

Predictive Modeling of Rice Production in South America Using Machine Learning

Renan Grion¹, Ana Bari Koifman¹, Laura O. Moraes¹,
Mariana Simoes Larraz Ferreira¹

¹Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
Rio de Janeiro – RJ – Brasil

renan.grion@edu.unirio.br, anabkoifman@edu.unirio.br,

laura@uniriotec.br, mariana.ferreira@unirio.br

Abstract. Research Context: Rice is a staple food for nearly half of the global population and is highly vulnerable to climate variability. In South America, accurate forecasting models are critical to ensure food security and guide agricultural decision-making. **Scientific and/or Practical Problem:** Traditional methods often fail to capture the complexity of agricultural systems shaped by environmental and socioeconomic factors. The lack of predictive systems adapted to South America restricts the ability of governments and producers to anticipate risks. This research addresses both a scientific challenge, by advancing predictive modeling, and a practical challenge, by supporting sustainable agricultural planning. **Proposed Solution and/or Analysis:** The study integrates agricultural production data from FAOSTAT (2004–2023) with climate variables from NASA POWER (2011–2023). It compares Linear Regression, Support Vector Regression (SVR), and XGBoost Regressor to identify the most reliable configuration for rice production forecasting. **Related IS Theory:** The work aligns with decision support systems (DSS) and information systems for sustainability, contributing to the Brazilian Grand Challenges in Information Systems (GranDSI-BR 2016–2026). **Research Method:** The approach combined exploratory analysis, training with multiple time windows (5, 10, 15 years), and cross-validation, using Mean Absolute Percentage Error (MAPE) as the main evaluation metric. **Summary of Results:** Longer windows (10 years) improved stability. XGBoost achieved the lowest MAPE and greater robustness, while climate data integration reduced errors in atypical years. **Contributions and Impact to IS area:** The study demonstrates the potential of machine learning-based predictive systems as components of agricultural information systems, reinforcing their role in enhancing food security, supporting public policies, and promoting sustainable resource management in South America and beyond.

1. Introdução

A expectativa mundial é de que a população cresça por mais 50 ou 60 anos, atingindo um pico de cerca de 10,3 bilhões de pessoas em meados da década de 2080 (ONU, 2024). Esse cenário projeta pressão crescente sobre os sistemas agrícolas, já que as taxas atuais de aumento na produção não suprem a demanda alimentar futura [Ray et al. 2013]. Com essa projeção, governos e setores produtivos precisam estar preparados para lidar com

choques de oferta e garantir segurança alimentar, especialmente no caso de culturas relevantes como arroz (*Oryza sativa*), milho (*Zea mays*), trigo (*Triticum aestivum*) e cevada (*Hordeum vulgare*) [Albahri et al. 2023].

O arroz (*Oryza sativa* L.), em particular, destaca-se por desempenhar um papel crucial como a principal fonte calórica para aproximadamente metade da população global, consolidando-se como uma cultura essencial não só na garantia da segurança alimentar, mas também intrinsecamente ligada ao desenvolvimento econômico e geração de empregos, principalmente em países do Sul Global [Dias et al. 2022]. Regiões como a Ásia (75,94 kg de consumo per capita entre 2016–2018) e a América do Sul (29,05 kg no mesmo período) figuram entre as maiores consumidoras [Samal et al. 2022].

Apesar de sua importância, o cultivo de arroz é extremamente vulnerável às variações climáticas, já que fatores como temperatura, regime de chuvas e ocorrência de eventos extremos exercem influência direta sobre sua produtividade. Essa condição impõe a necessidade de identificar regiões mais suscetíveis e desenvolver estratégias de adaptação que aumentem a resiliência das lavouras, assegurando a continuidade do fornecimento global — um dos principais desafios diante das mudanças climáticas [Li et al. 2024].

Garantir o crescimento sustentável dessa produção requer estratégias integradas que combinem uso eficiente dos recursos, aumento da produtividade por área cultivada e preservação dos recursos naturais [Parry and Hawkesford 2010]. Entre as ferramentas disponíveis, a previsão de produtividade agrícola tem se mostrado essencial para apoiar políticas públicas e decisões no campo. Essa tarefa é complexa, pois depende de múltiplos fatores — solo, clima, manejo e variáveis ambientais — de difícil mensuração e modelagem. Tecnologias de aprendizado de máquina (*machine learning* - ML) ampliaram as possibilidades nesse campo, aumentando a acurácia das estimativas e sendo aplicadas com sucesso em estudos de previsão de safras, análise de solo e gestão agrícola em escala global [Satpathi et al. 2023].

O aprendizado de máquina, de forma geral, consiste em desenvolver sistemas capazes de aprender e melhorar seu desempenho a partir de dados, identificando padrões e relações que nem sempre são visíveis de forma direta. Essas técnicas podem ser classificadas principalmente em aprendizado supervisionado, no qual o modelo é treinado com exemplos rotulados para gerar previsões sobre novos dados (como redes neurais artificiais, máquinas de vetores de suporte, árvores de decisão e redes bayesianas), e aprendizado não supervisionado, que busca identificar padrões ocultos e agrupar dados sem rótulos pré-definidos (como algoritmos k-means, agrupamentos hierárquicos e mapas auto-organizados) [Mishra et al. 2016].

A aplicação de modelos preditivos na agricultura tem avançado de forma acelerada, impulsionada pelo grande volume de dados disponíveis de fontes climáticas, agronômicas e socioeconômicas [Satpathi et al. 2023]. A integração da ciência da computação com a produção agrícola permite transformar essas informações em conhecimento para previsão de safras, otimização do uso de insumos, detecção de pragas e doenças, e análise de solo. Esses modelos, ao integrarem múltiplos fatores que influenciam o rendimento, auxiliam agricultores e gestores na tomada de decisões estratégicas, favorecendo uma produção sustentável e explorando ao máximo o potencial produtivo —

aspecto particularmente relevante diante das mudanças climáticas globais. A previsão de produtividade agrícola é essencial para apoiar políticas públicas e o planejamento sustentável, especialmente na América do Sul [Khan and Khare 2025].

No contexto de Sistemas de Informação, o estudo insere-se na tradição de *Decision Support Systems* (DSS), ao transformar dados agroclimáticos em insumos para o planejamento e políticas públicas [Turban et al. 2010]. A perspectiva *sociotécnica* orienta o desenho do artefato, reconhecendo que a utilidade do modelo depende tanto de sua acurácia quanto da integração aos fluxos decisórios de produtores e instituições [Bostrom and Heinen 1977]. À luz da *Teoria da Contingência*, o estudo adapta suas variáveis e janelas de análise às especificidades sul-americanas — caracterizadas por heterogeneidade climática e variabilidade produtiva — reforçando o princípio de adequação ao contexto [Daft and Lengel 1986]. A agenda de *Information Systems for Sustainability* (IS4S) permeia a motivação central, ao associar predição agrícola à segurança alimentar e à gestão sustentável de recursos [Watson et al. 2010]. Por fim, a contribuição se alinha aos Grandes Desafios de Pesquisa em Sistemas de Informação no Brasil [GranDSI-BR 2016], conectando ciência de dados, sustentabilidade e impacto social.

2. Revisão da Literatura

Esta seção apresenta os principais estudos e abordagens relacionadas à previsão da produção agrícola com técnicas de aprendizado de máquina, com ênfase em modelos aplicados ao arroz e na incorporação de variáveis climáticas. São discutidas as contribuições metodológicas mais relevantes, as lacunas ainda existentes e as tendências recentes observadas na literatura, que fundamentam o delineamento metodológico adotado neste trabalho.

2.1. Modelos de aprendizado de máquina

Diversos estudos recentes apontam a superioridade de abordagens baseadas em árvores de decisão e *boosting* em relação a métodos lineares e a séries univariadas (como ARIMA) na previsão de produção/rendimento de grãos. Uma revisão metodológica sintetiza evidências de ganhos de acurácia, escalabilidade e generalização com algoritmos como XGBoost e Random Forests, especialmente em cenários com múltiplos preditores agroclimáticos e heterogeneidade espacial [Hamim et al. 2025]. Em aplicações nacionais, trabalhos que compararam XGBoost a modelos tradicionais relataram vantagem consistente do *boosting* na predição anual, inclusive superando ARIMA em séries de arroz de Bangladesh [Noorunnahar et al. 2023]. Em paralelo, revisões e estudos aplicados reforçam a adoção de XGBoost como *baseline* forte em agricultura de precisão, pela eficiência computacional e pelo bom equilíbrio entre viés e variância [Chen and Guestrin 2016, Ujjwal et al. 2023].

Apesar da relevância nos países da América do Sul, há uma lacuna na literatura de estudos de previsão da produção de arroz nestes países. [Satpathi et al. 2023] investigaram a previsão da produtividade de arroz em três distritos do estado de Chhattisgarh, na Índia, utilizando dados climáticos e de produção agrícola coletados entre 1999 e 2019. O problema foi tratado como uma tarefa de regressão, sendo testadas abordagens estatísticas (SMLR, LASSO, Ridge, Elastic Net) e de ML (rede neural artificial e modelos em conjunto, como Random Forest, Cubist e GLM). O desempenho dos modelos foi avaliado

por meio das métricas R^2 , RMSE, MAE, MBE, nRMSE e o coeficiente de concordância de Willmott. Os resultados mostraram que a rede neural artificial obteve os melhores desempenhos em Raipur e Surguja, com $R^2 = 1,00$ em ambos os conjuntos (calibração e validação) em Raipur, e R^2 de 1,00 e 0,99 em Surguja, com valores muito baixos de RMSE. Já no distrito de Bastar, os métodos LASSO e Elastic Net se destacaram, com R^2 de validação de 0,568 e 0,480, respectivamente. Além disso, os modelos de *ensemble*, especialmente a Random Forest, apresentaram desempenho superior em Bastar (R^2 de 0,85 na calibração e 0,81 na validação), evidenciando o potencial dessas técnicas para aumentar a precisão na previsão da produtividade de arroz.

2.2. Clima como covariável

Há um consenso crescente de que incluir variáveis climáticas (temperatura, precipitação, radiação) melhora a capacidade preditiva e a robustez temporal. Estudos recentes empregaram dados do NASA POWER para compor *features* meteorológicas e edáficas em modelos de arroz, documentando ganhos na explicabilidade e na performance preditiva [Rodrigues and Braga 2021]. Em escala global, análises mostraram relações sistemáticas entre variabilidade climática (temperatura e chuva) e produção de arroz, motivando o uso de indicadores agregados anuais e sazonais em modelos de ML [Neog 2022]. Técnicas de pré-processamento também têm recebido atenção: por exemplo, a remoção de tendências de longo prazo (*de-trending*) melhora significativamente a acurácia de modelos de rendimento agrícola via ML, ao estabilizar o aprendizado em diferentes regiões e períodos [Kumar et al. 2024]. No âmbito institucional, o emprego de CatBoost pela FAO (*Food and Agriculture Organization*) para previsão de produtividade ilustra a maturidade das famílias de *gradient boosting* em aplicações operacionais [FAO 2025].

3. Metodologia

A Figura 1 apresenta o fluxograma da metodologia adotada neste estudo, abrangendo as etapas de obtenção e pré-processamento dos dados, definição dos atributos explicativos, divisão em conjuntos de treino e teste/validação cruzada, treinamento dos modelos de regressão e avaliação com base nas métricas estabelecidas.



Figura 1. Fluxograma da metodologia. Fonte: Elaborado pelo autor.

3.1. Análise Exploratória da Base de Dados

Os dados utilizados neste estudo foram obtidos a partir da plataforma oficial FAOSTAT [FAOSTAT 2025], mantida pela Organização das Nações Unidas para Agricultura e Alimentação (FAO), que fornece estatísticas agrícolas globais atualizadas, organizadas por país e por ano. O foco recaiu sobre a produção de arroz na América do Sul, abrangendo o período de 2011 a 2020 para treinamento dos modelos e de 2021 a 2023 para validação.

Cada dataset refere-se a um ano específico e contém informações detalhadas sobre a produção agrícola, com granularidade anual e escopo por país. Os principais atributos considerados foram:

- **Area:** país onde ocorreu a produção (ex.: Brasil, Argentina, Uruguai);
- **Year:** ano de referência;
- **Production (t):** quantidade total de arroz produzido, em toneladas;
- **Area harvested (ha):** área total colhida, em hectares;
- **Yield (kg/ha):** produtividade média, em quilos por hectare colhido.

Tais atributos permitem modelar a dinâmica da produção ao longo do tempo, visto que a variável *Production* é função direta da área colhida e da produtividade.

A análise estatística preliminar revelou que a variável *Production* apresenta valores médios em torno de 2.045.192 toneladas, com forte assimetria à direita em razão da concentração da produção em grandes produtores (como Brasil e Argentina). A área colhida segue comportamento semelhante, enquanto a variável *Yield* se mostrou mais homogênea entre os países, indicando menor dispersão. A Tabela 1 resume essas estatísticas descritivas, apresentando os valores médios, mínimos e máximos dos principais atributos.

Tabela 1. Estatísticas descritivas dos atributos. Fonte: Elaborado pelo autor.

Atributo	Média	Mínimo	Máximo
Production (t)	2,045,192.62	94,672.50	13,476,994.00
Area harvested (ha)	394,613.19	20,937.00	2,872,036.00
Yield (kg/ha)	5,302.70	2,200.10	8,687.60

O exame de outliers identificou valores extremos em *Production (t)* e *Area harvested (ha)*. Esses valores, entretanto, correspondem a casos reais (como países com larga escala de produção) e, portanto, foram mantidos no conjunto de dados para preservar a representatividade da amostra.

Quanto a dados faltantes, não foram identificados registros ausentes, o que assegura a completude da base. O único atributo categórico é *Area*, representando o país de produção. Como a modelagem foi conduzida por país, não houve necessidade de codificação (*one-hot encoding*).

A análise de correlação indicou forte relação linear entre *Production* e *Area harvested* (coeficiente acima de 0,9), confirmando que o volume total de arroz está fortemente associado à extensão cultivada. Já a produtividade (*Yield*) apresentou correlação moderada com as demais variáveis, evidenciando que sua influência é complementar, mas não determinante em magnitude comparável à área colhida.

3.2. Modelos Preditivos

Para a análise comparativa de desempenho, foram empregados três modelos de aprendizado de máquina:

- **Linear Regression (LR)** – modelo base linear [Montgomery et al. 2012];
- **XGBoost Regressor (XGB)** – técnica baseada em árvores de gradiente [Chen and Guestrin 2016];
- **Support Vector Regression (SVR)** – modelo baseado em margens de máxima separação em espaço de alta dimensão [Smola and Schölkopf 2004].

A escolha desses algoritmos buscou equilibrar modelos lineares e não lineares, bem como comparar técnicas estatísticas clássicas com métodos mais recentes baseados em árvores de decisão e margens máximas.

A Regressão Linear foi utilizada como modelo de referência (*baseline*), servindo para avaliar se modelos mais sofisticados apresentaram ganhos expressivos em termos de erro percentual. O XGBoost Regressor foi selecionado por sua consistência em problemas de previsão com dados tabulares e pela capacidade de capturar interações complexas entre atributos. Já o SVR foi empregado para explorar a modelagem de relações não lineares, utilizando margens suaves e kernel radial.

Ajuste de Hiperparâmetros

No processo de treinamento, os hiperparâmetros foram definidos com base em valores reportados na literatura e em testes preliminares, sem a aplicação de otimização exaustiva como *Grid Search*. O foco esteve na viabilidade do uso de variáveis climáticas combinadas com séries históricas de produção de arroz, e não na busca da configuração ótima de cada modelo.

- **Regressão Linear:** empregada em sua forma padrão, sem regularizações adicionais.
- **XGBoost Regressor:** configurado com número fixo de árvores (`n_estimators`), taxa de aprendizado (`learning_rate`) e profundidade máxima (`max_depth`). Esses valores foram definidos empiricamente em testes iniciais.
- **Support Vector Regressor (SVR):** adotou-se o kernel radial (RBF) e normalização dos dados de entrada.

Os resultados foram bons, com erros baixos, e não foi necessária a busca sistemática de hiperparâmetros.

3.3. Métricas de Avaliação

A avaliação foi conduzida utilizando o Erro Percentual Absoluto Médio (MAPE) [Hyndman and Koehler 2006] como métrica principal de desempenho, complementado por validação cruzada (*k-fold*) para estimar a consistência dos modelos entre países e períodos de análise.

O desvio-padrão do erro foi adotado como indicador de estabilidade, refletindo a capacidade do modelo de manter variações reduzidas de desempenho entre diferentes

amostras de validação, mesmo em cenários com grande variação de escala produtiva. Valores mais baixos indicam maior estabilidade, enquanto valores elevados refletem maior sensibilidade e instabilidade.

Reconhece-se, contudo, que o MAPE apresenta limitações conhecidas, especialmente em cenários com valores reais reduzidos, sendo que métricas alternativas como RMSE, MAE e MASE podem fornecer informações complementares e são indicadas como extensões em trabalhos futuros.

3.4. Treinamento com Múltiplas Janelas Temporais

Foram avaliadas três janelas históricas de treinamento (5, 10 e 15 anos) com o objetivo de analisar o impacto da extensão do histórico de dados na estabilidade e na acurácia das previsões. Cada configuração foi utilizada para gerar estimativas anuais no período de 2019 a 2023, permitindo comparar o comportamento dos modelos em diferentes horizontes temporais.

3.5. Treinamento dos modelos nos principais produtores mundiais

Utilizando a mesma configuração e os mesmos hiperparâmetros definidos para os experimentos da América do Sul, os modelos foram novamente treinados e avaliados com os dados dos principais países produtores de arroz. O objetivo foi analisar a capacidade de generalização dos algoritmos em contextos produtivos distintos e verificar a consistência dos parâmetros adotados em diferentes blocos regionais.

A seleção desses países seguiu um critério objetivo, baseado no levantamento mais recente do *United States Department of Agriculture – Foreign Agricultural Service* (USDA-FAS) para a safra 2024/2025, que relaciona os maiores produtores mundiais de arroz segundo sua participação na produção global [USDA 2025]. Foram considerados os dez primeiros colocados, apresentados na Tabela 2.

Tabela 2. Maiores países produtores de arroz 2024/2025. Fonte: USDA (2025) – Traduzido.

País	% da Produção Global	Produção Total (t)
Índia	28%	150.000.000
China	27%	145.280.000
Bangladesh	7%	36.600.000
Indonésia	6%	34.600.000
Vietnã	5%	26.950.000
Tailândia	4%	20.550.000
Filipinas	2%	12.350.000
Mianmar	2%	11.900.000
Paquistão	2%	9.750.000
Camboja	2%	8.470.000

Os dados históricos anuais de produção desses países foram obtidos na base FA-OSTAT, no mesmo formato e estrutura utilizados para a América do Sul, garantindo consistência entre os conjuntos de dados e possibilitando comparações diretas de desempenho entre os blocos regionais.

3.6. Integração de variáveis climáticas

Para incorporar fatores climáticos à previsão da produção de arroz na América do Sul, utilizou-se a API do NASA POWER¹, uma base amplamente empregada em estudos agroclimáticos por disponibilizar variáveis atmosféricas e de radiação solar em resolução diária e mensal [Stackhouse et al. 2018].

O processo metodológico adotado compreendeu as seguintes etapas:

Coleta de dados climáticos: foram extraídas variáveis de precipitação e temperatura média anual, abrangendo o período de 2011 a 2023, de modo compatível com a série histórica de produção de arroz obtida na FAOSTAT.

Pré-processamento: aplicaram-se ajustes de normalização para compatibilizar as escalas entre os atributos de produção e as variáveis climáticas. Países com divergências de nomenclatura foram padronizados, e os conjuntos de dados foram integrados em um único *dataset* consolidado.

Treinamento dos modelos: adotou-se o período de 2011 a 2020 para treinamento e os anos de 2021, 2022 e 2023 como horizonte de previsão. Foram empregados os mesmos algoritmos utilizados nas etapas anteriores: Regressão Linear (LR), XGBoost Regressor (XGB) e *Support Vector Regression* (SVR).

Validação: além da comparação entre valores reais e estimados, foi aplicada validação cruzada *k-fold*, utilizando o Erro Percentual Absoluto Médio (MAPE) como métrica principal de avaliação [Hyndman and Koehler 2006].

4. Resultados e Discussão

Os resultados obtidos demonstram como a extensão histórica dos dados, as variações regionais e as variáveis climáticas influenciam o desempenho dos modelos de aprendizado de máquina na previsão da produção de arroz. A discussão destaca padrões de acurácia, estabilidade e generalização observados entre diferentes países e configurações de treinamento.

4.1. Desempenho por janela de treinamento

Foram comparados LR, XGB e SVR em três janelas históricas (15, 10 e 5 anos) com previsões anuais de 2019 a 2023. Observou-se que a Regressão Linear obteve melhor desempenho em séries com trajetória mais regular, como Argentina e Uruguai, enquanto o XGBoost se mostrou mais consistente a diferentes contextos, com resultados particularmente consistentes na janela intermediária de 10 anos. O SVR mostrou-se o mais sensível à extensão do histórico, apresentando maior instabilidade nas janelas curtas — caracterizada por forte variação do erro entre anos e países, conforme indicado pelo desvio-padrão mais elevado.

No XGBoost, em alguns países surgiram previsões idênticas entre anos consecutivos. Esse comportamento decorre do próprio mecanismo de árvores de gradiente quando as variáveis preditoras oferecem pouca variação marginal para novas divisões, levando a estimativas constantes próximas ao ótimo encontrado [Chen and Guestrin 2016,

¹NASA POWER API – Prediction of Worldwide Energy Resources. Documentação disponível em: <https://power.larc.nasa.gov/docs/services/api/>.

Hamim et al. 2025]. Tal efeito reduz responsividade a oscilações anuais, mas aumenta a estabilidade em contextos de baixa volatilidade (Figura 2).

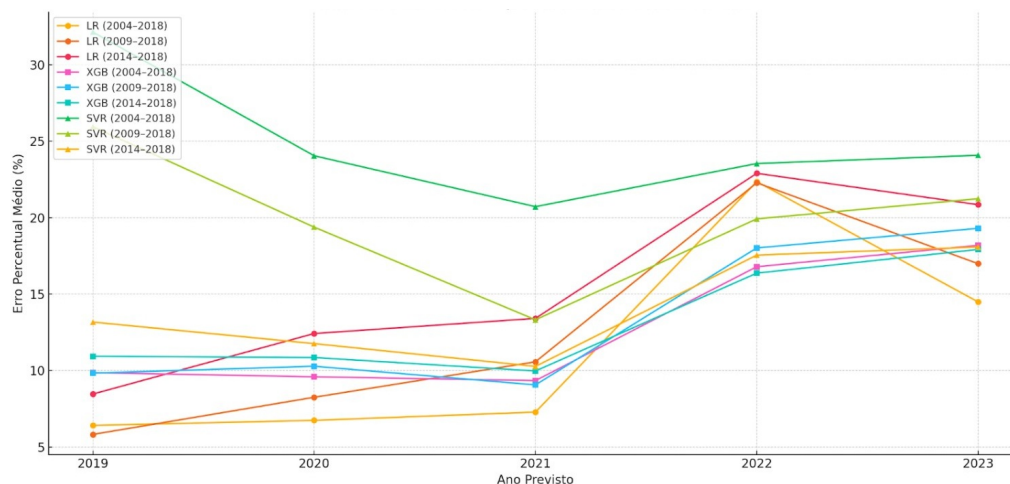


Figura 2. Erro percentual médio por ano previsto (2019-2023). Fonte: Elaborado pelo autor.

De forma geral, a janela intermediária de 10 anos apresentou o melhor equilíbrio entre volume histórico e relevância temporal, sendo a mais consistente para o XGBoost e mantendo desempenho competitivo também nos demais modelos.

4.2. Análise geral de desempenho dos modelos

A análise integrada dos resultados temporais, espaciais e de validação cruzada revelou padrões consistentes de comportamento entre os três algoritmos avaliados. Em termos temporais, observou-se aumento gradual do erro conforme a previsão se distanciava do período de treino, tendência comum a todos os modelos. O XGBoost manteve erros inferiores a 10% até 2021, degradando de forma controlada nos anos seguintes. A Regressão Linear apresentou crescimento mais acentuado do erro nos anos finais, enquanto o SVR mostrou maior instabilidade, com forte variação de desempenho entre anos e países, refletida nos elevados valores de desvio-padrão observados.

Sob a perspectiva espacial, a Regressão Linear apresentou melhores resultados em países com estabilidade histórica de produção, o XGBoost manteve desempenho consistente mesmo em contextos mais voláteis e o SVR foi o mais afetado por oscilações e limitação de dados, apresentando erros mais altos em países como Bolívia, Venezuela e Paraguai. Em alguns casos, o XGBoost produziu valores idênticos para anos consecutivos, fenômeno associado à baixa variação das variáveis preditoras, levando o algoritmo a estabilizar as estimativas próximas ao ponto ótimo encontrado durante o ajuste [Chen and Guestrin 2016, Hamim et al. 2025]. Embora tal comportamento reduza a sensibilidade a flutuações anuais sutis, ele também reforça a estabilidade do modelo em cenários de menor volatilidade, como ilustrado na Figura 3.

4.3. Validação cruzada

Os resultados da validação cruzada (5-fold) confirmaram essas tendências, reforçando a superioridade do XGBoost, que obteve o menor MAPE (12,47%) e o menor desvio-padrão (4%), evidenciando maior precisão e estabilidade entre diferentes amostras. A

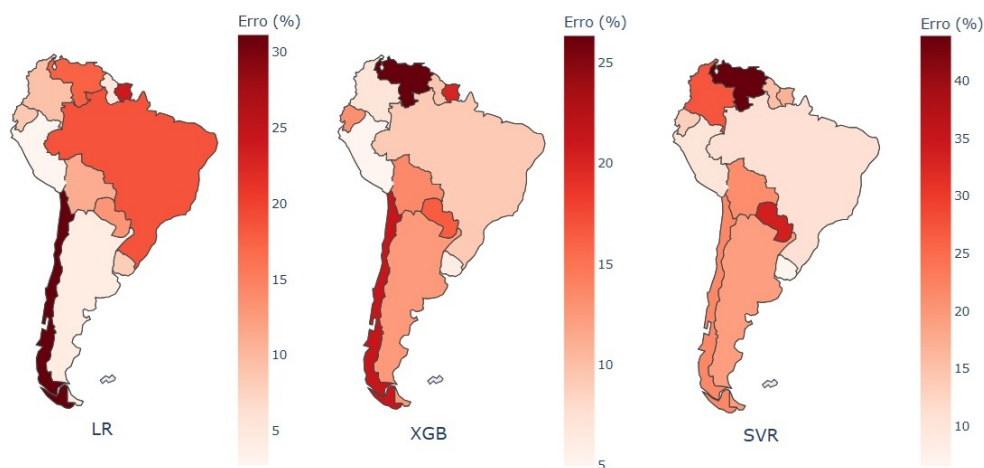


Figura 3. Erro percentual médio por país e por modelo (previsão 5 anos). Fonte: Elaborado pelo autor.

Regressão Linear apresentou desempenho satisfatório, especialmente como modelo-base, enquanto o SVR manteve comportamento mais sensível à disponibilidade e à variabilidade dos dados, apresentando erros mais elevados ao longo das rodadas de validação, conforme ilustrado na Figura 4.

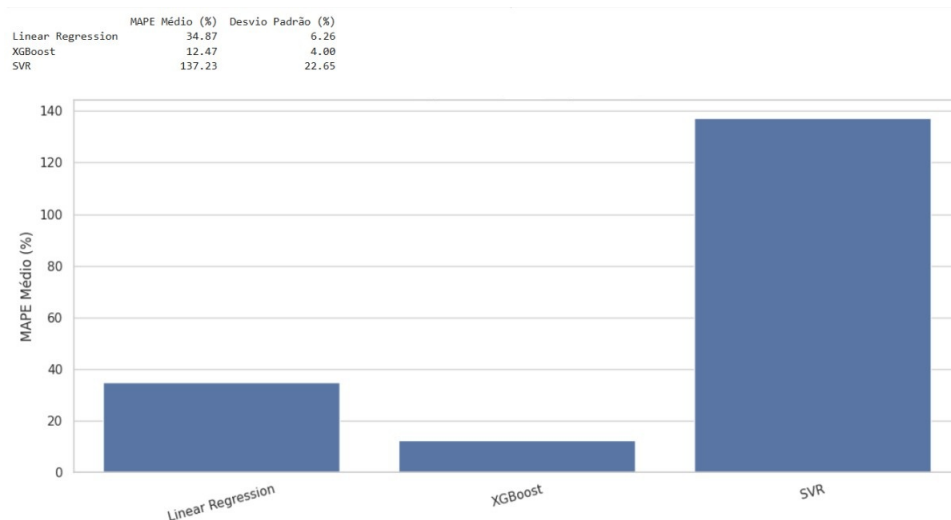


Figura 4. Validação cruzada (MAPE) por modelo. Fonte: Elaborado pelo autor.

4.4. Justificativa para a seleção da janela de 10 anos

Com base nos resultados comparativos, adotou-se a janela 2009–2018 como padrão nos experimentos subsequentes, por equilibrar abrangência histórica e atualidade e por apresentar menor degradação do erro nos anos previstos.

4.5. Desempenho do modelo otimizado para previsão de três anos

Com base na análise anterior, foi definido um cenário de previsão com horizonte de três anos para avaliar a capacidade dos algoritmos de manter desempenho consistente

em projeções de curto prazo, utilizando a produção de arroz na América do Sul como variável-alvo.

No comparativo geral, o XGBoost apresentou o melhor resultado, com MAPE médio de 13,72% e desvio-padrão de 4,35%, superando amplamente a Regressão Linear (MAPE 32,38%) e o SVR (MAPE 121,15%). Esse desempenho confirma a capacidade do XGBoost em capturar relações não lineares e interações entre variáveis, mantendo boa estabilidade entre diferentes anos e países, conforme ilustrado na Figura 5.

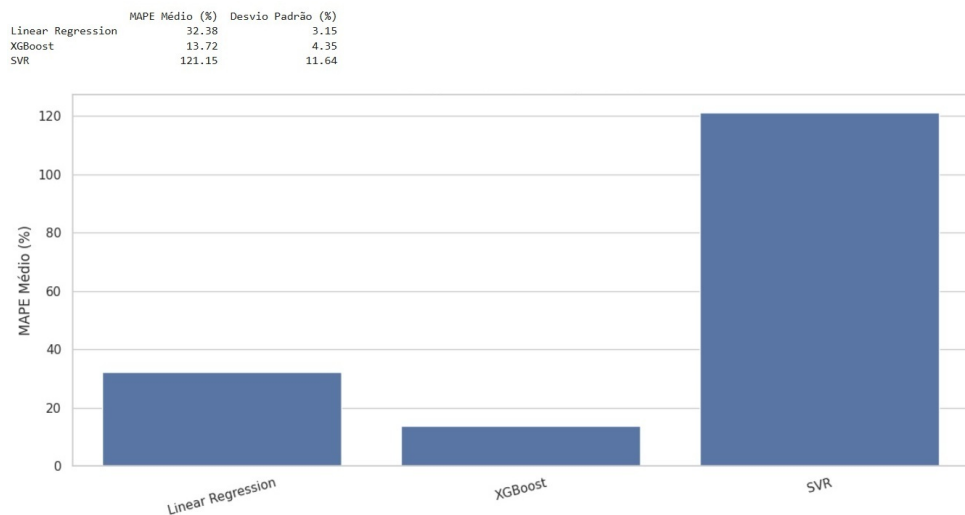


Figura 5. Comparativo do MAPE médio entre modelos — previsão de três anos (América do Sul). Fonte: Elaborado pelo autor.

A análise temporal mostrou que o erro médio do XGBoost manteve-se relativamente constante, variando de 6,12% em 2021 para 12,22% em 2023, evidenciando resiliência frente ao afastamento do período de treino. A Regressão Linear apresentou aumento mais acentuado de erro em 2022 (15,36%), antes de recuar para 9,86% em 2023, comportamento típico de modelos lineares sob séries não estacionárias. Já o SVR registrou desempenho mais irregular, com oscilações intensas e erros superiores a 17% em 2023, refletindo maior sensibilidade à escala e à dispersão das variáveis, conforme mostrado na Figura 6.

A análise por país reforça a consistência do XGBoost, com destaque para Venezuela (3,39%), Peru (3,94%) e Colômbia (4,15%). Em alguns casos, como Brasil, Paraguai e Suriname, as previsões mantiveram-se invariáveis ao longo dos três anos previstos — um comportamento característico do algoritmo quando a variação das variáveis preditoras é insuficiente para gerar novas divisões nas árvores de decisão [Chen and Guestrin 2016]. Embora isso reduza a sensibilidade a pequenas oscilações anuais, também evidencia a estabilidade do modelo em contextos de baixa volatilidade. A Regressão Linear apresentou maior variabilidade interanual, capturando flutuações específicas, mas com erro global mais elevado, enquanto o SVR permaneceu instável mesmo com normalização dos atributos (Figura 7).

Em síntese, o XGBoost mostrou-se o modelo mais equilibrado para o horizonte

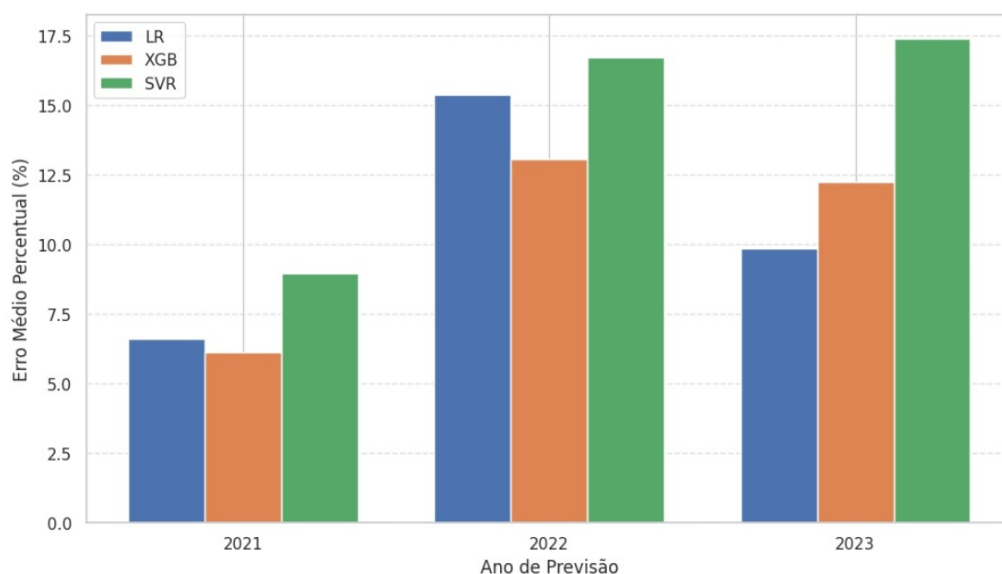


Figura 6. Erro percentual médio por modelo e ano. Fonte: Elaborado pelo autor.

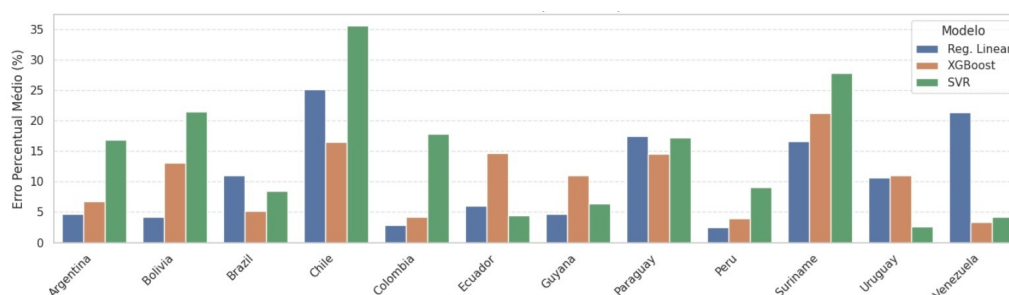


Figura 7. Erro percentual médio por país e por modelo. Fonte: Elaborado pelo autor.

de três anos, combinando boa acurácia média e menor dispersão de erros. Esses resultados reforçam sua adequação para cenários de previsão agrícola de curto prazo, nos quais é necessário conciliar estabilidade com capacidade de adaptação a diferentes contextos produtivos.

4.6. Análise do modelo aplicado aos maiores produtores mundiais de arroz

Usando a mesma configuração e os mesmos hiperparâmetros definidos para a América do Sul, os modelos foram novamente treinados com dados dos maiores produtores mundiais (USDA-FAS 2024/2025), representando cerca de 85% da produção global, a fim de avaliar o comportamento em um cenário de alta relevância produtiva.

O XGBoost apresentou, em diversos países, previsões idênticas entre 2021, 2022 e 2023, sem variação anual. Em termos anuais, a Regressão Linear registrou erros médios entre 2,46% e 3,82%; o XGBoost variou entre 4,68% e 6,41%; e o SVR superou 7%, chegando a 10,67% em 2023 (Figura 8).

Por país, o desempenho manteve-se consistente, com destaque para Bangladesh, China e Vietnã (erros inferiores a 2% na maioria dos modelos), e maiores erros para Paquistão, Camboja e Índia, especialmente no SVR (acima de 19% em alguns casos),

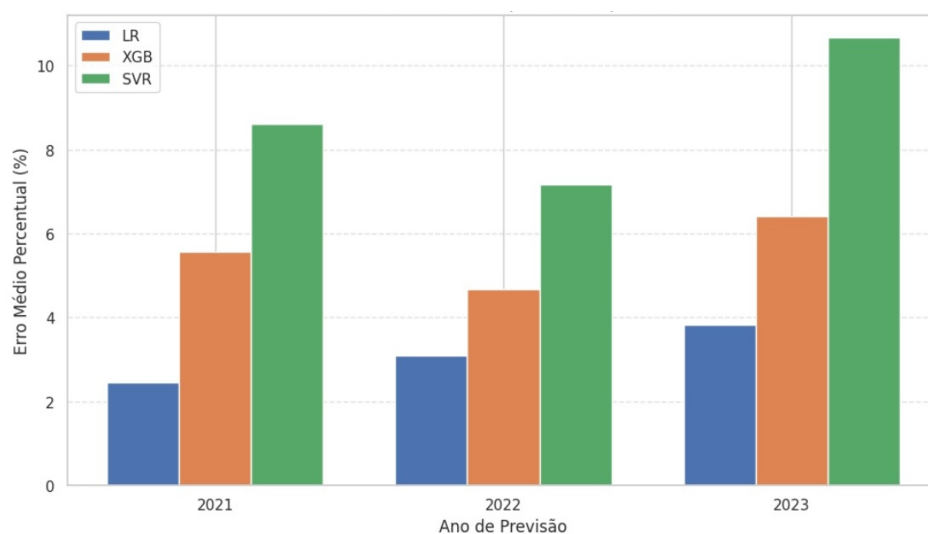


Figura 8. Erro percentual médio por modelo e por ano – Maiores Produtores.
Fonte: Elaborado pelo autor.

conforme mostrado na Figura 9.

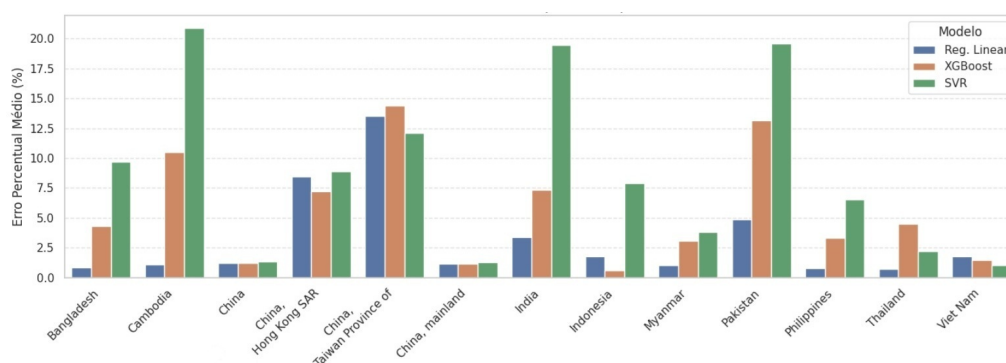


Figura 9. Erro percentual médio por país e por modelo – Maiores Produtores.
Fonte: Elaborado pelo autor.

Considerando a heterogeneidade de escalas e variância entre países, o MAPE mostrou-se pouco informativo em situações de valores reais muito baixos, pois pequenas diferenças absolutas geram grandes percentuais. Por isso, privilegiou-se aqui a leitura de erros percentuais médios por ano e por país, em complemento ao MAPE.

4.7. Comparação – América do Sul × Maiores Produtores

A comparação entre os dois blocos regionais revela diferenças importantes no comportamento dos modelos.

Estabilidade do modelo: nos maiores produtores, os erros foram, em geral, menores e mais estáveis, refletindo séries mais previsíveis.

Desempenho por modelo: o XGBoost manteve-se como abordagem mais equilibrada em ambos os contextos, embora tenha gerado valores idênticos de previsão em vários países entre 2021 e 2023. A Regressão Linear apresentou leve aumento de erro nos maiores

produtores, possivelmente pela presença de não linearidades, e o SVR teve desempenho inferior em cenários com forte oscilação.

Complexidade dos dados: a América do Sul apresentou maior variabilidade de erros entre países, sugerindo influência de fatores externos (climáticos, políticos, econômicos), enquanto os maiores produtores exibiram comportamento mais homogêneo.

Em síntese, as tendências gerais dos algoritmos se mantêm entre blocos, com o XGBoost mostrando robustez em ambientes de alta disponibilidade de dados e menor variabilidade anual (Figura 10).

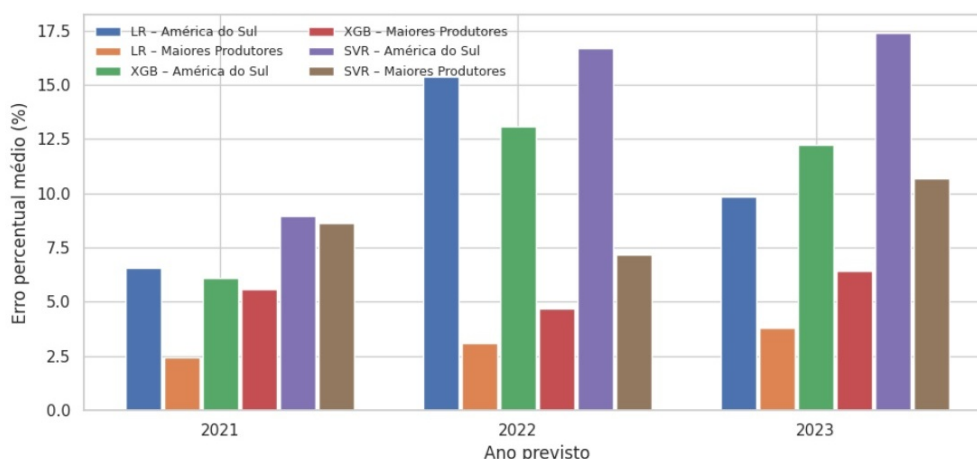


Figura 10. Comparação dos erros percentuais médios por modelo e ano – América do Sul vs. Maiores Produtores. Fonte: Elaborado pelo autor.

4.8. Cruzamento com dados climáticos

A inclusão das variáveis climáticas impactou os modelos de forma diferenciada.

O Linear Regression (LR) apresentou desempenho estável, com erros percentuais médios de 0,76% em 2021, 3,94% em 2022 e 1,69% em 2023. Esse resultado sugere que o modelo linear conseguiu capturar de forma eficiente a relação entre variáveis climáticas e produção, especialmente em anos com menor oscilação de fatores externos.

O XGBoost (XGB) obteve desempenho competitivo em 2021 (4,37%), mas degradou em 2022 e 2023 ($\approx 11\%$), indicando maior sensibilidade a variações abruptas nos dados climáticos.

O SVR manteve-se instável, com erros entre 9% e 17%, confirmando limitações do modelo mesmo com variáveis climáticas adicionais, como evidenciado na Figura 11.

Em relação aos países, Chile e Suriname apresentaram os maiores erros médios em XGB ($\approx 20\%$ e 16%) e SVR ($\approx 35\%$ e 28%), possivelmente devido ao menor volume de produção e maior sensibilidade a fatores climáticos locais, o que aumenta a variabilidade relativa. Brasil e Argentina destacaram-se pela estabilidade, com erros inferiores a 2% no LR e até 6% no XGB. Paraguai e Bolívia tiveram erros significativamente maiores nos modelos mais complexos, reforçando que para séries curtas e com forte oscilação anual, modelos lineares tendem a ser mais sólidos, como mostrado na Figura 12.

Já na avaliação cruzada o XGB apresentou MAPE médio de 11,3%, com desvio padrão de 3,1%, confirmando-se como o modelo mais equilibrado entre desempenho e

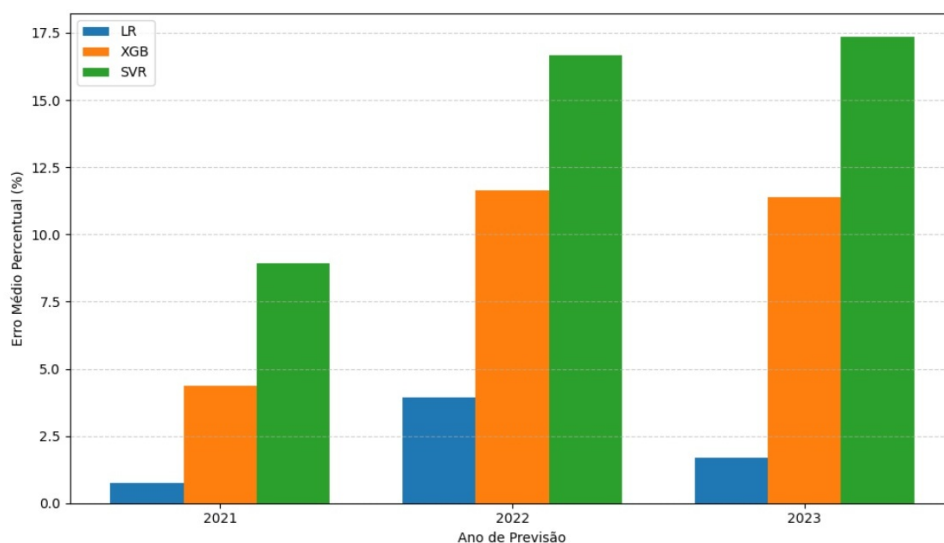


Figura 11. Erro percentual médio por modelo e por ano - Dados Climáticos.
Fonte: Elaborado pelo autor.

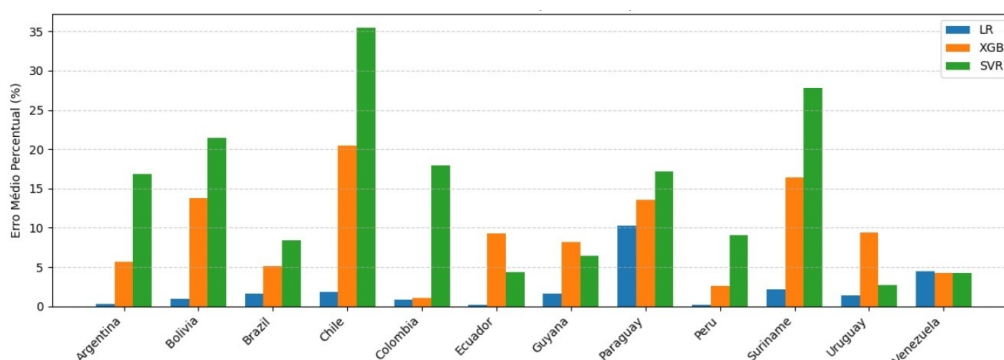


Figura 12. Erro percentual médio por país e por modelo - Dados Climáticos.
Fonte: Elaborado pelo autor.

acurácia. O LR manteve alta consistência (desvio de 2,2%), mas com MAPE médio elevado (34,6%) ao longo da validação cruzada, mostrando limitação estrutural quando extrapolado para diferentes *folds*. O SVR reafirmou desempenho inconsistente, com MAPE superior a 120% (ver Figura 13).

A integração de variáveis climáticas permitiu capturar melhor as oscilações interanuais na produção de arroz, especialmente em países de grande escala como Brasil e Argentina. Entretanto, em países com séries mais curtas ou produção instável (Chile, Suriname, Paraguai), os modelos apresentaram maior erro, sugerindo a necessidade de estratégias complementares, como: Inclusão de variáveis socioeconômicas (ex.: área colhida, investimentos em irrigação); Adoção de modelos híbridos (ML + ARIMA); Segmentação por clusters climáticos e produtivos antes do treinamento [Küçüktopcu et al. 2023].

5. Conclusão

O presente estudo investigou a aplicação de modelos estatísticos e de aprendizado de máquina para previsão da produção de arroz na América do Sul, considerando diferentes

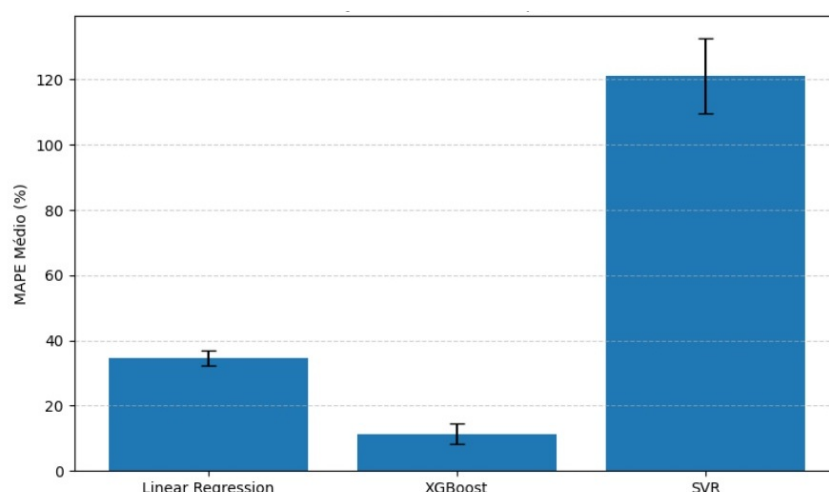


Figura 13. Validação cruzada (MAPE) por modelo - Dados Climáticos. Fonte: Elaborado pelo autor.

janelas temporais de treinamento e a integração de variáveis climáticas. Os resultados obtidos demonstraram que o desempenho dos modelos é altamente sensível tanto ao tamanho da janela histórica utilizada quanto à complexidade dos dados regionais.

Nos experimentos iniciais, verificou-se que o treinamento com apenas cinco anos de dados produziu resultados instáveis e inconsistentes, sobretudo em países com menor área cultivada ou maior variabilidade na série histórica. Esse achado reforça a limitação de janelas curtas, que não capturam adequadamente a sazonalidade agrícola e os choques externos (climáticos, econômicos e políticos). Em contrapartida, janelas mais longas, como a de 2009–2018, revelaram-se mais equilibradas, conciliando capacidade preditiva com menor tendência a sobreajuste. Esse resultado foi corroborado por diferentes métricas de avaliação, destacando a importância da quantidade e da qualidade das observações históricas para o aprendizado dos modelos.

A comparação entre algoritmos indicou o XGBoost como a abordagem com o melhor desempenho, apresentando menor erro percentual médio e maior estabilidade ao longo dos anos previstos. Em cenários com alta variabilidade ou dados mais ruidosos, a Regressão Linear mostrou desempenho satisfatório, especialmente em países com séries temporais regulares, ainda que limitada na captura de padrões não lineares. Já o SVR apresentou desempenho inferior, com erros elevados em diversos países, evidenciando maior sensibilidade à amplitude das séries e à heterogeneidade dos dados regionais.

Outro aspecto relevante foi a análise comparativa entre a América do Sul e os maiores produtores mundiais de arroz. Enquanto estes últimos apresentaram séries mais estáveis e previsíveis, refletindo cadeias agrícolas estruturadas, os países sul-americanos evidenciaram maior oscilação anual, reforçando o papel de fatores externos e a necessidade de modelos adaptativos mais sofisticados. Nessa comparação, observou-se que o XGBoost manteve baixa variação anual de erro nos dois blocos regionais, consolidando-se como a técnica mais promissora para cenários de produção agrícola.

A etapa de integração de variáveis climáticas, com dados extraídos via API NASA POWER, trouxe um avanço metodológico relevante. Ao incorporar indicadores como

precipitação e temperatura média anual, os modelos puderam capturar relações mais complexas entre fatores ambientais e produtividade agrícola. Essa integração resultou em ganhos de acurácia em diversos países, especialmente naqueles com forte dependência de condições meteorológicas sazonais. Contudo, os resultados também indicaram que a influência das variáveis climáticas não é homogênea: países com maior estabilidade produtiva apresentaram menor impacto das variáveis adicionais, ao passo que países com elevada variabilidade climática tiveram ganhos expressivos de explicabilidade e redução de erro. A literatura aponta que a produção de arroz é altamente sensível à variabilidade climática, com aumentos de temperatura e irregularidade de precipitação associados a reduções significativas de produtividade [Neog 2022].

Em termos de métricas, a validação cruzada reforçou a robustez do XGBoost (MAPE médio $\approx 11,3\%$), superando amplamente a Regressão Linear ($\approx 34,6\%$) e o SVR ($\approx 121\%$). Esses achados confirmam a superioridade de modelos baseados em árvores de decisão e boosting para problemas de previsão agrícola em séries temporais multivariadas. Além disso, observou-se que a acurácia dos modelos tende a se degradar após três anos de previsão, sugerindo que horizontes mais longos exigem maior cautela na interpretação dos resultados.

Por fim, este trabalho contribui metodologicamente ao demonstrar que a combinação de séries históricas ampliadas, técnicas de aprendizado de máquina e integração de dados climáticos constitui uma estratégia eficaz para previsão da produção agrícola. Os resultados obtidos oferecem subsídios valiosos para a formulação de políticas públicas, o planejamento setorial e a gestão de cadeias produtivas, ao mesmo tempo em que abrem espaço para investigações futuras, como a inclusão de variáveis socioeconômicas e o uso de arquiteturas mais avançadas, como redes neurais profundas.

Em síntese, conclui-se que a previsão da produção de arroz na América do Sul pode alcançar níveis satisfatórios de acurácia, desde que respaldada por séries históricas adequadas, pela integração de variáveis climáticas e pela escolha criteriosa de modelos, sendo o XGBoost a abordagem mais recomendada neste contexto.

A partir da ótica de Sistemas de Informação, os resultados reforçam que abordagens de ML podem compor artefatos de *Decision Support Systems* (DSS) eficazes quando fundamentadas em princípios *sociotécnicos* e de *contingência*. O modelo XGBoost, por exemplo, demonstra um bom desempenho técnico, mas sua adoção prática depende da integração com processos institucionais, infraestrutura e governança de dados. Além disso, a incorporação de variáveis climáticas fortalece a perspectiva de *Information Systems for Sustainability* (IS4S), ao fornecer subsídios analíticos para mitigação de riscos e políticas de segurança alimentar. A disseminação desse tipo de artefato em órgãos públicos e cooperativas tende a seguir determinantes previstos por modelos de aceitação tecnológica (TAM/UTAUT), como utilidade percebida, facilidade de uso e condições facilitadoras. Assim, o estudo não apenas apresenta um modelo preditivo eficiente, mas também contribui com conhecimento de design para DSS agrícolas, endereçando o GrandSI-BR 2016–2026 ao aproximar ciência de dados, decisão pública e sustentabilidade socioeconômica.

6. Trabalhos Futuros

A pesquisa realizada permitiu demonstrar a viabilidade da modelagem preditiva da produção de arroz na América do Sul, sobretudo com a utilização do algoritmo XGBoost e da integração de variáveis climáticas. No entanto, algumas perspectivas de aprofundamento se destacam como caminhos relevantes para trabalhos futuros:

Integração de variáveis climáticas mais complexas – Além de temperatura média e precipitação, pode-se ampliar a base climática para incluir variáveis como índices de seca (ex.: SPI, SPEI), frequência de eventos extremos e dados de anomalias climáticas (ENSO, La Niña). Essa ampliação poderá aumentar a sensibilidade do modelo a fatores ambientais críticos para a produtividade agrícola.

Exploração de modelos híbridos e redes neurais – Embora o XGBoost tenha se mostrado consistente, estudos futuros podem testar arquiteturas mais avançadas, como LSTM, GRU ou modelos híbridos que combinem regressão estatística, aprendizado de máquina e séries temporais. Isso possibilitaria capturar relações dinâmicas não lineares mais profundas entre produção e clima. Modelos híbridos, como a combinação de redes temporais convolucionais (TCN) com XGBoost, vêm sendo utilizados para reduzir erros em séries temporais agrícolas [Zhao et al. 2025].

Cruzamento com dados socioeconômicos e de políticas agrícolas – Outro vetor de avanço é integrar variáveis relacionadas a políticas públicas, subsídios, infraestrutura de irrigação, acesso a insumos e práticas de manejo. Essas informações permitiriam avaliar não apenas a influência climática, mas também o impacto de fatores estruturais no desempenho produtivo.

Construção de um sistema de suporte à decisão – A partir dos modelos desenvolvidos, é possível projetar uma plataforma interativa que ofereça previsões atualizadas para produtores, órgãos governamentais e entidades de planejamento agrícola, auxiliando na decisão estratégica e na antecipação de riscos relacionados à segurança alimentar.

Conformidade com Ciência Aberta

Os modelos, códigos e conjuntos de dados utilizados neste estudo estão disponíveis publicamente em: <https://github.com/rgrion/mlarrozprevisao>. Essa disponibilização garante transparência, replicabilidade e reprodutibilidade, permitindo que outros pesquisadores possam replicar os experimentos, validar os resultados e propor avanços metodológicos a partir da mesma base empregada.

Declaração de uso de IA Generativa

Ferramentas de Inteligência Artificial Generativa, como o ChatGPT, foram utilizadas apenas para apoiar a revisão textual e a organização de referências. Todo o conteúdo científico, incluindo definição de metodologia, experimentos, análises e conclusões, foi elaborado integralmente pelas autoras e autores, que assumem total responsabilidade.

Agradecimentos

Este estudo contou com financiamento parcial da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e da Universidade Federal do Estado do Rio de Janeiro (UNIRIO).

Referências

- Albahri, G., Alyamani, A. A., Badran, A., Hijazi, A., Nasser, M., Maresca, M., and Baydoun, E. (2023). Enhancing essential grains yield for sustainable food security and bio-safe agriculture through latest innovative approaches. *Agronomy*, 13(7):1709.
- Bostrom, R. P. and Heinen, J. S. (1977). Mis problems and failures: A socio-technical perspective. part i: The causes. *MIS quarterly*, pages 17–32.
- Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.
- Daft, R. L. and Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management science*, 32(5):554–571.
- Dias, L. G., Hacke, A., dos Santos Souza, E., Nath, S., Canesin, M. R., Vilella, O. V., Geloneze, B., Pallone, J. A. L., Cazarin, C. B. B., Blakeslee, J. J., Mariutti, L. R. B., and Bragagnolo, N. (2022). Comparison of chemical and nutritional compositions between aromatic and non-aromatic rice from brazil and effect of planting time on bioactive compounds. *Journal of Food Composition and Analysis*, 111:104608.
- FAO (2025). Yield forecasting with catboost: Fao pilot project on machine learning for agriculture. Technical report, Food and Agriculture Organization of the United Nations, Rome.
- FAOSTAT (2025). Crops and livestock products.
[urlhttps://www.fao.org/faostat](https://www.fao.org/faostat). Accessed: 2025-09-25.
- GranDSI-BR (2016). Grandes desafios de pesquisa em sistemas de informação no brasil (2016–2026). Technical report, SBC – Sociedade Brasileira de Computação.
- Hamim, A. A. I., Mohaimin, A., Ishmum, A. F., Ifty, R. A., and Patwary, M. J. A. (2025). Advances in machine learning for crop yield prediction: A comprehensive review of techniques, trends, and challenges. pages 1–6.
- Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688.
- Khan, I. and Khare, B. K. (2025). Integrating ai in plant science: A review of applications and future prospects. *Plant Gene*, page 100542.
- Kumar, A., Kumar, A., Singh, R. P., Kumar, P., and Singh, P. V. (2024). A comparative study of detrending methods on crop yield time series for drought studies. *Asian Research Journal of Agriculture*, 17(3):191–204.
- Küçüktopcu, E., Cemek, E., Cemek, B., and Simsek, H. (2023). Hybrid statistical and machine learning methods for daily evapotranspiration modeling. *Sustainability*, 15(7).
- Li, S., Wu, F., Zhou, Q., and Zhang, Y. (2024). Adopting agronomic strategies to enhance the adaptation of global rice production to future climate change: a meta-analysis. *Agronomy for Sustainable Development*, 44(3).
- Mishra, S., Mishra, D., and Santra, G. H. (2016). Applications of machine learning techniques in agricultural crop production: A review paper. *Indian Journal of Science and Technology*, 9(38).

- Montgomery, D. C., Peck, E. A., and Vining, G. G. (2012). *Introduction to Linear Regression Analysis*. Wiley, New Jersey, 5th edition.
- Neog, B. J. (2022). Temperature shocks and rural labour markets: evidence from india. *Climatic Change*, 171(16).
- Noorunnahar, M., Chowdhury, A. H., and Mila, F. A. (2023). A tree based extreme gradient boosting (xgboost) machine learning model to forecast the annual rice production in bangladesh. *PLOS ONE*, 18(3):e0283452.
- Parry, M. A. J. and Hawkesford, M. J. (2010). Food security: increasing yield and improving resource use efficiency. *Proceedings of The Nutrition Society*, 69(4):592–600.
- Ray, D. K., Mueller, N. D., West, P. C., and Foley, J. A. (2013). Yield trends are insufficient to double global crop production by 2050. *PLOS ONE*, 8(6):1–8.
- Rodrigues, G. C. and Braga, R. P. (2021). Evaluation of nasa power reanalysis products to estimate daily weather variables in a hot summer mediterranean climate. *Agronomy*, 11(6):1207.
- Samal, P., Babu, S. C., Mondal, B., and Mishra, S. N. (2022). The global rice agriculture towards 2050: An inter-continental perspective. *Outlook on Agriculture*, 51(2):164–172.
- Satpathi, A., Setiya, P., Das, B., Nain, A. S., Jha, P. K., Singh, S., and Singh, S. (2023). Comparative analysis of statistical and machine learning techniques for rice yield forecasting for chhattisgarh, india. *Sustainability*, 15(3):2786.
- Smola, A. J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222.
- Stackhouse, P. W. J., Whitlock, C. H., Chandler, W. S., Hoell, J. M., Zhang, T., Westberg, D., Hoell, A., Macpherson, B., and Mikovitz, J. C. (2018). Power release: Nasa prediction of worldwide energy resources. Technical report, NASA.
- Turban, E., Sharda, R., and Delen, D. (2010). *Decision Support and Business Intelligence Systems*. Prentice Hall Press, USA, 9th edition.
- Ujjwal, T. R., Murty, B. N. S., Lenka, N. K., and Lal, S. B. (2023). Comparative analysis of statistical and machine learning techniques for rice yield forecasting for chhattisgarh, india. *Sustainability*, 15(3):2786.
- USDA (2025). Production, supply and distribution (psd) online.
url<https://www.fas.usda.gov/data/production/commodity/0422110>. Accessed: 2025-09-25.
- Watson, R. T., Boudreau, M.-C., and Chen, A. J. (2010). Information systems and environmentally sustainable development: energy informatics and new directions for the community. *MIS quarterly*, pages 23–38.
- Zhao, T., Chen, G., Suraphee, S., Phoophiwfa, T., and Busababodhin, P. (2025). A hybrid tcn-xgboost model for agricultural product market price forecasting. *PLOS ONE*, 20.