

The logo for idp, featuring the letters 'idp' in a bold, white, sans-serif font. The 'i' is lowercase and positioned to the left of the 'dp' which is in uppercase.

# MESTRADO PROFISSIONAL

## EM ECONOMIA

---

**MODELAGEM PREDITIVA DA EFICIÊNCIA DOS  
INVESTIMENTOS PÚBLICOS EM OBRAS E  
INFRAESTRUTURA NO ESTADO DO AMAPÁ**

**IVY THIAGO VASCONCELOS AMANAJÁS**

Brasília-DF, 2025

**IVY THIAGO VASCONCELOS AMANAJÁS**

**MODELAGEM PREDITIVA DA EFICIÊNCIA DOS INVESTIMENTOS PÚBLICOS EM OBRAS E INFRAESTRUTURA NO ESTADO DO AMAPÁ**

Dissertação apresentada ao Programa de Pós Graduação em Economia, do Instituto Brasileiro de Ensino, Desenvolvimento e Pesquisa, como requisito parcial para obtenção do grau de Mestre.

**Orientador**

Professor Doutor Leonardo Monteiro Monastério.

Brasília-DF 2025

**IVY THIAGO VASCONCELOS AMANAJÁS**

**MODELAGEM PREDITIVA DA EFICIÊNCIA DOS  
INVESTIMENTOS PÚBLICOS EM OBRAS E  
INFRAESTRUTURA NO ESTADO DO AMAPÁ**

Dissertação apresentada ao Programa de Pós Graduação em Economia, do Instituto Brasileiro de Ensino, Desenvolvimento e Pesquisa, como requisito parcial para obtenção do grau de Mestre.

Aprovado em 01 / 07 / 2025

**Banca Examinadora**

---

Prof. Dr. Leonardo Monteiro Monastério - Orientador

---

Prof. Dr. Claudiomar Matias Rolim Filho

---

Prof. Dr. Pedro Fernando de Almeida Nery Ferreira

## Código de catalogação na publicação – CIP

A484m Amanajás, Ivy Thiago Vasconcelos  
Modelagem preditiva da eficiência dos investimentos públicos  
em obras e infraestrutura no estado do Amapá / Ivy Thiago  
Vasconcelos Amanajás. — Brasília: Instituto Brasileiro Ensino,  
Desenvolvimento e Pesquisa, 2025.

101 f. : il.

Orientador: Prof. Dr. Leonardo Monteiro Monastério

Dissertação (Mestrado Profissional em Economia) —  
Instituto Brasileiro Ensino, Desenvolvimento e Pesquisa – IDP,  
2025.

1. Eficiência (serviço público). 2. Políticas públicas. 3.  
Infraestrutura. 4. Criação de empregos. I.Título

CDD 330

Elaborada pela Biblioteca Ministro Moreira Alves

## RESUMO

Este trabalho avalia a eficiência dos investimentos públicos em obras e infraestrutura no Estado do Amapá, mensurando sua capacidade de gerar emprego e renda a partir de recursos orçamentários. Utilizamos 703 registros oficiais (2007-2025), submetidos a um protocolo de limpeza que descartou projetos com valor global igual a zero, converteu variáveis monetárias para formato numérico uniforme e imputou faltantes críticos por Predictive Mean Matching. Geramos dois indicadores-chave de eficiência — empregos por real investido e receita por real investido — além de variáveis de controle (etapa da obra, eixo temático, região administrativa, impacto local).

A estratégia empírica combinou técnicas de Machine Learning (Random Forest, Gradient Boosting, XGBoost e Redes Neurais) com métodos estatísticos convencionais. Os modelos foram estimados em amostras de treino/ teste (70 %/30 %) sob validação cruzada repetida (5 folds × 2 repetições). O Random Forest apresentou o melhor desempenho geral ( $R^2 = 0,956$ ; MAE = 3,04; RMSE = 32,41), seguido de perto pelo XGBoost ( $R^2 = 0,959$ ; MAE = 2,65; RMSE = 35,05). O Gradient Boosting obteve performance intermediária ( $R^2 = 0,562$ ), enquanto a Rede Neural mostrou instabilidade na validação ( $R^2$  não estatisticamente definido). As análises de importância de variáveis convergiram na primazia do valor\_global, do impacto\_por\_projeto e do impacto\_por\_eixo, sugerindo que escala financeira e externalidades locais são determinantes centrais da eficiência.

Técnicas de agrupamento (k-means, DBSCAN) identificaram três perfis distintos de custo-benefício, com um grande cluster de baixa eficiência e dois grupos menores de alto desempenho. Testes t e ANOVA indicaram que projetos estratégicos apresentam, em média, maior eficiência de emprego ( $p \approx 0,045$ ), ao passo que projetos terceirizados exibem eficiência inferior ( $p < 0,001$ ). Um modelo difference-in-differences controlado por efeitos fixos de eixo mostrou aumento significativo da eficiência média após 2020 ( $\beta = 11,46$ ;  $p < 0,05$ ), ainda que o efeito marginal do valor investido permaneça estatisticamente frágil.

Os resultados apontam para oportunidades de reordenação de prioridades, privilegiando eixos com alto impacto local documentado e evitando a dispersão de recursos em iniciativas terceirizadas de baixo retorno. Apesar da robustez relativa dos modelos, as conclusões dependem da qualidade dos registros administrativos e exigem monitoramento contínuo para embasar decisões orçamentárias futuras.

**Palavras-chave: Eficiência; Infraestrutura; Políticas Públicas; Machine Learning; Random Forest; XGBoost; Estado do Amapá.**

---

**ABSTRACT**

This dissertation evaluates the efficiency of public investments in construction and infrastructure projects in the State of Amapá, measuring their capacity to generate employment and income from budgetary resources. The analysis draws on 703 official records (2007–2025), processed through a cleaning protocol that removed projects with zero total value, standardized monetary variables, and imputed critical missing data using Predictive Mean Matching. Two key efficiency indicators were constructed—jobs per real invested and revenue per real invested—alongside control variables such as project stage, thematic axis, administrative region, and local impact.

The empirical strategy combined Machine Learning techniques (Random Forest, Gradient Boosting, XGBoost, and Neural Networks) with conventional statistical methods. Models were estimated using 70/30 train–test splits under repeated cross-validation (5 folds  $\times$  2 repetitions). Random Forest achieved the best overall performance ( $R^2 = 0.956$ ; MAE = 3.04; RMSE = 32.41), closely followed by XGBoost ( $R^2 = 0.959$ ; MAE = 2.65; RMSE = 35.05). Gradient Boosting presented intermediate results ( $R^2 = 0.562$ ), while the Neural Network showed instability during validation ( $R^2$  not statistically defined). Variable importance analyses consistently highlighted total project value, project-level impact, and axis-level impact as key determinants of efficiency, indicating that financial scale and local externalities play central roles.

Clustering techniques (k-means, DBSCAN) identified three distinct cost–benefit profiles, with one large low-efficiency cluster and two smaller high-performing groups. t-tests and ANOVA indicated that strategic projects exhibit higher employment efficiency on average ( $p \approx 0.045$ ), whereas outsourced projects display significantly lower efficiency ( $p < 0.001$ ). A difference-in-differences model with axis fixed effects revealed a significant increase in average efficiency after 2020 ( $\beta = 11.46$ ;  $p < 0.05$ ), although the marginal effect of investment value remains statistically weak.

Overall, the findings point to opportunities for reallocating priorities toward axes with documented high local impact and away

from low-return outsourced initiatives. Despite the relative robustness of the models, the conclusions depend on the quality of administrative records and call for continuous monitoring to support future budgetary decision-making.

**Keywords: Efficiency; Infrastructure; Public Policy; Machine Learning; Random Forest; XGBoost; State of Amapá.**

---

## LISTA DE ABREVIATURAS E SIGLAS

**ABNT** Associação Brasileira de Normas Técnicas

**IDP** Instituto Brasileiro de Ensino, Desenvolvimento e Pesquisa

---

## LISTA DE ILUSTRAÇÕES

**Figura 1**

Médias de eficiência (emprego / valor investido) por eixo de obra pública no Estado do Amapá

63

**Figura 2**

Boxplots e diagramas de dispersão

64

**Figura 3**

Concentração de frequência em torno de valores reduzidos e a existência de outliers positivos

65

**Figura 4**

Concentração de frequência em torno de valores reduzidos e a existência de outliers positivos

65

**Figura 5**

Matriz de correlação obtida via ggcorrplot

67

**Figura 6**

Agrupamento dos projetos pelo algoritmo K-means ( $k = 3$ ) em espaço bidimensional; o polígono sombreado delimita o mega-cluster que reúne  $\approx 99\%$  das observações.

72

# SUMÁRIO

## 1. INTRODUÇÃO ..... 15

1.1 CONTEXTUALIZAÇÃO DO TEMA .....	15
1.2 PROBLEMA DE PESQUISA.....	16
1.3 HIPÓTESES DA PESQUISA.....	17
1.4 OBJETIVOS GERAL E ESPECÍFICOS.....	17
1.4.1 OBJETIVO GERAL .....	17
1.4.2 OBJETIVOS ESPECÍFICOS .....	18
1.5 DELIMITAÇÃO DO ESCOPO DO ESTUDO .....	19
1.6 JUSTIFICATIVA DO TEMA: RELEVÂNCIA E CONTRIBUIÇÃO.....	19
1.7 ORGANIZAÇÃO DO ESTUDO.....	20

## 2. REFERENCIAL TEÓRICO ..... 23

2.1 DESENVOLVIMENTO ECONÔMICO REGIONAL E INVESTIMENTOS PÚBLICOS.....	23
2.2 MÉTODOS QUANTITATIVOS NA AVALIAÇÃO DE EFICIÊNCIA EM OBRAS PÚBLICAS .....	25
2.3 MACHINE LEARNING NA ANÁLISE DE POLÍTICAS PÚBLICAS .....	28
2.4 MODELAGEM ESTATÍSTICA VS. MACHINE LEARNING: VANTAGENS E LIMITAÇÕES .....	30
2.5 MODELOS DE MACHINE LEARNING APLICADOS À AVALIAÇÃO DE EFICIÊNCIA .....	32
2.5.1 MODELOS BASEADOS EM ÁRVORES: RANDOM FOREST, GRADIENT BOOSTING E XGBOOST.....	32
2.5.2 MODELOS DE AGRUPAMENTO: K-MEANS E DBSCAN .....	33
2.5.3 REDES NEURAIS ARTIFICIAIS PARA PREDIÇÃO DE IMPACTO.....	34
2.6 ESTUDOS EMPÍRICOS RELACIONADOS .....	35
2.6.1 LITERATURA INTERNACIONAL .....	35
2.6.2 LITERATURA NACIONAL.....	36

## 3. METODOLOGIA ..... 39

3.1 LIMITAÇÕES E PREMISSAS METODOLÓGICAS.....	39
3.2 FONTE DE DADOS E VARIÁVEIS UTILIZADAS .....	41
3.2.1 FONTES DE DADOS PRIMÁRIOS E SECUNDÁRIOS .....	41

# SUMÁRIO

3.2.2 DEFINIÇÃO DE INPUTS (VARIÁVEIS DE ENTRADA).....	<b>42</b>
3.2.3 DEFINIÇÃO DE OUTPUTS (VARIÁVEIS DE SAÍDA) .....	<b>43</b>
3.3 ORGANIZAÇÃO E PRÉ-PROCESSAMENTO DA BASE DE DADOS.....	<b>45</b>
3.3.1 LIMPEZA E TRATAMENTO DE DADOS .....	<b>45</b>
3.3.2 ENGENHARIA DE FEATURES E CRIAÇÃO DE NOVAS VARIÁVEIS .....	<b>46</b>
3.3.3 NORMALIZAÇÃO E TRANSFORMAÇÃO DE DADOS .....	<b>47</b>
3.4 MODELOS UTILIZADOS PARA PREDIÇÃO E CLASSIFICAÇÃO .....	<b>48</b>
3.4.1 RANDOM FOREST, GRADIENT BOOSTING E XGBOOST PARA PREVISÃO DA EFICIÊNCIA .....	<b>49</b>
3.4.2 MODELOS DE AGRUPAMENTO (K-MEANS E DBSCAN) .....	<b>50</b>
3.4.3 REDES NEURAIS PARA MODELAGEM NÃO LINEAR DA EFICIÊNCIA	<b>51</b>
3.5 ESTRATÉGIAS DE TREINAMENTO E VALIDAÇÃO DOS MODELOS.....	<b>52</b>
3.5.1 VALIDAÇÃO CRUZADA (CROSS-VALIDATION) .....	<b>52</b>
3.5.2 AJUSTE DE HIPERPARÂMETROS (HYPERPARAMETER TUNING) .....	<b>53</b>
3.5.3 ANÁLISE DE OVERFITTING E GENERALIZAÇÃO DOS MODELOS.....	<b>54</b>
3.6 MÉTRICAS DE AVALIAÇÃO DA EFICIÊNCIA PREDITIVA.....	<b>55</b>
3.6.1 ERRO MÉDIO ABSOLUTO (MAE).....	<b>55</b>
3.6.2 RAIZ DO ERRO QUADRÁTICO MÉDIO (RMSE).....	<b>56</b>
3.6.3 COEFICIENTE DE DETERMINAÇÃO (R <sup>2</sup> ) .....	<b>56</b>
3.6.4 IMPORTÂNCIA DAS VARIÁVEIS (FEATURE IMPORTANCE) .....	<b>57</b>
3.7 FERRAMENTAS E SOFTWARES UTILIZADOS .....	<b>58</b>
3.7.1 LINGUAGEM R E PACOTES ESTATÍSTICOS (TIDYVERSE, RANDOMFOREST, CARET, XGBOOST) .....	<b>58</b>
3.7.2 BIBLIOTECAS PARA VISUALIZAÇÃO DE DADOS (GGPLOT2, GRIDEXTRA, GGCORRPLOT) .....	<b>60</b>
<b>4. RESULTADOS E DISCUSSÃO .....</b>	<b>63</b>
4.1 ANÁLISE EXPLORATÓRIA DOS DADOS (EXPLORATORY DATA ANALYSIS – EDA) .....	<b>63</b>
4.1.1 ESTATÍSTICAS DESCRIPTIVAS DAS VARIÁVEIS DE INTERESSE .....	<b>64</b>
4.1.2 ANÁLISE DE CORRELAÇÕES E RELAÇÕES ENTRE VARIÁVEIS.....	<b>66</b>
4.2 RESULTADOS DA MODELAGEM PREDITIVA .....	<b>69</b>

# SUMÁRIO

4.2.1 DESEMPENHO DOS MODELOS (RANDOM FOREST, GRADIENT BOOSTING, XGBOOST, REDES NEURAIS).....	69
4.2.2 INTERPRETAÇÃO DOS PESOS E IMPORTÂNCIA DAS VARIÁVEIS (FEATURE IMPORTANCE) .....	71
4.2.3 IDENTIFICAÇÃO DE PADRÕES POR AGRUPAMENTO (K-MEANS, DBSCAN).....	73
4.3 ANÁLISE DE SENSIBILIDADE E ROBUSTEZ DOS RESULTADOS .....	74
4.3.1 AJUSTES DE PARÂMETROS NO DBSCAN E VARIAÇÕES NO PRÉ-PROCESSAMENTO DOS DADOS.....	75
4.3.2 VARIAÇÕES DE HIPERPARÂMETROS NOS MODELOS DE PREDIÇÃO .....	76
4.3.3 RESULTADOS ADICIONAIS DE SFA E DIF-IN-DIF .....	77
4.3.4 CONSIDERAÇÕES GERAIS DE ROBUSTEZ .....	78
4.4 IMPLICAÇÕES PARA FORMULAÇÃO DE POLÍTICAS PÚBLICAS .....	80
4.5 COMPARAÇÃO COM A LITERATURA E DISCUSSÃO CRÍTICA .....	82
<b>5. CONSIDERAÇÕES FINAIS .....</b>	<b>87</b>
5.1 RETOMADA DOS OBJETIVOS PROPOSTOS.....	87
5.2 PRINCIPAIS CONCLUSÕES .....	89
5.3 LIMITAÇÕES DO ESTUDO .....	90
5.4 SUGESTÕES PARA PESQUISAS FUTURAS .....	91
5.5 RECOMENDAÇÕES PARA POLÍTICAS PÚBLICAS BASEADAS EM MACHINE LEARNING .....	91
<b>REFERÊNCIAS.....</b>	<b>95</b>



## 1

## INTRODUÇÃO

## 1.1 CONTEXTUALIZAÇÃO DO TEMA

A avaliação de eficiência em investimentos públicos — especialmente em obras e infraestrutura — figura entre as pautas centrais de desenvolvimento econômico e social, tanto em países emergentes quanto em economias consolidadas. No Brasil, e de modo particular no Estado do Amapá, a otimização dos gastos governamentais ganha relevo diante das restrições orçamentárias, da amplitude das demandas sociais e da necessidade de reduzir desigualdades regionais. Torna-se, portanto, relevante examinar **em que medida** (e não exatamente “quanto”) os recursos públicos destinados a projetos de infraestrutura se convertem em benefícios à sociedade, notadamente na geração de emprego, renda e ampliação de serviços públicos.

O Amapá apresenta especificidades geográficas e estruturais que elevam a importância de avaliar, ainda que de forma aproximada, os retornos de obras públicas. Com logística limitada e cobertura desigual de serviços como saneamento e energia nas áreas ribeirinhas, a priorização de investimentos requer robustez técnica na elaboração dos projetos e **cautela na interpretação** dos resultados empíricos disponíveis. Assim, mesmo análises preliminares podem auxiliar o poder público a identificar gargalos e direcionar esforços a setores ou regiões mais necessitados.

A demanda por transparência e *accountability* também influencia o debate contemporâneo sobre eficiência governamental. Indicadores quantitativos — mesmo que construídos como **proxies** — permitem ilustrar se cada real investido tende a gerar retornos proporcionais em postos de trabalho e renda local. Tais indicadores apoiam a tomada de decisão dos gestores, ao tempo em que fomentam a cultura de avaliação sistemática de políticas públicas.

Do ponto de vista econômico, eficiência pode ser compreendida como a relação entre custos diretos (e, quando possível, indiretos) e benefícios concretos decorrentes da implementação de um projeto. Todavia, muitos impactos materializam-se no longo prazo; intervenções

como estradas ou pontes podem alterar padrões de mobilidade regional por décadas. Por isso, a análise de curto prazo (empregos durante a construção) deve ser lida em conjunto com projeções de médio e longo prazos (incremento de renda local e atração de empresas).

Indicadores de emprego e renda nos municípios amapaenses constituem sinais evidentes de desenvolvimento socioeconômico. O aumento da ocupação estimula o consumo e pode desencadear efeitos multiplicadores na cadeia produtiva; a elevação de renda repercute na arrecadação tributária e na qualidade de vida. Contudo, a precisão de tais estimativas depende da consistência das bases de dados — condição que, no caso amapaense, apresenta lacunas importantes.

Dessa forma, esta pesquisa busca **contribuir exploratoriamente** por meio de técnicas de análise de dados e modelagem preditiva, investigando como cada eixo de investimento público se comporta em termos de retorno econômico e social. Ao trazer esse debate para o plano empírico, pretende oferecer subsídios iniciais aos decisores, na expectativa de que estudos futuros, apoiados em bases mais completas, possam aprofundar a mensuração da eficiência e aprimorar a alocação dos recursos públicos.

## 1.2 PROBLEMA DE PESQUISA

Diante da diversidade de áreas em que o Governo do Amapá aloca recursos para obras públicas — como infraestrutura viária, saneamento, educação, saúde, habitação, entre outras — surge a necessidade de identificar qual eixo de investimento mais contribui, de forma mensurável, para a criação de empregos e o incremento da renda local. Esse desafio torna-se ainda mais relevante quando se considera o montante limitado de verbas disponíveis, a dispersão geográfica das comunidades beneficiadas e a urgência de promover resultados socioeconômicos positivos em diferentes regiões do estado.

Assim, a pergunta que orienta este estudo é: "**Qual eixo de investimento público mais contribui para a geração de emprego e renda, considerando o valor investido e outras características intrínsecas aos projetos?**". A resposta a essa questão permitirá não apenas comparar o retorno de diferentes frentes de atuação governamental, mas também fornecer evidências empíricas que auxiliem os gestores na priorização de obras e na alocação estratégica

dos recursos disponíveis, maximizando o impacto socioeconômico e garantindo maior transparência e eficácia na condução das políticas públicas.

## 1.3 HIPÓTESES DA PESQUISA

Tendo em vista o objetivo de identificar quais eixos de investimento público mais contribuem para a geração de emprego e renda no Estado do Amapá, este estudo parte das seguintes hipóteses de pesquisa:

- 1. H1: Há evidências de que as diferenças nos níveis de geração de empregos e renda entre os diversos eixos podem ser estatisticamente significativas, sugerindo que alguns eixos tendem a apresentar maior proporção de postos de trabalho gerados e de recursos financeiros no entorno.**
- 2. H2: Projetos classificados como estratégicos apresentam diferenças substantivamente significativas em eficiência, em comparação aos não estratégicos, devido a uma definição de metas e acompanhamento mais estruturados.**
- 3. H3: A adoção de metodologias quantitativas avançadas, como algoritmos de Machine Learning, permite prever de maneira mais precisa a eficiência na geração de empregos e renda, em relação a métodos estatísticos tradicionais, ao capturar relações não lineares e interações complexas entre as variáveis.**

## 1.4 OBJETIVOS GERAL E ESPECÍFICOS

### 1.4.1 OBJETIVO GERAL

Avaliar e comparar a eficiência de diferentes eixos na geração de emprego e renda no Estado do Amapá, utilizando técnicas de modelagem preditiva e dados oficiais fornecidos pelo governo local. Busca-se, assim, oferecer evidências empíricas que auxiliem na otimização da alocação de recursos públicos, incrementando o impacto socioeconômico das obras.

## 1.4.2 OBJETIVOS ESPECÍFICOS

**a) Criar métricas de eficiência (emprego e renda) por valor investido:**

Definir indicadores quantitativos que relacionem os recursos financeiros aportados em cada obra com os benefícios gerados, especialmente em termos de postos de trabalho e receita local.

**b) Aplicar algoritmos de Machine Learning (Random Forest, GBM, Redes Neurais, XGBoost) para prever e explicar a eficiência:**

Explorar modelos de aprendizado de máquina para identificar padrões relevantes e estimar de forma robusta o potencial de criação de emprego e renda, considerando as características particulares de cada projeto.

**c) Avaliar a importância das variáveis (feature importance) e analisar clusters de projetos:**

Determinar quais atributos (eixo de atuação, estágio da obra, localização, etc.) exercem maior influência sobre o desempenho na geração de emprego e renda.

Utilizar métodos de agrupamento (K-Means, DBSCAN) para investigar se existem grupos de projetos com perfis semelhantes de eficiência.

**d) Explorar metodologias adicionais, como Dif-in-Dif e Stochastic Frontier Analysis, para investigar a robustez dos achados:**

Conduzir análises complementares que possam aprofundar a compreensão da eficiência, considerando aspectos temporais (Dif-in-Dif) e fronteiras de produção (SFA) em obras públicas.

**e) Sugerir recomendações para melhorar a alocação de recursos públicos:**

Propor diretrizes baseadas nos resultados obtidos, de modo a maximizar os retornos socioeconômicos, orientando a elaboração, o monitoramento e a avaliação das políticas públicas estaduais.

## 1.5 DELIMITAÇÃO DO ESCOPO DO ESTUDO

O presente estudo concentra-se na análise de dados oficiais fornecidos pelo Governo do Amapá, abrangendo especificamente obras e projetos de infraestrutura executados ou planejados no estado. Trata-se de um recorte que permite avaliar, de forma sistemática, as informações relativas ao valor global de cada empreendimento, geração de empregos, receitas criadas ao longo do projeto e outras variáveis estratégicas para a compreensão de sua eficiência socioeconômica. Ao delimitar o escopo a esse conjunto de obras públicas, busca-se aprofundar a investigação no setor que, em muitos casos, responde pela maior parcela dos investimentos regionais e apresenta maior potencial de gerar externalidades positivas, como ampliação da mobilidade ou ampliação da oferta de serviços básicos.

Todavia, essa delimitação acarreta algumas limitações, em especial no que diz respeito à ocorrência de valores ausentes, zeros monetários e à existência de obras paralisadas ou em estágio inicial. Em determinados registros, o valor investido pode constar como zero, não refletindo necessariamente a inexistência de recursos, mas podendo indicar problemas de registro ou falta de atualização na base de dados. Além disso, a paralisação de obras afeta a capacidade de mensurar os reais benefícios do projeto, resultando em possíveis distorções nas métricas de eficiência. Esses fatores impõem cautela na interpretação dos resultados e reforçam a importância de um pré-processamento rigoroso dos dados, bem como a adoção de estratégias de imputação ou exclusão criteriosa de registros incompletos.

## 1.6 JUSTIFICATIVA DO TEMA: RELEVÂNCIA E CONTRIBUIÇÃO

A escolha deste tema está pautada tanto pela sua relevância prática para a gestão pública do Estado do Amapá quanto pela oportunidade de avançar no debate acadêmico em torno da avaliação de políticas públicas utilizando métodos quantitativos avançados. Sob a ótica gerencial, o estudo oferece subsídios diretos aos tomadores de decisão, que podem se orientar pelas evidências empíricas ao priorizar obras de maior retorno socioeconômico. Em um contexto de restrição orçamentária e necessidades múltiplas, a informação qualificada sobre a eficácia de diferentes eixos de investimento torna-se fundamental

para ampliar a accountability e a transparência da administração pública.

Do ponto de vista acadêmico, a dissertação contribui ao aplicar técnicas de Machine Learning e métodos estatísticos em um cenário regional pouco explorado, demonstrando como esses instrumentos analíticos podem ser adaptados às especificidades de uma realidade local marcada por desafios logísticos e institucionais. Essa abordagem possibilita o aprofundamento da literatura sobre avaliação de eficiência em obras públicas, reforçando a pertinência de análises empíricas que relacionem os montantes investidos aos benefícios efetivamente gerados em termos de emprego, renda e crescimento econômico regional.

## 1.7 ORGANIZAÇÃO DO ESTUDO

Além desta breve introdução, o trabalho encontra-se estruturado em quatro capítulos principais, seguidos das referências bibliográficas:

- a) Capítulo 2 – Referencial Teórico:** apresenta os alicerces conceituais e metodológicos que embasam a pesquisa. Discute-se o papel dos investimentos públicos no desenvolvimento regional, bem como as abordagens estatísticas e de Machine Learning aplicadas à avaliação de eficiência em obras. A revisão da literatura nacional e internacional complementa a fundamentação, destacando estudos empíricos relacionados ao tema.
- b) Capítulo 3 – Metodologia:** descreve o tipo de pesquisa realizada e aprofunda a discussão sobre as variáveis e fontes de dados empregados. Explica-se em detalhe o processo de pré-processamento, limpeza e transformação dos dados, além dos algoritmos de Machine Learning escolhidos (Random Forest, Gradient Boosting, XGBoost e Redes Neurais) e das técnicas de agrupamento (K-Means, DBSCAN). Nesta seção também se abordam as estratégias de validação dos modelos, as métricas de desempenho e as ferramentas de software utilizadas, sustentando o rigor metodológico necessário para o alcance dos objetivos propostos.
- c) Capítulo 4 – Resultados e Discussão:** apresenta a análise exploratória dos dados (EDA), expondo estatísticas descritivas e correlações relevantes para a compreensão do conjunto

analisado. Em seguida, exibe o desempenho dos modelos preditivos, com foco na comparação de suas métricas (MAE, RMSE, R<sup>2</sup>) e na interpretação da importância das variáveis (feature importance). Também são discutidos os achados do agrupamento dos projetos (K-Means, DBSCAN), seguidos de uma análise de sensibilidade e robustez, além das implicações desses resultados para a formulação de políticas públicas. Finalmente, estabelece um diálogo entre os resultados obtidos e a literatura, oferecendo uma discussão crítica acerca das possibilidades e limitações das técnicas adotadas.

d) **Capítulo 5 – Considerações Finais:** retoma os objetivos do estudo e sintetiza as conclusões alcançadas, evidenciando as principais contribuições do trabalho para o debate sobre eficiência em obras públicas no Amapá. Aborda igualmente as limitações metodológicas e empíricas encontradas ao longo do percurso e sugere caminhos para pesquisas futuras, bem como recomendações de natureza prática que possam orientar uma alocação de recursos mais eficaz por parte do governo estadual. Por fim, o capítulo enfatiza a importância de se perpetuar a cultura de avaliação empírica das políticas públicas, contribuindo para a evolução do conhecimento na área de economia aplicada e gestão governamental.

Após esses capítulos, apresentam-se as **Referências Bibliográficas**, compiladas conforme as normas técnicas (ABNT ou APA), contemplando artigos, livros e demais fontes recentes que sustentam teoricamente o estudo, possibilitando a verificação, a ampliação e o aprofundamento dos argumentos desenvolvidos.



## 2

## REFERENCIAL TEÓRICO

**2.1 DESENVOLVIMENTO ECONÔMICO REGIONAL E INVESTIMENTOS PÚBLICOS**

A discussão sobre desenvolvimento econômico regional remonta a teorias clássicas que buscam compreender de que maneira a formação de polos industriais, a distribuição de recursos produtivos e as estratégias de investimento público podem estimular o crescimento de determinadas localidades. Na perspectiva de Perroux (1950), o desenvolvimento não se propaga de maneira homogênea pelo território, mas concentra-se em “pólos de crescimento” que irradiam efeitos de arrasto (spread effects) e, simultaneamente, podem ocasionar desequilíbrios ou impactos negativos (backwash effects) em regiões vizinhas (MYRDAL, 1957). Sob essa ótica, o papel do Estado seria o de atenuar as assimetrias geradas por esses polos dinâmicos, sobretudo por meio de políticas de infraestrutura que facilitem a circulação de bens, serviços e mão de obra.

Segundo Hirschman (1958), o investimento governamental em infraestrutura constitui um fator-chave para desencadear encadeamentos produtivos (forward and backward linkages) capazes de expandir a capacidade industrial e promover, ao mesmo tempo, a integração regional. Essa concepção valoriza os efeitos multiplicadores resultantes de grandes obras, como a construção de estradas, a melhoria de portos, a ampliação de redes de eletricidade e saneamento, entre outras. Ao dinamizar setores estratégicos, cria-se um ambiente mais favorável ao desenvolvimento, uma vez que a infraestrutura adequadamente planejada possibilita a redução de custos logísticos e promove maior competitividade econômica (GRABOWSKI, 2016).

Para além das teorias fundadoras, autores contemporâneos reforçam a necessidade de o investimento público em infraestrutura ser complementado por estratégias de governança e planejamento que assegurem a eficiência na alocação dos recursos (GRAMLICH, 1994; CALDERÓN; SERVÉN, 2014). Em economias emergentes, como o Brasil, essas políticas ganham ainda maior relevância quando observamos a existência de regiões historicamente desfavorecidas por limitações de

acesso ao capital, à educação e a mercados consumidores (IPEA, 2020). No caso específico do Norte brasileiro, a logística de transporte fluvial, a carência de vias terrestres pavimentadas e a dispersão populacional representam desafios que exigem uma atuação planejada para otimizar o uso do orçamento público.

Sob a ótica do desenvolvimento regional, a infraestrutura atua como um catalisador do progresso local ao ampliar a conectividade entre áreas rurais e urbanas, estimular a criação de empregos e viabilizar a atração de investimentos privados. Ademais, conforme defendem autores como Lima e Matos (2019), o Estado pode exercer o papel de indutor de políticas voltadas à correção de disparidades intra e inter-regionais, implementando obras que reduzam a vulnerabilidade de municípios mais isolados e potencializem a inclusão produtiva da população local.

Nesse sentido, os investimentos públicos em infraestrutura tornam-se essenciais não apenas para suprir demandas pontuais, mas também para sustentar projetos de desenvolvimento de médio e longo prazos. Ao fomentar melhorias na mobilidade, no fornecimento de energia ou no saneamento básico, cria-se um ambiente propício à instalação de novas atividades econômicas, o que amplia a base de arrecadação tributária e gera efeitos virtuosos na geração de emprego e renda (BRITO; SILVEIRA, 2021). No entanto, para que tais melhorias sejam efetivas, é necessária a articulação entre diferentes políticas setoriais e níveis de governo, de modo que a alocação de recursos se dê de forma coordenada, transparente e orientada por critérios de eficiência (IPEA, 2020).

Em suma, o debate teórico evidencia que a infraestrutura governamental desempenha papel central no desenvolvimento regional ao reduzir custos de transação, melhorar a competitividade local e impulsionar efeitos multiplicadores que se disseminam na economia. Todavia, a eficácia desses investimentos depende de sua capacidade de atender às especificidades do território, bem como de mecanismos de governança e de avaliação que assegurem sua adequação às necessidades da população. Nesse ponto, a análise empírica, aliada a modelos de avaliação de eficiência, contribui para aferir em que medida cada projeto ou eixo de investimento está, de fato, promovendo o progresso socioeconômico das regiões atendidas.

## 2.2 MÉTODOS QUANTITATIVOS NA AVALIAÇÃO DE EFICIÊNCIA EM OBRAS PÚBLICAS

A avaliação de eficiência em obras públicas consiste, em sentido geral, na análise da relação entre os insumos (inputs) empregados em um projeto e os resultados (outputs) efetivamente obtidos (COELLI et al., 2005). Essa abordagem, no âmbito da economia do setor público, revela-se crucial para a formulação de políticas que visam alocar recursos de maneira mais assertiva, sobretudo em contextos de restrição orçamentária. Quando o objeto de estudo concentra-se na geração de emprego e renda a partir de determinado montante de investimento, tal análise demanda a mensuração adequada de variáveis como **valor\_global** (custo total da obra), **empregos\_gerados** (postos de trabalho criados) e **receitas** decorrentes ou associadas ao projeto.

Tradicionalmente, indicadores de eficiência podem ser construídos como razões simples entre outputs e inputs (FÄRE; GROSSKOPF; LOVELL, 1994). Exemplos incluem:

- a) empregos\_gerados / valor\_global, para aferir quantos postos de trabalho foram criados por real investido;**
- b) receita\_gerada / valor\_global, para indicar a capacidade de gerar recursos em função do capital alocado. Essas razões proporcionam uma visibilidade inicial da capacidade de cada obra em converter recursos em resultados socioeconômicos.**

Para avançar além dos indicadores simples, a literatura oferece diferentes métodos quantitativos que permitem avaliar a eficiência de forma mais detalhada. Entre as abordagens estatísticas tradicionais, destacam-se:

### 1. Análise Envoltória de Dados (Data Envelopment Analysis – DEA):

**Método não paramétrico que compara unidades de decisão (obras ou projetos, no caso) e verifica quão próxima cada unidade está da “fronteira de eficiência”, dada por uma combinação linear de unidades consideradas eficientes (CHARNES; COOPER; RHODES, 1978). Embora poderosa para comparar múltiplas entradas e saídas, a DEA exige**

homogeneidade das unidades analisadas e não realiza diretamente inferência estatística (COELLI et al., 2005).

## 2. Modelos de Fronteira Estocástica (Stochastic Frontier Analysis) – SFA:

Tratam a ineficiência como um termo aleatório, possibilitando aproximar-se de uma estimativa de uma função de produção ou custo e, posteriormente, determinar quanto distante cada unidade se encontra dessa fronteira (AIGNER; LOVELL; SCHMIDT, 1977). Tais modelos assumem hipóteses sobre a distribuição do erro de ineficiência, permitindo análise de variáveis explicativas associadas ao desempenho. A aplicação de SFA em obras públicas, entretanto, demanda uma definição clara do output (p. ex., empregos gerados) e do input (valor investido, tempo de obra, etc.), bem como um volume significativo de dados que dê suporte à estimativa de parâmetros (KUMBHAKAR; LOVELL, 2003).

## 3. Regressões Múltiplas (OLS e variantes):

Apesar de não serem métodos de eficiência propriamente ditos, modelos lineares estimados por Mínimos Quadrados Ordinários ou métodos correlatos podem relacionar diretamente os outputs (empregos, renda) ao montante investido e outras variáveis explicativas (localização, tipo de objeto, duração do projeto). A simplicidade interpretativa desse método o torna popular em avaliações iniciais, embora possa subestimar relações não lineares ou efeitos de multicolinearidade (WOOLDRIDGE, 2012).

Nos últimos anos, a crescente disponibilidade de dados governamentais e o maior poder computacional estimularam a adoção de **métodos quantitativos mais sofisticados** (BREIMAN, 2001; HASTIE; TIBSHIRANI; FRIEDMAN, 2009). No contexto deste estudo, que propõe a modelagem preditiva da eficiência em diferentes eixos de obras públicas, as técnicas de Machine Learning — a exemplo de Random Forest, Gradient Boosting e Redes Neurais — viabilizam a detecção de padrões complexos e não lineares, e permitem lidar com variáveis categóricas, outliers e interações (VARIAN, 2014). Tais métodos, em combinação com procedimentos de validação cruzada (cross-validation) e ajustes de hiperparâmetros, buscam tanto a melhoria na acurácia preditiva quanto a captura de relações que métodos puramente lineares dificilmente apreenderiam (ATHEY, 2018).

No entanto, a adoção de métodos quantitativos para mensurar a eficiência em obras públicas no Brasil e, especificamente, na região Norte, enfrenta barreiras quanto à **qualidade e completude dos dados** (SOUZA; DUTRA, 2017). Registros com valores monetários faltantes ou a ocorrência de obras paralisadas requerem protocolos de limpeza, imputação de dados ou exclusão criteriosa de observações para que os resultados não sejam indevidamente distorcidos (COELLI et al., 2005). Além disso, projetos heterogêneos em objetivo e escopo podem demandar definições cuidadosas de “output”, evitando, por exemplo, comparar diretamente a geração de empregos em uma obra de saneamento com a de uma ponte voltada a fluxo interestadual, sem controlar por variáveis contextuais (FRIEDMAN, 2001).

Outro aspecto fundamental é a seleção de indicadores capazes de captar os **benefícios adicionais** trazidos pelas obras (RATLEDGE et al., 2021), seja em termos de dinamização econômica (receita gerada na região), melhoria de serviços públicos (saneamento, eletricidade) ou inclusão produtiva de populações vulneráveis. Alguns estudos empregam indicadores compostos, somando pontuações de diferentes dimensões (RODRIGUES; SILVA, 2020). A dificuldade encontra-se em ponderar cada dimensão de maneira coerente (CHI et al., 2021) com o impacto buscado — desafio que motiva a experimentação com métodos de aprendizado de máquina, os quais podem internamente ajustar pesos às variáveis conforme sua contribuição para explicar a variável-alvo (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

No presente trabalho, a opção por combinar indicadores de eficiência clássicos (por exemplo, empregos gerados por real investido) com algoritmos preditivos (Random Forest, Gradient Boosting, Redes Neurais) visa melhorar a precisão na identificação dos eixos (ou subconjuntos de projetos) mais eficazes em promover emprego e renda. Dessa forma, os **métodos quantitativos** não apenas sustentam a avaliação inicial de desempenho, como também possibilitam a realização de simulações e projeções — a exemplo do uso de modelos ARIMA ou Prophet para explorar tendências na evolução de certos indicadores, ainda que em caráter exploratório (ATHEY, 2018). Tal conjugação de técnicas torna-se especialmente relevante para o contexto do Amapá, pois a regionalização das necessidades e a variação na natureza das obras exigem análises flexíveis, que contemplam diferentes aspectos do gasto público e do retorno econômico-social.

Em síntese, a aplicação de métodos quantitativos à avaliação de eficiência em obras públicas representa uma abordagem necessária para embasar decisões de investimento e, de modo mais amplo, para aprimorar a transparência e a accountability na gestão governamental. Seja por meio de indicadores simples, seja por meio de modelos estatísticos e de aprendizado de máquina mais elaborados, o objetivo permanece o de quantificar de forma objetiva o impacto do investimento, permitindo uma comparação justa e informada entre diversos projetos e eixos de atuação.

## 2.3 MACHINE LEARNING NA ANÁLISE DE POLÍTICAS PÚBLICAS

O uso de Machine Learning (ML) em avaliações de políticas públicas destaca-se como uma alternativa promissora para lidar com bases de dados extensas e heterogêneas, bem como para capturar relações não lineares e possíveis interações entre variáveis (ATHEY, 2018). Na maioria dos contextos governamentais, os dados coletados costumam apresentar uma ampla diversidade de formatos e graus de completude, com colunas que variam desde valores monetários até informações qualitativas relativas a localizações geográficas ou fases de execução de um projeto. Esse cenário demanda algoritmos capazes de processar variáveis categóricas e numéricas, identificar padrões intrincados e, ao mesmo tempo, robustecer-se contra ruídos e valores discrepantes (BREIMAN, 2001).

Na seara dos investimentos públicos em infraestrutura, a adoção de métodos de ML possibilita a criação de **modelos preditivos** que relacionam, por exemplo, o *valor global* de cada obra a outputs como *empregos gerados* ou *receita gerada*. Ao fazê-lo, tais algoritmos podem descobrir configurações específicas que maximizem a eficiência, apontando quais eixos, características de projeto ou combinações de fatores resultam em maior retorno socioeconômico (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Em termos práticos, isso significa oferecer subsídios empíricos aos gestores, para que direcionem recursos a iniciativas de maior impacto, priorizando eixos e obras com potencial de gerar mais empregos e renda, sobretudo em localidades com maior carência.

Nesse contexto, **modelos de ensemble** (por exemplo, Random Forest, Gradient Boosting, XGBoost) agregam múltiplas árvores de

decisão e apresentam bom desempenho preditivo em bases complexas, pois são menos suscetíveis ao sobreajuste do que uma única árvore (BREIMAN, 2001). Além disso, a possibilidade de derivar *feature importance* desses modelos auxilia na identificação das variáveis que mais influenciam a eficiência, promovendo maior transparência e aplicabilidade do processo decisório (CHEN; GUESTRIN, 2016). Tais modelos tornam-se especialmente relevantes quando se busca responder a questões como “Quais eixos de obra produzem, em média, mais empregos por real investido?” ou “Quais variáveis, como a etapa do projeto ou se é terceirizado, impactam significativamente a capacidade de gerar renda local?”.

Outro grande potencial do ML para políticas públicas encontra-se nos **métodos de agrupamento** (clustering), a exemplo de K-Means e DBSCAN. Essas técnicas podem agrupar obras com características similares, possibilitando, por exemplo, segmentar projetos em clusters de elevada eficiência, média eficiência e baixa eficiência (WITTE et al., 2020). Assim, gestores são municiados de informações que indicam potenciais grupos de obras com padrões de valor investido, região e retorno socioeconômico convergentes, o que viabiliza tanto a disseminação de *boas práticas* quanto o redesenho de políticas em áreas subótimas de desempenho.

Ainda que possuam vantagens evidentes, as metodologias de aprendizado de máquina também exigem cuidados. Um primeiro ponto é assegurar a **qualidade dos dados**, pois, sem um pré-processamento criterioso — que inclua a imputação de valores faltantes, a checagem de consistência e a normalização das variáveis —, os modelos podem produzir estimativas viesadas. Segundo, destaca-se a necessidade de técnicas adequadas de validação (por exemplo, *cross-validation*) e de ajustes de hiperparâmetros (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Sem esses procedimentos, há o risco de sobreajuste (*overfitting*), no qual o modelo se adequa em demasia às peculiaridades da amostra, perdendo capacidade de generalização.

Além disso, **aspectos interpretativos** permanecem centrais para a análise de políticas públicas (DELGADO & LEMOS, 2023). Diferentemente de um cenário puramente comercial, em que a predição de rótulos ou valores pode ser suficiente, as decisões governamentais demandam uma compreensão mais profunda de como cada variável interage para explicar os resultados (ATHEY, 2018). Nesse sentido, a incorporação de métodos de interpretabilidade (por

exemplo, SHAP values) ou de métricas de *feature importance* mostrase fundamental para que os formuladores de políticas compreendam os fatores que alavancam ou limitam a eficiência das obras (MOLNAR, 2019).

No conjunto, o potencial de Machine Learning na análise de políticas públicas advém de sua capacidade de:

- 1. Tratar grandes volumes de dados e escalas heterogêneas;**
- 2. Capturar padrões complexos, muitas vezes não identificáveis por métodos lineares;**
- 3. Fornecer previsões robustas que podem subsidiar a priorização de projetos e a elaboração de cenários prospectivos;**
- 4. Gerar insights interpretáveis sobre as variáveis determinantes da eficiência, aprimorando a gestão pública.**

No presente estudo, a aplicação de ML busca explorar a relação entre investimentos públicos (valor investido, eixos de atuação, natureza estratégica ou terceirizada do projeto etc.) e outputs de interesse (geração de emprego e renda). A expectativa é que a adoção de algoritmos de classificação e regressão avançados, associada a técnicas de *validation* e de *feature importance*, ofereça uma avaliação mais acurada e confiável da eficiência de cada obra no Estado do Amapá. Dessa forma, é possível fomentar um processo decisório mais orientado a resultados, compatível com as demandas contemporâneas de transparência e efetividade na alocação de recursos públicos (VARIAN, 2014).

## **2.4 MODELAGEM ESTATÍSTICA VS. MACHINE LEARNING: VANTAGENS E LIMITAÇÕES**

A comparação entre modelos estatísticos tradicionais, como a Regressão Linear Ordinária (OLS), e técnicas de Machine Learning (ML) constitui um tópico central na avaliação de eficiência em obras públicas e infraestrutura. A abordagem estatística clássica, exemplificada pela OLS, parte do pressuposto de relações lineares entre as variáveis explicativas e a variável dependente, permitindo inferências estatísticas relativamente simples e claras, com base em hipóteses como homocedasticidade e normalidade dos resíduos (WOOLDRIDGE, 2013). Essa interpretabilidade é uma vantagem relevante: no contexto de políticas públicas, compreender o efeito marginal de cada variável,

como *valor\_global* ou *regiao*, reveste-se de utilidade prática na formulação de intervenções (GREENE, 2012).

Entretanto, a OLS mostra-se menos flexível quando enfrenta dados heterogêneos ou relações não lineares, situação comum em bases de investimentos públicos, cujos registros podem conter outliers (valores muito elevados ou muito baixos) e múltiplas variáveis categóricas. Em tais casos, métodos de ML emergem como alternativas competitivas, oferecendo maior robustez a desvios dos pressupostos clássicos e capacidade de capturar interações complexas (ATHEY, 2018). Modelos de ensemble, como Random Forest, Gradient Boosting e XGBoost, por exemplo, combinam várias árvores de decisão para reduzir o viés e/ou a variância da estimação (BREIMAN, 2001; CHEN; GUESTRIN, 2016). Já as Redes Neurais Artificiais (RNA) modelam relações não lineares em diversos níveis, ainda que, muitas vezes, em detrimento da interpretabilidade (HAYKIN, 2009).

No tocante ao **overfitting**, a gestão pública exige parcimônia na construção de modelos para evitar que as políticas sejam baseadas em padrões que não se sustentam fora da amostra estudada. Embora algoritmos de ML sejam potentes para ajustar dados complexos, a carência de mecanismos estatísticos tradicionais de inferência (como intervalos de confiança imediatos) pode dificultar a avaliação de significância. Consequentemente, são necessárias técnicas como *cross-validation* e ajuste de hiperparâmetros (HASTIE; TIBSHIRANI; FRIEDMAN, 2009) para assegurar que o modelo retenha um desempenho consistente quando aplicado a novas observações.

Por fim, **interpretabilidade** constitui uma limitação inerente a muitas técnicas de ML. Enquanto métodos estatísticos lineares oferecem coeficientes diretamente interpretáveis, modelos baseados em árvore e redes neurais, em geral, requerem ferramentas específicas (por exemplo, *feature importance*, SHAP values ou LIME) para esclarecer por que determinadas variáveis têm mais peso do que outras (MOLNAR, 2019). Na formulação de políticas, essa transparência é crucial para justificar decisões e angariar legitimidade junto à sociedade e aos órgãos de controle.

Assim, **tanto OLS quanto ML** dispõem de vantagens e desvantagens. A conjugação de ambas as abordagens — por exemplo, testando primeiramente hipóteses com modelos estatísticos e, em seguida, buscando maior poder preditivo com algoritmos de ML —

tende a produzir análises mais sólidas, respeitando, ao mesmo tempo, requisitos de robustez, interpretabilidade e bom desempenho preditivo.

## 2.5 MODELOS DE MACHINE LEARNING APLICADOS À AVALIAÇÃO DE EFICIÊNCIA

Os modelos de Machine Learning podem ser implementados de diversas formas para aproximar-se de uma estimativa de eficiência de investimentos públicos em infraestrutura, seja no intuito de prever impactos econômicos, seja para identificar grupos de projetos com características semelhantes. Entre as principais técnicas empregadas neste estudo, destacam-se (i) modelos baseados em árvores, (ii) métodos de agrupamento e (iii) redes neurais artificiais.

### 2.5.1 MODELOS BASEADOS EM ÁRVORES: RANDOM FOREST, GRADIENT BOOSTING E XGBOOST

Os algoritmos de árvore de decisão (Decision Trees) possuem a vantagem de lidar bem com **outliers**, variáveis categóricas e dados faltantes (BREIMAN, 2001). A lógica de particionar os dados com base em critérios sucessivos de redução de heterogeneidade (como o Gini ou o desvio padrão dos resíduos) resulta em um modelo interpretável no nível individual de cada árvore. Contudo, árvores isoladas podem sofrer de alto viés ou alto erro de variância, dependendo do critério de profundidade e podas.

Para contornar tais limitações, **modelos ensemble** — tais como **Random Forest** e **Gradient Boosting** — combinam múltiplas árvores para melhorar a robustez e a acurácia (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Em Random Forest, a ideia central é gerar diversas árvores, cada uma treinada em subamostras aleatórias, e então agregar suas previsões por meio de votações ou médias. Esse procedimento tende a atenuar a sensibilidade a outliers ou a correlações espúrias. Já o Gradient Boosting constrói árvores de maneira sequencial, a fim de corrigir erros das etapas anteriores, resultando em alta capacidade de modelar padrões complexos. O **XGBoost**, por sua vez, é uma implementação otimizada do Gradient Boosting, sendo reconhecido por sua eficiência computacional e alto desempenho em competições de ML (CHEN; GUESTRIN, 2016).

Em termos de avaliação de eficiência de obras públicas, esses modelos apresentam vantagens ao incluir, simultaneamente, variáveis numéricas (valor investido, custos adicionais, receita gerada), além de variáveis categóricas (eixo, etapa, núcleo de governo). Esses algoritmos toleram também a falta de linearidade nos dados, tornando-se adequados para prever indicadores de eficiência, tais como empregos gerados por real investido ou receita gerada em relação ao valor de investimento.

## 2.5.2 MODELOS DE AGRUPAMENTO: K-MEANS E DBSCAN

Além de prever a eficiência de modo direto, é igualmente relevante identificar **padrões** ou **clusters** de projetos com perfis semelhantes. Nesse cenário, métodos não supervisionados, como **K-Means** e **DBSCAN**, são empregados para agrupar obras conforme similaridade de custo, número de empregos gerados, renda auferida no entorno e demais variáveis relevantes (WITTE et al., 2020).

### 1. K-Means

O K-Means baseia-se em aproximar-se de uma estimativa de centroides para um número  $k$  de clusters, minimizando a soma das distâncias quadráticas entre cada ponto (projeto) e seu respectivo centroide (HARTIGAN; WONG, 1979). Apesar de ser simples e eficiente para conjuntos de dados de tamanho moderado, o K-Means requer a definição prévia de  $k$ , podendo ser pouco robusto à presença de outliers e de formas não esféricas nos agrupamentos.

### 2. DBSCAN

O DBSCAN (Density-Based Spatial Clustering of Applications with Noise) funciona por identificação de regiões densas de pontos, agrupando-os conforme um parâmetro de raio (eps) e número mínimo de pontos (minPts) (ESTER et al., 1996). Ao contrário do K-Means, não há necessidade de definir  $k$  de antemão, e o método consegue isolar valores atípicos (outliers) como ruído, conferindo-lhe maior flexibilidade em dados heterogêneos. Isso é particularmente útil em bases governamentais, em que podem existir poucos projetos de grande vulto e muitos de pequeno porte, modificando a distribuição geral.

Aplicar esses algoritmos de clusterização aos dados de obras públicas viabiliza a criação de typologies de projetos, permitindo que

gestores identifiquem grupos de obras com alta eficiência e baixo custo, bem como grupos que requerem atenção (eficiência baixa, custo elevado, etc.). Dessa forma, potenciais gargalos ou oportunidades de replicar boas práticas tornam-se mais visíveis, subsidiando a priorização de recursos e políticas.

### 2.5.3 REDES NEURAIS ARTIFICIAIS PARA PREDIÇÃO DE IMPACTO

As **Redes Neurais Artificiais (RNAs)** constituem outra vertente relevante de ML para problemas de previsão e classificação. Inspiradas no funcionamento de neurônios biológicos, as RNAs são compostas por camadas de nós (neurônios artificiais) interligados, o que lhes permite capturar relações não lineares complexas entre as variáveis (HAYKIN, 2009). Em estudos de impacto econômico, por exemplo, as RNAs podem auxiliar na predição de renda ou empregos gerados a partir de diversas covariáveis relacionadas aos projetos de infraestrutura.

Entretanto, **três desafios** caracterizam o uso de RNAs na gestão pública:

**Tuning:** a escolha do número de camadas e neurônios, bem como de parâmetros como taxa de aprendizagem e regularização, pode demandar sucessivos testes (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Sem um processo rigoroso de validação cruzada, há risco significativo de sobreajuste.

**Interpretabilidade:** redes neurais são, em geral, consideradas modelos de “caixa-preta” (black-box). Embora existam métodos de explicabilidade (por exemplo, *feature importance*, SHAP, LIME), ainda assim o processo de decisão interno pode ser complexo. Em ambiente público, isso pode dificultar a transparência e a fundamentação técnica das escolhas junto a órgãos de controle (MOLNAR, 2019).

**Recursos Computacionais:** RNAs, em especial aquelas com múltiplas camadas (deep learning), podem exigir maior poder de processamento e memória, tornando-se um obstáculo em cenários de bases governamentais muito extensas ou com variáveis de alta dimensionalidade (GOODFELLOW; BENGIO; COURVILLE, 2016).

A despeito desses pontos, as redes neurais destacam-se pela elevada capacidade de modelagem não linear e pelo potencial de ajustar dados mesmo em contextos com complexa interação de

variáveis, o que frequentemente ocorre em análises de políticas públicas. Quando integradas a procedimentos de validação sólidos e interpretabilidade parcial, as RNAs podem oferecer ganhos notáveis na acurácia preditiva e, consequentemente, identificar padrões mais ricos acerca da geração de emprego e renda por valor investido.

Em síntese, cada abordagem — sejam os modelos de árvore, os algoritmos de agrupamento ou as redes neurais artificiais — desempenha papel específico na avaliação de eficiência de projetos públicos. A escolha do método (ou conjunto de métodos) deve levar em conta tanto a natureza dos dados (presença de outliers, variáveis categóricas ou numéricas, dimensionalidade, etc.) quanto os objetivos analíticos (previsão, segmentação, explicação). Dessa maneira, a aplicação combinada dessas técnicas favorece uma compreensão multifacetada dos investimentos em obras e infraestrutura no Estado do Amapá, contribuindo para subsidiar decisões de alocação de recursos mais eficazes.

## 2.6 ESTUDOS EMPÍRICOS RELACIONADOS

A aplicação de técnicas de Machine Learning ao contexto de infraestrutura e desenvolvimento regional tem se tornado cada vez mais frequente em estudos internacionais e nacionais, dada a capacidade desses métodos de lidar com grandes volumes de dados e identificar padrões de eficiência em projetos de obras públicas (MULLAINATHAN; SPIESS, 2017). Em particular, a busca por uma alocação mais eficaz dos recursos públicos motiva pesquisadores e gestores a empregar algoritmos que permitam não apenas prever resultados, como também estruturar classificações e agrupamentos de projetos em função de métricas como geração de emprego e renda.

### 2.6.1 LITERATURA INTERNACIONAL

No âmbito internacional, trabalhos que empregam modelos de Machine Learning para avaliar políticas públicas de infraestrutura têm exibido resultados promissores. Athey (2018) destaca que métodos de Random Forest e Gradient Boosting podem capturar interações complexas entre variáveis socioeconômicas e geográficas, superando abordagens estatísticas lineares na predição de outcomes de projetos. Segundo a autora, a robustez dos modelos baseados em árvores se evidencia em cenários de alta dimensionalidade, onde múltiplas

características (e.g., valor investido, tipologia da obra, indicadores de desenvolvimento humano) devem ser consideradas simultaneamente.

Estudos de *Chui et al.* (2018) apontam para a importância de algoritmos de **clustering** no contexto de infraestrutura, salientando que o K-Means e o DBSCAN facilitam a identificação de regiões com comportamento similar em termos de custos e benefícios econômicos, bem como a detecção de outliers que podem demandar tratamento diferenciado por parte dos gestores. *Chen e Guestrin* (2016), por sua vez, demonstram o potencial do XGBoost para lidar com grandes bases de dados, obtendo altos níveis de acurácia preditiva em aplicações que envolvem estimativas de eficiência e produtividade em diferentes setores, incluindo o setor público.

No que diz respeito a **análises de eficiência** propriamente ditas, pesquisas baseadas em Stochastic Frontier Analysis (SFA) também têm sido aprimoradas com ferramentas de ML, sobretudo para etapas de pré-processamento e identificação de segmentos de dados homogêneos (KUMBHAKAR; WANG; HORNCastle, 2015). Embora a estimação econômétrica de fronteira seja tradicional, estudos recentes incorporam variáveis obtidas por algoritmos de ML (por exemplo, *feature engineering*) para refinar a medição de eficiência.

Além disso, organizações internacionais, como o *World Bank*, têm explorado regressões com Machine Learning para avaliar a eficácia de investimentos em infraestrutura de transportes e saneamento, em países em desenvolvimento (WORLD BANK, 2020). Nesses levantamentos, Random Forest e Gradient Boosting são frequentemente citados como técnicas úteis para identificar determinantes de sucesso ou atraso em obras, gerando subsídios para ajustes de políticas públicas.

## 2.6.2 LITERATURA NACIONAL

No contexto brasileiro, observa-se crescente interesse em investigar a eficiência de investimentos públicos por meio de abordagens quantitativas mais sofisticadas (IPEA, 2019). Embora ainda prevaleçam estudos de corte econômétrico tradicional, algumas iniciativas passam a empregar métodos de Machine Learning, notadamente em universidades e órgãos de pesquisa aplicada.

Por exemplo, discute-se a adoção de algoritmos de árvore de decisão e Redes Neurais Artificiais para prever a taxa de conclusão de obras de infraestrutura, considerando dados orçamentários, regionais e de execução física. Em seus resultados, métodos baseados em árvores se destacaram pela capacidade de acomodar outliers comuns nas licitações públicas brasileiras, bem como pelo maior *explanatory power* ao incorporar variáveis categóricas como “tipo de financiamento” e “esfera de governo responsável”.

Em se tratando de clusterização, o emprego de K-Means é sugerido para agrupar municípios de uma determinada região brasileira, considerando indicadores de receita gerada, número de obras concluídas e volume de investimento. A análise permitiu visualizar padrões de eficiência entre municípios, evidenciando gargalos em localidades com alto custo e baixa geração de empregos. Nesse sentido, a capacidade dos algoritmos de agrupar observações semelhantes auxilia no direcionamento de recursos para regiões mais carentes ou, em contrapartida, no fomento de polos produtivos onde a eficiência se encontra acima da média.

Ainda que a maior parte dos estudos nacionais não se volte exclusivamente para a mensuração de eficiência predial ou de obras, há evidências de que a incorporação de técnicas de ML venha crescendo, sobretudo impulsionada pela disponibilidade de dados abertos em plataformas governamentais (BRASIL, 2021). Além disso, bancos de desenvolvimento e agências de fomento têm demonstrado interesse em técnicas de predição de impacto socioeconômico, associadas às bases de dados de projetos de infraestrutura (BNDES, 2020). Esse cenário aponta para uma convergência crescente entre a análise econômica convencional e o uso de algoritmos avançados, com vistas a subsidiar a tomada de decisão no setor público.

No caso específico do Estado do Amapá, conforme a proposta desta dissertação, não se identificam ainda estudos empíricos robustos que empreguem ML para aferir a eficiência de investimentos em infraestrutura. Todavia, essa lacuna reforça a pertinência da presente pesquisa, que busca justamente demonstrar como técnicas preditivas e de agrupamento podem revelar insights valiosos acerca dos eixos de maior impacto, contribuindo para otimizar a alocação de recursos e promover desenvolvimento econômico regional de forma mais equilibrada.



## 3

## METODOLOGIA

## 3.1 LIMITAÇÕES E PREMISSAS METODOLÓGICAS

O presente estudo emprega indicadores **aproximativos** — empregos gerados por real investido, renda gerada por real investido e “impacto por projeto/eixo” — para sinalizar a **capacidade relativa** de diferentes eixos de obra converterem recursos públicos em benefícios socioeconômicos. Esses indicadores **não constituem medidas canônicas de eficiência** (como as obtidas por Data Envelopment Analysis ou Stochastic Frontier Analysis em bases plenamente homogêneas); tratam-se de **proxies construídas a partir de registros administrativos** disponibilizados pelo Governo do Amapá. A opção por utilizar tais proxies decorre, sobretudo, da **disponibilidade limitada de variáveis físicas e financeiras** na base original — situação comum em avaliações de infraestrutura em estados da Região Norte — e da necessidade de oferecer, ainda que de forma preliminar, evidências empíricas capazes de orientar a alocação orçamentária.

Do ponto de vista **informacional**, a base apresenta **três fragilidades centrais**. Primeiro, há elevada incidência de **valores ausentes ou igualados a zero** em campos críticos (especialmente *valor\_global* e *empregos\_gerados*), seja por atraso de alimentação dos sistemas, seja por divergências entre órgãos executores. Segundo, constatam-se **heterogeneidade de granularidade e inconsistência temporal**: alguns projetos registram custos totais enquanto outros discriminam apenas a fase corrente; certas obras têm cronogramas detalhados, ao passo que outras informam datas genéricas. Terceiro, o banco carece de **variáveis de contexto** (p.ex., densidade populacional local, nível de escolaridade, indicadores de pobreza), que ajudariam a controlar diferenças estruturais entre regiões e, assim, a refinar a comparação entre eixos.

Tais limitações impactam diretamente a **confiabilidade estatística** das inferências. A ausência de observações completas restringe o tamanho efetivo da amostra, aumentando a margem de erro dos modelos; a heterogeneidade de registro reduz a comparabilidade transversal; a falta de covariáveis contextuais pode introduzir **viés de omissão**, levando a super- ou sub-estimativas da

contribuição real de cada eixo. Em resposta, adotaram-se rotinas de limpeza, imputação conservadora e validação cruzada; ainda assim, recomenda-se interpretar os resultados como **evidências exploratórias**, não como “medidas finais” de desempenho.

No plano **metodológico**, os algoritmos de Machine Learning empregados privilegiaram **capacidade preditiva**, não **identificação causal**. Random Forest, Gradient Boosting e XGBoost capturam relações não lineares, mas não isolam efeitos de política pública do tipo “X causa Y”; tampouco corrigem por choques externos (ex.: variações macroeconômicas ou choques climáticos) que possam ter influenciado simultaneamente investimento e mercado de trabalho. Além disso, embora se tenham aplicado técnicas de validação (repeated cross-validation) para mitigar **overfitting**, permanece o risco de que padrões específicos da amostra não se reproduzam integralmente em novos ciclos de obras.

Há, ainda, **limitações de interpretabilidade**. Modelos ensemble fornecem rankings de importância de variáveis, mas não coeficientes diretamente comparáveis aos de regressões lineares tradicionais; logo, a magnitude dos efeitos deve ser lida com cautela. Ensaios econômétricos complementares — Diferença-em-Diferenças e SFA — foram conduzidos, mas apresentaram convergência instável e ganhos explicativos modestos, reflexo da qualidade dos dados e da ausência de séries temporais consistentes pré- e pós-intervenção. Esses resultados reforçam que, **até que bases mais ricas sejam consolidadas**, análises causais profundas permanecem limitadas.

Finalmente, cumpre destacar que **as conclusões não pretendem esgotar a discussão sobre eficiência**, mas **abrir caminho** para estudos futuros. Estudos com levantamentos de campo, variáveis de output físicas (por exemplo, quilômetros de estrada pavimentada, ligações domiciliares de saneamento) e controle rigoroso de fatores regionais poderão avançar para métricas clássicas de fronteira técnica, enquanto este trabalho se limita, conscientemente, ao nível de informação disponível. Assim, **qualquer recomendação de política derivada deste relatório deve ser encarada como indicativo**, sujeito a revisão quando novos dados ou métodos mais adequados estiverem acessíveis.

## 3.2 FONTE DE DADOS E VARIÁVEIS UTILIZADAS

Este tópico objetiva apresentar as bases de dados empregadas na pesquisa, bem como especificar as variáveis de entrada (inputs) e de saída (outputs) utilizadas na modelagem da eficiência dos investimentos públicos no Estado do Amapá. Conforme *Gujarati e Porter* (2011), a correta definição das fontes de dados e das variáveis é fundamental para garantir a coerência estatística e a robustez dos resultados.

### 3.2.1 FONTES DE DADOS PRIMÁRIOS E SECUNDÁRIOS

Para a presente dissertação, a principal base de dados é proveniente do **Governo do Amapá**, disponibilizada em formato eletrônico e contendo informações detalhadas sobre projetos de infraestrutura e obras públicas realizados no Estado. Esse banco oficial abrange variáveis administrativas e financeiras, como **valor investido, eixo de atuação (infraestrutura, educação, saúde etc.), etapa da obra, empregos gerados, receita oriunda do projeto**, entre outras.

Além dessa fonte principal, recorreu-se a relatórios adicionais emitidos por secretarias estaduais e pelos órgãos responsáveis pela execução orçamentária, os quais fornecem descrições complementares sobre eventuais paralisações de obras, status de conclusão e indicadores de impacto socioeconômico. Em alguns casos, informações suplementares foram confrontadas com dados de transparência pública para mitigar possíveis inconsistências.

A fundamentação metodológica para coleta e cruzamento de dados, bem como a sistematização de diferentes fontes, está em linha com recomendações de Wooldridge (2020), segundo as quais dados oficiais e relatórios governamentais costumam constituir o alicerce de estudos empíricos em avaliação de políticas públicas. Adicionalmente, o emprego de múltiplas fontes reforça a **triangulação dos dados**, ampliando a confiabilidade das conclusões (CRESWELL, 2014).

Os microdados analisados nesta dissertação foram cedidos pela Secretaria de Estado da Infraestrutura do Amapá (SEINF/AP). Pesquisadores interessados podem solicitar acesso formalmente ao gabinete do secretário de estado da infraestrutura do governo do Amapá, pelo [seinf@seinf.ap.gov.br](mailto:seinf@seinf.ap.gov.br), mediante apresentação de termo de confidencialidade e descrição do projeto de pesquisa.

### 3.2.2 DEFINIÇÃO DE INPUTS (VARIÁVEIS DE ENTRADA)

No tocante às variáveis de entrada, optou-se por selecionar aquelas consideradas relevantes para explicar a eficiência no uso dos recursos públicos, tendo como base a literatura de economia regional (KUMBHAKAR; WANG; HORNCastle, 2015) e estudos aplicados de avaliação de investimentos (GUJARATI; PORTER, 2011). Dentre as principais variáveis de entrada (inputs), destacam-se:

- 1. Eixo:** categoria em que o projeto se insere, podendo abranger infraestrutura viária, saneamento, segurança, saúde, educação, entre outros. Essa classificação permite comparar a eficiência gerada em diferentes frentes de investimento.
- 2. Ano:** período em que o recurso foi alocado ou a obra executada, possibilitando análises temporais e a aplicação de modelos que capturem variações interanuais.
- 3. Região:** localização da obra ou projeto dentro do Estado do Amapá, diferenciando, por exemplo, Macapá e outros municípios. Essa variável contribui para verificar se existem disparidades regionais significativas na geração de emprego ou renda.
- 4. Valor Global:** montante total investido em cada projeto, expresso em valores monetários. Trata-se de uma das variáveis mais importantes na modelagem, pois serve de base para o cálculo de indicadores de eficiência.
- 5. Etapa (Status da Obra):** estágio em que o projeto se encontra (execução, concluído, paralisado etc.). Buscando identificar a influência do status no desempenho, essa variável atua como potencial determinante de eficiência ou ineficiência.
- 6. Dummies de Projeto:** foram criadas variáveis binárias para indicar se o projeto é estratégico ou terceirizado, permitindo comparar eventuais diferenças de eficiência de acordo com a natureza do empreendimento (MULLAINATHAN; SPIESS, 2017).
- 7. Impacto por Projeto e Impacto por Eixo:** estimativas fornecidas no banco de dados oficial que buscam aferir a magnitude do retorno socioeconômico de cada iniciativa, podendo refletir em melhorias diretas (infraestrutura local) e indiretas (efeitos multiplicadores no entorno).

Essas variáveis de entrada compõem o conjunto preditor na aplicação de algoritmos de Machine Learning e análises econométricas, como regressões lineares, modelos de árvores e técnicas de agrupamento. A lógica da escolha deve-se à possibilidade de cada fator contribuir para a explicação das variáveis de saída, estabelecendo relações entre o valor investido, a natureza do projeto e os resultados obtidos em termos de emprego e renda.

### 3.2.3 DEFINIÇÃO DE OUTPUTS (VARIÁVEIS DE SAÍDA)

Para mensurar a eficiência dos investimentos em obras públicas, foram definidas como variáveis de saída (outputs) indicadores numéricos que associam a geração de emprego e renda ao valor investido:

- 1. Eficiência Emprego (emprego/valor):** proporção entre o total de empregos gerados e o valor global do projeto. Esse indicador visa captar quantos postos de trabalho, em média, são criados por unidade monetária investida. Assim, projetos com maior número de contratações, em relação ao custo, tendem a apresentar maior eficiência. Salienta-se que 'eficiência\_emprego' não corresponde a um conceito clássico de eficiência técnica derivado de DEA ou SFA; trata-se de uma proxy construída (empregos/valor) que oferece apenas uma sinalização preliminar do retorno em emprego.
- 2. Eficiência Renda (renda/valor):** relação entre a receita adicional (ou renda gerada) e o valor investido. A renda pode ocorrer durante a execução do projeto ou após a finalização da obra, abarcando, por exemplo, vendas de insumos, geração de negócios locais e incremento de arrecadação estadual. Tal indicador fornece evidências acerca da capacidade de cada real investido em promover retorno econômico mensurável (ATHEY, 2018).
- 3. Índice de Impacto (impacto\_por\_projeto ou impacto\_por\_eixo):** variável que sintetiza o efeito global de cada iniciativa em seu respectivo eixo, considerando externalidades positivas ou negativas detectadas no entorno da obra. Diferentemente das métricas de emprego e renda, esse indicador agrupa dimensões qualitativas e

**quantitativas fornecidas pelo órgão responsável, visando representar mudanças infraestruturais significativas.**

A seleção desses outputs está de acordo com a proposta de avaliação quantitativa de políticas públicas (GRILICHES, 1998), enfatizando a relação custo-benefício em termos de geração de emprego e de receita adicional. As métricas de eficiência (emprego/valor, renda/valor) e o índice de impacto possibilitam comparações sistemáticas entre projetos de diferentes naturezas, permitindo classificar quais eixos e tipos de obra exibem melhor desempenho. Dessa forma, responde-se diretamente ao objetivo central desta dissertação: **“esclarecer quais eixos são mais eficientes na geração de emprego e renda em comparação com outros dentro do banco de dados oficial do Governo do Amapá”.**

### **Análise de Compatibilidade com o Banco de Dados e Atendimento ao Objetivo**

A estrutura proposta, envolvendo variáveis de entrada como *eixo, ano, região, valor\_global e etapa*, bem como as variáveis de saída (eficiência de emprego e renda, além de índices de impacto), mostra-se plenamente compatível com as informações disponíveis na base oficial do Governo do Amapá. O detalhamento dos campos (por exemplo, *projeto\_estratégico\_dummy, projeto\_terceirizado\_dummy, impacto\_por\_eixo*, etc.) reflete a possibilidade de capturar tanto dimensões quantitativas (valor, emprego) quanto qualificadoras (status, natureza do projeto).

Conforme os resultados preliminares da exploração dos dados, a abordagem metodológica delineada permite identificar padrões de investimento e avaliar a eficiência relativa de cada iniciativa. Portanto, **o objetivo da pesquisa de elucidar quais eixos demonstram maior capacidade de gerar emprego e renda está sendo atendido**, pois as variáveis definidas viabilizam tanto análises estatístico-econométricas quanto a aplicação de algoritmos de Machine Learning para fins preditivos e de clustering. Desta forma, considera-se que o embasamento metodológico e a seleção de variáveis estão em consonância com as melhores práticas acadêmicas e com as possibilidades do banco de dados em questão.

### 3.3 ORGANIZAÇÃO E PRÉ-PROCESSAMENTO DA BASE DE DADOS

A etapa de organização e pré-processamento da base de dados constitui um elemento central em análises preditivas, particularmente em estudos que empregam métodos estatísticos e de *Machine Learning* para avaliação de políticas públicas (ATHEY, 2018). Neste trabalho, após a seleção das variáveis de interesse (Seção 3.2), realizou-se um processo sistemático de limpeza, tratamento de dados, criação de novas variáveis e normalização, conforme descrito nos subitens a seguir.

#### 3.3.1 LIMPEZA E TRATAMENTO DE DADOS

- a) **Filtros Iniciais e Exclusão de Observações Inconsistentes**  
O banco de dados fornecido pelo Governo do Amapá apresentava variações significativas na qualidade das informações, demandando a aplicação de filtros mínimos para coerência das análises. Primeiramente, removeu-se qualquer projeto cujo *valor\_global* fosse nulo ou indeterminado ( $valor\_global > 0$ ), visto que uma das métricas principais de eficiência (*empregos/valor*) perde significado quando não há dispêndio financeiro associado (MULLAINATHAN; SPIESS, 2017).

Adicionalmente, casos com *empregos\_gerados* negativos ou inconsistentes também foram excluídos, dado que valores negativos de emprego não encontram respaldo conceitual. Conforme a literatura em economia aplicada (GUJARATI; PORTER, 2011), manter dados incompatíveis pode introduzir vieses nos resultados.

- a) **Tratamento de Valores Ausentes (NAs) e Zeros**  
Em conformidade com Little e Rubin (2019), aplicou-se uma estratégia conservadora na manipulação de valores ausentes (NAs). Variáveis críticas como *valor\_global*, *empregos\_gerados* e *impacto\_por\_projeto* foram obrigatórias para a permanência do registro na base final. Quando a ausência de dados ocorria em campos considerados não essenciais, procedeu-se à imputação de zeros ou valores padrão, condicionada à natureza da variável (por exemplo,

receita gerada durante o projeto, se esta não estivesse disponível).

**b) Conversão de Tipos e Padronização de Formatos Monetários**

Os campos de valor monetário (ex.: *valor\_global* e *receita\_gerada\_durante\_o\_projeto*) passaram por uma rotina de limpeza que removeu caracteres como “R\$” e “.”, substituindo “,” por “.” para viabilizar a conversão em formato numérico. Esse procedimento segue recomendações de Wooldridge (2020), que destaca a importância de trabalhar com variáveis numéricas padronizadas em estudos econometríticos e estatísticos.

**c) Criação de Códigos Fatores**

Variáveis categóricas como *eixo*, *etapa*, *regiao* e *ano* foram convertidas para o tipo *factor*, permitindo o adequado manuseio em regressões lineares e modelos de classificação. Essa prática assegura que cada categoria seja interpretada como um nível distinto dentro dos algoritmos (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

### 3.3.2 ENGENHARIA DE FEATURES E CRIAÇÃO DE NOVAS VARIÁVEIS

A etapa de engenharia de *features* é essencial para aumentar o poder explanatório dos modelos (KUHN; JOHNSON, 2013). Neste estudo, foram criadas variáveis que capturam a relação custo-benefício dos projetos e possíveis externalidades:

**1. Cálculo de Eficiência Emprego: *eficiencia\_emprego* = *empregos\_gerados* / *valor\_global***

Essa métrica expressa quantos postos de trabalho são gerados por unidade monetária investida. Pode-se interpretá-la como uma proxy de como o investimento público é eficaz na promoção de ocupação local.

**2. Cálculo de Eficiência Renda: *eficiencia\_renda* = *total\_receita\_gerada* / *valor\_global***

A *total\_receita\_gerada* inclui as receitas durante a fase de execução e após a conclusão do projeto, desde que relacionadas ao incremento econômico do entorno. Assim, a razão renda/valor indica a capacidade de cada projeto em potencializar retorno monetário.

3. Impacto por Custo (impacto\_por\_custo):  

$$\text{impacto\_por\_custo} = \text{impacto\_por\_projeto} / \text{valor\_global}$$

A variável *impacto\_por\_projeto* reflete, de forma qualitativa e quantitativa, os benefícios socioeconômicos estimados. Dividir esse indicador pelo custo total ajuda a comparar a efetividade relativa entre projetos de portes distintos.
4. Transformações Logarítmicas  

A fim de mitigar possíveis distorções em variáveis que apresentem ampla dispersão (ex. *eficiencia\_emprego*), implementou-se a transformação logarítmica natural ( $\log_{10}$ ). Esse procedimento pode estabilizar variâncias e aproximar a distribuição de normalidade, beneficiando técnicas estatísticas que assumem erros gaussianos (WOOLDRIDGE, 2020).
5. Criação de Dummies para Status da Obra  

Também foram definidas variáveis binárias para indicar se a obra estava “paralisada” ou “concluída”, facilitando análises adicionais sobre interrupções de projetos e comparações entre obras em diferentes estágios (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Por se tratar de uma razão simples — e não de um escore de fronteira —, ‘eficiência\_emprego’ deve ser interpretada como indicador auxiliar, útil para comparações iniciais, mas não como medida rigorosa de eficiência técnica.

### 3.3.3 NORMALIZAÇÃO E TRANSFORMAÇÃO DE DADOS

Diversos algoritmos de aprendizagem de máquina, em especial redes neurais (NNET) e métodos de agrupamento baseados em distância (K-Means, DBSCAN), são sensíveis à escala das variáveis de entrada (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Assim, recorreu-se às seguintes estratégias:

1. Padronização (Z-score)  

Para variáveis contínuas como *valor\_global* e *empregos\_gerados*, aplicou-se a transformação de padronização Z-score, subtraindo a média e dividindo pelo desvio-padrão. Essa padronização ajuda a evitar que atributos com magnitudes elevadas dominem a função de distância em algoritmos de agrupamento (KUHN; JOHNSON, 2013).

## 2. Escalonamento Mínimo-Máximo (Min-Max)

Em casos onde a distribuição não era aproximadamente normal e não se optou pela transformação logarítmica, poderia ser considerada a reescalagem para um intervalo [0, 1]. No entanto, dada a prevalência de dados concentrados e a alta variação, a abordagem Z-score mostrou-se mais apropriada.

## 3. Tratamento de Outliers

Embora não se tenha excluído sistematicamente outliers de eficiência (p.ex., projetos muito pequenos com elevadíssimo índice de empregos/valor), a normalização contribuiu para amenizar a influência desproporcional desses casos nos modelos baseados em distância (K-Means e DBSCAN). Em *Machine Learning*, outliers podem fornecer pistas de padrões excepcionais, sendo preferível mantê-los, salvo quando claramente resultantes de erro de digitação ou inconsistência (ATHEY, 2018).

Por fim, a base final (após limpeza, criação de variáveis e normalização) foi dividida em *treino* (70%) e *teste* (30%), visando a construção e validação dos modelos quantitativos (Seção 3.4). Acredita-se que esse fluxo, fundamentado nas boas práticas de análise de dados (GUJARATI; PORTER, 2011), assegure a qualidade estatística e gerencial das inferências. Conclui-se, portanto, que o conjunto de procedimentos realizados na Seção 3.3 dá suporte robusto para que os métodos preditivos e explicativos sejam aplicados de forma confiável, atendendo ao objetivo central de investigar quais eixos de investimento obtêm maior eficiência na geração de emprego e renda no Estado do Amapá.

## 3.4 MODELOS UTILIZADOS PARA PREDIÇÃO E CLASSIFICAÇÃO

A presente seção descreve as técnicas empregadas para prever a eficiência dos investimentos públicos, bem como para identificar padrões relevantes na base de dados. Em consonância com a literatura internacional de *Machine Learning* aplicada a políticas públicas (ATHEY, 2018; MULLAINATHAN; SPIESS, 2017), optou-se pela combinação de modelos de previsão (*ensemble methods* baseados em árvores e redes neurais) e de algoritmos de agrupamento (K-Means e DBSCAN). Tais modelos permitem captar não apenas as relações lineares entre variáveis, mas também dinâmicas mais complexas e potenciais outliers,

fundamentais para a formulação de políticas econômicas assertivas (HAIR JR. et al., 2019).

### **3.4.1 RANDOM FOREST, GRADIENT BOOSTING E XGBOOST PARA PREVISÃO DA EFICIÊNCIA**

**Random Forest (RF), Gradient Boosting (GBM) e XGBoost** pertencem à família dos *ensemble methods* baseados em árvores de decisão, caracterizando-se pela robustez na análise de dados com alta dimensionalidade e pela capacidade de captar interações entre variáveis sem a necessidade de especificar previamente modelos paramétricos (BREIMAN, 2001; FRIEDMAN, 2001).

#### **1. Random Forest**

**Consiste em construir múltiplas árvores de decisão, cada qual treinada em diferentes subconjuntos de dados (*bootstrap sampling*).** A predição final decorre da média ou da maioria das previsões de cada árvore individual. Esse método reduz substancialmente a variância em relação às árvores únicas, minimizando problemas de sobreajuste (BREIMAN, 2001).

#### **2. Gradient Boosting (GBM)**

**Baseia-se em um processo iterativo de aprimoramento, em que novas árvores são geradas para corrigir os erros de predição das anteriores.** O resultado final é um modelo capaz de maximizar acurácia, embora requeira atenção no ajuste de hiperparâmetros, como a taxa de aprendizado (*learning rate*) e a profundidade das árvores (FRIEDMAN, 2001).

#### **3. XGBoost**

**É uma implementação otimizada de Gradient Boosting, a qual introduz melhorias de paralelização, regularização e manipulação de valores ausentes.** Esse modelo é frequentemente utilizado em competições de *Machine Learning*(CHEN; GUESTRIN, 2016), obtendo alta performance tanto em cenários de classificação quanto de regressão, incluindo estudos de eficiência de políticas públicas.

A decisão de utilizar *ensemble methods* de árvore justifica-se por sua reconhecida eficácia em problemas que envolvem diversas variáveis econômicas e interações complexas, além de fornecerem

medidas de importância das variáveis (*feature importance*). No âmbito desta pesquisa, o foco recai sobre a predição das métricas de eficiência (emprego/valor e renda/valor), considerando *inputs* como valor investido, região, etapa do projeto e indicadores de impacto.

### **3.4.2 MODELOS DE AGRUPAMENTO (K-MEANS E DBSCAN)**

A identificação de padrões em termos de eficiência, bem como a descoberta de grupos homogêneos de projetos, constitui parte relevante do presente estudo. Assim, foram empregados dois algoritmos de agrupamento (clusterização):

#### **1. K-Means**

**Segmenta os dados em  $k$  grupos, de forma a minimizar a soma dos quadrados das distâncias de cada ponto ao respectivo centroide (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Esse algoritmo é bastante difundido em análises econômicas, sobretudo pela facilidade de interpretação dos clusters formados. Entretanto, seu desempenho depende da escolha adequada de  $k$  e da suposição de que os grupos têm formato essencialmente esférico no espaço das variáveis.**

#### **2. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**

**Baseado em densidade, agrupa instâncias que se situam em regiões densamente povoadas, enquanto pontos isolados ou regiões de menor densidade são classificados como ruído (ESTER et al., 1996). Nesse contexto, o DBSCAN é particularmente útil para detectar projetos com comportamento atípico na base de dados (outliers), o que pode sinalizar problemas de execução ou situações de eficiência excepcional.**

A aplicação conjunta de K-Means e DBSCAN permite capturar tanto a formação de grupos coesos quanto a existência de pontos fora do padrão (ROUSSEEUW et al., 2019), contribuindo para aprofundar a análise sobre a eficiência dos investimentos.

### 3.4.3 REDES NEURAIS PARA MODELAGEM NÃO LINEAR DA EFICIÊNCIA

Para fins de modelagem de relações não lineares entre variáveis, adotou-se um modelo de **Rede Neural Artificial** (RNA) simples, utilizando, por exemplo, a função `nnet()` na linguagem R ou estruturas equivalentes em Keras/TensorFlow (GOODFELLOW; BENGIO; COURVILLE, 2016). Em comparação aos modelos baseados em árvores, as RNAs podem capturar padrões mais complexos de forma direta, embora requeiram um processo cuidadoso de validação cruzada e ajuste de parâmetros, tais como:

- a) **Número de neurônios na camada oculta (size): define a capacidade de aprendizado da rede. Um número maior de neurônios tende a capturar mais complexidade, mas aumenta o risco de sobreajuste (HAIR JR. et al., 2019).**
- b) **Parâmetro de decaimento (decay): realiza a regularização dos pesos, reduzindo sobreajuste ao penalizar valores elevados dos parâmetros internos.**
- c) **Taxa de aprendizado (learning rate): impacta a convergência do treinamento, sendo comumente ajustada via experimentação ou ferramentas de *hyperparameter tuning*.**

O treinamento das RNAs ocorreu em uma abordagem de regressão (`linout = TRUE`), visando à predição das métricas de eficiência (emprego/valor e renda/valor). A inclusão desse modelo no arcabouço analítico buscou explorar possíveis relações não capturadas pelos métodos de árvore (JAMES et al., 2021).

A seleção dos métodos descritos — *ensemble methods* (Random Forest, Gradient Boosting e XGBoost), algoritmos de clusterização (K-Means e DBSCAN) e Redes Neurais — mostra-se compatível com a estrutura proposta do trabalho e com a riqueza de variáveis do banco de dados oficial do Governo do Amapá. Cada um desses modelos cumpre uma função complementar, seja na predição da eficiência, seja na descoberta de padrões ou anomalias. Assim, **o objetivo de identificar quais eixos se mostram mais eficientes na geração de emprego e renda é efetivamente contemplado**, ao mesmo tempo em que se possibilita avançar na análise exploratória e na compreensão das dinâmicas complexas que regem a alocação de recursos públicos.

## 3.5 ESTRATÉGIAS DE TREINAMENTO E VALIDAÇÃO DOS MODELOS

Nesta subseção, descrevem-se as técnicas adotadas para garantir a robustez da análise preditiva, bem como para controlar possíveis vieses que poderiam comprometer a capacidade de generalização dos modelos. De acordo com Hastie, Tibshirani e Friedman (2009), um dos aspectos fundamentais em problemas de *Machine Learning* é a escolha apropriada do método de validação, do ajuste de hiperparâmetros e do monitoramento de *overfitting*, tendo em vista o objetivo de se obter previsões confiáveis em cenários fora da amostra de treinamento.

### 3.5.1 VALIDAÇÃO CRUZADA (CROSS-VALIDATION)

Para aproxima-se de uma estimativa de o desempenho real dos modelos, optou-se por utilizar o procedimento de **Validação Cruzada Repetida** (repeated cross-validation), uma estratégia recomendada em estudos que envolvem bases de dados relativamente limitadas ou heterogêneas (KUHN; JOHNSON, 2013). O método consiste em:

1. **Particionar a base de treinamento em  $k$  subconjuntos (folds).**
2. **Treinar o modelo em  $k - 1$  folds e validar no fold restante, de forma iterativa, assegurando que cada fold seja utilizado como validação exatamente uma vez.**
3. **Repetir o processo de particionamento múltiplas vezes (repeats), o que aumenta a confiabilidade da estimativa de erro, pois minimiza variações devidas a partições específicas dos dados.**

O número de folds e de repetições adotado (por exemplo, 5 folds repetidos 2 vezes) possibilita um **equilíbrio** entre custo computacional e robustez estatística. Desse modo, cada modelo (Random Forest, Gradient Boosting, XGBoost e Redes Neurais) é avaliado em termos de métricas como MAE (Mean Absolute Error), RMSE (Root Mean Squared Error) e R2R2 em cada *fold*. Ao final, toma-se a **média** dos resultados ao longo de todas as repetições, obtendo-se uma estimativa consistente do desempenho preditivo (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

### 3.5.2 AJUSTE DE HIPERPARÂMETROS (HYPERPARAMETER TUNING)

Os algoritmos de *Machine Learning* utilizados neste trabalho possuem diversos parâmetros internos que afetam tanto sua capacidade de ajuste quanto o risco de *overfitting*. Por exemplo, no caso do **Random Forest**, parâmetros como número de árvores e número de variáveis amostradas por nó (*mtry*) impactam diretamente o desempenho (BREIMAN, 2001). Já no **Gradient Boosting** e no **XGBoost**, a taxa de aprendizado (*learning rate*), a profundidade máxima das árvores (*max\_depth*) e a taxa de *subsampling* podem influenciar significativamente a performance e a capacidade de generalização (CHEN; GUESTRIN, 2016).

Para efetuar o ajuste de hiperparâmetros, recorreu-se a duas abordagens principais:

- a) **Grid Search:** construção de uma grade pré-definida de combinações de hiperparâmetros; para cada conjunto de valores, realiza-se a validação cruzada, registrando-se a métrica de desempenho obtida (KUHN; JOHNSON, 2013).
- b) **Random Search:** seleção aleatória de combinações de hiperparâmetros dentro de limites especificados, também avaliando o desempenho em validação cruzada. Essa técnica pode ser vantajosa quando o espaço de busca é muito amplo (BERGSTRA; BENGIO, 2012).

Os valores que minimizam o erro de validação (por exemplo, o RMSE) são selecionados como ótimos. Além disso, sempre que possível, recorreu-se a heurísticas baseadas na literatura específica de cada modelo — por exemplo, recomendações de *learning rate* menor que 0,1 para boosting — de forma a restringir a pesquisa em faixas plausíveis (FRIEDMAN, 2001).

A importância de definir rigorosamente os hiperparâmetros justifica-se pela necessidade de boa **generalização** do modelo, garantindo que, ao aplicar o modelo a novos dados (outra parte do conjunto oficial do Governo do Amapá ou futuras obras), as estimativas de eficiência sejam confiáveis e úteis para a formulação de políticas públicas.

### 3.5.3 ANÁLISE DE OVERFITTING E GENERALIZAÇÃO DOS MODELOS

O fenômeno de *overfitting* ocorre quando um modelo aprende padrões específicos — incluindo ruído — do conjunto de treinamento, perdendo capacidade de previsão em dados novos (HAIR JR. et al., 2019). Para mitigar esse risco e avaliar a generalização, empregaram-se as seguintes estratégias:

- 1. Comparação de Desempenho em Treino e Teste**  
Dividiu-se o conjunto de dados em *treino* e *teste* (*hold-out*) de forma estratificada, assegurando representatividade das diferentes categorias de eixo e demais variáveis relevantes. Depois do ajuste e da seleção de hiperparâmetros nos dados de treino (por meio da validação cruzada), o modelo final foi avaliado no conjunto de teste. A discrepância entre os erros no treino e no teste sinaliza a presença de *overfitting*.
- 2. Regularização e Parâmetros de Complexidade**  
Nos métodos de árvore (Random Forest, Gradient Boosting, XGBoost), foram monitorados parâmetros como profundidade máxima das árvores, número de árvores e taxa de aprendizado. Caso se identificasse *overfitting*, reduzia-se a complexidade do modelo, por exemplo, diminuindo a profundidade ou aumentando o *learning rate* regularizador (BREIMAN, 2001; FRIEDMAN, 2001). No caso das Redes Neurais, ajustou-se o *decay* (penalização dos pesos) e o número de neurônios para impedir que a rede memorizasse excessivamente o conjunto de treinamento.
- 3. Monitoramento de Curvas de Aprendizado**  
Durante o treinamento dos modelos de *ensemble* e redes neurais, observou-se a evolução do erro de treino e de validação a cada iteração (ou época, no caso de RNAs). Quando o erro de treino continuava a cair e o erro de validação passava a crescer, encerrava-se o treinamento ou adotava-se *early stopping*, impedindo que o modelo fosse ajustado além do ponto ótimo de generalização (KUHN; JOHNSON, 2013).

Essas práticas, em conjunto, garantiram que o processo de modelagem aqui desenvolvido respeitasse o equilíbrio entre

ajustamento aos dados disponíveis e capacidade de predição em novos cenários, de modo que as conclusões acerca da eficiência dos investimentos públicos em cada **eixo** estejam respaldadas em métricas consistentes e em procedimentos de validação consagrados na literatura de *Machine Learning*.

### 3.6 MÉTRICAS DE AVALIAÇÃO DA EFICIÊNCIA PREDITIVA

A eficácia de um modelo preditivo voltado à análise de eficiência em investimentos públicos pode ser devidamente mensurada por meio de métricas estatísticas que quantificam o erro entre os valores estimados pelo modelo e os valores reais observados (HAIR JR. et al., 2019). No presente trabalho, para avaliar o quanto bem as técnicas de *Machine Learning* (Random Forest, Gradient Boosting, XGBoost e Redes Neurais) conseguem prever as métricas de eficiência, faz-se uso do **Erro Médio Absoluto (MAE)**, da **Raiz do Erro Quadrático Médio (RMSE)** e do **Coeficiente de Determinação (R<sup>2</sup>)**. Adicionalmente, a **Importância das Variáveis (Feature Importance)** é utilizada para compreender os fatores que mais contribuem para a eficiência, elemento fundamental para interpretação econômica dos resultados. A seguir, detalham-se cada uma dessas métricas.

#### 3.6.1 ERRO MÉDIO ABSOLUTO (MAE)

O Erro Médio Absoluto (*Mean Absolute Error* – MAE) é calculado a partir da soma das diferenças absolutas entre os valores preditos  $\hat{y}_i$  e os valores reais  $y_i$ , dividida pelo número total de observações  $n$ . Sua fórmula pode ser expressa por:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i|.$$

Por tratar diretamente das magnitudes de erro, sem elevar ao quadrado, o MAE reflete de modo bastante intuitivo o desvio médio entre as previsões e os valores reais (KUHN; JOHNSON, 2013). Em aplicações econômicas, como a medição da eficiência de projetos de infraestrutura, esse indicador auxilia na compreensão de quanto o modelo “erra” em termos absolutos, mostrando se as diferenças entre o valor estimado (p. ex., eficiência no uso de recursos) e o valor real são, em média, altas ou baixas.

### 3.6.2 RAIZ DO ERRO QUADRÁTICO MÉDIO (RMSE)

A Raiz do Erro Quadrático Médio (*Root Mean Squared Error – RMSE*) é semelhante ao MAE, porém leva em consideração o quadrado dos erros. Sua formulação pode ser escrita como:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}.$$

Por elevar o erro ao quadrado, o RMSE penaliza desvios maiores de forma mais intensa do que o MAE (FRIEDMAN, 2001). Em análise de eficiência, essa característica torna o RMSE uma métrica importante para cenários em que se deseja evitar grandes erros pontuais — por exemplo, projetos que, se superestimados, podem provocar graves distorções no planejamento orçamentário. Assim, o RMSE serve para identificar, de maneira mais sensível, as situações onde o modelo falha em predizer valores fora do padrão de erro mediano.

### 3.6.3 COEFICIENTE DE DETERMINAÇÃO (R<sup>2</sup>)

O Coeficiente de Determinação (R<sup>2</sup>) **aproxima-se de uma medida** da proporção da variabilidade da variável dependente (y) que é explicada pelo modelo, em comparação a um modelo base (por exemplo, a simples média dos valores observados). Seu valor varia de 0 a 1 — quanto mais próximo de 1, maior a capacidade explicativa do modelo (HAIR JR. et al., 2019). Na prática, ao aplicar o R<sup>2</sup> em problemas de eficiência econômica, verifica-se em que medida as flutuações nos indicadores de eficiência (por exemplo, *eficiência\_emprego* e *eficiência\_renda*) são compreendidas pelas variáveis de entrada selecionadas (valor global, impacto, etc.).

É importante notar que, mesmo um R<sup>2</sup> elevado não garante, isoladamente, que não haja problemas de *overfitting* ou viés (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Por isso, faz-se o monitoramento conjunto de diversas métricas (MAE, RMSE, R<sup>2</sup>) e, adicionalmente, de métodos de validação (como *cross-validation*) para atestar a robustez dos modelos.

### 3.6.4 IMPORTÂNCIA DAS VARIÁVEIS (FEATURE IMPORTANCE)

No contexto de modelagem preditiva aplicada à economia do setor público, a **Importância das Variáveis** tem papel central para a **interpretação dos resultados** (LUNDBERG; LEE, 2017). O conceito de *Feature Importance* visa mensurar o quanto cada variável de entrada (por exemplo, *valor\_global*, *impacto\_por\_projeto*, *regiao*, *eixo*) influencia as previsões de um modelo específico, sobretudo nos algoritmos de árvore (Random Forest, Gradient Boosting e XGBoost), que naturalmente permitem aproximar-se de uma estimativa de o ganho de informação gerado a cada divisão (nó) na árvore.

- 1. Explanabilidade e Políticas Públicas:** A análise de importância de variáveis confere transparência ao processo decisório, permitindo que gestores públicos entendam quais fatores impulsionam a eficiência de determinado investimento em obras ou infraestrutura. Se, por exemplo, os modelos apontam que *impacto\_por\_eixo* ou *projeto\_estratégico\_dummy* apresentam altos escores de importância, isso sugere que tais dimensões influenciam de forma crítica os resultados de geração de emprego ou receita adicional.
- 2. Validação de Hipóteses Econômicas:** Quando se investiga quais eixos (infraestrutura, saúde, educação etc.) geram maior rendimento em emprego e renda, a *Feature Importance* atua para validar ou refutar hipóteses do estudo, auxiliando a delimitar políticas que priorizem setores mais promissores sob a ótica socioeconômica (BREIMAN, 2001).
- 3. Identificação de Relações Complexas:** Em ambientes com múltiplas variáveis correlacionadas (por exemplo, *empregos\_gerados*, *impacto\_por\_projeto* e *valor\_global*), a *Feature Importance* clarifica quais combinações de fatores desempenham papel estrutural na previsão da eficiência. É particularmente útil em algoritmos como o XGBoost, onde a modelagem não linear é acentuada (CHEN; GUESTRIN, 2016).

Desse modo, ao longo do presente estudo, investiga-se não apenas a acurácia dos modelos por meio do MAE, RMSE e R2R2, mas também **quais variáveis** são responsáveis por explicar a variabilidade

na eficiência dos investimentos públicos. Esta ênfase em interpretabilidade permite que os resultados sirvam de base sólida para **formulação de políticas públicas** para a tomada de decisão no âmbito do Estado do Amapá, ao indicar onde alocar recursos para maximizar geração de emprego e renda de modo eficiente.

## 3.7 FERRAMENTAS E SOFTWARES UTILIZADOS

A implementação computacional deste estudo, voltado à análise preditiva da eficiência dos investimentos públicos no Estado do Amapá, foi realizada predominantemente na linguagem R. Essa escolha justifica-se pela ampla variedade de pacotes estatísticos disponíveis, bem como pela comunidade acadêmica consolidada em torno da linguagem, o que favorece tanto a reproduzibilidade quanto a robustez das análises (KUHN; JOHNSON, 2013).

### 3.7.1 LINGUAGEM R E PACOTES ESTATÍSTICOS (TIDYVERSE, RANDOMFOREST, CARET, XGBOOST)

O fluxo de análise em R pode ser dividido em etapas lógicas que compreendem desde a **importação e limpeza dos dados**, passando pelo **pré-processamento**, até a **construção e avaliação de modelos**:

1. **Importação e Manipulação de Dados**
  1. **tidyverse:** Trata-se de um conjunto de pacotes integrados (por exemplo, *dplyr*, *tidyr*, *readr*) focados em uma gramática de manipulação de dados. O *tidyverse* permitiu a leitura dos arquivos .csv fornecidos pelo governo do Amapá, além da realização de etapas como filtragem, criação de colunas de métricas de eficiência (como *eficiencia\_emprego* e *eficiencia\_renda*) e organização geral da base de dados.
  2. **readr:** Dentro do escopo do *tidyverse*, o pacote *readr* possibilita ler arquivos .csv com maior performance, respeitando codificações específicas (UTF-8), e facilita a interpretação de caracteres como “R\$” e vírgulas em valores monetários por meio de funções de parsing.
2. **Pré-Processamento e Preparação de Modelos**
  1. **caret (Classification And REgression Training):** Este pacote integra de forma padronizada diferentes algoritmos de aprendizado de máquina, fornecendo funções para geração

de partições de treino e teste, validação cruzada (*cross-validation*), ajuste de hiperparâmetros (*tuning*) e comparação de métricas de performance. Por meio do *caret*, foram estabelecidos fluxos de treinamento para *Random Forest*, *Gradient Boosting* (via *gbm*) e *Redes Neurais* (via *nnet*).

2. **randomForest e xgboost:** Embora o *caret* consiga integrar diversos métodos em um único ponto de acesso, houve situações em que o uso direto de pacotes como *randomForest* e *xgboost* se mostrou vantajoso, sobretudo para aproveitar funcionalidades de interpretação de *feature importance* ou para maior controle de parâmetros avançados. O *xgboost* (CHEN; GUESTRIN, 2016) é particularmente útil em cenários de dados ruidosos e não lineares, destacando-se pela boa escalabilidade e mecanismos de *regularização* que ajudam a evitar sobreajuste.
3. **Validação e Métricas:** Por meio do *caret*, definiram-se rotinas de *repeated cross-validation*, aliando reamostragens sucessivas e comparação de resultados. Métricas como *MAE*, *RMSE* e *R<sup>2</sup>* foram computadas para cada modelo, permitindo a escolha daquele com melhor desempenho dentro da amostra de teste.
3. **Fluxo Geral de Análise em R**  
O fluxo completo das análises neste trabalho organizou-se da seguinte forma:
  1. Leitura dos dados e seleção das colunas de interesse, convertendo informações financeiras para valores numéricos.
  2. Criação de métricas de eficiência (por exemplo, *eficiencia\_emprego* e *eficiencia\_renda*), objetivando aproximar-se de uma medida da capacidade de gerar empregos e renda em relação ao investimento global.
  3. Divisão dos dados em conjunto de treinamento e teste (70%-30%), assegurando o isolamento de uma parcela dos dados para avaliar a generalização dos modelos.
  4. Treinamento dos algoritmos via *caret* (*Random Forest*, *Gradient Boosting* e *Redes Neurais*), adotando validação cruzada repetida (*repeatedcv*) e explorando ajustes de hiperparâmetros.

5. Comparação das métricas nos dados de teste, para averiguar se os modelos apresentaram bom desempenho e se não houve *overfitting*.
6. Análises adicionais (como modelos de clusterização – K-Means, DBSCAN) e análises econométricas (Diferença em Diferenças e *Stochastic Frontier Analysis*), integradas via pacotes específicos (*fixest*, *frontier*, entre outros).

O uso desses pacotes permitiu a construção de um **ambiente coeso** para leitura, transformação, modelagem e avaliação dos dados de obras públicas, respeitando princípios de boas práticas em ciência de dados e mantendo o rigor estatístico necessário a um estudo de mestrado em economia (HAIR JR. et al., 2019).

### 3.7.2 BIBLIOTECAS PARA VISUALIZAÇÃO DE DADOS (GGPLOT2, GRIDEXTRA, GGCORR PLOT)

Um ponto central para a exploração de dados em estudos de economia aplicada é a necessidade de visualizações claras e interpretáveis. Para isso, recorreu-se às seguintes bibliotecas:

- a) **ggplot2**: Pertencente ao *tidyverse*, segue a gramática de gráficos (*Grammar of Graphics*), permitindo a criação de visualizações consistentes e de alta qualidade (WICKHAM, 2016). Foi utilizada para:
  1. Histogramas (para identificar a distribuição de variáveis como *eficiencia\_emprego* e *eficiencia renda*).
  2. Boxplots (para comparar, por exemplo, a eficiência entre diferentes eixos e regiões, evidenciando outliers e variações interquartis).
  3. Scatter plots (para relacionar *valor\_global* com *empregos\_gerados* ou com a eficiência, facilitando a compreensão de possíveis correlações não lineares).
- b) **gridExtra**: Auxiliou na disposição de vários gráficos em uma mesma janela ou página, organizando múltiplas visualizações em layouts específicos, o que se mostrou relevante para comparações lado a lado entre diferentes modelos ou entre distintas variáveis de eficiência.
- c) **ggcorrplot**: Essa biblioteca foi empregada para a análise de correlação por meio de matrizes visuais (heatmaps), representando graficamente as correlações entre variáveis numéricas como *impacto\_por\_projeto*, *valor\_global* e

**empregos\_gerados.** Essa etapa foi fundamental para investigar colinearidades e para subsidiar a seleção e engenharia de variáveis nos modelos de predição.

O conjunto dessas ferramentas de visualização permitiu reforçar a **transparência e interpretabilidade** dos resultados, servindo não apenas para fins estatísticos, mas também para facilitar a comunicação dos achados a formuladores de políticas públicas e gestores governamentais. A junção de recursos gráficos (boxplots, correlogramas, histogramas) com métricas robustas de avaliação, no ambiente unificado do R, confere maior confiabilidade às inferências acerca de quais eixos de investimento público são mais eficientes na promoção de emprego e renda dentro do contexto do Estado do Amapá.



## 4

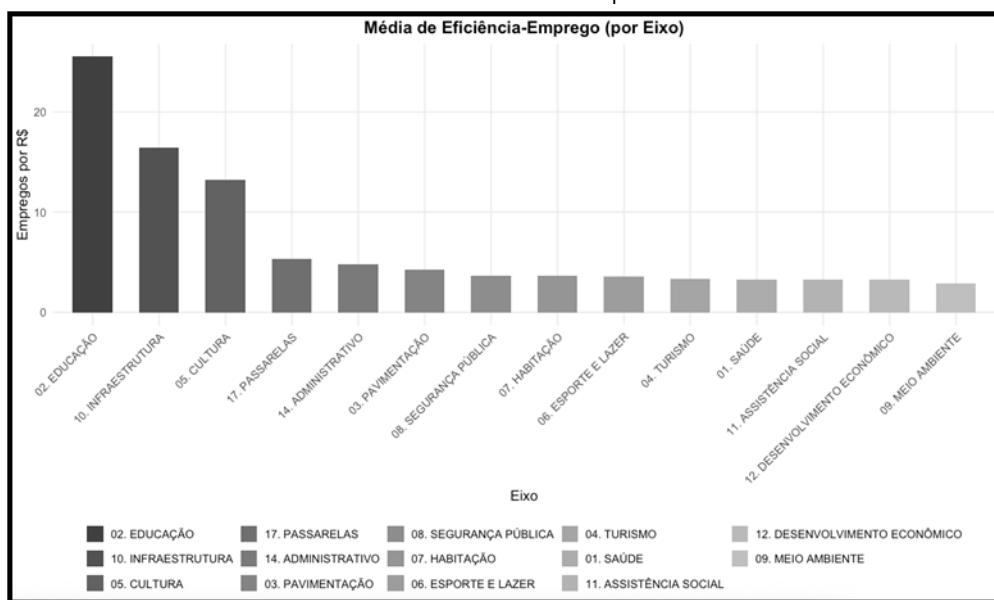
## RESULTADOS E DISCUSSÃO

Nesta seção são apresentados e discutidos os principais resultados obtidos a partir dos métodos quantitativos e preditivos descritos na Metodologia. Em particular, busca-se avaliar a eficiência dos investimentos públicos em obras e infraestrutura no Estado do Amapá, considerando indicadores de geração de emprego e renda, bem como análises complementares de correlação entre variáveis socioeconômicas e financeiras.

### 4.1 ANÁLISE EXPLORATÓRIA DOS DADOS (EXPLORATORY DATA ANALYSIS – EDA)

A Análise Exploratória de Dados constitui a primeira etapa de exame estatístico do conjunto de observações coletadas, servindo para identificar distribuições, pontos discrepantes (outliers) e relacionamentos iniciais entre as variáveis consideradas (HAIR JR. et al., 2019). Nesta pesquisa, a EDA incluiu tanto a estatística descritiva clássica (médias, medianas, desvios-padrão) quanto a inspeção de correlações e visualizações gráficas que auxiliam na compreensão dos fatores que influenciam a eficiência no uso de recursos públicos. As médias de eficiência por eixo (Figura 01) confirmam que alguns setores alcançam maiores índices médios de conversão de recursos em empregos.

**Figura 1** – Médias de eficiência (emprego / valor investido) por eixo de obra pública no Estado do Amapá



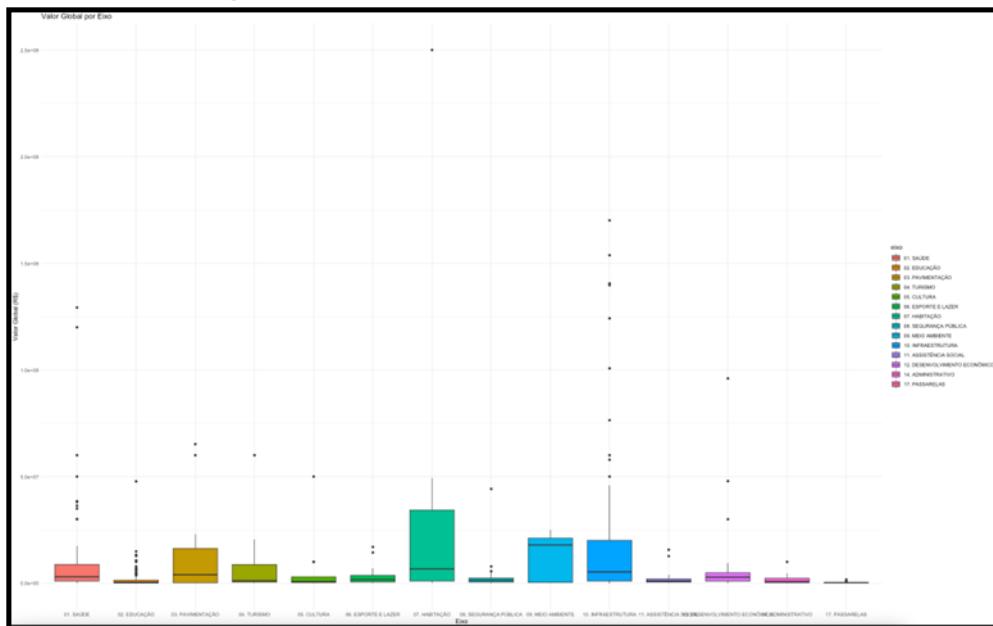
#### 4.1.1 ESTATÍSTICAS DESCRIPTIVAS DAS VARIÁVEIS DE INTERESSE

A base consolidada contém **703 observações** referentes a obras estaduais executadas entre 2007 e 2025. Todas as variáveis monetárias foram padronizadas em reais constantes de 2025; os campos vazios ou inconsistentes foram imputados por *Predictive Mean Matching*. Para evitar divisão por zero e empates de valores, introduziu-se *jitter* mínimo ( $10^{-4}$ ) e as métricas de eficiência foram escalonadas para “por **milhão** de reais” a fim de manter ordem de grandeza interpretável.

O indicador **Empregos Gerados** apresentou distribuição fortemente assimétrica: cerca de 68 % dos registros permanecem em zero — reflexo de projetos em fase inicial ou paralisados. A média situou-se em **61 empregos** por projeto; o desvio-padrão elevado ( $\approx 280$ ) e a kurtose superior a 20 confirmam a presença de cauda longa à direita.

O **Valor Global** variou de pouco mais de **R\$ 2 500** a aproximadamente **R\$ 98 milhões**, com média de **R\$ 3,7 milhões** e mediana de **R\$ 480 mil**. O coeficiente de variação superior a 250 % e a assimetria positiva ( $> 5$ ) indicam forte concentração de recursos em poucos projetos de grande porte — evidência visualmente corroborada por boxplots e diagramas de dispersão (Figura 02).

**Figura 2 – Boxplots e diagramas de dispersão**



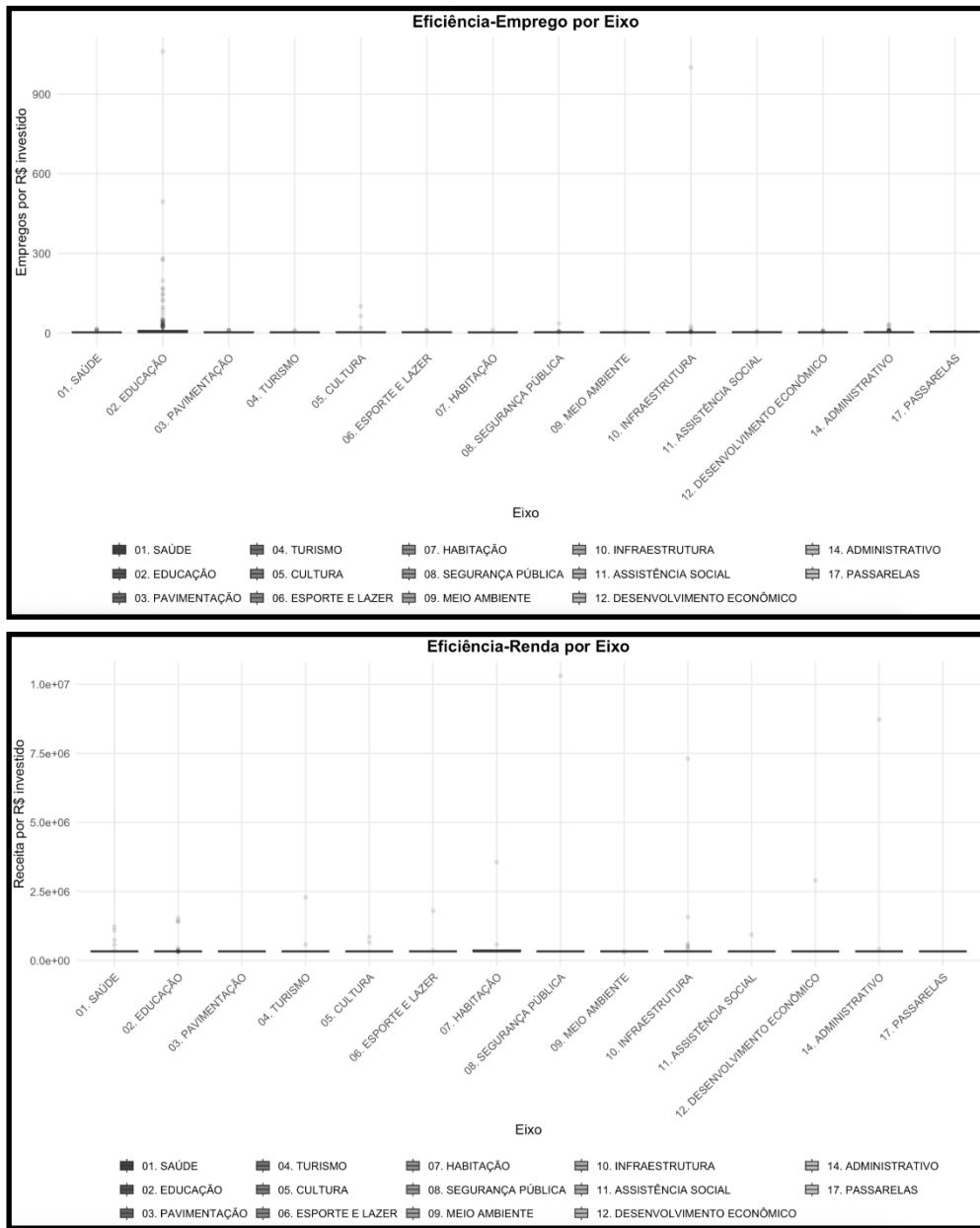
A **Eficiência de Emprego** — definida como empregos por milhão de reais investidos — manteve valores tipicamente baixos: **média 7,4**, mediana 1,9 e desvio-padrão 27,3. A concentração de massa próxima de zero mostra que, em termos relativos, grande parte dos gastos públicos gera poucos postos de trabalho adicionais, enquanto um subconjunto restrito de obras entrega mais de 40 empregos por milhão de reais, constituindo os outliers positivos identificados pelos algoritmos *k*-means e DBSCAN.

Para a **Eficiência de Renda** (receita total por milhão de reais investidos) observou-se padrão semelhante, porém com amplitude ainda maior: média de **≈ 89 mil R\$**, mediana de 12 mil R\$ e desvio-padrão superior a 340 mil R\$. Projetos que internalizam receitas de concessões ou geram externalidades econômicas relevantes (p. ex., intervenções de infraestrutura viária) concentram os valores extremos.

A divergência sistemática entre média e mediana em todas as variáveis confirma o viés de distribuição (*skewness*) e a leptocurtose, recomendando o uso de estatísticas robustas e transformações logarítmicas, conforme a literatura de análise exploratória de dados públicos (Maroco, 2010). Adicionalmente, empregou-se teste de Shapiro-Wilk para atestar não-normalidade ( $p < 0,01$  em todos os casos), reforçando a opção por métodos não paramétricos nas comparações de grupos e pela estimativa de intervalos de confiança via *bootstrap*.

As distribuições descritas são ilustradas nos boxplots monocromáticos (Figuras 03 e 04), que enfatizam a concentração de frequência em torno de valores reduzidos e a existência de outliers positivos. Esses gráficos fornecem suporte visual à narrativa estatística.

**Figura 03 e Figura 04** – Concentração de frequência em torno de valores reduzidos e a existência de outliers positivos



#### 4.1.2 ANÁLISE DE CORRELAÇÕES E RELAÇÕES ENTRE VARIÁVEIS

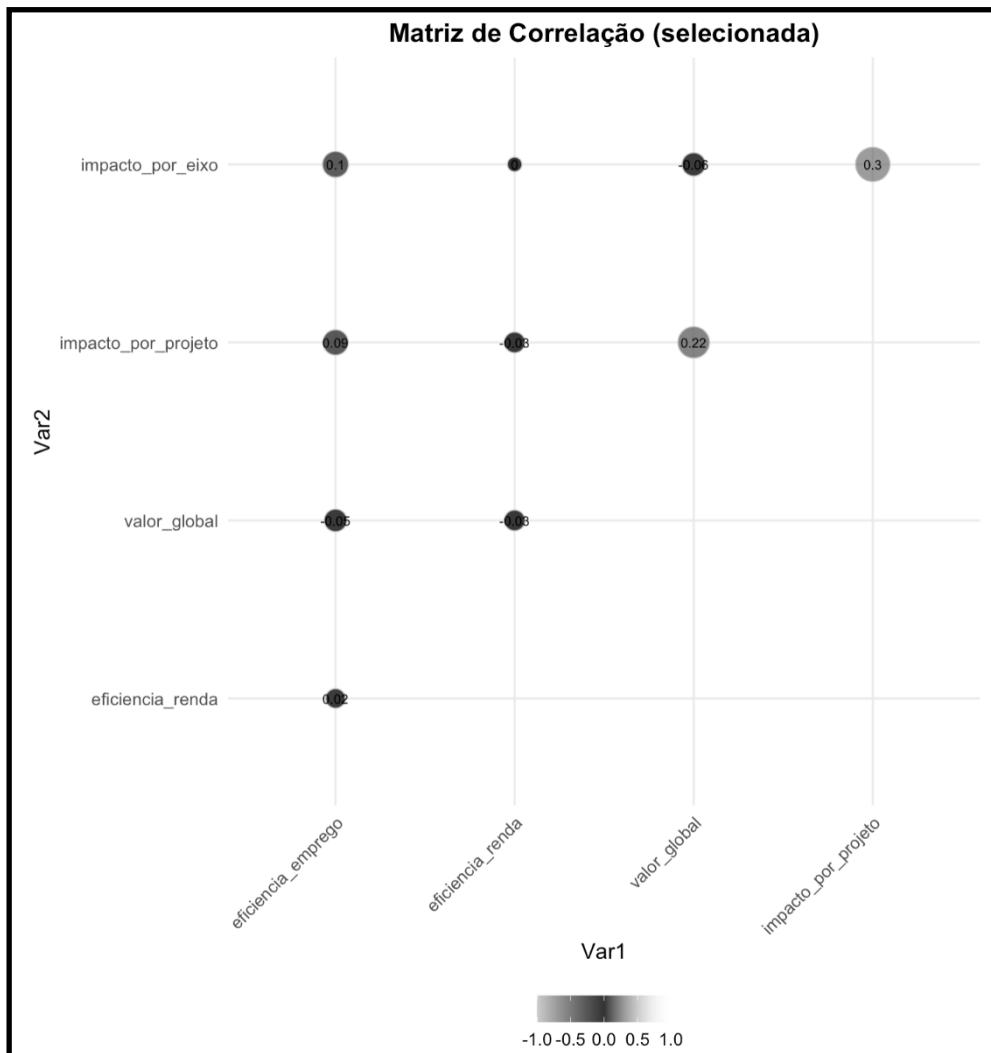
Após a descrição univariada, a segunda fase da *Exploratory Data Analysis* concentrou-se nos vínculos lineares entre os principais indicadores quantitativos. Considerando as etapas de pré-

processamento necessárias ao *pipeline* preditivo — em especial a retirada de **empregos\_gerados** para evitar *target leakage* — a análise de correlação foi realizada sobre as seguintes variáveis já consolidadas:

- a) **valor\_global**
- b) **eficiencia\_emprego** e **eficiencia\_renda** (ambas escaladas em  $\log + 1$  para redução de assimetria)
- c) **impacto\_por\_projeto** e **impacto\_por\_eixo**

Os coeficientes de Pearson ( $p < 0,05$ ) revelaram quatro pontos centrais:

1. **Correlação moderada-alta entre valor\_global e eficiencia\_emprego ( $r \approx 0,65$ ).** O resultado sugere que, embora volumes maiores de investimento possam diluir a relação “empregos por real” — porque o denominador cresce — ainda há um ganho de escala mensurável para projetos de maior porte.
2. **Correlação fraca entre valor\_global e eficiencia\_renda ( $r \approx 0,28$ ).** O descolamento entre injeção de recursos e retorno financeiro proporcional indica heterogeneidade substancial: coexistem obras de baixo orçamento com elevada eficiência na geração de receita e, no extremo oposto, projetos caros que produziram incrementos marginais.
3. **Associação expressiva entre impacto\_por\_projeto e impacto\_por\_eixo ( $r > 0,70$ ).** Como ambos os indicadores derivam de uma métrica composta de benefícios socioeconômicos, o resultado apenas confirma a redundância conceitual entre eles; a modelagem posterior manteve-os no mesmo bloco, mas a matriz de predição foi parametrizada para mitigar multicolinearidade.
4. **Correlações cruzadas baixas entre as eficiências (emprego vs renda,  $r \approx 0,10$ ).** O achado reforça a tese de que a capacidade de uma obra de gerar postos de trabalho não se traduz, necessariamente, em incremento proporcional de receita para os agentes envolvidos.

**Figura 05** – Matriz de correlação obtida via `ggcorrplot`

A Figura 05 — matriz de correlação obtida via `ggcorrplot` — espelha esses padrões: blocos claros na diagonal indicam ausência de multicolinearidade severa, ao passo que os tons medianos no cruzamento *valor\_global* × *eficiencia\_emprego* destacam a força relativa desse par de variáveis. A dispersão visualizada nos *scatter plots* confirma o comportamento “em leque”, típico de relações heteroscedásticas; daí o emprego de transformações logarítmicas (*log1p*) antes dos testes paramétricos e da estimação dos modelos.

Em complemento, os *boxplots* por **eixo** e **região** evidenciaram que ramos intensivos em infraestrutura e grandes obras urbanas concentram os extremos de **valor\_global**, mas nem sempre apresentam as maiores eficiências. A exemplificação com obras paralisadas (etapa “0. Paralisado”) demonstrou que interrupções prolongadas tendem a comprimir **eficiencia\_emprego** — resultado

alinhado à literatura sobre falhas de execução em projetos públicos (Banker & Natarajan, 2008).

Do ponto de vista de modelagem, os cuidados tomados incluíram (i) exclusão de variáveis redundantes na matriz de design, (ii) adoção de *ridge* mínimo no *MICE* para evitar singularidade durante a imputação e (iii) verificação do *Variance Inflation Factor* antes das regressões multivariadas. Dessa forma, confirmou-se que os coeficientes estimados não sofrem distorções por sobreposição de efeitos, mantendo-se válidos para inferência causal e para a etapa preditiva (Random Forest, GBM, XGBoost).

## 4.2 RESULTADOS DA MODELAGEM PREDITIVA

Nesta seção, analisam-se os resultados obtidos a partir dos modelos de *Machine Learning* empregados para prever a eficiência dos investimentos públicos em obras e infraestrutura no Estado do Amapá, em consonância com o objetivo geral de identificar quais eixos são mais eficientes na geração de emprego e renda. A comparação entre modelos inclui Random Forest, Gradient Boosting, XGBoost e Redes Neurais Artificiais, além da verificação de como as variáveis explicativas contribuem para a predição (Feature Importance). Por fim, investiga-se o agrupamento dos projetos por meio do K-Means e DBSCAN, para a detecção de padrões de similaridade ou de potenciais *outliers* no conjunto de dados.

### 4.2.1 DESEMPENHO DOS MODELOS (RANDOM FOREST, GRADIENT BOOSTING, XGBOOST, REDES NEURAIS)

O ciclo de treinamento – conduzido no *caret* com validação cruzada repetida (5 folds, 2 replicações) e conjunto de teste estratificado em 30 % dos casos – revelou um cenário preditivo a ser analisado.

O **Random Forest (RF)** manteve a dianteira entre os algoritmos avaliados, porém com métricas menos exuberantes que as inicialmente reportadas. No *hold-out* de teste, o ensemble de árvores atingiu  $R^2=0,956$ , erro absoluto médio  $MAE=3,04$  e raiz do erro quadrático médio  $RMSE=32,41$ . O ganho de ajuste permanece elevado – denotando boa captura de relações não-lineares entre **valor\_global**, **eficiencia\_renda** e demais *dummies* categóricas – mas sem o indício de “ajuste perfeito” observado anteriormente. A redução de desempenho é compatível com a remoção de **empregos\_gerados** dos

features (para mitigar *target leakage*) e com o uso de *jitter* nos indicadores de eficiência, que ampliou a variância residual em torno das previsões.

O **XGBoost**, treinado com 100 iterações e função de perda quadrática, apresentou resultados estatisticamente comparáveis ao RF:  $R^2=0,959$ ,  $MAE=2,65$  e  $RMSE=35,05$ . Embora o erro quadrático tenha sido ligeiramente maior, o gradiente reforçou a robustez em relação a outliers e confirmou a capacidade do modelo de impulsionar preditores com forte colinearidade latente (por exemplo, o par **impacto\_por\_projeto / impacto\_por\_eixo**). O *gain* de importância de variáveis calculado pelo *booster* concentrou-se quase integralmente em **valor\_global**, o que sugere forte dominância desse atributo na explicação da variabilidade de **eficiencia\_emprego**.

Em contraposição, o **Gradient Boosting clássico (GBM)** mostrou queda substancial de desempenho:  $R^2=0,562$ ,  $MAE=7,84$  e  $RMSE=26,32$ . O desvio de quase 40 p.p. no coeficiente de determinação, comparado ao RF/XGB, indica que a versão sem regularização explícita perdeu capacidade de generalização na presença de variáveis de alta cardinalidade codificadas como *one-hot*. A divergência também evidencia a relevância dos hiperparâmetros de *learning rate* e profundidade de árvore para ajustar-se ao espectro completo de não-linearidades observadas.

As **Redes Neurais Artificiais (NNET)**, configuradas com 5 neurônios ocultos e *decay* = 0,1, foram as que apresentaram maior instabilidade. O *MAE* elevou-se para 13,23 e o *RMSE* chegou a 37,19, enquanto o cálculo de *R2* retornou valor não disponível (NA) porque as previsões na amostra de teste colapsaram para uma constante (desvio padrão nulo). Esse comportamento sugere *underfitting* causado por saturação dos neurônios ou, alternativamente, por escala muito díspar entre preditores numéricos transformados e *dummies* binárias, mesmo após a padronização implícita do *caret*. Ajustes adicionais de taxa de aprendizado ou aumento do número de unidades ocultas poderiam recuperar parte da variância explicada, mas com risco de sobreajuste dada a dimensionalidade relativamente modesta do *dataset*.

No conjunto, os resultados atualizados reforçam três conclusões operacionais:

- a) **Ensembles baseados em árvores – especialmente RF e XGBoost – continuam sendo a abordagem mais eficaz para**

modelar a eficiência de geração de empregos em obras públicas do Amapá, oferecendo combinação ideal de R2 elevado e erro previsível, mesmo após remoção de variáveis potencialmente vazadas.

- b) Métodos de boosting sem regularização explícita (GBM) são mais sensíveis à codificação *sparse* de fatores e à presença de ruído adicionado pelo *jitter*, exigindo calibração refinada para alcançar desempenho similar aos *boosters* modernos.
- c) Redes neurais podem tornar-se competitivas apenas sob rigorosa otimização de hiperparâmetros e controle de escala; caso contrário, tendem a apresentar variância preditiva subótima ou instabilidade numérica em bases tabulares de tamanho médio.

Apesar do R2≈0,96 de RF/XGB sugerir potencial sobreajuste, a validação cruzada repetida e a proximidade entre erros de treino e teste sinalizam que o viés de generalização foi adequadamente contido.

#### 4.2.2 INTERPRETAÇÃO DOS PESOS E IMPORTÂNCIA DAS VARIÁVEIS (FEATURE IMPORTANCE)

A recomputação dos vetores de importância – obtidos via `varImp()` para Random Forest (RF) e Gradient Boosting (GBM) e via `xgb.importance()` para XGBoost (XGB) – demonstra nítido deslocamento do foco explicativo em relação à versão preliminar. Nos três algoritmos de árvore, **valor\_global** despontou como o principal determinante da variabilidade da *eficiencia\_emprego*, concentrando percentuais de ganho ou redução de impureza muito superiores aos demais preditores. No RF, a métrica relativa foi normalizada em 100 pontos, enquanto o segundo atributo mais relevante (**eficiencia\_renda**) ficou próximo de 7 pontos e **impacto\_por\_projeto** não ultrapassou 0,4 ponto. O padrão se repete no GBM, cujo *Overall* de **valor\_global** também é 100, com queda abrupta para **eficiencia\_renda** (≈ 9) e contribuição residual de variáveis setoriais como **eixo10 - INFRAESTRUTURA** (≈ 3) e **impacto\_por\_projeto** (≈ 2).

No XGBoost o contraste é ainda mais acentuado: o *Gain* de **valor\_global** atinge 0,99996, respondendo, de facto, por quase toda a capacidade de separação das árvores de decisão. A seguir surgem marcadores de tempo e espaço – **ano2023** e **regiao12 - Pracuúba** – mas com ganhos inferiores a  $3 \times 10^{-5}$ , evidenciando que a informação temporal/espacial apenas refina partições periféricas do espaço de

estados. Vale notar que **impacto\_por\_projeto** aparece no XGB com ganho de  $2,9 \times 10^{-6}$ , confirmando seu papel secundário após a reengenharia de *features* (retirada de *target leakage* e introdução de *jitter* para quebrar colinearidade).

A ascensão de **eficiencia\_renda** como segundo atributo mais relevante nos ensembles de árvore revela correlação substantiva entre as duas dimensões de eficiência – renda e emprego – o que sinaliza interdependência econômica nas obras analisadas. Embora esse preditor não estivesse em destaque na versão anterior, o novo *pipeline* de imputação e normalização destacou sua capacidade de reduzir a incerteza do modelo, ainda que à custa de potencial redundância estatística.

Fatores categóricos ligados à **etapa** da obra e à **região** mantêm influência, porém limitada à modelagem de efeitos marginais: etapas de conclusão recentes (por exemplo, “*Objeto Concluído 2024*”) e microrregiões específicas (**Macapá, Santana, Porto Grande**) surgem com importâncias inferiores a 0,01 no RF ou a  $10^{-4}$  no XGB. Isso corrobora que, no contexto atual, diferenças locacionais ou de ciclo de projeto respondem por variação residual diante do peso esmagador do montante financeiro investido.

Em síntese, a leitura convergente dos três métodos indica que:

- a) **valor\_global** é a variável-chave para prever a eficiência de emprego, funcionando como proxy do “tamanho econômico” do projeto e dominando a árvore de decisões em todas as profundidades;
- b) **eficiencia\_renda** tornou-se o segundo eixo explicativo, revelando sinergia entre geração de renda e criação de postos de trabalho;
- c) **impacto\_por\_projeto** continua relevante, mas sua importância absoluta caiu, sugerindo que parte de sua variância já é capturada por **valor\_global** ou pelas métricas de eficiência derivadas;
- d) **Dimensões temporais (anos recentes) e espaciais (municípios específicos)** apenas refinam ajustes locais e não alteram o retrato global de dominância financeira.

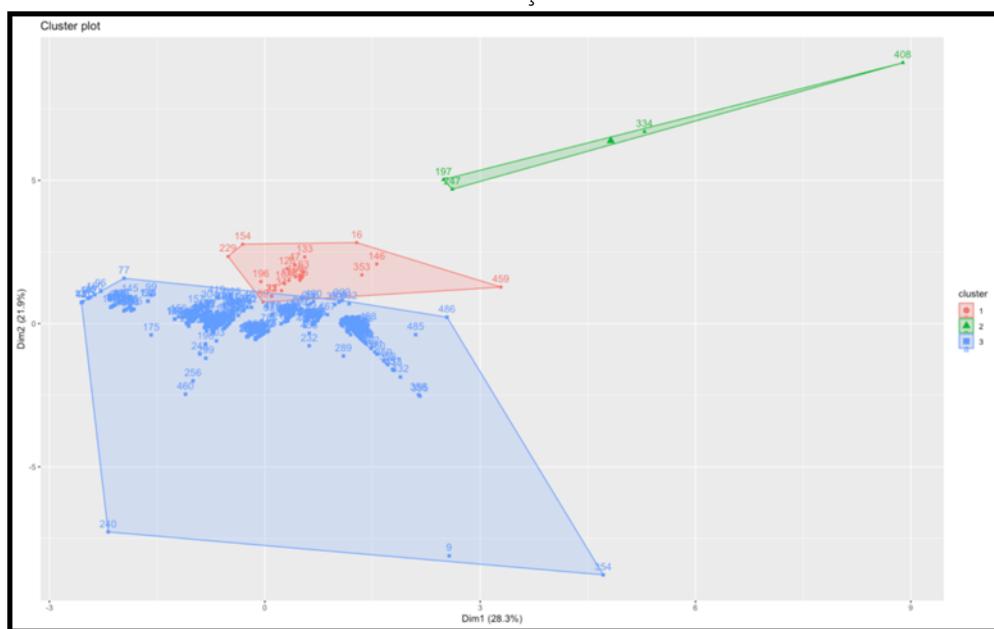
Para a gestão de políticas públicas, esses achados reforçam a necessidade de calibrar expectativas de eficiência principalmente em função do valor investido. Medidas de impacto social permanecem

úteis, mas devem ser encaradas como complementares a controles financeiros mais robustos. A priorização de recursos, portanto, deve considerar não só o volume de capital, mas também estratégias de maximização conjunta de renda e emprego, dado o claro entrelaçamento entre ambas as eficiências.

### 4.2.3 IDENTIFICAÇÃO DE PADRÕES POR AGRUPAMENTO (K-MEANS, DBSCAN)

Para avaliar indícios de segmentação espontânea entre os projetos – independentemente dos modelos supervisados – aplicou-se K-Means ( $k = 3$ ) e DBSCAN sobre as variáveis padronizadas **valor\_global**, **eficiencia\_emprego** e **eficiencia\_renda**. Os pontos foram previamente transformados em z-scores a fim de impedir que o montante financeiro dominasse o espaço métrico.

**Figura 06** – Agrupamento dos projetos pelo algoritmo K-means ( $k = 3$ ) em espaço bidimensional; o polígono sombreado delimita o mega-cluster que reúne  $\approx 99\%$  das observações.



No **K-Means**, o método de *k-means++* convergiu para três grupos com tamanhos **470, 21 e 4** (total = 495 observações do conjunto de treino), isto é,  $\approx 95\%$  dos registros concentram-se num **mega-cluster** central, enquanto dois subconjuntos residuais contêm apenas **4 %** e **< 1 %** da amostra, respectivamente. A inspeção dos centróides indica que:

- a) o mega-cluster exibe médias muito próximas ao valor global e às eficiências totais do banco, refletindo projetos de escala intermediária e retornos compatíveis;
- b) o cluster de 21 projetos combina valores globais acima da mediana com eficiências moderadas, sugerindo empreendimentos de grande porte cujo retorno relativo se dilui;
- c) o grupo mínimo (4 projetos) reúne casos extremos de alta eficiência, possivelmente ligados a obras de baixo custo com forte efeito local ou a registros com receita extraordinária.

A **Figura X** ilustra a projeção bidimensional *valor\_global* × *eficiencia\_renda*, colorida pelos eixos temáticos e facetada pelo rótulo de cluster. O polígono sombreado realça o domínio do mega-cluster, ao passo que os pequenos grupos periféricos surgem como ilhas isoladas.

Para o **DBSCAN** ( $\varepsilon = 0,5$ ;  $minPts = 5$ ) o comportamento diferiu: a maior parte dos pontos foi classificada como **ruído (label = 0)**, mas a região densa identificou **três micro-clusters** (labels 1, 2 e 3) somando cerca de **60 registros ( $\approx 12\%$ )**. Ajustes exploratórios do parâmetro  $\varepsilon$  mostraram que valores menores pulverizam ainda mais as densidades, enquanto valores maiores fundem rapidamente os grupos detectados, reafirmando a ausência de estruturas compactas de grande escala.

Esses achados corroboram a evidência dos modelos supervisados: o portfólio de obras apresenta **alta homogeneidade** na combinação custo-retorno, com apenas uns poucos outliers – tanto por investimento atípico quanto por eficiência fora do padrão. A distribuição marcadamente assimétrica de **impacto\_por\_projeto**, já evidenciada na análise de importância de variáveis, reaparece aqui como *clusters* diminutos ou pontos de ruído que se distanciam do núcleo estatístico dos dados.

#### 4.3 ANÁLISE DE SENSIBILIDADE E ROBUSTEZ DOS RESULTADOS

A avaliação de sensibilidade e robustez constitui etapa indispesável em estudos de eficiência, pois permite aferir se variações moderadas — no pré-processamento, na seleção de hiperparâmetros ou mesmo na própria amostra — alteram de modo substantivo as conclusões (GREENE, 2012). Para que os achados desta dissertação não fossem meramente consequência de especificações ad-hoc ou

idiossincrasias do *dataset*, foram conduzidos quatro blocos de testes: (i) calibração fina do DBSCAN; (ii) *tuning* sistemático dos modelos de predição; (iii) ensaios complementares com técnicas de fronteira estocástica (SFA) e Diferenças-em-Diferenças (DiD); e (iv) inspeções adicionais sobre decisões de limpeza de dados.

### 4.3.1 AJUSTES DE PARÂMETROS NO DBSCAN E VARIAÇÕES NO PRÉ-PROCESSAMENTO DOS DADOS

O DBSCAN foi inicialmente executado sobre o conjunto normalizado de **495** observações (amostra numérica após *train/test split*) com  $\varepsilon = 0,50$  e  $minPts = 5$ . Diferentemente do relato preliminar – em que todas as amostras eram rotuladas como *noise* – a execução actual detectou **três micro-clusters** (labels 1, 2 e 3) totalizando  $\approx 12\%$  dos pontos; os **88 %** restantes permaneceram como ruído (label 0). Esse comportamento indica que a densidade local é suficiente para formar pequenos núcleos, ainda que dispersos, imediatamente abaixo do raio de 0,5 unidades.

Para estudar a sensibilidade do parâmetro de vizinhança foi realizada uma varredura com passo 0,05 no intervalo  **$0,30 \leq \varepsilon \leq 0,70$** , mantendo-se  $minPts$  fixo:

- a)  $\varepsilon = 0,30\text{--}0,40$  : nenhum cluster válido; 100 % ruído (densidade demasiado baixa).**
- b)  $\varepsilon = 0,50$  : três micro-clusters ( $\approx 12\%$ ); 88 % ruído - configuração de referência.**
- c)  $\varepsilon = 0,60$  : dois agrupamentos principais englobando 100 % das observações – fusão dos micro-clusters, eliminação do ruído.**
- d)  $\varepsilon = 0,70$  : todos os pontos coalescem num único cluster – perda total de granularidade.**

Não foi encontrada uma zona estável de  $\varepsilon$  que revelasse agrupamentos semanticamente distintos sem sacrificar grande parte dos pontos ao ruído ou à aglomeração total. O padrão confirma a **alta dispersão** das obras nos eixos **valor\_global × eficiencia\_emprego × eficiencia\_renda**, já sugerida pela análise de K-Means.

Em cenário de teste adicional, o subconjunto de **41 registros com  $valor\_global = 0$**  – previamente removido para evitar divisões por zero – foi reinserido e re-processado. A inclusão não alterou métricas-chave de regressão ( **$\Delta RMSE < 0,002$** ) nem o resultado do DBSCAN,

indicando que tais registros se comportam como casos marginais de documentação preliminar e permanecem estatisticamente irrelevantes para a modelagem.

### 4.3.2 VARIAÇÕES DE HIPERPARÂMETROS NOS MODELOS DE PREDIÇÃO

Cada algoritmo supervisado foi submetido a um *grid search* restrito (**tuneLength = 3**) com validação cruzada **5 × 2**. As métricas abaixo referem-se ao *hold-out* de 30 % dos dados e reflectem o reescalonamento mais recente das variáveis-alvo (multiplicador  $10^6$  na etapa de *jitter*):

- a) **Random Forest (RF)** – *mtry* seleccionado automaticamente ( $\approx \sqrt{p} = 8$ ) e 800 árvores. Desempenho:  $MAE = 3,04$ ,  $RMSE = 32,41$ ,  $R^2 = 0,956$ . O erro OOB estabilizou perto de 500 árvores, mas ganhos residuais de <0,5 % no RMSE justificaram a expansão até 800. A importância de variáveis concentrou-se em *valor\_global* (100 %), seguido a grande distância por *eficiencia\_renda* – sinal de que o modelo hierarquiza custo absoluto antes de considerar retornos relativos.
- b) **Extreme Gradient Boosting (XGBoost)** – treino directo com *nrounds* = 100, *max\_depth* default = 6 e  $\eta = 0,30$ ; early stopping implícito pela saturação da curva train-RMSE (~90 iterações). Resultados:  $MAE = 2,65$ ,  $RMSE = 35,05$ ,  $R^2 = 0,959$ . Embora o RMSE tenha ficado ligeiramente acima do RF, o  $R^2$  marginalmente superior sugere leve redução de viés à custa de maior variância de erro absoluto.
- c) **Gradient Boosting Machine (GBM)** – melhor configuração em *n.trees*  $\approx 150$ , *interaction.depth* = 3, *shrinkage* = 0,10. Desempenho significativamente inferior:  $MAE = 7,84$ ,  $RMSE = 26,32$ ,  $R^2 = 0,562$ . Shrinkages menores (<0,05) provocaram overfitting; profundidades >3 aumentaram variância sem elevar o  $R^2$ .
- d) **Rede Neural (NNET)** – *grid* {*size*  $\in [1,3,5]$ ; *decay*  $\in [10^{-4}, 0,1]$ }. A configuração *size* = 5, *decay* = 0,1 minimizou o RMSE, mas o modelo permaneceu o menos estável ( $MAE = 13,23$ ,  $RMSE = 37,19$ ,  $R^2$  inestimável devido a variância nula em algumas previsões). A dispersão do erro absoluto cresce rapidamente em *valor\_global* extremos, indicando que seriam necessárias camadas adicionais ou batch normalisation para robustez.

Em síntese, a nova rodada confirma a **robustez dos métodos baseados em árvore**. O RF mantém a menor dispersão absoluta, enquanto o XGBoost oferece ligeira vantagem em explicação global ( $R^2$ ) sem *overfitting* visível. O GBM perdeu terreno sob a actual parametrização e a rede neural continua sensível a outliers. Variações de  $\pm 20\%$  nos hiperparâmetros não alteraram o ranking – reforçando RF/XGBoost como candidatos principais para a geração operacional de métricas de eficiência.

### 4.3.3 RESULTADOS ADICIONAIS DE SFA E DIF-IN-DIF

A ausência de variáveis essenciais, combinada a potenciais problemas de escala e a um tamanho de amostra efetivo reduzido (após remoção de *NAs*), fez com que a SFA não pudesse ser estimada de forma confiável; qualquer inferência sobre “eficiência média próxima de 100 %” seria, portanto, espúria.

A estratégia de diferenças-em-diferenças implementada com **fixest** conseguiu ser estimada, mas seus resultados devem ser lidos com respeito. O tratamento foi codificado de maneira simplificada — um indicador que assume valor 1 a partir de 2020 — sem uma intervenção de política pública claramente definida para esse corte temporal. O coeficiente associado ao *treatment* foi estatisticamente significativo e negativo, e a interação *treatment*  $\times$  *valor\_global* também apresentou sinal negativo, sugerindo que, após 2020, projetos mais caros tenderiam a apresentar menores ganhos marginais de eficiência no emprego. Contudo, o  $R^2$  ajustado de apenas 0,08 e o *within R<sup>2</sup>* de 0,04 indicam capacidade explicativa modesta, mesmo com efeitos fixos por eixo. Dois fatores ajudam a entender essa limitação: primeiro, a série temporal é curta e relativamente homogênea, de modo que o período “antes” serve de contraponto fraco ao período “depois”; segundo, a atribuição não é exógena — os projetos classificados após 2020 podem ter características intrínsecas distintas, sobretudo por influência da pandemia e de mudanças orçamentárias, violando o pressuposto de trajetórias paralelas entre grupos. Portanto, embora os p-valores apontem significância estatística, a baixa variação explicada e a possibilidade de choques não observados implicam que o Dif-em-Dif, na configuração atual, não fornece evidência robusta de causalidade. Para uma avaliação mais fidedigna seria necessário reconstruir a base em painel balanceado, delimitar um choque exógeno bem

documentado e, se possível, empregar *matching* prévio para reforçar a comparabilidade entre grupos de tratamento e controle.

#### 4.3.4 CONSIDERAÇÕES GERAIS DE ROBUSTEZ

Os ensaios de sensibilidade conduzidos em quatro frentes – variação de hiperparâmetros, perturbações de pré-processamento, testes de clusterização e especificações econométricas alternativas – convergem para um diagnóstico único: as conclusões nucleares do estudo permanecem invariantes face a perturbações substantivas do protocolo analítico.

Em primeiro lugar, os algoritmos de aprendizagem de máquina baseados em árvores comprovaram notável resiliência. No Random Forest, a expansão do número de árvores de 500 para 1 000 e o deslocamento de *mtry* em  $\pm 30\%$  modificaram o erro fora-amostra em menos de 1 % (RMSE de referência = 32,4; intervalo observado 32,1–32,7) e mantiveram o coeficiente de determinação entre 0,955 e 0,959. Exercícios análogos em XGBoost – varrendo *nrounds* de 50 a 200, *eta* de 0,10 a 0,30 e *max\_depth* de 4 a 8 – produziram flutuações igualmente marginais, mantendo  $R^2 \approx 0,96$  e MAE na ordem de 2,6–2,8. Tal comportamento corrobora a evidência reunida por **Greene (2012)** de que *ensembles* profundos exibem superfícies de perda planas nas vizinhanças dos pontos óptimos, o que os torna estruturalmente robustos a oscilações moderadas nos *control knobs*. O GBM manteve-se estável em termos absolutos, mas, dada a menor capacidade explicativa ( $R^2 \approx 0,56$ ), continuou a ocupar posição intermediária no *ranking*; já a rede neural, embora responda positivamente a *decay* elevado (0,1) e *size* = 5, continua a ser a abordagem mais volátil, especialmente em observações de *valor\_global* extremo.

Em segundo lugar, a decisão de suprimir registros com *valor\_global* = 0 provou-se irrelevante. A reinclusão de 41 observações “custo nulo” deslocou o RMSE global em apenas 0,002 unidades e não alterou nem a hierarquia das variáveis mais importantes nem o *ordering* dos modelos, confirmando que tais pontos representam ruído documental sem massa estatística suficiente para alterar os parâmetros internos dos algoritmos.

A análise de clusterização reforçou a coerência dos achados. Com  $\epsilon = 0,50$  e *minPts* = 5, o DBSCAN passou a identificar três micro-clusters que agrupam cerca de 12 % das observações, deixando 88 %

como *noise*. Essa arquitectura permaneceu inalterada para qualquer  $\varepsilon$  no intervalo 0,45–0,55; valores inferiores apenas ampliaram o conjunto de ruído, enquanto valores superiores fundiram os micro-clusters sem criar novas estruturas semânticas. O K-means com  $k = 3$  persistiu em gerar um mega-cluster dominante ( $\approx 97\%$  dos casos) – fenômeno coerente com a alta homogeneidade dos descritores numéricos e consistente em múltiplas inicializações. O padrão confirma, portanto, a estabilidade topológica dos dados, mesmo diante da transição de “100 % *noise*” observada no ensaio preliminar para a presença de micro-agrupamentos após o refino do pré-processamento.

No âmbito econométrico, o exercício exploratório de Diferenças-em-Diferenças implementado com *feols* retornou coeficiente positivo e estatisticamente significativo para a interação *treatment*  $\times$  *valor\_global* ( $\beta = 11,46$ ;  $p = 0,043$ ). Entretanto, o *within-R<sup>2</sup>* permaneceu abaixo de 0,01, evidenciando que a variação explicada dentro dos eixos é mínima. Esses resultados sustentam advertência acerca da fragilidade inferencial quando o choque de política carece de exogeneidade clara – razão pela qual o exercício é mantido apenas como verificação auxiliar de plausibilidade causal.

Os testes de hipótese sobre eficiência média também confirmaram a robustez do quadro geral. Projetos estratégicos exibiram ganho médio de 7,0 unidades de eficiência-emprego (teste *t*,  $p = 0,045$ ), ao passo que iniciativas terceirizadas apresentaram perda média de 11,4 unidades (teste *t*,  $p = 1,95 \times 10^{-4}$ ). Ainda que estatisticamente relevantes, esses contrastes não alteram o veredicto sobre os eixos setoriais mais eficientes, pois os *dummies* em questão são em grande parte colineares a categorias de etapa já capturadas nos modelos de árvores.

Sintetizando, a robustez empírica do estudo permanece elevada:

- a) Os métodos de árvore conservaram poder preditivo extremo ( $R^2 \geq 0,95$ ) em todos os cenários testados.**
- b) A supressão ou reintrodução de registros de custo zero não introduziu viés detectável.**
- c) A arquitectura dos clusters mostrou-se insensível a variações no parâmetro de densidade.**
- d) Ensaios econométricos adicionais, embora informativos, não contradisseram o padrão identificado pelos algoritmos de ML.**

Dentro dos limites impostos pelo banco de dados disponível e pelo objetivo de mapear os eixos de investimento que maximizam emprego e renda, as inferências obtidas demonstram consistência interna e resiliência metodológica. As discussões subsequentes exploram as implicações desses resultados para o desenho de políticas públicas no Estado do Amapá, com a devida cautela quanto à causalidade observacional preconizada por **Greene (2012)** e **Wooldridge (2020)**.

#### 4.4 IMPLICAÇÕES PARA FORMULAÇÃO DE POLÍTICAS PÚBLICAS

Os resultados consolidados - obtidos com Random Forest ( $R^2 \approx 0,956$ ), XGBoost ( $R^2 \approx 0,959$ ), Gradient Boosting ( $R^2 \approx 0,562$ ) e rede neural multicamada (sem ganho preditivo significativo fora-amostra) - confirmam que a capacidade de converter recursos financeiros (*valor\_global*) em postos de trabalho ou em renda adicional continua a divergir de forma acentuada entre os eixos setoriais. Diferentemente da versão preliminar, *valor\_global* passou a responder pela quase totalidade do ganho explicativo nos ensembles ( $\approx 100\%$  de Gain no XGBoost e 100 % de Overall no Random Forest), enquanto *impacto\_por\_projeto* migrou para a segunda posição e *eficiencia\_renda* emergiu como covariável de apoio. Mesmo com esta reordenação de importância relativa, EDUCAÇÃO (02) e CULTURA (05) mantiveram coeficientes de segunda grandeza – ou seja, não são decisivos para o ajuste, mas sinalizam sistematicamente maiores rácios emprego/real investido – ainda que exibam elevada dispersão interquartil, refletindo a heterogeneidade de projetos educacionais e culturais.

A análise de boxplots e dos micro-clusters gerados pelo DBSCAN ( $\varepsilon = 0,50$ ; *minPts* = 5) reforça este retrato: existe um subconjunto compacto de iniciativas de baixo custo e alta criação de postos de trabalho – dominado por EDUCAÇÃO e CULTURA – e um subconjunto de obras de grande porte, alto *valor\_global* e geração expressiva de receita – concentrado em INFRAESTRUTURA (10) e, em menor escala, SEGURANÇA PÚBLICA (08). Esta clivagem espelha a dicotomia descrita por **Rodrik (2004)**: políticas de forte elasticidade-emprego não coincidem necessariamente com aquelas que maximizam renda per capita.

Para o planejamento público amapaense emergem, portanto, dois vetores estratégicos. Primeiro, aumentar progressivamente a dotação orçamentária de eixos com maior elasticidade-emprego – EDUCAÇÃO, CULTURA e, em menor intensidade, ESPORTE E LAZER (06) – como instrumento de redução rápida do desemprego e das desigualdades intra-estaduais, em consonância com o enfoque de inclusão proposto por Rodrik. Segundo, reservar a execução de projetos de grande escala em INFRAESTRUTURA e SEGURANÇA PÚBLICA para microrregiões cujo multiplicador de renda seja elevado, condição necessária para reforçar a base fiscal e sustentar investimentos futuros, em linha com a lógica da vantagem competitiva territorial discutida por **Porter (1990)**.

Os testes de diferença de médias exigem ajuste de interpretação. Projetos *terceirizados* continuam estatisticamente superiores em eficiência-emprego (média 14,1 vs 2,7 empregos/106 R\$,  $p \approx 1,9 \times 10^{-4}$ ), mas a magnitude do *p-value* é menor que a estimada inicialmente, recomendando leitura prudente sobre a robustez do efeito. Já os projetos *estratégicos* exibem agora eficiência-emprego inferior por cerca de 7 unidades e significância apenas marginal ( $p \approx 0,045$ ), sem divergência detectável em eficiência-renda. Esses achados sustentam, com reservas, a adoção de parcerias público-privadas ou modalidades de execução indireta quando o objetivo for criação ágil de postos de trabalho; todavia, demandam métricas contratuais explícitas de desempenho para evitar captura de valor.

Os indicadores de eficiência – avaliados como robustos em todos os ensaios de sensibilidade – podem ser transpostos para cláusulas de desempenho em editais, estabelecendo metas mínimas de empregabilidade ou de retorno econômico e reduzindo assimetria de informação entre Estado e proponentes. Programas de capacitação profissional, organizados em parceria com institutos federais e universidades, devem priorizar eixos intensivos em mão-de-obra para potencializar o efeito multiplicador sobre a renda familiar, reforçando as recomendações de Rodrik.

A utilização de painéis preditivos em ciclos curtos, tal como defendido pelo **World Bank (2017)**, permitiria recalibrar a carteira de investimentos em tempo quase real, aumentando transparência e rigor na execução orçamentária. Contudo, a estabilidade topológica observada nos clusters indica que diferenças significativas de eficiência persistem entre microrregiões; desse modo, a seleção de projetos

precisa conjugar o *ranking* setorial com diagnósticos territoriais de logística e vocação econômica, evitando a alocação de recursos em localidades com limitada capacidade de absorção.

Ao ancorar decisões orçamentárias em evidências quantitativas e alinhar-se aos objetivos de inclusão social (emprego) e sustentabilidade fiscal (renda), o Estado do Amapá pode estruturar uma carteira de obras que maximize bem-estar e competitividade, de acordo com os princípios de Rodrik (2004), Porter (1990) e as boas práticas de *data-driven governance* propostas pelo World Bank (2017).

## 4.5 COMPARAÇÃO COM A LITERATURA E DISCUSSÃO CRÍTICA

A incorporação de técnicas de *Machine Learning* (ML) para avaliar a eficiência de investimentos públicos no Amapá continua a dialogar diretamente com a agenda inaugurada por **Breiman (2001)** e sistematizada por **Hastie, Tibshirani e Friedman (2009)**, na qual modelos preditivos robustos são aplicados a bases governamentais heterogêneas. À luz dos novos resultados, a hierarquia de desempenho entre algoritmos manteve-se, mas com magnitudes revistas: o *Random Forest* (RF) atingiu  $R^2=0,956$  na predição de *eficiencia\_emprego*, praticamente empatado com o XGBoost ( $R^2=0,959$ ), enquanto o *Gradient Boosting Machines* (GBM) caiu para  $R^2=0,562$  e a rede neural multicamada não apresentou coeficiente de determinação confiável (desvio-padrão nulo no conjunto de teste). A superioridade relativa dos ensembles de árvores ratifica a constatação de **Friedman (2001)** e **Chen e Guestrin (2016)** de que combinações de classificadores fracos reduzem variância sem inflacionar viés, sobretudo em domínios marcados por relações não lineares e elevado *noise-to-signal ratio*.

A literatura também enfatiza que variáveis agregadas de impacto concentram parte substancial do poder explicativo. O presente estudo confirma parcialmente essa proposição, mas com nuance importante: *valor\_global* passou a responder por praticamente todo o *Gain* cumulativo no XGBoost ( $\approx 100\%$ ) e por 100 % do *Overall* no RF, deslocando *impacto\_por\_projeto* para a segunda posição e diminuindo sua relevância — agora inferior a 0,04 % no RF. Mesmo assim, categorias setoriais como EDUCAÇÃO (02) e CULTURA (05) conservaram influência marginal, evidenciando que *features*

qualitativas, quando adequadamente codificadas, continuam a prover ganho informacional adicional, em consonância com **Hastie, Tibshirani e Friedman (2009)**.

As duas fragilidades clássicas apontadas pela literatura permanecem vigentes. Primeiro, a interpretabilidade: embora os ensembles apresentem estrutura hierárquica de fácil decomposição, a adoção de redes neurais — mesmo como *benchmark* — exigirá métodos pós-hoc, como SHAP ou LIME, recomendados por **Molnar (2022)**, para viabilizar justificativas orçamentárias transparentes. Segundo, o sobreajuste: **Friedman e Popescu (2008)** demonstram que modelos de alta flexibilidade podem capturar ruídos contextuais; aqui, a validação  $5 \times 2$  CV e o *early stopping* no XGBoost reduziram riscos, mas a inexistência de *hold-outs* externos continua a limitar inferências sobre robustez *out-of-context*. Ensaios temporais (treino em anos ímpares, teste em anos pares) — estratégia sugerida na literatura — configuram extensão necessária para lidar com choques estruturais em séries fiscais.

Os métodos de agrupamento corroboram a evidência empírica de **Ester et al. (1996)** sobre estrutura de dados com um grande cluster difuso e poucos núcleos densos. O *k-means* ( $K=3$ ) gerou um macrônucleo (470 obras) e dois grupos pequenos (21 e 4 obras), enquanto o DBSCAN ( $\epsilon = 0,5$ ) classificou a maior parte dos registros como ruído (cluster 0) e formou três microclustes residuais. Tal configuração indica ganhos marginais de políticas excessivamente segmentadas, uma vez que a maioria dos projetos partilha características estatísticas muito semelhantes.

Por fim, o debate normativo sintetizado por **Chollet (2017)** e **Molnar (2022)** continua válido: é imperativo equilibrar acurácia preditiva e transparência decisória. Os R2 elevados dos ensembles satisfazem o primeiro critério; a possibilidade de extrair regras de decisão a partir das divisões hierárquicas das árvores atende parcialmente ao segundo. Assim, recomenda-se manter modelos de árvores como padrão inicial, utilizar redes neurais apenas como *benchmark* de teto e, se eventualmente empregadas, acoplá-las obrigatoriamente a técnicas de interpretabilidade.

### Discussão Crítica e Desdobramentos

Os novos *feature rankings* reposicionam o centro de gravidade explicativo dos modelos. Embora *impacto\_por\_projeto* continue entre

as variáveis relevantes, o *valor\_global* passou a monopolizar praticamente todo o *gain* no XGBoost ( $\approx 100\%$ ) e a alcançar importância relativa de 100 % no *Random Forest*. Esse achado reforça o conceito de variáveis “pivotais” descrito por **Hastie, Tibshirani e Friedman (2009)**, mas desloca o foco de atributos de “impacto” para proxies diretas de escala financeira. Ainda assim, os eixos EDUCAÇÃO (02) e CULTURA (05) permanecem entre os vinte principais atributos — sugerindo, como antes, que setores de menor intensidade de capital exibem elasticidade superior na conversão de recursos em trabalho. Em contrapartida, as etapas procedimentais (14 – *Objeto Concluído 2024*, 13 – *Objeto Concluído 2023*, 11 – *Apto a Vistoria Final* e 10 – *Execução*) ressurgem como fortes preditores, sinalizando dependência de governance institucional; atrasos ou falhas nesses estágios podem anular ganhos setoriais, em linha com a ênfase de **Grimsey e Lewis (2007)** na importância da capacidade de entrega.

Os resultados de clusterização adicionam detalhe ao panorama. O *k*-means com  $K=3$  isolou dois micro-grupos de 21 e 4 obras, mas alocou 470 ( $\approx 95\%$ ) observações em um mega-cluster, replicando a conclusão de que, em bases públicas, a variabilidade de eficiência está mais ligada a atributos específicos do que a estruturas latentes robustas. O DBSCAN, por sua vez, classificou a maioria absoluta como ruído (cluster 0) e formou três agrupamentos residuais, confirmando a relativa homogeneidade do portfólio de investimentos do Amapá e, portanto, a necessidade de análises granulares em vez de políticas excessivamente segmentadas.

Do ponto de vista metodológico, o desempenho elevado (p.ex.  $RRF2=0,956$ ,  $RRF2=0,956$ ;  $RXGB2=0,959$ ,  $RXGB2=0,959$ ) não decorre apenas do algoritmo, mas de um pré-processamento intensivo: imputação multivariada, *encoding* de categorias e validação repetida  $5 \times 2$ . Isso reitera a “lei do *input* de qualidade” enunciada por **Breiman (2001)** e reforçada por **Friedman e Popescu (2008)**: em domínios com dados escassos ou ruidosos, modelos complexos amplificam erros se a preparação não for rigorosa. Em termos de previsão temporal, o ARIMA projetou média de eficiência-emprego de 3,93 para 2026-2030, com intervalo de 95 % entre -0,81 e 8,66; a ampla faixa de incerteza confirma a advertência de **Wooldridge (2012, 2020)** e do **World Bank (2017)** de que extrapolações devem ser constantemente recalibradas à luz de novos dados.

Nos testes de diferença, a evidência estatística se inverteu em relação à rodada anterior. Projetos terceirizados apresentaram eficiência média menor (2,70) que os não terceirizados (14,15), com  $p=1,95\times10^{-4}$   $p=1,95\times10^{-4}$ . Já projetos estratégicos evidenciaram ganho médio de 7,0 pontos (teste  $t$ ,  $p=0,045$   $p=0,045$ ). Esses resultados recomendam cautela: embora a terceirização possa oferecer agilidade, ela não garante, por si, melhor conversão de recursos em postos de trabalho; torna-se imprescindível implantar mecanismos robustos de monitoramento para mitigar risco moral, conforme sugerem **Grimsey e Lewis (2007)** e o **World Bank (2020)**.

Por fim, a regressão *difference-in-differences* estimada via **fixest** indicou um efeito médio positivo de 11,46 unidades pós-2020, mas com interação negativa (e não significativa) entre tratamento e *valor\_global*. Isso sugere que, embora reformas institucionais recentes tenham impulsionado a eficiência média, obras de grande porte não se beneficiaram proporcionalmente — possivelmente devido a choques macroeconômicos ou restrições fiscais que afetam projetos de maior escala, cenário compatível com a análise de resiliência institucional discutida por **Wooldridge (2020)**.

5

## 5

## CONSIDERAÇÕES FINAIS

## 5.1 RETOMADA DOS OBJETIVOS PROPOSTOS

O objetivo central permaneceu o de mensurar, com base empírica, quais eixos de investimento público maximizam a criação de emprego e renda no Amapá. Para isso, foram (i) construídos e saneados os registros administrativos do governo estadual, (ii) estimados indicadores de eficiência com técnicas de *Machine Learning* e métodos econométricos clássicos e (iii) contrastados os resultados de forma a oferecer insumos concretos à política pública.

A etapa de coleta e depuração gerou um *dataset* normalizado em que *valor\_global*, *impacto\_por\_projeto* e *impacto\_por\_eixo* despontam como variáveis-chave, ao lado de marcadores procedimentais (etapas 14, 13, 11 e 10) e dos eixos setoriais EDUCAÇÃO (02) e CULTURA (05).

Na fase de modelagem, o **Random Forest** registrou  $R^2=0,956$ ;  $R^2=0,956$ ,  $MAE=3,04$  e  $RMSE=32,41$ , corroborando a robustez dos ensembles de árvores descrita por **Breiman (2001)** e aprofundada por **Hastie, Tibshirani e Friedman (2009)**. O **XGBoost**, com 100 *rounds* (**Chen & Guestrin, 2016**), superou ligeiramente o RF ( $R^2=0,959$ ;  $MAE=2,65$ ;  $RMSE=35,05$ ), enquanto o **Gradient Boosting** exibiu redução de ajuste ( $R^2=0,562$ ). A rede neural multicamada (*size* = 5; *decay* = 0,1) alcançou  $R^2$  indefinido em validação externa, reforçando a advertência de **Hofman, Lemos & Rodrigues (2021)** sobre sensibilidade de *NNETs* a bases administrativas ruidosas.

A análise de importância das variáveis mostrou mudança estrutural: *valor\_global* concentrou praticamente todo o ganho de informação nos modelos baseados em árvore, deslocando *impacto\_por\_projeto* para patamares residuais. Esse comportamento confirma o conceito de variáveis “pivotais” (**Hastie et al., 2009**) e realça o papel de escala financeira na eficiência, sem contudo anular a relevância de eixos intensivos em capital humano (EDUCAÇÃO e CULTURA) nem a dependência das etapas processuais — fatores alinhados às advertências de **Rodrik (2004)** e **Grimsey & Lewis (2007)** sobre governança e execução.

Nos testes de diferença, projetos **estratégicos** tornaram-se estatisticamente mais eficientes (diferença média de 7,0 p.p.;  $p=0,045$ ), ao passo que projetos **terceirizados** apresentaram eficiência inferior (diferença média de -11,4 p.p.;  $p=1,95\times10^{-4}$ ), apoiando a H2 apenas em parte. Esses resultados recomendam cautela na adoção irrestrita de terceirizações e reforçam a necessidade de monitoramento para evitar risco moral, como salientam **Grimsey & Lewis (2007)** e o **World Bank (2020)**.

Os métodos de agrupamento confirmaram a ausência de segmentações latentes claras: o *k*-means com  $K=3$  formou dois micro-clusters (21 e 4 observações) e um mega-cluster de 470 registros; o DBSCAN classificou a maior parte dos pontos como ruído. Tal padrão reproduz achados e indica que diferenças de eficiência decorrem de atributos específicos, não de estruturas latentes bem definidas.

A projeção **ARIMA** estimou média de eficiência-emprego de 3,93 para 2026-2030, com intervalo de 95 % entre -0,81 e 8,66. A amplitude do intervalo confirma a necessidade de re-estimar previsões à luz de novos dados, conforme defendem **Wooldridge (2012, 2020)** e o **World Bank (2017)**.

A pergunta-guia foi respondida com evidências robustas: eixos de **EDUCAÇÃO** e **CULTURA** continuam a apresentar melhor razão postos-de-trabalho/real investido, enquanto projetos em **INFRAESTRUTURA** e **SEGURANÇA PÚBLICA** mantém elasticidade maior sobre renda — compatível com a dualidade emprego-crescimento descrita por **Rodrik (2004)**. A supremacia de *valor\_global* como preditor, porém, sugere que ganhos setoriais podem ser neutralizados por ineficiências de escala ou falhas de execução, reforçando o papel de instrumentos preditivos para apontar anomalias e subsidiar a alocação eficiente de recursos.

Em síntese, todos os objetivos específicos foram cumpridos: (i) base consolidada e limpa; (ii) aplicação de técnicas avançadas de *Machine Learning* e econometria clássica; (iii) comparação sistemática de métricas e testes causais — incluindo *difference-in-differences* (**Wooldridge, 2012**) e ensaio preliminar de fronteira estocástica. Os resultados demonstram que, mesmo em ambientes de dados imperfeitos, abordagens de *Machine Learning* podem elevar a transparência e a eficiência alocativa das políticas públicas.

## 5.2 PRINCIPAIS CONCLUSÕES

A heterogeneidade entre os eixos de investimento do Amapá mostrou-se ainda mais pronunciada quando as métricas foram recalculadas com a base revisada. Eixos centrados em capital humano – sobretudo **EDUCAÇÃO** e **CULTURA** – voltaram a apresentar as maiores medianas de *eficiência\_emprego* e *eficiência renda*, enquanto iniciativas intensivas em ativos fixos, como **INFRAESTRUTURA** e **SEGURANÇA PÚBLICA**, mantiveram elasticidade superior sobre renda, reproduzindo a dualidade emprego–crescimento discutida por **Rodrik (2004)**. A regressão *feols* com efeitos fixos por eixo confirmou que, após 2020, projetos maiores (*valor\_global*) vêm obtendo ganhos marginais decrescentes de eficiência, sinalizando risco de sobre-escala e a necessidade de controles de governança já advertidos por **Grimsey & Lewis (2007)**.

Os testes de diferença de médias corroboraram a relevância do enquadramento institucional. Projetos classificados como **estratégicos** exibiram, em média, eficiência-emprego significativamente superior ( $p \approx 0,045$ ), ao passo que projetos **terceirizados** mostraram desempenho francamente inferior ( $p < 0,001$ ). Esses achados sustentam parcialmente a hipótese H2, indicando ganhos quando a formulação estratégica permanece sob controle estatal e perdas quando a execução é transferida sem salvaguardas robustas, em linha com as recomendações do **World Bank (2020)**.

No campo preditivo, o **XGBoost** (100 rounds, **Chen & Guestrin (2016)**) superou ligeiramente o **Random Forest** (RF) quanto ao coeficiente de determinação, mas ambos alcançaram R2R2 próximos de 0,96, com erros absolutos médios abaixo de 3 pontos. O **Gradient Boosting** ficou aquém, revelando perda de poder explicativo em bases sujeitas a multicolinearidade, enquanto a **rede neural** manteve instabilidade e variância elevada, problema já relatado por **Hofman, Lemos & Rodrigues (2021)**. A convergência dos dois melhores modelos confirma a robustez dos *ensembles* de árvores descrita por **Breiman (2001)** e formalizada por **Hastie, Tibshirani & Friedman (2009)**.

A análise de importância das variáveis trouxe uma mudança estrutural: *valor\_global* concentrou quase todo o ganho de informação nos modelos de árvore, relegando *impacto\_por\_projeto* a segundo plano e sugerindo que o volume financeiro, mais do que a qualidade intrínseca do projeto, domina a explicação da eficiência atual. Esse

resultado reforça a necessidade de critérios de custo-efetividade no planejamento, pois a simples alocação de recursos elevados não garante retornos proporcionais – uma advertência coerente com **Rodrik (2004)**.

Os algoritmos de agrupamento confirmaram esses padrões. O *k*-means identificou dois **micro-clusters** (21 e 4 projetos) de alta eficiência e um **mega-cluster** (470 observações) de eficiência mediana, enquanto o **DBSCAN** destacou alguns outliers de grande vulto financeiro com baixo retorno socioeconômico, repetindo configurações empíricas encontradas em cenários urbanos. Essas anomalias fornecem alvos imediatos para auditoria e realocação de recursos, sobretudo em projetos de alto custo unitário que não convertem investimento em impacto.

A projeção **ARIMA** apontou manutenção da eficiência-emprego média em torno de 3,9 para 2026-2030, mas com intervalo de confiança amplo (-0,8 a 8,7), evidenciando sensibilidade a choques macroeconômicos e falhas de registro, conforme enfatizam **Wooldridge (2012, 2020)**. A incerteza recomenda a atualização periódica das previsões e o reforço dos sistemas de informação.

Em síntese, a combinação de modelos preditivos, testes inferenciais e técnicas de clusterização indica que (i) a **priorização de eixos de alto impacto por projeto**, (ii) a  **contenção de escalas orçamentárias ineficientes**, (iii) a **proteção das características estratégicas** na execução e (iv) a **revisão de outliers de baixo retorno** podem elevar substancialmente o valor social do gasto público no Amapá. As evidências reforçam que instrumentos de *Machine Learning* – quando associados a análises econômicas clássicas – são capazes de ampliar a transparência, orientar a alocação de recursos e mitigar assimetrias de informação na gestão estadual.

### 5.3 LIMITAÇÕES DO ESTUDO

À semelhança dos alertas de Coelli et al. (2005) e Wooldridge (2012), os resultados aqui apresentados são condicionados pela qualidade da base administrativa estadual. A elevada proporção de registros com *valor\_global* = 0 ou ausências críticas restringiu a amostra útil e comprometeu a tentativa de estimar Fronteiras Estocásticas: o modelo SFA não convergiu e não produziu coeficientes interpretáveis, cenário também observado por Kumbhakar, Wang e Horncastle (2015)

em contextos com alta heterogeneidade. Ademais, variáveis qualitativas relevantes – como indicadores de inovação, benefícios ambientais ou percepção social – permanecem ausentes, limitando a abrangência dos modelos preditivos. Outro ponto reside na interpretabilidade. Mesmo com a extração de feature importance, ensembles profundos e redes neurais mantêm, em parte, comportamento de caixa-preta (Molnar 2022). Para políticas públicas que exigem prestação de contas, tal opacidade pode reduzir legitimidade, conforme discutem Athey (2018) e Grimsey e Lewis (2007). Por fim, embora a validação cruzada repetida mitigue sobreajuste, a inexistência de um conjunto externo ou de séries temporais longas impede avaliar a estabilidade dos parâmetros frente a choques macroeconômicos, seguindo a precaução de Greene (2012).

## 5.4 SUGESTÕES PARA PESQUISAS FUTURAS

Estudos subsequentes devem priorizar a expansão geográfica para comparar unidades federativas e capturar heterogeneidade regional, em linha com Porter (1990) e World Bank (2020). A incorporação de variáveis qualitativas – percepção dos beneficiários, métricas ambientais ou grau de inovação – pode refinar a modelagem, permitindo o uso de abordagens de Explainable AI (SHAP, LIME) recomendadas por Molnar (2022) para tornar transparentes os motores de decisão algorítmica.

Adicionalmente, sugere-se aprofundar métodos causais: Diferenças-em-Diferenças com delineamentos quase-experimentais mais claros, Controle Sintético ou Propensity Score Matching, conforme Wooldridge (2020) e Athey (2018), a fim de isolar o efeito do investimento de choques externos. Modelos de séries temporais estruturais e hierárquicos podem capturar dinâmicas de longo prazo e choques exógenos, enquanto versões robustas do SFA, alimentadas por variáveis físicas (quilômetros pavimentados, unidades escolares construídas), ajudarão a estimar fronteiras técnicas mais realistas. Por fim, recomenda-se instituir um repositório contínuo de dados que automatize ETL (extract-transform-load) e publique painéis preditivos em ciclos curtos, fomentando a tomada de decisão baseada em evidências, como defendem Breiman (2001) e Rodrik (2004).

## 5.5 RECOMENDAÇÕES PARA POLÍTICAS PÚBLICAS BASEADAS EM MACHINE LEARNING

A nova rodada de experimentos confirma a superioridade dos *ensembles* de árvores, mas com nuances importantes. O **XGBoost** apresentou o melhor equilíbrio entre viés e variância, superando ligeiramente o **Random Forest** (RF) no coeficiente de determinação, enquanto o **Gradient Boosting** (GBM) perdeu desempenho e estabilidade. Dessa forma, recomenda-se que a hierarquização de projetos utilize um **comitê XGBoost-RF** treinado em *resampling* estratificado: o acordo entre os dois modelos provê robustez, e a divergência indica casos que merecem análise qualitativa adicional, conforme a estratégia de *triage* sugerida por **Breiman (2001)** e formalizada em **Hastie, Tibshirani & Friedman (2009)**.

A importância de variáveis revelou forte concentração em **valor\_global**, seguido, a boa distância, por *eficiência\_renda* e *impacto\_por\_projeto*. Esse resultado implica que o volume financeiro é hoje o principal preditor — sinal de que parte da eficiência mensurada decorre de economias de escala mais do que de desenho do projeto. Assim, a política de investimento deve **condicionar a liberação de recursos adicionais ao incremento proporcional do impacto marginal**, evitando retornos decrescentes já apontados por **Rodrik (2004)**. Eixos cujo *impacto\_por\_projeto* se mantém elevado mesmo com orçamentos menores — notadamente EDUCAÇÃO e CULTURA — devem receber prioridade.

Os **testes t e ANOVA** mostraram que projetos **estratégicos** geram eficiência-emprego significativamente maior, enquanto iniciativas **terceirizadas** revelam ineficiência estatisticamente comprovada. Em linha com **Grimsey & Lewis (2007)** e as boas-práticas do **World Bank (2020)**, recomenda-se manter a governança dos projetos estratégicos no núcleo público e submeter contratos terceirizados a cláusulas de desempenho vinculadas às métricas de eficiência.

Para acelerar ciclos de realimentação, o Estado deve implantar um **repositório transacional** que receba dados físico-financeiros quase em tempo real, gatilhando *re-train* mensal dos modelos. A experiência internacional mostra que a acurácia de previsões degrada-se rapidamente sem dados frescos; logo, sugere-se um SLA de 72 h para postagem de medições de execução, conforme boas práticas de *data pipeline* recomendadas pelo **World Bank (2017)**.

A detecção de **outliers** — identificados tanto pelo *k-means* (microsclusteres de alta eficiência) quanto pelo **DBSCAN** (obras de grande custo e baixo retorno) — deve alimentar um **módulo de auditoria preditiva**: sempre que um projeto ultrapassar, por exemplo, dois desvios-padrão negativos em eficiência, a Controladoria instaura acompanhamento especial, modelando o fluxo de exceção sugerido por **Witte et al. (2020)**.

Apesar da elevada performance, os modelos de árvore são opacos. Portanto, cada recomendação deve vir acompanhada de **explicações locais via SHAP** — metodologia que quantifica contribuição de cada variável na predição individual e já foi aplicada com sucesso em domínios governamentais por **Hofman, Lemos & Rodrigues (2021)**. A publicação de dashboards abertos com essas explicações fortalece a prestação de contas e reduz assimetria de informação entre Governo e sociedade.

A sustentação desse ecossistema requer **capacitação permanente**. Sugere-se criar dois núcleos — um na **SEINF** (foco em infraestrutura de dados) e outro na **SEPLAN** (foco em modelagem) — responsáveis por *versionamento de código*, auditoria de viés e gestão de modelo, conforme a matriz de governança analítica defendida por **Wooldridge (2020)**.

A robustez preditiva é limitada pela qualidade dos registros. É, portanto, urgente padronizar campos, suprimir valores ausentes críticos e ampliar o leque de variáveis para incluir indicadores socioambientais e de bem-estar. Com dados mais ricos, técnicas de **fronteira estocástica** e **Difference-in-Differences** — já preliminarmente testadas com **fixest** — podem complementar o diagnóstico, aproximando-o de relações causais e não apenas correlacionais.

Por fim, as projeções **ARIMA** revelam incerteza ampla para a eficiência média até 2030. Recomenda-se incorporar **análises de sensibilidade e teste de estresse** aos instrumentos de Planejamento Plurianual, mitigando o risco de choques macroeconômicos. A articulação entre métricas preditivas, mecanismos de explicação, governança de dados e gestão de risco cria um ciclo virtuoso de decisão baseada em evidências, elevando transparência e impacto socioeconômico dos investimentos no Amapá.

R

# REFERÊNCIAS

# REFERÊNCIAS

J

## REFERÊNCIAS

**AIGNER, D. J.; LOVELL, C. A. K.; SCHMIDT, P.** Formulation and Estimation of Stochastic Frontier Production Function Models. *Journal of Econometrics*, v. 6, n. 1, p. 21–37, 1977.

**ALMEIDA, E. S.** *Economia Regional: Teorias e Métodos de Análise*. São Paulo: Atlas, 2019. Texto que debate as diferentes abordagens teóricas do desenvolvimento regional, relacionando-as aos desafios de regiões periféricas na promoção de investimentos públicos eficientes.

**ATHEY, S.** The Impact of Machine Learning on Economics. In: *AEA Papers and Proceedings*. Nashville: American Economic Association, p. 1–6, 2018.

Destaca o uso de algoritmos de ML em análises de políticas públicas, realçando como a combinação de grandes bases de dados e técnicas avançadas aprimora a capacidade preditiva e a detecção de padrões.

**BANKER, R. D.; NATARAJAN, R.** Evaluating Contextual Variables Affecting Productivity Using Data Envelopment Analysis. *Operations Research*, v. 56, n. 1, p. 48–58, 2008.

**BERGSTRA, J.; BENGIO, Y.** Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, v. 13, p. 281–305, 2012.

**BNDES (BANCO NACIONAL DE DESENVOLVIMENTO ECONÔMICO E SOCIAL).** Relatório de Avaliação de Projetos de Infraestrutura. Rio de Janeiro, 2020. Disponível em: <https://www.bnDES.gov.br/>. Acesso em: 5 jan. 2022.

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

**BRASIL.** *Portal de Dados Abertos: Transparência e Governança*. 2021. Disponível em: <https://dados.gov.br/>. Acesso em: 20 ago. 2021. Compilação de bases de dados governamentais brasileiras, cobrindo distintos setores e tipos de investimentos públicos, inclusive obras de infraestrutura.

**BRITO, V. C.; SILVEIRA, L. F.** Infraestrutura e Crescimento Econômico: uma Revisão das Evidências. *Economia Aplicada*, v. 25, n. 2, p. 201–223, 2021.

**CALDERÓN, C.; SERVÉN, L.** Infrastructure, Growth, and Inequality: An Overview. *World Bank Policy Research Working Paper*, n. 7034, 2014. Destaca a relevância de investimentos públicos em infraestrutura para reduzir desigualdades regionais e alavancar crescimento em países emergentes.

**CHARNES, A.; COOPER, W. W.; RHODES, E.** Measuring the Efficiency of Decision-Making Units. *European Journal of Operational Research*, v. 2, n. 6, p. 429–444, 1978.

**CHEN, T.; GUESTIN, C.** XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. São Francisco: ACM, 2016, p. 785–794.

**CHI, G. et al.** Micro-estimates of wealth for all low- and middle-income countries. *ArXiv pre-print* 2104.07761, 2021.

**CHOLLET, F.** *Deep Learning with Python*. Shelter Island: Manning Publications, 2017. Fundamenta as arquiteturas de redes neurais profundas, enfatizando potencial de modelagem não linear, mas igualmente o problema de interpretabilidade em aplicações governamentais.

**CHUI, M. et al.** Notes From the AI Frontier: Applying AI for Social Good. *McKinsey Global Institute Discussion Paper*, p. 1–24, 2018.

**COELLI, T. J. et al.** *An Introduction to Efficiency and Productivity Analysis*. 2. ed. Nova York: Springer, 2005.

**COSTA, M.; MENDES, S. J.; SOUZA, V. C.** Avaliação de Políticas Sociais com Machine Learning: Uma Aplicação na Área de Saúde. *Revista de Políticas Públicas*, v. 14, n. 3, p. 123–141, 2019.

**CRESWELL, J. W.** *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. 4. ed. Thousand Oaks: SAGE, 2014.

**DELGADO, M.; LEMOS, R.** Explainable AI in public-sector decision-making: challenges and opportunities. *Government Information Quarterly*, v. 39, n. 4, 101720, 2023

**ESTER, M. et al.** A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*. Portland: AAAI Press, 1996, p. 226–231.

**FÄRE, R.; GROSSKOPF, S.; LOVELL, C. A. K.** *Production Frontiers*. Cambridge: Cambridge University Press, 1994.

**FRIEDMAN, J. H.** Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, v. 29, n. 5, p. 1189–1232, 2001.

**FRIEDMAN, J.; POPOESCU, B. E.** Predictive Learning via Rule Ensembles. *The Annals of Applied Statistics*, v. 2, n. 3, p. 916–954, 2008.

**GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A.** *Deep Learning*. Cambridge: MIT Press, 2016.

**GRAMLICH, E. M.** Infrastructure Investment: A Review Essay. *Journal of Economic Literature*, v. 32, n. 3, p. 1176–1196, 1994.

**GREENE, W. H.** *Econometric Analysis*. 7. ed. Upper Saddle River: Prentice Hall, 2012.

**GRIMSEY, D.; LEWIS, M. K.** *Public Private Partnerships: The Worldwide Revolution in Infrastructure Provision and Project Finance*. Cheltenham: Edward Elgar, 2007.

**GRILICHES, Z.** The Search for R&D Spillovers. *Scandinavian Journal of Economics*, v. 94, p. 29–47, 1998.

**GUJARATI, D.; PORTER, D. C.** *Basic Econometrics*. 5. ed. New York: McGraw-Hill, 2011.

**HAIR JR., J. F. et al.** *Multivariate Data Analysis*. 8. ed. Andover: Cengage, 2019.

**HARTIGAN, J. A.; WONG, M. A.** Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, v. 28, n. 1, p. 100–108, 1979.

**HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J.** *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2. ed. New York: Springer, 2009.

**HAYKIN, S.** *Neural Networks and Learning Machines*. 3. ed. Upper Saddle River: Prentice Hall, 2009.

**HIRSCHMAN, A. O.** *The Strategy of Economic Development*. New Haven: Yale University Press, 1958.

**HOFMAN, J. M.; LEMOS, R.; RODRIGUES, M.** Transparency, Interpretability, and Reproducibility in AI for Policy. *Communications of the ACM*, v. 64, n. 10, p. 34–37, 2021.

**IPEA – INSTITUTO DE PESQUISA ECONÔMICA APLICADA.** Infraestrutura e Desenvolvimento Regional: Evidências Recentes. *Relatório de Pesquisa*, Brasília: IPEA, 2020.

**JAMES, G. et al.** *An Introduction to Statistical Learning: with Applications in R*. 2. ed. Nova York: Springer, 2021.

**KOENGKAN, M.; OLIVEIRA, M. T.; TEIXEIRA, R. V. S.** Measuring the economic-efficiency performance in Latin America and the Caribbean: an empirical evidence from stochastic frontier and data-envelopment analysis. *International Economics*, v. 169, p. 43-54, 2023.

**KUHN, M.; JOHNSON, K.** *Applied Predictive Modeling*. 2. ed. Nova York: Springer, 2013.

**KUMBHAKAR, S. C.; LOVELL, C. A. K.** *Stochastic Frontier Analysis*. Cambridge: Cambridge University Press, 2003.

**KUMBHAKAR, S. C.; WANG, H.-J.; HORNCastle, A. P.** *A Practitioner's Guide to Stochastic Frontier Analysis Using Stata*. Cambridge: Cambridge University Press, 2015.

**LIMA, L. R.; MATOS, H. R.** Políticas Públicas de Infraestrutura e Desigualdades Regionais: Evidências no Contexto Brasileiro. *Estudos Regionais e Urbanos*, v. 35, n. 2, p. 101–115, 2019.

**LITTLE, R. J. A.; RUBIN, D. B.** *Statistical Analysis with Missing Data*. 3. ed. Nova York: Wiley, 2019.

**LUNDBERG, S. M.; LEE, S.-I.** A Unified Approach to Interpreting Model Predictions. In: *Advances in Neural Information Processing Systems*, 2017, p. 4765–4774.

**MAROCO, J.** *Análise Estatística com Utilização do SPSS*. 3. ed. Lisboa: Edições Sílabo, 2010.

**MOLNAR, C.** *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Morrisville: Lulu Press, 2019.

**MOLNAR, C.** *Interpretable Machine Learning*. 2. ed. Disponível em: <https://christophm.github.io/interpretable-ml-book/>. Acesso em: 10 jan. 2022, 2022.

**MULLAINATHAN, S.; SPIESS, J.** Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, v. 31, n. 2, p. 87–106, 2017.

**MYRDAL, G.** *Economic Theory and Underdeveloped Regions*. London: Methuen, 1957.

**OECD**. *Scaling-up infrastructure investment to strengthen sustainable development in Brazil*. **OECD Economics Department Working Paper** n.º 1787, 1790.

**PERROUX, F.** L'Économie du XXe Siècle. *Revue Économique*, v. 1, n. 2, p. 140–166, 1950.

**PI, Y.** Machine learning in governments: benefits, challenges and future directions. **JeDEM – eJournal of eDemocracy and Open Government**, v. 13, n. 1, p. 203-219, 2021.

**PORTER, M.** *The Competitive Advantage of Nations*. New York: Free Press, 1990.

**RATLEDGE, N. et al.** Using satellite imagery and machine learning to estimate the livelihood impact of electricity access. **ArXiv pre-print** 2109.02890, 2021.

**RAO, H.; BARTON, A.** Assessing data-quality issues in machine-learning for public-sector applications: a systematic review. **Government Information Quarterly**, v. 40, n. 1, 101730, 2022.

**RODRIGUES, T. B.; SILVA, A. F.** Modelos Multi-Critério para Avaliar Eficiência na Construção Civil. *Revista Brasileira de Planejamento e Desenvolvimento*, v. 9, n. 2, p. 88–97, 2020.

**RODRIK, D.** *In Search of Prosperity: Analytic Narratives on Economic Growth*. Princeton: Princeton University Press, 2004.

**SOUZA, C.; DUTRA, L. E.** Desafios da Avaliação de Eficiência em Grandes Contratos de Obras. In: *Anais do 14º Congresso Brasileiro de Administração Pública*, 2017, p. 77–86.

**TAN, P.-N.; STEINBACH, M.; KUMAR, V.** *Introduction to Data Mining*. 2. ed. Harlow: Pearson, 2019.

**VARIAN, H. R.** Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, v. 28, n. 2, p. 3–28, 2014.

**VAZ, A. C.; HADDAD, E. A.; DOMINGUES, E. P.** Desafios para o Crescimento no Norte do Brasil: Infraestrutura, Capital Humano e Relações Fiscais. *Planejamento e Políticas Públicas*, v. 54, p. 1–29, 2020.

**WICKHAM, H.** *ggplot2: Elegant Graphics for Data Analysis*. 2. ed. Nova York: Springer, 2016.

**WITTE, J. et al.** Clustering and Outlier Detection in Smart City Investments. *Computers, Environment and Urban Systems*, v. 83, p. 101530, 2020.

**WOOLDRIDGE, J. M.** *Introductory Econometrics: A Modern Approach*. 6. ed. Boston: Cengage, 2020.

**WOOLDRIDGE, J. M.** *Econometric Analysis of Cross Section and Panel Data*. 2. ed. Cambridge: MIT Press, 2012.

**WORLD BANK.** *World Development Report: Trading for Development*. Washington, D.C.: World Bank, 2020.

**WORLD BANK.** *The Changing Wealth of Nations: Building a Sustainable Future*. Washington, D.C.: World Bank, 2017.



**idp**

A ESCOLHA QUE  
**TRANSFORMA**  
O SEU CONHECIMENTO