Revista Pleiade – Centro Universitário Descomplica UniAmérica
ISSN 1980-8666 (Impresso) – ISSN 2674-8231 (Eletrônico) – Classificação Capes B3 (2017-2020)

Perfil Socioeconômico e Desempenho no ENADE: Análise Causal Aplicada aos Cursos de Tecnologia da Informação

Socioeconomic Profile and Performance in ENADE: Applied Causal Analysis to Information Technology Courses

Emílio Anastácio de Paula Correa, Ruminik Schmoeller e Isabel Fernandes

1. Acadêmico concluinte do curso de Bacharelado em Engenharia de Software do Centro Universitário Descomplica UniAmérica. 2. Mestre Tecnologias Computacionais para o Agronegócio. Docente do curso de Bacharelado em Engenharia de Software do Centro Universitário Descomplica UniAmérica e orientador do presente trabalho. https://orcid.org/0009-0006-5046-4390 3. Computação. Doutora em Ciências. Enga. da Produção. Professora Projeto Final de Curso e coordenadora do curso de Engenharia de Software, Centro Universitário Descomplica UniAmérica. https://orcid.org/0000-0002-6906-5756

emilioanastacio@gmail.com e *isabel.souza@descomplica.com.br*

Palavras-chave
Desempenho Acadêmico
ENADE
Inferência Causal
Machine Learning
Propensity Score Matching
Regressão Linear

Keywords
Academic Performance
Brazilian National Student Performance Exam
Causal Inference
Machine Learning
Propensity Score Matching
Linear Regression

Resumo:
Este estudo analisa, no nível curso-ano, a relação entre o perfil socioeconômico médio dos estudantes (índice SES_INDEX construído a partir de respostas do Questionário do Estudante) e o desempenho no ENADE (NT_GER_mean) em cursos de TI (2014, 2017, 2021). Como variáveis institucionais, consideram-se o tipo de IES (pública/privada) e um índice de metodologias ativas percebidas. Metodologicamente, implementa-se em Python um pipeline CRISP-DM com regressão ponderada (WLS com pesos pelo tamanho da turma e erros-padrão clusterizados por IES) e uma etapa de inferência causal para o efeito de IES pública: PSM (ATT), IPTW (ATE) e AIPW/DR (ATE). Os resultados indicam associação positiva entre IES pública e maior NT_GER_mean, condicionada ao perfil socioeconômico agregado do curso, e um efeito médio causal positivo de ser curso em IES pública, sob os pressupostos de ignorabilidade. As conclusões são estritamente no nível de curso-ano.

Abstract:
This study analyzes, at the course-year level, the relationship between the average socioeconomic profile of students (SES_INDEX constructed from responses to the Student Questionnaire) and performance on the ENADE exam (NT_GER_mean) in IT courses (2014, 2017, 2021). Institutional variables considered include the type of higher education institution (public/private) and an index of perceived active learning methodologies. Methodologically, a CRISP-DM pipeline with weighted regression (WLS weighted by class size and standard errors clustered by higher education institution) and a causal inference step for the effect of public higher education institutions were implemented in Python: PSM (ATT), IPTW (ATE), and AIPW/DR (ATE). The results indicate a positive association between public higher education institutions and higher NT_GER_mean, conditioned on the aggregate socioeconomic profile of the course, and a positive average causal effect of being a course in a public higher education institution, under the assumptions of ignorability. The conclusions are strictly at the course-year level.

Artigo recebido em: 15.10.2025.
Aprovado para publicação em: 07.11.2025.

INTRODUÇÃO

A educação sempre desempenhou um processo crucial no desenvolvimento da sociedade humana. Desde a Grécia Antiga, onde a sofística trouxe um alto significado à educação, trazendo uma abordagem mais consciente e racional, foi em decorrência das contribuições dos sofistas que hoje temos nossos próprios fundamentos da pedagogia moderna (ARAÚJO, 2013).

O SINAES (Sistema Nacional da Avaliação da Educação Superior) foi instituído pela lei nº 10.861, com o objetivo de "melhorar a qualidade da educação superior, a expansão da sua oferta e o aumento de sua eficiência como instituição e efetividade social e acadêmica" (BRASIL, 2004). Integrante desse sistema, o ENADE (Exame Nacional de Desempenho dos Estudantes) é uma avaliação periódica trienal do desempenho dos estudantes dos cursos de graduação, com a finalidade de mapear a evolução aos conteúdos programáticos previstos da grade curricular dos cursos das IES(Instituições de Ensino Superior). Além disso, busca a reunião de informações pessoais relacionadas às características socioeconômicas dos estudantes (LIMANA; BRITO, 2005). Não se limita a avaliar a performance final dos alunos, mas também a coletar informações socioeconômicas, permitindo que fatores externos à prova sejam analisados e proporcionando uma visão mais ampla sobre a qualidade das IES.

O ENADE, em conjunto com a avaliação institucional e a avaliação de cursos de graduação são os três pilares que garantem e reconhecem a qualidade das IES, sendo uma fonte importante de dados que permite identificar diferenças entre instituições públicas e privadas e as metodologias de ensino que empregam. Nessa discussão podemos observar um objetivo em comum das IES, promover uma educação de qualidade que seja capaz de formar pessoas competentes ao mercado de trabalho.

A literatura recente sobre ENADE em cursos de TI tem priorizado descrições e predição de desempenho (mineração de dados e modelos supervisionados), com menor ênfase na identificação causal dos efeitos institucionais e socioeconômicos sobre os resultados, especialmente quando se usam os microdados do Questionário do Estudante e agregações por curso-ano. Exemplos incluem análises exploratórias e preditivas em Ciência da Computação (VISTA; FIGUEIRÓ; CHICON, 2017; SOUZA et al., 2017), evidências sobre determinantes institucionais como composição do corpo docente (BRITO, 2016) e estudos metodológicos de análise de exames no contexto brasileiro (BARBOSA et al., 2024), enquanto a base teórica de inferência causal reforça a necessidade de estratégias de identificação para separar associação de efeito (PEARL, 2010).Nesse contexto, a questão central desta pesquisa é: o perfil socioeconômico agregado dos cursos, o tipo de IES (pública/privada) e a adoção de metodologias ativas causam variações significativas no desempenho médio dos cursos de TI no ENADE (2014, 2017 e 2021)?

Para responder a essa questão, o estudo parte de duas hipóteses orientadoras: H1: A intervenção socioeconômica — representada pelas características socioeconômicas dos estudantes (renda familiar, escolaridade dos pais, situação de trabalho) — afeta significativamente o desempenho no ENADE, sendo que o tipo de IES (pública ou privada) atua como um tratamento moderador, com expectativa de desempenho superior para cursos de IES públicas, especialmente entre aqueles com perfis socioeconômicos agregados mais baixos. H2: A intervenção socioeconômica dos alunos afeta diretamente o desempenho no ENADE e a adoção de metodologias ativas de ensino nas IES funciona como um tratamento moderador, potencializando o desempenho de cursos com perfil socioeconômico agregados mais desfavorecidos.

Em coerência com essas hipóteses, o objetivo deste estudo é analisar o desempenho dos cursos de Ciência da Computação e Análise e Desenvolvimento de Sistemas no ENADE (ENADE, 2022; ITCG, 2010), nos

anos de 2014, 2017 e 2021. A partir dos microdados do ENADE, busca-se identificar como o tipo de IES — pública ou privada — a metodologia de ensino adotada — ativa ou tradicional — e as características socioeconômicas dos estudantes influenciam os resultados avaliativos, utilizando análise estatística com ênfase em inferência causal.

Esta pesquisa se justifica pela escassez de análises atuais e aprofundadas sobre os cursos de Tecnologia da Informação (TI) no ENADE, apesar de o exame ser uma das principais ferramentas de avaliação da educação superior no Brasil. Em um contexto de demanda crescente por profissionais qualificados, é crucial que as IES alinhem formação, qualidade pedagógica e resultados acadêmicos. Persiste uma lacuna específica: faltam estudos que integrem, de forma sistemática, o tipo de IES (pública ou privada), a metodologia de ensino (tradicional ou ativa) e os dados socioeconômicos dos estudantes na explicação do desempenho no ENADE, especialmente em TI. O cenário institucional reforça a relevância do tema: o Censo da Educação Básica de 2023 (INEP/MEC) registra 2.580 IES, sendo 87,8% privadas e 12,2% públicas, o que sugere efeitos diferenciados sobre o desempenho discente. A literatura indica expansão do uso de metodologias ativas e pertinência para áreas aplicadas como TI, nas quais a autonomia discente e a aprendizagem prática são centrais (ROCHA; LEMOS, 2014; SILVA, 2018). Ao empregar análise de dados e inferência causal, este estudo pretende identificar relações e isolar efeitos entre IES, metodologia, perfil socioeconômico e desempenho, controlando fatores confundidores. Com isso, busca subsidiar decisões pedagógicas e institucionais e orientar políticas educacionais mais assertivas e inclusivas. Além de contribuir academicamente ao preencher a lacuna em TI/ENADE, o trabalho oferece aplicação prática, fornecendo evidências para melhoria da qualidade do ensino superior; sua execução ancora-se na minha formação em ADS e experiência com análise de dados, favorecendo uma abordagem crítica e tecnicamente robusta.

Estado da Arte / Trabalhos Relacionados

Como base teórica, Pearl (2010) consolida os Modelos Causais Estruturais e o operador do(·), formalizando condições de identificação (p. ex., critério do *back-door*) que permitem estimar efeitos médios de tratamento em dados observacionais quando se controla adequadamente por confundidores. Para estudos com microdados do ENADE, essa moldura é crucial: explicita a necessidade de declarar o tratamento (p. ex., tipo de IES; intensidade de metodologias ativas), mapear os nós de confusão (renda familiar, escolaridade parental, trabalho do estudante) e testar a robustez das estimativas. Em síntese, fornece o alicerce para mover-se de associação para causalidade ao analisar desempenho discente.

Em estudo sobre cursos de Ciência da Computação no RS, Vista, Figueiró e Chicon (2017) demonstram a operacionalização dos microdados do ENADE no R, articulando etapas de limpeza, seleção de atributos, mineração e visualização para avaliar padrões de desempenho discente. O mérito é metodológico-aplicado: mostra que os microdados são tratáveis e reprodutíveis em pipelines bem documentados. Como limitação, a análise é predominantemente associativa; não há esquema explícito de identificação causal, o que reforça a necessidade de arcabouço como o de Pearl para separar correlação de efeito.

Souza et al. (2017), em perspectiva preditiva, integra o perfil socioeconômico e trajetória no ENEM para antecipar o desempenho no ENADE, utilizando algoritmos de aprendizado supervisionado e validação empírica para aferir acurácia. Os achados indicam alto poder explicativo das variáveis socioeconômicas e antecedentes acadêmicos, sinalizando que tais fatores devem compor o núcleo de covariáveis em estudos sobre desempenho. A contribuição é dupla: (i) evidencia quais atributos carregam informação substantiva; (ii) mos-

tra a utilidade de modelagem supervisionada para diagnóstico institucional. Porém, assim como em Vista et al., trata-se de um exercício de previsão, não de efeito causal.

Barbosa et al. (2024), sistematizam boas práticas de análise de dados em exames de graduação brasileiros: Análise Exploratória de Dados (Exploratory Data Analysis) rigorosa, critérios de seleção de variáveis, visualização informativa e reprodutibilidade. Embora não seja um estudo de causalidade, o artigo oferece um framework de governança analítica que melhora a qualidade das inferências em bases como o ENADE. Sendo assim, reforça-se a importância de documentar decisões analíticas, verificar pressupostos (heterocedasticidade, ponderação, codificação de dummies) e assegurar transparência na construção dos modelos.

Liu et al. (2025), em levantamento recente, mapearam a colaboração entre aprendizado de máquina e inferência causal, cobrindo técnicas para descoberta causal, estimação com heterogeneidade de efeito e suporte de modelos de linguagem à documentação, auditoria e robustez. Para avaliações educacionais, a mensagem é pragmática: combinar modelos estatísticos clássicos (regressão com dummies/interações, estratificação) com ferramentas contemporâneas pode ampliar a capacidade de detectar efeitos condicionais (p. ex., impacto diferencial de metodologias ativas por estratos de renda) e melhorar a reprodutibilidade.

Os estudos aplicados em TI/ENADE (2017) comprovam a viabilidade técnica e destacam o peso socioeconômico, mas permanecem no plano associativo/preditivo. As contribuições recentes Barbosa et al (2024) e Liu et al (2025) consolidam boas práticas analíticas e caminhos para inferência com heterogeneidade, enquanto Pearl (2010) oferece o critério formal para identificação sob confusão. Falta, porém, um estudo que trate o perfil socioeconômico como intervenção central, considerando tipo de IES e metodologia como moderadores, reportando efeitos médios e condicionais com checagens de robustez em microdados do ENADE — exatamente o vazio que este trabalho busca preencher.

MATERIAIS E MÉTODOS

A metodologia desta pesquisa foi desenhada para, primeiramente, estruturar e modelar as associações entre os múltiplos fatores que influenciam o desempenho no ENADE e, em seguida, aplicar técnicas de inferência causal para isolar o efeito específico do tipo de instituição (pública vs. privada). O processo foi dividido em duas abordagens: um pipeline baseado no modelo CRISP-DM (*Cross-Industry Standard Process for Data Mining*) (Wirth e Hipp, 2000), para a análise de regressão ajustada e uma análise de Propensity Score para estimação causal.

1 ABORDAGEM 1: PIPELINE CRISP-DM E MODELO DE REGRESSÃO PONDERADA

Para a exploração inicial e modelagem das associações, foram adotados ciclos iterativos, segundo as fases do CRISP-DM, para (1) entendimento do negócio; (2) entendimento dos dados; (3) preparação dos dados; (4) modelagem; e (5) avaliação.

1.1 ENTENDIMENTO DO NEGÓCIO (BUSINESS UNDERSTANDING)

Esta fase inicial definiu o escopo da pesquisa, centrada na questão de como fatores socioeconômicos, o tipo de IES e as metodologias de ensino percebidas impactam o desempenho no ENADE para cursos de TI. A fonte de dados foram os microdados oficiais do ENADE, disponibilizados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), abrangendo os anos de 2014, 2017 e 2021 (INEP, ENADE Microdados, 2025).

1.2 Entendimento dos Dados (Data Understanding)

Nesta etapa a unidade de análise fundamental foi definida como o curso (CO_CURSO), exigindo que os dados, originalmente no nível de aluno, fossem agregados. Os dados brutos foram extraídos dos diversos arquivos (arq1, arq3, arq4, etc.) de cada ano, que detalham a estrutura dos cursos, as notas e as respostas ao Questionário do Estudante (QE).

Tabela 1: Principais Variáveis Brutas Extraídas dos Microdados do INEP

Arquivo(s) de Origem	Nomes das Variáveis	Descrição (Conforme Dicionário INEP)
arq1	CO_CURSO, CO_IES	Códigos de identificação do Curso e da IES.
arq1	CO_CATEGAD	Código da Categoria Administrativa (Pública/Privada).
arq1	CO_GRUPO	Código da Área do curso (ex: 4004 = C. da Computação).
arq1	CO_UF_CURSO	Código da UF (Estado) de localização do curso.
arq3	NT_GER	Nota Geral do aluno (Formação Geral + Específica).
arq14	QE_I08	Renda total da família (em faixas A, B, C...).
arq10/11	QE_I04, QE_I05	Escolaridade do Pai e da Mãe (em faixas A, B, C...).
arq16	QE_I10	Situação de trabalho/emprego do aluno.
arq4	QE_I29, QE_I32, etc.	Respostas (Likert) às questões de percepção sobre o curso.

A Tabela 1 descreve as variáveis-chave extraídas e os arquivos de origem, que serviram de base para a etapa de preparação.

1.3 Preparação dos Dados (Data Preparation)

Esta foi a etapa mais intensiva do pipeline, focada na transformação dos dados brutos em um conjunto de dados analítico e agregado por curso. Primeiramente, a base de dados de cada ano foi filtrada para incluir apenas os cursos de Tecnologia da Informação, identificados pelos códigos CO_GRUPO (72, 4004, 4005, 4006). Em seguida, procedeu-se à engenharia de variáveis, onde os dados dos alunos foram agregados por CO_CURSO:

Variável Dependente (Y): NT_GER_mean, a média da Nota Geral dos concluintes do curso.

Variável de Tratamento (T): IES_PUBLICA, uma variável binária (1=Pública, 0=Privada) criada a partir do CO_CATEGAD.

Covariável Socioeconômica (SES): O SES_INDEX, um índice contínuo criado via Análise de Componentes Principais (PCA) sobre as médias das respostas às questões do QE sobre renda, escolaridade dos pais e situação de trabalho (QE_I08, QE_I05, QE_I04, QE_I10). O PCA sumariza a variância comum dessas variáveis em um único fator. Na análise, o primeiro componente principal (SES_INDEX) se mostrou robusto,

Pleiade, 19(49): 05-18, Out.-Dez., 2025
DOI: 10.32915/pleiade.v19i49.1192

explicando, em média, 54.2% da variabilidade combinada dessas quatro questões (54.1% em 2014, 54.6% em 2017 e 53.8% em 2021).

Covariável (SES Binária): LOW_SES_curso, uma variável binária criada a partir do SES_INDEX. Ela recebe o valor 1 se o SES_INDEX do curso está abaixo do percentil 40 (P40) da amostra total, e 0 se está no P40 ou acima. A escolha do P40 é uma decisão metodológica para definir um grupo de vulnerabilidade que seja suficientemente amplo para análise, alinhando-se a práticas comuns em ciências sociais que definem os dois quintis inferiores (40%) da distribuição como o grupo de "baixo perfil" para análises de desigualdade.

Covariável de Metodologia: O METODO_ATIVA_SCORE, um índice da percepção discente sobre metodologias ativas, foi calculado pela média dos Z-scores das médias das respostas às questões QE_I29, QE_I32, QE_I41, QE_I42, QE_I48 e QE_I49. O uso de Z-scores (padronização pela média e desvio padrão) é uma técnica padrão para criar índices compostos, garantindo que cada questão contribua com peso igual para o score final, evitando que uma única questão com alta variância domine o índice.

Controles e Pesos: Foram incluídas variáveis categóricas de controle para CO_GRUPO, CO_UF_CURSO e ano. A contagem de alunos (n_alunos) foi mantida para ser usada como peso na modelagem.

Limpeza e Consolidação: Cursos com dados ausentes em colunas essenciais para a modelagem (como NT_GER_mean, CO_IES) ou que não registraram que oialunos válidos (n_alunos=0) foram removidos. Os DataFrames anuais limpos foram então concatenados na nova base de dados, totalizando 3.618 observações (cursos-ano) válidas.

1.4 Modelagem (Modelo WLS)

Para modelar as associações ajustadas, foi empregada uma Regressão por Mínimos Quadrados Ponderados (WLS) (WEISBERG, 2005). A escolha pelo WLS se justifica pela necessidade de corrigir a heterocedasticidade identificada nos dados; a média da nota de um curso com muitos alunos (n_alunos alto) é estatisticamente mais precisa (possui menor variância) do que a de um curso com poucos alunos. O modelo utiliza o n_alunos como peso, atribuindo maior importância às observações mais precisas (GREENE, 2012).

Para corrigir a não independência das observações (cursos diferentes podem pertencer à mesma IES), o modelo utilizou erros padrão robustos clusterizados pelo código da IES (CO_IES). A fórmula completa do modelo incluiu as variáveis principais, seus termos de interação e os controles geográficos, de área e de ano.

1.5 Avaliação (Evaluation)

A adequação do modelo foi avaliada através do R^2 *Ajustado* (poder explicativo) e da significância estatística (p-valores) dos coeficientes. Testes diagnósticos de heterocedasticidade (Breusch-Pagan e White) também foram executados para validar a necessidade do uso de WLS.

2 Abordagem 2: Metodologia de Inferência Causal (PSM, IPTW, DR)

Embora o modelo WLS estime associações robustas, ele não pode, por si só, garantir a causalidade. O principal desafio em dados observacionais como o ENADE é o viés de seleção (confundimento): fatores não medidos, como a habilidade acadêmica prévia do aluno (ex: nota do ENEM), influenciam tanto a escolha da IES (Pública ou Privada) quanto o resultado no ENADE.

Para tentar isolar o efeito causal do tipo de instituição (IES_PUBLICA, T) sobre a nota (NT_GER_mean, Y), foram aplicados métodos de Propensity Score (PS). O objetivo é criar uma "pseudo-população" estatisticamente balanceada onde os grupos de tratamento (Pública) e controle (Privada) sejam comparáveis em

Pleiade, *19*(49): 05-18, Out.-Dez., 2025
DOI: 10.32915/pleiade.v19i49.1192

relação a todo o vetor de covariáveis observadas (X). O vetor X incluiu o SES_INDEX, METODO_ATIVA_SCORE, n_alunos e os controles categóricos para CO_GRUPO, CO_UF_CURSO e ano.

2.1 Etapa 1: Estimativa do Propensity Score (PS)

O Propensity Score (PSi) é a probabilidade de um curso i receber o tratamento (ser público, Ti=1), dadas suas covariáveis observadas $Xi: PSi = P(Ti = 1 / Xi)$. Este score foi estimado para todas as 3.618 observações pooled usando uma Regressão Logística. A validade da análise (suposição de suporte comum) foi confirmada visualmente através da sobreposição das distribuições de PS entre os grupos (ROSENBAUM; RUBIN, 1983).

2.2 Etapa 2: Métodos de Estimação Causal

Com base nos scores calculados, três estimadores diferentes foram aplicados:

Propensity Score Matching (PSM): O PSM foi usado para estimar o Efeito Médio do Tratamento nos Tratados (ATT), focando nos cursos que efetivamente são públicos. Foi implementado um pareamento 1:1 por vizinho mais próximo (Nearest Neighbor) com *caliper* de 0.2 desvios padrão do logit do PS. Cursos públicos sem um "gêmeo" privado similar foram descartados. O sucesso do pareamento foi avaliado via Diferença Média Padronizada (SMD), onde um |SMD|<0.1 indica bom balanceamento. O ATT foi então calculado como a diferença simples das médias de NT_GER_mean entre os grupos pareados. (ROSENBAUM; RUBIN, 1983).

Inverse Probability of Treatment Weighting (IPTW): Para estimar o Efeito Médio do Tratamento (ATE) em toda a população, foi utilizado o IPTW. Este método utiliza todos os cursos, mas pondera cada um pelo inverso de sua probabilidade de tratamento, $Wi = \frac{PSi}{Ti} + \frac{i\,1-Ti}{1-PS}$(HERNÁN; ROBINS, 2020). Para garantir estabilidade, os pesos foram truncados no 99º percentil. O ATE foi estimado usando uma regressão WLS simples (Y∼T), ponderada por Wi.

Regressão Duplamente Robusta (AIPW): Como estimativa principal, foi aplicado o método *Augmented Inverse Probability Weighting* (AIPW) (BANG; ROBINS, 2005), que é Duplamente Robusto. Este método combina o modelo de propensity score (usado no IPTW) com dois modelos de resultado (regressões de **Y∼X**, um para T=1 e outro para T=0). A estimativa do ATE é consistente se *pelo menos um* dos modelos (propensão ou resultado) estiver correto. O ATE é calculado pela diferença das médias dos resultados potenciais estimados (Y^(1) e Y^(0)), conforme as fórmulas de AIPW.

Resultados e Discussão

Esta seção apresenta os resultados obtidos através da execução dos pipelines metodológicos descritos. Primeiramente, são detalhadas as estatísticas descritivas da amostra final. Em seguida, são apresentados e discutidos os achados do modelo de regressão ponderada (WLS), com foco nas interações entre as variáveis. Por fim, são expostos os resultados da análise de inferência causal, que busca isolar o efeito do tipo de IES no desempenho.

1 Análise Descritiva dos Dados

Após a execução do pipeline CRISP-DM (descrito na Metodologia, seção 3.1), os dados dos anos 2014, 2017 e 2021 foram limpos, processados e agregados por curso. A base de dados final consolidada é composta por 3.618 observações (cursos-ano).

Do total da amostra, 32% dos cursos-ano são de IES Públicas (IES_PUBLICA mean=0.32) e 41.2% foram classificados como de perfil socioeconômico baixo (LOW_SES_curso mean=0.412). Os índices SES_INDEX e METODO_ATIVA_SCORE foram padronizados durante sua criação, apresentando médias próximas de zero, como esperado.

A Tabela 2 apresenta as estatísticas descritivas das variáveis-chave utilizadas neste estudo. A nota geral média (NT_GER_mean) dos cursos de TI na amostra foi de 39.71, com um desvio padrão (DP) de **8.17**. Este valor de desvio padrão será usado como "régua" para interpretar a magnitude dos efeitos encontrados.

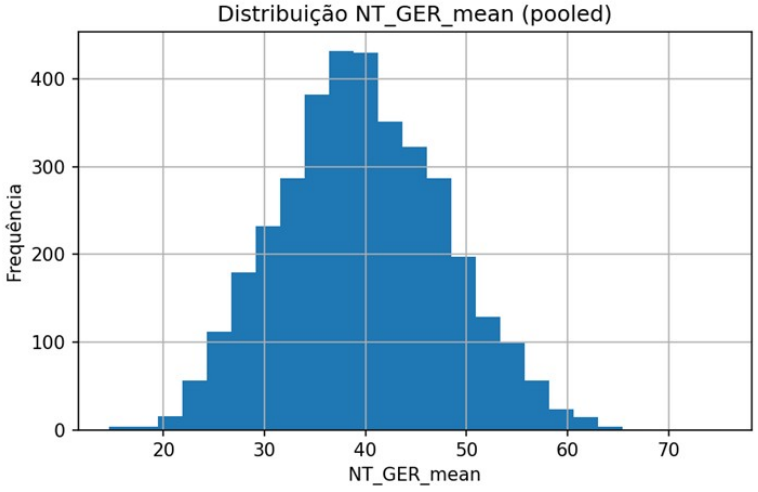
Tabela 2: Estatísticas Descritivas da Amostra Pooled (N=3618)

Variável	Média	Desvio Padrão (DP)	Mínimo	Mediana (50%)	Máximo
NT_GER_mean (Nota)	39.71	8.17	14.60	39.36	75.15
IES_PUBLICA (1=Pública)	0.320	0.467	0.00	0.00	1.00
LOW_SES_curso (1=Baixo SES)	0.412	0.492	0.00	0.00	1.00
SES_INDEX	0.006	1.461	-6.06	-0.03	5.73
METODO_ATIVA_SCORE	-0.003	0.886	-4.73	-0.01	1.88
n_alunos (Peso)	36.89	65.98	2.00	2.00	1721.00

Fonte: Elaborado pelo autor com base no arquivo anexos_enade.xlsx.

A distribuição das notas (Gráfico 1 e Gráfico 2), revela diferença nas medianas de desempenho entre IES Públicas e Privadas antes de qualquer ajuste.

Gráfico 1: Histograma de Distribuição NT_GER médio Pooled



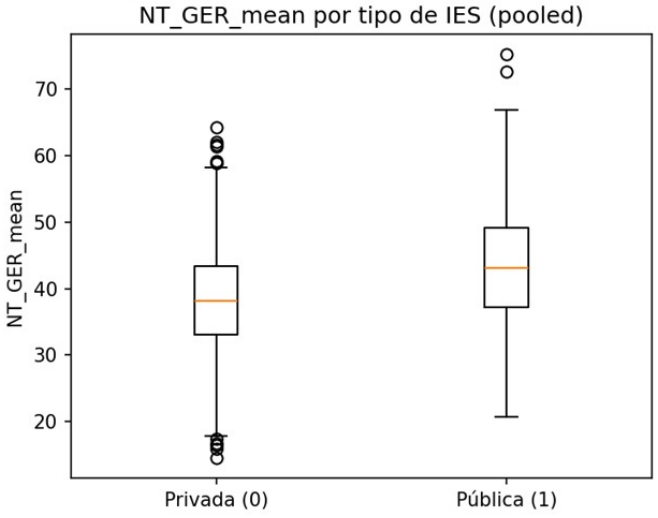
Fonte: Elaborado pelo autor

A análise descritiva visual corrobora os dados da Tabela 1. O Gráfico 1 demonstra que as notas médias dos cursos (NT_GER_mean) seguem uma distribuição aproximadamente normal, com a maioria dos cursos concentrada na faixa central de desempenho. O Gráfico 2 oferece a primeira evidência visual de uma das

principais questões desta pesquisa: há uma diferença clara nas distribuições de notas, onde os cursos de IES Públicas apresentam uma mediana e quartis visivelmente superiores aos dos cursos de IES privadas.

Contudo, esta diferença observada nas médias (a "diferença bruta") não considera o impacto de outros fatores, como o perfil socioeconômico dos curso. Para analisar essas associações de forma controlada e testar as interações propostas nas hipóteses, a próxima seção apresenta os resultados do modelo de regressão ponderada (WLS).

Gráfico 2: Boxplot de NT_GER médio por tipo de IES pooled



Fonte: Elaborado pelo autor

2 Resultados da Análise de Regressão (WLS)

A primeira abordagem de modelagem consistiu em uma Regressão Ponderada (WLS), que ajusta as associações observadas controlando simultaneamente por todas as covariáveis. O modelo pooled (N=3618) apresentou um alto poder explicativo (R^2 Ajustado = 0.672), indicando que 67,2% da variação nas notas dos cursos de TI pode ser explicada pelas variáveis do modelo.

Os testes de heterocedasticidade (Breusch-Pagan e White) apresentaram p-valores extremamente baixos ($p < 0.001$), confirmando que a variância dos erros não era constante e validando a escolha do WLS (ponderado por n_alunos) sobre um OLS simples. A Tabela 2 sumariza os coeficientes das variáveis de interesse.

Tabela 2: Resultados do Modelo WLS Pooled (N=3618)

Variável	Coeficiente	Erro Padrão (Cluster)	p-valor	Magnitude (em DP da Nota)
Intercepto	36.6385	1.345	0.000	-
LOW_SES_curso (1=Baixo SES)	-3.4316	0.349	0.000	-0.42 DP
IES_PUBLICA (1=Pública)	+8.4487	0.497	0.000	+1.03 DP
METODO_ATIVA_SCORE	+1.6709	0.233	0.000	+0.20 DP
LOW_SES_curso : IES_PUBLICA	-4.5924	0.689	0.000	+0.20 DP
LOW_SES_curso : METODO...	-0.3810	0.292	0.192	+0.20 DP
C(ano)[T.2021] (vs 2014)	-8.5501	0.296	0.000	-1.05 DP

Fonte: Elaborado pelo autor com base no arquivo anexos_enade.xlsx. Apenas coeficientes principais e interações de interesse são mostrados. Controles de CO_GRUPO e CO_UF_CURSO incluídos no modelo, mas omitidos por brevidade.

2.1 Interpretação dos Coeficientes WLS

O resultado mais importante da Tabela 2 é o termo de interação LOW_SES_curso:IES_PUBLICA, que é grande e estatisticamente significativo ($p < 0.001$). Isso significa que o "efeito" de ser uma IES pública não é constante, mas depende do perfil socioeconômico do curso.

Efeito Condicional da IES (Prêmio da Pública):

Para cursos de SES Não-Baixo (LOW_SES=0): A vantagem associada a ser uma IES Pública é o seu coeficiente principal: +8.45 pontos (ou +1.03 DP).

Para cursos de Baixo SES (LOW_SES=1): A vantagem é a soma do efeito principal e da interação: 8.45 + (-4.59) = +3.86 pontos (ou +0.47 DP).

Discussão: A associação positiva da IES pública existe para ambos os grupos, mas ela é significativamente menor (menos da metade) para cursos com perfil socioeconômico mais baixo.

Efeito Condicional do SES (Penalidade do Baixo SES):

Em IES Privadas (IES_PUBLICA=0): A penalidade associada ao baixo SES é o coeficiente principal: -3.43 pontos (ou -0.42 DP).

Em IES Públicas (IES_PUBLICA=1): A penalidade é a soma do efeito principal e da interação: -3.43 + (-4.59) = -8.02 pontos (ou -0.98 DP).

Discussão: Este é um achado crucial. A lacuna de desempenho associada ao perfil socioeconômico é mais que o dobro dentro das IES públicas em comparação com as privadas.

Outros Efeitos:

METODO_ATIVA_SCORE: Mostra uma associação positiva pequena, mas significativa, com a nota (+1.67 pontos por ponto no score, ou +0.18 DP).

LOW_SES_curso: METODO_ATIVA_SCORE: A interação não foi significativa (p=0.192), sugerindo que o benefício percebido das metodologias ativas é o mesmo para ambos os grupos de SES.

C(ano)[T.2021]: O ano de 2021 está associado a uma queda massiva de -8.55 pontos (mais de 1 DP) em relação a 2014, controlando por todos os outros fatores, o que pode refletir os impactos da pandemia da COVID-19.

3 Resultados da Análise de Inferência Causal

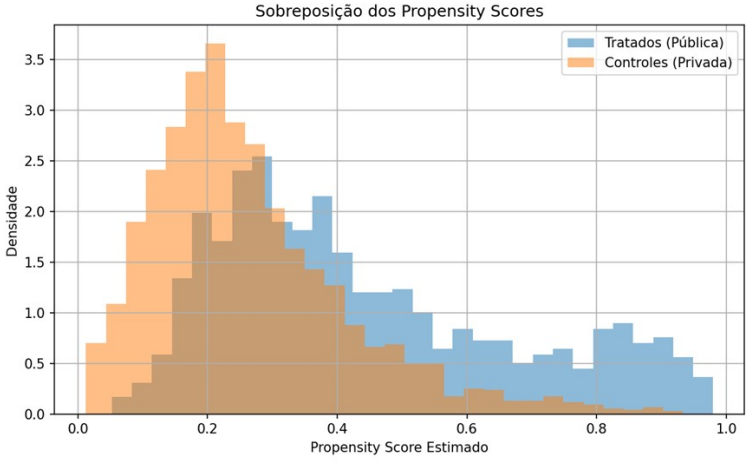
O modelo WLS é poderoso para mostrar associações ajustadas, mas o viés de seleção ainda é uma preocupação. Para tentar isolar o **efeito causal** médio de IES_PUBLICA, foram aplicados métodos de Propensity Score (PSM, IPTW e DR).

3.1 Balanceamento de Covariáveis (PSM)

A primeira etapa foi estimar a probabilidade (Propensity Score) de cada curso ser público, com base em *todas* as covariáveis (SES, metodologia, n_alunos, área, UF e ano). O gráfico de sobreposição (Gráfico 3) mostrou bom suporte comum entre os grupos.

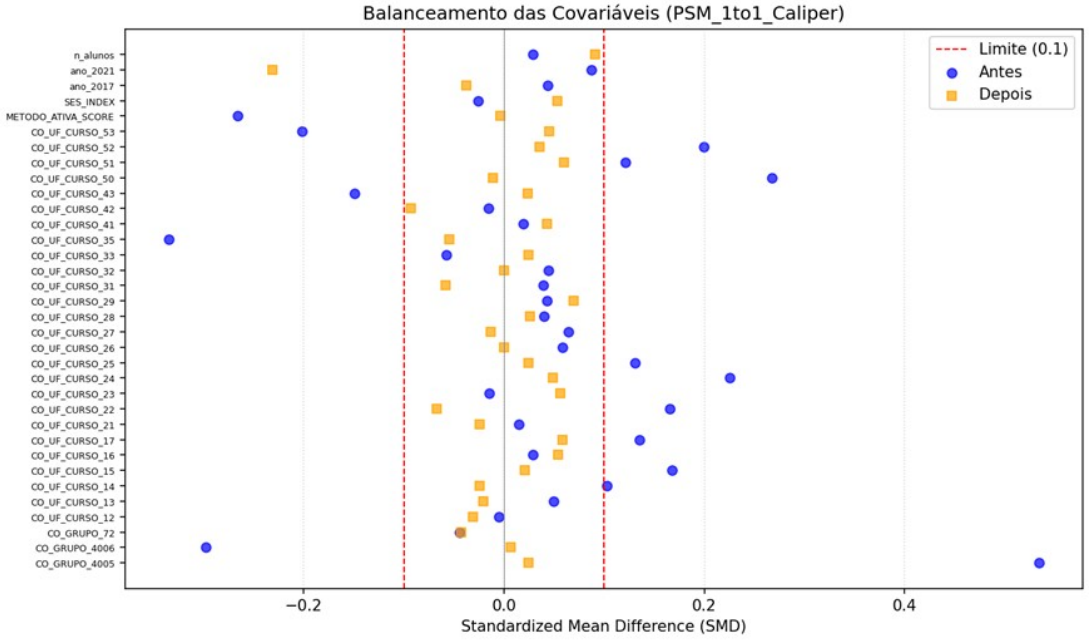
Em seguida, foi realizado o pareamento 1:1 com caliper, que selecionou 681 cursos públicos e 681 cursos privados "gêmeos". O sucesso deste pareamento é demonstrado no Gráfico 4, que mostra o balanceamento das covariáveis.

Gráfico 3: Sobreposição dos Propensity Scores



Fonte: Elaborado pelo autor

Gráfico 4: Gráfico de Balanceamento (Love Plot) das Covariáveis (PSM)



Fonte: Elaborado pelo autor

Como visto no Gráfico 3 e no Gráfico 4 de Balanceamento, antes do pareamento (pontos azuis), os grupos eram altamente desbalanceados (15 covariáveis com |SMD| > 0.1). Após o pareamento (pontos laranjas), quase todas as covariáveis ficaram perfeitamente balanceadas, com exceção de ano_2021, que permaneceu ligeiramente desbalanceada (SMD = -0.23). Isso valida a criação de um grupo de comparação robusto.

3.2 Estimativas do Efeito Causal

Com os grupos balanceados, foi estimado o efeito causal de IES_PUBLICA (Pública=1 vs. Privada=0) na NT_GER_mean. A Tabela 3 compara os resultados dos diferentes métodos.

Tabela 3: Estimativas do Efeito de IES Pública sobre a Nota Geral Média (NT_GER_mean)

Método de Análise	Estimativa (Pontos)	Magnitude (em DP da Nota)	Tipo de Efeito Estimado
1. Diferença Bruta (Viesado)	+4.94	+0.60 DP	Associação Simples
2. PSM (Matching 1:1)	+7.38	+0.90 DP	ATT (Efeito nos Tratados)
3. IPTW (Ponderação)	+5.53	+0.68 DP	ATE (Efeito na População)
4. DR (AIPW)	**+5.86**	**+0.72 DP**	**ATE (Duplamente Robusto)**

Fonte: Elaborado pelo autor com base no arquivo causal_effects_summary.csv. DP da Nota = 8.17.

4 Discussão e Síntese dos Resultados

Os resultados das duas abordagens metodológicas são consistentes e se complementam:

Confundimento Comprovado: A estimativa causal mais confiável (ATE via DR = +5.86 pontos) é substancialmente maior que a diferença bruta de médias (+4.94). Isso sugere que as covariáveis (como perfil SES, área, etc.) atuavam como um **confundidor negativo líquido**: os grupos "Público" e "Privado" não eram comparáveis, e o simples ajuste pela média escondia parte da vantagem das IES públicas.

Magnitude do Efeito: O efeito causal de ser uma IES Pública nos cursos de TI é grande, positivo e estatisticamente significativo, estimado em cerca de +5.86 pontos (ou +0.72 Desvios Padrão) na nota geral. Isso é corroborado pelo ATT do PSM (+7.38), que foca no efeito apenas para os cursos que já são públicos e possuem análogos privados.

Reconciliação dos Modelos (A História Completa):

a. A análise de Regressão WLS (Tabela 2) nos mostrou *por que* as estimativas de efeito médio (ATE/ATT) são como são. O modelo WLS revelou que a vantagem das IES Públicas não é uniforme; ela é muito maior para cursos de perfil SES mais alto (+8.45) e menor para cursos de perfil SES mais baixo (+3.86).

b. As estimativas de ATE (como +5.86 do DR) representam uma média ponderada desses dois efeitos, fazendo todo o sentido estarem posicionadas "entre" os dois extremos (+3.86 e +8.45) encontrados na análise de interação.

Considerações Finais

Os dados indicam que, mesmo após um rigoroso controle estatístico e causal para perfil socioeconômico, metodologia percebida, área e ano, os cursos de TI em IES Públicas apresentam um desempenho significativamente superior no ENADE, com um efeito médio estimado em +5.86 pontos (0.72 DP) na nota geral.

Contudo, esta vantagem é moderada pela desigualdade: a análise de interação (WLS) sugere que a lacuna de desempenho entre cursos de baixo e alto perfil SES é muito mais pronunciada dentro das IES Públicas (-8.02 pontos de penalidade SES) do que nas IES Privadas (-3.43 pontos).

É fundamental reiterar que esta análise, embora robusta, não pôde ser controlada pelo principal confundidor: a habilidade acadêmica prévia do aluno (ex: nota do ENEM de ingresso). É altamente provável que alunos com maior habilidade prévia se auto-selecionem para IES Públicas. Portanto, o efeito causal estimado (ex: +5.86) deve ser interpretado como um limite superior do verdadeiro efeito da instituição, pois ele provavelmente captura uma mistura do "efeito-IES" real e do "efeito-aluno" não medido.

REFERÊNCIAS

ARAÚJO, David Velanes de. AS CONTRIBUIÇÕES DOS SOFISTAS PARA O FENÔMENO DA EDUCAÇÃO NUMA PERSPECTIVA CONTEMPORÂNEA. Cadernos do PET Filosofia, [S. l.], v. 4, n. 7, p. 53–64, 2013. DOI: 10.26694/pet.v4i7.2088. Disponível em: https://periodicos.ufpi.br/index.php/pet/article/view/2088. Acesso em: 12 abr. 2025.

BANG, H.; ROBINS, J. M. Doubly robust estimation in missing data and causal inference models. Biometrics, [s.l.], v. 61, n. 4, p. 962–973, 2005. DOI: 10.1111/j.1541-0420.2005.00377.x. Disponível em: https://doi.org/10.1111/j.1541-0420.2005.00377.x. Acesso em: 20 out. 2025.

BARBOSA, P. L. S.; DAMAZIO, G. N. D. O.; CARVALHO, W. V. de; CARMO, R. A. F. do; OLIVEIRA, E. N. de. Explorando técnicas de análise de dados em exames de avaliação de estudantes de graduação no contexto brasileiro. Avaliação: Revista da Avaliação da Educação Superior, Campinas, v. 29, p. e024030, 2024. Disponível em: https://doi.org/10.1590/1982-57652024v29id279513. Acesso em: 20 out. 2025.

BRITO, Tainá Fernandes de. Corpo Docente: Fatores determinantes do desempenho discente no ENADE. 2015. 98 f. Dissertação (Mestrado em Administração) – Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo, 2015. Disponível em: https://teses.usp.br/teses/disponiveis/12/12139/tde-21032016-115045/pt-br.php. Acesso em: 20 out. 2025.

GREENE, W. H. Econometric Analysis. 7. ed. New York: Pearson Education, 2012. Disponível em: https://www.ctanujit.org/uploads/2/5/3/9/25393293/_econometric_analysis_by_greence.pdf. Acesso em: 20 out. 2025.

HERNÁN, M. A.; ROBINS, J. M. Causal Inference: What If. Boca Raton: Chapman & Hall/CRC, 2020. Disponível em:https://static1.squarespace.com/static/675db8b0dd37046447128f5f/t/68e5466ee0e24211c8a38383/1759856238731/hernanrobins_WhatIf_7oct25.pdf. Acesso em: 20 out. 2025.

LIMANA, Amir; BRITO, Márcia Regina F. de. O MODELO DE AVALIAÇÃO DINÂMICA E O DESENVOLVIMENTO DE COMPETÊNCIAS: ALGUMAS CONSIDERAÇÕES A RESPEITO DO ENADE. Avaliação: Revista da Avaliação da Educação Superior, Campinas; Sorocaba, SP, v. 10, n. 2, 2005. Disponível em: https://periodicos.uniso.br/avaliacao/article/view/1303. Acesso em: 12 abr. 2025.

LIU, Xiaoyu et al. Large Language Models and Causal Inference in Collaboration: A Comprehensive Survey. *In*: FINDINGS OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: NAACL 2025, Albuquerque. Findings of the Association for Computational Linguistics: NAACL 2025. Albuquerque: Association for Computational Linguistics, 2025. p. 7683-7699, 2025 Disponível em: https://aclanthology.org/2025.findings-naacl.427/. Acesso em: 20 out. 2025.

PEARL, Judea. Causal Inference. *In*: NIPS 2008 WORKSHOP ON CAUSALITY: OBJECTIVES AND ASSESSMENT, 2008, Whistler. Proceedings of Machine Learning Research (PMLR). [S.l.]: PMLR, 2010. v. 6, p. 39-58, 2010 Disponível em: https://proceedings.mlr.press/v6/pearl10a.html. Acesso em: 20 out. 2025.

ROSENBAUM, P. R.; RUBIN, D. B. The central role of the propensity score in observational studies for causal effects. Biometrika, Oxford, v. 70, n. 1, p. 41-55, 1983.Disponível em: https://doi.org/10.2307/2335942. Acesso em: 20 out. 2025.

SOUZA, Hugo Vieira Lucena de et al. Uma Análise preditiva de desempenho dos cursos no ENADE com base no perfil socioeconômico e desempenho no ENEM dos alunos. *In*: CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO (CBIE), 6., 2017, Recife. Anais dos Workshops do VI Congresso Brasileiro de Informática na Educação (WCBIE 2017). Porto Alegre: Sociedade Brasileira de Computação (SBC), 2017. p. 684-693. Disponível em: http://milanesa.ime.usp.br/rbie/index.php/wcbie/article/view/7454/5250. Acesso em: 20 out. 2025.

SILVA, Joyci Mesquita Rocha. Utilizando as metodologias ativas de aprendizagem com sucesso. 2018. Trabalho de Conclusão de Curso (Especialização em Educação: Métodos e Técnicas de Ensino) – Universidade Tecnológica Federal do Paraná, Medianeira, 2018. Disponível em: <http://repositorio.utfpr.edu.br/jspui/handle/1/21171>. acesso em 12 abr 2025.

ROCHA, Henrique Martins; LEMOS, Washington de Macedo. Metodologias ativas: do que estamos falando? Base conceitual e relato de pesquisa em andamento. In: IX Simpósio Pedagógico e Pesquisas em Comunicação. Resende, Brasil. Anais do… São Paulo: Associação Educacional Dom Boston, p. 12, 2014. acesso em 13 abr 2025.

VISTA, Nicolas P. Boa; FIGUEIRÓ, Michele F.; CHICON, Patricia M. M. Técnicas de mineração de dados aplicadas aos microdados do ENADE para avaliar o desempenho dos acadêmicos do curso de Ciência da Computação no Rio

Grande do Sul utilizando o software R. *In*: SEMINÁRIO DE PESQUISA CIENTÍFICA E TECNOLÓGICA, 1., 2017, [Cruz Alta?]. Anais [...]. [Cruz Alta?: Unicruz?], 2017. Disponível em: https://www.semanticscholar.org/paper/T%C3%A9cnicas-de-minera%C3%A7%C3%A3o-de-dados-aplicadas-aos-do-o-o-Vista-Figueir%C3%B3/f4d64fd50a1484134afe250503d22ab5699b1ac2. Acesso em: 20 out. 2025.

WEISBERG, Sanford. Applied Linear Regression *Volume 528 de Wiley Series in Probability and Statistics* 2005. Disponivel em : <https://books.google.com.br/books?id=xd0tNdFOOjcC> acesso em 13 abr 2025.

WIRTH, Rüdiger; HIPP, Jochen CRISP-DM: Towards a Standard Process Model for Data Mining, 2000. Disponível em: <https://api.semanticscholar.org/CorpusID:1211505>. Acesso em 20 out 2025.