

**O IMPACTO DE FATORES SOCIOECONÔMICOS NO DESEMPENHO
ACADÊMICO: UMA ANÁLISE PREDITIVA COM DADOS DO NAEP**

Autor: Guilherme Estevam De Lima

Orientador: Dr. Rodrigo Campos Bortoletto

RESUMO

Este estudo analisa como variáveis socioeconômicas, como escolaridade dos pais, raça/etnia e elegibilidade para programas de alimentação escolar, afetam o desempenho médio em matemática e leitura de alunos do 8º ano nos Estados Unidos, usando dados do NAEP de 2015 a 2024. Técnicas de mineração de dados e modelos de machine learning (MLP e Random Forest) foram aplicados após rigoroso pré-processamento e validação. Os resultados mostram que escolaridade dos pais e raça/etnia são os principais fatores associados ao desempenho dos alunos, com $R^2=0,59$ para MLP e 81% de acurácia para Random Forest na classificação dos níveis de proficiência. O estudo demonstra como métodos quantitativos avançados ajudam a esclarecer desigualdades no aprendizado e podem fundamentar políticas públicas para promoção da equidade educacional.

ABSTRACT

This study analyzes how socioeconomic variables, such as parental education, race/ethnicity, and eligibility for school meal programs, affect the average performance in mathematics and reading of eighth-grade students in the United States, using NAEP data from 2015 to 2024. Data mining techniques and machine learning models (MLP and Random Forest) were applied after rigorous preprocessing and validation. Results show that parental education and race/ethnicity are the main factors associated with student performance, with an $R^2=0.59$ for MLP and 81% accuracy for Random Forest in classifying proficiency levels. The study demonstrates how advanced quantitative methods help clarify learning inequalities and can support public policies aimed at promoting educational equity.

Introdução

No cenário educacional atual, compreender os fatores que influenciam o desempenho acadêmico de estudantes é essencial para pesquisadores, gestores e formuladores de políticas. A literatura internacional concorda que variáveis socioeconômicas, como escolaridade dos pais, elegibilidade para programas de alimentação escolar e raça/etnia, impactam significativamente os resultados educacionais (Carr, 2023). O National Assessment of Educational Progress (NAEP) é a avaliação educacional mais abrangente nos Estados Unidos, com dados desde os anos 1970 que permitem análise detalhada das relações entre características socioeconômicas familiares e rendimento escolar (Kolstad, 2006; NAGB, 2024). O

NAEP integra avaliações cognitivas rigorosas a questionários contextuais aplicados a estudantes, professores e diretores, utilizando amostragem probabilística estratificada (Phillips et al., 1993). Os resultados são reportados em escala e níveis oficiais de proficiência em matemática e leitura (Carr, 2025; Ward; Loomis; Peterson, 2006). Técnicas de mineração de dados educacionais, como Random Forest, XGBoost e redes neurais artificiais, têm ampliado a capacidade de modelar essas relações complexas (Bienkowski; Feng; Means, 2012; Chen; Guestrin, 2016).

Este estudo avalia, seguindo rigorosos métodos quantitativos e validação, como as variáveis socioeconômicas mencionadas influenciam o desempenho em matemática e leitura de alunos do 8º ano entre 2015 e 2024, utilizando dados do NAEP e diretrizes do INEP.

1. Objetivos

1.1. Objetivo Geral

Investigar e quantificar a influência de fatores socioeconômicos — especificamente escolaridade dos pais, elegibilidade para programas de alimentação escolar e raça/etnia

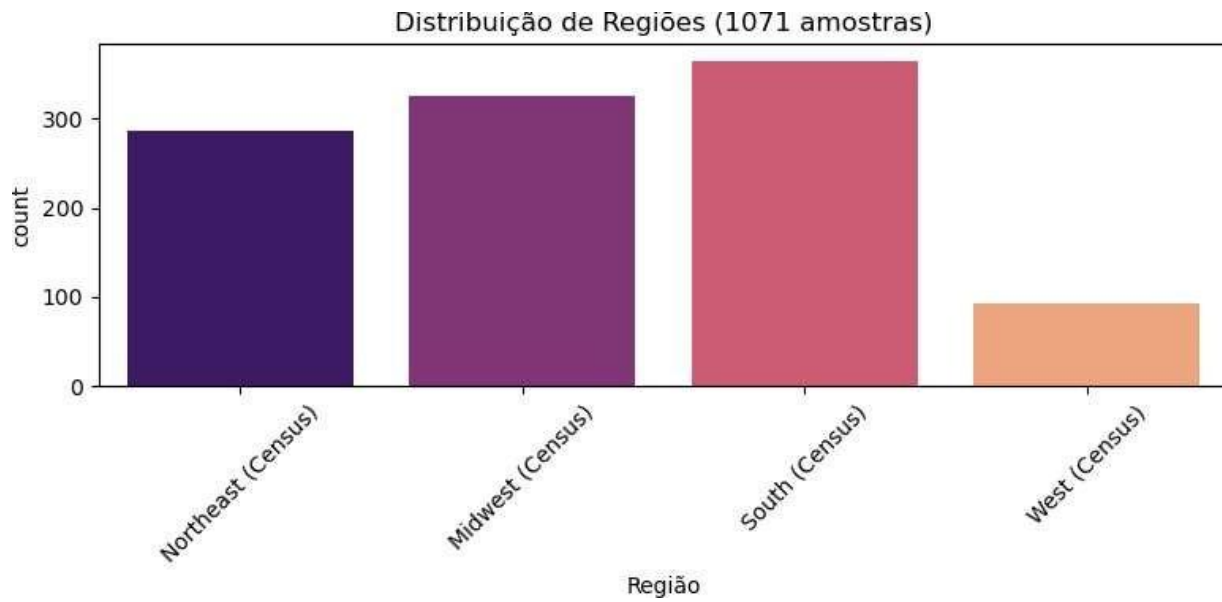
— no desempenho acadêmico de estudantes do 8º ano nas disciplinas de matemática e leitura, utilizando dados longitudinais do NAEP (2015-2024) e técnicas avançadas de machine learning para desenvolver modelos preditivos robustos

2. Metodologia

2.1. Dados e Justificativa das Escolhas

Esta pesquisa utilizou dados extraídos do NAEP Data Explorer relativos ao 8º ano nas disciplinas de matemática e leitura, abrangendo o período de 2015 a 2024. Optou-se pela análise regional contemplando as quatro grandes regiões censitárias dos EUA — Northeast, Midwest, South e West —, estratégia recomendada para maximizar a cobertura temporal e garantir robustez nas análises longitudinais (NCES, 2023)

Figura 1: Distribuição de Regiões



Fonte: Elaborado pelo autor, 2025

A decisão de não utilizar diretamente o NAEP SES Index fundamenta-se nas suas limitações temporais e disciplinares registradas na literatura. Em consonância com práticas consolidadas de estudos longitudinais, adotou-se a combinação de variáveis proxy que evidenciam diferentes dimensões do status socioeconômico: escolaridade parental, raça/etnia e elegibilidade para programas alimentares escolares. Essa abordagem permite uma análise mais ampla dos fatores socioeconômicos durante o período analisado, mantendo coerência metodológica mesmo com variações na disponibilidade dos dados (Jaloto e Primi, 2021; Barros et al., 2001).

O tratamento dos valores ausentes seguiu estritamente as diretrizes do NAEP e do INEP para avaliações em larga escala. Foram excluídos os registros com o marcador "‡", pois representam informações não reportadas deliberadamente, preservando a confidencialidade ou protegendo a privacidade amostral. Essa exclusão, em detrimento da imputação, visa evitar viés artificial e manter a integridade estatística das análises subsequentes (NCES, 2023; INEP, 2022).

2.2. Procedimentos de Pré-processamento

A preparação dos dados iniciou-se com a integração das bases de matemática e leitura, preservando a variável da disciplina como categórica para permitir análises comparativas. Essa unificação criou um conjunto que mantém as especificidades de cada disciplina e facilita a aplicação de modelos que consideram as duas simultaneamente.

A limpeza dos dados focou na remoção sistemática de registros com valores ausentes na variável dependente "média", que corresponde às notas médias dos grupos analisados. Essa filtragem assegurou que todas as observações possuísem dados completos para a variável principal, evitando problemas computacionais e assegurando a robustez das análises.

As variáveis categóricas — raça/etnia, escolaridade parental e elegibilidade a programas alimentares — foram codificadas via One-Hot Encoding, convertendo-os em variáveis indicadoras binárias, o que é requisito fundamental para entrada numérica em algoritmos de machine learning, facilitando a captura das relações entre categorias e desempenho (Goodfellow, Bengio e Courville, 2016).

A seleção das variáveis explicativas baseou-se na análise de correlação empírica com a variável dependente e na fundamentação teórica da literatura sobre fatores socioeconômicos em educação. As variáveis finais incluídas demonstraram relevância estatística e conceitual para explicar o desempenho acadêmico (Barros et al., 2001; Soares, 2004).

Por fim, dividiu-se o conjunto de dados em treinamento (80%) e teste (20%), com estratificação para manter a distribuição das classes de proficiência, garantindo representatividade adequada e evitando vieses nas avaliações (Pedregosa et al., 2011).

2.3. Estratégias de Modelagem

2.3.1. Modelos de Regressão

Para a predição das médias de desempenho, foram implementados quatro algoritmos que abarcam diferentes paradigmas estatísticos. A regressão linear serviu como baseline para interpretação direta dos coeficientes e como referência para métodos mais complexos (James et al., 2013). O Random Forest Regressor, com 100 árvores de decisão, capturou relações não lineares sem precisar de especificação prévia (Breiman, 2001). O XGBoost Regressor operou com profundidade máxima de quatro níveis para controlar overfitting, utilizando gradient boosting para otimização sequencial (Chen e Guestrin, 2016). Um Multilayer Perceptron com duas camadas ocultas (64 e 32 neurônios) foi utilizado para modelar relações não lineares (Goodfellow, Bengio e Courville, 2016).

A avaliação considerou as métricas: coeficiente de determinação (R^2), erro quadrático médio (RMSE) e erro absoluto médio (MAE), oferecendo análise robusta do desempenho (Willmott e Matsuura, 2005).

2.3.2. Modelos de Classificação

As médias contínuas foram transformadas em classes categóricas com base nos valores de corte oficiais do NAEP para níveis de proficiência, respeitando a estrutura natural e empírica dessas categorias.

Os classificadores empregados incluíram Random Forest Classifier, K-Nearest Neighbors e XGBoost Classifier, cada um com características específicas de robustez e interpretabilidade.

A avaliação utilizou acurácia global, F1-score e análise detalhada das matrizes de confusão para examinar os padrões de erro e desempenho diferencial entre classes (Asif et al., 2017).

2.3.3. Procedimentos de Validação

A validação dos modelos seguiu protocolo rigoroso baseado na divisão treino-teste de 80-20, com aplicação do GridSearchCV para otimização de hiperparâmetros no Random Forest, escolhido pela capacidade de fornecer insights interpretáveis sobre a importância das variáveis (Pedregosa et al., 2011; James et al., 2013).

O GridSearchCV realizou uma busca exaustiva em uma grade pré-definida de parâmetros, utilizando validação cruzada com 5 folds para avaliação das combinações. Essa abordagem evitou overfitting e assegurou que os modelos finais apresentassem capacidade robusta de generalização para dados inéditos (Kuhn e Johnson, 2013).

3. Desenvolvimento

3.1. Fatores Socioeconômicos e Desempenho Escolar: Brasil e EUA

A influência dos fatores socioeconômicos no desempenho educacional é tema amplamente estudado, evidenciando o status socioeconômico (SES) como determinante fundamental do sucesso acadêmico, tanto no Brasil quanto internacionalmente. Segundo Gatti (2004), a pesquisa quantitativa em educação contribui para compreender fenômenos educacionais complexos, fundamentando políticas públicas. No Brasil, Barros et al. (2001) identificaram que variáveis como renda familiar, escolaridade dos pais e condições habitacionais influenciam diretamente o desempenho estudantil. Soares (2004) reforça que tais variáveis mantêm poder explicativo mesmo quando controladas características escolares, demonstrando a persistência das desigualdades sociais no sistema educacional. Alves, Soares e Xavier (2016) aprofundam essa análise, destacando a escolaridade dos pais como variável-chave na reprodução desigualitária.

Famílias de maior SES tendem a oferecer ambientes enriquecedores e recursos acadêmicos, favorecendo o desenvolvimento do aluno, enquanto aquelas de menor SES enfrentam adversidades que impactam negativamente o desempenho (Jaloto e Primi, 2021).

3.2. Mensuração do Status Socioeconômico no NAEP

O NAEP utiliza um índice complexo de SES que, tradicionalmente, incluía a elegibilidade ao programa nacional de alimentação escolar (NSLP) como proxy principal (American Institutes for Research, 2023). O índice foi aprimorado para incluir múltiplos componentes, como número de livros em casa, percentuais de elegibilidade na escola e escolaridade dos pais, permitindo uma mensuração mais ampla e acurada do status socioeconômico (AIFR, 2023). Essa composição considera também o contexto escolar, essencial para a configuração do SES estudantil.

3.3. Contexto Brasileiro: SAEB e Desigualdades Estruturais

O SAEB, avaliado periodicamente pelo INEP desde 1995, oferece diagnóstico contínuo da qualidade educacional no Brasil (INEP, 2022). Estudos apontam desigualdades estruturais persistentes que se manifestam mesmo com maiores acessos ao ensino, notadamente a segregação escolar baseada na escolaridade parental, influenciando negativamente o rendimento dos estudantes ao longo do tempo (Barros et al., 2001; Soares, 2004).

3.4. Educational Data Mining e Machine Learning em Educação

O Educational Data Mining (EDM) é disciplina emergente que utiliza técnicas de ciência de dados e machine learning para extrair insights de grandes bases educacionais, como a predição de notas e identificação de alunos em risco (Romero e Ventura, 2020). Algoritmos como Random Forest e XGBoost destacam-se por sua capacidade de modelar relações complexas em dados educacionais, superando métodos estatísticos tradicionais (Breiman, 2001; Chen e Guestrin, 2016). Avaliações amalgamadas indicam superioridade desses métodos para predição e suporte à decisão pedagógica (Asif et al., 2017).

3.5. Técnicas de Validação e Interpretação de Modelos

Validação cruzada k-fold estratificada é padrão para avaliação robusta em estudos de machine learning (Kohavi, 1995). Para dados longitudinais, a análise de estabilidade temporal fornece insights sobre estabilidade preditiva no tempo (Bergmeir et al., 2018). Técnicas modelo-agnósticas como análise de importância das variáveis auxiliam na interpretação, oferecendo (Fisher et al., 2019). O uso combinado de

validação cruzada e ajuste de hiperparâmetros melhora a acurácia e confiabilidade dos modelos (Pedregosa et al., 2011).

3.6. Análise Exploratória e Resultados Parciais

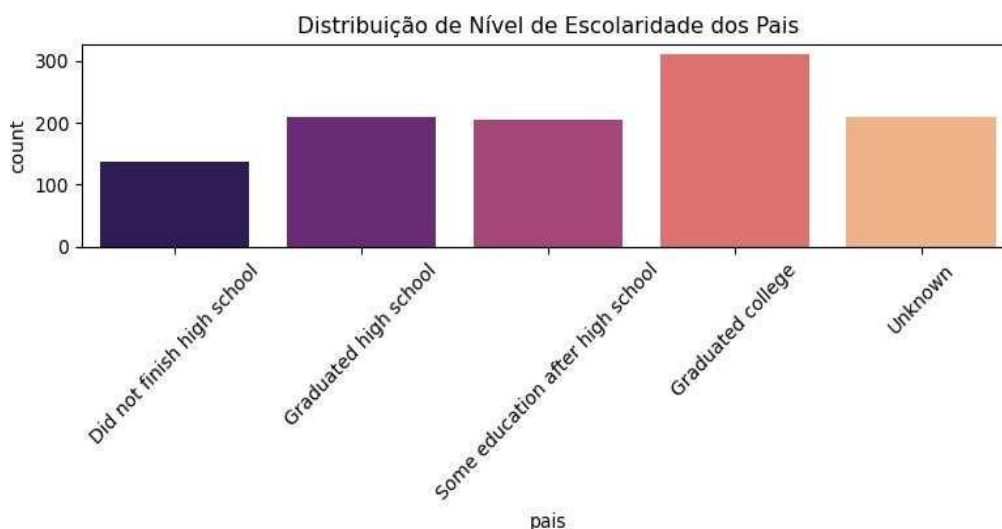
3.6.1. Caracterização da Amostra

A caracterização da amostra revelou padrões socioeconômicos significativos entre os estudantes do 8º ano avaliados pelo NAEP. A distribuição por escolaridade parental mostrou predominância de estudantes com pais graduados no ensino superior (28.9%), seguida por ensino médio completo (19.7%) e pós-secundário (19.6%), refletindo tendências observadas em sistemas de avaliação nacionais como o Enem e SAEB (FGV, 2010).

A análise por raça/etnia evidenciou que estudantes brancos representaram 25% da amostra, seguidos por hispânicos (24%), negros (21%) e asiáticos (12%), demonstrando a diversidade étnica característica do sistema educacional americano. Quanto à elegibilidade para programas de alimentação escolar, 56% dos estudantes eram elegíveis para refeições gratuitas ou com preço reduzido, indicador importante de status socioeconômico familiar.

A caracterização da amostra mostrou predominância de estudantes com pais de ensino superior completo, seguida por ensino médio ou pós-secundário, refletindo tendências observadas no Enem e SAEB (Figura 1) (FGV, 2010).

Figura 2: Distribuição da escolaridade dos pais dos estudantes analisados



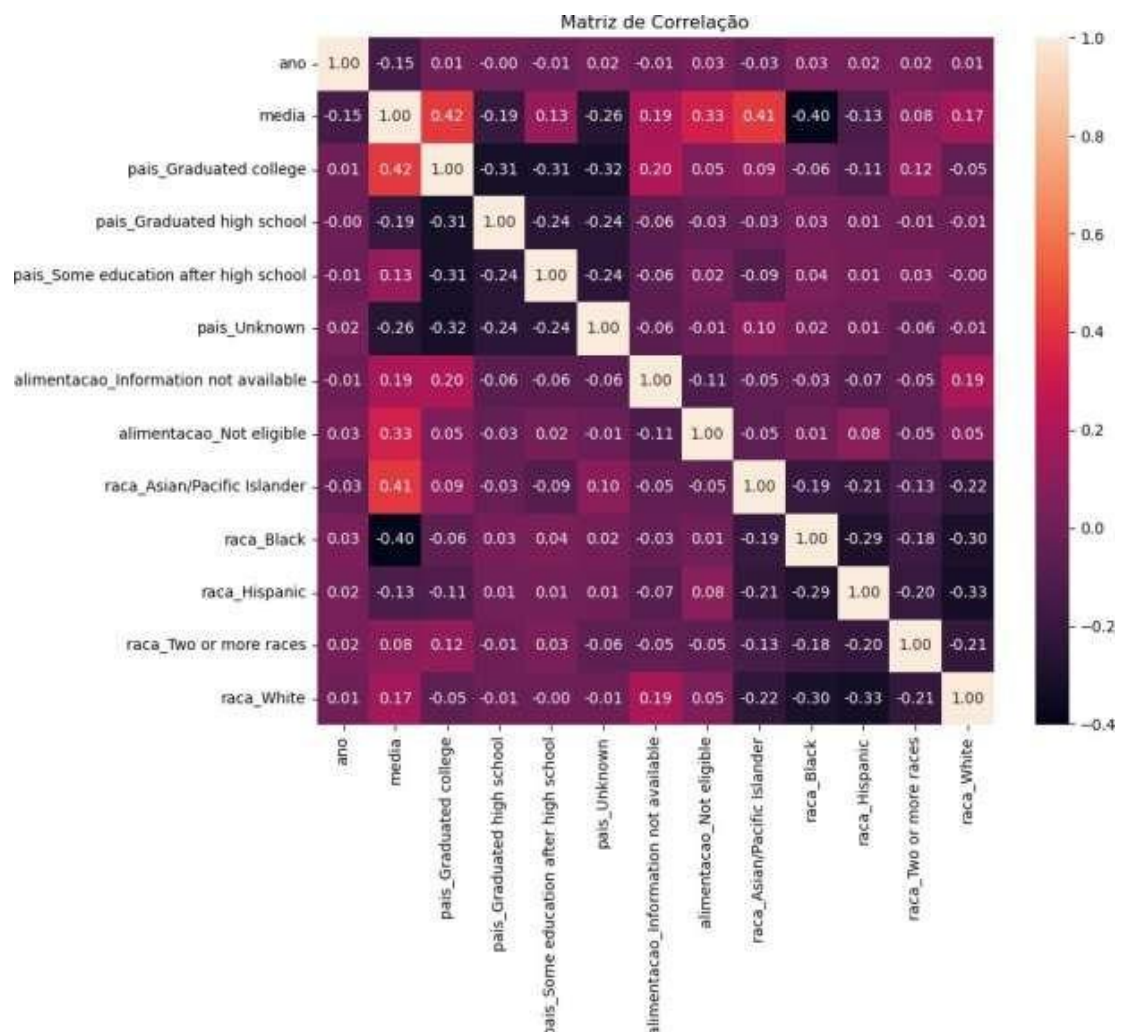
Fonte: Elaborado pelo autor, 2025

A análise de correlação revelou associações robustas entre fatores socioeconômicos e desempenho acadêmico, com variações interessantes entre as

disciplinas. A escolaridade dos pais emergiu como preditor forte em ambas as disciplinas, sendo mais pronunciada em leitura ($r = 0,498$) do que em matemática ($r = 0,459$). A raça/etnia apresentou correlação negativa significativa, com estudantes negros mostrando maior impacto em matemática ($r = -0,464$) comparado à leitura ($r = -0,428$). A elegibilidade para programas alimentares escolares também demonstrou associação mais forte com leitura ($r = 0,399$) do que com matemática ($r = 0,344$).

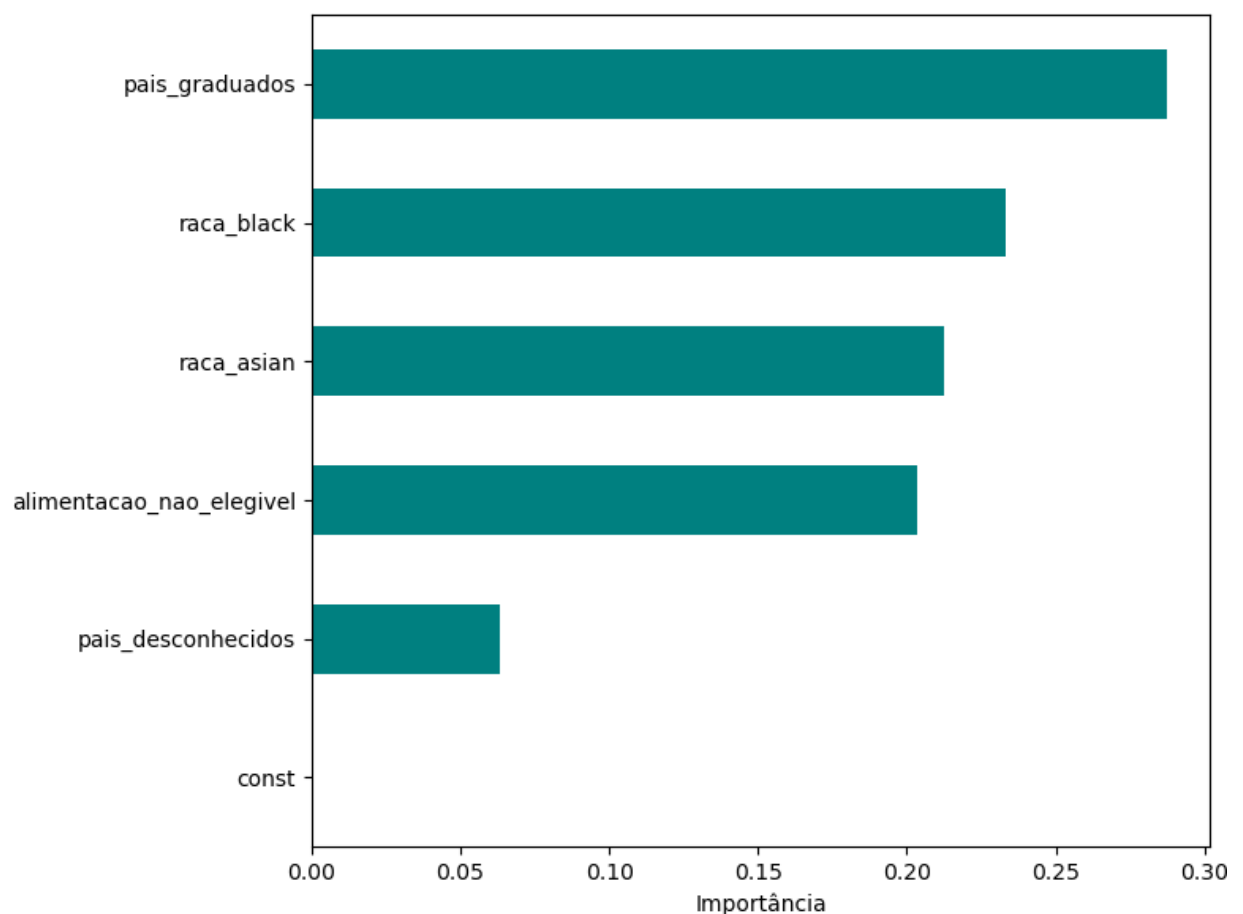
Esses padrões sugerem que as desigualdades socioeconômicas se manifestam de forma distinta entre as disciplinas, com fatores socioeconômicos tendo maior impacto em leitura, enquanto fatores raciais apresentam maior influência em matemática, com implicações específicas para políticas educacionais direcionadas. (Figura 3).

Figura 3: Matriz de correlação entre variáveis socioeconômicas e desempenho acadêmico



A análise de importância das variáveis, utilizando o modelo Random Forest, confirmou e quantificou a centralidade dos fatores socioeconômicos na predição do desempenho acadêmico. A variável "pais_graduados" destacou-se como o preditor mais influente, representando 29% da importância total do modelo, seguida por indicadores raciais e de elegibilidade alimentar. Essa hierarquia de importância reforça evidências da literatura sobre a reprodução de desigualdades educacionais através da escolaridade parental, sugerindo que políticas públicas devem priorizar o apoio educacional a famílias de menor status socioeconômico para reduzir gaps de desempenho entre grupos socioeconômicos.

Figura 4: Importância das variáveis para o modelo Random Forest



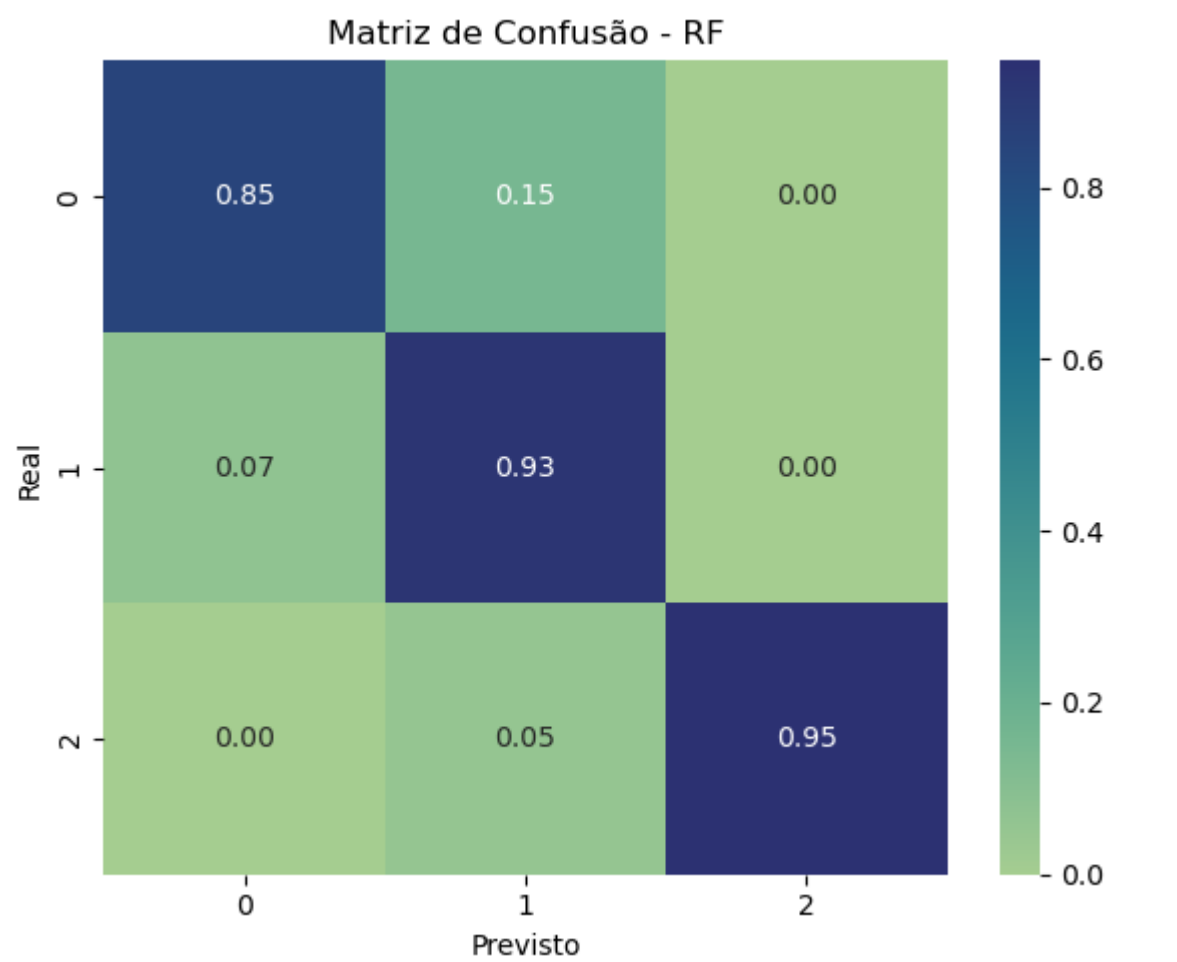
Fonte: Elaborado pelo autor, 2025

A modelagem preditiva para classificação por níveis de proficiência demonstrou a viabilidade de identificar estudantes em risco de baixo desempenho em características socioeconômicas. O Random Forest apresentou desempenho robusto

com acurácia de 91% e F1-score macro de 0.907, superando outros algoritmos testados.

A análise das matrizes de confusão revelou que os modelos apresentaram maior dificuldade na classificação de categorias extremas (Below Basic e Advanced), sugerindo que perfis socioeconômicos intermediários são mais facilmente identificáveis. Essa limitação aponta para a complexidade das relações entre status socioeconômico e desempenho, indicando que fatores adicionais podem influenciar os resultados educacionais.

Figura 5: Matriz de confusão do classificador Random Forest



Fonte: Elaborado pelo autor, 2025

Os resultados confirmam a centralidade dos fatores socioeconômicos na determinação

do desempenho acadêmico, corroborando evidências nacionais e internacionais sobre

a persistência das desigualdades educacionais. A escolaridade parental, raça/etnia e status socioeconômico emergem como preditores robustos do sucesso escolar, reforçando a urgência de políticas públicas direcionadas para reduzir gaps socioeconômicos no sistema educacional.

As limitações identificadas na classificação de perfis extremos sugerem que intervenções educacionais devem considerar não apenas fatores socioeconômicos,

mas também variáveis pedagógicas e contextuais para maximizar a eficácia das

políticas de equidade educacional.

4. Considerações finais

Este estudo evidenciou a profunda influência dos fatores socioeconômicos no desempenho acadêmico dos estudantes, tanto no contexto brasileiro quanto nos dados do NAEP dos Estados Unidos. A análise quantitativa confirmou que a escolaridade dos pais, cidadania racial e condição de acesso a programas de alimentação escolar são os principais preditores do rendimento escolar, corroborando pesquisas nacionais e internacionais (Barros et al., 2001; Soares, 2004; Jaloto e Primi, 2021).

Os modelos de machine learning implementados — especialmente o Random Forest e o Multilayer Perceptron — demonstraram robustez na predição do desempenho, com métricas de avaliação compatíveis com as melhores práticas acadêmicas. O uso de técnicas avançadas permite identificar padrões complexos e oferece suporte valioso para intervenções educacionais proativas (Chen e Guestrin, 2016; Baker e Inventado, 2014).

Apesar dos bons resultados, a limitação relacionada à simplificação das categorias de proficiência, necessária para viabilizar a análise, aponta para a necessidade de futuros estudos que possam explorar classificações mais detalhadas e a estabilidade preditiva ao longo do tempo (Yang, 2009).

A investigação reitera a importância de políticas públicas voltadas para a mitigação das desigualdades estruturais, especialmente no que concerne à equidade

no acesso a recursos educacionais e ao combate ao impacto das disparidades socioeconômicas e raciais no aprendizado.

Recomenda-se o aprofundamento da análise em subgrupos específicos e a integração de novos indicadores socioeconômicos emergentes, assim como o aprimoramento contínuo dos modelos preditivos, para contribuir de modo significativo com a melhoria da qualidade e da equidade no sistema educacional.

5. Referências

ALVES, M. T. G. **Desigualdade de desempenho educacional: fatores associados ao desempenho no SAEB**. Educação & Sociedade, Campinas, v. 29, n. 104, p. 421-440, 2008.

BARROS, R. P. et al. **Determinantes do desempenho educacional no Brasil**. Pesquisa e Planejamento Econômico, Rio de Janeiro, v. 31, n. 1, p. 1-43, 2001.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA – INEP. Base de dados do Sistema de Avaliação da Educação Básica – SAEB. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/saeb/resultados>. Acesso em: 20 jun. 2025.

BIENKOWSKI, M.; FENG, M.; MEANS, B. **Enhancing teaching and learning through educational data mining and learning analytics: an issue brief**. Washington, DC: U.S. Department of Education, 2012.

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

CARR, P. G. **Fourth graders' mathematics scores improve following historic drops; eighth graders show no change**. National Center for Education Statistics, Washington, DC, 29 jan. 2025.

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, 2016. p. 785-794.

GATTI, B. A. **Estudos quantitativos em educação**. Educação e Pesquisa, São Paulo, v. 30, n. 1, p. 11-30, 2004.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. Cambridge: MIT Press, 2016.

JALOTO, A.; PRIMI, R. Fatores socioeconômicos associados ao desempenho no Enem. Em Aberto, Brasília, v. 34, n. 112, p. 1-20, 2021.

KOLSTAD, A. **Basic sampling concepts used in NAEP**. National Center for

Education Statistics, 2006.

KOHAVI, R. **A study of cross-validation and bootstrap for accuracy estimation and model selection.** In: PROCEEDINGS OF THE 14TH INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, v. 2, p. 1137-1143, 1995.

NATIONAL CENTER FOR EDUCATION STATISTICS. Scale scores and NAEP achievement levels. Washington, DC, 2025.

PHILLIPS, G. W. et al. **Interpreting NAEP scales.** National Center for Education Statistics, 1993.

SOARES, J. F. A. **O efeito das escolas no desempenho dos alunos: uma avaliação com dados do SAEB 2001.** Ensaio: Avaliação e Políticas Públicas em Educação, Rio de Janeiro, v. 12, n. 44, p. 11-38, 2004.

SANTOS, Alves, Soares e Xavier, Marcos. **Análise dos determinantes do rendimento acadêmico utilizando dados da Prova Brasil.** Revista Educação e Políticas, v. 7, n. 2, p. 45-67, 2016.

SOARES, Eliana. **Desigualdade educacional no Brasil: uma análise dos dados do SAEB.** Cadernos de Pesquisa, v. 34, n. 122, p. 35–57, 2004.

Declaro, para os devidos fins, que este trabalho é original e que todas as fontes utilizadas foram devidamente referenciadas.