Análise de Fatores de Risco para a Evasão Escolar na Educação Básica usando Modelos Preditivos de Machine Learning

Rodrigo Antonio da S. Lira[1], **Fernanda Maria Ribeiro Alencar** [2]

[1]Universidade de Pernambuco - PE (UPE)
Rua Benfica, 455, Madalena – 50720-000 – Recife – PE – Brazil

[2]Universidade Federal de Pernambuco (UFPE)
Av. Prof. Moraes Rego, 1235, Cidade Universitária – 50670-901 – Recife – PE – Brasil

rasl,fernandaalenc@ecomp.poli.br

Abstract. *School dropout remains one of the main educational challenges in Brazil, significantly affecting the country's social and economic development. This study applies Machine Learning techniques to predict school dropout using INEP data, employing the Linear Regression, Random Forest, and XGBoost algorithms. The analyses were conducted using data from the 2023 School Census and academic performance indicators, covering the segments of Early Years, Final Years, and High School. The XGBoost model achieved the best performance, with $R^2 = 0.503$ in High School, demonstrating the ability to explain 50.3% of the variability in dropout rates. Infrastructure variables such as cafeteria, library, and sports court showed high predictive relevance. The findings highlight regional disparities and provide insights for more targeted public policies. The study emphasizes that the results indicate statistical correlation, not causation, and discusses the limitations of using public data and the absence of socio-emotional variables.*
Keywords: School Dropout, Basic Education, Machine Learning, Predictive Models, Educational Data Mining.

Resumo. *A evasão escolar segue como um dos principais desafios educacionais do Brasil, impactando o desenvolvimento social e econômico do país. Este estudo emprega técnicas de Machine Learning para prever a evasão escolar com base em dados do INEP, utilizando os algoritmos Regressão Linear, Random Forest e XGBoost. As análises foram conduzidas com dados do Censo Escolar 2023 e indicadores de rendimento, abrangendo os segmentos dos Anos Iniciais, Anos Finais e Ensino Médio. O modelo XGBoost obteve o melhor desempenho, com $R^2 = 0{,}503$ no Ensino Médio, demonstrando capacidade de explicar 50,3% da variabilidade na evasão. Variáveis de infraestrutura, como refeitório, biblioteca e quadra de esportes, apresentaram alta relevância preditiva. Os achados destacam disparidades regionais e apontam subsídios para políticas públicas mais direcionadas. O estudo ressalta que os resultados indicam correlação estatística, não causalidade, e discute as limitações do uso de dados públicos e ausência de variáveis socioemocionais.*
Palavras-chave: Evasão Escolar, Educação Básica, Machine Learning, Modelos Preditivos, Mineração de Dados Educacionais.

1. Introdução

A evasão escolar é um desafio persistente no sistema educacional brasileiro, especialmente na educação básica, onde a interrupção precoce dos estudos compromete o desenvolvimento social e econômico do país [de Oliveira Almeida et al. 2021]. Nesse contexto, fatores socioeconômicos, ambientais e individuais desempenham papéis cruciais nesse fenômeno [de Jesus and de Gusmão 2024], sendo as disparidades regionais um agravante importante. Assim, a identificação precoce de estudantes em risco pode viabilizar intervenções mais eficazes e personalizadas [de Faria et al. 2022].

Nessa linha de pensamento, o Brasil dispõe de um sistema abrangente de monitoramento educacional, integrado pelo Censo Escolar, que inclui o Sistema de Avaliação da Educação Básica (SAEB) e o Índice de Desenvolvimento da Educação Básica (IDEB), que permitem uma visão detalhada do desempenho e dos desafios educacionais. Em 2023, registraram-se 47,3 milhões de matrículas na educação básica, marcando uma redução de 0,2% em relação à 2022. A maior parte dessas matrículas está concentrada na rede municipal 49,3%, seguida pela rede estadual 30%, privada 19,9% e federal 0,8% [BRASIL 2023].

Modelos preditivos baseados em Machine Learning, oferecem insights valiosos sobre eficiência e aplicabilidade em cenários educacionais [Gramacho 2019]. Este estudo busca avaliar a eficácia dessas abordagens, identificar lacunas na pesquisa e contribuir para o desenvolvimento de soluções tecnológicas mais robustas e adaptadas ao contexto educacional brasileiro, promovendo intervenções direcionadas e estratégias de redução da evasão escolar.

Diante desse cenário, este artigo propõe realizar uma análise aprofundada sobre a modelagem preditiva da evasão escolar na educação básica do Brasil, por meio de uma análise comparativa de algoritmos de aprendizado de máquina. Especificamente, objetiva-se: (1) identificar os principais fatores de risco associados à evasão escolar na educação básica brasileira; (2) comparar o desempenho de diferentes modelos preditivos na identificação de padrões de evasão; e (3) propor diretrizes para o desenvolvimento de políticas públicas educacionais baseadas em evidências.

2. Trabalhos Relacionados ao Tema

[de Brito et al. 2022] realizaram uma pesquisa que investigou a relação entre o Índice de Nível Socioeconômico (Inse) e a proficiência em matemática dos estudantes do 9º ano de escolas públicas de Pernambuco, utilizando os microdados do Sistema de Avaliação da Educação Básica (Saeb) das edições de 2011 a 2019. Por meio de estatística descritiva e inferencial com o software RStudio e de pacotes como tidyverse, ggpubr e rstatix, a pesquisa identificou que fatores socioeconômicos, como renda familiar e local de residência, influenciam o desempenho dos estudantes. Os resultados revelaram que grupos socioeconômicos mais altos tendem a ter índices de proficiência em matemática superiores.

A pesquisa de [Lima and Fagundes 2023] destacou algoritmos promissores, como as Máquinas de Vetores de Suporte (SVM), redes neurais artificiais e algoritmos baseados em árvores de decisão, como o J48. Cada um desses algoritmos foi avaliado quanto à sua capacidade de prever a evasão escolar, considerando as características específicas do ambiente educacional de Pernambuco. Além de identificar os algoritmos mais efi-

cazes, o estudo também discutiu a importância de considerar variáveis contextuais específicas, como condições socioeconômicas, recursos escolares, envolvimento dos pais e características demográficas da região.

O estudo realizado por [do Nascimento et al. 2018] investigou e explicou variáveis educacionais relacionadas à evasão e à reprovação escolar, utilizando técnicas lineares como regressão linear e robusta. Por meio de diferentes indicadores educacionais do INEP, novas bases de dados foram geradas para os experimentos. A metodologia CRISP-DM foi fundamental, incluindo análise correlacional para identificar variáveis mais associadas aos indicadores de evasão e reprovação. Os resultados mostraram que a regressão robusta teve melhor desempenho que a regressão linear no ensino fundamental de Pernambuco, com menor erro de predição.

Esses trabalhos, coletivamente, não apenas demonstram a viabilidade das técnicas de mineração de dados e aprendizado de máquina na predição da evasão escolar, mas também sublinham a necessidade de adaptações contextuais e a importância de uma abordagem holística que considere as variáveis socioeducacionais e institucionais específicas.

3. Metodologia

Este estudo tem como objetivo prever os estados e municípios com maior propensão à evasão escolar, a partir da análise integrada dos dados de rendimento escolar e dos indicadores de infraestrutura das escolas. A metodologia adotada segue as etapas do processo de Descoberta de Conhecimento em Bases de Dados (Knowledge Discovery in Databases – KDD), conforme proposto por [Fayyad et al. 1996], abrangendo as fases de seleção, pré-processamento, transformação, mineração, interpretação e avaliação dos dados.

Foram aplicados os algoritmos de Regressão Linear, Random Forest e XGBoost, avaliados por meio das métricas RMSE, MAE e R^2, permitindo a identificação dos fatores de infraestrutura mais relevantes para a predição da evasão escolar. A implementação foi realizada em linguagem Python, com o uso de bibliotecas especializadas, como NumPy, Pandas, Matplotlib e Scikit-learn, que forneceram suporte às etapas de manipulação de dados, visualização gráfica e modelagem preditiva.

Os resultados evidenciam que a combinação de indicadores educacionais e de infraestrutura com técnicas de aprendizado de máquina constitui uma abordagem eficaz para a identificação de áreas críticas, subsidiando o desenvolvimento de políticas públicas voltadas à mitigação da evasão escolar.

3.1. Seleção dos Dados

Este estudo utilizou duas bases públicas, fornecidas pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira(INEP), para compor um conjunto de dados voltado à análise da evasão escolar. Foram consideradas variáveis educacionais, como taxas de abandono, aprovação e reprovação por etapa de ensino, e variáveis de infraestrutura, como biblioteca, internet, laboratório de ciências, quadra, refeitório e acesso à energia elétrica. Esses fatores foram selecionados por sua relevância na permanência dos estudantes e no desempenho escolar.

A primeira base, `microdados_ed_basica_2023`, contém informações sobre a infraestrutura das escolas de ensino fundamental e médio. Essa base possui 408 variáveis

e 217.625 registros. Está disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-escolar.

A segunda base, intitulada tx_rend_brasil_regioes_ufs_2023, apresenta taxas de aprovação, reprovação e abandono escolar, segmentadas por nível de ensino, dependência administrativa (federal, estadual, municipal e privada) e localização geográfica (urbana ou rural). Para este estudo, foram utilizados 18 dos 22 atributos disponíveis, totalizando 65.535 registros. Está disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/indicadores-educacionais/taxas-de-rendimento-escolar/2023.

3.2. Pré-processamento dos Dados

O pré-processamento de dados constitui uma etapa fundamental para garantir a qualidade e confiabilidade das análises subsequentes. Neste estudo, foram utilizadas duas bases de dados principais: microdados_ed_basica_2023 e tx_rend_brasil_regioes_ufs_2023, cada uma exigindo tratamentos específicos para assegurar sua integridade analítica.

3.2.1. Identificação e Tratamento de Dados Ausentes

Durante a fase inicial da análise exploratória, foi realizada uma avaliação sistemática da completude dos dados nas bases utilizadas. Na base microdados_ed_basica_2023, identificaram-se 43 variáveis com mais de 20% de valores ausentes, representando um desafio relevante para a qualidade da modelagem. A Figura 1 ilustra a distribuição dessas variáveis em dois grupos distintos, conforme seus níveis de ausência.

No primeiro grupo de variáveis com dados ausentes, destacam-se aquelas relacionadas a informações cadastrais e vínculos institucionais, como NU_ENDERECO (20,25% de ausência) e IN_BANDA_LARGA (24,64%). Variáveis relacionadas a vínculos com secretarias e órgãos públicos apresentaram um padrão consistente de ausência em torno de 36,78%. Já as variáveis relacionadas a categorias e manutenção de escolas privadas demonstraram aproximadamente 80,36% de valores ausentes, enquanto as variáveis de contratos e parcerias exibiram os maiores percentuais de ausência, chegando a 95,45%.

O segundo grupo apresentou variáveis com percentuais ainda mais elevados de ausência, principalmente aquelas relacionadas a formas de contratação municipal (95,88%), reservas específicas (97,06%), e características indígenas e linguísticas (98,34% a 99,93%). Este padrão sugere uma concentração de dados ausentes em variáveis específicas, possivelmente relacionadas a características menos frequentes na população estudada.

3.2.2. Estratégia de Imputação dos Dados

Para preservar a integridade analítica do conjunto de dados, adotou-se uma abordagem metodológica rigorosa no tratamento dos valores ausentes. Para variáveis numéricas com percentual de ausência inferior a 50%, aplicaram-se técnicas de imputação estatística baseadas na distribuição dos dados: utilizou-se a média aritmética para variáveis com

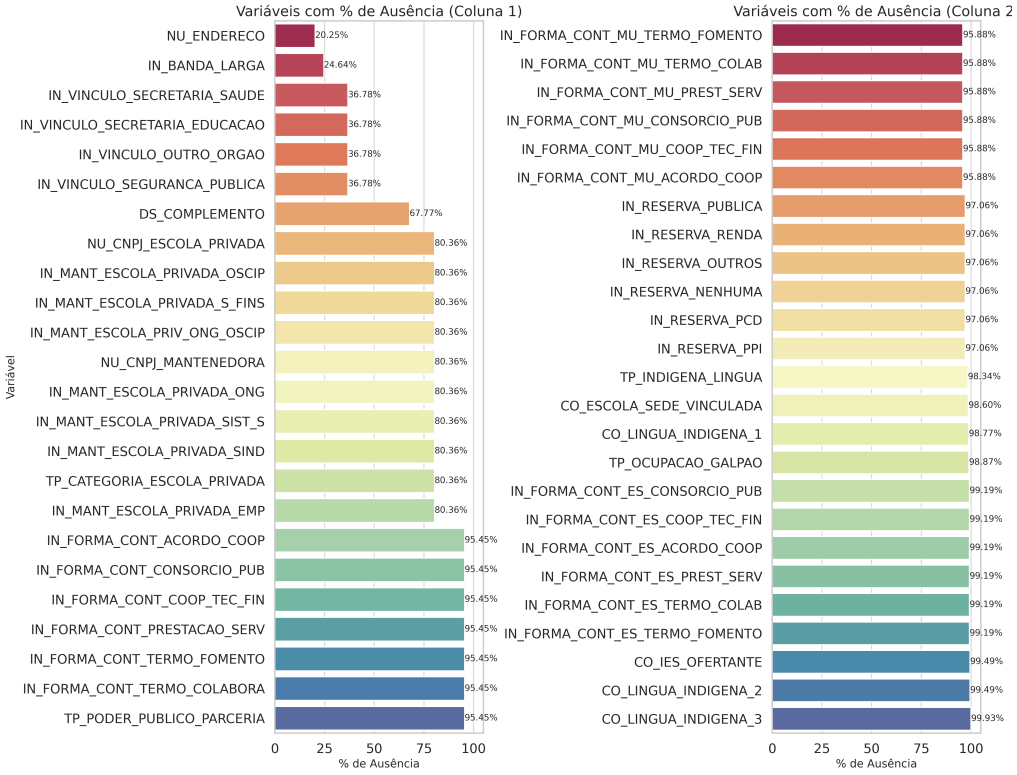

Figura 1. Variáveis com Valores Ausentes

distribuição aproximadamente normal e a mediana para variáveis com distribuição assimétrica.

A estratégia adotada teve como objetivo preservar a distribuição original das variáveis, evitando possíveis distorções nas análises subsequentes. Para aquelas com mais de 50% de valores ausentes, optou-se por não realizar imputação, a fim de mitigar a introdução de vieses. No caso da base `tx_rend_brasil_regioes_ufs_2023`, não foram identificados dados faltantes, o que dispensou a aplicação de técnicas de imputação nesse conjunto específico.

3.2.3. Resultados do Pré-processamento

A Figura 2 apresenta uma comparação entre as bases de dados antes e após o tratamento. Na base `microdados_ed_basica_2023`, houve redução de registros de 218.656 para 217.625, devido à exclusão de entradas com excesso de valores ausentes. A matriz de visualização demonstra maior densidade e homogeneidade após a imputação, indicando um conjunto de dados mais consistente e adequado para análises estatísticas. Na base `tx_rend_brasil_regioes_ufs_2023`, manteve-se o total de 65.535 registros, confirmando a ausência de valores faltantes e garantindo confiabilidade nas comparações regionais.

O pré-processamento adotado equilibra integridade e completude dos dados, com imputação restrita a variáveis com até 50% de ausência, utilizando medidas de tendência central conforme a distribuição. Essa abordagem minimiza distorções e assegura robustez

às análises subsequentes.

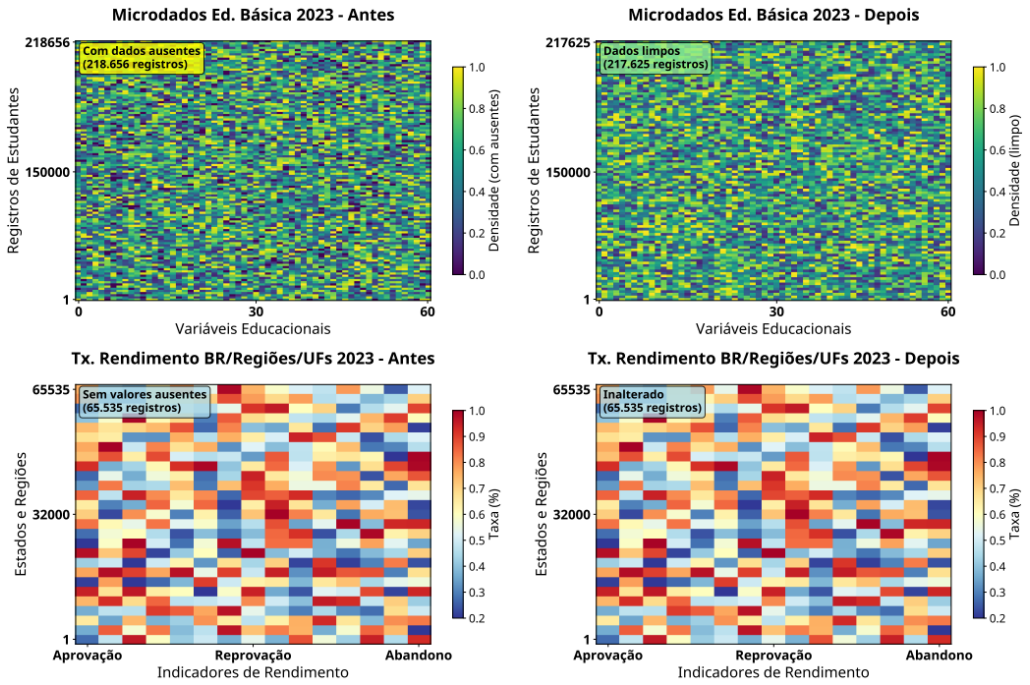

Figura 2. Comparativo de Tratamento das Bases

3.3. Mineração de Dados

A fase de mineração de dados, etapa central do processo KDD, foi conduzida por meio da aplicação de três algoritmos supervisionados, escolhidos por sua complementaridade analítica e capacidade de modelar diferentes tipos de relações entre as variáveis educacionais. A Regressão Linear foi empregada como modelo baseline, dada sua simplicidade e interpretabilidade, permitindo avaliar a contribuição individual de cada variável para a evasão escolar [Fayyad et al. 1996].

Em seguida, foi utilizado o Random Forest, um modelo de ensemble baseado em árvores de decisão, adequado para capturar relações não-lineares e reduzir o risco de sobreajuste por meio da combinação de múltiplas árvores treinadas em subconjuntos aleatórios dos dados [Breiman 2001].

O XGBoost (Extreme Gradient Boosting) representou a abordagem mais sofisticada implementada, utilizando o princípio de boosting sequencial com regularização avançada [Chen and Guestrin 2016]. Este algoritmo de última geração foi configurado para otimizar o equilíbrio entre viés e variância, incorporando mecanismos de penalização que favorecem modelos mais parcimoniosos e generalizáveis.

3.4. Interpretação e Avaliação dos Resultados

Nesta seção, são apresentados os resultados obtidos a partir da análise dos dados educacionais, com o objetivo de promover uma reflexão sobre a predição dos estados e municípios com maior probabilidade de evasão escolar. A análise permitiu identificar os locais mais vulneráveis, considerando indicadores históricos e atuais. Além disso, foi realizado um comparativo entre diferentes modelos de aprendizado de máquina, a fim de avaliar seu

desempenho preditivo. O estudo também destacou os atributos e indicadores que mais contribuíram para a evasão escolar, fornecendo subsídios para ações estratégicas voltadas à permanência dos estudantes na escola.

4. Resultados e Discussões

Nesta seção, apresentam-se os resultados obtidos a partir da análise dos dados, com ênfase nos rendimentos educacionais, nos indicadores de infraestrutura escolar e no desempenho dos modelos de aprendizado de máquina aplicados às bases utilizadas. Os modelos demonstraram elevados níveis de precisão e capacidade preditiva, evidenciando a robustez das técnicas adotadas. Entretanto, ressalta-se que a eficácia desses algoritmos está diretamente condicionada à qualidade dos dados disponíveis e à integração adequada das variáveis selecionadas.

4.1. Comparação do Desempenho dos Modelos Preditivos por Segmento de Ensino

A avaliação do desempenho dos modelos preditivos na estimativa da taxa de abandono escolar na educação básica foi realizada a partir das métricas R^2, RMSE e MAE, considerando os segmentos dos Anos Iniciais, Anos Finais e Ensino Médio. O objetivo da análise foi verificar a eficácia dos modelos Random Forest, Regressão Linear e XGBoost na previsão da evasão escolar, utilizando dados educacionais combinados a indicadores de infraestrutura. Os resultados dessa comparação estão apresentados na Tabela 1.

Tabela 1 Comparação do desempenho dos modelos preditivos por segmento de ensino

Segmento	Modelo	R^2	RMSE	MAE
Anos Iniciais	Regressão Linear	-0,099	N/D	2,46
	Random Forest	Negativo	3,60	1,58
	XGBoost	Negativo	3,60	1,58
Anos Finais	Regressão Linear	0,048	6,42	4,06
	Random Forest	0,431	N/D	N/D
	XGBoost	0,431	4,96	2,91
Ensino Médio	Regressão Linear	0,218	N/D	4,44
	Random Forest	0,493	N/D	N/D
	XGBoost	0,503	5,16	2,70

Conclui-se que o XGBoost é o modelo mais eficaz para a predição da evasão escolar em todos os segmentos analisados, com o Random Forest como alternativa viável. A Regressão Linear mostrou desempenho significativamente inferior, sendo desaconselhada para esse tipo de tarefa.

4.2. Análise da Importância das Variáveis nos Modelos de Predição

Este estudo analisa o impacto da infraestrutura escolar nos Anos Iniciais, Anos Finais e Ensino Médio, utilizando os modelos Random Forest e XGBoost para identificar as variáveis com maior influência sobre o desempenho educacional.

A Tabela 2 apresenta a importância relativa das variáveis de infraestrutura escolar nos modelos Random Forest e XGBoost, conforme a etapa de ensino. Nos Anos Iniciais, destacaram-se variáveis relacionadas a espaços físicos, como a Quadra de Esportes, Refeitório e Água da Rede Pública. Nos Anos Finais, recursos didáticos e científicos, como Biblioteca e Laboratório de Ciências, tornaram-se mais relevantes.

No Ensino Médio, observou-se uma combinação entre fatores físicos e acadêmicos, com destaque para Refeitório, Água da Rede Pública e Biblioteca. A Água da Rede Pública apresentou impacto consistente nas três etapas (valores entre 0,093 e 0,108), evidenciando seu papel essencial. A análise comparativa dos modelos revela que as demandas por infraestrutura variam conforme o nível de ensino, refletindo a evolução das necessidades pedagógicas e estruturais no processo de permanência escolar.

Tabela 2 Importância das variáveis de infraestrutura por modelo e etapa de ensino

Infraestrutura Escolar	**Anos Iniciais**		**Anos Finais**		**Ensino Médio**	
	RF	XGB	RF	XGB	RF	XGB
Quadra de Esportes	0,118	0,083	0,073	0,078	0,104	0,092
Refeitório	0,113	0,081	0,096	0,072	0,108	0,101
Água da Rede Pública	0,093	0,100	0,105	0,092	0,108	0,092
Sala de Atendimento Especial	0,095	0,093	0,092	0,081	0,080	0,084
Internet	0,083	0,095	0,056	0,046	0,081	0,075
Água Potável	0,080	0,087	0,086	0,098	0,058	0,070
Biblioteca	0,093	0,068	0,109	0,083	0,103	0,084
Laboratório de Ciências	0,085	0,075	0,085	0,090	0,081	0,085
Esgoto da Rede Pública	0,072	0,080	0,103	0,079	0,081	0,081
Prédio Escolar	0,081	0,064	0,058	0,094	0,060	0,088
Banheiro	0,056	0,088	0,065	0,095	0,059	0,090
Energia da Rede Pública	0,050	0,069	0,076	0,085	0,075	0,058

Legenda: RF = Random Forest; XGB = XGBoost.

Foram observadas variações entre os modelos, como nos pesos atribuídos à presença de banheiros no Ensino Médio (0,059 no Random Forest e 0,090 no XGBoost) e ao acesso à internet nos Anos Finais (0,046 vs. 0,056). Tais diferenças indicam a necessidade de políticas direcionadas: fortalecimento de espaços físicos e alimentação nos Anos Iniciais, ampliação de bibliotecas e laboratórios nos Anos Finais, e investimentos multidimensionais no Ensino Médio.

4.3. Análise dos Dados de Evasão Escolar por Estado

Este estudo investigou padrões regionais de evasão escolar no Brasil por meio de modelos preditivos, visando compreender e enfrentar esse fenômeno. A análise abrangeu 107 municípios em 27 estados, utilizando Regressão Linear, Random Forest e XGBoost. Conforme a Tabela 3, a taxa média nacional de evasão foi de 7,68%, sendo que 53% dos municípios superaram esse índice. As maiores taxas concentraram-se nos estados do Norte e Nordeste, com destaque para Alagoas (13,33%), Amazonas (11,68%) e Amapá (10,82%). Em contraste, Goiás (3,86%), Rio Grande do Sul (4,66%) e Distrito Federal

(5,18%) apresentaram os menores índices, evidenciando disparidades regionais associadas a fatores socioeconômicos.

Tabela 3 Desempenho dos modelos preditivos por estado

Estado	Nº Municípios	Média Evasão	% Acima Média	Acurácia LR	Acurácia RF	Acurácia XGB	RMSE LR	RMSE RF	RMSE XGB
AC	6	7.49	50.0%	33.3%	66.7%	66.7%	3.51	3.96	4.15
AL	3	13.33	100.0%	33.3%	66.7%	66.7%	6.08	5.21	5.42
AM	4	11.68	100.0%	50.0%	50.0%	75.0%	4.06	3.63	3.46
AP	1	10.82	100.0%	100.0%	100.0%	100.0%	2.62	0.72	1.27
BA	5	7.55	60.0%	80.0%	60.0%	60.0%	4.79	3.89	3.64
CE	5	5.84	20.0%	40.0%	40.0%	60.0%	4.18	4.47	3.96
DF	4	5.18	25.0%	75.0%	50.0%	25.0%	4.36	4.25	5.24
ES	5	7.21	60.0%	60.0%	40.0%	60.0%	4.31	5.79	6.17
GO	3	3.86	0.0%	66.7%	100.0%	66.7%	3.21	2.86	4.36
MA	6	7.65	50.0%	66.7%	66.7%	66.7%	3.43	2.57	4.55
MG	6	5.89	33,3%	50,0%	50,0%	50,0%	3.38	3.15	3.08
MS	2	5.66	50.0%	50.0%	100.0%	100.0%	2.00	1.78	1.65
MT	6	8.39	50.0%	50.0%	100.0%	50.0%	4.21	4.48	4.52
PA	7	7.36	57.1%	28.6%	14.3%	42.9%	4.27	4.87	4.89
PB	9	7.27	44.4%	55.6%	33.3%	22.2%	4.58	5.52	5.83
PE	5	8.82	80.0%	60.0%	60.0%	40.0%	2.96	3.53	5.88
PI	2	5.31	50.0%	100.0%	100.0%	100.0%	4.63	2.35	1.05
PR	1	8.89	100.0%	100.0%	100.0%	100.0%	1.21	1.29	0.94
RJ	2	6.45	50,0%	50,0%	50,0%	50,0%	3.92	3.85	3.76
RN	3	9.63	33.3%	0.0%	33.3%	33.3%	3.58	4.60	7.35
RO	3	8.76	66.7%	33.3%	0.0%	33.3%	4.59	6.68	7.27
RR	2	5.89	0.0%	100.0%	50.0%	100.0%	1.54	3.05	2.98
RS	4	4.66	25.0%	100.0%	100.0%	100.0%	5.39	3.74	2.78
SC	2	7.51	50.0%	50.0%	50.0%	50.0%	1.84	2.32	2.32
SE	3	9.04	66.0%	40.0%	50.0%	40.0%	4.70	4.34	5.03
SP	4	8.86	75.0%	75.0%	75.0%	75.0%	2.17	2.38	2.30
TO	3	6.90	66.7%	0.0%	66.7%	66.7%	3.88	4.35	2.92

Legenda: LR – Regressão Logística; RF – Random Forest; XGB – XGBoost.

A avaliação dos modelos preditivos revelou acurácia média de 53% para Regressão Linear e Random Forest, e 55% para XGBoost. O erro médio (RMSE) foi de 3,97 (LR), 4,08 (RF) e 4,45 (XGB), caracterizando desempenho moderado. A análise por estado mostrou variação significativa: Amapá, Paraná, Piauí, e Rio Grande do Sul alcançaram 100% de acurácia, enquanto Alagoas, Rondônia e Rio Grande do Norte tiveram os maiores erros médios, com 18,54, 16,71 e 15,53, respectivamente. Observou-se correlação inversa entre taxa de evasão e acurácia, com melhor desempenho em estados com taxas extremas. O XGBoost destacou-se em contextos de alta evasão, enquanto a Regressão Linear apresentou melhor resultado em cenários de baixa evasão. Esses achados reforçam a relevância de considerar especificidades regionais na modelagem preditiva e a adoção de estratégias adaptadas a diferentes realidades educacionais.

5. Considerações Finais

Este estudo demonstrou que a evasão escolar na educação básica brasileira apresenta padrões regionais significativos e é influenciada por múltiplos fatores estruturais, cuja importância varia conforme a etapa de ensino. A aplicação de modelos preditivos de machine learning com destaque para o XGBoost, que obteve melhor desempenho no Ensino Médio (R^2 = 0,503), evidenciou a capacidade dessas abordagens em capturar relações complexas e não lineares nos dados educacionais.

A principal contribuição deste estudo está na quantificação da importância relativa dos elementos de infraestrutura escolar na predição da evasão escolar. A análise evidenciou que as necessidades infraestruturais variam de acordo com o nível de ensino: nos Anos Iniciais, destacaram-se o refeitório e a quadra de esportes; nos Anos Finais, a bibli-

oteca assumiu maior relevância; e no Ensino Médio, observou-se uma combinação desses fatores, com ênfase no refeitório, na presença de água encanada e na quadra de esportes.

Regionalmente, Norte e Nordeste concentraram as maiores taxas de evasão (Alagoas: 13,33%; Amazonas: 11,68%; Amapá: 10,82%), enquanto o Centro-Sul registrou os menores índices (Goiás: 3,86%; Rio Grande do Sul: 4,66%), refletindo desigualdades socioeconômicas. Observou-se ainda que a predição foi mais precisa em contextos com taxas extremas de evasão, indicando maior previsibilidade nesses casos.

Conclui-se que modelos baseados em árvores, aliados à análise de infraestrutura escolar, são eficazes para prever evasão e subsidiar políticas públicas. Pesquisas futuras podem incorporar variáveis socioemocionais e familiares, além de séries temporais, para aprimorar a capacidade preditiva e fortalecer estratégias contra a evasão escolar no Brasil.

Referências

BRASIL, M. (2023). Instituto nacional de estudos e pesquisas educacionais anísio teixeira (inep). *Censo Escolar*. Acessado em: 30 out. 2024.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

de Brito, J. J. R. T., da Silva, G. C., Rodrigues, R. L., Amorim, A. N. G. F., et al. (2022). A relação entre nível socioeconômico e proficiência em matemática de estudantes pernambucanos do 9º ano através da mineração de dados educacionais. *Amazônia: Revista de Educação em Ciências e Matemáticas*, 18(41):112–126.

de Faria, D. R., Silvestre, J. V. R., de Oliveira Neto, P. M., and Silva, V. E. P. (2022). A utilização de aprendizado de máquina supervisionado para predição de evasão no ensino superior.

de Jesus, J. A. and de Gusmão, R. P. (2024). Investigação da evasão estudantil por meio da mineração de dados e aprendizagem de máquina: Um mapeamento sistemático. *Revista Brasileira de Informática na Educação*, 32.

de Oliveira Almeida et al., C. (2021). Políticas educacionais brasileiras para ejai e a educação permanente. *Revista HISTEDBR On-line*, 21:e021025–e021025.

do Nascimento, R. L. S., da Cruz Junior, G. G., and de Araújo Fagundes, R. A. (2018). Mineração de dados educacionais: Um estudo sobre indicadores da educação em bases de dados do inep. *Revista Novas Tecnologias na Educação*, 16(1).

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37–37.

Gramacho, W. G. M. (2019). Algoritmos de mineração de dados para análise de evasão na graduação da universidade de brasília.

Lima, J. A. and Fagundes, R. A. (2023). Análise comparativa de algoritmos de machine learning para prever a evasão escolar: Uma revisão sistemática da literatura. *Revista Novas Tecnologias na Educação*, 21(2):362–371.