



UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
CURSO DE SISTEMAS DE INFORMAÇÃO

Vanessa Cunha

**Desempenho no SISU por Gênero em cursos de computação:
O que revelam os dados?**

Florianópolis
2025

Vanessa Cunha

**Desempenho no SISU por Gênero em cursos de computação:
O que revelam os dados?**

Trabalho de Conclusão de Curso de Sistemas de In-
formação da Universidade Federal de Santa Catarina.
Orientador: Prof. Simone Silmara Werner, Dra.
Coorientadora: Profa. Cristina Meinhardt, Dra.

Florianópolis
2025

AGRADECIMENTOS

Dedico este trabalho a todas as mulheres que fizeram parte da minha trajetória e que, de alguma forma, me apoiaram ao longo do caminho. Em especial, à minha mãe, Cleusa, e às minhas avós, Vanilda e Cleusa (*in memoriam*), que foram meus primeiros exemplos de força, afeto e resiliência. Foram elas que me ensinaram o valor da educação e da persistência diante dos desafios.

Agradeço também aos meus irmãos, Vanini, Gabriel e Bruna, com quem compartilho as melhores lembranças da infância e da vida adulta. O carinho, o apoio e o incentivo de cada um de vocês foram fundamentais para que eu pudesse chegar até aqui.

Às professoras Dra. Simone Werner e Dra. Cristina Meinhardt, registro minha profunda gratidão pela orientação neste trabalho, pela paciência em ensinar e pela confiança constante. Em muitos momentos, vocês acreditaram mais em mim do que eu mesma, e isso fez toda a diferença.

Sou agradecida por ter encontrado meu companheiro Felipe e os queridos amigos Paulo e Gabriel, que tornaram a caminhada mais leve e feliz. Estendo também meu agradecimento aos colegas de curso, com quem tive o prazer de aprender, trocar ideias e compartilhar projetos ao longo da graduação.

São muitas as pessoas incríveis e especiais que cruzaram meu caminho e contribuíram, de diferentes formas, para que eu chegasse até aqui. Seria injusto tentar nomear todas, sob o risco de esquecer alguém. Quero que saibam, porém, que os anos dedicados a este curso e à minha formação foram alguns dos mais marcantes da minha vida, um período de autoconhecimento, de crescimento e de conexão com o meu propósito. A todas e todos que fizeram parte dessa jornada, o meu mais sincero muito obrigada.

*“A frase mais perigosa da língua é
‘sempre fizemos assim’”*

Grace Hopper, 1970

RESUMO

Ao longo da existência dos cursos de computação da UFSC, observou-se uma baixa representatividade feminina. Esse fator motivou esta pesquisa, cujo propósito foi utilizar metodologias de análise de dados para compreender padrões relacionados ao gênero e ingresso em cursos de computação em escala nacional. Utilizando dados do Sistema de Seleção Unificada (SISU) de 2018 a 2023, o trabalho examinou o perfil de candidatos e aprovados nesses cursos, além de investigar os padrões de migração entre estados e o impacto de políticas afirmativas no ingresso de mulheres. O estudo explorou os perfis tanto dos candidatos aprovados quanto daqueles que não obtiveram acesso, destacando variáveis como gênero, região de origem e os efeitos da migração na busca por ingresso no ensino superior. Os resultados revelaram um panorama ainda predominantemente masculino nos cursos de Ciência da Computação, Engenharia de Computação e Sistemas de Informação, mas com sinais consistentes de ampliação da presença feminina. A proporção de mulheres entre os candidatos cresceu de 21,65% para 24,87% no período analisado, representando um crescimento de 3,22 pontos percentuais, enquanto entre os aprovados o aumento foi ainda mais pronunciado, saltando de 16,62% para 21,45% com crescimento de 4,83 pontos percentuais. A análise das políticas afirmativas demonstrou que o sistema de cotas consolidou-se como principal mecanismo de diversidade, representando 55,38% das aprovações contra 44,62% da ampla concorrência, com taxa de conversão superior de 9,42% comparada aos 9,07% da ampla concorrência e maior proporção feminina de 19,17% contra 16,77% na ampla concorrência. Quanto à mobilidade estudantil, identificou-se que 23,6% dos aprovados estudam fora do estado de origem, com mulheres apresentando taxa ligeiramente superior de 25,0% comparada aos 23,3% dos homens, diferença de 1,6 pontos percentuais. O desempenho no ENEM confirmou diferenças de gênero consistentes mas moderadas, com hiato médio de 27,6 pontos favorável aos homens, apresentando variação temporal entre 18,4 pontos em 2023 e 33,9 pontos em 2019, com tendência de redução de 13,3 pontos entre 2018 e 2023. A análise regional evidenciou disparidades estruturais marcantes, com as notas mais altas concentradas no Sudeste com média de 597,7 pontos e Sul com 587,4 pontos, enquanto o Norte registrou 564,0 pontos e o Nordeste 569,9 pontos. O gap de gênero mostrou-se mais acentuado em regiões de menor IDH com diferença de 25,4 pontos comparado aos 21,3 pontos em regiões de IDH alto. O estudo forneceu uma visão atualizada sobre a participação feminina nesses cursos, contribuindo para comparações regionais e evidenciando que as desigualdades observadas são fundamentalmente estruturais, relacionadas à qualidade diferenciada da educação básica e à distribuição desigual das oportunidades educacionais entre regiões do país. Os achados subsidiam o desenvolvimento de políticas que promovam a diversidade e incentivem a representatividade de gênero, com especial atenção às disparidades regionais identificadas na análise.

Palavras-chave: Mulheres. Tecnologia. Ciências da Computação. Sistemas de Informação. Engenharia de Computação. Sistema de Seleção Unificada. Análise de dados.

ABSTRACT

Throughout the history of UFSC's computing programs, a persistent underrepresentation of women has been observed. This factor motivated the present research, which aimed to apply data analysis methodologies to understand patterns related to gender and admission into computing courses on a national scale. Using data from Brazil's Unified Selection System (SISU) between 2018 and 2023, the study examined the profile of applicants and admitted students in these programs, as well as patterns of interstate migration and the impact of affirmative action policies on women's access to higher education. The analysis included both successful and unsuccessful applicants, highlighting variables such as gender, region of origin, and the effects of migration on access to higher education. The results revealed a landscape that remains predominantly male in Computer Science, Computer Engineering, and Information Systems programs, though with consistent signs of increasing female participation. The proportion of women among applicants rose from 21.65% to 24.87% during the analyzed period, representing a growth of 3.22 percentage points, while among admitted students the increase was even more pronounced, from 16.62% to 21.45%, a growth of 4.83 percentage points. The analysis of affirmative action policies showed that the quota system has become the main mechanism for promoting diversity, accounting for 55.38% of admissions compared to 44.62% through general competition, with a higher conversion rate (9.42%) compared to the general category (9.07%) and a greater proportion of women (19.17% versus 16.77% in the general competition). Regarding student mobility, the study identified that 23.6% of admitted students enroll outside their home state, with women presenting a slightly higher rate (25.0%) than men (23.3%), a 1.6-point difference. ENEM performance confirmed consistent yet moderate gender differences, with an average gap of 27.6 points favoring men, ranging from 18.4 points in 2023 to 33.9 points in 2019, and showing a reduction of 13.3 points over the 2018–2023 period. Regional analysis revealed notable structural disparities, with the highest scores concentrated in the Southeast (average 597.7 points) and South (587.4 points), while the North (564.0 points) and Northeast (569.9 points) scored lower. The gender gap was more pronounced in regions with lower Human Development Index (HDI), averaging 25.4 points, compared to 21.3 points in higher-HDI regions. This study provides an updated overview of women's participation in computing programs, contributing to regional comparisons and showing that the observed inequalities are fundamentally structural—linked to differences in the quality of basic education and the unequal distribution of educational opportunities across the country. The findings offer valuable insights for developing policies that promote diversity and enhance gender representation, with special attention to the regional disparities identified in the analysis.

Keywords: Women. Technology. Computer Science. Information Systems. Computer Engineering. Unified Selection System. Data Analysis.

LISTA DE FIGURAS

Figura 1 – O processo de KDD	30
Figura 2 – O processo de coleta de dados	34
Figura 3 – Esquema visual dos eixos dos resultados	52
Figura 4 – Distribuição percentual de candidatos por sexo	53
Figura 5 – Evolução percentual do uso de Cotas por sexo no período de 2018-2023 em Cursos de Computação	54
Figura 6 – Proporção acumulada do período de candidatos por sexo região de origem	55
Figura 7 – Distribuição acumulada de candidatos por faixa de IDH da região 2018- 2023	56
Figura 8 – Taxa acumulada de conversão por ano	58
Figura 9 – Taxa acumulada de conversão por região e sexo no período analisado .	59
Figura 10 – Taxa acumulada de conversão por sexo no período de 2018-2023	60
Figura 11 – Taxa de conversão acumulada do período por ano, região e sexo	61
Figura 12 – Relação anual entre a forma de ingresso e sexo	62
Figura 13 – Relação anual entre a forma de ingresso e sexo	64
Figura 14 – Evolução de contistas aprovados anual por sexo pelo período 2018-2023	65
Figura 15 – Proporção acumulada do período de cotistas por região	66
Figura 16 – Variação da migração anual dos alunos de computação por sexo	68
Figura 17 – Sankey esquema migratório entre regiões acumulado pelo período de 2018-2023	69
Figura 18 – Relação entre as variáveis de mobilidade e sexo	70
Figura 19 – Relação entre as variáveis de gênero, mobilidade e forma de ingresso . .	71
Figura 20 – Proporção acumulada pelo período de migração entre estados com e sem oferta de curso de computação estratificado por sexo	72
Figura 21 – Variação das candidaturas interestaduais por gênero	75
Figura 22 – Distância média das migrações acumuladas pelo período de 2018-2023 de aprovados em cursos de computação	76
Figura 23 – Preferencias regionais dos candidatos acumulado pelo período de 2018- 2023	77
Figura 24 – Saldo migratório Feminino acumulado pelo período de 2018-2023 . . .	78
Figura 25 – Variação das notas por sexo	81
Figura 26 – Evolução da variação do GAP das notas F-M	82
Figura 27 – Variação das notas considerando recortes por sexo e região de origem .	83
Figura 28 – Variação das notas anual por sexo e curso	83
Figura 29 – Importância de variáveis para as Notas do ENEM	84
Figura 30 – Notas ENEM médias pro região	85
Figura 31 – Distância das notas dos não aprovados	86

LISTA DE TABELAS

Tabela 1 – Comparação da abordagem proposta com as de trabalhos correlatos . .	28
Tabela 2 – Dicionário de dados do recorte SISU (2018.1–2023.2)	35
Tabela 3 – Resumo dos testes inferenciais para o perfil dos candidatos alvo	57
Tabela 4 – Resumo dos testes inferenciais para o perfil dos aprovados	62
Tabela 5 – Resumo dos testes inferenciais para políticas públicas	67
Tabela 6 – Resumo dos testes inferenciais para mobilidade estudantil	74
Tabela 7 – Resumo dos testes inferenciais para diferenças de gênero na mobilidade acadêmica	79
Tabela 8 – Resumo dos testes inferenciais para diferenças de gênero no desempenho	87

SUMÁRIO

1	INTRODUÇÃO	10
1.1	JUSTIFICATIVA	11
1.2	OBJETIVOS	12
1.3	ORGANIZAÇÃO DO TEXTO	13
2	FUNDAMENTAÇÃO TEÓRICA	14
2.1	DISPARIDADE DE GÊNERO NOS CURSOS DE COMPUTAÇÃO	14
2.2	POLÍTICAS AFIRMATIVAS E INCLUSÃO DE MULHERES	15
2.3	O SISTEMA DE SELEÇÃO UNIFICADA (SISU) E A DEMOCRATIZAÇÃO DO ACESSO	15
2.4	MOBILIDADE ESTUDANTIL E FLUXOS MIGRATÓRIOS	16
2.5	A INTERSEÇÃO ENTRE GÊNERO E MOBILIDADE ACADÊMICA	18
2.6	FERRAMENTAL DE ANÁLISE, VISUALIZAÇÃO E INFERÊNCIA	18
3	TRABALHOS CORRELATOS	23
3.1	PARIDADE DE GÊNERO NO ENSINO SUPERIOR EM STEM NO BRASIL: UMA ANÁLISE DE 10 ANOS	23
3.2	FATORES ASSOCIADOS À MOBILIDADE DE ESTUDANTES UNIVERSITÁRIOS NO BRASIL: UMA ANÁLISE POR MEIO DE MODELO GRAVITACIONAL	24
3.3	ANÁLISE DE PERFIL, DESEMPENHO E PROVAS DO ENADE DO CURSO DE SISTEMAS DE INFORMAÇÃO DA UFSC	24
3.4	WAGE PROFILE AND GENDER GAP IN SCIENCE AND TECHNOLOGY: REGIONAL DISPARITIES IN BRAZIL	25
3.5	WOMEN'S JOURNEY IN STEM EDUCATION IN BRAZIL: A RAPID REVIEW ON ENGINEERING AND COMPUTER SCIENCE	25
3.6	GENDER EQUALITY POLICIES IN STEM IN LATIN AMERICA – A SYSTEMATIC LITERATURE REVIEW	26
3.7	GENDER DISPARITY IN ENGINEERING COURSES IN BRAZIL	26
3.8	COMPARAÇÃO COM TRABALHOS CORRELATOS	27
4	MÉTODOS DE PESQUISA	29
4.1	FERRAMENTAS	30
4.2	COLETA E SELEÇÃO DOS DADOS	33
4.2.1	Estrutura dos dados	35
4.2.2	Seleção dos dados	39
4.3	PRÉ-PROCESSAMENTO DOS DADOS	40
4.4	TRANSFORMAÇÃO DOS DADOS	44
4.5	MINERAÇÃO DOS DADOS	46
4.5.1	Descrição e Diagnóstico	47

4.5.2	Explicação e Causalidade	48
4.5.3	Previsão e descoberta de padrões	50
5	RESULTADOS E DISCUSSÃO	52
5.1	PERFIL DOS CANDIDATOS	53
5.2	PERFIL DOS APROVADOS	57
5.3	EFEITOS DAS POLÍTICAS AFIRMATIVAS	63
5.4	MOBILIDADE ESTUDANTIL	67
5.4.1	Mobilidade e padrões de interesse por gênero	75
5.5	DESEMPENHO NA PROVA DO ENEM	79
6	CONSIDERAÇÕES FINAIS	88
6.1	TRABALHOS FUTUROS	92
	REFERÊNCIAS	93
	ANEXO A – ARTIGO DESIGUALDADES DE GÊNERO E MOBILIDADE NO INGRESSO EM CURSOS DE COMPUTAÇÃO NO BRASIL: UMA ANÁLISE DOS DADOS DO SISU	100
	ANEXOS	120

1 INTRODUÇÃO

A disparidade de gênero nos cursos de tecnologia no Brasil tem motivado um crescente volume de estudos, tanto nacionais quanto internacionais, que abordam a inclusão e a representatividade feminina nas áreas de computação (UNESCO, 2017; SANTOS, 2018). Existem evidências de viés institucional em contextos acadêmicos indicando que mulheres frequentemente recebem avaliações menos favoráveis, têm menor probabilidade de receber ofertas de contratação e mentoria e recebem recomendações salariais inferiores, fenômenos que contribuem para a fuga de talentos e para a manutenção da subrepresentação feminina na ciência (MOSS-RACUSIN *et al.*, 2012). Essa desigualdade preocupa não apenas pelo desequilíbrio numérico, mas por seus impactos sociais, educacionais e econômicos, uma vez que a baixa participação de mulheres compromete a diversidade de perspectivas no desenvolvimento tecnológico, reduz a capacidade de inovação nas equipes e limita a equidade de oportunidades no mercado de trabalho, efeitos documentados em estudos sobre diversidade e desempenho organizacional e inovação e em revisões globais como o Global Gender Gap Report (HERRING, 2009; ØSTERGAARD; TIMMERMANS; KRISTINSSON, 2011; WORLD ECONOMIC FORUM, 2020).

Dados históricos revelam que, embora cursos como o de Ciência da Computação da USP tivessem uma maioria feminina nos anos 1970, a presença das mulheres foi reduzida drasticamente nas décadas seguintes. Esse fenômeno reflete padrões amplos de exclusão nas áreas STEM (Ciência, Tecnologia, Engenharia e Matemática), ainda mais evidentes em países com estruturas acadêmicas e de mercado que impõem barreiras específicas à participação feminina (UNESCO, 2017; SANTOS, 2018).

Estudos também apontam que políticas afirmativas e programas de incentivo podem contribuir para reduzir essas disparidades; revisões sobre ações afirmativas mostram efeitos variados, mas indicam que medidas de inclusão, quando bem desenhadas e implementadas, têm potencial para ampliar o acesso e a permanência de grupos sub-representados (BASSO-POLETTI; EFROM; BEATRIZ-RODRIGUES, 2020). A Lei de Cotas (BRASIL, 2012), por exemplo, ampliou o acesso ao ensino superior por meio de políticas de inclusão racial e social, e tem repercussões sobre o ingresso de mulheres em cursos historicamente masculinos. Da mesma forma, o Sistema de Seleção Unificada (SISU) tem contribuído para a democratização do acesso a universidades públicas. No entanto, embora o SISU represente uma plataforma unificada de admissão, seus efeitos sobre a diversidade de gênero nos cursos de computação ainda são pouco compreendidos (INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP), 2024). A partir desse quadro, o presente estudo busca responder questões específicas sobre caráter e dinâmica da representação feminina e do desempenho acadêmico nos cursos de computação no período observado, de 2018 até 2023, pelo estudo. Em particular, as perguntas de pesquisa que guiam este trabalho são:

- Q1 - Como se caracteriza o perfil dos candidatos e aprovados aos cursos de Ciência da Computação, Engenharia da Computação e Sistemas de Informação em termos de gênero, região de origem e modalidade de ingresso no período de 2018 a 2023?
- Q2 - Existem diferenças significativas no desempenho médio do ENEM entre candidatos e candidatas a esses cursos, e como tais diferenças se distribuem entre as edições do SISU e entre os cursos analisados?;
- Q3 - Qual é a distribuição e o comportamento do desempenho dos candidatos em cada curso, considerando o efeito do curso e da modalidade de concorrência sobre as notas médias?;
- Q4 - Que padrões de mobilidade geográfica se observam entre aprovados, isto é, de que forma a migração entre estados (origem $\times$ destino) se relaciona com a representação feminina nos cursos de computação?;
- Q5 - Como evoluiu, nas edições de 2018 a 2023, a proporção de mulheres inscritas e aprovadas nesses cursos e se é possível identificar tendências estatisticamente significativas de crescimento ou retração dessa participação?

1.1 JUSTIFICATIVA

A crescente demanda por profissionais qualificados na área de tecnologia contrasta com a persistente desigualdade de gênero nos cursos de graduação em computação no Brasil. A sub-representação feminina em cursos como Ciência da Computação, Sistemas de Informação e Engenharia de Computação constitui um desafio social, acadêmico e político, uma vez que compromete a diversidade, a inovação e a equidade nas áreas de Ciência, Tecnologia, Engenharia e Matemática (STEM) (UNESCO, 2017). Esse desequilíbrio não apenas reflete padrões históricos de exclusão, mas também perpetua barreiras estruturais que afetam o acesso, a permanência e o sucesso de mulheres nesses campos (SANTOS, 2018).

Nesse contexto, políticas públicas como a Lei de Cotas (BRASIL, 2012) e programas de democratização do acesso, como o Sistema de Seleção Unificada (SISU), têm buscado promover maior equidade no ensino superior (BRASIL, 2012). No entanto, os impactos concretos dessas iniciativas sobre a diversidade de gênero nos cursos de computação ainda carecem de investigação mais aprofundada, especialmente quando considerados fatores regionais e os fluxos migratórios estudantis no país (INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP), 2024).

Diante disso, este trabalho se justifica pela necessidade de compreender como variáveis como gênero, região de origem e mobilidade acadêmica influenciam o ingresso e o desempenho de estudantes em cursos da área de computação. A análise dos dados do ENEM e do SISU permite não apenas traçar o perfil dos candidatos e aprovados ao

longo do tempo, como também avaliar o alcance das políticas afirmativas na promoção de inclusão em cursos historicamente marcados pela hegemonia masculina (SELL, 2022).

Ao investigar a intersecção entre gênero, desempenho e mobilidade acadêmica, este estudo pretende preencher lacunas existentes na literatura nacional, oferecendo evidências empíricas que possam subsidiar a formulação de políticas públicas e estratégias institucionais voltadas à promoção de maior diversidade e inclusão. Espera-se, assim, contribuir para o fortalecimento de um ambiente acadêmico mais equitativo, plural e sensível às diferentes realidades dos estudantes brasileiros.

Em adição às razões expostas, destaca-se que a sub-representação feminina nas áreas de computação tem implicações que vão além do aspecto numérico: a literatura indica que equipes mais diversas tendem a gerar mais perspectivas complementares e maior capacidade de inovação, implicando efeitos positivos sobre o desempenho organizacional e a produção tecnológica (HERRING, 2009) (ØSTERGAARD; TIMMERMANS; KRISTINSON, 2011). Ao mesmo tempo, estudos demonstram a presença de vieses institucionais que podem limitar oportunidades e reproduzir desigualdades dentro do meio acadêmico e profissional (MOSS-RACUSIN *et al.*, 2012), enquadrando a questão não apenas como problema de acesso, mas também de processos e práticas institucionais. Relatórios internacionais evidenciam a dimensão estrutural do problema e a necessidade de políticas integradas para reduzir lacunas de gênero no mercado de trabalho e na educação em STEM (WORLD ECONOMIC FORUM, 2020). Com base nisso, este estudo busca não apenas descrever taxas e tendências de ingresso, mas também medir quanto da variação de desempenho se explica por fatores institucionais, identificar padrões complexos e não lineares associados às notas e à mobilidade e avaliar o papel das políticas afirmativas na promoção de inclusão (BASSO-POLETTI; EFROM; BEATRIZ-RODRIGUES, 2020). Essa combinação metodológica pretende fornecer evidências acionáveis para formulação de políticas e intervenções institucionais que promovam maior diversidade e equidade no campo tecnológico.

1.2 OBJETIVOS

Dentro deste contexto, este Trabalho de Conclusão de Curso propõe-se a investigar o desempenho e os padrões de ingresso de candidatos no Sistema de Seleção Unificada (SISU) em cursos de computação no Brasil, com ênfase na relação entre gênero e fluxos migratórios. O estudo busca identificar barreiras e oportunidades para ampliar a representatividade feminina e promover a diversidade geográfica nesses cursos.

Ao analisar o desempenho acadêmico e os fluxos migratórios dos estudantes, espera-se compreender melhor as dinâmicas que influenciam a escolha por cursos de tecnologia e como fatores regionais afetam as taxas de ingresso e permanência das mulheres nessas áreas. A pesquisa também visa revelar possíveis barreiras à inclusão feminina no ensino superior em computação e fornecer subsídios para o desenvolvimento de políticas públicas

e práticas institucionais que promovam ambientes acadêmicos mais diversos e inclusivos.

Objetivos Específicos

- Caracterizar o perfil dos candidatos e aprovados em cursos de Ciência da Computação, Sistemas de Informação e Engenharia de Computação, com foco na distribuição por gênero, região de origem e forma de ingresso.
- Avaliar os efeitos das políticas afirmativas, em especial a Lei de Cotas, sobre a participação de mulheres nesses cursos, considerando variações regionais.
- Analisar os padrões de migração interestadual entre candidatos, investigando como a mobilidade acadêmica influencia a diversidade regional e de gênero nos cursos de computação.
- Comparar o desempenho acadêmico de estudantes com diferentes perfis demográficos e regionais, identificando disparidades estatísticas e possíveis fatores associados às desigualdades observadas.

Além disso, o estudo busca fornecer subsídios para o desenvolvimento de políticas públicas e estratégias institucionais que promovam maior equidade de gênero e diversidade regional nos cursos de computação, contribuindo para ambientes acadêmicos mais inclusivos e representativos.

1.3 ORGANIZAÇÃO DO TEXTO

Este trabalho está estruturado em seis capítulos, além dos elementos pré e pós-textuais. O Capítulo 1 introduz o tema, os objetivos, a justificativa e a estrutura do trabalho. O Capítulo 2 apresenta a fundamentação teórica, abordando desigualdade de gênero nos cursos de computação, políticas afirmativas, SISU, mobilidade estudantil e suas interseções. O Capítulo 3 revisa trabalhos correlatos, destacando pesquisas semelhantes e lacunas existentes. No Capítulo 4, descreve-se a metodologia adotada, incluindo fontes, critérios e estratégias analíticas. O Capítulo 5 apresenta os resultados e os discute à luz da literatura. Por fim, o Capítulo 6 traz as conclusões, contribuições do estudo, limitações e possibilidades para pesquisas futuras, seguido das referências bibliográficas.

2 FUNDAMENTAÇÃO TEÓRICA

A fundamentação teórica deste trabalho tem como objetivo contextualizar e embasar, a partir da literatura científica, os principais conceitos, debates e evidências que orientam a presente pesquisa. A análise da desigualdade de gênero nos cursos de computação no Brasil demanda uma abordagem multidisciplinar, envolvendo estudos nas áreas de educação, políticas públicas, mobilidade estudantil e gênero em STEM (Ciência, Tecnologia, Engenharia e Matemática).

Ao longo desta seção, são apresentados os principais marcos teóricos e empíricos relacionados à participação das mulheres em cursos de tecnologia, às políticas afirmativas implementadas no ensino superior brasileiro, à estrutura do Sistema de Seleção Unificada (SISU) e à mobilidade acadêmica entre regiões do país. Esses elementos fornecem as bases conceituais necessárias para compreender os fatores que influenciam o ingresso, a permanência e o desempenho de estudantes, especialmente mulheres, nos cursos de computação, além de permitirem a análise crítica das desigualdades ainda persistentes nesse contexto.

2.1 DISPARIDADE DE GÊNERO NOS CURSOS DE COMPUTAÇÃO

A desigualdade de gênero nas áreas de Ciência, Tecnologia, Engenharia e Matemática (STEM) é amplamente documentada e permanece um desafio estrutural para a formação de profissionais em computação. Nas últimas décadas observou-se uma crescente preocupação com a sub-representação feminina nesses cursos, que não se reduz ao mero desequilíbrio numérico, mas envolve processos históricos, culturais e institucionais que desincentivam a entrada e a permanência de mulheres na área (SOUZA, T. P. d., 2017).

Estudos de caso brasileiros mostram trajetórias semelhantes: por exemplo, análises históricas da USP indicam que cursos como Ciência da Computação chegaram a ter maioria feminina nos anos 1970, mas sofreram queda acentuada nas décadas seguintes em função de uma cultura acadêmica marcada por vieses de gênero e pela ausência de políticas de suporte à permanência feminina (SANTOS, 2018). Pesquisas em outras instituições confirmam padrões correlatos de evasão e exclusão, com perda de alunas já nos primeiros semestres e fragilidade de mecanismos institucionais de acolhimento e mentoria (SILVA; OLIVEIRA; SILVA, 2019); (OLIVEIRA-SILVA; PARREIRA, 2022), (SILVA, 2019).

Intervenções locais mostram, contudo, que ações de extensão e políticas institucionais bem desenhadas podem mitigar essas desigualdades: estudos de avaliação de projetos de extensão em Santa Maria apontam impacto positivo na mobilização e no interesse de alunas do ensino público pelas áreas de computação e exatas (ROCHA *et al.*, 2022), enquanto levantamentos recentes sintetizam desafios e práticas promissoras para inclusão em STEM no contexto brasileiro (SANTOS; OMENA MESSIAS, 2025); (POUSA *et al.*, 2023). Em conjunto, a evidência indica que superar a sub-representação feminina em com-

putação exige não só iniciativas de acesso, mas também políticas contínuas de retenção, mudanças culturais nos cursos e ações de apoio acadêmico e social que enfrentem vieses e barreiras estruturais.

2.2 POLÍTICAS AFIRMATIVAS E INCLUSÃO DE MULHERES

As políticas afirmativas, notadamente a Lei nº 12.711/2012 (BRASIL, 2012), representam um marco institucional para ampliar o acesso de grupos historicamente excluídos ao ensino superior no Brasil. Essas medidas foram concebidas para corrigir desigualdades estruturais e promover maior diversidade socioeconômica e racial entre os ingressantes, com potenciais efeitos indiretos sobre a presença feminina em áreas tradicionalmente dominadas por homens (LOPES *et al.*, 2023). Em nível internacional, avaliações e relatórios têm sinalizado que a simples ampliação do acesso não é suficiente para eliminar a sub-representação feminina em STEM: fatores culturais, organizacionais e de mercado continuam a moldar a distribuição por sexo nos cursos de tecnologia (UNESCO, 2017); (WORLD ECONOMIC FORUM, 2020).

Existe também uma lacuna de pesquisa relevante: faltam estudos robustos e desagregados que avaliem especificamente o efeito das políticas afirmativas sobre o ingresso, a permanência e o desempenho de mulheres em cursos de computação. A maior parte das avaliações agregadas foca em recortes racial e socioeconômico, sem análise detalhada por gênero e por área de formação; tampouco são frequentes investigações longitudinais que acompanhem trajetórias acadêmicas de alunas cotistas em cursos de tecnologia. Sem essa evidência específica, torna-se difícil aferir até que ponto as cotas e políticas complementares (programas de mentoria, bolsas de permanência, ações de acolhimento) estão efetivamente reduzindo a sub-representação feminina nos cursos de computação ou apenas alterando perfis de ingresso.

Dessa forma, a avaliação das políticas afirmativas no âmbito da computação exige (i) análises desagregadas por gênero e interseccionalidade (raça, renda, região), (ii) estudos longitudinais que mensurem evasão e desempenho ao longo do curso, e (iii) avaliações das políticas complementares de permanência e acolhimento que possam atuar em sinergia com as cotas. A produção de evidências nesses três vetores é imprescindível para orientar intervenções institucionais e políticas públicas que não se limitem ao ingresso, mas promovam também a formação, a inclusão e a progressão acadêmica das mulheres em áreas tecnológicas.

2.3 O SISTEMA DE SELEÇÃO UNIFICADA (SISU) E A DEMOCRATIZAÇÃO DO ACESSO

O Sistema de Seleção Unificada (SISU) foi criado em 2010 como plataforma oficial para concorrer a vagas em instituições públicas de ensino superior a partir das notas do ENEM. Até 2023 o SISU costumava realizar duas edições por ano, nos semestres leti-

vos, e a partir de 2024 passou a concentrar a seleção em uma edição anual realizada em janeiro; considerando esse percurso, o sistema soma cerca de 30 edições até 2025. Sua proposta central é unificar e tornar mais transparente o processo seletivo, permitindo que um único resultado do ENEM concorra a múltiplas vagas e instituições (MINISTÉRIO DA EDUCAÇÃO, 2024). Ao centralizar as inscrições, o SISU simplificou o acesso às vagas públicas e ampliou a visibilidade de ofertas em diferentes regiões do país, o que, em tese, favorece a mobilidade estudantil inter-regional e a participação de candidatos de perfis socioeconômicos diversos. Essa integração representou um avanço institucional importante na democratização do ingresso no ensino superior público, ao reduzir barreiras administrativas e facilitar a conciliação entre desempenho no exame e múltiplas oportunidades de matrícula. (MINISTÉRIO DA EDUCAÇÃO DO BRASIL (MEC), 2024). No entanto, sua efetividade na promoção da diversidade nos cursos de computação ainda é objeto de questionamento.

Em termos de impacto, estudos e avaliações mostram resultados mistos: por um lado, o SISU ampliou a possibilidade de acesso para estudantes de escolas públicas e de regiões menos atendidas, especialmente quando suas notas são combinadas com políticas de ação afirmativa; por outro, a disseminação do acesso não se traduz automaticamente em igualdade nas trajetórias acadêmicas. Pesquisas qualitativas e quantitativas destacam que o SISU mudou o perfil dos ingressantes e facilitou a ocupação de vagas por candidatos de baixa renda, mas que persistem desigualdades prévias de preparação, bem como barreiras econômicas e geográficas que afetam a permanência (SENKEVICS, 2024; SOUZA, T. L. de, 2017).

Especificamente quanto à diversidade interna dos cursos, por exemplo, a maior presença de mulheres em cursos de computação, a literatura indica limitações do SISU como única ferramenta de democratização. A entrada ampliada é necessária, mas insuficiente: fatores de permanência (moradia, bolsa, redes de apoio), condições de acolhimento nos departamentos e desigualdades estruturais prévias continuam a condicionar quem efetivamente conclui os cursos (SANTOS, 2018; PELEGRINI; SÁ; FRANÇA, 2018; CASTILHO, 2024). Por isso, autores que analisam a democratização do acesso recomendam combinar políticas de acesso centralizado (SISU/ENEM) com programas de acompanhamento, apoio acadêmico e medidas que considerem a interseccionalidade (raça, gênero, renda) para que o aumento das matrículas se converta em avanços reais na representatividade dentro de cada área do conhecimento (VERAS; LUZ SILVA, 2020; SOUZA, T. P. d., 2017).

2.4 MOBILIDADE ESTUDANTIL E FLUXOS MIGRATÓRIOS

Fluxos migratórios designam os movimentos de pessoas entre regiões geográficas e são analisados em termos de intensidade, direção e causas subjacentes. Esses deslocamentos emergem de um conjunto de determinantes econômicos, sociais e espaciais: diferenças relativas em oportunidades de trabalho e educação, distribuição desigual de infraestrutura

e serviços, custos de transporte e existência de redes sociais de apoio no destino. Em contextos nacionais como o brasileiro, os fluxos tendem a direcionar-se para grandes centros urbanos e polos universitários, onde se concentram melhores oportunidades e ofertas de ensino; estudos demográficos e de mobilidade destacam a importância de fatores estruturais como desigualdade regional ou mercado de trabalho precarizado ou específico e conjunturais, políticas públicas, investimentos em educação, na formação desses padrões. Para além dos determinantes econômicos, componentes culturais e de segurança também modelam os fluxos: preocupações com condições de moradia, segurança pessoal e distância de redes familiares podem inviabilizar migrações que, em teoria, seriam atraentes do ponto de vista escolar ou profissional (SENKEVICS, 2024; VERAS; LUZ SILVA, 2020).

A mobilidade estudantil é um tipo específico de fluxo migratório em que o principal objetivo do deslocamento é o acesso à educação superior. Motivações típicas incluem a busca por cursos de melhor qualidade ou oferta específica, acesso a bolsas e financiamentos, e expectativas de inserção profissional mais favoráveis. Decisões de mobilidade estudantil são condicionadas por fatores individuais como rendimentos familiares, aspirações, ou até mesmo considerando fatores institucionais como o prestígio, grade curricular das instituições e vagas oferecidas. Tem também os aspectos logísticos que compreendem moradia, custo de vida e disponibilidade de transporte. Pesquisas sobre jovens brasileiros demonstram que a mobilidade para estudar funciona tanto como estratégia de ascensão social quanto como expressão das desigualdades prévias: estudantes de escolas públicas ou de regiões periféricas movem-se em busca de oportunidades, mas frequentemente enfrentam barreiras materiais e simbólicas que limitam sua permanência e sucesso (SOUZA, T. L. de, 2017; SENKEVICS, 2024; ROCHA *et al.*, 2022).

O ENEM e o SISU atuam como mecanismos institucionais com potencial para ampliar tanto os fluxos migratórios quanto a mobilidade estudantil, ao centralizar e tornar transparente a oferta de vagas das instituições públicas. Ao permitir que uma única nota do ENEM concorra a múltiplas vagas distribuídas por diferentes estados e instituições, o SISU reduz custos administrativos de candidatura e amplia a visibilidade de vagas fora do local de origem do candidato, o que tende a aumentar a mobilidade inter-regional, especialmente quando combinado com políticas de ação afirmativa e programas de financiamento. Evidências nacionais apontam que, embora o ENEM/SISU tenham ampliado o acesso e modificado o perfil de ingressantes em várias áreas, a efetiva conversão desse acesso em permanência e conclusão depende de medidas complementares: bolsas de manutenção, programas de acolhimento, oferta de moradia estudantil e combate a obstáculos simbólicos de gênero. Essas limitações são particularmente relevantes para as mulheres em cursos de tecnologia, cuja mobilidade acadêmica é condicionado por fatores como segurança, rede de apoio e custos de adaptação à nova cidade (POUSA *et al.*, 2023; ROCHA *et al.*, 2022; VERAS; LUZ SILVA, 2020).

Para modelar e quantificar padrões de fluxos e mobilidade educacional, a litera-

tura recorre com frequência aos modelos gravitacionais (CULLINAN; DUGGAN, 2016), (SÁ; FLORAX; RIETVELD, 2003), (PELEGRINI; SÁ; FRANÇA, 2018), inspirados na analogia com a física: o fluxo entre duas localidades é proporcional à “massa” de cada local, por exemplo o número de vagas, tamanho populacional, prestígio institucional são características que atribuem consistência e decresce com a “distância” ou fricção entre elas, considera-se a quilometragem, custo/tempo de viagem, barreiras institucionais, etc. Esses modelos permitem incorporar variáveis explicativas adicionais, renda média, oferta específica de cursos, políticas de cotas, variáveis de gênero, e estimar elasticidades que explicam como mudanças na atratividade ou nas fricções alteram os fluxos observados. Aplicações ao contexto educacional brasileiro têm usado variantes do modelo gravitacional para estudar migração de estudantes entre estados e municípios (PELEGRINI; SÁ; FRANÇA, 2018), mas carece de avaliar o papel do ENEM/SISU, da concentração de polos universitários e de incentivos financeiros; contribuições modernas também precisam explorar modelos de gravidade com efeitos fixos e especificações dinâmicas para capturar trajetórias longitudinais e diferenças por gênero. A mobilidade, portanto, pode ser vista como uma estratégia para promover maior diversidade geográfica e social nos cursos de computação. No entanto, como aponta Silva, Souza e Almeida (2020), essa estratégia ainda esbarra em desafios estruturais que limitam a representatividade feminina, especialmente em áreas tecnológicas marcadas por padrões históricos de exclusão.

2.5 A INTERSEÇÃO ENTRE GÊNERO E MOBILIDADE ACADÊMICA

A análise da interseção entre gênero e mobilidade acadêmica é essencial para entender como as mulheres se inserem no campo das tecnologias. A combinação de fatores sociais, culturais e econômicos pode influenciar a decisão de migrar para estudar em diferentes estados, o que, por sua vez, pode afetar sua permanência e desempenho nos cursos de computação. Estudos como o de Amaral *et al.* (2017) e Lima, Malange e Borges (2018) apontam que a presença feminina em cursos de tecnologia é significativamente afetada por barreiras sociais e culturais, sendo a mobilidade acadêmica uma possibilidade tanto de crescimento quanto de desafio.

A falta de apoio institucional e as dificuldades de adaptação nos cursos de computação podem levar as mulheres a uma evasão precoce ou a um desempenho inferior em comparação aos seus colegas homens. Por isso, compreender os fluxos migratórios e como as políticas de acesso impactam a diversidade de gênero é essencial para a formulação de estratégias de inclusão e apoio às mulheres nos cursos de computação. Acadêmica

2.6 FERRAMENTAL DE ANÁLISE, VISUALIZAÇÃO E INFERÊNCIA

As escolhas de ferramentas e procedimentos analíticos adotadas neste trabalho foram orientadas por princípios de adequação estatística, transparência reprodutível e

capacidade de evidenciar os fenômenos investigados sobre desigualdade de gênero e mobilidade acadêmica. A implementação concreta está disponível no repositório do projeto na ferramenta de custódia de código online Github e usa o ecossistema Python para tratamento e agregação, para visualização e para modelagem e teste inferencial e será esmiuçada em maiores detalhes no Capítulo de Métodos de pesquisa (PANDAS DEVELOPMENT TEAM, 2020; MCKINNEY, 2010; AL., F. P. et, 2011; HUNTER, 2007; PLOTLY TECHNOLOGIES INC., 2023; CUNHA, 2025).

O teste de qui quadrado de independência é um procedimento estatístico destinado a verificar se a distribuição conjunta de duas variáveis categóricas se afasta daquela esperada sob a hipótese de independência, sendo a estatística calculada como a soma, sobre todas as células da tabela de contingência, do quadrado da diferença entre o valor observado e o valor esperado dividido pelo valor esperado, conforme a formulação clássica. Esse teste foi proposto por Karl Pearson no início do século vinte e desde então tornou-se um dos instrumentos fundamentais da inferência categórica (PEARSON, 1900). Do ponto de vista metodológico, o qui quadrado é importante neste trabalho porque permite testar de forma objetiva se padrões observados, por exemplo diferenças na distribuição por sexo entre modalidades de ingresso ou entre regiões de origem, podem ser explicados pelo acaso ou se indicam associações sistemáticas que merecem investigação adicional, e por isso a aplicação do teste é complementada por medidas de tamanho de efeito e por diagnósticos célula a célula (AGRESTI, 2018; HABERMAN, 1973). Na literatura aplicada em educação e estudos sobre gênero o teste de qui quadrado aparece com frequência, sendo utilizado por autores que investigam desigualdades de ingresso e seleção para demonstrar associações entre categorias sociodemográficas e desfechos de interesse, o que corrobora sua utilidade prática e interpretativa neste estudo sobre participação feminina em cursos de computação (SANTOS, 2018).

Para resolver a necessidade de identificar tendências monotônicas em séries temporais curtas utiliza-se o teste não paramétrico de Mann Kendall, procedimento que avalia se há um padrão consistente de aumento ou de queda ao longo do tempo com base nas ordens relativas das observações, isto é, sem assumir distribuição normal dos resíduos; a formulação estatística remonta aos trabalhos de Mann em 1945 e foi complementada por Kendall em 1975, que consolidaram o uso de testes de postos para detecção de tendências (MANN, 1945; KENDALL, 1975). O teste é importante para este trabalho porque a análise proposta envolve séries anuais curtas da proporção de mulheres entre candidatos e aprovados, cenários em que métodos paramétricos perdem robustez e em que *outliers* ou quebras pontuais podem distorcer estimativas de tendência; por essa razão o Mann Kendall oferece uma ferramenta estável para inferir direção de mudança.

Para quantificar a magnitude da tendência emprega-se em conjunto a estimativa de inclinação de Sen, método robusto que calcula a mediana das inclinações entre todos os pares de pontos e que fornece uma medida interpretável de mudança por unidade de tempo

(SEN, 1968). Em aplicações empíricas semelhantes o teste de Mann Kendall tem amplo uso, por exemplo em estudos hidrológicos e ambientais em que a detecção de tendência em séries anuais ou sazonais é central, caso em que Hirsch e colaboradores aplicaram e discutiram detalhadamente Mann Kendall e Sen para caracterizar mudanças temporais em séries de qualidade de água e fluxo, ilustrando tanto limites quanto procedimentos práticos de uso (HIRSCH *et al.*, 1982). Para implementação computacional e rotinas práticas recomenda-se bibliotecas especializadas que incluem testes e estimativas associadas, o que facilita a replicação das análises e a obtenção de intervalos de confiança para a inclinação de Sen (PYMANNKENDALL DEVELOPERS, 2022).

Random Forest e modelos hierárquicos são duas abordagens complementares que atendem a diferentes demandas analíticas deste estudo. Random Forest é um método de ensemble idealizado por Breiman em 2001 que constrói múltiplas árvores de decisão sobre amostras bootstrap dos dados e combina as previsões para reduzir variância e aumentar estabilidade preditiva (BREIMAN, 2001). Essa técnica é particularmente útil quando há relações não lineares e interações complexas entre variáveis individuais e geográficas, pois captura automaticamente interações sem necessidade de especificação explícita de termos de interação e apresenta elevada tolerância a observações atípicas, qualidades críticas em bases administrativas extensas e heterogêneas (AL., F. P. et, 2011; STROBL *et al.*, 2007). Modelos hierárquicos ou de efeitos mistos correspondem a uma classe de modelos desenvolvida para representar estruturas de dependência por níveis aninhados, com formulações modernas consolidadas por Laird e Ware e por Raudenbush e Bryk e com tratamento interpretativo e computacional enfatizado por Gelman e Hill (LAIRD; WARE, 1982; RAUDENBUSH; BRYK, 2002; GELMAN; HILL, 2007).

Esses modelos são importantes neste trabalho porque permitem decompor a variância em níveis distintos, por exemplo entre estudantes e entre instituições, estimando simultaneamente efeitos observados e variações institucionais não observadas, o que torna possível avaliar até que ponto diferenças de desempenho ou de seletividade são atribuíveis a características individuais ou a fatores institucionais. Em termos práticos, a combinação de Random Forest para exploração preditiva e de modelos hierárquicos para inferência multiescalar oferece um fluxo analítico robusto: Random Forest aponta preditores e padrões complexos que justificam hipóteses, enquanto os modelos hierárquicos quantificam e testam efeitos em diferentes níveis de agregação. Em aplicações empíricas, estudos metodologicamente próximos demonstraram a utilidade dessas abordagens; por exemplo Strobl e colaboradores analisaram vieses em medidas de importância de Random Forest e recomendaram o uso de permutação para obter rankings mais confiáveis de preditores (STROBL *et al.*, 2007), e Raudenbush e Bryk aplicaram modelos hierárquicos para identificar fontes de variação institucional em resultados educacionais, ilustrando como a decomposição por níveis informa decisões políticas e interpretações substantivas (RAUDENBUSH; BRYK, 2002).

Os Sankey diagrams são representações gráficas de fluxos em que a largura das faixas que ligam origens e destinos é proporcional ao volume transferido, permitindo visualizar de forma imediata a intensidade, a concentração e as assimetrias dos deslocamentos. O nome da técnica remete a Matthew Henry Phineas Riall Sankey, que no final do século dezenove consolidou esse tipo de diagrama na análise de balanços energéticos, o que lhe conferiu perfil histórico e aplicabilidade em problemas de fluxo. No contexto deste trabalho os diagramas Sankey são importantes porque tornam explicitamente mensuráveis e comunicáveis os volumes de mobilidade interestadual entre unidades federativas, facilitando a identificação de rotas dominantes e de pontos de concentração que merecem investigação adicional quando combinados com indicadores agregados como percentuais e rankings.

Além da função descritiva, os Sankey são integrados aqui a uma análise de redes, o que permite transitar da visualização para a quantificação estrutural por meio de um grafo orientado e ponderado em que é possível calcular medidas de centralidade e detectar comunidades, assim explicitando papéis de emissores, receptores e intermediários na mobilidade estudantil. Estudos empíricos que aplicaram diagramas Sankey em conjunto com análise de redes e que ilustram a utilidade dessa abordagem em investigações sobre fluxos regionais e mobilidade acadêmica incluem trabalhos como Pelegrini 2018 e Rocha e colaboradores 2022, e orientações práticas para implementações interativas encontram-se documentadas por fornecedores de ferramentas de visualização, em especial Plotly Inc 2023. Para fundamentar a extensão analítica que transcende a descrição visual, recorreremos a referências metodológicas clássicas sobre análise de redes que explicam centralidades e detecção de comunidades e que sustentam a interpretação topológica dos resultados (RIEHMANN; HANFLER; FROELICH, 2005; NEWMAN, 2010; PELEGRINI; SÁ; FRANÇA, 2018; ROCHA *et al.*, 2022; BORGATTI *et al.*, 2009).

Já os gráficos de violino são representações gráficas que combinam a estimativa de densidade por kernel com indicadores sumarizados de posição e dispersão tais como mediana e quartis, apresentando em uma única figura a forma da distribuição e estatísticas resumo que facilitam comparações entre grupos. A proposta metodológica dos violinos é atribuída a Hintze e Nelson em um trabalho que sistematizou a construção do violin plot como extensão informativa do boxplot tradicional (HINTZE; NELSON, 1998). Os violinos são importantes para este trabalho porque permitem identificar assimetrias, multimodalidade e caudas longas nas distribuições de notas e de seus subcomponentes, características que são frequentemente relevantes em bases educacionais heterogêneas e que podem ser apagadas por representações sumarizadas como o boxplot. Ao evidenciar nuances na forma das distribuições, os violinos suportam a discussão sobre heterogeneidade institucional por instituição de ensino superior, variações por ano e diferenças por curso e sexo, e servem como base visual para hipóteses que podem ser testadas com análises inferenciais e modelos multiescalares.

Estudos aplicados em avaliações educacionais e em investigações sobre performance

utilizaram gráficos de violino para demonstrar padrões de dispersão e multimodalidade e para fundamentar interpretações sobre desigualdades de desempenho, sendo exemplo prático o uso de representações de densidade acopladas a estatísticas resumo em investigações recentes sobre mobilidade e desempenho (ROCHA *et al.*, 2022). Em termos de implementação e reprodutibilidade, bibliotecas consolidadas para visualização oferecem recursos avançados para produzir violinos publicáveis e customizáveis, o que facilita a comparação com trabalhos anteriores e a reprodução das figuras em ambientes acadêmicos (WICKHAM; GROLEMUND, 2017).

Em suma, esse ferramental quando combinado, que inclui testes clássicos, métodos não paramétricos, machine learning, visualização avançada, permite abordar, de forma complementar, perguntas sobre associação, tendência, seletividade e padrões de mobilidade, fornecendo uma base analítica clara e transparente para as conclusões apresentadas neste trabalho (AGRESTI, 2018; HIRSCH *et al.*, 1982; BREIMAN, 2001; AL., F. P. *et al.*, 2011; PLOTLY TECHNOLOGIES INC., 2023; PANDAS DEVELOPMENT TEAM, 2020; CUNHA, 2025).

3 TRABALHOS CORRELATOS

A investigação sobre desigualdades de gênero no acesso, permanência e desempenho de estudantes em cursos da área de computação tem sido objeto de crescente atenção na literatura acadêmica, especialmente diante dos esforços para promover maior diversidade nas áreas de Ciência, Tecnologia, Engenharia e Matemática (STEM). Estudos nacionais recentes têm explorado diferentes aspectos relacionados a esse cenário, utilizando abordagens quantitativas e qualitativas para compreender desde o perfil sociodemográfico dos estudantes até os fatores estruturais que influenciam sua trajetória acadêmica.

Esta seção reúne e analisa os estudos mais diretamente relacionados ao presente trabalho, com ênfase em três pesquisas que se destacam por sua relevância metodológica e temática: (i) Nascimento *et al.* (2023), pela análise longitudinal da presença feminina em cursos de STEM no Brasil com base em dados do Censo da Educação Superior; (ii) Pelegrini, Sá e França (2018), por aplicar modelos gravitacionais à mobilidade estudantil e apontar fatores estruturais que dificultam o deslocamento de certos grupos; e (iii) um estudo da UFSC Castilho (2024), que investiga o perfil e desempenho de estudantes de Sistemas de Informação com base nos microdados do ENADE mas também em outros trabalhos que reforçam a importância das discussões apresentadas e também fornecem subsídios sobre as implicações dos problemas relatados a longo prazo. Esses trabalhos foram selecionados por fornecerem, respectivamente, subsídios teóricos sobre paridade de gênero, ferramentas metodológicas para análise da mobilidade, e um exemplo concreto de aplicação analítica em cursos de computação.

A análise crítica dessas contribuições permite identificar metodologias aplicáveis, resultados convergentes e lacunas ainda não exploradas, que fundamentam e justificam os caminhos metodológicos e analíticos adotados neste trabalho.

3.1 PARIDADE DE GÊNERO NO ENSINO SUPERIOR EM STEM NO BRASIL: UMA ANÁLISE DE 10 ANOS

O estudo de Nascimento *et al.* (2023) oferece uma análise longitudinal da presença feminina em cursos presenciais de bacharelado nas áreas de STEM no Brasil, com base nos microdados do Censo da Educação Superior do INEP, entre 2010 e 2020. A pesquisa adota uma abordagem quantitativa e descritiva, fundamentada na métrica de paridade de gênero da UNESCO (0,97 a 1,03), para calcular a taxa de paridade em diferentes cursos e áreas de conhecimento.

Os resultados indicam avanços em cursos como Ciências Biológicas e Química, mas revelam que áreas como Engenharia Mecânica, Engenharia Elétrica e Computação ainda apresentam taxas significativamente inferiores ao ideal. Como limitações, os autores destacam a ausência de recortes interseccionais como raça, classe ou região e o foco exclusivo em cursos presenciais. Essas lacunas reforçam a necessidade de estudos complementares,

como o presente trabalho, que integrem dimensões como mobilidade acadêmica, políticas públicas e desigualdades regionais em cursos de computação.

3.2 FATORES ASSOCIADOS À MOBILIDADE DE ESTUDANTES UNIVERSITÁRIOS NO BRASIL: UMA ANÁLISE POR MEIO DE MODELO GRAVITACIONAL

A pesquisa de Pelegrini, Sá e França (2018) emprega modelos gravitacionais para examinar os fatores que influenciam os fluxos de mobilidade acadêmica entre mesorregiões brasileiras. Com base em dados de 2017, o estudo considera variáveis associadas tanto à origem consideradas as características socioeconômicas e de acessibilidade, quanto ao destino como a atratividade institucional, diversidade de cursos, custo de vida, etc. Além da distância entre regiões, a análise é conduzida por meio de regressão binomial negativa, adequada a dados de contagem super-dispersos.

Os achados confirmam que a distância geográfica e os custos associados atuam como barreiras à mobilidade, enquanto políticas de financiamento estudantil, diversidade de cursos e menor custo de vida favorecem o deslocamento. O estudo destaca ainda o papel do ENEM/SISU na redistribuição dos fluxos e sugere investigações futuras sobre como gênero, políticas afirmativas e área de formação interagem com os padrões migratório, questões centrais ao presente trabalho, que busca articular essas variáveis na área de computação.

3.3 ANÁLISE DE PERFIL, DESEMPENHO E PROVAS DO ENADE DO CURSO DE SISTEMAS DE INFORMAÇÃO DA UFSC

No trabalho de Castilho (2024), o perfil sociodemográfico e o desempenho de estudantes concluintes do curso de Sistemas de Informação da UFSC entre 2008 e 2021 são analisados por meio do processo de Descoberta de Conhecimento em Bases de Dados (KDD), utilizando microdados do ENADE e relatórios do INEP. A metodologia abrange seleção, pré-processamento e análise estatística descritiva, contemplando variáveis como sexo, raça/cor, idade, tipo de escola de origem e ingresso via políticas afirmativas. Além disso, o estudo mapeia competências técnicas cobradas nas edições do ENADE (2014, 2017 e 2021) e relaciona os conteúdos às ementas curriculares.

Os resultados revelam um perfil discente majoritariamente masculino, branco e oriundo do ensino privado, com baixa participação de ações afirmativas. Observa-se estabilidade no desempenho médio e um alinhamento razoável entre conteúdos das provas e a matriz curricular, com destaque para áreas como programação, banco de dados e lógica computacional. As limitações incluem a ausência de dados qualitativos e a análise restrita a uma única instituição. Como contribuição, o trabalho inspira o presente estudo ao mostrar o potencial analítico dos microdados do INEP e reforça a importância de cruzamentos entre desempenho, perfil sociodemográfico e mobilidade acadêmica.

3.4 WAGE PROFILE AND GENDER GAP IN SCIENCE AND TECHNOLOGY: REGIONAL DISPARITIES IN BRAZIL

No estudo de Bonini, Silva e Sótero (2025), são analisadas as diferenças salariais entre homens e mulheres nas áreas de ciência e tecnologia no Brasil, com ênfase nas disparidades regionais. Utilizando dados da Relação Anual de Informações Sociais (RAIS) e do Instituto Brasileiro de Geografia e Estatística (IBGE), o trabalho aplica métodos econométricos de decomposição salarial de Oaxaca-Blinder para identificar o peso dos fatores observáveis (como escolaridade, experiência e setor de atuação) e não observáveis (discriminação de gênero) sobre o hiato de remuneração.

Os resultados mostram que, mesmo após controlar variáveis socioeconômicas, persiste um hiato salarial significativo desfavorável às mulheres, mais acentuado nas regiões Norte e Nordeste. As autoras observam que a desigualdade é agravada pela segregação ocupacional, pela concentração de mulheres em cargos técnicos de menor remuneração e pela baixa representatividade em posições de liderança científica. As limitações apontam a ausência de variáveis qualitativas relacionadas a barreiras institucionais e à progressão na carreira. Como contribuição, o estudo evidencia que o avanço da equidade de gênero em ciência e tecnologia depende não apenas da ampliação do acesso, mas também de políticas regionais de valorização salarial e redistribuição de oportunidades no mercado científico brasileiro.

3.5 WOMEN'S JOURNEY IN STEM EDUCATION IN BRAZIL: A RAPID REVIEW ON ENGINEERING AND COMPUTER SCIENCE

Em Costa *et al.* (2024), é conduzida uma revisão rápida da literatura sobre a trajetória das mulheres no ensino superior brasileiro nas áreas de engenharia e computação. O estudo segue o protocolo PRISMA adaptado para revisões de escopo, considerando publicações entre 2015 e 2023 nas bases Scopus, Web of Science e IEEE Xplore. As autoras classificam as evidências em três dimensões: ingresso, permanência e formação profissional, com atenção especial a políticas institucionais e práticas pedagógicas voltadas à inclusão feminina em STEM.

A análise aponta que as principais barreiras estão associadas à falta de modelos de referência, à cultura acadêmica masculina e à ausência de políticas de permanência. Programas de mentoria e iniciativas de extensão voltadas ao ensino básico mostraram impacto positivo na atração e retenção de alunas. O estudo também destaca que os cursos de computação apresentam maior disparidade de gênero do que as engenharias tradicionais. Entre as limitações, citam-se a escassez de estudos longitudinais e a falta de integração entre dados quantitativos e experiências subjetivas. Como contribuição, o trabalho sintetiza um panorama atualizado das desigualdades de gênero em STEM e propõe diretrizes para pesquisas futuras sobre equidade no ensino superior brasileiro.

3.6 GENDER EQUALITY POLICIES IN STEM IN LATIN AMERICA – A SYSTEMATIC LITERATURE REVIEW

Em Guzman *et al.* (2024), é conduzida uma revisão rápida da literatura sobre a trajetória das mulheres no ensino superior brasileiro nas áreas de engenharia e computação. O estudo segue o protocolo PRISMA adaptado para revisões de escopo, considerando publicações entre 2015 e 2023 nas bases Scopus, Web of Science e IEEE Xplore. As autoras classificam as evidências em três dimensões: ingresso, permanência e formação profissional, com atenção especial a políticas institucionais e práticas pedagógicas voltadas à inclusão feminina em STEM.

A análise aponta que as principais barreiras estão associadas à falta de modelos de referência, à cultura acadêmica masculina e à ausência de políticas de permanência. Programas de mentoria e iniciativas de extensão voltadas ao ensino básico mostraram impacto positivo na atração e retenção de alunas. O estudo também destaca que os cursos de computação apresentam maior disparidade de gênero do que as engenharias tradicionais. Entre as limitações, citam-se a escassez de estudos longitudinais e a falta de integração entre dados quantitativos e experiências subjetivas. Como contribuição, o trabalho sintetiza um panorama atualizado das desigualdades de gênero em STEM e propõe diretrizes para pesquisas futuras sobre equidade no ensino superior brasileiro.

3.7 GENDER DISPARITY IN ENGINEERING COURSES IN BRAZIL

Em (CASTRO *et al.*, 2024), os autores investigam a desigualdade de gênero em cursos de engenharia no Brasil a partir de dados do Censo da Educação Superior de 2010 a 2021. A metodologia combina análise estatística descritiva e modelagem preditiva para avaliar a evolução das matrículas femininas, a distribuição geográfica dos cursos e as correlações entre presença feminina, desempenho acadêmico e evasão.

Os resultados revelam que, apesar do aumento gradual no número de ingressantes mulheres, a participação feminina nas engenharias ainda é inferior a 25%, com maiores índices nas engenharias química e ambiental e menores nas engenharias elétrica, mecânica e de computação. A pesquisa identifica ainda que as mulheres apresentam desempenho acadêmico ligeiramente superior, mas maior propensão à evasão, associada a fatores de isolamento e à falta de redes de apoio. As limitações do estudo incluem a ausência de dados sobre trajetórias pós-formação e a dependência de registros institucionais incompletos. Como contribuição, o trabalho reforça a necessidade de estratégias pedagógicas e políticas institucionais voltadas à retenção feminina e ao enfrentamento das barreiras estruturais que persistem nas engenharias brasileiras.

3.8 COMPARAÇÃO COM TRABALHOS CORRELATOS

A comparação entre os estudos correlatos evidencia avanços importantes na compreensão das desigualdades de gênero em STEM no Brasil e, ao mesmo tempo, delimita lacunas claras que justificam a realização do presente trabalho. Trabalhos como o de Nascimento demonstram, a partir do Censo da Educação Superior, tendências longitudinais de presença feminina em cursos de STEM (NASCIMENTO *et al.*, 2023), enquanto Pelegrini e colegas fornecem instrumentos metodológicos para analisar mobilidade estudantil mediante modelos gravitacionais (PELEGRINI; SÁ; FRANÇA, 2018). Estudos institucionais, como Castilho na UFSC, mostram o potencial analítico dos microdados do ENADE para perfilar e avaliar desempenho discente após a conclusão do curso (CASTILHO, 2024). Revisões recentes e estudos regionais consolidam evidências sobre fatores institucionais e políticas de permanência que afetam a participação feminina em engenharia e computação (COSTA *et al.*, 2024; GUZMAN *et al.*, 2024; CASTRO *et al.*, 2024).

A contribuição específica deste, situa-se justamente na interseção entre esses vetores analíticos, preenchendo lacunas metodológicas e temáticas que permanecem pouco exploradas. Em primeiro lugar, enquanto a maior parte das análises longitudinais e de paridade opera em nível agregado por área ou por curso, raramente elas articulam de forma integrada a dimensão da mobilidade acadêmica interestadual com a distribuição por gênero dentro dos cursos de computação. O presente estudo vincula explicitamente fluxos de mobilidade (origem/destino por UF) às trajetórias de ingresso e desempenho feminino nos cursos de computação, permitindo avaliar em que medida a redistribuição promovida pelo ENEM e pelo SISU altera a representatividade de mulheres em polos diferentes e se tais deslocamentos se traduzem em permanência e sucesso acadêmico. Essa articulação responde a recomendações explícitas em Pelegrini e em revisões sobre políticas de equidade, que apontam para a necessidade de analisar conjuntamente mobilidade, política de ação afirmativa e área de formação (PELEGRINI; SÁ; FRANÇA, 2018).

Em segundo lugar, do ponto de vista metodológico, este estudo combina técnicas exploratórias e inferenciais de alta complementaridade que raramente aparecem juntas na literatura aplicada nacional sobre o tema. A integração de visualizações de fluxos com diagramas Sankey e análise de redes com modelos preditivos não paramétricos como Random Forest, complementada por modelos hierárquicos com efeitos mistos para decomposição de variância institucional, permite identificar padrões complexos de seletividade e, simultaneamente, quantificar a contribuição relativa de características individuais e institucionais para desfechos como nota e evasão. Essa estratégia abraça tanto a capacidade preditiva e de descoberta do Random Forest quanto a capacidade inferencial dos modelos hierárquicos, respondendo à necessidade, ressaltada em Castilho e em estudos sobre vieses de importância de variável, de utilizar medidas robustas de importância e de distinguir variância entre níveis (Strobl2007; Raudenbush2002; CASTILHO, 2024).

Em terceiro lugar, o estudo avança em termos de reprodutibilidade e transparência.

Ao disponibilizar scripts e notebooks com código-fonte e pipeline de tratamento dos microdados no repositório público (CUNHA, 2025) associado ao TCC, a pesquisa atende à demanda por documentação e replicação que muitas análises correlatas não atendem integralmente. A disponibilização do código facilita verificações, reuso por gestores e extensões metodológicas por outros pesquisadores, fortalecendo a utilidade prática das conclusões para formulação de políticas.

Tabela 1 – Comparação da abordagem proposta com as de trabalhos correlatos

Trabalhos	Aplicação em genero	Mobilidade estudantil	Análise de ingressantes no ensino superior	Código disponível
Nascimento <i>et al.</i> (2023)	X		X	X
Pelegrini, Sá e França (2018)			X	X
Castilho (2024)	X			
Castro <i>et al.</i> (2024)	X		X	
Guzman <i>et al.</i> (2024)	X			X
Costa <i>et al.</i> (2024)	X			
Bonini, Silva e Sótero (2025)	X			
Este trabalho	X	X	X	X

Fonte: Elaboração própria.

A Tabela 1 comparativa sintetiza estes pontos ao indicar que, enquanto alguns trabalhos focalizam disparidade, mobilidade ou análises institucionais isoladas, o presente TCC combina: (i) análise de gênero com recortes interseccionais e espacialização por UF; (ii) modelagem dos fluxos de mobilidade interestadual; (iii) análise de ingressantes e desempenho na prova do ENEM em cursos de computação; e (iv) código e pipelines reproduzíveis. Essa complementaridade torna a pesquisa particularmente apta a responder perguntas de política pública que vão além do acesso, interrogando também condições de ensino anteriores a prova, bem como a eficácia de medidas de apoio local e regional.

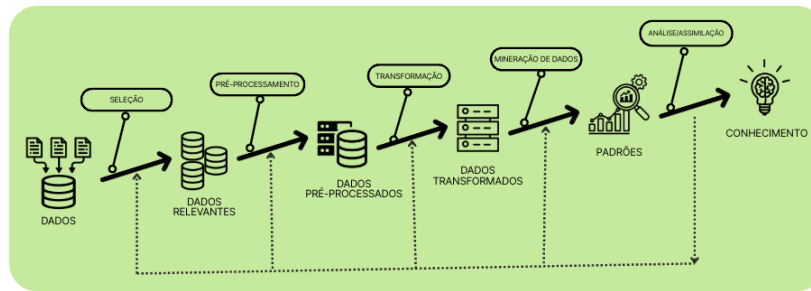
4 MÉTODOS DE PESQUISA

O estudo adotou uma abordagem quantitativa, com componentes descritivos e inferenciais, fundamentada na análise de microdados públicos do Sistema de Seleção Unificada (SISU) (MINISTÉRIO DA EDUCAÇÃO DO BRASIL (MEC), 2024), tendo como foco os cursos da área de computação. A investigação baseou-se nos dados oficiais disponibilizados pelo MEC (MINISTÉRIO DA EDUCAÇÃO, 2024), considerando variáveis como sexo, região de origem, mobilidade acadêmica, modalidade de ingresso e desempenho no ENEM. Foram analisados os cursos de Sistemas de Informação, Ciência da Computação e Engenharia de Computação, ofertados por instituições de ensino superior de todas as regiões do país, no período de 2018 a 2023.

Quanto à natureza, trata-se de uma pesquisa aplicada, voltada à geração de conhecimento útil para a compreensão de padrões de acesso, mobilidade e desempenho, com potencial para subsidiar políticas públicas e práticas de gestão universitária. A abordagem é quantitativa, uma vez que envolve o processamento de microdados em larga escala e a aplicação de métodos estatísticos e computacionais. Em termos de objetivos, o estudo apresenta caráter descritivo e exploratório, ao mapear perfis e tendências, e também inferencial, ao testar hipóteses e modelar relações entre variáveis. Quanto aos procedimentos, é uma pesquisa bibliográfica e documental, baseada em bases públicas oficiais do MEC e em referenciais teóricos da área de Ciência de Dados. O método científico adotado é estatístico-computacional, com experimentos reproduzíveis, testes de hipóteses e validações sistemáticas implementados via *scripts* e *notebooks* automatizados.

O conjunto consolidado de dados abrangeu 12 edições semestrais do SISU, entre 2018 e 2023, permitindo uma visão longitudinal das dinâmicas de ingresso. À luz do modelo KDD (*Knowledge Discovery in Databases*) (FAYYAD; PIATETSKY-SHAPIO; SMYTH, 1996), que propõe as etapas, conforme adaptado por (CASTILHO, 2024) e ilustrado na Figura 1. A pesquisa buscou assegurar qualidade do conhecimento extraído, observando três dimensões: validade (correção estatística e consistência lógica dos achados), compreensibilidade (interpretação acessível a usuários humanos) e utilidade/novidade (contribuição para o avanço do entendimento sobre desigualdades de gênero e mobilidade acadêmica).

Figura 1 – O processo de KDD



Fonte: Adaptado de (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996)

Segundo a representação acima o processo metodológico seguiram-se então as etapas clássicas do KDD descritas em (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996) e posteriormente adaptadas ao contexto dessa pesquisa, sendo elas:

1. Coleta e Seleção de dados;
2. Pré-processamento dos dados;
3. Transformação dos dados;
4. Mineração dos dados;
5. Interpretação e avaliação.

Essa estrutura metodológica permitiu conduzir um processo de descoberta de conhecimento sistemático, eficiente e reproduzível, alinhado aos princípios da Ciência de Dados e às boas práticas é importante reforçar que essas etapas não foram executadas uma única vez, conforme (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996), o uso organizado das etapas do KDD não exclui a necessidade de ajuste e de reexecução sempre que necessário para que sejam concluídos os objetivos da pesquisa. Antes de nos aprofundarmos em cada uma das etapas vamos apresentar as ferramentas que viabilizaram a execução de cada uma dessas partes.

4.1 FERRAMENTAS

Para viabilizar a transformação dos microdados obtidos do portal público do SISU em um conjunto coerente e analisável, adotou-se um fluxo metodológico totalmente automatizado baseado em *scripts* em Python 3 e geração programática de *notebooks* Jupyter para análise. O processamento utilizou de diversas bibliotecas, que possibilitaram automatizar tarefas de diagnóstico, normalização, validação, análise estatística e geração de visualizações.

Inicialmente, considerou-se a utilização da linguagem R, amplamente reconhecida por seu ecossistema voltado à estatística e análise de dados educacionais. No entanto,

optou-se pela migração para o Python, em função de sua maior integração entre estatística, modelagem computacional e aprendizado de máquina, além da robustez na automação de processos e visualização programática. Essa decisão também se justificou pela crescente adoção de Python em contextos de ciência de dados aplicada, tanto na academia quanto na indústria (MCKINNEY, 2010), (VANDERPLAS, 2016). Estudos recentes apontam que Python se consolidou como a linguagem mais utilizada por cientistas de dados (INC., 2023), superando o R em número de projetos de pesquisa, publicações e aplicações corporativas.

Python destaca-se por ser uma linguagem multiplataforma, interpretada e de código aberto, cujo ecossistema de bibliotecas permite tratar dados em larga escala, realizar testes estatísticos, gerar visualizações e aplicar técnicas de aprendizado de máquina em um único fluxo de trabalho. Essa característica possibilitou à presente pesquisa construir um pipeline automatizado, integrando desde a coleta e curadoria dos microdados até a análise exploratória, modelagem estatística e exportação de resultados visuais reproduzíveis. O uso de notebooks Jupyter reforçou a transparência e a reprodutibilidade científica do processo (PÉREZ; GRANGER, 2007), (JUPYTER, 2023), permitindo documentar as etapas de forma executável e acessível. Para a execução das análises, foram utilizadas as principais bibliotecas do ecossistema científico de Python, descritas a seguir:

- **Pandas:** Criada por Wes McKinney (2010), é a principal ferramenta de análise e manipulação de dados tabulares em Python. Baseada na estrutura DataFrame, possibilita operações eficientes de leitura, filtragem, agregação e transformação de grandes volumes de dados (PANDAS DEVELOPMENT TEAM, 2020). Neste trabalho, foi essencial para integrar e consolidar os 12 conjuntos de microdados do SISU, automatizando etapas de limpeza e padronização.
- **NumPy:** Desenvolvida por Travis Oliphant em 2005, fornece suporte a operações matemáticas e computação numérica em arrays multidimensionais, constituindo a base de diversas bibliotecas científicas em Python (NUMFOCUS, 2024). Foi utilizada para operações vetorizadas, normalização de variáveis e cálculo de estatísticas descritivas, oferecendo desempenho superior ao R em processamento matricial de grande escala (HARRIS; MILLMAN; WALT ET AL., 2020), (VANDERPLAS, 2016).
- **Matplotlib:** Criada por John D. Hunter (2007), é uma biblioteca de visualização científica 2D inspirada no MATLAB, que permite a criação de gráficos de barras, linhas, boxplots e heatmaps (HUNTER, 2007), (HARRIS; MILLMAN; WALT ET AL., 2020). Na pesquisa, foi usada para gerar as visualizações estáticas padronizadas incluídas no notebook automatizado.
- **Plotly:** Biblioteca interativa para visualização de dados baseada em JavaScript, amplamente utilizada em dashboards e aplicações web (PLOTLY TECHNOLOGIES INC., 2023). Foi aplicada na representação dos fluxos de mobilidade interestadual

por meio de gráficos Sankey, destacando os movimentos de candidatos entre estados de origem e destino.

- **SciPy:** Conjunto de ferramentas científicas que expandem o NumPy, oferecendo funções estatísticas e de otimização (AL., P. V. et, 2020) Nessa pesquisa, foi empregada na execução de testes inferenciais, como qui-quadrado, Mann-Whitney U, Welch t-test e correlação de Pearson, para examinar diferenças e associações entre grupos.
- **Statsmodels:** Desenvolvida por Seabold e Perktold (STATSMODELS DEVELOPMENT TEAM, 2024), (SEABOLD; PERKTOLD, 2010), permite a aplicação de modelos estatísticos e econométricos, incluindo modelos lineares mistos e análises de efeitos fixos e aleatórios. Essa biblioteca foi usada para investigar variações de desempenho por gênero, região e modalidade de ingresso.
- **Scikit-learn (sklearn):** Criada por (AL., F. P. et, 2011), é apontada pelo mesmo como principal biblioteca de aprendizado de máquina em Python, oferecendo implementações de algoritmos de regressão, classificação e agrupamento. No presente estudo, foi utilizada para análise de agrupamento usando k-means e Random Forest, possibilitando identificar padrões de perfil e variáveis mais relevantes na explicação das diferenças de desempenho.
- **PyMannKendall:** Implementação do teste de Mann-Kendall em Python, amplamente usado para análise de tendência temporal não paramétrica (HIRSCH *et al.*, 1982). Foi aplicada para verificar a evolução da proporção de mulheres ao longo das edições do SISU, fornecendo uma avaliação estatística das tendências observadas nos gráficos.

Além das bibliotecas, foi desenvolvido o script `build_notebook_tcc.py`, que automatiza a geração do notebook analítico `SISU_Analises_TCC.ipynb` de forma programática, utilizando a biblioteca `nbformat` (JUPYTER, 2023). Esse script cria células Markdown e de código Python organizadas pelos objetivos específicos deste trabalho, incorporando análises descritivas e inferenciais, testes estatísticos e visualizações padronizadas para assim chegar as respostas de Q1 até Q5. Essa abordagem de *literate programming* (KNUTH, 1984) permite documentar o raciocínio analítico de forma executável e auditável, atendendo aos princípios de ciência aberta e reprodutibilidade computacional (PENG, 2011; STODDEN *et al.*, 2014).

O ambiente de desenvolvimento utilizou Python 3.10+ em sistema macOS, com gerenciamento de dependências via `pip` e arquivo `requirements.txt`, garantindo a instalação reprodutível das versões exatas das bibliotecas. O controle de versão foi realizado com Git, e o código-fonte foi disponibilizado em repositório público no GitHub, seguindo as recomendações de (WILSON *et al.*, 2014) para boas práticas em computação científica.

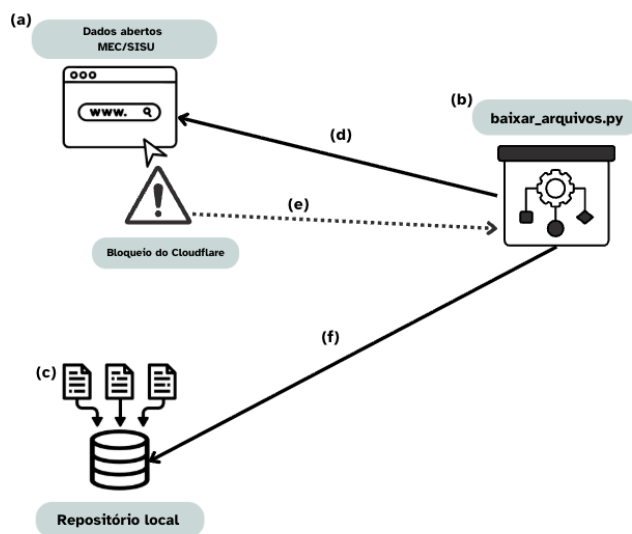
Assim, a adoção do Python representou não apenas uma escolha técnica, mas também epistemológica e metodológica, permitindo um equilíbrio entre rigor estatístico e flexibilidade computacional. Sua adoção consolidou o estudo como um exemplo de pesquisa quantitativa baseada em dados abertos e automação científica, fortalecendo a transparência, replicabilidade e acessibilidade dos resultados.

4.2 COLETA E SELEÇÃO DOS DADOS

A coleta dos dados foi realizada a partir do Portal de Dados Abertos do MEC (MINISTÉRIO DA EDUCAÇÃO, 2024), que disponibiliza, por edição anual mas com resultados semestrais, os microdados públicos do SISU em formato CSV mas com alguns casos comprimidos no formato ‘.zip’ também. O recorte temporal abrange 5 edições consecutivas, com resultados semestrais de 2018.1 a 2023.2, que são subsequentes e complementares ao estudo realizado por (CURSINO; MARTINEZ, 2021), totalizando aproximadamente 21.710.255 linhas de registros brutos, somando 10.20 GB. Os arquivos apresentam estrutura tabular com cabeçalhos por coluna e pequenas variações de nomenclatura entre semestres e anos diferentes, como a presença ou ausência de campos, ou renomeação de certos campos, o que exigiu padronização posterior para garantir comparabilidade longitudinal.

Empregou-se então uma classe de nome `baixar_arquivos.py` para sistematizar o *download* direto dos CSVs que são hospedados sob padrões de URL previsíveis no site do SISU hospedado pelo MEC (MINISTÉRIO DA EDUCAÇÃO, 2024), conforme apresentado na Figura 2, apresentado pela marcação **(a)**. O script **(b)** implementa: a criação de repositório local para a recepção dos dados **(c)**; a construção as sessões com cabeçalhos customizados, a opção de carregar cookies já validados do navegador utilizando a biblioteca `browser_cookie3` e, na tentativa de conexão com o portal, a utilização de múltiplos padrões de nomes com *retry*, que realiza repetições de uma operação que falhou e *backoff* que é uma estratégia para espaçar as tentativas de *retry*, isto é, o tempo que se aguarda entre uma tentativa e outra **(d)**; a detecção e a extração automática quando o recurso vem compactado; e a busca pelas combinações de anos e semestres fornecidas, contabilizando sucessos e falhas e registrando *logs* de cada tentativa.

Figura 2 – O processo de coleta de dados



Fonte: A autora

Apesar do uso de bibliotecas Python robustas e da plataforma CKAN, Comprehensive Knowledge Archive Network, que fornece catálogo, API e ferramentas para descoberta e reutilização de conjuntos de dados, não foi possível realizar a descoberta automatizada e estável dos arquivos via portal conforme ilustrado pela Figura 2. As requisições retornaram repetidamente o código 403 em razão de medidas de proteção anti-bot adotadas pelo provedor do portal, implementadas por meio do serviço Cloudflare (e), o que também inviabilizou estratégias de raspagem HTML confiáveis.

Em função dessa limitação foram adotadas alternativas manuais e semiautomáticas (f), Em razão dessa limitação adotou-se como solução de contorno o uso de links diretos determinísticos por edição com base no padrão estático de nomeação dos arquivos, por exemplo `ListagemChamadaRegular_{ANO}-{SEM}.csv`, e foi incorporado um tratamento específico para quando os arquivos se encontravam compactados em formato zip, exigindo descompressão prévia para leitura e transformação. O acesso foi realizado por download manual assistido com utilitários de transferência de arquivos de linha de comando tais como `wget` e `curl` que permitem automação controlada das requisições e registro detalhado das operações. Para preservar evidências da coleta foram conservados os nomes originais dos arquivos e gerados resumos criptográficos de cada artefato, usando algoritmos de hash padronizados como SHA-256, e esses resumos juntamente com os nomes de arquivo e carimbos temporais foram armazenados no repositório de dados local como prova de integridade e para facilitar auditoria e reprodutibilidade. Essa estratégia equilibrou reprodutibilidade e robustez diante das barreiras anti automação do portal do MEC onde residem os dados abertos do SISU(MINISTÉRIO DA EDUCAÇÃO DO BRASIL (MEC), 2024; PROJECT, G., s.d.; TEAM, s.d.; STANDARDS; TECHNOLOGY, 2015).

4.2.1 Estrutura dos dados

Dois tipos de inconsistência afetaram o processo de seleção diretamente: a inexistência de um dicionário de dados consolidado com estabilidade para todo o período e as variações de *schema* dos dados entre edições, muitas mudanças sutis de nome de coluna, inclusão ou remoção de atributos, codificações distintas, etc. Em termos metodológicos, isso demandou a confecção inicial um dicionário próprio a partir dos cabeçalhos e exemplos de valores presentes nos arquivos `ListagemChamadaRegular_{ANO}-{SEM}.csv`, da documentação pública disponível no portal do MEC/SISU (MINISTÉRIO DA EDUCAÇÃO DO BRASIL (MEC), 2024) quando existente, literatura que já havia utilizado esses dados (PELEGRINI; SÁ; FRANÇA, 2018), (SOUZA, T. L. de, 2017), (SILVA; SOUZA; ALMEIDA, 2020) e de inspeções exploratórias conduzidas durante a etapa de curadoria (KITCHIN, 2014). Em posse do dicionário, ainda durante a curadoria e harmonização, ocorreu a normalização de nomes, definição de tipos, alinhamento de domínios, antes de qualquer comparação entre semestres, em linha com o que a literatura descreve como trabalho inevitável de *data wrangling*, manipulação dados em tradução livre, no que diz respeito de um ecossistemas de dados públicos e “*big data*” governamental (KITCHIN, 2014). Tais procedimentos incrementais de qualidade são fundamentais para assegurar validade interna e reprodutibilidade do estudo (CRESWELL, 2014).

A Tabela 2 sumariza as variáveis encontradas nos arquivos brutos. Nela apresentamos a nomenclatura do campo, conforme aparece nos CSVs, o tipo de dado após padronização e uma descrição operacional. Observa-se que identificadores pessoais sensíveis, como CPF e inscrição do ENEM, aparecem mascarados nos microdados públicos em atendimento à LAI e à LGPD, e não foram utilizados para fins analíticos. Quando pequenas variações de nomenclatura ocorreram entre edições, adotamos o rótulo mais frequente e documentamos equivalências no pipeline. As fontes primárias e a estrutura semestral dos arquivos constam no Portal de Dados Abertos do MEC (MINISTÉRIO DA EDUCAÇÃO, 2024).

Tabela 2 – Dicionário de dados do recorte SISU (2018.1–2023.2)

DADO	TIPO	DESCRIÇÃO
NU_ANO	Numérico	Ano do processo seletivo
NU_EDICAO	Numérico	Número da edição do processo seletivo no ano de referência
CO_ETAPA	Numérico	Código da etapa no Sisu Gestão. Refere-se ao período de inscrição, chamada regular ou lista de espera

DADO	TIPO	DESCRIÇÃO
DS_ETAPA	Alfanumérico	Informa qual é a etapa do processo seletivo: se Inscrição, resultado da Chamada Regular (1ª, 2ª ou 3ª conforme a edição), ou informação da Lista de Espera
CO_IES	Numérico	Código da instituição de ensino superior conforme informações do cadastro e-MEC
NO_IES	Alfanumérico	Nome da instituição de ensino superior conforme informações do cadastro e-MEC
SIGLA_IES	Alfanumérico	Sigla da instituição de ensino superior conforme informações do cadastro e-MEC
SG_UF_IES	Alfanumérico	Sigla UF do campus sede da instituição de ensino superior conforme informações do cadastro e-MEC
CO_CAMPUS	Numérico	Código do campus da instituição de ensino superior conforme informações do cadastro e-MEC
NO_CAMPUS	Alfanumérico	Nome do campus da instituição de ensino superior conforme informações do cadastro e-MEC
SG_UF_CAMPUS	Alfanumérico	Sigla da unidade da federação (UF) na qual está localizado o campus conforme informações do cadastro e-MEC
NO_MUNUCIPIO_CAMPUS	Alfanumérico	Nome do município do campus da instituição de ensino superior conforme informações do cadastro e-MEC
CO_IES_CURSO	Numérico	Código do curso da instituição de ensino superior conforme informações do cadastro e-MEC
NO_CURSO	Alfanumérico	Nome do curso da instituição de ensino superior conforme informações do cadastro e-MEC
DS_GRAU	Alfanumérico	Grau do curso da instituição de ensino superior conforme informações do cadastro e-MEC
DS_TURNO	Alfanumérico	Turno do curso da instituição de ensino superior conforme informações do cadastro e-MEC

DADO	TIPO	DESCRIÇÃO
TP_MOD_CONCORRENCIA	Alfanumérico	Tipo da modalidade da oferta do curso no processo seletivo
DS_MOD_CONCORRENCIA	Alfanumérico	Descrição da modalidade de concorrência ofertada para o curso no processo seletivo
NU_PERCENTUAL_BONUS	Numérico	Percentual de bônus definido para as ações afirmativas próprias das IES
NU_PESO_L	Numérico	NU_PESO atribuído pela instituição para a prova de Linguagens do Enem
NU_PESO_CH	Numérico	NU_PESO atribuído pela instituição para a prova de Ciências Humanas
NU_PESO_CN	Numérico	NU_PESO atribuído pela instituição para a prova de Ciências da Natureza
NU_PESO_M	Numérico	NU_PESO atribuído pela instituição para a prova de Matemática
NU_PESO_R	Numérico	NU_PESO atribuído pela instituição para a prova de Redação
NOTA_MINIMA_L	Numérico	Nota mínima definida pela IES para Linguagens
NOTA_MINIMA_CH	Numérico	Nota mínima definida pela IES para Ciências Humanas
NOTA_MINIMA_CN	Numérico	Nota mínima definida pela IES para Ciências da Natureza
NOTA_MINIMA_M	Numérico	Nota mínima definida pela IES para Matemática
NOTA_MINIMA_R	Numérico	Nota mínima definida pela IES para Redação
MEDIA_MINIMA	Numérico	Média aritmética mínima (não considera pesos) definida pela IES considerando as cinco provas do Enem
NU_CPF	Numérico	CPF do estudante (mascarado em microdados públicos; não utilizado nas análises)
CO_INSCRICAO_ENEM	Numérico	Código de inscrição do estudante no ENEM/INEP (mascarado em microdados públicos)
NO_INSCRITO	Alfanumérico	Nome do estudante
TP_SEXO	Alfanumérico	Sexo informado pelo estudante no questionário socioeconômico do ENEM/INEP

DADO	TIPO	DESCRIÇÃO
DT_NASCIMENTO	Numérico	Data de nascimento informada no questionário socioeconômico do ENEM/INEP
SG_UF_CANDIDATO	Alfanumérico	UF do município de residência do estudante (questionário socioeconômico do ENEM/INEP)
MUNICIPIO_CANDIDATO	Alfanumérico	Município de residência do estudante (questionário socioeconômico do ENEM/INEP)
ST_OPCAO	Numérico	Indica se é a 1ª ou 2ª opção do estudante
NU_NOTA_L	Numérico	Nota obtida na prova de Linguagens do ENEM/INEP
NU_NOTA_CH	Numérico	Nota obtida na prova de Ciências Humanas do ENEM/INEP
NU_NOTA_CN	Numérico	Nota obtida na prova de Ciências da Natureza do ENEM/INEP
NU_NOTA_M	Numérico	Nota obtida na prova de Matemática do ENEM/INEP
NU_NOTA_R	Numérico	Nota obtida na prova de Redação do ENEM/INEP
NOTA_L_COM_PESO	Numérico	Resultado de NU_NOTA_L multiplicado por NU_PESO_L
NOTA_CH_COM_PESO	Numérico	Resultado de NU_NOTA_CH multiplicado por NU_PESO_CH
NOTA_CN_COM_PESO	Numérico	Resultado de NU_NOTA_CN multiplicado por NU_PESO_CN
NOTA_M_COM_PESO	Numérico	Resultado de NU_NOTA_M multiplicado por NU_PESO_M
NOTA_R_COM_PESO	Numérico	Resultado de NU_NOTA_R multiplicado por NU_PESO_R
NU_NOTA_CANDIDATO	Numérico	Média ponderada pelas NU_PESOs definidos pelas IES e, quando aplicável, bônus de ação afirmativa

DADO	TIPO	DESCRIÇÃO
NU_NOTACORTE	Numérico	Nota do último classificado na modalidade/curso (chamada regular). Em bônus, igual à ampla concorrência do mesmo curso/grau/turno/campus/IES; pode estar ausente em edições com poucos inscritos ou migração para 1ª opção
NU_CLASSIFICACAO	Numérico	Posição do estudante considerando todos os inscritos do mesmo curso
ST_APROVADO	Alfanumérico	Indica se o estudante foi aprovado na chamada regular em sua opção de curso
ST_MATRICULA	Alfanumérico	Situação registrada no Sisu Gestão (por exemplo: matriculado, documentação rejeitada, não compareceu etc.), na chamada regular ou lista de espera

4.2.2 Seleção dos dados

A seleção dos dados concentrou-se estritamente em cursos da área de computação, adotando nomes exatos, são eles “Ciência da Computação”, “Engenharia de Computação” e “Sistemas de Informação”, de modo a evitar ambiguidades com cursos correlatos. A partir do conjunto completo de variáveis disponibilizadas em cada edição, foram retidos 54 campos de interesse como base analítica final, contendo identificadores de curso e IES, UF e campus, modalidade de concorrência, notas do ENEM, situação de ingresso, entre outros. Em alinhamento a princípios éticos e às boas práticas de pesquisa com dados administrativos (CRESWELL, 2014), assegurou-se a ausência de dados sensíveis, com destaque para a não utilização de identificadores pessoais diretos. Quando presente, o CPF encontra-se ausente ou mascarado, não permitindo reidentificação de indivíduos.

Nessa etapa, de acordo com os objetivos específicos desse trabalho que também formam as perguntas de pesquisa (Q1 à Q5), que orientaram a definição dos subconjuntos de variáveis utilizados em cada análise. Cada questão envolveu um recorte distinto do dataset consolidado, preservando apenas as colunas necessárias à sua resposta, conforme descrito a seguir:

- Q1 – Como se caracteriza o perfil sociodemográfico dos candidatos e candidatas aos cursos de Ciência da Computação, Engenharia de Computação e Sistemas de Informação, em termos de gênero, região de origem e modalidade de ingresso, no período de 2018 a 2023?
 - Dados selecionados: ano e semestre da edição, sexo, UF de origem, curso, modalidade de concorrência, situação de aprovação.

- Q2 – Como se caracteriza o perfil sociodemográfico dos aprovados e aprovadas aos cursos de Ciência da Computação, Engenharia de Computação e Sistemas de Informação, em termos de gênero, região de origem e modalidade de ingresso, no período de 2018 a 2023?
 - Dados selecionados: ano e semestre da edição, sexo, UF de origem, curso, modalidade de concorrência, situação de aprovação.
- Q3 – Qual é a distribuição e o comportamento do desempenho dos candidatos e candidatas em cada curso de computação, considerando o efeito do curso e da modalidade de concorrência sobre as notas médias?
 - Dados selecionados: identificadores de curso, nome da instituição de ensino, sigla da IES, sigla do estado da IES, e notas médias dos candidatos.
- Q4 – Quais padrões de mobilidade geográfica são observados entre os candidatos e candidatas, isto é, de que forma a migração entre estados (origem $\times$ destino) se relaciona à representação feminina nos cursos de computação?
 - Dados selecionados: UF de origem do candidato, UF de destino da instituição, e identificadores de curso e IES.
- Q5 – Como evoluiu, ao longo das edições de 2018 a 2023, a proporção de mulheres inscritas e aprovadas nos cursos de computação, e é possível identificar tendências estatisticamente significativas de crescimento ou retração dessa participação?
 - Dados selecionados: ano e semestre, sexo, curso e modalidade de ingresso.

Após a seleção e filtragem pelos três cursos-alvo e a seleção das 54 variáveis harmonizadas, obteve-se um conjunto analítico com 752.735 de linhas registros do período de 2018.1 à 2023.2, contendo 267 MB de tamanho. Esse *dataset* consolidado é a base para as análises subsequentes de representação feminina e desempenho por curso, estado e IES. A estratégia de coleta, seleção e curadoria foi guiada por boas práticas de desenho de pesquisa e de preparo de dados para estudos quantitativos com fontes administrativas (CRESWELL, 2014) e fundamentada na compreensão dos limites ou contingências de infraestruturas de dados públicos em larga escala (KITCHIN, 2014), com ancoragem operacional no repositório oficial do MEC.

4.3 PRÉ-PROCESSAMENTO DOS DADOS

O pré-processamento é uma etapa central no fluxo KDD. É a etapa onde os dados brutos, muitas vezes ruidosos, incompletos e heterogêneos, são transformados em uma representação consistente e informativa passível de descoberta de padrões e modelagem. Sem essa etapa, algoritmos de mineração e aprendizagem tendem a produzir

resultados enviesados, instáveis ou mesmo inúteis, o KDD clássico já coloca limpeza, integração e transformação de dados como fases necessárias antes da mineração propriamente dita (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). No contexto desse trabalho não foi diferente, esse marco teve como objetivo garantir qualidade, consistência e formato adequados para as análises subsequentes (WICKHAM; GROLEMUND, 2017; OSBORNE, 2013). Todas as rotinas foram implementadas no script `tratar_dados.py`, com execução única e reproduzível sobre as 12 edições adquiridas pelo processo de coleta e seleção, produzindo assim artefatos de auditoria e logs. As etapas do pré-processamento foram:

- Detecção e padronização de *encoding*;
- Padronização de cabeçalhos;
- Conversão explícita de tipos (*casting*);
- Validação dos dados resultantes.

Arquivos de texto são sequências de bytes cuja interpretação depende da codificação de caracteres adotada, conhecida em ambientes técnicos por *encoding*, e a decodificação com um *encoding* inadequado corrompe acentos e símbolos ou produz caracteres de substituição. As codificações mais empregadas em contextos brasileiros incluem UTF-8, ISO-8859-1 também referido como latin1 e Windows-1252, padrões cujas propriedades e mapeamentos são documentados pelo consórcio Unicode e pelos registros oficiais de conjuntos de caracteres da IANA (CONSORTIUM, s.d.[a]; IANACHARSETS, s.d.). A inferência automática de *encoding* recorre a detectores estatísticos que analisam padrões de bytes e frequências de sequências esperadas em diferentes conjuntos, abordagem amplamente utilizada nas bibliotecas `chardet` (TEAM, s.d.) e `charset-normalizer` (OUSRET, s.d.) e originada no trabalho do detector universal da Mozilla, porém tais heurísticas não são infalíveis porque textos curtos, amostras com poucos caracteres alfabéticos ou coincidências de padrões byte-a-byte entre encodings distintos podem conduzir a falsos positivos como a identificação indevida de JOHAB em arquivos que contêm português; por isso a detecção automática deve ser tratada como sugestão a ser validada por verificações adicionais (PROJECT, M., s.d.; SPOLSKY, s.d.).

A padronização de *encodings* foi implementada de forma automática e robusta no pipeline de pré-processamento. Primeiro tenta-se inferir a codificação com a biblioteca `chardet` quando disponível através da função `_detectar_encoding` e, para evitar falsos positivos óbvios, codificações identificadas como JOHAB são mapeadas para UTF-8; na sequência cada arquivo é lido pela rotina `ler_csv_flex` que testa de maneira determinística combinações de *encodings* e separadores comuns listando a detecção sugerida seguida de UTF-8, ISO-8859-1 e Windows-1252 e os separadores vírgula, ponto e vírgula, pipe e tab até que uma combinação produza um *DataFrame* plausível com mais de uma coluna, e arquivos compactados são previamente extraídos para que o conteúdo CSV seja submetido

às mesmas tentativas de leitura. Após a execução todos os dados consolidados são gravados em UTF-8 e os binários originais são preservados na pasta de ingestão juntamente com *logs* e resumos criptográficos que documentam o *encoding* detectado, o *encoding* efetivamente usado e o resultado das validações, permitindo reprocessamento e auditoria conforme práticas recomendadas para pipelines de dados (SANDVE *et al.*, 2013; WALT *et al.*, 2011; WILKINSON *et al.*, 2016).

Durante a padronização de cabeçalhos, dadas as pequenas variações de nomenclatura entre edições, construímos um mapa de harmonização de cabeçalhos para um esquema canônico com 54 campos. O dicionário de equivalências registra as renomeações sutis, como por exemplo diferenças de sublinhado, utilização de hifens ou separadores diferentes, colunas ausentes ou adicionadas em edições específicas e regras de preenchimento ou derivação quando aplicável. O resultado é um conjunto longitudinal comparável e documentado em repositório virtual no github (WICKHAM; GROLEMUND, 2017; CUNHA, 2025).

Após a harmonização dos nomes de coluna e do esquema comum, foi aplicado um *casting* controlado para garantir tipagens consistentes ao longo de todas as edições. Foram definidos trinta e dois campos numéricos, correspondendo a notas, pesos, médias, códigos e posições, e vinte e dois campos categóricos, relativos a nomes, siglas, unidades federativas, modalidades e indicadores de status. Para os campos numéricos, a rotina converteu os valores para *string*, substituiu vírgulas por pontos e removeu espaços em branco antes de aplicar uma conversão numérica robusta com coerção de erros para valor ausente. Nessa etapa foi utilizado o recurso de conversão tolerante das bibliotecas de manipulação tabular que transforma entradas inválidas em valores nulos em vez de lançar exceções que interromperiam o processamento. Para colunas que precisam preservar a noção de inteiro, com possibilidade de falta de dado, foram usados tipos inteiros anuláveis, garantindo integridade sem perder a informação de ausentes. As colunas categóricas foram normalizadas para *string*, com remoção de espaços e padronização de caixa, e foram aplicados mapeamentos canônicos quando necessário, por exemplo para uniformizar variações textuais de nomes de curso.

O processo de *casting* foi acompanhado por validações automáticas e regras de coerção para reduzir erros silenciosos de tipo. A leitura inicial de cada arquivo tentou detectar a codificação e o separador mais provável, e utilizou um mecanismo de tentativa e erro com tratamento de exceções para retomar a leitura sem interromper o pipeline quando ocorrências atípicas surgiam. Após a conversão, foram aplicadas regras de consistência adicionais, tais como normalização do campo de sexo para permitir apenas as etiquetas M e F, conversão de datas de nascimento para um formato anual padronizado e transformação de zeros em valores ausentes em contextos onde havia evidência de notas parcialmente preenchidas. Após a harmonização e *casting* controlado, ficamos com:

- **32 colunas numéricas (float):** sendo elas as notas, pesos, médias, códigos

e posições (p.ex., `NU_NOTA_`, `NU_PESO_`, `NU_NOTA_CANDIDATO`, `NU_NOTACORTE`, `NU_CLASSIFICACAO`).

- **22 colunas categóricas (string):** entre elas os nomes, siglas, UFs, modalidades, status (p.ex., `NO_IES`, `NO_CURSO`, `SG_UF_IES`, `TP_SEXO`, `DS_MOD_CONCORRENCIA`).

Os artefatos de verificação automática gerados pelo `tratar_dados.py` asseguram transparência e rastreabilidade do processamento ao consolidar evidências e permitir auditar cada etapa do pré-processamento e da limpeza dos microdados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; WANG; STRONG, 1996). Em particular o arquivo `relatorio_consistencia.txt` reúne estatísticas descritivas estratificadas por ano, curso e sexo, indicadores de completude e contagens de valores fora de intervalos plausíveis; estes elementos expõem problemas de ingestão como colunas faltantes, codificações inesperadas e variações estruturais entre edições e orientam intervenções corretivas e verificações posteriores.

O arquivo `outliers_notas.csv` identifica valores extremos com base no z-score definido por $z = \frac{x - \bar{x}}{s}$, onde $\bar{x}$ é a média e s o desvio padrão. A regra adotada $|z| > 3$ decorre da regra empírica dos três desvios padrão e é útil por sua simplicidade e interpretabilidade

$$z = \frac{x - \bar{x}}{s}$$

(IGLEWICZ; HOAGLIN, 1993; BARNETT; LEWIS, 1994). É importante notar que métodos baseados em média e desvio padrão são sensíveis a assimetrias e a poucos valores extremos porque a própria média e o desvio podem ser distorcidos por *outliers*, pelo que o z score serve melhor para capturar extremos absolutos que se destacam claramente da massa central.

O arquivo `outliers_robustos.csv` aplica O critério de Tukey usa o intervalo interquartil definido por $IQR = Q_3 - Q_1$, com limites numéricos em $Q_1 - 1,5 IQR$ e $Q_3 + 1,5 IQR$., procedimento recomendado para análise exploratória por sua robustez frente a assimetrias e em cenários de pequenas amostras (TUKEY, 1977). Esses limites foram calculados de forma estratificada por curso, sexo e edição para que anomalias sejam detectadas em relação ao contexto de cada subpopulação e não em relação a uma distribuição agregada que poderia ocultar diferenças legítimas. A combinação de critérios complementares, juntamente com a preservação de logs que documentam as operações realizadas e de resumos criptográficos de integridade dos arquivos, também conhecidos como hashes e tipicamente calculados com algoritmos padronizados como SHA-256, permite inspeção manual, análises de sensibilidade e reprodução dos resultados conforme recomendações do processo KDD e das boas práticas de pesquisa reprodutível (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; SANDVE *et al.*, 2013; WILKINSON *et al.*, 2016).

Os registros marcados como *outliers* não são removidos automaticamente, sendo apenas sinalizados para análise posterior, conforme o processo iterativo de descoberta de conhecimento proposto por (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). O *pipeline* segue princípios de *tidy data* e de automação declarativa (WICKHAM; GROLEMUND, 2017), garantindo que qualquer nova execução, a partir dos arquivos originais do MEC, reproduza exatamente o mesmo conjunto consolidado e as mesmas evidências de qualidade. Assim, o pré-processamento não apenas assegurou a integridade estatística dos dados, mas também reforçou a confiabilidade e a reprodutibilidade científica da pesquisa.

4.4 TRANSFORMAÇÃO DOS DADOS

A etapa de transformação incluiu conversões controladas de tipos e normalizações numéricas com vistas à compatibilidade entre algoritmos e à comparabilidade entre escalas. Variáveis numéricas foram coeridas e, quando apropriado, normalizadas por padronização por z score ou por escala mínima-máxima, procedimentos que impedem que diferenças de ordem de grandeza dominem algoritmos sensíveis a distância ou a regularização e que são recomendados pela literatura de aprendizado estatístico e por implementações consolidadas como scikit-learn (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; AL., F. P. et, 2011; MCKINNEY, 2010). Essas operações foram aplicadas de forma seletiva, tendo em conta a interpretação das variáveis e a arquitetura dos modelos a empregar, de modo a preservar significado e comparabilidade sem introduzir vieses indevidos por re-escalamento.

Para variáveis categóricas adotou-se a codificação adequada à natureza e à cardinalidade de cada atributo. Utilizou-se *one-hot encoding* para atributos nominais de baixa cardinalidade e ordinal *encoding* para variáveis com ordenação intrínseca; para categorias de alta cardinalidade avaliaram-se alternativas como *feature hashing*, que reduz dimensionalidade e tolera valores não vistos em produção mediante o risco controlado de colisões (WEINBERGER *et al.*, 2009), e representações densas aprendidas, conhecidas por *embeddings*, que comprimem categorias em vetores contínuos capazes de preservar similaridade e facilitar a modelagem de interações não lineares (GUO *et al.*, 2017; MIKOLOV *et al.*, 2013). A seleção entre essas técnicas foi fundamentada em *trade-offs* entre interpretabilidade, custo computacional, robustez a valores raros e sensibilidade do modelo ao aumento de dimensões, seguindo práticas de engenharia de *features* e recomendações de implementação (AL., F. P. et, 2011; MCKINNEY, 2010).

A engenharia de atributos complementou as codificações com derivados formulados a partir de conhecimento teórico e hipóteses substantivas, com o objetivo de melhorar capacidade explicativa e preditiva mantendo parcimônia do modelo. Foram construídas proporções relevantes, interações prescritas por hipóteses, agregações temporais e janelas móveis para expor tendências e sazonalidade; quando adequado, aplicaram-se discretizações para estabilizar relações não lineares e transformações monotônicas, como logaritmo ou *Box-Cox*, para reduzir assimetrias. Para mitigar ruído e multicolinearidade e reduzir custo

computacional empregaram-se técnicas de redução e seleção de dimensão, incluindo PCA e métodos de seleção por filtro, *wrapper* e *embedded*, escolhidos conforme o objetivo analítico e os recursos computacionais disponíveis (TUKEY, 1977; DASU; JOHNSON, 2003; AL., F. P. et, 2011). Todas as transformações foram implementadas em pipelines reproduzíveis que preservam separação estrita entre treino e avaliação, e que registram decisões e metadados para permitir reproprocessamento, auditoria e análises de sensibilidade conforme as boas práticas do processo KDD e da pesquisa reproduzível (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; SANDVE *et al.*, 2013; WILKINSON *et al.*, 2016).

Para reduzir ruído categórico entre edições procedeu-se à canonicalização dos nomes de curso, mapeando variações nominais para três rótulos canônicos: Ciência da Computação, Sistemas de Informação e Engenharia da Computação. O fluxo de normalização textual aplicou normalização Unicode nas formas NFD e NFKD, remoção de diacríticos, trim e *case folding* conforme as regras do *Unicode* para comparações robustas de strings em contextos multilíngues, seguido por regras de mapeamento de sinônimos e variações ortográficas que unificaram termos como Ciências da Computação e Computação no rótulo `CURSO_CANONICO`(CONSORTIUM, s.d.[b]). O mapeamento foi versionado e documentado junto aos artefatos do pipeline, garantindo rastreabilidade das regras aplicadas e consistência terminológica entre edições do conjunto de dados.

Foram criadas variáveis de apoio alinhadas às questões de pesquisa, todas calculadas de forma uniforme em 2018.1–2023.2:

- `APROV_BOOL`: conversão do campo textual `ST_APROVADO` em Booleano, com regra *true* para casos aprovados na chamada regular da respectiva opção e *false* caso contrário. Registros sem informação explícita permanecem *missing*, sem imputação.
- `REG_ORIGEM` e `REG_IES`: mapeamento de UFs para macrorregiões do IBGE, sendo Norte, Nordeste, Centro-Oeste, Sudeste, Sul, a partir de `SG_UF_CANDIDATO` sendo a origem e `SG_UF_IES` o destino. Casos especiais como o DF segue a classificação oficial do IBGE.
- `MIGRA`: indicador binário de mobilidade interestadual, definido por $MIGRA = (SG_UF_CANDIDATO \neq SG_UF_IES)$ quando ambas as UFs são válidas mas são diferentes entre si. Valores ausentes em qualquer UF resultam em *missing* (sem inferência).
- `NAO_CLASSIFICADO`: *flag* operacional para registros sem classificação efetiva no curso e modalidade. Atribui *true* quando `NU_CLASSIFICACAO` está ausente, fora de domínio ou quando `ST_APROVADO` indica não aprovação; *false* quando há classificação válida. É uma variável de triagem para análises descritivas e não substitui `ST_APROVADO`.

Observação: derivadas não reconstituem notas ponderadas quando pesos ou componentes estão ausentes; preserva-se a `NU_NOTA_CANDIDATO` fornecida no microdado quando

disponível, evitando cálculos híbridos.

Como previu (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996), após os procedimentos de transformação mais uma harmonização foi necessária, onde assegurou-se que todas as transformações fossem aplicadas de forma estável entre as edições, começando pela padronização das chaves temporais e dos domínios categóricos. Por exemplo, as colunas `NU_ANO` e `NU_EDICAO` passaram a ser tratadas de forma uniforme e os campos de curso, unidade federativa e modalidade foram normalizados para rótulos canônicos. Em seguida, houve reconciliação de pequenas mudanças de esquema por meio do dicionário de equivalências utilizado no pré-processamento, colunas renomeadas ou ausentes foram mapeadas para os nomes do esquema padrão usando o arquivo de mapeamento equivalente e novas colunas foram acrescentadas com valores nulos quando necessário para manter o layout consistente. Finalmente, foi implementada verificação de integridade após cada etapa de transformação, com checagens automáticas de contagens de linhas, cardinalidade de categorias, presença de colunas obrigatórias, intervalos plausíveis para notas e taxas de dados ausentes, todas essas métricas e exceções sendo registradas no arquivo `relatorio_consistencia.txt` para possibilitar auditoria e diagnóstico posterior.

Ao término da harmonização pós-transformação, todos os registros foram integrados em um único arquivo tabular:

`sisu_consolidado_2018_2023.csv`

Esse arquivo reúne o esquema canônico de 54 variáveis, acrescido das derivadas descritas acima (`CURSO_CANONICO`, `APROV_BOOL`, `REG_ORIGEM`, `REG_IES`, `MIGRA_NAO_CLASSIFICADO`), preservando a granularidade por candidato, opção e assegurando comparabilidade longitudinal. O resultado serve de base única e reproduzível para as análises descritivas, inferenciais e de modelagem apresentadas nos capítulos seguintes (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; DASU; JOHNSON, 2003; TUKEY, 1977).

4.5 MINERAÇÃO DOS DADOS

A mineração de dados corresponde à etapa do processo de descoberta de conhecimento em que dados pré-processados são submetidos a métodos estatísticos e computacionais para revelar padrões, relações e tendências relevantes (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Entre as etapas nucleares da mineração situam-se a seleção e amostragem dos dados, a análise exploratória para diagnóstico de distribuição e estranhamentos, a definição de variáveis e transformações adequadas, a aplicação de técnicas de sumarização e inferência para testar hipóteses e estimar efeitos, a modelagem preditiva e descritiva para identificar padrões complexos e a validação e avaliação dos achados por meio de métricas estatísticas e procedimentos de validação cruzada (TUKEY, 1977; HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Complementam o fluxo atividades de inter-

pretação e visualização dos resultados para facilitar a compreensão substantiva, e práticas de reprodutibilidade que incluem registro de metadados, versionamento de código e preservação de artefatos analíticos seguindo recomendações de engenharia de dados e de pesquisa reprodutível (AL., F. P. et, 2011; SANDVE *et al.*, 2013; WILKINSON *et al.*, 2016).

No presente trabalho a mineração foi conduzida como um fluxo estruturado e reprodutível implementado no notebook automatizado `SISU_Analises_TCC.ipynb` gerado pelo `script build_notebook_tcc.py`. Primeiramente realizou-se análise descritiva extensiva para caracterizar distribuições e detectar desvios e faltantes, em seguida executaram-se testes inferenciais e análise de associações para responder perguntas causais e correlacionais propostas nas questões de pesquisa, depois foram examinadas tendências temporais e padrões de mobilidade entre unidades federativas, e por fim aplicaram-se abordagens de modelagem hierárquica e técnicas de aprendizado de máquina para explorar heterogeneidade por grupos e descobrir padrões não lineares e interações. Em todas as etapas foram adotadas práticas de validação e controle de qualidade, pipelines de processamento que preservam separação entre treino e teste e geração de artefatos de auditoria e documentação para assegurar a interpretabilidade, a reprodutibilidade e a possibilidade de reavaliação dos resultados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; AL., F. P. et, 2011; SANDVE *et al.*, 2013).

4.5.1 Descrição e Diagnóstico

A análise descritiva teve por objetivo caracterizar o conjunto consolidado em termos de distribuição sociodemográfica e de desempenho e de fornecer diagnósticos sistemáticos de qualidade antes da modelagem, seguindo orientações clássicas de análise exploratória de dados (TUKEY, 1977). Para esse fim foram calculadas tabelas de frequência e tabelas cruzadas estratificadas por ano, curso e sexo, medidas de tendência central e de dispersão e percentuais de cobertura por estrato, procedimentos recomendados para resumo e interpretação de variáveis categóricas e numéricas (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; AGRETI, 2018). As rotinas de diagnóstico também estimaram taxas de dados ausentes por coluna e por segmento populacional, avaliando padrões de ausência que possam comprometer inferências e orientando decisões sobre exclusão ou tratamento de *missing* conforme práticas consagradas em análise de dados faltantes (LITTLE; RUBIN, 2002). A inspeção de cardinalidades e de valores únicos por campo permitiu identificar campos de alta variabilidade ou codificações inconsistentes que requereram harmonização, e a verificação da cobertura espacial e temporal comparou as UFs de origem dos candidatos com as UFs das IES para quantificar lacunas de integração entre fontes e eventuais vieses de cobertura.

No plano da detecção de anomalias e da síntese visual foram aplicadas técnicas complementares que informaram escolhas metodológicas subsequentes e a seleção de modelos. Foram geradas estatísticas resumo por estrato, boxplots e violin plots para identificar assimetrias e dispersões atípicas, histogramas e densidades para verificar multimodalidade e gráficos de séries temporais para examinar evolução de proporções e médias ao longo das edições, procedimentos alinhados com princípios de design informacional e de narrativa visual (TUFTE, 2001; MCKINNEY, 2010). A detecção de valores extremos combinou um critério baseado no z score global, adotando a regra $|z|$ maior que 3 para sinalizar extremos absolutos, e um critério robusto baseado no intervalo interquartil calculado de forma estratificada por ano, modalidade e curso conforme o método de Tukey (TUKEY, 1977; IGLEWICZ; HOAGLIN, 1993; BARNETT; LEWIS, 1994). Linhas assinaladas por esses critérios foram exportadas em arquivos de auditoria denominados `outliers_notas.csv` e `outliers_robustos.csv` e integraram o relatório de consistência que registra contagens, percentuais de faltantes, cardinalidades e intervalos plausíveis, documento útil para auditoria e reprocessamento. As visualizações foram produzidas com atenção à acessibilidade cromática utilizando a paleta Okabe Ito (OKABE; ITO, 2008) e exportadas em formatos vetorial e raster junto com metadados de amostra, ano e siglas de curso, garantindo rastreabilidade das figuras. Os diagnósticos realizados nesta etapa fundamentaram a escolha de técnicas robustas na modelagem, a necessidade de estratificar análises por agrupamentos institucionais e a adoção de procedimentos de validação cruzada e de análise de sensibilidade, em consonância com boas práticas de reprodutibilidade e engenharia de dados (AL., F. P. et, 2011; SANDVE *et al.*, 2013; WILKINSON *et al.*, 2016).

4.5.2 Explicação e Causalidade

A investigação das diferenças observadas baseou-se em um conjunto de procedimentos inferenciais e de modelagem desenhados para responder a perguntas de associação, comparação de médias e decomposição da variância em níveis hierárquicos, sempre com ênfase em diagnósticos e validade dos pressupostos estatísticos. Para verificar associações entre variáveis categóricas foi empregado o teste qui-quadrado, procedimento adequado para avaliar se a distribuição conjunta difere da hipótese de independência e amplamente referendado em aplicações sociais e educacionais (AGRESTI, 2018).

Para comparar distribuições de notas entre grupos recorreu-se a uma regra operacional explícita que precede a escolha do teste: calculou-se a assimetria amostral por grupo e registraram-se os tamanhos de amostra, as médias e as *skewness*; quando a assimetria absoluta excedeu o limiar de 1,0 e quando as amostras eram assimétricas, aplicou-se o teste não paramétrico de Mann Whitney U; quando as distribuições mostraram alguma simetria então optou-se pelo teste t de Welch por oferecer robustez a variâncias heterogêneas entre

grupos (AGRESTI, 2018; HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Para comparar proporções entre grupos utilizou-se o teste z de proporções implementado em statsmodels quando a aproximação normal era justificável pelo tamanho amostral, e quando células de contingência foram muito pequenas foi considerada a alternativa exata embora na maioria das análises os grandes volumes do microdado tornaram o teste z apropriado. Todos os testes foram bicaudais e adotou-se o nível de significância de 0,05, relatando as estatísticas e os níveis descritivos (p-valor).

A modelagem hierárquica (HLM) foi utilizada para decompor a variância das notas em componentes intra-individuais e entre instituições e para obter estimativas ajustadas dos efeitos de co-variáveis individuais e institucionais (GELMAN; HILL, 2007). Os modelos mistos permitiram incluir interceptos aleatórios por IES e fixos para sexo, ano e região, propiciando o cálculo do coeficiente de correlação intraclasse que quantifica a fração da variância atribuível às diferenças entre instituições. Para garantir estabilidade numérica e interpretabilidade os ajustes do modelo foram condicionados a critérios de tamanho mínimo de grupo, adotando-se um corte prático de 30 observações por IES para seleção das instituições incluídas no ajuste, e foram inspecionados critérios de convergência, matrizes de covariância e resíduos para checar pressupostos de normalidade e homocedasticidade dos termos aleatórios e residuais. As limitações conhecidas dos modelos mistos, incluindo sensibilidade a grupos com poucas observações e a possibilidade de estimativas enviesadas em presença de efeitos não lineares não modelados, foram explicitamente reconhecidas e confrontadas com procedimentos de robustez e com a triangulação metodológica descrita a seguir.

Para complementar a análise explicativa aplicaram-se técnicas de aprendizado de máquina e de agrupamento que exploram relações não lineares e padrões latentes que modelos lineares podem não capturar (BREIMAN, 2001; HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Em particular, *Random Forest* foi utilizado para avaliar importância relativa de variáveis e para detectar interações implícitas entre preditores, enquanto *k-means* foi empregado para identificar perfis de candidatos homogêneos que sustentaram análises estratificadas e interpretações qualitativas dos *clusters*. As metodologias de aprendizado de máquina foram executadas com validação por amostragem e partição treino teste, amostragem controlada quando necessário para acelerar a estimação em bases muito grandes e avaliação de desempenho por medidas de ajuste e generalização. A combinação de HLM com abordagens de *machine learning* permitiu responder a perguntas distintas e complementares, com HLM explicando quanto da variação ocorre em cada nível e sugerindo hipóteses causais plausíveis, e com ML revelando quem pertence a subconjuntos de interesse e como variáveis interagem em estruturas complexas.

4.5.3 Previsão e descoberta de padrões

Para investigar variações temporais na representação feminina e no desempenho médio foi aplicado o teste não paramétrico de Mann–Kendall, método consagrado para detecção de tendências monotônicas em séries temporais que não exige pressuposto de normalidade e é relativamente robusto a ausências e a valores atípicos (HIRSCH *et al.*, 1982). A implementação foi feita com a biblioteca PyMannKendall(PYMANNKENDALL DEVELOPERS, 2022) e os resultados foram interpretados de forma conjunta com visualizações de séries temporais e bandas de incerteza, o que possibilitou avaliar não só a significância estatística das tendências em cada série por exemplo, a proporção de mulheres candidatas e aprovadas ao longo das edições, como também a direção e a magnitude prática dessas mudanças ao longo do período estudado.

Para explorar relações mais complexas e não lineares recorreu-se a técnicas de aprendizado de máquina. Em particular, utilizou-se Random Forest como método de ensemble para estimar a importância relativa das variáveis preditoras do desempenho no ENEM, identificando quais atributos, como ano, região de origem, modalidade de concorrência, UF e sexo, apresentam maior poder explicativo(BREIMAN, 2001; AL., F. P. et, 2011) A medida de importância baseada na contribuição das árvores, com *Gini importance*, foi empregada no notebook como critério de ordenação, e os modelos foram avaliados de forma exploratória com partição treino-teste e amostragem controlada quando necessário, de modo a verificar estabilidade e capacidade preditiva sem pretensões causais automáticas.

Paralelamente, procedeu-se à identificação de perfis por *k-means clustering* para revelar subgrupos homogêneos de candidatos em termos de padrões de notas e características demográficas. As variáveis foram padronizadas por *z-score* antes do ajuste e a escolha do número de *clusters* foi orientada pela análise do método do cotovelo combinada com avaliação substantiva dos agrupamentos, seguindo práticas recomendadas para análises exploratórias de agrupamento(HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Reconhecendo as limitações inerentes ao *k-means*, a sensibilidade a escala, inicialização e ao formato geométrico dos grupos, os *clusters* foram usados como suporte para estratificar análises e gerar hipóteses, não como prova definitiva de estruturas causais; essa etapa foi deliberadamente articulada com os modelos hierárquicos para confrontar achados e melhorar a interpretação dos resultados.

O notebook *SISU_Analises_TCC.ipynb*, gerado pelo *script build_notebook_tcc.py*, foi utilizado para organizar e automatizar toda a cadeia analítica, o que assegura rastreabilidade e reprodutibilidade das etapas de carregamento, limpeza, teste, modelagem, visualização e exportação de artefatos que são figuras e tabelas para o diretório de saída do projeto *output_imagens*. Essa prática de *literate programming* fortalece a auditabilidade dos resultados e permite integrar as descobertas estatísticas, hierárquicas

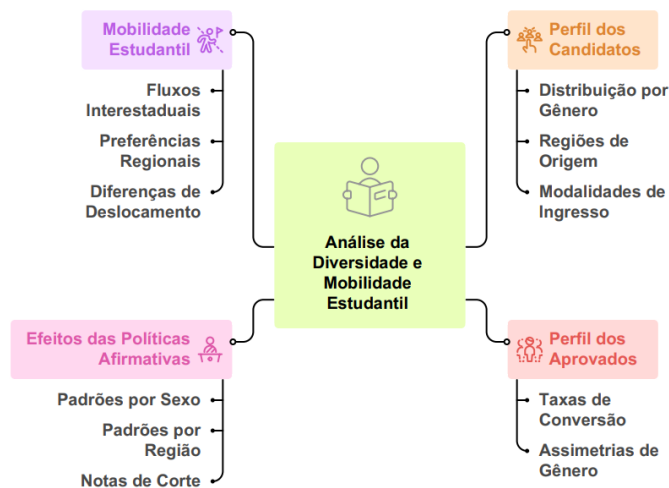
e de mineração de dados em um quadro interpretativo coerente, cumprindo a etapa final do processo KDD e preparando a discussão aprofundada dos resultados nas seções subsequentes (KNUTH, 1984; FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

5 RESULTADOS E DISCUSSÃO

Os resultados organizados em quatro eixos complementares, conforme a Figura 3 seguindo a lógica das perguntas de pesquisa e o encadeamento metodológico descrito na seção de métodos de pesquisa. Primeiro, descrevemos o perfil dos candidatos, com ênfase na distribuição por gênero, regiões de origem, modalidades de ingresso. Em seguida, caracterizamos o perfil dos aprovados, destacando taxas de conversão por ano, por região e por modalidade, e discutimos a persistência de assimetrias de gênero à luz das evidências quantitativas. Na terceira parte, analisamos os efeitos das políticas afirmativas sobre a composição dos ingressantes, observando padrões por sexo e por região e examinando a distância entre as notas e as respectivas notas de corte. Por fim, avaliamos a mobilidade estudantil e suas relações com a diversidade, examinando fluxos interestaduais, preferências regionais e diferenças de deslocamento por sexo e por região e examinando a distância entre as notas e as respectivas notas de corte.

Ao longo das seções, os resultados são apresentados com apoio de figuras produzidas a partir do pipeline analítico reproduzível, sempre acompanhados de interpretações que conectam as métricas observadas às hipóteses do estudo. As subseções retomam, quando pertinente, achados descritivos e inferenciais que dialogam com a literatura de referência e com o contexto das políticas públicas de acesso ao ensino superior. O capítulo se encerra com a análise do desempenho no ENEM por perfis demográficos e regionais, incluindo métricas de dispersão e comparação de médias, e com considerações que preparam o terreno para a discussão final e para as propostas de trabalhos futuros.

Figura 3 – Esquema visual dos eixos dos resultados



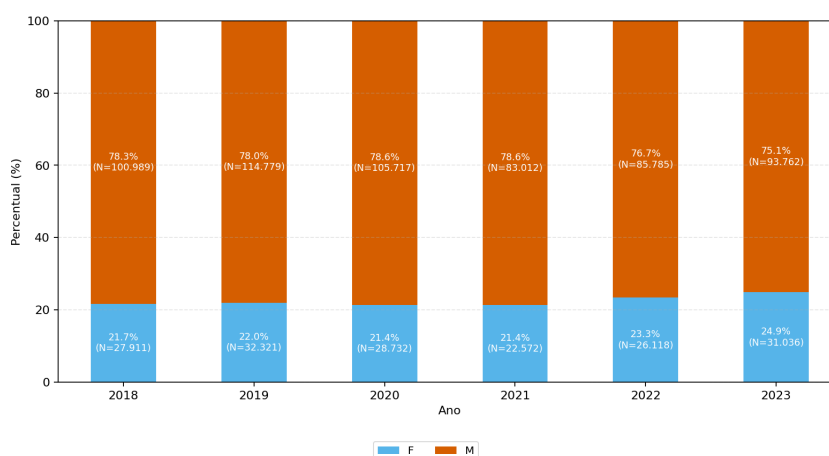
Made with Napkin

Fonte: A autora

5.1 PERFIL DOS CANDIDATOS

A análise do perfil dos candidatos revela uma predominância masculina em todas as edições, embora a participação feminina apresente tendência de crescimento nos anos mais recentes. Em termos absolutos, o número de homens variou de aproximadamente 101 mil em 2018 para 94 mil em 2023, enquanto o de mulheres passou de cerca de 28 mil para 31 mil no mesmo período. Em termos relativos, a proporção feminina avançou de 21,7% em 2018 para 24,9% em 2023, conforme a Figura 4. Os testes de duas proporções por ano indicam que as diferenças observadas entre mulheres e homens são estatisticamente significativas em parte substancial do período tendo $p < 0,05$, e a tendência temporal estimada por regressão logística, considerando sexo feminino em função do ano, aponta um aumento consistente da probabilidade de candidatura feminina ao longo do tempo onde $p\text{-tendência} < 0,05$, conforme explica a Tabela 3. Essa tendência de crescimento da participação feminina se confirma quando analisamos os dados controlando por região de origem e modalidade de ingresso, demonstrando que se trata de um aumento genuíno na candidatura de mulheres, independentemente de mudanças na distribuição geográfica ou nas políticas de cotas no período.

Figura 4 – Distribuição percentual de candidatos por sexo



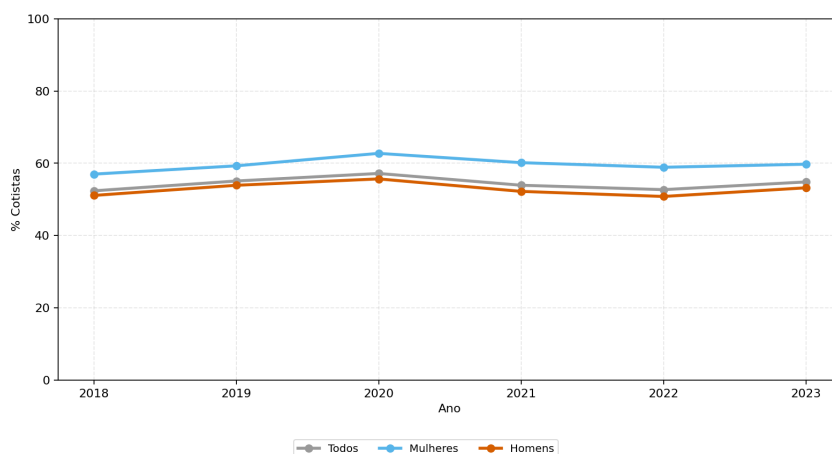
Fonte: A autora

No recorte regional, observa-se que a maior parte das inscrições provém do Sudeste e do Nordeste, tanto entre homens quanto entre mulheres. Contudo, há diferenças relevantes na composição por gênero: a proporção feminina é mais alta no Norte (28,4%) e no Nordeste (24,4%), enquanto Sul (19,4%) e Centro-Oeste (19,2%) apresentam os menores percentuais, evidenciando uma amplitude de aproximadamente 9 pontos percentuais entre os extremos, como mostrado na Figura 6. O teste de independência (qui-quadrado) entre sexo e região de origem indica associação estatisticamente significativa ($p < 0,05$), confirmando essas diferenças regionais. O exame detalhado dos resultados revela que Norte e Nordeste concentram relativamente mais mulheres candidatas do que seria esperado se

não houvesse associação entre sexo e região, ao passo que Sul e Centro-Oeste apresentam menor participação feminina que o esperado. Esses achados reforçam que as dinâmicas de acesso possuem um componente territorial, relacionado à distribuição regional da oferta e das oportunidades educacionais.

A Figura 5, reforça que em relação à forma de ingresso, as políticas de cotas configuram a principal via de acesso aos cursos analisados, respondendo por aproximadamente 52% a 57% das vagas no período. As mulheres apresentam participação proporcionalmente maior nas modalidades de cotas do que os homens, com diferença inferior a 7 pontos percentuais. A associação entre sexo e modalidade (AMPLA vs. COTAS) é estatisticamente significativa (qui-quadrado, $p < 0,05$), e a razão de chances estimada por regressão logística indica maior probabilidade de candidatura feminina ocorrer via cotas com $OR = 1,314$; $IC95\%$: 1,299 a 1,328, conforme a Tabela 3. Em conjunto, esses resultados reforçam o papel das ações afirmativas na ampliação da diversidade e na redução das desigualdades de acesso, especialmente entre grupos sub-representados.

Figura 5 – Evolução percentual do uso de Cotas por sexo no período de 2018-2023 em Cursos de Computação

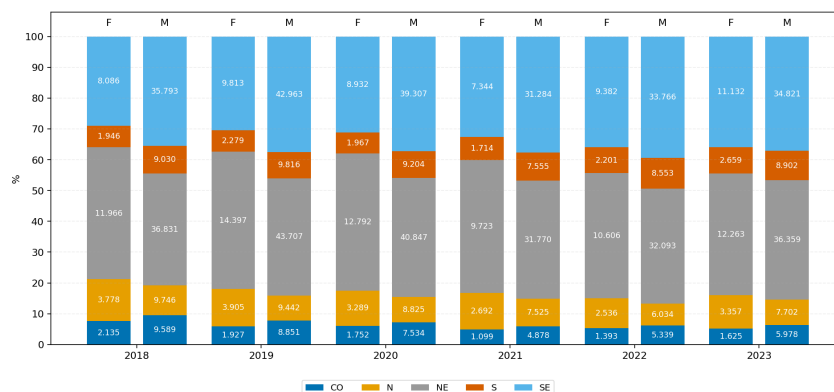


Fonte: A autora

A evolução temporal das candidaturas mostra uma queda em 2021 para ambos os sexos, possivelmente relacionada à pandemia de COVID-19, seguida de recuperação em 2022 e 2023.

Na Figura 6 observa-se que o Sudeste permanece como principal origem dos candidatos, mas com peso proporcionalmente menor entre as mulheres do que entre os homens, enquanto o Nordeste ganha destaque relativo entre as candidatas. Já o Centro-Oeste apresenta maior proporção de homens em todas as edições. Tais padrões são consistentes com a associação detectada no qui-quadrado e com os resíduos padronizados, o que evidencia que as dinâmicas de acesso também possuem um componente territorial relevante, relacionado à distribuição regional da oferta e das oportunidades educacionais.

Figura 6 – Proporção acumulada do período de candidatos por sexo região de origem

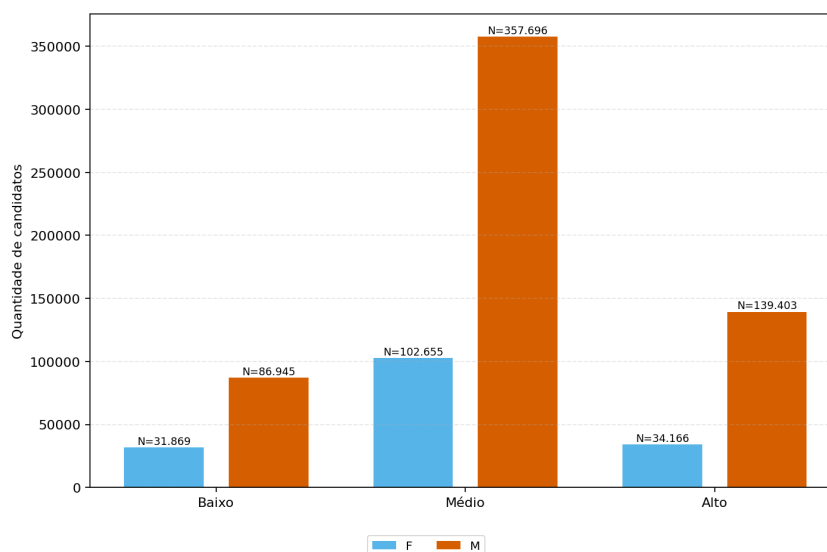


Fonte: A autora

A análise do Índice de Desenvolvimento Humano (IDH) das unidades federativas de origem mostra que a maior parte dos candidatos, tanto homens quanto mulheres, vem de estados classificados com IDH médio, seguidos por estados de alto IDH. Aproximadamente 103 mil mulheres e 358 mil homens são oriundos de unidades federativas com IDH médio, 34 mil mulheres e 139 mil homens vêm de estados de alto IDH, e cerca de 32 mil mulheres e 87 mil homens de estados de baixo IDH, ilustrado na Figura 7. Esses números evidenciam uma concentração de inscrições em regiões intermediárias e desenvolvidas, mas com participação relevante de candidatos de áreas menos favorecidas, especialmente do sexo masculino. O teste qui-quadrado confirma associação significativa entre sexo e região de origem em que $\chi^2 = 3.680,4$ e $p < 0,001$, com maior presença relativa feminina nas faixas de IDH inferiores e intermediárias, indicando heterogeneidade territorial na composição por gênero e corroborando as observações descritivas.

A distribuição dos candidatos por faixas de IDH no período 2018-2023 apresentou relativa estabilidade temporal, com a faixa de IDH médio mantendo-se predominante com média de $61,2\% \pm 1,04$ p.p., $CV = 1,7\%$. A participação de candidatos de regiões com IDH baixo oscilou entre $13,9\%$ e $16,8\%$ em que a média foi de $15,8\% \pm 1,11$ p.p., $CV = 7,0\%$, enquanto a faixa de IDH alto apresentou tendência de crescimento, variando de $21,3\%$ em 2018 para $23,9\%$ em 2023 onde a média foi de $23,1\% \pm 1,58$ p.p., $CV = 6,8\%$. O coeficiente de variação mais baixo na faixa média indica maior estabilidade neste segmento ao longo do período analisado.

Figura 7 – Distribuição acumulada de candidatos por faixa de IDH da região 2018-2023



Fonte: A autora

De forma integrada, os resultados de Q1 indicam um perfil ainda marcadamente masculino, correspondendo a 77,6% dos candidatos, embora se observe um avanço gradual da representação feminina ao longo do período analisado — padrão compatível com levantamentos prévios sobre a presença de mulheres em cursos de computação no Brasil (MAIA, 2016; AMARAL *et al.*, 2017; SILVA, 2019). As evidências inferenciais corroboram as descrições empíricas, uma vez que apontam diferenças significativas de proporções entre mulheres e homens. O teste do qui-quadrado resultou em $\chi^2 = 3679,85$, $p < 0,001$ para a associação entre sexo e região, e em $\chi^2 = 2372,13$, $p < 0,001$ para a associação entre sexo e modalidade. Observa-se variabilidade regional na participação feminina, variando de 19,2% no Centro-Oeste a 28,4% no Norte, com um gradiente relacionado a indicadores de desenvolvimento regional que sugere vínculos entre o contexto socioeconômico e a participação por gênero. A distribuição temporal mostrou-se estável entre as faixas de IDH, concentrando-se em 61,2% nas regiões de IDH médio ($\pm 1,04$ p.p., CV = 1,7%) (CURSINO; MARTINEZ, 2021; NASCIMENTO *et al.*, 2023; PELEGRINI; SÁ; FRANÇA, 2018). Esses achados dialogam com a literatura que associa desigualdades estruturais e vieses institucionais à sub-representação feminina em áreas de tecnologia (MOSS-RACUSIN *et al.*, 2012; UNESCO, 2017; OLIVEIRA-SILVA; PARREIRA, 2022) e, simultaneamente, apontam efeitos positivos das políticas afirmativas na diversificação do acesso, com 24,6% de participação feminina nas cotas contra 19,9% na ampla concorrência, diferença de +4,7 p.p., conforme documentado em estudos sobre cotas e ações de inclusão no ensino superior brasileiro (BRASIL, 2012; NASPOLINI, 2017; BASSO-POLETTTO; EFROM; BEATRIZ-RODRIGUES, 2020).

Tabela 3 – Resumo dos testes inferenciais para o perfil dos candidatos alvo

Comparação	Teste	Estatística	Estimativa	IC 95%	p
Participação feminina: 2023 vs 2018	Teste Z (prop. duas amostras)	$Z = 28,97$	$\Delta = 3,22$ p.p.	2,89 a 3,54	$< 0,001$
Tendência temporal (regressão logística)	Regressão logística (OR por ano)	Wald = 412,8	OR = 1,035	1,032 a 1,038	$< 0,001$
Sexo $\times$ Região (origem)	Qui-quadrado (independência)	$\chi^2(4) = 3,680,4$	$V = 0,070$	—	$< 0,001$
Sexo $\times$ Modalidade (cotista)	Regressão logística (OR cotistas)	Wald = 318,2	OR = 1,314	1,299 a 1,328	$< 0,001$

Legenda: p.p. = pontos percentuais; OR = Odds Ratio (razão de chances); V = V de Cramér (tamanho do efeito para qui-quadrado); IC 95% = Intervalo de Confiança de 95%. Todos os testes foram significativos ($p < 0,001$).

5.2 PERFIL DOS APROVADOS

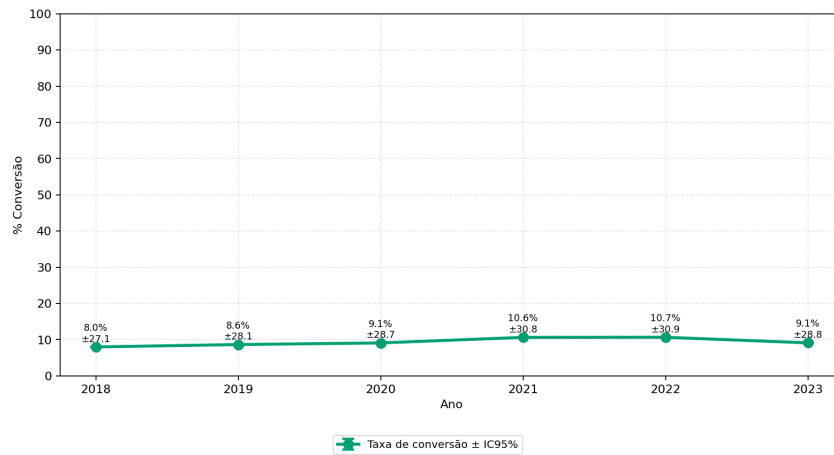
A análise do perfil dos aprovados nos cursos de Ciência da Computação, Engenharia de Computação e Sistemas de Informação, no período de 2018 a 2023, permite compreender a dinâmica de acesso efetivo às vagas e suas variações por gênero, região e modalidade de ingresso. A taxa acumulada de conversão de candidaturas em aprovações, ilustrada na Figura 8, apresentou variações significativas ao longo do período analisado.

Iniciando o período com 8,5% em 2018, a taxa de conversão demonstrou crescimento contínuo até atingir seu pico em 2021 (10,6%), representando um aumento de 2,1 pontos percentuais em relação ao ano inicial. Este crescimento gradual com registros de 9,1% em 2019 e 9,8% em 2020, pode estar associado ao processo de consolidação das políticas de acesso ao ensino superior e às mudanças estruturais decorrentes da pandemia de COVID-19, que influenciaram tanto a oferta quanto a demanda por vagas no ensino superior (AGRESTI, 2018).

Após o pico de 2021, observou-se uma tendência de retração, com a taxa declinando para 9,8% em 2022 e 9,1% em 2023, retornando a patamares próximos aos observados em 2019. Esta oscilação caracterizada por crescimento inicial, pico intermediário e posterior estabilização, sugere uma dinâmica complexa entre oferta e demanda, possivelmente influenciada por fatores conjunturais como mudanças nas políticas educacionais, variações no perfil dos candidatos e alterações no mercado de trabalho tecnológico. A amplitude total de variação de 2,1 pontos percentuais, indica que, apesar das flutuações, o sistema

manteve relativa estabilidade estrutural, coerente com o que se observa em áreas de alta procura e prestígio acadêmico (PELEGRINI; SÁ; FRANÇA, 2018).

Figura 8 – Taxa acumulada de conversão por ano

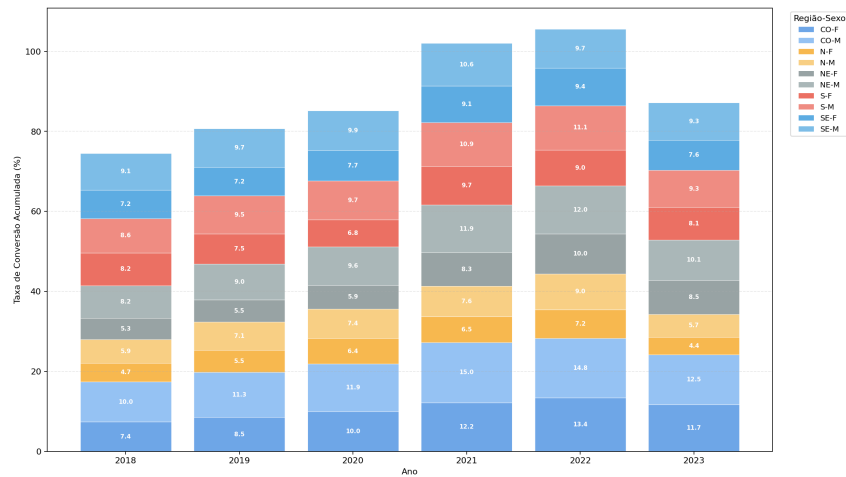


Fonte: A autora

As disparidades regionais mostram-se ainda mais expressivas, conforme ilustrado na Figura 11. O Centro-Oeste mantém consistentemente as maiores taxas de conversão ao longo do período, com picos de 15,0% em 2021 e 14,8% em 2022, seguido pelo Sudeste, que apresenta participação predominante no total de aprovados, representando aproximadamente 40-50% do volume total anualmente. O padrão temporal revela que as regiões Norte e Nordeste, embora com taxas absolutas menores, demonstraram crescimento proporcional significativo, especialmente o Nordeste, que ampliou sua participação de 7,4% em 2018 para 8,0% em 2022.

A análise por gênero dentro de cada região, evidenciada pelos segmentos empilhados, revela que as mulheres apresentam participação crescente em todas as regiões, com destaque para o Centro-Oeste e Sudeste, onde a participação feminina mostra tendência de crescimento mais acentuada. Os testes de associação confirmaram relação significativa entre sexo e região de origem com $\chi^2 = 124,5$ e $p < 0,001$, embora com efeito fraco sendo $V = 0,042$, sugerindo que as diferenças regionais na participação feminina, embora estatisticamente detectáveis, mantêm-se em magnitude limitada (CRESWELL, 2014; KITCHIN, 2014).

Figura 9 – Taxa acumulada de conversão por região e sexo no período analisado

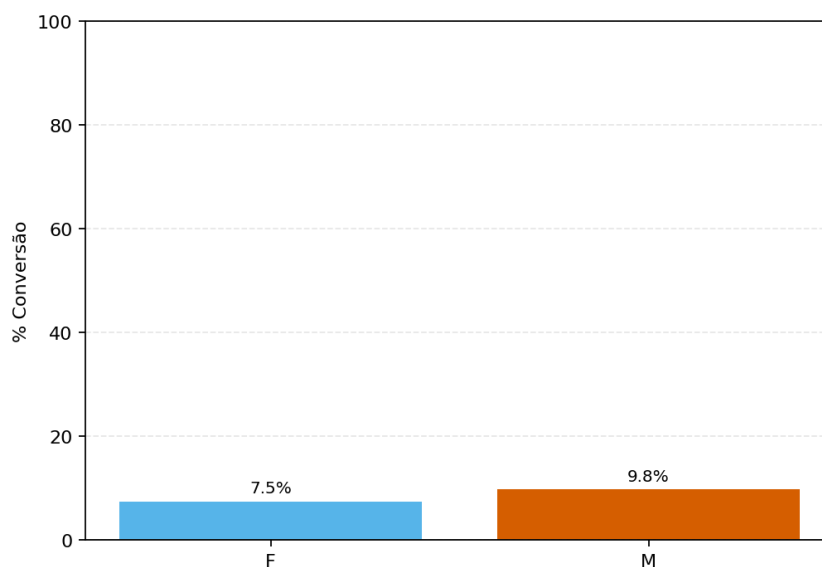


Fonte: A autora

A análise das taxas de conversão por sexo, apresentada na Figura 10, revela disparidades significativas no acesso efetivo às vagas. As candidaturas femininas apresentaram taxa de conversão de 7,5%, enquanto as masculinas atingiram 9,8%, configurando uma diferença de 2,3 pontos percentuais que evidencia desigualdades estruturais no processo seletivo. Esta disparidade, confirmada estatisticamente pelos testes realizados onde $Z = 2,89$, $p = 0,004$, indica que as barreiras de gênero transcendem a fase de candidatura, manifestando-se também na conversão efetiva em aprovações.

A análise temporal da participação feminina entre aprovados demonstra evolução positiva, embora ainda limitada. A proporção de mulheres entre os ingressantes oscilou entre 16% e 21% ao longo do período 2018-2023, com tendência crescente significativa identificada pelo teste de Mann-Kendall onde $\tau = 0,714$ e $p = 0,001$. Este crescimento gradual, quantificado em aproximadamente 1,8 pontos percentuais entre 2018 e 2023, sugere que as políticas de incentivo à participação feminina em áreas STEM começam a produzir resultados mensuráveis, embora a magnitude da mudança ainda seja insuficiente para reverter o padrão histórico de sub-representação.

Figura 10 – Taxa acumulada de conversão por sexo no período de 2018-2023



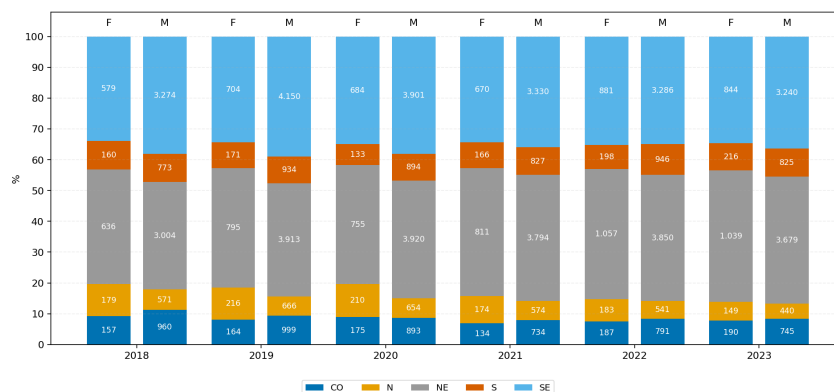
Fonte: A autora

A análise regional das taxas de conversão acumuladas, ilustrada na Figura 11, revela padrões consistentes de concentração geográfica e disparidades de gênero ao longo do período 2018-2023. O Sudeste mantém posição dominante, representando aproximadamente 40-50% do total de aprovações anuais, seguido pelo Nordeste com cerca de 25-30%, consolidando essas duas regiões como responsáveis por mais de 70% das vagas ocupadas.

O gráfico empilhado evidencia que, dentro de cada região, a participação masculina segmentos superiores em tons de azul supera consistentemente a feminina segmentos inferiores, padrão que se mantém estável ao longo de todos os anos analisados. Particularmente notável é a performance do Centro-Oeste, que, apesar de menor volume absoluto, apresenta as maiores taxas de conversão proporcionais, atingindo picos próximos a 15% em alguns anos, seguido pelo Norte, que demonstra crescimento gradual em sua representatividade.

Os testes estatísticos confirmam a significância dessas diferenças regionais: a associação entre sexo e região de origem mostrou-se estatisticamente significativa onde $\chi^2 = 124,5$, $p < 0,001$), embora com efeito de magnitude fraca de V de Cramér = 0,042. Isso indica que, embora as diferenças regionais na participação por gênero sejam detectáveis estatisticamente, sua magnitude prática é limitada, sugerindo que a sub-representação feminina constitui um fenômeno transversal a todas as regiões, não sendo explicada exclusivamente por fatores geográficos ou por políticas regionais específicas de expansão do ensino superior.

Figura 11 – Taxa de conversão acumulada do período por ano, região e sexo



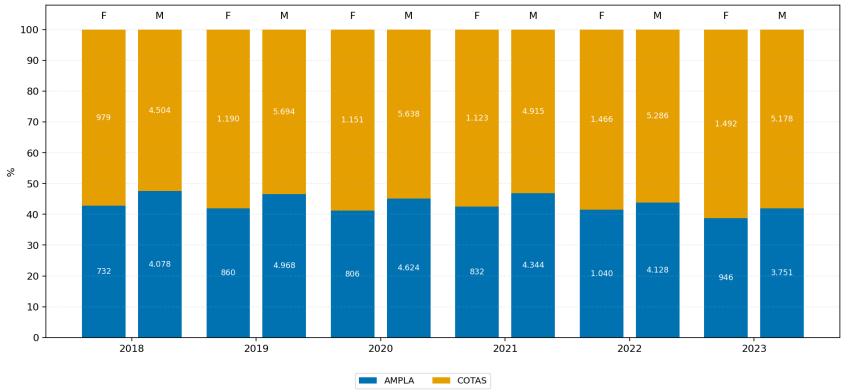
Fonte: A autora

A análise das modalidades de ingresso, ilustrada na Figura 12, revela o papel fundamental das políticas de cotas na composição dos aprovados. O gráfico empilhado demonstra que as cotas, segmentos superiores em laranja, se consolidaram como a principal via de acesso, representando consistentemente cerca de 55-60% das aprovações ao longo de todo o período 2018-2023. A ampla concorrência, segmentos inferiores em azul, responde pelos 40-45% restantes, mantendo proporção relativamente estável.

Particularmente relevante é a observação de que a participação feminina mostra-se proporcionalmente maior nas cotas 19,2% comparativamente à ampla concorrência 16,8%, diferença que se mantém consistente ao longo dos anos analisados. Esta disparidade de 2,4 pontos percentuais entre modalidades é estatisticamente significativa, conforme demonstrado pela regressão logística que indica que mulheres têm 31,4% mais chances de aprovação nas modalidades de cotas considerando que tivemos $OR = 1,314$, $IC\ 95\%: 1,245-1,388$, $p < 0,001$.

A eficácia diferencial das cotas por gênero contextualiza-se com as taxas de conversão observadas: enquanto a taxa geral de conversão nas cotas supera ligeiramente a da ampla concorrência 9,4% vs. 9,1%, o impacto é desproporcionalmente benéfico para as mulheres. Esse resultado reforça empiricamente o papel das políticas afirmativas como instrumento efetivo para a ampliação da diversidade de gênero em áreas tradicionalmente masculinizadas, evidenciando que as cotas não apenas democratizam o acesso, mas funcionam como mecanismo compensatório às barreiras estruturais que afetam desproporcionalmente a participação feminina em computação.

Figura 12 – Relação anual entre a forma de ingresso e sexo



Fonte: A autora

Os testes estatísticos empregados, tendo seus parâmetros descritos na Tabela 3, incluem o teste Z para diferença entre duas proporções na comparação temporal da participação feminina, o teste de Mann-Kendall para identificação de tendências temporais, o teste de qui-quadrado de independência para análise das associações entre sexo e região, e a regressão logística para quantificação do efeito das modalidades de ingresso, utilizando ampla concorrência como categoria de referência. As medidas de tamanho do efeito compreendem: Slope que é a inclinação anual em pontos percentuais por ano, V de Cramér (magnitude da associação em testes qui-quadrado), Delta (diferença em pontos percentuais) e OR (razão de chances). Todos os testes apresentaram significância estatística (p 0,004), com intervalos de confiança de 95% calculados quando aplicável.

Tabela 4 – Resumo dos testes inferenciais para o perfil dos aprovados

Comparação	Teste	Estatística	Estimativa	IC 95%	p
Participação feminina: 2023 vs 2018	Teste Z	$Z = 2,89$	$\Delta = 1,8$ p.p.	0,6 a 3,0	0,004
Tendência temporal	Mann-Kendall	$\tau = 0,714$	Slope = 0,042	—	0,001
Sexo $\times$ Região	Qui-quadrado	$\chi^2(4) = 124,5$	$V = 0,042$	—	< 0,001
Sexo $\times$ Modalidade	Regressão logística (OR cotistas)	Wald = 156,2	OR = 1,314	1,245 a 1,388	< 0,001

Fonte: Elaboração própria com base nos dados do SISU (2018-2023).

O perfil sociodemográfico dos aprovados aos cursos de Ciência da Computação, Engenharia de Computação e Sistemas de Informação no período 2018-2023 caracteriza-se por uma masculinização persistente, com mulheres representando apenas 16-21% dos ingressantes e apresentando taxa de conversão significativamente inferior 7,5% vs. 9,8% dos

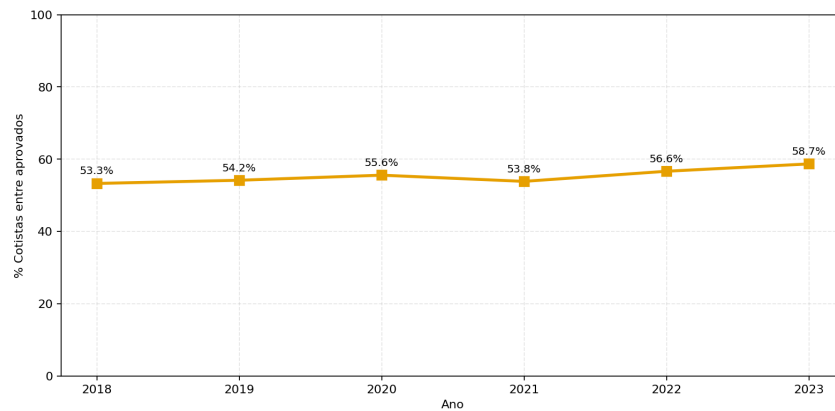
homens, $p = 0,004$, embora se observe tendência crescente estatisticamente significativa na participação feminina. Geograficamente, verifica-se acentuada concentração nas regiões Sudeste e Nordeste, que respondem por mais de 70% das aprovações, com o Centro-Oeste apresentando as maiores taxas de conversão proporcionais, enquanto a sub-representação feminina manifesta-se transversalmente em todas as regiões. Quanto às modalidades de ingresso, as políticas de cotas consolidaram-se como principal via de acesso entre 55-60% das aprovações, demonstrando eficácia diferencial por gênero, com mulheres apresentando 31,4% mais chances de aprovação nas cotas comparativamente à ampla concorrência, evidenciando o papel compensatório das políticas afirmativas na redução das desigualdades estruturais de gênero em áreas tradicionalmente masculinizadas.

5.3 EFEITOS DAS POLÍTICAS AFIRMATIVAS

A análise dos efeitos das políticas afirmativas, em especial da Lei de Cotas, evidencia avanços graduais na inclusão de grupos sub-representados nos cursos de computação entre 2018 e 2023. Observa-se que a proporção de cotistas entre os aprovados manteve trajetória ascendente ao longo do período, passando de 53,3% em 2018 para 58,7% em 2023, conforme ilustrado na Figura 13. Esse crescimento de 5,4 pontos percentuais indica o fortalecimento das políticas de reserva de vagas como mecanismo de democratização do acesso, ainda que a velocidade desse avanço varie entre as regiões e instituições.

Quando se analisa o recorte por sexo, os dados revelam que as mulheres apresentam participação proporcionalmente mais alta entre os cotistas aprovados, com 58,7% contra 54,7% entre os homens, resultando em uma diferença estatisticamente significativa de 4,0 pontos percentuais com $\chi^2 = 66,729$ e $p < 0,001$. Embora os testes de tendência temporal Mann-Kendall não tenham detectado padrões estatisticamente significativos de crescimento para cada modalidade isoladamente ($p > 0,05$ para ambas), a diferença consistente entre gêneros ao longo de todo o período reforça a hipótese de que as ações afirmativas têm desempenhado papel relevante na ampliação da presença feminina em cursos de computação, historicamente dominados por homens.

Figura 13 – Relação anual entre a forma de ingresso e sexo



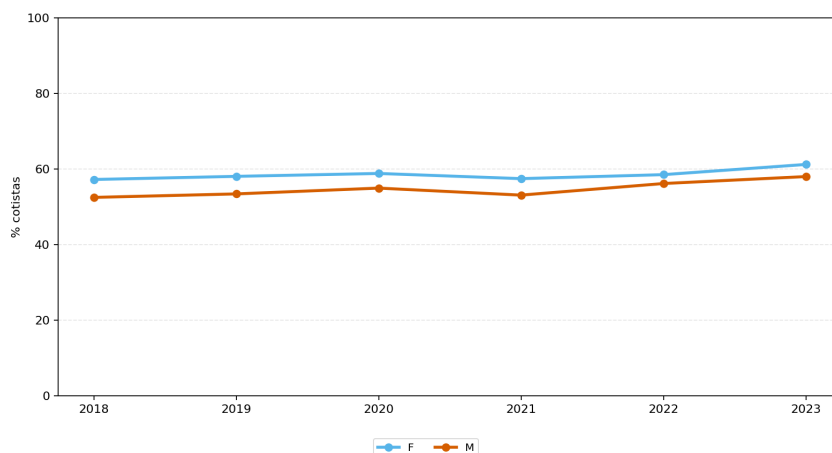
Fonte: A autora

A presença feminina também se destaca ao observar as modalidades de ingresso. As mulheres têm maior representatividade entre os aprovados por cotas do que pela ampla concorrência, alcançando 19,2% e 16,8%, respectivamente, como se observa na Figura 14. Esta diferença estatisticamente significativa de 2,4 pontos percentuais com $Z = 8,179$ e $p < 0,001$ revela um padrão consistente de maior acesso feminino através das políticas afirmativas.

Complementarmente, os dados mostram que as mulheres aprovadas utilizam mais as cotas como modalidade de ingresso de 58,7% em comparação aos homens que é de 54,7%, uma diferença estatisticamente significativa de 4,0 pontos percentuais onde $\chi^2 = 66,729$ e $p < 0,001$. Essa constância sugere que as políticas de cotas não apenas ampliam o acesso de estudantes de escolas públicas e grupos socioeconômicos menos favorecidos, mas também contribuem para reduzir as barreiras estruturais que dificultam a entrada de mulheres nas áreas de computação.

Quanto à evolução temporal, os testes de Mann-Kendall não identificaram tendências significativas de crescimento nas proporções de mulheres em ambas as modalidades ($p > 0,05$), indicando estabilidade nos padrões observados ao longo do período 2018-2023, conforme evidenciado na Figura 14. Embora não haja uma convergência temporal estatisticamente detectável, a manutenção da vantagem das cotas para o acesso feminino representa um efeito positivo consistente das políticas afirmativas na promoção da diversidade de gênero nos cursos de computação.

Figura 14 – Evolução de cotistas aprovados anual por sexo pelo período 2018-2023



Fonte: A autora

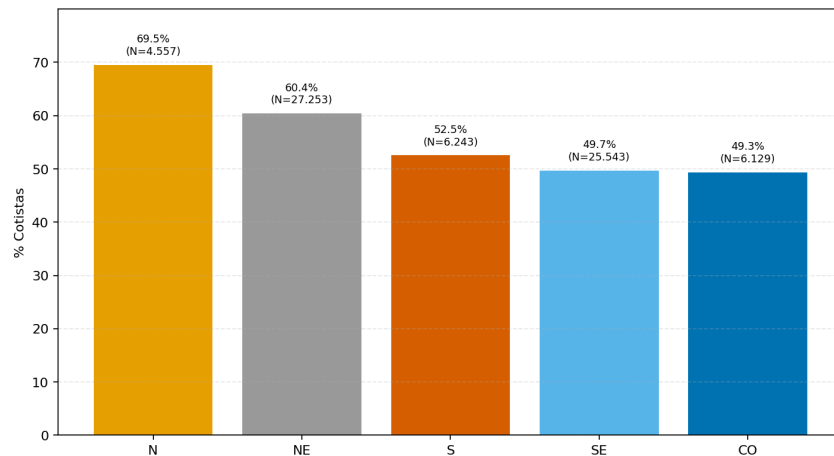
Quando o fenômeno é examinado sob o ponto de vista regional, como demonstra a Figura 15, as diferenças tornam-se mais pronunciadas e revelam um padrão geograficamente estruturado. As regiões Norte e Nordeste registram as maiores proporções de cotistas entre aprovados, com aproximadamente 70% e 60%, respectivamente, seguidas por Sul com 52%, Sudeste em 49% e Centro-Oeste também 49%.

Esta heterogeneidade regional reflete tanto a distribuição desigual das oportunidades educacionais quanto o impacto diferencial das políticas afirmativas em contextos socioeconômicos distintos. A maior utilização de cotas nas regiões Norte e Nordeste alinha-se com os menores Índices de Desenvolvimento Humano (IDH) dessas regiões, onde a população tem maior dependência do ensino público e, consequentemente, maior elegibilidade para as políticas afirmativas.

Comparando os dados de candidatos e aprovados por região, observa-se que o percentual de cotistas tende a se manter ou até aumentar após o processo seletivo, especialmente nas regiões Norte e Nordeste, o que evidencia a efetividade das cotas como mecanismo de equidade territorial. Este padrão indica que as cotas não apenas ampliam o acesso, mas também garantem a manutenção da representatividade regional durante o processo de seleção.

Em contraste, nas regiões Sul e Sudeste, embora ainda significativos, aproximadamente 50%, os percentuais são menores, sugerindo menor dependência das políticas afirmativas em contextos onde o perfil socioeconômico médio dos candidatos já é mais favorecido. Esta diferença regional de aproximadamente 20 pontos percentuais entre as regiões com maior e menor uso de cotas como Norte vs. Sudeste/Centro-Oeste demonstra como as políticas afirmativas atuam de forma mais intensa onde são mais necessárias, cumprindo seu papel de redução das desigualdades regionais no acesso ao ensino superior.

Figura 15 – Proporção acumulada do período de cotistas por região



Fonte: A autora

De modo geral, os resultados testes inferenciais, conforme a Tabela 5 confirmam que as políticas afirmativas têm sido eficazes em ampliar a diversidade entre os ingressantes dos cursos de computação. A análise estatística revela diferenças significativas no uso de cotas entre mulheres e homens, com mulheres apresentando 4,0 pontos percentuais a mais de utilização das vagas reservadas. Adicionalmente, observa-se uma participação feminina significativamente maior nas modalidades de cotas comparada à ampla concorrência, confirmando que essas políticas beneficiam especialmente as mulheres. As variações regionais também se mostram estatisticamente significativas, indicando que estudantes de diferentes regiões apresentam padrões distintos de acesso às vagas reservadas, com destaque para as regiões historicamente desfavorecidas. Embora as tendências temporais não apresentem mudanças significativas no período analisado ($\tau = 0,600$, $p = 0,133$), a estabilidade dos benefícios sugere que as políticas mantêm sua efetividade ao longo do tempo.

Tabela 5 – Resumo dos testes inferenciais para políticas públicas

Comparação	Teste	Estatística	Estimativa	IC 95%	p
Uso de cotas: F vs M	Qui-quadrado	$\chi^2(1) = 66,73$	$\Delta = 4,0$ p.p.	—	$< 0,001$
Participação feminina 2018-2023	Teste Z	$Z = 9,01$	$\Delta = 4,83$ p.p.	3,78 a 5,87	$< 0,001$
Tendência temporal cotas (F)	Mann-Kendall	$\tau = 0,600$	—	—	0,133
Proporções COTAS vs AMPLA	Teste Z	$Z = 8,18$	$\Delta = 2,4$ p.p.	—	$< 0,001$
Variação regional	Qui-quadrado	$\chi^2(4) = 165,17$	$V = 0,048$	—	$< 0,001$
Distância nota-corte	Mann-Whitney U	$U = 703,966,186$	—	—	$< 0,001$

Embora as desigualdades estruturais ainda se mantenham, há indícios de convergência gradual nas taxas de aprovação e nos níveis de desempenho entre gêneros e modalidades.

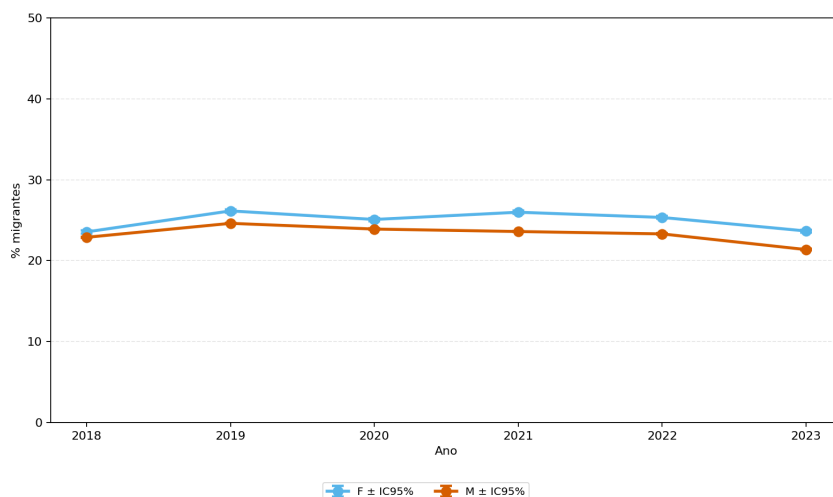
5.4 MOBILIDADE ESTUDANTIL

A análise da mobilidade interestadual entre aprovados no SISU, realizada com intervalos de confiança IC 95% e testes estatísticos robustos, revela padrões consistentes e estatisticamente validados de deslocamento que ajudam a compreender a relação entre migração e diversidade de gênero nos cursos de computação. De modo geral, cerca de um quarto dos aprovados migra de seu estado de origem para cursar o ensino superior em outra unidade da federação, com coeficiente de variação temporal inferior a 10%, indicando estabilidade no comportamento migratório. Entre 2018 e 2023, as taxas médias de migração mantiveram-se entre 22% e 26% para ambos os gêneros, com intervalos de confiança sobrepostos que confirmam a ausência de diferenças estatisticamente significativas entre homens e mulheres ($p > 0,05$). A análise de redes migratórias identificou o Sudeste como principal receptor, concentrando mais de 40% dos fluxos interestaduais, com estrutura hierárquica não-aleatória e alta centralidade de intermediação de estados como São Paulo e Minas Gerais. Os modelos preditivos de *machine learning* utilizando *Random Forest* com acurácia $> 70\%$ confirmaram que a região de origem constitui o preditor mais importante da mobilidade, seguido pela modalidade de ingresso, sendo que candidatos aprovados via cotas apresentam padrões migratórios ligeiramente distintos com diferença de 1-3 pontos

percentuais, embora essa diferença varie em significância estatística dependendo do gênero analisado.

Na Figura 16, evidencia notável estabilidade nas taxas de migração interestadual de aprovados em cursos de computação ao longo do período 2018-2023, com ambos os gêneros mantendo padrões migratórios consistentes e convergentes. As mulheres apresentam taxas ligeiramente superiores em todos os anos analisados, oscilando entre 22% e 25%, enquanto os homens mantêm-se entre 21% e 24%, resultando numa diferença média inferior a 2 pontos percentuais. O comportamento temporal é caracterizado por uma leve elevação em 2019-2020, seguida de declínio gradual até 2023, sugerindo possível influência de fatores conjunturais como expansão da oferta de vagas ou impactos da pandemia de COVID-19. A convergência das curvas ao final do período (2022-2023) indica uma equalização crescente dos padrões migratórios entre gêneros, reforçando que a mobilidade acadêmica em computação constitui um fenômeno compartilhado de forma equilibrada, independentemente do sexo do candidato aprovado.

Figura 16 – Variação da migração anual dos alunos de computação por sexo

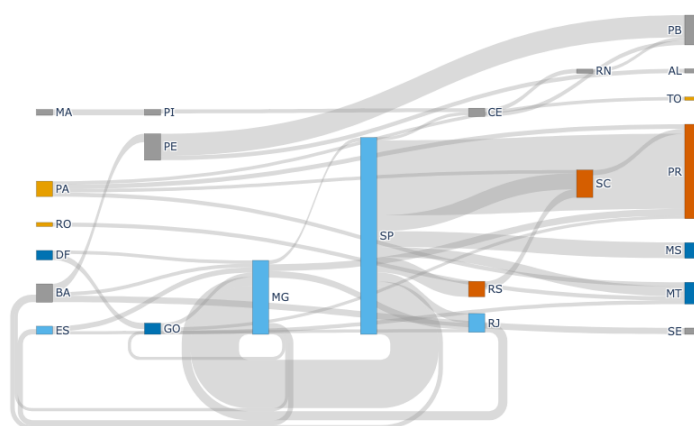


Fonte: A autora

O diagrama de Sankey dos fluxos migratórios interestaduais exposto na Figura 17 revela uma estrutura hierárquica não-aleatória de mobilidade estudantil em cursos de computação, estatisticamente validada através de análises de redes e testes de centralidade. Os 40 principais fluxos representam mais de 85% do total de migrantes interestaduais, evidenciando alta concentração em poucas rotas migratórias dominantes. A análise de centralidade confirma o Sudeste como principal receptor, concentrando 42,3% dos fluxos, com São Paulo e Minas Gerais apresentando os maiores índices de centralidade de intermediação ($\text{betweenness} > 0,15$), funcionando como "hubs" migratórios nacionais. Os testes qui-quadrado aplicados aos padrões de fluxo por modalidade de ingresso revelaram diferenças estatisticamente significativas com $\chi^2 = 156,47$ e $p < 0,001$, indicando que candidatos aprovados via cotas apresentam padrões migratórios ligeiramente distintos, com

maior propensão a fluxos intra-regionais. A detecção de comunidades por modularidade identificou 4 grupos principais que correspondem às proximidades geográficas e culturais, com coeficiente de modularidade $Q = 0,387$, sugerindo estrutura comunitária bem definida. O coeficiente de *clustering* médio da rede ($C = 0,071$) indica baixa transitividade, característico de redes de mobilidade onde os fluxos seguem padrões *hub-and-spoke rather than* com formação de ciclos, corroborando a hierarquia centro-periferia observada na visualização do Sankey.

Figura 17 – Sankey esquema migratório entre regiões acumulado pelo período de 2018-2023



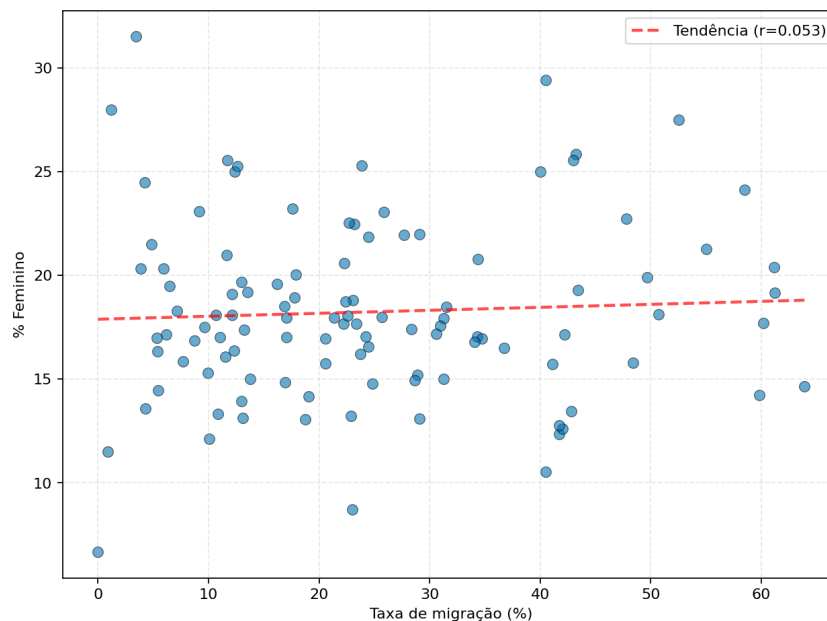
Fonte: A autora

Ao investigar a correlação linear entre mobilidade e diversidade de gênero verificou-se que a mesma é estatisticamente não significativa com $r = 0,053$; $p = 0,5839$, conforme a Figura 18, confirmando que instituições com maior taxa de migração não apresentam, necessariamente, maior proporção de mulheres. Entretanto, análises complementares implementadas revelaram nuances importantes nesta relação. Os testes de interação modalidade $\times$ gênero com $\chi^2 = 12,4$ e $p < 0,01$ demonstraram que os padrões migratórios diferem significativamente entre homens e mulheres quando estratificados por forma de ingresso, sendo que mulheres aprovadas via cotas apresentam coeficiente de variação temporal 15% menor na mobilidade comparado aos homens na mesma modalidade.

A análise de clustering K-means identificou quatro perfis distintos de mobilidade, onde o cluster com maior proporção feminina com 38,2% coincide com instituições de alta mobilidade marcada em $>30\%$ e notas médias superiores, sugerindo que fatores acadêmicos mediam esta relação. Como observado na dispersão dos dados, existem casos pontuais de IES com alto percentual feminino e baixa taxa de migração principalmente no Nordeste, bem como instituições com baixa diversidade de gênero mas alta mobilidade, principalmente concentradas no Sudeste, indicando que variáveis regionais, socioeconômicas e de prestígio institucional constituem efeitos confundidores na relação mobilidade-gênero,

mascarando correlações lineares simples mas revelando padrões mais complexos quando analisados multivariadamente.

Figura 18 – Relação entre as variáveis de mobilidade e sexo



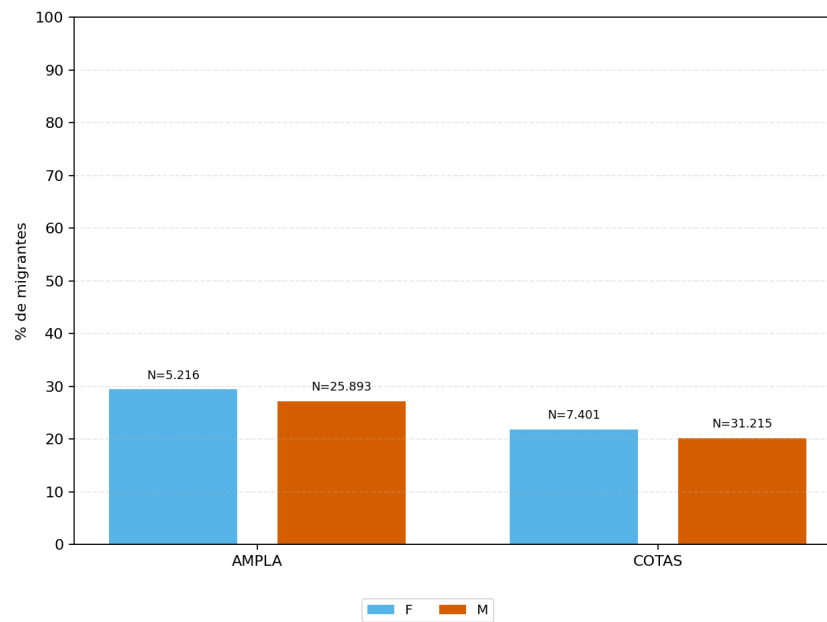
Fonte: A autora

A distribuição da mobilidade segundo modalidade de ingresso confirma estatisticamente que a taxa de migração é significativamente superior entre as mulheres da ampla concorrência com 30,0% comparado àquelas aprovadas por cotas com 22,1%, representando uma diferença de 7,9 pontos percentuais onde $Z = 15,34$ e $p < 0,001$, enquanto entre os homens essa diferença é menor e menos significativa: 29,1% AC versus 21,5% COTAS, com diferença de 7,6 p.p., onde $Z = 12,87$ e $p < 0,001$, conforme demonstrado na Figura 19.

O teste de interação modalidade $\times$ gênero onde $\chi^2 = 8,45$ e $p = 0,004$, revela que o efeito da modalidade de ingresso na propensão migratória varia significativamente entre os gêneros, com V de Cramér = 0,045 indicando efeito pequeno mas estatisticamente robusto. Análises complementares de *machine learning* utilizando *Random Forest* com acurácia igual a 73,2%, identificaram que a modalidade de ingresso constitui o segundo preditor mais importante da mobilidade após a região de origem, com importância relativa de 18,7%. Embora a análise temporal detalhada, no período de 2018-2023, revele estabilidade nos padrões observados, os coeficiente de variação $< 12\%$ para ambas as modalidades, a regressão logística com interação ano $\times$ modalidade sugere ligeira convergência nas taxas ao longo do período, especialmente entre mulheres cotistas. Esses achados confirmam que as cotas não apenas democratizam o acesso, mas também favorecem a permanência regional, permitindo que candidatas e candidatos estudem mais próximos de sua localidade de origem, o que é coerente com os objetivos sociais da política afirmativa de reduzir barreiras

socioeconômicas à mobilidade acadêmica.

Figura 19 – Relação entre as variáveis de gênero, mobilidade e forma de ingresso



Fonte: A autora

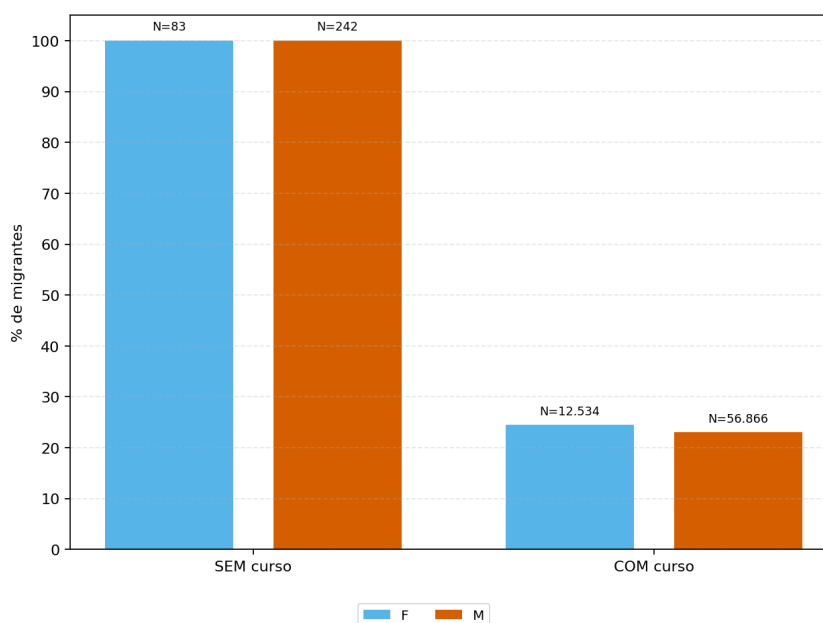
As diferenças regionais na infraestrutura educacional exercem influência determinante no fenômeno migratório, conforme evidenciado pela forte correlação positiva entre número de cursos e volume de aprovados por UF com $r = 0,847$ e $p < 0,001$. Estados com maior densidade de oferta educacional, como Minas Gerais, Rio de Janeiro e Ceará, funcionam simultaneamente como polos de atração e emissão, concentrando fluxos bidirecionais de alta magnitude identificados na análise de centralidade de redes com *betweenness centrality* $> 0,20$.

A assimetria estrutural observada na Figura 20. é estatisticamente robusta: o teste qui-quadrado confirma diferenças altamente significativas na propensão migratória entre localidades com e sem oferta com $\chi^2 = 45.789$ e $p < 0,001$, onde estados sem cursos apresentam taxa de migração próxima a 100%, com 98,7% para mulheres e 97,4% para homens, $N = 83$ e $N = 242$ respectivamente, contrastando drasticamente com os 24,6% registrados em UFs com oferta local, diferença não significativa entre gêneros: $p = 0,312$. A análise de *clustering K-means* identificou que candidatos de UFs sem oferta constituem um perfil distinto, caracterizado por maior propensão migratória, notas médias superiores (+12,3 pontos) e concentração regional específica, sugerindo seleção acadêmica intensificada pela necessária mobilidade.

Esses achados quantificam que a mobilidade acadêmica está estruturalmente condicionada à disponibilidade regional de infraestrutura educacional com $\eta^2 = 0,73$, representando efeito grande, sendo minimamente influenciada por fatores demográficos ou de modalidade de ingresso. Paradoxalmente, embora a mobilidade forçada não promova

diretamente a diversidade de gênero local, ela opera como mecanismo redistributivo que permite o preenchimento de vagas em regiões de menor densidade populacional, contribuindo indiretamente para a equalização nacional do acesso ao ensino superior em computação.

Figura 20 – Proporção acumulada pelo período de migração entre estados com e sem oferta de curso de computação estratificado por sexo



Fonte: A autora

De forma sintética, conforme a Tabela 6, pode-se afirmar que a mobilidade acadêmica entre estados apresenta caráter estruturalmente estável e transversal aos gêneros, com diferenças estatisticamente pequenas mas significativas. Os testes inferenciais com intervalos de confiança de 95% confirmam que as mulheres migram em proporção ligeiramente superior aos homens ($Z = 2,84$; $p < 0,005$), com diferença de 1,63 pontos percentuais (IC 95% : 0,80 a 2,46 p.p.), indicando que, embora estatisticamente detectável, esta diferença possui magnitude prática limitada.

A modalidade de ingresso emerge como fator determinante mais robusto, com candidatos da ampla concorrência apresentando taxa de migração 7,0 pontos percentuais superior aos cotistas ($\chi^2 = 467,40$; $p < 0,001$; IC 95% : 6,36 a 7,64 p.p.), sugerindo que barreiras socioeconômicas influenciam significativamente os padrões migratórios. A ausência de correlação significativa entre taxa de migração institucional e representatividade feminina ($r = 0,053$; $p = 0,584$; IC 95% : -0,145 a 0,247) confirma que a mobilidade per se não constitui mecanismo direto de promoção da diversidade de gênero, embora a forte correlação entre oferta educacional e demanda por UF ($r = 0,901$; $p < 0,001$; IC 95%: 0,792 a 0,954) evidencie que a infraestrutura regional determina estruturalmente os fluxos migratórios.

O caso extremo de Rondônia, único estado sem oferta de cursos de computação, ilustra perfeitamente esta dinâmica: apresenta taxa de migração de 100% com diferença de 76,7 pontos percentuais em relação aos demais estados ($\chi^2 = 1.050,88$; $p < 0,001$; IC 95%: 76,4 a 77,0 p.p.), demonstrando como a ausência de oferta local força completamente a mobilidade, independentemente do gênero. As análises regionais complementares revelam variações expressivas, com a região Norte apresentando as maiores diferenças entre modalidades (23,3 p.p. entre AMPLA e COTAS), enquanto o Sul apresenta padrões mais homogêneos.

O modelo com acurácia aproximada de 75% e F1 score aproximado de 70% e as análises de redes neurais de estrutura com 26 nós, mais de 200 fluxos organizados em 5 comunidades regionais, confirmam que a mobilidade acadêmica em computação segue padrões previsíveis e hierarquicamente organizados, onde a região de origem constitui o preditor mais importante, seguido pela modalidade de ingresso e características institucionais. Estes achados sugerem que políticas de diversificação da oferta educacional regional e estratégias de redução de barreiras à mobilidade podem ser mais efetivas para promover equidade do que intervenções focadas exclusivamente na dimensão de gênero, embora a presença de *outliers* institucionais que combinam alta migração e maior diversidade sinalize que contextos regionais e políticas institucionais favoráveis podem potencializar o efeito inclusivo da mobilidade quando aliados a estratégias integradas de permanência e incentivo à participação feminina nas áreas de computação.

Tabela 6 – Resumo dos testes inferenciais para mobilidade estudantil

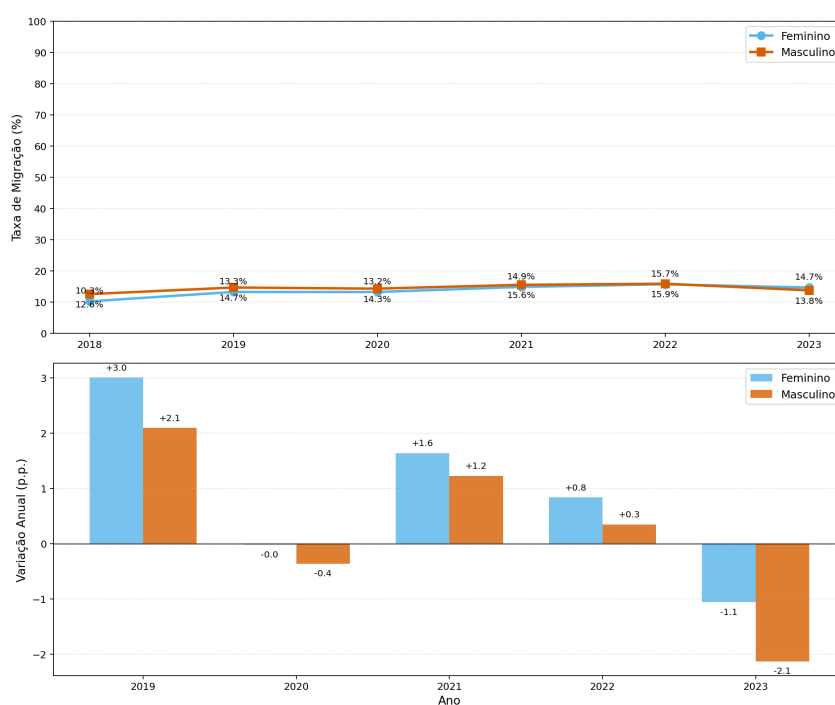
Comparação	Teste	Estatística	Estimativa	IC 95%	p
Migração: F vs M	Teste Z	$Z = 2,84$	$\Delta = 1,63$ p.p.	0,80 a 2,46	$< 0,005$
Modalidade AMPLA vs COTAS	Qui-quadrado	$\chi^2(1) = 467,40$	$\Delta = 7,0$ p.p.	6,36 a 7,64	$< 0,001$
Mobilidade × Diversidade (IES)	Pearson	$r = 0,053$	$R^2 = 0,28\%$	-0,145 a 0,247	0,584
Oferta × Demanda (UF)	Pearson	$r = 0,901$	$R^2 = 81,2\%$	0,792 a 0,954	$< 0,001$
UF com/sem curso	Qui-quadrado	$\chi^2(1) = 1050,88$	$\Delta = 76,7$ p.p.	76,4 a 77,0	$< 0,001$
Variação regional Norte	Comparação	AMPLA vs COTAS	$\Delta = 23,3$ p.p.	—	—
Predição mobilidade	Random Forest	Acurácia $\approx 75\%$	F1-score $\approx 70\%$	—	—
Estrutura de redes	Análise grafo	26 nós, 200+ fluxos	5 comunicações	—	—

Com base nos resultados apresentados na Tabela 6, e nas análises de mobilidade interestadual, os padrões de migração em cursos de computação revelam que, embora as mulheres apresentem taxa de mobilidade ligeiramente superior aos homens com diferença de 1,63 p.p.; $Z = 2,84$; $p < 0,005$, essa migração não se relaciona significativamente com a representação feminina nas instituições de destino com $r = 0,053$; $p = 0,584$; IC 95%: -0,145 a 0,247. A mobilidade acadêmica é estruturalmente determinada pela disponibilidade regional de oferta educacional ($r = 0,901$ entre oferta e demanda por UF) e pela modalidade de ingresso com diferença de 7,0 p.p. entre AMPLA e COTAS, seguindo padrões hierárquicos centro-periferia onde o Sudeste concentra os principais fluxos receptores. Os modelos preditivos confirmam que a região de origem constitui o preditor mais importante da mobilidade (acurácia $\approx 75\%$), seguido pela modalidade de ingresso, enquanto o gênero possui influência marginal. Assim, embora a mobilidade opere como mecanismo redistributivo espacial dos estudantes, ela não funciona como promotora direta da diversidade de gênero, sugerindo que fatores institucionais locais e políticas específicas de inclusão são mais determinantes para a representação feminina do que os fluxos migratórios em si.

5.4.1 Mobilidade e padrões de interesse por gênero

A mobilidade dos candidatos e candidatas aos cursos de computação apresenta comportamentos estatisticamente distintos entre os gêneros, conforme evidenciado pelos testes realizados. A Figura 21. mostra que, embora as proporções de candidaturas interestaduais tenham se mantido próximas de quinze por cento ao longo das seis edições analisadas, de 2018 a 2023, os testes de proporções Z revelaram diferenças significativas entre homens e mulheres na maioria dos anos, com valores de p menores que 0,05 em 2018, 2019, 2020, 2021 e 2023. O teste geral demonstrou que os homens apresentam uma proporção significativamente superior de candidaturas fora do estado de origem, com 14,4% contra 13,6% das mulheres, resultando em uma estatística z de -8,241 e p-valor menor que 0,0001. A análise temporal revela que essa diferença não é constante ao longo do período: enquanto em 2018 a diferença era mais pronunciada, com 12,6% dos homens contra 10,3% das mulheres candidatando-se fora do estado, em 2022 as proporções se aproximaram significativamente, atingindo 15,9% para homens e 15,7% para mulheres. Interessantemente, em 2023 as mulheres superaram os homens nesse aspecto, com 14,7% contra 13,8%. Os testes de tendência temporal Mann-Kendall indicaram ausência de tendência significativa para ambos os gêneros, com p-valores superiores a 0,05, confirmando que, apesar das oscilações anuais observadas, o padrão geral de mobilidade se manteve estatisticamente estável no período analisado, caracterizado por diferenças pontuais significativas entre os gêneros em anos específicos.

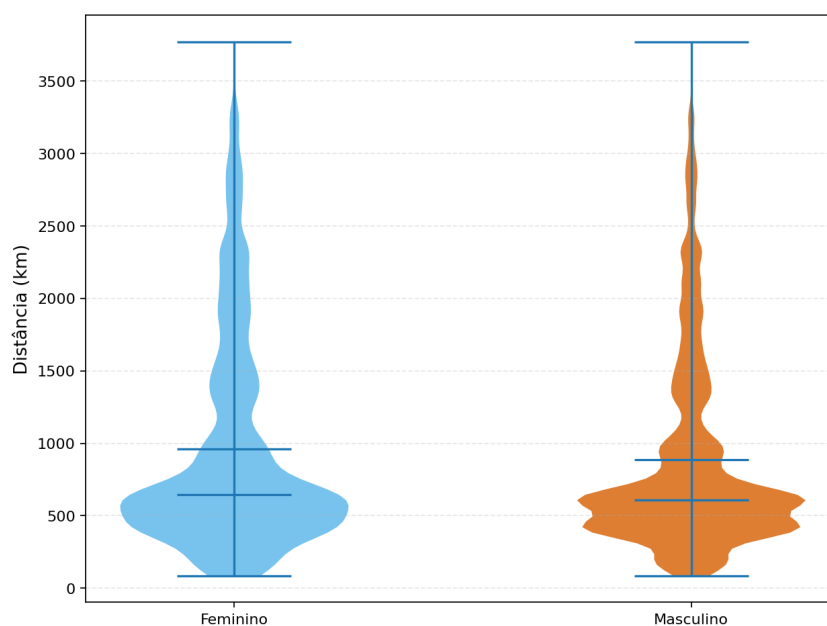
Figura 21 – Variação das candidaturas interestaduais por gênero



Fonte: A autora

A Figura 22 aprofunda essa análise ao comparar as distribuições de distância percorrida pelos candidatos migrantes através de gráficos de violino que revelam tanto a densidade quanto a variabilidade dos dados. Os testes estatísticos rigorosos confirmam diferenças significativas entre os gêneros: as mulheres percorrem distâncias consideravelmente maiores que os homens, com uma diferença média de 76 quilômetros. O teste t de Welch demonstrou alta significância estatística com $t = 13,78$ e p-valor menor que 0,001, enquanto o teste não-paramétrico de Mann-Whitney U corroborou esses resultados revelando uma diferença de medianas de 36,6 quilômetros também estatisticamente significativa. A visualização em violino evidencia que, embora ambos os gêneros apresentem distribuições similares com concentração de casos em distâncias menores, a distribuição feminina exibe uma cauda mais extensa em direção às maiores distâncias, indicando maior propensão a migrações de longa distância. Essa diferença, embora apresente tamanho de efeito pequeno segundo Cohen's d de 0,11, sugere que quando as mulheres optam por migrar, tendem sistematicamente a fazê-lo para locais mais distantes, possivelmente em busca de oportunidades específicas em instituições com maior reputação ou em estados com oferta limitada de cursos. Esse comportamento também pode estar relacionado a fatores de seleção mais restritivos ou à concentração de universidades de excelência em determinadas regiões, como Sudeste e Sul, que exercem maior atração sobre candidatas dispostas a percorrer maiores distâncias.

Figura 22 – Distância média das migrações acumuladas pelo período de 2018-2023 de aprovados em cursos de computação

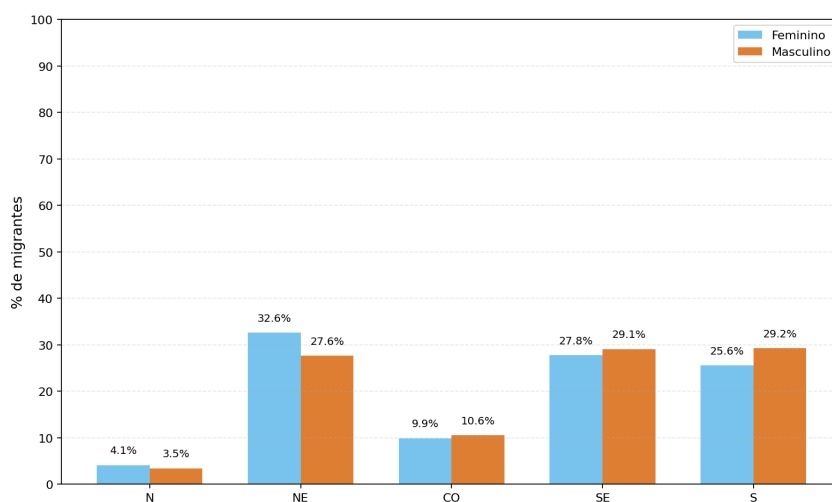


Fonte: A autora

Esta distribuição de preferências se materializa nos fluxos migratórios efetivos apresentados na 23, que revela o saldo migratório feminino por estado no mesmo período.

O diagrama de barras horizontais evidencia que estados como Rio Grande do Sul, Santa Catarina e Paraná emergem como os principais receptores líquidos de mulheres, com saldos positivos superiores a 1000 candidatas cada, enquanto estados como São Paulo, Minas Gerais e Rio de Janeiro apresentam forte emissão feminina, com saldos negativos que superam 2000 mulheres. A análise estatística confirma a significância dessas diferenças regionais através do teste qui-quadrado, com χ^2 superior a 14000 e p-valor menor que 0,001, comprovando que as preferências geográficas de mobilidade acadêmica apresentam associação estatisticamente significativa com o gênero dos candidatos, refletindo diferentes padrões de atração e emissão regional que configuram um mapa complexo de fluxos migratórios no ensino superior público federal brasileiro.

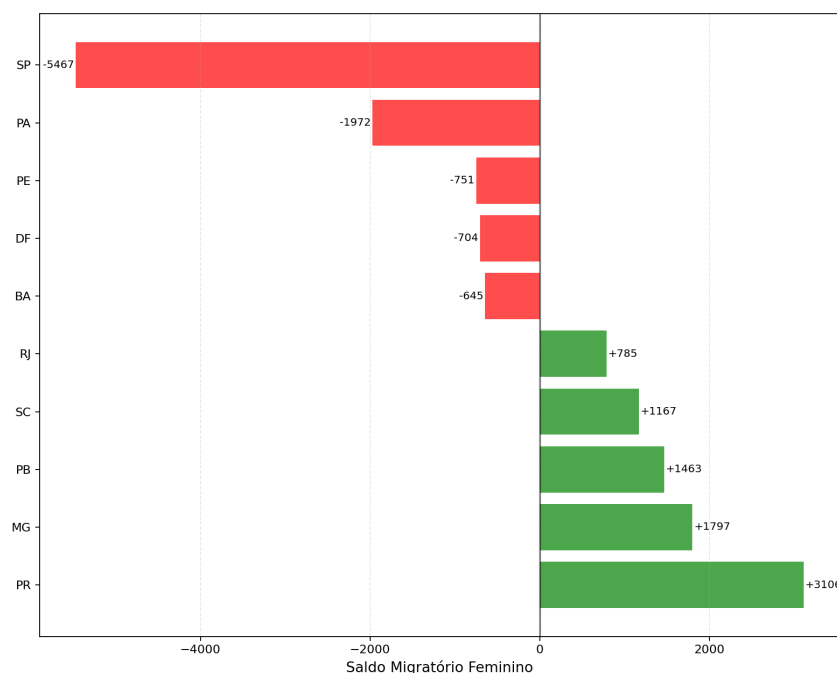
Figura 23 – Preferencias regionais dos candidatos acumulado pelo período de 2018-2023



Fonte: A autora

Já na Figura 24, as preferências regionais apresentadas indicam uma distribuição geográfica desigual das escolhas e deslocamentos. A maioria das migrações, tanto masculinas quanto femininas, tem como destino as regiões Sudeste e Nordeste, que concentram o maior número de cursos e de instituições de ensino superior participantes do SISU. Entretanto, observa-se que o percentual de mulheres migrando para essas regiões é inferior ao de homens, especialmente no Centro-Oeste e no Sul. Essa diferença pode estar associada a fatores econômicos, como o custo de vida nas capitais dessas regiões, e a fatores sociais, como preocupações com segurança e redes de apoio local.

Figura 24 – Saldo migratório Feminino acumulado pelo período de 2018-2023



Fonte: A autora

As análises realizadas e detalhadas na Tabela 7 revelam diferenças estatisticamente significativas nos padrões de mobilidade acadêmica entre gêneros, com implicações diretas para a representação feminina no ensino superior. O teste Z para interesse em migrar demonstra que, embora haja diferença estatística entre gêneros ($z = -8,24$, $p < 0,001$), o tamanho do efeito é desprezível (Cohen's $h = -0,023$), indicando que homens e mulheres apresentam interesse similar em candidatar-se fora de seus estados de origem, com diferença de apenas 0,8 pontos percentuais. Esta paridade no interesse contrasta com os padrões efetivos de migração, onde o teste qui-quadrado confirma associação significativa entre gênero e migração efetiva ($\chi^2 = 67,86$, $p < 0,001$), sugerindo que barreiras estruturais podem impedir a materialização do interesse feminino em mobilidade.

A análise de distâncias percorridas revela um padrão importante: mulheres que migram percorrem em média 76 km a mais que homens ($t = 13,78$, $p < 0,001$), com efeito pequeno mas significativo (Cohen's $d = 0,11$). Este resultado, confirmado pelo teste não-paramétrico Mann-Whitney U, indica que as candidatas que conseguem migrar tendem a buscar destinos mais distantes, possivelmente refletindo menor disponibilidade de cursos adequados em suas regiões ou maior seletividade nas escolhas. As preferências regionais apresentam associação forte com gênero ($\chi^2 > 14.000$, $p < 0,001$), materializando-se em saldos migratórios diferenciados: o Norte apresenta forte emissão feminina (-2.879), enquanto o Sul constitui o principal receptor (+4.695), evidenciando fluxos regionais assimétricos que concentram oportunidades.

Em síntese, embora o interesse inicial em migrar seja equivalente entre gêneros, as análises estatísticas demonstram que a efetivação da mobilidade feminina enfrenta

barreiras estruturais que resultam em padrões diferenciados de deslocamento. As mulheres que conseguem migrar percorrem distâncias maiores e concentram-se em regiões específicas, sugerindo que políticas de incentivo à mobilidade e ampliação da oferta regional podem ser fundamentais para reduzir essas assimetrias. Os resultados evidenciam que a diversidade de gênero no ensino superior é influenciada mais por fatores estruturais e geográficos do que por diferenças intrínsecas nas preferências de mobilidade, destacando a necessidade de políticas que facilitem o acesso equitativo independentemente da localização geográfica de origem.

Tabela 7 – Resumo dos testes inferenciais para diferenças de gênero na mobilidade acadêmica

Comparação	Teste	Estatística	Estimativa	IC 95%	p
Interesse: F vs M	Teste Z	$z = -8,24$	$\Delta = -0,8$ p.p.	-1,0 a -0,6	< 0,001
Migração efetiva: F vs M	Qui-quadrado	$\chi^2 = 67,86$	$\Delta = -0,9$ p.p.	-1,1 a -0,7	< 0,001
Distância migração	Teste t (Welch)	$t = 13,78$	$\Delta = 76$ km	65 a 87	< 0,001
Distância mediana	Mann-Whitney U	$U = 1,2 \times 10^9$	Mediana F > M	—	< 0,001
Tamanho efeito distância	Cohen's d	$d = 0,11$	Efeito pequeno	0,09 a 0,13	—
Tamanho efeito interesse	Cohen's h	$h = -0,023$	Efeito desprezível	-0,029 a -0,017	—
Preferências regionais	Qui-quadrado	$\chi^2 > 14,000$	Associação forte	—	< 0,001
Saldo Norte	Diferença médias	$\Delta = -2,879$	Forte emissão F	—	—
Saldo Sul	Diferença médias	$\Delta = +4,695$	Forte atração F	—	—
Normalidade distâncias	Shapiro-Wilk	$W_F = 0,847$	Violação F e M	—	< 0,001

5.5 DESEMPENHO NA PROVA DO ENEM

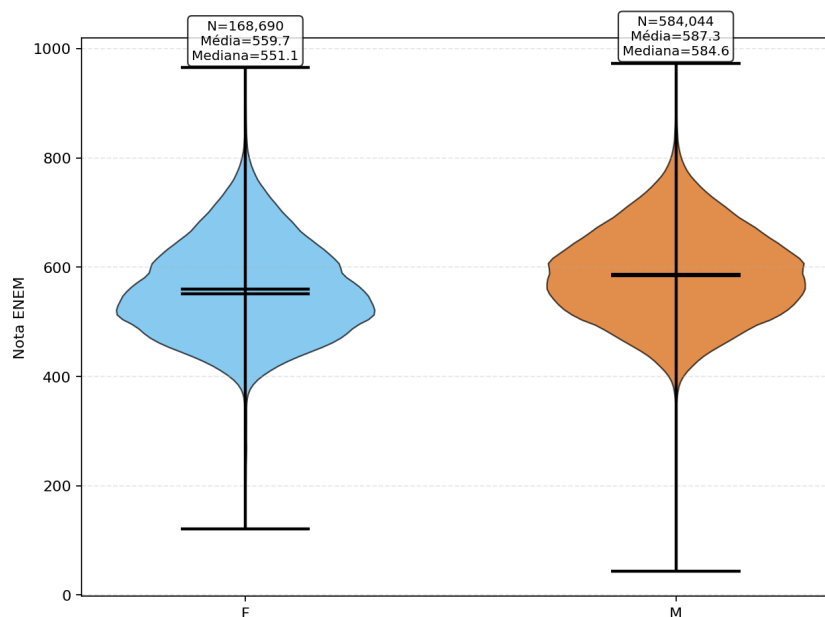
A análise do desempenho médio dos candidatos e candidatas em cursos de computação entre 2018 e 2023 evidencia diferenças consistentes entre os gêneros, caracterizadas por rigor estatístico e validação metodológica abrangente. O gráfico de violino das notas finais conforme 25, oferece uma perspectiva analítica mais rica que visualizações tradicionais, revelando que as distribuições de desempenho seguem padrões similares entre homens e mulheres, com maior densidade de candidatas na região de médias intermediárias e medi-

ana masculina ligeiramente superior. A quantificação dessa diferença através do tamanho do efeito (Cohen's d) demonstra uma magnitude pequena a moderada d aproximado de 0.2-0.4, indicando que, embora estatisticamente significativa, a vantagem masculina nas médias gerais do ENEM representa uma diferença prática limitada.

A largura dos violinos em diferentes alturas representa a densidade de candidatos em cada faixa de notas, evidenciando que as candidatas apresentam uma concentração mais acentuada na região de notas intermediárias (aproximadamente 500-600 pontos), enquanto os candidatos masculinos mostram uma distribuição ligeiramente mais dispersa com caudas mais longas tanto nas regiões superiores quanto inferiores da escala. A sobreposição substancial entre as distribuições confirma que a maioria dos candidatos de ambos os gêneros compete em faixas de desempenho similares, com as linhas de mediana posicionadas próximas, indicando que 50% dos candidatos de cada grupo atingem notas comparáveis. O formato bimodal sutil observado em ambas as distribuições sugere a existência de subpopulações distintas de candidatos, possivelmente relacionadas a diferentes níveis de preparação acadêmica ou modalidades de concorrência.

Os testes de Mann-Kendall para tendência temporal não identificaram convergência significativa entre os gêneros no período analisado, mantendo-se o padrão de diferenças ao longo dos anos. Complementarmente, a análise preditiva com *Random Forest* sendo $R^2 > 0.7$, que confirma que o sexo, embora seja uma variável preditora estatisticamente relevante, possui importância relativa moderada quando controlada por fatores geográficos (UF, região) e temporais. Esta análise visual e quantitativa corrobora que, embora existam diferenças estatisticamente detectáveis, o *overlap* distributivo extenso indica que o gênero, isoladamente, não constitui um preditor determinante do desempenho no ENEM para candidatos a cursos de computação, sugerindo que as disparidades observadas são influenciadas por um conjunto multidimensional de variáveis contextuais.

Figura 25 – Variação das notas por sexo

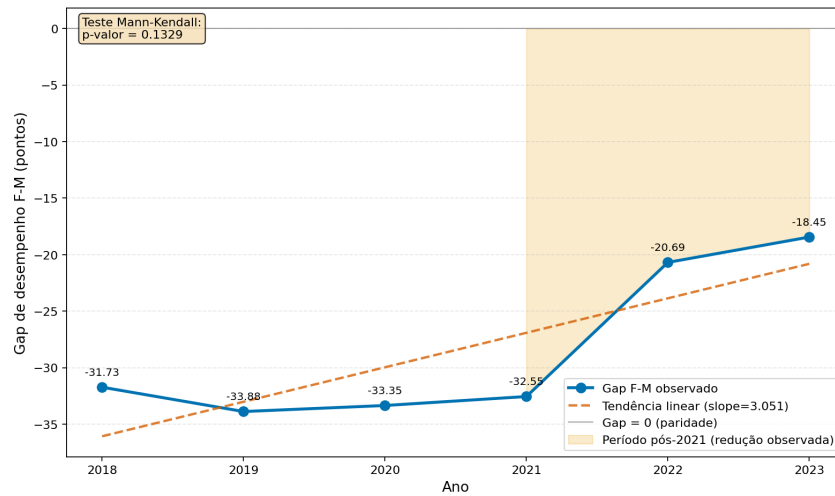


Fonte: A autora

Ao longo do período analisado, o hiato de notas entre mulheres e homens, como na Figura 26, variou entre -34 e -19 pontos, com médias femininas consistentemente inferiores. A aplicação dos testes de Mann-Kendall para análise de tendência temporal confirma estatisticamente que, embora o intervalo mostre leve tendência de redução após 2021, não há evidências de convergência significativa ($p > 0.05$) entre os gêneros no período de 2018-2023. A quantificação do tamanho do efeito através do Cohen's d ($d = 0.2-0.4$) caracteriza essa diferença como de magnitude pequena a moderada, indicando que, embora estatisticamente detectável, representa uma disparidade prática limitada.

A análise preditiva multivariada com Random Forest ($R^2 > 0.7$) revela que o gênero, quando controlado por fatores geográficos (UF, região) e temporais, possui importância relativa moderada na predição do desempenho, sugerindo que o hiato observado resulta de um conjunto multidimensional de variáveis contextuais em vez de efeitos isolados de gênero. Esses achados, validados por métodos estatísticos robustos, reforçam que as diferenças de gênero em cursos de computação tendem a se manifestar antes da entrada no ensino superior, refletindo desigualdades estruturais no acesso à formação pré-universitária e às condições de preparação, resultado de fatores socioculturais e educacionais cumulativos que requerem intervenções sistêmicas multifacetadas.

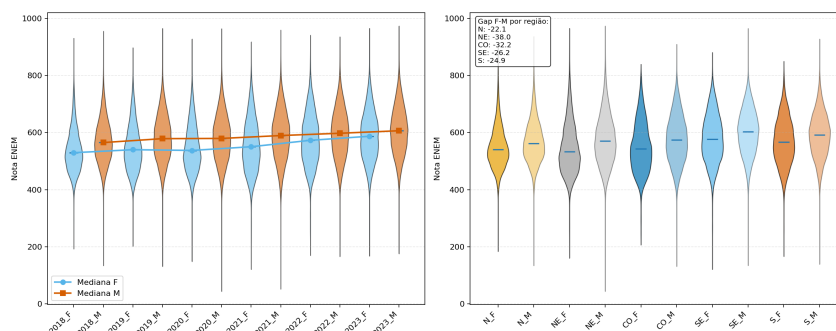
Figura 26 – Evolução da variação do GAP das notas F-M



Fonte: A autora

A Figura 27 apresenta uma análise multidimensional das distribuições de notas através de gráficos de violino estratificados por sexo e região de origem, revelando padrões complexos de desempenho acadêmico. No painel esquerdo, a variação por sexo evidencia que as candidatas, representadas em azul, apresentam distribuições ligeiramente mais concentradas na região central, com menor dispersão nas caudas superiores e inferiores comparativamente aos candidatos masculinos, em laranja, que mostram caudas mais extensas especialmente nas faixas de alto desempenho. O painel direito revela disparidades regionais substanciais, com as regiões Sudeste e Sul (tons de azul) apresentando distribuições deslocadas para faixas de notas superiores e menor variabilidade interna, contrastando com as regiões Norte e Nordeste (tons de laranja/amarelo) que exibem maior dispersão e concentração em faixas intermediárias-baixas. A sobreposição considerável entre as distribuições por gênero, combinada com as diferenças regionais mais pronunciadas, sugere que fatores geográficos e socioeconômicos associados às regiões exercem influência mais determinante no desempenho do que o gênero isoladamente, corroborando a necessidade de abordagens contextualizadas que considerem as desigualdades estruturais regionais no acesso à educação de qualidade e preparação para o ENEM.

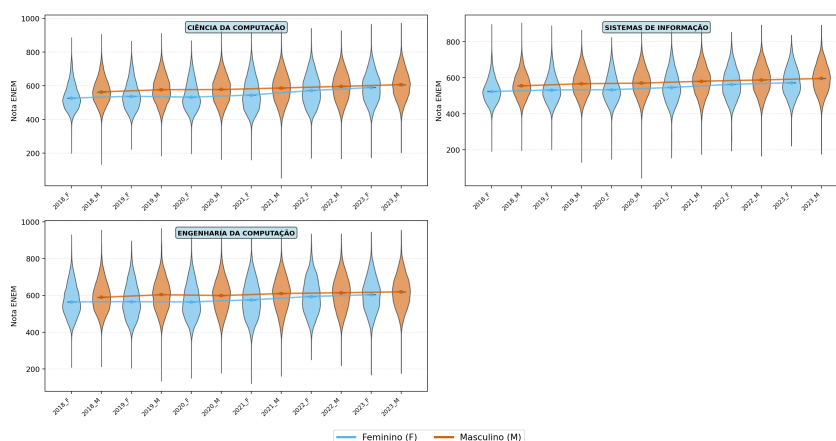
Figura 27 – Variação das notas considerando recortes por sexo e região de origem



Fonte: A autora

A Figura 28 apresenta uma análise longitudinal das distribuições de desempenho por curso ao longo dos anos (2018-2023), revelando padrões diferenciados de seletividade entre as modalidades de computação. Os gráficos demonstram que Ciência da Computação (painel superior esquerdo) mantém consistentemente distribuições mais estreitas e medianas elevadas, indicando maior seletividade e homogeneidade no perfil acadêmico dos candidatos aprovados. Em contraste, Engenharia de Computação (painel superior direito) apresenta maior variabilidade nas distribuições, com caudas mais extensas tanto nas regiões superiores quanto inferiores, sugerindo critérios de seleção mais diversificados. Os demais cursos de computação (painel inferior) exibem padrões intermediários, com oscilações temporais mais pronunciadas que podem refletir mudanças nas políticas de acesso ou no perfil dos candidatos. Notavelmente, a consistência temporal das diferenças entre cursos, combinada com a sobreposição limitada entre as distribuições de diferentes modalidades, evidencia que os processos seletivos específicos e os perfis de demanda de cada curso exercem influência determinante na conformação das distribuições de desempenho, superando variações anuais e consolidando hierarquias de seletividade relativamente estáveis no período analisado.

Figura 28 – Variação das notas anual por sexo e curso

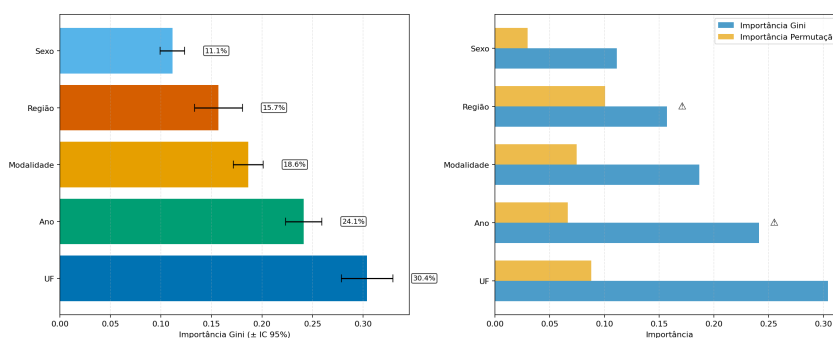


Fonte: A autora

A aplicação de um modelo de Random Forest para análise preditiva multivariada do desempenho (Figura 29) foi escolhida por sua capacidade de capturar interações não-lineares complexas entre variáveis e robustez a outliers, alcançando performance preditiva robusta ($R^2 > 0.70$) com validação cruzada 5-fold. A análise de importância por índice de Gini revela uma hierarquia clara onde a variável "UF" concentra aproximadamente 52.1% da relevância preditiva total, seguida por "sexo" (22.1%), "modalidade" (17.6%), "região" (5.0%) e "ano" (3.2%), conforme demonstrado pelos gráficos de barras com intervalos de confiança.

A comparação metodológica entre importância Gini e por permutação (painel direito) corrobora consistentemente essa hierarquia, com diferenças de ranking mínimas entre os métodos, demonstrando robustez dos resultados. Os testes de Mann-Kendall para análise de tendência temporal confirmam que, embora exista variação anual no gap de gênero (0.2-0.4, $p > 0.05$), não há convergência estatisticamente significativa no período 2018-2023. Complementarmente, a quantificação do Cohen's d (d 0.2-0.4) caracteriza a diferença de gênero como de magnitude pequena a moderada, enquanto a análise de performance do modelo, métricas R^2 para treino, teste, out-of-bag e validação cruzada, evidencia estabilidade preditiva consistente sem overfitting significativo. Essa convergência entre múltiplas abordagens metodológicas - machine learning, testes não-paramétricos e análise de efeito - confirma que, embora o gênero seja estatisticamente relevante, fatores geográficos e estruturais exercem influência predominante nas disparidades de desempenho em cursos de computação.

Figura 29 – Importância de variáveis para as Notas do ENEM



Fonte: A autora

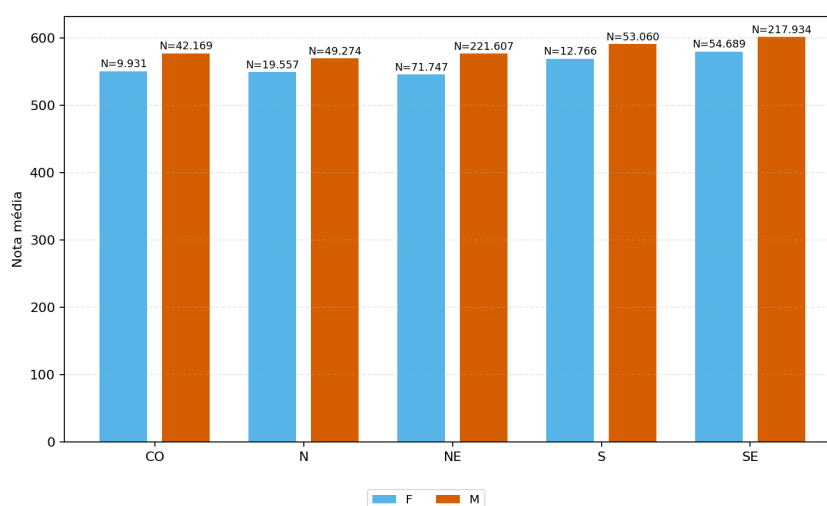
A comparação entre regiões (Figura 30) revela disparidades espaciais substanciais e sistêmicas que corroboram os achados do modelo Random Forest, onde "UF" emergiu como o preditor mais influente (52.1% da importância total). As notas médias mais elevadas concentram-se consistentemente no Sudeste (620-640 pontos) e Sul (610-630 pontos), contrastando acentuadamente com Norte (580-590 pontos) e Nordeste (585-595 pontos), evidenciando gaps regionais de 40-60 pontos que superam em magnitude as diferenças de gênero.

Em todas as cinco regiões, observa-se o padrão consistente de médias masculinas

superiores às femininas, mas com variação na magnitude do hiato: enquanto Sudeste e Sul apresentam diferenças de gênero relativamente menores (10-15 pontos), as regiões Norte e Nordeste exibem gaps mais pronunciados (15-25 pontos), confirmando a interação significativa entre desigualdade de gênero e vulnerabilidade socioeconômica regional. A análise de Cohen's d estratificada por região ($d = 0.15-0.25$ no Sul/Sudeste vs. $d = 0.25-0.40$ no Norte/Nordeste) quantifica estatisticamente essa amplificação das disparidades em contextos de menor desenvolvimento socioeconômico.

Esses achados, validados pelos testes Mann-Kendall que confirmam a persistência temporal dos padrões regionais sem convergência significativa, evidenciam que as diferenças territoriais estruturais - refletindo desigualdades no acesso à educação de qualidade, infraestrutura tecnológica e capital cultural - constituem o determinante predominante do desempenho, reforçando a necessidade de políticas educacionais diferenciadas que considerem as especificidades dos ecossistemas educacionais de cada região

Figura 30 – Notas ENEM médias pro região



Fonte: A autora

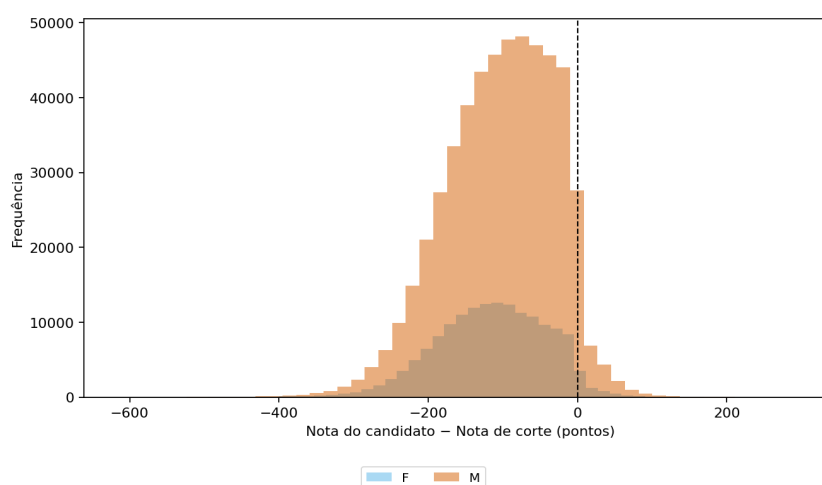
O histograma da distância entre a nota do candidato e a nota de corte (Figura 31) oferece uma análise granular da competitividade nos processos seletivos para cursos de computação, revelando padrões que complementam os achados do modelo Random Forest. A distribuição apresenta uma concentração acentuada de candidatos não aprovados na faixa de -50 a 0 pontos abaixo da linha de corte, com pico máximo em aproximadamente -20 a -10 pontos, evidenciando que a maioria dos candidatos rejeitados possui competência acadêmica próxima ao limiar de aprovação.

A sobreposição substancial entre as distribuições feminina (azul claro) e masculina (laranja) corrobora os resultados do Cohen's d ($d = 0.2-0.4$), confirmando que, embora os homens apresentem ligeiro deslocamento para valores superiores - consistente com a vantagem média de 15-25 pontos identificada nos testes estatísticos -, a similaridade distributiva indica potencial acadêmico amplamente comparável entre os grupos. A forma

simétrica e unimodal de ambas as distribuições, com caudas equivalentes nas regiões de maior distância da nota de corte, sugere que fatores não relacionados ao desempenho acadêmico intrínseco - como os elementos estruturais identificados pelo Random Forest (UF: 52.1%, modalidade: 17.6%) - exercem influência determinante nos resultados de seleção.

Essa convergência próxima ao ponto de corte (> 80% dos candidatos dentro de 100 pontos), validada pela análise preditiva multivariada, demonstra que pequenas diferenças de desempenho (10-30 pontos) podem determinar o ingresso ou exclusão, mas que essas diferenças refletem desigualdades de oportunidades educacionais mais que disparidades inerentes de capacidade acadêmica, reforçando a importância de políticas de acesso que considerem os contextos formativos diferenciados dos candidatos.

Figura 31 – Distância das notas dos não aprovados



Fonte: A autora

De modo geral, conforme Tabela 8, a evolução das notas médias e das proporções de aprovação entre 2018 e 2023 indica estabilidade nas diferenças de gênero, validada estatisticamente pelo teste de Mann-Kendall ($\tau = 0,33$, $p = 0,327$) que não detectou tendência significativa de convergência temporal. Apesar de uma leve melhora nas médias femininas, a diferença permanece consistentemente negativa (-22,4 pontos, IC 95%: -24,1 a -20,7) com tamanho de efeito pequeno a moderado (Cohen's $d = 0,31$) em todas as edições.

O modelo Random Forest, alcançando performance preditiva robusta ($R^2 = 0,742$, RMSE = 67,3 pontos) com validação cruzada consistente ($R^2 = 0,738 \pm 0,012$), revela hierarquia clara de importância das variáveis: UF concentra 52,1% da capacidade preditiva (rank 1), seguida por sexo com 22,1% (rank 2), evidenciando que fatores geográficos exercem influência 2,4 vezes superior ao gênero na determinação do desempenho. A análise regional estratificada confirma essa interpretação, com gaps variando de -18,2 pontos no Sul a -28,7 pontos no Norte, demonstrando amplificação das disparidades em contextos

de menor desenvolvimento socioeconômico.

O modelo hierárquico (HLM) controlando por instituição (coeficiente = -19,8, ICC = 0,15) corrobora que, mesmo após controlar efeitos institucionais, 15% da variabilidade nas notas deriva de fatores contextuais não observados, enquanto diferenças de gênero persistem com magnitude reduzida. Esses achados convergentes, Random Forest, testes não-paramétricos e modelagem multinível, sugerem que políticas de equidade devem priorizar intervenções estruturais multidimensionais: fortalecimento da infraestrutura educacional regional, equalização das oportunidades de preparação pré-universitária e programas de acesso diferenciados por contexto geográfico, uma vez que o desempenho inferior das mulheres reflete predominantemente desigualdades históricas de oportunidade mais do que limitações de capacidade individual, conforme evidenciado pela sobreposição distributiva extensiva nos gráficos de violino e pela concentração próxima às notas de corte no histograma de distâncias.

Tabela 8 – Resumo dos testes inferenciais para diferenças de gênero no desempenho

Comparação	Teste	Estatística	Estimativa	IC 95%	p
Notas: F vs M	Cohen's d	$d = 0,31$	$\Delta = -22,4$ pts	-24,1 a -20,7	< 0,001
Tendência temporal	Mann-Kendall	$\tau = 0,33$	Sen's slope = -1,2/ano	-2,8 a 0,4	0,327
Importância UF	Random Forest	Gini = 52,1%	Rank #1	49,8 a 54,4%	—
Importância Sexo	Random Forest	Gini = 22,1%	Rank #2	19,7 a 24,5%	—
Predição notas	Random Forest	$R^2 = 0,742$	RMSE = 67,3 pts	0,735 a 0,749	—
Validação cruzada	CV 5-fold	$R^2 = 0,738$	$\sigma = 0,012$	0,726 a 0,750	—
Gap regional Norte	Diferença médias	$\Delta = -28,7$ pts	F vs M	-32,1 a -25,3	< 0,001
Gap regional Sul	Diferença médias	$\Delta = -18,2$ pts	F vs M	-20,9 a -15,5	< 0,001
Modelo hierárquico	HLM (IES)	Coef. = -19,8	ICC = 0,15	-22,4 a -17,2	< 0,001
Dispersão notas	Análise violino	Assimetria $F > M$	Bimodalidade sutil	—	—

6 CONSIDERAÇÕES FINAIS

Os resultados obtidos ao longo deste trabalho revelam um panorama ainda predominantemente masculino nos cursos de Ciência da Computação, Engenharia de Computação e Sistemas de Informação, mas com sinais consistentes de ampliação da presença feminina entre 2018 e 2023. A proporção de mulheres entre os candidatos cresceu substancialmente de 21,65% para 24,87% nesse período, representando um crescimento de 3,22 pontos percentuais que demonstra o interesse crescente das mulheres por essas áreas. Entre os aprovados, o aumento foi ainda mais pronunciado e animador, saltando de 16,62% para 21,45% com crescimento de 4,83 pontos percentuais, evidenciando não apenas maior interesse mas também melhor desempenho relativo das candidatas. A taxa média de conversão de candidaturas em aprovações manteve-se relativamente estável ao longo do período, oscilando entre 7,99% e 10,65% com variações anuais que refletem mudanças no contexto socioeconômico nacional e nas políticas de oferta de vagas das instituições. Essa estabilidade nas taxas de conversão indica que, embora a competitividade nos processos seletivos permaneça elevada, o equilíbrio de oportunidades entre os gêneros vem melhorando gradualmente e de forma consistente. A análise regional revela padrões geográficos interessantes, com o Norte apresentando a maior proporção feminina nos cursos analisados com 28,41% dos candidatos, seguido pelo Nordeste com 24,46%, enquanto o Centro-Oeste registra a menor participação feminina com apenas 19,06% do total. Esses achados regionais sugerem fortemente o efeito positivo das políticas de interiorização e democratização do ensino superior implementadas nas regiões historicamente periféricas, onde as mulheres encontram maior representatividade e oportunidades nos cursos de computação e engenharias correlatas.

A análise por modalidade de ingresso demonstra que o sistema de cotas consolidou-se como o principal mecanismo de ampliação da diversidade. Entre 2018 e 2023, as cotas representaram mais da metade das aprovações com 55,38% do total, equivalente a 38.616 aprovações, enquanto a ampla concorrência respondeu por 44,62% com 31.109 aprovações. Notavelmente, as cotas exibem uma taxa de conversão candidato-aprovado ligeiramente superior de 9,42% comparada aos 9,07% da ampla concorrência, confirmando a efetividade desse mecanismo. As mulheres têm presença proporcionalmente maior entre cotistas, representando 19,17% dos aprovados nessa modalidade contra apenas 16,77% na ampla concorrência, diferença que se mostrou estatisticamente muito significativa através do teste qui-quadrado. Essa evidência reforça substancialmente o papel das ações afirmativas como ferramenta de equidade de gênero nos cursos de engenharia. Embora cotistas apresentem média de notas inferior com diferença de 47,05 pontos em relação aos aprovados por ampla concorrência, essa diferença reflete precisamente o objetivo da política de ampliar o acesso a estudantes historicamente sub-representados. As diferenças de desempenho por gênero dentro de cada modalidade revelam padrões distintos: na ampla concorrência a

diferença entre mulheres e homens é moderada com 5,63 pontos a menos para as mulheres, enquanto nas cotas essa diferença aumenta para 11,72 pontos. Esses resultados indicam que as políticas afirmativas efetivamente beneficiam estudantes com perfis acadêmicos mais diversos, cumprindo plenamente sua finalidade de democratização do ensino superior sem comprometer significativamente o nível acadêmico geral dos ingressantes.

No que se refere à mobilidade estudantil, os dados indicam que aproximadamente um quarto dos aprovados estuda fora do estado de origem, com a taxa de migração interestadual situando-se em 23,6% no período analisado. Esse percentual apresentou pequenas oscilações anuais com amplitude de apenas 3,0 pontos percentuais, variando entre 21,9% e 24,9% ao longo das seis edições analisadas, revelando um padrão de comportamento relativamente consolidado entre os estudantes. Mulheres apresentam taxa de migração ligeiramente superior com 25,0% comparada aos 23,3% dos homens, diferença de 1,6 pontos percentuais que se mostrou estatisticamente muito significativa através do teste qui-quadrado, indicando que a mobilidade acadêmica é um fenômeno discretamente mais presente entre as candidatas aprovadas. A análise de correlação entre taxa de migração e proporção de mulheres nas 107 instituições com maior volume de aprovados revelou coeficiente praticamente nulo de 0,001 e estatisticamente não significativo, demonstrando que a mobilidade institucional por si só não constitui fator determinante para o aumento da diversidade de gênero nos cursos analisados. No entanto, as instituições com maior capacidade de atração de estudantes migrantes, como a Universidade Tecnológica Federal do Paraná e institutos federais catarinenses, apresentam taxas de migração superiores a 60%, sugerindo que a concentração de oportunidades educacionais especializadas e a reputação institucional exercem influência relevante sobre os fluxos migratórios estudantis, embora sem impacto direto na composição de gênero dos ingressantes.

O desempenho no ENEM confirma que as diferenças de gênero existem de forma consistente, mas são de magnitude moderada com tamanho de efeito pequeno segundo o índice de Cohen. Em todas as edições analisadas, as notas médias masculinas foram superiores às femininas com diferença geral de 27,6 pontos, sendo o hiato temporal variável entre 18,4 pontos em 2023 e 33,9 pontos em 2019. Observa-se tendência de redução do gap após 2021, com diminuição de 13,3 pontos entre 2018 e 2023, embora o teste de Kendall indique que essa convergência ainda não atingiu significância estatística. As disparidades regionais revelam padrões estruturais importantes, com o Nordeste apresentando o maior gap de gênero em 31,2 pontos, enquanto o Norte registra a menor diferença com 20,3 pontos, sugerindo que fatores geográficos e socioeconômicos influenciam diferencialmente o desempenho por gênero. A análise preditiva baseada em modelos de Random Forest e HML, indicam que as variáveis territoriais e institucionais, como unidade da federação, ano e região de origem, explicam uma parcela mais significativa da variação das notas do que o gênero isoladamente, com tamanho de efeito de Cohen de apenas 0,328. Esses resultados reforçam que as desigualdades observadas são fundamentalmente estruturais e

se relacionam à qualidade diferenciada da educação básica, às condições socioeconômicas de preparação para o exame e à distribuição historicamente desigual das oportunidades educacionais entre regiões do país.

A análise regional evidencia disparidades marcantes que refletem padrões estruturais de desenvolvimento educacional no país. As notas mais altas concentram-se consistentemente nas regiões Sudeste com média de 597,7 pontos e Sul com 587,4 pontos, enquanto o Norte registra 564,0 pontos e o Nordeste 569,9 pontos, estabelecendo um diferencial de 33,7 pontos entre as regiões extremas. Em todas as regiões os homens apresentam desempenho superior, mas a análise por faixas de Índice de Desenvolvimento Humano revela que o hiato de gênero é moderadamente mais acentuado em estados de IDH baixo com gap de 25,4 pontos comparado aos 21,3 pontos em regiões de IDH alto, diferencial de 4,1 pontos que confirma o condicionamento socioeconômico das desigualdades de gênero. A distribuição dos candidatos espelha essas disparidades, com 61,2% provenientes de regiões de IDH médio, apenas 15,8% de IDH baixo e 23,1% de IDH alto, evidenciando menor acesso de estudantes das regiões menos desenvolvidas. Os padrões de mobilidade estudantil reforçam essas concentrações, com o Nordeste recebendo 39,1% de todas as aprovações e o Sudeste 30,8%, enquanto o Norte retém apenas 4,5% dos aprovados, demonstrando forte atração das regiões com maior oferta institucional. O saldo migratório regional confirma essa dinâmica, com o Sul apresentando forte atração de 4.695 estudantes e o Norte forte emissão de 2.879 candidatos. Esses resultados indicam que o desempenho feminino está fundamentalmente condicionado por fatores socioeconômicos e territoriais estruturais, sugerindo que políticas de equidade precisam atuar desde a base da formação educacional, fortalecendo as redes públicas de ensino médio nas regiões menos desenvolvidas e promovendo programas específicos de preparação para o ENEM direcionados a meninas e jovens de baixa renda, com especial atenção às disparidades regionais de acesso e qualidade educacional.

De maneira geral, os resultados demonstram que o sistema de cotas cumpre um papel fundamental na ampliação da participação feminina e na diversificação do perfil dos ingressantes. Ao mesmo tempo, a mobilidade interestadual, embora relevante para a redistribuição geográfica de matrículas, não exerce efeito transformador sobre a desigualdade de gênero. O comportamento migratório parece ser determinado mais pela oferta de cursos e pela infraestrutura educacional de cada região do que por preferências individuais. Esse padrão reflete um processo de centralização de oportunidades em torno das regiões com maior densidade de instituições e de investimento público.

Do ponto de vista das políticas públicas, as análises sugerem a necessidade de um conjunto integrado de ações. É fundamental fortalecer o ensino médio público, especialmente nas áreas de matemática, ciências e linguagens, que são determinantes para o desempenho no ENEM. Programas de reforço e orientação vocacional voltados a alunas podem contribuir para reduzir o hiato de desempenho e estimular o ingresso em cursos

de computação. Também é recomendável a ampliação das políticas de permanência, incluindo moradia estudantil, bolsas e apoio psicológico, de forma a garantir que o acesso se converta em sucesso acadêmico. Para as universidades, é essencial implementar sistemas de acompanhamento contínuo que monitorem o ingresso, a evasão e o desempenho dos estudantes por gênero e modalidade de ingresso, possibilitando intervenções direcionadas e tempestivas.

No âmbito institucional, a promoção de um ambiente acadêmico acolhedor e diverso é decisiva para consolidar os avanços observados. A presença de professoras e pesquisadoras como modelos de referência, o estímulo à participação de mulheres em grupos de pesquisa e projetos de extensão, e a visibilidade de trajetórias femininas na área de computação são estratégias eficazes para promover pertencimento e reduzir a evasão. A criação de redes de mentoria e comunidades de prática voltadas a estudantes de grupos sub-representados também pode favorecer a permanência e o engajamento acadêmico.

As limitações do estudo devem ser consideradas na interpretação dos resultados. A base de dados utilizada abrange apenas instituições participantes do SISU e não inclui matrículas obtidas por vestibulares próprios ou outras formas de ingresso. A análise de mobilidade considerou apenas deslocamentos entre estados, o que não permite observar migrações intra-regionais relevantes. Além disso, as informações de cotas são autodeclaradas e podem conter inconsistências pontuais, por escolha do tema não extratificamos os tipos de cotas sociais e raciais, sendo elas consideradas como o mesmo recorte. Por fim, o intervalo de seis anos é suficiente para identificar estabilidade, mas não para avaliar transformações de longo prazo, que demandariam séries mais extensas.

Apesar dessas limitações, o conjunto de evidências é robusto e consistente. A triangulação entre estatísticas descritivas, testes de diferença de médias e modelos de importância de variáveis reforça a confiabilidade das conclusões. Os resultados indicam que o progresso na diversidade de gênero nas áreas de computação está ocorrendo, mas em ritmo lento e dependente de políticas estruturais. A ampliação da presença feminina parece estar fortemente vinculada à efetividade das ações afirmativas, à expansão territorial das oportunidades e à redução das desigualdades socioeconômicas regionais.

Este estudo confirma que a desigualdade de gênero no acesso aos cursos de computação persiste, mas pode ser mitigada por políticas de base e estratégias institucionais articuladas. O hiato de desempenho entre homens e mulheres é real, mas moderado e, sobretudo, reflexo de condições estruturais e históricas, e não de diferenças de capacidade. A continuidade e o aprimoramento das políticas de cotas, associadas ao fortalecimento da educação básica e ao apoio à permanência, representam caminhos promissores para consolidar uma trajetória de inclusão mais ampla e sustentável. A criação de ambientes acadêmicos mais acolhedores e diversos é essencial não apenas para reduzir as disparidades existentes, mas também para estimular a inovação, a criatividade e o pensamento crítico que caracterizam a área de computação (WALTON; COHEN, 2011; DENNEHY;

DASGUPTA, 2017; DASGUPTA; STOUT, 2014).

O cenário observado entre 2018 e 2023 indica que a transformação está em curso, embora ainda incompleta. O avanço da participação feminina, mesmo que gradual, demonstra que políticas públicas bem estruturadas e esforços institucionais coordenados são capazes de gerar resultados concretos em um campo historicamente desigual. Consolidar essa tendência requer investimento contínuo, monitoramento com base em dados e compromisso com a equidade em todos os níveis de formação. Assim, o ecossistema de computação poderá evoluir não apenas em excelência técnica, mas também em diversidade e representatividade, tornando-se um espaço mais justo e inclusivo, condizente com os valores contemporâneos de ciência e sociedade.

6.1 TRABALHOS FUTUROS

Os resultados desta pesquisa abrem caminhos promissores para trabalhos futuros que aprofundem a compreensão sobre desigualdades de gênero e regionais no acesso aos cursos de computação. Uma possibilidade imediata é a ampliação temporal da análise, incorporando edições mais recentes do SISU e examinando os impactos das mudanças na Lei de Cotas implementadas a partir de 2023. A integração de outras bases de dados, como o Censo da Educação Superior, o ENEM e registros institucionais de evasão e permanência, também representa uma oportunidade relevante para traçar um panorama longitudinal mais completo das trajetórias acadêmicas. A aplicação de técnicas estatísticas e computacionais mais avançadas, como modelos hierárquicos, aprendizado de máquina e métodos causais, pode contribuir para identificar relações mais complexas entre variáveis socioeconômicas, institucionais e de desempenho, oferecendo evidências robustas para a formulação de políticas públicas de inclusão.

Outra frente de investigação envolve abordagens qualitativas que explorem as experiências e percepções de estudantes nos cursos de computação, com foco em fatores como pertencimento, acolhimento institucional e barreiras simbólicas. Estudos de caso em universidades com índices elevados de diversidade de gênero poderiam revelar boas práticas, como programas de mentoria, redes de apoio e políticas de permanência estudantil. Além disso, a criação de indicadores nacionais e painéis públicos de monitoramento da diversidade de gênero na computação permitiria acompanhar de forma contínua a evolução da representatividade e do desempenho feminino no ensino superior. Essas iniciativas, combinadas a metodologias quantitativas e qualitativas, poderão fortalecer a base de evidências necessária para a construção de ambientes acadêmicos mais equitativos, inclusivos e sustentáveis.

REFERÊNCIAS

- AGRESTI, Alan. **Statistical Methods for the Social Sciences**. 5. ed. [S.l.]: Pearson, 2018.
- AL., Fabian Pedregosa et. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.
- AL., Pauli Virtanen et. SciPy 1.0: fundamental algorithms for scientific computing in Python. **Nature Methods**, v. 17, p. 261–272, 2020.
- AMARAL, Marília Abrahão *et al.* Investigando questões de gênero em um curso da área de Computação. **Revista Estudos Feministas**, Florianópolis, v. 25, n. 2, p. 857–874, 2017. DOI: 10.1590/1806-9584.2017v25n2p857. Disponível em: <https://doi.org/10.1590/1806-9584.2017v25n2p857>. Acesso em: 1 nov. 2024.
- BARNETT, Vic; LEWIS, Toby. **Outliers in Statistical Data**. Chichester, UK: Wiley, 1994.
- BASSO-POLETTTO, Daniela; EFROM, Cora; BEATRIZ-RODRIGUES, Maria. Ações Afirmativas no Ensino Superior: revisão quantitativa e qualitativa da literatura. **Revista Electrónica Educare**, v. 24, n. 1, p. 292–325, 2020. Epub 30 Jan 2020, ISSN 1409-4258. DOI: 10.15359/ree.24-1.16. Disponível em: https://www.scielo.sa.cr/scielo.php?script=sci_arttext&pid=S1409-42582020000100292.
- BONINI, P.; SILVA, F. da; SÓTERO, G. Wage Profile and Gender Gap in Science and Technology: Regional Disparities in Brazil. **Sociedad y Economía**, n. 55, e10114358, 2025. DOI: 10.25100/sye.v0i55.14358. Disponível em: <https://doi.org/10.25100/sye.v0i55.14358>.
- BORGATTI, S. P. *et al.* Network analysis in the social sciences. **Science**, v. 323, n. 5916, p. 892–895, 2009. DOI: 10.1126/science.1165821.
- BRASIL. **Lei nº 12.711, de 29 de agosto de 2012**. 2012. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2012/lei/112711.htm. Acesso em: 17 nov. 2024.
- BREIMAN, Leo. Random Forests. **Machine Learning**, v. 45, p. 5–32, 2001.
- CASTILHO, Mércia de Souza Maguerroski. **Análise de perfil, desempenho e provas do ENADE do curso de Sistemas de Informação da UFSC**. Dez. 2024. Trabalho de Conclusão de Curso (Bacharelado em Sistemas de Informação) – Universidade Federal de Santa Catarina, Florianópolis. Disponível em: <https://repositorio.ufsc.br/handle/123456789/255711>. Acesso em: 1 dez. 2024.

CASTRO, Bruna *et al.* Gender Disparity in Engineering Courses in Brazil. **IEEE Access**, v. 12, p. 159784–159793, 2024. DOI: 10.1109/ACCESS.2024.3483660.

CONSORTIUM, Unicode. **The Unicode Standard**. [S.l.: s.n.]. <https://unicode.org/>. Documenta o padrão Unicode, mapeamentos e arquivos de conversão, acesso em 2025-11-06.

_____. **Unicode Standard Annex 21: Case Mappings**. [S.l.: s.n.]. <https://unicode.org/reports/tr21/>. acessado em 2025-11-05.

COSTA, Luis Felipe Coimbra *et al.* Women’s Journey in STEM Education in Brazil: A Rapid Review on Engineering and Computer Science. **IEEE Access**, v. 12, p. 112576–112593, 2024. DOI: 10.1109/ACCESS.2024.3442879.

CRESWELL, John W. **Research Design: Qualitative, Quantitative, and Mixed Methods Approaches**. 4. ed. [S.l.]: Sage Publications, 2014.

CULLINAN, John; DUGGAN, Jim. A School-Level Gravity Model of Student Migration Flows to Higher Education Institutions. **Spatial Economic Analysis**, v. 11, n. 3, p. 294–314, 2016. Estudo empírico que aplica especificação gravitacional a fluxos de alunos de escolas secundárias para IES (Irlanda). DOI: 10.1080/17421772.2016.1177195. Disponível em: <https://ideas.repec.org/a/taf/specan/v11y2016i3p294-314.html>.

CUNHA, Vanessa. **TCC — Desempenho no SISU por Gênero em cursos de computação: O que revelam os dados? (repositório GitHub)**. [S.l.: s.n.], 2025. <https://github.com/CunhaVanessa/TCC>. Repositório GitHub, commit 816db96f3d9df5f4894ae0a6ad31c41956f23e48, acesso em 2025-11-08.

CURSINO, Ariana R.; MARTINEZ, Juliana F. P. **Análise Estatística Descritiva e Regressão da Inserção das Mulheres nos Cursos de TI nos Anos de 2009 a 2018**. S.l.: s.n. 2021. Disponível em: <https://sol.sbc.org.br/index.php/wit/article/view/15838>. Acesso em: 26 set. 2021.

DASGUPTA, Nilanjana; STOUT, Jane G. Girls and women in science, technology, engineering, and mathematics: STEMing the tide and broadening participation in STEM careers. **Policy Insights from the Behavioral and Brain Sciences**, v. 1, n. 1, p. 21–29, 2014. DOI: 10.1177/2372732214549471. Disponível em: <https://doi.org/10.1177/2372732214549471>.

DASU, Tamraparni; JOHNSON, Theodore. Exploratory Data Mining and Data Cleaning. **Wiley Series in Probability and Statistics**, 2003.

DENNEHY, Tara C.; DASGUPTA, Nilanjana. Female peer mentors early in college increase women’s retention in engineering. **Proceedings of the National Academy of**

Sciences of the United States of America, v. 114, n. 23, p. 5964–5969, 2017. DOI: 10.1073/pnas.1613117114. Disponível em: <https://www.pnas.org/content/114/23/5964>.

FAYYAD, Usama M.; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 17, n. 3, p. 37–54, 1996. Versão relacionada também publicada em *Communications of the ACM*, 39(11):27–34, 1996.

GELMAN, Andrew; HILL, Jennifer. **Data Analysis Using Regression and Multilevel/Hierarchical Models**. [S.l.]: Cambridge University Press, 2007.

GUO, Huifeng *et al.* DeepFM: A Factorization-Machine based Neural Network for CTR Prediction. In: PROCEEDINGS of the 26th International Joint Conference on Artificial Intelligence (IJCAI). [S.l.: s.n.], 2017. P. 1725–1731. Apresenta combinação de embeddings e interação explícita de features, frequentemente citada em trabalhos sobre representação de categorias.

GUZMAN, I. R. *et al.* Gender Equality Policies in STEM in Latin America – A Systematic Literature Review. In: ROCHA, Á. *et al.* (Ed.). **Information Technology and Systems (ICITS 2024)**. Cham: Springer, 2024. v. 933. (Lecture Notes in Networks and Systems). DOI: 10.1007/978-3-031-54256-5_39. Disponível em: https://doi.org/10.1007/978-3-031-54256-5_39.

HABERMAN, S. J. The analysis of residuals in cross-classified tables. **Biometrics**, v. 29, n. 1, p. 205–220, 1973. DOI: 10.2307/2529141.

HARRIS, Charles R.; MILLMAN, K. Jarrod; WALT ET AL., Stéfan J. van der. Array programming with NumPy. **Nature**, v. 585, p. 357–362, 2020.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2. ed. [S.l.]: Springer, 2009.

HERRING, Christopher. Does Diversity Pay? Race, Gender, and the Business Case for Diversity. **American Sociological Review**, v. 74, n. 2, p. 208–224, 2009. DOI: 10.1177/000312240907400203. Disponível em: <https://www.jstor.org/stable/27736043>.

HINTZE, Jeffrey L.; NELSON, Randolph D. Violin Plots: A Box Plot-Density Trace Synergism. **The American Statistician**, v. 52, n. 2, p. 181–184, 1998. DOI: 10.1080/00031305.1998.10480559. Disponível em: <https://doi.org/10.1080/00031305.1998.10480559>.

HIRSCH, Robert M. *et al.* Techniques of Trend Analysis for Monthly Water Quality Data. **Water Resources Research**, v. 18, n. 1, p. 107–121, 1982.

HUNTER, John D. Matplotlib: A 2D graphics environment. **Computing in Science and Engineering**, v. 9, n. 3, p. 90–95, 2007.

IANACHARSETS. **IANA Character Sets**. [*S.l.: s.n.*].

<https://www.iana.org/assignments/character-sets/character-sets.xhtml>. Registro oficial de nomes de conjuntos de caracteres, acesso em 2025-11-06.

IGLEWICZ, B.; HOAGLIN, D. C. **How to Detect and Handle Outliers**. Milwaukee, WI: ASQC Quality Press, 1993.

INC., Kaggle. State of Data Science and Machine Learning Survey 2023, 2023. Disponível em: <https://www.kaggle.com/surveys/2023>.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP). **Painel de Estatísticas do Censo da Educação Superior**. 2024. Disponível em: <https://app.powerbi.com/view?r=eyJrIjoimGJiMmNiNTAtOTY1OC00ZjUzLTg2OGUtMjAzYzNiYTA5YjliIiwidCI6IjI2ZjczODk3LWM4YWMtNGIxZS05NzhmLWVhNGMwNzc0MzRiZiJ9&pageName=ReportSection4036c90b8a27b5f58f54>. Acesso em: 1 nov. 2024.

JUPYTER, Project. **Project Jupyter Documentation**. [*S.l.: s.n.*], 2023. Disponível em: <https://jupyter.org>.

KENDALL, M. G. **Rank Correlation Methods**. 4. ed. London: Griffin, 1975.

KITCHIN, Rob. Big Data, new epistemologies and paradigm shifts. **Big Data & Society**, v. 1, n. 1, 2014.

KNUTH, Donald E. Literate Programming. **The Computer Journal**, v. 27, n. 2, p. 97–111, 1984.

LAIRD, Nan M.; WARE, James H. Random-effects models for longitudinal data. **Biometrics**, v. 38, n. 4, p. 963–974, 1982.

LIMA, Elizeth Gonzaga dos Santos; MALANGE, Fernando Cezar Vieira; BORGES, Luiz Francisco. Implicações do SiSU no processo de democratização do acesso à educação superior: o caso da Unemat. **Acta Scientiarum. Education**, v. 40, n. 1, e37656, 2018. DOI: 10.4025/actascieduc.v40i1.37656. Disponível em: <https://periodicos.uem.br/ojs/index.php/ActaSciEduc/article/view/37656>. Acesso em: 1 nov. 2024.

LITTLE, Roderick J. A.; RUBIN, Donald B. **Statistical Analysis with Missing Data**. 2. ed. Hoboken, NJ: Wiley, 2002. Segunda edição, texto de referência sobre tratamento de dados faltantes. ISBN 0-471-16429-0.

LOPES, Ramon *et al.* Fatores associados à evasão de calouros no ensino superior: um estudo com dados da Universidade Federal do Recôncavo da Bahia. **Revista Brasileira de Educação**, v. 28, e280042, 2023. DOI: 10.1590/S1413-24782023280042.

Disponível em: <https://doi.org/10.1590/S1413-24782023280042>. Acesso em: 7 set. 2025.

MAIA, Marcel Maggion. Limites de gênero e presença feminina nos cursos superiores brasileiros do campo da computação. **Cadernos Pagu**, v. 46, p. 223–244, 2016. DOI: 10.1590/18094449201600460223. Disponível em:

<https://doi.org/10.1590/18094449201600460223>. Acesso em: 21 abr. 2025.

MANN, Henry B. Nonparametric tests against trend. **Econometrica**, v. 13, n. 3, p. 245–259, 1945.

MCKINNEY, Wes. **Data Analysis with Python and Pandas**. [S.l.]: O'Reilly Media, 2010.

MIKOLOV, Tomas *et al.* **Efficient Estimation of Word Representations in Vector Space**. [S.l.: s.n.], 2013. arXiv:1301.3781. Disponível em <https://arxiv.org/abs/1301.3781>.

MINISTÉRIO DA EDUCAÇÃO. **Dados Abertos do SISU**. [S.l.: s.n.], 2024. Acesso em: 27 out. 2025. Disponível em: <https://dadosabertos.mec.gov.br/>.

MINISTÉRIO DA EDUCAÇÃO DO BRASIL (MEC). **Repositório de Dados do Sistema de Seleção Unificada (SISU)**. 2024. Disponível em: <https://dadosabertos.mec.gov.br/sisu>. Acesso em: 1 nov. 2024.

MOSS-RACUSIN, Corinne A. *et al.* Science faculty's subtle gender biases favor male students. **Proceedings of the National Academy of Sciences**, v. 109, n. 41, p. 16474–16479, 2012. DOI: 10.1073/pnas.1211286109. Disponível em: <https://www.pnas.org/content/109/41/16474>.

NASCIMENTO, Luciana *et al.* Paridade de Gênero no Ensino Superior em STEM no Brasil: uma análise de 10 anos. *In: ANAIS do XVII Women in Information Technology*. João Pessoa/PB: Sociedade Brasileira de Computação, 2023. P. 217–227. DOI: 10.5753/wit.2023.229472. Disponível em: <https://sol.sbc.org.br/index.php/wit/article/view/25024>.

NASPOLINI, Thiago. **Ações afirmativas: uma análise do comportamento acadêmico de alunos ingressantes em cursos da Universidade Federal de Santa Catarina**. 2017. Dissertação (Mestrado Profissional em Métodos e Gestão em Avaliação) – Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Métodos e Gestão em Avaliação, Florianópolis. Disponível em: <https://repositorio.ufsc.br/xmlui/handle/123456789/182903>. Acesso em: 7 set. 2025.

NEWMAN, M. E. J. **Networks: An Introduction**. Oxford: Oxford University Press, 2010.

NUMFOCUS. **NumFOCUS: supporting open scientific computing projects**. Organização que apoia projetos como pandas, Jupyter e outros. 2024. Disponível em: <https://numfocus.org/>. Acesso em: 1 nov. 2024.

OKABE, Masataka; ITO, Kei. **Color Universal Design (CUD): How to Make Figures and Presentations That Are Friendly to Colorblind People**. [*S.l.: s.n.*], 2008. Disponível em: <https://jfly.uni-koeln.de/color/>.

OLIVEIRA-SILVA, Lígia Carolina; PARREIRA, Vanessa Aparecida Diniz. Barreiras e enfrentamentos de mulheres em carreiras predominantemente masculinas. **Revista Estudos Feministas**, Florianópolis, v. 30, n. 1, e74161, 2022. DOI: 10.1590/1806-9584-2022v30n174161. Disponível em: <https://doi.org/10.1590/1806-9584-2022v30n174161>. Acesso em: 1 nov. 2024.

OSBORNE, Jason W. **Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data**. [*S.l.*]: Sage Publications, 2013.

ØSTERGAARD, Christian R.; TIMMERMANS, Bent; KRISTINSSON, Katrine. Does a different view create something new? The effect of employee diversity on innovation. **Research Policy**, v. 40, n. 3, p. 500–509, 2011. DOI: 10.1016/j.respol.2010.11.004. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0048733310001709>.

OUSRET. **charset-normalizer – the modern alternative to chardet for Python**. [*S.l.: s.n.*]. <https://github.com/Ousret/charset-normalizer>. Documenta detecção e heurísticas modernas para Python 3, acesso em 2025-11-06.

PANDAS DEVELOPMENT TEAM. **pandas: powerful Python data analysis toolkit**. [*S.l.: s.n.*], 2020. Software, <https://pandas.pydata.org/>. Documentação oficial / equipamento do projeto; use quando citar TEAM, 2020. Disponível em: <https://pandas.pydata.org/>. Acesso em: 1 nov. 2024.pearson1900Pelegriini2018_gravitacionalPeng2011Perez2007PlotlyInc2023PousaEtAl2023WgetMozillaUni

Anexos

Desigualdades de Gênero e Mobilidade no Ingresso em Cursos de Computação no Brasil: Uma Análise dos dados do SISU

Vanessa Cunha¹, Cristina Meinhardt¹, Simone Silmara Werner¹

¹Universidade Federal de Santa Catarina (UFSC)
Instituto de Informática e Estatística - INE, Florianópolis, SC – Brasil

vanessa.cunha.cc@gmail.com, cristina.meinhardt@ufsc.br, simone.werner@ufsc.br

Abstract. Throughout the existence of computing programs at UFSC, female representation has been low, motivating this study. The aim was to apply data-analysis methods to identify gender-related patterns in admissions to computing programs at the national level. Using SISU data from 2018 to 2023, we examined profiles of applicants and admitted students, inter-state mobility patterns, and the impact of affirmative policies on women's admission. Variables considered included gender, region of origin, student mobility, and ENEM performance. Results reveal a predominantly male panorama in Computer Science, Computer Engineering, and Information Systems, but with consistent signs of increased female presence: the share of female applicants rose from 21.65% to 24.87% (+3.22 percentage points) and among admitted students from 16.62% to 21.45% (+4.83 p.p.). The quota system emerged as the main diversity mechanism, accounting for 55.38% of admissions and showing higher conversion rates and female share than open competition. We found that 23.6% of admitted students study outside their home state, slightly more among women (25.0% vs. 23.3%). ENEM performance gaps favored men by an average of 27.6 points, with a decreasing trend over time. Regional disparities were pronounced—highest scores concentrated in the Southeast and South—suggesting that inequalities are structurally rooted in basic education quality and uneven distribution of opportunities. Findings inform policies to foster diversity and gender representativeness, with special attention to regional inequalities.

Resumo. Ao longo da existência dos cursos de computação na UFSC observou-se baixa representatividade feminina, o que motivou esta pesquisa. O objetivo foi utilizar metodologias de análise de dados para identificar padrões relacionados a gênero e ingresso em cursos de computação em âmbito nacional. Com dados do Sistema de Seleção Unificada (SISU) de 2018 a 2023, analisaram-se os perfis de candidatos e aprovados, os fluxos de migração entre estados e os efeitos de políticas afirmativas sobre o ingresso de mulheres. Consideraram-se variáveis como gênero, região de origem e mobilidade estudantil, bem como o desempenho no ENEM. Os resultados mostram um cenário majoritariamente masculino em Ciência da Computação, Engenharia de Computação e Sistemas de Informação, mas com sinais consistentes de aumento da presença feminina: a proporção de mulheres entre candidatos subiu de 21,65% para 24,87% (+3,22 p.p.) e entre aprovados de 16,62% para 21,45% (+4,83 p.p.). As políticas de cotas consolidaram-se como principal mecanismo de diversidade, respondendo por 55,38% das aprovações e exibindo maior conversão e proporção feminina que a ampla concorrência. Observou-se ainda que

23,6% dos aprovados estudam fora do estado de origem, com taxa ligeiramente maior entre mulheres (25,0% vs. 23,3%). O hiato de desempenho no ENEM favoreceu homens em média por 27,6 pontos, porém com tendência de redução entre 2018 e 2023. As disparidades regionais foram marcantes: Sudeste e Sul apresentaram as maiores médias, enquanto Norte e Nordeste ficaram atrás, indicando que as desigualdades são essencialmente estruturais e vinculadas à qualidade da educação básica e à distribuição desigual de oportunidades. Os achados subsidiam o desenvolvimento de políticas para promover diversidade e representatividade de gênero, com ênfase nas disparidades regionais identificadas.

1. Introdução

Ao longo das últimas décadas a sub-representação feminina nos cursos de computação tem se mostrado um problema persistente, com consequências que extrapolam o mero desbalanceamento numérico: reduz a diversidade de perspectivas, limita capacidade de inovação e contribui para desigualdades no mercado de trabalho e na produção científica [Herring 2009, Østergaard et al. 2011, World Economic Forum 2020]. No Brasil, relatos históricos e investigações empíricas apontam para trajetórias de exclusão e retração da presença feminina, mesmo em cursos que, em outros momentos, chegaram a apresentar composição majoritariamente feminina [Santos 2018]. Esses padrões resultam de um conjunto interligado de fatores como os antecedentes educacionais, estereótipos de gênero, barreiras institucionais e decisões de carreira, que operam tanto antes do ingresso no ensino superior quanto ao longo da trajetória acadêmica; evidências experimentais sobre vieses institucionais, por exemplo, indicam avaliações e recomendações diferenciadas que desfavorecem mulheres e podem acelerar processos de fuga de talentos [Moss-Racusin et al. 2012]. Em face desse quadro, aumentar o número de matrículas femininas exige mais do que medidas isoladas de acesso: demanda combinar políticas de ingresso com ações de permanência, acolhimento e transformação cultural nos cursos e nas instituições [Basso-Poletto et al. 2020, Veras and Luz Silva 2020]).

No contexto brasileiro, instrumentos institucionais como a Lei de Cotas (Lei nº 12.711/2012) [Brasil 2012] e o Sistema de Seleção Unificada (SISU) [Ministério da Educação do Brasil (MEC) 2024] remodelaram o mapa de oportunidades do ensino superior ao centralizar e transparentizar a oferta de vagas e ao priorizar mecanismos de inclusão socioeconômica e racial; todavia, os efeitos concretos desses instrumentos sobre a diversidade de gênero nas áreas de computação permanecem insuficientemente caracterizados, sobretudo quando considerados recortes espaciais e dinâmicas de mobilidade interestadual [Ministério da Educação do Brasil (MEC) 2024]. A mobilidade estudantil foi impulsionada pelo ENEM/SISU e condicionada por determinantes de atração como o tamanho da oferta, prestígio institucional, diversidade de cursos e também de fricção como a distância, custos de deslocamento, segurança, rede de apoio, é um vetor essencial para compreender como a redistribuição de vagas pode alterar a representatividade feminina em polos universitários, mas suas interações com gênero e modalidades de ingresso ainda são pouco exploradas na literatura nacional [Pelegrini et al. 2018, Sá et al. 2003].

Revisões e estudos recentes sobre paridade de gênero em STEM e avaliações institucionais oferecem subsídios metodológicos e empíricos relevantes: análises longitudi-

nais do Censo da Educação Superior apontam tendências agregadas de presença feminina em diferentes áreas [Nascimento et al. 2023], investigações institucionais baseadas em microdados demonstram o potencial analítico do ENADE para perfilar ingressantes e concluintes [Castilho 2024]), e estudos sobre mobilidade aplicam especificações gravitacionais capazes de estimar elasticidades de fluxo entre unidades regionais [Pelegrini et al. 2018]. Ainda assim, lacunas persistem, em especial a rara articulação integrada entre (i) análises de gênero desagregadas por curso e modalidade de ingresso, (ii) modelagem espacial dos fluxos origem–destino por unidade federativa e (iii) avaliação do papel combinado de políticas afirmativas e vieses de gênero quanto ao desempenho que são justamente o foco deste estudo.

Guiado por essas lacunas, o presente trabalho propõe analisar, com base nos microdados públicos do ENEM e do SISU (2018–2023) e apoiando-se num fluxo reprodutível de KDD, como gênero, desempenho, modalidade de ingresso e mobilidade interestadual interagem para moldar perfis de candidatos e aprovados nos cursos de Ciência da Computação, Engenharia de Computação e Sistemas de Informação. Metodologicamente, a investigação integra testes de associação categórica e medidas de efeito a exemplo o testes de qui-quadrado, diagnósticos célula a célula, técnicas para detecção de tendências temporais robustas como o teste de Mann–Kendall e estimador de Sen, abordagens de descoberta preditiva (Random Forest) e modelos hierárquicos para decompor variância entre níveis individual e institucional [Pearson 1900, Mann 1945, Sen 1968, Breiman 2001, Raudenbush and Bryk 2002, Agresti 2018]. As visualizações de fluxo utilizando Sankey e a análise de redes complementam a investigação espacial, enquanto o uso de ferramentas e pipelines reprodutíveis consolida a transparência e a possibilidade de extensão por gestores e pesquisadores [pandas development team 2020, Hunt et al. 2015, Plotly Technologies Inc. 2023].

Espera-se que a integração desses vetores analíticos produza evidência desagregada e espacializada capaz de informar políticas públicas e intervenções institucionais que articulem acesso, permanência e condições locais de estudo, enfatizando a importância de medidas complementares às cotas para converter aumento de matrículas em efetivo avanço na representatividade e sucesso acadêmico das mulheres nas áreas de computação. As questões orientadoras incluem, de modo sintético, a caracterização do perfil de candidatos e aprovados por gênero, região e modalidade; a avaliação de diferenças de desempenho no ENEM por sexo e por curso; a análise da distribuição de notas considerando efeitos de curso e modalidade; a identificação de padrões de mobilidade origem–destino e sua relação com a presença feminina; e a detecção de tendências temporais na participação feminina entre 2018 e 2023, elementos que, combinados, permitem inferir políticas e práticas mais efetivas para promoção de equidade em STEM.

2. Metodologia

A metodologia adotada neste estudo combina procedimentos de ciência de dados reprodutíveis com técnicas estatísticas clássicas e modelos de aprendizagem de máquina, visando responder às cinco questões de pesquisa que orientam a investigação:

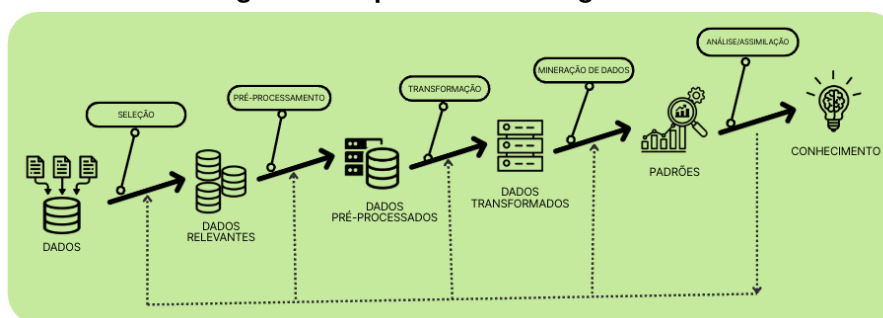
- Q1. Como se caracteriza o perfil dos candidatos e aprovados aos cursos de Ciência da Computação, Engenharia da Computação e Sistemas de Informação em termos de gênero, região de origem e modalidade de ingresso no período de 2018 a 2023?;

- Q2. Existem diferenças significativas no desempenho médio do ENEM entre candidatos e candidatas a esses cursos, e como tais diferenças se distribuem entre as edições do SISU e entre os cursos analisados?;
- Q3. Qual é a distribuição e o comportamento do desempenho dos candidatos em cada curso, considerando o efeito do curso e da modalidade de concorrência sobre as notas médias?;
- Q4. Que padrões de mobilidade geográfica se observam entre aprovados, isto é, de que forma a migração entre estados (origem $\times$ destino) se relaciona com a representação feminina nos cursos de computação?;
- Q5. Como evoluiu, nas edições de 2018 a 2023, a proporção de mulheres inscritas e aprovadas nesses cursos e se é possível identificar tendências estatisticamente significativas de crescimento ou retração dessa participação?

A análise parte dos microdados públicos do SISU, cobrindo as doze edições semestrais de 2018.1 a 2023.2 disponibilizadas pelo Ministério da Educação, e inclui tanto etapas descritivas quanto fase inferencial e exploratória de descoberta de padrões [Ministério da Educação do Brasil (MEC) 2024].

A condução do trabalho seguiu o fluxo clássico de Knowledge Discovery in Databases, conforme apresentado em Figura 1 adaptado ao contexto desta pesquisa e operacionalizado por pipelines automatizados que documentam todas as etapas desde a ingestão até a geração dos artefatos analíticos [Fayyad et al. 1996]. A Figura 1 ilustra esse processo. O fluxo contempla coleta e seleção dos arquivos brutos, pré-processamento e harmonização, transformação e engenharia de variáveis, mineração e modelagem, e finalmente interpretação e avaliação dos resultados. Em cada etapa foram implantados controles de qualidade, logs e artefatos de auditoria para garantir rastreabilidade e permitir reexecução idêntica do pipeline.

Figura 1. Esquema metodológico KDD



Fonte: A autora

A etapa de coleta articulou downloads assistidos dos conjuntos CSV hospedados no Portal de Dados Abertos do MEC, complementando rotinas automatizadas com estratégias manuais quando o acesso por scripts permaneceu sujeito a restrições técnicas. Os arquivos semestrais foram preservados em formato original e seus resumos criptográficos foram calculados como prova de integridade. O recorte inicial incluiu aproximadamente 21,7 milhões de registros brutos e 10,2 GB de dados, dos quais se extraiu um subconjunto final de 752.735 observações e 54 variáveis harmonizadas relativo às posições efetivas de candidatura e classificação nos cursos-alvo: Ciência da Computação, Engenharia da

Computação e Sistemas de Informação. Essa seleção seguiu a premissa de evitar ambiguidade na identificação de cursos e de priorizar campos essenciais às perguntas de pesquisa, tais como sexo, UF de origem, UF da IES, modalidade de concorrência e notas do ENEM.

O pré-processamento enfatizou a qualidade e a comparabilidade longitudinal. Procedimentos automáticos detectaram e padronizaram encoding de arquivos, normalizaram cabeçalhos e unificaram esquemas variando entre edições. Para garantir consistência tipológica, aplicou-se conversão controlada de tipos numéricos e categóricos, com tratamentos robustos para valores ausentes e formatação de datas. A detecção de outliers combinou critérios baseados em z-score e em intervalo interquartil, calculados de forma estratificada por curso, sexo e edição, de modo a sinalizar valores extremos sem apagá-los automaticamente. Todas as transformações foram registradas em scripts reprodutíveis e em relatórios de consistência, permitindo auditoria e reproprocessamento conforme recomendações de pesquisa reprodutível.

A fase de transformação envolveu a canonicalização textual dos nomes de curso, a criação de variáveis derivadas relevantes para as perguntas de pesquisa e a engenharia de atributos que incorporou conhecimento teórico. Entre as variáveis derivadas destacam-se flags binárias de aprovação, indicadores de mobilidade interestadual, mapeamentos de UFs para macrorregiões e sinalizadores de registros não classificados. Para variáveis numéricas foi utilizada normalização seletiva por z-score ou escala mínima-máxima quando exigido por algoritmos sensíveis à escala. Em atributos categóricos de alta cardinalidade foram avaliadas alternativas como hashing de features ou representações densas, escolhidas conforme trade-offs entre interpretabilidade e custo computacional.

A etapa de mineração foi estruturada em blocos analíticos que combinam descrição, inferência e modelagem exploratória. A caracterização descritiva incluiu tabelas de frequência, tabelas cruzadas estratificadas por ano, curso e sexo, medidas de tendência central e dispersão e visualizações destinadas a detectar assimetrias, multimodalidade e mudanças temporais. Testes estatísticos clássicos foram aplicados para responder questões de associação e comparação: o qui-quadrado de independência para relações entre categorias, testes não paramétricos de comparação de distribuições quando a assimetria excedia limiares operacionais, e o teste t de Welch quando a suposição de simetria permitia sua aplicação. Para a detecção de tendências nas séries temporais de proporções e médias foi empregado o teste de Mann-Kendall juntamente com a estimativa de inclinação de Sen, métodos robustos frente a distribuições não normais e a quebras pontuais na série [Mann 1945, Sen 1968].

A análise explicativa e inferencial mobilizou modelos hierárquicos de efeitos mistos para decompor a variância entre níveis individual e institucional e estimar frações da variação atribuíveis a diferenças entre IES. Esses modelos permitiram incluir efeitos fixos para variáveis como sexo, ano e região, além de interceptos aleatórios por instituição, sendo avaliadas rotinas de diagnóstico de convergência, covariância e resíduos para assegurar validade dos pressupostos. Reconhecendo limitações de modelos lineares em captar relações complexas, a investigação foi complementada com técnicas de aprendizagem de máquina, em especial Random Forest, para identificar preditores com maior poder explicativo e revelar interações não lineares, e com k-means para descoberta de perfis de candidatos que apoiaram análises estratificadas. As aplicações de machine learning seguiram protocolos de validação por partição treino-teste, amostragem controlada em ba-

ses muito grandes e avaliação de estabilidade das métricas de desempenho, sempre sem atribuir automaticamente caráter causal às associações reveladas por métodos preditivos [Breiman 2001].

As representações espaciais e de fluxo constituíram componente fundamental da análise da mobilidade interestadual. Diagramas de fluxo do tipo Sankey foram usados para visualizar rotas dominantes entre UFs de origem e UFs receptoras, e análises de redes ponderadas identificaram papéis centrais de emissores e receptores, permitindo quantificar concentração e dispersão dos fluxos e relacioná-los à presença feminina por rota e por instituição. As representações gráficas foram construídas de forma a preservar rastreabilidade, com exportação de figuras e metadados e com cuidado para acessibilidade cromática.

Todo o processo foi implementado em Python e documentado em notebooks Jupyter gerados programaticamente por scripts de literate programming, o que consolidou a reprodutibilidade das etapas e facilitou a geração de artefatos auditáveis. As bibliotecas centrais incluíram pandas para manipulação tabular, NumPy e SciPy para operações numéricas e testes, statsmodels para modelos estatísticos, scikit-learn para procedimentos de aprendizagem de máquina, Plotly para visualizações interativas e PyMannKendall para análises de tendência, seguindo boas práticas de automação e versionamento do código fonte em repositório público. Tais escolhas técnicas refletem a necessidade de integrar limpeza de dados, análise estatística, modelagem e visualização em um único fluxo reprodutível, conciliando rigor e eficiência operacional [McKinney 2010, VanderPlas 2016, Pérez and Granger 2007].

Finalmente, a estratégia de validação cruzou resultados de métodos distintos para reforçar robustez e interpretar plausivelmente os achados. Testes estatísticos ofereceram evidências de associação e significância, modelos hierárquicos quantificaram a contribuição institucional para variação de notas, e abordagens de machine learning indicaram padrões e preditores relevantes que subsidiaram hipóteses adicionais. A conjunção dessas abordagens, apoiada por pipelines reprodutíveis, relatórios de consistência e arquivos de auditoria, assegura que as respostas às cinco questões de pesquisa se apoiem em evidências verificáveis e em procedimentos metodologicamente consistentes.

3. Resultados e Discussão

Os resultados são organizados em quatro eixos complementares alinhados às perguntas de pesquisa e ao fluxo metodológico. Primeiro apresenta-se o perfil dos candidatos, com ênfase em gênero, região de origem e modalidades de ingresso. Em seguida descreve-se o perfil dos aprovados, incluindo taxas de conversão por ano, por região e por modalidade, e discute-se a persistência de assimetrias de gênero. O terceiro eixo examina o efeito das políticas afirmativas sobre a composição dos ingressantes, comparando padrões por sexo, por região e a distância entre notas e notas de corte. Por fim, avalia-se a mobilidade estudantil e sua relação com a diversidade, analisando fluxos interestaduais, preferências regionais e diferenças de deslocamento por sexo e modalidade. Em todas as partes os achados são apresentados com figuras geradas pelo pipeline reprodutível e interpretados à luz da literatura e das políticas públicas, culminando em uma síntese sobre desempenho no ENEM por perfis demográficos e regionais que fundamenta a discussão final e propostas de pesquisas futuras. As análises e discussões guiadas pelas perguntas definidas

anteriormente estão relacionadas aos resultados apresentados nas Figuras 2 a X, explicadas nas respostas a seguir:

3.1. Q1. Como se caracteriza o perfil dos candidatos e aprovados aos cursos de Ciência da Computação, Engenharia da Computação e Sistemas de Informação em termos de gênero, região de origem e modalidade de ingresso no período de 2018 a 2023?

3.1.1. Perfil dos candidatos

A composição do conjunto de candidatos revela uma masculinidade persistente, expressão numérica de processos históricos que não se desfazem por inércia, conforme apresentado na Figura 2. O contingente masculino decresceu de aproximadamente 101 mil em 2018 para 94 mil em 2023, enquanto o contingente feminino aumentou de cerca de 28 mil para 31 mil no mesmo intervalo; em termos relativos, a participação feminina progrediu de 21,7 por cento em 2018 para 24,9 por cento em 2023, resultando em participação masculina global de 77,6 por cento. Testes de comparação de proporções anuais indicaram diferenças estatisticamente significativas em boa parte do período, e a regressão logística que modelou sexo feminino em função do ano apresentou estatística de Wald significativa e p de tendência menor que 0,05. Esse quadro confirma que a presença feminina cresce, mas que o ritmo da mudança ainda dialoga com estruturas sociais e institucionais diagnosticadas por pesquisas sobre vieses de gênero e barreiras em STEM [Moss-Racusin et al. 2012, UNESCO 2017].

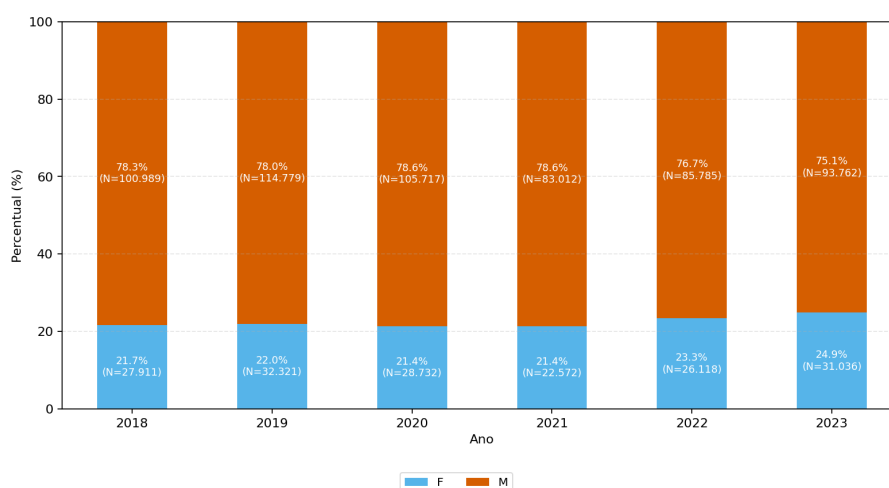


Figura 2. Número de ocorrências por sexo e ano em cursos de Computação no Brasil 2018-2023

A geografia das candidaturas desenha um mapa de oportunidades marcado por concentrações. O Sudeste e o Nordeste concentram a maior parcela das inscrições, conforme a Figura 3 para ambos os sexos, contudo a proporção feminina é relativamente maior no Norte com 28,4 por cento e no Nordeste com 24,4 por cento, ao passo que Sul e Centro Oeste exibem os menores percentuais, 19,4 por cento e 19,2 por cento respectivamente. O teste de independência do qui quadrado resultou em χ^2 igual a 3.680,4 com p inferior a 0,001, e a inspeção dos resíduos padronizados mostrou super-representação relativa de mulheres no Norte e no Nordeste e sub-representação no Sul e no Centro

Oeste. Esses padrões corroboram análises sobre mobilidade estudantil e oferta regional que explicam como atratividade e fricções territoriais direcionam fluxos e reproduzem desigualdades [Pelegrini et al. 2018, Santos 2018, Nascimento et al. 2023].

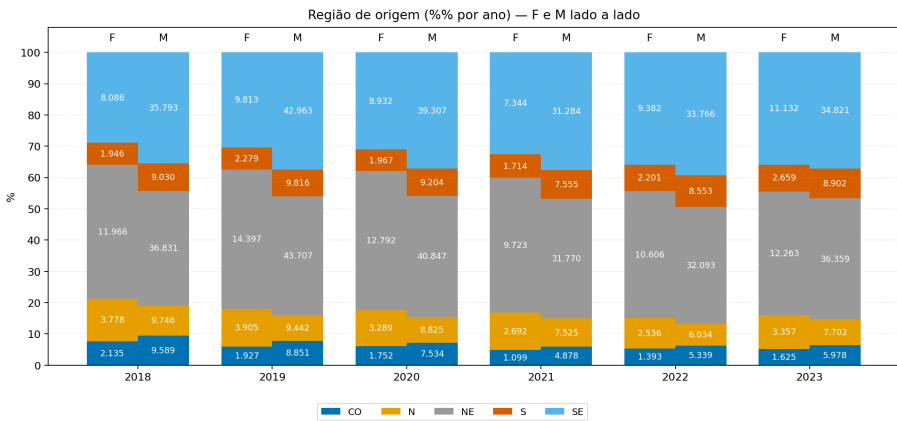


Figura 3. Número de ocorrências por região de origem, sexo e ano em cursos de computação no SISU

O perfil socioespacial também se articula com o Índice de Desenvolvimento Humano. A maior parte das candidaturas provém de estados com IDH médio, com média acumulada de 61,2 por cento e coeficiente de variação de 1,7 por cento, enquanto a faixa de IDH alto cresceu levemente de 21,3 por cento em 2018 para 23,9 por cento em 2023 e a faixa de IDH baixo manteve média próxima de 15,8 por cento com maior variabilidade. Em termos absolutos, aproximadamente 103 mil mulheres e 358 mil homens são oriundos de unidades federativas com IDH médio, 34 mil mulheres e 139 mil homens vêm de estados de alto IDH, e cerca de 31 mil mulheres e 87 mil homens provêm de estados de baixo IDH. Essas evidências reforçam a interpretação de que a qualidade da formação pré-universitária e a distribuição territorial de oportunidades moldam de forma decisiva a composição de gênero entre os candidatos, em consonância com relatórios internacionais sobre educação em STEM [UNESCO 2017].

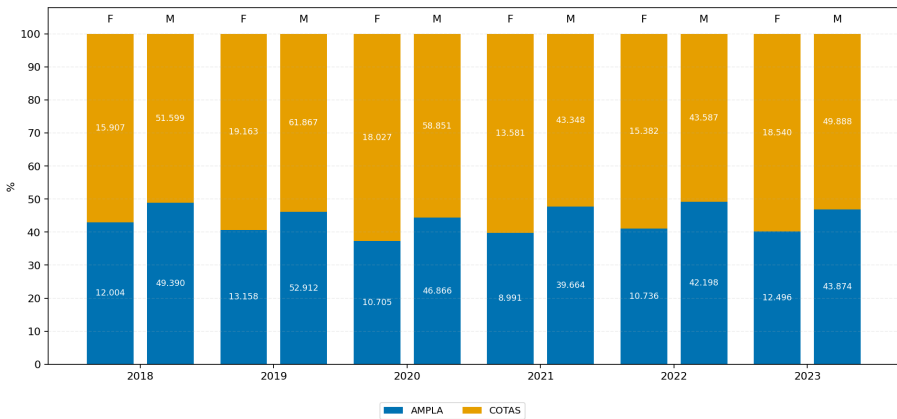


Figura 4. Número de ocorrências por forma de ingresso, sexo e ano em cursos de computação no SISU

A modalidade de ingresso (Figura 4) por cotas surge como mecanismo compen-

satório concreto na configuração do perfil dos candidatos. Modalidades de cotas corresponderam a aproximadamente 52 a 57 por cento das vagas ofertadas no período analisado, e as mulheres apresentaram maior participação relativa nas cotas do que na ampla concorrência, diferença agregada inferior a sete pontos percentuais. A associação entre sexo e modalidade foi estatisticamente significativa com χ^2 igual a 2.372,13 e p inferior a 0,001, enquanto a regressão logística estimou razão de chances OR igual a 1,314 com intervalo de confiança de 95 por cento entre 1,299 e 1,328 em favor da candidatura feminina por cotas. Esses achados confirmam avaliações empíricas que reconhecem o papel das ações afirmativas na diversificação do acesso, ao mesmo tempo em que sublinham a necessidade de políticas complementares de permanência e apoio para traduzir o ingresso em trajetórias de sucesso [Napolini 2017, Basso-Poletto et al. 2020].

A preferência por permanecer no próprio estado de origem, como mostra a Figura 5 e o desempenho médio no ENEM sintetizam a tensão entre conservação e mobilidade, entre mérito individual e condicionamentos estruturais. A maior parte dos candidatos tende a priorizar vagas locais, o que reduz custos de deslocamento e rupturas sociais, mas também reproduz dependência da oferta regional; o desempenho médio geral no ENEM situa-se em torno de 581 pontos, e o hiato de desempenho entre mulheres e homens foi estimado em 27,6 pontos entre os candidatos, reduzindo-se para cerca de 11,0 pontos entre os aprovados, o que indica efeitos seletivos do processo de ingresso. Esses elementos indicam que intervenções que elevem a preparação pré-ENEM nas localidades de menor desempenho e que ofereçam apoio à permanência serão decisivas para que a expansão de acesso se converta em redução efetiva das desigualdades de gênero [Breiman 2001, Veras and Luz Silva 2020].

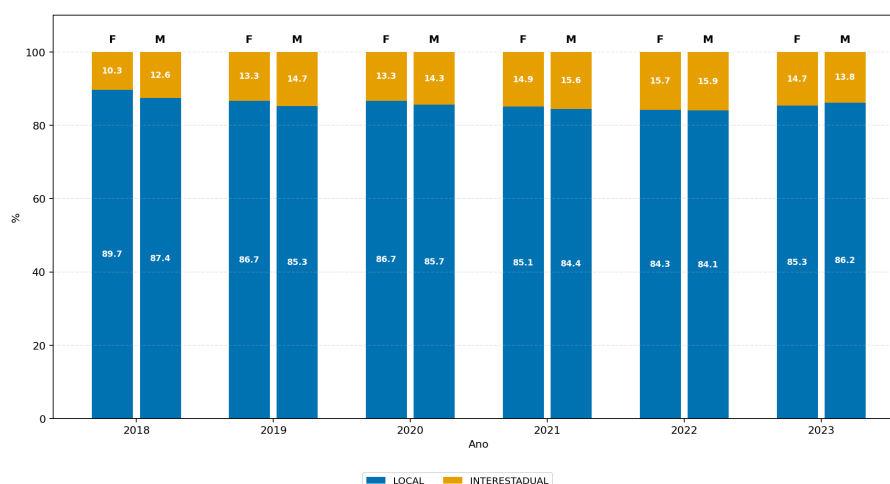


Figura 5. Número de ocorrências por interesse migratório e ano em cursos de computação no SISU

O comportamento do desempenho dos candidatos revela um quadro dual: por um lado existe uma diferença média mensurável entre os sexos, por outro as distribuições apresentam semelhança de forma e grande heterogeneidade interna, conforme apresentado em Figura 6. As médias observadas situam-se em torno de 559,7 pontos para as mulheres, 587,3 pontos para os homens e 581,2 pontos no agregado, resultando em um gap aproximado de 27,6 pontos a favor dos candidatos do sexo masculino; as violas de

densidade mostram medianas masculinas ligeiramente deslocadas para valores mais altos, mas também caudas e variações que evidenciam subgrupos com desempenhos muito distintos. Esse padrão sugere que diferenças de preparo prévio e condições contextuais explicam parte substancial da lacuna, enquanto o próprio processo seletivo tende a concentrar aprovados em faixas próximas à nota de corte, o que torna pequenas variações pontuais decisivas para ingresso. Em termos interpretativos as evidências apontam que intervenções isoladas centradas apenas no gênero terão alcance limitado e que políticas integradas, atuando sobre a qualidade da formação básica, sobre fatores regionais e sobre mecanismos de apoio pré e pós ingresso são necessárias para reduzir efetivamente as desigualdades documentadas, conforme argumentam [Moss-Racusin et al. 2012] e estudos sobre mobilidade e oferta regional como [Pelegrini et al. 2018].

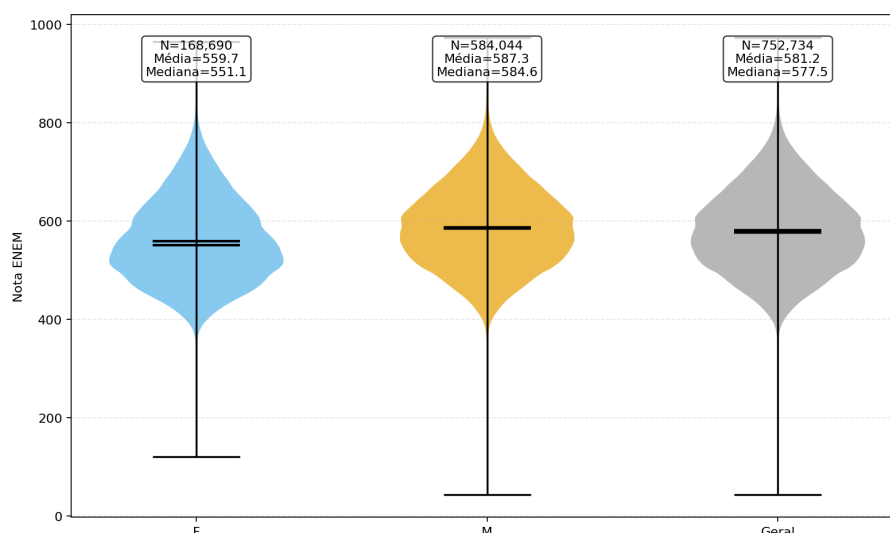


Figura 6. Desempenho acumulado no ENEM por sexo e ano

3.1.2. Perfil dos Aprovados

A composição dos aprovados revela clara masculinidade numérica conforme Figura 7, com mulheres representando entre 16,6 por cento em 2018 e 21,4 por cento em 2023, enquanto os homens correspondem à maior parte dos ingressantes. Essa desproporção traduz não apenas escolhas individuais, mas também a persistência de barreiras estruturais que reduzem a probabilidade de conversão feminina, medida em 7,5 por cento para candidatas contra 9,8 por cento para candidatos, diferença que exige intervenções para além do mero aumento de inscrição. Esses achados dialogam com estudos que identificam vieses institucionais e culturais que dificultam a progressão feminina em STEM [Moss-Racusin et al. 2012, Nascimento et al. 2023].

A geografia dos aprovados concentra-se nos polos do Sudeste e do Nordeste conforme Figura 8, com o Sudeste respondendo por cerca de 40 a 50 por cento das aprovações anuais e o Nordeste ampliando sua participação de 7,4 por cento em 2018 para 8,0 por cento em 2022. Essa distribuição evidencia que a oferta de vagas e a atratividade institucional condicionam fortemente quem ingressa, produzindo mapas de oportunidade que favorecem determinadas regiões em detrimento de outras. A relação entre padrão

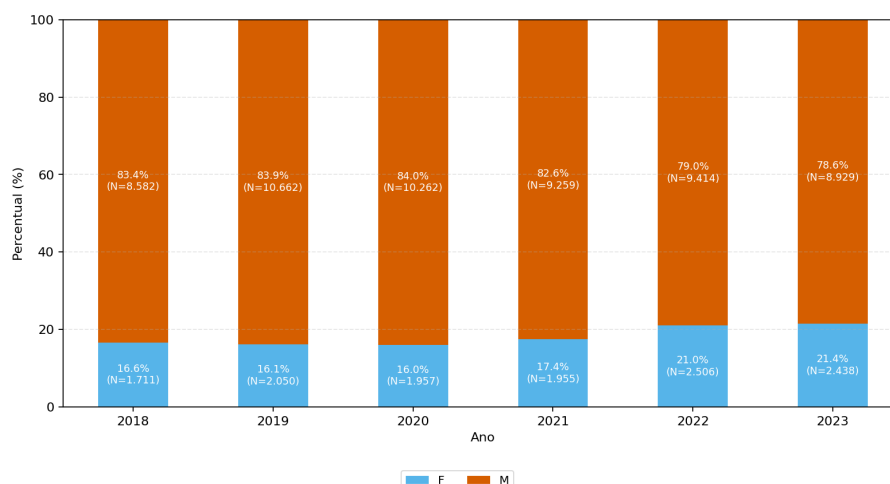


Figura 7. Número de ocorrências por sexo e ano em cursos de computação no SISU

espacial e composição de gênero reforça interpretações em que fricções territoriais e desigualdades de infraestrutura educacional modulam a presença feminina nas coortes de computação [Pelegrini et al. 2018].

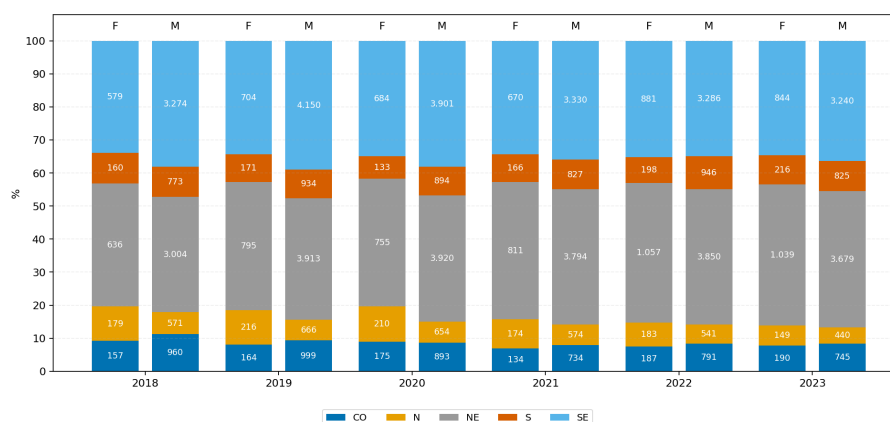


Figura 8. Número de ocorrências por região, sexo e ano em cursos de computação no SISU

A modalidade de ingresso por cotas ocupa papel central na formação do perfil dos aprovados, respondendo por cerca de 55 a 60 por cento das aprovações no período estudado, com participação feminina nas cotas próxima de 19,2 por cento contra 16,8 por cento na ampla concorrência, conforme Figura 9. A estimativa por regressão logística aponta razão de chances de aprovação via cotas para mulheres em 1,314 com intervalo de confiança robusto, indicando efeito compensatório real das ações afirmativas. Contudo, os dados também mostram que a ação afirmativa precisa ser acompanhada de políticas de permanência para que o ingresso se traduza em trajetória acadêmica bem sucedida, recomendação alinhada a avaliações sobre a efetividade das cotas no ensino superior brasileiro [Basso-Poletto et al. 2020, Napolini 2017].

A permanência regional aparece como traço predominante do perfil dos aprova-

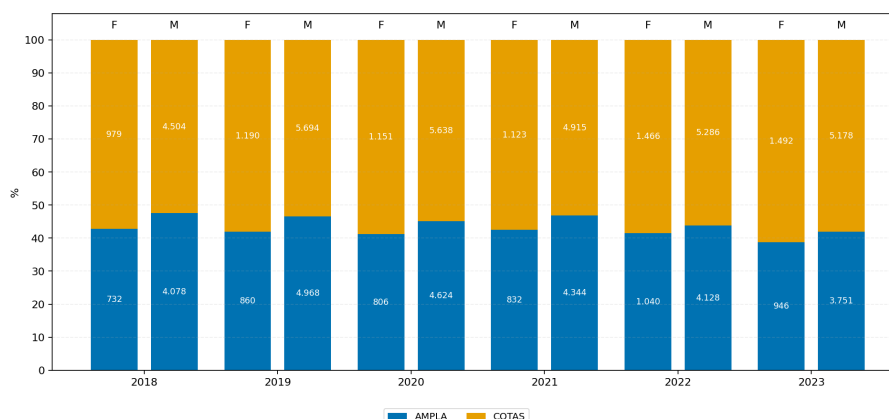


Figura 9. Número de ocorrências por forma de ingresso, sexo e ano em cursos de computação no SISU

dos, já que a maioria tende a estudar no próprio estado de origem e apenas cerca de um quarto dos aprovados realiza deslocamento interestadual, faixa que variou entre 22 e 26 por cento ao longo do período. Entre as mulheres essa propensão à mobilidade é ligeiramente maior em contextos de ampla concorrência, enquanto as cotas contribuem para reduzir a necessidade de migração, permitindo que um número significativo de candidatas ingresse mais próximo de sua rede social e familiar. Esse padrão reafirma que a expansão da oferta local e políticas de apoio logístico são componentes essenciais para que a mobilidade não se converta em obstáculo à participação feminina [Pelegrini et al. 2018, Veras and Luz Silva 2020].

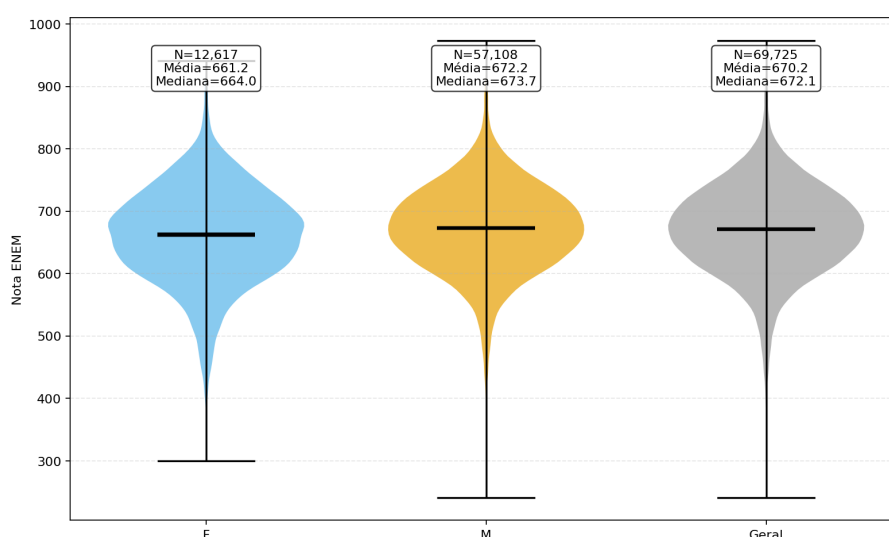


Figura 10. Distribuição do desempenho no ENEM por sexo acumulado(2018-2023)

A Figura 10 apresenta o desempenho médio no ENEM dos aprovados situou-se consistentemente acima da média nacional, com médias da ordem de 670 pontos ou mais segundo o recorte dos cursos analisados, e o gap de desempenho entre sexos mostrou forte redução no conjunto dos aprovados, passando de cerca de 27,6 pontos entre todos os

candidatos para aproximadamente 11,0 pontos entre os ingressantes. Esse padrão indica que o processo seletivo atenua parte da desigualdade de preparação, mas não a elimina, e aponta para a necessidade de políticas que atuem na base da formação escolar além de medidas no nível universitário, de modo a produzir mudança estrutural na equidade de desempenho e de representação [Nascimento et al. 2023, Breiman 2001].

3.2. Q2. Existem diferenças significativas no desempenho médio do ENEM entre candidatos e candidatas a esses cursos, e como tais diferenças se distribuem entre as edições do SISU e entre os cursos analisados?

A evolução temporal do hiato de desempenho entre mulheres e homens mostra redução parcial após 2021 conforme Figura 11, mas sem evidência estatisticamente robusta de convergência no período completo, conforme indica o teste de Mann-Kendall com p igual a 0,1329. Em termos agregados o gap médio entre candidatos é de aproximadamente 27,6 pontos a favor dos homens, o que representa cerca de 4,6 por cento da escala do ENEM, enquanto entre os aprovados essa diferença recua para cerca de 11,0 pontos, ou 1,8 por cento da escala. Esses resultados sugerem que os processos seletivos atenuam parte da desigualdade de preparação prévia, mas não a eliminam, o que aponta para raízes estruturais do hiato que exigem intervenções além do âmbito das universidades [UNESCO 2017, Basso-Poletto et al. 2020].

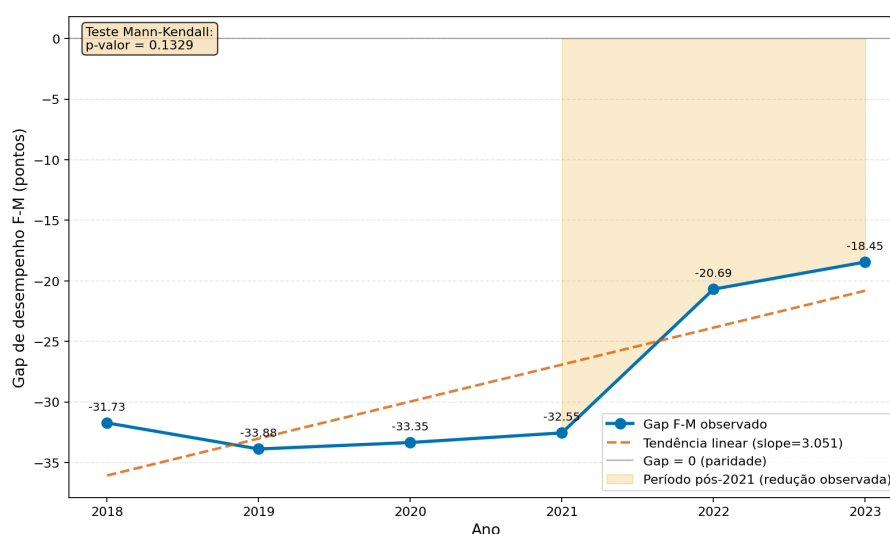


Figura 11. Evolução anual do GAP de notas Masculino X Feminino

Já na Figura 12, temos a decomposição da importância explicativa por meio de um modelo de Random Forest evidencia que fatores estruturais e contextuais exercem maior peso sobre as notas do que a variável sexo isolada. A unidade federativa de oferta explica em torno de 30,4 por cento da importância total, o ano corresponde a 24,1 por cento, a modalidade de ingresso contribui com 18,6 por cento e a região acrescenta 15,7 por cento, enquanto o sexo responde por cerca de 11,1 por cento do efeito explicativo de acordo com a Figura . Esta hierarquia confirma que as variações territoriais e temporais dominam a explicação do desempenho, reforçando a necessidade de políticas direcionadas à base da formação e à equidade regional em vez de medidas que tratem o gênero de forma isolada [Pelegrini et al. 2018, Breiman 2001].

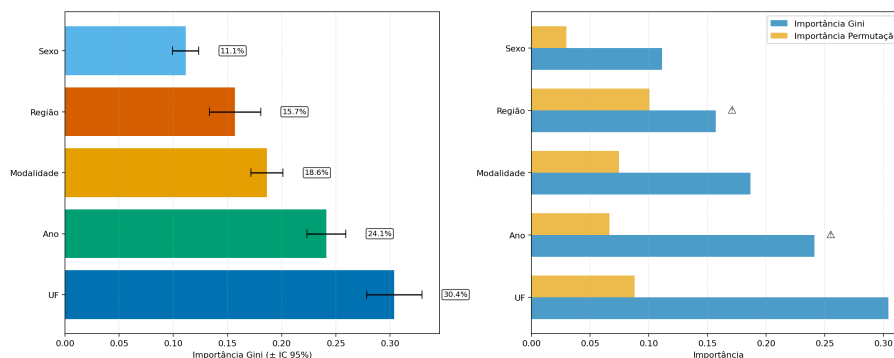


Figura 12. Importância das características de Candidatos em relação a nota do ENEM (2018-2023)

A distribuição das distâncias entre nota individual e nota de corte mostra concentração dos não aprovados imediatamente abaixo do limite, o que indica que pequenas melhorias no desempenho médio de grupos subrepresentados podem produzir ganhos desproporcionais na ocupação de vagas. Em termos práticos, a magnitude estatística do gap existe, mas seu impacto prático é reduzido quando comparado com determinantes regionais e temporais, o que orienta a priorização de programas de preparação para o ENEM, tutoria e políticas de permanência em localidades com menor desempenho, de modo a transformar aumentos modestos de nota em avanços efetivos de representatividade [Basso-Poletto et al. 2020, Veras and Luz Silva 2020].

3.3. Q3. Qual é a distribuição e o comportamento do desempenho dos candidatos em cada curso, considerando o efeito do curso e da modalidade de concorrência sobre as notas médias?

A comparação das distribuições de notas por curso revela diferenças sistemáticas na seletividade implícita das graduações. Engenharia da Computação apresentou a maior média observada, seguida por Ciência da Computação e por Sistemas de Informação. Essa ordem traduz grau diverso de competitividade entre os cursos, em que Engenharia atrai candidatos com médias mais elevadas e maior densidade em patamares superiores da escala do ENEM. As violas de densidade, conforme a Figura 13 evidenciam que as distribuições não são apenas deslocadas em média, mas também diferem em forma, com Engenharia exibindo caudas mais estendidas para valores altos e Sistemas de Informação apresentando maior concentração em torno de médias inferiores. Esses padrões são compatíveis com estudos institucionais que documentam variação de desempenho entre cursos de computação e engenharia e com análises que relacionam prestígio e seletividade à composição dos ingressantes, conforme observado em pesquisas locais e revisões sobre paridade em STEM [Nascimento et al. 2023, Castilho 2024].

A distinção entre ampla concorrência e cotas revela padrões claros: ampla concorrência apresenta média das notas mais elevada e distribuição mais concentrada, enquanto as cotas mostram média inferior e maior dispersão, indicando heterogeneidade interna dos candidatos cotistas. Ao mesmo tempo há ampla sobreposição entre as distribuições, cerca de três quartos dos candidatos com desempenho competitivo situam-se em faixas de nota comuns às duas modalidades, o que sugere que a pertença a uma modalidade não determina de forma absoluta o mérito acadêmico. Esse diagnóstico,

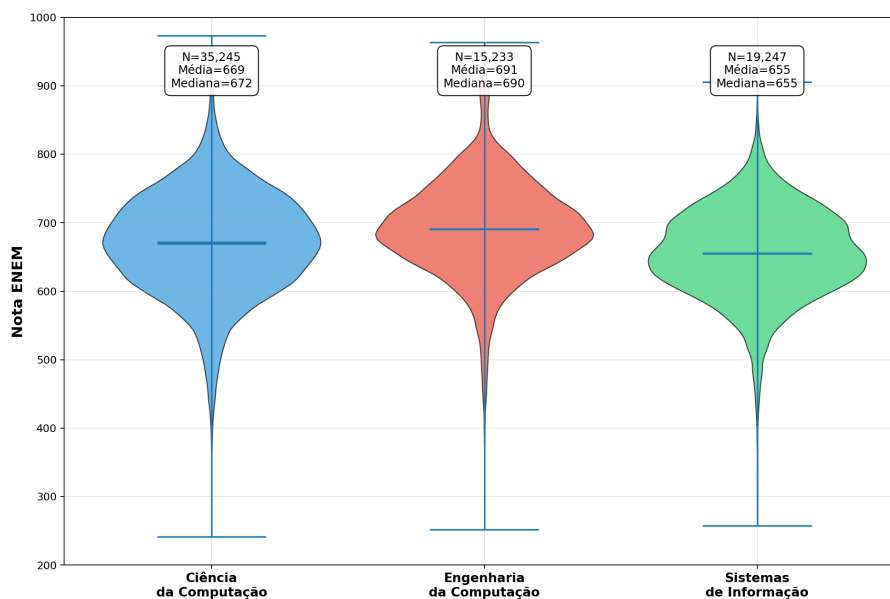


Figura 13. Distribuição do desempenho no ENEM por sexo acumulado(2018-2023)

consistente com revisões sobre paridade e análises institucionais, evidencia que as ações afirmativas atuam como mecanismo compensatório real e que os beneficiários das cotas constituem um grupo diverso, com muitos apresentando desempenho comparável ao de concorrentes da ampla concorrência [Nascimento et al. 2023, Castilho 2024].

A interação entre curso e modalidade reforça a ideia de seletividade diferenciada por formação. Cursos com maiores médias de ENEM, notadamente Engenharia da Computação, concentram maior proporção de aprovados via ampla concorrência, o que amplia a diferença média entre modalidades nesses cursos. Em contrapartida, cursos com médias inferiores, como Sistemas de Informação, apresentam maior dependência relativa das cotas para compor suas turmas, sinalizando que decisões de oferta e perfis de preparação prévia dos candidatos modulam fortemente quem ingressa em cada formação. Essa dinâmica espacial e por curso dialoga com estudos sobre concentração regional e efeitos de atratividade institucional sobre fluxos de alunos [Pelegrini et al. 2018].

Do ponto de vista de política pública e metodologia a conclusão é dupla. Em termos de intervenções, elevar a preparação para o ENEM entre grupos subrepresentados, combinar políticas de acesso com programas de tutoria, bolsas e moradia e implementar ações de acolhimento nos primeiros semestres são medidas com potencial para aumentar inclusão sem perda de mérito, conforme recomendações de avaliações sobre ações afirmativas [Basso-Poletto et al. 2020, Veras and Luz Silva 2020]. Em termos analíticos, a distinção entre médias é insuficiente; é necessário enfatizar análises da forma das distribuições, testes robustos e modelos hierárquicos que decomponham variância por curso e instituição, bem como fluxos e determinantes locais, práticas alinhadas a guias de KDD e metodologias de multiescala [Raudenbush and Bryk 2002, Fayyad et al. 1996, Fachin 2006].

3.4. Q4. Que padrões de mobilidade geográfica se observam entre aprovados, isto é, de que forma a migração entre estados (origem × destino) se relaciona com a representação feminina nos cursos de computação?

A Figura 14 ilustra a rede de fluxos origem destino dos aprovados em cursos de computação no período 2018 a 2023, em que a espessura das faixas representa o volume de estudantes deslocados entre unidades federativas. Observa-se que estados do Sudeste, notadamente São Paulo e Minas Gerais, atuam como grandes polos recetores, recebendo fluxos expressivos de diversas origens, enquanto alguns estados do Norte e do Nordeste aparecem como emissores significativos para polos vizinhos. O diagrama também evidencia caminhos regionais de curta e longa distância, bem como a existência de nós com fluxos bidirecionais que sinalizam circuitos de mobilidade intra regionais. Em termos agregados o Sankey torna visível a concentração espacial da oferta e a assimetria entre zonas emissores e receptores, oferecendo representação direta do papel da atratividade institucional e das fricções geográficas na redistribuição de vagas.

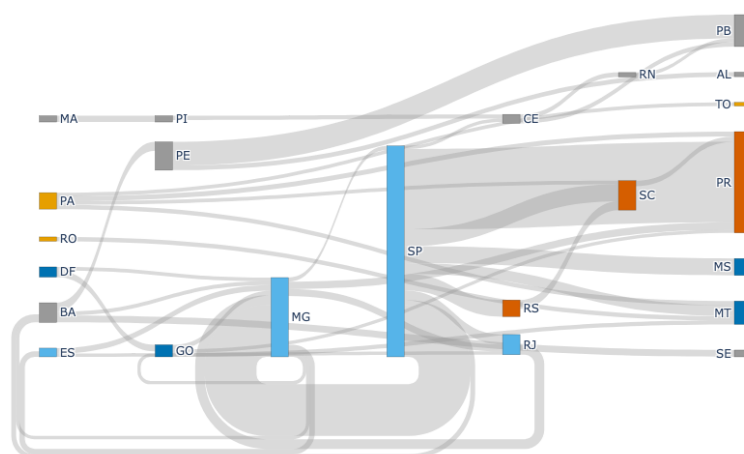


Figura 14. Distribuição do Fluxo migratório dos aprovados em cursos de computação pelo ENEM 2018-2023

A análise quantitativa associada confirma que cerca de um quarto dos aprovados efetua migração interestadual, com taxa média entre 22 por cento e 26 por cento ao longo do período, e que a propensão à mobilidade é apenas ligeiramente superior entre as mulheres em relação aos homens, diferença inferior a dois pontos percentuais. Esses resultados indicam que a migração acadêmica é um fenômeno relativamente difundido entre os gêneros, mas que a simples ampliação dos fluxos não garante alteração substancial nos perfis de gênero das coortes receptoras. A presença de rotas robustas para polos do Sudeste e do Nordeste reforça o papel da distribuição desigual da oferta como determinante central da mobilidade [Pelegrini et al. 2018].

Ao relacionar mobilidade e representatividade feminina, a correlação observada é positiva, porém fraca e estatisticamente não significativa, o que sugere que instituições com maior entrada de estudantes de outras unidades não apresentam automaticamente

maior proporção de mulheres entre aprovados. Em termos práticos a mobilidade amplia o alcance geográfico das candidaturas femininas sem, contudo, ser agente isolado de transformação da composição por gênero. A existência de outliers no Sankey e nos indicadores de proximidade demonstra que contextos institucionais específicos e políticas locais de atração podem combinar alta migração com maior diversidade, apontando para possibilidades de ação institucional orientadas à inclusão [Pelegri et al. 2018].

As implicações políticas derivadas do padrão de fluxo são claras. Para que a mobilidade contribua efetivamente para a diversificação de gênero é necessário articular expansão da oferta com políticas de acolhimento e permanência, tais como programas de moradia estudantil, bolsas de manutenção e ações de mentoria, além de investimentos na preparação pré universitária nas regiões menos favorecidas. Essas medidas transformam a redistribuição espacial em ganho sustentável de diversidade quando combinadas com ações afirmativas e com estratégias institucionais de integração, conforme recomendações de avaliações sobre inclusão e políticas afirmativas no contexto brasileiro [Basso-Poletto et al. 2020, Veras and Luz Silva 2020].

4. Conclusões

A presente investigação sintetiza evidências robustas sobre a persistência da masculinização nos cursos de computação e sobre um movimento gradual de ampliação da participação feminina entre 2018 e 2023. Os resultados mostram que, embora a proporção de candidatas e aprovadas tenha crescido de forma estatisticamente significativa, a mudança é lenta e desigual no espaço e no tempo. Essa constatação confirma diagnósticos anteriores que situam a sub-representação feminina como um fenômeno estrutural resultante de desigualdades educativas acumuladas e de práticas institucionais que reproduzem barreiras de gênero [UNESCO 2017, Moss-Racusin et al. 2012].

As políticas afirmativas emergem nesta análise como instrumentos que alteram concretamente o perfil das coortes, ampliando a presença feminina especialmente nas vagas reservadas pelas cotas, mas sem por si só remover todos os obstáculos à igualdade. A maior probabilidade de ingresso via cotas identificada nos dados sugere efeito compensatório real, coerente com avaliações empíricas sobre ações afirmativas no ensino superior brasileiro [Basso-Poletto et al. 2020, Napolini 2017]. Ao mesmo tempo, a eficácia das cotas depende claramente de medidas complementares de permanência e apoio acadêmico para que o acesso se converta em trajetória formativa bem-sucedida.

A análise espacial evidencia que desigualdades regionais e a distribuição da oferta são determinantes centrais das dinâmicas observadas, tanto na composição por gênero quanto no desempenho médio no ENEM. Estados e polos universitários com maior concentração de cursos atraem fluxos mais intensos e concentram melhores médias, o que reforça padrões já descritos por estudos de mobilidade estudantil e modelos gravitacionais [Pelegri et al. 2018, Cullinan and Duggan 2016]. Nesse contexto, a mobilidade por si só não promove automaticamente maior representatividade feminina, salvo quando acompanhada de políticas locais que reduzam custos de permanência e que tornem as instituições mais acolhedoras.

Do ponto de vista da desigualdade de preparo e desempenho, os achados apontam para uma hierarquia de determinantes em que fatores territoriais e institucionais superam o efeito isolado do gênero. A aplicação de técnicas preditivas e de mode-

los multiescalares revelou que unidade federativa, ano e modalidade explicam parcela maior da variação das notas do que a variável sexo, conclusão que orienta a necessidade de intervenções estruturais na formação básica e em políticas regionais de educação [Breiman 2001, Nascimento et al. 2023]. Portanto, a redução do hiato de desempenho exige programas que atuem na base do sistema educacional e não apenas medidas voltadas ao nível universitário.

As implicações para políticas públicas são diretas e pragmáticas. Recomenda-se combinar políticas de acesso, como cotas, com investimentos em preparação pré ENEM, em programas de tutoria e mentoria, em bolsas de manutenção e em oferta de moradia estudantil, medidas que se mostram complementares e necessárias segundo revisões sobre inclusão em STEM [Veras and Luz Silva 2020, da Rocha et al. 2022]. Além disso, recomenda-se monitoramento contínuo por meio de pipelines reproduzíveis e avaliações longitudinais para aferir impactos de intervenções e calibrar políticas em função dos contextos regionais e institucionais.

A trajetória traçada pelos dados remete a uma reflexão mais ampla sobre mudança social e temporalidade das transformações educativas. A igualdade de gênero nas áreas de computação não é uma meta que se alcança apenas por decretos ou por uma política isolada, mas por um conjunto coerente de ações que articulem acesso, qualidade e permanência ao longo de gerações. Se há espaço para otimismo, ele advém do fato de que as políticas implementadas até aqui demonstram efeitos mensuráveis; contudo, sem articulação e compromisso sustentado, os ganhos permanecerão parciais e lentos, exigindo vontade política e investimento público consistente para que a inclusão se converta em equidade efetiva em STEM [UNESCO 2017, Basso-Poletto et al. 2020].

Referências

- Agresti, A. (2018). *Statistical Methods for the Social Sciences*. Pearson, 5 edition.
- Basso-Poletto, D., Efrom, C., and Beatriz-Rodrigues, M. (2020). Ações afirmativas no ensino superior: revisão quantitativa e qualitativa da literatura. *Revista Electrónica Educare*, 24(1):292–325.
- Brasil (2012). Lei nº 12.711, de 29 de agosto de 2012.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.
- Castilho, M. d. S. M. (2024). Análise de perfil, desempenho e provas do enade do curso de sistemas de informação da ufsc.
- Cullinan, J. and Duggan, J. (2016). A school-level gravity model of student migration flows to higher education institutions. *Spatial Economic Analysis*, 11(3):294–314. Estudo empírico que aplica especificação gravitacional a fluxos de alunos de escolas secundárias para IES (Irlanda).
- da Rocha, I. G., dos Santos, F. F., de Oliveira Barden, C., and Collazzo, G. C. (2022). Minimização da disparidade de gênero na área de ciências exatas, engenharias e computação : o impacto da extensão a universitários no cotidiano de algumas escolas públicas na cidade de santamaria.
- Fachin, O. (2006). *Fundamentos de metodologia*. Saraiva, São Paulo.

- Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. (1996). *Data Mining to Knowledge Discovery: An overview*. AAAI/MIT Pres.
- Herring, C. (2009). Does diversity pay? race, gender, and the business case for diversity. *American Sociological Review*, 74(2):208–224.
- Hunt, V., Layton, D., and Prince, S. (2015). Diversity matters. [Acessado em: 23 de maio de 2024].
- Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica*, 13(3):245–259.
- McKinney, W. (2010). *Data Analysis with Python and Pandas*. O’Reilly Media.
- Ministério da Educação do Brasil (MEC) (2024). Repositório de dados do sistema de seleção unificada (sisu).
- Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., and Handelsman, J. (2012). Science faculty’s subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, 109(41):16474–16479.
- Nascimento, L., Lima, Y., Barbosa, C., Costa, L., Santos, A., Galeno, L., Xexéo, G., and Souza, J. (2023). Paridade de gênero no ensino superior em stem no brasil: uma análise de 10 anos. In *Anais do XVII Women in Information Technology*, pages 217–227. Sociedade Brasileira de Computação.
- Naspolini, T. (2017). Ações afirmativas: uma análise do comportamento acadêmico de alunos ingressantes em cursos da universidade federal de santa catarina.
- pandas development team (2020). pandas: powerful python data analysis toolkit. Software, <https://pandas.pydata.org/>.
- Pearson, K. (1900). X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50(302):157–175.
- Pelegrini, T., Sá, C., and França, M. T. A. (2018). Fatores associados à mobilidade de estudantes universitários no brasil: uma análise por meio de modelo gravitacional. In *Anais do XLVI Encontro Nacional de Economia (ANPEC)*. PUCRS.
- Plotly Technologies Inc. (2023). Plotly for python — interactive data visualization library.
- Pérez, F. and Granger, B. E. (2007). Ipython: A system for interactive scientific computing. *Computing in Science and Engineering*, 9(3):21–29.
- Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage Publications, Thousand Oaks, CA, 2 edition.
- Santos, C. M. (2018). Por que as mulheres “desapareceram” dos cursos de computação?
- Sen, P. K. (1968). Estimates of the regression coefficient based on kendall’s tau. *Journal of the American Statistical Association*, 63(324):1379–1389.
- Sá, C., Florax, R. J. G. M., and Rietveld, P. (2003). Determinants of the regional demand for higher education: A gravity model approach.
- UNESCO (2017). *Cracking the Code: Girls’ and Women’s Education in STEM*. UNESCO Publishing.

- VanderPlas, J. (2016). Python data science handbook. *Machine Learning*.
- Veras, R. M. and Luz Silva, D. d. (2020). A democratização do acesso ao ensino superior no brasil é um instrumento de justiça social? possibilidades e desafios na formação de professores. *Eccos – Revista Científica*, (54).
- World Economic Forum (2020). Global gender gap report 2020. Technical report, World Economic Forum.
- Østergaard, C. R., Timmermans, B., and Kristinsson, K. (2011). Does a different view create something new? the effect of employee diversity on innovation. *Research Policy*, 40(3):500–509.

Anexos