



Universidade Federal do Sul da Bahia - UFSB
Programa de Pós-Graduação em Ciências e Sustentabilidade - PPGCS

NATAN BRITO SANTOS

**SENSORIAMENTO REMOTO E APRENDIZADO DE MÁQUINA NA
IDENTIFICAÇÃO DE *Lecythis pisonis* EM FRAGMENTOS FLORESTAIS DA MATA
ATLÂNTICA, SUL DA BAHIA**

TEIXEIRA DE FREITAS - BA

2025

NATAN BRITO SANTOS

**SENSORIAMENTO REMOTO E APRENDIZADO DE MÁQUINA NA
IDENTIFICAÇÃO DE *Lecythis pisonis* EM FRAGMENTOS FLORESTAIS DA MATA
ATLÂNTICA, SUL DA BAHIA**

Dissertação submetida ao Programa de Pós-graduação em
Ciências e Sustentabilidade da Universidade Federal do Sul
da Bahia, como parte das exigências para obtenção do título
de mestre em Ciências e Sustentabilidade.

Orientador: Prof. Dr. Gerson Dos Santos Lisboa

Coorientadora: Dra. Viviane Maria Barazetti

Coorientador: Prof. Dr. João Batista L. da Silva

TEIXEIRA DE FREITAS - BA

2025

Dados Internacionais de Catalogação na Publicação (CIP)
Universidade Federal do Sul da Bahia
Sistema de Bibliotecas

S237s Santos, Natan Brito.
Sensoriamento remoto e aprendizado de máquina na identificação de *Lecythis pisonis* em fragmentos florestais da mata atlântica, sul da Bahia / Natan Brito Santos. Teixeira de Freitas, 2025.
80 f.

Dissertação (Mestrado) – Universidade Federal do Sul da Bahia, Campus Paulo Freire, Centro de Formação em Desenvolvimento Territorial, Programa de Pós-Graduação em Ciências e Sustentabilidade, 2025.
Orientador: Prof. Dr. Gerson dos Santos Lisboa.
Coorientadora: Prof. Dra. Viviane Maria Barazetti.
Coorientador: Prof. Dr. João Batista Lopes da Silva.

1. Sensoriamento remoto. 2. Aprendizado de máquina. 3. Sapucaia.
I. Título. II. Lisboa, Gerson dos Santos.

CDD – 910.285

NATAN BRITO SANTOS

**SENSORIAMENTO REMOTO E APRENDIZADO DE MÁQUINA NA
IDENTIFICAÇÃO DE *Lecythis pisonis* EM FRAGMENTOS FLORESTAIS DA MATA
ATLÂNTICA, SUL DA BAHIA**

Dissertação submetida ao Programa de Pós-graduação em
Ciências e Sustentabilidade da Universidade Federal do Sul
da Bahia, como parte das exigências para obtenção do título
de mestre em Ciências e Sustentabilidade.

Orientador: Prof Dr. Gerson Dos Santos Lisboa

Coorientadora: Dra. Viviane Maria Barazetti

Coorientador: Prof Dr. João Batista L. da Silva

Este trabalho foi submetido à avaliação e julgado aprovado em: 08/ Outubro/ 2025.

BANCA EXAMINADORA

Documento assinado digitalmente
 **EMANUEL ARAUJO SILVA**
Data: 16/11/2025 11:02:09-0300
Verifique em <https://validar.iti.gov.br>

Dr. EMANUEL ARAÚJO SILVA
UFSM

Documento assinado digitalmente
 **THIAGO FLORIANI STEPKA**
Data: 12/11/2025 11:35:26-0300
Verifique em <https://validar.iti.gov.br>

Dr. THIAGO FLORIANI STEPKA
UDESC

Documento assinado digitalmente
 **VICENTE TOLEDO MACHADO DE MORAIS JUNIO**
Data: 17/11/2025 15:50:56-0300
Verifique em <https://validar.iti.gov.br>

Dr. VICENTE TOLEDO MACHADO DE
MORAIS JÚNIOR
UFU

Documento assinado digitalmente
 **GERSON DOS SANTOS LISBOA**
Data: 12/11/2025 11:28:13-0300
Verifique em <https://validar.iti.gov.br>

ORIENTADOR
Dr. Gerson Dos Santos Lisboa
UFG

TEIXEIRA DE FREITAS - BA

2025

À minha filha, Ana Júlia, que me lembra quem sou,
e me ensina sempre a ver o mundo com novos olhos.

AGRADECIMENTOS

Em primeiro lugar, agradeço a Deus pela saúde e pela sabedoria que me sustentaram nesta trajetória.

E à minha esposa Marina Barbosa Souza, pela paciência, compreensão e apoio incondicional em todos os momentos.

Ao Programa *Arboretum*, pela parceria, disponibilidade e suporte fundamentais para a execução desta pesquisa.

Ao Professor Dr. Gerson Dos Santos Lisboa, pela orientação, dedicação e valiosas contribuições para o desenvolvimento deste trabalho.

À Dra. Viviane Maria Barazetti, pelo incentivo constante aos estudos.

À Derlyn Johana Solano López, pelos conselhos ao longo desta jornada.

A todos que, direta ou indiretamente, contribuíram para a concretização deste sonho, deixo aqui minha sincera gratidão.

“Não fomos feitos para temer o desconhecido, mas para explorá-lo.”

Inspirado em *Interestelar*

SANTOS, Natan Brito. **Sensoriamento remoto e aprendizado de máquina na identificação de *Lecythis pisonis* em fragmentos florestais da mata atlântica, sul da Bahia**. Orientador: Gerson Dos Santos Lisboa. 2025. Dissertação (Mestrado em Ciências e Sustentabilidade) – Universidade Federal do Sul da Bahia, Teixeira de Freitas, 2025.

RESUMO

Lecythis pisonis Cambess é uma espécie arbórea, que possui relevância tanto ecológica quanto econômica, sendo reconhecida pelo valor de suas sementes comestíveis e por aspectos de sua fenologia. Considerando essa importância, este estudo aplicou técnicas de aprendizado de máquina para identificar a ocorrência de *L. pisonis* em remanescentes florestais localizados no extremo sul da Bahia. A estratégia metodológica explorou a característica fenológica da renovação foliar com tonalidade rosada. Para identificar esse estágio fenológico, foram analisadas imagens multiespectrais do satélite Sentinel-2 adquiridas em outubro de 2020 e setembro de 2023, abrangendo cerca de 70 mil hectares nos municípios de Itamaraju, Guaratinga e Itabela. Para a classificação supervisionada, foram testados Random Forest, Support Vector Machine, K-Nearest Neighbors e dois ensembles votação (*voting*) e empilhamento (*stacking*). O treinamento considerou 300 pontos de referência (100 da espécie alvo e 200 de outras classes), aplicando-se um filtro que reteve agrupamentos de 3–4 pixels, compatíveis com copas de árvores adultas. A avaliação, realizada por meio de validação cruzada, indicou um desempenho de alta acurácia para os modelos supervisionados, com o índice Kappa variando entre 0,862 (RF) e 0,939 (atingido pelos modelos Stacking e SVM). A aplicação dos modelos revelou uma expressiva variação interanual, com o número de agrupamentos detectados pelo modelo Stacking, por exemplo, dobrando de 3.224 em 2020 para 6.490 em 2023. A análise espacial indicou uma forte tendência de concentração das detecções em altitudes mais baixas e próximas a cursos d'água, com os 35 pontos de maior consenso entre os modelos apresentando uma altitude mediana de apenas 21,8 metros. Conclui-se que a metodologia é tecnicamente viável para a detecção da *L. pisonis* baseada em sua fenologia, mas sua eficácia é influenciada pelas condições fenológicas da espécie em cada ano.

Palavras-chave: sapucaia; conservação; detecção de árvores.

SANTOS, Natan Brito. **Remote sensing and machine learning in the identification of *Lecythis pisonis* in forest fragments of the Atlantic Forest, southern Bahia**. Advisor: Gerson Dos Santos Lisboa. 2025. Dissertation (Master's in Science and Sustainability) – Universidade Federal do Sul da Bahia, Teixeira de Freitas, 2025.

ABSTRACT

Lecythis pisonis Cambess is a tree species that has both ecological and economic relevance, being recognized for the value of its edible seeds and aspects of its phenology. Given this importance, this study applied machine learning techniques to identify the occurrence of *L. pisonis* in forest remnants located in the extreme south of Bahia. The methodological strategy explored the phenological characteristic of pinkish leaf renewal. To identify this phenological stage, multispectral images from the Sentinel-2 satellite acquired in October 2020 and September 2023 were analyzed, covering approximately 70,000 hectares in the municipalities of Itamaraju, Guaratinga, and Itabela. For supervised classification, Random Forest, Support Vector Machine, K-Nearest Neighbors, and two ensembles: *voting* and *stacking* were tested. The training considered 300 reference points (100 of the target species and 200 of other classes), applying a filter that retained clusters of 3–4 pixels, consistent with the crowns of mature trees. The evaluation, performed through cross-validation, indicated high accuracy for the supervised models, with the Kappa index ranging from 0.862 (RF) to 0.939 (achieved by the Stacking and SVM models). The application of the models revealed significant interannual variation, with the number of clusters detected by the Stacking model, for example, doubling from 3,224 in 2020 to 6,490 in 2023. Spatial analysis indicated a strong tendency for detections to be concentrated at lower altitudes and near watercourses, with the 35 points with the highest consensus among the models having a median altitude of only 21.8 meters. It is concluded that the methodology is technically feasible for the detection of *L. pisonis* based on its phenology, but its effectiveness is influenced by the phenological conditions of the species in each year.

Keywords: sapucaia; conservation; tree detection.

LISTA DE FIGURAS

Figura 1 – Indivíduos da espécie de <i>L. pisonis</i> localizadas em Mucuri - BA.....	18
Figura 2 – Informações sobre a reflectância do espectro eletromagnético.....	21
Figura 3 – Fluxograma deste trabalho.....	26
Figura 4 – Delimitação da área de estudo.....	28
Figura 5 – Ocupação e uso da terra da área de estudo.....	29
Figura 6 – Modelo digital de elevação (MDE) da área de estudo.....	30
Figura 7 – Recorte da área de estudo com imagem Sentinel.....	31
Figura 8 – Matriz cadastrada pelo Programa <i>Arboretum</i>	33
Figura 9 – Esquematização K-fold Cross-Validation.....	36
Figura 10 – Imagem recortada com indivíduos de <i>L. Pisonis</i>	40
Figura 11 – Exemplo da aplicação da condicionante de tamanho.....	41
Figura 12 – Comparativo das métricas de classificação ao aplicar o modelo treinado.....	44
Figura 13 – Gráfico da distribuição por elevação, comparando os modelos em 2020.....	49
Figura 14 – Gráfico da distribuição por elevação, comparando os modelos em 2023.....	50
Figura 15 – Gráfico da distribuição por elevação, comparando os modelos em 2020-2023	50
Figura 16 – Mapa distribuição das detecções estáveis - altitude.....	51
Figura 17 – Gráfico da distribuição comparando os modelos em 2020.....	52
Figura 18 – Gráfico da distribuição comparando os modelos em 2023.....	53
Figura 19 – Gráfico da distribuição comparando os modelos em ambos anos 2020-2023.....	54
Figura 20 – Mapa distribuição das detecções estáveis - hidrografia.....	55
Figura 21 – <i>L. pisonis</i> confirmado em campo.....	61

LISTA DE TABELAS

Tabela 1 – Características das bandas do satélite Sentinel 2A.....	22
Tabela 2 – Estrutura da matriz de confusão.....	37
Tabela 3 – Sumário da classificação por algoritmo e ano.....	45
Tabela 4 – Sumário da análise de interseção de detecções estáveis.....	46
Tabela 5 – Comparativo das métricas de avaliação dos classificadores.....	47
Tabela 6 – Distribuição dos agrupamentos estáveis (2020-2023) por faixas de elevação.....	49
Tabela 7 – Distribuição dos agrupamentos estáveis por proximidade a cursos d'água.....	54
Tabela 8 – Comparativo dos resultados obtidos a partir da proximidade em APP (0-30m)	56
Tabela 9 – Resultados das detecções dos modelos e as classes de uso e cobertura do solo.....	57
Tabela 10 – Resultados da interseção entre as detecções estáveis dos modelos.....	58
Tabela 11 – Análise descritiva dos 276 pontos em comum.....	59
Tabela 12 – Análise descritiva dos 35 pontos em comum.....	60
Tabela 13 – Indivíduos de <i>L. pisonis</i> com dados biométricos coletados.....	61
Tabela 14 – Análise comparativa de desempenho com a literatura.....	63

LISTA DE ABREVIATURAS E SIGLAS

AHGN	Atlas Hidrogeomorfológico Nacional
ANA	Agência Nacional de Águas
APP	Área de Preservação Permanente
CAP	Circunferência a Altura do Peito
ESA	Agência Espacial Européia
KNN	K-Nearest Neighbors
ha	Hectare
INPE	Instituto Nacional de Pesquisas Espaciais
MDE	Modelo Digital de Elevação
RF	Random Forest
RGB	Sistema de cores aditivas (Red, Green, Blue) utilizado para a composição de imagens coloridas.
SVM	Support Vector Machine

SUMÁRIO

1. INTRODUÇÃO GERAL	14
2. OBJETIVO GERAL	16
2.1. Objetivos Específicos	16
2.2. Hipótese da Pesquisa.	16
3. FUNDAMENTAÇÃO TEÓRICA	17
3.1. <i>Lecythis pisonis</i> Cambess - Sapucaia	17
3.2. Sensoriamento remoto	19
3.3. Sentinel 2	22
3.4. Ferramentas computacionais para o tratamento de dados geoespaciais	23
3.5. Classificação supervisionada com aprendizado de máquina	24
3.5.1. Métodos de <i>ensemble</i> – votação e empilhamento	25
4. METODOLOGIA	26
4.1. Delineamento geral e área de estudo	27
4.2. Aquisição e pré-processamento de dados	30
4.3. Construção do conjunto de dados de referência	32
4.4. Classificação supervisionada com modelos individuais e métodos de <i>ensemble</i>	34
4.5. Avaliação de desempenho	35
4.6. Métricas baseadas na Matriz de Confusão	37
4.6.1. Precisão e Recall e F1 Score	37
4.6.2. Índice Kappa	38
4.7. Implementação da condicionante de tamanho	39
4.8. Análises espaciais	41
4.9. Análise da detecção por classes de cobertura do solo	43
4.10. Caracterização em campo e coleta de dados morfométricos	43
5. RESULTADOS	43
5.1. Análise de interseção e consistência dos modelos	46
5.2. Avaliação de desempenho dos classificadores com validação cruzada	47
5.3. Distribuição por elevação	48
5.4. Distribuição próximos a cursos hídricos	51
5.5. Distribuição em áreas de APP	55
5.6. Ocorrência de <i>L. pisonis</i> por classes de cobertura do solo	56
5.7. Análise da interseção dos modelos	59
5.8. Verificação e caracterização em campo	60
6. DISCUSSÃO	62
7. CONCLUSÕES	66
8. PERSPECTIVAS	67

9. REFERÊNCIAS	68
APÊNDICES	78
APÊNDICE A – DISPONIBILIZAÇÃO DE CÓDIGOS E DADOS	78
APÊNDICE B - TABELA COM GRUPOS E COORDENADAS IDENTIFICADOS NOS ANOS DE 2020 E 2023 E EM TODOS OS MODELOS.	78

1. INTRODUÇÃO GERAL

Representando um conjunto de recursos tecnológicos aplicados às ciências da Terra, as geotecnologias proporcionam avanços significativos na pesquisa, planejamento e gestão do espaço geográfico (Fitz, 2008). Dentre elas, o sensoriamento remoto se destaca por permitir a aquisição de informações precisas de áreas de difícil acesso e pela organização de grandes volumes de dados espaciais (Borlachenco, 2018; Santori, 2021). Sua técnica baseia-se no registro da radiação eletromagnética emitida ou refletida pela superfície, o que possibilita a análise das informações coletadas (Weiss et al. 2020).

Essa capacidade se mostra valiosa tanto para investigar a degradação de ecossistemas vegetais quanto para monitorar outros indicadores de transformações ambientais (Abad-segura et al., 2020). A aplicação de tais técnicas confere agilidade, eficiência e confiabilidade aos processos de avaliação.

No contexto florestal, a tecnologia é fundamental para o monitoramento de ecossistemas (Kennedy et al., 2007), abrindo caminho para abordagens mais refinadas, como o acompanhamento individual de árvores por meio de indicadores temporais, a exemplo das variações sazonais na coloração foliar.

Para transformar essas oportunidades em processos analíticos eficientes, Python é uma linguagem de programação que oferece recursos para a automação na manipulação de dados, otimizando o tratamento e a verificação de informações (Vicentainer; Mattedi; Mello, 2020). Essa capacidade de automação é, portanto, fundamental para viabilizar o mapeamento em larga escala de espécies florestais com sinais fenológicos claros.

Nesse sentido, espécies com características fenológicas distintivas que apresentam sinais detectáveis se tornam apropriadas para o mapeamento automatizado, como por exemplo a *L. pisonis*. Este fato, somado à sua importância ecológica e socioeconômica, justifica o esforço de monitoramento, por se tratar de uma espécie arbórea endêmica, com ocorrência na Amazônia, Mata Atlântica e Cerrado, em diferentes fitofisionomias (Mori & Prance, 1990; Amorim; Bicalho, 2022).

Socioeconomicamente, seu valor reside no comércio de suas sementes, utilizadas na alimentação humana. Suas sementes oleaginosas são uma fonte de alimento de alto valor nutritivo, sendo deliciosas como as da castanha-do-brasil (*Bertholletia excelsa*) (Carvalho, 2008). São consumidas por diversos povos indígenas, como os Juruna e Arara na Volta Grande do Xingu (Sartorelli et al., 2018), e comercializadas em mercados locais (Carvalho, 2008).

Ecologicamente, *L. pisonis*, possui interação significativa com a fauna, uma vez que suas sementes são dispersas por macacos e morcegos (Carvalho, 2008), servindo também como alimento para outros mamíferos (Sartorelli et al., 2018).

Essa relevância se consolida quando se analisa as grandes perdas de biodiversidade e cobertura florestal nas últimas décadas, causadas principalmente pela fragmentação (Silva, 2017), tornando monitoramento de suas populações prejudicado pela extensão territorial e pela inacessibilidade de fragmentos remanescentes.

Esse cenário de fragmentação contribui para uma incerteza sobre a situação populacional da espécie. Conforme a avaliação do Centro Nacional de Conservação da Flora por não haver documentação sobre declínios populacionais, foi classificada como "Menos Preocupante" (LC). A própria justificativa da avaliação ressalta que essa categoria se deve à falta de dados e demanda de ações de pesquisa como tendências e números populacionais, a fim de se ampliar o conhecimento disponível e garantir sua perpetuação na natureza no futuro (Amorim; Bicalho, 2022).

Essa necessidade é fortalecida pela inclusão da espécie em Planos de Ação Nacional (PAN) para a conservação da flora, como o do Rio de Janeiro e o da Serra do Espinhaço Meridional (Amorim; Bicalho, 2022). De fato, a utilização da fenologia como ferramenta para a biologia da conservação é uma abordagem estratégica para entender e proteger espécies e ecossistemas (Morellato et al., 2016), alinhando a presente pesquisa às prioridades de conservação nacionais

Dentro dessa realidade, o problema central desta pesquisa reside na ausência de um método eficaz e escalável para o mapeamento da espécie, o que dificulta a avaliação de seu real estado de conservação.

Para superar essa limitação, a presente pesquisa abordou análise espectral com imagens do satélite Sentinel-2, consistindo no processamento dessas imagens por meio de algoritmos de aprendizado de máquina, com a comparação entre cinco modelos, para identificar pixels correspondentes a árvores de *L. pisonis*.

2. OBJETIVO GERAL

Avaliar métodos de detecção remota para localizar a espécie *L. pisonis* em fragmentos florestais, no Sul da Bahia, com foco no período de renovação foliar.

2.1. Objetivos Específicos

- Identificar a localização de árvores de *L. pisonis* em imagens de satélite, visando mapear a sua distribuição espacial.
- Aplicar e comparar o uso de algoritmos de aprendizado de máquina, utilizando dados multiespectrais Sentinel-2, para a detecção de *L. pisonis*.
- Identificar e quantificar a distribuição dos indivíduos detectados em relação às diferentes classes de uso e cobertura do solo, avaliando sua ocorrência e a aplicabilidade do método de detecção.

2.2. Hipótese da Pesquisa.

A assinatura espectral distintiva de *Lecythis pisonis* durante seu período de renovação foliar é detectável de forma eficaz com o uso de imagens de satélite e aprendizado de máquina.

3. FUNDAMENTAÇÃO TEÓRICA

3.1. *Lecythis pisonis* Cambess - Sapucaia

Lecythis pisonis Cambess, popularmente conhecida como sapucaia, é uma espécie arbórea, pertencente à família Lecythidaceae. Seu nome, *Lecythis*, deriva do grego *lecythos* (panela de óleo), uma alusão ao seu fruto que se assemelha a uma panela com tampa (Mori & Prance, 1983), enquanto o epíteto específico *pisonis* é uma homenagem ao médico holandês Guilherme Piso (1611-1678), que chefiou uma missão científica no Brasil (Reitz, 1981). A espécie recebe ainda diversos outros nomes populares, como castanha-de-sapucaia no Amazonas, marmita-de-macaco na Bahia e sapucaia-vermelha no Espírito Santo.

Ela se estende além do bioma Mata Atlântica, ocorrendo também nos domínios fitogeográficos da Amazônia e do Cerrado (Smith et al., 2013). Sua distribuição natural abrange latitudes de 1°N no Amapá a 24°S em São Paulo, e altitudes que variam de 15 m na Paraíba a 900 m no Ceará (Carvalho, 2006). A espécie foi descrita em 1829, com ocorrência no estado do Espírito Santo (Cambesedes, 1829), e mais tarde foi registrada em vegetação de área antrópica, Floresta Estacional Semidecidual e Floresta Ombrófila Densa (Flora do Brasil, 2020).

Na região da Volta Grande do Xingu, sua presença é marcante em uma diversidade de habitats, incluindo floresta de terra firme, áreas de várzea, beira de grotas, beiradões, ilhas e igapós. Ecologicamente, é classificada como uma espécie secundária inicial que alcança o dossel superior ou emergente nas florestas primárias, ocorrendo também em florestas secundárias com 20 a 40 anos (Almeida & Vieira, 2001).

De acordo com Mori & Prance (1983), o gênero apresenta inflorescências de racemos ou panículas compostas, flores com pétalas e sépalas hexâmeras, às vezes com ductos mucilaginosos, androceu zigomorfo, capuz não enrolado, com apêndices que podem ou não conter anteras; ovário dividido em quatro lóculos; frutos geralmente deiscentes e sementes com arilo basal ou desprovidas de arilo.

As folhas, que medem de 5 a 18 cm de comprimento por 3 a 8 cm de largura (Carvalho, 2006), em seu estágio de desenvolvimento inicial, costumam apresentar tonalidades cremes, verde-claras ou avermelhadas, e verde quando adultas, conferindo à árvore uma beleza ornamental singular, especialmente durante a brotação (Figura 1) (Carvalho, 2006; Guimarães et al., 1993). As flores são violáceas, vistosas, com até 7 cm de diâmetro, odoríferas, carnosas e com numerosos estames amarelos (Guimarães et al., 1993).



Figura 1: Indivíduos da espécie de L. pisonis localizadas em Mucuri - BA. Registro fotográfico de campo mostrando a coloração rosada das folhas jovens, o principal alvo espectral da classificação.

(Fonte: Programa Arboretum, 2024)

Segundo Mori e Prance (1990), a altura pode atingir até 50 m e 150 cm de diâmetro à altura do peito, com um fuste que pode chegar a 15 m de comprimento. A madeira da sapucaia, por sua vez, é densa ($0,85$ a $1,00 \text{ g.cm}^{-3}$) e resistente, sendo utilizada para construções externas (postes, mourões) e internas (vigas, caibros, assoalhos) (Mainieri & Chimelo, 1989), e a casca possui propriedades medicinais, sendo usada na cultura popular no combate à tosse e como diurético (Soares, 1990).

A espécie é monóica (Carvalho, 2006) e fornece habitat e alimento a uma ampla gama de organismos (Sartorelli et al., 2018). Sua casca externa, de cor pardo-escuro e profundamente gretada-sulcada (Carvalho, 2006), serve como suporte para diversas plantas epífitas, como bromélias e orquídeas.

A polinização é realizada principalmente por abelhas de grande porte; estudos apontam a mamangava-de-toco (*Xylocopa frontalis*) como o provável agente polinizador (Mori & Orchard, 1979). A flor da sapucaia possui um sistema de polinização especializado, no qual

oferece um pólen diferenciado como recompensa: as abelhas coletam um tipo de pólen nutritivo, mas que não germina, para alimentar suas larvas e, durante essa atividade, o pólen fértil, localizado em outra parte da flor, adere às suas costas, garantindo a polinização cruzada na visita a outras flores (Mori, 1988; Ormond et al., 1981).

A floração ocorre, de modo geral, entre a primavera e o verão, variando conforme a localidade, com registros de agosto a janeiro na Bahia (Alvim, Alvim, 1978) e de julho a novembro no Espírito Santo. A dispersão de suas sementes é zoocórica, efetuada principalmente por morcegos e macacos, que consomem os arilos carnosos (Mori et al., 1983).

Uma das características mais notáveis de *L. pisonis* é seu fruto, um grande pixídio lenhoso e deiscente que pode medir de 12 a 30 cm de comprimento e até 18 cm de diâmetro, assemelhando-se a uma cumbuca com uma tampa (opérculo). No interior, contém de 10 a 40 sementes de 4 a 6 cm de comprimento, grandes e nutritivas, conhecidas como castanhas-de-sapucaia (Guimarães et al., 1993; Lorenzi, 2002; Mori; Prance, 1990). As sementes são usadas na alimentação humana há séculos, tendo grande apreço na culinária (Vallilo et al., 1998). Além do consumo humano, as sementes e flores são um recurso alimentar vital para a fauna; as sementes são consumidas por psitacídeos (WikiAves, 2025).

Um comportamento curioso de macacos jovens, e que por vezes ficam com a mão presa dentro do fruto ao tentar retirar as sementes, deu-se origem ao conhecido ditado popular "macaco velho não mete a mão em cumbuca" (Soares, 1990).

Fica evidente, portanto, que a relevância de *L. pisonis* vai além de seus usos materiais e se integra aos padrões culturais e ecológicos de seu ambiente, o que reforça a necessidade, já apontada por estudos anteriores, de se considerar os aspectos socioeconômicos e culturais na gestão dos recursos florestais da Mata Atlântica (Pinto; Roderjan; Galvão, 2006).

3.2. Sensoriamento remoto

O termo "Sensoriamento" define o processo de adquirir dados ou informações sobre algo usando um sensor. "Remoto" indica que essa aquisição ocorre à distância, sem contato direto com o objeto (Borges et al., 2015). O uso dessas ferramentas permite adquirir dados consistentes de observação da Terra a longo prazo, em múltiplas escalas, do local ao global (Camacho-Velasco, et al., 2015).

A captura de dados por sensoriamento remoto é realizada por meio de uma variedade de plataformas e sensores (Sheykhmousa et al., 2020). Em nível orbital, satélites como o Sentinel-2 fornecem imagens de alta resolução espacial e temporal, permitindo o monitoramento

contínuo de mudanças na superfície terrestre e da fenologia das plantas em grandes áreas (Francini et al., 2024; Wagner, 2021). Em nível aéreo, a coleta de dados é feita por meio de aeronaves tripuladas, que capturam imagens aéreas, e por Veículos Aéreos Não Tripulados (VANTs), utilizados para mapeamento de altíssima resolução de alvos específicos, como florestas urbanas ou espécies de árvores (Sheykhmousa et al., 2020; Cano; Marcato Junior, 2021; Francini et al., 2024).

Em cada um desses níveis, o processo envolve registrar como os alvos interagem com a radiação, o que inclui a medição da energia refletida e a análise das variações nos comprimentos de onda e nas diversas faixas espectrais (Steffen, 2025). Os sensores classificados como passivos necessitam de luz externa (geralmente solar), enquanto sensores ativos emitem sua própria luz. A energia resultante dessa interação é descrita como reflectância, absorbância ou transmitância (Costa, 2019).

O sensoriamento remoto por satélite demonstra ser uma alternativa confiável para o monitoramento de recursos naturais (INPE, 2017). Em aplicações florestais, ele possibilita a obtenção de informações da vegetação em larga escala (Costa, 2014), servindo como um método alternativo para estimar biomassa, mapear o desmatamento e identificar a distribuição de populações de árvores, a partir da análise da energia solar refletida ou absorvida pela cobertura vegetal (Ponzoni e Rezende, 2004). O monitoramento detalhado da cobertura florestal, por exemplo, tem se beneficiado enormemente do uso de imagens de alta resolução espacial (Silva; Oliveira; Costa, 2018). Recentemente, o acesso facilitado a imagens de maior resolução espacial, permite identificar e caracterizar objetos menores, com precisão antes restrita a outras plataformas (Conceição, 2023).

Segundo Moraes (2002), cada tipo de cobertura terrestre interage com a radiação de forma distinta, gerando um padrão de reflectância característico ao longo dos comprimentos de onda, conhecido como assinatura espectral. Moraes (2002), demonstra as diferenças marcantes entre as assinaturas de elementos como água, solo e vegetação (Figura 2).

A vegetação saudável, por exemplo, absorve fortemente a luz visível (azul e vermelho, pela clorofila) e reflete intensamente no infravermelho próximo (pela estrutura celular), permitindo sua distinção de outros alvos e, potencialmente, entre diferentes tipos vegetais ou espécies.

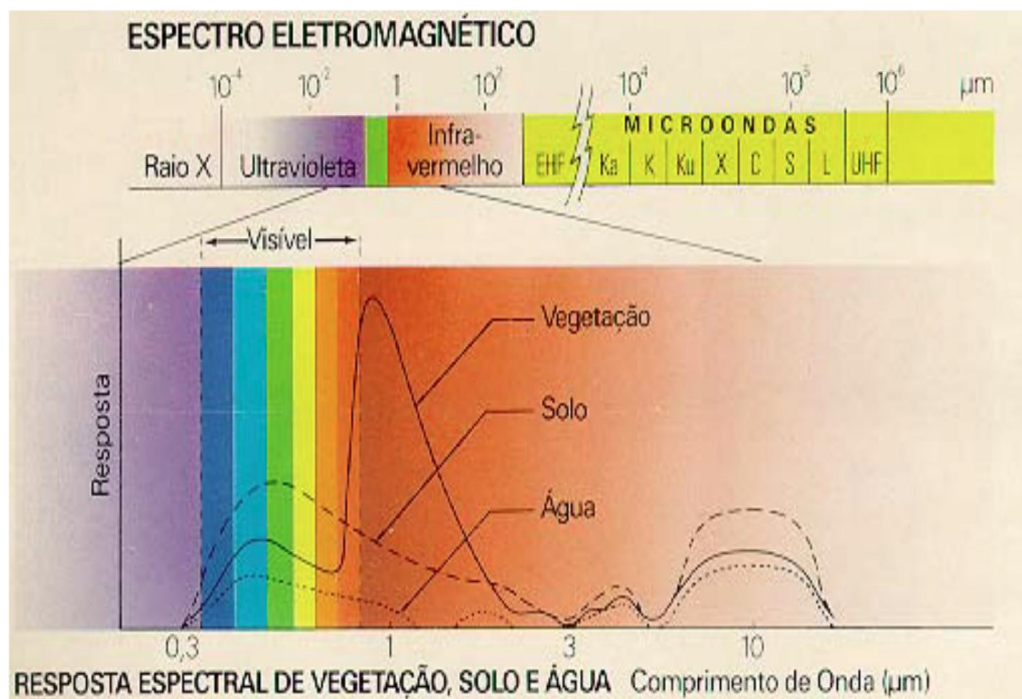


Figura 2: Informações sobre a reflectância de alguns objetos dentro do espectro eletromagnético. A figura demonstra como a reflectância da vegetação varia ao longo do espectro, justificando o uso de bandas específicas do satélite.

(Fonte: Moraes 2002)

Essas assinaturas espectrais podem variar devido a mudanças na estrutura foliar associadas à fenologia, ou seja, às fases do ciclo de vida das plantas como brotação, floração e senescência, influenciadas pelo teor de água e pigmentos alteram a interação com a luz.

O sensoriamento remoto multitemporal explora essas variações temporais. As informações captadas pelos sensores são armazenadas na forma de imagens digitais. Estas são constituídas por uma grade regular de elementos mínimos, os pixels, cada um representando uma área específica na superfície terrestre (Boscolo, Namikawa, 2024).

A cada pixel é atribuído um valor numérico, ou número digital, que corresponde à intensidade média da energia eletromagnética (refletida ou emitida) naquela área, dentro de uma faixa espectral específica. O poder dessa abordagem se revela quando múltiplas bandas são capturadas simultaneamente: combiná-las nos permite criar composições coloridas e, mais importante, desvendar padrões que o olho nu não veria na paisagem (Zhao et al., 2021; Boscolo, Namikawa, 2024).

3.3. Sentinel 2

Segundo a ESA e o relatório Copernicus (2015), o programa Sentinel-2, componente do programa Copernicus da União Europeia, tem como principal característica o fornecimento de dados ópticos. Estes se distinguem pela alta qualidade, gratuidade, disponibilidade global e por cobrirem diferentes comprimentos de onda. A captação desses dados multiespectrais é realizada pela constelação Sentinel-2, formada pelos satélites 2A e 2B, que organizam os dados em variadas bandas espectrais. As bandas Visíveis (azul B2, verde B3, vermelho B4) e juntamente com a banda do infravermelho próximo (B8) possuem uma resolução nominal de 10 metros.

As bandas de transição do vermelho (Red-Edge) e as de infravermelho de ondas curtas (SWIR) são fornecidas com 20 metros de resolução. Já as bandas de 60 metros são otimizadas para aplicações que envolvem a caracterização atmosférica e a observação de zonas litorâneas. (Copernicus, 2015).

Conforme indicado na Tabela 1, as bandas do Sentinel utilizadas neste estudo apresentam várias coberturas de comprimentos de onda e resolução. Essas bandas, combinadas, permitem uma variedade de aplicações em monitoramento ambiental, agrícola, florestal e urbano. (Copernicus, 2015).

Banda	Faixa Espectral (nm)	Faixa Espectral Mínima (nm)	Faixa Espectral Máxima (nm)	Resolução Espacial (m)	Aplicação Principal
Banda 2 (B2)	490	463	511	10	Mapeamento de vegetação, identificação de tipos de solo e detecção de poluição (verde)
Banda 3 (B3)	560	534	586	10	Mapeamento de vegetação, monitoramento de cobertura terrestre e análise de uso da terra (vermelho)
Banda 4 (B4)	665	653	673	10	Avaliação da saúde da vegetação, detecção de áreas urbanas e monitoramento de recursos hídricos (vermelho próximo)
Banda 8 (B8)	842	789	896	10	Detecção de composição do solo, mapeamento de vegetação e monitoramento de mudanças no uso da terra (infravermelho de curto alcance)

Tabela 1: Características das bandas do satélite Sentinel 2A. A tabela detalha a resolução espacial, o comprimento de onda e a designação de cada banda utilizada no estudo (Fonte: Copernicus - 2015)

3.4. Ferramentas computacionais para o tratamento de dados geoespaciais

No mesmo sentido ao sensoriamento remoto a computação tem feito um progresso significativo no desenvolvimento das geociências, permitindo que ferramentas inteligentes, elaboradas para ajudar a manipular vastos conjuntos de informações de base espacial e auxiliar em análises e identificações complicadas, tenham se tornado mais acessíveis (Grohmann, 2004; Kobal et al., 2013; Omuto & Vargas, 2014; Metcalfe et al., 2015; Brown, 2016).

Dentre elas o Python é uma das linguagens considerada promissora no campo de programação computacional científica, principalmente pela sua simplicidade, facilidade de aprendizado e aplicabilidade (McKinney, 2018). Outro ponto relevante é que ela se destaca por disponibilizar um grande número de pacotes numéricos e bibliotecas para fins específicos. Novos pacotes com funcionalidades aprimoradas são constantemente lançados e desenvolvidos por uma comunidade ativa (Python, 2024).

Para operacionalizar as análises em sensoriamento remoto, a linguagem Python tem se destacado, permitindo segmentação e análise espectral por meio de bibliotecas específicas, das quais se destacam Numpy, Rasterio, Scipy, Scikit-Learn e Geopandas. Dentre as bibliotecas, voltadas para o geoprocessamento, pode-se citar o Numpy e o Rasterio devido à sua eficiência e grande capacidade de processamento de dados. Essas bibliotecas permitem automatizar todo o processo de segmentação na identificação das árvores e permitindo reduzir significativamente o tempo necessário para a análise (McKinney, 2018; Virtanen et al., 2020; Gillies 2024).

De acordo com Gillies (2024), o criador da biblioteca Rasterio, sua especificidade é o geoprocessamento e a manipulação de dados raster, com interface simples e eficiência para ler e escrever dados rasters em vários formatos, permitindo o acesso aos metadados atribuídos aos registros de dados raster, como informações de projeção e valores de pixel.

Para manipulação de dados numéricos a biblioteca Numpy permite a execução de operações numéricas em dados geoespaciais em matrizes multidimensionais (Numpy, 2024) possibilitando também realizar funções para álgebra linear, transformações de Fourier e geração de números aleatórios, tornando-a uma escolha ideal para processamento de dados científicos e análise numérica (Walt, Colbert, Varoquaux, 2011).

3.5. Classificação supervisionada com aprendizado de máquina

A classificação de padrões é uma tarefa comum na análise de imagens de sensoriamento remoto, com o objetivo de associar cada pixel a uma categoria temática com base nas características espectrais (Mather, 2004). Realizar essa classificação manualmente com grandes volumes de dados por área seria extremamente custoso e demorado. Por essa razão, algoritmos de aprendizado de máquina são amplamente utilizados, pois conseguem identificar padrões, processar diversos tipos de dados e automatizar decisões com mínima intervenção humana (Maxwell et al., 2018; Li; Stein, 2020). Neste estudo, uma análise comparativa abrangente foi conduzida, empregando cinco abordagens distintas: três classificadores individuais, *Random Forest* (RF), *Support Vector Machine* (SVM) e *K-Nearest Neighbors* (KNN) e duas técnicas de *ensemble* (*Voting Classifier* e *Stacking Classifier*), que combinam as predições dos modelos individuais para aprimorar a precisão final.

O *Random Forest* (RF) opera construindo múltiplas árvores de decisão durante a fase de treinamento. Cada árvore é gerada com base em uma subamostra aleatória dos dados de referência, e cada ponto de decisão considera apenas um subconjunto aleatório das variáveis. Essa abordagem de conjunto resulta em um modelo final mais confiável e preciso (Breiman, 2001). No contexto do sensoriamento remoto com dados multiespectrais de satélite, como os do Sentinel-2, a aplicação conjunta do RF permite identificar padrões fenológicos das espécies ao longo do tempo (Francini et al., 2024).

O *Support Vector Machine* (SVM) é um classificador que melhor separa as classes, sendo reconhecido por sua eficácia em dados de alta dimensionalidade (Mironczuk; Hościło, 2017). O SVM opera buscando o hiperplano ótimo que melhor separa as diferentes classes de dados (Sheykhmousa et al., 2020), o que o torna particularmente poderoso para lidar com a variabilidade espectral encontrada em imagens de sensoriamento remoto (Cano; Marcato Junior, 2021).

Diversos estudos confirmam o sucesso do SVM na classificação de cobertura arbórea e uso do solo com dados Sentinel-2, alcançando altas acurácias (Mironczuk; Hościło, 2017; Sheykhmousa et al., 2020). A comparação entre RF e SVM é relevante, pois não há um consenso sobre a superioridade universal de um sobre o outro, com o desempenho dependendo das características específicas dos dados e da paisagem (Sheykhmousa et al., 2020).

O *K-Nearest Neighbors* (KNN) é um dos algoritmos de aprendizado supervisionado mais intuitivos e amplamente utilizados, classificado como um método baseado em instância ou de aprendizado preguiçoso ou *lazy learning*, pois não constrói um modelo interno, mas

armazena todo o conjunto de treinamento para realizar a classificação (Altman, 1992). O funcionamento do KNN é classificar uma nova amostra de dados com base na classe majoritária de seus 'k vizinhos mais próximos no espaço de atributos.

A proximidade é geralmente calculada por meio de uma métrica de distância, como a distância euclidiana. A escolha do número de vizinhos a serem considerados, e que define o hiperparâmetro 'k', para o desempenho do modelo: valores pequenos de 'k' podem tornar o modelo sensível a ruído, enquanto valores grandes podem simplificar demais a fronteira de decisão (Cover & Hart, 1967). Uma variação comum e eficaz, é a ponderação por distância *weights='distance'*, na qual o voto de cada vizinho é ponderado pelo inverso de sua distância, conferindo maior influência aos vizinhos mais próximos (Dudani, 1976). Por ser um algoritmo baseado em distância, o desempenho do KNN é sensível à escala dos dados, tornando o pré-processamento de padronização das feições uma etapa fundamental para sua correta aplicação.

3.5.1. Métodos de *ensemble* – votação e empilhamento

Métodos de *Ensemble* ou conjunto de algoritmos são técnicas que buscam melhorar a performance preditiva e a confiabilidade de modelos de aprendizado de máquina através da combinação estratégica de múltiplos classificadores (Sagi & Rokach, 2018). A premissa fundamental é que um comitê de modelos, mesmo que individualmente fracos ou com diferentes vieses, pode tomar uma decisão coletiva mais acurada do que qualquer um de seus membros isoladamente (Dietterich, 2000).

Uma das abordagens de *ensemble* mais diretas, é a *Voting Classifier* ou Classificador por Votação. Ela agrega as predições de múltiplos classificadores de base e define a classe final com base em uma regra de combinação. Na modalidade *hard voting*, a classe com o maior número de votos é a vencedora. Na modalidade *soft voting*, utilizada nesta pesquisa, as probabilidades de classe previstas por cada modelo são coletadas, ponderadas por pesos pré-definidos, e a classe com a maior probabilidade média ponderada é selecionada. O *soft voting* é frequentemente superior, pois considera a confiança de cada modelo em sua própria predição (Sagi & Rokach, 2018).

O *Stacking Classifier* ou Empilhamento ou Generalização Empilhada é uma técnica de *ensemble* mais sofisticada que introduz um sistema de aprendizado em dois níveis (Wolpert, 1992). No primeiro nível, um conjunto de classificadores como RF e SVM é treinado sobre o conjunto de dados original. Em seguida, as predições geradas por esses modelos de base são

utilizadas como um novo conjunto de atributos para treinar um classificador de segundo nível, conhecido como meta-modelo, combinando as previsões dos modelos de base, corrigindo seus erros e explorando suas forças complementares para produzir uma classificação final mais precisa, como a regressão logística, cuja função é aprender a melhor forma de ponderar as previsões dos modelos de base (Breiman, 1996).

4. METODOLOGIA

Para identificar e mapear a espécie arbórea *Lecythis pisonis*, esta pesquisa implementou e avaliou um fluxo metodológico específico (Figura 3). Todo o processamento e as análises foram realizados no ambiente Python (versão 3.13, IDLE), empregando bibliotecas científicas e geoespaciais especializadas.

A totalidade do código, incluindo os scripts de pré-processamento, classificação, análise espacial e exportação, está disponível publicamente no repositório GitHub (APÊNDICE A). Esse fluxo de trabalho foi estruturado em etapas sequenciais, que abrangem desde a definição da área e coleta de dados até o pós-processamento dos resultados, a validação estatística e a caracterização em campo (Figura 3).

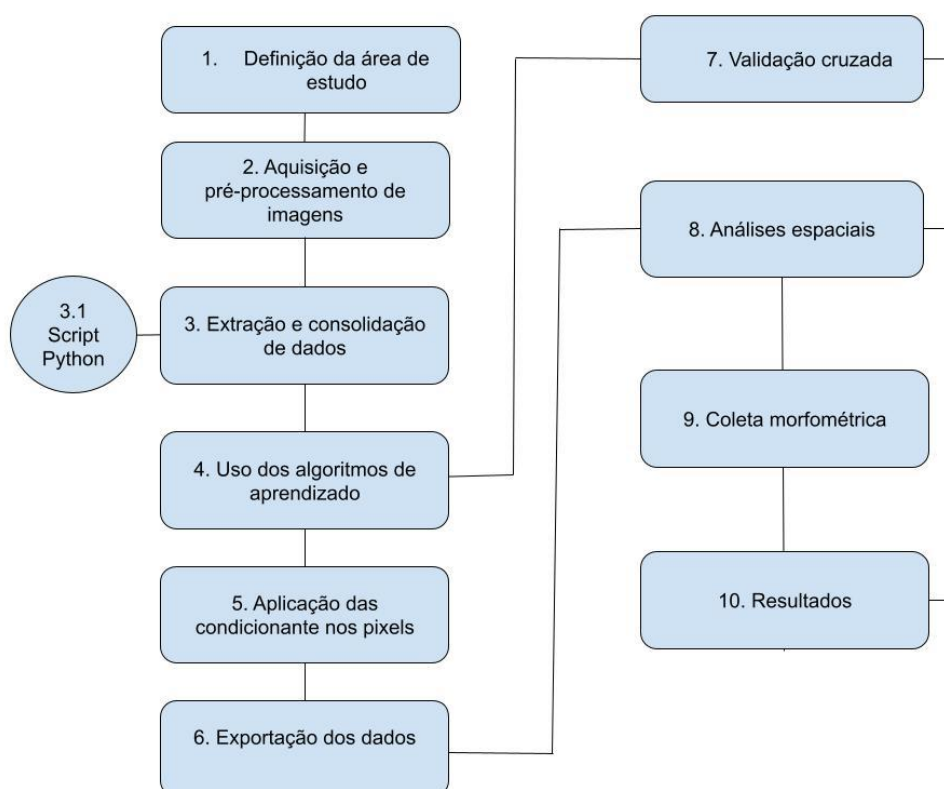


Figura 3: Fluxograma operacional da metodologia com os processos executados. Representação visual de todas as etapas metodológicas da pesquisa, desde a obtenção dos dados até a análise dos resultados.

(Fonte: Autor, 2025)

A abordagem consistiu na segmentação de imagens e na análise espectral, com foco em imagens capturadas pelo satélite Sentinel-2 durante a fase de renovação foliar. As imagens foram processadas por análise comparativa por cinco algoritmos de aprendizado de máquina, utilizando três classificadores individuais: *Random Forest* (RF), *Support Vector Machine* (SVM) e *K-Nearest Neighbors* (KNN), e duas abordagens que combinam algoritmos, denominada *ensemble*, com *Voting Classifier* (Votação) e *Stacking Classifier* (Empilhamento), para identificar grupos de pixels correspondentes a árvores de *L. pisonis*. Com essa abordagem foi possível realizar a classificação em si, e a automação do fluxo de trabalho, a construção do script e a otimização da análise (Vicentiner; Mattedi; Mello, 2020).

Após a classificação, foram avaliados a confiabilidade e a estabilidade temporal do método. A eficácia dos algoritmos foi mensurada por meio da técnica de validação cruzada K-Fold, aplicada a um conjunto de dados de referência coletados em campo. Foram calculadas métricas de desempenho, como o Índice Kappa e a Acurácia Balanceada com as amostras do treinamento, para quantificar a precisão dos modelos.

Para complementar a identificação remota, realizaram-se visitas de campo, nas quais foram coletados dados morfométricos das árvores, como altura, circunferência à altura do peito (CAP). Esses dados de campo foram utilizados para caracterizar os indivíduos detectados remotamente e entender a capacidade do modelo em identificar árvores de grande porte, fornecendo subsídios para futuras ações de manejo e conservação.

4.1. Delineamento geral e área de estudo

A área de estudo localiza-se no Território de Identidade do Extremo Sul da Bahia, uma das porções mais significativas e ameaçadas do bioma Mata Atlântica no Brasil. Historicamente, a economia local foi impulsionada pelo extrativismo de madeira, mas hoje é dominada por duas atividades principais: a pecuária extensiva e a silvicultura de eucalipto em larga escala, que abastece a indústria de celulose. (SEPLAN, 2016).

Os limites específicos da análise foram definidos por um recorte vetorial (shapefile) englobando fragmentos florestais de interesse, totalizando aproximadamente 70.000 hectares (Figura 4). Entre os três municípios analisados, Itamaraju apresenta a maior área com 50.371,45 hectares. Em seguida, Itabela possui uma área de 17.660,95 hectares. Por fim, o município de Guaratinga tem a menor extensão territorial do grupo, com 1.694,535 hectares.

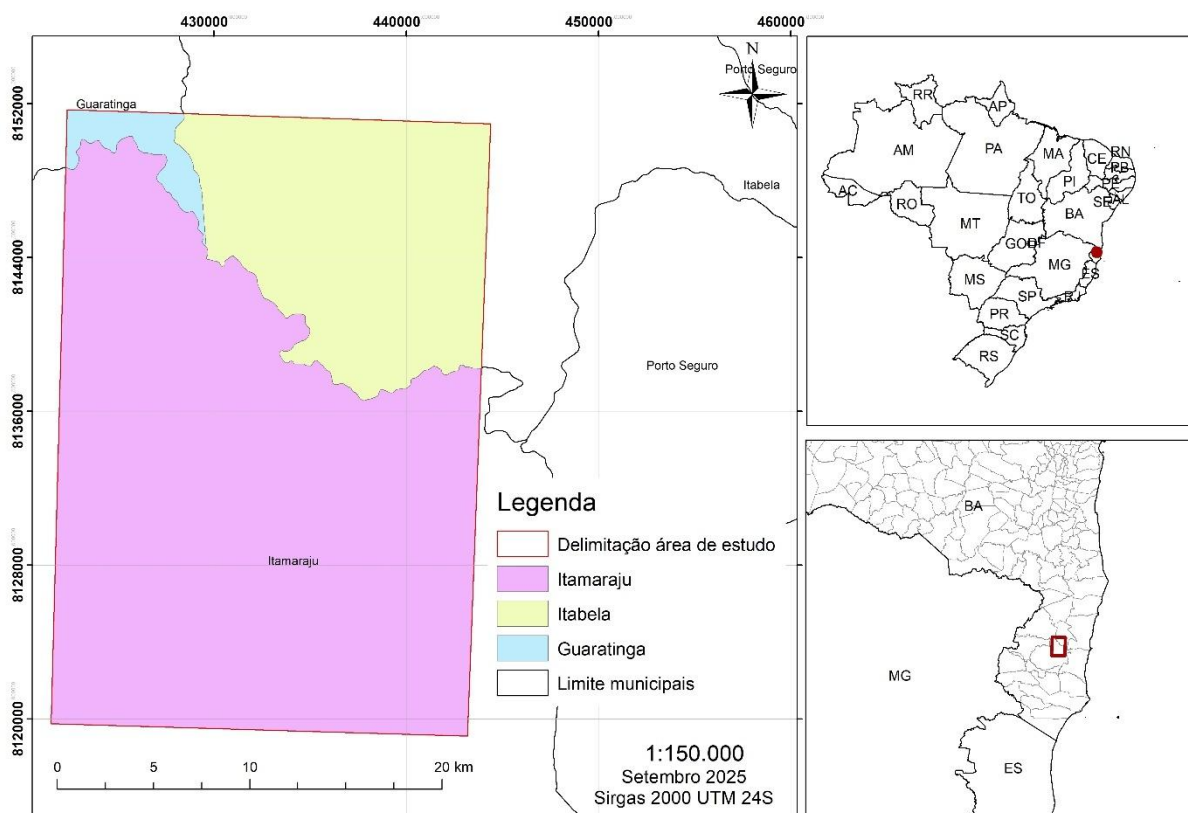


Figura 4: Delimitação da área de estudo da pesquisa. Mapa de localização apresentando os limites geográficos da área de estudo no contexto regional do Extremo Sul da Bahia.

(Fonte: Autor, 2025)

Para contextualizar a área de detecção de *L. pisonis*, foi utilizado o mapeamento de uso e cobertura da terra realizado pelo Fórum Florestal do Sul da Bahia para o ano de 2022, (Fórum Florestal Bahia, 2024) (Figura 5). Este mapeamento revela um mosaico complexo composto por diferentes fisionomias e usos. As classes predominantes neste recorte são a floresta ombrófila densa em estágio médio de regeneração (aproximadamente 28.396 ha), representando os remanescentes florestais mais conservados, e pastagens, tanto limpas quanto sujas (somando cerca de 32.458 ha). Outras classes relevantes incluem floresta em estágio inicial de regeneração (cerca de 3.252 ha), comunidades aluviais (matas de baixada,

aproximadamente 2.282 ha), plantios de eucalipto (cerca de 1.172 ha), afloramentos rochosos (aproximadamente 806 ha) e café (cerca de 600 ha). Classes como agricultura perene, cacau em sistema de cabruca, instalações rurais, seringal, sistema viário e áreas urbanas ocupam porções menores da área.

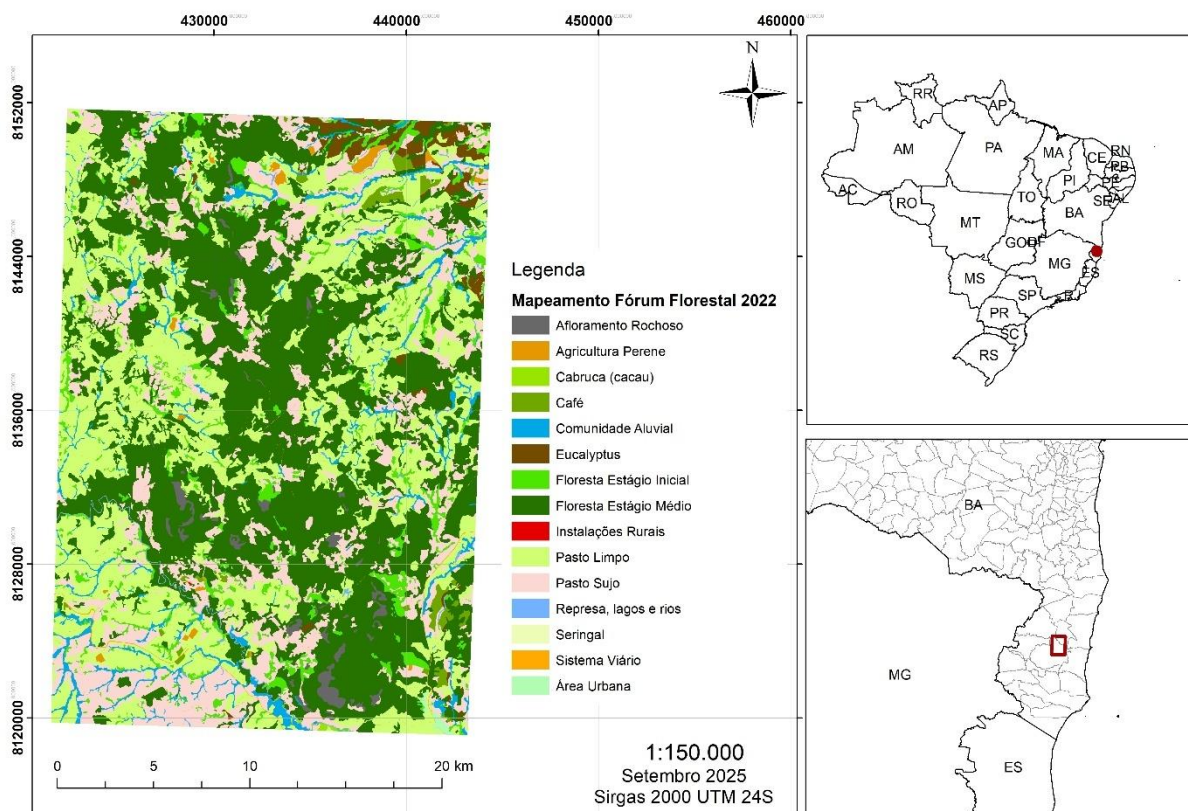


Figura 5: Ocupação e uso da terra da área de estudo da pesquisa. Mapa temático que caracteriza a paisagem da área de estudo, destacando as principais classes de uso e cobertura da terra. (Fonte: Autor, 2025)

A caracterização da paisagem evidencia que a detecção de *L. pisonis* ocorre em fragmentos florestais inseridos em uma matriz diversificada, e não em um ambiente homogêneo. Nesse caso os classificadores precisam distinguir a assinatura espectral da *L. pisonis* não apenas de outras espécies arbóreas nativas dentro das florestas, mas também de todas as demais classes de uso e cobertura presentes, incluindo pastagens, eucaliptos e áreas de regeneração. A complexidade, portanto, reside em isolar um sinal fenológico em meio a um mosaico variado.

A área de estudo apresenta um relevo variado, com altitudes que se estendem da elevação de 5 metros até um máximo de 683 metros. Esta caracterização topográfica foi obtida a partir do Modelo Digital de Elevação (MDE) (Figura 6) derivado de dados do sensor ALOS PALSAR, com resolução espacial de 12,5 metros (ASF, 2025). Essa variação altimétrica define

diferentes ambientes na paisagem, influenciando diretamente a distribuição das classes de uso e cobertura do solo analisadas.

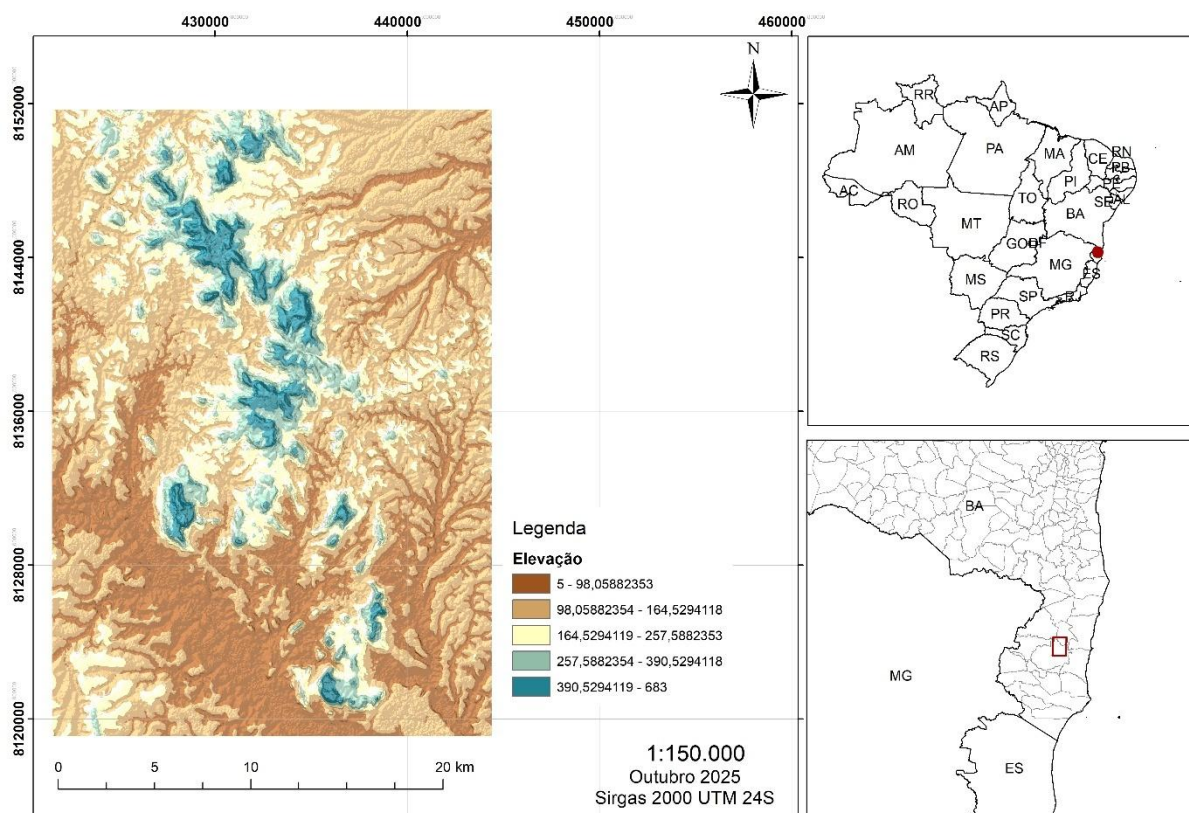


Figura 6: Modelo Digital de Elevação (MDE) da área de estudo, com variação altimétrica entre 5 e 683 metros, derivado de dados ALOS PALSAR (resolução espacial de 12,5 m). (Fonte: Autor, 2025).

4.2. Aquisição e pré-processamento de dados

A utilização exclusiva de sensores ópticos, em especial das bandas RGB e do infravermelho próximo, foi fundamentada na capacidade desses sensores de capturar as alterações de coloração associadas à fenologia da espécie estudada sendo perceptível nas bandas do visível e do infravermelho, além da disponibilidade e simplicidade na aquisição e no processamento dos dados. (Moraes, 2002; Ponzoni, Rezende, 2004).

Foram utilizadas imagens multiespectrais do Sentinel-2 devido à sua alta disponibilidade temporal, sendo selecionado o produto Nível 2A (L2A), que oferece reflectância de superfície com correção atmosférica, ideais para análises temporais. As imagens, em formato JPEG 2000 (.JP2), foram obtidas pelo Copernicus Open Access Hub. A seleção das cenas específicas foi guiada por dois critérios principais:

1. Intervalo temporal fenológica: buscou-se imagens adquiridas durante o pico do período de renovação foliar da *L. pisonis* na região de estudo, que ocorre tipicamente entre os meses de setembro e novembro. Neste período, a espécie apresenta uma característica fenológica alvo, conforme descrito por (Guimarães et al., 1993), que oferece um intervalo temporal favorável à detecção ideal, uma vez que altera temporariamente a sua assinatura espectral.

2. Qualidade da imagem: foram selecionadas cenas com um máximo de 30% de cobertura pelo próprio portal do Copernicus, para minimizar a obstrução da visibilidade da superfície e garantir a qualidade dos dados de reflectância visando maximizar a probabilidade de captura da espécie em sua fase de máxima expressividade fenológica.

As imagens específicas utilizadas no estudo, ambas pertencentes ao Tile Sentinel-2 T24KVG, foram datadas de 07 de outubro de 2020 e 27 de setembro de 2023, (Figura 7). Para a tarefa de classificação, foram selecionadas quatro bandas espectrais do Sentinel-2 que possuem resolução espacial de 10 metros, a mais alta disponível para este sensor (Tabela 1).

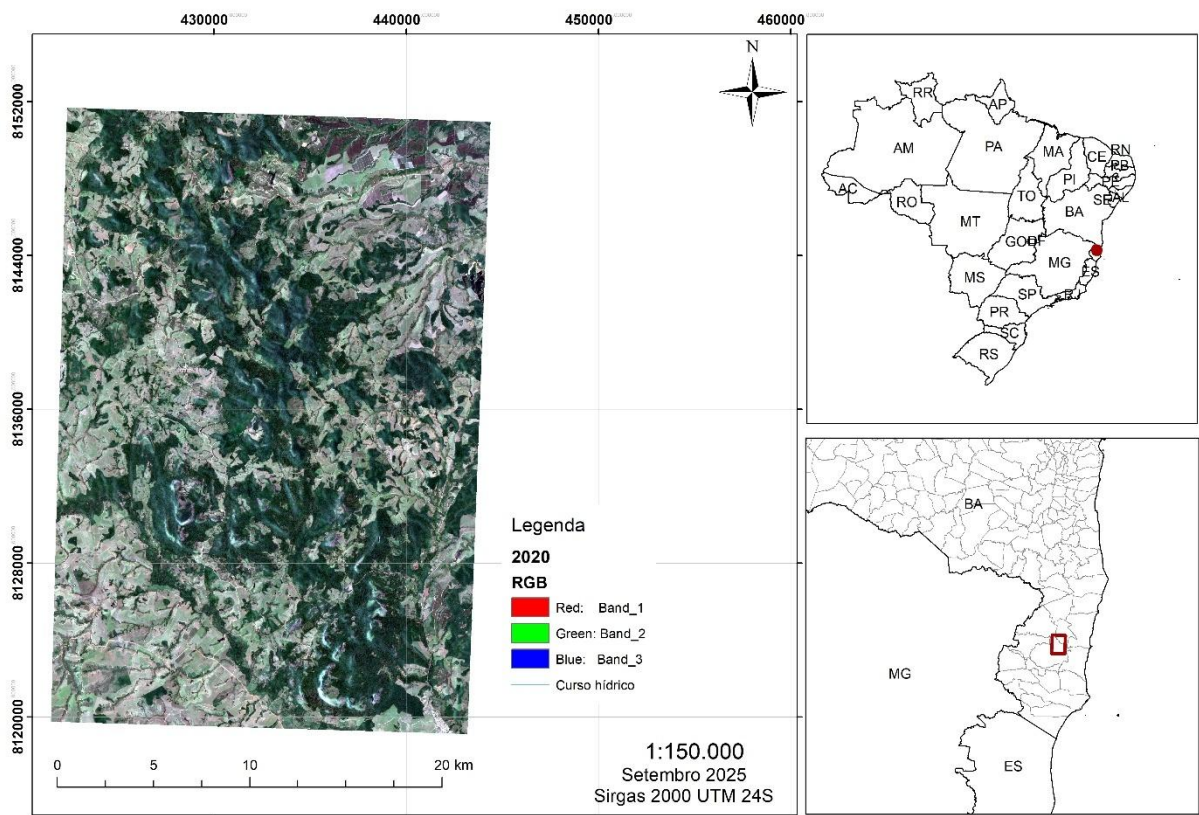


Figura 7: Recorte da área de estudo com imagem Sentinel. Visualização da imagem de satélite utilizada, destacando a aparência da vegetação na área. (Fonte: Autor, 2025)

A escolha dessas datas específicas e a ausência de dados de anos anteriores e intermediários, como 2021 e 2022, ocorreu devido as dificuldades encontradas na aquisição de imagens. Durante esse período, a região de estudo foi impactada pela atuação do fenômeno La Niña, que intensificou a umidade na região e contribuiu para o aumento da formação de nuvens e chuvas no Nordeste brasileiro (NOAA, 2025). Essa condição restringiu a disponibilidade de imagens com a visibilidade necessária, o que comprometeu a possibilidade de uma análise temporal contínua.

4.3. Construção do conjunto de dados de referência

Após a aquisição das imagens e seu pré-processamento, foi necessário construir um conjunto de dados de referência para treinar os classificadores. Dados de referência com boa qualidade e representatividade são importantes para o sucesso de modelos de aprendizado de máquina supervisionado (Lanaras et al., 2018). Isso é feito comparando a classificação do modelo com dados de referência independentes ou *ground truth* (Congalton & Green, 2009).

Para este estudo, a base de treinamento e validação foi construída a partir de um arquivo vetorial com 300 localizações geográficas amostrais, distribuídas na área analisada.

Os pontos da classe-alvo, *L. pisonis*, foram extraídos de dados de matrizes já cadastradas e georreferenciadas, cedidos pelo Programa *Arboretum* de Conservação e Restauração da Diversidade Florestal (Figura 8). Já os pontos da classe *Não Alvo* foram gerados a partir da criação de pontos aleatórios sobre o mapeamento de uso e cobertura do solo de 2022, produzido pelo Fórum Florestal do Sul da Bahia (Fórum Florestal Bahia, 2022).

A amostragem da classe *Não Alvo* foi distribuída sobre as principais fisionomias da paisagem para garantir que os modelos fossem treinados para distinguir a assinatura espectral da *L. pisonis* de outras classes de cobertura na imagem. As classes específicas sobre as quais estes pontos aleatórios incidiram foram: afloramento rochoso, floresta em estágio médio, pasto sujo e represas, lagos e rios.

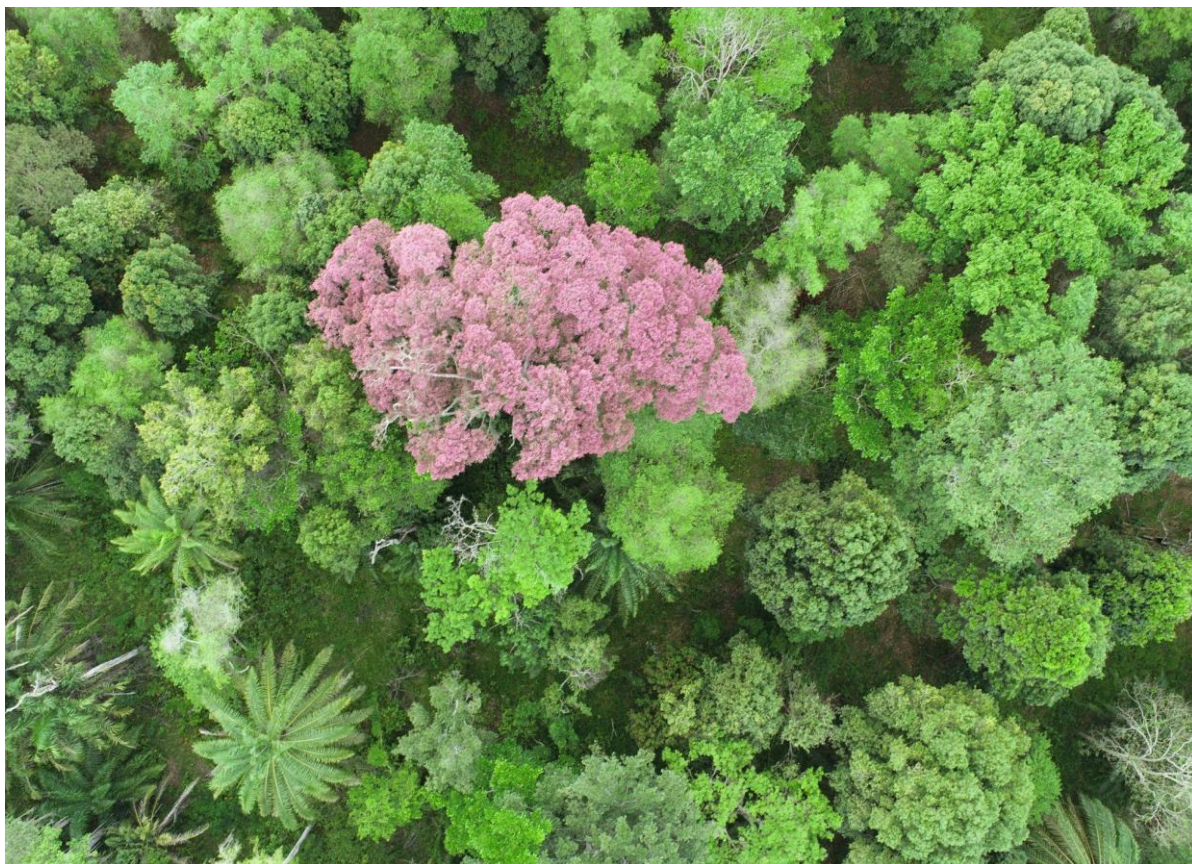


Figura 8: Matriz cadastrada pelo Programa Arboretum. Exemplo de dado de campo cedido pelo Programa Arboretum, confirmando a ocorrência da espécie na área de estudo.

(Fonte: Programa Arboretum, 2025)

A preparação dos dados de referência iniciou-se com a importação do arquivo vetorial, ajustando seu sistema de referência de coordenadas (CRS) para coincidir com o das imagens Sentinel-2. No processamento, os pontos da espécie-alvo receberam o rótulo numérico *1 Alvo*, enquanto todos os pontos pertencentes às outras classes foram agrupados na categoria *Não Alvo*, recebendo o rótulo *0*.

Essa organização dos dados gerou um conjunto final com 100 exemplos da Classe 1 *L. pisonis* e 200 da Classe 0 *Não Alvo*, caracterizando um desbalanceamento inerente de 1 para 2, uma proporção que reflete a realidade de campo onde a espécie alvo é menos abundante que o conjunto de outras coberturas.

Para mitigar o risco de o modelo ser enviesado em favor da classe majoritária, o parâmetro *class_weight=balanced* foi utilizado. Essa configuração ajusta automaticamente os pesos das amostras durante o treinamento, conferindo maior peso à classe minoritária para compensar o desbalanceamento nos dados.

Para cada um dos 300 pontos de referência, os respectivos valores de reflectância nas quatro bandas espectrais (B2, B3, B4 e B8) foram extraídos tanto da imagem de 2020 quanto da imagem de 2023. As amostras de ambos os anos foram então consolidadas em um único conjunto de dados, totalizando 600 registros. O objetivo dessa consolidação foi criar um modelo mais eficiente, capaz de reconhecer a assinatura espectral da *L. pisonis* considerando a variabilidade interanual. Esse arquivo conjunto de dados tabular (*arrays X_all_final e y_all_final*) formou a base para o treinamento supervisionado e a validação cruzada dos modelos.

4.4. Classificação supervisionada com modelos individuais e métodos de *ensemble*

Visando a classificação pixel a pixel da área estudada para detectar a *L. pisonis*, o presente trabalho empregou e comparou cinco algoritmos de aprendizado de máquina. A análise incluiu três classificadores individuais, sendo o *Random Forest* (RF), que para este estudo, utilizou-se a implementação *RandomForestClassifier* com 100 árvores $n\_estimators=100$ (Scikit-Learn, 2025).

O *Support Vector Machine* (SVM), que utilizou o *SVC* do Scikit-learn com um kernel de base radial $kernel=rbf$, adequado para relações não-lineares. E o *K-Nearest Neighbors* (KNN), algoritmo que classifica um pixel com base na classe majoritária entre seus 'k vizinhos mais próximos no espaço de atributos (Altman, 1992). Foi utilizada a implementação *KNeighborsClassifier* com $n\_neighbors=5$ e o parâmetro $weights=distance$, que atribui maior peso aos vizinhos mais próximos, refinando a predição.

E os métodos de *ensemble* foram implementados para investigar se a combinação das forças dos classificadores individuais poderia levar a um resultado mais confiável. Conforme descrito por Sagi & Rokach (2018), tais técnicas frequentemente superam o desempenho de modelos únicos. Foram testadas duas arquiteturas:

- *Voting Classifier* (Votação Ponderada): Combinou as predições de RF e SVM por meio de *soft voting*. Nessa modalidade, as probabilidades de classe previstas por cada modelo são ponderadas e a classe final é definida pela maior probabilidade média. Os pesos para o *Voting Classifier* foram definidos como 0.6 para o SVM e 0.4 para o RF, refletindo o desempenho individual superior do SVM nas métricas de validação cruzada, notadamente no Índice Kappa e na precisão para a classe *Alvo*.

- *Stacking Classifier* (Empilhamento): Utilizou-se uma abordagem de *ensemble*, onde as previsões dos modelos de base (RF e SVM) são usadas como entrada para treinar um classificador final, que neste caso, uma Regressão Logística (*LogisticRegression*) que aprende a melhor forma de combinar as previsões (Wolpert, 1992).

Para garantir a comparabilidade e mitigar o desbalanceamento de classes nos dados de treinamento, o hiperparâmetro *class_weight=balanced* foi ativado em todos os classificadores supervisionados. Adicionalmente, para os algoritmos baseados em distância (SVM e KNN), que são sensíveis à escala dos dados, foi aplicada uma etapa de padronização das feições com *StandardScaler*. A escolha dos modelos, se deu por serem reconhecidos por sua eficácia em sensoriamento remoto (Sheykhmousa et al., 2020), e busca para identificar o classificador mais preciso, e avaliar a eficácia de técnicas de combinação de modelos para esta tarefa (Gomes et al., 2016).

O treinamento de cada classificador foi realizado uma única vez, utilizando o conjunto de dados combinado de 600 amostras (representando os anos de 2020 e 2023). Para garantir a reprodutibilidade dos resultados, foi adotada uma semente de aleatoriedade fixa (*random_state = 42*) no algoritmo *Random Forest*. A escolha do valor 42 segue uma convenção amplamente utilizada na literatura de ciência de dados, sem impacto direto no desempenho do modelo, mas com função essencial na padronização dos experimentos.

A avaliação de desempenho dos classificadores foi conduzida por meio de validação cruzada estratificada do tipo *K-Fold*, com $k = 5$, procedimento que divide o conjunto de dados em cinco subconjuntos balanceados entre as classes, alternando-os entre treino e teste. Essa configuração foi escolhida por proporcionar um equilíbrio adequado entre a estabilidade estatística das métricas e a eficiência computacional do processo de validação (Pacheco, 2021).

4.5. Avaliação de desempenho

Para avaliar o desempenho dos cinco, adotou-se o método de Validação Cruzada K-Fold Estratificada. Essa abordagem foi aplicada sobre o conjunto de dados combinado com as amostras de 2020 e 2023, para gerar um conjunto único de métricas que refletisse a capacidade de generalização de cada modelo diante da variabilidade temporal.

A escolha por este método é justificada por sua superioridade em relação a uma divisão simples entre treino e teste *Hold-Out*. A validação cruzada é uma técnica de reamostragem que

avalia a capacidade de generalização de um modelo em um conjunto de dados limitado, fornecendo uma estimativa de desempenho mais robusta (Refaeilzadeh; Tang; Liu, 2009). Ao aproveitar todos os dados de referência tanto para treinar quanto para validar o modelo ao longo de k iterações, o método atenua o risco de um resultado enviesado que poderia surgir de uma única partição dos dados (Kohavi, 1995; Pacheco, 2021).

O processo de validação cruzada K-Fold consiste em segmentar o conjunto de dados em k subconjuntos ou *folds* de tamanhos aproximados. O treinamento do modelo é então repetido k vezes. A cada iteração, um *fold* é reservado para teste, enquanto os $k-1$ *folds* restantes são utilizados para o treinamento, garantindo que cada amostra seja usada para teste exatamente uma vez. A abordagem estratificada é um refinamento para dados desbalanceados como os deste estudo, pois assegura que a proporção das classes *Alvo* e *Não Alvo* seja mantida em cada um dos *folds*. Para a implementação, utilizou-se a classe StratifiedKFold da biblioteca Scikit-learn. O número de subconjuntos foi definido como $5 \text{ } K\_FOLD\_N\_SPLITS = 5$, resultando em uma divisão de 70% dos dados para treinamento e 30% para validação a cada iteração (Figura 9) (Fassnacht et al. 2014).

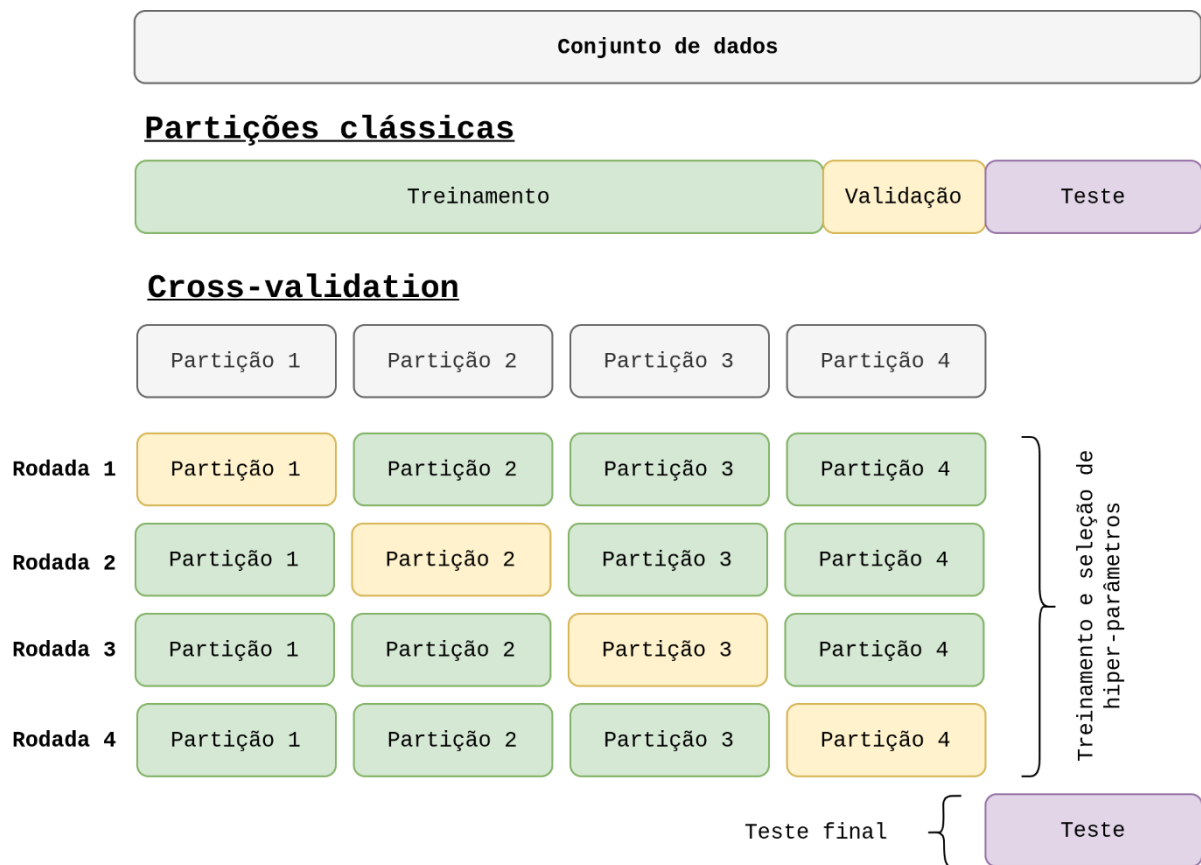


Figura 9: Esquematização K-fold Cross-Validation. Representação esquemática dos processos de validação empregados para garantir a confiabilidade estatística dos modelos.

(Fonte: Pacheco, 2021)

4.6. Métricas baseadas na Matriz de Confusão

A matriz de confusão é uma ferramenta importante para essa análise. Conforme descrito por Pedregosa et al. (2011), ela fornece uma tabela onde se pode visualizar quantas previsões foram feitas correta e incorretamente para cada classe. Esta visualização permite identificar quais classes estão sendo confundidas entre si, ajudando a diagnosticar pontos onde o método pode apresentar dificuldades. Sendo utilizada para avaliar a precisão de modelos de classificação (Aghdam; Jahani-Habibi, 2017). Com a matriz de confusão podemos calcular a acurácia como mostra a tabela 2.

Predição/ Real	Positivo	Negativo
Positivo	TP	FP
Negativo	FN	TN

Tabela 2: Estrutura da Matriz de Confusão. A tabela ilustra a estrutura da matriz, definindo os componentes usados para calcular as métricas de desempenho dos modelos.

(Fonte: Hamed Habibi Aghdam, 2017)

4.6.1. Precisão e Recall e F1 Score

Precisão e Recall (Hamed Habibi Aghdam, 2017) são duas medidas quantitativas importantes para avaliar um classificador. A Precisão calcula a fração de amostras previstas positivas e Recall a fração de amostras realmente positivas conforme demonstrado nas equações abaixo:

$$\text{precisão} = \frac{TP}{TP+FP}$$

$$\text{recall} = \frac{TP}{TP+FN}$$

Em um classificador perfeito FP e FN devem ser zero para que precisão e recall sejam 1, neste caso podemos dizer que é um classificador perfeito. Caso precisão e recall sejam próximos de zero significa que o classificador é impreciso (Jahani-Habibi, Aghdam, 2017).

F1 Score é uma métrica eficiente para avaliar modelos de classificação, utilizando a combinação de precisão e recall conforme definido pela Equação 23 (Aghdam; Jahani-Habibi, 2017). O valor de $F1$ varia em $[0,1]$ sendo 1 o resultado de um classificador perfeito.

$$F1 = \frac{2}{\frac{1}{\text{precisão}} + \frac{1}{\text{recall}}} = \frac{2TP}{2T+FP+FN}$$

Brodersen et al. (2010) descrevem as métricas a Acurácia Balanceada e a Acurácia Geral como importantes para avaliação de algoritmos. A Acurácia Balanceada é a média aritmética da sensibilidade para cada classe, sendo recomendada para conjuntos de dados desbalanceados por atribuir o mesmo peso ao desempenho de cada classe sendo utilizada por meio da função *balanced_accuracy_score*. A Acurácia Geral, por sua vez, corresponde à proporção de classificações corretas em relação ao total de amostras, e pode ser obtida pela função *accuracy_score* do Scikit-learn.

4.6.2. Índice Kappa

Segundo Cohen (1960) O índice Kappa de Cohen mede o grau de concordância entre as previsões do modelo e as classes reais, corrigido para a concordância que poderia ocorrer ao acaso, sendo utilizado para medir a concordância entre a máscara automática e a manual. É definido como a fórmula abaixo:

$$k = \frac{p_o - p_e}{1 - p_e}$$

Para entender essa métrica, é importante considerar dois componentes principais:

1. Precisão observada (p_o): Esta é a proporção de casos em que o modelo fez previsões que concordam com as classes reais. Em outras palavras, representa a exatidão efetiva do modelo ao identificar corretamente as categorias reais.
2. Precisão esperada (p_e): Esta é a proporção de casos em que a concordância entre o modelo e as classes reais poderia ser esperada apenas por acaso. É uma estimativa da concordância que ocorreria sem que o modelo contribuísse de fato para a acurácia.

Ainda de acordo com Cohen (1960), quando o índice de Kappa atinge um valor de 1, é a combinação perfeita entre a previsão do modelo e a classe real. Se o Kappa estiver em 0, isso indica que a correspondência entre o modelo e a classe de observação é casual. Isso significa que o desempenho do modelo não excede o esperado das previsões aleatórias. Um valor negativo para Kappa indica que o modelo produz previsões menos confiáveis do que as obtidas com uma seleção completamente aleatória.

4.7. Implementação da condicionante de tamanho

Os resultados da classificação inicial gerados pelos cinco algoritmos (RF, SVM, KNN, Voting e Stacking) embora baseados em modelos estatísticos, podem conter imperfeições espaciais, como pixels isolados classificados incorretamente (Figura 10). Para refinar o resultado, aumentar a confiabilidade das detecções e focar em padrões espaciais consistentes como copas de árvores, foi aplicada uma condicionante espacial de pós-processamento sobre o resultado dos cinco classificadores, implementado via funções da biblioteca *Scipy*.

A etapa consistiu na utilização da função *label*, da biblioteca *scipy.ndimage*, que identifica e rotula componentes conectados em uma matriz (Scipy, 2023). Inicialmente, gerou-se uma máscara booleana contendo apenas os pixels classificados como *L. pisonis*.

Em seguida, a função *label* foi aplicada considerando conectividade de 8 vizinhos, de modo que pixels fossem considerados parte de um mesmo grupo quando compartilhassem uma borda ou um vértice, independentemente da posição, direção ou orientação espacial dos pixels, permitindo que grupos com formato irregular, inclusive em padrão diagonal, como a letra “L” sejam identificados como uma única entidade contínua. O resultado foi uma matriz com rótulos inteiros para cada grupo conectado, além do número total de grupos detectados.

Após essa identificação, calculou-se o tamanho de cada grupo. Apenas os que apresentaram entre 3 e 4 pixels foram mantidos, critério definido para representar a assinatura espacial de um indivíduo de *L. pisonis* de grande porte (Figura 11). Essa filtragem segue prática comum no pós-processamento de imagens classificadas para redução de ruído (Mather, 2004).

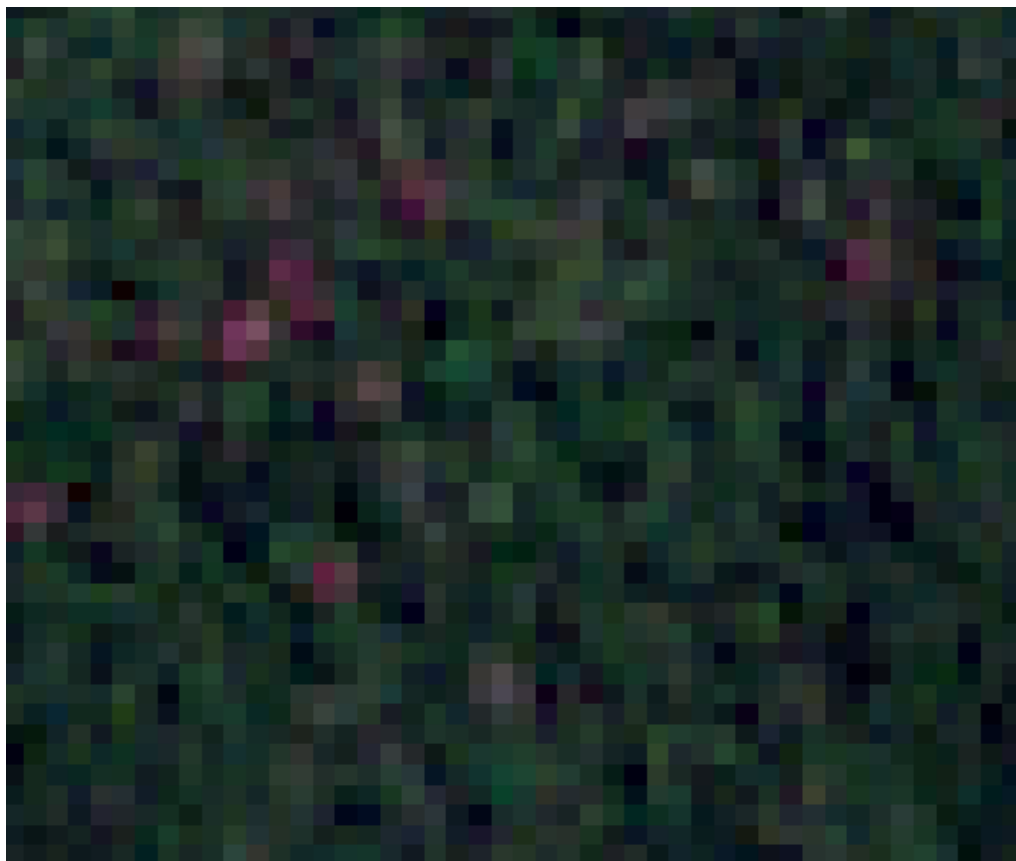


*Figura 10: Imagem recortada com indivíduos de *L. pisonis*. A imagem ilustra como os indivíduos de *L. pisonis* aparecem na imagem Sentinel-2, servindo como base para a classificação. (Fonte: Copernicus, 2025)*

A escolha desse critério se justifica pelo fenômeno do pixel misto, no qual a reflectância de pixels isolados, especialmente em bordas de copas, tende a ser uma combinação do sinal da árvore e de elementos vizinhos (Lu et al., 2003). Em imagens de resolução espacial de 10 metros, um único pixel isolado pode representar apenas uma pequena porção da copa ou ser um falso positivo gerado por essa mistura espectral, incluindo sombra, solo ou folhagem de outras espécies. Exigir pequenos agrupamentos garante, portanto, que a assinatura espectral da espécie seja dominante e mais representativa da copa.

O uso de pixels conectados como unidade de análise é respaldado por estudos com dados Sentinel-2. Wagner (2021), por exemplo, demonstrou que a detecção eficaz de árvores de gênero *Pleroma* ocorreu quando estas formavam manchas contínuas, evidenciando que, na escala de 10 m, eventos fenológicos raramente se manifestam em um único pixel, mas em padrões espaciais agregados.

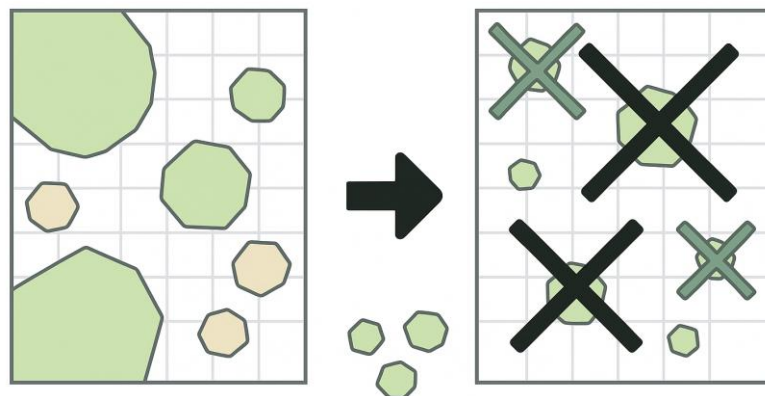


Figura 11: Exemplo da aplicação da condicionante de tamanho. Ilustração do passo de pós-processamento, onde a condicionante é aplicada para remover ruídos e reter apenas agrupamentos significativos.

(Fonte: Autor, 2025)

Ao exigir um agrupamento contíguo de 3 a 4 pixels, o método busca identificar o centro de uma copa de grande porte, onde o sinal espectral de *L. pisonis* é mais puro e dominante. Dessa forma, a condicionante atua como um filtro espacial, conectando o pixel misto das bordas à detecção de um sinal mais confiável no núcleo da copa. Isso minimiza a influência de pixels periféricos e aumenta a confiança de que cada detecção final corresponde a um indivíduo arbóreo maduro, e não a ruídos de classificação.

Portanto, a condicionante não apenas reduz ruídos, mas constitui uma premissa metodológica para a detecção de árvores individuais compatíveis com a resolução do sensor. Os pixels que não atenderam ao critério foram reclassificados como *Não Alvo*, resultando em um raster final contendo exclusivamente agrupamentos válidos da classe *L. pisonis*.

4.8. Análises espaciais

Após a validação estatística, os resultados do mapa de classificação final pós-condicionante espacial foram preparados para análises espaciais e interpretação. O mapa *raster* foi exportado como GeoTIFF. Os pixels fora da área de análise receberam o valor 255 para representar a ausência de dados. Os pixels classificados como *L. pisonis* também foram convertidos em pontos vetoriais (representando o centro de cada pixel detectado) e exportados como um arquivo ESRI Shapefile. Este shapefile incluiu atributos informativos para cada

ponto, como o ano de referência (*ano*), e as coordenadas originais da imagem (*lin_img*, *col_img*).

A distribuição espacial destes pontos/grupos detectados como *L. pisonis* foi então analisada em relação a variáveis ambientais relevantes, buscando identificar possíveis padrões ou preferências de nicho ecológico. Especificamente, analisou-se a frequência das detecções em diferentes faixas de:

- **Altitude:** Com a utilização do Modelo Digital de Elevação (MDE) derivado dos dados do sensor ALOS PALSAR, com resolução espacial de 12,5 metros, disponibilizado pela Alaska Satellite Facility (ASF, 2025). O MDE foi processado para extrair a altitude de cada ponto de detecção utilizando o método de interpolação por vizinho mais próximo.
- **Proximidade a Cursos Hídricos:** A análise da relação espacial das ocorrências da *L. pisonis* com os sistemas de drenagem foi realizada a partir da base de dados do Atlas Hidrogeomorfológico Nacional (AHGN). Esse atlas, derivado da integração da Base Hidrográfica Atlas-Estudos (BHAE) e do Modelo Digital de Elevação ANADEM, (ANA, 2021) que oferece informações detalhadas sobre a rede de drenagem, seus atributos hidrogeomorfológicos. O AHGN foi utilizado para a delimitação da rede hidrográfica, possibilitando compreender a distribuição das detecções em relação à sua proximidade dos cursos d'água.
- **Análise de Sobreposição com Áreas de Preservação Permanente (APPs):** Para avaliar a ocorrência da espécie dentro das faixas de APPs hídricas, conforme a legislação. Adotou-se como critério de APP a faixa de 30 metros a partir da margem dos cursos d'água, conforme definido pela Lei nº 12.651/2012 (Código Florestal) para rios com menos de 10 metros de largura. Utilizando a base de dados vetorial da rede de drenagem (AHGN) e os pontos de detecção da *L. pisonis*, foi calculada a distância de cada ponto até o curso d'água mais próximo. Foram contabilizados todos os pontos cuja distância ao rio era menor ou igual a 30 metros, sendo estes considerados como localizados dentro da APP. O resultado foi expresso em número absoluto de detecções e em percentual sobre o total para cada cenário (cada modelo e ano), permitindo uma avaliação quantitativa da associação da espécie com as matas ciliares legalmente protegidas.

4.9. Análise da detecção por classes de cobertura do solo

Para compreender a distribuição espacial das diferentes classes de uso e cobertura do solo, foi realizado o cruzamento dos dados de detecção gerados pelos modelos com mapeamento de uso e cobertura do solo. Inicialmente, os *shapefiles* de pontos detectados por cada modelo (*KNN*, *RF*, *Stacking*, *SVM* e *Voting*) para os anos de 2020 e 2023 foram integrados, via operação de interseção geográfica, ao *shapefile* de referência contendo as classes de cobertura do solo categorizadas no campo *NIVEL_III*.

Para cada ponto de detecção intersectado, foi extraída a informação da classe de uso do solo correspondente. Em seguida, realizou-se a contabilização das ocorrências de detecções para cada combinação de modelo, ano e classe. A frequência de detecções por classe foi quantificada para cada modelo e ano.

4.10. Caracterização em campo e coleta de dados morfométricos

Foram selecionados pontos geográficos correspondentes às detecções dos indivíduos para verificação em campo. Devido à vasta extensão da área de estudo e às restrições de acesso, a seleção dos pontos para verificação priorizou locais próximos a vias e acessíveis logisticamente. Em campo, a navegação até as coordenadas foi realizada com um receptor GNSS (modelo Garmin Etrex 30), onde a equipe procedeu à confirmação visual da presença ou ausência da espécie.

Para os locais onde a presença da espécie foi confirmada, realizou-se a coleta de dados morfométricos dos indivíduos encontrados. Utilizando fita métrica, foi medido circunferência à altura do peito (CAP), altura total da árvore, além de avaliações visuais sobre a posição sociológica da copa, tipo de fuste, fitossanidade geral e vigor da copa.

Essas informações permitem identificar árvores de destaque, como matrizes porta semente, fundamentais para estratégias de coleta, manejo e conservação genética. A coleta desses dados fornece, ainda, subsídios para futuras análises sobre a relação entre as dimensões das árvores e sua detecção por sensoriamento remoto (Costa et al., 2014).

5. RESULTADOS

Nesta seção, são apresentados os resultados da classificação e das análises espaciais. É importante ressaltar que, para fins de interpretação, cada agrupamento de 3 a 4 pixels que atendeu à condicionante espacial é considerado a representação de um indivíduo arbóreo

detectável. Portanto, os totais quantitativos e as análises subsequentes referem-se a esses agrupamentos, que funcionam como um indicador de árvores dominantes de dossel visível.

Com base na metodologia proposta, os resultados a seguir demonstram a eficácia da abordagem para detectar a *L. pisonis*. Esta seção apresenta os resultados quantitativos da classificação (Figura 12), a análise da distribuição espacial das detecções e os achados da caracterização em campo (Tabela 3). Os resultados quantitativos, obtidos após a aplicação da condicionante, revelaram diferenças de desempenho tanto entre os algoritmos quanto entre os anos.

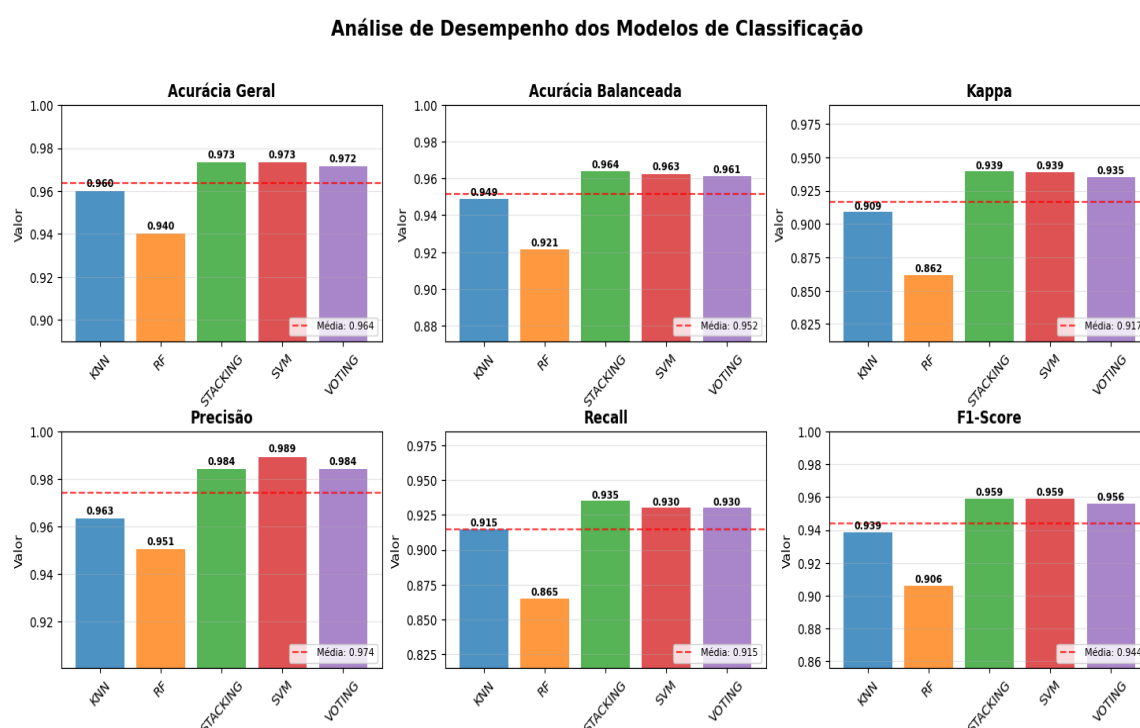


Figura 12: Comparativo das métricas de classificação ao aplicar o modelo treinado nos dados de 2020 e 2023. (Fonte: Autor, 2025)

Para o ano de 2020, o *Random Forest* (RF) foi o modelo que identificou o maior número de agrupamentos (5.517), seguido pelo KNN (3.897), *Stacking* (3.224), *Voting* (2.993) e, por último, o SVM (2.646). Em 2023, observou-se um aumento no número de detecções para todos os modelos, com o KNN, identificando 11.010 agrupamentos, seguido de perto pelo RF, com 10.182. Os modelos de *ensemble*, *Stacking* e *Voting*, apresentaram um número intermediário de detecções, com 6.490 e 6.044 grupos de pixels, respectivamente. O SVM novamente foi o mais conservador, com 6.014 grupos

Mesmo com a condicionante removendo grande parte das classificações espúrias iniciais, o número final de detecções em 2023 permaneceu consideravelmente superior ao de 2020 para ambos os algoritmos, reforçando a influência da variabilidade interanual na detecção da espécie.

Ano	Modelo	Pixels antes da condicionante	Pixels depois da condicionante	Total de Grupos encontrados
2020	RF	158.251	18.440	5.517
	SVM	95.153	8.874	2.646
	KNN	115.276	13.010	3.897
	<i>Voting</i>	107.887	10.038	2.993
	<i>Stacking</i>	114.269	10.802	3.224
2023	RF	228.613	33.964	10.182
	SVM	200.326	20.152	6.014
	KNN	238.648	36.753	11.010
	<i>Voting</i>	197.332	20.287	6.044
	<i>Stacking</i>	221.526	21.715	6.490

Tabela 3: Sumário da classificação por algoritmo e ano. A tabela quantifica o número total de detecções da espécie alvo para cada algoritmo (RF, SVM, KNN, Stacking e Voting) e para cada ano analisado. (Fonte: Autor, 2025)

A densidade de detecções foi calculada em relação à área florestal total do estudo (28.395,57 ha), expressa em grupos arbóreos por hectare. A análise revelou uma variação expressiva tanto entre os modelos quanto entre os anos. No ano de 2020, as densidades foram relativamente baixas, variando de 0,093 grupos/ha para o *Support Vector Machine* (SVM), o mais conservador, a 0,194 grupos/ha para o *Random Forest* (RF), que apresentou o maior número de detecções. Os demais modelos registraram densidades intermediárias: 0,137 grupos/ha (KNN), 0,114 grupos/ha (*Stacking*) e 0,105 grupos/ha (*Voting*).

Em comparação, o ano de 2023 apresentou um aumento expressivo neste índice para todos os algoritmos. As densidades variaram de 0,212 grupos/ha (SVM) a um pico de 0,388 grupos/ha para o KNN, que se tornou o modelo com maior volume de detecções. O RF registrou 0,358 grupos/ha, enquanto os modelos de *ensemble*, *Stacking* e *Voting*, apresentaram 0,229 e 0,213 grupos/ha, respectivamente. Essa tendência de aumento geral reforça a influência da variabilidade fenológica interanual na eficácia da detecção.

5.1. Análise de interseção e consistência dos modelos

A análise da interseção das detecções ao longo do tempo (2020 e 2023) fornece uma medida da estabilidade e da consistência de cada modelo. A Tabela 4 resume o número de agrupamentos que foram identificados em ambos os anos por cada um dos cinco algoritmos. O modelo *Voting Classifier* foi o que apresentou o maior número de detecções estáveis (224 grupos), seguido pelo *Stacking* (207), KNN (188) e SVM (186), indicando uma alta confiabilidade dos métodos de *ensemble* na identificação recorrente de alvos.

Tipo de Interseção	Modelo	Total de detecções dos grupos estáveis
Interseção Anual (Grupos estáveis 2020-2023 por modelo)	<i>Voting</i>	224
	<i>Stacking</i>	207
	KNN	188
	SVM	186
	RF	175
Ambos os anos	Todos os modelos	35

Tabela 4: Sumário da análise de interseção de detecções estáveis. A tabela apresenta o número de agrupamentos detectados em ambos os anos (2020 e 2023) para cada modelo (RF, SVM, KNN, Stacking e Voting). (Fonte: Autor, 2025)

5.2. Avaliação de desempenho dos classificadores com validação cruzada

A performance dos classificadores foi avaliada quantitativamente por meio da validação cruzada *K-Fold* Estratificada ($k=5$), realizada sobre o conjunto de dados combinado, a fim de avaliar a confiabilidade dos modelos à variabilidade interanual (Tabela 5).

Os resultados revelam uma performance alta para os modelos de *ensemble* e para o SVM. O *Stacking Classifier* e o SVM se destacaram como os mais precisos, alcançando um Índice Kappa de 0,939 e 0,939, respectivamente. O modelo *Stacking* demonstrou a maior capacidade de identificar corretamente a espécie com o Recall 0,935, com a menor taxa de omissão apresentando 13 Falsos Negativos. O *Voting Classifier* seguiu de perto, com um Kappa de 0,935, confirmando a eficácia da combinação de modelos.

O KNN apresentou um desempenho muito forte (Kappa 0,908), superando o RF e se mostrando uma alternativa confiável. O *Random Forest*, embora com um resultado sólido (Kappa 0,861), foi o que apresentou a maior taxa de omissão entre os classificadores supervisionados (27 Falsos Negativos), indicando uma menor sensibilidade para a classe *Alvo*.

Métrica	<i>Stacking</i>	SVM	<i>Voting</i>	KNN	RF
Acurácia Geral	0,973	0,973	0,971	0,960	0,940
Acurácia Balanceada	0,963	0,962	0,961	0,948	0,921
Índice de Kappa	0,939	0,939	0,935	0,908	0,861
Classe Alvo: <i>L. pisonis</i>	-	-	-	-	-
Precisão	0,984	0,989	0,984	0,963	0,950
Sensibilidade (Recall)	0,935	0,930	0,930	0,915	0,865
F1-Score	0,959	0,958	0,956	0,938	0,905

Verdadeiros Positivos (TP)	187	186	186	183	173
Falsos Positivos (FP)	3	2	3	7	9
Falsos Negativos (FN)	13	14	14	17	27
Verdadeiros Negativos (TN)	397	398	397	393	391

Tabela 5: Comparativo das métricas de avaliação dos classificadores. Desempenho obtido via validação cruzada ($k=5$) sobre o conjunto de dados combinado. (Fonte: Autor, 2025)

5.3. Distribuição por elevação

A análise da frequência de detecções por faixas de altitude foi expandida para incluir os cinco modelos e revelou um padrão de distribuição espacial consistente (Figura 16). Conforme resumido na Tabela 6, os resultados de todos os algoritmos e ambos os anos, demonstram uma concentração predominante das ocorrências de *L. pisonis* em faixas de altitude que podem ser classificadas como médias-baixas (entre 80 e 180 metros) (Figura 15).

De forma similar, a análise dos dados de 2020 (Figura 13), embora com um número absoluto menor de detecções, revela um padrão de distribuição por altitude idêntico ao observado em 2023 (Figura 13).

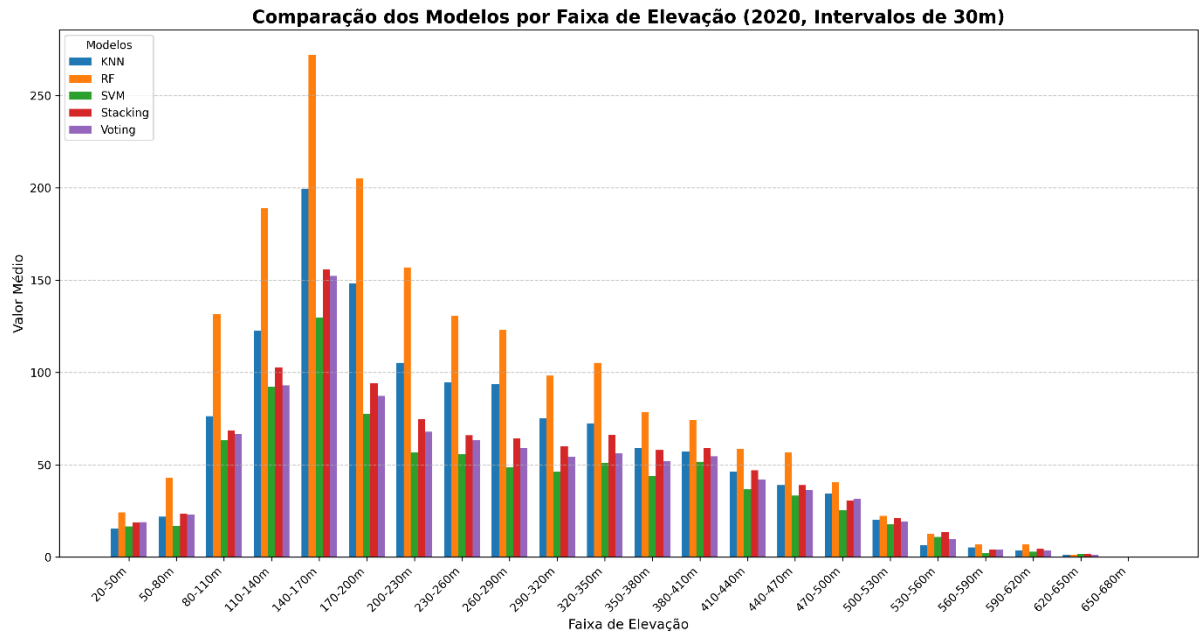


Figura 13: Gráfico da distribuição dos indivíduos de *L. pisonis* por elevação, comparando os modelos em 2020. O gráfico permite visualizar a relação da ocorrência da espécie com o gradiente de altitude e a soma das detecções (eixo Y) em intervalos de 50 m. (Fonte: Autor, 2025)

Modelo	Baixa (<80m)	Média-Baixa (80-180m)	Média-Alta (180-300m)	Alta (>300m)
RF	19	88	53	15
SVM	18	91	48	29
KNN	16	88	53	21
Voting	21	93	48	23
Stacking	21	91	51	18

Tabela 6: Distribuição dos agrupamentos estáveis (2020-2023) por faixas de elevação. A tabela resume o número de detecções em faixas de altitude consolidadas para cada modelo. (Fonte: Autor, 2025)

Para o ano de 2023, que apresentou o maior número de detecções, essa faixa de 80-180m concentrou a grande maioria dos indivíduos identificados por todos os modelos, desde o RF (4.621 agrupamentos) até o *Stacking* (2.543 agrupamentos). Picos de frequência foram

particularmente notáveis na faixa entre 130 e 160 metros para todos os classificadores (Figura 14 e 15).

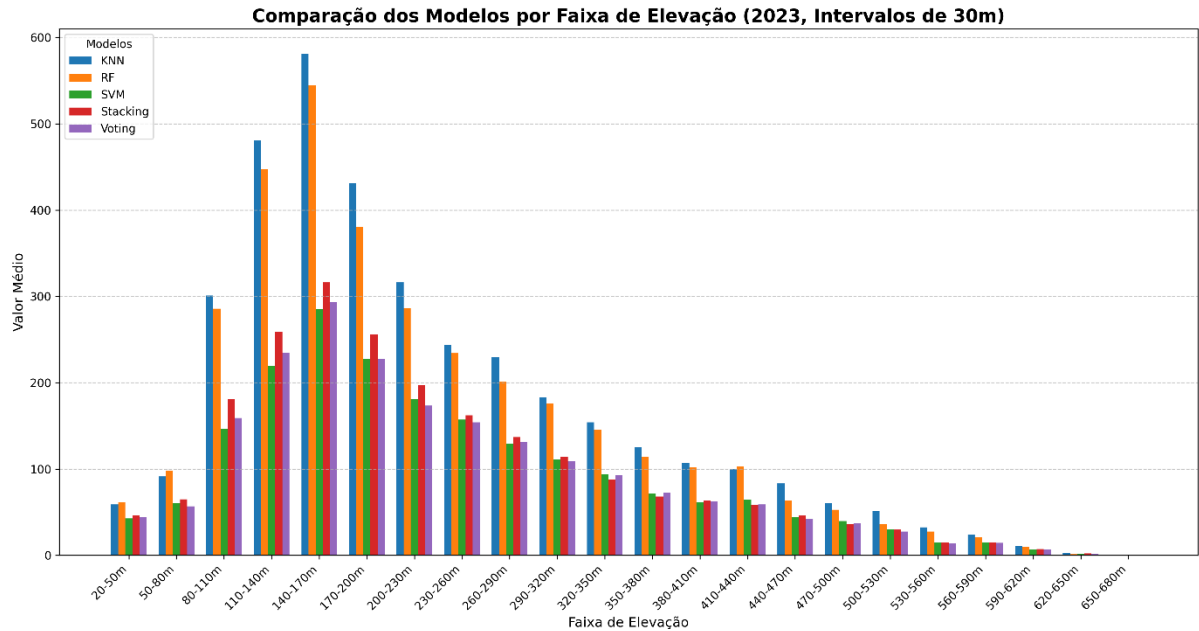


Figura 14: Gráfico da distribuição dos indivíduos de *L. pisonis* por elevação, comparando os modelos em 2023. O gráfico permite visualizar a relação da ocorrência da espécie com o gradiente de altitude e a soma das detecções (eixo Y) em intervalos de 50 m. (Fonte: Autor, 2025)

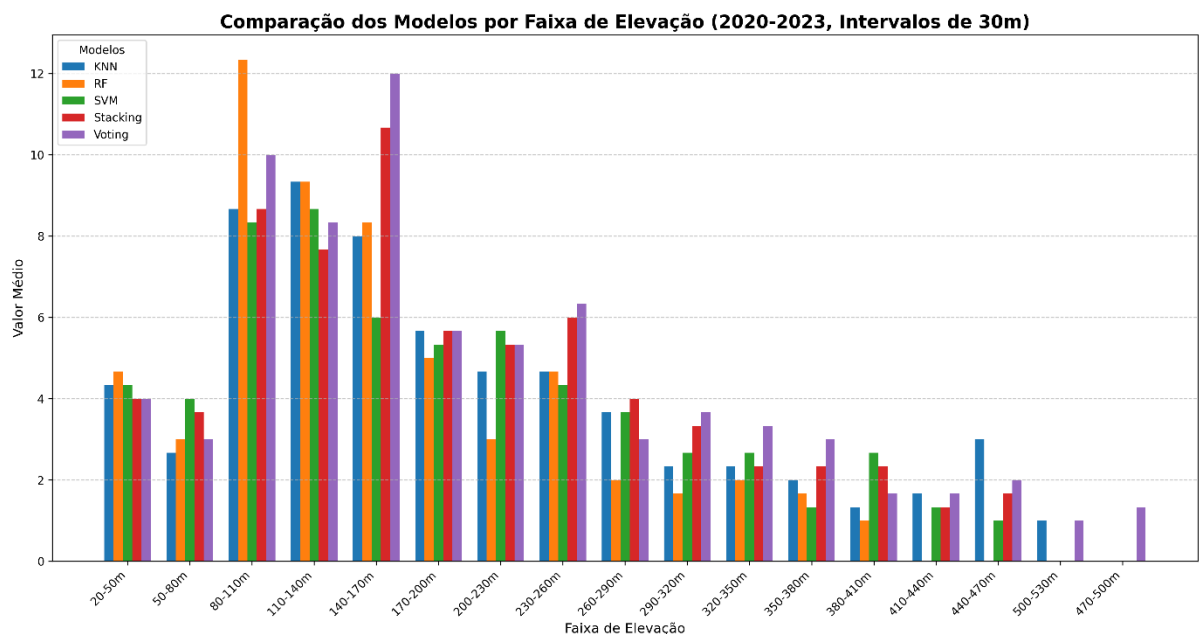


Figura 15: Gráfico da distribuição dos indivíduos de *L. pisonis* por elevação, comparando os modelos em ambos anos 2020-2023. O gráfico permite visualizar a relação da ocorrência da

espécie com o gradiente de altitude e a soma das detecções (eixo Y) em intervalos de 50 m. (Fonte: Autor, 2025)

Essa consistência demonstra que a preferência da *L. pisonis* por essas elevações tem um padrão ecológico e recorrente na área de estudo. O fato de os indivíduos com a assinatura espectral mais clara e repetível, estarem localizados predominantemente nesta faixa, solidifica a conclusão de que este é o principal local de preferência de ocorrência da espécie na paisagem analisada.

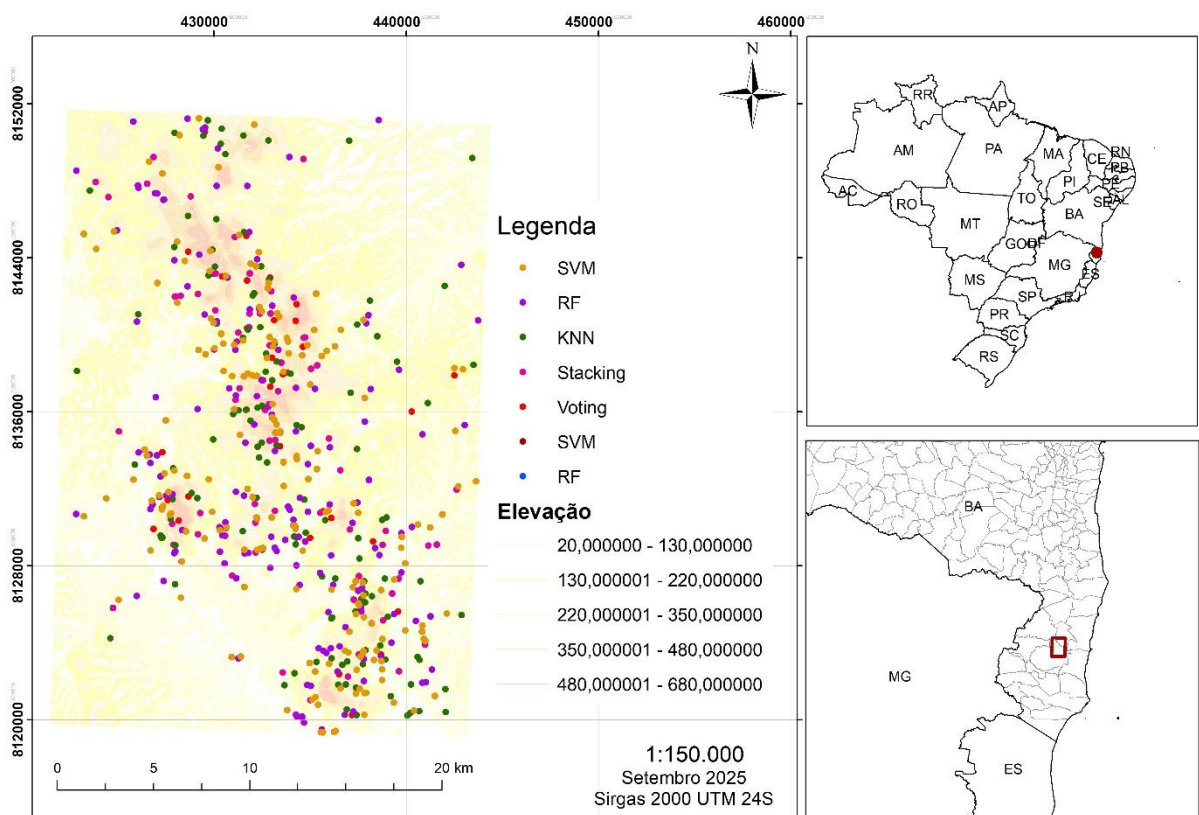


Figura 16: Mapa distribuição das detecções estáveis - altitude. Mapa de distribuição espacial que sobrepõe as detecções dos modelos a um MDE, confirmando a preferência por baixas altitudes.

(Fonte: Autor, 2025)

5.4. Distribuição próximos a cursos hídricos

A análise da proximidade das detecções em relação à rede hidrográfica demonstrou um padrão de agrupamento espacial consistente e marcante, que se manteve para todos os cinco

modelos e em ambos os anos analisados. Conforme detalhado na Tabela 7, a maior frequência de ocorrências da *L. pisonis* concentra-se de forma expressiva na faixa de até 100 metros dos cursos d'água.

No ano de 2020, que apresentou um número geral de detecções inferior a 2023, o padrão de forte associação com a rede hidrográfica já era proeminente. Conforme detalhado na Tabela 8, todos os cinco modelos concentraram uma parcela significativa de suas detecções nas faixas mais próximas aos cursos d'água. O *Random Forest* (RF) e o KNN, modelos com maior volume de detecção neste ano, identificaram 733 e 483 agrupamentos, respectivamente, a uma distância de até 100 metros da drenagem (Figura 17). Este comportamento, consistente entre todos os algoritmos, demonstra que a preferência da espécie por ambientes úmidos é uma característica marcante e detectável, mesmo em anos de menor expressão fenológica.

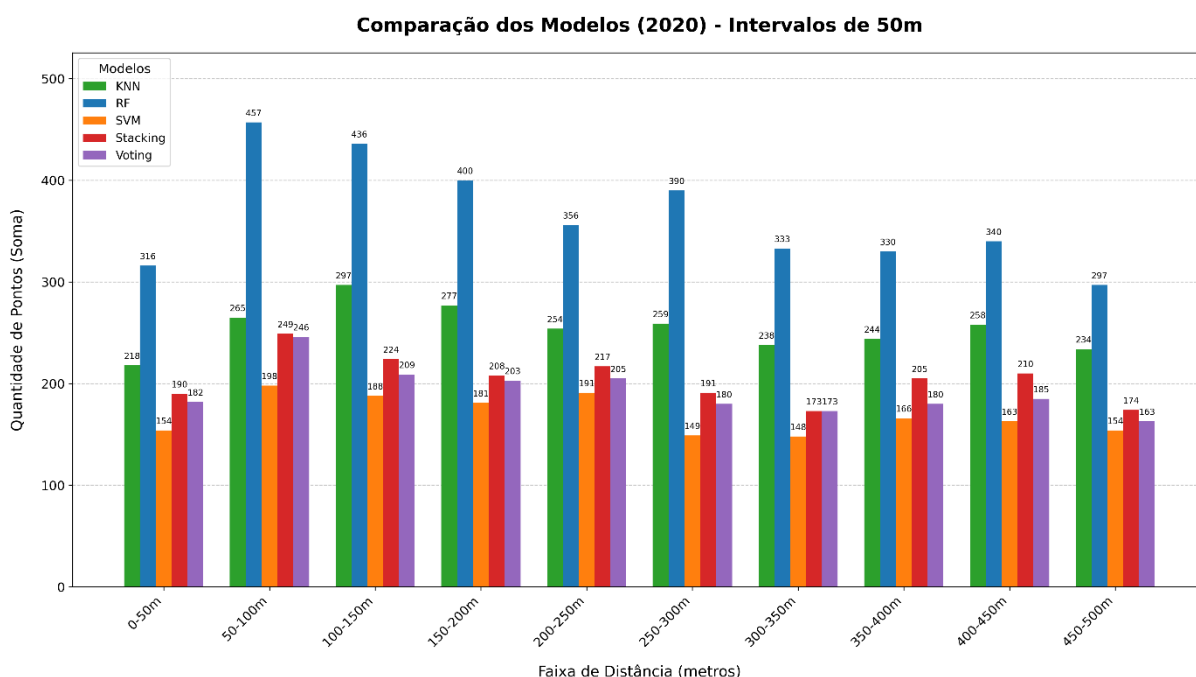


Figura 17: Gráfico da distribuição comparando os modelos em 2020. Gráfico de frequência que analisa a proximidade das detecções com a rede de drenagem e a soma das detecções (eixo Y) em intervalos de 50 m, indicando a afinidade da espécie por áreas úmidas.

(Fonte: Autor, 2025)

Para o ano de 2023, por exemplo, o modelo KNN detectou 1.560 agrupamentos nesta faixa, enquanto o *Stacking* e o *Voting* identificaram 1.022 e 931, respectivamente. Este padrão de alta concentração permanece elevado nas faixas mais próximas, apresentando um

decréscimo gradual até a distância de aproximadamente 400 metros, com uma queda mais significativa em distâncias superiores (Figura 18 e 19).

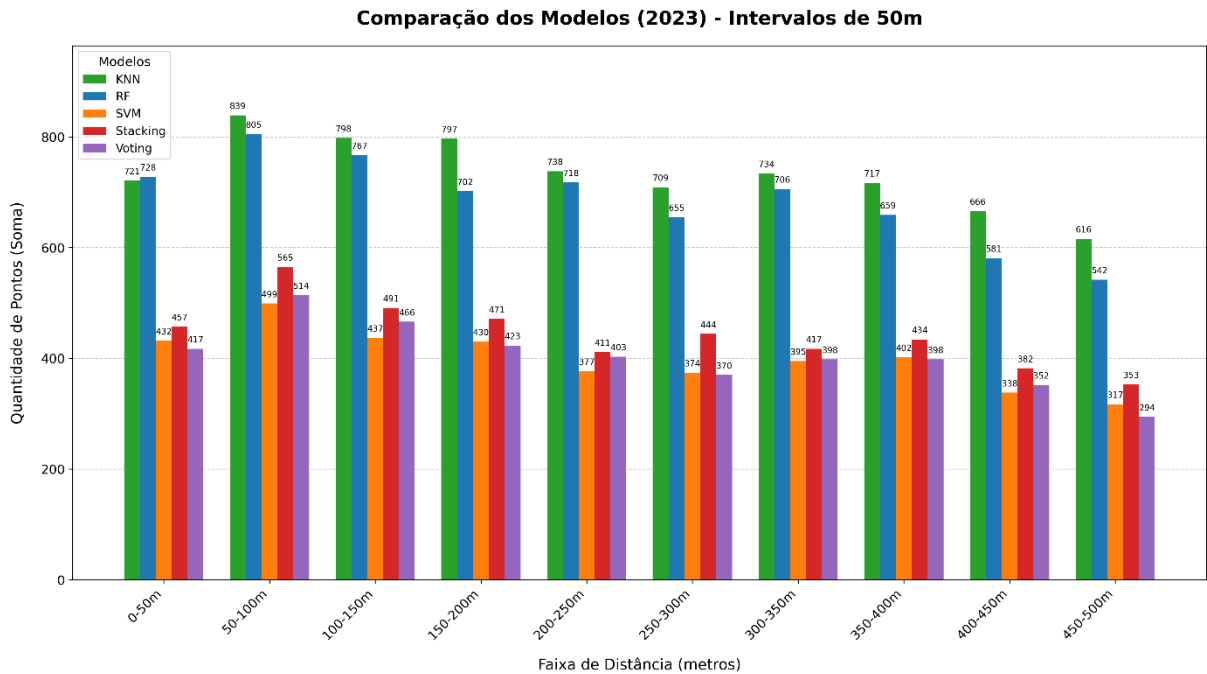


Figura 18: Gráfico da distribuição comparando os modelos em 2023. Gráfico de frequência que analisa a proximidade das detecções com a rede de drenagem, e a soma das detecções (eixo Y) em intervalos de 50 m, indicando a afinidade da espécie por áreas úmidas.

(Fonte: Autor, 2025)

A consistência deste resultado entre algoritmos com lógicas tão distintas, desde um classificador baseado em distância como o KNN até os complexos *ensembles Stacking e Voting*, reforça a conclusão de que a espécie possui uma forte preferência por habitats com alta disponibilidade hídrica.

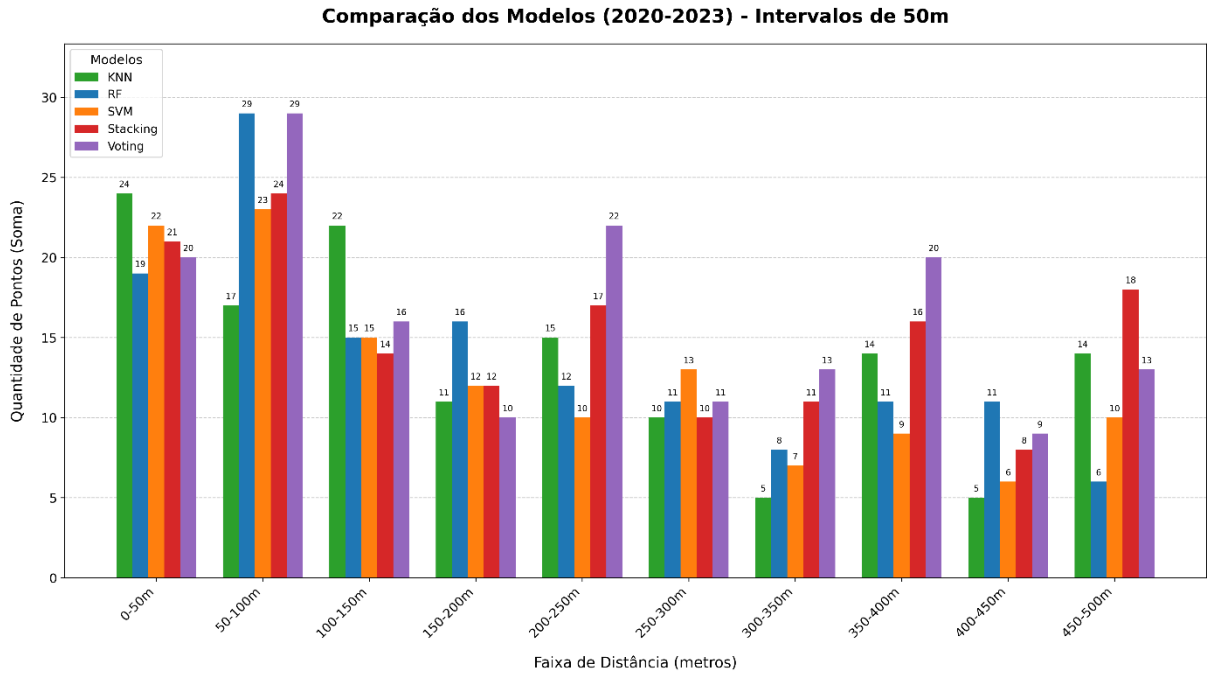


Figura 19: Gráfico da distribuição comparando os modelos em ambos os anos 2020-2023. Gráfico de frequência que analisa a proximidade das detecções com a rede de drenagem, e a soma das detecções (eixo Y) em intervalos de 50 m, indicando a afinidade da espécie por áreas úmidas.

(Fonte: Autor, 2025)

A análise focada nos indivíduos detectados de forma estável em ambos os anos demonstra que os indivíduos com a assinatura espectral mais confiável, mostrou-se ainda mais concentrado nas imediações dos rios e córregos (Tabela 7), (Figura 20).

Para os modelos de melhor desempenho, SVM e *Stacking*, mais da metade de todas as detecções estáveis ocorreram a uma distância de até 200 metros dos cursos d'água.

Modelo	Até 30m (APP)	Até 100m	Até 200m	Até 400m
RF	14	43	74	125
SVM	17	45	86	138
KNN	16	43	74	122
<i>Voting</i>	17	57	74	131
<i>Stacking</i>	17	48	85	136

Tabela 7: Distribuição dos agrupamentos estáveis (2020-2023) por proximidade a cursos d'água. A tabela mostra o número acumulado de detecções em faixas de distância para cada modelo. (Fonte: Autor, 2025)

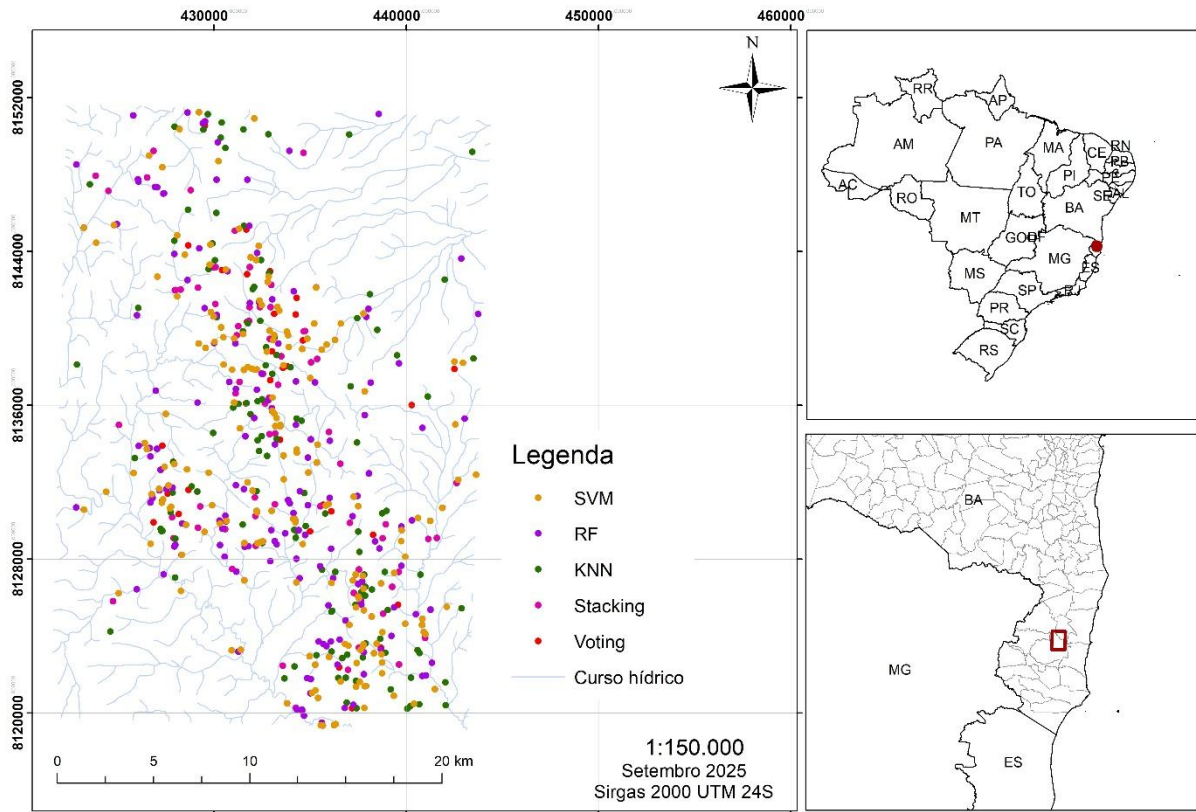


Figura 20: Mapa distribuição das detecções estáveis - hidrografia. Mapa de distribuição que correlaciona a localização dos indivíduos detectados de todos os modelos com a rede hidrográfica.

(Fonte: Autor, 2025)

5.5. Distribuição em áreas de APP

A análise de sobreposição espacial revelou uma associação entre a ocorrência de *L. pisonis* e as Áreas de Preservação Permanente (APPs) hídricas para todos os modelos avaliados. Utilizando como critério a faixa legal de 30 metros a partir dos cursos d'água, os resultados da Tabela 8 indicam que uma parcela significativa das detecções está localizada dentro das APPs, com um aumento no número de indivíduos entre 2020 e 2023.

Para o modelo Stacking, o número de detecções em APP cresceu de 142 para 344, enquanto para o Voting, o aumento foi de 135 para 304. O modelo KNN também seguiu a mesma tendência, saltando de 164 para 540 indivíduos detectados na faixa de APP.

Esses resultados sugerem que, embora a *L. pisonis* não possua um mecanismo de proteção legal específico, uma porção relevante de sua população na área de estudo se beneficia indiretamente da proteção conferida às matas ciliares.

Ano	Modelo	Total de Grupos	Em APP (0-30m)	0-10m	10-20m	20-30m
2020	RF	5.517	237	71	72	94
	SVM	2.646	119	34	37	48
	KNN	3.897	164	54	48	62
	<i>Voting</i>	2.993	135	40	39	56
	<i>Stacking</i>	3.224	142	41	45	56
2023	RF	10.182	543	196	175	172
	SVM	6.014	323	107	98	118
	KNN	11.010	540	188	177	175
	<i>Voting</i>	6.044	304	91	100	113
	<i>Stacking</i>	6.490	344	101	116	127

Tabela 8: Comparativo dos resultados obtidos a partir da proximidade em APP (0-30m). A tabela quantifica as detecções que ocorrem dentro das Áreas de Preservação Permanente (APP) para cada modelo e ano. (Fonte: Autor, 2025)

5.6. Ocorrência de *L. pisonis* por classes de cobertura do solo

A análise espacial dos pontos detectados permitiu avaliar a ocorrência de *L. pisonis* em diferentes classes de uso e cobertura do solo nos anos de 2020 e 2023 (Tabela 9). De maneira geral, observou-se que a maioria dos indivíduos está concentrada em áreas de vegetação natural, especialmente nas classes floresta estágio médio e floresta estágio inicial, indicando todos os

modelos identificaram essas classes como predominantes, embora com diferenças na quantidade de pontos detectados. Por exemplo, o *KNN* e o *RF* tendem a detectar um número maior de indivíduos, enquanto o *Stacking* e o *Voting* apresentam valores menores.

Em áreas agrícolas, como cabruca (cacau) e café, foram detectados poucos indivíduos, sugerindo que a espécie pode ocorrer de forma mais esparsa nesses contextos. Observou-se também a presença esporádica em pastagens e áreas de eucalipto, indicando algum grau de presença residual da espécie fora das florestas naturais. Classes com pouca vegetação natural, como sistema viário, área urbana ou represas, apresentaram detecções praticamente inexistentes.

Modelo	KNN	KNN	RF	RF	<i>Stacking</i>	<i>Stacking</i>	SVM	SVM	<i>VOTING</i>	<i>VOTING</i>
Ano	2020	2023	2020	2023	2020	2023	2020	2023	2020	2023
CLASSE										
Afloramento rochoso	7	40	14	43	6	31	6	28	7	30
Agricultura perene	13	14	8	15	25	18	18	26	23	16
Cabruca (cacau)	6	4	6	6	5	8	6	7	5	6
Café	28	69	45	76	33	81	34	68	28	75
Comunidade aluvial	1	27	9	35	1	18	1	18	2	17
Eucalipto	273	283	154	272	191	122	248	133	193	148
Floresta estágio inicial	114	386	207	456	106	314	92	241	106	279
Floresta estágio médio	3379	9829	4973	8839	2779	5531	2170	5183	2564	5138
Pasto limpo	20	114	17.0	137	22	124	19	99	19	117
Pasto sujo	53	231	81	300	53	229	50	199	45	209
Represa	0	5	0	5	0	5	0	3	0	4
Seringal	0	2	0	2	0	0	0	0	0	0
Sistema viário	2	9	4	10	3	10	1	11	1	8
Área urbana	0	2	0	2	0	4	0	3	0	3

Tabela 9: Resultados das detecções dos modelos (RF, KNN, Stacking, SVM e Voting) e as classes de uso e cobertura do solo (Mapeamento Fórum Florestal - 2022). Os valores representam o número total de agrupamentos detectados para cada classe e por ano. (Fonte: Autor, 2025)

Ao considerar detecções dos modelos de ambos os anos que se intersectaram, elas ocorreram predominantemente em áreas de floresta estágio médio, por todos os classificadores. Os modelos de *Stacking* e *Voting*, que combinam algoritmos, foram os mais expressivos na identificação, com 187 e 206 detecções, respectivamente, nas áreas de floresta estágio médio (Tabela 10).

Em menor grau, a espécie também foi encontrada em locais de floresta estágio inicial e em algumas áreas de pasto sujo, indicando sua presença em diferentes estágios de regeneração e em mosaicos de paisagem.

Modelo/ Classe	<i>Stacking</i>	<i>Voting</i>	SVM	RF	KNN
Floresta estágio médio	187	206	168	158	168
Floresta estágio inicial	11	12	9	7	9
Cabruca (cacau)	3	2	4	2	1
Eucalipto	0	0	0	0	3
Pasto limpo	2	1	2	2	3
Pasto sujo	4	3	2	4	3
Café	0	0	1	1	1

Tabela 10: Resultados da interseção entre as detecções estáveis dos modelos (RF, KNN, Stacking, SVM e Voting) e as classes de uso e cobertura do solo (Mapeamento Fórum Florestal - 2022). Os valores representam o número de agrupamentos detectados em ambos os anos (2020 e 2023) para cada classe. (Fonte: Autor, 2025)

5.7. Análise da interseção dos modelos

Para os algoritmos que apresentaram melhor precisão (*SVM, Stacking e Voting*), foram encontradas 276 detecções que se repetiu em ambos os anos, representando os pontos com maior confiabilidade na identificação da espécie (Tabela 11).

A altitude média dos 276 pontos foi de 437,6 metros, com uma mediana de 238,9 metros, indicando que, embora a espécie seja predominantemente encontrada em áreas baixas a intermediárias, há registros esparsos em elevações mais elevadas. A distância média dos cursos d'água foi de 23,4 metros, com uma mediana de apenas 12,6 metros, o que evidencia a forte afinidade da espécie por habitats próximos a corpos d'água.

Ao examinar a distribuição dos indivíduos em relação à altitude, percebe-se que aproximadamente metade dos pontos está abaixo de 250 metros, reforçando a tendência da *L. pisonis* ocupar áreas mais baixas, possivelmente devido à maior disponibilidade hídrica e solos mais úmidos. Em relação à proximidade de cursos d'água, mais de 50% dos indivíduos estão localizados a menos de 15 metros de um corpo hídrico.

Métrica	Altitude (m)	Distância de curso hídrico (m)
Média	344,39	16,55
Mediana	256,44	10,74
Mínimo	8,52	0,17
Máximo	1252,19	87,38
Desvio Padrão	309,32	19,25

Tabela 11: Análise descritiva dos 276 pontos em comun. Estatísticas das variáveis ambientais para os indivíduos detectados pelos algoritmos que apresentaram melhor precisão (SVM, Stacking e Voting). (Fonte: produção autoral, 2025)

Focando nas detecções coincidentes entre os cinco algoritmos e ambos os anos (2020 e 2023), foi identificado um subconjunto de 35 indivíduos consistentemente reconhecidos por todos os modelos. Esse grupo de detecções por sua alta estabilidade interanual e entre algoritmos, representa os indivíduos com a assinatura espectral mais distintiva e consistente, solidificando a conclusão de que a espécie tem forte preferência por habitats próximos à água (APÊNDICE B).

Uma análise estatística descritiva das variáveis ambientais para este grupo (Tabela 12) revela uma preferência de habitat ainda mais acentuada. A altitude média destes 35 pontos foi de 43,4 metros, com uma mediana de apenas 21,8 metros, indicando que metade das detecções mais seguras está localizada em áreas muito baixas. A distância média dos cursos d'água foi de 214 metros, porém, este valor é influenciado por alguns pontos mais distantes, já que a mediana foi de apenas 118,9 metros.

Esses resultados fornecem a evidência da afinidade da espécie por terrenos de baixa elevação e alta umidade, reforçando o conhecimento ecológico sobre a *L. pisonis* ser uma espécie característica de matas de baixada e ciliares.

Métrica	Altitude (m)	Distância de curso hídrico (m)
Média	43,40	214,04
Mediana	21,83	118,96
Mínimo	0,28	0,78
Máximo	136,61	725,83
Desvio Padrão	41,51	215,79

Tabela 12: Análise descritiva dos 35 pontos em comun. Estatísticas das variáveis ambientais para os indivíduos detectados por todos os cinco modelos. (Fonte: produção autoral, 2025)

5.8. Verificação e caracterização em campo

Durante as expedições de campo, realizadas em março de 2025, foram visitadas coordenadas selecionadas com base na acessibilidade logística para verificar as detecções do modelo (Tabela 9). Dentre os pontos visitados, a presença da espécie *L. pisonis* foi confirmada

visualmente em dois locais, sendo margem da estrada de acesso ao distrito de São Paulinho em Itamaraju – BA. Sendo visualizado alguns indivíduos e coletado dados de apenas 2 indivíduos, devido a acessibilidade.

A análise revelou uma altura total variando entre 18 e 22 metros e diâmetros de copa estimados entre 15 e 20 metros (Figura 21). Os valores de circunferência à altura do peito (CAP) medidos variaram entre 2,5 e 3,8 metros.



Figura 21: L. pisonis Confirmado em Campo. Registro fotográfico de uma das detecções, onde um ponto classificado como positivo pelo algoritmo foi confirmado como um indivíduo da espécie alvo.

(Fonte: produção autoral, 2025)

Indivíduo	Altura (m)	CAP (m)	Altura da 1º Bifurcação (m)	x	y	Município
-----------	------------	---------	-----------------------------	---	---	-----------

100	22	3,8	2	430285	8130005	Itamaraju
105	18	2,5	3	439745	8129685	Itamaraju

Tabela 13: Indivíduos de *L. pisonis* com dados biométricos coletados.

(Fonte: produção autoral, 2025)

Os dados morfométricos coletados em campo (Tabela 3) são valiosos para caracterizar a estrutura das árvores que o método efetivamente detecta. As medições de circunferência à altura do peito (CAP) e altura indicam que os indivíduos confirmados eram árvores de grande porte. Esta observação é consistente com a metodologia, combinando a resolução espacial de 10 metros do Sentinel-2 com o filtro de pós-processamento, que favorece a detecção de árvores maiores, com copas expansivas e expostas ao sol, enquanto indivíduos menores, sob dossel denso, ou que durante a fenologia não tenha atingido completamente a tonalidade alvo no momento da passagem do satélite, têm maior probabilidade de serem omitidos. As observações em campo dos indivíduos validados eram, em sua maioria, árvores com fuste bem formado e copas largas e saudáveis, indicando bom estado de vigor. A lista completa dos grupos identificados em ambos os anos e por ambos os modelos está detalhada no APÊNDICE C.

6. DISCUSSÃO

Este estudo demonstrou a viabilidade de detectar e mapear indivíduos de *L. pisonis* com base em sua assinatura fenológica, utilizando uma análise comparativa de cinco algoritmos de aprendizado de máquina. A performance dos modelos supervisionados confirma a hipótese central de que o evento de renovação foliar da espécie gera um sinal espectral suficientemente distinto para ser capturado por imagens Sentinel-2. A avaliação quantitativa revelou uma diferença de desempenho, com os métodos que combinam múltiplos modelos e modelos únicos.

A análise de desempenho posiciona o *Stacking Classifier* e o SVM como as abordagens mais precisas, com resultados parecidos (Kappa = 0,939). O fato de o *Stacking*, ter alcançado a performance mais alta está em conformidade com a literatura, que indica que essas arquiteturas podem aprender com os vieses dos classificadores de base para refinar a decisão final (Wolpert, 1992). A performance do *Voting Classifier* (Kappa = 0,935) e do KNN (Kappa = 0,909) também foi significativa, superando o *Random Forest* (Kappa = 0,862).

Para contextualizar a relevância desses resultados, é fundamental posicioná-los em relação a outros trabalhos na área de sensoriamento remoto e classificação de vegetação. A

Tabela 14 apresenta uma análise comparativa entre o desempenho dos modelos desenvolvidos neste estudo e os de outras pesquisas relevantes na literatura.

Estudo	Algoritmo	Tarefa de Classificação	Acurácia Geral	Índice Kappa
Este Estudo	<i>Stacking</i>	Detecção de <i>L. pisonis</i> vs. Outras classes	0,973	0,939
	SVM	Detecção de <i>L. pisonis</i> vs. Outras classes	0,973	0,939
	<i>Voting</i>	Detecção de <i>L. pisonis</i> vs. Outras classes	0,972	0,935
	KNN	Detecção de <i>L. pisonis</i> vs. Outras classes	0,96	0,909
	RF	Detecção de <i>L. pisonis</i> vs. Outras classes	0,94	0,862
Literatura Comparativa	-	-	-	-
Li et al. (2024)	Métodos de Ensemble	Classificação de cobertura em terrenos complexos	> 0,933	N/A
Pham et al. (2019)	KNN	Mapeamento de cobertura do solo	0,88	0,85
Mirończuk & Hościło (2017)	SVM	Mapeamento de "cobertura arbórea" vs. "não arbórea"	0,967 - 0,996	0,933 - 0,992
Rodriguez-Galiano et al. (2012)	RF	Classificação de cobertura do solo com imagens multitemporais	0,928	0,9
Francini et al. (2024)	RF	Mapeamento de 7 grupos funcionais de árvores	0,819	0,789
Cano & Marcato Junior (2021)	RF	Mapeamento de árvores em área urbana	94,2%	N/A (F1-Score: 85,13%)

Tabela 14: Análise Comparativa de Desempenho com a Literatura. Os valores para este estudo são baseados na validação cruzada sobre o conjunto de dados combinado.

(Fonte: produção autoral, 2025)

Como observado na Tabela 14, o desempenho dos modelos deste trabalho é competitivo. O Índice Kappa obtido (até 0,939) é equivalente ao de Mironczuk & Hościło (2017), que obtiveram um Kappa entre 0,933 e 0,992 para a tarefa mais simples de separar "cobertura arbórea" de "não arbórea". Além disso, a abordagem aqui utilizada, focada em uma única espécie com base em sua fenologia, se alinha a estudos recentes que também obtiveram sucesso na classificação de múltiplas espécies florestais utilizando imagens Sentinel-2 combinadas com dados auxiliares, como modelos de elevação (Xu et al., 2022). A performance dos métodos *Stacking* está de acordo com as conclusões de Li et al. (2024), que confirmam que tais abordagens consistentemente melhoram a acurácia em tarefas complexas. O desempenho do KNN neste estudo também superou o de trabalhos como o de Pham et al. (2019), que alcançou um Kappa de 0,85 para mapeamento de cobertura do solo.

Um dos achados centrais deste estudo é a acentuada variabilidade interanual na eficácia da detecção. Conforme detalhado na Tabela 3, houve um aumento expressivo no número de agrupamentos detectados de 2020 para 2023 em todos os modelos. O KNN, por exemplo, saltou de 3.897 para 11.010 grupos. Este fenômeno, a princípio pode indicar um evento de renovação foliar menos intenso ou síncrono em 2023. Uma hipótese para essa variação é que o evento de renovação foliar em 2023 tenha sido menos sincronizado entre os indivíduos do que em 2020. Em um evento dessincronizado, um número maior de árvores pode apresentar uma coloração intermediária no dia da passagem do satélite, o que levaria a um volume maior de detecções, ainda que a assinatura espectral não fosse a ideal.

Neste segundo cenário, a assinatura espectral média da espécie se torna "menos distinta" em relação ao alvo perfeito, mas permanece suficientemente diferente das outras classes de vegetação. Isso faz com que o limiar de decisão do classificador seja cruzado por um número maior de pixels, resultando em um volume de detecções inflado, que pode incluir tanto indivíduos em fases fenológicas subótimas quanto um aumento na taxa de falsos positivos.

Este desafio é bem documentado em estudos fenológicos, como o de Wagner (2021), que ilustra que eventos como a floração não são uniformes ano a ano. Fatores climáticos, como precipitação e temperatura, podem ter efeito direto sobre a intensidade e a duração da renovação foliar da *L. pisonis*. A abordagem com imagens únicas para cada ano é sensível a essa

variabilidade, o que destaca um contraste com métodos de séries temporais densas, como o utilizado por Francini et al. (2024), que são mais resilientes a essas diferenças.

Apesar dos resultados promissores, a metodologia possui algumas limitações. A análise da matriz de confusão do modelo de melhor desempenho, o *Stacking* (TP=187, FP=3, FN=13, TN=397), revela que, embora o modelo seja confiável, ele ainda omite cerca de 6,5% dos verdadeiros locais. Esta omissão é uma consequência direta da resolução espacial de 10 metros, que, conforme observado por Wagner (2021), requer agrupamentos de pixels para uma detecção segura. Essa característica, combinada com o filtro de pós-processamento, otimiza a detecção de árvores grandes e de dossel exposto, mas sistematicamente omite indivíduos jovens ou sob o dossel.

Portanto, os mapas gerados focam em indivíduos maduros, úteis para localizar árvores matrizes, mas não representam um censo completo da população. A abordagem com dados multiespectrais públicos também se difere de trabalhos como o de Ferreira et al. (2016), que utilizaram imagens hiperespectrais, ou de Venancio (2009), que empregou imagens de altíssima resolução para mapeamento individual.

Apesar dessas limitações, os padrões de distribuição espacial identificados corroboram o conhecimento ecológico documentado. A clara preferência da espécie por altitudes mais baixas e maior proximidade a cursos d'água, um padrão consistente entre todos os cinco modelos, confirma a classificação da *L. pisonis* como uma espécie de matas de baixada e ambientes úmidos (Mori & Prance, 1990; Lorenzi, 1992; Carvalho, 2008; Silva et al, 2014). A capacidade do sensoriamento remoto de validar e mapear esses nichos ecológicos demonstra o poder da tecnologia para uma compreensão espacialmente explícita da autoecologia da espécie.

No campo da conservação e restauração, as implicações são diretas. A característica da metodologia de detectar preferencialmente árvores de grande porte é uma de suas aplicações mais valiosas, pois esses indivíduos são, com alta probabilidade, árvores matrizes porta sementes. A capacidade de mapeá-las remotamente representa um avanço para potencializar a coleta de sementes e colaborar com programas como o Plano Nacional de Recuperação da Vegetação Nativa (PLANAVEG) (BRASIL, 2017). A metodologia apresentada demonstra potencial de aplicação em planos de manejo florestal, por sua capacidade de identificar indivíduos dominantes na paisagem, sobretudo considerando que *L. pisonis* já é reconhecida em Planos de Ação Nacional (PAN) para a conservação da flora, o que reforça a relevância desse tipo de abordagem. Além de auxiliar na detecção de indivíduos em estágio adulto, a metodologia pode complementar ou agregar valor aos inventários florestais tradicionais, oferecendo uma alternativa de monitoramento mais ágil e espacialmente explícita, capaz de

apoiar decisões estratégicas de manejo e conservação. Adicionalmente, o mapeamento pode subsidiar uma reavaliação de seu status de conservação, atualmente classificada como Menos Preocupante (LC) pela IUCN.

As limitações identificadas neste estudo podem ser superadas com a utilização de séries temporais mais densas de imagens Sentinel-2. A aplicação de técnicas de modelagem harmônica, em plataformas de computação em nuvem como o Google Earth Engine, permitiria capturar a curva fenológica completa, tornando o modelo mais confiável às variações no pico do evento (Francini et al., 2024). Adicionalmente, é necessário aprimorar a validação em campo, incluindo a visita a locais classificados como positivos e negativos em áreas de difícil acesso, a fim de obter estimativas mais precisas de erros de comissão e omissão.

Finalmente, o sucesso desta abordagem para a *L. pisonis* sugere seu potencial de adaptação para outras espécies da Mata Atlântica que também apresentam eventos fenológicos visualmente distintos, como a floração dos ipês (*Handroanthus*) (Wagner, 2021).

Ao mapear a distribuição de uma espécie-chave como a *L. pisonis*, o método pode ser usado indiretamente para avaliar a qualidade do habitat para a fauna associada, como psitacídeos que consomem suas sementes. Este trabalho, portanto, não apenas supre uma lacuna de dados sobre a espécie, mas também oferece uma ferramenta escalável e de baixo custo para monitorar a biodiversidade florestal.

7. CONCLUSÕES

Esta pesquisa desenvolveu e avaliou com sucesso um método combinando sensoriamento remoto e aprendizado de máquina para detectar árvores de *L. pisonis* em fragmentos da Mata Atlântica, confirmando a hipótese central de que a assinatura espectral da renovação foliar é uma característica eficaz para o mapeamento da espécie.

A abordagem utilizada, ao comparar cinco algoritmos distintos, demonstrou alta acurácia, com destaque para o *Stacking Classifier* e o SVM, que apresentaram o desempenho mais elevado. A eficácia dos métodos combinando múltiplos modelos, reforça que a combinação de múltiplos classificadores é uma estratégia poderosa para aumentar a confiabilidade em tarefas de sensoriamento remoto.

No entanto, o estudo também revelou que a eficácia da detecção é sensível à variabilidade interanual. Fatores ligados ao desenvolvimento fenológico da planta, como a intensidade e a sincronia da coloração das folhas, influenciam diretamente a capacidade dos modelos em discriminar a espécie alvo de forma consistente ano após ano. Adicionalmente,

reconhece-se que a condicionante espacial de tamanho de 3 ou 4 pixels agrupados, implementada para refinar os resultados, seleciona preferencialmente indivíduos maiores e de dossel dominante.

Longe de ser uma limitação, essa característica se mostrou uma aplicação prática valiosa, consolidando a metodologia como uma ferramenta confiável para a prospecção de árvores matrizes e porta sementes, de grande interesse para programas de conservação e restauração florestal.

8. PERSPECTIVAS

A pesquisa atual demonstrou a viabilidade de usar sensoriamento remoto e aprendizado de máquina para identificar *L. pisonis* com base em sua assinatura espectral durante a renovação foliar. No entanto, a análise de dados pontuais, restrita a dois anos, expôs a sensibilidade do método à variabilidade interanual da fenologia da espécie. Um estudo futuro deve explorar o uso de séries temporais mais densas de imagens Sentinel-2.

Em vez de utilizar apenas uma imagem por ano, a coleta contínua de dados permitiria a modelagem da curva fenológica completa da espécie. A aplicação tornaria o modelo mais robusto contra as variações anuais no pico do evento fenológico. Essa abordagem, sugerida por outros trabalhos na literatura, como o de Francini et al. (2024), tornaria a detecção mais confiável, independentemente das condições em um dia específico da passagem do satélite.

A validação de campo é uma etapa fundamental para a pesquisa, mas foi limitada a locais de fácil acesso. Para aprofundar a validação do método, a continuidade do trabalho deve incluir a verificação de um número maior de pontos, a fim de obter estimativas sobre a confiabilidade do método. A coleta de dados morfométricos adicionais, como altura e circunferência, forneceria informações significativas para entender a relação entre a biomassa da árvore e a detecção por sensoriamento remoto, além de ajudar a identificar árvores matrizes para futuros esforços de conservação.

9. REFERÊNCIAS

- ABAD-SEGURA, E. et al. Remote sensing applied in forest management to optimize ecosystem services: advances in research. **Forests**, v. 11, n. 9, p. 1-28, 2020.
- AGHDAM, H. H.; JAHANI-HABIBI, E. **Guide to Convolutional Neural Networks: A practical application to traffic-sign detection and classification**. Cham: Springer, 2017.
- ALASKA SATELLITE FACILITY Distributed Active Archive Center (ASF DAAC). **ALOS PALSAR RTC Project**. [S. l.], 2025. Disponível em: <https://www.earthdata.nasa.gov/data/projects/alos-palsar-rtc-project>. Acesso em: 1 ago. 2025.
- ALMEIDA, A. S. de; VIEIRA, I. C. G. Padrões florísticos e estruturais de uma cronosequência de florestas no Município de São Francisco do Pará, Região Bragantina, Pará. **Boletim do Museu Paraense Emílio Goeldi**, Belém, v. 17, n. 1, p. 209-240, 2001.
- ALVIM, P. de T.; ALVIM, R. Relation of climate to growth periodicity in tropical trees. In: CABOT SYMPOSIUM OF THE TROPICAL TREES AS LIVING SYSTEMS, 4., 1976, Petersham. **Proceedings**. Cambridge: Cambridge University Press, 1978. p. 445-464.
- ALTMAN, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. **The American Statistician**, v. 46, n. 3, p. 175-185, 1992.
- AMORIM, E.; BICALHO, M. *Lecythis pisonis* (Lecythidaceae). In: **Lista Vermelha da Flora Brasileira**: Centro Nacional de Conservação da Flora/ Instituto de Pesquisas Jardim Botânico do Rio de Janeiro. Rio de Janeiro: CNCFlora, 2022.
- BELGIU, M.; DRĂGUȚ, L. Random Forest in remote sensing: A review of applications and future research directions. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 114, p. 24-31, 2016. DOI: <https://doi.org/10.1016/j.isprsjprs.2016.01.011>.
- BAHIA. Secretaria do Planejamento. **Plano Territorial de Desenvolvimento Sustentável: território Extremo Sul**. Salvador: SEPLAN, 2016. Disponível em: https://www.seplan.ba.gov.br/wp-content/uploads/PTDS_Territorio_Extremo_Sul.pdf. Acesso em: 8 jan. 2025.
- BRASIL. Ministério do Meio Ambiente. **Plano Nacional de Recuperação da Vegetação Nativa – PLANAVEG**. Brasília: MMA, 2017. Disponível em: https://www.gov.br/mma/pt-br/assuntos/biodiversidade-e-biomas/biomas-e-ecossistemas/conservacao-1/politica-nacional-de-recuperacao-da-vegetacao-nativa/planaveg_plano_nacional_recuperacao_vegetacao_nativa.pdf. Acesso em: 23 ago. 2025.
- BRASIL. Agência Nacional de Águas e Saneamento Básico – ANA. **Atlas Hidrogeomorfológico Nacional – AHGN**. Brasília: ANA, 2021. Disponível em:

<https://metadados.snirh.gov.br/geonetwork/srv/por/catalog.search#/metadata/79f92322-f806-4e52-ab6c-4ef348fae938>. Acesso em: 17 jun. 2025.

BORGES, G. M.; PACHÊCO, A. da P.; SANTOS, F. K. S. dos. Sensoriamento remoto: avanços e perspectivas. **Revista de Geografia**, Recife, v. 32, n. 2, p. 267-292, 2015.

BORLACHENCO, N. G. C. **Avaliação da recuperação de uma área degradada com espécies arbóreas melíferas**. 2018. Tese (Doutorado em Ciências Ambientais e Sustentabilidade Agropecuária) – Programa de Pós-Graduação em Ciências Ambientais e Sustentabilidade Agropecuária, Universidade Católica Dom Bosco, Campo Grande, 2018.

BOSCOLO, H. K.; NAMIKAWA, L. As cores em imagens de sensoriamento remoto. [S. l.], 2024. Disponível em: <https://periodicos.unb.br/index.php/dasquestoes/article/download/22052/20199/39265>. Acesso em: 17 jun. 2024.

BRODERSEN, K. H. et al. The balanced accuracy and its posterior distribution. In: IEEE INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION, 2010, Istambul. **Anais...** Istambul: IEEE, 2010. p. 3121-3124.

BREIMAN, L. Stacked regressions. **Machine Learning**, v. 24, n. 1, p. 49-64, 1996.

BREIMAN, L. Random forests. **Machine Learning**, [S. l.], v. 45, n. 1, p. 5-32, 2001. DOI: 10.1023/A:1010933404324.

BROWN, R. Geomorphons: landform and property predictions in a glacial moraine in Indiana landscapes. **Catena**, v. 142, p. 66-76, 2016.

CANO, P. L. G.; MARCATO JUNIOR, J. Aplicação de aprendizado de máquina com dados de sensoriamento remoto para o mapeamento de florestas urbanas. **Revista UNG - Geociências**, v. 20, n. 2, p. 16-27, 2021.

CARVALHO, P. E. R. **Espécies arbóreas brasileiras**. v. 2: Sapucaia (*Lecythis pisonis*). Colombo, PR: Embrapa Florestas; Brasília, DF: Embrapa Informação Tecnológica, 2006. Disponível em: <https://www.alice.cnptia.embrapa.br/alice/bitstream/doc/1140784/1/Especies-Arboreas-Brasileiras-vol-2-Sapucaia.pdf>. Acesso em: 8 jul. 2025.

CAMACHO-VELASCO, A. et al. Aplicações e desafios do sensoriamento remoto hiperespectral na geologia colombiana. **Revista Facultad de Ingeniería**, v. 24, n. 40, p. 17-29, 2015.

CAMBESSEDES, J. L. *pisonis*. In: SAINT-HILAIRE, A. de; CAMBESSEDES, J.; JUSSIEU, A. de. **Flora Brasiliae Meridionalis**. 4. ed. v. 2, n. 20. [S.l.: s.n.], 1829. p. 377.

COHEN, J. A coefficient of agreement for nominal scales. **Educational and Psychological Measurement**, v. 20, n. 1, p. 37-46, 1960.

CONCEIÇÃO, I. H. da. **Histórico do uso de drones na agricultura: uma revisão**. 2023. Trabalho de conclusão de curso (Bacharelado em Agronomia) - Universidade Federal do Recôncavo da Bahia, Cruz das Almas, 2023.

CONGALTON, R. G.; GREEN, K. **Assessing the accuracy of remotely sensed data: principles and practices**. 2. ed. Boca Raton, FL: CRC Press, 2009.

COPERNICUS. **Sentinel-2 User Handbook**. [S. l.], 2024. Disponível em: https://sentinels.copernicus.eu/documents/247904/685211/Sentinel-2_User_Handbook. Acesso em: 6 abr. 2024.

COPERNICUS [Sistema de Observação da Terra da União Europeia]. **Descrição das Fórmulas de Renderização de Imagens no EO Browser Sentinel**. [S. l.], 2024. Disponível em: <https://apps.sentinel-hub.com/eo-browser/>. Acesso em: 6 abr. 2024.

COPERNICUS. **Sentinel-2**. [S. l.], 2024. Disponível em: <https://sentinels.copernicus.eu/web/sentinel/copernicus/sentinel-2>. Acesso em: 6 abr. 2024.

COSTA, E. A. et al. Relação entre o diâmetro de copa e o diâmetro à altura do peito de *Araucaria angustifolia* (Bertol.) Kuntze, Lages, SC. In: SIMPÓSIO BRASILEIRO DE PÓS-GRADUAÇÃO EM CIÊNCIAS FLORESTAIS, 8., 2014, Recife. **Anais**. Recife, 2014. p. 696-700. DOI: 10.12702/VIII.SimposFloresta.2014.250-607-1.

COSTA, W. C. A. da. **Caracterização agrônômica da cultura do café com auxílio de ferramentas do controle estatístico e sensoriamento remoto terrestre**. 2019. Dissertação (Mestrado em Agronomia) – Programa de Pós-Graduação em Agronomia (Produção Vegetal), Universidade Estadual Paulista "Júlio de Mesquita Filho", Jaboticabal, 2019.

COVER, T.; HART, P. Nearest neighbor pattern classification. **IEEE Transactions on Information Theory**, v. 13, n. 1, p. 21-27, 1967.

DIETTERICH, T. G. Ensemble methods in machine learning. In: INTERNATIONAL WORKSHOP ON MULTIPLE CLASSIFIER SYSTEMS, 1., 2000, Cagliari. **Proceedings...** Cagliari: Springer, 2000. p. 1-15.

DUDANI, S. A. The distance-weighted k-nearest-neighbor rule. **IEEE Transactions on Systems, Man, and Cybernetics**, n. 4, p. 325-327, 1976.

ESA - AGÊNCIA ESPACIAL EUROPEIA. **Programa Sentinel-2**. [S. l.], 2024. Disponível em: https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-2. Acesso em: 6 abr. 2024.

FITZ, P. R. **Geoprocessamento sem complicação**. São Paulo: Oficina de Textos, 2008.

FLORA DO BRASIL. **Lecythidaceae in Flora do Brasil 2020**. Rio de Janeiro: Jardim Botânico do Rio de Janeiro, 2024. Disponível em: <https://floradobrasil.jbrj.gov.br/consulta/ficha.html?idDadosListaBrasil=145>. Acesso em: 17 jun. 2024.

FÓRUM FLORESTAL. **Monitoramento de uso e cobertura do solo no sul da Bahia: O Monitoramento Independente da cobertura vegetal e uso do solo dos territórios de identidade Costa do Descobrimento, Extremo Sul e Parcela do Litoral Sul**. [S. l.]: Forum Florestal Bahia, 2024. Disponível em: <https://forumflorestalbahia-worldresources.hub.arcgis.com/>. Acesso em: 17 jun. 2024.

FASSNACHT, F. et al. Importance of sample size, data type and prediction method for remote sensing-based estimations of aboveground forest biomass. **Remote Sensing of Environment**, v. 154, p. 102-114, 2014. DOI: 10.1016/j.rse.2014.07.028.

FRANCINI, S. et al. Forest species mapping and area proportion estimation combining Sentinel-2 harmonic predictors and national forest inventory data. **International Journal of Applied Earth Observation and Geoinformation**, v. 131, 2024. DOI: 10.1016/j.jag.2024.103935.

GOMES, H. M. et al. A survey of stream data mining. **ACM Computing Surveys (CSUR)**, v. 49, n. 4, p. 1-37, 2016.

GUIMARÃES, E. F. et al. **Árvores do Jardim Botânico do Rio de Janeiro**. Rio de Janeiro: Jardim Botânico, 1993. 234 p.

GILLIES, S. et al. **Rasterio Documentation**. Release 1.4b1.dev0. [S. l.: s. n.], 2024. Disponível em: <https://readthedocs.org/projects/Rasterio/downloads/pdf/latest/>. Acesso em: 5 maio 2025.

GROHMANN, C. H. Morphometric analysis in geographic information systems: applications of free software GRASS and R. **Computers & Geosciences**, v. 30, n. 9-10, p. 1055-1067, 2004. INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS (INPE). **PRODES Project: Monitoring the Brazilian Amazon forest by satellite**. São José dos Campos: INPE, 2017. Disponível em: http://www.inpe.br/amazonia1/en/uses_applications.php. Acesso em: 6 abr. 2025.

KENNEDY, R. E.; COHEN, W. B.; SCHROEDER, T. A. Trajectory-based change detection for automated characterization of forest disturbance dynamics. **Remote Sensing of Environment**, v. 110, n. 3, p. 370-386, 2007.

KOBAL, M. et al. On the use of R programming language in the analyses of spatial data. **Acta Silvae et Ligni**, v. 102, p. 55-62, 2013.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, 14., 1995, Montreal. **Proceedings** [...]. Montreal: [s.n.], 1995. v. 2, p. 1137–1143.

LANARAS, C. et al. Super-resolution of Sentinel-2 images: learning a globally applicable deep neural network. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 146, p. 305–319, 2018.

LI, M.; STEIN, A. Mapping land use from high resolution satellite images by exploiting the spatial arrangement of land cover objects. **Remote Sensing**, v. 12, n. 24, p. 4158, 2020.

LI, J. et al. Ensemble Learning for the Land Cover Classification of Loess Hills in the Eastern Qinghai–Tibet Plateau Using GF-7 Multitemporal Imagery. **Remote Sensing**, v. 16, n. 14, p. 2556, 2024.

LORENZI, H. **Árvores brasileiras: manual de identificação e cultivo de plantas arbóreas nativas do Brasil**. 4. ed. Nova Odessa: Instituto Plantarum, 2002. v. 1, 368 p.

LORENZI, H. **Árvores brasileiras: manual de identificação e cultivo de plantas arbóreas nativas do Brasil**. 1. ed. Nova Odessa, SP: Instituto Plantarum, 1992. v. 1. 368 p.

LU, D.; MORAN, E.; BATISTELLA, M. Linear mixture model applied to Amazonian vegetation classification. **Remote Sensing of Environment**, v. 87, n. 4, p. 456–469, 2003.

MAINIERI, C.; CHIMELO, J. P. **Fichas de características das madeiras brasileiras**. São Paulo: IPT, 1989. 418 p.

MATHER, P. M. **Computer Processing of Remotely-Sensed Images: An Introduction**. [S.l.]: John Wiley & Sons, 2004.

MAXWELL, A. et al. Implementation of machine-learning classification in remote sensing: An applied review. **International Journal of Remote Sensing**, v. 39, n. 9, p. 2784-2817, 2018.

MCKINNEY, W. **Python para análise de dados: Tratamento de dados com Pandas, Numpy e IPython**. São Paulo: Novatec Editora, 2018.

METCALFE, P.; BEVEN, K.; FREER, J. Dynamic TOPMODEL: A new implementation in R and its sensitivity to time and space steps. **Environmental Modelling & Software**, v. 72, p. 155-172, 2015.

MIRONCZUK, A.; HOŚCIŁO, A. Mapping tree cover with Sentinel-2 data using the Support Vector Machine (SVM). **Geoinformation Issues**, v. 9, n. 1, p. 5-15, 2017. DOI: 10.34867/gi.2017.1..

MORAES, E. C. Fundamentos de sensoriamento remoto. In: CURSO DE USO DE SENSORIAMENTO REMOTO NO ESTUDO DO MEIO AMBIENTE. São José dos Campos: INPE, 2002. cap. 1.

MORELLATO, L. P. C. et al. Linking plant phenology to conservation biology. **Biological Conservation**, v. 195, p. 60-72, 2016.

MORI, S. A.; ORCHARD, J. E. Fenologia, biologia floral e evidência sobre dimorfismo do pólen de *Lecythis pisonis* Cambess. (Lecythidaceae). In: CONGRESSO NACIONAL DE BOTÂNICA, 31., 1979, Campinas. **Anais**. Campinas: Sociedade Botânica do Brasil, 1979. p. 109-116.

MORI, S. A.; PRANCE, G. T. **Lecitidáceas**: família da castanha-do-pará. Salvador: Centro de Pesquisa do Cacau, 1983. (Boletim Técnico, 116).

MORI, S. A.; PRANCE, G. T. **Lecythidaceae** - Part II: The zygomorphic-flowered New World genera (*Bertholletia*, *Corythophora*, *Couratari*, *Couroupita*, *Eschweilera*, and *Lecythis*). **Flora Neotropica Monographs**, v. 21, n. 2, p. 1-376, 1990.

NOAA. Climate Prediction Center. **Oceanic Niño Index (ONI)**, v5. [S. l.], 2025. Disponível em: https://origin.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ONI_v5.php. Acesso em: 17 jun. 2025.

NUMPY. **NumPy Documentation**. [S. l.], 2024. Disponível em: <https://docs.scipy.org/doc/>. Acesso em: 13 jun. 2025.

OMUTO, C. T.; VARGAS, R. R. Re-tooling of regression kriging in R for improved digital mapping of soil properties. **Geoscience Journal**, v. 19, n. 1, p. 157-165, 2014.

ORMOND, W. T.; PINHEIRO, M. C. B.; CORTELLA DE CASELLS, A. R. A contribution to the floral biology and reproductive system of *Couroupita guianensis* Aubl. (Lecythidaceae). **Annals of the Missouri Botanical Garden**, St. Louis, v. 68, p. 514-523, 1981.

PACHECO, A. Avaliação de modelos, cross-validation e data leakage. **Computação Inteligente**, Vitória (ES), 25 jun. 2021. Disponível em:

<https://computacaointeligente.com.br/conceitos/avaliando-performance-cross-validation/>. Acesso em: 22 ago. 2025.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825-2830, 2011. Disponível em: <http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf>. Acesso em: 18 ago. 2024.

PHAM, H. T. et al. A comparative study of machine learning algorithms for land use/land cover classification using Landsat-8 OLI. **International Journal of Remote Sensing**, v. 40, n. 24, p. 9452-9472, 2019.

PINTO, J. R. R.; RODERJAN, C. V.; GALVÃO, F. Introdução à fitossociologia e ecologia de florestas da Mata Atlântica. In: FANTINI, A. C.; GURSKI, C. (Org.). **Essências Nativas da Mata Atlântica: Identificação, Propagação e Utilização**. São Carlos: RiMa Editora, 2006. p. 23-46. Disponível em: https://www.academia.edu/download/54049819/Pinto_et_al_Capitulo_MataAtlantica_Essencias_RiMa_Editora_2006.pdf. Acesso em: 24 fev. 2025.

PIRES, C. S. S.; TOREZANI, K. R. de S. **Seleção de espécies de abelhas nativas para avaliação de risco de agrotóxicos**. Brasília: Ibama, 2018. 84 p. Disponível em: https://www.ibama.gov.br/phocadownload/agrotoxicos/reavaliacao-ambiental/2018/Selecao_Especies_Abelhas_Nativas_para_Avaliacao_de_Risco_de_Agrotoxicos.pdf. Acesso em: 24 ago. 2025.

PONZONI, F. J.; REZENDE, A. C. P. Caracterização espectral de estágios sucessionais de vegetação arbórea secundária arbórea em Altamira (PA), através de dados orbitais. **Revista Árvore**, v. 28, n. 4, p. 535-545, 2004.

PRANCE, G. T. **Lecythidaceae** – Part II. The zygomorphic-flowered New World genera (*Couroupita*, *Corythophora*, *Bertholletia*, *Couratari*, *Eschweilera*, & *Lecythis*). **Flora Neotropica**, v. 21, n. 2, p. 1-376, 1990.

PYTHON SOFTWARE FOUNDATION. **Python**. [S. l.], 2025. Disponível em: <https://www.python.org/>. Acesso em: 24 ago. 2025.

PYTHON. **Documentação da linguagem Python**. [S. l.], 2025. Disponível em: <https://docs.python.org/pt-br/3/>. Acesso em: 24 ago. 2025.

REFAEILZADEH, P.; TANG, L.; LIU, H. Cross-validation. In: LIU, L.; ÖZSU, M. T. (eds.). **Encyclopedia of database systems**. Boston: Springer US, 2009. p. 532-538.

REITZ, P. R. **Lecitidáceas**. Itajaí: Herbário Barbosa Rodrigues, 1981. 32 p.

RODRIGUEZ-GALIANO, V. F. et al. An assessment of the effectiveness of a random forest classifier for land-cover classification. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 67, p. 93-104, 2012.

SAGI, O.; ROKACH, L. Ensemble learning: A survey. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 8, n. 4, p. e1249, 2018.

SARTORELLI, P. A. R. et al. **Plantas dos povos da Volta Grande do Xingu**: Terras Indígenas Paquçamba, Arara da VGX e Área Indígena Juruna do Km 17. [S. l.]: Baobá Florestal, 2018. Disponível em: https://www.baobaflorestal.com.br/pesquisas/downloads/guia_vgx.pdf. Acesso em: 24 ago. 2025.

SANTORI, A. R. **Geotecnologias aplicadas à análise ambiental**: aquisição e avaliação de variáveis espectrais da vegetação entre os anos de 2018 e 2019 na bacia hidrográfica do Ribeirão das Cruzes, Selvíria/MS. 2021. Dissertação (Mestrado em Geografia) – Universidade Federal de Mato Grosso do Sul, Três Lagoas, 2021.

SCIKIT-LEARN. **RandomForestClassifier** — **scikit-learn 1.7.0 documentation**. [S. l.], 2025. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>. Acesso em: 22 ago. 2025.

SCIPY. **scipy.ndimage.label** — **Scipy Documentation**. 2023. Disponível em: <https://docs.scipy.org/doc/scipy/reference/generated/scipy.ndimage.label.html>. Acesso em: 22 ago. 2025.

SHEYKHOUMOUS, M. et al. Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, v. 13, p. 6308-6325, 2020.

SILVA, J. A.; SILVA, M. F. F. da; CARVALHO, J. E. U. de. *Conhecendo espécies de plantas da Amazônia: Sapucaia (Lecythis pisonis Cambess. – Lecythidaceae)*. Belém, PA: Embrapa Amazônia Oriental, 2014. (Comunicado Técnico, 250).

SILVA, J. P.; OLIVEIRA, M. L.; COSTA, R. F. Uso de imagens de alta resolução espacial para o monitoramento da cobertura florestal na região central do Rio Grande do Sul. **Revista do Instituto de Geociências**, v. 41, n. 3, p. 123-135, 2018. Disponível em: <https://revistas.ufrj.br/index.php/aigeo/article/view/31325>. Acesso em: 24 abr. 2025.

SILVA, J. M. C. da et al. **Conservação da Mata Atlântica brasileira**: um balanço dos últimos dez anos. [S. l.]: SOS Mata Atlântica, [2017]. Disponível em: <https://www.sosma.org.br/artigos/conservacao-da-mata-atlantica-brasileira-um-balanco-dos-ultimos-dez-anos/>. Acesso em: 14 abr. 2022.

SMITH, N. P.; MORI, S. A.; PRANCE, G. T. *Lecythis*. In: JARDIM BOTÂNICO DO RIO DE JANEIRO. **Lista de Espécies da Flora do Brasil**. Rio de Janeiro: Jardim Botânico do Rio de Janeiro, 2013.

SOARES, C. B. L. da V. **Árvores nativas do Brasil**. Rio de Janeiro: Salamandra, 1990. 115 p.

STEFFEN, C. A. **Introdução ao Sensoriamento Remoto**. São José dos Campos: Instituto Nacional de Pesquisas Espaciais (INPE), [s.d.]. Disponível em: <http://www3.inpe.br/unidades/cep/atividadescep/educasere/apostila.htm>. Acesso em: 6 maio 2025.

VALLILO, M. I. et al. Caracterização química parcial das sementes de *L. pisonis* camb. (Sapucaia). **Acta Amazônica**, v. 28, p. 131-140, 1998. Disponível em: <https://doi.org/10.1590/1809-43921998282140>. Acesso em: 6 abr. 2024.

VENANCIO, D. L. **Mapeamento de *Araucaria angustifolia* utilizando DGPS e imagem de satélite Quickbird-2**. 2009. 64 f. Dissertação (Mestrado em Ciências Florestais) – Universidade Estadual do Centro-Oeste, Irati, 2009.

VICENTAINER, D.; MATTEDI, M.; MELLO, B. Aplicação das Bibliotecas Python para tratamento de dados em tempo real: A análise dos dados de isolamento social em Santa Catarina. **Meta: Avaliação**, v. 12, n. 23, p. 80-90, 2020. Disponível em: <https://publicacoes.ifc.edu.br/index.php/metapre/article/view/1392/1061>. Acesso em: 17 jun. 2024.

VIRTANEN, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. **Nature Methods**, v. 17, n. 3, p. 261–272, mar. 2020. Disponível em: <https://doi.org/10.1038/s41592-019-0686-2>. Acesso em: 6 maio 2025.

WALT, S. van der; COLBERT, S. C.; VAROQUAUX, G. The Numpy array: a structure for efficient numerical computation. **Computing in Science & Engineering**, v. 13, n. 2, p. 22-30, 2011.

WAGNER, F. H. et al. **The flowering of Atlantic Forest *Pleroma* trees. *Scientific Reports***, v. 11, n. 20437, 2021. DOI: <https://www.nature.com/articles/s41598-021-99304-x>.

WEISS, M.; JACOB, F.; DUVEILLER, G. Remote sensing for agricultural applications: A meta-review. **Remote Sensing of Environment**, v. 236, p. 1-19, 2020.

WIKIAVES. **Sapucaia**. [S. l.], 2025. Disponível em: <https://www.wikiaves.com.br/wiki/flora:sapucaia>. Acesso em: 24 ago. 2025.

XU, L. et al. Forest Tree Species Classification Based on Sentinel-2 Images and Auxiliary Data. **Forests**, v. 13, n. 9, p. 1432, 2022. DOI: <https://doi.org/10.3390/f13091432>.

WOLPERT, D. H. Stacked generalization. **Neural Networks**, v. 5, n. 2, p. 241-259, 1992.

ZHAO, Y. L. I. et al. Meta-optics achieves RGB-achromatic focusing for virtual reality. **Science Advances**, v. 7, n. 5, 2021. Disponível em: <https://www.science.org/doi/full/10.1126/sciadv.abe4458>. Acesso em: 19 fev. 2025.

APÊNDICES

APÊNDICE A – DISPONIBILIZAÇÃO DE CÓDIGOS E DADOS

Este trabalho disponibiliza um repositório público no GitHub contendo:

- Scripts em Python para pré-processamento de imagens Sentinel-2;
- Implementação completa do modelo Random Forest com Validação Cruzada K-Fold;
- Aplicação da condicionante espacial para filtragem dos agrupamentos;
- Scripts para extração de métricas de avaliação (Kappa, Acurácia, Recall, etc.);
- Dados vetoriais (.shp) com pontos de amostragem utilizados para treinamento e validação;

O repositório pode ser acessado no seguinte endereço:

<https://github.com/natanbritoeng/Sapucaia>

APÊNDICE B - TABELA COM GRUPOS E COORDENADAS IDENTIFICADOS NOS ANOS DE 2020 E 2023 E EM TODOS OS MODELOS.

Ano	X	Y
2020 2023	423235	8145235
2020 2023	423235	8145225
2020 2023	435315	8142115
2020 2023	436295	8139405
2020 2023	436285	8139395
2020 2023	436295	8139395
2020 2023	434585	8138925
2020 2023	434595	8138925
2020 2023	434585	8138915
2020 2023	433185	8134905
2020 2023	433195	8134905
2020 2023	428585	8132655
2020 2023	428595	8132655
2020 2023	432005	8131555
2020 2023	432015	8131555
2020 2023	423225	8130575
2020 2023	423225	8130565
2020 2023	423235	8130565
2020 2023	441245	8129985
2020 2023	441235	8129975
2020 2023	441245	8129975
2020 2023	432535	8128915

2020	2023	432545	8128915
2020	2023	432395	8128865
2020	2023	432405	8128865
2020	2023	432395	8128855
2020	2023	431575	8128815
2020	2023	431585	8128815
2020	2023	439945	8128115
2020	2023	439945	8128105
2020	2023	440825	8124575
2020	2023	440815	8124565
2020	2023	440825	8124565
2020	2023	434965	8122945
2020	2023	434965	8122935

APÊNDICE C - DISTRIBUIÇÃO DOS INDIVÍDUOS DE AMBOS OS MODELOS E ANOS.

Foram detectados 35 indivíduos de *L. pisonis* por todos ambos os modelos em ambos os anos, sobre o Modelo Digital de Elevação (MDE) e sua proximidade com os cursos d'água.

