



Universidade Federal do Piauí
Centro de Ciências da Natureza
Programa de Pós-Graduação em Ciência da Computação

Modelagem Preditiva para o IDEB Municipal na Educação Básica Piauiense por Meio de Algoritmos de Aprendizado de Máquina e Aprendizado Profundo

Maria Eva Clemencia Fonseca de Castro Silva

Número de Ordem PPGCC: M001

Teresina-PI, Fevereiro de 2026

Maria Eva Clemencia Fonseca de Castro Silva

Modelagem Preditiva para o IDEB Municipal na Educação Básica Piauiense por Meio de Algoritmos de Aprendizado de Máquina e Aprendizado Profundo

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da UFPI (área de concentração: Inteligência Artificial), como parte dos requisitos necessários para a obtenção do Título de Mestre em Ciência da Computação.

Universidade Federal do Piauí – UFPI
Centro de Ciências da Natureza
Programa de Pós-Graduação em Ciência da Computação

Orientador: Prof. Dr. Ivan Saraiva Silva
Coorientador: Prof. Dr. Vinícius Ponte Machado

Teresina-PI
Fevereiro de 2026

FICHA CATALOGRÁFICA
Universidade Federal do Piauí
Sistema de Bibliotecas UFPI - SIBi/UFPI
Biblioteca Setorial do CCN

S586m Silva, Maria Eva Clemencia Fonseca de Castro.
 Modelagem preditiva para o IDEB Municipal na Educação
 Básica Piauiense por meio de algoritmos de aprendizado de
 máquina e aprensizado profundo / Maria Eva Clemencia
 Fonseca de Castro Silva. -- 2026.
 127 f. : il.

 Dissertação (Mestrado) - Universidade Federal do Piauí.
 Centro de Ciências da Natureza. Programa de Pós-Graduação
 em Ciências da Computação, Teresina, 2026.
 “Orientador: Prof. Dr. Ivan Saraiva Silva, Coorientador:
 Prof. Dr. Vinicius Ponte Machado”.

 1. Aprendizado de Máquina. 2. Rede Neural Convolucional.
 3. Índice de Desenvolvimento da Educação Básica - IDEB. 4.
 Simulação Contrafactual. I. Silva, Ivan Saraiva. II. Machado,
 Vinicius Ponte. III. Título.

CDD 005.43

Bibliotecária: Caryne Maria da Silva Gomes - CRB3/1461

Maria Eva Clemencia Fonseca de Castro Silva

Modelagem Preditiva para o IDEB Municipal na Educação Básica Piauiense por Meio de Algoritmos de Aprendizado de Máquina e Aprendizado Profundo

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da UFPI (área de concentração: Inteligência Artificial), como parte dos requisitos necessários para a obtenção do Título de Mestre em Ciência da Computação.

Trabalho aprovado. Teresina-PI, 09 de fevereiro de 2026:

Prof. Dr. Ivan Saraiva Silva
Orientador

Prof. Dr. Vinícius Ponte Machado
Co-Orientador

Prof. Dr. Raimundo Santos Moura
(PPGCC - UFPI)
Avaliador Interno ao PPGCC

Prof^a. Dra. Luisa Xavier de Oliveira
(FUN-UFPI)
Avaliadora Externa ao PPGCC

Prof. Gustavo Girão Barreto da Silva
(IMD-UFRN)
Avaliador Externo à UFPI

Teresina-PI
Fevereiro de 2026

*Aos meus pais Pedro Fonseca de Castro e Maria de Jesus da Silva Santos,
por serem minha base, minha força e minha motivação em cada etapa desta jornada.*

Agradecimentos

Agradeço a Deus, fonte de sabedoria e inspiração, por me conceder força, serenidade e perseverança ao longo desta caminhada acadêmica. Por estar sempre comigo, iluminando meus passos, concedendo-me entendimento e renovando minhas forças diante dos desafios, sem jamais permitir que eu desistisse.

Aos meus pais, Pedro Fonseca de Castro e Maria de Jesus da Silva Santos, pelo amor incondicional, apoio e incentivo constante em todas as etapas da minha vida.

Aos meus irmãos e familiares, pela compreensão e pelas palavras de encorajamento nos momentos mais desafiadores.

Ao meu orientador, Prof. Dr. Ivan Saraiva Silva, e ao meu coorientador, Prof. Dr. Vinícius Ponte Machado, pela orientação firme, pelos ensinamentos compartilhados e pela confiança depositada em mim durante o desenvolvimento desta pesquisa.

Aos professores e colegas do Programa de Pós-Graduação em Ciência da Computação da Universidade Federal do Piauí (PPGCC/UFPI), pela troca de conhecimentos e pelo ambiente de aprendizado e colaboração.

Aos amigos que estiveram presentes ao longo desta jornada, oferecendo apoio, amizade e motivação, contribuindo para tornar este percurso mais leve e significativo.

Por fim, a todos que, de alguma forma, contribuíram para a concretização deste trabalho, registro minha sincera e profunda gratidão.

“Em Deus nós confiamos; todos os demais devem apresentar dados.”
(William Edwards Deming)

Resumo

As desigualdades no desempenho educacional entre municípios brasileiros permanecem amplas e persistentes, conforme evidenciado pelas variações do Índice de Desenvolvimento da Educação Básica (IDEB) nas diferentes etapas da educação básica. Apesar da crescente disponibilidade de informações sobre os sistemas de ensino, ainda há limitações na utilização integrada desses insumos para prever o desempenho municipal e subsidiar decisões estratégicas. Essa lacuna é particularmente relevante em estados como o Piauí, onde os municípios apresentam desempenho heterogêneo, com alguns registrando avanços consistentes e outros mantendo resultados inferiores à média estadual. Neste contexto, o presente trabalho tem como objetivo principal aplicar e avaliar modelos preditivos baseados em aprendizado de máquina (Machine Learning - ML) e aprendizado profundo (Deep Learning - DL) para prever o desempenho educacional dos municípios piauienses, representado pelo IDEB, nas três etapas da educação básica (anos iniciais, anos finais e ensino médio), integrando variáveis educacionais, socioeconômicas e financeiras. Para tanto, foram construídas três bases de dados, compreendendo o período de 2013 a 2023, considerando os anos ímpares. Foram testados e comparados dez algoritmos clássicos de regressão supervisionada, incluindo regressões penalizadas, modelos baseados em árvore, como Floresta Aleatória (Random Forest - RF) e Extreme Gradient Boosting (XGBoost) e métodos baseados em instâncias, além de sete arquiteturas de Redes Neurais Artificiais (Artificial Neural Networks - ANN) profundas, como Perceptron Multicamadas (Multi-layer Perceptron - MLP), Rede Neural Convolucional (Convolutional Neural Network - CNN), Memória de Longo e Curto Prazo (Long Short-Term Memory - LSTM) e híbridas. O desempenho dos modelos foi avaliado por meio de métricas de erro e coeficiente de determinação (R^2), calculadas nos conjuntos de treino, teste e validação cruzada com 5 *folds*. Os resultados indicaram que o modelo XGBoost apresentou o melhor desempenho global nas três etapas da educação básica, com R^2 de 0,5455 nos anos iniciais, 0,5205 nos anos finais e 0,4488 no ensino médio, segundo validação cruzada. Destacou-se para sua robustez nos anos iniciais e finais e sua relativa estabilidade no ensino médio, etapa notoriamente mais difícil de prever devido à maior variabilidade e evasão. A análise de importância das variáveis identificou determinantes centrais para o IDEB. Nos anos iniciais e finais, destacaram-se o PIB per capita, a taxa de distorção idade-série, o percentual de docentes sem curso superior, variáveis da agropecuária local e repasses do programa Criança Feliz. No ensino médio, assumiram relevância variáveis ligadas à infraestrutura escolar, gestão e esforço docente. Com base no modelo XGBoost, foi desenvolvido um sistema de simulação de cenários, que permite testar, de forma contrafactual, o impacto de alterações paramétricas nas variáveis explicativas sobre os valores previstos do IDEB municipal de 2023. Esse módulo opera com interface amigável e recursos de exportação, gráficos e visualizações personalizadas, representando uma contribuição concreta para o planejamento educacional baseado em evidências. Os resultados obtidos demonstram que é possível antecipar, com razoável grau de acurácia, o comportamento do IDEB em nível municipal. O sistema de simulação desenvolvido consolida-se como ferramenta de apoio à tomada de decisão, permitindo que gestores visualizem os efeitos hipotéticos de políticas públicas educacionais. Como perspectivas futuras, recomenda-se a inclusão de novas variáveis e a replicação da metodologia em outros contextos federativos.

Palavras-chaves: Machine learning. Deep learning. Educação básica. IDEB. Políticas públicas. Simulação contrafactual.

Abstract

Inequalities in educational performance among Brazilian municipalities remain wide and persistent, as evidenced by variations in the Basic Education Development Index (IDEB) across different stages of basic education. Despite the increasing availability of information on education systems, there are still limitations in the integrated use of these inputs to predict municipal performance and support strategic decisions. This gap is particularly relevant in states like Piauí, where municipalities exhibit heterogeneous performance, with some registering consistent progress and others maintaining results below the state average. In this context, the present work aims to apply and evaluate predictive models based on machine learning (ML) and deep learning (DL) to predict the educational performance of municipalities in Piauí, represented by the IDEB (Basic Education Development Index), in the three stages of basic education (early years, later years, and high school), integrating educational, socioeconomic, and financial variables. To this end, three databases were constructed, covering the period from 2013 to 2023, considering odd-numbered years. Ten classic supervised regression algorithms were tested and compared, including penalized regressions, tree-based models, such as Random Forest (RF) and Extreme Gradient Boosting (XGBoost), and instance-based methods, in addition to seven deep Artificial Neural Network (ANN) architectures, such as Multilayer Perceptron (MLP), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and hybrid architectures. The performance of the models was evaluated using error metrics and coefficient of determination (R^2), calculated on the training, test, and cross-validation sets with 5 folds. The results indicated that the XGBoost model presented the best overall performance in the three stages of basic education, with an R^2 of 0.5455 in the early years, 0.5205 in the later years, and 0.4488 in high school, according to cross-validation. It stood out for its robustness in the early and later years and its relative stability in high school, a stage notoriously more difficult to predict due to greater variability and dropout rates. The analysis of the importance of the variables identified central determinants for the IDEB (Basic Education Development Index). In the early and later years, the following stood out: GDP per capita, the age-grade distortion rate, the percentage of teachers without a higher education degree, variables from local agriculture, and transfers from the Criança Feliz (Happy Child) program. In high school, variables related to school infrastructure, management, and teacher effort became relevant. Based on the XGBoost model, a scenario simulation system was developed, which allows testing, in a counterfactual way, the impact of parametric changes in explanatory variables on the predicted values of the municipal IDEB (Basic Education Development Index) for 2023. This module operates with a user-friendly interface and export features, customized graphs and visualizations, representing a concrete contribution to evidence-based educational planning. The results obtained demonstrate that it is possible to anticipate, with a reasonable degree of accuracy, the behavior of the IDEB at the municipal level. The developed simulation system consolidates itself as a tool to support decision-making, allowing managers to visualize the hypothetical effects of educational public policies. As future perspectives, the inclusion of new variables and the replication of the methodology in other federative contexts is recommended.

Keywords: Machine learning. Deep learning. Basic education. IDEB. Public policies. Simulation.

Lista de ilustrações

Figura 1 – Classificação dos principais tipos de aprendizado de máquina	20
Figura 2 – Alguns modelos supervisionados de ML	20
Figura 3 – Relação entre IA, ML e DL	23
Figura 4 – Arquiteturas fundamentais de DL em ordem crescente de complexidade . . .	24
Figura 5 – Relação entre qualidade e desempenho educacional	28
Figura 6 – Vantagens da simulação contrafactual para o planejamento educacional	43
Figura 7 – Etapas metodológicas da pesquisa em modelagem preditiva do IDEB .	45
Figura 8 – Resumo do processo de seleção de variáveis e configuração final por etapa de ensino	54
Figura 9 – Evolução histórica do IDEB médio piauiense nos anos iniciais do ensino fundamental (2013–2023)	58
Figura 10 – Evolução histórica do IDEB médio piauiense nos anos finais do ensino fundamental (2013–2023)	58
Figura 11 – Evolução histórica do IDEB médio piauiense no ensino médio (2017–2023)	59
Figura 12 – Mapa de correlação entre variáveis selecionadas e o IDEB médio piauiense dos anos iniciais do ensino fundamental	60
Figura 13 – Mapa de correlação entre variáveis selecionadas e o IDEB médio piauiense dos anos finais do ensino fundamental	61
Figura 14 – Mapa de correlação entre variáveis selecionadas e o IDEB médio piauiense do ensino médio	62
Figura 15 – Comparação dos melhores modelos de ML (XGBoost) e DL (MLP nos anos iniciais e finais; CNN 1D no ensino médio), segundo o R^2 , por etapa da educação básica	68
Figura 16 – Comparação dos melhores modelos de ML (XGBoost) e DL (MLP nos anos iniciais e finais; CNN 1D no ensino médio), segundo o MAE, por etapa da educação básica	69
Figura 17 – Comparação dos melhores modelos de ML (XGBoost) e DL (MLP nos anos iniciais e finais; CNN 1D no ensino médio), segundo o RMSE, por etapa da educação básica	69
Figura 18 – Fluxo metodológico da previsão algorítmica à simulação contrafactual	74
Figura 19 – Interface interativa do módulo de simulação contrafactual implementada em Python (Colab).	75

Lista de quadros

Quadro 1 – Critérios metodológicos adotados para o estado do conhecimento .	35
Quadro 2 – Critérios analíticos para avaliação de estudos empíricos em modelagem preditiva da educação	37
Quadro 3 – Estudos com modelos de regressão aplicados à predição do desempenho educacional	40
Quadro 4 – Estudos com modelos de classificação aplicados à predição do desempenho educacional	41
Quadro 5 – Principais hiperparâmetros dos algoritmos de ML e DL aplicados neste estudo	49
Quadro 6 – Etapas da modelagem preditiva e estratégias de seleção de variáveis nos modelos (continua)	53
Quadro 7 – Seleção dos municípios segundo o desempenho no IDEB de 2023 .	81

Lista de tabelas

Tabela 1 – IDEB observado nos anos iniciais do ensino fundamental por região (2013–2023)	30
Tabela 2 – Desempenho dos melhores modelos de ML para os anos iniciais do ensino fundamental com as variáveis selecionadas	63
Tabela 3 – Desempenho dos modelos de DL para os anos iniciais do ensino fundamental com as variáveis selecionadas	64
Tabela 4 – Desempenho dos melhores modelos de ML para os anos finais do ensino fundamental com as variáveis selecionadas	65
Tabela 5 – Desempenho dos modelos de DL para os anos finais do ensino fundamental com as variáveis selecionadas	66
Tabela 6 – Desempenho dos melhores modelos de ML para o ensino médio com as variáveis selecionadas	66
Tabela 7 – Desempenho dos modelos de DL para o ensino médio com as variáveis selecionadas	67
Tabela 8 – Impacto marginal das variáveis nos anos iniciais para o município de Caracol (IDEB real 2023 = 3,5; IDEB previsto sem alteração = 3,8798)	83
Tabela 9 – Impacto marginal das variáveis nos anos iniciais para o município de Francisco Macedo (IDEB real 2023 = 5,3; IDEB previsto sem alteração = 5,8084)	85
Tabela 10 – Impacto marginal das variáveis nos anos iniciais para o município de Domingos Mourão (IDEB real 2023 = 8,4; IDEB previsto sem alteração = 6,6108)	86
Tabela 11 – Impacto marginal das variáveis nos anos finais para o município de Barreiras do Piauí (IDEB real 2023 = 2,6; IDEB previsto sem alteração = 3,6210)	87
Tabela 12 – Impacto marginal das variáveis nos anos finais para o município de São João da Serra (IDEB real 2023 = 4,7; IDEB previsto sem alteração = 4,9859)	88
Tabela 13 – Impacto marginal das variáveis nos anos finais para o município de Buriti dos Montes (IDEB real 2023 = 6,5; IDEB previsto sem alteração = 5,2890)	89
Tabela 14 – Impacto marginal das variáveis de infraestrutura escolar no ensino médio para o município de Capitão de Campos (IDEB real 2023 = 3,1; IDEB previsto sem alteração = 4,4100)	90

Tabela 15 – Impacto marginal das variáveis de características docentes e pedagógicas no ensino médio para o município de Capitão de Campos (IDEB real 2023 = 3,1; IDEB previsto sem alteração = 4,4100)	91
Tabela 16 – Impacto marginal das variáveis de rendimento e jornada escolar no ensino médio para o município de Capitão de Campos (IDEB real 2023 = 3,1; IDEB previsto sem alteração = 4,4100)	92
Tabela 17 – Impacto marginal das variáveis de fatores socioeconômicos e produtivos no ensino médio para o município de Capitão de Campos (IDEB real 2023 = 3,1; IDEB previsto sem alteração = 4,4100)	93
Tabela 18 – Impacto marginal das variáveis de infraestrutura escolar no ensino médio para o município de Simplício Mendes (IDEB real 2023 = 4,1; IDEB previsto sem alteração = 4,5053)	94
Tabela 19 – Impacto marginal das variáveis de características docentes e pedagógicas no ensino médio para o município de Simplício Mendes (IDEB real 2023 = 4,1; IDEB previsto sem alteração = 4,5053)	94
Tabela 20 – Impacto marginal das variáveis de rendimento e jornada escolar no ensino médio para o município de Simplício Mendes (IDEB real 2023 = 4,1; IDEB previsto sem alteração = 4,5053)	95
Tabela 21 – Impacto marginal das variáveis de fatores socioeconômicos e produtivos no ensino médio para o município de Simplício Mendes (IDEB real 2023 = 4,1; IDEB previsto sem alteração = 4,5053)	96
Tabela 22 – Impacto marginal das variáveis de infraestrutura escolar no ensino médio para o município de Cocal dos Alves (IDEB real 2023 = 7,0; IDEB previsto sem alteração = 4,9919)	97
Tabela 23 – Impacto marginal das variáveis de características docentes e pedagógicas no ensino médio para o município de Cocal dos Alves (IDEB real 2023 = 7,0; IDEB previsto sem alteração = 4,9919)	98
Tabela 24 – Impacto marginal das variáveis de rendimento e jornada escolar no ensino médio para o município de Cocal dos Alves (IDEB real 2023 = 7,0; IDEB previsto sem alteração = 4,9919)	98
Tabela 25 – Impacto marginal das variáveis de fatores socioeconômicos e produtivos no ensino médio para o município de Cocal dos Alves (IDEB real 2023 = 7,0; IDEB previsto sem alteração = 4,9919)	99

Lista de abreviaturas e siglas

ANN	<i>Artificial Neural Network</i>
BLEU	<i>Bilingual Evaluation Understudy</i>
CAQ	Custo Aluno Qualidade
CAQi	Custo Aluno Qualidade Inicial
CNN	<i>Convolutional Neural Network</i>
DL	<i>Deep Learning</i>
DT	<i>Decision Tree</i>
FID	<i>Frechet Inception Distance</i>
FPE	Fundo de Participação dos Estados
FPM	Fundo de Participação dos Municípios
FUNDEB	Fundo de Manutenção e Desenvolvimento da Educação Básica e de Valorização dos Profissionais da Educação
FUNDEF	Fundo de Manutenção e Desenvolvimento do Ensino Fundamental e de Valorização do Magistério
GBM	<i>Gradient Boosting Machine</i>
GPT	<i>Generative Pre-trained Transformer</i>
IA	Inteligência Artificial
IBGE	Instituto Brasileiro de Geografia e Estatística
ICMS	Imposto sobre Circulação de Mercadorias e Serviços
IDEB	Índice de Desenvolvimento da Educação Básica
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
INSE	Índice de Nível Socioeconômico
IPTU	Imposto Predial e Territorial Urbano
IRRF	Imposto de Renda Retido na Fonte

ISS	Imposto Sobre Serviços
ITS	<i>Intelligent Tutoring Systems</i>
KNN	<i>K-Nearest Neighbors</i>
LLM	<i>Large Language Models</i>
LDB	Lei de Diretrizes e Bases da Educação Nacional
LSTM	<i>Long Short-Term Memory</i>
MAE	<i>Mean Absolute Error</i>
ML	<i>Machine Learning</i>
MLP	<i>Multilayer Perceptron</i>
OCDE	Organização para a Cooperação e Desenvolvimento Econômico
PCA	<i>Principal Component Analysis</i>
PISA	<i>Programme for International Student Assessment</i>
PLN	Processamento de Linguagem Natural
PNAE	Programa Nacional de Alimentação Escolar
PNATE	Programa Nacional de Apoio ao Transporte Escolar
PNE	Plano Nacional de Educação
PQR	Padrão de Qualidade de Referência
R^2	Coefficiente de Determinação
RF	<i>Random Forest</i>
RMSE	<i>Root Mean Squared Error</i>
RNN	<i>Recurrent Neural Network</i>
SAEB	Sistema de Avaliação da Educação Básica
SHAP	<i>SHapley Additive exPlanations</i>
SIDRA	Sistema IBGE de Recuperação Automática
SIOPE	Sistema de Informações sobre Orçamentos Públicos em Educação
SimCAQ	Simulador de Custo Aluno Qualidade

UFPR	Universidade Federal do Paraná
UFG	Universidade Federal de Goiás
VAB	Valor Adicionado Bruto
VAAR	Valor Aluno Ano Resultado
VAAT	Valor Aluno Ano Total
VIF	<i>Variance Inflation Factor</i>
XGBoost	<i>Extreme Gradient Boosting</i>

Sumário

1	INTRODUÇÃO	12
1.1	Contextualização e Justificativa	13
1.2	Objetivos	16
1.2.1	Objetivo Geral	16
1.2.2	Objetivos Específicos	16
1.3	Hipóteses	16
1.4	Questões de Pesquisa	17
1.5	Contribuições Esperadas	17
1.6	Estrutura da Proposta de Dissertação	17
2	FUNDAMENTAÇÃO TEÓRICA	19
2.1	Aprendizado de Máquina (Machine Learning)	19
2.2	Aprendizado Profundo (Deep Learning)	23
2.3	Inteligência Artificial Aplicada à Educação	26
2.4	Qualidade e Desempenho Educacional	27
2.5	Financiamento Público na Educação Brasileira	31
3	REVISÃO DA LITERATURA	34
3.1	Estado do Conhecimento	34
3.1.1	Estratégia de Busca e Critérios de Seleção	34
3.1.2	Trabalhos Relacionados e Lacunas Identificadas	36
4	METODOLOGIA	44
4.1	Fontes dos Dados	45
4.2	Coleta e Integração dos Dados	46
4.3	Tratamento dos Dados Ausentes e Estimativas	47
4.4	Ferramentas e Softwares	48
4.5	Estratégia de Modelagem Preditiva por Etapas	50
4.6	Modelagem Preditiva com ML e DL	54
5	RESULTADOS E DISCUSSÃO	57
5.1	Análise Descritiva dos Dados	57
5.2	Resultado dos Modelos	62
5.3	Avaliação Comparativa dos Modelos e Considerações Críticas	67
5.4	Análise dos Resultados Considerando os Objetivos, Questões de Pesquisa e Hipóteses	71

6	SIMULAÇÃO DE CENÁRIOS CONTRAFACTUAIS	74
6.1	Fundamentação e Implementação da Simulação Contrafactual . . .	75
6.2	Variáveis e Percentuais de Simulação por Etapa de Ensino	76
6.2.1	Variáveis do Ensino Fundamental (Anos Iniciais e Finais)	77
6.2.2	Variáveis do Ensino Médio	78
6.3	Testes de Simulações Contrafactuais	80
6.4	Resultados das Simulações	82
6.4.1	Simulações do Ensino Fundamental	82
6.4.2	Simulações do Ensino Médio	89
7	CONSIDERAÇÕES FINAIS	101
	REFERÊNCIAS	103
	APÊNDICE A – SCRIPTS DE MODELAGEM PREDITIVA E SI- MULAÇÃO DE CENÁRIOS	110
	ANEXOS	112
	ANEXO A – FONTES DE DADOS UTILIZADAS NA DISSERTAÇÃO	113
	ANEXO B – LISTA DETALHADA DE VARIÁVEIS E DESCRIÇÕES	116

1 Introdução

A educação básica pública, gratuita e de qualidade constitui um direito assegurado pela Constituição Federal de 1988 e um vetor estratégico para o desenvolvimento social e econômico do país (Brasil, 1988). Apesar dos avanços significativos, persistem desigualdades no acesso, permanência e aprendizagem dos estudantes, sobretudo nas regiões Norte e Nordeste (SILVA; GRACIANO, 2022). Os resultados do Índice de Desenvolvimento da Educação Básica (IDEB) de 2023 ilustram essas disparidades: nos anos iniciais do ensino fundamental, as médias foram de 5,2 no Norte e 5,6 no Nordeste, contra 6,3 no Sudeste e 6,4 no Sul; no ensino médio, os índices caíram para 4,2 e 4,1, enquanto Sudeste e Sul registraram 4,3 e 4,4, respectivamente (INEP, 2024).

Os dados do Censo Escolar de 2024 reforçam esse cenário. A taxa de distorção idade-série no ensino fundamental alcançou 17,6% no Norte e 14,1% no Nordeste, diante de 7,5% no Sudeste e 9,3% no Sul. No ensino médio, os percentuais chegaram a 26,8% e 22,2% nas duas primeiras regiões, contrastando com 13,9% no Sudeste e 16% no Sul (INEP, 2024).

O IDEB, criado em 2007, é o principal indicador de qualidade da educação básica no Brasil e resulta da combinação entre o fluxo escolar, medido pelas taxas de aprovação do Censo Escolar, e o desempenho dos estudantes no Saeb (Sistema de Avaliação da Educação Básica) em Língua Portuguesa e Matemática (INEP, 2023b). Embora amplamente utilizado, o indicador evidencia disparidades significativas entre municípios, mesmo diante de políticas de financiamento como o Fundeb Permanente, instituído pela Emenda Constitucional nº 108/2020, e do ICMS Educação (VAAR – Valor Anual por Aluno Resultado), mecanismo de redistribuição que vincula parte dos recursos ao cumprimento de critérios de desempenho e gestão, com o objetivo de incentivar a melhoria dos resultados educacionais (Brasil, 2020).

Estudos apontam que o financiamento educacional é condição indispensável para a qualidade da educação básica, pois viabiliza investimentos em áreas estratégicas como infraestrutura escolar, formação e valorização docente, além do fortalecimento da gestão pedagógica (VIANA; SANTOS et al., 2025). Contudo, o impacto do financiamento sobre os resultados de aprendizagem ocorre de forma indireta, mediado por essas variáveis, e não de maneira automática. Assim, fatores como a eficácia da gestão escolar e as condições socioeconômicas locais também exercem influência decisiva sobre o desempenho educacional, o que explica por que maiores níveis de investimento não garantem, por si só, melhor desempenho (HANUSHEK; WOESSMANN, 2020a; MAIA; BUENO; SATO, 2021).

A compreensão da qualidade educacional requer métodos analíticos capazes de integrar múltiplas variáveis. Nesse sentido, os modelos de aprendizado de máquina (*Machine Learning* – ML) e aprendizado profundo (*Deep Learning* – DL) têm se destacado por

identificar fatores determinantes e prever indicadores, superando limitações das análises tradicionais e oferecendo subsídios mais precisos para a formulação de políticas educacionais (CHEN; DING, 2023; ZAWACKI-RICHTER et al., 2019).

Diante disso, esta pesquisa teve como objetivo central aplicar e avaliar modelos preditivos do desempenho educacional dos municípios piauienses, mensurado pelo IDEB, com base em dados educacionais provenientes do INEP, financeiros extraídos do SIOPE, SICONFI/FINBRA e FUNDEB, bem como indicadores socioeconômicos disponibilizados pelo IBGE. Para tanto, foram aplicados e comparados diferentes algoritmos de ML e DL, selecionando-se o de melhor desempenho. Além da construção e validação dos modelos, realizou-se simulações de cenários exploratórios mediante manipulação controlada das variáveis mais influentes sobre o IDEB, a fim de explorar o comportamento do modelo frente a possíveis alterações nessas variáveis e fornecer subsídios técnicos para estratégias educacionais mais eficazes e políticas públicas.

1.1 Contextualização e Justificativa

A educação constitui dimensão central do desenvolvimento humano e econômico, com impactos diretos na renda, na participação cívica e na igualdade de oportunidades (IPEA, 2021). Contudo, a qualidade educacional no Brasil ainda enfrenta desafios, especialmente em regiões economicamente desfavorecidas. No Piauí, os resultados permanecem abaixo das médias nacionais: em 2023, o IDEB registrou 5,9 pontos nos anos iniciais do ensino fundamental e 5,2 nos anos finais, contra 6,0 e 5,5 observados no país, respectivamente (INEP, 2023a). Além disso, em 2024, apenas 59,8% das crianças do 2º ano estavam alfabetizadas ao final da etapa, percentual que, embora superior ao de anos anteriores, ainda não atende à Meta 2 do Plano Nacional de Educação (2014–2024), que estabelece a universalização da escolarização de 6 a 14 anos e a conclusão de 95% dos estudantes na idade recomendada (Brasil, 2014; PIAUÍ, 2024; IBGE, 2023).

Embora relacionados, qualidade educacional e desempenho educacional são conceitos distintos e complementares. Para Dourado e Oliveira (2009), a qualidade educacional deve ser compreendida a partir do princípio da justiça social, assegurando que todos os estudantes, independentemente de sua origem socioeconômica, tenham acesso a oportunidades educacionais equivalentes. Essa concepção ultrapassa a mera dimensão técnica e quantitativa, incorporando aspectos políticos, sociais e culturais que configuram o direito à educação. Já o desempenho é aferido por indicadores quantitativos, como taxas de aprovação, evasão escolar e os resultados em avaliações padronizadas, como o IDEB, que mede tanto o aprendizado dos estudantes quanto o fluxo escolar (HANUSHEK; WOESSMANN, 2020a).

Para efetivar o direito à educação, garantido pela Constituição Federal do Brasil de 1988, é necessário criar condições que promovam equidade no acesso, permanência,

inclusão e sucesso escolar, incluindo indivíduos fora da idade regular e grupos historicamente marginalizados (SAVIANI, 2013; DOURADO; OLIVEIRA, 2009). Isso implica atender a dimensões intra e extraescolares. Entre as intraescolares, destacam-se a infraestrutura física, os recursos pedagógicos, a formação e valorização docente, o tamanho adequado das turmas e a carga horária compatível com o currículo. Entre as extraescolares, incluem-se fatores como a escolaridade dos pais, as condições socioeconômicas das famílias, o acesso a serviços públicos básicos e o contexto comunitário em que a escola está inserida (CURY, 2010; UNESCO, 2015).

Tais condições exigem critérios objetivos para mensuração e monitoramento, tendo a equidade como princípio norteador, de modo a respeitar a diversidade cultural e eliminar barreiras discriminatórias (UNESCO, 2015). Nesse sentido, o Padrão de Qualidade de Referência (PQR) define parâmetros mensuráveis para orientar financiamento e gestão, incluindo infraestrutura, formação docente, dimensionamento das turmas, carga horária e assistência ao aluno (SILVEIRA; SCHNEIDER; ALVES, 2023).

Instrumentos como o Custo Aluno Qualidade (CAQ) e o Custo Aluno Qualidade Inicial (CAQi) operacionalizam esses padrões, quantificando os recursos necessários para garantir qualidade educacional em todas as etapas da educação básica (anos iniciais e finais do ensino fundamental e ensino médio) (CARREIRA; PINTO, 2007). O CAQi fornece uma referência inicial e emergencial para infraestrutura e insumos básicos; o CAQ propõe valores mais robustos e permanentes. Ambos foram incorporados ao Plano Nacional de Educação (PNE) e reforçados pela Emenda Constitucional n.º 108/2020 (Brasil, 2020).

Para viabilizar a aplicação dos parâmetros definidos pelo PQR, foi desenvolvido o Simulador de Custo Aluno Qualidade (SimCAQ), pelo Laboratório de Dados Educacionais da Universidade Federal do Paraná (UFPR) e da Universidade Federal de Goiás (UFG). Essa ferramenta contextualiza as necessidades reais das escolas públicas ao estimar os recursos por aluno, etapa e modalidade, orientando gestores na alocação de recursos e contribuindo para a justiça social ao reduzir desigualdades entre municípios e regiões (SILVEIRA; SCHNEIDER; ALVES, 2023). Contudo, sem o financiamento adequado para implementar as recomendações do SimCAQ e os padrões do PQR, esses instrumentos permanecem como referências normativas, sem concretizar plenamente o direito à educação de qualidade.

A relação entre qualidade e desempenho educacional é influenciada por uma série de fatores, que envolvem tanto o contexto interno da escola quanto variáveis externas ao ambiente educacional. Fatores internos incluem infraestrutura física, disponibilidade de recursos pedagógicos, formação e valorização docente e gestão escolar, todos componentes fundamentais para garantir processos de ensino e aprendizagem efetivos (CURY, 2010). Já no âmbito externo, aspectos como o contexto socioeconômico das famílias, o grau de escolaridade dos pais, a localização das escolas e a eficácia das políticas públicas voltadas à equidade educacional exercem influência significativa sobre os resultados escolares

(HANUSHEK; WOESSMANN, 2020a). Nesse contexto, o financiamento da educação é fundamental para sustentar as condições adequadas de ensino (UNESCO, 2015).

No Brasil, o financiamento da educação combina receitas municipais e transferências intergovernamentais, como o Fundo de Participação dos Municípios (FPM) e o Fundo de Manutenção e Desenvolvimento da Educação Básica e de Valorização dos Profissionais da Educação (FUNDEB), destinados a reduzir desigualdades regionais (ENCINAS; DUE-NHAS, 2020). Apesar desses mecanismos, persistem ineficiências na alocação dos recursos, sobretudo em municípios com menor capacidade fiscal, o que reforça a necessidade de instrumentos que apoiem decisões mais eficientes.

Estudos recentes demonstram que modelos de ML aplicados à previsão de indicadores educacionais oferecem previsões e subsídios para decisões estratégicas baseadas em evidências (CARAVAGGIO; RESCE; VAQUERO-PIÑEIRO, 2025; WANDERA; MARIVATE; SENGEH, 2020; AULCK; NAMBI; WEST, 2019). Esses trabalhos mostram que os modelos não apenas identificam os determinantes do desempenho, mas também permitem simular políticas públicas sob restrições orçamentárias, promovendo equidade e eficiência.

Caravaggio, Resce e Vaquero-Piñeiro (2025) aplicaram nove algoritmos supervisionados para prever concessões do Fundo Europeu Agrícola de Desenvolvimento Rural (FEADER), destacando o *Random Forest* (RF) como o mais eficaz (acurácia de 73,76%, F1-score de 75,42% e AUC de 80,47%) e evidenciando a importância de fatores territoriais. Wandera, Marivate e Sengeh (2020) utilizaram algoritmos de ML, como RF, XGBoost, Máquina de Vetores de Suporte (Support Vector Machine - SVM) e regressão logística, para prever o desempenho de escolas secundárias da África do Sul e Serra Leoa, classificando-as como “aprovadas” ou “reprovadas” com base em dados institucionais e comunitários. O XGBoost obteve os melhores resultados, com AUC de 77,3% (Serra Leoa) e 82,1% (África do Sul), e acurácia de 65,5% e 75,5%, respectivamente, destacando a influência de infraestrutura, qualificação docente e gestão como fatores mais influentes.

Complementarmente, Aulck, Nambi e West (2019) avaliaram sete modelos de ML para alocação de bolsas de estudo em uma universidade pública norte-americana, com o objetivo de maximizar as matrículas confirmadas entre os alunos admitidos. Os algoritmos testados foram RF, XGBoost, SVM, K-Vizinhos Mais Próximos (*K-Nearest Neighbors* – KNN), regressão logística regularizada, *bagging* e rede neural de três camadas ocultas. O XGBoost apresentou o melhor desempenho, com AUC de 0,85, acurácia de 0,93 e F1-score de 0,91.

Essas evidências reforçam a utilidade dos modelos de ML para lidar com grandes volumes de dados, identificar padrões complexos e projetar cenários alternativos, especialmente em contextos como o piauiense, marcado por desigualdades estruturais e restrições orçamentárias. Apesar dos avanços, a literatura ainda carece de estudos que integrem modelagem preditiva com simulações de cenários exploratórios para estimar os efeitos de estratégias dos fatores influentes em nível municipal. Este trabalho, portanto, propõe a

combinação de ML, DL e simulação de cenários exploratórios para não apenas prever o IDEB municipal, mas também projetar os impactos de alterações nas variáveis influentes, fornecendo evidências para decisões mais eficazes e equitativas.

1.2 Objetivos

1.2.1 Objetivo Geral

Aplicar e avaliar modelos preditivos de ML e DL para estimar o desempenho educacional dos municípios piauienses, representado pelo IDEB, nas três etapas da educação básica, integrando variáveis educacionais, socioeconômicas e financeiras.

1.2.2 Objetivos Específicos

1. Testar e comparar o desempenho de algoritmos de ML e DL para prever o IDEB, utilizando as métricas de Erro Médio Absoluto (MAE), Raiz do Erro Quadrático Médio (RMSE) e Coeficiente de Determinação (R^2), a fim de selecionar os modelos mais acurados para cada etapa da educação básica.
2. Identificar os fatores explicativos com maior relevância para a variação do IDEB, por meio de técnicas de seleção de atributos e análise de importância de variáveis nos modelos de melhor desempenho.
3. Simular cenários exploratórios manipulando variáveis influentes para estimar o impacto de alterações controladas sobre os valores previstos do IDEB municipal em 2023.

1.3 Hipóteses

Esta pesquisa parte do pressuposto de que é possível aplicar modelos de ML e DL capazes de prever o desempenho educacional dos municípios piauienses, a partir de variáveis públicas de natureza educacional, socioeconômica e financeira. Com base nesse marco teórico e nas evidências da literatura, foram formuladas as seguintes hipóteses que orientaram a análise empírica por meio da aplicação de modelos de ML e DL:

- H_1 : Modelos de ML e DL são capazes de prever o IDEB dos municípios piauienses com desempenho avaliado por métricas específicas de erro e ajuste, utilizando exclusivamente variáveis públicas de natureza educacional, socioeconômica e financeira.
- H_2 : Os algoritmos de ML, tais como RF, XGBoost e Gradient Boosting Machine (GBM), apresentam desempenho superior aos modelos lineares tradicionais, como regressão linear, na previsão do IDEB municipal.

- H_3 : A análise de importância das variáveis realizada pelo melhor modelo de ML testado permitirá identificar as mais relevantes para a previsão do IDEB.

1.4 Questões de Pesquisa

Com a finalidade de investigar a aplicação de modelos de ML e DL para a previsão e compreensão do desempenho educacional nos municípios do estado do Piauí, esta pesquisa propõe-se a responder às seguintes questões orientadoras:

1. Quais são os principais desafios para integrar e preparar os dados para modelagem?
2. Qual algoritmo apresentou melhor desempenho na previsão do IDEB municipal?
3. Como interpretar os resultados do modelo de melhor desempenho testado para entender os fatores que influenciam o desempenho educacional dos municípios piauienses?
4. Como o resultado do melhor modelo preditivo pode subsidiar a formulação e o aprimoramento de políticas públicas educacionais nos municípios piauienses?

1.5 Contribuições Esperadas

Esta pesquisa apresenta contribuições em dois eixos principais: metodológico e prático. No âmbito metodológico, destaca-se a integração de diferentes algoritmos de ML e DL aplicados a uma base de dados pública multivariada e de alta dimensionalidade, composta por variáveis educacionais, socioeconômicas e financeiras. Inclui, ainda, a construção de um pipeline analítico replicável e adaptável a diferentes contextos regionais, ampliando sua aplicabilidade a outras unidades federativas com realidades educacionais similares.

Do ponto de vista prático, o estudo oferece subsídios analíticos por meio das previsões do IDEB e das simulações de cenários exploratórios, possibilitando examinar como alterações nas variáveis identificadas pelos modelos como mais influentes podem se refletir no desempenho municipal. A principal contribuição prática concentra-se em apoiar o monitoramento e o acompanhamento do desempenho educacional, favorecendo a identificação de áreas que demandam maior atenção e a discussão de estratégias fundamentadas em evidências. Espera-se, assim, que este estudo contribua para ampliar o uso de análises quantitativas em contextos educacionais, especialmente em unidades federativas que enfrentam desigualdades estruturais, como o estado do Piauí.

1.6 Estrutura da Proposta de Dissertação

Esta dissertação está organizada em seis capítulos. O Capítulo 1 – Introdução apresenta o contexto da pesquisa, a justificativa, os objetivos, as hipóteses, as questões

de pesquisa, as contribuições esperadas e a estrutura do trabalho. O Capítulo 2 – Fundamentação Teórica discute os conceitos centrais de inteligência artificial (IA), com ênfase em ML e DL, além de abordar aspectos relacionados à qualidade e ao desempenho educacional e às políticas de financiamento da educação básica. O Capítulo 3 – Revisão da Literatura reúne o estado do conhecimento, contemplando a estratégia de busca e critérios de seleção, aplicações de ML na educação, estudos sobre simulação de cenários e trabalhos relacionados, de modo a identificar tendências e lacunas.

O Capítulo 4 – Metodologia descreve as fontes de dados utilizadas, os procedimentos de coleta e tratamento, os softwares empregados, a estratégia de modelagem preditiva por etapas e a aplicação dos algoritmos de ML e DL. O Capítulo 5 – Resultados apresenta a análise descritiva dos dados, o desempenho dos modelos testados, a avaliação comparativa e a análise crítica dos achados em relação aos objetivos, hipóteses e questões de pesquisa. O Capítulo 6 – Simulação de Cenários Contrafactuais detalha a fundamentação teórica, a implementação do módulo de simulação, o conjunto de variáveis empregadas, os testes realizados e a análise dos resultados obtidos nos diferentes cenários testados. Por fim, o Capítulo 7 – Considerações Finais sintetiza os principais resultados, discute o alcance dos objetivos propostos, apresenta as limitações da pesquisa e sugere direções para estudos futuros e aprimoramentos da ferramenta desenvolvida.

2 Fundamentação Teórica

Este capítulo apresenta a fundamentação teórica que sustenta o desenvolvimento da pesquisa, abordando os conceitos essenciais de IA, com ênfase em suas subáreas de ML e DL, aplicadas no contexto educacional. A compreensão desses fundamentos é imprescindível para a construção adequada dos modelos preditivos propostos e para a análise crítica dos resultados. Além disso, discute-se os conceitos de qualidade e desempenho educacional, em suas múltiplas dimensões, bem como as especificidades do financiamento público da educação brasileira, aspectos que se relacionam diretamente com os fatores que influenciam o desempenho escolar. A exposição teórica fornece, portanto, o embasamento necessário para a análise dos dados e a interpretação dos resultados, garantindo o rigor científico da investigação e o seu alinhamento com o estado do conhecimento na área.

2.1 Aprendizado de Máquina (Machine Learning)

O ML, subcampo da IA, refere-se ao desenvolvimento de sistemas capazes de aprender a partir de dados, realizando previsões ou decisões sem programação explícita. Essa aprendizagem ocorre por meio da identificação automática de padrões, com ajuste interno de parâmetros a partir da experiência com exemplos anteriores (FEITOSA; SANTOS; SOUSA, 2024). A eficácia dos modelos depende diretamente da qualidade e diversidade dos dados de treinamento, sendo essa dependência um dos principais fatores para sua capacidade de generalização (LUDERMIR, 2021).

Conforme Bigonha (2025), o ML pode ser compreendido como a ciência que capacita máquinas a tomar decisões autônomas, aprendendo associações entre variáveis de entrada (*features*) e saídas esperadas (*labels*), com o objetivo de aplicar esse conhecimento a novos dados. Essa aprendizagem é usualmente classificada em quatro tipos: supervisionado, não supervisionado, semi-supervisionado e por reforço. O aprendizado supervisionado envolve exemplos rotulados e é amplamente utilizado em tarefas de classificação (e.g., detecção de spam) e regressão (e.g., previsão de preços), com técnicas como Regressão Linear, SVM, DT e ANN (MURPHY, 2012).

Por sua vez, no aprendizado não supervisionado, os dados não possuem rótulos e o modelo busca identificar padrões ou estruturas ocultas, geralmente por meio de técnicas de agrupamento (*clustering*) ou redução de dimensionalidade. Métodos como *K-means*, Análise de Componentes Principais (PCA) e Algoritmos de Agrupamento Hierárquico são amplamente empregados para segmentação e exploração de dados (LUDERMIR, 2021). O aprendizado semi-supervisionado combina exemplos rotulados e não rotulados, sendo útil quando a rotulação completa dos dados é inviável (BIGONHA, 2025).

Já o aprendizado por reforço baseia-se em um sistema de recompensas e penalidades,

no qual o agente aprende por tentativa e erro, técnica com aplicações bem-sucedidas em jogos e robótica (LUDERMIR, 2021; FEITOSA; SANTOS; SOUSA, 2024). As principais classificações do aprendizado de máquina estão sintetizadas na Figura 1.

Figura 1 – Classificação dos principais tipos de aprendizado de máquina



Fonte: Elaborado pela autora (2025).

Os modelos de ML fundamentam-se em princípios matemáticos sólidos, como estatística, álgebra linear, cálculo e teoria da probabilidade. A Regressão Linear utiliza mínimos quadrados para estimar relações entre variáveis; a Regressão Logística aplica a função sigmoide para modelar probabilidades; e as SVM resolvem problemas de otimização para maximizar margens entre classes (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). As DT modelam decisões em árvores, as RF, que combinam várias DT por *bagging* para reduzir sobreajuste, e os algoritmos de *boosting*, como o GBM e o XGBoost, corrigem erros iterativamente (CHEN; GUESTRIN, 2016).

Figura 2 – Alguns modelos supervisionados de ML



Fonte: Elaborado pela autora (2025).

A Figura 2 acima, apresenta alguns modelos supervisionados de ML, que variam quanto à complexidade, capacidade preditiva, interpretabilidade e requisitos computacionais. Enquanto modelos lineares são mais transparentes e de fácil interpretação, algoritmos mais avançados, como RF e XGBoost, tendem a oferecer desempenho preditivo superior, embora exijam maior custo computacional e ajustes mais refinados.

A avaliação dos modelos é essencial para mensurar sua capacidade de generalização. Em problemas de classificação, as métricas mais empregadas são acurácia, precisão, revocação, F1-score e área sob a curva ROC (AUC-ROC), cada qual destacando diferentes aspectos do desempenho (SOKOLOVA; LAPALME, 2009). Já em tarefas de regressão, as métricas mais relevantes incluem o MAE, o Erro Quadrático Médio (MSE), a RMSE e o R^2 (MONTGOMERY; PECK; VINING, 2012; WILLMOTT; MATSUURA, 2005), cujas fórmulas são apresentadas a seguir.

- **MAE:** Mede a média dos erros absolutos entre os valores preditos ($\hat{y}_i$) e os valores reais (y_i), fornecendo uma medida intuitiva do desvio médio.

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

- **MSE:** Calcula a média dos erros elevados ao quadrado, penalizando discrepâncias maiores.

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

- **RMSE:** Representa a raiz quadrada do MSE, expressando o erro na mesma unidade da variável predita.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (3)$$

- **R^2 :** Mede a proporção da variabilidade dos dados explicada pelo modelo, sendo um indicador do ajuste do modelo.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (4)$$

onde $\bar{y}$ representa a média dos valores reais da variável dependente.

A combinação dessas métricas oferece uma avaliação abrangente do desempenho dos modelos, mitigando vieses e permitindo escolhas mais assertivas (POWERS, 2011). Além disso, a construção de modelos exige um processo metodológico cuidadoso, incluindo: (i) coleta e seleção de dados representativos e de qualidade; (ii) pré-processamento com

tratamento de ausências e normalização; (iii) treinamento supervisionado ou não supervisionado, a depender da natureza do problema; (iv) validação do modelo por meio de técnicas como a validação cruzada (*k-fold cross-validation*); e (v) aplicação prática do modelo em contextos reais (FEITOSA; SANTOS; SOUSA, 2024).

No contexto do tratamento de dados ausentes, diferentes técnicas podem ser aplicadas conforme a natureza da variável e o comportamento da distribuição. A imputação pela média, embora simples, é sensível a valores extremos, podendo deslocar artificialmente o centro da distribuição. A mediana reduz esse efeito em distribuições assimétricas, enquanto a moda é indicada para variáveis categóricas, por preservar a categoria mais frequente. A escolha adequada da técnica de imputação contribui para maior consistência estatística e precisão nas análises subsequentes (ALAM et al., 2023).

A seleção e ajuste de hiperparâmetros, o controle do sobreajuste e a avaliação da interpretabilidade são fundamentais para garantir tanto o desempenho preditivo quanto a transparência dos modelos. Esse aspecto se torna crítico no caso dos modelos mais complexos, como os baseados em DL, que frequentemente se comportam como “caixas-pretas”. Para mitigar essa opacidade, técnicas como SHAP e LIME (*Local Interpretable Model-agnostic Explanations*) têm sido empregadas, promovendo maior interpretabilidade e responsabilização técnica, especialmente em áreas críticas como educação, saúde e justiça.

Entre os métodos frequentemente utilizados na etapa de pré-processamento e seleção de variáveis em ML e DL, destaca-se o *Variance Inflation Factor* (VIF), métrica que quantifica a magnitude da multicolinearidade entre variáveis independentes. Conforme estabelecido por Montgomery, Peck e Vining (2012), valores de VIF superiores a 10 indicam correlação elevada, podendo comprometer a estabilidade dos coeficientes e a interpretabilidade do modelo.

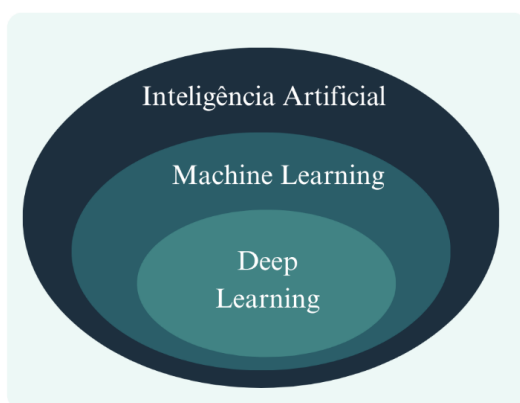
Outro método amplamente empregado é o Lasso (*Least Absolute Shrinkage and Selection Operator*), técnica de regressão com penalização L_1 , proposta por Tibshirani (1996), que permite realizar simultaneamente regressão e seleção de variáveis ao reduzir a zero os coeficientes de menor influência. A utilização da validação cruzada para escolha do parâmetro de regularização (como no LassoCV) é recomendada para prevenir sobreajuste e garantir melhor desempenho preditivo (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

No campo da interpretabilidade, os valores de SHAP fornecem uma decomposição consistente da contribuição de cada variável para as predições de um modelo, fundamentando-se na teoria dos jogos cooperativos e sendo compatíveis com diferentes algoritmos de ML e DL (LUNDBERG; LEE, 2017; MOLNAR, 2020). Em aplicações práticas, é comum adotar limiares de importância como $\text{SHAP} \geq 0,01$ para indicar contribuição mínima relevante (MOLNAR, 2020). Já o algoritmo Boruta realiza seleção de atributos por meio de RF, comparando variáveis reais a versões permutadas (*shadow features*) e mantendo apenas aquelas com importância estatisticamente superior às sombras (KURSER; RUDNICKI, 2010).

2.2 Aprendizado Profundo (Deep Learning)

O DL constitui uma das subáreas mais expressivas e inovadoras da IA contemporânea. Como ilustrado na Figura 3, ele se insere hierarquicamente dentro do campo do ML, que por sua vez compõe o escopo mais amplo da IA. O DL se caracteriza pelo uso de ANN com múltiplas camadas ocultas, as chamadas ANN profundas. Essas redes possuem a capacidade de aprender representações hierárquicas de dados, extraindo características de alto nível a partir de dados brutos de maneira progressiva e automática (LECUN; BENGIO; HINTON, 2015).

Figura 3 – Relação entre IA, ML e DL



Fonte: Elaborado pela autora (2025).

Inspiradas na estrutura do cérebro humano, as ANNs consistem em unidades de processamento chamadas neurônios artificiais, organizados em camadas de entrada, ocultas e de saída. Os pesos das conexões entre os neurônios são ajustados durante o treinamento, geralmente por retropropagação do erro combinada com métodos de otimização como o gradiente descendente (GOODFELLOW; BENGIO; COURVILLE, 2016).

O destaque do DL como paradigma dominante na IA resulta de uma conjunção de fatores tecnológicos e metodológicos. Em primeiro lugar, a crescente disponibilidade de grandes volumes de dados, impulsionada por dispositivos conectados à internet e pelo armazenamento em nuvem. Em segundo lugar, o avanço do poder computacional, sobretudo por meio de unidades de processamento gráfico (GPUs) e, mais recentemente, TPUs (Tensor Processing Units), reduziu significativamente o tempo de treinamento. Por fim, melhorias em algoritmos de otimização, como Adam e RMSprop, e em funções de ativação, como a ReLU, contribuíram para a viabilidade prática de arquiteturas profundas (GOODFELLOW; BENGIO; COURVILLE, 2016).

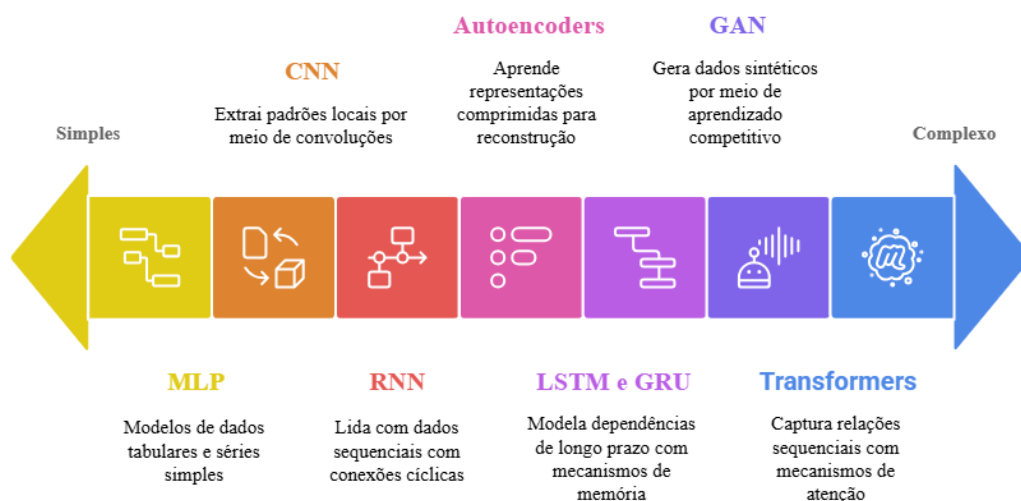
Diferentemente dos algoritmos clássicos de ML, que exigem um esforço substancial em engenharia de atributos (feature engineering), os modelos de DL aprendem automaticamente representações úteis diretamente dos dados, o que é especialmente vantajoso em tarefas com dados não estruturados, como imagens, sinais sonoros e textos. Por exemplo, enquanto abordagens tradicionais de classificação de imagens demandam extração

manual de bordas, texturas e formas, uma rede convolucional aprende essas características de maneira autônoma a partir dos exemplos anotados (KRIZHEVSKY; SUTSKEVER; HINTON, 2017).

No entanto, o DL também apresenta algumas limitações consideráveis. Sua eficácia depende de grandes volumes de dados rotulados, cuja obtenção pode ser onerosa e sujeita a erros. Além disso, os modelos de DL são frequentemente criticados pela sua natureza de “caixa-preta”, ou seja, pela dificuldade de interpretar os critérios que conduzem às suas decisões. Essa opacidade se mostra problemática em aplicações críticas, como decisões judiciais ou diagnósticos médicos. Em resposta a essas preocupações, emergiram técnicas de interpretabilidade como o SHAP e o LIME, que buscam explicar as decisões dos modelos complexos (LUNDBERG; LEE, 2017).

Entre as principais arquiteturas de DL, conforme sistematizado na Figura 4, destaca-se a Rede Neural Convolucional (Convolutional Neural Network – CNN), amplamente utilizada em tarefas de visão computacional. As CNNs se distinguem pela capacidade de extrair automaticamente padrões espaciais por meio de operações de convolução e reduzir a dimensionalidade com camadas de pooling, mantendo a representatividade da informação. Essa arquitetura tem demonstrado elevado desempenho em contextos como diagnóstico por imagem, reconhecimento facial, classificação de imagens satelitais e sistemas de monitoramento visual (KRIZHEVSKY; SUTSKEVER; HINTON, 2017).

Figura 4 – Arquiteturas fundamentais de DL em ordem crescente de complexidade



Fonte: Elaborado pela autora (2025).

Outro grupo importante é a Rede Neural Recorrente (Recurrent Neural Network – RNN), voltada para o processamento de dados sequenciais. Por meio de conexões cíclicas, essas redes conseguem manter uma memória dos estados anteriores, o que as torna adequadas a tarefas que envolvem dependências temporais, como previsão de séries temporais, reconhecimento de fala e tradução automática. Contudo, RNNs convencionais apresentam limitações associadas ao desaparecimento ou explosão dos gradientes durante o

treinamento, dificultando a modelagem de dependências de longo prazo (HOCHREITER; SCHMIDHUBER, 1997).

Para contornar essas limitações, foram desenvolvidas arquiteturas mais sofisticadas, como as LSTM e as GRU (Gated Recurrent Unit), que introduzem mecanismos de controle sobre o fluxo de informações dentro das células recorrentes. Essas variantes têm demonstrado desempenho superior em diversas tarefas de Processamento de Linguagem Natural (PLN) e modelagem de linguagem (HOCHREITER; SCHMIDHUBER, 1997).

Uma das evoluções mais disruptivas do DL foi a introdução da arquitetura Transformer, por Vaswani et al. (2017), que revolucionou o PLN ao eliminar a necessidade de recorrência. Em vez disso, os Transformers utilizam mecanismos de atenção para capturar as relações entre todas as palavras de uma sequência simultaneamente, permitindo maior paralelismo e escalabilidade.

Outros modelos notáveis incluem os Autoencoders e os Autoencoders Variacionais (VAEs), que aprendem representações comprimidas dos dados para reconstrução ou geração. Além disso, as Redes Gerativas Adversariais (Generative Adversarial Networks – GANs), introduzidas por Goodfellow et al. (2014), permitiram avanços impressionantes na geração de imagens sintéticas, estilização e em aplicações artísticas. As GANs consistem em dois modelos: o gerador, que tenta criar dados realistas, e o discriminador, que tenta distinguir entre dados reais e sintéticos, em uma dinâmica de aprendizado competitivo.

A avaliação dos modelos de DL varia conforme a tarefa. Em classificação, são comuns métricas como acurácia, precisão, revocação, medida F1 e AUC-ROC. Para tarefas de regressão, aplicam-se o MAE, o RMSE e o R^2 . Em PLN, como tradução ou geração de texto, utilizam-se métricas como BLEU (Bilingual Evaluation Understudy) e perplexidade. Em geração de imagens, como nas GANs, utiliza-se frequentemente a métrica FID (Frechet Inception Distance), que avalia a qualidade visual das imagens sintéticas. A escolha adequada da métrica é fundamental para garantir que o desempenho do modelo seja mensurado de forma coerente com os objetivos da tarefa, e sua definição deve considerar, além da acurácia bruta, os *trade-offs* entre diferentes tipos de erro (SOKOLOVA; LAPALME, 2009; BORJI, 2019).

Cabe ressaltar que, além das métricas convencionais, outros aspectos também são considerados na prática do DL, como o tempo de treinamento, o consumo computacional, a robustez frente a ruídos nos dados e a capacidade de generalização. O sucesso de um modelo depende não apenas da arquitetura escolhida, mas também do ajuste adequado de seus hiperparâmetros, do uso de regularizações como *dropout* e *batch normalization*, e da qualidade dos dados de entrada. Estratégias como *early stopping*, normalização de dados e técnicas de *data augmentation* também são essenciais para mitigar o sobreajuste e garantir um desempenho consistente do modelo em dados reais (SRIVASTAVA et al., 2014).

2.3 Inteligência Artificial Aplicada à Educação

A IA aplicada à educação configura-se como uma área emergente e promissora, dedicada ao desenvolvimento de tecnologias capazes de simular aspectos do comportamento humano, como o raciocínio, a tomada de decisão e a aprendizagem, para apoiar, potencializar e transformar os processos educativos. De modo geral, busca compreender como sistemas inteligentes podem promover melhorias nos contextos pedagógicos e institucionais, seja por meio da personalização do ensino, da automação de tarefas ou da análise preditiva do desempenho discente (HOLMES et al., 2019).

Na literatura especializada, a IA na educação (IAEd) é definida como a aplicação de algoritmos, modelos e técnicas de IA, como ML, DL e PLN, para interpretar dados educacionais e gerar intervenções ou decisões que visem otimizar a aprendizagem (WANG; LUO, 2024). Esses sistemas aprendem a partir de interações em ambientes digitais, adaptando-se ao progresso dos alunos, diagnosticando dificuldades, sugerindo estratégias de estudo personalizadas e, em alguns casos, interagindo de forma autônoma com os estudantes, como ocorre com os tutores inteligentes (Intelligent Tutoring Systems – ITS). Tais sistemas utilizam modelos computacionais do conhecimento e do comportamento dos aprendizes para fornecer instruções adaptativas e feedback em tempo real, com base em técnicas de ML e análise cognitiva, sendo considerados uma das aplicações mais eficazes da IA na educação (VANLEHN, 2011).

Entre as principais aplicações destacam-se os tutores inteligentes, sistemas de recomendação de conteúdos, correção automática de avaliações, análise preditiva do desempenho, detecção de evasão, assistentes virtuais, análise de sentimentos em discussões online, transcrição e tradução automática de aulas, e suporte à tomada de decisão institucional (AKINWALERE; IVANOV, 2022; ROMERO; VENTURA, 2020). No contexto da avaliação, a IA tem se mostrado eficaz na correção automatizada de textos, como redações e fóruns de discussão, utilizando modelos de PLN e ANN para identificar aspectos linguísticos, sintáticos e semânticos nas produções escritas dos alunos (TAGHIPOUR; NG, 2016).

O uso de IA tem se mostrado promissor na previsão do desempenho acadêmico em instituições, sobretudo por meio da aplicação de algoritmos de ML. Conforme revisão sistemática realizada por Karim-Abdallah et al. (2025), técnicas como DT, SVM, ANN, Naïve Bayes e RF têm sido amplamente utilizadas para estimar o desempenho de estudantes com base em variáveis institucionais, como histórico acadêmico, frequência e participação em atividades educacionais. Esses modelos permitem a identificação precoce de alunos com baixo rendimento, possibilitando intervenções pedagógicas mais eficazes. Ainda segundo os autores, a literatura tem concentrado esforços em utilizar dados educacionais para alimentar modelos computacionais de monitoramento do progresso discente, apoiando a gestão acadêmica e orientando estratégias de apoio educacional baseadas em evidências (KARIM-ABDALLAH et al., 2025).

Apesar de seus potenciais, a implementação da IA na educação enfrenta desafios

significativos. Destacam-se questões éticas e de privacidade dos dados, pois os sistemas inteligentes dependem do acesso a grandes volumes de dados pessoais dos estudantes para funcionar adequadamente. Assim, consentimento informado, transparência algorítmica, governança de dados e mitigação de vieses são temas centrais no debate atual (HOLMES et al., 2023). Além disso, a formação docente representa um entrave relevante, pois muitos professores ainda não possuem preparo para utilizar ou mediar tecnologias baseadas em IA, o que exige investimentos em capacitação e letramento digital. Soma-se a isso o risco de exclusão digital, especialmente em contextos de vulnerabilidade socioeconômica, onde o acesso a dispositivos e conectividade permanece limitado. Portanto, a adoção da IA deve estar associada a políticas públicas que promovam equidade, inclusão e justiça social, garantindo que seus benefícios sejam amplamente distribuídos (GREGORI; SILVA, 2021).

No cenário contemporâneo, observa-se um crescimento expressivo das pesquisas sobre IA na educação, com publicações em periódicos de alto impacto, como *Journal of Educational Data Mining* e *Computers & Education: Artificial Intelligence*, além de iniciativas internacionais voltadas ao desenvolvimento de diretrizes para o uso ético e responsável da IA na educação, como as promovidas pela UNESCO e pela OCDE. Essas iniciativas reforçam que a IA deve ser utilizada como ferramenta de potencialização da aprendizagem, sem substituir a dimensão humana do ensino, reconhecendo o papel insubstituível do professor como mediador do conhecimento (UNESCO, 2021; OECD, 2021).

Dessa forma, a IA aplicada à educação configura-se como uma ferramenta poderosa, desde que utilizada de forma crítica, ética e contextualizada. Seu desenvolvimento deve estar alinhado aos princípios pedagógicos, às especificidades locais e às diretrizes de inclusão digital, para que a inovação tecnológica seja, de fato, um vetor de transformação educacional e promoção do direito à educação de qualidade para todos (UNESCO, 2021).

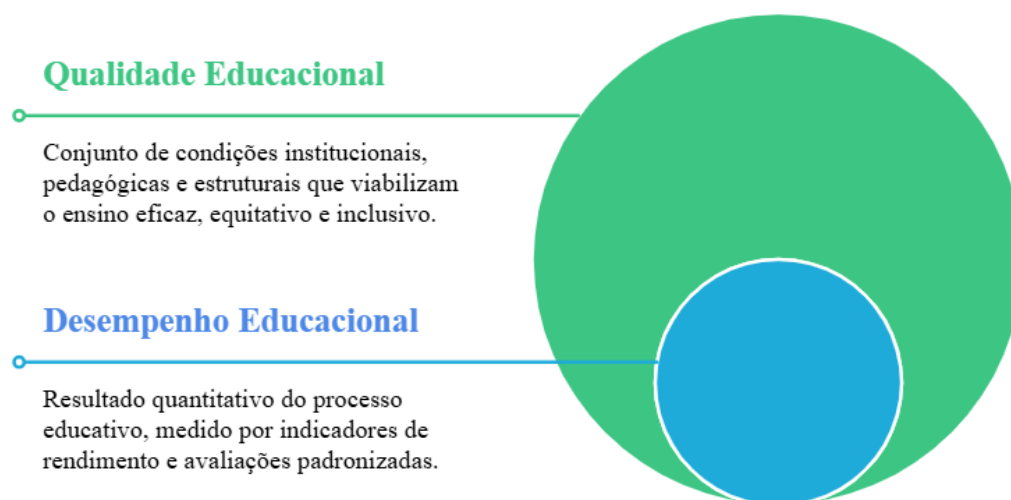
2.4 Qualidade e Desempenho Educacional

A qualidade da educação constitui um conceito amplo e multifacetado, que abrange não apenas o acesso à escola, mas também a permanência, a aprendizagem efetiva e a equidade no ensino. É fundamental, entretanto, distinguir qualidade de desempenho educacional, pois, embora relacionados, não são sinônimos.

A qualidade educacional refere-se ao conjunto de condições institucionais, pedagógicas e estruturais que garantem a efetividade do processo de ensino-aprendizagem, com ênfase em princípios como equidade, inclusão, formação docente e relevância curricular. Já o desempenho educacional diz respeito aos resultados alcançados pelos estudantes, geralmente mensurados por meio de exames padronizados e indicadores de rendimento escolar (SOARES; ARAÚJO, 2006; FREITAS, 2007). Essa distinção conceitual é sistematizada na Figura 5, que apresenta, de forma comparativa, as principais diferenças entre essas duas

dimensões centrais da avaliação educacional.

Figura 5 – Relação entre qualidade e desempenho educacional



Fonte: Elaborado pela autora (2025).

O desempenho educacional é uma expressão quantitativa dos resultados da aprendizagem formal, refletindo tanto os conhecimentos adquiridos quanto a capacidade do sistema educacional de desenvolver competências e habilidades ao longo das etapas de ensino. Conforme Soares e Santos (2020), esse desempenho está fortemente condicionado por fatores contextuais, como o nível socioeconômico dos estudantes, a infraestrutura escolar e o perfil do corpo docente, sendo imprescindível considerá-los na análise dos resultados.

A Lei nº 13.005/2014, que institui o PNE, reconhece o monitoramento do desempenho escolar como instrumento estratégico para a melhoria da qualidade do ensino, sendo operacionalizado por meio de avaliações em larga escala e indicadores sintéticos, como o IDEB, o PISA e as provas do SAEB. Contudo, sua interpretação exige abordagem crítica, atenta às desigualdades regionais, às condições socioeconômicas e às limitações estruturais do sistema educacional.

A qualidade educacional refere-se à capacidade do sistema de ensino em garantir acesso universal, permanência escolar e aprendizagem para todos, independentemente de seu contexto socioeconômico (DOURADO; OLIVEIRA, 2009). Essa visão dialoga com as diretrizes da UNESCO (2022), que enfatiza uma educação inclusiva, relevante e voltada ao desenvolvimento integral dos estudantes, considerando as especificidades culturais, sociais e econômicas de cada território.

Nesse contexto, instrumentos de aferição e monitoramento, fundamentados em evidências, tornam-se essenciais para orientar as políticas públicas. No Brasil, desde os anos 1990, os sistemas de avaliação em larga escala, como o Sistema de Avaliação da Educação Básica (SAEB) e, posteriormente, o IDEB, instituído em 2007 pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), adquiriram centralidade nas

políticas educacionais.

O IDEB é o principal indicador sintético da qualidade da educação básica no Brasil. Seu cálculo resulta do produto entre o aprendizado dos estudantes, aferido pelas proficiências médias em Língua Portuguesa e Matemática no SAEB, e o fluxo escolar, mensurado pela taxa média de aprovação. A nota varia de 0 a 10 e reflete, de forma integrada, tanto o desempenho acadêmico quanto a eficiência do sistema em garantir a progressão escolar adequada (INEP, 2023a). Formalmente, o cálculo para uma unidade escolar j no ano i é representado por:

$$IDEB_{ji} = P_{ji} \times N_{ji} \quad (5)$$

em que P_{ji} é a taxa de aprovação e N_{ji} a média padronizada das proficiências em Língua Portuguesa e Matemática. A padronização baseia-se nos dados do SAEB de 1997, com limites fixados para cada disciplina e etapa (INEP, 2023b). A consolidação do IDEB como parâmetro oficial foi reforçada pela Lei nº 13.005/2014, que institui o PNE para o decênio 2014–2024. A Meta 7 do referido plano estabelece explicitamente a elevação da qualidade da educação básica, com ênfase na melhoria do IDEB, propondo metas específicas para cada etapa de ensino e para os diferentes entes federados (Brasil, 2014).

Apesar de sua relevância, a adoção do IDEB como medida de qualidade educacional tem sido alvo de críticas no meio acadêmico. Segundo Gesqui (2023), a institucionalização do IDEB como parâmetro legal restringe o debate teórico sobre o que constitui educação de qualidade, reduzindo-o a uma lógica tecnocrática baseada em resultados estatísticos. O autor argumenta que a qualidade da educação é uma construção histórica, polissêmica e socialmente situada, e que indicadores como o IDEB, embora úteis para fins de monitoramento, não dão conta da complexidade envolvida nos processos pedagógicos e nas desigualdades estruturais que permeiam o sistema educacional brasileiro. Essa crítica dialoga com reflexões anteriores de Freitas (2007), que já advertia sobre o risco de mascaramento da má qualidade educacional por meio de indicadores quantitativos, além de chamar atenção para os usos políticos desses instrumentos.

A Constituição Federal de 1988 e a Lei de Diretrizes e Bases da Educação Nacional (Lei nº 9.394/1996) garantem que a educação seja direito de todos e dever do Estado, a ser oferecida com qualidade, equidade e valorização dos profissionais da educação (Brasil, 1988; Brasil, 1996). O PNE, nesse contexto, funciona como plano de ação para concretizar esses princípios. Entretanto, conforme Shiroma, Moraes e Evangelista (2004), políticas educacionais brasileiras das últimas décadas têm sido moldadas também por diretrizes internacionais, muitas vezes vinculando financiamento e políticas a metas e indicadores alinhados às demandas do mercado global.

Estudos têm demonstrado que o desempenho escolar resulta da interação entre fatores internos, como formação docente, recursos pedagógicos, uso de tecnologias, tamanho das turmas e tempo de ensino, e fatores externos, como renda familiar, escolaridade dos

país, acesso à saúde, alimentação e moradia digna (SOARES; ARAÚJO, 2006). Essas interações produzem trajetórias educacionais desiguais, especialmente em contextos de vulnerabilidade social.

As desigualdades regionais impactam o desempenho educacional no Brasil. Os resultados do IDEB revelam que as regiões Norte e Nordeste permanecem sistematicamente abaixo da média nacional, enquanto Sul e Sudeste concentram os melhores indicadores, refletindo desigualdades históricas na distribuição de investimentos públicos e na capacidade arrecadatória dos municípios. Essas condições limitam a manutenção de infraestrutura adequada e a contratação de docentes qualificados. Embora políticas como o Fundeb tenham contribuído para atenuar tais disparidades, ainda persistem desafios políticos e administrativos relacionados à redistribuição de recursos (ENCINAS; DUENHAS, 2020).

Tabela 1 – IDEB observado nos anos iniciais do ensino fundamental por região (2013–2023)

Ano	Norte	Nordeste	Centro-Oeste	Sudeste	Sul	Brasil
2013	4,3	4,3	5,5	5,9	5,8	5,2
2015	4,7	4,8	5,7	6,1	6,0	5,5
2017	4,9	5,1	6,0	6,4	6,2	5,8
2019	5,0	5,4	6,1	6,5	6,3	5,9
2021	5,0	5,3	5,8	6,1	6,2	5,8
2023	5,2	5,6	6,1	6,3	6,4	6,0

Fonte: INEP (2024).

A Tabela 1 evidencia essa trajetória desigual ao apresentar o IDEB observado nos anos iniciais do ensino fundamental por região, entre 2013 e 2023. Nota-se que, enquanto Norte e Nordeste evoluíram de 4,3 para 5,2 e 5,6, respectivamente, no mesmo período, Sudeste e Sul mantiveram desempenho superior, alcançando 6,3 e 6,4 em 2023, valores acima da média nacional de 6,0 (INEP, 2024).

Além do IDEB, outros instrumentos complementam a avaliação da educação básica. O SAEB fornece dados detalhados das provas aplicadas a estudantes do 5º e 9º anos do ensino fundamental e da 3ª série do ensino médio, servindo também para análises específicas de aprendizagem (INEP, 2023b). O Índice de Nível Socioeconômico (INSE), calculado com base na escolaridade dos pais, posse de bens, renda familiar e acesso à internet, permite contextualizar os resultados, identificando desigualdades de oportunidade (INEP, 2022). Outros indicadores relevantes são a média de alunos por turma, a carga horária diária, a proporção de professores com formação adequada e a estabilidade do vínculo docente (INEP, 2022).

Portanto, embora o IDEB represente o principal parâmetro de avaliação do desempenho educacional no Brasil, sua utilização requer uma leitura crítica e contextualizada, articulada a outros indicadores que captem a complexidade da realidade educacional. Essa necessidade de compreensão integrada reforça a importância de abordagens analíticas capazes de explorar múltiplas dimensões simultaneamente, como aquelas utilizadas neste estudo, que emprega modelos de ML e DL para estimar o IDEB municipal a partir de

variáveis associadas às condições educacionais, socioeconômicas e estruturais dos municípios. Tal perspectiva permite examinar de forma sistemática como diferentes fatores se relacionam com o desempenho observado, contribuindo para diagnósticos mais precisos e alinhados às particularidades regionais.

Assim, promover a qualidade da educação e compreender o desempenho escolar dos estudantes demanda um olhar sistêmico e multidimensional, que reconheça a influência dos contextos sociais, econômicos, culturais e estruturais nos processos de ensino e aprendizagem. A consolidação de uma educação pública de qualidade, democrática e inclusiva depende da articulação entre financiamento adequado, valorização e formação dos profissionais da educação, gestão escolar eficiente e formulação de políticas sensíveis às diversidades regionais. Nesse sentido, análises quantitativas como as desenvolvidas nesta pesquisa oferecem subsídios para interpretar o IDEB de maneira contextualizada, contribuindo para o entendimento das desigualdades históricas que permeiam o sistema educacional brasileiro.

2.5 Financiamento Público na Educação Brasileira

O financiamento público da educação no Brasil constitui uma das bases estruturantes para a garantia do direito à educação de qualidade, conforme estabelece a Constituição da República Federativa do Brasil de 1988. O artigo 211 determina a responsabilidade solidária entre os entes federativos quanto à oferta educacional, cabendo à União, aos estados, ao Distrito Federal e aos municípios a cooperação na gestão e no custeio do sistema educacional (Brasil, 1988). Essa diretriz foi aprofundada pela Lei de Diretrizes e Bases da Educação Nacional (Lei nº 9.394/1996), que define os princípios de equidade, qualidade e valorização do magistério como orientadores da distribuição de recursos e da oferta educacional (Brasil, 1996).

No que se refere à composição das fontes de custeio, a educação básica brasileira é financiada predominantemente por recursos vinculados à arrecadação de impostos, conforme disposto no artigo 212 da Constituição Federal. Estados, o Distrito Federal e os municípios devem investir, no mínimo, 25% da receita resultante de impostos na manutenção e desenvolvimento do ensino (MDE), enquanto à União cabe aplicar ao menos 18%. Complementam esse financiamento as transferências constitucionais, legais e voluntárias, como o Fundo de Participação dos Estados (FPE), Fundo de Participação dos Municípios (FPM), cota-parte do ICMS, Cota do Salário-Educação e o repasse do FUNDEB.

Um marco importante na estruturação do financiamento da educação foi a criação do Fundo de Manutenção e Desenvolvimento do Ensino Fundamental e de Valorização do Magistério (FUNDEF), por meio da Emenda Constitucional nº 14/1996. Em 2007, foi instituído o FUNDEB, ampliando a abrangência para toda a educação básica. Tornado

permanente pela Emenda Constitucional nº 108/2020, o novo FUNDEB introduziu mecanismos redistributivos como o Valor Aluno Ano Total (VAAT) e o Valor Aluno Ano Resultados (VAAR), além de uma complementação progressiva da União, que chegará a 23% até 2026 (Brasil, 2020).

No que tange à repartição federativa das responsabilidades, observa-se que os governos estaduais concentram-se predominantemente na oferta do ensino médio, ainda que algumas redes mantenham matrículas no ensino fundamental. O ICMS constitui a principal fonte de financiamento estadual, complementado por recursos provenientes do FUNDEB e por transferências voluntárias da União. Por sua vez, os municípios, responsáveis pela educação infantil e pelos anos iniciais do ensino fundamental, têm no FUNDEB sua principal fonte de custeio, seguido do FPM, da cota-parte do ICMS e de receitas próprias, como ISS e IPTU. Em média, o FUNDEB responde por mais de 60% do financiamento educacional municipal, sobretudo nos municípios de pequeno porte (SIOPE, 2023).

Cabe ainda destacar, no contexto piauiense, a implementação do ICMS Educação – VAAR, mecanismo instituído pela Lei nº 7.429/2020 no Piauí, que estabelece a distribuição de recursos com base em indicadores educacionais. Esse modelo vincula parte do repasse do ICMS ao desempenho dos municípios, considerando tanto resultados de aprendizagem quanto critérios de equidade. A adesão ao ICMS Educacional é condição necessária para que o estado receba a complementação-VAAR do Fundeb, reforçando a necessidade de cooperação entre o governo estadual e os municípios. Desde 2022, 5% do ICMS estadual passaram a ser distribuídos com base nesse critério, percentual que aumentou para 10% a partir de 2023, calculado a partir do Índice de Qualidade em Educação Municipal (IQEM), que combina a taxa de aprovação do 1º ao 5º ano e os resultados do SAEPI (Piauí, 2020; FNDE, 2023).

O caso do estado do Piauí ilustra os desafios enfrentados por redes municipais e estaduais em contextos de alta dependência de transferências constitucionais e baixa capacidade arrecadatória. Com um dos menores PIBs per capita do país, mais de 70% dos municípios com menos de 20 mil habitantes e elevada vulnerabilidade fiscal, o estado conta com o FUNDEB como principal pilar do custeio da educação (IBGE, 2023). Estudos do INEP (2023a) indicam que aproximadamente 95% das receitas municipais para a educação no Piauí provêm de transferências intergovernamentais, com investimento médio por aluno abaixo da média nacional, em torno de R\$ 6.500/ano. Desse total, mais de 80% são destinados ao pagamento de pessoal, limitando investimentos em infraestrutura, formação docente e inovação pedagógica (SIOPE, 2023).

Além da predominância dos gastos com pessoal, os orçamentos educacionais contemplam subfunções como educação infantil, ensino fundamental, educação especial, educação de jovens e adultos, administração do ensino e transporte escolar, custeadas por receitas tributárias vinculadas e transferências intergovernamentais. A correta alocação desses

investimentos é fundamental para assegurar a manutenção das escolas, a provisão de materiais didáticos e o funcionamento de serviços essenciais à permanência e ao aprendizado dos estudantes (Brasil, 1996; PINTO, 2018).

Em relação ao FUNDEB, pelo menos 70% dos recursos devem ser aplicados na remuneração dos profissionais da educação básica em exercício, enquanto os 30% restantes destinam-se à manutenção e desenvolvimento do ensino, incluindo construção e reforma de escolas, aquisição de mobiliário, equipamentos e tecnologias, transporte escolar e formação continuada. Esse modelo visa garantir tanto a valorização profissional quanto condições adequadas de funcionamento da rede pública. Contudo, sua efetividade depende da capacidade de gestão dos entes federativos, especialmente nos municípios mais vulneráveis (Brasil, 2020; HANUSHEK; WOESSMANN, 2020a).

A literatura especializada ressalta que o volume de recursos, embora relevante, não assegura por si só a qualidade educacional. Segundo Hanushek e Woessmann (2020b), a eficácia do financiamento decorre do uso estratégico dos recursos, sobretudo em contextos de vulnerabilidade social. Alves e Frio (2020) demonstram que municípios com maior investimento por aluno tendem a alcançar melhores resultados no IDEB quando aliados à boa gestão, formação docente e infraestrutura adequada. No cenário nacional, observa-se retração dos investimentos: relatório da OCDE (2023) indica que, entre 2015 e 2021, o Brasil registrou redução média anual de 2,5% nos investimentos públicos em educação, em contraste com um aumento médio de 2,1% nos demais países-membros, o que ameaça a sustentabilidade das políticas de expansão e qualidade, sobretudo em estados como o Piauí.

Diante desses desafios, é necessário adotar mecanismos que permitam acompanhar e avaliar de forma sistemática a aplicação dos recursos educacionais. O uso de tecnologias digitais, sistemas de informação e indicadores de desempenho pode apoiar a gestão pública educacional, tornando o monitoramento mais claro e acessível (ZAWACKI-RICHTER et al., 2019). Nesse contexto, análises quantitativas e modelos preditivos disponíveis na literatura podem ser aplicados para examinar como diferentes características dos municípios se relacionam com os resultados educacionais representados pelo IDEB. Essa perspectiva se articula com a proposta deste estudo, que emprega modelos já consolidados de ML e DL para estimar o IDEB municipal a partir de variáveis que refletem condições educacionais, socioeconômicas e estruturais observadas nos municípios, sem desenvolver novas arquiteturas ou métodos próprios.

Assim, o fortalecimento da gestão pública em educação depende da ampliação da capacidade administrativa dos estados e municípios, do uso adequado dos recursos disponíveis e da análise criteriosa dos fatores que influenciam o desempenho escolar. No Piauí, essa necessidade é ainda mais evidente devido às desigualdades regionais. A aplicação dos modelos preditivos neste estudo contribui para essa análise ao permitir observar como diferentes variáveis se associam às variações do IDEB municipal.

3 Revisão da Literatura

A revisão da literatura apresentada neste capítulo tem como finalidade consolidar o embasamento teórico necessário para a condução da pesquisa. Busca-se contextualizar as principais abordagens contemporâneas em IA aplicadas à educação, delinear os fundamentos conceituais de técnicas como ML e DL, bem como discutir aspectos relacionados à qualidade educacional e ao financiamento público da educação básica no Brasil. Além disso, são explorados estudos empíricos recentes que tratam da modelagem preditiva do desempenho educacional, com vistas à identificação de tendências, lacunas e oportunidades de avanço científico na área.

3.1 Estado do Conhecimento

Esta seção apresenta o estado do conhecimento acerca da aplicação de técnicas de ML e DL em contextos educacionais. O termo é aqui empregado na acepção de Morosini (2015), que o define como um mapeamento sistematizado e delimitado da produção científica existente. Diferentemente de um inventário descritivo, essa abordagem busca explicitar como e em que condições o tema tem sido investigado, quais contribuições foram consolidadas, quais tendências emergem no campo e quais lacunas metodológicas permanecem abertas. Esse procedimento permite situar criticamente a presente pesquisa, evidenciando seu diálogo com estudos anteriores e destacando os aspectos em que avança ou se diferencia.

3.1.1 Estratégia de Busca e Critérios de Seleção

A busca bibliográfica foi realizada nas bases SciELO, Portal de Periódicos CAPES, Scopus, Web of Science, SpringerOpen, MDPI, Wiley Online Library e Google Scholar. A estratégia foi conduzida em dois ciclos: inicialmente com descritores em português e, posteriormente, com a mesma combinação de termos traduzida para o inglês, com o objetivo de ampliar a cobertura internacional da revisão.

Os descritores foram organizados em três eixos temáticos: (i) técnicas computacionais, (ii) contexto educacional e (iii) indicadores e avaliação. As combinações foram realizadas por meio de operadores booleanos (AND, OR), integrando termos relacionados a ML e DL ao contexto da educação básica e à previsão de indicadores educacionais. Entre os principais descritores utilizados destaca-se: “aprendizado de máquina”, “aprendizado profundo”, “inteligência artificial na educação”, “educação básica”, “desempenho acadêmico”, “resultados de aprendizagem”, “desempenho escolar”, “indicadores educacionais”,

“IDEB”, “avaliação educacional”, “desempenho educacional” e “previsão de indicadores educacionais”.

O recorte temporal concentrou-se no período de 2019 a 2025, marcado pela intensificação das aplicações de ML e DL na educação. Entretanto, estudos publicados a partir de 2015 também foram incorporados quando apresentavam relevância teórica ou metodológica. Para trabalhos de caráter conceitual, não se adotou limite temporal rígido, considerando-se autores de maior impacto acadêmico.

Nesse contexto, foram incluídos estudos empíricos que aplicaram ML e/ou DL em ambientes educacionais, com ênfase na previsão de indicadores de desempenho por meio de modelos de classificação ou regressão. A seleção priorizou artigos publicados em periódicos revisados por pares, admitindo-se teses e dissertações como material complementar quando pertinentes ao escopo da pesquisa. Embora o foco recaia sobre a educação básica, também foram considerados estudos no ensino técnico e superior quando relevantes para evidenciar a eficiência das técnicas.

Quadro 1 – Critérios metodológicos adotados para o estado do conhecimento

Dimensão	Descrição
Bases de dados	SciELO, Periódicos CAPES, Scopus, Web of Science, SpringerOpen, MDPI, Wiley e o Google Scholar.
Idiomas	Português e inglês.
Descritores	“aprendizado de máquina”, “aprendizado profundo”, “inteligência artificial na educação”, “educação básica”, “desempenho acadêmico”, “resultados de aprendizagem”, “desempenho escolar”, “indicadores educacionais”, “IDEB”, “avaliação educacional”, bem como expressões como “desempenho educacional” e “previsão de indicadores educacionais”.
Recorte temporal	2019–2025 (com inclusão seletiva a partir de 2015 para trabalhos de alta relevância).
Inclusão	Estudos empíricos que aplicaram ML/DL em contextos educacionais, voltados à previsão de indicadores (classificação ou regressão); preferência por artigos revisados por pares; teses e dissertações admitidas como complemento.
Exclusão	Trabalhos sem aplicação preditiva; áreas externas à educação (saúde, indústria); estudos restritos ao ensino superior em áreas específicas sem relação com indicadores; ensaios teóricos, revisões sem resultados empíricos, relatos de experiência, duplicatas e textos incompletos.
Aspectos observados	Modelos utilizados; nível de ensino; nível de aplicação (aluno, escola, município); tipos de variáveis empregadas.

Fonte: Elaborado pela autora (2025).

Por outro lado, foram excluídos trabalhos que não apresentavam aplicação preditiva de ML ou DL, restringindo-se a análises descritivas ou exploratórias, bem como estudos aplicados a áreas externas ao campo educacional, como saúde e indústria. Além disso, não foram contempladas pesquisas voltadas exclusivamente a áreas profissionais específicas, sem

conexão com indicadores educacionais; ensaios teóricos; revisões sem resultados empíricos; relatos de experiência sem validação metodológica.

A sistematização desses parâmetros encontra-se no Quadro 1 acima, que reúne as dimensões metodológicas adotadas nesta revisão. O quadro não apenas resume as bases consultadas, idiomas, descritores e recorte temporal, mas também explicita os critérios de inclusão e exclusão, assegurando transparência ao processo de seleção e coerência entre a revisão da literatura e os objetivos desta pesquisa.

3.1.2 Trabalhos Relacionados e Lacunas Identificadas

O uso de técnicas de ML no contexto educacional tem se expandido progressivamente, impulsionado pela crescente digitalização dos ambientes de aprendizagem e pela ampla disponibilidade de dados educacionais. Modelos supervisionados e não supervisionados vêm sendo empregados para diversas finalidades, como previsão de desempenho acadêmico, detecção precoce de risco de evasão, personalização de trajetórias de aprendizagem e apoio à gestão escolar (DUTT; ISMAIL; HERAWAN, 2017; ZAWACKI-RICHTER et al., 2019).

No âmbito da modelagem preditiva, destaca-se abordagens supervisionadas, como Regressão Linear, RF e Redes Neurais Artificiais (Artificial Neural Networks - ANN), amplamente utilizadas para prever notas, taxas de aprovação e outros indicadores acadêmicos a partir de variáveis socioeconômicas, demográficas e pedagógicas. Já as técnicas não supervisionadas, como *clustering* e análise de agrupamento, auxiliam na identificação de padrões ocultos nos dados, permitindo a segmentação de estudantes em perfis de risco ou potencial de desempenho (ROMERO; VENTURA, 2020). Além disso, observa-se a incorporação de abordagens baseadas em aprendizado por reforço e em DL para aplicações mais complexas, como sistemas de tutoria inteligente, análise automatizada de textos e previsão de trajetórias educacionais (HOLMES et al., 2019).

Apesar desse avanço metodológico, os estudos empíricos em modelagem preditiva aplicada à educação apresentam significativa heterogeneidade quanto às variáveis utilizadas, às unidades de análise adotadas e aos modelos empregados. Tal diversidade demanda, nesta dissertação, a definição de critérios analíticos que permitam avaliar sistematicamente a aderência dos trabalhos ao escopo da pesquisa. Conforme sintetizado no Quadro 2, esses critérios abrangem o recorte municipal como unidade de análise, a integração de múltiplas dimensões explicativas, o uso de modelos regressivos supervisionados e a aplicação de simulações contrafactuais, orientando a leitura crítica da literatura e permitindo identificar com clareza as lacunas que este estudo busca suprir.

Dessa forma, a sistematização desses critérios fornece a base para a análise dos trabalhos relacionados, garantindo maior consistência metodológica e coerência epistemológica. A partir dessa estrutura, torna-se possível compreender como a literatura tem explorado a previsão de indicadores educacionais e em quais pontos ainda persistem limitações.

No Brasil, o IDEB consolidou-se como o principal indicador de qualidade da

Quadro 2 – Critérios analíticos para avaliação de estudos empíricos em modelagem preditiva da educação

Critério Avaliado	Justificativa
Unidade de análise em nível municipal	Alinha-se ao foco territorial desta pesquisa, permitindo que os modelos preditivos gerem subsídios diretos para a gestão educacional local.
Uso de múltiplas dimensões (educacional, socioeconômica e financeira)	Amplia o poder explicativo dos modelos ao considerar fatores interdependentes que afetam o desempenho educacional.
Modelagem preditiva do tipo regressiva	Permite maior granularidade na análise do IDEB, favorecendo intervenções mais específicas do que classificações categóricas.
Inclusão dos Anos Iniciais do Ensino Fundamental	Etapla estratégica na formação escolar, frequentemente negligenciada nos estudos preditivos.
Simulação de cenários contrafactuais	Inovação metodológica que possibilita testar virtualmente o impacto de políticas educacionais, promovendo a aplicabilidade dos resultados.

Fonte: Elaborado pela autora (2025).

educação básica, articulando dimensões de rendimento escolar e desempenho acadêmico em avaliações externas (INEP, 2023b). Nesse contexto, a aplicação de técnicas preditivas para antecipar sua evolução tem se consolidado como linha estratégica de pesquisa, em virtude do potencial de subsidiar políticas públicas mais eficazes. O avanço do ML e do DL, aliado à crescente disponibilidade de dados educacionais, institucionais e socioeconômicos, impulsionou novas abordagens analíticas em larga escala, aplicadas desde a educação básica até o ensino superior. Na educação básica, em particular, sobressaem os estudos voltados à previsão de indicadores como notas escolares, taxas de evasão e, sobretudo, o IDEB e os resultados da Prova Brasil.

Entre essas contribuições, destaca-se o trabalho de Maia, Bueno e Sato (2021), que sobressai como referência central desta dissertação, ao propor modelos preditivos do IDEB em nível agregado, país, regiões e municípios, com base em 61 variáveis educacionais, estruturais e socioeconômicas. Utilizando técnicas como regressões penalizadas (Lasso, Ridge, Elastic Net) e GBM, este último apresentou os melhores desempenhos: R^2 de 0,421 e RMSE de 0,761 nos anos iniciais, com o percentual de professores com ensino superior como principal variável preditiva; R^2 de 0,305 e RMSE de 0,826 nos anos finais, onde o PIB educacional per capita teve maior impacto; e no ensino médio, R^2 de 0,459 com RMSE de 0,763 no ensino médio, sendo a média de horas de aula diária a variável mais relevante. Esses resultados evidenciam a necessidade de abordagens preditivas sensíveis às especificidades de cada ciclo educacional, bem como o desafio representado pelos elevados erros residuais decorrentes da heterogeneidade socioeducacional dos municípios brasileiros.

De forma complementar, Soares e Santos (2024) investigaram o impacto de variáveis educacionais e socioeconômicas sobre o IDEB de 2019 em 222 escolas estaduais de ensino

médio no Espírito Santo, utilizando Regressão Linear Múltipla. Assim, a unidade de análise e predição adotada foi a escola, buscando explicar as variações do IDEB entre diferentes instituições educacionais da rede pública estadual. O modelo atingiu um R^2 ajustado de 0,508, indicando que pouco mais de 50% da variação do IDEB foi explicada por 13 variáveis independentes relativas ao perfil dos alunos, dos docentes e à infraestrutura escolar. Destacaram-se a distorção idade-série como fator de maior impacto negativo e, positivamente, o nível socioeconômico e o percentual de alunos brancos. Esse estudo reforça a influência das variáveis contextuais e a necessidade de políticas sensíveis às desigualdades estruturais.

Outros estudos, nacionais e internacionais, também têm contribuído para o avanço da modelagem preditiva na educação, aplicando uma diversidade de modelos em distintos níveis e modalidades educacionais. Benevento e Meirelles (2023) avaliaram 21 algoritmos de classificação, incluindo RF, GBM, SVM, Naïve Bayes, Árvore de Decisão (Decision Tree - DT) e ANN, na predição do desempenho discente, com base em um conjunto de dados composto por 900 registros de alunos de uma escola online de desenvolvimento de software. Entre os modelos testados, a DT e GBM apresentaram o melhor desempenho, atingindo acurácia de 100%.

Em outro estudo, Benevento e Meirelles (2023) aplicou GBM, RF e Logistic Ridge em dados de uma escola técnica em Joinville-SC, com RF obtendo acurácia de 95,28% e AUC-ROC de 95,95%. Rodrigues et al. (2024) exploraram Transformers e XGBoost em dados do ensino fundamental e médio, oriundos de avaliações contínuas em uma plataforma educacional, com o XGBoost alcançando AUC-PR (área sob a curva precisão-recall) de 0,9856.

Com abordagem voltada ao município de Teotônio Vilela (AL), Pinto, Júnior e Costa (2019) analisaram o impacto de variáveis educacionais e socioeconômicas sobre o IDEB do ensino fundamental entre 2015 e 2017. A variável-alvo foi categorizada como “acima” ou “abaixo” da média municipal. Utilizando o método SMOTE para balanceamento da base e técnicas de seleção de atributos, os autores aplicaram oito algoritmos de classificação (J48, OneR, JRip, LibSVM, RF, Naïve Bayes, IBK e REPTree). Os melhores desempenhos foram obtidos com os modelos OneR, LibSVM e J48, cujas acurácias ultrapassaram 98% tanto em Língua Portuguesa quanto em Matemática.

O estudo de Rocha, Teixeira e Melo (2015) aplicou Regressão Linear Múltipla para analisar os fatores que influenciam o desempenho escolar de alunos do 5º e 9º ano do ensino fundamental na rede pública do estado do Rio Grande do Norte. A base de dados utilizada foi composta por alunos de 1.502 escolas, extraída dos microdados do SAEB 2011. As variáveis consideradas incluíram o contexto familiar (como escolaridade da mãe e posse de bens), formação e experiência docente, além de infraestrutura escolar. Os resultados indicaram que os fatores familiares apresentaram maior poder explicativo sobre as notas em Língua Portuguesa e Matemática, enquanto variáveis relacionadas aos professores

não foram estatisticamente significativas. A presença de biblioteca e de equipamentos tecnológicos teve efeito positivo em ambas as disciplinas. Os modelos apresentaram R^2 entre 0,415 e 0,468, indicando bom ajuste.

No ensino superior, Teodoro e Kappel (2020) aplicaram técnicas de ML para prever o risco de evasão escolar em instituições públicas de ensino superior no Brasil, utilizando dados do INEP com informações acadêmicas e socioeconômicas dos estudantes. Utilizaram uma base com mais de 376 mil registros de estudantes, construída a partir de dados do Censo Escolar e do INEP. Foram selecionadas 160 variáveis preditoras, incluindo características sociodemográficas, estruturais, desempenho acadêmico (notas, frequência) e variáveis institucionais (localidade da escola, regime de ensino). Foram empregados cinco modelos preditivos: Naive Bayes, KNN, DT, RF e ANN. O modelo RF apresentou o melhor desempenho, com acurácia próxima de 80%.

Freitas e Salgado (2020) desenvolveram uma ferramenta de simulação de cenários para apoiar educadores na compreensão dos resultados de modelos preditivos aplicados à evasão em cursos de ensino a distância. A base de dados utilizada incluiu 432 estudantes, sendo 323 evadidos e 109 não evadidos, com variáveis como notas, número de acessos à plataforma, participação em fóruns e dados sociodemográficos. Foram testados três algoritmos: KNN, com acurácia final de 83,8%; uma DT enviesada propositalmente, com acurácia de 100% (devido a *overfitting*); e uma Rede Neural Perceptron Multicamadas (Multilayer Perceptron – MLP), cuja melhor configuração com o otimizador Adam resultou em acurácia de 88,5%. O diferencial do estudo foi o desenvolvimento de uma interface interativa onde educadores puderam simular alterações nas variáveis dos alunos e observar o impacto nas previsões, promovendo uma reflexão crítica sobre os vieses e limitações dos modelos.

Silva (2023) investigou o uso de modelos de ML para prever a evasão de alunos em cursos técnicos de nível médio, com base em dados socioeconômicos e no desempenho dos estudantes no processo seletivo. A base de dados construída reuniu 789 registros de alunos com 12 atributos, incluindo idade, renda, curso, afrodescendente e variável indicadora de conclusão do curso. Foram testados cinco algoritmos: Naive Bayes, KNN, SVM, DT, RF e ANN (MLP), utilizando validação cruzada com 10 divisões. A acurácia média inicial foi de 73,3%, sendo posteriormente melhorada por meio da normalização dos dados e ajuste de hiperparâmetros via *RandomizedSearchCV*. Os melhores resultados por curso foram: Administração (RF, 82,3%), Informática (SVM, 72,1%), Nutrição e Dietética (RF, 74,4%) e Química (SVC, 66,5%).

No âmbito internacional, Chen e Ding (2023) analisaram o impacto de diferentes modelos de ML na previsão do desempenho acadêmico em escolas da Pensilvânia, utilizando dados educacionais e socioeconômicos. Foram testados cinco modelos: DT, RF, Regressão Logística, SVM e Rede Neural. Destes, a Rede Neural obteve a maior precisão nos testes, com 60%. Wang e Luo (2024) desenvolveram um modelo preditivo aplicado a 87 estudantes

universitários de um curso de japonês em Wuhan, analisando algoritmos como RF, Rede Neural Artificial, SVM e XGBoost. O XGBoost combinado com SHAP (*SHapley Additive exPlanations*) apresentou o melhor resultado, com MAE de 0,6 e R^2 de 0,82. A análise de SHAP evidenciou que os fatores mais influentes foram habilidades de autoaprendizagem, frequência nas aulas, participação em atividades extracurriculares e métodos de ensino adotados.

Guanin-Fajardo, Guaña-Moya e Casillas (2024) investigaram o desempenho acadêmico de estudantes utilizando técnicas de ML aplicadas ao contexto universitário. Para isso, os autores utilizaram a metodologia CRISP-DM, analisando um conjunto de dados composto por 6.690 registros e 21 variáveis, abrangendo informações acadêmicas e socioeconômicas dos alunos. Foram empregados os algoritmos de classificação DT, SVM, RF e ANN, bem como GBM, XGBoost, Bagging e Naïve Bayes. O XGBoost apresentou o melhor desempenho, com precisão de 79,49% e AUC de 87,75%.

Por sua vez, Pereira e Seruffo (2024) aplicaram o algoritmo MLP para prever o desempenho em Matemática e Português de 650 estudantes de uma escola secundária em Portugal e assim classificar suas performances em “baixo”, “médio” e “alto”. As variáveis consideradas foram nota por disciplina, histórico de reprovação e gênero, e os modelos obtiveram acurácia de 86,08% para Matemática e 80,77% para Português, demonstrando a viabilidade de ANN nesse contexto.

O Quadro 3 apresenta uma síntese dos principais estudos que utilizaram modelos de regressão para estimar o desempenho educacional em diferentes níveis de ensino. Importa destacar que os modelos listados correspondem àqueles que obtiveram os melhores resultados em cada estudo, embora todos os autores tenham avaliado múltiplas alternativas de modelagem. Observa-se uma tendência à aplicação de técnicas estatísticas e algoritmos mais avançados, como o GBM e o XGBoost, com incorporação de variáveis educacionais e socioeconômicas. No entanto, a maioria das investigações adota a escola ou o aluno como unidade de análise, com limitada exploração do município como recorte espacial.

Quadro 3 – Estudos com modelos de regressão aplicados à predição do desempenho educacional

Autores (ano)	Melhor Modelo	Nível de Ensino	Nível de Aplicação	Tipo de Variáveis	Acurácia
Maia, Bueno e Sato (2021)	GBM	Educação Básica	Município, região e país	Educacionais e socioeconômicas	$R^2 = 0,305$ a $0,459$
Soares e Santos (2024)	Regressão Linear Múltipla	Ensino Médio	Escola	Educacionais e socioeconômicas	$R^2 = 0,508$
Rocha, Teixeira e Melo (2015)	Regressão Linear Múltipla	Anos Finais	Escola	Educacionais e socioeconômicas	$R^2 = 0,468$
Wang e Luo (2024)	XGBoost	Ensino Superior	Aluno	Educacionais	$R^2 = 0,82$

Fonte: Elaborado pela autora (2025).

Ademais, a utilização de abordagens como a simulação de cenários ainda é pouco explorada, sendo identificada apenas no estudo de Freitas e Salgado (2020), cuja proposta

não contempla o IDEB como variável de interesse. Por fim, nota-se que os anos iniciais do ensino fundamental seguem com baixa representatividade nos estudos revisados, a despeito de sua relevância para o desenvolvimento das competências básicas de aprendizagem.

O Quadro 4 apresenta a sistematização de estudos que empregaram modelos de classificação no campo educacional, predominantemente voltados à predição de evasão, aprovação e categorização de desempenho estudantil. Embora muitos desses trabalhos reportem acurácias elevadas, observa-se que a maior parte adota como unidade de análise o estudante ou a escola, utilizando bases de dados de pequena escala e conjuntos de variáveis essencialmente educacionais.

Quadro 4 – Estudos com modelos de classificação aplicados à predição do desempenho educacional

Autores (ano)	Melhor Modelo	Nível de ensino	Nível de aplicação	Tipo de variáveis	Acurácia
Benevento e Meirelles (2023)	DT e GBM	Ensino Técnico	Escola	Educacionais	100%
Benevento (2023)	RF	Ensino Técnico	Escola	Educacionais	95,28%
Rodrigues et al. (2024)	XGBoost	Fund. e Médio	Aluno	Educacionais	AUC = 0,9856
Pinto, Freitas Jr. e Costa (2019)	OneR, LibSVM e J48	Anos Finais	Município	Educacionais e socioeconômicas	> 98%
Teodoro e Kappel (2020)	RF	Ensino Superior	Aluno	Educacionais e socioeconômicas	~80%
Freitas e Salgado (2020)	MLP	Ensino Superior	Aluno	Educacionais e socioeconômicas	88,5%
Silva (2022)	RF	Técnico Médio	Aluno	Educacionais e socioeconômicas	82,3%
Chen e Ding (2023)	ANN	Não especificado	Aluno	Educacionais e socioeconômicas	60%
Guanin-Fajardo, Guaña-Moya e Casillas (2024)	XGBoost	Ensino Superior	Aluno	Educacionais e socioeconômicas	Precisão = 79,49%
Pereira e Seruffo (2024)	MLP	Ensino Médio	Aluno	Educacionais	80,77% a 86,08%

Fonte: Elaborado pela autora (2025).

Essa delimitação metodológica reduz o potencial de generalização dos resultados e restringe a aplicabilidade dos modelos em políticas públicas com foco territorial ampliado. Evidencia-se, assim, a escassez de estudos que adotem o município como unidade analítica e que articulem, de maneira integrada, dimensões educacionais, socioeconômicas e estruturais na modelagem preditiva.

Apesar dos avanços na aplicação de técnicas de ML e DL para previsão de indicadores educacionais, a literatura ainda se concentra predominantemente na estimação de valores futuros a partir de padrões históricos. A incorporação sistemática de simulações contrafactuais como técnica analítica complementar permanece limitada. Embora a simulação de cenários seja amplamente utilizada em áreas como economia, planejamento urbano e saúde pública para subsidiar decisões baseadas em evidências (SCHWARTZ,

1991; CHERMACK, 2011), sua aplicação estruturada ao campo da modelagem preditiva educacional ainda é incipiente.

A simulação de cenários, amplamente empregada em áreas como economia, planejamento urbano e saúde pública para subsidiar decisões fundamentadas em evidências (SCHWARTZ, 1991; CHERMACK, 2011), permite explorar trajetórias alternativas por meio da manipulação controlada de variáveis relevantes. Diferentemente da previsão algorítmica tradicional (que projeta o comportamento futuro a partir de dados pretéritos) a simulação possibilita a avaliação de hipóteses contrafactuais, isto é, a estimativa dos efeitos potenciais decorrentes de alterações em políticas ou estratégias específicas.

No contexto educacional, essa abordagem permite examinar como mudanças em variáveis relacionadas à infraestrutura escolar, financiamento e valorização docente podem impactar indicadores de desempenho, como o IDEB e as taxas de rendimento.

Entretanto, a maioria dos estudos revisados restringe-se à previsão algorítmica, sem explorar parametrizações hipotéticas das variáveis explicativas para projetar efeitos associados a diferentes configurações de políticas públicas. Mesmo iniciativas que integram modelagem e experimentação hipotética, como Freitas e Salgado (2020), permanecem circunscritas a bases específicas e não contemplam indicadores agregados em escala municipal. Estudos como os de Maia, Bueno e Sato (2021), Chen e Ding (2023), Wang e Luo (2024) e Guanin-Fajardo, Guaña-Moya e Casillas (2024) demonstram a eficácia de modelos preditivos, mas não incorporam de forma estruturada a lógica da simulação contrafactual como instrumento complementar de análise.

Essa lacuna metodológica torna-se particularmente relevante em contextos marcados por desigualdades estruturais e restrições fiscais, como o do estado do Piauí. Segundo o IPEA (2021), as disparidades regionais comprometem a capacidade de financiamento dos entes subnacionais. Dados da Fundação Getúlio Vargas (2024) mostram que, em 2023, a arrecadação própria dos municípios nordestinos representava apenas 5,17% das receitas correntes, e 58% das prefeituras receberam notas C ou D no indicador de Capacidade de Pagamento (CAPAG). Como resposta, o governo do Piauí tem promovido iniciativas como o programa “Moderniza Seduc”, que até dezembro de 2024 havia investido R\$ 1,5 bilhão na reforma de 324 escolas, com climatização de salas, bibliotecas, laboratórios e quadras cobertas, beneficiando mais de 300 mil estudantes (VIAGORA, 2024; PIAUÍ, 2024).

A Figura 6 destaca as principais contribuições da simulação contrafactual para a análise e formulação de políticas públicas educacionais. Ao possibilitar a exploração de hipóteses, essa abordagem permite testar, de forma controlada, o impacto de intervenções antes de sua implementação real. Além disso, favorece a antecipação de efeitos, ao simular cenários alternativos que ajudam a compreender os possíveis resultados decorrentes de mudanças em variáveis-chave.

Adicionalmente, os Anos Iniciais do ensino fundamental seguem com baixa representatividade nos estudos revisados, apesar de sua relevância estratégica para a consolidação

Figura 6 – Vantagens da simulação contrafactual para o planejamento educacional

Fonte: Elaborado pela autora (2025).

das competências básicas de aprendizagem. Persistem, ainda, lacunas na análise integrada do impacto combinado de fatores educacionais, socioeconômicos e financeiros sobre o desempenho educacional agregado.

Em síntese, observa-se na literatura: (i) predominância de unidades de análise centradas no aluno ou na escola; (ii) limitada exploração do recorte municipal; (iii) escassa integração multidimensional de variáveis explicativas; e (iv) baixa incorporação de simulações contrafactuais associadas à modelagem preditiva.

Diante desse cenário, esta dissertação propõe o desenvolvimento de modelos preditivos para o IDEB nos 224 municípios piauienses, integrando dimensões educacionais, socioeconômicas e financeiras e incorporando a simulação de cenários como ferramenta analítica complementar. Busca-se, assim, ampliar o potencial explicativo das abordagens existentes e oferecer subsídios técnicos ao planejamento educacional orientado por evidências.

4 Metodologia

Este capítulo apresenta a metodologia adotada para o desenvolvimento da pesquisa, que se fundamenta em uma abordagem quantitativa, com ênfase na análise de dados secundários e na aplicação de algoritmos de ML e de DL. O objetivo central consiste na aplicação de modelos preditivos para estimar o IDEB nos municípios do estado do Piauí e na identificação das relações entre esse indicador e variáveis de natureza educacional, socioeconômica e financeira.

O delineamento metodológico contempla a análise e o tratamento de dados secundários provenientes de bases oficiais, seguido da construção de modelos preditivos e da realização de simulações de cenários exploratórios. Ressalta-se que, embora a base inicial de dados seja composta por um conjunto abrangente de variáveis educacionais, socioeconômicas e financeiras, nem todas essas variáveis foram mantidas nos modelos finais. A definição do subconjunto de variáveis explicativas utilizou critérios estatísticos e computacionais para eliminar redundâncias, variáveis derivadas diretamente do IDEB e atributos colineares, priorizando-se apenas os mais relevantes sob a ótica preditiva.

O percurso metodológico da pesquisa foi sistematizado em seis etapas complementares, conforme representado na Figura 7. A primeira etapa consistiu na coleta e integração das bases de dados, obtidas a partir de diferentes fontes oficiais. Em seguida, procedeu-se ao tratamento dos dados, com aplicação de rotinas de limpeza, padronização, unificação de formatos e exclusão de inconsistências. A terceira etapa compreendeu a seleção das variáveis explicativas, por meio de métodos estatísticos e computacionais, visando à redução da dimensionalidade e à eliminação de atributos redundantes ou colineares.

Na quarta etapa, os modelos preditivos foram construídos com base nos dados tratados e nas variáveis selecionadas, abrangendo algoritmos de ML e DL. No campo do ML, foram utilizados modelos como Regressão Linear Múltipla e as Regressões Penalizadas (Lasso, Ridge e Elastic Net), KNN, DT, RF, Gradiente Boosting e o algoritmo XGBoost. Para os modelos de DL, foram implementadas ANN do tipo MLP, CNNs e RNNs com unidades de memória de longo curto prazo (LSTM). Todos os modelos foram submetidos à validação cruzada do tipo *k-fold*.

Na quinta etapa, todos os modelos foram avaliados por meio das métricas MAE, RMSE e R^2 , que são apropriadas para problemas de regressão. Como os algoritmos utilizados nesta pesquisa foram aplicados em configurações regressivas para estimar valores contínuos do IDEB, essas métricas foram calculadas externamente a partir das previsões de cada modelo, independentemente de seus critérios internos de otimização. Dessa forma, modelos como RF e DT, embora não utilizem RMSE em seus processos de treinamento, também têm essa métrica apresentada por se tratar de uma medida de desempenho aplicada uniformemente a todos os algoritmos.

Além disso, foi realizada a análise da importância das variáveis explicativas por meio de abordagens interpretativas, incluindo técnicas baseadas em permutação e valores SHAP. Na sexta etapa, foram conduzidas simulações de cenários exploratórios com base no modelo preditivo que apresentou o melhor desempenho.

Essas simulações consistiram na manipulação controlada dos valores das variáveis mais influentes, previamente identificadas, com o propósito de examinar a sensibilidade dos resultados e estimar os efeitos potenciais de alterações hipotéticas nos fatores determinantes do IDEB. Tal abordagem possibilitou a análise condicional do comportamento do modelo diante de diferentes configurações de contexto, oferecendo subsídios para a formulação de políticas públicas baseadas em evidências.

Figura 7 – Etapas metodológicas da pesquisa em modelagem preditiva do IDEB



Fonte: Elaborado pela autora (2025).

Na sequência, apresenta-se detalhadamente as fontes de dados utilizadas, os procedimentos de coleta e tratamento das informações, a construção e validação dos modelos preditivos.

4.1 Fontes dos Dados

A base empírica desta pesquisa foi construída a partir da consolidação de dados provenientes de sistemas públicos e institucionais de reconhecida credibilidade, organizados em dois eixos: variáveis internas, relacionadas às condições das escolas e redes de ensino, e variáveis externas, associadas ao contexto socioeconômico e financeiro dos municípios. O estudo abrange os 224 municípios do estado do Piauí.

As variáveis internas foram extraídas dos microdados do INEP, notadamente do Censo Escolar, SAEB e IDEB, contemplando indicadores de infraestrutura escolar, formação e regularidade docente, gestão, acessibilidade, rendimento escolar, distorção idade-série e matrícula por etapa de ensino, com agregação municipal. As variáveis externas incluíram dados econômicos e fiscais, sistematizados pela Prof^a Dra. Rosana Evangelista da Cruz (UFPI), no âmbito do projeto “*Criação do Observatório Educacional do Piauí*”. As informações foram obtidas em bases oficiais como SIOPE, SICONFI/FINBRA, FUNDEB e relatórios orçamentários. Foram considerados indicadores de receitas próprias (IPTU, ISS, ITBI, IRRF), transferências constitucionais (FPM, ICMS, IPVA, entre outros), repasses federais vinculados à educação, componentes do FUNDEB e despesas com manutenção e desenvolvimento do ensino (MDE).

Complementarmente, foram incluídos indicadores socioeconômicos obtidos junto ao IBGE, por meio do Censo Demográfico, SIDRA e Contas Regionais. Foram considerados dados como população residente, Produto Interno Bruto (total, setorial e per capita), valor adicionado por atividade econômica, efetivo de rebanhos, área plantada ou destinada à colheita (lavouras temporárias e permanentes), valor da produção na extração vegetal, além de indicadores de assistência social. Estes últimos incluem informações sobre o Benefício de Prestação Continuada (BPC), a Renda Mensal Vitalícia (RMV) e os repasses via Fundos Municipais e Estaduais de Assistência Social, obtidos do sistema SAGICAD (Secretaria de Avaliação, Gestão da Informação e Cadastro Único).

Para assegurar consistência analítica e compatibilidade entre séries históricas, o recorte temporal foi delimitado aos anos ímpares entre 2013 e 2023, que correspondem aos ciclos bienais do SAEB e à divulgação oficial do IDEB. Essa escolha permitiu integrar os dados educacionais e contextuais em análises longitudinais coerentes com os objetivos da pesquisa.

Por fim, foram organizadas três bases analíticas distintas, cada uma referente a uma etapa da educação básica, com estrutura multivariada e longitudinal. A integração dos indicadores escolares e contextuais conferiu uma abordagem sistêmica à análise dos determinantes do desempenho educacional, fornecendo subsídios para a formulação de políticas públicas mais eficazes e fundamentadas em evidências.¹

4.2 Coleta e Integração dos Dados

A etapa de coleta e tratamento dos dados constitui um dos pilares metodológicos deste estudo, assegurando a consistência das bases consolidadas e a confiabilidade dos modelos preditivos e das simulações de cenários. Foram utilizados dados secundários extraídos de plataformas oficiais em formatos estruturados (CSV, XLSX e TXT), organizados em

¹ Os links das fontes referentes a todas as bases de dados utilizadas nesta pesquisa estão apresentados no Anexo A.

séries temporais compatíveis com os ciclos de divulgação do IDEB e das demais variáveis analisadas.

A estrutura analítica foi segmentada em três bases principais, conforme cada etapa da educação básica. Para os anos iniciais e finais do ensino fundamental, foram utilizados os anos ímpares entre 2013 e 2023; já para o ensino médio, foram considerados os anos ímpares entre 2017 e 2023, única série histórica disponível para o IDEB nessa etapa. A unificação das bases utilizou os códigos do IBGE como identificadores únicos, assegurando o pareamento adequado das informações por município, etapa de ensino e ano.

O processo de integração, agregação e transformação dos dados foi automatizado por meio de scripts desenvolvidos em linguagem Python, com o uso de bibliotecas² como *Pandas*, *NumPy* e *Scikit-learn*, assegurando reprodutibilidade, transparência e controle rigoroso sobre cada etapa do processo.

No caso específico dos microdados do Censo Escolar, originalmente organizados por unidade escolar, foi necessário agregá-los à unidade de análise municipal. As variáveis quantitativas, como número de matrículas, salas, equipamentos e docentes, foram somadas aritmeticamente em cada agrupamento (município, etapa e ano). Para variáveis binárias, representando a presença ou ausência de recursos (como água potável, energia elétrica, biblioteca, laboratório, quadra, acessibilidade, conselho escolar, entre outros), aplicou-se também a soma, refletindo a quantidade de escolas que dispunham de determinado recurso. Por fim, foi criada a variável “quantidade de escolas”, que contabiliza o número total de escolas públicas em cada agrupamento.

4.3 Tratamento dos Dados Ausentes e Estimativas

Para variáveis sem atualização disponível para o ano de 2023, foram adotadas estratégias de estimativa. A estimativa da população foi realizada com base na metodologia de projeção geométrica de curto prazo, amplamente utilizada em estudos demográficos. A taxa de crescimento foi calculada a partir da variação populacional entre 2021 e 2022. A estimativa para 2023 foi obtida aplicando-se essa taxa ao valor da população de 2022, segundo a fórmula:

$$P_{2023} = P_{2022}(1 + r) \quad (6)$$

Em que a taxa é calculada a partir da razão entre a variação absoluta da população no período e o valor da população no ano base, conforme a seguinte expressão:

$$r = \frac{P_{2022} - P_{2021}}{P_{2021}} \quad (7)$$

² Documentação oficial das bibliotecas utilizadas: Pandas — <https://pandas.pydata.org/>; NumPy — <https://numpy.org/>; Scikit-learn — <https://scikit-learn.org/>

Esse procedimento foi implementado diretamente em planilhas eletrônicas, com base nas abordagens descritas por Barclay (1958), Miller e Blair (2009) e Creswell (2014).

Para estimar os valores do PIB municipal no ano de 2023, foi aplicada regressão linear simples sobre séries históricas de 2010 a 2021, utilizando o ano como variável independente e o valor econômico como variável dependente. A modelagem ocorreu em duas etapas: primeiro, foi estimado o valor de 2022, que foi incorporado à série para o ajuste final da projeção de 2023. Esse procedimento foi replicado para os componentes do PIB, como o Valor Adicionado Bruto (VAB) da agropecuária, indústria, serviços e administração pública, os impostos líquidos de subsídios, o PIB total e o PIB per capita.

4.4 Ferramentas e Softwares

As etapas analíticas desta pesquisa foram conduzidas com um conjunto integrado de ferramentas computacionais, selecionadas por sua ampla aceitação acadêmica, robustez técnica e adequação às demandas específicas da modelagem preditiva em contextos educacionais. A linguagem de programação adotada foi o Python, destacada por sua sintaxe acessível, modularidade e extenso ecossistema de bibliotecas voltadas à ciência de dados, ML e DL.

O Quadro 5 apresenta os algoritmos de ML e DL utilizados no estudo, acompanhados dos principais hiperparâmetros configurados para cada modelo. No grupo de ML, exploraram-se técnicas lineares, penalizadas, baseadas em instâncias, árvores de decisão e métodos de *ensemble*. Para DL, foram implementadas sete arquiteturas de redes neurais, combinando camadas densas, convolucionais e recorrentes. As redes neurais foram treinadas com a função de perda *mean squared error* (MSE) e com a métrica adicional MAE.

Para o desenvolvimento e avaliação dos modelos de ML, foi utilizada a biblioteca *Scikit-learn*, que oferece implementações consolidadas de algoritmos supervisionados e não supervisionados, além de funcionalidades para pré-processamento de dados, técnicas de validação cruzada (*cross-validation*), ajuste de hiperparâmetros e cálculo de métricas de desempenho preditivo.

No âmbito do DL, foram empregadas as bibliotecas³ *Keras* e *TensorFlow*, adotadas para a construção de arquiteturas de ANN, tais como o MLP, CNN, RNNs do tipo LSTM, e modelos híbridos CNN+LSTM. Essas bibliotecas permitem alto grau de personalização na definição das camadas, funções de ativação, otimizadores e critérios de parada antecipada, viabilizando o treinamento eficiente mesmo em bases de dados mais complexas.

As tarefas de manipulação, padronização e transformação dos dados foram conduzidas com o apoio das bibliotecas *Pandas* e *NumPy*, as quais proporcionam estruturas de dados otimizadas, como *DataFrames* e *arrays* multidimensionais, indispensáveis para operações vetorizadas e engenharia de atributos. A etapa de exploração e visualização dos

³ Keras — <https://keras.io/>; TensorFlow — <https://www.tensorflow.org/>

Quadro 5 – Principais hiperparâmetros dos algoritmos de ML e DL aplicados neste estudo

Grupo	Modelo / Arquitetura	Principais Hiperparâmetros
ML	Regressão Linear Múltipla	—
	Lasso	$\alpha = 0,1$
	Ridge	$\alpha = 1,0$
	Elastic Net	$\alpha = 0,1$; $l1_ratio = 0,5$
	KNN	$n_neighbors = 7$
	SVR	kernel = 'rbf'; $C = 1,0$; $\epsilon = 0,1$
	DT	$max_depth = 3$; $min_samples_split = 15$; $min_samples_leaf = 10$
	RF	$n_estimators = 300$; $max_depth = 4$; $min_samples_split = 10$; $min_samples_leaf = 10$; $max_features = 'sqrt'$
	GBM	$n_estimators = 200$; $learning_rate = 0,01$; $max_depth = 3$; $subsample = 0,8$; $min_samples_leaf = 5$
DL	XGBoost	$n_estimators = 1000$; $learning_rate = 0,01$; $max_depth = 2$; $subsample = 0,8$; $colsample_bytree = 0,8$; $reg_alpha = 1,0$; $reg_lambda = 10,0$
	MLP	64–32 neurônios; ativação ReLU nas camadas ocultas; saída linear; otimizador Adam; $learning_rate = 0,001$
	MLP com Dropout	128–64 neurônios; $dropout = 0,3$ após cada camada oculta; ativação ReLU; saída linear; Adam; $learning_rate = 0,001$
	MLP com Batch Normalization	128–64 neurônios; BatchNorm após cada camada oculta; ativação ReLU; saída linear; Adam; $learning_rate = 0,001$
	Deep MLP	256–128–64–32 neurônios; $dropout = 0,4$ e $0,3$ em camadas intermediárias; ativação ReLU; saída linear; Adam; $learning_rate = 0,0005$
	CNN	Conv1D: filters = 32; $kernel_size = 2$; MaxPooling1D; camada densa final com 50 neurônios e ativação ReLU; saída linear; Adam; $learning_rate = 0,001$
	LSTM	50 unidades LSTM; ativação tanh; saída linear; Adam; $learning_rate = 0,001$
DL	CNN+LSTM	Conv1D: 64 filtros; $kernel_size = 2$; MaxPooling1D; 50 unidades LSTM; ativação tanh; saída linear; Adam; $learning_rate = 0,001$

Fonte: Elaborado pela autora (2025).

dados foi realizada com as bibliotecas⁴ *Matplotlib* e *Seaborn*, utilizadas para a criação de gráficos estatísticos, histogramas, mapas de correlação, curvas de aprendizado e outras representações visuais que apoiam tanto a compreensão dos dados quanto a comunicação clara dos resultados.

Na fase preliminar de inspeção dos dados brutos, o *Microsoft Excel* foi utilizado para conferência manual, verificação de integridade e organização inicial das bases oriundas de diferentes fontes. Por fim, os scripts analíticos foram desenvolvidos no *Google Colab*⁵, para treinamento em nuvem de ANN, e no *Visual Studio Code*⁶, para edição local, organização e versionamento, garantindo controle metodológico e reprodutibilidade.

4.5 Estratégia de Modelagem Preditiva por Etapas

Com o propósito de identificar um subconjunto parcimonioso de variáveis que maximizasse o desempenho preditivo dos modelos, partiu-se de uma base composta por 128 atributos originais, organizados para os três recortes deste estudo: anos iniciais, anos finais do ensino fundamental e ensino médio. Desse total, três atributos correspondiam a identificadores administrativos (ano, código e nome do município), utilizados apenas para referência e, portanto, excluídos das etapas de modelagem.

Também foram removidas cinco variáveis que integram diretamente o cálculo do IDEB (taxa de aprovação, indicador de rendimento (P), nota SAEB em matemática, nota SAEB em língua portuguesa e nota SAEB – nota média padronizada (N)), pois sua inclusão poderia gerar *data leakage* no processo de treinamento. Por fim, a variável IDEB foi mantida como representação direta do objeto de interesse, constituindo a variável-alvo a ser prevista pelos modelos em cada etapa de ensino.

Assim, o conjunto efetivamente disponível para análise reduziu-se a 119 variáveis explicativas, englobando dimensões educacionais, socioeconômicas, financeiras. Contudo, essas variáveis apresentavam elevados níveis de correlação entre si, o que implicava riscos de multicolinearidade e redundância informacional. Um exemplo típico é a forte associação entre taxa de aprovação, taxa de reprovação e taxa de abandono, que descrevem diferentes facetas de um mesmo fenômeno — o fluxo escolar. Esse cenário reforçou a necessidade de aplicar métodos sistemáticos de seleção de variáveis, capazes de identificar aquelas mais relevantes e eliminar redundâncias, garantindo maior robustez estatística e interpretabilidade aos modelos de ML e DL testados.

O processo principal foi organizado em sete etapas. A Etapa 1 consistiu na utilização de todas as variáveis disponíveis, com exceção de “Nota SAEB em Matemática” e “Nota SAEB em Língua Portuguesa”, que apresentavam 282 observações faltantes e cuja informação já estava sintetizada na variável agregada “Nota SAEB – Nota Média

⁴ Matplotlib — <https://matplotlib.org/>; Seaborn — <https://seaborn.pydata.org/>

⁵ Google Colab: <https://colab.research.google.com/>

⁶ Visual Studio Code: <https://code.visualstudio.com/>

Padronizada (N)” e sem os identificadores (código e nome dos municípios e ano). Essa etapa teve como propósito servir de linha de base, utilizando o maior número possível de informações, ainda que com elevada redundância e potencial *data leakage*. Na Etapa 2, foram excluídas todas as variáveis diretamente relacionadas ao cálculo do IDEB, como “Nota SAEB em Matemática”, “Nota SAEB em Língua Portuguesa”, “Nota SAEB – Nota Média Padronizada (N)”, “Indicador de Rendimento (P)” e “Taxa de Aprovação”, bem como variáveis de identificação. Essa configuração visou reduzir o sobreajuste e eliminar a influência direta da variável dependente sobre os resultados.

Em seguida, foi implementado um *pipeline* de seleção de variáveis que subsidiou as etapas seguintes. Esse *pipeline* combinou quatro técnicas complementares: VIF, para diagnóstico e exclusão iterativa de variáveis com alta colinearidade ($VIF > 10$); regressão LassoCV, com penalização L_1 para induzir esparsidade no vetor de coeficientes e ajuste automático do parâmetro por validação cruzada; valores SHAP, para mensurar a contribuição marginal das variáveis e selecionar aquelas com impacto médio superior a 0,01; e algoritmo Boruta, fundamentado em RF, que retém apenas variáveis estatisticamente superiores a suas contrapartes embaralhadas (*shadow features*). Com base nesse *pipeline*, definiram-se as Etapas 3 a 6.

Na Etapa 3, os modelos foram treinados com as variáveis selecionadas exclusivamente pelo VIF; na Etapa 4, com as variáveis indicadas pela LassoCV; na Etapa 5, com aquelas destacadas pelos valores SHAP; e na Etapa 6, com as variáveis selecionadas pelo algoritmo Boruta. Por fim, a Etapa 7 utilizou as variáveis pertencentes à interseção das quatro técnicas, que coincidiram com as selecionadas na Etapa 6, evidenciando a robustez desse conjunto.

Os resultados obtidos ao longo das etapas indicaram que, nos anos iniciais e finais do ensino fundamental, os modelos de ML apresentaram, na Etapa 1, métricas elevadas ($R^2 > 0,98$; RMSE e MAE $< 0,06$), inflacionadas pelo *data leakage*, em razão da inclusão de variáveis que compõem o cálculo do IDEB, o que configurou um caso clássico de sobreajuste estrutural. Os modelos de DL, mesmo com essas variáveis, apresentaram desempenho inferior, possivelmente devido à alta colinearidade e à menor capacidade do DL para lidar com dados tabulares redundantes. Na Etapa 2, com a remoção das variáveis do IDEB, houve queda acentuada do desempenho para ambos os paradigmas, com aumento da discrepância entre treino e teste e persistência de multicolinearidade.

Já as Etapas 3 a 5 resultaram em desempenhos insatisfatórios, com acurácia reduzida e altos erros preditivos, sugerindo que os subconjuntos isolados gerados por VIF, LassoCV e SHAP eram excessivamente restritivos. Em contrapartida, as Etapas 6 e 7 produziram os melhores resultados para os anos iniciais e finais, sobretudo para os modelos de ML, que apresentaram maior estabilidade e capacidade preditiva. Os modelos de DL também alcançaram seu melhor desempenho nessas etapas, embora ainda inferiores aos de ML.

No caso do ensino médio, cujo tamanho amostral era menor, nenhuma das sete

etapas iniciais produziu resultados satisfatórios quando comparados aos obtidos nas etapas do ensino fundamental. Em face dessa limitação, foi adotado um procedimento adicional, que consistiu, primeiramente, na remoção das variáveis financeiras, uma vez que nenhuma das técnicas de seleção previamente aplicadas as indicou como relevantes. Em seguida, procedeu-se à exclusão das variáveis com coeficiente de correlação superior a 0,8, valor estipulado para reduzir redundâncias sem comprometer informações substantivas. Essa estratégia resultou em um conjunto final de 17 variáveis, que proporcionou um desempenho superior para os modelos dessa etapa, tanto para ML quanto para DL, em relação às demais configurações testadas.

Somente após a conclusão do processo completo foi possível identificar, para os anos iniciais e finais do ensino fundamental, um conjunto final de seis variáveis, selecionadas de forma convergente pelas técnicas do *pipeline*. As variáveis selecionadas foram: PIB per capita; taxa de distorção idade-série; percentual de docentes sem formação superior (classificados no Grupo 5 de adequação da formação); área plantada ou destinada à colheita (lavouras temporárias e permanentes); valor da produção na extração vegetal; e valor repassado por meio do programa Criança Feliz 1⁷. Esse conjunto sintetiza dimensões socioeconômicas, estruturais e pedagógicas associadas ao desempenho educacional.

Para o ensino médio, o conjunto final de 17 variáveis contemplou aspectos educacionais relacionados à estrutura física e tecnológica da escola, qualificação e condições dos professores, organização e gestão escolar, indicadores de fluxo e rendimento escolar e variáveis socioeconômicas, como PIB, área plantada, rebanhos e valores repassados no âmbito da Proteção Social Básica. O Quadro 6 apresenta um resumo das etapas, técnicas aplicadas e principais resultados para ML e DL.

A Figura 8 sintetiza o processo de seleção de variáveis conduzido em sete etapas sequenciais, aplicado a todas as etapas de ensino (anos iniciais, anos finais e ensino médio). Após as exclusões preliminares (Etapas 1 e 2), o *pipeline* envolveu: (i) diagnóstico de multicolinearidade pelo VIF; (ii) regressão LassoCV; (iii) análise de valores SHAP; (iv) algoritmo Boruta; e (v) interseção das técnicas (Etapa 7). Em todas as etapas também foi verificada a influência da correlação entre variáveis, de modo a reduzir redundâncias.

Os resultados evidenciaram que, para o ensino fundamental, o *pipeline* completo (Etapas 1–7), detalhado no Quadro 6, apresentou o melhor desempenho, culminando em um conjunto final de seis variáveis explicativas. No ensino médio, por sua vez, as sete etapas iniciais não resultaram em configurações satisfatórias, em razão do menor tamanho amostral e da maior redundância entre atributos. Nesse caso, a solução mais robusta foi obtida pela exclusão manual de variáveis com correlação superior a 0,8, o que resultou em um conjunto final de 17 variáveis explicativas.

Em síntese, os resultados mostraram que, para o ensino fundamental, o conjunto

⁷ O Programa Criança Feliz foi instituído pelo Decreto nº 9.579/2018, com a finalidade de promover o desenvolvimento integral na primeira infância por meio de visitas domiciliares e ações intersetoriais de apoio a gestantes, crianças de até seis anos e suas famílias (BRASIL, 2018).

de variáveis selecionado por meio do *pipeline* permitiu aos modelos alcançar desempenho satisfatório, conciliando parcimônia e capacidade preditiva. Para o ensino médio, por sua vez, a solução mais adequada foi obtida por meio de ajustes manuais e da exclusão de variáveis altamente correlacionadas, em razão das limitações impostas pelo menor tamanho amostral. Ademais, verificou-se que as variáveis financeiras, aqui utilizadas, embora relevantes do ponto de vista orçamentário, não apresentaram poder explicativo significativo para a variação do IDEB nos municípios piauienses.

Quadro 6 – Etapas da modelagem preditiva e estratégias de seleção de variáveis nos modelos (continua)

Etapas	Descrição e Técnicas Aplicadas	Principais Resultados (ML e DL)
Etapa 1	Modelo inicial com 122 variáveis, excluindo apenas “Nota SAEB em Matemática” e “Nota SAEB em Língua Portuguesa” e identificadores (ano, nome e código do município).	ML: desempenho inflacionado por <i>data leakage</i> ($R^2 > 0,98$; RMSE e MAE $< 0,06$). DL: baixo desempenho ($R^2 < 0,01$), prejudicado pela alta colinearidade das variáveis.
Etapa 2	Exclusão das variáveis que compõem diretamente o IDEB: “Nota SAEB em Matemática”, “Nota SAEB em Língua Portuguesa”, “Nota SAEB – Nota Média Padronizada (N)”, “Indicador de Rendimento”, “Taxa de Aprovação” e identificadores, resultando em 119 variáveis.	ML: queda no desempenho, alta discrepância treino–teste, com o melhor desempenho entre as etapas de ensino, no treino $R^2 = 0,7222$, MAE = 0,2381. No teste $R^2 = 0,4832$ e MAE = 0,3437, e na validação cruzada $R^2 = 0,2352$. DL: desempenho inferior aos de ML, no treino $R^2 = 0,6101$ e no teste $R^2 = 0,4154$, e na validação cruzada $R^2 = 0,3827$.
Etapa 3	Aplicação do VIF, com remoção iterativa (<i>stepwise</i>) de variáveis com VIF > 10 , reduzindo o conjunto da etapa anterior em 38 variáveis.	ML: treino $R^2 = 0,6735$, MAE = 0,4138; teste $R^2 = 0,5574$, MAE = 0,4987; validação cruzada $R^2 = 0,2995$. DL: treino $R^2 = 0,6414$, teste $R^2 = 0,4737$, validação cruzada $R^2 = 0,3554$.
Etapa 4	Redução por meio da regressão LassoCV (penalização L_1), ajustado por 5-fold CV, reduzindo o conjunto de variáveis em 15.	ML: treino $R^2 = 0,6593$, MAE = 0,4221; teste $R^2 = 0,5669$, MAE = 0,4973; validação cruzada $R^2 = 0,3215$. DL: treino $R^2 = 0,6598$, teste $R^2 = 0,3624$, validação cruzada $R^2 = 0,3777$.
Etapa 5	Seleção baseada em valores SHAP, mantendo apenas variáveis com impacto médio $> 0,01$ (12 variáveis).	ML: treino $R^2 = 0,6575$, MAE = 0,4237; teste $R^2 = 0,5663$, MAE = 0,4977; validação cruzada $R^2 = 0,3234$. DL: treino $R^2 = 0,6756$, teste $R^2 = 0,5195$, validação cruzada $R^2 = 0,4118$.

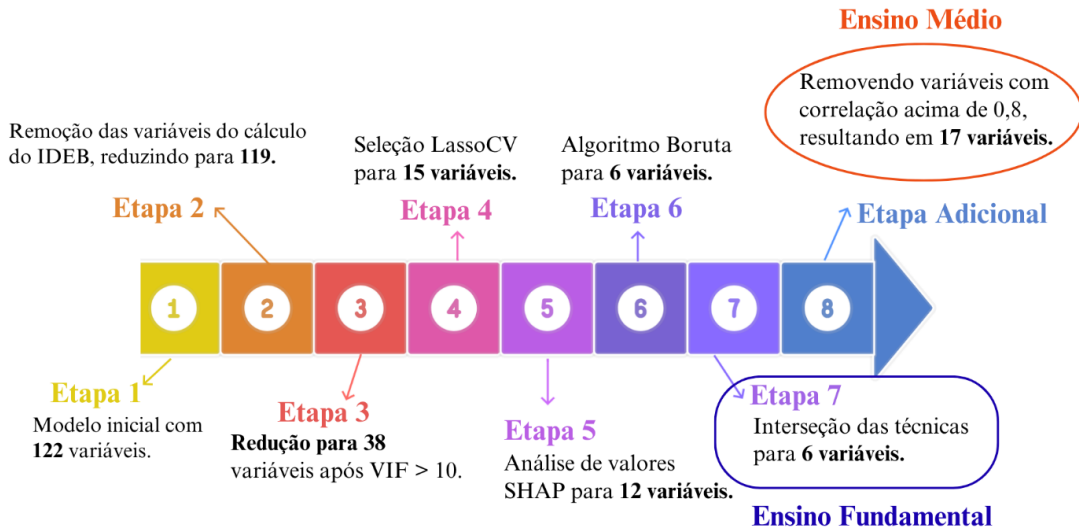
Fonte: Elaborado pela autora (2025).

Quadro 6 – continuação: Etapas da modelagem preditiva e estratégias de seleção de variáveis nos modelos (conclusão)

Etapas	Descrição e Técnicas Aplicadas	Principais Resultados (ML e DL)
Etapa 6	Aplicação do algoritmo Boruta, baseado em RF e <i>shadow features</i> , identificando variáveis estatisticamente relevantes (6 variáveis).	ML: melhores resultados. Treino $R^2 = 0,6385$, MAE = 0,4354; teste $R^2 = 0,5542$, MAE = 0,5050; validação $R^2 = 0,5455$. DL: treino $R^2 = 0,5922$, teste $R^2 = 0,4570$, validação $R^2 = 0,5288$.
Etapa 7	Interseção das quatro técnicas de seleção (VIF, Lasso, SHAP e Boruta). Conjunto final idêntico à Etapa 6.	ML e DL: equivalentes à Etapa 6.
Etapa adicional (Ensino Médio)	Exclusão manual de variáveis com correlação $> 0,8$, resultando em 17 variáveis específicas para o Ensino Médio, já que com as etapas anteriores os modelos não obtiveram um bom desempenho em comparação com essa etapa.	ML: melhor desempenho em comparação às sete etapas anteriores. Treino $R^2 = 0,6272$, MAE = 0,2754; teste $R^2 = 0,4186$, MAE = 0,3665; validação $R^2 = 0,4488$. DL: treino $R^2 = 0,4508$, teste $R^2 = 0,2683$, validação $R^2 = 0,2747$.

Fonte: Elaborado pela autora (2025).

Figura 8 – Resumo do processo de seleção de variáveis e configuração final por etapa de ensino



Fonte: Elaborado pela autora (2025).

4.6 Modelagem Preditiva com ML e DL

A modelagem preditiva constituiu o núcleo analítico deste estudo, com o objetivo de estimar, de forma contínua e quantitativa, o IDEB a partir de um conjunto de atributos educacionais, socioeconômicos e financeiros. Trata-se de uma tarefa de regressão supervi-

onada, que busca modelar a relação funcional entre variáveis independentes e a variável dependente (IDEB), expressa em escala contínua. Para isso, foram utilizados algoritmos de ML e DL, ambos explorados sob abordagens regressivas, a fim de capturar padrões lineares e não lineares explicativos do desempenho educacional nos municípios piauienses. A escolha por modelos regressivos fundamenta-se na natureza contínua da variável-alvo e na comprovada eficácia dessas técnicas em fenômenos complexos, com múltiplos fatores interdependentes.

Com vistas à robustez estatística e à capacidade de generalização, adotou-se estratégias rigorosas de avaliação, adaptadas às especificidades de cada paradigma. Para os modelos de ML, a base foi particionada em 70% para treino e 30% para teste, com estratificação por IDEB para manter a distribuição dos escores. Em seguida, aplicou-se validação cruzada *k-fold* ($k=5$) no conjunto de treino, com rotações sucessivas entre *folds* para estimar a variabilidade preditiva e reduzir o risco de *overfitting*.

Para os modelos de DL, devido à maior complexidade computacional, utilizou-se divisão em 80% para treino e 20% para teste, igualmente estratificada, com validação cruzada implementada sobre o conjunto de treino por meio da classe *KFold*. Cada instância da rede era inicializada do zero a cada iteração, assegurando consistência e evitando memorização indesejada. Em ambos os paradigmas, as métricas de avaliação incluíram MAE, RMSE e R^2 , calculadas separadamente para os conjuntos de treino, teste e validação cruzada.

Foram inicialmente aplicados modelos lineares, como a Regressão Linear Múltipla e as regressões penalizadas Lasso, Ridge e Elastic Net, destacando-se pela capacidade de selecionar variáveis relevantes, reduzir colinearidades e permitir a interpretação direta dos coeficientes. Além das métricas preditivas, verificou-se os pressupostos estatísticos clássicos (linearidade, homocedasticidade, independência, normalidade dos resíduos) e a ausência de multicolinearidade (VIF). Em seguida, aplicou-se modelos baseados em árvores de decisão, como o DT e o RF, capazes de capturar relações não lineares e interações entre variáveis. Esses modelos oferecem robustez frente à variabilidade dos dados e fornecem medidas interpretáveis de importância relativa das variáveis, acessadas por meio do atributo *feature_importances_*.

Modelos de *boosting*, como o GBM e o XGBoost, também foram utilizados, dada sua reconhecida capacidade de produzir previsões precisas por meio da agregação sequencial de múltiplas árvores com minimização de erro residual. Para algoritmos que não possuem interpretabilidade direta dos parâmetros, como o KNN e o Regressor de Vetores de Suporte (Support Vector Regressor - SVR), empregou-se a técnica de importância por permutação (Permutation Importance). Esse procedimento consiste em embaralhar os valores de uma variável por vez e observar a queda no R^2 , interpretando-se essa degradação como uma estimativa do impacto da variável na acurácia da predição.

No âmbito do DL, foram exploradas arquiteturas profundas do tipo MLP, bem

como CNN, RNN, LSTM e modelos híbridos (CNN+LSTM). A análise de importância das variáveis foi estimada com uma do método de permutação, calculando-se a diferença entre o desempenho original (R^2) e o desempenho após permutação de cada variável de entrada, uma a uma. Adicionalmente, foram utilizadas curvas de aprendizado (função de perda ao longo das épocas) para monitorar sinais de sobreajuste ou subajuste. Para mitigar o *overfitting*, aplicou-se o critério de parada antecipada (*early stopping*), interrompendo o treinamento quando não houvesse melhora na perda de validação após número predefinido de épocas consecutivas.

5 Resultados e Discussão

Neste capítulo são apresentados e discutidos os resultados obtidos a partir da aplicação dos modelos preditivos desenvolvidos. Inicialmente, realiza-se uma análise descritiva das bases de dados, com o objetivo de caracterizar o conjunto de variáveis utilizadas e oferecer uma visão geral sobre o contexto educacional analisado. Em seguida, são expostos os desempenhos dos modelos de ML e DL, considerando as métricas de avaliação estabelecidas na metodologia.

A análise contempla tanto os resultados individuais de cada modelo quanto a comparação entre eles, de forma a identificar padrões de desempenho e destacar as abordagens mais adequadas para a previsão do IDEB nos diferentes níveis da educação básica. Posteriormente, são discutidas as simulações de cenários exploratórios, que permitem avaliar os impactos potenciais de alterações nas variáveis explicativas sobre o IDEB. Por fim, os resultados são confrontados com as hipóteses e questões de pesquisa definidas na introdução, bem como articulados com a literatura revisada nos capítulos anteriores, de modo a oferecer uma interpretação dos achados e evidenciar suas contribuições para a compreensão do fenômeno investigado.

5.1 Análise Descritiva dos Dados

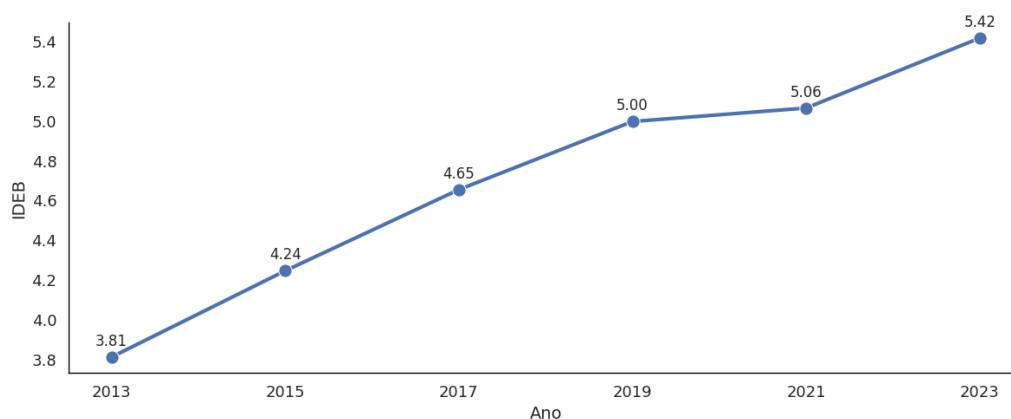
A construção das bases de dados utilizadas neste estudo envolveu etapas de limpeza, imputação e seleção de variáveis, a fim de garantir consistência analítica e confiabilidade estatística dos modelos preditivos. Inicialmente, a base referente aos anos iniciais do ensino fundamental apresentava 1.344 registros e 128 variáveis. Foram removidas variáveis com elevada proporção de valores ausentes, como “Nota SAEB em Matemática” e “Nota SAEB em Língua Portuguesa”, uma vez que essas informações já estavam representadas de forma agregada na variável “Nota SAEB – Nota Média Padronizada (N)”. Além disso, observações com valores ausentes na variável dependente (IDEB) e Nota SAEB padronizada também foram excluídas. Após esse procedimento, a base final foi reduzida para 1.262 registros e 126 variáveis.

A mesma lógica foi aplicada à base dos anos finais do ensino fundamental, que passou a conter 1.271 registros e 126 variáveis, incluindo identificadores como ano, nome e código dos municípios. Observou-se ainda que algumas variáveis explicativas apresentavam ausências residuais (entre 0,08% e 2,05%). Considerando a presença de distribuições assimétricas e outliers, optou-se pela imputação da mediana, realizada de forma individual por variável, por ser uma medida robusta e menos suscetível à distorção causada por valores extremos.

Já a base do ensino médio apresentava inicialmente 896 registros, mas 151 observa-

ções possuíam valores ausentes na variável dependente, sendo, portanto, removidas. Após a imputação da mediana nas variáveis explicativas com dados faltantes, a base final foi composta por 745 registros e 126 colunas. Importante destacar que, diferentemente das etapas anteriores, o IDEB do ensino médio é divulgado apenas a partir do ano de 2017, o que restringiu o recorte temporal da análise ao período 2017–2023.

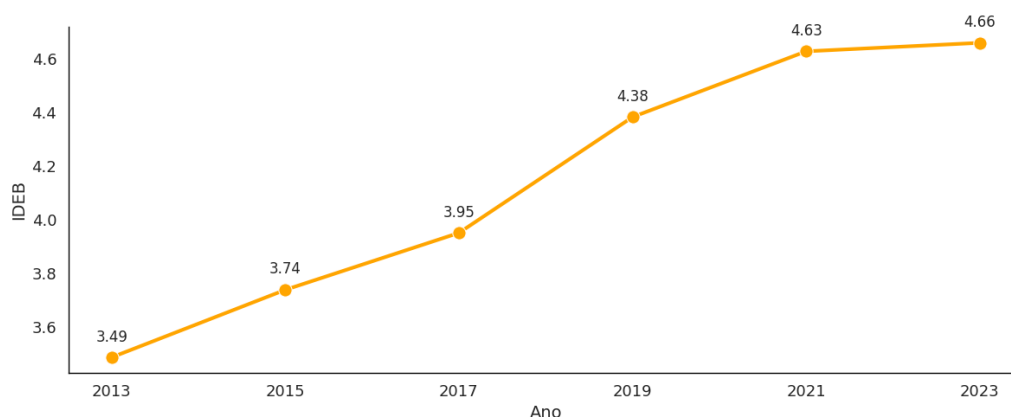
Figura 9 – Evolução histórica do IDEB médio piauiense nos anos iniciais do ensino fundamental (2013–2023)



Fonte: Elaborado pela autora (2025).

A Figura 9 acima evidencia uma trajetória ascendente contínua no desempenho médio dos anos iniciais ao longo da série histórica. O IDEB médio cresceu de 3,81 em 2013 para 5,42 em 2023, revelando avanços expressivos nesta etapa. A evolução é praticamente linear, com destaque para o salto entre 2015 e 2019, indicando possíveis efeitos de políticas estruturantes de alfabetização e formação docente no período.

Figura 10 – Evolução histórica do IDEB médio piauiense nos anos finais do ensino fundamental (2013–2023)

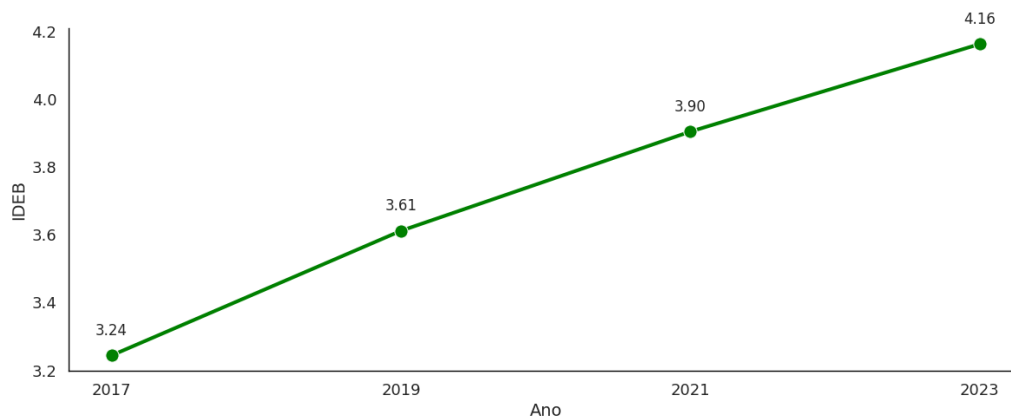


Fonte: Elaborado pela autora (2025).

A Figura 10 evidencia que o IDEB médio dos anos finais do ensino fundamental cresceu de forma contínua ao longo da última década, passando de 3,49 em 2013 para 4,66 em 2023. Entretanto, enquanto o período de 2017 a 2021 registrou um salto expressivo (de 3,95 para 4,63), indicando avanços significativos, o intervalo entre 2021 e 2023 apresentou

variação mínima (de 4,63 para 4,66). Essa desaceleração pode estar associada a desafios estruturais da etapa, como a maior complexidade dos conteúdos e a ampliação das desigualdades de aprendizagem acumuladas nos anos iniciais. Soma-se a isso o impacto residual da pandemia de Covid-19, que provocou longos períodos de suspensão das aulas presenciais e acentuou dificuldades de continuidade escolar, afetando a implementação de políticas eficazes para a fase final do ensino fundamental.

Figura 11 – Evolução histórica do IDEB médio piauiense no ensino médio (2017–2023)



Fonte: Elaborado pela autora (2025).

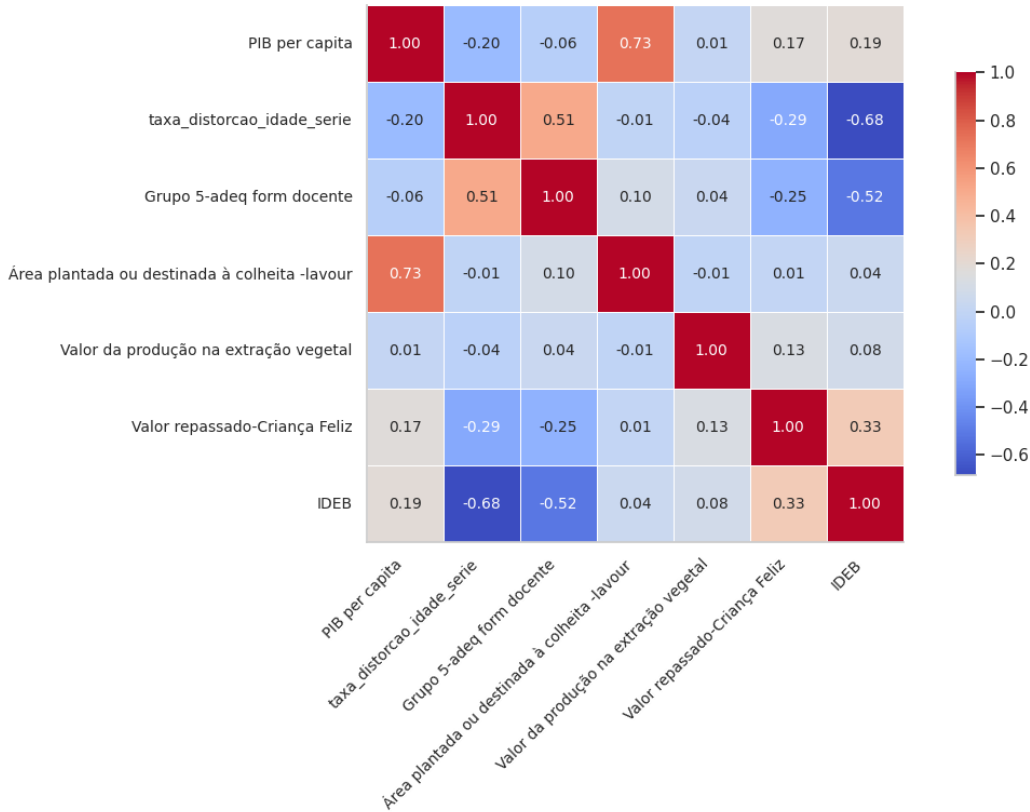
A Figura 11 apresenta um panorama distinto. O IDEB do ensino médio cresceu de 3,24 em 2017 para 4,16 em 2023, com aceleração visível a partir de 2019. Este comportamento pode estar associado a mudanças no modelo de avaliação, expansão de tempo integral e reforço no acompanhamento pedagógico. Ainda assim, o patamar geral continua inferior ao das demais etapas, evidenciando que o ensino médio permanece como o maior desafio da política educacional piauiense.

A Figura 12 abaixo, referente aos anos iniciais, revela correlações estatisticamente significativas entre fatores educacionais, socioeconômicos e estruturais. Destaca-se a forte correlação negativa da distorção idade-série ($r = -0,68$) e da inadequação na formação docente ($r = -0,52$) com o IDEB, sugerindo impacto negativo do atraso escolar e da baixa qualificação docente. Em contrapartida, o PIB per capita ($r = 0,19$) e os repasses do programa Criança Feliz ($r = 0,33$) apresentam correlação positiva, ainda que modesta, indicando que melhores condições econômicas e políticas sociais favorecem o desempenho escolar. Já a área plantada ou destinada à colheita das lavouras temporárias e permanentes ($r = 0,04$) e o valor da produção na extração vegetal ($r = 0,08$) revelam correlações próximas de zero, sinalizando que o impacto direto desses fatores sobre o IDEB é pouco relevante no contexto analisado.

Em conjunto, esses resultados reforçam a relevância de políticas integradas que atuem tanto sobre fatores internos à escola, como a redução da distorção idade-série e a valorização docente, quanto sobre dimensões externas de caráter socioeconômico, destacando o papel das políticas de assistência social e de desenvolvimento econômico no

fortalecimento do desempenho educacional.

Figura 12 – Mapa de correlação entre variáveis selecionadas e o IDEB médio piauiense dos anos iniciais do ensino fundamental



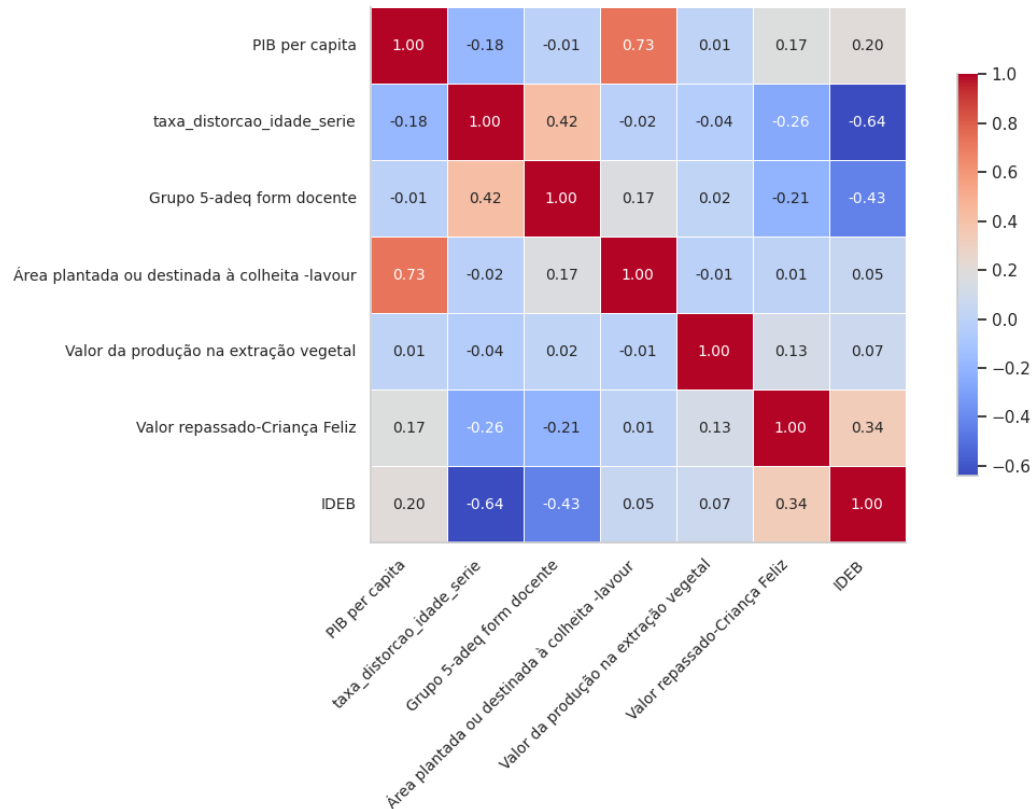
Fonte: Elaborado pela autora (2025).

A Figura 13, referente aos anos finais do ensino fundamental, evidencia a predominância de correlações negativas entre o IDEB e fatores internos da rede, especialmente a distorção idade-série ($r = -0,64$) e o percentual de docentes sem formação superior ($r = -0,43$), confirmando que atraso escolar e baixa qualificação docente prejudicam o desempenho. Em contrapartida, o repasse do programa Criança Feliz ($r = 0,34$) e o PIB per capita ($r = 0,20$) mostram correlações positivas, ainda que de menor magnitude, indicando que políticas sociais e condições econômicas contribuem de forma complementar. Esses achados reforçam a necessidade de estratégias educacionais integradas para enfrentar os desafios estruturais e qualificar a prática docente.

A Figura 14, referente ao mapa de correlação das variáveis selecionadas para a base do ensino médio, revela um padrão mais disperso de correlações entre as variáveis selecionadas e o IDEB, com associações majoritariamente de baixa a moderada magnitude. Destacam-se, entre as variáveis com maior força explicativa, a média de horas-aula, positivamente correlacionada com o IDEB ($r = 0,45$), e a distorção idade-série, com correlação negativa expressiva ($r = -0,50$), reforçando a relevância da permanência regular e do tempo efetivo de instrução para o desempenho dos estudantes.

A heterogeneidade dos coeficientes sugere que o desempenho no ensino médio é

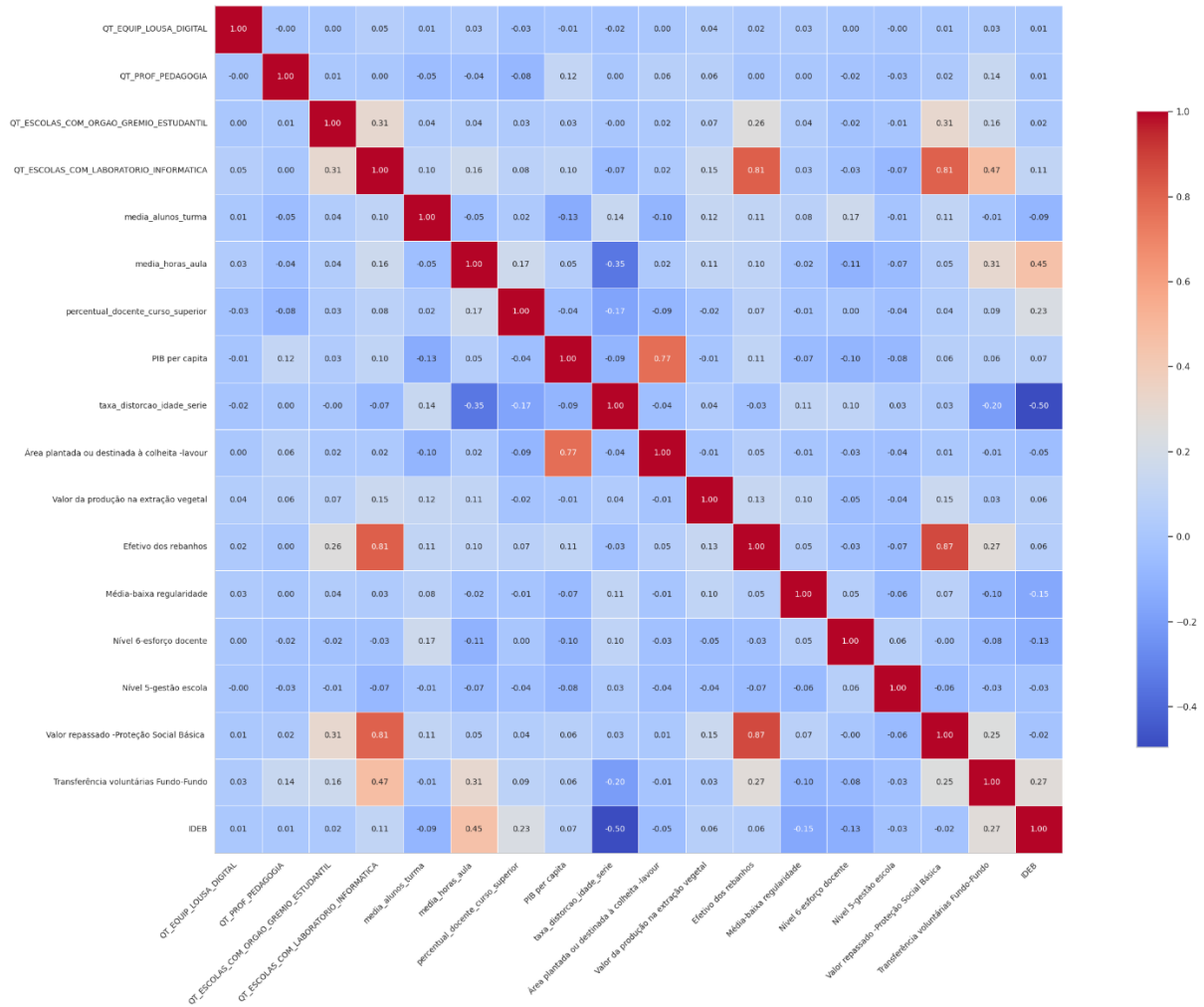
Figura 13 – Mapa de correlação entre variáveis selecionadas e o IDEB médio piauiense dos anos finais do ensino fundamental



Fonte: Elaborado pela autora (2025).

determinado por um conjunto mais difuso de fatores, exigindo abordagens analíticas integradas para captar a complexidade dessa etapa.

A análise descritiva consolidou um panorama da qualidade da educação básica nos municípios piauienses, articulando a evolução do IDEB com a caracterização estrutural, socioeconômica e institucional das redes de ensino. Os mapas de correlação revelaram relações lineares relevantes, destacando o impacto negativo da distorção idade-série e da formação docente inadequada, e a influência positiva, ainda que modesta, de indicadores socioeconômicos. Esses achados fundamentam a etapa seguinte da pesquisa, dedicada à modelagem preditiva do IDEB com técnicas de aprendizado de máquina.

Figura 14 – Mapa de correlação entre variáveis selecionadas e o IDEB médio piauiense do ensino médio

Fonte: Elaborado pela autora (2025).

5.2 Resultado dos Modelos

Após o pré-processamento e a seleção das variáveis explicativas, foram aplicados dez algoritmos clássicos de ML e sete arquiteturas de DL às três bases de dados. Para maior clareza, são apresentados apenas os modelos com melhor desempenho em cada etapa, avaliados por MAE, RMSE e R^2 , com validação cruzada estratificada em cinco dobras para garantir robustez e reduzir a variabilidade, especialmente relevante no ensino médio devido ao menor tamanho amostral. Os resultados são apresentados a seguir, iniciando pelos anos iniciais e finais do ensino fundamental e, em seguida, pelo ensino médio.

Na Tabela 2, o XGBoost se destacou como o modelo mais eficaz para prever o IDEB nos anos iniciais, com $R^2 = 0,5542$ no teste e $0,5455 (\pm 0,0293)$ na validação cruzada, além dos menores MAE (0,5050) e RMSE (0,6431), evidenciando sua robustez e capacidade de generalização. Outros ensembles, como RF e GBM, também superaram os modelos lineares, embora com leve perda de acurácia. Já os modelos lineares (Regressão Linear, Ridge, Lasso e Elastic Net) tiveram R^2 médios em torno de 0,50, limitados pela simplicidade da

abordagem.

A DT apresentou o pior desempenho, com $R^2 = 0,4275$ no teste, indicando sobreajuste e alta sensibilidade aos dados. Modelos como SVR e KNN foram excluídos por resultados insatisfatórios e instabilidade, não sendo recomendados neste contexto.

Tabela 2 – Desempenho dos melhores modelos de ML para os anos iniciais do ensino fundamental com as variáveis selecionadas

Modelo	Métricas	Treino	Teste	Validação Cruzada
Regressão Linear	R^2	0,5336	0,5037	$0,5227 \pm 0,0351$
	MAE	0,4968	0,5240	$0,5014 \pm 0,0180$
	RMSE	0,6368	0,6785	$0,6430 \pm 0,0261$
Lasso	R^2	0,5063	0,4575	$0,5027 \pm 0,0308$
	MAE	0,5084	0,5447	$0,5099 \pm 0,0177$
	RMSE	0,6552	0,7094	$0,6565 \pm 0,0259$
Ridge	R^2	0,5336	0,5037	$0,5228 \pm 0,0351$
	MAE	0,4968	0,5240	$0,5014 \pm 0,0180$
	RMSE	0,6368	0,6785	$0,6429 \pm 0,0261$
Elastic Net	R^2	0,5214	0,4835	$0,5171 \pm 0,0310$
	MAE	0,5007	0,5321	$0,5023 \pm 0,0192$
	RMSE	0,6451	0,6922	$0,6469 \pm 0,0237$
RF	R^2	0,6012	0,5222	$0,5311 \pm 0,0292$
	MAE	0,4548	0,5192	$0,4917 \pm 0,0245$
	RMSE	0,5889	0,6657	$0,6373 \pm 0,0193$
GBM	R^2	0,6102	0,5123	$0,5270 \pm 0,0256$
	MAE	0,4535	0,5254	$0,4973 \pm 0,0211$
	RMSE	0,5821	0,6726	$0,6403 \pm 0,0219$
XGBoost	R^2	0,6385	0,5542	$0,5455 \pm 0,0293$
	MAE	0,4354	0,5050	$0,4867 \pm 0,0197$
	RMSE	0,5606	0,6431	$0,6275 \pm 0,0216$
DT	R^2	0,5400	0,4275	$0,4864 \pm 0,0361$
	MAE	0,4944	0,5628	$0,5224 \pm 0,0220$
	RMSE	0,6324	0,7288	$0,6669 \pm 0,0216$

Fonte: Elaborado pela autora (2025).

Na Tabela 3, com os resultados de DL nos anos iniciais, observa-se que o melhor desempenho entre os modelos de DL para os anos iniciais foi obtido pelo MLP simples ($R^2 = 0,5462$ no teste e $0,5288$ na validação cruzada; MAE = $0,5210$; RMSE = $0,6708$), evidenciando estabilidade e boa capacidade preditiva. As arquiteturas híbridas CNN+LSTM e CNN 1D apresentaram resultados próximos ($R^2 \approx 0,52$), mas inferiores ao XGBoost. O LSTM puro teve desempenho ligeiramente menor ($R^2 = 0,5291$ no teste e $0,4839$ na validação).

Em contraste, modelos mais complexos, como Deep MLP e MLP com BatchNorm, apresentaram os piores resultados, indicando que a maior profundidade ou normalização não beneficiou a performance em dados tabulares com amostras limitadas. Comparativamente, o XGBoost superou todos os modelos de DL, confirmando a superioridade dos ensembles para regressão em contextos educacionais com estrutura multivariada.

Tabela 3 – Desempenho dos modelos de DL para os anos iniciais do ensino fundamental com as variáveis selecionadas

Modelo	Métricas	Treino	Teste	Validação Cruzada
MLP	R^2	0,5922	0,5462	0,5288 $\pm$ 0,0645
	MAE	0,4570	0,5210	0,4957 $\pm$ 0,0295
	RMSE	0,5924	0,6708	0,6329 $\pm$ 0,0422
MLP com Dropout	R^2	0,5388	0,5186	0,5044 $\pm$ 0,0553
	MAE	0,4834	0,5343	0,5053 $\pm$ 0,0262
	RMSE	0,6300	0,6909	0,6503 $\pm$ 0,0417
MLP com BatchNorm	R^2	0,4280	0,3826	0,0667 $\pm$ 0,2124
	MAE	0,5007	0,6108	0,6513 $\pm$ 0,0783
	RMSE	0,7016	0,7824	0,8935 $\pm$ 0,1418
Deep MLP	R^2	0,4259	0,4186	0,3956 $\pm$ 0,1002
	MAE	0,5403	0,5807	0,5455 $\pm$ 0,0297
	RMSE	0,7029	0,7593	0,7142 $\pm$ 0,0344
CNN 1D	R^2	0,5421	0,5234	0,5081 $\pm$ 0,0614
	MAE	0,4922	0,5378	0,5086 $\pm$ 0,0300
	RMSE	0,6278	0,6875	0,6471 $\pm$ 0,0393
LSTM	R^2	0,5114	0,5291	0,4839 $\pm$ 0,0675
	MAE	0,5072	0,5353	0,5239 $\pm$ 0,0270
	RMSE	0,6484	0,6833	0,6623 $\pm$ 0,0393
CNN + LSTM	R^2	0,5418	0,5255	0,5125 $\pm$ 0,0613
	MAE	0,4925	0,5376	0,5067 $\pm$ 0,0291
	RMSE	0,6280	0,6859	0,6442 $\pm$ 0,0394

Fonte: Elaborado pela autora (2025).

Na Tabela 4, com os resultados de ML nos anos finais, observa-se que o XGBoost apresentou o melhor desempenho na predição do IDEB para os anos finais do ensino fundamental ($R^2 = 0,4823$ no teste e $0,5205$ na validação cruzada; $MAE = 0,4314$; $RMSE = 0,5581$), destacando-se pela robustez e estabilidade entre os folds. Os ensembles GBM e RF tiveram resultados próximos, porém ligeiramente inferiores ($R^2 \approx 0,46$ no teste e $\approx 0,50$ na validação), mantendo boa consistência preditiva.

Os modelos lineares (Regressão Linear, Ridge, Lasso e Elastic Net) alcançaram R^2 em torno de $0,44$ no teste e $0,48$ na validação, com limitações para capturar a complexidade dos dados. Por fim, a DT apresentou o pior desempenho ($R^2 = 0,36$ no teste e $0,43$ na validação), com alta variabilidade nos erros, confirmando sua baixa adequação para o problema analisado.

Na Tabela 5, os modelos de DL aplicados aos anos finais do ensino fundamental não superaram o desempenho do XGBoost ($R^2 = 0,4823$ no teste e $0,5205$ na validação), embora alguns apresentassem resultados próximos. O MLP simples foi a melhor arquitetura entre as redes, com $R^2 = 0,4516$ no teste e $0,4634$ na validação, além de erros compatíveis com os melhores modelos de DL. A CNN+LSTM apresentou desempenho estável ($R^2 \approx 0,45$ no teste e $\approx 0,42$ na validação), seguida pelo MLP com Dropout e pelas arquiteturas LSTM pura, CNN 1D e Deep MLP, todas com R^2 entre $0,42$ e $0,43$ na validação.

Tabela 4 – Desempenho dos melhores modelos de ML para os anos finais do ensino fundamental com as variáveis selecionadas

Modelo	Métricas	Treino	Teste	Validação Cruzada
Regressão Linear	R^2	0,4859	0,4428	$0,4812 \pm 0,0748$
	MAE	0,4280	0,4496	$0,4311 \pm 0,0342$
	RMSE	0,5419	0,5790	$0,5433 \pm 0,0536$
Lasso	R^2	0,4453	0,4084	$0,4447 \pm 0,0609$
	MAE	0,4471	0,4657	$0,4486 \pm 0,0308$
	RMSE	0,5629	0,5966	$0,5625 \pm 0,0472$
Ridge	R^2	0,4859	0,4428	$0,4813 \pm 0,0747$
	MAE	0,4280	0,4496	$0,4311 \pm 0,0342$
	RMSE	0,5419	0,5790	$0,5433 \pm 0,0535$
Elastic Net	R^2	0,4680	0,4313	$0,4665 \pm 0,0686$
	MAE	0,4365	0,4543	$0,4385 \pm 0,0327$
	RMSE	0,5513	0,5849	$0,5512 \pm 0,0507$
RF	R^2	0,5675	0,4573	$0,5032 \pm 0,0545$
	MAE	0,3885	0,4415	$0,4181 \pm 0,0289$
	RMSE	0,4970	0,5714	$0,5319 \pm 0,0427$
GBM	R^2	0,5794	0,4653	$0,5005 \pm 0,0619$
	MAE	0,3835	0,4371	$0,4206 \pm 0,0303$
	RMSE	0,4902	0,5671	$0,5332 \pm 0,0462$
XGBoost	R^2	0,6128	0,4823	$0,5205 \pm 0,0630$
	MAE	0,3681	0,4314	$0,4124 \pm 0,0335$
	RMSE	0,4703	0,5581	$0,5225 \pm 0,0477$
DT	R^2	0,4965	0,3621	$0,4352 \pm 0,0733$
	MAE	0,4164	0,4851	$0,4468 \pm 0,0322$
	RMSE	0,5363	0,6195	$0,5670 \pm 0,0510$

Fonte: Elaborado pela autora (2025).

O MLP com Batch Normalization apresentou o pior resultado ($R^2 = 0,1091$). Em síntese, os algoritmos de ML, especialmente o XGBoost, mostraram maior robustez e generalização, enquanto os modelos de DL apresentaram maior variabilidade e sensibilidade aos parâmetros. Assim, para os anos finais, métodos baseados em árvores com boosting continuam mais eficazes para predição do IDEB, considerando as variáveis do estudo.

A Tabela 6 apresenta os resultados dos modelos de ML aplicados ao ensino médio, etapa marcada por maior complexidade preditiva. O XGBoost destacou-se como o melhor modelo, com R^2 de 0,4186 no teste e 0,4488 na validação cruzada, demonstrando robustez e superioridade frente aos demais. Os ensembles GBM e RF também obtiveram resultados próximos (R^2 entre 0,4054 e 0,4111 no teste e em torno de 0,42 na validação), confirmando sua adequação para capturar relações não lineares.

Em contrapartida, a Regressão Linear apresentou o pior desempenho, com R^2 de apenas 0,3324 no teste, evidenciando limitações para modelar fenômenos multifatoriais. A DT simples também apresentou resultados insatisfatórios ($R^2 = 0,3357$ no teste e 0,2714 na validação), reforçando a fragilidade dos métodos menos sofisticados diante da complexidade dos dados do ensino médio.

Tabela 5 – Desempenho dos modelos de DL para os anos finais do ensino fundamental com as variáveis selecionadas

Modelo	Métricas	Treino	Teste	Validação Cruzada
MLP	R^2	0,5562	0,4516	0,4634 ± 0,0727
	MAE	0,3936	0,4507	0,4286 ± 0,0241
	RMSE	0,5068	0,5672	0,5513 ± 0,0278
MLP com Dropout	R^2	0,5091	0,4684	0,4202 ± 0,0922
	MAE	0,4128	0,4456	0,4425 ± 0,0292
	RMSE	0,5330	0,5584	0,5724 ± 0,0330
MLP com BatchNorm	R^2	0,3970	0,2608	0,1091 ± 0,1968
	MAE	0,4514	0,5241	0,5449 ± 0,0511
	RMSE	0,5907	0,6585	0,7099 ± 0,0878
Deep MLP	R^2	0,4041	0,3508	0,2981 ± 0,0817
	MAE	0,4585	0,4935	0,4901 ± 0,0253
	RMSE	0,5872	0,6171	0,6313 ± 0,0369
CNN 1D	R^2	0,4811	0,4330	0,4337 ± 0,0639
	MAE	0,4296	0,4605	0,4455 ± 0,0195
	RMSE	0,5480	0,5767	0,5666 ± 0,0160
LSTM	R^2	0,4565	0,4283	0,4274 ± 0,0694
	MAE	0,4417	0,4599	0,4509 ± 0,0255
	RMSE	0,5608	0,5791	0,5697 ± 0,0257
CNN + LSTM	R^2	0,4760	0,4487	0,4246 ± 0,0599
	MAE	0,4319	0,4519	0,4492 ± 0,0220
	RMSE	0,5507	0,5686	0,5715 ± 0,0208

Fonte: Elaborado pela autora (2025).

Tabela 6 – Desempenho dos melhores modelos de ML para o ensino médio com as variáveis selecionadas

Modelo	Métricas	Treino	Teste	Validação Cruzada
Regressão Linear	R^2	0,4012	0,3324	–
	MAE	0,3567	0,3764	–
	RMSE	0,4716	0,5000	–
RF	R^2	0,5162	0,4111	0,4207 ± 0,0678
	MAE	0,3095	0,3727	0,3448 ± 0,0483
	RMSE	0,4239	0,4696	0,4620 ± 0,0784
GBM	R^2	0,5809	0,4054	0,4180 ± 0,0793
	MAE	0,2951	0,3700	0,3464 ± 0,0525
	RMSE	0,3946	0,4719	0,4631 ± 0,0823
XGBoost	R^2	0,6272	0,4186	0,4488 ± 0,0868
	MAE	0,2754	0,3665	0,3399 ± 0,0500
	RMSE	0,3721	0,4666	0,4502 ± 0,0807
DT	R^2	0,4249	0,3357	0,2714 ± 0,1028
	MAE	0,3416	0,3887	0,3993 ± 0,0553
	RMSE	0,4622	0,4987	0,5172 ± 0,0869

Fonte: Elaborado pela autora (2025).

A Tabela 7 apresenta os resultados dos modelos de DL no ensino médio, evidenciando desempenho globalmente inferior ao registrado nas demais etapas da educação básica. O

melhor resultado entre as arquiteturas foi obtido pela CNN 1D, com R^2 de 0,2683 no teste e 0,2747 na validação cruzada, destacando-se no grupo por sua relativa estabilidade, ainda que modesta.

Tabela 7 – Desempenho dos modelos de DL para o ensino médio com as variáveis selecionadas

Modelo	Métricas	Treino	Teste	Validação Cruzada
MLP	R^2	0,5462	-0,0088	-0,0939 $\pm$ 0,2751
	MAE	0,3095	0,4709	0,4693 $\pm$ 0,0448
	RMSE	0,4119	0,6181	0,6271 $\pm$ 0,0733
MLP com Dropout	R^2	0,3565	0,0773	0,0448 $\pm$ 0,2206
	MAE	0,3743	0,4598	0,4548 $\pm$ 0,0515
	RMSE	0,4904	0,5912	0,5914 $\pm$ 0,0902
MLP com BatchNorm	R^2	-0,0793	-0,2491	-1,5274 $\pm$ 1,4985
	MAE	0,4047	0,5368	0,6498 $\pm$ 0,0947
	RMSE	0,6351	0,6878	0,9215 $\pm$ 0,2265
Deep MLP	R^2	0,2180	0,1127	0,0368 $\pm$ 0,0866
	MAE	0,4094	0,4607	0,4650 $\pm$ 0,0361
	RMSE	0,5406	0,5797	0,5955 $\pm$ 0,0587
CNN 1D	R^2	0,4508	0,2683	0,2747 $\pm$ 0,1061
	MAE	0,3449	0,3981	0,4035 $\pm$ 0,0485
	RMSE	0,4531	0,5264	0,5160 $\pm$ 0,0575
LSTM	R^2	0,1763	0,1890	0,1227 $\pm$ 0,0337
	MAE	0,4333	0,4392	0,4472 $\pm$ 0,0337
	RMSE	0,5549	0,5542	0,5688 $\pm$ 0,0527
CNN + LSTM	R^2	0,3214	0,2687	0,1092 $\pm$ 0,0637
	MAE	0,3898	0,4015	0,4487 $\pm$ 0,0413
	RMSE	0,5036	0,5263	0,5739 $\pm$ 0,0652

Fonte: Elaborado pela autora (2025).

Em contraste, o pior desempenho foi observado no MLP simples, cujo R^2 foi negativo tanto no teste ($-0,0088$) quanto na validação cruzada ($-0,0939 \pm 0,2751$), indicando sobreajuste severo e incapacidade de modelar adequadamente o fenômeno. Esse resultado, aliado ao desempenho instável das arquiteturas mais complexas, reforça a fragilidade dos modelos de DL diante do reduzido tamanho amostral e da elevada heterogeneidade dos dados do ensino médio, que impõem limitações significativas à generalização e à convergência dos algoritmos.

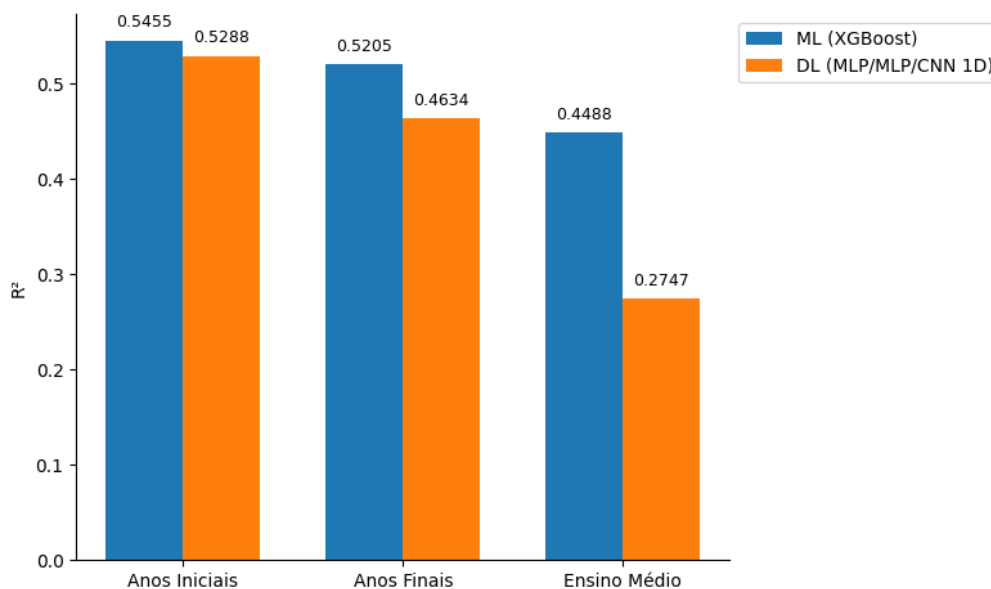
5.3 Avaliação Comparativa dos Modelos e Considerações Críticas

A presente seção realiza uma análise comparativa do desempenho dos modelos preditivos avaliados nas três etapas da educação. A comparação fundamenta-se nos resultados apresentados nas Tabelas 2 a 7, com foco nas métricas de desempenho (R^2 , MAE e RMSE) obtidas nos conjuntos de treinamento, teste e validação cruzada.

Já os gráficos apresentados a seguir, correspondentes às Figuras 15, 16 e 17, sintetizam os resultados apenas da validação cruzada. A Figura 15 apresenta a comparação

do R^2 obtido por meio da validação cruzada para os modelos com melhor desempenho em cada etapa: XGBoost para ML e MLP (anos iniciais e finais) e CNN 1D (ensino médio) para DL.

Figura 15 – Comparação dos melhores modelos de ML (XGBoost) e DL (MLP nos anos iniciais e finais; CNN 1D no ensino médio), segundo o R^2 , por etapa da educação básica



Fonte: Elaborado pela autora (2025).

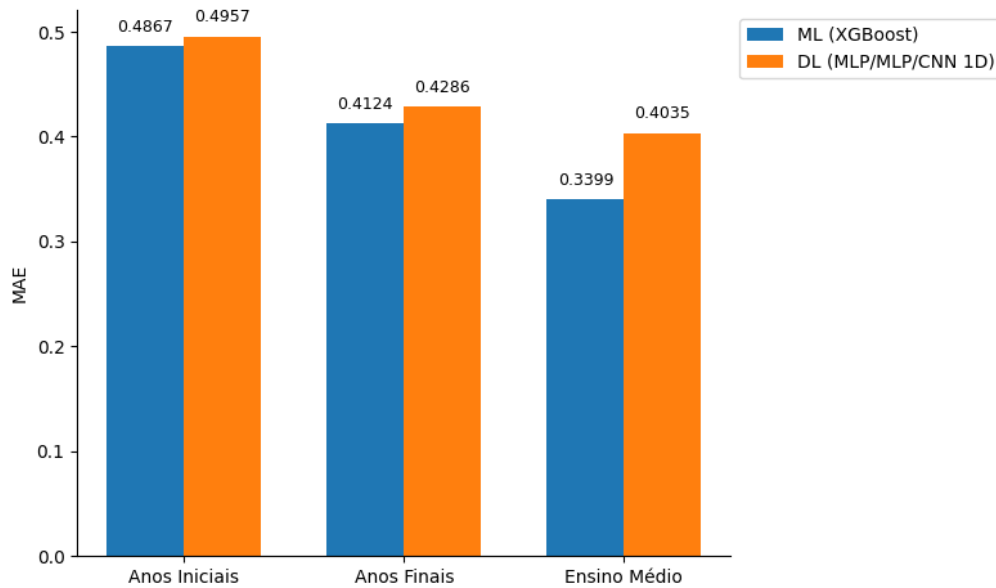
Observa-se que, em todas as etapas, o modelo XGBoost apresenta maior poder explicativo, com destaque para o ensino médio, onde a diferença entre ML e DL é mais expressiva (0,4488 vs. 0,2747). Nos anos iniciais e finais, embora o MLP apresente resultados competitivos, o XGBoost mantém a superioridade.

A Figura 16 apresenta os valores médios do MAE obtidos na validação cruzada para os modelos com melhor desempenho em cada etapa da educação básica. Essa métrica, que representa o desvio médio entre valores previstos e observados, é menos sensível a outliers do que o RMSE e de fácil interpretação. Em todas as etapas, o XGBoost obteve os menores MAEs, indicando maior precisão nas previsões. Nos anos iniciais, o MAE foi de 0,4867 para o XGBoost e 0,4957 para o MLP; nos anos finais, 0,4124 e 0,4286, respectivamente. No ensino médio, a diferença foi mais acentuada: o XGBoost alcançou MAE de 0,3399, enquanto a CNN 1D apresentou 0,4035, um aumento superior a 18% no erro do modelo de DL.

Já a Figura 17 apresenta os valores médios da RMSE, métrica que penaliza mais fortemente os grandes desvios entre valores previstos e observados. Os resultados reiteram a superioridade do modelo XGBoost nas três etapas da educação básica, com menor dispersão nos erros preditivos em comparação às arquiteturas de DL.

Nos anos iniciais e finais, as diferenças foram discretas, mas ainda favoráveis ao modelo de ML. No ensino médio, a discrepância foi mais acentuada: o XGBoost alcançou RMSE de 0,4502, enquanto a CNN 1D obteve 0,5160, indicando que o modelo de ensemble

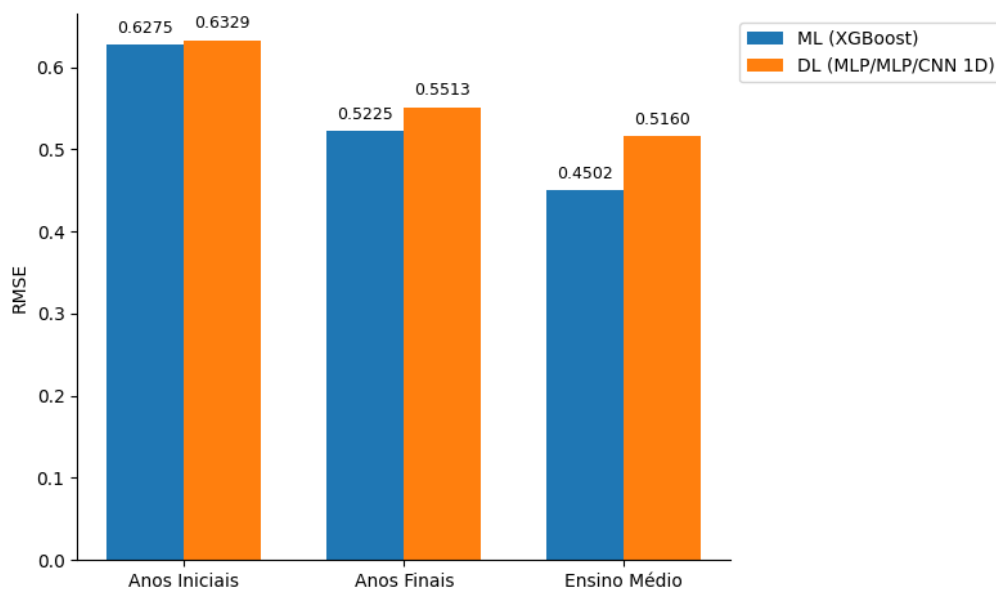
Figura 16 – Comparação dos melhores modelos de ML (XGBoost) e DL (MLP nos anos iniciais e finais; CNN 1D no ensino médio), segundo o MAE, por etapa da educação básica



Fonte: Elaborado pela autora (2025).

produziu previsões mais consistentes e menos influenciadas por erros extremos. Esse padrão reforça a robustez do XGBoost frente à complexidade e à variabilidade dos dados educacionais.

Figura 17 – Comparação dos melhores modelos de ML (XGBoost) e DL (MLP nos anos iniciais e finais; CNN 1D no ensino médio), segundo o RMSE, por etapa da educação básica



Fonte: Elaborado pela autora (2025).

Complementarmente aos gráficos de desempenho médio obtidos por validação cruzada, realiza-se aqui uma análise detalhada dos resultados de todos os modelos testados, considerando as métricas em treinamento, teste e validação cruzada. Nos anos iniciais do

ensino fundamental, o XGBoost apresentou superioridade clara em relação a todos os demais modelos, incluindo as arquiteturas de DL. Com R^2 de 0,5542 no teste e $0,5455 \pm 0,0293$ na validação cruzada, além dos menores MAE (0,5050) e RMSE (0,6431), o algoritmo demonstrou elevada capacidade de generalização. O MLP simples apresentou desempenho competitivo ($R^2 = 0,5462$ no teste; 0,5288 na validação), e arquiteturas híbridas como a CNN+LSTM também se aproximaram, mas sem superar o XGBoost, reforçando seu potencial para modelagem de dados educacionais tabulares.

Nos anos finais do ensino fundamental, o XGBoost novamente se destacou, com R^2 de 0,4823 no teste e $0,5205 \pm 0,0630$ na validação cruzada, superando as demais abordagens. As arquiteturas de DL com melhor desempenho, MLP simples e CNN+LSTM, alcançaram R^2 de 0,4516 e 0,4487 no teste, respectivamente, valores notavelmente inferiores aos do modelo de ML. Além disso, as redes neurais apresentaram maior instabilidade entre os *folds*, sobretudo nas versões com normalização ou maior profundidade, o que reforça a robustez dos métodos baseados em *boosting* frente a bases moderadas e variáveis predominantemente quantitativas.

No ensino médio, observou-se um decréscimo geral no desempenho dos modelos, reflexo da menor amostra ($n = 745$) e da maior complexidade dessa etapa. Ainda assim, o XGBoost manteve-se como o algoritmo mais eficaz, com R^2 de 0,4186 no teste e $0,4488 \pm 0,0868$ na validação cruzada, superando as redes neurais. Entre as arquiteturas de DL, a CNN 1D apresentou os melhores resultados, embora modestos ($R^2 = 0,2683$ no teste; 0,2747 na validação). Os demais modelos de DL, como o MLP, apresentaram R^2 negativos, indicando sobreajuste e instabilidade preditiva. Esses achados evidenciam a insuficiência dos dados disponíveis para capturar a complexidade dos fenômenos educacionais do ensino médio, especialmente em relação a variáveis subjetivas como engajamento ou permanência escolar.

Por fim, cabe destacar a comparação com o estudo de Maia, Bueno e Sato (2021), que empregou regressão linear, Lasso, Ridge, Elastic Net e GBM para prever o IDEB em todos os municípios brasileiros. O GBM, no teste, obteve R^2 de 0,421 (anos iniciais), 0,305 (anos finais) e 0,459 (ensino médio), com MAE entre 0,593 e 0,647, resultados inferiores aos desta dissertação. Nesta pesquisa, o XGBoost alcançou R^2 de 0,5542 (anos iniciais), 0,4823 (anos finais) e 0,4186 (ensino médio), com menores erros em todas as etapas. Além disso, enquanto o estudo de Maia, Bueno e Sato (2021) não utilizou validação cruzada, esta dissertação aplicou *k-fold cross-validation*, conferindo maior robustez e reduzindo o risco de sobreajuste.

Esses resultados não apenas superam os de uma referência consolidada, como também ilustram o potencial de abordagens metodológicas que integram validação cruzada e técnicas complementares de seleção de variáveis. Ainda que os achados estejam restritos ao contexto piauiense, eles sugerem caminhos promissores para futuras análises regionais baseadas em *pipelines* estatísticos robustos.

5.4 Análise dos Resultados Considerando os Objetivos, Questões de Pesquisa e Hipóteses

Os resultados obtidos nesta pesquisa evidenciam que o objetivo geral — aplicar e avaliar modelos preditivos baseados em ML e DL para estimar o IDEB dos municípios piauienses nas três etapas da educação básica, integrando variáveis educacionais, socioeconômicas e financeiras — foi plenamente alcançado. Foram implementados modelos, com *pipeline* metodológico replicável, que se mostraram capazes de prever o desempenho educacional com diferentes graus de acurácia e estabilidade, conforme a etapa analisada.

Quanto aos objetivos específicos, todos foram atingidos. O primeiro, testar e comparar o desempenho de algoritmos de ML e DL para prever o IDEB com base em métricas (MAE, RMSE e R^2), foi cumprido por meio da aplicação de dez modelos clássicos de ML e sete arquiteturas de DL. O XGBoost destacou-se sistematicamente como o mais eficaz em todas as etapas, apresentando superioridade em robustez, estabilidade e acurácia frente às redes neurais, justificando sua adoção nas etapas subsequentes, incluindo a simulação de cenários exploratórios.

O segundo objetivo, identificar os fatores explicativos mais relevantes para a variação do IDEB, também foi atendido. Para os anos iniciais e finais, seis variáveis-chave (como PIB per capita, taxa de distorção idade-série e percentual de docentes sem formação superior) foram identificadas com técnicas como VIF, LassoCV, SHAP e Boruta. Já no ensino médio, dada sua maior complexidade, recorreu-se à análise de correlação, resultando em 17 variáveis selecionadas. Esses achados confirmam a importância de fatores socioeconômicos, pedagógicos e estruturais no desempenho educacional.

O terceiro objetivo, simular cenários exploratórios, foi realizado com a construção de um sistema interativo fundamentado no modelo XGBoost. Por meio desse sistema, foi possível estimar os impactos contrafactuais de intervenções hipotéticas em variáveis-chave sobre os valores previstos do IDEB de 2023, reforçando a aplicabilidade prática dos resultados para subsidiar políticas educacionais territorialmente sensíveis e fundamentadas em evidências.

As hipóteses foram, em sua maioria, confirmadas. A H_1 se confirmou, já que os modelos previram o IDEB com desempenho consistente, exclusivamente com variáveis públicas. A H_2 também foi validada: os algoritmos de ML (como XGBoost, RF e GBM) superaram os modelos lineares tradicionais em todas as métricas. A H_3 foi parcialmente confirmada: embora tenha sido possível identificar variáveis explicativas relevantes nos melhores modelos, a ausência de atributos subjetivos (como engajamento discente e clima escolar) limitou a completude das inferências.

Por fim, no que tange às questões de pesquisa, a primeira, relacionada aos desafios de integração e preparação dos dados, foi amplamente endereçada durante a etapa de pré-processamento. Entre os principais obstáculos enfrentados, destaca-se a necessidade de

consolidar dados provenientes de diferentes fontes e sistemas, muitas vezes em formatos e granularidades distintas. Como os microdados educacionais, financeiros e socioeconômicos são disponibilizados separadamente por diferentes órgãos (INEP, IBGE, SIOPE, entre outros), tornou-se imprescindível a adoção de uma estratégia sistemática de junção, padronização e harmonização das bases, respeitando filtros de ano, município e etapa de ensino. Esse processo exigiu não apenas o cruzamento de identificadores comuns, mas também a construção de um *pipeline* robusto de tratamento de variáveis para viabilizar análises estratificadas consistentes.

A segunda questão, sobre o melhor algoritmo, foi respondida de forma conclusiva: o XGBoost apresentou os melhores resultados em todas as etapas e métricas avaliadas. A terceira questão de pesquisa, referente à interpretação dos resultados do modelo de melhor desempenho, foi respondida com base na análise de importância das variáveis no XGBoost para cada etapa da educação básica. Nos anos iniciais e finais, destacaram-se fatores socioeconômicos e estruturais, como a taxa de distorção idade-série, o PIB per capita e o percentual de docentes sem formação superior. No ensino médio, o modelo apontou um conjunto mais heterogêneo de variáveis, incluindo aspectos tecnológicos, administrativos, pedagógicos e socioeconômicos, refletindo a complexidade dessa etapa. Os resultados evidenciam que o desempenho educacional é influenciado por múltiplos fatores interdependentes, exigindo políticas articuladas e sensíveis ao contexto.

A quarta questão, relativa à aplicabilidade do modelo preditivo como suporte à formulação de políticas públicas educacionais, foi abordada por meio da simulação de cenários contrafactuais. Essa etapa buscou demonstrar de que forma os resultados do melhor modelo identificado podem subsidiar decisões estratégicas no âmbito da gestão educacional. Ainda que a simulação tenha utilizado os mesmos dados empregados no treinamento — caracterizando-se, portanto, como um exercício condicional de validação funcional — sua estrutura computacional foi concebida para ser replicável em contextos reais, com dados prospectivos. Assim, a ferramenta proposta poderá ser apropriada por analistas e gestores para avaliação prévia de intervenções, contribuindo para o planejamento de políticas mais eficazes, contextualizadas e orientadas por evidências.

Por fim, constata-se que os objetivos desta pesquisa foram plenamente alcançados, tanto na construção de modelos preditivos com desempenho satisfatório quanto na implementação de uma aplicação prática voltada à análise de políticas públicas educacionais. A simulação de cenários contrafactuais, apresentada no próximo capítulo, exemplifica o potencial dessa abordagem para investigar a sensibilidade do IDEB a intervenções hipotéticas em variáveis explicativas observadas. É importante, contudo, destacar que os dados de 2023 foram utilizados simultaneamente no treinamento dos modelos e na simulação, o que caracteriza essa etapa como uma análise condicional e exploratória, e não como uma previsão futura. Portanto, os resultados obtidos devem ser interpretados exclusivamente como demonstrações do funcionamento do sistema de simulação, sem

extrapolação temporal.

A proposta metodológica, contudo, está estruturada para ser replicável com dados novos e independentes, de modo que, em futuras atualizações da base de dados, a simulação possa ser aplicada de fato a observações não utilizadas no treinamento, ampliando sua aplicabilidade como ferramenta de apoio à tomada de decisão por parte de gestores públicos. Esse direcionamento futuro reforça a contribuição prática da pesquisa, ao integrar modelagem preditiva com estratégias de análise orientadas por dados, dentro de uma perspectiva aplicada à realidade educacional dos municípios piauienses.

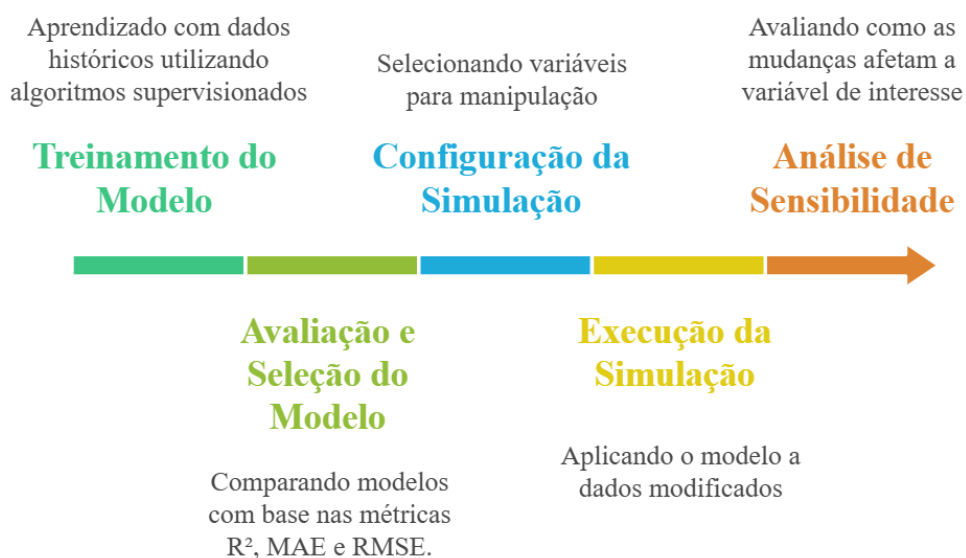
6 Simulação de Cenários Contrafactuais

Este capítulo apresenta a aplicação dos modelos preditivos na simulação de cenários contrafactuais do IDEB, avaliando como alterações hipotéticas nas variáveis selecionadas para cada etapa de ensino, conforme detalhado nos capítulos anteriores, podem modificar o IDEB municipal piauiense. A simulação contrafactual permite estimar impactos marginais de intervenções virtuais sem reestimar os modelos, aproximando a análise da prática do planejamento educacional orientado por evidências.

As simulações podem ser realizadas para variáveis isoladas ou para combinações de variáveis, mediante a aplicação de percentuais de aumento ou redução definidos de acordo com sua natureza e plausibilidade empírica. O objetivo é identificar os fatores mais sensíveis na predição do IDEB e compreender como seus efeitos variam entre as etapas da educação básica e entre diferentes perfis municipais. Dessa forma, o capítulo integra a modelagem preditiva a uma análise aplicada, oferecendo subsídios objetivos para a formulação de estratégias no âmbito da gestão educacional.

A Figura 18 representa as etapas metodológicas que estruturam o processo de modelagem preditiva e simulação contrafactual adotado neste estudo. O processo inicia-se com o tratamento dos dados e a seleção das variáveis explicativas, etapas desenvolvidas nos capítulos anteriores, seguidas do treinamento dos modelos preditivos. Posteriormente, realiza-se a avaliação e a seleção do modelo com melhor desempenho, considerando as métricas obtidas tanto na validação cruzada quanto na base de teste.

Figura 18 – Fluxo metodológico da previsão algorítmica à simulação contrafactual



Fonte: Elaborado pela autora (2025).

A partir do modelo selecionado, procede-se à configuração da simulação, definindo-se as variáveis explicativas a serem manipuladas e os percentuais de alteração a serem aplicados.

A execução da simulação ocorre mediante a aplicação do modelo fixo aos dados modificados, gerando diferentes cenários hipotéticos. Por fim, a análise de sensibilidade permite avaliar o impacto de cada alteração simulada sobre a variável de interesse, contribuindo para decisões educacionais fundamentadas em evidências quantitativas.

6.1 Fundamentação e Implementação da Simulação Contrafactual

A simulação contrafactual constitui a etapa analítica destinada a estimar os efeitos potenciais de alterações paramétricas nas variáveis explicativas sobre o IDEB. Essa estratégia combina os modelos preditivos ao módulo de simulação parametrizada e permite explorar, de forma experimental, como cenários alternativos poderiam influenciar o desempenho educacional projetado.

O XGBoost foi adotado como modelo padrão nas três etapas de ensino, por apresentar melhor desempenho nas métricas de R^2 e RMSE, além de maior estabilidade frente à variabilidade dos dados. Como princípio da simulação contrafactual, os parâmetros do modelo são mantidos fixos, assegurando que os efeitos estimados derivem exclusivamente das modificações aplicadas às variáveis explicativas.

O módulo de simulação foi implementado em Python, utilizando as bibliotecas *ipywidgets*, *pandas* e *matplotlib*. A interface interativa permite ao usuário selecionar a etapa de ensino, escolher variáveis a serem modificadas, definir percentuais de variação e nomear cada cenário. A Figura 19 ilustra essa interface.

A interface interativa do módulo de simulação contrafactual implementada em Python (Colab) apresenta os seguintes elementos:

- Tipo de En...:** Dropdown menu com a opção "Anos Iniciais" selecionada.
- Variável:** Dropdown menu com a opção "PIB per capita" selecionada.
- Percentual:** Campo de entrada com o valor "10".
- Operação:** Radio buttons para "Aumento" (selecionado) e "Redução".
- Botões de Ação:** "Adicionar variável" (verde) e "Limpar variáveis" (vermelho).
- Nome do cenário:** Campo de entrada com o texto "Cenário Personalizado".
- Visualizar:** Radio buttons para "Ver IDEB médio previsto" (selecionado) e "Ver tabela completa por município".
- Botão de Ação:** "Gerar Cenário" (verde).
- Tabela Final dos Cenários Gerados:** Seção para exibir os resultados.
- Botões de Ação:** "Gerar Gráficos" (azul) e "Exportar Resultados" (laranja).
- Botão de Ação:** "Ver Métricas do Modelo" (azul).

Figura 19 – Interface interativa do módulo de simulação contrafactual implementada em Python (Colab).

Fonte: Elaborado pela autora (2025)

O sistema executa automaticamente: (i) filtragem dos dados de 2023 para a etapa selecionada; (ii) aplicação dos ajustes paramétricos; (iii) cálculo das previsões; e (iv) comparação entre o cenário original e o cenário modificado, quantificando o impacto marginal de cada intervenção simulada.

Os coeficientes do modelo permanecem fixos durante todo o processo, o que preserva a natureza contrafactual da análise e garante que as variações observadas resultem exclusivamente das modificações aplicadas às variáveis explicativas. Os resultados podem ser exportados, visualizados em gráficos comparativos e sintetizados em tabela contendo: nome do cenário, variáveis alteradas, IDEB médio de 2023, IDEB previsto sem alterações, IDEB simulado e os respectivos deltas.

Ressalta-se que as simulações foram realizadas com base nos dados de 2023, os quais também compuseram o treinamento do modelo. Assim, não se tratam de previsões futuras, mas de análises internas ao conjunto observado, realizadas para assegurar consistência estatística na calibragem e na geração dos cenários. Os resultados devem ser interpretados como avaliações condicionais que estimam efeitos marginais de intervenções hipotéticas, seguindo uma lógica de experimentação empírica.

Como exemplo, apresenta-se a simulação “Cenário 1 – Avanço Socioeducacional”, referente aos anos finais do ensino fundamental, que combinou: aumento de 10% no PIB per capita, redução de 15% na taxa de distorção idade-série e redução de 20% na proporção de docentes sem formação adequada. O IDEB simulado foi de 4,9164, representando acréscimos de 0,193 pontos em relação ao cenário sem alterações e de 0,2559 pontos frente ao IDEB real de 2023. O resultado ilustra que melhorias coordenadas em fatores estruturais podem gerar ganhos relevantes, evidenciando a utilidade da modelagem preditiva para apoiar decisões educacionais.

Na sequência, apresenta-se o dicionário das variáveis utilizadas nas simulações, seguido dos testes contrafactuais realizados em municípios selecionados conforme sua heterogeneidade territorial. Tais testes permitem avaliar a sensibilidade do modelo a diferentes intervenções e identificar os cenários de maior impacto, subsidiando análises orientadas à formulação de políticas educacionais mais equitativas e fundamentadas em evidências.

6.2 Variáveis e Percentuais de Simulação por Etapa de Ensino

As variáveis utilizadas nas simulações contrafactuais, tanto no ensino fundamental quanto no ensino médio, correspondem aos fatores explicativos que apresentaram maior relevância preditiva, conforme demonstrado nos capítulos anteriores. Assim, foram mantidos apenas os indicadores com maior poder explicativo e estabilidade, que passaram a compor os modelos e, posteriormente, os cenários simulados.

Para a simulação, adotou-se variações percentuais controladas, representando intervenções factíveis no contexto dos municípios piauienses. Foram aplicadas reduções de -2% , -5% , -10% , -20% e -30% às variáveis em que a melhora corresponde a uma diminuição do indicador (como taxa de distorção idade-série). Para variáveis em que a evolução se expressa por incremento positivo (área plantada), aplicaram-se aumentos de $+2\%$, $+5\%$,

+10%, +20% e +30%.

No caso específico do PIB per capita, adotou-se intervalos mais conservadores (+1%, +3%, +5%, +7% e +10%), de modo a refletir trajetórias de crescimento economicamente plausíveis e consistentes com o comportamento recente das economias municipais. A padronização dos percentuais garante comparabilidade entre cenários e preserva a coerência metodológica da análise.

Observou-se que algumas variáveis apresentaram valor igual a zero em determinados municípios. Quando esse valor representava uma condição desejável, como ausência de distorção idade-série, nenhuma modificação foi aplicada, uma vez que o indicador já se encontrava em sua situação ideal. De forma semelhante, quando o valor zero indicava inexistência de operação ou repasse, como no caso do Programa Criança Feliz em municípios que não participaram da política em 2023, não foram aplicados acréscimos percentuais, pois qualquer variação sobre zero permanece igual a zero.

Em síntese, sempre que uma variável apresentou valor igual a zero, o simulador manteve esse valor inalterado, independentemente de sua natureza. As variações percentuais foram aplicadas exclusivamente aos municípios com valores positivos, assegurando coerência empírica, consistência lógica e realismo metodológico nos cenários simulados.

6.2.1 Variáveis do Ensino Fundamental (Anos Iniciais e Finais)

As variáveis incluídas nas simulações do ensino fundamental representam fatores estruturais e socioeconômicos associados ao desempenho municipal no IDEB. A seguir, apresentam-se os seis indicadores utilizados.

1. Taxa de distorção idade-série

Expressa o percentual de alunos matriculados em determinada série ou etapa de ensino que apresentam idade superior à recomendada para aquela etapa escolar. O cálculo é realizado dividindo-se o número de alunos com dois anos ou mais de atraso escolar pelo total de matrículas da série considerada, multiplicado por 100 para obtenção do valor em percentual. Em termos práticos, uma taxa de distorção de 25% significa que, a cada quatro estudantes, um encontra-se em situação de atraso escolar, seja por repetência ou ingresso tardio, o que impacta diretamente o fluxo educacional e a aprendizagem.

2. Percentual de docentes sem formação superior

Expressa a proporção de professores em exercício que não possuem diploma de curso superior. É calculado dividindo-se o número de docentes sem ensino superior pelo total de docentes e multiplicando por 100.

3. PIB per capita

Corresponde ao valor total do Produto Interno Bruto (PIB) de um município dividido pelo número de habitantes, expresso em reais. Representa a média da riqueza econômica produzida por pessoa, sendo um indicador relevante das condições socioeconômicas locais.

4. Valor da produção na extração vegetal

Refere-se ao valor monetário obtido pelas atividades de exploração de produtos vegetais nativos, como madeira, lenha, castanha, babaçu, carnaúba, borracha e outros itens não cultivados em lavouras. Esse dado é fornecido pelo IBGE e expressa a importância econômica do extrativismo vegetal na economia municipal.

5. Área plantada ou destinada à colheita das lavouras temporárias e permanentes

Refere-se à soma das áreas (em hectares) ocupadas por lavouras temporárias (como arroz, milho, feijão e mandioca) e lavouras permanentes (como banana, coco, laranja e castanha de caju). O dado é levantado pelo IBGE por meio da Pesquisa Agrícola Municipal (PAM) e expressa o grau de uso agrícola da terra em cada município.

6. Valor repassado por meio do programa Criança Feliz

Corresponde ao montante financeiro transferido pelo governo federal aos municípios para execução do Programa Criança Feliz, política intersetorial voltada à primeira infância (crianças de 0 a 6 anos em situação de vulnerabilidade social). Os recursos são aplicados em visitas domiciliares, acompanhamento do desenvolvimento infantil e integração de ações de saúde, educação, assistência social, cultura e direitos humanos.

6.2.2 Variáveis do Ensino Médio

As variáveis consideradas nas simulações para o ensino médio refletem tanto aspectos internos do ambiente escolar (infraestrutura, corpo docente e condições pedagógicas) quanto fatores externos de natureza socioeconômica e produtiva. A seguir, descrevem-se as 17 variáveis utilizadas:

1. Quantidade de escolas com lousa digital

Indica o número absoluto de escolas que possuem lousas digitais instaladas, segundo dados do Censo Escolar. Trata-se de um indicador de infraestrutura tecnológica voltada ao apoio pedagógico e à modernização do ensino.

2. Quantidade de escolas com laboratório de informática

Refere-se ao número absoluto de escolas que possuem laboratório de informática em funcionamento, de acordo com o Censo Escolar. Representa um indicador de infraestrutura tecnológica voltada ao ensino de informática, ao uso pedagógico de computadores e à inclusão digital dos alunos.

3. Média de alunos por turma

Expressa o número médio de estudantes matriculados em cada turma, segundo os dados do Censo Escolar. Indica a relação entre a quantidade de alunos e a capacidade física e pedagógica da escola, sendo um importante indicador de qualidade.

4. Média de horas-aula diária

Indica a média de tempo de permanência dos estudantes em atividades escolares por dia, expressa em horas, de acordo com os dados do Censo Escolar. Reflete a carga

horária efetiva ofertada pelas escolas e está diretamente associada ao tempo de exposição dos alunos aos conteúdos curriculares.

5. Taxa de distorção idade-série

O mesmo do ensino fundamental.

6. Regularidade média do corpo docente da escola, “Média-Baixa” (IRD médio maior que 2 até 3)

Indica o percentual de estudantes matriculados em escolas cujo corpo docente apresenta regularidade média classificada como “média-baixa”, segundo o Indicador de Regularidade Docente (IRD) do INEP. Esse indicador mede a permanência e estabilidade dos professores em determinada escola: quanto menor a rotatividade, maior a regularidade. A classificação “média-baixa” significa que a escola apresenta instabilidade no quadro de docentes, com impactos negativos na continuidade do processo de ensino-aprendizagem.

7. PIB per capita

O mesmo do ensino fundamental.

8. Área plantada ou destinada à colheita das lavouras temporárias e permanentes

O mesmo do ensino fundamental.

9. Valor da produção na extração vegetal

O mesmo do ensino fundamental.

10. Efetivo total dos rebanhos

Refere-se ao número total de animais de criação registrados no município (bovinos, suínos, caprinos, ovinos, aves etc.), conforme dados do IBGE.

11. Valor total repassado — Proteção Social Básica

Refere-se ao montante de recursos transferidos pela União aos municípios para financiar serviços, programas e benefícios socioassistenciais no âmbito do Sistema Único de Assistência Social (SUAS), especialmente voltados à proteção de famílias em situação de vulnerabilidade social. Esse valor é contabilizado anualmente e varia conforme o porte do município, número de famílias atendidas e critérios de cofinanciamento.

12. Transferências voluntárias realizadas - Fundo a Fundo

Refere-se aos recursos financeiros repassados pela União ou pelo Estado diretamente ao Fundo Municipal de Educação, sem necessidade de convênio, ou seja, transferências automáticas que podem ser aplicadas em manutenção e desenvolvimento do ensino. Esses valores são fundamentais para ampliar a capacidade de investimento dos municípios na infraestrutura escolar, capacitação docente e programas educacionais.

13. Quantidade de escolas com órgão de grêmio estudantil

Indica o número de escolas em um município que possuem grêmio estudantil formalmente constituído. O grêmio é uma entidade representativa dos estudantes que promove participação cidadã, organização de atividades culturais, esportivas e sociais, além de fortalecer o vínculo entre alunos, escola e comunidade.

14. Percentual de escolas no Nível 5 de gestão escolar

Refere-se à proporção de escolas que se enquadram no Nível 5 de gestão escolar. Esse nível caracteriza instituições de porte intermediário (150 a 1000 matrículas), funcionamento em três turnos, oferta de duas ou três etapas de ensino e presença da Educação de Jovens e Adultos (EJA) como etapa mais elevada. Essas escolas geralmente apresentam maior complexidade administrativa e demandam práticas de gestão mais estruturadas.

15. Professores com formação em Pedagogia

Representa a proporção de docentes que possuem graduação em Pedagogia, formação tradicionalmente voltada à docência na educação básica, sobretudo nos anos iniciais do ensino fundamental. Esse indicador expressa o nível de qualificação específica para a prática pedagógica.

16. Percentual de docentes com curso superior completo

Expressa a proporção de professores em exercício que concluíram um curso de graduação em nível superior, independentemente da área de formação. Esse indicador reflete o grau de qualificação acadêmica do corpo docente, sendo diretamente associado à valorização profissional e à melhoria da prática pedagógica.

17. Percentual de docentes no Nível 6 de esforço docente

Indica a proporção de professores que atuam em condições de sobrecarga extrema: geralmente atendem mais de 400 alunos, trabalham nos três turnos, lecionam em duas ou três escolas e em duas ou três etapas de ensino. Esse nível caracteriza situações de desgaste profissional elevado, com impacto negativo no tempo disponível para planejamento pedagógico, acompanhamento individualizado e qualidade do ensino.

6.3 Testes de Simulações Contrafactuais

Para a operacionalização da simulação de cenários contrafactuais, foram selecionados, em cada etapa de ensino, três municípios piauienses representativos, correspondentes ao pior, mediano e melhor desempenho no IDEB de 2023. Essa estratégia visa avaliar os impactos das intervenções simuladas em contextos contrastantes, abrangendo desde realidades educacionais mais vulneráveis até municípios com alto desempenho consolidado. Essa diversidade permite examinar a sensibilidade do modelo preditivo às variações contextuais e compreender de que modo os efeitos das intervenções hipotéticas se manifestam em diferentes perfis territoriais. O Quadro 7 apresenta os municípios selecionados e seus respectivos valores do IDEB de 2023 para cada etapa de ensino.

Quadro 7 – Seleção dos municípios segundo o desempenho no IDEB de 2023

Etapas de Ensino	Pior Desempenho	Desempenho Mediano	Melhor Desempenho
Anos Iniciais	Caracol – 3,5	Francisco Macedo – 5,3	Domingos Mourão – 8,4
Anos Finais	Barreiras do Piauí – 2,5	São João da Serra – 4,7	Buriti dos Montes – 6,5
Ensino Médio	Capitão de Campos – 3,1	Simplicio Mendes – 4,1	Cocal dos Alves – 7,0

Fonte: INEP (2023), elaborado pela autora.

Após a seleção dos municípios representativos, procedeu-se à caracterização de seus contextos socioeconômicos e educacionais, utilizando as seis variáveis de maior relevância preditiva identificadas para o ensino fundamental: PIB per capita, taxa de distorção idade-série, percentual de docentes sem formação superior, valor da produção extrativista vegetal, área plantada ou destinada à colheita e repasses do Programa Criança Feliz. Os dados referem-se ao ano de 2023, assegurando consistência temporal entre o conjunto observado e as projeções do modelo.

Nos anos iniciais, o município de Caracol, que apresentou o pior desempenho em 2023, registrou PIB per capita de R\$ 10.162,04, taxa de distorção idade-série de 42,1%, 34,4% de docentes sem formação superior, produção extrativista de R\$ 79 mil, área plantada de 5.855 hectares e repasse de R\$ 143.298 pelo Programa Criança Feliz. Em posição mediana, Francisco Macedo apresentou, também em 2023, PIB per capita de R\$ 12.366,08, distorção idade-série de 4,2%, inexistência de docentes sem formação superior, produção extrativista de R\$ 161 mil, área plantada das lavouras temporárias e permanentes de 1.208 hectares e ausência de repasses pelo Programa Criança Feliz. Já Domingos Mourão, município de melhor desempenho nos anos iniciais em 2023, destacou-se com PIB per capita de R\$ 11.299,76, distorção idade-série de 0,6%, 100% de docentes com formação superior, produção extrativista de R\$ 2.312 mil, área plantada de 2.074 hectares e repasse de R\$ 73.620 pelo Programa Criança Feliz.

Nos anos finais, Barreiras do Piauí obteve o pior resultado em 2023, com PIB per capita de R\$ 34.496,82, distorção idade-série de 35,1%, 9% de docentes sem formação superior, produção extrativista de R\$ 68 mil, área plantada das lavouras temporárias e permanentes de 13.649 hectares e repasse de R\$ 141.192 pelo Programa Criança Feliz. Em posição intermediária, São João da Serra apresentou PIB per capita de R\$ 11.151,51, distorção idade-série de 4,3%, corpo docente plenamente formado, produção extrativista de R\$ 4.502 mil, área plantada de 774 hectares e repasse de R\$ 107.370, todos referentes ao ano de 2023. Por sua vez, Buriti dos Montes, município de melhor desempenho, destacou-se em 2023 com PIB per capita de R\$ 78.255,64, distorção idade-série de 1,0%, 2,1% de docentes sem formação superior, produção extrativista de R\$ 1.743 mil, área plantada das lavouras temporárias e permanentes de 2.275 hectares e repasse de R\$ 50.220 pelo Programa Criança Feliz.

No ensino médio, Capitão de Campos registrou, em 2023, o pior desempenho, com PIB per capita de R\$ 10.638,81, distorção idade-série de 6,2%, 24,1% de docentes sem formação superior, produção extrativista de R\$ 426 mil, área plantada das lavouras temporárias e permanentes de 2.276 hectares e repasse de R\$ 108.045 pelo Programa Criança Feliz. Em posição mediana, Simpício Mendes apresentou PIB per capita de R\$ 16.390,86, distorção idade-série de 1,7%, apenas 2,5% de docentes sem formação superior, produção extrativista de R\$ 1.520 mil, área plantada das lavouras temporárias e permanentes de 2.239 hectares e repasse de R\$ 70.623, todos dados de 2023. Finalmente, Cocal dos Alves, município de melhor desempenho no ensino médio, registrou em 2023 PIB per capita de R\$ 10.414,14, distorção idade-série de 3,5%, corpo docente totalmente formado, produção extrativista de R\$ 796 mil, área plantada das lavouras temporárias e permanentes de 6.313 hectares e repasse de R\$ 149.625 pelo Programa Criança Feliz.

6.4 Resultados das Simulações

As simulações contrafactuais permitiram mensurar o impacto marginal de variações controladas em variáveis-chave sobre o IDEB previsto. Em cada cenário, apenas uma variável foi alterada, mantendo-se as demais constantes, de modo a isolar seu efeito específico. Os resultados são apresentados em dois blocos: ensino fundamental e ensino médio.

De modo geral, os impactos foram maiores nos municípios de menor desempenho, indicando maior sensibilidade a mudanças estruturais e pedagógicas. Nos municípios com resultados mais elevados, os efeitos foram reduzidos, o que é compatível com indicadores já próximos de limites superiores. Os resultados devem ser interpretados como exercícios contrafactuais baseados no modelo preditivo, indicando direção e magnitude esperadas das alterações simuladas, sem implicar causalidade.

6.4.1 Simulações do Ensino Fundamental

A Tabela 8 apresenta os resultados das simulações para Caracol, cujo IDEB real em 2023 foi 3,5 e o valor previsto pelo modelo, sem intervenção, foi 3,8798. As simulações avaliam o efeito de variações percentuais em variáveis estruturais, socioeconômicas e educacionais sobre o IDEB estimado.

Os resultados indicam que a redução do percentual de docentes sem formação superior é a variável de maior sensibilidade. Pequenas reduções entre 2% e 5% elevam o IDEB simulado em cerca de 0,32 ponto em relação ao cenário base, alcançando ganho de aproximadamente 0,70 ponto em relação ao valor real. A distorção idade-série também apresenta efeito significativo, uma vez que reduções de 20% e 30% resultam em aumentos de 0,20 e 0,25 ponto no IDEB previsto, respectivamente.

Variáveis econômicas e produtivas, como PIB per capita, área plantada e valor da produção vegetal, apresentaram efeitos marginais discretos. Por exemplo, um aumento de 10% no PIB per capita elevou o IDEB simulado em até 0,028 ponto, indicando que avanços econômicos isolados não se traduzem automaticamente em melhorias educacionais expressivas.

Os repasses do Programa Criança Feliz apresentaram pequenas reduções no IDEB previsto, entre -0,01 e -0,02 ponto. Esse efeito sugere que aumentos financeiros isolados não alteram substancialmente o desempenho médio, podendo refletir associações com maior vulnerabilidade social.

Tabela 8 – Impacto marginal das variáveis nos anos iniciais para o município de Caracol (IDEB real 2023 = 3,5; IDEB previsto sem alteração = 3,8798)

Variável	Cenário	Prev. cenário	Δ em relação ao IDEB Prev. sem alt.	Δ em relação ao IDEB Real
Distorção idade-série	-2%	3,8876	0,0078	0,3876
	-5%	3,8876	0,0078	0,3876
	-10%	3,9063	0,0265	0,4063
	-20%	4,0786	0,1988	0,5786
	-30%	4,1294	0,2496	0,6294
% de docentes sem formação superior	-2%	4,2024	0,3226	0,7024
	-5%	4,2024	0,3226	0,7024
	-10%	4,2024	0,3226	0,7024
	-20%	4,2024	0,3226	0,7024
	-30%	4,2024	0,3226	0,7024
PIB per capita	+1%	3,8812	0,0014	0,3812
	+3%	3,8802	0,0004	0,3802
	+5%	3,8831	0,0033	0,3831
	+7%	3,8831	0,0033	0,3831
	+10%	3,9080	0,0282	0,4080
Valor da produção na extração vegetal	+2%	3,8798	0	0,3798
	+5%	3,8798	0	0,3798
	+10%	3,8798	0	0,3798
	+20%	3,8798	0	0,3798
	+30%	3,8798	0	0,3798
Área plantada ou destinada à colheita (lavouras temporárias e permanentes)	+2%	3,8808	0,0010	0,3808
	+5%	3,8808	0,0010	0,3808
	+10%	3,8824	0,0026	0,3824
	+20%	3,8812	0,0014	0,3812
	+30%	3,8812	0,0014	0,3812
Valor repassado por meio do Programa Criança Feliz	+2%	3,8622	-0,0176	0,3622
	+5%	3,8707	-0,0091	0,3707
	+10%	3,8727	-0,0071	0,3727
	+20%	3,8688	-0,0110	0,3688
	+30%	3,8707	-0,0091	0,3707

Fonte: Elaborado pela autora (2025).

A Tabela 9, referente ao município de Francisco Macedo, indica um padrão de estabilidade nos anos iniciais do ensino fundamental. Em 2023, o IDEB real foi de 5,3, enquanto o valor previsto pelo modelo, sem intervenção, foi de 5,8084. As simulações mostram que as variações percentuais aplicadas às variáveis analisadas produzem efeitos

marginais próximos de zero, sugerindo baixa sensibilidade do desempenho educacional a mudanças pontuais nesses fatores.

A redução da distorção idade-série gera apenas oscilações positivas inferiores a 0,01 ponto no IDEB previsto em todos os cenários testados, indicando que o fluxo escolar já se encontra em nível elevado de regularidade. De forma semelhante, a proporção de docentes sem formação superior não apresenta efeito mensurável, o que aponta para um quadro docente consolidado.

As variáveis econômicas e produtivas, como PIB per capita e valor da produção na extração vegetal, mantêm impacto praticamente nulo sobre o IDEB simulado. A área plantada ou destinada à colheita apresenta pequenas variações positivas e negativas, sem efeito consistente sobre o desempenho educacional.

Os repasses do Programa Criança Feliz permanecem inalterados em todos os cenários simulados, reforçando a ausência de associação direta com o IDEB municipal. Em conjunto, os resultados indicam que o desempenho de Francisco Macedo apresenta baixa resposta a variações moderadas nas variáveis consideradas, refletindo um contexto educacional relativamente estabilizado.

A Tabela 10, referente ao município de Domingos Mourão, indica subestimação do modelo, uma vez que o IDEB previsto sem alteração (6,61) é inferior ao valor real observado em 2023 (8,4). Essa diferença sugere que, embora o modelo capture a tendência geral de desempenho, não incorpora plenamente fatores contextuais específicos associados ao resultado elevado do município.

As simulações mostram que a distorção idade-série e a proporção de docentes sem formação superior não geram variações no IDEB estimado, indicando estabilidade desses componentes e baixo potencial de ganho adicional nessas dimensões. O PIB per capita e o valor da produção vegetal apresentam reduções discretas nos cenários de aumento, com variações negativas inferiores a 0,04 ponto, caracterizando uma associação inversa de baixa magnitude.

A área plantada ou destinada à colheita apresenta oscilações residuais, sem padrão consistente. De forma semelhante, os repasses do Programa Criança Feliz exibem variações pequenas e irregulares. Em conjunto, os resultados indicam que, em contextos de alto desempenho como Domingos Mourão, o IDEB é pouco sensível às variáveis socioeconômicas consideradas, sendo mais influenciado por fatores institucionais e pedagógicos não explicitamente modelados.

Tabela 9 – Impacto marginal das variáveis nos anos iniciais para o município de Francisco Macedo (IDEB real 2023 = 5,3; IDEB previsto sem alteração = 5,8084)

Variável	Cenário	Prev. cenário	Δ em relação ao IDEB Prev. sem alt.	Δ em relação ao IDEB Real
Distorção idade-série	-2%	5,8067	-0,0017	0,5067
	-5%	5,8161	0,0077	0,5161
	-10%	5,8161	0,0077	0,5161
	-20%	5,8151	0,0067	0,5151
	-30%	5,8076	-0,0008	0,5076
% de docentes sem formação superior	-2%	5,8084	0	0,5084
	-5%	5,8084	0	0,5084
	-10%	5,8084	0	0,5084
	-20%	5,8084	0	0,5084
	-30%	5,8084	0	0,5084
PIB per capita	+1%	5,8074	-0,0010	0,5074
	+3%	5,8023	-0,0061	0,5023
	+5%	5,8023	-0,0061	0,5023
	+7%	5,8023	-0,0061	0,5023
	+10%	5,8023	-0,0061	0,5023
Valor da produção na extração vegetal	+2%	5,8084	0	0,5084
	+5%	5,8084	0	0,5084
	+10%	5,8084	0	0,5084
	+20%	5,8084	0	0,5084
	+30%	5,8069	-0,0015	0,5069
Área plantada ou destinada à colheita (lavouras temporárias e permanentes)	+2%	5,8154	0,0070	0,5154
	+5%	5,8188	0,0104	0,5188
	+10%	5,8333	0,0249	0,5333
	+20%	5,8020	-0,0064	0,5020
	+30%	5,8070	-0,0014	0,5070
Valor repassado por meio do Programa Criança Feliz	+2%	5,8084	0	0,5084
	+5%	5,8084	0	0,5084
	+10%	5,8084	0	0,5084
	+20%	5,8084	0	0,5084
	+30%	5,8084	0	0,5084

Fonte: Elaborado pela autora (2025).

A Tabela 11, referente ao município de Barreiras do Piauí, apresenta os resultados das simulações para os anos finais do ensino fundamental. Em 2023, o IDEB real foi de 2,6, enquanto o valor previsto pelo modelo, sem intervenção, foi de 3,6210. As simulações indicam que a redução da distorção idade-série é a variável com maior efeito marginal, produzindo incrementos de até 0,53 ponto no IDEB estimado, evidenciando a centralidade do fluxo escolar regular para a melhoria do desempenho.

A qualificação docente também apresenta impacto positivo, com ganhos de aproximadamente 0,23 ponto no IDEB simulado, indicando que avanços na formação dos professores tendem a gerar efeitos diretos em contextos de baixo desempenho. Os repasses do Programa Criança Feliz exibem efeito positivo de pequena magnitude, sugerindo contribuição complementar, possivelmente associada à melhoria indireta das condições familiares dos estudantes.

Em contraste, variáveis econômicas e produtivas, como PIB per capita, área plantada

Tabela 10 – Impacto marginal das variáveis nos anos iniciais para o município de Domingos Mourão (IDEB real 2023 = 8,4; IDEB previsto sem alteração = 6,6108)

Variável	Cenário	Prev. cenário	Δ em relação ao IDEB Prev. sem alt.	Δ em relação ao IDEB Real
Distorção idade-série	-2%	6,6108	0	-1,7892
	-5%	6,6108	0	-1,7892
	-10%	6,6108	0	-1,7892
	-20%	6,6108	0	-1,7892
	-30%	6,6108	0	-1,7892
% de docentes sem formação superior	-2%	6,6108	0	-1,7892
	-5%	6,6108	0	-1,7892
	-10%	6,6108	0	-1,7892
	-20%	6,6108	0	-1,7892
	-30%	6,6108	0	-1,7892
PIB per capita	+1%	6,6108	0	-1,7892
	+3%	6,6021	-0,0087	-1,7979
	+5%	6,6021	-0,0087	-1,7979
	+7%	6,6021	-0,0087	-1,7979
	+10%	6,6011	-0,0097	-1,7989
Valor da produção na extração vegetal	+2%	6,5830	-0,0278	-1,8170
	+5%	6,5765	-0,0343	-1,8235
	+10%	6,5765	-0,0343	-1,8235
	+20%	6,5765	-0,0343	-1,8235
	+30%	6,5873	-0,0235	-1,8127
Área plantada ou destinada à colheita (lavouras temporárias e permanentes)	+2%	6,6108	0	-1,7892
	+5%	6,6108	0	-1,7892
	+10%	6,6206	0,0098	-1,7794
	+20%	6,6003	-0,0105	-1,7997
	+30%	6,5876	-0,0232	-1,8124
Valor repassado por meio do Programa Criança Feliz	+2%	6,6108	0	-1,7892
	+5%	6,6108	0	-1,7892
	+10%	6,6108	0	-1,7892
	+20%	6,7024	0,0916	-1,6976
	+30%	6,5471	-0,0637	-1,8529

Fonte: Elaborado pela autora (2025).

e valor da produção vegetal, apresentam efeitos marginais reduzidos ou neutros. Em conjunto, os resultados indicam que, em Barreiras do Piauí, os principais determinantes do desempenho educacional estão associados a fatores internos ao sistema de ensino, especialmente aqueles relacionados ao fluxo escolar e à qualificação docente.

A Tabela 12, referente ao município de São João da Serra, apresenta os resultados das simulações para os anos finais do ensino fundamental. Em 2023, o IDEB real foi de 4,7, enquanto o valor previsto sem intervenção foi de 4,9859. As simulações indicam que a formação docente é a variável com maior efeito positivo, produzindo ganhos de até 0,14 ponto no IDEB estimado, evidenciando sua importância em contextos de desempenho intermediário.

O PIB per capita e a área plantada ou destinada à colheita apresentam efeitos positivos de pequena magnitude, sugerindo associação indireta com o desempenho educacional. Esses resultados indicam que fatores econômicos exercem influência limitada e tendem a

Tabela 11 – Impacto marginal das variáveis nos anos finais para o município de Barreiras do Piauí (IDEB real 2023 = 2,6; IDEB previsto sem alteração = 3,6210)

Variável	Cenário	Prev. cenário	Δ em relação ao IDEB Prev. sem alt.	Δ em relação ao IDEB Real
Distorção idade-série	-2%	3,6296	0,0086	1,0296
	-5%	3,6210	0	1,0210
	-10%	3,7370	0,1160	1,1370
	-20%	3,9688	0,3478	1,3688
	-30%	4,1569	0,5359	1,5569
% de docentes sem formação superior	-2%	3,8525	0,2315	1,2525
	-5%	3,8525	0,2315	1,2525
	-10%	3,8525	0,2315	1,2525
	-20%	3,8525	0,2315	1,2525
	-30%	3,8525	0,2315	1,2525
PIB per capita	+1%	3,6210	0	1,0210
	+3%	3,6210	0	1,0210
	+5%	3,6210	0	1,0210
	+7%	3,6210	0	1,0210
	+10%	3,6261	0,0051	1,0261
Valor da produção na extração vegetal	+2%	3,6210	0	1,0210
	+5%	3,6210	0	1,0210
	+10%	3,6058	-0,0152	1,0058
	+20%	3,6058	-0,0152	1,0058
	+30%	3,6071	-0,0139	1,0071
Área plantada ou destinada à colheita (lavouras temporárias e permanentes)	+2%	3,6210	0	1,0210
	+5%	3,6210	0	1,0210
	+10%	3,6210	0	1,0210
	+20%	3,6210	0	1,0210
	+30%	3,6210	0	1,0210
Valor repassado por meio do Programa Criança Feliz	+2%	3,6345	0,0135	1,0345
	+5%	3,6345	0,0135	1,0345
	+10%	3,6887	0,0677	1,0887
	+20%	3,6930	0,0720	1,0930
	+30%	3,6966	0,0756	1,0966

Fonte: Elaborado pela autora (2025).

atuar de forma estrutural, sem impacto expressivo em variações pontuais.

A distorção idade-série não gerou alterações relevantes no IDEB simulado, indicando que o fluxo escolar já se encontra em nível satisfatório. Os repasses do Programa Criança Feliz apresentaram efeitos levemente negativos, possivelmente associados a características contextuais dos municípios atendidos, sem evidência de impacto direto sobre o desempenho educacional.

De forma geral, os resultados indicam que, em São João da Serra, o aprimoramento da formação docente constitui o principal fator associado a avanços no IDEB, enquanto as variáveis econômicas e sociais apresentam efeitos secundários ou inconsistentes.

Tabela 12 – Impacto marginal das variáveis nos anos finais para o município de São João da Serra (IDEB real 2023 = 4,7; IDEB previsto sem alteração = 4,9859)

Variável	Cenário	Prev. cenário	Δ em relação ao IDEB Prev. sem alt.	Δ em relação ao IDEB Real
Distorção idade-série	-2%	4,9859	0	0,2859
	-5%	4,9859	0	0,2859
	-10%	4,9861	0,0002	0,2861
	-20%	4,9870	0,0011	0,2870
	-30%	4,9870	0,0011	0,2870
% de docentes sem formação superior	-2%	5,1296	0,1437	0,4296
	-5%	5,1296	0,1437	0,4296
	-10%	5,1296	0,1437	0,4296
	-20%	5,1296	0,1437	0,4296
	-30%	5,1296	0,1437	0,4296
PIB per capita	+1%	4,9859	0	0,2859
	+3%	4,9859	0	0,2859
	+5%	4,9859	0	0,2859
	+7%	4,9866	0,0007	0,2866
	+10%	4,9906	0,0047	0,2906
Valor da produção na extração vegetal	+2%	4,9859	0	0,2859
	+5%	4,9859	0	0,2859
	+10%	4,9859	0	0,2859
	+20%	4,9859	0	0,2859
	+30%	4,9859	0	0,2859
Área plantada ou destinada à colheita (lavouras temporárias e permanentes)	+2%	4,9859	0	0,2859
	+5%	4,9828	-0,0031	0,2828
	+10%	4,9841	-0,0018	0,2841
	+20%	5,0171	0,0312	0,3171
	+30%	5,0245	0,0386	0,3245
Valor repassado por meio do Programa Criança Feliz	+2%	4,9844	-0,0015	0,2844
	+5%	4,9844	-0,0015	0,2844
	+10%	4,9816	-0,0043	0,2816
	+20%	4,9816	-0,0043	0,2816
	+30%	4,9577	-0,0282	0,2577

Fonte: Elaborado pela autora (2025).

A Tabela 13, referente ao município de Buriti dos Montes, apresenta os resultados das simulações para os anos finais do ensino fundamental. Em 2023, o IDEB real foi de 6,5, enquanto o valor previsto sem intervenção foi de 5,2890, caracterizando subestimação do modelo. Essa diferença indica a influência de fatores locais não capturados pelas variáveis consideradas.

As variáveis relacionadas à formação docente e aos repasses do Programa Criança Feliz apresentaram efeitos positivos de pequena magnitude. A melhoria na qualificação dos professores elevou o IDEB estimado em cerca de 0,01 ponto, enquanto os repasses do programa social resultaram em acréscimos próximos de 0,07 ponto, indicando contribuição complementar, porém insuficiente para explicar o desempenho observado.

Em sentido oposto, o valor da produção vegetal apresentou efeito negativo consistente, com reduções graduais no IDEB previsto à medida que a variável aumenta. O PIB per capita e a área plantada exibiram efeitos marginais e instáveis, sem padrão definido.

Em conjunto, os resultados sugerem que o desempenho superior de Buriti dos Montes está associado a fatores institucionais e pedagógicos não explicitamente modelados, reforçando a necessidade de análises complementares para compreender contextos de alto desempenho.

Tabela 13 – Impacto marginal das variáveis nos anos finais para o município de Buriti dos Montes (IDEB real 2023 = 6,5; IDEB previsto sem alteração = 5,2890)

Variável	Cenário	Prev. cenário	Δ em relação ao IDEB Prev. sem alt.	Δ em relação ao IDEB Real
Distorção idade-série	-2%	5,2890	0	-1,2110
	-5%	5,2890	0	-1,2110
	-10%	5,2890	0	-1,2110
	-20%	5,2823	-0,0067	-1,2177
	-30%	5,2823	-0,0067	-1,2177
% de docentes sem formação superior	-2%	5,3003	0,0113	-1,1997
	-5%	5,3003	0,0113	-1,1997
	-10%	5,3003	0,0113	-1,1997
	-20%	5,3003	0,0113	-1,1997
	-30%	5,3003	0,0113	-1,1997
PIB per capita	+1%	5,2890	0	-1,2110
	+3%	5,2890	0	-1,2110
	+5%	5,2891	0,0001	-1,2109
	+7%	5,2891	0,0001	-1,2109
	+10%	5,2912	0,0022	-1,2088
Valor da produção na extração vegetal	+2%	5,2890	0	-1,2110
	+5%	5,2622	-0,0268	-1,2378
	+10%	5,2534	-0,0356	-1,2466
	+20%	5,2508	-0,0382	-1,2492
	+30%	5,2442	-0,0448	-1,2558
Área plantada ou destinada à colheita (lavouras temporárias e permanentes)	+2%	5,2896	0,0006	-1,2104
	+5%	5,2904	0,0014	-1,2096
	+10%	5,2975	0,0085	-1,2025
	+20%	5,2853	-0,0037	-1,2147
	+30%	5,2829	-0,0061	-1,2171
Valor repassado por meio do Programa Criança Feliz	+2%	5,3595	0,0705	-1,1405
	+5%	5,3595	0,0705	-1,1405
	+10%	5,3595	0,0705	-1,1405
	+20%	5,3609	0,0719	-1,1391
	+30%	5,3609	0,0719	-1,1391

Fonte: Elaborado pela autora (2025).

6.4.2 Simulações do Ensino Médio

A análise das simulações contrafactuais referentes ao ensino médio foi organizada em três blocos correspondentes aos municípios selecionados com base no critério de desempenho no IDEB de 2023, representando, respectivamente, o pior, o desempenho mediano e o melhor desempenho dessa etapa: Capitão de Campos, Simplicio Mendes e Cocal dos Alves. Cada município é apresentado por meio de quatro tabelas, estruturadas de modo a permitir leitura sistemática e comparação entre dimensões distintas.

Os resultados para Capitão de Campos estão apresentados nas Tabelas 14 a 17, distribuídos em: (i) infraestrutura escolar, (ii) características docentes e pedagógicas, (iii) rendimento e jornada escolar e (iv) fatores socioeconômicos e produtivos. As Tabelas 18 a 21 seguem a mesma organização para Simplício Mendes, enquanto as Tabelas 22 a 25 sintetizam os resultados para Cocal dos Alves. Essa padronização assegura comparabilidade entre os diferentes perfis territoriais analisados.

As Tabelas 14 a 17 referem-se ao município de Capitão de Campos, que apresentou o menor IDEB no ensino médio em 2023 (3,1), com valor previsto sem intervenção de 4,4100. Observa-se que, embora algumas variáveis indiquem potencial de melhoria, o desempenho real permanece inferior ao estimado, sugerindo a influência de fatores contextuais não contemplados no modelo.

As variáveis de infraestrutura escolar (Tabela 14) não produziram alterações significativas no IDEB simulado, mesmo em cenários de ampliação de até 30%. O aumento no número de escolas com lousa digital, laboratório de informática ou grêmio estudantil não resultou em variações perceptíveis, indicando baixa sensibilidade do modelo a esse conjunto de fatores.

Tabela 14 – Impacto marginal das variáveis de infraestrutura escolar no ensino médio para o município de Capitão de Campos (IDEB real 2023 = 3,1; IDEB previsto sem alteração = 4,4100)

Variável	Cenário	Prev. cenário	Δ em relação ao IDEB Prev. sem alt.	Δ em relação ao IDEB Real
Quantidade de escolas com lousa digital	+2%	4,4100	0	1,3100
	+5%	4,4100	0	1,3100
	+10%	4,4100	0	1,3100
	+20%	4,4100	0	1,3100
	+30%	4,4100	0	1,3100
Quantidade de escolas com laboratório de informática	+2%	4,4100	0	1,3100
	+5%	4,4100	0	1,3100
	+10%	4,4100	0	1,3100
	+20%	4,4100	0	1,3100
	+30%	4,4100	0	1,3100
Quantidade de escolas com órgão de grêmio estudantil	+2%	4,4100	0	1,3100
	+5%	4,4100	0	1,3100
	+10%	4,4100	0	1,3100
	+20%	4,4100	0	1,3100
	+30%	4,4100	0	1,3100

Fonte: Elaborado pela autora (2025).

As variáveis relacionadas ao corpo docente (Tabela 15) apresentaram os efeitos mais evidentes, ainda que de pequena magnitude. O aumento do percentual de professores com curso superior completo elevou o IDEB estimado em até 0,05 ponto, enquanto a redução do esforço docente também gerou ganhos discretos. A diminuição da regularidade média do corpo docente produziu efeito negativo, indicando a importância da permanência dos professores na escola. A variável associada ao nível de gestão escolar apresentou efeito nulo.

Tabela 15 – Impacto marginal das variáveis de características docentes e pedagógicas no ensino médio para o município de Capitão de Campos (IDEB real 2023 = 3,1; IDEB previsto sem alteração = 4,4100)

Variável	Cenário	Prev. cenário	Δ em relação ao IDEB Prev. sem alt.	Δ em relação ao IDEB Real
% de docentes no Nível 6 de esforço docente	-2%	4,4100	0	1,3100
	-5%	4,4100	0	1,3100
	-10%	4,4100	0	1,3100
	-20%	4,4100	0	1,3100
	-30%	4,4235	0,0135	1,3235
Professores com formação em pedagogia	+2%	4,4100	0	1,3100
	+5%	4,4084	-0,0016	1,3084
	+10%	4,4084	-0,0016	1,3084
	+20%	4,4084	-0,0016	1,3084
	+30%	4,4084	-0,0016	1,3084
% de docentes com curso superior completo	+2%	4,4274	0,0174	1,3274
	+5%	4,4482	0,0382	1,3482
	+10%	4,4526	0,0426	1,3526
	+20%	4,4609	0,0509	1,3609
	+30%	4,4609	0,0509	1,3609
Regularidade média do corpo docente da escola (“Média-Baixa”)	-2%	4,3900	-0,0200	1,2900
	-5%	4,3900	-0,0200	1,2900
	-10%	4,3900	-0,0200	1,2900
	-20%	4,3900	-0,0200	1,2900
	-30%	4,3941	-0,0159	1,2941
% de escolas no Nível 5 de gestão escolar	+2%	4,4100	0	1,3100
	+5%	4,4100	0	1,3100
	+10%	4,4100	0	1,3100
	+20%	4,4100	0	1,3100
	+30%	4,4085	-0,0015	1,3085

Fonte: Elaborado pela autora (2025).

No eixo de rendimento e jornada escolar (Tabela 16), o aumento da média de horas-aula diárias foi a variável com maior impacto, elevando o IDEB simulado em até 0,21 ponto. A redução da distorção idade-série e da média de alunos por turma gerou apenas variações positivas de pequena magnitude.

Por fim, as variáveis socioeconômicas e produtivas (Tabela 17) apresentaram efeitos marginais. O aumento da área plantada e do efetivo de rebanhos resultou em impactos positivos discretos, enquanto o crescimento das transferências da Proteção Social Básica produziu efeito levemente negativo.

Em síntese, os resultados indicam que, no ensino médio de Capitão de Campos, o modelo é pouco sensível a variações em infraestrutura e gestão escolar, respondendo de forma mais consistente a mudanças relacionadas à formação docente, estabilidade profissional e jornada escolar. Esses achados sugerem que os desafios locais extrapolam a dimensão estrutural, envolvendo aspectos pedagógicos e institucionais que demandam políticas integradas voltadas à permanência, ao engajamento estudantil e ao fortalecimento do tempo pedagógico efetivo.

Tabela 16 – Impacto marginal das variáveis de rendimento e jornada escolar no ensino médio para o município de Capitão de Campos (IDEB real 2023 = 3,1; IDEB previsto sem alteração = 4,4100)

Variável	Cenário	Prev. cenário	Δ em relação ao IDEB Prev. sem alt.	Δ em relação ao IDEB Real
Taxa de distorção idade-série	-2%	4,4100	0	1,3100
	-5%	4,4105	0,0005	1,3105
	-10%	4,4118	0,0018	1,3118
	-20%	4,4118	0,0018	1,3118
	-30%	4,4131	0,0031	1,3131
Média de alunos por turma	-2%	4,4100	0	1,3100
	-5%	4,4110	0,0010	1,3110
	-10%	4,4057	-0,0043	1,3057
	-20%	4,3678	-0,0422	1,2678
	-30%	4,3632	-0,0468	1,2632
Média de horas-aula diária	+2%	4,4145	0,0045	1,3145
	+5%	4,4145	0,0045	1,3145
	+10%	4,4313	0,0213	1,3313
	+20%	4,4303	0,0203	1,3303
	+30%	4,6201	0,2101	1,5201

Fonte: Elaborado pela autora (2025).

As Tabelas 18 a 21 apresentam os resultados das simulações para o município de Simplício Mendes, que ocupou posição intermediária no IDEB do ensino médio em 2023, com valor real de 4,1. O IDEB previsto pelo modelo, sem intervenção, foi de 4,5053, indicando leve superestimação. As variações simuladas produziram efeitos muito reduzidos, em torno de $\pm 0,02$ ponto, evidenciando estabilidade das condições educacionais locais.

As variáveis de infraestrutura escolar (Tabela 18) não apresentaram impacto perceptível no IDEB simulado, mesmo em cenários de ampliação de até 30%. O aumento do número de escolas com lousa digital, laboratório de informática ou grêmio estudantil não resultou em alterações relevantes, indicando baixa sensibilidade do desempenho a esse conjunto de fatores.

No eixo docente e pedagógico (Tabela 19), a regularidade do corpo docente foi a variável com maior efeito, ainda que de pequena magnitude, elevando o IDEB previsto em até 0,0289 ponto. As demais variáveis, como formação superior, esforço docente e nível de gestão escolar, apresentaram efeitos nulos ou residuais, sugerindo equilíbrio do quadro docente municipal.

As variáveis de rendimento e jornada escolar (Tabela 20) mostraram leve sensibilidade à ampliação do tempo pedagógico. O aumento da média de horas-aula diárias foi o fator com maior impacto, elevando o IDEB simulado em até 0,21 ponto. A redução da média de alunos por turma e da distorção idade-série produziu variações próximas de zero.

Por fim, as variáveis socioeconômicas e produtivas (Tabela 21) apresentaram efeitos marginais. A ampliação da área plantada gerou pequenos acréscimos no IDEB previsto, enquanto o aumento dos repasses da Proteção Social Básica resultou em reduções discretas, padrão também observado em outros municípios.

Tabela 17 – Impacto marginal das variáveis de fatores socioeconômicos e produtivos no ensino médio para o município de Capitão de Campos (IDEB real 2023 = 3,1; IDEB previsto sem alteração = 4,4100)

Variável	Cenário	Prev. cenário	Δ em relação ao IDEB Prev. sem alt.	Δ em relação ao IDEB Real
PIB per capita	+1%	4,4100	0	1,3100
	+3%	4,4100	0	1,3100
	+5%	4,4115	0,0015	1,3115
	+7%	4,4115	0,0015	1,3115
	+10%	4,4105	0,0005	1,3105
Área plantada ou destinada à colheita (lavouras temporárias e permanentes)	+2%	4,4245	0,0145	1,3245
	+5%	4,4245	0,0145	1,3245
	+10%	4,4245	0,0145	1,3245
	+20%	4,4245	0,0145	1,3245
	+30%	4,4269	0,0169	1,3269
Valor da produção na extração vegetal	+2%	4,4100	0	1,3100
	+5%	4,4100	0	1,3100
	+10%	4,4135	0,0035	1,3135
	+20%	4,4135	0,0035	1,3135
	+30%	4,4145	0,0045	1,3145
Efetivo total dos rebanhos	+2%	4,4100	0	1,3100
	+5%	4,4100	0	1,3100
	+10%	4,4225	0,0125	1,3225
	+20%	4,4285	0,0185	1,3285
	+30%	4,4162	0,0062	1,3162
Valor total repassado – Proteção Social Básica	+2%	4,4100	0	1,3100
	+5%	4,4099	-0,0001	1,3099
	+10%	4,4045	-0,0055	1,3045
	+20%	4,3834	-0,0266	1,2834
	+30%	4,3773	-0,0327	1,2773
Transferências voluntárias realizadas – Fundo a Fundo	+2%	4,4100	0	1,3100
	+5%	4,4100	0	1,3100
	+10%	4,4100	0	1,3100
	+20%	4,4100	0	1,3100
	+30%	4,4100	0	1,3100

Fonte: Elaborado pela autora (2025).

Em conjunto, os resultados indicam que Simpício Mendes apresenta um cenário de estabilidade institucional, no qual o modelo responde de forma limitada às variações simuladas. A regularidade do corpo docente e a ampliação do tempo pedagógico constituem os principais fatores associados a possíveis ganhos no desempenho do ensino médio.

Tabela 18 – Impacto marginal das variáveis de infraestrutura escolar no ensino médio para o município de Simplício Mendes (IDEB real 2023 = 4,1; IDEB previsto sem alteração = 4,5053)

Variável	Cenário	Prev. cenário	Δ em relação ao IDEB Prev. sem alt.	Δ em relação ao IDEB Real
Quantidade de escolas com lousa digital	+2%	4,5053	0	0,4053
	+5%	4,5053	0	0,4053
	+10%	4,5053	0	0,4053
	+20%	4,5053	0	0,4053
	+30%	4,5053	0	0,4053
Quantidade de escolas com laboratório de informática	+2%	4,5053	0	0,4053
	+5%	4,5053	0	0,4053
	+10%	4,5053	0	0,4053
	+20%	4,4991	-0,0062	0,3991
	+30%	4,4991	-0,0062	0,3991
Quantidade de escolas com órgão de grêmio estudantil	+2%	4,5053	0	0,4053
	+5%	4,5053	0	0,4053
	+10%	4,5053	0	0,4053
	+20%	4,5053	0	0,4053
	+30%	4,5053	0	0,4053

Fonte: Elaborado pela autora (2025).

Tabela 19 – Impacto marginal das variáveis de características docentes e pedagógicas no ensino médio para o município de Simplício Mendes (IDEB real 2023 = 4,1; IDEB previsto sem alteração = 4,5053)

Variável	Cenário	Prev. cenário	Δ em relação ao IDEB Prev. sem alt.	Δ em relação ao IDEB Real
% de docentes no Nível 6 de esforço docente	-2%	4,5053	0	0,4053
	-5%	4,5053	0	0,4053
	-10%	4,5053	0	0,4053
	-20%	4,5053	0	0,4053
	-30%	4,5053	0	0,4053
Professores com formação em pedagogia	+2%	4,5053	0	0,4053
	+5%	4,5053	0	0,4053
	+10%	4,5053	0	0,4053
	+20%	4,5053	0	0,4053
	+30%	4,5053	0	0,4053
% de docentes com curso superior completo	+2%	4,5053	0	0,4053
	+5%	4,5053	0	0,4053
	+10%	4,5053	0	0,4053
	+20%	4,5053	0	0,4053
	+30%	4,5053	0	0,4053
Regularidade média do corpo docente da escola (“Média-Baixa”)	-2%	4,5053	0	0,4053
	-5%	4,5065	0,0012	0,4065
	-10%	4,5205	0,0152	0,4205
	-20%	4,5342	0,0289	0,4342
	-30%	4,5265	0,0212	0,4265
% de escolas no Nível 5 de gestão escolar	+2%	4,5053	0	0,4053
	+5%	4,5053	0	0,4053
	+10%	4,5053	0	0,4053
	+20%	4,5053	0	0,4053
	+30%	4,5053	0	0,4053

Fonte: Elaborado pela autora (2025).

Tabela 20 – Impacto marginal das variáveis de rendimento e jornada escolar no ensino médio para o município de Simplício Mendes (IDEB real 2023 = 4,1; IDEB previsto sem alteração = 4,5053)

Variável	Cenário	Prev. cenário	Δ em relação ao IDEB Prev. sem alt.	Δ em relação ao IDEB Real
Taxa de distorção idade-série	-2%	4,5053	0	0,4053
	-5%	4,5053	0	0,4053
	-10%	4,5053	0	0,4053
	-20%	4,5063	0,0010	0,4063
	-30%	4,5127	0,0074	0,4127
Média de alunos por turma	-2%	4,5063	0,0010	0,4063
	-5%	4,5063	0,0010	0,4063
	-10%	4,4897	-0,0156	0,3897
	-20%	4,4348	-0,0705	0,3348
	-30%	4,4348	-0,0705	0,3348
Média de horas-aula diária	+2%	4,5053	0	0,4053
	+5%	4,5028	-0,0025	0,4028
	+10%	4,5076	0,0023	0,4076
	+20%	4,7077	0,2024	0,6077
	+30%	4,7157	0,2104	0,6157

Fonte: Elaborado pela autora (2025).

As Tabelas 22 a 25 sintetizam os resultados das simulações para o município de Cocal dos Alves, que apresentou o melhor desempenho no ensino médio em 2023, com IDEB igual a 7,0 pontos. O modelo estimou um valor previsto de 4,9919, indicando subestimação do desempenho real e revelando que o alto rendimento educacional do município decorre de fatores qualitativos e institucionais não capturados integralmente pelas variáveis estruturais e socioeconômicas. De modo geral, as variações simuladas produziram efeitos marginais e estáveis, característica esperada em sistemas educacionais já consolidados.

Nas variáveis de infraestrutura escolar (Tabela 22), não se observaram alterações no IDEB previsto, mesmo em cenários de ampliação de até 30% no número de escolas com lousa digital, laboratório de informática e grêmio estudantil. Esse comportamento indica maturidade e estabilidade dessa dimensão, sugerindo que a infraestrutura material atingiu um ponto de saturação em que incrementos adicionais não se traduzem em ganhos de aprendizagem mensuráveis. Esse resultado é coerente com o fato de Cocal dos Alves ser um caso consolidado de eficiência educacional, no qual o desempenho já reflete o uso otimizado dos recursos existentes e não apenas sua disponibilidade quantitativa.

As variáveis docentes e pedagógicas (Tabela 23) apresentaram pequena sensibilidade negativa a reduções na regularidade do corpo docente e no esforço docente, resultando em decréscimos inferiores a 0,01 ponto no IDEB previsto. Esses valores reforçam a resiliência do sistema educacional local, capaz de manter alta performance mesmo diante de variações moderadas em indicadores de pessoal. Esse padrão sugere que o município possui uma cultura pedagógica institucionalizada e mecanismos de gestão escolar capazes de mitigar eventuais flutuações na composição do corpo docente. O percentual de professores com formação superior e a proporção de escolas em níveis elevados de gestão mostraram efeito

Tabela 21 – Impacto marginal das variáveis de fatores socioeconômicos e produtivos no ensino médio para o município de Simplício Mendes (IDEB real 2023 = 4,1; IDEB previsto sem alteração = 4,5053)

Variável	Cenário	Prev. cenário	Δ em relação ao IDEB Prev. sem alt.	Δ em relação ao IDEB Real
PIB per capita	+1%	4,5053	0	0,4053
	+3%	4,5053	0	0,4053
	+5%	4,5019	-0,0034	0,4019
	+7%	4,5053	0	0,4053
	+10%	4,5008	-0,0045	0,4008
Área plantada ou destinada à colheita (lavouras temporárias e permanentes)	+2%	4,5053	0	0,4053
	+5%	4,5199	0,0146	0,4199
	+10%	4,5199	0,0146	0,4199
	+20%	4,5186	0,0133	0,4186
	+30%	4,5213	0,0160	0,4213
Valor da produção na extração vegetal	+2%	4,5053	0	0,4053
	+5%	4,5053	0	0,4053
	+10%	4,5053	0	0,4053
	+20%	4,5053	0	0,4053
	+30%	4,5053	0	0,4053
Efetivo total dos rebanhos	+2%	4,5053	0	0,4053
	+5%	4,5053	0	0,4053
	+10%	4,5053	0	0,4053
	+20%	4,5053	0	0,4053
	+30%	4,5053	0	0,4053
Valor total repassado – Proteção Social Básica	+2%	4,5024	-0,0029	0,4024
	+5%	4,5018	-0,0035	0,4018
	+10%	4,5003	-0,0050	0,4003
	+20%	4,4696	-0,0357	0,3696
	+30%	4,4696	-0,0357	0,3696
Transferências voluntárias realizadas – Fundo a Fundo	+2%	4,5053	0	0,4053
	+5%	4,5053	0	0,4053
	+10%	4,5053	0	0,4053
	+20%	4,5053	0	0,4053
	+30%	4,5053	0	0,4053

Fonte: Elaborado pela autora (2025).

nulo, o que evidencia homogeneidade e estabilidade nas condições docentes.

As variáveis de rendimento e jornada (Tabela 24) permaneceram praticamente inalteradas, com exceção de uma leve redução do IDEB simulado quando a média de alunos por turma diminuiu em 30%. Esse comportamento pode estar associado à perda de efeito de escala em escolas de pequeno porte, nas quais a redução do número de estudantes por turma não necessariamente implica melhoria das práticas pedagógicas. No conjunto, as evidências indicam que o tempo de instrução, a composição das turmas e o fluxo escolar encontram-se em níveis de equilíbrio que sustentam o bom desempenho observado.

Por outro lado, as variáveis socioeconômicas e produtivas (Tabela 25) apresentaram variações ligeiramente negativas, especialmente nos cenários de aumento da área plantada, da produção extrativista e do efetivo de rebanhos. Esse resultado pode refletir o perfil rural do município e a limitação dessas variáveis em explicar ganhos adicionais de desempenho

escolar em contextos já de alta eficácia pedagógica. A ampliação de atividades produtivas pode inclusive estar associada a deslocamentos de foco ou de mão de obra juvenil, o que reduz a correlação positiva entre crescimento econômico e aprendizado escolar.

Em síntese, as simulações para Cocal dos Alves confirmam o comportamento esperado de um sistema educacional consolidado, no qual o modelo estatístico demonstra estabilidade e baixa sensibilidade a variações incrementais. Esse resultado reforça a importância de combinar abordagens estatísticas com análises qualitativas para compreender as dinâmicas de excelência em sistemas educacionais de referência.

Tabela 22 – Impacto marginal das variáveis de infraestrutura escolar no ensino médio para o município de Cocal dos Alves (IDEB real 2023 = 7,0; IDEB previsto sem alteração = 4,9919)

Variável	Cenário	Prev. cenário	Δ em relação ao IDEB Prev. sem alt.	Δ em relação ao IDEB Real
Quantidade de escolas com lousa digital	+2%	4,9919	0	-2,0081
	+5%	4,9919	0	-2,0081
	+10%	4,9919	0	-2,0081
	+20%	4,9919	0	-2,0081
	+30%	4,9919	0	-2,0081
Quantidade de escolas com laboratório de informática	+2%	4,9919	0	-2,0081
	+5%	4,9919	0	-2,0081
	+10%	4,9919	0	-2,0081
	+20%	4,9919	0	-2,0081
	+30%	4,9919	0	-2,0081
Quantidade de escolas com órgão de grêmio estudantil	+2%	4,9919	0	-2,0081
	+5%	4,9919	0	-2,0081
	+10%	4,9919	0	-2,0081
	+20%	4,9919	0	-2,0081
	+30%	4,9919	0	-2,0081

Fonte: Elaborado pela autora (2025).

Tabela 23 – Impacto marginal das variáveis de características docentes e pedagógicas no ensino médio para o município de Cocal dos Alves (IDEB real 2023 = 7,0; IDEB previsto sem alteração = 4,9919)

Variável	Cenário	Prev. cenário	Δ em relação ao IDEB Prev. sem alt.	Δ em relação ao IDEB Real
% de docentes no Nível 6 de esforço docente	-2%	4,9919	0	-2,0081
	-5%	4,9849	-0,0070	-2,0151
	-10%	4,9849	-0,0070	-2,0151
	-20%	4,9835	-0,0084	-2,0165
	-30%	4,9835	-0,0084	-2,0165
Professores com formação em pedagogia	+2%	4,9919	0	-2,0081
	+5%	4,9919	0	-2,0081
	+10%	4,9919	0	-2,0081
	+20%	4,9888	-0,0031	-2,0112
	+30%	4,9756	-0,0163	-2,0244
% de docentes com curso superior completo	+2%	4,9919	0	-2,0081
	+5%	4,9919	0	-2,0081
	+10%	4,9919	0	-2,0081
	+20%	4,9919	0	-2,0081
	+30%	4,9919	0	-2,0081
Regularidade média do corpo docente da escola (“Média-Baixa”)	-2%	4,9898	-0,0021	-2,0102
	-5%	4,9898	-0,0021	-2,0102
	-10%	4,9824	-0,0095	-2,0176
	-20%	4,9826	-0,0093	-2,0174
	-30%	4,9897	-0,0022	-2,0103
% de escolas no Nível 5 de gestão escolar	+2%	4,9919	0	-2,0081
	+5%	4,9919	0	-2,0081
	+10%	4,9919	0	-2,0081
	+20%	4,9919	0	-2,0081
	+30%	4,9919	0	-2,0081

Fonte: Elaborado pela autora (2025).

Tabela 24 – Impacto marginal das variáveis de rendimento e jornada escolar no ensino médio para o município de Cocal dos Alves (IDEB real 2023 = 7,0; IDEB previsto sem alteração = 4,9919)

Variável	Cenário	Prev. cenário	Δ em relação ao IDEB Prev. sem alt.	Δ em relação ao IDEB Real
Taxa de distorção idade-série	-2%	4,9919	0	-2,0081
	-5%	4,9919	0	-2,0081
	-10%	4,9919	0	-2,0081
	-20%	4,9919	0	-2,0081
	-30%	4,9919	0	-2,0081
Média de alunos por turma	-2%	4,9930	0,0011	-2,0070
	-5%	4,9930	0,0011	-2,0070
	-10%	5,0001	0,0082	-1,9999
	-20%	4,9919	0	-2,0081
	-30%	4,9336	-0,0583	-2,0664
Média de horas-aula diária	+2%	4,9919	0	-2,0081
	+5%	4,9919	0	-2,0081
	+10%	4,9919	0	-2,0081
	+20%	4,9919	0	-2,0081
	+30%	4,9919	0	-2,0081

Fonte: Elaborado pela autora (2025).

Tabela 25 – Impacto marginal das variáveis de fatores socioeconômicos e produtivos no ensino médio para o município de Cocal dos Alves (IDEB real 2023 = 7,0; IDEB previsto sem alteração = 4,9919)

Variável	Cenário	Prev. cenário	Δ em relação ao IDEB Prev. sem alt.	Δ em relação ao IDEB Real
PIB per capita	+1%	4,9919	0	-2,0081
	+3%	4,9912	-0,0007	-2,0088
	+5%	4,9912	-0,0007	-2,0088
	+7%	4,9919	0	-2,0081
	+10%	4,9915	-0,0004	-2,0085
Área plantada ou destinada à colheita (lavouras temporárias e permanentes)	+2%	4,9906	-0,0013	-2,0094
	+5%	4,9892	-0,0027	-2,0108
	+10%	4,9268	-0,0651	-2,0732
	+20%	4,9040	-0,0879	-2,0960
	+30%	4,9013	-0,0906	-2,0987
Valor da produção na extração vegetal	+2%	4,9879	-0,0040	-2,0121
	+5%	4,9879	-0,0040	-2,0121
	+10%	4,9870	-0,0049	-2,0130
	+20%	4,9870	-0,0049	-2,0130
	+30%	4,9870	-0,0049	-2,0130
Efetivo total dos rebanhos	+2%	4,9768	-0,0151	-2,0232
	+5%	4,9768	-0,0151	-2,0232
	+10%	4,9463	-0,0456	-2,0537
	+20%	4,9476	-0,0443	-2,0524
	+30%	4,9476	-0,0443	-2,0524
Valor total repassado – Proteção Social Básica	+2%	4,9853	-0,0066	-2,0147
	+5%	4,9853	-0,0066	-2,0147
	+10%	4,9853	-0,0066	-2,0147
	+20%	4,9840	-0,0079	-2,0160
	+30%	4,9614	-0,0305	-2,0386
Transferências voluntárias realizadas – Fundo a Fundo	+2%	4,9919	0	-2,0081
	+5%	4,9919	0	-2,0081
	+10%	4,9919	0	-2,0081
	+20%	4,9919	0	-2,0081
	+30%	4,9919	0	-2,0081

Fonte: Elaborado pela autora (2025).

De forma geral, as simulações contrafactuais indicam maior sensibilidade do ensino fundamental a intervenções em variáveis educacionais diretas, especialmente formação docente e distorção idade-série. No ensino médio, os efeitos estimados são mais discretos, concentrando-se sobretudo na jornada escolar, o que reflete maior consolidação estrutural dessa etapa.

Nos municípios de baixo desempenho, como Caracol e Barreiras do Piauí, as maiores variações positivas do IDEB simulado decorrem da redução da distorção idade-série e da qualificação docente, com incrementos superiores a 0,5 ponto. Em contextos intermediários, como Francisco Macedo e Simplício Mendes, os impactos foram marginais, indicando menor margem para ganhos incrementais por meio de alterações isoladas. Já nos municípios de alto desempenho, como Domingos Mourão e Cocal dos Alves, as simulações não produziram mudanças relevantes, evidenciando saturação do indicador e limites dos modelos em captar

fatores qualitativos não observáveis.

Em síntese, os resultados confirmam que o potencial de ganho no IDEB é fortemente condicionado ao estágio de desenvolvimento educacional municipal. Formação docente, distorção idade-série e carga horária diária permanecem como os principais vetores de intervenção, especialmente em localidades de baixo desempenho. Destaca-se que os impactos estimados são marginais e isolados. A combinação simultânea de políticas pode gerar efeitos não lineares, hipótese que extrapola o escopo deste capítulo e aponta para investigações futuras.

Por fim, ressalta-se que as métricas de erro do modelo aplicam-se apenas ao cenário observado. Nos cenários simulados, os valores representam variações contrafactuais estimadas, devendo ser interpretados como exercícios analíticos de apoio à tomada de decisão, e não como previsões empíricas sujeitas a critérios de acurácia.

7 Considerações Finais

Esta dissertação teve como objetivo central o desenvolvimento e aplicação de modelos preditivos para estimar o IDEB piauiense em nível municipal, utilizando técnicas de ML e DL. Para tanto, foram organizadas bases de dados específicas para os anos iniciais, finais e ensino médio, abrangendo o período de 2013 a 2023 sendo que, no caso do ensino médio, os dados disponíveis iniciam-se em 2017.

As variáveis explicativas contemplaram dimensões educacionais, socioeconômicas e financeiras, extraídas de fontes oficiais nacionais, como INEP, IBGE e SIOPE. Foram avaliados dez modelos clássicos de ML e sete arquiteturas de DL, por meio de um pipeline de avaliação sistemática, que incluiu divisão treino/teste, padronização das variáveis, validação cruzada estratificada (*5-fold*) e análise comparativa com base em RMSE, MAE e R^2 .

Os resultados demonstraram que o XGBoost apresentou o melhor desempenho global em todas as etapas, com R^2 na validação cruzada de 0,5455 nos anos iniciais, 0,5205 nos anos finais e 0,4488 no ensino médio, e MAE sempre inferior a 0,51 pontos do IDEB em todos os casos. Verificou-se, contudo, maior dificuldade preditiva no ensino médio, possivelmente associada à evasão escolar, instabilidade institucional e menor volume amostral. Os modelos de DL apresentaram desempenho inferior, sobretudo em função da sensibilidade dessas arquiteturas a bases de dados reduzidas, o que justificou a não adoção como modelo final.

Como etapa complementar, foi desenvolvido um sistema interativo de simulação de cenários contrafactuais, implementado em Python. O sistema permite estimar variações no IDEB a partir de modificações em variáveis-chave, aplicadas aos dados de 2023. Embora a simulação tenha utilizado os mesmos dados do treinamento, seu propósito foi demonstrar o funcionamento metodológico e explorar possibilidades de aplicação prática. A ferramenta está estruturada para futura replicação com dados ainda não observados, podendo ser utilizada como subsídio à análise antecipada do impacto de políticas públicas educacionais.

A análise de importância das variáveis evidenciou que os determinantes do IDEB variam por etapa da educação básica. Nos anos iniciais e finais, destacaram-se PIB per capita, taxa de distorção idade-série, % de docentes sem formação superior, produção agrícola e repasses do programa Criança Feliz. No ensino médio, os fatores mais relevantes englobaram aspectos de infraestrutura (lousa digital, gestão escolar), pedagógicos (horas-aula, tamanho das turmas, qualificação docente) e socioeconômicos (efetivo de rebanhos, frequência dos alunos, transferências sociais).

As simulações contrafactuais, conforme os testes realizados no Capítulo 6, indicaram que os municípios com baixo IDEB tendem a apresentar maior sensibilidade a intervenções pedagógicas e de gestão escolar, especialmente em ações voltadas à qualificação docente e à

redução da defasagem idade-série. Em contrapartida, os municípios com alto desempenho apresentaram resultados estáveis, isto é, mesmo após alterações nas variáveis simuladas, o IDEB manteve-se praticamente inalterado. Esse comportamento sugere que essas localidades já alcançaram um patamar de desempenho próximo ao limite esperado e que novos avanços dependem menos de fatores estruturais e mais de elementos de natureza qualitativa, como o engajamento da comunidade, a cultura escolar, as práticas pedagógicas adotadas pelos docentes ou, ainda, de outros indicadores não contemplados neste estudo.

Os resultados alcançados superam os de trabalhos correlatos, como o de Maia, Bueno e Sato (2021), cujo modelo de melhor desempenho (GBM) obteve R^2 entre 0,305 e 0,459 no teste e MAE superior a 0,59 em todas as etapas. Apesar de o estudo citado abranger maior número de estados e anos, a presente pesquisa obteve métricas superiores ao concentrar-se no Piauí e adotar metodologia mais estruturada, incluindo validação cruzada e simulações contrafactuais, o que confere maior consistência aos achados.

Entre as limitações identificadas, destaca-se a ausência de variáveis qualitativas e socioemocionais, cuja inclusão futura poderia ampliar a capacidade explicativa dos modelos. Ainda assim, os resultados obtidos confirmam a viabilidade técnica e metodológica da proposta, bem como sua relevância prática para o aprimoramento da gestão pública educacional.

Como perspectivas futuras, sugere-se a ampliação do conjunto de variáveis explicativas, incluindo novos indicadores educacionais, institucionais e socioeconômicos, além da replicação do modelo com dados de outros estados brasileiros, a fim de avaliar sua capacidade de generalização territorial. Recomenda-se, ainda, o aprimoramento do sistema de simulação contrafactual, com a realização de testes adicionais voltados à identificação de pontos de melhoria. Ademais, propõe-se o desenvolvimento de uma versão online da ferramenta de simulação, com interface intuitiva e ambiente seguro, que possibilite sua utilização direta por gestores e analistas educacionais. Essa versão deverá permitir acesso facilitado aos cenários preditivos e favorecer o uso da modelagem como instrumento de apoio à tomada de decisão nas políticas públicas de educação.

Conclui-se, portanto, que a abordagem proposta nesta dissertação constitui uma contribuição significativa tanto para o campo da Ciência da Computação quanto para o aprimoramento da gestão educacional baseada em evidências. Do ponto de vista técnico, destaca-se a aplicação de um pipeline estruturado de aprendizado supervisionado, incorporando validação cruzada e comparação sistemática entre modelos clássicos e ANN's, bem como o desenvolvimento de um sistema interativo de simulação de cenários contrafactuais. A validação do modelo XGBoost como solução final evidencia o potencial das técnicas de ML para apoiar o diagnóstico e o planejamento educacional em escala municipal, oferecendo subsídios quantitativos para a formulação de políticas públicas mais eficazes, contextualizadas e orientadas por dados.

Referências

- AKINWALERE, S. N.; IVANOV, V. Artificial intelligence in higher education: challenges and opportunities. *Border Crossing*, London, v. 12, n. 1, p. 1–15, 2022. Disponível em: <<https://doi.org/10.33182/bc.v12i1.2015>>.
- ALAM, S. et al. An investigation of the imputation techniques for missing values in ordinal data enhancing clustering and classification analysis validity. *Decision Analytics Journal*, v. 9, p. 100295, 2023.
- ALVES, P. J. H.; FRIO, G. S. Uma análise dos resultados educacionais dos municípios brasileiros. *Política e Planejamento Econômico*, Brasília: IPEA, v. 48, n. 1, p. 1–30, 2020. Disponível em: <<https://www.ipea.gov.br/ppp/index.php/PPP/article/view/1465/647>>.
- AULCK, L.; NAMBI, D.; WEST, J. Using machine learning and genetic algorithms to optimize scholarship allocation for student yield. In: *Proceedings of the 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining – KDD '19*. Anchorage, Alaska: ACM, 2019.
- BARCLAY, G. W. *Techniques of Population Analysis*. New York: John Wiley & Sons, 1958.
- BENEVENTO, M.; MEIRELLES, F. d. S. Prever e melhorar o desempenho dos alunos com o uso combinado de aprendizagem de máquina e gpt. *Regae: Revista de Gestão e Avaliação Educacional*, Santa Maria, v. 12, n. 21, p. 1–22, 2023.
- BIGONHA, R. d. S. *Fundamentos do Aprendizado de Máquina*. Belo Horizonte, MG: s.n., 2025. ISBN 978-65-01-28912-0.
- BORJI, A. Pros and cons of gan evaluation measures. *Computer Vision and Image Understanding*, v. 179, p. 41–65, 2019.
- Brasil. *Constituição da República Federativa do Brasil de 1988*. Brasília, DF: [s.n.], 1988. Disponível em: <https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm>.
- Brasil. *Lei nº 9.394, de 20 de dezembro de 1996. Estabelece as diretrizes e bases da educação nacional*. Brasília, DF: [s.n.], 1996. Acesso em: 24 mar. 2025. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/l9394.htm>.
- Brasil. *Lei nº 13.005, de 25 de junho de 2014: aprova o Plano Nacional de Educação – PNE e dá outras providências*. Brasília, DF: [s.n.], 2014. Diário Oficial da União: seção 1. Publicada em: 26 jun. 2014. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l13005.htm>.
- Brasil. *Emenda Constitucional n.º 108, de 26 de agosto de 2020*. Brasília, DF, 2020. Disponível em: <https://www.planalto.gov.br/ccivil_03/constituicao/emendas/emc/emc108.htm>.
- CARAVAGGIO, N.; RESCE, G.; VAQUERO-PIÑEIRO, C. Predicting policy funding allocation with machine learning. *Socio-Economic Planning Sciences*, v. 98, p. 102175, 2025.

- CARREIRA, D.; PINTO, J. M. R. *Custo Aluno Qualidade Inicial: rumo à educação pública de qualidade no Brasil*. São Paulo: Global; Campanha Nacional pelo Direito à Educação, 2007. Disponível em: <https://media.campanha.org.br/acervo/documentos/CAQiRoxo_final_23out2007.pdf>.
- CHEN, S.; DING, Y. A machine learning approach to predicting academic performance in pennsylvania's schools. *Social Sciences*, v. 12, n. 3, p. 118, 2023.
- CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [s.n.], 2016. p. 785–794. Disponível em: <<https://arxiv.org/abs/1603.02754>>.
- CHERMACK, T. J. *Scenario Planning in Organizations: How to Create, Use, and Assess Scenarios*. San Francisco: Berrett-Koehler Publishers, 2011. ISBN 978-1-60509-413-7.
- CRESWELL, J. W. *Research design: qualitative, quantitative, and mixed methods approaches*. 4. ed. Thousand Oaks: Sage Publications, 2014.
- CURY, C. R. J. Direito à educação: direito à igualdade, direito à diferença. *Cadernos de Pesquisa*, São Paulo, v. 40, n. 141, p. 351–377, 2010. Disponível em: <<https://www.scielo.br/j/cp/a/x6g8nsWJ4MSk6K58885J3jd/?format=pdf&lang=pt>>.
- DOURADO, L. F.; OLIVEIRA, J. F. d. A qualidade da educação: perspectivas e desafios. *Cadernos CEDES*, Campinas, v. 78, p. 201–215, 2009. Disponível em: <<https://www.scielo.br/j/ccedes/a/Ks9m5K5Z4Pc5Qy5HRVgssjg/>>.
- DUTT, A.; ISMAIL, M. A.; HERAWAN, T. A systematic review on educational data mining. *IEEE Access*, v. 5, p. 15991–16005, 2017.
- ENCINAS, R.; DUENHAS, R. A. O fundeb e a desigualdade educacional nos municípios do estado do paran . *Educa  o & Sociedade*, v. 41, 2020. Disponível em: <<https://www.scielo.br/j/es/a/VCTPCGyrND7HZmDqV99YZhS/>>.
- FEITOSA, R. D. F.; SANTOS, D. T. d.; SOUSA, R. T. *Introdu  o a Machine Learning e redes neurais*. Goi  nia: Cegraf UFG, 2024. E-book.
- FNDE. *Cartilha do Novo Fundeb: Lei n  14.113/2020 e Emenda Constitucional n  108/2020*. Bras lia, 2023.
- FRANCO, C.; ALVES, F.; BONAMINO, A. Qualidade do ensino fundamental: pol ticas, suas possibilidades, seus limites. *Educa  o & Sociedade*, Campinas, v. 28, n. 100, p. 989–1014, 2007. Disponível em: <<https://www.scielo.br/j/es/a/ZqbnNkCG3YtpSs9VGhYzKxn/>>.
- FREITAS, L. C. Elimina  o adiada: o ocaso das classes populares no interior da escola e a oculta  o da (m ) qualidade de ensino. *Educa  o e Sociedade*, Campinas, v. 28, n. 100, p. 965–987, 2007. Disponível em: <<https://www.scielo.br/j/es/a/h8RgCZ6JvYpJNLr8MXxvNMf/?format=pdf&lang=pt>>.
- FREITAS, R.; SALGADO, L. C. d. C. Educators in the loop: using scenario simulation as a tool to understand and investigate predictive models of student dropout risk in distance learning. In: *International Conference on Human-Computer Interaction*. [s.n.], 2020. Disponível em: <<https://www.researchgate.net/publication/342834887>>.

- Fundação Getulio Vargas. *Boletim macroeconômico regional – Nordeste, edição n.º 2*. São Paulo, 2024. Disponível em: <https://portalibre.fgv.br/sites/default/files/2024-10/boletim-macro-regional-ne-ed-02-vf_0.pdf>.
- GESQUI, L. C. O ideb como parâmetro de qualidade da educação básica no Brasil: algumas preocupações. *Educar em Revista*, Curitiba, v. 23, n. 3, p. 88–99, 2023.
- GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. Cambridge, MA: The MIT Press, 2016. ISBN 978-0262035613.
- GOODFELLOW, I. et al. Generative adversarial nets. In: *Advances in Neural Information Processing Systems (NIPS 2014)*. [S.l.: s.n.], 2014. v. 27, p. 2672–2680.
- GREGORI, P.; SILVA, M. Mira da. Artificial intelligence in higher education: challenges and opportunities from a student-centered perspective. *Education and Information Technologies*, v. 26, p. 2261–2284, 2021.
- GUANIN-FAJARDO, J. H.; GUAÑA-MOYA, J.; CASILLAS, J. Predicting academic success of college students using machine learning techniques. *Data*, v. 9, n. 4, p. 60, 2024.
- HANUSHEK, E. A.; WOESSMANN, L. The economics of international differences in educational achievement. In: HANUSHEK, E. A.; MACHIN, S.; WOESSMANN, L. (Ed.). *Handbook of the Economics of Education*. Amsterdam: Elsevier, 2020. p. 89–200.
- HANUSHEK, E. A.; WOESSMANN, L. Education, knowledge capital, and economic growth. In: SLEEPER, S. (Ed.). *The economics of education*. 2. ed. Amsterdam: Elsevier, 2020. p. 171–182.
- HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer, 2009.
- HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. *Neural Computation*, v. 9, n. 8, p. 1735–1780, 1997.
- HOLMES, W. et al. *Artificial Intelligence in Education: Promise and Implications for Teaching and Learning*. Paris: OECD Publishing, 2019.
- HOLMES, W. et al. *Ethics of Artificial Intelligence in Education: Practices, Challenges, and Debates*. New York: Routledge, 2023.
- IBGE. *Produto Interno Bruto dos Municípios: 2021*. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística, 2023. Disponível em: <<https://www.ibge.gov.br>>.
- INEP. *Indicadores Educacionais: metodologia e conceitos*. Brasília, DF: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2022. Acesso em: 12 abr. 2025. Disponível em: <<https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/indicadores-educacionais>>.
- INEP. *Apresentação do IDEB 2023*. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2023. Acesso em: 6 abr. 2025. Disponível em: <https://download.inep.gov.br/ideb/apresentacao_ideb_2023.pdf>.
- INEP. *Nota Técnica nº 3/2023 – Cálculo do IDEB*. Brasília, DF: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2023.

- INEP. *Relatório do 5º Ciclo de Monitoramento das Metas do Plano Nacional de Educação – 2024*. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2024.
- IPEA. *Diálogos para uma Política Nacional de Desenvolvimento Regional: Desafios do financiamento urbano e da governança interfederativa. Volume 2*. Brasília: Instituto de Pesquisa Econômica Aplicada, 2021. Disponível em: <https://repositorio.ipea.gov.br/bits/tream/11058/12711/1/Dialogos_para_uma_Politica_Nacional_v.2Book.pdf>.
- KARIM-ABDALLAH, B. et al. Application of machine learning algorithms in predicting academic performance of students in higher education institutes (heis): a systematic review and bibliographic analysis. *African Journal of Applied Research*, Sunyani, v. 11, n. 1, p. 536–559, jan 2025. Disponível em: <<https://doi.org/10.26437/ajar.v11i1>>.
- KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, v. 60, n. 6, p. 84–90, 2017.
- KURSER, M. B.; RUDNICKI, W. R. Feature selection with the boruta package. *Journal of Statistical Software*, v. 36, n. 11, p. 1–13, 2010. Disponível em: <<https://www.jstatsoft.org/article/view/v036i11>>.
- LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *Nature*, v. 521, p. 436–444, 2015. Disponível em: <https://www.researchgate.net/publication/277411157_Deep_Learning>.
- LUDERMIR, T. B. Inteligência artificial e aprendizado de máquina: estado atual e tendências. *Estudos Avançados*, São Paulo, v. 35, n. 101, p. 85–94, 2021.
- LUNDBERG, S. M.; LEE, S. I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, v. 30, p. 4765–4774, 2017.
- MAIA, J. d. S. Z.; BUENO, A. P. A.; SATO, J. R. Assessing the educational performance of different brazilian school cycles using data science methods. *PLoS ONE*, v. 16, n. 3, p. e0248525, 2021.
- MILLER, R. E.; BLAIR, P. D. *Input–output analysis: foundations and extensions*. 2. ed. Cambridge: Cambridge University Press, 2009.
- MOLNAR, C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2. ed. [s.n.], 2020. Disponível em: <<https://christophm.github.io/interpretable-ml-book/>>.
- MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. *Introduction to Linear Regression Analysis*. 5. ed. Hoboken: Wiley, 2012.
- MOROSINI, M. C. Estado do conhecimento: natureza, alcance e desafios. *Educação em Revista*, Belo Horizonte, v. 31, n. 1, p. 1–21, 2015.
- MURPHY, K. P. *Machine Learning: A Probabilistic Perspective*. Cambridge: MIT Press, 2012.
- OCDE. *Education at a Glance 2023: OECD Indicators*. Paris: OECD Publishing, 2023.

- OECD. *AI and the Future of Skills: Capabilities and Assessments*. Paris: Organisation for Economic Co-operation and Development, 2021. Disponível em: <<https://www.oecd.org/education/ai-future-of-skills.htm>>.
- PEREIRA, L. G. d. C.; SERUFFO, M. C. R. Previsão de desempenho escolar aplicando perceptron de multicamadas. *RevistaFT*, Rio de Janeiro, v. 28, n. 136, jul. 2024. Acesso em: 30 mar. 2025. Disponível em: <<https://revistaft.com.br/previsao-de-desempenho-escolar-aplicando-perceptron-de-multicamadas/>>.
- PIAUÍ. *Secretaria da Educação do Piauí orienta gestores sobre pactuação de metas 2023*. Teresina: [s.n.], 2024. Secretaria de Estado da Educação – SEDUC. Acesso em: 13 set. 2025. Disponível em: <<https://www.seduc.pi.gov.br/noticias/noticia/12896/secretaria-da-educacao-do-piaui-orienta-gestores-sobre-pactuacao-de-metas-2023>>.
- Piauí. *Lei nº 7.429, de 2 de setembro de 2020: Dispõe sobre o ICMS Educacional no Estado do Piauí e dá outras providências*. Teresina: [s.n.], 2020. Diário Oficial do Estado do Piauí.
- PINTO, G. d. S.; JÚNIOR, O. d. G. F.; COSTA, E. d. B. Identificação dos fatores de melhorias no ideb pelo uso de mineração de dados: Um estudo de caso em escolas municipais de teotônio vilela-alagoas. *Revista Novas Tecnologias na Educação – RENOTE*, Porto Alegre, v. 17, n. 3, p. 183–192, dez. 2019. Acesso em: 30 mar. 2025. Disponível em: <<https://www.researchgate.net/publication/339035763>>.
- PINTO, J. M. d. R. O financiamento da educação na constituição federal de 1988: 30 anos de mobilização social. *Educação & Sociedade*, Campinas, v. 39, n. 145, p. 1105–1124, dez. 2018.
- POWERS, D. M. Evaluation: From precision, recall and f-measure to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, v. 2, n. 1, p. 37–63, 2011.
- ROCHA, F. A. F. d.; TEIXEIRA, J. C. M.; MELO, F. L. N. B. d. Análise dos fatores que influenciam o desempenho escolar dos alunos do ensino fundamental no estado do rio grande do norte. *Revista Interface*, [S.l.], v. 1, n. 1, p. 89–105, 2015. ISSN 2237-7506. Jan./jul.
- ROMERO, C.; VENTURA, S. Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, v. 10, n. 3, p. e1355, 2020. Disponível em: <<https://arxiv.org/abs/2402.07956>>.
- ROMERO, C.; VENTURA, S. Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, v. 10, n. 3, p. e1355, 2020. Acesso em: 06 abr. 2025. Disponível em: <<https://arxiv.org/abs/2402.07956>>.
- SAVIANI, D. *História das ideias pedagógicas no Brasil*. 4. ed. Campinas: Autores Associados, 2013. (Coleção Memória da Educação). ISBN 978-85-7496-322-8.
- SCHWARTZ, P. *The Art of the Long View: The Path to Strategic Insight for Yourself and Your Company*. 1. ed. New York: Doubleday, 1991. ISBN 0-385-26731-2.

- SHIROMA, E. O.; MORAES, M. C.; EVANGELISTA, O. *Política educacional*. 2. ed. Rio de Janeiro: DP&A, 2004.
- SILVA, A. L. d.; GRACIANO, A. B. As políticas públicas de educação: o caso do ensino médio na região nordeste do Brasil. *Educação e Sociedade*, Campinas, v. 43, p. e023002856, 2022. Acesso em: 5 maio 2025. Disponível em: <<https://doi.org/10.1590/S0104-40362022003002856>>.
- SILVA, J. C. L. Definição de modelos de aprendizado de máquina para predição de evasão de alunos do curso técnico. *Refas – Revista Fatec Zona Sul*, São Paulo, v. 9, n. 3, fev. 2023. ISSN 2359-182X.
- SILVEIRA, A. A. D.; SCHNEIDER, G.; ALVES, T. *Simulador de Custo-Aluno Qualidade: Padrão de Qualidade de Referência, versão 02.2023*. Curitiba; Goiânia, 2023. Acesso em: 25 fev. 2025. Disponível em: <https://simcaq.c3sl.ufpr.br/assets/SIMCAQ_PQR_2023.pdf>.
- SIOPE. *Dados municipais consolidados*. Brasília, 2023. Acesso em: 10 abr. 2025. Disponível em: <<https://siope.fnde.gov.br>>.
- SOARES, D. J. M.; SANTOS, W. Indicadores de avaliação de contexto e resultados educacionais no IDEB: uma análise das escolas estaduais de ensino médio no Espírito Santo. *Revista Brasileira de Estudos Pedagógicos*, Brasília, v. 101, n. 259, p. 297–323, 2020. Acesso em: 12 abr. 2025. Disponível em: <<https://www.scielo.br/j/rbeped/a/J8BpZyQBxbtwvYTL7jZmzwc/?format=pdf&lang=pt>>.
- SOARES, J. F.; ARAÚJO, R. J. d. Nível socioeconômico, qualidade e equidade das escolas de Belo Horizonte. *Ensaio: Avaliação e Políticas Públicas em Educação*, Rio de Janeiro, v. 14, n. 50, p. 107–126, jan./mar. 2006. Acesso em: 12 abr. 2025. Disponível em: <<https://www.redalyc.org/pdf/3995/399537942008.pdf>>.
- SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, v. 45, n. 4, p. 427–437, 2009.
- SRIVASTAVA, N. et al. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, v. 15, p. 1929–1958, 2014. Disponível em: <<https://www.jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf>>.
- TAGHIPOUR, K.; NG, H. T. A neural approach to automated essay scoring. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. [s.n.], 2016. p. 1882–1891. Disponível em: <<https://doi.org/10.18653/v1/D16-1191>>.
- TEODORO, L. d. A.; KAPPEL, M. A. A. Aplicação de técnicas de aprendizado de máquina para predição de risco de evasão escolar em instituições públicas de ensino superior no Brasil. *Revista Brasileira de Informática na Educação – RBIE*, [S.l.], v. 28, n. 2, p. 27–48, 2020. Acesso em: 30 mar. 2025. Disponível em: <<http://br-ie.org/pub/index.php/rbie>>.
- TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, v. 58, n. 1, p. 267–288, 1996.
- UNESCO. *Education for All 2000–2015: Achievements and Challenges*. 2. ed. Paris: United Nations Educational, Scientific and Cultural Organization, 2015. Acesso em: 21 mar. 2025. ISBN 978-92-3-100085-0. Disponível em: <<https://unesdoc.unesco.org/ark:/48223/pf0000232205>>.

UNESCO. *Recommendation on the Ethics of Artificial Intelligence*. Paris: Organização das Nações Unidas para a Educação, a Ciência e a Cultura, 2021. Acesso em: 6 abr. 2025. Disponível em: <<https://unesdoc.unesco.org/ark:/48223/pf0000381137>>.

UNESCO. *Reimaginar juntos nossos futuros: um novo contrato social para a educação*. Paris: Organização das Nações Unidas para a Educação, a Ciência e a Cultura, 2022. Acesso em: 6 abr. 2025. Disponível em: <<https://unesdoc.unesco.org/ark:/48223/pf0000381560>>.

VANLEHN, K. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, v. 46, n. 4, p. 197–221, 2011. Disponível em: <<https://doi.org/10.1080/00461520.2011.611369>>.

VASWANI, A. et al. Attention is all you need. In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. p. 5998–6008. Disponível em: <<https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>>.

VIAGORA. *Seduc-PI realiza investimento de R\$ 1,5 bilhão em infraestrutura das escolas*. Teresina: [s.n.], 2024. Publicado em 22 dez. 2024. Disponível em: <<https://www.viagora.com.br/piaui/noticia/2024/12/22/seduc-pi-realiza-investimento-de-r-15-bilhao-em-infraestrutura-das-escolas-118621.html>>.

VIANA, A. R. M.; SANTOS, c. a. o. d. a. s. d. et al. Educação no brasil: análise do financiamento, dilemas e perspectivas futuras. *Revista Aracê*, v. 7, n. 6, p. 1–16, 2025. Acesso em: 12 out. 2025. Disponível em: <<https://doi.org/10.56238/arev7n6-150>>.

WANDERA, H.; MARIVATE, V.; SENGEH, D. Investigating similarities and differences between south african and sierra leonean school outcomes using machine learning. *arXiv preprint*, 2020. Acesso em: 6 maio 2025. Disponível em: <<https://arxiv.org/abs/2004.11369>>.

WANG, S.; LUO, B. Academic achievement prediction in higher education through interpretable modeling. *PLOS ONE*, v. 19, n. 9, p. e0309838, 2024. Acesso em: 15 fev. 2025. Disponível em: <<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0309838>>.

WILLMOTT, C. J.; MATSUURA, K. Advantages of the mean absolute error (mae) over the root mean square error (rmse). *Climate Research*, v. 30, p. 79–82, 2005. Acesso em: 21 mar. 2025. Disponível em: <<https://www.int-res.com/articles/cr2005/30/c030p079.pdf>>.

ZAWACKI-RICHTER, O. et al. Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, v. 16, n. 39, 2019. Disponível em: <<https://educationaltechnologyjournal.springeropen.com/articles/10.1186/s41239-019-0171-0>>.

APÊNDICE A – Scripts de Modelagem Preditiva e Simulação de Cenários

Este apêndice apresenta os códigos desenvolvidos pela autora para a implementação e validação dos modelos de *Machine Learning* e *Deep Learning* aplicados à predição do Índice de Desenvolvimento da Educação Básica (IDEB) nos municípios piauienses, bem como o simulador interativo de cenários educacionais. Os scripts foram elaborados em ambiente *Google Colab* e estão disponíveis para acesso público, de modo a garantir a transparência e a reprodutibilidade dos resultados.

Os códigos encontram-se organizados por etapa da educação básica e contemplam todo o processo de modelagem, desde o pré-processamento das bases de dados até a avaliação preditiva e a análise de sensibilidade. As implementações utilizam a linguagem Python (versão 3.10) e bibliotecas amplamente reconhecidas, tais como **pandas**, **numpy**, **matplotlib**, **scikit-learn**, **xgboost** e **keras**, conforme detalhado no Capítulo 4.

Links de acesso aos scripts

- **Modelagem completa – Anos Iniciais:**
<<https://colab.research.google.com/drive/1Aj9dep8j1UuFJw6sx5VEfyPn9r9Zyvt3?usp=sharing>>
- **Modelagem completa – Anos Finais:**
<https://colab.research.google.com/drive/1O69OHw9hOv_winay4-KC206wk-0RAVw-?usp=sharing>
- **Modelagem completa – Ensino Médio:**
<<https://colab.research.google.com/drive/1Aa8sKx3uOPUWaOeaE2PKXHGUBTNZkfJy?usp=sharing>>
- **Simulador de Cenários Preditivos Interativo:**
<https://colab.research.google.com/drive/1UpbaAZ_DqYcYXkTOvFyTB22nS7Vg8Pil?usp=sharing>
- **Dashboard Interativo – Simulações do Capítulo 6:**
<<https://app.powerbi.com/view?r=eyJrIjoia0NkNzYxZmQtMWVhZC00MzU4LTk0NDgtY2NmM2M1ZjhiZmIzIiwidCI6ImIwNTgyNWUyLTAYZGItNDI0YS04Y2UyLTA3OWJmMDE2ZDIwMCJ9&pageName=ReportSectionef86303cf72904fab66d>>

Cada notebook contém células de código, gráficos e comentários explicativos que descrevem o funcionamento das etapas de tratamento, seleção de variáveis, modelagem, validação cruzada e simulação. O simulador de cenários permite ao usuário ajustar percentuais de variação em variáveis explicativas selecionadas e visualizar o impacto previsto sobre o IDEB médio municipal, possibilitando análises hipotéticas de políticas públicas educacionais.

Elaborado pela autora (2025). Última atualização: outubro de 2025.

Anexos

ANEXO A – Fontes de Dados Utilizadas na Dissertação

Este anexo apresenta as bases de dados, documentos e sistemas oficiais que serviram de fonte para a construção das variáveis e indicadores empregados na modelagem preditiva do Índice de Desenvolvimento da Educação Básica (IDEB). As informações foram obtidas de fontes públicas, com ênfase em bases do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), Instituto Brasileiro de Geografia e Estatística (IBGE), Fundo Nacional de Desenvolvimento da Educação (FNDE), Secretaria do Tesouro Nacional (STN) e outros sistemas de acesso aberto.

Os dados financeiros foram gentilmente fornecidos pela Prof^a. Dra. Rosana Evangelista da Cruz, responsável pelos estudos orçamentários que subsidiaram esta pesquisa.

1. Links de acesso às bases consolidadas

- **Base consolidada – Anos Iniciais:**

<https://docs.google.com/spreadsheets/d/16caciVQwEEEHgZqT-zbheLUb2M5mrbzw/edit?usp=drive_link&ouid=115357405006404344826&rtpof=true&sd=true>

- **Base consolidada – Anos Finais:**

<https://docs.google.com/spreadsheets/d/1xlJxQ-bfE8pWbSYYh-vC7ISP3nOF0Ziv/edit?usp=drive_link&ouid=115357405006404344826&rtpof=true&sd=true>

- **Base consolidada – Ensino Médio:**

<https://docs.google.com/spreadsheets/d/1oC2L5tZtXwqvtxwyaMJsphTX9da5Y_f4/edit?usp=drive_link&ouid=115357405006404344826&rtpof=true&sd=true>

2. Principais fontes oficiais consultadas

1. Censo Escolar da Educação Básica – Microdados

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP).
<<https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/microdados/censo-escolar>>

2. Indicadores Educacionais do INEP

<<https://www.gov.br/inep/pt-br/acao-a-informacao/dados-abertos/indicadores-educacionais>>

- Adequação da Formação Docente
- Média de Alunos por Turma
- Média de Horas-aula Diária
- Percentual de Docentes com Curso Superior
- Taxas de Rendimento e de Distorção Idade-Série
- Complexidade de Gestão Escolar
- Esforço Docente
- Regularidade do Corpo Docente

3. IDEB – Resultados por Município

<<https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/ideb/resultados>>

4. Sistema de Informações sobre Orçamentos Públicos em Educação (SIOPE)

Fundo Nacional de Desenvolvimento da Educação (FNDE/MEC). <<https://siope.fn.de.gov.br/>>

5. Finanças do Brasil – FINBRA

Secretaria do Tesouro Nacional (STN). <<https://www.tesourotransparente.gov.br/>>

6. FUNDEB – Complementações e Indicadores de Financiamento

<<https://www.gov.br/fnde/pt-br/acao-a-informacao/dados-abertos/fundeb>>

7. IBGE – Sistema IBGE de Recuperação Automática (SIDRA)

<<https://sidra.ibge.gov.br/>>

- Tabela 5457 – Lavouras Temporárias e Permanentes
- Tabela 289 – Extração Vegetal
- Tabela 74 – Produção de Origem Animal
- Tabela 3939 – Efetivo de Rebanhos

8. Secretaria de Avaliação, Gestão da Informação e Cadastro Único (SAGI-CAD) – VIS DATA 3 Beta

3. Observações gerais

- Todos os dados analisados referem-se exclusivamente às escolas públicas.
- O período considerado abrange os anos de 2013 a 2023 para o ensino fundamental e de 2017 a 2023 para o ensino médio, utilizando-se apenas os anos ímpares, uma vez que o IDEB é divulgado bianualmente e, no caso do ensino médio, sua disponibilização inicia-se apenas a partir de 2017.
- As bases foram tratadas, agregadas e harmonizadas pela autora, preservando a integridade dos identificadores municipais (*ano, código e nome do município*).

Elaborado pela autora (2025). Última atualização: outubro de 2025.

ANEXO B – Lista Detalhada de Variáveis e Descrições

Este anexo apresenta a relação completa das variáveis utilizadas na dissertação, organizadas por categoria temática. Incluem-se os fatores internos (educacionais), externos (socioeconômicos e financeiros), com base nas fontes descritas no Anexo A. Em todas as bases estão presentes as variáveis de identificação Ano, Código do Município e Nome do Município, assegurando a padronização e a integridade das chaves de agregação. No que se refere aos dados educacionais, ressalta-se que todas as informações correspondem exclusivamente às escolas públicas, conforme os critérios de filtragem adotados na consolidação da base analítica.

O detalhamento a seguir tem como finalidade documentar a amplitude da base consolidada, garantindo replicabilidade e transparência metodológica.

1. Fatores internos – Bases educacionais (INEP)

1.1 IDEB e Indicadores Associados

- Tipo de ensino: Anos Iniciais, Anos Finais e Ensino Médio;
- Taxa de aprovação;
- Indicador de rendimento (P);
- Nota SAEB em Matemática;
- Nota SAEB em Língua Portuguesa;
- Nota SAEB padronizada (N);
- Nota final do IDEB.

1.2 Média de Alunos por Turma

- Tipo de ensino;
- Média de alunos por turma.

1.3 Taxas de Distorção Idade-Série

- Tipo de ensino;
- Taxa de distorção idade-série.

1.4 Taxas de Rendimento Escolar

- Tipo de ensino;
- Taxas de aprovação, reprovação e abandono.

1.5 Percentual de Docentes com Curso Superior

- Tipo de ensino;
- Percentual de docentes com curso superior completo.

1.6 Média de Horas-Aula Diária

- Tipo de ensino;
- Média diária de horas-aula.

1.7 Adequação da Formação Docente

Classificação dos docentes conforme sua formação em relação à disciplina que leciona:

1. **Grupo 1:** licenciatura (ou bacharelado com complementação pedagógica) na mesma área da disciplina que leciona;
2. **Grupo 2:** bacharelado (sem complementação pedagógica) na mesma área da disciplina que leciona;
3. **Grupo 3:** licenciatura (ou bacharelado com complementação pedagógica) em área distinta;
4. **Grupo 4:** formação superior não enquadrada nas categorias anteriores;
5. **Grupo 5:** docentes sem formação superior.

1.8 Complexidade de Gestão Escolar

Classificação das escolas em seis níveis de complexidade, conforme porte, número de turnos e etapas de ensino ofertadas:

- **Nível 1:** até 50 matrículas, único turno, oferta restrita à Educação Infantil ou Anos Iniciais;
- **Nível 2:** entre 50 e 300 matrículas, até dois turnos, etapas até os Anos Iniciais;
- **Nível 3:** entre 50 e 500 matrículas, até dois turnos, etapas até os Anos Finais;
- **Nível 4:** entre 150 e 1.000 matrículas, até três turnos, incluindo Ensino Médio, Educação Profissional ou EJA;
- **Nível 5:** porte médio, até três turnos, predominância da EJA;
- **Nível 6:** mais de 500 matrículas, três turnos, quatro ou mais etapas de ensino, com predominância da EJA.

1.9 Esforço Docente

Classificação dos docentes conforme carga de trabalho, número de alunos e turmas atendidas:

- **Nível 1:** até 25 alunos, um turno, uma escola e uma etapa;
- **Nível 2:** 25 a 150 alunos, um turno, uma escola e etapa;
- **Nível 3:** 25 a 300 alunos, até dois turnos, uma escola e etapa;
- **Nível 4:** 50 a 400 alunos, dois turnos, até duas escolas e etapas;
- **Nível 5:** mais de 300 alunos, três turnos, até três escolas e etapas;
- **Nível 6:** mais de 400 alunos, três turnos, até três escolas e etapas.

1.10 Regularidade do Corpo Docente (IRD)

Classificação das escolas segundo a média do Indicador de Regularidade Docente (IRD):

- Baixa regularidade ($IRD \leq 2$);
- Média-baixa ($2 < IRD \leq 3$);
- Média-alta ($3 < IRD \leq 4$);
- Alta regularidade ($IRD > 4$).

1.11 Microdados do Censo Escolar – INEP

Agrupamento das variáveis de oferta educacional, infraestrutura, recursos humanos e participação comunitária:

- Quantidade de escolas e matrículas;
- Infraestrutura física: salas, climatização, acesso a água, energia, esgoto, coleta de lixo, acessibilidade e laboratórios;
- Equipamentos tecnológicos: lousas digitais, computadores, tablets e multimídia;
- Recursos humanos escolares: pedagogos, gestores, secretários, monitores e profissionais de saúde;
- Participação da comunidade: existência de conselho escolar, associação de pais e mestres e grêmio estudantil.

2. Fatores Externos – Bases Socioeconômicas (IBGE, SAGICAD)

- População residente (Censo 2022 e estimativas 2013–2023);
- Produto Interno Bruto (PIB) e valor adicionado bruto setorial (Agropecuária, Indústria, Serviços, Administração Pública);
- Benefício de Prestação Continuada (BPC) — número de beneficiários (idosos e PCDs) e valores repassados;
- Renda Mensal Vitalícia (RMV) — número de beneficiários e valores repassados;
- Transferências via Fundos de Assistência Social — Proteção Social Básica, Criança Feliz, Auxílio Brasil, SUAS e Fundo a Fundo;
- Produção agropecuária — lavouras temporárias e permanentes, extração vegetal e produção animal;
- Efetivo de rebanhos — total de cabeças por espécie.

3. Fatores Financeiros – Bases Fiscais (SIOPE, FINBRA, FUNDEB)

- Impostos próprios municipais — IPTU, ITBI, ISS e IRRF;
- Transferências constitucionais — FPM, ICMS, IPVA, IPI-Exp, IOF-Ouro;

- Transferências via FNDE — Salário-Educação, PDDE, PNAE e PNATE;
- FUNDEB — contribuições, redistribuições, complementações da União (VAAF, VAAT, VAAE) e receitas financeiras;
- Manutenção e Desenvolvimento do Ensino (MDE) — valores aplicados, despesas com profissionais e destinação por etapa (Educação Infantil, Fundamental, Médio, Profissional).

Elaborado pela autora (2025). Última atualização: outubro de 2025.