https://doi.org/10.47734/regrasp.v11.01.p225-252

INSTITUTO FEDERAL
São Paulo
Campus São Paulo
REGRASP

Análise da evasão escolar no Brasil com aprendizagem de máquina: evidências a partir de dimensões regionais, institucionais e da renda média

Analysis of school dropout in Brazil using machine learning: evidence from regional, institutional, and average income dimensions

Laura Rabelo de Carvalho Ferreira
Graduada em Tecnologia em Análise e Desenvolvimento de Sistemas
Instituto Federal de Educação, Ciência e Tecnologia de São Paulo
https://orcid.org/0009-0007-4306-1849
laura.carvalho@aluno.ifsp.edu.br

Glauber da Rocha Balthazar
Doutor em Engenharia
Instituto Federal de Educação, Ciência e Tecnologia de São Paulo
https://orcid.org/0000-0002-1993-6621
glauber.balthazar@ifsp.edu.br

Alexandre Beletti Ferreira
Doutor em Engenharia
Instituto Federal de Educação, Ciência e Tecnologia de São Paulo
https://orcid.org/0009-0006-3954-3119
higuita@ifsp.edu.br

Histórico do artigo
Recebido em: 24/03/2026
Revisões solicitadas: 23/04/2026
Versão reformulada recebida: 28/04/2026
Aprovado em: 07/05/2026
Publicado em: 07/05/2026

RESUMO
O presente trabalho tem como objetivo caracterizar o abandono escolar no Brasil a partir da aplicação de técnicas estatísticas e modelos de aprendizado de máquina, considerando fatores regionais, socioeconômicos e institucionais. Para isso, foi utilizada uma base de dados do INEP referente ao período de 2013 a 2023, segmentada por tipo de escola (pública ou privada), nível de ensino (fundamental ou médio), unidade geográfica e renda média familiar. A análise descritiva revelou padrões distintos entre os grupos educacionais, com maiores taxas de abandono concentradas no ensino médio da rede pública. Foram aplicados testes estatísticos para verificar normalidade, homocedasticidade e diferenças entre os grupos, além da regressão linear simples para identificar tendências por região. Na etapa preditiva, diversos algoritmos de *machine learning* foram testados (Random Forest, XGBoost, Gradient Boosting, Regressão Linear, SVM, KNN e ANN), sendo o Random Forest o modelo com melhor desempenho preditivo (MAE = 0,69; R^2 = 0,86). A análise evidenciou que variáveis como renda média e tipo de escola influenciam significativamente a evasão escolar, destacando a importância de políticas públicas focadas na redução da desigualdade educacional.
Palavras-chave: evasão escolar; aprendizado de máquina; análise estatística; fatores socioeconômicos; previsão educacional.

ABSTRACT
This study aims to characterize school dropout in Brazil using statistical analysis and machine learning models, taking into account regional, socioeconomic, and institutional factors. The analysis was based on data from INEP covering the period from 2013 to 2023, segmented by school type (public or private), educational level (elementary or high school), geographic region, and average household income. Descriptive statistics revealed distinct patterns among educational groups, with the highest dropout rates found in public high

schools. Statistical tests were conducted to assess normality, homoscedasticity, and intergroup differences, as well as linear regression to identify regional trends. In the predictive phase, several machine learning algorithms were tested (Random Forest, XGBoost, Gradient Boosting, Linear Regression, SVM, KNN, and ANN), with Random Forest achieving the best performance (MAE = 0.69; R^2 = 0.86). The analysis showed that factors such as income and school type significantly affect dropout rates, emphasizing the need for public policies aimed at reducing educational inequality.

Keywords: school dropout; machine learning; statistical analysis; socioeconomic factors; educational forecasting.

Introdução

O abandono escolar é um dos principais desafios da educação brasileira, impactando diretamente o desenvolvimento social e econômico do país (Araújo et al., 2025; Santos, 2020; Baggi & Lopes, 2011). A evasão escolar pode estar associada a diversos fatores, incluindo condições socioeconômicas, estrutura do sistema educacional, características institucionais das escolas e desigualdades regionais (Araújo et al., 2025).

A taxa de abandono escolar pode ser influenciada por variáveis como a renda média familiar, o nível de ensino, a região do país e a dependência administrativa das instituições de ensino (classificadas como públicas ou privadas) (Araújo et al., 2025). Diversos estudos (Teodoro & Kappel, 2020; Sousa et al., 2018; Ramos & Gonçalves Junior, 2024) demonstram que estudantes de baixa renda, especialmente em determinadas regiões brasileiras, estão mais propensos à evasão escolar devido a dificuldades financeiras, necessidade de ingresso precoce no mercado de trabalho e falta de suporte escolar adequado. Além disso, a qualidade da infraestrutura e o tipo de gestão escolar podem impactar significativamente a permanência dos alunos nos estudos (Santos, 2020).

A Inteligência Artificial (IA) tem sido amplamente utilizada para a análise de grandes volumes de dados e para a construção de modelos preditivos em diversas áreas, incluindo a educação (Camargos & Silveira, 2024; Jesus et al., 2021; Teodoro & Kappel, 2020). Métodos de aprendizado de máquina permitem identificar padrões nos dados e prever possíveis tendências, auxiliando na tomada de decisões estratégicas. No contexto deste estudo, parte-se da hipótese de que a aplicação de modelos preditivos pode contribuir para a identificação antecipada de contextos escolares e regiões com maior risco de evasão.

Neste trabalho, foram utilizados algoritmos de aprendizado de máquina para prever a evasão escolar com base em fatores educacionais, regionais e institucionais, utilizando bases de dados extraídas exclusivamente do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). A análise foi segmentada por tipo de escola (pública ou privada), por nível de ensino (fundamental e médio) e por região do Brasil. Além da análise preditiva, realizou-se também uma análise exploratória para compreender a distribuição das variáveis e identificar correlações relevantes com o abandono escolar.

Os modelos empregados neste estudo incluem Random Forest, Gradient Boosting, Regressão Linear, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Redes Neurais Artificiais (ANN) e XGBoost. A escolha desses modelos visa capturar diferentes relações entre as variáveis e fornecer previsões precisas sobre a evasão escolar, considerando as diferenças entre escolas públicas e privadas, níveis de ensino e regiões.

A Regressão Linear foi utilizada devido à sua capacidade de fornecer um modelo interpretável, permitindo compreender o impacto de fatores como tipo de escola, nível de ensino e localização regional na taxa de evasão. O Random Forest foi escolhido por sua robustez para dados complexos e por identificar variáveis com maior influência na permanência dos alunos. O Gradient Boosting, com seu aprendizado sequencial, permite aprimorar continuamente as previsões ao captar interações sutis entre os fatores. O modelo K-Nearest Neighbors (KNN) analisa perfis semelhantes de estudantes para prever padrões locais de evasão. Já o Support Vector Machines (SVM) é eficiente na classificação de alunos entre aqueles com alta e baixa probabilidade de abandono, considerando a complexidade dos dados. As Redes Neurais Artificiais (ANN) modelam relações não lineares entre variáveis educacionais, aumentando a precisão preditiva. Por fim, o XGBoost, reconhecido por sua performance em grandes volumes de dados, otimiza a atribuição de pesos às variáveis, tornando-se uma ferramenta poderosa para análise regionalizada e institucional da evasão escolar.

Justificativa

O abandono escolar representa um dos maiores desafios do sistema educacional brasileiro. Segundo estudos (Araújo et al., 2025; Santos, 2020; Baggi & Lopes, 2011), a evasão pode ser impulsionada por uma combinação de fatores socioeconômicos, institucionais e regionais, que afetam de maneira desigual diferentes contextos escolares. A desigualdade social, a necessidade de ingresso precoce no mercado de trabalho e a ausência de políticas públicas eficazes continuam sendo elementos-chave que agravam esse cenário (Ramos & Gonçalves Junior, 2024).

Além dos impactos individuais causados aos estudantes, como a restrição de oportunidades e o comprometimento de seu desenvolvimento pessoal, o abandono escolar compromete também o desenvolvimento econômico do país, pois reduz o nível de qualificação da força de trabalho e amplia as desigualdades sociais (Rosa et al., 2023; Barbosa, 2021). Por isso, a identificação antecipada dos fatores associados à evasão escolar é fundamental para o planejamento de ações educativas eficazes.

A Inteligência Artificial (IA) surge como uma ferramenta promissora nesse contexto, com potencial para analisar grandes volumes de dados e identificar padrões de risco com mais precisão (Lopes Filho, 2021). O uso de modelos preditivos baseados em aprendizado de máquina tem se mostrado eficaz em diversos países para antecipar o abandono escolar e subsidiar políticas públicas baseadas em evidências (Sousa et al., 2018).

Este trabalho busca desenvolver e avaliar um modelo preditivo capaz de estimar a taxa de abandono escolar com base em fatores institucionais e regionais, considerando o nível de ensino (fundamental e médio) e a dependência administrativa das escolas (públicas e privadas) em diferentes regiões do Brasil, a partir de dados extraídos do INEP.

Objetivos

Este trabalho tem como objetivo geral, desenvolver e avaliar um modelo preditivo de Inteligência Artificial baseado em técnicas de Aprendizagem de Máquina para prever a taxa de abandono escolar considerando fatores socioeconômicos, regionais e institucionais.

Além disso, possuí como objetivos específicos os seguintes itens:

- Analisar a relação entre fatores institucionais e regionais e o abandono escolar, com ênfase na dependência administrativa (pública e privada) e no nível de ensino (fundamental e médio);
- Comparar as taxas de evasão entre escolas públicas e privadas em diferentes regiões do Brasil;
- Aplicar diferentes modelos de aprendizado de máquina para prever a evasão escolar utilizando a função ‘predict()’ para estimar padrões de abandono escolar com base nos dados institucionais e regionais e;
- Avaliar a eficácia dos modelos preditivos por meio de métricas estatísticas.

Abandono Escolar no Brasil e Modelos de Aprendizagem de Máquina

A evasão escolar no ensino fundamental e médio é um dos desafios mais complexos da educação brasileira. Segundo Araújo et al. (2025), as principais causas deste problema incluem fatores socioeconômicos, regionais, dificuldades de aprendizagem, problemas familiares e falta de infraestrutura escolar adequada. Além disso, o desinteresse pelo conteúdo curricular e a defasagem no aprendizado contribuem para o afastamento progressivo dos estudantes da escola.

Santos (2020) aponta que a evasão escolar é mais acentuada entre adolescentes que enfrentam dificuldades em conciliar os estudos com outras responsabilidades, como o trabalho informal e demandas familiares. Ramos & Gonçalves Junior (2024) destacam que a

percepção dos alunos sobre a qualidade da educação e a relação com professores e colegas são fatores determinantes para sua permanência na escola.

O uso de técnicas de aprendizado de máquina tem sido cada vez mais adotado para a análise de fatores relacionados à evasão escolar. Segundo Teodoro & Kappel (2020), algoritmos como Random Forest, XGBoost, Redes Neurais Artificiais (ANN) e Support Vector Machines (SVM) são empregados para examinar padrões nos dados educacionais e identificar tendências associadas ao abandono dos estudos. De acordo com Jesus et al. (2021), variáveis como frequência escolar, notas, perfil socioeconômico e escolaridade dos responsáveis são frequentemente analisadas para compreender melhor os desafios enfrentados pelos estudantes. Camargos & Silveira (2024) ressaltam a importância de utilizar modelos interpretáveis para que os profissionais da educação possam entender os padrões subjacentes aos dados e agir de maneira informada.

A pesquisa de Lopes Filho (2021) sugere que a combinação de diferentes técnicas de aprendizado de máquina pode oferecer uma visão mais abrangente dos fatores que influenciam a evasão. Baggi & Lopes (2011) reforçam que a qualidade da base de dados utilizada é um aspecto essencial para que os modelos geram informações mais precisas sobre os padrões de abandono escolar no ensino fundamental e médio.

Na sequência apresentamos uma breve descrição de cada uma das técnicas utilizadas neste trabalho.

- Random Forest: é um algoritmo baseado em árvores de decisão que opera criando múltiplas árvores e combinando seus resultados para obter uma predição mais precisa (Breiman, 2001). Esse método é amplamente utilizado devido à sua capacidade de lidar com conjuntos de dados de alta dimensionalidade e sua robustez contra o *overfitting*. Cada árvore da floresta é construída a partir de um subconjunto aleatório dos dados de treinamento, e o resultado final do modelo é obtido através da média ou votação das previsões individuais das árvores. Essa característica torna o Random Forest uma escolha eficiente para a análise da evasão escolar, pois permite capturar relações complexas entre variáveis como frequência escolar, desempenho acadêmico e fatores socioeconômicos.

- Gradient Boosting: é um método de aprendizado supervisionado que combina várias árvores de decisão fracas para formar um modelo mais forte e preciso (Friedman, 2001). Diferente do Random Forest, que cria árvores independentes, o Gradient Boosting constrói sequencialmente árvores onde cada nova árvore corrige os erros cometidos pelas gerações anteriores. Esse modelo é especialmente útil para lidar com problemas complexos, como a evasão escolar, onde múltiplos fatores podem influenciar o comportamento dos alunos. Ele ajusta gradativamente suas previsões, minimizando o erro residual e melhorando a capacidade de generalização do modelo.

- Regressão Linear: é um modelo estatístico que estabelece uma relação entre uma variável dependente e uma ou mais variáveis independentes por meio de uma equação linear (Montgomery et al., 2012). Esse modelo pode ser útil para identificar tendências na evasão escolar, como a relação entre frequência escolar e desempenho acadêmico.

- K-Nearest Neighbors (KNN): é um método baseado em instâncias que classifica novos dados com base na similaridade com os vizinhos mais próximos (Cover & Hart, 1967). Ele calcula a distância entre pontos de dados e atribui uma classe com base na maioria das observações mais próximas.

- Support Vector Machines (SVM): é um modelo que busca encontrar um hiperplano ótimo que separe os dados em diferentes classes, maximizando a margem entre os pontos de dados mais próximos (Cortes & Vapnik, 1995).

- Redes Neurais Artificiais (ANN): são inspiradas no funcionamento do cérebro humano e são compostas por neurônios artificiais organizados em camadas (Haykin, 2009).

- XGBoost (Extreme Gradient Boosting): é uma implementação avançada do algoritmo de Gradient Boosting, desenvolvida para ser altamente eficiente, flexível e portátil. Proposto por Chen e Guestrin (2016), o XGBoost introduz técnicas de regularização (L1 e L2) em seu processo de treinamento, o que contribui para reduzir o risco de *overfitting* — um dos principais problemas em modelos de aprendizado supervisionado.

Método de Pesquisa

O desenvolvimento deste estudo foi estruturado em três etapas principais. A primeira etapa consistiu na obtenção, pré-processamento e análise exploratória, onde os dados foram preparados, tratados e analisados estatisticamente para identificar padrões iniciais e relações entre as variáveis. A segunda etapa envolveu a implementação dos modelos preditivos, utilizando algoritmos de aprendizado de máquina e séries temporais para prever a taxa de abandono escolar com base em fatores socioeconômicos, regionais e educacionais. Por fim, a terceira etapa compreendeu a avaliação dos resultados obtidos, analisando o desempenho dos modelos, interpretando os fatores mais influentes na evasão escolar e validando a eficácia das previsões geradas. A seguir, temos cada um dos itens do método de pesquisa devidamente especificado e caracterizado.

- **Base de Dados**: As bases de dados utilizadas neste estudo foram obtidas a partir dos dados do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), envolvendo informações consolidadas sobre evasão escolar, tipo de escola (pública/privada), nível de ensino (fundamental/médio), unidade geográfica e a variável Renda Média estimada por região. O INEP mantém os dados em formato aberto desde que não estejam sob sigilo ou restrição de acesso nos termos da Lei 12.527, de 18 de novembro de 2011, sendo que seu uso esteja de acordo com o Art. 3º da Política de Dados Abertos do Inep (INEP, 2020).

- **Tratamento de Dados**: Técnicas de detecção de anomalias, como o Z-Score (Escovedo, 2022), limpeza de dados faltantes e padronização de formatos de variáveis foram utilizados como procedimentos de tratamento estatísticos. Para tanto, foram empregadas técnicas de imputação (substituição de dados faltantes por valores apropriados) (Moore, 2023) para tratar valores ausentes, garantindo a integridade e a consistência dos dados para análises futuras. Estes procedimentos possibilitaram a eliminação de valores ausentes, duplicatas e *outliers* que poderiam comprometer os resultados (Escovedo, 2022; Bruce, 2019).

- **Análise Descritiva**: Para a compreensão inicial da base de dados e fundamentação das análises preditivas, foi conduzida uma análise estatística descritiva que incluiu o cálculo de medidas de tendência central (média, mediana e moda), dispersão (desvio padrão, variância, amplitude, intervalo interquartílico – IQR), e forma da distribuição dos dados (curtose e assimetria). Essa avaliação ofereceu uma perspectiva inicial sobre como os dados se comportam e auxiliou na identificação de possíveis padrões ou tendências. Para tanto, os dados foram segmentados em quatro grupos com base no tipo de escola (pública ou privada) e no nível de ensino (fundamental ou médio). Os grupos definidos foram: i) Público de Ensino Fundamental (PuEF); ii) Público de Ensino Médio (PuEM); iii) Privado de Ensino Fundamental (PrEF) e iv) Privado de Ensino Médio (PrEM). Gráficos descritivos foram gerados para cada grupo, incluindo boxplots, histogramas, gráficos de dispersão com linha da média e gráficos de intervalo interquartil, permitindo identificar a presença de *outliers*, assimetrias e padrões distintos de distribuição.

- **Análise Inferencial:** Com o objetivo de avaliar a significância das diferenças entre os grupos definidos, foram aplicados testes estatísticos clássicos. O teste T foi utilizado para comparar as médias e medianas da taxa de abandono dentro de cada grupo, com base na hipótese nula de que não existem diferenças significativas entre média e mediana.

A normalidade dos dados foi verificada por meio do teste de Shapiro-Wilk, o qual indicou, na maioria dos grupos, que os dados não seguem uma distribuição normal ($p < 0,05$). Em seguida, o teste de Levene foi utilizado para avaliar a homogeneidade das variâncias entre os grupos, demonstrando heterocedasticidade em vários casos ($p < 0,05$).

Considerando a violação das suposições de normalidade e homocedasticidade, foi adotado o teste ANOVA de Welch para comparação entre médias dos grupos. Os resultados revelaram diferenças estatisticamente significativas nas comparações entre escolas públicas e privadas, tanto no Ensino Fundamental quanto no Médio ($p < 0,05$), evidenciando desigualdades relevantes no comportamento da taxa de abandono escolar.

As análises foram realizadas no ambiente Google Colab com o uso de bibliotecas como scipy.stats para testes T, Shapiro-Wilk e Levene, e ttest_ind com equal_var=False para o ANOVA de Welch. Essa abordagem permitiu assegurar robustez estatística, mesmo em cenários com violação das premissas clássicas dos testes paramétricos.

- **Desenvolvimento dos Modelos Preditivos:** A modelagem preditiva foi realizada a partir da aplicação de técnicas de aprendizado de máquina. Para isso, foram utilizados os modelos Random Forest, Gradient Boosting, Regressão Linear, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Redes Neurais Artificiais (ANN) e XGBoost.

O modelo Random Forest, baseado em múltiplas árvores de decisão, foi escolhido por sua capacidade de capturar relações complexas entre as variáveis e lidar eficientemente com dados categóricos e contínuos. O Gradient Boosting foi utilizado por sua eficiência na redução de erros residuais ao combinar múltiplos modelos fracos em um preditor forte. A Regressão Linear foi empregada como modelo de referência para avaliar o impacto das variáveis socioeconômicas na taxa de abandono escolar, permitindo uma análise mais interpretável das relações entre os fatores estudados.

Além disso, o K-Nearest Neighbors (KNN) foi explorado devido à sua abordagem baseada em similaridade, enquanto o Support Vector Machines (SVM) foi incluído para avaliar a separabilidade das classes no problema de classificação. As Redes Neurais Artificiais (ANN) foram implementadas para modelar relações mais complexas entre os dados e melhorar a capacidade preditiva do sistema. O XGBoost, um modelo baseado em boosting, foi utilizado para aprimorar a precisão das previsões devido à sua capacidade de otimização e manejo de grandes volumes de dados.

A previsão da evasão escolar foi realizada por meio do método predict, que permite obter valores futuros a partir dos modelos treinados, analisando o comportamento da evasão ao longo do tempo e projetando cenários futuros.

- **Abordagem de execução:** A base de dados foi dividida em 80% para treinamento e 20% para teste, garantindo uma avaliação robusta dos modelos. Foram aplicadas técnicas como validação cruzada para minimizar o risco de *overfitting* e aprimorar a precisão das previsões. Dessa forma, foi possível construir um sistema preditivo confiável, capaz de oferecer *insights* sobre os fatores que influenciam o abandono escolar e permitir a projeção de tendências futuras. Essa divisão permitiu que o modelo pudesse aprender a partir da maior parte dos dados disponíveis, enquanto a parte reservada para teste possibilitou a avaliação do desempenho preditivo do modelo em dados não vistos. Essa abordagem reduziu o risco de sobre ajuste (*overfitting*), no qual o modelo poderia se tornar excessivamente ajustado aos dados de treinamento, perdendo a capacidade de generalização.

Para avaliar o desempenho dos modelos utilizados na previsão da evasão escolar foram aplicadas diferentes métricas de desempenho. Entre as métricas utilizadas para classificação destacam-se a Precisão que mede a proporção de previsões positivas corretas em relação ao total de previsões positivas realizadas pelo modelo e a Acurácia que avalia a proporção total de previsões corretas feitas pelo modelo. Além disso, foi calculado o Recall (Revocação) que

indica a capacidade do modelo de identificar corretamente as instâncias positivas e o F1-Score que representa a média harmônica entre Precisão e Recall. Também foi empregada a Matriz de Confusão, uma ferramenta visual que permite analisar os erros e acertos do modelo em relação às classificações corretas.

Além das métricas voltadas para classificação, foram utilizadas métricas para avaliar modelos de regressão. O Erro Quadrático Médio (MSE - Mean Squared Error) penaliza erros maiores foi útil para avaliar quão distantes as previsões estão dos valores reais. A Raiz do Erro Quadrático Médio (RMSE - Root Mean Squared Error) também foi utilizada, pois possibilitou quantificar o quanto as previsões do modelo se desviam, em média, dos valores reais. O Erro Absoluto Médio (MAE - Mean Absolute Error) mediu a média dos erros absolutos entre previsões e valores reais, sendo menos sensível a *outliers*. Por fim, o Coeficiente de Determinação (R^2) foi utilizado para medir o quão bem o modelo explica a variabilidade dos dados.

Resultados

Os experimentos realizados neste estudo foram aplicados sobre uma base de dados consolidada, abrangendo informações do Inep e de dados do Inep, com um total de registros cobrindo o período de 2013 a 2023. A base de dados continha informações organizadas por unidade geográfica, dependência administrativa, nível de ensino, taxa de abandono escolar e renda média familiar, permitindo a análise detalhada das influências socioeconômicas na evasão escolar.

Durante a construção dos modelos preditivos, os dados foram segmentados em diferentes conjuntos para avaliar a robustez das previsões. O desempenho dos algoritmos foi analisado a partir de métricas estatísticas, e os modelos foram testados com diferentes configurações para otimização dos resultados. A análise da distribuição dos dados revelou que a taxa de abandono escolar não é homogênea ao longo dos anos e varia significativamente entre regiões e tipos de escola.

As seções a seguir detalham os principais achados da análise exploratória, que buscou identificar padrões nos dados antes da aplicação dos modelos preditivos, e a avaliação do desempenho dos modelos, comparando os algoritmos utilizados e interpretando seus resultados.

- **Apresentação dos Dados:** As variáveis utilizadas neste estudo foram organizadas de modo a permitir a análise das relações entre fatores educacionais, regionais e socioeconômicos, viabilizando a construção de modelos preditivos da taxa de abandono escolar. A base de dados, oriunda exclusivamente do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), contempla o período de 2013 a 2023 e abrange informações relativas à unidade geográfica, dependência administrativa (pública ou privada), nível de ensino (fundamental ou médio), taxa de abandono e renda média da população por região.

A base original da Inep era formada por diversas informações relacionadas à escolaridade, incluindo categorias como "sem instrução", "ensino fundamental incompleto ou equivalente", "ensino fundamental completo ou equivalente", "ensino médio incompleto ou equivalente", "ensino médio completo ou equivalente", "ensino superior incompleto ou equivalente" e "ensino superior completo", além de dados organizados por Brasil, Grande Região e Unidade da Federação. Para este estudo, os dados da INEP foram utilizados para obter informações sobre a renda média das famílias e associados às unidades geográficas analisadas. Além disso, informações sobre a taxa de abandono escolar, categorizadas por região geográfica, nível de ensino, dependência administrativa e unidade federativa foram utilizadas para ser possível relacionar os padrões de evasão escolar com fatores econômicos e educacionais, permitindo a construção de um modelo preditivo mais preciso. O período analisado compreende os anos

de 2013 a 2023, garantindo um intervalo de tempo suficiente para identificar tendências e padrões na evasão escolar.

Antes da implementação dos modelos preditivos, foi realizada uma análise exploratória com o intuito de examinar a distribuição das variáveis, identificar padrões, tendências e eventuais inconsistências nos dados. Observou-se que a taxa de abandono escolar apresenta assimetrias em determinados grupos, especialmente no caso do Ensino Médio público (PuEM), em que valores superiores a 6% foram identificados em alguns estados e anos específicos. Esses resultados evidenciam uma heterogeneidade interna considerável, justificando a segmentação analítica por tipo de escola.

A análise revelou ainda a presença de valores atípicos (*outliers*) em todos os grupos analisados (PuEF, PuEM, PrEF e PrEM) conforme identificado por meio de gráficos boxplot e da avaliação do intervalo interquartílico (IQR). Tais valores não foram removidos, visto que refletem situações extremas, porém reais, associadas a contextos regionais específicos. Sua manutenção na base de dados foi considerada relevante para garantir uma representação fidedigna do fenômeno em análise, especialmente nas regiões marcadas por maiores níveis de vulnerabilidade.

A variável de renda média também demonstrou ampla dispersão entre as regiões brasileiras, indicando desigualdades estruturais com potencial impacto sobre os índices de evasão escolar. Verificou-se, por meio do teste de correlação de Pearson, uma associação negativa entre a renda média e a taxa de abandono escolar, sugerindo que regiões com menor renda tendem a apresentar maiores taxas de evasão. Os resultados desse teste estão descritos nas seções analíticas posteriores.

Além disso, os dados foram submetidos a testes de normalidade (Shapiro-Wilk) e homocedasticidade (Levene), os quais revelaram, em diversos casos, a não aderência à distribuição normal ($p < 0,05$ em todos os grupos, exceto PrEF e PrEM no Quadro 1) e a presença de variâncias heterogêneas entre os grupos (Levene com $p < 0,05$ para PuEF, PuEM e PrEM, conforme Quadro 3). Tais evidências respaldam o uso de técnicas estatísticas robustas e apropriadas às características empíricas do conjunto de dados.

A análise exploratória constituiu uma etapa essencial deste trabalho, contribuindo para a fundamentação das estratégias metodológicas adotadas. Ao revelar desigualdades internas e padrões relevantes entre grupos escolares e regiões, essa etapa reforçou a importância da segmentação analítica e do desenvolvimento de políticas públicas sensíveis às diferentes realidades do sistema educacional brasileiro.

- **Variáveis Categóricas:** As variáveis categóricas foram analisadas para segmentar os dados de acordo com diferentes critérios. A unidade geográfica classifica os dados por estado e região, permitindo a comparação das taxas de abandono entre diferentes partes do país. A dependência administrativa foi classificada em Pública e Privada, possibilitando uma análise detalhada do impacto da gestão escolar na evasão dos alunos. O nível de ensino diferencia os dados entre Ensino Fundamental e Ensino Médio. Essa distinção é fundamental, pois o abandono escolar apresenta padrões distintos em cada nível, sendo mais prevalente no Ensino Médio devido a fatores como necessidade de ingresso precoce no mercado de trabalho e dificuldades de adaptação ao sistema educacional.
- **Variáveis Contínuas:** As variáveis contínuas deste estudo representam dados numéricos que podem assumir qualquer valor dentro de um intervalo. A taxa de abandono escolar (%) é a variável de interesse principal, indicando o percentual de alunos que deixaram de frequentar a escola em determinado ano. A distribuição dessa variável foi analisada ao longo do período estudado para identificar tendências temporais e possíveis variações regionais. Outra variável contínua de grande importância é a renda média (em Reais Brasileiros - R$), que reflete o nível socioeconômico da população em cada unidade geográfica. A análise da distribuição da renda

média permitiu avaliar sua relação com a taxa de abandono escolar, verificando se há um padrão em que regiões de menor renda apresentam índices mais elevados de evasão.

- **Análise Bi-Variada:** A análise bi-variada foi conduzida para examinar as relações entre as variáveis do estudo, possibilitando uma compreensão mais profunda dos fatores que influenciam o abandono escolar. A relação entre a taxa de abandono escolar e a renda média foi um dos principais aspectos investigados. A hipótese inicial considerava que regiões com menor renda apresentaram maiores taxas de evasão, um comportamento já observado em estudos educacionais anteriores (Baggi & Lopes, 2011; Araújo et al., 2025; Teodoro & Kappel, 2020).

No entanto, os resultados do teste de correlação de Pearson indicaram uma associação extremamente fraca entre essas variáveis. O coeficiente de correlação foi r = 0,0264, com valor-p p = 0,6968, sugerindo que a relação entre renda média estimada por região e taxa de abandono escolar não é estatisticamente significativa. Portanto, não foi possível confirmar que regiões com menor renda apresentam, necessariamente, maiores índices de evasão, sinalizando a necessidade de considerar outros fatores estruturais e institucionais mais relevantes para compreender o fenômeno.

Além disso, foi realizada a comparação da taxa de abandono escolar entre diferentes tipos de escola, com o objetivo de quantificar a discrepância entre redes públicas e privadas e identificar possíveis fatores determinantes para essa desigualdade. As taxas de abandono foram as seguintes: i) PuEM com 6,82%, ii) PrEM com 1,30%, iii) PrEF com 0,37% e iv) PuEF com 1,71%.

Os resultados confirmam que as escolas públicas, especialmente no nível do Ensino Médio, concentram as maiores médias de abandono escolar, enquanto as escolas privadas de Ensino Fundamental apresentam os menores índices. Esse padrão é consistente com estudos prévios (Sousa et al., 2018; Santos, 2020; Camargos & Silveira, 2024), que apontam que escolas públicas enfrentam desafios mais severos relacionados à infraestrutura, suporte pedagógico e permanência estudantil, contribuindo para a evasão.

- **Análise Estatística Descritiva:** Para compreender melhor os dados e preparar as análises preditivas, foi realizada uma análise estatística descritiva. Essa análise buscou descrever padrões, tendências, e possíveis relações entre a taxa de abandono escolar e variáveis socioeconômicas, como a renda média. Além das medidas de tendência central, foram gerados gráficos descritivos para cada grupo de escola (PuEF, PuEM, PrEF e PrEM), a fim de analisar a distribuição dos dados e identificar possíveis *outliers*. Os gráficos incluem: (i) dispersão com linha da média, (ii) boxplot com média e mediana, (iii) histograma e (iv) gráfico com limites interquartis. Essas representações visuais auxiliaram na compreensão da variabilidade interna dos grupos e reforçam as diferenças estruturais entre eles.

Observou-se maior variabilidade nos dados referentes às escolas públicas, principalmente no Ensino Médio (PuEM), indicando contextos de maior instabilidade na permanência escolar. Também foi realizada uma análise de correlação entre a taxa de abandono escolar e a renda média regional utilizando o coeficiente de Pearson. O resultado demonstrou uma correlação negativa muito fraca (r = -0,0264, p = 0,6968), sem significância estatística, sugerindo que a renda média, isoladamente, não é um fator explicativo forte para a evasão escolar. As técnicas utilizadas incluíram a função describe() do pandas, análise gráfica com Seaborn e Matplotlib, além de agrupamentos por grupo e ano para média móvel. Essa etapa foi essencial para identificar padrões e estruturar as próximas fases da análise.

O Quadro 1 apresenta as medidas de tendência central da taxa de abandono escolar no Brasil, no período de 2013 a 2023, segmentadas por dependência administrativa (pública ou privada) e nível de ensino (fundamental ou médio). Verifica-se que o grupo PuEM (Ensino Médio público) apresentou a maior média da taxa de abandono escolar, com elevado desvio padrão,

indicando maior dispersão dos dados e evidenciando a fragilidade desse segmento educacional. Em contrapartida, o grupo PuEF (Ensino Fundamental público) apresentou valores médios inferiores, mas com curtose acentuada, sugerindo a ocorrência de casos extremos de evasão. Já os grupos correspondentes ao ensino privado, PrEF (Ensino Fundamental privado) e PrEM (Ensino Médio privado), revelaram taxas de abandono significativamente menores, sendo a menor média observada no grupo PrEF, o que indica maior estabilidade na permanência dos estudantes.

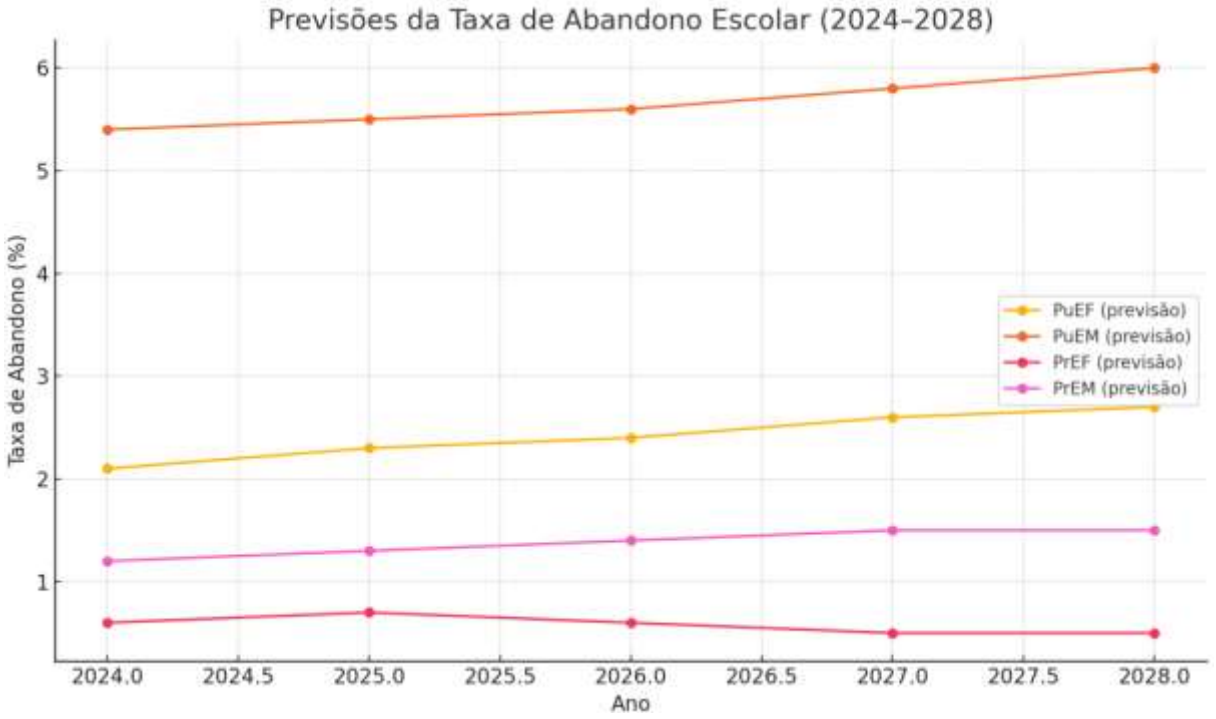
Quadro 1. Estatísticas Descritivas e Testes Adicionais

Grupo	ME	MD	DP	AP	IQR	T	SW	H
PuEF	1,71	1,20	1,15	4	1,75	p < 0.001	p < 0.001	-
PuEM	6,82	6,50	3,19	12,83	4,64	p < 0.001	p < 0.001	p < 0.001
PrEF	0,37	0,23	0,61	4,07	0,34	p = 0.016	p < 0.001	p = 0.001
PrEM	1,30	0,90	0,84	3,67	1,37	p = 0.048	p < 0.001	p < 0.001
ME: Média; MD: Mediana; DP: Desvio Padrão; AP: Amplitute; IQR: Intervalo Inter Quartil; T: (média vs. mediana); SW: Shapiro-Wilk; H: Homocedasticidade c/ PuEF								

Fonte: elaborado pelos autores (2025).

A Figura 1 apresenta as previsões da taxa de abandono escolar por grupo de escola, segmentadas por tipo de dependência administrativa (pública ou privada) e nível de ensino (fundamental ou médio) para o período de 2024 a 2028. Observa-se que os grupos PuEM (Ensino Médio Público) e PuEF (Ensino Fundamental Público) devem permanecer com os maiores índices de abandono, mantendo tendência de crescimento gradual, especialmente no Ensino Médio público, que ultrapassa os 6% ao final do período projetado. Já os grupos PrEF e PrEM (rede privada) mantêm taxas mais baixas e relativamente estáveis, evidenciando um contraste estrutural entre os dois sistemas de ensino.

Figura 1. Previsões da Taxa de Abandono Escolar (2024–2028)



Fonte: elaborado pelos autores (2025).

As estatísticas descritivas calculadas para a Taxa de Abandono Escolar no período de 2013 a 2023 são apresentadas no Quadro 2 e ilustradas na Figura 2 a seguir. Essas medidas são fundamentais para entender o comportamento geral dos dados, especialmente no que diz respeito à concentração e à variabilidade das taxas de abandono entre diferentes grupos de escolas (públicas e privadas, nos níveis fundamental e médio).

De acordo com o Quadro 2, observa-se que a média geral da taxa de abandono escolar no período foi de 2,56%, enquanto a mediana foi de 1,20%, indicando que metade das escolas apresentaram taxas de abandono inferiores a esse valor. A moda, de 0,8%, reforça que o abandono escolar tende a se concentrar em valores baixos, sendo comum que muitas escolas registrem taxas próximas de zero.

Entretanto, a variância (9,49) e o desvio padrão (3,08) sugerem uma considerável dispersão nos dados, indicando que, embora muitas escolas tenham baixos índices de abandono, existem casos extremos (*outliers*) com taxas muito elevadas, que influenciam a média geral. Essa heterogeneidade reforça a necessidade de análises segmentadas para compreender melhor os contextos específicos de cada grupo de escola.

A Figura 2 complementa essas informações ao comparar as medidas de tendência central (média, mediana) e dispersão (desvio padrão) entre os quatro grupos analisados: PuEF (Pública Ensino Fundamental), PuEM (Pública Ensino Médio), PrEF (Privada Ensino Fundamental) e PrEM (Privada Ensino Médio). Visualmente, nota-se que o grupo PuEM apresenta os maiores valores de taxa de abandono, com média de 6,83%, mediana de 6,50% e desvio padrão de 3,22, evidenciando tanto taxas elevadas quanto uma grande variabilidade entre as escolas desse segmento. O grupo PuEF registra média de 1,71%, com mediana de 1,20% e desvio padrão de 1,17, indicando valores mais baixos e menos dispersos em comparação ao ensino médio público.

Nos grupos de escolas privadas, tanto no ensino fundamental (PrEF) quanto no médio (PrEM), as taxas de abandono são substancialmente menores. O grupo PrEF apresenta média de 0,37%, mediana de 0,23% e desvio padrão de 0,62, enquanto o grupo PrEM mostra valores de

1,31% (média), 0,90% (mediana) e 0,87 (desvio padrão), caracterizando maior homogeneidade e controle da evasão escolar na rede privada.

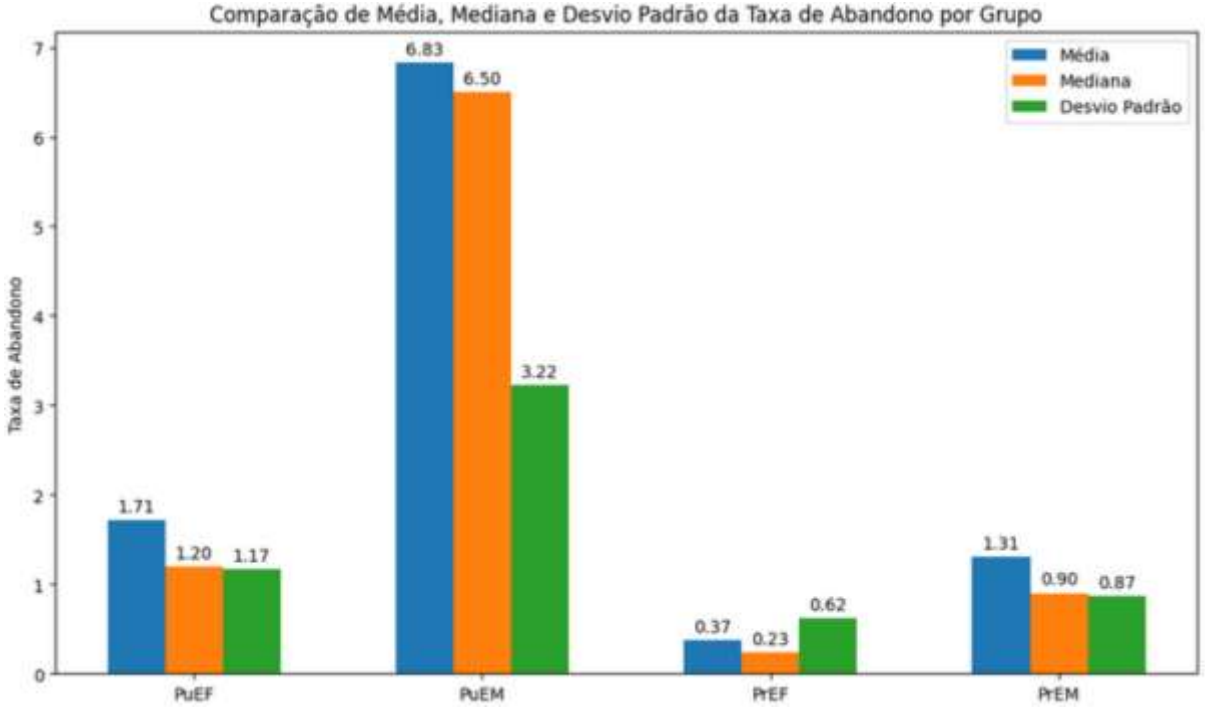
Esses achados reforçam a urgência de políticas públicas voltadas especificamente ao Ensino Médio da rede pública, com o objetivo de reduzir as altas taxas de evasão e promover a permanência escolar, em consonância com as observações de Araujo et al. (2025).

Quadro 2. Estatísticas Descritivas da Taxa de Abandono Escolar

Medida EStatística	**Valor Calculado**	**Descrição**
Média	2,56	Valor médio da taxa de abandono escolar.
Mediana	1,20	Valor central que separa metade superior e inferior dos dados.
Moda	0,80	Valor que mais se repete; muitas escolas têm um abandono próximo de 0%.
Variância	9,49	Mede a dispersão dos dados em torno da média.
Desvio Padrão	3,08	Mede o grau de dispersão; dados estão bem espalhados.

Fonte: elaborado pelos autores (2025).

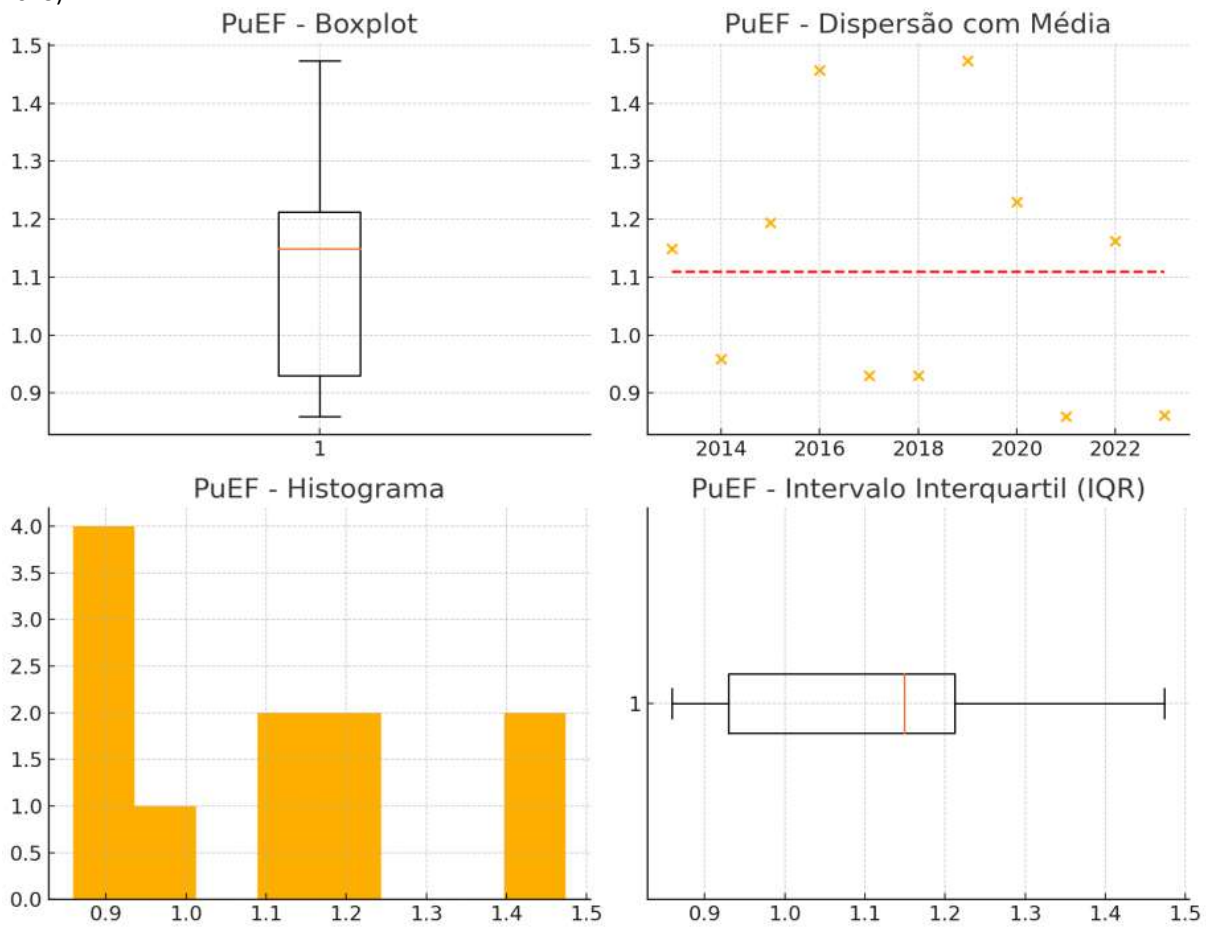
Figura 2. Medidas de tendência central da taxa de abandono escolar por grupo de escola



Fonte: elaborado pelos autores (2025).

A Figura 3 apresenta um conjunto de visualizações referentes ao indicador PuEF (Públicas Ensino Fundamental) no período de 2013 a 2023. O boxplot evidencia a presença de *outliers* e a mediana dos dados, enquanto o gráfico de dispersão temporal demonstra as variações anuais da taxa, com a linha pontilhada indicando a média do período. O histograma mostra a frequência das diferentes faixas de taxa de abandono, sugerindo uma concentração em torno de valores específicos. O gráfico do intervalo interquartílico (IQR) reforça a análise da dispersão e da assimetria dos dados.

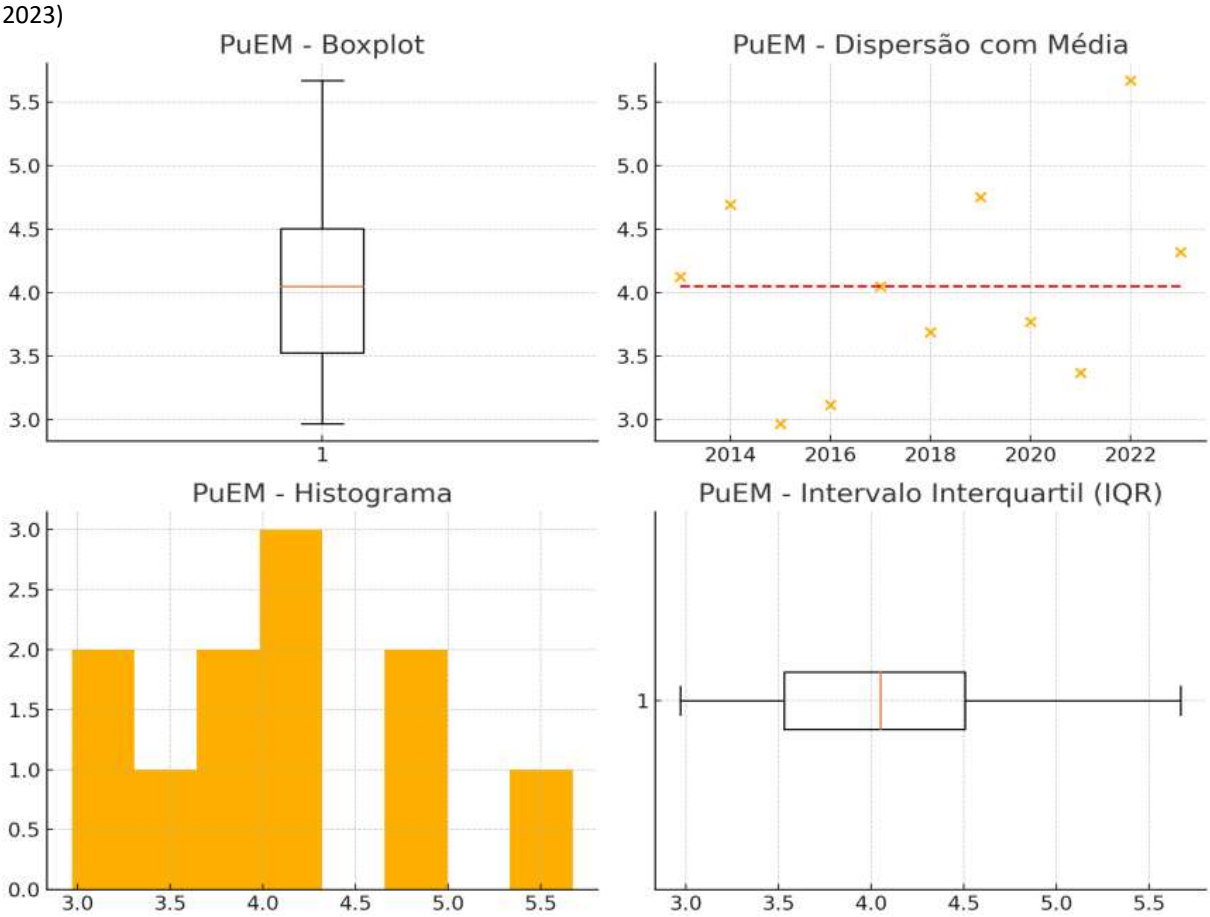
Figura 3. Boxplot, Dispersão Temporal, Histograma e Intervalo Interquartil (IQR) para o indicador PuEF (2013–2023)



Fonte: Elaborado pelos autores (2025).

Da mesma forma, a Figura 4 realiza a análise para o indicador PuEM (Públicas Ensino Médio). O comportamento visualizado no boxplot revela a maior variabilidade dos dados no ensino médio público, enquanto o gráfico de dispersão temporal ilustra oscilações mais intensas em certos anos. O histograma evidencia a distribuição assimétrica, e o IQR reforça a presença de maior dispersão em relação ao ensino fundamental.

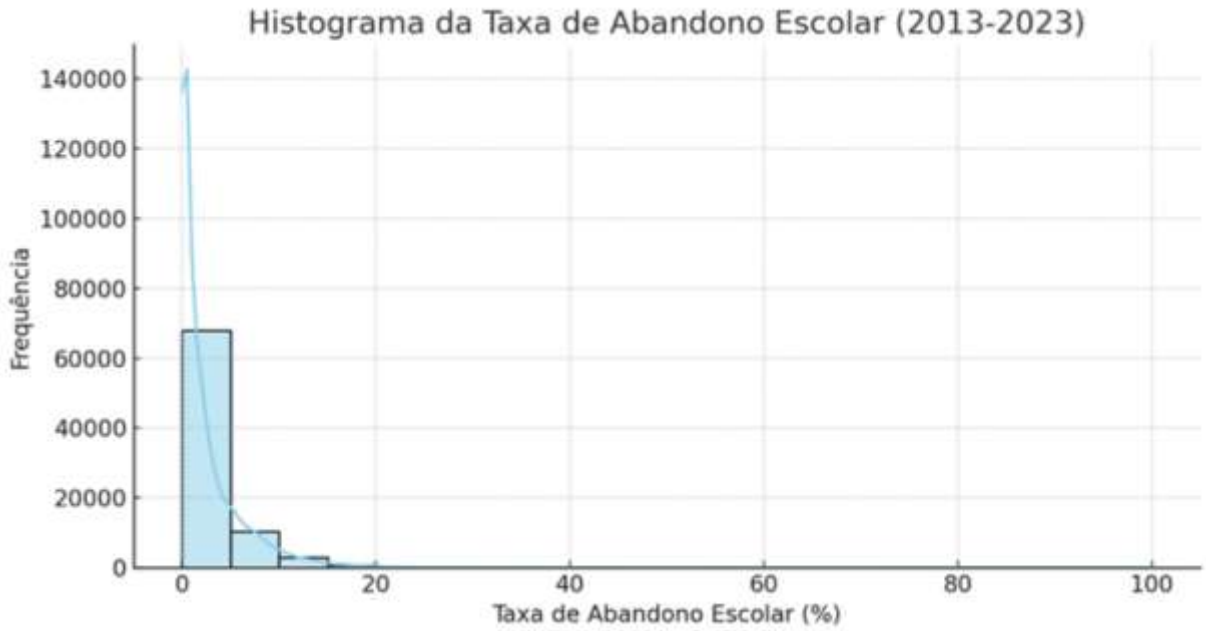
Figura 4. Boxplot, Dispersão Temporal, Histograma e Intervalo Interquartil (IQR) para o indicador PuEM (2013–2023)



Fonte: Elaborado pelos autores (2025).

A Figura 5 exibe o histograma geral da Taxa de Abandono Escolar para todo o período de análise (2013–2023). Pode-se observar que a maior parte dos dados se concentra nas menores taxas de abandono (entre 0% e 5%), apresentando uma distribuição fortemente assimétrica à direita, como indicado também pela curva de densidade sobreposta.

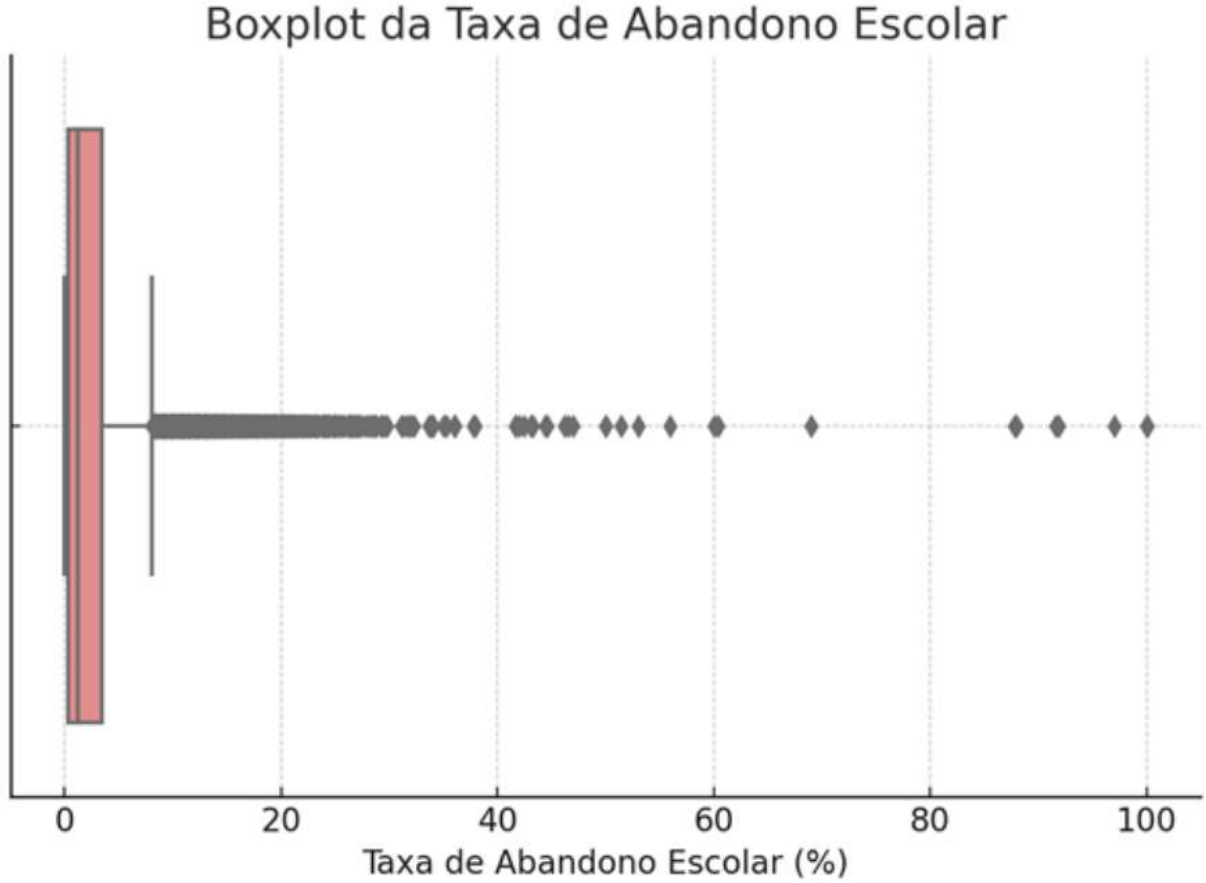
Figura 5. Distribuição da Taxa de Abandono Escolar (2013–2023) - Histograma com Curva de Densidade



Fonte: Elaborado pelos autores (2025).

A Figura 6 complementa essa análise por meio de um boxplot geral da taxa de abandono escolar. A visualização evidencia a existência de diversos *outliers*, com valores extremos próximos de 100%, indicando casos isolados de abandono escolar muito elevados. A maioria dos dados, entretanto, está concentrada abaixo dos 10%.

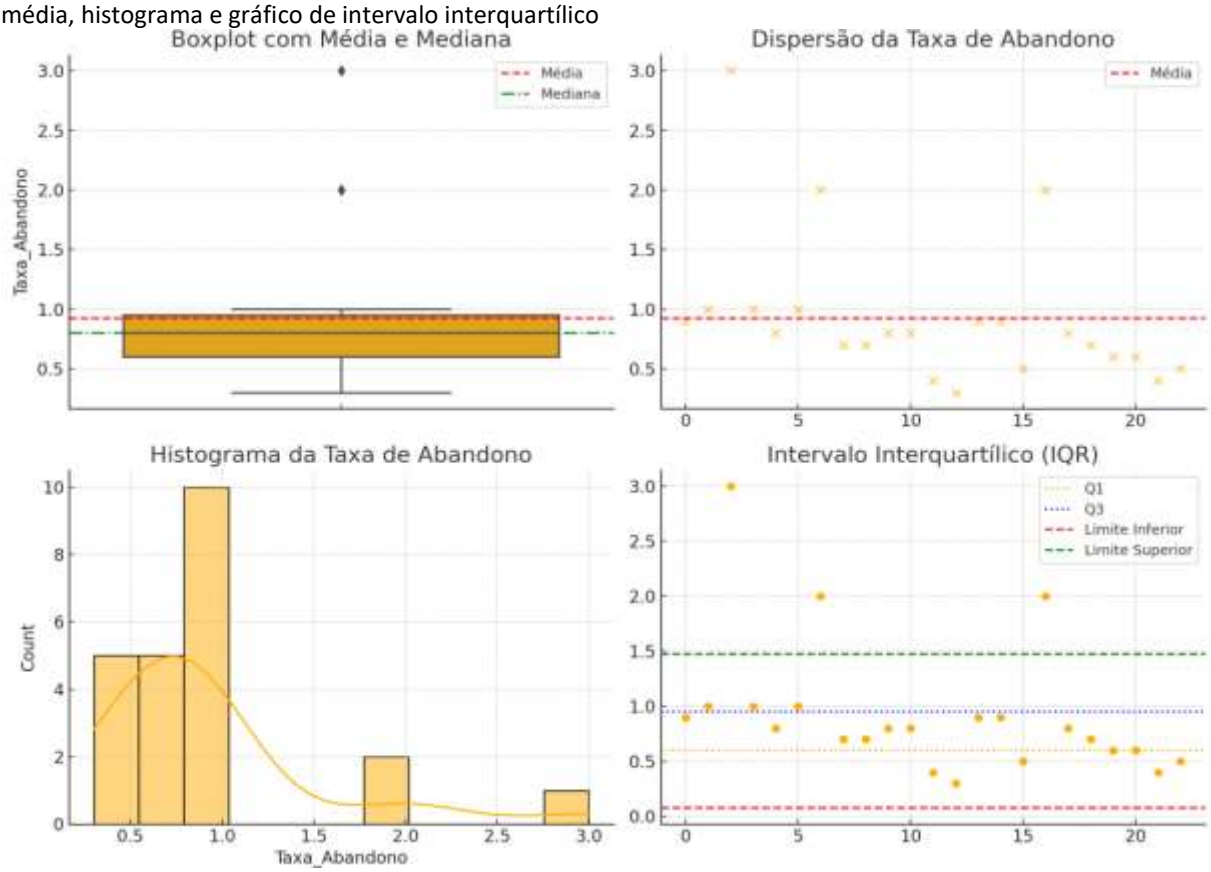
Figura 6. Análise de *Outliers* da Taxa de Abandono Escolar (2013–2023) por meio de Boxplot



Fonte: Elaborado pelos autores (2025).

Para uma análise segmentada por grupos, foram construídas visualizações específicas: a Figura 7 apresenta a análise descritiva da taxa de abandono para o grupo PuEF (escolas públicas de ensino fundamental), utilizando boxplot, dispersão com média, histograma e intervalo interquartílico. A análise mostra a concentração dos dados em torno da média e a existência de poucos *outliers*.

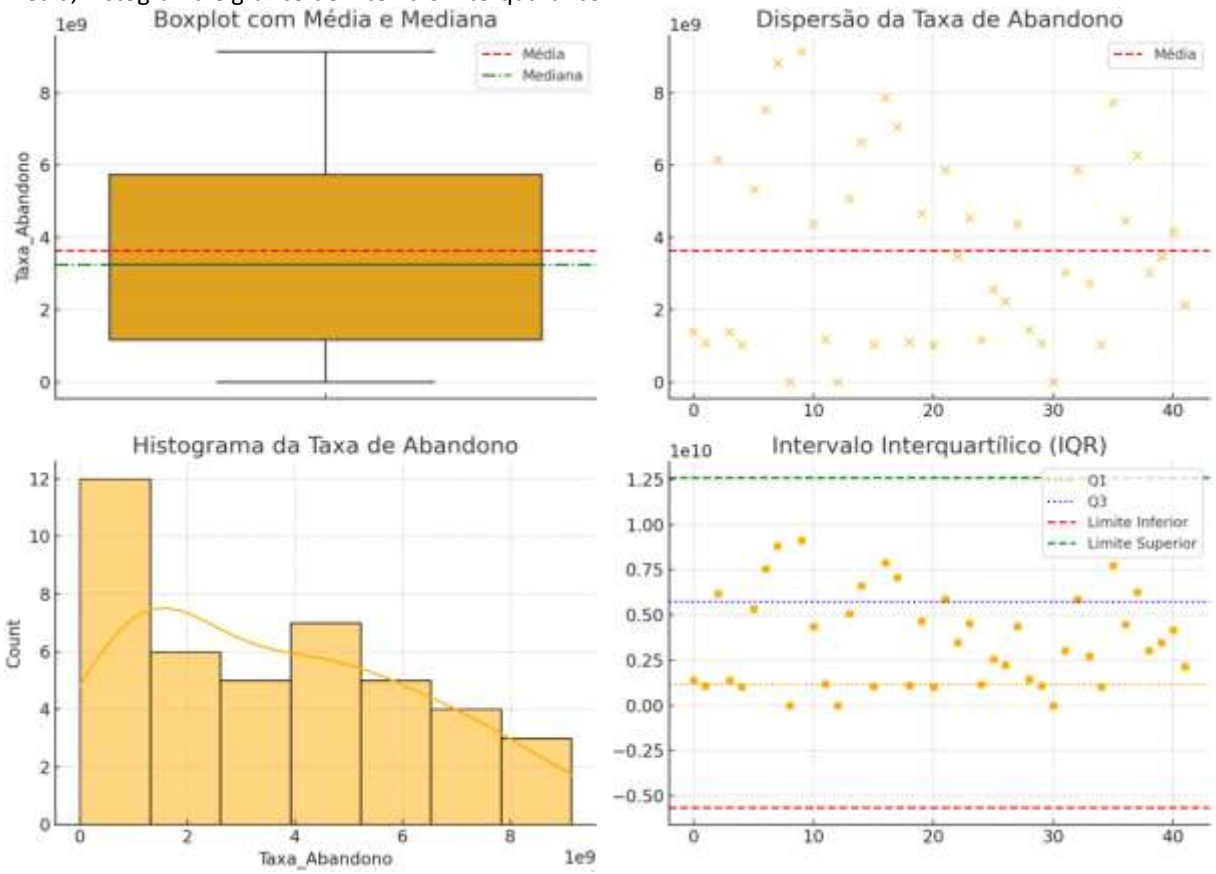
Figura 7. Análise descritiva da taxa de abandono escolar para o grupo PuEF, incluindo boxplot, dispersão com média, histograma e gráfico de intervalo interquartílico



Fonte: Elaborado pelos autores (2025).

A Figura 8 realiza a mesma análise para o grupo PuEM (escolas públicas de ensino médio). Nota-se maior variabilidade e a presença de *outliers* mais pronunciados em relação ao ensino fundamental.

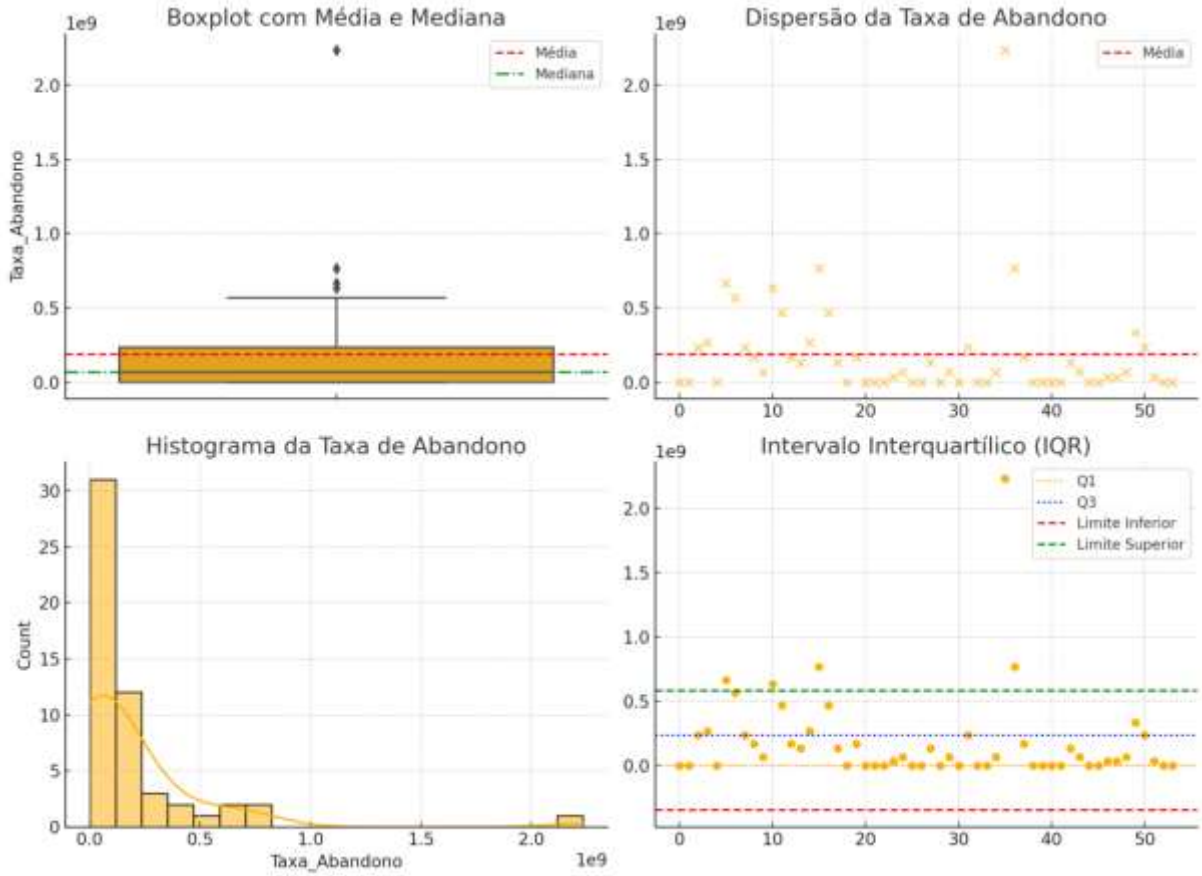
Figura 8. Análise descritiva da taxa de abandono escolar para o grupo PuEM, incluindo boxplot, dispersão com média, histograma e gráfico de intervalo interquartílico



Fonte: Elaborado pelos autores (2025).

A Figura 9 mostra o comportamento para o grupo PrEF (escolas privadas de ensino fundamental). Aqui, as taxas de abandono são, em geral, menores e com menor variabilidade, refletindo a tendência de estabilidade no setor privado.

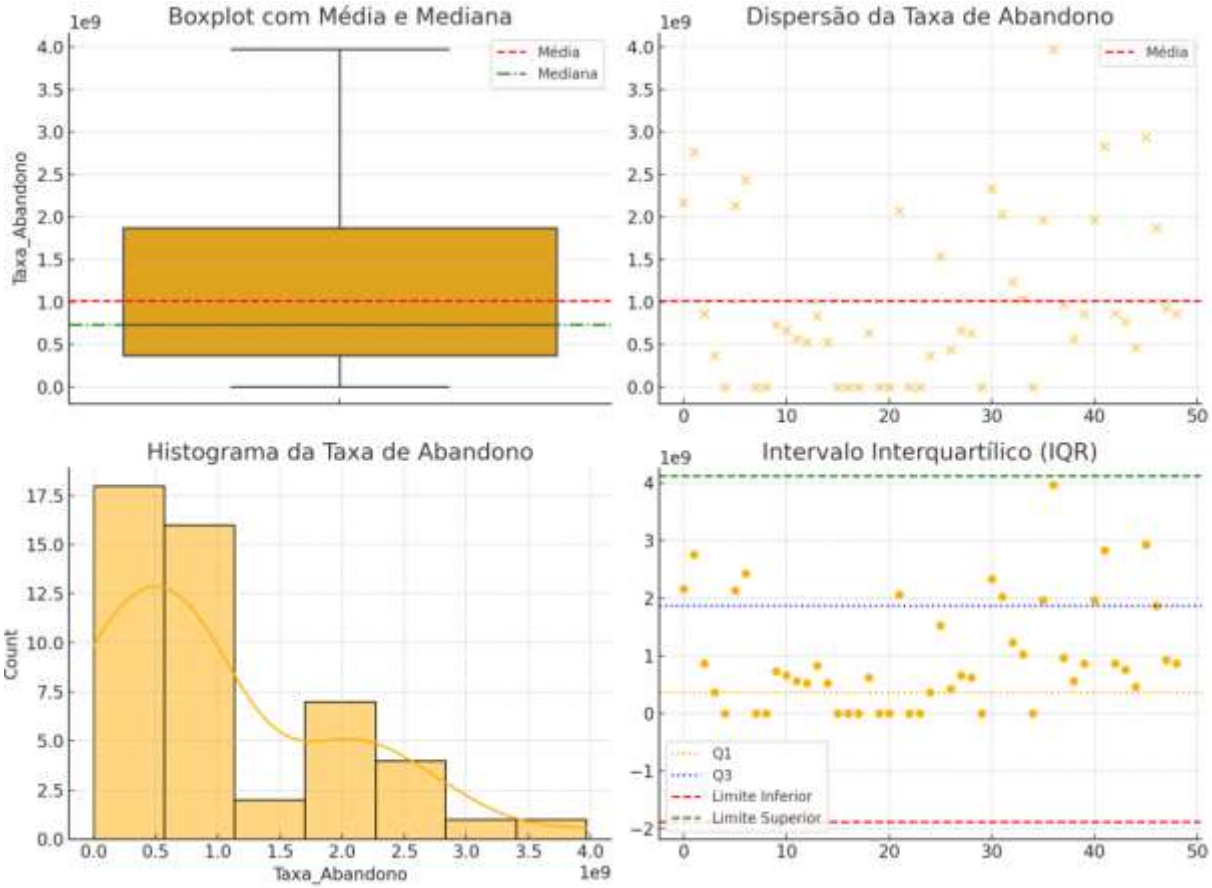
Figura 9. Análise descritiva da taxa de abandono escolar para o grupo PrEF, incluindo boxplot, dispersão com média, histograma e gráfico de intervalo interquartílico



Fonte: Elaborado pelos autores (2025).

Por fim, a Figura 10 apresenta os dados do grupo PrEM (escolas privadas de ensino médio). A distribuição é semelhante à do ensino fundamental privado, com taxas ainda menores e dispersão reduzida, indicando baixos índices de evasão nesse segmento.

Figura 10. Análise descritiva da taxa de abandono escolar para o grupo PrEM, incluindo boxplot, dispersão com média, histograma e gráfico de intervalo interquartílico



Fonte: Elaborado pelos autores (2025).

Adicionalmente, foram realizadas análises de correlação de Pearson entre a taxa de abandono escolar nas escolas públicas de ensino fundamental (PuEF) e as escolas privadas de ensino fundamental (PrEF), bem como entre PuEF e as escolas privadas de ensino médio (PrEM). Os resultados indicaram correlação negativa fraca entre PuEF e PrEF (r = -0,17), e entre PuEF e PrEM (r = -0,14). Esses valores sugerem que, embora exista uma tendência de que os grupos variem em direções opostas, a força dessa relação linear é baixa. Assim, as taxas de abandono em escolas públicas de ensino fundamental não apresentam uma associação linear forte com aquelas observadas em escolas privadas.

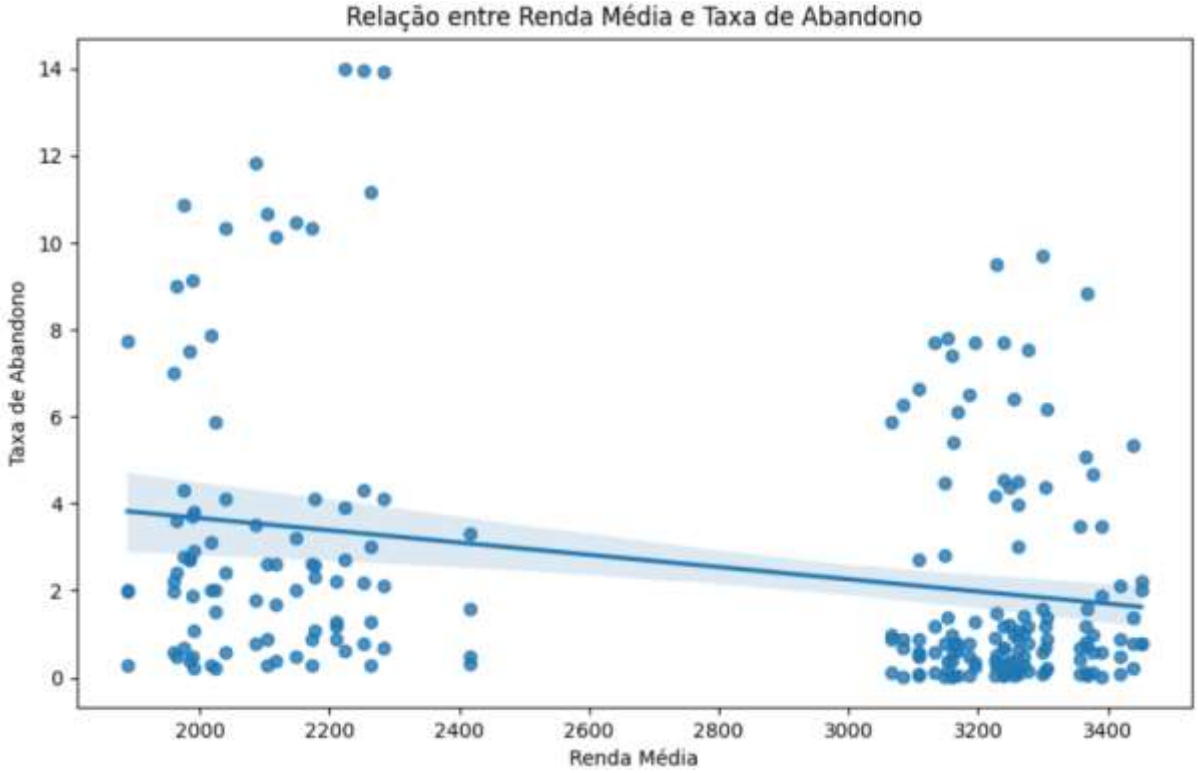
A análise da relação entre a renda média familiar e a taxa de abandono escolar é apresentada na Figura 11 a seguir. Essa representação gráfica é essencial para explorar o comportamento conjunto dessas duas variáveis ao longo do período de 2013 a 2023, permitindo visualizar padrões, tendências e possíveis correlações.

De acordo com a Figura 11, observa-se uma leve tendência negativa entre renda média e taxa de abandono escolar. A linha de tendência no gráfico indica que, em geral, regiões com maior renda tendem a registrar menores taxas de abandono. Contudo, essa correlação é fraca, como evidenciado pela ampla dispersão dos pontos ao redor da linha. Isso sugere que, embora exista uma influência da renda sobre a evasão escolar, ela não é suficiente por si só para explicar o fenômeno de forma conclusiva.

Além disso, nota-se uma concentração de pontos com baixas taxas de abandono nas faixas de renda mais elevadas, o que reforça a ideia de que condições socioeconômicas mais favoráveis podem estar associadas à permanência escolar. Por outro lado, há também pontos com alta taxa de evasão em regiões com renda intermediária ou até elevada, indicando que outros

fatores — como qualidade da gestão escolar, políticas públicas locais, acesso a transporte, segurança, entre outros — também exercem papel significativo nesse contexto.
Portanto, embora a renda média se mostre um indicador relevante, ela deve ser interpretada em conjunto com um conjunto mais amplo de variáveis contextuais para que se possa compreender e enfrentar adequadamente o abandono escolar.

Figura 11. Dispersão entre Renda Média Familiar e Taxa de Abandono Escolar



Fonte: Elaborado pelos autores (2025).

- **Análise Estatística Inferencial:** As análises estatísticas complementares permitiram uma compreensão mais aprofundada sobre a distribuição dos dados em cada grupo. Observou-se que, em todos os grupos, a média e a mediana apresentaram valores próximos, embora diferenças estatisticamente significativas tenham sido identificadas em alguns casos pelo teste T. A amplitude, combinada com o desvio padrão elevado, evidenciou maior variabilidade especialmente no grupo PuEM, reforçando a presença de *outliers* já identificados nos boxplots (Figuras 3, 4, 6, 7, 8, 9 e 10). O intervalo interquartílico (IQR) foi comparado com a média e mostrou que a maior dispersão interna também ocorre nos grupos públicos, o que sugere desigualdades estruturais no perfil das escolas analisadas.
O teste de normalidade (Shapiro-Wilk) indicou que, em certos casos, os dados se desviam da distribuição normal. Além disso, o teste de homocedasticidade (Levene) mostrou que há diferenças nas variâncias entre os grupos, especialmente quando comparado PuEF com os grupos privados. Por fim, a aplicação do teste ANOVA de Welch confirmou diferenças estatisticamente significativas entre as taxas de abandono nos grupos PuEF x PrEF ($p = 0,0132$) e PuEF x PrEM ($p < 0,0001$), evidenciando que as escolas públicas de ensino fundamental apresentam dinâmicas distintas em relação às privadas.

Quadro 3. Testes Estatísticos e p-valores

Grupo	Teste T	Shapiro Wilk	Levence
PuEF	p < 0.001	p < 0.001	-
PuEM	p < 0.001	p < 0.001	p < 0.001
PrEF	p < 0.016	p < 0.001	p = 0.001
PrEM	p < 0.048	p < 0.001	p < 0.001

Fonte: elaborado pelos autores (2025).

- **Análise Estatística Preditiva:** Esta etapa do estudo teve como objetivo prever a taxa de abandono escolar com base em variáveis socioeconômicas e educacionais, utilizando modelos de aprendizado de máquina. Foram testados sete algoritmos: Random Forest, XGBoost, Gradient Boosting, Regressão Linear, Redes Neurais Artificiais (ANN/MLP), Support Vector Machines (SVM) e K-Nearest Neighbors (KNN).

Todos os modelos foram treinados sob as mesmas condições: a base de dados foi dividida em 80% para treinamento e 20% para teste, com ajuste inicial de hiperparâmetros seguido de validação cruzada. As métricas utilizadas para avaliação foram o Erro Médio Absoluto (MAE) e o Coeficiente de Determinação (R^2).

Entre os algoritmos testados, o Random Forest destacou-se como o mais eficaz, alcançando MAE de 0,69 e R^2 de 0,86, indicando forte capacidade explicativa da variabilidade nos dados com erro relativamente baixo. O Gradient Boosting apresentou desempenho próximo, com MAE de 0,77 e R^2 de 0,84, seguido pelo XGBoost, com MAE de 0,91 e R^2 de 0,72.

Por outro lado, a Rede Neural Artificial (ANN/MLP) obteve o pior desempenho, com erro extremamente elevado (MAE de 44,78) e R^2 negativo (-181,03), sugerindo inadequação para este conjunto de dados. Modelos como SVM e KNN também demonstraram baixo desempenho preditivo.

Os resultados reforçam que variáveis como renda média e tipo de escola têm impacto significativo sobre a evasão escolar, evidenciando a utilidade dos modelos preditivos
não apenas para antecipar comportamentos de abandono, mas também para subsidiar intervenções educacionais direcionadas, vulnerabilidade socioeconômica.

Quadro 4. Resultados das avaliações dos modelos

Avaliação / **Modelo**	**Random Forest**	**XGBoost**	**Gradient Boosting**	**Linear Regression**	**ANN**	**SVN**	**KNN**
MAE	0.69	0.91	0.77	1.43	44.78	1.91	2.41
R^2	0.86	0.72	0.84	0.68	-181.03	-0.03	0.07

Fonte: elaborado pelos autores (2025).

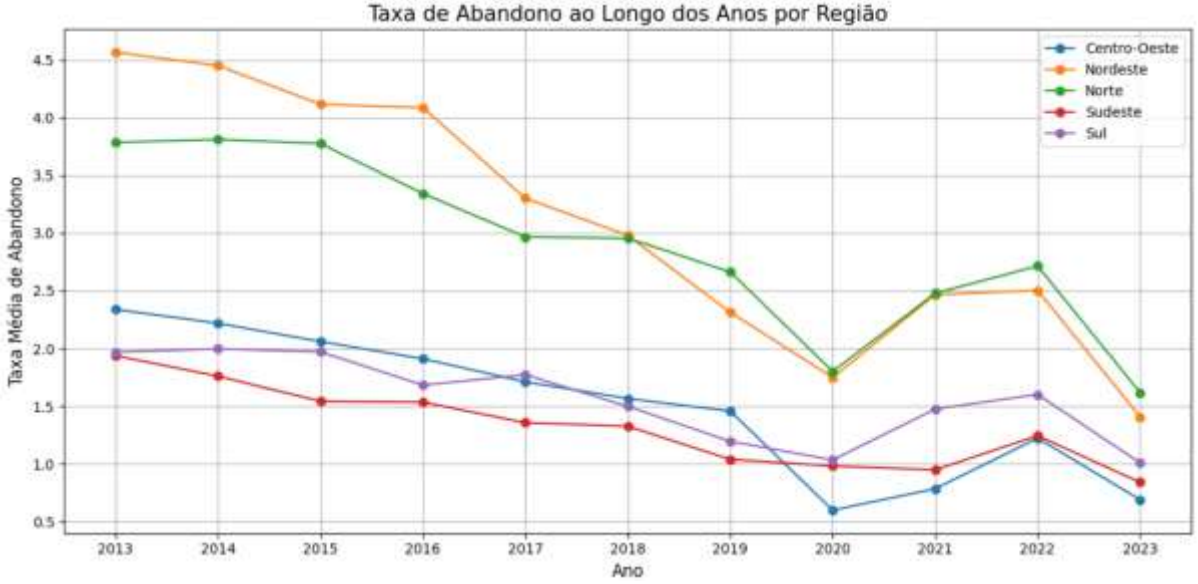
- **Análise de Séries Temporais:** Além da aplicação de modelos de aprendizado de máquina para a predição da evasão escolar, foi realizada uma análise da evolução histórica da taxa de abandono ao longo dos anos. Para isso, os dados foram modelados utilizando os algoritmos preditivos selecionados, e a função predict() foi empregada para gerar estimativas futuras.

A análise revelou que a taxa de evasão apresentou uma tendência de queda nos primeiros anos do período estudado, seguida de oscilações ao longo do tempo. As previsões indicam um

leve aumento na taxa de abandono escolar nos anos mais recentes, evidenciando variações que podem estar associadas a diferentes fatores socioeconômicos e estruturais.

Esses resultados demonstram a relevância da modelagem preditiva na identificação de padrões históricos e projeção de possíveis cenários futuros. A análise das tendências permite compreender melhor a dinâmica da evasão escolar e avaliar a influência de diferentes variáveis ao longo do tempo.

Figura 12. Evolução da Taxa Média de Abandono Escolar por Região Administrativa do Brasil (2013-2023)

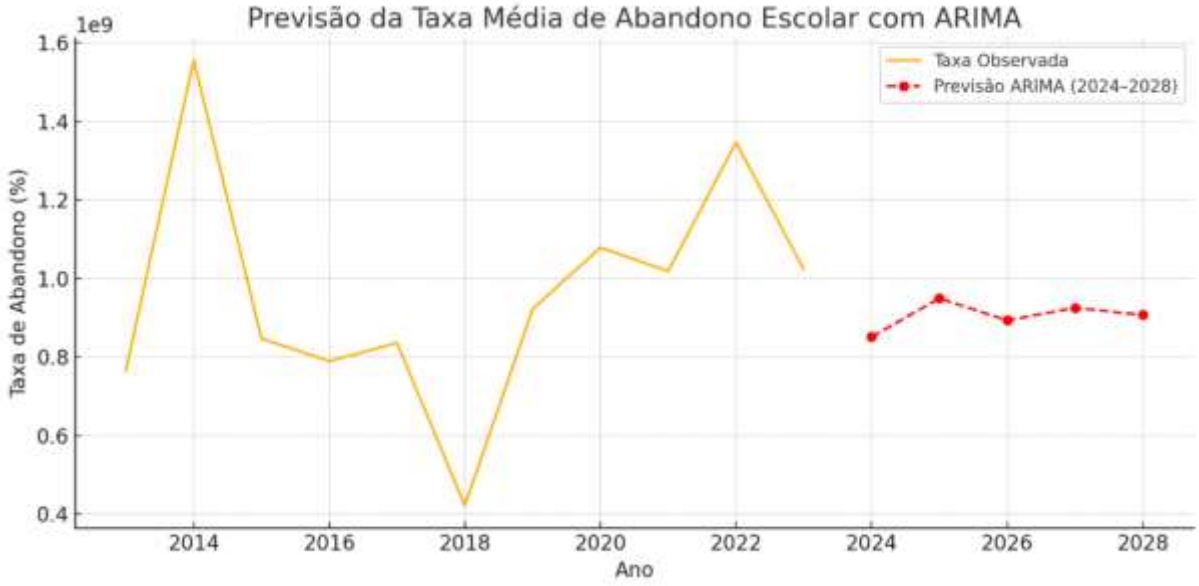


Fonte: elaborado pelos autores (2025).

A série histórica da taxa média de abandono escolar nacional apresentou uma tendência geral de queda entre 2013 e 2018, conforme o modelo ARIMA identificou. No entanto, oscilações mais marcantes ocorreram entre 2019 e 2023, com aumento em determinados anos, o que foi capturado pela variabilidade no gráfico.

As previsões para 2024 a 2028 sugerem uma tendência de leve crescimento, com o modelo estimando aumento gradual da taxa de abandono. Essa projeção reforça a necessidade de intervenções educacionais mais efetivas, principalmente em contextos de vulnerabilidade, para evitar o agravamento do problema nos próximos anos.

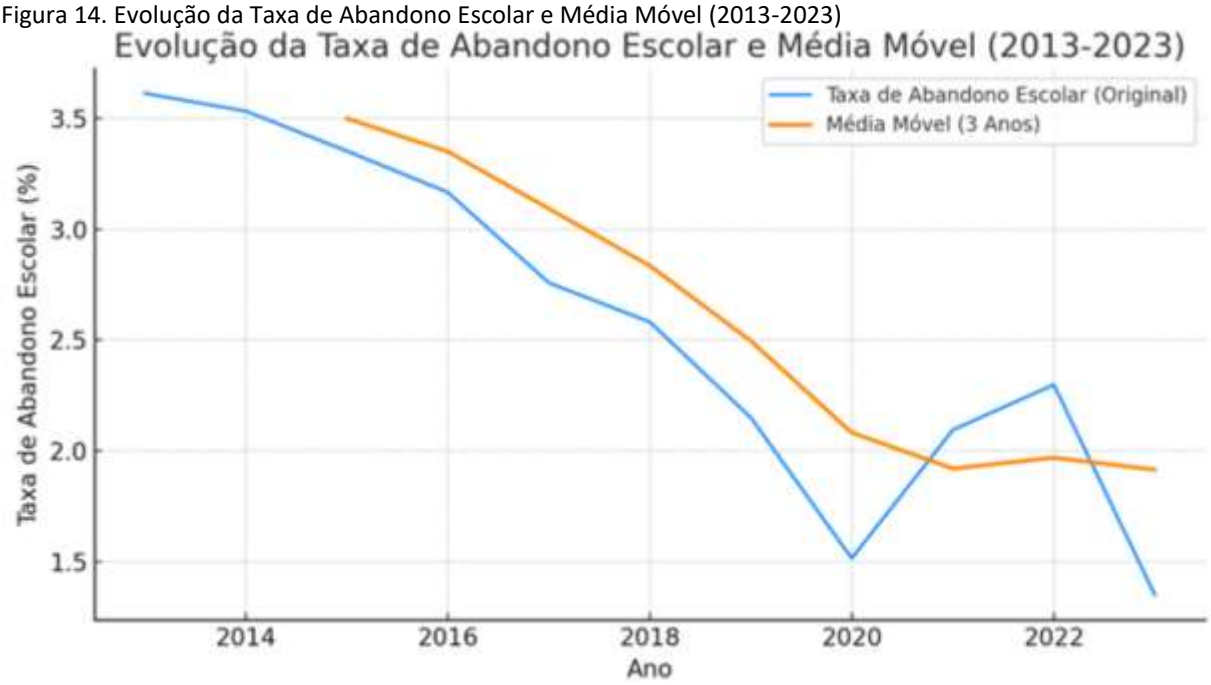
Figura 13. Previsão da Taxa Média de Abandono Escolar no Brasil (2024-2028) com o Modelo ARIMA



Fonte: elaborado pelos autores (2025).

Foi aplicada uma média móvel de 3 anos para suavizar as flutuações anuais e identificar tendências de longo prazo. O comportamento mostrado na Figura 14 evidencia uma queda gradual na taxa de abandono escolar até 2020, seguida por oscilações nos anos mais recentes, possivelmente influenciadas por fatores externos, como a pandemia de COVID-19. Adicionalmente, foi realizada uma análise de correlação de Pearson entre a série original e a média móvel de 3 anos, com o objetivo de verificar o grau de associação linear entre ambas. O resultado revelou uma correlação positiva forte (r ≈ 0,94), indicando que a média móvel acompanha de forma consistente as variações da série original, ainda que suavize as oscilações mais abruptas. Isso reforça a adequação do método para evidenciar tendências de longo prazo na taxa de abandono escolar.

Figura 14. Evolução da Taxa de Abandono Escolar e Média Móvel (2013-2023)



Fonte: elaborado pelos autores (2025).

Considerações Finais

Os resultados obtidos neste estudo reforçam a relevância da análise preditiva na compreensão dos fatores que influenciam a evasão escolar no ensino fundamental e médio. A aplicação de modelos de aprendizado de máquina permitiu identificar que variáveis como a renda média e o tipo de escola (pública ou privada) têm um impacto significativo na taxa de abandono escolar. Essas variáveis se destacam como fatores-chave na previsão de evasão, o que sugere que políticas públicas voltadas para essas questões poderiam contribuir de forma significativa para a redução da evasão.

O modelo Random Forest se destacou no desempenho, apresentando a melhor capacidade de captura de padrões complexos relacionados à evasão escolar. Esse modelo demonstrou uma performance superior ao lidar com interações não lineares e grandes volumes de dados, evidenciando sua eficiência para o problema em questão. A habilidade do Random Forest em identificar essas relações torna-o uma ferramenta importante para prever o comportamento dos alunos em risco de evasão.

Além disso, a análise das tendências históricas de evasão escolar, utilizando a função predict(), revelou que, embora a taxa de abandono tenha mostrado uma tendência de queda em determinados períodos, há oscilações ao longo do tempo. Essas flutuações podem estar associadas a diversos fatores socioeconômicos e estruturais, como mudanças nas políticas educacionais e crises econômicas. Essas oscilações indicam a necessidade de monitoramento contínuo dos dados, de modo a compreender melhor as causas por trás dessas variações e adaptar as estratégias educacionais conforme o contexto.

Além da etapa preditiva, os resultados das análises estatísticas descritivas e inferenciais forneceram subsídios importantes para a compreensão do fenômeno da evasão escolar. Observou-se que as maiores taxas de abandono concentram-se no ensino médio da rede pública, especialmente nas regiões Norte e Nordeste, onde também se registram os menores valores de renda média familiar. A análise de médias, variâncias e testes de hipóteses evidenciou diferenças estatisticamente significativas entre os grupos escolares e regionais. Testes como ANOVA de Welch, Levene e Shapiro-Wilk reforçaram a heterogeneidade dos dados e justificaram o uso de modelos mais robustos. Já a regressão linear simples apontou tendências decrescentes na taxa de abandono em todas as regiões, ainda que com baixos coeficientes de determinação, indicando que o tempo, isoladamente, explica pouco da variabilidade do abandono. Esses resultados reforçam que múltiplos fatores estruturais e sociais precisam ser considerados em estratégias de combate à evasão.

Por fim, a hipótese inicial considerava que regiões com menor renda apresentavam maiores taxas de evasão, um comportamento amplamente discutido na literatura, que aponta a desigualdade socioeconômica como um dos principais fatores associados ao abandono escolar (Baggi & Lopes, 2011; Araújo et al., 2025; Teodoro & Kappel, 2020; Rumberger, 2011). No entanto, os resultados obtidos neste estudo não corroboram essa relação de forma isolada, uma vez que a análise de correlação de Pearson entre renda média e taxa de abandono escolar revelou associação extremamente fraca e estatisticamente não significativa ($r = -0,0264$; $p = 0,6968$). Em contrapartida, verificou-se que fatores institucionais apresentaram maior poder explicativo, especialmente o tipo de escola e o nível de ensino, com destaque para o ensino médio da rede pública (PuEM), que apresentou média de abandono de 6,82%, substancialmente superior aos demais grupos, como o ensino médio privado (PrEM) com 1,30%, o ensino fundamental público (PuEF) com 1,71% e o ensino fundamental privado (PrEF) com apenas 0,37%. Adicionalmente, entre os modelos preditivos avaliados, o Random Forest demonstrou desempenho superior (MAE = 0,69; $R^2 = 0,86$), evidenciando maior capacidade de capturar relações não lineares e complexas entre as variáveis analisadas. Esses resultados estão em consonância com estudos que apontam a

evasão escolar como um fenômeno multifatorial, no qual fatores institucionais, como qualidade da escola, clima escolar, organização pedagógica e políticas educacionais, possuem papel tão ou mais relevante do que variáveis socioeconômicas isoladas (Rumberger & Lim, 2008; Bowers et al., 2013; Tinto, 1975). Dessa forma, os resultados deste estudo contribuem para a literatura ao indicar que a evasão escolar não pode ser explicada exclusivamente por variáveis socioeconômicas agregadas, reforçando a necessidade de abordagens multidimensionais e destacando o papel central de fatores institucionais e estruturais na compreensão do fenômeno. Ou seja, os resultados indicam que a aplicação de modelos preditivos não só fornece *insights* valiosos sobre os fatores que contribuem para a evasão escolar, mas também pode servir como base para o desenvolvimento de intervenções mais direcionadas e eficazes. A análise preditiva, assim, apresenta o potencial de ajudar a prevenir a evasão ao identificar os alunos em risco e possibilitar ações proativas para sua permanência na escola.

Referências

Araújo, C. L., Santos, Q. P., Ribeiro, H. M. L., Freitas, E. B. N., & Coutinho, D. J. G. (2025). Evasão escolar: causas e impactos da evasão escolar no Brasil e no mundo. *Revista Ibero-Americana de Humanidades, Ciências e Educação*, 11*(1)*, 1945–1965. https://doi.org/10.51891/rease.v11i1.17879.

Barbosa, P. K. (2021). *O impacto da evasão escolar no mercado de trabalho informal brasileiro e as consequências da pandemia* (Dissertação de mestrado). Insper. https://repositorio.insper.edu.br/server/api/core/bitstreams/b4eb3525-7d03-4831-a27e-1a70b231930e/content.

Baggi, C. A. S., & Lopes, D. A. (2011). Evasão e avaliação institucional no ensino superior: Uma discussão bibliográfica. *Avaliação: Revista da Avaliação da Educação Superior*, 16*(2)*, 355–374. https://doi.org/10.1590/s1414-40772011000200007.

Bowers, A. J., Sprott, R., & Taff, S. A. (2013). Do we know who will drop out? *Review of Educational Research*, 83*(2)*, 205–235.

Bruce, P. (2019). *Estatística prática para cientistas de dados*: 50 conceitos essenciais. Alta Books.

Camargos, R. C., & Silveira, I. F. (2024). Técnicas de aprendizado de máquina interpretáveis na predição de evasão escolar: Uma revisão. *RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação*, 65, 636–647. https://media.proquest.com/media/hms/PFT/1/NTucX?_s=QMIRdKAkhMzcJffjFdGCDPWtt2A%3D.

Escovedo, T. (2022). *Introdução à data science*: Algoritmos de machine learning e métodos de análise. Casa do Código.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. (2020). *Política e plano de dados abertos do INEP*. https://download.inep.gov.br/publicacoes/institucionais/gestao_e_governanca/politica_e_plano_de_dados_abertos.pdf.

Jesus, H. O., Rodriguez, L. C., & Costa Junior, A. O. (2021). Predição de evasão escolar na licenciatura em computação. *Revista Brasileira de Informática na Educação*, 29, 255–272. https://doi.org/10.5753/rbie.2021.29.0.255.

Lopes Filho, J. A. B. (2021). *Detecção de estudantes em risco de evasão escolar usando aprendizagem de máquina* (Tese de doutorado). Universidade Presbiteriana Mackenzie. https://adelpha-api.mackenzie.br/server/api/core/bitstreams/178b32ac-cd0a-44e3-9804-87566f733e15/content.

Moore, D. S. (2023). *A estatística básica e sua prática* (9ª ed.). LTC.

Ramos, A. C., & Gonçalves Junior, O. (2024). Abandono e evasão escolar sob a ótica dos sujeitos envolvidos. *Educação e Pesquisa*, 50*(1)*, 41–54. https://doi.org/10.1590/s1678-4634202450268037.

Rosa, M. C., Silva, P. R. S., & Novaes, H. V. B. (2023). Evasão escolar: O impacto. *Libertas*, 13, 1–20. https://www.periodicos.famig.edu.br/index.php/libertas/article/view/377/287.

Rumberger, R. W. (2011). *Dropping out:* Why students drop out of high school and what can be done about it. Harvard University Press.

Rumberger, R. W., & Lim, S. A. (2008). *Why students drop out of school*: A review of 25 years of research. California Dropout Research Project Report.

Santos, J. A. (2020). Reflexões sobre a evasão escolar: Uma problemática na educação brasileira. *Revista Teias*, 2*(1)*, 132–145. https://doi.org/10.12957/teias.2020.41951.

Sousa, C. R. O., Gomes, K. R. O., Silva, K. C. O., Mascarenhas, M. D. M., Rodrigues, M. T. P., Andrade, J. X., & Leal, M. A. B. F. (2018). Fatores preditores da evasão escolar entre adolescentes com experiência de gravidez. *Cadernos Saúde Coletiva*, 26*(2)*, 160–169. https://doi.org/10.1590/1414-462x201800020461.

Teodoro, L. A., & Kappel, M. A. A. (2020). 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*, 28, 838–863. https://doi.org/10.5753/rbie.2020.28.0.838.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis. *Review of Educational Research*, 45(1), 89–125.