ADILSON ROSA LOPES

SELEÇÃO PRECOCE DE CRUZAMENTOS DE EUCALIPTO TOLERANTES À SECA E PRODUTIVOS UTILIZANDO INTELIGÊNCIA ARTIFICIAL

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Magister Scientiae*.

Orientador: José Augusto Miranda Nacif

Coorientadores: Jean Marcel Sousa Lira
Glêison Augusto dos Santos

FLORESTAL - MINAS GERAIS
2023

Ficha catalográfica elaborada pela Biblioteca da Universidade Federal de Viçosa - Campus Florestal

T

L864s 2023

Lopes, Adilson Rosa, 1985-

Seleção precoce de cruzamentos de eucalipto tolerantes à seca e produtivos utilizando inteligência artificial / Adilson Rosa Lopes. – Florestal, MG, 2023.

1 dissertação eletrônica (105 f.): il. (algumas color.).

Orientador: José Augusto Miranda Nacif.

Dissertação (mestrado) - Universidade Federal de Viçosa, Instituto de Ciências Exatas e Tecnológicas, 2023.

Referências bibliográficas: f. 96-105.

DOI: https://doi.org/10.47328/ufvcaf.2023.014

Modo de acesso: World Wide Web.

1. Eucalipto - Melhoramento genético. 2. Plantas - Efeito da seca. 3. Aprendizagem do computador. 4. Inteligência artificial. I. Nacif, José Augusto Miranda, 1978-. II. Universidade Federal de Viçosa. Instituto de Ciências Exatas e Tecnológicas. Programa de Pós-Graduação em Ciência da Computação. III. Título.

CDD 23. ed. 006.3

ADILSON ROSA LOPES

SELEÇÃO PRECOCE DE CRUZAMENTOS DE EUCALIPTO TOLERANTES À SECA E PRODUTIVOS UTILIZANDO INTELIGÊNCIA ARTIFICIAL

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Magister Scientiae*.

APROVADA: 14 de dezembro de 2023.

Assentimento:

gov.br Documento assinado digitalmente
ADILSON ROSA LOPES
Data: 22/12/2023 06:02:24-0300
Verifique em https://validar.iti.gov.br

Adilson Rosa Lopes
Autor

gov.br Documento assinado digitalmente
JOSE AUGUSTO MIRANDA NACIF
Data: 22/12/2023 06:53:56-0300
Verifique em https://validar.iti.gov.br

José Augusto Miranda Nacif
Orientador

Dedico este trabalho à amada família que constituí, Maria Rita (filha) e Paula (esposa), razão dos meus esforços.

AGRADECIMENTOS

Primeiramente, agradeço a Deus pelo dom da vida, pela sustentação nos momentos difíceis e iluminação nas situações de dúvida e desânimo durante minha caminhada acadêmica. Aos meus pais, Onélia e Antônio, por todo amor e cuidado e por me ensinarem a importância da dedicação ao trabalho. A minha querida irmã Vanda, que, pelo rigor dos seus ensinamentos, me mostrou a importância do empenho nos estudos e, certamente, foram fundamentais nesta conquista. Aos meus irmãos Ailton, José Antônio e Reinaldo pela amizade e bons momentos nas reuniões de família, das quais poucas participei durante o mestrado. Agradeço também a minha esposa Paula por ter sido grande incentivadora para a realização deste trabalho e por sua compreensão nos momentos de ausência e conforto nas horas difíceis. Não poderia deixar de agradecer aquela que é a pessoa mais importante da minha vida, minha filha Maria Rita, por ter sido tão compreensiva nas ausências do papai, sempre reagindo com um triste “tudo bem papai...”, depois de uma negativa a um pedido de brincadeira...esta conquista é nossa! Ao meu orientador, professor José Augusto Miranda Nacif, pela oportunidade, orientação, paciência e amizade nesta empreitada, muito obrigado, Nacif! Ao professor Glêison dos Santos, pela oportunidade e confiança de fazer parte do projeto "Tolerância à Seca em eucalipto” e pela coorientação, na pessoa de quem agradeço aos demais colegas do projeto. Da mesma forma, agradeço ao professor Jean Marcel Souza Lira pelas valiosas contribuições no trabalho, amizade e coorientação. A Sociedade de Investigações Florestais e a Empresa Brasileira de Pesquisa e Inovação Industrial pelo financiamento da pesquisa. A todos os colegas do Laboratório de Engenharia e Sistemas de Computação (LESC) agradeço as conversas, orientações e os momentos de descontração. Ao Laboratório de Visão, Robótica e Imagens da Universidade Federal do Paraná (UFPR), na pessoa do professor David Menotti Gomes, por gentilmente ceder acesso a recursos computacionais empregados para processamento de parte dos experimentos deste trabalho. Agradeço também aos meus colegas de trabalho do Serviço de Tecnologia da Informação da UFV *Campus* Florestal, por todo apoio e por terem "segurado as pontas"durante a minha ausência. À Universidade Federal de Viçosa, pela oportunidade de realizar a pós-graduação. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

"É permanecendo firmes que ireis ganhar a vida!" **(Lc 21,19)**

RESUMO

Lopes, Adilson Rosa, M.Sc., Universidade Federal de Viçosa, dezembro de 2023. **Seleção Precoce de Cruzamentos de Eucalipto Tolerantes à Seca e Produtivos utilizando Inteligência Artificial**. Orientador: José Augusto Miranda Nacif. Coorientadores: Jean Marcel Sousa Lira e Glêison Augusto dos Santos .

A ocorrência cada vez mais frequente de episódios de seca severa e prolongada tem levado as empresas do setor florestal a buscar o desenvolvimento de materiais genéticos tolerantes à seca e com alta produtividade. Para isso, o setor tem utilizado o melhoramento genético florestal. Contudo, no âmbito florestal, o processo de melhoramento para obtenção dos genótipos desejados é mais demorado em comparação com o das culturas agrícolas. Este trabalho de pesquisa teve como objetivo reduzir o tempo necessário para a seleção desses materiais, aplicando técnicas de inteligência artificial (IA) para a classificação precoce de cruzamentos de eucalipto quanto à tolerância à seca e produtividade. Coletamos dados de campo de diferentes cruzamentos de eucaliptos em fases iniciais de desenvolvimento, obtidos por meio de um teste de progênies. Os dados incluíram a Área Foliar Específica, o Potencial Hídrico Foliar, a Área Foliar, a Largura e o Comprimento Foliar, o Incremento Médio Anual Volumétrico e imagens das folhas ao longo de 42 meses. Utilizamos esses dados como entrada nos modelos de IA para prever o comportamento dos materiais genéticos em relação à tolerância à seca e produtividade. Testamos dois grupos de modelos: os clássicos (Floresta Aleatória, Redes Neurais Artificiais, Máquina de Vetores de Suporte e XGBoost) e as redes neurais convolucionais (MobileNetV2, ResNet50 e Xception). Para os dados analisados, observamos que os modelos de redes neurais convolucionais são promissores, com o modelo Xception alcançando uma acurácia de teste de 72%. Esse resultado é importante, pois destaca a IA como uma ferramenta útil no processo de seleção precoce nos programas de melhoramento, além de demonstrar sua aplicação na previsão do comportamento de genótipos, utilizando características funcionais das folhas, que são obtidas de maneira mais rápida e simples.

Palavras-chave: Tolerância à seca em eucalipto; Aprendizado de Máquina; Inteligência Artificial;

ABSTRACT

Lopes, Adilson Rosa, M.Sc., Universidade Federal de Viçosa, december, 2023. **Early Selection of Drought Tolerant and Productive Eucalyptus Crosses Using Artificial Intelligence**. Advisor: José Augusto Miranda Nacif. Co-advisors: Jean Marcel Sousa Lira e Glêison Augusto dos Santos .

The increasingly frequent occurrence of severe and prolonged drought episodes has led companies in the forestry sector to seek the development of drought-tolerant and highly productive genetic materials. For this purpose, the forestry sector has employed forest genetic improvement. However, in the forestry field, the improvement process to obtain desired genotypes is more time-consuming compared to agricultural crops. This research aimed to reduce the time required for selecting these materials by applying artificial intelligence (AI) techniques for early classification of eucalyptus crosses regarding drought tolerance and productivity. We gathered field data from different eucalyptus crosses in early development stages, obtained through a progeny test. The data included Specific Leaf Area, Leaf Water Potential, Leaf Area, Leaf Width and Length, Annual Volumetric Mean Increment, and leaf images over 42 months. We utilized this data as input for AI models to predict the behavior of genetic materials concerning drought tolerance and productivity. We tested two groups of models: classic ones (Random Forest, Artificial Neural Networks, Support Vector Machine, and XGBoost) and convolutional neural networks (MobileNetV2, ResNet50, and Xception). For the analyzed data, we observed promising results in convolutional neural network models, with the Xception model achieving a test accuracy of 72%. This outcome is significant as it underscores AI as a useful tool in the early selection process within improvement programs, demonstrating its application in predicting genotype behavior using functional leaf traits obtained more quickly and simply.

Keywords: Drought Tolerance in Eucalyptus; Machine Learning; Artificial intelligence;

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

Capítulo 1

Introdução

As espécies do gênero *Eucalyptus ssp. (Myrtaceae)* são amplamente cultivadas em todo o mundo, especialmente em regiões tropicais e subtropicais (Myburg et al., 2014). No Brasil, o cultivo de eucalipto ocupou uma área de 7,53 milhões de hectares em 2021, representando 75,8% do total de florestas plantadas no país (IBÁ, 2021). Essa preferência pelas espécies de *Eucalyptus* é impulsionada por seu crescimento rápido, qualidade da madeira, forma do fuste e alta adaptabilidade fenotípica, além de sua capacidade de fornecer uma ampla variedade de produtos madeireiros e não madeireiros para mercados nacionais e internacionais (Zaiton et al., 2020, IBÁ, 2021).

A despeito das melhorias significativas em práticas de manejo, preparo do solo, fertilização e controle de pragas e doenças, o desenvolvimento e a produtividade do eucalipto continuam a ser afetados pela variabilidade climática. A escassez de água e temperaturas extremas, fora da faixa ideal de 8,5 ºC a 40 ºC, emergem como fatores críticos que impactam a sobrevivência e o rendimento das plantações de eucalipto (Florêncio et al., 2022, Oliveira, 2021). Um exemplo notório dessa influência ocorreu em 2015, quando Minas Gerais perdeu 200.000 hectares de florestas plantadas de eucalipto devido a uma prolongada seca (Gonçalves et al., 2017). Diante da crescente frequência e intensidade prevista para eventos climáticos extremos devido às mudanças climáticas globais (Shukla et al., 2019), as empresas do setor florestal brasileiro têm sido alertadas para a necessidade de encontrar soluções que garantam a sustentabilidade de suas operações.

Para enfrentar esses desafios, o setor florestal tem investido no desenvolvimento de materiais genéticos de eucalipto que sejam tolerantes à seca, ao mesmo tempo que mantêm níveis viáveis de produtividade. Um programa de melhoramento genético florestal abordou essa questão cruzando espécies de eucalipto reconhecidas no meio florestal por seu bom desempenho em relação à tolerância à seca e produtividade. Como parte desse projeto, foi estabelecido um teste de progênie em 2019 na cidade de Buritizeiro-MG, uma região com histórico de baixos índices pluviométricos, com uma duração planejada de sete anos, correspondendo a um ciclo completo de rotação do eucalipto. O objetivo é o desenvolvimento de genótipos superiores aos disponíveis no mercado.

O desenvolvimento de novos genótipos no melhoramento florestal é um pro-

cesso de longo prazo devido aos ciclos de reprodução das espécies. No entanto, é possível acelerar esse processo por meio da seleção precoce de cruzamentos (Castro et al., 2021). Essa abordagem envolve a identificação de características de interesse em plantas jovens que possam prever o comportamento de indivíduos adultos em idade produtiva, antecipando os ganhos genéticos (Moraes et al., 2014, Corrêa et al., 2017). Algumas variáveis fisiológicas, como o potencial hídrico foliar (ψf) e a área foliar específica (AFE), têm sido apontadas na literatura como parâmetros robustos de tolerância à seca, além de serem medidas de forma relativamente simples e rápida (Conti Junior, 2019, Oliveira, 2021). Assim, a aplicação de modelos preditivos com base nessas variáveis mostra-se uma oportunidade promissora, já que esses modelos têm a capacidade de identificar padrões e características sutis nos dados. Realizamos uma busca detalhada na literatura científica utilizando plataformas como o *Google Scholar* e aplicamos diversas *strings* de busca, em português e em inglês, incluindo termos como 'IA aplicada ao melhoramento florestal', 'aprendizado de máquina aplicado ao melhoramento florestal', 'aprendizado de máquina para predição de tolerância à seca em eucalipto' e 'previsão de produtividade em eucalipto usando aprendizado de máquina'. No entanto, até o momento de redação deste trabalho, não encontramos referências ou relatos que descrevam o uso específico desses modelos para prever o comportamento futuro de cruzamentos de eucalipto em relação ao seu padrão de tolerância à seca e produtividade. Essa ausência de literatura existente sugere uma lacuna significativa na aplicação de técnicas de IA para essa finalidade específica no contexto do melhoramento florestal, ressaltando a originalidade e contribuição deste estudo para a área.

Este trabalho de pesquisa teve como objetivo ajustar um modelo para automatizar a seleção precoce de cruzamentos de eucalipto com base em variáveis fisiológicas e imagens das folhas de plantas jovens coletadas em campo aos seis meses de plantio, no âmbito do projeto de melhoramento genético florestal mencionado anteriormente. Esses dados foram rotulados de acordo com os critérios de tolerância à seca e produtividade em diferentes idades posteriores aos seis meses e, em seguida, usados como entrada para treinar os modelos a fim de classificar os cruzamentos nessas idades mais avançadas, incluindo 18, 30, 36 e 42 meses (a última idade de coleta de dados disponível). Dessa forma, desenvolvemos a modelagem da seleção precoce, antecipando o tempo necessário para a seleção de materiais em apenas seis meses. Foram testados modelos de Aprendizado de Máquina (*Machine Learning - ML*) clássicos, *Random Forest (RF)*, *XGBoost*, *Support Vector Machines (SVM)* e *Artificial Neural Network (MLP)*, utilizando as variáveis fisiológicas contínuas como entrada de dados. Além destes, ajustamos três modelos de Redes Neurais Convolucionais (CNN), *MobileNetV2*, *Xception* e *ResNet50*, com base nas imagens das folhas. Também buscou-se identificar as variáveis que mais influenciaram os modelos clássicos e as características das folhas

que foram mais relevantes nas previsões das CNN.

Os resultados alcançados destacam o potencial do uso da IA na seleção precoce de materiais genéticos de eucalipto resistentes à seca e produtivos. O modelo Random Forest alcançou uma acurácia de 53,75% na idade de 42 meses para os dados tabulados, enquanto o modelo Xception obteve uma acurácia de 72% para os dados de imagem. A variável mais importante identificada pelo Random Forest foi a Área Foliar, enquanto a área da folha se destacou como a característica mais relevante detectada pelo melhor modelo de CNN.

Em resumo, este estudo teve como objetivo avaliar o desempenho de algoritmos de aprendizado de máquina e aprendizado profundo para apoiar os programas de melhoramento florestal nos ciclos de seleção de materiais genéticos de eucalipto resistentes à seca e produtivos, visando a seleção precoce destes materiais para redução de custos financeiros e de tempo. Além disso, buscou identificar as variáveis-chave e características das folhas com potencial preditivo das características de interesse, contribuindo, assim, para a efetividade dos programas de melhoramento genético florestal. Até onde sabemos, esta pesquisa representa uma das primeiras tentativas de aplicar Inteligência Artificial para prever o comportamento futuro de cruzamentos de eucalipto quanto ao seu potencial de tolerância ao estresse hídrico e produtividade.

1.1 Problema

A manutenção de altos níveis de produtividade das plantações de eucalipto diante de eventos climáticos extremos, como secas intensas, tem sido uma preocupação permanente do setor florestal. Uma técnica amplamente empregada pelas empresas do setor para tentar mitigar este problema é o melhoramento genético florestal, uma estratégia que envolve a manipulação e seleção de características genéticas em árvores com o objetivo de criar novos genótipos que apresentem atributos desejáveis para fins comerciais (Resende et al., 2010). Portanto, esta é uma abordagem empregada para o desenvolvimento de materiais genéticos adaptados a estes eventos, ajudando a garantir a viabilidade econômica das plantações florestais mesmo em regiões com condições climáticas desfavoráveis.

Embora o melhoramento genético florestal seja uma solução eficaz para contornar as condições desafiadoras impostas pelas secas cada vez mais intensas e duradouras, alguns problemas ainda impõe um gargalo na sua aplicação. Um deles se refere à identificação de características a serem exploradas em indivíduos superiores, que sejam potenciais bioindicadoras de tolerância à seca e produtividade, de modo que possam ser transmitidas para aprimoramento das mesmas em futuras gerações. Diversos autores na literatura descrevem a complexidade de se entender os fatores relacionados aos efeitos da deficiência hídrica sobre as plan-

tas (Oliveira, 2021, Pita-Barbosa et al., 2023, Bianchi et al., 2016), não havendo um mecanismo genérico de tolerância à seca, uma vez que as plantas têm respostas condicionadas por diferentes processos adaptativos quando submetidas ao estresse hídrico (Mittler, 2006). As pesquisas citadas acima buscam avaliar o desenvolvimento de atributos citados na literatura como potenciais bioindicadores de tolerância como, por exemplo, a área foliar, área foliar específica, potencial hídrico foliar, densidade estomática, densidade de venação, tamanho das raízes, entre outros, em plantas submetidas à seca. No entanto, os trabalhos recomendam certas características das plantas como potenciais preditoras de tolerância à seca, mas reforçam que este assunto ainda demanda mais estudos, pois, ainda que os processos de tolerância à seca sejam avaliados muitas vezes de forma isolada, as espécies combinam diferentes estratégias para a sua sobrevivência em condições desfavoráveis de disponibilidade hídrica (Vellini et al., 2008).

Outro problema que impacta fortemente a seleção de genótipos de interesse nos programas de melhoramento florestal são os longos ciclos de rotação das culturas florestais, podendo durar de sete a oito anos (Paludeto et al., 2017). Portanto, é um processo que consome considerável quantidade de tempo e recursos financeiros para seleção dos materiais de interesse. Vários estudos têm sido publicados tendo como finalidade a melhora da eficiência da seleção em espécies florestais e convergem no sentido de que, para aumentar o ganho genético por unidade de tempo, a seleção precoce é importante na maioria dos programas de melhoramento (Tambarussi et al., 2017). A ferramenta que tem norteado estes estudos é a correlação estimada fenotípica ou genotipicamente, definida como uma quantificação da magnitude da associação genotípica de caracteres entre indivíduos em diferentes idades (Paludeto et al., 2017).

Diante do exposto, este trabalho de pesquisa aborda duas questões relevantes no melhoramento genético florestal, a identificação de variáveis que têm relevância na predição da tolerância à seca e produtividade e os longos períodos de tempo necessários para a seleção de materiais genéticos de interesse, no contexto do trabalho, os tolerância à seca e produtivos. Tornar a seleção destes materiais mais eficiente é essencial para enfrentar os desafios impostos pelas mudanças climáticas, garantir a sustentabilidade da indústria florestal e manter o fornecimento de produtos essenciais para a sociedade. Assim, a pesquisa busca contribuir para solução destes problemas aproveitando as ferramentas do aprendizado de máquina para identificar as variáveis com maior importância relativa na predição da tolerância ao estresse hídrico e produtividade e classificar precocemente os genótipos de eucalipto de alto desempenho.

1.2 Objetivo

O objetivo central desta pesquisa foi avaliar o desempenho de modelos de aprendizado de máquina, incluindo abordagens clássicas e redes neurais convolucionais, para apoiar a seleção precoce de cruzamentos de eucalipto que sejam tolerantes à seca e produtivos, bem como na identificação de variáveis com maior importância preditiva neste processo. Isso foi realizado buscando antecipar o desenvolvimento de genótipos de interesse por meio da predição do comportamento desses cruzamentos, formados por plantas jovens, em relação a essas duas características em idade produtiva. Para atingir esse objetivo central, os objetivos específicos desta pesquisa foram os seguintes:

1. Realizar uma análise exploratória dos dados fisiológicos e de imagem a fim de obter um entendimento mais aprofundado dos dados e do problema em estudo, identificando tendências, dados faltantes, limitações e outras informações relevantes.
2. Realizar o pré-processamento dos dados, incluindo o tratamento de dados faltantes, normalização e curadoria das imagens, para prepará-los adequadamente para serem utilizados como entrada para os modelos de aprendizado de máquina.
3. Rotular os dados de entrada considerando quatro classes que representam diferentes níveis de tolerância à seca e produtividade.
4. Aplicar os modelos clássicos de Aprendizado de Máquina (ML), como *Random Forest*, *XGBoost*, *SVM* e *MLP*, aos dados fisiológicos.
5. Aplicar modelos de Rede Neural Convolucional (CNN) aos dados de imagem para realizar a predição das características de interesse.
6. Avaliar o desempenho dos modelos por meio de métricas amplamente citadas na literatura, além de utilizar a matriz de confusão e acompanhar as curvas de acurácia e perda durante o treinamento (para as arquiteturas de CNN).
7. Investigar e identificar quais das variáveis de fisiologia tiveram maior importância relativa na predição dos modelos, contribuindo para a compreensão das características com maior potencial preditivo de tolerância à seca e produtividade.
8. Visualizar e analisar quais características das folhas chamaram mais atenção da CNN durante o treinamento, fornecendo insights sobre as informações que a rede neural considerou relevantes para a predição, o que também pode contribuir com a identificação de variáveis importantes para a determinação de tolerância à seca e produtividade.

Esses objetivos específicos foram desenvolvidos para guiar a pesquisa e permitir uma abordagem abrangente na avaliação do desempenho dos modelos de aprendizado de máquina na seleção precoce de cruzamentos de eucalipto.

1.3 Contribuições

Este trabalho busca contribuir com a área de melhoramento genético florestal, ao aplicar modelos de IA visando a automação da seleção precoce de cruzamentos de eucalipto tolerantes à seca e produtivos. Embora haja uma extensa literatura sobre o uso de ML em diferentes áreas da ciência florestal, até onde sabemos, a aplicação dessas técnicas na predição precoce do comportamento de plantas de eucalipto quanto ao seu potencial de tolerância à seca e produtividade não foi explorada anteriormente. Objetivamente, as contribuições da pesquisa foram:
- Ajuste de quatro modelos de aprendizado de máquina clássicos a partir dos dados fisiológicos coletados em campo para definição do que obteve o melhor desempenho para o problema estudado.
- Definição das variáveis trabalhadas que tiveram maior importância relativa nas predições dos modelos.
- Ajuste de três modelos de rede neural convolucional (CNN) a partir de imagens das folhas das amostras coletadas em campo, a fim de selecionar o que alcançou a melhor performance preditiva.
- Visualização das áreas da folha que mais chamaram a atenção da CNN com melhor desempenho, visando a identificação de potenciais atributos da folha que poderiam contribuir para a identificação de materiais tolerantes.
Além dos resultados acima descritos, foi criado um painel de visualização para a etapa de análise exploratória dos dados do projeto de melhoramento florestal, com o intuito de facilitar o entendimento dos dados, obtenção de insights e identificação de tendências e padrões nos dados. Além do painel, durante o período do mestrado foi realizado um experimento de predição precoce de uma variável silvicultural de produtividade volumétrica, o Incremento Médio Anual Volumétrico (IMAVol m^3/ha/ano), por meio da aplicação dos mesmos quatro modelos de aprendizado de máquina clássicos já mencionados. Este trabalho de pesquisa resultou em uma publicação de artigo em conferência (Lopes et al., 2023), intitulado Predição do Incremento Médio Anual Volumétrico de Eucalyptus com Aprendizado de Máquina.

1.4 Estrutura da dissertação

A dissertação está organizada na forma de capítulos e, a partir da introdução, o texto está estruturado da seguinte maneira: o Capítulo 2 apresenta a contextualização, descrevendo o embasamento teórico das técnicas de aprendizado de máquina e o cenário de aplicação na pesquisa. O Capítulo 3 aborda os trabalhos relacionados com este estudo. Já o Capítulo 4, mostra a metodologia utilizada para desenvolver a dissertação. No Capítulo 5 são mostrados e discutidos os resultados obtidos, tanto para os mo-

delos aplicados aos dados fisiológicos quanto para os que foram que tem as imagens com entrada. As conclusões e trabalhos futuros que podem ser desenvolvidos para cada tipo de modelo estão descritas no Capítulo 6.

Capítulo 2

Contextualização

O melhoramento genético florestal é uma ferramenta amplamente utilizada para a manipulação e seleção de características genéticas em espécies florestais, visando criar novos genótipos com características desejáveis para aplicação comercial ou de interesse específico (Resende et al., 2010). No âmbito dos programas de melhoramento, o teste de progênie se destaca como um recurso essencial para avaliar o desempenho dos descendentes de um progenitor em diferentes ambientes (VanRaden, 2008).

No entanto, o período de aproximadamente sete anos para conclusão do teste de progênie representa um desafio significativo para os programas de melhoramento florestal. Esta duração prolongada gera um gargalo no tempo necessário para a seleção de materiais de interesse nestes programas, resultando em custos financeiros e temporais elevados. Este longo período de tempo é crítico, especialmente quando se busca identificar características relacionadas à tolerância à seca e à produtividade.

A principal forma de reduzir esse tempo é por meio da seleção precoce. Assim, os especialistas em melhoramento florestal têm buscado identificar características em plantas mais jovens que estejam relacionadas com aquelas de interesse econômico na fase de corte, ou seja, predizer pelas árvores em estágios o mais juvenil possível o comportamento de um indivíduo adulto, reduzindo, assim, o tempo necessário para completar o ciclo de seleção, gerando maior ganho genético por unidade de tempo (Massaro, 2008). Diversas técnicas vêm sendo empregadas a fim de avaliar a viabilidade e a eficiência da seleção precoce, tais como a estimativa da correlação genética nas diferentes idades (Reis et al., 2015, Moraes et al., 2014, Massaro, 2008), a estimativa dos parâmetros genéticos e fenotípicos (Paludeto et al., 2017, Garuzzo, 2022) e a identificação de variáveis potenciais bioindicadoras de tolerância à seca para auxílio na seleção precoce (Oliveira, 2021, Pita-Barbosa et al., 2023).

Nesse contexto, esta pesquisa propõe o uso de técnicas de aprendizado de máquina para prever antecipadamente quais genótipos possuirão as características desejadas em idades produtivas. O objetivo é acelerar a identificação desses materiais, reduzindo o tempo necessário para os testes de progênies em futuros programas de melhoramento genético florestal, focados em tolerância à seca e produtividade. Para isso, modelos de aprendizado de máquina clássicos (*Random Forest*, *SVM*, *XGBoost* e *MLP*) e modelos de redes neurais convolucionais (*MobileNetV2*, *Xception* e *Resnet50*)

foram ajustados e treinados utilizando variáveis fisiológicas e imagens das folhas das plantas, respectivamente, coletadas aos seis meses de plantio (fase de plantas jovens) em um teste de progênies. A modelagem da seleção precoce foi feita rotulando essas amostras com as características de tolerância e produtividade das plantas aos 42 meses, dados de plantas em idade próxima à fase produtiva disponíveis. Assim, buscamos prever a tolerância à seca e a produtividade aos 42 meses usando dados coletados aos 6 meses.

2.1 Aprendizado de máquina

Um algoritmo de aprendizado de máquina é um processo computacional que utiliza dados de entrada para executar uma tarefa desejada sem ser rigidamente programado para produzir um resultado específico. Esses algoritmos adaptam automaticamente sua arquitetura por meio de repetição, tornando-se mais especializados na realização da tarefa desejada. Esse processo de adaptação é conhecido como treinamento, no qual dados de entrada são fornecidos com resultados esperados e o algoritmo se ajusta para produzir o resultado o mais próximo possível do esperado para as entradas de treinamento e para dados novos não vistos. O treinamento é a parte "aprendizado"da aprendizagem de máquina (El Naqa and Murphy, 2015).

O objetivo da aprendizagem de máquina é imitar a forma como os seres humanos aprendem a processar sinais sensoriais para atingir um objetivo. Isso pode envolver tarefas como reconhecimento de padrões, como, por exemplo, distinguir cachorros de gatos. Em vez de programar a máquina com representações exatas destes objetos, ela é treinada repetidas vezes com exemplos dos mesmos. Esse treinamento envolve a capacidade de aprender automaticamente a partir dos dados, em vez de seguir regras explicitamente definidas (Hamedianfar et al., 2022).

De maneira geral, o aprendizado de máquina pode ser dividido de acordo com a maneira como os dados são rotulados em aprendizado supervisionado, não supervisionado e semi-supervisionado (Hamedianfar et al., 2022). Na abordagem supervisionada, onde os dados de entrada são fornecidos juntamente com seus rotulados conhecidos, como no exemplo acima em que os dados de entrada são rotulados nas classes gato e cachorro. No aprendizado não supervisionado, o algoritmo tenta encontrar padrões nos dados sem ter saídas esperadas informadas, ou seja, os dados de entrada não são rotulados. Uma forma de aprendizado de máquina não supervisionado muito aplicada é o aprendizado por reforço, no qual um agente tenta realizar uma sequência de ações que podem gerar uma recompensa ou penalidade, de forma cumulativa, a fim de maximizar os acertos de algoritmo durante o treinamento (Sutton and Barto, 1998). A aprendizagem semi-supervisionada é uma combinação de ambas, supervisionada e não supervisionada, em que parte dos dados é parcialmente

rotulada, e essas informações rotuladas são utilizadas para auxiliar na inferência e aprendizado da parte não rotulada. Isso permite aproveitar o conhecimento disponível nos dados rotulados, ao mesmo tempo que explora padrões e estruturas nos dados não rotulados (El Naqa and Murphy, 2015).

O pesquisador Arthur Samuel, da IBM, foi o primeiro a usar o termo "aprendizado de máquina", ainda na década de 50 (Samuel, 1959). Em 1958, Rosenblatt desenvolveu uma das primeiras arquiteturas de redes neurais, conhecida como o *Perceptron* (Rosenblatt, 1958). No entanto, o *Perceptron* tinha limitações, pois sua capacidade de classificação se restringia a problemas linearmente separáveis, não sendo adequado para problemas não lineares, limitação que foi superada em 1975 com a criação do *Perceptron* de Camada Múltipla (*Multilayer Perceptron - MLP*) por Werbos (Werbos, 1974). Posteriormente, em 1986, Quinlan desenvolveu árvores de decisão (*Decision Tree*) (Quinlan, 1986), e em 1995, Cortes e Vapnik introduziram as máquinas de vetores de suporte (*Support Vector Machine - SVM*) (Cortes and Vapnik, 1995). Mais tarde, surgiram algoritmos de aprendizado de máquina de conjunto, como as florestas aleatórias (*Random Forest - RF*) (Breiman, 2001) e o *XGBoost* mais recentemente (Chen and Guestrin, 2016). Atualmente, os algoritmos de aprendizado profundo demonstraram uma capacidade impressionante para aprender representações de dados que facilitam a extração de informações úteis na construção de preditores. Todas essas técnicas compartilham o objetivo comum de treinar computadores para realizar tarefas de forma inteligente, indo além da computação tradicional, e isso é alcançado por meio de exemplos repetidos (El Naqa and Murphy, 2015).

2.2 Aprendizado profundo

Aprendizado Profundo (*Deep Learning - DL*) é uma subcategoria do Aprendizado de Máquina . No contexto do Aprendizado Profundo, máquinas são treinadas para executar tarefas complexas, como o reconhecimento de padrões em imagens, com base em modelos de aprendizado estatístico e algoritmos. A detecção desses padrões é feita por uma técnica chamada visão computacional. Nas abordagens clássicas de visão computacional, há necessidade de uma etapa manual de execução de algoritmos específicos para pré-processar as imagens de modo a extrair suas características antes de usá-las como entrada para treinar um classificador. Já nos modelos de DL, estas etapas são encapsuladas em um único processo, no entanto, geram modelos complexos que frequentemente requerem mais recursos computacionais, como GPUs (dos Santos et al., 2022).

O Aprendizado Profundo se baseia em redes neurais artificiais, que são estruturas computacionais inspiradas na biologia dos neurônios cerebrais. Em uma rede neural, informações são transferidas de uma camada de entrada para uma camada de

saída por meio de camadas ocultas, permitindo que o modelo adquira gradualmente características de nível superior à medida que aprende a representação dos dados. Esse processo de aprendizado de atributos abstratos nos dados é fundamental para o sucesso do Aprendizado Profundo (LeCun et al., 2015).

2.3 Redes Neurais Convolucionais - CNN

As Redes Neurais Convolucionais (CNNs) são uma categoria de técnicas de Aprendizado Profundo (*Deep Learning - DL*) projetadas para processar dados, como imagens, que são representados na forma de matrizes. Quando aplicadas a uma rede de classificação de imagens, uma entrada composta por imagens coloridas será representada por três matrizes 2D, sendo duas dimensões relacionadas à altura e largura da imagem, e uma matriz para cada um dos três canais de cor RGB (vermelho, verde e azul). Cada matriz contém intensidades de pixels para a respectiva cor representada pelo canal (LeCun et al., 2015).

As CNNs exploram a estrutura espacial dos dados por meio de camadas de convolução, desempenhando um papel crucial na detecção de padrões e características específicas e complexas nas imagens. Essa capacidade torna as CNNs altamente eficazes na captura de informações relevantes, tornando-as especialmente adequadas para tarefas de visão computacional (Hamedianfar et al., 2022).

A camada convolucional em uma CNN utiliza filtros com pesos, em forma de uma matriz 2D, para calcular mapas de características que destacam padrões relevantes na entrada. Durante o treinamento da rede, os pesos dos filtros são ajustados para que a camada possa aprender a identificar padrões importantes, como bordas, texturas ou características de objetos. Após a aplicação dos filtros, os mapas de características gerados passam por uma função de ativação não-linear, como a ReLU, antes de serem usados para representar informações nas camadas subsequentes da rede (LeCun et al., 2015).

A camada de *pooling* é aplicada para reduzir a dimensionalidade dos mapas de características. Essa operação combina regiões de ativação em um único valor, ajudando a tornar as representações mais invariantes a pequenos deslocamentos. Geralmente, é aplicada como uma camada de transição entre duas camadas de convolução, realizando uma subamostragem da entrada para extrair as partes mais relevantes e diminuir o tamanho da imagem de entrada. Isso reduz a carga computacional, uma vez que menos parâmetros precisam ser ajustados durante o treinamento (Goodfellow et al., 2016).

O último segmento da CNN, a camada de classificação, consiste, em geral, em uma rede neural multicamadas que recebe como entrada um vetor de características extraídas pelas camadas anteriores e faz a predição da classe para a amostra infor-

mada. Normalmente, essa rede é uma *Multi-Layer Perceptron (MLP)* com todos os neurônios conectados entre as camadas, ou seja, uma rede totalmente conectada. A camada de classificação utiliza uma função de ativação Softmax. A função Softmax converte as saídas dos neurônios em probabilidades, garantindo que a soma das probabilidades para todas as classes seja igual a 1, e a saída da rede para a amostra de entrada será a classe com maior probabilidade atribuída (Krizhevsky et al., 2012). A Figura 2.1 ilustra uma arquitetura típica de CNN, destacando segmentos gerais, como a extração de características e a camada de classificação.

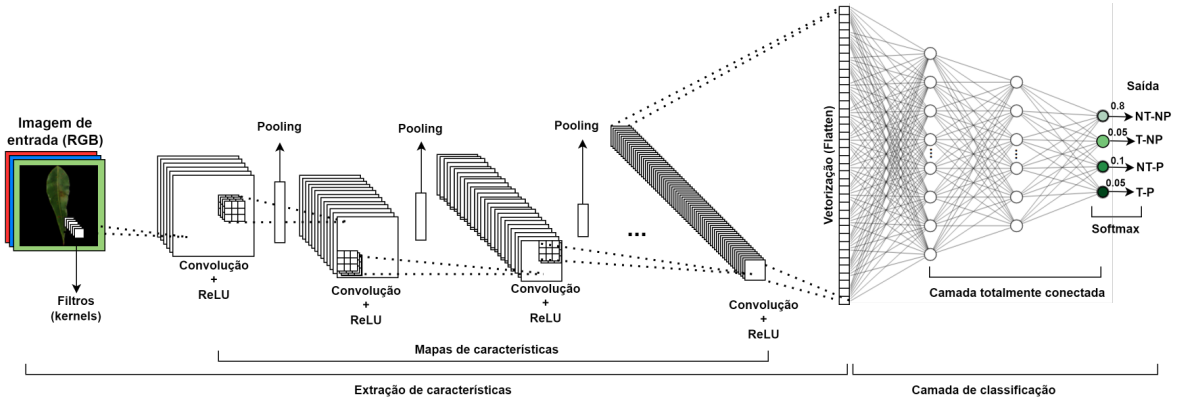

Figura 2.1: Elementos básicos da arquitetura de uma CNN

2.4 Ajuste fino e transferência de aprendizado

É comum realizar ajustes na estrutura e nos parâmetros das redes neurais para adaptá-las a problemas e contextos específicos, um procedimento conhecido como ajuste fino (*fine-tuning*) (Käding et al., 2017). O ajuste fino da arquitetura de um classificador como uma CNN, envolve adicionar ou remover camadas convolucionais, camadas de *pooling*, camadas totalmente conectadas ou até mesmo mudar a arquitetura base da rede. Isso pode ocorrer quando a tarefa-alvo exige uma adaptação da arquitetura da rede. Já o ajuste fino dos parâmetros é o processo de ajustar ou atualizar os pesos de uma rede neural pré-treinada em uma tarefa específica, enquanto se mantém a arquitetura da rede inalterada. Isso pode incluir a alteração dos pesos em camadas específicas ou de todas as camadas da rede, bem como a camada totalmente conectada (topo da rede), para se adaptar à tarefa em questão (He et al., 2016, Simonyan and Zisserman, 2014).

Além dessa flexibilidade de configurações, pode-se ainda aproveitar o aprendizado previamente adquirido por uma CNN em tarefas relacionadas ou mais abrangentes, aplicadas a grandes conjuntos de dados como o ImageNet, por meio da técnica de transferência de aprendizado (*transfer learning*) (Kaya et al., 2019). Por meio desta

técnica, o conhecimento adquirido em uma tarefa específica em um determinado domínio pode ser utilizado para melhorar o aprendizado da função preditiva em outra tarefa, em outro domínio. Essas estratégias permitem adaptar a CNN para solucionar uma variedade de problemas que envolvem o aprendizado de representações visuais (Tetila et al., 2020).

No presente estudo, as CNNs foram aplicadas com e sem transferência de aprendizado, com o intuito de avaliar a configuração com melhor resultado. Ao ajustar os modelos com transferência de aprendizado, os pesos das redes foram inicializados com os pesos pré-treinados no conjunto de dados do ImageNet (ImageNet, 2016), com o ajuste fino dos mesmos durante o processo de treinamento. A configuração com transferência de aprendizado obteve melhor resultado nos experimentos, portanto, os resultados apresentados são referentes à aplicação das CNNs com transferência de aprendizado.

Capítulo 3

Trabalhos Relacionados

Na ciência florestal as técnicas de ML vêm sendo aplicadas em vários estudos de diversas áreas, como exemplificam os estudos descritos nesta seção, mostrando o potencial de uso do aprendizado de máquina na área florestal. No entanto, até o momento desta pesquisa, não foram identificados estudos na literatura com o mesmo escopo e objetivo de antecipar a identificação de cruzamentos de espécies florestais quanto à tolerância à seca e produtividade utilizando modelos de aprendizado de máquina. Esta lacuna na pesquisa motiva e justifica a presente investigação, representando uma contribuição importante para esta área emergente e promissora do conhecimento.

3.1 Modelos clássicos

Há considerável quantidade de pesquisas para predição de produtividade florestal, como as descritas nos trabalhos de (da Silva Tavares Júnior et al., 2020) no qual foram aplicados os modelos *RF, MLP, SVM* para predizer o incremento do diâmetro de *Eucalyptus* em fragmentos de floresta de mata atlântica. Os modelos *XGBoost*, *RF*, *SVM*, Àrvores de Decisão e Regressão Linear foram aplicados nas pesquisas de (Li et al., 2022, Li et al., 2020) para predição de biomassa florestal acima do solo. (Cordeiro et al., 2022, Nunes and Görgens, 2016) usaram *RF*, *SVM* e *MLP* para estimar volume em plantações de *Eucalyptus*. (Lopes et al., 2023) aplicaram *RF*, *XGBoost*, *MLP* e *SVM* para predição do Incremento Médio Anual volumétrico (IMAvol) de *Eucalyptus*. Outra tarefa onde ML é amplamente utilizado é na fenotipagem florestal, como abordado pelos trabalhos de (Alba et al., 2022) que trabalharam na identificação de florestas sazonalmente secas com *KNN (K-Nearest Neighbors*), *RF*, *SVM* e *MLP*. Os modelos *XGBoost*, *LightGBM* e *CatBoost* para caracterização de produtividade de florestas em larga escala (Bombrun et al., 2020). Para classificação de espécies de árvores baseado em imagens de sensoriamento remoto, (Welle et al., 2022, Łoś et al., 2021) aplicaram com *XGBoost*, *LightGBM*, *RF*, *SVM* e *KNN*. (Su et al., 2018) aplicaram *MLP* e aprendizado profundo a partir de imagens aéreas de florestas para predição da altura de árvores de crescimento rápido, a fim de evitar acidentes com linhas de transmissão de energia; (Liao et al., 2022) usaram *RF*, a partir de imagens hiperes-

pectrais adquiridas por meio de drone, para detecção de diferentes níveis de doença em plantações de eucalipto. Também a partir de imagens RGB obtidas por meio de drone, (dos Santos et al., 2022) compararam a abordagem clássica *MLP* e o modelo de Aprendizado Profundo (*Deep Learning - DL*) *YOLOV5* para identificar e medir ninhos de formiga cortadeira em plantações de eucalipto. Outros trabalhos se destacam pela previsão de condições ambientais, como os trabalhos de (Ghafarian et al., 2022) para previsão de regimes de temperatura em florestas com *XGBoost*, *RF*, *SVM* e *MLP* e (Ismail and Mutanga, 2010) com previsão de estresse hídrico em florestas de *Pinus Patula* com *RF*. (Csillik et al., 2019) monitoraram estoque de carbono em florestas tropicais usando *RF*.

3.2 Aprendizado profundo

A aplicação do aprendizado profundo na área da ciência florestal tem sido extensivamente explorada, como evidenciado por estudos relevantes na literatura. Por exemplo, revisões como a de (Hamedianfar et al., 2022) destacam a diversidade de tarefas nas quais o aprendizado profundo tem sido aplicado para apoiar o inventário florestal. Além disso, levantamentos bibliográficos, como o realizado por (Liu et al., 2018), exploraram o emprego de técnicas de Machine Learning na ecologia florestal, oferecendo uma visão abrangente das aplicações nesse campo. Adicionalmente, a revisão de (Coelho Eugenio et al., 2021) investigaram o uso de dados obtidos por sistemas aéreos não tripulados em conjunto com o aprendizado de máquina em aplicações no contexto da ciência florestal. Da mesma forma, o trabalho de (Jain et al., 2020) abordou aplicações específicas do aprendizado de máquina no âmbito da ciência e gerenciamento de incêndios florestais. Nos estudos de (García-Gutiérrez et al., 2016, Zhang et al., 2019) foi explorado o uso de redes neurais profundas para estimar produção de biomassa de florestas, fazendo uso da extração de características de imagens aéreas LiDAR, alcançando um R^2 de 81% e 93%, respectivamente. No estudo de (Talebiesfandarani and Shamsoddini, 2022) foram aplicadas arquiteturas de CNN para extração de características de imagens de sensoriamento remoto e modelos clássicos para estimativa biomassa de florestas, obtendo R^2 de 88%. No trabalho de (Guan et al., 2015) foi realizada a classificação de espécies de árvores a partir de imagens LiDAR com aplicação de Deep Boltzmann Machines para extração de características e um classificador Support Vector Machine, alcançando 86% de acurácia. Aplicando CNN para as tarefas de segmentação de imagens de árvores registradas no solo e estimativa do volume de estoque florestal a partir das imagens segmentadas, (Liu et al., 2019) chegaram a uma acurácia na segmentação de 96% e um R^2 de 81%. Em trabalhos de fenotipagem florestal empregando aprendizado profundo, foram realizadas a detecção de copas de árvores (Weinstein et al., 2019) com 81% de

recall e 69% de precisão, o reconhecimento de pinheiros doentes (Hu et al., 2020) com 91% de acurácia, e a distinção entre árvores coníferas e decíduas (Hamraz et al., 2019) alcançando 91% de acurácia.

Tabela 3.1: Exemplos de aplicações de modelos clássicos de ML na área florestal

Aplicações	**Modelos**	**Estudos**
Predição de incremento de produtividade	*RF*, *SVM*, *MLP*	da Silva Tavares Júnior et al., 2020
Predição de biomassa florestal acima do solo	*XGBoost*, *RF*, *SVM*, *DT*, *LR*	Li et al., 2022; Li et al., 2020
Estimativa de volume em florestas plantadas	*RF*, *SVM* e *MLP*	Cordeiro et al., 2022; Nunes e Görgens, 2016
Predição do Incremento Médio Anual volumétrico (IMAvol) de *Eucalyptus*	*RF*, *XGBoost*, *MLP* e *SVM*	Lopes et al., 2023
Ecologia florestal	*RF*, *SVM* e *MLP*	Liu et al., 2018
Identificação de florestas tropicais sazonalmente secas	*KNN*, *RF*, *SVM* e *MLP*	Alba et al., 2022
Fenotipagem de florestas	*XGBoost*, *LightGBM* e *CatBoost*	Bombrun et al., 2020
Classificação de espécies de árvores baseado em dados de sensoriamento remoto	*XGBoost*, *LGBM*, *RF*, *SVM* e *KNN*	Łoś et al., 2021; Welle et al. 2022
Previsão de regimes de temperatura dentro de florestas a partir de dados meteorológicos	*XGBoost*, *RF*, *SVM* e *MLP*	Ghafarian et al., 2022
Previsão do estresse hídrico em florestas de *Pinus patula*	*RF* e outros de *Bagging e Boosting*	Ismail and Mutanga, 2010
Monitoramento de estoque de carbono em florestas tropicais	*RF*	Csillik et al., 2019
Predição da altura de árvores de crescimento rápido, a fim de evitar acidentes com linhas de transmissão de energia	*MLP* e aprendizado profundo	Su et. al., 2018
Detecção de diferentes níveis de doença em plantações de eucalipto por meio de imagens hiperespectrais	*RF*	Liao et al., 2022
Identificação e medida de ninhos de formiga cortadeira em plantações de eucalipto	*MLP* e Aprendizado profundo	dos Santos et al., 2022

Tabela 3.2: Exemplos de aplicações de modelos de aprendizado profundo na área florestal

Aplicações	**Modelos**	**Estudos**
Revisão sobre aplicações de Aprendizado Profundo em inventário florestal	*CNN e outros*	Hamedianfar et al., 2022
Revisão sobre aplicações de Aprendizado Profundo em ecologia florestal	*CNN e outros*	Liu et al., 2018
Revisão sobre aplicações de Aprendizado Profundo em ciência florestal	*CNN e outros*	Coelho Eugenio et al., 2021
Revisão sobre aplicações de Aprendizado Profundo no gerenciamento de incêndios florestais	*CNN e outros*	Jain et al., 2020
Estimativa da produção de biomassa de florestas a partir de imagens aéreas LiDAR	Autoencoders	García-Gutiérrez et al., 2016, Zhang et al., 2019
Estimativa de biomassa florestal a partir de imagens de sensoriamento remoto	*CNN*	Talebiesfandarani and Shamsoddini, 2022
Classificação de espécies de árvores	*Deep Boltzmann Machines*	Guan et al., 2015
Estimativa de volume de estoque florestal	*CNN*	Liu et al., 2019
Classificação do dossel	*CNN*	Weinstein et al., 2019
Reconhecimento de pinheiros doentes	*CNN, Deep Convolutional Generative Adversarial Networks (DCGANs)*	Hu et al., 2020
Classificação de árvores coníferas e dedículas	*CNN*	Hamraz et al., 2019

Capítulo 4

Materiais e Métodos

4.1 Caracterização da área

A área experimental do teste de progênies de híbridos de irmãos completos de *Eucalyptus* foi instalada com o objetivo de selecionar genótipos com alta produtividade e potencial de tolerância à seca. O teste foi instalado em março de 2019 no município de Buritizeiro, estado de Minas Gerais (17º 05′ 49"S 44º 53′ 09"O), local que tem clima tropical, classificado como sendo do tipo Aw (inverno seco), altitude de aproximadamente 570 m, com precipitação anual média de 1102 mm, com temperatura média anual de 24,5ºC, com máxima de 41ºC e mínima de 5,4°C. Com solo do tipo argiloso (15-35%) de textura média (Fonte: Cia Ferroligas Minas Gerais MinasLigas).

Clones cultivados em tubetes foram transferidos para vasos de 21 Litros com solo e areia na proporção 1:1. A adubação inicial foi feita com Super Simples, seguida por adubações periódicas com NPK 20-00-20 e micronutrientes (ácido bórico, cloreto de cobre e sulfato de zinco), baseadas na análise de solo. A adubação de NPK ocorreu a cada 30 dias e a de micronutrientes a cada 20 dias. Após o preparo do solo e o transplante, as mudas foram aclimatadas às condições ótimas, atingindo cerca de 100% da capacidade de campo (CC) ao longo de aproximadamente 55 dias.

As coletas de dados iniciaram em setembro de 2019 (6 meses de plantio) e foram realizadas também as 18, 30, 36 e 42 meses de plantio, sempre ao final do ciclo anual de seca, permitindo observar os contrastes entre os cruzamentos sob disponibilidade hídrica limitada. O projeto terá duração de sete anos, em julho de 2023 foi realizada a última coleta de dados, com 42 meses de plantio.

O gráfico da Figura 4.1 mostra a precipitação mensal acumulada e a variação das temperaturas mínimas, médias e máximas ao longo das coletas de dados (Dados: Cia Ferroligas Minas Gerais MinasLigas).

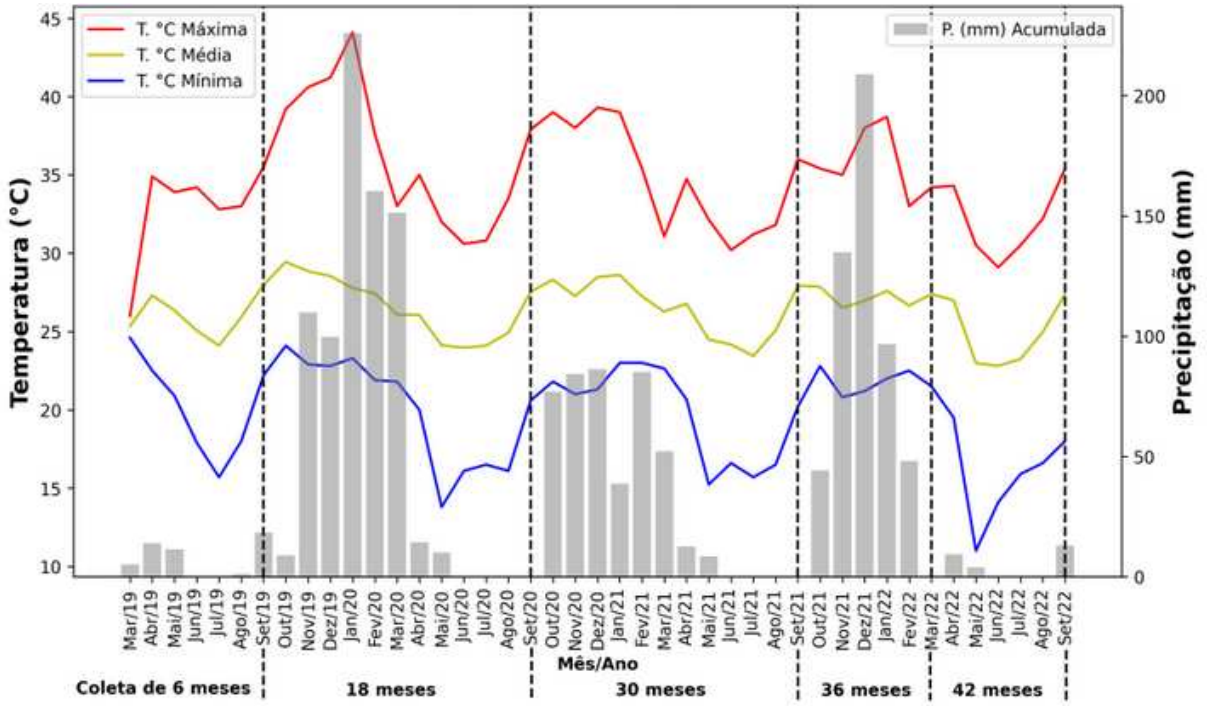

Figura 4.1: Variações da precipitação mensal acumulada e das temperaturas mínima, média e máximas ao longo das coletas de dados.

4.2 Material vegetal

Para as análises foram utilizados 214 cruzamentos (progênies), com genitores não aparentados, de híbridos de *Eucalyptus* gerados por meio da técnica de hibridação - Protoginia Artificialmente Induzida (PAI) (Assis et al., 2005). Para cada cruzamento, foram plantados 20 indivíduos no espaçamento 3,5 × 2,57 m (9 m²), sendo que, cada indivíduo constituiu uma parcela (*Single Tree Plot - STP*). Além disso, foram plantados 6 clones da testemunha (GG2673 [E. urophylla], GG1923 [E. urophylla], VM1 [E. urophylla × E. camaldulensis], AEC1528 [E. grandis × E. urophylla], I144 [E. urophylla HE], GG1980 [E. urophylla]) para comparação dos resultados. Para evitar o efeito de borda, foram plantadas, em fileira dupla, indivíduos do clone I144 (*E. urophylla HE*) nos limites das áreas. Portanto, o experimento é composto por 4400 (quatro mil e quatrocentos) indivíduos.

4.3 Conjunto de dados

Os dados trabalhados para aplicação dos modelos foram divididos em duas categorias: tabulados e imagens. Os dados tabulados são compostos por cinco variáveis fisiológicas, coletadas aos seis meses de plantio: comprimento do limbo foliar (CF), largura do limbo foliar (LF), área foliar (AF), área foliar específica (AFE) e potencial

hídrico foliar (ψf). Além destas, também foi fornecido o incremento médio anual do volume (IMAVol), que, por ser utilizado em um dos critérios de rotulagem (classificação) das amostras, não foi considerado como variável de entrada para os modelos.

Para o cálculo da AFE, aproximadamente 15 folhas foram digitalizadas em scanner (*HP ScanJet 200*) para cálculo da área foliar (cm²), em seguida, estas folhas foram secas em estufa de circulação forçada de ar a 65 ºC até atingir massa seca (g) constante. A partir da razão entre a área da folha e a massa seca foi determinada a AFE (cm². g-1). Das imagens também foram obtidas a AF, CF e LF, através do software *Image-Pro Plus*®. O ψf foi determinado com o auxílio de uma Bomba de Scholander em folhas completamente expandidas (3º ou 4º par de folhas) do terço médio da planta. Os contrastes observados durante a medição ao meio-dia foram avaliados no período da antemanhã (aproximadamente 06:00) para verificar o continuum solo-planta-atmosfera (Silva et al., 2017). O IMAVol foi calculado utilizando o volume da árvore produzido no espaçamento de 3,5 × 2,57 m (9 m²), extrapolado para 1 ha e dividido pela respectiva idade em meses, conforme descrito pela equação (Rodriguez et al., 1997):

$$IMAVol = \frac{(VOL \times 10.000)}{idade} \tag{4.1}$$

VOL é o Volume da árvore em m³ no espaçamento de 9 m². O Volume foi calculado pela equação proposta por (Schumacher e Halll, 1933):

$$VOL = \frac{(\pi \times DAP^2 \times Altura \times f)}{40.000} \tag{4.2}$$

DAP é o diâmetro à altura do peito (cm), *Altura* é a altura da árvore em metros, f é o fator de forma adotado (0,45) e π é a razão entre a circunferência e diâmetro de um círculo (3,14159).

Os dados fisiológicos e as imagens foram coletadas de um total de 326 indivíduos, pertencentes a um total de 55 cruzamentos diferentes, dos quais 52 cruzamentos com 6 indivíduos cada, dois cruzamentos com 5 indivíduos e um cruzamento foi representado por 4 indivíduos. A sobrevivência das plantas, juntamente com o IMAVol, foram medidas para todos os indivíduos e em todas as coletas, por meio do inventário florestal do experimento.

A sobrevivência dos indivíduos foi uma variável utilizada para rotulagem (classificação) das amostras em relação à característica de tolerante ou não tolerante à seca, definida de acordo com a observação da planta em campo, sendo atribuído 1 (um) à amostra viva e 0 (zero) para a que foi considerada morta.

O conjunto de dados de imagem fornecido originalmente é composto por imagens das folhas das plantas amostradas, com uma quantidade não padronizada de folhas para cada indivíduo. As folhas coletadas foram digitalizadas utilizando um scanner *HP Scanjet G2410*, todas juntas, gerando uma imagem com várias folhas para

cada amostra. Após realizado o pré-processamento, conforme descrito na **Seção 4.6.3**, as folhas de cada planta coletada foram desmembradas em uma folha por imagem (**Figura 4.5**), totalizando 4.443 imagens no conjunto de dados.

O objetivo do trabalho de pesquisa foi analisar o desempenho dos modelos para classificar precocemente os **cruzamentos** quanto a tolerância à seca e produtividade. Assim, para utilizar as variáveis fisiológicas como entrada dos modelos considerando os cruzamentos, foi calculada a média por variável para os indivíduos amostrados de cada cruzamento, de modo que a média passou a ser o valor da variável para o cruzamento. Ao final, o conjunto de dados fisiológicos foi comporto por 55 amostras, que são justamente os 55 cruzamentos amostrados aos seis meses.

Em relação às imagens, elas foram utilizadas individualmente como entrada para os modelos, mas rotuladas (classificadas) conforme a tolerância e produtividade dos cruzamentos, conforme discutido na Seção 4.4.

4.4 Rotulagem das amostras

As amostras coletadas aos 6 meses, tanto de dados tabulados (fisiológicos) quanto de imagens, por serem as plantas mais jovens (precoces) foram utilizadas como dados de entrada dos modelos. Elas foram rotuladas com base nas características de tolerância à seca e produtividade observadas nas coletas de dados de 18, 30, 36 e 42 meses, a fim de modelar a predição precoce. Esta forma de rotulagem permitiu a criação de quatro conjuntos de dados diferentes, o que possibilitou avaliar se os modelos são capazes de capturar padrões nos dados que representam as características específicas do desenvolvimento das plantas e das condições ambientais em cada idade de coleta.

O critério para definir um cruzamento como tolerante (**T**) foi que o percentual de mortalidade de seus indivíduos não tenha ultrapassado o limiar de 10%. Caso contrário, o cruzamento é considerado não tolerante (**NT**). Este limiar, embora rigoroso, foi adotado por ser comumente adotado em plantios comerciais, onde uma mortalidade de plantas acima de 10% representa perdas significativas. Em relação à produtividade, foi calculada a média do IMAVol dos indivíduos vivos do cruzamento e ela foi comparada à média geral do IMAVol de todos os indivíduos vivos do experimento, se a média do cruzamento for maior ou igual à geral, o cruzamento é produtivo (**P**), senão, não produtivo (**NP**).

A partir dos conjuntos de dados gerados e das classes criadas, foram experimentadas quatro abordagens para rotular os cruzamentos, considerando os critérios tolerância e produtividade e as classes (T, NT, P e NP):
I) Tolerância e Produtividade - 4 classes: Não Tolerante e Não Produtivo (NT-NP), Tolerante e Não Produtivo (T-NP), Não Tolerante e Produtivo (NT-P) e Tolerante e Produtivo (T-P).

II) Tolerância e Produtividade - 2 classes: Tolerante e Produtivo (T-P), Não Tolerante ou Não Produtivo (NT-NP).
III) Tolerância - 2 classes: Tolerante (T), Não Tolerante (NT).
IV) Produtividade - 2 classes: Produtivo (P), Não Produtivo (NP).

Em I foram consideradas todas as possibilidades de categorizar os cruzamentos quanto às duas características de interesse, portanto, permite um melhor entendimento do comportamento dos materiais no experimento de melhoramento genético. Na abordagem II, a ideia foi isolar os cruzamentos de maior interesse da indústria, tolerantes e produtivos, dos demais. As abordagens III e IV são aplicadas em conjunto, de forma paralela, a fim de separar o problema de classificação multi classe (I) em dois binários, de modo que seja utilizado um modelo para cada critério de classificação e no final seja possível classificar um cruzamento como feito em I, ou seja, nas quatro classes, permitindo uma melhor análise do ponto de vista do melhoramento genético como ocorre na abordagem I.

O intuito de trabalhar com estas 4 estratégias foi tentar construir um modelo com foco em auxiliar os programas de melhoramento florestal (abordagem I isoladamente e a III e IV em conjunto), e outro mais adequado ao emprego comercial (abordagem II) visando isolar a classe de maior interesse industrial. Ademais, também foi uma tentativa de explorar configurações para ajustar um modelo de ML com desempenho satisfatório para o problema.

4.5 Dados de fisiologia

A Tabela 4.1 mostra como ficou a divisão das amostras de cruzamentos para as quatro formas de rotular, em cada idade alvo e para cada classe.

Tabela 4.1: Distribuição das amostras para cada forma de rotulagem, nas idades alvo e por classe.

Total de amostras (cruzamentos) coletadas aos 6 meses: 55					
Rotulagem	**Idade Alvo**	**NT-NP**	**T-NP**	**NT-P**	**T-P**
Tolerância e Produtividade - 4 classes	18	4	17	2	32
	30	24	-	29	2
	36	23	-	30	2
	42	24	-	29	2
	Idade Alvo	**NT-NP**	**T-P**		
Tolerância e Produtividade - 2 classes	18	23	32		
	30	53	2		
	36	53	2		
	42	53	2		
	Idade Alvo	**NT**	**T**		
Tolerância - 2 classes	18	6	49		
	30	53	2		
	36	53	2		
	42	53	2		
	Idade Alvo	**NP**	**P**		
Produtividade - 2 classes	18	21	34		
	30	24	31		
	36	23	32		
	42	24	31		

Rótulos usados para classificar as amostras:
NT-NP: Não **T**olerante e **N**ão **P**rodutiva
T-NP: Tolerante e **N**ão **P**rodutiva
NT-P: Não **T**olerante e **P**rodutiva
T-P: Tolerante e **P**rodutiva

As etapas da aplicação dos algoritmos de aprendizado de máquina aos dados de fisiologia estão resumidas no fluxo de trabalho da Figura 4.2, desde o entendimento do problema, possibilidades de aplicação de ML, tratamento e análise exploratória dos dados, ajuste e avaliação dos modelos.

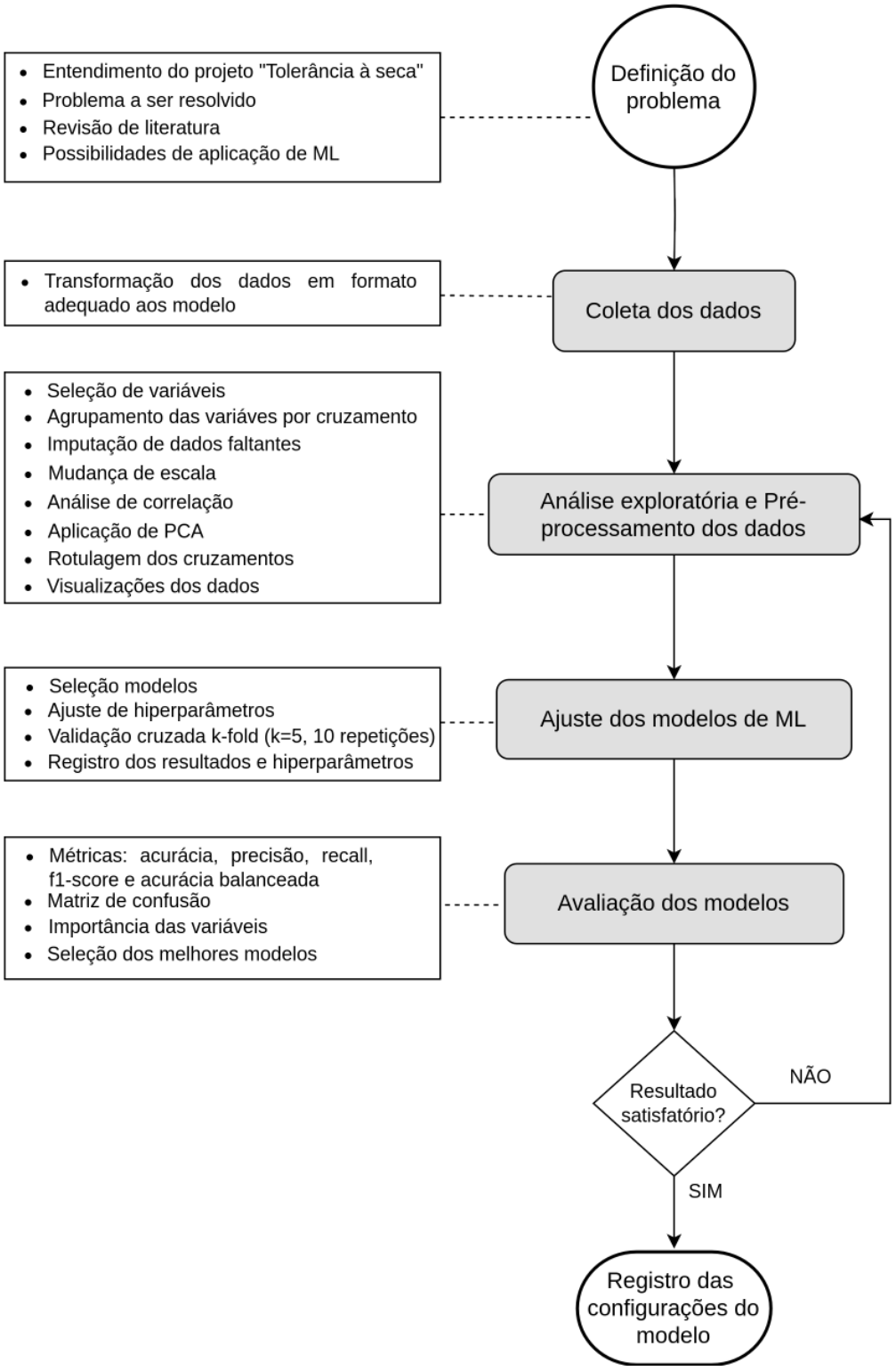

Figura 4.2: Fluxo de trabalho da aplicado aos dados tabulados de fisiologia

4.5.1 Modelos aplicados aos dados fisiológicos

A escolha dos algoritmos de aprendizado de máquina empregados se deu pelo fato de serem amplamente abordados na literatura acerca da aplicação de ML na área

florestal. Os algoritmos de aprendizado de máquina aplicados foram: *Random Forest (RF)*, *Extreme Gradient Boost (XGBoost)*, Redes Neurais Artificiais Multicamadas (*MLP*) e *Support Vector Machine (SVM)*.

Random Forest

O algoritmo *Random Forest* (Breiman, 2001) faz parte de uma categoria de técnicas de aprendizado de máquina chamada ensemble ou conjunto. Os métodos de ensemble combinam múltiplos modelos de aprendizado de máquina para melhorar o desempenho preditivo geral. O RF combina uma grande quantidade de árvores de decisão pouco correlacionadas para realizar tarefas de classificação e regressão, onde cada árvore individual é construída de maneira independente, tendo como dados de treinamento e variáveis de entrada uma amostra aleatória do conjunto de dados original.

A estimativa de erro *Out-of-Bag (OOB)* das árvores de decisão utiliza as amostras que não foram utilizadas para treinamento como um conjunto de validação independente. Isso permite estimar o erro de generalização do modelo e oferece uma alternativa imparcial e robusta de avaliação. A predição final do modelo RF é baseada na predição individual de cada árvore, de modo que na classificação a classe mais frequente prevista pelas árvores é selecionada como a predição final. Na regressão, será a média das previsões das árvores. Além disso, o RF não exige tratamento prévio dos dados de entrada com diferenças significativas de escala e dados faltantes.

Multilayer Perceptron (MLP)

Redes Neurais Artificiais (RNAs) são estruturas inspiradas no cérebro humano, capazes de aproximar funções para lidar com relações lineares e não lineares dos dados. O *Multilayer Perceptron* (*MLP*) é um dos tipos mais conhecidos de RNAs. O *MLP* consiste em várias camadas de neurônios interconectados: a camada de entrada, camadas intermediárias (ou camadas ocultas) e a camada de saída, que produz as previsões desejadas. Durante a propagação para frente (*forward propagation*), os dados de entrada são transmitidos através da rede neural para realizar previsões. Cada neurônio em uma camada recebe os sinais de entrada, realiza um somatório da multiplicação do sinal de entrada pelos pesos sinápticos correspondentes, acrescenta um viés, e aplica uma função de ativação a este resultado, o valor resultante desta operação será a saída do neurônio. Este sinal de saída é transmitido para todos os neurônios da próxima camada da rede e assim sucessivamente (Braga et al., 2000).

Durante a fase de aprendizado, o *MLP* ajusta os pesos das conexões entre os neurônios e o viés para minimizar o erro entre as previsões e os valores conhecidos. Isso é feito usando o algoritmo de retropropagação do erro (*backpropagation*), que calcula os gradientes dos pesos em relação ao erro. Os gradientes são utilizados para realizar

ajustes iterativos nos pesos usando um algoritmo de otimização, como o gradiente descendente estocástico, por exemplo. A estrutura em camadas e a utilização de funções de ativação não lineares permitem que o algoritmo *MLP* modele relações complexas entre as variáveis. A arquitetura da rede pode ser ajustada quanto ao número de camadas ocultas e o número de neurônios em cada camada, permitindo adaptá-la a diferentes problemas até que se obtenha um desempenho adequado ao problema (Taud and Mas, 2018).

Support Vector Machine (SVM)

O objetivo do algoritmo SVM é encontrar um hiperplano no espaço de variáveis de entrada que melhor separa as diferentes classes de dados (classificação) ou estima o valor numérico baseado na posição do ponto em relação ao hiperplano (regressão). O SVM busca encontrar os vetores de suporte, que são os pontos de dados mais próximos do hiperplano de separação. O hiperplano é selecionado de forma que a margem entre as instâncias de treinamento e o hiperplano seja maximizada. Durante o treinamento, o SVM otimiza os coeficientes do hiperplano para maximizar a margem e minimizar erros de classificação. Se os dados não forem linearmente separáveis, eles podem ser transformados em um espaço de maior dimensão usando funções de kernel. As funções de kernel mapeiam os dados para um espaço onde eles se tornam linearmente separáveis (Smola and Schölkopf, 2004).

Extreme Gradient Boost (XGBoost)

O *Extreme Gradient Boosting (XGBoost)* (Chen and Guestrin, 2016) é um algoritmo de aprendizado de máquina que, como o *Random Forest*, pertence à categoria métodos de conjunto (*essemble*), pois utiliza várias árvores de decisão em conjunto para tarefas de classificação e regressão. Ele faz parte da técnica de *Boosting*, que combina modelos "fracos"para criar um modelo agregador forte.

O *XGBoost* calcula o resíduo entre o valor real e a previsão da árvore atual para cada instância de treinamento, construindo novas árvores para capturar os resíduos restantes de forma iterativa e sequencial. Para evitar o sobreajuste (*overfitting*), o algoritmo emprega técnicas de regularização, como limitar a profundidade das árvores, e aplica pesos aos termos de perda. As previsões de todas as árvores são ponderadamente combinadas para obter a previsão final. O ajuste iterativo das árvores no *XGBoost* permite que o modelo se concentre nos padrões de dados não capturados pelas árvores anteriores, melhorando a precisão das previsões à medida que mais árvores são adicionadas. Além disso, elimina a necessidade de pré-processamento dos dados de entrada, como normalização e tratamento de valores faltantes.

4.5.2 Análise exploratória e pré-processamento

A análise exploratória e o pré-processamento dos dados foram conduzidos de forma conjunta, pois são interdependentes. Ambos são essenciais em projetos de aprendizado de máquina, possibilitando a compreensão e a preparação dos dados antes do treinamento dos modelos (Witten et al., 2016). As visualizações de dados são ferramentas vitais na análise exploratória, pois permitem a extração de informações relevantes, identificação visual de padrões e tendências dos dados (Few, 2021). Essas visualizações são apresentadas na Seção 5. O pré-processamento realizado inclui a mudança de escala das variáveis, o tratamento de dados faltantes e a análise de componentes principais (*Principal Component Analysis - PCA*). Todas essas ações têm como objetivo explorar diferentes configurações com os dados para encontrar aquela que resultaria em melhores desempenhos.

Mudança de escala

Na intenção de testar o efeito da mudança de escala dos dados no resultado dos modelos, que segundo (Huang et al., 2020) pode levar a estabilidade e eficiência do treinamento e consequente melhora na capacidade de generalização, foram aplicadas a padronização e a normalização dos dados de entrada. Ambas transformações têm o objetivo de mudar a escala das variáveis de entrada para que tenham intervalos semelhantes entre si.

A normalização das amostras foi calculada pela fórmula:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{4.3}$$

X' é o variável normalizada, X é o valor original, X_{min} é o menor valor, X_{max} é o maior valor. A escala dos valores ficará entre -1 e 1, pois o potencial hídrico tem valores negativos.

A padronização (ou normalização Z-score) redimensiona os valores das variáveis para que elas tenham as propriedades de uma distribuição Gaussiana com média (μ) = 0 e desvio-padrão da média (σ) = 1. A fórmula de cálculo é:

$$X' = \frac{X - \mu}{\sigma} \tag{4.4}$$

Os dados na escala original também foram testados e resultaram em melhor desempenho dos modelos.

Tratamento de dados faltantes

Para todas as variáveis havia dados faltantes na coleta de 6 meses, o percentual de amostras com dados faltantes por variável foi de: AFE: 13%; AFI: 12%; CF: 11%; LF: 11%; ψf: 13%. Para tratar este problema, que pode afetar o negativamente a predição dos modelos (García Laencina et al., 2010), foram adotadas duas estratégias: remoção das amostras de indivíduos com dados faltantes em alguma variável e a imputação da média da variável faltante considerando os indivíduos do mesmo cruzamento e da mesma idade da amostra com dado faltante, a fim de obter um valor médio mais representativo. Os modelos tiveram melhor desempenho com a imputação da média.

Análise de correlação e aplicação de PCA

As variáveis medidas tiveram sua correlação analisadas na Figura 4.3. Desta forma, foi observado que as variáveis Largura do limbo foliar (LF) e Área foliar (AF) apresentaram a maior correlação entre as variáveis testadas. Isto implica que, para as progênies testadas houve maior ganho em área foi em função do aumento da largura em comparação com o comprimento. correlação entre elas, uma vez que uma correlação muito forte pode significar a ocorrência de multicolinearidade, fator que pode afetar as previsões dos modelos (Yakubu, 2010). O mapa de calor da mostra a correlação entre as variáveis.

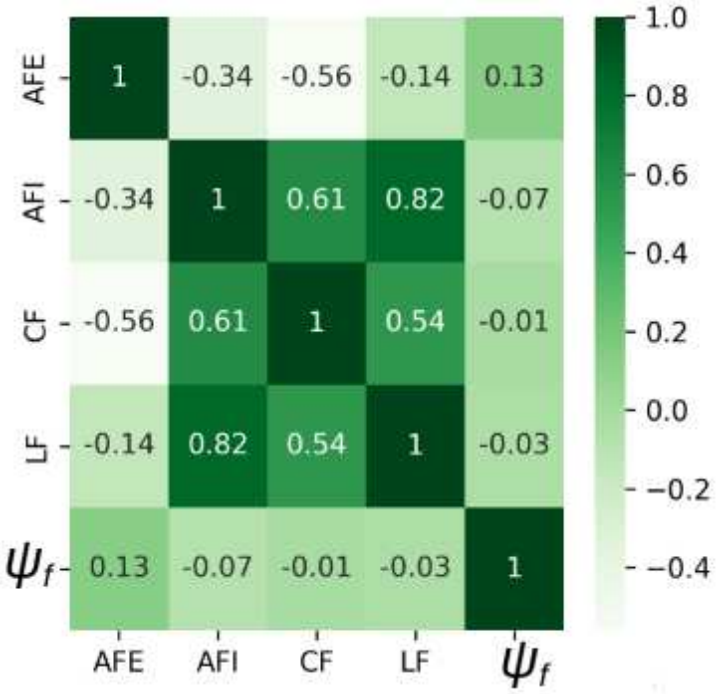

Figura 4.3: Mapa de calor com a correlação de Pearson entre as variáveis fisiológicas.

Há uma correlação maior entre Largura do limbo Foliar (LF) e AF, também entre

Comprimento do limbo Foliar (CF) e AF, o que se justifica pois a AF foi calculada usando estas duas variáveis.

Visando mitigar os efeitos de uma possível ocorrência de multicolinearidade, foi aplicada a redução de dimensionalidade dos dados por meio da técnica de Análise de Componentes principais (PCA). A redução de dimensionalidade implica em reduzir o número de variáveis independentes incluídas no modelo, transformando um conjunto de variáveis correlacionadas em um novo conjunto de variáveis não correlacionadas, chamadas de componentes principais. Estas são combinações lineares das variáveis originais de modo a representar duas ou mais variáveis que representam a mesma informação em um única variável (Jolliffe and Cadima, 2016).

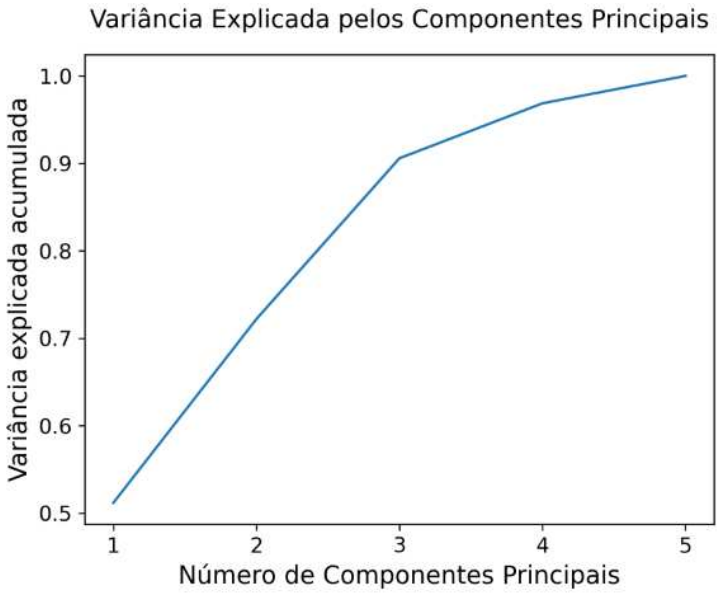

Figura 4.4: Variância explicada acumulada para diferentes números de componentes principais.

O gráfico da figura 3 mostra a explicação da variância dos dados acumulada de acordo com o número de componentes principais usados (Nadif et al., 2021). Pelo Gráfico 4.4, com três componentes principais já se explica mais de 90% da variância dos dados e quatro explicam mais de 95%. Os modelos de ML foram testados, seguindo a mesma metodologia aplicada às variáveis originais do conjunto de dados, com três e quatro componentes principais como preditores, no entanto, os resultados não superaram o uso das variáveis originais.

4.5.3 Ajuste dos modelos

Com o objetivo de ajustar um modelo com bom desempenho para o problema de pesquisa, além de explorar diferentes transformações nos dados, foram experimentadas configurações variadas para os modelos, tarefa conhecida como ajuste de hi-

perparâmetros. O ajuste de hiperparâmetros dos modelos de ML se deu pela busca de hiperparâmetros em grade, onde uma lista pré-definida (empiricamente) de hiperparâmetros e um espaço de busca de valores para os mesmos foi definida. A busca da melhor configuração dos hiperparâmetros foi feita por meio da função *GridSearchCV*, da biblioteca *Scikit-learn* (Pedregosa et al., 2011) da linguagem Python (Python Software Foundation, 2019). A tabela 4.2 mostra os hiperparâmetros testados e o espaço de busca definido para os modelos.

Tabela 4.2: Detalhes da pesquisa em grade de hiperparâmetros e espaço de busca relacionado a cada modelo aplicado.

Modelo	**Hiperparâmetro**	**Espaço de busca**
Random Forest	*n_estimators*	100, 500, 1000
	max_features	auto, sqrt
	max_depth	10, 15, 20
MLP	*hidden_layer_sizes*	(128, 32), (128,64), (32, 64), (32, 128), (128, 32)
	activation	*relu, logistic*
	solver	*adam*
	max_iter	100, 150
	learning_rate	*invscaling, adaptive*
	learning_rate_init	0.001, 0.0001, 0.00001
SVM	*C*	1, 10, 100
	Kernel	rbf
	Gamma	*scale, auto*
XGBoost	*n_estimators*	100, 500, 1000
	gamma	5, 30, 50
	max_depth	10, 15, 20

Tanto na busca de hiperparâmetros quanto no treinamento e avaliação dos modelos foi aplicada a técnica de validação cruzada *k-fold* (*k-fold cross-validation*). Essa abordagem faz a divisão aleatória do conjunto de dados em k partições de tamanho aproximadamente igual. A cada iteração, uma partição é considerada um conjunto de validação e as demais são o conjunto de treinamento (k menos uma partição) (Berrar, 2018). O resultado final da validação cruzada é média das acurácias de validação para cada partição (*fold*). Durante o treinamento dos modelos foram realizadas 10 iterações de validação cruzada, sendo o resultado do modelo a média da acurácia da validação cruzada das 10 iterações. Durante o treinamento, todos os resultados e configurações dos modelos são registrados.

A codificação foi desenvolvida na linguagem Python, versão 3.8, utilizando os modelos clássicos disponíveis na biblioteca Scikit-learn. A execução dos códigos foi

realizada em uma máquina, gentilmente cedida pelo professor David Menotti Gomes, do Laboratório de Visão, Robótica e Imagens da Universidade Federal do Paraná (UFPR). Este servidor está equipado com um processador AMD Ryzen 9 5900X 12-Core Processor, operando a 2.2 GHz, 64GB de memória principal, 35 TB de memória secundária e 2 GPUs NVIDIA TITAN V, cada uma com 12 GB de RAM.

4.5.4 Avaliação dos modelos

A melhor configuração de cada algoritmo foi avaliada pelas médias das métricas de Acurácia, Acurácia Balanceada, Precisão, Revocação e F1-score, após 10 iterações de validação cruzada . A avaliação da acurácia geral pode, em um cenário de conjunto de dados desbalanceado, proporcionar uma impressão enganosa do desempenho da generalização, uma vez que um classificador pode tendenciar a fazer a maioria das predições para a classe majoritária. Nesse contexto, a acurácia geral pode induzir a conclusões equivocadas sobre a performance do algoritmo, conforme destacado por (Brodersen et al., 2010). Em função disso, buscando uma métrica que permitisse uma avaliação mais adequada dos modelos, a acurácia balanceada foi escolhida por considerar o desequilíbrio entre as classes ao calcular a acurácia dos modelos (Brodersen et al., 2010). A métrica F1-score faz uma média harmônica entre Precisão e Revocação (*recall*) e também considera o desbalanceamento entre as classes, sendo adequada para o cenário estudado (Castro and Braga, 2011a).

$$Acurácia = \frac{TN + VN}{VP + FN + VN + FP} \tag{4.5}$$

$$Acurácia \;\; Balanceada = \frac{1}{2}\left(\frac{VP}{P} + \frac{VN}{N}\right) \tag{4.6}$$

$$Precisão = \frac{VP}{VP + FP} \tag{4.7}$$

$$Revocação = \frac{VP}{VP + FN} \tag{4.8}$$

$$F1 - score = 2 * \frac{Precisão * Revocação}{Precisão + Revocação} \tag{4.9}$$

As métricas acima são calculadas com base nos erros - Falso Positivo (FP) e Falso Negativo (FN) - e acertos - Verdadeiro Positivo (VP) e Verdadeiro Negativo (VN) - de cada classe. Estes resultados das predições do modelo podem ser descritos graficamente por meio de uma matriz de confusão ou tabela de contingência (Tabela 4.3) (Castro and Braga, 2011a). Nela, cada elemento $m_{k,j}$ dessa matriz indica a quantidade de exemplos cuja classe verdadeira era c_k e que foi classificada como c_j. Dessa forma, os elementos ao longo da diagonal principal representam as decisões corretas, isto

é, o número de verdadeiros negativos (VN) e verdadeiros positivos (VP), enquanto os elementos fora dessa diagonal indicam os erros cometidos, ou seja, o número de falsos positivos (FP) e falsos negativos (FN).

Tabela 4.3: Matriz de confusão para um modelo de classificação binário.

	predição (y=0)	predição (y=1)
real (y=0)	VN	FP
real (y=1)	FN	VP

As células da diagonal principal da matriz são as que concentram os resultados de interesse, ou seja, os acertos do modelo, seja as amostras da classe positiva corretamente classificadas como as positivas, sejam as da classe negativa acertadamente classificadas como negativas.

4.6 Dados de imagem

A Tabela 4.4 mostra como ficou a divisão das amostras de cruzamentos para as quatro formas de rotular, em cada idade alvo e para cada classe.

Tabela 4.4: Número de imagens coletadas aos 6 meses separadas nos conjuntos de treinamento, com (90%) dos dados para validação cruzada, e teste, (10%) dos dados para validação final, agrupadas pela forma de rotulagem, idade alvo e classe.

Rotulagem	**Idade Alvo**	**NT-NP**[0]	**T-NP**[1]	**NT-P**[2]	**T-P**[3]
		Tr-Te[4]	**Tr-Te**	**Tr-Te**	**Tr-Te**
T e P [5] - 4 classes	18	252-27	1.259-139	150-16	2.340-260
	30	1.746-193	-	2.106-234	148-16
	36	1.665-185	-	2.187-242	148-16
	42	1.745-193	-	2.107-234	148-16
		NT-NP	**T-P**		
T e P - 2 classes	18	1659-184	2.340-260		
	30	3.852-427	148-16		
	36	3.852-427	148-16		
	42	3.852-427	148-16		
		NT	**T**		
T - 2 classes	18	401-44	3.599-399		
	30	3.852-427	148-16		
	36	3.852-427	148-16		
	42	3.852-427	148-16		
		NP	**P**		
P - 2 classes	18	1.510-167	2.490-276		
	30	1.746-193	2.254-250		
	36	1.665-185	2.334-259		
	42	1.745-193	2.255-250		

[0] Classe **N**ão **T**olerante e **N**ão **P**rodutiva
[1] Classe **T**olerante e **N**ão **P**rodutiva
[2] Classe **N**ão **T**olerante e **P**rodutiva
[3] Classe **T**olerante e **P**rodutiva
[4] Nº de imagens nos conjuntos de **Tr**einamento (Tr) e **Te**ste (Te)
[5] **T**olerância e **P**rodutividade

A Figura 4.5 mostra alguns exemplos de imagens, já devidamente pré-processadas, pertencentes a cada uma das classes definidas (NT-NP, T-NP, NT-P e T-P). Pelas imagens nota-se que prever o comportamento dos cruzamentos quanto à tolerância à seca e produtividade a partir das imagens das folhas das plantas é uma tarefa desafiadora, pois não há uma diferença significativa entre as imagens que permita uma identificação visual da classe da amostra a partir da imagem. Uma característica que pode ser detectada visualmente é a área foliar. Esta característica morfofisiológica do eucalipto

é amplamente mencionada na literatura como um bioindicador de tolerância à seca, pois, sob condição de limitada disponibilidade hídrica, a alteração da área foliar é uma primeira linha de defesa das plantas como resposta adaptativa (Pita and Pardos, 2001, Maseda and Fernández, 2016, Chaves et al., 2004, Tatagiba et al., 2007, Li et al., 2020). Além disso, de acordo com (Larcher, 2006), a resposta inicial à deficiência hídrica nas plantas inclui a diminuição da turgescência, o fechamento dos estômatos, a redução na fotossíntese e a diminuição do alongamento celular, o que impacta diretamente o crescimento.

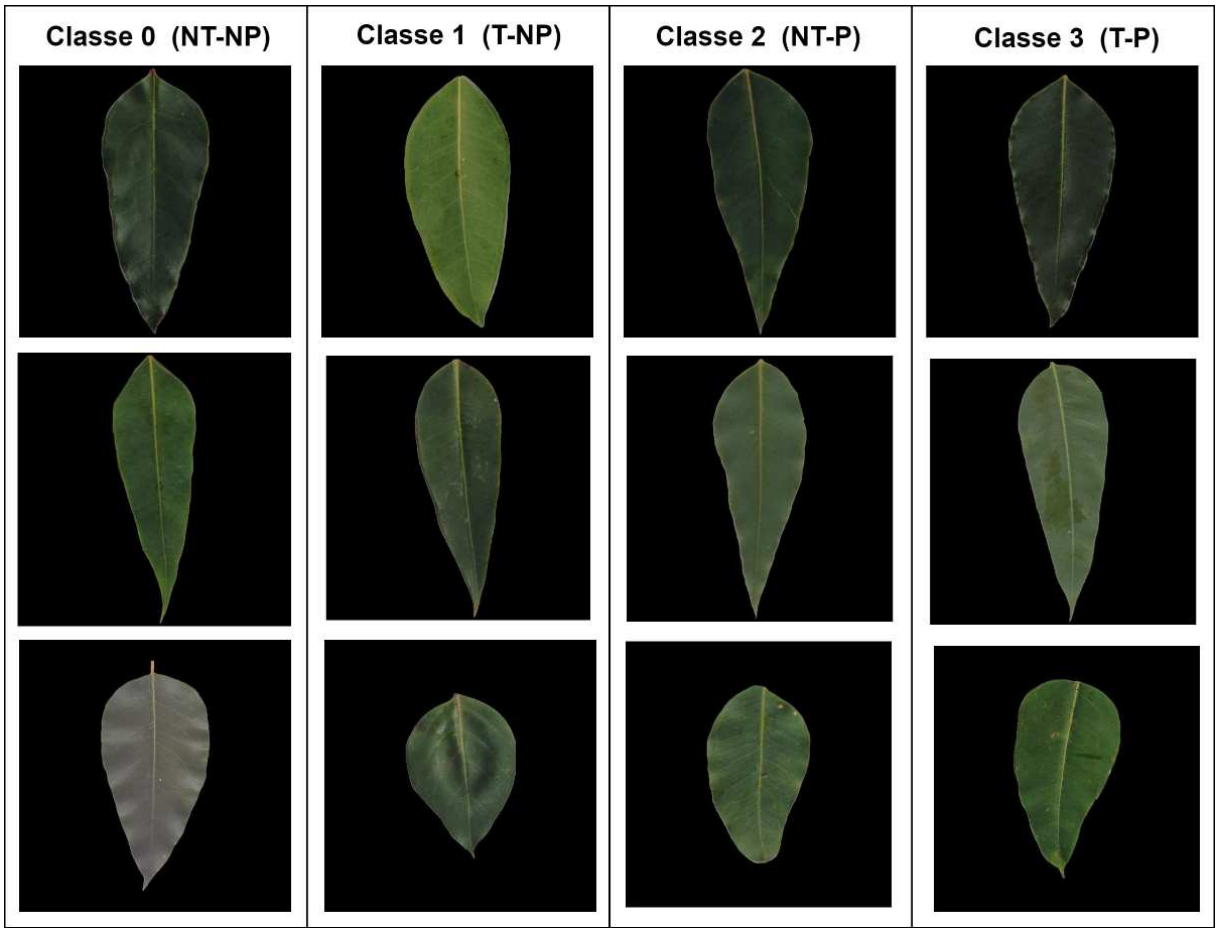

Figura 4.5: Diferentes imagens pertencentes a cada uma das 4 classes, rotuladas pelos critérios de Tolerância e Produtividade: 0 - Não Tolerante e Não Produtiva; 1 - Tolerante e Não Produtiva; 2 - Não Tolerante e Produtiva; 3 - Tolerante e Produtiva;

4.6.1 Modelos aplicados às imagens

As arquiteturas de CNN aplicadas foram a *MobileNetV2* (Sandler et al., 2018), *Xception* (Chollet, 2017) e *Resnet50* (He et al., 2015), disponíveis na biblioteca *Keras* (Chollet et al., 2015) da linguagem *Python* (Python Software Foundation, 2019). A escolha das arquiteturas se deu pelo fato de serem amplamente empregadas em tarefas de visão computacional e por representarem o estado da arte neste tipo de aplicação.

ResNet50

A ResNet-50 (He et al., 2015), uma variante das Redes Residuais (ResNet), foi desenvolvida em 2015 com o intuito de resolver os desafios enfrentados ao criar redes mais profundas. Um dos problemas principais era o "desaparecimento do gradiente"("*vanishing gradient*"), que ocorre ao criar redes mais profundas, com a adição de muitas camadas, resultando em gradientes muito pequenos que impactam o treinamento, tornando-o impreciso e dificultando a atualização dos pesos das camadas iniciais na retropropagação (He et al., 2015).

Para contornar esse problema, introduziu-se uma inovação: a conexão residual. Essa conexão permite que os gradientes fluam diretamente através das camadas da rede, evitando a degradação das derivadas. Em vez de aprender diretamente a função de mapeamento da entrada para a saída, as conexões residuais aprendem a diferença entre a entrada e a saída e, em seguida, adicionam essa diferença à entrada original (He et al., 2015).

A ResNet-50 possui um total de 50 camadas. Inicia com uma camada de convolução de 64 filtros usando um *kernel* de 7x7, seguida por uma camada de *pooling*. São então aplicados quatro "estágios"de convolução, onde são empilhadas três camadas de convolução (conv. 1x1, conv. 3x3, conv. 1x1), repetidas várias vezes. Após cada repetição, é incluído um bloco residual. Ao final dos estágios de convolução, há uma camada de *pooling* seguida por uma camada de classificação totalmente conectada, com ativação *softmax* para a probabilidade das classes de saída.

A arquitetura completa é representada na Figura 4.6. Para introduzir não linearidade, é aplicada a ativação *ReLU* após as camadas convolucionais e residuais. Além disso, a normalização em lote é utilizada após cada camada para estabilizar e acelerar o treinamento da rede, normalizando as ativações antes da aplicação da função de ativação.

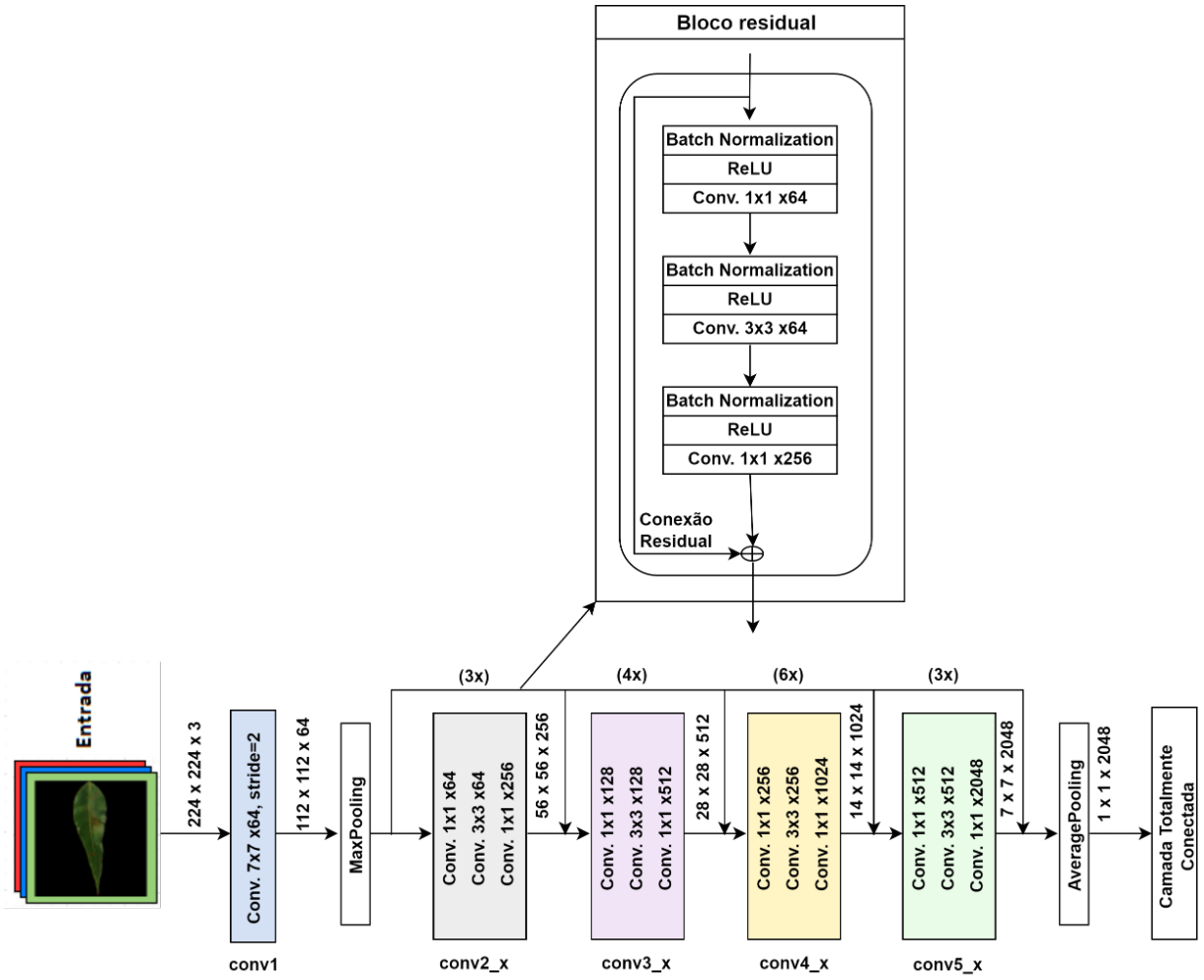

Figura 4.6: Arquitetura da *ResNet50*. Os blocos Convolucionais são nomeados como "conv" seguido do número do bloco, onde há uma repetição de blocos convolucionais, segue-se "_x" ao nome do bloco. A redução da resolução é realizada por conv3_1, conv4_1 e conv5_1 com um *stride*=2. Os blocos residuais são empilhados após a repetição dos blocos convolucionais.

Xception

A arquitetura da rede neural convolucional (CNN) chamada *Xception* (Chollet, 2017) é uma inovação proposta por François Chollet, criador da biblioteca *Keras*, que se concentra em melhorar a eficiência e a capacidade de aprendizado das CNNs. O nome "*Xception*"é uma contração de "*Extreme Inception*", uma vez que a rede foi baseada em uma arquitetura de rede anterior chamada *Inception* (Szegedy et al., 2015).

As Convoluções Separáveis em Profundidade (*Depthwise Separable Convolutions*) são a principal inovação da *Xception*, são duas camadas de convolução distintas:
a. Convolução em profundidade (*Depthwise Convolution*): Nessa etapa, cada canal de entrada é convoluído individualmente, reduzindo a quantidade de parâmetros e, consequentemente, o custo computacional.

b. Convolução em ponto (*Pointwise Convolution*): Após a convolução em profundidade, uma camada de convolução em ponto (convolução 1x1) é aplicada para combinar os canais resultantes.
A *Xception* utiliza módulos residuais semelhantes aos encontrados na arquitetura *ResNet50* e *MobileNet*. A *Xception* usa filtros menores em comparação com outras arquiteturas, o que permite capturar detalhes finos e aumenta a eficiência computacional.

Em vez de usar blocos convolucionais separados em camadas diferentes, a rede usa camadas residuais em cascata, o que ajuda a reduzir a necessidade de parâmetros e acelera o treinamento. As camadas de ativação *ReLU* são aplicadas após cada camada de convolução, bem como a normalização em lote. A rede termina com uma camada de *pooling* global médio que produz uma única representação vetorial da imagem de entrada, permitindo a classificação da imagem pela rede totalmente conectada. A Figura 4.7 é um diagrama da arquitetura da *Xception*.

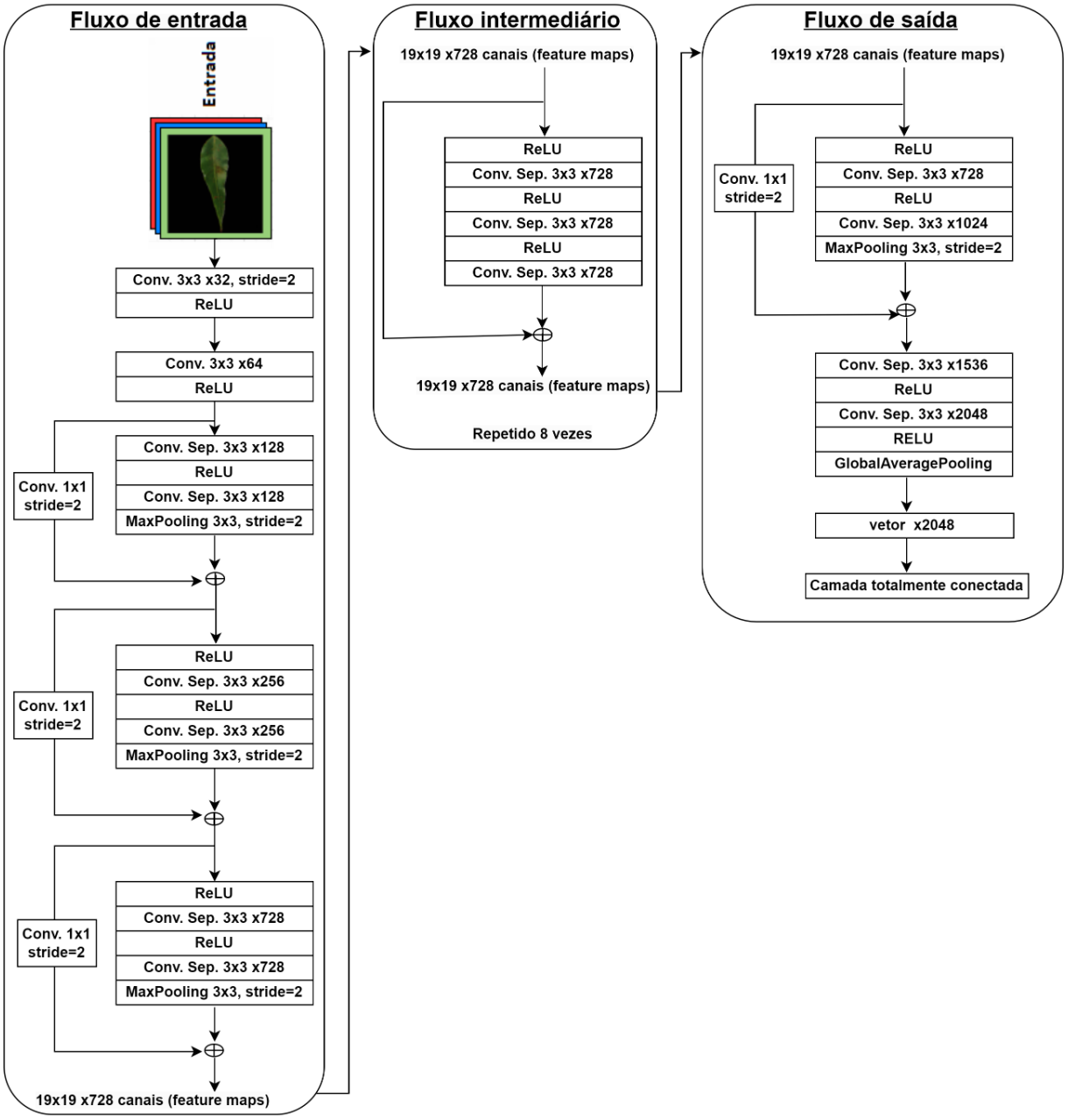

Figura 4.7: Arquitetura da *Xception*. Os dados passam primeiro pelo fluxo de entrada, depois pelo fluxo intermediário, que é repetido oito vezes, e por fim pelo fluxo de saída. Em todas as camadas de Convolução (Conv.) e Convolução Separável (Conv. Sep.) são seguidas pela normalização em lote (não inclusa na imagem). Todas as camadas de Convolução Separável utilizam um multiplicador de profundidade de 1 (sem expansão de profundidade).

MobileNetV2

Segundo (Sandler et al., 2018), a MobileNetV2 foi projetada para oferecer eficiência e desempenho em tarefas de visão computacional em dispositivos móveis. Para alcan-

çar este objetivo, ela incorpora uma técnica chamada blocos residuais de estrutura invertida com gargalos lineares (*inverted residual blocks with linear bottlenecks*). A rede inclui várias camadas destes blocos residuais invertidos, cuja estrutura interna inclui:

- Convolução pontual (*pointwise convolution*): é a primeira camada do bloco, uma convolução 1x1 com 24 canais, seguida de normalização em lote e aplicação da função *ReLU*6 como não linearidade devido à sua robustez quando usada em computação de baixa precisão.

- Convolução em profundidade (*depthwise convolution*): Esta camada aplica convolução com 144 filtros 3x3, expandido o volume da entrada e ao mesmo tempo reduzindo a dimensionalidade dos dados enquanto extrai as características.

- Gargalo linear (*linear bottleneck*): é a última camada do bloco. Ela executa uma transformação linear aplicando convoluções 1x1 com 24 filtros, reduzindo a quantidade de canais de saída. Isso resulta em uma diminuição na carga computacional, uma vez que há menos parâmetros a serem ajustados. Importante notar que essa camada não aplica uma função de ativação não linear, como a *ReLU*, apenas uma normalização em lote após a convolução, permitindo uma melhor preservação das informações úteis extraídas e do gradiente durante a propagação para trás (*backpropagation*) no treinamento da rede.

- Conexão residual: conecta a entrada do bloco à sua saída, de modo a evitar a perda de informação da entrada e ajuda a mitigar o problema do desvanecimento do gradiente de erro durante o treinamento. Isso cria a conexão de atalho que é característica dos blocos residuais.

Conforme o exposto, os blocos residuais invertidos combinam várias camadas, incluindo camadas lineares e de convolução em profundidade, que trabalham juntas para extrair características de maneira eficaz e reduzir o custo computacional. A Figura 4.8 é uma representação visual da arquitetura da *MobileNetV2*.

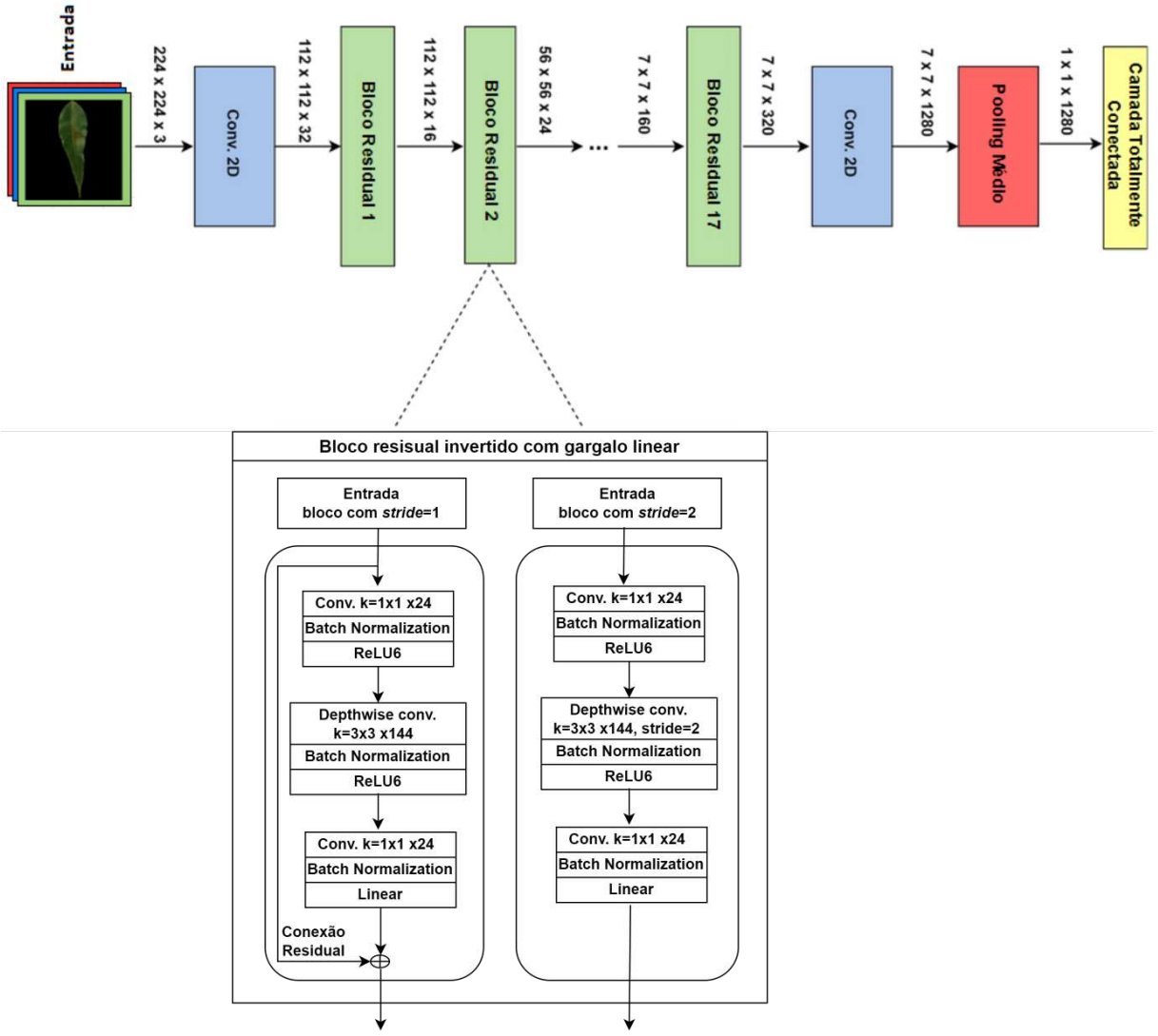

Figura 4.8: Representação da arquitetura da MobileNetV2

A tabela 4.5 descreve com mais detalhes a arquitetura da *MobilenetV2*. Cada linha descreve a sequência de camadas idênticas, repetidas n vezes. Todas as camadas na mesma sequência têm o mesmo número c de canais de saída. A primeira camada de cada sequência tem um *stride* s=2 e todas as outras usam s=1. Todas as convoluções fora dos blocos residuais usam filtro de tamanho 3×3. O fator de expansão t é aplicado às camadas internas do bloco residual a fim de expandir em número de canais na entrada do bloco.

Tabela 4.5: Detalhes da arquitetura da *MobileNetV2*

Entrada	**Operador**	***t***	***c***	***n***	***s***
224^2 x 3	Convolução	-	32	1	2
112^2 x 32	Bloco residual (*bottleneck*)	1	16	1	1
112^2 x 16	Bloco residual (*bottleneck*)	6	24	2	2
56^2 x 24	Bloco residual (*bottleneck*)	6	32	3	2
28^2 x 32	Bloco residual (*bottleneck*)	6	64	4	2
14^2 x 64	Bloco residual (*bottleneck*)	6	96	3	1
14^2 x 96	Bloco residual (*bottleneck*)	6	160	3	2
7^2 x 160	Bloco residual (*bottleneck*)	6	320	1	1
7^2 x 320	Convolução 1x1	-	1280	1	1
7^2 x 1280	*Pooling* Médio 7x7	-	-	1	-
1 x 1 x 1280	Convolução 1x1	-	k	-	-

O termo "invertido" se refere ao fato de as conexões residuais acontecerem entre blocos com pequeno número de canais (blocos "finos"), de maneira inversa aos blocos residuais tradicionais, nos quais a conexão acontece entre blocos com relativamente elevado número de canais (blocos "grossos"). No bloco residual da *MobilenetV2* a conexão acontece entre blocos com relativamente poucos canais, como o da Figura 4.8 que tem 24 canais.

4.6.2 Aplicação dos modelos

A Figura 4.9 representa o processo de aplicação dos algoritmos de aprendizado profundo aos dados de imagem.

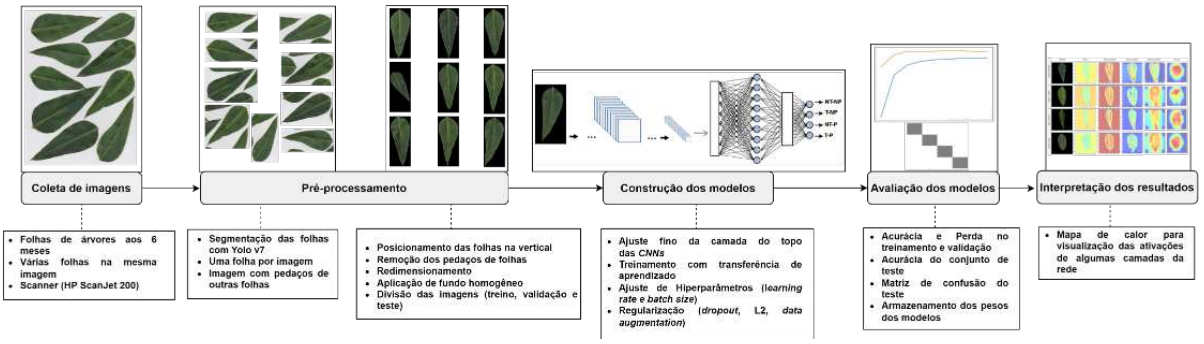

Figura 4.9: Processo de aplicação das CNNs às imagens das folhas

4.6.3 Pré-processamento das imagens

O processo se inicia com a digitalização das imagens contendo várias folhas coletadas em campo. Essas imagens brutas são então submetidas a um pré-processamento que envolve:
i) segmentação de folhas: as imagens são segmentadas para isolar individualmente cada folha, garantindo que apenas uma folha seja presente em cada imagem resultante.
ii) padronização de resolução e posicionamento: as imagens são redimensionadas para uma resolução padronizada e seu posicionamento é ajustado, garantindo uniformidade nos dados de entrada para os modelos.
iii) manipulação do fundo: as variações no fundo das imagens são tratadas para eliminar distrações e garantir foco nas folhas, assim, todas as imagens ficaram com fundo preto.

As imagens originais continham várias folhas de cada planta coletada, além de apresentarem objetos que não eram folhas, como pode ser visto na Figura 4.10 a presença de pecíolos junto das folhas. Com o objetivo de obter imagens com folhas individuais, de modo a diminuir o ruído das imagens e aumentar o conjunto de dados, foram realizadas algumas etapas de pré-processamento das imagens, como pode ser observado no fluxo da Figura 4.10.

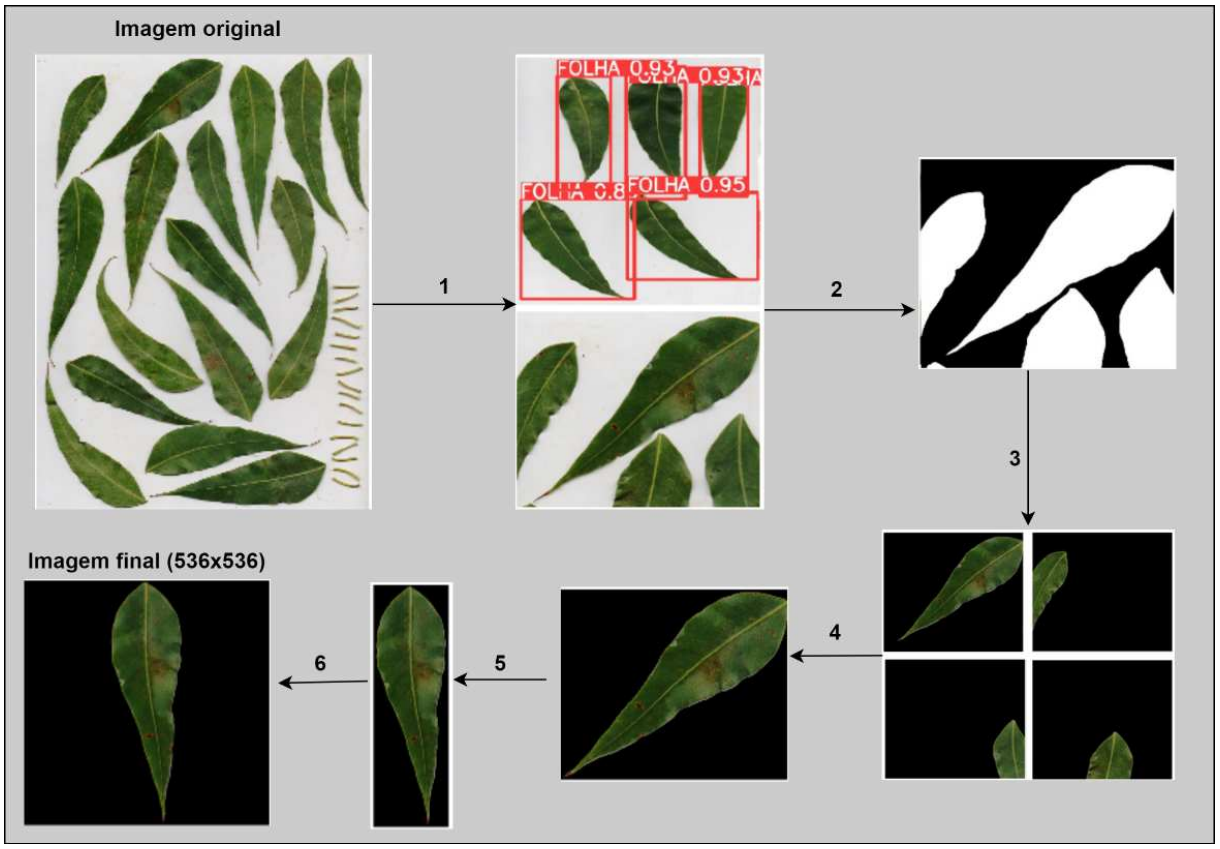

Figura 4.10: Etapas do pré-processamento das imagens

A seguir são detalhadas as etapas do pré-processamento:

1. Detecção e segmentação das folhas com o modelo de CNN *Yolo V5* (Redmon et al., 2016), o ajuste do *Yolo* teve acurácia de 93% para detecção das folhas individuais na imagem.

2. Após a segmentação, algumas imagens das folhas individuais ainda tinham pedaços de folhas, então foi gerada uma máscara para uma segunda segmentação de modo a diferenciar o que era folha do fundo da imagem.

3. Após a geração da máscara, as folhas foram separadas por meio da criação de uma matriz na mesma resolução da máscara. Cada ponto da massa da máscara é verificado até que um pixel branco seja encontrado, indicando a presença de uma possível folha. Em seguida, todos os pontos brancos próximos ao pixel identificado são mapeados para essa matriz. Ao concluir o mapeamento, os pixels correspondentes na imagem original, marcados na matriz, são extraídos e combinados em uma nova imagem com fundo preto, resultando na preservação apenas do objeto identificado na imagem. Após o mapeamento de um objeto, a verificação dos pixels continua do ponto onde o primeiro foi encontrado até que

outro pixel branco ainda não mapeado na matriz seja localizado. Esse procedimento se repete até que todos os pontos brancos da imagem sejam verificados. O resultado final é uma imagem de fundo preto para cada folha identificada.

4. O próximo passo envolve a seleção do objeto com o maior número de pixels, que representa a maior folha identificada.

5. Essa folha é verticalizada, reduzindo sua largura gradualmente por meio de rotações de 1 grau e removendo o excesso de fundo até que a largura mínima seja alcançada.

6. Após a minimização da largura, a imagem é redimensionada para uma largura de 536 pixels, igualando-a em largura e altura para obter um formato quadrado. Posteriormente, a folha é centralizada em uma imagem de fundo preto de 536 x 536 pixels. Este procedimento visa assegurar que a maior folha seja apresentada de forma verticalizada, mantendo um tamanho uniforme e quadrada

Antes das imagens serem utilizadas pelos modelos, os valores dos pixels das imagens foram normalizados (divididos por 255) para ficarem no intervalo entre 0 e 1. Essa normalização é frequentemente aplicada para ajudar no treinamento da rede, tornando o processo de aprendizado mais estável e rápido.

Com as imagens preparadas, o conjunto de dados foi dividido em duas partes: treinamento (90%) e teste (10%). O conjunto de treinamento foi utilizado para validação cruzada *k-fold* (k=5) e os outros 10% para um teste final de generalização dos modelos.

4.6.4 Aumento de dados

Uma técnica amplamente adotada nos projetos de aprendizado profundo para tentar suprir a demanda adequada de dados para o treinamento é o aumento de dados (*data augmentation*). Esta técnica se baseia em aplicar transformações nos dados originais de modo a criar amostras sintéticas a partir deles, permitindo o aumento na quantidade de amostras para o treinamento dos modelos. O aumento de dados no processo de treinamento é usado para garantir que a rede possa extrair as informações estruturais das amostras e conferir ao modelo a capacidade de generalização, ou seja, manter a acurácia de predição do treinamento nas previsões com dados não vistos anteriormente, evitando assim o sobreajuste (*overfitting*), sendo, assim, um técnica de regularização (Zhang et al., 2019). Em geral, à medida que a quantidade, a qualidade e a diversidade dos dados no conjunto de dados aumentam, a eficácia do modelo pode ser melhorada (Zhang et al., 2019).

O aumento de dados foi feito apenas com as imagens de treinamento, a fim de aumentar as imagens das classes minoritárias para balanceamento das classes . Segundo (Lones, 2021), o aumento de dados deve ser feito apenas no conjunto de treinamento para evitar o "vazamento" de dados do treinamento para a validação ou testes, prejudicando a generalização do modelo. O aumento de dados foi realizado durante a validação cruzada.

A Figura 4.11 mostra um exemplo das imagens geradas a partir das transformações realizadas em uma imagem original (a) do conjunto de treinamento. A partir de uma imagem, depois de aplicadas as transformações, foram geradas outras sete.

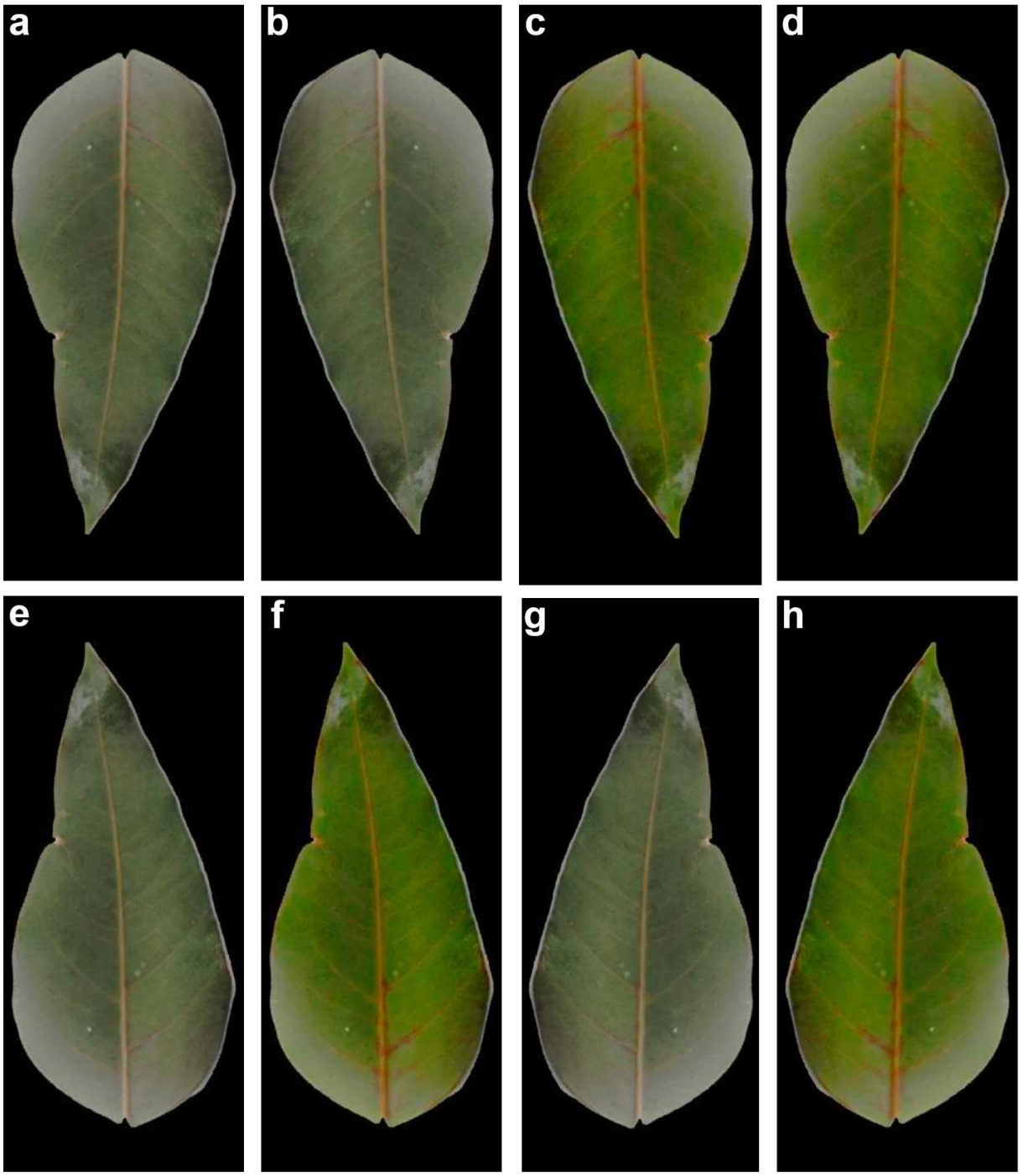

Figura 4.11: Resultado do aumento de dados: a) imagem original; b) inversão horizontal de a; c) imagem b com saturação (2x); d) imagem a com saturação; e) inversão vertical de a; f) imagem e com saturação; g) inversão horizontal de e; h) imagem g com saturação.

4.6.5 Ajuste dos modelos

Os modelos de CNN foram criados usando arquiteturas pré-existentes na biblioteca *Keras* (Chollet et al., 2015). Foi aplicada a técnica de transferência de aprendizado do

conjunto de dados *ImageNet* (ImageNet, 2016), ajustando todos os pesos da rede para o novo conjunto de dados. Durante esse processo, nenhum peso da rede foi mantido fixo em relação ao aprendizado anterior com o *ImageNet*. As CNNs foram treinadas tanto com transferência de aprendizado quanto sem, e os melhores resultados foram obtidos com a técnica de transferência, apresentados na Seção "Resultados e Discussão".

Para ajustar os modelos, foram testadas diferentes configurações de hiperparâmetros, incluindo a taxa de aprendizado (learning rate), tamanho do lote (batch size), a quantidade de camadas escondidas e neurônios na camada totalmente conectada, além de testar diferentes valores para as regularizações dropout e L2. Esses ajustes foram realizados da seguinte maneira:

- A taxa de aprendizagem (*learning rate*) foi inicializada em 1.6970e-04 e reduzida experimentalmente em 10% e 20% a cada três épocas de treinamento sem melhora na acurácia. A redução em 10% demonstrou o melhor desempenho.
- O tamanho do lote (*batch size*) foi testado com os valores 32 e 64, sendo que o valor 64 produziu resultados melhores.
- Na camada totalmente conectada, foram testadas diferentes configurações de camadas ocultas: duas camadas, com 64 e 128 neurônios; 128 e 256 neurônios; 256 e 512 neurônios. Uma camada com 64; 128; 256; 512; A configuração com **uma camada de 256 neurônios** demonstrou o melhor desempenho.
- Para a regularização, os valores de *dropout* testados foram 0.1, 0.25 e 0.5, e de *L2* de 0.001 e 0.01. Os resultados mais satisfatórios foram obtidos com os valores de 0.25 para *dropout* e 0.01 para *L2*.

Além disso, após a última camada da rede convolucional, foi adicionada uma camada de *Pooling* global médio (*GlobalAveragePooling*), seguido por uma camada de classificação com uma camada oculta de 256 neurônios, função de ativação *ReLU* e regularização *L2* (0.01). Em seguida, incluímos uma camada de *dropout* de 25%. A camada final de classificação conta com quatro neurônios e função de ativação *Softmax*, conforme ilustrado na Figura 4.12 com um exemplo da *MobileNetV2*. O treinamento foi conduzido por 30 épocas para todas as CNNs, mantendo a mesma configuração descrita anteriormente em todas elas.

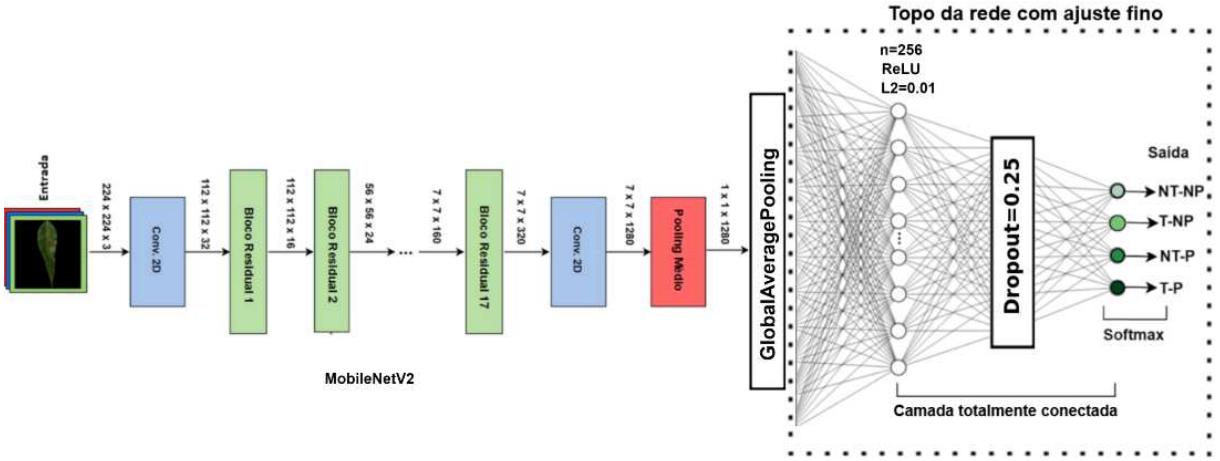

Figura 4.12: Representação da CNN com ajuste aplicado a todas as CNNs, com a *MobileNEetV2* de exemplo.

Os experimentos realizados com os modelos de aprendizado profundo foram realizados utilizando o ambiente de computação em nuvem fornecido pelo **Google Colaboratory (Colab Pro+)**. Os códigos foram executados em uma instância virtual do *Colab*, equipada com um processador Intel(R) Xeon(R) 2.00GHz 38MB de cache L3, 12GB de RAM, SSD 64 GB de armazenamento e uma GPU NVIDIA A100 com 40 GB de RAM.

4.6.6 Avaliação dos modelos

Por fim, após o treinamento, os modelos foram avaliados com base nas métricas, curvas de acurácia e perda no treinamento e validação e matriz de confusão, discutidas em mais detalhes na Seção Resultados e Discussão. No intuito de interpretar os resultados dos modelos, foram criados mapas de calor para as ativações de determinadas camadas das redes neurais convolucionais. Essa abordagem visa a visualização das áreas da imagem que foram ativadas pelos filtros convolucionais em camadas específicas, permitindo a identificação das características das folhas que exerceram maior influência nas decisões da rede.

As métricas de avaliação dos modelos foram acurácia (4.5), precisão (4.7), recall (4.8), f1-score (4.9) e perda no conjunto de teste, com a acurácia sendo calculada também nos conjuntos de treinamento e validação (Hossin and Sulaiman, 2015). As métricas são calculadas com base nos valores de Verdadeiro Positivo (VP), Verdadeiro Negativo (VN), Falso Positivo (FP) e Falso Negativo (FN) registrados na matriz de confusão obtida durante as predições. VP representa o número de imagens de folhas corretamente classificadas em suas respectivas classes, enquanto VN se refere ao número total de imagens de outras categorias que foram classificadas corretamente. FN indica quantas vezes o classificador errou ao classificar imagens da classe posi-

tiva como negativas, enquanto FP revela quantas vezes o classificador cometeu o erro oposto, classificando imagens da classe negativa como positivas (Atila et al., 2021).

A perda é calculada conforme a Fórmula 4.10:

$$Perda(loss) = -\sum_{i=1}^{k} y_i * \log(\hat{y}_i) \tag{4.10}$$

k é cada classe dos dados, m é a média, y^j representa a probabilidade *Softmax* 4.11 na classe i, e y_i representa o valor verdadeiro correspondente. Essa perda é uma medida de quão distantes estão duas probabilidades diferentes. O sinal negativo garante que a perda diminua à medida que as distribuições se aproximam uma da outra (Narayana and Ramana, 2022).

A função *Softmax* é calculada pela Fórmula 4.11:

$$\begin{aligned} P(y = j|z^{(i)}) = \phi_{\text{softmax}}(z^{(i)}) = \frac{e^{z^{(i)}}}{\sum_{j=0}^{k} e^{z_k^{(i)}}}, \\ z = w_0 x_0 + w_1 x_1 + \cdots + w_m x_m = \sum_{i=0}^{m} w_i x_i = \mathbf{W}^T \mathbf{X}. \end{aligned} \tag{4.11}$$

A função *Softmax* (4.11) calcula as probabilidades de pertencimento de uma amostra z_i a diferentes classes j, dadas as entradas z_i e os pesos associados. Isso é feito calculando o exponencial de cada entrada z_i e normalizando esses valores pela soma dos exponenciais de todas as entradas, resultando em probabilidades para cada classe (Low and Choo, 2018).

Capítulo 5

Resultados e Discussão

5.1 Dados fisiológicos

5.1.1 Análise dos cruzamentos

Após a rotulagem dos cruzamentos nas quatro classes definidas (NT-NP, T-NP, NT-P e T-P), buscou-se analisar como o comportamento dos cruzamentos em relação à produtividade e tolerância à seca variou ao longo do tempo. As visualizações das figuras 5.1 e 5.2 mostram o comportamento dos cruzamentos rotulados ao longo das coletas de dados.

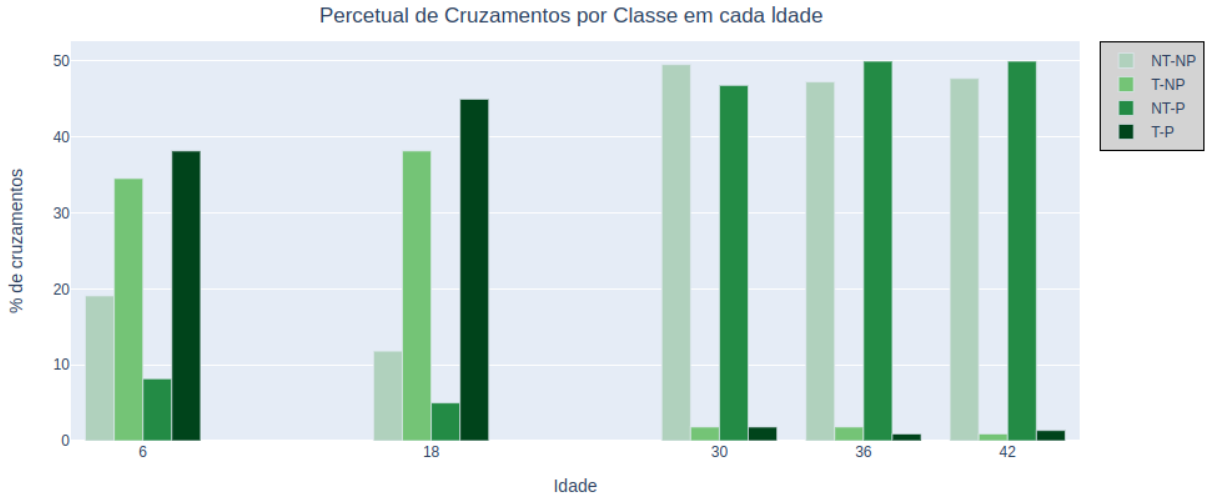

Figura 5.1: Gráfico de barras com o percentual de cruzamentos por classe em cada idade de coleta.

No Gráfico da Figura 5.1, aos 6 meses de idade, a maioria dos materiais (76%) foram classificados como tolerantes e, dentre estes, 38% dos cruzamentos foram também produtivos, ou seja, classe **T-P**. Em relação à produtividade, 54% foram considerados não produtivos, enquanto 46% produtivos. O fato da grande maioria dos cruzamentos serem tolerantes nesta idade, mesmo com a baixa precipitação acumulada no período entre o plantio e a coleta de 6 meses, média de 7.91mm de precipitação acumulada, mostra que as plantas priorizaram a tolerância ao estresse hídrico à produtividade.

Isso pode ser em função da menor sensibilidade das plantas em desenvolvimento inicial aos danos causados pela seca, pois a planta geralmente possui uma área foliar menor em comparação com estágios mais avançados. Isso resulta em menor perda de água por transpiração, pois há menos superfície foliar exposta. Além disso, o desenvolvimento do sistema radicular durante o crescimento inicial da planta permitem às raízes penetrarem em camadas mais profundas do solo em busca de água, o que pode ajudar a obter um suprimento hídrico adicional (Taiz et al., 2015). Nestas condições, as mudas tendem a reduzir a capacidade fotossintética em função da menor absorção do CO2, consequentemente, têm menor crescimento (Nogueira and Silva, 2002).

Aos 18 meses, foi observado o aumento da temperatura máxima e da precipitação acumulada, com média de precipitação acumulada no período de 97,52 mm e pico de temperatura máxima de 42º, com um aumento nos percentuais de cruzamentos tolerantes (**T**-NP e **T**-P) e não produtivos (NT-**NP** e T-**NP**), ambos totalizando 50%. Esse aumento nos materiais tolerantes e produtivos (T-P) e a redução nos não tolerantes e não produtivos (NT-NP) refletem o impacto positivo da maior disponibilidade de água nesse período (entre 6 e 18 meses ocorreu o pico de precipitação acumulada) no desenvolvimento das plantas. No entanto, durante o período da coleta dos 30 meses, observa-se o oposto, com a predominância dos cruzamentos não tolerantes à seca (NT-NP e NT-P), coincidindo com um longo período de menor disponibilidade hídrica, conforme indicado no gráfico de dados climáticos da Figura 4.1. Notavelmente, os percentuais de cruzamentos não tolerantes e produtivos (NT-P) foram os mais altos, evidenciando uma estratégia de adaptação das plantas que priorizou a produtividade em detrimento da tolerância à seca.

O mapa de calor da Figura 5.2 mostra uma visão geral do teste de progênies, nele é possível visualizar todos os 220 cruzamentos do experimento devidamente rotulados e em todas as idades. O materiais genéticos que são testemunha estão destacados em verde e vermelho, com este último sendo apenas o clone testemunha VM1, que obteve o melhor desempenho entre os seis clones testemunhas.

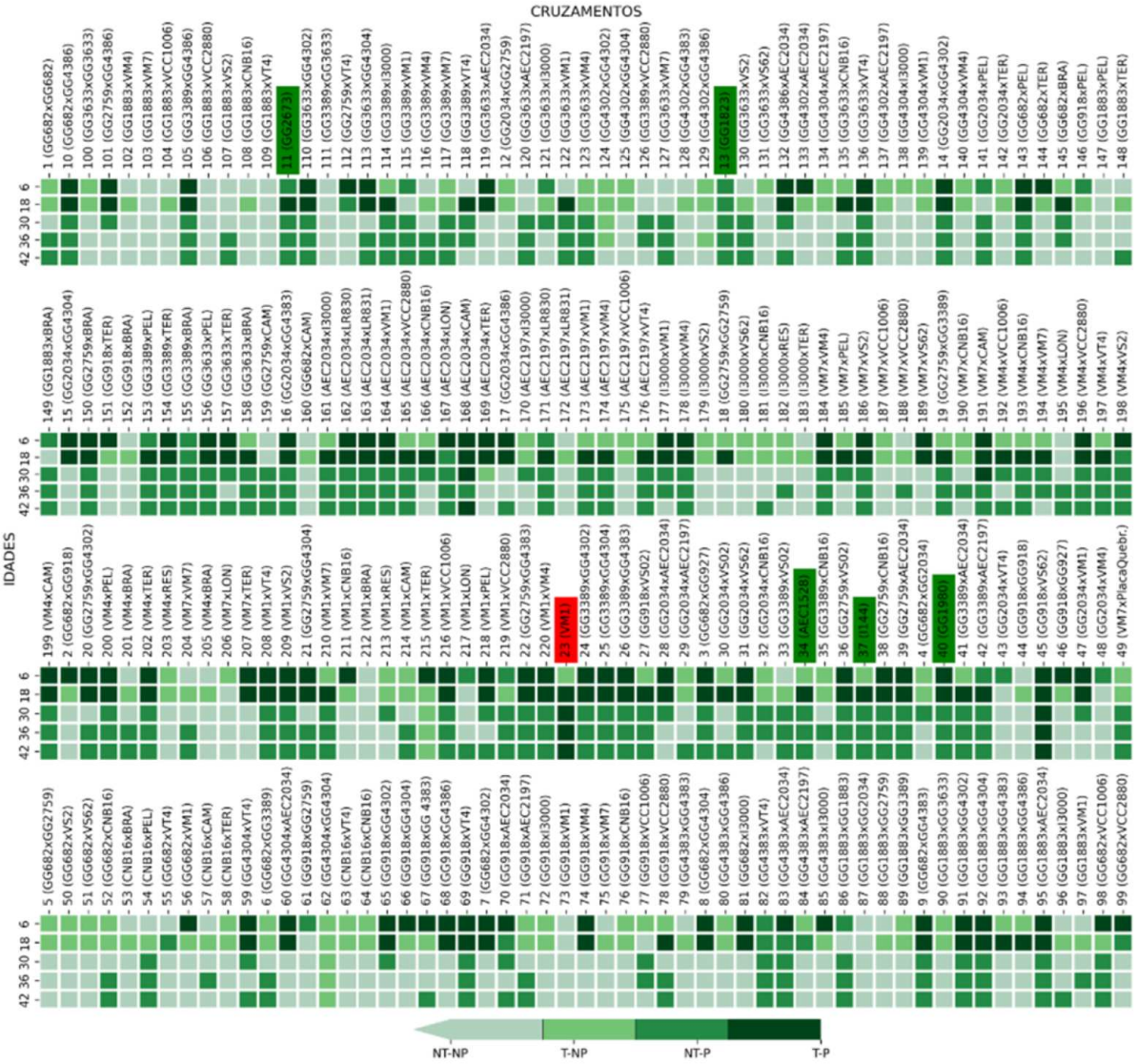

Figura 5.2: Mapa de calor com a visão geral do Teste de Progênies: classificação dos cruzamentos quanto à tolerância à seca e produtividade nas quatro idades de coleta (6, 18, 30, 36 e 42 meses), destacando os cruzamentos testemunhais (verde) e o cruzamento VM1 (vermelhor).

Na Figura 5.2 pode-se observar que os cruzamentos 23 (VM1 - testemunha) e 45 (GG918xvs62) mantiveram-se consistentemente na classe de tolerantes e produtividade (TP) ao longo dos 42 meses de avaliação, com exceção do 23 aos seis meses, quando foi classificado como tolerante e não produtivo (T-NP). O cruzamento 45 (GG918xvs62) manteve-se tolerante e produtivo em todas as idades.

A análise aos seis meses revelou que apenas uma das testemunhas, a 40 (GG1980), demonstrou ser tolerante e produtiva (TP), enquanto três foram classificadas como T-NP (23, 34, 37) e duas como NT-P (11 e 13). Nas idades subsequentes, todas as testemunhas foram produtivas, no entanto, apenas o VM1 foi categorizado como T-P. Para os materiais que não eram testemunha, a maioria foi rotulado na categoria de

não tolerante a partir dos 30 meses, com exceção das progênies 62 (GG4304xGG4304) e 215 (VM1xTER), que que foram rotuladas na categoria (T-NP).

Aos 42 meses apenas os cruzamentos 168 (AEC2034xCAM), 45 (GG918xV562) e a testemunha 23 (VM1) estavam na categoria T-P, ou seja, 1,39% do total de cruzamentos. No entanto, quando levamos em conta os cruzamentos apenas tolerantes, este valor sobe para 2,31%. Esse padrão sugere que a maioria dos materiais sofreu fortemente com o estresse hídrico a partir dos 30 meses, levando à perda da característica de tolerância.

A variação de comportamento das diferentes progênies em relação à produtividade está diretamente relacionada à capacidade de crescimento dos indivíduos para aumentar o volume, como descrevem (KLIPPEL et al., 2014) ao mostrarem que o desenvolvimento de raízes profundas é uma estratégia que as plantas utilizam para se proteger da escassez de água. Ao expandir suas raízes em direção a camadas mais profundas do solo, as plantas conseguem evitar a falta de água que poderia prejudicar seu crescimento em condições de campo. Enquanto a maior sobrevivência depende das estratégias de tolerância à seca de cada espécie (Mulkey et al., 2012). Ambos comportamentos, crescimento e sobrevivência, são influenciados por mecanismos a nível subcelular, como ajuste osmótico e eficiência do uso da água que permitem às plantas manter o turgor sob potenciais hídricos mais baixos (Taiz and Zeiger, 2007) e um bom desempenho momentâneo das trocas gasosas na folha (Larcher, 2006), respectivamente.

Um comportamento interessante a ser destacado é a transição de não tolerância para tolerância à seca em alguns cruzamentos. Embora os cruzamentos originalmente classificados como não tolerantes (NT) tenham registrado mais de 10% das plantas como "não sobreviventes"em campo, pela ausência de folhas nas plantas, é importante notar que a perda total de folhas não necessariamente indica a morte das plantas. De acordo com (KLIPPEL et al., 2014), a perda total de folhas pode ser uma resposta adaptativa das plantas a ambientes com restrição hídrica. Portanto, plantas que perderam totalmente as folhas não estavam necessariamente mortas e puderam recuperar a folhagem assim que a disponibilidade de água aumentou.

5.1.2 Análise das variáveis

Os gráficos de *boxplot* das Figuras 5.3, 5.4, 5.5, 5.6 e 5.7 mostram a distribuição das variáveis fisiológicas utilizadas com entrada para os modelos ao longo das coletas, agrupadas por classe. Os pontos na cor laranja representam a média da variável em cada idade de coleta, e a linha mostra a tendência da média ao longo do tempo.

Na Figura 8 vemos que a AFE diminuiu entre 6 e 18 meses, alguns estudos sugerem que plantas de eucalipto submetidas ao déficit hídrico apresentam redução nesta

variável (Oliveira, 2021, Pita-Barbosa et al., 2023, Tatagiba et al., 2007), tal estratégia pode estar associada à redução da área de recebimento de radiação solar, reduzindo a perda de água por transpiração, o que pode ser considerada a primeira defesa contra a deficiência hídrica (Chaves et al., 2004). Embora o período no qual ocorreu a redução da área foliar específica tenha sido também o que teve maior precipitação acumulada, igualmente mais altas foram as temperaturas, o que exige a evaporação de grandes quantidades de água para manutenção da temperatura foliar abaixo da temperatura do ar (Taiz and Zeiger, 2007). Logo, nestas condições, menor área foliar exige menos perda de água para controle de temperatura da folha.

Após os 18 meses, a média da variável se manteve constante até a coleta de 30 meses e teve um leve aumento dos 30 aos 36 meses. Esta estabilização da AFE pode indicar outra maneira da planta se adaptar às mudanças ambientais, no sentido de não reduzir significativamente a área foliar a ponto de prejudicar a fotossíntese (pela diminuição da área de recepção de luz) e as trocas gasosas para assimilação de CO2 (pelo fechamento estomático), impactando na produtividade (Peixoto, 2020). Portanto, isso pode evidenciar uma estratégia das plantas para equilíbrio entre tolerância à seca e produtividade em um ambiente de maior escassez de água. Um aspecto que chama a atenção nos gráficos das variáveis fisiológicas é que aos 36 meses nenhum cruzamento foi classificado como Tolerante e Não Produtiva (T-NP), isso mostra que as plantas menos produtivas também não suportam a escassez de água.

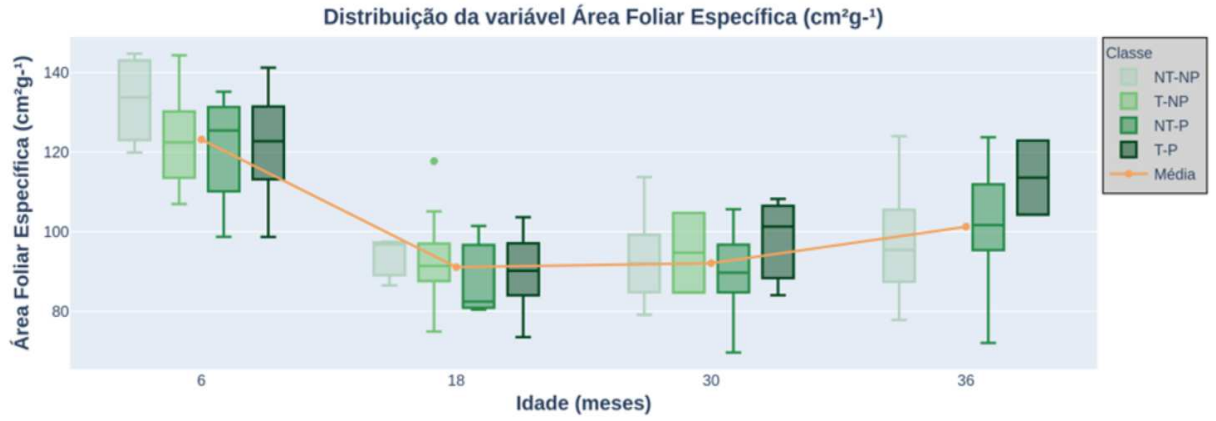

Figura 5.3: Distribuição da Área Foliar Específica dos cruzamentos, agrupados por classe, ao longo das coletas.

A distribuição do Potencial Hídrico Foliar, mostrada na Figura 5.4, mostra um comportamento descrito na literatura por (Pritzkow et al., 2020), registrando forte queda no valor médio e na distribuição geral desta variável em todas as classes dos 18 aos 30 meses, período no qual houve seca mais intensa e altas temperaturas. Pode-se observar que até os 18 meses, intervalo no qual a precipitação acumulada foi mais elevada, o potencial hídrico tanto de indivíduos tolerantes, quanto não tolerantes se

manteve muito próximo (-2,5 MPa). Já aos 30 meses, com agravamento da seca, os indivíduos tolerantes apresentaram um potencial hídrico um pouco mais elevado que os não tolerantes. Aos 36 meses, com a maior disponibilidade hídrica, o padrão da variável voltou a ser parecido entre cruzamentos T-P e NT-NP.

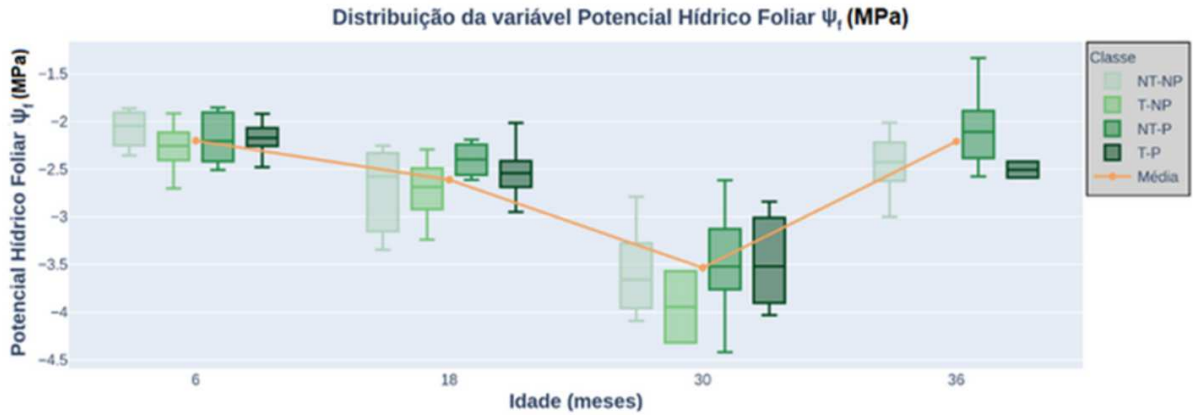

Figura 5.4: Distribuição do Potencial Hídrico dos cruzamentos, agrupados por classe, ao longo das coletas.

A Figura 5.5 mostra que a partir dos 30 meses, com a intensificação da condição de seca, as plantas mais tolerantes e produtivas tiveram menor área foliar individual, já as não tolerantes tiveram distribuição parecida, com maior área foliar, evidenciando que sob seca a plantas tolerantes tiveram como resposta fisiológica a redução da área da folha, muito provavelmente, para reduzir a superfície transpiracional (Carignato et al., 2019, Oliveira, 2021).

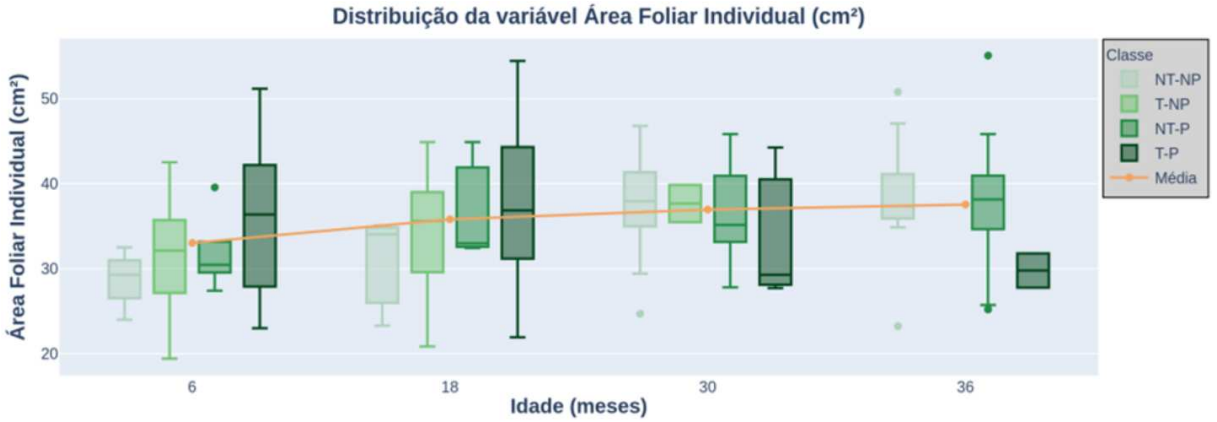

Figura 5.5: Distribuição do Área Foliar dos cruzamentos, agrupados por classe, ao longo das coletas.

Os gráficos das Figuras 11 e 12 evidenciam que as plantas, em um primeiro momento, aumentaram o comprimento do limbo foliar e reduziram a largura do limbo foliar, variando pouco estas características após os 18 meses.

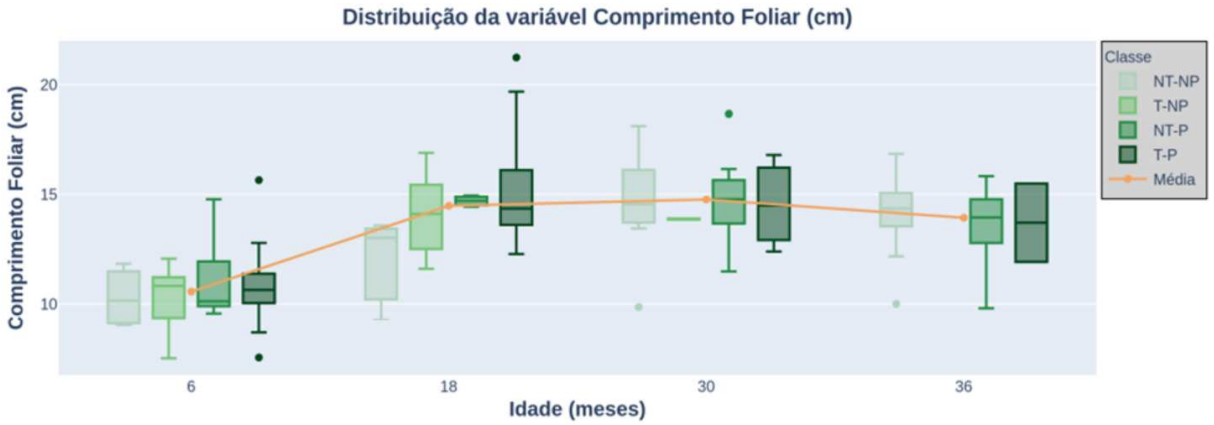

Figura 5.6: Distribuição do Comprimento Foliar dos cruzamentos, agrupados por classe, ao longo das coletas.

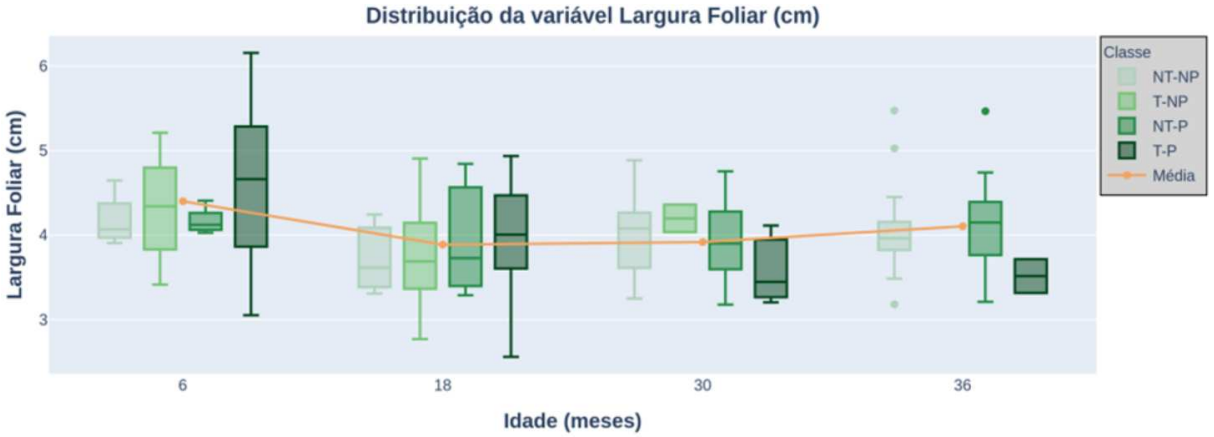

Figura 5.7: Distribuição do Largura Foliar dos cruzamentos, agrupados por classe, ao longo das coletas.

A Figura 5.8 mostra a evolução da produtividade dos cruzamentos medida pelo IMAVol (m^3/ha/ano). Entre 6 e 18 meses houve um ganho mais significativo de produtividade dos cruzamentos, com os rotulados como produtivos ficando com a mediana próxima a 20 m^3/ha/ano. Esse aumento pode ser explicado pela maior oferta de água no solo, pois coincide com o período de maior precipitação. Este comportamento foi observado no estudo de (Xavier et al., 2011), citado por (KLIPPEL et al., 2014), com plantas de híbridos de *Eucalyptus urophylla x Eucalyptus grandis*, cultivados em vasos e submetidos a diferentes níveis de déficit hídrico, apresentaram maior crescimento sob maior disponibilidade hídrica, sendo o diâmetro a medida mais dependente da disponibilidade de água.

Já entre os 18 e 30 meses, onde houve severo estresse hídrico, o IMAVol ficou estável, evidenciando que as plantas sofreram com esta condição, com estagnação de produtividade volumétrica. Tal resposta à condição de seca foi observada em

(Correia et al., 2016), onde verificou-se redução no crescimento e declínios na fotossíntese em plantas de eucalipto submetidas à seca. No estudo de (Martins, 2007), sobre a influência da deficiência hídrica em *Eucalyptus grandis e Eucalyptus saligna*, observou-se menor crescimento em altura e diâmetro nas plantas sob condições de seca. Para a variável IMAVol a idade de coleta vai até os 42 meses, diferentemente das fisiológicas que vão até 36, isso porque ela foi coletada para todos os indivíduos do teste de progênies e em todas as idades.

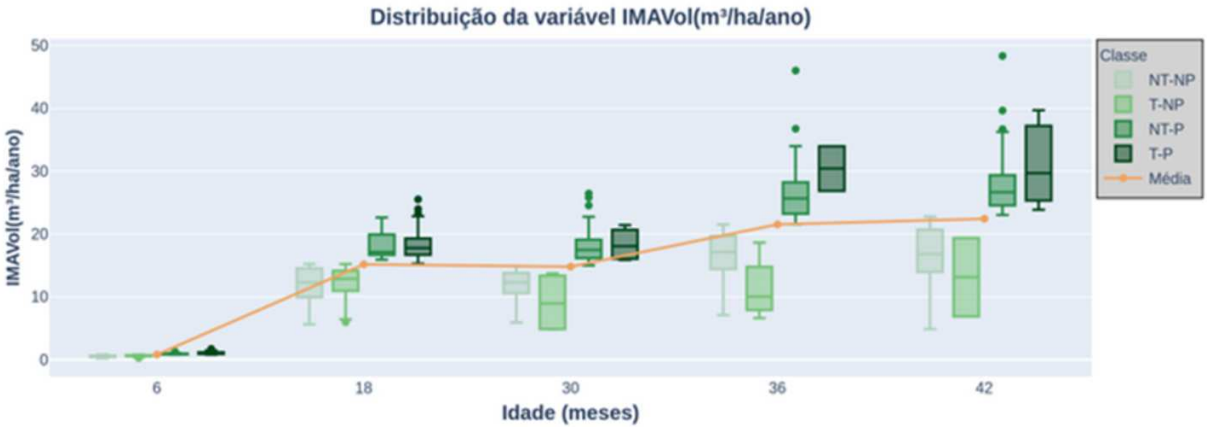

Figura 5.8: Distribuição do IMAVol dos cruzamentos, agrupados por classe, ao longo das coletas.

O gráfico de dispersão da Figura 5.9 mostra a relação entre a Área Foliar Específica e o Potencial Hídrico Foliar na separação das classes dos cruzamentos. Uma informação adicional é a produtividade, expressa pelo IMAVol por meio do tamanho dos pontos. É possível observar que é uma tarefa desafiadora separar os cruzamentos tolerantes (verdes mais escuro e intermediário) por estas duas variáveis fisiológicas, que são amplamente citadas na literatura como potenciais indicadoras de tolerância à seca. Embora a maioria dos cruzamentos tolerantes apresentem valores de ψf entre -2.4 e -2 MPa, ainda há relativa dispersão ao longo de toda a faixa de valores da coleta de seis meses. O mesmo ocorre com a AFE, onde cruzamentos tolerantes se concentram em sua maioria entre menos de 110 e 130 cm^2g^{-1}, com o mesmo padrão de dispersão geral da AFE.

Ademais, fica clara a mesma dificuldade quando se analisa a produtividade, os cruzamentos produtivos (verdes mais escuros) apresentam ψf maior que -2.2 MPa e AFE entre 110 e 130 cm^2g^{-1}, com uma dispersão relativamente significativa. A despeito da dispersão, o fato de a maioria dos pontos estarem agrupados nos dois intervalos mencionados acima, pode se justificar pela adaptação das plantas no sentido de equilibrar tolerância à seca e produtividade.

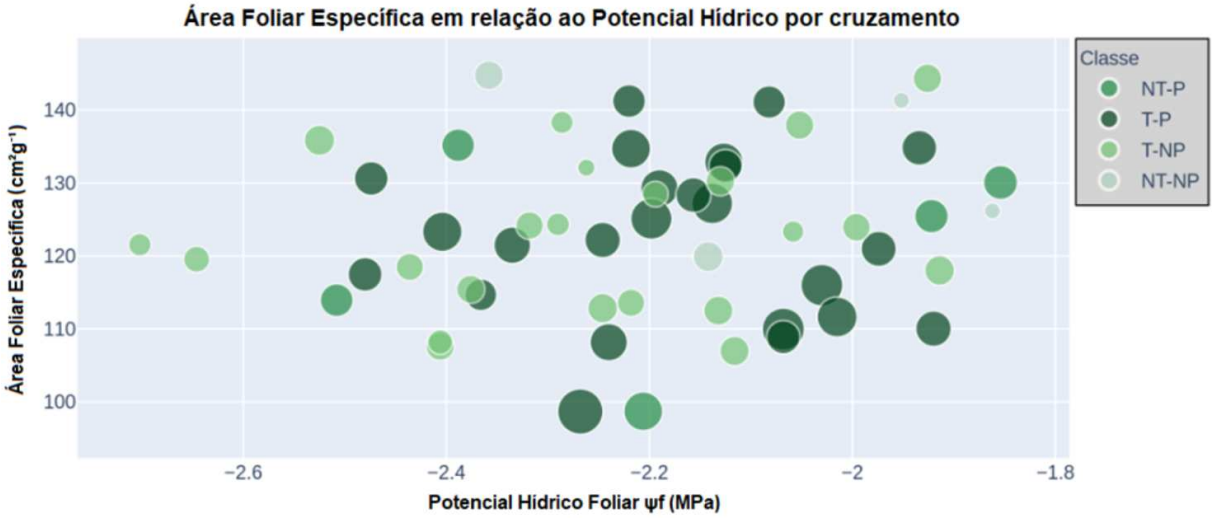

Figura 5.9: Gráfico de dispersão com a distribuição dos cruzamentos pelo potencial hídrico em função da Área Foliar Específica, com o tamanho dos pontos definido pela produtividade (IMAVol), na coleta de 6 meses.

5.1.3 Resultados dos modelos

Os quatro modelos foram avaliados pelas métricas da Tabela 5.1, nas idades alvo de 18, 30, 36 e 42 meses, para as quatro abordagens de rotulagem (Tolerância e Produtividade - 4 Classes; Tolerância e Produtividade - 2 Classes; Tolerância - 2 Classes; Produtividade - 2 Classes). O resultado de cada métrica corresponde à média das 10 iterações de validação cruzada *k-fold* (k=5) de cada métrica. Além das métricas apresentadas na tabela, os modelos foram avaliados pela matriz de confusão e também foi medida a importância relativa das variáveis para cada modelo, a fim de tentar identificar as que tiveram maior influência na predição.

Os resultados apresentados são para a abordagem de Tolerância e Produtividade - 4 classes, por ser a abordagem mais interessante do ponto de vista do programa de melhoramento florestal, uma vez que permite uma análise mais detalhada da classificação dos cruzamentos, considerando as quatro classes quanto a tolerância à seca e produtividade.

Tabela 5.1: Comparação do resultados dos algoritmos com os dados rotulados em duas classes de acordo com o critério de **Tolerância e Produtividade - 4 classes** em cada idade alvo.

Tolerância e Produtividade - 4 classes					
Modelo	**Acurácia**	**Precisão**	**Recall**	**F1-Score**	**Acurácia B.[1]**
Idade Alvo: 18 meses					
Random Forest	**58.0±2.3**	**51.4±4.9**	**58.0±2.3**	**51.9±3.0**	**37.6±2.1**
SVC	58.54±1.67	46.57±4.45	58.54±1.67	48.64±2.13	35.85±1.37
MLP	53.45±3.34	40.68±4.31	53.45±3.34	44.20±2.63	32.61±1.97
XGBoost	58.18±0.0	34.05±0.0	58.18±0.0	42.90±0.0	33.33±0.0
Idade Alvo: 30 meses					
Random Forest	**57.64±4.11**	**55.44±3.98**	**57.64±4.11**	**54.94±4.15**	**50.63±3.75**
SVC	50.00±6.0	48.95±7.54	50.00±6.0	47.83±6.17	44.08±5.22
MLP	46.73±7.28	45.18±8.53	46.73±7.28	44.15±7.52	41.23±6.51
XGBoost	52.36±1.15	27.99±0.17	52.36±1.15	36.4±0.27	43.17±0.53
Idade Alvo: 36 meses					
Random Forest	**60.18±5.52**	**58.76±7.24**	**60.18±5.52**	**57.73±6.02**	**52.97±5.38**
MLP	58.18±2.42	57.17±4.5	58.18±2.42	56.13±2.35	51.53±2.73
SVC	55.09±3.21	53.18±5.63	55.09±3.21	48.54±3.39	46.11±2.74
XGBoost	54.0±0.88	30.08±0.85	54.0±0.88	38.54±0.41	43.11±0.39
Idade Alvo: 42 meses					
SVC	**59.27±4.47**	**60.26±6.0**	**59.27±4.47**	**57.98±4.9**	**52.9±4.5**
MLP	57.45±2.3	57.86±2.46	57.45±2.3	55.61±2.6	51.66±2.66
Random Forest	53.82±3.45	52.05±4.1	53.82±3.45	51.37±3.5	47.92±3.64
XGBoost	52.73±0.0	27.93±0.0	52.73±0.0	36.48±0.0	43.33±0.0

1 - Acurácia Balanceada

O modelo *Random Forest* alcançou os melhores resultados para todas as métricas, em quase todas as idades alvo, com exceção dos 42 meses, onde o *SVC (Support Vector Machine for Classification)* obteve o melhor desempenho, conforme a Tabela 5.1. A idade de 42 meses é a idade mais longeva com dados coletados, portanto, a mais interessante do ponto de vista do propósito desta pesquisa. Neste sentido, o modelo SVC foi o modelo com melhor desempenho, com 57.98% ±4.9 de F1-score e 52.9% ±4.5 de Acurácia Balanceada, com leve diferença em relação ao *MLP*, considerando o desvio padrão. O modelo *XGBoost* teve a pior performance em todos os cenários.

Rotular as amostras pelos critérios de tolerância à seca e produtividade em cada idade de coleta teve como objetivo analisar a performance preditiva dos modelos ao longo do tempo, de modo a observar as tendências de cada modelo e poder comparar

o desempenho com os novos dados que ainda surgirão ao longo do projeto de melhoramento genético florestal. Neste sentido, os resultados da idade alvo de 18 meses se diferenciam mais acentuadamente das demais idades, inclusive apresentando os piores resultados para a métrica de Acurácia Balanceada, com o *Random Forest* chegando a 37.6% ±2.1. Para a métrica F1-score os resultados também foram piores, com 51.9% ±3 para o *Random Forest*. O mesmo se observa com a Precisão, 51.4% ±4.9. Estas duas últimas métricas tiveram este resultado em função da alta taxa de Falsos Positivos que pode ser observada na matriz de confusão para a idade alvo de 18 meses Figura 5.10.

Em todas as idades houve um desbalanceamento entre as classes dos cruzamentos, sendo que aos 18 meses houve uma prevalência dos tolerantes, como pode ser observado na Tabela 4.1 e na Figura 5.1. De maneira oposta, após os 30 meses, a classe não tolerante foi majoritária. Este desbalanceamento pode ter afetado a performance dos modelos (Castro and Braga, 2011b). Além deste fator, as variáveis avaliadas, embora amplamente mencionadas na literatura como boas preditoras de tolerância à seca e produtividade, podem não ter sido representativas ou em quantidade suficiente para um ajuste satisfatório dos modelos de classificação para o problema proposto. Ademais, pelo fato de os dados serem provenientes de um experimento de campo, sujeito a condições climáticas extremas, eles estão mais sujeitos a variações irregulares e imprecisões por vícios de medição.

Todas estas questões trazem para os modelos uma maior dificuldade de aprender os padrões que refletem estas condições. Ao se analisar os modelos ao longo do tempo, os resultados mostram certa instabilidade para todas as métricas em função do valor relativamente alto do desvio-padrão, principalmente a partir dos 30 meses. O *Random Forest* é o modelo que teve uma variação relativamente pequena dos valores das métricas ao longo das idades alvo.

Matriz de confusão

As matrizes de confusão da Figura 5.10 trazem a soma das previsões para cada classe ao longo das 10 iterações de validação cruzada. Pelas matrizes de confusão podemos observar que os modelos tiveram melhor desempenho nas predições das classes com mais amostras. Na idade alvo de 18 meses, as classes majoritárias foram as duas tolerantes (T-P e T-NP), com a T-P sendo a mais representada, sendo a que os modelos mais acertaram as previsões. O *Random Forest* foi o modelo que mais variou as predições entre as classes, embora os acertos tenham se concentrado apenas nas duas classes tolerantes, com 60% de acertos na classe T-P e 44% para a classe T-NP. O *SVM* teve 59% de acertos para a classe T-P e 52% para a T-NP. O *MLP* teve resultado parecido: 54% para T-P e 42% na T-NP. O *XGBoost* classificou todas as amostras como T-P, acertando 55% das predições. As previsões para as classes minoritárias foram

todas incorretas para todos os modelos na idade alvo de 18 meses.

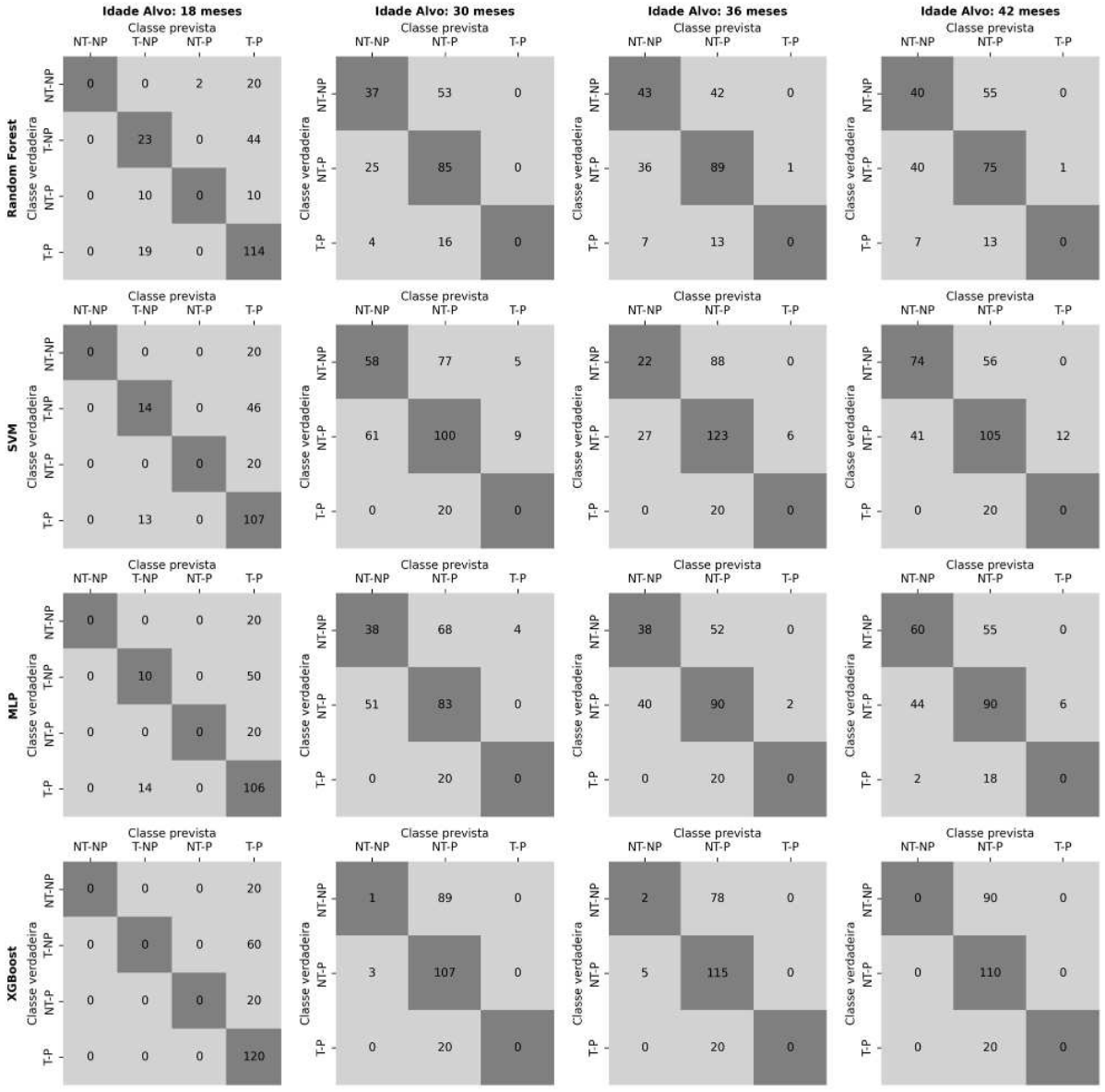

Figura 5.10: Matriz de confusão: Tolerância e Produtividade - 4 classes.

A partir da idade alvo de 30 meses, com os cruzamentos não tolerantes sendo mais representados, inclusive sem nenhum rotulado como tolerante e não produtivo (T-NP), o comportamento dos modelos se manteve, com os acertos concentrados apenas nestas classes (NT-NP e NT-P). A diferença para a idade de 18 meses foi que as amostras aqui ficaram bem equilibradas entre as duas classes, consequentemente, o percentual de acertos dos modelos para ambas acompanhou esta divisão.

O *RF* teve os percentuais de acurácia por classes de 56% (NT-NP) e 55% (NT-P) aos 30 meses, 50% (NT-NP) e 61% (NT-P) aos 36 meses, 45% (NT-NP) e 52% (NT-P) aos 42 meses. Na mesma sequência de idades alvo e classes, o *SVM* obteve as acurácias de 49% - 51%, 45% - 53% e 64% - 58%, respectivamente. Enquanto o *MLP*,

segundo a mesma ordem, alcançou 43% - 49%, 49% - 56% e 57% - 55%. A exceção foi o *XGBoost*, pois concentrou praticamente todas todas as suas previsões na classe com mais cruzamentos, a NT-P, sendo o modelo com a menor capacidade preditiva.

Em todos os cenários, os modelos tiveram dificuldade com as previsões para as classes minoritárias, o que reforça a necessidade de explorar mais variáveis e adquirir mais amostras para aumentar o conjunto de dados.

Importância relativa das variáveis

Para o cálculo da importância relativa das variáveis foi utilizada a biblioteca *permutation_importance*, da biblioteca *scikit-learning*, da linguagem *Python*. Seu funcionamento baseia-se na permuta aleatória dos valores de uma variável de cada vez no conjunto de teste, medindo como essa permutação afeta a acurácia. Quanto mais a acurácia diminui quando uma variável é permutada, mais importante essa variável é considerada. O processo é repetido para todas as variáveis, resultando em uma classificação das variáveis com base em sua contribuição para o desempenho do modelo, o que ajuda na compreensão das relações entre elas e as previsões feitas (Pedregosa et al., 2011).

Ao avaliar a importância relativa das variáveis preditoras, de acordo com a Figura 5.11, as variáveis de área foliar (AFI), largura do limbo foliar (LF) e área foliar específica (AFE) mostraram-se mais importantes para as previsões do modelo de melhor desempenho, o *SVM*, na idade alvo de 42 meses. Nas idades-alvo de 36 e 18 meses, as duas medidas de área foram as mais significativas para o *SVM*. No caso do RF, o potencial hídrico foliar foi a variável mais importante aos 18 meses, enquanto aos 30 e 36 meses, foi o CF, e, por fim, a LF e as AFE e AF.

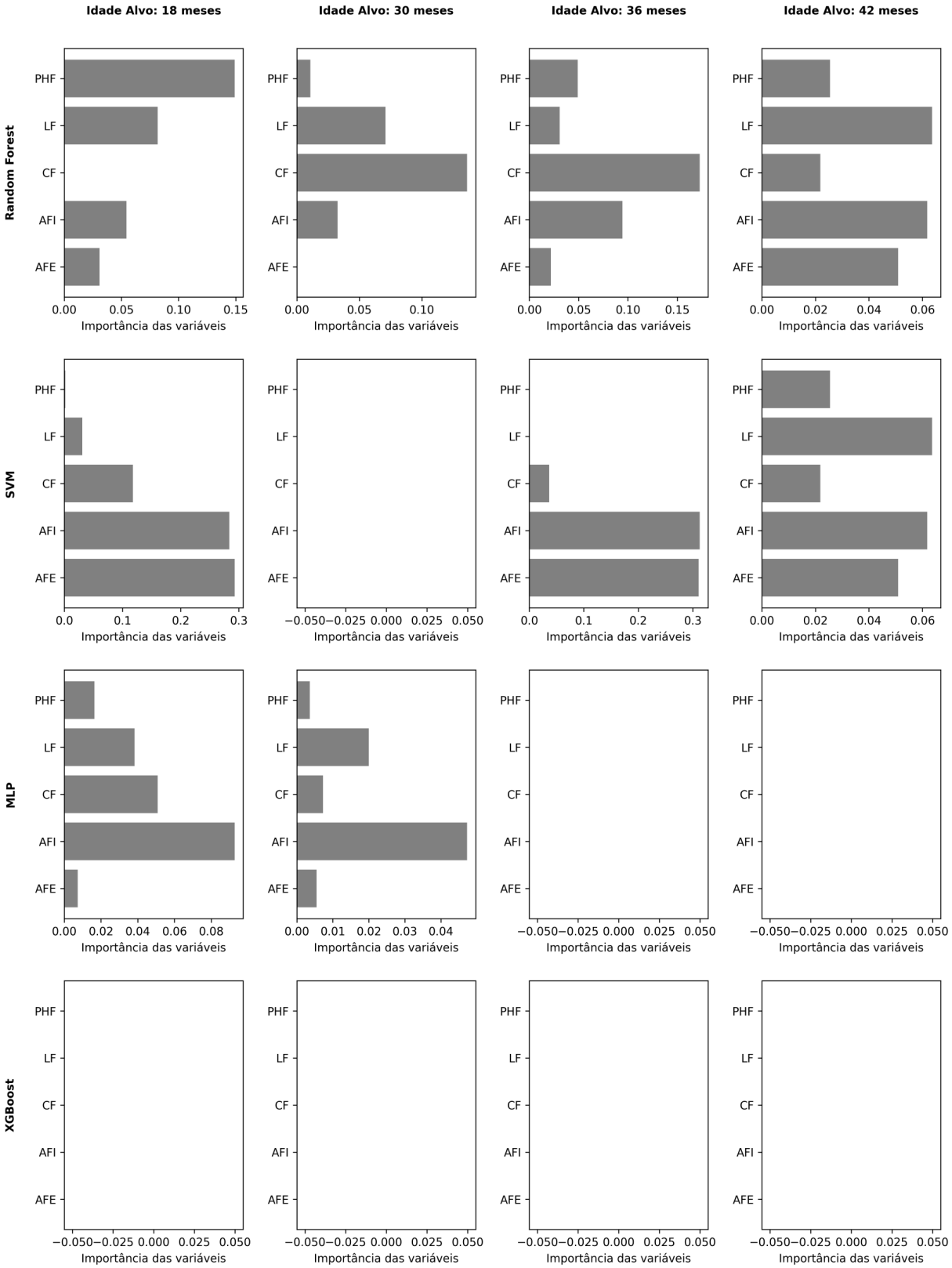

Figura 5.11: Importância relativa das variáveis para cada modelo, em cada idade alvo.

Nos gráficos da Figura 5.11 que ficaram em branco (*SVM* - idade alvo de 30 meses; *MLP* - 36 e 42 meses; *XGBoost*), conforme a documentação da biblioteca *scikit-learning*,

significa que a importância relativa de todas as variáveis é muito próxima de zero ou que não há variação nas métricas ao permutar as variáveis. Isso pode ter como causa o fato de as variáveis serem pouco informativas ou pelo fato de o conjunto de dados ter tamanho insuficiente para a análise da permutação das variáveis, podendo não produzir variações significativas nas métricas de desempenho quando permutadas as variáveis (Pedregosa et al., 2011).

Configuração de hiperparâmetros

Os resultados dos melhores conjuntos de hiperparâmetros encontrados por meio da busca em grade utilizando a biblioteca *GridSearchCV* do Python são apresentados na Tabela 5.2. Estes resultados correspondem aos melhores modelos em cada idade e estratégia de rotulagem.

Tabela 5.2: Hiperparâmetros selecionados pelo *GridsearchCV* e escala dos dados para o melhor modelo em cada forma de rotulagem e idade alvo.

Rotulagem	**Idade Alvo**	**Melhor modelo**	**Hiperparâmetros**
TP[1] - 4 classes	18	Random Forest	n_estimators: 500; max_features: auto; max_depth: 10
	30	Random Forest	n_estimators: 1000; max_features: auto; max_depth: 15
	36	Random Forest	n_estimators: 1000; max_features: auto; max_depth: 10
	42	SVC	C: 100; kernel: rbf; gamma: scale

[1] **T**olerância e **P**rodutividade

5.2 Dados de Imagem

Esta seção descreve os resultados obtidos por meio da abordagem proposta, bem como a discussão acerca dos mesmos. Os resultados apresentados são referentes à forma de rotulagem "Tolerância e Produtividade - 4 classes", enquanto os resultados para outras formas de rotulagem estão disponíveis no material suplementar. Optamos por discutir essa forma de rotulagem em particular devido à sua capacidade de detalhar melhor a seleção de materiais genéticos, permitindo a classificação em 4 classes, sendo, assim, de maior interesse para o programa de melhoramento florestal.

5.2.1 Resultados dos modelos

Os resultados da Tabela 5.3 incluem a acurácia nos conjuntos de treinamento, validação e teste, juntamente com a precisão, *recall*, *f1-score* e a perda (*loss*) no conjunto de teste, ao aplicar as arquiteturas *MobileNetV2*, *Xception* e *ResNet50* com transferência de aprendizado em cada idade alvo. Os melhores valores de acurácia nos testes para cada idade alvo foram destacados em negrito.

Tabela 5.3: Resultados das métricas (em %) da aplicação das redes neurais convolucionais nos conjuntos de treinamento (90% dos dados com validação cruzada k-fold) e teste (10% dos dados) considerando o critério de rotulagem de **Tolerância e Produtividade - 4 Classes** para todas idades alvo.

Tolerância e Produtividade - 4 classes						
Arquitetura	**Acurácia**^{Tr}	**Acurácia**^{Va}	**Acurácia**^{Te}	**Precisão**	**Recall**	**F1-score**
Idade Alvo: 18 meses						
MobileNetV2	99.96	65.23	68.55	75.92	52.79	57.61
Resnet50	99.99	67.19	66.97	66.45	55.96	59.31
Xception	99.99	68.88	67.19	67.08	52.62	57.32
Idade Alvo: 30 meses						
MobileNetV2	99.99	67.45	67.04	60.08	61.86	60.87
Resnet50	99.99	62.76	66.59	66.42	57.27	60.45
Xception	99.99	66.54	63.21	57.49	52.53	54.16
Idade Alvo: 36 meses						
MobileNetV2	99.99	66.41	66.14	65.06	62.37	63.58
Resnet50	99.99	65.63	67.49	63.26	67.38	65.06
Xception	99.99	70.70	62.98	54.92	55.87	55.09
Idade Alvo: 42 meses						
MobileNetV2	99.99	68.49	67.49	64.41	59.96	61.84
Resnet50	99.99	68.23	67.04	63.78	61.62	62.62
Xception	**99.99**	**72.00**	**71.56**	**69.87**	**60.68**	**63.86**

^{Tr} Acurácia de treino da validação cruzada

^{Va} Acurácia de validação da validação cruzada

^{Te} Acurácia do conjunto de teste

Precisão, *Recall* e *F1-score* são para o conjunto de teste

Na análise, a CNN *MobileNetV2* alcançou a melhor acurácia de teste nas idades de 18 e 30 meses, atingindo 68.55% e 67.04%, respectivamente. Aos 36 meses, a *ResNet50* registrou a melhor acurácia de teste, com 67.49%. Entretanto, aos 42 meses, a idade mais avançada, o modelo *Xception* superou todas as métricas no conjunto de testes, alcançando 71.56%, 69.87%, 60.68% e 63.86% para acurácia, precisão, *recall* e *f1-score*, respectivamente. Além disso, a perda foi a menor para este modelo, atingindo 1.11, sendo inclusive a menor em relação aos outros modelos e em todas as idades. Vale destacar que a acurácia de validação da *Xception* aos 42 meses também superou a de todos os modelos em todas as idades, com 72%.

Os resultados favoráveis da *Xception* aos 42 meses, a idade mais avançada com dados disponíveis e, portanto, a mais relevante para a previsão precoce do comportamento futuro dos cruzamentos, demonstram que o modelo ajustado obteve um desempenho superior em relação aos demais na predição precoce (usando folhas de plantas com 6 meses de plantio) dos cruzamentos em relação à tolerância à seca e produtividade em idade mais próxima da produtiva.

A Figura 5.12 (treliça de gráficos de barras) ilustra os resultados das métricas para os modelos ajustados no conjunto de testes em cada idade alvo, conforme os dados detalhados na Tabela 5.3, que apresenta os resultados da forma de rotulagem dos dados 'Tolerância e Produtividade - 4 classes'.

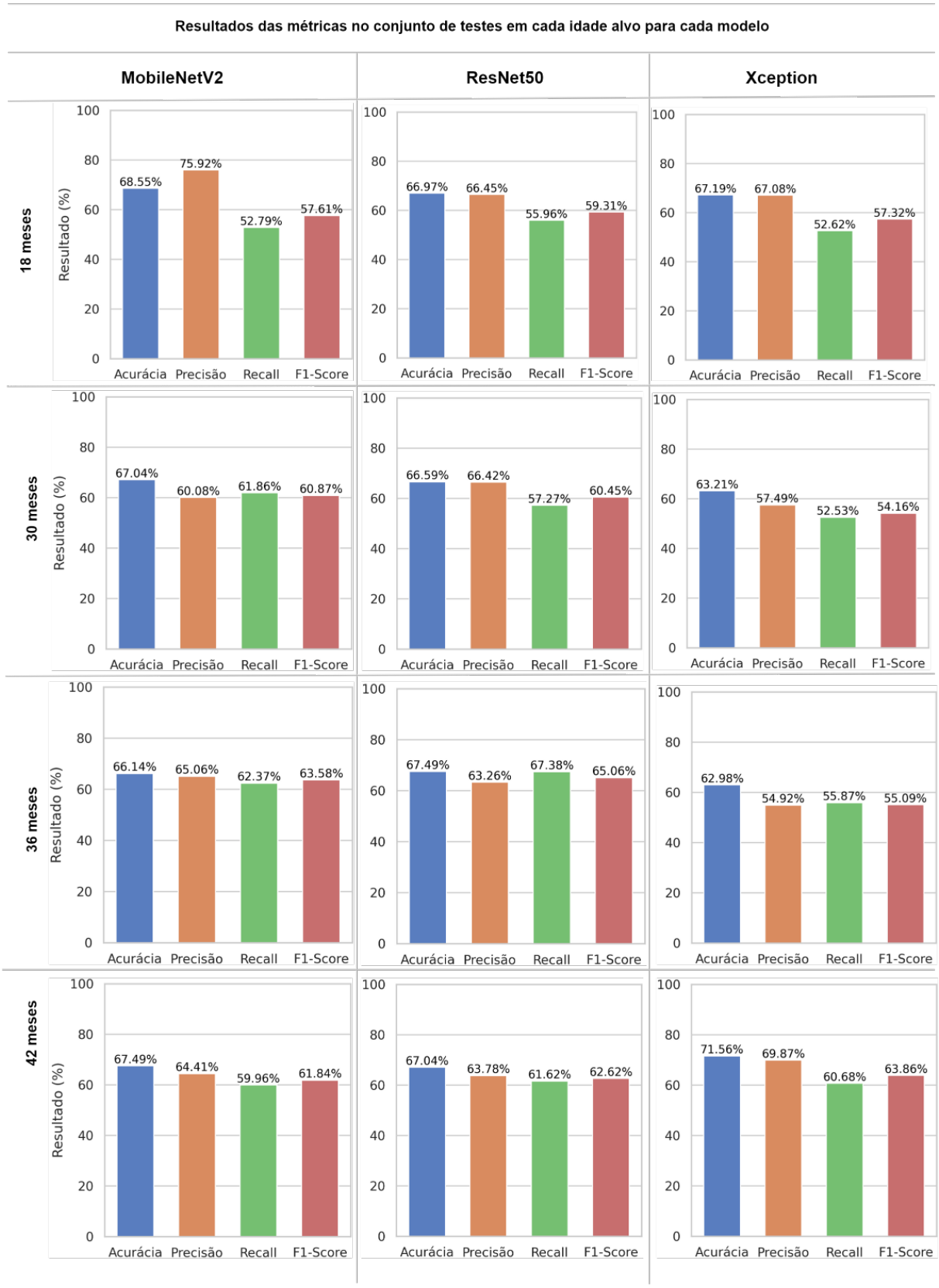

Figura 5.12: Treliça de gráficos de barras com as métricas para as CNNs nas idades alvo para a rotulagem em 4 classes.

Os gráficos da Figura 5.12 mostram que não houve diferenças muito significati-

vas nas métricas para o conjunto de testes entre os modelos nas idades alvo, o que pode indicar que as mudanças no desenvolvimento das plantas, afetadas em grande parte pelas condições ambientais desafiadoras (Pita-Barbosa et al., 2023) no decorrer do teste de progênie, foram bem assimiladas pelos modelos.

A acurácia é uma das métricas de avaliação de modelos mais amplamente utilizadas na literatura. No entanto, do ponto de vista comercial, quando se trata da seleção precoce de materiais de interesse (tolerantes e produtivos) para a propagação de plantios, a métrica de **precisão** se apresenta como uma medida mais eficaz. Isso porque uma alta precisão indica taxas reduzidas de falsos positivos (FP), ou seja, o modelo faz poucas previsões de cruzamentos que não são, na verdade, tolerantes e produtivos. Em outras palavras, se o modelo comete poucos erros ao identificar precocemente um material de interesse, ele não indicará erroneamente materiais inadequados para o plantio, o que, por sua vez, evita prejuízos decorrentes da propagação de plantas que não são tolerantes e produtivas.

Neste sentido, a *MobileNetV2* obteve o melhor resultado de precisão aos 18 (75.92%) e 36 meses (65.06%), com a *ResNet50* sendo a melhor aos 30 meses (66.42%) e a *Xception* superando as duas aos 42 meses, com 69.87%.

Curvas de aprendizado

Os gráficos de aprendizado das Figuras 5.13 e 5.14 apresentam o comportamento dos modelos em relação à acurácia e perda, respectivamente, durante o treinamento e validação. A análise das curvas de acurácia é crucial para avaliar a capacidade do modelo de generalizar previsões em dados vistos e não vistos. Um modelo que se ajusta bem aos dados de treinamento, com a curva de acurácia crescendo e estabilizando em um valor considerado elevado, mas não à validação, pode indicar sobreajuste (overfitting). Se houver um desempenho ruim tanto no treinamento quanto na validação, isso pode sugerir underfitting (Géron, 2019). A perda de treinamento mede a diferença entre as previsões e os rótulos reais. Se diminuir ao longo do tempo, isso indica melhorias nas previsões (Géron, 2019). Portanto, são desejadas curvas de acurácia ascendentes e de perda descendentes durante o treinamento e validação, sem grandes divergências entre as linhas.

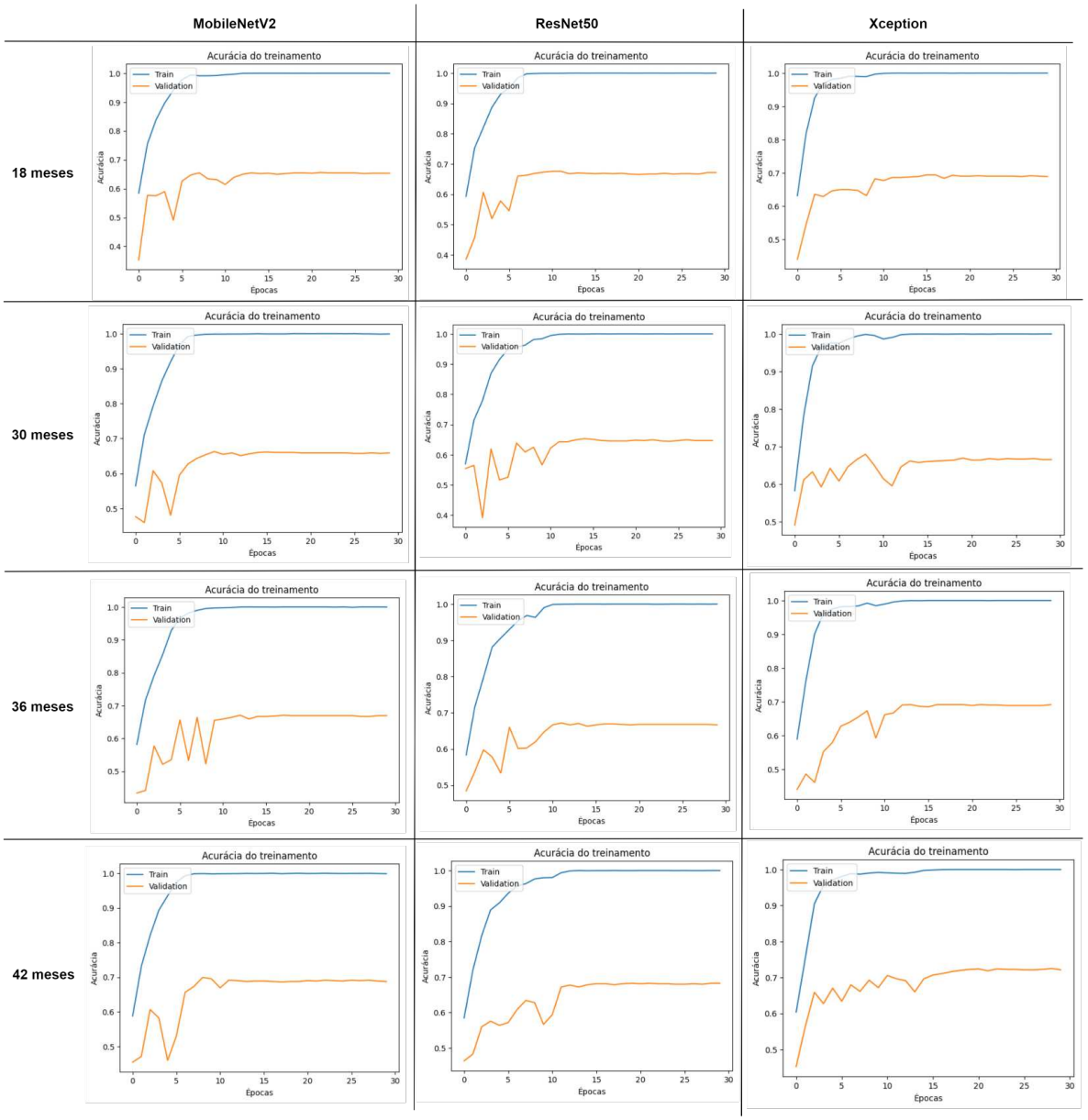

Figura 5.13: Curvas de acurácia para os modelos aplicados separadas por idade alvo para a rotulagem em 4 classes.

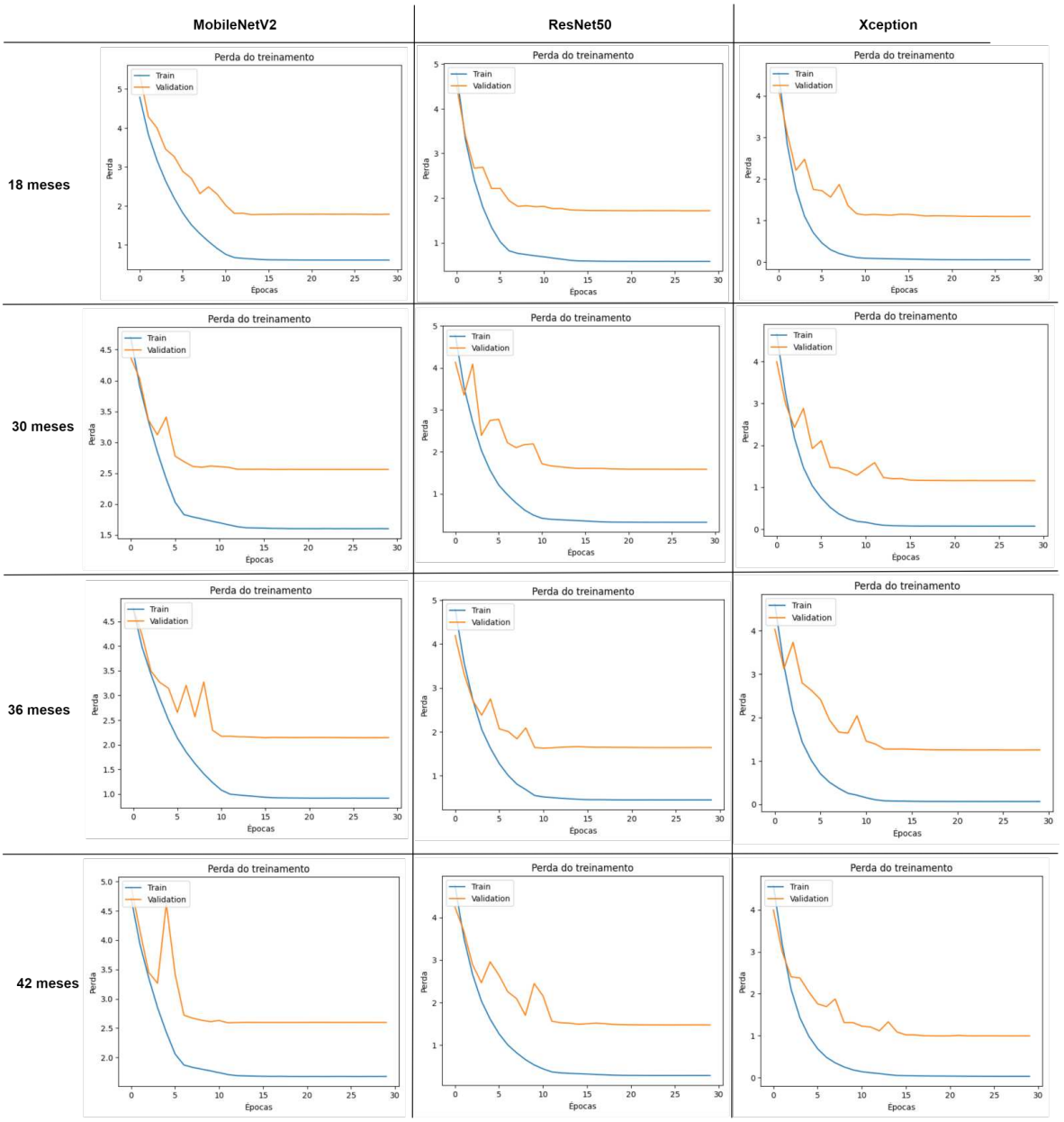

Figura 5.14: Curvas de perda para os modelos aplicados separadas por idade alvo para a rotulagem em 4 classes.

Pelas curvas pode-se observar que os modelos apresentaram desempenho consistente ao longo do treinamento, com as curvas de acurácia aumentando e as de perda diminuindo, indicando uma melhoria gradual no aprendizado dos modelos. Apesar das relativas diferenças entre as curvas de treinamento e validação, estas convergiram para valores próximos de 100% no treinamento e 70% na validação por volta das 10 épocas, evidenciando uma estabilização nos resultados.

É importante notar que, apesar das variações entre acurácias de treinamento e validação, os modelos mantiveram desempenho relativamente estável nos conjuntos de validação, sugerindo uma capacidade aceitável de generalização para a tarefa pro-

posta, especialmente considerando a complexidade do problema. Contudo, a ausência de diferenças visíveis entre as imagens das folhas das classes definidas pode ter afetado a precisão dos modelos em distinguir as classes, evidenciada pela distância entre as curvas de acurácia de treinamento e validação após a convergência.

Essa falta de distinção visual pode indicar uma representatividade insuficiente do conjunto de dados para uma modelagem mais precisa. Nesse sentido, a exploração de outros tipos de imagens, como imagens de microscopia das folhas das raízes das árvores (Pita-Barbosa et al., 2023) ou da copa das árvores (Xu et al., 2018), pode ser uma estratégia interessante para enriquecer o conjunto de dados. Essa abordagem pode permitir que os modelos aprendam com uma variedade maior de informações visuais, melhorando a generalização e distinção entre as classes, contribuindo para um aprimoramento nos resultados de classificação.

Matriz de confusão

As Figuras 5.15, 5.16 e 5.17 exibem as matrizes de confusão para cada modelo de CNN nas idades-alvo, usando a forma de rotulagem 'Tolerância e Produtividade - 4 classes' e o conjunto de dados de teste. Na idade alvo de 18 meses foi a única em que as imagens coletadas foram classificadas nas 4 classes definidas, com base na tolerância à seca e produtividade dos cruzamentos aos quais pertencem. Nas demais idades alvo, as classes das amostras coletadas foram não tolerante e não produtiva (NT-NP), não tolerante e produtiva (NT-P) e tolerante e produtiva (T-P).

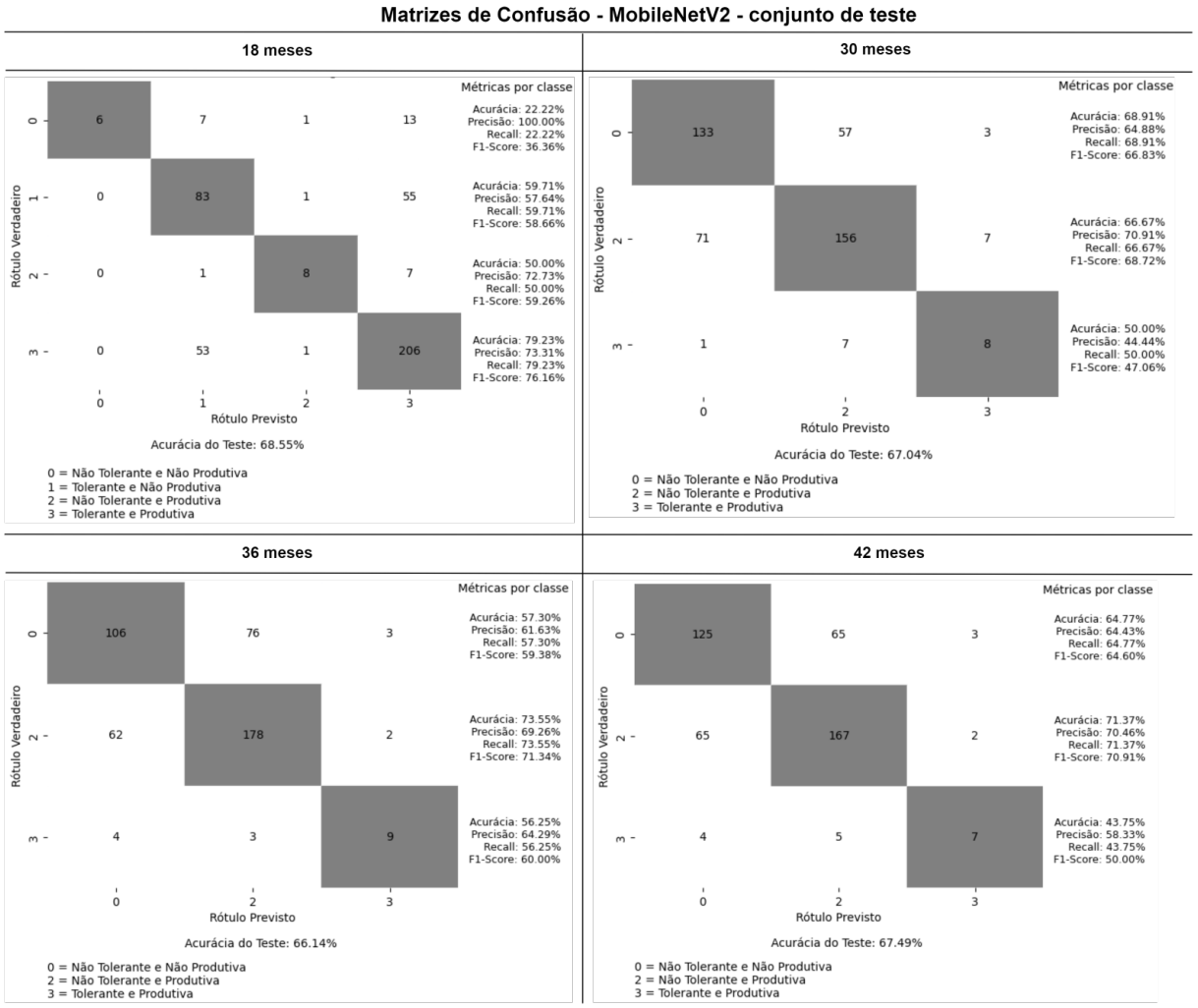

Figura 5.15: Matrizes de confusão da *MobileNetV2* em cada idades alvo.

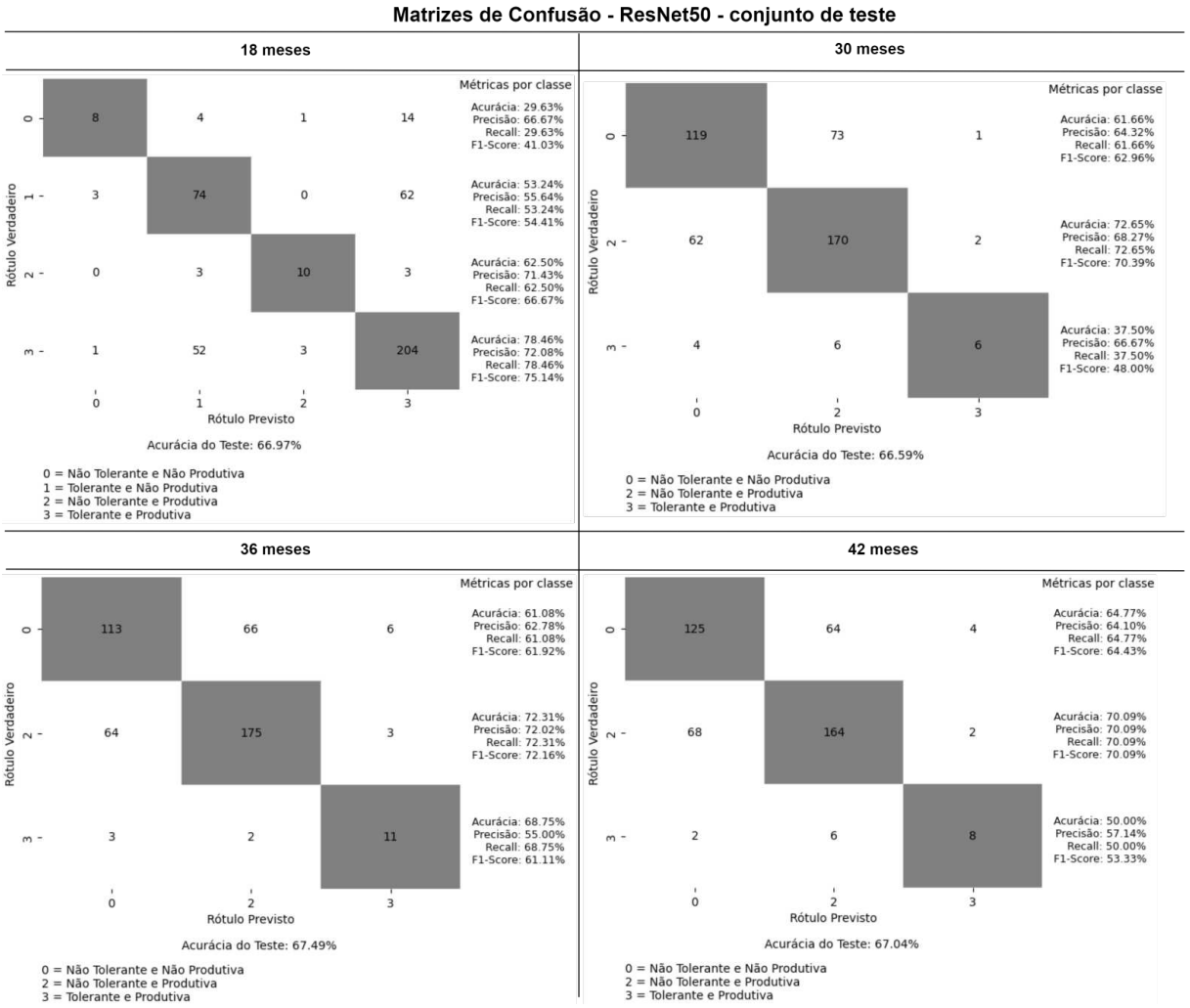

Figura 5.16: Matrizes de confusão da *ResNet50* em cada idades alvo.

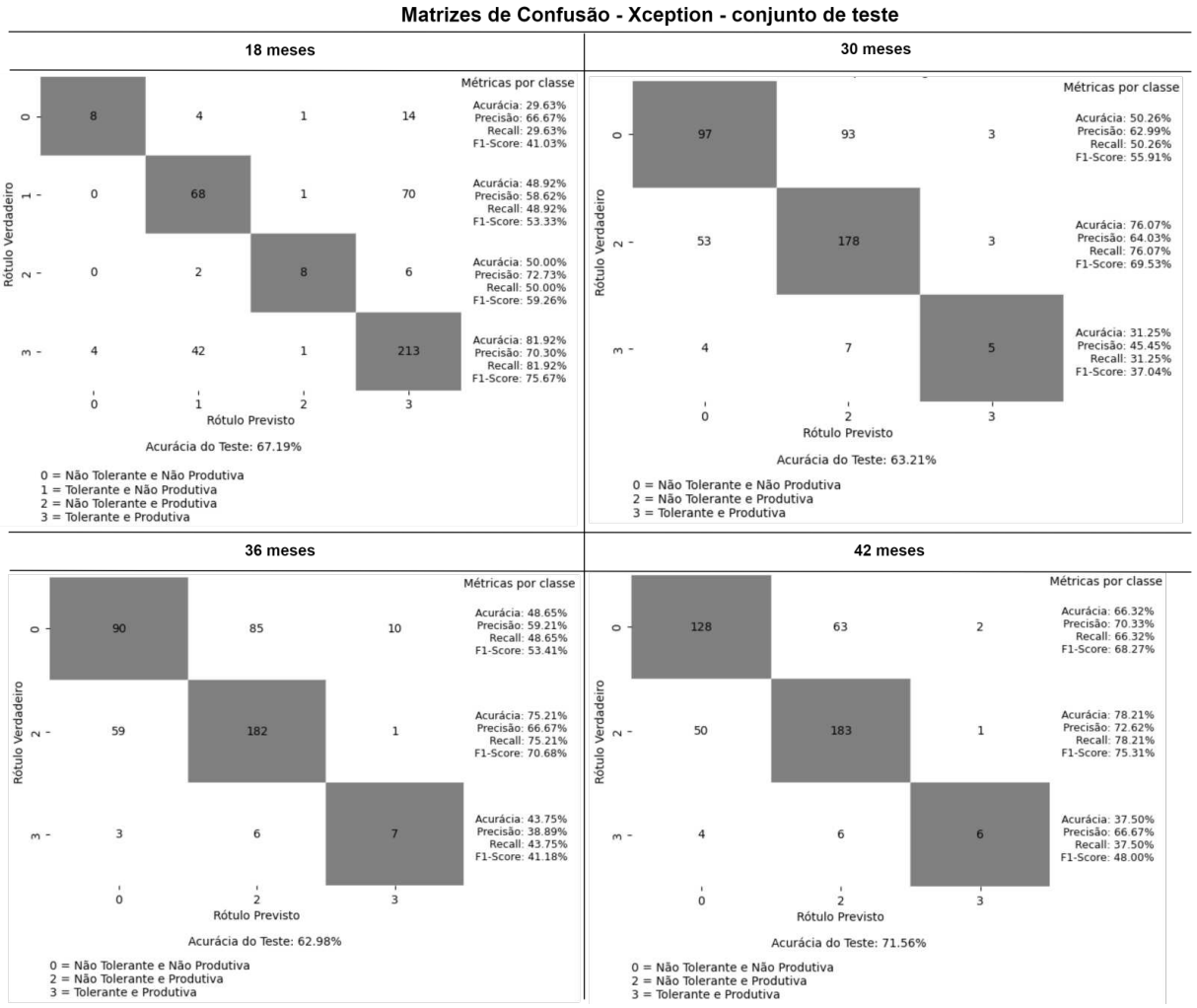

Figura 5.17: Matrizes de confusão da *Xception* em cada idades alvo.

A diagonal principal da matriz de confusão mostra os acertos do modelo, por isso está destacada. Embora o conjunto de testes esteja desbalanceado em relação à quantidade de amostras em cada classe, durante o treinamento esta diferença não foi significativa devido a aplicação do aumento de dados no conjunto de treino, o que equilibrou a quantidade de amostras das classes. No entanto, as matrizes de confusão mostram que, na maioria das vezes, as métricas de resultado foram piores para as classes em idades nas quais foram menos representadas.

Neste sentido, um resultado de exceção foi a *ResNet50* na idade alvo de 18 meses, a classe 2 teve uma quantidade menor de amostras, mas alcançou resultados melhores que a classe 0 e 1. O mesmo aconteceu com a mesma CNN aos 36 meses, com a classe 3 obtendo resultados muito próximos aos das classes 0 e 2, chegando a superar a 0 em algumas métricas. Em relação à *MobileNetV2*, apenas na idade alvo de 36 meses a classe minoritária teve resultados parecidos com as outras classes, nas demais, as classes mais representadas alcançaram desempenho melhor.

Outro resultado que chama a atenção são as métricas para a classe 2, em pratica-

mente todos os cenários experimentais, com exceção da rede *Xception* na idade alvo de 18 meses, os modelos tiveram melhor desempenho para classificar a classe 2. Isso pode indicar que os modelos tiveram mais facilidade em identificar padrões nas folhas relacionados a plantas produtivas, mesmo quando a classe 2 foi menos representada, que ocorreu na idade alvo de 18 meses na *Xception* e *ResNet50*.

As matrizes de confusão proporcionaram *insights* valiosos sobre o desempenho dos modelos de CNN no processo de classificação das imagens das folhas de eucalipto em relação à tolerância à seca e produtividade. Os resultados indicam desafios associados à representatividade das classes em diferentes idades-alvo, afetando as métricas de desempenho dos modelos. Apesar das variações e desafios identificados, a capacidade dos modelos em identificar padrões relevantes nas imagens das folhas, principalmente relacionados a plantas produtivas, sugere a promissora aplicabilidade dos modelos ajustados nesse contexto. No entanto, os resultados das matrizes reforçam a necessidade de enriquecer o conjunto de dados com uma variedade maior de imagens como uma estratégia para melhorar a capacidade de generalização dos modelos, permitindo uma classificação mais precisa e robusta das folhas de eucalipto.

5.2.2 Ativações da rede

As redes neurais têm a característica de serem modelos "caixa-preta", uma vez que seus resultados não são de fácil explicação. Uma abordagem para interpretar os resultados da CNN é visualizar as regiões da imagem que mais chamaram a atenção da rede. Neste contexto, as Figuras 5.18, 5.19 e 5.20 exibem mapas de calor para um exemplo de imagem de cada classe, para cada modelo. Eles mostram a ativação média de todos os canais em uma camada específica, com a escala de cores indicando a intensidade da ativação em diferentes regiões da imagem.

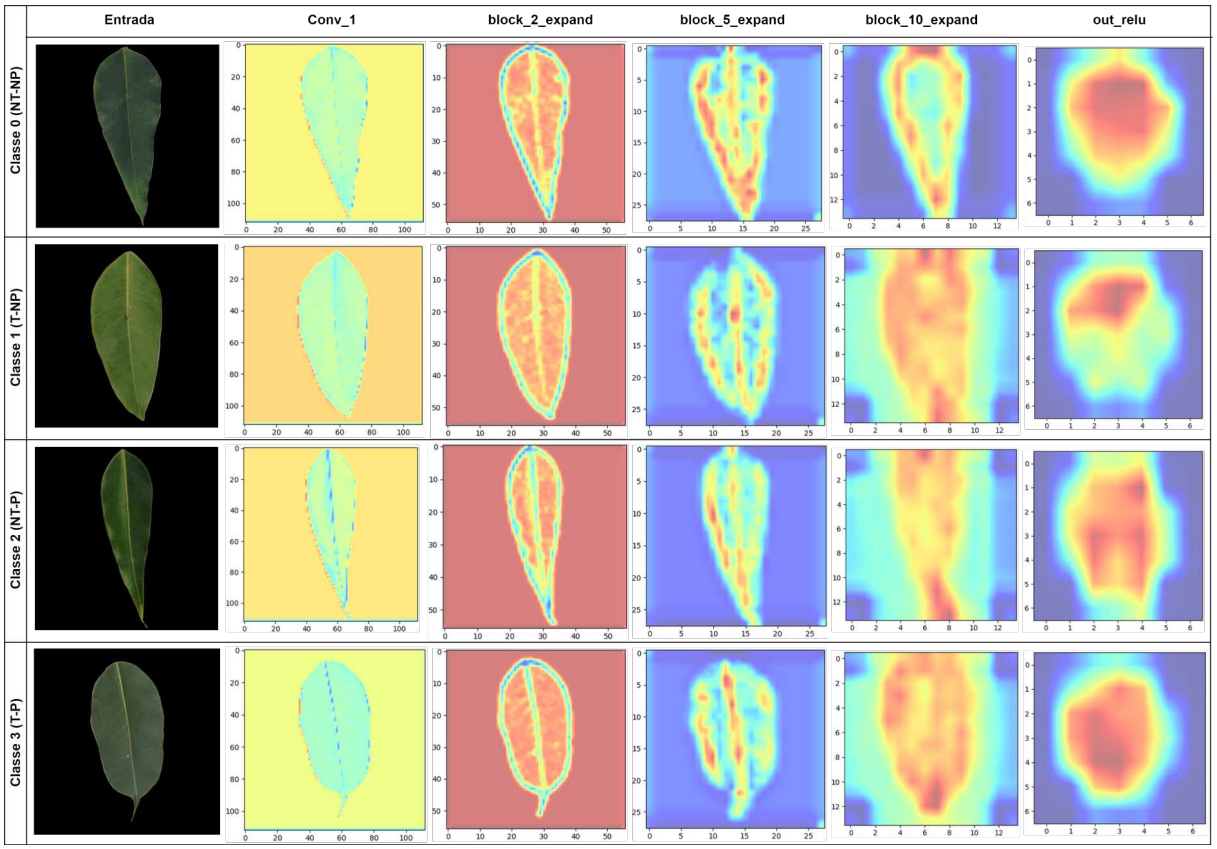

Figura 5.18: Um exemplo de imagem para cada classe com a ativação média de todos os canais de algumas camadas da rede convolucional *MobileNetV2*. Os eixos x e y das imagens representam os pixels de altura e largura, respectivamente.

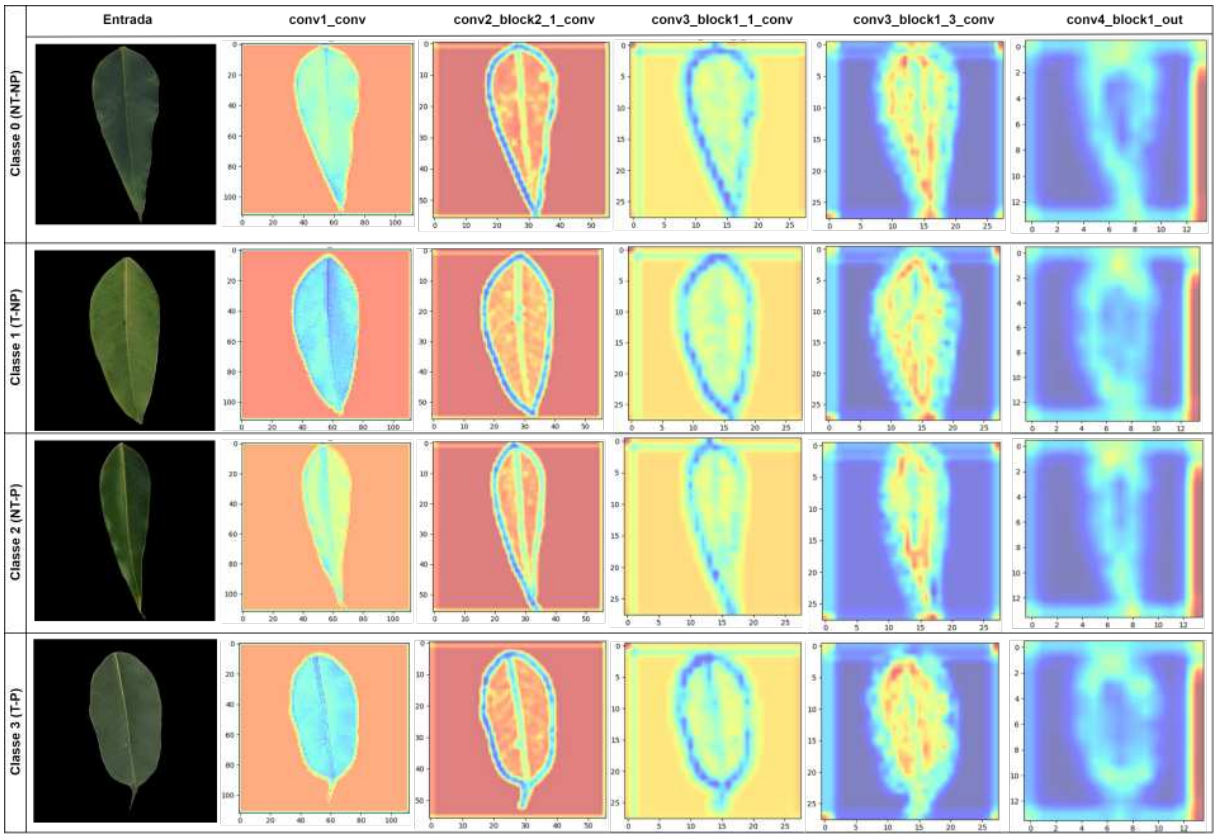

Figura 5.19: Um exemplo de imagem para cada classe com a ativação média de todos os canais de algumas camadas da rede convolucional *ResNet50*. Os eixos x e y das imagens representam os pixels de altura e largura, respectivamente.

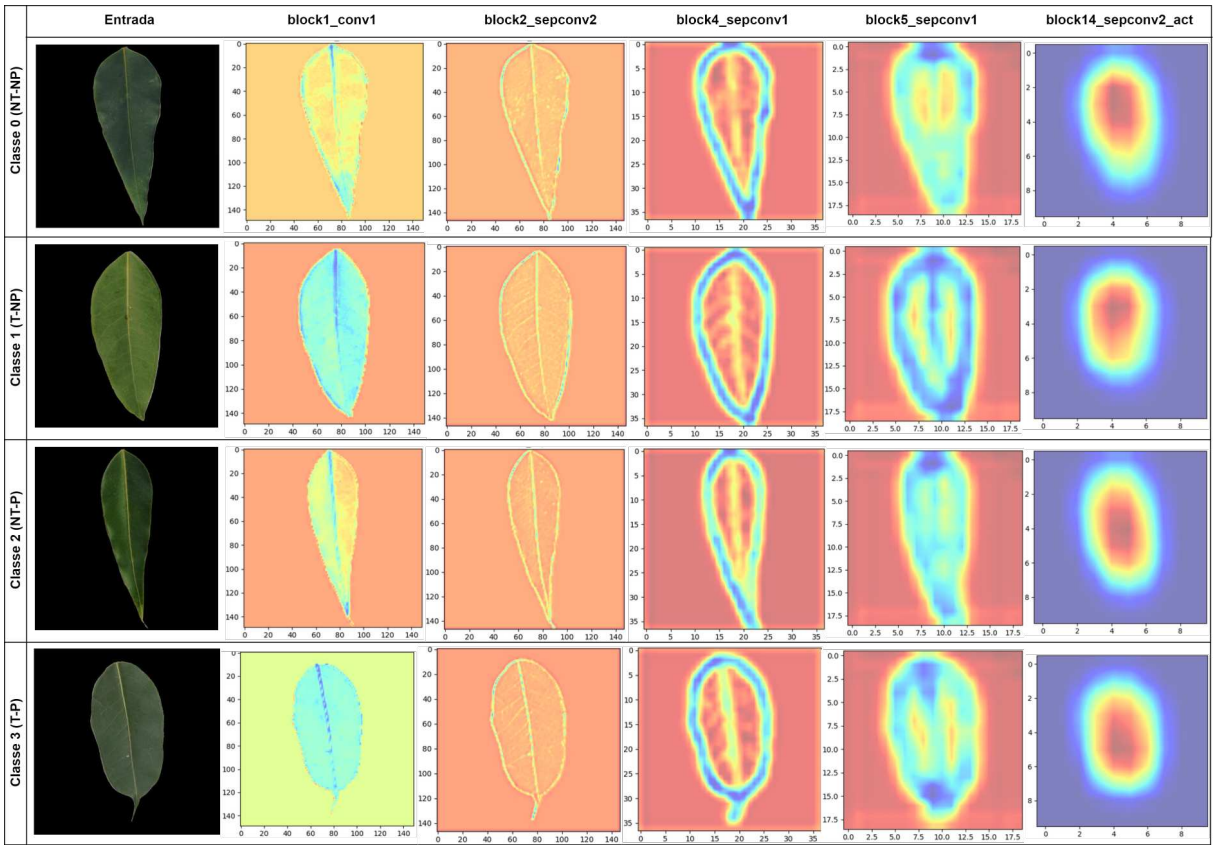

Figura 5.20: Um exemplo de imagem para cada classe com a ativação média de todos os canais de algumas camadas da rede convolucional *Xception*. Os eixos x e y das imagens representam os pixels de altura e largura, respectivamente.

À medida que se avança para camadas mais profundas, as ativações se tornam progressivamente mais abstratas e difíceis de serem interpretadas visualmente. Na última camada da rede, começam a ser codificados conceitos de nível superior, como manchas e texturas nas folhas. Essas representações mais abstratas contêm menos informações sobre o conteúdo visual da imagem e cada vez mais informações sobre a classe à qual a imagem pertence (Karlekar and Seal, 2020).

As ativações apresentadas são de algumas camadas das CNNs aplicadas. Os pontos de atenção das redes seguem basicamente o mesmo padrão, na primeira camada, as bordas das folhas foram detectadas, preservando praticamente todas as informações presentes na imagem inicial. Na camada seguinte, continuaram a ser detectadas as bordas e as estruturas internas das folhas, como as veias (venação). Até a quarta camada visualizada da rede, houve uma leve perda de informações nas bordas, mas aspectos visuais relacionados à área das folhas foram mantidos.

O fato de os modelos terem reconhecido bem a área foliar pode indicar que essa característica desempenhou um papel determinante na classificação das amostras. Isso está em consonância com a literatura, que faz diversas associações entre as alterações na área das folhas e o potencial de tolerância à seca e produtividade em euca-

lipto (Conti Junior, 2019, Oliveira, 2021, Pita-Barbosa et al., 2023, Tatagiba et al., 2007, Chaves et al., 2004, Taiz and Zeiger, 2007, Peixoto, 2020).

Capítulo 6

Conclusão e Trabalhos Futuros

Os resultados obtidos pela pesquisa abrem perspectivas importantes para a aplicação de técnicas de inteligência artificial para predição precoce do comportamento de cruzamentos de eucalipto quanto à tolerância à seca e produtividade em idade produtiva, a partir de dados de plantas joves. Principalmente pela exploração de dois tipos de dados de natureza bem distinta, o que exigiu o estudo e a aplicação de modelos adequados aos dados. Ambas abordagens de IA, modelos clássicos e as CNNs, apresentaram resultados promissores, considerando que o conjunto de dados é proveniente de um experimento de campo, onde as condições climáticas foram severas em termos de baixa disponibilidade hídrica e altas temperaturas. Além disso, foram relativamente poucas variáveis e um conjunto relativamente reduzido de amostas.

No contexto dos modelos clássicos aplicados aos dados tabulados, notamos que a predição precoce dos cruzamentos obteve resultados interessantes, ainda que os modelos revelaram uma tendência a realizar predições para as classes majoritárias, sugerindo uma influência do desbalanceamento de classes nos padrões aprendidos. Apesar disso, as análises exploratórias prévias permitiram uma compreensão mais profunda das mudanças morfofisiológicas das plantas ao longo do tempo, contribuindo para a interpretação dos resultados. Os desafios enfrentados nesta pesquisa ofereceram uma base para futuros estudos, que podem se concentrar na ampliação do conjunto de variáveis ou no desenvolvimento de abordagens mais sofisticadas para tentar melhorar a predição.

Em relação às CNNs aplicadas às imagens de folhas de eucalipto, os resultados mostraram um bom potencial preditivo, evidenciado pela acurácia de teste superior a 70% na idade mais avançada do conjunto de dados. A *Xception* apresentou um desempenho ligeiramente superior, embora as diferenças entre as redes não tenham sido significativa. No entanto, as curvas de acurácia e perda durante o treinamento sugeriram um possível sobreajuste, destacando a dificuldade dos modelos em identificar padrões distintivos nas imagens das folhas, indicando uma possível baixa representatividade do conjunto de dados. As matrizes de confusão revelaram que o tamanho das amostras por classe pode ter influenciado os resultados dos modelos.

Diante dos desafios encontrados, destacamos a importância de explorar uma maior variedade e quantidade de imagens, incluindo sensoriamento remoto, térmicas, mi-

croscopia e das raízes. A inclusão desses tipos de imagens pode ajudar a mitigar problemas de representatividade e melhorar a predição da tolerância à seca e produtividade. Além disso, expandir a análise a nível de indivíduos é uma abordagem interessante, pois permite analisar os materiais individualmente e selecionar os melhores indivíduos precocemente para clonagem. Ademais, aumentar o número de variáveis preditoras e incorporar dados futuros do projeto "Tolerância à seca em eucalipto"são perspectivas relevantes para estudos posteriores. Por fim, uma abordagem interessante para trabalhos futuros é a combinação de dados tabulados com imagens como entrada dos modelos de CNN.

Em síntese, embora os resultados não tenham atingido valores expressivos em termos de acurácia, este trabalho representa um passo importante na investigação científica para aprimorar a seleção precoce de eucaliptos tolerantes à seca e produtivos. As lições aprendidas e os desafios enfrentados neste estudo oferecem *insights* valiosos para pesquisas futuras, visando avançar na compreensão e predição do comportamento das plantas frente ao estresse hídrico.

Referências Bibliográficas

Alba et al., 2022 Alba, E., de Souza Alexandre, M. L., Marchesan, J., de Souza, L. S. B., Bezerra, A. C., and Silva, E. A. (2022). Comparação entre algoritmos de aprendizado de máquina para a identificação de floresta tropical sazonalmente seca. *Anuário do Instituto de Geociências*, 45:1–10.

Assis et al., 2005 Assis, T. F., Warburton, P., and Harwood, C. (2005). Artificially induced protogyny: an advance in the controlled pollination of eucalyptus. *Australian Forestry*.

Atila et al., 2021 Atila, Ü., Uçar, M., Akyol, K., and Uçar, E. (2021). Plant leaf disease classification using efficientnet deep learning model. *Ecological Informatics*, 61:101182.

Berrar, 2018 Berrar, D. (2018). *Cross-Validation*. Encyclopedia of Bioinformatics and Computational Biology.

Bianchi et al., 2016 Bianchi, L., Germino, G. H., and de Almeida Silva, M. (2016). Adaptação das plantas ao déficit hídrico. *Acta iguazu*, 5(4):15–32.

Bombrun et al., 2020 Bombrun, M., Dash, J. P., Pont, D., Watt, M. S., Pearse, G. D., and Dungey, H. S. (2020). Forest-scale phenotyping: Productivity characterisation through machine learning. *Frontiers in Plant Science*, 11:99.

Braga et al., 2000 Braga, A. d. P., Ludermir, T. B., and Carvalho, A. C. P. d. L. F. (2000). Redes neurais artificiais: teoria e aplicações. *LTC*.

Breiman, 2001 Breiman, L. (2001). Random forests. *Machine Learning*.

Brodersen et al., 2010 Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*, pages 3121–3124. IEEE.

Carignato et al., 2019 Carignato, A., Vázquez-Piqué, J., Tapias, R., Ruiz, F., and Fernández, M. (2019). Variability and plasticity in cuticular transpiration and leaf permeability allow differentiation of eucalyptus clones at an early age. *Forests*, 11(9).

Castro et al., 2021 Castro, C. A. O., dos Santos, G. A., Takahashi, E. K., Nunes, A. C. P., Souza, G. A., and de Resende, M. D. (2021). Accelerating eucalyptus breeding strategies through top grafting applied to young seedlings. *Industrial Crops and Products*.

Castro and Braga, 2011a Castro, C. L. d. and Braga, A. P. (2011a). Aprendizado supervisionado com conjuntos de dados desbalanceados. *Sba: Controle Automação Sociedade Brasileira de Automatica*, 22(5):441–466.

Castro and Braga, 2011b Castro, C. L. d. and Braga, A. P. (2011b). Aprendizado supervisionado com conjuntos de dados desbalanceados. *Sba: Controle & Automação Sociedade Brasileira de Automatica*, 22:441–466.

Chaves et al., 2004 Chaves, J. H., Reis, G. G. d., Reis, M. d. G. F., Neves, J. C. L., Pezzopane, J. E. M., and Polli, H. Q. (2004). Seleção precoce de clones de eucalipto para ambientes com disponibilidade diferenciada de água no solo: relações hídricas de plantas em tubetes. *Revista Árvore*, 28(3):333–341.

Chen and Guestrin, 2016 Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Chollet, 2017 Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions.

Chollet et al., 2015 Chollet, F. et al. (2015). keras.

Coelho Eugenio et al., 2021 Coelho Eugenio, F., Badin, T. L., Fernandes, P., Mallmann, C. L., Schons, C., Schuh, M. S., Soares Pereira, R., Fantinel, R. A., and Pereira da Silva, S. D. (2021). Remotely piloted aircraft systems (rpas) and machine learning: A review in the context of forest science. *International Journal of Remote Sensing*, 42(21):8207–8235.

Conti Junior, 2019 Conti Junior, J. L. F. (2019). Parâmetros fisiológicos como indicadores de tolerância à seca em clones de eucalyptus spp. *Universidade Estadual Paulista (Unesp)*.

Cordeiro et al., 2022 Cordeiro, M. A., Arce, J. E., Guimarães, F. A. R., Bonete, I. P., Silva, A. V. d. S., Abreu, J. C. d., and Binoti, D. H. B. (2022). Estimativas volumétricas em povoamentos de eucalipto utilizando máquinas de vetores de suporte e redes neurais artificiais. *Madera y bosques*, 28(1).

Corrêa et al., 2017 Corrêa, T. R., de Toledo Picoli, E. A., de Souza, G. A., Conde, S. A., Silva, N. M., Lopes-Mattos, K. L. B., ..., and Oda, S. (2017). Phenotypic markers in early selection for tolerance to dieback in eucalyptus. *Industrial Crops and Products*.

Correia et al., 2016 Correia, B., Valledor, L., Hancock, R. D., Renaut, J., Pascual, J., Soares, A. M., and Pinto, G. (2016). Integrated proteomics and metabolomics to unlock global and clonal responses of eucalyptus globulus recovery from water deficit. *Metabolomics*, 12:1–12.

Cortes and Vapnik, 1995 Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20:273–297.

Csillik et al., 2019 Csillik, O., Kumar, P., Mascaro, J., et al. (2019). Monitoring tropical forest carbon stocks and emissions using planet satellite data. *Scientific Reports*, 9(1):17831.

da Silva Tavares Júnior et al., 2020 da Silva Tavares Júnior, I., Torres, C. M. M. E., Leite, H. G., de Castro, N. L. M., Soares, C. P. B., Castro, R. V. O., and Farias, A. A. (2020). Machine learning: Modeling increment in diameter of individual trees on atlantic forest fragments. *Ecological Indicators*, 117:106685.

dos Santos et al., 2022 dos Santos, A., Biesseck, B. J. G., Latte, N., de Lima Santos, I. C., dos Santos, W. P., Zanetti, R., and Zanuncio, J. C. (2022). Remote detection and measurement of leaf-cutting ant nests using deep learning and an unmanned aerial vehicle. *Computers and Electronics in Agriculture*, 198:107071.

El Naqa and Murphy, 2015 El Naqa, I. and Murphy, M. J. (2015). *What Is Machine Learning?*, pages 3–11. Springer International Publishing.

Few, 2021 Few, S. (2021). *Now You See It: An Introduction to Visual Data Sensemaking*. Analytics Press, Oakland, CA.

Florêncio et al., 2022 Florêncio, G. W. L., Martins, F. B., and Fagundes, F. F. A. (2022). Climate change on eucalyptus plantations and adaptive measures for sustainable forestry development across brazil. *Industrial Crops and Products*, 188:115538.

García-Gutiérrez et al., 2016 García-Gutiérrez, J., González-Ferreiro, E., Mateos-García, D., and Riquelme-Santos, J. C. (2016). A preliminary study of the suitability of deep learning to improve lidar-derived biomass estimation. In *Hybrid Artificial Intelligent Systems: 11th International Conference, HAIS 2016, Seville, Spain, April 18-20, 2016, Proceedings 11*, pages 588–596. Springer.

García Laencina et al., 2010 García Laencina, P., Morales-Sánchez, J., Verdú-Monedero, R., Larrey-Ruiz, J., Sancho-Gómez, J. L., and Figueiras-Vidal, A. (2010). Classification with missing data. *Neural Computing and Applications*.

Garuzzo, 2022 Garuzzo, M. d. S. P. B. (2022). Análise dialélica e seleção genética de híbridos de eucalyptus com potencial de tolerância à seca. *UFV*.

Géron, 2019 Géron, A. (2019). *Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow*. Alta Books.

Ghafarian et al., 2022 Ghafarian, F., Wieland, R., Lüttschwager, D., and Nendel, C. (2022). Application of extreme gradient boosting and shapley additive explanations to predict temperature regimes inside forests from standard open-field meteorological data. *Environmental Modelling Software*, 156:105466.

Gonçalves et al., 2017 Gonçalves, J. L., Alvares, C. A., Rocha, J. H., Brandani, C. B., and Hakamada, R. (2017). Eucalypt plantation management in regions with water stress. *Southern Forests: a Journal of Forest Science*.

Goodfellow et al., 2016 Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Guan et al., 2015 Guan, H., Yu, Y., Ji, Z., Li, J., and Zhang, Q. (2015). Deep learning-based tree classification using mobile lidar data. *Remote Sensing Letters*, 6(11):864–873.

Hamedianfar et al., 2022 Hamedianfar, A., Mohamedou, C., Kangas, A., and Vauhkonen, J. (2022). Deep learning for forest inventory and planning: a critical review on the remote sensing approaches so far and prospects for further applications. *Forestry: An International Journal of Forest Research*, 95(4):451–465.

Hamraz et al., 2019 Hamraz, H., Jacobs, N. B., Contreras, M. A., and Clark, C. H. (2019). Deep learning for conifer/deciduous classification of airborne lidar 3d point clouds representing individual trees. *ISPRS Journal of Photogrammetry and Remote Sensing*, 158:219–230.

He et al., 2015 He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition.

He et al., 2016 He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hossin and Sulaiman, 2015 Hossin, M. and Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1.

Hu et al., 2020 Hu, G., Yin, C., Wan, M., Zhang, Y., and Fang, Y. (2020). Recognition of diseased pinus trees in uav images using deep learning and adaboost classifier. *Biosystems Engineering*, 194:138–151.

Huang et al., 2020 Huang, L., Qin, J., Zhou, Y., Zhu, F., Liu, L., and Shao, L. (2020). Normalization techniques in training dnns: Methodology, analysis and application. *arXiv preprint arXiv:2009.12836*.

IBÁ, 2021 IBÁ (2021). *Relatório Anual IBÁ - Indústria Brasileira de Árvores, 2021*. Revista da Indústria Brasileira de Árvores.

ImageNet, 2016 ImageNet (2016). About ImageNet. <http://www.image-net.org/about-overview>.

Ismail and Mutanga, 2010 Ismail, R. and Mutanga, O. (2010). A comparison of regression tree ensembles: Predicting sirex noctilio induced water stress in pinus patula forests of kwazulu-natal, south africa. *International Journal of Applied Earth Observation and Geoinformation*, 12:S45–S51.

Jain et al., 2020 Jain, P., Coogan, S. C., Subramanian, S. G., Crowley, M., Taylor, S., and Flannigan, M. D. (2020). A review of machine learning applications in wildfire science and management. *Environmental Reviews*, 28(4):478–505.

Jolliffe and Cadima, 2016 Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202.

Käding et al., 2017 Käding, C., Rodner, E., Freytag, A., and Denzler, J. (2017). Fine-tuning deep neural networks in continuous learning scenarios. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part III 13*, pages 588–605. Springer.

Karlekar and Seal, 2020 Karlekar, A. and Seal, A. (2020). Soynet: Soybean leaf diseases classification. *Computers and Electronics in Agriculture*, 172:105342.

Kaya et al., 2019 Kaya, A., Keceli, A. S., Catal, C., Yalic, H. Y., Temucin, H., and Tekinerdogan, B. (2019). Analysis of transfer learning for deep neural network based plant classification models. *Computers and Electronics in Agriculture*, 158:20–29.

KLIPPEL et al., 2014 KLIPPEL, V. H., PEZZOPANE, J. E. M., PEZZOPANE, J. R. M., and TOLEDO, J. V. (2014). Impacto da deficiência hídrica no crescimento inicial de eucalipto. *Revista Científica Eletrônica de Engenharia Florestal*, 23(1):48–59.

Krizhevsky et al., 2012 Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Larcher, 2006 Larcher, W. (2006). *Ecofisiologia Vegetal*. Rima, São Carlos, 3ª ed. edition.

LeCun et al., 2015 LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Li et al., 2020 Li, Y., Li, M., Li, C., and et al. (2020). Forest aboveground biomass estimation using landsat 8 and sentinel-1a data with machine learning algorithms. *Sci Rep*, 10:9952.

Li et al., 2022 Li, Y., Wang, R., Shi, W., Yu, Q., Li, X., and Chen, X. (2022). Research on accurate estimation method of eucalyptus biomass based on airborne lidar data and aerial images. *Sustainability*, 14(17).

Liao et al., 2022 Liao, K., Yang, F., Dang, H., Wu, Y., Luo, K., and Li, G. (2022). Detection of eucalyptus leaf disease with uav multispectral imagery. *Forests*, 13(8):1322.

Liu et al., 2019 Liu, J., Wang, X., and Wang, T. (2019). Classification of tree species and stock volume estimation in ground forest images using deep learning. *Computers and Electronics in Agriculture*, 166:105012.

Liu et al., 2018 Liu, Z., Peng, C., Work, T., Candau, J.-N., DesRochers, A., and Kneeshaw, D. (2018). Application of machine-learning methods in forest ecology: recent progress and future challenges. *Environmental Reviews*, 26(4):339–350.

Lones, 2021 Lones, M. A. (2021). How to avoid machine learning pitfalls: a guide for academic researchers. *arXiv preprint arXiv:2108.02497*.

Lopes et al., 2023 Lopes, A. R., Lira, J. M. S., Oliveira, L. A., Garuzzo, M. d. S. P. B., de Sá Barbalho, M. V., de Araújo, P. O. C., dos Santos, G. A., and Nacif, J. A. (2023). Predição do incremento médio anual volumétrico de eucalyptus com aprendizado de máquina. In *Anais do XIV Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais*, pages 81–90. SBC.

Low and Choo, 2018 Low, J. X. and Choo, K. W. (2018). Classification of heart sounds using softmax regression and convolutional neural network. In *Proceedings of the 2018 International Conference on Communication Engineering and Technology*, pages 18–21.

Martins, 2007 Martins, F. B. (2007). Desenvolvimento e estresse hídrico em mudas de *Eucalyptus grandis* (hill ex maiden) e *Eucalyptus saligna* (smith). Master's thesis, Universidade Federal de Santa Maria, Santa Maria. Dissertação de Mestrado, Programa de Pós-graduação em Engenharia Agrícola.

Maseda and Fernández, 2016 Maseda, P. H. and Fernández, R. J. (2016). Growth potential limits drought morphological plasticity in seedlings from six eucalyptus provenances. *Tree physiology*, 36(2):243–251.

Massaro, 2008 Massaro, R. A. M. (2008). Viabilidade de aplicação da seleção precoce e tamanho de parcelas em testes clonais de eucalyptus spp. *Universidade Estadual Paulista*.

Mittler, 2006 Mittler, R. (2006). Abiotic stress, the field environment and stress combination. *Trends in plant science*, 11(1):15–19.

Moraes et al., 2014 Moraes, C. B. d., de Freitas, M., Casella, T., Pieroni, G. B., Vilela de Resende, M. D., Zimbacks, L., and Mori, E. S. (2014). Estimativas de parâmetros genéticos para seleção precoce de clones de eucalyptus para região com ocorrência de geadas. *Scientia Forestalis*, pages 219–227.

Mulkey et al., 2012 Mulkey, S. S., Chazdon, R. L., and Smith, A. P. (2012). *Tropical forest plant ecophysiology*. Springer Science & Business Media.

Myburg et al., 2014 Myburg, A. A., Grattapaglia, D., Tuskan, G. A., Hellsten, U., Hayes, R. D., Grimwood, J., and Goodstein, D. M. (2014). The genome of eucalyptus grandis. *Nature*.

Nadif et al., 2021 Nadif, M., El Abbassi, N., and Tairi, H. (2021). Analysis of data mining techniques for online advertising. *Journal of Theoretical and Applied Information Technology*.

Narayana and Ramana, 2022 Narayana, C. L. and Ramana, K. V. (2022). Plant leaf classification using fine-tuned transfer learning. *Mathematical Statistician and Engineering Applications*, 71(4):6051–6070.

Nogueira and Silva, 2002 Nogueira, R. J. M. C. and Silva, E. C. d. (2002). Comportamento estomático em plantas jovens de schinopsis brasiliensis engl. cultivadas sob estresse hídrico. *Iheringia, Série Botânica*, 57(1):31–38.

Nunes and Görgens, 2016 Nunes, M. H. and Görgens, E. B. (2016). Artificial intelligence procedures for tree taper estimation within a complex vegetation mosaic in brazil. *PLOS ONE*, 11:1–16.

Oliveira, 2021 Oliveira, F. S. (2021). Aspectos morfoanatômicos e metabólicos envolvidos na tolerância à seca em eucalipto. *Universidade Federal de Viçosa*.

Paludeto et al., 2017 Paludeto, J. G. Z., Estopa, R. A., and Tambarussi, E. V. (2017). Eficiência da seleção precoce em clones de eucalyptus grandis w. hill ex maiden x eucalyptus urophylla st blake. *Revista do Instituto Florestal*, 29(2):169–179.

Pedregosa et al., 2011 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*.

Peixoto, 2020 Peixoto, C. P. (2020). *Princípios de Fisiologia Vegetal: Teoria e Prática*. Editora Pod, Rio de Janeiro, RJ, 1 edition.

Pita and Pardos, 2001 Pita, P. and Pardos, J. A. (2001). Growth, leaf morphology, water use and tissue water relations of eucalyptus globulus clones in response to water deficit. *Tree Physiology*, 21(9):599–607.

Pita-Barbosa et al., 2023 Pita-Barbosa, A., Oliveira, L. A., de Barros, N. F., Hodecker, B. E. R., Oliveira, F. S., Araújo, W. L., and Martins, S. C. (2023). Developing a roadmap to define a potential ideotype for drought tolerance in eucalyptus. *Forest Science*, 69(1):101–114.

Pritzkow et al., 2020 Pritzkow, C., Szota, C., Williamson, V. G., and Arndt, S. K. (2020). Phenotypic plasticity of drought tolerance traits in a widespread eucalypt (eucalyptus obliqua). *Forests*, 11(12):1371.

Python Software Foundation, 2019 Python Software Foundation (2019). Python programming language (version 3.8). <https://www.python.org/downloads/release/python-380/>. Accessed on March 2, 2023.

Quinlan, 1986 Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1:81–106.

Redmon et al., 2016 Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.

Reis et al., 2015 Reis, C. A. F., Gonçalves, F. M. A., Ramalho, M. A. P., and Rosado, A. M. (2015). Estratégias na seleção simultânea de vários caracteres no melhoramento do eucalyptus. *Ciência Florestal*, 25(2):457–467.

Resende et al., 2010 Resende, M. D. V. d., Lopes, P. S., Silva, R. L. d., and Pires, I. E. (2010). Seleção genômica ampla (gws) e maximização da eficiência do melhoramento genético. *Pesquisa Florestal Brasileira*, 56(56):63.

Rodriguez et al., 1997 Rodriguez, L. C. E., Bueno, A. R. S., and Rodrigues, F. (1997). Rotações de eucaliptos mais longas: análise volumétrica e econômica. *Scientia forestalis*, 51(1):15–28.

Rosenblatt, 1958 Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

Samuel, 1959 Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229.

Sandler et al., 2018 Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520.

Shukla et al., 2019 Shukla, P. R., Skea, J., Calvo Buendia, E., Masson-Delmotte, V., Pörtner, H. O., Roberts, D. C., Zhai, P., Slade, R., Connors, S., and van Diemen, R. (2019). *Climate Change and Land: An IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems*. Intergovernmental Panel on Climate Change (IPCC).

Silva et al., 2017 Silva, M., Rubilar, R., Espinoza, J., Yánez, M., Emhart, V., and Quiroga, J. J. (2017). Respuesta en parámetros de intercambio gaseoso y supervivencia en plantas jóvenes de genotipos comerciales de eucalyptus spp sometidas a déficit hídrico. *Bosque (valdivia)*, 38(1):79–87.

Simonyan and Zisserman, 2014 Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Smola and Schölkopf, 2004 Smola, A. J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*.

Su et al., 2018 Su, C., Wu, X., Tang, X., and Hu, J. (2018). Growth height prediction for the trees under overhead lines based on deep learning algorithm. In *2018 International Conference on Power System Technology (POWERCON)*, pages 3693–3699. IEEE.

Sutton and Barto, 1998 Sutton, R. S. and Barto, A. G. (1998). The reinforcement learning problem. In *Reinforcement learning: An introduction*, pages 51–85. MIT Press Cambridge, MA.

Szegedy et al., 2015 Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Taiz and Zeiger, 2007 Taiz, L. and Zeiger, E. (2007). *Fisiologia vegetal*, volume 10. Universitat Jaume I.

Taiz et al., 2015 Taiz, L., Zeiger, E., Møller, I. M., Murphy, A., et al. (2015). *Plant physiology and development.* Sinauer Associates Incorporated.

Talebiesfandarani and Shamsoddini, 2022 Talebiesfandarani, S. and Shamsoddini, A. (2022). Global-scale biomass estimation based on machine learning and deep learning methods. *Remote Sensing Applications: Society and Environment*, 28:100868.

Tambarussi et al., 2017 Tambarussi, E. V., De Lima, B. M., Da Costa Queiroz, R., Peres, F. S. B., Da Costa Dias, D., Pagliarini, M. K., Pereira, F. B., Rosa, J. R. B. F., and Peçanha Rezende, G. D. S. (2017). Estimativas de parâmetros genéticos para a seleção precoce em clones de eucalyptus spp. *Scientia Forestalis/Forest Sciences*.

Tatagiba et al., 2007 Tatagiba, S. D., Pezzopane, J. E. M., and dos Reis, E. F. (2007). Avaliação do crescimento e produção de clones de eucalyptus submetidos a diferentes manejos de irrigação. *Cerne*, 13(1):1–9.

Taud and Mas, 2018 Taud, H. and Mas, J. (2018). *Multilayer Perceptron (MLP)*, chapter 27, pages 451–455. Springer International Publishing.

Tetila et al., 2020 Tetila, E. C., Machado, B. B., Astolfi, G., de Souza Belete, N. A., Amorim, W. P., Roel, A. R., and Pistori, H. (2020). Detection and classification of soybean pests using deep learning with uav images. *Computers and Electronics in Agriculture*, 179:105836.

VanRaden, 2008 VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of dairy science*, 91(11):4414–4423.

Vellini et al., 2008 Vellini, A. L. T. T., Paula, N. F. d., Alves, P. L. d. C. A., Pavani, L. C., Bonine, C. A. V., Scarpinati, E. A., and Paula, R. C. d. (2008). Respostas fisiológicas de diferentes clones de eucalipto sob diferentes regimes de irrigação. *Revista Árvore*, 32(4):651–663.

Weinstein et al., 2019 Weinstein, B. G., Marconi, S., Bohlman, S., Zare, A., and White, E. (2019). Individual tree-crown detection in rgb imagery using semi-supervised deep learning neural networks. *Remote Sensing*, 11(11).

Welle et al., 2022 Welle, T., Aschenbrenner, L., Kuonath, K., Kirmaier, S., and Franke, J. (2022). Mapping dominant tree species of german forests. *Remote Sensing*, 14(14).

Werbos, 1974 Werbos, P. J. (1974). *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. PhD thesis, Harvard University.

Witten et al., 2016 Witten, I. H., Frank, E., and Hall, M. A. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers.

Xavier et al., 2011 Xavier, T. M. T., Pezzopane, J. E. M., Penchel, R. M., Caldeira, M. V. W., Klippel, V. H., Toledo, J. V., and Silva, M. R. (2011). Crescimento do eucalipto sob diferentes níveis de déficit hídrico. In *CONGRESSO BRASILEIRO DE FISIOLOGIA, Anais*, Búzios-RJ.

Xu et al., 2018 Xu, P., Zhou, T., Yi, C., Fang, W., Hendrey, G., and Zhao, X. (2018). Forest drought resistance distinguished by canopy height. *Environmental Research Letters*, 13(7):075003.

Yakubu, 2010 Yakubu, A. (2010). Fixing multicollinearity instability in the prediction of body weight from morphometric traits of white fulani cows. *Journal of Central European Agriculture*.

Zaiton et al., 2020 Zaiton, S., Sheriza, M. R., Ainishifaa, R., Alfred, K., and Norfaryanti, K. (2020). Eucalyptus in malaysia: Review on environmental impacts. *Journal of Landscape Ecology*.

Zhang et al., 2019 Zhang, L., Shao, Z., Liu, J., and Cheng, Q. (2019). Deep learning based retrieval of forest aboveground biomass from combined lidar and landsat 8 data. *Remote Sensing*, 11(12):1459.

Łoś et al., 2021 Łoś, H., Mendes, G. S., Cordeiro, D., Grosso, N., Costa, H., Benevides, P., and Caetano, M. (2021). Evaluation of xgboost and lgbm performance in tree species classification with sentinel-2 data. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 5803–5806.