Uso de *Machine Learning* e Variáveis Socioeconômicas na Previsão de Desempenho do ENEM: potencialidades, limites e implicações éticas

David Melo da Luz, ICMC/USP, davidluz@usp.br, 0009-0002-1158-9239
Natália de Mesquita Matheus, Instituto Par, na.matheus@gmail.com, 0000-0003-2274-4222

Resumo: Este artigo apresenta o resultado de uma pesquisa que treinou modelos de *machine learning* para predizer se estudantes do ENEM ficariam abaixo da média nacional, utilizando exclusivamente variáveis socioeconômicas. No treinamento do modelo foram utilizados microdados do ENEM 2023, submetidos a procedimentos de limpeza e organização, bem como à construção de variáveis mais explicativas a partir de variáveis originais. Modelos de referência foram comparados a algoritmos de *gradient boosting*, com destaque para *CatBoost* e *XGBoost*, incluindo versões enriquecidas com *embeddings* semânticos. Os resultados indicam boa capacidade de identificação de estudantes em situação de maior risco de baixo desempenho, com desempenho preditivo consistente para esse tipo de tarefa, com *F1-score* em torno de 0,76. Em vez de apoiar decisões individuais, os resultados são interpretados de forma agregada, como indicadores de desigualdades estruturais associadas a fatores como renda, raça/cor, tipo de escola, gênero e território. O artigo discute as potencialidades e os limites do uso desses modelos no planejamento educacional e apresenta considerações éticas relacionadas à governança, equidade e transparência no uso de tecnologias preditivas em políticas públicas.

Palavras-chave: enem; *machine learning*; predição de desempenho; ética; viés algorítmico

Use of Machine Learning and Socioeconomic Variables in Predicting ENEM Performance: potentialities, limits, and ethical implications

Abstract: This article presents the results of a study that trained machine learning models to predict the risk of students in the Brazilian National High School Exam (ENEM) achieving performance below the national average, using exclusively socioeconomic variables. Microdata from the 2023 ENEM were analyzed and subjected to data cleaning and organization procedures, as well as feature construction based on the socioeconomic questionnaire. Baseline models were compared with gradient boosting algorithms, with emphasis on CatBoost and XGBoost, including versions enriched with semantic embeddings. The results indicate a good ability to identify students at higher risk of low performance, achieving consistent predictive performance for this type of task, with an F1-score around 0.76. Rather than supporting individual decision-making, the results are interpreted in an aggregated manner, as indicators of structural inequalities associated with factors such as income, race/ethnicity, type of school, gender, and territory. The article discusses the potentialities and limitations of using these models for educational planning and presents ethical considerations related to governance, equity, and transparency in the use of predictive technologies in public policies.

Keywords: enem; *machine learning*; *performance prediction*; *ethics*; *algorithmic bias*

1. Introdução

O Exame Nacional do Ensino Médio (ENEM) consolidou-se como uma das principais avaliações educacionais de larga escala no Brasil. O exame estima a proficiência de estudantes em quatro áreas do conhecimento, além da redação. Para além de sua função avaliativa, os resultados do exame também são amplamente utilizados como critério de acesso ao ensino superior em instituições públicas e privadas em todo o país, influenciando trajetórias educacionais e subsidiando decisões em políticas de acesso e permanência. Somente em 2024, o ENEM registrou 4,3 milhões de inscritos, dos quais cerca de 1,6 milhão eram egressos do ensino médio (INEP, 2024).

Além das questões do próprio exame, os participantes do ENEM respondem, previamente à aplicação da prova, a um questionário socioeconômico. As respostas dos estudantes, devidamente anonimizadas, bem como os dados provenientes do formulário socioeconômico, são disponibilizados anualmente no portal do Ministério da Educação do Brasil em formato de microdados, o que permite a pesquisadores realizar estudos e análises sobre o desempenho educacional e suas associações com diferentes características socioeconômicas. Esses microdados, em função do volume de dados, abrangência, capilaridade e disponibilidade anual, tem sido amplamente utilizado por pesquisadores da área social para investigar padrões de desigualdade educacional e a relação entre condições socioeconômicas e desempenho, com evidências consistentes de efeitos associados à renda, à escolaridade dos pais e ao acesso a recursos tecnológicos, além de subsidiar o planejamento de políticas públicas educacionais (NOVAES; PITON-GONÇALVES, 2021; PINTO; TENÓRIO, 2021; RUBIM; MANTOVANI; ALAVARSE, 2025).

A cada ano, multiplicam-se propostas de ferramentas e instrumentos de apoio à tomada de decisão baseados em técnicas computacionais e análise de dados, inclusive nas ciências sociais aplicadas e na área de educação (SEVILLA et al., 2022; RAHAL; VERHAGEN; KIRK, 2022). Esse movimento está diretamente relacionado à crescente disponibilidade de bases de dados em larga escala, como os microdados do ENEM, e à necessidade de métodos capazes de lidar com a complexidade e o volume dessas informações. Nesse contexto, observa-se o uso crescente de técnicas de *machine learning*, ou Aprendizado de Máquina, entendidas como um conjunto de métodos que permitem a um computador aprender padrões a partir dos dados para realizar tarefas específicas e avaliar seu próprio desempenho (MITCHELL, 1997).

O avanço do poder computacional tem ampliado a viabilidade dessas abordagens, que vêm sendo empregadas para expandir a capacidade explicativa e preditiva das análises educacionais, ao mesmo tempo em que abrem novas frentes de investigação sobre desigualdades educacionais e a eficácia de intervenções.

Contudo, a incorporação desses métodos no campo educacional demanda cautela. Modelos de *machine learning* geram previsões que podem apoiar a organização de prioridades e a exploração de hipóteses sobre combinações de fatores, mas não substituem abordagens voltadas à identificação de relações causais, tipicamente apoiadas por métodos inferenciais clássicos. Nesse sentido, a escolha entre técnicas de estatística clássica, modelos baseados em *machine learning* ou estratégias híbridas deve ser orientada pela natureza do problema, pela estrutura e qualidade dos dados disponíveis e pela finalidade da modelagem, explicação, monitoramento, priorização ou apoio a decisões operacionais (ALI et al., 2024). Quando o objetivo é estimar parâmetros, testar hipóteses e comunicar incerteza, métodos inferenciais tradicionais tendem a ser mais adequados; quando a meta é maximizar desempenho preditivo, abordagens de *machine learning* podem ser preferíveis; e, em muitos cenários, combinações dos dois paradigmas oferecem um compromisso útil.

Além das questões metodológicas, a literatura recente tem destacado que modelos de *machine learning* podem reproduzir e amplificar vieses históricos presentes nos dados, afetando de forma desigual grupos definidos por raça, gênero e renda (ALMALAWI et al., 2024). Esse risco torna indispensável que a adoção dessas tecnologias seja orientada por princípios de ética, transparência e equidade, especialmente quando aplicadas a contextos educacionais e a políticas públicas.

Apesar do potencial desses métodos e da ampla disponibilidade de microdados como os do ENEM, ainda é relevante delimitar, em cenários de larga escala, até que ponto variáveis estritamente socioeconômicas sustentam modelos preditivos capazes de identificar estudantes em maior risco de baixo desempenho, bem como explicitar limites de uso e cuidados éticos necessários para evitar interpretações indevidas e efeitos distributivos indesejados. Assim, o

problema investigado pode ser sintetizado na seguinte pergunta de pesquisa: em que medida variáveis socioeconômicas do ENEM permitem predizer o risco de um estudante obter desempenho abaixo da média e quais são as potencialidades, limitações e implicações éticas associadas a esse uso?

É neste contexto que se insere o presente trabalho, cujo objetivo foi empregar modelos de *machine learning* baseados exclusivamente em variáveis socioeconômicas para prever o risco de um estudante obter desempenho abaixo da média no ENEM, conforme definição operacional detalhada na seção de Materiais e Métodos, e discutir, a partir dessa predição, as potencialidades, os limites e as implicações éticas do uso desse tipo de modelo no contexto educacional (LUZ, 2025).

2. Referencial teórico

O uso de modelos preditivos para estimar o desempenho acadêmico resulta da convergência entre (1) evidências consolidadas de associação entre condições socioeconômicas e desempenho, (2) a disponibilidade de microdados públicos em larga escala, como os do ENEM, e (3) o amadurecimento recente de métodos e infraestrutura para análise de grandes volumes de dados, impulsionado por avanços em computação em nuvem e hardware e pela consolidação de abordagens de *deep learning* (BERISHA; MËZIU; SHABANI, 2022; LECUN; BENGIO; HINTON, 2015). No contexto brasileiro, a literatura documenta de forma consistente que desigualdades sociais e territoriais se refletem em resultados acadêmicos distintos (NOVAES; PITON-GONÇALVES, 2021; PINTO; TENÓRIO, 2021). Revisões também indicam que algoritmos de *machine learning* podem oferecer ganhos preditivos em cenários com relações múltiplas e não lineares (FERNÁNDEZ-DELGADO et al., 2014; ALMALAWI et al., 2024).

2.1 Formulação do problema e métricas de avaliação

Em estudos preditivos de desempenho de avaliações como o ENEM, a variável-alvo define a formulação do problema e, por consequência, os indicadores de desempenho. Quando a meta é estimar uma nota ou pontuação numérica (problema de regressão), trabalha-se com variáveis contínuas. Quando a meta é classificar perfis de desempenho (por exemplo, alto/regular/baixo definidos por faixas percentuais), o problema torna-se categórico (classificação) e é usualmente avaliado por métricas como *accuracy* (acurácia), *precision* (precisão) , *recall (*revocação ou sensibilidade), *ROC-AUC* (área sob a curva ROC), *balanced accuracy* (acurácia balanceada) e *F1-score*, especialmente quando há interesse em comparar desempenho entre classes. É comum, em ciência de dados e *machine learning*, referir-se a essas métricas pelos nomes em inglês, por padronização na literatura e nas ferramentas da área; a tradução para o português é possível, mas muitos autores mantêm o inglês para facilitar comparação com estudos e implementações

A *accuracy* (acurácia) indica, entre todas as previsões do modelo, qual proporção ele acertou “no geral”. A *precision* (precisão) indica, entre os casos que o modelo classificou como pertencentes à classe de interesse, quantos realmente pertencem a essa classe. A *recall* (revocação ou sensibilidade) indica, entre todos os casos que de fato pertencem à classe de interesse, quantos o modelo conseguiu identificar. A *ROC-AUC* (área sob a curva ROC) é usada principalmente quando o modelo fornece probabilidades: em vez de avaliar um único ponto de corte, ela resume o desempenho do classificador em vários limiares e indica a capacidade do modelo de separar (ranquear) corretamente os casos da classe de interesse em relação aos demais. Já a *balanced accuracy* (acurácia balanceada) é uma variação da acurácia, útil quando as classes são desbalanceadas, pois calcula o desempenho de forma mais equilibrada entre as classes, em vez de deixar a classe mais frequente “dominar” o resultado.

O *F1-score* é uma métrica que resume, em um único número, o quão bem o modelo classifica corretamente os casos da classe de interesse, equilibrando dois aspectos: *precision* e *recall*. Assim, o F1-score ajuda a evitar situações em que o modelo acerta "no geral", mas erra justamente o que importa: ou marca muitos casos sem necessidade (baixa *precision*) ou deixa passar muitos casos que deveriam ser identificados (baixa *recall*). Quanto mais próximo de 1, melhor o desempenho.

SOUZA e KLUG (2025), por exemplo, aplicam o algoritmo de classificação *random forest* a microdados do ENEM, para classificar desempenho em faixas (alto/regular/baixo) e reportam métricas elevadas (*accuracy*, *precision*, *recall* e *F1-score*), destacando renda, escolaridade e ocupação dos responsáveis entre os principais preditores, em perspectiva agregada para análise de desigualdades.

2.2 Técnicas de preparação de dados e modelos de *machine learning* para dados tabulares

A construção de modelos preditivos com microdados educacionais envolve "escolhas" em diferentes etapas. A primeira etapa é preparar os dados para que o algoritmo consiga "ler" as informações de forma consistente. Isso inclui, por exemplo, transformar respostas categóricas (como tipo de escola, município ou língua escolhida) em um formato numérico adequado (codificação), e também lidar com situações em que há muito mais casos de uma classe do que de outra (por exemplo, muito mais estudantes acima da média do que abaixo, ou vice-versa). Quando as classes estão desbalanceadas, o modelo pode "aprender" a favorecer a classe mais comum; por isso, podem ser usadas estratégias como o *SMOTE*, que cria exemplos sintéticos da classe menos frequente para equilibrar o treinamento (CHAWLA et al., 2002)..

A segunda etapa é escolher o modelo de *machine learning*. Modelos baseados em algoritmos de *gradient boosting*, como *XGBoost* e *CatBoost*, tendem a ter excelente desempenho em dados tabulares e escalam bem com muitas variáveis, inclusive categóricas, com desempenho consistente reportado em estudos anteriores (JOSHI et al., 2021; JOSHI et al., 2022; ALMALAWI et al., 2024). Em termos simples, esses modelos funcionam combinando várias árvores de decisão: cada árvore faz uma "tentativa" de separar as classes, e as árvores seguintes vão corrigindo os erros das anteriores, aumentando a qualidade final da classificação. O *CatBoost* tem uma vantagem prática importante neste contexto: ele foi projetado para tratar variáveis categóricas de forma mais direta e robusta, o que é útil quando há muitas categorias possíveis, reduzindo a necessidade de transformações complexas e, em muitos casos, melhorando a estabilidade do resultado. Evidências recentes reforçam a eficiência do *CatBoost*: no trabalho de FAN; GOU; WENG (2025), uma variação do *CatBoost* (*C-CatBoost*) que complementa o modelo com a correção do erro residual apresentou ganhos consistentes na predição de desempenho estudantil.

A literatura sobre *machine learning* em educação enfatiza que modelos podem reproduzir vieses históricos e produzir efeitos desiguais entre grupos, sobretudo quando variáveis sensíveis ou proxies estão presentes nos dados. Por isso, recomenda-se que aplicações em políticas públicas sejam avaliadas sob critérios de equidade, transparência e responsabilização, com atenção a métricas por subgrupos e à clareza sobre finalidades e limites de uso (ALMALAWI et al., 2024; SIMÕES-GOMES; ROBERTO; MENDONÇA, 2020).

3. Material e Métodos

Foram utilizados os microdados do ENEM 2023 (INEP, 2024), disponibilizados publicamente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), em formato CSV, com 3.933.955 registros e 76 variáveis.

A sequência de ações de tratamento e uso base está esquematizada na Figura 01, a seguir.

Etapa	Entrada	Procedimento
1. Dados brutos	Microdados ENEM 2023 (CSV)	Base original
2. Filtragem de participantes	Dados brutos	Remoção de estudantes que não realizaram alguma das provas ou que são treineiros
3. Seleção de variáveis + engenharia de features	Base filtrada	Manutenção apenas das colunas com variáveis do formulário socioeconômico + feature engineering + criação da feature "média"
4. Base final pós-tratamento	Base preparada	Exportação da base final (CSV)

Figura 01 - Sequência de ações de tratamento e uso da base.

Para garantir consistência, manteve-se apenas participantes com notas nas quatro provas objetivas e redação registrada, o que removeu 1.348.840 registros; adicionalmente, excluíram-se os candidatos "treineiros" (489.141). Após esses critérios, a base final foi composta por 2.095.974 participantes.

A nota do estudante (nota_aluno) foi definida como a média aritmética simples das quatro provas objetivas (Linguagens, Ciências da Natureza, Ciências Humanas e Matemática), sem incluir a redação. A média de referência (média_ENEM) foi calculada como a média de nota_aluno na base já filtrada. A variável-alvo ABAIXO_DA_MEDIA foi então definida de forma binária: 1 se nota_aluno $\leq$ média_ENEM e 0 caso contrário; após sua criação, as colunas de notas individuais por prova (e identificadores sem potencial analítico) foram removidas para evitar vazamento de informação.

Na engenharia de atributos, removeram-se colunas com mais de 30% de valores ausentes e criaram-se indicadores derivados do questionário socioeconômico: computador_por_pessoa (Q024/Q005) e renda_per_capita (Q006/Q005). Além disso, CO_MUNICIPIO_PROVA (código IBGE) foi decomposto para derivar UF e Região, preservando a dimensão territorial. Valores ausentes foram tratados por imputação simples (mediana para variáveis numéricas e moda para categóricas). Como os modelos empregados são baseados em árvores, não foi necessária padronização/normalização das variáveis.

A estratégia de modelagem foi organizada em três etapas. (i) Linha de base: foram treinados modelos de comparação para balizar a tarefa e reduzir o risco de superinterpretação de ganhos: *dummy classifier*, regressão logística, *random forest* e *LightGBM* (KE et al., 2017). (ii) Modelos principais sem enriquecimento semântico: foram ajustados *XGBoost* e *CatBoost* (CHEN; GUESTRIN, 2016; PROKHORENKOVA et al., 2018), explorando-se sua adequação a dados tabulares; o *CatBoost* foi avaliado também em uma variante com pequenas variações de pré-processamento (C-*CatBoost*). (iii) Modelos principais com enriquecimento semântico: repetiram-se os treinamentos de *XGBoost* e *CatBoost* com a adição de embeddings (gerados

via *API* da OpenAI) para representar, em um espaço contínuo, informações categóricas/textuais selecionadas do questionário; esses vetores foram concatenados ao conjunto de atributos tabulares.

Em todos os casos, o alvo foi binarizado como "abaixo da média geral do ENEM" (classe positiva). O *CatBoost* tratou variáveis categóricas de forma nativa (codificação ordenada com controle de vazamento), enquanto os demais modelos utilizaram codificação apropriada (por exemplo, *one-hot* para cardinalidades moderadas e técnicas parcimoniosas quando necessário). Quando a cardinalidade foi elevada, empregaram-se codificações de menor dimensionalidade para evitar aumento excessivo no número de atributos. Os hiperparâmetros do *XGBoost* e do *CatBoost* foram otimizados com *RandomizedSearchCV* e *Optuna*, respectivamente (AKIBA et al., 2019)..

Os experimentos foram executados em ambiente local com Python 3.10 e Jupyter Notebook em máquina com Intel Core i9 (13ª geração), 32 GB de RAM e GPU GeForce RTX 3060 (12 GB); quando necessário, utilizou-se aceleração em *GPU* para otimizar o ajuste e as buscas de hiperparâmetros. O código-fonte (*scripts*, *notebooks* e artefatos, bases) foi organizado para reprodutibilidade, registro das versões de dados/modelos e *pipelines*, e encontra-se versionado no repositório GitHub do projeto disponível publicamente.

4. Resultados e Discussão

A tabela 1 apresenta a comparação de desempenho entre os 11 modelos avaliados. Inclui-se também um *dummy classifier*, utilizado como linha de base. Trata-se de um classificador deliberadamente simples, que não realiza aprendizado a partir dos dados e produz previsões com base em uma regra elementar (predição da classe majoritária ou a atribuição de rótulos segundo a distribuição observada das classes). Dessa forma, seu desempenho funciona como um limiar mínimo esperado para o problema. Observa-se que todos os modelos testados superaram esse *baseline*, sugerindo que capturam padrões informativos presentes nos dados e apresentam capacidade preditiva superior à de uma estratégia ingênua.

Posição	Modelo	Acurácia balanceada	ROC-AUC	Acurácia global	F1-score (false)	F1-score (true)
1	C-CatBoost	0,736	0,816	0,739	0,716	0,758
2	CatBoost	0,736	0,815	0,738	0,716	0,757
3	CatBoost + *embeddings*	0,736	0,815	0,738	0,716	0,757
4	C-CatBoost + *embeddings*	0,736	0,815	0,738	0,715	0,757
5	LightGBM	0,733	0,811	0,740	0,710	0,760
6	LightGBM + *embeddings*	0,733	0,811	0,740	0,710	0,760
7	XGBoost + *embeddings*	0,733	0,811	0,734	0,719	0,748
8	Regressão Logística	0,728	0,805	0,730	0,710	0,750
9	XGBoost	0,727	0,805	0,730	0,710	0,740
10	Random Forest	0,724	0,801	0,730	0,700	0,750
11	*Embeddings* "puros"	0,708	0,781	0,711	0,685	0,733
12	Dummy Classifier	0,500	0,500	0,500	0,470	0,530

Tabela 01 - Desempenho dos modelos

O modelo com melhor desempenho global foi o *C-CatBoost*, que apresentou o maior *ROC-AUC* (0,816) e desempenho elevado na classe positiva (*F1-score* (*true*) = 0,758 ≈ 0,76), indicando bom equilíbrio entre precisão e revocação na identificação de estudantes abaixo da média, sem explosão de falsos positivos. Esse patamar é coerente com o teto empírico

relatado na literatura de aproximadamente 75% quando se utilizam apenas variáveis socioeconômicas (LUZ, 2025).

Uma vez definido o modelo de melhor desempenho, cabe identificar as variáveis mais importantes. Na Figura 02, a seguir, estão apresentadas as dez variáveis mais importantes para o modelo *C-CatBoost*. Aqui, "importância" significa o quanto cada variável contribuiu para reduzir erros, não causalidade. Em outras palavras, indica quais sinais o modelo mais utilizou para acertar suas previsões (LUZ, 2025).

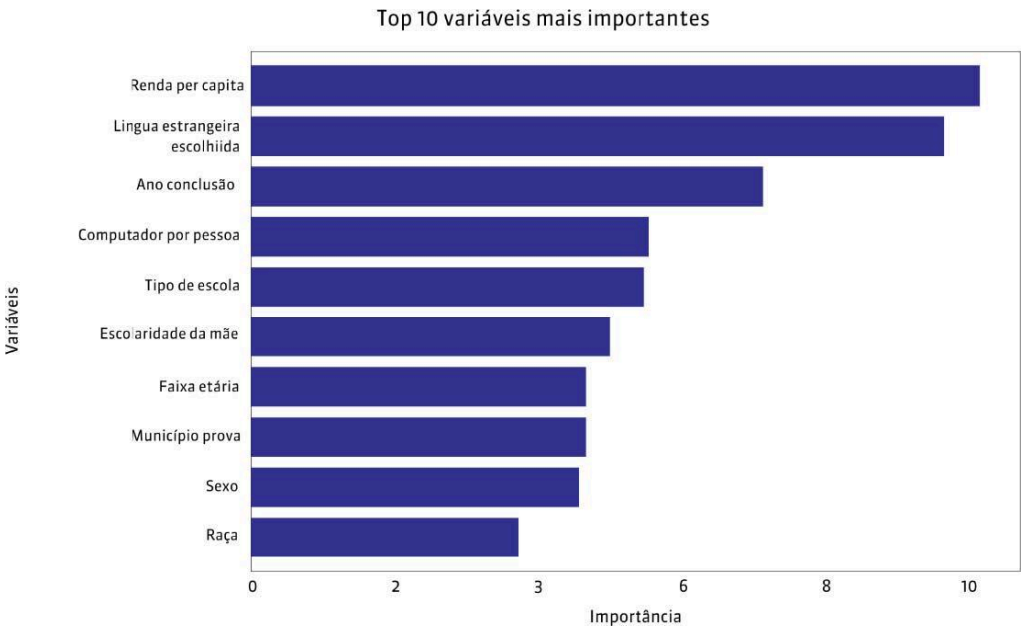

Figura 02 - Variáveis mais importantes para o modelo

De modo consistente com a literatura, renda familiar per capita aparece no topo. A renda sintetiza acesso a recursos materiais: dispositivos e conectividade, locais silenciosos para estudar, cursos extracurriculares, transporte, tempo livre para estudo. No contexto do modelo, a renda funciona como *proxy* de oportunidades educacionais acumuladas.

Em seguida, surge a língua estrangeira escolhida no ENEM (Inglês ou Espanhol). A hipótese é que essa decisão possivelmente codifica diferenças de exposição prévia e capital cultural: optar por inglês (menos próximo ao português do que o espanhol) tende a sinalizar maior familiaridade com a língua, seja por escola com maior carga horária, curso de idiomas ou consumo de mídia em inglês.

O ano de conclusão do ensino médio mostra um efeito relevante: quanto mais recente a conclusão, melhor tende a ser o desempenho. A defasagem e o tempo fora da escola se correlacionam negativamente com a performance.

O indicador de computadores por pessoa no domicílio captura o acesso efetivo à tecnologia e funciona como *proxy* de renda: mais aparelhos costumam significar maior possibilidade de fazer simulados, assistir a videoaulas e realizar pesquisas. O tipo de escola (rede ou administração) agrega diferenças de infraestrutura, tempo de instrução, currículo e orientação para exames, amplamente documentadas.

Já a escolaridade da mãe sintetiza capital cultural e expectativas; níveis mais altos costumam vir acompanhados de maior acompanhamento escolar e de circulação de repertórios de leitura.

Outros preditores mostram contribuição menor, porém ainda relevante: faixa etária e município de realização da prova (que codificam heterogeneidades individuais e territoriais), bem como sexo e raça/cor (que operam como marcadores de desigualdades estruturais no

acesso a oportunidades educacionais e econômicas). Importante frisar que a presença de sexo e raça/cor entre os preditores não aponta para capacidades intrínsecas, mas para barreiras sistêmicas ainda ativas no país.

A interpretação das importâncias exige cautela. Elas dependem do conjunto de variáveis usado, dos hiperparâmetros e do alvo (neste caso, "abaixo da média") e não medem o efeito de intervenções no mundo real. Também cabe destacar que essas variáveis contribuem para padrões de grupo e não para predição individual.

Parte do que aparece como "renda", "tipo de escola", "língua", "escolaridade materna" e marcadores demográficos pode refletir também fatores não observados, como horas de estudo, qualidade de instrução e redes de apoio. Além disso, a ordem das importâncias é específica ao algoritmo e pode mudar com outras especificações de modelo (por exemplo inclusão de variáveis acadêmicas). Portanto, a leitura é associativa e deve ser contextualizada no desenho observacional (LUZ, 2025).

Embora a variável raça apareça em décimo lugar de importância no modelo, ela está intimamente relacionada à renda, que está em primeiro lugar. A demonstração dessa relação pode ser observada na Figura 03, que deve ser analisada sob dois aspectos: no primeiro, há um gradiente de renda nítido: em todas as raças/cores, as médias aumentam conforme cresce a renda familiar. Segundo, mesmo entre pessoas da mesma faixa de renda, persistem diferenças sistemáticas entre grupos: em geral, brancos e amarelos apresentam médias mais altas; pardos e pretos ficam abaixo; indígenas tendem aos menores valores. Nas faixas superiores, esses gaps diminuem, mas não desaparecem.

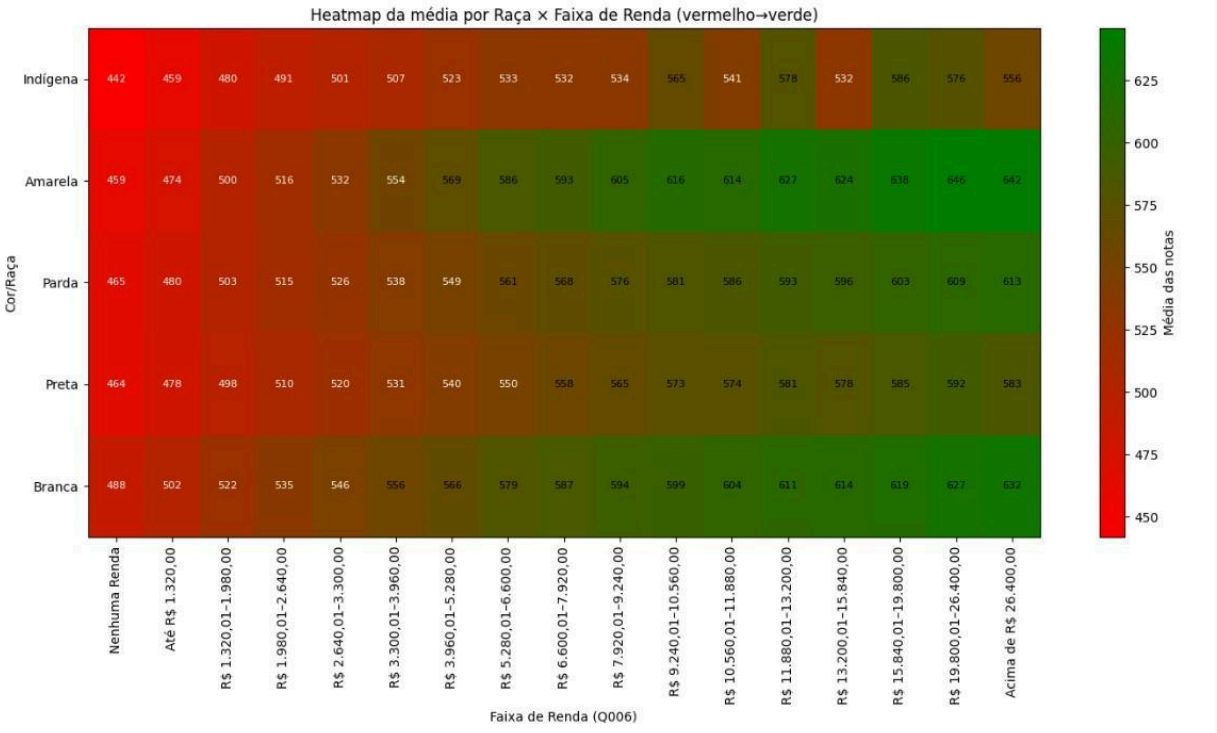

Figura 03 – Mapa de calor da média das notas por cor/raça e faixa de renda

Esses padrões não indicam causas individuais; funcionam como um mapa das desigualdades estruturais que atravessam raça/cor no Brasil, diferenças de trajetória escolar, oportunidades territoriais, capital cultural e barreiras institucionais que o gráfico apenas retrata, não explica. Portanto, a leitura é associativa: útil para orientar políticas públicas de mitigação de risco e focalização de recursos, sem atribuir qualquer traço de desempenho a

características intrínsecas dos grupos.

4.1. Usos recomendados para modelos

As Figuras 02 e 03 indicam que os principais preditores capturam desigualdades estruturais. Assim, as previsões deste estudo devem ser interpretadas em nível agregado e usadas exclusivamente para planejamento e monitoramento, não como base para decisões individuais. Esse enquadramento exige governança explícita ao longo do ciclo de vida do modelo (finalidade, base legal, minimização de dados e segurança), auditoria de desempenho por subgrupos e documentação pública de usos permitidos e proibidos.

A maior potencialidade dessa tecnologia está no uso por governos e tomadores de decisão em políticas públicas educacionais. Nesse cenário, as predições devem orientar a alocação de recursos em escala. O "mapa de risco" pode priorizar territórios, redes e grupos com maior concentração de vulnerabilidades e dimensionar respostas universalistas proporcionais, como fortalecimento de infraestrutura de estudo (dispositivos, conectividade e espaços adequados), programas territoriais de preparação para o ENEM (tutoria, apoio logístico e transporte) e ações de retomada para egressos com maior tempo desde a conclusão do ensino médio. Também se destacam iniciativas de ampliação da exposição linguística, especialmente em inglês, nas redes públicas, e ações sensíveis a gênero em Matemática. Em todos os casos, recomenda-se formato inclusivo, sem listas nominais e sem usos excludentes, reforçando que o alvo das intervenções são desigualdades estruturais, não indivíduos isolados.

Para operacionalizar essas salvaguardas, uma boa prática, especialmente ao lidar com raça/cor e sexo, é explicitar no contexto de uso que tais variáveis não servem para "rotular" pessoas, mas para acompanhar a justiça do processo. Isso pode assumir um arranjo em duas camadas: o modelo que orienta políticas não utiliza variáveis sensíveis, enquanto um modelo paralelo de auditoria as incorpora para verificar diferenças injustificadas entre grupos e, se necessário, corrigi-las. Nesse contexto, é importante definir metas explícitas de equidade e monitorá-las em painéis agregados, com métricas como *F1*, calibração e taxas de erro por subgrupo, sempre sem rastreio individual.

A governança desses modelos exige transparência e documentação acessível. Uma "Carta do Modelo" pode sintetizar a finalidade científica e social, os dados mínimos utilizados, os limites de decisão, a frequência de revisão, às responsabilidades institucionais e os usos proibidos, entre eles a produção de listas nominais para punição ou exclusão. Além disso, são essenciais salvaguardas éticas e legais, como minimização e anonimização de dados, retenção limitada, protocolos de segurança da informação e mecanismos claros de contestação e revisão de decisões. Dessa forma, o uso de modelos preditivos se mantém alinhado a princípios de equidade, proteção de direitos e prestação de contas à sociedade.

A comunicação com a comunidade escolar e com a sociedade precisa ser clara e pública. Mapas e tabelas agregadas (por escola, município, estado) ajudam a explicitar prioridades, metas para os próximos anos e o que já foi feito. Relatórios curtos, sem jargões técnicos, que apresentem resultados, limites e próximos passos, criam confiança e permitem correções de rota. Trata-se de encarar o risco como grau, calibrar e revisar periodicamente, cuidar da equidade entre grupos com acompanhamento simples e difundir as boas práticas. Assim, a previsão se torna "capacidade" de cuidar: priorizar melhor, apoiar melhor e reduzir desigualdades de aprendizagem com responsabilidade. Recomenda-se implementar as ações em ondas (pilotos regionais), com monitoramento e ajustes periódicos.

Adicionalmente, uma agenda de pesquisa translacional pode identificar territórios resilientes e intervenções com melhor relação custo-efetividade na redução de desigualdades, subsidiando a escala de políticas baseadas em evidências.

5. Conclusões

Ao demonstrar que predições de *machine learning* baseadas exclusivamente em variáveis socioeconômicas alcançam *F1* = 0,76 na identificação do risco de ficar abaixo da média no ENEM, este artigo sugere o uso deste tipo de tecnologia como instrumento de política pública orientado à mitigação de risco em escala, não a decisões individuais.

O principal valor está em mapear necessidades territoriais e sociodemográficas e dimensionar políticas universais proporcionais, como infraestrutura de estudo (dispositivos e conectividade), programas territoriais de preparação para o ENEM, ações de retomada para egressos e ampliação da exposição linguística.

Para que esse potencial se traduza em benefício público, o uso desses modelos deve ocorrer sob governança responsável e comunicável, com finalidades explícitas e interpretação agregada dos resultados. Do ponto de vista ético, as predições devem orientar a priorização de políticas com foco territorial, evitando personalização punitiva e garantindo que nenhuma pessoa seja excluída por um algoritmo.

Referências Bibliográficas

AKIBA, T.; SANO, S.; YANASE, T.; OHTA, T.; KOYAMA, M. *Optuna: A Next-generation Hyperparameter Optimization Framework*. arXiv:1907.10902, 2019.

ALI, A.; JAYARAMAN, R.; AZAR, E.; MAALOUF, M. *A comparative analysis of machine learning and statistical methods for evaluating building performance: a systematic review and future benchmarking framework*. *Building and Environment*, v. 252, art. 111268, 2024. Disponível em: https://doi.org/10.1016/j.buildenv.2024.111268. Acesso em: 11 jul. 2025.

ALMALAWI, A.; SOH, B.; LI, A.; SAMRA, H. *Predictive models for educational purposes: a systematic review*. *Big Data and Cognitive Computing*, v. 8, art. 187, 2024. Disponível em: https://doi.org/10.3390/bdcc8120187. Acesso em: 6 jul. 2025.

BERISHA, M.; MËZIU, E.; SHABANI, I. *Big data analytics in cloud computing: an overview*. *Journal of Cloud Computing*, v. 11, art. 24, 2022. Disponível em: https://doi.org/10.1186/s13677-022-00301-w. Acesso em: 16 out. 2025.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. *SMOTE: Synthetic Minority Over-sampling Technique*. *Journal of Artificial Intelligence Research*, v. 16, p. 321–357, 2002. DOI: 10.1613/jair.953.

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 (KDD 2016)*, 2016. p. 785–794. Disponível em: arXiv:1603.02754.

FAN, Z.; GOU, J.; WENG, S. Complementary CatBoost based on residual error for student performance prediction. *Pattern Recognition*, v. 161, 111265, 2025. Acesso em: 27 dez. 2025.

FERNÁNDEZ-DELGADO, M.; CERNADAS, E.; BARRO, S.; AMORIM, D. *Do we need hundreds of classifiers to solve real world classification problems? Journal of Machine Learning Research*, v. 15, p. 3133–3181, 2014. Acesso em: 16 jul. 2025.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP). *Microdados do Enem 2023*. Brasília: INEP, 2024.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP). Enem 2024 tem 4,3 milhões de inscritos confirmados. 2024. Publicado em: 23 out. 2024. Atualizado em: 30 jun. 2025. Disponível em: https://www.gov.br/inep/pt-br/centrais-de-conteudo/noticias/enem/enem-2024-tem-4-3-milhoes-de-inscritos-confirmados. Acesso em: 27/12/2025

JOSHI, A.; SAGGAR, P.; JAIN, R.; SHARMA, M.; GUPTA, D.; KHANNA, A. *CatBoost, an ensemble machine learning model for prediction and classification of student academic performance*. *Advances in Data Science and Adaptive Analysis*, v. 13, n. 2, 2021. DOI: 10.1142/S2424922X21410023. Acesso em: 11 jul. 2025.

JOSHI, Y.; MALLIBHAT, K.; VIJAYALAKSHMI, M. *Students' performance prediction using multimodal machine learning*. In: *IEEE IFEES World Engineering Education Forum – Global Engineering Deans Council (WEEF-GEDC)*, 2022. Anais... IEEE, 2022. Disponível em: https://doi.org/10.1109/WEEFGEDC54384.2022.9996212. Acesso em: 6 jul. 2025.

KE, G. et al. *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*. In: *Advances in Neural Information Processing Systems (NeurIPS 2017)*, v. 30, 2017. p. 3146–3154.

LECUN, Y.; BENGIO, Y.; HINTON, G. *Deep learning*. *Nature*, v. 521, n. 7553, p. 436–444, 2015.

LUZ, D. M. *Previsão do desempenho dos estudantes no ENEM por meio de modelos de aprendizado de máquina e embeddings semânticos: uma abordagem baseada em variáveis*. 2025. Dissertação (Mestrado em Matemática, Estatística e Computação Aplicadas à Indústria) — Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação, São Carlos, 2025.

MITCHELL, T. M. *Machine Learning*. 1. ed. New York: McGraw-Hill, 1997. Disponível em: https://www.cs.cmu.edu/~tom/mlbook.html. Acesso em: 20 out. 2025.

NOVAES, A. A.; PITON-GONÇALVES, J. *Predição do desempenho de estudantes por regressão múltipla*. *Sigmae*, v. 10, n. 1, p. 82–98, 2021. Disponível em: https://publicacoes.unifalmg.edu.br/revistas/index.php/sigmae/article/view/1552. Acesso em: 28 jul. 2025.

RAHAL, C.; VERHAGEN, M.; KIRK, D. *The rise of machine learning in the academic social sciences*. *AI & Society*, v. 39, p. 799–801, 2022.

RUBIM, J. A. D. O.; MANTOVANI, D.; ALAVARSE, O. M. *Exclusão Digital e seus impactos sobre a proficiência no ENEM: um estudo com concluintes do Ensino Médio de baixa renda entre 2015 e 2023*. *Educação em Revista*, v. 41, e59221, 2025. Acesso em: 21 dez. 2025.

PINTO, J. C. A.; TENÓRIO, R. M. *A influência dos fatores socioeconômicos no desempenho acadêmico dos estudantes de ensino médio integrado do IFBA/Campus Barreiras*. *Revista Sigma*, v. 15, n. 3, p. 45–58, 2021.

PROKHORENKOVA, L.; GUSEV, G.; VOROBEV, A.; DOROGUSH, A. V.; GULIN, A. *CatBoost: unbiased boosting with categorical features*. In: *Advances in Neural Information Processing Systems (NeurIPS 2018)*, v. 31, 2018.

SEVILLA, J.; BELTRÁN, M.; CABALLERO, J. *Machine learning in education: global trends and ethical challenges*. *Computers & Education Review*, v. 8, p. 101–118, 2022.

SIMÕES-GOMES, L.; ROBERTO, E.; MENDONÇA, J. *Viés algorítmico: um balanço provisório*. *Estudos de Sociologia*, Araraquara, v. 25, n. 48, 2020. Disponível em: https://doi.org/10.52780/res.13402. Acesso em: 16 jul. 2025.

SOUZA, M. de; KLUG, D. L. *Desigualdades educacionais no Enem: uma perspectiva baseada em variáveis socioeconômicas e aprendizagem de máquina*. *Regae: Revista de Gestão e Avaliação Educacional*, Santa Maria, v. 14, n. 23, e93251, 2025. doi:10.5902/2318133893251.