

Análise da Política de Cotas do ENEM através da Clusterização de Dados de Desempenho

**Humberto Nunes de Lira¹, Vivianny Khatly Medeiros Pereira¹,
Yuri de Almeida Malheiros Barbosa¹**

¹ Universidade Federal da Paraíba (UFPB) – João Pessoa, Paraíba, Brasil

humberto.nunes@academico.ufpb.br, vivianny.khatly@academico.ufpb.br

yuri@ci.ufpb.br

Abstract. *Identifying vulnerable groups is essential for effective educational policies. This study maps socio-academic profiles of incoming students at higher education institutions and assesses the impact of quotas on access equity. A clustering technique was applied to 497,070 records from the 2023 and 2024 admission cycles, obtained through public information access requests, using scores across five knowledge areas of the National Upper Secondary Education Exam. Results reveal four profiles associated with socioeconomic background and show that ethnic-racial quotas produce the greatest reduction in entry barriers: 11.30% versus 4.31% for the group without this criterion. Such policies are central to democratizing access, reinforcing need for continuous monitoring.*

Resumo. *A identificação de grupos vulneráveis é essencial para políticas educacionais eficazes. Este estudo mapeia perfis socioacadêmicos de ingressantes em instituições públicas e avalia o impacto das cotas na equidade de acesso. Aplicou-se uma técnica de agrupamento a 497.070 registros dos ciclos de 2023 e 2024, obtidos por acesso à informação pública, com notas nas cinco áreas do Exame Nacional do Ensino Médio. Os resultados revelam quatro perfis associados à origem socioeconômica e mostram que as cotas com recorte étnico-racial promovem a maior redução na barreira de entrada: 11,30% frente a 4,31% no grupo sem esse recorte. Conclui-se que tais políticas são centrais para a democratização do acesso, reforçando a necessidade de monitoramento contínuo.*

1. Introdução

Instituído em 1998, o Exame Nacional do Ensino Médio (ENEM) tornou-se uma das principais vias de acesso ao ensino superior público brasileiro, sobretudo via Sistema de Seleção Unificada (SISU) [Oliveira 2013], gerando um dos mais relevantes repositórios de informações educacionais do país [Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) 2024]. Apesar desse potencial, não foram identificados estudos que explorem esses dados de forma integrada. [Romero and Ventura 2010], em revisão internacional sobre mineração de dados educacionais, apontam que a área ainda concentra esforços em dimensões isoladas do desempenho, como disciplina, evasão ou rendimento geral, sem considerar simultaneamente fatores socioeconômicos, geográficos e de política educacional. Essa fragmentação se reproduz no Brasil: [Campos et al. 2014] analisa o desempenho de cotistas no ENEM e [Artes and Ricoldi 2015] investiga disparidades socioeconômicas

no acesso, mas nenhum dos dois integra o impacto das cotas, as dinâmicas geográficas e a formação de perfis multidimensionais dos estudantes. Essa lacuna motiva a presente análise exploratória dos dados do ENEM, com base em 497.070 registros de ingressantes em Instituições de Ensino Superior (IES) públicas dos ciclos de 2023 e 2024, contemplando identificação institucional e do curso, notas individuais nas cinco áreas de conhecimento, nota média, modalidade de acesso e distribuição geográfica por unidade federativa. Para identificar os perfis, utilizam-se técnicas de agrupamento, que organizam participantes por características semelhantes sem categorias prévias e são adequadas a dados educacionais multivariados [Jain 2010, Hair Jr. et al. 2019]. O objetivo é identificar perfis socioacadêmicos dos ingressantes no ENEM 2023–2024, investigando o impacto das cotas na equidade de acesso e os padrões regionais de desempenho.

2. Fundamentação Teórica e Trabalhos Relacionados

2.1. Mineração de Dados Educacionais

A Mineração de Dados Educacionais (MDE) consiste na aplicação de técnicas computacionais à extração de padrões em grandes volumes de dados provenientes de contextos de ensino e avaliação. Em revisão abrangente da área, [Romero and Ventura 2010] evidenciam que grande parte dos estudos concentra esforços em objetivos isolados, como predição de desempenho, evasão ou modelagem de trajetórias, sem articular as múltiplas dimensões que caracterizam os estudantes. Como já apontado na Introdução, as técnicas de agrupamento mostram-se especialmente adequadas a esse tipo de dado multivariado, por revelarem estruturas que métodos estatísticos tradicionais tendem a não evidenciar [Jain 2010, Hair Jr. et al. 2019].

2.2. Políticas de Cotas e Equidade no Acesso ao Ensino Superior

O acesso ao ensino superior público brasileiro é atravessado por desigualdades históricas, cujo enfrentamento motivou a adoção de políticas de ação afirmativa, mecanismo de promoção da equidade que amplia as oportunidades de grupos historicamente sub-representados [Sen 2000, Wainer and Melguizo 2018]. A menor pontuação média entre cotistas é comumente associada ao acúmulo de desvantagens na educação básica, e não a uma incapacidade inerente [Bourdieu 1998, Santos 2013]. Ainda assim, as políticas afirmativas têm contribuído para diversificar o perfil dos ingressantes, inclusive em cursos de alta seletividade [Ristoff 2014, Artes and Ricoldi 2015]. Nesse cenário, os estudos de [Campos et al. 2014] e [Artes and Ricoldi 2015] tratam isoladamente o desempenho de cotistas e as disparidades de acesso, sem integrar dinâmicas geográficas e perfis multidimensionais a partir dos microdados do ENEM; é essa lacuna que a presente pesquisa busca preencher, combinando a identificação de perfis socioacadêmicos via agrupamento, a análise das dinâmicas regionais de desempenho e a avaliação do impacto diferencial das modalidades de cota.

3. Metodologia

O delineamento é quantitativo-exploratório, voltado à caracterização dos perfis socioeconômicos e acadêmicos dos ingressantes, análise crucial para identificar estruturas ocultas em dados multivariados [Hair Jr. et al. 2019]. Para isso, utilizou-se o agrupamento por similaridade, técnica que organiza registros em grupos coerentes, sem rótulos pré-definidos [Jain 2010] (Figura 1).

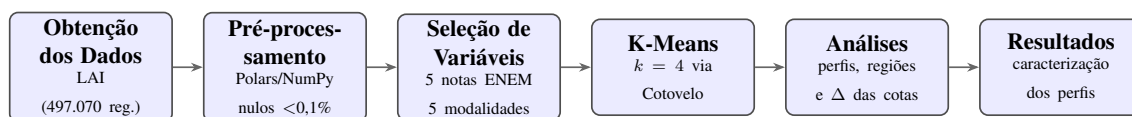


Figure 1. Fluxo metodológico do estudo, das etapas de obtenção e tratamento dos dados até a geração dos resultados.

3.1. Dados e Coleta

Os dados foram obtidos via Lei de Acesso à Informação (LAI), compreendendo 497.070 registros de ingressantes em IES públicas nos ciclos de 2023 e 2024, com ausência de dados para Rondônia (RO). A escolha dessa base decorre de uma lacuna na literatura correlata: [Garcez Filho et al. 2025] investigam o efeito da renda sobre o desempenho no ENEM ao longo de uma série histórica (1998 a 2023), mas sem incorporar a modalidade de cota nem o recorte institucional; [Moura et al. 2025] aplicam aprendizado de máquina para prever o desempenho estudantil, restringindo, porém, a análise ao Maranhão, o que compromete a generalização nacional. Diante dessa fragmentação, optou-se por uma base que reunisse abrangência nacional, informações sobre modalidade de acesso e identificação institucional e de curso, viabilizando uma análise multidimensional. Cabe destacar que a base compreende exclusivamente ingressantes via ENEM/SISU, não contemplando outras vias de acesso, como vestibulares próprios, processos seletivos específicos (e.g., PAS/UnB) ou editais de transferência, que também podem adotar critérios de cota distintos; os perfis e diferenciais aqui identificados são, portanto, representativos apenas do fluxo ENEM/SISU. A opção pelos ciclos de 2023 e 2024 se justifica tanto pela contribuição inédita frente à literatura, que cobre majoritariamente ciclos anteriores, quanto pela disponibilidade efetiva dos dados fornecidos pelas instituições respondentes via LAI, refletindo uma limitação do próprio processo de solicitação. Conforme a Tabela 1, as médias por área situam-se entre 552,2 (Ciências da Natureza) e 634,9 (Matemática), com Redação destacando-se com 776,5 pontos. Redação e Matemática apresentam os maiores desvios padrão (140,0 e 120,5), enquanto Linguagens exibe a maior homogeneidade ($dp = 62,2$). As distribuições aproximadamente simétricas sustentam o uso do K-Means nas etapas subsequentes.

Table 1. Estatísticas descritivas das notas por área de conhecimento.

Área	Média	DP	Mediana	Mín	Máx
Linguagens	572,9	62,2	578,1	0,0	820,8
C. Humanas	590,9	70,0	596,0	0,0	839,2
C. Natureza	552,2	79,1	553,0	0,0	875,3
Matemática	634,9	120,5	640,9	0,0	985,7
Redação	776,5	140,0	800,0	40,0	1000,0
Média Geral	625,5	73,9	626,1	184,0	862,6

A base de dados consolidada agrupa variáveis fundamentais para a caracterização socioacadêmica dos estudantes, estruturadas nos seguintes eixos:

- **Identificação Institucional e Acadêmica:** Ano de ingresso, Sigla, Nome da IES, nome do curso, grau acadêmico e turno.
- **Desempenho no ENEM:** Notas individuais nas cinco áreas de conhecimento (Linguagens, Ciências Humanas, Ciências da Natureza, Matemática e Redação), além do cálculo da nota média.

- **Modalidade de Acesso:** Tipo de concorrência e modalidade de cota, cujas categorias analisadas são detalhadas na sequência.

3.2. Tratamento e Infraestrutura de Dados

Dada a magnitude do *corpus*, o pré-processamento foi realizado em Python com Polars e NumPy, garantindo eficiência computacional no tratamento de valores inconsistentes e na padronização dos atributos. Foram removidos 296 registros com valores nulos ($<0,1\%$ da base), procedimento seguro para amostras de grande porte com menos de 5% de *missing data* [Hair Jr. et al. 2019]. Selecionaram-se as cinco modalidades de concorrência mais representativas, cada uma respondendo por ao menos 6,08% dos registros (Figura 2), sendo as demais excluídas para preservar a robustez estatística. As definições adotadas neste estudo são:

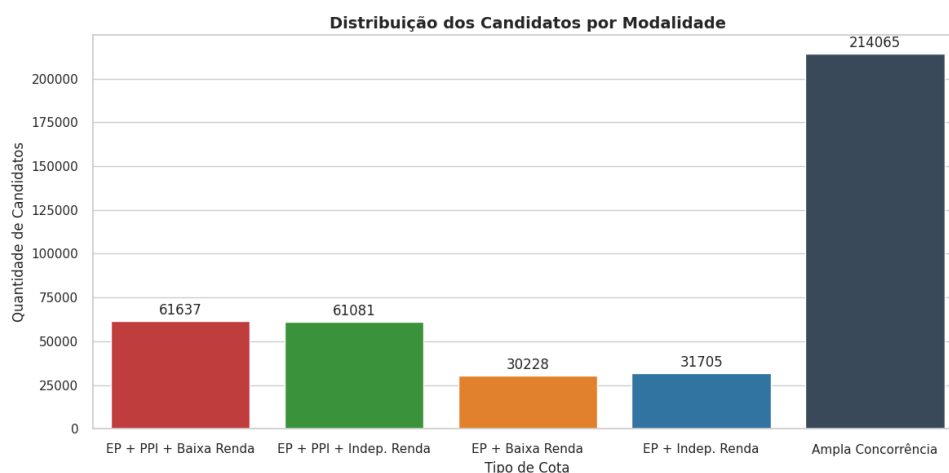


Figure 2. Distribuição de frequência das modalidades de concorrência selecionadas para análise (Ciclos 2023-2024).

Ampla Concorrência: Candidatos que concorrem sem reserva de vagas, independentemente de critérios socioeconômicos ou escolares.

EP + PPI + Baixa Renda: Escola Pública, Pretos/Pardos/Indígenas, Renda $\leq 1,5$ salários.

EP + PPI + Indep. Renda: Escola Pública, Pretos/Pardos/Indígenas, Independente de Renda.

EP + Baixa Renda: Escola Pública, Renda $\leq 1,5$ salários, sem cota racial.

EP + Indep. Renda: Escola Pública, Independente de Renda, sem cota racial.

3.3. Agrupamento por Similaridade

Para segmentar os perfis de ingresso, utilizou-se o algoritmo **K-Means**, cuja eficiência no processamento de grandes volumes de dados multidimensionais e capacidade de organizar registros por distâncias euclidianas o tornam adequado ao contexto educacional [Jain 2010, Hair Jr. et al. 2019]. O número ideal de agrupamentos (k) foi definido pelo **Método do Cotovelo** [Bholowalia and Kumar 2014], cujo ponto de inflexão indicou $k = 4$ como equilíbrio ótimo entre compactação e separação. Como validação complementar, calculou-se o **Coeficiente de Silhueta** [Rousseeuw 1987] sobre uma amostra representativa, medindo a coesão interna e a separação entre os agrupamentos. Conforme

a Figura 3, o score decresce à medida que k aumenta, atingindo seu valor máximo em $k = 2$ (0,325) e um valor moderado em $k = 4$ (0,24). Optou-se pela manutenção de $k = 4$, pois essa configuração oferece maior granularidade analítica e permite caracterizar perfis substantivamente distintos (Seção 4.2), sem comprometer excessivamente a separação entre os grupos, que permanece dentro da faixa considerada razoável pela literatura [Rousseeuw 1987].

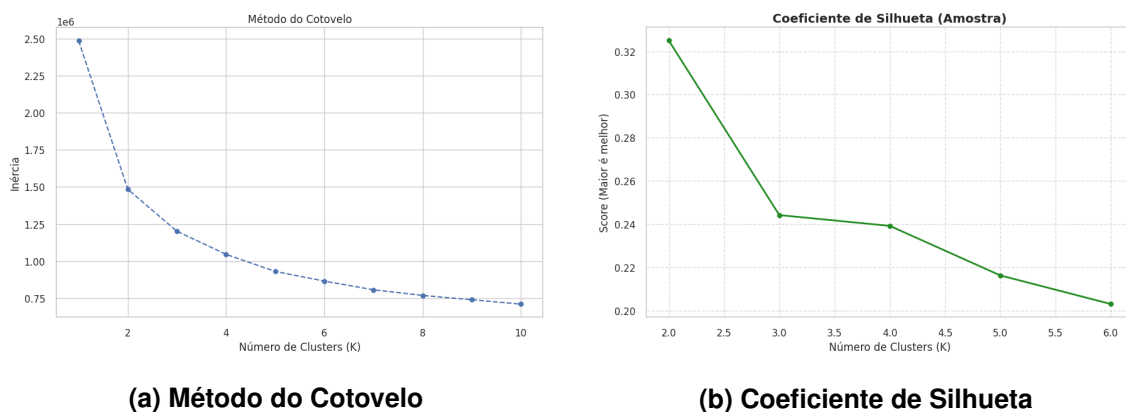


Figure 3. Comparação entre o Método do Cotovelo e o Coeficiente de Silhueta para definição do número de *clusters* (k).

Perfil de Notas por Grupo (Assinatura das Áreas)

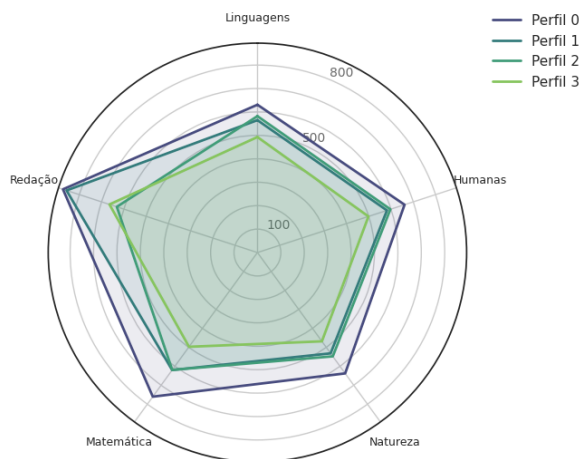


Figure 4. Perfil de desempenho acadêmico (Assinatura das Áreas) baseada nas cinco competências do ENEM.

A segmentação resultou na identificação de quatro perfis hierárquicos (Figura 4), organizados pela magnitude das notas obtidas no exame:

- **Perfil 0:** Agrupa discentes com os maiores desempenhos médios nas cinco áreas de conhecimento do ENEM.
- **Perfis 1 e 2:** Representam faixas de desempenho intermediário, com variações específicas em áreas disciplinares distintas.
- **Perfil 3:** Reúne discentes com os menores índices de desempenho médio registrados na base de dados.

3.4. Avaliação do Impacto e Diferencial Relativo

A Ampla Concorrência (AC) foi adotada como grupo de referência (*baseline*) para avaliar a disparidade nas notas de corte das demais categorias, identificando quais modalidades de cota apresentam impacto significativo no acesso e quais enfrentam desafios para cumprir seu papel de garantia de acessibilidade [Romero and Ventura 2010]. Para isso, utilizou-se o cálculo do Diferencial Relativo (Δ), métrica aplicada por [Campos et al. 2014] para quantificar o percentual de redução da nota de corte da cota em relação ao patamar da AC, conforme definido pela Equação 1:

$$\Delta = \left(\frac{NC_{AC} - NC_{cota}}{NC_{AC}} \right) \times 100 \quad (1)$$

onde NC_{AC} representa a nota de corte da Ampla Concorrência e NC_{cota} a da respectiva modalidade de reserva de vagas; valores positivos indicam o percentual em que a nota da cota é inferior à da AC, evidenciando o *gap* de acesso que a política afirmativa visa mitigar. Esta abordagem torna-se ainda mais relevante diante das atualizações da Lei n.º 14.723/2023, que reestruturou os critérios de renda e a dinâmica de concorrência no SISU [Brasil 2023]. Para assegurar a robustez da análise e mitigar variações sazonais ou *outliers* [Hair Jr. et al. 2019], aplicou-se um critério de refinamento estatístico baseado no limiar de representatividade institucional, considerando apenas cursos ofertados em, no mínimo, cinco universidades distintas. Esse filtro garante que o diferencial reflita um padrão recorrente e consolidado na base de 2023–2024, e não distorções locais ou casos isolados.

4. Resultados

Os resultados estão organizados em quatro eixos complementares: panorama geral do desempenho por modalidade de concorrência; tipologia dos perfis acadêmicos identificados pela clusterização; desigualdades regionais na distribuição desses perfis; e impacto diferencial das modalidades de cotas sobre as barreiras de acesso. Cabe destacar que o desenho deste estudo é observacional e transversal, restrito aos ciclos de 2023–2024; assim, os padrões identificados pela clusterização e pelo Diferencial Relativo são interpretados à luz da literatura sobre desigualdade educacional, sem que isso implique relações de causalidade testadas diretamente pelos dados.

4.1. Panorama Geral e Caracterização da Demanda

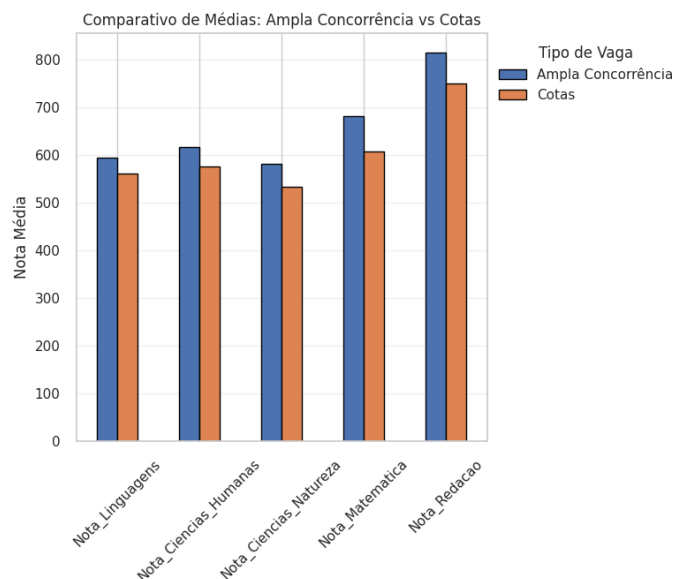


Figure 5. Comparativo de médias por área do conhecimento entre Ampla Concorrência e Cotas (Ciclos 2023-2024).

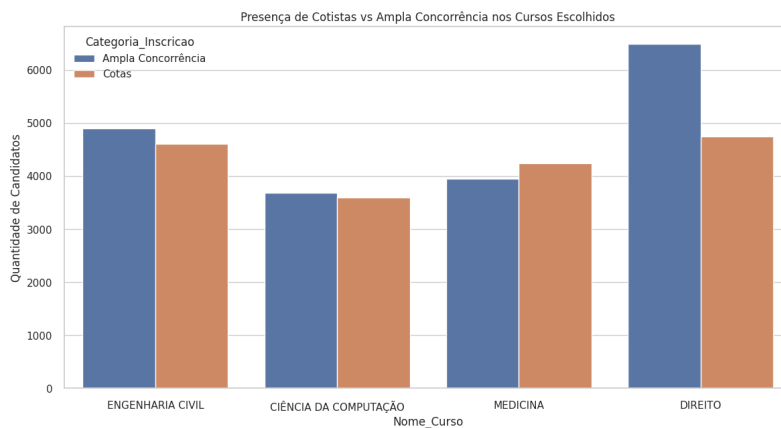


Figure 6. Proporção de ingressantes por modalidade de concorrência em cursos de alta seletividade (destaque para Medicina).

Como ilustrado na Figura 5, o desempenho médio dos cotistas é sistematicamente inferior ao da ampla concorrência em todas as áreas [Campos et al. 2014, Artes and Ricoldi 2015], coerente com as desigualdades de origem discutidas na Seção 2. Apesar dessa defasagem inicial, a Figura 6 mostra que, em cursos como Medicina, o número de ingressantes via cotas já se aproxima ou supera o da ampla concorrência, achado compatível com a literatura que associa políticas afirmativas à ampliação do acesso de grupos sub-representados a cursos de elevada demanda [Ristoff 2014, Santos 2013], embora o desenho transversal dos dados não permita atribuir causalidade direta a essa tendência.

4.2. Tipologia dos Perfis Acadêmicos

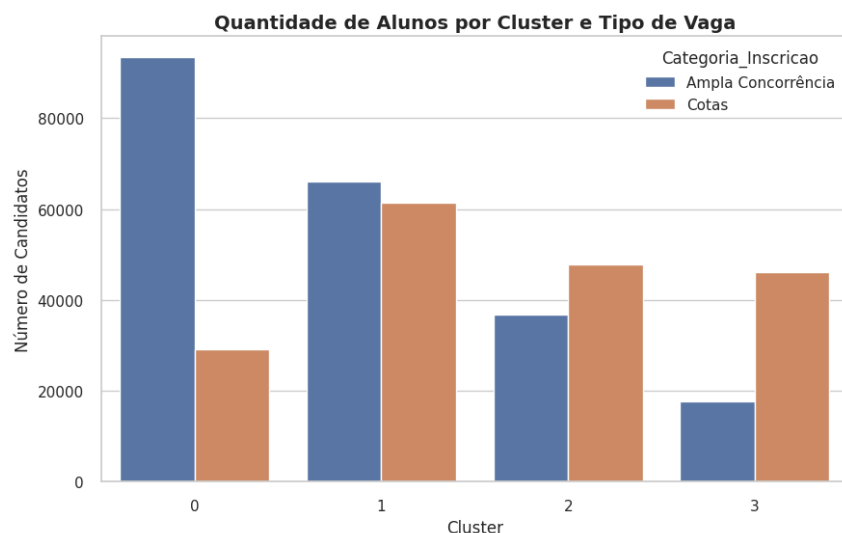
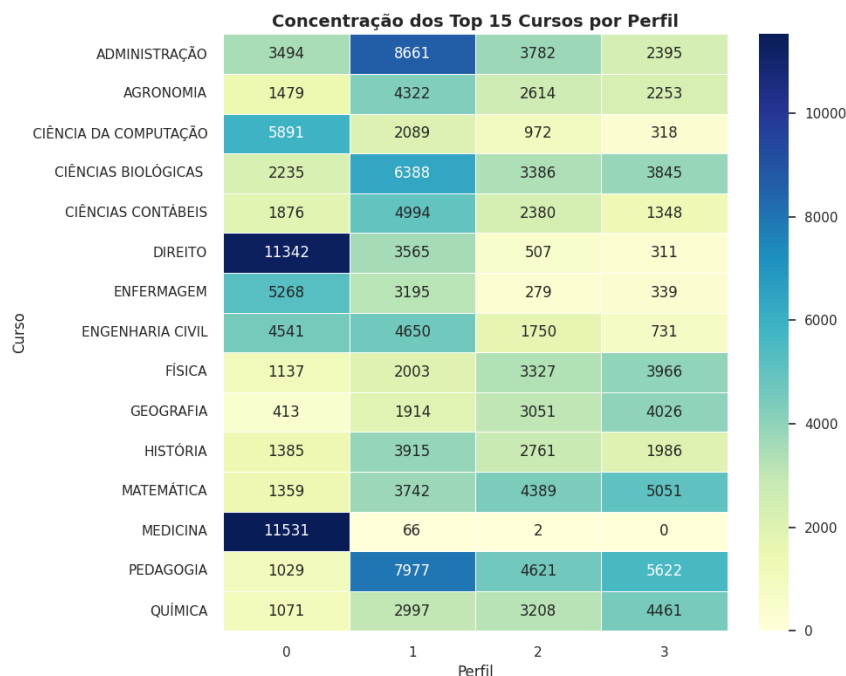


Figure 7. Distribuição das modalidades de concorrência por *Perfil* de desempenho acadêmico.

Table 2. Quantidade de ingressantes por perfil acadêmico nos 15 cursos com maior volume de registros.



Para segmentar os estudantes em perfis similares, aplicou-se o algoritmo K-Means tendo como variáveis de entrada as cinco notas do ENEM, definindo-se quatro *clusters*, cuja validação foi realizada via Análise de Componentes Principais (PCA) [Jolliffe 2002], indicando que os agrupamentos capturam estruturas distintas nos dados [Jain 2010, Hair Jr. et al. 2019]. A Figura 7 revela clara estratificação no perfil de

ingresso: o Perfil 0, de maiores médias, concentra predominantemente candidatos da ampla concorrência, enquanto o Perfil 3 agrupa proporcionalmente mais cotistas com menores médias de entrada, padrão consistente com a literatura que associa desigualdades de desempenho no ingresso ao hiato educacional acumulado na educação básica [Wainer and Melguizo 2018, Santos 2013], ainda que os dados aqui analisados não incluam variáveis diretas sobre a trajetória escolar prévia dos estudantes. Em contrapartida, as Licenciaturas concentram perfis de menor desempenho, o que encontra respaldo em estudos sobre a baixa atratividade da carreira docente e a desvalorização profissional e salarial [Gatti 2010, Louzano et al. 2010]. A Tabela 2 torna esse padrão mais concreto: em cursos como Física, Geografia, Matemática e Química, o Perfil 3 é isoladamente o grupo mais numeroso, respondendo por 35%–43% dos ingressantes, enquanto em Medicina e Direito o Perfil 0 concentra, respectivamente, 99% e 72% das vagas. Essa polarização é consistente com a hipótese de que cursos de formação de professores atraem um perfil de ingressante distinto do observado em carreiras de alta demanda e prestígio social.

4.3. Análise Espacial e Desigualdades Regionais

Com o intuito de mapear as disparidades regionais no ingresso ao ensino superior brasileiro, desenvolveu-se um índice de desempenho por estado calculado a partir da distribuição proporcional dos ingressantes em cada perfil. Para cada unidade federativa, computa-se a proporção de estudantes em cada perfil e multiplica-se pelo peso correspondente, conforme a Equação 2:

$$I_{UF} = \sum_{k=0}^3 p_{k,UF} \cdot w_k \quad (2)$$

onde $p_{k,UF}$ é a proporção de ingressantes do estado no perfil k e w_k é o peso atribuído ao perfil ($w_0 = 4$, $w_1 = 3$, $w_2 = 2$, $w_3 = 1$).

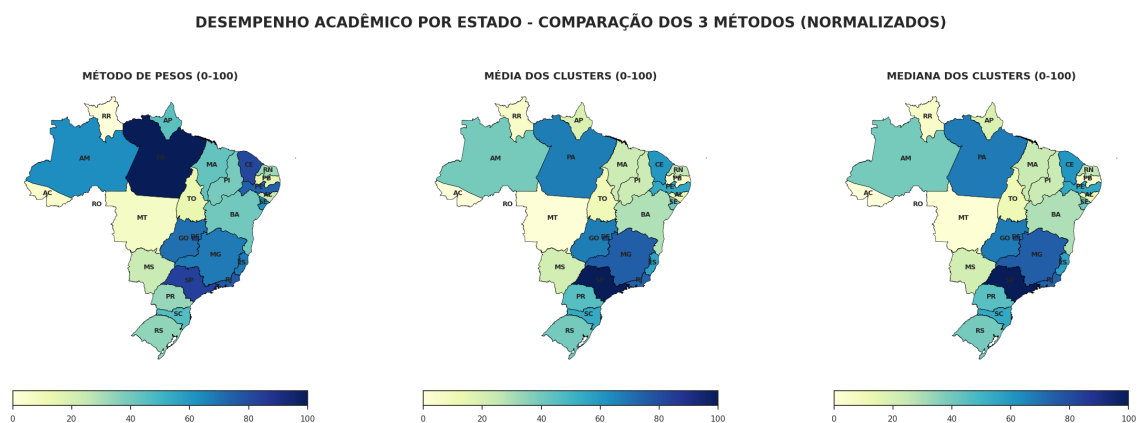


Figure 8. Distribuição regional do desempenho: índice de *Perfis* ponderado vs. médias absolutas por Unidade da Federação.

Para comparação, a Figura 8 apresenta esse índice normalizado (0–100) ao lado de dois métodos alternativos: a média e a mediana das notas brutas por estado, também normalizados. A análise revela dinâmicas geográficas que divergem das estatísticas

absolutas: sob a ótica do índice ponderado, estados como Pará, Ceará e São Paulo apresentam as maiores concentrações relativas de ingressantes nos perfis superiores, enquanto Roraima, Acre e Mato Grosso concentram proporcionalmente mais ingressantes nos perfis de menor desempenho. Ao confrontar essa métrica com médias e medianas absolutas, observa-se que as regiões Sul e Sudeste recuperam posições de destaque, sugerindo maior volume absoluto de notas elevadas, mas não necessariamente maior concentração proporcional nos perfis superiores [Soares and Delgado 2016, Menezes-Filho 2007].

4.4. Barreiras de Acesso e Impacto das Cotas

Os cursos com menores discrepâncias no Δ (Equação 1) apresentam variação estreita entre $-0,6$ e $0,5$; valores negativos indicam que, nesses casos, a nota de corte do cotista supera a da Ampla Concorrência. Predomina nesse grupo a subcategoria “EP + Independente de Renda”, padrão que a literatura associa à presença de estudantes de escolas públicas de excelência, como Institutos Federais e Colégios de Aplicação [Barros 2014, Campos et al. 2014]. Como a base não identifica o tipo específico de escola pública frequentada, essa leitura permanece como hipótese interpretativa. Em contrapartida, os cursos com as maiores disparidades de delta (Figura 9) são dominados pelas modalidades que conjugam escola pública e recorte étnico-racial (PPI: Pretos, Pardos e Indígenas), com destaque para “EP + PPI + Baixa Renda” e “EP + PPI + Independente de Renda”. Isso indica que a desvantagem observada transcende a dimensão socioeconômica, manifestando-se de forma independente mesmo quando controlada pela origem escolar, consistente com estudos que tratam o pertencimento étnico-racial como variável autônoma no acesso ao ensino superior [Almeida 2019, Santos 2013, Feres Júnior et al. 2018]. Isso sugere que políticas de recorte exclusivamente social tendem a capturar de forma incompleta as desigualdades impostas pela interseccionalidade de raça e classe, reforçando a necessidade de monitoramento contínuo dessas modalidades.

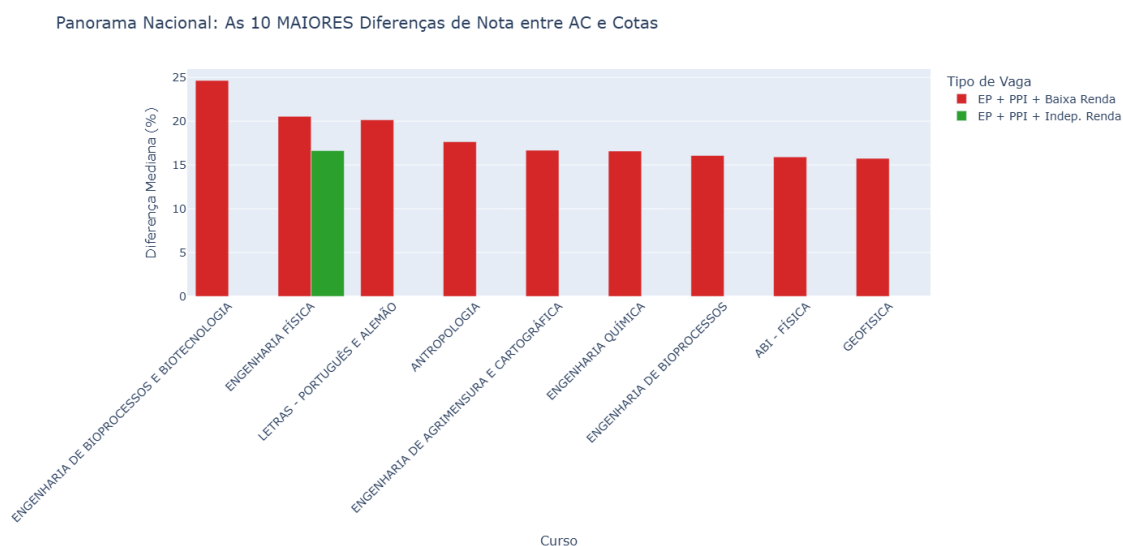


Figure 9. Cursos com as maiores diferenças de mediana (*delta*) entre AC e Cotas, com destaque para as modalidades PPI.

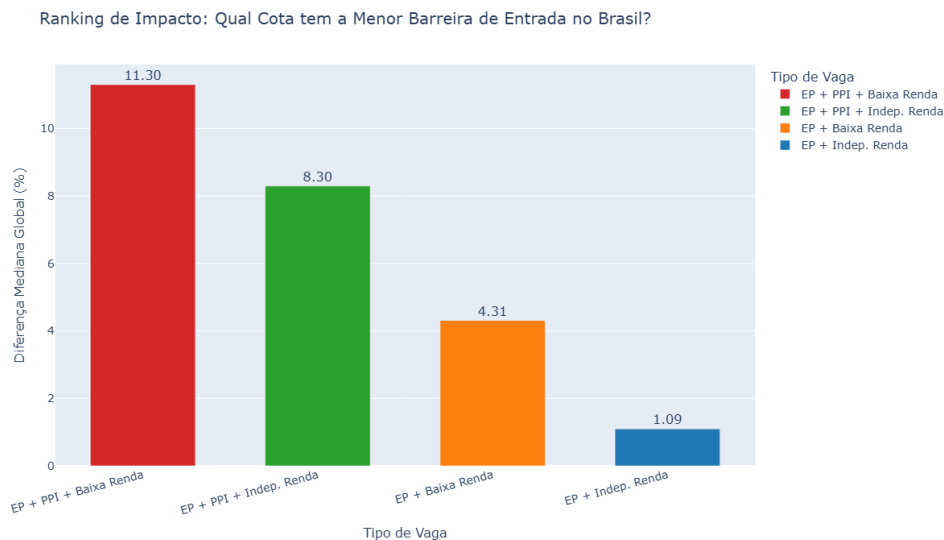


Figure 10. Ranking de impacto das modalidades de cotas, destacando a efetividade das reservas PPI.

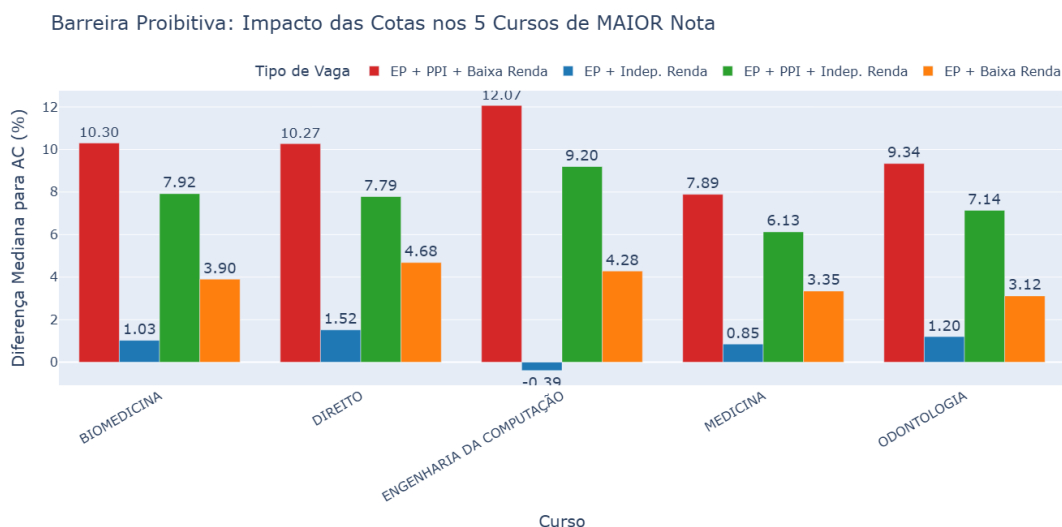


Figure 11. Impacto das cotas (*delta*) nos cinco cursos de maior seletividade acadêmica, evidenciando a disparidade nas modalidades PPI.

A análise do ranking de impacto das ações afirmativas (Figura 10) confirma que as modalidades PPI apresentam os maiores diferenciais de nota de corte frente à Ampla Concorrência, indicando que esses grupos enfrentam as maiores barreiras relativas de acesso à educação superior no Brasil. A literatura interpreta esse tipo de padrão como evidência da persistência de estruturas de exclusão e do racismo estrutural [Almeida 2019, Feres Júnior et al. 2018, Santos 2013]; contudo, cabe ressaltar que o desenho transversal deste estudo não permite estimar o que ocorreria na ausência dessas políticas, questão que exigiria análise longitudinal ou quase-experimental dedicada. Nos 5 cursos de maior média, a Figura 11 confirma que PPI + Baixa Renda atinge os maiores

diferenciais, evidenciando que a interseccionalidade de raça e classe constitui a barreira mais íngreme em carreiras de prestígio. Já “EP + Independente de Renda” registra impactos marginais ou *deltas* negativos, inclusive em Engenharia da Computação, sugerindo capital acadêmico convergente ao da Ampla Concorrência e questionando sua efetividade na inclusão dos grupos mais vulneráveis [Zoghbi et al. 2013, Feres Júnior et al. 2018].

5. Conclusão

A presente pesquisa demonstrou que avaliar o desempenho educacional apenas pela “média geral” esconde a verdadeira complexidade e as desigualdades do sistema. A aplicação do K-Means revelou quatro perfis distintos de ingressantes: os Perfis 0 e 1 concentram o alto desempenho acadêmico, enquanto os Perfis 2 e 3 apresentam os menores índices, padrão compatível com as lacunas estruturais herdadas da educação básica [Wainer and Melguizo 2018, Santos 2013]. Os dados indicam maior presença proporcional de cotistas em cursos de alta seletividade (Figura 6); contudo, a forte associação entre a Ampla Concorrência e o Perfil 0, em contraste com a concentração de cotistas nos Perfis 2 e 3 (Figura 7), sugere que as desigualdades de desempenho entre os grupos persistem, compatível com a leitura de que as ações afirmativas atuam sobre o acesso sem alterar as condições de formação nas etapas anteriores, o que reforça a relevância de políticas complementares voltadas à educação básica. A análise das modalidades de cotas confirma que as categorias com recorte étnico-racial (PPI) apresentam os maiores *deltas* de acesso (Figura 10), indicando que candidatos pretos, pardos e indígenas enfrentam as maiores diferenças relativas de nota de corte em relação à Ampla Concorrência, mesmo comparados a outros cotistas. Geograficamente, Pará, Ceará e Pernambuco destacam-se no índice ponderado de perfis (Figura 8), enquanto Roraima, Acre, Mato Grosso e Paraíba concentram proporcionalmente mais ingressantes nos perfis de menor desempenho, sugerindo a necessidade de investigações futuras sobre os fatores regionais associados. Em suma, a abordagem proposta integra Mineração de Dados Educacionais (K-Means) à análise espacial e macrosocial, oferecendo subsídios para a gestão pública com evidências baseadas em dados abertos. Os achados são limitados ao período analisado (2023–2024) e ao desenho transversal do estudo; o acompanhamento de trajetórias em ciclos futuros seria necessário para avaliar tendências de médio e longo prazo na equidade de acesso ao ensino superior brasileiro.

Cabe ressaltar, como limitação deste estudo, que a base compreende exclusivamente ingressantes aprovados via ENEM/SISU em IES públicas, não contemplando outras vias de acesso, como vestibulares próprios e processos seletivos específicos, que também podem adotar critérios de cotas distintos. Além disso, a base não representa a totalidade dos participantes do ENEM: como apontado por [Garcez Filho et al. 2025], a participação no exame caiu de 70,5% para 43,7% de registros válidos entre 2019 e 2020, em razão da pandemia de Covid-19, o que pode indicar sub-representação de segmentos mais vulneráveis já na etapa de inscrição. Consequentemente, os perfis e diferenciais de acesso aqui identificados refletem a realidade de quem consegue transpor as barreiras de acesso ao ENEM e ao SISU, e não a totalidade dos ingressantes no ensino superior público brasileiro nem da população concluinte do ensino médio.

Disponibilidade de Dados e Código

O código-fonte, os scripts de pré-processamento e a base de dados consolidada necessários para a reprodução dos resultados (clusterização e gráficos) estão disponíveis publicamente em: <https://doi.org/10.17605/OSF.IO/5U628>.

Declaração sobre Uso de Inteligência Artificial

Durante a elaboração deste artigo, o modelo de linguagem Claude Sonnet 4.6 (Anthropic) foi utilizado pelos autores como ferramenta de assistência tecnológica. O uso da ferramenta restringiu-se exclusivamente ao aprimoramento do estilo de escrita acadêmica, revisão ortográfica e gramatical, sugestões de sinonímia para ganho de fluidez no texto e auxílio na formatação de comandos da linguagem LaTeX (como estruturação de tabelas, equações e inserção de figuras). Ressalta-se que toda a concepção metodológica, a coleta e tratamento dos dados, a execução dos algoritmos de *Machine Learning* (como K-Means e PCA), a geração e interpretação dos gráficos, bem como o levantamento do referencial teórico e a validação das conclusões científicas, são de autoria e responsabilidade exclusiva dos pesquisadores envolvidos. Os autores assumem total e irrestrita responsabilidade pelo conteúdo final, integridade e veracidade dos dados apresentados neste trabalho.

References

- Almeida, S. (2019). *Racismo Estrutural*. Pólen Livros, São Paulo.
- Artes, A. and Ricoldi, A. M. (2015). Acesso de negros no ensino superior: o que mudou entre 2000 e 2010. *Cadernos de Pesquisa*, 45(158):858–881.
- Barros, A. S. X. (2014). Vestibular e cotas: o desempenho de alunos de escolas públicas e privadas. *Revista Brasileira de Estudos Pedagógicos*, 95(239):64–83.
- Bholowalia, P. and Kumar, A. (2014). EBK-means: A clustering technique based on elbow method and K-means in WSN. *International Journal of Computer Applications*, 105(9):17–24.
- Bourdieu, P. (1998). *Escritos de Educação*. Vozes, Petrópolis.
- Brasil (2023). Lei n.º 14.723, de 13 de novembro de 2023. Altera a Lei n.º 12.711/2012, para dispor sobre o programa especial para o acesso às instituições federais de educação superior. Diário Oficial da União.
- Campos, L. A., Feres Júnior, J., and Daflon, V. T. (2014). O desempenho dos cotistas no ENEM: comparando as notas de corte no SISU. Textos para Discussão GEMAA 4, GEMAA.
- Feres Júnior, J., Campos, L. A., Daflon, V. T., and Venturini, A. C. (2018). *Ação afirmativa: conceito, história e debates*. EdUERJ, Rio de Janeiro.
- Garcez Filho, A. C. L., Santos, S. P. d., Trevisan, M., Santos, L. A. d., and Kocourek, S. (2025). Influência da renda no desempenho dos estudantes no enem: Uma análise longitudinal. *Gestão e Desenvolvimento*, 22(2):202–225.
- Gatti, B. A. (2010). Formação de professores no Brasil: características e problemas. *Educação & Sociedade*, 31(113):1355–1379.

- Hair Jr., J. F., Black, W. C., Babin, B. J., and Anderson, R. E. (2019). *Multivariate Data Analysis*. Cengage Learning, 8 edition.
- Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) (2024). Microdados do exame nacional do ensino médio (enem). <https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/microdados>. Acesso em: 16 jul. 2026.
- Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8):651–666.
- Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer-Verlag New York, 2 edition.
- Louzano, P., Rocha, V., Moriconi, G. M., and Oliveira, R. P. d. (2010). Quem quer ser professor? Perfil dos concluintes de pedagogia e licenciaturas. *Estudos em Avaliação Educacional*, 21(47):543–568.
- Menezes-Filho, N. A. (2007). *Os determinantes do desempenho escolar do Brasil*. Instituto de Ensino e Pesquisa (Insper), São Paulo.
- Moura, E. G., Cunha, H. L., and Moraes, B. R. S. d. (2025). Aplicação de técnicas de aprendizado de máquina para classificação do desempenho acadêmico no enem: Um estudo com dados do maranhão. In *Knowledge Networks: Education as a Multidisciplinary Field*. Seven Publicações Acadêmicas.
- Oliveira, J. d. P. (2013). Acesso à educação superior pelo ENEM/SiSU: uma análise da implementação nas universidades sul-mato-grossenses. Dissertação (mestrado em educação), Universidade Federal da Grande Dourados, Dourados.
- Ristoff, D. (2014). O novo perfil do campus brasileiro: uma análise do perfil socioeconômico do estudante de graduação. *Avaliação: Revista da Avaliação da Educação Superior (Campinas)*, 19(3):723–747.
- Romero, C. and Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618.
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.
- Santos, J. T. d. (2013). *O impacto das cotas nas universidades brasileiras (2004–2012)*. Centro de Estudos Afro-Orientais (CEAO/UFBA), Salvador.
- Sen, A. (2000). *Desenvolvimento como Liberdade*. Companhia das Letras, São Paulo.
- Soares, J. F. and Delgado, V. M. S. (2016). Medida das desigualdades de aprendizado entre estudantes de ensino fundamental. *Estudos em Avaliação Educacional*, 27(66):754–780.
- Wainer, J. and Melguizo, T. (2018). Políticas de inclusão no ensino superior: avaliação do desempenho dos alunos baseado no Enade de 2012 a 2014. *Educação e Pesquisa*, 44:e162807.
- Zoghbi, A. C. P., Rocha, F., and Mattos, E. (2013). Education production efficiency: evidence from Brazilian universities. *Economic Modelling*, 31:94–103.