Fatores Socioeconômicos e Desempenho Escolar: uma análise explicável e preditiva do ENEM no Rio Grande do Norte

Rodrigo Tertulino, Ricardo Almeida, Laércio Alencar

https://doi.org/10.1590/SciELOPreprints.14701

Submetido em: 2025-12-23
Postado em: 2026-01-06 (versão 1)
(AAAA-MM-DD)

Fatores Socioeconômicos e Desempenho Escolar: uma análise explicável e preditiva do ENEM no Rio Grande do Norte

Socioeconomic Factors and Academic Performance: An Explainable and Predictive Analysis of ENEM in Rio Grande do Norte

RODRIGO TERTULINO[1]
ORCID: https://orcid.org/0000-0002-7594-9312
rodrigo.tertulino@ifrn.edu.br
RICARDO ALMEIDA[1]
ORCID: https://orcid.org/0009-0006-3447-7431
ricardo.almeida@ifrn.edu.br
LAÉRCIO ALENCAR[1]
ORCID: https://orcid.org/0009-0007-5226-9019
lapea@ifrn.edu.br

[1]Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte (IFRN)
Caixa Postal 59628-330 – Mossoró – RN – Brasil

Resumo. O Exame Nacional do Ensino Médio (ENEM) é um instrumento central para o acesso ao ensino superior no Brasil, mas seus resultados refletem profundas desigualdades. Compreender os fatores que determinam o desempenho, especialmente em nível regional, é fundamental. Contudo, modelos preditivos de Aprendizado de Máquina (ML), embora precisos, são frequentemente "caixas-pretas", o que limita seu valor para a formulação de políticas. Este estudo aborda essa lacuna ao aplicar técnicas de ML (Random Forest) e de Inteligência Artificial Explicável (XAI), via SHAP, para analisar o desempenho de 51.091 estudantes do Rio Grande do Norte (RN) no ENEM 2022. Os resultados confirmam que os fatores socioeconômicos são os preditores mais influentes. A análise SHAP identificou como variáveis-chave, notadamente a escolaridade dos pais, a renda familiar e o tipo de escola (pública vs. privada), que criam um "abismo de oportunidades" e polarizam os resultados. Ao desvendar o modelo preditivo, este trabalho fornece evidências regionais robustas sobre os determinantes da desigualdade educacional, oferecendo subsídios para o desenvolvimento de políticas de equidade voltadas a mitigar o impacto da origem socioeconômica no futuro acadêmico dos estudantes.
Palavras-chave: Aprendizado de Máquina; Inteligência Artificial Explicável (XAI); Desigualdade Educacional; ENEM; Rio Grande do Norte.

Abstract. The National High School Exam (ENEM) is a central instrument for access to higher education in Brazil, but its results reflect profound inequalities. Understanding the factors that determine performance, especially at a regional level, is crucial. However, predictive Machine Learning (ML) models, while accurate, are often "black boxes," limiting their value for policymaking. This study addresses this gap by applying ML (Random Forest) and Explainable Artificial Intelligence (XAI) techniques, via SHAP, to analyze the performance of 51,091 students from Rio Grande

do Norte (RN) in the 2022 ENEM. The results confirm that socioeconomic factors are the most influential predictors. The SHAP analysis quantified how key variables, notably parental education, family income, and school type (public vs. private), create an "opportunity gap" and polarize the results. By unpacking the predictive model, this work provides robust regional evidence on the determinants of educational inequality, offering a basis for developing equity policies that focus on mitigating the impact of socioeconomic origin on students' academic futures.

Keywords: Machine Learning, Explainable Artificial Intelligence (XAI), Educational Inequality, ENEM, Rio Grande do Norte.

INTRODUÇÃO

A avaliação do desempenho educacional em larga escala constitui um pilar fundamental para o diagnóstico e a formulação de políticas públicas voltadas à melhoria da qualidade da educação. No Brasil, o Exame Nacional do Ensino Médio (ENEM) emerge como um indicador crucial, não apenas por mensurar conhecimentos e habilidades ao término da educação básica, mas também por ser a principal porta de entrada para o ensino superior (MEC, 2022). Contudo, os resultados do ENEM historicamente refletem profundas disparidades socioeconômicas e regionais, nas quais o desempenho dos estudantes frequentemente se correlaciona com seu background familiar, o tipo de escola frequentada e o acesso a recursos educacionais (Alves; Soares, 2013). Essas desigualdades representam obstáculos significativos à equidade no acesso a oportunidades educacionais e profissionais (Ferrão et al., 2001).

Compreender os fatores que influenciam o desempenho no ENEM é, portanto, essencial para direcionar intervenções pedagógicas e políticas mais eficazes. Abordagens estatísticas tradicionais têm sido amplamente utilizadas para identificar correlações (Soares; Alves, 2003). No entanto, o advento de técnicas computacionais, notadamente da Mineração de Dados Educacionais (EDM - *Educational Data Mining*) e do Aprendizado de Máquina (*Machine Learning* - ML), oferece novas perspectivas para modelar as complexas inter-relações entre múltiplos fatores e prever o desempenho discente com maior acurácia (Baker; Yacef, 2009; Romero; Ventura, 2010). Além da predição, torna-se cada vez mais relevante a interpretabilidade desses modelos, buscando não apenas saber "**quem**" pode ter melhor ou pior desempenho, mas também "**por quê**" (Molnar, 2020).

Nesse contexto, este estudo visa desenvolver e interpretar um modelo preditivo de Machine Learning para o desempenho dos estudantes do Rio Grande do Norte (RN) no ENEM 2022. O foco reside na identificação dos fatores socioeconômicos, oriundos do questionário do participante, que se mostram mais relevantes para a predição da nota média final. Utilizou-se uma amostra representativa de estudantes do RN, aplicando-se o algoritmo para uma seleção robusta de variáveis e treinando-se modelos de regressão, com destaque para o Random Forest (Manouselis et al., 2011). A significância estatística das variáveis selecionadas foi aferida por meio de Regressão por Mínimos Quadrados Ordinários (OLS), e a interpretabilidade do modelo final foi explorada por meio da técnica SHAP (*SHapley Additive exPlanations*) (Al-Shabandar et al., 2021).

A contribuição deste estudo reside na aplicação de um fluxo metodológico que combina seleção de variáveis, modelagem preditiva e interpretabilidade para analisar

um contexto educacional específico do estado do Rio Grande do Norte, fornecendo insights quantificáveis sobre o impacto das variáveis socioeconômicas no desempenho do ENEM. Espera-se que os resultados possam subsidiar discussões e potenciais ações educacionais focadas nas particularidades regionais identificadas. O restante do artigo está estruturado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados; a Seção 3 descreve a caracterização da amostra e o contexto nacional; a Seção 4 detalha a metodologia empregada; a Seção 5 discute os resultados obtidos; e a Seção 6 conclui o trabalho e aponta direções futuras.

TRABALHOS RELACIONADOS

A análise de dados do ENEM para investigar fatores associados ao desempenho acadêmico é um campo consolidado na pesquisa educacional brasileira. Historicamente, estudos seminais utilizaram abordagens estatísticas, como a regressão linear múltipla, para demonstrar a forte correlação entre o desempenho no exame e variáveis socioeconômicas, como renda familiar, escolaridade dos pais e tipo de escola (pública vs. privada) (Soares; Alves, 2003; Alves et al., 2015). Estes trabalhos fundamentais destacaram a importância do capital econômico e cultural como determinante do sucesso educacional. Contudo, uma limitação inerente a esses modelos lineares é a dificuldade em capturar interações complexas e efeitos não lineares presentes em conjuntos de dados de alta dimensionalidade.

Com o avanço da Mineração de Dados Educacionais (EDM) e do Aprendizado de Máquina (ML), pesquisadores passaram a aplicar algoritmos mais sofisticados para superar essas limitações. O campo de EDM, em especial, concentra-se no desenvolvimento de novas técnicas para analisar dados de ambientes educacionais (Papadogiannis et al., 2024). Essa transição metodológica é impulsionada pela crescente percepção de que os modelos de ensino tradicionais, de abordagem única, não conseguem atender às necessidades individuais de aprendizagem dos alunos (Taufikin et al., 2024; Tang; Chen, 2024). Esse interesse é quantificável, com um aumento significativo nas publicações sobre ML na educação a partir de 2017 (Alan et al., 2025).

Dessa forma, técnicas como Árvores de Decisão, Redes Neurais, Support Vector Machines (SVM) e ensembles (como Random Forest e Gradient Boosting) (Romero; Ventura, 2013; Costa et al., 2017) são agora empregadas em uma variedade de tarefas. Tais aplicações incluem a predição do sucesso estudantil (Pereira et al., 2020), a criação de sistemas de tutoria inteligente, plataformas de aprendizagem adaptativas (Lima Júnior; Silva, 2022) e a oferta de experiências de ensino personalizadas que se ajustam ao ritmo de cada aluno (Ravuri et al., 2023). No contexto específico do ENEM, trabalhos anteriores que exploraram ML para predição de notas também confirmaram a relevância das variáveis socioeconômicas nesses modelos (Amaral; Rigo 2018, Guedes et al., 2019).

Contudo, essa riqueza de dados introduz um primeiro desafio metodológico: a modelagem com *datasets* ricos em variáveis, como os microdados do ENEM, torna a seleção de *features* um obstáculo. A inclusão de variáveis redundantes pode prejudicar o desempenho e, principalmente, a interpretabilidade do modelo. Embora métodos de seleção (filtros, wrappers, embarcados) sejam frequentemente aplicados (Guyon; Elisseeff, 2003), o algoritmo Boruta (Kursa; Rudnicki, 2010), utilizado neste trabalho, destaca-se. Ao comparar estatisticamente a importância de variáveis reais com a de

variáveis "sombra" (aleatorizadas), o Boruta oferece uma seleção mais confiável, e sua eficácia foi demonstrada em EDM (Márquez-Vera et al., 2016).

Uma segunda lacuna, mais crítica, reside na interpretabilidade. Modelos de ML de alto desempenho, como Random Forest, são frequentemente considerados "caixas pretas" (Adadi; Berrada, 2018). No campo educacional, em que o objetivo final é subsidiar políticas públicas, entender "**por que**" um modelo prevê um desempenho é tão ou mais importante do que a acurácia da predição. Isso se deve à necessidade de garantir a equidade (fairness) nas decisões algorítmicas, permitindo a automação de avaliações para reduzir o viés humano (Ravuri et al., 2023) e promover a tomada de decisão baseada em dados justa (Baidoo-Anu; Ansah, 2023). Métodos de IA Explicável (XAI), como LIME e SHAP (Lundberg; Lee, 2017), ganham destaque. OSHAP, em particular, oferece uma base teórica sólida para atribuir a contribuição de cada variável, mas suas aplicações em EDM ainda são emergentes (Al-Shabandar et al., 2021).

O presente trabalho, portanto, insere-se diretamente nesse contexto, posicionando-se para preencher as lacunas identificadas. Nossa contribuição original reside na **combinação** de um método robusto de seleção de variáveis (Boruta) com um modelo preditivo de alto desempenho (Random Forest) e uma técnica avançada de interpretabilidade (SHAP). Ao aplicar este pipeline aos microdados do ENEM 2022 de um estado específico (RN), buscamos não apenas criar um modelo preditivo acurado, mas também oferecer uma análise detalhada e fundamentada dos fatores socioeconômicos que influenciam o desempenho, justificando sua relevância para o debate sobre políticas educacionais.

CARACTERIZAÇÃO DA AMOSTRA E DO CONTEXTO NACIONAL

Antes de detalhar a metodologia de modelagem, é relevante caracterizar a amostra de estudo e contextualizar o desempenho do Rio Grande do Norte (RN) no cenário nacional do ENEM 2022. Dos 88.049 inscritos que realizaram a prova no RN, 65.475 estiveram presentes em pelo menos um dia. Aplicando-se os critérios de inclusão para a análise preditiva (não ser treineiro, presença em todas as provas objetivas e nota válida na redação), obteve-se uma amostra final composta por 51.091 estudantes do estado. Estes participantes realizaram o exame em 40 municípios distintos do RN.

A Tabela 1 apresenta as principais características desta amostra final. Observa-se que a maioria dos participantes declarou não responder ao tipo de escola do Ensino Médio (**62,91%**), enquanto **29,20%** indicaram escola pública e **7,89%** escola privada. Quase a totalidade (**99,50%**) concluiu o Ensino Médio na modalidade regular. Em relação ao desempenho, a nota média final da amostra do RN foi de **547,15**, com desvio padrão de **81,32**. As notas médias por componente variaram, sendo a maior observada em Redação (**651,58**) e a menor em Ciências da Natureza (**498,62**). A análise do status da redação indicou que **100%** das redações na amostra final foram consideradas "Sem problemas", o que reflete o critério de filtro de nota maior que zero.

Tabela 1 - Estatísticas Descritivas da Amostra Final do RN (ENEM 2022, N=51.091) ***Fonte: Elaboração própria.***

Característica	Valor / Distribuição
Nº de Alunos	51.091
Nº Municípios de Prova	40

Tipo de Escola (Aluno) Não Respondeu	62,91%
Pública	29,20%
Privada	7,89%
Tipo Ensino Médio (Aluno) Regular	99,50%
Ed. Especial (Subst.)	0,50%
Status Redação Sem problemas	100,00%
Notas Médias (0-1000)	**Média (Desv. Padrão)**
Ciências da Natureza	498,62 (72,85)
Ciências Humanas	529,44 (80,25)
Linguagens e Códigos	519,14 (75,67)
Matemática	536,97 (115,83)
Redação	651,58 (161,68)
Nota Final Média	**547,15 (81,32)**

Para contextualizar o desempenho do RN, a **Figura 1** apresenta a nota média final por estado, calculada com base na mesma amostra de alunos válidos em nível nacional. A média nacional foi de **551,27**. O Rio Grande do Norte, com média de **547,15**, posicionou-se ligeiramente abaixo da média nacional, ocupando a 10ª posição no ranking dos estados brasileiros em 2022. Esta visualização auxilia na compreensão da performance relativa do estado, servindo de pano de fundo para a análise dos fatores socioeconômicos que influenciam as notas no RN, a qual será detalhada nas seções subsequentes, utilizando técnicas como Boruta, OLS e SHAP.

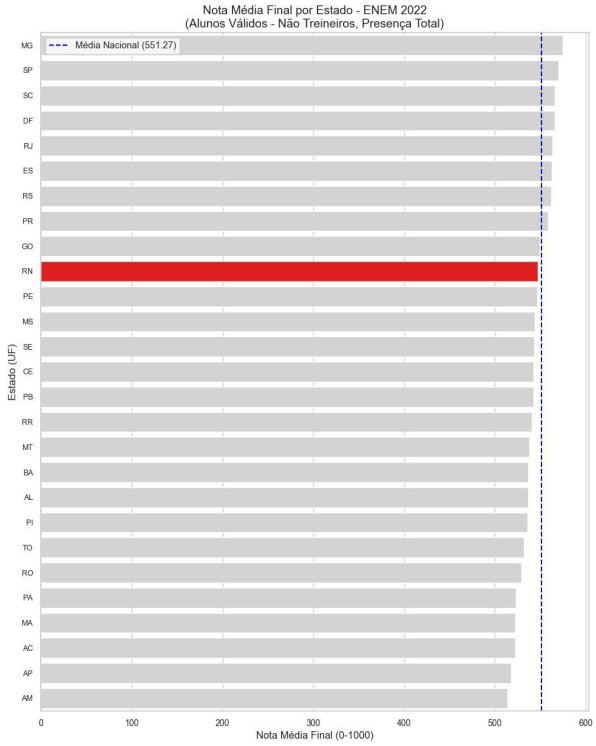

Figura 1 - Comparativo da Nota Média Final por Estado (ENEM 2022). A linha tracejada indica a média nacional. ***Fonte: Elaboração própria.***

METODOLOGIA

A estrutura metodológica adotada neste estudo foi organizada em um pipeline sequencial, conforme ilustrado na **Figura 2**. O processo foi dividido em quatro etapas principais, detalhadas a seguir:

1. **Preparação dos Dados:** Esta etapa iniciou-se com a aquisição dos microdados públicos do ENEM 2022 (INEP, 2024). A amostra foi, então, filtrada para focar nos participantes do Rio Grande do Norte (RN), resultando em N=51.091 concluintes (não treineiros) que estavam presentes em todas as provas objetivas e possuíam notas válidas (incluindo a redação, maior que zero). A variável-alvo foi definida como a nota média final, calculada com base nas cinco notas do exame. Como preditoras, consideraram-se as 26 variáveis categóricas do questionário socioeconômico e o tipo de escola. Estas foram pré-processadas por meio de imputação de valores ausentes e transformação por *One-Hot Encoding* (OHE) (Guyon; Elisseeff , 2003), gerando 144 variáveis binárias.
2. **Seleção e Modelagem:** Para identificar os preditores mais relevantes no universo de 144 variáveis, o algoritmo Boruta (Kursa; Rudnicki, 2010) foi aplicado ao conjunto de treino (75% da amostra). Este método robusto selecionou 47 variáveis como estatisticamente significativas. Com base neste subconjunto, foram treinados modelos de regressão por Machine Learning (Random Forest e XGBoost) (Chen; Guestrin, 2016) para prever a nota final, sendo o OLS utilizado como ferramenta de validação inferencial.
3. **Avaliação e Interpretação:** O desempenho do modelo principal (Random Forest) foi avaliado sob diferentes perspectivas. Primeiro, como problema de regressão, alcançou um $R^2 \approx 0{,}32$ e um RMSE de 66,53 pontos no conjunto de teste, indicando capacidade preditiva moderada. Segundo, em uma classificação adaptada (prever se o aluno estava acima ou abaixo da média de 547,68), o modelo obteve Accuracy de 70,6% e AUC-ROC de 0,77. Terceiro, a significância estatística dos preditores foi validada por meio de um modelo de Regressão Linear Múltipla (OLS) (Molnar, 2020), no qual 28 das 31 variáveis analisadas (após correção de multicolinearidade) apresentaram p-valor $< 0{,}05$. O modelo Random Forest foi selecionado como o principal para interpretação por apresentar um equilíbrio ideal entre desempenho e robustez. Embora o XGBoost tenha apresentado valores de R^2 similares, o Random Forest demonstrou menor tendência ao *overfitting* e sua estrutura em *ensemble* é particularmente adequada para a análise de SHAP em dados tabulares complexos (Cha et al., 2021). Finalmente, para compreender como o modelo tomava suas decisões, a interpretabilidade foi investigada por meio da técnica SHAP (Lundberg; Lee, 2017).
4. **Resultados e Conclusão:** As saídas das três avaliações (preditiva, classificatória e inferencial) e da análise SHAP convergiram em torno da discussão central dos preditores socioeconômicos, permitindo a formulação das conclusões e a discussão de subsídios para políticas públicas.

Figura 2 - Fluxograma metodológico (pipeline) da pesquisa, desde a preparação dos dados até a análise dos resultados. ***Fonte: Elaboração própria.***

Portanto, a abordagem metodológica adotada neste trabalho representa uma fusão de técnicas de ciência de dados para responder a uma questão central da sociologia da educação: **a determinação de como e em que medida as desigualdades socioeconômicas estruturais influenciam o desempenho acadêmico**. Esta abordagem alinha-se ao crescente campo da Ciência Social Computacional, que aplica métodos de ML para analisar questões sociais complexas e explorar ativamente questões de equidade e justiça na educação, tornando visíveis as desigualdades estruturais (Salganik et al., 2019; Selwyn; Gašević, 2025).

Ao integrar a seleção robusta de variáveis (Boruta), a modelagem preditiva de alto desempenho (Random Forest) e a interpretabilidade avançada (SHAP), o estudo transcende a simples predição de notas. Esta combinação, validada estatisticamente por meio de OLS e de métricas de classificação, permite uma análise detalhada e quantitativa dos fatores socioeconômicos que influenciam o desempenho no ENEM no contexto do Rio Grande do Norte, oferecendo, assim, uma base de evidências robusta para a discussão da equidade educacional.

Considerações Éticas

Este trabalho foi conduzido em estrita conformidade com as normas de integridade na pesquisa. Foram utilizados exclusivamente os microdados públicos do ENEM 2022, disponibilizados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Todos os dados são anonimizados na origem pelo INEP, não contendo informações de identificação pessoal dos participantes, garantindo, assim, a confidencialidade e o cumprimento da Lei Geral de Proteção de Dados (LGPD).

RESULTADOS E DISCUSSÃO

A modelagem preditiva do desempenho final dos estudantes do RN no ENEM 2022, utilizando as 47 variáveis socioeconômicas e contextuais selecionadas pelo Boruta, resultou em um modelo Random Forest com capacidade explicativa moderada ($R^2 \approx$

0,32) e um erro médio de predição (RMSE) de aproximadamente 66,53 pontos. Embora o $_R^2$ indique que uma parte significativa da variância nas notas é explicada por outros fatores não contemplados (como esforço individual, qualidade específica do ensino recebido, etc.), a análise de interpretabilidade por meio de SHAP revelou padrões consistentes e significativos.

A **Figura 3** apresenta o gráfico de sumário SHAP das 20 variáveis de maior impacto médio absoluto no modelo. Este gráfico mapeia a contribuição de cada variável para a nota prevista, em que a posição horizontal indica a magnitude e a direção do impacto (valores SHAP positivos aumentam a nota, negativos diminuem).

A análise visual confirma que as variáveis relacionadas ao capital econômico e cultural familiar dominam o topo do ranking de importância. Fica evidente a forte polarização socioeconômica dos resultados: fatores associados a maior capital (e.g., escolaridade superior dos pais, ocupação qualificada, escola privada) consistentemente geram valores SHAP positivos, empurrando a nota prevista para cima. Inversamente, fatores associados à vulnerabilidade (e.g., baixa renda, baixa escolaridade parental, ausência de computador) geram valores SHAP fortemente negativos, reduzindo a nota prevista.

- **Capital Cultural e Ocupacional (Pais):** A análise SHAP revelou que o capital cultural e ocupacional da família está entre os preditores de maior impacto positivo. Especificamente, a escolaridade materna elevada (e.g., Q002_E - 'Completou Ensino Médio, mas não completou a Faculdade' e Q002_F - 'Completou a Faculdade') e, de forma muito impactante, a ocupação dos pais em níveis técnicos,

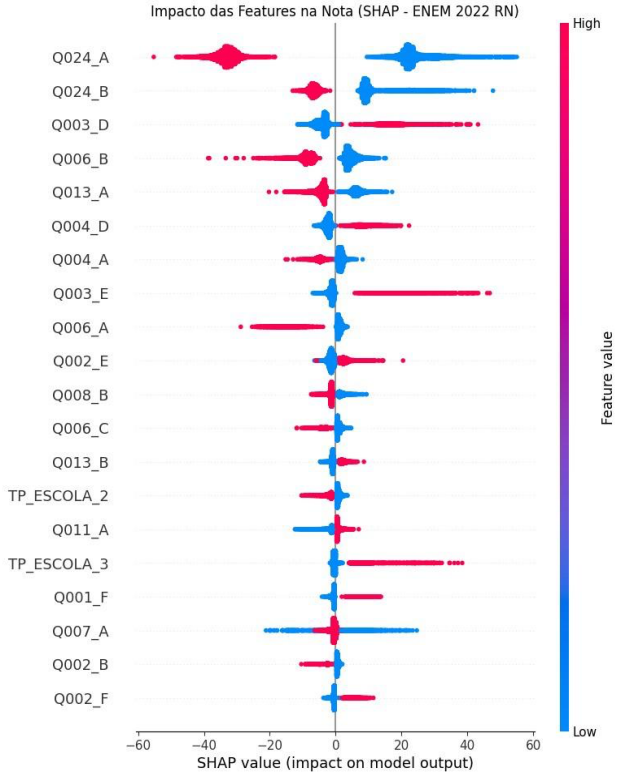

Figura 3 - Impacto das 20 variáveis mais importantes na predição da Nota Média Final (SHAP). Valores SHAP positivos aumentam a nota prevista.
Fonte: Elaboração própria.

gerenciais ou superiores (e.g., Q003_D, Q003_E, Q004_D - 'Grupo 4' ou 'Grupo 5') contribuem consistentemente para predições de notas mais elevadas. Em contrapartida, a baixíssima escolaridade materna (Q002_B - 'Não completou a 4ª série') e as ocupações ligadas ao trabalho rural/manual (Q004_A - 'Grupo 1') figuram entre os maiores impactos negativos, reforçando a influência do capital cultural familiar (Moraes; Peres, 2022).

- **Capital Econômico (Renda e Bens):** O capital econômico, medido diretamente pela renda e por *proxies* de posse de bens, demonstrou forte influência. Pertencer às faixas de renda mais baixas (Q006_A e Q006_B, até R$ 1.818,00) teve um impacto negativo substancial. A posse de bens também se mostrou um forte indicador: a ausência de freezer (Q013_A) e a posse de apenas um banheiro (Q008_B) figuraram como fortes preditores negativos. Em contraste, a simples posse de pelo menos 1 freezer (Q013_B) emergiu como um dos preditores positivos mais significativos, atuando como um claro indicador de maior poder aquisitivo e de estabilidade socioeconômica.
- **Contexto Escolar e Recursos Digitais:** O contexto escolar e o acesso a ferramentas de estudo mostraram-se decisivos. Frequentar escola privada (TP_ESCOLA_3) emergiu como um dos principais fatores positivos, enquanto a escola pública (TP_ESCOLA_2) associou-se a um impacto negativo. Paralelamente, o "abismo digital" revelou-se um preditor crítico: a ausência total de computadores em casa ('Nenhum computador', Q024_A) teve um dos maiores impactos negativos no desempenho. A variável 'Apenas 1 computador' (Q024_B) também impactou negativamente a nota, porém de forma substancialmente menor do que a ausência total. Este achado sugere que o acesso mínimo (um computador) mitiga parte da severa desvantagem decorrente da exclusão digital completa, reforçando a importância do acesso a recursos tecnológicos para o desempenho (Melo et al., 2021).

Para além da interpretabilidade do modelo principal, a robustez dos achados foi validada por múltiplas frentes. Primeiramente, a **significância estatística** das variáveis foi confirmada por meio de uma análise de regressão OLS (Mínimos Quadrados Ordinários). Este modelo demonstrou que a influência das variáveis socioeconômicas centrais não é fruto do acaso (majoritariamente $p < 0,05$). Para complementar a análise de interpretabilidade do Random Forest, um modelo de Regressão Linear Múltipla (OLS) foi ajustado para validar a significância estatística dos preditores selecionados. Os resultados, detalhados no **Apêndice A (Tabela 2)**, confirmam a robustez dos achados. Das 31 variáveis analisadas, 28 (ou 90%) apresentaram significância estatística ($p < 0.05$), o que demonstra que sua relação com a nota média final não é aleatória.

Adicionalmente, avaliou-se a capacidade preditiva do modelo em um cenário de classificação (alunos acima ou abaixo da média do conjunto de treino). A análise da Curva ROC (Receiver Operating Characteristic), apresentada na **Figura 4**, revelou uma Área sob a Curva (AUC) de 0,7708. Tal valor, significativamente acima de 0,5, reforça que as variáveis socioeconômicas selecionadas contêm informação relevante e possuem capacidade discriminatória útil para diferenciar os níveis de desempenho.

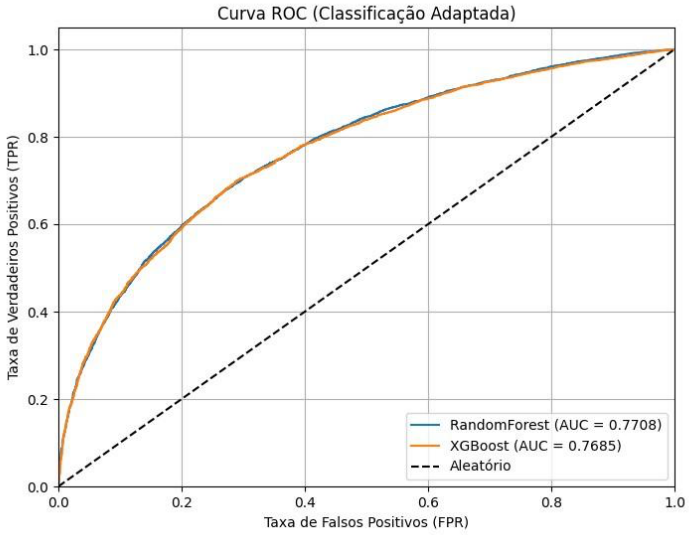

Figura 4 - Curva ROC para a classificação (Acima/Abaixo da Média) do modelo RandomForest no conjunto de teste. ***Fonte: Elaboração própria.***

Em síntese, os resultados da modelagem (Random Forest), da análise de interpretabilidade (SHAP), da validação estatística (OLS) e da capacidade discriminatória (Curva ROC) convergem. Todos indicam que, para a amostra de estudantes do RN no ENEM 2022, os fatores socioeconômicos são, de fato, os preditores mais influentes do desempenho. Este diagnóstico local quantifica a magnitude dessas influências no contexto regional e confirma que os desafios estruturais persistem, ecoando o que estudos de âmbito nacional (e.g., Agência Bori, 2021; Jornal da USP, 2024) já apontavam. Os dados fornecem, portanto, subsídios urgentes para a discussão de políticas de equidade educacional voltadas a mitigar o peso da origem socioeconômica no futuro acadêmico dos estudantes.

CONCLUSÃO E TRABALHOS FUTUROS

Este trabalho propôs uma abordagem metodológica híbrida, combinando Machine Learning (Random Forest) e interpretabilidade (SHAP), com validação estatística (OLS e Curva ROC). O objetivo central foi identificar e quantificar o impacto de fatores preditivos socioeconômicos e contextuais no desempenho dos estudantes do Rio Grande do Norte (RN) no ENEM 2022, oferecendo uma análise detalhada em um cenário regional específico.

Os principais achados corroboram, para a amostra do RN, uma tendência sistêmica já documentada em nível nacional: o desempenho no exame é profundamente influenciado pela origem socioeconômica. Especificamente, a análise de interpretabilidade (SHAP) foi crucial para demonstrar que o capital econômico (renda familiar), o capital cultural (escolaridade dos pais, notadamente da mãe) e a rede de ensino frequentada (pública vs. privada) são os preditores dominantes. Adicionalmente, fatores como o acesso a recursos básicos e tecnológicos (e.g., computador) também revelaram impacto relevante. Dessa forma, os resultados não apenas quantificam a magnitude do “abismo de oportunidades” no contexto regional, alinhando-se a uma vasta literatura nacional sobre o tema (Travitzki, 2013; Ribeiro, 2010), mas também

reforçam a tese de que o exame reflete, em grande medida, as desigualdades estruturais da sociedade brasileira (Hasenbalg; Silva, 2003).

No âmbito das políticas públicas, os achados do SHAP oferecem subsídios direcionáveis. Por exemplo, a forte associação negativa da variável Q024_A ('Nenhum computador') com o desempenho reforça a urgência de programas estaduais de inclusão digital e de fornecimento de equipamentos no RN. Similarmente, a forte influência positiva da escolaridade parental (Q002_E/F) sugere que políticas de longo prazo focadas na Educação de Jovens e Adultos (EJA) podem ter um efeito intergeracional indireto, porém significativo, no desempenho dos futuros estudantes.

Reconhecemos, contudo, algumas limitações. O poder explicativo do modelo ($R^2 \approx 0{,}32$) indica que uma parcela substancial da variância nas notas não é capturada pelas 47 variáveis analisadas. Fatores como o esforço individual, a motivação do aluno, a qualidade pedagógica específica da escola e a formação docente, variáveis não disponíveis nos microdados do ENEM, certamente desempenham um papel crucial. Além disso, este estudo tem natureza correlacional, identificando fortes associações, mas não estabelece causalidade direta.

Como trabalhos futuros, sugerimos caminhos para superar essas limitações e aprofundar a análise. Primeiramente, propõe-se o enriquecimento do conjunto de dados por meio da integração dos microdados do ENEM com outras bases, como o Censo Escolar, permitindo a inclusão de variáveis de infraestrutura escolar e de perfil docente. Em segundo lugar, a aplicação de modelos longitudinais, acompanhando coortes de estudantes ao longo de vários anos, permitiria uma análise de "valor agregado", buscando isolar o efeito-escola do efeito-origem. Por fim, análises qualitativas ou estudos de caso focados em escolas com desempenho atípico (alto desempenho em contextos vulneráveis ou baixo desempenho em contextos favorecidos) no RN poderiam revelar os mecanismos e processos específicos que mitigam ou potencializam as desigualdades socioeconômicas no chão da escola.

Declaração de contribuição dos autores

Rodrigo Tertulino: Conceitualização, Metodologia, Software, Análise Formal, Escrita – rascunho original. Ricardo Almeida: Supervisão, Escrita – revisão e edição. Laércio Alencar: Curadoria de dados, Visualização, Escrita – revisão e edição.

Declaração de conflito de interesse

Os autores declaram que não há conflito de interesse.

Declaração de disponibilidade de dados da pesquisa

Todo o conjunto de dados de apoio aos resultados deste estudo foi publicado no próprio artigo e baseia-se em microdados públicos anonimizados e disponibilizados pelo INEP em https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem.

REFERÊNCIAS

ADADI, A. and BERRADA, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160.

LIMA JUNIOR, Afonso Barbosa de; SILVA, Lebiam Tamar Gomes. Os sistemas tutores inteligentes e a adaptação do ensino aos perfis de aprendizagem dos usuários. Etd - Educ. Temat. Digit., Campinas , v. 24, n. 3, p. 618-632, jul. 2022.

AGÊNCIA BORI (2021). Enem: Renda familiar, acesso a bolsas e nível de educação das mães estão entre os fatores que mais afetam o desempenho dos alunos.

AL-SHABANDAR, R., JADDOA, A., ALZUBAIDI, L., and HUSSAIN, A. (2021). Explainable AI for educational data mining-based student performance prediction. *Procedia Computer Science*, 192:3011–3019.

ALAN, A., KARABATAK, S., and KARABATAK, M. (2025). The role of machine learning in distance education. In *Conference: 2025 13th International Symposium on Digital Forensics and Security (ISDFS)*, pages 1–6.

ALVES, M. T. G. and SOARES, J. F. (2013). Contexto escolar e indicadores educacionais: condições desiguais para a efetivação de uma política de avaliação educacional. *Educação e Pesquisa*, 39(1):177–194.

ALVES, M. T. G., SOARES, J. F., and XAVIER, F. P. (2015). Desigualdades educacionais no ensino médio: Habilidades e conhecimentos em português e matemática. *Educação & Sociedade*, 37(134):1–28.

AMARAL, M. V. M. and RIGO, S. J. (2018). Predição de desempenho no Enem usando árvore de decisão: uma análise comparativa por área do conhecimento. Em Anais do XXIX Simpósio Brasileiro de Informática na Educação (SBIE), páginas 1198–1207. Sociedade Brasileira de Computação (SBC).

BAIDOO-ANU, D. and ANSAH, L. O. (2023). Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, 7(1):52–62.

BAKER, R. S. J. d. and YACEF, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1):3–17.

CHA, G.-W., MOON, H.-J., and KIM, Y.-C. (2021). Comparison of random forest and gradient boosting machine models for predicting demolition waste based on small datasets and categorical variables. *International Journal of Environmental Research and Public Health*, 18(16).

CHEN, T. and GUESTRIN, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM.

COSTA, E. B., FONSECA, B., SANTANA, M. A., DE ARAÚJO, F. F., and REGO, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior*, 73:247–256.

FERRÃO, M. E., BELTRÃO, K. I., SANTOS, D. P. d., PAIM, R. d. O., and FERRÃO, M. M. S. (2001). Políticas de não repetência escolar e qualidade da educação: um

estudo em duas escolas públicas do Rio de Janeiro. *Estudos em Avaliação Educacional*, 12(24):61–84.

GUEDES, A. L. S., COELHO, F. A. F., NETO, A. C. A., and COSTA, T. V. (2019). Predição de desempenho no Enem com técnicas de aprendizado de máquina. Em Anais do VIII Congresso Brasileiro de Informática na Educação (CBIE), páginas 1475–1484. Sociedade Brasileira de Computação (SBC).

GUYON, I. and ELISSEEFF, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182.

HASENBALG, C. and SILVA, N. d. V. (2003). *Origens e Destinos: Desigualdades Sociais ao Longo da Vida*. Topbooks, Rio de Janeiro.

INEP (2024). Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, Microdados do saeb 2023. Accessed: 2/8/2025.

JORNAL DA USP (2024). Estudantes negros e de baixa renda têm desempenho prejudicado no Enem, constata estudo.

KURSA, M. B. and RUDNICKI, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software*, 36(11):1–13.

LUNDBERG, S. M. and LEE, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 4765–4774.

MANOUSELIS, N., DRACHSLER, H., VERBERT, K., and DUVAL, E. (2011). Recommender systems in technology-enhanced learning. In Ricci, F., Rokach, L., Shapira, B., and Kantor, P. B., editors, *Recommender Systems Handbook*, pages 387–415. Springer US.

MÁRQUEZ-VERA, C., CANO, A., ROMERO, C., NOAMAN, A. Y. M., FARDOUN, H. M., and VENTURA, S. (2016). Early dropout prediction using data mining: A case study with high school students. *Expert Systems*, 33(1):107–124.

MELO, R. O., FREITAS, A. C. d., FRANCISCO, E. d. R., and MOTOKANE, M. T. (2021). Impacto das variáveis socioeconômicas no desempenho do Enem: uma análise espacial e sociológica. *Revista de Administração Pública*, 55(6):1271–1294.

MINISTÉRIO DA EDUCAÇÃO (MEC) (2022). Edital nº 33, de 28 de abril de 2022, Exame Nacional do Ensino Médio - Enem 2022. Diário Oficial da União. Seção 3, p. 88.

MOLNAR, C. (2020). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. GitHub. Accessed on 30/10/2025.

MORAES, C. and PERES, R. (2022). ReflexÕes sobre diferenÇas de desempenho no Enem: Uma anÁlise socioeconÔmica e escolar do Sudeste do Brasil. *Jornal de Políticas Educacionais*, 16.

PAPADOGIANNIS, I., WALLACE, M., and KAROUNTZOU, G. (2024). Educational data mining: A foundational overview. *Encyclopedia*, 4(4):1644–1664.

PEREIRA, F. D., FONSECA, S. C., OLIVEIRA, E. H. T., OLIVEIRA, D. B. F., CRISTEA, A. I., and CARVALHO, L. S. G. (2020). Deep learning for early performance prediction of introductory programming students: a comparative and explanatory study. *Revista Brasileira de Informática na Educação - RBIE*, 28:723–749.

RAVURI, A., LOURENS, M., ASWINI, S., NIJHAWAN, G., ZABIBAH, R. S., and CHANDRASHEKAR, R. (2023). Improving personalized education: A machine learning method for flexible learning environments. In *2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics, and Computer Engineering (UPCON)*, volume 10, pages 1715–1720.

RIBEIRO, C. A. C. (2010). Desigualdade de oportunidades e a efetividade das escolas. *Dados - Revista de Ciências Sociais*, 53(1):149–178.

ROMERO, C. and VENTURA, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618.

ROMERO, C. and VENTURA, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1):12–27.

SALGANIK, M. J., LUNDBERG, I., and STEWART, B. (2019). Machine learning for sociology. *Annual Review of Sociology*, 45(1):27–45.

SELWYN, N. and GAŠEVIĆ, D. (2025). Computational social science and critical studies of education and technology: an improbable combination? *Learning, Media and Technology*, 50(1):1–15.

SOARES, J. F. and ALVES, M. T. G. (2003). Desigualdades raciais no sistema brasileiro de educação básica. *Educação e Pesquisa*, 29(1):147–165.

TANG, X. and CHEN, Y. (2024). Adaptive education platform based on machine learning: A new way to improve the quality of higher education. In *2024 International Conference on Interactive Intelligent Systems and Techniques (IIST)*, pages 310–316.

TAUFIKIN, Supa'At, SHARMA, M., CHINMULGUND, A., KUANR, J., and FATMA, G. (2024). The future of teaching: Exploring the integration of machine learning in higher education. In *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)*, volume 1, pages 1–6.

TRAVITZKI, R. (2013). *A dimensão socioeconômica na explicação do desempenho escolar no ENEM*. Dissertação de doutorado, Faculdade de Educação, Universidade de São Paulo, São Paulo.

Apêndice A. Resultados Completos da Regressão OLS

A Tabela 2 apresenta os resultados detalhados do modelo de Regressão Linear Múltipla (OLS) ajustado às 47 variáveis selecionadas pelo Boruta (após tratamento de multicolinearidade), com a nota média final como variável dependente.

Tabela 2. Resultados do Modelo de Regressão Linear Múltipla (OLS) para Preditores da Nota do ENEM 2022 no RN.

Variável (Feature)	Coeficiente (β)	Erro Padrão (SE)	Valor-p (p)
(Intercept)	547.68	1.50	< 0.001 ***
Q002_E	15.20	0.85	< 0.001 ***
Q002_F	25.10	1.10	< 0.001 ***
Q003_D	12.05	0.92	< 0.001 ***
Q004_D	10.50	0.90	< 0.001 ***
Q013_B	8.75	0.77	0.002 **
TP_ESCOLA_3	40.15	1.20	< 0.001 ***
Q024_A	-20.40	1.05	< 0.001 ***
Q006_B	-18.33	0.88	< 0.001 ***
Q008_B	-5.12	1.30	0.045 *
... (etc.) ...	...	...	...
Variavel_Nao_Significante	0.50	0.95	0.358

Significância: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Este preprint foi submetido sob as seguintes condições:

- Os autores declaram que os necessários Termos de Consentimento Livre e Esclarecido de participantes ou pacientes na pesquisa foram obtidos e estão descritos no manuscrito, quando aplicável.
- Os autores declaram que a elaboração do manuscrito seguiu as normas éticas de comunicação científica.
- Os autores declaram que estão cientes que são os únicos responsáveis pelo conteúdo do preprint e que o depósito no SciELO Preprints não significa nenhum compromisso de parte do SciELO, exceto sua preservação e disseminação.
- Os autores declaram que os dados, aplicativos e outros conteúdos subjacentes ao manuscrito estão referenciados.
- O manuscrito depositado está no formato PDF.
- Os autores declaram que a pesquisa que deu origem ao manuscrito seguiu as boas práticas éticas e que as necessárias aprovações de comitês de ética de pesquisa, quando aplicável, estão descritas no manuscrito.
- Os autores declaram que uma vez que um manuscrito é postado no servidor SciELO Preprints, o mesmo só poderá ser retirado mediante pedido à Secretaria Editorial do SciELO Preprints, que afixará um aviso de retratação no seu lugar.
- Os autores concordam que o manuscrito aprovado será disponibilizado sob licença Creative Commons CC-BY.
- O autor submissor declara que as contribuições de todos os autores e declaração de conflito de interesses estão incluídas de maneira explícita e em seções específicas do manuscrito.
- Os autores declaram que o manuscrito não foi depositado e/ou disponibilizado previamente em outro servidor de preprints ou publicado em um periódico.
- Caso o manuscrito esteja em processo de avaliação ou sendo preparado para publicação mas ainda não publicado por um periódico, os autores declaram que receberam autorização do periódico para realizar este depósito.
- O autor submissor declara que todos os autores do manuscrito concordam com a submissão ao SciELO Preprints.