



INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PIAUÍ
CAMPUS TERESINA ZONA SUL
CURSO LICENCIATURA EM COMPUTAÇÃO

JOÃO VICTOR DOS SANTOS MENDES

MODELO PREDITIVO PARA AVALIAÇÃO DE DESEMPENHO ACADÊMICO

TERESINA
2026

JOÃO VICTOR DOS SANTOS MENDES

MODELO PREDITIVO PARA AVALIAÇÃO DE DESEMPENHO ACADÊMICO

Trabalho de Conclusão de Curso (Artigo Científico) apresentado como exigência parcial para obtenção do diploma do Curso de Licenciatura em Computação do Instituto Federal de Educação, Ciência e Tecnologia do Piauí, Campus Teresina Zona Sul.

Orientador: Prof. Dr. Kelson Carvalho Santos

TERESINA
2026

MODELO PREDITIVO PARA AVALIAÇÃO DE DESEMPENHO ACADÊMICO

João Victor dos Santos Mendes¹

Prof. Dr. Kelson Carvalho Santos²

RESUMO

O desempenho acadêmico no ensino médio público é condicionado por fatores sociológicos como pobreza, trabalho juvenil, exclusão digital e distorção idade-série, mas o acesso a micro-dados individuais é restringido por barreiras legais e institucionais. Este trabalho propõe um modelo preditivo de tendência de desempenho para estudantes da rede pública estadual do Piauí, fundamentado em fatores sociológicos e treinado sobre uma população sintética representativa do perfil de uma turma média dessa rede. Cada parâmetro de geração é rastreável a estatísticas oficiais e à literatura, e o gerador incorpora mecanismos anti-deterministas: a assistência financeira deriva de regra legal, as vulnerabilidades interagem de forma não linear e perfis de resiliência acadêmica são preservados. Os indicadores acadêmicos observáveis correspondem a parciais do primeiro semestre, e a variável-alvo representa a tendência ao final do ano, rotulada pelos critérios de aprovação da legislação educacional. Sobre três mil registros, três algoritmos de classificação e dois *baselines* foram comparados sob validação cruzada estratificada, com sobreamostragem confinada a cada partição de treino e teste estatístico pareado. O XGBoost obteve acurácia de 84,11%, equivalente à *Random Forest* e superior à árvore única e aos *baselines*, sem erros entre as classes extremas. Após os sinais acadêmicos parciais, os atributos mais determinantes foram a distorção idade-série, a assistência financeira, o trabalho juvenil e a renda *per capita*, recuperando o peso dos fatores sociológicos no gerador. Conclui-se que a articulação entre dados sintéticos rastreáveis e classificadores de conjunto constitui alternativa metodológica válida para sistemas de alerta precoce em contextos de dados restritos.

Palavras-chave: desempenho acadêmico; fatores sociológicos; dados sintéticos; aprendizado de máquina; ensino médio público.

ABSTRACT

Academic performance in Brazilian public high schools is strongly conditioned by sociological factors such as poverty, youth labor, digital exclusion, and age-grade distortion, yet access to individual student records is restricted by legal and institutional barriers. This work proposes a predictive model of academic performance trends for students of the state public school network of Piauí, grounded in sociological factors and trained on a synthetic population representative of the profile of an average class of that network. Every generation parameter is traceable to official statistics and academic literature, and the generator incorporates anti-deterministic mechanisms: student financial assistance is derived from a legal eligibility rule, vulnerabilities interact non-linearly, and academic resilience profiles are preserved. Observable academic

¹ Graduando do Curso de Licenciatura em Computação. Instituto Federal de Educação, Ciência e Tecnologia do Piauí (IFPI), Campus Teresina Zona Sul.

² Professor orientador. Instituto Federal de Educação, Ciência e Tecnologia do Piauí (IFPI), Campus Teresina Zona Sul.

indicators correspond to first-semester partial records, and the target variable represents the end-of-year trend, labeled according to Brazilian educational legislation on approval criteria. Over three thousand records, three classification algorithms and two reference *baselines* were compared under stratified cross-validation, with oversampling confined to each training partition and paired statistical testing. XGBoost achieved 84.11% accuracy, statistically equivalent to *Random Forest* and superior to the single tree and to the *baselines*, with no errors between the extreme classes. After the partial academic signals, the most determinant attributes were age-grade distortion, financial assistance, youth labor, and per capita income, recovering the relative weight of the sociological factors in the generating phenomenon. It is concluded that the articulation between traceable synthetic data and ensemble classifiers constitutes a valid methodological alternative for early-warning systems in restricted-data contexts.

Keywords: academic performance; sociological factors; synthetic data; machine learning; public high school.

Data de aprovação: 11/08/2026

1 INTRODUÇÃO

O cenário atual da educação básica brasileira enfrenta desafios significativos relacionados à melhoria da qualidade do ensino e à redução dos índices de evasão, abandono e retenção (D'ABREU; MARTURANO, 2010; CAMPOS *et al.*, 2024). No ensino médio público, etapa final e mais crítica da educação básica, esses desafios assumem contornos estruturais: estima-se que apenas 74,3% dos jovens brasileiros concluem o ensino médio até os 19 anos de idade, patamar que cai para 69,5% entre jovens pretos, pardos e indígenas e para 70,5% nos domicílios de menor renda (Todos Pela Educação, 2023). Nesse contexto, a análise e a compreensão dos fatores que condicionam o sucesso ou o insucesso acadêmico dos estudantes tornam-se imprescindíveis para a implementação de estratégias pedagógicas e de políticas públicas eficazes.

A predição de desempenho acadêmico constitui um campo emergente na interseção entre as ciências da educação e da computação, consolidado na literatura como Mineração de Dados Educacionais (EDM) (BAKER; ISOTANI; CARVALHO, 2011; ROMERO; VENTURA, 2010). A possibilidade de identificar precocemente estudantes em trajetória de risco, a partir de dados demográficos, socioeconômicos e acadêmicos, representa uma oportunidade concreta para a gestão educacional baseada em evidências. A literatura sociológica, contudo, adverte que o desempenho escolar é um fenômeno multifatorial, fortemente condicionado pela origem social dos estudantes (BERLATO; MENDES; ANDRETTA, 2020; MARTURANO; PIZATO, 2015), o que exige que os modelos preditivos incorporem explicitamente os condicionantes sociológicos do rendimento (renda, trabalho juvenil, exclusão digital, distorção idade-série e políticas de permanência) sem, entretanto, reproduzir um determinismo social fatalista.

A dimensão dessas desigualdades é bem documentada pelas estatísticas oficiais, e assume contornos particularmente agudos no Piauí, objeto desta investigação. No estado, 45,3% da população geral encontrava-se em situação de pobreza em 2023 (Instituto Brasileiro de Geografia e Estatística, 2024b), e a literatura demonstra que a pobreza entre crianças e adolescentes é sistematicamente superior à da população geral: cerca de 60% dos estudantes brasileiros de 15 a 17 anos da rede pública vivem em domicílios com renda *per capita* de até meio salário mínimo (Fundação Abrinq, 2023). A exclusão digital piauiense é acentuada: embora 92,5% dos domicílios do estado disponham de acesso à internet, apenas 25,5% possuem microcomputador ou *tablet* (pouco mais da metade da média nacional de 40,9%) (Instituto Brasileiro de Geografia e Estatística, 2025; Instituto Brasileiro de Geografia e Estatística, 2024a). A distorção idade-série (proporção de alunos com dois ou mais anos de atraso escolar) attingia 21,7% na rede pública estadual do Piauí, após patamar histórico de 33,2% (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2025; Todos Pela Educação, s.d.); no mesmo período de consolidação das políticas de permanência, a taxa de abandono da rede recuou de 8,6% para 0,5% (Minuto MT, 2026; SANTOS; COSTA; COSTA, 2026). A Tabela 1 consolida esses indicadores, que fundamentam quantitativamente a relevância do problema investigado.

Delimitando-se o escopo desta investigação, o estudo concentra-se na elaboração de um modelo preditivo de **tendência de desempenho acadêmico** para estudantes do **ensino médio da rede pública estadual do Piauí**, fundamentado em fatores sociológicos. Em virtude das restrições de acesso a microdados individuais reais de estudantes (protegidos pela Lei Geral de Proteção de Dados e condicionados a autorizações institucionais e éticas cujos prazos de tramitação excedem o cronograma desta etapa da pesquisa), adotou-se uma estratégia metodologicamente validada pela literatura: a construção de uma **população sintética representativa do perfil de uma turma média** dessa rede, na qual cada distribuição é ancorada em estatísticas oficiais desagregadas para o Piauí sempre que disponíveis (IBGE, INEP, PNAD Contínua) e em estudos acadêmicos sobre os condicionantes sociais do desempenho escolar. Estudos recentes em EDM sustentam a validade preditiva de bases sintéticas de qualidade para o desenvolvimento e a validação de modelos de classificação (KOSTOPOULOS; TSIKAKI; KOTSIANTIS, 2026; SHABNAM; RAMACHANDRIAH; HALADAPPA, 2025), ressalvadas as diferenças entre técnicas de geração (discutidas no Referencial Teórico e retomadas nas limitações deste trabalho).

A problemática que orienta esta pesquisa emerge da constatação de que, apesar da crescente disponibilidade de dados educacionais e da consolidação das técnicas de aprendizado de máquina, há uma lacuna na construção de modelos preditivos que incorporem explicitamente, e com fundamentação estatística rastreável, os fatores sociológicos que condicionam o desempenho no ensino médio público. Questiona-

Tabela 1 – Indicadores educacionais e socioeconômicos que fundamentam o estudo.

Indicador	Valor	Nível	Fonte
Conclusão do ensino médio até os 19 anos	74,3%	Brasil	Todos Pela Educação (2023)
Matrículas do ensino médio na rede pública	82,0%	Brasil	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2025)
Estudantes de 15–17 anos da rede pública em situação de pobreza (renda <i>per capita</i> $\leq 0,5$ SM)	~60%	Brasil	Fundação Abrinq (2023)
População em situação de pobreza (2023)	45,3%	Piauí	Instituto Brasileiro de Geografia e Estatística (2024b)
Domicílios com acesso à internet	92,5%	Piauí	Instituto Brasileiro de Geografia e Estatística (2025)
Domicílios com microcomputador ou <i>tablet</i>	25,5%	Piauí	Instituto Brasileiro de Geografia e Estatística (2025)
Distorção idade-série (rede pública estadual)	21,7%	Piauí	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2025), Todos Pela Educação (s.d.)
Abandono no ensino médio (2022 → 2025)	8,6% → 0,5%	Piauí	Minuto MT (2026)
Estudantes do ensino médio de baixa renda que trabalham	26,3%	Brasil	Berlato, Mendes & Andretta (2020)

Fonte: Elaborado pelo autor (2026).

se, assim: *como construir um modelo de previsão de tendência de desempenho acadêmico para o ensino médio da rede pública estadual do Piauí, fundamentado em fatores sociológicos documentados pela literatura, e quais desses fatores se mostram mais determinantes na classificação dos estudantes?*

Dessa questão derivam quatro hipóteses de pesquisa, retomadas individualmente nas Considerações Finais:

H1. Classificadores de *ensemble* (*Random Forest* e *XGBoost*) superam a *decision tree* única e os *baselines* de referência (classificador majoritário e Regressão Logística) na predição da tendência de desempenho.

H2. Os fatores sociológicos parametrizados no fenômeno gerador são recuperáveis pelos modelos como atributos de importância não redundante, mesmo na presença de indicadores acadêmicos diretos.

- H3.** Sinais acadêmicos parciais do primeiro semestre são suficientes para prever a tendência do desfecho anual com desempenho substancialmente superior ao *baseline* majoritário.
- H4.** A regra legal de elegibilidade da assistência financeira estudantil produz, por si só, proporções coerentes de beneficiários entre cenários socioeconômicos contrastantes, sem parâmetro dedicado de calibração.

O presente trabalho propõe, portanto, a construção de um conjunto de dados sintéticos sociologicamente fundamentado e o desenvolvimento comparativo de modelos de classificação (*Decision Tree*, *Random Forest* e *XGBoost*) para prever a tendência de desempenho ao final do ano letivo (Aprovado, Em Risco ou Reprovado), a partir de sinais acadêmicos parciais do primeiro semestre e do contexto socioeconômico, identificando as variáveis mais determinantes e fornecendo subsídios para intervenções pedagógicas preventivas.

1.1. OBJETIVO GERAL

Desenvolver um modelo preditivo de tendência de desempenho acadêmico para estudantes do ensino médio da rede pública estadual do Piauí, com base em fatores sociológicos, utilizando um conjunto de dados sintéticos fundamentado em estatísticas oficiais e na literatura acadêmica.

1.2. OBJETIVOS ESPECÍFICOS

1. Construir e parametrizar um gerador de dados sintéticos que reproduza o perfil demográfico, socioeconômico e acadêmico de uma turma média do ensino médio da rede pública estadual do Piauí, com rastreabilidade documentada entre cada parâmetro de distribuição e sua fonte estatística ou acadêmica, incluindo o nível de desagregação de cada fonte.
2. Testar e avaliar comparativamente algoritmos de aprendizado de máquina (*Decision Tree*, *Random Forest* e *XGBoost*) frente a *baselines* de referência, para identificar aquele com melhor desempenho na predição da tendência de rendimento acadêmico, sob tratamento adequado de desbalanceamento de classes, prevenção de vazamento de dados e teste estatístico pareado entre os modelos.
3. Analisar e interpretar os resultados obtidos, verificando se os fatores sociológicos parametrizados no fenômeno gerador são recuperados pelos classificadores como atributos determinantes e discutindo suas implicações para intervenções pedagógicas e políticas de permanência, com salvaguardas explícitas contra vieses discriminatórios.

2 REFERENCIAL TEÓRICO

O referencial teórico está organizado de maneira a facilitar a compreensão dos conceitos que fundamentam a pesquisa. Inicialmente, discutem-se os fatores sociológicos que condicionam o desempenho escolar no ensino médio público brasileiro (fundamento da parametrização do conjunto de dados deste estudo) e o uso de dados sintéticos em Mineração de Dados Educacionais. Na sequência, apresenta-se a abordagem de descoberta de conhecimento em bancos de dados (KDD), seguida da aplicação da mineração de dados educacionais para a análise do desempenho acadêmico. Por fim, são detalhadas as técnicas de seleção de atributos, os métodos de validação de modelos, essenciais para a construção e a avaliação dos modelos preditivos, e os principais classificadores.

O desempenho acadêmico é um fenômeno multifatorial, influenciado por dimensões sociais, psicológicas, pedagógicas e individuais. Estudos como os apresentados por Marturano & Pizato (2015) destacam que variáveis socioeconômicas, como renda familiar, acesso a recursos educacionais e contexto geográfico, desempenham papel crítico na determinação do sucesso discente.

No âmbito psicológico, fatores como motivação, autoeficácia e saúde mental têm sido amplamente discutidos. Segundo D'Abreu & Marturano (2010), evidencia-se que a co-ocorrência de baixo desempenho escolar e problemas externalizantes sugere a influência de variáveis antecedentes, como condições adversas na família e baixo nível socioeconômico, corroborando a hipótese de que aspectos subjetivos devem ser considerados em análises preditivas. Além disso, características acadêmicas prévias são apontadas como preditores relevantes em modelos de inteligência artificial, conforme demonstrado em pesquisas com base em dados institucionais (CURI, 2006).

A aplicação de técnicas de mineração de dados e *machine learning* na educação tem ganhado destaque na literatura internacional e nacional. Algoritmos como *decision trees* (árvores de decisão), redes neurais e regressão logística são frequentemente utilizados para identificar padrões em grandes conjuntos de dados educacionais (BAKER; ISOTANI; CARVALHO, 2011). Esses modelos permitem não apenas prever resultados acadêmicos, mas também mapear grupos de risco, facilitando intervenções personalizadas.

2.1. FATORES SOCIOLÓGICOS DO DESEMPENHO ESCOLAR NO ENSINO MÉDIO PÚBLICO

A sociologia da educação demonstra, desde os estudos de reprodução social de Pierre Bourdieu, que o sistema escolar tende a reproduzir as condições sociais de origem dos estudantes, mais do que a transformá-las. Em investigação quantitativa

com 3.287 estudantes do ensino médio público brasileiro, Berlato, Mendes & Andretta (2020) constataram que as diferenças entre agrupamentos de jovens estavam mais vinculadas às reproduções familiares e de classe social do que ao papel exercido pela escola, e que 26,3% dos alunos de baixa renda precisavam conciliar trabalho e estudo. As desigualdades educacionais brasileiras, documentadas por Wenczenovicz & Santos (2023) e Palominio & Bitencourt (2025), expressam-se em barreiras estruturais de acesso, permanência e êxito escolar que incidem de forma desproporcional sobre estudantes pobres, pretos, pardos e residentes nas regiões Norte e Nordeste.

Entre os condicionantes mensuráveis do desempenho, destacam-se quatro eixos documentados pelas estatísticas oficiais e pela literatura. O primeiro é a **pobreza**: cerca de 60% dos estudantes de 15 a 17 anos da rede pública brasileira vivem em domicílios com renda *per capita* de até meio salário mínimo (Fundação Abrinq, 2023); no Piauí, 45,3% da população geral encontrava-se em situação de pobreza em 2023 (Instituto Brasileiro de Geografia e Estatística, 2024b), sendo que a pobreza infanto-juvenil é sistematicamente superior à da população geral (Fundação Abrinq, 2023). O segundo é a **exclusão digital**, que no Piauí assume caráter qualitativo: 92,5% dos domicílios dispõem de acesso à internet, mas apenas 25,5% possuem microcomputador ou *tablet* (contra 40,9% na média nacional), de modo que os acessos dependem majoritariamente de aparelhos celulares, comprometendo atividades que exigem proficiência digital (Instituto Brasileiro de Geografia e Estatística, 2025; Instituto Brasileiro de Geografia e Estatística, 2024a). O terceiro é a **distorção idade-série**, percentual de alunos com dois ou mais anos de atraso em relação à idade recomendada: associada ao gênero masculino, à matrícula em áreas de pobreza e à necessidade de geração de renda (LIMA; GOMES, 2013; CHIAPINOTO; LUNARDI, 2024), a distorção desloca a distribuição etária das turmas para a direita, com cauda longa que se estende até os 24 anos. No Piauí, o indicador recuou de 33,2% para 21,7% na rede estadual (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2025; Todos Pela Educação, s.d.). O quarto é o **trabalho juvenil**, que subtrai tempo de estudo e está diretamente associado ao abandono temporário e à retenção (BERLATO; MENDES; ANDRETTA, 2020; CAMPOS *et al.*, 2024); no Piauí, a taxa de trabalho infantil (5 a 17 anos) situava-se entre 6,8% e 8,6% no biênio 2023–2024 (Ministério do Trabalho e Emprego, 2024) (conceito adjacente, porém distinto, do estudante de ensino médio que trabalha).

Em contraposição a esses fatores de risco, as **políticas públicas de permanência** operam como mecanismo de reversão de tendência. O programa Pé-de-Meia, instituído pela Lei nº 14.818, de 2024, constitui incentivo financeiro-educacional na modalidade de poupança destinado a estudantes do ensino médio público de 14 a 24 anos, inscritos em famílias do Cadastro Único com renda *per capita* de até meio salário mínimo, condicionado à frequência mínima de 80% das aulas (Brasil, 2024; Ministério do Desenvolvimento e Assistência Social, Família e Combate à Fome, 2024). Análises iniciais do programa indicam

forte efeito de retenção sobre estudantes em vulnerabilidade (SANTOS; COSTA; COSTA, 2026): no Piauí, onde o programa atende 187.620 estudantes, a taxa de abandono da rede estadual recuou de 8,6% para 0,5% no período de consolidação das políticas de permanência (Todos Pela Educação, s.d.; Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2025; Minuto MT, 2026). Programas estaduais anteriores, como o Poupança Jovem Piauí, exibiram efeito análogo (PALOMINIO; BITENCOURT, 2025).

A literatura adverte, contudo, contra o **determinismo social** em modelos preditivos educacionais: vulnerabilidade severa não implica insucesso inevitável. O caso da Unidade Escolar Augustinho Brandão, em Cocal dos Alves (PI) (escola pública que atende comunidade de baixa renda e superou milhares de escolas públicas e privadas no Exame Nacional do Ensino Médio), ilustra que a sinergia entre engajamento discente e infraestrutura pedagógica pode romper as barreiras da origem social (Secretaria de Estado da Educação do Piauí, s.d.). Modelos preditivos responsáveis devem, portanto, capturar interações não lineares (como o efeito de reversão da assistência financeira) e tolerar perfis de resiliência acadêmica, sem penalizar fatalisticamente os estudantes vulneráveis, uma diretriz incorporada à construção do conjunto de dados deste estudo.

2.2. DADOS SINTÉTICOS EM MINERAÇÃO DE DADOS EDUCACIONAIS

O acesso a microdados educacionais individuais é crescentemente restringido por legislações de proteção de dados pessoais e por trâmites institucionais e éticos. Nesse cenário, a geração de dados sintéticos (registros artificiais que preservam as distribuições estatísticas e as relações entre variáveis da população de interesse) consolidou-se como alternativa metodológica válida em EDM (KOSTOPOULOS; TSIKMAKI; KOTSIANTIS, 2026). Em avaliação comparativa (*benchmarking*) de modelos generativos estatísticos e profundos aplicados a dados estudantis, Kostopoulos, Tsiakmaki & Kotsiantis (2026) demonstram que bases sintéticas de qualidade preservam a utilidade preditiva dos dados originais, com a vantagem de eliminar riscos de reidentificação de estudantes.

Resultados convergentes são reportados por Shabnam, Ramachandriah & Haladappa (2025), que compararam dados reais, sintéticos e híbridos na predição de desempenho de estudantes com algoritmos de regressão e *ensemble*: os dados sintéticos igualaram os reais em capacidade preditiva, e conjuntos híbridos alcançaram acurácia de até 87,76%. Técnicas baseadas em redes adversárias generativas condicionais e transformadores tabulares atingem entre 96% e 98% de equivalência preditiva em relação aos dados autênticos (KOSTOPOULOS; TSIKMAKI; KOTSIANTIS, 2026; SHABNAM; RAMACHANDRIAH; HALADAPPA, 2025), sustentando o uso de bases sintéticas para o desenvolvimento e a validação de modelos de classificação quando o acesso primário é vedado.

É necessário, contudo, distinguir duas famílias de dados sintéticos com estatutos

epistemológicos diferentes. Os estudos de equivalência preditiva citados referem-se a dados sintéticos **derivados de microdados reais**: modelos generativos (estatísticos ou profundos, como redes adversárias condicionais e transformadores tabulares) são ajustados sobre uma base autêntica e produzem réplicas que preservam suas distribuições conjuntas (KOSTOPOULOS; TSIKMAKI; KOTSIANTIS, 2026; SHABNAM; RAMACHANDRIAH; HALADAPPA, 2025). Este trabalho, por indisponibilidade de microdados, situa-se na segunda família: a **simulação baseada em regras**, na qual as distribuições marginais são parametrizadas a partir de estatísticas agregadas oficiais e as relações entre variáveis são especificadas pelo pesquisador com fundamento na literatura. Nessa família, a validade da base não é herdada de dados reais, mas depende integralmente da qualidade da parametrização e da auditabilidade das fontes, e as métricas obtidas medem a capacidade dos classificadores de recuperar um fenômeno cuja estrutura é, em parte, conhecida por construção.

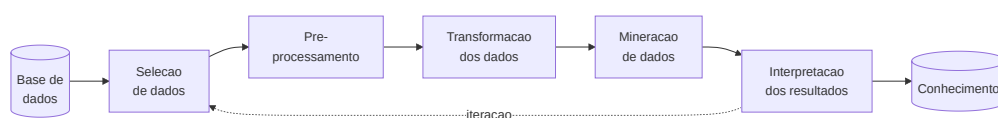
Dessa distinção decorrem os dois requisitos que orientam a metodologia deste trabalho: (i) a **rastreabilidade** entre cada parâmetro de geração e sua fonte estatística ou acadêmica (incluindo o nível de desagregação da fonte), de modo que as distribuições simuladas sejam auditáveis; e (ii) a **não trivialidade** do problema resultante, garantida pela injeção de variância estocástica e de casos-limite que impeçam os classificadores de memorizar a regra geradora.

2.3. DESCOBERTA DE CONHECIMENTO EM BANCOS DE DADOS (KDD)

Para lidar com o desafio da crescente quantidade e volume de dados disponíveis, Fayyad, Piatetsky-Shapiro & Smyth (1996) propuseram o processo de Descoberta de Conhecimento em Banco de Dados (KDD (*Knowledge Discovery in Databases*)). Esse processo visa identificar padrões explicáveis nos dados, permitindo interpretações e previsões para auxiliar na tomada de decisões (BAKER; ISOTANI; CARVALHO, 2011).

O processo de KDD é dinâmico e iterativo, podendo exigir retornos a etapas anteriores conforme novas descobertas são feitas (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). A Figura 1 apresenta a visão geral de suas etapas.

Figura 1 – Visão geral das etapas que compõem o processo de KDD.



Fonte: Adaptado de Fayyad, Piatetsky-Shapiro & Smyth (1996).

Ele é composto pelas seguintes etapas:

1. **Seleção de Dados:** Define-se a origem das informações e selecionam-se as características relevantes para a mineração de dados.
2. **Pré-processamento:** Garante a qualidade dos dados, identificando e corrigindo inconsistências, como valores ausentes, *outliers* e erros.
3. **Transformação dos Dados:** Converte os dados para formatos mais adequados à análise, incluindo discretização de variáveis contínuas e criação de novas variáveis.
4. **Mineração de Dados:** Utiliza algoritmos e técnicas estatísticas para identificar padrões e relações nos dados, sendo a etapa central do processo.
5. **Interpretação dos Resultados:** Analisa as descobertas para apoiar a tomada de decisões.

O KDD inicia-se com uma base de dados relacionada ao objetivo da análise e gera conhecimento aplicável à tomada de decisão. Sendo um processo cíclico, ele pode ser aprimorado continuamente com novas iterações (MAIMON; ROKACH, 2010).

2.4. MINERAÇÃO DE DADOS EDUCACIONAIS

A Mineração de Dados (*Data Mining*, DM) emerge como resposta fundamental às limitações da análise manual frente ao crescimento do volume de dados. Esta abordagem transcende as fronteiras de diversas áreas, incluindo estatística, aprendizado de máquina e visualização de dados, constituindo elemento essencial do processo de KDD (CASTRO; FERRARI, 2016).

No contexto educacional, a aplicação específica denomina-se Mineração de Dados Educacionais (*Educational Data Mining*, EDM). Esta vertente especializada explora informações sobre comportamentos, avaliações e conquistas dos alunos (COSTA *et al.*, 2012). A EDM estrutura-se conforme as necessidades do usuário final, seja ele aluno, educador, administrador ou pesquisador (ROMERO; VENTURA, 2010).

Peña-Ayala (2014) identificou diversos objetivos gerais associados ao processo de EDM, incluindo a modelagem de aluno, a análise do comportamento discente e a previsão do desempenho acadêmico, que visa antecipar notas finais e outros resultados de aprendizagem com base nos dados das atividades realizadas durante o curso.

Nessa linha, a literatura de EDM destaca o valor preditivo dos **sinais de engajamento contínuo**: além das notas de provas e da frequência, o desempenho e a regularidade nas atividades avaliativas ao longo do período capturam a dimensão comportamental do engajamento escolar (um construto multidimensional que antecede o desfecho acadêmico) (FREDRICKS; BLUMENFELD; PARIS, 2004). Costa *et al.* (2012) demonstram que a combinação de dados demográficos coletados no início do período

com métricas de engajamento pedagógico (notas de avaliações contínuas e frequência) provê predições com acurácia superior a 80% ainda em estágio inicial do curso (fundamento da inclusão, neste trabalho, de variáveis de desempenho e de entrega de atividades como sinais parciais do primeiro semestre).

2.5. TÉCNICAS DE SELEÇÃO DE ATRIBUTOS

A seleção de características desempenha papel fundamental na otimização de modelos preditivos. Conforme Liu & Motoda (2007), existem três categorias principais de técnicas: *Filter*, *Wrapper* e *Embedded*. Cada abordagem apresenta equilíbrio distinto entre eficiência computacional e qualidade dos resultados.

A abordagem *Embedded* integra a seleção de características ao processo de treinamento do modelo, constituindo parte fundamental de algoritmos como *Decision Tree* e *Random Forest*. A técnica *Filter* remove características sem utilidade, apresentando baixo custo computacional. A abordagem *Wrapper* permite experimentar subconjuntos de características em conjunto com o algoritmo específico, embora demande maior processamento.

2.6. MÉTODOS DE VALIDAÇÃO DE MODELOS

A validação cruzada (*Cross-Validation*) representa uma abordagem robusta para avaliação da capacidade de generalização de modelos preditivos. Esta técnica divide os dados em k subconjuntos, utilizando cada um como conjunto de teste enquanto os demais funcionam como treinamento. A configuração $k=10$ é frequentemente adotada (GÉRON, 2019).

Uma abordagem alternativa é o método *Holdout*, que consiste na divisão dos dados em conjuntos distintos de treinamento e teste, fornecendo uma estimativa do erro em relação a dados desconhecidos (GÉRON, 2019).

2.7. DESBALANCEAMENTO DE DADOS E ESTRATÉGIAS DE CORREÇÃO

O desbalanceamento de dados ocorre quando um conjunto apresenta quantidade significativamente maior de instâncias em uma classe comparativamente a outra. Para superar este desafio, técnicas de balanceamento são empregadas, destacando-se a subamostragem e a sobreamostragem (HE; GARCIA, 2009).

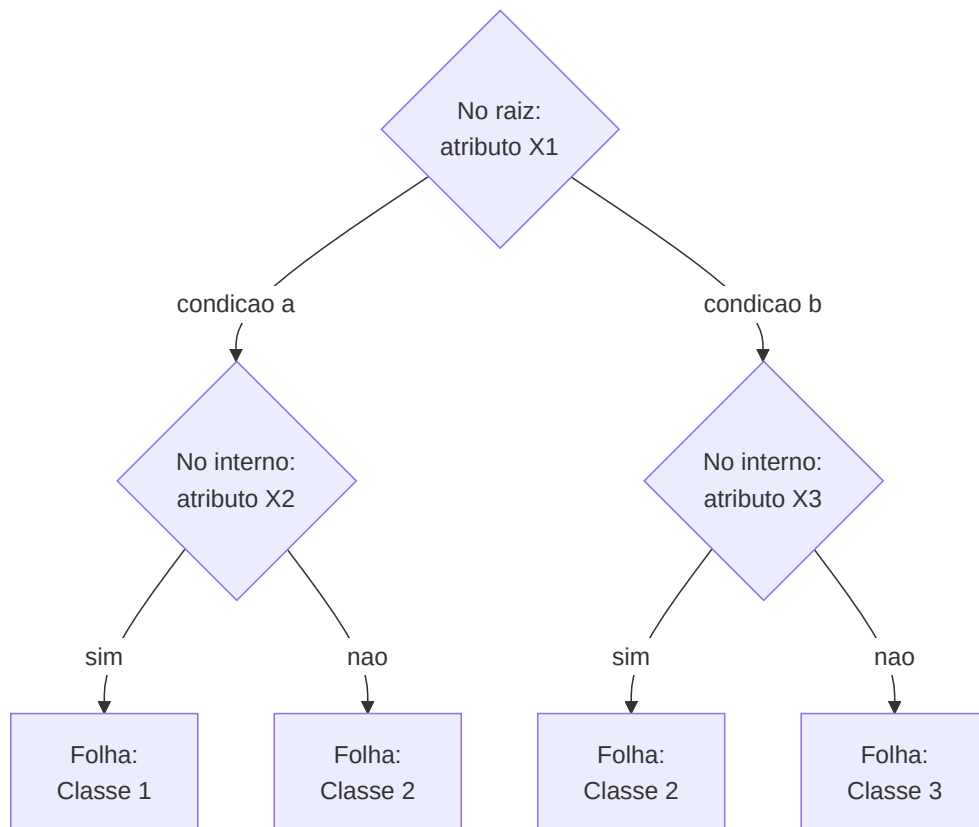
A sobreamostragem via SMOTE (*Synthetic Minority Oversampling Technique*) emprega abordagens estatísticas para gerar instâncias minoritárias sintéticas, enriquecendo a diversidade dos dados sem mera replicação (CHAWLA *et al.*, 2002). Contudo, a aplicação incorreta do SMOTE constitui uma das fontes mais recorrentes de vazamento de dados (*data leakage*) em estudos de aprendizado de máquina (MADUGULA; Esteve de

la Rosa; SHANKAR, 2026). O vazamento ocorre quando a sobreamostragem é aplicada sobre o conjunto de dados completo antes da divisão entre treino e teste, fazendo com que informações sintéticas geradas a partir de instâncias de teste contaminem o conjunto de treinamento. Para mitigar esse problema, a técnica deve ser confinada estritamente ao conjunto de treinamento dentro de cada partição da validação cruzada, utilizando-se *pipelines* que encapsulam o SMOTE como etapa interna ao *fold*, uma abordagem adotada neste trabalho por meio do `imblearn.Pipeline` (MADUGULA; Esteva de la Rosa; SHANKAR, 2026).

2.8. DECISION TREES

Decision Trees (árvores de decisão) constituem classificadores que, a partir de um vetor de atributos, definem por meio de trajetórias decisórias um resultado final (RUSSEL; NORVIG, 2013). Na aplicação voltada ao desempenho acadêmico, este modelo revela os fatores subjacentes que influenciam as previsões, permitindo interpretação clara dos processos decisórios (GRUS, 2016). A Figura 2 ilustra a estrutura genérica do classificador.

Figura 2 – Estrutura genérica de uma *decision tree*.



Fonte: Elaborado pelo autor (2026), com base em Russel & Norvig (2013).

Como se observa na Figura 2, a classificação parte de um nó raiz, no qual o atributo mais discriminante é avaliado, e percorre nós internos com condições sucessivas sobre os demais atributos, até alcançar um nó folha, que atribui a classe final à instância. Essa estrutura hierárquica de decisões é o que torna o modelo diretamente interpretável: cada predição corresponde a uma trajetória legível de regras.

2.9. RANDOM FOREST

A *Random Forest* (Floresta Aleatória) é uma abordagem robusta que agrega diversas *decision trees* para tarefas de classificação e regressão. Cada árvore é construída independentemente a partir de uma amostra aleatória do conjunto de dados (ARTERO, 2009).

No processo de treinamento, utiliza-se a técnica de *bagging* (*bootstrap aggregating*), que gera múltiplas amostras por meio de reamostragem com reposição. Em cada divisão de nó, seleciona-se aleatoriamente um subconjunto de atributos para avaliação, contribuindo para diminuir tanto o viés quanto a variância do modelo (GÉRON, 2019). A diversidade induzida pela aleatoriedade dupla (amostras *bootstrap* e subconjuntos de características) é o mecanismo central que confere à *Random Forest* resistência ao sobreajuste (*overfitting*) e capacidade de capturar interações complexas entre variáveis (DENISKO; HOFFMAN, 2018). A Figura 3 ilustra a estrutura do *ensemble*.

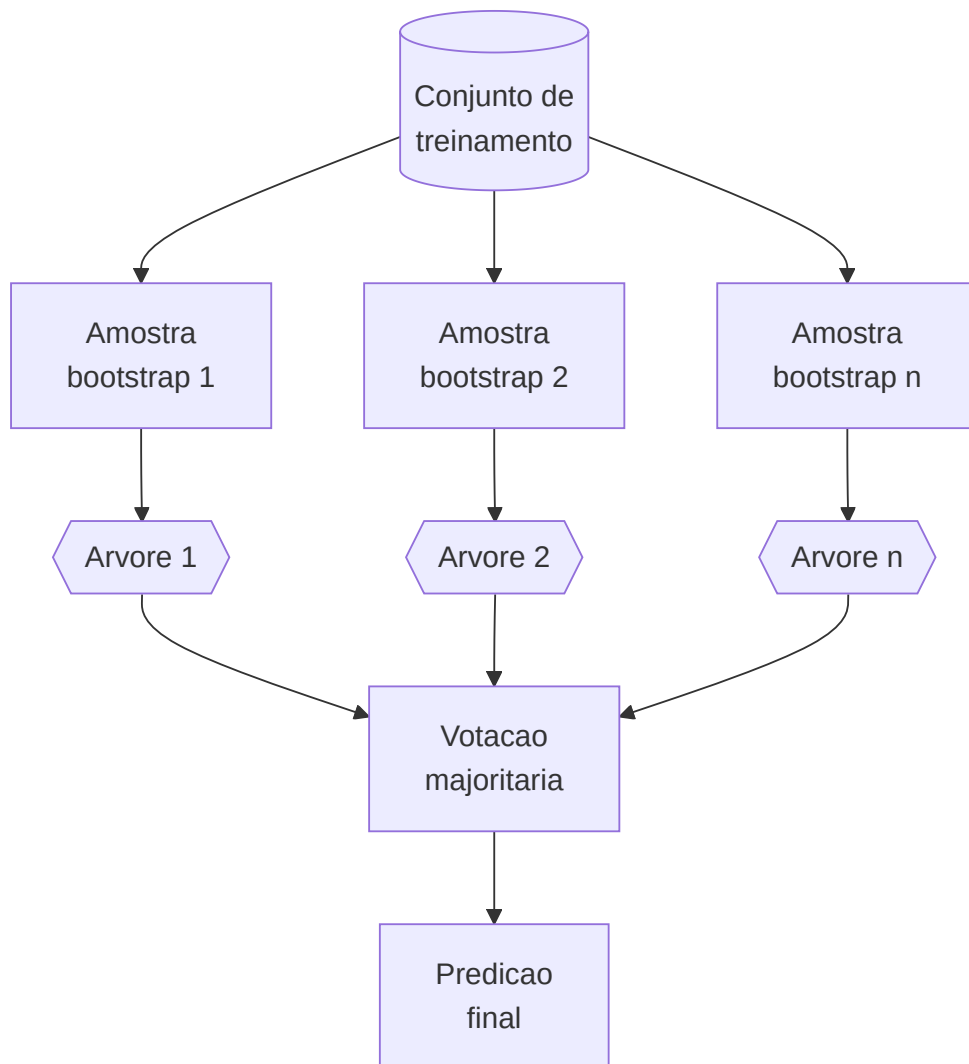
A Figura 3 evidencia o funcionamento do *ensemble*: a partir do conjunto de treinamento, são extraídas múltiplas amostras *bootstrap*, cada uma originando uma árvore treinada de forma independente; na predição, as saídas individuais das árvores são combinadas por votação majoritária, produzindo a classificação final com variância reduzida em relação a uma árvore única.

2.10. GRADIENT BOOSTING E XGBOOST

Enquanto a *Random Forest* baseia-se no paradigma de *bagging* (construindo árvores independentes em paralelo e combinando-as por votação), o *Gradient Boosting* adota uma estratégia sequencial, na qual cada nova árvore é treinada para corrigir os erros residuais das anteriores. Essa abordagem foi formalizada matematicamente por Friedman (2001), que demonstrou que o *boosting* pode ser interpretado como um algoritmo de otimização numérica no espaço de funções, utilizando o gradiente descendente para minimizar uma função de perda arbitrária.

O XGBoost (*eXtreme Gradient Boosting*) constitui uma implementação otimizada e regularizada do *Gradient Boosting*, proposta por Chen & Guestrin (2016). O algoritmo incorpora dois avanços fundamentais em relação à formulação original de Friedman. Primeiro, emprega uma aproximação de Taylor de segunda ordem da função de perda, permitindo capturar a curvatura do erro e direcionar a otimização com maior precisão:

Figura 3 – Estrutura genérica de uma *Random Forest*.



Fonte: Elaborado pelo autor (2026), com base em Géron (2019).

enquanto o *Gradient Boosting* clássico utiliza apenas a informação do gradiente de primeira ordem, o XGBoost incorpora também o Hessiano, isto é, a derivada de segunda ordem da função de perda em relação à predição. Segundo, introduz um termo de regularização explícito que penaliza a complexidade das árvores, controlando simultaneamente o número de folhas e a magnitude dos pesos, o que reduz o sobreajuste de forma mais eficaz que a mera limitação de profundidade (CHEN; GUESTRIN, 2016).

Adicionalmente, o XGBoost apresenta otimizações algorítmicas que o tornam particularmente eficiente em cenários com dados esparsos e valores ausentes, incluindo um algoritmo de divisão ponderada por quantis e tratamento nativo de registros incompletos (características relevantes para domínios educacionais, nos quais lacunas de informação são frequentes) (CHEN; GUESTRIN, 2016).

2.11. MÉTRICAS DE AVALIAÇÃO PARA CLASSIFICAÇÃO MULTICLASSE

A avaliação do desempenho de modelos preditivos em cenários com classes desbalanceadas exige métricas que vão além da acurácia simples. Em problemas nos quais uma ou mais classes são sub-representadas (como é o caso da classe de alunos *Reprovados* frente aos *Aprovados*), a acurácia tende a favorecer a classe majoritária, mascarando deficiências críticas na identificação dos grupos minoritários (HE; GARCIA, 2009).

Diante desse desafio, emprega-se o *F1-Score*, métrica que combina precisão e *recall* por meio da média harmônica, penalizando modelos que sacrificam uma das dimensões em detrimento da outra. Para cenários multiclasse, adota-se a variante ponderada (*weighted F1-Score*), que calcula o *F1* de cada classe individualmente e agrega os resultados ponderando-os pela proporção de instâncias reais de cada classe no conjunto de dados. Essa variante é particularmente adequada quando as frequências das classes refletem a distribuição natural do fenômeno estudado (como ocorre com o desempenho acadêmico), pois atribui maior peso às classes mais populosas sem ignorar o desempenho nas minoritárias.

Complementarmente, a análise por meio da Curva ROC (*Receiver Operating Characteristic*) e da Área sob a Curva (AUC) constitui uma abordagem consolidada para avaliação de classificadores (FAWCETT, 2006). Em problemas binários, a curva ROC plota a taxa de verdadeiros positivos em função da taxa de falsos positivos para diferentes limiares de decisão, e a AUC sintetiza essa informação em um escalar no intervalo $[0, 1]$. Para o contexto multiclasse, Hand & Till (2001) propuseram uma generalização do AUC por meio da estratégia *One-vs-Rest* (OvR), na qual cada classe é tratada como positiva contra a união das demais, produzindo uma curva ROC por classe. A média ponderada das AUCs individuais fornece uma medida global da capacidade discriminativa do modelo, mantendo a interpretabilidade do caso binário (HAND; TILL, 2001).

2.12. BASELINES DE REFERÊNCIA E COMPARAÇÃO ESTATÍSTICA DE CLASSIFICADORES

A magnitude de uma métrica, isoladamente, não informa se um modelo é bom: um classificador com 80% de acurácia pode ser trivial se a classe majoritária representar 80% das instâncias. Por isso, boas práticas de avaliação recomendam a inclusão de **baselines de referência**: um classificador trivial que sempre prediz a classe majoritária, estabelecendo o piso de desempenho do problema, e um modelo linear simples (como a Regressão Logística), que quantifica a parcela do fenômeno explicável por relações lineares. A diferença entre os modelos de interesse e esses *baselines* mede a contribuição efetiva da capacidade não linear dos algoritmos (GÉRON, 2019).

Adicionalmente, diferenças numéricas entre classificadores podem decorrer da variabilidade amostral das partições de avaliação. Para verificar se a diferença observada entre dois modelos é estatisticamente significativa, Demšar (2006) recomenda o **teste pareado de postos sinalizados de Wilcoxon** sobre os resultados das mesmas partições de validação cruzada: por ser não paramétrico, o teste não pressupõe normalidade das diferenças, sendo mais robusto que o teste t pareado para comparação de classificadores. Quando o valor- p excede o nível de significância adotado, a diferença entre os modelos não pode ser distinguida da flutuação amostral, e conclusões de superioridade devem ser evitadas ou moderadas.

3 TRABALHOS CORRELATOS

A previsão do desempenho escolar e o combate à evasão utilizando algoritmos de aprendizado de máquina têm sido amplamente investigados na literatura nacional. Para contextualizar a presente proposta e identificar o estado da arte na área de Mineração de Dados Educacionais (EDM), foram selecionados trabalhos de referência cujas características estruturais de modelagem tangenciam as abordagens deste estudo.

Silva *et al.* (2019) avaliaram o desempenho de estudantes universitários utilizando regressão logística e *decision trees*. Os autores destacaram que variáveis socioeconômicas (como renda e nível de escolaridade dos pais) operaram como fortes determinantes de retenção nos períodos acadêmicos iniciais.

Por sua vez, Costa *et al.* (2012) propuseram um sistema preditivo voltado à detecção de alunos em risco de reprovação em disciplinas introdutórias de programação por meio de algoritmos de classificação Bayesiana e *decision trees*. Constatou-se que o uso de dados demográficos no início do semestre combinado a dados de engajamento pedagógico provê predições com acurácia superior a 80%.

Nessa linha de análise, Baker, Isotani & Carvalho (2011) e Romero & Ventura (2010) discutiram a importância da integração de múltiplas fontes de dados (registros acadêmicos e bases contextuais) e o emprego de técnicas para contornar problemas de desbalanceamento de classes, como o SMOTE. Argumenta-se que a correção do desbalanceamento propicia uma sensibilidade substancialmente maior no mapeamento de alunos em risco de evasão.

A Tabela 2 resume as similaridades e contrastes metodológicos entre as principais iniciativas da literatura e a presente pesquisa.

A Tabela 2 evidencia que, embora os trabalhos correlatos compartilhem algoritmos e famílias de atributos com esta pesquisa, nenhum deles combina a construção fundamentada do próprio conjunto de dados com a predição de tendência de desempenho no ensino médio público estadual, lacuna que o presente estudo busca preencher.

Diferentemente dos trabalhos correlatos apresentados (majoritariamente voltados à predição de evasão ou reprovação em disciplinas do ensino superior, a partir de bases institucionais preexistentes), o presente estudo une, em uma única investigação, a

Tabela 2 – Comparação metodológica entre trabalhos correlatos da literatura e o presente estudo.

Trabalho / Autor	Algoritmos Usados	Principais Atributos	Relação com este Trabalho
Silva <i>et al.</i> (2019)	Regressão Logística, <i>Decision Tree</i>	Sociodemográficos (renda, escolaridade familiar, cota)	Compartilha o eixo socioeconômico, aqui expandido com conectividade, trabalho e assistência financeira.
Costa <i>et al.</i> (2012)	<i>Naive Bayes</i> , <i>Decision Tree</i>	Notas de avaliações contínuas, frequência, dados demográficos	Adota a formulação de métricas de engajamento e a classificação multiclasse do rendimento discente.
Peña-Ayala (2014)	Redes Neurais Alternativas, SVM, <i>Random Forest</i>	Histórico escolar prévio, taxas de evasão municipais	Utiliza <i>Random Forest</i> como referencial robusto, integrando-a com técnicas avançadas de SMOTE.
Este Estudo (rede pública estadual do PI)	<i>Decision Tree</i> , <i>Random Forest</i> , XGBoost + <i>baselines</i>	Fatores sociológicos (renda <i>per capita</i> , trabalho, exclusão digital, distorção idade-série, Pé-de-Meia) + sinais acadêmicos parciais do 1º semestre (nota, frequência, atividades)	Constrói o próprio conjunto de dados sintéticos com rastreabilidade parâmetro-fonte; SMOTE dentro do CV k=10; alvo de tendência de desempenho anual (critérios da LDB), não evasão.

Fonte: Elaborado pelo autor (2026).

construção do conjunto de dados e a modelagem preditiva: o conjunto de dados sintéticos é gerado com parametrização sociologicamente fundamentada e rastreável às fontes oficiais, e a variável-alvo é a **tendência de desempenho acadêmico** (e não a evasão) de estudantes do ensino médio da rede pública estadual do Piauí. Essa combinação permite avaliar de forma controlada o peso relativo dos fatores sociológicos na classificação, algo inviável em bases institucionais nas quais tais variáveis são incompletas ou inacessíveis.

4 METODOLOGIA

A metodologia deste trabalho fundamenta-se no processo de Descoberta de Conhecimento em Bancos de Dados (KDD), estruturado para desenvolver um modelo preditivo de tendência de desempenho acadêmico de estudantes do ensino médio da

rede pública estadual do Piauí, com base em fatores sociológicos. Como o acesso a microdados individuais reais é vedado por restrições legais e institucionais, a fonte de dados do estudo é um conjunto sintético construído pelo próprio trabalho (uma população representativa do perfil de uma turma média dessa rede), no qual cada distribuição é parametrizada a partir de estatísticas oficiais desagregadas para o Piauí sempre que disponíveis (IBGE, INEP, PNAD Contínua) e da literatura acadêmica apresentada no Referencial Teórico. Trata-se de simulação baseada em regras, nos termos da distinção estabelecida no Referencial: a validade da base depende da qualidade da parametrização e da auditabilidade das fontes, requisitos operacionalizados pela matriz de rastreabilidade apresentada adiante.

4.1. CLASSIFICAÇÃO DA PESQUISA

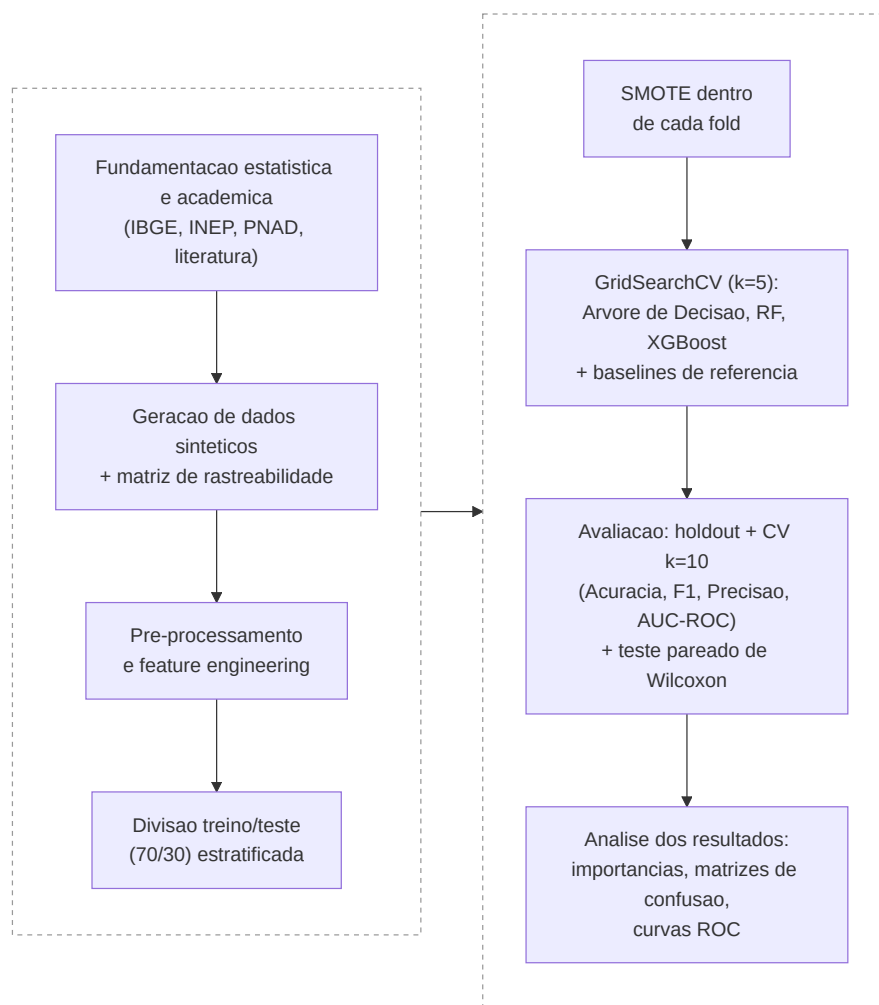
Seguindo os referenciais metodológicos de Gil (2017), esta pesquisa classifica-se, quanto à natureza, como aplicada; quanto à abordagem, como quantitativa; e, quanto aos objetivos, como exploratória e descritiva. Quanto aos procedimentos técnicos, combina o levantamento documental (utilizado na fundamentação estatística dos parâmetros de geração, a partir de relatórios oficiais e estudos acadêmicos publicados) com a pesquisa experimental, materializada na variação controlada de parâmetros do gerador de dados e na comparação sistemática de algoritmos de classificação sob condições idênticas de treinamento e avaliação.

4.2. PROCEDIMENTOS TÉCNICOS

Os procedimentos técnicos organizam-se em etapas sequenciais e reproduzíveis, executadas sob semente aleatória fixa e parâmetros documentados: geração dos dados sintéticos (com exportação da matriz de rastreabilidade parâmetro–fonte), pré-processamento, treinamento e otimização dos três classificadores e dos *baselines* de referência, avaliação por *holdout* e validação cruzada com teste estatístico pareado, e análise dos resultados. A Figura 4 apresenta a visão geral das etapas.

A Figura 4 organiza o fluxo metodológico em dois blocos. O primeiro compreende a preparação dos dados: a fundamentação estatística e acadêmica dos parâmetros, a geração do conjunto sintético acompanhada da matriz de rastreabilidade, o pré-processamento com engenharia de atributos e a divisão estratificada entre treinamento e teste. O segundo bloco compreende a modelagem e a avaliação: o balanceamento por SMOTE confinado a cada *fold*, o ajuste de hiperparâmetros dos três classificadores e dos *baselines* via *GridSearchCV*, a avaliação por *holdout* e validação cruzada com teste pareado de Wilcoxon e, por fim, a análise dos resultados por importâncias, matrizes de confusão e curvas ROC. As seções seguintes detalham cada uma dessas etapas.

Figura 4 – Etapas metodológicas do estudo.



Fonte: Elaborado pelo autor (2026).

4.3. CONSTRUÇÃO E PARAMETRIZAÇÃO DOS DADOS SINTÉTICOS

A primeira etapa da pesquisa consiste na construção do conjunto de dados. O gerador sintético produz 3.000 registros de estudantes do ensino médio da rede pública estadual do Piauí, cada um descrito por atributos demográficos (idade, genero, cor_raca, estado_civil), socioeconômicos (renda_per_capita, trabalha, acesso_internet, possui_computador), institucionais (serie, tipo_escola_origem, distorcao_idade_serie, assistencia_pe_de_meia) e acadêmicos (nota_media, frequencia, media_atividades, taxa_entrega_atividades, reprovacoes), além da variável-alvo situacao. O dicionário completo dos atributos, com tipos e descrições, é apresentado na Tabela 9 do Apêndice A.

4.3.1. Semântica Temporal dos Atributos

Os atributos acadêmicos observáveis representam **parciais do primeiro semestre letivo**: *nota_media* e *frequencia* correspondem ao acumulado até o fechamento do semestre, e as variáveis de engajamento *media_atividades* (média das notas de atividades, normalizada para a escala 0–10) e *taxa_entrega_atividades* (proporção de atividades entregues) capturam a regularidade do engajamento contínuo (uma decisão de modelagem fundamentada na literatura de EDM sobre sinais de engajamento discutida no Referencial Teórico (COSTA *et al.*, 2012; FREDRICKS; BLUMENFELD; PARIS, 2004)). A variável-alvo, por sua vez, refere-se à **situação projetada ao final do ano letivo**. Essa separação temporal entre preditores (primeiro semestre) e desfecho (fim do ano) confere sentido prático ao modelo: a predição ocorre em momento no qual a intervenção pedagógica preventiva ainda é possível.

4.3.2. Rastreabilidade Parâmetro–Fonte

Cada parâmetro de distribuição do gerador é ancorado em uma estatística oficial ou em um achado da literatura, e o gerador exporta automaticamente, junto ao arquivo de dados, uma matriz de rastreabilidade em formato legível e processável, registrando também o **nível de desagregação** de cada fonte (estadual, nacional, normativo ou decisão de modelagem). Quando o valor piauiense não existe para o conceito exato do parâmetro, adota-se um *proxy* (valor de conceito adjacente ou de agregação diferente) com nota explícita. A Tabela 3 consolida essa matriz.

Tabela 3 – Matriz de rastreabilidade entre parâmetros do gerador sintético e suas fontes.

Parâmetro	Valor	Evidência	Fonte	Desagregação
Origem em escola pública	85%	PI: 715.748 estudantes na rede pública (2025); BR: 82,0% das matrículas do EM	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2025), Todos Pela Educação (s.d.)	Estadual (PI)

(continua na próxima página)

Tabela 3 – Matriz de rastreabilidade entre parâmetros do gerador sintético e suas fontes. (continuação)

Parâmetro	Valor	Evidência	Fonte	Desagregação
Renda <i>per capita</i> até 0,5 SM	60%	PI: 45,3% da população geral em pobreza (2023); pobreza de 15–17 anos é superior à geral (~60% no BR)	Instituto Brasileiro de Geografia e Estatística (2024b), Fundação Abrinq (2023)	Brasileiro de Geografia e Estatística (PI) (proxy: população geral)
Gênero / pretos e pardos	51% / 77%	PI (Censo 2022): 64,8% pardos, 12,3% pretos; razão de sexo 95,81 homens por 100 mulheres	Instituto Brasileiro de Geografia e Estatística (2023)	Brasileiro de Geografia e Estatística (PI)
Estudantes trabalhadores	26,3%	Alunos de EM de baixa renda que conciliam trabalho e estudo; PI 5–17 anos: 6,8–8,6% em trabalho infantil (conceito adjacente)	Berlato, Mendes & Andretta (2020), Ministério do Trabalho e Emprego (2024)	Decisão de modelagem (proxy)
Acesso doméstico à internet	92,5%	PI: domicílios com utilização de internet (2025)	Instituto Brasileiro de Geografia e Estatística (2025)	Brasileiro de Geografia e Estatística (PI)
Posse de computador	25,5%	PI: domicílios com microcomputador ou <i>tablet</i> (2025); BR: 40,9%	Instituto Brasileiro de Geografia e Estatística (2025), Instituto Brasileiro de Geografia e Estatística (2024a)	Brasileiro de Geografia e Estatística (PI)

(continua na próxima página)

Tabela 3 – Matriz de rastreabilidade entre parâmetros do gerador sintético e suas fontes. (continuação)

Parâmetro	Valor	Evidência	Fonte	Desagregação
Distorção idade-série	21,7%	Rede estadual do PI (histórico: 33,2%); cauda etária até 24 anos	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2025), Todos Pela Educação (s.d.)	Estadual (PI)
Pé-de-Meia (regra)	-	Elegibilidade: 14–24 anos, renda <i>per capita</i> $\leq 0,5$ SM, frequência $\geq 80\%$; 187.620 atendidos no PI	Brasil (2024), Ministério do Desenvolvimento e Assistência Social, Família e Combate à Fome (2024), Minuto MT (2026)	Normativo (nacional), efeito documentado no PI
Limiares de aprovação	6,0 / 75%	Média final mínima e frequência mínima de 75% da carga horária	Brasil (1996)	Normativo (nacional)
<i>Outliers</i> de resiliência	3%	Perfis vulneráveis de alto desempenho (ex.: Cocal dos Alves–PI)	Secretaria de Estado da Educação do Piauí (s.d.)	Decisão de modelagem, com âncora estadual

Fonte: Elaborado pelo autor (2026).

Parâmetros sem estatística oficial correspondente em qualquer nível de desagregação (como a variância idiossincrática do segundo semestre, as bandas de fronteira da classe intermediária e as variáveis de engajamento) são registrados na matriz como **decisões de modelagem**, com a justificativa correspondente, e discutidos nas subseções seguintes.

4.3.3. Regras de Geração Não Lineares

Em atenção à diretriz anti-determinista discutida no Referencial Teórico, três mecanismos afastam o gerador de uma soma linear de vulnerabilidades:

1. **Assistência financeira como regra de elegibilidade:** o benefício do Pé-de-Meia não é sorteado, mas derivado das regras da Lei nº 14.818 (idade entre 14 e 24 anos e renda *per capita* de até meio salário mínimo, atuando como *proxy* do Cadastro Único), com taxa de adesão de 85% entre os elegíveis. O incentivo eleva a frequência dos aderentes em direção à exigência mínima de 80% mensais (efeito de retenção), e o beneficiário efetivo é aquele que mantém a frequência exigida. Beneficiários recebem bônus no score latente, revertendo a tendência de queda do perfil vulnerável (SANTOS; COSTA; COSTA, 2026).
2. **Interações entre fatores:** o estudante-trabalhador em distorção idade-série sofre penalidade adicional (dupla vulnerabilidade), enquanto variáveis de conectividade são geradas condicionalmente à faixa de renda, preservando o gradiente observado na PNAD Contínua.
3. **Outliers de resiliência acadêmica:** 3% dos perfis de alta vulnerabilidade (renda mínima e sem computador) recebem desempenho acadêmico elevado (nota 8,5–9,7; frequência 90–100%), inspirados em casos reais como o da U.E. Augustinho Brandão (Secretaria de Estado da Educação do Piauí, s.d.), impedindo que os classificadores aprendam um determinismo social absoluto.

Adicionalmente, gênero e cor/raça são gerados apenas para caracterização demográfica da turma, com **peso zero** no fenômeno gerador, uma decisão de salvaguarda contra vieses discriminatórios, complementada pela exclusão desses atributos do treinamento dos modelos.

4.3.4. Definição da Variável-Alvo

A variável-alvo *situacao* representa a **tendência de desempenho acadêmico** do estudante ao final do ano letivo, em três classes ordinais: *Aprovado* (trajetória consolidada), *Em Risco* (zona de atenção pedagógica) e *Reprovado* (vulnerabilidade acadêmica severa). A unidade de análise é o estudante individual de uma turma do ensino médio da rede pública estadual do Piauí, e o horizonte de predição é o fechamento do ano letivo, a partir dos sinais parciais do primeiro semestre (Seção 4.3.1).

A rotulagem opera em duas etapas. Na primeira, projeta-se o **desfecho anual** de cada estudante: à média e à frequência parciais do primeiro semestre aplica-se uma deriva de segundo semestre, composta por um componente sistemático (função do contexto socioeconômico (renda, conectividade, trabalho, distorção idade-série), do efeito de retenção do Pé-de-Meia e das interações não lineares entre vulnerabilidades) e por um componente idiossincrático: um ruído gaussiano de média zero cujo desvio-padrão é controlado pelo parâmetro `NOISE_LEVEL`. Esse componente representa, fenomenologicamente, os fatores não observáveis que alteram trajetórias escolares

entre semestres (saúde, eventos familiares, motivação), e sua magnitude é uma decisão de modelagem registrada na matriz de rastreabilidade (o capítulo de Experimentos e Discussão dos Resultados apresenta a análise de sensibilidade que caracteriza o impacto dessa escolha).

Na segunda etapa, a projeção anual é discretizada pelos **critérios reais de aprovação do ensino médio brasileiro**: média final mínima de 6,0 (padrão das redes estaduais) e frequência mínima de 75% da carga horária, exigência do art. 24 da Lei de Diretrizes e Bases da Educação Nacional (Brasil, 1996). O estudante cuja projeção atende a ambos os critérios com folga é rotulado *Aprovado*; aquele que descumpre algum critério de forma clara, *Reprovado*; e a classe *Em Risco* é atribuída à banda pedagógica de fronteira (projeção de média entre 5,0 e 6,5 ou frequência entre 75% e 80%), uma decisão de modelagem que operacionaliza a zona de atenção na qual a intervenção preventiva é mais efetiva. As proporções das três classes **emergem** dessa regra, não sendo calibradas por limiares arbitrários de distribuição.

Ressalta-se que os modelos de classificação **não têm acesso à projeção anual nem à deriva**: aprendem exclusivamente a partir dos atributos observáveis do primeiro semestre (Tabela 9), de modo que o componente idiossincrático estabelece um teto teórico de acurácia (erro de Bayes controlado) que impede a memorização da regra geradora.

4.3.5. Cenários de Contraste

Para avaliar a robustez do gerador e dos modelos frente a diferentes realidades socioeconômicas, foram configurados, além do cenário base (rede pública estadual do Piauí), três perfis contrastantes em arquivos de configuração próprios: **média nacional** (parâmetros agregados do ensino médio público brasileiro), **alta vulnerabilidade** (75% dos estudantes com renda *per capita* de até 0,5 SM, distorção idade-série de 33,2%, conectividade reduzida) e **rede privada de renda alta** (65% acima de 1 SM *per capita*, distorção de 5%, acesso pleno à internet). Nos cenários de contraste, a proporção de beneficiários do Pé-de-Meia decorre exclusivamente da regra de elegibilidade (no perfil de renda alta, apenas ~6,5% dos estudantes tornam-se beneficiários, sem qualquer parâmetro dedicado, o que evidencia a coerência interna da regra, em linha com a hipótese H4).

4.4. PRÉ-PROCESSAMENTO DOS DADOS

O pré-processamento visa garantir a qualidade e a adequação dos dados aos algoritmos. A estratégia distingue variáveis nominais de binárias:

1. Codificação de variáveis categóricas: *One-Hot Encoding* para nominais (*estado_civil*, *renda_per_capita*, *cor_raca*) e *Label Encoding* para binárias

(Sim/Não → 1/0)

2. Remoção de atributos protegidos e de baixa importância: `genero`, `cor_raca` e `tipo_escola_origem` são excluídos do treinamento (os dois primeiros como salvaguarda contra vieses discriminatórios, o terceiro por baixa contribuição preditiva observada)
3. Verificação de valores ausentes e *outliers* (por construção, o gerador não produz valores ausentes; os *outliers* de resiliência são preservados intencionalmente)

A exclusão de gênero e cor/raça do treinamento é decisão de desenho ético, e não o resultado de um *ranking* de importância: ambos os atributos entram no gerador com peso zero, de modo que não há sinal racial ou de gênero a reconstituir a partir dos demais preditores. Já `tipo_escola_origem` foi avaliada em execuções exploratórias restritas ao conjunto de treino (na mesma lógica do estudo de ablação das variáveis derivadas) e removida por baixa contribuição observada: 85% da população sintética origina-se de escola pública, o que reduz a variância preditiva da variável.

4.5. TRANSFORMAÇÃO DOS DADOS (*FEATURE ENGINEERING*)

Criação de variáveis derivadas que capturam relações não lineares entre os atributos originais:

1. Interação nota-frequência (`nota_x_frequencia`): produto entre nota média e frequência parciais
2. Reprovações por série (`reprov_por_serie`): razão entre reprovações acumuladas e série atual

A seleção do conjunto final de variáveis derivadas foi validada por um **estudo de ablação**, no qual o desempenho dos modelos é medido com e sem cada variável candidata. Uma terceira variável derivada (um escore acadêmico composto ponderando nota, frequência e reprovações) foi avaliada e descartada, por não agregar desempenho ao conjunto final; o experimento e sua discussão são apresentados no capítulo de Experimentos e Discussão dos Resultados.

4.6. BALANCEAMENTO DO CONJUNTO DE DADOS

A técnica SMOTE (*Synthetic Minority Over-sampling Technique*) é aplicada para equilibrar a representação das classes minoritárias. O SMOTE é aplicado exclusivamente dentro de cada *fold* de treino (via `imblearn.Pipeline`), garantindo a prevenção de *data leakage*.

4.7. APLICAÇÃO DOS ALGORITMOS DE CLASSIFICAÇÃO

São implementados e comparados três algoritmos:

1. **Decision Tree** (`DecisionTreeClassifier`): modelo interpretável baseado em regras de divisão hierárquica
2. **Random Forest** (`RandomForestClassifier`): *ensemble* de múltiplas árvores com *bagging*
3. **Gradient Boosting** (`XGBClassifier`): *ensemble* sequencial com *boosting*

Os algoritmos são implementados com *scikit-learn* e *XGBoost* em Python. O ajuste de hiperparâmetros é realizado via *GridSearchCV* com validação cruzada estratificada de 5 *folds*, utilizando *F1-Score* ponderado como métrica de otimização.

Complementarmente, conforme a fundamentação apresentada no Referencial Teórico, dois **baselines de referência** são avaliados sob o mesmo protocolo: um classificador trivial de classe majoritária (`DummyClassifier`), que estabelece o piso de desempenho do problema, e uma **Regressão Logística** multinomial, que quantifica a parcela do fenômeno explicável por relações lineares.

4.8. VALIDAÇÃO E AVALIAÇÃO DOS MODELOS

A avaliação considera múltiplas métricas:

1. Acurácia: proporção de previsões corretas
2. Precisão, *Recall* e *F1-Score* ponderados: métricas por classe agregadas
3. Área sob a curva ROC (AUC-ROC) multiclasse *One-vs-Rest*
4. Matriz de confusão

Além do conjunto de teste isolado (*holdout*, 30% dos dados com estratificação), utiliza-se validação cruzada *k-fold* estratificada ($k=10$) para garantir robustez estatística. As diferenças de desempenho entre pares de modelos são submetidas ao **teste pareado de postos sinalizados de Wilcoxon** sobre os *F1-Scores* das mesmas partições da validação cruzada, ao nível de significância de 5%, conforme a recomendação de Demšar (2006): diferenças cujo valor-*p* exceda esse nível são tratadas como estatisticamente não distinguíveis da flutuação amostral.

Complementarmente, realiza-se uma **análise de sensibilidade ao ruído idiossincrático**: o parâmetro `NOISE_LEVEL` do gerador é variado em uma grade de valores e, para cada valor, o desempenho de um classificador de referência é medido, permitindo

caracterizar como a previsibilidade do desfecho anual degrada conforme cresce a variância não observável do fenômeno (análise apresentada no capítulo de Experimentos e Discussão dos Resultados). Ressalta-se que a varredura não busca uma acurácia-alvo: ela caracteriza o problema, e o valor adotado no estudo é uma decisão fenomenológica registrada na matriz de rastreabilidade.

4.9. ANÁLISE E INTERPRETAÇÃO DOS RESULTADOS

A interpretação dos resultados é conduzida pelo pesquisador em quatro frentes complementares: (i) comparação das métricas agregadas (acurácia, *F1-Score*, precisão, *recall* e AUC-ROC) entre os classificadores e os *baselines*, nos ambientes de *holdout* e validação cruzada, com verificação estatística das diferenças; (ii) análise do desempenho por classe, por meio das curvas ROC *One-vs-Rest* e das matrizes de confusão, com atenção à classe crítica *Reprovado* e à estrutura ordinal dos erros; (iii) análise da importância dos atributos em cada modelo, confrontando os pesos recuperados pelos classificadores com os fatores sociológicos parametrizados no fenômeno gerador; e (iv) verificação da coerência do gerador nos cenários de contraste. Todos os artefatos quantitativos da análise (tabelas de métricas, figuras comparativas e matriz de rastreabilidade) são produzidos de forma automatizada e reproduzível, com semente aleatória fixa.

4.10. ASPECTOS ÉTICOS

Esta pesquisa não envolve seres humanos nem dados pessoais: todos os registros são sintéticos, gerados por simulação parametrizada a partir de estatísticas agregadas públicas, o que dispensa a submissão a Comitê de Ética em Pesquisa e afasta a incidência da Lei Geral de Proteção de Dados. Ainda assim, o desenho experimental adota salvaguardas de equidade algorítmica: gênero e cor/raça são gerados apenas para caracterização demográfica, com peso zero no fenômeno gerador, e excluídos do treinamento dos modelos. Ressalta-se, por fim, a diretriz de uso responsável: em eventual aplicação real, atributos socioeconômicos sensíveis devem operar como critérios de **priorização de apoio** ao estudante, jamais como instrumentos de exclusão, rotulação ou penalização.

5 EXPERIMENTOS E DISCUSSÃO DOS RESULTADOS

Este capítulo apresenta e discute os resultados em duas frentes. Primeiro, caracteriza-se o conjunto de dados sintéticos gerado, verificando a aderência das distribuições realizadas às fontes estatísticas que as fundamentam, a coerência dos cenários de contraste e a sensibilidade do problema ao componente idiossincrático do gerador. Em seguida, apresentam-se os resultados da modelagem preditiva com os classifica-

dores *Decision Tree*, *Random Forest* e *XGBoost* frente aos *baselines* de referência, contemplando o estudo de ablação das variáveis derivadas, a comparação de métricas nos conjuntos de *holdout* e validação cruzada ($k = 10$) com verificação estatística das diferenças, a análise por classe via curvas ROC e matrizes de confusão, e a interpretação da importância dos atributos.

5.1. CARACTERIZAÇÃO DA BASE SINTÉTICA E ADERÊNCIA ÀS FONTES

O conjunto de dados definitivo contém 3.000 registros gerados sob o cenário base (rede pública estadual do Piauí), com semente aleatória fixa para reprodutibilidade. A Tabela 4 compara, para os principais parâmetros, o valor-alvo derivado das fontes com o valor efetivamente realizado na base gerada.

Tabela 4 – Aderência entre os parâmetros-alvo (fundamentados nas fontes) e as distribuições realizadas na base sintética.

Indicador	Alvo	Realizado
Renda <i>per capita</i> até 0,5 SM	60,0%	58,0%
Acesso doméstico à internet	92,5%	92,6%
Posse de computador ou <i>tablet</i>	25,5%	26,1%
Estudantes trabalhadores	26,3%	27,3%
Distorção idade-série	21,7%	22,7%

Fonte: Elaborado pelo autor (2026).

Todas as distribuições realizadas situam-se a menos de dois pontos percentuais dos alvos documentados, atestando a fidelidade da parametrização. A distribuição etária apresenta a assimetria positiva esperada (coeficiente de assimetria de +1,39; mediana de 16 anos; máximo de 24), reproduzindo a cauda longa da distorção idade-série em vez de uma distribuição normal artificial. A regra de elegibilidade do Pé-de-Meia resultou em 47,7% de beneficiários efetivos (coerente com um cenário em que 58% dos estudantes se encontram na faixa de renda elegível e a adesão é de 85% condicionada à frequência mínima). As variáveis acadêmicas parciais do primeiro semestre apresentaram médias plausíveis para o contexto: nota média de 6,4; frequência média de 84,8%; média de 5,9 nas atividades, com taxa de entrega de 63,8%; e 0,62 reprovação acumulada por estudante.

A variável-alvo, rotulada pelos critérios da LDB sobre a projeção anual (média mínima de 6,0 e frequência mínima de 75%, com banda pedagógica de fronteira para a classe intermediária), distribuiu-se em 45,0% de *Aprovados*, 31,7% *Em Risco* e 23,3% de *Reprovados*. Destaca-se que essas proporções **emergem** da regra de rotulagem (não foram calibradas por limiares de distribuição), e configuram um desbalanceamento moderado e realista, no qual nenhuma classe é dominante, justificando o emprego do SMOTE no treinamento.

Nos cenários de contraste, o gerador comportou-se de forma coerente. No perfil de **média nacional**, as classes distribuíram-se em 46,7% / 30,1% / 23,1%, próximas do cenário base (consistente com o fato de os indicadores piauienses de conectividade e distorção se aproximarem das médias nacionais após as políticas recentes). No perfil de **alta vulnerabilidade**, a proporção de reprovados subiu para 36,0% e a de beneficiários do Pé-de-Meia alcançou 57,6%. No perfil de **rede privada de renda alta**, 81,4% dos estudantes projetam-se aprovados e apenas 6,5% tornam-se beneficiários (resultado emergente exclusivamente da regra de elegibilidade, sem qualquer parâmetro dedicado, o que confirma a hipótese H4).

5.2. SENSIBILIDADE AO RUÍDO IDIOSSINCRÁTICO

Como descrito na Metodologia, a deriva do segundo semestre possui um componente idiossincrático (ruído gaussiano que representa os fatores não observáveis, como saúde, eventos familiares e motivação, capazes de alterar trajetórias escolares). O parâmetro `NOISE_LEVEL` controla a magnitude dessa variância e, por consequência, o teto teórico de acurácia alcançável por qualquer classificador (erro de Bayes do problema). Para caracterizar essa relação, o desempenho da *Random Forest* (hiperparâmetros fixos, validação cruzada estratificada $k = 5$ com SMOTE no *fold*) foi medido sob diferentes níveis de ruído, mantidos constantes todos os demais parâmetros (Tabela 5).

Tabela 5 – Análise de sensibilidade: desempenho da *Random Forest* em função da variância idiossincrática do gerador.

NOISE_LEVEL	σ da média (pontos)	σ da frequência (p.p.)	Acurácia (CV $k=5$)
0,0	0,00	0,0	90,7% $\pm$ 0,8%
1,0	0,10	0,8	90,2% $\pm$ 0,6%
2,0	0,20	1,6	88,1% $\pm$ 0,8%
3,0	0,30	2,4	84,3% $\pm$ 0,6%
4,0	0,40	3,2	80,5% $\pm$ 1,6%
5,0	0,50	4,0	77,5% $\pm$ 1,2%
6,0	0,60	4,8	74,4% $\pm$ 1,2%
8,0	0,80	6,4	68,7% $\pm$ 1,4%

Fonte: Elaborado pelo autor (2026).

A curva evidencia a degradação monotônica do desempenho com o aumento da variância não observável, confirmando que os classificadores não memorizam a regra geradora, apenas aproximam o sinal observável. Ressalta-se que a varredura **caracteriza o problema**, não busca uma acurácia-alvo: o valor adotado na base definitiva (`NOISE_LEVEL = 3,0`, equivalente a um desvio-padrão de 0,3 ponto na média anual projetada e de 2,4 pontos percentuais na frequência) é uma decisão fenomenológica, registrada na matriz de rastreabilidade, que representa uma magnitude plausível de variação semestral atribuível a fatores idiossincráticos. Sob essa escolha, o desempenho

esperado situa-se na faixa de 80–85%, compatível com os resultados reportados pela literatura de EDM em problemas análogos (COSTA *et al.*, 2012; SHABNAM; RAMACHANDRIAH; HALADAPPA, 2025).

5.3. ESTUDO DE ABLAÇÃO DAS VARIÁVEIS DERIVADAS

Antes da comparação final, avaliou-se a contribuição de cada variável derivada candidata por meio de um estudo de ablação (validação cruzada estratificada $k = 10$, SMOTE no *fold*, hiperparâmetros fixos moderados). A Tabela 6 resume as configurações principais.

Tabela 6 – Estudo de ablação das variáveis derivadas (acurácia média na validação cruzada, $k = 10$).

Configuração	<i>Decision Tree</i>	<i>Random Forest</i>	XGBoost
Sem variáveis derivadas	80,1%	84,9%	84,5%
+ nota_x_frequencia	80,9%	84,7%	84,7%
+ reprov_por_serie	80,3%	84,7%	84,5%
+ escore acadêmico composto	79,6%	84,2%	84,1%
Conjunto final (duas derivadas)	79,7%	84,6%	84,9%
Conjunto final + escore composto	80,2%	84,5%	84,9%

Fonte: Elaborado pelo autor (2026).

Os resultados indicam que o escore acadêmico composto (combinação linear ponderada de nota, frequência e reprovações) **não agrega desempenho** em nenhuma configuração: isoladamente, chega a degradar levemente os três modelos, e somado ao conjunto final produz diferenças dentro da margem de flutuação. A variável foi, portanto, excluída do conjunto final, por dois motivos convergentes: a ausência de ganho empírico e sua proximidade estrutural com a lógica de rotulagem (uma combinação de nota e frequência, os mesmos insumos dos critérios da LDB), que tornaria a interpretação das importâncias menos transparente. O conjunto final contém apenas as duas variáveis derivadas de interação (*nota_x_frequencia* e *reprov_por_serie*).

5.4. COMPARAÇÃO GERAL DAS MÉTRICAS

Os três modelos e os dois *baselines* foram treinados sobre a base definitiva de $N = 3.000$ amostras, dividida em proporção estratificada de 70/30 para treinamento e teste. O ajuste fino de hiperparâmetros deu-se por meio do *GridSearchCV* com validação cruzada de 5 subconjuntos e o SMOTE foi confinado a cada *fold* de treino. A Tabela 7 sumariza as métricas nos ambientes de *holdout* e validação cruzada ($k = 10$), e a Tabela 8 apresenta os *baselines* de referência.

Tabela 7 – Métricas comparativas entre os estimadores nos conjuntos de *holdout* e validação cruzada.

Métrica de Desempenho	<i>Decision Tree</i>	<i>Random Forest</i>	XGBoost
Acurácia (<i>Holdout</i>)	78,22%	82,00%	84,11%
<i>F1-Score</i> Ponderado (<i>Holdout</i>)	78,58%	81,94%	84,10%
Precisão Ponderada (<i>Holdout</i>)	79,97%	81,90%	84,12%
<i>Recall</i> Ponderado (<i>Holdout</i>)	78,22%	82,00%	84,11%
AUC-ROC OvR (<i>Holdout</i>)	90,32%	95,13%	95,63%
Acurácia (CV $k = 10$)	80,33% $\pm$ 1,55%	85,05% $\pm$ 2,09%	85,52% $\pm$ 1,99%
<i>F1-Score</i> Ponderado (CV $k = 10$)	80,50% $\pm$ 1,70%	85,03% $\pm$ 2,09%	85,54% $\pm$ 1,98%
Precisão Ponderada (CV $k = 10$)	81,07% $\pm$ 1,84%	85,17% $\pm$ 2,08%	85,64% $\pm$ 1,95%
<i>Recall</i> Ponderado (CV $k = 10$)	80,33% $\pm$ 1,55%	85,05% $\pm$ 2,09%	85,52% $\pm$ 1,99%

Fonte: Elaborado pelo autor (2026).

Tabela 8 – *Baselines* de referência sob o mesmo protocolo de avaliação.

Métrica de Desempenho	Classe Majoritária	Regressão Logística
Acurácia (<i>Holdout</i>)	45,00%	80,22%
<i>F1-Score</i> Ponderado (<i>Holdout</i>)	27,93%	80,26%
Acurácia (CV $k = 10$)	45,05% $\pm$ 0,23%	83,57% $\pm$ 2,60%
<i>F1-Score</i> Ponderado (CV $k = 10$)	27,98% $\pm$ 0,24%	83,69% $\pm$ 2,55%

Fonte: Elaborado pelo autor (2026).

Três leituras emergem das tabelas. Primeiro, o problema **não é trivial**: o *baseline* de classe majoritária alcança apenas 45% de acurácia (e 28% de *F1-Score*), de modo que os classificadores de *ensemble* o superam em cerca de 39 pontos percentuais de acurácia no *holdout* (confirmando a hipótese H3, de que os sinais parciais do primeiro semestre carregam informação preditiva substancial sobre o desfecho anual). Segundo, a Regressão Logística captura a maior parte do sinal (80,22% no *holdout*), indicando que o fenômeno tem forte componente linear; ainda assim, os *ensembles* agregam entre 2 e 4 pontos percentuais sobre ela, parcela atribuível às interações não lineares do fenômeno gerador. Terceiro, os modelos de *ensemble* superaram consistentemente a *Decision Tree* em todas as dimensões de medição, em conformidade com a hipótese H1.

Para verificar se as diferenças entre os modelos são estatisticamente significativas, aplicou-se o teste pareado de postos sinalizados de Wilcoxon sobre os *F1-Scores* dos mesmos *folds* da validação cruzada (DEMŠAR, 2006). A superioridade dos *ensembles* sobre a *Decision Tree* é significativa ao nível de 5% (*Random Forest* $\times$ *Decision Tree*: $p = 0,002$; XGBoost $\times$ *Decision Tree*: $p = 0,002$). Já a diferença entre XGBoost e *Random Forest* **não é estatisticamente significativa** ($p = 0,232$): a vantagem numérica do XGBoost (85,54% contra 85,03% de *F1* médio) não pode ser distinguida da flutuação amostral. O XGBoost é adotado como modelo final por sua vantagem numérica consistente em todas as métricas do *holdout* (em especial na classe intermediária, como se

verá adiante), registrando-se, contudo, que os dois *ensembles* são estatisticamente equivalentes neste problema.

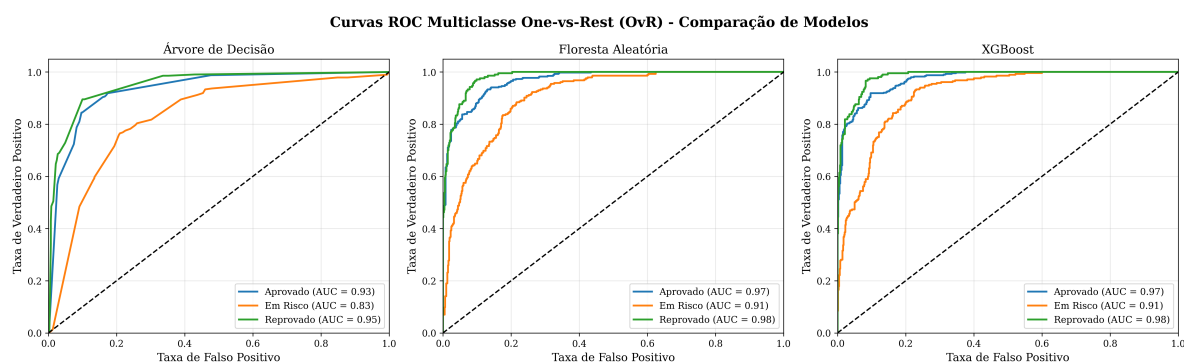
A superioridade dos métodos de comitê em relação à árvore única admite leitura arquitetural, além da métrica. O *bagging* da *Random Forest* reduz a variância ao combinar árvores treinadas sobre amostras *bootstrap* e subconjuntos aleatórios de atributos, o que confere resistência ao sobreajuste (DENISKO; HOFFMAN, 2018). O *boosting* do XGBoost encadeia árvores que corrigem sequencialmente os erros das anteriores e penaliza a complexidade via regularização explícita (CHEN; GUESTRIN, 2016), o que favorece a captura das interações não lineares do fenômeno gerador (em especial a dupla penalidade entre trabalho juvenil e distorção idade-série). A Regressão Logística, ao atingir 80,22% no *holdout*, confirma que o sinal linear já é forte; os dois a quatro pontos percentuais adicionais dos *ensembles* correspondem precisamente a essas interações.

Quanto à generalização, a acurácia de *holdout* de cada modelo situa-se dentro (ou ligeiramente abaixo) da faixa observada na validação cruzada (o XGBoost, por exemplo, obteve 84,11% no *holdout* contra $85,52\% \pm 1,99\%$ na validação cruzada), não havendo, portanto, indícios de sobreajuste. A regularização do *gradient boosting* (taxa de aprendizado de 0,1, profundidade máxima 3, subamostragem de 80%) e a poda da árvore única contribuíram para esse comportamento. Os três modelos situaram-se na faixa de 78–86%, próxima do teto estabelecido pela análise de sensibilidade, o que sugere que o desempenho restante é limitado pelo componente idiossincrático do fenômeno, e não pela capacidade dos algoritmos.

5.5. ANÁLISE DE DESEMPENHO POR CLASSE E CURVA ROC

Para subsidiar intervenções pedagógicas qualificadas, o sistema preditivo deve demonstrar sensibilidade nas três categorias de prognóstico. A Figura 5 ilustra o comportamento comparativo das curvas ROC sob o paradigma *One-vs-Rest* (OvR).

Figura 5 – Comparação das curvas ROC *One-vs-Rest* (OvR) dos classificadores propostos.



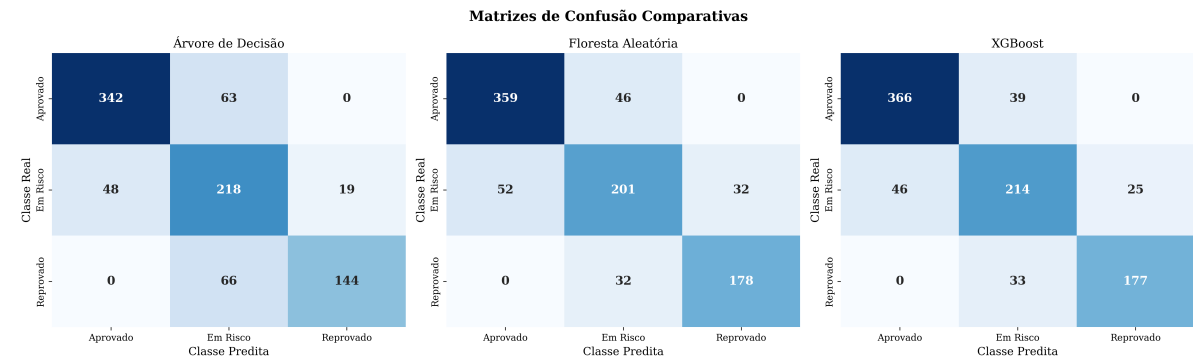
Fonte: Elaborado pelo autor (2026).

No XGBoost, as classes extremas apresentaram excelente separabilidade (AUC de 0,971 para *Aprovado* e 0,983 para *Reprovado*), enquanto a classe intermediária *Em Risco* registrou AUC de 0,915. O mesmo padrão repete-se nos demais modelos (*Decision Tree*: 0,932 / 0,827 / 0,951; *Random Forest*: 0,968 / 0,905 / 0,981, na ordem *Aprovado* / *Em Risco* / *Reprovado*). A menor separabilidade da classe intermediária é esperada e estruturalmente explicável: ela ocupa a banda de fronteira da rotulagem, comprimida por duas fronteiras de decisão simultâneas, e os estudantes próximos aos limiares da LDB compartilham perfis observáveis com as classes vizinhas. Do ponto de vista pedagógico, o desempenho elevado na classe *Reprovado* (*recall* de 84,3% no XGBoost, com precisão de 87,6%) é o resultado mais relevante: o modelo raramente deixa de sinalizar o estudante em vulnerabilidade acadêmica severa, minimizando os falsos negativos que mais comprometeriam intervenções preventivas.

5.6. ESTUDO DETALHADO DAS CLASSES VIA MATRIZ DE CONFUSÃO

Para investigar a origem granular dos erros de predição, elaborou-se o mapa de calor das matrizes de confusão (Figura 6).

Figura 6 – Matrizes de confusão obtidas nos ensaios de *holdout* (teste).



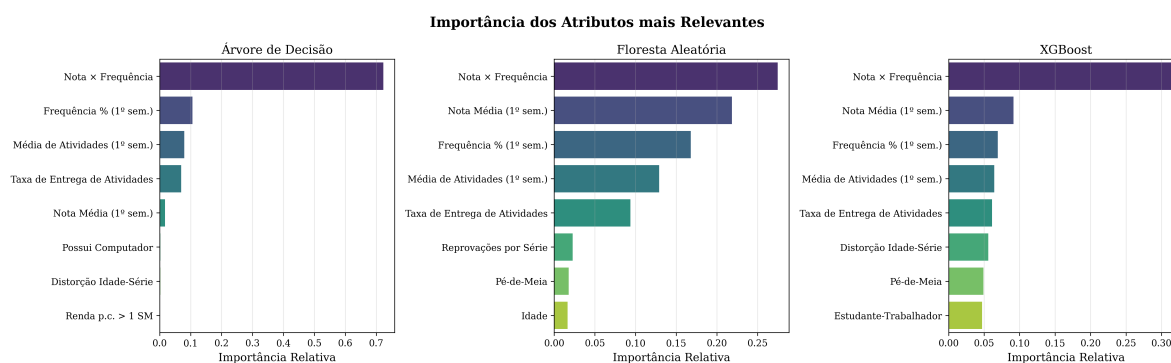
Fonte: Elaborado pelo autor (2026).

O achado mais significativo é a **ausência total de erros entre classes extremas** nos três modelos: nenhum estudante *Aprovado* foi classificado como *Reprovado*, nem o inverso. Todos os erros concentram-se em adjacências ordinais (*Aprovado* ↔ *Em Risco* e *Em Risco* ↔ *Reprovado*), o que indica que os classificadores capturaram a ordinalidade latente da tendência de desempenho. No XGBoost, dos 143 erros cometidos em 900 instâncias de teste, 85 ocorreram na fronteira *Aprovado–Em Risco* e 58 na fronteira *Em Risco–Reprovado* (erros de gravidade operacional reduzida, pois deslocam o estudante para uma categoria vizinha de atenção, sem inverter o prognóstico).

5.7. ANÁLISE DIMENSIONAL E IMPORTÂNCIA DOS ATRIBUTOS

A sustentabilidade ética e científica de intervenções educacionais assistidas por aprendizado de máquina exige interpretabilidade. A Figura 7 apresenta as contribuições relativas dos atributos em cada classificador.

Figura 7 – Critério de importância relativa dos atributos na predição de desempenho.



Fonte: Elaborado pelo autor (2026).

A variável derivada de interação nota–frequência (*nota_x_frequencia*) consolidou-se como o preditor de maior relevância em todas as estruturas (cerca de 72% na *Decision Tree*, 28% na *Random Forest* e 32% no XGBoost), corroborando a literatura educacional: nota e frequência não operam isoladamente, mas de maneira sinérgica na representação do rendimento discente (CURI, 2006). Na sequência aparecem os demais sinais acadêmicos parciais, incluindo as variáveis de engajamento contínuo introduzidas neste trabalho: no XGBoost, a média das atividades responde por 6,4% e a taxa de entrega por 6,1% da importância (evidência de que a regularidade do engajamento agrega informação além da nota média e da frequência, em linha com a literatura de EDM (COSTA *et al.*, 2012; FREDRICKS; BLUMENFELD; PARIS, 2004)).

O resultado central para a questão de pesquisa, contudo, está na camada seguinte de importâncias do XGBoost: os **fatores sociológicos** dominam as posições subseqüentes: distorção idade-série (5,6%), recebimento do Pé-de-Meia (4,9%), trabalho juvenil (4,7%), renda *per capita* de até meio salário mínimo (3,5%) e acesso à internet (3,1%). Cumpre interpretar esse achado nos seus devidos termos: como o fenômeno gerador é sintético, ele não constitui evidência empírica independente de que tais fatores condicionam o desempenho real (essa relação é premissa do gerador, fundamentada na literatura sociológica (BERLATO; MENDES; ANDRETTA, 2020; MARTURANO; PIZATO, 2015)). O que o resultado demonstra é de natureza **metodológica**: os classificadores foram capazes de *recuperar*, a partir exclusivamente dos atributos observáveis, o peso não redundante que os fatores sociológicos receberam no fenômeno gerador, mesmo na presença de indicadores acadêmicos diretos, o que confirma a hipótese H2 e valida a capacidade da arquitetura de capturar essa estrutura quando ela existir nos dados. A posição de destaque do indicador de assistência financeira, em particular,

mostra que o modelo capturou o efeito de reversão de tendência do Pé-de-Meia embutido nas regras de geração, cuja fundamentação está na literatura (SANTOS; COSTA; COSTA, 2026).

Por fim, ressalta-se a salvaguarda anti-viés: gênero e cor/raça foram gerados com peso zero no fenômeno e excluídos do treinamento, juntamente com a origem escolar, de modo que nenhuma decisão do modelo se apoia em atributos demográficos protegidos, uma decisão alinhada às boas práticas de equidade algorítmica em contextos educacionais. As limitações que condicionam a leitura desses resultados são discutidas, em conjunto com suas implicações, na seção de Limitações e Implicações das Considerações Finais.

6 CONSIDERAÇÕES FINAIS

O presente trabalho propôs o desenvolvimento de um modelo preditivo de tendência de desempenho acadêmico para estudantes do ensino médio da rede pública estadual do Piauí, fundamentado em fatores sociológicos. Para tanto, foram realizadas duas entregas articuladas: a construção de um conjunto de dados sintéticos com parametrização rastreável a estatísticas oficiais e à literatura acadêmica (no qual os preditores representam sinais parciais do primeiro semestre e o alvo, rotulado pelos critérios de aprovação da LDB, representa a tendência ao final do ano letivo), e a comparação sistemática de três algoritmos de classificação (*Decision Tree*, *Random Forest* e XGBoost) frente a *baselines* de referência, sob o processo KDD, com validação cruzada estratificada ($k = 10$), balanceamento por SMOTE confinado a cada *fold* e teste estatístico pareado entre os modelos.

Retomam-se, a seguir, as hipóteses formuladas na Introdução:

H1 (confirmada parcialmente). Os classificadores de *ensemble* superaram a árvore única e os *baselines* com significância estatística (Wilcoxon pareado, $p = 0,002$ em ambas as comparações): a *Decision Tree* atingiu 78,22% de acurácia no *holdout*, a *Random Forest* 82,00% e o XGBoost 84,11%, contra 45,00% do *baseline* majoritário e 80,22% da Regressão Logística. Entre os dois *ensembles*, contudo, a diferença não é estatisticamente significativa ($p = 0,232$): o XGBoost foi adotado como modelo final por sua vantagem numérica consistente, registrando-se a equivalência estatística com a *Random Forest*.

H2 (confirmada). Os fatores sociológicos parametrizados no fenômeno gerador foram recuperados pelos modelos como atributos de importância não redundante: após os sinais acadêmicos parciais, as variáveis mais determinantes no XGBoost foram a distorção idade-série, o recebimento do Pé-de-Meia, o trabalho juvenil, a renda *per capita* e o acesso à internet, o que serve como validação metodológica de que a arquitetura captura essa estrutura quando presente nos dados.

H3 (confirmada). Os sinais parciais do primeiro semestre foram suficientes para prever a tendência do desfecho anual com desempenho cerca de 39 pontos percentuais acima do *baseline* majoritário, em um cenário com deriva sistemática e idiossincrática de segundo semestre.

H4 (confirmada). A regra legal de elegibilidade da assistência financeira produziu, sem qualquer parâmetro dedicado, proporções coerentes de beneficiários entre os cenários: 47,7% no cenário base, 57,6% no perfil de alta vulnerabilidade e 6,5% no perfil de renda alta.

Nos três modelos, nenhum erro ocorreu entre as classes extremas (*Aprovado* e *Reprovado*): todos os equívocos concentraram-se em adjacências ordinais, e a classe crítica *Reprovado* obteve *recall* de 84,3% no XGBoost (um desempenho relevante para a finalidade preventiva do modelo), pois minimiza os falsos negativos no grupo que mais demanda intervenção pedagógica, em momento do ano letivo no qual a ação preventiva ainda é possível.

O bom comportamento do XGBoost é coerente com suas características arquiteturas: a otimização por aproximação de Taylor de segunda ordem, que incorpora a curvatura da função de perda via Hessiano, e o termo de regularização explícito que penaliza a complexidade das árvores (CHEN; GUESTRIN, 2016). Essas propriedades permitem ao algoritmo capturar interações não lineares complexas (como a sinergia entre desempenho acadêmico e contexto socioeconômico) que modelos mais simples não representam adequadamente (FRIEDMAN, 2001), ainda que, neste problema, a *Random Forest* tenha alcançado patamar estatisticamente equivalente.

6.1. CONTRIBUIÇÕES DO TRABALHO

A principal contribuição desta pesquisa reside na demonstração de que é possível unir, em uma única investigação, a construção de um conjunto de dados sintéticos sociologicamente fundamentado (com rastreabilidade documentada entre cada parâmetro de geração e sua fonte estatística ou acadêmica, incluindo o nível de desagregação e as notas de *proxy*, exportada automaticamente pelo gerador) e a modelagem preditiva comparativa da tendência de desempenho acadêmico. O gerador incorpora mecanismos anti-deterministas ausentes em abordagens ingênuas de simulação: a assistência financeira é derivada de regra de elegibilidade legal (e não de sorteio), as interações entre vulnerabilidades são não lineares, perfis de resiliência acadêmica impedem que os modelos aprendam um determinismo social absoluto, e a rotulagem deriva dos critérios reais de aprovação da LDB, com proporções de classe emergentes.

Adicionalmente, o trabalho contribui para a consolidação de boas práticas metodológicas em Mineração de Dados Educacionais: prevenção de vazamento de dados por meio do SMOTE aplicado exclusivamente no interior de cada partição da validação

cruzada (MADUGULA; Esteva de la Rosa; SHANKAR, 2026); adoção do *F1-Score* ponderado como métrica de otimização em cenário desbalanceado (HE; GARCIA, 2009); inclusão de *baselines* de referência e verificação estatística das diferenças entre modelos (DEMŠAR, 2006); seleção de variáveis derivadas validada por estudo de ablação; caracterização explícita do teto de acurácia via análise de sensibilidade ao ruído idiossincrático; e salvaguardas de equidade algorítmica, com exclusão de gênero e cor/raça do treinamento. Todo o procedimento é reproduzível e parametrizável, permitindo a simulação de cenários socioeconômicos contrastantes por simples troca de configuração.

6.2. LIMITAÇÕES E IMPLICAÇÕES

Não obstante os resultados promissores, algumas limitações devem ser consideradas. A primeira e mais significativa é a natureza sintética da base (em particular, por se tratar de simulação baseada em regras, e não de dados sintéticos derivados de microdados reais): embora cada distribuição marginal seja ancorada em fontes oficiais auditadas, as relações conjuntas entre variáveis refletem as regras especificadas pelo pesquisador. As métricas reportadas medem, portanto, a capacidade dos classificadores de recuperar um fenômeno cuja estrutura (ainda que sociologicamente fundamentada, auditada quanto às fontes e dotada de componente estocástico irreduzível) é conhecida por construção, e o desempenho absoluto não deve ser extrapolado diretamente para populações reais. O fenômeno gerador é conhecido por construção, e a transposição para registros reais de estudantes pode revelar padrões, vieses e relações não capturados pela parametrização. A validação com dados reais constitui etapa imprescindível antes de qualquer aplicação prática dos modelos, e as conclusões sobre a importância dos fatores sociológicos devem ser lidas como validação metodológica, não como evidência empírica independente. Como implicação, o valor da evidência produzida está na validação metodológica da arquitetura de classificação, na caracterização do comportamento comparativo dos algoritmos (incluindo a equivalência estatística entre os *ensembles*) e na demonstração de que a estrutura sociológica parametrizada é recuperável pelos modelos a partir dos atributos observáveis.

A segunda limitação refere-se ao recorte do fenômeno: o modelo prevê a tendência do desfecho anual a partir dos sinais do primeiro semestre, sem capturar a dinâmica temporal mais longa da trajetória escolar (transições entre séries, efeitos cumulativos de reprovações sucessivas, sazonalidade da frequência). A terceira concerne à técnica de balanceamento: a literatura recente aponta que a sobreamostragem sintética introduz riscos de privacidade em dados reais (particularmente ataques de inferência de associação), recomendando, para implantações futuras, técnicas de geração com garantias formais de privacidade diferencial (MADUGULA; Esteva de la Rosa; SHANKAR, 2026; KOSTOPOULOS; TSIKMAKI; KOTSIANTIS, 2026).

Quanto à suficiência dos parâmetros, o conjunto adotado cobre o recorte declarado

(sinais parciais intra-ano, contexto socioeconômico individual e desfecho do mesmo ano) e está ancorado nas fontes oficiais e na literatura. Há, contudo, espaço de refinamento: novos atributos de engajamento e de trajetória, substituição do SMOTE pelo SMOTENC nas colunas binárias, repetição da validação cruzada com múltiplas sementes e, sobretudo, a validação com registros reais. Esses refinamentos não invalidam o recorte atual; delimitam a evolução natural do gerador e do procedimento experimental.

Como desdobramentos futuros da pesquisa, destaca-se sobretudo a validação da arquitetura de classificação com registros reais de estudantes (mediante autorização institucional e ética), direção que dialoga diretamente com os trabalhos do grupo de pesquisa voltados à construção de conjuntos de dados institucionais sobre trajetória acadêmica, bem como a integração do modelo a sistemas de gestão escolar, com alertas precoces às equipes pedagógicas, em sinergia com os sistemas *web* de apoio à gestão educacional desenvolvidos no mesmo grupo. Delineiam-se ainda, como extensões metodológicas, a modelagem ordinal e temporal do desfecho, a incorporação de técnicas de IA explicável (SHAP e LIME) para explicações individualizadas das predições, o uso do gerador parametrizável para simulação de contrafactuais de políticas públicas e a adoção de geração sintética com garantias formais de privacidade diferencial nas etapas que envolverem dados reais (KOSTOPOULOS; TSIKMAKI; KOTSIANTIS, 2026).

REFERÊNCIAS

ARTERO, A. O. **Inteligência Artificial: teórica e prática**. São Paulo: Livraria da Física, 2009. ISBN 8578610296. Citado na página 14.

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineração de dados educacionais: Oportunidades para o Brasil. **Revista Brasileira de Informática na Educação**, v. 19, n. 2, p. 03, 2011. Citado 4 vezes nas páginas 3, 7, 10 e 17.

BERLATO, H.; MENDES, L.; ANDRETTA, D. Perspectivas de carreira de jovens do ensino médio de escolas públicas: transgressão ou reprodução das condições sociais? **Cadernos EBAPE.BR**, v. 18, n. 4, p. 865–876, 2020. Citado 5 vezes nas páginas 3, 5, 8, 22 e 35.

Brasil. **Lei nº 9.394, de 20 de dezembro de 1996. Estabelece as diretrizes e bases da educação nacional**. Brasília: [s.n.], 1996. Diário Oficial da União. Disponível em: <https://www.planalto.gov.br/ccivil_03/leis/l9394.htm>. Acesso em: 11 jul. 2026. Citado 2 vezes nas páginas 23 e 25.

Brasil. **Lei nº 14.818, de 16 de janeiro de 2024. Institui incentivo financeiro-educacional, na modalidade de poupança, para a permanência e conclusão escolar de estudantes matriculados no ensino médio público**. Brasília: [s.n.], 2024. Diário Oficial da União. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2024/lei/l14818.htm>. Acesso em: 11 jul. 2026. Citado 2 vezes nas páginas 8 e 23.

CAMPOS, A. B. T. *et al.* Evasão e abandono escolar: análise das consequências sociais, econômicas e dos fatores contribuintes. **Anais da IX Mostra Científica da Faculdade Multivix Cachoeiro de Itapemirim**, v. 9, n. 1, 2024. Citado 2 vezes nas páginas 3 e 8.

CASTRO, L. N. d.; FERRARI, D. G. **Introdução à Mineração de Dados: Conceitos Básicos, Algoritmos e Aplicações**. [S.l.: s.n.], 2016. ISBN 9788547200985. Citado na página 11.

CHAWLA, N. V. *et al.* Smote: Synthetic minority over-sampling technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321–357, 2002. Citado na página 12.

CHEN, T.; GUESTRIN, C. XGBoost: A scalable tree boosting system. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. San Francisco, CA, USA: ACM, 2016. (KDD '16), p. 785–794. Citado 4 vezes nas páginas 14, 15, 33 e 37.

CHIAPINOTO, A. M.; LUNARDI, E. M. Análise das produções sobre distorção idade/ano e políticas públicas sobre a defasagem na aprendizagem e correção do fluxo escolar. **Lumen et Virtus**, v. 15, n. 38, p. 1068–1085, 2024. Citado na página 8.

COSTA, E. *et al.* Mineração de dados educacionais: Conceitos, técnicas, ferramentas e aplicações. v. 1, n. 1, p. 1–29, 2012. Citado 6 vezes nas páginas 11, 17, 18, 21, 31 e 35.

- CURI, A. Z. **A relação entre o desempenho escolar e os salários no Brasil**. Dissertação (Mestrado) — Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo, São Paulo, 2006. Citado 2 vezes nas páginas 7 e 35.
- D'ABREU, L. C. F.; MARTURANO, E. M. Associação entre comportamentos externalizantes e baixo desempenho escolar: uma revisão de estudos prospectivos e longitudinais. **Estudos de Psicologia (Natal)**, v. 15, n. 1, p. 43–51, 2010. Citado 2 vezes nas páginas 3 e 7.
- DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. **Journal of Machine Learning Research**, v. 7, p. 1–30, 2006. Citado 4 vezes nas páginas 17, 27, 32 e 38.
- DENISKO, D.; HOFFMAN, M. M. Classification and interaction in random forests. **Proceedings of the National Academy of Sciences**, v. 115, n. 8, p. 1690–1692, 2018. Citado 2 vezes nas páginas 14 e 33.
- FAWCETT, T. An introduction to ROC analysis. **Pattern Recognition Letters**, v. 27, n. 8, p. 861–874, 2006. Citado na página 16.
- FAYYAD, U.; PIATETSKY-SHAPIO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, v. 17, n. 3, p. 37, 1996. Citado na página 10.
- FREDRICKS, J. A.; BLUMENFELD, P. C.; PARIS, A. H. School engagement: potential of the concept, state of the evidence. **Review of Educational Research**, v. 74, n. 1, p. 59–109, 2004. Citado 3 vezes nas páginas 11, 21 e 35.
- FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. **Annals of Statistics**, v. 29, n. 5, p. 1189–1232, 2001. Citado 2 vezes nas páginas 14 e 37.
- Fundação Abrinq. **Um Retrato da Infância e Adolescência no Brasil 2023**. São Paulo, 2023. Disponível em: <<https://fadc.org.br/sites/default/files/2023-10/um-retrato-infancia-adolescencia-2023.pdf>>. Acesso em: 11 jul. 2026. Citado 4 vezes nas páginas 4, 5, 8 e 22.
- GIL, A. C. **Como Elaborar Projetos de Pesquisa**. 6. ed. São Paulo: Atlas, 2017. Citado na página 19.
- GRUS, J. **Data Science do Zero**. Rio de Janeiro: Alta Books, 2016. Citado na página 13.
- GÉRON, A. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow**. 2. ed. [S.l.]: O'Reilly Media, Inc., 2019. ISBN 1492032646. Citado 4 vezes nas páginas 12, 14, 15 e 16.
- HAND, D. J.; TILL, R. J. A simple generalisation of the area under the ROC curve for multiple class classification problems. **Machine Learning**, v. 45, p. 171–186, 2001. Citado na página 16.
- HE, H.; GARCIA, E. A. Learning from imbalanced data. **IEEE Transactions on Knowledge and Data Engineering**, v. 21, n. 9, p. 1263–1284, 2009. Citado 3 vezes nas páginas 12, 16 e 38.

Instituto Brasileiro de Geografia e Estatística. **Censo Demográfico 2022: identificação étnico-racial da população, por sexo e idade: resultados do universo**. Rio de Janeiro, 2023. Disponível em: <<https://www.ibge.gov.br/estatisticas/sociais/populacao/22827-censo-demografico-2022.html?edicao=38698>>. Acesso em: 11 jul. 2026. Citado na página 22.

Instituto Brasileiro de Geografia e Estatística. **PNAD Contínua: Acesso à Internet e à televisão e posse de telefone móvel celular para uso pessoal 2023**. Rio de Janeiro, 2024. Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv102107_informativo.pdf>. Acesso em: 11 jul. 2026. Citado 3 vezes nas páginas 4, 8 e 22.

Instituto Brasileiro de Geografia e Estatística. **Síntese de Indicadores Sociais: uma análise das condições de vida da população brasileira 2024**. Rio de Janeiro, 2024. Disponível em: <<https://www.ibge.gov.br/estatisticas/sociais/populacao/9221-sintese-de-indicadores-sociais.html>>. Acesso em: 11 jul. 2026. Citado 4 vezes nas páginas 4, 5, 8 e 22.

Instituto Brasileiro de Geografia e Estatística. **PNAD Contínua TIC: tabelas 7302 (existência de microcomputador ou tablet no domicílio) e 7307 (utilização de internet no domicílio), desagregadas por Unidade da Federação**. 2025. Sistema IBGE de Recuperação Automática (SIDRA). Disponível em: <<https://sidra.ibge.gov.br/tabela/7302>>. Acesso em: 11 jul. 2026. Citado 4 vezes nas páginas 4, 5, 8 e 22.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Resumo Técnico do Censo Escolar da Educação Básica 2024**. Brasília, 2025. Disponível em: <https://download.inep.gov.br/publicacoes/institucionais/estatisticas_e_indicadores/resumo_tecnico_censo_escolar_2024.pdf>. Acesso em: 11 jul. 2026. Citado 6 vezes nas páginas 4, 5, 8, 9, 21 e 23.

KOSTOPOULOS, G.; TSIKMAKI, M.; KOTSIANTIS, S. Benchmarking statistical and deep generative models for privacy-preserving synthetic student data in educational data mining. **Algorithms**, v. 19, n. 1, p. 39, 2026. Citado 5 vezes nas páginas 4, 9, 10, 38 e 39.

LIMA, L. C. A.; GOMES, C. A. Ensino médio para todos: oportunidades e desafios. **Revista Brasileira de Estudos Pedagógicos**, v. 94, n. 238, p. 745–769, 2013. Citado na página 8.

LIU, H.; MOTODA, H. **Computational methods of feature selection**. [S.l.]: CRC Press, 2007. Citado na página 12.

MADUGULA, S. S.; Esteva de la Rosa, P.; SHANKAR, D. Data leakage-free explainable AI for decentralized credit scoring: A SHAP-interpretable approach to default prediction. **Preprints.org**, 2026. Pré-publicação. Citado 2 vezes nas páginas 13 e 38.

MAIMON, O.; ROKACH, L. **Data Mining and Knowledge Discovery Handbook**. [S.l.]: Springer Science & Business Media, 2010. v. 14. Citado na página 11.

MARTURANO, E. M.; PIZATO, E. C. G. Preditores de desempenho escolar no 5º ano do ensino fundamental. **Psico**, Porto Alegre, v. 46, n. 1, p. 16–24, 2015. Citado 3 vezes nas páginas 3, 7 e 35.

Ministério do Desenvolvimento e Assistência Social, Família e Combate à Fome. **Informe Cadastro Único nº 50: Pé-de-Meia beneficiará estudantes do Cadastro Único com renda de até meio salário mínimo.** 2024. MDS. Disponível em: <https://www.mds.gov.br/webarquivos/MDS/2_Acoes_e_Programas/Cadastro_Unico/Informes/2024/Informe_Cadastro_Unico_N_50.pdf>. Acesso em: 11 jul. 2026. Citado 2 vezes nas páginas 8 e 23.

Ministério do Trabalho e Emprego. **Diagnóstico Ligeiro: o trabalho infantil na PNAD Contínua 2024.** Brasília, 2024. Disponível em: <<https://www.gov.br/trabalho-e-emprego/pt-br/pdfs/diagnostico-ligeiro-trabalho-infantil-na-pnadc-2024-v-1.pdf>>. Acesso em: 11 jul. 2026. Citado 2 vezes nas páginas 8 e 22.

Minuto MT. **Piauí acompanha avanço do ensino médio público.** 2026. Minuto MT. Replicação jornalística de divulgação oficial MEC/INEP, cujas páginas originais encontravam-se indisponíveis na data de acesso. Disponível em: <<https://minutomt.com.br/educacao/piaui-acompanha-avanco-do-ensino-medio-publico/>>. Acesso em: 11 jul. 2026. Citado 4 vezes nas páginas 4, 5, 9 e 23.

PALOMINIO, R. N.; BITENCOURT, C. M. Políticas públicas de enfrentamento a vulnerabilidade educacional no Brasil. **Cadernos Cajuína**, v. 10, n. 6, 2025. Citado 2 vezes nas páginas 8 e 9.

PEÑA-AYALA, A. Educational data mining. **Studies in Computational Intelligence**, Springer, v. 524, 2014. Citado 2 vezes nas páginas 11 e 18.

ROMERO, C.; VENTURA, S. Educational data mining: A review of the state of the art. **IEEE Transactions on Systems, Man, and Cybernetics, Part C**, v. 40, n. 6, p. 601–618, 2010. Citado 3 vezes nas páginas 3, 11 e 17.

RUSSEL, S. J.; NORVIG, P. **Inteligência Artificial**. Rio de Janeiro: Elsevier Editora Ltda, 2013. Citado na página 13.

SANTOS, S. d. S.; COSTA, K. T. S. d.; COSTA, N. T. S. d. O programa “pé-de-meia” e a evasão escolar: articulações pertinentes. **Cadernos Cajuína**, v. 11, n. 1, 2026. Citado 4 vezes nas páginas 4, 9, 24 e 36.

Secretaria de Estado da Educação do Piauí. **Escola pública do Piauí é destaque nacional.** s.d. SEDUC-PI. Disponível em: <<https://seduc.pi.gov.br/noticias/noticia/443/escola-publica-do-piaui-e-destaque-nacional>>. Acesso em: 11 jul. 2026. Citado 3 vezes nas páginas 9, 23 e 24.

SHABNAM, A. S. J.; RAMACHANDRIAH, T.; HALADAPPA, M. S. Predictive model to analyze real and synthetic data for learners’ performance prediction using regression techniques. **Online Learning**, v. 29, n. 1, p. 108–129, 2025. Citado 4 vezes nas páginas 4, 9, 10 e 31.

SILVA, P. M. D. *et al.* Ensemble regression models applied to dropout in higher education. In: IEEE. **2019 8th Brazilian Conference on Intelligent Systems (BRACIS)**. [S.l.], 2019. p. 120–125. Citado 2 vezes nas páginas 17 e 18.

Todos Pela Educação. **Conclusão da Educação Básica: avanços e desigualdades na última década.** 2023. Todos Pela Educação. Disponível em: <<https://todospelaeducacao.org.br/noticias/>>

conclusao-da-educacao-basica-estudo-sobre-avancos-e-desigualdades-na-ultima-decada/>. Acesso em: 11 jul. 2026. Citado 2 vezes nas páginas 3 e 5.

Todos Pela Educação. **Educação no Piauí: indicadores educacionais**. s.d. Todos Pela Educação. Disponível em: <https://www.todospelaeducacao.org.br/_uploads/_posts/490.pdf>. Acesso em: 11 jul. 2026. Citado 6 vezes nas páginas 4, 5, 8, 9, 21 e 23.

WENCZENOVICZ, T. J.; SANTOS, S. M. C. d. Desigualdades educacionais e direito à educação: uma reflexão necessária. **Revista de Sociologia, Antropologia e Cultura Jurídica**, v. 9, n. 1, p. 76–92, 2023. Citado na página 8.

APÊNDICE A – DICIONÁRIO DE ATRIBUTOS DO CONJUNTO DE DADOS SINTÉTICOS

A Tabela 9 apresenta o dicionário completo dos atributos gerados pelo gerador de dados sintéticos, referenciado no capítulo de Metodologia.

Tabela 9 – Dicionário descritivo dos atributos do conjunto de dados sintéticos.

Atributo	Tipo	Descrição
serie	Numérico (1–3)	Série do ensino médio em curso.
idade	Numérico (14–24)	Idade do estudante, com assimetria positiva decorrente da distorção idade-série.
genero	Categórico (binário)	Gênero autodeclarado (paridade ~50/50).
cor_raca	Categórico (nominal)	Cor/raça autodeclarada (Parda, Preta, Branca, Outra).
estado_civil	Categórico (nominal)	Estado civil (Solteiro, Casado/União, Outro).
renda_per_capita	Categórico (nominal)	Faixa de renda familiar <i>per capita</i> (até 0,5 SM; 0,5–1 SM; acima de 1 SM).
trabalha	Categórico (binário)	Indica se o estudante concilia trabalho e estudo.
acesso_internet	Categórico (binário)	Acesso doméstico à internet.
possui_computador	Categórico (binário)	Disponibilidade de computador para os estudos (<i>proxy</i> da exclusão digital qualitativa).
tipo_escola_origem	Categórico (binário)	Rede de origem no ensino fundamental (pública ou privada).
distorcao_idade_serie	Categórico (binário)	Indica atraso escolar de dois ou mais anos.
assistencia_pe_de_meia	Categórico (binário)	Beneficiário efetivo do programa Pé-de-Meia (derivado por regra de elegibilidade).
nota_media	Numérico (0–10)	Média das notas parciais do 1º semestre letivo.
frequencia	Numérico (%)	Percentual de frequência às aulas acumulado no 1º semestre.
media_atividades	Numérico (0–10)	Média das notas obtidas nas atividades avaliativas do 1º semestre, normalizada para a escala 0–10 (ex.: atividade valendo 2,0 pontos com nota 1,0 equivale a 5,0).
taxa_entrega_atividades	Numérico (%)	Proporção de atividades entregues no 1º semestre (sinal de regularidade do engajamento).
reprovacoes	Numérico (0–3)	Reprovações acumuladas no ensino médio.
situacao	Categórico (alvo)	Tendência de desempenho projetada para o final do ano letivo, rotulada pelos critérios da LDB: <i>Aprovado</i> , <i>Em Risco</i> ou <i>Reprovado</i> .

Fonte: Elaborado pelo autor (2026).