

XGBoost e a Prática Orçamentária do MAPA: Influência no IDH e no PIB do Agronegócio

Jorge Castro

ORCID: [0000-0003-0612-4576](https://orcid.org/0000-0003-0612-4576)

Disponível em: https://github.com/Jotta-se/MBA_USP

RESUMO

A alocação eficiente de recursos públicos representa um desafio estrutural para governos em todo o mundo, especialmente em setores estratégicos como o agropecuário. O planejamento orçamentário do Ministério da Agricultura, Pecuária e Abastecimento (MAPA) carece de instrumentos preditivos capazes de identificar quais variáveis orçamentárias, socioeconômicas e regionais exercem maior influência sobre o desenvolvimento humano e o desempenho do agronegócio. Este estudo aplica o algoritmo *XGBoost*, com otimização de hiperparâmetros via *GridSearchCV*, a um *dataset* de 465 observações oriundas de fontes abertas do governo federal, do Censo Agropecuário de 2017 (IBGE) e do CEPEA/USP, cobrindo o período entre 2018 e 2022. São modeladas duas variáveis-alvo, a variação percentual do PIB do Agronegócio (o %PIB-AGRO) e o Índice de Desenvolvimento Humano (IDH) das unidades federativas brasileiras. A seleção de *features* empregou *Recursive Feature Elimination* (RFE) e análise de multicolinearidade. O modelo para o percentual do PIB-AGRO alcançou o Coeficiente de Determinação (R^2) igual a 0,8764 e *Root Mean Squared Error* (RMSE) igual a 0,00956, enquanto o modelo para IDH obteve R^2 igual a 0,9632 e RMSE igual a 0,00712. Em ambos os alvos, o percentual de propriedades rurais com acesso a orientação técnica emergiu como a variável de maior importância relativa (28,18% e 52,11%, respectivamente), seguida, no modelo do PIB, pelos investimentos em Projetos (análogo ao CAPEX, 21,59%), e, no modelo do IDH, pela localização geográfica (região Sudeste, 31,89%). Os resultados oferecem evidências empíricas para a formulação de políticas públicas *data-driven*, demonstrando que a extensão rural e os investimentos de capital são determinantes críticos da competitividade do agronegócio brasileiro e do bem-estar das populações rurais.

Palavras-chave: Orçamento público; Aprendizado de máquina; *Extreme Gradient Boosting*; *XGBoost*; Agronegócio; Desenvolvimento humano; IDH; PIB Agronegócio; MAPA; Políticas públicas.

ABSTRACT

The efficient allocation of public resources represents a structural challenge for governments worldwide, particularly in strategic sectors such as agribusiness. The budgetary planning of Brazil's Ministry of Agriculture, Livestock and Food Supply (MAPA) lacks predictive tools capable of identifying which budgetary, socioeconomic, and regional variables most influence human development and agribusiness performance. This study applies the *XGBoost* algorithm,

with hyperparameter optimization via GridSearchCV, to a dataset of 465 observations drawn from federal government open data, the 2017 Agricultural Census (IBGE), and CEPEA/USP, covering the period from 2018 to 2022. Two target variables are modeled, the percentage variation of the Agribusiness GDP (%GDP-AGRO) and the Human Development Index (HDI) of Brazilian states. Feature selection employed Recursive Feature Elimination (RFE) and multicollinearity analysis. The %GDP-AGRO model achieved a Coefficient of Determination (R^2) of 0.8764 and a Root Mean Squared Error (RMSE) of 0.00956, while the HDI model obtained an R^2 of 0.9632 and an RMSE of 0.00712. In both targets, the percentage of rural properties with access to technical assistance emerged as the most important feature (28.18% and 52.11%, respectively), followed, in the GDP model, by investments in Projects (analogous to CAPEX, 21.59%), and, in the HDI model, by geographic location (Southeast region, 31.89%). The results provide empirical evidence for data-driven public policy formulation, demonstrating that rural extension services and capital investments are critical determinants of Brazilian agribusiness competitiveness and rural populations' well-being.

Keywords: Public budget; Machine learning; Extreme Gradient Boosting; XGBoost; Agricultural sector; Human development; HDI; GDP Agribusiness; MAPA; Public policy.

1. INTRODUÇÃO

A gestão eficiente de recursos públicos é um desafio global que se intensifica à medida que governos precisam conciliar pressões fiscais crescentes com demandas sociais cada vez mais complexas [1]. No contexto brasileiro, o setor agropecuário ocupa posição estratégica. Em 2022, o agronegócio respondeu por aproximadamente 24,8% do Produto Interno Bruto nacional, consolidando o Brasil como um dos maiores produtores e exportadores agrícolas do mundo [2]. Apesar desse protagonismo econômico, a tradução de investimentos públicos em ganhos mensuráveis de produtividade, competitividade e desenvolvimento humano permanece insuficientemente compreendida na literatura nacional.

O Ministério da Agricultura, Pecuária e Abastecimento (MAPA) é o principal vetor institucional dessas políticas, coordenando entidades como EMBRAPA, INCRA, CONAB e SFB. Entre 2018 e 2022, o orçamento total autorizado do ministério variou de R\$ 4,04 bilhões a R\$ 7,23 bilhões anuais, mas a taxa média de conversão de empenhos em pagamentos efetivos permaneceu em 47,9% [3]. Essa lacuna crônica entre o planejado e o executado revela não apenas ineficiências operacionais, mas também a ausência de ferramentas analíticas que identifiquem quais categorias de despesa geram maior impacto nos resultados pretendidos.

Nesse contexto, o avanço das técnicas de aprendizado de máquina, em especial o *Extreme Gradient Boosting* (XGBoost), oferece uma oportunidade concreta para preencher essa lacuna. O XGBoost destaca-se por sua robustez a multicolinearidade, capacidade de capturar relações não lineares e eficiência computacional em *datasets* de dimensão moderada [4, 5]. Sua aplicação a dados orçamentários e socioeconômicos permite identificar, de forma objetiva e reproduzível, quais variáveis exercem maior influência sobre os resultados de políticas públicas.

Este artigo apresenta as seguintes contribuições originais:

- (i) Aplicação do *XGBoost* para modelagem preditiva de dois alvos distintos, percentual do PIB-AGRO e IDH, a partir de dados orçamentários abertos do MAPA (2018–2022), integrando variáveis do Censo Agropecuário de 2017 e da série histórica CEPEA/CNA;
- (ii) Mapeamento quantitativo da importância relativa de *features* orçamentárias, socioeconômicas e regionais, revelando a centralidade da orientação técnica rural como determinante de ambos os alvos;
- (iii) Evidências empíricas para a formulação de políticas públicas *data-driven* no setor agropecuário, com implicações diretas para a priorização do orçamento federal;
- (iv) Disponibilização pública do código-fonte e dos *datasets* em repositório *GitHub* [6], garantindo transparência e replicabilidade.

As *Research Questions* (RQs) que orientam o estudo são:

RQ1. Quais variáveis orçamentárias, socioeconômicas e regionais exercem maior influência sobre o percentual do PIB do Agronegócio no Brasil entre 2018 e 2022?

RQ2. Em que medida essas mesmas variáveis explicam a variação do IDH das unidades federativas?

RQ3. A aplicação do *XGBoost* produz poder explicativo suficiente para subsidiar decisões de alocação orçamentária no setor agropecuário?

O artigo está organizado da seguinte forma. A Seção 2 apresenta o referencial teórico e o estado da arte. A Seção 3 descreve os dados e a metodologia. A Seção 4 reporta os resultados. A Seção 5 discute as implicações. A Seção 6 aborda as ameaças à validade. A Seção 7 conclui o trabalho.

2. REFERENCIAL TEÓRICO E ESTADO DA ARTE

2.1 *Aprendizado de máquina em orçamento público e gestão financeira*

A aplicação de técnicas de inteligência artificial à gestão de recursos públicos tem sido crescentemente documentada na literatura internacional. Haytham e Elmousalami [7] comparam diferentes técnicas de IA para estimativa de custos conceituais em projetos, evidenciando o potencial de métodos de aprendizado de máquina como alternativa a abordagens paramétricas tradicionais. Puhach [8] aplica econometria de *machine learning* à simulação de fatores que influenciam despesas orçamentárias locais, evidenciando que o *gradient boosting* captura relações não lineares que escapam aos modelos de regressão clássica. Capone et al. [4] confirmam o potencial do *XGBoost* para previsão de despesas governamentais, ressaltando sua capacidade de processar grandes volumes de dados heterogêneos. Zainab et al. [9] desenvolvem modelo integrado de inteligência de dados para

estimativa de custos em construção, demonstrando que a combinação de técnicas estatísticas clássicas com modelos preditivos avançados é superior a qualquer abordagem isolada.

2.2 Investimento público, produtividade agrícola e crescimento econômico

A relação entre investimento público no setor agropecuário e crescimento econômico é bem estabelecida na literatura. Torres [10] demonstra que os investimentos em pesquisa e desenvolvimento são vitais para elevar a produtividade agrícola, com retornos superiores aos de outras categorias de gasto público. Diallo e Wouterse [11] documentam, para economias africanas, que investimentos em infraestrutura rural e extensão técnica são os maiores catalisadores do PIB agrícola e da redução da pobreza rural. No Brasil, Corcioli, Medina e Arrais [12] revelam que 75% do crédito operacional governamental destina-se a produtores de soja, milho e gado, sendo que 76,1% da cadeia produtiva da soja é controlada por multinacionais estrangeiras, o que sugere que o modelo atual de alocação pode não maximizar o retorno social dos recursos investidos.

2.3 Agronegócio, IDH e disparidades regionais

A relação entre desenvolvimento agrícola e Índice de Desenvolvimento Humano é multidimensional e geograficamente heterogênea. Piras, Wesz e Ghinoi [13] observam que a expansão da soja está associada a melhorias no IDH em municípios brasileiros, embora os benefícios sejam mais pronunciados na dimensão de renda do que em educação e longevidade. Soleman et al. [14] demonstram, para o contexto indonésio, que o IDH exerce impacto positivo e significativo sobre o PIB, sugerindo uma relação recíproca entre desenvolvimento humano e crescimento econômico. Ondetti [15] aponta que a concentração fundiária histórica no Brasil produz efeitos negativos de longo prazo sobre o IDH, particularmente nas dimensões de longevidade e renda. Tahu et al. [16] enfatizam o papel das despesas de capital como variável interveniente entre receitas regionais e crescimento econômico.

2.4 Posicionamento do presente estudo

A Tabela 1 sintetiza o posicionamento do presente estudo em relação às três linhas de pesquisa identificadas. A primeira linha abrange estudos sobre aprendizado de máquina (ML) aplicado ao orçamento público [4, 7, 8, 9], nos quais técnicas preditivas são aplicadas a dados orçamentários sem, contudo, conectar suas *features* a variáveis de impacto socioeconômico como IDH ou PIB setorial. A segunda linha reúne trabalhos sobre investimento agrícola e crescimento econômico [10, 11, 12], cujo foco recai sobre os efeitos de políticas de crédito e gasto público na produtividade rural, predominantemente por meio de econometria clássica e sem uso de modelos preditivos de ML. A terceira linha concentra estudos sobre IDH, distribuição fundiária e disparidades regionais [13, 14, 15], que examinam os determinantes do desenvolvimento humano em contextos rurais, também sem recorrer a ML aplicado a dados orçamentários. O presente artigo integra as três perspectivas ao utilizar o *XGBoost* sobre dados orçamentários federais combinados com variáveis socioeconômicas e do Censo Agropecuário, modelando simultaneamente o percentual do PIB-AGRO e o IDH como variáveis-alvo.

Tabela 1 – Posicionamento do presente estudo na literatura

Linha temática	Representantes	Método predominante	Diferencial deste estudo
ML em orçamento público	[4, 7, 8, 9]	XGBoost, GBM, redes neurais	Conecta features orçamentárias a IDH e PIB-AGRO
Investimento agrícola e crescimento	[10, 11, 12]	Econometria clássica	<i>Feature importance</i> quantifica canais de transmissão
IDH, distribuição fundiária e região	[13, 14, 15]	Painel, regressão múltipla	Integra dados do Censo Agropecuário 2017 ao modelo

3. MATERIAL E MÉTODOS

3.1 Fontes de dados e período de análise

O *dataset* foi construído a partir de quatro fontes abertas de acesso público:

- (i) Acompanhamento Físico-Financeiro das Ações Orçamentárias e Orçamento Fiscal e da Seguridade Social, disponíveis no Portal de Dados Abertos do Governo Federal [3, 17];
- (ii) Censo Agropecuário 2017 do IBGE [18], o mais atual disponível, que forneceu dados sobre estrutura fundiária e acesso à orientação técnica;
- (iii) série histórica do PIB do Agronegócio (CEPEA/ESALQ/USP–CNA) [2]; e
- (iv) cotações do dólar americano (USD) para conversão de valores correntes (CEPEA/ESALQ/USP) [2]. O período analisado compreende os exercícios fiscais de 2018 a 2022.

A unidade de análise é o par (exercício fiscal \times região/UF \times unidade orçamentária \times subfunção \times tipo de ação), resultando em 465 observações válidas após o processo de limpeza e integração. O *dataset* final não apresentou valores ausentes (null ou NA), dispensando imputação.

3.2 Variáveis e pré-processamento

O *dataset* incorpora cinco variáveis categóricas, sendo elas região, unidade federativa, unidade orçamentária (MAPA, EMBRAPA, INCRA, CONAB, SFB), subfunção (22 categorias) e tipo de ação (Atividade/OPEX, Projeto/CAPEX, Operação Especial). Complementarmente, foram incluídas variáveis numéricas abrangendo os valores orçamentários efetivamente disponíveis nas bases públicas (autorizado e pago, em reais e em dólares norte-americanos), indicadores macroeconômicos (PIB e PIB-AGRO setorial) e variáveis extraídas do Censo Agropecuário (total de propriedades, taxa de orientação técnica e distribuição populacional por propriedade). As variáveis categóricas foram codificadas via *one-hot encoding* e as variáveis numéricas foram padronizadas por meio do *StandardScaler*, técnica que transforma cada variável subtraindo sua média e dividindo pelo desvio padrão, resultando em média zero e desvio padrão unitário, de modo que nenhuma variável exerça influência desproporcional sobre o modelo em

razão de sua escala original. Adicionalmente, a transformação logarítmica `Log_Pop_Por_Propriedade` foi criada para capturar relações não lineares entre a distribuição populacional e a estrutura fundiária.

A seleção de *features* adotou abordagem teórica dirigida por hipótese (*hypothesis-driven*), na qual as variáveis foram definidas a priori a partir de um modelo conceitual dos determinantes do desenvolvimento agropecuário, e não por garimpo automático de correlações. Cada *feature* responde a uma dimensão explicativa, a saber, capacidade de investimento (`Pago_USD` e tipo de ação `Projeto/CAPEX`), riqueza regional (`PIB`), pressão demográfica sobre a terra (logaritmo da população por propriedade), acesso à assistência técnica (percentual de propriedades que recebem orientação técnica) e controles estruturais de região e de unidade orçamentária. Essa estratégia foi escolhida por três razões. Primeiro, a seleção teórica é metodologicamente mais conservadora que a automática, pois garante que cada variável tenha fundamento conceitual prévio, evitando que o algoritmo trate correlações espúrias como poder explicativo. Segundo, um conjunto fixo de *features* assegura interpretabilidade e comparabilidade entre os dois modelos do estudo (`IDH` e percentual do `PIB-AGRO`), permitindo confrontar lado a lado a importância das mesmas variáveis, o que é central para a leitura proposta. Terceiro, diante de um conjunto de dimensão moderada (465 observações), restringir os graus de liberdade do processo de seleção reduz o risco de sobreajuste. O conjunto final compõe 16 *features* para o modelo de percentual do `PIB-AGRO` e 14 *features* para o modelo de `IDH`. O *Recursive Feature Elimination* (RFE) com regressão linear permaneceu encadeado ao *pipeline* por consistência mecânica, mas, sob o conjunto teórico fixo, não atuou como redutor adicional. Como diagnóstico complementar, a multicolinearidade foi inspecionada pelo critério $|r| > 0,70$ (Seção 6).

3.3 Modelagem com XGBoost e otimização de hiperparâmetros

O *XGBoost Regressor* foi selecionado como algoritmo principal em razão de sua robustez a multicolinearidade residual, eficiência com *datasets* de tamanho moderado e capacidade de capturar interações não lineares entre *features* [4, 5]. A otimização de hiperparâmetros empregou *GridSearchCV* com validação cruzada *k-fold* ($k=5$), varrendo uma grade que incluiu $n_estimators \in \{100, 200\}$, $max_depth \in \{3, 5, 7\}$, $learning_rate \in \{0,05, 0,1, 0,2\}$, $reg_alpha \in \{0, 0,1\}$ e $reg_lambda \in \{0, 0,1\}$. Os dados foram divididos em conjuntos de treinamento (80%) e teste (20%).

A Tabela 2 resume os hiperparâmetros selecionados e sua justificativa. Para o modelo de `IDH`, a ausência de regularização adicional ($\alpha = \lambda = 0$) é justificada pelo menor número de *features* e pela menor dimensionalidade do problema. Para o modelo de percentual do `PIB-AGRO`, a Regularização L2 com $\lambda = 0,1$ (*Ridge*) contribuiu para a estabilidade das estimativas diante do maior número de variáveis.

Tabela 2 – Hiperparâmetros selecionados por *GridSearchCV* para cada modelo

Hiperparâmetro	Modelo PIB-AGRO	Modelo IDH	Justificativa
n_estimators	200	100	Capacidade preditiva vs. custo computacional
max_depth	5	5	Controle de <i>overfitting</i>
learning_rate	0,1	0,2	Convergência estável
reg_alpha (L1)	0	0	Sem esparsidade forçada
reg_lambda (L2)	0,1	0	Ridge moderado; robusto a multicolinearidade

3.4 Avaliação do modelo e análise de importância de features

O desempenho dos modelos foi avaliado mediante as métricas R^2 (Coeficiente de Determinação), MSE (*Mean Squared Error*), RMSE (*Root Mean Squared Error*) e MAE (*Mean Absolute Error*). A análise de importância de *features* utilizou o critério nativo do *XGBoost* (ganho médio de divisão ponderado por cobertura), permitindo identificar quais variáveis mais contribuíram para a redução do erro de previsão. Análise geográfica e temática complementar examinou a distribuição orçamentária por região e por subfunção ao longo do período.

4. RESULTADOS E DISCUSSÃO

4.1 Desempenho dos modelos

A Tabela 3 sumariza as métricas de desempenho dos dois modelos nos conjuntos de teste. Ambos apresentam poder explicativo elevado, validando a adequação do *XGBoost* para este domínio e respondendo afirmativamente à RQ3.

Tabela 3 – Métricas de desempenho dos modelos *XGBoost* (conjunto de teste, 20%)

Modelo	R^2	RMSE	MAE	MSE
XGBoost – % PIB-AGRO	0,8764	0,00956	0,00358	$9,14 \times 10^{-5}$
XGBoost – IDH	0,9632	0,00712	0,00374	$5,07 \times 10^{-5}$

O modelo para percentual do PIB-AGRO alcança R^2 igual a 0,8764, indicando que aproximadamente 87,6% da variabilidade do percentual do PIB agropecuário é explicada pelas *features* selecionadas. O RMSE de 0,00956 é expressivo dado o intervalo da variável-alvo situado entre 0,20 e 0,27. O modelo para IDH supera esse resultado, com R^2 igual a 0,9632, indicando que 96,3% da variabilidade do IDH é capturada pelas *features*, com RMSE de 0,00712 sobre uma escala de 0,676 a 0,859. Os baixos valores de MAE em ambos os modelos (0,00358 e 0,00374, respectivamente) corroboram que as previsões são muito próximas dos valores observados, sem distorções sistemáticas atribuíveis a *outliers*. Para aferir a

sensibilidade desses resultados à estratégia de seleção, a Seção 4.4 reporta um experimento de seleção automática aplicado ao modelo de IDH.

4.2 Importância relativa das features

A Tabela 4 apresenta o ranking de importância das *features* para os dois modelos, respondendo às RQs 1 e 2.

Tabela 4 – Importância relativa das *features* para os modelos XGBoost

Painel A – Modelo XGBoost para % PIB-AGRO ($R^2 = 0,8764$)		
Rank	Feature	Importância
1	% Receber Orientação Técnica	0,2818
2	Tipo de Ação Projeto (CAPEX)	0,2159
3	IDH	0,1175
4	Log. População por Propriedade	0,1059
5	PIB geral	0,0891
6	Unidade Orçamentária – EMBRAPA	0,0511
7	Região Sudeste	0,0400
8	Região Nordeste	0,0389
Painel B – Modelo XGBoost para IDH ($R^2 = 0,9632$)		
Rank	Feature	Importância
1	% Receber Orientação Técnica	0,5211
2	Região Sudeste	0,3189
3	Região Centro-Oeste	0,1067
4	Log. População por Propriedade	0,0179
5	PIB geral	0,0110
6	Região BR (nacional)	0,0109
7	Tipo Ação – Projeto (CAPEX)	0,0048
8	Região Norte	0,0036

- Orientação técnica rural (RQ1 e RQ2)

A variável %_Receber_Orientacao_Tecnica emerge como o fator de maior importância em ambos os modelos, respondendo por 28,18% no modelo de percentual do PIB-AGRO e por 52,11% no modelo de IDH. Este resultado é especialmente relevante diante da realidade revelada pelo Censo Agropecuário 2017 [18], segundo o qual apenas 20,2% das propriedades rurais brasileiras recebem algum tipo de orientação técnica, com disparidades regionais marcantes, sendo 48,6% no Sul, 10,4% no Norte e 8,2% no Nordeste. A convergência de ambos os modelos em torno dessa variável sugere que a extensão rural opera como o principal canal de transmissão entre o investimento público e os resultados socioeconômicos no campo, corroborando Torres [10], Diallo e Wouterse [11] e Piras et al. [13].

- **Investimentos em Projetos/CAPEX (RQ1)**

O tipo de ação Projeto, equiparado ao *Capital Expenditure* (CAPEX) no setor privado, é o segundo fator mais relevante para o percentual do PIB-AGRO, com importância relativa de 21,59%, o que reforça o papel dos investimentos de capital como motor do desempenho do setor. Esse achado é particularmente significativo quando confrontado com o cenário de execução orçamentária descrito na Tabela 5, no qual a deterioração da taxa de conversão de empenhos em pagamentos revela uma dificuldade crescente em efetivar justamente o tipo de investimento que o modelo identifica como mais determinante para o agronegócio. Há, portanto, um contraste entre o que os dados apontam como prioritário e o que o orçamento efetivamente executa.

- **Disparidades geográficas (RQ2)**

Para o modelo de IDH, a localização geográfica ocupa o segundo e terceiro lugares, com a região Sudeste respondendo por 31,89% e a região Centro-Oeste por 10,67%. Essa distribuição reflete as assimetrias históricas documentadas por Ondetti [15] e Yuliana et al. [19] e ressalta a necessidade de políticas regionalmente diferenciadas. As variáveis econômicas tradicionais, PIB e pagamentos em USD, mostraram importância relativamente baixa, de 1,10% e 0,35%, respectivamente, desafiando a premissa de que o crescimento econômico agregado é o principal motor do desenvolvimento humano.

4.3 Análise da execução orçamentária do MAPA

A análise da série orçamentária 2018-2022, apresentada na Tabela 5, revela um descompasso estrutural entre a autorização e a entrega de recursos. Embora o orçamento autorizado tenha crescido de R\$ 4,04 bilhões em 2018 para R\$ 7,22 bilhões em 2021, os valores efetivamente pagos permaneceram no patamar de R\$ 1,6 a R\$ 2,7 bilhões em todo o período. A taxa de execução, medida pela razão entre o pago e o autorizado de cada recorte, oscilou em faixa baixa e estável, de 39,7% em 2018, recuando para 36,5% em 2021 e recuperando-se para 43,5% em 2022, com média de 39,2% no quinquênio. A leitura regional reforça o diagnóstico, pois, enquanto a execução agregada nacional atinge 43,5%, as unidades regionais apresentam taxas substancialmente inferiores, de 11,3% no Centro-Oeste a 2,6% no Norte. Por órgão executor, a EMBRAPA alcança 56,9% de execução e o MAPA 46,6%, ao passo que a CONAB executa apenas 16,7% do autorizado. Esse padrão indica que o gargalo não reside na alocação orçamentária, mas na capacidade operacional de execução, sobretudo nas unidades regionais e em órgãos específicos, em consonância com a combinação de fatores econômicos, políticos e estratégicos identificada por Monteiro et al. [20].

Tabela 5 – Execução orçamentária do MAPA por exercício fiscal (2018-2022). Execução = razão entre o pago e o autorizado de cada exercício.

Ano	Autorizado (R\$ bi)	Pago (R\$ bi)	Execução (Pago/Autorizado)
2018	4,04	1,60	39,7%
2019	5,96	2,22	37,2%
2020	6,27	2,46	39,2%

2021	7,22	2,64	36,5%
2022	6,16	2,68	43,5%
Média	5,93	2,32	39,2%

4.4 Análise de robustez por seleção automática de features

Para avaliar a sensibilidade dos resultados à estratégia de seleção de *features*, aplicou-se ao modelo de IDH um procedimento de seleção automática em duas camadas, sem vazamento de informação. Na primeira camada, calculou-se a multicolinearidade exclusivamente sobre o conjunto de treino, descartando, em cada par com $|r| > 0,70$, a variável de maior redundância média. Na segunda camada, o RFE foi acoplado ao *pipeline* e o número de *features* a manter foi varrido pela validação cruzada do *GridSearchCV*, conjuntamente com os hiperparâmetros, com reescalonamento dentro de cada *fold*. Partindo de 42 *features* candidatas, a poda por multicolinearidade removeu 24 variáveis, restando 18, das quais o RFE selecionou 14.

O modelo resultante atinge R^2 de 0,9959 no conjunto de teste, com RMSE de 0,00239. Esse desempenho não deve ser lido como evidência de um modelo superior, mas como sintoma da dominância da variável de orientação técnica discutida na Seção 5.3. Ao remover a âncora teórica, o procedimento concentra ainda mais a importância nessa variável, que passa de 52,1% no modelo teórico para 66,2% na seleção automática. O achado relevante é de consistência, uma vez que a hierarquia de importância permanece essencialmente a mesma sob ambas as filosofias de seleção, com a orientação técnica e a região Sudeste ocupando as primeiras posições nos dois casos. A conclusão central do estudo, portanto, sobrevive à mudança de estratégia de seleção. O experimento foi conduzido apenas para o modelo de IDH, de caráter ilustrativo, uma vez que a estratégia adotada no estudo é a seleção teórica fixa; replicá-lo para o modelo de percentual do PIB-AGRO não acrescentaria evidência metodológica e comprometeria a comparabilidade entre os dois modelos que motiva o conjunto fixo de *features*.

5. DISCUSSÃO

5.1 Implicações para políticas públicas

Os resultados convergem para três recomendações de política pública, cada uma ancorada nos percentuais de importância produzidos pelos modelos, e reforçam a necessidade de consolidar uma cultura *data-driven* na gestão orçamentária do setor público:

- (i) A ampliação do acesso à orientação técnica rural deve ser tratada como prioridade orçamentária de primeira ordem, e não como linha residual. Com apenas 20,2% das propriedades cobertas [18] e uma importância relativa de 28% a 52% nos modelos, o retorno marginal de cada real alocado nesse canal supera o de praticamente qualquer outra categoria de despesa analisada.
- (ii) Os investimentos em Projetos/CAPEX, que por definição abrangem bens de capital voltados à expansão e à melhoria operacional, devem ser protegidos dos contingenciamentos, dada sua importância relativa de 21,59% no modelo de PIB-AGRO.

- (iii) As políticas de desenvolvimento devem incorporar dimensão regional explícita, dado que a localização geográfica explica 42,6% da variação do IDH no modelo, evidência de que transferências e programas de caráter universal têm impacto limitado na redução de desigualdades regionais estruturais.

5.2 Interpretação à luz do referencial teórico

A centralidade da orientação técnica nos dois modelos é consistente com o argumento de Soleman et al. [14] sobre a relação recíproca entre desenvolvimento humano e crescimento econômico, na medida em que, ao elevar o IDH via extensão rural, o setor agropecuário cria as condições para seu próprio crescimento futuro. A baixa importância das variáveis econômicas tradicionais, como o PIB e os pagamentos em USD, nos modelos de IDH confirma a advertência de Piras et al. [13] de que o crescimento do agronegócio, sem a mediação de políticas distributivas, não se traduz automaticamente em melhoria de bem-estar. O resultado alinha-se a Corcioli et al. [12], que denunciam o viés de alocação em favor de grandes produtores, sugerindo que a fragmentação dos benefícios explica parte da dissociação entre o crescimento do PIB-AGRO e o IDH regional.

5.3 Limitações do estudo

Quatro limitações merecem reconhecimento explícito. A primeira é que o *dataset* cobre 465 observações em cinco exercícios fiscais, dimensão determinada pela própria disponibilidade das bases públicas oficiais do governo federal e que não pode ser ampliada artificialmente. Cabe ressaltar que houve aproveitamento integral das observações disponíveis, sem descarte de registros válidos, e que a seleção teórica de *features*, ao restringir os graus de liberdade do processo de modelagem, atenua os riscos associados a um conjunto de dados de dimensão moderada. Ainda assim, mudanças de governo, choques climáticos e revisões de política agrícola ocorridas após 2022 podem alterar a hierarquia de importância das *features*, o que recomenda cautela na generalização temporal dos modelos. A segunda é que o modelo não incorpora variáveis diretas de saúde e educação, dimensões constitutivas do IDH, em razão de sua ausência no escopo de coleta, de modo que sua inclusão em trabalhos futuros poderia refinar o diagnóstico. A terceira é que os dados do Censo Agropecuário são de 2017, o que introduz defasagem temporal relevante em um setor de dinâmica rápida, uma vez que o próximo censo, em preparação pelo IBGE para o ciclo 2026/2027, ainda não está disponível. A quarta, e mais sensível do ponto de vista interpretativo, é a forte dominância da variável de orientação técnica, que responde por 52,1% da importância no modelo de IDH e por 28,2% no modelo de percentual do PIB-AGRO. A análise de multicolinearidade revelou que as variáveis derivadas da mesma fonte de assistência técnica (número de propriedades que recebem e que não recebem orientação, em valores absolutos e percentuais) são fortemente colineares entre si, e que a própria variável de orientação técnica apresenta correlação de 0,814 com o IDH no modelo de percentual do PIB-AGRO. Esse padrão sugere que parte do elevado poder explicativo pode decorrer de uma associação de origem entre cobertura de assistência técnica e nível de desenvolvimento estrutural das unidades federativas, e não de uma relação causal direta. O elevado R^2 deve, portanto, ser lido como evidência de associação robusta, não de causalidade, o que motiva a investigação causal proposta na agenda futura.

6. AMEAÇAS À VALIDADE

- Validade interna

A principal ameaça reside na possibilidade de variáveis omitidas que se correlacionem simultaneamente com as *features* e com as variáveis-alvo. Para mitigar esse risco, o desenho do estudo adotou três estratégias complementares.

- (i) Em primeiro lugar, a multicolinearidade foi avaliada já na etapa de concepção do conjunto de variáveis, com a exclusão de variáveis redundantes derivadas da mesma fonte. O diagnóstico por matriz de correlação ($|r| > 0,70$), conduzido separadamente para cada modelo, confirmou ausência de pares colineares no conjunto do modelo de IDH e identificou um único par no conjunto do modelo de percentual do PIB-AGRO, entre o IDH e a variável de orientação técnica, cuja interpretação é tratada na validade de construto e na dominância discutida na Seção 5.3.
- (ii) Em segundo lugar, a seleção teórica dirigida por hipótese restringiu o conjunto a variáveis com fundamento conceitual prévio, evitando que o processo de modelagem capturasse correlações espúrias como poder explicativo.
- (iii) Por fim, a regularização L2 foi empregada no modelo de percentual do PIB-AGRO para conferir estabilidade às estimativas diante do maior número de preditores.

- Validade de construto

A equiparação de Projetos ao CAPEX e de Atividades ao OPEX, embora fundamentada em Tahu et al. [16] e Puhach [8], constitui uma aproximação, uma vez que o orçamento público segue lógicas distintas das que regem o balanço patrimonial corporativo. As conclusões sobre a importância do CAPEX devem, portanto, ser interpretadas dentro desse contexto analógico. Cabe ressaltar, contudo, que essa ressalva é específica do dado público utilizado neste estudo. Aplicado a uma base de origem corporativa, na qual CAPEX e OPEX constituem categorias contábeis nativas e não aproximações, o mesmo *pipeline* preservaria sua coerência analítica, o que evidencia que a limitação reside na natureza do dado, e não no método empregado.

- Validade externa

Os modelos foram treinados e testados com dados de uma única instituição, o MAPA, em um único país. Análises comparativas com ministérios de agricultura de outros países de perfil agrícola similar, como Argentina, Estados Unidos e Índia, poderiam ampliar a generalidade das conclusões. Vale destacar, ainda, que a arquitetura do *pipeline* não é específica do contexto agropecuário, de modo que o mesmo método pode ser estendido à análise orçamentária de outros ministérios, desde que as variáveis-alvo e o conjunto de *features* sejam devidamente ajustados ao domínio de cada pasta. Esse potencial de transferência alcança também o setor privado, conforme discutido na análise de validade de construto, no qual as categorias contábeis estão nativamente

definidas. Em conjunto, essas possibilidades reforçam o potencial do método como ferramenta geral de apoio à decisão *data-driven*, aplicável tanto à gestão pública quanto à gestão orçamentária corporativa.

- **Validade de conclusão**

Os modelos alcançam coeficientes de determinação elevados, com R^2 de 0,8764 para o percentual do PIB-AGRO e de 0,9632 para o IDH, o que confere robustez estatística às conclusões sobre a importância relativa das *features*. Cabe ressalva quanto à dominância da variável de orientação técnica (Seção 5.3), pois parte do poder explicativo pode decorrer de associação de origem entre cobertura de assistência técnica e nível de desenvolvimento estrutural, de modo que os coeficientes de determinação sustentam afirmações de associação, não de causalidade. A análise de robustez por seleção automática (Seção 4.4) corrobora a estabilidade da hierarquia de importância sob filosofia de seleção distinta. Além de seu valor imediato, o estudo abre caminho para a investigação das relações causais subjacentes às associações identificadas e a replicação do método em outros contextos institucionais e setoriais.

7. CONCLUSÃO

Os resultados deste estudo evidenciaram que o *XGBoost*, aplicado a dados orçamentários abertos do MAPA, integrados a variáveis do Censo Agropecuário 2017 e da série CEPEA/CNA, produz modelos preditivos com alto poder explicativo para o percentual do PIB-AGRO ($R^2 = 0,8764$) e para o IDH das unidades federativas ($R^2 = 0,9632$), respondendo afirmativamente às três RQs propostas.

As principais contribuições empíricas são:

- (i) a orientação técnica rural é o fator de maior influência em ambos os alvos, superando em importância variáveis econômicas tradicionais como o PIB e os valores orçamentários pagos;
- (ii) os investimentos em Projetos/CAPEX são o segundo determinante do percentual do PIB-AGRO, com importância relativa de 21,59%, o que reforça seu papel estratégico na alocação de recursos;
- (iii) as disparidades geográficas explicam 42,6% da variação do IDH, evidenciando a insuficiência de políticas uniformes para a redução de desigualdades regionais.

As limitações do estudo, a saber, a janela temporal de cinco anos, a ausência de variáveis diretas de saúde e educação e a defasagem do Censo Agropecuário, definem a agenda para pesquisas futuras:

- (i) a incorporação de variáveis climáticas e ambientais, cuja relevância cresce com as pressões do aquecimento global sobre a produtividade agrícola;
- (ii) comparações internacionais com países de perfil agrícola semelhante e a extensão do método a outros ministérios e ao setor privado;

- (iii) a atualização contínua dos modelos com dados anuais, garantindo sua relevância para ciclos orçamentários futuros;
- (iv) a investigação das relações causais subjacentes às associações identificadas, confirmando em que medida a expansão da cobertura de orientação técnica produz ganhos mensuráveis de IDH e de PIB-AGRO.

Em síntese, a integração de técnicas de aprendizado de máquina ao planejamento orçamentário federal representa uma evolução metodológica concreta em direção a uma gestão pública mais transparente, eficiente e orientada por dados. Os resultados aqui apresentados oferecem base empírica sólida para que gestores e formuladores de política priorizem a extensão rural e os investimentos de capital como alavancas fundamentais do desenvolvimento sustentável do agronegócio brasileiro.

AGRADECIMENTOS

O autor agradece à Universidade de São Paulo / ESALQ pelo suporte ao Programa MBA em Data Science & Analytics de 2024, bem como às entidades que disponibilizaram os dados em formato aberto: Governo do Brasil (Portal de Dados Abertos), IBGE, CEPEA e Confederação Nacional da Agricultura e Pecuária (CNA).

REFERÊNCIAS

- [1] MINENHLE, H. AI in Finance: Challenges, Techniques, and Opportunities. ACM Computing Surveys, 2022. DOI: 10.1145/3502289.
- [2] CEPEA/ESALQ/USP; CNA. PIB do agronegócio brasileiro de 1996 a 2024. Disponível em: <https://www.cepea.esalq.usp.br>. Acesso em: 15 jul. 2024.
- [3] GOVERNO DO BRASIL. Dados do Orçamento Federal – Orçamento Fiscal e da Seguridade Social (2018–2022). Disponível em: <https://dados.gov.br>. Acesso em: 15 fev. 2024.
- [4] CAPONE, C. et al. Artificial Intelligence Models for Predicting Budget Expenditures. Eurasian Journal of Economic and Business Studies, 2024.
- [5] DARMAWAN, H.; YULIANA, M.; HADI, M.Z.S. GRU and XGBoost Performance with Hyperparameter Tuning Using GridSearchCV and Bayesian Optimization on an IoT-Based Weather Prediction System. International Journal on Advanced Science, Engineering and Information Technology, 2023.
- [6] CASTRO, J. MBA_USP – Repositório público de código e dados. GitHub, 2024. Disponível em: https://github.com/Jotta-se/MBA_USP.
- [7] HAYTHAM, H.; ELMOUSALAMI. Comparison of Artificial Intelligence Techniques for Project Conceptual Cost Prediction. arXiv: Learning, 2019.
- [8] PUHACH, Y.V. Modern econometric approaches of machine learning in the simulation of factors influencing local budget expenditures. Actual Problems of Economics, 2023.
- [9] ZAINAB, H.A. et al. Developing an Integrative Data Intelligence Model for Construction Cost Estimation. Complexity, 2022. DOI: 10.1155/2022/4285328.
- [10] TORRES, D.A.P. Investimento em pesquisa e a produtividade agropecuária: uma revisão da literatura recente. Revista de Política Agrícola, v. 32, p. 106-120, 2023.
- [11] DIALLO, M.; WOUTERSE, F. Agricultural development promises more growth and less poverty in Africa. Development Policy Review, 2022. DOI: 10.1111/dpr.12669.

- [12] CORCIOLI, G.; MEDINA, G.S.; ARRAIS, C.A. Missing the Target: Brazil's Agricultural Policy Indirectly Subsidizes Foreign Investments. *Frontiers in Sustainable Food Systems*, 2022.
- [13] PIRAS, S.; WESZ, V.J.; GHINOI, S. Soy Expansion, Environment, and Human Development: An Analysis across Brazilian Municipalities. *Sustainability*, v. 13, n. 13, 2021. DOI: 10.3390/su13137246.
- [14] SOLEMAN, R. et al. The Effect of HDI and Macroeconomic Variables on Economic Growth in Indonesia 2015–2020. *Jurnal Ilmu Ekonomi Terapan*, 2022.
- [15] ONDETTI, G. Ideational Bases of Land Reform in Brazil: 1910 to the Present. 2020.
- [16] TAHU, G.P. et al. The Influence of Regional Original Income and General Allocation Funds on Economic Growth with Capital Expenditures as an Intervening Variable. *Revista de Gestão Social e Ambiental*, v. 18, 2024.
- [17] GOVERNO DO BRASIL. Acompanhamento Físico-Financeiro das Ações Orçamentárias. Disponível em: <https://dados.gov.br>. Acesso em: 15 fev. 2024.
- [18] IBGE. Censo Agropecuário 2017: Resultados definitivos. Rio de Janeiro: IBGE, 2019.
- [19] YULIANA; PRIHATININGSIH, A.; ISPA, A.Y. Dynamics Of Economic Development Disparities: A New Perspective In Regional Analysis. 2024.
- [20] MONTEIRO, J.J. et al. Reflexos da COVID-19 no orçamento público do governo federal. *Revista de Gestão e Organizações*, v. 14, 2021.
- [21] SILVA, L.A.P. et al. Vetores de degradação das pastagens no Cerrado norte mineiro. *Raega – O Espaço Geográfico em Análise*, v. 57, p. 66-80, 2023.
- [22] REIS, C.C. et al. Sistemas integrados de lavoura-pecuária-floresta no Cerrado. 2020.