UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Diagnóstico precoce de ataque de pragas em plantas usando imagens de fluorescência

Luiz Gonzaga da Silva Junior

Dissertação de Mestrado do Programa de Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria (MECAI)

ICMC USP
SÃO CARLOS

Luiz Gonzaga da Silva Junior

Diagnóstico precoce de ataque de pragas em plantas usando imagens de fluorescência

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria. *EXEMPLAR DE DEFESA*

Área de Concentração: Matemática, Estatística e Computação

Orientador: Prof. Dr. Paulino Ribeiro Villas-Boas

USP – São Carlos
Novembro de 2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

G642d

Gonzaga da Silva Junior, Luiz
Diagnóstico precoce de ataque de pragas em plantas usando imagens de fluorescência / Luiz Gonzaga da Silva Junior; orientador Paulino Ribeiro Villas-Boas. -- São Carlos, 2023.
95 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. Infestação. 2. Identificação precoce. 3. Aprendizado de máquina. 4. Espectroscopia de fluorescência. 5. Data augmentation. I. Ribeiro Villas-Boas, Paulino, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Luiz Gonzaga da Silva Junior

Early diagnosis of pest attack on plants with fluorescence spectroscopy data

Dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP – in accordance with the requirements of the Professional Master's Program in Mathematics Statistics and Computing Applied to Industry, for the degree of Master in Science. *EXAMINATION BOARD PRESENTATION COPY*

Concentration Area: Mathematics, Statistics and Computing

Advisor: Prof. Dr. Paulino Ribeiro Villas-Boas

USP – São Carlos
November 2023

Este trabalho é dedicado à memória dos meus pais, que, mesmo com toda simplicidade, empenharam-se para garantir uma boa educação, contribuindo para a formação da pessoa que sou hoje.

AGRADECIMENTOS

Gostaria de expressar meus agradecimentos aos meus familiares e amigos que me apoiaram nessa jornada. Em especial, quero reconhecer o apoio das minhas irmãs **Luciana Souza da Silva** e **Luana da Silva Vicentini** e meu amigo **Anderson Henrique Rodrigues Ferreira**, que estiveram ao meu lado desde o começo.

A minha companheira de vida, **Bruna Roque Loureiro**, por sua constante presença ao meu lado, sempre incentivando e oferecendo apoio incondicional ao longo desta jornada.

Ao meu orientador, **Prof. Dr. Paulino Ribeiro Villas-Boas**, pela orientação impecável e apoio constante. Sua experiência interdisciplinar foi fundamental na construção de insights valiosos, contribuindo significativamente para o desenvolvimento e conclusão desse trabalho.

A minha colega de trabalho, **Bianca Batista Barreto**, pela condução dos experimentos que deram origem aos dados analisado nesse trabalho, pelas sugestões e criticas que foram essenciais para aprimorar o trabalho.

A **Embrapa São Carlos** por disponibilizar a estrutura e os equipamentos necessários para aquisição dos dados nessa pesquisa.

Aos **professores do ICMC**, tive a oportunidade de absorver um pouco do valioso conhecimento que compartilharam.

Aos **funcionários do ICMC**, agradeço pela assistência nos trâmites quando foi necessário. Sua colaboração foi fundamental.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

“A mais profunda emoção que podemos experimentar
é inspirada pelo senso de mistério”
(Albert Einstein)

RESUMO

JUNIOR, L. G. **Diagnóstico precoce de ataque de pragas em plantas usando imagens de fluorescência**. 2023. 95 p. Dissertação (Mestrado – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

Um dos maiores desafios na agropecuária em grande escala é o controle e monitoramento de doenças e pragas que acometem as plantações. Sem o controle adequado, podem comprometer a produtividade gerando grandes prejuízos. O monitoramento normalmente é feito de forma visual, um processo ineficiente e propenso a falhas. Para contornar o problema da baixa eficiência no monitoriamento e evitar a disseminação descontrolado de doenças, o produtores costumam realizar aplicações regulares de defensivos químicos. No entanto, quando a inspeção visual detecta uma infestação, os danos causados à plantação frequentemente são elevados, deixando os produtores com opções limitadas de manobra. Além disso, a aplicação regular de defensivos pode acarretar em outros problemas, tais como o desenvolvimento de resistência em pragas e doenças, impactos na saúde humana e contaminação ambiental. Este trabalho tem como objetivo desenvolver modelos de aprendizado de máquina para classificação precoce de pragas no milho usando dados de imagens de fluorescência da clorofila. Para atingir esse objectivo, utilizamos técnicas de " data augmentation" para expandir o conjunto de dados inicial. Essa abordagem permitiu uma representação mais abrangente dos atributos, aumentando a capacidade de generalização dos modelos. Plantas de milho das variedade Zapalotes chico (LE) e Sintético Spodoptera (SE) foram cultivadas em vasos e preservadas em casa de cultivo até o momento da infestação com as pragas e coleta dos dados. As imagens de fluorecência foram obtidas através do equipamento Closed FluorCam FC800-C e processadas pelo programa FlourCam7 para extração dos atributos. Foram avaliadas dois tipos de infestação, ataque inicial de Spodoptera frugiperda (lagarta) e Dichelops melacanthus (percevejo). Para identificar qual viés de representação mais adequado para o conjunto de dados, exploramos quatro métodos de classificação: baseados em distâncias, como o KNN; métodos simbólicos, exemplificados pela Árvore de Decisão; métodos conexionistas, como Redes Neurais; e métodos de maximização de margens, como o SVM. As redes neurais e o Adaboost demonstraram os melhores desempenhos na classificação, alcançando uma acurácia de 83% na detecção de percevejos e 75% na detecção de lagartas, respectivamente. Este estudo evidenciou o potencial transformador ao integrar dados reais e sintéticos no treinamento de modelos de aprendizado de máquina, resultando em melhorias significativas na identificação precoce de pragas no cultivo de milho.

Palavras-chave: Infestação, Identificação precoce, Aprendizado de máquina, Espectroscopia de fluorescência, Data augmentation.

ABSTRACT

JUNIOR, L. G. **Early diagnosis of pest attack on plants with fluorescence spectroscopy data**. 2023. 95 p. Dissertação (Mestrado – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

One of the major challenges in large-scale agriculture is the control and monitoring of diseases and pests affecting crops. Without proper control, they can compromise productivity, leading to significant losses. Typically, diagnosis is done visually, an inefficient and error-prone process. To prevent uncontrolled disease spread due to inefficient monitoring, farmers often resort to regular applications of chemical pesticides. However, when visual inspection detects an infestation, the resulting damage to the crops is often extensive, leaving farmers with limited options. Moreover, regular pesticide use can lead to other issues such as pest and disease resistance, impacts on human health, and environmental contamination. This study aims to develop machine learning models for early pest classification in corn using chlorophyll fluorescence imaging data. To achieve this, we employed data augmentation techniques to expand the initial dataset, allowing a more comprehensive representation of attributes and enhancing the models' generalization capability. Corn plants from the Zapalotes chico (LE) and Synthetic Spodoptera (SE) cultivars were grown in pots, maintained in a greenhouse until infestation with pests, and data collection. Fluorescence images were captured using the Closed FluorCam FC800-C equipment and processed by the FlourCam7 program for attribute extraction. Two types of infestation were evaluated: the initial attack of Spodoptera frugiperda (caterpillar) and Dichelops melacanthus (bug). To determine the most suitable bias for dataset representation, we explored four classification methods: distance-based methods like KNN, symbolic methods exemplified by Decision Trees, connectionist methods such as Neural Networks, and margin maximization methods like SVM. Neural networks and Adaboost demonstrated the best performance in classification, achieving an accuracy of 83% in detecting bugs and 75% in detecting caterpillars, respectively. This study highlighted the transformative potential of integrating real and synthetic data in machine learning model training, resulting in significant improvements in early pest identification in corn cultivation.

Keywords: Infestation, Early identification, Machine learning, Fluorescence spectroscopy, Data augmentation.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

1.1 Importância do assunto

O milho é uma das principais culturas do país cuja produção foi estimada em 130 milhões de toneladas no ciclo 2022/2023 CONAB (2023), com boa parte dessa produção atendendo o mercado interno. Apesar do grande potencial produtivo, os agricultores de milho no país enfrentam ataques frequentes de pragas e patógenos que limitam a produtividade e aumentam os custos de produção. Dentre essas pragas e patógenos, a lagarta-do-cartucho Spodoptera frugiperda e o percevejo barriga verde Dichelops melacanthus vem recebendo atenção devido às perdas de produção e dificuldade de controle, podendo gerar perdas de até 60% Ivan Cruz and Vasconcelos (1999). Enquanto o primeiro tem chamado a atenção dos produtores devido ao desenvolvimento de resistência à variedade BT (uma variedade geneticamente modificada resistente a insetos) Horikoshi et al. (2016), o último tem se tornado uma preocupação na produção de milho de segunda safra no cerrado brasileiro. Apesar de D. melacanthus ser uma praga de menor importância na cultura da soja, após a colheita, seu inseto adulto pode causar danos severos nas fases iniciais do desenvolvimento vegetativo (estágios fenológicos v1-v3) do milho de segunda safra. Os danos causados por S. frugiperda estão associados a lesões de desfolhamento nas folhas e na parte central da planta, cuja redução na área foliar pode resultar em perdas de produtividade de até 57%, especialmente se a infestação ocorrer durante as fases v1-v7 Ivan Cruz and Vasconcelos (1999). O D. melacanthus é um inseto sugador de seiva, cujas ninfas e adultos colonizam os vasos da planta na região do colmo. Na cultura da soja, esse inseto praga causa danos nas vagens e nos grãos na fase final do ciclo da cultura. Por outro lado, na cultura do milho, os danos ocorrem principalmente nos estágios fenológicos v1 a v3, causados pelas ninfas de quarto e quinto instares e pelos adultos, o que pode resultar em perdas de produtividade de até 50% (Fernandes et al., 2018).

Embora existam vários defensivos químicos, o problema tem se agravado devido ao surgimento de resistência aos inseticidas e fungicidas e ao uso contínuo do sistema de plantio

direto, que favorece as pragas, proporcionando local de refúgio na palhada e alimento o ano todo. Além disso, o uso excessivo e continuado de defensivos químicos tem agravado problemas ambientais, como poluição de lençóis freáticos e redução da população de insetos benéficos à cultura, tais como abelhas e inimigos naturais de pragas. Neste contexto, é essencial que novas formas de manejo de pragas e doenças sejam desenvolvidas para melhorar o controle e reduzir o impacto ambiental.

1.2 Qual o estado atual

A principal forma de detecção de pragas nas culturas de milho ainda é a inspeção visual dos padrões de cores e estrutura das folhas. Essa abordagem envolve a observação direta das plantas e a identificação de quaisquer sinais visíveis de infestação por pragas. Esses sinais podem incluir mudanças na coloração das folhas, manchas, mordeduras, lesões, furos, presença de insetos ou seus ovos, entre outros. Essa técnica é frequentemente usada pelos agricultores e técnicos agrícolas para monitorar a saúde das plantas e identificar a presença de pragas precocemente.

A detecção precoce é crucial, por permitir a implementação de medidas de controle apropriadas antes que a infestação se torne grave e cause danos significativos à colheita. No entanto, a inspeção visual tem suas limitações, uma vez que algumas pragas podem não ser facilmente identificadas a olho nu, especialmente em estágios iniciais de infestação. Além disso, a inspeção visual pode ser demorada e requer conhecimento e experiência na identificação de pragas específicas. Porém, gradualmente, essas práticas vêm sendo substituída por tecnologias não-invasivas, rápidas e acuradas, principalmente aquelas baseadas em sistema de imagens Lu et al. (2021); Yang et al. (2019).

Pesquisas anteriores mostraram que alterações fisiológicas relacionadas à fotossíntese afetam a estrutura e o aparato fotossintético das folhas que sofrem algum tipo de ataque que, por sua vez, se relacionam com alterações da fluorescência da clorofila Yao et al. (2018); Dong et al. (2020). A luz reemitida pela clorofila, chamada fluorescência da clorofila a, está sendo considerada uma técnica valiosa para caracterizar o desempenho fotossintético da cultura Weng et al. (2021).

1.3 Qual nossa proposta

O diagnóstico de problemas de saúde em plantas, como ataques de pragas e patógenos, desempenha um papel crítico na agricultura moderna. A detecção precoce desses problemas é essencial para a tomada de medidas eficazes e a prevenção de perdas significativas na produção de culturas, como o milho. Nesse contexto, o uso da tecnologia de imagem de fluorescência, como o aparelho Fluorcam, oferece uma abordagem promissora para o diagnóstico automatizado.

O aparelho Fluorcam é capaz de gerar imagens de fluorescência emitidas pelas plantas e, por meio de algoritmos de processamento de imagens, gerar um conjunto de métricas relacionadas a desempenho fotossintética. Com a premissa de que a fluorescência pode ser influenciada por diversos fatores, incluindo o estado de saúde da planta, a presença de pragas e patógenos, bem como estresses ambientais. Portanto, ao analisar as métricas geradas pelo Fluorcam, é possível identificar padrões e características que indicam problemas de saúde nas plantas de milho.

Conforme comentado, os métodos de diagnóstico automatizado envolvem o uso de algoritmos de processamento de imagem, iniciada pelo aparelho a fim de extrair métricas relevantes, e posteriormente o uso de técnicas de aprendizado de máquina para analisar os dados gerado pelo aparelho. Esses algoritmos podem identificar variações anormais na fluorescência associadas a ataques de pragas, patógenos ou outros estresses.

O objetivo final desses métodos é fornecer aos agricultores e pesquisadores uma ferramenta automatizada e precisa para identificar rapidamente problemas de saúde em plantas de milho. Isso permite uma resposta mais eficaz, como o tratamento direcionado ou a implementação de medidas preventivas, contribuindo para a produtividade e a segurança alimentar. Além disso, a abordagem automatizada economiza tempo e recursos em comparação com métodos manuais de diagnóstico.

1.4 Hipótese e objetivo

A análise das métricas de fluorescência tem demonstrado resultados promissores na classificação das plantas, como será visto com mais detalhes posteriormente. Portanto, partimos da hipótese de que é possível desenvolver métodos de classificação eficazes para distinguir plantas sob ataques de pragas de plantas saudáveis com base na análise de imagens de fluorescência. Acreditamos que as diferenças no sinal de fluorescência entre plantas saudáveis e estressadas são distintivas o suficiente para permitir a classificação precisa. Além disso, esperamos que a melhoria dos dados de entrada, a seleção de atributos relevantes, o aumento do conjunto de dados disponíveis e a exploração de diversos modelos de classificação contribuirão para o sucesso dessa tarefa.

Entretanto, temos consciência das restrições que nosso conjunto de dados possui. Devido ao procedimento de preparação das amostras empregado nos experimentos, assim como a coleta de dados, que essencialmente consiste em capturar a imagem da planta em teste, revelou-se um processo bastante demorado. Cada medida podendo demandar até uma hora para ser concluída, o que impõe limitações à quantidade de dados que podemos gerar. Esta limitação, por sua vez, compromete a capacidade de alimentar os algoritmos de aprendizado com uma quantidade significativa de dados, afetando diretamente a eficácia de aprendizado do algoritmo.

O principal objetivo deste estudo consiste em desenvolver métodos para a produção de dados sintéticos que, quando combinados com os dados reais originados dos experimentos, possam resultar em classificadores robustos e confiáveis para identificar plantas sob a influência

de pragas, utilizando como base as métricas de fluorescência da clorofila. Para alcançar esse objetivo, delineamos as seguintes etapas a serem seguidas:

1. **Melhorar Dados de Entrada**: Aperfeiçoar a aquisição e o pré-processamento das métricas de fluorescência, garantindo a qualidade dos dados de entrada. Etapa com forte dependência nas condições do aparelho e seu protocolo utilizado.

2. **Selecionar Atributos Relevantes**: Identificar e reter os atributos mais informativos a serem usados na tarefa de classificação, eliminando informações irrelevantes ou redundantes, que possam confundir o algoritmo ou contribuir para uma situação de superajuste.

3. **Aumentar o Conjunto de Dados Disponíveis**: Aumentar o conjunto de dados original com a inclusão de dados gerados por redes GAN, tornando-o mais completo e diversificado. Essa abordagem oferece a vantagem de aumentar o tamanho do conjunto de dados de maneira controlada, enquanto mantém a qualidade e a representatividade das métricas geradas, tornando-as valiosas para a tarefa de classificação de plantas sob estresse.

4. **Testar Diferentes Modelos de Classificação**: Avaliar a eficácia de uma variedade de modelos de aprendizado de máquina, como redes neurais, árvores de decisão, SVM (Support Vector Machine), entre outros, e determinar o modelo mais adequado para a tarefa de classificação.

Ao alcançar esses objetivos, pretendemos fornecer uma ferramenta valiosa para a detecção precoce de estresse em plantas, o que pode ter implicações significativas na agricultura, na conservação ambiental e na segurança alimentar.

1.5 Principais contribuições

A principal contribuição desse trabalho destaca-se na integração de dados reais com dados sintéticos gerados por redes GAN (Generative Adversarial Networks), enfocando especialmente o processo de data augmentation em conjuntos de dados tabulares para o treinamento de modelos de aprendizado de máquina, processo pouco explorado nesse tipo de dados. Essa abordagem não apenas ampliou significativamente a diversidade e a complexidade do conjunto de dados, mas também permitiu uma representação mais abrangente das características intrínsecas ao cultivo do milho. Ao integrar dados artificiais, observamos melhorias notáveis na identificação precoce de pragas, evidenciando a eficácia dessa estratégia inovadora para otimizar o desempenho dos modelos em condições diversas e desafiadoras na agricultura.

1.6 Estrutura do trabalho

Este trabalho está estruturado em cinco capítulos: Introdução, Revisão bibliográfica, Materiais e métodos, Resultados e Conclusão. No próximo capítulo faremos uma revisão biblio-

gráfica das principais técnicas de detecção de pragas e doenças em plantas. No terceiro capítulo falaremos dos materiais e métodos estatísticos utilizados na elaboração desse trabalho, as plantas e insetos utilizados, os tratamentos para adequação dos dados, os modelos de aprendizado de máquina e as métricas de avaliação. No quarto capítulo apresentamos os resultados do trabalho, iniciando pela exploração do conjunto de dados para melhor entendimento e avaliação da qualidade dos dados, seguindo para o processo de "data augmentation", onde o conjunto de dados inicial será expandido e finalmente comparação das métricas obtidas antes e após aumento dos dados de treino. O último capítulo faz a conclusão dos resultados alcançados pelo estudo e apresenta algumas sugestões para trabalhos futuros.

CAPÍTULO

2

REVISÃO BIBLIOGRÁFICA

2.1 Principais pragas e doenças no Milho

O milho, uma cultura agrícola disseminada globalmente, se destaca como o cereal de maior área de cultivo no planeta (Erenstein et al., 2022). No Brasil, entretanto, certos hábitos têm desempenhado um papel relevante no surgimento de novas pragas e doenças associadas a essa cultura. Entre esses costumes, inclui-se o cultivo sequencial do milho (tanto na safra regular quanto na safrinha), a prática de monocultura, a aplicação indiscriminada de defensivos agrícolas e a adoção do sistema de plantio direto sem a devida rotação de culturas (Casela et al., 2006). Esses fatores têm contribuído significativamente para o aumento da incidência de doenças.

A produtividade desses grãos é vulnerável a uma variedade de pragas e doenças que surgem ao longo do ciclo de crescimento da cultura, desde o momento da semeadura até a colheita dos grãos, como ilustrado na figura 1. Esses desafios podem resultar em uma redução drástica na produção.

As principais pragas que afetam essa cultura são aquelas que se manifestam durante a fase larval do inseto, como a lagarta-do-cartucho, lagarta-elasmo e as larvas da broca-do-colmo. Esses insetos têm a capacidade de se alimentar das folhas, espigas e outras partes da planta, conforme destacado por Américo et al. (2016). Além disso, outras pragas que podem causar impacto nessa cultura incluem os pulgões e os percevejos, que são insetos de hábitos sugadores. Eles se alimentam retirando nutrientes das plantas, o que pode resultar em danos consideráveis.

Além dos danos causados por pragas que se alimentam da planta, a cultura do milho também sofre com diversas outras doenças, muitas delas causadas por fungos, como a Ferrugem Polissora e a mancha-branca que agem nas folhas da planta reduzindo a capacidade fotossintética da planta, levando uma redução na produção de energia.

Por experiência, sabemos que as pragas que atacam a parte aérea das plantas normalmente são mais fáceis de ser identificadas. Porém, os danos das pragas e doenças de hábito subterrâneo podem ser confundidos com deficiências de nutrientes, dificuldades climáticas ou baixa qualidade

das sementes.

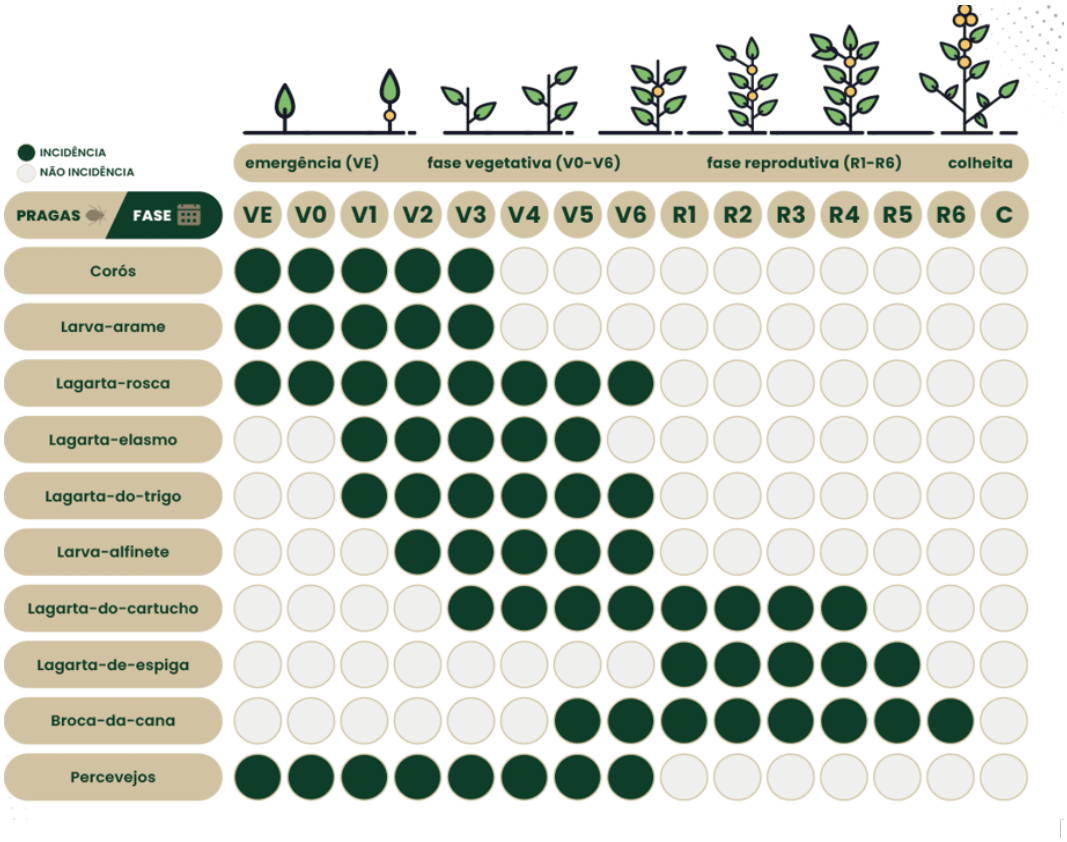


Figura 1 – Calendário de pragas e doenças na cultura do milho. Fonte: agrointeli figura Disponível em: https://blog.agrointeli.com.br/blog/como-fazer-o-manejo-integrado-do-milho/. Acesso em: 17 ago. 2023

O manejo eficaz de pragas e doenças no milho envolve práticas de manejo integrado de pragas e cultivo, incluindo a escolha de variedades resistentes, rotação de culturas, monitoramento preventivo e regular das plantações, uso controlado de pesticidas quando necessário e adoção de práticas de cultivo adequado para minimizar o risco de infestações. Quando bem empregado, o conjunto de ações de manejo integrado de pragas limita os efeitos potenciais que os defensivos químicos podem causar na saúde pública e no meio ambiente.

Recentemente, técnicas de diagnóstico precoce de patógenos tem ganhado bastante espaço. Identificar problemas de saúde ou estresses em plantas em estágios iniciais pode ter diversos benefícios para o agricultor. Detectar a doença ou infestações de pragas nas plantas antes que se espalhem é essencial para evitar a disseminação e minimizar os danos. O diagnóstico precoce permite a implementação imediata de medidas de controle, como o uso de pesticidas específicos, prevenindo o uso indiscriminado do produto, evitando assim a necessidade de tratamentos mais agressivos ou o risco de perda completa da colheita.

2.1.1 D. Melacanthus

O Diceraeus Melacanthus, também conhecido como percevejo barriga verde, é uma praga agrícola que afeta diversas culturas no mundo e ultimamente este inseto tem sido encontrado com mais frequência nas plantações de soja e milho no Brasil. Reconhecido como um problema em 1985 (Bessin, 2019), o caso recebeu pouca atenção, pois uma pequena porcentagem dos campos eram afetadas por este inseto. Hoje considerado como uma praga-chave de período inicial na cultura do milho (Bueno et al., 2021), especialmente quando ocorre após o cultivo da soja, o D. Melacanthus prejudica o desenvolvimento das plantas, reduzindo o rendimento das colheitas (Fernandes et al., 2020). Em casos mais severos podendo causar a perda total da lavoura gerando grandes prejuízos para os produtores.

Figura 2 – Diferentes espécies do percevejo e fase da planta com maior incidência. Fonte: agrolink foto Disponível em: https://www.agrolink.com.br/problemas/percevejo-barriga-verde_512.html. Acesso em: 12 ago. 2023

A praga passa por diferentes estágios de desenvolvimento, incluindo ovo, ninfa e percevejo na sua fase adulta (Pereira et al., 2007). Tanto adultos como ninfas causam danos ao perfurar as plantas para se alimentar sugando a seiva, porém os insetos na fase adulta são mais nocivos para o milho (Gomes et al., 2020). O ataque dos insetos compromete a capacidade de desenvolvimento da planta, afetando seu crescimento. Nas plantações de milho o ataque da praga resulta na redução das espigas, comprometendo também a qualidade dos grãos. Pela forma que o inseta se alimenta, o diagnóstico da doença envolve uma observação cuidadosa da planta, pois os primeiros sinais visuais da infestação demoram para aparecer. Esses sinais incluem folhas danificadas e deformação das espigas. O controle dessa praga é normalmente realizado por meio de uma combinação de técnicas: incluindo aplicação de inseticidas e uso de variedade da planta resistente ao inseto.

2.1.2 S. Frugiperda

Spodoptera Frugiperda, também identificada como lagarta-do-cartucho ou lagarta-do-milho, é considerada uma das pragas mais prejudiciais às plantações de milho e sorgo, causando grandes prejuízos econômicos (Tay et al., 2022). Além do sorgo e milho, a lagarta-do-cartucho pode atacar uma ampla variedade de plantações, incluindo outras culturas como arroz e trigo, bem como algumas plantas forrageiras. As lagartas causam danos ao alimentarem-se das folhas, espigas e outras partes das plantas. Os danos podem levar à perda de produtividade e qualidade das culturas que impacta diretamente os produtores. A resistência contra inseticidas e a notável capacidade de se adaptar desse inseto, são fatores importantes no ponto de vista de controle eficaz dessa praga (Paredes-Sánchez et al., 2021).

Figura 3 – Diferentes espécies do lagarta-do-cartucho e diferentes estágios da infestação. Fonte: agrolink foto Disponível em: https://www.agrolink.com.br/problemas/lagarta-do-cartucho_252.html. Acesso em: 16 ago. 2023

Assim como o percevejo, a lagarta-do-cartucho passa por diferentes estágios de desenvolvimento, incluindo ovo, larva, pupa e mariposa (fase adulta da espécie). Com uma capacidade reprodutiva bastante alta, a fêmea é capaz de colocar de 200 a 300 ovos por oviposição (Beserra et al., 2005). A fase de larva ou lagarta do inseto é a fase que causa os problemas para os agricultores, sendo essa compreendendo metade da vida do inseto, que gira em torno de 30 dias. Os danos causados na cultura do milho são mais intensos que os causados pelo percevejo. Os sintomas mais comuns são: folhas raspadas e perfuradas, cartucho destruído, espigas danificadas e excreções das lagartas nas plantas como mostra a figura 3. A principais formas de controle desse inseto são: tratamento de sementes, uso de variedades transgênicas da planta, controle químico através da aplicação sistêmica de inseticidas e controle biológico com uso espécies capazes de parasitar os ovos, como, por exemplo, o uso da espécie de vespa Trichogramma, capaz de parasitar de 20 a 120 ovos por fêmea (Cruz et al., 1999). Há uma preferência por este último visando boas práticas agrícolas, pois é uma forma de reduzir resíduos químicos, empecilhos para a expansão das exportações.

2.2 Técnicas de detecção de pragas e doenças em plantas

O problema causado das pragas e doenças afeta praticamente todos os tipos de cultivos que são conhecidos e praticados na atualidade. No contexto brasileiro, o cultivo de produtos como algodão, soja, milho e frutas em geral emerge como setores particularmente dinâmicos na produção agrícola do país (Beserra et al., 2005). Sendo, portanto, os produtos que sofrem o maior impacto econômico. Em escala global, essas doenças trazem aos produtores prejuízos que podem chegar 40% de perda anual em culturas economicamente importante (FAO, 2019).

Portanto, com o intuito de reduzir os impactos causados e, por conseguinte, reduzir as perdas no rendimento, é crucial que a doença seja diagnosticada o mais precocemente possível, uma vez que quanto antes for identificada, menores serão os danos. Com objetivo de diminuir os prejuízos causados por essas infestações, diferentes técnicas de detecção foram desenvolvidas ao longo dos tempos para atacar o problema.

Hoje há uma grande variedade de técnicas desenvolvidas com objetivo de detectar patogêneos em plantas. Isso inclui métodos manuais que dependem de pouco ou zero recursos tecnológicos, até métodos sofisticados como análise em laboratório apoiada por sistemas e modelos computacionais complexos para análise dos dados (Venbrux et al., 2023). De modo geral, podemos dividir as técnicas tradicionais de diagnóstico de patógenos em dua classes principais:

(1) Técnicas manuais dependem da perícia de um profissional especializado no tipo de problema, que possui conhecimento suficiente para identificar visualmente a presença de doenças ou pragas ao analisar as plantas diretamente no campo. Embora ainda amplamente utilizadas, essas técnicas são demoradas e suscetíveis a erros devido à subjetividade do processo. Além disso, apresentam limitações, sendo eficazes apenas para alguns tipos de infestações, especialmente aquelas que manifestam sintomas visuais nos primeiros dias. No entanto, sua eficácia diminui quando se trata de doenças assintomáticas ou enfermidades que demoram a exibir sinais iniciais.

(2) Técnicas de laboratórios como a Reação em Cadeia da Polimerase (PCR), sequenciamento genético e imunologia são reconhecidas como as abordagens mais eficazes disponíveis para detecção de doenças em plantas. No entanto, apesar da sua eficácia, essas técnicas são frequentemente desafiadoras de aplicar devido ao alto custo e significativa complexidade nas análises. Um dos principais obstáculos reside no custo financeiro associado a essas abordagens, que pode ser proibitivo para muitos agricultores. Além disso, a necessidade de equipamentos especializados, reagentes caros e pessoal treinado torna a implementação dessas técnicas um desafio logístico e financeiro. Além disso, outro aspecto a ser considerado é o perfil invasivo dessas técnicas, que frequentemente requerem amostragem em campo, seguida pelo transporte das amostras para um laboratório onde as análises podem ser conduzidas.

Em vista das desvantagens acima, temos observado um aumento expressivo no uso de técnicas não invasivas para a coleta de dados em campo, por exemplo, o emprego de biossensores e dispositivos ópticos. A vantagem do uso desses dispositivos é que eles podem fornecer análises

rápidas, são fáceis de usar e, o mais importante, podem ser usados para diagnóstico no local, permitindo que os agricultores tomem decisões mais rápidas. A figura 4 ilustra uma visão geral das técnicas de detecção de patógenos de plantas discutidas nessa sessão.

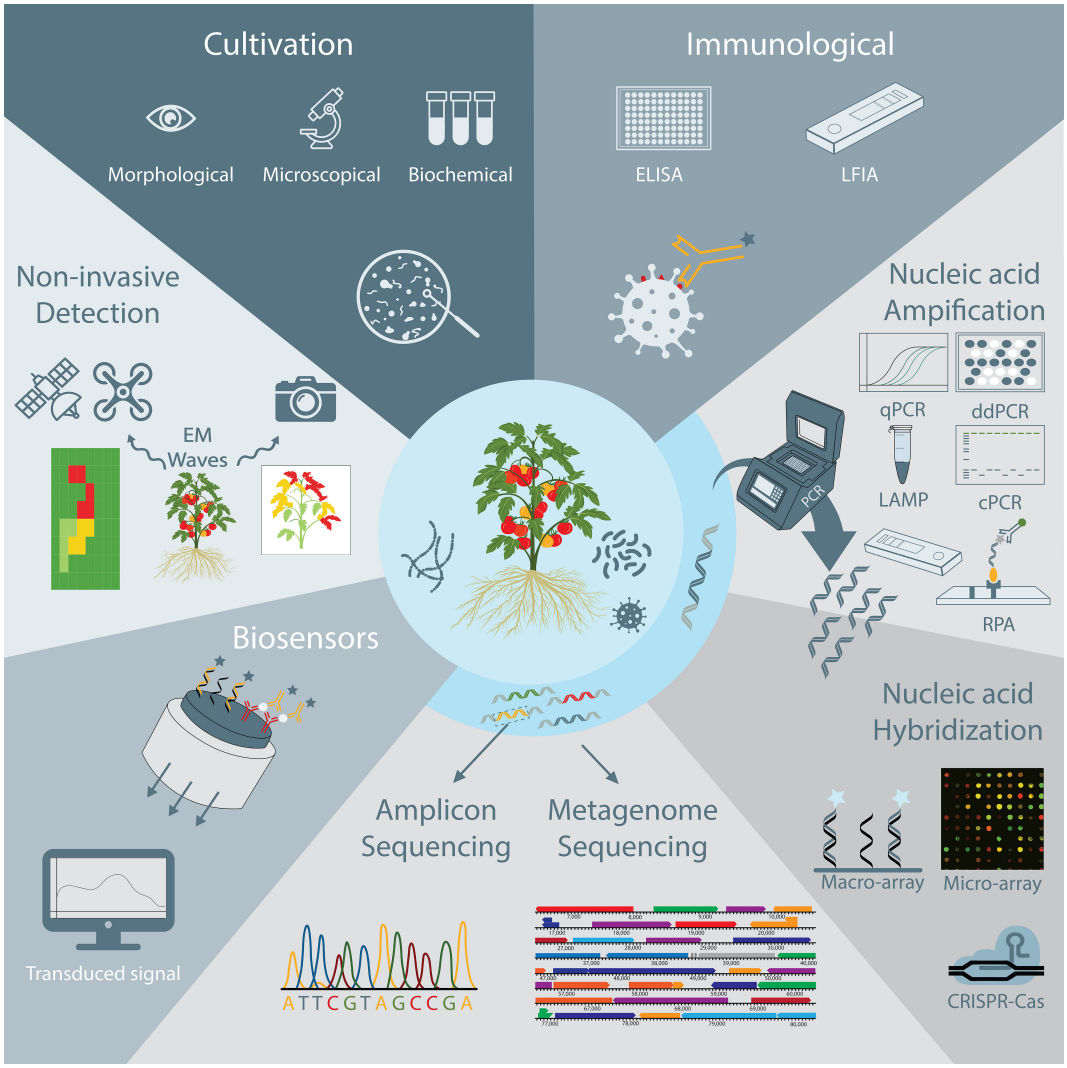


Figura 4 – Visão geral das técnicas de detecção de patógenos de plantas discutidas nessa sessão, incluindo monitoramento não invasivo, baseado em cultivo e técnicas imunológicas, técnicas de amplificação e hibridização de ácidos nucleicos, técnicas de sequenciamento de DNA e biossensores. Fonte: Venbrux et al. (2023)

A utilização de dispositivos ópticos oferece uma abordagem flexível e eficaz para a análise de patógenos em plantas, permitindo diversas formas de aplicação. Além disso, a combinação dos dados gerados por esses dispositivos com técnicas de aprendizado de máquina tem demonstrado ser promissora na identificação de patógenos em plantas. Essa sinergia entre tecnologias ópticas e algoritmos de aprendizado de máquina tem contribuído significativamente para avanços na detecção e monitoramento de patógenos, representando um campo de pesquisa promissor e inovador. Entre as mais utilizadas temos: câmeras hiper-espectrais, câmeras térmicas, espectroscopia de fluorescência induzida por laser (LIFS), espectroscopia de infravermelho e fluorescência da Clorofila.

2.2.1 Câmeras hiper-espectrais

As câmeras hiper-espectrais são instrumentos muito semelhante às câmaras convencionais, que geram imagem no espectro visível. Porém, nesse caso, as imagens desse tipo é composta por uma série de imagens revelando a reflectância observada em diferentes comprimentos de onda que, na maioria das vezes, cobre a faixa que vai do espectro visível até o infravermelho próximo. Esses instrumentos desempenham papel importante em várias áreas, incluindo agricultura, monitoramento ambiental e até na detecção de problemas de saúde humana (Lu and Fei, 2014). Na agricultura, as câmeras hiper-espectrais têm sido usadas para monitorar a saúde das culturas, detectar estresse hídrico, avaliar a qualidade do solo e até mesmo identificar doenças e pragas em estágios iniciais (Garcia and Barbedo, 2015). Ao analisar os padrões espectrais das plantas, os agricultores e pesquisadores podem identificar assinaturas específicas associadas às condições saudáveis e anormais, o que ajuda na tomada de decisões para o manejo das culturas.

Behmann et al. (2014) fizeram a hipótese que a partir de uma imagem hiper-espectral de curto alcance é capaz de descobrir processos relacionados ao estresse de forma não destrutiva nos estágios iniciais, sintomas que normalmente são invisíveis ao olho humano. Para isso, eles combinaram aprendizado supervisionado e não-supervisionado para identificar os vários estágios do desenvolvimento do stress progressivo a partir de uma série de imagens hiper-espectral. O aprendizado não supervisionado foi utilizado para separar assinaturas hiper-espectrais em clusters relacionados, enquanto maquinas de suporte vetorial (SVM) foram utilizadas no processo de classificação. O método foi aplicado em dois experimentos envolvendo plantas de cevada em vasos sob condições bem irrigadas e de estresse hídrico em uma estufa.

Apesar do potencial, as câmeras hiper-espectrais podem ser bastante caras, devido à tecnologia complexa e à necessidade de muitos sensores e filtros especializados para capturar uma ampla variedade de comprimentos de onda. Isso pode limitar sua acessibilidade para algumas aplicações ou orçamentos. Além disso. as imagens hiper-espectrais geradas por essas câmeras consistem em dados em várias bandas espectrais. Isso resulta em grandes conjuntos de dados que precisam ser processados e analisados de maneira complexa para extrair informações úteis. O processamento e a análise de dados podem exigir recursos computacionais significativos e conhecimento especializado.

2.2.2 Câmeras térmicas

As câmeras térmicas, também conhecidas como câmeras de imagem térmica ou termográficas, são dispositivos que permitem visualizar e capturar imagens com base nas diferenças de temperatura dos objetos. Elas operam no espectro infravermelho, invisível para o olho humano, mas que representa o calor radiante emitido por todos os corpos e objetos que estão com a temperatura acima do zero absoluto (Gade and Moeslund, 2014). Esse tipo de câmera foi originalmente desenvolvido com foco em vigilância e como uma ferramenta para visão noturna para os militares, mas com a popularização e queda no preço do equipamento, abrindo

significantemente um campo mais amplo de aplicações como construção, detecção de gases, indústria, em geral, aplicações militares, bem como detecção e reconhecimento de humanos.

Na agricultura, as câmeras térmicas têm se mostrado particularmente úteis na detecção precoce de doenças e pragas em plantas. A temperatura das plantas afetadas por patógenos muitas vezes difere daquelas saudáveis, e essas discrepâncias podem ser detectadas pelas câmeras térmicas. Isso ocorre porque as áreas afetadas frequentemente apresentam mudanças em seu metabolismo, o que resulta em diferentes padrões de emissão de calor.

Com essa abordagem simples Zhu et al. (2018) conseguiram mostrar que a medida que a doença evolui na planta, a diferença de temperatura máxima (MTD) também apresenta variações. No estudo eles acompanharam a evolução da doença do mosaico nas plantações de tomate e a doença do ferrugem que ataca as folhas do trigo. Os resultados mostraram que o MTD na doença do tomate variou entre 0.2°C e 1.7°C, e na doença do trigo a varição ficou entre 0.4°c e 2°C. Mostrando, portanto, que a medida a doença evolui, o valor do MTD apresenta uma tendência de aumento. Os testes realizados revelaram a detecção da doença 5 dias antes que os sintomas do mosaico se tornassem visíveis a olho nu nos tomates e 7 dias antes dos primeiros sinais de ferrugem nas folhas de trigo.

A combinação de imagens de infravermelho e imagens no espectro visível em RGB permitiu identificar a presença da Fusarium head blight (FHB) em plantações de trigo quando o nível de contaminação se encontrava entre 20% e 60% das espigas comprometidas (Francesconi et al., 2021). Em contraste, a depender de uma identificação visual que normalmente é possível quando a infestação ultrapassou 80% da plantação. Resultados preliminares também mostraram o potencial que o método tem em discriminar se a planta esta passando por estresse hídrico ou se há uma infestação de FHB.

Apesar do potencial e facilidade em analisar os dados gerados pelas câmeras térmicas, ao comparar com as câmeras convencionais, as câmeras térmicas geralmente têm uma resolução mais baixa. Isso significa que os detalhes nas imagens térmicas podem ser menos nítidos, o que pode dificultar a identificação precisa de objetos ou pessoas, especialmente em distâncias maiores.

2.2.3 Espectroscopia de infravermelho

A espectroscopia de infravermelho (IV) é uma técnica analítica que explora a interação entre a radiação eletromagnética na região do infravermelho e as moléculas dos materiais. Essa técnica fornece informações sobre a estrutura molecular, identificação de substâncias e características químicas, utilizando o fato que moléculas orgânicas absorve a radiação no infravermelho e converte em energia de vibração molecular. Sendo amplamente utilizada em várias áreas, como química, farmacologia e agricultura. Teve bastante aplicações na indústria alimentícia como uma importante ferramenta na detecção de produtos adulterado e autenticação de marcas (Capuano and van Ruth, 2016). No contexto das plantas, a espectroscopia de infravermelho pode ajudar na detecção antecipada de doenças. A análise dos espectros das plantas pode apresentar mudanças

nas assinaturas moleculares que ocorrem quando as plantas estão sob infestação por patógenos ou qualquer outro tipo de estresse. Essa abordagem possibilita a identificação precoce de problemas, permitindo aos agricultores adotar medidas corretivas antes que os danos se tornem extensos (Zahir et al., 2022).

Abu-Khalaf and Salman (2014) utilizou espectroscopia do visível e infravermelho próximo (VIS/NIR) e análise multivariada de dados (MVDA) na identificação e quantificação da doença olive leaf spot (OLS) que ataca as folhas das oliveiras e que possui longo período de incubação. Na análise foi utilizado dois métodos distintos: Partial Least Squared-Discrimination Analysis (PLS-DA), que é um método linear, foi utilizado para separar os seis níveis de severidade da doença em dois grupos principais (primeiro grupo 0, 1, 2, 3 e segundo grupo com 4,5), com essa configuração foi possível obter taxar de acerto acima de 95%. O segundo método utilizado foi o Support Vector Machine (SVM), neste caso com kernel não linear, foi possível aplicar a classificação nos seis níveis da doença 0-5 obtendo como taxa de classificação 94, 90, 73, 79, 83 e 100%, respectivamente.

Em um estudo mais recente, Bienkowski et al. (2019) empregaram análises de espectro visível e infravermelho próximo (400-1000nm) para a detecção e diferenciação de diversas doenças de grande relevância econômica na cultura da batata. Para realizar essa tarefa, eles aplicaram técnicas como o "Partial Least Square" (PLS) e Redes Neurais Artificiais no processo de classificação. Dessa forma, os modelos conseguiram identificar e distinguir doenças, inclusive aquelas sem sintomas foliares visíveis, mesmo em estágios pré-sintomáticos. Os espectros coletados em experimentos realizados em casas de vegetação foram classificados com uma acurácia de 84,6%.

Uma desvantagem dessa técnica está na interpretação dos espectros infravermelhos, que pode ser complexa, exigindo conhecimento especializado para correlacionar as bandas de absorção com grupos funcionais específicos e estruturas moleculares. A sobreposição de bandas de absorção pode dificultar a identificação de componentes individuais. Outro ponto importante que deve ser considerado é que algumas amostras precisam ser preparadas de maneira específica para análise no infravermelho, como mistura com matrizes ou dissolução em solventes. A preparação inadequada pode afetar os resultados espectrais. Essa etapa de preparação pode ser um impeditivo em muitos casos, principalmente quando há necessidade que as medidas sejam feitas em campo.

2.2.4 Técnicas baseadas em fluorescência

A técnica LIFS (Laser-Induced Fluorescence Spectroscopy), ou Espectroscopia de Fluorescência Induzida por Laser, é uma abordagem utilizada na detecção e identificação de patógenos em plantas. A técnica explora as propriedades de fluorescência das moléculas presentes nas folhas das plantas para fornecer informações sobre composição e possivelmente a presença de patógenos. Como já apresentado na seção anterior, esta também é uma técnica de espectroscopia, na qual estuda a interação da luz com a matéria, pode ser classificada em molecular (visível,

infravermelho, ressonância magnética nuclear, espectroscopia de massa e impedância elétrica) ou atômica (espectroscopia de fluorescência) que depende da natureza da interação da luz com a matéria (Khaled et al., 2018).

Trabalhos realizados aplicando a técnica mostram que a análise desses dados somada com técnicas de aprendizado de máquina é possível fazer a distinção de plantas saudáveis e plantas infectadas em plantações de citros (Ranulfi et al., 2016). No estudo eles utilizaram um sistema com excitação em 405nm para geração de dados em três situações distintas: plantas saudáveis, infestada com HLB e infestada com CVC. Sendo o HLB separado em duas subclasses: sintomático e não sintomático. Para classificação dos dados foi utilizado a combinação de Regressão e Partial Least Square Regression (PLSR), que são métodos caracterizados por tentar encontrar uma relação linear entre as variáveis preditoras que responda da melhor forma a variável alvo. Os resultados desse trabalho mostraram acurácias acima de 90% na detecção da doença Huanglongbing plantas assintomáticas 21 meses antes dos sintomas aparecerem.

No entanto, assim como qualquer técnica analítica, o LIFS também possui algumas desvantagens técnicas a serem consideradas: geralmente requer equipamentos complexos, como fontes de luz específicas, monocromadores para selecionar comprimentos de onda de excitação e emissão, detectores sensíveis, entre outros. Além disso, pode ser difícil identificar moléculas no espectro de fluorescência, pois é comum haver sobreposição. A análise de dados requer conhecimento especializado para discernir os padrões relevantes.

Outra variação da técnica baseada na fluorescência, é a fluorescência da clorofila, fundamental na compreensão da saúde das plantas e sua capacidade de realizar a fotossíntese. A análise dos dados tornou-se uma das ferramentas mais poderosa e amplamente utilizada no estudo da fisiologia das plantas, nenhuma investigação a respeito do desempenho fotossintética é vista como completa sem algum apoio de dados de fluorescência (Maxwell1 and Johnson2, 2000).

Quando a luz do sol atinge as folhas das plantas, a clorofila absorve a energia dessa luz, na qual pode seguir um dos três possíveis caminhos: ser utilizada no processo de fotossíntese, ser dissipada na forma de calor ou ser re-emitida como luz novamente – fluorescência da clorofila, geralmente na faixa do vermelho e infravermelho próximo (Govindjee and Papageorgiou, 2004).

Esses três processos ocorrem de forma competitiva, sendo que qualquer aumento na eficiência de um resultará diminuição no rendimento dos outros dois. Portanto, por medir o rendimento da fluorescência da clorofila, informações sobre mudanças na eficiência fotossintético e dissipação de calor podem ser obtidas. Embora a quantidade da luz remitida no formato de fluorecência seja muito pequena, representando entre 1 e 2% da luz total absorvida, a medida é relativamente simples.

Dessa forma é possível entender como a análise da fluorescência da clorofila pode fornecer informações valiosas sobre o estado das plantas. Por exemplo, a fluorescência da clorofila pode ser usada para medir a eficiência da fotossíntese, a saúde das plantas e sua resposta a fatores ambientais como estresse hídrico, deficiência nutricional e doenças (Maxwell1 and Johnson2,

2000). Ela oferece uma maneira não invasiva e sensível de avaliar a saúde das plantas em tempo real, o que é crucial para a produção agrícola eficiente e a gestão sustentável dos ecossistemas. Embora seja uma ferramenta valiosa para estudar a eficiência fotossintética e a saúde das plantas, também apresenta algumas desvantagens técnicas: fatores ambientais como a intensidade da luz, a temperatura e a umidade podem afetar a fluorescência da clorofila. Isso pode dificultar a comparação direta entre diferentes amostras em condições diferentes. A fluorescência da clorofila pode variar ao longo do dia, com picos durante o período de máxima atividade fotossintética. Isso deve ser levado em consideração ao realizar medições em diferentes horários do dia (Venkat and Muneer, 2022). A interpretação dos dados de fluorescência da clorofila pode ser complexa e exige conhecimento especializado. A análise de dados envolve a compreensão de diferentes parâmetros, como a Fv/Fm (fluorescência variável máxima), que mede a eficiência fotossintética. Como forma de contornar a dificuldade de análise dos dados gerados pela fluorescência da clorofila, alguns grupos interdisciplinar vem empregando técnicas avançadas de estática, tais como métodos de classificação binaria e seleção de atributos importante (Cen et al., 2017), obtendo taxas de acerto próximas de 97% na identificação da Citrus Huanglongbing.

CAPÍTULO

3

MATERIAIS E MÉTODOS

3.1 Insetos

A escolha do método de obtenção dos insetos para realização dos ensaios, desempenha um papel crucial nos estudos relacionado ao diagnóstico de doenças. Existem várias maneiras de adquirir esses insetos, no entanto, as mais comuns incluem a coleta na natureza, a criação em laboratório ou a aquisição por meio de fornecedores especializados. A coleta direto na natureza acaba limitando os possíveis casos de estudo, sendo frequentemente usada para estudar a infestação em seu ambiente natural. No nosso caso estamos interessados em controlar principalmente a idade dos insertos, onde a criação em laboratório permite maior controle sobre a população.

Os percevejos barriga verde (*D. melacanthus*), foram criados em laboratório na Embrapa Pecuária sudeste, na unidade de São Carlos, SP, Brasil. Para os testes biológicos, os percevejos foram colocados em uma câmara de demanda bioquímica de oxigênio (DBO) a $25^\circ \pm 2^\circ C$, com 12 horas de exposição a luz e uma umidade relativa de $70 \pm 5\%$, onde ovos e ninfas do 1° ao 3° instar foram mantidos em potes de plástico transparente, forrados com papel filtro e tampas com furos para ventilação. A alimentação das ninfas até o 3° instar consistia em vagens de feijão e fornecimento de água via algodão umedecido. As ninfas do 4° e 5° instar, bem como os insetos adultos, foram mantidos em recipientes de plástico com um volume de 4 litros. A alimentação das ninfas do 4° e 5° instar e dos insetos adultos consistia em amendoim e girassol, vagens de feijão e grãos de soja, além de algodão umedecido para fornecimento de água aos insetos. A criação teve início com ovos provenientes do laboratório de entomologia da Embrapa Recursos Genéticos e Biotecnologia, Brasília, DF, Brasil. Os insetos usados nos testes eram adultos não-sexados de *D. melacanthus*, com idades entre quatro e sete dias após a emergência.

Os insetos de *S. frugiperda* foram doados pela fazenda de insetos da Estação Experimental Monsanto, em Santa Cruz das Palmeiras, SP, Brasil. Os insetos foram incubados em uma câmara de demanda bioquímica de oxigênio (DBO) a $25^\circ \pm 2^\circ C$, com exposição à luz solar de 12 horas a cada ciclo de 24 horas e uma umidade relativa de $70 \pm 5\%$, mantidos em recipientes de plástico

transparente e alimentados com dieta artificial. As larvas de *S. frugiperda* foram usadas nos testes biológicos quando atingiram o primeiro estágio de desenvolvimento, onde sabemos que é a fase que o inseto mais causa prejuízos nas plantações.

3.2 Cultivo e infestação das plantas

A coleta de dados para esta análise foi conduzida por meio de dois experimentos realizados em épocas diferentes. No primeiro experimento, infestamos o percevejo-barriga-verde em plantas saudáveis de milho, enquanto no segundo, as plantas foram expostas à lagarta cartucho. Em ambos os experimentos, metade das plantas disponíveis foi deliberadamente contaminada pelas pragas, enquanto a outra metade foi mantida como grupo de controle, destinada a fornecer dados de referência sobre plantas saudáveis.

Neste bioensaio, foram utilizadas duas cultivares de milho: *Zapalotes chico* (LE) e *Spodoptera sintético* (SE), a fim de avaliar se os parâmetros medidos das plantas variavam com a cultivar. As plantas foram cultivadas em recipientes de plástico preto com volume de 430 mL, contendo substrato para plantas Vivatto Slim Plus®. O cultivo foi realizado em estufa até que as plantas atingissem o desenvolvimento fenológico V2 a V3 (duas a três folhas expandidas), momento em que a infestação foi realizada com cinco insetos por planta. Para o bioensaio, foram utilizadas lagartas de primeiro instar de *S. frugiperda* ou adultos de *D. melacanthus* com 4 a 7 dias de emergência.

A coleta das medidas de fluorescência foram feitas em um ambiente com temperatura entre 19 e 32$^{\circ}C$ e umidade relativa entre 30 e 42%. As medidas foram feitas nos períodos de 4, 8, 12, 24, 48 e 72 horas após a inoculação dos insetos nas plantas. Quatro plantas distintas foram avaliadas em cada medição. As plantas infestadas com insetos foram identificadas como LE e SE, para ser possível a distinção entre cultivares, enquanto as não infestadas, chamadas de testemunhas, foram identificadas com os tratamentos LET e SET. A distinção entre os tipos de infestação é realizada por meio da inclusão de uma coluna ao final das medidas, denominada "percevejo" para dados relacionados à infestação por percevejos e "lagarta" para dados associados à infestação por lagartas. Antes da coleta dos dados, as plantas foram colocadas em um ambiente escuro por 30 minutos.

3.3 Equipamento

Os dados de fluorescência da clorofila foram gerados com o equipamento Closed FluorCam FC 800-C, Photon System Instruments (Figura 5), na qual tem a capacidade de medir a cinética da fluorescência da clorofila nas plantas por meio de imagens geradas da amostra. O equipamento é composto de uma câmera CCD, quatro painéis de LED fixos e uma roda de filtro equipada com cinco filtros ópticos. Os painéis de LED fornecem irradiância uniforme em uma área de 90 x 90 mm. O equipamento foi projetado de forma que possa estimular a

emissão da fluorescência por uma fonte de excitação. As imagens são geradas a partir do sinal de fluorescência emitida pela planta e captado pelo sensor.

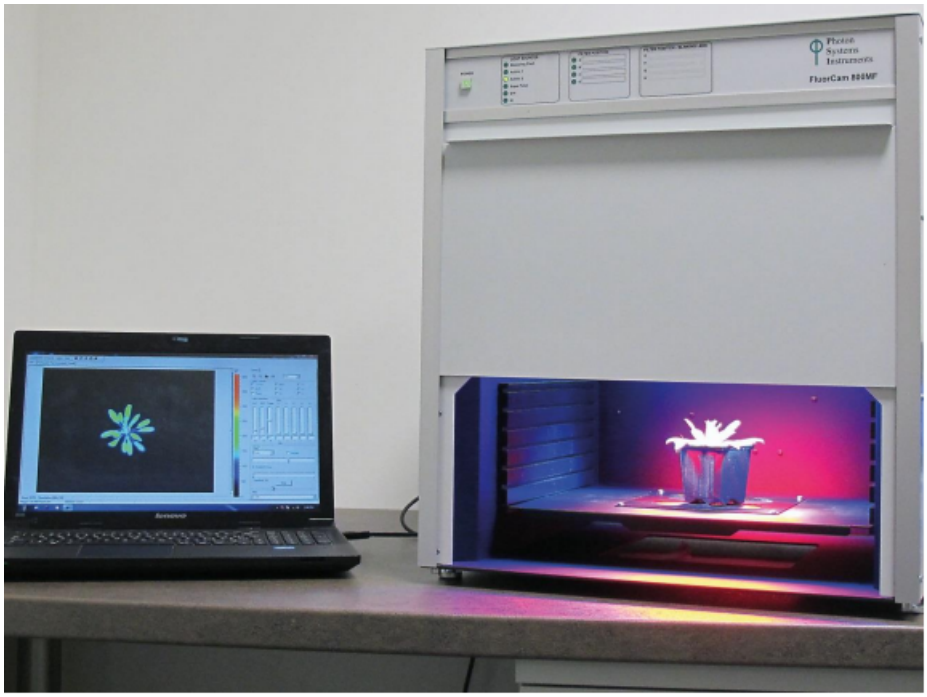


Figura 5 – Sistema de imagem, FluorCam FC800-C, usado para medir o sinal de fluorescência das folhas de milho

Em todas as aplicações, a câmera permite a captura de imagens de fluorescência induzidas por luz actinic (vermelho-laranja em 617 nm) ou flashes de saturação (cool white 6500 K). O tempo e a amplitude da irradiância actinic são determinados por protocolos pré-definidos, cujos parâmetros são escolhidos no programa FluorCam 7.0 do próprio equipamento.

Neste trabalho utilizamos dois protocolos diferentes definidos pelos Protocolo I e II. O protocolo I foi o Actinic Light Curve II, com filtro de clorofila e os parâmetros pré-definidos, porém ajustados de acordo com cada amostra e para não saturar a imagem. Os parâmetros ajustados foram: i) Shutter da câmera em 20μs; ii) ajuste da sensibilidade com a luz flash, buscando manter a cor da escala da amostra em 500, azul-escuro na escala de cores; iii) ajuste da intensidade da luz Act 2, buscando manter a cor da amostra na escala entre 1.000 e 1.500; iv) ajuste do pulso de saturação Super, buscando manter a imagem da amostra branca. Com os parâmetros ajustados para cada amostra, a intensidade da luz actinic variou de $300-2.000\mu mol(\text{fótons}).m^{-2}.s^{-1}$, e a intensidade do super pulso não ultrapassou $4.000\mu mol(\text{fótons}).m^{-2}.s^{-1}$. Como as amostras se diferem umas das outras em altura, tamanho e geometria, permite-se um ajuste de parâmetro para cada amostra. Com todos os parâmetros ajustados para cada amostra, os dados foram adquiridos.

No Protocolo II, empregamos o modo Snapshot - Imaging of Fluorescence Proteins and Fluorescent Dyes para complementar as métricas relacionadas à cinética da fluorescência. Nesse modo, o sistema captura imagens da fluorescência emitida. Em folhas verdes, as principais bandas de emissão ocorrem nos comprimentos de onda f440, f520, f690 e f740.

Como resultado da utilização, o equipamento oferece um software que possibilita a integração do equipamento com um computador pessoal onde é possível visualizar a imagem de fluorescência da clorofila capturada (Figura 6).

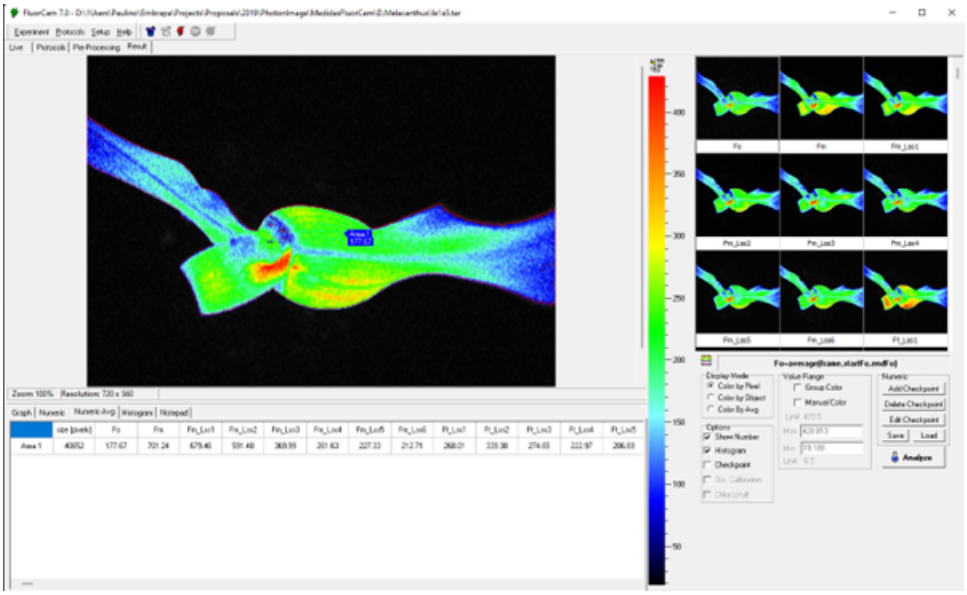


Figura 6 – Imagem de fluorescência da clorofila acompanhada dos parâmetros calculados pelo software

3.4 Métricas de fluorescência

Conforme mencionado anteriormente, o aparelho realiza o cálculo das métricas de fluorescência da clorofila com base em uma imagem da planta capturada na câmara. O software associado ao dispositivo prossegue com o processamento da imagem, selecionando automaticamente uma pequena área que será utilizada para estimar os parâmetros da fluorescência da clorofila. Dentro dessa área, são calculadas a média e a variância para cada métrica que será extraída da imagem, levando em consideração a intensidade do sinal em cada ponto da área selecionada. A Figura 7 mostra o resultado de um experimento que mede o efeito Kautsky com luz modulada.

Como consequência da criação e análise da imagem, são obtidos dois conjuntos de métricas: as métricas básicas, que consistem em valores calculados diretamente a partir da imagem capturada pela câmera (ver Tabela 1), e um segundo grupo composto pelas métricas derivadas (ver Tabela 2). Essas métricas derivadas são calculadas combinando dois ou mais parâmetros básicos. No total, são geradas 90 métricas da eficiência da fotossíntese.

3.5 Processamento dos dados

Todo processamento dos dados e implementação do modelo foram realizados com ajuda da linguagem de programação Python, utilizando principalmente as bibliotecas numpy e pandas

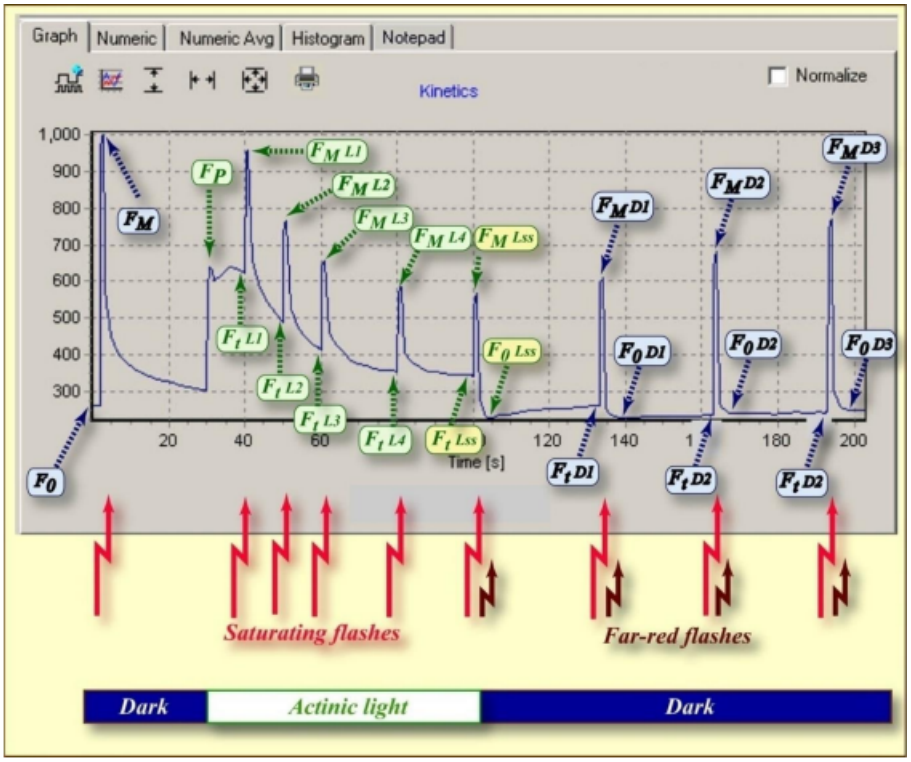


Figura 7 – Janela do FluorCam mostrando o resultado de um experimento que mede o efeito Kautsky. As etiquetas azuis marcam os níveis de fluorescência medidos no escuro (F0, FM) ou durante o relaxamento no escuro $(Ft\_Dn, FM\_Dn, F0\_Dn)$. As etiquetas verdes marcam os níveis de fluorescência medidos durante a adaptação à luz (Ft_Ln, FM_Ln). As etiquetas amarelas mostram os níveis de fluorescência em estado estacionário atingidos em luz contínua $(Ft\_Lss, FM\_Lss, F0\_Lss)$. As setas vermelho-claro indicam o momento dos flashes de saturação. As setas vermelhas-escuras indicam flashes de infravermelho distante que excitam seletivamente o Fotossistema I

Tabela 1 – parâmetros da fluorescência da clorofila - métricas básicas

Parâmetro	Descrição
	Protocolo I
fo	fluorescência miníma no estado adaptado ao escuro
fm	fluorescência máxima no estado adaptado ao escuro
fm_lss	fluorescência máxima em estado estacionário na luz
fo_lss	fluorescência miníma em estado estacionário na luz
ft_lss	fluorescência em estado estacionário na luz
	Protocolo II
f440	fluorescência na região do azul
f520	fluorescência na região do verde
f690	fluorescência na região do vermelho
f740	fluorescência na região do vermelho distante

para manipulação e visualização dos dados e a biblioteca de Data Science *scikit-learn* para o processamento dos dados e implementação dos modelos de classificação. O Google Colab foi adotado como editor principal para escrita dos códigos, pois além de permitir a execução

Tabela 2 – parâmetros da fluorescência da clorofila - métricas calculadas

Parâmetro	Fórmula	Descrição
	Protocolo I	
fv	$fm - fo$	variação fluorescência no estado adaptado ao escuro
qy_max	$\frac{fv}{fm}$	rendimento máximo do sistema PSII
fv_lss	$fm\_lss - fo\_lss$	variação da fluorescência no estado estacionário
fq_lss	$fm\_lss - ft\_lss$	variação da fluorescência sob luz actínica
qp_lss	$\frac{fm\_lss - ft\_lss}{fm\_lss - fo\_lss}$	coeficiente de extinção fotoquímica no estado estacionário
qn_lss	$\frac{fm - fm\_lss}{fm - fo\_lss}$	–
ql_lss	$\frac{qp\_lss}{fo\_lss/ft\_lss}$	eficiência máxima do sistema PSII no estado estacionário
npq_lss	$\frac{fm - fm\_lss}{fm\_lss}$	extinção não fotoquímica no estado estacionário
qy_lss	$\frac{fm\_lss - ft\_lss}{fm\_lss}$	rendimento quântico no estado estacionário do sistema PSII
fv/fm_lss	$\frac{fm\_lss - fo\_lss}{fm\_lss}$	rendimento quântico do sistema PSII no estado estacionado adaptado a luz
	Protocolo II	
f440_f520	$\frac{f440}{f520}$	relação fluorescência azul/verde
f440_f690	$\frac{f440}{f690}$	relação fluorescência azul/vermelho
f440_f740	$\frac{f440}{f740}$	relação fluorescência azul/vermelho distante
f520_f690	$\frac{f520}{f690}$	relação fluorescência verde/vermelho
f520_f740	$\frac{f520}{f740}$	relação fluorescência verde/vermelho distante
f690_f740	$\frac{f690}{f740}$	relação fluorescência vermelho/vermelho distante

interativa dos comandos é possível manter todo o fluxo de trabalho documentado mesclando células contendo código com células contendo linguagem de marcação. Diferente dos notebooks tradicionais como jupyter e zeppelin que executam localmente ocupando recursos da máquina, o Colab roda sobre uma máquina virtual compartilhada hospedada na cloud pública do Google. O Colab roda além da versão sem custos, que atende boa parte das cargas de trabalho, possui modelos de pagamento flexíveis se houver a necessidade de um poder computacional mais expressivo.

3.5.1 Remoção de outliers

Para evitar distorções nas classificações e análises, os valores atípicos que se desviam significativamente da maioria (*outliers*) foram substituídos com auxilia da análise via quartis, técnica estatística comumente utilizada para identificar valores errantes. Os quartis são medidas estatísticas que dividem um conjunto de dados ordenado em quatro partes iguais. O primeiro quartil (Q1) é o valor do conjunto que delimita os 25% menores valores: 25% dos valores são menores do que Q1 e 75% são maiores. O segundo quartil coincide com a mediana, que separa o conjunto pela metade. O terceiro quartil (Q3) é o valor que delimita os 25% maiores valores, como ilustrado na figura 8.

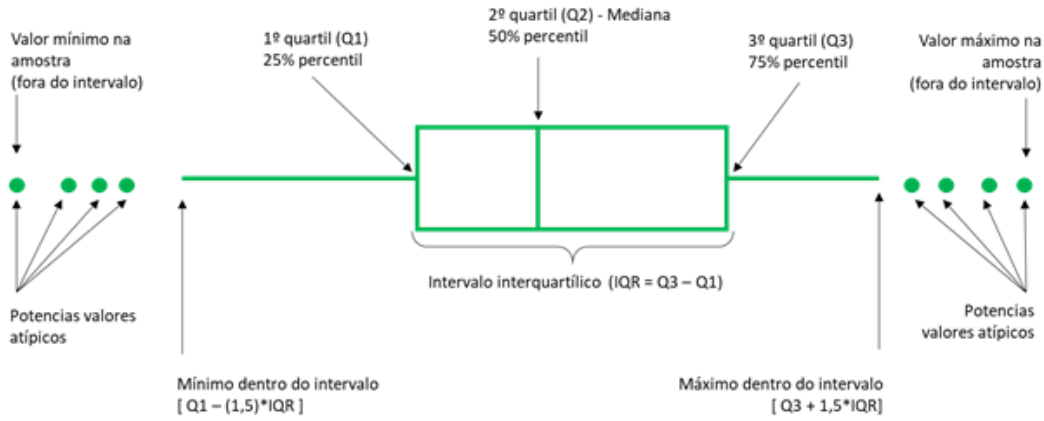


Figura 8 – Representação gráfica do quartil utilizado na identificação de outliers

Como possuímos um conjunto reduzido de amostras, remover as linhas contendo os *outliers* deixaria o conjunto de amostras mais restrito ainda. Como solução, os *outliers* foram substituídos pela mediana da variável em questão. Também pelas características do experimento, executado de forma sistemática, o conjunto não possui dados faltantes, dispensando, portanto, qualquer tipo de tratamento nesse sentido.

3.5.2 Normalização dos dados

Com o objeto de obter um diagnóstico precoce do ataque das pragas antes que os sintomas sejam visíveis aos olhos, os dados gerados pelo dispositivo precisam passar por alguns

tratamentos para ser viável a aplicação de algoritmos de classificação no contexto de aprendizado de máquina. Um primeiro ponto a ser definido são quais tipos de métricas geradas pelo aparelho serão utilizadas no treinamento do modelo. O equipamento fornece os dados médios e o desvio padrão de cada métrica. Portanto, é possível desenvolver modelos de classificação usando as médias, os desvios padrões ou ambos. A dúvida surgiu porque o desempenho dos modelos dependeu da estatística utilizada. Após definido a estatística a ser utilizada nas análises, o próximo tratamento é qual tipo de normalização será feita sobre os dados.

Essa etapa é importante, pois as medidas são geradas em ordem de grandeza distintas e, quando há uma diferença expressiva entre medidas, o modelo pode favorecer indevidamente atributos com maior ordem e dar pesos baixos nas demais. Portanto, a normalização dos dados é conveniente, pois permite a comparação de variáveis que possui diferentes escalas e unidades. Além de que, muitos modelos de aprendizado de máquina funcionam melhor quando os dados estão na mesma escala, sendo um tratamento obrigatório em muitos estimadores de aprendizado de máquina para alcançar a convergência, como as redes neurais.

Para o conjunto de dados em análise foi utilizado o método StandardScaler da biblioteca scikit-learn. Este método, na sua forma mais básica, padroniza os atributos removendo a média e ajusta a escala para variação unitária.

$$z = \frac{X - \bar{X}}{S},$$

onde $\bar{X}$ e S são a média e o desvio padrão respectivamente.

3.5.3 Seleção de atributos

Com um total de 90 métricas fornecidas pelo aparelho, a maioria delas derivadas a partir de medidas base, é esperado que haja uma forte correlação entre diferentes métricas. Outro ponto é que possuímos um conjunto reduzido de amostras e isso faz com que a densidade do conjunto de dados diminua com o aumento da dimensionalidade. Portanto, não vai ajudar muito fornecer todas as métricas para o algoritmo de treinamento. Forçar tal situação, além de aumentar o tempo de processamento, tem grande chance do algoritmo realizar um super ajuste nos dados de treino e falhar na classificação do conjunto de testes.

Uma forma rápida de visualizar este aspecto dos dados é mediante um mapa de calor aplicado sobre a matriz de correlação das métricas. A matriz de correlação fornece estatisticamente relacionamentos significativos entre as variáveis no conjunto de dados, onde os coeficientes da matriz exibe a força da dependência linear entre duas variáveis. No presente trabalho, foram utilizados os coeficientes de correlação de Pearson (Benesty et al., 2009) para calcular a matriz de correlação, onde os coeficientes podem variar entre +1 e -1. Sendo que um valor próximo de zero indica que não existe relação entre duas variáveis ou a relação é muito pequena, a medida que o valor do coeficiente se aproxima de 1 há um indicativo de uma relação proporcional entre as variáveis e quando o coeficiente de correlação se aproxima de -1 indica que a relação

entre as variáveis são do tipo $y = -x$, ou seja, quando uma aumenta a outra diminui. Dado os conjuntos $X = [x_1, x_2, ..., x_n]$ e $Y = [y_1, y_2, ..., y_n]$ calcula-se o coeficiente de correlação de Person pela seguinte fórmula:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(x_i - \bar{y})^2}}$$

A partir da matriz de correlação é possível concluir que deve haver uma redução na quantidade de atributos, mas uma dúvida que surge é identificar quais atributos podemos descartar e quais serão mantidos no modelo. Essa tarefa, conhecida como redução de dimensionalidade, tem por objetivo reduzir a dimensionalidade do espaço de features considerando a obtenção de um conjunto de features principais. Para isso diversas técnicas foram desenvolvidas que auxiliam na seleção de variáveis, essas técnicas que variam desde aplicação de algoritmos sofisticados utilizando modelos de aprendizado como modelo alvo até metodologias simples como análise do gráfico de distribuição.

Neste trabalho empregamos inicialmente a análise gráfica da distribuição das métricas para ter uma noção rápida de quais regiões do espectro apresenta uma melhor separação. Apesar do método apresentar bons resultados, é pouco eficiente, pois são necessárias duas etapas: o primeiro passo é partir de um gráfico de distribuição, selecionar as variáveis que melhor separam as classes e em seguida avaliar a correlação entre as variáveis selecionadas na primeira etapa. Para um conjunto de dados com poucas variáveis o método até pode ser aplicado e apresentar bons resultados, no entanto, para um conjunto com muitos atributos, como no presente caso, o método deixa de ser viável, podendo até mesmo gerar resultados indesejados.

Uma técnica mais robusta e que entregou melhores resultados é a eliminação recursiva de variáveis (RFE) (Granitto et al., 2006). O objetivo desse algoritmo é selecionar as variáveis considerando recursivamente conjuntos cada vez menores de variáveis. Primeiramente devemos escolher um estimador que atribui diferentes pesos para cada variável do problema, em seguida o estimador é treinado no conjunto inicial contendo todas as variáveis e é obtido um valor que representa a importância de cada variável. Em seguida, as variáveis com menor importância são removidas do conjunto inicial. O processo é repetido recursivamente em conjunto de dados cada vez menor até atingir o número de variáveis desejada seja alcançado.

3.6 Agrupamento

Análise de agrupamento foi utilizada com objetivo de identificar a formação de possíveis grupos no conjunto de dados que pudessem ser utilizados em algum processo de personalização da classificação. Análise de formação de grupos é uma forma de obter *insights* valiosos a partir de um conjunto de dados complexo. Para esse tipo de tarefa é comum utilizar algoritmos de aprendizado de máquina, como K-Means, Hierarchical Clustering e DBSCAN, porém no presente trabalho optamos em trabalhar a formação de cluster com PCA (Análise de Componentes Principais). Técnica muita utilizada na redução de dimensionalidade, tem como principio

encontrar os componentes principais, que são as direções ao longo das quais os dados possui maior variação. Ao projetar os dados nas direções dos componentes principais mais importantes, criamos novas características que capturam a maior parte da variabilidade nos dados, mas com uma dimensão menor do que o conjunto de dados original.

3.7 Amostragem

Amostragem dos dados é uma técnica comum usada em ciência de dados para dividir um conjunto de dados em duas partes distintas, normalmente reservam-se 80% dos dados para treinar o modelo e o restante (20%) para testar o modelo. Existem funções prontas para auxiliar nessa tarefa, como por exemplo train_test_split da biblioteca scikit-learn.model_selection. A função, além de receber o conjunto de dados que desejamos separar e a proporção que será reservada para o conjunto de testes, possui um parâmetro chamado stratify que faz uma divisão para que a proporção de valores na amostra produzida seja igual à proporção de valores fornecidos pelo parâmetro. Por exemplo, se a variável alvo (y) é uma variável categórica com valores 0 e 1 e existem 40% de zeros e 60% de valores um, stratify=y irá produzir uma divisão aleatória contendo 40% de zeros e 60% de um em ambos os conjuntos de dados. No entanto, a característica dos nossos dados não é favorecida com esse tipo de divisão, isso porque há uma dependência com o horário que as medidas foram realizadas: 4, 8, 24, 36, 48 e 72 horas após a infestação. Por outro lado, em cada horário as medidas foram repetidas 4 vezes para cada planta. Dessa forma, escolhendo uma repetição específica para o conjunto de teste (por exemplo, repetição = 2) é possível manter a proporção dos dados originais e teremos assim 75% dos dados para treinar o modelo e 25% para teste.

3.8 Métodos de aprendizado de máquina

Os métodos de aprendizado de máquina podem ser divididos em duas classes principais: métodos preditivos e métodos descritivos. Nas tarefas preditivas, o algoritmo é aplicado em um conjunto de exemplos de treinamento rotulado para inferir um modelo preditivo capaz de predizer, para um novo exemplo não utilizado no treinamento, o valor do atributo alvo. Nessas tarefas são utilizados algoritmos que estão dentro do contexto de aprendizado supervisionado e se diferenciam pelo valor de etiqueta a ser predita: discreto, quando trabalhamos com tarefas de classificação; e contínuo, quando se trata de tarefas de regressão. No atual trabalho estamos interessados em fazer a distinção entre plantas infectada e saudável, portanto neste caso se trata de tarefa de classificação. Em tarefas descritivas, os algoritmos tentam identificar estruturas a partir dos valores preditivos de um determinado conjunto de dados. Por não fazer uso de um supervisor externo, esses algoritmos estão enquadrados no contexto de aprendizado não supervisionado. Uma das principais aplicações dessa classe de algoritmo é encontrar grupos

de objetos similares entre si no conjunto de dados. A figura 9 ilustra de forma hierárquica as categorias de aprendizado e as tarefas associada.

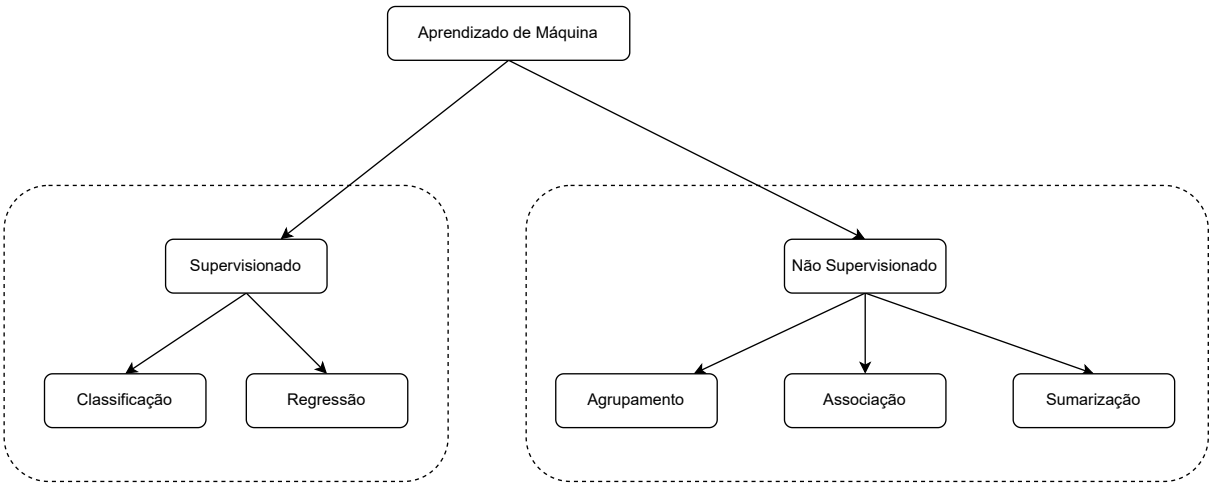


Figura 9 – Representação hierárquica das categorias de aprendizado e as tarefas associada

Em uma tarefa de classificação, a primeira etapa consiste em treinar o modelo a partir dos dados de treino, nessa etapa o algoritmo aprende ou tenta inferir a relação entre as características do indivíduo (X) e a variável alvo (y). Um ponto importante nessa etapa é a escolha do modelo que será utilizado. Há disponível na literatura uma vasta quantidade de algoritmos de aprendizado de máquina, cada um com propósito e aplicação distinta e a escolha do algoritmo é um fator crucial no sucesso da análise. Pelas características do problema devemos olhar para modelos de classificação, pois estamos preocupados em identificar qual planta possui determinada doença, dessa forma podemos associar plantas saudáveis com o valor zero (y=0) e plantas contaminadas com o valor um (y=1). Mesmo dentro do espaço dos algoritmos de classificação ainda existe a tarefa em escolher qual algoritmo utilizar. Dado nosso conjunto reduzido, devemos dar preferência para modelos que trabalhem bem com poucos dados, dessa forma o estudo iniciou aplicando dois modelos principais: System Vector Machine (SVM) e Decision Trees (DT), para validar se o conjunto de dados é aplicável no contexto de aprendizado de máquinas. Prosseguimos com a inclusão de outros modelos no estudo, como K-Nearest Neighbors (KNN), AdaBoobt e Multi Layers Percepton (MLP), visando ampliar nossa compressão sobre a performance de modo geral.

3.9 Métricas de avaliação

Para avaliação da performance dos modelos de classificação propostos, utilizamos a acurácia como principal medida de desempenho, na qual reporta a taxa de acerto do modelo resumindo toda informação em uma única métrica, variando entre 0 e 1, e valores próximos de 1 são considerados melhores. Como complemento, a matriz de confusão foi empregada para analisar as classes que o algoritmo de classificação tem maior dificuldade em classificar. Para um conjunto de dados em análise, as linhas dessa matriz representam as classes verdadeiras, e

as colunas, as classes preditas pelo modelo. Na figura 3 é apresentado uma matriz de confusão para um problema de classificação binária, onde verdadeiro positivo (VP) são os casos que o modelo previu corretamente a classe positiva, falso positivo (FP) são os casos que o modelo previu incorretamente a classe positiva, falso negativo (FN) os casos em que o modelo previu incorretamente a classe negativa e verdadeiro negativo (VN) são os casos em que o modelo previu corretamente classe negativa.

Tabela 3 – Matriz de confusão para problema de classificação binário

		Predição		
		Negativo	Positivo	Métricas
Verdadeiro	Negativo	**Verdadeiro Negativo (VN)**	**Falso Positivo (FP)**	Especificidade $\frac{VN}{VN+FP}$
	Positivo	**Falso Negativo (FN)**	**Verdadeiro Positivo (VP)**	Sensibilidade $\frac{VP}{VP+FN}$
	Métricas	Confiabilidade $\frac{VN}{VN+FN}$	Precisão $\frac{VP}{VP+FP}$	Acurácia $\frac{VN+VP}{VN+FN+VP+FP}$

A partir da matriz de confusão podemos definir uma série de outras medidas de desempenho (Monard and Baranauskas, 2003). Entre elas, daremos ênfase na Precisão, que fornece uma medida da proporção de previsões corretas em relação ao número total de previsões. Sensibilidade que fornece uma medida da proporção de instâncias positivas que foram corretamente identificadas em relação ao total de instâncias positivas. Especificidade que mede a proporção de instâncias negativas que foram corretamente identificadas pelo modelo em relação ao total de instâncias negativas.

Para análise e comparação de resultado em análise de agrupamento é interessante validar as estruturas encontradas pelos algoritmos. Considerando que a análise de agrupamento é uma tarefa não supervisionada, devemos estar ciente que não existe um resultado verdadeiramente correto, na qual estamos buscando atingir. Entretanto, é importante ter mecanismos para identificar se o algoritmo está realmente encontrando uma estrutura apropriada. Neste caso foi utilizado estruturas já conhecida e o algoritmo foi avaliado com respeito a sua habilidade em reproduzir essas estruturas.

3.10 Aumento artificial de dados

Redes Adversárias Generativas (GANs) foram utilizadas no processo de sintetização de dados, como forma de complementar os dados iniciais gerados no experimento. Isso permitiu o fornecimento de informações adicionais para algoritmos de aprendizado e modelos estatísticos,

que frequentemente requerem uma quantidade substancial de exemplos para alcançar uma convergência satisfatória e assegurar o aprendizado eficiente.

As GANs são arquiteturas de redes neurais profundas compostas por duas redes, a rede geradora G na qual gera uma instância de objeto, podendo ser uma imagem, video, texto ou dados tabulares, e uma rede discriminadora D na qual recebe a instância criada pela rede geradora e tenta descobrir se o dado faz parte do conjunto de exemplos ou se o dado foi criado pela rede geradora. Durante a execução do algoritmo, cada modelo tenta obter vantagem em relação ao outro, com a rede geradora criando exemplos cada vez mais realistas, pois recebe *feedbacks* constantes da rede discriminadora. Do outro lado, a rede discriminadora fica cada vez mais especializada na tarefa de diferenciar dados reais e dados fictícios. Portanto, esses modelos competem um contra o outro, daí chamados de redes adversárias. De forma resumida, a rede G tenta aproximar a distribuição dos dados gerados à distribuição dos dados reais (Figura 10).

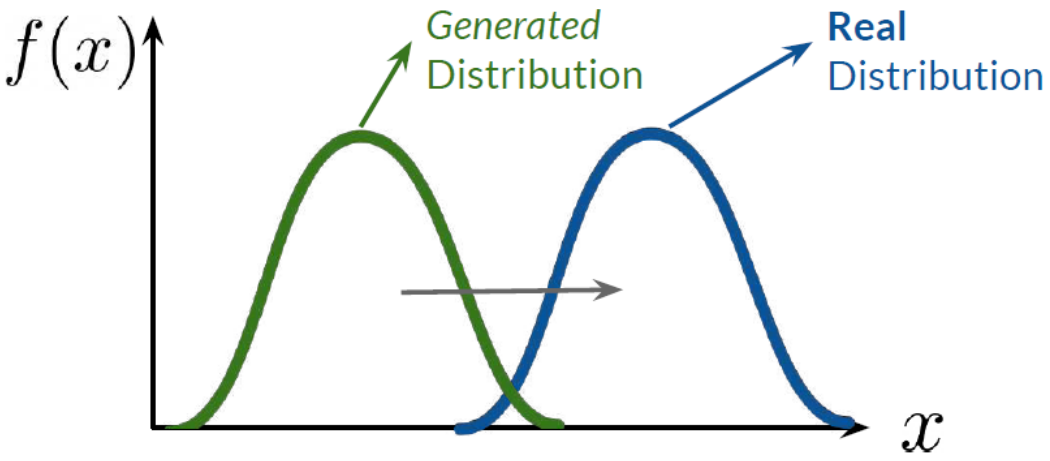


Figura 10 – Representação da distribuição dos dados reais e dados gerado pela rede.

Os *feedbacks* mencionados anteriormente representam a maneira pela qual a rede, como um todo, adquire conhecimento. Esses *feedbacks* são obtidos por meio de um processo de otimização da função de custo, que mede o quão próxima à previsão está do valor real. Como resultado, a rede G (geradora) procura maximizar a função de custo, enquanto a rede D (discriminadora) busca minimizá-la.

$$J(\Theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log h(x^{(i)},\Theta) + (1-y^{(i)})\log(1-h(x^{(i)},\Theta))],$$

onde h é o valor predito pelo modelo, y classe da instância, x atributo de entrada e Θ os parâmetros da rede.

Embora existam apenas desde 2014 (Goodfellow et al., 2014), as GANs já alcançaram um desempenho impressionante em diversas tarefas. Das tarefas mais impressionantes estão a geração de rostos humanos realista, animação de obras de arte famosas e translação de imagens, que significa que elas podem pegar uma imagem de um domínio e transformar em outro domínio, por exemplo, transformar um cavalo em uma zebra e vice-versa.

Além das redes GANs, existem outras técnicas desenvolvidas exclusivamente para auxiliar no processo de síntese de dados. Entre as técnicas disponíveis, podemos citar duas abordagens: o método Bootstrap e o aumento de dados com Autoencoder.

Método *Bootstrap*: O método *Bootstrap* é uma das abordagens mais simples. Ele cria novos dados selecionando aleatoriamente exemplos do conjunto de dados original. Geralmente conhecido como amostragem ou re-amostragem com reposição, essa técnica permite que novos exemplos sejam selecionados várias vezes, enquanto outros podem não ser escolhidos.

Aumento de dados com *Autoencoder*: Uma alternativa amplamente utilizada é o aumento de dados com *Autoencoder*. Esse método envolve a utilização de dois modelos, um codificador e um decodificador, normalmente construídos com redes neurais. O codificador recebe um exemplo do conjunto de dados e tenta representá-lo em um espaço de dimensão menor, conhecido como espaço latente. Em seguida, o *Autoencoder* pega essa representação latente, ou um ponto próximo, e a decodifica. O objetivo do decodificador é reconstruir o exemplo inicial que o codificador recebeu anteriormente.

A escolha do método GAN no processo de síntese de dados é justificada pelo rápido avanço que essa técnica tem experimentado em um curto período. Isso se deve, na maioria, à crescente comunidade de pesquisadores que estão desenvolvendo novas arquiteturas. Além disso, é importante observar que modelos de classificação têm apresentado melhorias na acurácia ao incluir dados sintetizados via GANs, em comparação com o conjunto de dados original ou através do método Bootstrap (Nakhwan and Duangsoithong, 2022).

Nessa etapa de sintetização dos dados, devemos levar em conta que nem todos os parâmetros que o aparelho entrega como saída são extraídos diretamente da imagem de fluorescência. A maioria dos parâmetros são calculados a partir de parâmetros básicos. Portanto, no processo de sintetização dos dados, devemos levar para a rede GAN apenas os parâmetros calculados diretamente da imagem e após a sintetização dos dados adicionais, calcular os demais parâmetros. Conforme comentado anteriormente, o aparelho seleciona uma área sobre a folha da planta onde as medidas serão calculas ponto a ponto para obter um valor médio (medidas com sufixo **_med** no nome) e variância (medidas como sufixo **_var** no nome). Como consequência desse método, cada parâmetro fica caraterizado por duas grandezas, média e variância. Para calcular os valores médias dos parâmetros derivados basta aplicar as fórmulas descritas na tabela 2. Porém, para calcular os valores da variância para os parâmetros derivados, devemos levar em consideração a propagação das incertezas data pela tabela 4.

Tabela 4 – Propagação de incerteza utilizada na reprodução das medidas que representam a variância do parâmetro, medidas do tipo "_var"

Função	**Variância**
$f = A - B$	$\sigma_f^2 \approx \sigma_A^2 + \sigma_B^2 - 2\sigma_{AB}$
$f = \frac{A}{B}$	$\sigma_f^2 \approx f^2[(\frac{\sigma_A}{A})^2 + (\frac{\sigma_B}{B})^2 - 2\frac{\sigma_{AB}}{AB}]$

Um ponto importante que devemos destacar, é que não conseguiremos calcular exata-

mente a covariância entre duas variáveis ($\sigma_{AB} = \rho_{AB}\sigma_A\sigma_B$), dado que não temos condições de calcular de forma precisa a correlação dessas variáveis (ρ_{AB}). Uma forma que encontramos para contornar o problema foi estimar este valor a partir do conjunto de dados inicial e aplicar no conjunto de dados sintetizados. Para isso fizemos uso da técnica de minimização da soma dos quadrados, onde o método **Nelder-Mead** apresentou os melhores resultados (Lagarias et al., 1998).

CAPÍTULO

4

RESULTADOS E DISCUSSÕES

4.1 Conjunto de dados

No início, não temos conhecimento suficiente para decidir sobre quais medidas são relevantes para o problema em análise. Pelo menos, não podemos afirmar com total certeza antes de realizar uma análise quantitativa dos dados. Portanto, é necessário iniciar com uma análise exploratória para obter uma compreensão mais aprofundada do conjunto de dados e entender como as métricas de fluorescência da clorofila podem interagir da melhor forma para determinar as classes de interesse no problema, ou seja, se a planta está saudável ou infectada.

A tabela 5 apresenta uma amostra dos dados organizados em um *dataframe* pandas mantendo praticamente o layout original do arquivo de saída do aparelho. Cada instância do conjunto de dados é representado por duas linhas consecutivas, as linhas de cores mais claras representam os valores médios do parâmetro e as linhas de cores mais escura representam a variância.

Algumas colunas no *dataframe* foram adicionadas com objetivo de marcar determinadas características do experimento como: **variedade** (SE ou LE), **tempo** a partir da inoculação da praga na planta até o instante que a imagem foi registrada e **rep**. A coluna **área** é a área extraída da figura original para geração das métricas. Essas métricas são as colunas seguintes começando por **fo** e terminado em **f740_f690** e **classe** se a planta foi infectada com a praga ou se trata de uma planta testemunha. Por questões de melhor visualização, a figura contém apenas uma pequena parte dos parâmetros disponíveis no *dataframe*.

4.2 Principais parâmetros de fluorescência da clorofila

Na recente história da análise da fluorescência da clorofila, uma vasta quantidade de diferente coeficientes foram propostos na tentativa de quantificar a redução fotoquímica e não fotoquímica em plantas. Os parâmetros que relatam a extinção fotoquímica sempre estão de

Tabela 5 – Amostra do conjunto de dados utilizado na análise organizado em um dataframe pandas apresentando linhas verdes para as médias do parâmetro e linhas vermelhas para identificar a variância

variedade	tempo	rep	área	fo	fm	fv	...	classe
SE	4.0	1.0	57601.0	149.55	477.24	327.7	...	1.0
SE	4.0	1.0	nan	39.24	101.73	74.46	...	1.0
SE	4.0	2.0	60559.0	163.41	407.76	244.35	...	1.0
SE	4.0	2.0	nan	49.96	114.24	74.99	...	1.0
SE	4.0	3.0	49969.0	205.81	698.68	492.87	...	1.0
SE	4.0	3.0	nan	53.96	177.4	128.77	...	1.0
SE	4.0	4.0	38167.0	156.16	473.48	317.32	...	1.0
SE	4.0	4.0	nan	45.77	120.7	88.38	...	1.0
SE	8.0	1.0	55504.0	317.85	597.31	279.46	...	1.0
SE	8.0	1.0	nan	80.63	142.9	70.4	...	1.0

alguma forma relacionados com o valor relativo de $F_m - F_o$. Entre todos os parâmetros dessa classe, o mais utilizado é o que mede a eficiência fotossintética do sistema II $\Phi_{PSII} = F_v/F_m$ (Genty et al., 1989). É considerado tão importante, pois mede a proporção da luz absorvida pela clorofila associada com o PSII utilizado na fotoquímica (Maxwell1 and Johnson2, 2000). No caso de uma planta saudável ela possui uma ótima capacidade fotossintética, uma ótima capacidade em converter energia luminosa em forma de energia fotoquímica. Por outro lado, quando a planta está passando por algum estresse, sua capacidade fotossintética torna-se comprometida. Assim é esperado um rendimento maior da fluorescência quando a via da fotossíntese está comprometida. Para a maioria de espécie de plantas, o valor ótimo para eficiência fotossintética fica em torno de **0.83** para uma planta saudável (Johnson et al., 1993).

Outro parâmetro bastante importante, **NPQ**, na qual quantifica a extinção não-fotoquímica da fluorescência, mede a proporção na mudança de Fm para o seu valor final e está linearmente relacionado com a dissipação de energia na forma de calor. Este parâmetro pode assumir valores que variam entre $0 - \infty$, porém em uma planta típica e sobre intensidades de luz de saturação os valores esperados devem estar próximo de **2.33** (Buschmann, 1999).

Tabela 6 – Valores médios dos parâmetros mais comum em análise de fluorescência: *QY_max* medindo a eficiência fotoquímica e *NPQ* medindo a eficiência não-fotoquímica para plantas infectadas com o Percevejo separada pela variedade.

Infestação	Variedade	*QY_max*	NPQ
percevejo	LE	0.740 ± 0.023	2.790 ± 0.826
	SE	0.677 ± 0.107	2.283 ± 0.476
lagarta	LE	0.787 ± 0.022	3.031 ± 0.989
	SE	0.774 ± 0.028	2.204 ± 0.607

A tabela 6 apresenta a média e variância para os parâmetros de performance mencionados acima. Como o objetivo aqui é comparar os valores obtidos em ambos os experimentos, percevejo e lagarta, com os valores encontrados na literatura, estamos omitindo a separação dos resultados por classe de planta (infectada ou controle) e agrupando todas em uma única linha. Isso, pois queremos ter uma ideia global da validade dos nossos dados e consequentemente maior segurança para prosseguir com a demais análises. Inicialmente podemos ver que os valores obtidos para *QY_max* estão a baixo do valor esperado (0.83), principalmente no experimento com a presença do percevejo. Essa diferença está muito relacionada a dificuldade encontrada em colocar as amostras em tempo de adaptação ao escuro necessária. Além disso, podemos comentar também a diferença notável que existe ao comparar o valor do parâmetro nos dois experimentos, apesar de se tratar de experimentos muito semelhantes, onde se aplicou o mesmo protocolo na aquisição dos dados, os experimentos ocorreram em épocas distintas, onde a diferença de temperatura pode ser um fator determinante nos resultados.

Seguindo na exploração dos dados, o próximo passo é analisar como esses parâmetros estão respondendo após stress introduzido com a infestação. Esperamos encontrar algum indicativo de que a presença do percevejo ou da lagarta tenha causado alterações significativas no funcionamento da planta a ponto de afetar o padrão de emissão da florescência. A questão aqui não é tentar encontrar um único parâmetro que resolva todo o problema, ao invés disso, dada as características do experimento e a complexidade do problema, o esperado é que um conjunto razoável de parâmetros trabalhando juntos tenha mais êxito nessa tarefa. Portanto, nesse momento o objetivo é destacar como determinados parâmetros podem auxiliar no entendimento do problema por meio de um apelo visual.

A figura 11 apresenta um gráfico de força comparando as médias relativas entre plantas saudáveis e plantas infectadas, para o parâmetro que reporta o desempenho do fotossistema PSII (QY). Ao analisar os dados de plantas infestadas pelo Percevejo, para ambos tipos de plantas (LE e SE), os gráficos do conjunto apresentam o mesmo comportamento, ocorrendo sobreposição das curvas dificultando a identificação de alguma separação visual. Porém, ao analisar os dados de plantas infectadas com a lagarta, é possível visualizar que a maioria dos valores são maiores para uma planta saudável quando se trata do tipo SE, indicando o potencial desse conjunto.

A figura 12 apresenta o mesmo gráfico de força, mas neste caso comparando as médias para os parâmetros que reportam a eficiência da dissipação de calor (NPQ), neste caso podemos

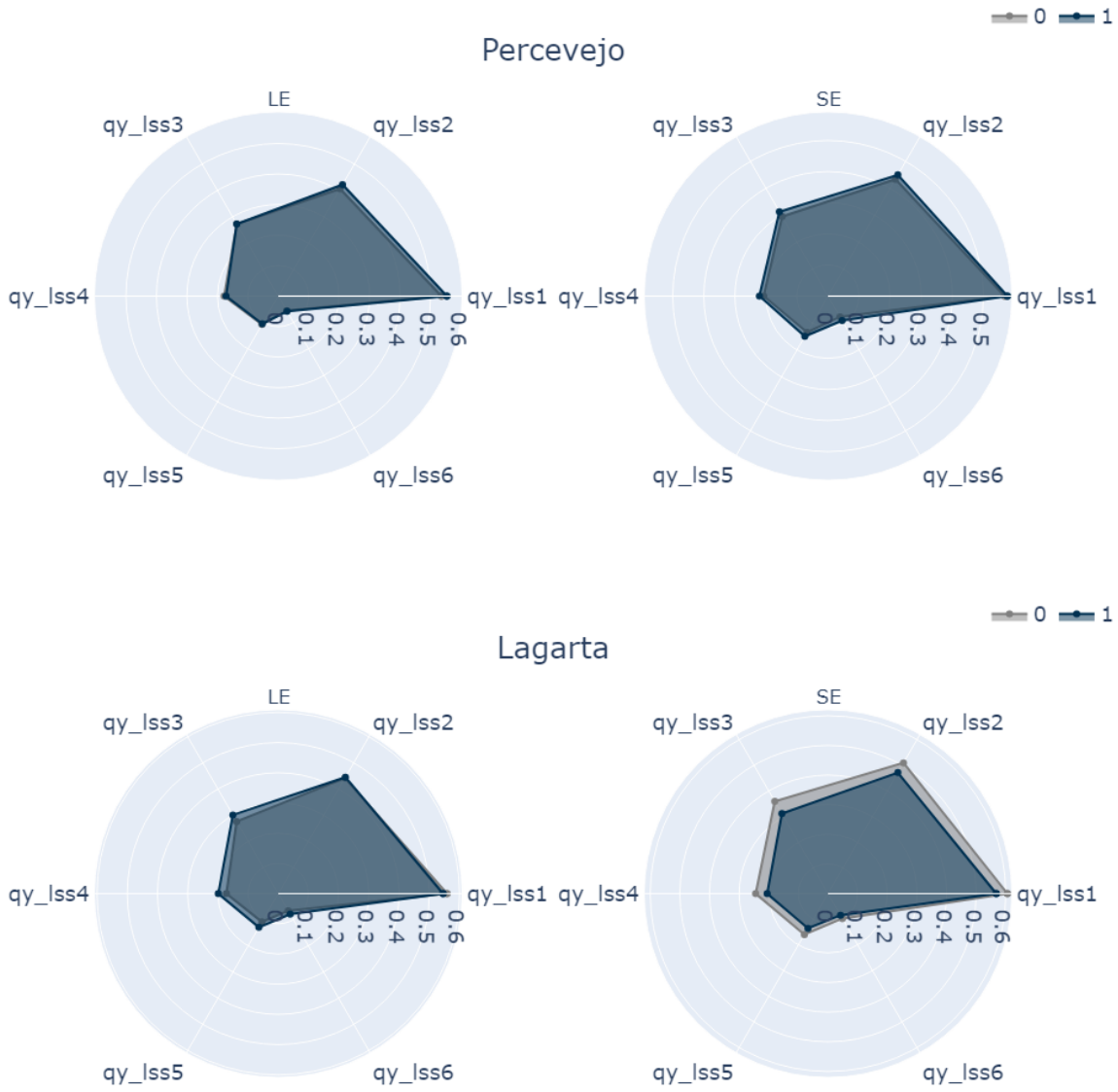


Figura 11 – Gráfico de força das médias relativas entre plantas saudáveis e plantas infectadas pelos dois tipos de infestação, com base nos parâmetros comumente utilizados em análise de fluorescência e segmentado pela variedade da planta(LE/SE)

visualizar que em ambos os tipos de infestação e para os dois tipo de plantas analisada, o comportamento apresentado é o mesmo, a maioria dos valores são maiores para uma planta saudável, com valores mais expressivos para plantas do tipo SE infectadas com o percevejo, indicando o potencial desse conjunto.

Nesse caso o resultado está mais homogêneo, o comportamento dos parâmetros segue a mesma tendencia independente do tipo de infestação ou da variedade da planta, apresentando valores mais altos para as amostras de controle e mostrando maior flexibilidade que o conjunto anterior, pois consegue generalizar para qualquer que seja a variedade da planta.

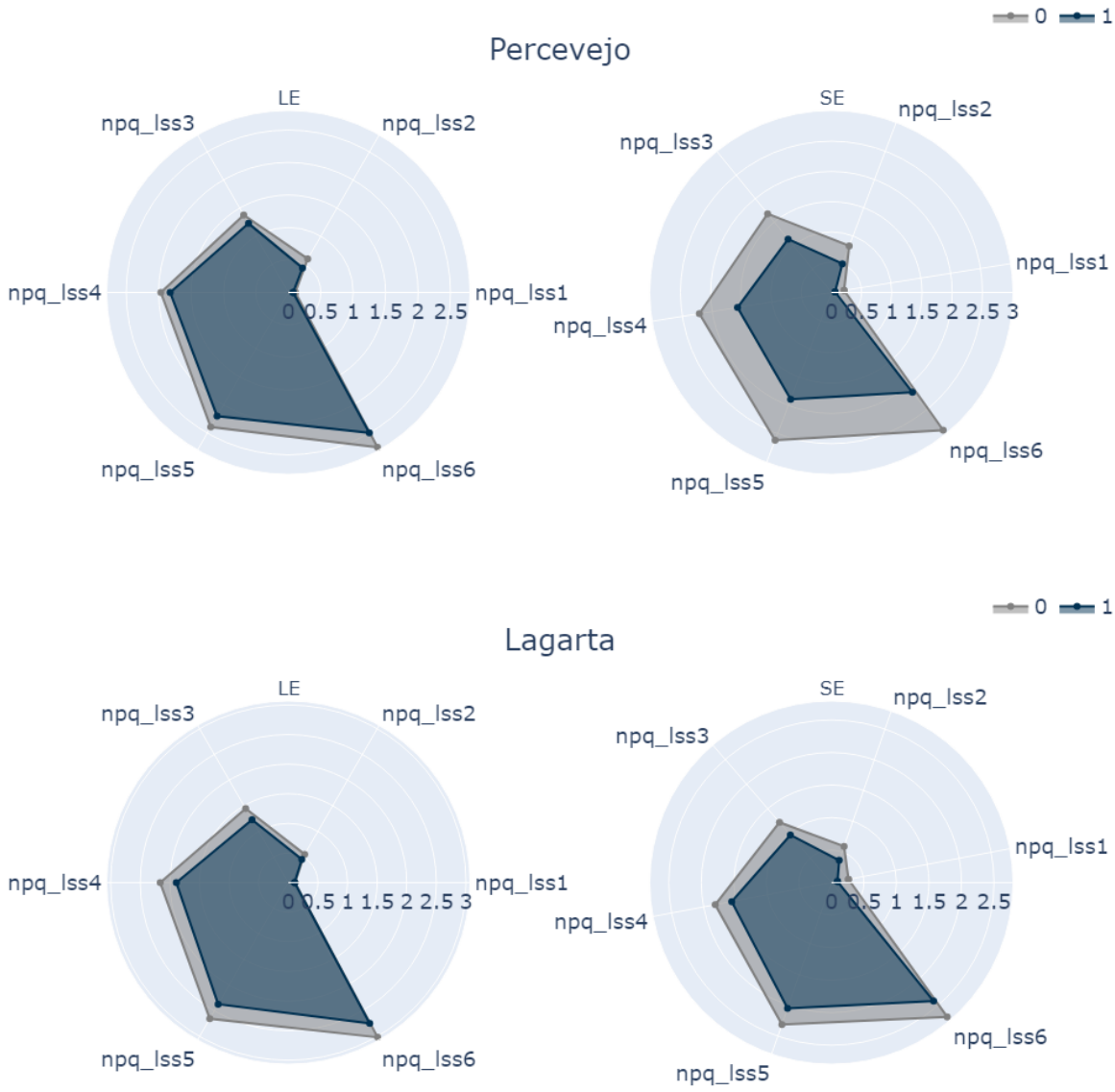


Figura 12 – Gráfico de força das médias relativas entre plantas saudáveis e plantas infectadas pelos dois tipos de infestação, com base nos parâmetros comumente utilizados em análise de fluorescência e segmentado pela variedade da planta(LE/SE)

Como visto nos exemplos anteriores, o apelo visual nos ajuda a conhecer melhor os dados que estamos analisando e se realmente estão atendendo nossas expectativas. Porém, seus resultados não são conclusivos, ainda depende da subjetividade de quem está analisando e dos métodos empregado na criação da visualização. Dessa forma, determinar qual métrica está separando melhor pelo simples exame de superioridade/inferioridade das médias não é aconselhável. Muitas vezes, as diferenças verificadas não são significativas, e podem-se considerar as médias obtidas equivalentes. Portanto, é necessário conduzir um teste de hipóteses para uma comparação dos resultados. A tabela 7 apresenta os resultados do teste de hipótese para as métricas de desempenho. Neste caso temos como hipótese nula que as médias dos conjuntos podem ser consideradas equivalentes.

Tabela 7 – Teste T para hipótese nula que o conjunto de exemplos de plantas saudáveis e o conjunto de plantas infectadas possuem a mesma média.

Percevejo				
	LE		SE	
	estatística	valor-p	estatística	valor-p
qy_lss1	-2.2558	**0.0285**	-0.5496	0.5853
qy_lss2	-1.5768	0.1212	-0.8917	0.3772
qy_lss3	-0.0767	0.9391	-0.9495	0.3473
qy_lss4	0.5053	0.6155	-0.8203	0.4163
qy_lss5	0.1816	0.8566	-0.9747	0.3348
qy_lss6	-0.3004	0.7651	-0.9607	0.3417
npq_lss1	3.0758	**0.0034**	1.2174	0.2297
npq_lss2	3.8202	**0.0004**	1.2667	0.2116
npq_lss3	1.8129	0.0758	1.4885	0.1434
npq_lss4	1.2656	0.2115	1.5217	0.1349
npq_lss5	1.4215	0.1614	1.5908	0.1185
npq_lss6	1.5548	0.1263	1.6680	0.1021
Lagarta				
	LE		SE	
	estatistica	valor-p	estatistica	valor-p
qy_lss1	1.8245	0.0704	7.5176	**0.0000**
qy_lss2	-0.3863	0.6999	5.8893	**0.0000**
qy_lss3	-2.8511	**0.0051**	4.9727	**0.0000**
qy_lss4	-3.3467	**0.0011**	4.3912	**0.0000**
qy_lss5	-3.4037	**0.0009**	3.5865	**0.0005**
qy_lss6	-3.1754	**0.0019**	2.7535	**0.0068**
npq_lss1	0.6201	0.5363	1.1120	0.2682
npq_lss2	2.9994	**0.0033**	1.0090	0.3149
npq_lss3	3.1787	**0.0019**	0.9147	0.3621
npq_lss4	2.6542	**0.0090**	0.8250	0.4109
npq_lss5	2.1932	**0.0301**	0.7507	0.4542
npq_lss6	1.8464	0.0672	0.8101	0.4194

Considerando a população de plantas saudáveis e plantas infectadas e um intervalo de confiança de 5%, podemos afirmar que existe diferença entre as médias dos conjuntos somente quando consideramos as métricas que reportam a eficiência do fotossistema PSII (qy_lss) no conjunto de dados da lagarta quando olhamos para plantas do tipo SE. Nos demais casos não podemos afirmar nada dado que o *p-value* desses casos ultrapassam o limiar escolhido.

4.3 Importância da variância

Naturalmente tendemos a considerar como atributos válidos para entrada do modelo apenas os valores médios dos parâmetros, porém um estudo pre-liminar mostrou que o uso da variância acrescenta melhoras significantes na determinação de plantas infectadas. Para isso foi necessário combinar as médias e variâncias dos parâmetros de fluorescência em um único objeto, ou seja, em uma única linha para compor o conjunto de dados, portanto foi acrescentado os sufixos **_med** e **_var** nos nomes dos atributos para ser possível identificar quem representa média e quem representa a variância, respectivamente. A tabela 8 sintetiza esse estudo apresentando as acurácias da classificação no conjunto de dados gerado no experimento com percevejo. Os resultados foram obtidos utilizando árvores de decisão simples, sem normalização dos atributos e utilizando *leave-one-out* para o método de amostragem.

Tabela 8 – Acurácia da classificação utilizando árvore de decisão simples para dados do experimento com percevejo, na escala original, mostrando a importância da variância como atributo de entrada no modelo

	Média	Variância	Média+Variância
LE/SE	57,6	58,7	62,1
LE	71,8	80,1	76,6
SE	61,6	44,2	47,2

Na tabela podemos ver que em alguns casos o uso da variância em contraste com a média apresentou uma melhor performance, como é o caso para o tipo de planta LE, porém há caso que a performance teve uma piora, como as plantas do tipo SE. E por fim, como resultado geral, a combinação dos dois tipos de medida (média e variância) apresentou ganhos positivo.

Uma forma de visualizar a importância desses atributos é apresentado pela figura 13 onde a árvore de decisão foi criada após treinamento incluindo todos os atributos do experimento com percevejo e planta da variedade LE. Podemos observar que a árvore é formada por uma mistura de ambos os tipos de medidas.

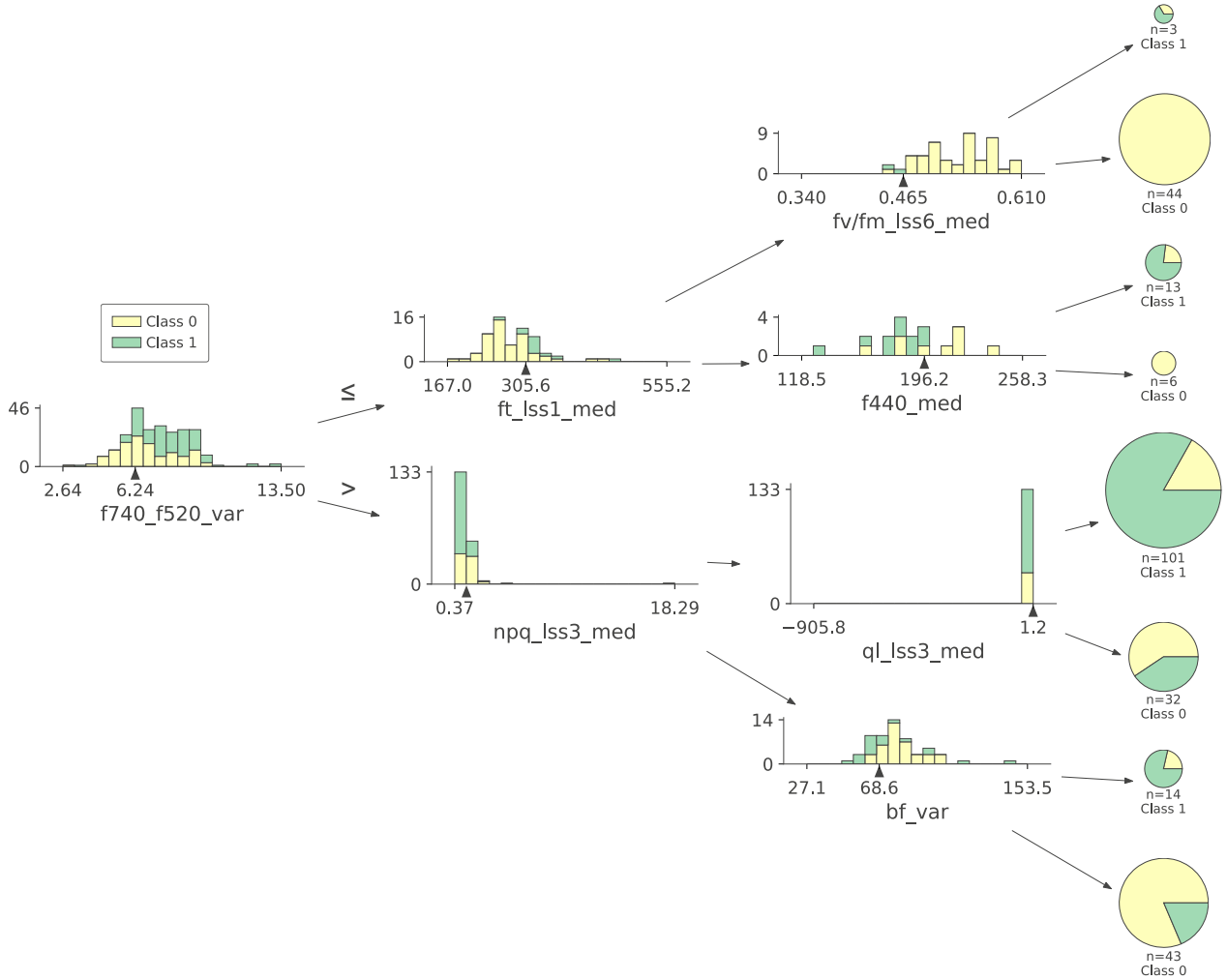


Figura 13 – Árvore de decisão criada após treinamento incluindo todos 172 atributos do experimento com percevejo e planta da variedade LE

4.4 Dependência entre parâmetros

Apesar da grande quantidade de parâmetros quando comparado com a quantidade de instâncias do problema, muitos desses parâmetros podem ser eliminadas, pois estão correlacionadas de alguma forma, algumas possuem dependência linear por serem geradas como combinação de outros, como as métricas derivadas, outras possuem dependência devido à semelhança na forma que o sinal foi gerado para produzir a imagem. Essa dependência entre variáveis é possível ser visualizada através da matriz de correlação calculada pelo método de Pearson como mostra a figura 14. Devido à quantidade de parâmetros, foi omitido os rótulos nos eixos do gráfico com a finalidade de manter boa visualização.

A partir da matriz de correlação é possível identificar no gráfico regiões próximas com cores vermelhas, essas são as medidas que possuem correlação próximo do valor 1, e regiões próximas com cores azuis, medidas que possuem correlação próximo de -1. Essas regiões mostram alto grau de associação linear entre as medidas e indicam que podemos eliminar

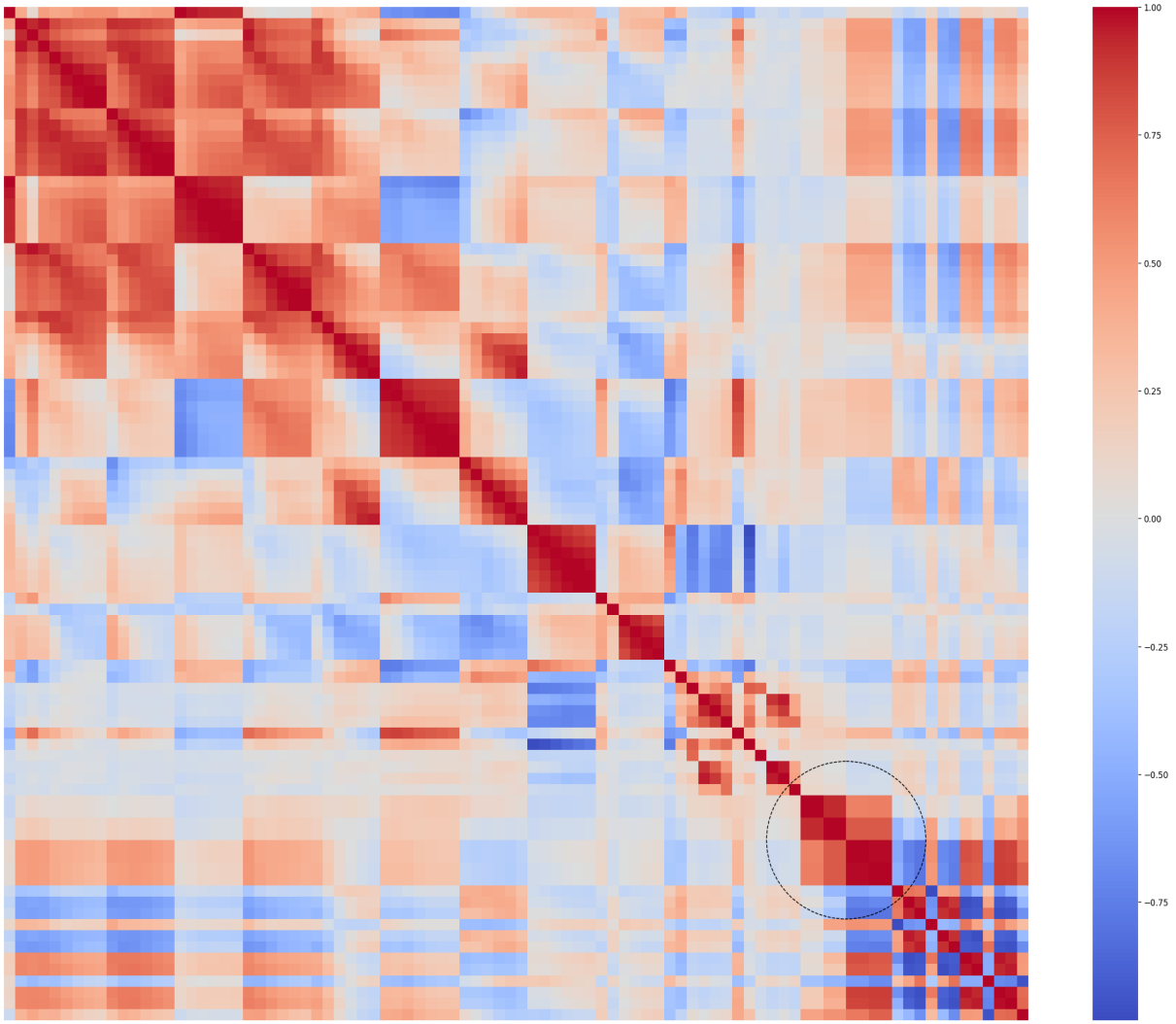


Figura 14 – Correlação do parâmetros da florescência da clorofila através do método Person aplicado no conjunto do Percevejo

algumas delas sem perda significativa de informação.

Ao dar um zoom no canto inferior direito da imagem, região delimitada pelo círculo tracejado, podemos identificar um grupo relevante envolvendo as medidas **f440**, **f520**, **f690**, **f740**. A figura 15 exibe a matriz de correlação desse grupo, deixando mais evidente a existência de correlação entre essas variáveis. De modo geral, o conjunto em destaque apresenta uma forte dependência entre os valores, no entanto, tal situação fica ainda mais evidente próximo da diagonal principal, como é caso das combinações (f440, f520) e (f690, f740).

A forte correlação nesses casos é facilmente explicada com ajuda do espectro da luz na região do visível. Na primeira combinação, temos as cores azul e verde, que são cores consecutivas no espectro de luz visível. Na segunda combinação, ainda mais próximas no espectro, estão o vermelho e o vermelho distante (conforme ilustrado na Figura 16).

Embora essa análise não produza resultados diretos, ela apresenta evidências substanciais de que é possível reduzir o número de atributos necessários para definir cada instância e manter a informação inicial praticamente inalterada.

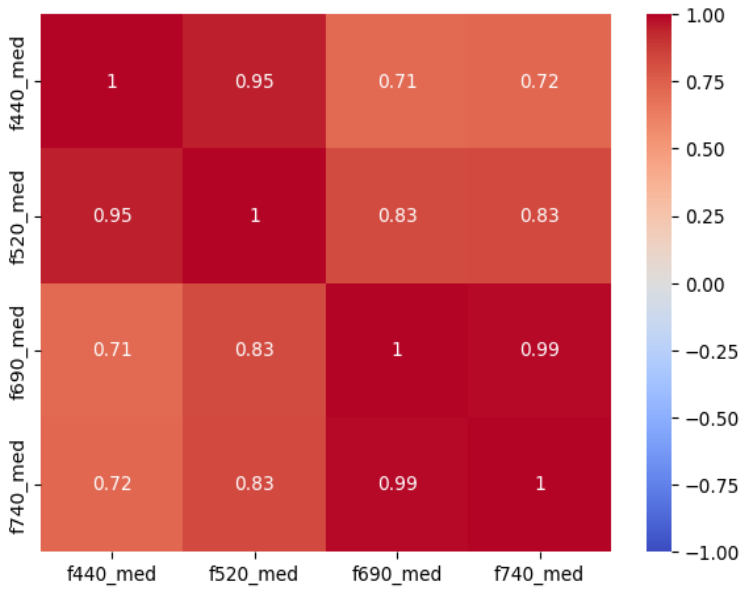


Figura 15 – Matriz de correlação de Person para medidas que estão fortemente correlacionadas.

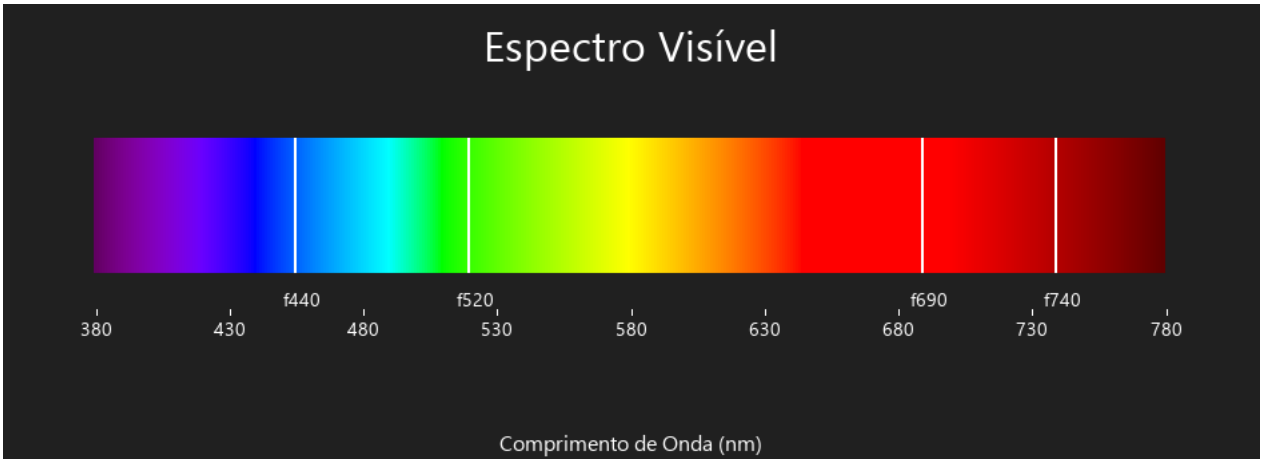


Figura 16 – Espectro da luz na região do visível contendo a marcação das tuplas com maior dependência linear apresentada pela correlação.

4.5 Formação de grupos

Para avaliar a formação de grupos, primeiramente aplicamos PCA sobre os dois conjunto de dados, que foram previamente normalizados separadamente através do método standardscaler, em seguida avaliamos a importância de cada componente principal com base na quantidade de variação explicada por cada uma delas. Selecionamos, então, as duas primeiras componentes com a maior variação explicada e geramos um gráfico no plano XY, uma abordagem que proporciona uma representação visual dos grupos identificados pelo modelo. Na Tabela 9, encontramos uma listagem dos componentes principais, que exibe a variação individual e a variação acumulada das componentes que, juntas, abrangem 80% da variação total do conjunto.

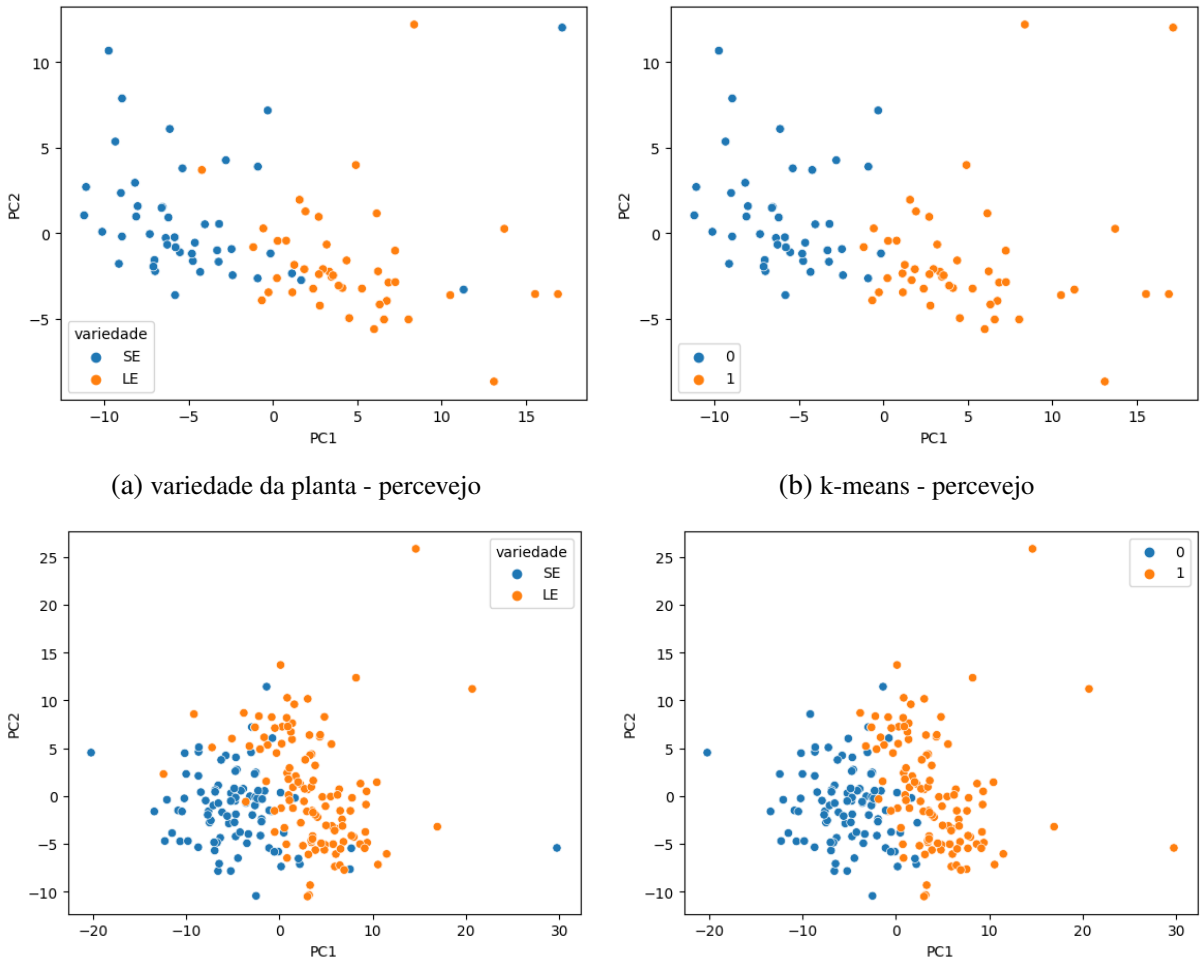
Com o propósito de comparar as estruturas identificadas pelo algoritmo com estruturas previamente conhecidas, realizamos variações nas marcações do conjunto de dados, usando

Tabela 9 – Listagem dos componentes principais, apresentando a variação e variação acumulada dos componentes que juntos detém 80% da variação.

Percevejo		
Componente	**Variação (%)**	**Variação Acumulada (%)**
PC1	25.86	25.86
PC2	17.72	43.57
PC3	11.82	55.40
PC4	8.16	63.55
PC5	5.57	69.12
PC6	5.00	74.13
PC7	3.97	78.09
PC8	3.20	81.29
Lagarta		
Componente	**Variação (%)**	**Variação Acumulada (%)**
PC1	22.92	22.92
PC2	16.78	39.70
PC3	11.34	51.04
PC4	9.18	60.23
PC5	5.08	65.31
PC6	4.84	70.15
PC7	3.49	73.64
PC8	3.32	76.96
PC9	2.99	79.95
PC10	2.32	82.31

algumas das características mencionadas anteriormente na seção 4.1. Em resumo, conduzimos testes de plotagem de gráficos com base na variável alvo (classe), tempo de coleta dos dados (tempo) e a variedade da planta (variedade). Na maioria dos casos, não observamos evidências de formação de grupos, havendo uma considerável sobreposição dos dados. A exceção a esta tendência foi observada com relação à variedade das plantas, como pode ser visto nas Figuras 17(a) e 17(c), onde é possível identificar a formação de dois grupos com pouca sobreposição entre as classes.

Em seguida, aplicamos o método K-Means a cada conjunto reduzido de componentes principais (conforme mostrado na Tabela 9), resultando em um vetor que atribui uma classe a cada instância do conjunto de dados. Como configuramos o número de grupos para 2, as classes possíveis são 0 ou 1. As Figuras 17(b) e 17(d) apresentam as estruturas encontradas pelo algoritmo. Os resultados obtidos nessa análise, além de validar a eficácia do algoritmo na identificação de estruturas que inicialmente podem estar ocultas, no caso, uma estrutura que segmenta a variedade da planta, os resultados são importantes porque oferecem a possibilidade de criar um classificador único capaz de generalizar para ambas as variedades da planta (LE e SE).



(a) variedade da planta - percevejo

(b) k-means - percevejo

(c) variedade da planta - lagarta

(d) k-means - lagarta

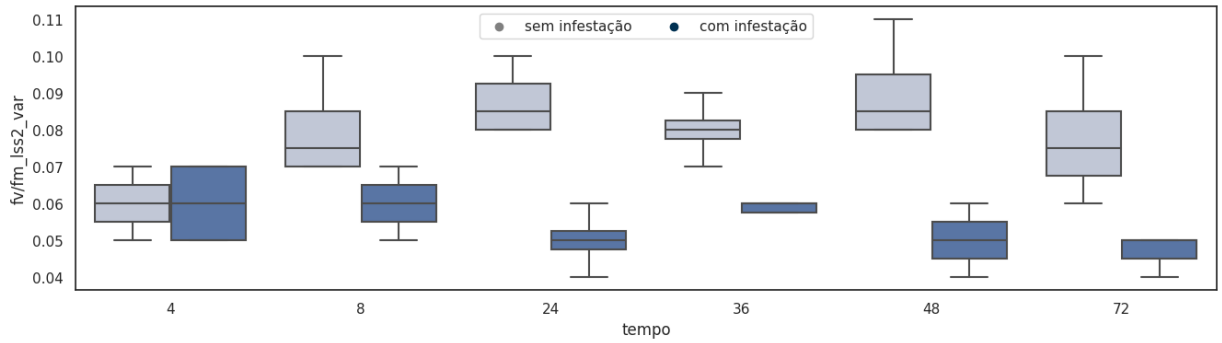
Figura 17 – Formação de grupos com base nos dois primeiros componentes principais: (a-b) para os dados do percevejo e (c-d) para os dados da lagarta. Comparações entre grupos identificados com base nas variedades das plantas e grupos identificados por meio do algoritmo K-Means.

4.6 Evolução no tempo

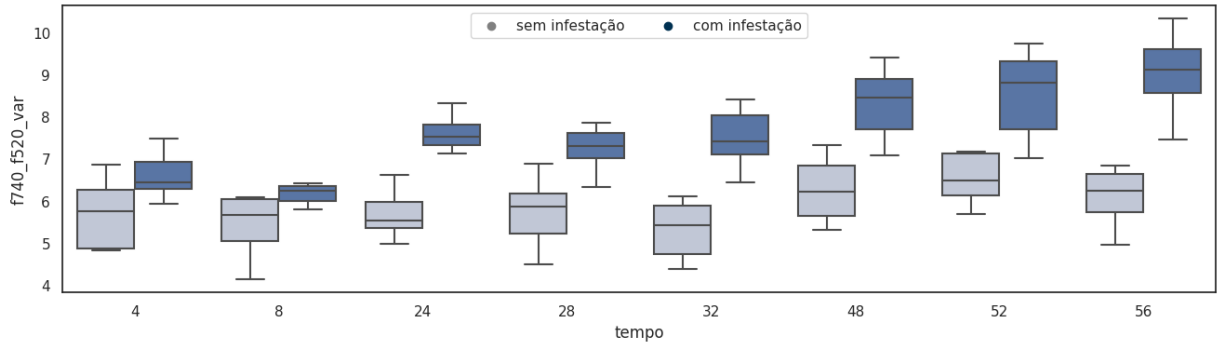
Além do nosso interesse central em desenvolver métodos que permitam o diagnóstico precoce do ataque das pragas antes que os sintomas sejam visíveis aos olhos, também estamos preocupados em determinar quando os efeitos da infestação começam a se manifestar. Sabemos que esses efeitos não ocorrem de imediato; cada planta possui seu próprio tempo de resposta e, em muitos casos, a doença é assintomática nos primeiros dias da infestação. Devido à natureza controlada da infestação neste experimento e à coleta de medidas em intervalos específicos, temos a oportunidade de estudar como a doença se desenvolve ao longo do tempo, analisando a evolução dos parâmetros de fluorescência da clorofila ao longo do período de observação.

Após tratamento dos *outliers*, foi realizado uma análise da evolução dos parâmetros no tempo através das suas distribuições. Utilizando gráficos de *boxplot*, separamos as classes em duas distribuições distintas evoluindo no tempo. Como resultado, podemos observar que

alguns parâmetros apresentaram separação das classes com poucas horas após infestação, por exemplo, o parâmetro **fv/fm_lss2_var** que começa separar as classes de plantas saudáveis e plantas infectadas com percevejo em apenas 8 horas, apresentando separações máximas nos tempos 24 e 48 horas. No caso da infestação com a lagarta o parâmetro **f740_f520_var** começa apresentar separação visual após 24 horas da infestação, com máximas em 24 e 56 horas (figura 18).



(a) Percevejo

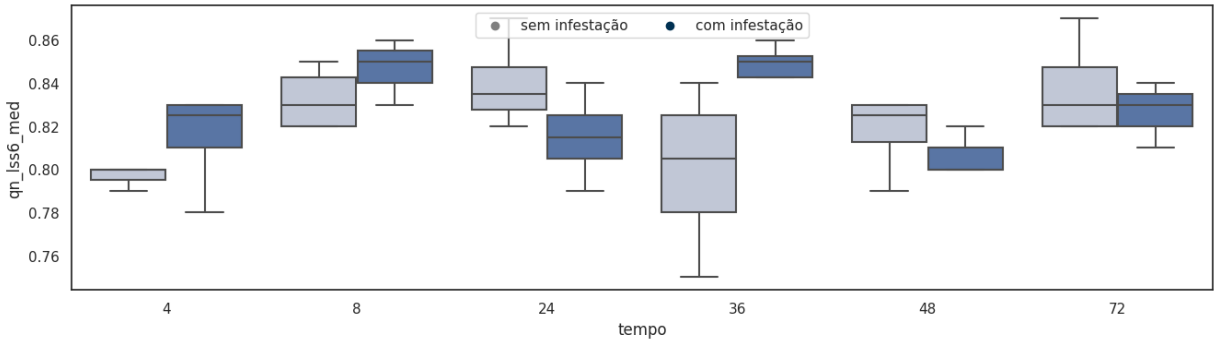


(b) Lagarta

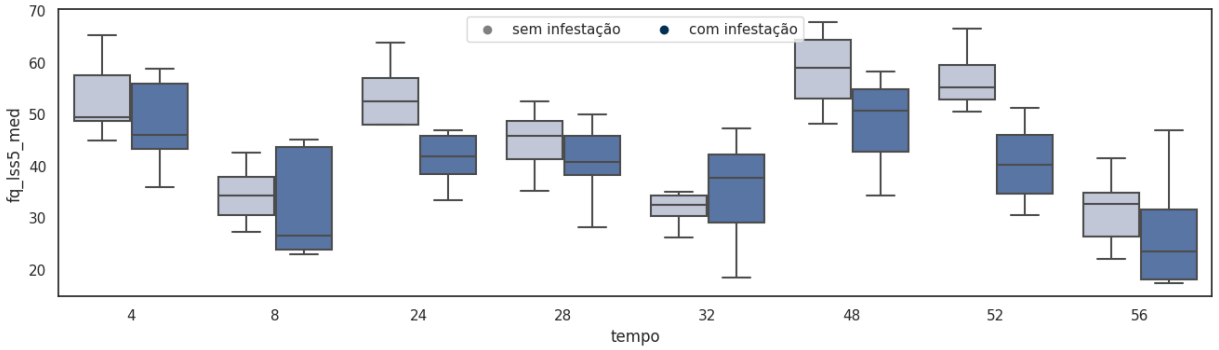
Figura 18 – Boxplot que ilustra a evolução da distribuição ao longo do tempo, destacando parâmetros que apresentam mudanças ocorridas 8 horas após a infestação pelo percevejo e 24 horas após a infestação pela lagarta.

Durante a seleção dos parâmetros, observamos um padrão comum na evolução temporal que se aplica a ambos os conjuntos de dados. A maioria dos parâmetros que demonstram separação visual e mantêm consistência ao longo do período observado pertence à categoria '_var', ou seja, são medidas que representam a variância do parâmetro. Por outro lado, as medidas do tipo '_med', que representam a média do parâmetro, além de raramente apresentarem separações visuais, exibem um comportamento com um padrão recorrente, como ilustrado na Figura 19.

Há uma suspeita inicial de que o padrão recorrente esteja de alguma forma relacionado ao ciclo circadiano da planta, na qual confere uma variação nas funções biológicas em intervalos regulares de aproximadamente 24hs (Venkat and Muneer, 2022). Ao analisar os padrões gerados a partir dos dados do percevejo, é possível identificar a presença de dois mínimos consecutivos nos momentos 4 e 24hs, assim como dois máximos consecutivos nos momentos 8 e 36hs. Nos dados



(a) Percevejo



(b) Lagarta

Figura 19 – Boxplot ilustrando evolução no tempo, destacando parâmetros da fluorescência da clorofila que apresentam padrão recorrente, e que na maior parte estão representados pelas medidas do tipo '_med'

da lagarta, os primeiros mínimos ocorrem nos tempos 8 e 32hs, com duas máximas ocorrendo nos momentos 24 e 48hs. É importante notar que os intervalos observados nos experimentos não concordam precisamente com o intervalo de 24hs observado na literatura. Essas divergências estão provavelmente associadas a limitação dos experimentos e dificuldade em coletar os dados de forma mais tempestiva e em intervalos regulares.

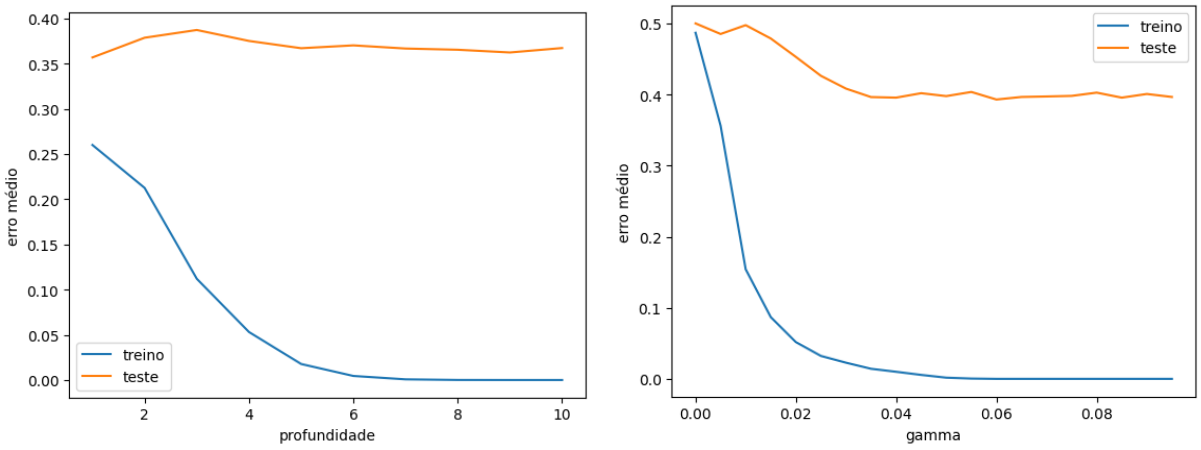
4.7 Generalização

A generalização refere-se a capacidade de um modelo de aprendizado de máquina se comportar bem em dados não utilizados no processo de treinamento. Geralmente, um modelo tende a ter um desempenho melhor na generalização quando é treinado com uma grande quantidade de dados de treinamento. Isso porque o modelo tem mais exemplos para aprender e ajustar seus parâmetros. Em uma situação com poucos dados de treinamento, os modelos podem ser propensos ao *overfitting*, situação quando o modelo se ajusta demais aos dados de treinamento, capturando o ruído e não consegue generalizar bem. Por outro lado, quando uma grande quantidade de dados é utilizada no treinamento, é menos provável que ocorra o *overfitting*, pois o modelo tem a oportunidade de aprender relações mais significativas.

Conforme discutido na Seção 4.3, já conduzimos análises iniciais utilizando modelos

simples de aprendizado de máquina. Essas análises foram úteis para confirmar a relevância dos atributos que representam a variância do parâmetro de fluorescência na previsão do atributo alvo. No nosso estudo, empregamos uma abordagem básica de árvore de decisão, sem realizar otimizações nos hiper-parâmetros para aprimorar os resultados, pois o objetivo não era aprimorar a acurácia e sim comparar os dois grupos de atributos disponíveis. De qualquer forma, o modelo apresentou boa generalização, apresentando acurácias próximas de 80% nos dados de teste. Apesar dos resultados apresentados, temos poucas ferramentas para tentar melhorá-los. Principalmente devido à quantidade limitada de dados disponíveis para treinar o modelo.

Visando aprofundar nossa compreensão do desempenho de generalização de nossos modelos, incorporamos o modelo SVM como um segundo classificador em nosso estudo. Essa inclusão visa fornecer mais *insights* que orientem melhorias por meio da comparação entre o erro de teste e o erro de treinamento. Alguns hiper-parâmetros do modelo desempenham um papel crucial na transição de um modelo que sofre *underfitting* para um modelo que sofre *overfitting*. Buscamos alcançar um equilíbrio adequado entre esses dois extremos. Para adquirir esse conhecimento, plotamos uma curva conhecida como “curva de validação”. Para controlar o *trade-off* entre underfitting e overfitting, ajustamos o hiper-parâmetro de profundidade na árvore de decisão e o hiper-parâmetro *gamma* no SVM, conforme ilustrado na Figura 26.

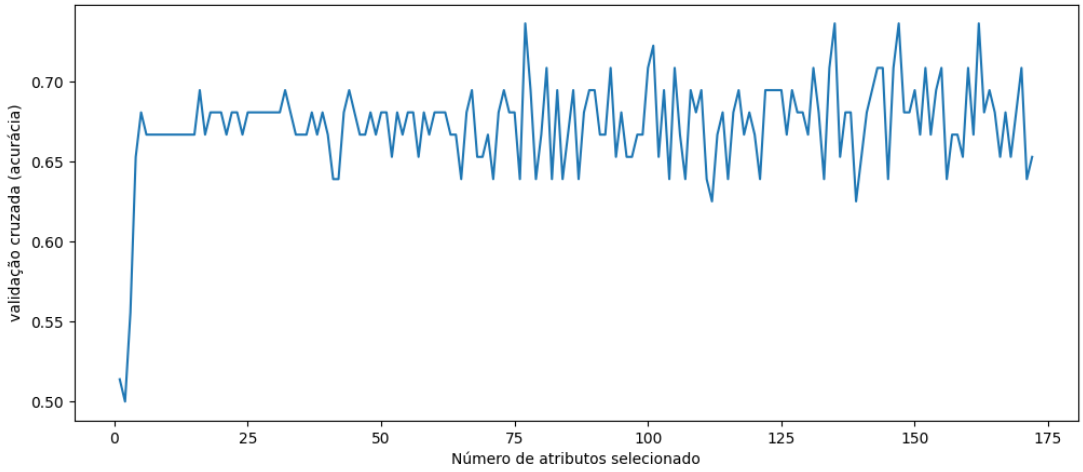


(a) Arvore de Decisão (b) SVM

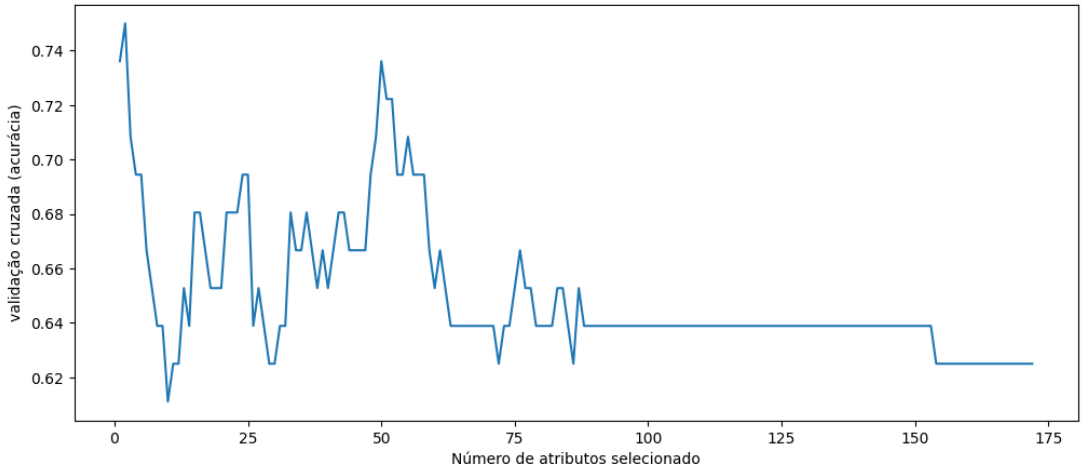
Figura 20 – Curvas de validação para análise de desempenho da generalização

Podemos ver que para valores baixos dos hiper-parâmetros, ambos modelos fazem um subajuste. O erro do treino e por consequência o erro do teste são ambos altos. Nesses casos os modelos estão tão restritos que não conseguem capturar a variabilidade que existe na variável alvo. Por outro lado, para elevados dos hiper-parâmetros, os modelos fazem um superajuste dos dados. O erro do treino se torna muito baixo enquanto o erro no teste aumenta. Essa situação sugere que os modelos criam decisões específicas para amostra ruidosas, prejudicando a capacidade de generalização para os dados de teste. Não é possível identificar alguma oportunidade para reduzir o erro apenas variando os hiper-parâmetros.

A dificuldade em generalização dos modelos traz outras consequências que impactam nas demais etapas do pipeline de dados, principalmente as que dependem de uma quantidade significativa de exemplos para atingir um bom desempenho. Um exemplo que pode ser utilizado para ilustrar é a redução de variáveis através da técnica RFE utilizada na seleção de atributos. O comportamento esperado ao aplicar a técnica é um aumento gradual na acurácia conforme mais atributos são adicionados no modelo, esperamos que esse aumento atinga um valor máximo e após isso, mantenha-se constante a partir de uma certa quantidade de atributos. Porém, dada a baixa quantidade de exemplos, o comportamento difere do esperado, conforme apresentado nas figuras 21.



(a) Arvore de Decisão



(b) SVM

Figura 21 – Evolução da acurácia com aumento na quantidade de atributos no modelo.

No caso da árvore de decisão, ainda podemos observar um aumento no desempenho do modelo, embora não seja um aumento gradual. Inicialmente, o modelo apresenta um comportamento aleatório quando está sendo treinado com poucos atributos. No entanto, observamos um aumento rápido na acurácia quando adicionamos alguns atributos. No caso do modelo SVM, o comportamento difere significativamente das expectativas. Inicialmente, o modelo apresenta

alta acurácia com poucos atributos, mas à medida que aumentamos a quantidade de atributos, o desempenho do modelo começa a piorar.

4.8 Engenharia de atributos

A análise exploratória do conjunto de dados forneceu as informações necessárias para iniciar o processo de construção do algoritmo que será utilizado como nosso classificador final. Antes de chegarmos à etapa de classificação propriamente dita, é necessário que os dados passem por alguns tratamentos a fim de torná-los mais adequados para servirem como entrada nos modelos de indução. Essa fase, conhecida como pré-processamento ou engenharia de atributos, envolve normalmente as etapas de normalização, tratamento de dados faltantes, balanceamento, conversão de dados simbólicos em numéricos e redução de variáveis.

Devido às características controlada dos nossos experimentos, algumas etapas de pré-processamento não são necessárias, como tratamento de dados faltantes ou conversão simbólico-numérico. No entanto, neste trabalho, estamos avaliando uma nova etapa, a inclusão do *data augmentation* como forma de contornar alguns problemas que surgem quando estamos lidando com um conjunto de dados reduzido, técnica pouco explorada nos pipelines de dados tabular.

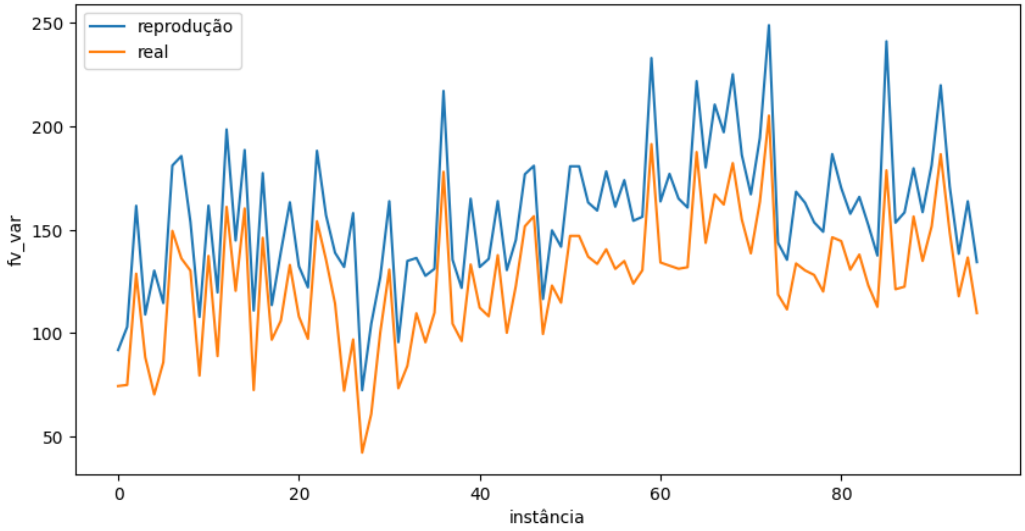
4.8.1 Data Augmentation

Visando aprimorar a capacidade de aprendizado de nossos modelos, complementamos os dados obtidos a partir dos experimentos com dados sintéticos gerados por uma Rede Generativa Adversarial (GAN). Essa abordagem permite a replicação dos componentes estatísticos dos dados reais, contribuindo para a melhoria do desempenho dos modelos.

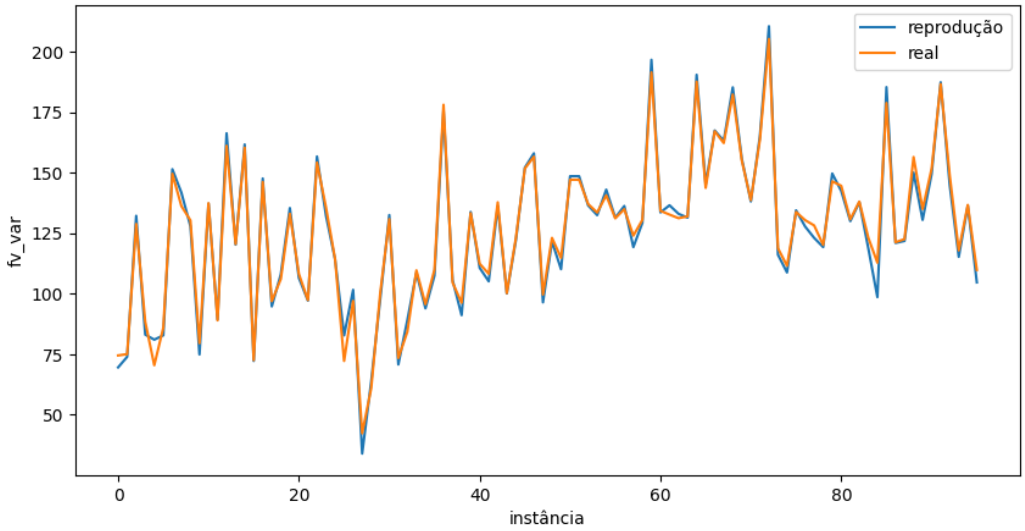
Considerando todos os atributos do problema, temos um total de 172 atributos que descrevem a instância do problema. Dentre esses, 86 atributos representam as médias e 86 atributos representam as variâncias dos parâmetros de fluorescência da clorofila. Além disso, há um atributo alvo que identifica a classe da instância. Durante o processo de geração de dados sintéticos, optamos por não incluir atributos que são derivados de outros atributos (conforme Tabela 2). A justificativa para essa escolha reside no fato de que, ao tentar sintetizar esses valores, poderíamos encontrar situações em que o cálculo da variável, a partir dos dados sintéticos, não coincidiria com o valor gerado pelo processo de síntese original, resultando em inconsistências. Portanto, nossa abordagem consiste em gerar valores apenas para os atributos básicos e, em seguida, calcular os atributos derivados.

Após sintetização dos atributos básicos, a reprodução dos atributos derivados que representam as médias é necessária apenas aplicar as fórmulas descritas na tabela 2. Porém, para reprodução dos atributos que representam a variância devemos considerar a propagação da incerteza (Tabela 4). Como não temos condições de calcular a correlação entre dois atributos de forma precisa, pois nesse caso precisaríamos ter acesso aos valores de cada ponto mapeado dentro da área selecionado pelo aparelho, seguimos duas abordagens. A primeira, seguir com

a propagação do erro considerando as correlações calculadas na Figura 14, na qual traz uma estimativa subestimada da correlação real, pois neste caso a correlação está sendo calculada sobre uma média de valores. Na segunda abordagem fizemos as estimativas dos valores através do método Nelder-Mead, na qual apresentou melhor resultado ao reproduzir os atributos do conjunto de dados reais, conforme ilustrado na Figura 22.



(a) abordagem utilizando média



(b) abordagem baseada em otimização

Figura 22 – Abordagens para estimativa da correlação entre variáveis: (a) abordagem utilizando correlação calculada a partir da média. (b) abordagem baseada em otimização da soma do quadrado do erro (Nelder-Mead)

Empregamos uma GAN do tipo WGAN-GP (Wasserstein Generative Adversarial Networks with Gradient Penalty) para gerar os atributos básicos dos parâmetros de fluorescência. A rede geradora foi configurada com 5 camadas de neurônios e empregou a função de ativação Leaky-ReLU. Quanto à rede discriminadora, esta consistiu em 7 camadas de neurônios, intercalando entre camadas densas e camadas de dropout. Utilizamos LeakyReLU como função de ativação

nas camadas ocultas e a função sigmoide na camada final. A figura 34 apresenta o sumário de cada rede.

Layer (type)	Output Shape	Param #
input_5 (InputLayer)	[(12, 32)]	0
dense_16 (Dense)	(12, 64)	2112
dense_17 (Dense)	(12, 128)	8320
dense_18 (Dense)	(12, 256)	33024
dense_19 (Dense)	(12, 48)	12336

Total params: 55792 (217.94 KB)
Trainable params: 55792 (217.94 KB)
Non-trainable params: 0 (0.00 Byte)

(a) tabela de resuma da rede geradora

Layer (type)	Output Shape	Param #
input_6 (InputLayer)	[(12, 48)]	0
dense_20 (Dense)	(12, 256)	12544
dropout_4 (Dropout)	(12, 256)	0
dense_21 (Dense)	(12, 128)	32896
dropout_5 (Dropout)	(12, 128)	0
dense_22 (Dense)	(12, 64)	8256
dense_23 (Dense)	(12, 1)	65

Total params: 53761 (210.00 KB)
Trainable params: 53761 (210.00 KB)
Non-trainable params: 0 (0.00 Byte)

(b) tabela de resuma da rede discriminadora

Figura 23 – tabela de resumo das redes utilizadas na estrura da GAN

No processo de treinamento, utilizamos lotes de tamanho 12 (batch size), com um limite de 4200 épocas. Adotamos a técnica de normalização por lote para mitigar variações na covariância interna e acelerar o processo de aprendizado. Obtivemos resultados satisfatórios ao estabelecer a taxa de aprendizado em 5e-4.

Para ilustrar o funcionamento da Rede Generativa Adversária (GAN), a Figura 24 apresenta a evolução no aprendizado do algoritmo comparando os dados reais com dados

sintéticos criados a cada etapa do aprendizado. Na figura é apresentado a relação dos parâmetros f44, ft_lss1, fm_lss1, fo_lss1 e fm em função do parâmetro fo. Nas primeiras execuções (passo 0 e 100) os dados artificias possuem uma distribuição muito distante dos dados reais, concentrando principalmente no ponto médio dos dados reais. Conforme iteramos o algoritmo, a distribuição dos dados artificias vai se aproximando da distribuição que os dados reais possuem (step 800).

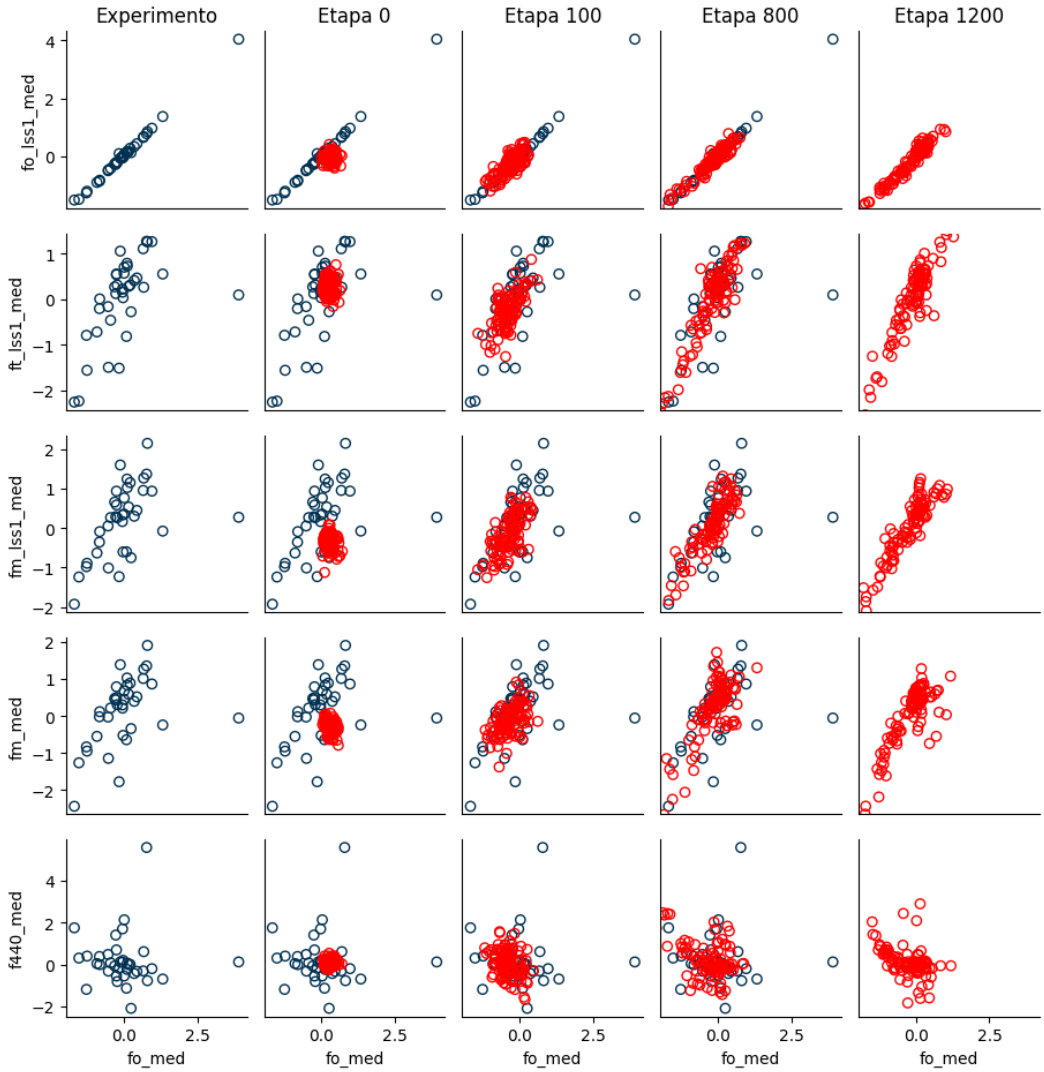
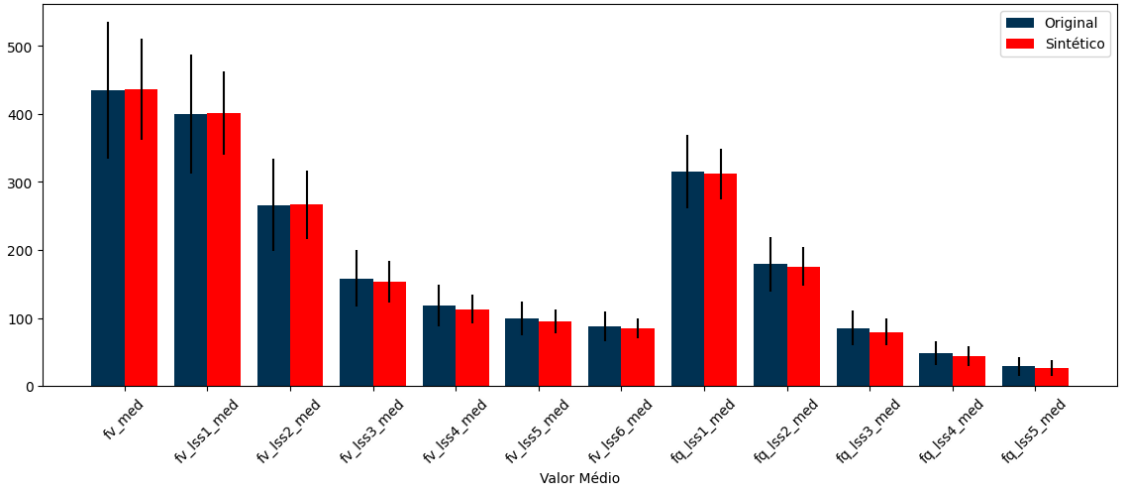


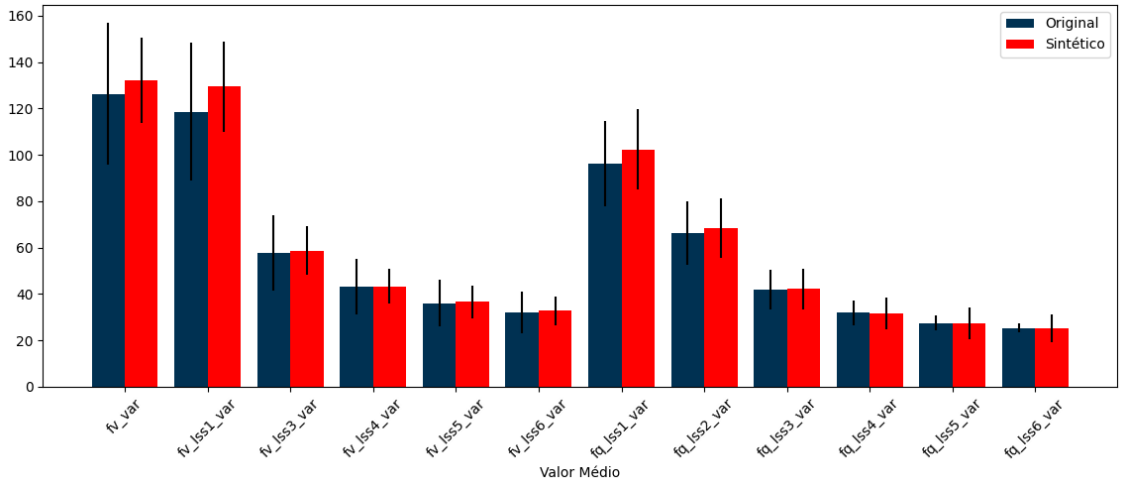
Figura 24 – Evolução no aprendizado da rede em gerar dados sintéticos

Após o treinamento da rede GAN, seguimos com a geração dos dados adicionais. Durante o processo de geração de dados, sintetizamos 1024 instâncias para cada classe que está sendo analisada, totalizando 2048 dados artificiais que serão incorporados ao conjunto de dados original. No entanto, esses dados compreendem apenas os 48 parâmetros básicos. Antes de prosseguirmos com as análises, é necessário calcular os demais parâmetros.

A Figura 25 apresenta uma comparação entre as médias dos dados reais e os dados sintetizados. A partir dessa comparação, podemos observar que a reprodução das medidas que representam as médias é mais consistente com os dados reais em comparação com as medidas que representam a variância do parâmetro.



(a) Medidas representando as médias



(b) Medidas representando as variância

Figura 25 – Comparação das médias dos atributos calculados após sintetização.

Retomando nosso argumento inicial que motivou a geração de dados adicionais, relacionado à capacidade de nossos modelos de generalizar os resultados em dados previamente não observados, podemos realizar uma comparação utilizando as mesmas curvas de validação utilizadas na seção 4.7. Isso nos permitirá avaliar como o aumento dos dados pode aprimorar a capacidade de generalização dos modelos, conforme ilustrado na Figura 26.

Pode-se observar uma notável melhora na capacidade de generalização do modelo após a inclusão de dados sintéticos para complementar o conjunto de dados original. Um pequeno ajuste no valor do hiper-parâmetro gamma resulta em uma rápida redução do erro médio nos dados de treinamento, atingindo valores próximos a 0.2. Essa redução se mantém estável até que gamma atinja aproximadamente 0.02, momento a partir do qual o erro começa a aumentar novamente. No entanto, quando gamma é ajustado para 0.006, o erro volta a diminuir para o mesmo patamar.

O aumento de dados desempenha um papel importante no processo de treinamento do

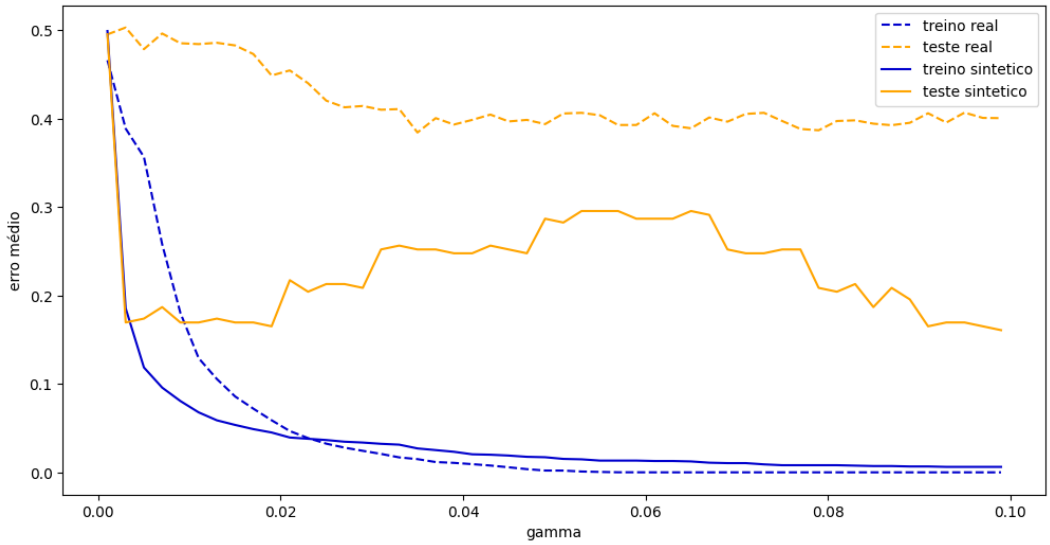


Figura 26 – Curvas de validação comparando desempenho da generalização do modelo SVM antes e depois do data augmentation.

algoritmo. Isso ocorre porque, ao reduzir as restrições do modelo e, consequentemente, aumentar o espaço de possíveis funções, o modelo tem a oportunidade de explorar uma variedade mais ampla de exemplos no conjunto de dados. Isso, por sua vez, permite ao modelo aprender relações mais significativas e complexas. A Figura 27 resume o ganho de performance obtido após adição de dados artificiais no treinamento de alguns modelos de classificação.

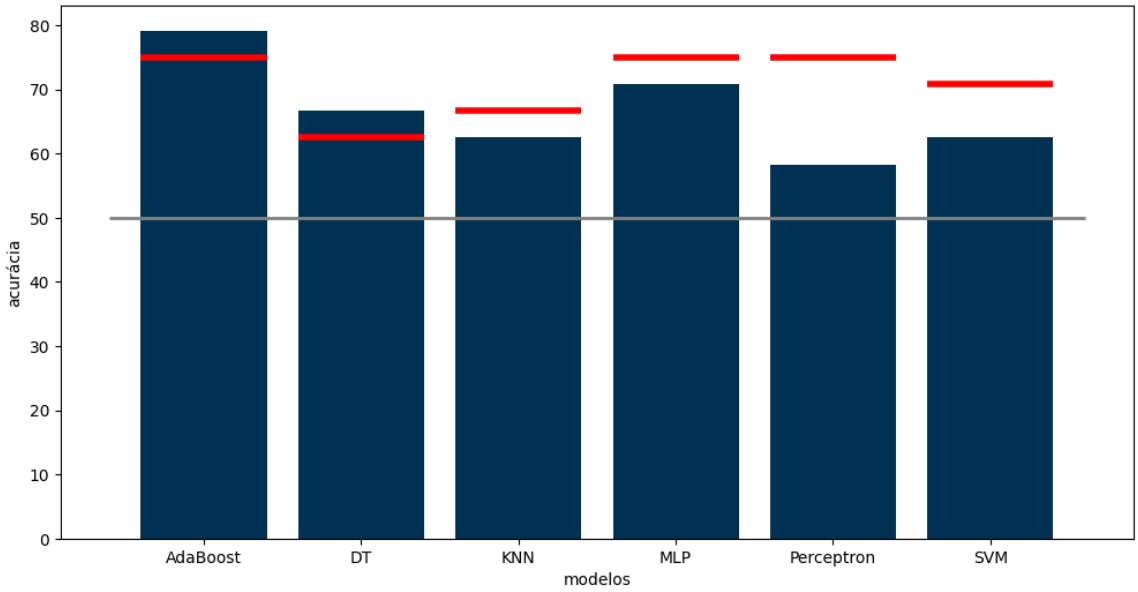


Figura 27 – Seleção de variáveis por RFE aplicando sobre o conjunto de dados sintéticos

4.8.2 Seleção de Atributos

Da mesma forma que ocorre quando se trabalha com um conjunto de dados reduzido, o excesso de atributos pode levar o modelo a cometer overfitting. Isso ocorre devido à maior

probabilidade de o modelo se ajustar excessivamente aos dados de treinamento, capturando o ruído nos dados em vez dos padrões reais. Essa situação ocorre porque muitos atributos podem conter informações irrelevantes ou redundantes, como o caso de pares de atributos apresentando alta correlação. Além disso, a inclusão de atributos em excesso no treinamento do modelo requer mais recursos computacionais, tornando o processo mais lento. Portanto, é de suma importância estabelecer um método para a seleção dos atributos mais relevantes na tarefa de determinação da classe.

Portanto, o objetivo aqui é identificar entre todas as variáveis um subconjunto que seja realmente relevante para o problema, removendo possíveis redundâncias que possa existir entre atributos. Para essa tarefa temos à disposição algumas técnicas que podem auxiliar no processo de seleção: técnicas menos rigorosas, como seleção manual das variáveis a partir do conhecimento prévio obtido por uma análise exploratória dos dados: como, por exemplo, análise dos gráficos de força, apresentados na seção 4.2, ou análise da distribuição dos valores com segmentação na classe alvo. Na figura 28 temos a distribuição de oito variáveis, na primeira fila separamos quatro variáveis que apresentam uma pequena distinção na curva de distribuição quando comparamos plantas saudáveis e infectadas. Na segunda fila separamos mais quatro variáveis, mas neste caso quase não há distinção entre as curvas, sugerindo que tais variáveis trazem pouco ou nenhuma informações a respeito do problema.

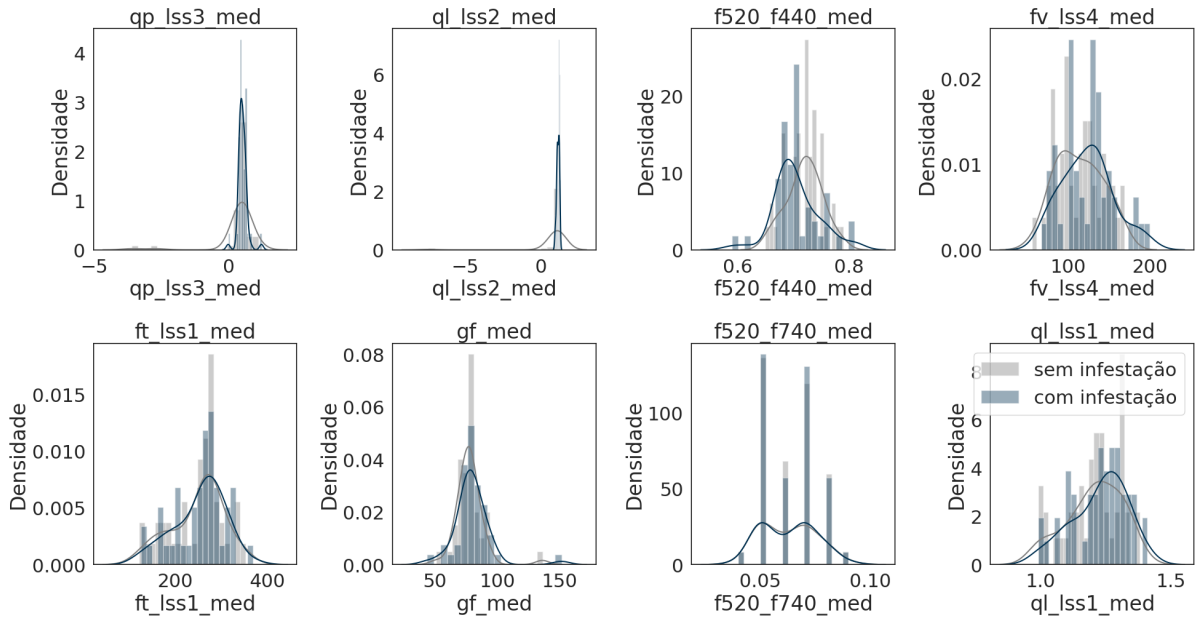


Figura 28 – Distribuição dos atributos discriminado pela classe alvo - as quatro figuras na parte superior mostram atributos que apresentam uma separação visual, em contraste as quatro na parte inferior onde não há separação aparente

Outro ponto que devemos levar em consideração nessa fase da análise é a correlação que existe entre variáveis. Como vimos na seção exploratória, uma grande parcela desses atributos estão fortemente correlacionados, seja por se tratar de medidas derivadas ou por se tratar de medidas geradas por pulsos de luz semelhante. A Figura 29 ilustra uma situação onde temos

dois sinais consecutivos, fo_lss3_med e fo_lss4_med, que estão fortemente correlacionados. Os atributos foram selecionadas manualmente por apresentarem separação visual com relação à classe (diagonal principal). Ao analisar a relação entre os atributos, percebemos que alguns pares apresentam uma relação significativa, como acontece para os atributos fo_lss3_med e fo_lss4_med. Isso nos leva a questionar a necessidade de todas as variáveis, uma vez que apenas uma delas pode ser suficiente.

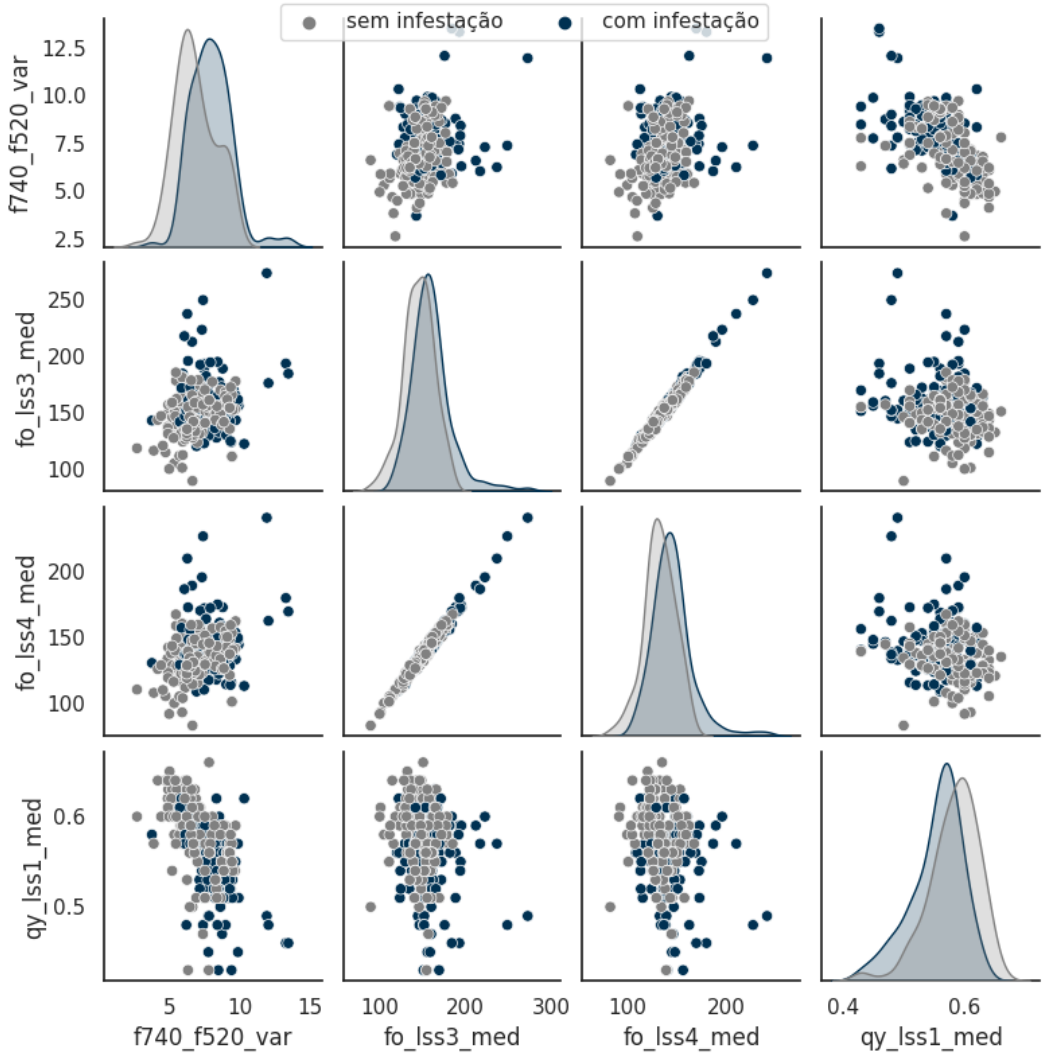


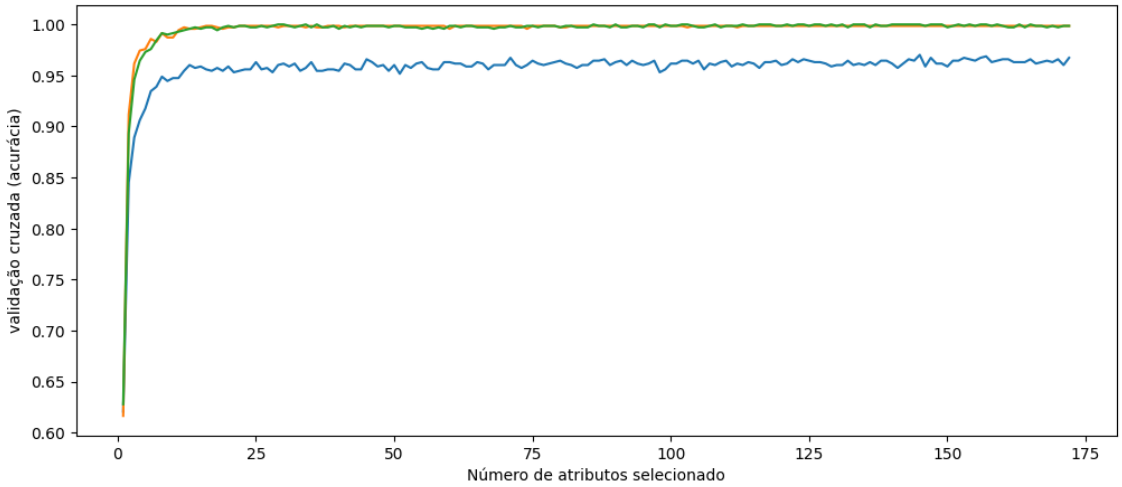
Figura 29 – Seleção manual dos atributos a partir da curva de distribuição segmentada pela classe alvo, mostrando o relacionamento par a par do subconjunto de dados da lagarta

Sabendo da existência desses casos particulares entre variáveis, o processo de seleção precisa, além de manter o máximo de variáveis relevantes dentro problema, também deve ter condições para identificar e dar a solução mais apropriada. Dessa forma devemos proteger o algoritmo contra esse tipo de informação duplicada e evitar a especialização nos dados de treino.

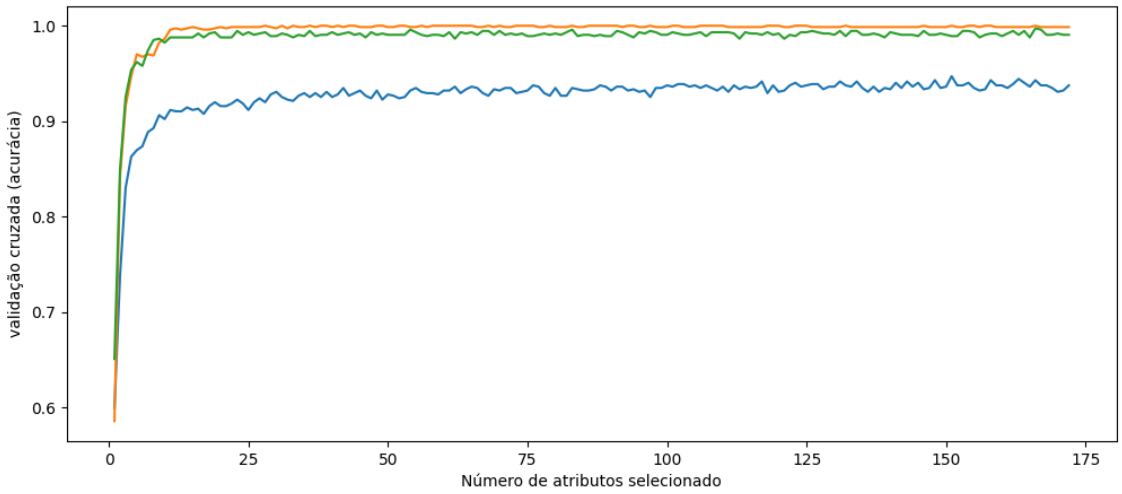
Para um problema com poucas variáveis um processo de seleção manual ou semi-automatizado em muitos casos seria uma opção viável, mas para nosso caso seria uma tarefa tanto pouco cansativa dado que temos 172 variáveis para analisar. Para automatizar o processo de seleção de variáveis, utilizamos a técnica de Eliminação de Variáveis por Recursão com Validação

Cruzada (RFECV), na qual já demos uma pequena introdução na seção “Generalização” ao enfatizar problemas decorrentes de um conjunto reduzido de dados.

O objetivo principal aqui é encontrar um ponto de corte na qual mantenha o compromisso entre um nível de acurácia aceitável e um número reduzido de atributos retidos. Verificamos que essa situação acontece próxima aos pontos em que a acurácia apresenta aumento expressivo com a adição de um atributo. Essa situação está ilustrada na figura 31.



(a) Percevejo

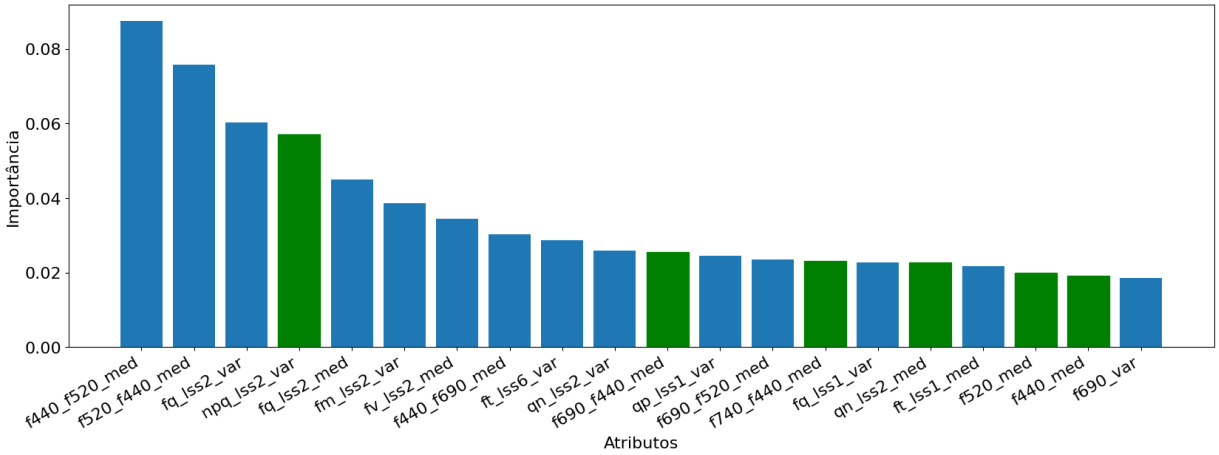


(b) Lagarta

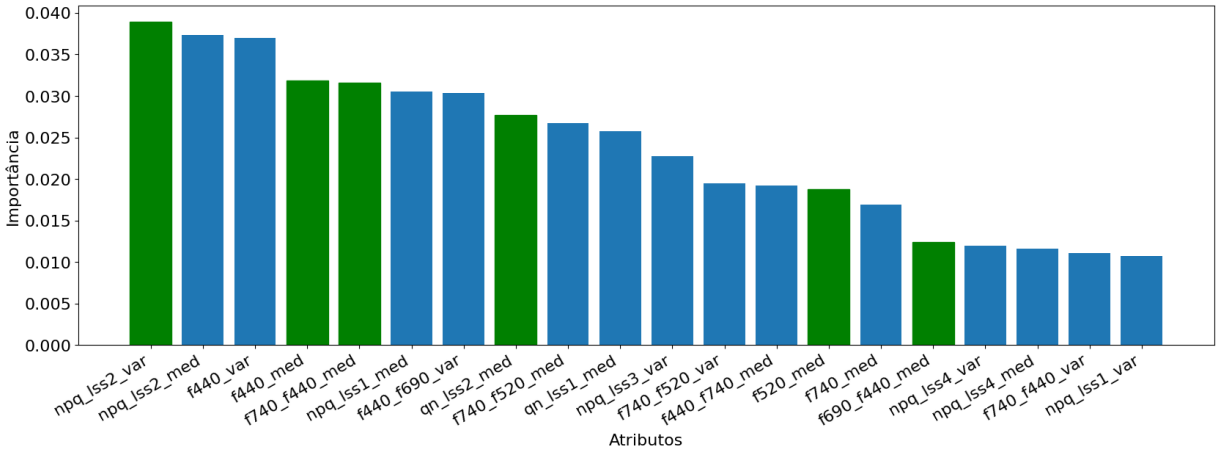
Figura 30 – Seleção de variáveis por RFE aplicado sobre o conjunto de dados sintéticos, sugerindo o valor 145 como o número ótimo de atributos

Quando o modelo é testado com poucas variáveis, temos uma classificação que podemos dizer quase aleatória, onde o valor da acurácia fica em torno de 50%, conforme aumentamos a quantidade de atributos a classificação perde gradualmente a aleatoriedade melhorando o valor da acurácia até atingir os valores máximos quando o número de atributos retidos é igual 40 para os dados do Percevejo e 166 para dados da Lagarta. As quantidades sugerida pelo modelo é baseada no valor máximo da acurácia obtido naquele conjunto de dados. No entanto, ainda

podemos obter valores satisfatórios quando a quantidade de atributos está em torno de 30, pois a partir desse valor não obtemos ganhos expressivos na acurácia.



(a) Percevejo



(b) Lagarta

Figura 31 – Gráfico de barras listando em ordem decrescente os 15 atributos com maior importância, destacando em cor verde 6 atributos que são comuns nos dois conjuntos dedados

4.9 Resultados da classificação

Após complementar os dados originais com dados gerados artificialmente por meio de uma rede GAN e selecionar os atributos mais relevantes, prosseguimos com o processo final de classificação. Antes de nos aprofundarmos em um modelo de classificação específico, vamos avaliar o desempenho dos dados sintéticos em alguns modelos conhecidos, visando identificar aquele que melhor se adapta à nossa finalidade. Nesse estudo foi empregado variações dos modelos KNN, DT, Ensemble, MLP e SVM sem a preocupação em ajustar os hiper-parâmetros e mantendo todos os atributos do conjunto de dados. A Tabela 10 detalha os principais resultados desse estudo. Como esperado, na maior parte dos modelos há um aumento na acurácia quando o modelo é treinado com os dados sintéticos. Algumas exceções ocorrem para modelos aplicados

nos dados do percevejo, na qual podemos ver uma diminuição na performance do modelo, porém na média o aumento é positivo.

Tabela 10 – Acurácia da classificação nos dados testes comparando o resultado do modelo treinado com os dados originais e com dados artificiais

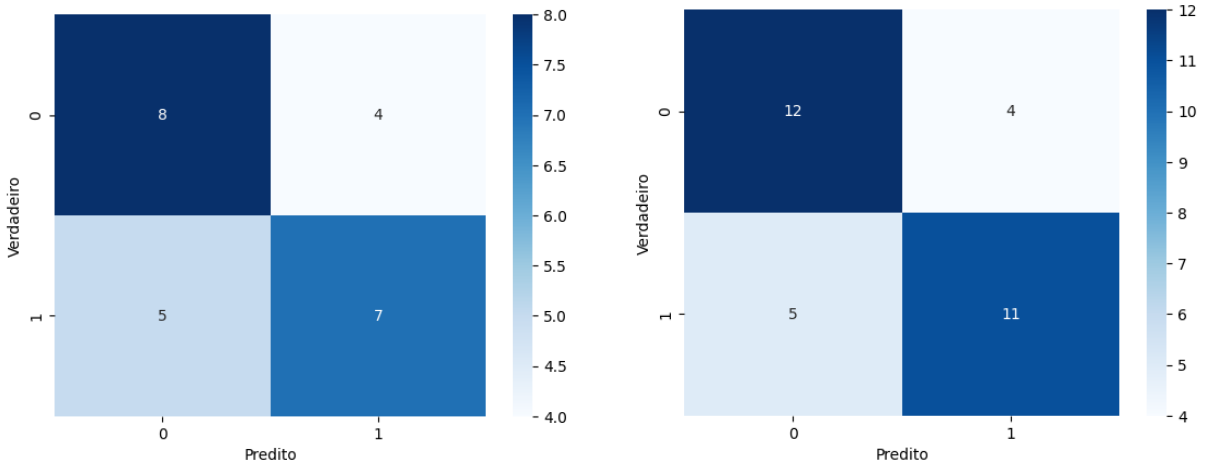
	Percevejo		**Lagarta**	
Modelo	Original	Sintético	Original	Sintético
KNN 5 vizinhos	54.17	75.00	40.62	59.38
KNN 7 vizinhos	50.00	70.83	46.88	59.38
KNN 9 vizinhos	62.50	70.83	50.00	62.50
DT (gini, best)	70.83	70.83	43.75	59.38
DT (gini, random)	54.17	62.50	21.88	62.50
DT (entropy, best)	54.17	66.67	53.12	65.62
DT (entropy, random)	70.83	54.17	59.38	46.88
AdaBoostClassifier	79.17	**79.17**	53.12	**71.88**
Perceptron	58.33	**79.17**	56.25	65.62
MLP (5,) camadas	54.17	66.67	50.00	56.25
MLP (3,) camadas	70.83	62.50	50.00	56.25
MLP (2,1) camadas	54.17	62.50	53.12	59.38
MLP (7,5,3) camadas	58.33	62.50	34.38	46.88
SVM Linear	62.50	70.83	46.88	65.62
SVM RBF	50.00	50.00	50.00	50.00
SVM Sigmoid	58.33	50.00	40.62	71.88
SVM Polinomial Grau 3	50.00	62.50	43.75	50.00
Média	**59.56**	**65.69**	**46.69**	**59.38**

A princípio não sabemos quais resultados vamos obter com a classificação após um exaustivo processo exploratório, sendo importante a definição de um **baseline** para nortear o resultado que desejamos. Com um baseline como referência podemos buscar melhorias com objetivo de chegar em um modelo que supere o resultado do baseline. Com base nos resultados da análise feita anteriormente, selecionamos como baseline o modelo MLP para dados do percevejo e o *adaboost* para os dados da Lagarta. A Figura 32 apresenta a matriz de confusão com esses modelos aplicado em cada conjunto de dados e a Tabela 11 detalha as principais métricas calculada a partir da matriz de confusão.

Tabela 11 – Principais métricas do baseline

	Acurácia	Precisão	Sensibilidade	Especificidade
MLP (percevejo)	62.50	63.64	58.33	66.67
AdaBoost (lagarta)	71.88	73.33	68.75	75.00

Na matriz de confusão, a diagonal principal apresenta a quantidade de instâncias que o algoritmo conseguiu acertar, ou seja, qual número de indivíduos da classe 1 o algoritmo realmente conseguiu classificar como 1 e qual número de indivíduos da classe 0 o algoritmo



(a) MLP (percevejo)

(b) AdaBoost (lagarta)

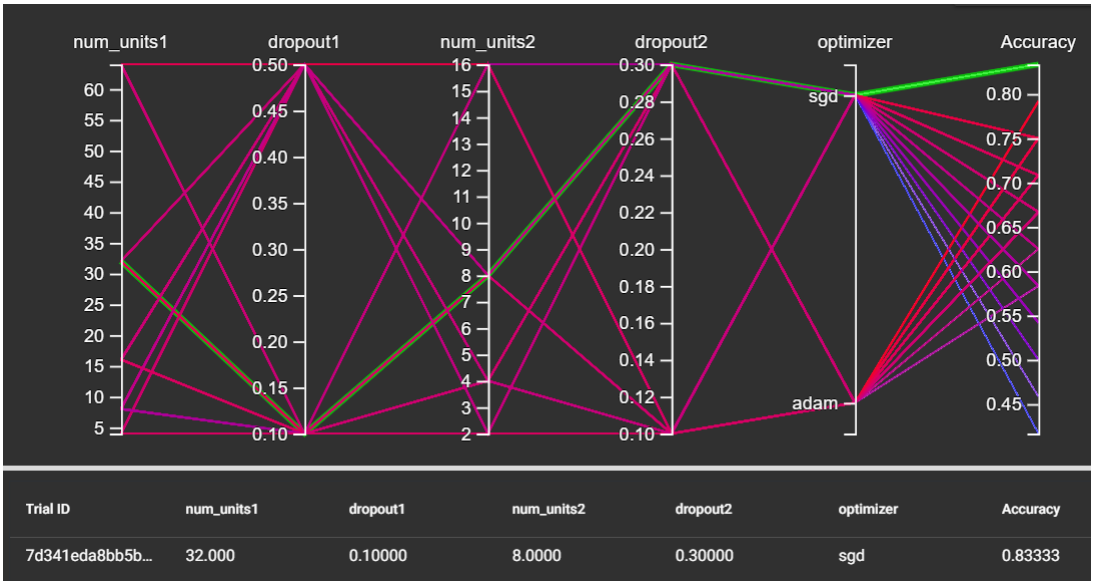
Figura 32 – Matriz de confusão dos Baseline selecionado para cada tipo de infestação

classificou como 0. Na diagonal secundária temos a quantidade de instâncias que o algoritmo faz a classificação errada. Podemos observar comportamentos similares entre os resultados dos dois modelos, em ambos os casos temos uma taxa de acerto na classe negativa (classe=0) aceitável, especificidade igual 67% para classificação nos dados do percevejo e 75% para os dados da lagarta. Porém, neste caso que estamos trabalhando com dados relacionados a infestação por uma praga, devemos dar mais atenção para a sensibilidade, na qual mede a taxa de acerto na classe positiva (classe=1), pois é preferível classificar uma classe negativa como positiva do que o contrário. Ao classificarmos erroneamente uma classe positiva como negativa, arriscamos manter uma planta contaminada, o que pode resultar na disseminação da infestação por toda a plantação.

Com baseline em mãos, podemos buscar por melhorias incrementais nos nossos modelos com objetivo em obter melhores resultados. Nessa próxima fase da análise, onde seguiremos com um processo de experimentação para tentar identificar os melhores hiper-parâmetros para o nosso problema, adotamos dois *frameworks* distintos: *TensorFlow* para otimização do modelo MLP e continuamos utilizando o framework scikit-learn no modelo adaboost. Para otimização dos hiper-parâmetros do modelo MLP, utilizamos uma extensão do TensorFlow chamada TensorBoard como complemento para facilitar a configuração do espaço de pesquisa. A extensão oferece um painel para ajudar no processo de identificação do melhor experimento ou dos conjuntos mais promissores de hiper-parâmetros. A Figura 33 mostra o resultado dos experimentos executados, destacando em cor verde o experimento que gerou melhor acurácia (83%).

A tabela de resumo para rede neural *feedforward* para classificação das folhas contaminada com percevejo pode se vista na Figura 34. O modelo RNN inclui uma camada de Dense(32), Dropout(0.1), Dense(8), Dropout(0.3), com todas as camadas ocultas sendo ativada função ReLu e uma camada de saída contendo um nerônio com função de ativação sigmoid.

Após ajustar os parâmetros da rede, o modelo foi treinando novamente com os parâmetros



Trial ID	num_units1	dropout1	num_units2	dropout2	optimizer	Accuracy
7d341eda8bb5b...	32.000	0.10000	8.0000	0.30000	sgd	0.83333

Figura 33 – Painel do TensorBoard

Layer (type)	Output Shape	Param #
dense_44 (Dense)	(None, 32)	5536
dropout_33 (Dropout)	(None, 32)	0
dense_45 (Dense)	(None, 16)	528
dropout_34 (Dropout)	(None, 16)	0
dense_46 (Dense)	(None, 4)	68
dropout_35 (Dropout)	(None, 4)	0
dense_47 (Dense)	(None, 1)	5

Total params: 6137 (23.97 KB)
Trainable params: 6137 (23.97 KB)
Non-trainable params: 0 (0.00 Byte)

Figura 34 – Tabela de resumo da rede neural utilizada na classificação das folhas infectadas com Percevejo

obtidos na busca com objetivo de encontrar o número ótimo de épocas para treinar o modelo, para isso foi analisado as taxas de perda e precisão de validação nos conjuntos de treinamento e validação foram examinadas para evitar que os modelos sofram overfitting. Conforme a figura 35, após muitas épocas o modelo aumenta a propensão de cometer overfitting, sugerindo um número ideal entre 10 e 20 épocas.

Fixamos os parâmetros e executamos o modelo várias vezes para garantir a variação no aprendizado.

O mesmo procedimento, utilizando redes neurais, foi replicado nos dados de infestação

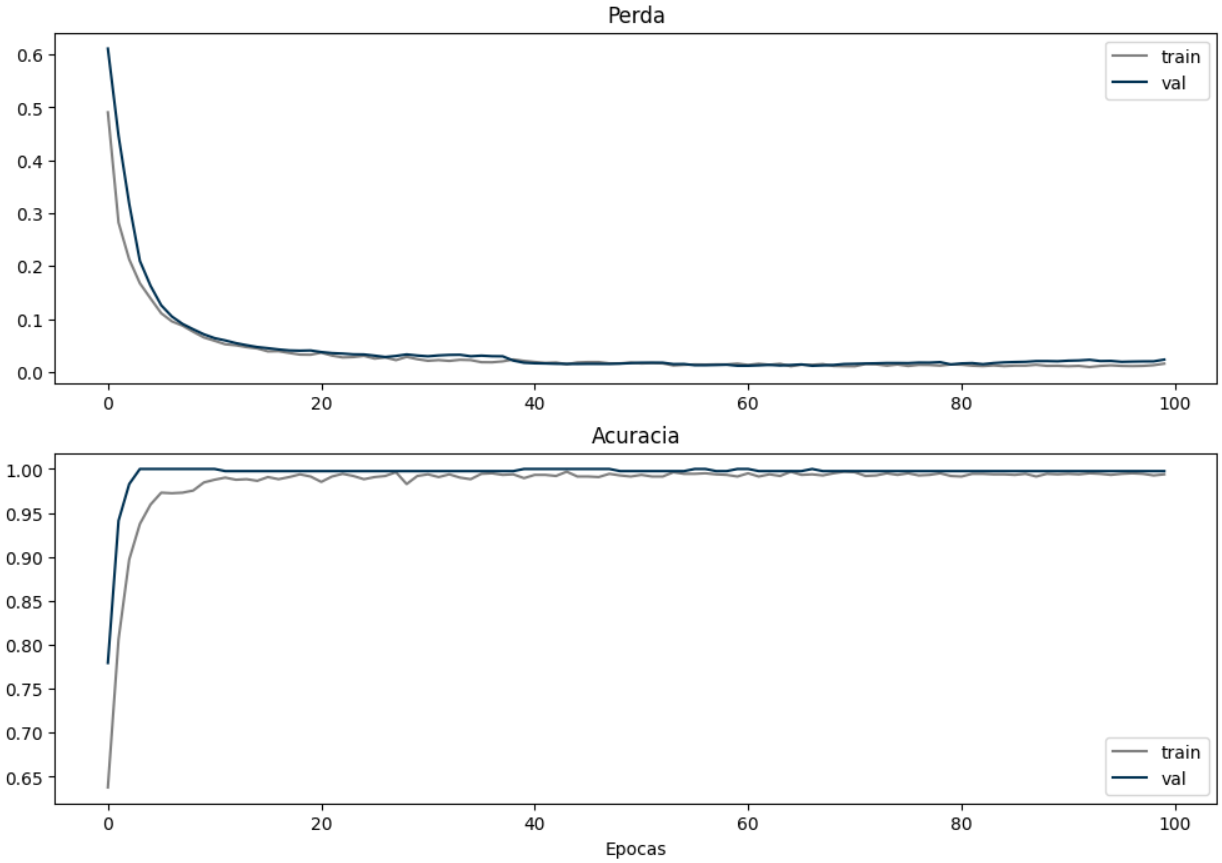
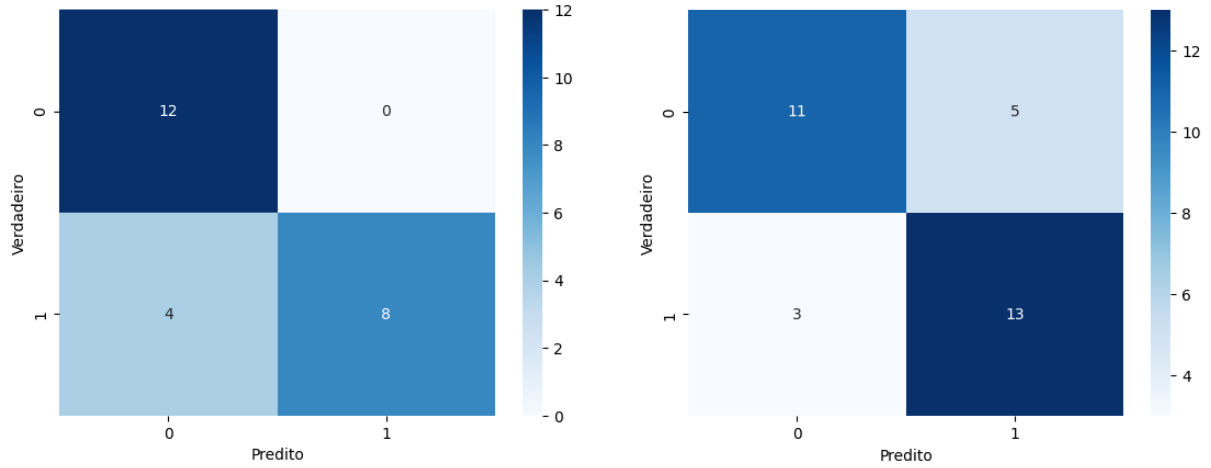


Figura 35 – Evolução da função perda e da acurácia durante o processo de treinamento da rede neural nos dados do Percevejo

por lagarta, porém não obtivemos melhoras significativas no modelo, demandando a expansão da análise para outros modelos de aprendizado de máquina. Neste contexto, optamos por empregar o modelo adaboost para a otimização dos parâmetros. Para esses casos, recorremos ao método GridSearchCV, que faz parte da biblioteca scikit-learn. Como resultado da análise obtivemos os seguintes valores para os parâmetros:

```
Optimal hyperparameter combination: {'algorithm': 'SAMME.R', '
    learning_rate': 1, 'n_estimators': 150}
```

Como resultado principal dos nossos estudos, apresentamos novamente a matriz de confusão (Figura 36) e um resumo das principais métricas alcançadas (Tabela 12).



(a) MLP (percevejo)

(b) AdaBoost (lagarta)

Figura 36 – Matriz de confusão dos modelos otimizado selecionado para cada tipo de infestação

Tabela 12 – Principais métricas dos modelos após otimização dos parâmeros

	Acurácia	Precisão	Sensibilidade	Especificidade
MLP (percevejo)	83.33	100.00	66.67	100.00
AdaBoost (lagarta)	75.00	72.22	81.25	68.75

CAPÍTULO

5

CONCLUSÃO

Este trabalho teve como objetivo principal construir classificadores capazes de fornecer diagnóstico precoce e confiável sobre o estado de saúde de uma planta em análise. Para esse objetivo, propomos uma nova etapa no pipeline de aprendizado de máquina: a inclusão do "data augmentation", um método comumente utilizado para gerar imagens artificias que se assemelham a imagens reais. Essa abordagem foi adotada para complementar dados reais no formato tabular, gerados por meio do bioensaio, como forma de contornar a limitação na quantidade de dados que podemos obter nesses tipos de experimentos.

As redes GANs foram escolhidas por apresentar boa capacidade de gerar dados que se assemelham de perto aos dados reais. Elas aprendem a distribuição dos dados de treinamento, permitindo a criação de amostras artificialmente geradas com características semelhantes. Além de trabalhar bem com dados complexos e de alta dimensionalidade, como é caso de imagens de alta resolução, tornando o modelo adequado para uma variedade de domínios, desde visão computacional até dados tabulares complexos, como é o nosso caso.

Ao integrar os dados gerados pelas GANs ao conjunto de treinamento inicial, conseguimos ampliar e diversificar o conjunto de dado original. Essa abordagem teve um impacto positivo no desempenho e na capacidade de generalização da maioria dos modelos de aprendizado de máquina selecionados para este estudo. Destacando principalmente as Redes Neurais aplicadas ao conjunto de dados do percevejo, que elevou a taxa de acerto de 62.5% para 83.3%. Além disso, o Adaboost, ao ser aplicado aos dados da lagarta, viu uma melhoria de 71.8% para 75%. Atribuímos essa melhora não ao fato de estamos acrescentando realmente dados novos no conjunto de treinamento, acreditamos que a melhora vem do fato que estamos alimentando os modelos com variações do ruído que são diferentes daquelas do conjunto original. Propiciando aos modelos maior variedade e oportunidade para aprender relações mais significativas.

A fase de exploração dos dados teve papel importante para dar o direcionamento e esclarecer os pontos mais importantes que deveriam ser levados em consideração durante a fase de pre-processamento e escolha dos modelos. Nessa fase foi possível identificar os parâmetros

mais relevantes para identificação da classe alvo, assim como os tempos aproximados onde os primeiros sinais da infestação começam aparecer, apresentando os primeiros sinais 8 horas após a infestação do percevejo e 24 horas após infestação com a lagarta.

A análise de formação de grupos revelou a possibilidade de evolução dos modelos no sentido de construção de classificadores mais genéricos, que possam trabalhar com diferentes variedades do milho. O resultado mostrou que é possível combinar os dois paradigmas de aprendizado: iniciando por uma classificação não supervisionada para definir qual variedade a planta se enquadra e na sequência direcionando a instância para um classificador específico para seguir com a classificação final no contexto de aprendizado supervisionado.

Ao final, foi possível confirmar nossa hipotética inicial, na qual diz que é possível desenvolver métodos de classificação eficazes, com base nos parâmetros de fluorescência da clorofila, para diagnóstico precoce de plantas sob ataque de pragas e que este modelo tende a ter um desempenho melhor na generalização quando treinado com uma quantidade representativa dos dados. Entretanto, estudos adicionais podem ser feitos para expandir os resultados encontrados. Por exemplo, combinar as pragas em um único conjunto de dados e testar se é possível diagnosticar qual a doença a planta foi infectada. Além disso, não foi possível estudar os efeitos do ciclo circadiano da planta nos parâmetros medidos devido à coleta não uniforme dos dados, ficando como sugestão para trabalhos futuros, o desenho de uma coleta com esparsamento mais uniforme, tentando manter os mesmos horários de coleta do dia anterior.

REFERÊNCIAS

aaaa, a. Nenhuma citação no texto.

aaaa, b. Nenhuma citação no texto.

aaaa, c. Nenhuma citação no texto.

aaaa, d. Nenhuma citação no texto.

aaaa, e. Nenhuma citação no texto.

aaaa, f. Nenhuma citação no texto.

CONAB. Produção de grãos, 8 2023. URL <https://www.conab.gov.br/ultimas-noticias/5116-producao-de-graos-e-estimada-em-320-1-milhoes-de-toneladas-com-ganhos-de-area/produtividade>. Citado na página 23.

Antonio C. Oliveira Ivan Cruz, M. L. C. Figueiredo and Carlos A. Vasconcelos. Damage of spodoptera frugiperda (smith) in different maize genotypes cultivated in soil under three levels of aluminium saturation. *International Journal of Pest Management*, 45(4):293–296, 1999. doi: <10.1080/096708799227707>. URL <https://doi.org/10.1080/096708799227707>. Citado na página 23.

Renato J Horikoshi, Daniel Bernardi, Oderlei Bernardi, José B Malaquias, Daniela M Okuma, Leonardo L Miraldo, Fernando S de A E Amaral, and Celso Omoto. Effective dominance of resistance of spodoptera frugiperda to bt maize and cotton varieties: implications for resistance management. *Scientific reports*, 6:34864, October 2016. ISSN 2045-2322. doi: <10.1038/srep34864>. URL <https://europepmc.org/articles/PMC5056508>. Citado na página 23.

Jinzhu Lu, Lijuan Tan, and Huanyu Jiang. Review on convolutional neural network (cnn) applied to plant leaf disease classification. *Agriculture*, 11(8), 2021. ISSN 2077-0472. doi: <10.3390/agriculture11080707>. URL <https://www.mdpi.com/2077-0472/11/8/707>. Citado na página 24.

Wei Yang, Ce Yang, Ziyuan Hao, Chuanqi Xie, and Minzan Li. Diagnosis of plant cold damage based on hyperspectral imaging and convolutional neural network. *IEEE Access*, 7:118239–118248, 2019. doi: <10.1109/ACCESS.2019.2936892>. Citado na página 24.

Jieni Yao, Dawei Sun, Haiyan Cen, Haixia Xu, Haiyong Weng, Fang Yuan, and Yong He. Phenotyping of arabidopsis drought stress response using kinetic chlorophyll fluorescence and multicolor fluorescence imaging. *Frontiers in Plant Science*, 9, 2018. ISSN 1664-462X. doi: <10.3389/fpls.2018.00603>. URL <https://www.frontiersin.org/articles/10.3389/fpls.2018.00603>. Citado na página 24.

Zhenfen Dong, Yuheng Men, Zhenzhen Liu, Jinpeng Li, and Jianwei Ji. Application of chlorophyll fluorescence imaging technique in analysis and detection of chilling injury of tomato

seedlings. *Computers and Electronics in Agriculture*, 168:105109, 2020. ISSN 0168-1699. doi: <https://doi.org/10.1016/j.compag.2019.105109>. URL <https://www.sciencedirect.com/science/article/pii/S0168169919321957>. Citado na página 24.

Haiyong Weng, Yunshi Liu, Ishimwe Captoline, Xiaobin Li, Dapeng Ye, and Renye Wu. Citrus huanglongbing detection based on polyphasic chlorophyll a fluorescence coupled with machine learning and model transfer in two citrus cultivars. *Computers and Electronics in Agriculture*, 187:106289, 2021. ISSN 0168-1699. doi: <https://doi.org/10.1016/j.compag.2021.106289>. URL <https://www.sciencedirect.com/science/article/pii/S0168169921003069>. Citado na página 24.

Olaf Erenstein, Moti Jaleta, Kai Sonder, Khondoker Mottaleb, and B. M. Prasanna. Global maize production, consumption and trade: trends and rd implications, 10 2022. ISSN 18764525. Citado na página 29.

Carlos Roberto Casela, Alexandre da Silva Ferreira, and Nicésio Filadelfo J. de Almeida Pinto. Doenças na cultura do milho. 2006. ISSN 1679-1150. Citado na página 29.

João Américo, Wordell Filho, Leandro Do Prado, Ribeiro Luis, Antônio Chiaradia, José Carlos, Madalóz Cristiano, and Nunes Nesi. Pragas e doenÇas do milho, 2016. ISSN 0100-7416. Citado na página 29.

Ric Bessin. Stink bug damage to corn ric bessin, 2019. URL <https://entomology.ca.uky.edu/categories/corn-pests.> Citado na página 31.

Nádia Maebara Bueno, Edson Luiz Lopes Baldin, Vinicius Fernandes Canassa, Leandro do Prado Ribeiro, Ivana Fernandes da Silva, André Luiz Lourenção, and Robert Lee Koch. Characterization of antixenosis and antibiosis of corn genotypes to dichelops melacanthus dallas (hemiptera: Pentatomidae). *Gesunde Pflanzen*, 73:67–76, 3 2021. ISSN 14390345. doi: <10.1007/s10343-020-00529-z>. Citado na página 31.

Paulo Henrique Ramos Fernandes, Crébio José Ávila, Ivana Fernandes da Silva, and Daniele Zulin. Damage by the green-belly stink bug to corn. *Pesquisa Agropecuaria Brasileira*, 55, 2020. ISSN 16783921. doi: <10.1590/S1678-3921.PAB2020.V55.01131>. Citado na página 31.

Paulo Pereira, Paulo Roberto Valle, Silva Pereira, Lucas Simionato Tonello, and José Roberto Salvadori. Ministério da agricultura, pecuária e abastecimento caracterização das fases de desenvolvimento e aspectos da biologia do percevejo barriga-verde dichelops melacanthus (dallas, 1851). 2007. ISSN 1517-4964. Citado na página 31.

Emerson Crivelaro Gomes, Rafael Hayashida, and Adeney de Freitas Bueno. Dichelops melacanthus and euschistus heros injury on maize: Basis for re-evaluating stink bug thresholds for ipm decisions. *Crop Protection*, 130, 4 2020. ISSN 02612194. doi: <10.1016/j.cropro.2019.105050>. Citado na página 31.

Wee Tek Tay, Robert L Meagher Jr, Cecilia Czepak, and Astrid T Groot. Spodoptera frugiperda: Ecology, evolution, and management options of an invasive species. *Annual Review of Entomology*, 68:299–317, 2022. doi: <10.1146/annurev-ento-120220>. URL <https://doi.org/10.1146/annurev-ento-120220->. Citado na página 32.

Francisco A. Paredes-Sánchez, Gildardo Rivera, Virgilio Bocanegra-García, Hadassa Y. Martínez-Padrón, Martín Berrones-Morales, Nohemí Niño-García, and Verónica Herrera-Mayorga. Advances in control strategies against spodoptera frugiperda. a review, 9 2021. ISSN 14203049. Citado na página 32.

Eduardo Barbosa Beserra, José Roberto, and Postali Parra. Beserra parra impact of the number of spodoptera frugiperda egg layers on parasitism by trichogramma atopovirilia, 2005. Citado nas páginas 32 e 33.

Ivan Cruz, Maria Lourdes Figueiredo, and Marcos Matoso. Controle biológico de spodptera frugiperda utilizando o parasitóide de ovos trichogramma, 1999. ISSN 01 00-801 3. Citado na página 32.

FAO. New standards to curb the global spread of plant pests and diseases, 2019. URL <https://www.fao.org/news/story/en/item/1187738/icode/>. Citado na página 33.

Marc Venbrux, Sam Crauwels, and Hans Rediers. Current and emerging trends in techniques for plant pathogen detection, 2023. ISSN 1664462X. Citado nas páginas 33 e 34.

Guolan Lu and Baowei Fei. Medical hyperspectral imaging: a review. *Journal of Biomedical Optics*, 19, 2014. ISSN 1083-3668. doi: <10.1117/1.jbo.19.1.010901>. Citado na página 35.

Jayme Garcia and Arnal Barbedo. Uso de dados multiespectrais e hiperespectrais na detecção, medição e diagnóstico de doenças na agricultura, 2015. URL <www.embrapa.br/informatica-agropecuaria>. Citado na página 35.

Jan Behmann, Jörg Steinrücken, and Lutz Plümer. Detection of early plant stress responses in hyperspectral images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 93:98–111, 7 2014. ISSN 09242716. doi: <10.1016/j.isprsjprs.2014.03.016>. URL <https://linkinghub.elsevier.com/retrieve/pii/S092427161400094X>. Citado na página 35.

Rikke Gade and Thomas B. Moeslund. Thermal cameras and applications: A survey. *Machine Vision and Applications*, 25, 2014. ISSN 09328092. doi: <10.1007/s00138-013-0570-5>. Citado na página 35.

Wenjing Zhu, Hua Chen, Izabela Ciechanowska, and Dean Spaner. Application of infrared thermal imaging for the rapid diagnosis of crop disease. volume 51, pages 424–430. Elsevier B.V., 1 2018. doi: <10.1016/j.ifacol.2018.08.184>. Citado na página 36.

Sara Francesconi, Antoine Harfouche, Mauro Maesano, and Giorgio Mariano Balestra. Uav-based thermal, rgb imaging and gene expression analysis allowed detection of fusarium head blight and gave new insights into the physiological responses to the disease in durum wheat. *Frontiers in Plant Science*, 12, 4 2021. ISSN 1664462X. doi: <10.3389/fpls.2021.628575>. Citado na página 36.

E. Capuano and S.M. van Ruth. Infrared spectroscopy: Applications. In Benjamin Caballero, Paul M. Finglas, and Fidel Toldrá, editors, *Encyclopedia of Food and Health*, pages 424–431. Academic Press, Oxford, 2016. ISBN 978-0-12-384953-3. doi: <https://doi.org/10.1016/B978-0-12-384947-2.00644-9>. URL <https://www.sciencedirect.com/science/article/pii/B9780123849472006449>. Citado na página 36.

Siti Anis Dalila Muhammad Zahir, Ahmad Fairuz Omar, Mohd Faizal Jamlos, Mohd Azraie Mohd Azmi, and Jelena Muncan. A review of visible and near-infrared (vis-nir) spectroscopy application in plant stress detection, 5 2022. ISSN 09244247. Citado na página 37.

Nawaf Abu-Khalaf and Mazen Salman. Visible/near infrared (vis/nir) spectroscopy and multivariate data analysis (mvda) for identification and quantification of olive leaf spot (ols) disease, 2014. URL <http://www.ptuk.edu.ps>. Citado na página 37.

Damian Bienkowski, Matt J. Aitkenhead, Alison K. Lees, Christopher Gallagher, and Roy Neilson. Detection and differentiation between potato (solanum tuberosum) diseases using calibration models trained with non-imaging spectrometry data. *Computers and Electronics in Agriculture*, 167, 12 2019. ISSN 01681699. doi: <10.1016/j.compag.2019.105056>. Citado na página 37.

Alfadhl Yahya Khaled, Samsuzana Abd Aziz, Siti Khairunniza Bejo, Nazmi Mat Nawi, Idris Abu Seman, and Daniel Iroemeha Onwude. Early detection of diseases in plant tissue using spectroscopy–applications and limitations, 1 2018. ISSN 1520569X. Citado na página 38.

Anielle C. Ranulfi, Marcelo C.B. Cardinali, Thiago M.K. Kubota, Juliana Freitas-Astúa, Ednaldo J. Ferreira, Barbara S. Bellete, Maria Fátima G.F. da Silva, Paulino R. Villas Boas, Aida B. Magalhães, and Débora M.B.P. Milori. Laser-induced fluorescence spectroscopy applied to early diagnosis of citrus huanglongbing. *Biosystems Engineering*, 144:133–144, 2016. ISSN 15375110. doi: <10.1016/j.biosystemseng.2016.02.010>. Citado na página 38.

Kate Maxwell1 and Giles N Johnson2. Chlorophyll fluorescence-a practical guide, 2000. Citado nas páginas 38 e 58.

Govindjee Govindjee and George Papageorgiou. *Chlorophyll A Fluorescence: A Signature of Photosynthesis*, volume 19. Springer Netherlands, 2004. ISBN 978-1-4020-3217-2. doi: <10.1007/978-1-4020-3218-9>. URL <http://link.springer.com/10.1007/978-1-4020-3218-9>. Citado na página 38.

Ajila Venkat and Sowbiya Muneer. Role of circadian rhythms in major plant metabolic and signaling pathways, 4 2022. ISSN 1664462X. Citado nas páginas 39 e 69.

Haiyan Cen, Haiyong Weng, Jieni Yao, Mubin He, Jingwen Lv, Shijia Hua, Hongye Li, and Yong He. Chlorophyll fluorescence imaging uncovers photosynthetic fingerprint of citrus huanglongbing. *Frontiers in Plant Science*, 8, 8 2017. ISSN 1664462X. doi: <10.3389/fpls.2017.01509>. Citado na página 39.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. *Pearson correlation coefficient*, volume 2, pages 1–4. Springer Science and Business Media B.V., 2009. doi: <10.1007/978-3-642-00296-0_5>. Citado na página 48.

Pablo M. Granitto, Cesare Furlanello, Franco Biasioli, and Flavia Gasperi. Recursive feature elimination with random forest for ptr-ms analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*, 83(2):83–90, 2006. ISSN 0169-7439. doi: <https://doi.org/10.1016/j.chemolab.2006.01.007>. Citado na página 49.

Maria Carolina Monard and José Augusto Baranauskas. *Conceitos sobre Aprendizado de Máquina*. 2003. Citado na página 52.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. 6 2014. URL <http://arxiv.org/abs/1406.2661>. Citado na página 53.

Mukrin Nakhwan and Rakkrit Duangsoithong. Comparison analysis of data augmentation using bootstrap, gans and autoencoder. pages 18–23. Institute of Electrical and Electronics Engineers Inc., 2022. ISBN 9781665400145. doi: <10.1109/KST53302.2022.9729065>. Citado na página 54.

Jeffrey C Lagarias, James A Reeds, Margaret H Wright, and Paul E Wright. Convergence properties of the nelder–mead simplex method in low dimensions. *SIAM Journal on Optimization*, 9:112–147, 1998. doi: <10.1137/S1052623496303470>. URL <https://doi.org/10.1137/S1052623496303470>. Citado na página 55.

Bernard Genty, Jean Marie Briantais, and Neil R. Baker. The relationship between the quantum yield of photosynthetic electron transport and quenching of chlorophyll fluorescence. *Biochimica et Biophysica Acta - General Subjects*, 990:87–92, 1989. ISSN 03044165. doi: <10.1016/S0304-4165(89)80016-9>. Citado na página 58.

G N Johnson, A J Young, J D Scholes, and P Horton'. The dissipation of excess excitation energy in british plant species. *Plant. Cell and Environment*, 16:673–679, 1993. Citado na página 58.

C. Buschmann. Photochemical and non-photochemical quenching coefficients of the chlorophyll fluorescence: Comparison of variation and limits. *Photosynthetica*, 37:217–224, 9 1999. ISSN 03003604. doi: <10.1023/A:1007003921135>. Citado na página 58.


ICMC
USP
SÃO CARLOS