechado [0,1] e a validação cruzada estratificada *K-Fold* (com 5 seções) para avaliar a capacidade de generalização

DOI: 10.59254/sbpo-2024-193805
https://proceedings.science/p/193805?lang=pt-br

do modelo no conjunto de treinamento (80%) e teste (20%), garantindo aleatoriedade, mitigação de *overfitting* (ou *underfitting*) e equilíbrio amostral.

As execuções dos dois métodos (*Random Forest* e *XGBoost*) ocorreu com a mesma semente aleatória de inicialização. O cenário inicial considera apenas o uso de parâmetros *default*, isto é, as melhores configurações segundo o estado atual da arte, enquanto o segundo, *HyperTuning*, tem o acréscimo da hiperparametrização (Tabela 4).

Tabela 4. Parametrização dos métodos de AM de acordo com o cenário analisado.

Algoritmos	Utilizado no Cenário *"Default"*	Subespaço de busca do otimizador *OpenBox*	Melhores configurações obtidas para "*HyperTuning*"
RF	N° de Estimadores = 100; Critério Gini; Profundidade máxima = "*None*"; Quantidade de características = "Auto"; Divisão mínima de amostras = 2 e; Folhas mínimas = 1	N° de estimadores = [10; 2.000] e profundidade máxima = [5; 1.500]	Profundidade Máxima = 1400 e N° de estimadores = 70
XGBoost	Taxa de Aprendizado = 0,1; Profundidade máxima da árvore = 3 e; Número de estimadores = 100	Taxa de aprendizado = [0.1; 1.9]; profundidade máxima = [20;500] e; N° de estimadores = [200; 2.500]	Taxa de Aprendizado = 1.8879; Profundidade Máxima = 476 e N° Estimadores = 2.500

Fonte: Os autores [2024].

É importante destacar que o número de execuções foi definido igual a 100 e que a dimensão assumida pelo domínio de generalização do *XGBoost*, para a taxa de aprendizado, é, por padrão, igual a 0,3, ou limitada no intervalo de 0,1 até 1,0. A lógica se resume em desviar de eventuais problemas de *overfitting* e a acondicionar a estabilidade. Todavia, na iniciativa de mapear o seu funcionamento na busca por melhores resultados, projetou-se um limite ainda maior, variando de 0,1 a 1,9.

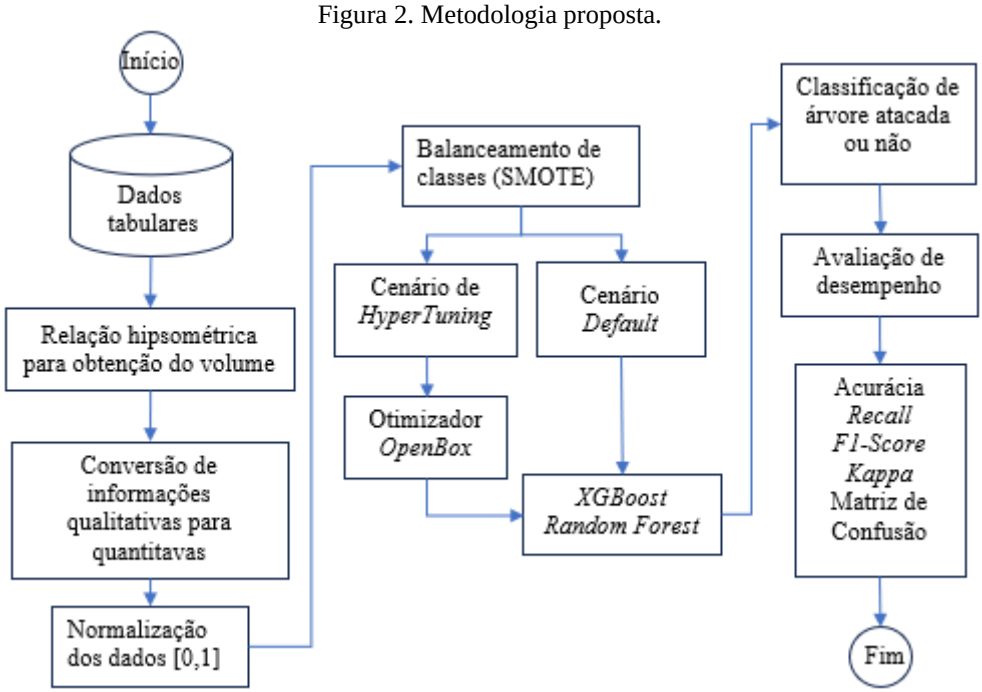
Ao final, os resultados foram avaliados de acordo com a matriz de confusão e quatro medidas principais, a Acurácia, coeficiente Kappa de Cohen ($Kappa$), $F1_Score$ e $Recall$. Essas e outras informações podem ser observadas de acordo com a Figura 2.

A Acurácia (Acc) é um índice utilizado como uma métrica referente a “taxa de acerto” dos modelos de classificação de AM, esta medida de desempenho é apresentada na Equação (3). O $Recall$, também conhecido como sensibilidade, mede a capacidade de um modelo captar todas as instâncias relevantes de uma classe positiva. Valores altos sinalizam a proficiência do sistema em evitar falsos negativos, tornando-o, assim, particularmente útil nos cenários em que a detecção de casos positivos é importante, sua definição está na Equação (4).

O $F1_Score$ combina $Recall$ com a Acc, conforme apresentado na Equação (5), oferecendo uma visão mais completa em situações em que ambas as métricas são igualmente necessárias. É aplicado em casos em que há desiquilíbrios significativos entre as classes, evitando extremos e fornecendo um resultado mais equitativo para avaliação. Por fim o último coeficiente de desempenho utilizada neste trabalho é o coeficiente $Kappa$, um algoritmo que trabalha bem com desbalanceamentos nas classes por conta de avaliações nos níveis de reprodutibilidade entre dois conjuntos de dados. O mecanismo de funcionamento faz referência aos testes de hipóteses e intervalos de confiança, na qual há uma taxa de aceitação relativa e hipotética responsáveis por determinar a concordância dos resultados, conforme Equação (6).

DOI: 10.59254/sbpo-2024-193805
https://proceedings.science/p/193805?lang=pt-br

Figura 2. Metodologia proposta.



Fonte: Os autores [2024].

Nas Equações (4) a (6), tp, fn e fp representam os verdadeiros positivos, falsos negativos e falsos positivos, respectivamente. Já $\hat{p}_0$ está associado à proporção da concordância observada entre os avaliadores e $\hat{p}_e$ é a proporção da concordância esperada ao acaso.

$$Acc = \frac{tp}{tp + fp} \tag{3}$$

$$Recall = \frac{tp}{tp + fp} \tag{4}$$

$$F1_Score = 2 * \frac{Acc * Recall}{Acc + Recall} \tag{5}$$

$$Kappa = \frac{\hat{p}_0 - \hat{p}_e}{1 - \hat{p}_e} \tag{6}$$

4. Resultados

Os recursos computacionais utilizados para a implementação dos métodos contou com um processador Intel i5-13400f (4,3 GHz) de 13ª geração, com 16 GB de memória de acesso aleatório (RAM), um disco de estado sólido (SSD) de 512GB, uma placa de vídeo RTX 3060 de 12GB e sistema operacional Windows 11. O ambiente de codificação foi o Python na versão 3.10 (MiniConda) pelo Visual Studio (VSCode).

Ao examinar individualmente as experimentações, é possível generalizar as conclusões de que os modelos obtiveram razoáveis índices de eficiência para o cenário *default*. O *Random Forest* é o principal preditor, apresentando Acurácia de 71,6% e $Recall$ de 94,9% durante o treinamento. Isso indica a categorização correta de 30.758 instâncias com uma taxa razoavelmente alta de identificação correta dos casos positivos, com uma margem muito pequena de falsos positivos.

O Coeficiente $Kappa$ de 43,3% representa uma concordância moderada entre os avaliadores, deixando em aberto possibilidades de melhoria em termos de consistência. Por outro lado, a apresentação de novos valores ao modelo expõe não só a precisão de 70,19% e $F1_Score$ de 71,12%, refletindo em um modelo não excessivamente inclinado para um lado ou para outro

DOI: 10.59254/sbpo-2024-193805
https://proceedings.science/p/193805?lang=pt-br

(equilíbrio entre Acurácia e *Recall*) como, também, no acerto de 66,99% das árvores atacadas e 73,40% das não atacadas.

Já o *XGBoost* manteve-se muito próximo do *Random Forest*, com Acurácia de 69,5% e *Recall* de 91,4% durante o treinamento. O efeito comparativo sugere que as maiores nuances se concentraram na menor precisão da classe de árvores atacadas (91,44%), enquanto a classificação das árvores não atacadas apresentou similaridades (47,57%). Por outro lado, em termos de teste, o algoritmo também obteve resultados razoáveis, atingindo 69,95% de Acurácia e 70,67% de $F1_Score$.

Em ambos os casos é visível a dificuldade em classificar os rótulos de não ataques. Conforme Liebsch et al. [2015], os ataques não seguem um padrão propriamente dito, mas há indícios de maior incidência em indivíduos adultos vigorosos (geralmente no terço superior) e em épocas que o alimento primário dos macacos-prego é escasso (frutos e sementes, principalmente).

Para o cenário *HyperTuning*, a composição dos resultados é feita pela interconectividade entre o espaço de busca (*OpenBox*) para os parâmetros e o número de execuções, igual a 100, na qual designa-se em aperfeiçoar a eficiência dos classificadores.

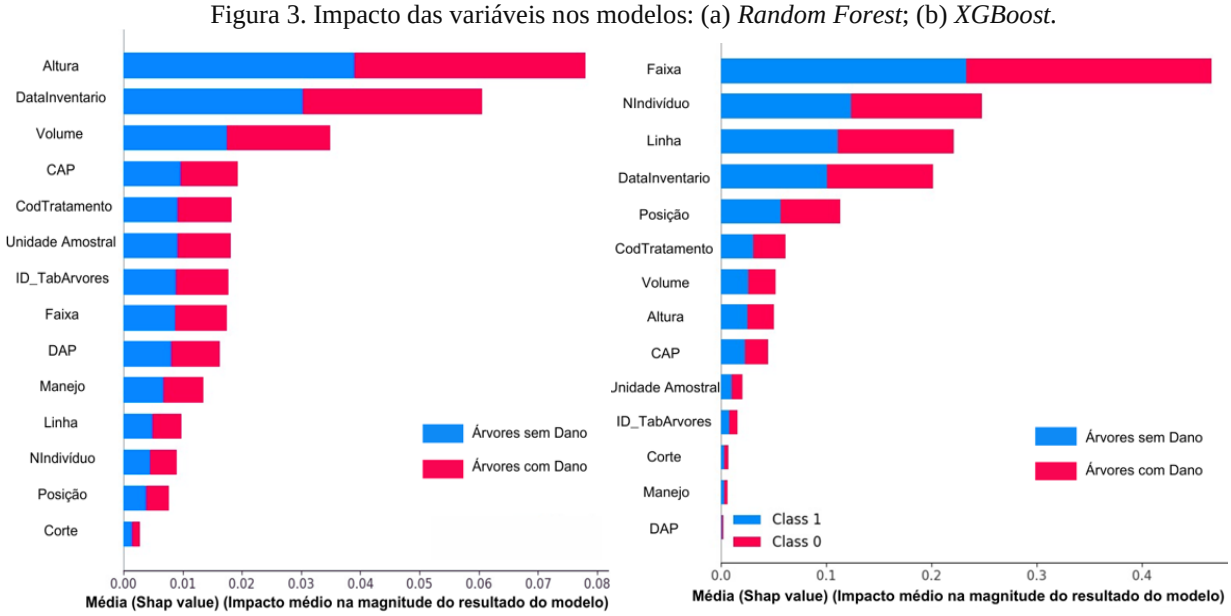
Os ensaios projetam aspectos variados de desempenho. A reprodutibilidade de detalhes traz em perspectiva uma sutil melhora dos algoritmos em termos de Acurácia (73,5% e 72,4% para *XGBoost* e *Random Forest*, respectivamente), tanto para a classe atacada quanto não atacada. Novamente, para o conjunto de teste, os valores são moderados. Diferentemente do primeiro cenário, o *XGBoost* foi o melhor algoritmo, atingindo 72,17% de precisão contra 70,20% do *Random Forest*.

Analisando por outro espectro, a ferramenta *Shap Value* demonstra, visualmente, a contribuição de cada atributo na saída desejada em termos de magnitude média individual, conforme apresentado na Figura 3. As barras horizontais (azul e vermelho) ao longo das abscissas designam, respectivamente, a ausência e a eminência dos eventos de ataque em total balanceamento. A hierarquia dos dados traz em foco a significante influência da data da altura (0,078) e o volume comercializável (0,034). Essas novas tendências podem significar níveis de preferência por árvores ditas vigorosas e, também, na seleção direta por indivíduos dominantes, codominantes ou dominados, corroborando com algumas afirmações previamente evidenciadas por [Liebsch et al. 2015]. Já as diretrizes de corte (0,002) continuam sendo pouco significativas, assim como a posição na cruz de malta (0,007) e o número do indivíduo (0,009), como apresentado na Figura 3 (a).

As interpretações preservam uma nova classe de componentes a serem priorizados, 185 como a faixa (0,49), o número do indivíduo (0,25) e a linha (0,215), reproduzindo a hipótese de que o algoritmo trabalha bem com informações de espaço e, eventualmente, a possibilidade de existir correlação entre a distribuição geográfica da árvore na cruz de malta com os padrões de ataque. A intensificação nas demais magnitudes, como o volume na quarta posição (0,20), determina superioridade ao primeiro elemento do RF e, ao mesmo tempo, um conjunto maior de parâmetros impactantes, tornando-se um ponto de vista útil na compreensão aprofundada do problema. Não obstante, vale ressaltar que mesmo que o DAP e a CAP sendo matematicamente proporcionais, o número de amostras utilizadas no experimento (n = 2.000) não extraiu 100% este relacionamento, como pode-se notar na Figura 3 (b).

DOI: 10.59254/sbpo-2024-193805
https://proceedings.science/p/193805?lang=pt-br

Figura 3. Impacto das variáveis nos modelos: (a) *Random Forest*; (b) *XGBoost*.



(a) (b)

Fonte: Os autores [2024].

5. Conclusões

Ao longo do estudo foram conduzidas análises envolvendo métodos computacionais na busca por modelos capazes de gerar classificações precisas em meio a um tema de grande relevância nos espectros econômicos, políticos e ambientais. Os ataques de *Sapajus nigritus* (Goldfuss,1809) atingem magnitudes significativas de impacto na qualidade e produtividade de plantios de espécies comerciais como pinus e eucalipto.

As configurações básicas de pré-processamento, balanceamento amostral da classe minoritária (SMOTE) e validação cruzada estratificada, foram igualmente mantidas para cada cenário, de modo que as comparações sejam coerentes. Além disso, é importante destacar que a não exclusão de outliers preserva a integridade e a variabilidade das informações registradas em campo, mesmo que em outros contextos estes valores possam distorcer alguns resultados.

Um dos maiores desafios foi, sem dúvida, a discriminação das classes para o conjunto de teste. Por mais que tenham sido adicionadas modificações e incrementos, as medidas estatísticas ressaltam um comportamento similar para os 406 padrões utilizados nos contextos *default* e *HyperTuning*, uma vez que a Acurácia dos métodos oscila de 69,95% a 72,17%. Por outro lado, o uso de gráficos expõe que as variáveis que mais influenciam na decisão de uma árvore ser atacada, ou não, estão associadas aos conteúdos existentes na "data do inventário" (confirmando a importância temporal), a altura dos indivíduos e a faixa (o que indica para uma possível relação geográfica da ação dos macacos-prego em áreas distantes de estradas ou instalações).

Como o conteúdo deste trabalho é inovador para o tema, torna-se importante explorar outros modelos de AM, hiperparâmetros e técnicas complementares capazes de melhorar não só a qualidade das classificações como, também, em casos considerando a predição (como a mensuração de perdas de volume). Além disso, a inclusão de análises envolvendo dados climáticos passa a ser fundamental para a pesquisa, uma vez que as alterações no clima impactam diretamente na ação do Macaco-prego.

Agradecimentos

Os autores agradecem o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), processos: 310195/2022-5, 312688/2021-0 referentes a bolsa de produtividade em pesquisa PQ-2 e Projeto Universal 408164/2021-2, pelo financiamento desta pesquisa.

DOI: 10.59254/sbpo-2024-193805
https://proceedings.science/p/193805?lang=pt-br

Referências

Alckmin, G. T., Lucieer, A., Rawnsley, R. e Kooistra L. (2022) Perennial ryegrass biomass retrieval through multispectral UAV data. *Computers and Electronics in Agriculture*, 193, 106574.

Breiman, L. (2001). Random forests. *Machine learning*, 5:1, p. 5-32.

Chen, T. e Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Anais do 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).

Embrapa Florestas - Programa Macaco-prego. (2024). Embrapa Florestas. https://www.embrapa.br/florestas/programa-macaco-prego. Acessado: 05/05/2024.

Embrapa Florestas. (2024). Softwares, aplicativos e planilhas para manejo florestal. https://www.embrapa.br/florestas/transferencia-de-tecnologia/softwares-florestais. Acessado: 03/05/2024.

Freund, Y., Schapire, R. E. e Hill, M. (1996) Experiments with a new boosting algorithm rooms. 2B-428, 2A-424 g.

Galetti, M. e Pedroni, F. (1994). Seasonal diet of capuchin monkeys (Cebus apella) in a semideciduous forest in south-east Brazil. *Journal of Tropical Ecology*, 10:1, p. 27-39.

Hill, C. M. (2000). Conflict of interest between people and baboons: Crop raiding in Uganda. *International Journal of Primatology*. 21:2, p. 299-315.

Huang, T., Ou, G., Xu, H., Zhang, X., Wu, Y., Liu, Z. e Xu, C. (2023). Comparing Algorithms for Estimation of Aboveground Biomass in Pinus yunnanensis. *Forests*, 14:9, 1742.

IBÁ. (2020). Indústria Brasileira de Árvores. Relatório Anual 2020. https://iba.org/datafiles/publicacoes/relatorios/relatorio-iba-2020.pdf. Acessado: 03/05/2024.

Li, Y., Shen, Y., Zhang, W., Chen, Y., Jiang, H., Liu, M., Jiang, J., Gao, J., Wu, W., Yang, Z., Zhang, C. e Cui, B. (2021). OpenBox: A generalized black-box optimization service. Anais do 27th ACM SIGKDD *Conference on Knowledge Discovery & Data Mining* (pp. 3209-3219).

Liebsch, D., Mikich, S. B., Oliveira, E. B. e Moreira, J. M. M. A. P. (2015). Pinus taeda bark stripping by capuchin monkeys (Sapajus nigritus): type and intensity of damage and its impact on tree growth. *Scientia Forestalis*, 43:105, p. 37-49.

Mikich, S. B. e Liebsch, D. (2009). O Macaco-prego e os Plantios de *Pinus* spp. *Comunicado Técnico*. Embrapa Florestas. 234, (pp. 1-5). Colombo, PR.

Mikich, S. B. e Liebsch, D. (2014). Assessment of food supplementation and surveillance as techniques to reduce damage caused by black capuchin monkeys Sapajus nigritus to forest plantations. *Current Zoology*, 60:5, p. 581-590.

Mikich, S. B. e Liebsch, D. (2014b) Damage to forest plantations by tufted capuchins (Sapajus nigritus): Too many monkeys or not enough fruits? *Forest Ecology and Management*, 314, p. 9-16.

DOI: 10.59254/sbpo-2024-193805
https://proceedings.science/p/193805?lang=pt-br

Mikich, S. B., Liebsch, D., Almeida, A. D., Oliveira, E. B. e Moreira, J. M. M. A. P. (2015). O programa macaco-prego da Embrapa Florestas: avanços e perspectivas. *Anais do Congresso Florestal Paranaense*, 5, 2015, Curitiba. Novas tecnologias florestais, Associação Paranaense de Empresas de Base Florestal.

Miller, P. J., Mcartor, D. B. e Lubke, G. H. (2017). A gradient boosting machine for hierarchically clustered data. *Multivariate Behavioral Research*, v. 52:1, p. 117-117.

Miranda, E. N., Barbosa, B. H. G., Silva, S. H. G., Monti, C. A. U., Tng, D. Y. P. e Gomide, L. R. (2022). Variable selection for estimating individual tree height using genetic algorithm and random forest. *Forest Ecology and Management*, 504, 119828.

Miskell, G., Pattinson, W., Weissert, L. e Willams, D. (2019). Forecasting short-term peak concentrations from a network of air quality instruments measuring PM2. 5 using boosted gradient machine models. *Journal of Environmental Management*, 242, p. 56-64.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1, p. 81-106.

Santos, C. V., Morais Jr. M. M., Oliveira, M. M., Mikich, S. B., C. R. Ruiz-Miranda, C. R. e Moore, K. P. L. (2007). Ecologia, comportamento e manejo de primatas invasores e populações-problema. BICCA-MARQUES, J. C. (Ed.). *A Primatologia no Brasil*. EDIPUCRS, 10, p.101-118, Porto Alegre, RS.

Stoffels, A. e Soest, J. V. (1953). The main problems in sample plots. *Height regression*, 1, p. 190-199.

Watt, M. S., Palmer, D. J., Leonardo, E. M. C. e Bombrun, M. (2021). Use of advanced modelling methods to estimate radiata pine productivity indices. *Forest Ecology and Management*, 479, 118557.

DOI: 10.59254/sbpo-2024-193805
https://proceedings.science/p/193805?lang=pt-br