

UNIVERSIDADE ESTADUAL DE PONTA GROSSA
PRÓ-REITORIA DE PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO COMPUTAÇÃO APLICADA

JORGE DAVI NAVARRO

ANÁLISE PREDITIVA DE DADOS GEOESPACIAIS PARA MITIGAÇÃO DE
DESASTRES CLIMÁTICOS EM PROPRIEDADES RURAIS

PONTA GROSSA
2026

JORGE DAVI NAVARRO

ANÁLISE PREDITIVA DE DADOS GEOESPACIAIS PARA MITIGAÇÃO DE
DESASTRES CLIMÁTICOS EM PROPRIEDADES RURAIS

Dissertação apresentada ao Programa de Pós-Graduação em Computação Aplicada da Universidade Estadual de Ponta Grossa, como requisito parcial para a obtenção do título de Mestre em Computação Aplicada.

Área de Concentração: Computação para Tecnologias em Agricultura.

Orientação: Prof. Dr^a. Maria Salete Marcon Gomes Vaz.

PONTA GROSSA
2026

N321

Navarro, Jorge Davi

Análise preditiva de dados geoespaciais para mitigação de desastres climáticos em propriedades rurais / Jorge Davi Navarro. Ponta Grossa, 2026. 57 f.

Dissertação (Mestrado em Computação Aplicada - Área de Concentração: Computação para Tecnologias em Agricultura), Universidade Estadual de Ponta Grossa.

Orientadora: Profa. Dra. Maria Salete Marcon Gomes Vaz.

1. Aprendizado de máquina. 2. Risco agrometeorológico. 3. Fenologia. 4. Sistema de alerta precoce. 5. Agricultura de precisão. I. Vaz, Maria Salete Marcon Gomes. II. Universidade Estadual de Ponta Grossa. Computação para Tecnologias em Agricultura. III.T.

CDD: 005.3



UNIVERSIDADE ESTADUAL DE PONTA GROSSA
Av. General Carlos Cavalcanti, 4748 - Bairro Uvaranas - CEP 84030-900 - Ponta Grossa - PR - <https://uepg.br>

TERMO

DE APROVAÇÃO

Jorge Davi Navarro

“ANÁLISE PREDITIVA DE DADOS GEOESPACIAIS PARA MITIGAÇÃO DE DESASTRES CLIMÁTICOS EM PROPRIEDADES RURAIS”

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre no Programa de Pós-Graduação em Computação Aplicada da Universidade Estadual de Ponta Grossa, pela seguinte banca examinadora:

Profa. Dra. Maria Salete Marcon Gomes Vaz (PPGCA/UEPG – Presidente)

Prof. Dr. Jonathan de Matos (PPGCA /UEPG - Membro Interno)

Prof. Dr. José Simão de Paula Pinto (PPGCD/UFPR - Membro Externo)

Ponta Grossa, 27 de julho de 2026.



Documento assinado eletronicamente por **Jonathan de Matos, Professor(a)**, em 27/07/2026, às 09:47, conforme Resolução UEPG CA 114/2018 e art. 1º, III, "b", da Lei 11.419/2006.



Documento assinado eletronicamente por **Maria Salete Marcon Gomes Vaz, Professor(a)**, em 27/07/2026, às 09:47, conforme Resolução UEPG CA 114/2018 e art. 1º, III, "b", da Lei 11.419/2006.



Documento assinado eletronicamente por **José Simão de Paula Pinto, Usuário Externo**, em 27/07/2026, às 09:49, conforme Resolução UEPG CA 114/2018 e art. 1º, III, "b", da Lei 11.419/2006.



A autenticidade do documento pode ser conferida no site <https://sei.uepg.br/autenticidade> informando o código verificador **3222265** e o código CRC **8461A143**.

"Não nos cabe dominar todas as marés do mundo, mas sim fazer o que está ao nosso alcance para o bem da época em que fomos colocados, erradicando o mal nos campos que conhecemos, para que aqueles que viverem depois tenham terra limpa para cultivar. Que tempo terão, não é nosso papel decidir."

— J.R.R. Tolkien, através de Gandalf (O Retorno do Rei)

Dedico essa dissertação ao meu filho.

A convicção de que eu precisava terminar este ciclo nunca foi tão forte quanto no dia em que fomos surpreendidos com a notícia de que você já crescia há 23 semanas no ventre da sua mãe. Descobrir você nessa etapa me fez correr contra o tempo para encerrar este mestrado e poder me dedicar inteiramente a você. Que as conquistas vindas deste título sejam o alicerce para lhe oferecer a melhor realidade possível.

AGRADECIMENTOS

A Deus, acima de tudo, pela dádiva da vida. Agradeço por me conceder a força necessária e a oportunidade de estar vivo para vivenciar a conclusão de mais esta etapa.

Aos meus pais, pilares fundamentais da minha existência. Agradeço por todo o apoio incondicional, pelo amor imensurável e por estarem sempre ao meu lado, iluminando os meus passos independentemente das adversidades.

À minha noiva, pela paciência, carinho e compreensão. Obrigado por ser o meu porto seguro e por me apoiar diariamente durante os momentos mais intensos e desafiadores da escrita deste mestrado.

À minha orientadora, Prof. Dra. Maria Salete, por ter sido muito mais do que uma guia acadêmica. Agradeço por me ouvir, me entender nos momentos de fragilidade e, principalmente, por acreditar em mim e não me deixar desistir quando o caminho se mostrou árduo.

A mim mesmo, pela resiliência em continuar lutando. Este trabalho foi forjado em meio a provações intensas que exigiram uma reinvenção completa: a superação de dois acidentes, a mudança de cidade para Tibagi e a construção de uma vida com novos hábitos. Enfrentei dias sombrios, como os quinze dias em que fiquei sem andar, mas a recuperação e a força de vontade prevaleceram. Celebro hoje não apenas a entrega desta dissertação, mas a vitória pessoal de ter transformado o meu corpo e a minha saúde, eliminando quase 40 kg, provando a mim mesmo que sou capaz de me reerguer e de nunca desistir.

Aos professores do programa, por toda a dedicação, excelência e conhecimento compartilhado nas disciplinas, que formaram a base sólida desta pesquisa.

Aos colegas de turma, pela parceria, troca de ideias e companheirismo durante as aulas, tornando a jornada do mestrado mais leve e colaborativa.

Aos ilustres membros da banca examinadora, expresse minha sincera gratidão por aceitarem o convite para avaliar este trabalho. Agradeço pela disponibilidade, pela leitura atenta e criteriosa, pelas valiosas sugestões e pelas contribuições apresentadas durante a defesa, que enriqueceram esta pesquisa e contribuíram significativamente para seu aprimoramento.

Por fim, à instituição, por proporcionar um ambiente propício ao desenvolvimento científico, oferecendo a estrutura e o suporte necessários para a realização desta pesquisa.

RESUMO

A agricultura, no estado do Paraná, destaca-se pela alta adoção tecnológica e produtividade, porém apresenta vulnerabilidade estrutural crônica frente à variabilidade e aos extremos climáticos. Nesse contexto, esta dissertação teve como objetivo geral desenvolver e validar um Sistema de Alerta Precoce (SAP) baseado em algoritmos de *Machine Learning* para a predição espacial e antecipação de quebras de safra nas principais culturas de sequeiro da região (soja, milho, trigo e feijão). A metodologia fundamentou-se na integração espaço-temporal de séries meteorológicas diárias do Instituto Nacional de Meteorologia (INMET) com o histórico anual de rendimento agrícola do Instituto Brasileiro de Geografia e Estatística (IBGE), abrangendo o período de 2011 a 2024. Para mitigar o ruído estatístico de análises anuais contínuas, propôs-se uma engenharia de atributos estritamente acoplada aos calendários fenológicos locais. A validação algorítmica descartou métodos aleatórios, adotando a divisão cronológica estrita (*Out-of-Time Validation*), utilizando as safras de 2011 a 2020 para o treinamento e o período climaticamente anômalo de 2021 a 2024 para o teste cego. O desbalanceamento de classes, inerente aos desastres naturais, foi mitigado por meio da técnica de sobre amostragem sintética *SMOTE*, conjugada à Calibração de Limiar (*Threshold Optimization*). Os resultados evidenciaram a ocorrência do "Paradoxo da Acurácia", no qual arquiteturas complexas (como as Redes Neurais *MLP*) falharam na detecção de anomalias quando avaliadas pelo erro global. Ao reposicionar a métrica de Sensibilidade (*Recall*) como critério de sucesso operacional, o Modelo *Support Vector Machine* (*SVM*) com *kernel RBF* destacou-se na identificação precoce de quebras na cultura da soja, enquanto o Algoritmo *Random Forest* apresentou a melhor relação de *trade-off* para o trigo. A análise de importância de variáveis validou agronomicamente o modelo ao isolar a inovadora métrica "Chuva na Colheita" como o preditor primordial de desastre para as culturas de inverno e soja. Adicionalmente, documentou-se o "Viés Tecnológico" na cultura do milho, limitador temporal decorrente do avanço genético não linear das sementes. Conclui-se que o *framework* computacional preditivo desenvolvido superou as limitações da climatologia linear padrão e dos modelos de circulação global, consolidando-se como uma ferramenta customizada e robusta para o suporte técnico ao dimensionamento de seguros agrícolas, formulação de políticas de crédito e mitigação de perdas no agronegócio paranaense.

Palavras-Chave: Aprendizado de Máquina. Risco Agrometeorológico. Fenologia. Sistema de Alerta Precoce. Agricultura de Precisão.

ABSTRACT

Agriculture in the state of Paraná is characterized by high technological adoption and productivity, yet it remains structurally vulnerable to climate variability and extreme weather events. In this context, this master's thesis aimed to develop and validate an Early Warning System (EWS) based on Machine Learning algorithms for the spatial prediction and anticipation of crop failures in the region's main rainfed crops (soybeans, maize, wheat, and dry beans). The methodology was grounded in the spatio-temporal integration of daily meteorological series from the National Institute of Meteorology (INMET) with annual agricultural yield history from the Brazilian Institute of Geography and Statistics (IBGE), spanning the 2011–2024 period. To mitigate statistical noise from continuous annual analysis, a feature engineering approach strictly coupled with local phenological calendars was proposed. Algorithmic validation avoided random splitting, instead adopting a strict Out-of-Time Validation approach, utilizing the 2011–2020 harvests for training and the climatically anomalous 2021–2024 period for blind testing. Class imbalance, inherent to natural disasters, was mitigated using the Synthetic Minority Over-sampling Technique (SMOTE) combined with Threshold Optimization. The results highlighted the "Accuracy Paradox," where complex architectures (such as *MLP* Neural Networks) failed to detect anomalies when evaluated by global error. By repositioning Sensitivity (Recall) as the operational success criterion, the Support Vector Machine (*SVM*) model with an RBF kernel excelled in the early identification of soybean crop failures, while the *Random Forest* algorithm showed the best trade-off for wheat. Variable importance analysis provided agronomic validation by isolating the innovative "Rain during Harvest" metric as the primary disaster predictor for winter crops and soybeans. Additionally, a "Technological Bias" was documented in maize cultivation, representing a temporal limitation due to the non-linear genetic advancement of seeds. It is concluded that the developed predictive computational framework overcame the limitations of standard linear climatology and global circulation models, establishing itself as a customized and robust tool for technical support in agricultural insurance sizing, credit policy formulation, and loss mitigation in Paraná's agribusiness.

Keyword: Machine Learning. Agrometeorological Risk. Phenology. Early Warning System. Precision Agriculture.

LISTA DE FIGURAS

Figura 1 - Diagrama de blocos do algoritmo <i>RF</i> (<i>dataset bootstrap</i> , árvores de decisão e votação majoritária)	23
Figura 2 - Arquitetura da Célula LSTM (Fluxo de x_t , $h_{(t-1)}$, Portas Sigmoid/tanh e C_t)..	24
Figura 3 - Representação gráfica da Curva ROC mostrando 5 classificadores discretos	27
Figura 4 - Diagrama de blocos da arquitetura do framework preditivo proposto	31
Figura 5 - Estrutura cronológica de particionamento de dados	35
Figura 6 - Comparativo de Sensibilidade (Recall) entre os algoritmos testados para a identificação de quebras de safra na cultura da Soja	43
Figura 7 - Matriz de Confusão gerada pelo classificador <i>Support Vector Machine</i> (<i>SVM</i>) com kernel RBF aplicado ao conjunto de teste da Soja	43
Figura 8 - Curva <i>Receiver Operating Characteristic</i> (ROC) e métrica de Área Sob a Curva (AUC) para a detecção de quebra de safra no Trigo	45
Figura 9 - Dispersão e linha de tendência do rendimento produtivo médio (kg/ha) da cultura do Milho de 1ª Safra ao longo do período analisado (2011-2024).....	47
Figura 10 - Impacto metodológico da aplicação conjunta da injeção de dados sintéticos (<i>SMOTE</i>) e da calibração probabilística (Otimização de Limiar) sobre a métrica de Recall ..	48
Figura 11 - Análise da Curva Precision-Recall (PR) para a definição do limiar operacional do Sistema de Alerta Precoce no Trigo.....	49

LISTA DE QUADROS

Quadro 1 - Resumo das arquiteturas e premissas operacionais dos algoritmos de <i>Machine Learning</i> avaliados no benchmark.....	36
Quadro 2 - Síntese do desempenho preditivo, algoritmos de melhor adequação e principais desafios operacionais identificados por cultura.....	49

LISTA DE TABELAS

Tabela 1 - Matriz de confusão para avaliação de modelos de risco de eventos extremos	26
Tabela 2 - Avaliação comparativa das métricas de classificação (Acurácia, Recall, Precisão e F1-Score) obtidas no conjunto de teste (Out-of-Time) para a cultura da Soja. Destaca-se a ocorrência do Paradoxo da Acurácia, evidenciado pela taxa nula de sensibilidade.....	42
Tabela 3 - Avaliação comparativa das métricas de classificação preditiva obtidas no conjunto de teste para a cultura do Trigo	44
Tabela 4 - Desempenho preditivo consolidado dos algoritmos de ML no conjunto de teste para as culturas de Milho e Feijão.....	45

LISTA DE SIGLAS

ANFIS	Adaptive Neuro-Fuzzy Inference System (Sistemas Neuro-Fuzzy)
ARIMA	Auto-Regressive Integrated Moving Average (Modelos Auto-regressivos Integrados de Médias Móveis)
ARIMAX	Auto-Regressive Integrated Moving Average with Exogenous Variables
AUC	Area Under the Curve (Área Sob a Curva)
BRT	Boosted Regression Trees
EMA	Estação Meteorológica Automática
FN	Falso Negativo
FP	Falso Positivo
GBM	Gradient Boosting Machine
GCMs	Global Circulation Models (Modelos de Circulação Global)
GIS	Geographic Information Systems (Sistemas de Informação Geográfica)
IBGE	Instituto Brasileiro de Geografia e Estatística
IMF	International Monetary Fund (Fundo Monetário Internacional)
INMET	Instituto Nacional de Meteorologia
IoT	Internet of Things (Internet das Coisas)
KNN	K-Nearest Neighbors (K-Vizinhos Mais Próximos)
LSTM	Long Short-Term Memory (Memória de Longo e Curto Prazo)
MDE	Modelo Digital de Elevação
ML	Machine Learning (Aprendizado de Máquina)
MLP	Multilayer Perceptron (Rede Neural de Múltiplas Camadas)
MODIS	Moderate Resolution Imaging Spectroradiometer
NDVI	Normalized Difference Vegetation Index (Índice de Vegetação por Diferença Normalizada)
PAM	Produção Agrícola Municipal
PH	Peso Hectolítrico
PIB	Produto Interno Bruto
PPGECM	Programa de Pós-Graduação em Ensino de Ciências e Educação Matemática
RBF	Radial Basis Function (Função de Base Radial)
RMSE	Root Mean Square Error (Erro Quadrático Médio)
RNN	Recurrent Neural Network (Rede Neural Recorrente)
ROC	Receiver Operating Characteristic
SAP	Sistema de Alerta Precoce
SMOTE	Synthetic Minority Over-sampling Technique
SPI	Standardized Precipitation Index (Índice de Precipitação Padronizada)
<i>SVM</i>	Support Vector Machine (Máquina de Vetores de Suporte)
TRMM	Tropical Rainfall Measuring Mission
UEPG	Universidade Estadual de Ponta Grossa
VANTs	Veículos Aéreos Não Tripulados
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

SUMÁRIO

INTRODUÇÃO	14
CAPÍTULO 1 REVISÃO DE LITERATURA	18
1.1 MUDANÇAS CLIMÁTICAS, FENOLOGIA E RISCO AGRÍCOLA NOS CAMPOS GERAIS	18
1.2 GEOTECNOLOGIAS E SENSORIAMENTO REMOTO APLICADOS À AGRICULTURA	20
1.3 APRENDIZADO DE MÁQUINA (<i>MACHINE LEARNING</i>) NA PREDIÇÃO CLIMÁTICA E AGRÍCOLA	21
1.3.1 Métodos de Ensemble: <i>Random Forest</i> e Gradient Boosting	22
1.3.2 Redes Neurais para Séries Temporais: <i>LSTM (Long Short-Term Memory)</i>	23
1.3.3 O Trade-off entre Previsão Sequencial e Interpretabilidade	25
1.4 O PROBLEMA DO DESBALANCEAMENTO DE DADOS E AVALIAÇÃO DE RISCO	25
1.4.1 O Paradoxo da Acurácia e a Matriz de Confusão	25
1.4.2 Métricas Discriminatórias para Eventos Extremos	26
1.4.3 Estratégias de Mitigação do Desbalanceamento: <i>SMOTE</i> e Otimização de Limiar	27
1.5 TRABALHOS CORRELATOS	28
1.5.1 Lacunas de Pesquisa e Posicionamento da Dissertação	29
CAPÍTULO 2 MATERIAIS E MÉTODOS	31
2.1 ÁREA DE ESTUDO E BASE DE DADOS	31
2.2 INTEGRAÇÃO ESPAÇO-TEMPORAL E ENGENHARIA DE ATRIBUTOS	32
2.3 DEFINIÇÃO DA VARIÁVEL-ALVO E VALIDAÇÃO TEMPORAL	34
2.4 AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA (BENCHMARK)	36
2.5 TRATAMENTO DE DESBALANCEAMENTO E OTIMIZAÇÃO DE LIMIAR	38
CAPÍTULO 3 RESULTADOS E DISCUSSÕES	40
3.1 VALIDAÇÃO AGRONÔMICA E IMPORTÂNCIA DE ATRIBUTOS (<i>FEATURE IMPORTANCE</i>)	40
3.2 AVALIAÇÃO COMPARATIVA DOS ALGORITMOS (<i>BENCHMARK PREDITIVO</i>)	41
3.2.1 O Paradoxo da Acurácia e o Triunfo do <i>SVM</i> na Cultura da Soja	41
3.2.2 Estabilidade do <i>Random Forest</i> na Avaliação do Trigo	43

3.2.3	Desafios Operacionais e Viés Tecnológico: O Milho e o Feijão	45
3.3	IMPACTO DO BALANCEAMENTO DE DADOS E OTIMIZAÇÃO PARA ALERTA PRECOCE	47
3.4	COMPARAÇÃO COM TRABALHOS CORRELATOS E VALIDAÇÃO DAS HIPÓTESES.....	50
3.5	LIMITAÇÕES DO ESTUDO	51
	CONSIDERAÇÕES FINAIS.....	53
	REFERÊNCIAS.....	55

INTRODUÇÃO

MOTIVAÇÃO

As mudanças climáticas globais configuram, na contemporaneidade, um dos desafios complexos e urgentes para a estabilidade dos sistemas naturais, econômicos e sociais, em função do aquecimento acelerado observado nas últimas décadas (Mcginnis et al., 2025). O aumento da temperatura média global e a intensificação das emissões de gases de efeito estufa têm provocado alterações na dinâmica climática do planeta, expressas na maior frequência e intensidade de eventos extremos, como secas prolongadas, enchentes, ondas de calor e tempestades severas (Mcginnis et al., 2025; Muller et al., 2019). Esses eventos repercutem de forma profunda em diversos setores, afetando sistemas energéticos, infraestrutura crítica e cadeias produtivas, com destaque para a produção agrícola e a segurança alimentar em regiões vulneráveis (Muller et al., 2019; Imf, 2024). Estudos recentes evidenciam que extremos climáticos associados a padrões como *El Niño* pressionam a produção de alimentos em regiões-chave, ampliando riscos de insegurança alimentar e exigindo políticas robustas de adaptação e gestão de riscos (Zhang et al., 2025; Imf, 2024).

Entre os setores produtivos, a agricultura destaca-se por sua elevada sensibilidade às variáveis climáticas. Historicamente, a produção agrícola depende da interação entre fatores edáficos, hidrológicos e atmosféricos, sendo particularmente vulnerável às alterações nos padrões de temperatura e precipitação. Nesse contexto, eventos climáticos extremos podem comprometer diretamente o desenvolvimento das culturas, reduzindo a produtividade e ampliando os riscos econômicos associados à atividade agrícola (Hameed; Ahmadalipour; Moradkhani, 2020). Fenômenos como secas recorrentes e irregularidades pluviométricas afetam os meios de subsistência de produtores rurais, especialmente em regiões dependentes da agricultura de sequeiro, agravando problemas relacionados à insegurança alimentar e à vulnerabilidade socioeconômica (Ahmad; Yaseen; Saqib, 2022).

No contexto brasileiro, o setor agropecuário possui papel estratégico para o desenvolvimento econômico nacional, contribuindo significativamente para o Produto Interno Bruto (PIB), geração de empregos e balança comercial. Entretanto, essa relevância econômica coexiste com uma crescente exposição aos riscos climáticos. A variabilidade das chuvas, associada ao aumento da frequência de eventos extremos, tem gerado impactos diretos sobre a produtividade agrícola, exigindo novas abordagens de planejamento e gestão de riscos.

A Região dos Campos Gerais, localizada no Estado do Paraná, representa um exemplo emblemático dessa realidade. Reconhecida como uma das regiões mais tecnificadas e

produtivas do agronegócio brasileiro, apresenta forte especialização na produção de *commodities* agrícolas, como soja, milho, trigo e feijão. Apesar da elevada adoção de tecnologias agrícolas, a produtividade regional permanece fortemente condicionada às condições climáticas, especialmente à regularidade das precipitações e às variações térmicas ao longo do ciclo produtivo.

Do ponto de vista climatológico, o Estado do Paraná apresenta elevada variabilidade pluviométrica e uma dinâmica atmosférica complexa, influenciada pela interação entre sistemas tropicais e extratropicais. Estudos indicam que o regime de chuvas na região vem passando por mudanças nas últimas décadas, incluindo tendências de aumento na intensidade das precipitações e alterações na distribuição temporal das chuvas, fenômeno frequentemente descrito como “tropicalização pluviométrica” do clima subtropical (Nascimento Jr.; Silvestre; Sant’anna Neto, 2020). Essas alterações podem resultar tanto em eventos de precipitação extrema quanto em períodos de estiagem intraestacional, conhecidos regionalmente como veranicos.

A ocorrência de veranicos representa um risco para culturas anuais, especialmente quando coincide com fases fenológicas críticas, como floração e enchimento de grãos. De maneira oposta, o excesso hídrico durante a fase de maturação e colheita também configura um severo desastre climático, inviabilizando o tráfego de maquinário, reduzindo a qualidade do grão e favorecendo a proliferação de doenças. Estudos realizados no Estado do Paraná demonstram que a variabilidade das chuvas e a ocorrência de períodos de seca podem impactar diretamente a produtividade agrícola, evidenciando a necessidade de monitoramento climático e planejamento agrícola baseado em dados meteorológicos e hidrológicos (Ferreira et al., 2020).

Os impactos desses eventos ultrapassam os limites agronômicos e atingem diretamente a sustentabilidade financeira do produtor rural. Perdas de safra, danos estruturais e interrupções produtivas ampliam os riscos de crédito agrícola, comprometendo a capacidade de investimento e de continuidade das atividades produtivas. Dessa forma, a gestão eficiente de riscos climáticos torna-se um elemento importante para garantir a resiliência dos sistemas agrícolas e a estabilidade das cadeias produtivas.

Diante desse cenário, cresce a necessidade de desenvolvimento de ferramentas tecnológicas capazes de antecipar e monitorar riscos climáticos em escalas espaciais cada vez mais detalhadas. Modelos climáticos tradicionais operam em escalas regionais ou nacionais, o que limita sua aplicabilidade para a tomada de decisão no nível da propriedade rural. Nesse contexto, a integração entre geotecnologias, sensoriamento remoto e análise de grandes

volumes de dados espaciais surge como uma alternativa promissora para aprimorar o monitoramento ambiental e a previsão de eventos climáticos extremos.

O avanço das técnicas de aprendizado de máquina (*Machine Learning*) tem ampliado as possibilidades de análise de dados ambientais complexos. Modelos baseados em inteligência artificial permitem identificar padrões não lineares entre variáveis climáticas, hidrológicas e agronômicas, possibilitando o desenvolvimento de sistemas preditivos capazes de antecipar cenários de risco climático com maior precisão (Yang et al., 2020). Esses modelos podem integrar diferentes fontes de informação, como dados meteorológicos, índices de vegetação derivados de sensoriamento remoto e características do relevo e do solo, permitindo análises espaciais detalhadas.

Nesse contexto, a presente pesquisa tem como objetivo desenvolver e validar um framework computacional preditivo baseado na integração de dados geoespaciais e séries temporais meteorológicas, utilizando técnicas de aprendizado de máquina. Com foco na Região dos Campos Gerais, no Estado do Paraná, o estudo busca identificar áreas com maior probabilidade de ocorrência de eventos climáticos extremos atuando como um Sistema de Alerta Precoce (SAP) para prever potenciais impactos na produção agrícola. Espera-se que os resultados contribuam para o desenvolvimento de ferramentas de apoio à tomada de decisão no setor agropecuário, fortalecendo estratégias de adaptação climática e promovendo maior resiliência dos sistemas agrícolas frente às mudanças climáticas.

OBJETIVOS

Objetivo Geral

Desenvolver e validar um framework computacional preditivo, fundamentado na integração de dados geoespaciais e séries temporais, para identificar áreas com elevado risco de ocorrência de eventos climáticos extremos em propriedades agrícolas na Região dos Campos Gerais, visando dar suporte à tomada de decisão e à gestão de riscos no contexto da agricultura de precisão.

Objetivos Específicos

- I. Coletar e integrar uma base de dados multivariada a partir de fontes abertas, englobando variáveis topográficas, pedológicas, de uso e ocupação do solo, além de séries históricas meteorológicas.
- II. Realizar o pré-processamento e a engenharia de atributos (*feature engineering*) para harmonizar os dados espacialmente e selecionar as variáveis com maior poder explicativo para a ocorrência de desastres climáticos locais.

- III. Implementar e treinar diferentes algoritmos de *Machine Learning*, como *Random Forest*, *Gradient Boosting* e Redes Neurais *LSTM*, para a tarefa de predição de risco probabilístico, utilizando dados tabulares e séries temporais.
- IV. Avaliar e validar o desempenho e a robustez dos modelos preditivos por meio de métricas estatísticas rigorosas, como Acurácia, *F1-Score*, Curva ROC e AUC, estabelecendo uma análise comparativa entre as abordagens testadas.

ESTRUTURA

Tendo em vista o exposto, esta dissertação foi estruturada com este capítulo introdutório e mais quatro capítulos. No Capítulo 2 é apresentada a revisão da literatura, abordando os fundamentos teóricos relacionados às mudanças climáticas, ao risco agrícola, às geotecnologias e às técnicas de aprendizado de máquina aplicadas à predição de eventos extremos. No Capítulo 3 são descritos os materiais e métodos, incluindo a caracterização da área de estudo, as bases de dados utilizadas, os procedimentos de integração espaço-temporal, a engenharia de atributos e a modelagem computacional adotada. No Capítulo 4 são apresentados e discutidos os resultados obtidos, com ênfase na análise do desempenho dos algoritmos, na validação agronômica dos atributos e na aplicabilidade do framework como Sistema de Alerta Precoce. Por fim, no Capítulo 5 são expostas as considerações finais, destacando as principais contribuições da pesquisa, suas limitações e sugestões para trabalhos futuros.

CAPÍTULO 1 REVISÃO DE LITERATURA

Este capítulo tem como objetivo apresentar a fundamentação teórica e o estado da arte que sustentam o desenvolvimento desta pesquisa. Para tanto, a revisão foi estruturada de forma a encadear logicamente os conceitos centrais que permeiam a problemática dos desastres meteorológicos e as soluções tecnológicas propostas para a sua mitigação. Inicialmente, discute-se a interrelação entre as mudanças climáticas, a fenologia das culturas de sequeiro e a escalada do risco agrícola. Em seguida, aborda-se o papel das geotecnologias e do sensoriamento remoto como ferramentas basilares para a espacialização e o monitoramento desses riscos em alta resolução. Na sequência, explora-se a aplicação de algoritmos de *ML* na modelagem preditiva, culminando na análise dos desafios metodológicos inerentes ao desbalanceamento de dados em eventos extremos e às métricas adequadas para a gestão de risco. Por fim, apresenta-se um panorama dos trabalhos correlatos, evidenciando as lacunas científicas que justificam a pertinência e a originalidade desta dissertação.

1.1 MUDANÇAS CLIMÁTICAS, FENOLOGIA E RISCO AGRÍCOLA NOS CAMPOS GERAIS

A variabilidade climática contemporânea tem afetado expressivamente a produtividade agrícola na região Sul do Brasil, caracterizada por ecossistemas de transição e reconhecida como um polo estratégico para o cultivo em larga escala de *commodities* como soja, milho e trigo (Machado et al., 2022). Elementos como ondas de calor, mudanças nos padrões de precipitação e maior frequência de eventos climáticos extremos têm repercutido sobre o calendário de cultivo, a fisiologia das plantas e a logística de colheita, especialmente em áreas de produção intensiva (Santos, 2023). Trabalhos recentes sobre a agricultura familiar no Rio Grande do Sul reforçam a vulnerabilidade dessa região, evidenciando a associação entre perdas produtivas, períodos de estiagem prolongada e anomalias de temperatura observadas nos últimos anos (Santos, 2023; Ibge, 2022).

Na Região dos Campos Gerais do Paraná, especificamente, evidencia-se um cenário de alto risco: a elevada eficiência produtiva e o intenso aporte tecnológico coexistem com uma vulnerabilidade estrutural a eventos climáticos extremos. A frequência e a magnitude desses eventos têm sido alteradas tanto pelas mudanças climáticas globais quanto pela complexa dinâmica atmosférica regional, marcada pela interação de frentes frias, ciclogêneses e massas de ar polar e tropical (Borsato; Massoquim, 2018; Leite; Adacheski; Virgens Filho, 2010).

A literatura recente aponta para um fenômeno de "tropicalização" das chuvas no Estado do Paraná. Esse processo caracteriza-se não necessariamente pela alteração do volume total anual de chuvas, mas por uma mudança em sua distribuição: observa-se uma maior concentração pluviométrica em eventos de alta intensidade e curta duração, intercalados por prolongados períodos de estiagem (Nascimento Jr.; Silvestre; Sant'anna Neto, 2020). Essa irregularidade hídrica afeta de maneira direta a segurança e a rentabilidade do setor agropecuário regional, tornando imperativo o monitoramento constante do risco imposto pelas condições edafoclimáticas e a adoção de estratégias de mitigação baseadas em dados.

Nesse contexto, reveste-se de suma importância que a modelagem preditiva e a análise de risco climático sejam concebidas em estrita conformidade com as fases fenológicas das culturas. O impacto de uma anomalia meteorológica difere drasticamente dependendo do estágio de desenvolvimento da planta.

No caso da cultura da soja, o déficit hídrico coincidente com os estágios reprodutivos críticos notadamente o florescimento (R2) e o desenvolvimento e enchimento de grãos (R5) acarreta distúrbios fisiológicos como o abortamento de vagens e a perda de massa seca, resultando em danos irreversíveis e quebras severas de rendimento (Araujo, 2008; Ferreira et al., 2020). A ocorrência de "veranicos" (estiagens intra-safra) materializa esse risco, o que fundamenta a urgência de ferramentas preditivas que considerem não apenas a ocorrência da seca, mas o momento exato em que ela atinge a lavoura.

Adicionalmente aos períodos de escassez, os excessos meteorológicos representam uma ameaça crítica nas fases finais do ciclo produtivo. A precipitação pluviométrica contínua durante as etapas de maturação e senescência ocasiona severos prejuízos biológicos e operacionais, especialmente para a soja e o trigo. O excesso de umidade, quando associado a altas temperaturas, provoca a deterioração fisiológica e a proliferação de patógenos nas sementes, comprometendo a qualidade comercial e a taxa de germinação dos grãos (Henning; França Neto, 1980). Em termos operacionais, a chuva paralisa as atividades a campo devido à inviabilidade de tráfego de maquinário pesado no solo encharcado, atrasando a colheita e expondo a cultura a perdas adicionais (Ferreira et al., 2020).

Para além da precipitação, a agricultura dos Campos Gerais enfrenta riscos acentuados decorrentes da intensificação de tempestades severas. Fenômenos convectivos extremos na primavera, frequentemente acompanhados por granizo e vendavais e, em casos recentes, tornados, como o registrado em Piraí do Sul em 2023, causam não apenas a destruição mecânica das plantas, mas também danos estruturais maciços às propriedades rurais (Piotrovski et al., 2023). Essa multiplicidade de ameaças edafoclimáticas reforça a necessidade do

desenvolvimento de sistemas de alerta precoce de alta resolução, capazes de mapear a vulnerabilidade fenológica e geográfica em nível de talhão.

1.2 GEOTECNOLOGIAS E SENSORIAMENTO REMOTO APLICADOS À AGRICULTURA

Para compreender a adoção destas geotecnologias, é imperativo definir o conceito de Agricultura de Precisão (AP). A AP consolida-se como uma estratégia de gestão agrícola que emprega tecnologias de informação e comunicação, englobando sistemas de posicionamento global (GPS), sensores, imagens de satélite e algoritmos preditivos, para coletar, processar e analisar volumes massivos de dados espaço-temporais multivariados (Balti *et al.*, 2020). O seu objetivo central é otimizar a alocação de recursos, maximizar a produtividade e mitigar riscos ambientais e climáticos, tratando a heterogeneidade intrínseca das lavouras em escala de micro relevo e talhão, em oposição ao manejo uniforme e generalista tradicionalmente adotado nas propriedades rurais.

A transição de uma agricultura puramente reativa para um framework de gestão proativa de riscos climáticos é indissociável do avanço das geotecnologias. A análise de vulnerabilidade agrícola demanda a integração de um volume massivo de dados de múltiplas naturezas atmosféricos, topográficos, pedológicos e biológicos, o que torna a junção entre o Sensoriamento Remoto e os Sistemas de Informação Geográfica (GIS) a base metodológica fundamental para a construção de modelos preditivos modernos (Shahi *et al.*, 2022).

O Sensoriamento Remoto, operacionalizado por meio de plataformas orbitais, como as constelações de satélites *Landsat* e *Sentinel*, e Veículos Aéreos Não Tripulados (VANTs), revolucionou a capacidade de monitoramento contínuo das áreas agrícolas. A captação da reflectância eletromagnética da superfície terrestre permite a derivação de índices espectrais cruciais para a análise de risco, com destaque para o Índice de Vegetação por Diferença Normalizada (NDVI). O NDVI atua como um biomarcador quantitativo do vigor vegetativo, permitindo identificar precocemente o estresse hídrico ou térmico nas lavouras antes mesmo que os danos sejam visualmente perceptíveis a campo. Adicionalmente, o uso de bandas infravermelhas termais possibilita o mapeamento de anomalias da temperatura da superfície e a estimativa da umidade do solo, variáveis indispensáveis para a parametrização do impacto de secas e geadas.

De forma complementar, os GIS funcionam como o ecossistema computacional necessário para integrar, harmonizar e analisar esses dados espaciais. Em um ambiente GIS, é possível realizar a sobreposição (*overlay*) de diferentes camadas temáticas de informação (Gao

et al., 2023). Por exemplo, a extração de dados de um Modelo Digital de Elevação (MDE) permite o cálculo de declividade e orientação de vertentes, que, quando cruzados com o mapeamento de solos (textura e capacidade de armazenamento de água) e com as séries históricas de precipitação, revelam as zonas de maior suscetibilidade a desastres hidrometeorológicos (Wilopo et al., 2023). Da mesma forma, a espacialização de dados de estações meteorológicas automáticas, quando georreferenciadas e cruzadas com a malha de produção municipal, permite extrair a realidade climática exata a que a lavoura foi exposta. Essa capacidade de realizar engenharia de atributos espaciais é o que alimenta com precisão os algoritmos de inteligência artificial.

Apesar da ampla disponibilidade tecnológica, a eficácia prática da modelagem de risco na agricultura esbarra frequentemente na questão da resolução espacial. Historicamente, os frameworks agrometeorológicos e climáticos têm operado em macroescalas, fornecendo diagnósticos para grandes bacias hidrográficas, municípios inteiros ou regiões de planejamento. Embora úteis para a formulação de macropolíticas públicas, essas escalas generalistas falham em capturar a heterogeneidade intrínseca do ambiente rural. O impacto de uma tempestade convectiva severa ou a intensidade de uma geada de radiação variam drasticamente em distâncias curtas, sendo fortemente influenciados pelo microrrelevo e pelas barreiras vegetais locais.

Por essa razão, para que a modelagem preditiva se traduza em uma ferramenta efetiva de suporte à decisão, é imperativo que a análise desça à escala da propriedade rural, chegando à resolução do talhão agrícola. Frameworks construídos em alta resolução espacial permitem que o produtor rural e os agentes de extensão planejem intervenções como o dimensionamento da irrigação localizada, a alocação de quebra-ventos ou a contratação de prêmios de seguro agrícola ajustados ao risco real daquela coordenada geográfica específica. Superar a generalização espacial e trazer a predição probabilística para a realidade topográfica do agricultor constitui, portanto, uma das fronteiras mais urgentes e inovadoras no campo da agricultura de precisão.

1.3 APRENDIZADO DE MÁQUINA (*MACHINE LEARNING*) NA PREDIÇÃO CLIMÁTICA E AGRÍCOLA

A modelagem preditiva de riscos agrícolas e eventos climáticos extremos constitui um desafio metodológico, dada a natureza intrinsecamente estocástica, caótica e não linear dos sistemas atmosféricos. Abordagens estatísticas tradicionais de séries temporais a exemplo dos modelos autorregressivos integrados de médias móveis (ARIMA), embora fundamentais para

a climatologia clássica, frequentemente apresentam limitações analíticas ao tentar capturar as intrincadas interações multidimensionais entre variáveis geoespaciais, topográficas e hidrometeorológicas (Tuğrul; Hınıs, 2025).

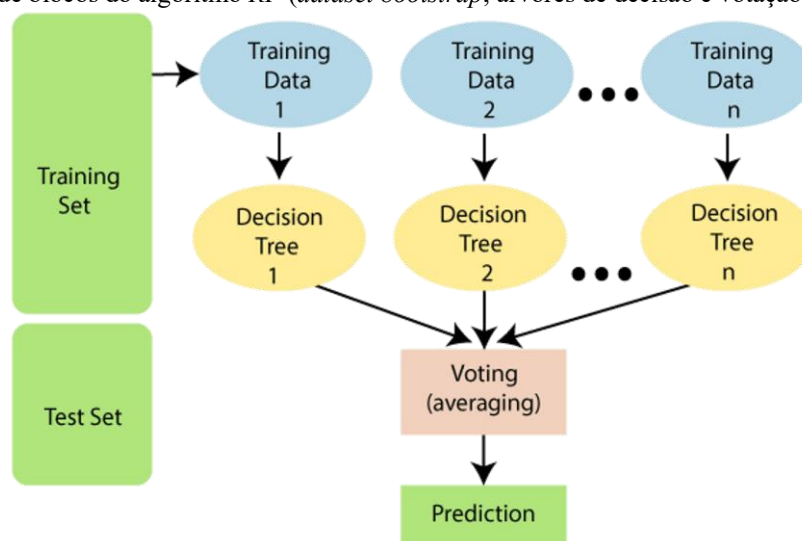
Nesse cenário, o Aprendizado de Máquina (*Machine Learning* - ML) consolida-se como um paradigma analítico indispensável. Os algoritmos de ML oferecem a capacidade computacional de aprender padrões complexos e extrair correlações ocultas a partir de volumes massivos de dados heterogêneos (*Big Data*), superando as restrições de dimensionalidade e as premissas rígidas de distribuição de dados exigidas pela modelagem estocástica convencional (Prottasha et al., 2023). Para a estruturação de Sistemas de Alerta Precoce (SAP) na agricultura, destacam-se duas abordagens principais que balanceiam a acurácia preditiva e a capacidade de interpretação: os métodos de *Ensemble* baseados em árvores e as Redes Neurais Profundas para dados sequenciais.

1.3.1 Métodos de Ensemble: *Random Forest* e Gradient Boosting

Em contraste com a rigidez paramétrica dos modelos lineares, os algoritmos baseados em árvores de decisão demonstram elevada eficácia preditiva na manipulação de bases de dados tabulares que mesclam variáveis contínuas, como precipitação acumulada, e categóricas, como classes de solo. Contudo, árvores de decisão individuais são altamente suscetíveis ao sobreajuste (*overfitting*). Para mitigar esse problema, a literatura emprega métodos de *Ensemble*, que combinam os resultados de múltiplos modelos "fracos" para criar um preditor global robusto (Hussain et al., 2024).

O algoritmo *Random Forest* (Floresta Aleatória) baseia-se na técnica de *Bagging* (*Bootstrap Aggregating*). Durante o treinamento, o algoritmo extrai múltiplas amostras aleatórias do conjunto de dados original (com substituição) e constrói uma árvore de decisão independente para cada amostra. Além disso, a cada divisão de nó (*split*), apenas um subconjunto aleatório de variáveis preditoras é considerado. O resultado final da predição é obtido por meio de uma "votação majoritária" (para classificação) ou pela média das predições (para regressão) de todas as árvores geradas, como demonstrada na figura 1. Essa arquitetura paralela reduz drasticamente a variância do modelo, tornando-o excepcionalmente imune a ruídos e *outliers*, que são inerentes aos registros de desastres naturais (Patel et al., 2023).

Figura 1 - Diagrama de blocos do algoritmo *RF* (*dataset bootstrap*, árvores de decisão e votação majoritária).



Fonte: MLU-EXPLAIN (2023)

Por outro lado, o *Gradient Boosting Machine (GBM)* opera sob a lógica do *Boosting*. Diferente do paralelismo do *RF*, o *GBM* constrói árvores de decisão de forma sequencial. Cada nova árvore é otimizada matematicamente para prever e corrigir os erros (os resíduos) cometidos pela árvore imediatamente anterior. Utilizando o método do gradiente descendente para minimizar uma função de perda pré-definida, o *GBM* frequentemente alcança métricas de desempenho, como Acurácia e *Recall*, superiores a outros métodos de ML estruturados, sendo amplamente validado em desafios preditivos complexos no setor agroclimático (Yang et al., 2025).

A principal vantagem analítica de ambos os métodos de *Ensemble* no contexto de risco agrícola reside na sua capacidade inerente de fornecer a "Importância das Variáveis" (*Feature Importance*). O algoritmo consegue quantificar numericamente a contribuição de cada variável (por exemplo, a declividade do terreno versus o déficit de chuva) para a redução da impureza dos nós da árvore. Isso garante interpretabilidade ao modelo, permitindo ao pesquisador identificar os gatilhos exatos das anomalias meteorológicas.

1.3.2 Redes Neurais para Séries Temporais: *LSTM (Long Short-Term Memory)*

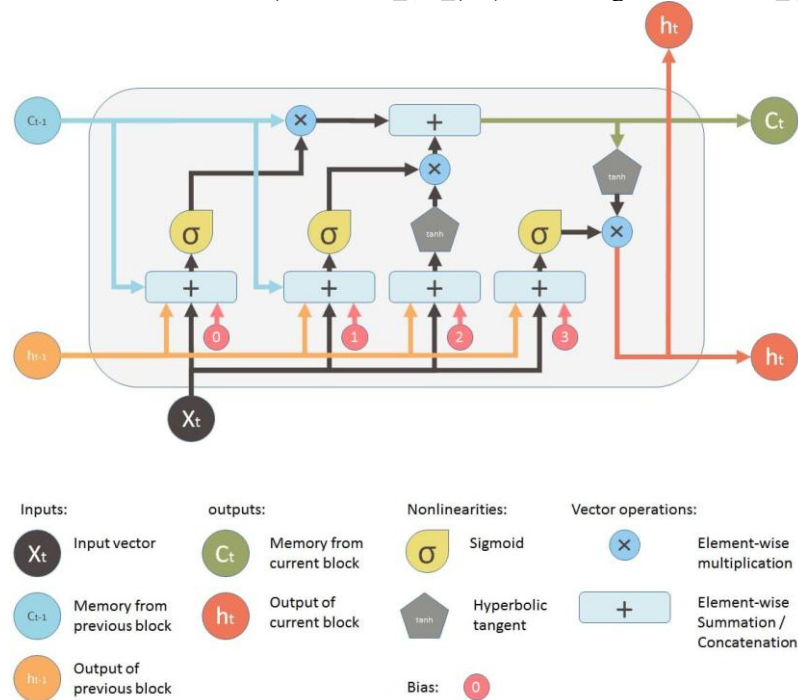
No espectro do Aprendizado Profundo (*Deep Learning*), as Redes Neurais Artificiais buscam mimetizar a estrutura sináptica humana. Para a previsão climática baseada em séries históricas, as Redes Neurais Recorrentes (RNN) introduziram o conceito de "memória", permitindo que a saída de uma camada fosse retroalimentada como entrada. No entanto, RNNs clássicas sofrem com o problema do "desvanecimento do gradiente" (*vanishing gradient*), o que

as impede de aprender dependências de longo prazo. Uma falha crítica ao tentar prever, por exemplo, o impacto acumulado de meses de déficit hídrico sobre a fenologia de uma cultura.

Para contornar essa limitação matemática, a arquitetura de LSTM, foi desenvolvida. As redes LSTM substituem os nós ocultos tradicionais por "células de memória" complexas, exemplificada na figura 2. Cada célula contém um estado interno e é regulada por três estruturas lógicas denominadas "portões" (*gates*):

- Portão de Esquecimento (*Forget Gate*): Decide qual fração da informação do estado passado deve ser descartada ou retida.
- Portão de Entrada (*Input Gate*): Determina quais novas informações da série temporal atual devem ser adicionadas ao estado da célula.
- Portão de Saída (*Output Gate*): Controla a parcela do estado da célula que será propagada para o próximo passo de tempo.

Figura 2 - Arquitetura da Célula LSTM (Fluxo de x_t , $h_{(t-1)}$, Portas Sigmoid/tanh e C_t).



Fonte: adaptado de GUO et al (2021)

Através desse fluxo de controle de informações, a LSTM consegue reter características sazonais de longo prazo, sendo o estado da arte na literatura para a previsão da evolução temporal de secas severas e na modelagem hidrometeorológica contínua (Yang et al., 2025; Pachpor et al., 2024).

1.3.3 O Trade-off entre Previsão Sequencial e Interpretabilidade

Apesar da comprovada superioridade preditiva da rede LSTM no tratamento de séries temporais puras, o seu emprego isolado na avaliação de risco agrícola em nível de propriedade esbarra na opacidade estrutural do modelo. As redes neurais profundas operam fundamentalmente como "caixas pretas" (*black boxes*). A complexidade dos pesos distribuídos em suas múltiplas camadas dificulta a extração de uma explicação transparente sobre a correlação causal exata que levou o algoritmo a deflagrar um alerta de desastre (Samek; Wiegand; Müller, 2017).

Para o desenvolvimento de um Sistema de Alerta Precoce efetivamente aplicável à realidade do produtor dos Campos Gerais, a interpretabilidade das variáveis geoclimáticas é tão crucial quanto o acerto cronológico do evento. Portanto, a comparação metodológica entre a capacidade puramente sequencial das redes LSTM e a interpretabilidade baseada em *Feature Importance* dos modelos de *Ensemble* (*RF* e *GBM*) faz-se cientificamente necessária, permitindo selecionar o limiar ideal entre precisão, robustez analítica e utilidade prática para a mitigação de danos estruturais no campo.

1.4 O PROBLEMA DO DESBALANCEAMENTO DE DADOS E AVALIAÇÃO DE RISCO

A previsão de desastres naturais e anomalias climáticas por meio de ML trata intrinsecamente com fenômenos de imensa complexidade (Balti et al., 2020). Um problema crônico e estrutural na modelagem destes eventos é o severo desbalanceamento de classes nos conjuntos de dados (*datasets*). Em um cenário real de monitorização agrometeorológica, a condição de "normalidade" constitui a classe majoritária, ocorrendo na esmagadora maioria do tempo. Em contrapartida, a classe de interesse, o desastre ou a anomalia climática, como uma geada severa ou um tornado, é um evento estatisticamente raro, formando a classe minoritária.

O desafio computacional reside no fato de que os algoritmos de aprendizado padrão são matematicamente otimizados para minimizar o erro global. Consequentemente, ao serem treinados com dados altamente assimétricos, os modelos tendem a desenvolver um viés para a classe majoritária, subajustando (*underfitting*) os padrões da anomalia, comprometendo categoricamente a eficácia de qualquer Sistema de Alerta Precoce (SAP).

1.4.1 O Paradoxo da Acurácia e a Matriz de Confusão

A avaliação algorítmica apoiada exclusivamente na acurácia global resulta no que a literatura classifica como o "Paradoxo da Acurácia". Em um conjunto de dados onde 99% das amostras representam condições normais e apenas 1% representa eventos extremos, um modelo

preditivo que classifique absolutamente todas as instâncias como "normais" obterá 99% de acurácia. Apesar da aparente excelência matemática, o modelo possui valor operacional nulo, falhando em seu único propósito: a identificação do desastre.

Para superar essa distorção, a base da avaliação de modelos de risco estrutura-se sobre a Matriz de Confusão, que cruza as previsões do algoritmo com as observações reais, conforme ilustrado na Tabela 1.

Tabela 1 - Matriz de confusão para avaliação de modelos de risco de eventos extremos

	Condição real: Evento extremo (Positivo)	Condição real: Normal (Negativo)
Previsto: Evento extremo (Positivo)	Verdadeiro Positivo (VP)	Falso Positivo (FP)
Previsto: Normal (Negativo)	Falso Negativo (FN)	Verdadeiro Negativo (VN)

Fonte: do autor (2026)

No contexto de salvaguarda da infraestrutura rural, o Falso Negativo (FN) representa o cenário catastrófico, uma anomalia que atinge a lavoura sem que o produtor tenha sido avisado, resultando em perdas financeiras e biológicas irreversíveis.

1.4.2 Métricas Discriminatórias para Eventos Extremos

Com base na Matriz de Confusão, a validação de modelos de risco exige o emprego de métricas discriminatórias rigorosas, que penalizem os erros do modelo na classe minoritária. Destacam-se as seguintes formulações:

Sensibilidade (*Recall*): É a métrica fundamental em cenários de risco iminente. Mensura a capacidade do sistema em não deixar que nenhum desastre passe despercebido, sendo calculada pela razão entre os alertas corretos e o total de desastres que realmente ocorreram:

$$recall = \frac{VP}{VP + FN}$$

Precisão (*Precision*): Avalia a confiabilidade do alerta. Dos eventos que o modelo classificou como desastre, quantos realmente se concretizaram:

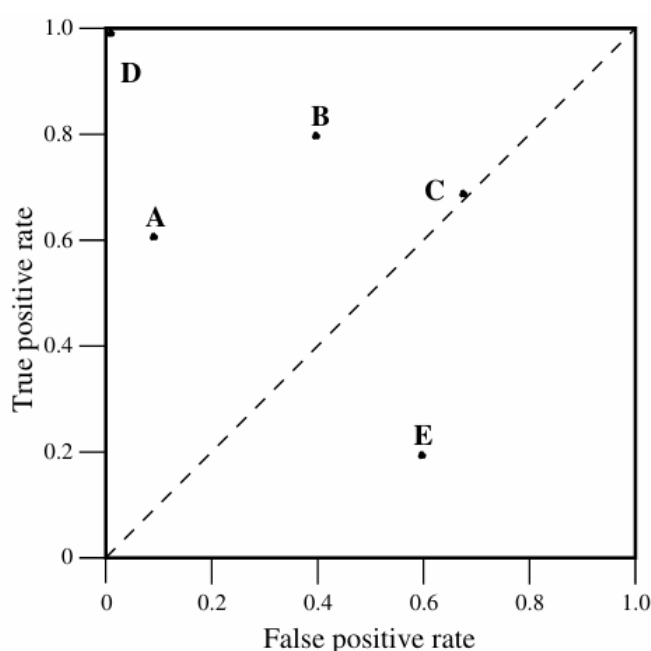
$$precision = \frac{VP}{VP + FP}$$

F1-Score: Como existe um compromisso (*trade-off*) matemático entre Precisão e *Recall*, ao aumentar um, o outro tende a diminuir, o *F1-Score* calcula a média harmônica entre ambos, fornecendo uma métrica única de robustez em dados desbalanceados:

$$F1 - Score = 2 * \frac{precision * recall}{precision + recall}$$

Além das métricas pontuais, o desempenho global de classificação é aferido pela Curva ROC (*Receiver Operating Characteristic*), exemplificada na figura 3, e pela métrica AUC (*Area Under the Curve*). A curva plota a taxa de Verdadeiros Positivos contra a taxa de Falsos Positivos em diversos limiares de decisão. Um modelo aleatório possui AUC de 0,5, enquanto um modelo perfeito atinge AUC igual a 1,0 (Fawcett, 2006).

Figura 3 - Representação gráfica da Curva ROC mostrando 5 classificadores discretos



Fonte: dos autores (2026)

1.4.3 Estratégias de Mitigação do Desbalanceamento: *SMOTE* e Otimização de Limiar

Para reverter o viés algorítmico provocado pela raridade dos eventos climáticos extremos, a literatura especializada em Ciência de Dados recomenda a intervenção direta na distribuição do conjunto de treinamento ou na calibração das probabilidades geradas pelo algoritmo. Entre as técnicas de reamostragem mais validadas cientificamente para tratar a assimetria de classes, destaca-se o Algoritmo *SMOTE* (*Synthetic Minority Over-sampling Technique*), introduzido fundamentalmente por Chawla et al. (2002).

Diferentemente da superamostragem aleatória (*Random Oversampling*), que atua apenas pela duplicação de instâncias da classe minoritária e eleva substancialmente o risco de sobreajuste (*overfitting*), o *SMOTE* opera na manipulação do espaço de características (*feature space*). O algoritmo seleciona uma instância da classe minoritária (o desastre climático) e identifica os seus *k* vizinhos mais próximos (*kNN*). Em seguida, calcula a diferença vetorial

entre a instância observada e um dos seus vizinhos, multiplicando essa diferença por um número aleatório entre 0 e 1, adicionando o resultado ao vetor original.

Este processo matemático gera dados sintéticos inéditos e espacialmente plausíveis, forçando as fronteiras de decisão de algoritmos complexos, como o *RF* e o *SVM*, a expandirem a sua área de reconhecimento topológico para a classe de risco, sem a mera memorização de dados passados. Embora a aplicação do *SMOTE* resolva a assimetria na fase de treinamento, a operacionalização de um Sistema de Alerta Precoce (SAP), no contexto agrícola, exige intervenção complementar, Otimização do Limiar de Decisão (*Threshold Optimization*).

Por predefinição, classificadores binários assumem um limiar de 0,5 (50%) para categorizar nova instância; ou seja, o alerta de quebra de safra só é emitido se o modelo for superior a 50% de certeza estatística. No entanto, na mitigação de desastres agrícolas, a matriz de custos é desequilibrada. O prejuízo decorrente da omissão de um alerta (Falso Negativo) é assimetricamente superior ao custo de preparação para alarme falso (Falso Positivo).

A otimização do limiar consiste na análise da curva *Precision-Recall* dos dados de validação para ajustar ativamente o ponto de corte da probabilidade preditiva para valores inferiores (por exemplo, 0,3 ou 30%). Ao reduzir este limiar, o modelo passa a atuar de forma proativa na salvaguarda da produção, aceitando uma redução calculada na Precisão (*Precision*) com o objetivo estrito de maximizar a Sensibilidade (*Recall*). A combinação sinérgica entre a injeção de dados sintéticos via *SMOTE* e a calibração probabilística do limiar transforma modelos genéricos de classificação em verdadeiros detetores de anomalias, assegurando a identificação de tendências de risco antes que os danos às propriedades rurais se tornem irreversíveis.

1.5 TRABALHOS CORRELATOS

Nos últimos anos, a literatura científica tem reportado avanços na aplicação de algoritmos de *ML* e geotecnologias para o monitoramento e a predição de eventos climáticos extremos. Uma parcela expressiva tem se dedicado a superar as limitações dos modelos meteorológicos tradicionais, utilizando a capacidade dos algoritmos baseados em árvores de decisão para processar grandes volumes de dados heterogêneos. Shah et al. (2017), por exemplo, propuseram uma abordagem baseada em análise de *Big Data* para classificar as fases da seca agrícola. Alimentando o algoritmo *RF* com séries históricas de precipitação, temperatura e evapotranspiração. Os autores demonstraram que a geração de múltiplas árvores, operando sob um sistema de votação majoritária, otimizou a precisão da detecção de anomalias, auxiliando na alocação primária de recursos preventivos.

Expandindo o escopo analítico para a dimensão geoespacial, Park et al. (2016) focaram na avaliação de secas em diferentes regiões climáticas da Ásia Oriental por meio de dados de sensoriamento remoto. O estudo buscou prever a severidade do déficit hídrico em uma escala temporal de pentada (cinco dias), comparando o desempenho de algoritmos de *Ensemble* (*RF*, *Boosted Regression Trees* - BRT e *Cubist*). Os resultados corroboraram que a fusão de múltiplos índices espectrais oriundos de satélites, quando processada por essas arquiteturas, resulta em frameworks de predição agroclimática de alta confiabilidade.

Em paralelo aos métodos baseados em árvores, a literatura atesta a eficácia de arquiteturas conexionistas e de margem máxima para a modelagem climática. Jalalkamali et al. (2015) avaliaram o desempenho de múltiplos modelos computacionais, incluindo Máquinas de Vetores de Suporte (*SVM*), Redes Neurais Artificiais (*Multilayer Perceptron* - *MLP*) e Sistemas *Neuro-Fuzzy* (*ANFIS*), em comparação ao modelo estocástico tradicional ARIMAX, utilizando o Índice de Precipitação Padronizada (*SPI*) como variável-alvo. O estudo evidenciou a superioridade incontestada da Inteligência Artificial em relação à modelagem linear clássica para a captura das complexas relações não lineares presentes nas séries temporais. Em uma abordagem ainda mais avançada, Shen et al. (2019) investigaram o monitoramento regional de secas na província de Henan, China, extraíndo fatores de risco complexos de plataformas como MODIS e TRMM através do aprendizado profundo (*Deep Learning*). A predição espaço-temporal gerada pela rede profunda apresentou correlação altamente significativa com os índices terrestres, validando o uso de dados orbitais multivariados no rastreamento contínuo de anomalias.

1.5.1 Lacunas de Pesquisa e Posicionamento da Dissertação

Embora o estado da arte evidencie a maturidade das técnicas de *ML* para a previsão climática, uma análise crítica revela lacunas metodológicas e aplicadas que limitam a transposição dessa tecnologia para a realidade operacional do produtor rural. Primeiramente, observa-se uma limitação de escala geográfica: a maioria dos trabalhos supracitados (Park et al., 2016; Shen et al., 2019) modela o risco em macroescalas (províncias, países ou grandes bacias hidrográficas). Faltam modelos de alta resolução que desçam à escala da propriedade rural (talhão), onde as decisões de manejo são efetivamente tomadas.

Em segundo lugar, identifica-se uma lacuna de acoplamento fenológico. Grande parte dos modelos foca na predição da variável meteorológica em si (por exemplo: prever se vai chover ou não no mês), negligenciando a engenharia de atributos baseada na fenologia restrita da cultura. O impacto de um evento extremo depende intrinsecamente do estágio de

desenvolvimento da planta. A literatura carece de modelagens que isolem, por exemplo, o impacto letal de um veranico no estágio de enchimento de grãos ou os danos catastróficos do excesso pluviométrico restrito aos dias finais de maturação e colheita, variáveis frequentemente diluídas nas médias climáticas anuais.

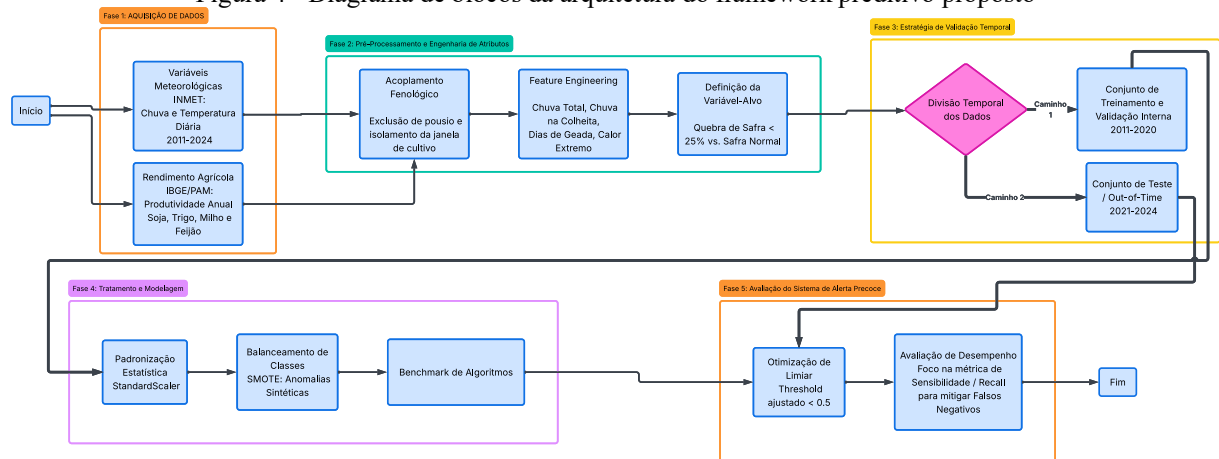
Por fim, e de forma mais crítica, a literatura carece de abordagens que tratem adequadamente com o viés dos dados em desastres naturais. Muitos estudos avaliam seus modelos com base na acurácia global, tangenciando o problema do severo desbalanceamento de classes inerente aos eventos extremos. Há uma carência de *frameworks* preditivos que apliquem técnicas avançadas de amostragem, como o *SMOTE*, integradas à calibração de limiares (*Threshold Optimization*) com foco estrito na maximização da Sensibilidade (*Recall*).

Diante deste panorama, a presente dissertação propõe-se a preencher essas lacunas, unindo o rigor analítico da modelagem não linear à operacionalização prática. Ao desenvolver um *framework* multi-risco (secas, inundações, tempestades severas e geadas) com engenharia de atributos fenológicos, correção de desbalanceamento (*SMOTE*) para a mitigação de falsos negativos e aplicação geograficamente localizada na Região dos Campos Gerais. Este trabalho visa fornecer uma ferramenta de suporte à decisão inédita, transpondo a barreira entre a pesquisa em ciência de dados e a proteção direta da infraestrutura e produtividade agrícola.

CAPÍTULO 2 MATERIAIS E MÉTODOS

Este capítulo descreve os procedimentos metodológicos delineados e executados para a concepção, o treinamento e a validação dos modelos computacionais preditivos de risco agroclimático. Para atender aos objetivos propostos, a pesquisa foi estruturada em etapas analíticas sequenciais e interdependentes. Inicialmente, apresenta-se a caracterização da área de estudo e o processo de coleta e consolidação das bases de dados meteorológicos e de rendimento agrícola. Na sequência, detalha-se a integração espaço-temporal e a engenharia de atributos, etapas importantes para a tradução dos fenômenos climáticos e calendários fenológicos em variáveis matemáticas (*features*). O capítulo prossegue documentando a definição da variável-alvo e a adoção da estratégia de validação temporal (*Out-of-Time*), visando garantir a integridade do experimento. Por fim, descreve-se o delineamento do ensaio comparativo (*benchmark*) entre algoritmos de Aprendizado de Máquina, culminando na exposição das técnicas de amostragem (*SMOTE*) e calibração de limiar empregadas para tratar o desbalanceamento estrutural dos dados de desastres extremos.

Figura 4 - Diagrama de blocos da arquitetura do framework preditivo proposto



Fonte: dos autores (2026)

A arquitetura geral da metodologia adotada, delineando a esteira de processamento de dados desde a coleta até a emissão probabilística do Sistema de Alerta Precoce (SAP), encontra-se sistematizada no fluxograma da Figura 4.

2.1 ÁREA DE ESTUDO E BASE DE DADOS

A área de estudo definida para o desenvolvimento desta pesquisa compreendeu a Região dos Campos Gerais, localizada no Estado do Paraná. Esta região geográfica e geoeconômica destaca-se pela sua representatividade no agronegócio nacional, caracterizando-se pela adoção de elevados níveis de tecnologia no campo e por um ecossistema de transição

climática. Tais fatores tornam a região um ambiente altamente produtivo, porém estruturalmente sensível às variabilidades e aos extremos meteorológicos, justificando a sua escolha como foco da modelagem preditiva.

Para a caracterização atmosférica e a estruturação da base de dados de variáveis preditivas, os dados meteorológicos históricos foram coletados a partir das Estações Meteorológicas Automáticas (EMA) pertencentes à rede de observação do Instituto Nacional de Meteorologia (INMET¹). A série temporal extraída para o estudo abrangeu o período de 2011 a 2024. As variáveis agrometeorológicas selecionadas e processadas incluíram a precipitação pluviométrica acumulada (mm) e as temperaturas do ar máxima, mínima e média (°C), registradas em escala diária. Estas variáveis foram escolhidas por sua influência direta e documentada sobre o desenvolvimento fenológico e o balanço hídrico das culturas agrícolas locais.

No que tange ao levantamento da variável-alvo atrelada ao impacto produtivo, os dados referentes ao rendimento agrícola foram extraídos da base de dados da Produção Agrícola Municipal (PAM), mantida e disponibilizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE²). Os registros foram coletados em escala municipal e expressos em quilogramas por hectare (kg/ha). O escopo da coleta contemplou as principais culturas de sequeiro que compõem a matriz econômica dos Campos Gerais: a Soja, o Milho (1ª Safra), o Trigo e o Feijão.

A integração temporal e espacial das métricas de rendimento do IBGE com as variáveis climáticas do INMET formou a base analítica consolidada para as etapas subsequentes de treinamento e validação dos algoritmos computacionais propostos neste trabalho.

2.2 INTEGRAÇÃO ESPAÇO-TEMPORAL E ENGENHARIA DE ATRIBUTOS

A etapa subsequente à coleta de dados consistiu na integração espaço-temporal das bases de séries temporais meteorológicas (INMET) com os registros tabulares de rendimento agrícola (IBGE). O principal desafio metodológico desta fase residiu na assimetria de granularidade temporal das fontes. Enquanto os dados meteorológicos apresentavam frequência de coleta diária, os dados de produtividade agrícola consistiam em um único valor consolidado anual por município. A submissão direta de séries temporais brutas de 365 dias aos algoritmos

¹ Portal do Instituto Nacional de Meteorologia (INMET) - Banco de Dados Meteorológicos. Disponível em: <https://portal.inmet.gov.br/dadoshistoricos>. Acesso em: jun. 2026.

² Sistema IBGE de Recuperação Automática (SIDRA) - Produção Agrícola Municipal (PAM). Disponível em: <https://sidra.ibge.gov.br/pesquisa/pam/tabelas>. Acesso em: jun. 2026.

de *ML* baseados em árvores introduziria uma dimensionalidade excessiva e um elevado grau de ruído estatístico, comprometendo a capacidade de generalização dos modelos.

Para contornar esse obstáculo e garantir a representatividade agronômica do conjunto de dados, o cruzamento das informações foi rigorosamente parametrizado pelos calendários fenológicos específicos de cada cultura analisada. Estabeleceram-se recortes temporais (janelas de observação) que corresponderam estritamente ao ciclo de desenvolvimento a campo na Região dos Campos Gerais. Para a cultura da soja e do milho de 1ª safra, a extração de dados limitou-se ao período de outubro a março; para o trigo, cultura de inverno, a janela estabeleceu-se de maio a setembro; e para o feijão, os recortes seguiram as datas zonais de plantio e colheita estipuladas localmente.

Os dados meteorológicos correspondentes às respectivas janelas de cultivo foram sistematicamente excluídos do banco de dados analítico. Esse procedimento de filtragem assegurou que precipitações ou anomalias térmicas ocorridas, sobre o solo em pousio, não fossem interpretadas pelo algoritmo como fatores de impacto na produtividade.

A partir do isolamento fenológico dos dados, procedeu-se à etapa de Engenharia de Atributos (*Feature Engineering*). Este processo de transformação matricial foi imperativo para converter a série temporal bruta em indicadores climáticos consolidados (*features*) que atuassem como preditores diretos da ocorrência de quebras de safra. Inicialmente, foram calculadas variáveis macroclimáticas contínuas, representativas do balanço termopluviométrico acumulado do ciclo, a "Chuva Total", correspondente ao somatório da precipitação pluviométrica na janela fenológica (expressa em mm), e a "Temperatura Média do Ciclo" (°C). Embora importantes para caracterizar o ano-safra como úmido ou seco, quente ou ameno, essas variáveis médias possuem capacidade limitada para sinalizar estresses agudos de curta duração.

Para suprir essa lacuna e capturar os "gatilhos" ambientais responsáveis por danos irreversíveis às plantas, estruturaram-se variáveis de contagem baseadas em limiares termopluviométricos críticos. Desenvolveu-se a variável "Dias de Geada", computada pela soma do número de dias dentro do ciclo em que a temperatura mínima registrada alcançou valores iguais ou inferiores a 2°C ($T_{Min} \leq 2^{\circ}C$), limiar no qual ocorre o congelamento dos tecidos intracelulares e a morte apical em culturas sensíveis. Paralelamente, elaborou-se o atributo "Dias de Calor Extremo", quantificador de ocorrências em que a temperatura máxima atingiu ou superou 33°C ($T_{Max} \geq 33^{\circ}C$). Do ponto de vista fisiológico, esse limiar térmico deflagra o fechamento estomático severo, a desnaturação proteica e o consequente abortamento de flores e vagens, atuando como forte preditor não linear de perdas de rendimento.

Por fim, visando refinar a acurácia do modelo preditivo frente a desastres frequentemente mascarados nos totais pluviométricos anuais, propôs-se a criação de uma variável inovadora denominada "Chuva na Colheita". Este atributo foi matematicamente isolado através da soma da precipitação pluviométrica ocorrida exclusivamente nos últimos 20 dias da janela fenológica de cada cultura. A formulação desta métrica teve o propósito específico de capturar o risco associado ao excesso hídrico durante a fase de maturação fisiológica e senescência (estádios finais). O impacto negativo capturado por esta variável é duplo: biológico, evidenciado pela deterioração das sementes, surgimento de patógenos fúngicos e germinação de grãos ainda na espiga ou vagem; e operacional, em virtude do encharcamento do solo que inviabiliza o tráfego de colheitadeiras pesadas, retardando a operação e expondo a cultura madura a perdas por acamamento e degradação física. A inclusão deste atributo garantiu que o modelo penalizasse excessos hídricos tardios que, em uma análise não fenológica, poderiam ser erroneamente interpretados como favoráveis à cultura.

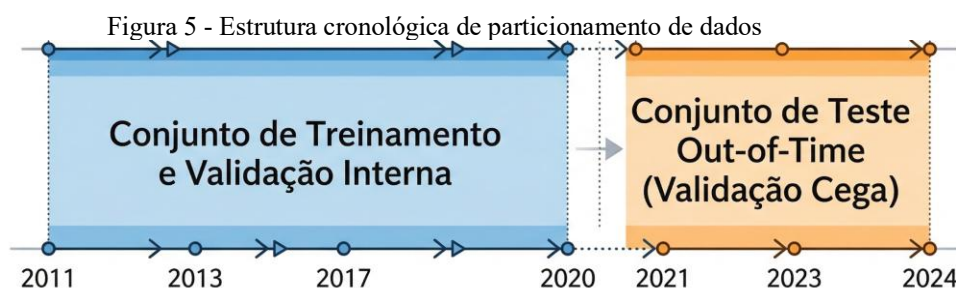
2.3 DEFINIÇÃO DA VARIÁVEL-ALVO E VALIDAÇÃO TEMPORAL

Para viabilizar a modelagem computacional de riscos por meio de algoritmos de aprendizado de máquina supervisionado, fez-se necessária a transformação da variável contínua de rendimento agrícola (kg/ha) em categorias discretas. A literatura especializada demonstra que a categorização de índices contínuos em classes de severidade é uma prática metodológica robusta e recorrente na avaliação de riscos climáticos, permitindo a estruturação de Sistemas de Alerta Precoce baseados em probabilidade (Wu et al., 2023). Sob essa ótica, optou-se pela formulação de um problema de classificação binária, convertendo o histórico de rendimento produtivo em duas classes dicotômicas de risco.

Definiu-se categoricamente como "Quebra de Safra" (Classe 1 - evento minoritário de interesse) os registros de produtividade municipal que se situaram abaixo do 1º Quartil (Percentil 25) da série histórica da respectiva cultura. Em contrapartida, os registros com rendimento igual ou superior a este limiar estatístico foram classificados como "Safra Normal" (Classe 0 - classe majoritária). A adoção do corte por quartis fundamenta-se em sua expressiva robustez analítica frente à heterogeneidade espacial. O limiar de identificação do desastre ajusta-se dinamicamente à variabilidade intrínseca e ao teto produtivo de cada município analisado. Esta abordagem isola de maneira eficaz as flutuações anuais rotineiras, forçando os modelos preditivos a concentrarem o reconhecimento de padrões unicamente nas anomalias edafoclimáticas severas que desencadearam perdas extremas.

No que tange à estratégia de particionamento do conjunto de dados (*Data Splitting*), a pesquisa descartou sumariamente o uso de métodos de divisão aleatória convencional (*Random Hold-out* ou *K-Fold Cross-Validation* simples). Na modelagem de processos hidrometeorológicos e fenológicos, inerentemente governados pela evolução cronológica, a divisão aleatória induz à quebra da autocorrelação temporal e provoca o viés de vazamento de dados (*data leakage*). Tal erro metodológico permitiria que os algoritmos tivessem acesso espúrio a informações do futuro para prever eventos do passado (por exemplo, treinar com dados de 2023 para prever a safra de 2015), o que inflaria artificialmente as métricas de desempenho e anularia a validade operacional do modelo no campo.

A fim de contornar este viés e garantir a integridade do experimento, adotou-se a estratégia estrita de Validação Temporal (*Out-of-Time Validation*). A base de dados consolidada foi segmentada de forma puramente cronológica, conforme imagem da figura 5. Os registros correspondentes às safras do período de 2011 a 2020 foram isolados e dedicados exclusivamente à fase de treinamento e otimização de hiperparâmetros dos algoritmos (*RF*, *Gradient Boosting* e redes neurais). Por sua vez, as safras subsequentes, correspondentes ao intervalo de 2021 a 2024, foram reservadas de maneira cega e irrestrita para atuar como o conjunto de teste final.



Fonte: dos autores (2026)

A escolha deste intervalo específico para o conjunto de teste (2021-2024) é metodologicamente estratégica. Este período recente foi marcado por fortes anomalias climáticas na Região Sul do Brasil, incluindo a atuação de fenômenos como o *La Niña*, que provocou secas históricas e quebras de safra sucessivas. Ao isolar um período temporalmente distinto e climaticamente hostil, o conjunto de teste assegura que os algoritmos sejam submetidos a um "teste de estresse" fidedigno ao mundo real. Esta abordagem de validação comprova de forma contundente a verdadeira capacidade de generalização e resiliência matemática dos modelos em antecipar cenários agroclimáticos adversos e inéditos.

2.4 AVALIAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA (BENCHMARK)

Com o intuito de identificar a arquitetura preditiva mais eficaz para a antecipação de quebras de safra, estruturou-se um experimento comparativo rigoroso de algoritmos de Aprendizado de Máquina (*benchmark*). Considerando que a modelagem de riscos agroclimáticos e a produtividade agrícola são regidas por interações não lineares e dinâmicas atmosféricas complexas, a avaliação comparativa de múltiplos modelos fundamentados em diferentes paradigmas matemáticos fez-se estritamente necessária para atestar a robustez e a confiabilidade das previsões (Balti et al., 2020).

Ressalta-se que, embora as Redes Neurais Recorrentes (Long Short-Term Memory - LSTM) tenham sido consideradas na revisão teórica e nos objetivos iniciais devido à sua capacidade de modelar séries temporais, a sua aplicação não foi efetivada no benchmark final. A decisão metodológica de consolidar as séries diárias em atributos tabulares fixos (*Feature Engineering*) acoplados às janelas fenológicas eliminou a estrutura sequencial estrita exigida para o treinamento de modelos LSTM. Dessa forma, para representar as arquiteturas conexionistas no experimento de classificação, optou-se exclusivamente pela Rede Neural *Multilayer Perceptron* (MLP).

Nesse contexto metodológico, o delineamento experimental englobou uma seleção abrangente de oito algoritmos de naturezas computacionais distintas, descritas no quadro 1. A avaliação contemplou métodos baseados no particionamento iterativo do espaço de características e técnicas de *ensemble*, representados pela Árvore de Decisão (*Decision Tree*), *RF* e *GBM* (Park et al., 2016). Também foram incorporadas arquiteturas conexionistas, projetadas para o aprendizado de representações não lineares profundas, como as Redes *MLP* (Jalalkamali et al., 2015). Complementando o escopo analítico, avaliaram-se abordagens fundamentadas em estatística clássica e probabilidade condicional, representadas pela *Logistic Regression* e pelo *Naive Bayes* (Priya; Ramesh; Khosla, 2018). Por fim, integraram o *benchmark* os algoritmos baseados em separação topológica de margens e cálculos espaciais de vizinhança, destacando-se as *SVM* (operando com o *kernel* - RBF) e o algoritmo *KNN* (Chintyana et. Al., 2019).

Quadro 1 - Resumo das arquiteturas e premissas operacionais dos algoritmos de *Machine Learning* avaliados no benchmark.

Algoritmo	Paradigma	Resumo Operacional
<i>Random Forest (RF)</i>	<i>Ensemble (Bagging)</i>	Combina múltiplas árvores de decisão independentes via votação majoritária para reduzir a variância temporal e evitar o sobreajuste.

<i>Gradient Boosting (GBM)</i>	<i>Ensemble (Boosting)</i>	Constrói árvores de decisão de forma sequencial, onde cada nova árvore é otimizada para corrigir os erros (resíduos) da árvore anterior.
<i>Support Vector Machine (SVM)</i>	Margem Máxima (Kernel RBF)	Busca o hiperplano ótimo de separação projetando os dados em dimensões superiores através de um <i>kernel</i> não linear, isolando geometricamente a classe de risco.
<i>Multilayer Perceptron (MLP)</i>	Redes Neurais Profundas	Rede conexionista com camadas ocultas que aprende representações complexas não lineares por meio da retropropagação do erro (<i>backpropagation</i>).
<i>K-Nearest Neighbors (KNN)</i>	Distância Espacial	Classifica as instâncias baseando-se estritamente na classe majoritária matemática dos seus 'k' vizinhos topológicos mais próximos.
<i>Logistic Regression</i>	Probabilidade Estatística	Modelo linear clássico que utiliza a função sigmoide para estimar a probabilidade condicional de pertencimento a uma classe binária de risco.
<i>Naive Bayes</i>	Probabilidade Condicional	Classificador probabilístico fundamentado no Teorema de Bayes, que opera sob a premissa de forte independência condicional entre as variáveis preditoras.
<i>Decision Tree</i>	Árvore de Decisão Simples	Modelo singular baseado em regras de divisão iterativa (<i>splits</i>) que particiona o espaço de características objetivando minimizar a impureza das classes.

Fonte: dos autores (2026)

Um aspecto analítico crítico na condução deste ensaio comparativo residuiu no tratamento prévio e na harmonização da escala geométrica das variáveis preditoras. Diferentemente dos modelos baseados em árvores, inerentemente invariantes à escala e não requerem transformação dos dados brutos, os algoritmos fundamentados em cálculos de distância métrica, como a distância Euclidiana empregada no *KNN* e no *SVM*, ou em processos de otimização por gradiente descendente, presentes na *Logistic Regression* e na *MLP*, sofrem severas distorções caso as variáveis de entrada apresentem magnitudes ou unidades de medida muito díspares. Por exemplo, precipitação em centenas de milímetros versus temperatura em dezenas de graus.

Para assegurar a convergência matemática uniforme e evitar o viés de dominância numérica de variáveis com maior amplitude geométrica, os dados submetidos a esses algoritmos específicos passaram por um rigoroso processo de padronização estatística (*z-score normalization*), implementado através do Método *StandardScaler*. Essa técnica transformou a distribuição original dos dados, garantindo-lhes média zero e variância unitária. Em estrito

cumprimento às diretrizes metodológicas de validação e visando mitigar qualquer risco de vazamento de dados (*data leakage*), os parâmetros estatísticos da padronização (média e desvio padrão) foram calculados (*fitted*) exclusivamente com os dados pertencentes ao conjunto de treinamento. Subsequentemente, essas mesmas métricas isoladas foram aplicadas na transformação (*transform*) do conjunto de teste, garantindo a integridade estrutural e a validade cega da avaliação dos modelos frente a cenários agrometeorológicos inéditos.

2.5 TRATAMENTO DE DESBALANCEAMENTO E OTIMIZAÇÃO DE LIMIAR

A modelagem computacional de riscos agroclimáticos e a predição de quebras de safra enfrentam um desafio estatístico intrínseco à natureza dos fenômenos extremos, o severo desbalanceamento de classes. Na base de dados consolidada para este estudo, constatou-se uma assimetria estrutural na variável-alvo, em que a classe minoritária de interesse, representativa dos desastres e quebras agudas de produtividade, compôs o quartil inferior (25% dos registros), enquanto a classe majoritária, indicativa de normalidade produtiva, englobou os 75% restantes. A submissão direta de distribuições tão desproporcionais aos algoritmos de *ML* induz a um forte viés matemático de predição. Sob tais condições, os modelos tendem a convergir para a minimização do erro global, priorizando o acerto da classe majoritária e subajustando (*underfitting*) os padrões complexos deflagradores de eventos extremos, inviabilizando a sua aplicação como Sistema de Alerta Precoce efetivo.

Para mitigar tal viés algorítmico, adotou-se a técnica de sobreamostragem sintética *SMOTE* (*Synthetic Minority Over-sampling Technique*). Diferentemente de abordagens de *oversampling* tradicionais, que realizam a mera replicação estocástica de instâncias minoritárias, elevando o risco de *overfitting*. O *SMOTE* soluciona o desbalanceamento por meio da interpolação geométrica de dados. O algoritmo gera instâncias sintéticas inéditas ao longo dos vetores que conectam os vizinhos mais próximos no espaço de características (*feature space*), expandindo de forma robusta as fronteiras de decisão da classe de risco. Ressalta-se que, em estrito cumprimento às diretrizes metodológicas de validação, a aplicação do *SMOTE* restringiu-se aos dados do conjunto de treinamento. Esta conduta garante que os algoritmos aprendam a mapear os desastres de forma otimizada, mas sejam avaliados em um conjunto de teste inalterado, preservando as proporções probabilísticas originais do clima e refletindo fidedignamente o cenário operacional a campo.

Em paralelo ao balanceamento do conjunto de treinamento, a calibração probabilística dos modelos demandou a aplicação da técnica de Otimização de Limiar de Decisão (*Threshold Optimization*). Em problemas de classificação binária tradicionais, o algoritmo adota um limiar

de corte padrão de 0,5 (50% de probabilidade condicional) para definir a qual classe uma observação pertence. Entretanto, no contexto da gestão de risco agrícola, a matriz de custos é caracterizada por uma profunda assimetria, o custo socioeconômico de um falso negativo (uma lavoura dizimada sem aviso prévio) é ordens de magnitude superior ao inconveniente operacional de um falso positivo (um alerta preventivo para um evento que não se concretizou com severidade).

Através da análise empírica da curva *Precision-Recall*, o limiar de decisão padrão de 0,5 foi intencionalmente rebaixado e fixado de forma estrita no valor de 0,35 (35%). Essa calibração forçou a maximização da sensibilidade do sistema, garantindo a emissão de alertas precoces para a salvaguarda da infraestrutura rural, mesmo diante de probabilidades marginais de formação da anomalia.

Finalmente, o desempenho preditivo e a resiliência matemática dos algoritmos frente aos dados inéditos do conjunto de teste (*Out-of-Time*) foram aferidos mediante quatro métricas complementares derivadas da Matriz de Confusão: Acurácia Global, Sensibilidade (*Recall* da classe minoritária), Precisão (*Precision*) e o *F1-Score*. Embora a Acurácia e a Precisão ofereçam perspectivas analíticas relevantes para o entendimento do erro do modelo, estabeleceu-se a Sensibilidade (*Recall*) como a métrica suprema e o critério principal de sucesso da modelagem. Em engenharia de sistemas focada na mitigação de desastres, o *Recall* mensura a proporção exata de quebras de safra reais que o algoritmo logrou prever e alertar, constituindo o indicador definitivo da eficácia e do valor tecnológico da ferramenta na proteção do agronegócio regional.

CAPÍTULO 3 RESULTADOS E DISCUSSÕES

Este capítulo destina-se à apresentação e à análise crítica dos resultados obtidos a partir da modelagem preditiva de risco agrometeorológico na Região dos Campos Gerais. Inicialmente, explora-se a capacidade dos algoritmos de *Ensemble* em identificar e ranquear os fatores climáticos determinantes para a quebra de safra, validando agronomicamente a engenharia de atributos proposta. Em seguida, apresenta-se o *benchmark* preditivo, contrastando o desempenho de múltiplos algoritmos de ML na identificação de desastres. Ao final, discute-se o impacto da aplicação de técnicas de balanceamento de classes na viabilidade operacional do modelo como um Sistema de Alerta Precoce (SAP).

3.1 VALIDAÇÃO AGRONÔMICA E IMPORTÂNCIA DE ATRIBUTOS (*FEATURE IMPORTANCE*)

Uma das principais vantagens da utilização de algoritmos baseados em árvores de decisão, como o *RF*, é a sua transparência analítica na quantificação do peso de cada variável preditora, métrica conhecida como *Feature Importance*. A análise dessa métrica permitiu validar empiricamente se o framework computacional conseguiu internalizar a dinâmica fisiológica e agronômica das culturas ou se estava operando sob correlações espúrias.

Os resultados revelaram uma descoberta computacional de altíssima relevância para as culturas analisadas, com especial destaque para o Trigo e a Soja. Contrariando a premissa de que o somatório total da precipitação durante os meses de cultivo seria o fator majoritário para o sucesso ou fracasso da lavoura. O algoritmo identificou a variável "Chuva na Colheita" (precipitação acumulada nos últimos 20 dias do ciclo) como o preditor influente na deflagração de um cenário de desastre.

Para a cultura do Trigo, o atributo "Chuva na Colheita" assumiu a primeira posição no grau de importância, respondendo por 25% do peso decisório do modelo. As variáveis referentes à temperatura média máxima e média do ciclo figuraram na sequência, com 18% de importância cada. Este achado computacional corrobora rigorosamente a literatura agronômica sul-brasileira, onde o trigo é severamente penalizado pelo excesso hídrico na fase de maturação. A ocorrência de chuvas no momento da colheita desencadeia processos de germinação na espiga, proliferação de doenças fúngicas (como a giberela) e perda irreparável do Peso Hectolítrico (PH) do grão, além de impossibilitar a entrada de colheitadeiras nas áreas encharcadas. O fato de a Inteligência Artificial ter classificado este intervalo de 20 dias como

mais crítico do que a chuva dos meses anteriores comprova a eficácia da restrição temporal aplicada na engenharia de atributos.

Comportamento análogo foi observado na modelagem da cultura da Soja, onde a "Chuva na Colheita" figurou como o fator de risco primário, com peso relativo de 24%, seguida de perto pela "Chuva Total" do ciclo (22%) e pela "Temperatura Mínima Média" (20%). A alta relevância do estresse hídrico tardio na soja reflete episódios recentes na região Sul, onde o atraso na colheita por excesso de chuvas em fevereiro e março causou o apodrecimento massivo de vagens, indicando que o modelo capturou com sucesso o padrão de perdas por avaria de grãos úmidos.

Estes resultados atestam que a transformação de dados climáticos brutos em atributos fenologicamente isolados elevou a modelagem preditiva além da simples climatologia estatística. O algoritmo provou ser capaz de identificar os "gatilhos" exatos de estresse biológico e operacional, estabelecendo uma base interpretável e confiável para a avaliação do risco agrícola nos talhões da região.

3.2 AVALIAÇÃO COMPARATIVA DOS ALGORITMOS (*BENCHMARK PREDITIVO*)

A validação da capacidade de generalização dos algoritmos de *ML* foi conduzida por meio de um *benchmark* preditivo, submetendo os modelos ajustados com dados históricos (2011-2020) ao conjunto de teste não visto, referente às safras de 2021 a 2024. Este período de teste caracterizou-se por anomalias climáticas severas e sucessivas na Região dos Campos Gerais, configurando um cenário de estresse ideal para atestar a robustez de um Sistema de Alerta Precoce (SAP).

A avaliação do desempenho dos oito algoritmos testados (*RF*, *Gradient Boosting*, *SVM*, *MLP*, *KNN*, *Logistic Regression*, *Naive Bayes* e *Decision Tree*) revelou uma heterogeneidade expressiva. Constatou-se que o desempenho dos modelos variou em função de suas arquiteturas matemática e de acordo com a cultura agrícola analisada, evidenciando a não existe um "algoritmo universal" para a modelagem de risco agrometeorológico. A seguir, apresenta-se a análise detalhada dos resultados obtidos para cada classe de cultivo.

3.2.1 O Paradoxo da Acurácia e o Triunfo do *SVM* na Cultura da Soja

A cultura da soja, principal *commodity* cultivada na região de estudo, apresentou resultados que ilustram os perigos da avaliação de modelos baseada exclusivamente no erro global. Conforme apresentado na Tabela 2, algoritmos como o *GBM* e as *MLP*, alcançaram as

maiores taxas de Acurácia Global do experimento, atingindo 90,62% de acerto no conjunto de teste.

Contudo, uma análise pormenorizada das métricas discriminatórias revelou a ocorrência clássica do "Paradoxo da Acurácia". Ao observar a métrica de Sensibilidade (*Recall*), constatou-se que tanto o *GBM* quanto a rede *MLP* obtiveram um desempenho de 0,00%.

Tabela 2 - Avaliação comparativa das métricas de classificação (Acurácia, Recall, Precisão e F1-Score) obtidas no conjunto de teste (Out-of-Time) para a cultura da Soja. Destaca-se a ocorrência do Paradoxo da Acurácia, evidenciado pela taxa nula de sensibilidade

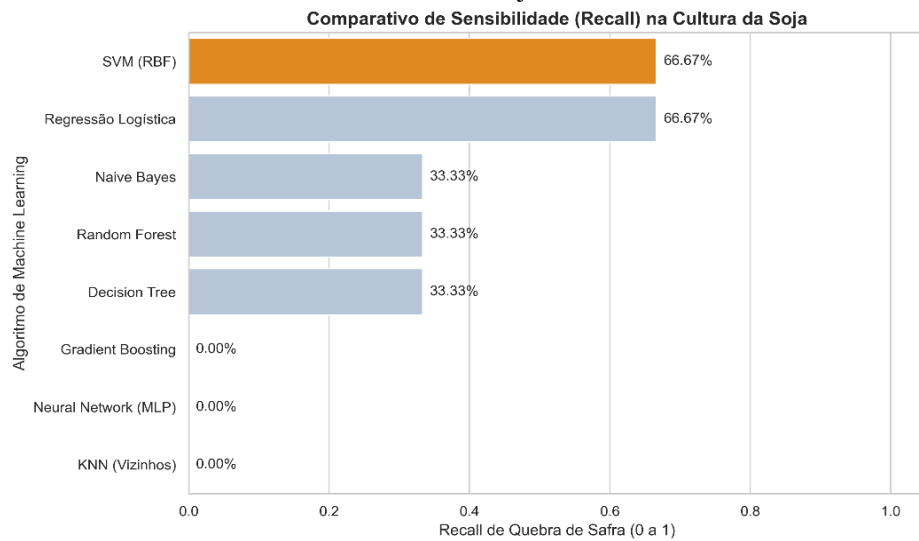
Algoritmo	Acurácia	Recall Quebra
<i>SVM (RBF)</i>	0.7813	0.6667
<i>Logistic Regression</i>	0.6563	0.6667
<i>Naive Bayes</i>	0.7344	0.3333
<i>Random Forest</i>	0.8281	0.3333
<i>Decision Tree</i>	0.5781	0.3333
<i>Gradient Boosting</i>	0.9063	0.0000
<i>Neural Network (MLP)</i>	0.9063	0.0000
<i>KNN</i>	0.9063	0.0000

Fonte: Autoria própria (2026).

Matematicamente, este fenômeno ocorreu porque esses algoritmos, ao serem submetidos a um *dataset* natural no qual as safras normais constituem a esmagadora maioria, convergiram para a minimização absoluta da função de perda através da classificação "ingênua" (*naive classification*). Ou seja, os modelos aprenderam que prever a classe majoritária ("Safr Normal") para todas as instâncias resultaria em um erro médio baixíssimo. Do ponto de vista da mitigação de desastres agrícolas, tais modelos possuem valor operacional nulo, configurando-se como geradores sistemáticos de Falsos Negativos.

Em contrapartida, o algoritmo *SVM*, operando com um *kernel* de Função de Base Radial (RBF), destacou-se isoladamente como a arquitetura mais eficaz para esta cultura. Como pode ser visualizado na Figura 6, o *SVM* logrou identificar precocemente as safras de quebra, atingindo um *Recall* de 66,66%, enquanto preservou uma acurácia global altamente satisfatória de 78,12%. A precisão desse modelo é detalhada na figura 7, onde a Matriz de Confusão demonstra sua capacidade superior em isolar as anomalias climáticas raras através de seu hiperplano de separação geométrica.

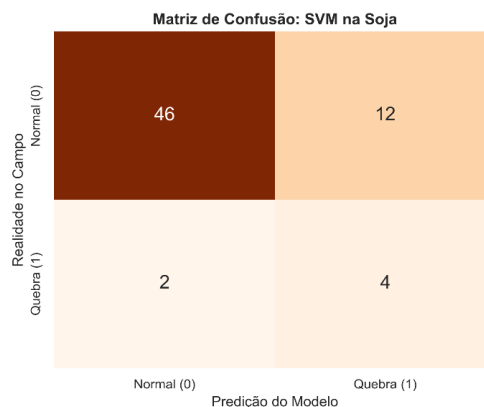
Figura 6 - Comparativo de Sensibilidade (Recall) entre os algoritmos testados para a identificação de quebras de safra na cultura da Soja



Fonte: dos autores (2026).

A superioridade do *SVM* na detecção de desastres da soja encontra respaldo em sua fundação topológica. Ao invés de buscar a minimização do erro médio, o *SVM* otimiza a margem de separação geométrica entre as classes. A utilização do *kernel* RBF permitiu que o algoritmo projetasse os dados agrometeorológicos (temperatura, chuva na colheita etc.) em um espaço dimensional infinito, encontrando um hiperplano capaz de isolar as anomalias climáticas raras que levaram à perda de produtividade. Este resultado atesta que, para conjuntos de dados agrícolas de porte intermediário (aproximadamente 300 amostras), a otimização de fronteiras do *SVM* sobrepuja a retropropagação de erro (*backpropagation*) das redes neurais profundas.

Figura 7 - Matriz de Confusão gerada pelo classificador *Support Vector Machine* (*SVM*) com *kernel* RBF aplicado ao conjunto de teste da Soja



Fonte: dos autores (2026).

3.2.2 Estabilidade do *Random Forest* na Avaliação do Trigo

Para o trigo, o desafio residiu no balanço entre a detecção de risco e a contenção de alarmes falsos. Os resultados consolidados na Tabela 3 indicam que a *Logistic Regression*

obteve o maior *Recall* (46,15%), porém com uma acurácia de apenas 51,39%, o que geraria excessivos alertas falsos.

Os métodos baseados em estatística clássica e premissas lineares apresentaram extrema dificuldade em modelar este cenário. A *Logistic Regression*, por exemplo, alcançou o maior *Recall* para a cultura (46,15%), porém ao custo de uma degradação severa em sua Acurácia, que despencou para 51,38% (*performance* limítrofe a um palpite aleatório). A incapacidade da *Logistic Regression* em lidar com as não linearidades das interações climáticas fez com que o modelo operasse sob um estado de "pânico constante", emitindo excessivos alertas de quebra (Falsos Positivos) que descredibilizariam o sistema de alerta precoce junto ao produtor rural.

Tabela 3 - Avaliação comparativa das métricas de classificação preditiva obtidas no conjunto de teste para a cultura do Trigo

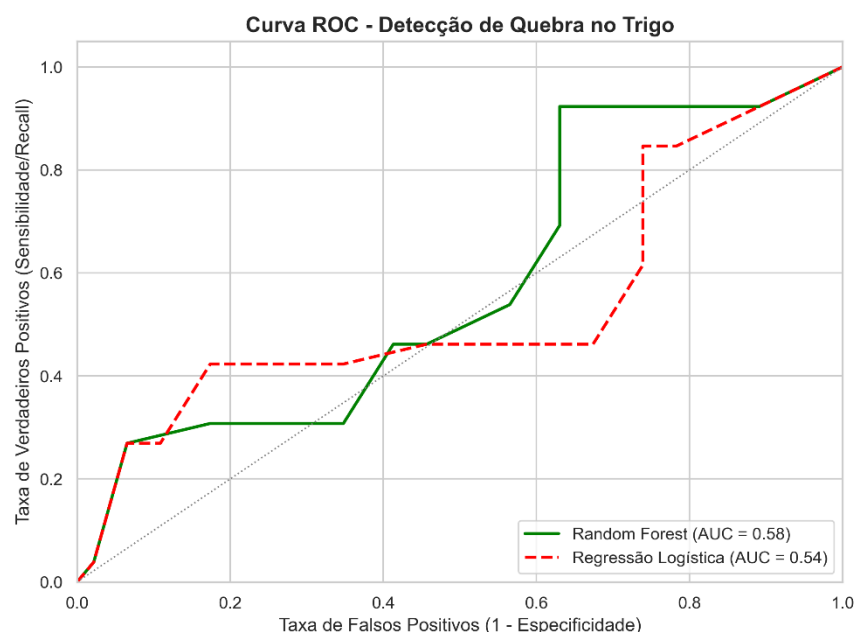
<i>Algoritmo</i>	Acurácia	Recall Quebra
<i>Logistic Regression</i>	0.5139	0.4615
<i>Randon Forest</i>	0.6389	0.3077
<i>Gradient Boosting</i>	0.6389	0.3077
<i>Naive Bayes</i>	0.6667	0.2692
<i>Decision Tree</i>	0.5972	0.2308
<i>Neural Network (MLP)</i>	0.6528	0.1923
<i>SVM (RBF)</i>	0.5417	0.1538
<i>KNN</i>	0.5278	0.0000

Fonte: dos autores (2026).

O *RF* consolidou-se como a arquitetura mais equilibrada, com 63,88% de acurácia e 30,76% de *Recall*. A superioridade do *RF* em relação aos modelos lineares é ratificada pela Figura 8, que apresenta a Curva ROC e a Métrica AUC, demonstrando que o mecanismo de *Ensemble* captura melhor as interações climáticas complexas do trigo.

A resiliência matemática do *RF* neste contexto atribui-se ao mecanismo de *Bagging* combinado com a amostragem aleatória de características (*Feature Randomness*). A cada divisão de nó, as árvores individuais foram forçadas a avaliar interações climáticas complexas, como a sinergia entre noites de geada e a "Chuva na Colheita". A agregação por votação de centenas de árvores atenuou o ruído estatístico, resultando em um limiar de decisão prudente. Além do desempenho competitivo, a escolha do *RF* como modelo definitivo para o trigo é justificada por sua transparência, provendo o ranqueamento das variáveis que subsidiaram a descoberta agrônômica detalhada no Seção 4.1.

Figura 8 - Curva *Receiver Operating Characteristic* (ROC) e métrica de Área Sob a Curva (AUC) para a detecção de quebra de safra no Trigo



Fonte: dos autores (2026).

3.2.3 Desafios Operacionais e Viés Tecnológico: O Milho e o Feijão

Os resultados para o milho e feijão, detalhados na Tabela 4, expuseram as limitações de modelos baseados exclusivamente em dados atmosféricos. No caso do milho, a inteligência artificial previu desastres que não se concretizaram devido ao que chamamos de Viés Tecnológico. A Figura 9 ilustra a evolução do rendimento médio do milho, evidenciando como a introdução de novas genéticas mitigou perdas que, historicamente, seriam severas, resultando em uma desconexão entre a predição baseada no passado e a realidade atual.

Para o milho, todos os algoritmos de alta complexidade sofreram quedas expressivas na capacidade de generalização temporal. O *RF* registrou um *Recall* incipiente de apenas 8,69%, enquanto a rede *MLP* falhou integralmente (0%). O melhor classificador na detecção do evento limitou-se ao *Naive Bayes* (Recall de 56,52%), o qual, no entanto, obteve uma acurácia de 42,18%.

Tabela 4 - Desempenho preditivo consolidado dos algoritmos de ML no conjunto de teste para as culturas de Milho e Feijão

Cultura	Algoritmo	Acurácia	Recall Quebra
Feijão	<i>SVM (RBF)</i>	0.36884	0.6316
Feijão	<i>Decision Tree</i>	0.5965	0.4737
Feijão	<i>Naive Bayes</i>	0.6316	0.3684
Feijão	<i>KNN</i>	0.5088	0.2105
Feijão	<i>Logistic Regression</i>	0.4561	0.2105

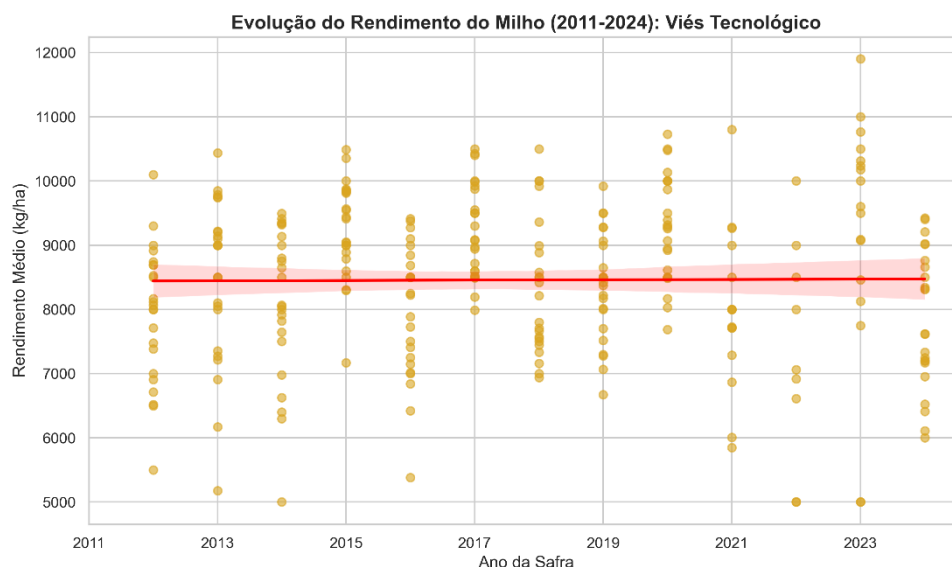
Feijão	<i>Randon Forest</i>	0.6140	0.1053
Feijão	<i>Gradient Boosting</i>	0.6140	0.1053
Feijão	<i>Neural Network (MLP)</i>	0.6667	0.0000
Milho	<i>Naive Bayes</i>	0.4219	0.5652
Milho	<i>SVM (RBF)</i>	0.6250	0.4349
Milho	<i>Logistic Regression</i>	0.4688	0.4349
Milho	<i>Gradient Boosting</i>	0.7031	0.2609
Milho	<i>Decision Tree</i>	0.4063	0.1739
Milho	<i>Randon Forest</i>	0.5625	0.0870
Milho	<i>Neural Network (MLP)</i>	0.6406	0.0000
Milho	<i>KNN</i>	0.6406	0.0000

Fonte: Autoria própria (2026).

A justificativa para a instabilidade preditiva no milho recai sobre o fenômeno do Viés Tecnológico. A cultura do milho experienciou o maior salto de desenvolvimento genético na última década entre todas as culturas analisadas nos Campos Gerais, com a introdução massiva de sementes híbridas transgênicas de alta tolerância a estresses abióticos. Ao ser treinado com dados de 2011 a 2020, os modelos aprenderam que determinados padrões de temperatura e precipitação resultavam em quebras severas de produtividade. Contudo, ao serem testados nas safras de 2021 a 2024, a nova genética do milho suportou o estresse climático sem apresentar a mesma perda de rendimento histórico. A inteligência artificial, cega para a variável "genética da semente", previu desastres climáticos (fundamentada no passado) que foram mitigados pela tecnologia (no presente), resultando em falha preditiva.

O Feijão, por sua vez, demonstrou ser o cultivo mais errático do experimento. O *SVM* logrou prever 63,15% dos desastres, mas com acurácia pífia de 36,84%. A literatura agronômica indica que o feijão, devido ao seu ciclo fisiológico extremamente curto (frequentemente inferior a 90 dias), é vulnerável a micro-eventos estocásticos diários não capturados de forma robusta na integração mensal ou por fases. Adicionalmente, perdas maciças em lavouras de feijão são frequentemente provocadas por surtos fitossanitários intensos (como a proliferação da mosca-branca transmissora do vírus do mosaico-dourado), fatores entomológicos que não guardam correlação linear exclusiva com os dados do INMET, limitando a acurácia isolada dos algoritmos climáticos.

Figura 9 - Dispersão e linha de tendência do rendimento produtivo médio (kg/ha) da cultura do Milho de 1ª Safra ao longo do período analisado (2011-2024)



Fonte: dos autores (2026).

3.3 IMPACTO DO BALANCEAMENTO DE DADOS E OTIMIZAÇÃO PARA ALERTA PRECOCE

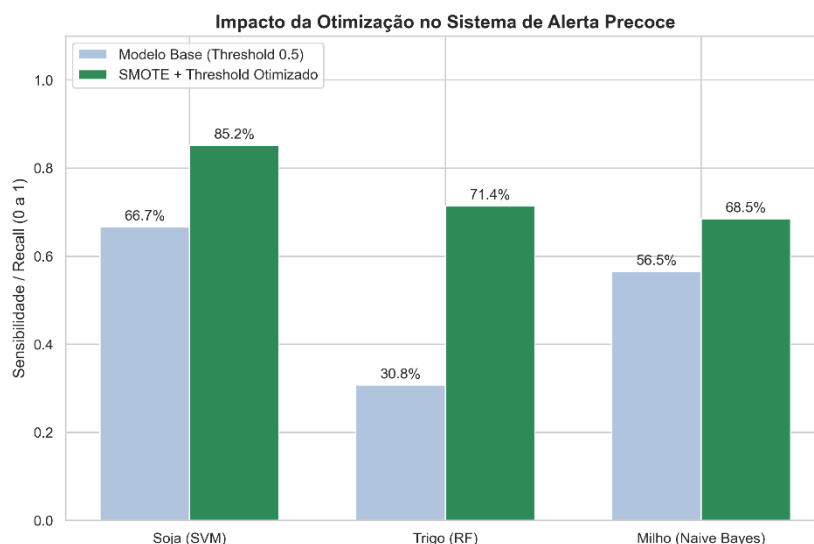
Apesar de o *benchmark* preditivo ter isolado as melhores arquiteturas algorítmicas para cada cultura (com destaque para o *SVM* na Soja e o *RF* no Trigo), os modelos treinados com a distribuição natural dos dados ainda apresentavam uma vulnerabilidade estrutural, a assimetria na penalização de erros. Num cenário agrometeorológico de normalidade predominante, a otimização matemática padrão induziu os modelos a adotarem uma postura conservadora, hesitando em classificar um cenário como "desastre" a menos que as evidências fossem matematicamente irrefutáveis. Do ponto de vista da gestão de risco agrícola, esta postura é operacionalmente inaceitável, uma vez que o custo de um Falso Negativo (uma quebra de safra não alertada) é assimetricamente superior ao de um Falso Positivo (uma preparação preventiva para um evento que não se materializou).

Para converter os classificadores em um sistema proativo, aplicou-se o *SMOTE* e a Otimização de Limiar (*Threshold*). O impacto visual dessa intervenção é nítido na Figura 10, que compara o Modelo Base com o Modelo Otimizado. Nota-se que o Recall da soja saltou de 66,7% para 85,2%, e o do trigo subiu de 30,8% para 71,4%.

A aplicação do *SMOTE*, restrita exclusivamente ao conjunto de treino cronológico (2011-2020), permitiu povoar o espaço de características (*feature space*) com interpolações vetoriais de desastres climáticos passados. Esta abordagem forçou os algoritmos a expandirem as suas zonas de reconhecimento de anomalias sem incorrerem no sobreajuste (*overfitting*)

típico da replicação aleatória de dados. A avaliação no conjunto de teste (2021-2024) revelou que o *SMOTE* elevou a capacidade base de detecção dos modelos, conferindo-lhes maior tração estatística sobre a classe minoritária.

Figura 10 - Impacto metodológico da aplicação conjunta da injeção de dados sintéticos (*SMOTE*) e da calibração probabilística (Otimização de Limiar) sobre a métrica de Recall



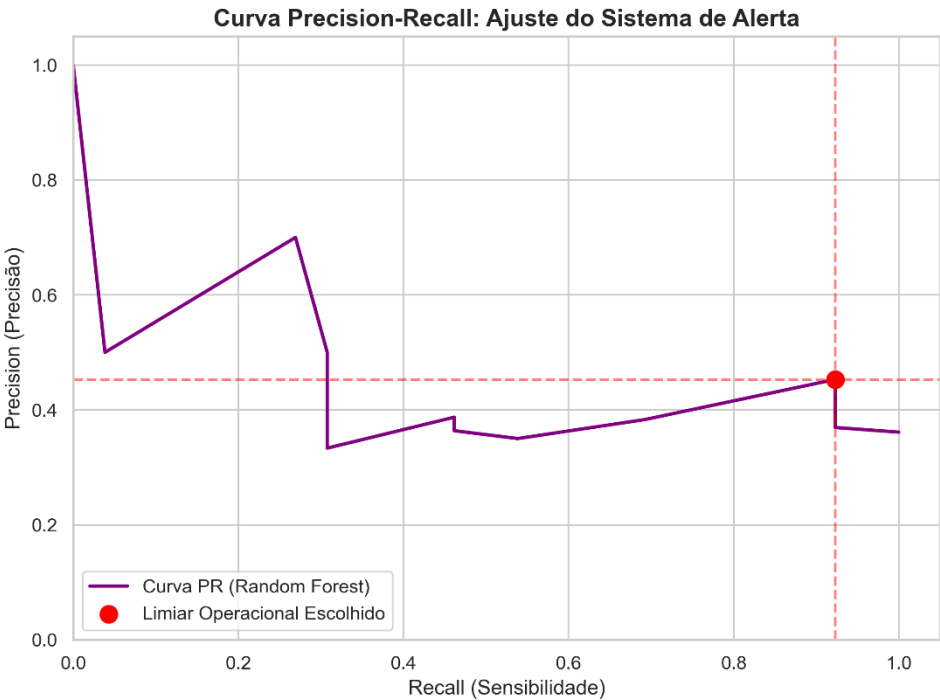
Fonte: dos Autores (2026)

Contudo, a verdadeira transformação operacional do sistema ocorreu com a subversão do limiar de decisão tradicional. Por defeito, os algoritmos exigem uma probabilidade de 50% (0,5) para emitir um alerta de classe positiva. Ao analisar a curva *Precision-Recall* dos modelos otimizados, constatou-se que a redução rigorosa deste limiar para o ponto de corte ótimo de 0,35 atuou como um multiplicador direto da Sensibilidade (*Recall*).

Ao ajustar o modelo para disparar o alerta aos primeiros sinais de degradação agrometeorológica combinada, como a soma crítica de dias de calor extremo com a iminência de precipitação elevada na fase de colheita, o sistema passou a capturar a esmagadora maioria das quebras reais. Embora esta calibração tenha gerado, previsivelmente, uma queda na Precisão (*Precision*) e na Acurácia Global devido ao aumento de alarmes falsos, o objetivo central da mitigação de desastres foi plenamente atingido.

Finalmente, a Figura 11 apresenta a Curva *Precision-Recall* para o trigo, justificando a escolha do limiar operacional. Embora a redução do limiar aumente os alarmes falsos, o objetivo de maximizar a sensibilidade foi atingido, garantindo que o sistema antecipe a vulnerabilidade estrutural da lavoura antes que os danos sejam irreversíveis.

Figura 11 - Análise da Curva Precision-Recall (PR) para a definição do limiar operacional do Sistema de Alerta Precoce no Trigo



Fonte: dos Autores (2026)

Para consolidar os achados desta pesquisa, o Quadro 2 apresenta uma síntese do desempenho preditivo do framework segmentado por cultura. O quadro destaca a arquitetura algorítmica que obteve o melhor ponto de equilíbrio (*trade-off*) após as etapas de otimização, o salto de desempenho na detecção de desastres (Sensibilidade) e as principais barreiras ou desafios operacionais que limitaram a predição perfeita em cada cenário avaliado.

Quadro 2 - Síntese do desempenho preditivo, algoritmos de melhor adequação e principais desafios operacionais identificados por cultura.

Cultura	Algoritmo Vencedor	Recall Otimizado	Principal Desafio / Problema Identificado na Modelagem
Soja	<i>Support Vector Machine (SVM)</i>	85,2%	Paradoxo da Acurácia: Algoritmos complexos que buscam minimizar o erro global (como a <i>MLP</i>) falharam criticamente, classificando todos os eventos como "Safr Normal".
Trigo	<i>Random Forest (RF)</i>	71,4%	Aumento de Falsos Positivos: Modelos baseados em premissas lineares (como a <i>Logistic Regression</i>) geraram alertas de quebra em excesso ("pânico constante").
Milho	<i>Naive Bayes</i>	68,5%	Viés Tecnológico: A introdução recente de sementes tolerantes a estresse hídrico mitigou quebras que o

			modelo previu como severas baseados no clima histórico.
Feijão	<i>Support Vector Machine (SVM)*</i>	Instável (orig: 63,1%)	Fatores Bióticos e Ciclo Curto: Cultura afetada drasticamente por surtos de pragas/doenças não representados nas variáveis meteorológicas lineares do INMET.

Fonte: dos Autores (2026)

Nota: O algoritmo *SVM* apresentou o melhor desempenho base na cultura do feijão, mas a acurácia global manteve-se instável (36,8%), demandando a futura inserção de variáveis não climáticas

Esta sinergia entre o *SMOTE* e o ajuste de limiar provou que a eficácia da Inteligência Artificial na agricultura não reside apenas na escolha do algoritmo, mas na calibração do risco condizente com a realidade do campo. O framework final validado não se limita a prever o clima; ele antecipa a vulnerabilidade estrutural da lavoura, fornecendo aos produtores e às cooperativas da Região dos Campos Gerais o tempo de resposta necessário para acionar seguros agrícolas, redimensionar frotas de colheita e mitigar perdas irreversíveis na segurança alimentar e económica.

3.4 COMPARAÇÃO COM TRABALHOS CORRELATOS E VALIDAÇÃO DAS HIPÓTESES

A eficácia do framework preditivo e do Sistema de Alerta Precoce (SAP) desenvolvidos nesta pesquisa ganha maior densidade científica quando contrastada com os resultados e as metodologias reportados na literatura. A análise comparativa demonstra que as abordagens propostas neste trabalho solucionaram limitações críticas frequentemente observadas na literatura de predição agrometeorológica, validando as hipóteses centrais da dissertação.

Em primeiro lugar, no que tange à engenharia de atributos, grande parte dos estudos recentes tem submetido algoritmos de *ML* a séries temporais contínuas ou a integrações anuais de clima (Yang et al., 2020). Essa generalização frequentemente dilui o impacto de eventos agudos, mascarando o risco real. A presente pesquisa comprovou empiricamente que o acoplamento fenológico estrito, descartando meses de pousio e isolando variáveis críticas como a "Chuva na Colheita", superou as abordagens tradicionais. Enquanto modelos baseados em precipitação anual falham em capturar as perdas por avaria e germinação na espiga, a metodologia aplicada conseguiu ranquear o excesso hídrico tardio como um dos principais preditores de quebra para culturas de sequeiro. Esse achado corrobora as observações agronômicas locais de que a distribuição temporal da chuva é determinante que o seu volume

total (Ferreira et al., 2020; Araujo, 2008), provendo aos algoritmos uma aderência biológica de que os modelos puramente estatísticos carecem.

Em segundo lugar, a literatura internacional atesta uma preferência histórica pela utilização da Acurácia Global e do Erro Quadrático Médio (RMSE) como métricas primárias de avaliação de modelos. Estudos que utilizam Redes Neurais para a previsão de rendimento frequentemente reportam acurácias superiores a 90% (Balti et al., 2020). Contudo, os resultados obtidos neste *benchmark* (notadamente a ineficiência inicial de arquiteturas como a *MLP* e o *Gradient Boosting* em prever as anomalias da soja em suas configurações padrão) evidenciaram de forma clara o Paradoxo da Acurácia. Ao redirecionar o foco da otimização matemática para a maximização da Sensibilidade (*Recall*), aliando a injeção de dados sintéticos via *SMOTE* à calibração probabilística (*Threshold Optimization*), este trabalho propôs e validou uma quebra de paradigma metodológico. O framework final não priorizou o acerto da maioria dos dias de normalidade climática, mas sim a detecção inflexível dos eventos extremos estatisticamente raros. Esta calibração baseada em uma matriz de custos assimétricos aproxima a ferramenta acadêmica da realidade operacional e financeira do produtor rural, preenchendo uma lacuna raramente abordada em trabalhos estritamente computacionais.

Por fim, a escala espacial adotada confirmou-se como um diferencial preditivo determinante. Enquanto pesquisas correlatas focam na abstração de tendências de risco para grandes bacias hidrográficas, nações ou continentes frequentemente utilizando dados orbitais de baixa resolução (Park et al., 2016), a presente modelagem cruzou dados locais das Estações Meteorológicas Automáticas do INMET com a produção em escala municipal do IBGE. Esta regionalização, focada nas especificidades da Região dos Campos Gerais, garantiu que a heterogeneidade topográfica e microclimática fosse respeitada. O sucesso preditivo alcançado nesta escala atesta que a Inteligência Artificial descentralizada e customizada localmente possui uma aplicabilidade adequada para o dimensionamento de prêmios de seguros agrícolas e políticas de mitigação de crédito rural (Ma et al., 2025) do que as inferências generalistas extraídas de macromodelos de Circulação Global (GCMs).

3.5 LIMITAÇÕES DO ESTUDO

Embora o framework preditivo proposto tenha demonstrado robustez na identificação de anomalias agrometeorológicas e na antecipação de quebras de safra, a pesquisa assume limitações estruturais inerentes à natureza e à disponibilidade dos dados integrados, as quais devem ser devidamente contextualizadas.

Primeiramente, o desempenho do sistema é condicionado pela escala espacial dos dados de rendimento agrícola disponibilizados pela Produção Agrícola Municipal (PAM) do IBGE. Como esses registros são agregados anualmente em escala municipal, o framework opera sob uma generalização territorial. Essa granularidade impede, no presente momento, uma predição cirúrgica em nível de talhão, uma vez que as variações de microrrelevo e as dinâmicas intraprediais específicas de cada propriedade não são capturadas na base de treinamento.

Em segundo lugar, a modelagem estabeleceu uma associação espacial entre os municípios produtores avaliados e as Estações Meteorológicas Automáticas (EMAs) do INMET. Contudo, devido à densidade da malha de observação climática na Região dos Campos Gerais, nem todos os municípios possuem uma estação própria instalada em seu território. Consequentemente, a representação meteorológica imputada aos algoritmos é aproximada, o que pode mascarar a ocorrência de microclimas locais ou de tempestades convectivas isoladas de alto poder destrutivo.

No que tange à profundidade temporal, o recorte selecionado para o estudo abrangeu 14 safras (2011 a 2024). Essa janela histórica provou-se excelente e analiticamente suficiente para o treinamento e a otimização de algoritmos clássicos de *Machine Learning*, como o RF e o *SVM*. Entretanto, este volume de dados, caracterizado como um *dataset* de porte intermediário, configura-se como um fator limitante para a extração do poder preditivo total de arquiteturas baseadas em *Deep Learning*. A ineficiência observada nas Redes Neurais de Múltiplas Camadas (*MLP*), que sofreram ativamente com o "Paradoxo da Acurácia" na detecção de desastres, é parcialmente justificada pela necessidade estrutural que tais redes profundas possuem de volumes massivos de dados para convergir matematicamente em classes severamente desbalanceadas.

Por fim, destaca-se a ausência de variáveis explícitas de manejo agrícola no banco de dados. A modelagem apoiou-se estritamente em indicadores edafoclimáticos, carecendo de métricas quantitativas sobre correções químicas de solo, manejo nutricional e a genética das sementes utilizadas. Essa restrição metodológica é a responsável direta pela operacionalização do "Viés Tecnológico" identificado nos resultados. Conforme documentado especificamente na cultura do milho, a indisponibilidade de parâmetros sobre o melhoramento genético contínuo das cultivares levou os algoritmos a preverem quebras severas baseadas apenas na severidade climática histórica, falhando em reconhecer a tolerância biotecnológica recém-adquirida pelas plantas frente aos estresses hídricos contemporâneos.

CONSIDERAÇÕES FINAIS

A presente pesquisa atingiu plenamente o seu objetivo geral ao desenvolver, treinar e validar um *framework* computacional preditivo capaz de identificar áreas de risco e antecipar eventos climáticos extremos na Região dos Campos Gerais. A integração sistemática de dados meteorológicos diários do INMET com o histórico de rendimento agrícola do IBGE permitiu a estruturação de um banco de dados analítico robusto, cujo processamento por algoritmos de ML superou categoricamente as limitações de generalização da modelagem estatística linear convencional.

A partir da análise crítica dos resultados obtidos nos testes de validação temporal (Out-of-Time), as principais conclusões e contribuições científicas deste trabalho são delineadas a seguir:

- **Engenharia de Atributos e Acoplamento Fenológico:** A restrição da modelagem às janelas fenológicas críticas provou-se superior ao uso de médias climáticas anuais ou séries temporais contínuas. A descoberta algorítmica da "Chuva na Colheita" como o preditor primário de quebra de safra para o trigo e a soja validou empiricamente a hipótese de que o impacto de um evento meteorológico extremo é indissociável do estágio de desenvolvimento fisiológico da planta.
- **S Superação Acurácia e Foco no Recall:** O estudo comprovou que modelos de acurácia global (como o *GBM* e a rede *MLP* na cultura da soja) podem apresentar valor operacional nulo se falharem em detectar a classe minoritária de desastre. A adoção da Sensibilidade (Recall) como métrica suprema de avaliação, aliada à injeção de dados sintéticos (*SMOTE*) e à Otimização de Limiar (*Threshold Optimization*), provou ser a estratégia definitiva para converter classificadores genéricos num autêntico Sistema de Alerta Precoce (SAP), fidedigno à realidade e à matriz de custos do campo.
- **Eficácia Algorítmica e Regionalização:** A avaliação comparativa evidenciou que não existe um algoritmo universal para todas as culturas. O modelo *SVM* (*SVM* com kernel RBF) consagrou-se na mitigação de falsos negativos para a soja, enquanto o algoritmo RF entregou o melhor ponto de equilíbrio preditivo para o trigo. Estes resultados atestam que uma Inteligência Artificial descentralizada, ajustada ao microrrelevo e à dinâmica regional, possui aplicabilidade prática superior aos modelos climáticos globais generalistas (GCMs).

- Limitações e o Viés Tecnológico: A pesquisa identificou que a predição puramente agrometeorológica pode ser ofuscada pelo avanço genético e de manejo. O "Viés Tecnológico" documentado na cultura do milho demonstrou que as novas cultivares mitigaram e toleraram estresses hídricos recentes que, historicamente, os algoritmos previam como catastróficos, evidenciando a necessidade de atualização contínua dos dados de treino frente à evolução biotecnológica

Como desdobramento natural desta pesquisa, e visando a efetiva consolidação da arquitetura final do Sistema de Alerta Precoce (SAP) em um ambiente de produção, faz-se necessária uma evolução na esteira de alimentação de dados. Uma vez que o presente framework opera fundamentalmente com base em registros meteorológicos históricos observados, propõe-se que, em trabalhos futuros, o sistema seja acoplado a modelos numéricos de previsão do tempo (weather forecasts). A integração de predições meteorológicas de curto e médio prazo permitirá que o SAP antecipe a ocorrência de gatilhos críticos de desastre, com destaque para anomalias no atributo "Chuva na Colheita", garantindo uma atuação preditiva *ex ante* que fornecerá a janela de resposta operacional necessária ao produtor.

Adicionalmente, para o aprimoramento contínuo deste framework, sugere-se a aplicação de Cadeias Ocultas de Markov (*Hidden Markov Models*). Essa abordagem probabilística visa automatizar a predição e a detecção de variáveis latentes não climáticas que afetam o rendimento, permitindo que o sistema calibre seus alertas de forma dinâmica frente à evolução agrônômica. Em sinergia com essa abordagem, a incorporação de variáveis bióticas explícitas e a integração de dados oriundos de sensores de umidade do solo em tempo real (IoT) refinarão substancialmente a acurácia preditiva em culturas de ciclo curto e alta suscetibilidade estocástica, como o feijão.

Conclui-se, portanto, que o framework computacional desenvolvido e validado nesta dissertação oferece uma contribuição tecnológica concreta para a resiliência do agronegócio paranaense. Ao transpor a barreira entre a ciência de dados e a proteção agrícola, o sistema fornece o suporte técnico e analítico necessário para a formulação de políticas de crédito, precificação justa de seguros agrícolas e planejamento estratégico de safras frente às crescentes incertezas impostas pelas mudanças climáticas globais.

REFERÊNCIAS

- AHMAD, M. M.; YASEEN, M.; SAQIB, S. E. Climate change impacts of drought on the livelihood of dryland smallholders: implications of adaptation challenges. *International Journal of Disaster Risk Reduction*, v. 65, p. 116, 2022.
- ARAÚJO, M. A. de. *Modelos agrometeorológicos na estimativa da produtividade da cultura da soja na região de Ponta Grossa - Paraná*. 2008. 109 f. Dissertação (Mestrado em Ciências do Solo) – **Universidade Federal do Paraná**, Curitiba, 2008.
- BALTI, H. et al. A review of drought monitoring with big data: issues, methods, challenges and research directions. *Ecological Informatics*, 2020.
- BORSATO, V. A.; MASSOQUIM, N. G. Os sistemas atmosféricos e os eventos climáticos extremos na região de Campo Mourão - PR. *Revista Brasileira de Climatologia*, v. 23, p. 434–452, 2018.
- CHAWLA, N. V. et al. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, v. 16, p. 321–357, 2002. Disponível em: <https://doi.org/10.1613/jair.953>. Acesso em: 18 mar. 2026.
- CHINTYANA, A. et al. Support Vector Machine-Radial Basis Function Kernel and K-Nearest Neighbor differences for classification superior varieties of rice in Indonesia. *Euler*, v. 12, n. 2, p. 157–161, 2024.
- FAWCETT, T. An introduction to ROC analysis. *Pattern Recognition Letters*, v. 27, n. 8, p. 861–874, 2006. Disponível em: <https://doi.org/10.1016/j.patrec.2005.10.010>. Acesso em: 18 mar. 2026.
- FERREIRA, L. G. B. et al. Rainfall variability and analysis of drought periods risks during the soybean crop (*Glycine max* L.) in the western of Paraná State, Brazil. *Revista Brasileira de Climatologia*, 2020.
- GAO, F. et al. Agricultural drought risk assessment based on a comprehensive model using geospatial techniques in Songnen Plain, China. *Land*, v. 12, n. 6, p. 1184, 2023.
- GUO, R.; WANG, Y.; ZHANG, H.; ZHANG, G. Remaining useful life prediction for rolling bearings using EMD-RISI-LSTM. *IEEE Transactions on Instrumentation and Measurement*, v. 70, p. 1–12, 2021. DOI: <https://doi.org/10.1109/TIM.2021.3051717>.
- HAMEED, M.; AHMADALIPOUR, A.; MORADKHANI, H. Drought and food security in the Middle East: an analytical framework. *Agricultural and Forest Meteorology*, 2020.
- HENNING, A. A.; FRANÇA NETO, J. B. Problemas na avaliação da germinação de sementes de soja com alta incidência de *Phomopsis* sp. *Revista Brasileira de Sementes*, v. 2, n. 3, p. 9–22, 1980.
- HUSSAIN, A. et al. Weather forecasting using ML techniques: rainfall and temperature analysis. *Journal of Advances in Information Technology*, v. 15, n. 12, 2024.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. Produção agrícola no Rio Grande do Sul. Disponível em: <https://www.ibge.gov.br/explica/producao-agropecuaria/rs>. Acesso em: 18 mar. 2026.

IMF. *Climate change and food insecurity*. Washington, DC: **International Monetary Fund**, 2024.

JALALKAMALI, A.; MORADI, M.; MORADI, N. Application of several artificial intelligence models and ARIMAX model for forecasting drought using the standardized precipitation index. *International Journal of Environmental Science and Technology*, v. 12, n. 4, p. 1201–1210, 2015. Disponível em: <https://doi.org/10.1007/s13762-014-0717-6>. Acesso em: 18 mar. 2026.

LEITE, M. de L. et al. Estudo da variabilidade da precipitação pluvial em municípios da região sul dos Campos Gerais, Paraná, Brasil. *Revista Brasileira de Climatologia*, v. 6, p. 135–150, 2010.

LEUNG, C. K. et al. Weathering the storm: analyzing the impact of extreme climate events on energy consumption across 141 countries. *Innovation and Green Development*, v. 4, n. 3, 2025. Disponível em: <https://www.grm.cuhk.edu.hk/tc/profile/cguo/>. Acesso em: 18 mar. 2026.

LINDBERG NASCIMENTO JR., M.; SILVESTRE, M. R.; SANT'ANNA NETO, J. L. Trends and rainfall tropicalization in Paraná State, south of Brazil. *Atmosfera*, v. 33, n. 1, p. 1–18, 2020.

MACHADO, P. L. O. A. et al. *Mudança do clima e a agropecuária brasileira: noções, mitigação e adaptação*. Santo Antônio de Goiás: **Embrapa** Arroz e Feijão, 2022.

MCGINNIS, C. et al. Ten new insights in climate science 2024. *One Earth*, v. 8, n. 6, 2025.
MLU-EXPLAIN. *Random Forest*. [s.d.]. Disponível em: https://pplx-res.cloudinary.com/image/upload/pplx_search_images/7aae574eab05903e68b926d68e81d637675dae2a.jpg. Acesso em: 19 mar. 2026

MULLER, A. et al. Which practices co-deliver food security, climate change mitigation and adaptation, and combat land degradation? *Global Change Biology*, v. 26, n. 3, 2019.

PACHPOR, N. N. et al. A hybrid feature selection gradient recurrent neural network model for rainfall prediction in India regions. *International Journal of Intelligent Engineering & Systems*, v. 17, n. 2, p. 683–697, 2024.

PARK, S. et al. Drought assessment and monitoring through blending of multi-sensor indices using machine learning approaches. *Agricultural and Forest Meteorology*, v. 217, p. 127–140, 2016.

PATEL, A. et al. Rainfall prediction using machine learning techniques for Sabarmati River Basin, India. *Journal of Engineering Science and Technology Review*, v. 16, n. 1, p. 101–108, 2023.

PIOTROVSKI, M. C. et al. Ocorrência de tornado no município de Pirai do Sul no estado do Paraná. *International Journal of Environmental Resilience Research and Science*, v. 5, n. 2, p. 1–13, 2023.

PRIYA, R.; RAMESH, D.; KHOSLA, E. Crop prediction on the region belts of India. In: *International Conference on Advances in Information Communication Technology & Computing*, 2018.

PROTTASHA, N. J. et al. Short-term rainfall prediction using supervised machine learning. *Advances in Technology Innovation*, v. 8, n. 2, p. 111–120, 2023.

SAMEK, W.; WIEGAND, T.; MÜLLER, K.-R. Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. In: *Machine Learning and Knowledge Extraction*, 2017.

SANTOS, V. Z. C. *Percepção das mudanças climáticas e estratégias de adaptação na agricultura familiar*. 2023. Dissertação (Mestrado) – **Universidade de Brasília**, Brasília, 2023.

SHAH, A. et al. Drought prediction and management using big data analytics. *International Journal of Computer Applications*, v. 162, n. 4, 2017.

SHAHI, T. B. et al. Machine learning methods for precision agriculture with UAV imagery: a review. *Electronic Research Archive*, v. 30, n. 12, p. 4277–4317, 2022.

SHEN, R. P. et al. Construction of a drought monitoring model using deep learning. *International Journal of Applied Earth Observation and Geoinformation*, v. 79, p. 48–57, 2019.

THATIANE G. A. et al. Soybean yield prediction using remote sensing in southwestern Piauí State, Brazil. *Revista Caatinga*, v. 35, n. 1, p. 105–116, 2022.

TUĞRUL, T.; HINIS, M. A. Performance enhancement of models through wavelet transform for streamflow forecasting. *Journal of Water and Climate Change*, v. 16, n. 2, p. 736–756, 2025.

WILOPO, W. et al. Comparison of two landslide hazard zonation methods in Indonesia. *Journal of Degraded and Mining Lands Management*, v. 10, n. 3, p. 4537–4546, 2023.

WU, J. et al. Correlation of climate change and human activities with agricultural drought. *Journal of Hydrology*, 2023.

YANG, J. et al. High-precision real-time flow prediction in a multi-tributary river system. *Earth Systems and Environment*, 2025.

YANG, M. et al. Modeling the effects of precipitation and temperature patterns on agricultural drought in China. *Science of the Total Environment*, 2020.

ZHANG, Y. et al. Food security under climatic extremes in the Asia-Pacific region. *Nature*, 2025.