v7.n9.2026

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

FATORES DOCENTES E DISCENTES NA PROFICIÊNCIA ESCOLAR: UMA ABORDAGEM MULTINÍVEL DE APRENDIZADO DE MÁQUINA E IA EXPLICÁVEL NOS MICRODADOS DO SAEB

TEACHER AND STUDENT-LEVEL FACTORS IN SCHOOL PROFICIENCY: A MULTILEVEL MACHINE LEARNING AND EXPLAINABLE AI APPROACH TO SAEB MICRODATA

FACTORES DOCENTES Y DISCENTES EN LA COMPETENCIA ESCOLAR: UN ENFOQUE MULTINIVEL DE APRENDIZAJE AUTOMÁTICO E IA EXPLICABLE EN LOS MICRODATOS DEL SAEB

Rodrigo Ronner Tertulino da Silva[1], Laércio Alencar[2]

e799005
https://doi.org/10.47820/recima21.v7i9.9005

PUBLICADO: 09/2026

RESUMO

Estudos em mineração de dados educacionais e em sociologia da educação costumam examinar o desempenho escolar sob a dicotomia entre fatores individuais e fatores sistêmicos, sem isolar a contribuição específica das camadas docente e discente como dimensões próprias de análise. Este trabalho revisita, sob esse recorte, um modelo multinível de aprendizado de máquina construído a partir dos microdados do SAEB 2023, que integra o perfil socioeconômico discente, o perfil profissional docente, indicadores escolares e o perfil de gestão diretiva, com mais de 6,4 milhões de registros de estudantes do 9º ano e do Ensino Médio. Entre os quatro algoritmos comparados, o Random Forest apresentou o melhor desempenho, com acurácia de 90,17% e AUC de 0,967 no conjunto de teste. A interpretação via SHAP evidenciou que, na camada discente, a escolaridade dos pais e os bens do domicílio influenciam a proficiência, enquanto, na camada docente, o percentual de professores com formação adequada e o vínculo trabalhista figuram entre os preditores institucionais mais relevantes, ainda que subordinados ao nível socioeconômico médio da escola. Esses achados dialogam com a literatura sobre percepções docentes, clima escolar e desempenho, que indica capacidade preditiva limitada das variáveis subjetivas de percepção em relação às variáveis objetivas de perfil aqui analisadas. O estudo estende a análise sistêmica original de Tertulino e Alencar (2026), evidenciando as camadas docente e discente do modelo multinível e oferecendo subsídios para políticas de formação docente e de apoio socioeconômico ao corpo discente.

PALAVRAS-CHAVE: Mineração de dados educacionais. Perfil docente. Perfil discente. Aprendizado de máquina. IA explicável. SAEB.

ABSTRACT

Studies in educational data mining and in the sociology of education often examine school performance through the dichotomy between individual and systemic factors, without isolating the specific contribution of the teacher and student layers as analytical dimensions in their own right. This work revisits, through that lens, a multilevel machine learning model built on the 2023 SAEB microdata, integrating student socioeconomic profiles, teacher professional profiles, school indicators, and principal management profiles, comprising more than 6.4 million records of 9th- and high school students. Among four compared algorithms, Random Forest achieved the best performance, with 90.17% accuracy and an AUC of 0.967 on the test set.

[1] IFRN, doutor.
[2] IFRN, graduado.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

SHAP interpretation showed that, at the student layer, parental education and household assets influence proficiency, while, at the teacher layer, the percentage of adequately trained teachers and employment status rank among the most relevant institutional predictors, although subordinate to the school's average socioeconomic level. These findings are discussed alongside the literature on teacher perceptions, school climate, and performance, which indicates that subjective perception variables have limited predictive capacity relative to the objective profile variables analyzed here. The study extends the original systemic analysis by Tertulino and Alencar (2026), foregrounding the teacher and student layers of the multilevel model and providing evidence to inform teacher-training policies and socioeconomic support for the student body.

KEYWORDS: *Educational data mining. Teacher profile. Student profile. Machine learning. Explainable AI. SAEB.*

RESUMEN
Los estudios de minería de datos educativos y de sociología de la educación suelen examinar el rendimiento escolar bajo la dicotomía entre factores individuales y sistémicos, sin aislar la contribución de las capas docente y discente como dimensiones propias de análisis. Este trabajo revisita, bajo ese enfoque, un modelo multinivel de aprendizaje automático construido sobre los microdatos del SAEB 2023, que integra el perfil socioeconómico discente, el perfil profesional docente, indicadores escolares y el perfil de gestión directiva, con más de 6,4 millones de registros de estudiantes de 9° año y de Enseñanza Media. Entre cuatro algoritmos comparados, Random Forest presentó el mejor desempeño, con una precisión del 90,17% y un AUC de 0,967 en el conjunto de prueba. La interpretación mediante SHAP evidenció que, en la capa discente, la escolaridad de los padres y los bienes del hogar influyen en la competencia escolar, mientras que, en la capa docente, el porcentaje de profesores con formación adecuada y el vínculo laboral figuran entre los predictores institucionales más relevantes, aunque subordinados al nivel socioeconómico medio de la escuela. Estos hallazgos dialogan con la literatura sobre percepciones docentes, clima escolar y rendimiento, que indica capacidad predictiva limitada de las variables subjetivas de percepción frente a las variables objetivas de perfil aquí analizadas. El estudio amplía el análisis sistémico original de Tertulino y Alencar (2026), poniendo de relieve las capas docente y discente del modelo multinivel, con subsidios para políticas de formación docente y de apoyo socioeconómico al alumnado.

PALABRAS CLAVE: *Minería de datos educativos. Perfil docente. Perfil del alumnado. Aprendizaje automático. IA explicable. SAEB.*

INTRODUÇÃO

O desempenho acadêmico dos estudantes da educação básica é determinado por uma teia complexa de fatores que transcendem o esforço individual de cada aluno. Compreender quais variáveis, em quais níveis de análise, exercem maior influência sobre os resultados escolares é uma questão central tanto para a pesquisa em Tecnologia Educacional quanto para a formulação de políticas públicas no Brasil. O Sistema de Avaliação da Educação Básica (SAEB), conduzido pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), constitui a principal fonte de dados para essa investigação, ao medir a qualidade do

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

ensino por meio de testes de proficiência em Língua Portuguesa e Matemática e de questionários contextuais respondidos por alunos, professores, diretores e escolas (INEP, 2025).

Boa parte da literatura em mineração de dados educacionais (MDE) e em sociologia da educação examina o desempenho escolar sob uma dicotomia ampla entre fatores individuais e fatores sistêmicos (COLEMAN *et al.*, 1966; BOURDIEU; PASSERON, 1977), sem isolar, de forma dedicada, a contribuição específica das camadas docente e discente enquanto dimensões próprias de análise. Em estudo anterior dos autores, publicado na revista Discover Education, um modelo multinível de aprendizado de máquina foi construído com base nos microdados do SAEB 2023 para identificar os fatores sistêmicos associados à proficiência escolar, revelando a centralidade do nível socioeconômico médio da escola como preditor dominante (TERTULINO; ALENCAR, 2026). O presente trabalho parte desse modelo multinível para aprofundar, especificamente, a análise das variáveis de perfil docente (formação, experiência e vínculo trabalhista dos professores) e de perfil discente (escolaridade dos pais e bens do domicílio dos alunos), discutindo-as à luz da literatura sobre percepções docentes, clima escolar e desempenho (MIRANDA; BERTAGNA; FREITAS, 2019; TEIXEIRA; ALMEIDA; CASSUCE, 2022; KHINE *et al.*, 2024).

O problema de pesquisa que orienta este trabalho pode ser assim formulado: em que medida os perfis objetivos de professores e alunos, capturados pelos microdados do SAEB, explicam a proficiência escolar quando comparados aos fatores institucionais mais amplos, e como esses achados dialogam com a literatura sobre percepções docentes e discentes acerca do processo de ensino-aprendizagem? O objetivo geral deste trabalho é analisar, sob as óticas docente e discente, os fatores associados à proficiência acadêmica de estudantes do 9º ano e do Ensino Médio, a partir de um modelo multinível de aprendizado de máquina e de IA explicável aplicado aos microdados do SAEB 2023. Como objetivos específicos, busca-se: (i) comparar o desempenho preditivo de quatro algoritmos *ensemble* sobre os microdados multinível do SAEB; (ii) interpretar, por meio do *framework* SHAP, a contribuição relativa das variáveis de perfil docente e de perfil discente para a proficiência escolar; e (iii) discutir esses achados em diálogo com a literatura sobre percepções docentes e discentes e clima escolar, identificando convergências, limites e implicações para políticas de formação e valorização docente.

A justificativa deste recorte decorre da constatação de que, embora a formação, a experiência e o vínculo trabalhista dos professores, assim como a escolaridade dos pais e os bens do domicílio dos alunos, integrem o modelo multinível original, sua contribuição específica para a proficiência escolar não havia sido objeto de discussão dedicada, tampouco confrontada com a literatura que investiga como professores e alunos percebem os fatores que afetam o

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

ensino e a aprendizagem. Compreender essas duas camadas separadamente, e não apenas como componentes de um índice agregado de nível escolar, é relevante tanto para a pesquisa em MDE quanto para a formulação de políticas de formação continuada docente e de apoio socioeconômico ao corpo discente.

O artigo está organizado da seguinte forma: a Seção 1 apresenta o referencial teórico sobre mineração de dados educacionais, o SAEB, aprendizado de máquina e IA explicável, e a literatura sobre percepções docentes e discentes e fatores sistêmicos; a Seção 2 detalha a metodologia; a Seção 3 apresenta e discute os resultados; e a Seção 4 apresenta as considerações finais.

1. REFERENCIAL TEÓRICO

1.1. Mineração de dados educacionais e trabalhos relacionados

A predição do desempenho e da evasão escolar é um dos focos mais consolidados na MDE (PAPADOGIANNIS *et al.*, 2024). No contexto brasileiro, estudos aplicaram modelos de classificação para prever evasão em instituições de ensino superior, utilizando registros acadêmicos institucionais (SALES *et al.*, 2015; CARNEIRO *et al.*, 2022; COLPO *et al.*, 2024). Embora relevantes, esses trabalhos restringem-se a contextos institucionais específicos e ao ensino superior, não abordando a educação básica em escala nacional. Outro conjunto de estudos explorou técnicas de aprendizado semissupervisionado para contornar a escassez de dados rotulados (MELO *et al.*, 2023) e comparações entre aprendizado profundo e modelos tradicionais (PEREIRA *et al.*, 2020), sinalizando a busca por maior poder preditivo a custos computacionais manejáveis.

No âmbito internacional, há uma atenção crescente à interpretabilidade dos modelos. Jang *et al.* (2022) incorporaram XAI para explicar predições de desempenho em ambientes de aprendizagem *online*, demonstrando como a combinação de alta acurácia preditiva e transparência interpretativa agrega valor à prática educacional. Mastour *et al.* (2025) desenvolveram um *framework* de XAI para avaliar estudantes de medicina em avaliações de alto impacto. Ujkani *et al.* (2024) identificaram estudantes em risco utilizando técnicas de IA explicável em contexto universitário. Khine *et al.* (2024) combinaram variáveis de clima escolar, características do aluno e status socioeconômico, social e cultural em um modelo interpretável para prever o desempenho matemático de estudantes australianos no PISA 2022, encontrando que a autoeficácia do estudante e o status socioeconômico superam as variáveis de clima escolar em poder preditivo. Nenhum desses estudos, contudo, adota uma abordagem multinível e

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

sistêmica em um *dataset* de avaliação nacional da educação básica da dimensão do SAEB, com recorte dedicado às camadas docente e discente — lacuna que o presente trabalho busca preencher.

Tabela 1. Síntese dos trabalhos relacionados

Referência	Contribuição Principal	Limitação / Lacuna
Sales *et al.* (2015)	Predição de evasão em universidade pública brasileira com registros acadêmicos.	Restrito a uma instituição de ensino superior; foco em evasão, não em desempenho.
Carneiro *et al.* (2022)	MDE aplicada à prevenção de evasão em cursos de programação no Brasil.	Domínio específico; escopo de uma única disciplina.
Melo *et al.* (2023)	Aprendizado semissupervisionado para predição de evasão com dados parcialmente rotulados.	Foco em técnica específica para evasão; não aborda desempenho multinível.
Jang *et al.* (2022)	Predição precoce de desempenho com XAI para explicar decisões do modelo em *e-learning*.	Restrito a ambiente *online*; não aborda dados nacionais multinível.
Ujkani *et al.* (2024)	Identificação de estudantes em risco com técnicas de XAI em curso universitário.	Limitado a curso único; não integra dados de múltiplos agentes educacionais.
Mastour *et al.* (2025)	*Framework* XAI para predição de desempenho de estudantes de medicina em avaliações.	Domínio específico; foco institucional.
Miranda, Bertagna e Freitas (2019)	Percepções de professores sobre os fatores que compõem o clima escolar.	Abordagem qualitativa; não relaciona percepções a modelos preditivos de desempenho.
Teixeira, Almeida e Cassuce (2022)	Modelagem multinível com dados do SAEB 2015 para testar se percepções docentes explicam o desempenho.	Poder explicativo limitado das percepções subjetivas; não incorpora aprendizado de máquina nem XAI.
Khine *et al.* (2024)	Aprendizado de máquina interpretável (XGBoost/SHAP) com clima escolar e status socioeconômico no PISA.	Contexto australiano; escala nacional, mas não multinível com quatro agentes educacionais.
Tertulino e Alencar (2026)	Modelo multinível integrando dados de aluno, professor, escola e diretor do SAEB 2023 com XAI/SHAP para identificar fatores sistêmicos.	Trata fatores docentes e discentes como parte de um índice agregado, sem análise dedicada a cada camada.
Este trabalho	Recorte dedicado às camadas docente e discente do modelo multinível do SAEB 2023, em diálogo com a literatura sobre percepções docentes e discentes.	Não incorpora, ainda, as variáveis de percepção subjetiva dos questionários do SAEB (trabalho futuro).

Fonte: elaborada pelos autores.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

1.2. O SAEB como fonte de dados para a MDE

O SAEB é composto por avaliações externas, aplicadas periodicamente pelo INEP, para aferir a qualidade do ensino nas etapas da educação básica. Os microdados da edição de 2023, utilizados neste trabalho, abrangem registros de alunos do 9º ano do Ensino Fundamental e das séries do Ensino Médio, acompanhados de questionários contextuais respondidos por professores, diretores e pelas próprias escolas. Além dos blocos de caracterização objetiva (formação, experiência, vínculo trabalhista), o questionário do professor inclui blocos dedicados a percepções subjetivas, como concepções profissionais, condições de trabalho e clima escolar (colaboração entre pares, relação com as famílias e ocorrências de violência), que não integram o modelo multinível aqui analisado, mas fundamentam a discussão sobre percepções docentes retomada na Seção 1.4. A riqueza e a escala dos microdados do SAEB tornam-no uma fonte privilegiada para investigações em MDE orientadas a políticas públicas, possibilitando análises longitudinais e multiníveis dos determinantes do desempenho escolar em todo o território nacional (BONAMINO; SOUSA, 2012).

1.3. Aprendizado de máquina na educação e XAI

Os algoritmos de aprendizado de máquina baseados em árvore, como Random Forest, XGBoost, LightGBM e CatBoost, demonstraram alto desempenho em dados tabulares de larga escala (BREIMAN, 2001; CHEN; GUESTRIN, 2016). O Random Forest, em particular, caracteriza-se pela metodologia de *bagging*, que constrói múltiplas árvores de decisão sobre subamostras distintas dos dados e agrega suas predições, conferindo robustez natural ao ruído presente em dados de pesquisa survey (BENTÉJAC; CSÖRGŐ; MARTÍNEZ-MUÑOZ, 2021). Para tornar esses modelos interpretáveis, o *framework* SHAP (SHapley Additive exPlanations) calcula a contribuição de cada variável para cada predição individual com base na teoria dos jogos cooperativos (LUNDBERG; LEE, 2017). No contexto educacional, o SHAP permite transformar um modelo preditivo em uma ferramenta diagnóstica, identificando quais fatores, em quais direções, influenciam o desempenho de cada grupo de estudantes, incluindo os fatores associados especificamente ao perfil docente e ao discente.

1.4. Fatores sistêmicos, perfil docente e perfil discente na literatura

A literatura em sociologia da educação aponta, há décadas, que o desempenho escolar não é resultado exclusivo de atributos individuais dos estudantes. O relatório seminal de Coleman *et al.* (1966) estabeleceu que o contexto familiar e socioeconômico dos estudantes explica uma parcela maior da variância no desempenho do que os recursos escolares disponíveis, achado

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

complementado pela teoria da reprodução cultural de Bourdieu e Passeron (1977), segundo a qual o sistema educacional tende a valorizar o capital cultural herdado nas famílias mais privilegiadas. No Brasil, estudos com dados do SAEB confirmam que o nível socioeconômico da escola, medido pelo perfil agregado de seus alunos, é um dos preditores mais robustos do desempenho em avaliações nacionais (ALVES; SOARES, 2013).

Paralelamente a essa tradição, um conjunto distinto de estudos investiga como as percepções de professores e alunos sobre o ambiente escolar se relacionam com o desempenho. Miranda, Bertagna e Freitas (2019) analisaram, sob a perspectiva de professores da rede pública paulista, os fatores que compõem o clima escolar, destacando a relação entre colaboração docente, disciplina e qualidade do ensino. Utilizando dados do próprio SAEB (edição de 2015) e modelagem multinível, Teixeira, Almeida e Cassuce (2022) investigaram se as percepções docentes sobre obstáculos à aprendizagem no ensino médio brasileiro explicam o desempenho dos alunos, encontrando que apenas uma pequena parcela das percepções examinadas apresentou coeficientes estatisticamente significativos, o que os autores interpretam como uma capacidade limitada dos professores de identificar, subjetivamente, as causas intraescolares mais determinantes da qualidade da aprendizagem. Em contexto internacional, Khine *et al.* (2024) chegaram a um resultado convergente: entre variáveis de clima escolar, características do aluno e status socioeconômico, social e cultural, foram estas últimas, e não as percepções sobre o clima escolar, que mais contribuíram para a predição do desempenho matemático de estudantes australianos.

Esses achados sugerem uma distinção relevante para o presente trabalho: enquanto variáveis subjetivas de percepção tendem a apresentar poder explicativo limitado sobre o desempenho quando comparadas a variáveis objetivas de perfil (TEIXEIRA; ALMEIDA; CASSUCE, 2022), variáveis objetivas de perfil docente e discente — formação, experiência e vínculo trabalhista dos professores; escolaridade dos pais e bens do domicílio dos alunos — mantêm-se como preditores relevantes em modelos de aprendizado de máquina de larga escala (KHINE *et al.*, 2024; TERTULINO; ALENCAR, 2026). É justamente essa lacuna — a ausência de uma análise dedicada às camadas docente e discente de um modelo multinível nacional de larga escala, em diálogo com a literatura sobre percepções — que o presente trabalho busca preencher, estendendo a análise sistêmica original de Tertulino e Alencar (2026).

2. METODOLOGIA

2.1. Fonte de dados e integração multinível

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Os dados utilizados provêm da edição de 2023 do SAEB, a mais recente disponível no momento da coleta (acessada em janeiro de 2025). Foram integrados quatro arquivos de microdados: TS_ALUNO (estudantes), TS_PROFESSOR (professores), TS_ESCOLA (escolas) e TS_DIRETOR (diretores), por meio das chaves de identificação ID_ESCOLA e ID_TURMA. Após a integração, a remoção dos identificadores e a exclusão de registros com valores verdadeiramente ausentes por exclusão em lista (*listwise deletion*), o corpus final compreende 6.482.168 registros válidos de estudantes, provenientes de 70.151 escolas. Cabe destacar que a categoria de resposta 'não sei/não informado' (código _F em variáveis como a escolaridade dos pais) representa uma resposta válida do questionário do INEP e foi mantida como categoria codificada via *one-hot encoding*, pois seu impacto preditivo negativo, identificado pelo SHAP, constitui um achado substantivo sobre a condição socioeconômica dos estudantes que não sabem informar a escolaridade dos responsáveis.

2.2. Seleção de variáveis, engenharia de *features* e camadas analíticas

De um conjunto inicial de 98 variáveis candidatas, o algoritmo Boruta confirmou estatisticamente 44 como relevantes por meio de uma comparação iterativa com variáveis-sombra aleatórias. Desse conjunto, 17 variáveis foram selecionadas por meio de curadoria estruturada com dois pesquisadores em política educacional e dois profissionais da educação básica, seguindo critérios de relevância teórica, interpretabilidade para gestores, não redundância entre indicadores correlacionados e representatividade dos quatro níveis de análise. Para os fins deste trabalho, as 17 variáveis finais foram organizadas em quatro camadas analíticas: (a) camada discente — escolaridade materna e paterna e bens do domicílio; (b) camada docente — percentual de docentes com formação adequada, experiência e formação do corpo docente, e tipo de vínculo trabalhista do professor; (c) camada escolar — nível socioeconômico médio da escola e taxa de participação no SAEB; e (d) camada de gestão — experiência do diretor. Essa organização em camadas não altera o modelo preditivo original, mas orienta a leitura interpretativa apresentada na Seção 3, com ênfase nas camadas docente e discente.

2.3. Variável-Alvo, modelos e avaliação

A variável-alvo binária foi construída a partir da média dos escores de proficiência em Língua Portuguesa e Matemática: estudantes acima da média do *dataset* foram rotulados como 'Acima da Média' e os demais como 'Abaixo da Média'. O conjunto foi dividido em treino (80%, n = 5.185.734) e teste (20%, n = 1.296.434) por estratificação. O StandardScaler foi ajustado

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

exclusivamente sobre o conjunto de treino para evitar vazamento de dados (*data leakage*). Quatro modelos foram comparados: Random Forest (RF), XGBoost, LightGBM e CatBoost. A interpretabilidade foi garantida pelo SHAP TreeExplainer aplicado a uma amostra estratificada de 20.000 registros do conjunto de teste, cuja estabilidade foi verificada por meio de replicação em três amostras independentes.

3. RESULTADOS E DISCUSSÃO

3.1. Desempenho comparativo dos modelos

A comparação entre os quatro algoritmos, apresentada na Tabela 2, demonstrou a superioridade expressiva do Random Forest em relação aos modelos de *gradient boosting* com hiperparâmetros padrão. O RF alcançou acurácia de 90,17% e AUC de 0,967 no conjunto de teste, métricas consideravelmente superiores às dos demais modelos, cujas acurácias variaram entre 62,48% e 64,17%, com AUC entre 0,674 e 0,698. Essa diferença é atribuível principalmente à sensibilidade dos modelos de *boosting* à otimização de hiperparâmetros: em configuração padrão, o Random Forest apresenta desempenho de linha de base notoriamente mais robusto (BENTÉJAC; CSÖRGŐ; MARTÍNEZ-MUÑOZ, 2021). A diferença entre a acurácia de treino do RF (99,98%) e de teste (90,17%) é característica do método de *bagging* e não indica sobreajuste prejudicial, o que é confirmado pelo alto AUC de teste.

Tabela 2. Desempenho comparativo dos modelos nos conjuntos de treino e teste

Modelo	Acurácia Treino	Acurácia Teste	AUC Treino	AUC Teste
Random Forest	0,9998	0,9017	1,0000	0,9669
CatBoost	0,6521	0,6417	0,7089	0,6984
XGBoost	0,6418	0,6315	0,6951	0,6833
LightGBM	0,6354	0,6248	0,6862	0,6743

Fonte: elaborada pelos autores.

3.2. Classificação detalhada do modelo Random Forest

A Tabela 3 detalha o relatório de classificação do modelo Random Forest, demonstrando que o desempenho elevado se distribui de forma equânime entre as duas classes, sem viés em favor de nenhuma delas. As métricas simétricas de precisão, revocação e F1-score para as classes 'Acima da Média' e 'Abaixo da Média' confirmam a robustez e a imparcialidade do

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

classificador, uma qualidade especialmente relevante em aplicações educacionais, onde erros assimétricos de classificação poderiam gerar intervenções inadequadas.

Tabela 3. Relatório detalhado de classificação do Random Forest no conjunto de teste

Classe	Precisão	Revocação	F1-Score
Acima da Média	0,91	0,89	0,90
Abaixo da Média	0,90	0,92	0,91
Média Ponderada	0,91	0,91	0,91

Fonte: elaborada pelos autores.

3.3. Importância das variáveis: camada discente

A análise de importância das variáveis e a interpretação por meio do SHAP, sintetizadas na Tabela 4 e ilustradas na Figura 1, revelaram uma hierarquia clara entre os preditores da camada discente. A escolaridade materna e paterna ocupa posições de alta relevância, refletindo a transmissão de capital cultural descrita por Bourdieu (1986): níveis mais elevados de escolaridade dos pais associam-se a valores SHAP positivos, que deslocam as predições em direção à classificação 'Acima da Média'. Quando a escolaridade dos pais é classificada como 'desconhecida', o impacto no modelo é negativo, o que evidencia que estudantes que não sabem informar a escolaridade dos responsáveis pertencem a subgrupos socioeconômicos particularmente vulneráveis. Os bens do domicílio seguem um padrão semelhante, com maior capital econômico familiar associado a um desempenho superior.

3.4. Importância das variáveis: camada docente e institucional

Na camada docente, o percentual de professores com formação adequada à disciplina lecionada e o tipo de vínculo trabalhista do professor figuram entre os principais preditores do nível institucional, o que indica que a qualidade e a estabilidade do corpo docente atuam como fatores protetivos mesmo em contextos socioeconômicos desfavoráveis. A experiência e a formação continuada dos professores também contribuem positivamente para a proficiência, embora com magnitude inferior à dos preditores de nível discente e escolar. Ainda assim, o nível socioeconômico médio da escola (NIVEL_SOCIO_ECONOMICO) — indicador composto calculado pelo INEP a partir das respostas do questionário do aluno sobre a escolaridade dos pais e os bens do domicílio, e não da infraestrutura física da escola (INEP, 2021) — permanece como o preditor isolado de maior impacto, superando, com ampla margem, tanto as variáveis docentes quanto as discentes, quando consideradas individualmente. Esse resultado aponta para efeitos composicionais: o contexto coletivo dos estudantes de uma escola é mais determinante do que qualquer característica isolada de professores ou de alunos.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

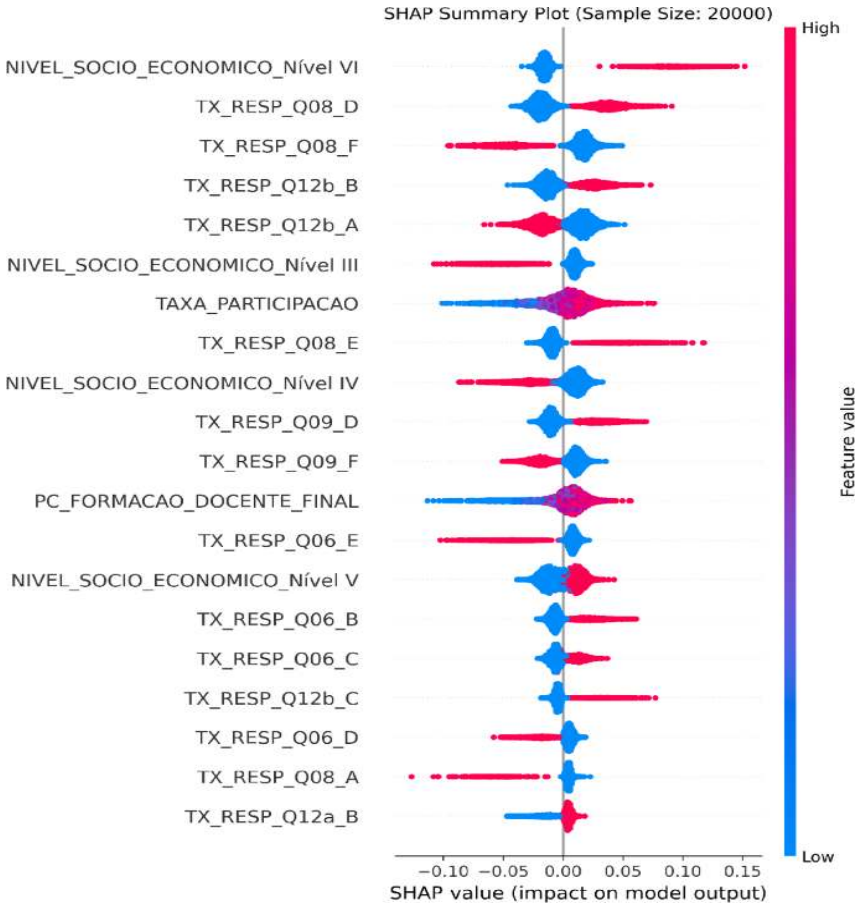
Tabela 4. Síntese das variáveis mais importantes por camada de análise

Camada / Variável	Descrição	Direção inferida
Escola: Nível Socioeconômico	Indicador composto do perfil socioeconômico agregado dos alunos da escola (INEP).	Níveis mais elevados associados a melhor desempenho; preditor dominante.
Docente: Formação Adequada	Percentual de professores com formação adequada à disciplina lecionada.	Maior percentual correlaciona-se positivamente com o desempenho.
Escola: Taxa de Participação	Percentual de alunos do 9º ano que realizaram o SAEB.	Maior participação associada a maior engajamento escolar e melhor desempenho.
Discente: Escolaridade dos Pais	Nível de escolaridade da mãe e do pai ou responsável.	Níveis mais elevados correlacionam-se com melhor desempenho; categoria 'desconhecida' impacta negativamente.
Discente: Bens do Domicílio	Número de quartos e acesso a computador no domicílio do aluno.	Indicadores de maior capital econômico associados a melhor desempenho.
Docente: Experiência e Formação	Anos de experiência, grau acadêmico e participação em formação continuada.	Maior qualificação e experiência associadas positivamente ao desempenho.
Docente: Vínculo Trabalhista	Tipo de vínculo empregatício do professor com a rede de ensino.	Vínculos mais estáveis associados positivamente ao desempenho.
Gestão: Experiência do Diretor	Anos de experiência do diretor na função de liderança escolar.	Maior experiência de gestão associada a melhores indicadores escolares.

Fonte: elaborada pelos autores.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Figura 1. Sumário SHAP dos principais preditores do modelo Random Forest, por camada de análise (amostra de 20.000 registros do conjunto de teste)



Fonte: reproduzida de Tertulino e Alencar (2026), Discover Education, v. 5, art. 653, publicado sob licença Creative Commons Atribuição 4.0 Internacional (CC-BY).

3.5. Diálogo com a literatura sobre percepções docentes e discentes

O achado central deste estudo, quando lido pela chave docente-discente, é que as camadas discente e docente contribuem de formas distintas e complementares para a proficiência escolar: a camada discente concentra o maior poder preditivo individual, por meio da escolaridade dos pais e dos bens do domicílio, enquanto a camada docente contribui de forma mais modesta, porém consistente, por meio da formação, da experiência e do vínculo trabalhista dos professores. Essa hierarquia é compatível com o achado de Teixeira, Almeida e Cassuce (2022), para quem apenas uma pequena parcela das percepções docentes sobre obstáculos à

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

aprendizagem apresentou relação estatisticamente significativa com o desempenho dos alunos no SAEB: os dados aqui analisados sugerem que variáveis objetivas de perfil docente — e não necessariamente as percepções subjetivas dos professores sobre o contexto escolar — são as que efetivamente carregam poder preditivo no modelo multinível. Esse padrão também converge com Khine *et al.* (2024), para quem as variáveis de clima escolar percebido apresentaram menor poder preditivo do que as características objetivas dos estudantes e do status socioeconômico, social e cultural. Já o estudo de Miranda, Bertagna e Freitas (2019), de natureza qualitativa, indica que os próprios professores associam colaboração entre pares e disciplina em sala a um clima escolar mais favorável à aprendizagem — uma dimensão de processo que os microdados objetivos do SAEB analisados aqui não capturam diretamente, mas que é coerente com a constatação de que a estabilidade e a qualificação do corpo docente (variáveis objetivas de perfil) se associam positivamente ao desempenho.

Em segundo lugar, o peso do nível socioeconômico agregado da escola dialoga com a teoria do conflito aplicada à educação, elaborada por Bowles e Gintis (1976), que postula que as escolas reforçam as estruturas de classe existentes ao oferecer experiências educativas radicalmente distintas a estudantes de diferentes origens sociais. No plano internacional, esses achados encontram eco em grandes meta-análises que identificam o contexto socioeconômico entre os fatores mais influentes no desempenho escolar (HATTIE, 2008) e em avaliações do PISA, que consistentemente mostram a forte associação entre a origem social e o desempenho acadêmico em países de renda média e alta (OECD, 2023). A particularidade do caso brasileiro, revelada pelo modelo, é a intensidade dessa associação, mesmo quando as camadas docente e discente são examinadas separadamente da camada escolar agregada.

3.6. Implicações para formação e valorização docente e para o apoio discente

Do ponto de vista das políticas de formação e valorização docente, o principal *insight* é que intervenções centradas exclusivamente no estudante individual ou exclusivamente na percepção subjetiva dos professores sobre o clima escolar produzem impacto limitado em um sistema marcado por desigualdades estruturais entre escolas. As evidências objetivas de perfil docente e discente aqui analisadas reforçam a necessidade de políticas que atuem simultaneamente nas duas camadas: alocação equitativa de professores qualificados e com vínculo trabalhista estável, programas de formação continuada docente, e políticas de apoio socioeconômico e de engajamento familiar voltadas a contextos discentes vulneráveis. A abordagem multinível, integrando dados de alunos, professores, escolas e diretores em um único *pipeline* analítico, mantém-se como contribuição metodológica relevante para a área de MDE

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

aplicada à educação básica, ampliada aqui pela leitura dedicada às camadas docente e discente e pelo diálogo com a literatura sobre percepções (MIRANDA; BERTAGNA; FREITAS, 2019; TEIXEIRA; ALMEIDA; CASSUCE, 2022).

3.7. Limitações da abordagem

Algumas limitações merecem reconhecimento. A construção da variável-alvo com base na média do próprio *dataset* implica que a classificação 'Acima da Média' é relativa ao conjunto analisado, e não a um padrão absoluto de proficiência definido pela política educacional. O uso de *listwise deletion*, embora preserve a integridade dos registros analisados, pode introduzir viés de seleção se os registros com valores ausentes não forem distribuídos aleatoriamente entre as escolas. A análise concentra-se em um único corte temporal, o que impede inferências sobre causalidade ou sobre a dinâmica temporal dos fatores identificados. Os modelos de *gradient boosting* foram avaliados sem ajuste de hiperparâmetros, o que limita a comparação com o Random Forest a um cenário de configuração padrão. Por fim, e de forma central para o recorte deste trabalho, os blocos de percepção subjetiva dos questionários do professor e do aluno — concepções profissionais, condições de trabalho e clima escolar, no caso do questionário docente — não foram integrados ao modelo preditivo, que se apoia exclusivamente em variáveis objetivas de perfil. A incorporação futura dessas variáveis de percepção representa uma extensão natural e promissora deste trabalho, permitindo comparar diretamente, no mesmo *pipeline* multinível, o poder preditivo de fatores objetivos e subjetivos das camadas docente e discente.

4. CONSIDERAÇÕES FINAIS

Este trabalho revisitou, sob a ótica das camadas docente e discente, um modelo multinível de aprendizado de máquina para classificar a proficiência de estudantes do 9º ano e do Ensino Médio a partir dos microdados do SAEB 2023, estendendo a análise sistêmica original de Tertulino e Alencar (2026). O Random Forest apresentou desempenho expressivamente superior aos demais algoritmos comparados, com métricas equilibradas entre as duas classes-alvo. A interpretação via SHAP evidenciou que a camada discente (escolaridade dos pais e bens do domicílio) concentra maior poder preditivo individual do que a camada docente (formação, experiência e vínculo trabalhista dos professores), ainda que ambas permaneçam subordinadas ao nível socioeconômico médio da escola.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

A conclusão central do trabalho é que, no recorte docente e discente aqui proposto, o desempenho acadêmico na educação básica brasileira continua sendo, antes de tudo, um fenômeno sistêmico e coletivo, mas com contribuições distintas e mensuráveis de cada camada: a família e o aluno pesam mais na predição individual do que o professor, e as variáveis objetivas de perfil docente pesam mais do que as percepções subjetivas dos professores sobre o clima escolar, conforme sugerem estudos anteriores baseados no próprio SAEB (TEIXEIRA; ALMEIDA; CASSUCE, 2022). Esse achado, consistente com a teoria da reprodução cultural de Bourdieu e Passeron (1977) e com os resultados do relatório de Coleman *et al.* (1966), fornece evidência quantitativa de que políticas educacionais centradas apenas no estudante individual ou apenas na percepção docente do clima escolar são estruturalmente insuficientes para romper o ciclo de desigualdade educacional no país.

Do ponto de vista das contribuições para a área de Tecnologia Educacional, este trabalho demonstra como a leitura dedicada das camadas docente e discente de um modelo multinível de larga escala, combinada à IA Explicável e confrontada com a literatura sobre percepções docentes e discentes, pode transformar avaliações nacionais em instrumentos de diagnóstico sistêmico orientados à ação, tanto para a formação e valorização docente quanto para o apoio socioeconômico ao corpo discente. Como trabalhos futuros, destacam-se: a incorporação dos blocos de percepção subjetiva dos questionários do professor e do aluno do SAEB ao modelo preditivo, permitindo comparar diretamente fatores objetivos e subjetivos; a incorporação de dados longitudinais de múltiplas edições do SAEB para análise de tendências nas camadas docente e discente; o uso de técnicas avançadas de imputação para reduzir perdas por exclusão em lista; e a aplicação dos modelos de *boosting* com ajuste sistemático de hiperparâmetros para comparação equânime de desempenho.

REFERÊNCIAS

ALVES, M. T. G.; SOARES, J. F. Contexto escolar e indicadores educacionais: condições desiguais para a efetivação de uma política de avaliação educacional. Educação e Pesquisa, v. 39, n. 1, p. 177-194, 2013.

BENTÉJAC, C.; CSÖRGŐ, A.; MARTÍNEZ-MUÑOZ, G. A comparative analysis of *gradient boosting* algorithms. Artificial Intelligence Review, v. 54, n. 3, p. 1937-1967, 2021.

BONAMINO, A.; SOUSA, S. Z. Três gerações de avaliação da educação básica no Brasil: interfaces com o currículo da/na escola. Educação e Pesquisa, v. 38, n. 2, p. 373-388, 2012.

BOURDIEU, P. The forms of capital. In: RICHARDSON, J. (ed.). Handbook of theory and research for the sociology of education. New York: Greenwood, 1986. p. 241-258.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

BOURDIEU, P.; PASSERON, J. C. Reproduction in education, society and culture. London: Sage, 1977.

BOWLES, S.; GINTIS, H. Schooling in capitalist America: educational reform and the contradictions of economic life. New York: Basic Books, 1976.

BREIMAN, L. Random forests. Machine Learning, v. 45, n. 1, p. 5-32, 2001.

CARNEIRO, T. C. J. *et al.* Educational data mining applied to predict dropout in introductory programming courses in Brazil. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 33., 2022, Manaus. Anais [...]. Porto Alegre: SBC, 2022.

CHEN, T.; GUESTRIN, C. XGBoost: a scalable tree *boosting* system. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 22., 2016, San Francisco. Proceedings [...]. New York: ACM, 2016.

COLEMAN, J. S. *et al.* Equality of educational opportunity. Washington, D.C.: U.S. Department of Health, Education, and Welfare, 1966.

COLPO, M. *et al.* Tendências, oportunidades e desafios da mineração de dados educacionais para predição de evasão: uma revisão sistemática. RENOTE, v. 22, n. 1, 2024.

HATTIE, J. Visible learning: a synthesis of over 800 meta-analyses relating to achievement. London: Routledge, 2008.

INEP. Nota técnica: indicador de nível socioeconômico das escolas de educação básica. Brasília: INEP, 2021. Disponível em: https://www.gov.br/inep. Acesso em: 10 jan. 2025.

INEP. Microdados do SAEB 2023. Brasília: INEP, 2025. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/saeb. Acesso em: 15 jan. 2025.

JANG, Y. *et al.* A practical approach to explainable artificial intelligence for early prediction of academic performance. Expert Systems with Applications, v. 203, p. 117501, 2022.

KHINE, M. S.; LIU, Y.; PALLIPURAM, V. K.; AFARI, E. A machine-learning approach to predicting the achievement of Australian students using school climate, learner characteristics, and economic, social, and cultural status. Education Sciences, v. 14, n. 12, art. 1350, 2024.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 30., 2017, Long Beach. Proceedings [...]. Red Hook: Curran Associates, 2017.

MASTOUR, H. *et al.* Explainable AI *framework* for predicting medical students' performance in high-stakes assessments. Education and Information Technologies, 2025.

MELO, G. *et al.* Semi-supervised learning to improve dropout prediction with limited labeled data in a Brazilian context. In: BRAZILIAN SYMPOSIUM ON COMPUTERS IN EDUCATION, 34., 2023. Anais [...]. Porto Alegre: SBC, 2023.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

MIRANDA, A. C.; BERTAGNA, R. H.; FREITAS, L. C. Fatores que afetam o clima da escola: a visão dos professores. Pro-Posições, v. 30, e20160102, 2019.

OECD. PISA 2022 results: the state of learning and equity in education. Paris: OECD Publishing, 2023.

PAPADOGIANNIS, I. *et al.* A systematic review of educational data mining: concepts, techniques and applications. Journal of *Big Data*, v. 11, n. 1, 2024.

PEREIRA, F. D. *et al.* Deep learning for early dropout prediction in a Brazilian university programming course. Computers and Education: Artificial Intelligence, v. 1, 2020.

SALES, A. *et al.* Exploiting the use of recurrent neural networks for student performance prediction. In: INTERNATIONAL EDUCATIONAL DATA MINING SOCIETY CONFERENCE, 8., 2015. Proceedings [...]. 2015.

TEIXEIRA, O. H.; ALMEIDA, L. V.; CASSUCE, F. C. C. Percepções docentes e aprendizagem no ensino médio brasileiro: qual a relação? Estudos em Avaliação Educacional, v. 33, e07473, 2022.

TERTULINO, R.; ALENCAR, L. A multi-level analysis of factors associated with student performance: a machine learning approach to the SAEB microdata. Discover Education, v. 5, art. 653, 2026.

UJKANI, M. *et al.* Course success prediction and at-risk student identification using explainable AI. International Journal of Advanced Computer Science and Applications, v. 15, n. 2, 2024.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.