ANA LETÍCIA BECKER GOMES

MODELOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE DINÂMICAS POPULACIONAIS DE PLANTAS DANINHAS EM SISTEMAS ILP

Itajaí (SC), abril de 2025

UNIVERSIDADE DO VALE DO ITAJAÍ
CURSO DE MESTRADO ACADÊMICO EM COMPUTAÇÃO APLICADA

MODELOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE DINÂMICAS POPULACIONAIS DE PLANTAS DANINHAS EM SISTEMAS ILP

por

Ana Letícia Becker Gomes

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Computação Aplicada.
Orientadora: Anita Maria da Rocha Fernandes, Dra.
Co-orientador: Maurílio Fernandes de Oliveira, Dr.

Itajaí (SC), abril de 2025

AGRADECIMENTOS

Ao meu noivo e aos meus pais por sempre me apoiarem e me incentivarem, em especial durante esta fase.

A minha orientadora, Profª Anita (Dra.), pela orientação e disposição em ensinar. Ao meu co-orientador Maurílio (Dr.), por todas as trocas de conhecimento e pela colaboração.

Aos professores da banca, por todas as suas contribuições. A Fábio Volkman Coelho por auxiliar o projeto.

A EMBRAPA de Milho e Sorgo por disponibilizar os dados, em particular, a Ramon Costa Alvarenga.

A FAPESC, projetos da chamada pública FAPESC N° 54/2022; ao Programa de Ciência, Tecnologia e Programa de Ciência, Tecnologia e Inovação para apoio aos Grupos de Pesquisa da Acafe, N° 2023TR000875; ao CNPq, projeto da chamada CNPq/MCTI n° 10/2023, faixa A, edital universal, processo N° 404755/2023-2; por apoiarem e financiarem este trabalho.

MODELOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE DINÂMICAS POPULACIONAIS DE PLANTAS DANINHAS EM SISTEMAS ILP

Ana Letícia Becker Gomes

Abril / 2025

Orientadora: Anita Maria da Rocha Fernandes, Dra.

Co-orientador: Maurílio Fernandes de Oliveira, Dr.

Área de Concentração: Computação Aplicada

Linha de Pesquisa: Inteligência Aplicada ao Meio Ambiente

Palavras-chave: Aprendizado de Máquina; Plantas daninhas; Sistemas ILP.

Número de páginas: 146

RESUMO

Até o ano de 2050 estima-se que a população mundial será de 9 bilhões de pessoas. Este constante crescimento populacional vem intensificando a necessidade de ampliar a produção alimentícia. Dentre os diversos obstáculos enfrentados pelo sistema agrícola, destacam-se as plantas daninhas. Tem-se que estas possuem diversos métodos de manejo, sendo o controle químico o mais utilizado. Contudo, ao mesmo tempo que procura-se aumentar a produção de alimentos, busca-se também reduzir a poluição ambiental causada pelos herbicidas. Neste contexto, este trabalho busca aplicar algoritmos de aprendizado de máquina e técnicas de análise de dados no manejo de plantas daninhas em sistemas Integração Lavoura-Pecuária (ILP). O objetivo é encontrar padrões no comportamento das plantas daninhas em relação às culturas e às épocas em que estas se desenvolvem, de modo a prever seu aparecimento. Em parceria com a Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA), unidade Milho e Sorgo, foi elaborada uma base de dados unificada contendo informações de solo, clima e das populações amostradas na área experimental, os quais passaram por um pré-processamento. Na sequência foi feita uma análise estatística, na qual foi possível identificar diferenças entre as variáveis estudadas para os diversos elementos relacionados às plantas daninhas. Por fim, foram desenvolvidos alguns modelos de aprendizado de máquina. Dentre estes, destacam-se o modelo de predição de cultura e de predição da época de amostragem. A partir destes modelos, foi possível identificar as culturas e os períodos em que certas espécies tinham maior probabilidade de aparecer. Sendo que, nestes casos, os algoritmos com melhor desempenho foram a Árvore de Decisão – com 99% de acurácia – e o KNN – com 98,92% de acurácia –, respectivamente. Assim, este trabalho destaca que é viável utilizar algoritmos de aprendizado de máquina no manejo de plantas daninhas em sistemas ILP.

MACHINE LEARNING MODELS FOR PREDICTING WEED POPULATION DYNAMICS IN ICL SYSTEMS

Ana Letícia Becker Gomes

April / 2025

Advisor: Anita Maria da Rocha Fernandes, Dr.

Co-advisor: Maurílio Fernandes de Oliveira, Dr.

Area of Concentration: Applied Computer Science

Research Line: Applied Intelligence for the Environment

Keywords: Machine Learning; Weeds; ICL Systems.

Number of pages: 146

ABSTRACT

By the year of 2050, it is estimated that the world's population will be nine billion people. Continuous population growth has intensified the need to expand food production. Among the various obstacles faced by the agricultural system, one of the most notable is weeds. There are various weed management methods, with chemical control being the most widely used. However, while trying to increase food production, the environmental pollution caused by herbicides must also be reduced. This work applies machine learning algorithms and data analysis techniques to weed management in integrated crop-livestock systems (ICLS). The aim is to find patterns in the behaviour of certain weed species present in particular crops, in order to predict the appearance of these plants. In partnership with the Brazilian Agricultural Research Corporation (EMBRAPA), Maize and Sorghum unit, a unified database was created containing information on soil, climate and the populations sampled in the experimental area, which passed through a pre-processing stage. This was followed by a statistical analysis, in which it was possible to identify differences between the variables studied for the different elements related to weeds. Finally, some machine learning models were developed. Notable among these are the crop prediction model and the sampling season prediction model. Using these models, it was possible to identify the crops and periods in which certain species were most likely to appear. In these cases, the best performing algorithms were Decision Tree, with 99% accuracy and KNN, with 98.92 accuracy. This work highlights the feasibility of using machine learning algorithms in weed management in ICL systems.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AC	Árvore Complexa
AD	Árvore de Decisão
AM	Aprendizado de Máquina
ANN	*Artificial Neural Network*
BPNN	*Back-Propagation Neural Network*
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CNN	*Convolutional Neural Network*
DBSCAN	*Density based spatial clustering of applications with noise*
DCNN	*Deep Convolutional Neural Network*
DL	*Deep Learning*
DSSL	*Deep Semi-Supervised Learning*
EMBRAPA	Empresa Brasileira de Pesquisa Agropecuária
FR	Floresta Randômica
GAN	*Generative Adversarial Network*
IA	Inteligência Artificial
ICLS	*Integrated Crop-Livestock Systems*
ILP	Integração Lavoura-Pecuária
INMET	Instituto Nacional de Meteorologia
KELM	*Kernel Extreme Learning Machine*
KNN	*K-Nearest Neighbors*
MCA	Mestrado em Computação Aplicada
MLP	*Multilayer Perceptron*
MV	*Machine Vision*
OBIA	*Object-based image analysis*
PCA	Análise de Comportamentos Principais
PRISMA	Principais Itens para Relatar Revisões Sistemáticas e Meta-análises
PSO	*Particle Swarm Optimization*
RBF	*Radial Basis Function*
RCNN	*Region-Based Neural Network*
RL	Regressão Logística
SIMCA	*Soft Independent Modelling by Class Analogy*
SVM	*Support Vector Machin*e
TL	*Transfer Learning*
UNIVALI	Universidade do Vale do Itajaí
YOLO	*You only look once*

SUMÁRIO

1 INTRODUÇÃO

Nas últimas décadas, o crescimento populacional vem promovendo a intensificação no uso das áreas agrícolas para expandir a produção de alimentos. Diversos são os fatores que dificultam este processo, um destes é a presença de plantas daninhas nas plantações. Tem-se que as plantas daninhas são um empecilho na produção agrícola, pois competem por nutrientes, água, espaço, CO_2 e luz, além de causar alelopatia (KAUR *et al.*, 2018). Em adição podem ser hospedeiras de doenças, pragas e insetos (OLIVEIRA JUNIOR *et al.*, 2011). Há diversos métodos de manejo para essas espécies, tais como, cultural, mecânico (arranquio manual), físico, biológico e químico (uso de herbicidas); sendo o controle químico o mais utilizado (MONTEIRO & SANTOS, 2022).

Entretanto, observa-se que em Sistemas Integração Lavoura-Pecuária (ILP) – os quais caracterizam-se pela adoção da rotação e consórcio no cultivo de grãos com pastagens e, o pastejo por animais – a presença de plantas daninhas é menor do que em sistemas de lavoura contínua (OLIVEIRA JUNIOR *et al.*, 2011). Isso ocorre porque as plantas forrageiras formam uma cobertura do solo, o que previne a emergência das plantas daninhas (SCHUSTER *et al*., 2019). Ademais, tem-se que sistemas de rotação ajudam na redução do banco de sementes de plantas daninhas (IKEDA *et al*., 2007). Estes fatores contribuem para a redução no uso de herbicidas, minimizando os danos ambientais e os custos. Assim, para evitar a matocompetição e para minimizar os custos por reduzir a quantidade de produtos aplicados nos sistemas ILP, é preciso monitorar essas áreas de modo a caracterizar a dinâmica (aparecimento/desaparecimento) das diversas espécies. Para isso, ferramentas de análise de dados e aprendizado de máquina apresentam grande contribuição (JHA *et al*., 2019).

Existem diversos estudos que apresentam o uso de tecnologias aplicadas no manejo de plantas daninhas, tais como o trabalho de Monteiro *et al*. (2018), o qual identifica o período certo para realizar o controle das plantas daninhas. Também é possível fazer o reconhecimento e a classificação de espécies de plantas daninhas utilizando algoritmos de aprendizado de máquina,

como demonstrado por Costello *et al*. (2022); e, com isso, utilizar aplicadores inteligentes para fazer uma aplicação local do herbicida (YU *et al.*, 2019).

Observa-se, no entanto, que os trabalhos existentes e as tecnologias apresentadas, focam em realizar o manejo das plantas daninhas após estas já terem emergido e se desenvolvido. Dentro desse contexto, esta pesquisa aplica técnicas de análise de dados e aprendizado de máquina para compreender o comportamento de plantas daninhas em culturas em sistemas ILP. Como resultado, espera-se identificar os fatores ambientais de solo, de clima, e das culturas que contribuem para o crescimento ou supressão das espécies de plantas daninhas, de modo a antever possíveis interferências e adotar práticas de controle, incluindo as preventivas. Espera-se melhorar o manejo das plantas daninhas em sistemas ILP, reduzindo-as e, por consequência, aumentando a produtividade das áreas cultivadas, diminuindo o uso de herbicidas, e também os custos.

Os dados foram coletados no sistema ILP da Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA) de Milho Sorgo (Sete Lagoas, MG). As informaçẽos fornecidas são sobre as plantas daninhas, clima, solo, herbicidas e práticas culturais nos tratamentos. O período de análise dos dados é referente aos anos de 2006 e de 2015 até 2023 (OLIVEIRA *et al.* 2018).

1.1 PROBLEMA DE PESQUISA

Entender o comportamento das populações de plantas daninhas em culturas em sistemas ILP é um fator importante para evitar a matocompetição, aumentar a produtividade, e diminuir a utilização de herbicidas. Afinal, as plantas daninhas são responsáveis por 45% das perdas nas plantações (MONTEIRO & SANTOS, 2022) e – por consequência – por 4 milhões de agroquímicos aplicados nas culturas anualmente (OLIVEIRA *et al.* 2018).

Assim, compreender a dinâmica populacional destas espécies a partir de modelos preditivos de aprendizado de máquina, é algo a ser explorado. Com base neste problema, foram elaboradas as seguintes Perguntas de Pesquisa:

1. Qual é o comportamento das plantas daninhas em culturas em sistemas ILP?

2. Como os modelos preditivos de aprendizado de máquina podem ajudar no manejo das plantas daninhas?
3. Quais fatores influenciam no crescimento ou supressão destas populações?

1.1.1 Solução Proposta

Este trabalho propõe uma solução que tem como objetivo encontrar padrões no comportamento das plantas daninhas presentes em culturas, de modo a antever o surgimento ou não destas em sistemas ILP. Para isso são utilizadas técnicas de análise de dados e algoritmos de aprendizado de máquina supervisionado.

No Capítulo 3 é apresentada a revisão sistemática da literatura, na qual foram identificados os principais algoritmos de aprendizagem de máquina usados no contexto do manejo de plantas daninhas. Todavia, observa-se que a maioria destes modelos foca em processamento de imagens, pois uma grande parte dos trabalhos correlatos focam na identificação e classificação das espécies de plantas daninhas, não sendo utilizados para análise de dinâmicas populacionais.

Para a solução proposta, foi desenvolvida uma base de dados unificada, contendo informações de épocas de amostragem das plantas daninhas nas áreas cultivadas, de plantio/cultura, de clima e de solo. Tais dados foram aplicados em algoritmos de aprendizado de máquina supervisionado de classificação. Assim, este trabalho busca verificar a Hipótese: “É possível, a partir de certos fatores ambientais, compreender o comportamento de determinadas espécies de plantas daninhas em sistemas ILP, utilizando algoritmos de aprendizado de máquina.”

1.1.2 Delimitação de Escopo

Os algoritmos de aprendizado de máquina utilizados nesta pesquisa têm as seguintes delimitações: *(i)* análise de dados apenas de culturas em Sistemas ILP e *(ii)* análise de dados apenas das espécies de plantas daninhas e culturas presentes no sistema.

Neste trabalho não será abordado o comportamento de plantas daninhas em sistemas de lavoura contínua, nem serão analisados outros tipos de culturas. Ademais, não serão consideradas no modelo preditivo outras espécies de plantas daninhas além das observadas no sistema ILP na EMBRAPA de Milho e Sorgo. Os dados utilizados para a execução do trabalho são provenientes da EMBRAPA, dos anos de 2006 e de 2015 até 2023. Estas informações referem-se às populações das diferentes espécies de plantas daninhas; das amostras de solo; e os dados de clima do período. Dados de outras fontes, que datam de outros períodos, ou que não são sobre os aspectos analisados, não serão levados em consideração.

Os modelos serão treinados em um ambiente local, por não haver a necessidade de tantos recursos de *hardware*, nem de processamento gráfico. O tempo de execução dos algoritmos não será avaliado, já que o foco do estudo é a acurácia dos modelos e a análise de seus resultados.

1.1.3 Justificativa

Estima-se que até o ano de 2050 a população mundial será de 9 bilhões de pessoas (Organização das Nações Unidas, 2024). Esse aumento gera diversos desafios para os sistemas agrícolas, que precisam garantir alimento para todos os habitantes em meio a: mudanças climáticas; falta de disponibilidade de água e terras para produção de comida; doenças e pragas. Dentre todos esses problemas destacam-se as plantas daninhas, as quais são uma das principais e mais caras ameaças agrícolas (MONTEIRO & SANTOS, 2022). Apesar de tais plantas poderem ser controladas por meio de herbicidas, nos últimos anos houve uma pressão ambiental para reduzir a quantidade de agroquímicos – e também os custos de manejo. Visto que os herbicidas são responsáveis por contaminar solo, água e ar, além de causarem doenças em animais e seres humanos (RIBAS & MATSUMURA, 2009).

Para lidar com todos esses desafios, e garantir a segurança alimentar e a preservação do meio ambiente, observou-se que nos últimos anos houve um aumento do uso da computação no controle das plantas daninhas. Com relação a isso, há diversos estudos sobre a utilização de algoritmos de aprendizado de máquina no controle de plantas daninhas. Contudo, observa-se que a

maioria destes trabalhos usa tais modelos para o processamento de imagens, de modo a identificar e classificar as espécies de plantas daninhas.

As tecnologias disponíveis focam em sistemas capazes de localizar e identificar as plantas daninhas, às vezes em tempo real, de modo a realizar a aplicação localizada do herbicida sobre tais espécies. Esta aplicação pode ser feita por drones ou robôs de modo autônomo, e sem a necessidade de espalhar os herbicidas por toda a plantação.

Logo, percebe-se que o manejo de plantas daninhas por meio da análise do comportamento destas – utilizando técnicas de análise de dados e aprendizagem de máquina – é algo a ser desenvolvido. Desta forma, a criação de um modelo preditivo que auxilie no controle das plantas daninhas pode ser justificado pelo fato de não haver propostas de soluções com a mesma abordagem que a deste trabalho, e por contribuir com informações que permitam um manejo adequado e antecipado por parte dos produtores.

1.2 OBJETIVOS

Esta seção formaliza os objetivos do trabalho, conforme apresentado nas seguintes subseções.

1.2.1 Objetivo Geral

Implementar algoritmos de aprendizado de máquina capazes de prever a cultura e a época de amostragem na qual haverá a emergência de determinadas espécies de plantas daninhas em sistemas ILP, contribuindo para o entendimento de quais fatores ambientais influenciam na dinâmica populacional dessas plantas.

1.2.2 Objetivos Específicos

Os objetivos específicos deste trabalho são:

1. Elaborar uma base de dados unificada com as informações de plantio, clima e solo;
2. Identificar os algoritmos de aprendizado de máquina mais apropriados;
3. Avaliar o desempenho dos modelos quanto a acurácia de prever as culturas e as épocas em que as espécies de plantas daninhas irão aparecer.

1.3 METODOLOGIA

Esta seção apresenta a metodologia de pesquisa utilizada, bem como os procedimentos metodológicos adotados para alcançar a solução proposta no trabalho.

1.3.1 Metodologia da Pesquisa

Esta pesquisa usa o método hipotético-dedutivo, visto que foi desenvolvida uma hipótese para nortear o trabalho, a qual será aceita ou refutada com base nas evidências empíricas. Tem-se que este trabalho é de natureza aplicada, pois objetiva produzir conhecimentos para aplicações práticas destinados à solução de um problema específico. Neste projeto, busca-se gerar conhecimento por meio do desenvolvimento de algoritmos de aprendizado de máquina no contexto do manejo de plantas daninhas em sistemas ILP.

Em relação a abordagem do problema, esta é quantitativa, pois os resultados serão avaliados por meio de métodos estatísticos. Este trabalho procura verificar os algoritmos com melhor desempenho, utilizando métricas de acurácia, obtidas a partir de experimentos realizados com os modelos de aprendizado de máquina.

No que diz respeito aos objetivos da pesquisa, tem-se que esta é de caráter exploratório. Foi realizado um levantamento bibliográfico com o intuito de verificar as tecnologias de aprendizado de máquina utilizadas no manejo de plantas daninhas em sistemas ILP. Além disso, o trabalho também tem objetivos de pesquisa explicativos, os quais referem-se à análise de comportamento das espécies de plantas daninhas – em sistemas ILP – em relação aos fatores ambientais. Ademais, há objetivos preditivos, que são definidos pela construção de um modelo a base de inteligência artificial, capaz de predizer as culturas e a época nas quais aparecerão as plantas daninhas.

1.3.2 Procedimentos Metodológicos

Nesta pesquisa foram adotados os seguintes procedimentos metodológicos:

1. Revisão sistemática da literatura: tem como objetivo identificar o estado da arte sobre o tema da pesquisa, de modo a determinar as técnicas de aprendizado de máquina e os algoritmos utilizados no contexto do manejo de plantas daninhas em sistemas ILP.

2. Revisão bibliográfica: visa proporcionar a fundamentação teórica necessária para o desenvolvimento do trabalho, apresentando e discutindo os temas mais relevantes para esta pesquisa.

3. Preparação dos dados: nesta etapa é realizada a preparação dos dados de plantas daninhas nos períodos de amostragem; de plantio das culturas; de clima e de solo para serem utilizados – por meio de uma base unificada – no treinamento dos algoritmos. Nesta fase são feitas: a validação do formato dos dados; a remoção de *outliers*; o tratamento de dados faltantes; e a normalização dos dados. A preparação dos dados contempla o objetivo específico do item 1.

4. Desenvolvimento: primeiro, o conjunto de dados é dividido em dados de treinamento e teste. Seguindo, os algoritmos de aprendizado de máquina são aplicados, de forma a selecionar os modelos que melhor se adequam aos tipos de dados e aos objetivos propostos. Nesta fase, também é realizada uma análise estatística para ter uma compreensão mais geral dos dados.

Esta etapa, em conjunto com a revisão sistemática da literatura, atende ao objetivo específico 2.

5. Avaliação dos modelos de aprendizado de máquina: avaliar o desempenho dos algoritmos de aprendizado de máquina com base em algumas métricas selecionadas, adotando como padrão os resultados dessas métricas descritos nos trabalhos relacionados. O objetivo específico 3 está contemplado nesse estágio.

6. Análise dos resultados: analisar os resultados alcançados na fase anterior, de modo a verificar o cumprimento da hipótese da pesquisa, e comparar com os resultados obtidos pelos trabalhos correlatos e das observações empíricas.

1.4 ESTRUTURA DA DISSERTAÇÃO

O presente trabalho está dividido em cinco capítulos. O Capítulo 1 contempla a Introdução da pesquisa – justificativa do problema, solução proposta, objetivos e metodologia. O Capítulo 2 apresenta a Fundamentação Teórica sobre os temas: manejo de plantas daninhas, sistemas ILP, estatística, e aprendizado de máquina. No Capítulo 3 está detalhado o processo feito para a Revisão Sistemática da Literatura e os trabalhos correlatos selecionados. O Capítulo 4 mostra o desenvolvimento do projeto, desde a criação das bases de dados até as métricas obtidas nos algoritmos de aprendizado de máquina desenvolvidos, além das análises estatísticas realizadas. O Capítulo 5 contempla os Resultados, isto é, as informações que foram extraídas tanto dos modelos de aprendizado de máquina quanto das análises estatísticas. Por último, no Capítulo 6 há as Considerações Finais do trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

Para uma melhor compreensão do trabalho, serão apresentados neste capítulo os principais conceitos considerados relevantes para o desenvolvimento do projeto. A Seção 2.1 aborda o problema das plantas daninhas na agricultura; a Seção 2.2 apresenta o funcionamento dos sistemas ILP; e a Seção 2.3 explica como estes assuntos estão relacionados. Já a Seção 2.4 e 2.5 trazem os principais conceitos de estatística e aprendizado de máquina, respectivamente. Por último, a Seção 2.5 descreve o funcionamento dos algoritmos de aprendizado de máquina utilizados no desenvolvimento da solução proposta.

2.1 PLANTAS DANINHAS

Planta daninha é uma planta não semeada e que emerge nas culturas, acarretando perdas nas lavouras, de modo que requer um gerenciamento para que não interfira na plantação (MONTEIRO & SANTOS, 2022). As plantas daninhas competem com as culturas por diversos recursos naturais, tais como: nutrientes, água, espaço, CO_2 e luz; além de serem hospedeiras de insetos e patógenos, como fungos e bactérias, e causarem alelopatia (efeito que uma planta tem sobre outra) (OLIVEIRA JUNIOR *et al.*, 2011). Isto é, algumas espécies de plantas daninhas produzem aleloquímicos para desordenar os processos fisiológicos das culturas e, assim, inibir seu crescimento e desenvolvimento (KUBIAK *et al.*, 2022).

De acordo com Monteiro e Santos (2022), as plantas daninhas diminuem a produtividade e a qualidade das culturas produzidas. A presença das plantas daninhas nas lavouras pode resultar em 100% de perda da plantação, caso não sejam controladas (CHAUHAN, 2020). Além disso, algumas espécies de plantas daninhas produzem substâncias que causam alergias em seres humanos e animais (KUBIAK *et al.*, 2022).

Por outro lado, as plantas daninhas – tanto emergidas quanto do banco de sementes – estão presentes na natureza e não podem, nem devem, ser completamente erradicadas (MONTEIRO & SANTOS, 2022; KUBIAK *et al.*, 2022). Afinal, também deve-se considerar que as plantas daninhas

são um indicador valioso de biodiversidade e podem ter um efeito positivo no ecossistema. Como exemplo, tem-se que as plantas daninhas estão envolvidas no ciclo e balanceamento de nutrientes; previnem a erosão do solo; e podem servir de habitat para insetos e microrganismos benéficos, o que aumenta a biodiversidade da microfauna e da microflora (KUBIAK *et al.*, 2022).

Também tem-se que a redução excessiva de plantas daninhas pode causar alterações nos ecossistemas. Por exemplo, Monteiro e Santos (2022) afirmam que uma grande proporção do declínio de pássaros em fazendas, tem sido associada à diminuição de ocorrência de plantas daninhas em plantações aráveis.

Atualmente, há uma busca por novos métodos de controle das plantas daninhas sustentáveis. Diversos são os métodos de manejo existentes, sendo que os mais utilizados são químico, mecânico e manual. Todavia, há outros meios de controle não tão convencionais. Todos estes processos serão brevemente explicados na sequência, bem como serão apresentadas suas vantagens e limitações.

O controle manual ou a catação, comum em áreas pequenas ou em horticultura, consiste na remoção das plantas daninhas do solo por uma pessoa. No entanto, além de ser um processo tedioso e consumir muito tempo, também é ineficiente e economicamente inviável (YADURAJAU & RAO, 2013). *Mulching,* ou cobertura morta, é uma técnica não convencional em que o solo é coberto com resíduos e restos de plantas, formando uma cobertura morta. O que previne a germinação das sementes e a emergência das espécies de plantas daninhas. Contudo, tem um custo elevado e pode causar alterações no solo se usada continuamente com o mesmo material (MONTEIRO & SANTOS, 2022).

Também há a opção de plantações de coberturas vivas. Estes resíduos vivos são plantados previamente ou ao mesmo tempo que a cultura principal, impedindo as plantas daninhas de emergirem. Porém, deve-se tomar cuidado para que estas plantas forrageiras não se tornem plantas daninhas e comecem a competir pelos recursos naturais com as culturas. Já o método de solarização do solo consiste em colocar uma cobertura de plástico sobre a superfície do solo, para prender a radiação solar e promover o aumento da temperatura. Pode ser feito em estufas ou em campos

abertos, já que é restrito apenas a culturas menores. Todavia, a indução de altas temperaturas pode ser letal para fungos e bactérias presentes no solo (MONTEIRO & SANTOS, 2022).

O controle termal é uma técnica que usa fogo, água quente e vapor; fornecendo um controle rápido das plantas daninhas, sem deixar resíduos químicos. Entretanto, o uso de água e combustíveis fósseis é muito alto (MONTEIRO & SANTOS, 2022), além de que o último também contamina o ar. Outro método não convencional para reduzir as plantas daninhas é o pastoreio de gado, o qual tem como objetivo manipular os padrões de desfolhação para colocar uma determinada espécie em desvantagem competitiva, em relação às demais plantas da comunidade (MONTEIRO & SANTOS, 2022).

Já o controle mecânico requer o uso de ferramentas, de modo que estas destroem/removem as plantas daninhas do solo ou reduzem sua habilidade competitiva, mas não é eficiente para obter um controle total e efetivo. Este método apresenta como desvantagem o revolvimento do solo, que pode prejudicar diversos aspectos do deste, como estrutura, diversidade biológica e armazenamento de água (MONTEIRO & SANTOS, 2022).

Por último, tem-se que o controle químico por meio de herbicidas é o método de manejo mais utilizado atualmente. Tais agroquímicos contribuem para redução de plantas daninhas pois afetam: o processo de fotossíntese; o metabolismo; a divisão e o crescimento das células das plantas; entre outros (KUBIAK *et al.*, 2022). Podem ser classificados pela época de aplicação, mecanismos de ação ou seletividade das espécies (MONTEIRO & SANTOS, 2022).

No entanto, deve-se tomar cuidado, pois se aplicados em doses menores do que o necessário tornam-se ineficientes (SABZI & ABBASPOUR-GILANDEH, 2018). Enquanto se aplicados em doses muito altas, os herbicidas causam a contaminação de diversos recursos naturais – como água, solo e ar –, deixam resíduos na cadeia alimentar, e podem causar doenças em seres humanos e animais (RIBAS & MATSUMURA, 2009). De acordo com Partel *et al.* (2019), 98,9% dos produtos alimentícios contém resíduos de agroquímicos.

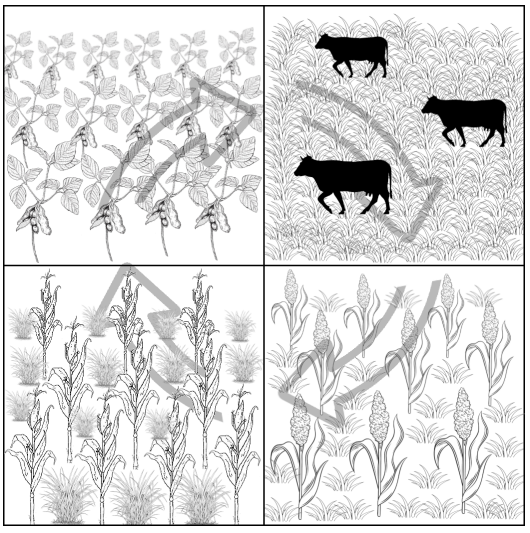
Além disso, o uso contínuo dos mesmos agroquímicos pode levar as plantas daninhas a desenvolverem resistência a estas substâncias, de modo que seu controle não pode mais ser feitos por meio dos herbicidas disponíveis, sendo assim necessário buscar outros métodos de manejo (KUBIAK *et al.*, 2022).

Assim, tem-se que o controle de plantas daninhas de modo eficiente e sustentável é um assunto relevante, com impacto significativo em toda a sociedade. Diante deste cenário, o uso de diferentes tecnologias no manejo de plantas daninhas está crescendo (JHA *et al*., 2019).

2.2 SISTEMAS ILP

O Sistema Integração Lavoura-Pecuária (ILP) engloba dois sistemas de produção: agricultura e pecuária (DUARTE *et al.*, 2018). De acordo com Balbinot Junior *et al*. (2009), o ILP constitui de um sistema de rotação e consórcio, que alterna na mesma área o cultivo de pastagens anuais ou perenes, destinadas à alimentação animal; e culturas destinadas à produção vegetal, sobretudo grãos. Tal funcionamento pode ser visualizado na Figura 1.

Figura 1: Esquema de funcionamento do Sistema ILP.



Fonte: Elaborada pela autora.

A produção animal pode ser de bovinos de corte, bovinos de leite, ovinos ou caprinos. Enquanto a produção vegetal é constituída por culturas como soja, milho, fumo, feijão, entre outras (BALBINOT JUNIOR, *et al.*, 2009). Os países que mais utilizam o sistema ILP são Brasil, Austrália e Nova Zelândia (GARRETT *et al.,* 2017). Contudo, há outros países que também estão implementando esse tipo de modelo de cultivo, tais como Estados Unidos, França e Índia (SEKARAN *et al.*, 2021).

Os sistemas ILP proporcionam diversos benefícios tanto biológicos quanto econômicos. Como vantagens, tem-se a elevada velocidade de ciclagem dos nutrientes em um curto intervalo de tempo, tanto via urina e fezes dos animais quanto pela degradação da palhada. Essa aceleração de ciclagem dos nutrientes também pode reduzir perdas por erosão e lixiviação (BALBINOT JUNIOR, *et al.*, 2009). Assim, o sistema ILP é efetivo para renovação de pastagens degradadas (KLUTHCOUSKI *et al.*, 2007).

Também tem-se que, devido ao crescimento contínuo de plantas na área, a quantidade de carbono orgânico aumenta ao longo do tempo, o que melhora a qualidade do solo (BALBINOT JUNIOR, *et al.*, 2009). Ademais as plantas forrageiras são resistentes a muitas pragas e doenças, quebrando os ciclos de agente bióticos nocivos às culturas. Não obstante, o sistema ILP é efetivo para redução de plantas daninhas. Tudo isso contribui para diminuição do uso de herbicidas em sistemas ILP (KLUTHCOUSKI *et al.*, 2007). Por último, os sistemas ILP contribuem no condicionamento do solo, na qualidade das fontes de água, e na conservação biológica (DUARTE *et al.*, 2018).

Como vantagens econômicas pode-se apontar: diversificação da renda, resultante da produção animal e vegetal na mesma área; aumento da renda, devido ao uso contínuo das áreas agrícolas; redução dos custos, em função dos benefícios biológicos (BALBINOT JUNIOR, *et al.*, 2009); e redução da necessidade de ampliação de novas terras para cultivo (DUARTE *et al.*, 2018).

A maior preocupação sobre o sistema ILP é a compactação do solo, devido ao pisoteio dos animais. A compactação altera a estrutura do solo deixando-o menos poroso, o que ocasiona a baixa infiltração de água, de nutrientes e a difusão de gases. Tudo isso pode gerar erosão superficial e

prejudicar o desenvolvimento das plantas (BALBINOT JUNIOR, *et al.*, 2009; DUARTE *et al.*, 2018).

Além disso, Duarte *et al.* (2018) também alertam que apesar das diversas vantagens do sistema ILP, implantação e manejos errôneos produzirão efeitos negativos no solo, os quais influenciarão a produtividade da plantação.

2.3 PLANTAS DANINHAS EM SISTEMAS ILP

Como mencionado anteriormente, a presença de plantas daninhas em sistemas ILP é menor do que em sistemas de lavoura contínua. Afinal, a constituição dos sistemas ILP inclui o pastoreio por animais e as coberturas vegetais que tendem a reduzir a população de plantas daninhas.

Assim, tem-se que as plantas forrageiras cultivadas junto com as culturas formam uma cobertura do solo, impedindo as plantas daninhas de emergirem (KLUTHCOUSKI *et al.*, 2007; CONCENÇO *et al.*, 2015). Ademais, os sistemas ILP ajudam a reduzir o banco de sementes de plantas daninhas no solo (IKEDA *et al.*, 2007).

2.4 ANÁLISE ESTATÍSTICA

A análise estatística é um conjunto de métodos e técnicas utilizadas para revisar, analisar e ter uma melhor compreensão acerca dos dados utilizados. Nesta seção serão apresentadas algumas das técnicas empregadas no desenvolvimento do projeto.

2.4.1 Análise descritiva

É uma abordagem estatística cujo objetivo é descrever e resumir as principais características de um conjunto de dados. A partir desta análise é possível ter uma visão mais geral dos dados, contribuindo para uma melhor compreensão destes, antes de realizar análises mais complexas (FREUND, 2009). A estatística descritiva é composta por: medidas de tendência central; medidas de dispersão; e distribuição de frequência. Dentre estas, serão citadas apenas as utilizadas ao longo do trabalho.

As medidas de tendência central focam em descrever o centro dos dados. A primeira delas é a Média Aritmética – muitas vezes denominada apenas de média –, a qual é o indicador mais comum de tendência central de uma variável (FREUND, 2009). Ela é a soma dos valores das observações dividida pelo número total de observações:

$$\mu_x = \sum_{i=1}^{n} x_i \quad (1)$$

Para a Mediana, primeiro, classifica-se os dados do menor para o maior. Assim, a mediana separa a metade superior da metade inferior dos dados, ou seja, é o valor do meio. Caso a quantidade de observações seja par, calcula-se a média aritmética dos dois valores do meio e esse resultado passa a ser a mediana. Por último, a Moda refere-se ao valor ou categoria que ocorre com maior frequência. Esta medida não exige nenhum cálculo, apenas contagem; e pode-se utilizá-la tanto para dados numéricos quanto para categóricos (FREUND, 2009).

Já as medidas de dispersão apresentam a variabilidade dos dados. A primeira destas é o Desvio Padrão, o qual indica o quão próximo os dados estão da média:

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu_x)^2}{n}} \quad (2)$$

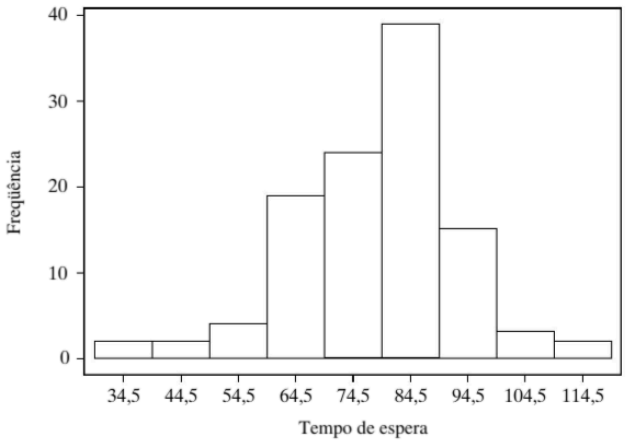
O desvio padrão mostra o quanto um conjunto de dados é uniforme, logo, valores altos desta medida indicam que os dados estão espalhados por uma ampla gama de valores, enquanto valores

baixos apontam que os dados estão perto da média. A Variância também mede a dispersão dos dados em relação à média. Valores altos de variância indicam que os dados não são homogêneos, já valores baixos demonstram o contrário (FREUND, 2009). A variância é dada pelo quadrado do desvio padrão:

$$\sigma_x^{\,2} = \frac{\sum_{i=1}^{n} (x_i - \mu_x)^2}{n} \quad (3)$$

Quanto à distribuição de frequência, tem-se que estas são utilizadas para visualizar a distribuição dos dados. Uma das maneiras de visualizar os dados é por meio de histogramas, que mostram a distribuição de frequência de dados contínuos. Em outras palavras, é um gráfico de barras em que a base de cada barra representa uma classe e a altura representa a frequência absoluta de cada classe (FREUND, 2009). A Figura 2 traz um exemplo de histograma.

Figura 2: Exemplo de histograma dos tempos de espera entre erupções de gêiseres.



Fonte: Freund, 2009.

A partir dos histogramas é possível verificar que tipo de distribuição os dados seguem. Por exemplo, pode-se verificar se o conjunto de dados apresenta uma distribuição normal. Tem-se que a distribuição normal (ou distribuição gaussiana) é uma distribuição de probabilidade contínua que é simétrica ao redor da média, de modo que a maior parte das observações estão agrupadas ao redor do pico central (FREUND, 2009). Os dados são normalmente distribuídos quando:

[i] A curva tem formato de sino;

[ii] Há uma simetria em torno do centro;

[iii] Os valores da média, mediana e moda são iguais;

[iv] A área total embaixo da curva é 1;

[v] Há apenas uma moda;

[vi] A curva nunca toca o eixo x e;

[vii] Quando calcula-se o desvio padrão da média, tem-se que: 68% dos dados estão dentro de um desvio padrão da média, 95% dos dados estão dentro de dois desvios padrões da média e, 99% dos dados estão dentro de três desvios padrões da média.

Muitos fenômenos naturais seguem uma distribuição (aproximadamente) normal como: altura das pessoas, erros das medições, pressão sanguínea, entre outros. Além disso, ela possibilita realizar aproximações para calcular probabilidades de muitas variáveis aleatórias que tem outras distribuições. Vale destacar que a distribuição normal é simétrica, mas nem toda distribuição simétrica é normal (FREUND, 2009).

2.4.2 Testes Não-Paramétricos

Testes não-paramétricos são utilizados para testar hipóteses sobre um conjunto de dados. Em particular, são usados quando a distribuição da população não é normal e/ou quando algum parâmetro não pode ser estimado. As vantagens dos testes não-paramétricos é que estes podem ser usados em amostras de tamanho pequeno e com dados do tipo ordinal. Já as desvantagens são que eles tendem a desperdiçar informações e – caso a normalidade possa ser assumida –, os testes paramétricos são mais poderosos (FREUND, 2009).

Há diversos tipos de testes não-paramétricos, como exemplo pode-se citar o Teste de Sinais; Teste de Wilcoxon; Teste de Mann-Whitney; Teste de Kruskal-Wallis e; Teste de Friedmann. Será abordado em mais detalhes apenas o Teste de Kruskal-Wallis, já que este foi o único teste não-paramétrico utilizado no desenvolvimento do trabalho.

O Teste de Kruskal-Wallis (ou Teste H), é um teste de soma de postos que serve para verificar se k amostras aleatórias independentes provêm ou não de populações idênticas. Isto é, testa-se a hipótese nula de que $\mu_1 = \mu_2 = \ ... = \mu_k$ contra a hipótese alternativa de que essas médias não são todas iguais – ou seja, pelo menos uma delas é diferente das demais. Assim, tem-se que esse teste é utilizado para testar hipóteses sobre três ou mais amostras independentes. Vale ressaltar que tais amostras não precisam ter o mesmo tamanho (FREUND, 2009).

O teste funciona da seguinte forma: escreve-se todos os dados de todas as amostras em ordem crescente, como se fosse uma só amostra; atribui-se os postos; soma-se os postos das i-ésimas amostras, e também a quantidade de registros de cada uma das amostras; e a partir desses valores calcula-se H conforme segue:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1) \quad (4)$$

Tem-se que a distribuição amostral de H se aproxima de uma distribuição qui-quadrado com k-1 graus de liberdade. Isso significa que rejeita-se a hipótese nula quando o valor obtido para H é maior ou igual ao X^2_{α} para k-1 graus de liberdade (FREUND, 2009). Isto é

$$H \geq X^2_{\alpha} \Rightarrow rejeitar\ a\ hipótese\ nula,$$

$$H < X^2_{\alpha} \Rightarrow aceitar\ a\ hipótese\ nula.$$

Outra maneira de obter o resultado do teste de Kruskal-Wallis, é calcular o p-valor, por meio de algum programa computacional. Desta forma, se o p-valor for menor que o nível de significância – que vale 0,05 normalmente – rejeita-se a hipótese nula, caso contrário aceita-a.

Todavia, tem-se que o teste de Kruskal-Wallis informa se há uma diferença nas médias das amostras ou não, mas não explicita quais são as amostras que possuem essa diferença. Para encontrar estes casos, pode-se utilizar as Comparações Múltiplas Dwass-Steel-Critchlow-Fligner (DSCF).

O teste DSCF também é empregado no caso da distribuição dos dados não ser normal. O objetivo deste método é verificar quais grupos são significativamente diferentes entre si. Para isso, as comparações são feitas com base nos valores da mediana e/ou na distribuição dos dados de cada grupo. Neste caso, também tem-se que, caso o p-valor seja menor que o nível de significância, rejeita-se a hipótese nula, ou seja, concluí-se que há uma diferença entre os grupos comparados. Em outras palavras, o teste de Kruskal-Wallis afirma que há uma diferença entre os grupos, enquanto as comparações DSCF mostram em quais instâncias está essa diferença.

2.4.3 Correlação

A análise de correlação é uma técnica estatística que mostra como as variáveis estão relacionadas, ou seja, ela é utilizada para verificar se uma característica pode ser usada para prever outra. Ademais, tem-se que a força e a direção das correlações também podem ser determinadas pela análise de correlação. Assim, tem-se que os coeficientes de correlação variam entre -1 e 1, indicando o quão forte é a correlação e a direção em que ela vai (GRUS, 2021).

As correlações podem ser positivas ou negativas. Correlações positivas referem-se aos casos em que quanto maior os valores da variável A, maiores os valores da variável B. Caso o resultado do coeficiente seja 1, diz-se que há uma correlação positiva perfeita. Já nas correlações negativas, enquanto os valores da variável A aumentam, os da variável B diminuem (ou vice-versa). Nesse caso, se o resultado do coeficiente for -1, afirma-se que a correlação é negativa perfeita (GRUS, 2021).

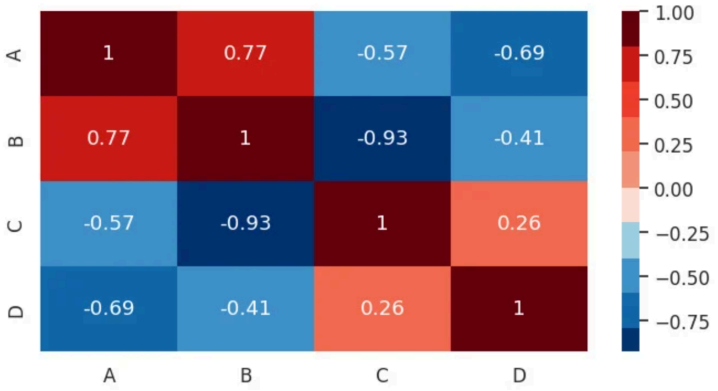
Há dois tipos de coeficientes que podem ser utilizados para calcular a correlação entre as variáveis. O primeiro destes é o Coeficiente de Pearson, o qual quantifica a relação linear entre duas

variáveis contínuas. Assim, tem-se que este coeficiente mede apenas relações lineares. Caso a relação entre as variáveis não seja linear, então este método não é adequado para calcular as correlações. Além disso, ele só pode ser utilizado caso a distribuição dos dados seja normal (GRUS, 2021).

Também pode-se utilizar o Coeficiente de Spearman para calcular a correlação entre duas variáveis ordinais ou contínuas. Este coeficiente é empregado quando os dados não estão distribuídos normalmente, sendo assim uma medida não-paramétrica. Ou também quando a relação entre as variáveis não é linear (GRUS, 2021).

Vale destacar que a correlação não implica em causalidade, isto é, quando uma ação A causa um resultado B. As maneiras mais comuns de visualizar as correlações entre os dados é utilizando matrizes de correlação e mapas de calor (GRUS, 2021). A Figura 3 apresenta um exemplo de mapa de calor.

Figura 3: Exemplo de mapa de calor entre as variáveis A, B, C e D.



Fonte: Alura, 2024.

2.5 APRENDIZADO DE MÁQUINA

Russell e Norvig (2022) apresentam o seguinte cenário: um certo agente aprenderá a melhorar o seu desempenho depois de realizar observações sobre o mundo. Porém, quando o agente é um computador, tal processo denomina-se Aprendizado de Máquina (AM) – ou *Machine*

Learning do termo original em inglês – no qual um computador recebe e analisa alguns dados; e com base nesses dados, monta um modelo para resolver uma determinada tarefa.

Assim, tem-se que AM é uma área da Inteligência Artificial (IA), cujo objetivo é o desenvolvimento de programas de computador capazes de aprender por conta própria, a partir de um conjunto de dados – o qual representa as experiências passadas –, e com isso executar uma dada tarefa (COELHO *et al.*, 2018).

O aprendizado de máquina engloba conhecimentos de probabilidade e estatística; teoria da informação; teoria da complexidade computacional; visão computacional; processamento de linguagem natural; entre outros (COELHO *et al.*, 2018).

Os modelos de AM podem ser classificados de acordo com o tipo de aprendizagem, isto é, se são treinados ou não com supervisão humana. Os tipos de aprendizado são: supervisionado, não-supervisionado, semi-supervisionado e por reforço.

2.5.1 Tipos de aprendizado

Nas seções seguintes são apresentados os tipos de supervisão que os sistemas de AM podem receber durante seu treinamento.

2.5.1.1 Aprendizado Supervisionado

Conforme Russell e Norvig (2022), na aprendizagem supervisionada o agente recebe alguns exemplos de pares de entrada e saída e, com isso, aprende uma função que faz o mapeamento entre elas. Para isso, é necessário que a classe de cada observação – também chamada de rótulo – seja conhecida.

Um exemplo clássico é o filtro de *spam*. Neste caso, o algoritmo é treinado para classificar os *e-mails* como *spam* ou não-*spam*, com base em exemplos rotulados destas duas classes. Assim,

por meio das características dos *e-mails* – como comprimento do texto e presença de *links* – o modelo aprende a rotular os *e-mails* (BHATIA & KALUZA, 2018).

Problemas desse tipo são chamados de problemas de classificação. Em casos assim, o modelo identifica os padrões das classes, de modo que novos dados de entrada possam ser rotulados. O aprendizado supervisionado também pode ser utilizado em problemas de regressão, cujo objetivo é prever um rótulo numérico. Por exemplo, o valor de um empréstimo que um banco faria com base nas características do cliente (BHATIA & KALUZA, 2018).

De acordo com Geron (2019), os principais algoritmos de aprendizado supervisionado são:

[i] K-Vizinhos mais Próximos (KNN);

[ii] Regressão Logística;

[iii] Árvore de Decisão;

[iv] Floresta Randômica;

[v] Redes Neurais Artificiais; e

[vi] Máquinas de Vetores de Suporte (SVM).

2.5.1.2 Aprendizado Não-Supervisionado

No aprendizado não-supervisionado o agente aprende os padrões dos dados de entrada sem que seja informada a classe a qual o dado pertence. Isto é, os algoritmos são treinados com dados não rotulados (RUSSELL & NORVIG, 2022).

O tipo de problema mais comum que utiliza aprendizado não-supervisionado é o agrupamento (*clustering*), o qual pode ser descrito como o processo de agrupar e categorizar grupos de dados. Por exemplo, em *sites* de *e-commerce* os sistemas de recomendação de produtos usam

técnicas de agrupamento para descobrir os padrões de compra dos clientes (BHATIA & KALUZA, 2018).

Assim, tem-se que o agrupamento – também denominado de clusterização – consiste em comparar dados não rotulados com base em suas características comuns. O aprendizado não-supervisionado também pode ser utilizado em problemas de regra de associação e redução de dimensionalidade (BHATIA & KALUZA, 2018).

Os principais algoritmos de aprendizado não-supervisionado são (GERON, 2019):

[i] *K-Means*;

[ii] Apriori;

[iii] Análise de Componentes Principais (PCA); e

[iv] DBSCAN (*Density based spatial clustering of applications with noise*).

2.5.1.3 Aprendizado Semi-Supervisionado

O aprendizado semi-supervisionado é uma combinação das abordagens supervisionada e não-supervisionada. Neste caso, são dados para o modelo poucos exemplos rotulados e muitos exemplos não rotulados (RUSSELL & NORVIG, 2022). Pode-se dizer que este tipo de aprendizado utiliza a lógica “tentativa e erro”, ou seja, a aprendizagem do sistema é feita com base na sua experiência e ele aprende com os seus próprios erros (BHATIA & KALUZA, 2018).

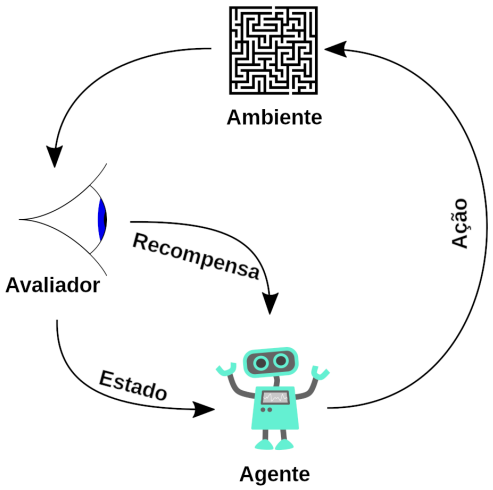
Essa forma de aprendizado é utilizada quando o custo de rotulação dos dados é muito alto (BHATIA & KALUZA, 2018). Por exemplo, não seria viável que uma pessoa lesse uma grande quantidade de textos inteiros para poder classificá-los. Assim, o aprendizado semi-supervisionado seria aplicado para fazer um classificador de documentos de texto, a partir de alguns exemplos de textos rotulados.

2.5.1.4 Aprendizado por Reforço

Já no aprendizado por reforço, o agente é treinado a partir de uma série de reforços, os quais podem ser recompensas ou punições. O objetivo é que o agente consiga maximizar as recompensas, aprendendo, assim, quais são as melhores ações a serem tomadas. Dessa forma, uma ação com resultados positivos será reforçada, enquanto ações de resultados desfavoráveis serão desencorajadas (RUSSELL & NORVIG, 2022).

Tal aprendizado é bastante utilizado em jogos e robótica. Por exemplo, em um jogo de xadrez o agente (computador) é comunicado se ele ganhou (recompensa) ou se ele perdeu (punição). Assim, o agente analisa quais as ações tomadas antes do reforço que produziram o resultado obtido, de modo a alterar suas ações em situações futuras para obter a recompensa (RUSSELL & NORVIG, 2022) , tal como pode ser visto na Figura 4.

Figura 4: Exemplo de aprendizado por reforço.



Fonte: Wikiversidade (2023)

2.5.2 Métricas de avaliação

Tem-se que a situação problema apresentada nesta dissertação pode ser considerada como um problema de classificação. Portanto, para o desenvolvimento da solução proposta serão

utilizados algoritmos de aprendizado supervisionado. Os modelos desta forma de aprendizagem tem seu desempenho avaliado pelas métricas descritas a seguir.

2.5.2.1 Matriz de Confusão

Para obter uma melhor visualização da performance dos algoritmos, pode-se utilizar a matriz de confusão. Tem-se que as linhas representam as instâncias verdadeiras enquanto as colunas representam as instâncias preditivas, assim como demonstra o Quadro 1. O contrário também poderia acontecer, isto é, as preditivas estarem nas linhas e as verdadeiras nas colunas (BHATIA & KALUZA, 2018).

Quadro 1: Modelo de matriz de confusão.

		Predito como positivo?	
		Sim	Não
Realmente positivo?	Sim	VP - Verdadeiro Positivo	FN - Falso Negativo
	Não	FP - Falso Positivo	VN - Verdadeiro Negativo

Fonte: Bhatia e Kaluza (2018).

As variáveis VP (Verdadeiro Positivo), VN (Verdadeiro Negativo), FP (Falso Positivo) e FN (Falso Negativo) da imagem, são as comparações entre as instâncias verdadeiras e preditivas. Observa-se que os valores verdadeiros ficam sempre na diagonal principal (BHATIA & KALUZA, 2018).

2.5.2.2 Acurácia

A acurácia é a proporção entre os valores verdadeiros e o total:

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \qquad (5)$$

Entretanto, tal métrica não é recomendada quando há um grande desequilíbrio nas classes. Pois, apesar de obter bons resultados, o modelo fará a previsão de acordo com a classe majoritária; podendo não ser útil para o contexto da situação problema. Esse fenômeno é chamado de Paradoxo da Acurácia (BHATIA & KALUZA, 2018).

2.5.2.3 Precisão

Precisão – ou do inglês *Precision* – é uma medida de relevância da previsão (GAO *et al.*, 2018). O objetivo é calcular a taxa de classificações positivas. Dessa forma, tem-se que a Precisão descreve a proporção de identificações positivas que estão corretas (COSTELLO *et al*., 2022). Isto é, dentre todas as classes que realmente eram positivas, quantas o modelo considerou como positivas, ou seja:

$$Precisão = \frac{VP}{VP + FP} \qquad (6)$$

Valores altos de Precisão indicam que das identificações feitas, poucas destas estão erradas (COSTELLO *et al.*, 2022)

2.5.2.4 Sensibilidade

Já a Sensibilidade – em inglês denominado *Recall* – descreve a proporção de objetos reais que foram identificados. Ou seja, o objetivo é verificar dentre todas as classes positivas identificadas quantas foram classificadas corretamente (COSTELLO *et al.*, 2022), isto é:

$$Sensibilidade = \frac{VP}{VP + FN} \qquad (7)$$

Valores de Sensibilidade altos indicam que dada uma determinada classe, a maior parte de seus registros será classificada corretamente (COSTELLO *et al.*, 2022).

2.5.2.5 F1 Score

O *F1 Score* é a média harmônica entre a Precisão e Sensibilidade:

$$F1\ Score\ =\ 2(\frac{Precisão \cdot Sensibilidade}{Precisão + Sensibilidade}) \qquad (8)$$

Assim, o *F1 Score* é uma medida de compromisso entre a Precisão e a Sensibilidade, atribuindo o mesmo grau de importância para ambas (BHATIA & KALUZA, 2018). Vale ressaltar que alguns sistemas podem ter Precisão elevada e Sensibilidade baixa, ou vice-versa. Quando isto ocorre, o *F1 Score* é impactado negativamente (COSTELLO *et al.*, 2022).

O valor do *F1 Score* varia entre 0 e 1, sendo que quanto mais próximo de 1, melhor o modelo (BHATIA & KALUZA, 2018). Valores altos de *F1 Score* implicam em valores elevados de Precisão e Sensibilidade, o que sugere valores altos para VP e VN, e valores baixos para FP e FN (COSTELLO *et al.*, 2022).

2.6 ALGORITMOS DE APRENDIZADO DE MÁQUINA

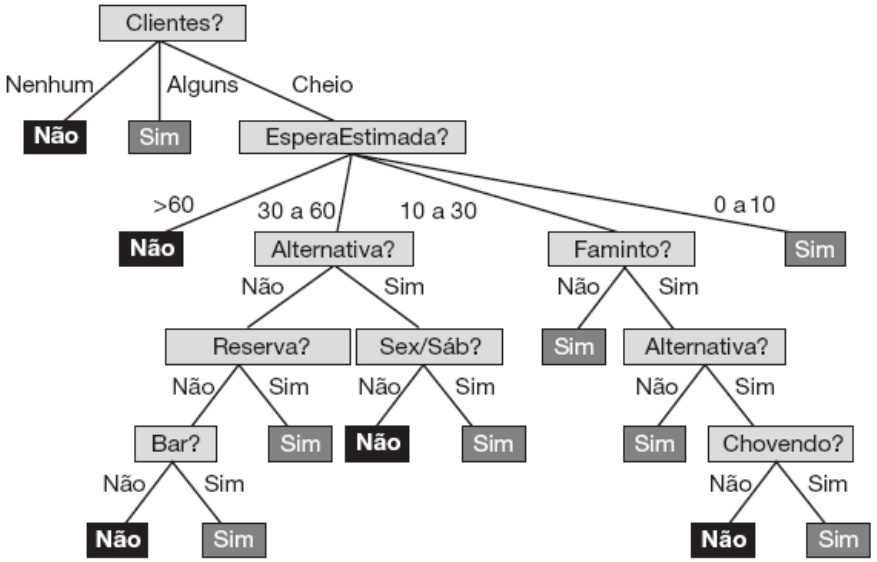
Nesta seção são apresentados os algoritmos selecionados para o desenvolvimento da solução proposta. Como mencionado anteriormente, o problema abordado é um problema de classificação, de modo que modelos de aprendizado supervisionado foram elegidos. Estes são: Árvore de Decisão, Floresta Randômica, Máquinas de Vetores de Suporte (SVM) e K-Vizinhos mais Próximos (KNN).

2.6.1 Árvore de Decisão

Uma árvore de decisão – do inglês *decision tree* – recebe como entrada um conjunto de dados, no formato de vetor ou matriz, e com base nesses dados consegue retornar uma decisão em forma de um único valor de saída. Para o modelo tomar uma decisão, ele realiza uma série de testes, começando na raiz da árvore e seguindo por um ramo adequado até chegar numa folha. Cada nó interno corresponde a um teste de valor de um dos atributos de entrada. Os ramos dos nós são

possíveis valores que o atributo pode assumir e os nós da folha são o valor que deve ser retornado pela função (RUSSELL & NORVIG, 2022). A Figura 5 ilustra um exemplo de árvore de decisão para decidir a espera ou não de uma mesa de restaurante.

Figura 5: Exemplo de árvore de decisão, sobre espera por uma mesa de restaurante.



Fonte: Russell e Norvig (2022).

Para muitos problemas o resultado da árvore de decisão é conciso. Todavia, algumas funções não podem ser representadas de forma sucinta, pois seu resultado produz uma árvore de decisão exponencialmente grande. Entretanto, mesmo para problemas de dimensão maior, pode-se encontrar uma solução aproximada. Afinal tem-se que a árvore de decisão utiliza a estratégia de um algoritmo guloso: “dividir para conquistar”, testando primeiro o atributo mais importante (RUSSELL & NORVIG, 2022).

Dessa forma, o problema é dividido em subproblemas menores, que podem ser resolvidos de maneira recursiva. Vale destacar que o atributo “mais importante” é aquele que mais fez diferença na classificação de um exemplo. Assim, com um pequeno número de testes, espera-se obter a classificação correta, de modo que todos os caminhos da árvore serão curtos e a árvore será pouco profunda (RUSSELL & NORVIG, 2022).

Como mencionado anteriormente, o algoritmo seleciona o atributo com importância mais alta, a qual é definida em termos da entropia. Tem-se que a entropia mede a incerteza de uma variável aleatória (RUSSELL & NORVIG, 2022). Em outras palavras, é a quantidade de certeza que existe sobre o valor que uma variável pode assumir. Quanto menos incerteza sobre seus valores, mais importante a variável é.

Portanto, quanto mais certeza houver, menor será a entropia. Por exemplo, uma moeda viciada não tem incerteza, de modo que sua entropia é zero. A entropia de uma variável aleatória *v*, com *k* valores valendo v_k e com probabilidade $P(v_k)$ é definida pela seguinte equação:

$$Entropia = -\sum_{k=1}^{n} P(v_k) log_2 \cdot P(v_k) \qquad (9)$$

Outra medida que avalia o grau de desordem dos dados é o coeficiente *gini*, o qual é dado pela fórmula abaixo:

$$Gini = 1 - \sum_{i=1}^{n} p_i^2 \qquad (10)$$

onde o cálculo do coeficiente é feito para cada i-ésimo nó da árvore, com p_i sendo a probabilidade de cada uma das classe (BEYELER, 2017).

Também vale ressaltar, que para casos em que não há um padrão a ser encontrado, o algoritmo produzirá uma árvore muito grande. Tal problema é denominado superadaptação. A superadapatção é mais provável de ocorrer conforme o espaço de hipótese e a quantidade de atributos de entrada aumente; e menos provável quando a quantidade de exemplos de treinamento é grande (RUSSELL & NORVIG, 2022).

Uma técnica utilizada para combater a superadaptação é a poda de árvore de decisão, a qual consiste na eliminação de nós que não são relevantes. Isto é, para um nó de teste que tem somente nós folha como descendentes, é verificado a relevância do teste. Caso esse seja considerado irrelevante, o teste é eliminado e substituído por um nó folha. Esse processo é repetido para cada

teste que tem apenas descendentes folhas, até que cada um destes seja podado ou considerado relevante (RUSSELL & NORVIG, 2022).

2.6.2 Floresta Randômica

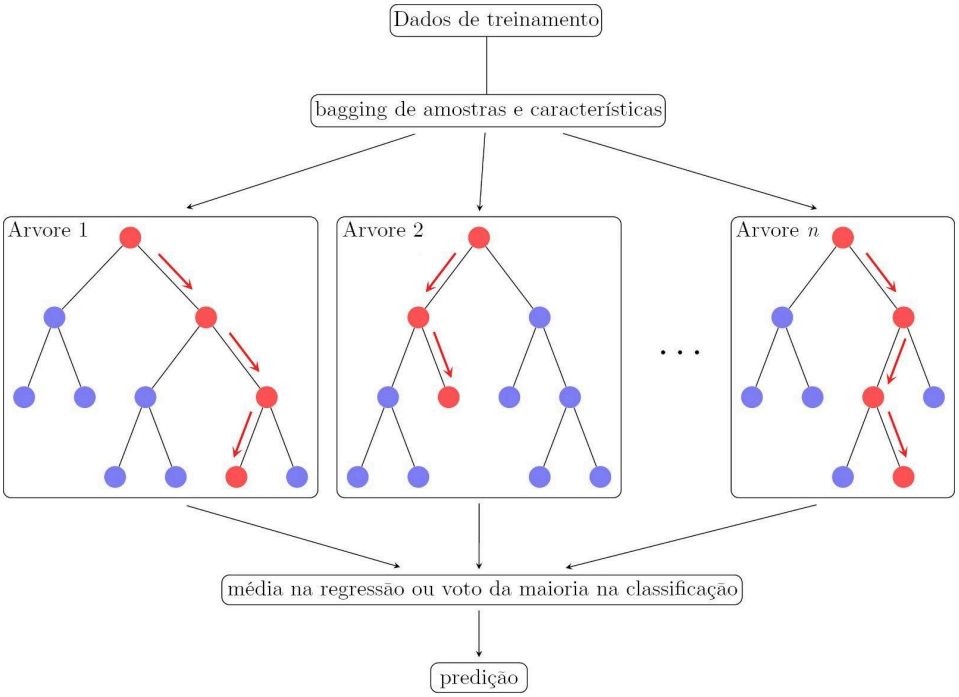
Para compreender melhor como funciona a floresta randômica – do inglês *random forest* – é preciso, primeiro, compreender o conceito de *bagging*. No *bagging* são gerados *K* conjuntos de treino diferentes por amostragem. Isto é, são escolhidos aleatoriamente *N* exemplos de conjuntos de treino, os quais podem ser um exemplo que já foi selecionado anteriormente (RUSSELL & NORVIG, 2022).

Na sequência, um algoritmo de aprendizado de máquina é executado nos *N* exemplos para obter uma hipótese. Esse processo é repetido *K* vezes, de modo que *K* hipóteses são obtidas. Assim, para fazer a previsão de um novo valor de entrada, as previsões de todas as *K* hipóteses são geradas (RUSSELL & NORVIG, 2022).

O *bagging* é utilizado quando os dados são limitados ou quando o modelo-base apresenta *overfitting* – isto é, quando o modelo "decorou" as respostas do conjunto de treinamento ao invés de aprender; obtendo, assim, um desempenho ruim quando confrontado com novos dados do conjunto de teste –, o que reduz a variância (RUSSELL & NORVIG, 2022).

Com isso, tem-se que o algoritmo floresta randômica é uma forma de *bagging* de árvore de decisão, em que são executadas algumas etapas extras para diversificar o grupo de *K* árvores e diminuir a variância (RUSSELL & NORVIG, 2022). Em outras palavras, a floresta randômica é uma coleção de árvores de decisão, onde cada árvore é diferente das outras, assim como demonstra a Figura 6. Cada árvore da floresta é treinada com um conjunto de dados levemente diferente (BEYELER, 2017).

Figura 6: Exemplo gráfico de uma floresta randômica.



Fonte: TikZ.net (2021).

No entanto, o *bagging* de árvores de decisão acaba gerando *K* árvores com correlação muito alta. De modo que um atributo com ganho de informação muito alto provavelmente será a raiz da maioria das árvores. Assim, para reduzir a correlação, o objetivo é variar aleatoriamente as escolhas dos atributos, ao invés dos exemplos de treino (RUSSELL & NORVIG, 2022).

Também pode-se fazer uma seleção aleatória do ponto de divisão, ou seja, para cada atributo são amostrados vários possíveis valores; dentre estes seleciona-se o valor com maior ganho de informação. Esse processo aumenta a chance de que todas as árvores da floresta sejam diferentes (RUSSELL & NORVIG, 2022).

Conforme mais árvores são adicionadas, a taxa de erro do conjunto de validação tende a não aumentar. Assim, tem-se que o modelo floresta randômica é mais resistente ao *overfitting* (BREIMAN, 2001). Vale destacar que este algoritmo pode ser usado tanto para problemas de regressão, quanto para problemas de classificação.

2.6.3 Máquinas de Vetores de Suporte

Nas máquinas de vetores de suporte (SVM), o principal objetivo é encontrar uma linha (ou um plano) que separe os pontos de dados de diferentes categorias da melhor forma possível. Para isso, constroi-se um separador de margem máxima, o qual é um limite de decisão com a maior distância possível dos pontos de exemplos (RUSSELL & NORVIG, 2022). Portanto, quanto maior a margem melhor o modelo tende a funcionar em novos dados, pois ele se torna mais confiável e menos propenso a cometer erros. (BEYELER, 2017).

O SVM também cria um hiperplano de separação linear, mas consegue incorporar dados em um espaço de dimensão superior. Isso porque dados que não são separáveis linearmente no espaço de entrada original são, muitas vezes, facilmente separáveis em um espaço de dimensão superior (RUSSELL & NORVIG, 2022).

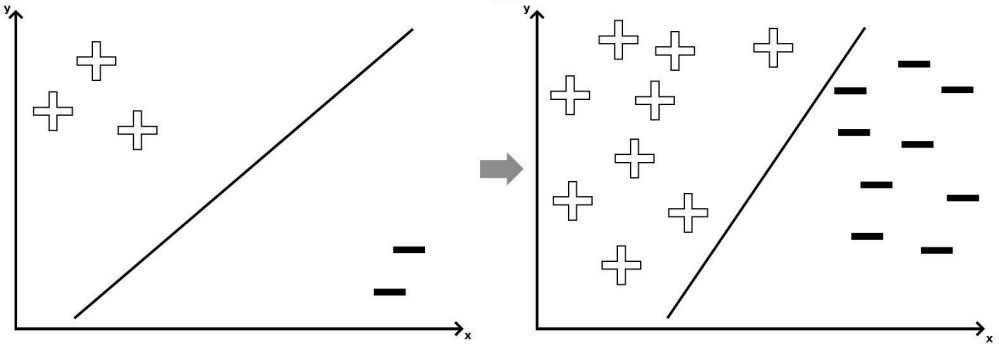
De acordo com Russell e Norvig (2022), o SVM é um método não paramétrico, já que o hiperplano de separação é definido por um conjunto de pontos de exemplo e não por uma coleção de valores de parâmetro. Enquanto alguns modelos precisam manter todos os exemplos, o SVM mantém apenas os exemplos que estão mais próximos do plano de separação.

Dessa forma, tem-se que o SVM possui vantagens tanto de modelos paramétricos como de não-paramétricos, tendo flexibilidade para representar funções complexas e sendo resistente ao *overfitting* (RUSSELL & NORVIG, 2022).

Beyeler (2017) apresenta o seguinte problema de classificação binária: há algumas amostras de treinamento com apenas dois atributos (valores x e y) e um rótulo de destino correspondente (positivo (+) ou negativo (-)). Em problemas como esse, um limite de decisão ideal seria aquele em que todos os dados de uma classe estão de um lado do limite de decisão e todos os dados da outra classe estão do outro lado. Por exemplo, todos os (+) ficariam do lado esquerdo e todos os (-) do lado direito.

A escolha de um limite de decisão ocorre durante todo o processo de treinamento. Isso porque no início do treinamento o modelo recebeu apenas alguns pontos de dados, de modo a posicionar o limite de decisão da melhor maneira que separasse as duas classes. Conforme o treinamento avança, o modelo vai recebendo mais exemplos e, portanto, continua atualizando o limite de decisão em cada etapa (BEYELER, 2017). Tal processo pode ser visto na Figura 7.

Figura 7: Exemplo de mudança do limite de decisão do SVM.

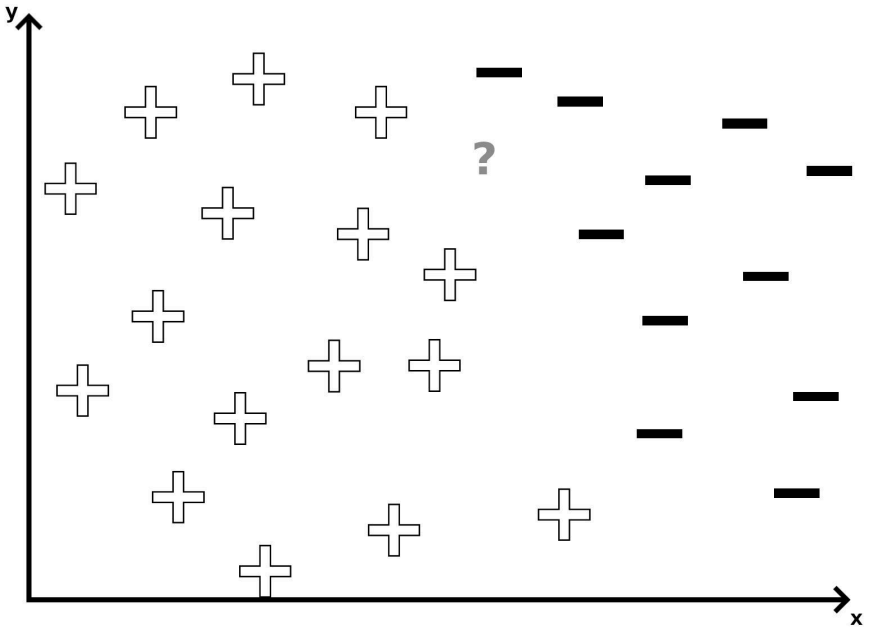


Fonte: Beyeler (2017).

Assim, à medida que o treinamento procede o algoritmo recebe cada vez mais dados e obtém uma ideia melhor de onde o limite de decisão ideal deve estar. Após o treinamento, o modelo de classificação não é mais alterado. De modo que o modelo terá de prever o rótulo-alvo de novos pontos de dados usando o limite de decisão obtido durante o treinamento (BEYELER, 2017).

Ainda no exemplo de Beyeler (2017), o autor questiona qual a classe do símbolo (?) (Figura 8), baseada no limite de decisão que o modelo aprendeu durante o treinamento.

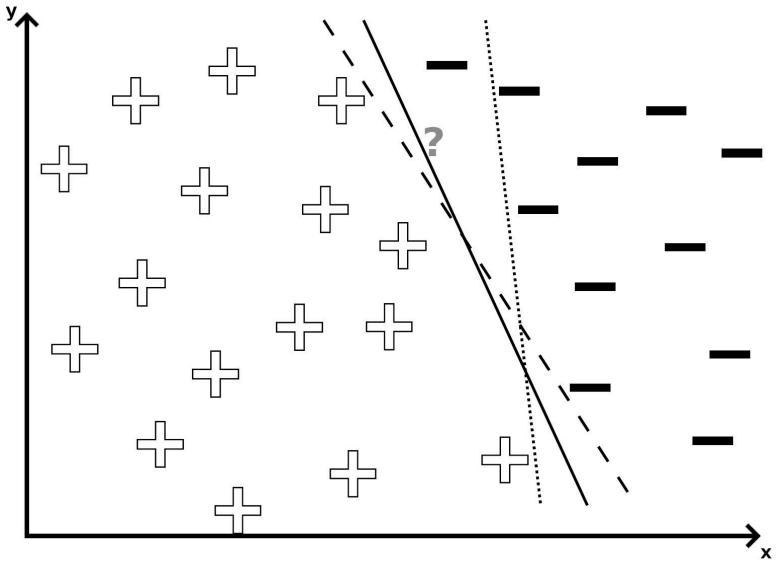
Figura 8: Exemplo de predição do limite de decisão do SVM.



Fonte: Beyeler (2017).

Há várias maneiras de desenhar o limite de decisão, como ilustra a Figura 9.

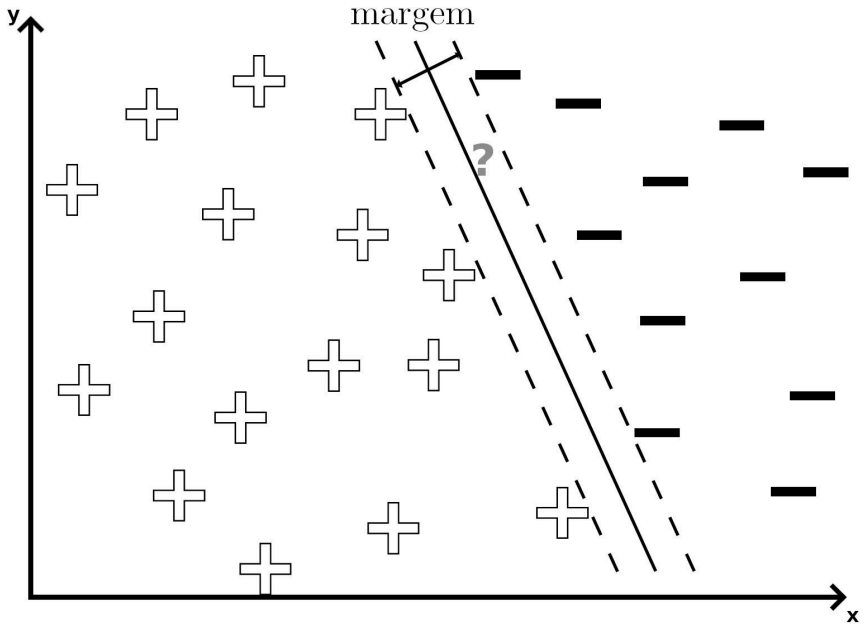
Figura 9: Exemplo de possíveis limites de decisão do SVM.



Fonte: Beyeler (2017).

Entretanto, para este caso tem-se que o SVM escolheria a linha sólida, pois é o limite de decisão que maximiza a margem e os pontos dos dados das duas classes (BEYELER, 2017), assim como pode ser visto na Figura 10.

Figura 10: Exemplo de limite de decisão do SVM.



Fonte: Beyeler (2017).

Por isso que tal limite de decisão é denominado separador de margem máxima. Além disso, tem-se que a margem (área delimitada pelas linhas tracejadas) é duas vezes a distância do separador até o ponto de exemplo mais próximo (RUSSELL & NORVIG, 2022). Assim, para encontrar a margem máxima é necessário apenas considerar os pontos de dados que estão nas margens das classes. Esses pontos são chamados vetores de suporte (BEYELER, 2017).

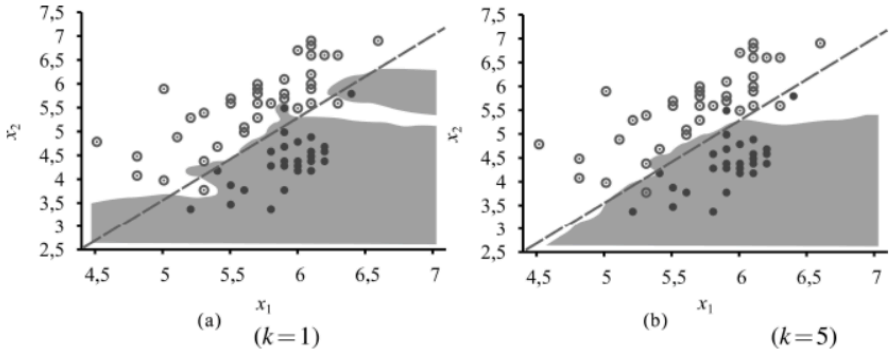
2.6.4 K-Vizinhos mais Próximos

O K-vizinhos mais próximos – ou *K-nearest neighbors* (KNN) – é um dos algoritmos mais simples de aprendizado de máquina. Isso porque para esse modelo é necessário apenas armazenar o conjunto de dados de treinamento. Logo, para fazer uma previsão para um novo ponto de dados, é preciso – somente – encontrar o ponto de dados mais próximo no conjunto de treinamento, ou seja, o seu vizinho mais próximo. Todavia para algumas vizinhanças tal processo é mais complicado. Em casos assim, não será considerado apenas o vizinho mais próximo (k = 1), mas os k vizinhos mais próximos. Portanto, tem-se que k é a quantidade de vizinhos (BEYELER, 2017).

Denota-se VP(k, x_q) o conjunto de k vizinhos mais próximos de uma dada consulta x_q. Em outras palavras, tem-se que: dado um conjunto de N exemplos e uma consulta x_q, todos os exemplos são percorridos e a distância de cada um deles até x_q é calculada, de modo a selecionar os melhores k (RUSSELL & NORVIG, 2022).

Para problemas de classificação deve-se, primeiro, encontrar o conjunto VP(k, x_q) e considerar o valor de saída mais comum. No caso de uma classificação binária, k normalmente é escolhido como número ímpar para evitar empates. Para problemas de regressão, calcula-se a média ou a mediana dos k vizinhos, ou resolve-se um problema de regressão linear sobre os vizinhos. A Figura 11 apresenta um exemplo de KNN.

Figura 11: Exemplo de problema de classificação por k-vizinhos mais próximos para k = 1 e k = 5.



Fonte: Russell e Norvig (2022)

A expressão “mais próximos” remete a uma medida de distância. Para medir a distância a partir de um ponto de consulta x_q até um ponto de exemplo x_j, utiliza-se a distância de Minkowski ou norma L^p, a qual é definida como:

$$L^p(x_j, x_q) = (\sum_i |x_{j,i} - x_{q,i}|^p)^{1/p} \quad (11)$$

Se $p = 2$ tem-se a distância euclidiana, e se $p = 1$ tem-se a distância de Manhattan. Com valores de atributos booleanos, o número de atributos em que dois pontos diferem é chamado distância Hamming. Na maioria dos casos, tem-se que a distância euclidiana é utilizada se as dimensões estiverem medindo propriedades semelhantes, como largura, altura e profundidade. Já a distância de Manhattan é usada se as propriedades forem distintas, tais como idade, peso e gênero de um paciente, por exemplo (RUSSELL & NORVIG, 2022).

Vale ressaltar que mudanças de unidade em qualquer dimensão afetam a distância total. Por exemplo, se a dimensão da propriedade “altura” for alterada de metros para milhas, mas “largura” e “profundidade” permanecerem com as mesmas dimensões, obter-se-á vizinhos mais próximos diferentes. Agora, para comparar propriedades distintas, aplica-se a normalização para medições em cada dimensão (RUSSELL & NORVIG, 2022).

Em espaços de baixa dimensionalidade com bastantes dados, a técnica dos vizinhos mais próximos funciona de maneira eficiente, de modo que há dados suficientes nas proximidades para obter uma boa resposta. Porém, conforme o número de dimensões cresce, os vizinhos mais próximos ficam muito próximos. Tal problema é denominado maldição da dimensionalidade (RUSSELL & NORVIG, 2022).

3 REVISÃO SISTEMÁTICA DA LITERATURA

Este capítulo apresenta a Revisão Sistemática da Literatura, a qual objetiva selecionar os trabalhos correlatos, de modo a identificar as principais técnicas de aprendizado de máquina no contexto do controle de plantas daninhas. Tem-se que revisões sistemáticas possuem requisitos rigorosos para elaborar estratégias de pesquisa e fazer a seleção de artigos que serão incluídos na revisão, além de ser um jeito efetivo de sintetizar o que a coleção de estudos está evidenciando (SNYDER, 2019).

O procedimento da revisão da literatura consiste de algumas etapas: planejamento; execução; análise dos resultados e redação (SNYDER, 2019). Nesta pesquisa, a revisão da literatura foi baseada na metodologia PRISMA (Principais Itens para Relatar Revisões Sistemáticas e Meta-análises), conforme descrito em Liberati *et al.* (2009). Vale ressaltar que a etapa de meta-análise não foi realizada.

Segue o modo como o capítulo está organizado: a Seção 3.1 apresenta as perguntas de análise e os repositórios utilizados na pesquisa; na Seção 3.2 é feita a seleção dos artigos com base nos critérios de inclusão e exclusão; na Seção 3.3 são apresentados os trabalhos correlatos, na Seção 3.4 é feita a análise comparativa entre os estudos selecionados, e a seção 3.5 exibe as considerações da revisão sistemática.

3.1 DEFINIÇÃO DOS CRITÉRIOS DE BUSCA

Para localizar os trabalhos correlatos, às perguntas de pesquisa definidas na Seção 1.1 foram utilizadas para elaborar as perguntas de análise e, com isso, a *string* de busca. As perguntas de análise são:

1. Quais técnicas e algoritmos de aprendizado de máquina são utilizados no controle de plantas daninhas em sistemas ILP?

2. Quais as soluções desenvolvidas com o auxílio de técnicas de aprendizado de máquina para o manejo de plantas daninhas em sistemas ILP?

3. Há alguma disparidade no número de estudos encontrados sobre os modelos de aprendizado de máquina e as soluções desenvolvidas usados para o manejo de plantas daninhas em sistemas ILP?

4. Qual a taxa de redução das plantas daninhas em sistemas ILP que as soluções e/ou modelos desenvolvidos atingem?

Com base nessas perguntas, foram extraídas as palavras-chave para criação de uma *string* de busca para, então, realizar a pesquisa pelos trabalhos relacionados. As palavras-chave foram traduzidas para o inglês, considerando seus possíveis sinônimos. Assim, a expressão de busca definida é:

- [(weed(s) OR weed control OR weed management) AND (machine learning OR artificial intelligence) AND (ICLS OR ICL system)]

Observa-se que a tradução de sistema ILP ficou como *ICL system* ou *ICLS*, isso porque Sistema Integração Lavoura-Pecuária em inglês é *Integrated Crop-Livestock System*, de modo que a sigla fica *ICL system* ou *ICLS*.

Também vale ressaltar que a pesquisa teve que ser feita em duas partes, já que a composição dos três termos não teve retorno nas bases. Logo, a expressão de busca ampla foi dividida em duas partes: [(*weed(s)* OR *weed control* OR *weed management*) AND (*machine learning* OR *artificial intelligence*)] e [(*weed(s)* OR *weed control* OR *weed management*) AND (*ICL system* OR *ICLS*)].

No Quadro 2 pode-se verificar os repositórios escolhidos e as *strings* de busca adaptadas para cada biblioteca.

Quadro 2: Bases de dados, expressões de busca adaptadas e endereços de acesso.

Repositório	***String* de busca**
IEEE Xplore	[TITLE-ABS-KEY(weed(s)) OR TITLE-ABS-KEY(weed control) OR TITLE-ABS-KEY(weed management) AND TITLE-ABS-KEY(machine learning) OR TITLE-ABS-KEY(artificial intelligence)] OR [TITLE-ABS-KEY(weed(s) OR TITLE-ABS-KEY(weed control) OR TITLE-ABS-KEY(weed management) AND TITLE-ABS-KEY(ICLS) OR TITLE-ABS-KEY(ICL system))]
Mendeley	[(“weed(s)” OR “weed control” OR “weed management”) AND (“machine learning” OR “artificial intelligence”)] OR [(“weed(s)” OR “weed control” OR “weed management”) AND (“ICLS” OR “ICL system”)]
Science Direct	[(“weed(s)” OR “weed control” OR “weed management”) AND (“machine learning” OR “artificial intelligence”)] OR [(“weed(s)” OR “weed control” OR “weed management”) AND (“ICLS” OR “ICL system”)]
Scopus	[((“All Metadata” : “weed(s)”) OR (“All Metadata” : “weed control”) OR (“All Metadata” : “weed management”)) AND ((“All Metadata” :“machine learning”) OR (“All Metadata” :“artificial intelligence”))] OR [((“All Metadata” :“weed(s)”) OR (“All Metadata” :“weed control”) OR (“All Metadata” :“weed management”)) AND ((“All Metadata” :“ICLS”) OR (“All Metadata” : “ICL system”))]

Fonte: Elaborada pela autora.

3.2 SELEÇÃO DOS TRABALHOS RELACIONADOS

Embora o objetivo seja incluir a maior quantidade possível de trabalhos relacionados com as perguntas de análise, é necessário definir critérios de inclusão e exclusão para verificar quais estudos podem contribuir de modo efetivo para a pesquisa. O Quadro 3 apresenta os critérios de inclusão e exclusão.

Quadro 3: Critérios de Inclusão e Exclusão.

Inclusão	Exclusão
CI1: Artigos em inglês ou português.	CE1: Artigos duplicados.
CI2: Artigos acessíveis via Portal CAPES.	CE2: Artigos que não permitam acesso ao seu texto completo
CI3: Artigos publicados entre 01/01/2013 até 31/03/2023.	CE3: Artigos que não tenham “weed management”, “weed control” ou “weed(s)” como palavras-chave.

Fonte: Elaborada pela autora.

Além dos critérios de inclusão e exclusão, foram também analisados o título, o resumo e as palavras-chave de cada artigo para verificar se estes estavam de acordo com os objetivos do estudo. Tal critério de qualidade foi aplicado, já que as pesquisas retornaram muitos artigos que traziam aplicações de aprendizado de máquina na agricultura, mas que não envolviam o manejo de plantas daninhas.

O processo de revisão da literatura foi dividido em três etapas. A primeira consistia em identificar e remover as duplicatas dos estudos (CE1), mantendo apenas uma versão para análise. Na segunda etapa foram aplicados os critérios de inclusão e exclusão CI1, CI3 e CE3, além do critério de qualidade. Por último, na terceira etapa foram empregados os critérios CI2 e CE2, isto é, foi verificado quais trabalhos tinham seu texto completo disponível nos próprios repositórios ou no Portal da CAPES. Este procedimento foi feito com o auxílio da ferramenta Parsifal. O Quadro 4 apresenta a quantidade de artigos encontrados e selecionados em cada base nas respectivas etapas.

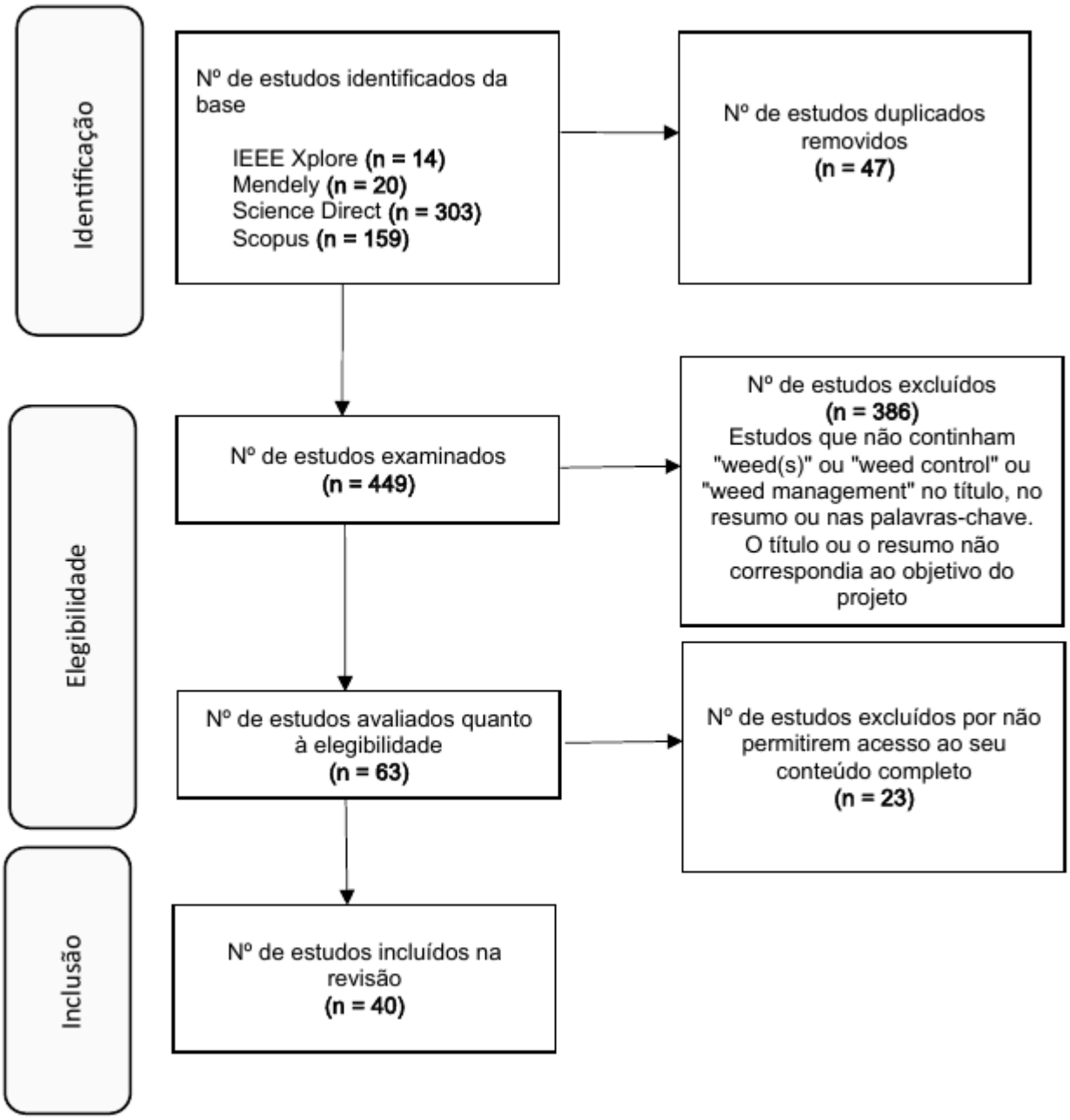
Quadro 4: Quantidade de estudos encontrados e selecionados por repositório.

Base de Dados	Encontrados	1ª Etapa	2ª Etapa	3ª Etapa
IEEE Xplore	14	12	7	7
Mendeley	20	13	4	4
Science Direct	303	276	17	15
Scopus	159	148	35	14
Total	**496**	**449**	**63**	**40**

Fonte: Elaborada pela autora.

No total havia 47 duplicatas em todas as bases, sendo que a maioria delas era da Science Direct. Já na segunda etapa, a maioria dos resultados foram removidos por não terem “weed(s)”, “weed control” ou “weed management” no título, no resumo ou nas palavras-chave. Também tem-se que para as bases IEEE Xplore e Mendeley todos os artigos tinham seu texto na íntegra, enquanto os trabalhos do repositório Scopus foram o que menos tinham seu texto completo – tanto na base de pesquisa quanto no Portal da CAPES. Portanto, ao final de todo o processo foram selecionados 40 trabalhos correlatos. A Figura 12 demonstra o processo completo da seleção dos artigos.

Figura 12: Fluxograma de identificação e seleção dos estudos da revisão sistemática da literatura.



Fonte: Elaborada pela autora.

A seguir, o Quadro 5 apresenta todos os artigos selecionados, com sua identificação, título e ano de publicação. Os estudos estão ordenados por ordem cronológica, do mais antigo para o mais recente.

Quadro 5: Trabalhos Relacionados.

Identificação	Título	Ano
Torres-Sospedra e Nebot (2014)	Two-stage procedure based on smoothed ensembles of neural networks applied to weed detection in orange groves	2014
Pérez-Ortiz *et al*. (2016)	Selecting patterns and features for between- and within- crop-row weed mapping using UAV-imagery	2016
Schuster *et al*. (2016)	Grazing intensities affect weed seedling emergence and the seed bank in an integrated crop–livestock system	2016
Lutosa *et al*. (2016)	Floristic and phytosociology of weed in response to winter pasture sward height at Integrated Crop-Livestock in Southern Brazil	2016
Chavan e Nandedkar (2018)	AgroAVNET for crops and weeds classification: A step forward in automatic farming	2018
Sabzi e Abbaspour-Gilandeh (2018)	Using video processing to classify potato plant and three types of weed using hybrid of artificial neural network and particle swarm algorithm	2018
Sandino e Gonzalez (2018)	A Novel Approach for Invasive Weeds and Vegetation Surveys using UAS and Artificial Intelligence	2018
Zhang *et al*. (2018)	Broad-Leaf Weed Detection in Pasture	2018
Abouzahir *et al*. (2018)	Enhanced Approach for Weeds Species Detection Using Machine Vision	2018
Gao *et al*. (2018)	Recognising weeds in a maize crop using a random forest machine-learning algorithm and near-infrared snapshot mosaic hyperspectral imagery	2018
Yu *et al*. (2019)	Weed Detection in Perennial Ryegrass With Deep Learning	2019

	Convolutional Neural Network	
Partel *et al*. (2019)	Development and evaluation of a low-cost and smart technology for precision weed management utilizing artificial intelligence	2019
Sudars *et al*. (2020)	Dataset of annotated food crops and weed images for robotic computer vision control	2020
Qiao *et al*. (2020)	MmNet: Identifying Mikania micrantha Kunth in the wild via a deep Convolutional Neural Network	2020
Souza *et al*. (2020)	Spectral differentiation of sugarcane from weeds	2020
Yan *et al*. (2020)	Classification of weed species in the paddy field with DCNN-Learned features	2020
Wang *et al*. (2020)	Semantic Segmentation of Crop and Weed using an Encoder-Decoder Network and Image Enhancement Method under Uncontrolled Outdoor Illumination	2020
Yu *et al*. (2020)	Detection of grassy weeds in bermudagrass with deep convolutional neural networks	2020
Sabzi *et al*. (2020)	An automatic visible-range video weed detection, segmentation and classification prototype in potato field	2020
Hussain *et al*. (2021)	Application of deep learning to detect Lamb's quarters (Chenopodium álbum L.) in potato fields of Atlantic Canada	2021
Fawakherji *et al*. (2021)	Multi-Spectral Image Synthesis for Crop/Weed Segmentation in Precision Farming	2021
Siddiqui *et a*l. (2021)	Neural Network based Smart Weed Detection System	2021
Monteiro *et al*. (2021)	A new alternative to determine weed control in agricultural systems based on artificial neural networks (ANNs)	2021
Etienne *et al*. (2021)	Deep Learning-Based Object Detection System for Identifying Weeds Using UAS Imagery	2021
Shorewala *et al*. (2021)	Weed Density and Distribution Estimation for Precision Agriculture Using Semi-Supervised Learning	2021

Subeesh *et al*. (2022)	Deep convolutional neural network models for weed detection in polyhouse grown bell peppers	2022
Nasiri *et al*. (2022)	Deep learning-based precision agriculture through weed recognition in sugar beet fields	2022
Alrowais *et al*. (2022)	Hybrid leader based optimization with deep learning driven weed detection on internet of things enabled smart agriculture environment	2022
Razfar *et al*. (2022)	Weed detection in soybean crops using custom lightweight deep learning models	2022
Costello *et al*. (2022)	Detection of Parthenium Weed (Parthenium hysterophorus L.) and Its Growth Stages Using Artificial Intelligence	2022
Dominschek *et al*. (2022)	Diversification of traditional paddy field impacts target species in weed seedbank	2022
Ni *et al*. (2022)	A deep convolutional neural network-based method for identifying weed seedlings in maize fields	2022
Ngo *et al*. (2022)	Automated Weed Detection System for Bok Choy Using Computer Vision	2022
Jose *et al*. (2022)	Classification of Weeds and Crops using Transfer Learning	2022
Wang e Leelapatra (2022)	Weeding Robot Based on Lightweight Platform and Dual Cameras	2022
Firmansyah *et al*. (2022)	Real-time Weed Identification Using Machine Learning and Image Processing in Oil Palm Plantations	2022
Meena *et al*. (2023)	Crop Yield Improvement with Weeds, Pest and Disease Detection	2023
Ajayi e Ashi (2023)	Effect of varying training epochs of a Faster Region-Based Convolutional Neural Network on the Accuracy of an Automatic Weed Classification Scheme	2023
Dang *et al*. (2023)	YOLOWeeds: A novel benchmark of YOLO object detectors for multi-class weed detection in cotton production systems	2023

Raja *et al*. (2023)	Real-time control of high-resolution micro-jet sprayer integrated with machine vision for precision weed control	2023

Fonte: Elaborada pela autora.

3.3 TRABALHOS RELACIONADOS

Nesta seção será apresentada uma breve descrição dos trabalhos relacionados elegidos, ressaltando as técnicas de aprendizado de máquina utilizadas, tipos de culturas nas quais foram realizados os estudos, possíveis soluções desenvolvidas, entre outros.

3.3.1 Two-stage procedure based on smoothed ensembles of neural networks applied to weed detection in orange groves

Torres-Sospedra e Nebot (2014) apresentam um sistema de detecção de plantas daninhas, baseado num procedimento de duas etapas. Primeiro, os principais elementos da plantação são identificados na imagem: árvores, troncos, céu e solo. Depois, para os casos classificados como solo, faz-se a detecção da presença ou ausência de plantas daninhas. A pesquisa foi realizada numa plantação de laranjas na Espanha, e no total foram coletadas 140 imagens.

Foi utilizado no desenvolvimento do sistema a rede neural (*Artificial Neural Networks* - ANN) *Multilayer Perceptron* (MLP). Também foi usado o algoritmo *Backpropagation Neural Network* (BPNN) para refinar o modelo para alguns casos. Para outros cenários foi aplicado um redutor de ruído espacial para lidar com as variações de luminosidade. Ambos os métodos foram comparados, e percebeu-se que a aplicação do redutor de ruído teve resultados melhores. Porém, a performance da detecção de plantas daninhas foi menor do que a classificação do solo.

3.3.2 Selecting patterns and features for between- and within- crop-row weed mapping using UAV-imagery

Os autores abordam o problema de mapeamento de plantas daninhas usando imagens fornecidas por um drone em plantações de milho e girassol na Espanha. Utilizaram o algoritmo SVM e também a técnica de análise de imagens baseada em objetos (*Object-based image analysis* - OBIA).

A utilização do OBIA foi boa para reduzir o efeito "sal e pimenta" para as imagens das duas culturas. Quanto ao algoritmo, este teve resultados satisfatórios para a cultura de girassol, detectando os *pixels* de plantas daninhas de forma precisa mesmo quando perto das linhas da plantação. Entretanto, o modelo apresentou alguns problemas de classificação para o milho, devido às sombras das plantas. De modo geral, tem-se que o algoritmo conseguiu detectar plantas daninhas não só entre as fileiras das culturas, mas também nas linhas cultivadas.

3.3.3 Grazing intensities affect weed seedling emergence and the seed bank in an integrated crop–livestock system

Schuster *et al.* investigam os efeitos de diferentes intensidades de pastoreio na emergência de plantas daninhas e de seu banco de sementes em sistemas ILP. Os autores levantam a hipótese de que reduzir a intensidade do pastoreio ajuda a reduzir a interferência de espécies de plantas daninhas e seus bancos de sementes. Também verificam se a composição das plantas daninhas, sua emergência e seu banco de sementes, muda com a intensidade do pastoreio.

Para isso, foram coletadas e analisadas sementes das plantas daninhas encontradas no pasto e na plantação de soja do sistema ILP, localizado no sul do Brasil. Os autores observaram que diminuir a intensidade de pastoreio ajuda a reduzir a quantidade de plantas daninhas, a densidade das sementes do banco e, também, a densidade das plantas emergidas. Todavia, se a intensidade de pastoreio é aumentada, o banco de sementes das espécies de plantas daninhas também aumenta.

3.3.4 Floristic and phytosociology of weed in response to winter pasture sward height at Integrated Crop-Livestock in Southern Brazil

Assim como o trabalho anterior, Lutosa *et al.* (2016) averiguam aspectos sobre os sistemas ILP: os autores descrevem a diversidade e a comunidade de plantas daninhas devido a modificações na altura do pasto em sistemas de rotação de culturas. Em outras palavras, é verificada a composição florística, a fitossociologia e a diversidade da comunidade de plantas daninhas em pastagens com diferentes alturas. A pesquisa foi conduzida em plantações de soja, milho, trigo, aveia preta e tipos de pastagens; localizadas em um sistema ILP no sul do Brasil.

Foram coletados e identificados 813 indivíduos de 55 espécies diferentes, pertencentes a 18 famílias. A densidade de cada planta também foi estimada. Os resultados mostram que a uniformidade das espécies com o manejo da altura do pasto teve ligação direta com a estrutura da comunidade de plantas daninhas. Assim, a densidade das plantas daninhas e o complexo de comunidades podem ser reduzidos pelo aumento da altura do pasto.

3.3.5 AgroAVNET for crops and weeds classification: A step forward in automatic farming

Chavan e Nandedkar (2018) tem como objetivo classificar culturas e plantas daninhas em estágio inicial de crescimento, de modo que ações de controle possam ser tomadas e o rendimento das colheitas possa ser melhorado. Os autores usam um *dataset* de diferentes culturas – trigo, milho, tipos de grama e beterraba – e espécies de plantas daninhas para criar um sistema mais genérico de classificação destes dois grupos. Utilizando *Convolutional Neural Networks* (CNN) propõem o AgroAVNET, o qual consiste de uma combinação das arquiteturas AlexNet e VGGNET. Assim, o sistema usa o conceito de normalização do AlexNet, enquanto a profundidade dos filtros é escolhida com base no VGGNET.

Como resultado, tem-se que o AgroAVNET obteve uma acurácia de 90%; a qual foi maior na classificação de diferentes tipos de plantas do que AlexNet, VGGNET e também outros sistemas existentes na literatura. Apesar de não ter sido desenvolvido, os pesquisadores afirmam que o sistema pode ser aplicado em um robô, que realizaria a aplicação de herbicidas nas plantas daninhas identificadas.

3.3.6 Using video processing to classify potato plant and three types of weed using hybrid of artificial neural network and particle swarm algorithm

Com o objetivo de localizar e identificar três espécies de plantas daninhas em plantações de batata, Sabzi e Abbaspour-Gilandeh (2018) propõem um sistema de *Machine Vision* (MV) que compreende outros dois subsistemas: um de processamento de vídeo capaz de detectar plantas verdes em cada frame; e o outro para classificar as plantas como planta daninhas ou vegetais (batata), utilizando uma abordagem híbrida de ANN e *Particle Swarm Optimization* (PSO) para fazer a classificação. Vale ressaltar que no primeiro subsistema é feito dois tipos de segmentações, para separar o fundo das plantas verdes e depois para identificar os objetos.

Com as plantas completamente separadas do solo e das outras partes da imagem, é feita a extração de características. Foram extraídas 30 características de cada *frame*, as quais foram divididas em 5 grupos: cor, textura baseada em histograma, momento de invariantes, textura baseada em matriz de co-ocorrência de nível de cinza (referente a posição de *pixels* com luminosidade parecida) e características de formato (área, perímetro, comprimento, largura e compressão.)

As seis características mais importantes que diferenciam as plantas daninhas das plantas de batatas foram selecionadas pelo algoritmo Árvore de Decisão (AD). Quanto ao modelo híbrido (ANN-PSO), este teve acurácia acima de 98% na classificação das três espécies de plantas daninhas e da cultura batata. Apesar de não desenvolverem, os autores afirmam que o sistema pode ser usado em aplicadores inteligentes.

3.3.7 A Novel Approach for Invasive Weeds and Vegetation Surveys using UAS and Artificial Intelligence

Sandino e Gonzales (2018) apresentam um *framework* para detectar e mapear plantas daninhas usando um drone, com uma câmera RGB para coletar as imagens, e técnicas de ML para o processamento dos dados. O experimento foi realizado na Austrália, sendo que as imagens utilizadas para detecção de plantas daninhas continham tipos de grama, vegetação local, solo e arbustos. O *framework* compreende as fases de aquisição, pré-processamento, treinamento e predição dos dados.

Primeiro, as imagens são adquiridas por meio do drone e da câmera RGB. Em seguida, as fotos são extraídas e pré-processadas para obter amostras com as características essenciais. Essas imagens são rotuladas e processadas num classificador supervisionado de AM, que é ajustado e otimizado posteriormente. Por último, a imagem orto-mosaica é processada para mapear os locais previstos de plantas daninhas.

O modelo teve problemas de classificação para os arbustos, que em vários casos foram considerados como ruído de imagem. Contudo, a média da métrica *F1 Score* foi de 96,5% para classificação das plantas daninhas.

3.3.8 Broad-Leaf Weed Detection in Pasture

Zhang *et al.* (2018) desenvolvem um método para reconhecer plantas daninhas de folha larga em pastos, de forma que o controle de precisão de plantas daninhas possa ser alcançado e o uso de herbicidas reduzido. Durante o processo, foram comparadas técnicas tradicionais de AM e técnicas de *Deep Learning* (DL), para testar a acurácia dos modelos em um ambiente real. Os autores citam que uma dificuldade foi que o pasto e as plantas daninhas tem cores bem semelhantes. Assim, a textura foi usada como informação para diferenciar as plantas daninhas do pasto.

Os dados foram rotulados em três categorias: “grama”, “plantas daninhas” e “incerto”. Aqueles considerados como “incertos” foram descartados por não serem confiáveis. Além disso, como as plantas daninhas são muito pequenas nas fases iniciais de crescimento, e não há como detectá-las nesse tamanho, foi decidido que a detecção seria apenas para plantas daninhas que cobriam mais de 5% do total da imagem. O objetivo para cada imagem era verificar se havia ou não uma planta daninha na foto.

Foram capturadas e rotuladas 6087 imagens, das quais 4080 eram do pasto e 2007 eram de plantas daninhas. Os algoritmos convencionais de ML utilizados foram SVM, KNN, Árvore Complexa (AC) e Regressão Logística (RL). Já como método baseado em DL foi usado uma CNN. Dos modelos tradicionais o SVM foi o que teve melhor resultado com 89,4% de acurácia. No entanto, a CNN teve um melhor desempenho, com 96,8% de acurácia. Também vale destacar que áreas de muita sobra causaram falsas detecções para todos os modelos.

3.3.9 Enhanced Approach for Weeds Species Detection Using Machine Vision

Baseado na literatura, Abouzahir *et al.* (2018) observaram que uma limitação da identificação de plantas daninhas é o tempo computacional. Conforme o número de características aumenta, o algoritmo fica mais complexo e o uso do computador fica mais intenso. De modo que é difícil incorporar um modelo com custo computacional baixo para tarefas em tempo real, apesar dos algoritmos terem uma acurácia alta. Além disso, características de textura e formatos são limitadas devido a oclusão das plantas e sobreposição, que é o caso para aplicações reais. Outro aspecto que dificulta a determinação das espécies de culturas e plantas daninhas é devido a similaridade de na intensidade de cores.

Assim, os autores exploram o uso de índices de cores de vegetação para determinar pontos de plantas daninhas para passar um *spray* automático. Utilizando técnicas e algoritmos como MV, BPNN e SVM, são geradas imagens monocromáticas baseadas nos índices de cor. Essas imagens são divididas em intervalos para obter a escala de cinza baseada nos índices. Depois os histogramas das imagens são gerados e normalizados com a representação final das amostras. O modelo é

treinado com esses histogramas para diferenciar plantas daninhas de folha larga, soja, solo e resíduos. Com isso, o sistema gerado é usado para prever exemplos rotulados para examinar a generalização.

Algumas considerações a serem feitas, é que os índices de cor podem ser sensíveis à mudanças na iluminação externa, o que afeta a performance dos algoritmos. Além disso, o estudo foi realizado por grupo marroquino e o *dataset* de imagens utilizado é brasileiro.

3.3.10 Recognising weeds in a maize crop using a random forest machine-learning algorithm and near-infrared snapshot mosaic hyperspectral imagery

Gao *et al.* (2018) analisam o uso de uma câmera hiperespectral para classificar milho e três tipos de plantas daninhas. O estudo foi conduzido na Bélgica e foram usados os algoritmos KNN e Floresta Randômica (FR). Foram obtidas das imagens 185 características, das quais 30 foram consideradas como as mais importantes pelo RF, baseado no máximo de acurácia acumulada.

Tem-se que o FR com as 30 características mais importantes (usadas na construção do modelo) teve um desempenho melhor do que o KNN. A média de acurácia do FR foi de 80%. No geral, os resultados mostram que é possível reconhecer os três tipos de plantas daninhas e milho usando imagens hiperespectrais. Entretanto, os autores alertam que há algumas limitações e considerações que devem ser levadas em conta para transferir o experimento do laboratório para as condições do mundo real.

3.3.11 Weed Detection in Perennial Ryegrass With Deep Learning Convolutional Neural Network

Yu *et al.* (2019) verificam a viabilidade de usar *Deep Convolutional Neural Networks* (DCNN) para detectar plantas daninhas de folha larga em *perennial reygrass*. Foram utilizadas

quatro arquiteturas de DCNN: AlexNet, DetectNet,VGGNet e GoogleNet. Também vale ressaltar que as imagens empregadas são de países diferentes (Estados Unidos e Canadá), e que as plantas daninhas presentes nos *datasets* de treinamento de teste estavam em diferentes fases de crescimento – o que aumenta a complexidade dos algoritmos.

No que se refere aos resultados, a arquitetura GoogleNet não foi efetiva para detectar as plantas planta daninhas selecionadas, devido a valores muito baixos de Precisão. Exceto para uma espécie específica, as DCNNs produziram resultados consistentes para detecção de plantas daninhas, mesmo em regiões geográficas diferentes. A exceção do DetectNet, que foi eficiente para identificação dessa espécie.

Ainda sobre a planta daninha em particular, tem-se que AlexNet e VGGNet tiveram excelentes valores de Precisão e Sensibilidade para o primeiro *dataset* de treino, mas para o segundo *dataset* os valores de Sensibilidade baixaram. Sugerindo que a rede tende a classificar as plantas daninhas erroneamente como grama. Com relação a detecção simultânea de várias espécies de plantas daninhas, apenas o VGGNet foi efetivo nesta tarefa, mesmo quando as plantas daninhas tinham estruturas morfológicas diferentes.

3.3.12 Development and evaluation of a low-cost and smart technology for precision weed management utilizing artificial intelligence

Neste trabalho, os autores desenvolveram um aplicador inteligente para diferenciar plantas daninhas de outros objetos, os quais são as culturas, e aplicar o herbicida precisamente nos alvos pretendidos. Houve dois cenários de experimentos para simular as condições de campo: o primeiro continha ambas plantas e plantas daninhas artificiais; já o segundo tinha plantas daninhas e culturas de verdade, a qual eram plantas de pimenta. Para cada um dos cenários foram feitos testes que simulam as condições de campo, onde as espécies plantas daninhas e as culturas estão dispostas aleatoriamente.

Um GPS foi conectado ao aplicador e um sistema foi desenvolvido usando DL, MV e CNN, para gerar e visualizar os dados coletados. Ademais, foram utilizadas e analisadas dois tipos de GPUs – GTX 1070 e Jetson TX2. No experimento com as plantas artificiais, havia 30 plantas daninhas, as quais eram os alvos. Para esse caso, a GTX 1070 teve um melhor desempenho do que a TX2, em específico quando comparando a porcentagem de alvos não atingidos e parcialmente atingidos. Para esse cenário, não houve nenhum caso em que uma cultura foi tomada como planta daninha.

Assim, como na experiência anterior, a GTX 1070 teve uma performance superior no cenário com as plantas reais. Já a TX2 perdeu metade dos alvos, de modo que essa GPU pode não ser forte o suficiente para processar as imagens das culturas e das 20 plantas daninhas alvo. No entanto, houve muitos casos em que o aplicador foi aplicado erroneamente na cultura, para ambas as GPUs. Em muitos desses casos em que o aplicador não atingiu o alvo, a planta daninha estava logo ao lado.

3.3.13 Dataset of annotated food crops and weed images for robotic computer vision control

Sudars *et al.* (2020) produzem um *dataset* de imagens de plantas daninhas e de diferentes culturas – beterraba, cenoura, abóbora e rabanete – manualmente anotado e acesso aberto. Parte das imagens foram tiradas em um ambiente controlado e a outra foi tirada em condições de campo, com as plantas daninhas e as culturas em diferentes estágios de crescimento.

As fotos foram tiradas uma vez ao dia, e ao total o *dataset* conta com 1118 imagens de 8 espécies diferentes de plantas daninhas e 7853 anotações, as quais foram feitas por especialistas da área.

3.3.14 MmNet: Identifying Mikania micrantha Kunth in the wild via a deep Convolutional Neural Network

Utilizando um drone, Quiao *et al.* (2020) obtém imagens de uma planta daninha específica – *Mikania micrantha* – e propõem uma CNN chamada MnNet para identificar essa espécie de planta daninha. Foram usadas 400 amostras para validar o modelo, sendo que destas 378 foram reconhecidas corretamente.

Diversos fatores ambientais influenciam na identificação desta planta daninha, tais como luz, forma do relevo e escala do objeto na imagem. Para lidar com a interferência desses elementos e resistir à semelhança visual entre essa espécie e as plantas de fundo, foi desenvolvida uma CNN baseada em DL para identificar com precisão os alvos de *Mikania micrantha.* Tal CNN emprega elementos das arquiteturas GoogleLeNet, AlexNet e VGG-VD16. Comparada com as arquiteturas usadas, MnNet tem vantagens: não apenas sua acurácia é maior (94,5%), mas o tempo de processamento é menor, e sua construção é mais simples.

3.3.15 Spectral differentiation of sugarcane from weeds

Souza *et al*. (2020) apresentam a possibilidade de diferenciar cana-de-açúcar de plantas daninhas pelo comportamento espectral das folhas. Foram consideradas na pesquisa 11 espécies de plantas daninhas, divididas em dois grupos: monocotiledôneas (capins/folha estreita) e dicotiledôneas (folha larga). Essas duas categorias foram escolhidas para testar a sensibilidade das técnicas em diferenciar a cana-de-açúcar de espécies de plantas daninhas similares (capins) ou bem distintas (folhas largas).

De algoritmos foram selecionados os modelos FR e *Soft Independent Modelling by Class Analogy* (SIMCA), ambos também foram comparados em termos de performance e capacidade de discriminação. Os dados foram distribuídos na proporção 70% para treinamento e 30% para validação. Foi demonstrado que a seleção adequada de bandas e a calibração local usando uma

abordagem de classificação espectral podem permitir o mapeamento de plantas daninhas e facilitar a aplicação local do herbicida.

Para todas as bandas de classificação o FR foi levemente melhor. Com exceção de quando foram usadas apenas 4 variáveis (bandas), neste cenário o RF teve um desempenho pior. Apesar do RF ter sido um pouco melhor no geral, o SIMCA ainda pode ser considerado uma técnica eficiente. Assim, os autores consideram que o SIMCA e o FR podem ser usados de modo complementar. Portanto, o estudo mostra que o desenvolvimento de câmeras que funcionem com apenas algumas bandas espectrais, podem ser usadas para distinguir cana-de-açúcar de plantas daninhas, desde que uma seleção apropriada das bandas espectrais seja feita.

3.3.16 Classification of weed species in the paddy field with DCNN-Learned features

Nesta pesquisa, Yan *et al.* (2020) investigam um método de identificação de plantas daninhas que aprende as características das imagens originais usando DCNN. Os autores afirmam que as características aprendidas pela DCNN (AlexNet) podem evitar o complicado e demorado processo de desenvolvimento e ajuste necessários para obter características que seriam extraídas manualmente.

Assim, foram comparadas as características extraídas de seis espécies de plantas daninhas pela DCNN, e as características marcadas manualmente utilizando os algoritmos SVM e KNN. Também foram aferidos os resultados de classificação das imagens de plantas daninhas com fundo complexo e com fundo removido, para verificar a sensibilidade das características aprendidas pela DCNN com relação a mudanças de fundo e iluminação. No total foram usadas 923 imagens, com uma proporção de 70% dos dados para treino e 30% para teste.

A acurácia de classificação das características manuais no SVM foi de 87,8% e das aprendidas pela DCNN foi de 94,5%. Demonstrando que a DCNN teve um desempenho melhor na identificação de plantas daninhas com fundo complexo e alterações de iluminação. Afinal, foi

concluído pelos pesquisadores que não houve mudanças significativas com relação aos fundos. O que demonstra que a segmentação do fundo de uma imagem de planta daninha tem pequeno efeito na acurácia de classificação, pois o modelo DCNN consegue aprender automaticamente as características efetivas da imagem.

3.3.17 Semantic Segmentation of Crop and Weed using an Encoder-Decoder Network and Image Enhancement Method under Uncontrolled Outdoor Illumination

Neste trabalho, um *encoder-decoder* de DL foi investigado para segmentação semântica por *pixels* de culturas e plantas daninhas. Diferentes representações de entradas, incluindo diferentes transformações do espaço de cores e dos índices de cores foram comparados para otimizar a entrada da rede. Também foram analisados três métodos de aprimoramento de imagens – que melhoram o brilho e o contraste das fotos –, para aperfeiçoar o modelo diante de diferentes condições de iluminação.

Foram utilizados dois *datasets*: um de fotos de beterraba e o outro de sementes oleaginosas. No geral, para o conjunto de plantas de beterraba o aprimoramento das imagens melhorou a qualidade das fotos e os modelos de segmentação, em termos de diferentes condições de iluminação. Para o das sementes oleaginosas, o aprimoramento não diminuiu nem aumentou a performance de segmentação.

3.3.18 Detection of grassy weeds in bermudagrass with deep convolutional neural networks

Yu *et al.* (2020) exploram a viabilidade de usar classificação de imagens com DCNNs – incluindo AlexNet, GoogleLeNet e VGGNet – para detectar certas espécies de plantas daninhas (capins) em sistemas de grama. Os modelos de DCNN foram treinados usando um conjunto de

dados de uma única espécie de planta daninha ou de várias espécies. A rede neural de várias espécies foi testada para ver como seria usar apenas uma DCNN para identificar vários tipos de plantas daninhas. Enquanto as imagens de uma única espécie foram utilizadas para treinar a rede neural de vários tipos de plantas daninhas.

Os autores observaram que a densidade de plantas daninhas teve influência na detecção das plantas daninhas pelos modelos, em particular, pelo AlexNet e GoogleLeNet. Já o VGGNet teve excelente performance em detectar as espécies selecionadas para ambos os casos de baixa e alta densidade. No geral, VGGNet obteve um desempenho melhor do que as outras duas arquiteturas para ambos os casos estudados: para imagens de apenas uma espécies teve um resultado ótimo; e para imagens com vários tipos de plantas daninhas também teve bons resultados para algumas plantas daninhas.

No geral, os resultados apresentados são promissores, pois demonstram a viabilidade de usar DCNNs para fazer a detecção de plantas daninhas em gramas usando um subsistema de MV para um *spray* inteligente.

3.3.19 An automatic visible-range video weed detection, segmentation and classification prototype in potato field

Os autores propõem um protótipo de MV baseado em processamento de vídeo e classificadores de meta heurística para identificação e classificação *on-line* de batatas e cinco tipos de plantas daninhas. O objetivo deste protótipo é identificar e classificar corretamente as plantas de batata e as espécies de plantas daninhas; detectando, segmentando e classificando para aplicar os herbicidas localmente com um aplicador autônomo.

Para treinar propriamente o sistema de MV, vários vídeos foram feitos em plantações de batata no Irã. Depois de extrair características de cor, descritores espectrais de textura, momentos invariantes e aspectos de formato; seis destas características foram selecionadas. Foi utilizado um modelo híbrido de ANN e algoritmo cultural (ANN-CA) para classificar as diferentes

características selecionadas anteriormente, de modo a identificar diferentes plantas baseadas nesses elementos de entrada. Também foram empregados outros modelos de classificadores para comparar com o ANN-CA, são estes: FR, SVM e método LDA.

No que tange aos resultados, o modelo ANN-CA teve performance superior do que os demais algoritmos analisados, com uma acurácia de classificação de 98%. O protótipo foi testado em campo sob condições reais e foi capaz de detectar, segmentar e classificar propriamente as plantas daninhas e as plantas de batata.

3.3.20 Application of deep learning to detect Lamb's quarters (Chenopodium álbum L.) in potato fields of Atlantic Canada

Hussain *et al.* (2021) investigam a viabilidade do uso de DCNNs para detectar a espécie planta daninha *Chenopodium album L.* em plantações de batata. O conjunto de imagens é proveniente de cinco plantações diferentes, e contém imagens de plantas daninhas e plantas de batata em vários estágios de crescimento, diversas iluminações externas, diferentes condições de sombra, e com variações de espaço e tempo.

As imagens foram treinadas com os seguintes modelos de DCNN: GoogleNet, VGG-16 e EfficentNet. Também foram avaliados dois *frameworks* no treinamento, teste e durante a inferência da DCNN, os quais são PyTorch e TensorFlow. A proporção de treino, validação e teste foi de 70%, 20% e 10%, respectivamente.

Todos os modelos de DCNN tiveram um desempenho melhor com o *framework* PyTorch. Em particular, EfficientNet foi o algoritmo com melhor resultado, com acurácia máxima variando entre 92% e 97% para todas as fases de crescimento das plantas. Com isso, os autores concluem que os modelos treinados serão usados no desenvolvimento de um aplicador automático inteligente, que fará a aplicação local do herbicida.

3.3.21 Multi-Spectral Image Synthesis for Crop/Weed Segmentation in Precision Farming

Usando imagens reais, Fawakherji *et al.* (2021) criam amostras semi-artificiais, substituindo as classes de objetos mais relevantes (culturas e plantas daninhas) por suas contrapartes sintetizadas. Para isso, utilizaram *Generative Adversial Networks* (GAN) condicional, onde o modelo generativo é treinado condicionando a forma do objeto gerado. Também tem-se que este estudo introduz uma estratégia de aumento de dados que favorece uma GAN condicional a gerar cenas agrícolas inteiras, sintetizando apenas objetos mais relevantes para o propósito da segmentação.

Para avaliar o desempenho do modelo, foram utilizados dois tipos de avaliação quantitativa: a primeira testa a generalização de propriedades da GAN condicional, provando que com um número pequeno de imagens é possível gerar boas amostras sintéticas de plantas; a segunda tem como objetivo demonstrar que os *datasets* aumentados de GAN condicional podem melhorar a performance do estado da arte de arquiteturas de segmentação.

Assim, os resultados mostram que a qualidade da segmentação aumenta usando o *dataset* original aumentado com imagens sintéticas, referentes apenas ao *dataset* original. O método criado serve como ferramenta para criação de um *framework* de treinamento para problemas de segmentação.

3.3.22 Neural Network based Smart Weed Detection System

O estudo de Siddiqui *et al.* (2021) visa demonstrar um sistema automático de identificação de plantas daninhas e culturas, para realizar a aplicação local de herbicidas. O modelo desenvolvido usa uma CNN para extrair características das imagens e classificar o tipo e a porcentagem da planta detectada, de modo a fornecer uma identificação antecipada das espécies de plantas daninhas logo nos estágio iniciais. Além disso, também é implementada uma técnica de aumento de dados para

melhorar a acurácia dos resultados. Por último, é proposto um aplicador automático para aplicação do herbicida.

O *dataset* utilizado nesta pesquisa é proveniente da plataforma Kaggle e contém imagens de 4 espécies de plantas daninhas e de milho. Ademais, os dados foram separados para treinamento, validação e teste na proporção de 60%, 20% e 20%, respectivamente. A metodologia proposta foi comparada com uma CNN sem o aumento de dados.

Para cada entrada – seja milho ou uma das quatro espécies de plantas daninhas – a acurácia de predição do modelo proposto foi maior do que 80%. Portanto, os resultados mostram que a CNN com aumento de dados teve acurácia maior do que a CNN sem esse processo.

3.3.23 A new alternative to determine weed control in agricultural systems based on artificial neural networks (ANNs)

Baseado na pergunta "Quando começar o controle de plantas daninhas?", o estudo de Monteiro *et al.* (2021) tem como objetivos: avaliar a habilidade de uma ANN para estimar o começo do controle de plantas daninhas para diferentes perdas de classes de rendimento e; validar uma nova alternativa para modelar e prever a competição entre as plantas daninhas e as culturas.

A ANN escolhida foi a MLP, com a proporção de dados sendo 50%, 30% e 20% para treinamento, validação e teste, respectivamente. Para ambas as culturas de gergelim e melão, foram considerados três métodos: destrutivo, não-destrutivo e misto. Nessa etapa, os autores alertam para a necessidade de selecionar as variáveis apropriadas.

No que se refere aos resultados, tem-se que a ANN conseguiu determinar o momento ideal para fazer o controle de plantas daninhas, baseado em variáveis destrutivas e não-destrutivas. As acurácias dos métodos não-destrutivo, destrutivo e misto, foram maiores do que 95%, 90% e 95%, respectivamente. Ademais, tem-se que as técnicas de ML podem ser usadas para modelar a matocompetição.

3.3.24 Deep Learning-Based Object Detection System for Identifying Weeds Using UAS Imagery

Um grande desafio no desenvolvimento de sistemas de detecção de plantas daninhas é o requisito de um conjunto de dados propriamente rotulados, para diferenciar culturas e plantas daninhas sob condições de campo. Sendo assim, Etienne *et al.* (2021) criaram um *dataset* anotado de 374 imagens RGB organizado em classes de plantas daninhas mono e dicotiledôneas. As imagens foram adquiridas de um local de pesquisa com plantações de milho e soja nos Estados Unidos, utilizando um drone. As imagens obtidas foram tiradas em diferentes alturas (de 10m a 30m). Além disso, os autores avaliaram a performance do modelo YOLOv3 para detectar estes tipos de plantas daninhas durante os estágios iniciais de crescimento das culturas.

Durante a pesquisa, foram desenvolvidos 4 conjuntos de treinamento de imagens. Ademais, o modelo de DL utilizado foi útil em detectar várias instâncias de plantas daninhas, especialmente quando as culturas emergentes eram de cor e tamanhos semelhantes. Outro aspecto importante do processo, é que imagens de menores alturas tiveram resultados mais precisos. Por último, os pesquisadores ressaltam que os conjuntos de treinamento com apenas uma cultura foram consideravelmente melhores do que os de culturas mistas, o que demonstra a necessidade de criar *datasets* separados para as culturas de interesse.

Dessa forma, o trabalho demonstrou que conjuntos de imagens de treinamento manualmente anotados e rotulados de classes mono e dicotiledôneas na rede YOLOv3, tem resultados promissores para automatizar a detecção de plantas daninhas, por meio de drones.

3.3.25 Weed Density and Distribution Estimation for Precision Agriculture Using Semi-Supervised Learning

Shorewala *et al.* (2021) propõem uma abordagem baseada em *Deep Semi-Supervised Learning* (DSSL) para estimar a densidade e distribuição de plantas daninhas em plantações de

cenoura e beterraba, usando apenas imagens coloridas obtidas por robôs agrícolas. A distribuição e densidade das espécies de plantas daninhas pode ser útil no gerenciamento local de plantas daninhas para um tratamento seletivo de áreas infectadas usando robôs autônomos.

Utilizando uma CNN de segmentação não supervisionada, os *pixels* são – primeiro – classificados como vegetação ou fundo e, em seguida, os que foram considerados como vegetação são identificados como culturas ou plantas daninhas. Depois, as regiões infestadas por planas daninhas são reconhecidas usando uma CNN ajustada. Eliminando, assim, a necessidade de fazer uma anotação por *pixels* das características baseadas em morfologia e textura visual das plantas.

O método proposto é capaz de localizar as regiões infestadas de plantas daninhas com uma Sensibilidade máxima de 99% e estimar a densidade das espécies de plantas daninhas com uma acurácia máxima de 82%.

3.3.26 Deep convolutional neural network models for weed detection in polyhouse grown bell peppers

Subeesh *et al.* (2022) avaliam a viabilidade de diferentes técnicas de DL – AlexNet, GoogleLeNet, InceptionV3 e Xception – na identificação de plantas daninhas em imagens RGB em plantações de pimentão. O *dataset* utilizado continha 685 imagens de pimentão e 421 imagens de diversas espécies de plantas daninhas.

As variações de ruído e iluminação foram removidas durante o pré-processamento dos dados. Também foi aplicado um aumento de dados sobre o conjunto de imagens, para aumentar o tamanho e a qualidade do *dataset* de treinamento e prevenir o *overfitting*. O experimento foi repetido com número de épocas de 10, 20 e 30; este aumento na quantidade de épocas melhorou significativamente a acurácia dos modelos.

Todos os algoritmos tiveram uma performance satisfatória, com acurácia variando entre 94% e 97%. Não obstante, o modelo InceptionV3 foi o que teve melhores resultados, com acurácia,

Precisão, e Sensibilidade de 97,7%, 98,5% e 97,8%, respectivamente. Também vale destacar que não houve casos de *overfitting* e underfitting para nenhum dos modelos.

3.3.27 Deep learning-based precision agriculture through weed recognition in sugar beet fields

Neste estudo, os autores usam a arquitetura U-Net como um *encoder-decoder* da CNN ResNet50, para realizar uma segmentação semântica por *pixels* em plantas de beterraba e plantas daninhas. O *dataset* coletado continha 1385 imagens coletadas sob diferentes condições (de estágio de crescimento, de grau de cobertura das plantas daninhas, de texturas dos solos, de tempos de coleta, de altura em que as fotos foram tiradas e de várias variedades de espécies de plantas daninhas).

Durante a fase de treinamento, combinações aleatórias de aumento de dados foram aplicadas, tais como: rotação, inversão de largura e altura, cortes, *zoom* e, *blurring*; para evitar *overfitting* e estender artificialmente o *dataset*. Aumentando, assim, a habilidade do modelo de generalizar problemas do mundo real e melhorar o processo de aprendizado no geral.

Houve alguns pontos em que o modelo confundiu a segmentação das classes beterraba, solo e plantas daninhas. Esse problema ocorreu principalmente quando as plantas daninhas e as plantas de beterraba estavam sobrepostas. Mesmo assim, os resultados mostram que modelos baseados em DL são eficientes para tarefas de segmentação; obtendo uma acurácia de 96%.

3.3.28 Hybrid leader based optimization with deep learning driven weed detection on internet of things enabled smart agriculture environment

Alrowais *et al.* (2022) introduzem um novo modelo de otimização híbrida baseada no líder com detecção de plantas daninhas orientada por DL em *smart agriculture* habitada por Internet das

coisas (*Internet of Things* - IoT). Ou seja, o objetivo do modelo é coletar imagens usando dispositivos IoT e fazer o reconhecimento automático das plantas daninhas.

Inicialmente, o modelo faz com que os dispositivos capturem imagens das plantações e as transmitam para um servidor *cloud* para serem examinadas. Em seguida, é aplicado o modelo YOLOv5 para fazer a detecção de plantas daninhas, de modo que o algoritmo de otimização híbrido baseado no líder é explorado como um otimizador de hiperparâmetros. Por último, o modelo *Kernel Extreme Learning Machine* (KELM) é aplicado para realizar a classificação das espécies de plantas daninhas.

O modelo proposto foi experimentalmente validado, usando um *dataset* com 287 imagens de culturas e 2713 imagens de plantas daninhas. Também foi feita uma análise comparativa entre o sistema apresentado e outros algoritmos existentes. Tem-se que KNN teve uma performance de classificação pior; SVM teve um resultado ligeiramente superior, seguido por FR, ResNet-101 e VGG-16. Poreḿ, o modelo proposto superou os outro métodos, produzindo valores máximos de acurácia, Precisão e Sensibilidade; os quais foram 98,8%, 93,4% e 95%, respectivamente.

3.3.29 Weed detection in soybean crops using custom lightweight deep learning models

Assim como em outros dos trabalhos correlatos, Razfar *et al.* (2022) propoẽm um sistema de detecção de plantas daninhas usando modelos de DL, neste caso em plantações de soja. Foram usados cinco modelos de DL: MobileNetV2, ResNet50 e três CNNs customizadas. O objetivo é identificar plantas daninhas na soja com relação a grama, espécies de folha larga e solo.

O modelo MobileNetV2 foi o que teve menor acurácia de todos os algoritmos avaliados. Já o ResNet50 performou bem melhor, com 82% de acurácia na validação. Entretanto, a CNN customizada de 5 camadas superou o desempenho dos outros modelos, inclusive das outras redes personalizadas; com 97,7% de acurácia. A segunda melhor foi a CNN de 8 camadas, com 97,1% de acurácia. Por último, a CNN de 4 camadas teve acurácia de 90,3% na validação. Portanto, tem-se

que a CNN de 5 camadas capturou as características essenciais para identificar as classes, apesar do *dataset* não estar bem balanceado.

3.3.30 Detection of Parthenium Weed (Parthenium hysterophorus L.) and Its Growth Stages Using Artificial Intelligence

Costello *et al.* (2002) apresentam um novo método de detectar e mapear populações da espécie *Parthenium hysterophorus L.* em um ambiente de pasto simulado, usando imagens RGB ou hiperespectrais com auxílio de uma IA, nas quais as plantas daninhas – da espécie *Parthenium* e outras – estão em diversos estágios de crescimento (flora e não-flora).

As imagens RGB foram processadas com a CNN YOLOv4, atingindo uma acurácia de 95% para detecção e 86% para classificação dos estágios de flora e não-flora das plantas daninhas. Também foi utilizado um classificador XGBoost para classificação de *pixels* do *dataset* hiperespectral, obtendo uma acurácia de classificação de 99% para cada classe de estágio de crescimento da espécie *Parthenium*. Além disso, quando plantas *Parthenium* e de outras espécies foram artificialmente combinadas em várias permutações, a acurácia de classificação por *pixel* foi de 99% para ambas as classes.

3.3.31 Diversification of traditional paddy field impacts target species in weed seedbank

Dominschek *et al.* (2022) tem como objetivo avaliar o impacto de um arrozal tradicional e quatro sistemas ILP sobre o banco de sementes de plantas daninhas em um experimento de longa duração, no sul do Brasil. Os quatro sistemas ILP constituem de dois sistemas de integração arroz-pecuária, um sistema de integração soja-arroz-pecuária e um sistema de integração soja-milho-arroz-pecuária.

Para realizar o experimento, foram coletadas amostras de solo de diferentes profundidades. Todas as amostras foram processadas para remover pedras e raízes. Depois foram espalhadas em bandejas e colocadas em uma estufa pelo período de 12 meses. Durante esse tempo, foram quantificadas apenas as sementes germináveis, mas não as em estado de dormência. As mudas emergidas foram periodicamente identificadas e removidas das bandejas. Quando não era possível identificar a espécie, a muda era transplantada para um vaso, até que crescesse o suficiente para ser reconhecida. As contagens totais de todas as espécies foram somadas para calcular o tamanho do banco de sementes de plantas daninhas.

Observou-se que a diminuição de sementes de plantas daninhas é mais pronunciada em sistemas ILP que integram diferentes culturas de verão e compostagens hibernais. Em particular, os sistemas de cultivo arroz-pecuária tiveram as sementes das espécies planta daninhas mais localizadas entre 0 cm e 5 cm, que – embora sejam mais prováveis de germinar – estão mais suscetíveis a dissecação, predação e variações do clima; o que favorece a redução dessas sementes. Já para os outros dois sistemas ILP, a redução no banco de sementes foi devido a diversificação de culturas e, por consequência, ao uso de diferentes herbicidas. Outro aspecto que os autores destacam sobre o ILP, é a presença de resíduos (palha) antes da semeadura da safra de verão, que é considerado um dos principais fatores que regula a comunidade de plantas daninhas.

3.3.32 A deep convolutional neural network-based method for identifying weed seedlings in maize fields

Utilizando diferentes DCNNs, Ni *et al.* (2022) desenvolvem um modelo para ajudar no reconhecimento e remoção de plantas daninhas. As DCNNs foram usadas para extração de características na camada convolucional. Os modelos selecionados foram AlexNet, InceptionV3, DenseNet e VGG-16.

Dentre os algoritmos, VGG-16 foi o modelo com melhor capacidade de reconhecimento de plantas daninhas, com *F1 Score* acima de 90% para todas as espécies de plantas daninhas

consideradas. Os autores acrescentam que para trabalhos futuros o sistema será integrado em um *spray* inteligente.

3.3.33 Automated Weed Detection System for Bok Choy Using Computer Vision

Ngo *et al.* (2022) apresentam uma segmentação de instâncias usando *Region-based Convolutional Neural Networks* (RCNN), para criar um sistema de detecção de plantas daninhas capaz de diferenciar repolhos de outras plantas. O objetivo dos autores é desenvolver uma tecnologia de MV que segmente objetos para gerar a forma de plantas daninhas e, com isso, calcular e localizar a centroide para fazer a aplicação de um *laser* na espécie planta daninha. Os dados foram coletados a partir de um ambiente controlado, que continha solo, repolho e outras plantas.

Os modelos de RCNN ResNet50 e ResNet101, foram treinados cada um com três variações do *dataset* original, num total de seis combinações. O primeiro conjunto consiste de imagens reais do ambiente controlado, o segundo contém imagens sintéticas geradas a partir de elementos do *dataset* real, e o terceiro é uma combinação dos outros dois. Os dados foram separados para treino, validação e teste na proporção de 70%, 15% e 15%, respectivamente.

Quanto aos *datasets*, os resultados foram melhores para as imagens reais, seguido das imagens combinadas e, as fotos geradas foram as que tiveram o pior resultado. No que se refere aos algoritmos, o ResNet50 teve performance superior na detecção de plantas daninhas do que o ResNet101. Assim, o ResNet50 foi o melhor modelo para identificação de objetos, enquanto o ResNet101 foi melhor para segmentação de instâncias.

3.3.34 Classification of Weeds and Crops using Transfer Learning

Nesta pesquisa, Jose *et al.* (2022) propoẽm um método para classificação automática de imagens em culturas e plantas daninhas, baseado em *Transfer Learning* (TL). Foram utilizadas as arquiteturas de CNN MobileNetV2, NASNetLarge e NASNetMobile.

O *dataset* empregado contém 347 imagens de tomate e duas espécies de plantas daninhas. Para aumentar a variabilidade do conjunto de dados, foi feito o processo de aumento de dados. Como resultado, o modelo MobileNetV2 foi o que teve melhor desempenho, com acurácia de teste de 97%.

3.3.35 Weeding Robot Based on Lightweight Platform and Dual Cameras

Wang e Leelapatra (2022) apresentam um robô capinador de trilhos com câmeras duplas para detecção e remoção de plantas daninhas. Em comparação com os robôs existentes que causam compactação do solo, o novo método pode evitar esse fenômeno e – ao mesmo tempo – manter a estabilidade do robô durante o trabalho. Baseado em técnicas de MV, DL e CNN, uma nova rede de identificação leve é proposta: GF-YOLO. Vale ressaltar que a CNN também foi usada na rede para garantir a extração eficiente das características das imagens.

Os autores demonstram que o uso de câmeras duplas para detectar plantas daninhas em diferentes ângulos evita problemas como oclusão e reflexão. Assim, os resultados mostram que um robô de câmera dupla baseado em uma plataforma leve tem acurácia alta, comparado com o método de câmera única. O modelo proposto obteve Precisão de 98% e Sensibilidade de 83%.

3.3.36 Real-time Weed Identification Using Machine Learning and Image Processing in Oil Palm Plantations

Neste estudo, os autores propoẽm um sistema baseado em MV para reconhecer plantas daninhas em plantações de óleo de palma e integram o sistema a um aplicativo de celular para

recomendar o planejamento apropriado para o manejo das espécies planta daninhas baseado em herbicidas.

O fluxo do trabalho consiste da coleta dos dados de plantas daninhas e de herbicidas, rotulação dos dados, configuração do modelo e treinamento do algoritmo. Durante o processo também são utilizadas técnicas de TL e CNN.

As imagens baixadas no sistema são processadas usando o modelo salvo. Depois disso, o algoritmo lê as imagens para extrair padrões para o processo de reconhecimento, e os resultados são salvos em um arquivo. O sistema continua para rotular os resultados do modelo, baseado nos rótulos. Esses consistem do nome da espécie, nome comum, nível de perigo e sugestões para o controle com herbicidas.

3.3.37 Crop Yield Improvement with Weeds, Pest and Disease Detection

Meena *et al.* (2023) desenvolveram um modelo neural utilizando diferentes tipos de DCNN, para detectar características de doenças, pestes e plantas daninhas presentes em diferentes culturas. O *dataset* utilizado na pesquisa consiste de 4 tipos de plantas daninhas, 38 doenças e 9 grupos de pestes. As arquiteturas selecionadas foram DenseNet201, MobileNet, VGG-16, InceptionV3 e busca de hiperparâmetros. Por último, tem-se que o modelo produzido foi implementado em um *site*.

Para a detecção de pestes, o modelo MobileNet teve a melhor acurácia: 88,75%. Já o DenseNet021 foi o que melhor performou para identificação de doenças, com acurácia de 99,06%. Quanto às plantas daninhas, a acurácia máxima de detecção foi de 97,56% pela busca de hiperparâmetros.

3.3.38 Effect of varying training epochs of a Faster Region-Based Convolutional Neural Network on the Accuracy of an Automatic Weed Classification Scheme

Ajayi e Ashi (2023) implementam uma RCNN para identificação e classificação de plantas daninhas em uma fazenda de culturas mistas (cana-de-açúcar, espinafre, banana e pimenta). As imagens foram adquiridas por um drone e, na sequência, foram anotadas.

Foi observado que a performance do modelo melhorou significativamente conforme o número de épocas aumentava. Mas ficou saturada depois de 242000 épocas. Sendo que a melhor acurácia de classificação foi de 98,4% com 200000 épocas. Logo, é demonstrado pelos resultados que a RCNN é capaz de identificar plantas daninhas em plantações mistas.

3.3.39 YOLOWeeds: A novel benchmark of YOLO object detectors for multi-class weed detection in cotton production systems

Dang *et al*. apresentam um *dataset* de plantas daninhas em plantações de algodão nos Estados Unidos. Foram coletadas 5648 imagens de 12 classes de plantas daninhas num total de 9730 anotações de *bounding boxes*. As imagens foram adquiridas sob diferentes condições de iluminação natural e com as plantas daninhas em diferentes estágios de crescimento.

Na sequência, foi avaliado o desempenho de diferentes modelos YOLO na detecção de plantas daninhas. As versões do modelo, em diferentes tamanhos são: YOLOv3, YOLOv4, Scaled-YOLOv4, YOLOR, YOLOv5, YOLOv6 e YOLOv7. Para facilitar o treinamento do algoritmo, todos os detectores de objetos YOLO foram treinados via TL.

De resultados, todos os modelos YOLO mostraram grande potencial para detecção de plantas daninhas em tempo real. Em particular, tem-se que o YOLOv3-tiny teve 88,1% de acurácia, enquanto o YOLOv4 95,2%. No geral, o YOLOv4 teve acurácia maior que o YOLOv3 e o

YOLOv5. Todavia, o YOLOv5-nano e o YOLOv5-small mostraram vantagens em tempos de inferência rápidos, ao mesmo tempo que obtiveram acurácia de detecção comparáveis.

3.3.40 Real-time control of high-resolution micro-jet sprayer integrated with machine vision for precision weed control

Raja *et al.* (2023) propoẽm um sistema inteligente de controle de plantas daninhas usando o conceito de sinalizador de culturas com um sistema de MV, e integrado com um micro-*spray* para fazer a aplicação local dos herbicidas. Tem-se que a sinalização de culturas é uma tecnologia inventada para que máquinas façam a leituras das plantas cultivadas, para simplificar a tarefa de diferenciar culturas de plantas daninhas e, assim, fazer o controle seletivo de espécies de plantas daninhas em tempo real.

Para isso, o sistema adquire – primeiro – imagens digitais por meio de uma câmera e o algoritmo de MV usa a sinalização de culturas para fazer a identificação de alface e de plantas daninhas, de várias espécies e tamanhos. Criando, assim, um mapa de localização de culturas e espécies de plantas daninhas. Com base no mapa, o micro-*spray* foi constituído e controlado em tempo-real para atingir as plantas daninhas com o herbicida.

Os resultados indicam a eficiência do modelo proposto, já que 99% das culturas e 84% das plantas daninhas foram corretamente detectadas e classificadas. Dentre estas, 98% foram assertivamente pulverizadas e, apenas 46% das plantas de alface foram incorretamente pulverizadas.

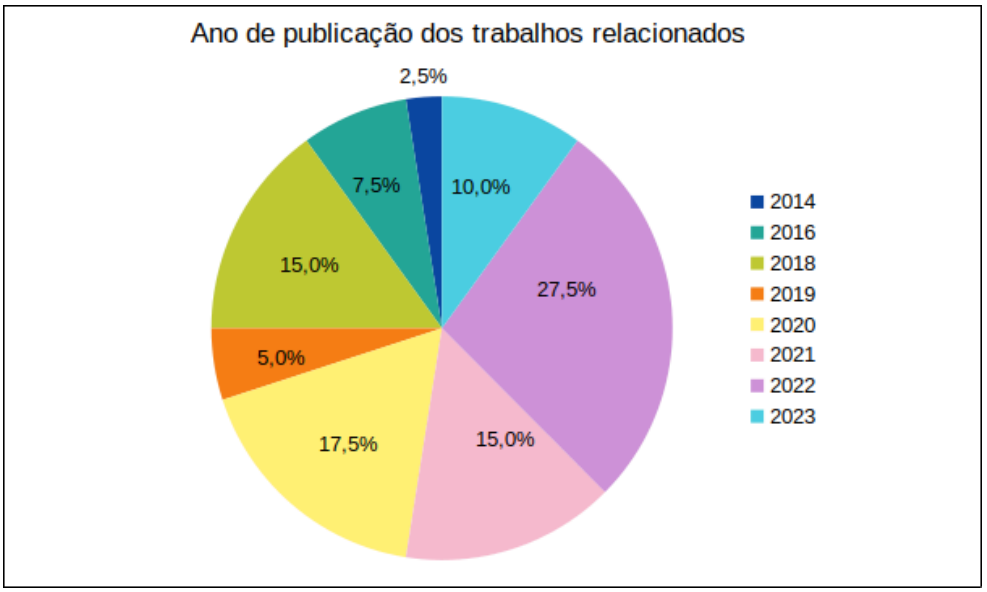
3.4 ANÁLISE COMPARATIVA

A partir da revisão sistemática da literatura que foi realizada, é preciso fazer uma análise comparativa entre os trabalhos selecionados. Podem ser comparados nas pesquisas aspectos como: as técnicas e os algoritmos de aprendizado de máquina utilizados; os anos e locais de

desenvolvimentos dos estudos; as soluções confeccionadas a partir dos modelos usados e as culturas em que os experimentos foram conduzidos.

É possível observar que a partir de 2020 houve um aumento na quantidade de pesquisas sobre esse tema. Em particular, observa-se que apesar dos estudos realizados em 2023 corresponderem apenas aos trabalhos publicados entre os meses de janeiro e março, estes representam 10% de todas as pesquisas selecionadas. Além disso, antes de 2014 não há registros do uso de inteligência artificial no manejo de plantas daninhas. Assim como ilustra a Figura 13.

Figura 13: Ano de publicação dos trabalhos relacionados.



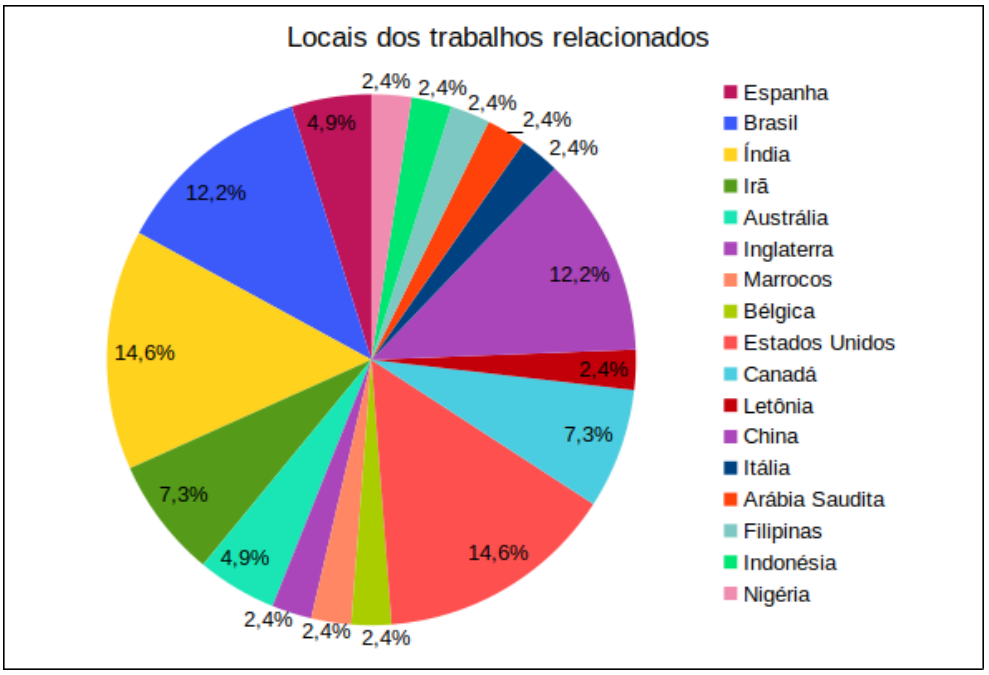
Fonte: Elaborada pela autora.

Tal fato demonstra que a utilização de técnicas de aprendizado de máquina no controle de plantas daninhas está crescendo. Isso ocorre pois há uma maior disponibilidade de dados sobre o assunto, bem como os algoritmos de AM estão ficando cada vez mais evoluídos e populares.

Outra observação, é que a maioria desses estudos foram realizados nos Estados Unidos e na Índia – tal como demonstra a Figura 14. Também é possível observar que o continente com maior quantidade de pesquisas é a Ásia. Além disso, tem-se que a maioria dos artigos brasileiros

selecionados são sobre o sistema ILP, e não sobre a aplicação de aprendizado de máquina no manejo das plantas daninhas.

Figura 14: Locais dos trabalhos relacionados.



Fonte: Elaborada pela autora.

O quadro a seguir apresenta as culturas, os algoritmos de aprendizado de máquina e as soluções desenvolvidas em cada um dos trabalhos relacionados. Também é possível verificar estas informações nas Figuras 15, 16 e 17.

Quadro 6: Culturas, técnicas de AM e soluções desenvolvidas dos trabalhos correlatos.

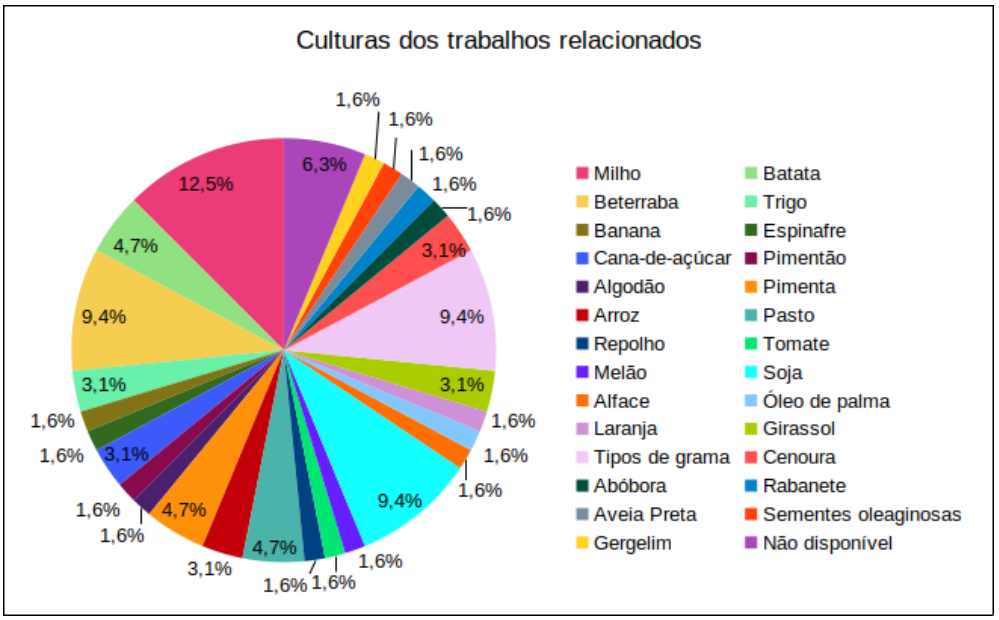
Identificação	**Culturas**	**Técnicas e Algoritmos de ML**	**Soluções desenvolvidas**
Torres-Sospedra e Nebot (2014)	Laranja	ANN, MLP e BPNN	Não disponível
Pérez-Ortiz et al. (2016)	Milho e Girassol	SVM	Não disponível
Schuster et al. (2016a)	Pasto e Soja	Não disponível	Não disponível

Lutosa et al. (2016)	Tipos de grama, Soja, Milho, Trigo e Aveia Preta	Não disponível	Não disponível
Chavan e Nandedkar (2018)	Tipos de grama, Milho, Trigo e Beterraba	CNN e TL	Não disponível
Sabzi e Abbaspour-Gileeh (2018)	Batata	ANN, MV e PSO	Não disponível
Sandino e Gonzalez (2018)	Tipos de grama	MV	*Framework*
Zhang et al. (2018)	Pasto	DL, CNN, SVM, KNN, RL e AC	Não disponível
Abouzahir et al. (2018)	Soja	SVM, MV e BPNN	Aplicadores inteligentes
Gao et al. (2018)	Milho	KNN e RF	Câmera hiperespectral
Yu et al. (2019)	Tipos de grama	MV e DCNN	Aplicadores inteligentes
Partel et al. (2019)	Pimenta	DL, MV e CNN	Aplicadores inteligentes
Sudars et al. (2020)	Cenoura, Abóbora, Rabanete e Beterraba	MV	Dataset de imagens
Qiao et al. (2020)	Não disponível	CNN, MV e DL	Não disponível
Souza et al. (2020)	Cana-de-açúcar	RF e SIMCA	Não disponível
Yan et al. (2020)	Arroz	SVM, DCNN e KNN	Não disponível
Wang et al. (2020)	Beterraba e Sementes oleaginosas	DL e TL	Não disponível
Yu et al. (2020)	Tipos de grama	MV e DCNN	Aplicadores inteligentes
Sabzi et al. (2020)	Batata	MV, ANN, LDA, RF e SVM	Aplicadores inteligentes
Hussain et al. (2021)	Batata	MV, DL, DCNN e TL	Spray inteligente
Fawakherji et al. (2021)	Beterraba e Girassol	GAN	Robô agrícola e Framework
Siddiqui et al. (2021)	Milho	CNN	Aplicadores

			inteligentes
Monteiro et al. (2021)	Melão e Gergelim	ANN e MLP	Não disponível
Etienne et al. (2021)	Soja e Milho	MV, DL e TL	Dataset de imagens
Shorewala et al. (2021)	Cenoura e Beterraba	DSSL, CNN, DL, SVM, RF, MV e TL	Robô agrícola
Subeesh et al. (2022)	Pimentão	CNN, MV, DL e DCNN	Não disponível
Nasiri et al. (2022)	Beterraba	CNN e DL	Não disponível
Alrowais et al. (2022)	Não disponível	DL, KELM, KNN, SVM, RF, CNN e MV	Não disponível
Razfar et al. (2022)	Soja	CNN, DL e TL	Não disponível
Costello et al. (2022)	Tipos de grama	CNN, DL e DT	Não disponível
Dominschek et al. (2022)	Arroz, Milho, Pasto e Soja	Não disponível	Não disponível
Ni et al. (2022)	Milho	MV e DCNN	Não disponível
Ngo et al. (2022)	Repolho	MV, RCNN e TL	Laser inteligente
Jose et al. (2022)	Tomate	CNN, DL e TL	Não disponível
Wang e Leelapatra (2022)	Não disponível	CNN, MV e DL	Robô agrícola
Firmansyah et al. (2022)	Óleo de palma	CNN, MV e TL	Aplicativo de celular
Meena et al. (2023)	Não disponível	DL e DCNN	Não disponível
Ajayi e Ashi (2023)	Banana, Espinafre, Cana-de-açúcar e Pimenta	CNN, DL e RCNN	Não disponível
Dang et al. (2023)	Algodão	MV, DL e TL	Dataset de imagens
Raja et al. (2023)	Alface	MV e TL	Aplicadores inteligentes

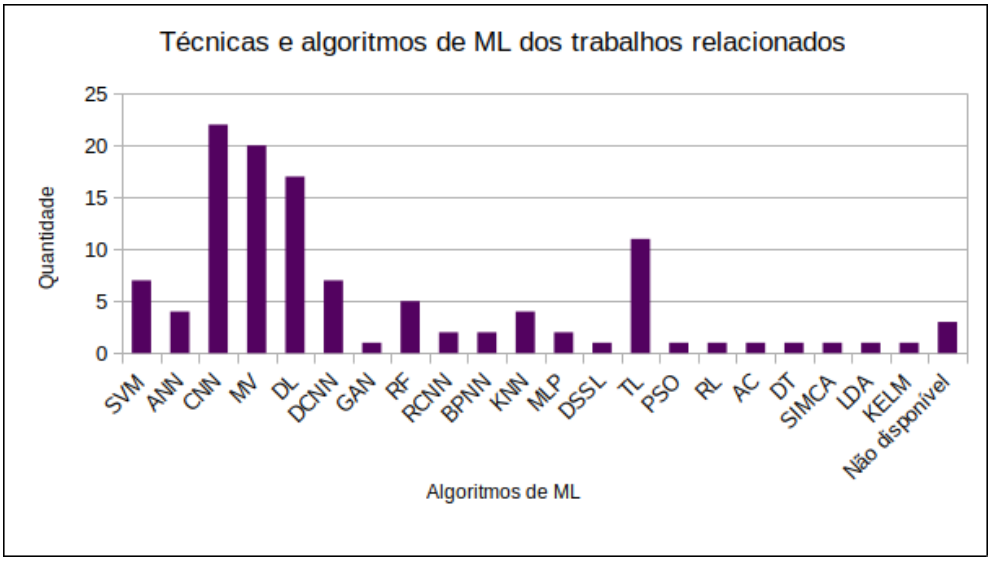
Fonte: Elaborada pela autora.

Figura 15: Culturas dos trabalhos relacionados.



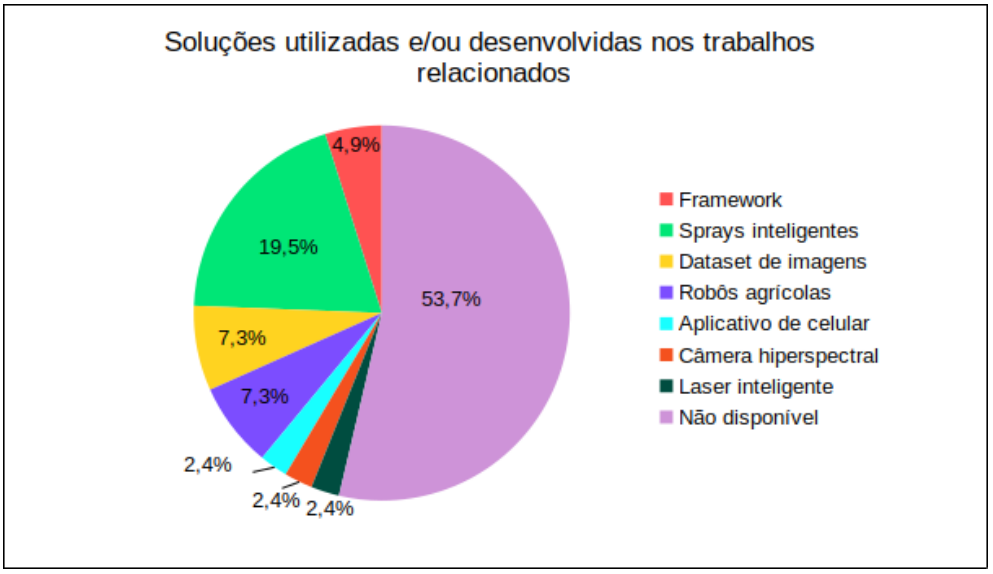
Fonte: Elaborada pela autora.

Figura 16: Técnicas e algoritmos de aprendizado de máquina dos trabalhos relacionados.



Fonte: Elaborada pela autora.

Figura 17: Soluções desenvolvidas nos trabalhos relacionados.



Fonte: Elaborada pela autora.

Em relação às plantações, a cultura que mais apareceu nos trabalhos correlatos foi o milho. Isso porque o milho é uma das culturas mais produzidas (ERENSTEIN *et al.*, 2022). Ademais, há uma dificuldade em manter o controle de plantas daninhas de folha estreita nessas plantações, já que o suprimento de herbicidas para esse tipo de espécies é baixo e, por consequência, os custos são altos (VARGAS *et al.,* 2006). Também vale ressaltar que para algumas pesquisas a cultura foi definida como "Não disponível", pois esta era mencionada de maneira genérica como "vegetais" ou "culturas".

Ademais, tem-se que a classificação "Tipos de grama" inclui algumas variedades de gramas mencionadas nos trabalhos, como *Bermudagrass; Ryegrass; Desert bluegrass;* entre outras. Apesar de ser uma das "culturas" mais utilizadas nas pesquisas, esta não se refere a áreas de produção de alimentos, mas sim a campos de golfe. Isso se deve ao fato de que países como Estados Unidos utilizam boa parte das terras – a qual seria destinada para plantações – em atividades de lazer (YU *et al.*, 2019).

No que diz respeito aos modelos de aprendizado de máquina, tem-se que o algoritmo mais utilizado foi o CNN e as técnicas mais empregadas foram MV e DL. Contudo, tem-se que CNN é

um tipo de modelo de DL e também é usado para problemas de MV. Logo, como a maioria dos trabalhos correlatos foca na identificação e classificação de plantas daninhas, tem-se que a escolha por modelos de processamento de imagens faz sentido. Assim, tem-se a resposta da pergunta de análise de item 1, elaborada na Seção 3.1.

Os artigos classificados como “Não Informado” referem-se aos estudos sobre o sistema ILP e as plantas daninhas, de forma que não envolvem modelos de aprendizado de máquina. Afinal, não foi possível identificar os algoritmos de AM usados em sistemas ILP no manejo de espécies de plantas daninhas, já que não há estudos sobre o tema.

Ainda em relação aos algoritmos de aprendizado de máquina, tem-se que o intuito em identificar as técnicas mais empregadas nos trabalhos correlatos também era para selecionar e empregar tais algoritmos no desenvolvimento do projeto. Todavia, tem-se que foi optado não utilizar o CNN neste trabalho, já que este não envolve o processamento de imagens.

Assim, tem-se que os algoritmos elegidos foram: árvore de decisão, floresta randômica, máquinas de vetores de suporte e k-vizinhos mais próximos. A escolha destes se deve ao fato de que, a maioria dos trabalhos correlatos utilizou o CNN para realizar o processamento de imagens, mas quando tinham algum processo de classificação para fazer durante seu desenvolvimento, utilizavam os modelos citados acima. Logo, como tais algoritmos foram os mais empregados para tarefas de classificação, estes foram os selecionados para serem utilizados no trabalho.

Quanto às soluções desenvolvidas, tem-se que muitos trabalhos apenas investigaram a performance dos algoritmos de aprendizado de máquina no controle de plantas daninhas, mas não os aplicaram em nenhuma solução prática. Em alguns casos, os autores até propuseram ideias de soluções, mas não as desenvolveram. Além disso, deve-se considerar que os registros “Não disponível” também referem-se às pesquisas sobre espécies de plantas daninhas e sistemas ILP.

Todavia, para os estudos que produziram alguma aplicação prática, a mais comum foi aplicadores inteligentes, seguida de *datasets* de imagens e robôs agrícolas – tais soluções práticas respondem a pergunta de análise de item 2. Os *datasets* propostos são de imagens de plantas

daninhas e culturas, e tem como objetivo serem usados para treinar e melhorar os algoritmos de aprendizado de máquina, para terem um melhor desempenho na detecção de espécies de plantas daninhas.

Já os robôs agrícolas, são criados para fazer a remoção das plantas daninhas e/ou aplicar agroquímicos. Do mesmo modo, os aplicadores inteligentes – também chamados de *sprays* inteligentes ou *smart sprayers* –, os quais normalmente são drones, também são usados para realizar a aplicação local do herbicida. Ambas as aplicações precisam ser capazes de distinguir as culturas das plantas daninhas para poderem realizar suas funções, seja removendo mecanicamente a espécie de planta daninha ou aplicando o herbicida. Por isso que métodos de AM de processamento de imagens, tais como MV e DL, são os mais empregados.

Apesar de tais soluções apresentarem vantagens, como reduzir a contaminação da comida e do meio ambiente, estas requerem um investimento financeiro muito alto. Além disso, tem-se que tais tecnologias podem trazer diversas consequências para o mercado de trabalho da área rural (OECD, 2021). Vale destacar que todas estas aplicações não se referem ao manejo de plantas daninhas em sistemas ILP, apenas a sistemas de lavoura contínua.

Observa-se tanto pelos modelos de aprendizado de máquina, bem como pelas soluções apresentadas nos trabalhos correlatos, que há uma disparidade quanto à natureza dessas tecnologias. Tem-se que 51% dos sistemas de aprendizado de máquina e 26% das aplicações propostas são sobre processamento de imagens. O que demonstra o grande enfoque dado para esse tipo de tecnologia, e responde a pergunta de análise de item 3.

Com relação à pergunta de análise de item 4, tem-se que muitos dos trabalhos relacionados propuseram diversos sistemas para o manejo de plantas daninhas, mas não as colocaram em prática para averiguar a possível taxa de redução destas. Quanto às pesquisas que fizeram testes em campo, estas apenas avaliaram a eficiência de seus modelos e/ou soluções, mas não verificaram a taxa de redução de plantas daninhas.

3.5 COMPLEMENTO DA REVISÃO SISTEMÁTICA DA LITERATURA

Os resultados apresentados anteriormente referem-se aos trabalhos encontrados na literatura até o primeiro trimestre de 2023 – período em que a revisão foi realizada. Todavia, com o intuíto de verificar se há estudos mais recentes sobre o tema, foi feito um complemento da revisão sistemática da literatura, relativo ao intervalo de 01/04/2023 até 31/04/2025.

Tem-se que esta segunda revisão também foi baseada na metodologia PRISMA. Da mesma forma, as perguntas de análise, as *strings* de busca, e os repositórios escolhidos também são os mesmos apresentados na Seção 3.1. Em relação aos critérios de inclusão e exclusão, tem-se que estes tiveram algumas modificações, conforme pode ser visto na sequência:

Quadro 7: Critérios de Inclusão e Exclusão da segunda Revisão Sistemática da Literatura.

Inclusão	Exclusão
CI1: Artigos em inglês ou português.	CE1: Artigos duplicados.
CI2: Artigos acessíveis via Portal CAPES.	CE2: Artigos que não permitam acesso ao seu texto completo
CI3: Artigos publicados entre 01/04/2023 até 31/04/2025.	CE3: Artigos que não tenham “weed management”, “weed control” ou “weed(s)” como palavras-chave.
	CE4: Artigos que já foram selecionados na primeira revisão

Fonte: Elaborada pela autora.

Assim, a data de análise foi alterada para complementar a primeira revisão, conforme já havia sido mencionado; e o critério de exclusão CE4 foi adicionado justamente para garantir que nenhum trabalho selecionado na primeira revisão fosse escolhido de novo. Os demais critérios de inclusão e exclusão permanecem os mesmos, bem como o critério de qualidade, no qual os títulos, resumos e palavras-chave foram analisados para averiguar se os estudos retornados pelas bases eram condizentes com os objetivos do trabalho.

As etapas que compõem esta segunda revisão são as iguais às da primeira, apenas tem-se que na segunda etapa o critério CE4 também foi aplicado. Isto é: na primeira etapa foram removidas todas as duplicatas (CE1); na segunda foram empregados os critérios CI1, CI3, CE3 e CE4; e na terceira e última etapa foi averiguado para quais trabalhos era possível acessar o seu conteúdo completo. O Quadro 8 ilustra a quantidade de estudos selecionados em cada etapa para cada repositório. Vale ressaltar, que aspectos como técnicas de aprendizado de máquina, culturas, países e soluções desenvolvidas não serão analisados nesta segunda revisão, visto que o objetivo é apenas verificar se houve mais estudos sobre o tema nos últimos anos.

Quadro 8: Quantidade de estudos encontrados e selecionados por repositório na segunda Revisão Sistemática da Literatura.

Base de Dados	**Encontrados**	**1ª Etapa**	**2ª Etapa**	**3ª Etapa**
IEEE Xplore	374	373	162	12
Mendeley	8	5	2	0
Science Direct	1856	1856	103	103
Scopus	6	6	0	0
Total	**2244**	**2240**	**267**	**115**

Fonte: Elaborada pela autora.

Na base de dados Mendeley foram encontrados apenas 8 artigos, sendo que todas as etapas tiveram trabalhos eliminados por algum dos critérios, não sobrando nenhum ao final. O mesmo ocorreu para a base Scopus, a qual teve poucos retornos, e todos os estudos foram eliminados por serem de antes da data de análise; de modo que não há trabalhos mais recentes sobre este tema em tal biblioteca. Entretanto, enquanto estas bases indicam uma falta de pesquisas sobre o assunto, os repositórios IEEE Xplore e Science Direct demonstram o contrário: que de 2023 em diante houve um grande aumento de pesquisas sobre aprendizado de máquina para controle de plantas daninhas.

Tanto no IEEE Xplore quanto no Science Direct, muitos artigos foram rejeitados por não se enquadrarem nos objetivos do projeto. Pela análise do resumo destes, observou-se que há muitos

estudos sobre o uso de aprendizado de máquina na agricultura no geral, tais como para detecção de doenças e pestes em plantas. Também apareceram bastante trabalhos sobre identificação, classificação e contagem de culturas. Porém, a maioria das pesquisas que eram sobre plantas daninhas, utilizavam técnicas de processamento de imagens. Quanto ao assunto sistemas ILP e plantas daninhas, em ambos os repositórios foram encontrados alguns artigos sobre sistemas de rotação de culturas para ajudar no manejo de plantas daninhas, mas que não necessariamente eram o sistema ILP.

Portanto, a partir desta segunda revisão sistemática da literatura – cujo objetivo é servir de complemento a primeira – foi possível perceber que, enquanto repositórios como Mendeley e Scopus parecem estar desatualizados sobre o tema, bases como IEEE Xplore e Science Direct exibem um notável crescimento de aplicações de aprendizado de máquina nos últimos anos, não só no controle de plantas daninhas mas na agricultura como um todo. Novamente, com um enfoque maior para o processamento de imagens.

3.6 CONSIDERAÇÕES

Os trabalhos selecionados mostram que a partir de 2020 houve um aumento no uso de técnicas de aprendizado de máquina no controle de plantas daninhas, em particular, nos Estados Unidos e na Índia. Também é evidenciado que a principal cultura utilizada nas pesquisas de manejo de plantas daninhas foi o milho.

Quanto aos algoritmos de AM, tem-se que o modelo mais utilizado foi *Convolutional Neural Networks* (CNN), por consequência, as técnicas mais aplicadas foram o *Machine Vision* (MV) e *Deep Learning* (DL). Já as soluções práticas desenvolvidas nas pesquisas que mais apareceram foram os aplicadores inteligentes. Observa-se então que as tecnologias utilizadas no manejo de plantas daninhas estão focadas na área de processamento de imagens. Diante disso, há uma carência na literatura sobre outros aspectos do controle de plantas daninhas que poderiam ser explorados.

Também vale ressaltar que a maioria dos trabalhos selecionados referem-se ao uso de algoritmos de aprendizado de máquina em sistemas de lavoura contínua. Nenhuma pesquisa sobre tecnologias aplicadas no controle de espécies de plantas daninhas em sistemas ILP foi identificada. Na verdade, percebeu-se uma falta de estudos sobre sistemas ILP de modo geral.

Neste capítulo foi apresentado o processo de busca, exclusão e inclusão de trabalhos relevantes para a pesquisa. Por meio de uma expressão de busca ampla e critérios de seleção, foram escolhidas pesquisas que auxiliaram na *(i)* identificação dos modelos de aprendizado de máquina mais utilizados no manejo de plantas daninhas e *(ii)* no reconhecimento das características dos trabalhos selecionados.

A Revisão Sistemática da Literatura desta dissertação foi publicada como um artigo pela revista Advances in Weed Science. O título do artigo é *Machine learning algorithms applied to weed management in integrated crop-livestock systems: a systematic literature review*. Tal artigo foi escrito por Ana Letícia Becker Gomes, Anita Maria da Rocha Fernandes, Bruno Araújo Cautiero Horta e Maurílio Fernandes de Oliveira.

4 DESENVOLVIMENTO

Este capítulo apresenta as etapas realizadas no desenvolvimento do projeto, de forma a cumprir com os objetivos apresentados na Seção 1.2. Primeiro, os *datasets* disponibilizados pela EMBRAPA foram organizados e compilados em uma base de dados unificada. Na segunda fase, foi realizado o pré-processamento dos dados. Com isso, foi possível fazer uma análise estatística dos dados e também implementar e avaliar os algoritmos de aprendizado de máquina.

4.1 ORGANIZAÇÃO DOS *DATASETS*

Os dados disponibilizados pela EMBRAPA consistiam de informações das plantas daninhas, do solo e do clima. Cada um destes conjuntos de dados era composto por diversas planilhas EXCEL. De modo que foi necessário – primeiro – organizar cada um destes *datasets* individualmente, para então montar a base de dados unificada.

4.1.1 Base de dados das plantas daninhas

Com relação às plantas daninhas, tem-se que as planilhas referiam-se aos dados coletados das espécies, o que ocorria em determinadas épocas dos anos. Assim, os arquivos enviados são das coletas feitas entre 2015 e 2023. Ademais, foram extraídas de um artigo da EMBRAPA as informações das coletas do ano de 2006 (GAMA *et al.*, 2007). Vale ressaltar, que há anos em que só foram realizadas uma amostragem – portanto há apenas um arquivo para este período –, enquanto em outros foram feitas mais de uma amostra, gerando, assim, mais de uma planilha para aquele ano.

Assim, todas essas planilhas de informações das plantas daninhas foram, primeiro, compiladas em um único arquivo. As variáveis dessa base de dados são: “Data”, “Invasora”, “Tipo de folha”, “Quantidade”, “Peso verde (g)”, “Peso seco (g)”, “Coleta da Amostra”, “Plantação”, “Pasto”, “Número de Quadros”, “Área do Quadro (m²)” e, “Área Total (m²)”. A Tabela 1 exibe um

recorte da base de dados das plantas daninhas, de modo que é possível verificar algumas das variáveis deste conjunto de dados.

Tabela 1: Recorte da base de dados das plantas daninhas.

Data	Invasora	Tipo de Folha	Quantidade	Peso Verde	Peso Seco	Coleta da Amostra	Plantação
01/10/2006	mentrasto	larga	1	0,985	0,394	na lavoura	soja
01/10/2006	caruru	larga	4	7,471	2,988	na lavoura	soja
01/10/2006	picão	larga	2	1,625	0,65	na lavoura	soja

Fonte: Elaborada pela autora.

A variável “Data” refere-se à quando foi feita aquela coleta; “Invasora” é o nome comum da espécie amostrada; “Tipo de folha” concerne à morfologia da folha daquela planta daninha, ou seja, se é de folha larga ou folha estreita; “Quantidade” é o número de plantas coletadas daquela espécie, naquele dia. Depois disso essas plantas são pesadas e, assim, é aferido o seu “Peso Verde”. Na sequência as plantas passam por um processo de desidratação e são pesadas novamente, obtendo, dessa forma, o “Peso Seco”. A “Coleta da Amostra” diz respeito à epóca em que foi feita a coleta; enquanto a “Plantação” refere-se a cultura que estava no local da amostragem.

O sistema ILP de onde os dados são provenientes contém 4 campos de plantações, de modo que a variável “Pasto” é a numeração destes campos. Em cada pasto são delimitados vários quadros de coleta (“Número de Quadros”), cada quadro tem uma área específica (“Área do Quadro”) de forma que as áreas destes quadros somadas dão a área total amostrada (“Área Total”).

4.1.2 Base de dados do solo

Quanto às amostras do solo, tem-se que essas foram feitas uma vez por ano. Sendo assim, há somente um arquivo para cada ano amostrado. Porém, estas amostragens de solo não foram realizadas exatamente no mesmo período que o das plantas daninhas. Portanto, os dados do solo são

referentes aos anos de 2005, 2006, 2012, 2014, 2015, 2016, 2017, 2018, 2019, 2020 e 2022. Todas estas planilhas foram compiladas em um único arquivo.

As variáveis deste conjunto de dados são: “Ano”, “Pasto”, “Profundidade (cm)”, “pH (H_2O)”, “Hidrogênio mais Alumínio: H+Al (cmolc/dm³)”, “Alumínio: Al (cmolc/dm³)”, “Cálcio: Ca (cmolc/dm³)”, “Magnésio: Mg (cmolc/dm³)”, “Potássio: K (mg/dm³)”, “Fósforo: P (mg/dm³)”, “Matéria Orgânica: MO (dag/kg)”, “Soma de Bases: SB (dag/kg)”, “Capacidade de Troca de Cátions: CTC (cmolc/dm³)”, “Saturação por Bases: V (%)”, “Saturação de Alumínio: Sat.Al (%)”, “Boro: B (mg/dm³)”, “Zinco: Zn (mg/dm³)”, “Ferro: Fe (mg/dm³)”, “Cobre: Cu (mg/dm³)”, “Manganês: Mn (mg/dm³)”, “Cálcio por Magnésio: Ca/Mg (cmolc/dm³)”, “Cálcio por Potássio: Ca/K (cmolc/dm³)” e, “Cálcio mais Magnésio por Potássio: Ca+Mg/K (cmolc/dm³)”.

A Tabela 2 apresenta um recorte da base de dados do solo, em que é possível identificar algumas das variáveis citadas acima. Não foi possível fazer um recorte em que mostrasse todas as variáveis pois, como são muitas colunas, a qualidade da figura ficaria muito baixa.

Tabela 2: Recorte da base de dados do solo.

Ano	Pasto	Profundidade	pH (H20)	H+Al	Al	Ca	Mg	K	P
2005	1	0-20	5,4	6,01	0,20	4,49	0,72	140	20,00
2005	1	20-40	5,4	6,30	0,30	1,06	0,61	89	14,00
2005	2	0-20	5,2	6,00	0,55	3,50	0,48	92	39,00
2005	2	20-40	5,4	6,00	0,30	3,64	0,44	52	11,00

Fonte: Elaborada pela autora.

Assim como na base das plantas daninhas, tem-se que a variável “Pasto” refere-se à numeração do campo. Já a “Profundidade” diz respeito ao ponto de extração da amostra, enquanto “pH (H_2O)” é o pH da água do solo. A saturação por bases é oriunda da razão entre a soma de bases e a capacidade de troca de cátions. As demais variáveis são elementos químicos encontrados no solo da região.

4.1.3 Base de dados do clima

Com relação aos dados do clima, estes são provenientes de duas fontes diferentes: de 2000 a 2016 as informações são da Estação Convencional de Sete Lagoas, e de 2017 em diante são do Instituto Nacional de Meteorologia (INMET), o qual é uma estação automática. Os dados da estação convencional estavam todos agrupados em uma mesma planilha, enquanto os do INMET estavam em planilhas separadas para cada ano. Assim, todos estes arquivos foram compilados em um único conjunto de dados.

Para poder unificar os dados das duas fontes em uma só base, foi necessário selecionar as variáveis comuns a ambas, as quais são: “Ano”, “Mês”, “Dia”, “Pressão (hPa)”, “Temperatura Média (ºC)”, “Temperatura Máxima (ºC)”, “Temperatura Mínima (ºC)”, “Umidade Relativa (%)”, “Velocidade do Vento (m/s)”, “Direção do Vento (°)”, “Precipitação (mm)”. Dentre estas, vale apenas destacar que a “Direção do Vento” está em graus de orientação da bússola. A Tabela 3 mostra as variáveis em comum selecionadas das bases da EMBRAPA e do INMET, as quais compõem a base de dados do clima.

Tabela 3: Recorte da base de dados do clima.

Ano	Mês	Dia	Pressão	Temp. Média	Temp. Máxima	Temp. Mínima	Umidade Relativa	Velocidade do Vento	Direção do Vento	Precipitação
2000	1	1	923	20,9	22,2	19,8	94	3,0666667	105	7,7
2000	1	2	923,166	21,53333	23,1	19,4	91	2,9	0	31,9
2000	1	3	923,133	22,5	26,9	19,2	84	2,5666667	120	7,5

Fonte: Elaborada pela autora.

4.2 PRÉ-PROCESSAMENTO DOS DADOS

Para cada um dos *datasets* desenvolvidos acima, foi necessário aplicar técnicas de pré-processamento de dados e realizar alguns ajustes na nomenclatura das variáveis. A seguir, são detalhados os procedimentos realizados em cada uma das bases e também na base unificada.

4.2.1 Base de dados das plantas daninhas

Em relação a esse conjunto de dados, tem-se que foi preciso padronizar os nomes das plantas daninhas, isso porque mesmas espécies possuem nomes comuns diferentes. Por exemplo, “vassoura” e “guanxuma” são a mesma espécie de planta daninha. Além disso, mesmas nomenclaturas apareciam escritas de maneira diferentes, tal como “joá de capote” e “joá-de-capote”. Assim, para que registros desse tipo não fossem considerados como informações diferentes, foi feita uma padronização dos nomes das plantas daninhas.

Para determinar os valores da variável “Coleta de Amostra”, o profissional da EMBRAPA que acompanha a pesquisa participou dessa etapa do projeto, já que as descrições das planilhas referentes a essa informação não estavam claras e foram preenchidas erroneamente na primeira vez que essa base foi elaborada.

Ademais, havia valores faltantes para as variáveis “Peso Verde” e “Peso Seco”. Para lidar com estes casos, os valores ausentes do “Peso Verde” foram preenchidos com a mediana (BRUCE *et al.*, 2020). Já o “Peso Seco” foi calculado a partir dos valores do “Peso Verde” – usando uma regra de três. Isso porque o “Peso Seco” é aproximadamente 40% do “Peso Verde” (ABDULLAH *et al*., 2020).

4.2.2 Base de dados do solo

Para os dados do solo foi realizada uma análise dos valores para identificar e remover *outliers* – tal como exemplificado na Figura 18. Além disso, os valores faltantes foram preenchidos com a mediana. (BRUCE *et al.*, 2020).

Figura 18: Valores dos registros de Zinco.

Fonte: Elaborada pela autora.

4.2.3 Base de dados do clima

Como mencionado anteriormente, para juntar todos os arquivos do clima, foram selecionadas as variáveis em comum das planilhas da EMBRAPA e do INMET. Uma destas era a "Direção do Vento", porém, nos arquivos da estação convencional os valores estavam escritos como pontos cardeais, enquanto nos arquivos da estação automática os valores estavam em graus. Desta forma, os registros que estavam como pontos cardeais foram traduzidos para seus respectivos valores em graus. Por exemplo, O= 270°, NO = 315°, e assim por diante. Também foram identificados e removidos os *outliers*, bem como foram preenchidos os valores faltantes com a mediana (BRUCE *et al.*, 2020).

4.2.4 Base de dados unificada

Primeiro, foi feito um *enconding* dos dados, transformando os valores de texto em numéricos, de modo que a base final ficasse adequada para os algoritmos de AM. Para poder juntar todos os conjuntos de dados em um só *dataset*, as bases individuais foram otimizadas.

Para o conjunto de dados das plantas daninhas, a quantidade de registros foi minimizada. Para fazer isso, os registros foram agrupados por pasto, por data e por espécie. Na sequência, esses registros de mesmas espécies (para um mesmo pasto e data) foram somados. Para a mesma data de coleta da espécie de planta daninha, foram selecionadas as colunas do clima desse mesmo dia, e estas foram concatenadas com as informações das plantas daninhas.

A quantidade de registros da base do solo também foi minimizada. Os registros foram agrupados por pasto, por ano e por profundidade. Em seguida, foram calculadas as médias de todos os valores para uma mesma profundidade (de um mesmo pasto e ano), para ficar um só registro. Por último, foi feita a junção “de todos para todos” da planilha clima/plantas-daninhas com os dados do solo.

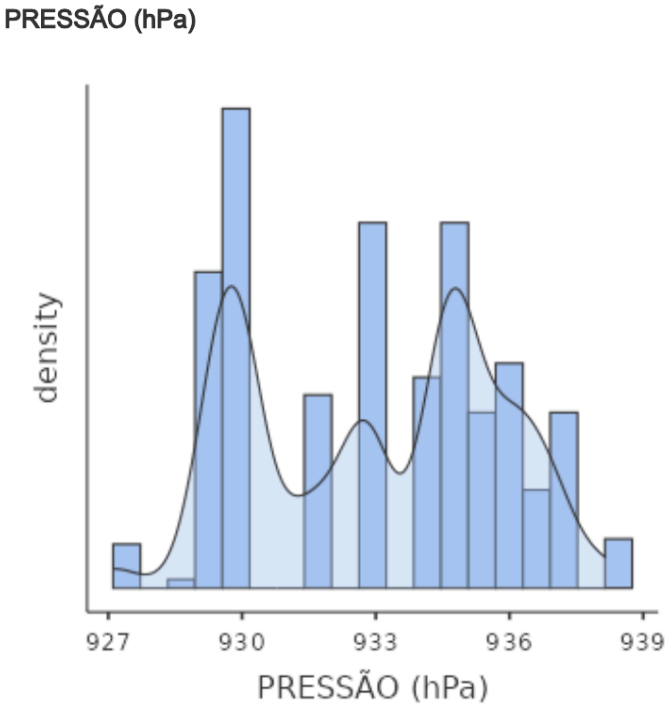
Vale ressaltar que devido aos registros do solo serem por ano, os dados da base clima/plantas-daninhas tiveram que ser agrupados por esta variável, a partir de uma coluna auxiliar que continha apenas a informação do ano, para que a junção final dos *datasets* pudesse ser feita; não obstante, a base unificada ainda continha o registro da data completa (ano, mês, dia).

Entretanto, como as amostragens de solo e de plantas daninhas não ocorreram exatamente nos mesmos anos, alguns dados acabaram sendo perdidos durante o processo de junção das bases. Portanto, para cada espécie coletada em um determinado dia, este registro acabou sendo repetido para os diferentes tipos de profundidades. No total, a base unificada contém 1541 registros.

4.3 ANÁLISE ESTATÍSTICA

Com a base de dados unificada pronta, foi possível fazer uma análise estatística sobre os dados para tentar observar algum padrão ou alguma relação entre as variáveis. Ademais, tem-se que tal etapa foi realizada para entender melhor os dados e – por consequência – ter uma melhor compreensão do que poderia ser feito nos algoritmos de aprendizado de máquina a partir de tais informações. Primeiro, foram plotados os histogramas de cada uma das variáveis numéricas para analisar a distribuição dos dados, conforme mostra a Figura 19.

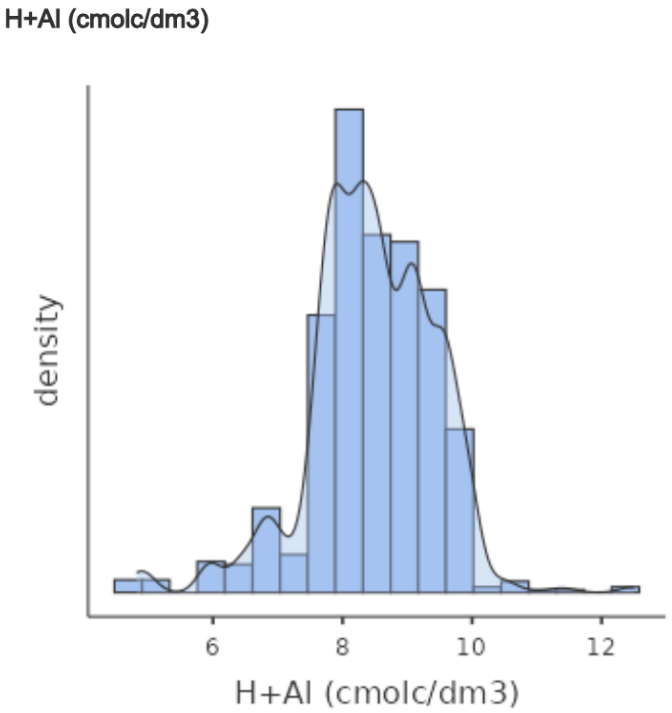
Figura 19: Distribuição dos dados da Pressão.



Fonte: Elaborada pela autora.

Apesar de algumas variáveis possuírem uma distribuição aproximadamente normal (Figura 20), a maioria delas não tinha os dados distribuídos normalmente.

Figura 20: Distribuição dos dados do H+Al (cmolc/dm³).



Fonte: Elaborada pela autora.

Logo, como os dados não apresentavam uma distribuição normal – na etapa seguinte –, foi optado por executar testes não-paramétricos. Em específico, foi feito o Teste de Kruskal-Wallis para averiguar se havia alguma diferença entre os valores de cada uma das variáveis numéricas para cada uma das variáveis nominais. Por exemplo, foi verificado se havia alguma diferença entre os valores da Umidade Relativa (%) para as diferentes espécies de plantas daninhas (Tabela 4).

Tabela 4: Teste de Kruskal-Wallis para a variável Umidade Relativa (%) em relação às espécies de plantas daninhas.

ANOVA a um fator (não-paramétrico) - Kruskal-Wallis			
	X^2	gl	p-valor
UMIDADE RELATIVA (%)	147	36	<.001

Fonte: Elaborada pela autora.

Caso o resultado do teste fosse menor que o nível de significância, isto é, menor que 5%, concluía-se que havia uma diferença entre os valores da variável numérica para a variável nominal analisada. Na sequência, foram calculadas as Comparações Múltiplas Dwass-Steel-Critchlow-Fligner (DSCF) para descobrir quais eram os casos que apresentavam diferença. Assim como no Teste de Kruskal-Wallis, caso o p-valor fosse menor que 5% é possível afirmar que há uma diferença entre os casos, em relação a variável investigada. A Tabela 5 mostra algumas comparações DSCF referentes a Umidade Relativa (%) para algumas espécies de plantas daninhas.

Tabela 5: Comparações DSCF referentes a Umidade Relativa (%) para as diferentes espécies de plantas daninhas.

Comparações múltiplas - UMIDADE RELATIVA (%)			
		W	p-valor
pé-de-galinha	vassoura rabo de tatu	-5,305	0,07
serralha	soja perene	-1,3494	1
serralha	sorgo selvagem	2,5437	0,999
serralha	tiririca	-1,4876	1
serralha	trapoeraba	1,3352	1
serralha	vassoura de bruxa	-2,2648	1
serralha	vassoura rabo de tatu	-5,7343	0,024
soja perene	sorgo selvagem	2,3206	1
soja perene	tiririca	0,4246	1
soja perene	trapoeraba	2,8229	0,993
soja perene	vassoura de bruxa	-2,6458	0,998
soja perene	vassoura rabo de tatu	-3,385	0,914
sorgo selvagem	tiririca	-3,6936	0,795
sorgo selvagem	trapoeraba	-2,0372	1
sorgo selvagem	vassoura de bruxa	-3,61	0,834
sorgo selvagem	vassoura rabo de tatu	-7,1624	<0,001

Fonte: Elaborada pela autora.

Seguindo, foi feita uma análise da estatística descritiva, isto é, foram analisados os valores das medidas de tendência central – média, moda e mediana – e também das medidas de dispersão, como o desvio padrão e a variância. Além disso, também foram averiguados o mínimo, o máximo e a quantidade de registros. Vale ressaltar que tais parâmetros foram calculados para cada variável numérica em função de uma variável nominal. A Tabela 6 ilustra um exemplo das estatísticas descritivas realizadas.

Tabela 6: Estatística descritiva para o Peso Verde (g) referente à Plantação.

Estatística Descritiva									
	PLANTAÇÃO	N	Média	Mediana	Moda	Desvio-padrão	Variância	Mínimo	Máximo
PESO VERDE (g)	milho	251	127,916	6,075	1	875,066	765740,625	0,368	7927,82
	milho/sorgo	222	13,016	2,763	1	20,787	432,079	0	85,22
	pasto	132	19,501	2,1	1	40,579	1646,64	0	168,53
	pasto/soja	28	0,143	0	0	0,356	0,127	0	1
	soja	251	157,186	10,175	1	827,212	684280,445	0	6890,06
	soja/milho	468	22,381	11,917	0,55	37,273	1389,277	0,2	265,71
	sorgo	144	4,803	1,188	1	10,513	110,526	0,125	50,88
	sorgo/pasto	45	10,88	0,85	0	19,521	381,069	0	62,6

Fonte: Elaborada pela autora.

Tal processo foi efetuado para tentar extrair informações sobre os dados, e também para tentar verificar a direção das diferenças encontradas nas comparações DSCF, já que estas são calculadas em termos da mediana e/ou da distribuição da população.

Também foram elaboradas matrizes de correlação, bem como um mapa de calor, para identificar possíveis correlações entre as variáveis. Por causa dos dados não serem normalmente

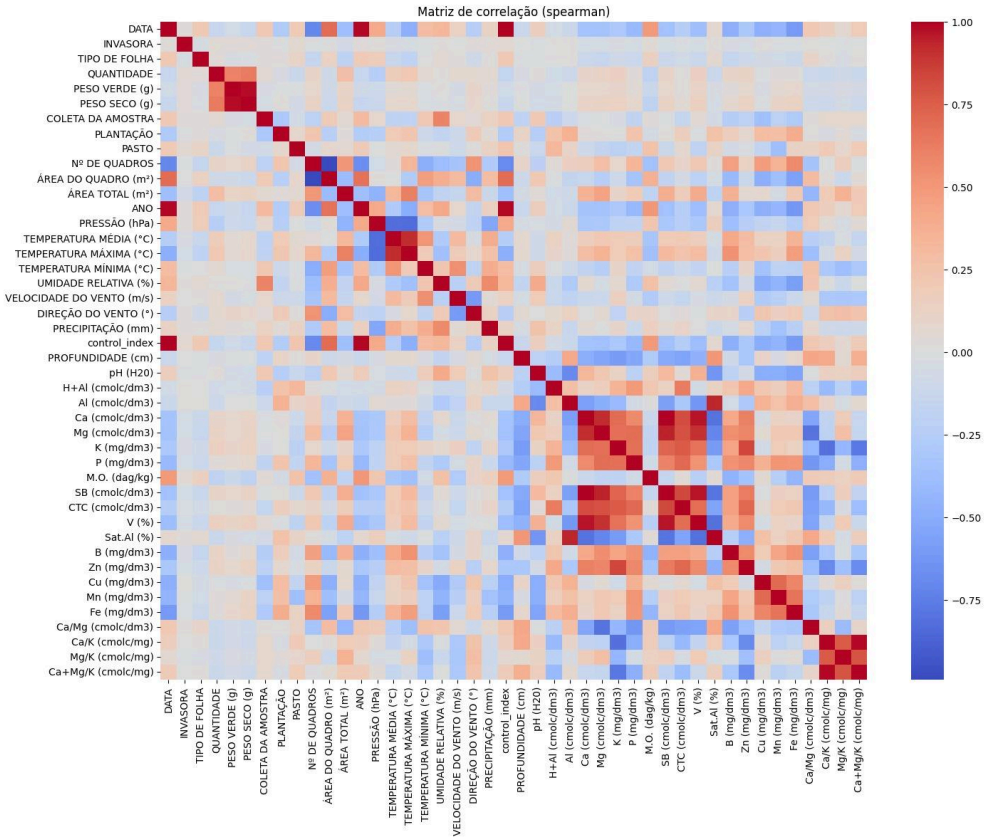
distribuídos, foi utilizado o coeficiente de Spearman para calcular as correlações. A Tabela 7 e a Figura 21 exibem um exemplo de matriz de correlação e o mapa de calor, respectivamente.

Tabela 7: Matriz de correlação entre a Coleta da Amostra e as variáveis de plantio.

Matriz de Correlação								
		COLETA DA AMOSTRA	PESO VERDE (g)	PESO SECO (g)	PASTO	N° DE QUADROS	ÁREA DO QUADRO (m²)	ÁREA TOTAL (m²)
COLETA DA AMOSTRA	Rho de Spearman	-						
	gl	-						
	p-valor	-						
PESO VERDE (g)	Rho de Spearman	-0,031	-					
	gl	1539	-					
	p-valor	0,229	-					
PESO SECO (g)	Rho de Spearman	-0,05	0,984	-				
	gl	1539	1539	-				
	p-valor	0,052	<0,001	-				
PASTO	Rho de Spearman	-0,067	-0,014	-0,064	-			
	gl	1539	1539	1539	-			
	p-valor	0,009	0,577	0,012	-			

Fonte: Elaborado pela autora.

Figura 21: Mapa de calor.



Fonte: Elaborado pela autora.

4.4 ALGORITMOS DE AM

Para implementação dos algoritmos de aprendizado de máquina foi definida a linguagem de programação *Python*. Para leitura e manipulação do *dataset* foi utilizado a biblioteca *Pandas*, e para o treinamento dos modelos de AM também foi utilizado a biblioteca *Scikit-Learn*. Foram implementados nos algoritmos diferentes tipos de predições, como tentativas para verificar qual seria melhor abordagem do problema. As predições de quantidade, tipo de folha e de cultura foram desenvolvidas antes da análise estatística; enquanto o modelo de predição da época de amostragem foi feito depois. Todas estas estão descritas na sequência.

4.4.1 Modelo de predição de quantidade de plantas daninhas

Nesta tentativa, o objetivo era prever a quantidade de plantas daninhas com base em intervalos. Isto é, ao invés de prever a quantidade exata, o modelo tentaria prever o intervalo da possível quantidade de plantas daninhas. Os intervalos definidos foram: de 0 a 25, 25 a 50, 50 a 75, 75 a 100 e mais de 100. Tal escolha foi feita, pois os modelos estavam tendo um desempenho ruim para a predição da quantidade exata. Os algoritmos selecionados – conforme citados no Capítulo 2 – foram Árvore de Decisão, Floresta Randômica e SVM. Os resultados obtidos estão apresentados na tabela abaixo:

Tabela 8: Métricas dos modelos para a predição da quantidade de plantas daninhas.

Algoritmo	Acurácia (%)	Precisão (%)	Sensibilidade (%)
Árvore de Decisão	88	87,5	93,75
Floresta Randômica	62	41,75	37,5
SVM	75	41,75	50

Fonte: Elaborada pela autora.

Tem-se que para a Floresta Randômica os valores da Precisão são a média de todos os intervalos utilizados na predição, o mesmo vale para a Sensibilidade. Para o SVM tanto o *kernel* polinomial quanto o *radial basis function* (RBF) alcançaram 75% de acurácia. Os valores de Precisão e Sensibilidade referem-se à média de todos os intervalos, para ambos os tipos de *kernel*. Já a Árvore de Decisão foi o algoritmo com melhores resultados, neste caso a Precisão e a Sensibilidade também se referem à média dos valores dos intervalos. Entretanto, esta era uma árvore viciada – isto é, foi encontrada a melhor árvore e esta foi reaplicada até chegar nas métricas apresentadas acima – assim, seus resultados tiveram de ser desconsiderados, já que este procedimento foi feito incorretamente.

Os modelos desenvolvidos nesta tentativa apresentaram resultados razoáveis mas não excelentes. Além disso, a predição da quantidade de plantas daninhas não foi considerada como relevante pela EMBRAPA. Assim, este experimento foi desconsiderado. Também vale ressaltar que quando estes modelos foram implementados, a base de dados ainda não contava com as informações do solo; o *dataset* do clima estava incompleto; e a base das plantas daninhas estava com diversos erros. Tudo isso pode ter influenciado os resultados apresentados pelos algoritmos.

4.4.2 Modelo de predição do tipo de folha

Depois de refazer as bases de dados e realizar o pré-processamento destas, foi feita uma nova tentativa de predição. A ideia original era tentar prever as espécies de plantas daninhas, porém, como havia muitas possibilidades de resposta – afinal são mais de 60 espećies – os algoritmos acabaram tendo resultados muito ruins: menos que 5% de acurácia. Desta forma, para ainda fazer uma predição sobre os tipos de plantas daninhas, foi optado por tentar prever as espécies de acordo com a morfologia da folha.

Os algoritmos selecionados foram Árvore de Decisão, Floresta Randômica, SVM e KNN. Para melhorar a performance dos modelos foi feito um embaralhamento dos dados, de modo a obter o melhor resultado possível. Entretanto, durante o processo percebeu-se que, como o nome das plantas daninhas era uma variável de entrada, não fazia sentido a saída ser o tipo de folha. Afinal, no momento em que determinada espécie foi coletada em campo, apenas com a sua identificação já seria possível determinar o tipo de folha, sem a necessidade de um algoritmo para prever tal informação.

Apesar de não ter sido registrado os melhores resultados obtidos pelos algoritmos, tem-se que as métricas obtiveram bons valores. Também foi feita uma tentativa de prever o tipo de folha, sem ter o nome da planta daninha como um dado de entrada. Todavia, desta maneira os modelos estavam sempre atingindo 100% de acurácia, o que não fazia sentido. Assim, por não conseguir identificar o que estava causando a máxima acurácia, esse experimento foi desconsiderado.

4.4.3 Modelo de predição da cultura

Na sequência, foi feita uma tentativa para prever a cultura na qual determinadas plantas daninhas iriam emergir, tendo como base as informações de solo, clima e plantio. Os algoritmos selecionados foram Árvore de Decisão, Floresta Randômica, SVM e KNN. Novamente foi feito o embaralhamento dos dados, para aperfeiçoar o processo de aprendizado dos modelos. Os melhores resultados obtidos pelos algoritmos foram salvos e estão descritos na tabela a seguir:

Tabela 9: Métricas dos modelos para a predição da cultura.

Algoritmo	Acurácia (%)	Precisão (%)	Sensibilidade (%)	F1 Score (%)
Árvore de Decisão	99	98,62	99,12	97,64
Floresta Randômica	99	97	96,62	99,08
SVM	98	98,5	95,5	91,83
KNN	95	95,25	86,75	87,47

Fonte: Elaborada pela autora.

No caso do SVM tem-se que as métricas referem-se ao *kernel* linear, já que este foi o que apresentou o melhor desempenho. Além disso, as previsões feitas por cada algoritmo foram salvas em um arquivo de texto para poderem ser analisadas. Junto com estas também foram salvas as *feature importance* dos modelos (Árvore de Decisão de Floresta Randômica), para verificar quais variáveis mais impactaram nas escolhas feitas por estes. Tem-se que todos os modelos performaram de maneira satisfatória, sendo que a Árvore de Decisão foi a que teve o melhor desempenho.

4.4.4 Modelo de predição da época de amostragem

Após a análise estatística percebeu-se que não haviam muitos fatores que diferenciavam as espécies de plantas daninhas, sendo assim, foi optado por desenvolver um modelo de predição da época de amostragem. Os algoritmos utilizados foram KNN e SVM. Tem-se que a Árvore de Decisão e a Floresta Randômica foram descartados, pois sua acurácia estava sempre dando 100%. Vale ressaltar, que nesta tentativa também foi feito o embaralhamento dos dados. A tabela a seguir descreve o desempenho obtido pelos algoritmos.

Tabela 10: Métricas dos modelos para a predição da época de amostragem.

Algoritmo	Acurácia (%)	Precisão (%)	Sensibilidade (%)	F1 Score (%)
SVM	98,92	99,04	99,19	99,11
KNN	99,78	99,8	99,85	99,82

Fonte: Elaborada pela autora.

Os resultados apresentados para o SVM são referentes ao *kernel* linear, o qual teve o melhor desempenho. Novamente, as predições feitas por cada algoritmo foram salvas em um arquivo de texto para, na sequência, serem analisadas. Ambos os modelos tiveram uma boa performance, sendo que o KNN foi o melhor entre os dois.

4.5 CONSIDERAÇÕES

Neste capítulo foram apresentadas todas as etapas realizadas para o desenvolvimento do projeto: iniciando pela elaboração das bases de dados da EMBRAPA, pré-processamento dos dados, criação da base unificada, execução das análises estatísticas, e desenvolvimento dos modelos de aprendizado de máquina.

A elaboração das bases individuais e os tratamentos dos dados destas permitiram a construção da base de dados unificada e as demais etapas. Os modelos desenvolvidos foram

desenvolvidos com base nos algoritmos apresentados no Capítulo 2, bem como sua avaliação foi feita segundo as métricas identificadas na literatura.

5 RESULTADOS

A proposta deste trabalho é aplicar modelos de aprendizado de máquina e técnicas de análise de dados no manejo de plantas daninhas em sistemas ILP. O objetivo deste capítulo é apresentar os resultados obtidos em cada algoritmo e nas análises estatísticas. Ademais, tem-se que o intuito também é avaliar as contribuições deste trabalho.

O capítulo está estruturado como segue: a Seção 5.1 exibe as conclusões obtidas na fase de análise estatística; e a Seção 5.2 traz os resultados atingidos pelos modelos de predição de cultura e de época de amostragem, já que ambos tiveram os melhores desempenhos e foram considerados como pertinentes pela EMBRAPA. Por último, a seção 5.3 apresenta as considerações sobre essa etapa do trabalho.

5.1 RESULTADOS DA ANÁLISE ESTATÍSTICA

A seguir são apresentados individualmente os resultados de cada etapa da análise estatística realizada. Vale ressaltar que toda esta parte do projeto foi feita utilizando a plataforma Jamovi.

5.1.1 Resultados do Teste de Kruskal-Wallis e das Comparações DSCF

Tem-se que o Teste de Kruskal-Wallis e as Comparações DSCF foram aplicados para seguintes variáveis nominais: Invasora; Tipo de Folha; Coleta da Amostra; Plantação; e Profundidade. Todas estas tiveram seus valores comparados em relação às seguintes variáveis numéricas: Peso Verde; Peso Seco; Pasto; Pressão; Temperatura Média; Temperatura Máxima; Temperatura Mínima; Umidade Relativa; Velocidade do Vento; Direção do Vento; Precipitação; pH; H+Al; Al; Ca; M; K; P; MO; SB; CTC; V; Sat.Al; B; Zn; Fe; Cu; Mn; Ca/Mg; Ca/K e; Ca+Mg/K. Vale destacar que as Comparações DSCF mostram em quais casos há diferença, porém, não indicam qual é a direção desta.

No que se refere às plantas daninhas (Invasoras), tem-se que para as variáveis de plantio o Teste de Kruskal-Wallis apontou que há diferença nos valores destas. Para o Peso Verde e o Peso Seco faz sentido que haja diferença nos valores entre as diferentes espécies, afinal como cada espécie de planta daninha é distinta das outras, os pesos vão ser diferentes por consequência. Já as diferenças encontradas em relação ao Pasto, podem indicar quais espécies de plantas daninhas não aparecem numa mesma área.

Quanto aos fatores de clima, tem-se que as espécies losna branca e erva quente foram as que mais apresentaram diferença para as outras plantas daninhas em relação à Pressão, Temperatura Média e Temperatura Máxima. Além disso, as Comparações DSCF também mostraram que: para a Umidade Relativa a vassoura rabo de tatu foi a que mais teve diferenças; o capim braquiária foi a espécie com mais desigualdades de valores no que tange a Velocidade do Vento; e ambas espécies apresentaram valores distintos para com as outras em relação à Direção do Vento.

Todavia, para os valores da Temperatura Mínima e da Precipitação, não foram encontrados muitos casos de diferenças entre as plantas daninhas. Indicando que todas as espécies apresentam valores muito semelhantes para tais elementos. No entanto, isso não faz sentido para a Precipitação, pois – como informado pelo profissional da EMBRAPA – a chuva é um elemento que interfere no aparecimento de certas plantas daninhas, pelo o que é observado na prática.

Por último, para os elementos de solo tem-se que para as variáveis H+Al e Ca/Mg foram encontradas poucas diferenças, de modo que a maioria das espécies se beneficia de níveis semelhantes destes componentes. Já para o Ca+Mg/K, Mg/K, e Ca/K o Teste de Kruskal-Wallis apontou que os valores destas variáveis são distintos entre as diferentes espécies de plantas daninhas, porém, não observou-se nenhuma diferença nas comparações do teste DSCF. Isso pode ter acontecido devido ao Teste de Kruskal-Wallis ser calculado primeiro. Assim, ao determinar as comparações, o programa pode ter feito uma análise mais profunda e visto que – na verdade – não haviam diferenças.

Para as demais variáveis de solo, algumas espécies de plantas daninhas apresentaram diferença nos seus valores em relação às outras. Os casos encontrados não serão todos descritos, por

causa da quantidade de variáveis e espécies a se citar. Contudo, merece destaque a losna branca, que foi a planta daninha com mais diferenças em relação às outras para grande parte dos fatores de solo. A vassoura rabo de tatu também teve uma presença considerável nas Comparações DSCF.

Entretanto, os casos que mais chamam a atenção são referentes às variáveis B e Fe. Tem-se que para estes dois elementos foram encontradas muitas diferenças entre a maioria das espécies e os casos em que não havia registro de plantas daninhas (indicados por "-----"). Isso indica que tais componentes estão diretamente relacionados com o aparecimento de plantas daninhas. As Tabela 11 e 12 demonstram isso.

Tabela 11: Comparações DSCF entre as plantas daninhas em relação ao B.

Comparações múltiplas - B (mg/dm3)			
		W	p-valor
-	apaga-fogo	-5,7911	0,02
-	assa-peixe	-2,5776	0,999
-	beldroega	-6,4146	0,003
-	buva	-6,5969	0,002
-	capim amargoso	0,6754	1
-	capim braquiária	-3,9173	0,676
-	capim carrapicho	-2,9548	0,986
-	capim colchão	-3,0119	0,981
-	caím guiné	-1,632	1
-	capim marmelada	-2,121	1
-	capim mombaça	-1,4166	1
-	carrapicho de carneiro	-4,1874	0,515
-	caruru	-6,7098	0,001
-	corda de viola	-5,6241	0,032
-	cordão de frade	-7,6417	<0,001
-	erva de santa luzia	-5,554	0,038

-	erva de touro	-5,957	0,013
-	erva moura	-5,1863	0,092
-	erva quente	-5,6949	0,026
-	fedegoso	-6,7035	0,001
-	guanxuma	-6,6667	0,001
-	joá de capote	-0,3841	1
-	leiteiro	-6,0235	0,011
-	losna branca	-8,335	<0,001
-	mamona	-5,1931	0,091
-	mentrasto	-6,9936	<0,001
-	picão	-6,1168	0,008
-	poaia	-6,5902	0,002
-	pé-de-galinha	-2,4392	1
-	serralha	-6,1011	0,008
-	soja perene	-2,7103	0,997
-	sorgo selvagem	-5,2682	0,076
-	tiririca	-0,0672	1
-	trapoeraba	-6,5415	0,002

Fonte: Elaborada pela autora.

Tabela 12: Comparações DSCF entre as plantas daninhas em relação ao Fe.

Comparações múltiplas - Fe (mg/dm3)			
		W	p-valor
-	apaga-fogo	-6,9132	<0,001
-	assa-peixe	-3,2672	0,944
-	beldroega	-5,9392	0,014
-	buva	-6,9096	<0,001
-	capim amargoso	-0,8389	1

-	capim braquiária	-1,9249	1
-	capim carrapicho	-3,1785	0,96
-	capim colchão	-4,5874	0,294
-	caím guiné	-4,0956	0,57
-	capim marmelada	-2,3179	1
-	capim mombaça	-0,9356	1
-	carrapicho de carneiro	-5,0873	0,114
-	caruru	-8,6541	<0,001
-	corda de viola	-7,1795	<0,001
-	cordão de frade	-8,3661	<0,001
-	erva de santa luzia	-5,0531	0,123
-	erva de touro	-6,1091	0,008
-	erva moura	-7,9869	<0,001
-	erva quente	-5,374	0,06
-	fedegoso	-8,2357	<0,001
-	guanxuma	-9,2059	<0,001
-	joá de capote	-1,7661	1
-	leiteiro	-6,254	0,005
-	losna branca	-6,6348	0,002
-	mamona	-4,4281	0,375
-	mentrasto	-9,196	<0,001
-	picão	-7,7111	<0,001
-	poaia	-7,632	<0,001
-	pé-de-galinha	-3,9269	0,67
-	serralha	-6,1053	0,008
-	soja perene	0,4881	1
-	sorgo selvagem	-7,534	<0,001
-	tiririca	-1,4902	1

-	trapoeraba	-8,3613	<0,001

Fonte: Elaborado pela autora.

No que concerne ao Tipo de Folha, tem-se que foram encontradas diferenças sobre os Pesos (Verde e Seco) tanto entre as de folha larga e folha estreita, quanto entre os casos em que não havia plantas daninhas com os outros dois tipos. Tal informação é condizente, afinal, quando não há plantas daninhas o peso é zero e quando há alguma espécies o peso é diferente de zero, gerando – assim – uma diferença. Entre os pesos das plantas daninhas de folha estreita e folha larga também é evidente que haverá diferença nos valores, pois como a morfologia da folha é diferente isso acaba influenciando o peso da planta.

Para as variáveis de clima tem-se que a maioria das diferenças encontradas foram entre as plantas daninhas de folha larga e os registros em que não havia nenhuma espécie. Isso pode indicar quais fatores climáticos favorecem ou não as plantas daninhas de folha larga. Ademais, a Pressão, a Temperatura Média, a Temperatura Máxima e a Direção do Vento apresentaram diferença de valores entre os tipos folha larga e folha estreita – as Tabelas 13, 14 e 15 demonstram alguns destes exemplos. Podendo ser um indício de quais elementos influenciam no aparecimento de espécies de um tipo ou de outro.

Tabela 13: Comparações DSCF entre os tipos de folha em relação à Pressão.

Comparações múltiplas - PRESSÃO(hPa)			
		W	p-valor
-	estreita	2,92	0,0097
-	larga	5,31	<0,001
estreita	larga	8,57	<0,001

Fonte: Elaborada pela autora.

Tabela 14: Comparações DSCF entre os tipos de folha em relação à Temperatura Média.

Comparações múltiplas - TEMPERATURA MÉDIA (°C)			
		W	p-valor
-	estreita	-2,03	0,325
-	larga	-4,79	0,002
estreita	larga	-6,32	<0,001

Fonte: Elaborada pela autora.

Tabela 15: Comparações DSCF entre os tipos de folha em relação à Temperatura Máxima.

Comparações múltiplas - TEMPERATURA MÁXIMA (°C)			
		W	p-valor
-	estreita	-4,38	0,006
-	larga	-6,63	<0,001
estreita	larga	-8,4	<0,001

Fonte: Elaborada pela autora.

Também vale ressaltar que não foram encontradas diferenças para a Precipitação – tanto no Teste de Kruskal-Wallis quanto nas Comparações DSCF. Novamente demonstrando que tal variável beneficia todas as espécies de maneira igual, o que não é condizente com as observações empíricas. Quanto às variáveis do solo, tem-se que as plantas daninhas de folha larga e folha estreita tiveram diferença de valores entre: Ca; Mg; K; P; MO; SB; V; Sat. Al; B; Zn; e Ca/Mg. De modo que tais elementos podem ser fatores determinantes no aparecimento de uma espécie ou de outra.

Em relação às épocas de amostragem (Coleta da Amostra), há diferença entre os Pesos Verde e Seco para as diversas épocas. Tal fato pode indicar que há uma diferença entre as plantas daninhas encontradas em cada período – seja na quantidade ou nas espécies –, de modo a influenciar nos valores dos pesos. Para as variáveis de clima, todas as comparaçẽos feitas entre as épocas de amostragem tiveram diferenças. Afinal, como as épocas de amostragem ocorrem em

momentos distintos do ano, e cada estação do ano tem características climáticas próprias, os períodos de coleta também têm diferença nos fatores de clima, por consequência.

Já para os elementos do solo, as épocas de amostragem que mais apresentaram diferenças foram Entre-Safra e Na-Lavoura; tanto no geral quanto quando comparadas entre si. Tais comparações podem mostrar as alterações que ocorrem no solo para cada período. As Tabelas 16 e 17 exibem algumas das Comparações DSCF entre as épocas de amostragem.

Tabela 16: Comparações DSCF entre as épocas de amostragem em relação ao pH.

Comparações múltiplas - pH (H2O)			
		W	p-valor
entre-safra	na colheita	2,56	0,167
entre-safra	na lavoura	13,66	<0,001
na colheita	na lavoura	10,55	<0,001

Fonte: Elaborada pela autora.

Tabela 17: Comparações DSCF entre as épocas de amostragem em relação ao H+Al.

Comparações múltiplas - H+Al (cmol/dm3)			
		W	p-valor
entre-safra	na colheita	-7,37	<0,001
entre-safra	na lavoura	-11,92	<0,001
na colheita	na lavoura	-6,91	<0,001

Fonte: Elaborada pela autora.

As diferenças encontradas entre os valores do Peso Verde e do Peso Seco das Plantações apontam que há uma diferença entre as plantas daninhas encontradas em cada cultura, sendo que tais diferenças podem ser em relação à quantidade ou às espécies; já que ambos os casos influenciam no peso. As Comparações DSCF feitas neste caso podem ser observadas na Tabela 18.

Tabela 18: Comparações DSCF entre as culturas em relação ao Peso Verde.

Comparações múltiplas - PESO VERDE (g)			
		W	p-valor
milho	milho/sorgo	-3,921	0,102
milho	pasto	-4,051	0,08
milho	pasto/soja	-11,728	<0,001
milho	soja	4,535	0,029
milho	soja/milho	5,425	0,003
milho	sorgo	-9,787	<0,001
milho	sorgo/pasto	-5,534	0,002
milho/sorgo	pasto	-1,165	0,992
milho/sorgo	pasto/soja	-11,277	<0,001
milho/sorgo	soja	7,947	<0,001
milho/sorgo	soja/milho	9,65	<0,001
milho/sorgo	sorgo	-5,956	<0,001
milho/sorgo	sorgo/pasto	-4,181	0,062
pasto	pasto/soja	-10,333	<0,001
pasto	soja	6,878	<0,001
pasto	soja/milho	8,028	<0,001
pasto	sorgo	-3,543	0,193
pasto	sorgo/pasto	-3,399	0,24
pasto/soja	soja	11,831	<0,001
pasto/soja	soja/milho	12,279	<0,001
pasto/soja	sorgo	10,807	<0,001
pasto/soja	sorgo/pasto	7,93	<0,001
soja	soja/milho	-0,615	1
soja	sorgo	-12,095	<0,001
soja	sorgo/pasto	-6,951	<0,001

soja/milho	sorgo	-16,351	<0,001
soja/milho	sorgo/pasto	-7,015	<0,001
sorgo	sorgo/pasto	-0,976	0,997

Fonte: Elaborada pela autora.

Não serão descritas todas as diferenças encontradas entre as culturas para as variáveis de clima e solo, devido a quantidade de informações. Vale apenas destacar que o milho foi uma das culturas mais presentes nas diferenças de clima encontradas e que a soja foi a plantação que mais apresentou diferença em relação a Precipitação.

Por fim, o Teste de Kruskal-Wallis e as Comparações DSCF foram calculados tendo a Profundidade como variável nominal. Para os Pesos Verde e Seco, não foram encontradas nenhuma diferença para as diversas profundidades. Já os elementos de clima tiveram a maioria das diferenças entre as profundidades mais superficiais (de 0 - 5 cm e de 5 - 10 cm). Quanto às variáveis de solo, tem-se que para todas elas quase todas as comparações apresentaram diferenças. Os poucos casos em que não houve diferença nos valores foram entre profundidades bem próximas ou bem distantes. Isso indica que a quantidade dos elementos presentes no solo varia bastante de acordo com a profundidade. A Tabela 19 traz as Comparações DSCF em relação ao pH.

Tabela 19: Comparações DSCF entre as profundidades em relação ao pH.

Comparações múltiplas - pH (H2O)			
		W	p-valor
0-10	0-5	-0,776	1
0-10	10-20	2,575	0,669
0-10	15-20	-13,994	<0,001
0-10	20-40	6,03	<0,001
0-10	25-30	-10,607	<0,001
0-10	40-60	-1,545	0,976
0-10	45-50	-13,155	<0,001

0-10	5-10	-15,491	<0,001
0-5	10-20	1,958	0,904
0-5	15-20	-12,937	<0,001
0-5	20-40	5,591	0,003
0-5	25-30	-10,203	<0,001
0-5	40-60	-0,492	1
0-5	45-50	-12,542	<0,001
0-5	5-10	-14,615	<0,001
10-20	15-20	-14,322	<0,001
10-20	20-40	0,177	1
10-20	25-30	-10,115	<0,001
10-20	40-60	-4,401	0,049
10-20	45-50	-12,619	<0,001
10-20	5-10	-16,167	<0,001
15-20	20-40	14,133	<0,001
15-20	25-30	10,625	<0,001
15-20	40-60	14,281	<0,001
15-20	45-50	11,322	<0,001
15-20	5-10	10,467	<0,001
20-40	25-30	-9,863	<0,001
20-40	40-60	-5,987	<0,001
20-40	45-50	-11,599	<0,001
20-40	5-10	-15,743	<0,001
25-30	40-60	9,992	<0,001
25-30	45-50	8,529	<0,001
25-30	5-10	-0,767	1
40-60	45-50	-12,806	<0,001
40-60	5-10	-15,821	<0,001

45-50	5-10	1,486	0,981

Fonte: Elaborada pela autora.

5.1.2 Resultados da estatística descritiva

Na fase de estatística descritiva foram analisados os seguintes parâmetros das variáveis: quantidade de registros; média; moda; mediana; desvio padrão; variância; mínimo e máximo. Conforme mencionado anteriormente, todas estas métricas foram calculadas para as variáveis numéricas em função das variáveis nominais. Tais variáveis – tanto as numéricas bem como as nominais – são as mesmas que foram apresentadas na seção anterior.

Começando pelas plantas daninhas (Invasora), tem-se que para os fatores de plantio (Peso Verde e Peso Seco) algumas espécies – como a tiririca – apresentam desvio padrão e variância altos. Isso significa que os valores destas variáveis são bem diversificados, sugerindo que ao longo do tempo houve uma mudança na presença e/ou quantidade desta espécie e, por consequência, no peso também.

Além disso, tem-se que os valores da mediana foram analisados com mais atenção. Isso porque as Comparações DSCF são calculadas com base na mediana ou na distribuição dos dados. Logo, tem-se que – por meio da mediana – foi possível estabelecer algumas das direções das diferenças encontradas nas comparações.

As espécies losna branca e erva quente foram as que mais apresentaram diferença em relação a Temperatura Média e a Temperatura Máxima nas Comparações DSCF. Tem-se que os valores da mediana dessas espécies são menores que os das demais plantas daninhas. Indicando que tais espécies são favorecidas por temperaturas mais baixas. O mesmo ocorre para a espécie vassoura rabo de tatu em relação à Umidade: a mediana desta planta daninha é menor que a das outras, sugerindo que tal espécie é beneficiada por uma umidade menor.

Em relação a Velocidade do Vento, tem-se que a mediana da espécie capim braquiária é maior que as demais. Confirmando os resultados das Comparações DSCF, de que há uma diferença nos valores desta variável para tal planta daninha; e apontando a direção desta diferença. Quanto à Direção do Vento, tem-se que as diferenças encontradas nas Comparações DSCF e as diferenças de valores das medianas não são as mesmas. De maneira geral, os valores das medianas da maioria das espécies – incluindo a do capim braquiária e da vassoura rabo de tatu – são bem próximos. Porém, as espécies vassoura de bruxa e capim amargoso apresentam uma mediana bem maior que as demais. Para trabalhos futuros será necessário verificar se, neste caso, as diferenças estão sendo calculadas em termos da distribuição dos dados e não da mediana.

Já para a Temperatura Mínima, os valores da mediana de todas as espécies são bem próximos. Isso corrobora com o fato de que não foram encontradas muitas diferenças nas Comparações DSCF para esta variável. De modo que a Temperatura Mínima não é um fator que favorece uma espécie ou outra, mas que – no geral – tem uma influência igual para todas as plantas daninhas. O mesmo pode ser afirmado sobre a Precipitação: como a mediana é semelhante para todas as espécies, conclui-se que a influência da chuva é a mesma para todas elas. Ademais, como os valores dessa variável são bem baixos, poderia-se dizer que a falta de chuva ou pouca chuva é o que beneficia as plantas daninhas, porém, isso não condiz com o observado em prática.

No que se refere às variáveis de solo, tem-se que para todas elas as diferenças das medianas estavam de acordo com as diferenças encontradas nas Comparações DSCF. De modo que foi possível estabelecer a direção das diferenças. Entretanto, também houve muitos casos em que os valores da mediana eram consideravelmente maiores ou menores que os das outras plantas daninhas, mas estes não haviam aparecido nas Comparações Múltiplas. Isso pode ter ocorrido devido ao teste DSCF ter feito os cálculos com base na distribuição dos dados.

Com relação às variáveis nominais Coleta da Amostra, Tipo de Folha, Plantação e Profundidade, tem-se que muitas das diferenças encontradas nas Comparações DSCF não puderam ser confirmadas pelos valores das medianas. Nestes casos será necessário investigar a distribuição

dos dados – para verificar se de fato tais diferenças existem e a direção delas –, ou se há algum outro fator interferindo como: o tamanho da amostra ou a proporção entre as classes.

Também tem-se que a quantidade de registros de plantas daninhas de folha larga é bem maior que os de folha estreita (Tabela 20). Como o tipo de folha está diretamente associado às plantas daninhas, pode-se afirmar que as classes desta variável não estão desbalanceadas. Sendo assim, é possível concluir que há mais plantas daninhas de folha larga no sistema ILP estudado.

Tabela 20: Estatística descritiva do Tipo de Folha.

Estatística Descritiva									
	TIPO DE FOLHA	N	Média	Mediana	Moda	Desvio-padrão	Variância	Mínimo	Máximo
PESO VERDE (g)	-	42	0	0	0	0	0	0	0
	estreita	286	240,2	9,53	1	1118,1	1,25e+6	0,275	7928
	larga	1213	16,5	5,1	1	28,9	836	0,07	269

Fonte: Elaborada pela autora.

Além disso, a quantidade de registros da entre-safra (variável Coleta da Amostra) também é maior que a das demais épocas de amostragem (Tabela 21). Todavia, visto que há mais registros deste tipo do que os outros na base de dados, e por este fator não estar diretamente associado às plantas daninhas, pode-se afirmar que há um desbalanceamento das classes. Não sendo possível concluir se há de fato mais plantas daninhas na época da entre-safra.

Tabela 21: Estatística descritiva da Coleta da Amostra.

Estatística Descritiva									
	COLETA DA AMOSTRA	N	Média	Mediana	Moda	Desvio-padrão	Variância	Mínimo	Máximo
PESO VERDE	entre-safra	763	18,2	6,8	1	32,1	1032	0	266
	na colheita	496	19,1	4,39	1	36,8	1356	0	269

(g)	na lavoura	282	231,8	4,19	1	1127,3	1,27e+6	0,45	7928

Fonte: Elaborada pela autora.

5.1.3 Resultados das correlações

Para analisar se havia alguma correlação entre as variáveis, tem-se que foram utilizadas as ferramentas mapa de calor e matrizes de correlação – conforme mencionado anteriormente. Tem-se que no mapa de calor foram analisadas as correlações de "todas para todas" as variáveis. Já nas matrizes de correlação (elaboradas na plataforma Jamovi), foi verificado se havia alguma correlação entre uma variável nominal e as demais variáveis numéricas. Para que a matriz não ficasse muito grande, foram feitas matrizes separadas para cada um dos grupos das variáveis numéricas: plantio, clima e solo. Ademais, tem-se que as variáveis nominais selecionadas foram: Coleta da Amostra, Invasora, Plantação, Profundidade e Tipo de Folha.

Para a Coleta da Amostra e o Tipo de Folha, tem-se que os resultados obtidos nas matrizes de correlação e no mapa de calor nem sempre condiziam; ou as correlações atingidas no mapa de calor não eram tão significativas quanto apresentado pelas matrizes de correlação. Contudo, dentre as correlações que apareceram em ambas as análises, pode-se citar que a época de amostragem tem correlação com: Pressão; Temperatura Máxima; MO; V; B; e Ca/Mg. Quanto ao tipo de folha tem correlação com: Nº de Quadros; Área do Quadro; e Temperatura Mínima.

Para a Profundidade, também tem-se que as correlações apresentadas nas matrizes de correlação – para as variáveis de plantio e clima – eram bem mais significativas do que as encontradas no mapa de calor. Porém, para as variáveis de solo, os resultados estavam em concordância. Neste caso, tem-se que a profundidade possui correlação com: Al; Ca; SB; V; Ca/K; e Ca+Mg/K. Já a Plantação, teve grande parte dos resultados – de ambas as análises – em consonância. Assim, tem-se que as culturas apresentam correlação com: Nº de Quadros; Área do Quadro; Pressão; Temperatura Média; Temperatura Máxima; Direção do Vento; pH; H+Al; Al; P; Sat.Al; Cu; Mn; e Fe.

Por último, tem-se que para a Invasora os resultados das matrizes de correlação e do mapa de calor eram os mesmos: não há correlação entre as plantas daninhas e as demais variáveis de clima, solo e plantio. Tal informação parece ser contraditória com o observado em prática, afinal, são vários os fatores ambientais que influenciam no aparecimento das plantas daninhas. Algumas das causas que podem ter interferido nesta análise são a baixa quantidade de dados, de modo que não foi possível estabelecer algum padrão no comportamento das plantas daninhas.

Ou também o alto número de espécies investigadas, de tal forma que pode não haver fatores em comum para todas as plantas daninhas, mas sim que cada espécie tem correlação com variáveis diferentes. Todavia, é preciso considerar que em muitos casos – no Teste de Kruskal-Wallis e nas Comparações DSCF – não foram encontradas diferenças nos valores das variáveis dentre as diferentes espécies de plantas daninhas. Assim, é necessário para trabalhos futuros investigar a correlação de cada planta daninha com os demais elementos, para ver se de fato não há nenhuma correlação. Também poderia-se tentar aumentar a quantidade de dados e verificar novamente os métodos já utilizados para encontrar algum padrão.

5.2 RESULTADO DOS MODELOS DE PREDIÇÃO

Como mencionado anteriormente, tem-se que foram feitos diferentes tipos de predições para tentar identificar qual seria a melhor abordagem. Com base nas tentativas iniciais e nos resultados obtidos na análise estatística, foi possível concluir que os modelos de predição de cultura e de época de amostragem eram os mais adequados – considerando os objetivos do trabalho e as demandas da EMBRAPA. Além de serem os que apresentaram melhor desempenho nas métricas de avaliação.

5.2.1 Resultados do modelo de predição de cultura

Tem-se que as predições feitas por cada algoritmo foram salvas e analisadas. Logo, com base nos resultados dos quatro algoritmos utilizados, foi possível observar algumas relações entre as

espécies de plantas daninhas (entrada) e as culturas (saída). Vale destacar que tais análises foram feitas a partir do conjunto de dados de teste, isto é, com 30% dos dados.

Para grande parte das predições que tiveram como saída a soja, tem-se que as plantas daninhas de entrada eram caruru e trapoeraba. Indicando que tais espécies têm uma probabilidade maior de aparecer nesta cultura. O mesmo ocorreu para o milho. Já o pasto, teve a espécie pé-de-galinha como a principal planta daninha de entrada. Por último, tem-se que o mentrasto e o fedegoso foram as espécies que mais apareceram como entrada das predições de sorgo.

Para todos os registros que tinham como resultado pasto/soja, observou-se que – em todos os algoritmos – a espécie de entrada sempre era "---", ou seja, os casos em que não haviam sido coletadas plantas daninhas. Isto pode ser um indício de que as condições nesta cultura não são favoráveis para o aparecimento de plantas daninhas. Na soja/milho, as espécies buva e trapoeraba foram as que mais apareceram nos dados de entrada. Já o milho/sorgo e sorgo/pasto foram, em maioria, predições das espécies trapoeraba e vassoura rabo de tatu, respectivamente.

A partir destas análises, foi possível identificar em quais culturas determinadas plantas daninhas tem mais chance de aparecer. Além disso, também pôde-se perceber que algumas espécies não são propensas a se desenvolver em uma cultura específica, mas que podem surgir em qualquer plantação. Um exemplo disso é a espécie cordão de frade, a qual teve todas as culturas como seu resultado nas predições, em todos os algoritmos.

Outro fato notável, é que a trapoeraba foi uma das plantas daninhas que mais apareceu nas predições que tiveram como resultado: a soja, o milho e as rotações de entre-safra destas duas culturas. Isso indica que as condições geradas por estas culturas podem beneficiar essa espécie em particular. Com relação às *feature importance*, tem-se que as variáveis que mais importam no processo de escolha dos algoritmos Árvore de Decisão e Floresta Randômica foram: Coleta da Amostra e Ferro.

5.2.2 Resultados do modelo de predição de época de amostragem

A partir das previsões feitas por ambos os algoritmos – KNN e SVM –, pôde-se perceber em quais épocas (saída) certas espécies de plantas daninhas (entrada) iriam aparecer. Novamente é importante destacar que tais observações foram feitas com base no conjunto de teste, ou seja, com base em 30% dos dados. Também deve-se lembrar que há um desbalanceamento nas classes para a variável época de amostragem.

Tem-se que para o período da entre-safra, a planta daninha que mais apareceu como dado de entrada foi a poaia. Já na época na colheita, a caruru foi a espécie com maior presença. Por último, para as predições de na lavoura, o cordão de frade foi a planta daninha que mais apareceu como dado de entrada. Tudo isso indica que tais espécies têm mais chance de aparecer durante estes períodos.

Todavia, ainda com o intuito de encontrar algum padrão entre as espécies de plantas daninhas para com os fatores ambientais analisados, tem-se que foi feito um experimento investigativo usando um algoritmo de clusterização para tal tarefa. Tem-se que o algoritmo selecionado foi o K-*means*, com 31 *clusters*. No entanto, não foi possível identificar nenhuma semelhança entre as plantas daninhas a partir dos agrupamentos, pois em nenhum *cluster* houve uma espécie que apareceu em maior quantidade.

5.3 CONSIDERAÇÕES

Os modelos preditivos desenvolvidos e seus resultados demonstraram que é viável prever certos fatores relacionados às plantas daninhas, tais como as culturas e as épocas em que essas espécies podem se desenvolver. Apesar de serem poucos casos, foi possível determinar algumas relações entre certas espécies de plantas daninhas e alguns elementos ambientais. Contudo, os resultados obtidos nas tentativas de predições – inclusive na clusterização – e nas análises

estatística, indicam que a quantidade e qualidade dos dados, bem como o desbalanceamento entre certas classes, está afetando o desempenho dos algoritmos.

6 CONCLUSÕES

Este trabalho abordou o problema do manejo de plantas daninhas em sistemas ILP, e como o uso de aprendizado de máquina pode contribuir num melhor entendimento sobre a dinâmica populacional destas espécies.

Assim, tem-se que o objetivo geral deste projeto foi implementar algoritmos de aprendizado de máquina para prever as culturas e as épocas de amostragem em que certas espécies de plantas daninhas iriam aparecer em sistemas ILP, verificando quais elementos ambientais influenciam nas dinâmicas populacionais dessas espécies. Tal objetivo foi cumprido com o desenvolvimento dos modelos de predição de cultura e de época de amostragem, os quais tiveram boas métricas de desempenho. Todavia, a parte do objetivo de compreender quais fatores ambientais mais influenciam nessas dinâmicas populacionais, foi parcialmente atendida. Isso porque apenas em alguns casos foi possível estabelecer uma relação entre certas espécies de plantas daninhas e alguns dos elementos ambientais.

Além disso, também foram definidos três objetivos específicos. O primeiro objetivo específico referia-se a elaboração de uma base de dados unificada, com todas as informações necessárias de solo, clima, e plantio. Tem-se que os dados foram disponibilizados pela EMBRAPA, e que cada uma das bases passou pelas etapas de limpeza, transformação e enriquecimento, durante o pré-processamento dos dados. De modo a formar a base unificada.

Já o segundo objetivo específico abordava a identificação dos algoritmos de aprendizado de máquina mais adequados para o problema. Todavia, como constatou-se na revisão da literatura, os modelos desenvolvidos eram voltados apenas para o processamento de imagens. Assim, foi optado por implementar algoritmos de classificação, devido a natureza do problema.

Por último, no terceiro objetivo específico, o propósito era avaliar o desempenho dos algoritmos construídos, quanto a sua acurácia em prever as informações solicitadas. Isto é, prever a cultura e a época de amostragem em que as espécies de plantas daninhas iriam emergir. Tem-se que estes foram os modelos selecionados, devido a sua relevância para a EMBRAPA e também por seu

desenvolvimento ser viável pelo que constatou-se na análise estatística e nas primeiras tentativas de desenvolvimento feitas previamente.

Ao final, todos os algoritmos (Árvore de Decisão, Floresta Randômica, SVM e KNN) tiveram um desempenho satisfatório na predição das culturas. Os resultados deste modelo mostraram quais plantas daninhas eram mais propensas a aparecer em cada cultura. Quanto à predição da época de amostragem, tem-se que os algoritmos KNN e SVM também performaram bem, com destaque para o KNN. A partir desse modelo, observou-se as espécies com maior probabilidade de se desenvolverem em cada época.

Desta forma, este trabalho demonstra o uso de algoritmos de aprendizado de máquina para prever certos aspectos das dinâmicas populacionais de determinadas espécies de plantas daninhas, em sistemas ILP. Contudo, ainda existem diversos empecilhos que devem ser superados para um melhor desenvolvimento do trabalho. Dado que o objetivo inicial do projeto era elaborar um modelo de predição das espécies de plantas daninhas, isso, no entanto, não foi possível devido a quantidade de dados ser baixa e também devido a qualidade destes, conforme ficou evidente na análise estatística e nas tentativas de predições feitas.

Logo, a composição de um protocolo de coleta de dados é algo a ser proposto para a EMBRAPA; visando, assim, melhorar a quantidade e qualidade dos dados, e habilitar um melhor aproveitamento destes. Por exemplo, a realização de coletas periódicas e estratégicas de todas as informações necessárias.

Além disso, a estruturação do trabalho também tem que ser revista. Pois, deve-se considerar, que as informações de solo e clima referem-se às datas das coletas das plantas daninhas. Entretanto, tais plantas começaram a se desenvolver algum tempo antes disso. Sendo assim, os dados de solo e clima analisados deveriam concernir à época de germinação destas espécies. Para, então, poder compreender quais destes fatores favoreceram a emergência de tais plantas.

Portanto, tem-se que as principais contribuições científicas desta pesquisa são:

[i] Estudo sobre a utilização de algoritmos de aprendizado de máquina no manejo de plantas daninhas, em sistemas ILP;

[ii] Melhor compreensão sobre as limitações e oportunidades de trabalho referentes ao uso de aprendizado de máquina no manejo de plantas daninhas, com base nas conclusões obtidas na revisão da literatura;

[iii] Elaboração de uma base de dados unificada com informações de clima, solo e plantio;

[iv] Desenvolvimento dos modelos preditivos de culturas e épocas de amostragem das espécies de plantas daninhas, os quais podem servir como base para trabalhos futuros.

REFERÊNCIAS

ABDULLAH, Y.; BALOCH, M.; SHAH, A.N.; HASHIM, M.M.; NADIM, M.A.; ULLAH, G.; KHAN, A.A.; SHAHZAD, M.F. Weed Management in Wheat by Cuscuta Alone and in Combination with Commercial Weedicides Allymax and Axial. **Planta Daninha**, v. 38, n. 1, p. 1-12, nov. 2020. http://dx.doi.org/10.1590/s0100-83582020380100030.

ABOUZAHIR, S.; SADIK, M.; SABIR, E. Enhanced Approach for Weeds Species Detection Using Machine Vision. **International Conference On Electronics, Control, Optimization And Computer Science**, 2018 p. 1-6. http://dx.doi.org/10.1109/icecocs.2018.8610505.

AJAYI, O.G.; ASHI, J. Effect of varying training epochs of a Faster Region-Based Convolutional Neural Network on the Accuracy of an Automatic Weed Classification Scheme. **Smart Agricultural Technology**, v. 3, n. 100128, p. 1-14, fev. 2023. http://dx.doi.org/10.1016/j.atech.2022.100128.

ALROWAIS, F.; ASIRI, M.M.; ALABDAN, R.; MARZOUK, R.; HILAL, A.M.; ALKHAYYAT, A.; GUPTA, D. Hybrid leader based optimization with deep learning driven weed detection on internet of things enabled smart agriculture environment. **Computers And Electrical Engineering**, v. 104, n. 108411, p. 1-14, dez. 2022. http://dx.doi.org/10.1016/j.compeleceng.2022.108411.

ALURA. **Dicas de como escolher o tipo de visualização de dados para sua análise**. Disponível em: https://www.alura.com.br/artigos/tipo-de-visualizacao-de-dados. Acesso em: 11 mar. 2025.

ALVARENGA, R.C.; GONTIJO NETO, M.M.; RAMALHO, J.H.; GARCIA, J.C.; VIANA, M.C.M.; CASTRO, A.A.D.N. Sistema de Integração Lavoura-Pecuária: O modelo implantado na Embrapa Milho e Sorgo. Sete Lagoas: **Embrapa**, 2007. 9 p.

BALBINOT JUNIOR, A.A.; MORAES, A.; VEIGA, M.; PELISSARI, A.; DIECKOW, J. Integração lavoura-pecuária: intensificação de uso de áreas agrícolas. **Ciência Rural**, v. 39, n. 6, p. 1925-1933, maio 2009. http://dx.doi.org/10.1590/s0103-84782009005000107.

BEYELER, M. **Machine Learning for OpenCV: a practical introduction to the world of machine learning and image processing using opencv and python**. Birmingham: Packt Pub Ltd, 2017. 368 p.

BHATIA, A.S.; KALUZA, B. **Machine Learning in Java**. 2. ed. Birmingham: Packt Publishing Ltd., 2018. 290 p.

BREIMAN, L. Random forests. **Machine learning**, v. 45, n. 1, p. 5–32, 2001.

BRUCE, P.; BRUCE, A.; GEDECK, P. **Practical Statistics for Data Scientists: 50+ essential concepts using R and Python**. 2. ed. Sebastopol: O'Reilly Media, 2020. 360p.

CHAUHAN, Bhagirath Singh. Grand Challenges in Weed Management. **Frontiers In Agronomy**, v. 1, n. 3, p. 1-4, 22 jan. 2020. http://dx.doi.org/10.3389/fagro.2019.00003.

CHAVAN, T.R.; NANDEDKAR, A.V. AgroAVNET for crops and weeds classification: a step forward in automatic farming. **Computers And Electronics In Agriculture**, v. 154, p. 361-372, nov. 2018. http://dx.doi.org/10.1016/j.compag.2018.09.021.

CHESS ONLINE. 2024. Disponível em: https://chess-online-duel-friends-online.br.uptodown.com/android. Acesso em: 15/07/2024.

COELHO, L.P.; RICHERT, W.; BRUCHER, M. **Building Machine Learning Systems With Python**. 3. ed. Birmingham: Packt Publishing Ltd, 2018. 395 p.

CONCENÇO, G.; SALTON, J.C.; MARQUES, R.F.; PALHARINI, W.G.; ALVES, M.E.S.; SANTOS, S.A.; GALON, L. Weed suppression in sustainable integrated agricultural systems. **Pakistan Journal Of Weed Science Research**, v. 21, n. 1, p. 1-14, jan. 2015. https://www.embrapa.br/en/busca-de-publicacoes/-/publicacao/1030094/weed-suppression-in-sustainable-integrated-agricultural-systems

COSTELLO, B.; OSUNKOYA, O.O.; SANDINO, J.; MARINIC, W.; TROTTER, P.; SHI, B.; GONZALEZ, F.; DHILEEPAN, K. Detection of Parthenium Weed (Parthenium hysterophorus L.) and Its Growth Stages Using Artificial Intelligence. **Agriculture**, v. 12, n. 11, p. 2-23, 2 nov. 2022. http://dx.doi.org/10.3390/agriculture12111838.

DANG, F.; CHEN, D.; LU, Y.; LI, Z. YOLOWeeds: a novel benchmark of yolo object detectors for multi-class weed detection in cotton production systems. **Computers And Electronics In Agriculture**, v. 205, n. 107655, p. 1-13, fev. 2023. http://dx.doi.org/10.1016/j.compag.2023.107655.

DOMINSCHEK, R.; SCHUSTER, M.Z.; BARROSO, A.A.M.; MORAES, A.; ANGHINONI, I.; CARVALHO, P.C.F. Diversification of traditional paddy field impacts target species in weed seedbank. **Revista Ciência Agronômica**, v. 53, n. 1, p. 1-10, mar. 2022. http://dx.doi.org/10.5935/1806-6690.20220030.

DUARTE, P.M.; SANTANA, V.T.P.; DALMAS, A.D.; FERRI, I.E.B. Integração Lavoura-Pecuária (ILP): uma revisão literária. **Uniciências**, v. 22, n. 2, p. 106-109, dez. 2018. http://dx.doi.org/10.17921/1415-5141.2018v22n2p106-109.

EMBRAPA MILHO E SORGO. Sistema ILP: produção de grãos, forragens e carne na região central de Minas Gerais. Sete Lagoas: **Embrapa**, 2018. https://ainfo.cnptia.embrapa.br/digital/bitstream/item/193357/1/Sistema-ILP.pdf

ERENSTEIN, O.; JALETA, M.; SONDER, K.; MOTTALEB, K.; PRASANNA, B.M. Global maize production, consumption and trade: trends and r&d implications. **Food Security**, v. 14, n. 5, p. 1295-1319, mai. 2022. http://dx.doi.org/10.1007/s12571-022-01288-7.

ETIENNE, A.; AHMAD, A.; AGGARWAL, V.; SARASWAT, D. Deep Learning-Based Object Detection System for Identifying Weeds Using UAS Imagery. **Remote Sensing**, v. 13, n. 24, p. 1-22, dez. 2021. http://dx.doi.org/10.3390/rs13245182.

FAWAKHERJI, M.; POTENA, C.; PRETTO, A.; BLOISI, D.D.; NARDI, D. Multi-Spectral Image Synthesis for Crop/Weed Segmentation in Precision Farming. **Robotics And Autonomous Systems**, v. 146, n. 103861, p. 1-13, dez. 2021. http://dx.doi.org/10.1016/j.robot.2021.103861.

FIRMANSYAH, E.; SUPARYANTO, T.; HIDAYAT, A.A.; PARDAMEAN, B. Real-time Weed Identification Using Machine Learning and Image Processing in Oil Palm Plantations. **Iop Conference Series: Earth and Environmental Science**, 2022 p. 1-9. http://dx.doi.org/10.1088/1755-1315/998/1/012046.

FREUND, John E. **Estatística Aplicada:** economia, administração e contabilidade. 11. ed. Porto Alegre: Bookman, 2009.

GAMA, J.C.M.; JESUS, L.L.; KARAM, D. Fitossociologia de plantas espontâneas em sistema de integração lavoura- pecuária. **Revista Brasileira de Agroecologia**, v. 2, n. 2, p. 929-932, out. 2007.

GAO, J.; NUYTTENS, D.; LOOTENS, P.; HE, Y.; PIETERS, J.G. Recognising weeds in a maize crop using a random forest machine-learning algorithm and near-infrared snapshot mosaic hyperspectral imagery. **Biosystems Engineering**, v. 170, n. 1, p. 39-50, jun. 2018. http://dx.doi.org/10.1016/j.biosystemseng.2018.03.006.

GARRETT, R.; NILES, M.; GIL, J.; DY, P.; REIS, J.; VALENTIM, J. Policies for Reintegrating Crop and Livestock Systems: a comparative analysis. **Sustainability**, v. 9, n. 3, p. 473, mar. 2017. http://dx.doi.org/10.3390/su9030473.

GÉRON, A. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: hands-on machine learning with scikit-learn, keras, and tensorflow**. 2. ed. Sebastopol: O'Reilly Media, 2019. 848 p.

GRUS, Joel. **Data Science do Zero:** noções fundamentais com python. Rio de Janeiro: Atla Books, 2021.

HUSSAIN, N.; FAROOQUE, A.A.; SCHUMANN, A.W.; ABBAS, F.; ACHARYA, B.; MCKENZIE-GOPSILL, A.; BARRETT, R.; AFZAAL, H.; ZAMAN, Q.U.; CHEEMA, M.J.M. Application of deep learning to detect Lamb's quarters (Chenopodium album L.) in potato fields of Atlantic Canada. **Computers And Electronics In Agriculture**, v. 182, n. 1, p. 1-9, mar. 2021. http://dx.doi.org/10.1016/j.compag.2021.106040.

IKEDA, F.S.; MITJA, D.; VILELA, L.; CARMONA, R. Banco de sementes no solo em sistemas de cultivo lavoura-pastagem. **Pesquisa Agropecuária Brasileira**, v. 42, n. 11, p. 1545-1551, nov. 2007. http://dx.doi.org/10.1590/s0100-204x2007001100005.

JHA, K.; DOSHI, A.; PATEL, P.; SHAH, M. A comprehensive review on automation in agriculture using artificial intelligence. **Artificial Intelligence In Agriculture**, v. 2, n. 1, p. 1-12, jun. 2019. http://dx.doi.org/10.1016/j.aiia.2019.05.004.

JOSE, J.A.; SHARMA, A.; SEBASTIAN, M.; DENSIL, R.V.F. Classification of Weeds and Crops using Transfer Learning. **International Conference On Advances In Computing, Communication And Applied Informatics**, 2022 p. 1-7. http://dx.doi.org/10.1109/accai53970.2022.9752477.

KAUR, S.; KAUR, R.; CHAUHAN, B.S. Understanding crop-weed-fertilizer-water interactions and their implications for weed management in agricultural systems. **Crop Protection**, v. 103, n. 1, p. 65-72, jan. 2018. http://dx.doi.org/10.1016/j.cropro.2017.09.011.

KLUTHCOUSKI, J.; AIDAR, H.; COBUCCI, T. Opções e vantagens da Integração Lavoura-Pecuária e a produção de forragens na entressafra. **Informe Agropecuário**, v. 28, n. 240, p. 16-27, out. 2007.

KUBIAK, A.; WOLNA-MARUWKA, A.; NIEWIADOMSKA, A.; PILARSKA, AA. The Problem of Weed Infestation of Agricultural Plantations vs. the Assumptions of the European Biodiversity Strategy. **Agronomy**, v. 12, n. 8, p. 1-29, jul. 2022. http://dx.doi.org/10.3390/agronomy12081808.

LIBERATI, A.; ALTMAN, D.G.; TETZLAFF, J.; MULROW, C.; GØTZSCHE, P.C.; IOANNIDIS, J.P.A.; CLARKE, M.; DEVEREAUX, P.J.; KLEIJNEN, J.; MOHER, D. The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: explanation and elaboration. **Plos Medicine**, v. 6, n. 7, p. 1-29, jul. 2009. http://dx.doi.org/10.1371/journal.pmed.1000100.

LUSTOSA, S.B.C.; SCHUSTER, M.Z.; MARTINICHEN, D.; PELISSARI, A.; GAZZIERO, D.L.P. Floristic and phytosociology of weed in response to winter pasture sward height at Integrated Crop-Livestock in Southern Brazil. **Revista Brasileira de Tecnologia Aplicada nas Ciências Agrárias**, v. 9, n. 2, p. 19-26, 2016. http://dx.doi.org/10.5935/paet.v9.n02.02.

MEENA, S.D.; SUSANK, M.; GUTTULA, T.; CHANDANA, S.H.; SHEELA, J. Crop Yield Improvement with Weeds, Pest and Disease Detection. **Procedia Computer Science**, v. 218, n. 1, p. 2369-2382, jan. 2023. http://dx.doi.org/10.1016/j.procs.2023.01.212.

MONTEIRO, A.L.; SOUZA, M.F.; LINS, H.A.; TEÓFILO, T.M.S.; BARROS JÚNIOR, A.P.; SILVA, D.V.; MENDONÇA, V. A new alternative to determine weed control in agricultural systems based on artificial neural networks (ANNs). **Field Crops Research**, v. 263, p. 108075, abr. 2021. http://dx.doi.org/10.1016/j.fcr.2021.108075.

MONTEIRO, A.; SANTOS, S. Sustainable Approach to Weed Management: the role of precision weed management. **Agronomy**, v. 12, n. 1, p. 1-14, 4 jan. 2022. http://dx.doi.org/10.3390/agronomy12010118.

NASIRI, A.; OMID, M.; TAHERI-GARAVAND, A.; JAFARI, A. Deep learning-based precision agriculture through weed recognition in sugar beet fields. **Sustainable Computing: Informatics and Systems**, v. 35, n. 100759, p. 1-11, set. 2022. http://dx.doi.org/10.1016/j.suscom.2022.100759.

NGO, K.; CHUA, J.; CHUN, B.; AI, R.T. Automated Weed Detection System for Bok Choy Using Computer Vision. **14Th International Conference On Humanoid, Nanotechnology, Information Technology, Communication And Control, Environment, And Management**, 2022 p. 1-6. http://dx.doi.org/10.1109/hnicem57413.2022.10109618.

NI, C.; TIAN, B.; WANG, X.; SUN, Y.; FEI, C. A deep convolutional neural network-based method for identifying weed seedlings in maize fields. **5Th Advanced Information Management, Communicates, Electronic And Automation Control Conference**, 2022 p. 776-779. http://dx.doi.org/10.1109/imcec55388.2022.10019943.

OECD - Organização para a Cooperação e Desenvolvimento Econômico. Artificial Intelligence and Employment: new evidence from occupations most exposed to ai. Paris: **Oecd - Organização Para a Cooperação e Desenvolvimento Econômico**, 2023.

OLIVEIRA, M.F.; SILVA, C.H.L.; ALVARENGA, R.C.; SILVA, A.F. Monitoramento de Plantas Daninhas em Sistema Integrado entre Lavoura e Pecuária em Sete Lagoas, MG. Sete Lagoas: **Embrapa**, 2018. 19 p.

OLIVEIRA JUNIOR, R.S.; CONSTANTIN, J.; INOUE, M.H. **Biologia e Manejo de Plantas Daninhas**. Curitiba: Omnipax Editora Ltda, 2011.

ONU. População mundial chegará a 9,9 bilhões em 2054. 2024. **Disponível em:** https://news.un.org/pt/story/2024/04/1830966#:~:text=Popula%C3%A7%C3%A3o%20mundial%20chegar%C3%A1%20a%209%2C9%20bilh%C3%B5es%20em%202054%20%7C%20ONU%20News. Acesso em: 25 fev. 2025.

PÉREZ-ORTIZ, M.; PEÑA, J.M.; GUTIÉRREZ, P.A.; TORRES-SÁNCHEZ, J.; HERVÁS-MARTÍNEZ, C.; LÓPEZ-GRANADOS, F. Selecting patterns and features for between- and within- crop-row weed mapping using UAV-imagery. **Expert Systems With Applications**, v. 47, n. 1, p. 85-94, abr. 2016. http://dx.doi.org/10.1016/j.eswa.2015.10.043.

PARTEL, V.; KAKARLA, S.C.; AMPATZIDIS, Y. Development and evaluation of a low-cost and smart technology for precision weed management utilizing artificial intelligence. **Computers And Electronics In Agriculture**, v. 157, n. 1, p. 339-350, fev. 2019. http://dx.doi.org/10.1016/j.compag.2018.12.048.

QIAO, X.; LI, Y.Z.; SU, G.Y.; TIAN, H.K.; ZHANG, S.; SUN, Z.Y.; YANG, L.; WAN, F.H.; QIAN, W.Q. MmNet: identifying mikania micrantha kunth in the wild via a deep convolutional neural

network. **Journal Of Integrative Agriculture**, v. 19, n. 5, p. 1292-1300, mai. 2020. http://dx.doi.org/10.1016/s2095-3119(19)62829-7.

RAJA, R.; SLAUGHTER, D.C.; FENNIMORE, S.A.; SIEMENS, M.C. Real-time control of high-resolution micro-jet sprayer integrated with machine vision for precision weed control. **Biosystems Engineering**, v. 228, n. 1, p. 31-48, abr. 2023. http://dx.doi.org/10.1016/j.biosystemseng.2023.02.006.

RAZFAR, N.; TRUE, J.; BASSIOUNY, R.; VENKATESH, V.; KASHEF, R. Weed detection in soybean crops using custom lightweight deep learning models. **Journal Of Agriculture And Food Research**, v. 8, n. 100308, p. 1-10, jun. 2022. http://dx.doi.org/10.1016/j.jafr.2022.100308.

RIBAS, P.P.; MATSUMURA, A.T.S. A química dos agrotóxicos: impacto sobre a saúde e meio ambiente. **Revista Liberato**, v. 10, n. 14, p. 149-158, jul. 2009. http://www.revista.liberato.com.br/index.php/revista/article/view/142

RONQUIM, C.C. Conceitos de fertilidade do solo e manejo adequado para as regiões tropicais. Campinas: **Embrapa**, 2010. 30 p.

RUSSELL, S.; NORVIG, P. **Inteligência Artificial**. 4. ed. Rio de Janeiro: Livros Técnicos e Científicos Editora, 2022. 969 p.

SABZI, S.; ABBASPOUR-GILANDEH, Y. Using video processing to classify potato plant and three types of weed using hybrid of artificial neural network and partincle swarm algorithm. **Measurement**, v. 126, n. 1, p. 22-36, out. 2018. http://dx.doi.org/10.1016/j.measurement.2018.05.037.

SABZI, S.; ABBASPOUR-GILANDEH, Y.; ARRIBAS, J.I. An automatic visible-range video weed detection, segmentation and classification prototype in potato field. **Heliyon**, v. 6, n. 5, p. 1-17, mai. 2020. http://dx.doi.org/10.1016/j.heliyon.2020.e03685.

SANDINO, J.; GONZALEZ, F. A Novel Approach for Invasive Weeds and Vegetation Surveys Using UAS and Artificial Intelligence. **23Rd International Conference On Methods & Models In Automation & Robotics**, 2018, p. 515-520. http://dx.doi.org/10.1109/mmar.2018.8485874.

SCHUSTER, M.Z.; LUSTOSA, S.B.C.; PELISSARI, A.; HARRISON, S.K.; SULC, R.M.; DEISS, L.; LANG, C.R.; CARVALHO, P.C.F.; GAZZIERO, D.L.P.; MORAES, A. Optimizing forage allowance for productivity and weed management in integrated crop-livestock systems. **Agronomy For Sustainable Development**, v. 39, n. 2, p. 1-10, 5 mar. 2019. http://dx.doi.org/10.1007/s13593-019-0564-4.

SCHUSTER, M.Z.; PELISSARI, A.; MORAES, A.; HARRISON, S.K.; SULC, R.M.; LUSTOSA, S.B.C.; ANGHINONI, I.; CARVALHO, P.C.F. Grazing intensities affect weed seedling emergence and the seed bank in an integrated crop–livestock system. **Agriculture, Ecosystems & Environment**, v. 232, n. 1, p. 232-239, set. 2016. http://dx.doi.org/10.1016/j.agee.2016.08.005.

SEKARAN, U.; LAI, L.; USSIRI, D.A.N.; KUMAR, S.; CLAY, S. Role of integrated crop-livestock systems in improving agriculture production and addressing food security – A review. **Journal Of Agriculture And Food Research**, v. 5, n. 100190, p. 1-10, set. 2021. http://dx.doi.org/10.1016/j.jafr.2021.100190.

SHOREWALA, S.; ASHFAQUE, A.; SIDHARTH, R.; VERMA, U. Weed Density and Distribution Estimation for Precision Agriculture Using Semi-Supervised Learning. **Ieee Access**, v. 9, n. 1, p. 27971-27986, fev. 2021. http://dx.doi.org/10.1109/access.2021.3057912.

SIDDIQUI, S.A.; FATIMA, N.; AHMAD, A. Neural Network based Smart Weed Detection System. **International Conference On Communication, Control And Information Sciences**, 2021, p. 1-5. http://dx.doi.org/10.1109/iccisc52257.2021.9484925.

SOUZA, M.F.; AMARAL, L.R.; OLIVEIRA, S.R.M.; COUTINHO, M.A.N.; NETTO, C.F. Spectral differentiation of sugarcane from weeds. **Biosystems Engineering**, v. 190, n. 1, p. 41-46, fev. 2020. http://dx.doi.org/10.1016/j.biosystemseng.2019.11.023.

SUBEESH, A.; BHOLE, S.; SINGH, K.; CHANDEL, N.; RAJWADE, Y.A.; RAO, K.V.R.; KUMAR, S.P.; JAT, D. Deep convolutional neural network models for weed detection in polyhouse grown bell peppers. **Artificial Intelligence In Agriculture**, v. 6, n. 1, p. 47-54, jan. 2022. http://dx.doi.org/10.1016/j.aiia.2022.01.002.

SUDARS, K.; JASKO, J.; NAMATEVS, I.; OZOLA, L.; BADAUKIS, N. Dataset of annotated food crops and weed images for robotic computer vision control. **Data In Brief**, v. 31, n. 105833, p. 1-6, ago. 2020. http://dx.doi.org/10.1016/j.dib.2020.105833.

SNYDER, H. Literature review as a research methodology: an overview and guidelines. **Journal Of Business Research**, v. 104, n. 1, p. 333-339, nov. 2019. http://dx.doi.org/10.1016/j.jbusres.2019.07.039.

TikZ.net. 2021. **Disponível em:** https://tikz.net/random-forest/. Acesso em: 18/07/2024

TORRES-SOSPEDRA, J.; NEBOT, P. Two-stage procedure based on smoothed ensembles of neural networks applied to weed detection in orange groves. **Biosystems Engineering**, v. 123, n. 1, p. 40-55, jul. 2014. http://dx.doi.org/10.1016/j.biosystemseng.2014.05.005.

VARGAS, L.; PEIXOTO, C.M.; ROMAN, Erivelton Scherer. Manejo de plantas daninhas na cultura do milho. **Embrapa**, [s. l], v. 1, n. 1, p. 1-67, set. 2006.

WANG, M.Y.; LEELAPATRA, W. Weeding Robot Based on Lightweight Platform and Dual Cameras. **14Th International Conference On Software, Knowledge, Information Management And Applications**, 2022, p. 157-163. http://dx.doi.org/10.1109/skima57145.2022.10029527.

WANG, A.; XU, Y.; WEI, X.; CUI, B. Semantic Segmentation of Crop and Weed using an Encoder-Decoder Network and Image Enhancement Method under Uncontrolled Outdoor

Illumination. **Ieee Access**, v. 8, n. 1, p. 81724-81734, abr. 2020. http://dx.doi.org/10.1109/access.2020.2991354.

WIKIVERSIDADE. DC-UFRPE/Bacharelado em Ciência da Computação/Inteligência Artificial/aprendizado por reforco. 2023. **Disponível em:** https://pt.wikiversity.org/wiki/DC-UFRPE/Bacharelado_em_Ci%C3%AAncia_da_Computa%C3%A7%C3%A3o/Intelig%C3%AAncia_Artificial/aprendizado_por_reforco. Acesso em: 25 fev. 2025.

YADURAJAU, N.T.; RAO, A.N. **Implications of weeds and weed management on food security and safety in the asia-pacific region**. Em: BAKAR, Baki Hj; KURNIADIE, Denny; TJITROSOEDIRDJO, Soekisman. The role of weed science in supporting food security by 2020. Bandung: Asian-Pacific Weed Science Society, Weed Science Society Of Indonesia And Padjadjaran University Bandung, 2020.

YAN, X.; DENG, X.; JIN, J. Classification of weed species in the paddy field with DCNN-Learned features. **5Th Information Technology And Mechatronics Engineering Conference**, 2020 p. 336-340. http://dx.doi.org/10.1109/itoec49072.2020.9141894.

YU, J.; SCHUMANN, A.W.; CAO, Z.; SHARPE, S.M.; BOYD, N.S. Weed Detection in Perennial Ryegrass With Deep Learning Convolutional Neural Network. **Frontiers In Plant Science**, v. 10, n. 1, p. 1-9, out. 2019. http://dx.doi.org/10.3389/fpls.2019.01422.

YU, J.; SCHUMANN, A.W.; SHARPE, S.M.; LI, X.; BOYD, N.S. Detection of grassy weeds in bermudagrass with deep convolutional neural networks. **Weed Science**, v. 68, n. 5, p. 545-552, 8 jun. 2020. http://dx.doi.org/10.1017/wsc.2020.46.

ZHANG, W.; HANSEN, M.F.; VOLONAKIS, T.N.; SMITH, M.; SMITH, L.; WILSON, J.; RALSTON, G.; BROADBENT, L.; WRIGHT, G. Broad-Leaf Weed Detection in Pasture. 3Rd **International Conference On Image, Vision And Computing**, 2018, p. 101-105. http://dx.doi.org/10.1109/icivc.2018.8492831.