UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

CARLOS VINÍCIUS MARTINS DA SILVA

De Vilão a Solução: O Sobreajuste na Geografia da Desigualdade Oculta do ENEM

Fevereiro de 2026

Carlos Vinícius Martins da Silva

De Vilão a Solução: O Sobreajuste na Geografia da Desigualdade Oculta do ENEM

Artigo apresentado ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção, do título de Bacharel em Sistemas de Informação.

Universidade Federal Rural de Pernambuco – UFRPE
Departamento de Estatística e Informática
Curso de Bacharelado em Sistemas de Informação

Orientador: Gabriel Alves De Albuquerque Junior

Recife
Fevereiro de 2026

CARLOS VINÍCIUS MARTINS DA SILVA

DE VILÃO A SOLUÇÃO: O OVERFITTING E XAI NA GEOGRAFIA DA DESIGUALDADE OCULTA DO ENEM

Artigo apresentado ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Aprovado em: 19 de fevereiro de 2026.

BANCA EXAMINADORA

Prof. Gabriel Alves De Albuquerque Junior (Orientador)
Departamento de Estatística e Informática – UFRPE

Prof. Lucas Fernando da Silva Cambuim
Departamento de Estatística e Informática – UFRPE

De Vilão a Solução: O Sobreajuste na Geografia da Desigualdade Oculta do ENEM

Carlos Vinícius Martins da Silva [1], **Gabriel Alves de Albuquerque Junior** [1]

[1]Departamento de Estatística e Informática – Universidade Federal Rural de Pernambuco
Rua Dom Manuel de Medeiros, s/n, - CEP: 52171-900 – Recife – PE – Brasil

carlos.vmsilva@ufrpe.br, gabriel.alves@ufrpe.br

Resumo. *O ENEM é uma base essencial para a análise educacional brasileira, mas aplicações de aprendizado de máquina na área costumam ter como objetivo a predição. Este trabalho propõe uma abordagem distinta: utiliza modelos interpretáveis para mapear padrões de desempenho e perfis socioeconômicos. A metodologia combina estatística descritiva com árvores de decisão submetidas ao sobreajuste (overfitting) intencional. O objetivo não é a generalização, mas a exaustão descritiva da base original, utilizando a renda familiar como alvo (target) instrumental para extrair regras que influenciam o desempenho em cada estrato social. Para quantificar a relevância dos fatores, aplicou-se a técnica SHAP (XAI) em cenário multiclasse. Os resultados confirmam que o desempenho acadêmico cresce proporcionalmente à renda, mas revelam nuances críticas: embora a posse de automóvel seja o principal determinante global de renda, a exclusão digital superou a imobilidade física como principal marcador de vulnerabilidade no estrato rural feminino em 2023. Adicionalmente, candidatos rurais têm maior dificuldade em converter renda em notas superiores, embora mulheres nesse contexto apresentem maior eficiência nessa conversão que homens. Conclui-se que esta abordagem revela desigualdades interseccionais que modelos preditivos convencionais ocultam.*

Abstract. *ENEM is an essential database for Brazilian educational analysis, but machine learning applications in the field typically aim for prediction. This work proposes a distinct approach: it utilizes interpretable models to map performance patterns and socioeconomic profiles. The methodology combines descriptive statistics with decision trees subjected to intentional overfitting. The goal is not generalization, but rather the descriptive exhaustion of the original database, using family income as an instrumental target to extract rules that influence performance within each social stratum. To quantify the relevance of the factors, the SHAP (XAI) technique was applied in a multiclass scenario. The results confirm that academic performance increases proportionally with income, but reveal critical nuances: although car ownership is the main global determinant of income, digital exclusion surpassed physical immobility as the primary marker of vulnerability in the female rural stratum in 2023. Additionally, rural candidates face greater difficulty in converting income into higher scores, although women in this context demonstrate greater efficiency in this conversion than men. It is concluded that this approach reveals intersectional inequalities that conventional predictive models obscure.*

1. Introdução

O Exame Nacional do Ensino Médio (ENEM) constitui uma das principais avaliações educacionais em larga escala no Brasil, sendo amplamente utilizado como instrumento de acesso ao ensino superior e como fonte de dados para estudos sobre o desempenho educacional dos estudantes. Os microdados públicos do ENEM oferecem um conjunto abrangente de informações que incluem características demográficas, educacionais e socioeconômicas dos participantes, possibilitando análises aprofundadas sobre os fatores associados ao desempenho acadêmico.

Nos últimos anos, técnicas de aprendizado de máquina têm sido cada vez mais aplicadas à análise de dados educacionais, com foco predominante na construção de modelos preditivos capazes de estimar resultados acadêmicos ou identificar riscos de baixo desempenho. Entretanto, abordagens exclusivamente preditivas tendem a limitar a compreensão dos padrões internos dos dados, dificultando a interpretação dos fatores que influenciam o desempenho dos estudantes.

Apesar da crescente disponibilidade de dados, observa-se que grande parte dos estudos que empregam técnicas de aprendizado de máquina no contexto educacional concentra-se na tarefa de previsão de desempenho, evasão ou aprovação, frequentemente priorizando métricas de acurácia [Baker and Inventado 2019]. Embora essas abordagens sejam relevantes, elas tendem a utilizar modelos complexos e de difícil interpretação, o que limita a compreensão dos padrões subjacentes aos dados e dificulta a análise crítica dos fatores que influenciam o desempenho dos estudantes. Nesse sentido, existe uma lacuna na literatura quanto à utilização de modelos de aprendizado de máquina com finalidade explicitamente descritiva e interpretável, voltada à exploração e compreensão dos dados, em vez da previsão.

A ausência de interpretações claras pode comprometer a utilização dos resultados por gestores, educadores e formuladores de políticas públicas, que necessitam de explicações transparentes para embasar decisões [Doshi-Velez and Kim 2017, Lipton 2018]. Diante dessa lacuna, torna-se relevante investigar abordagens que conciliem técnicas de aprendizado de máquina com interpretabilidade e explicabilidade, termos que, embora possuam distinções teóricas, são frequentemente utilizados de forma intercambiável na literatura de modelos preditivos aplicados a dados tabulares [Molnar 2022]. A adoção de modelos inerentemente interpretáveis, como árvores de decisão, aliada a métodos de explicabilidade, permite não apenas identificar padrões recorrentes nos dados, mas também compreender de forma clara e acessível as relações entre variáveis socioeconômicas e desempenho educacional, justificando a proposta deste trabalho.

1.1. Objetivos

O objetivo geral deste trabalho é realizar uma análise descritiva e interpretável dos fatores que caracterizam o perfil dos estudantes do ENEM, investigando as relações de dependência entre variáveis educacionais e de infraestrutura, tendo como variável alvo (*target*) a renda mensal familiar. Para isso, o modelo de árvore de decisão foi sobreajustado (*overfitting*) para atingir a exaustão descritiva dos dados, permitindo identificar os divisores específicos de cada estrato social.

Como objetivos específicos, destacam-se:

- Realizar análise estatística descritiva dos microdados do ENEM (2018, 2019 e 2023) para caracterização do perfil dos estudantes;
- Selecionar preditores socioeconômicos e demográficos para analisar o impacto da estrutura de renda no desempenho dos candidatos;
- Utilizar a média geral das notas do ENEM como métrica central para avaliar o impacto do contexto socioeconômico no desempenho dos estudantes;
- Utilizar árvores de decisão sobreajustadas (*overfitting*) para extrair padrões de desempenho e perfil socioeconômico, tendo a renda familiar como eixo central de análise;
- Avaliar a importância das variáveis utilizando métricas internas das árvores de decisão e valores SHAP para garantir a interpretabilidade do modelo;
- Comparar os resultados entre os anos analisados para identificar padrões ou variações temporais na relação entre renda e desempenho;

2. Referencial Teórico

Este capítulo apresenta a fundamentação teórica necessária para o embasamento desta pesquisa, estabelecendo um diálogo entre os indicadores socioeconômicos da educação brasileira e as ferramentas avançadas da ciência de dados. Compreender o desempenho dos estudantes no ENEM exige mais do que uma análise estatística convencional; demanda uma infraestrutura tecnológica robusta e flexível. Para tanto, aborda-se primeiramente a linguagem de programação Python, base do ecossistema de desenvolvimento deste trabalho devido à sua vasta gama de bibliotecas voltadas ao tratamento de dados e aprendizado de máquina. Em seguida, exploram-se os conceitos de aprendizado de máquina, com foco em modelos de árvores de decisão, e as técnicas de Inteligência Artificial Explicável (XAI), especificamente os valores SHAP. Essa base teórica é essencial para sustentar a transição de um modelo puramente preditivo para uma abordagem interpretável e descritiva, permitindo que os padrões de desigualdade identificados nos microdados sejam compreendidos em sua totalidade.

2.1. Aprendizado de Máquina e Modelos de Árvores de Decisão

O aprendizado de máquina (*Machine Learning*) representa um subcampo da inteligência artificial que se concentra no desenvolvimento de algoritmos capazes de aprender padrões a partir de dados e realizar predições ou classificações sem serem explicitamente programados para cada tarefa [Géron 2021]. Diferente da inferência estatística clássica, que muitas vezes assume distribuições de dados predefinidas, o aprendizado de máquina adota uma abordagem orientada por dados, buscando otimizar uma função de desempenho através da experiência [Mitchell 1997].

Dentre as diversas arquiteturas de modelos, as árvores de decisão (*Decision Trees*) destacam-se por sua capacidade de mapear relações não lineares de forma intuitiva. Tecnicamente, uma árvore de decisão realiza uma partição recursiva do espaço de atributos em regiões disjuntas, denominadas nós, onde cada divisão é escolhida para maximizar a pureza dos dados resultantes [James et al. 2013]. Esse processo de divisão é governado por critérios matemáticos como a impureza de gini ou a entropia (Ganho de Informação), que medem a homogeneidade das classes dentro de cada subconjunto [Faceli et al. 2011].

A estrutura de uma árvore é composta por um nó raiz, nós internos (que representam testes em atributos específicos) e nós folha, que contêm a classificação final ou o

valor predito. Uma característica fundamental desse modelo é a sua transparência, o que o classifica como um modelo de caixa-branca, permitindo que o caminho lógico da decisão seja rastreado desde a raiz até a folha [Molnar 2022]. Embora árvores muito profundas possam apresentar riscos de sobreajuste (*overfitting*), capturando ruídos do conjunto de treinamento, essa profundidade também permite o mapeamento minucioso de interações complexas entre variáveis, o que as torna ferramentas poderosas para a análise descritiva de grandes bases de dados [Hastie et al. 2009].

2.2. Inteligência Artificial Explicável (XAI) e Valores SHAP

Com o aumento da complexidade dos modelos de Aprendizado de Máquina, surgiu o desafio da caixa-preta, onde modelos de alta performance oferecem pouca ou nenhuma visibilidade sobre como chegam a determinadas conclusões. Nesse contexto, a inteligência artificial explicável (*eXplainable Artificial Intelligence* - XAI) emerge como uma área de pesquisa dedicada a tornar os resultados dos algoritmos compreensíveis para seres humanos, garantindo transparência, confiança e auditabilidade [Adadi and Berrada 2018]. A XAI é particularmente relevante em domínios sensíveis, como a educação e análise socioeconômica, onde entender os vieses e os fatores determinantes é tão importante quanto a precisão da predição em si [Arrieta et al. 2020]. Dentre as técnicas de XAI, destaca-se o uso de métodos *post-hoc* agnósticos ao modelo, que permitem interpretar predições de qualquer algoritmo após o seu treinamento. O *framework* SHAP (*SHapley Additive exPlanations*), proposto por Lundberg e Lee [Lundberg and Lee 2017], fundamenta-se na teoria dos jogos cooperativos para atribuir a cada atributo o seu devido valor de contribuição para o resultado final. A essência do SHAP reside nos Valores de Shapley, que distribuem o pagamento (a predição) entre os jogadores (os atributos), considerando todas as interações possíveis entre eles [Shapley 1953].

A grande vantagem da utilização dos valores SHAP em análises descritivas é a garantia de propriedades matemáticas desejáveis, como a consistência e a aditividade local. Enquanto métodos tradicionais de importância de variáveis podem fornecer apenas uma visão global, o SHAP permite uma análise tanto global quanto local, identificando como uma variável específica (ex: posse de computador ou renda familiar) aumenta ou diminui a probabilidade de um candidato pertencer a uma determinada classe socioeconômica ou faixa de desempenho [Molnar 2022]. Assim, a XAI transforma modelos estatísticos em ferramentas de diagnóstico social, permitindo a identificação precisa de barreiras e facilitadores no desempenho acadêmico.

2.3. Mineração de Dados Educacionais (MDE)

A Mineração de Dados Educacionais (*Educational Data Mining* - EDM) é uma disciplina emergente que se ocupa do desenvolvimento de métodos para explorar tipos únicos de dados provenientes de contextos educacionais. O objetivo central da MDE é utilizar essas técnicas para compreender melhor os estudantes e os ambientes nos quais eles aprendem, fornecendo subsídios para a melhoria dos processos educacionais e a tomada de decisões baseada em evidências [Romero and Ventura 2010]. Diferente da mineração de dados convencional, a MDE lida com dados que possuem uma hierarquia intrínseca e complexidades sociais que exigem uma interpretação contextualizada [Baker and Yacef 2009].

As aplicações de MDE são amplas, abrangendo desde a predição do desempenho acadêmico e a identificação de riscos de evasão até a descoberta de padrões de com-

portamento em ambientes virtuais de aprendizagem. No contexto de exames de larga escala, como o ENEM, a MDE permite realizar a análise de populações heterogêneas, identificando como variáveis demográficas, sociais e econômicas se inter-relacionam para moldar a trajetória do estudante [Castro et al. 2007]. Segundo Romero e Ventura [Romero and Ventura 2020], a mineração de dados na educação não deve focar apenas na precisão algorítmica, mas também na interpretabilidade dos modelos, para que os gestores educacionais possam implementar políticas públicas eficazes.

Nesta perspectiva, a MDE divide-se frequentemente em abordagens preditivas e descritivas. Enquanto a predição busca antecipar resultados futuros, a mineração descritiva, foco desta área de estudo, busca caracterizar as propriedades dos dados e encontrar associações ocultas entre os perfis socioeconômicos e o rendimento escolar [Peña-Ayala 2014]. Assim, a MDE atua como uma ponte entre a Ciência da Computação e as Ciências da Educação, transformando grandes volumes de dados brutos em conhecimento acionável sobre a realidade educacional de um país.

2.4. O ENEM e a Sociologia da Educação: Interseccionalidade e Novas Exclusões

O Exame Nacional do Ensino Médio (ENEM), embora consolidado como o principal mecanismo de democratização do acesso ao ensino superior no Brasil, atua também como um observatório das desigualdades estruturais do sistema educacional brasileiro. Sob a ótica da sociologia da educação, o desempenho acadêmico não é um fenômeno puramente meritocrático, mas sim o reflexo do acúmulo de capital cultural e econômico das famílias [Bourdieu and Passeron 2007]. Segundo a teoria da reprodução de Bourdieu, a escola tende a converter privilégios sociais em êxitos escolares, perpetuando a estratificação social através de exames padronizados.

A análise dessas desigualdades exige uma perspectiva interseccional, conceito desenvolvido por Kimberlé Crenshaw, que propõe que as categorias sociais como raça, classe e gênero não operam de forma isolada, mas se sobrepõem, criando sistemas únicos de opressão ou privilégio [Crenshaw 1989]. No contexto do ENEM, a interseccionalidade permite compreender como o isolamento geográfico do estudante rural, somado ao gênero e à cor/raça, potencializa barreiras que não seriam plenamente explicadas por uma única variável. Como aponta [Silva 2020], a desigualdade educacional brasileira é multidimensional e afeta de forma mais severa estudantes que acumulam múltiplas vulnerabilidades.

Ademais, o cenário contemporâneo introduziu o que a literatura denomina como novas exclusões, com destaque para o fosso digital e as barreiras de mobilidade. A exclusão digital, manifestada pela falta de acesso a dispositivos e conexão de qualidade, tornou-se um determinante crítico de desempenho, especialmente no período pós-pandemia, configurando-se como uma nova forma de capital cultural técnico [Sorj 2003]. Somado a isso, a carência de infraestrutura física e mobilidade no meio rural impõe um custo de acesso que diferencia o estudante do campo do estudante urbano. Portanto, o uso de modelos interpretáveis para analisar o ENEM fundamenta-se na necessidade de desvelar como esses fatores interseccionais e as novas exclusões tecnológicas moldam a pirâmide social e educacional do país.

3. Trabalhos relacionados

Diversos estudos têm aplicado modelos de aprendizado de máquina e de técnicas de explicabilidade na área da educação, especialmente no contexto da análise do desempenho

acadêmico e da identificação de fatores associados ao aprendizado. Essas abordagens têm sido empregadas tanto para fins preditivos quanto para a interpretação dos padrões presentes em dados educacionais. Esta seção apresenta uma revisão de trabalhos que adotam metodologias semelhantes à proposta deste estudo, com foco na aplicação de modelos de classificação, na análise da relevância de variáveis e no uso de técnicas explicáveis para interpretar resultados relacionados ao desempenho escolar. A seguir, são descritos os principais trabalhos relacionados, destacando seus objetivos, métodos utilizados e contribuições para a área, bem como suas diferenças em relação à abordagem descritiva e interpretável adotada neste trabalho.

No estudo de [Villas Boas 2023], os autores exploram a interpretabilidade de modelos de *Machine Learning* aplicados aos microdados do ENEM por meio da metodologia de *Shapley Values* (SHAP). O trabalho destaca que, embora modelos baseados em boosting — implementados no referido estudo através do algoritmo LightGBM — ofereçam alto poder preditivo, sua arquitetura complexa exige técnicas de explicabilidade para revelar como as características socioeconômicas influenciam na probabilidade de um participante alcançar notas mínimas no exame. Conforme as conclusões do autor, a estratificação por renda emerge como o fator de maior peso na probabilidade de êxito do candidato. O modelo revela que essa influência é potencializada pela combinação de indicadores socioeconômicos e demográficos, tais como o acesso a computadores, a idade, o grau de instrução familiar e a cor da pele.

A abordagem de [Villas Boas 2023] foca em identificar os fatores que mais impactam o funcionamento desse modelo de *boosting*. Em convergência com esse estudo, a presente pesquisa também adota os valores SHAP como ferramenta central de explicabilidade. Contudo, este trabalho diferencia-se ao deslocar o foco da performance preditiva para uma finalidade estritamente descritiva. Enquanto o estudo referenciado utiliza o SHAP para interpretar as predições de um modelo de classificação baseado em *boosting*, este TCC utiliza o método para realizar uma exaustão analítica sobre árvores de decisão deliberadamente sobreajustadas. Essa escolha permite uma investigação mais profunda sobre a interseccionalidade de gênero e localização geográfica, detalhando como o peso das variáveis se altera em recortes populacionais específicos, sem a necessidade de recorrer a modelos de alta complexidade para validar os padrões observados.

A literatura também tem destacado a importância da explicabilidade de modelos de aprendizado de máquina no contexto educacional, especialmente em aplicações voltadas à análise de desempenho acadêmico e suporte à tomada de decisão. Na revisão sistemática conduzida por [Silva and Santana 2024], os autores investigaram o panorama da *eXplainable Artificial Intelligence* (XAI) na educação entre 2012 e 2024, evidenciando um crescimento expressivo no interesse da comunidade científica pelo tema nos últimos seis anos. O estudo aponta que modelos baseados em Árvores de Decisão são os mais recorrentes na literatura (correspondendo a 40% dos trabalhos analisados), seguidos por Redes Neurais. Além disso, os métodos SHAP e LIME consolidaram-se como as principais técnicas de explicabilidade *post-hoc*, estando presentes em 40% e 35% das pesquisas, respectivamente.

Apesar desse avanço, [Silva and Santana 2024] ressaltam que a explicabilidade ainda é tratada frequentemente como um aspecto secundário em relação à performance preditiva, o que pode limitar a utilidade prática desses modelos em cenários que exigem

transparência e decisões seguras. Em consonância com essas discussões, o presente estudo reforça a relevância da XAI ao adotar uma abordagem onde a explicabilidade não é apenas um acessório, mas o núcleo da análise. Ao utilizar árvores de decisão e o método SHAP de forma centralizada sobre os microdados do ENEM, este trabalho alinha-se às tendências tecnológicas identificadas pela revisão de Silva e Santana, diferenciando-se, contudo, ao priorizar a exaustão descritiva e a extração de regras claras em detrimento da mera acurácia de predição.

Estudos recentes também têm explorado o uso de técnicas baseadas em valores de Shapley para aumentar a interpretabilidade de modelos de aprendizado de máquina aplicados à educação, como demonstrado por [Choi et al. 2024]. Em sua pesquisa sobre o desempenho de estudantes em cursos de programação *online*, os autores compararam diversos algoritmos, incluindo *Random Forest*, *Extra Trees*, *CatBoost* e *XGBoost*, identificando que o modelo *Extra Trees* apresentou a melhor performance preditiva. O diferencial do estudo reside na aplicação do método SHAP para elevar a transparência do modelo, permitindo a transição de uma análise puramente métrica para uma compreensão das contribuições de variáveis individuais e globais por meio de *summary*, *bar* e *dependence plots*. Os achados de [Choi et al. 2024] reforçam que mecanismos explicáveis são essenciais para evitar conclusões simplificadas ou enviesadas em sistemas educacionais baseados em IA.

Em convergência com esses autores, o presente trabalho adota o SHAP como ferramenta central de explicabilidade e utiliza visualizações análogas para interpretar o comportamento do modelo. Todavia, este estudo distancia-se da abordagem de [Choi et al. 2024] em dois pontos fundamentais: primeiro, ao substituir algoritmos de *ensemble* de alta complexidade por modelos de árvore de decisão intrinsecamente interpretáveis; e segundo, ao redirecionar o foco da eficácia da predição para uma análise estritamente descritiva e multiclasse, voltada à extração de padrões e regras de associação vinculadas às faixas de renda dos participantes do ENEM.

A aplicação de técnicas de *Machine Learning* para a predição e explicação do desempenho acadêmico em nível universitário também é discutida por [Vinces-Vinces and Flores-Sanchéz 2025]. No estudo de [Vinces-Vinces and Flores-Sanchéz 2025], os autores avaliaram o impacto de fatores sociodemográficos e acadêmicos sobre o rendimento de estudantes com histórico de reprovação, comparando algoritmos como Regressão Logística, Árvores de Decisão e *Random Forest*. O trabalho concluiu que o modelo *Random Forest* apresentou a melhor capacidade preditiva, e utilizou o método SHAP para conferir transparência às decisões do modelo, identificando que o histórico acadêmico prévio e a frequência escolar são os principais preditores de sucesso. Os resultados de [Vinces-Vinces and Flores-Sanchéz 2025] reforçam que o uso de técnicas de explicabilidade torna os modelos mais aceitáveis e úteis em ambientes educacionais ao evitar conclusões simplificadas. Em convergência com essa investigação, o presente estudo também enfatiza a importância da XAI para a compreensão de fenômenos educacionais complexos. Contudo, este trabalho diferencia-se ao deslocar a escala da análise do nível universitário para o ensino básico nacional (microdados do ENEM) e ao inverter a lógica da variável *target*. Enquanto o estudo referenciado foca na predição direta do rendimento acadêmico para prevenir a evasão, este TCC utiliza a faixa de renda como eixo central, aplicando o SHAP em um

cenário multiclasse para descrever os padrões socioeconômicos que definem o perfil dos participantes, priorizando a fidelidade descritiva em detrimento da predição de notas.

A investigação de padrões educacionais por meio de técnicas de aprendizado de máquina também é explorada por [de Souza and Klug 2025], que analisaram as desigualdades no ENEM 2023 sob a ótica de variáveis socioeconômicas. Utilizando o algoritmo *Random Forest* e técnicas de balanceamento de dados como o SMOTE, os autores desenvolveram modelos preditivos com alta performance, alcançando acurácias próximas a 94%. O estudo destaca que a escolaridade e ocupação dos pais, somadas à renda familiar, emergem como os principais preditores do sucesso acadêmico, fornecendo subsídios para políticas educacionais mais equitativas. Em convergência com os achados de [de Souza and Klug 2025], o presente estudo também identifica a centralidade da renda e a escolaridade e ocupação dos pais na configuração do perfil dos estudantes. Entretanto, a presente pesquisa diferencia-se metodologicamente ao abdicar da finalidade estritamente preditiva e do uso de técnicas de reamostragem sintética (como o SMOTE), optando por uma abordagem explicitamente descritiva e exploratória. Enquanto [de Souza and Klug 2025] priorizam a eficácia classificatória e a generalização dos modelos, este trabalho utiliza árvores de decisão deliberadamente sobreajustadas aos microdados originais para realizar uma mineração de regras exaustiva. Além disso, a integração do método SHAP em um cenário multiclasse neste TCC permite uma interpretação granular de como cada faixa de renda é caracterizada, indo além da importância global de atributos apresentada em modelos de *Random Forest* convencionais.

4. Metodologia

Esta seção descreve os procedimentos metodológicos que fundamentam este trabalho, estruturados em um fluxo que integra o tratamento de dados, a modelagem descritiva e a inteligência artificial explicável. A abordagem possui caráter exploratório e descritivo com base quantitativa, distanciando-se do foco preditivo tradicional para priorizar a compreensão das regras intrínsecas e dos padrões internos do conjunto de dados analisado.

Para proporcionar uma visão clara do processo, o método foi estruturado nas seguintes etapas:

1. **Tratamento e Caracterização:** Limpeza e pré-processamento dos microdados do ENEM, seguidos pela aplicação de estatística descritiva para traçar o perfil inicial dos estudantes e calcular a média aritmética de desempenho.
2. **Modelagem Instrumental:** Implementação de modelos de árvores de decisão deliberadamente submetidos ao sobreajuste (*overfitting*). Nesta etapa, a renda familiar é utilizada como alvo (*target*) instrumental para que o modelo esgote as associações entre o contexto socioeconômico e o desempenho acadêmico.
3. **Extração de Conhecimento (XAI):** Aplicação da técnica *SHAP* (*Explainable AI*) sobre os modelos gerados, permitindo quantificar e interpretar o impacto de cada variável explicativa na definição dos estratos sociais e de desempenho.
4. **Análise Comparativa Temporal:** Cruzamento e comparação dos padrões identificados nos anos de 2018, 2019 e 2023, visando identificar a evolução de desigualdades estruturais e nuances interseccionais.

As análises foram conduzidas utilizando a linguagem de programação *Python*, com o suporte das bibliotecas `pandas` e `numpy` para manipulação de dados,

`scikit-learn` para a construção dos modelos de árvore e `shap` para a implementação das técnicas de explicabilidade.

4.1. Base de Dados

O conjunto de dados utilizado neste trabalho é composto pelos microdados públicos do Exame Nacional do Ensino Médio (ENEM), referentes aos anos de 2018, 2019 e 2023, disponibilizados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Os microdados do ENEM consistem em registros individuais dos participantes do exame e incluem informações demográficas, educacionais, socioeconômicas, bem como as notas obtidas nas áreas avaliadas e na redação.

A escolha desses três anos teve como objetivo ampliar a base de análise, permitindo a observação de padrões recorrentes ao longo do tempo e reduzindo a influência de particularidades específicas de uma única edição do exame. Os anos de 2018 e 2019 representam o período imediatamente anterior à pandemia de COVID-19, servindo como base comparativa de estabilidade, enquanto o ano de 2023 fornece o panorama mais recente após as transformações impostas pelo ensino remoto. Cada edição do ENEM apresenta estrutura semelhante de variáveis, o que possibilitou a integração dos dados após a aplicação de procedimentos consistentes de pré-processamento. Adicionalmente, a delimitação até o ano de 2023 fundamenta-se na manutenção da consistência metodológica dos microdados. A partir de 2024, o questionário socioeconômico e os critérios de classificação das escolas entre as zonas urbana e rural sofreram alterações pelo INEP, o que poderia comprometer a comparabilidade direta dos padrões de inferência e os recortes populacionais estabelecidos nesta pesquisa.

Para cada ano analisado, foram consideradas as notas das quatro áreas do conhecimento — Ciências Humanas, Ciências da Natureza, Linguagens e Matemática — além da nota da redação, a média utilizada nas análises deste trabalho considera a média aritmética dessas 5 notas. Adicionalmente, foram utilizadas variáveis provenientes do questionário socioeconômico do participante, abrangendo aspectos como renda familiar, características domiciliares e posse de bens. Cabe destacar que todos os dados utilizados são de acesso público, anonimizados e disponibilizados para fins de pesquisa, não havendo qualquer identificação individual dos participantes.

4.2. Pré-processamento dos Dados

No pré-processamento dos dados, foi realizada a seleção das variáveis relevantes para a análise, considerando informações demográficas, educacionais, socioeconômicas e de desempenho dos participantes. Foram incluídas variáveis demográficas, como faixa etária, sexo, estado civil, cor/raça e nacionalidade; variáveis educacionais relacionadas à trajetória escolar, como tipo de escola, situação de conclusão do ensino médio, ano de conclusão e localização da escola; variáveis de desempenho correspondentes às notas das áreas avaliadas e da redação; e variáveis socioeconômicas provenientes do questionário do participante, relacionadas à renda familiar e à posse de bens e infraestrutura domiciliar. A descrição das variáveis utilizadas é apresentada na Tabela 1. A descrição detalhada das questões Q001 a Q025 do questionário socioeconômico, consideradas nesta análise, é apresentada no Apêndice A.

Tabela 1. Grupos de variáveis utilizadas na análise

Grupo de variáveis	Variáveis	Descrição
Demográficas	TP_FAIXA_ETARIA TP_SEXO TP_ESTADO_CIVIL TP_COR_RACA TP_NACIONALIDADE	Características pessoais e demográficas dos participantes
Educacionais	TP_ST_CONCLUSAO TP_ANO_CONCLUIU TP_ESCOLA TP_ENSINO TP_LOCALIZACAO_ESC	Informações sobre a trajetória escolar e o tipo de instituição
Desempenho	NU_NOTA_CN NU_NOTA_CH NU_NOTA_LC NU_NOTA_MT NU_NOTA_REDACAO	Notas obtidas nas áreas avaliadas do ENEM
Socioeconômicas	Q001 a Q025	Informações sobre renda familiar, bens e infraestrutura domiciliar

As variáveis categóricas selecionadas foram transformadas por meio da técnica de One-Hot Encoding, com o objetivo de representá-las adequadamente para a construção do modelo de árvore de decisão. Essa transformação foi realizada utilizando um pré-processador integrado ao pipeline do modelo, possibilitando a criação de variáveis binárias (*Dummies*) indicativas das categorias observadas nos dados. As variáveis numéricas foram mantidas em sua forma original, sem aplicação de normalização ou padronização, uma vez que modelos baseados em árvores de decisão não são sensíveis à escala dos atributos, realizando particionamentos com base em limiares dos valores observados.

Além da seleção das variáveis, foi realizado o tratamento de valores nulos presentes nos microdados. A ocorrência de valores ausentes está associada a não respostas dos participantes ou a registros incompletos no processo de coleta dos dados. Considerando o caráter descritivo da pesquisa, adotou-se uma abordagem conservadora para o tratamento desses valores, de modo a não introduzir distorções artificiais nos padrões observados. No tratamento dos valores nulos, adotou-se uma estratégia diferenciada de acordo com a natureza das variáveis, seguindo práticas recomendadas para a preparação de dados em aprendizado de máquina [Géron 2021]. As variáveis categóricas tiveram seus valores ausentes preenchidos pela moda, por se tratar da categoria mais frequente e por preservar a distribuição observada dos dados. Para as variáveis numéricas contínuas, correspondentes às notas das áreas avaliadas e da redação, os valores ausentes foram tratados por meio da imputação pela média. Essa abordagem, fundamentada em [Géron 2021], foi realizada previamente à etapa de modelagem, durante a preparação dos dados. De acordo com a abordagem de [Géron 2021], a adoção de critérios distintos para o tratamento dos valores nulos permite preservar as características essenciais de cada tipo de variável, minimizando a introdução de distorções e garantindo a coerência entre os dados utilizados e os objetivos descritivos do estudo.

Para tornar a variável de renda familiar mais interpretável e adequada aos objetivos descritivos deste estudo, as categorias originais da questão Q006 do questionário

socioeconômico do ENEM foram consolidadas em faixas de renda expressas em salários mínimos.

O Instituto Brasileiro de Geografia e Estatística (IBGE), por meio de levantamentos como a Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua) e da publicação *Síntese de Indicadores Sociais*, utiliza classes de rendimento baseadas em múltiplos do salário mínimo como forma de caracterizar as condições de vida da população brasileira [IBGE 2022]. Alinhado a essas práticas, o presente estudo adotou o reagrupamento das categorias originais da variável Q006 em faixas de renda mais amplas. O mapeamento realizado, bem como as correspondências entre as categorias originais do questionário e as faixas de renda adotadas neste trabalho, estão apresentados na Tabela 2.

Tabela 2. Mapeamento da variável Q006 para faixas de renda em salários mínimos

Categorias originais (Q006)	Faixa de renda
A–D	Até 2 salários mínimos
E	De 2 a 3 salários mínimos
F–J	De 3 a 6 salários mínimos
K–N	De 6 a 10 salários mínimos
O–P	De 10 a 15 salários mínimos
Q	Acima de 15 salários mínimos

Esse procedimento permitiu reduzir a granularidade da variável, preservando sua interpretação socioeconômica e tornando-a mais adequada às análises descritivas e à aplicação dos modelos de aprendizado de máquina utilizados. A descrição detalhada das categorias originais da questão Q006 do questionário socioeconômico do ENEM é apresentada no Apêndice B.

4.3. Análise Estatística Descritiva e Variável de Desempenho

Como parte fundamental da abordagem metodológica, realiza-se uma análise estatística descritiva com o objetivo de caracterizar a distribuição dos dados e identificar tendências centrais e dispersões no perfil dos estudantes. Esta etapa fornece a fundamentação quantitativa para a compreensão das desigualdades presentes nos microdados e valida a consistência das variáveis selecionadas para compor o estudo. Para representar o desempenho acadêmico, utiliza-se a Nota Média Geral (NMG), que consiste na média aritmética simples das cinco competências avaliadas. A utilização de siglas simplificadas na Equação 1 visa facilitar a leitura matemática, mantendo a correspondência direta com as colunas extraídas da base de dados do INEP:

$$NMG = \frac{N_{CH} + N_{CN} + N_{LC} + N_{MT} + N_{RED}}{5}, \tag{1}$$

Onde cada termo corresponde a uma variável específica da base de dados original:

- N_{CH}: Refere-se à nota em Ciências Humanas (`NU_NOTA_CH`);
- N_{CN}: Refere-se à nota em Ciências da Natureza (`NU_NOTA_CN`);
- N_{LC}: Refere-se à nota em Linguagens e Códigos (`NU_NOTA_LC`);
- N_{MT}: Refere-se à nota em Matemática (`NU_NOTA_MT`);

- N_{RED}: Refere-se à nota da Redação (`NU_NOTA_REDACAO`).

A análise descritiva contempla o cálculo de medidas de posição (média e mediana) e de dispersão (desvio padrão), permitindo observar as flutuações de desempenho conforme os recortes de sexo, localização da escola e as faixas de renda familiar. Complementarmente, são geradas tabelas de frequência para as variáveis categóricas do questionário socioeconômico, estabelecendo os subsídios necessários para a inferência descritiva realizada pelos modelos de árvore de decisão intencionalmente sobreajustados.

4.4. Modelo de Árvore de Decisão e Inferência Descritiva

Para a construção dos modelos, a faixa de renda familiar (extraída e reagrupada a partir da questão Q006 do questionário socioeconômico) foi definida como a variável-alvo (*target*). É imperativo destacar que, embora o modelo utilize uma estrutura típica de classificação, o *target* é empregado exclusivamente como um **eixo norteador para a análise descritiva**. O objetivo não é prever a renda de novos indivíduos, mas sim utilizar a arquitetura da árvore de decisão como um instrumento de engenharia reversa para mapear como as características demográficas, educacionais e de desempenho se estratificam entre as diferentes classes socioeconômicas.

Em consonância com essa finalidade, o conjunto de dados foi utilizado em sua totalidade para a indução dos modelos, abdicando-se da tradicional divisão entre conjuntos de treino e teste. Essa escolha metodológica justifica-se pelo fato de o estudo não buscar a validação de performance preditiva, mas sim a **exaustão descritiva dos padrões internos** da base de dados. Para viabilizar tal profundidade analítica, os modelos foram submetidos a um **sobreajuste integral (*overfitting*) intencional**. Ao permitir que as árvores cresçam sem as restrições usuais de poda (*pruning*) ou profundidade máxima, garante-se que o modelo capture as correlações mais minuciosas e as interações complexas entre as variáveis. Esta abordagem transforma a árvore de decisão em um mapa detalhado da população observada, priorizando a fidelidade aos dados presentes em detrimento da generalização estatística para dados não observados.

4.5. Técnicas de Explicabilidade (SHAP)

Como forma de complementar a interpretação das árvores de decisão, foram aplicadas técnicas de explicabilidade provenientes da área de *Explainable Artificial Intelligence* (XAI). Em especial, utilizou-se o método SHAP (*SHapley Additive exPlanations*) em um cenário multiclasse. A escolha deste método permite quantificar de forma equitativa a contribuição de cada variável para o resultado final do modelo.

Enquanto a árvore de decisão intencionalmente sobreajustada fornece um mapeamento exaustivo e complexo dos dados, o SHAP atua como uma ferramenta de síntese analítica, permitindo identificar a importância global e local das variáveis. Especificamente, o SHAP possibilitará:

- **Ranking de Relevância:** Identificar quais perguntas do questionário socioeconômico e quais áreas do conhecimento possuem maior peso na definição de cada faixa de renda familiar;
- **Direcionalidade do Impacto:** Analisar se a presença de um determinado fator (como acesso à internet ou escolaridade dos pais) aumenta ou diminui a probabilidade de o estudante pertencer a uma classe social específica;

- **Análise Interseccional Detalhada:** Comparar como o peso das variáveis se altera quando o modelo é aplicado aos diferentes recortes (sexo e localização da escola), revelando se os determinantes da renda se manifestam de forma distinta entre estudantes urbanos e rurais.

Assim, o SHAP transforma a saída técnica do modelo em uma explicação interpretável para o contexto educacional, permitindo que a análise descritiva identifique não apenas o que define o perfil do estudante, mas o quanto cada fator contribui para a desigualdade observada nos microdados do ENEM.

5. Resultados

Nesta seção são apresentados e analisados os principais resultados obtidos a partir da aplicação dos modelos de Árvore de Decisão aos dados do Exame Nacional do Ensino Médio (ENEM). Os resultados têm como objetivo evidenciar os fatores que mais influenciam o desempenho dos participantes, com ênfase na interpretabilidade dos modelos e na compreensão das relações entre as variáveis socioeconômicas e os resultados alcançados no exame.

Além das métricas de desempenho dos modelos, são discutidos os padrões identificados nas regras de decisão geradas, permitindo uma análise transparente e compreensível dos critérios utilizados pelo modelo para distinguir diferentes níveis de desempenho. Quando aplicável, testes de hipóteses estatísticos são empregados para verificar a significância das relações observadas, reforçando a validade dos achados e contribuindo para uma interpretação mais robusta dos resultados.

5.1. Visão Geral da Análise

A análise dos resultados foi estruturada de forma comparativa e estratificada, integrando os microdados do ENEM referentes aos anos de 2018, 2019 e 2023 em uma única perspectiva analítica. Em vez de uma abordagem cronológica isolada, optou-se por uma organização baseada em estratos populacionais, permitindo observar a evolução e a permanência de padrões socioeconômicos ao longo do tempo.

A apresentação dos achados segue três eixos principais: inicialmente, examina-se a **amostra consolidada**, que oferece uma visão global dos fatores de estratificação de renda para cada ano. Na sequência, os dados são segmentados entre o **estrato urbano** e o **estrato rural**. Essa estrutura facilita a identificação de disparidades regionais e infraestruturais, permitindo confrontar como as mesmas variáveis se comportam em diferentes contextos geográficos e temporais.

Dentro de cada um desses estratos, a investigação aprofunda-se por meio da análise de gênero, explorando as nuances entre participantes do sexo masculino e feminino. Essa abordagem progressiva permite que o estudo identifique desde tendências macroestruturais até exceções singulares de grupos específicos, mantendo o foco na interpretabilidade das árvores de decisão e na contribuição das variáveis reveladas pelo método SHAP, sem a fragmentação excessiva dos resultados por subseções anuais.

5.2. Análise da amostra consolidada

A análise conjunta das estatísticas descritivas para a amostra consolidada nos anos de 2018, 2019 e 2023 revela um padrão de desigualdade estrutural altamente resiliente. O

fenômeno mais evidente, transversal a todo o período estudado, é a hierarquia rígida entre capital econômico e desempenho acadêmico: cada incremento na faixa de renda familiar traduz-se num ganho direto e positivo na média e na mediana das notas.

No que concerne às semelhanças estatísticas, três pontos destacam-se como constantes históricas:

- **Deslocamento Sistêmico da Distribuição:** Observa-se que o aumento da renda não beneficia apenas os candidatos de elite, mas empurra toda a distribuição para cima. O movimento concomitante dos quartis Q1 e Q3 indica que o suporte socioeconômico eleva o patamar de desempenho de todo o grupo.
- **Paradoxo da Dispersão:** Em todos os anos, a heterogeneidade dos resultados (medida pelo desvio padrão) cresce proporcionalmente à renda. Isto sugere que, embora os estratos mais baixos apresentem notas menores, estas são mais concentradas e previsíveis, enquanto os estratos mais abastados, apesar das médias elevadas, exibem uma volatilidade de desempenho significativamente superior.
- **Robustez e Precisão:** A consistência dos dados é ratificada pelos intervalos de confiança (IC 95%) extremamente estreitos, fruto do elevado volume amostral, especialmente nas faixas de renda mais baixas, o que confere segurança estatística às disparidades apontadas.

As divergências entre os anos revelam nuances na dinâmica da desigualdade. Enquanto os períodos de 2018 e 2019 apresentam flutuações ligeiras nos patamares de média do topo (com uma breve retração em 2019), os dados de 2023 introduzem uma complexidade maior através do recorte de gênero. Nota-se que o gênero atua como um modulador da vantagem econômica: no cenário mais recente, as mulheres demonstram uma capacidade de conversão de recursos em desempenho superior à dos homens nos estratos de elite, alcançando as maiores medianas e o terceiro quartil (Q3) mais elevado de toda a amostra. Um ponto crítico revelado pela evolução temporal e análise de quartis é a existência de um abismo de mobilidade. Em 2023, por exemplo, o patamar inferior de desempenho (Q1) das candidatas no topo da pirâmide já supera a média global dos estudantes da base. Este dado reforça que a desigualdade acumulada criou barreiras onde mesmo o rendimento menos expressivo da elite econômica permanece acima do sucesso médio dos estratos populares, consolidando o que a literatura descreve como estabilidade na exclusão. Esses achados são evidenciados nas Tabelas 3 e 4.

Tabela 3. Estatísticas descritivas - Estrato Feminino (2023)

Faixa de renda	Mediana	Q1	Q3	Média	Desvio padrão	Tamanho (N)	IC 95%
Até 2 salários mínimos	537,78	480,48	537,78	517,47	69,07	167.480	[517,37 ; 517,58]
De 2 a 3 salários mínimos	537,78	515,94	590,10	549,34	73,09	17.175	[549,00 ; 549,69]
De 3 a 6 salários mínimos	564,24	537,78	631,78	576,30	79,79	4.050	[576,05 ; 576,55]
De 6 a 10 salários mínimos	611,74	537,78	670,93	609,56	82,54	928	[609,03 ; 610,10]
De 10 a 15 salários mínimos	629,34	553,56	686,18	623,48	82,70	415	[622,69 ; 624,28]
Acima de 15 salários mínimos	635,26	556,16	690,96	627,40	83,83	251	[626,37 ; 628,44]

Fonte: Autoria própria a partir dos microdados do ENEM 2023.

Tabela 4. Estatísticas descritivas - Estrato Masculino (2023)

Faixa de renda	Mediana	Q1	Q3	Média	Desvio padrão	Tamanho (N)	IC 95%
Até 2 salários mínimos	537,78	483,62	540,24	519,54	74,74	92.682	[519,38 ; 519,69]
De 2 a 3 salários mínimos	537,78	514,76	584,46	544,95	76,18	12.223	[544,53 ; 545,38]
De 3 a 6 salários mínimos	548,16	535,66	621,22	567,43	82,27	3.280	[567,15 ; 567,71]
De 6 a 10 salários mínimos	597,60	537,78	664,52	599,72	87,53	815	[599,12 ; 600,32]
De 10 a 15 salários mínimos	619,30	537,78	682,54	616,07	88,50	378	[615,18 ; 616,97]
Acima de 15 salários mínimos	627,88	537,78	692,62	622,17	91,54	263	[621,07 ; 623,28]

Fonte: Autoria própria a partir dos microdados do ENEM 2023.

A visualização da distribuição das notas através de diagramas de caixa (*boxplots*) nos anos de 2018, 2019 e 2023 ratifica os achados das estatísticas descritivas, a renda como um filtro de estabilidade de desempenho. Observa-se que o aumento do nível socioeconômico não apenas eleva as medianas, mas altera a própria arquitetura da dispersão dos dados o que pode ser observado na Figura 1

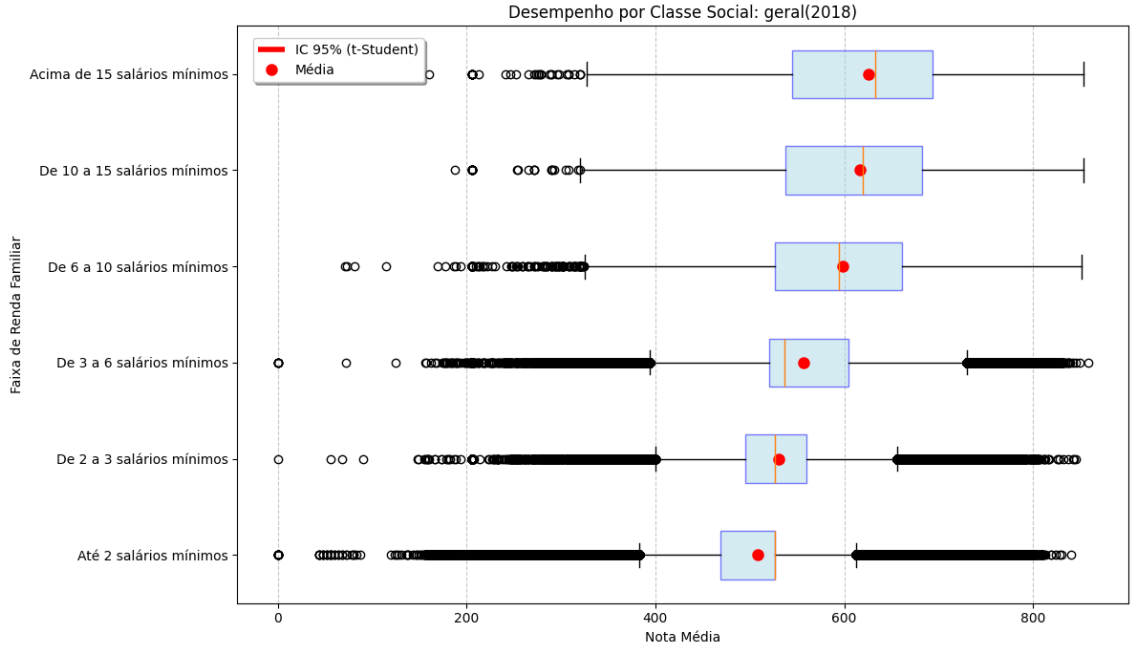


Figura 1. Distribuição das médias das notas por faixa de renda - Amostra consolidada (2018)

Fonte: Autoria própria.

No que tange às semelhanças fundamentais entre os períodos, destacam-se:

- **A Riqueza como Estabilizador:** Em todos os anos, as faixas de alta renda apresentam caixas mais compactas e situadas em patamares elevados. Este padrão sugere que o capital econômico mitiga dispersões extremas, garantindo uma proficiência uniforme no topo da pirâmide.
- **Heterogeneidade na Base:** Inversamente, os estratos de menor renda exibem uma massa densa de *outliers* inferiores e maior amplitude interquartílica relativa. Isso indica que a vulnerabilidade social amplia a incerteza do desempenho, embora a presença

de *outliers* superiores nesses grupos confirme a existência de trajetórias de superação excepcionais.
- **Segregação de Performance:** Um fenômeno crítico observado é que o terceiro quartil (Q3) das faixas de baixa renda raramente alcança o primeiro quartil (Q1) das faixas de elite. Esse isolamento de performance demonstra que mesmo os melhores estudantes da base dificilmente atingem o patamar dos estudantes menos produtivos do topo.

Quanto às diferenças e evoluções, o fator gênero surge como um diferencial de resiliência. Em 2018 e 2019, o impacto da renda mostrou-se transversal, com os homens apresentando uma dispersão ligeiramente superior de resultados excepcionais na base. Contudo, em 2023, nota-se uma consolidação da vantagem feminina nos extremos. Enquanto o grupo masculino apresenta maior volatilidade e uma densidade superior de notas próximas a zero nos estratos vulneráveis, as mulheres demonstram maior homogeneidade e consistência. No topo da pirâmide (Acima de 15 salários mínimos) em 2023, a vantagem feminina torna-se visualmente clara: as medianas e os limites superiores das hastes (*whiskers*) superam os valores masculinos. Em suma, os dados sugerem que, ao longo do tempo, as candidatas têm conseguido converter o acesso a recursos econômicos de elite em patamares de excelência de forma mais estável e resiliente do que seus pares masculinos sob as mesmas condições.

A extração da importância das variáveis (*Feature Importance*) a partir dos modelos de árvores de decisão permite identificar os pilares que sustentam a estratificação de renda no ENEM. Ao confrontar os anos de 2018, 2019 e 2023, observa-se uma estabilidade notável na hierarquia dos preditores, consolidando um binômio entre patrimônio material e desempenho acadêmico.

No que concerne às semelhanças estruturais, destacam-se:

- **A Primazia da Mobilidade:** A variável **Q010** (posse de automóvel) isola-se em todos os anos como o discriminador mais robusto, com pesos que chegam a superar $0,18$. Este achado sugere que o automóvel não é apenas um bem de consumo, mas o filtro primário de distinção de classe no cenário nacional, superando sistematicamente o peso individual de qualquer nota acadêmica.
- **Matemática como Validador de Renda:** Entre as áreas do conhecimento, a nota de Matemática (**NU_NOTA_MT**) consolidou-se como o segundo ou terceiro fator mais influente no ranking global. A proficiência em exatas atua como o sinalizador intelectual mais sensível para a classificação socioeconômica, especialmente em 2023.
- **Capital Cultural e Infraestrutura:** Variáveis como a escolaridade dos pais (**Q001** e **Q002**) e a infraestrutura doméstica (como a quantidade de banheiros, **Q008**) figuram consistentemente no topo, indicando que o modelo associa a renda a um equilíbrio entre bens duráveis e capital instrucional familiar.

As divergências revelam nuances importantes sobre como o gênero modula a percepção de renda pelo modelo. Historicamente, em 2018 e 2019, observou-se uma especialização por área: enquanto a renda masculina era melhor identificada pela proficiência em Matemática, a renda feminina apresentava uma inversão singular, tendo a nota de Linguagens (**NU_NOTA_LC**) como principal indicador acadêmico. Esta nuance sugere que, para as candidatas, o capital cultural manifestado na fluência verbal e na proficiência linguística foi um preditor de classe mais refinado. Visualmente podemos

verificar essas informações através das Figuras 2 e 3

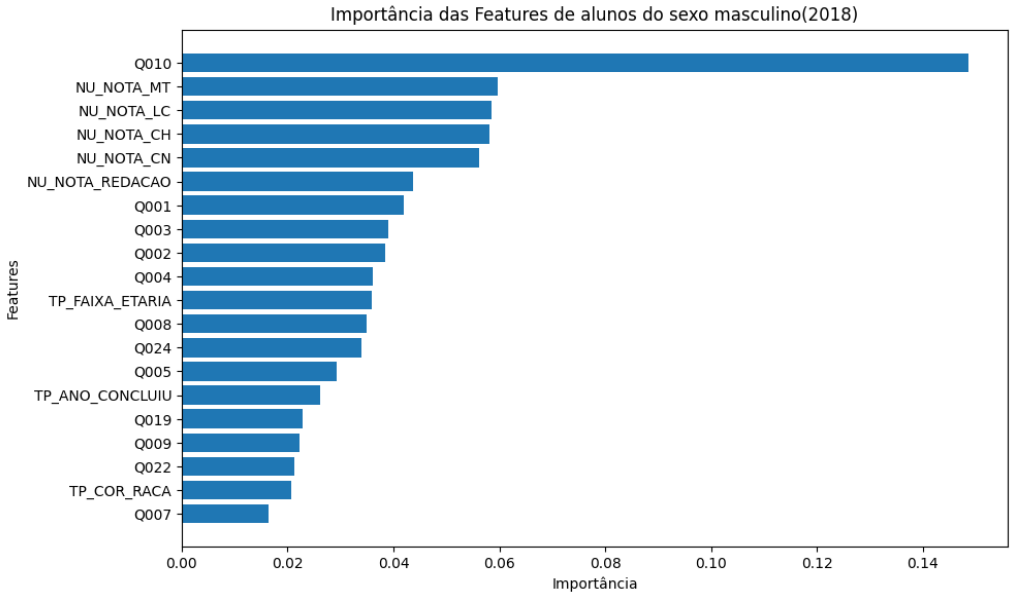


Figura 2. Importância agregada das variáveis da árvore de decisão - Estrato masculino (2018)

Fonte: Autoria própria.

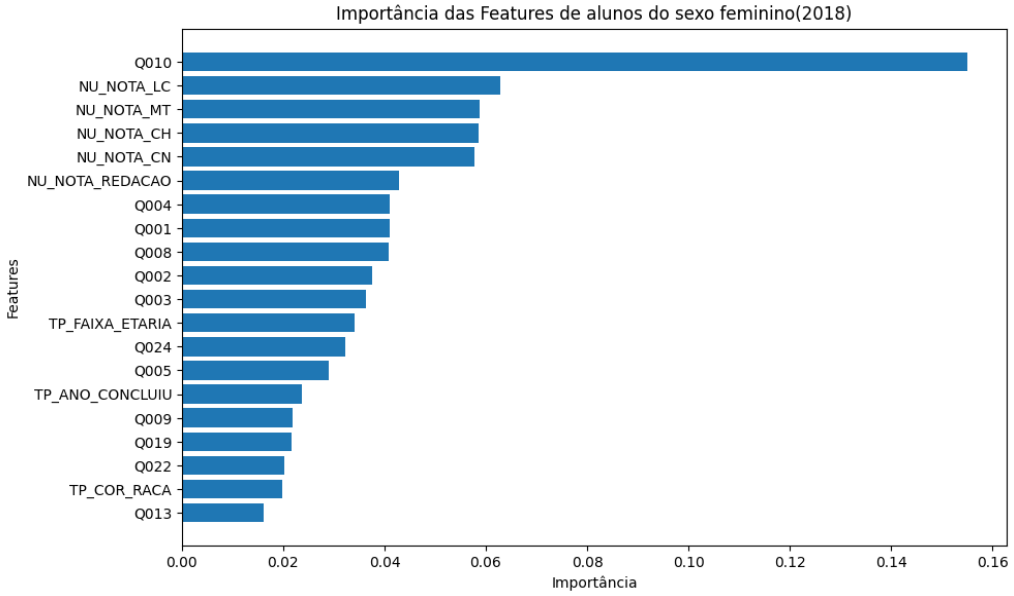


Figura 3. Importância agregada das variáveis da árvore de decisão - Estrato feminino (2018)

Fonte: Autoria própria.

Contudo, no cenário de 2023, essa distinção de gênero nas notas tendeu a uma convergência em direção à Matemática, embora as mulheres tenham demonstrado um peso

superior em indicadores de infraestrutura doméstica (**Q008**) comparado ao acesso tecnológico (**Q024**). Por fim, é relevante notar que fatores puramente demográficos, como a cor ou raça (**TP_COR_RACA**), e a nota da Redação apresentam sistematicamente menor poder preditivo na classificação de renda do que as variáveis patrimoniais e as notas de provas objetivas, reforçando a natureza material da estratificação revelada pelas árvores de decisão. Visualmente pode ser verificado através da Figura 4

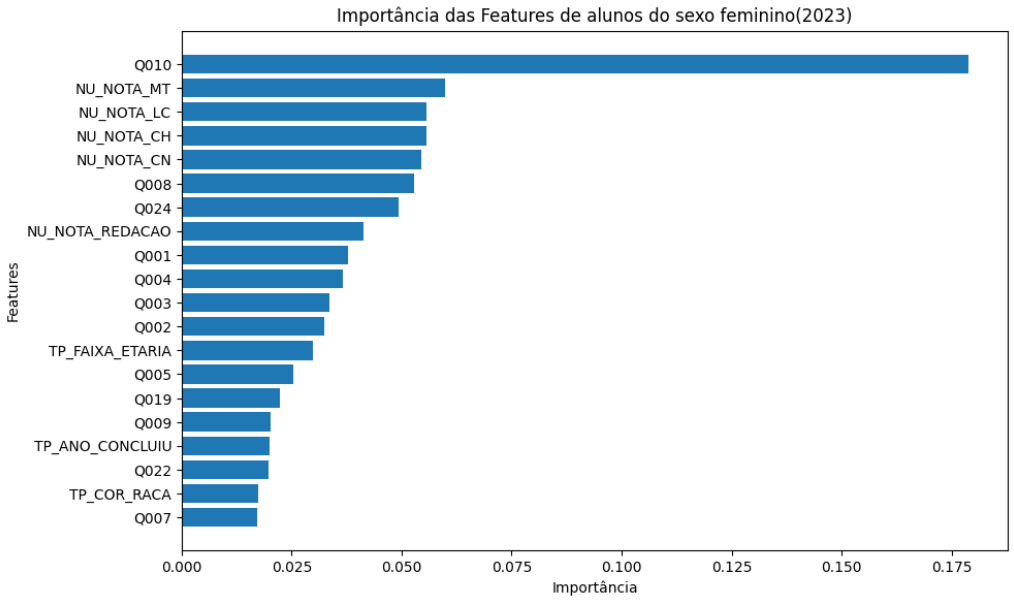


Figura 4. Importância agregada das variáveis - Estrato Feminino(2023)

Fonte: Autoria própria.

A decomposição da importância das variáveis em *dummies* permite identificar os estados específicos que governam as decisões dos modelos. Ao confrontar os anos de 2018, 2019 e 2023, observa-se uma consistência notável no papel da privação material como o principal marcador de estratificação socioeconômica no Brasil.

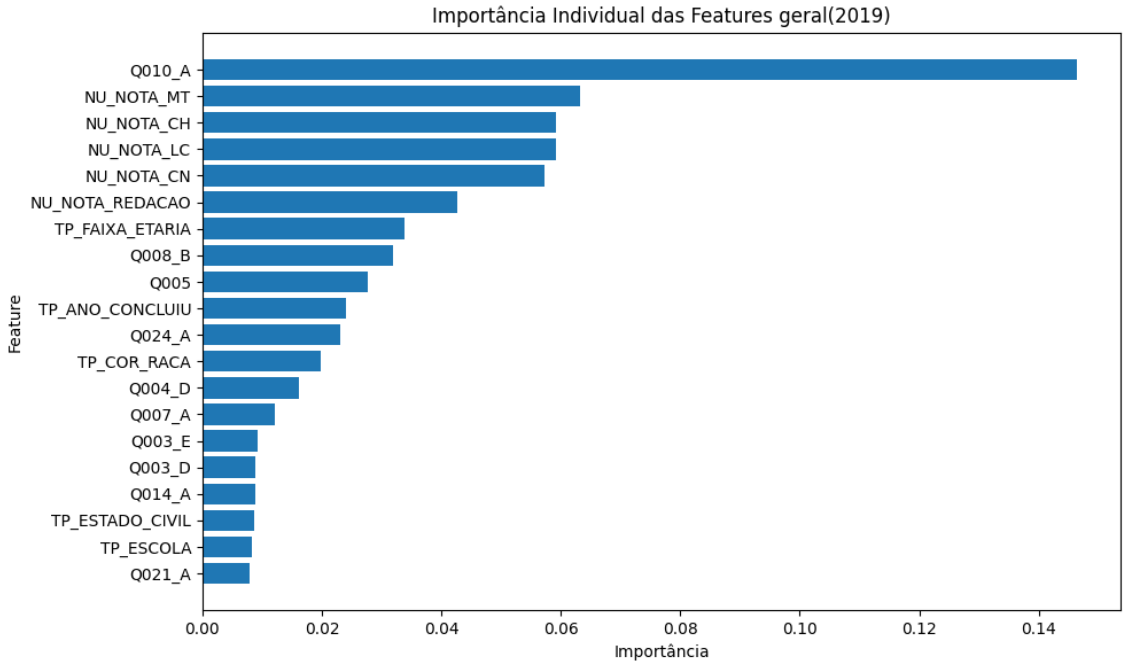
- **A Marca da Vulnerabilidade (Q010_A):** Em todos os anos e estratos analisados, a ausência de automóvel (**Q010_A**) consolidou-se como o preditor individual mais potente, com pesos variando entre $0,14$ e $0,17$. Este dado sugere que a restrição de mobilidade privada precede o desempenho acadêmico como o indicador mais preciso para classificar estratos de baixa renda, funcionando como um divisor de águas estatístico universal.
- **Matemática como Termômetro de Capital Escolar:** No bloco educacional, a nota de Matemática (**NU_NOTA_MT**) manteve-se estavelmente como o segundo fator individual mais influente. A persistência deste padrão ao longo dos anos indica que a proficiência em exatas é o sinalizador intelectual mais sensível à renda, independentemente das flutuações nas políticas educacionais do período.
- **Evolução das Nuances de Gênero:** Observa-se uma trajetória de mudança na percepção do capital cultural. Em 2018 e 2019, o modelo identificava o sucesso em áreas distintas como marcadores de renda específicos: Linguagens (**NU_NOTA_LC**) para as mulheres e Matemática para os homens. Contudo, em 2023, nota-se uma con-

vergência acadêmica, com a Matemática ganhando protagonismo em ambos os sexos, embora o estrato feminino tenha passado a exibir a infraestrutura doméstica (como a ausência de máquina de lavar, **Q013_A** ou **Q008_B**) como um marcador de classe mais nítido do que o acesso tecnológico (**Q024_A**).

- **Trajetória vs. Demografia:** Um ponto de convergência importante é a relevância da faixa etária (**TP_FAIXA_ETARIA**) e do tempo de conclusão (**TP_ANO_CONCLUIU**), especialmente para o grupo feminino. Isso sugere que o hiato temporal nos estudos é um marcador de vulnerabilidade econômica mais forte para as mulheres. Por outro lado, variáveis como cor/raça (**TP_COR_RACA**) ocupam consistentemente a base do ranking, reafirmando que a classificação de renda é governada primordialmente pelo binômio carência material e desempenho escolar.

Em suma, a evolução dos dados entre 2018 e 2023 demonstra que, enquanto o cenário educacional brasileiro passou por transformações, os marcadores de classe permaneceram centrados na posse de bens de consumo duráveis e na capacidade de conversão de renda em notas de exatas. O modelo de inteligência artificial revela, assim, uma estrutura de desigualdade onde a falta (de carro, de computador ou de pontos em matemática) é o sinalizador mais preciso da posição social do participante. A Figura 5 apresenta as features mais importantes para o modelo do ano de 2019.

Figura 5. Importância individual das variáveis da árvore de decisão - Amostra consolidada(2019)



Fonte: Autoria própria.

A aplicação do método SHAP (*SHapley Additive exPlanations*) permite decompor a contribuição de cada variável na classificação das faixas de renda, revelando como o modelo interpreta a transição entre a vulnerabilidade e a elite socioeconômica. A comparação entre 2018, 2019 e 2023 evidencia uma pirâmide de acessos estável, onde a base é defi-

nida pela privação material e o topo pela conversão de capital cultural em desempenho que pode ser verificada na Figura 6

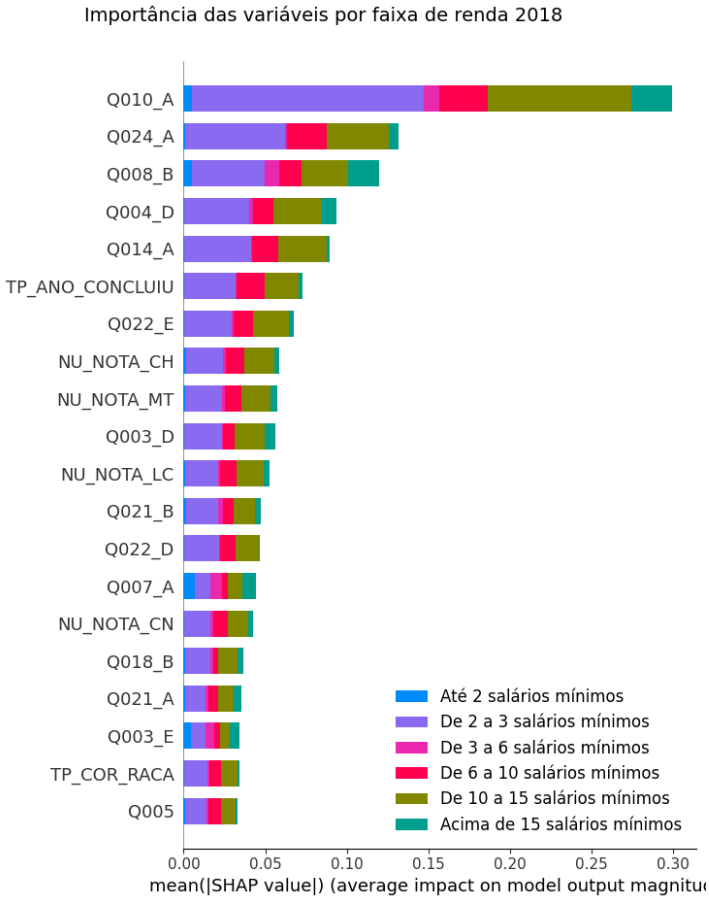


Figura 6. Importância das variáveis SHAP - Amostra consolidada (2018)

Fonte: Autoria própria.

No que tange às semelhanças e constantes históricas, destacam-se:

- **A Magnitude do Impacto Material (Q010_A):** Em todos os anos, a ausência de automóvel (**Q010_A**) manteve-se como o fator de maior impacto global, com valores SHAP situados na casa de $0,30$. Este indicador atua como o divisor crítico que isola as classes de menor renda das elites, demonstrando que a mobilidade privada é o sinalizador mais nítido da hierarquia social brasileira.
- **Triade de Privação:** A exclusão digital (**Q024_A**) e as limitações de infraestrutura básica (**Q008_B**) figuram sistematicamente como os preditores secundários que consolidam os perfis de vulnerabilidade (até 3 salários mínimos), funcionando como barreiras físicas ao desempenho.
- **Matemática como Marcador de Privilégio:** A nota de Matemática (**NU_NOTA_MT**) consolidou-se como o principal indicador acadêmico, com impacto concentrado quase exclusivamente nas faixas de renda média-alta e alta (acima de 10 salários mínimos). Isso sugere que a proficiência em exatas é o ativo intelectual que melhor sinaliza o acesso a recursos econômicos de elite.

As divergências e a evolução dos dados revelam mudanças significativas nas nuan-

ces de gênero e capital social. Nos anos de 2018 e 2019, observou-se uma especialização acadêmica: a renda masculina era sinalizada pela Matemática, enquanto a feminina era definida pela proficiência em Linguagens (**NU_NOTA_LC**), indicando que o capital cultural manifestado na fluência verbal era um marcador de classe mais sensível para as mulheres o que pode ser visualizado através da Figura 7

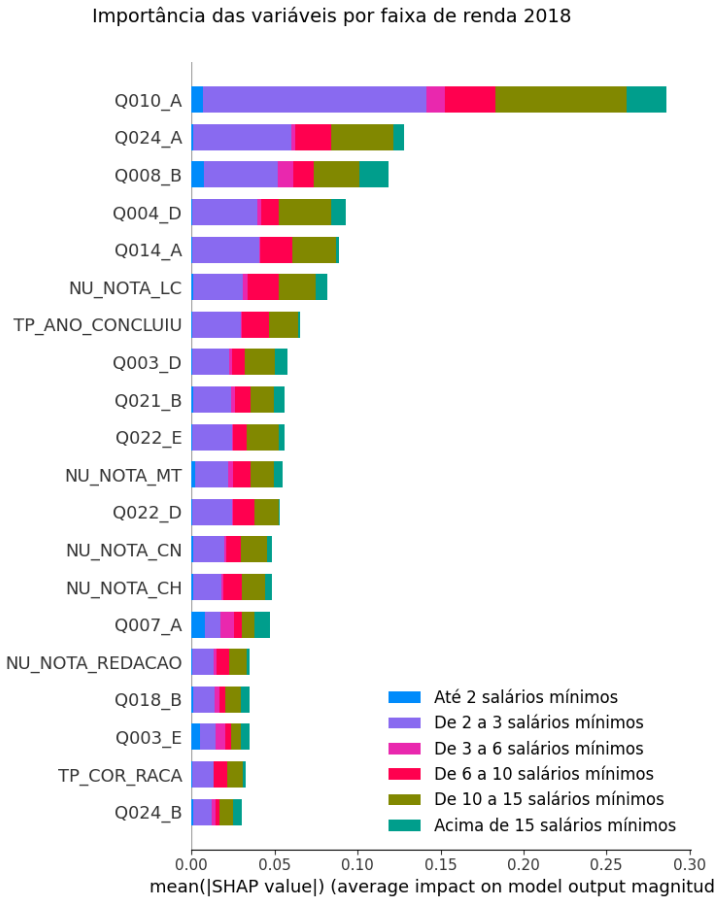


Figura 7. Importância das variáveis SHAP - Estrato feminino (2018)

Fonte: Autoria própria.

Contudo, no cenário de 2023, nota-se uma convergência acadêmica e a ascensão de fatores de herança social. A ocupação qualificada do pai (**Q004_D**) emergiu como um lastro de classe fundamental no topo do ranking, precedendo até mesmo o desempenho escolar em algumas competências. Adicionalmente, o hiato temporal nos estudos (**TP_ANO_CONCLUIU**) e a faixa etária deixaram de ser marcadores genéricos para se tornarem indicadores de vulnerabilidade específicos, refletindo como a trajetória escolar irregular está intrinsecamente ligada à restrição econômica. Em suma, os gráficos SHAP confirmam que, enquanto a carência patrimonial define a base da pirâmide, a estabilidade ocupacional familiar e a excelência em exatas são os passaportes estatísticos para o topo.

5.3. Análise do estrato urbano

A análise do estrato urbano nos anos de 2018, 2019 e 2023 ratifica a cidade como o epicentro da estratificação educacional brasileira. Observa-se que a densidade de recursos

urbanos não atenua o abismo socioeconômico; pelo contrário, a amplitude entre as faixas de rendas mais baixas e mais altas permanece superior a 100 pontos em todos os períodos analisados como pode ser verificado na Tabela 5.

Tabela 5. Estatísticas descritivas das médias das notas por faixa de renda - Estrato Urbano (2018)

Faixa de renda	Mediana	Q1	Q3	Média	DP	N	IC 95%
Até 2 salários mínimos	526.46	469.12	526.46	508.25	59.57	3 505 804	[508.19 ; 508.31]
De 2 a 3 salários mínimos	526.46	495.84	559.68	530.22	63.77	506 886	[530.05 ; 530.40]
De 3 a 6 salários mínimos	537.12	520.28	604.28	557.31	74.77	1 082 282	[557.17 ; 557.45]
De 6 a 10 salários mínimos	594.86	526.46	661.24	598.30	82.43	223 969	[597.96 ; 598.64]
De 10 a 15 salários mínimos	620.22	537.46	682.12	616.73	83.30	89 700	[616.18 ; 617.27]
Acima de 15 salários mínimos	632.64	545.04	693.60	626.11	85.13	55 701	[625.40 ; 626.81]

Fonte: Autoria própria a partir dos microdados do ENEM 2018.

No que tange às constantes e semelhanças, destacam-se:

- **Hierarquia Rígida e Isolamento de Elite:** Em 2023, o abismo tornou-se visualmente mensurável: a mediana do estrato superior supera o terceiro quartil (Q3) das três faixas de renda iniciais. Isso indica que a maioria dos candidatos de elite urbana supera 75% dos estudantes das classes populares, um padrão de segregação que se manteve estável desde 2018.
- **Crescimento da Heterogeneidade:** Em todos os anos, o desvio padrão cresce proporcionalmente à renda. Embora as médias urbanas sejam as mais altas do país, os resultados no topo são mais voláteis, sugerindo que o ambiente urbano oferece múltiplas trajetórias de sucesso, mas também maiores disparidades internas entre os mais abastados.
- **Estabilidade na Base:** Em 2019, observou-se que ganhos marginais de renda nas faixas iniciais não deslocavam o centro da distribuição, sugerindo que, no contexto citadino, apenas mudanças estruturais profundas de renda conseguem alterar o patamar de desempenho.

Quanto às diferenças e evoluções das nuances de gênero, o cenário urbano revela uma trajetória de domínio feminino:

- **Liderança Feminina de Topo:** Em 2018 e 2023, as mulheres urbanas de alta renda alcançaram as maiores médias registradas no estudo (chegando a $627,46$ pontos em 2023), superando sistematicamente o grupo masculino e a média consolidada.
- **Eficiência e Consistência:** O diferencial feminino urbano reside na consistência acadêmica. Enquanto os homens urbanos apresentam a maior volatilidade e desvios padrão mais elevados (atingindo $91,51$ em 2023), as mulheres exibem resultados mais homogêneos.
- **A Vantagem Marginal Masculina:** Nota-se um padrão recorrente onde os homens apresentam médias ligeiramente superiores apenas na base da pirâmide urbana, vantagem que é rapidamente dissipada e revertida a favor das mulheres conforme a renda familiar aumenta.

Tais evidências são corroboradas pela distribuição das notas por faixa de renda em 2023, conforme ilustrado nos diagramas de caixa das Figuras 8 e 9.

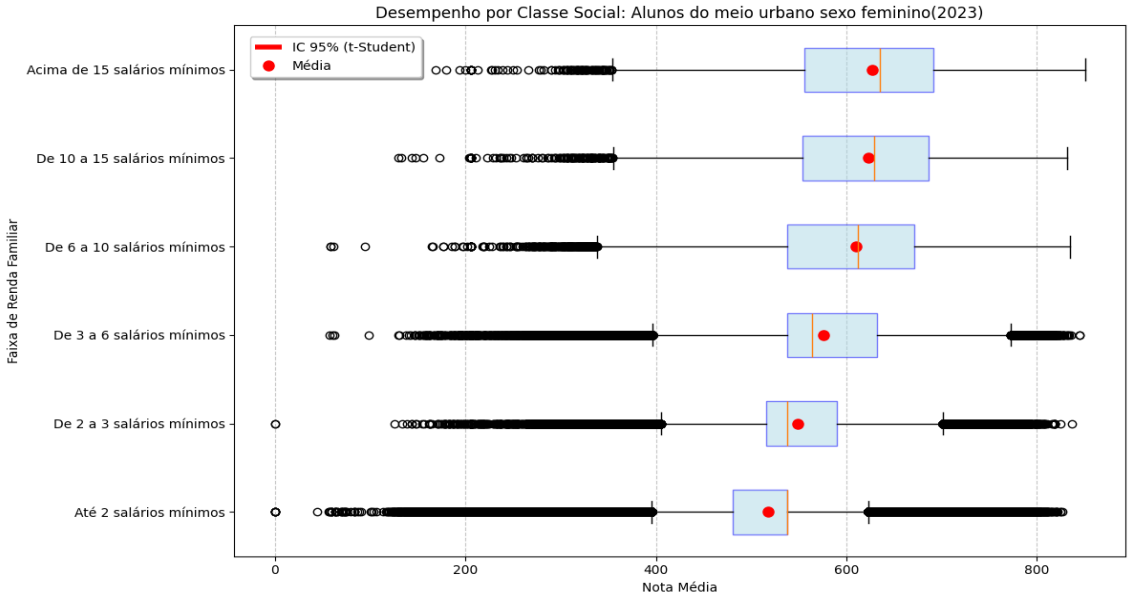


Figura 8. Distribuição das médias das notas por faixa de renda - Estrato urbano feminino

Fonte: Autoria própria.

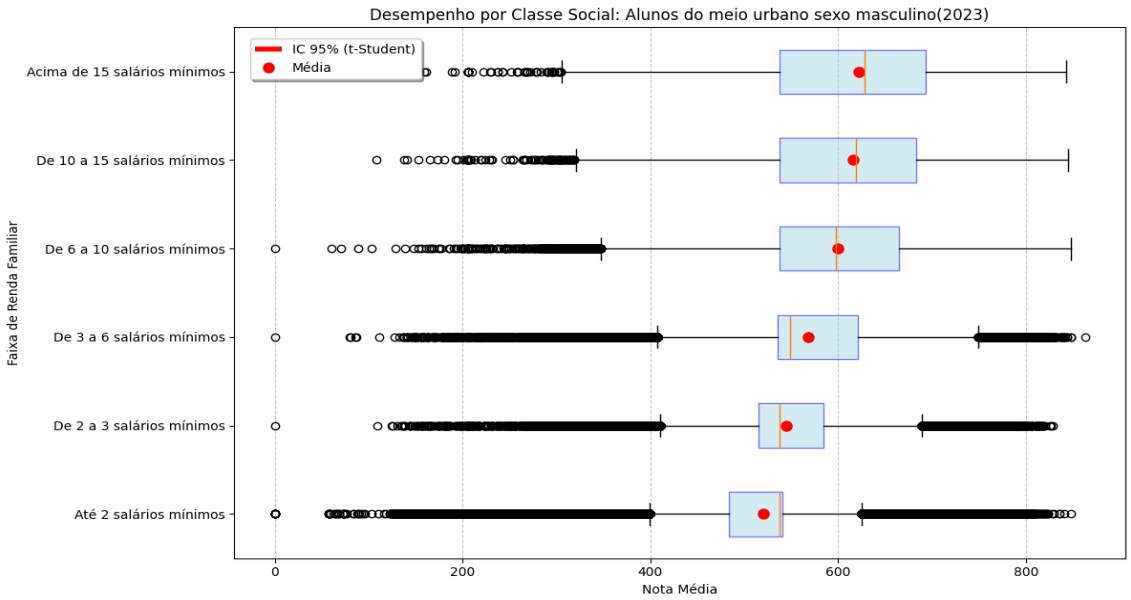


Figura 9. Distribuição das médias das notas por faixa de renda - Estrato urbano masculino

Fonte: Autoria própria.

Em suma, o cenário urbano brasileiro demonstra que o capital econômico e cultural traduz-se em ganhos de desempenho mais nítidos e estáveis para o público feminino. Enquanto os candidatos masculinos enfrentam uma heterogeneidade de desempenho mais acentuada nas cidades, as candidatas consolidam-se como o motor de excelência e a verdadeira elite acadêmica no topo da pirâmide social urbana.

A hierarquia de importância das variáveis no estrato urbano revela que a classificação de renda nas cidades está ancorada em um tripé fundamental: patrimônio acumulado, infraestrutura doméstica e capital cultural. A evolução dos modelos entre 2018 e 2023 demonstra uma estabilidade estrutural onde os bens de consumo duráveis sobrepõem-se sistematicamente aos indicadores acadêmicos subjetivos como apresenta a Figura 10.

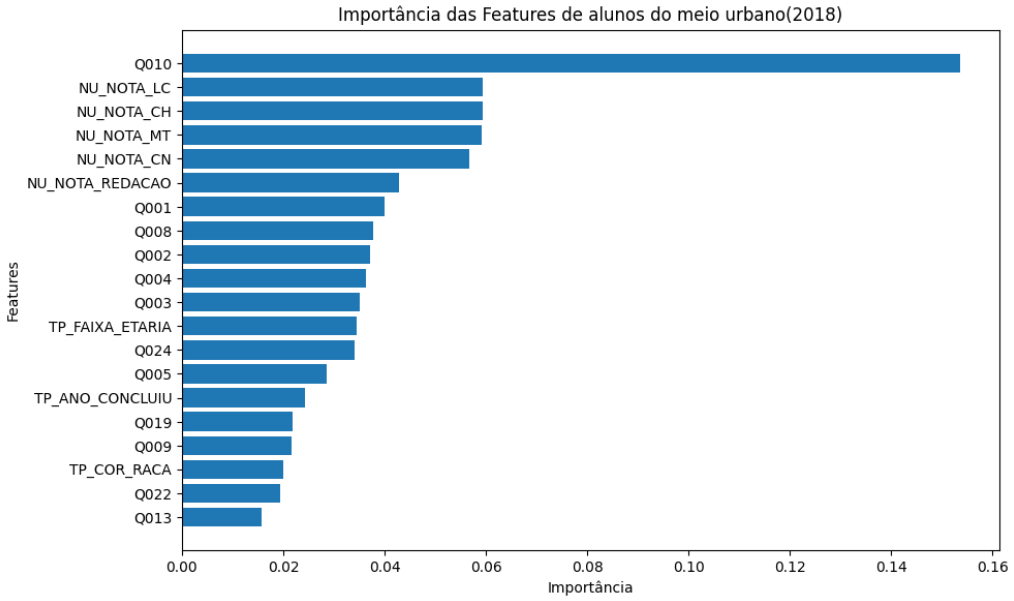


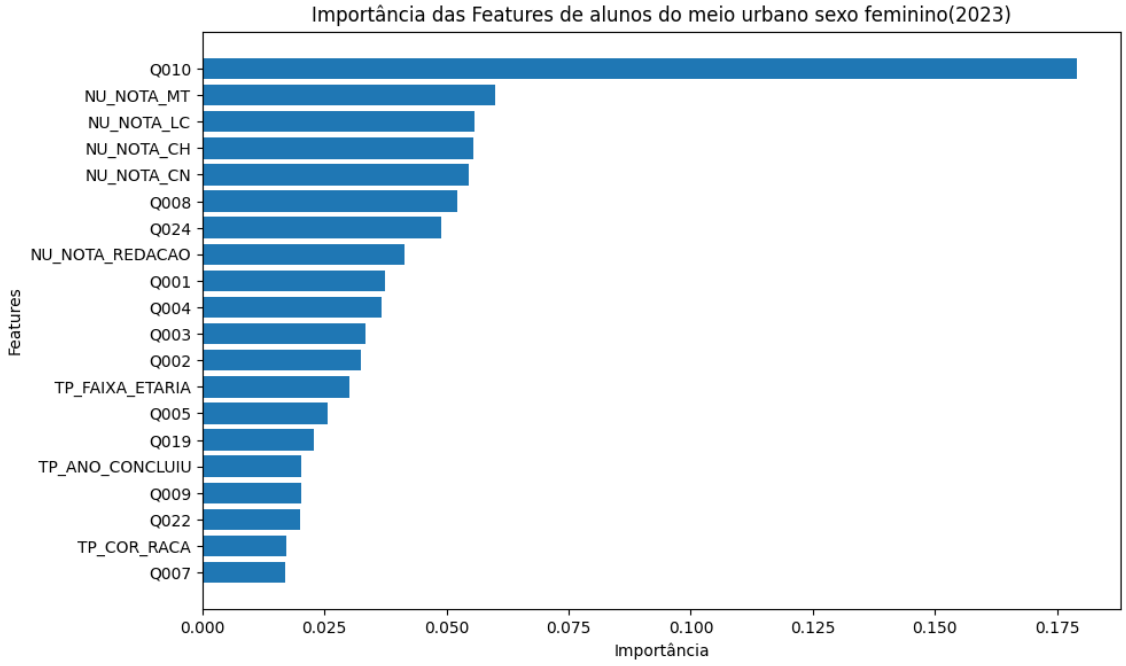
Figura 10. Importância agregada das variáveis - Estrato urbano consolidado (2018)

- **O Automóvel como Divisor Urbano (Q010):** Em todos os anos analisados, a posse de automóvel (**Q010**) isolou-se como o preditor mais robusto, apresentando um crescimento na sua importância relativa, que saltou de aproximadamente $0,15$ em 2018 para cerca de $0,18$ em 2023. Este achado consolida a mobilidade privada como o filtro socioeconômico primário no contexto citadino, sendo o marcador mais nítido de distinção entre as classes.
- **Evolução do Bloco Acadêmico:** Observa-se uma alternância na importância das áreas do conhecimento. Enquanto em 2018 as notas objetivas apresentavam pesos equilibrados, a partir de 2019 a nota de Matemática (**NU_NOTA_MT**) consolidou-se como o principal sinalizador educacional de renda no ambiente urbano, sugerindo que o desempenho técnico em exatas tornou-se um validador de classe mais sensível para o algoritmo.
- **Infraestrutura vs. Escrita:** Um padrão persistente é a precedência de variáveis de infraestrutura, como a quantidade de banheiros (**Q008**) e a posse de computador (**Q024**), sobre a nota da Redação. Isso indica que, para o modelo, a carência de bens duráveis

sinaliza a vulnerabilidade urbana de forma mais imediata e precisa do que competências subjetivas de escrita.

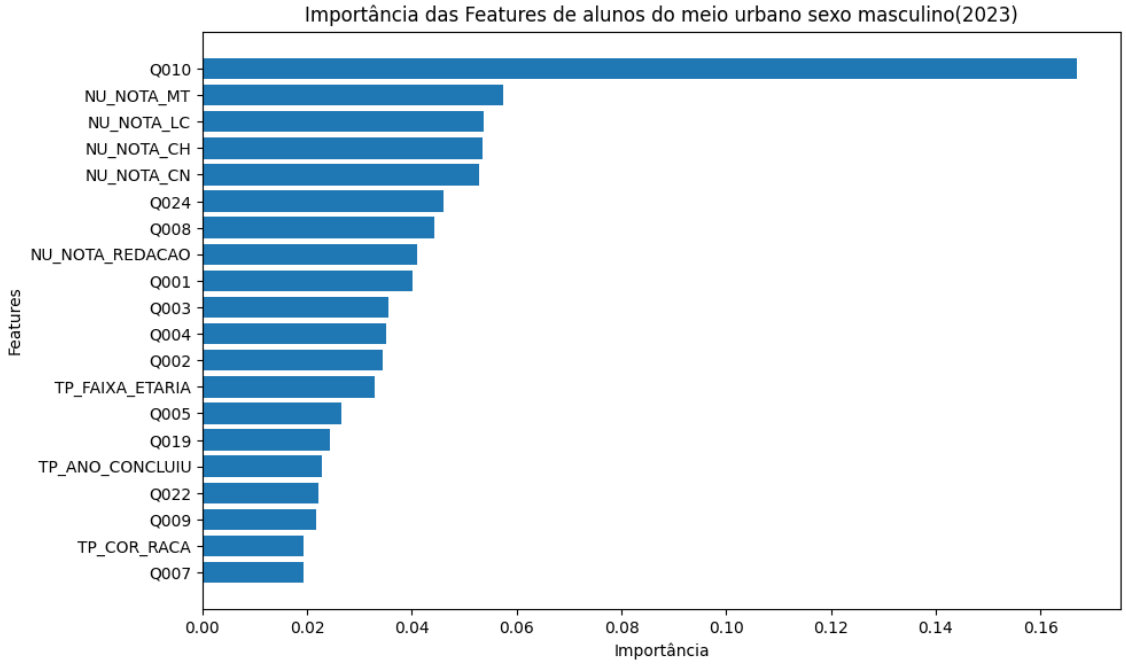
As nuances de gênero no ambiente urbano revelam trajetórias de sinalização de renda que evoluíram para uma convergência acadêmica em 2023. Historicamente (2018-2019), o modelo identificava o capital cultural de forma especializada: a renda feminina era fortemente associada à proficiência em Linguagens (**NU_NOTA_LC**), enquanto a masculina era atrelada à Matemática. No entanto, os dados de 2023 revelam uma mudança significativa: para as mulheres urbanas, a nota de Matemática (**NU_NOTA_MT**) ganhou protagonismo, alinhando-se ao padrão masculino. Esta convergência sugere que, no cenário contemporâneo, a proficiência em exatas tornou-se o validador universal de capital econômico para ambos os gêneros no contexto urbano, embora as candidatas ainda apresentem uma dependência superior de indicadores de infraestrutura doméstica (**Q008**) e mobilidade (**Q010**) para a definição de seus estratos sociais pelo modelo. Essa hierarquia de relevância é corroborada pela importância agregada das variáveis para o estrato urbano, conforme ilustrado nas Figuras 11 e 12.

Figura 11. Importância agregada das variáveis da árvore de decisão - Estrato Urbano feminino (2023)



Fonte: Autoria própria.

Figura 12. Importância agregada das variáveis da árvore de decisão - Estrato Urbano masculino (2023)



Fonte: Autoria própria.

Em suma, a análise comparativa urbana ratifica que, independentemente do ano, a classificação de renda é governada pelo binômio patrimônio-desempenho. A baixa relevância de fatores demográficos isolados, como cor/raça (**TP_COR_RACA**), reforça que a desigualdade urbana é decodificada pela inteligência artificial primordialmente através do acesso a bens que reduzem o esforço doméstico ou facilitam a mobilidade e a conectividade tecnológica.

A decomposição da importância das variáveis em categorias de resposta específicas permite identificar os marcadores de vulnerabilidade que governam a classificação socioeconômica nas cidades. A análise comparativa entre 2018 e 2023 revela uma estrutura de desigualdade urbana que se tornou mais materialista, onde a privação de bens duráveis sobrepõe-se à trajetória cronológica do estudante.

No que tange às constantes e à evolução dos indicadores, destacam-se:

- **A Hegemonia da Imobilidade (Q010_A):** Em todos os anos, a ausência de automóvel (**Q010_A**) consolidou-se como o divisor estatístico mais decisivo, apresentando uma trajetória ascendente de impacto (de $0,14$ em 2018 para mais de $0,16$ em 2023). Este dado ratifica que, no meio urbano, a restrição de mobilidade privada é o traço que melhor distingue a base da pirâmide social. Essa evidência estatística é corroborada pela distribuição da importância individual das variáveis, conforme ilustrado na Figura 13.

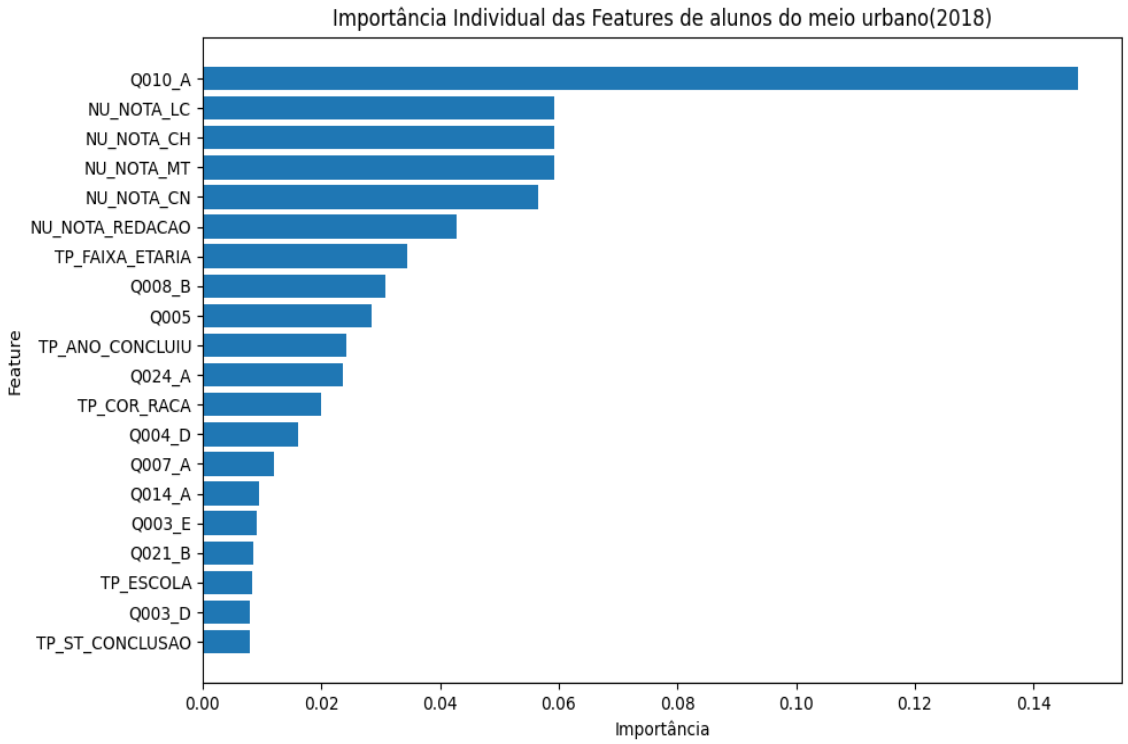


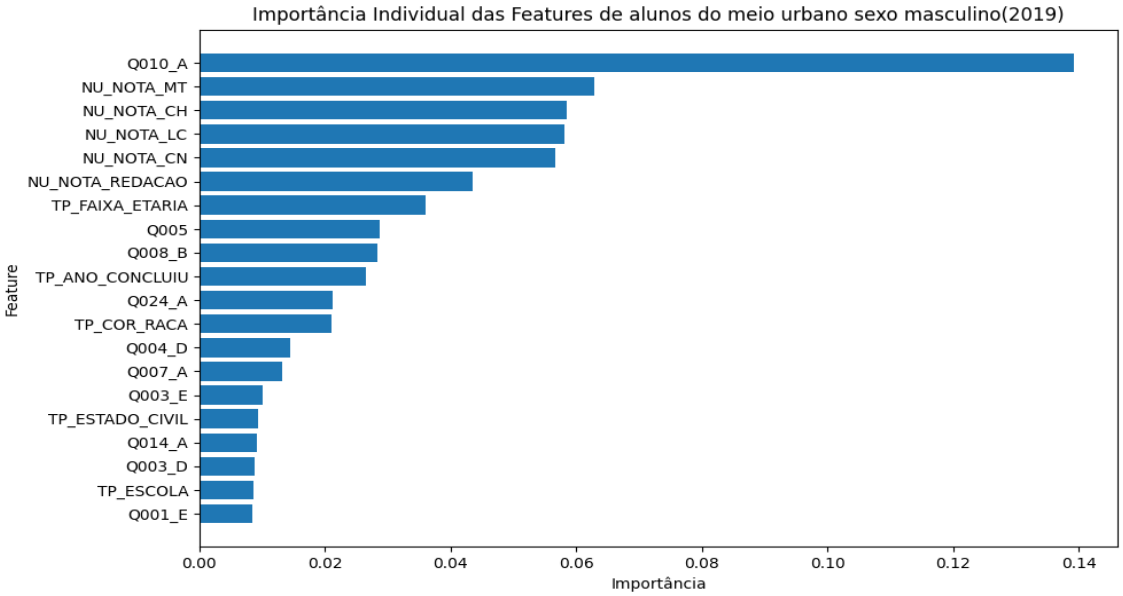
Figura 13. Importância individual das variáveis - Estrato urbano consolidado (2018)

Fonte: Autoria própria.

- **Evolução do Bloco Acadêmico:** Em 2018, as competências interpretativas (Linguagens e Humanas) eram marcadores de classe mais sensíveis no meio urbano. Contudo, a partir de 2019, a nota de Matemática (**NU_NOTA_MT**) assumiu a liderança acadêmica, sugerindo que o domínio das exatas tornou-se o sinalizador intelectual mais robusto de renda elevada nas cidades.
- **Infraestrutura e Exclusão Digital:** Um achado consistente em 2023 é que a presença de apenas um banheiro na residência (**Q008_B**) e a ausência de computador no domicílio (**Q024_A**) superam sistematicamente o poder preditivo da nota de Redação. Isso indica que o modelo identifica a vulnerabilidade urbana prioritariamente por indicadores de infraestrutura habitacional e de exclusão tecnológica, que funcionam como marcadores de classe mais nítidos para o algoritmo do que competências subjetivas de escrita.

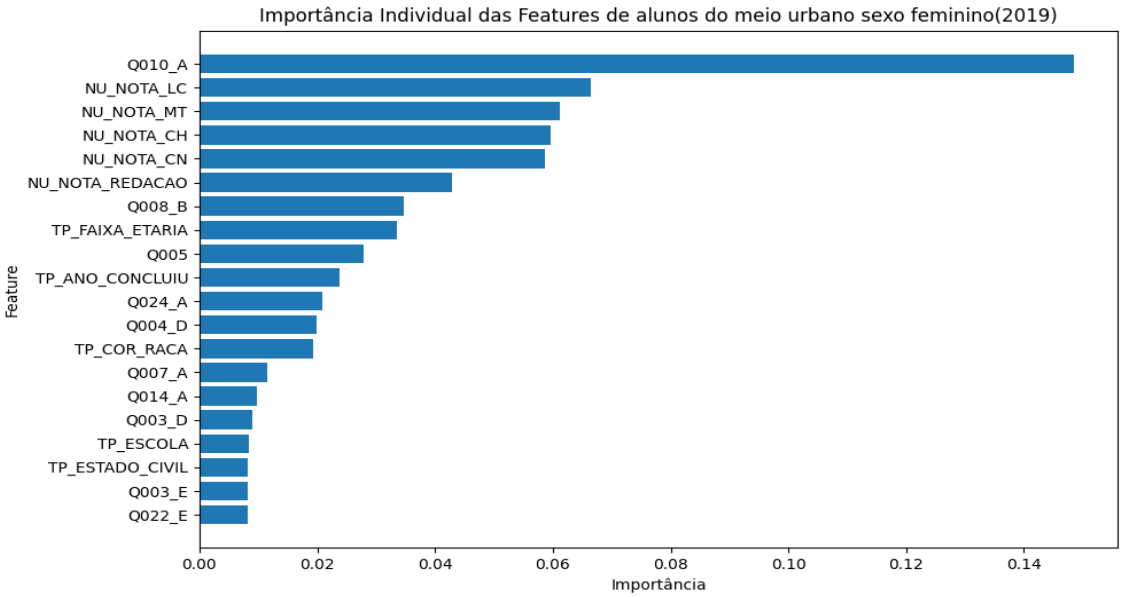
As divergências de gênero revelam como o contexto doméstico e a trajetória escolar impactam os grupos de forma distinta. Para os homens urbanos, observa-se que o fator temporal (faixa etária e ano de conclusão) possui um peso preditivo elevado, sugerindo que o atraso escolar é um marcador de vulnerabilidade mais nítido para este grupo. Já para as mulheres urbanas, a hierarquia é governada pela estrutura do lar: a infraestrutura sanitária e o adensamento familiar são preditores mais rígidos de renda do que a idade. Tais evidências são corroboradas pelos rankings de importância individual de 2019, conforme ilustrado nas Figuras 14 (estrato masculino) e 15 (estrato feminino).

Figura 14. Importância individual das variáveis da árvore de decisão - Estrato masculino



Fonte: Autoria própria.

Figura 15. Importância individual das variáveis da árvore de decisão - Estrato feminino



Fonte: Autoria própria.

Em suma, a evolução dos rankings individuais demonstra que a estratificação urbana em 2023 é definida por um teto de acesso material. Enquanto o sucesso nas provas objetivas (especialmente Matemática e Linguagens) sinaliza a permanência no topo, é a falta de itens básicos de infraestrutura doméstica que ancora os candidatos na base da pirâmide, independentemente do gênero ou da trajetória demográfica.

A aplicação dos valores SHAP (*SHapley Additive exPlanations*) ao estrato urbano permite quantificar como a infraestrutura doméstica e o desempenho escolar empurram as predições para os extremos da pirâmide social. A análise comparativa revela uma estrutura de desigualdade urbana que se tornou mais polarizada e dependente de marcadores de consumo durável.

- **A Dominância da Imobilidade (Q010_A):** Em todos os ciclos, a ausência de automóvel (**Q010_A**) manteve-se como o fator de maior impacto absoluto, com valores SHAP que evoluíram de $\approx 0,25$ em 2019 para $> 0,30$ em 2023. Este indicador é o principal responsável por ancorar as predições nas faixas de renda mais baixas (Até 2 salários), evidenciando que a privação de mobilidade é o traço mais nítido da vulnerabilidade urbana brasileira.
- **Triade de Exclusão Habitacional:** A ausência de computador (**Q024_A**) e a presença de apenas um banheiro (**Q008_B**) atuam como marcadores indissociáveis da baixa renda nas cidades. Em 2023, observou-se que a falta de conforto doméstico (incluindo a ausência de máquina de lavar ou empregado doméstico) passou a atuar de forma mais binária na separação entre as classes populares e a classe média.
- **Matemática como Passaporte de Elite:** Academicamente, a nota de Matemática (**NU_NOTA_MT**) consolidou-se como o principal diferenciador para os estratos superiores (acima de 10 salários). Enquanto as outras notas apresentam impactos distribuídos, o sucesso em exatas funciona como um sinalizador concentrado de privilégio econômico no ambiente urbano.

Tais achados são corroborados pela distribuição dos valores SHAP para o estrato urbano em 2019, conforme ilustrado na Figura 16, que quantifica o impacto de cada variável na classificação das faixas de renda.

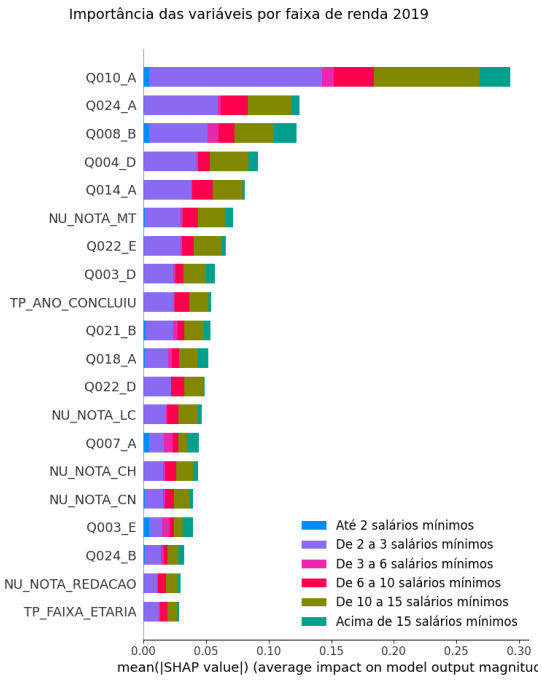


Figura 16. Importância das variáveis SHAP - Estrato Urbano Consolidado

Fonte: Autoria própria.

As divergências de gênero no cenário urbano demonstram como o capital cultural é interpretado de forma distinta. Historicamente (2018-2019), o modelo identificava a posição social feminina prioritariamente através da proficiência em Linguagens (**NU_NOTA_LC**). No entanto, em 2023, observa-se uma ascensão da Matemática no estrato feminino, embora as mulheres ainda apresentem uma dependência maior da combinação entre excelência acadêmica transversal (Linguagens e Redação) e infraestrutura tecnológica para a definição de seus perfis socioeconômicos.

Em suma, a evolução dos valores SHAP demonstra que a segmentação urbana é governada por uma lógica de exclusão: enquanto a falta de bens físicos (carro, computador e banheiro) define com precisão a base da pirâmide, é a acumulação de capital ocupacional familiar (**Q004_D**) e a alta performance em exatas que determinam o acesso ao topo.

5.4. Análise do estrato rural

A análise interseccional do estrato rural entre 2018 e 2023 revela uma estrutura de desigualdade distinta do meio urbano, caracterizada por um teto de vidro acadêmico que limita a conversão de capital econômico em desempenho escolar. Ao contrário da progressão linear urbana, o campo apresenta uma saturação de resultados nos estratos superiores.

Os pontos de convergência e evolução histórica indicam:

- **Vulnerabilidade Geográfica Cumulativa:** Em todos os ciclos analisados, as médias rurais são sistematicamente inferiores às urbanas. Em 2023, o pico de rendimento rural (589,22) permanece abaixo do patamar da classe média urbana, reforçando que o isolamento geográfico e a escassez de infraestrutura pedagógica de elite atuam como barreiras estruturais que o capital financeiro isolado não consegue transpor. Tais achados são corroborados através da Tabela 6

Tabela 6. Estatísticas descritivas das médias das notas por faixa de renda - Estrato Rural Consolidado (2023)

Faixa de renda	Mediana	Q1	Q3	Média	Desvio padrão	Tamanho (N)	IC 95%
Até 2 salários mínimos	537,78	455,56	537,78	502,55	78,45	28.679	[501,64 ; 503,45]
De 2 a 3 salários mínimos	537,78	491,26	578,55	534,36	77,53	2.051	[531,00 ; 537,71]
De 3 a 6 salários mínimos	551,87	518,51	615,11	559,27	81,08	414	[556,80 ; 561,74]
De 6 a 10 salários mínimos	592,98	537,78	646,98	588,37	79,80	557	[581,73 ; 595,01]
De 10 a 15 salários mínimos	589,84	537,78	652,58	589,22	81,32	149	[576,06 ; 602,39]
Acima de 15 salários mínimos	537,78	506,33	591,97	546,92	82,92	68	[526,85 ; 566,99]

Fonte: Autoria própria a partir dos microdados do ENEM 2023.

- **A Anomalia da Elite Rural:** Observa-se uma tendência de estagnação ou declínio no topo da pirâmide (acima de 15 salários mínimos). Especialmente em 2019 e 2023, as médias da classe mais abastada foram inferiores às das faixas intermediárias (6 a 15 salários). Esse fenômeno sugere que os candidatos da elite rural que permanecem vinculados ao campo durante o exame possuem trajetórias educacionais mais heterogêneas ou priorizam a sucessão ocupacional em detrimento da excelência acadêmica no ENEM.
- **Resiliência e Superação Feminina:** O estrato feminino rural consolida-se como o motor de eficiência do campo. Em 2018 e 2023, as mulheres alcançaram as maiores médias

e mantiveram uma evolução mais estável que a masculina. Enquanto os homens rurais sofrem quedas acentuadas de rendimento nos estratos superiores, as mulheres conseguem extrair retornos de proficiência mais consistentes, indicando que o capital cultural familiar é melhor aproveitado pelas candidatas em contextos de restrição geográfica.

Em suma, a ruralidade impõe um limite ao impacto da renda. A desigualdade no campo não é apenas uma questão de ter ou não ter, mas de um sistema educacional que não oferece suporte para que mesmo os estudantes mais ricos atinjam os níveis de elite observados nas metrópoles, deixando para as mulheres o papel de protagonistas da resiliência acadêmica rural.

A Figura 17 ilustra a distribuição das notas por faixa de renda, confirmando visualmente a tendência de ascensão das médias conforme o incremento do capital econômico. Contudo, é fundamental observar que os intervalos de confiança de 95% (representados pelas barras vermelhas) tornam-se progressivamente maiores nas faixas de renda superiores. Esse alargamento é reflexo direto da baixa densidade amostral nesses estratos (conforme detalhado na Tabela 6), o que resulta em estimativas menos precisas para as elites rurais. Adicionalmente, nota-se uma significativa sobreposição dos quartis entre as faixas intermediárias e superiores; embora as médias se desloquem, a amplitude das caixas indica que uma parcela considerável de candidatas de diferentes estratos socioeconômicos compartilha faixas de desempenho semelhantes, sugerindo que a ruralidade atua como um fator de homogeneização parcial dos resultados acadêmicos.

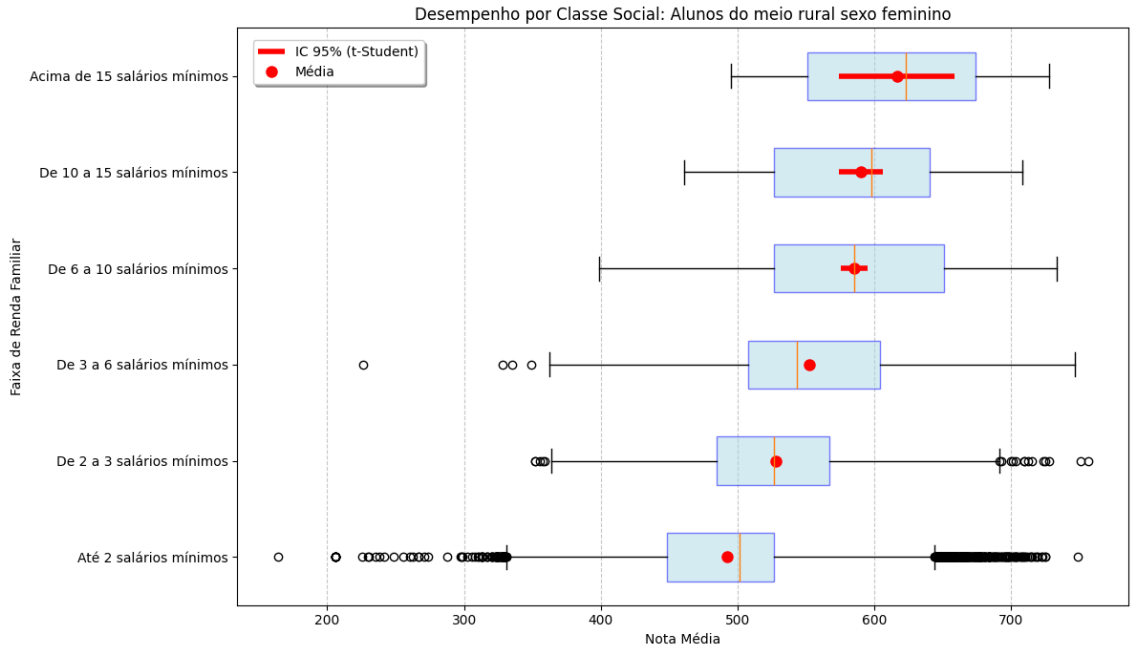


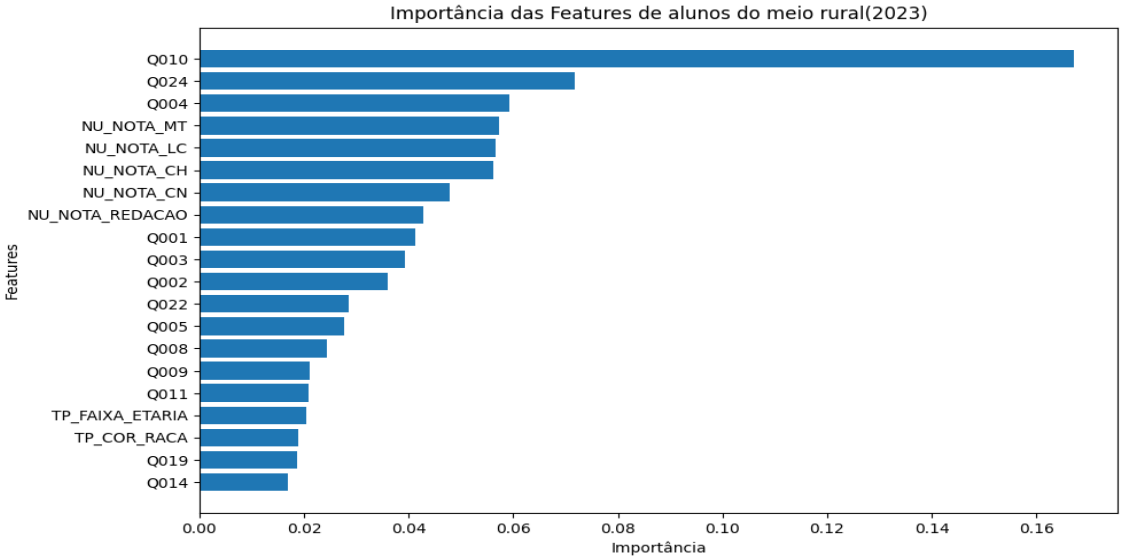
Figura 17. Distribuição das médias das notas por faixa de renda - Estrato rural feminino(2018)

Fonte: Autoria própria.

A hierarquia de importância das variáveis no estrato rural revela que a estratificação socioeconômica no campo é regida pelo binômio mobilidade e conectividade. Conforme ilustrado na Figura 18, observa-se que, no ciclo de 2023, o isolamento

geográfico e digital consolidou-se como o principal filtro de renda, superando inclusive o peso do desempenho acadêmico direto na classificação do modelo.

Figura 18. Importância agregada das variáveis da árvore de decisão - Estrato Rural (2023)

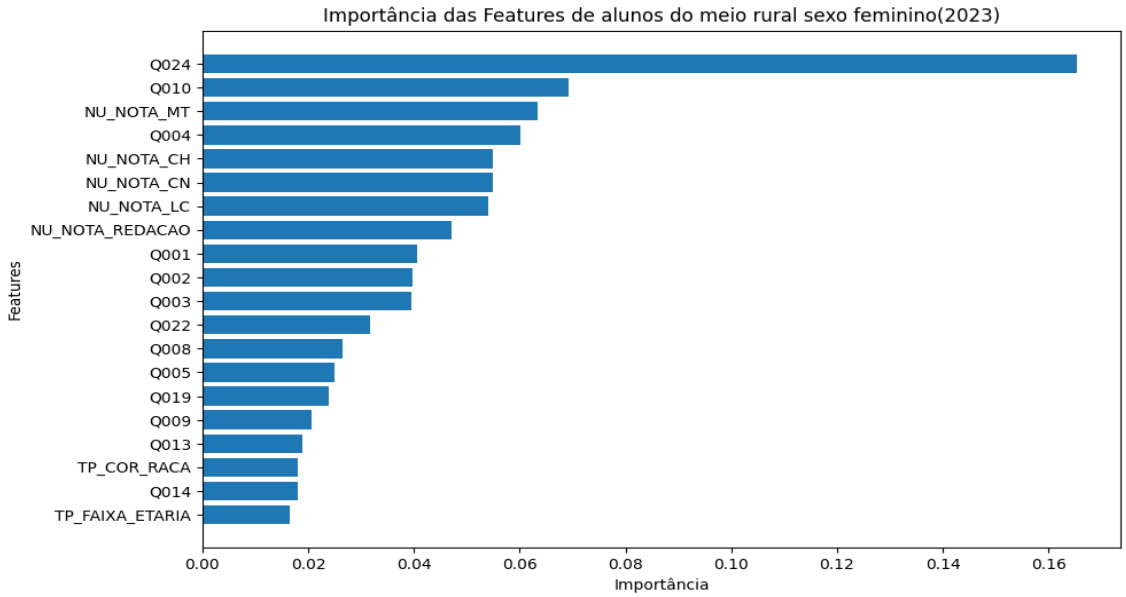


Fonte: Autoria própria.

Os principais eixos de evolução indicam:

- **Primazia da Mobilidade (Q010):** A posse de automóvel consolidou-se como o preditor mais robusto em todos os ciclos, com um crescimento notável em sua importância relativa (atingindo $> 0,175$ em 2023). No contexto rural, o transporte motorizado é o divisor de águas entre a vulnerabilidade e o acesso a mercados e polos educacionais, funcionando como o marcador de renda mais estável do modelo.
- **Transição do Capital Físico para o Digital:** Enquanto nos modelos de 2018 a infraestrutura habitacional mínima (banheiros) possuía relevância central, em 2023 a posse de computador (**Q024**) e a atividade de ocupação da mãe (**Q004**) assumiram o protagonismo. Esse deslocamento sugere que o acesso à tecnologia e a inserção profissional dos responsáveis são agora os filtros que melhor distinguem a classe média rural das famílias em situação de subsistência.
- **Convergência no Bloco Acadêmico:** Observou-se uma evolução nas nuances de gênero. Se em 2018 a distinção acadêmica de renda era pautada por Linguagens (feminino) e Matemática (masculino), os ciclos seguintes revelam uma mudança de padrão. A partir de 2019, consolidando-se em 2023, a nota de Matemática (**NU_NOTA_MT**) tornou-se o sinalizador universal de renda elevada para ambos os sexos no campo, conforme evidenciado pela posição de destaque desta variável na Figura 19.

Figura 19. Importância agregada das variáveis da árvore de decisão - Estrato Rural Feminino (2023)

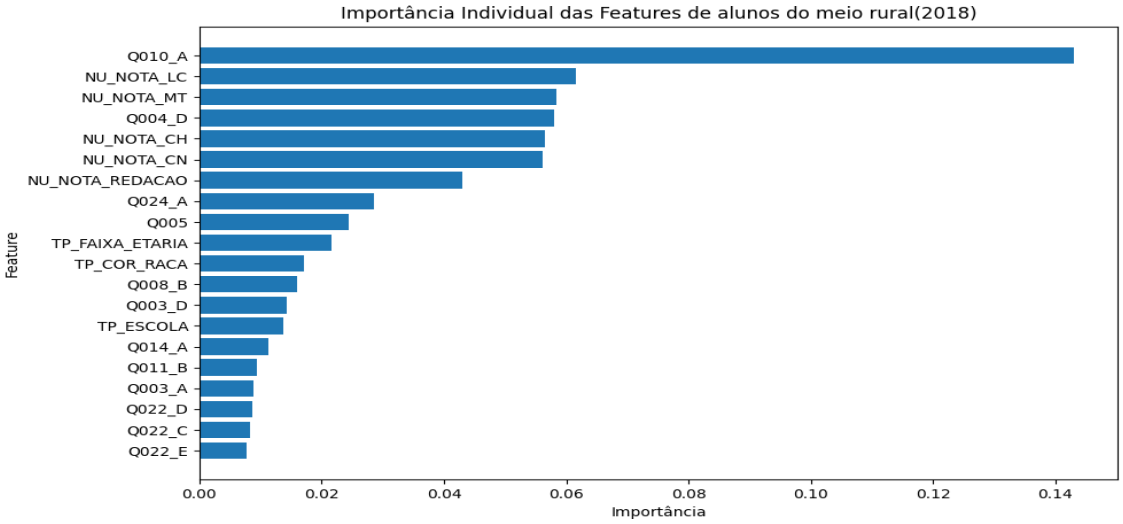


Fonte: Autoria própria.

Diferente do meio urbano, variáveis demográficas e de trajetória, como cor/raça e idade, perdem força preditiva no campo para ativos materiais e profissionais. Em suma, o perfil rural demonstra que a classificação de renda é governada pela capacidade da família de prover os meios físicos (transporte) e tecnológicos (computador) para superar o isolamento, sendo o sucesso em exatas o principal reflexo acadêmico desse capital acumulado.

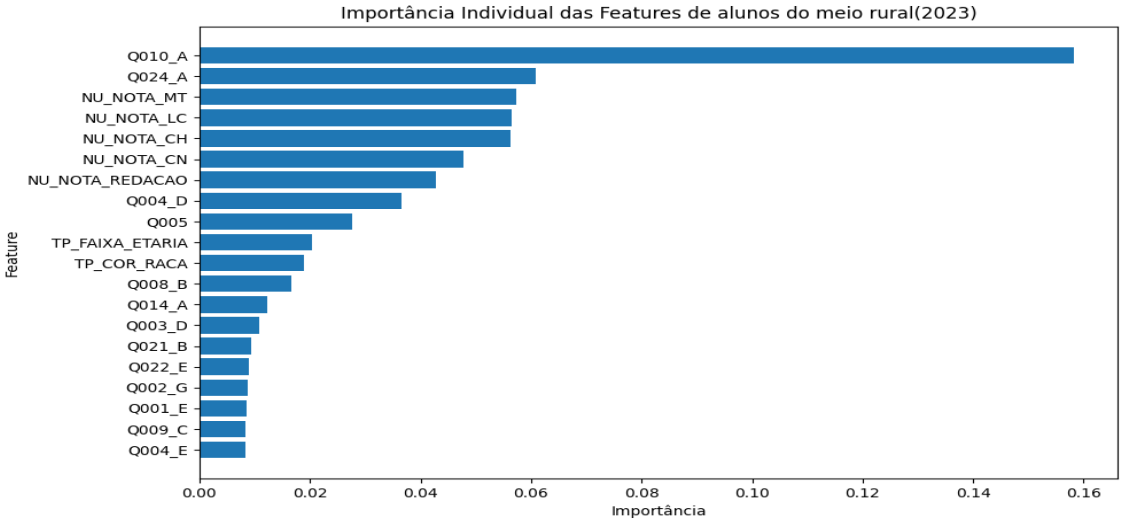
A decomposição em *dummies* da importância das variáveis no estrato rural identifica os marcos de privação que ancoram os candidatos na base da pirâmide socioeconômica. A evolução temporal demonstra uma transição significativa nos divisores de classe: enquanto em 2018 o isolamento geográfico, representado pela ausência de automóvel (**Q010_A**), detinha a hegemonia absoluta como preditor de vulnerabilidade (Figura 20), os dados de 2023 revelam que a exclusão digital (**Q024_A**) passou a rivalizar diretamente com esse fator, conforme ilustrado na Figura 21. Essa mudança sugere que a conectividade tecnológica tornou-se um marcador de renda tão crítico quanto a mobilidade física para a classificação socioeconômica no campo.

Figura 20. Importância individual das variáveis da árvore de decisão - Estrato rural(2018)



Fonte: Autoria própria.

Figura 21. Importância individual das variáveis da árvore de decisão - Estrato Rural Consolidado (2023)



Fonte: Autoria própria.

Os principais eixos de análise revelam:

- **A Hegemonia da Privação Material (Q010_A):** Historicamente, a ausência de automóvel (**Q010_A**) consolidou-se como o marcador de vulnerabilidade mais decisivo, especialmente para o grupo masculino, onde seu peso atingiu $0,17$ em 2023. No campo, a falta de transporte privativo atua como um divisor estatístico nítido, isolando a base da pirâmide dos demais estratos.
- **A Ascensão da Exclusão Digital (Q024_A):** Uma mudança paradigmática ocorre no estrato feminino em 2023. Diferente dos anos anteriores, a ausência de computador (**Q024_A**) assumiu a liderança como o principal preditor individual $(0,16)$, deslocando a mobilidade física para uma posição secundária. Este achado sugere que, para as

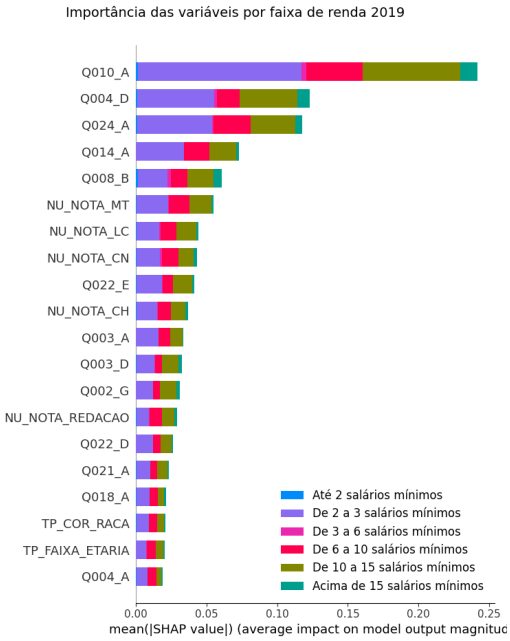
mulheres rurais, o acesso à tecnologia tornou-se o divisor de classe mais sensível no cenário pós-pandêmico.

- **Sinalização Acadêmica de Classe:** Observou-se um deslocamento na relevância das competências. Se em 2018 a proficiência verbal (Linguagens) era o principal sinalizador de capital cultural, a partir de 2019 a nota de Matemática (**NU_NOTA_MT**) assumiu o protagonismo, especialmente entre os homens rurais, funcionando como o indicador intelectual de renda mais robusto do modelo.

Diferente do meio urbano, o estrato rural feminino em 2023 apresenta uma estrutura de estratificação mais complexa, onde variáveis de trajetória (faixa etária) e interseccionalidade (cor/raça) ganham relevância na base do ranking. Em suma, os dados sugerem que a pobreza rural masculina permanece ancorada na falta de ativos físicos e no desempenho técnico, enquanto a vulnerabilidade feminina rural é cada vez mais definida pela exclusão digital e por barreiras sociais estruturais.

A aplicação dos valores SHAP ao cenário rural permite quantificar o peso de cada variável na classificação dos candidatos ao longo da pirâmide social. A evolução entre 2018 e 2023 revela uma transição crítica nos mecanismos de estratificação do campo; nesse contexto, a Figura 22 ilustra o cenário de 2019, evidenciando como a imobilidade física e o capital ocupacional ainda detinham o maior poder preditivo.

Figura 22. Importância das variáveis SHAP - Estrato rural(2019)



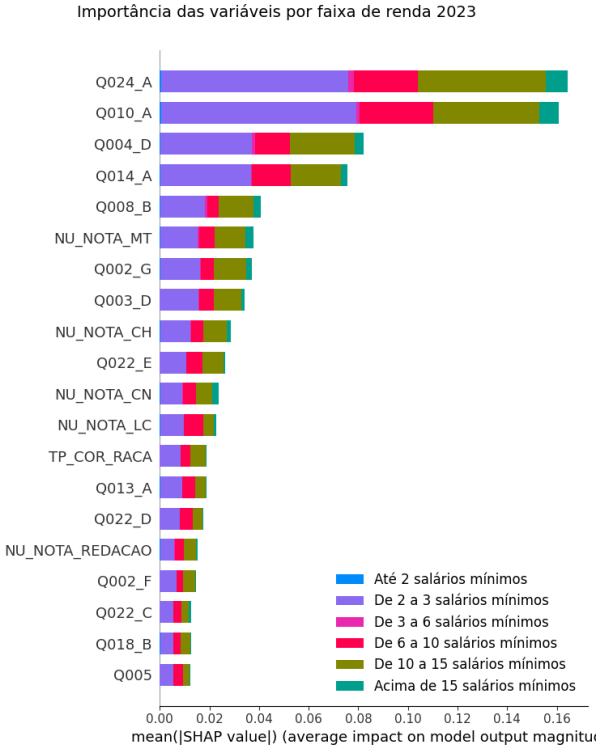
Fonte: Elaboração própria.

- **A Âncora da Vulnerabilidade (Q010_A):** Em 2018 e 2019, a imobilidade física (ausência de automóvel) era o principal fator que empurrava o candidato rural para a base da pirâmide, com impactos SHAP próximos a $0,25$. Em 2023, esse impacto intensificou-se no grupo masculino ($> 0,30$), consolidando a posse de transporte como o critério mais rígido de distinção para os homens no campo que pode ser visto no grafico shap de 2019 atraves da Figura 22.

- **O Paradigma Digital Feminino (Q024_A):** O achado mais disruptivo de 2023 ocorre no estrato feminino, onde a exclusão digital (Q024_A) sobrepôs-se à falta de mobilidade como o principal fator de impacto. Para as mulheres rurais, o acesso tecnológico tornou-se o divisor de águas mais preciso para a classificação socioeconômica, superando barreiras geográficas tradicionais.
- Matemática como Validador de Elite: No bloco acadêmico, observa-se que a nota de Matemática (NU_NOTA_MT) evoluiu de um papel secundário em 2018 para o posto de principal indicador educacional em 2023. O SHAP demonstra que, no topo da pirâmide rural, o impacto das notas de exatas torna-se proporcionalmente superior ao dos bens de consumo básico, funcionando como um refinador de classe média-alta e alta.

A divergência nas trajetórias de ascensão rural é ratificada pelos valores SHAP: no estrato masculino, prevalece o impacto do transporte e das exatas; já no estrato feminino, observa-se um protagonismo do capital ocupacional (Q004_D) e da conectividade. Esse novo padrão de estratificação feminina, mais dependente do acesso tecnológico e do desempenho escolar, é ilustrado graficamente na Figura 23.

Figura 23. Importância das variáveis SHAP - Estrato Rural Feminino (2023)



Fonte: Elaboração própria.

6. Conclusão

Este trabalho teve como objetivo realizar uma análise descritiva e interpretável da relação entre variáveis educacionais, demográficas e socioeconômicas presentes nos microdados do ENEM nos anos de 2018, 2019 e 2023. Diferenciando-se das abordagens predominantemente preditivas, a pesquisa utilizou modelos de árvore de decisão intencionalmente sobreajustados e técnicas de explicabilidade (*Explainable AI* - XAI) para mapear os padrões

internos dos dados e identificar os determinantes que caracterizam os diferentes estratos de renda familiar dos participantes.

6.1. Principais Achados e Interpretações

Os resultados deste estudo permitem concluir que o sistema de avaliação brasileiro, representado pelo ENEM, ainda atua como um mecanismo de reprodução de privilégios, onde o desempenho acadêmico é indissociável da infraestrutura material prévia do estudante. A análise evidencia que o capital tecnológico e o patrimônio acumulado não são apenas auxiliares, mas preditores determinantes que precedem a proficiência escolar.

A hegemonia da variável Q010 (posse de automóvel) como o indicador mais robusto de renda e desempenho em 2018, 2019 e 2023 revela que a mobilidade física e a estabilidade econômica familiar continuam sendo as barreiras estruturais primárias para a maioria dos candidatos. No entanto, a inversão desse padrão observada no grupo de mulheres rurais em 2023 — onde a posse de computador (Q024) assumiu o protagonismo — permite concluir que houve uma reconfiguração das necessidades estruturais no cenário pós-pandemia.

Essa transição sugere que, para grupos em situações de múltiplas vulnerabilidades, a exclusão digital tornou-se um fator de isolamento mais severo do que a própria falta de transporte, transformando o acesso à informação no novo divisor de águas da desigualdade educacional brasileira. Assim, o trabalho demonstra que a interseccionalidade entre gênero, localização e acesso técnico é fundamental para entender por que certas políticas públicas genéricas falham ao não considerar que o obstáculo de uma mulher rural em 2023 é qualitativamente diferente do obstáculo de um estudante urbano.

6.2. Limitações e Trabalhos Futuros

A principal limitação deste estudo reside na impossibilidade de generalização preditiva, dada a ausência de separação entre conjuntos de treino e teste, além de restrições temporais impostas por mudanças metodológicas nos microdados do INEP a partir de 2024. Para o aprimoramento desta linha de pesquisa, sugerem-se os seguintes trabalhos futuros:

- **Análise de Séries Temporais:** Investigar métodos de harmonização para incluir edições do ENEM a partir de 2024;
- **Justiça Algorítmica:** Incorporar métricas voltadas à análise de viés, investigando se os padrões identificados diferem sistematicamente entre grupos minoritários;
- **Ferramentas de Apoio à Decisão:** Desenvolver *dashboards* de explicabilidade que permitam a gestores públicos explorarem as regras extraídas pelas árvores de decisão para a formulação de políticas baseadas em evidências.
- **Aprofundamento Interseccional:** Explorar outros cruzamentos, como cor/raça e dependência administrativa da escola (pública vs. privada), utilizando o SHAP para identificar se a exclusão digital atinge de forma diferente mulheres rurais brancas e negras.

Por fim, conclui-se que a abordagem proposta mostrou-se eficaz para a análise interpretável de dados em larga escala. Ao priorizar a explicabilidade, o estudo oferece subsídios críticos para pesquisadores e formuladores de políticas públicas, reforçando que a Inteligência Artificial pode ser uma poderosa aliada da sociologia da educação ao tornar compreensíveis os complexos fatores que moldam a realidade educacional no Brasil.

7. Repositório do Projeto

O código-fonte desenvolvido neste trabalho, bem como os scripts de pré-processamento dos dados, modelos implementados e materiais auxiliares para reprodução dos experimentos, encontram-se disponíveis em um repositório público no GitHub. O acesso ao repositório pode ser realizado por meio do seguinte endereço:

`https://github.com/vinicarlosss/enem-explainable-ml`

Referências

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

Arrieta, A. B., Rodríguez, N., Del Ser, J., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

Baker, R. and Inventado, P. (2019). Educational data mining and learning analytics. *The Cambridge Handbook of the Learning Sciences*.

Baker, R. S. 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.

Bourdieu, P. and Passeron, J.-C. (2007). *A Reprodução: elementos para uma teoria do sistema de ensino*. Vozes, Petrópolis.

Castro, F., Vellido, A., Nebot, A., and Mugica, F. (2007). Data mining in education: An experimental study. In *Proceedings of the 16th IASTED International Conference on Applied Simulation and Modelling*, pages 145–149.

Choi, W. C., Lam, C. T., and Mendes, A. J. (2024). Analyzing the interpretability of machine learning prediction on student performance using shapley additive explanations. In *TALE 2024 - IEEE International Conference on Teaching, Assessment and Learning for Engineering*.

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, 1989(1):139–167.

de Souza, M. and Klug, D. L. (2025). Desigualdades educacionais no enem: uma perspectiva baseada em variáveis socioeconômicas e aprendizagem de máquina. *Revista de Gestão e Avaliação Educacional*.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. In *Workshop on Human Interpretability in Machine Learning (WHI)*.

Faceli, K., Lorena, A. C., Gama, J., and Carvalho, A. C. P. L. F. (2011). *Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina*. LTC, Rio de Janeiro.

Géron, A. (2021). *Mãos à Obra: Aprendizado de Máquina com Scikit-Learn, Keras & TensorFlow*. O'Reilly/Alta Books, Rio de Janeiro, 2 edition.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media, New York.

IBGE (2022). *Síntese de indicadores sociais: uma análise das condições de vida da população brasileira*. INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA, Rio de Janeiro.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer, New York.

Lipton, Z. (2018). The mythos of model interpretability. *Communications of the ACM*.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, New York.

Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Lulu.com, 2nd edition. Acessado em: 06 fev. 2026.

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining process. *Expert Systems with Applications*, 41(4):1432–1459.

Romero, C. and Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics*.

Romero, C. and Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3):e1355.

Shapley, L. S. (1953). A value for n-person games. In *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press.

Silva, F. d. C. and Santana, A. M. (2024). Explicabilidade dos modelos de aprendizado de máquina no cenário educacional: uma revisão sistemática. *RENOTE*, 22(1):477–486.

Silva, M. (2020). Desigualdades educacionais no brasil: uma análise sobre o acesso ao ensino superior. *Revista Brasileira de Educação*, 25.

Sorj, B. (2003). *Brasil @ digital.br: exclusão digital, estratificação social e políticas públicas*. Jorge Zahar, Rio de Janeiro.

Villas Boas, P. R. (2023). Interpretabilidade de modelos aplicados aos dados do enem. In *Anais do II Workshop de Matemática, Estatística e Computação Aplicadas à Indústria (WMECAI)*.

Vinces-Vinces, F. V. and Flores-Sanchéz, M. (2025). Application of machine learning to predict and explain university academic performance. *Comunicar*.

A. Descrição das Variáveis do Questionário Socioeconômico

Este apêndice apresenta a descrição das questões do questionário socioeconômico do Exame Nacional do Ensino Médio (ENEM) utilizadas neste trabalho, correspondentes às variáveis Q001 a Q025, cujas formulações completas encontram-se organizadas na Tabela 7. Essas questões fornecem informações relacionadas à renda familiar, posse de bens e condições de infraestrutura domiciliar dos participantes, sendo fundamentais para a caracterização do perfil socioeconômico analisado. As descrições apresentadas seguem o dicionário de dados oficial disponibilizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP).

Tabela 7. Descrição das questões do questionário socioeconômico utilizadas

Variável	Descrição da questão
Q001	Até que série seu pai, ou o homem responsável por você, estudou?
Q002	Até que série sua mãe, ou a mulher responsável por você, estudou?
Q003	A partir da apresentação de algumas ocupações divididas em grupos ordenados, indique o grupo que contempla a ocupação mais próxima da ocupação do seu pai ou do homem responsável por você. (Se ele não estiver trabalhando, escolha uma ocupação pensando no último trabalho dele).
Q004	A partir da apresentação de algumas ocupações divididas em grupos ordenados, indique o grupo que contempla a ocupação mais próxima da ocupação da sua mãe ou da mulher responsável por você. (Se ela não estiver trabalhando, escolha uma ocupação pensando no último trabalho dela).
Q005	Incluindo você, quantas pessoas moram atualmente em sua residência?
Q006	Qual é a renda mensal de sua família? (Some a sua renda com a dos seus familiares).
Q007	Em sua residência trabalha empregado(a) doméstico(a)?
Q008	Na sua residência tem banheiro?
Q009	Na sua residência tem quartos para dormir?
Q010	Na sua residência tem carro?
Q011	Na sua residência tem motocicleta?
Q012	Na sua residência tem geladeira?
Q013	Na sua residência tem freezer (independente ou segunda porta da geladeira)?
Q014	Na sua residência tem máquina de lavar roupa? (O tanquinho NÃO deve ser considerado).
Q015	Na sua residência tem máquina de secar roupa (independente ou em conjunto com a máquina de lavar roupa)?
Q016	Na sua residência tem forno micro-ondas?
Q017	Na sua residência tem máquina de lavar louça?
Q018	Na sua residência tem aspirador de pó?
Q019	Na sua residência tem televisão em cores?
Q020	Na sua residência tem aparelho de DVD?
Q021	Na sua residência tem TV por assinatura?
Q022	Na sua residência tem telefone celular?
Q023	Na sua residência tem telefone fixo?
Q024	Na sua residência tem computador?
Q025	Na sua residência tem acesso à Internet?

B. Descrição das categorias da questão Q006

A Tabela 8 apresenta a descrição original das categorias da questão Q006 do questionário socioeconômico do ENEM, conforme disponibilizado pelo INEP.

Tabela 8. Descrição das categorias da variável Q006 — Renda mensal familiar

Código	Descrição da renda mensal familiar
A	Nenhuma renda
B	Até R$ 1.212,00
C	De R$ 1.212,01 até R$ 1.818,00
D	De R$ 1.818,01 até R$ 2.424,00
E	De R$ 2.424,01 até R$ 3.030,00
F	De R$ 3.030,01 até R$ 3.636,00
G	De R$ 3.636,01 até R$ 4.848,00
H	De R$ 4.848,01 até R$ 6.060,00
I	De R$ 6.060,01 até R$ 7.272,00
J	De R$ 7.272,01 até R$ 8.484,00
K	De R$ 8.484,01 até R$ 9.696,00
L	De R$ 9.696,01 até R$ 10.908,00
M	De R$ 10.908,01 até R$ 12.120,00
N	De R$ 12.120,01 até R$ 14.544,00
O	De R$ 14.544,01 até R$ 18.180,00
P	De R$ 18.180,01 até R$ 24.240,00
Q	Acima de R$ 24.240,00