Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
https://sol.sbc.org.br/journals/index.php/rbie

Submission: dd/Mmm/yyyy; 1st round notif.: dd/Mmm/yyyy; New version: dd/Mmm/yyyy; 2nd round notif.: dd/Mmm/yyyy;
Camera ready: dd/Mmm/yyyy; Edition review: dd/Mmm/yyyy; Available online: dd/Mmm/yyyy; Published: dd/Mmm/yyyy;

Modelagem Estatística da Evasão no Ensino Médio Brasileiro: Uma Investigação Baseada em Variáveis Institucionais e Socioeconômicas

Title: Statistical Modeling of Dropout Rates in Brazilian High Schools: An Investigation Based on Institutional and Socioeconomic Variables

Título: Modelización estadística de la deserción escolar en la enseñanza secundaria brasileña: una investigación basada en variables institucionales y socioeconómicas

Bruno Alexandre Dias da Silva
Universidade de São Paulo
ORCID: 0000-0000-0000-0000
Brunoalexdias20@usp.br

Lucas Gurgel do Amaral
Universidade de São Paulo
ORCID: 0000-0000-0000-0000
lucasgurgel@usp.br

Rafael de França
Universidade de São Paulo
ORCID: 0000-0000-0000-0000
rafaeldefranca@usp.br

Richard Pereira do Nascimento
Universidade de São Paulo
ORCID: 0000-0000-0000-0000
rcdwoods@usp.br

Resumo

A evasão escolar no ensino brasileiro vem se mostrando como um grande desafio na qualificação e formação educacional do público infanto-juvenil, principalmente nas camadas menos favorecidas da sociedade brasileira. Jovens que não concluem o ensino médio não conseguem especializar-se em cursos superiores e, portanto, submetem-se a empregos com mão de obra barata, perdurando um ciclo de pobreza e baixos indicadores socioeconômicos. Para isso, este trabalho tem como objetivo avaliar e desenvolver modelos probabilísticos de aprendizado de máquina com o intuito de prever a probabilidade de um aluno evadir o ensino médio da rede de ensino de São Paulo baseado em técnicas como Floresta Aleatória, Árvores de Decisão e Redes Neurais. Por meio da base de dados da Pesquisa Nacional por Amostra de Domicílios (PNAD) provenientes do Instituto Brasileiro de Geografia e Estatística (IBGE), serão extraídas variáveis que, de acordo com a literatura de estudos acerca de evasão, mais expliquem o abandono escolar contínuo. Os resultados deste artigo podem auxiliar docentes e escolas a rastrearem e prestarem mais apoio àqueles alunos que apresentam alta chance de evasão por motivos socioeconômicos.
Palavras-chave: *Evasão escolar; Aprendizado de Máquina; PNAD; Variáveis Socioeconômicas; Ensino médio.*

Abstract

School dropout rates in Brazil have proven to be a major challenge in the educational qualification and training of children and young people, especially in the less privileged segments of Brazilian society. Young people who do not complete high school are unable to specialize in higher education courses and, therefore, take on low-wage jobs, perpetuating a cycle of poverty and low socioeconomic indicators. Therefore, this study aims to evaluate and develop probabilistic machine learning models to predict the probability of a student dropping out of high school in the São Paulo school system based on techniques such as Random Forest, Decision Trees, and Neural Networks. Using data from the Brazilian Institute of Geography and Statistics (IBGE) National Household Sample Survey (PNAD), variables will be extracted that, according to the literature on dropout, best explain continuous school abandonment. The results of this article can help teachers and schools track and provide more support to those students who are at high risk of dropping out for socioeconomic reasons.
Keywords: *School dropout; Machine learning; PNAD; Socioeconomic variables; Secondary education.*

Cite as: SILVA, B. A. D.; GURGEL, L.; FRANÇA, R. D.; NASCIMENTO, R. P. do. Aprendizado de Máquina Aplicado à Predição de Evasão no Ensino Médio em São Paulo. Revista Brasileira de Informática na Educação, vol , pp-pp, 2025. .

Resumen

La deserción escolar en la educación brasileña se ha convertido en un gran desafío para la calificación y la formación educativa de los niños y jóvenes, especialmente en los sectores menos favorecidos de la sociedad brasileña. Los jóvenes que no completan la educación secundaria no pueden especializarse en cursos superiores y, por lo tanto, se ven obligados a aceptar empleos con mano de obra barata, perpetuando un ciclo de pobreza y bajos indicadores socioeconómicos. Por ello, este trabajo tiene como objetivo evaluar y desarrollar modelos probabilísticos de aprendizaje automático con el fin de predecir la probabilidad de que un estudiante abandone la educación secundaria en la red educativa de São Paulo, basándose en técnicas como bosques aleatorios, árboles de decisión y redes neuronales. A partir de la base de datos de la Encuesta Nacional por Muestra de Hogares (PNAD) del Instituto Brasileño de Geografía y Estadística (IBGE), se extraerán variables que, según la literatura de estudios sobre el abandono escolar, explican mejor el abandono escolar continuo. Los resultados de este artículo pueden ayudar a los docentes y a las escuelas a rastrear y brindar más apoyo a aquellos alumnos que presentan un alto riesgo de deserción por motivos socioeconómicos.

Palabras clave: *Absentismo escolar; Aprendizaje automático; PNAD; Variables socioeconómicas; Educación secundaria.*

1 Introdução

A evasão escolar constitui um dos principais desafios para a educação brasileira, representando não apenas a interrupção de trajetórias individuais, mas também a perpetuação de desigualdades sociais e o comprometimento do desenvolvimento econômico do país. No contexto do ensino médio, etapa final da educação básica, o fenômeno assume contornos particularmente preocupantes: segundo indicadores de fluxo escolar do Censo Escolar (INEP, 2023), o ensino médio apresenta taxa de abandono de aproximadamente 6%, com disparidades significativas entre as redes de ensino. A rede pública concentra os maiores índices de evasão, enquanto a rede privada apresenta taxas substancialmente inferiores, evidenciando desigualdades estruturais que ultrapassam questões meramente pedagógicas.

Compreender os fatores associados à evasão exige reconhecer sua natureza multidimensional. A literatura nacional e internacional aponta para a convergência de elementos socioeconômicos — como baixa renda familiar, trabalho precoce e insegurança alimentar —, fatores acadêmicos — repetência, distorção idade-série e baixo desempenho — e variáveis institucionais — infraestrutura escolar, formação docente e tamanho de turmas (SILVA, 2016; ARAQUE; ROLDÁN; SALGUERO, 2009). Essa complexidade demanda abordagens analíticas capazes de quantificar a contribuição relativa de cada dimensão e orientar políticas públicas baseadas em evidências.

Dados do IBGE (2024) mostram que cerca de 8,7 milhões de jovens entre 14 e 29 anos abandonaram os estudos ou nunca frequentaram a escola, sendo que parcela significativa não havia concluído o ensino médio. As consequências desse abandono são duradouras: a renda média de trabalhadores com ensino médio completo é significativamente superior à de quem abandonou os estudos antes dessa etapa, perpetuando ciclos de vulnerabilidade e limitando oportunidades de mobilidade social. Em perspectiva comparada, o Brasil apresenta taxas de abandono escolar superiores à média de países da América Latina e muito distantes das registradas em nações com sistemas educacionais mais consolidados, evidenciando a necessidade urgente de políticas públicas de permanência e combate à evasão.

Apesar da relevância do tema, ainda persiste lacuna de estudos que integrem, em escala nacional, dados educacionais e socioeconômicos para análise sistemática dos determinantes da evasão. Grande parte da literatura brasileira concentra-se em recortes locais ou institucionais, com ênfase em modelos preditivos de risco individual voltados ao ensino superior. Investigações que articulem bases de dados oficiais — como os microdados do Censo Escolar, do Sistema de Avaliação da Educação Básica (SAEB) e da Pesquisa Nacional por Amostra de Domicílios (PNAD) — para analisar o fenômeno em nível municipal, identificando padrões territoriais e desigualdades regionais, permanecem escassas.

Nesse contexto, este trabalho propõe uma análise explicativa da evasão no ensino médio brasileiro a partir de modelagem estatística baseada em Regressão Linear Múltipla. O objetivo é investigar em que medida variáveis institucionais — como formação docente, infraestrutura e composição de turmas —, variáveis acadêmicas — taxa de repetência, distorção idade-série e desempenho em avaliações — e variáveis socioeconômicas — renda per capita, trabalho precoce e acesso a programas sociais — contribuem para explicar a variação das taxas de evasão entre municípios brasileiros.

A escolha da Regressão Linear Múltipla justifica-se por sua capacidade de quantificar o

impacto individual de cada variável sobre a taxa de evasão, oferecendo coeficientes diretamente interpretáveis que facilitam a compreensão dos mecanismos subjacentes ao fenômeno e a aplicação dos resultados em políticas educacionais. Diferentemente de abordagens orientadas à predição de risco individual, este estudo privilegia a análise agregada em nível municipal, permitindo identificar contextos territoriais de maior vulnerabilidade e orientar estratégias de intervenção direcionadas.

Para tanto, foram integradas bases de dados do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) — incluindo indicadores educacionais do Censo Escolar e microdados do SAEB — e da Pesquisa Nacional por Amostra de Domicílios (PNAD) do IBGE. Essa articulação possibilita examinar simultaneamente dimensões escolares e sociais, ampliando a compreensão sobre como fatores estruturais, pedagógicos e socioeconômicos interagem na determinação da evasão escolar no ensino médio brasileiro. Os resultados obtidos visam subsidiar gestores públicos, pesquisadores e educadores na formulação de políticas mais efetivas para a redução da evasão e promoção da equidade educacional.

2 Fundamentos Teóricos

O presente capítulo tem como objetivo apresentar a base conceitual que sustenta esta pesquisa, fornecendo o embasamento necessário para compreender as principais variáveis que influenciam a evasão escolar no Brasil em nível municipal. A fundamentação teórica busca situar o estudo no contexto das pesquisas já existentes, permitindo identificar os fatores socioeconômicos, educacionais e estruturais que impactam a permanência dos estudantes, bem como os modelos e abordagens utilizados para analisar tais fenômenos. Além disso, o capítulo descreve os fundamentos teóricos, as bases de dados e as técnicas estatísticas e computacionais empregadas na pesquisa, estabelecendo o alicerce metodológico para as análises realizadas nos capítulos seguintes.

2.1 Evasão Escolar e Fatores Socioeconômicos

A evasão escolar é um fenômeno amplamente estudado nas áreas da Educação e das Ciências Sociais, podendo ser associado a fatores de ordem social, econômica e cultural. Segundo Araque, Roldán e Salguero (2009), as principais razões que influenciam o abandono da escola estão relacionadas a variáveis socioeconômicas, familiares, institucionais e psicológicas. Nesse sentido, compreender esse fenômeno requer considerar tanto o contexto social do aluno quanto as condições oferecidas pelo sistema educacional.

2.2 Mineração de Dados Educacionais e Aprendizado de Máquina

No campo da análise de dados, o termo Mineração de Dados Educacionais (*Educational Data Mining — EDM*) refere-se ao uso de métodos computacionais para explorar grandes volumes de dados gerados em ambientes educacionais, buscando padrões relevantes (BAKER; ISOTANI; CARVALHO, 2011). O objetivo disso é apoiar a tomada de decisão em políticas públicas e institucionais, fornecendo evidências que permitam compreender e reduzir problemas como a própria evasão escolar.

Associado à EDM, o Aprendizado de Máquina (*Machine Learning*) é um subcampo da Inteligência Artificial que possibilita a criação de modelos preditivos a partir de dados. Esses modelos são capazes de identificar relações complexas entre variáveis e gerar previsões com base em informações históricas (MITCHELL, 1997). No contexto da evasão escolar, técnicas de aprendizado supervisionado permitem estimar a probabilidade de um estudante abandonar ou concluir seus estudos, dadas suas características socioeconômicas, acadêmicas e pessoais (TEODORO; KAPPEL, 2020).

2.3 Principais Técnicas Utilizadas

Entre as abordagens utilizadas para a análise de fatores que influenciam fenômenos complexos, destaca-se a Regressão Linear Múltipla, que permite investigar a relação entre uma variável dependente contínua e múltiplas variáveis independentes. Esse método possibilita compreender de forma quantitativa o impacto de diferentes fatores sobre o resultado observado, identificando quais variáveis apresentam maior influência e em que magnitude (MONTGOMERY; PECK; VINING, 2012). No contexto da evasão escolar, a Regressão Linear Múltipla é aplicada para estimar como aspectos socioeconômicos, educacionais e estruturais contribuem para a variação das taxas de evasão entre municípios, permitindo uma análise explicativa e comparativa das desigualdades regionais.

2.4 Técnicas Computacionais Complementares

Embora este trabalho foque exclusivamente na aplicação da Regressão Linear Múltipla, outras técnicas computacionais também têm sido amplamente empregadas em estudos voltados à predição e análise de dados educacionais. Entre essas abordagens, destacam-se métodos como Regressão Logística, Redes Neurais Artificiais e Florestas Aleatórias de Classificação, amplamente utilizados em pesquisas que buscam estimar a probabilidade de evasão escolar e identificar padrões complexos de comportamento estudantil (HOSMER; LEMESHOW, 2000; GARDNER; DORLING, 1998; BREIMAN, 2001).

Apesar da diversidade de métodos existentes, a escolha da Regressão Linear Múltipla neste estudo se justifica por sua capacidade de quantificar o impacto individual de múltiplas variáveis explicativas sobre a taxa de evasão escolar, permitindo identificar quais fatores exercem maior influência sobre o fenômeno. Essa abordagem fornece não apenas um modelo preditivo, mas também uma interpretação direta dos coeficientes, facilitando a compreensão e a aplicação dos resultados em políticas públicas e estratégias educacionais.

2.5 A Pesquisa Nacional por Amostra de Domicílios (PNAD)

A Pesquisa Nacional por Amostra de Domicílios (PNAD), realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE), constitui uma das principais fontes de dados socioeconômicos no Brasil. Seu objetivo é coletar informações abrangentes sobre características demográficas, educacionais, ocupacionais e de rendimento da população brasileira, por meio de entrevistas domiciliares aplicadas em amostras representativas em nível nacional, regional e estadual (IBGE, 2022).

A PNAD Contínua, em vigor desde 2012, aprimorou o levantamento ao adotar coleta trimestral, permitindo análises mais atualizadas e consistentes acerca da dinâmica social e econômica do país. Entre suas variáveis, destacam-se renda familiar per capita, inserção no mercado de trabalho, características do domicílio, composição familiar, escolaridade e acesso a programas sociais. Tais informações são de grande relevância para estudos sobre evasão escolar, pois possibilitam identificar relações entre vulnerabilidade socioeconômica e permanência na escola.

2.6 Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP)

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), autarquia federal vinculada ao Ministério da Educação (MEC), tem como missão promover estudos, pesquisas e avaliações sobre o sistema educacional brasileiro. Fundado em 1937, o INEP é responsável pela produção e disseminação de informações estatísticas e avaliativas que subsidiam a formulação e o monitoramento de políticas públicas educacionais em âmbito nacional (INEP, 2023).

O Instituto realiza levantamentos censitários anuais em todas as etapas da educação básica e superior, além de coordenar avaliações de larga escala, como o Sistema de Avaliação da Educação Básica (SAEB) e o Exame Nacional do Ensino Médio (ENEM). Os dados produzidos pelo INEP constituem a principal fonte oficial de informações sobre matrícula, rendimento escolar, infraestrutura, profissionais da educação e fluxo escolar no país, sendo essenciais para a análise de fenômenos como evasão, repetência e distorção idade-série.

2.7 Microdados do Censo Escolar

O Censo Escolar, realizado anualmente pelo INEP em regime de colaboração com as secretarias estaduais e municipais de educação, constitui o principal instrumento de coleta de informações sobre a educação básica brasileira. Seu objetivo é reunir dados detalhados sobre estudantes, turmas, escolas e profissionais da educação em todas as redes de ensino – públicas e privadas –, abrangendo as etapas de educação infantil, ensino fundamental e ensino médio (INEP, 2023).

Os microdados do Censo Escolar são disponibilizados publicamente pelo INEP e permitem análises granulares sobre características individuais dos estudantes, como idade, sexo, raça/cor, situação de matrícula, modalidade de ensino e deficiências, além de informações sobre infraestrutura escolar, localização, dependência administrativa e recursos disponíveis. Esses dados são fundamentais para estudos sobre evasão escolar, pois possibilitam acompanhar longitudinalmente a trajetória estudantil, identificar fatores associados ao abandono e à reprovação, bem como analisar disparidades regionais e socioeconômicas que impactam o acesso e a permanência na escola.

2.8 Sistema de Avaliação da Educação Básica (SAEB)

O Sistema de Avaliação da Educação Básica (SAEB), também coordenado pelo INEP, constitui uma das principais avaliações externas e padronizadas da qualidade da educação no Brasil. Instituído em 1990, o SAEB tem como objetivo diagnosticar a educação básica brasileira e produzir indicadores sobre o desempenho dos estudantes em Língua Portuguesa e Matemática, além de coletar informações contextuais sobre condições de aprendizagem, perfil docente e clima escolar (INEP, 2023).

A partir de 2019, o SAEB passou a ser censitário para estudantes dos anos finais do ensino fundamental e do ensino médio, permitindo avaliações em nível de escola, município e estado. Os microdados do SAEB, disponibilizados junto aos do Censo Escolar, incluem resultados de proficiência, questionários contextuais aplicados a estudantes, professores e diretores, além de indicadores socioeconômicos e de infraestrutura. Tais informações são relevantes para estudos sobre evasão escolar, pois possibilitam relacionar desempenho acadêmico, contexto familiar e escolar com a permanência ou abandono dos estudantes, oferecendo subsídios para políticas educacionais mais direcionadas e eficazes.

2.9 Síntese

Dessa forma, os fundamentos apresentados evidenciam a relevância de associar teorias sobre evasão escolar com métodos de mineração de dados e aprendizado de máquina, bem como o uso de bases de dados socioeconômicas e educacionais amplas como a PNAD, Censo Escolar e SAEB. Essa integração permite analisar grandes quantidades de dados e construir modelos preditivos capazes de apoiar políticas públicas e institucionais na área da Educação, contribuindo para a redução da evasão escolar e para a promoção da equidade educacional.

3 Trabalhos Relacionados

Diversos estudos buscaram modelar de forma probabilística ou com técnicas de aprendizado de máquina o fenômeno da evasão escolar no Brasil. Como exemplo, destacam dois trabalhos acerca da evasão no âmbito acadêmico.

Em Mello et al. (2023) foram realizadas análises exploratórias acerca das características consideradas (variáveis independentes) como coeficiente de rendimento, tipo de escola de origem, percentual de frequência, nível de ensino, cor ou raça, modalidade de curso, sexo, renda familiar bruta, situação acadêmica, dentre outros aspectos, de alunos do Instituto Federal do Pernambuco (IFPE). Com a análise constatou-se que, como alguns exemplos, alguns cursos e períodos apresentam evasão maior que outros; o percentual de evasão do sexo masculino era maior em comparação ao feminino. Contudo, não houve análise acerca da renda per capita dos indivíduos, que é reiterada por autores que exploraram o tema de forma qualitativa, como Silva (2016) e Ferreira & Oliveira (2020). Os autores utilizaram como técnica de aprendizado de máquina para a previsão de evasão XGBoost, Florestas Aleatórias e Árvores de decisão, para predizer a situação de matrícula do estudante conforme a base de dados disponível, e obtiveram, respectivamente, 82%, 83% e 80% de acurácia. Esses resultados sugerem que tais modelos apresentam ótimo desempenho preditivo para estimar as probabilidades de evasão escolar, em concordância com o objetivo deste trabalho.

Por outro lado, o trabalho de Teodoro e Kappel (2020) também segue a mesma lógica do trabalho supracitado, com a ressalva de que tem ênfase em atributos diferentes e apresenta uma análise exploratória mais robusta acerca das características correlacionadas à evasão. Teodoro e Kappel optaram por desenvolver modelos preditivos baseados em Naive Bayes, KNN, Árvores de Decisão, Florestas Aleatórias de Classificação e Redes Neurais, com acurácias de 60%, 75%, 77%, 79% e 78%, respectivamente, no geral. A apuração mostra que, em termos de acurácia, Florestas Aleatórias e Redes Neurais são as técnicas mais adequadas para estimação de evasão.

Porém, assim como o estudo anterior, tem como principal objeto de estudo a evasão no ensino superior.

Apesar de trabalhos aplicados ao ensino médio/técnico (por exemplo, Barbosa et al., 2023), observa-se que a literatura nacional permanece concentrada em bases institucionais e recortes locais. Em âmbito brasileiro, ainda há escassez de investigações que integrem variáveis socioeconômicas de abrangência nacional — como renda per capita, trabalho precoce e insegurança alimentar — a modelos preditivos para o ensino médio nas redes pública e privada. Este artigo procura preencher essa lacuna ao combinar PNAD e aprendizado de máquina para estimar probabilidades de risco de evasão.

Além das características individuais e domiciliares capturadas pela PNAD, a literatura indica que fatores estruturais ao nível da escola também se associam ao risco de evasão no ensino médio. Evidências reportam relações entre escolas de maior porte, turmas numerosas e razões aluno-professor elevadas com maior probabilidade de abandono, sugerindo que variáveis de contexto educacional podem complementar os atributos socioeconômicos no modelo preditivo. Em um desenho nacional, tais elementos podem ser operacionalizados por indicadores derivados de bases administrativas educacionais, agregados por unidade da federação ou município e vinculados aos registros individuais via chaves geográficas.

Outro conjunto de determinantes envolve práticas e fluxos escolares. A literatura brasileira relaciona a repetência/atraso escolar, a disponibilidade de vagas/turnos e a qualificação docente a maiores taxas de abandono, o que respalda a inclusão de medidas como a taxa de reprovação no ensino médio e a proporção de docentes com formação adequada como variáveis de contexto. Esses indicadores, agregados por recortes territoriais, permitem capturar diferenças institucionais relevantes sem exigir a identificação da escola no nível do indivíduo, preservando a modelagem em escala nacional.

Condições de infraestrutura educacional — como conectividade, laboratórios e manutenção predial — também são associadas a níveis de engajamento estudantil e, por consequência, à evasão. Uma estratégia pragmática consiste em sintetizar um índice de infraestrutura (via combinação simples de itens ou redução de dimensionalidade) por município ou unidade da federação e integrá-lo como variável de contexto ao lado dos determinantes socioeconômicos da PNAD. Esse acoplamento amplia a cobertura explicativa do modelo ao articular condições de vida e ambiente escolar.

Em termos de interpretação e uso, trabalhos aplicados destacam o papel de técnicas de *ensemble* e análises de *feature importance* para orientar intervenções pedagógicas e de gestão, ao mesmo tempo em que estudos qualitativos em ensino médio ajudam a elucidar por que variáveis como renda per capita, trabalho precoce e insegurança alimentar emergem como relevantes. Essa combinação de previsibilidade e interpretabilidade é coerente com o objetivo de apoiar decisões em escala nacional, mantendo atenção às diferenças regionais e às especificidades das redes.

Em linha com a ênfase em determinantes contextuais, evidências regionais analisando o ensino médio público do Ceará, com modelos logísticos multiníveis, exploram simultaneamente características dos estudantes e das escolas. Os resultados destacam a contribuição de fatores como desinteresse declarado, histórico de repetência e defasagem idade-série para o aumento do risco de abandono, enquanto a participação em programas de transferência de renda apresenta associação negativa com a evasão. Esse tipo de achado complementa o enfoque nacional ao indicar

que variáveis de fluxo escolar e proteção social ajudam a explicar variações do risco para além das condições domiciliares.

Já em aplicações baseadas em dados administrativos, há estudo recente com a rede estadual paulista que estrutura um sistema de alerta precoce com classificadores de aprendizado de máquina (por exemplo, florestas, *gradient boosting* e regressão logística) a partir de histórico escolar e registros de frequência. Além de reforçar o bom desempenho de métodos de *ensemble*, o trabalho relata ganhos práticos ao combinar importância de atributos com regras de priorização de atendimento, o que dialoga com a proposta deste artigo de aliar previsibilidade e utilidade para a gestão em escala nacional.

Questões metodológicas recorrentes. A literatura destaca desafios que também norteiam nossa abordagem: (i) desbalanceamento de classes — a evasão costuma ser minoria; (ii) métricas alinhadas ao objetivo de intervenção — acurácia isolada pode mascarar falsos negativos; (iii) explicabilidade — necessária para a adoção por equipes pedagógicas; e (iv) transferência temporal — modelos treinados em um período podem degradar com mudanças de coorte/currículo. Em linha com essas recomendações, adotamos métricas como F1-score, AUC-ROC e, quando pertinente, AUC-PR, priorizando sensibilidade/*recall* na classe de risco; e técnicas de reamostragem (subamostragem, conforme nossa Metodologia) quando cabível. A análise de *feature importance* é tratada como insumo interpretável para orientar ações, em consonância com práticas relatadas em estudos aplicados (Barbosa et al., 2023; Teodoro & Kappel, 2020).

Implicações para variáveis e desenho de intervenções. Estudos qualitativos ajudam a explicar por que determinadas variáveis aparecem entre as mais relevantes nos modelos. Fatores como renda, trabalho e alimentação são recorrentemente associados à decisão de abandonar a escola no ensino médio brasileiro. Ao incorporar essas dimensões — operacionalizadas aqui via PNAD — produz-se um modelo que não apenas prediz, mas também dialoga com mecanismos plausíveis de evasão descritos na literatura (Silva, 2016; Ferreira & Oliveira, 2020), o que favorece o desenho de intervenções focalizadas (auxílio financeiro, merenda/reforço de programas de alimentação, mediação de estágio/trabalho protegido, acompanhamento social).

Considerações de adoção e equidade. A transferência dos resultados para a prática escolar exige atenção à calibração de probabilidades (para que o limiar de alerta reflita custos/benefícios de intervenção) e à equidade (evitar que o modelo reforce desigualdades ao superidentificar grupos específicos). O uso de métricas sensíveis à classe minoritária (Strauss; Villas Bôas Júnior; Ferreira, 2022) e a verificação de viés por subgrupos (sexo, raça/cor) ajudam a balizar decisões responsáveis. Tais cuidados metodológicos aumentam a aceitabilidade do sistema por gestores e docentes e estão alinhados às melhores práticas de *Learning Analytics* aplicadas à educação básica.

Nota de escopo. Embora a revisão sintetize achados nacionais e internacionais, o objetivo empírico deste trabalho permanece delimitado ao ensino médio brasileiro, abrangendo as redes pública e privada, em escala nacional. As referências mais amplas servem para contextualizar o estado da arte e não implicam generalização automática para períodos futuros ou outros contextos; discutimos validade externa e possíveis diferenças regionais nas seções de Resultados e Conclusões.

Em complemento, estudos qualitativos voltados especificamente ao ensino médio contribuem para fundamentar a seleção de variáveis e o enquadramento do problema. Batista, Souza e

Oliveira (2009) apresentam um estudo de caso que discute fatores estruturais, familiares e socioeconômicos associados ao abandono na etapa final da educação básica. Embora não empregue técnicas de aprendizado de máquina, o trabalho oferece evidências sobre a relevância de indicadores como renda per capita, trabalho precoce e insegurança alimentar, que dialogam diretamente com as variáveis extraídas da PNAD neste estudo. Assim, a literatura qualitativa auxilia a interpretabilidade dos modelos, ao delinear mecanismos plausíveis pelos quais o contexto social do discente influencia a probabilidade de evasão.

Do ponto de vista metodológico, a literatura também enfatiza a necessidade de métricas adequadas para avaliar modelos preditivos em cenários com classes desbalanceadas, como a evasão (menos frequente do que a permanência). Strauss, Villas Bôas Júnior e Ferreira (2022) discutem as limitações da acurácia como métrica isolada e recomendam o uso de F1-score, AUC-ROC e, quando pertinente, AUC-PR para capturar de forma mais fiel o desempenho sobre a classe minoritária. Essas diretrizes sustentam as escolhas de avaliação adotadas neste trabalho e reforçam a importância de reportar métricas sensíveis a falsos negativos, dada a natureza de risco e intervenção associada ao problema.

4 Metodologia

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) dispõe uma vasta base de microdados e dados abertos baseados em variáveis institucionais, acadêmicas e socioeconômicas do sistema de ensino brasileiro no decorrer das últimas 2 décadas. Como complemento, a Pesquisa Nacional por Amostra de Domicílios (PNAD), realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE), é um conjunto de informações detalhadas sobre o cenário socioeconômico da sociedade brasileira. A partir dela, é possível extrair dados e informações das características gerais da população.

Apoiado em tais conjuntos de dados, a metodologia deste artigo consiste no processamento desses dados e a utilização de técnicas estatísticas que possibilitem uma análise quantitativa acerca da evasão de alunos do ensino médio brasileiro, e quais fatores explicam tal fenômeno, dadas as suas características de cunho social, econômico, acadêmico e institucional. Para tal, foram utilizadas as linguagens R, Python e a biblioteca Statsmodels.

Conjuntos de dados provenientes do INEP foram extraídos do site de dados abertos do instituto na aba de "Indicadores Educacionais"e "Microdados". Bases de dados como a taxa de repetência e de evasão, taxa de distorção idade-série, índice do esforço docente, número médio de alunos por turma, nível de adequação do docente e microdados da educação básica (todos a nível municipal e do ano de 2017) foram coletados, extraindo apenas ocorrências do ensino médio, em formato de planilhas xlsx ou arquivo com valores separados por vírgula (*CSV*) e lidos posteriormente em Python com a biblioteca de manipulação de dados Pandas.

Em sequência, foi realizada a junção das bases supracitadas por meio das colunas, renomeadas a partir da amostra original, UF (sigla da unidade federativa), COD_MUNICIPIO (código do município), LOCALIZACAO (urbana ou rural) E DEP_ADMINISTRATIVA (instituição privada ou pública), consideradas chave primária composta do conjunto de dados. Deste modo, foi construído uma única estrutura de dados tabular (*dataframe*) contendo as variáveis institucionais

relevantes à manifestação da evasão na fase final da educação básica fundamento na literatura da evasão escolar no Brasil e no mundo (BANAAG ET AL., 2024; SHIRASU & ARRAES, 2018; SOUSA ET AL., 2025).

Para modelagem estatística ulterior isolada também foi coletada uma amostra a nível escolar do Sistema de Avaliação da Educação Básica (SAEB), com variáveis sociais e acadêmicas no ano de 2017, também oriunda dos microdados do INEP.

A base de dados da PNAD foi obtida diretamente a partir do website do IBGE na seção PNAD Contínua e lida preliminarmente em R para interpretação dos dados de largura fixa a largura variável com o auxílio da biblioteca PNADcIBGE. Foram filtrados apenas os resultados pertinentes, a partir do de ano de 2016 até 2024, à analise de ocorrência de pessoas que não frequentam mais a escola, frequentaram escola alguma vez, cursaram como grau mais elevado o ensino médio regular ou 2º grau, e com idade até 18 anos. Com base nesses dados é possível rotular ocorrência de evasão ou conclusão do ensino médio como variável binária exclusiva dependente baseada na conclusão do curso.

Após o procedimento de filtragem, foi feita uma seleção das características, ou colunas, mais relevantes para explicar o fenômeno da evasão escolar. Características como sexo, cor ou raça, se já trabalhou ou estagiou por pelo menos 1 hora em alguma atividade remunerada em dinheiro, remunerada em mercadorias e bens, não remunerada ou atividade ocasional ("bico"), número de componentes do domicílio (exclusive as pessoas cuja condição no domicílio era pensionista, empregado doméstico ou parente do empregado doméstico), se recebe bolsa família ou outro auxílio governamental e rendimento domiciliar per capita (habitual de todos os trabalhos e efetivo de outras fontes), as quais são algumas das características que mais se correlacionam qualitativamente ao abandono e à evasão escolar no ensino médio (FERREIRA; OLIVEIRA, 2020).

Foi gerado um arquivo CSV com os dados já filtrados e características selecionadas, e posteriormente este foi lido em Python para limpeza de dados faltantes, e verificação dos tipos de variáveis com o objetivo de utilizá-lo como base de dados complementar à análise de evasão. Foi também criada a coluna "evasao"codificada de forma binária para classificar o aluno como evasão (1) e conclusão (0).

4.1 Pré-processamento dos Dados

As taxas obtidas a partir da base do INEP, inicialmente no formato 0 a 100 foram transformadas para formato decimal (0 a 1) como:

$$T_k = \frac{t_k}{100} \tag{1}$$

Em que T_k representa o vetor modificado e t_k representa o vetor (coluna) da taxa original

Baseado nas limitações de tipo do modelo e da biblioteca utilizada, variáveis categóricas foram separadas em novas colunas de acordo com o número de valores únicos de texto presentes na coluna única original, transformadas cada coluna nova em vetores binários e na sequência a re-

moção de uma ou mais colunas para evitar correlação linear entre as características artificialmente separadas.

Variáveis ordinais numéricas foram tranformadas em variáveis categóricas textuais e, em sequência, o fluxo de transformação citado anteriormente para variáveis categóricas foi aplicado.

Células vazias ou nulas tiveram suas linhas removidas por completo da base de modelagem para integridade da análise estatística.

4.2 Análise de Multicolinearidade

Multicolinearidade pode ser compreendida como uma relação linear entre duas ou mais variáveis dependentes. Conforme Paul (2006), quando há importância na investigação dos impactos dos regressores na variável dependente, a multicolinearidade pode ser um problema, visto que p-valores podem se mostrar equivocadamente elevados e em alguns casos pode interferir na interpretação dos coeficientes. Uma das maneiras de analisar se há multicolinearidade no conjunto de dados é calcular o fator de inflação de variância (*VIF*), baseado em R^2, que indica o quanto uma variável independente é explicada pelos demais regressores para cada umas das i variáveis indepedentes (MILOCA & CONEJO, 2008).

No presente trabalho será aplicado o *VIF* para verificar a mutlicolinearidade entre as variáveis advindas das bases do INEP:

$$F_i = \frac{1}{1 - R_i^2} \tag{2}$$

F_i se refere ao fator de inflação de variância do i-ésimo regressor; R_i^2 se refere ao R^2 da i-ésima variável independente.

Um *VIF* maior que 10 indica que a multicolinearidade influencia fortemente o valor dos coeficientes do modelo, como proposto por Johnson e Wichern (1988; apud MILOCA & CONEJO, 2008), e algumas medidas em relação ao conjunto de dados ou ao modelo devem ser tomadas com o objetivo de preservar a interpretação dos dados na modelagem estatística.

4.3 Modelo Estatístico

Este artigo visa aplicar métodos estatísticos com o intuito de investigar o impacto das variáveis socioeconômicas e institucionais na evasão escolar e em variáveis que corroboram tal fenômeno. Para tanto, conforme Green et al (2011), que salientam a regressão linear como ferramenta amplamente usada para examinar as relações estatísticas entre variáveis, nesta pesquisa será utilizado, primordialmente, o modelo de regressão linear múltipla, expresso por:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_i X_i + \varepsilon \tag{3}$$

Em que Y é a variável dependente a ser modelada, X_i são as variáveis independentes, ou regressores, β_i representa os coeficientes atrelados a cada variável independente e ε representa o erro aleatório.

Com o modelo, e o auxílio da técnica dos mínimos quadrados para quantificar os coeficientes, é possível mensurar a significância das diferentes variáveis independentes, e o quanto sua variação impacta na variável dependente, de acordo com o contexto das amostras e das variáveis coletadas.

4.4 Métricas de Desempenho

Com o objetivo de extrair resultados acerca do desempenho do modelo deste trabalho, foram utilizadas as técnicas R^2 e a média dos quadrados dos resíduos (*MSE*). Por meio destas é possível ter ciência do quanto o modelo explica a variação da variável dependente que varia de 0 a 1 (sendo 1 o ajuste perfeito dos dados pelo modelo) e a média da magnitude de erros do modelo, respectivamente:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{4}$$

$$E_{qm} = \frac{1}{n} \sum_i (y_i - \hat{y}_i)^2 \tag{5}$$

De forma que y_i é o valor real, $\hat{y}_i$ é valor predito pelo modelo e $\bar{y}$ é a média dos valores da variável dependente.

5 Resultados Parciais

Com base nos dados coletados a partir dos dados abertos do INEP e na regressão linear múltipla resumida na tabela 1:

O modelo tem erro quadrático médio $E_{qm} = 0,0026$ o que indica que erra cerca de 5 pontos percentuais para fins de predição. Ao mesmo tempo, apresenta um $R^2 = 0,514$, explicando 51% da variação dos dados, e uma $Prob(F) = 0$, que indica que o modelo é significativo para explicar a variável dependente.

Os resultados mostram que, ao manter as outras variáveis constantes, quando TX_DI_TOTAL, a taxa de defasagem idade-série, aumenta em 1 ponto percentual, a taxa de evasão total tende

Tabela 1: Resultados da regressão linear MQO para a variável dependente `TX_EV_TOTAL`.

Variável	**Coef.**	**Erro Padrão**	**t**	**P>|t|**
const	-0,0029	0,003	-0,898	0,369
DEP_ADMINISTRATIVA_Pública	0,0273	0,002	16,393	0,000
Grupo 2	0,0572	0,009	6,338	0,000
Grupo 3	-0,0115	0,004	-3,189	0,001
Grupo 4	-0,0168	0,006	-2,989	0,003
Grupo 5	0,0049	0,005	1,079	0,281
IN_BIBLIOTECA	0,0016	0,002	0,694	0,488
IN_COZINHA	-0,0031	0,003	-1,143	0,253
IN_INTERNET	0,0178	0,003	6,740	0,000
IN_LABORATORIO_CIENCIAS	-0,0052	0,004	-1,253	0,210
IN_LABORATORIO_INFORMATICA	0,0011	0,003	0,413	0,680
IN_QUADRA_ESPORTES	0,0025	0,003	0,881	0,378
NUM_ALUNO_TURMA_TOTAL	0,0000	0,000	0,642	0,521
TX_DI_TOTAL	0,1853	0,004	41,654	0,000
TX_IED_N5	0,0070	0,004	1,767	0,077
TX_IED_N6	-0,0112	0,006	-1,939	0,053
TX_REP_TOTAL	0,0528	0,009	5,622	0,000

$R^2 = 0,514$, Estatística-F = 455,7, Prob(F) = 0,00, $E_{qm} = 0,0026$

Observações = 6915

a aumentar em 0,1853 pontos percentuais. O coeficiente é estatisticamente significativo com $p < 0.001$ e se mostra como o coeficiente mais alto que se relaciona com a variável dependente TX_EV_TOTAL.

De maneira semelhante, TX_REP_TOTAL, que demonstra a taxa de repetência total do município, também apresenta alto coeficiente em comparação com as outras variáveis, sendo $0,0528$, e estatisticamente significativo com $p < 0,0001$.

Os resultados também sugerem que, com $N = 6915$ observações, as variáveis de infraestrutura coletadas não se mostram, em sua maioria, como estatisticamente significadas, e aquela que apresentam p-valor menor que 5% (IN_INTERNET) exibe coeficiente $\beta_i < 0,0178$, o que caracteriza uma ínfima variação de pontos percentuais na taxa de evasão total. Taxas do índice de esforço do docente (TX_IED_N5 e TX_IED_N5) de forma análoga não se mostram estatisticamente significativos, com $p > 0,05$.

Em resumo, as variáveis Grupo 2, TX_DI_TOTAL, TX_REP_TOTAL, são evidenciadas como as mais relevantes no presente estudo.

com base, em particular, na taxa de repetência, decidiu-se utilizar a base do Sistema de Avaliação do Educação básica (SAEB) do INEP, para investigar quais variáveis mais impactam na taxa de repetência e, de forma implícita, na taxa de defasagem idade-série, dado que a repetência leva à defasagem em relação à série cursada.

Com base na tabela 2, que resume os resultados da regressão linear múltipla da base de dados do SAEB:

Tabela 2: Resultados da regressão linear MQO para a variável dependente `TX_RESP_Q041`.

Variável	Coeficiente	Erro padrão	t	p>|t|	[0,025 ; 0,975]
const	0,6980	0,039	17,716	0,000	[0,621 ; 0,775]
PROFICIENCIA_MT	-0,1567	0,003	-61,733	0,000	[-0,162 ; -0,152]
TX_RESP_Q001	0,0841	0,010	8,203	0,000	[0,064 ; 0,104]
TX_RESP_Q002_Indígena	0,1083	0,018	5,879	0,000	[0,072 ; 0,144]
TX_RESP_Q002_Parda	0,0474	0,007	6,716	0,000	[0,034 ; 0,061]
TX_RESP_Q002_Preta	0,1422	0,012	11,985	0,000	[0,119 ; 0,165]
TX_RESP_Q027	-0,4332	0,039	-11,121	0,000	[-0,510 ; -0,357]
TX_RESP_Q038	0,0595	0,007	8,035	0,000	[0,045 ; 0,074]
TX_RESP_Q044	-0,1006	0,010	-10,433	0,000	[-0,120 ; -0,082]
TX_RESP_Q052	0,0965	0,009	11,240	0,000	[0,080 ; 0,113]

$R^2 = 0,272$, Estatística F = 782,4, Prob(F) = 0,00, $E_{qm} = 0,0299$
Número de observações = 18846

O modelo se mostra estatisticamente significativo dada $Prob(F) = 0$, e explica 27,2% da variação dos dados coletados.

Pode-se observar que das variáveis coletadas todas são significativas com $p < 0,001$, e se destacam as variáveis TX_RESP_Q027, taxa de alunos incentivados pelos pais aos estudos, com um alto coeficiente de $\beta_i = -0,4332$, PROFICIENCIA_MT ($\beta_i = -0,1567$), que mede a proficiência de matemática média da instituição; TX_RESP_Q002_Preta ($\beta_i = 0,1422$), que refere-se à taxa de alunos autodeclarados pretos na instituição de ensino; TX_RESP_Q002_Indígena ($\beta_i = 0,1083$), que faz jus à taxa de alunos autodeclarados indígenas. Tais variáveis possuem os maiores coeficientes do conjunto de regressores.

De acordo com os resultados, o aumento de 1 ponto percentual na taxa de alunos incentivados ao estudo está associado à diminuição de 0,4332 pontos percentuais na taxa de repetência. De modo semelhante, o aumento de 1 ponto percentual na proficiência em matemática tem associação com a diminuição em 0,1567 pontos percentuais da variável dependente.

Por outro lado, o aumento de 1 ponto percentual no grupo de alunos autodeclarados pretos nas instituições tende a elevar a taxa de repetência em 0,1422. E o mesmo pode ser observado para o grupo de alunos indígenas, cujo aumento de 1 ponto percentual implica em $+0.1083$ pontos percentuais na taxa de repetência. Ademais, o aumento de 1 ponto percentual na taxa de alunos que trabalham (TX_RESP_Q038) está associado a $+0,0595$ pontos percentuais na taxa de repetência.

A variável TX_RESP_Q044, que faz referência à taxa de alunos que gostam de estudar língua portuguesa, também mostra que com o aumento de 1 ponto percentual, a taxa de repetência tende a $-0,1006$ pontos percentuais. E curiosamente, o aumento de 1 ponto percentual na taxa de alunos que demonstram interesse por matemática (TX_RESP_Q052) está associado a um aumento de aproximandamente 0,09 pontos percentuais na taxa de repetência.

A correlação entre evasão, repetência e defasagem idade-série é corroborada pela cenário do ensino médio brasileiro. De acordo com dados da PNAD, pesquisa nacional realizada pelo IBGE, coletados nesta pesquisa para análise complementar, a taxa de evasão entre pretos em idade escolar adequada ao ensino secundário é de 22,87%, entre pardos 22,26%, e entre indígenas equivalente a 25,64% (com base na lógica de cálculo de evasão deste trabalho). Brancos apresentam o menor

número com cerca de 16,16%.

No índice de repetência, o grupo de pretos e pardos é cerca de 2 a 4 pontos percentuais acima do percentual de brancos que se mantêm na mesma série em todo o Brasil, de acordo com dados do INEP. Ainda segundo a plataforma inepdata, a taxa de jovens pretos e pardos que estão defasados no quesito idade-série, chega a ser de 10 pontos percentuais em comparação com jovens brancos.

A evasão e dificuldade escolar de tais grupos fragilizados também são refletidos em suas condições socioeconômicas e o inverso também é verdadeiro. Jovens pretos em idade escolar que já estão fora do ensino médio devido à conclusão ou evasão apresentam renda média mensal de 790 reais, jovens pardos 780 reais, jovens indígenas 727 reais, e jovens brancos 1230 reais em média, segundo dados da PNAD no contexto desta pesquisa.

Ante o exposto, os resultados parciais reforçam que, em alguma medida, fatores de infra-estutura e institucionais não são signficativos para explicar a evasão e o índice de repetência tal como fatores acadêmicos e sociais, como a raça e o incentivo doméstico aos estudos e manutenção de presença na instituição de ensino.

5.1 Descrição de Variáveis Selecionadas

Para analisar estatisticamente a evasão e a repetência, foram selecionadas as seguintes variáveis descritas nas tabelas 3 (regressão da taxa de evasão como variável dependente) e 4 (regressão com taxa de repetência como variável dependente):

Tabela 3: Variáveis explicativas do modelo de regressão da taxa de evasão total (TX_EV_TOTAL).

Variável	Descrição
DEP_ADMINISTRATIVA_Pública	Escola pública (1 = sim, 0 = não)
Grupo 2	Docentes com bacharelado na mesma área (Grupo 2)
Grupo 3	Docentes com licenciatura ou bacharelado em área diferente (Grupo 3)
Grupo 4	Docentes com formação superior não enquadrada (Grupo 4)
Grupo 5	Docentes sem formação superior (Grupo 5)
IN_BIBLIOTECA	Percentual de escolas do município com biblioteca
IN_COZINHA	Percentual de escolas do município com cozinha
IN_INTERNET	Percentual de escolas do município com acesso à Internet
IN_LABORATORIO_CIENCIAS	Percentual de escolas do município com laboratório de ciências
IN_LABORATORIO_INFORMATICA	Percentual de escolas do município com laboratório de informática
IN_QUADRA_ESPORTES	Percentual de escolas do município com quadra de esportes
NUM_ALUNO_TURMA_TOTAL	Número médio de alunos por turma
TX_DI_TOTAL	Taxa de distorção idade-série
TX_IED_N5	Percentual de docentes nível 5 do Índice de Esforço Docente
TX_IED_N6	Percentual de docentes nível 6 do Índice de Esforço Docente
TX_REP_TOTAL	Taxa total de reprovação

Fonte: Elaborado pelo autor a partir do Censo Escolar e do INEP.

Tabela 4: Variáveis da regressão da taxa de repetência (TX_RESP_Q041).

Variável	Descrição
PROFICIENCIA_MT	Proficiência média em Matemática da escola
TX_RESP_Q001	Percentual de alunos do sexo masculino
TX_RESP_Q027	Percentual de alunos que são incentivados pelos pais ao estudo
TX_RESP_Q002_Indígena	Percentual de alunos autodeclarados indígenas
TX_RESP_Q002_Parda	Percentual de alunos autodeclarados pardos
TX_RESP_Q002_Preta	Percentual de alunos autodeclarados pretos
TX_RESP_Q038	Percentual de alunos que trabalham fora de casa atualmente
TX_RESP_Q044	Percentual de alunos que gostam de estudar Língua Portuguesa
TX_RESP_Q052	Percentual de alunos que gostam de estudar Matemática

Fonte: Elaborado pelo autor a partir da base SAEB do INEP.