



Programa de Pós-Graduação em

**Computação Aplicada**

**Mestrado/Doutorado Acadêmico**

Tiago Boechel

UM MODELO DE MACHINE-LEARNING PARA PREDIÇÃO  
DO TEMPO DE COLHEITA DE ÁRVORES MACIEIRAS COM  
BASE EM DADOS FENOLÓGICOS E PARÂMETROS  
CLIMÁTICOS

São Leopoldo, 2022



Tiago Boechel

**UM MODELO DE MACHINE-LEARNING PARA PREDIÇÃO DO TEMPO DE  
COLHEITA DE ÁRVORES MACIEIRAS COM BASE EM DADOS FENOLÓGICOS E  
PARÂMETROS CLIMÁTICOS**

Tese apresentada como requisito parcial para a  
obtenção do título de Doutor pelo Programa de  
Pós-Graduação em Computação Aplicada da  
Universidade do Vale do Rio dos Sinos —  
UNISINOS

Orientador:  
Prof. Dr. Rodrigo da Rosa Righi

Coorientador:  
Prof. Dr. Gabriel de Oliveira Ramos

São Leopoldo, RS  
2022

B669m Boechel, Tiago.

Um modelo de machine-learning para predição do tempo de colheita de árvores macieiras com base em dados fenológicos e parâmetros climáticos / por Tiago Boechel. – 2022.

87 f. : il. ; 30 cm.

Tese (doutorado) — Universidade do Vale do Rio dos Sinos, Programa de Pós-Graduação em Computação Aplicada, São Leopoldo, RS, 2022.

Orientador: Dr. Rodrigo da Rosa Righi.

Coorientador: Dr. Gabriel de Oliveira Ramos

1. Estádios fenológicos de macieiras. 2. Predição de data da colheita. 3. Redes neurais (Computação). I. Título.

CDU: 004.8:634.11

ATA DE BANCA EXAMINADORA DE TESE DE DOUTORADO Nº 09/2022

Aluno: Tiago Boechel

Título da Tese: "UM MODELO DE MACHINE-LEARNING PARA PREDIÇÃO DO TEMPO DE COLHEITA DE ÁRVORES MACIEIRAS COM BASE EM DADOS FENOLÓGICOS E PARÂMETROS CLIMÁTICOS"

Banca: Prof. Dr. Rodrigo da Rosa Righi (Orientador, UNISINOS)  
Prof. Dr. Gabriel de Oliveira Ramos (Coorientador, UNISINOS)  
Prof. Dr. Jorge Luis Victória Barbosa (Avaliador, UNISINOS)  
Prof. Dr. Gustavo Pessin (Avaliador, UFOP)  
Prof. Dr. Adenauer Correa Yamin (Avaliador, UFPEL)

Aos vinte e dois dias do mês de abril do ano de 2022, às 8h30 reuniu-se a Comissão Examinadora de Defesa de tese composta pelos professores: Prof. Dr. Rodrigo da Rosa Righi, Orientador – UNISINOS (por webconferência); Prof. Dr. Gabriel de Oliveira Ramos, Coorientador – UNISINOS; Prof. Dr. Jorge Luis Victória Barbosa, Avaliador - UNISINOS; Prof. Dr. Gustavo Pessin, Avaliador – UFOP (por webconferência) e Prof. Dr. Adenauer Correa Yamin, Avaliador – UFPEL (por webconferência) para analisar e avaliar a Tese apresentada pelo(a) aluno(a) **Tiago Boechel** (por webconferência).

A Banca Examinadora, em cumprimento ao requisito exigido para a obtenção do Título de Doutor em Computação Aplicada, julga esta tese:

APROVADA     REPROVADA

São Leopoldo, 22 de abril de 2022.

Prof. Dr. Rodrigo da Rosa Righi  
Orientador



## **AGRADECIMENTOS**

A minha esposa Danúbia e minha filha Mariana pelo apoio e paciência nos muitos momentos de ausência. Vocês são muito especiais para mim e me fazem acreditar, a cada dia que passa, que todos os sacrifícios valem a pena.

Aos meus pais que sempre acreditaram em meus sonhos e objetivos, me apoiaram e incentivaram nos meus estudos.

Ao Instituto Federal do Rio Grande do Sul (IFRS) - Câmpus Vacaria por possibilitar aos seus servidores a participação em programas de qualificação em pós-graduação.

Ao Diretor Geral do IFRS Câmpus Vacaria. Prof. Dr. Gilberto Luiz Putti, pela amizade e boa vontade em ajudar sempre que solicitado.

Ao Engenheiro Agrônomo José Luiz Petri, pela paciência em esclarecer minhas dúvidas e por todo conhecimento e ensinamentos transmitidos que foram fundamentais para a realização deste trabalho.

Ao meu orientador Prof. Dr. Rodrigo da Rosa Righi, e meu coorientador Prof. Dr. Gabriel de Oliveira Ramos pela paciência, dedicação e orientação.

Ao colega, do programa de Pós Graduação em Computação Aplicada da Unisinos, Lucas Micol Policarpo pelas valiosas sugestões e contribuições no trabalho.

Aos colegas, do Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul (IFRS) - Câmpus Vacaria, Prof. Gabriel Nachtigall Marques, Prof. Rogério Ricalde Torres e Prof. Raphael Rodrigues Cunha pelas valiosas contribuições.

A Empresa de Pesquisa Agropecuária e Extensão Rural de Santa Catarina (EPAGRI), Estação Experimental de Caçador.

Ao Departamento de Diagnóstico e Pesquisa Agropecuária (DDPA) da Secretaria Estadual da Agricultura, Pecuária e Desenvolvimento Rural do Rio Grande do Sul.

A Engenheira Agrônoma, Ma. Camila Cargnino.

A Engenheira Agrônoma, Dra. Rosa Maria Valdebenito Sanhueza e a empresa Proterra Engenharia Agronômica.

A todas as pessoas, que embora não tenham sido citadas, contribuíram de alguma forma para a conclusão deste trabalho.



## RESUMO

Abordagens de aprendizado de máquina têm sido usadas em várias áreas. No campo da pesquisa agropecuária, o aprendizado de máquina tem sido utilizado para aumentar a produtividade agrícola e minimizar seu impacto ambiental, mostrando-se uma importante ferramenta de apoio à tomada de decisões. Diferentes estratégias são encontradas na literatura para prever estágios fenológicos de diferentes culturas. Do estado da arte atual, observamos poucos trabalhos que abordam a previsão da data de colheita. Não encontramos trabalhos com abordagem semelhante ao proposto. Prever a época da colheita é um desafio para desenvolver a produção de frutas de forma sustentável e reduzir o desperdício de alimentos. As frutas são perecíveis, de alto valor e sazonais, e os preços de venda geralmente são sensíveis ao tempo, o que torna as previsões de colheita extremamente valiosas para os produtores. Este estudo propõe o modelo PredHarv, um modelo de aprendizado de máquina que utiliza redes neurais recorrentes, para prever a data de início da colheita da maçã, dadas as condições climáticas relacionadas à temperatura, esperadas para o período. As previsões são feitas a partir do estágio fenológico da plena floração, com base nas séries históricas da fenologia e dados meteorológicos. O modelo computacional contribui com a capacidade de antecipar informações sobre a data de colheita, possibilitando ao produtor planejar melhor as atividades, evitando custos e melhorando a produtividade. A utilização de métodos de ML visa tornar mais eficaz a capacidade preditiva de modelos baseados em soma térmica voltados para a fruticultura, permitindo simular as mudanças climáticas no período. O modelo PredHarv é baseado em modelos de soma térmica, mas usa uma abordagem multivariada. Usamos a soma térmica relacionando-a com a duração do período e outras variáveis relacionadas a temperatura do período. Usamos um método de aprendizado de máquina, explorando o potencial das redes LSTM para tratar problemas que envolvem séries temporais. A saída do modelo retorna à duração do período em dias de calendário, dadas as condições climáticas relacionadas a temperatura esperadas para o período. Adicionalmente uma metodologia de utilização do modelo é proposta com o objetivo de ampliar a capacidade preditiva, como forma de reduzir a incerteza implícita nas informações fornecidas pelo usuário, necessárias para o cálculo da previsão. Desenvolvemos um protótipo do modelo PredHarv e realizamos experimentos com dados reais de instituições agrícolas. A combinação das variáveis utilizadas no modelo demonstrou uma estratégia de predição eficaz. Avaliamos as métricas e os resultados obtidos nos cenários de avaliação demonstram que o modelo é eficiente, com boa generalização e capaz de obter resultados com melhor acurácia comparado ao modelo linear baseado em acúmulo térmico.

**Palavras-chave:** Estádios fenológicos de macieiras. Predição de data da colheita. Redes neurais recorrentes.



## ABSTRACT

Machine learning approaches have been used in several areas. In the field of agricultural research, machine learning has been used to increase agricultural productivity and minimize its environmental impact, proving to be an important tool to support decision making. Different strategies are found in the literature to predict phenological stages of different cultures. From the current state of the art, we observed few works that address the prediction of the harvest date. We did not find works with an approach similar to the one proposed. Forecasting the time of harvest is a challenge to develop fruit production sustainably and reduce food waste. Fruits are perishable, of high value and seasonal, and sales prices are generally time sensitive, which makes harvest forecasts extremely valuable to growers. This study proposes the PredHarv model, a machine learning model that uses recurrent neural networks to predict the start date of the apple harvest, given the temperature-related weather conditions expected for the period. Predictions are made from the phenological stage of full bloom, based on historical series of phenology and meteorological data. The computational model contributes with the ability to anticipate information about the harvest date, enabling the producer to better plan activities, avoiding costs and improving productivity. The use of ML methods aims to make the predictive capacity of models based on thermal summation aimed at fruit growing more effective, allowing for the simulation of climate changes in the period. The PredHarv model is based on thermal sum models, but uses a multivariate approach. We use the thermal sum relating it to period length and other variables related to period temperature. We use a machine learning method, exploring the potential of LSTM networks to deal with problems involving time series. The model output returns the period length in calendar days, given the expected temperature-related weather conditions for the period. Additionally, a methodology for using the model is proposed in order to expand the predictive capacity, as a way to reduce the uncertainty implicit in the information provided by the user, necessary for calculating the forecast. We developed a prototype of the PredHarv model and performed experiments with real data from agricultural institutions. The combination of variables used in the model demonstrated an effective prediction strategy. We evaluated the metrics and the results obtained in the evaluation scenarios demonstrate that the model is efficient, with good generalization and capable of obtaining results with better accuracy compared to the linear model based on thermal accumulation.

**Keywords:** Phenological stages of apple trees. Prediction harvest date. Recurrent neural networks.



## LISTA DE FIGURAS

Figura 1 – Pomares de maçã na cidade de Vacaria, RS. . . . .	21
Figura 2 – Ciclo anual da macieira . . . . .	24
Figura 3 – Sequência de estádios de desenvolvimento fenológico da macieira . . . . .	27
Figura 4 – Visão simplificada da estrutura típica de um neurônio . . . . .	30
Figura 5 – Diagrama de uma rede neural artificial . . . . .	31
Figura 6 – (a) Representação de uma RNN com <i>loops</i> ; (b) Representação de uma RNN desdobrada . . . . .	32
Figura 7 – (a) Representação de uma rede <i>feedforward</i> ; (b) Representação de uma RNN	32
Figura 8 – Arquitetura da rede LSTM . . . . .	33
Figura 9 – O estado da célula da arquitetura da rede LSTM . . . . .	33
Figura 10 – Protocolo da revisão da literatura com abordagem em 3 etapas . . . . .	37
Figura 11 – Ciclo de desenvolvimento da macieira . . . . .	49
Figura 12 – Arquitetura do modelo . . . . .	50
Figura 13 – Boxplot das variáveis de entrada - Conjunto de dados DS1 . . . . .	51
Figura 14 – Boxplot das variáveis de entrada - Conjunto de dados DS2 . . . . .	52
Figura 15 – Linha do tempo . . . . .	53
Figura 16 – Fluxograma da execução do modelo . . . . .	54
Figura 17 – Diagrama de sequência dos eventos do processo de previsão . . . . .	55
Figura 18 – Séries temporais das variáveis de temperatura 2000 a 2018 - Conjunto de dados ds1 . . . . .	57
Figura 19 – Séries temporais das variáveis de temperatura 2016 a 2020 - Conjunto de dados ds2 . . . . .	58
Figura 20 – Matriz de correlação entre as variáveis de entrada. . . . .	59
Figura 21 – Esquema de divisão de dados de treinamento e teste . . . . .	60
Figura 22 – Representação da metodologia de validação <i>Walk-forward</i> . . . . .	64
Figura 23 – Dinâmica das previsões obtidas com o modelo PredHarv . . . . .	70
Figura 24 – Resultados das previsões do modelo PredHarv e do modelo Linear - Con- junto de dados DS1 . . . . .	73
Figura 25 – Resíduos das previsões do modelo PredHarv e do modelo Linear - Conjunto de dados DS1 . . . . .	73
Figura 26 – Resultados previsões do PredHarv e modelo Linear - DS2 . . . . .	74
Figura 27 – Resíduos das previsões do modelo PredHarv e do modelo Linear - Conjunto de dados DS2 . . . . .	74
Figura 28 – Duração do período em função de graus-dia acumulado . . . . .	87



## LISTA DE TABELAS

Tabela 1 – Resumo dos artigos resultantes da revisão bibliográfica . . . . .	44
Tabela 2 – Tabela comparativa das abordagens nos trabalhos relacionados . . . . .	46
Tabela 3 – Exemplo de dados de entrada usados na implementação . . . . .	56
Tabela 4 – Resultados da metodologia de validação <i>Walk Forward</i> . . . . .	68
Tabela 5 – Resultados das previsões obtidas com o segundo cenário de avaliação - Conjunto de dados DS1 . . . . .	68
Tabela 6 – Resultados das previsões obtidas com o segundo cenário de avaliação - Conjunto de dados DS2 . . . . .	69
Tabela 7 – Melhores resultados categoria GML usando AutoSklearn. . . . .	70
Tabela 8 – Melhores resultados categoria DL usando AutoKeras. . . . .	71
Tabela 9 – Número de dias dos eventos fenológicos da macieira cultivar Gala nas safras 2000 a 2018. EPAGRI Caçador, SC. . . . .	83
Tabela 10 – Valores e médias anuais da temperatura mínima (Tmin), temperatura máxima (Tmax), temperatura média (Tmed), amplitude (Ampl), média temperatura mínima (mmed), média temperatura máxima (mmax), média de graus-dia (mGD), graus-dia acumulado (GDA) e duração de amostras do período de 2000 a 2018. Temperatura base de 2°. Fonte: EPAGRI Caçador, SC. . . . .	84
Tabela 11 – Valores e médias anuais da temperatura mínima (Tmin), temperatura máxima (Tmax), temperatura média (Tmed), amplitude (Ampl), média temperatura mínima (mmed), média temperatura máxima (mmax), média de graus-dia (mGD), graus-dia acumulado (GDA) e duração do período de amostras 2016 a 2020. Temperatura base de 2°. Fonte dados meteorológicos: Departamento de Diagnóstico e Pesquisa Agropecuária (DDPA) da Secretaria Estadual da Agricultura, Vacaria, RS. . . . .	85
Tabela 12 – Tabela de Graus-dia acumulado com diferentes valores de temperatura base. Safras 2000 a 2018. EPAGRI Caçador, SC. . . . .	86



## LISTA DE SIGLAS

ABNT	Associação Brasileira de Normas Técnicas
ABPM	Associação Brasileira de Produtores de Maçã
AM	Aprendizado de máquina
AGAPOMI	Associação Gaúcha de Produtores de Maçã
AP	Aprendizagem Profunda
FAPERGS	Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul
CPU	Unidade central de processamento
CSV	<i>Comma-separated values</i>
GPU	Unidade de processamento gráfico
HF	Horas de frio
IFRS	Instituto Federal do Rio Grande do Sul
LSTM	Redes de memória de longo prazo - ( <i>Long short-term Memory</i> )
ML	<i>Machine learning</i>
MLP	<i>Multi-layer perceptron</i>
MLR	Regressão linear múltipla ( <i>Multiple linear regression</i> )
SEBRAE	Serviço Brasileiro de Apoio às Micro e Pequenas Empresas
SVM	<i>Support vector machine</i>
RF	Floresta aleatória ( <i>Randon forest</i> )
RNA	Rede neural artificial
RNN	Rede neural recorrente



## SUMÁRIO

<b>1</b>	<b>INTRODUÇÃO</b>	<b>19</b>
1.1	Motivação	21
1.2	Problema de Pesquisa e Hipótese	24
1.3	Objetivos	25
1.4	Organização do Documento	26
<b>2</b>	<b>REFERENCIAL TEÓRICO</b>	<b>27</b>
2.1	Fenologia da Macieira	27
2.1.1	Acumulação Térmica	28
2.2	Aprendizagem Profunda	29
2.3	Redes Neurais Recorrentes	31
2.3.1	Redes LSTM	33
2.4	Funções de Ativação	34
<b>3</b>	<b>TRABALHOS RELACIONADOS</b>	<b>37</b>
3.1	Metodologia para Seleção dos Trabalhos	37
3.2	Apresentação do Estado da Arte	38
3.3	Discussão	43
<b>4</b>	<b>PREDHARV: UM MODELO PARA PREDIÇÃO DA DATA DA COLHEITA DE MACIEIRAS</b>	<b>47</b>
4.1	Premissas e Decisões de Projeto	48
4.2	Arquitetura do Modelo	48
4.3	Estratégia de Predição	50
4.4	Conjunto de dados	54
4.4.1	Pré-processamento	56
4.4.2	Seleção de Variáveis de Entrada	59
4.4.3	Divisão do Conjunto de Dados em Treino e Teste	60
<b>5</b>	<b>METODOLOGIA DE AVALIAÇÃO</b>	<b>61</b>
5.1	Protótipo	61
5.2	Métricas de Avaliação	62
5.3	Cenários de Avaliação	63
5.3.1	Cenário 1: Validação <i>Walk-forward</i>	63
5.3.2	Cenário 2: Dados Reais x Predições	63
5.3.3	Cenário 3: Dados Sintéticos	64
5.3.4	Cenário 4: Comparativo com Outros Métodos	64
<b>6</b>	<b>RESULTADOS</b>	<b>67</b>
6.1	Cenário de Avaliação 1: Validação <i>Walk-forward</i>	67
6.2	Cenário de Avaliação 2: Dados Reais x Predições	67
6.3	Cenário de Avaliação 3: Dados Sintéticos	69
6.4	Comparativo de Outros Métodos Aplicados ao Problema	69
6.5	Discussão	71

<b>7</b>	<b>CONSIDERAÇÕES FINAIS</b>	<b>75</b>
<b>7.1</b>	<b>Contribuições</b>	<b>75</b>
<b>7.2</b>	<b>Limitações</b>	<b>76</b>
<b>7.3</b>	<b>Trabalhos Futuros</b>	<b>76</b>
<b>7.4</b>	<b>Trabalhos Publicados</b>	<b>76</b>
	<b>REFERÊNCIAS</b>	<b>77</b>
	<b>APÊNDICE A – INFORMAÇÕES COMPLEMENTARES</b>	<b>83</b>
<b>A.1</b>	<b>Tabela da duração dos eventos fenológicos da macieira cultivar Gala nas safras 2000 a 2018. Caçador, SC.</b>	<b>83</b>
<b>A.2</b>	<b>Tabela de valores e médias anuais do período de 2000 a 2018. Caçador, SC.</b>	<b>84</b>
<b>A.3</b>	<b>Tabela de valores e médias anuais do período de 2016 a 2020. Vacaria, RS.</b>	<b>85</b>
<b>A.4</b>	<b>Tabela de graus-dia acumulado com diferentes valores de temperatura base. EPAGRI Caçador, SC.</b>	<b>86</b>
<b>A.5</b>	<b>Modelo linear ajustado para estimar a duração do período da plena floração ao início da maturação em função de graus-dia acumulado. Dados: EPAGRI Caçador, SC.</b>	<b>87</b>

## 1 INTRODUÇÃO

A Inteligência Artificial (IA) é um campo da Ciência da Computação que se esforça para construir entidades inteligentes e também entendê-las (RUSSELL; NORVIG, 2021). Essas entidades inteligentes são interessantes e úteis por si mesmas e podem produzir muitos produtos significativos e impressionantes causando grande impacto em nossas vidas diárias. O uso de IA é capaz de contribuir para a solução de diversos problemas na atualidade. Ferramentas computacionais que usam IA como base conseguem realizar tarefas de forma bem mais eficiente superando humanos em diversas áreas.

O Aprendizado de Máquina, do inglês, *Machine Learning* (ML) é um campo de estudo definido por Arthur Lee Samuel, cientista da computação pioneiro no campo de aprendizado de máquina como: "Campo de estudo que dá aos computadores a habilidade de aprender sem serem explicitamente programados". O Aprendizado de Máquina é uma área de IA dedicada à criação de algoritmos e técnicas que permitem ao sistema aprender com as experiências no ambiente em que está situado. Fundamentalmente, de acordo com Data Science Academy (2022), ML é a utilização de algoritmos para extrair informações de dados brutos e representá-los através de algum tipo de modelo matemático. Segundo Alpaydin (2020), ML se preocupa com a questão de como construir programas de computador que melhoram automaticamente com a experiência. À medida que o agente aprende mais sobre o ambiente, ele é capaz de desempenhar melhor tarefas nesse ambiente. O interesse pelo ML cresceu na última década. O mundo está passando por uma transformação e vemos uma interação cada vez maior das aplicações de computador com os seres humanos. No início dos anos 2000, o poder computacional expandiu exponencialmente e o mercado viu uma "explosão" de técnicas computacionais que não eram possíveis antes disso. Os algoritmos de aprendizado de máquina provaram ser de grande valor prático em uma variedade de domínios de aplicação. ML é uma tecnologia onde os computadores tem a capacidade de aprender de acordo com as respostas esperadas por meio de associações de diferentes dados, os quais podem ser imagens, números e tudo que essa tecnologia possa identificar.

No cenário da agricultura de precisão, muitas iniciativas são observadas, que empregam diversas tecnologias para oferecer melhores serviços aos produtores, considerando variadas culturas. Abordagens que fazem uso de ML têm sido utilizadas em várias áreas incluindo, por exemplo, aplicações na indústria da construção em Xu et al. (2021), no campo da medicina em Garg e Mago (2021), na meteorologia em Schultz et al. (2021), e na bioquímica em Wang et al. (2020). No campo de pesquisa da agricultura, ML tem sido usado em uma variedade de aplicações, como pode ser observado em Liakos et al. (2018); Kamilaris e Prenafeta-Boldú (2018); Elavarasan et al. (2018); Tripathi e Maktedar (2020). A agro-tecnologia e a agricultura de precisão surgiram como novos campos científicos que usam abordagens intensivas em dados para aumentar a produtividade agrícola e minimizar seu impacto ambiental. Os dados gerados nas operações agrícolas modernas permitem um melhor entendimento do ambiente operacional

(uma interação da cultura dinâmica, solo e condições climáticas) e da operação, levando a uma tomada de decisão mais rápida (LIAKOS et al., 2018). Neste contexto, ML é uma importante ferramenta de apoio à tomada de decisão, auxiliando no planejamento, no manejo e na gestão, na previsão, na detecção de doenças e na qualidade da produção agrícola.

A agricultura de precisão está começando a mudar drasticamente a produção de alimentos no mundo todo. Existem soluções comerciais para inúmeras tarefas agrícolas, como o monitoramento de campos de cultivo, semeadura automática, colheita e empacotamento. No entanto, o domínio de aplicação da agricultura de precisão limitou-se principalmente às culturas de commodities, como trigo, arroz e milho (HÄNI; ROY; ISLER, 2018). O desenvolvimento de soluções para culturas especializadas, como frutas e vegetais, tem sido desafiador devido à complexa geometria dos pomares em comparação às plantações em fileiras. Embora o avanço tecnológico e a evolução da eletrônica tenha permitido um avanço nos sistemas de coleta, monitoramento, armazenamento e tratamento de dados, a fruticultura diferente da cultura de grãos e fibra, é altamente exigente no uso de mão de obra, apresentando uma série de restrições ao amplo uso de mecanização e tecnologias. Estas restrições acabam por gerar demandas específicas cujas soluções, em sua maior parte, ainda se encontram em fase de experimental (AGAPOMI. ASSOCIAÇÃO GAÚCHA DE PRODUTORES DE MAÇÃ, 2018).

A introdução da cultura da maçã no Brasil teve início nos anos 70 e desde então vem apresentando uma crescente expansão. O Brasil alcançou na safra de 2017/2018 a marca de 1,25 milhão de toneladas produzidas da fruta. Na região Sul a produção de maçãs tem uma participação significativa na economia. Segundo relatório da Associação Gaúcha de Produtores de Maçã - AGAPOMI a produção da safra de 2019/2020 no estado do Rio Grande do Sul foi de 485.357 toneladas (AGAPOMI. ASSOCIAÇÃO GAÚCHA DE PRODUTORES DE MAÇÃ, 2021). No Brasil, a maçã é a segunda fruta de clima temperado mais produzida, sendo os estados do Rio Grande do Sul e Santa Catarina os principais produtores da fruta no país (CEPA. CENTRO DE SOCIOECONOMIA E PLANEJAMENTO AGRÍCOLA, 2017). Nas últimas três décadas, a produção de maçãs teve um impressionante aumento de mais de 6.000%. De importador o país passou não apenas a abastecer todo o mercado interno, como também a exportar 15% de sua colheita (ABPM. ASSOCIAÇÃO BRASILEIRA DE PRODUTORES DE MAÇÃ, 2018). A maçã passou a ser a terceira fruta mais consumida atualmente e o consumidor, não só do Brasil mas em nível mundial, está mudando o seu comportamento em busca de hábitos saudáveis, o que fortalece ainda mais o segmento (KIST et al., 2018).

A maçã é o pseudofruto pomáceo da árvore macieira (*Malus pumila*). É uma frutífera típica de clima temperado, da família *Rosaceae*, com pequena exigência de frio, apresentando-se apta para produzir satisfatoriamente em condições de inverno brando, o que favorece o cultivo na região sul (SEBRAE. SERVIÇO BRASILEIRO DE APOIO ÀS MICRO E PEQUENAS EMPRESAS, 2018). A macieira é uma árvore caducifólia, que perde suas folhas sazonalmente. Atinge em média uma altura de 4 metros. As flores são produzidas na primavera. Os frutos amadurecem em brotos no final do verão ou outono, dependendo da variedade. Existem mais

de 7000 cultivares conhecidos. A Figura 1 mostra pomares de maçã, com cobertura de proteção anti-granizo (Figura 1 acima) e sem cobertura (Figura 1 abaixo), na cidade de Vacaria, região dos Campos de Cima da Serra, localizada no extremo nordeste do estado do Rio Grande do Sul.

Figura 1 – Pomares de maçã na cidade de Vacaria, RS.



Fonte: Imagens do autor

## 1.1 Motivação

A macieira sofre a ação de diversos fatores que podem exercer alguma influência de modo direto ou indireto sobre o crescimento, a produção e a qualidade da fruta. De acordo com Fioravanço e Santos (2013), entre as principais variáveis, destacam-se como mais impactantes: temperatura, precipitação, vento e radiação solar. Considerando a região de cultivo no Brasil e as exigências da macieira, a temperatura apresenta-se como uma das variáveis climáticas mais importantes (FIORAVANÇO; SANTOS, 2013).

O ramo da Ecologia que estuda os fenômenos periódicos dos seres vivos e suas relações com o ambiente é chamado de Fenologia (BERGAMASCHI, 2007). Segundo Petri et al. (2012), observações das alterações fenológicas têm sido intensamente estudadas e classificadas de acordo com seu desenvolvimento desde a dormência até o frutificação, e esse conhecimento é importante do ponto de vista prático para o manejo de algumas práticas culturais e tratamentos fitossanitários. O monitoramento da fenologia de plantas agrícolas é um entendimento crítico na agricultura de precisão. Melhorias vitais podem ser alcançadas com a detecção precisa da fenologia da planta que, pode resultar em melhoria no tempo de colheita, controle de pragas, previsão de produção, monitoramento, entre outros. Inferir informações fenológicas contribui para um melhor entendimento das relações entre produtividade, saúde da vegetação e condições ambientais. (YALCIN, 2017). Detectar mudanças na fenologia vegetal tornou-se um tema recorrente e a agricultura de precisão tem atraído cada vez mais a atenção de agricultores, governos e pesquisadores, uma vez que o monitoramento de plantas é uma das ferramentas importantes para o gerenciamento e otimização de recursos agrícolas (BROICH et al., 2015). As complexas interações entre o meio ambiente e as plantas representam uma fonte significativa de incerteza para produtores. O impacto do clima na produção pode mudar muito rapidamente e ter implicações significativas na produção de frutas. A incerteza em torno desses fatores (como temperatura, chuva, vento, geada, etc) representa custos para produtores, que são incapazes de planejar e gerenciar adequadamente os recursos pré e pós-colheita.

A previsão do momento da colheita é um desafio para desenvolver a produção de frutas de forma sustentável e reduzir o desperdício de alimentos (LEE et al., 2020). Frutos são perecíveis, de alto valor e sazonais, e os preços de venda são geralmente sensíveis ao tempo, o que torna as previsões de colheita extremamente valiosas para os produtores (MORRIS et al., 2017). Além dos fatores relacionados ao produto final, existe um custo relacionado a colheita da fruta. Durante esse período são contratados trabalhadores temporários para a realização da colheita nos pomares. A mão de obra implica além da remuneração pelo trabalho e encargos, em outros custos com infraestrutura de alojamentos e custos com alimentação. O atraso no período de colheita implica em mais tratamentos com agroquímicos. Outro ponto crítico que torna a previsão do momento da colheita valioso é o fato da fruta não resistir por muito tempo na árvore, após a maturação. Macieiras da cultivar *Royal Gala* apresentam rápida maturação no período de colheita e alta sensibilidade à queda de frutos na pré-colheita, podendo resultar em perdas significativas de produção se não forem colhidas no momento adequado, diminuindo a vida pós-colheita dos frutos. A queda dos frutos implica em redução no lucro final. Embora essa fruta possa ser aproveitada pela indústria o custo de venda comparado a uma fruta que foi colhida no momento adequado é muito menor.

Modelos de soma térmica, na forma de graus-dia tem sido ajustados para estimativas do acúmulo de calor. Segundo Bergamaschi (2007), estes modelos representam a integração das temperaturas efetivas para o crescimento das plantas, fixadas entre os limites inferior e superior. A soma de graus-dia que a planta necessita para completar parcial ou totalmente o ciclo tem

sido utilizada para caracterizar as fases fenológicas das plantas (SCHWARTZ et al., 2013). O conceito de graus-dia, considera a existência de uma temperatura base, o qual abaixo desta temperatura a planta não se desenvolve. Segundo Miller, Lanier e Brandt (2001), essa temperatura base (limite inferior) é determinado pela fisiologia da planta e é independente do método usado para calcular graus-dia.

Na literatura, encontramos diversos trabalhos que abordam o tema relacionado ao estudo da fenologia Putti e Petri (2002); Cardoso (2011); Chitu e Paltineanu (2020). Em Petri et al. (2012), Rivero et al. (2017) e Darbyshire et al. (2017), entre outros apresentados com detalhes no Capítulo 3, observamos trabalhos com abordagens no sentido de analisar, ou ainda fazer previsões da fenologia em função do clima ou de outros fatores que alteram a ocorrência dos fenômenos. Em sua maioria estes trabalhos utilizam abordagens estatísticas como análises de regressão e correlação. Os trabalhos como de Lee et al. (2020) e Safa et al. (2004) são relacionados a previsão de colheita e o foco é na previsão de produtividade. Observamos uma quantidade muito pequena de trabalhos que enfocam a previsão da época ou período de colheita como em Chen et al. (2020); McCormick et al. (2020); Blazek e Pistekova (2017). Abordagens fazendo uso de métodos de ML estão presentes em muitos trabalhos. Alguns autores aplicaram redes neurais para predição de estágios fenológicos, como em Yazdanpanah, Ohadi e Soleimani (2010) e Safa et al. (2004). Outros trabalhos como de Dai et al. (2019) e Czernecki, Nowosad e Jablonska (2018) apresentam abordagens com outros métodos de ML. Mais detalhes podem ser observados na Tabela 1 do Capítulo 3.

Neste trabalho estamos interessados em fazer previsões em relação a data de colheita de macieiras, dadas as condições climáticas esperadas para o período, baseado no conhecimento implícito nos dados históricos da fenologia e meteorológicos. Tecnicamente, trata-se de séries temporais modeladas como um problema de aprendizado supervisionado. Abordar séries temporais como uma tarefa de aprendizado supervisionado, permite que sejam utilizados uma série de algoritmos de ML existentes, abrindo-se um leque de possibilidades.

Uma série temporal é um conjunto de observações ordenadas no tempo, e devido a essa existência de uma dependência temporal entre as observações das séries consideramos o uso de Redes Neurais Recorrentes (RNN) por serem adequadas a análise de sequências temporais. Segundo McCormick et al. (2020) essas redes têm o potencial de capturar o impacto dependente da sequência de estímulos ambientais, como um período de clima frio, no desenvolvimento da planta. Algumas iniciativas, como em Yazdanpanah, Ohadi e Soleimani (2010), fazem uso de redes neurais para predição de estágios fenológicos. O problema das redes neurais tradicionais, chamadas *feed-forward*, é que se assume uma certa independência dos valores da amostra, ou seja, o fato dos dados serem ordenados sequencialmente não são considerados, de certa forma ignorando a temporalidade existente nas séries. Embora o uso de redes neurais artificiais para modelar problemas de fenologia não seja novo, o uso de redes recorrentes em problemas de predição de período colheita foi pouco explorado até então.

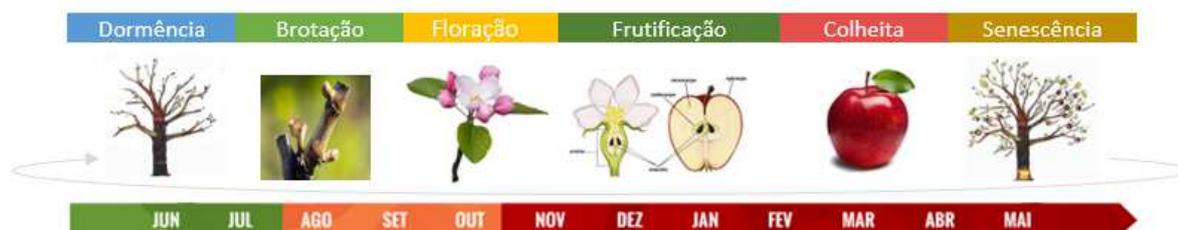
Analisando os trabalhos selecionados na revisão da literatura, encontramos trabalhos com

técnicas de inteligência computacional, mas não encontramos nenhuma abordagem semelhante à proposta neste trabalho, que utiliza redes LSTM para prever a data de colheita da maçã. Diante da lacuna identificada na revisão da literatura, o desenvolvimento deste trabalho contribui com a área de fruticultura, possibilitando antecipar informações sobre a data de colheita. Considera-se que antecipar informações sobre a data de colheita, traz benefícios ao produtor, possibilitando um melhor planejamento, economia financeira, evitando custos e conseqüentemente melhoria na produtividade. É uma importante ferramenta de apoio à tomada de decisão, gerando economia financeira para o fruticultor. O desenvolvimento deste trabalho contribui ainda com uma metodologia, utilizando ML, com uma abordagem multivariada que faz uso de dados históricos da fenologia e meteorológicos para realizar a predição da data da colheita de árvores macieiras, que torna mais efetiva a capacidade preditiva, possibilitando a prospecção de cenários futuros.

## 1.2 Problema de Pesquisa e Hipótese

O ciclo anual da macieira (Figura 2) compreende dois períodos distintos: repouso e crescimento vegetativo. Durante o período de repouso, também chamado de dormência, a planta não mostra crescimento visual, no entanto, as atividades metabólicas continuam, com intensidade reduzida, o que lhe permite resistir a baixas temperaturas, mesmo abaixo de zero graus centígrados. As baixas temperaturas de outono e inverno são os principais fatores ambientais que influenciam a entrada e saída da dormência. Quando a condição de frio é satisfeita, a planta precisa de acúmulo de calor para iniciar o período de crescimento vegetativo e produtivo. No início da primavera, as gemas passam por uma série de sucessivos estádios fenológicos, sendo sua caracterização importante para a realização de determinadas práticas. A evolução dos estádios fenológicos da macieira, durante a fase vegetativa, são influenciados pelas condições climáticas e tem influência na produtividade e qualidade da produção. A disponibilidade térmica tem influência direta na fenologia da macieira e de outras plantas. Temperaturas mais elevadas aceleram o desenvolvimento vegetal, enquanto que baixas temperaturas prolongam o ciclo (BERGAMASCHI, 2007).

Figura 2 – Ciclo anual da macieira



Fonte: O próprio autor

A questão de pesquisa que trata este trabalho pode ser resumida com a seguinte pergunta:

*Como estimar a data da colheita de macieiras a partir do estágio fenológico da plena floração baseado exclusivamente em dados históricos da fenologia e parâmetros climáticos de temperatura usando ferramentas de Machine Learning?*

Fases fenológicas são as transformações de órgãos nas plantas, como germinação, brotação, florescimento, maturação, etc. Os chamados estádios fenológicos, são fases específicas ou subdivisões de um subperíodo. Segundo Bergamaschi (2007), os estádios podem coincidir com as fases, quando envolvem mudanças importantes, ou apenas caracterizam uma condição qualquer da planta. Os dados históricos da fenologia são registros dos acontecimentos fenológicos que foram influenciados pelos parâmetros climáticos.

Segundo Bergamaschi (2007) a fenologia de muitas espécies responde à variação sazonal da duração do dia (fotoperíodo). Essa resposta das plantas ao fotoperíodo é chamada fotoperiodismo. Árvores macieiras são classificadas como plantas fotoneutras, insensíveis ou de muito pouca resposta ao fotoperíodo. A temperatura é reconhecida como sendo a principal variável que regula o tempo de ocorrência dos fenômenos fenológicos (BERGAMASCHI, 2007). De acordo com Miller, Lanier e Brandt (2001), embora a precipitação também possa influenciar a fenologia, na maioria das culturas, incluindo as espécies frutíferas, a soma térmica tem relação direta com o desenvolvimento das plantas. Supõe-se que existe uma relação direta entre a temperatura e o desenvolvimento da planta, uma vez que não haja limitações para outros fatores.

Assumimos como hipótese básica, que norteou o desenvolvimento deste trabalho, a seguinte sentença:

*A data de colheita da macieira pode ser estimada através de modelos multivariados baseados em soma térmica, utilizando métodos de ML com base em dados históricos da fenologia e parâmetros climáticos de temperatura, podendo melhorar a acurácia das predições, explicando melhor a relação entre as variáveis de temperatura e a quantidade de calor necessária para a planta completar seu ciclo.*

### **1.3 Objetivos**

Este estudo propõe um modelo de ML para estimar o período de ocorrência do início da colheita de árvores macieiras. A seguinte declaração define o objetivo principal desta pesquisa:

*Desenvolver um modelo de ML para prever a data de ocorrência do início da colheita de árvores macieiras, baseado nas séries históricas da fenologia e dados meteorológicos de temperatura, dadas as condições climáticas esperadas para o período.*

Como complemento ao objetivo principal, apoiando o desenvolvimento do modelo e aplicações necessárias para a sua validação, os seguintes objetivos secundários foram identificados:

- (i) Melhorar a acurácia da predição dos modelos lineares baseados em soma térmica, predizendo a duração do período em dias do calendário;
- (ii) Realizar o levantamento dos fundamentos teóricos e evidências que justifiquem o uso de determinados parâmetros e entrada multivariada;
- (iii) Explorar a utilização de redes neurais recorrentes que permitam capturar o impacto dependente da sequência de estímulos ambientais considerando a dependência temporal entre as observações das series;
- (iv) Tornar o uso do modelo preditivo mais eficaz através de uma metodologia que possibilite a prospecção de cenários;
- (v) Permitir flexibilidade de utilização do modelo, como forma de diminuir a incerteza da predição, permitindo que possam ser realizadas predições de todo o período ou de subperíodos.

#### **1.4 Organização do Documento**

O restante deste documento está estruturado da seguinte forma: o Capítulo 2 apresenta o referencial teórico incluindo os conceitos fundamentais abordados por esta pesquisa; em seguida, o Capítulo 3 apresenta a revisão da literatura com os trabalhos que foram selecionados na literatura e que de alguma forma estão relacionados ao tema deste trabalho; o Capítulo 4 apresenta o núcleo central deste documento, que é o modelo proposto, as premissas e decisões de projeto, os detalhes da arquitetura do modelo, a estratégia de predição e os detalhes sobre o conjunto de dados; o Capítulo 5 apresenta a metodologia de avaliação do modelo, incluindo, os detalhes sobre o protótipo, as métricas de avaliação e a descrição dos cenários de avaliação do modelo; o Capítulo 6 apresenta os resultados obtidos com o modelo submetido aos cenários de avaliação propostos; e, finalizando, o Capítulo 7 apresenta as considerações finais, contribuições, limitações, trabalhos futuros e os trabalhos publicados.

## 2 REFERENCIAL TEÓRICO

Este capítulo apresenta os conceitos essenciais e as terminologias utilizadas neste trabalho. O capítulo está segmentado em quatro seções. Na Seção 2.1 são tratados os conceitos relacionados a fenologia da macieira e sobre acumulação térmica. Na Seção 2.2 são apresentados os conceitos sobre aprendizagem profunda e redes neurais. Na Seção 2.3 são tratados dos conceitos relacionados as RNNs e redes LSTM. Por fim, na seção 2.4, é apresentado um breve panorama sobre diferentes funções de ativação.

### 2.1 Fenologia da Macieira

Fenologia é o estudo dos eventos biológicos periódicos influenciados pelo meio ambiente, especialmente mudanças de temperatura causadas pelo tempo e pelo clima (SCHWARTZ et al., 2013). De acordo com Bergamaschi (2007), as fases fenológicas são as transformações de órgãos nas plantas, como germinação, brotação, florescimento, desfolhação, maturação, etc. Algumas fases são facilmente observadas e outras somente perceptíveis por exames como microscopia ou análises químicas. Os chamados estádios fenológicos são fases específicas ou subdivisões de um subperíodo. Estes estádios são identificados por letras e subdivididos em subestádios. A Figura 3 mostra graficamente a sequência de estádios fenológicos da macieira.

Figura 3 – Sequência de estádios de desenvolvimento fenológico da macieira

Gema dormente	Gemas inchadas	Ponta verde	Meia polegada verde		Botão verde	Botão rosado	Início floração	
A	B	C	C3	D	D2	E	E2	F
								
Floração plena	Final da floração	Queda de pétalas		Frutificação	Frutos verdes	Frutos maduros		
F2	G	H		I	J	L		
								

Fonte: Adaptado de (PUTTI; PETRI, 2002).

O monitoramento da fenologia de plantas agrícolas é um entendimento crítico na agricultura

de precisão. Melhorias vitais podem ser alcançadas com a detecção precisa da fenologia da planta que, pode resultar em melhoria no tempo de colheita, controle de pragas, previsão de produção, monitoramento, alerta de desastres, entre outros. Inferir informações fenológicas contribui para um melhor entendimento das relações entre produtividade, saúde da vegetação e condições ambientais (YALCIN, 2017).

A macieira é uma planta caducifólia que perde as suas folhas no inverno. Durante este período entra em repouso vegetativo e requer um determinado acúmulo de horas de frio. Segundo Petri (2006), ao iniciar o período de repouso vegetativo ou dormência, as plantas limitam ou cessam seu crescimento, de forma a permitir a sobrevivência em baixas temperaturas. Quando a macieira tem sua exigência de frio plenamente satisfeita, ocorre um período de desenvolvimento vegetativo e produtivo. Em regiões como no sul do Brasil, a brotação das macieiras ocorre no final do inverno ou início da primavera, dependendo da cultivar, à medida que ocorre aumento da temperatura do ar (CARDOSO, 2011).

A duração dos eventos fenológicos podem variar entre anos devido à variação dos elementos climáticos, principalmente a temperatura. Para uma mesma espécie ou variedade, a duração dos estádios fenológicos varia entre anos e locais, dependendo das condições térmicas do ambiente. De acordo com Monteiro et al. (2009), as variáveis climatológicas que mais influenciam a produtividade da macieira são: temperatura, precipitação pluviométrica, luminosidade, vento e radiação solar. A temperatura e a precipitação são os fatores mais significativos, e dependendo da região cultivada podem ser limitantes. Segundo Cardoso (2011), as cultivares da macieira diferem amplamente no tempo de maturação e os elementos climáticos que tem maior influência, são a temperatura do ar e a radiação solar. As baixas temperaturas de primavera e verão retardam a maturação. A exposição à radiação solar tem influencia na coloração e no teor de açúcar da maçã.

### 2.1.1 Acumulação Térmica

De acordo com Bergamaschi (2007), a disponibilidade térmica tem influência direta no crescimento e no desenvolvimento das plantas. Temperaturas mais elevadas aceleram o desenvolvimento vegetal, enquanto que baixas temperaturas prolongam o ciclo (BERGAMASCHI, 2007). Em geral, a fenologia das plantas responde à temperatura do ar na forma de soma térmica. A acumulação térmica, é um índice utilizado para estabelecer a relação entre a quantidade de calor que a planta necessita para seu desenvolvimento e a temperatura média diária. Este acúmulo é geralmente expresso em graus-dia ( $GD$ ), e representa a soma térmica acima da temperatura base mínima ( $T_{base}$ ).

Vários modelos ou expressões foram propostos para o cálculo dos  $GD$  (ARNOLD, 1959; VILLA NOVA et al., 1972). O conceito de  $GD$  admite que há uma temperatura base mínima e que abaixo desse limite, o crescimento e o desenvolvimento da planta são interrompidos ou extremamente reduzidos. Segundo Petri et al. (2012), o processo fisiológico e as funções da planta

ocorrem sob limites térmicos no seu desenvolvimento ambiental. Além disso, assume-se uma relação direta entre temperatura e desenvolvimento da planta, uma vez que não há limitações para outros fatores (BRUNINI et al., 1976). O índice da unidade de calor  $GD$  pode ser calculado usando a Equação (2.1), onde  $T_{max_i}$  e  $T_{min_i}$  são a temperatura diária máxima e mínima, e  $T_{base}$  é a temperatura base. Com base no valor  $GD$ , a acumulação térmica de um determinado período, chamado de graus-dia cumulativo  $GDA$ , pode ser calculado usando a Equação (2.2).

Diferentes valores atribuídos à temperatura base podem ser encontrados na literatura. Neste trabalho utilizamos a temperatura base de 2 °C. Realizamos testes com temperaturas de 2 °C, 4 °C e 10 °C (Ver Anexo A.3). O melhor resultado foi obtido com a temperatura de 2 °C, onde apresentou um coeficiente de variação de 7,89, sendo 17% inferior aos resultados obtidos com 10 °C e 1,9 % inferior aos resultados obtidos com 4 °C. Estes resultados corroboram com os resultados obtidos em (PUTTI; PETRI, 2002).

$$GD_i = \frac{T_{max_i} + T_{min_i}}{2} - T_{base} \quad (2.1)$$

$$GDA = \sum_{i=1}^n GD_i \quad (2.2)$$

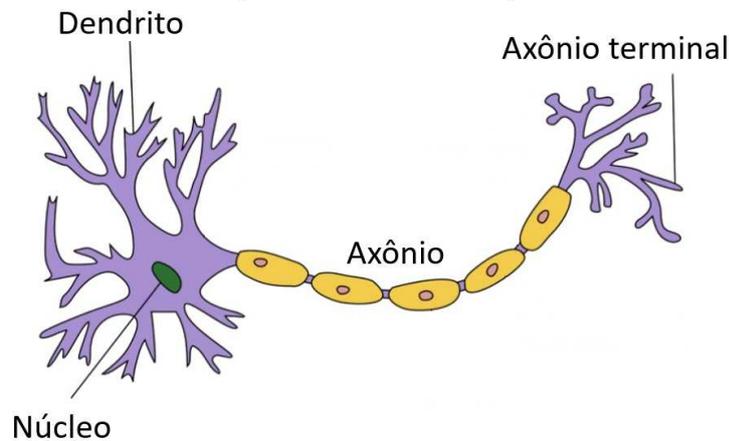
## 2.2 Aprendizagem Profunda

Aprendizagem Profunda (AP) ou *Deep Learning*, é uma sub-área da Aprendizagem de Máquina, que emprega algoritmos para processar dados e imitar o processamento feito pelo cérebro humano. Baseada no conceito de redes neurais artificiais, ou sistemas computacionais que imitam a maneira como o cérebro humano funciona, a aprendizagem profunda é responsável por avanços recentes em visão computacional, reconhecimento de fala, processamento de linguagem natural e reconhecimento de áudio (GOODFELLOW; BENGIO; COURVILLE, 2016).

Neurônios biológicos são os principais componentes do cérebro humano. Um neurônio consiste de um corpo celular, dendritos e um axônio. Ele processa e transmite informações para outros neurônios emitindo sinais elétricos. Cada neurônio recebe sinais de entrada de seus dendritos e produz sinais de saída ao longo de seu axônio. O axônio se ramifica e se conecta via sinapses aos dendritos de outros neurônios. A Figura 4 apresenta uma visão simplificada da estrutura típica de um neurônio. Cada sinapse tem uma força que é aprendida e controla a força de influência de um neurônio em outro. Os dendritos levam os sinais para o corpo do neurônio alvo, onde são somados. Se a soma final estiver acima de um certo limite, o neurônio será ativado, enviando um pico ao longo de seu axônio (DATA SCIENCE ACADEMY, 2022).

Os neurônios artificiais são inspirados nos neurônios biológicos e tentam formular a estrutura apresentada na Figura 4 em uma forma computacional. Um neurônio artificial tem um número finito de entradas com pesos associados a elas e uma função de ativação (também chamada de função de transferência). A saída do neurônio é o resultado da função de ativação

Figura 4 – Visão simplificada da estrutura típica de um neurônio



Fonte: Adaptado de (JAROSZ, 2017)

aplicada à soma ponderada das entradas. Neurônios artificiais são conectados uns aos outros para formar redes neurais artificiais (RNAs) comumente referenciadas apenas como "rede neural".

Segundo Haykin (2008), as redes neurais artificiais têm raízes em disciplinas como neurociência, matemática, estatística, física, ciência da computação e engenharia. Suas aplicações podem ser encontradas em campos tão diversos quanto modelagem, análise de séries temporais, reconhecimento de padrões, processamento de sinais e controle. O autor fornece as bases para o entendimento das redes neurais reconhecendo a natureza multidisciplinar do tema.

Na definição de Haykin (2008):

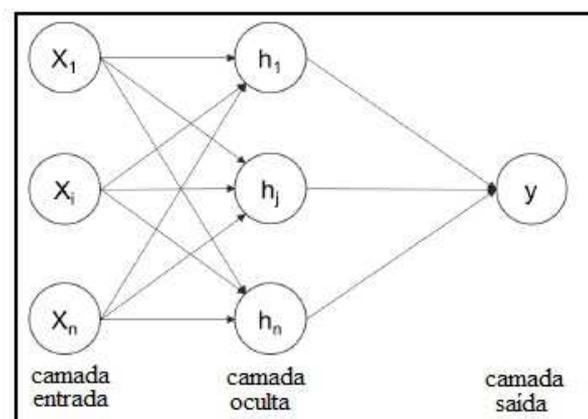
"Uma rede neural é um processador distribuído maciçamente paralelo composto de unidades de processamento simples que tem uma propensão natural para armazenar conhecimento experiencial e torná-lo disponível para uso. Ele se assemelha ao cérebro em dois aspectos: 1. O conhecimento é adquirido pela rede de seu ambiente por meio de um processo de aprendizagem. 2. As intensidades de conexão do interneurônio, conhecidas como pesos sinápticos, são usadas para armazenar o conhecimento adquirido.

Uma RNA é uma coleção de neurônios artificiais e funções de ativação, inspirada no comportamento do cérebro humano, definida como um modelo que adquire conhecimento através da experiência. A Figura 5, apresenta um diagrama de uma RNA. Os neurônios artificiais estão conectados para formar camadas ou unidades ocultas. As unidades ocultas formam a base não linear que mapeia as camadas de entrada para as camadas de saída em um espaço de menor dimensão, o que também é chamado de redes neurais artificiais. Os valores de peso e de polarização, juntamente com a arquitetura, são chamados de modelo (SHANMUGAMANI, 2018).

O processo de treinamento da rede determina os valores de peso. Os valores do modelo são inicializados com valores aleatórios durante o início do treinamento. O erro é calculado usando

uma função de perda. Com base na perda computada, os pesos são ajustados a cada etapa. O treinamento é interrompido quando o erro não pode ser reduzido. No treinamento de uma RNA envolve vários parâmetros que precisam ser otimizados. O procedimento de atualização dos pesos é chamado de retropropagação (*backpropagation*). O procedimento para minimizar o erro é chamado de otimização. O número de camadas ocultas e perceptrons pode ser definido conforme o problema que está sendo tratado.

Figura 5 – Diagrama de uma rede neural artificial



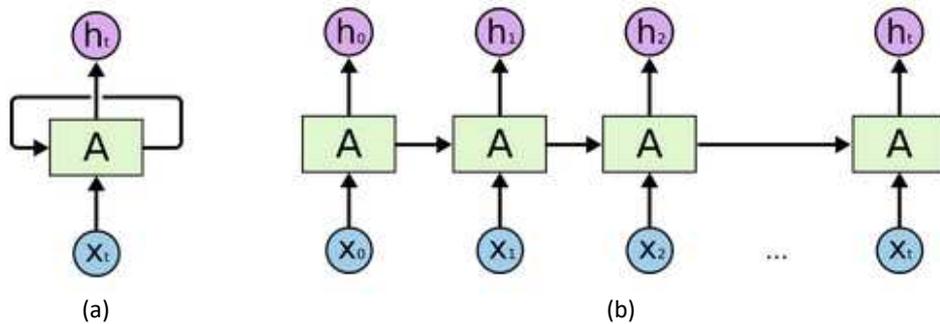
Fonte: Adaptado de (SHANMUGAMANI, 2018)

### 2.3 Redes Neurais Recorrentes

Uma Rede Neural Recorrente (RNN) é um tipo particular de RNA que possui loops (ver Figura 6), permitindo que as informações persistam. Uma rede neural recorrente se distingue de uma rede neural *feedforward* por ter pelo menos um *loop* de *feedback*. A presença de loops de feedback, tem um impacto profundo na capacidade de aprendizagem da rede e em seu desempenho (HAYKIN, 2008). Na Figura 6a, o digrama representa uma entrada  $X_t$  ao qual a rede neural  $A$  produz uma saída  $h_t$ . Um loop permite que as informações sejam passadas de uma etapa da rede para a próxima. Uma RNN pode ser considerada como várias cópias da mesma rede. A Figura 6b representa através de um diagrama a rede  $A$  desdobrada, com cada uma dessas cópias passando uma mensagem a um sucessor.

Em uma implementação de uma RNA tradicional, chamada *feedforward*, é assumida uma certa independência dos valores da amostra, ou seja, o fato de os dados serem ordenados sequencialmente não é considerado, e o aspecto temporal existente na série é de certa forma ignorado (GOODFELLOW; BENGIO; COURVILLE, 2016). Uma série temporal é um conjunto de observações ordenadas no tempo, e devido à existência de uma dependência temporal entre as observações da séries, o uso de RNN se torna adequado para problemas com essa característica. Segundo Data Science Academy (2022), as redes recorrentes são um tipo de RNA projetada para reconhecer padrões em sequências de dados, como texto, genomas, caligrafia, palavra fa-

Figura 6 – (a) Representação de uma RNN com *loops*; (b) Representação de uma RNN desdobrada

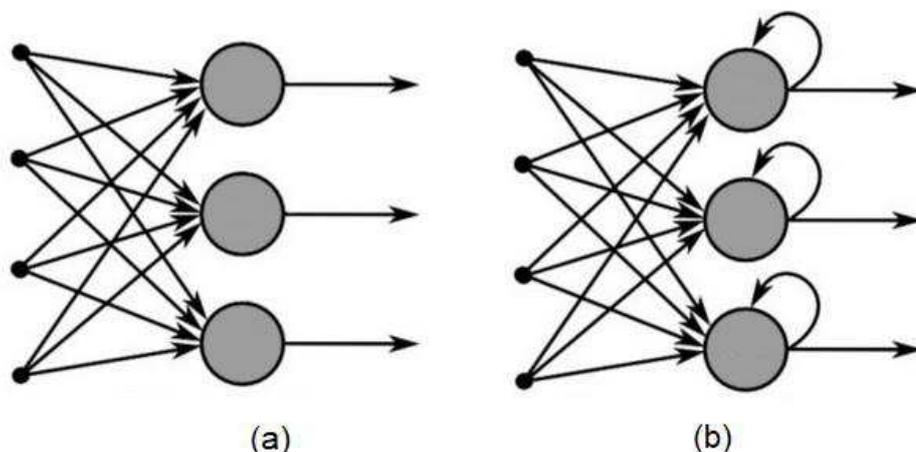


Fonte: Adaptado de Olah (2015)

lada ou dados de séries numéricas. Esses algoritmos consideram tempo, sequência e dimensão temporal.

As RNNs, tomam como entrada não apenas o exemplo de entrada atual, mas também o que perceberam anteriormente no tempo. A Figura 7 apresenta um diagrama com a representação de uma rede *feedforward* em 7a, e de uma RNN em 7b. As RNNs são diferenciadas das redes *feedforward* pelo loop de *feedback* conectado às suas decisões anteriores, ingerindo suas próprias saídas. Costuma-se dizer que as redes recorrentes têm memória, e essa adição de memória tem uma finalidade: há informações na própria sequência e as RNN conseguem utilizar essa informação para executar tarefas, o que a diferencia das redes *feedforward*. Essa informação sequencial é preservada no estado oculto da RNN, que consegue passar por muitas etapas de tempo à medida que ela avança no processamento de cada novo exemplo. Essas correlações entre eventos são separadas por muitos momentos, e essas correlações são chamadas de “dependências de longo prazo”, porque um evento no tempo depende e é uma função de um ou mais eventos que vieram antes (DATA SCIENCE ACADEMY, 2022).

Figura 7 – (a) Representação de uma rede *feedforward*; (b) Representação de uma RNN

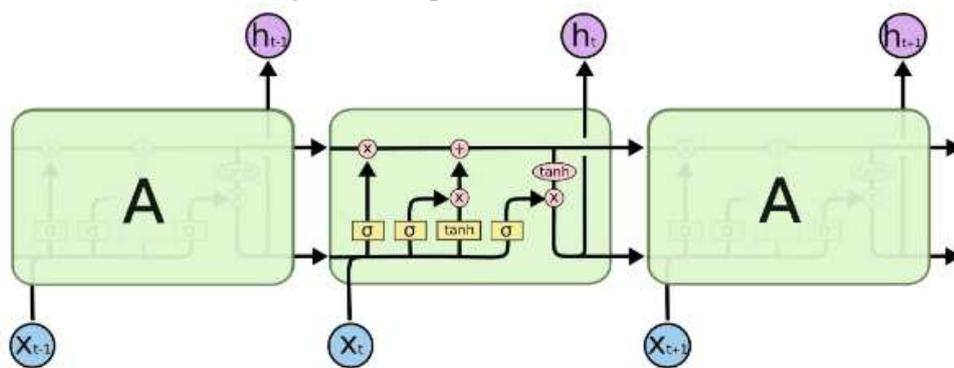


Fonte: Adaptado de Data Science Academy (2022)

### 2.3.1 Redes LSTM

Redes de memória de longo prazo, do inglês *Long short-term memory* - geralmente chamadas apenas de “LSTM” - são um tipo especial de RNN. Elas foram introduzidas pela primeira vez por Hochreiter e Schmidhuber (1997). As redes LSTM foram projetadas para evitar a dependência de longo prazo, funcionando muito bem em uma ampla variedade de problemas e amplamente utilizadas atualmente. A ideia central por trás da arquitetura LSTM é uma memória continuamente atualizada. Consiste em uma cadeia de módulos que se repetem, e ao invés de ter apenas uma única camada de rede neural, como nas RNNs padrão, nas redes LSTMs existem quatro camadas que interagem. A Figura 8 apresenta visualmente este conceito.

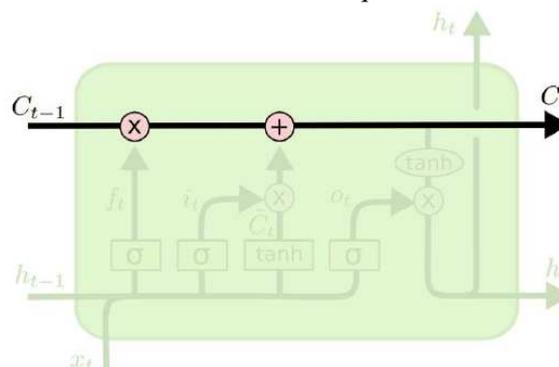
Figura 8 – Arquitetura da rede LSTM



Fonte: Olah (2015)

Na Figura 8 a entrada da rede é representada por " $x_t$ " e a saída por " $h_t$ ". As caixas amarelas são camadas de rede neural e os círculos rosa representam operações pontuais, como adição de vetores. A principal ideia das redes LSTM é criar uma representação do estado da célula " $C_t$ ", que corresponde a linha horizontal na parte superior do diagrama na Figura 9. Esse estado da célula percorre toda a cadeia da célula, sofrendo apenas algumas interações lineares, fazendo com que a informação possa fluir sem muitas alterações.

Figura 9 – O estado da célula da arquitetura da rede LSTM



Fonte: Olah (2015)

A rede LSTM tem a capacidade de remover ou adicionar informações ao estado da célula por meio de *gates* (portões). Estes portões são uma forma opcional de permitir a passagem de informação, controlando a passagem entre as etapas da rede, não limitando a informação apenas a etapa imediatamente anterior. Dessa forma, a rede pode aprender com informações de etapas muito anteriores que podem influenciar na predição da etapa atual da série. Esses portões são: *forget gate* (portão de esquecimento), *input gate* (portão de entrada) e *output gate* (portão de saída). O portão de esquecimento decide quais informações serão mantidas e quais serão descartadas dos estados anteriores. O portão de entrada decide quais informações devem ser atualizadas e o portão de saída é baseado no estado atual da célula e decide quais informações devem ser produzidas.

## 2.4 Funções de Ativação

As redes neurais são estruturas que promovem transformações matemáticas nos dados que recebem para processar. Em cada neurônio de cada camada, as redes multiplicam o valor de entrada pelo peso do neurônio correspondente, somam com o *bias* associado ao neurônio, e passam esse valor adiante. O *bias* é um elemento que serve para aumentar o grau de liberdade dos ajustes dos pesos. Essas operações são lineares, ou seja, por mais complexa que seja a rede neural, ela só pode captar relações lineares entre as variáveis de entrada e a variável de saída. Para torná-las capazes de modelar também relações não-lineares, os resultados de saída de cada camada passam por funções chamadas funções de ativação. A função de ativação é um importante componente de uma RNA. É um componente matemático incluído na estrutura de redes neurais artificiais a fim de permitir a solução de problemas complexos. Existem diversos tipos de funções de ativação. Dependendo das propriedades do problema, diferentes funções podem ser utilizadas. A seguir são descritas algumas das funções e suas características:

- Ativação linear - Geralmente utilizada nas camadas de saída em redes neurais de regressão, é a função de ativação mais básica utilizada porque não altera a saída do neurônio.
- Sigmóide - A função sigmoide é uma das funções de ativação mais comuns usadas na construção de redes neurais. É definida como uma função estritamente crescente que exibe um equilíbrio elegante entre o comportamento linear e não linear. Um exemplo da função Sigmóide é a função Logística. A função Logística produz valores no intervalo entre 0 e 1.
- Tangente hiperbólica - A função de ativação Tangente Hiperbólica é também muito utilizada em redes neurais. A função produz saídas cujos valores devem estar entre -1 e 1.
- Softmax - A função de ativação Softmax é bastante usada em redes neurais de classificação. Ela força a saída de uma rede neural a representar a probabilidade dos dados serem de uma das classes definidas.

- ReLU - A função é uma abreviação para *Rectified Linear Unit*, ou unidade linear retificada. A função retorna 0 para todos os valores negativos, e o próprio valor para valores positivos.
- Leaky ReLU - É uma modificação da função ReLU, que ao invés de zerar os valores negativos, aplica um fator de divisão, tornando-os próximos de zero.



### 3 TRABALHOS RELACIONADOS

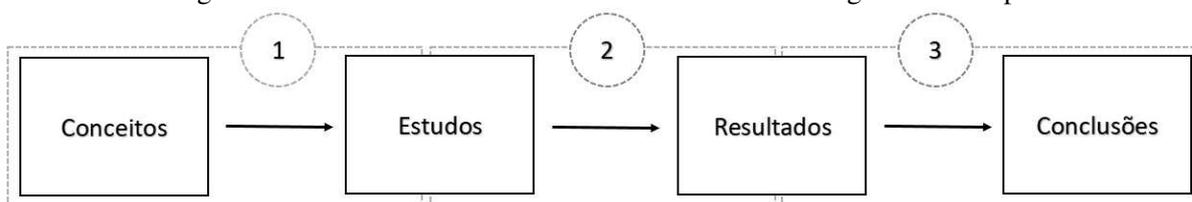
Este capítulo apresenta os trabalhos que foram selecionados na literatura e que de alguma forma estão relacionados ao tema deste trabalho. A Seção 3.1 descreve a metodologia adotada para seleção de trabalhos. Em seguida, a Seção 3.2 descreve os artigos, seus objetivos e abordagens. Por fim, a Seção 3.3 apresenta uma análise dos artigos selecionados e discute os aspectos considerados relevantes.

#### 3.1 Metodologia para Seleção dos Trabalhos

A revisão de literatura deste trabalho foi realizada utilizando algumas das diretrizes para realização de revisões sistemáticas propostas por Kitchenham (2004) e Biolchini et al. (2005). Seu objetivo é identificar os trabalhos existentes e fornecer uma visão geral das estratégias utilizadas para predição de estádios fenológicos e de colheita, e ainda técnicas de ML utilizadas para problemas relacionados ao tema deste trabalho.

Um protocolo (ver Figura 10) foi estabelecido para identificar, avaliar e interpretar os resultados relevantes para o escopo da pesquisa.

Figura 10 – Protocolo da revisão da literatura com abordagem em 3 etapas



Fonte: Adaptado de Biolchini et. al. 2005

A primeira etapa da pesquisa parte dos conceitos que explicitamente e formalmente representam o tema em questão, e vai para os estudos, material que potencialmente contém as informações que podem fornecer evidências sobre o tópico específico da investigação. A segunda etapa, parte desses estudos. Seus conteúdos são comparados entre si levando a resultados, que podem representar o surgimento de um novo tipo de evidência. A terceira e última etapa parte destes resultados, através de um processo de análise e síntese dos novos arranjos de dados que são possíveis através desta metodologia, para as conclusões. Estas conclusões implicam na aquisição de novos conhecimentos sobre o ponto em questão, e também como apoio a decisões relacionadas ao tema.

Para identificar os artigos acerca do assunto, realizou-se buscas nas seguintes bases de dados: ACM (Association for Computing Machinery) <sup>1</sup>, EBSCO <sup>2</sup>, Elsevier Science Direct <sup>3</sup>,

<sup>1</sup><https://www.acm.org/>

<sup>2</sup><https://www.ebsco.com/>

<sup>3</sup><https://www.sciencedirect.com/>

Elsevier Scopus <sup>4</sup>, Google Acadêmico <sup>5</sup>, IEEE Xplore Biblioteca Digital <sup>6</sup>, Scielo <sup>7</sup>, Springer <sup>8</sup>, Web of Science <sup>9</sup> e Portal de periódicos CAPES <sup>10</sup>.

Para o levantamento de artigos nas bases de dados, foram utilizadas as seguintes palavras-chave: "*Harvest forecast*"; "*Harvest prediction*", "*Machine learning*", "*Artificial intelligence*", "*Apple*".

Para seleção dos artigos foi considerado o intervalo de tempo das publicações entre 2001 a 2021, ou seja, nos últimos 20 anos. Os estudos considerados nesta revisão foram publicados em inglês ou português e publicação em periódicos, desconsiderando capítulos de livros, anais e resumos de eventos e seminários.

Após a consulta às bases de dados e a aplicação das estratégias de busca, foram eliminados alguns artigos de acordo com os seguintes critérios de exclusão: artigos em duplicidade, artigos com foco fora do escopo e artigos com abordagem ou método não relacionado ao escopo.

Com a análise das publicações e a aplicação dos critérios de exclusão, tendo em vista os artigos alinhados com o contexto da pesquisa, o conjunto final dos trabalhos selecionados é apresentado na Seção 3.2

### 3.2 Apresentação do Estado da Arte

O cenário da agricultura apresenta muitas iniciativas que empregam diversas tecnologias para oferecer melhores serviços aos produtores, considerando variadas culturas. Os trabalhos a seguir apresentam abordagens relacionadas ao tema desta pesquisa.

Em Cardoso (2011) foram ajustados modelos para estimar a ocorrência de estádios fenológicos de macieiras "*Royal Gala*" e "*Fuji Suprema*" em função do clima, na região produtora de Vacaria. Dados fenológicos foram obtidos em seis pomares, nos ciclos de 2003/2004 a 2009/2010. Foram ajustados modelos para estimar o número de dias entre o tratamento de quebra de dormência e eventos fenológicos posteriores, em função de horas de frio e temperatura do ar de cada período. Os dados foram tratados por uma abordagem estatística e os resultados demonstraram que é possível estimar o número de dias entre o tratamento de quebra de dormência e eventos fenológicos posteriores em função de graus-dia acumulados, e que as estimativas são mais precisas com o emprego de modelos que utilizam como variáveis preditoras graus-dia acumulados e temperaturas mínimas e médias do ar. Petri et al. (2012) publicaram um estudo da fenologia da maçã em condições de clima subtropical, na região de Caçador, Santa Catarina, Brazil. O estudo utiliza dados da fenologia e climáticos, usando uma abordagem estatística, apresenta importantes conclusões sobre os impactos na produção de frutas em condições de

---

<sup>4</sup><https://www.scopus.com>

<sup>5</sup><https://scholar.google.com.br/>

<sup>6</sup><https://ieeexplore.ieee.org/>

<sup>7</sup><https://www.scielo.org/>

<sup>8</sup><https://www.springer.com/br/>

<sup>9</sup><https://clarivate.com/>

<sup>10</sup><https://www.periodicos.capes.gov.br/>

clima subtropical.

Rivero et al. (2017) avaliaram o impacto do aquecimento global no tempo de floração das macieiras no clima nórdico. O trabalho buscou analisar a relação das fenofases o acúmulo de calor, usando análises estatísticas, de correlação e regressão. Os resultados mostram que, no clima nórdico, o período entre a floração e o início do surgimento de novas flores é consideravelmente mais curto do que em climas mais quentes. Darbyshire et al. (2017) apresentam uma avaliação de modelos de fenologia de floração de maçã usando dados de 14 locais. Dois estágios de floração, foram investigados permitindo uma avaliação com base no desempenho estatístico e suposições biológicas. Este trabalho avalia as mudanças climáticas e as épocas de floração e ilustra os diferentes efeitos das temperaturas contrastantes do inverno e da primavera.

Em Chitu e Paltineanu (2020) os autores examinam as consequências das mudanças climáticas em maçãs e pêras, em quatro estágios fenológicos no clima temperado do sul da Romênia. Dados meteorológicos e fenológicos foram coletados de 1969 a 2018 e processados estatisticamente. A hipótese testada é quanto as datas de início dos estágios fenológicos sofreram modificações devido às mudanças climáticas. Os resultados apontam um aumento significativo na temperatura do ar durante os primeiros meses do ano e também um aumento nas horas de sol no mês de abril. As consequências dessa tendência avançada para os estágios fenológicos em macieiras e pereiras é a ocorrência de acidentes climáticos como geadas tardias, e o aumento em sua frequência e severidade. Outras consequências referem-se à dificuldade de polinização dos insetos e frutificação, à aplicação mais precoce de tratamentos fitossanitários, água de irrigação e ao avanço da colheita.

Czernecki, Nowosad e Jablonska (2018) desenvolveram um estudo com o objetivo criar e avaliar diferentes modelos estatísticos para reconstruir, prever e melhorar a qualidade do monitoramento das fases fenológicas com o uso de dados de satélite e meteorológicos. Foram utilizadas técnicas de ML baseadas em regressão mais comumente aplicadas, como Regressão Linear Múltipla, Laço, Regressão de Componente Principal, *Generalized Boosting Model* e *Random Forest*. Segundo os autores a estratégia desenvolvida mostra potencial para estimar fenologia de sensoriamento remoto usando algoritmos de aprendizado de máquina.

No trabalho de Liu e Wang (2018) foram empregados dados de satélite para investigar o estágio do trigo de inverno sobre uma planície no norte da China. Um dos objetivos do estudo é investigar os padrões espaço-temporais do trigo de inverno e suas tendências nas últimas três décadas. Este estudo destaca papéis importantes dos dados de sensoriamento remoto e algoritmos fenológicos baseados em satélite para o monitoramento e gerenciamento do crescimento da cultura do trigo.

Dai et al. (2019) aplicaram três algoritmos de ML: *random forest*, redes neurais e *gradient boosting decision tree*. O objetivo foi prever o desdobramento das folhas para quarenta espécies lenhosas durante o qual o clima de Harbin, na China, foi substancialmente mais quente. Entre os objetivos, um deles foi comparar o desempenho de três tipos de modelos de ML, cinco modelos ecofisiológicos e dois modelos lineares simples e investigado a sensibilidade à temperatura do

desdobramento das folhas. Foram investigadas ainda mudanças temporais durante cada período com uma janela móvel de 15 anos e relacionadas a variações em outros fatores fisiológicos e ambientais. Os autores concluem que a pesquisa mostra que os algoritmos de ML são aplicáveis a estudos de modelos fenológicos, e os modelos que usam algoritmos de ML podem melhorar o desempenho do modelo em conjuntos de dados de validação externa.

Em Yazdanpanah, Ohadi e Soleimani (2010), os autores utilizaram dados de estatísticas fenológicas e da meteorologia para antecipar diferentes fases fenológicas de macieiras usando uma rede neural. Os dados são de 1999 a 2005, obtidos em Golmakan, província de Khorasan, Iran. Neste trabalho foi estabelecida uma matriz de dados de entrada com os parâmetros climáticos de temperatura mínima, temperatura máxima, a média da temperatura diária, temperatura mínima absoluta e temperatura máxima absoluta. A matriz de dados de saída são as datas de ocorrência dos diferentes estágios fenológicos de macieiras. Os resultados mostraram os valores estimados da duração dos estágios inferiores a 3 dias e diferentes taxas de erro com variações bastante altas dependendo do estágio avaliado.

El Yaacoubi et al. (2019) realizaram um estudo com o objetivo de selecionar um modelo com melhor ajuste em três locais contrastantes do Mediterrâneo. O modelo avaliado é baseado no modelo sequencial (Chuine, (2000); Legave et al., (2013)). O estágio da plena floração foi investigado no estudo. Os autores concluem que com a falta de estudos sobre a modelagem fenológica da maçã para a área do Mediterrâneo, o estudo identificou um modelo fenológico válido caracterizado por padrões de temperatura contrastantes. A previsão dos estágios fenológicos com base na modelagem sequencial destacou um bom ajuste do modelo e consegue prever os estágios de floração de macieiras na região do Mediterrâneo.

Lee et al. (2020) investigaram a eficácia de uma nova estrutura de modelagem para previsões de fenologia e produtividade de frutos de baga, usando morangueiros. No trabalho foram coletados dados ambientais e de produção ao longo da estação de cultivo usando microsensores em rede de alta resolução. O objetivo foi melhorar a compreensão da fenologia de frutos de baga e fornecer um caminho a seguir no uso de “big data” para parametrizar ferramentas de aprendizado de máquina para apoiar o desenvolvimento sustentável da indústria de frutos de baga.

Klompenburg, Kassahun e Catal (2020) apresentaram uma revisão sistemática da literatura, para extrair e sintetizar os algoritmos e recursos que têm sido usados em estudos de previsão de safra. De acordo com a análise dos artigos investigados, o estudo mostra que as publicações selecionadas usam uma variedade de recursos, dependendo do escopo da pesquisa e da disponibilidade de dados. Cada artigo investiga a previsão de rendimento com aprendizado de máquina, mas difere dos recursos. Os estudos também diferem em escala, posição geológica e cultura. A escolha dos recursos depende da disponibilidade do conjunto de dados e do objetivo da pesquisa. Estudos também afirmam que modelos com mais recursos nem sempre fornecem o melhor desempenho para a previsão de produtividade. Os resultados mostram que nenhuma conclusão específica pode ser tirada sobre qual é o melhor modelo, mas eles mostram clara-

mente que alguns algoritmos de aprendizado de máquina são usados mais do que outros. Os algoritmos mais usados são *random forest*, redes neurais, regressão linear e *gradient boosting tree*. As características mais utilizadas são temperatura, precipitação e tipo de solo. Os algoritmos de aprendizado profundo mais preferidos são Redes Neurais Convolucionais (CNN), *Long-Short Term Memory* (LSTM) e *Deep Neural Networks* (DNN).

Em Rea e Eccel (2006) os autores desenvolveram um modelo de previsão da data de floração a partir de seis séries disponíveis de maçã "*Golden Delicious*" em seis locais em Trentino, região alpina de fruticultura. Trata-se de uma versão modificada do modelo "Utah", utilizando unidades de frio e soma térmica de graus-hora de crescimento (GDH). Em Marra et al. (2001) os autores avaliaram o crescimento dos frutos para cultivares de pêssego e nectarina em relação ao tempo térmico acumulado do florescimento à colheita dos frutos para cultivares de pêssego e nectarina. O período de desenvolvimento dos frutos variou de 70 a 150 dias. A soma térmica foi calculada usando graus-dia e graus-hora de crescimento. Em Mendes et al. (2019), o objetivo principal do estudo foi a documentação e a caracterização dos diferentes estágios fenológicos, bem como a definição das exigências térmicas, da Atemóia (*Annona cherimola* Mill. e *Annona squamosa* L.) para duas safras agrônomicas. O número de dias e as necessidades térmicas para a realização de cada estágio fenológico foram diferentes entre as duas estações agrônomicas da atemóia.

Safa et al. (2004) apresentam um estudo onde foi aplicada uma Rede Neural Artificial para prever a produção de trigo de cultivo seco. Um dos objetivos é realizar a predição da produtividade do trigo pelo menos dois meses antes do estágio de maturação completa onde ocorre a formação do final do caule. Os resultados demonstram precisão e eficiência com um erro máximo entre 45 e 60 kg/ha. Souza, Leonel e Silva (2011) avaliaram a temperatura basal, a soma térmica acumulada em diferentes fases fenológicas, a duração das fenofases, a produtividade e a sazonalidade do ciclo de uma cultivar de nectarina e de 14 cultivares de pêssego. Os dados são tratados estatisticamente. Os resultados demonstram que as temperaturas mínimas basais variam entre os estágios fenológicos da planta, com pouca diferença entre as cultivares de nectarina e pêssego. Algumas cultivares selecionadas são mais adaptadas as condições climáticas locais, com grande variação quanto a requisitos térmicos e suas fases fenológicas.

McCormick et al. (2020) propôs um modelo híbrido usando um modelo orientado a dados com o uso de predições baseadas em conhecimento. Neste trabalho, um conjunto de dados de fenologia da soja foi usado para demonstrar que uma estratégia de otimização multimodal pode ser usada para treinar um modelo de crescimento de safra e que os recursos do modelo baseado em conhecimento podem melhorar os processos de aprendizado de máquina. No trabalho de Wang et al. (2019) os autores apresentam um modelo de rede de aprendizagem profunda para prever a densidade da vegetação em uma determinada área e o clima futuro é inferido a partir da densidade da vegetação prevista.

Chen et al. (2020) propõem um método híbrido combinando redes neurais com aprendizagem integrada para prever o período de floração de macieiras da variedade "Red Fuji". O método utiliza redes LSTM e as funções de classificação *Random Forest* e *Adaboost*. Em Qian

et al. (2018) os autores propõem um framework de distribuição que consiste na aquisição de imagens de árvores frutíferas, previsão de produtividade em cliente smartphone, processamento de dados e cálculo de modelo em cliente servidor para estimativa da produtividade potencial de fruta. Um modelo de estimativa de rendimento de árvore individual em uma rede neural artificial foi desenvolvido com três camadas, quatro parâmetros de entrada, 14 neurônios ocultos e um parâmetro de saída. O sistema foi utilizado em um pomar experimental de maçãs da variedade "*Fuji*". Vinte e seis amostras de árvores foram selecionadas de um total de 80 árvores de acordo com os múltiplos do número.

Em Blazek e Pistekova (2017) o crescimento dos frutos foi avaliado durante três épocas de cultivo em quatro cultivares de maçã. O período de acompanhamento foi definido pelo início da floração, a data de obtenção do estágio T pelos pequenos frutos e o início do amadurecimento ideal da colheita, que para cada cultivar foi determinado com base no índice de amido, firmeza da polpa e sólidos solúveis. As relações próximas foram detectadas entre o tempo de obtenção do estágio T e o prazo de maturação de colheita para três das quatro cultivares estudadas. O objetivo deste trabalho foi definir um método adequado para a previsão do prazo de colheita para quatro cultivares de maçã com base em estádios de desenvolvimento típicos dos frutos e uma correlação entre o diâmetro dos pequenos frutos no estágio T e os frutos no amadurecimento da colheita. Em Cheng et al. (2017), os autores apresentam uma nova abordagem para prever o rendimento inicial da produção de macieiras utilizando imagens da planta e características do dossel das árvores. O rendimento por árvore foi obtido com o uso de uma rede neural de retropropagação (BPNN). Os resultados mostraram que BPNN pode ser usado para previsão de produção de maçã e que essas quatro características de dossel selecionadas são adequadas para previsão de produção antecipada e apresenta uma maneira elegante de prever a produção de frutas usando visão computacional e aprendizado de máquina para maçã e possivelmente outras frutas.

No artigo de Sabzi et al. (2019), os autores desenvolveram um novo algoritmo de visão computacional para detectar os frutos existentes em imagens aéreas de uma cultivar de maçã da variedade "*Red Delicious*" e estimar seu estágio de maturação entre quatro classes possíveis: verde, meio maduro, maduro e maduro demais. O método proposto é baseado em uma combinação das características de cor mais eficazes e um classificador baseado em redes neurais artificiais *multilayer perceptron* (MLP) otimizadas com algoritmos genéticos. No trabalho de Haider et al. (2019) o objetivo se concentra no desenvolvimento de um modelo de previsão de produção de trigo usando *Robust-LOWESS* como uma função de suavização em conjunto com um modelo de redes neurais LSTM. Os resultados mostram que o modelo proposto atinge melhor desempenho em termos de previsão, e revela que enquanto a produção de trigo aumentará gradualmente nos próximos dez anos, a relação produção/consumo continuará a cair e representará uma ameaça para a economia em geral. A estrutura proposta pode ser usada como diretrizes para a produção de trigo em particular, e é favorável a outras culturas também, levando ao desenvolvimento da agricultura sustentável em geral.

### 3.3 Discussão

A revisão de literatura resultou em um conjunto de artigos, que enfocam as iniciativas de predição de tempo de colheita, e outras relacionadas a predição de estádios fenológicos em geral, de macieiras e de outras culturas. As técnicas aplicadas são variadas, desde métodos puramente estatísticos ao uso de algoritmos de ML, passando por redes neurais até métodos de aprendizagem profunda. A Tabela 1 apresenta um resumo dos trabalhos descritos na Seção 3.2. Os trabalhos foram organizados, destacando a cultura, o objetivo, características e a abordagem utilizada.

Considerando o conjunto de trabalhos relacionados, podemos enfatizar algumas características. Alguns trabalhos, como Rivero et al. (2017); Darbyshire et al. (2017); Chitu e Paltineanu (2020), avaliam os impactos da elevação da temperatura e mudanças climáticas no tempo de floração das macieiras devido ao aquecimento global. Esses trabalhos utilizam abordagens estatísticas e relacionam dados meteorológicos e da fenologia apontando consequências dessa tendência, e quanto as datas de início dos estádios fenológicos sofreram modificações devido às mudanças climáticas. Estes trabalhos na literatura se relacionam com o tema proposto neste trabalho no que diz respeito as variáveis climáticas e a relação com a ocorrência dos fenômenos fenológicos, aspecto comum entre ambos.

Trabalhos como o de Marra et al. (2001); Mendes et al. (2019); Souza, Leonel e Silva (2011); Petri et al. (2012) tem foco na caracterização dos diferentes estágios fenológicos, bem como a definição das exigências térmicas de diferentes culturas, ou ainda na ocorrência de estádios fenológicos de macieiras como em Cardoso (2011) onde foram ajustados modelos para estimar o número de dias entre a quebra de dormência de macieiras e os eventos fenológicos posteriores, em função do acúmulo de frio e calor. Alguns trabalhos como de Lee et al. (2020) e Saha et al. (2004) objetivam a predição de produtividade de culturas. Embora tenham sido aplicados em culturas de morango e trigo, tem relação com este trabalho pelas abordagens utilizadas. Uma grande quantidade de artigos são encontrados na literatura relacionados ao tema, alguns deles foram selecionados e depois descartados na seleção de trabalhos da metodologia, por tratarem de predição de colheita com foco específico na produtividade em aspectos qualitativos e quantitativos.

Trabalhos utilizando imagens de satélite e dados de sensoriamento remoto incluem Wang et al. (2019); Czernecki, Nowosad e Jablonska (2018); Liu e Wang (2018). Estes trabalhos, embora aplicados a outras culturas e com dados de sensoriamento remoto, demonstram o potencial de algoritmos de ML usados para estimar fenologia. Trabalhos como Sabzi et al. (2019); Qian et al. (2018) fazem uso de imagens, aplicando métodos de visão computacional e algoritmos de aprendizagem profunda. O uso destas técnicas permite explorar uma combinação de outras características, e tem sido amplamente utilizado como mostra a revisão apresentada em Klompenburg, Kassahun e Catal (2020).

Visando uma análise comparativa entre os trabalhos analisados, a Tabela 2 apresenta uma se-

Tabela 1 – Resumo dos artigos resultantes da revisão bibliográfica

Artigo/Ano	Cultura	Objetivo	Características	Abordagem utilizada
Cardoso (2011)	Maçã	Predição de estádios fenológicos	Dados fenologia/meteorológicos	Estatística
Petri et al. (2012)	Maçã	Avaliação de estádios fenológicos	Dados fenologia e meteorológicos	Estatística
Rivero et al. (2017)	Maçã	Predição de floração	Dados fenologia e meteorológicos	Estatística
Darbyshire et al. (2017)	Maçã	Predição de floração	Dados fenologia e meteorológicos	Estatística
Chitu e Paltineanu (2020)	Maçã e Pera	Avaliação de estádios fenológicos	Dados fenologia e meteorológicos	Estatística
Qian et al. (2018)	Maçã	Predição de produtividade	Imagens	Redes Neurais Artificiais
Czernecki, Nowosad e Jablonska (2018)	Maçã	Predição de produtividade	Imagens	Regressão linear múltipla, <i>Generalized Boosting Model</i> , <i>Random forest</i>
Liu e Wang (2018)	Trigo	Monitorar crescimento	Dados de satélite	<i>Polynomial-NDVI algorithm</i> , <i>polynomial-cumNDVI</i> e estatística
Dai et al. (2019)	Lenhosas	Predição de abertura folha	Dados fenologia e meteorológicos	RNA, <i>Random Forest</i> e <i>Gradient boosting decision tree</i>
Yazdanpanah, Ohadi e Soleimani (2010)	Maçã	Predição de estádios fenológicos	Dados fenologia/meteorológicos	Redes Neurais Artificiais
El Yaacoubi et al. (2019)	Maçã	Predição de floração	Dados fenologia e meteorológicos	Modelo Sequencial
Lee et al. (2020)	Morango	Predição de produtividade	Dados sensores/fenologia/meteorológicos	Generalised linear models
Rea e Eccel (2006)	Maçã	Predição de floração	Dados fenologia/meteorológicos	Modelo de Utah
Marra et al. (2001)	Pêssego e nectarina	Predição de estádios fenológicos	Dados fenologia/meteorológicos	Estatística
Safa et al. (2004)	Trigo	Predição de produtividade	Dados fenologia e meteorológicos	Redes Neurais Artificiais
Souza, Leonel e Silva (2011)	Pêssego e nectarina	Avaliação de estágios fenológicos	Dados fenologia e meteorológicos	Estatística
McCormick et al. (2020)	Soja	Predição de rendimento safra	Dados fenologia e meteorológicos	LSTM, modelo Bayesiano, <i>swarm of particles</i>
Boechel et al. (2021)	Maçã	Predição de estágios fenológicos	Dados fenologia e meteorológicos	<i>Fuzzy Time Series</i>
Wang et al. (2019)	Vegetação	Predição de densidade da vegetação	Imagens de satélite	<i>Fully convolutional networks (FCN)</i> , LSTM
Chen et al. (2020)	Maçã	Predição de floração	Dados fenologia e meteorológicos	LSTM, <i>Random Forest</i> e Adaboost
Qian et al. (2018)	Maçã	Predição de produtividade	Imagens aéreas	Rede Neural Artificial
Blazek e Pistekova (2017)	Maçã	Predição de época de colheita	Dados do teor de amido e diâmetro da fruta	Estatística
Sabzi et al. (2019)	Maçã	Predição de estágio de maturação	Imagens aéreas	Redes Neurais Artificiais
Cheng et al. (2017)	Maçã	Predição de produtividade	Imagens e dados do docel	Back propagation Neural Network (BPNN)
Haider et al. (2019)	Trigo	Predição de produtividade	Dados fenologia e meteorológicos	<i>Robust-LOWESS</i> e LSTM

Fonte: Elaborado pelo autor.

leção de trabalhos relacionados, considerando algumas características mais intimamente ligadas ao tema deste trabalho. Todos os trabalhos apresentam algum tipo de solução para problemas relacionados, embora não tenham o mesmo enfoque, têm uma relação muito forte com o problema de interesse. A coluna Cultura mostra o trabalho relacionado à cultura da macieira. A coluna ML mostra trabalhos que usam técnicas de ML. A coluna Outra abordagem mostra trabalhos que usam alguma outra estratégia para prever estágios fenológicos e previsão de colheita. As colunas ANN e RNN apresentam trabalhos que usam essas técnicas.

Considerando o conjunto de artigos avaliados, este estudo mostra que as publicações selecionadas usam uma variedade de técnicas. Os artigos que mais se aproximam do objetivo proposto neste trabalho são os de Chen et al. (2020) e McCormick et al. (2020). Porém ambos utilizam técnicas híbridas. O primeiro aplicado a macieiras, aspecto comum a este trabalho, mas com objetivo bastante distinto. E o segundo, aplicado a cultura da soja, e com finalidade também distinta. O trabalho de Blazek e Pistekova (2017) apresenta similaridade com o tema em questão, se considerado o objetivo de predizer a época de colheita, porém o método proposto utiliza técnica distinta.

Não foram encontrados trabalhos fazendo uso de redes recorrentes aplicados ao problema de predição da data de colheita de macieiras. Com base nesta revisão, reforçamos a evidência do ineditismo do tema proposto neste trabalho.

Tabela 2 – Tabela comparativa das abordagens nos trabalhos relacionados

Referência	Cultura (maçã)	ML	ANN	RNN	Outra abordagem	Predição data colheita
Blazek e Pistekova (2017)	✓				✓	✓
Boechel et al. (2021)	✓				✓	
Cardoso (2011)	✓				✓	
Chen et al. (2020)	✓	✓		✓		
Chitu e Paltineanu (2020)	✓				✓	
Czernecki, Nowosad e Jablonska (2018)		✓				
Dai et al. (2019)		✓	✓			
Darbyshire et al. (2017)	✓				✓	
Haider et al. (2019)		✓		✓		
McCormick et al. (2020)		✓		✓		
Petri et al. (2012)	✓				✓	
Putti, Mendez e Petri (2000)	✓				✓	
Rivero et al. (2017)	✓				✓	
Sabzi et al. (2019)	✓	✓	✓			
Safa et al. (2004)		✓	✓			
Yazdanpanah, Ohadi e Soleimani (2010)	✓	✓	✓			

## 4 PREDHARV: UM MODELO PARA PREDIÇÃO DA DATA DA COLHEITA DE MACIEIRAS

Este capítulo apresenta o modelo PredHarv. A Seção 4.1 apresenta as premissas e decisões de projeto. A Seção 4.2 apresenta os detalhes da arquitetura do modelo. A Seção 4.3 apresenta os detalhes da estratégia de predição. Por fim, a Seção 4.4 descreve os detalhes dos conjuntos de dados utilizados na implementação do protótipo.

Considerando os trabalhos apresentados no Capítulo 3, observou-se que o cenário atual apresenta diversas iniciativas que empregam diferentes tecnologias para oferecer melhores serviços aos produtores, considerando diferentes culturas. A evolução dos estágios fenológicos da macieira durante o estágio vegetativo e produtivo é influenciada pelas condições climáticas e influencia na produtividade e na qualidade da produção. De acordo com Bergamaschi (2007), a disponibilidade térmica tem influência direta na fenologia da macieira. As temperaturas mais altas aceleram o desenvolvimento da planta, enquanto as temperaturas mais baixas prolongam o ciclo. Neste trabalho propomos um modelo de ML para estimar um período específico. A partir da floração total, queremos estimar a data de início da colheita das macieiras. O modelo PredHarv é baseado em modelos de soma térmica, mas usa uma abordagem multivariada. Usamos a soma térmica relacionando-a com a duração do período e outras variáveis relacionadas a temperatura do período. Usamos um método ML para prever a duração do período, com base em redes LSTM. Uma metodologia de utilização do modelo é proposta com o objetivo de ampliar a capacidade preditiva, como forma de reduzir a incerteza implícita nas informações fornecidas pelo usuário, necessárias para o cálculo da predição.

A motivação para o desenvolvimento do presente trabalho tem como base as seguintes afirmações:

- As consequências das mudanças climáticas tem impactos na produção de frutas. As incertezas que envolvem as complexas interações do meio ambiente e o desenvolvimento das plantas podem representar custos significativos para os produtores;
- A evolução dos métodos de ML tem permitido o desenvolvimento de diversas aplicações, incluindo a área agrícola. A agricultura baseada no conhecimento, de acordo com o conceito de Agricultura Inteligente, com a ajuda de ML vem sendo capaz de aumentar os níveis de produção e a qualidade dos produtos, tornando a agricultura mais eficiente e eficaz;
- Considerando a importância da cultura da maçã na economia da região Sul do Brasil, o modelo proposto contribui no aspecto de antecipar informações para os produtores e pode representar melhorias na lucratividade para o setor da fruticultura.

## 4.1 Premissas e Decisões de Projeto

Os requisitos descritos abaixo cobrem um conjunto de tarefas a serem executadas pelo modelo. Eles são considerados como condições ou recursos que precisam ser atendidos. O modelo PredHarv foi modelado considerando os seguintes requisitos e decisões como ponto de partida para o seu desenvolvimento:

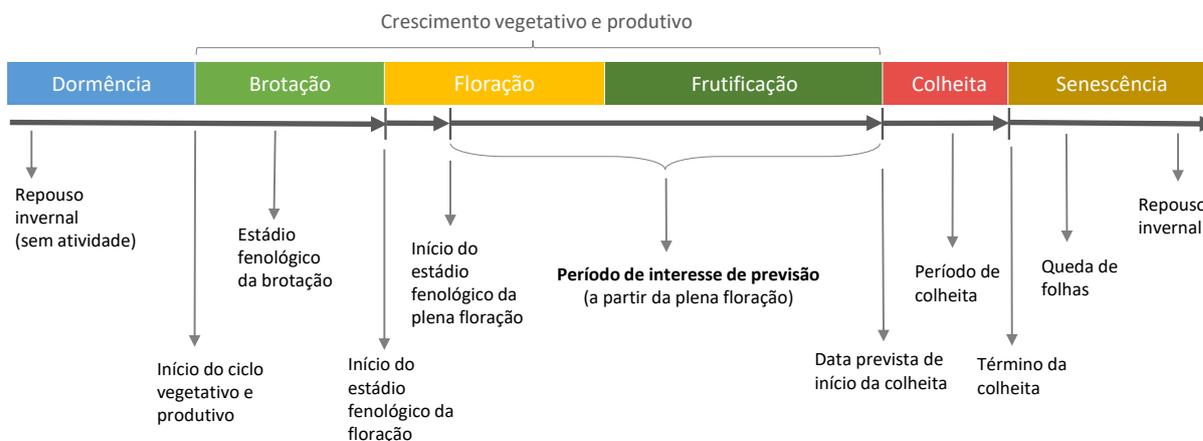
- o modelo precisa ser treinado antes de fazer previsões;
- os dados de treinamento provêm de séries históricas de fenologia e temperatura, organizados em períodos de acordo com o ano (safra);
- o treinamento do item é realizado pelo usuário apenas uma vez, exceto quando novos dados são incorporados aos dados de treinamento;
- o usuário precisa informar os cenários de predição que são utilizados pelo modelo para fazer as predições, ou seja, esses cenários referem-se aos valores que são atribuídos ao conjunto de variáveis preditoras utilizadas pelo modelo e que representam as condições meteorológicas para um certo período;
- as entradas do modelo são multivariadas e compreendem um conjunto de variáveis preditoras usadas para obter a variável de resposta;
- as saídas do modelo correspondem à duração (número de dias) do período entre o início da floração e o momento da colheita;
- a duração do período pode variar entre as previsões, pois o acúmulo de calor pode ser maior ou menor naquele período.

## 4.2 Arquitetura do Modelo

O objetivo do modelo PredHarv é obter a previsão da duração do período referente à fase fenológica desde a floração plena até o início da maturação. Dependendo da variação climática, a duração do período pode variar para mais ou para menos. Dadas as condições climáticas relacionadas à previsão de temperatura para o período a partir do momento da plena floração, queremos prever quanto tempo levará até a fase de maturação do fruto, que caracteriza o momento do início da colheita.

A macieira é uma árvore perene e sua estrutura vegetativa é resultante do crescimento acumulado de vários anos. No mesmo ano apresenta um período de repouso, no inverno, e um período de desenvolvimento vegetativo e produtivo, na primavera e no verão. Este ciclo é representado graficamente na Figura 11. Ao término do período de repouso, quando as exigências de frio são plenamente satisfeitas, ocorre a brotação que dá início ciclo vegetativo e reprodutivo, caracterizada pelos estádios fenológicos (MONTEIRO et al., 2009).

Figura 11 – Ciclo de desenvolvimento da macieira



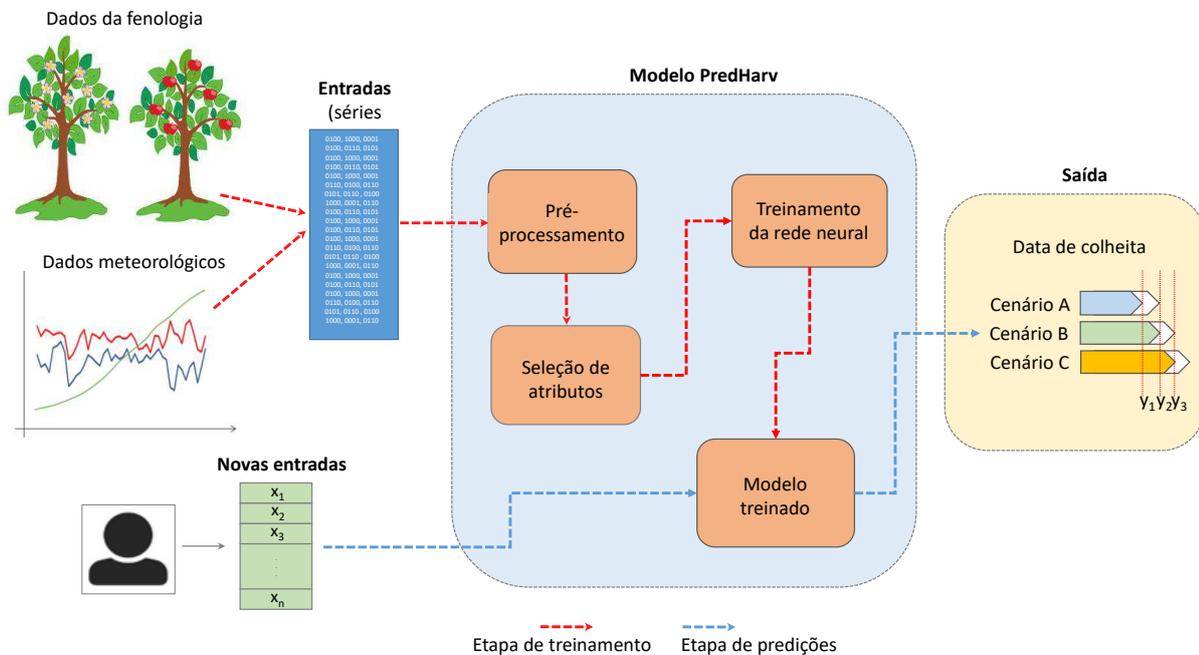
Fonte: O próprio autor

Temos um contexto caracterizado como um problema de regressão, com entrada multivariada e saída única. Tecnicamente, a modelagem de regressão preditiva é a tarefa de aproximar uma função de mapeamento ( $f$ ) de variáveis de entrada ( $X$ ) para uma variável de saída contínua ( $y$ ). A estratégia de previsão que usamos aborda a previsão de séries temporais como uma tarefa de aprendizagem supervisionada. Os métodos de ML, quando usados para prever séries temporais, precisam que os dados sejam reformulados como problemas de aprendizagem supervisionada. Essa reformulação organiza os dados de uma sequência em pares de sequências de entrada e saída. Um problema de aprendizado supervisionado é composto de padrões de entrada ( $X$ ) e padrões de saída ( $y$ ), de forma que os algoritmos possam aprender como prever padrões de saída a partir de padrões de entrada.

A Figura 12 apresenta a arquitetura do modelo PredHarv e seus principais elementos. Inicialmente, o modelo deve ser treinado antes de fazer as previsões, com dados da série histórica de fenologia e temperatura. O modelo só precisa ser treinado uma vez, a menos que haja necessidade de incorporar novos dados ao conjunto de dados. Resumidamente, os dados fenológicos identificam o início e o final do período, e os dados meteorológicos são os valores diários das variáveis neste período. A consistência dos dados é verificada no módulo de pré-processamento, garantindo que todos os dados estejam no formato adequado para serem apresentados ao modelo. Nesta etapa, também são realizadas operações de imputação de dados faltantes, quando necessário, normalização e classificação dos dados nos padrões de entrada e saída. Após o treinamento, presume-se que o modelo esteja treinado e minimizando o erro. Com base no que é aprendido na etapa de treinamento, são feitas inferências a partir das novas entradas informadas pelo usuário. As setas vermelhas na Figura 12 representam o fluxo da etapa de previsão após o treinamento. Os cenários A, B e C na etapa de saída da Figura 12 representam a saída do modelo, ou seja, a duração do período. Esses cenários são perspectivas, ou seja, pontos de vista sobre situações específicas relacionadas a janela de tempo futura, como por exemplo variações

na temperatura para mais ou para menos. Essas perspectivas são possíveis devido à flexibilidade do modelo, que com base em variações nos dados de entrada submetidos ao modelo treinado permitem a prospecção de cenários futuros.

Figura 12 – Arquitetura do modelo



Fonte: O próprio autor

As entradas do modelo são séries temporais de dados de temperatura, ordenados no tempo, que estão relacionados e contribuem para a ocorrência do fenômeno. Usamos seis variáveis que servem de entrada para o modelo: temperatura mínima diária ( $t_{min}$ ), temperatura máxima diária ( $t_{max}$ ), temperatura média diária ( $t_{med}$ ), temperatura mínima acumulada ( $t_{min\_acc}$ ), temperatura média acumulada ( $t_{med\_acc}$ ) e graus-dia acumulados ( $GDA$ ). As Figuras 13 e 14 apresentam diagramas boxplot com as variáveis utilizadas dos datasets DS1 e DS2.

As variáveis de entrada, usam os valores das observações anteriores ( $t-n, \dots, t-2, t-1$ ). Essas observações anteriores são chamadas de defasagens, que correspondem ao valor imediatamente anterior à série. Esta técnica é chamada de janela deslizante (KHANDELWAL; KONAR; BRAHMA, 2020). Usamos sete defasagens para cada variável de entrada, ou seja, os valores correspondentes a sete defasagens imediatamente anteriores ao tempo atual ( $t$ ).

### 4.3 Estratégia de Predição

O modelo PredHarv faz previsões a partir das informações fornecidas pelo usuário. A função do usuário é extremamente restrita, mas de suma importância no resultado da previsão, uma vez que a rede responde ao contexto que está sendo alimentado. Considerando que as informações alimentadas no início do período, próximo à floração plena, têm um horizonte de predição

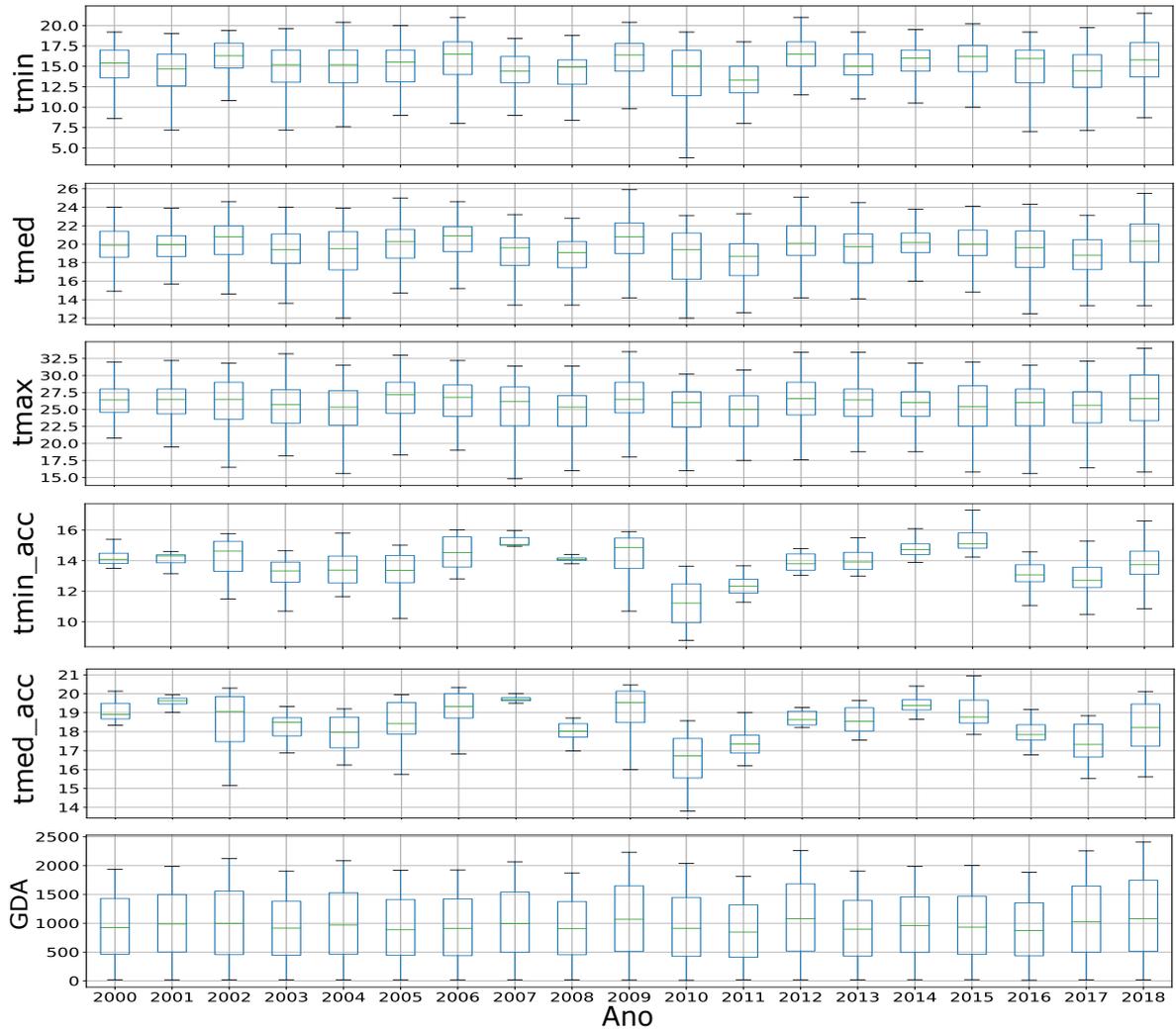


Figura 13 – Boxplot das variáveis de entrada - Conjunto de dados DS1

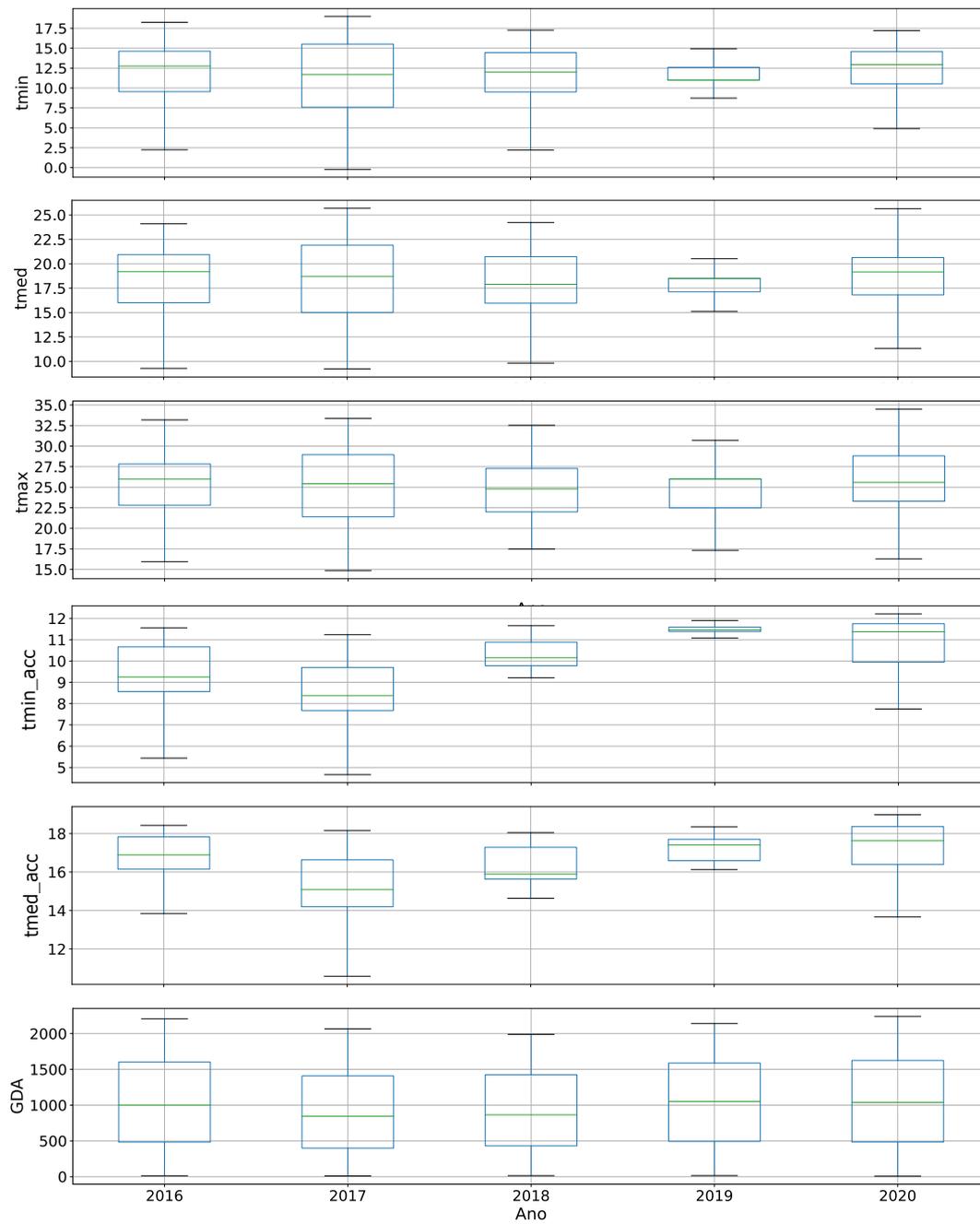
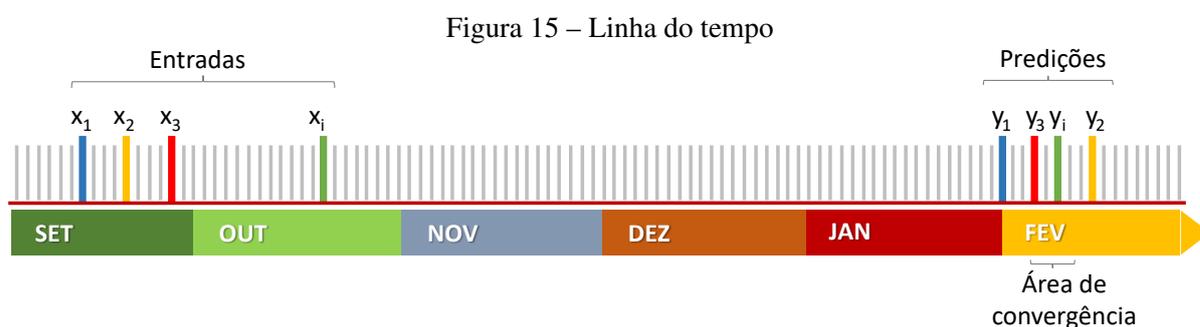


Figura 14 – Boxplot das variáveis de entrada - Conjunto de dados DS2

mais longo devido ao tamanho da janela, e com isso, um maior nível de incerteza do cenário futuro, o modelo PredHarv foi modelado para permitir flexibilidade, permitindo que outras previsões sejam feitas a qualquer momento dentro do período. Essas previsões permitem que a estimativa seja ajustada à medida que os dados de entrada são atualizados com informações sobre os dias que já ocorreram e uma janela futura menor. A Figura 15 ilustra o processo de atualização da previsão conforme novas entradas de dados são alimentadas. Na Figura 15, novas entradas são alimentadas pelo usuário representado por  $x$ , e as previsões são representadas por  $y$ . À medida que as previsões são feitas, há uma tendência de convergir para um ponto comum, representado na Figura 15 pela área de convergência.



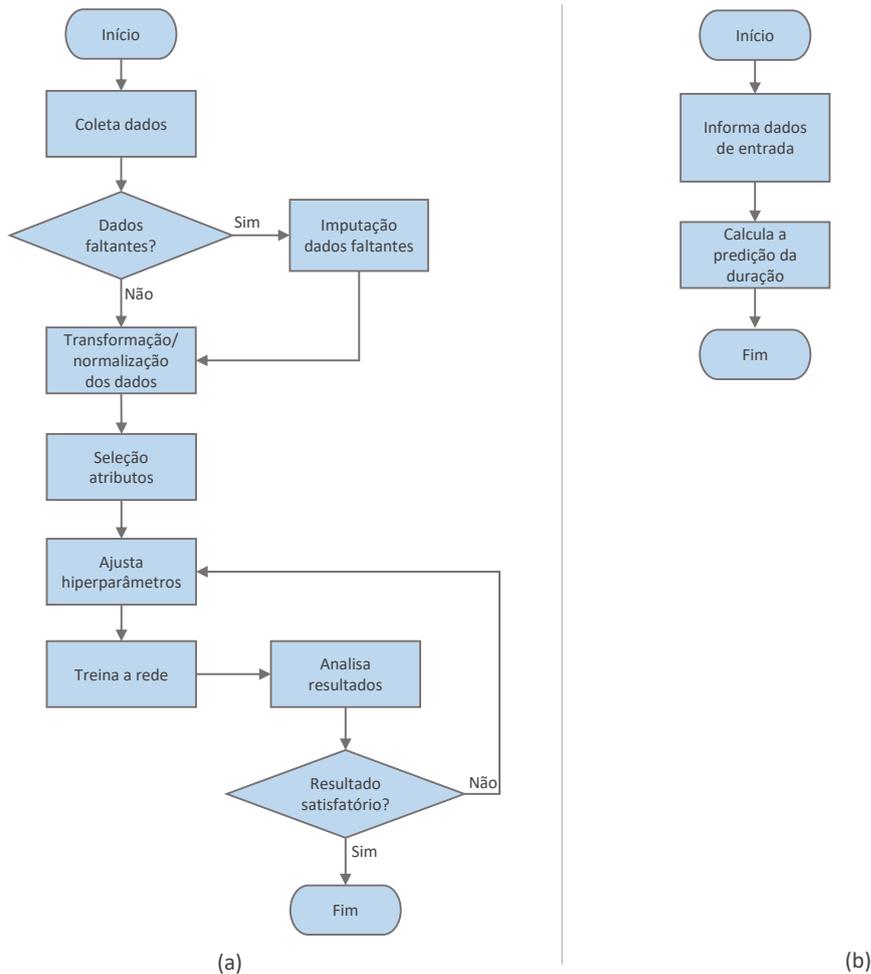
Fonte: O próprio autor

A Figura 16 apresenta o fluxo de processamento do modelo PredHarv nas etapas de treinamento e previsão. Na etapa de treinamento (Figura 16-a), após a coleta de dados, os dados ausentes são analisados. Esse é um problema comum ao trabalhar com séries temporais envolvendo cenários reais. A técnica utilizada para imputação de dados é descrita na Subseção 4.4.1.1. No módulo de transformação e normalização de dados, as operações foram realizadas para garantir que os dados estejam todos no mesmo formato e na mesma ordem de magnitude. Outros detalhes sobre a transformação e normalização dos dados são descritos na Subseção 4.4.1.2. No módulo de seleção de recursos, foram selecionados os atributos de entrada do modelo. No ajuste de hiperparâmetros, foram estabelecidos os principais parâmetros usados na configuração da rede LSTM. Esses parâmetros são descritos na seção 5.1. Em seguida, no módulo de treinamento, é realizado o treinamento da rede. Se os resultados do treinamento não forem considerados satisfatórios, avaliados a partir dos resíduos de treinamento da série de treinamento e teste, os valores dos hiperparâmetros são ajustados e o modelo pode ser retreinado. Se os resultados forem satisfatórios, o modelo treinado é considerado apto para realizar as previsões.

Na etapa de previsão (Figura 16-b), no módulo Dados de entrada, o usuário informa os valores dos atributos de entrada que são utilizados para realizar a previsão. Em seguida, esses dados de entrada são submetidos ao modelo treinado para o cálculo da previsão.

A Figura 17 ilustra graficamente, através de um diagrama de sequência, o comportamento do modelo, destacando a interação entre os componentes e determinando a sequência em que os eventos ocorrem. Inicialmente na Etapa 1, após serem carregados os dados de treinamento é

Figura 16 – Fluxograma da execução do modelo

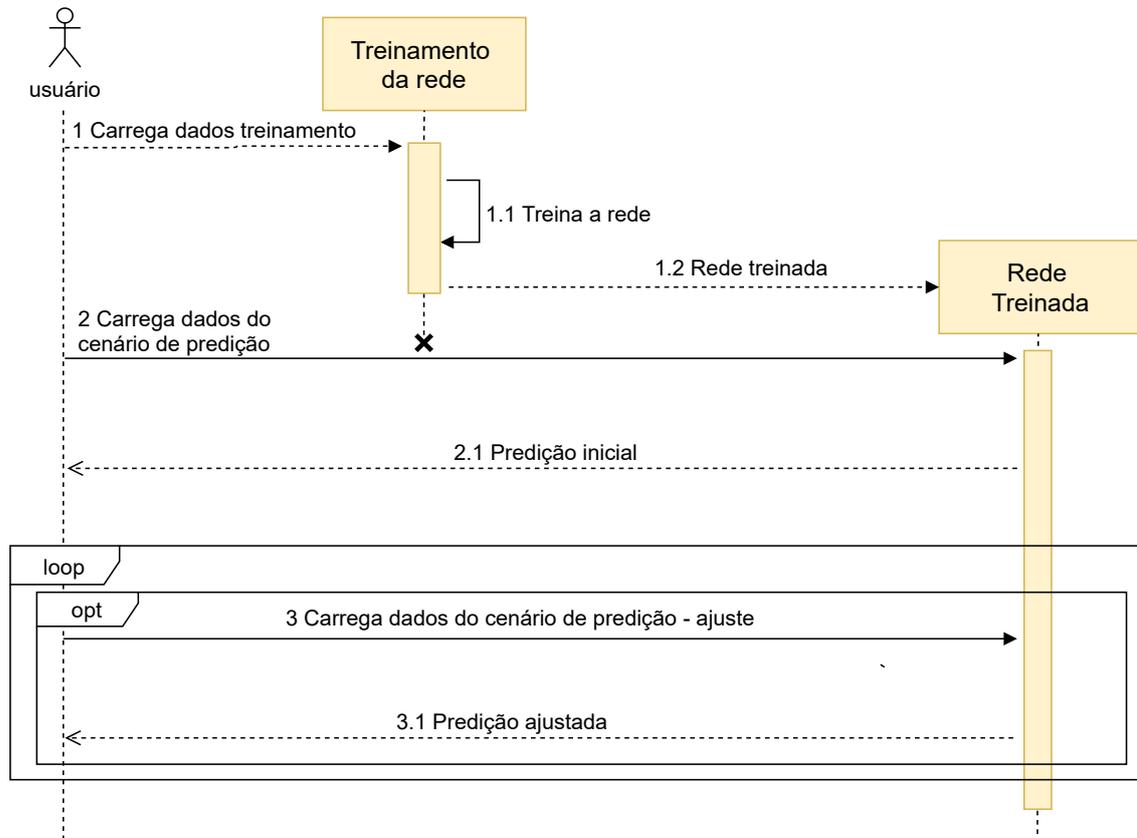


realizado o treinamento da rede. Com a rede treinada, na etapa 2, são carregados os dados do cenário de predição. A rede retorna em 2.1 a predição inicial. Na etapa 3, opcionalmente, são carregados os dados do cenário de predição novamente onde é obtido como resposta em 3.1 a predição ajustada para os dados carregados.

#### 4.4 Conjunto de dados

Os conjuntos de dados utilizados neste trabalho são compostos por dados coletados em pomares de macieiras. O primeiro conjunto (DS1) é composto com dados da fenologia e dados meteorológicos de um pomar administrado pela Empresa de Pesquisa Agropecuária e Extensão Rural de Santa Catarina (EPAGRI), Estação Experimental de Caçador, SC, localizado no sul do Brasil. O município de Caçador está situado no Oeste Catarinense (26° 49' 28" S, 50° 59' 45" O), com altitude de 960 metros. O clima é do tipo Cfb - temperado úmido, com verões amenos, segundo a classificação de Köppen. A precipitação pluvial média anual é de 1.653 mm. A temperatura média anual é de 16,6 °C. Os dados correspondem a um período de 2000 a 2018.

Figura 17 – Diagrama de sequência dos eventos do processo de previsão



O segundo conjunto de dados (DS2) é composto de informações contendo dados da fenologia de uma empresa comercial produtora de maçãs localizada na região de Vacaria, RS, e dados meteorológicos disponibilizados pelo Departamento de Diagnóstico e Pesquisa Agropecuária (DDPA) da Secretaria Estadual da Agricultura, Pecuária e Desenvolvimento Rural do Rio Grande do Sul. O município de Vacaria está situado na região dos Campos de Cima da Serra, localizado no extremo nordeste do estado do Rio Grande do Sul ( $29^{\circ} 32' 30''$  S,  $50^{\circ} 54' 51''$  O), com altitude de 962 metros. O clima é do tipo Cfb - temperado úmido, segundo a classificação de Köppen. A precipitação pluvial média anual é de 1.897 mm. A temperatura média anual é de  $16^{\circ}\text{C}$ . Os dados correspondem a um período de 2016 a 2020.

As séries de dados fenológicos consistem de anotações de observações visuais da planta e de dados obtidos a partir de cadernos de campo do Sistema de Produção Integrada de Maçã – PIM, que segue normas estabelecidas pelo Ministério da Agricultura Pecuária e Abastecimento. As macieiras em ambos os conjuntos são da cultivar *Royal Gala*. Essa cultivar foi escolhida devido a sua grande área de cultivo, representando 60% da produção brasileira de maçãs. Para a seleção das amostras utilizadas neste trabalho foram considerados apenas os locais onde não foram realizadas aplicações de agroquímicos para retardamento da maturação dos frutos, como redutores de formação do etileno, que é um hormônio vegetal relacionado à maturação dos frutos. Por questões de manejo, em alguns devido a escassez de mão-de-obra no período de

colheita, é uma prática comum a aplicação de produtos químicos visando à redução do processo de maturação e retenção de frutos nas plantas por maior período de tempo.

Os dados meteorológicos utilizados neste trabalho são provenientes de estações meteorológicas localizadas próximas aos pomares do qual foram utilizados os dados da fenologia. A partir dos dados da fenologia, foram organizados os dados meteorológicos diários referentes aos respectivos períodos. Estes dados meteorológicos foram organizados em um arquivo no formato *comma-separated values*, também conhecido como CSV. Um pequeno trecho é apresentado como exemplo na Tabela 3. De forma a contribuir com a reprodutibilidade dos resultados deste trabalho, os conjuntos de dados encontra-se disponíveis em um repositório publicamente disponível.<sup>1</sup>

Tabela 3 – Exemplo de dados de entrada usados na implementação

Data	$T_{min}$	$T_{max}$	$T_{med}$	$T_{min\_acc}$	$T_{med\_acc}$	$GD$	$GDA$
22/10/2006	12,0	25,0	18,5	12,0	18,5	15,5	15,5
23/10/2006	13,0	23,8	18,4	25,0	36,9	15,4	30,9
24/10/2006	12,5	25,0	18,7	37,5	55,6	15,8	46,7
25/10/2006	14,0	25,6	19,8	51,5	75,4	16,8	63,5
26/10/2006	13,0	29,0	21,0	64,5	96,4	18,0	81,5

#### 4.4.1 Pré-processamento

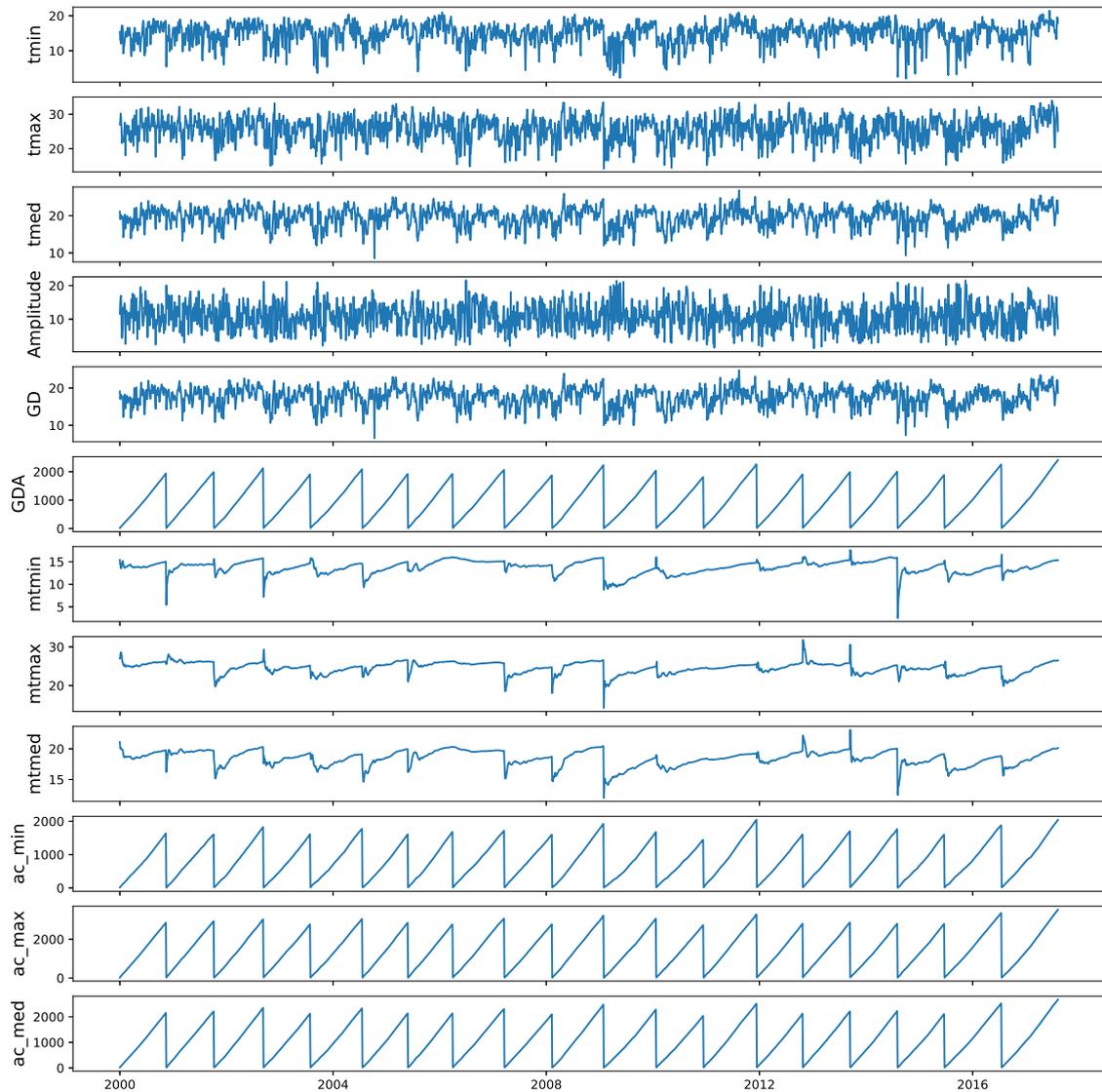
Inicialmente, verificou-se a consistência dos dados diários através de análises visuais. Em seguida, foi realizada uma análise exploratória dos dados através da plotagem das séries (ver Figura 18) e de diagramas *boxplot* (ver Figura 13). Estes diagramas possibilitam uma análise visual de toda a série de dados, mostrando aspectos úteis, como variabilidade, dispersão, simetria da distribuição, além de permitir a identificação de valores discrepantes (*outliers*) do conjunto de dados.

##### 4.4.1.1 Dados Faltantes

Um problema comum ao trabalhar com séries temporais envolvendo cenários reais é a ocorrência de dados faltantes ou dados perdidos (*missing data*). O ideal seria que essas séries estivessem completas, no entanto, em cenários reais essa situação é bastante rara. Vários problemas podem causar a perda de informações importantes, como interrupções ocasionais de estações automáticas e mau funcionamento dos instrumentos de medição. Para contornar esse problema, são usadas técnicas estatísticas que envolvem a substituição de dados ausentes por estimativas de valor. Uma estratégia é imputar os valores ausentes, ou seja, inferi-los da parte conhecida

<sup>1</sup>Disponível em: <https://github.com/tboechel/dataset>

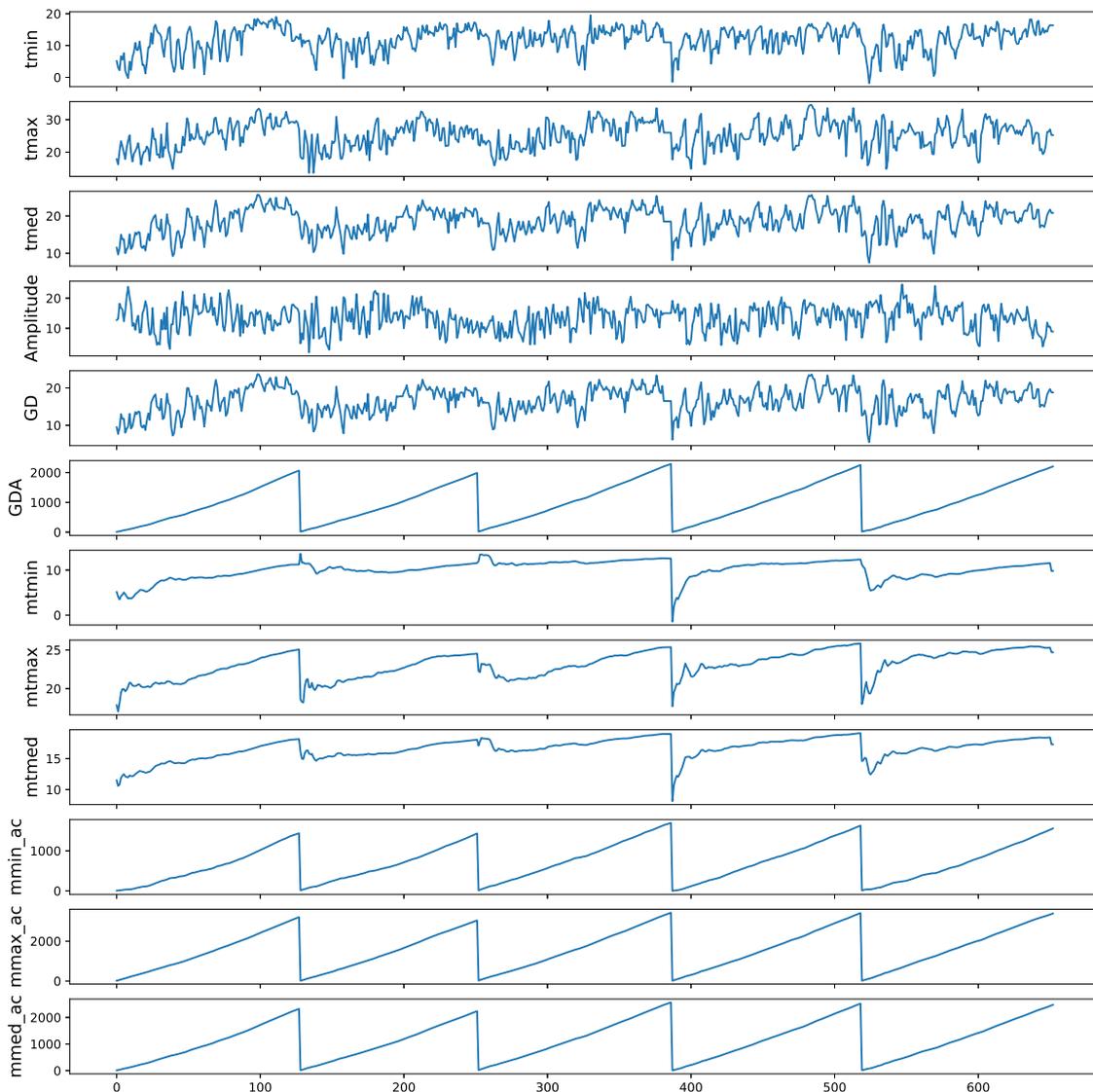
Figura 18 – Séries temporais das variáveis de temperatura 2000 a 2018 - Conjunto de dados ds1



Fonte: Imagens do autor

dos dados. Essas técnicas visam completar o conjunto de dados, possibilitando o trabalho com todos os dados em estudo. Após a análise dos dados constatou-se falhas nas séries de dados, e observou-se que a maioria dos dados faltantes do conjunto utilizado neste trabalho são das séries de dados meteorológicos, e ainda em número bastante reduzido. Considerando este contexto, utilizamos a classe *SimpleImputer* da biblioteca Scikit-learn que trabalha com a média, a mediana e os atributos mais frequentes. O valor padrão da classe *SimpleImputer* utiliza a média. Essa técnica mostrou-se eficiente no preenchimento das lacunas, e o método aplicado na imputação preserva a estrutura, sendo observado que não compromete as características da série original.

Figura 19 – Séries temporais das variáveis de temperatura 2016 a 2020 - Conjunto de dados ds2



Fonte: Imagens do autor

#### 4.4.1.2 Transformação e Normalização dos Dados

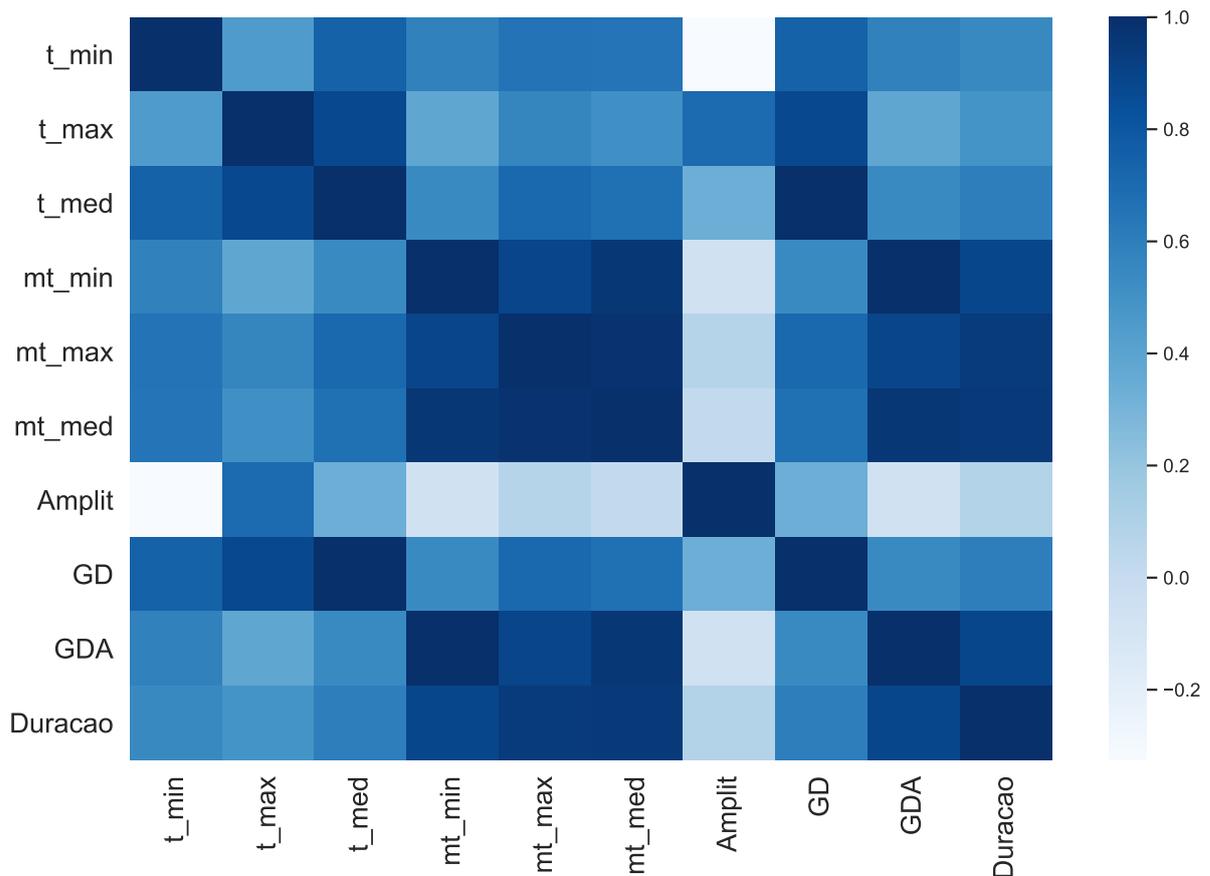
A transformação de dados é uma prática para evitar que o algoritmo seja tendencioso para variáveis com uma ordem de magnitude mais alta. A normalização é uma técnica geralmente aplicada como parte da preparação de dados para aprendizado de máquina. O objetivo da normalização é alterar os valores das colunas numéricas no conjunto de dados para usar uma escala comum, sem distorcer as diferenças nos intervalos de valores ou perder informações. Ao utilizar redes neurais, precisamos que os dados estejam dentro da escala da função de ativação usada pela rede. Utilizamos a classe *MinMaxScaler* da biblioteca *Scikit-learn* para realizar o processo de transformação e normalização do conjunto de dados.

#### 4.4.2 Seleção de Variáveis de Entrada

Nesta etapa, inicialmente, foram identificadas as variáveis de entrada, também chamadas de *features*, consideradas elegíveis para servirem de entrada no modelo. Foram selecionadas 12 variáveis descritas a seguir: temperatura mínima diária ( $t_{min}$ ), temperatura máxima diária ( $t_{max}$ ), temperatura média diária ( $t_{med}$ ), amplitude (*amplit*), média da temperatura mínima diária ( $mt_{min}$ ), média da temperatura máxima diária ( $mt_{max}$ ), temperatura média do período ( $mt_{med}$ ), temperatura mínima acumulada ( $act_{min}$ ), temperatura média acumulada ( $act_{med}$ ), temperatura máxima acumulada ( $act_{tmax}$ ), Graus-dia (*GD*) e Graus-dia acumulado (*GDA*).

Avaliamos o grau de relacionamento entre os pares de variáveis utilizando o coeficiente de correlação de Pearson. Este coeficiente é uma medida de dependência entre duas variáveis quantitativas. A análise de correlação mostra quanto uma variável está, de alguma forma, relacionada com a outra. A Figura 20 mostra graficamente o mapa de calor de correlação de recursos.

Figura 20 – Matriz de correlação entre as variáveis de entrada.



Ajustamos um modelo de regressão linear (ver Anexo A.5) buscando observar a duração do período em função da variável *GDA*. Ao observar o resultado da análise realizada verificou-se que as variáveis de entrada apresentam uma correlação positiva fraca em relação a duração do estágio, que é variável de interesse. A variável que apresenta maior grau de correlação com

a variável duração, é a variável *GDA*. O resultado mostrou uma forte correlação positiva, e o coeficiente de determinação  $R^2$  indicou que 78% da variabilidade na duração do período (número de dias) pode ser explicada pelo acúmulo térmico *GDA*. Através desta análise concluiu-se que o somatório de graus-dia não é o único fator determinante para a ocorrência do fenômeno. Observamos a influência da inclusão de outras variáveis além de *GDA*. Consideramos a análise de correlação na seleção das variáveis, mas realizamos testes para avaliar a importância de outros atributos como preditores, com o objetivo de observar se a variável contribui ou não para melhorar a acurácia do modelo. Os melhores resultados foram obtidos com a combinação de seis variáveis (ver Figura 13), sendo as demais descartadas.

#### 4.4.3 Divisão do Conjunto de Dados em Treino e Teste

De acordo com Hyndman e Athanasopoulos (2021), para a construção de modelos confiáveis de redes neurais é importante avaliar a precisão das previsões usando previsões genuínas. Consequentemente, o tamanho dos resíduos não é uma indicação confiável de quão grandes os erros de previsão podem ser. A precisão das previsões só pode ser determinada considerando o desempenho de um modelo em novos dados que não foram usados ao ajustar o modelo. É uma prática comum separar os dados em duas partes: dados de treinamento e teste. A Figura 21 ilustra essa técnica de divisão. O tamanho do conjunto de teste é normalmente definido com cerca de 20% da amostra total, mas esta proporção pode variar conforme as especificidades do conjunto de dados. Os dados de treinamento são usados para treinar o modelo e os dados de teste são usados para avaliar sua precisão. Os dados de teste fornecem uma indicação de quão bem o modelo provavelmente fará as previsões com base em novos dados. Algumas referências descrevem o conjunto de teste como conjunto de validação, porque esses dados são mantidos fora dos dados usados para o ajuste do modelo. Neste trabalho a divisão do conjunto de dados foi feita separando o conjunto de treino e testes, usando como ponto de corte a safra correspondente.

Figura 21 – Esquema de divisão de dados de treinamento e teste



Fonte: Adaptado de (HYNDMAN; ATHANASOPOULOS, 2021).

## 5 METODOLOGIA DE AVALIAÇÃO

Este capítulo apresenta inicialmente, na subseção 5.1 os detalhes sobre o protótipo do modelo PredHarv. Na seção 5.2 são apresentadas as métricas utilizadas para a avaliação dos resultados e por fim, na seção 5.3 são apresentados os cenários de avaliação.

### 5.1 Protótipo

Um protótipo do modelo foi implementado utilizando a linguagem de programação Python, versão 3.8, utilizando a ferramenta de código aberto Jupyter Notebook e a biblioteca Keras. A biblioteca Keras <sup>1</sup> é uma biblioteca de rede neural de código aberto escrita em Python que pode ser executada no TensorFlow <sup>2</sup>. Projetada para permitir experimentação rápida com redes neurais, se concentra em ser fácil de usar, modular e extensível. Além das redes neurais padrão, Keras tem suporte para redes neurais convolucionais e recorrentes. Tensorflow é uma biblioteca de código aberto para computação numérica e aprendizado de máquina disponibilizada pela equipe do Google Brain <sup>3</sup>. É uma das principais plataformas de aprendizagem profunda da atualidade e fornece implementação em múltiplas CPUs (unidade central de processamento) ou GPUs (unidade de processamento gráfico). Atualmente é utilizada por diversas empresas como: Google, Airbnb, Uber, Dropbox, Intel, entre outras. Os testes do protótipo foram executados em um computador Desktop com processador Intel Core i7 3.60GHz, com 16GB de memória RAM, Sistema operacional Windows 10 64 bits e adaptador de vídeo NVIDIA GeForce GTX 1060 com 6GB de memória.

Usamos dados reais de instituições agrícolas para treinar e testar o modelo. Os testes foram realizados com dados reais, separados em conjuntos de treinamento e testes. O conjunto de dados de teste é porção dos dados que não é conhecida pelo modelo, ou seja, que não foram apresentados ao modelo durante a fase de treinamento da rede. Esses dados são usados para avaliar o comportamento das previsões em relação aos dados reais.

A estrutura da rede neural implementada no protótipo usa um modelo sequencial do Keras, fornecendo uma estrutura em série onde a saída de uma camada serve como entrada para a próxima. Na implementação da rede LSTM, usamos uma camada LSTM com 200 neurônios. O parâmetro `input_shape` foi definido de acordo com a dimensão da camada de entrada, ou seja, o número de variáveis de entrada selecionadas no conjunto de dados e o número de defasagens. A função de ativação usada nesta camada foi a função de ativação linear retificada (ReLU).

Em seguida, outra camada do tipo Densa foi usada, com 1024 neurônios e função de ativação Rectified Linear Unit (ReLU). Finalmente, outra camada do tipo Densa com um neurônio e função de Ativação Linear é usada para retornar um único valor contínuo. Na etapa de com-

---

<sup>1</sup>Disponível em: <https://keras.io/>

<sup>2</sup>Disponível em: <https://www.tensorflow.org/>

<sup>3</sup>Disponível em: <https://research.google.com/teams/brain>

pilação do modelo foi utilizada a função *Adaptive Moment Estimation* (Adam) no parâmetro otimizador. Esta função define como os pesos da rede neural são atualizados. O parâmetro de perda foi definido como Erro Médio Absoluto (MAE). O treinamento da rede nesta etapa foi realizado com 300 épocas.

As funções de ativação são componentes essenciais para dar capacidade representativa às redes neurais artificiais, introduzindo um componente de não linearidade. A função de ativação faz com que as redes possam aprender mais do que relações lineares entre as variáveis dependentes e independentes. Realizamos experimentos com outras funções de ativação, porém os resultados apresentados com ReLU foram superiores.

## 5.2 Métricas de Avaliação

Para que seja possível avaliar o modelo é necessário medir a qualidade das predições, possibilitando a análise das predições, verificando se os valores obtidos com as predições estão com os valores próximos dos dados reais, e ainda comparar os resultados de implementações com diferentes parâmetros. Uma forma utilizada com frequência é fazer a medição do erro, também chamado de resíduo. As métricas de erro mais usadas consistem em sumarizar os erros residuais das previsões, ou seja, agrupar as diferenças entre o valor esperado e a previsão obtida para cada instante de tempo previsto. Avalia-se o erro nos dados de treino para verificar se o modelo tem boa assertividade, e valida-se o modelo verificando o erro nos dados de teste, dados estes que não são conhecidos pelo modelo.

Para medir e avaliar a precisão do modelo, utilizamos as métricas *Mean Absolute Error* (MAE) e *Root Mean Square Error* (RMSE). As métricas são descritas nas Equações 5.1 e 5.2. O conjunto de dados com os valores das observações reais é representado por  $Y$ , e  $\hat{Y}$  representa os valores de previsão. Esses indicadores são baseados em operações para quantificar a diferença entre os valores reais observados e os valores previstos pelo método.

$$MAE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (5.1)$$

$$RMSE = \sqrt{\frac{1}{n} \times \sum_{i=1}^n (Y_i - \hat{Y}_i)^2} \quad (5.2)$$

Existem diferentes métricas para diferentes tipos de tarefas, cada uma delas com características diferentes. RMSE é uma das métricas mais usadas para problemas de regressão. É uma métrica interpretativa. A métrica MAE nos permite visualizar a média do erro absoluto.

### 5.3 Cenários de Avaliação

Para avaliar o modelo PredHarv, foram estabelecidos diferentes cenários de avaliação. A utilização destes cenários visa realizar uma estimativa de desempenho, fornecer informações sobre o funcionamento do modelo e ainda como ele pode ser aprimorado. A avaliação do modelo foi realizada considerando os cenários descritos a seguir.

#### 5.3.1 Cenário 1: Validação *Walk-forward*

O primeiro cenário de avaliação foi definido usando a metodologia de validação chamada *Walk-forward*. Essa metodologia consiste em dividir os dados de treinamento e teste, movendo uma janela, uma etapa de cada vez, ao longo da série temporal, por meio da técnica chamada de janela deslizante (*window sliding*). Ao avaliar um modelo para predição de série temporal, estamos interessados no desempenho do modelo em dados que não foram usados para treiná-lo, chamados de dados fora da amostra. A Figura 22 ilustra o funcionamento da metodologia. As observações na cor cinza, chamadas de dados dentro da amostra, são dados do conjunto de dados de treinamento, e as observações na cor azuis, chamadas de dados fora da amostra, são os dados do conjunto de testes que não foram apresentados ao modelo. O conceito *Walk-forward* amplia a utilização do método de validação cruzada, permitindo avaliar o modelo com menor interferência de ruídos e padrões que podem ocorrer em um determinado período, mas que não fazem parte do todo. Outro aspecto a ser considerado é a manutenção da temporalidade dos dados, uma vez que nas séries temporais é necessário respeitar a ordem temporal das análises. A validação cruzada *K-fold* é uma forma básica de validação cruzada para avaliar a generalização de um modelo a partir de um conjunto de dados (REFAEILZADEH; TANG; LIU, 2015). A validação cruzada *K-fold* divide sistematicamente os dados em  $k$  grupos. Esse método pode ser problemático quando fornecido diretamente com dados de séries temporais, pois assume que cada observação é independente e que, de certa forma, não há relação entre elas.

Na implementação do cenário proposto, o conjunto de dados é dividido em subconjuntos compostos com dados de dez anos para treinamento e um ano para teste. O ponto de corte que define o subconjunto de teste é exatamente o início do período da floração plena, até o final do período de maturação, que equivale a uma safra. Na amostra seguinte, outro ano é selecionado como subconjunto de teste. No final, para cada amostra, as métricas RMSE e MAE descritas na seção 5.2 são calculadas.

#### 5.3.2 Cenário 2: Dados Reais x Predições

No segundo cenário de avaliação, executamos testes usando dados reais dos dois conjuntos de dados. Usamos amostras dos conjuntos de dados para comparar a duração dos períodos com os valores reais e com os valores de predição retornados pelo modelo. Neste cenário estamos

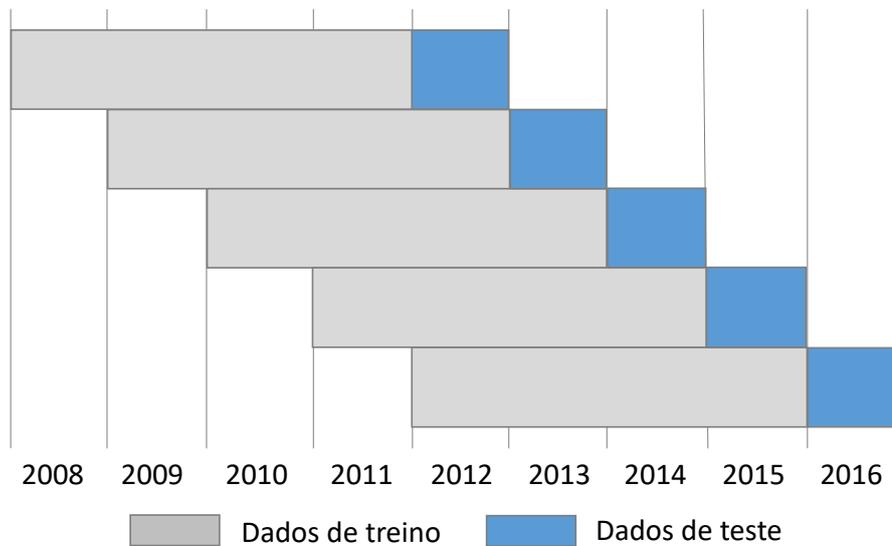


Figura 22 – Representação da metodologia de validação *Walk-forward*

avaliando a duração de todo o período. As amostras são equivalentes ao início do período da floração plena, até o final do período de maturação, que equivale a uma safra (ano). O ponto de corte que define o subconjunto de treinamento e teste, é referente ao início do período da floração plena até o final do período de maturação, da amostra que está sendo utilizada. No subconjunto de dados de treinamento são utilizados dados de todo o conjunto de dados, exceto do período da amostra que está sendo prevista. No final, para cada amostra, as métricas RMSE e MAE descritas na seção 5.2 são calculadas.

### 5.3.3 Cenário 3: Dados Sintéticos

No terceiro cenário de avaliação, usamos dados sintéticos. Seleccionamos uma amostra do conjunto de dados DS1. Foram feitas variações em algumas janelas de tempo em determinados períodos, com aumento ou diminuição dos valores das variáveis de temperatura. No subconjunto de dados de treinamento são utilizados dados de todo o conjunto de dados, exceto do período da amostra que está sendo prevista. O objetivo deste cenário de avaliação é visualizar o comportamento do método com mudanças guiadas nos dados de entrada.

### 5.3.4 Cenário 4: Comparativo com Outros Métodos

Nesse cenário de avaliação, realizamos testes com outros algoritmos de ML aplicados ao problema. Este cenário de avaliação se concentra na seleção do melhor algoritmo de ML supervisionado e no ajuste de seus hiperparâmetros. Usamos métodos *Automated Machine Learning* (AutoML) para selecionar os melhores modelos de um conjunto de modelos e otimizar seus hiperparâmetros. Para produzir uma comparação justa entre os modelos, as avaliações foram

divididas em categorias: Seleção de algoritmos de ML geral (GML) e *Deep Learning* (DL). Na categoria GML, usamos Auto-Sklearn <sup>1</sup>, que é uma biblioteca AutoML construída sobre a estrutura Scikit-Learn ML. A escolha de algoritmos e hiperparâmetros implementados pelo Auto-Sklearn aproveita os recentes avanços em otimização Bayesiana, meta-aprendizagem e *Ensemble Learning* (FEURER et al., 2015). Usamos o Auto-Sklearn versão 0.14.4. Na categoria DL, usamos Auto-Keras <sup>2</sup>, que é uma biblioteca escrita em *Python* baseada no Keras. A biblioteca é focada em *Neural Architecture Search* (NAS) que realiza pesquisa automática de arquitetura de redes neurais (JIN; SONG; HU, 2019). A busca é realizada usando uma otimização Bayesiana, com a ferramenta ajustando automaticamente o número de camadas densas, unidades, tipo de funções de ativação usadas, valores de dropout e outros hiperparâmetros. Neste trabalho, adotamos o Auto-Keras versão 1.0.18.

---

<sup>1</sup>disponível em: <https://www.automl.org/automl/auto-sklearn/> acessado em 13 de fevereiro de 2022

<sup>2</sup>disponível em: <https://autokeras.com/> acessado em 13 de fevereiro de 2022



## 6 RESULTADOS

Este capítulo apresenta os resultados dos experimentos realizados nos cenários descritos no capítulo 5. Na seção 6.1 são apresentados os resultados obtidos com a metodologia de validação *Walk-forward*. Em seguida na seção 6.2 são apresentados os resultados obtidos no segundo cenário de avaliação que estabelece um comparativo entre os dados reais e os resultados das previsões. Na seção 6.3 são apresentados os resultados com dados sintéticos. Na seção 6.4 os resultados obtidos com outros algoritmos de ML aplicados ao problema. Por fim, na seção 6.5 são discutidos aspectos relacionados aos resultados de forma geral.

### 6.1 Cenário de Avaliação 1: Validação *Walk-forward*

A tabela 4 mostra os resultados do teste do primeiro cenário de avaliação obtidos com o modelo PredHarv. A coluna Dados de treino mostra os dados usados no treinamento e a coluna Dados de teste mostra os dados usados nos testes. Estes dados referem-se a valores relacionados as safras (em anos), de acordo com a divisão do conjunto de dados descrita na seção 5.3. Os valores das métricas RMSE e MAE são mostrados para os testes realizados usando a implementação com MLP e LSTM. É possível observar que os resultados com MLP são consideravelmente inferiores aos obtidos com LSTM em ambas as etapas. Este resultado inferior das redes MLP é esperado, uma vez que a comparação é injusta se considerarmos as características das redes *feedforward*. Observou-se nos resultados com MLP, muito evidente na métrica RMSE a ocorrência de um sobreajuste (*overfitting*). Isso ocorre quando a lacuna entre o erro de treinamento e o erro de teste é muito grande. Bons resultados foram observados em erros de treinamento, mas ele não respondeu de forma eficiente à generalização, que é a capacidade de executar bem em entradas não apresentadas anteriormente ao modelo. Os resultados dos testes usando LSTM são significativamente melhores. Os valores da métrica RMSE permaneceram na média de 0,81, com desvio padrão de 0,28, e os valores da métrica MAE permaneceram na média de 0,75 com desvio padrão de 0,26. O desvio padrão é uma medida que expressa o grau de dispersão de um conjunto de dados, ou seja, indica o quanto estes dados são uniformes. Na média geral, os resultados da métrica RMSE apresentam uma diferença de 9,34 e na métrica MAE de 9,41, comparando os resultados obtidos com MLP e com LSTM. Observando os resultados com o LSTM, podemos concluir que o modelo possui uma boa capacidade de generalização e que os resultados apresentam uniformidade.

### 6.2 Cenário de Avaliação 2: Dados Reais x Previsões

A Tabela 5 mostra os resultados dos testes do segundo cenário de avaliação, obtidos com o conjunto de dados DS1. A coluna Real mostra a duração dos dados observados de amostras no conjunto de dados reais. Esses valores referem-se à duração do período em dias referente a uma

Tabela 4 – Resultados da metodologia de validação *Walk Forward*

Amostra	Dados de treino	Dados de teste	MLP		LSTM	
			RMSE	MAE	RMSE	MAE
1	2000-2010	2011	6,02	5,88	<b>0,95</b>	<b>0,89</b>
2	2001-2011	2012	22,76	25,04	<b>0,99</b>	<b>0,96</b>
3	2002-2012	2013	17,42	18,43	<b>0,96</b>	<b>0,77</b>
4	2003-2013	2014	12,07	9,69	<b>0,31</b>	<b>0,28</b>
5	2004-2014	2015	5,40	5,16	<b>1,13</b>	<b>1,01</b>
6	2005-2015	2016	7,18	6,85	<b>0,99</b>	<b>0,93</b>
7	2006-2016	2017	5,91	5,57	<b>0,49</b>	<b>0,41</b>
8	2007-2017	2018	4,83	4,51	<b>0,71</b>	<b>0,66</b>
Média			10,19	10,14	<b>0,81</b>	<b>0,73</b>
Desvio Padrão			6,65	7,53	<b>0,28</b>	<b>0,26</b>

safr (ano). A coluna Predição mostra os dados previstos pelo modelo PredHarv. Esses valores na coluna Predição, bem como as métricas RMSE e MAE, são calculados após dez execuções de cada amostra. Comparando os resultados dos valores reais com os valores previstos pelo modelo PredHarv, apresentados na Tabela 5, observa-se que os resultados são muito próximos, apresentando um erro absoluto entre 0,3 e 1,9 nas amostras selecionadas. Os valores da métrica RMSE permaneceram na média de 1,41, com desvio padrão de 0,82, e os valores da métrica MAE permaneceram na média de 1,23 com desvio padrão de 0,70. A métrica RMSE teve valor máximo de 2,43 e MAE de 2,19. Estes resultados demonstram que o modelo PredHarv obtém uma boa capacidade preditiva diante deste cenário de avaliação.

Tabela 5 – Resultados das predições obtidas com o segundo cenário de avaliação - Conjunto de dados DS1

Amostra	Real	Predição	RMSE	MAE
1	104	104,3	0,35	0,34
2	108	108,5	0,99	0,94
3	110	108,7	2,37	1,97
4	115	114,3	1,02	0,89
5	122	121,2	1,32	1,08
Média			<b>1,21</b>	<b>1,04</b>
Desvio Padrão			<b>0,74</b>	<b>0,59</b>

A Tabela 6 mostra os resultados dos testes do segundo cenário de avaliação obtidos com o conjunto de dados DS2. A coluna Real mostra a duração dos dados observados de amostras no conjunto de dados reais. Esses valores referem-se à duração do período em dias referente a uma safra (ano). A coluna Predição mostra os dados previstos pelo modelo PredHarv. Esses valores na coluna Predição, bem como as métricas RMSE e MAE, são calculados após dez execuções

de cada amostra.

Tabela 6 – Resultados das previsões obtidas com o segundo cenário de avaliação - Conjunto de dados DS2

Amostra	Real	Predição	RMSE	MAE
1	128	127,4	0,45	0,43
2	124	123,3	0,50	0,39
3	135	135,8	0,41	0,33
4	132	132,7	1,32	1,08
5	134	133,3	1,31	1,02
Média			<b>0,79</b>	<b>0,65</b>
Desvio Padrão			<b>0,47</b>	<b>0,36</b>

### 6.3 Cenário de Avaliação 3: Dados Sintéticos

A Figura 23 mostra a dinâmica das previsões obtidas com o modelo PredHarv. Nesse cenário, foram utilizados dados sintéticos com variações de temperatura em determinados períodos. A previsão inicial foi feita em 08/10, data que marca o início do período de plena floração. Com base nos dados alimentados, a duração estimada foi de 121 dias. A linha azul, na Figura 23, identifica o valor previsto no estágio inicial. Posteriormente, novas previsões foram feitas, chamadas de previsões de ajuste. A primeira previsão de ajuste foi feita em 8/11, 31 dias após o início do período. Foi simulado um período de duas semanas, entre as datas 20/10 e 02/11 com temperaturas 20% superiores às inicialmente previstas. A segunda previsão de ajuste foi feita em 20/11, 43 dias após o início do período. Nesse período, as temperaturas entre 12/11 e 17/11 foram simuladas 10% abaixo do previsto inicialmente. A terceira previsão de ajuste foi feita em 30/12, 83 dias após o início do período. Entre 01/12 e 19/12, as temperaturas foram simuladas 10% abaixo do inicialmente previsto. Na quarta e última previsão de ajuste, realizada em 20/01, 107 dias após o início do período, foram simuladas temperaturas 20% abaixo da inicialmente prevista. Como pode ser observado na Figura 23, a linha vermelha mostra a duração do período correspondente à data em que a previsão foi feita. Esses valores variam e pode-se perceber que o modelo responde ao estímulo de mudanças nos dados de entrada, ajustando a previsão conforme os dados vão sendo introduzidos. Variações de temperatura em determinados períodos, resultam em um maior acúmulo de calor, encurtando a duração do período. Em outras ocasiões, ocorre menos acúmulo, resultando em um aumento na duração conforme o esperado.

### 6.4 Comparativo de Outros Métodos Aplicados ao Problema

Neste cenário de avaliação, usamos métodos AutoML para selecionar o melhor algoritmo bem o melhor ajuste de hiperparâmetros aplicados ao conjunto de dados DS1. Usamos amostras

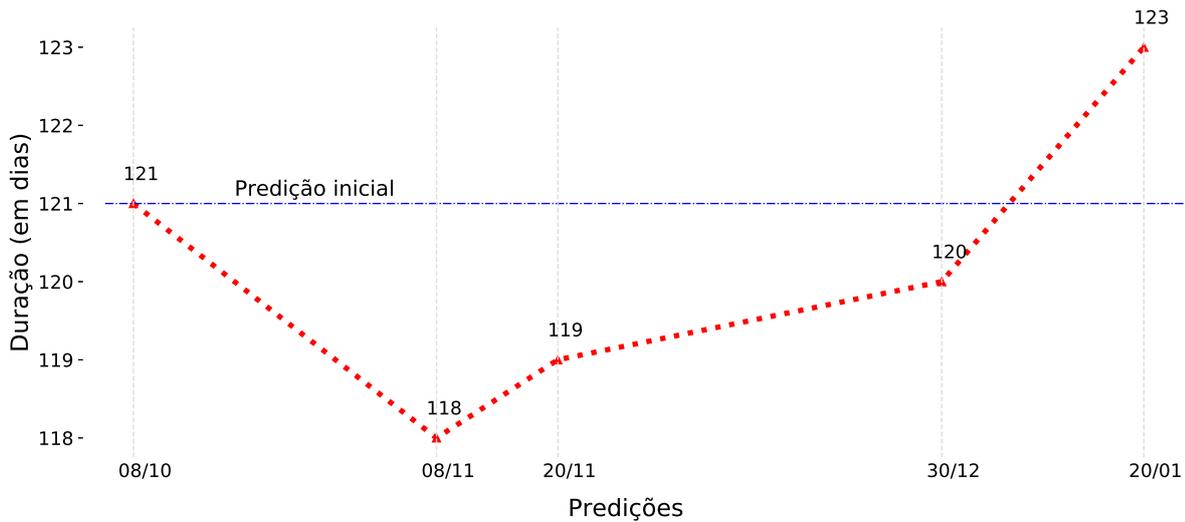


Figura 23 – Dinâmica das previsões obtidas com o modelo PredHarv

do conjunto de dados utilizando o esquema de validação cruzada *K-fold*, com  $k = 10$ . Dividimos as soluções em duas categorias. Na categoria GML, usamos o *Auto-Sklearn*. O melhor resultado no ranking dos modelos testados pelo *AutoSklearn* foi com Regressão Linear Múltipla (MR). A Tabela 7 apresenta os resultados obtidos a partir do modelo.

Tabela 7 – Melhores resultados categoria GML usando AutoSklearn.

Amostra	RMSE	MAE
1	3,68	3,47
2	1,54	1,16
3	1,23	0,95
4	3,92	3,41
5	2,56	2,16
Média	<b>2,58</b>	<b>2,23</b>
Desvio padrão	<b>1,21</b>	<b>1,19</b>

Na categoria DP, usamos *AutoKeras*. A biblioteca é focada na busca automática da arquitetura neural DL (NAS), ajustando automaticamente o número de camadas densas, unidades, funções de ativação usadas, valores de dropout e outros hiperparâmetros. Essa tarefa é frequentemente chamada de pesquisa de arquitetura e é frequentemente chamada de arquitetura neural pesquisar. A estrutura de rede neural retornada pelo *Autokeras* é composta por um modelo sequencial composto por uma camada Densa com 128 neurônios e função de ativação ReLU. Em seguida, há outra camada do tipo Dense com 1024 neurônios e função de ativação ReLU. Na etapa de saída, há uma camada Densa com apenas um neurônio. O ajuste de hiperparâmetros foram testados 134 parâmetros. O treinamento da rede foi realizado com 300 épocas. A Tabela 8 apresenta os resultados das métricas obtidas das amostras submetidas ao modelo.

Tabela 8 – Melhores resultados categoria DL usando AutoKeras.

Amostra	RMSE	MAE
1	4,13	3,78
2	2,94	3,31
3	1,97	1,69
4	4,46	3,67
5	1,96	1,55
Média	<b>3,09</b>	<b>2,80</b>
Desvio padrão	<b>1,17</b>	<b>1,09</b>

Nos resultados obtidos na categoria GML usando AutoSklearn com MR (Tabela 7), a métrica RMSE obteve um valor médio de 2,58 nas amostras selecionadas.

As métricas da categoria DL obtidas com AutoKeras (Tabela 8) obteve um valor médio da métrica RMSE de 3,09. Comparando esses resultados, o método da categoria GML obteve melhor resultado com diferença de 0,51. No entanto, observou-se que a medida de desvio padrão, que é uma medida de dispersão em torno da média, mostra que os resultados obtidos na categoria DL são mais uniformes.

Os métodos de AutoML testados, Auto-sklearn e Autokeras, são métodos totalmente automatizados. São modelos "caixa preta", e sua estrutura interna é desconhecida, limitando-se a medidas de relações de entrada e saída. No entanto, esses métodos buscam um vasto espaço de modelos e constroem conjuntos complexos de alta precisão, construindo modelos baseados nos dados, o que colabora com a finalidade para a qual foram submetidos neste trabalho. As investigações de outros métodos que podem ser aplicados ao problema demonstram que a escolha do método do modelo proposto é adequado à natureza do problema.

## 6.5 Discussão

Modelos lineares de soma térmica utilizando somatório de graus-dia tem sido utilizados para estimar a duração de períodos fenológicos de macieiras e de outras culturas. Estes modelos quando ajustados para estimar subperíodos fenológicos de macieiras, apresentam resultados com uma menor variabilidade no acúmulo de graus-dia para a maioria dos estádios fenológicos entre safras, em comparação com o número de dias, respondendo melhor a soma térmica do que ao tempo cronológico (dias de calendário) (CARDOSO, 2011) (ANZANELLO; CHRISTO, 2019). De acordo com GOODFELLOW; BENGIO; COURVILLE (2016), modelos lineares têm o defeito óbvio de que a capacidade do modelo é limitada a uma função linear, e então, o modelo pode não explicar a interação entre duas variáveis de entrada quaisquer. O modelo computacional PredHarv utiliza como linha de base modelos de soma térmica, mas usa uma abordagem multivariada. Usamos a soma térmica relacionando o acúmulo de graus-dia com a duração do período e outras variáveis relacionadas a temperatura. A inclusão de outras

variáveis, com a estratégia multivariada associada ao uso de um método de ML como proposto neste trabalho, melhora significativamente a resposta, como pode ser observado nos gráficos comparativos de previsões com o modelo PredHarv e o modelo linear, nas Figuras 24 e 26, conseguindo explicar melhor a ocorrência do fenômeno. Este desempenho superior da abordagem multivariada era esperado, considerando que as estratégias multivariadas têm mais informações sobre o problema. Uma propriedade importante das redes neurais é a habilidade de aprender de seu ambiente e com isso melhorar seu desempenho. As variáveis selecionadas, que servem de entrada para o modelo PredHarv, embora algumas não apresentem significância estatística quando observadas as relações lineares, ao serem incorporadas ao modelo apresentam melhora no resultado final da previsão. Essa melhora no resultado pode ser atribuída em partes as funções de ativação utilizadas na construção de redes neurais. As funções introduzem nas redes um componente não linear, o que faz com que elas possam aprender mais do que relações lineares entre as variáveis dependentes e independentes.

Outra característica do modelo proposto, é que método de ML utilizado explora o potencial das redes LSTM, um tipo especial de RNN para tratar problemas que envolvem séries temporais. As previsões do modelo PredHarv são feitas a partir do estágio fenológico da plena floração, com base nas séries históricas da fenologia e dados meteorológicos. Essas séries temporais são um conjunto de observações ordenadas no tempo, e possuem uma dependência temporal entre as observações. As redes recorrentes aplicadas a este tipo de problema têm essa capacidade de capturar parte da informação implícita nas séries, refletindo o impacto de uma sequência de estímulos ambientais no desenvolvimento da planta.

Os cenários de avaliação propostos e os testes comparativos entre os modelos evidenciam que o modelo PredHarv é uma alternativa útil, capaz de melhorar a acurácia do resultado das previsões. Na média observada dos resíduos das amostras selecionadas do conjunto de dados DS1, obteve-se 0,7 dias com o modelo PredHarv e 4,3 dias com o modelo linear, comparados aos dados reais. Nos resultados obtidos com o conjunto de dados DS2 obteve-se 0,7 dias com o modelo PredHarv e 1,7 dias com o modelo linear, comparados aos dados reais. Os gráficos apresentados nas Figuras 25 e 27 ilustram os resíduos representados pelo erro absoluto. O erro absoluto é o valor absoluto da diferença entre o valor real e o valor predito da amostra.

Devido a característica das redes neurais de aprenderem com o ambiente, o modelo proposto tem potencial para ser adaptado na previsão de outros estágios fenológicos da macieira ou ainda em diferentes culturas. Uma vez que se tenha disponibilidade de dados históricos para treinamento do modelo, este pode ser facilmente adaptado para o uso com o objetivo de melhorar a acurácia do resultado das previsões.

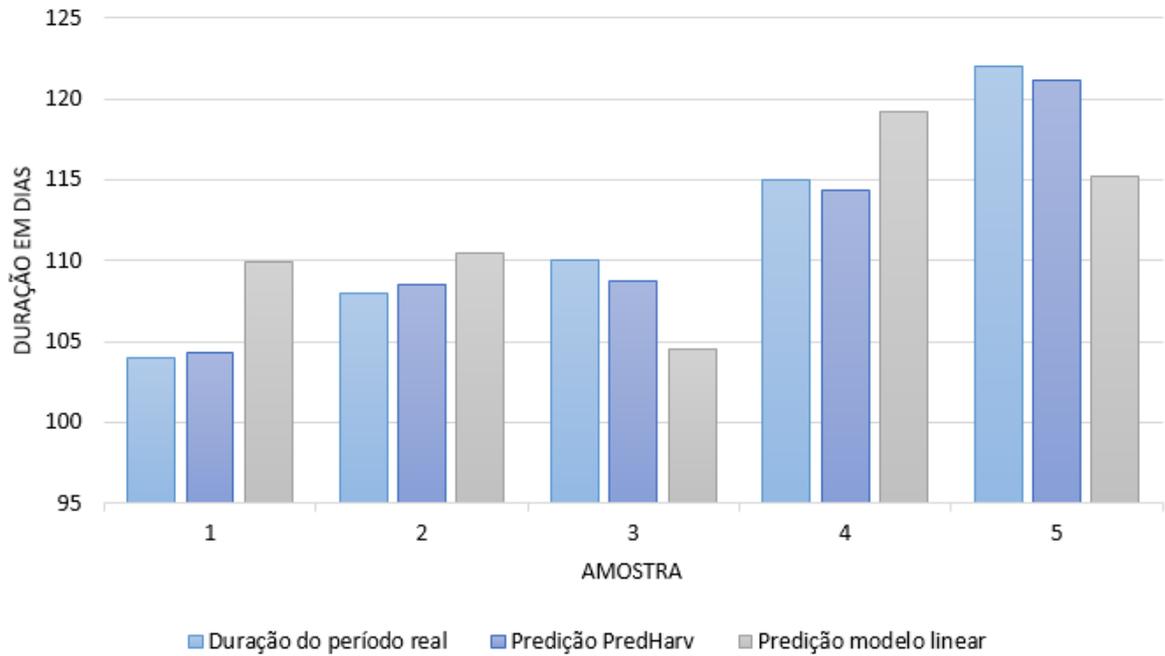


Figura 24 – Resultados das previsões do modelo PredHarv e do modelo Linear - Conjunto de dados DS1

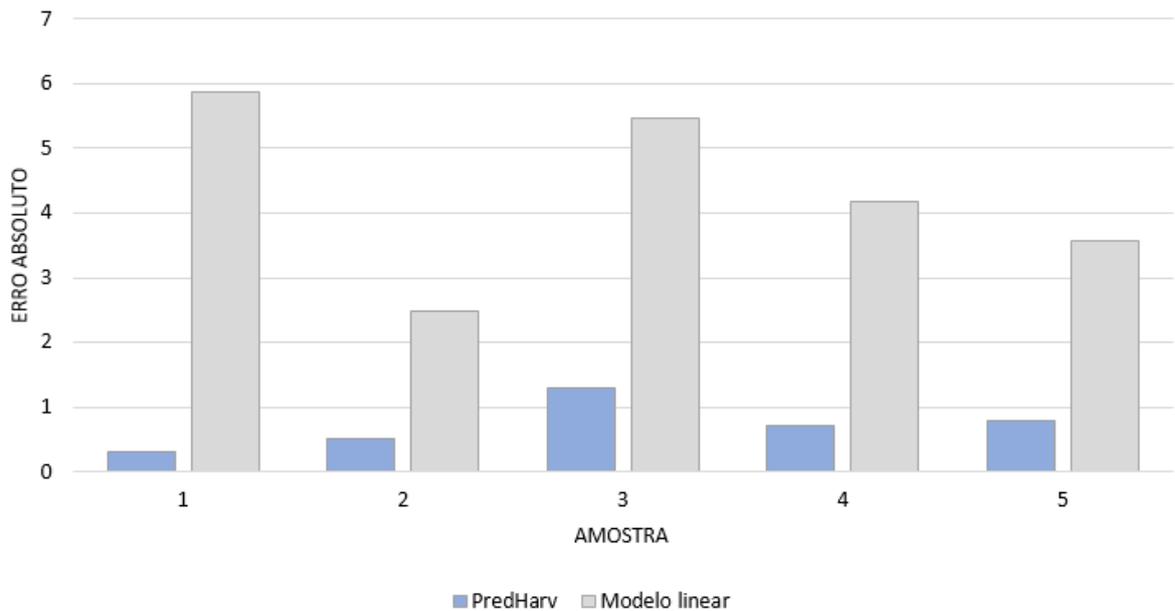


Figura 25 – Resíduos das previsões do modelo PredHarv e do modelo Linear - Conjunto de dados DS1

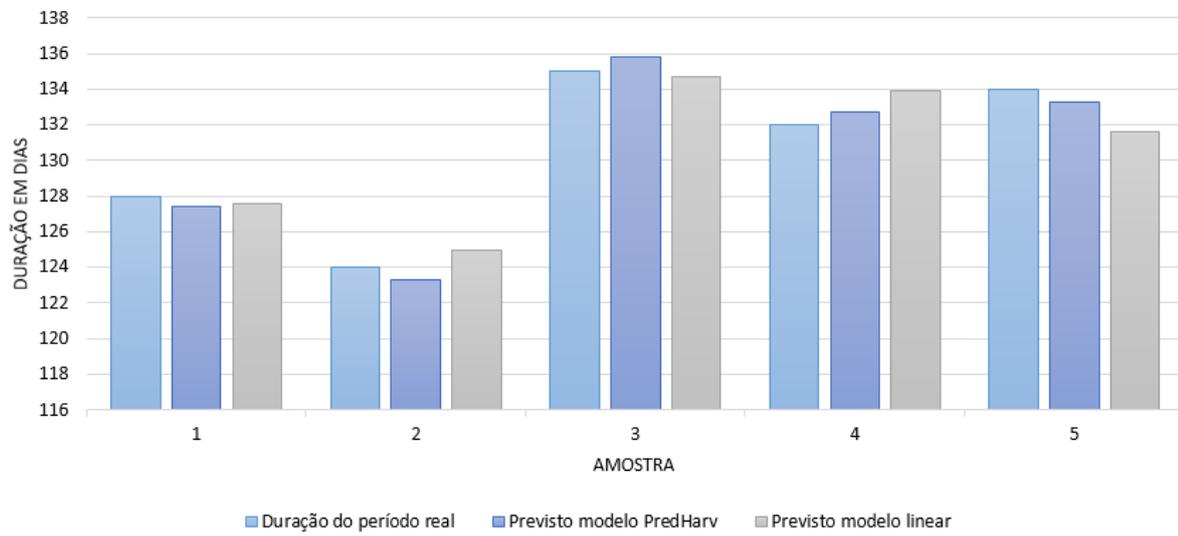


Figura 26 – Resultados predições do PredHarv e modelo Linear - DS2

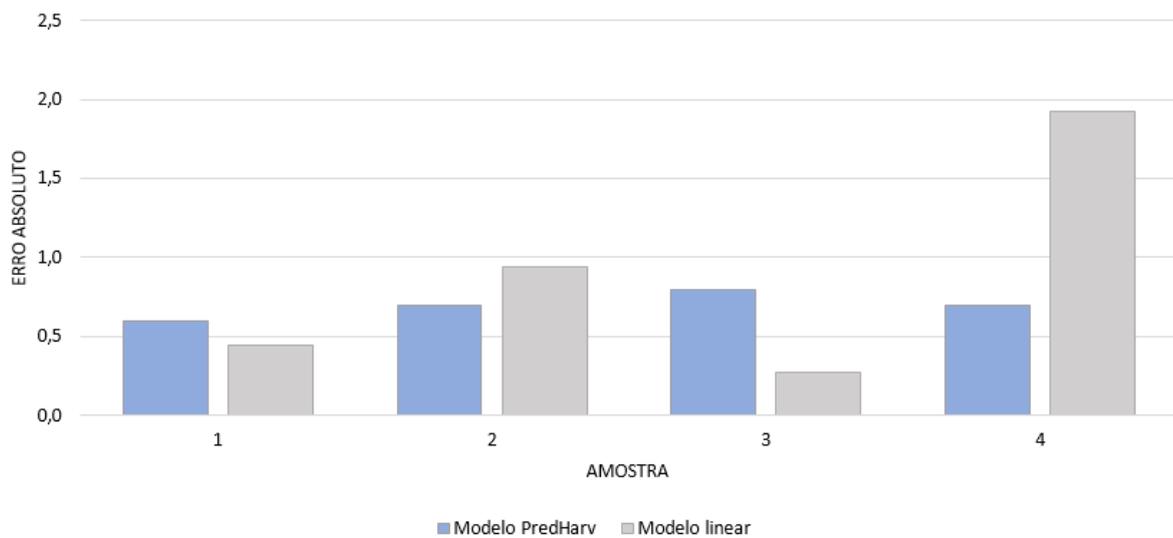


Figura 27 – Resíduos das predições do modelo PredHarv e do modelo Linear - Conjunto de dados DS2

## 7 CONSIDERAÇÕES FINAIS

Este trabalho propôs uma nova abordagem para predição da data de início da colheita de macieiras dadas as condições climáticas relacionadas à temperatura esperadas para o período. O modelo computacional PredHarv é baseado em modelos de soma térmica, mas usa uma abordagem multivariada. Usamos a soma térmica relacionando-a com a duração do período e outras variáveis relacionadas a temperatura do período. Abordagens usando ML têm sido usadas em várias áreas. No campo da pesquisa agropecuária, ML tem sido utilizado para aumentar a produtividade agrícola e minimizar seu impacto ambiental, mostrando-se uma importante ferramenta de apoio à tomada de decisões.

As predições do modelo são feitas a partir do estágio fenológico da plena floração, com base nas séries históricas da fenologia e dados meteorológicos. Utilizamos um método de ML, explorando o potencial das redes LSTM, um tipo especial de RNN para tratar problemas que envolvem séries temporais.

Um protótipo do modelo PredHarv foi desenvolvido e submetido a diferentes cenários de avaliação. Os resultados foram avaliados de acordo com as métricas RMSE e MAE. Nos testes do primeiro cenário de avaliação do PredHarv, o modelo demonstrou uma boa capacidade de generalização quando submetido a novos dados e resultados superiores à implementação com MLP. No segundo cenário, em testes com dados reais, PredHarv apresentou resultados muito próximos dos valores reais. No terceiro cenário, quando submetido aos dados sintéticos, os resultados mostraram que o modelo PredHarv apresentou uma resposta positiva aos estímulos causados pelas alterações feitas nos dados de entrada, ajustando a previsão à medida que os dados vão sendo introduzidos. No quarto cenário, utilizamos ferramentas de AutoML, para selecionar os melhores modelos de um conjunto de modelos, bem como a otimização de hiperparâmetros. Dividimos as avaliações em duas categorias e os resultados obtidos em ambas as categorias foram inferiores aos obtidos com o modelo proposto.

A utilização de RNNs, em especial redes LSTM, mostrou-se uma alternativa viável para predição da data de início da colheita de macieiras. Observando os resultados obtidos com os cenários de avaliação propostos validamos a sentença descrita como hipótese de que a data de colheita da macieira pode ser estimada através de modelos multivariados baseados em soma térmica, utilizando métodos de ML com base em dados históricos da fenologia e parâmetros climáticos de temperatura, podendo melhorar a acurácia das predições, explicando melhor a relação entre as variáveis de temperatura e a quantidade de calor necessária para a planta completar seu ciclo.

### 7.1 Contribuições

As contribuições científicas deste trabalho incluem:

- O desenvolvimento de um modelo computacional de ML utilizando uma abordagem mul-

tivariada baseada na soma térmica, contribui com a área da fruticultura, possibilitando antecipar informações sobre a data de colheita, possibilitando ao produtor um melhor planejamento das atividades no pomar, evitando custos, gerando economia financeira e melhorando a produtividade;

- A aplicação de métodos de ML com a capacidade de capturar a dependência temporal implícita nas séries históricas, visando melhorar a eficácia de modelos baseados em soma térmica voltados para a fruticultura;
- O modelo contribui com uma metodologia de utilização, com o objetivo de ampliar a capacidade preditiva, reduzindo a incerteza implícita nas informações fornecidas pelo usuário, permitindo flexibilidade de utilização do modelo;
- Tornar o uso de modelos preditivos na fruticultura mais efetivos possibilitando a prospecção de cenários futuros e simulações de alterações climáticas.

## 7.2 Limitações

Problemas que envolvem predição de séries temporais usando redes neurais tem a necessidade de dados históricos para o treinamento da rede, limitando seu uso a pomares com algum histórico, o que impossibilita sua aplicação em áreas novas de plantio, ainda sem histórico de produção. A utilização apenas de dados de temperatura pode ser considerada uma limitação do modelo proposto, uma vez que não é possível determinar com precisão a influência de outros fatores externos que podem influenciar nas complexas interações entre o ambiente e a planta.

## 7.3 Trabalhos Futuros

Para trabalhos futuros, uma possível extensão poderia ser desenvolvida para usar dados da própria previsão do tempo. O modelo PredHarv pode ser facilmente adaptado para prever outras fases fenológicas e para diferentes culturas. Também pode ser considerado o uso de outros dados que podem ser incorporados ao modelo para melhorar a confiabilidade da resposta, como o uso de imagens e informações sobre parâmetros biológicos da planta.

## 7.4 Trabalhos Publicados

Boechel, T. et al. Fuzzy time series for predicting phenological stages of apple trees. In: SAC 2021 - 36TH ACM/SIGAPP SYMPOSIUM ON APPLIED COMPUTING, MAR 2021, GWANGJU / VIRTUAL, SOUTH KOREA, 2021. p. 934–941.

Boechel, T.; Policarpo, L.M.; Ramos, G.d.O.; da Rosa Righi, R.; Singh, D. Prediction of Harvest Time of Apple Trees: An RNN-Based Approach. Algorithms 2022.

## REFERÊNCIAS

- ABPM. ASSOCIAÇÃO BRASILEIRA DE PRODUTORES DE MAÇÃ. Inovação em pauta. Fraiburgo, 2018. Disponível em: <<http://www.abpm.org.br>>. Acesso em: Nov. 2018.
- AGAPOMI. ASSOCIAÇÃO GAÚCHA DE PRODUTORES DE MAÇÃ. Produção de maçã - safra 2017/2018. Vacaria-RS, 2018. Disponível em: <<http://agapomi.com.br/>>. Acesso em: nov. 2018.
- AGAPOMI. ASSOCIAÇÃO GAÚCHA DE PRODUTORES DE MAÇÃ. Produção de maçã - safra 2019/2020. Vacaria-RS, 2021. Disponível em: <<http://www.agapomi.com.br/>>. Acesso em: jan. 2021.
- ALPAYDIN, E. Introduction to machine learning. 4th ed.. ed. ]: MIT Press, 2020.
- ANZANELLO, R.; CHRISTO, M. C. de. Temperatura base inferior, soma térmica e fenologia de cultivares de videira e quivizeiro. Revista de Ciências Agroveterinárias, v. 18, n. 3, p. 313–322, 2019.
- ARNOLD, C. Y. The determination and significance of the base temperature in a linear heat unit system. In: SOCIETY FOR HORTICULTURAL SCIENCE, 1959. Proceedings... 1959. v. 74, n. 1, p. 430–445.
- BERGAMASCHI, H. O clima como fator determinante da fenologia das plantas. Fenologia: ferramenta para conservação, melhoramento e manejo de recursos vegetais arbóreos. Colombo: Embrapa Florestas, v. 1, p. 291–310, 2007.
- BIOLCHINI, J. et al. Systematic review in software engineering. System Engineering and Computer Science Department COPPE/UFRJ, Technical Report ES, v. 679, n. 05, p. 45, 2005.
- BLAZEK, J.; PISTEKOVA, I. Prediction of the harvesting time for four apple cultivars on the basis of beginning of flowering and attaining of t-stage of fruitlets and dependence of diameter of fruitlets at t-stage and fruits at ripening stage. Journal of Horticultural Research, v. 25, n. 1, p. 55–59, 2017.
- BOECHEL, T. et al. Fuzzy time series for predicting phenological stages of apple trees. In: SAC 2021 - 36TH ACM/SIGAPP SYMPOSIUM ON APPLIED COMPUTING, MAR 2021, GWANGJU / VIRTUAL, SOUTH KOREA., 2021. Proceedings... 2021. p. 934–941.
- BROICH, M. et al. A spatially explicit land surface phenology data product for science, monitoring and natural resources management applications. Environmental Modelling & Software, v. 64, p. 191–204, 2015.
- BRUNINI, O. et al. Temperatura-base para alface cultivar "white boston", em um sistema de unidades térmicas. Bragantia, v. 35, n. 1, p. 213–219, 1976.
- CARDOSO, L. S. Modelagem aplicada à fenologia de macieiras 'royal gala' e 'fugi suprema' em função do clima, na região de vacaria, rs. 2011. Dissertação (Mestrado em Ciência da Computação) — , 2011.

CEPA. CENTRO DE SOCIOECONOMIA E PLANEJAMENTO AGRÍCOLA. Síntese anual da agricultura de santa catarina 2016-2017. Florianópolis-SC, 2017. Disponível em: <<https://bit.ly/2CUOj9A/>>. Acesso em: fev. 2019.

CHEN, C. et al. Research on dynamic forecast of flowering period based on multivariable lstm and ensemble learning classification task. Agricultural Sciences, v. 11, n. 09, p. 777, 2020.

CHENG, H. et al. Early yield prediction using image analysis of apple fruit and tree canopy features with neural networks. Journal of Imaging, v. 3, n. 1, p. 6, 2017.

CHITU, E.; PALTINEANU, C. Timing of phenological stages for apple and pear trees under climate change in a temperate-continental climate. International journal of biometeorology, v. 64, n. 8, p. 1263–1271, 2020.

CZERNECKI, B.; NOWOSAD, J.; JABLONSKA, K. Machine learning modeling of plant phenology based on coupling satellite and gridded meteorological dataset. International journal of biometeorology, v. 62, n. 7, p. 1297–1309, 2018.

DAI, W. et al. Detecting temporal changes in the temperature sensitivity of spring phenology with global warming: application of machine learning in phenological model. Agricultural and Forest Meteorology, v. 279, p. 107702, 2019.

DARBYSHIRE, R. et al. A global evaluation of apple flowering phenology models for climate adaptation. Agricultural and Forest Meteorology, v. 240, p. 67–77, 2017.

DATA SCIENCE ACADEMY, . Deep learning book. 2022. Disponível em: <http://www.deeplearningbook.com.br>. Acesso em: 14 mar.2022.

EL YAACOUBI, A. et al. Validated model for apple flowering prediction in the mediterranean area in response to temperature variation. Scientia horticultrae, v. 249, p. 59–64, 2019.

ELAVARASAN, D. et al. Forecasting yield by integrating agrarian factors and machine learning models: a survey. Computers and Electronics in Agriculture, v. 155, p. 257–282, 2018.

FEURER, M. et al. Efficient and robust automated machine learning. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 2015. Anais... Curran Associates: Inc., 2015. v. 28.

FIORAVANÇO, J.; SANTOS, R. dos. Maçã: o produtor pergunta, a embrapa responde. Área de Informação da Sede-Col Criar Plantar ABC 500P/500R Saber (INFOTECA-E), 2013.

GARG, A.; MAGO, V. Role of machine learning in medical research: a survey. Computer Science Review, v. 40, p. 100370, 2021.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. Deep learning. ]: MIT Press, 2016. <http://www.deeplearningbook.org>.

HAIDER, S. A. et al. Lstm neural network based forecasting model for wheat production in pakistan. Agronomy, v. 9, n. 2, p. 72, 2019.

HÄNI, N.; ROY, P.; ISLER, V. A comparative study of fruit detection and counting methods for yield mapping in apple orchards. arXiv preprint arXiv:1810.09499, 2018.

- HAYKIN, S. Neural networks and learning machines. 3a. ed. ]: Prentice Hall, 2008.
- HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. Neural computation, v. 9, n. 8, p. 1735–1780, 1997.
- HYNDMAN, R. J.; ATHANASOPOULOS, G. Forecasting: principles and practice. 3rd. ed. ]: OTexts: Melbourne, Australia, 2021.
- JAROSZ, Q. Neuron. The figure is attributed to Quasar Jarosz at English Wikipedia, CC BY-SA 3.0.
- JIN, H.; SONG, Q.; HU, X. Auto-keras: an efficient neural architecture search system. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY & DATA MINING, 25., 2019. Proceedings... 2019. p. 1946–1956.
- KAMILARIS, A.; PRENAFETA-BOLDÚ, F. X. Deep learning in agriculture: a survey. Computers and electronics in agriculture, v. 147, p. 70–90, 2018.
- KHANDELWAL, P.; KONAR, J.; BRAHMA, B. Training rnn and it's variants using sliding window technique. In: IEEE INTERNATIONAL STUDENTS' CONFERENCE ON ELECTRICAL, ELECTRONICS AND COMPUTER SCIENCE (SCEECS), 2020., 2020. Anais... 2020. p. 1–5.
- KIST, B. et al. Anuário brasileiro da maçã 2018. Santa Cruz do Sul: Editora Gazeta Santa Cruz, 2018.
- KITCHENHAM, B. Procedures for performing systematic reviews. Keele, UK, Keele University, v. 33, n. 2004, p. 1–26, 2004.
- KLOMPENBURG, T. van; KASSAHUN, A.; CATAL, C. Crop yield prediction using machine learning: a systematic literature review. Computers and Electronics in Agriculture, v. 177, p. 105709, 2020.
- LEE, M. A. et al. A framework for predicting soft-fruit yields and phenology using embedded, networked microsensors, coupled weather models and machine-learning techniques. Computers and Electronics in Agriculture, v. 168, p. 105103, 2020.
- LIAKOS, K. G. et al. Machine learning in agriculture: a review. Sensors, v. 18, n. 8, p. 2674, 2018.
- LIU, Z.; WANG, S. Detecting changes of wheat vegetative growth and their response to climate change over the north china plain. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, v. 11, n. 12, p. 4630–4636, 2018.
- MARRA, F. et al. Thermal time requirement and harvest time forecast for peach cultivars with different fruit development periods. In: V INTERNATIONAL PEACH SYMPOSIUM 592, 2001. Anais... 2001. p. 523–529.
- MCCORMICK, R. F. et al. Intercontinental prediction of soybean phenology via hybrid ensemble of knowledge-based and data-driven models. bioRxiv, 2020.
- MENDES, D. S. et al. Phenology and thermal requirements of the atemoya tree (*annona cherimola* mill. x *annona squamosa* l.). Revista Ceres, v. 66, n. 3, p. 200–209, 2019.

- MILLER, P.; LANIER, W.; BRANDT, S. Using growing degree days to predict plant stages. Ag/Extension Communications Coordinator, Communications Services, Montana State University-Bozeman, Bozeman, MO, v. 59717, n. 406, p. 994–2721, 2001.
- MONTEIRO, J. E. et al. Agrometeorologia dos cultivos: o fator meteorológico na produção agrícola. ]: MAPA/INMET, 2009. 450–464 p. v. 1.
- MORRIS, J. et al. Essential irrigation and the economics of strawberries in a temperate climate. Agricultural Water Management, v. 194, p. 90–99, 2017.
- OLAH, C. Understanding lstm networks. Accessed 2021-03-12.
- PETRI, J. L. Formação de flores, polinização e fertilização. EPAGRI. A cultura da macieira., p. 229–260, 2006.
- PETRI, J. L. et al. Apple phenology in subtropical climate conditions. Embrapa Agroindústria Tropical-Capítulo em livro científico (ALICE), 2012.
- PUTTI, G. L.; PETRI, J. L. Estádios fenológicos da macieira nas cultivares gala, fuji e golden delicious. Revista Agropecuária Catarinense, v. 15, p. 22–26, 2002.
- PUTTI, G.; MENDEZ, M. E. G.; PETRI, J. L. Unidades de frio e de calor para a brotação de macieira (*malus domestica*, borck), gala e fuji. Rev. Bras. de Agrociência, Pelotas, RS - Brasil, v. 6, p. 194–196, 2000.
- QIAN, J. et al. A smartphone-based apple yield estimation application using imaging features and the ann method in mature period. Scientia Agricola, v. 75, n. 4, p. 273–280, 2018.
- REA, R.; ECCEL, E. Phenological models for blooming of apple in a mountainous region. International Journal of Biometeorology, v. 51, n. 1, p. 1–16, 2006.
- REFAEILZADEH, P.; TANG, L.; LIU, H. Cross-validation. Accessed 2021-03-12.
- RIVERO, R. et al. Flowering phenology and the interrelations between phenological stages in apple trees (*malus domestica* borkh.) as influenced by the nordic climate. Acta Agriculturae Scandinavica, Section B—Soil & Plant Science, v. 67, n. 4, p. 292–302, 2017.
- RUSSELL, S.; NORVIG, P. Artificial intelligence: a modern approach. 4th ed.. ed. ]: Pearson Education Limited, 2021.
- SABZI, S. et al. An automatic non-destructive method for the classification of the ripeness stage of red delicious apples in orchards using aerial video. Agronomy, v. 9, n. 2, p. 84, 2019.
- SAFA, B. et al. Artificial neural networks application to predict wheat yield using climatic data. In: INTERNATIONAL CONFERENCE ON IIPS, 20., 2004. Proceedings... 2004. p. 1–39.
- SCHULTZ, M. et al. Can deep learning beat numerical weather prediction? Philosophical Transactions of the Royal Society A, v. 379, n. 2194, p. 20200097, 2021.
- SCHWARTZ, M. D. et al. Phenology: an integrative environmental science. ]: Springer, 2013.

SEBRAE. SERVIÇO BRASILEIRO DE APOIO ÀS MICRO E PEQUENAS EMPRESAS. O cultivo e o mercado da maçã. 2018. Disponível em: <<http://www.sebrae.com.br/sites/PortalSebrae/artigos/ocultivoemercadodamaca>> Acesso em: Nov 2018.

SHANMUGAMANI, R. Deep learning for computer vision: expert techniques to train advanced neural networks using tensorflow and keras. ]: Packt Publishing Ltd, 2018.

SOUZA, A. P. d.; LEONEL, S.; SILVA, A. C. d. Basal temperature and thermal sum in phenological phases of nectarine and peach cultivars. Pesquisa Agropecuária Brasileira, v. 46, n. 12, p. 1588–1596, 2011.

TRIPATHI, M. K.; MAKTEDAR, D. D. A role of computer vision in fruits and vegetables among various horticulture products of agriculture fields: a survey. Information Processing in Agriculture, v. 7, n. 2, p. 183–203, 2020.

VILLA NOVA, N. A. et al. Estimativa de graus-dia acumulados acima de qualquer temperatura base, em função das temperaturas máxima e mínima. ]: Instituto de geografia, 1972.

WANG, H. et al. Using machine learning to develop an autoverification system in a clinical biochemistry laboratory. Clinical Chemistry and Laboratory Medicine (CCLM), v. 1, n. ahead-of-print, 2020.

WANG, Y. et al. Phenological prediction algorithm based on deep learning. In: IEEE INTERNATIONAL CONFERENCE ON MECHATRONICS AND AUTOMATION (ICMA), 2019., 2019. Anais... 2019. p. 589–593.

XU, Y. et al. Machine learning in construction: from shallow to deep learning. Developments in the Built Environment, p. 100045, 2021.

YALCIN, H. Plant phenology recognition using deep learning: deep-pheno. In: INTERNATIONAL CONFERENCE ON AGRO-GEOINFORMATICS, 2017., 2017. Anais... 2017. p. 1–5.

YAZDANPANA, H.; OHADI, D.; SOLEIMANI, T. M. Forecasting different phenological phases of apple using artificial neural network. Indian Journal of Horticulture, v. 67, n. 4, p. 567–573, 2010.



## APÊNDICE A – INFORMAÇÕES COMPLEMENTARES

### A.1 Tabela da duração dos eventos fenológicos da macieira cultivar Gala nas safras 2000 a 2018. Caçador, SC.

Tabela 9 – Número de dias dos eventos fenológicos da macieira cultivar Gala nas safras 2000 a 2018. EPAGRI Caçador, SC.

Ano	Início da brotação	Floração plena	Floração final	Maturação início	Maturação plena
2000	9	19	26	127	141
2001	7	14	19	125	138
2002	3	12	19	127	140
2003	5	12	23	121	130
2004	7	26	32	147	159
2005	5	21	41	114	135
2006	3	14	47	118	127
2007	12	23	28	143	161
2008	7	29	39	140	151
2009	4	21	27	110	146
2010	5	19	25	141	158
2011	11	19	24	129	140
2012	5	18	21	142	151
2013	12	20	26	127	138
2014	6	22	30	132	148
2015	4	24	37	134	149
2016	12	24	30	133	143
2017	12	19	23	152	150
2018	11	18	23	150	146
Média	7,4	1,7	28,4	132,2	144,8
Desvio Padrão	3	5	8	12	9

**A.2 Tabela de valores e médias anuais do período de 2000 a 2018. Caçador, SC.**

Tabela 10 – Valores e médias anuais da temperatura mínima (Tmin), temperatura máxima (Tmax), temperatura média (Tmed), amplitude (Ampl), média temperatura mínima (mmed), média temperatura máxima (mmax), média de graus-dia (mGD), graus-dia acumulado (GDA) e duração de amostras do período de 2000 a 2018. Temperatura base de 2°. Fonte: EPAGRI Caçador, SC.

Ano	Tmin	Tmax	Tmed	Ampl	mmin	mmax	mGD	GDA_t2°	Duração
2000	8,4	32,0	19,8	23,6	15,0	26,2	17,8	1938	108
2001	5,4	32,2	19,7	26,8	14,3	26,1	17,7	1986	111
2002	6,0	31,8	20,3	25,8	15,8	26,1	18,3	2122	115
2003	6,0	33,2	19,3	27,2	14,6	25,2	17,3	1905	109
2004	3,6	31,5	19,1	27,9	14,6	25	17,1	2085	121
2005	6,2	33,0	19,7	26,8	14,7	26,3	17,7	1920	107
2006	4,0	32,2	20,3	28,2	16,0	26,3	18,3	1925	104
2007	4,2	31,4	19,1	27,2	14,2	25,4	17,1	2065	120
2008	4,4	31,4	18,7	27,0	14,3	24,7	16,7	1872	111
2009	6,6	33,4	20,1	26,8	15,5	26,4	18,1	2233	121
2010	2,2	30,2	18,6	28,0	13,7	24,8	16,6	2038	122
2011	4,5	30,8	18,3	26,3	13,0	24,6	16,3	1812	110
2012	6,5	33,4	20,1	26,9	16,4	26,2	18,1	2264	124
2013	6,6	33,4	19,7	26,8	14,9	26,0	17,7	1906	107
2014	8,0	31,8	19,9	23,8	15,4	25,8	17,9	1989	110
2015	10,0	32,0	20,0	22,0	15,9	25,2	18,0	2003	110
2016	2,0	31,5	19,1	29,5	14,6	31,5	17,1	1886	109
2017	3,0	32,1	18,8	29,1	14,0	25,0	16,8	2256	133
2018	5,9	34,0	20,1	28,1	15,4	26,5	18,1	2409	132
Média	5,4	32,1	19,5	26,7	14,8	25,6	17,5	2032	114
Desvio Padrão	2,1	1,0	0,6	1,9	0,9	0,6	0,6	160,4	8,55

### A.3 Tabela de valores e médias anuais do período de 2016 a 2020. Vacaria, RS.

Tabela 11 – Valores e médias anuais da temperatura mínima (Tmin), temperatura máxima (Tmax), temperatura média (Tmed), amplitude (Ampl), média temperatura mínima (mmed), média temperatura máxima (mmax), média de graus-dia (mGD), graus-dia acumulado (GDA) e duração do período de amostras 2016 a 2020. Temperatura base de 2°. Fonte dados meteorológicos: Departamento de Diagnóstico e Pesquisa Agropecuária (DDPA) da Secretaria Estadual da Agricultura, Vacaria, RS.

Ano	Tmin	Tmax	Tmed	Ampl	mmin	mmax	mGD	GDA_t2°	Duração
2016	0,2	33,0	16,6	32,8	11,2	25,1	14,6	2067	128
2017	0,4	32,5	16,4	32,1	11,6	24,5	14,4	1989	124
2018	2,4	32,6	17,5	30,2	11,3	24,4	15,5	2140	135
2019	-1,5	34,5	16,5	36	12,2	25,7	14,5	2240	132
2020	-1,8	33,2	15,7	35	11,6	25,3	13,7	2187	133
Média	0,5	33,2	16,6	33,1	11,6	25,0	14,6	2125	130
Desvio Padrão	1,5	0,8	0,5	2,2	0,4	0,5	0,5	98,9	4,4

#### A.4 Tabela de graus-dia acumulado com diferentes valores de temperatura base. EPA-GRI Caçador, SC.

Tabela 12 – Tabela de Graus-dia acumulado com diferentes valores de temperatura base. Safras 2000 a 2018. EPAGRI Caçador, SC.

Ano	Temp. base 2°	Temp. base 4°	Temp. base 10°
2000	1938	1720	1066
2001	1986	1762	1090
2002	2122	1890	1194
2003	1905	1685	1025
2004	2085	1841	1109
2005	1920	1706	1064
2006	1925	1715	1085
2007	2065	1823	1097
2008	1872	1648	976
2009	2233	1991	1265
2010	2038	1792	1054
2011	1812	1590	924
2012	2264	2014	1264
2013	1906	1690	1042
2014	1989	1767	1101
2015	2003	1781	1115
2016	1886	1666	1006
2017	2256	1988	1184
2018	2409	2143	1345
Média	2032	1800	1105
Desvio Padrão	160,4	145,5	105,1
Coefficiente de variação	7,89	8,08	9,51

**A.5 Modelo linear ajustado para estimar a duração do período da plena floração ao início da maturação em função de graus-dia acumulados. Dados: EPAGRI Caçador, SC.**

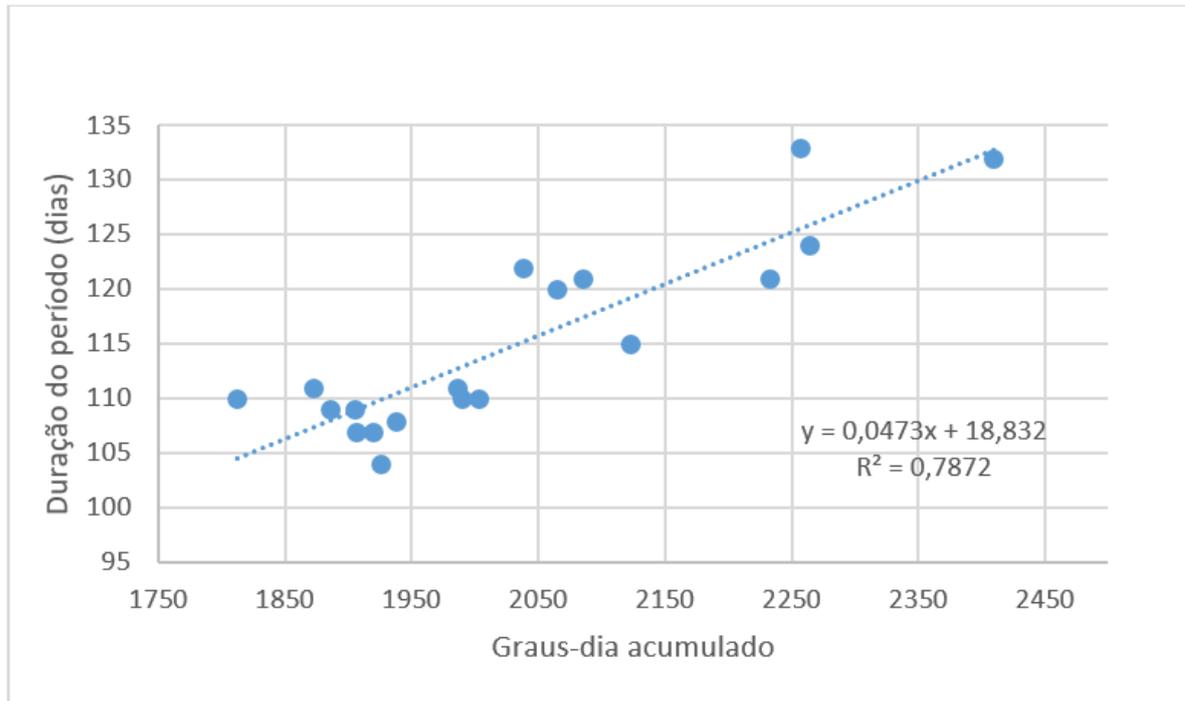


Figura 28 – Duração do período em função de graus-dia acumulados