



UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO
PRÓ-REITORIA DE GRADUAÇÃO
DEPARTAMENTO DE ENGENHARIAS E TECNOLOGIA
CURSO INTERDISCIPLINAR EM TECNOLOGIA DA INFORMAÇÃO

João Felype Feitosa Rêgo

Perfil Socioeconômico e Desempenho no ENEM 2023: uma análise com microdados

PAU DOS FERROS

2025

João Felype Feitosa Rêgo

Perfil Socioeconômico e Desempenho no ENEM 2023: uma análise com microdados

Monografia apresentada a Universidade Federal Rural do Semi-Árido como requisito para obtenção do título de Bacharel em Tecnologia da Informação.

Orientadora: Laysa Mabel de Oliveira Fontes,
Profa. Dra.

PAU DOS FERROS

2025

© Todos os direitos estão reservados a Universidade Federal Rural do Semi-Árido. O conteúdo desta obra é de inteira responsabilidade do (a) autor (a), sendo o mesmo, passível de sanções administrativas ou penais, caso sejam infringidas as leis que regulamentam a Propriedade Intelectual, respectivamente, Patentes: Lei nº 9.279/1996 e Direitos Autorais: Lei nº 9.610/1998. O conteúdo desta obra tomar-se-á de domínio público após a data de defesa e homologação da sua respectiva ata. A mesma poderá servir de base literária para novas pesquisas, desde que a obra e seu (a) respectivo (a) autor (a) sejam devidamente citados e mencionados os seus créditos bibliográficos.

R343p Rêgo, João Felype Feitosa.
Perfil socioeconômico e desempenho no enem
2023: uma análise com microdados / João Felype
Feitosa Rêgo. - 2025.
55 f. : il.

Orientadora: Laysa Mabel de Oliveira Fontes.
Monografia (graduação) - Universidade Federal
Rural do Semi-árido, Curso de Tecnologia da
Informação, 2025.

1. Análise de dados. 2. Desempenho estudantil.
3. ENEM. 4. Fatores socioeconômicos. 5. Políticas
públicas. I. Fontes, Laysa Mabel de Oliveira,
orient. II. Título.

Ficha catalográfica elaborada por sistema gerador automático em
conformidade com AACR2 e os dados fornecidos pelo)
autor(a). Biblioteca Campus Pau dos Ferros
Bibliotecária: Erlanda Maria Lopes da Silva
CRB: 15/966

O serviço de Geração Automática de Ficha Catalográfica para Trabalhos de Conclusão de Curso (TCC's) foi desenvolvido pelo Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (USP) e gentilmente cedido para o Sistema de Bibliotecas da Universidade Federal Rural do Semi-Árido (SISBI-UFERSA), sendo customizado pela Superintendência de Tecnologia da Informação e Comunicação (SUTIC) sob orientação dos bibliotecários da instituição para ser adaptado às necessidades dos alunos dos Cursos de Graduação e Programas de Pós-Graduação da Universidade.

João Felype Feitosa Rêgo

Perfil Socioeconômico e Desempenho no ENEM 2023: uma análise com microdados

Monografia apresentada a Universidade Federal Rural do Semi-Árido como requisito para obtenção do título de Bacharel em Tecnologia da Informação.

Defendida em: 22/07/2025.

BANCA EXAMINADORA

Laysa Mabel de Oliveira Fontes, Profa. Dra. (UFERSA)
Presidente

João Batista de Souza Neto, Prof. Dr. (UFERSA)
Membro Examinador

Reudismam Rolim de Sousa, Prof. Dr. (UFERSA)
Membro Examinador

AGRADECIMENTOS

Agradeço primeiramente a Deus e a Nossa Senhora, por me dar força e sabedoria, em cada um e em todos os momentos, para que eu pudesse chegar até aqui.

Agradeço a minha mãe, Olímpia, e a minha avó, Maria, pelo apoio incondicional, por sempre acreditarem em mim e por serem a base que me faz levantar todos os dias. Amo vocês do tamanho do mundo. Não. Do tamanho do universo.

Agradeço a minha orientadora, Profa. Mabel Fontes, pela confiança, dedicação e, acima de tudo, pela paciência em todo o processo de realização deste trabalho.

Agradeço as minhas amigas Neilla, Hadija e Andrey, por estarem sempre comigo desde o início da minha jornada acadêmica. Amo vocês.

Agradeço ao meu amigo Davi, por estar e compartilhar comigo todo esse processo de estudo e escrita deste trabalho.

Agradeço à banca examinadora, por disponibilizarem seu tempo e pela valiosa contribuição durante o processo de avaliação do meu trabalho.

Agradeço a todos os meus amigos e círculos de amizade conquistados durante a minha jornada acadêmica. Obrigado pelos conhecimentos e, acima de tudo, pelos laços de amizade compartilhados.

RESUMO

Este trabalho investigou a influência de fatores socioeconômicos no desempenho dos participantes do Exame Nacional do Ensino Médio (ENEM) 2023. Para isso, foi utilizado um conjunto de microdados disponibilizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), com informações pessoais, socioeconômicas e de desempenho dos candidatos. Como a média geral não é fornecida oficialmente, foi calculada uma média aritmética a partir das notas das cinco áreas avaliadas. As variáveis foram previamente classificadas em dependentes e independentes e convertidas em *dummies* para aplicação dos métodos de análise adotados: correlação de Spearman, agrupamento com o algoritmo K-Means e árvore de decisão com o modelo *DecisionTreeRegressor*. Os resultados evidenciaram que fatores como a dependência administrativa da escola de conclusão do ensino médio, a renda familiar e a posse de computador na residência estão diretamente relacionados ao desempenho dos estudantes. Constatou-se, ainda, que a prova de matemática e suas tecnologias apresentou maior sensibilidade a esses fatores, enquanto a redação foi a menos influenciada. As análises reforçam a importância de políticas públicas voltadas à redução das desigualdades educacionais, com foco na ampliação do acesso a recursos tecnológicos, no fortalecimento da infraestrutura escolar e na promoção da equidade de oportunidades. O estudo também sugere caminhos para investigações futuras que considerem séries históricas do ENEM e dados de diferentes contextos educacionais.

Palavras-chave: análise de dados; desempenho estudantil; ENEM; fatores socioeconômicos; políticas públicas.

ABSTRACT

This study investigated the influence of socioeconomic factors on the performance of participants in the 2023 edition of the National High School Examination. To this end, a dataset of microdata made available by the National Institute for Educational Studies and Research Anísio Teixeira was used, containing personal, socioeconomic, and performance information of the candidates. Since the overall average score is not officially provided, an arithmetic mean was calculated based on the scores of the five assessed areas. The variables were previously classified as dependent and independent and converted into dummy variables for the application of the analytical methods adopted: Spearman's correlation, clustering with the K-Means algorithm, and decision tree using the DecisionTreeRegressor model. The results showed that factors such as the administrative dependency of the high school of completion, household income, and ownership of a computer at home are directly related to student performance. It was also found that the Mathematics and its Technologies section was more sensitive to these factors, while the essay was the least influenced. The analyses reinforce the importance of public policies aimed at reducing educational inequalities, with a focus on expanding access to technological resources, strengthening school infrastructure, and promoting equal opportunities. The study also suggests directions for future research that consider historical series of the ENEM and data from different educational contexts.

Keywords: data analysis; student performance; ENEM; socioeconomic factors; public policy.

LISTA DE FIGURAS

Figura 1 – Matriz de correlação de Spearman	30
Figura 2 – Árvore de decisão de dois casos extremos	36

LISTA DE GRÁFICOS

Gráfico 1 – Diferenças de desempenho entre os três grupos.....	34
Gráfico 2 – Distribuição dos candidatos nos três grupos	35
Gráfico 3 – Média das áreas do conhecimento por sexo	38
Gráfico 4 – Média dos candidatos por renda	39
Gráfico 5 – Média dos candidatos por posse de computador	40
Gráfico 6 – Média dos candidatos por dependência administrativa da escola	41

LISTA DE QUADROS

Quadro 1 – Resumo dos trabalhos relacionados.....	19
Quadro 2 – Variáveis analisadas	24
Quadro 3 – Escala qualitativa de interpretação da magnitude do coeficiente ρ	26
Quadro 4 – Coeficientes ρ entre as variáveis independentes e MEDIA_NOTAS	31
Quadro 5 – Coeficientes ρ entre as variáveis independentes e NU_NOTA_CN	31
Quadro 6 – Coeficientes ρ entre as variáveis independentes e NU_NOTA_CH	32
Quadro 7 – Coeficientes ρ entre as variáveis independentes e NU_NOTA_LC	32
Quadro 8 – Coeficientes ρ entre as variáveis independentes e NU_NOTA_MT	32
Quadro 9 – Coeficientes ρ entre as variáveis independentes e NU_NOTA_REDACAO.....	33
Quadro 10 – Frequência das variáveis nas técnicas empregadas	42

LISTA DE ABREVIATURAS E SIGLAS

CadÚnico	Cadastro Único
EJA	Educação de Jovens e Adultos
ENEM	Exame Nacional do Ensino Médio
FIES	Fundo de Financiamento Estudantil
INEP	Instituto Nacional de Estudos e Pesquisa Educacional Anísio Teixeira
OCDE	Cooperação e Desenvolvimento Econômico
PISA	Programa Internacional de Avaliação de Estudantes
PROUNI	Programa Universidade Para Todos
SISU	Sistema de Seleção Unificada

SUMÁRIO

1	INTRODUÇÃO	13
1.1	Problematização	13
1.2	Justificativa	14
1.3	Objetivos.....	15
1.3.1	Objetivo Geral	15
1.3.2	Objetivos Específicos	15
1.4	Organização do Trabalho	15
2	REFERENCIAL TEÓRICO.....	16
2.1	Enem	16
2.2	Ciência de Dados	17
2.3	Trabalhos Relacionados.....	18
3	METODOLOGIA	22
3.1	Coleta dos Dados	22
3.2	Preparação dos Dados.....	22
3.3	Variáveis.....	23
3.4	Técnicas de Análise de Dados.....	25
3.4.1	Classificação das Variáveis	25
3.4.2	Coeficiente de Correlação de Spearman	26
3.4.3	Agrupamento com K-Means.....	27
3.4.4	Árvore de Decisão com DecisionTreeRegressor	27
3.4.5	Análise Exploratória	28
3.5	Ambiente Computacional e Tecnologias Utilizadas	28
4	RESULTADOS E DISCUSSÃO	30
4.1	Análise da Correlação de Spearman	30
4.2	Análise de Agrupamento com K-Means.....	34
4.3	Análise da Árvore de Decisão com DecisionTreeRegressor	36
4.4	Análise Exploratória	37
4.5	Discussão Integrada dos Resultados	41
5	CONSIDERAÇÕES FINAIS	44
5.1	Limitações do Estudo	44
5.2	Trabalhos Futuros.....	45

REFERÊNCIAS	46
APÊNDICE A – CONVERSÃO EM VARIÁVEIS <i>DUMMIES</i>.....	49

1 INTRODUÇÃO

Este capítulo apresenta uma visão geral deste trabalho e está organizado da seguinte forma: a Seção 1.1 apresenta a problematização; a Seção 1.2 explicita os argumentos que justificam o desenvolvimento deste trabalho; a Seção 1.3 apresenta os objetivos; e, por fim, a Seção 1.4 apresenta a organização deste trabalho.

1.1 Problematização

Segundo o Art. 205 da Constituição da República Federativa do Brasil, a educação é um direito de todos, que deve ser garantida pelo estado e encorajada pela sociedade para promover o pleno desenvolvimento da pessoa, seu preparo para o exercício da cidadania e sua qualificação para o trabalho (Brasil, 2016). Veloso et al. (2022) destacam que a educação desempenha um papel central no desenvolvimento socioeconômico do país, uma vez que trabalhadores mais qualificados tendem a ser mais produtivos, a receber salários mais elevados e a se adaptar com maior facilidade às novas tecnologias, fatores esses que contribuem significativamente para a redução do trabalho informal.

No entanto, uma análise social e histórica da educação brasileira revela profundas desigualdades no acesso a um ensino de qualidade. Essas disparidades geram elevados índices de evasão escolar e comprometem a formação dos estudantes nas três áreas fundamentais do conhecimento: leitura, escrita e matemática (Melo et al., 2021).

Um dos principais indicadores internacionais de desempenho educacional é o Programa Internacional de Avaliação de Estudantes (PISA), promovido pela Organização para a Cooperação e Desenvolvimento Econômico (OCDE). Aplicado a cada três anos em 81 países, o exame avalia os conhecimentos de estudantes de aproximadamente 15 anos de idade nas áreas de matemática, leitura e ciências.

Em 2022, o Brasil obteve uma pontuação inferior à registrada em 2018 no PISA, mesmo considerando que essa edição apresentou a menor média global desde a criação do programa, no ano 2000. Ainda assim, o desempenho dos estudantes brasileiros permaneceu abaixo da média nas três áreas avaliadas. O secretário-geral da OCDE, Mathias Cormann, destacou que os resultados educacionais estão diretamente relacionados a fatores como as condições econômicas e sociais das famílias (INEP, 2022). Tal observação evidencia as barreiras enfrentadas por estudantes em situação de vulnerabilidade socioeconômica para alcançar um bom desempenho escolar, o que pode comprometer seu acesso ao ensino superior (INEP, 2022).

Já o Exame Nacional do Ensino Médio (ENEM) é aplicado anualmente com o objetivo de avaliar o desempenho dos estudantes da educação básica, tendo como base os conteúdos previstos no currículo do ensino médio. Durante o processo de inscrição, os participantes também respondem a um questionário socioeconômico, que coleta informações pessoais, familiares e relacionadas às condições de vida e estudo dos candidatos (Iensen, 2022).

Neste sentido, emerge a questão central de pesquisa, que busca responder: quais aspectos socioeconômicos influenciaram o desempenho dos participantes no ENEM 2023?

1.2 Justificativa

Melo et al. (2021) afirmam que determinadas variáveis socioeconômicas impactaram o desempenho dos candidatos na prova objetiva do ENEM. Observou-se que municípios com maior percentual de famílias com renda elevada e de mães com maior nível de escolaridade apresentaram, em média, notas mais altas. Contudo, nesses mesmos municípios, verificou-se uma maior dispersão das notas entre os candidatos. No que diz respeito às variáveis raciais, municípios com maior proporção de estudantes autodeclarados brancos apresentaram melhor desempenho médio, ao passo que aqueles com maior presença de estudantes autodeclarados pretos registraram desempenho inferior e maior dispersão nas notas. Na prova de redação, os resultados seguiram a mesma tendência, com um agravante: municípios com maior número de candidatos indígenas apresentaram desempenho inferior em comparação aos demais.

De maneira semelhante, Brito e Pedroso (2023) chegaram a conclusões convergentes. Os autores identificaram que candidatos autodeclarados pretos, pardos ou indígenas obtiveram, em média, desempenho inferior em relação aos demais, sendo que essas desigualdades se intensificaram durante o período da pandemia de COVID-19. Um dos achados mais expressivos do estudo indica que os candidatos indígenas tiveram um desempenho até 196% inferior ao dos demais. Além disso, os autores destacaram que níveis mais elevados de escolaridade dos pais estiveram associados a melhores resultados dos candidatos.

Outro fator relevante apontado por Brito e Pedroso (2023) foi o acesso à internet no domicílio, o qual se mostrou estatisticamente significativo para o desempenho dos estudantes. Aqueles que dispunham de internet em casa apresentaram melhores resultados, especialmente, durante a pandemia, período em que o acesso às aulas remotas tornou-se imprescindível para a continuidade dos estudos.

Diante dos dados apresentados, evidencia-se que o desempenho dos estudantes no ENEM é influenciado por diversas variáveis socioeconômicas. Assim, justifica-se a

importância de investigar como o perfil socioeconômico dos candidatos impacta seus resultados nesse exame, contribuindo para a compreensão das desigualdades educacionais no Brasil.

1.3 Objetivos

Esta seção está organizada da seguinte forma: a Subseção 1.3.1 define o objetivo geral deste trabalho e a Subseção 1.3.2 apresenta os objetivos específicos.

1.3.1 Objetivo Geral

Este trabalho tem como objetivo geral analisar a relação entre o perfil socioeconômico dos estudantes e seu desempenho no ENEM 2023, identificando de que forma variáveis como renda familiar, escolaridade dos pais, acesso a recursos educacionais e condições de moradia influenciam os resultados obtidos no exame.

1.3.2 Objetivos Específicos

Para alcançar o objetivo geral, foram definidos os seguintes objetivos específicos:

- Verificar a existência de correlação entre variáveis socioeconômicas dos candidatos do ENEM 2023 e as notas obtidas nas quatro áreas do conhecimento, na redação e na média geral.
- Analisar individualmente as variáveis socioeconômicas, observando se diferentes categorias influenciam os resultados do ENEM 2023 e se há variância estatisticamente significativa entre os grupos.
- Comparar o perfil dos participantes do ENEM 2023 com diferentes características socioeconômicas para verificar a existência de desigualdades nos resultados.

1.4 Organização do Trabalho

Este trabalho está estruturado da seguinte forma: o Capítulo 2 apresenta a fundamentação teórica que embasa a pesquisa; o Capítulo 3 descreve os procedimentos metodológicos adotados; o Capítulo 4 expõe os resultados obtidos e a respectiva discussão; e, por fim, o Capítulo 5 reúne as considerações finais, limitações do estudo e sugestões para trabalhos futuros.

2 REFERENCIAL TEÓRICO

Este capítulo descreve a base teórica que fundamenta a elaboração deste trabalho, estando organizado da seguinte forma: a Seção 2.1 apresenta uma contextualização sobre o ENEM; a Seção 2.2 aborda os principais conceitos relacionados à ciência de dados; e a Seção 2.3 apresenta trabalhos correlatos ao tema desta pesquisa.

2.1 Enem

Tendo sua origem em meados do século VII, na China, os testes em larga escala tornaram-se comuns ao longo da história e, atualmente, são aplicados em diversos países (Melo et al., 2021). No Brasil, o ENEM foi criado em 1998 com o objetivo inicial de avaliar o nível de aprendizagem dos concluintes da educação básica. A partir de 2009, o exame passou a assumir também a função de principal porta de entrada para o ensino superior no país (INEP, 2025).

Atualmente, os estudantes podem utilizar a nota do ENEM para ingressar no ensino superior por meio de programas como o Sistema de Seleção Unificada (SISU) e o Programa Universidade para Todos (PROUNI). Além disso, é possível utilizar a nota para solicitar financiamento estudantil por meio do Fundo de Financiamento Estudantil (FIES).

Qualquer cidadão que tenha concluído o ensino médio, seja de forma regular ou por programas como a Educação de Jovens e Adultos (EJA), pode utilizar a nota do ENEM para acesso à educação superior. Estudantes que ainda não concluíram essa etapa podem participar do exame na condição de “treineiros”, sendo suas notas utilizadas apenas para fins estatísticos.

A aplicação do ENEM ocorre em dois dias e é composta por 180 questões objetivas, distribuídas entre quatro áreas do conhecimento: (i) Linguagens, Códigos e suas Tecnologias; (ii) Ciências Humanas e suas Tecnologias; (iii) Ciências da Natureza e suas Tecnologias; e (iv) Matemática e suas Tecnologias. Além disso, os candidatos devem elaborar uma redação dissertativo-argumentativa, com base em uma situação-problema apresentada na proposta (INEP, 2025).

Estudantes da rede pública ou em situação de vulnerabilidade socioeconômica, comprovada por inscrição em programas como o Cadastro Único (CadÚnico) ou o Bolsa Família, têm direito à isenção da taxa de inscrição.

Além de aferir o desempenho acadêmico dos participantes, o ENEM também coleta informações por meio de um questionário socioeconômico preenchido no ato da inscrição.

Esses dados compõem os microdados do exame, disponibilizados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) e são amplamente utilizados em estudos e pesquisas acadêmicas voltadas ao monitoramento das desigualdades educacionais e sociais no Brasil (INEP, 2024).

2.2 Ciência de Dados

Com o constante crescimento do acesso à internet, a quantidade de informações disponíveis aumenta exponencialmente a cada dia. Nesse contexto, torna-se evidente a necessidade de ferramentas e métodos que possibilitem a manipulação e a análise eficiente desses dados, a fim de evitar que informações valiosas fiquem ocultas. Tais informações podem ser fundamentais para diversas áreas, como a educação, em que gestores e formuladores de políticas públicas podem tomar decisões mais assertivas voltadas à melhoria do sistema educacional (Souza, 2021).

A partir disso, a ciência de dados surge por meio da aplicação de técnicas e métodos que viabilizam, de forma eficiente, a coleta, tratamento, análise e interpretação de grandes conjuntos de dados (Moutinho et al., 2024).

Segundo Souza (2021), o processo de trabalho na ciência de dados pode ser dividido em seis etapas principais:

1. Entendimento do problema: etapa inicial em que se define o problema a ser investigado, orientando os meios e objetivos da pesquisa;
2. Coleta de dados: ocorre após a definição do problema e envolve a obtenção dos dados relevantes para a análise;
3. Tratamento dos dados: envolve a limpeza e a padronização dos dados, que podem estar em formatos estruturados ou não estruturados;
4. Exploração dos dados: etapa em que se estudam as hipóteses e os métodos de análise mais adequados para o conjunto de dados;
5. Análise profunda dos dados: consiste na aplicação de modelos estatísticos ou computacionais para validar as hipóteses formuladas;
6. Comunicação dos resultados: momento em que os achados são apresentados de forma clara e compreensível, possibilitando a tomada de decisões informadas.

2.3 Trabalhos Relacionados

Os aspectos socioeconômicos relacionados ao desempenho no ENEM têm sido amplamente discutidos na literatura acadêmica, conforme demonstram os estudos a seguir.

Lucena e Santos (2020) concluíram que candidatos cujos pais possuem maior escolaridade, com acesso à internet em casa, autodeclarados brancos e do sexo masculino, tendem a obter melhores desempenhos no exame. Outros fatores associados a um desempenho superior incluem não trabalhar durante os estudos, ter cursado parte significativa do ensino médio em escolas privadas e possuir renda familiar mais elevada.

Fernandes et al. (2021) corroboram os achados anteriores, mas destacam variáveis que impactam negativamente o desempenho dos estudantes, como idade mais avançada e a necessidade de conciliar trabalho e estudo.

Henares et al. (2021) também identificaram correlação positiva entre o desempenho no ENEM e o nível de escolaridade dos pais, sendo o impacto da escolaridade paterna ligeiramente superior ao da materna. Além disso, os autores apontam que os candidatos com melhores notas são, em geral, mais jovens, do sexo masculino, autodeclarados brancos ou amarelos, matriculados no turno diurno de escolas federais ou com ensino médio profissionalizante, localizadas em áreas urbanas.

Com foco em variáveis agregadas por município, Melo et al. (2021) analisaram separadamente os resultados das provas objetivas e da redação. Os autores identificaram que a escolaridade e o nível profissionalizante dos pais ou responsáveis, a renda familiar média e a cor/raça do estudante são determinantes significativos para o desempenho médio municipal. Adicionalmente, variáveis macroestruturais, como renda e perfil social, apresentaram maior capacidade explicativa do que aquelas diretamente ligadas à infraestrutura escolar.

Na análise geográfica do desempenho, Souza (2021) observou que os estudantes da região Centro-Oeste obtiveram os melhores resultados, sendo esta também a segunda região com o maior número de participantes no exame. Em contrapartida, as regiões Norte e Nordeste apresentaram os desempenhos mais baixos.

Em um estudo voltado ao período da pandemia de COVID-19, Brito e Pedroso (2023) analisaram os impactos das variáveis socioeconômicas no desempenho dos candidatos nos dois primeiros anos da crise sanitária. Os autores concluíram que estudantes com renda familiar inferior a três salários mínimos e oriundos de escolas públicas foram os mais prejudicados. A variável cor/raça também se mostrou relevante, com candidatos autodeclarados pretos, pardos

e indígenas apresentando desempenhos inferiores em comparação aos brancos e amarelos, sendo os indígenas os mais afetados, segundo a amostra analisada.

Diante dos estudos citados, este trabalho se distingue por utilizar microdados mais recentes, referentes ao ENEM 2023, disponibilizados pelo INEP. Além disso, foram analisadas 38 variáveis, permitindo uma abordagem mais abrangente. Por meio da linguagem de programação Python e do ambiente Jupyter Notebook, foi possível realizar análises interativas e aplicar diferentes métodos estatísticos e computacionais que ampliaram a compreensão das relações entre variáveis socioeconômicas e o desempenho no exame. Por fim, os diversos métodos utilizados também possibilitaram verificar a consistência e a confiabilidade dos dados, bem como identificar discrepâncias e limitações em relação a abordagens anteriores, conforme discutido no Capítulo 4. O Quadro 1 apresenta um comparativo entre os trabalhos relacionados e a proposta aqui desenvolvida, com o intuito de evidenciar suas singularidades e avanços.

Quadro 1 – Resumo dos trabalhos relacionados

REFERÊNCIA	OBJETIVO	METODOLOGIA	VARIÁVEL	RESULTADO
Lucena & Santos (2020)	Examinar como os fatores socioeconômicos interferem na nota geral do ENEM.	Software Stata versão 13.0, regressão por Mínimos Quadrados Ordinários (MQO) robusta. Dados do ENEM 2016.	Renda familiar, cor/raça, escolaridade do pai e da mãe, se trabalha, tipo de escola.	Candidatos com pai e mãe mais escolarizados, que estudaram em escola privada, que não trabalham e de cor branca foram os que tiveram um melhor desempenho.
Henares et al. (2021)	Analisar como características socioeconômicas impactam em todas as cinco áreas e a média geral do ENEM.	Regressões econômicas. Dados do ENEM 2015 e Censo Escolar 2015.	Escolaridade do pai e da mãe, idade, sexo, dependência administrativa da escola, tipo de escola, localização da escola, cor/raça e turno de estudo.	Candidatos com pai e mãe mais escolarizados, mais jovens, do sexo masculino, que estudaram no período diurno, em escolas federais, de área urbana com ensino profissionalizante foram os que tiveram um melhor desempenho.

(Continua)

(Continuação)

Melo et al. (2021)	Identificar quais variáveis tiveram maior impacto no desempenho dos municípios no ENEM.	Software R, Python, análise exploratória, regressão linear, espacial e de explicação local. Dados do ENEM 2018, microdados do Censo Escolar 2018 e estatísticas do Instituto Brasileiro de Geografia e Estatística (IBGE) 2018.	Renda familiar, nível de escolaridade e profissionalizante das mães, cor/raça, tipo de escola, infraestrutura e ensino das escolas.	As variáveis nível de escolaridade e nível profissionalizante da mãe, renda familiar e cor do candidato foram as que apresentaram relevância significativa no desempenho do município.
Fernandes et al. (2021)	Analisar os fatores que têm impacto na nota de matemática e suas tecnologias no ENEM em Barra do Corda, MA.	Análise de correlação, regressão linear múltipla, Regressão por <i>Stepwise</i> . Dados do ENEM 2021 (Barra do Corda, MA).	Renda familiar, escolaridade do pai e da mãe, profissão do pai e da mãe, cor/raça, sexo, tipo de escola, se tem acesso à internet, idade, quantidade de computadores em casa, se tem empregada doméstica.	Os candidatos com pais mais escolarizados, de cor branca, com renda familiar mais alta e o com acesso à internet em casa apresentaram resultados mais satisfatórios. No entanto, a idade e se conciliam trabalho e estudo apresentaram influência negativa no desempenho.
Souza (2021)	Identificar características relevantes dos participantes no ENEM e se suas relações entre si interferem no desempenho do exame.	Linguagem Python, Google Colab, correlação, regressão linear, estatística descritiva. Dados do ENEM 2019.	Renda familiar, escolaridade dos pais, idade, sexo, cor/raça, tipo de escola, posse de celular, computador, acesso à internet e região.	Os candidatos com pais mais escolarizados, com uma maior renda, de escola privada e com acesso à tecnologia (computador, celular e internet) foram os com o melhor desempenho. As regiões também

(Continua)

(Conclusão)				
				mostraram relevância estatística.
Brito & Pedroso (2023)	Avaliar o impacto que a pandemia de COVID-19 teve no desempenho no ENEM relacionado aos fatores socioeconômicos.	Software R, modelos de regressão multinível. Dados do ENEM 2018 a 2021 (Paraná).	Renda familiar, escolaridade do pai e da mãe, sexo, cor/raça, tipo de escola e acesso à internet em casa.	Os candidatos com renda familiar menor (de até três salários mínimos) e de escola pública foram os que mais sofreram com os impactos negativos em seu desempenho. Além dos candidatos indígenas ou pretos e uma variabilidade entre o desempenho de municípios distintos.
Este trabalho	Investigar quais fatores socioeconômicos exerceram influência sobre o desempenho dos participantes nas quatro áreas do conhecimento, na redação e na média geral do ENEM 2023.	Linguagem Python, Jupyter Notebook, coeficiente de correlação de Spearman, agrupamento com K-Means, árvore de decisão com DecisionTree-Regressor, análise exploratória. Dados do ENEM 2023.	Foram analisadas 39 variáveis, conforme especificado no Quadro 2.	Os candidatos com renda familiar elevada, oriundos de escolas privadas e com posse de computador no domicílio foram os que apresentaram os melhores desempenhos. Em contraste, aqueles sem renda familiar, provenientes de escolas estaduais e sem computador no domicílio obtiveram os piores desempenhos.

Fonte: Autoria própria (2025)

3 METODOLOGIA

Este capítulo descreve o percurso metodológico adotado nesta pesquisa, estando organizado da seguinte forma: a Seção 3.1 descreve a coleta dos dados; a Seção 3.2 descreve o processo de preparação dos dados; a Seção 3.3 define as variáveis selecionadas; a Seção 3.4 apresenta as técnicas de análise de dados adotadas; e, por fim, a Seção 3.5 expõe o ambiente computacional e as tecnologias utilizadas neste trabalho.

3.1 Coleta dos Dados

Os dados deste estudo foram obtidos a partir do acesso aos microdados do ENEM 2023, disponibilizados publicamente pelo INEP. Esses dados correspondem a microdados de segundo nível, ou seja, registros individuais anonimizados de cada participante, fornecendo uma base rica para análises estatísticas e inferenciais com foco em variáveis socioeconômicas e de desempenho.

A partir do acesso a essa base, composta por 3.933.956 candidatos inscritos, foi possível identificar as variáveis disponíveis, com o suporte do dicionário de microdados também fornecido pelo INEP. Esse material auxiliar permitiu selecionar, interpretar e organizar os dados relevantes para este estudo.

Os microdados abrangem um amplo conjunto de dados, incluindo:

- Dados do participante: faixa etária, sexo, estado civil, cor/raça etc.;
- Informações socioeconômicas: renda familiar, número de pessoas na residência, escolaridade e ocupação dos pais etc.;
- Posse de bens: carro, moto, televisão, acesso à internet, computador etc.;
- Dados das provas: notas obtidas nas quatro áreas do conhecimento e na redação, além das respostas da prova objetiva.

O uso dos microdados está amparado em sua disponibilização pública para fins científicos e educacionais, conforme previsto nos termos do INEP. Ainda assim, foram adotadas medidas adicionais de proteção, conforme descrito da Seção 3.3.

3.2 Preparação dos Dados

Dado o grande volume de dados e a presença de registros incompletos, especialmente, de candidatos que não compareceram a pelo menos um dos dias de aplicação das provas, foi

necessário realizar um processo de limpeza e filtragem dos dados. Essa etapa teve como objetivo remover registros inconsistentes, garantir a integridade do conjunto de dados e adequá-lo às análises previstas.

Além disso, foram eliminados dados sensíveis, como nomes, números de inscrição e outras identificações pessoais, assegurando o anonimato dos participantes, em conformidade com princípios éticos e de proteção de dados.

Após essas etapas, o *dataframe*¹ final passou a conter 708.614 candidatos, cujos registros estavam completos e consistentes com os critérios definidos.

Em conformidade com os procedimentos metodológicos adotados por Lucena e Santos (2020), Henares et al. (2021) e Iensen (2022), as variáveis categóricas foram convertidas em variáveis *dummies*. Nessa conversão, cada categoria foi transformada em uma variável binária, assumindo o valor 1 para presença e 0 para ausência da resposta. Essa transformação foi essencial para viabilizar a aplicação das técnicas estatísticas e de aprendizado de máquina utilizadas neste trabalho. A estrutura dos dados após essa etapa está ilustrada no Apêndice A.

3.3 Variáveis

Este estudo selecionou 38 variáveis com o objetivo de ampliar a profundidade das análises, conforme descrito no Quadro 2. Essas variáveis englobam aspectos socioeconômicos, dados pessoais e os resultados dos participantes nas diferentes áreas do conhecimento avaliadas pelo ENEM.

Adicionalmente, foi criada uma nova variável denominada *MEDIA_NOTAS*, correspondente à média aritmética das notas obtidas nas quatro áreas do conhecimento e na redação. Essa inclusão segue a mesma abordagem utilizada por Lucena e Santos (2020), Henares et al. (2021) e Iensen (2022), que também empregaram essa média como indicador geral de desempenho dos candidatos.

A variável *MEDIA_NOTAS* foi incorporada como uma nova coluna no *dataframe* e calculada por meio da seguinte fórmula:

$$MEDIA_NOTAS = \frac{NOTA_CN + NOTA_CH + NOTA_LC + NOTA_MT + NOTA_REDACAO}{5}$$

¹ Neste trabalho, *dataframe* refere-se a uma estrutura bidimensional de organização dos dados, semelhante a uma tabela composta por linhas e colunas.

Onde:

- NOTA_CN: representa a nota da prova de Ciências da Natureza;
- NOTA_CH: representa a nota da prova de Ciências Humanas;
- NOTA_LC: representa a nota da prova de Linguagens e Códigos;
- NOTA_MT: representa a nota da prova de Matemática;
- NOTA_REDACAO: representa a nota da Redação.

Quadro 2 – Variáveis analisadas

VARIÁVEL	DESCRIÇÃO
TP_FAIXA_ETARIA	Faixa etária.
TP_COR_RACA	Cor/raça.
TP_SEXO	Sexo.
TP_ESTADO_CIVIL	Estado civil.
TP_ENSINO	Tipo de instituição que concluiu ou concluirá o ensino médio.
TP_ESCOLA	Tipo de escola do ensino médio (pública ou privada).
TP_LOCALIZACAO_ESC	Localização da escola (urbana ou rural).
TP_DEPENDENCIA_ADM_ESC	Dependência administrativa da escola (federal, estadual, municipal ou privada).
Q001	Até que série seu pai, ou o homem responsável por você, estudou?
Q002	Até que série sua mãe, ou a mulher responsável por você, estudou?
Q003	A partir da apresentação de algumas ocupações, divididas em grupos ordenados, indique o grupo que contempla a ocupação mais próxima da ocupação do seu pai ou do homem responsável por você. Se ele não estiver trabalhando, escolha uma ocupação pensando no último trabalho dele.
Q004	A partir da apresentação de algumas ocupações, divididas em grupos ordenados, indique o grupo que contempla a ocupação mais próxima da ocupação da sua mãe ou da mulher responsável por você. Se ela não estiver trabalhando, escolha uma ocupação pensando no último trabalho dela.
Q005	Incluindo você, quantas pessoas moram atualmente em sua residência?
Q006	Qual é a renda mensal de sua família? Some a sua renda com a dos seus familiares.
Q007	Em sua residência trabalha empregado(a) doméstico(a)?
Q008	Na sua residência tem banheiro?

(Continua)

(Conclusão)

Q009	Na sua residência tem quartos para dormir?
Q010	Na sua residência tem carro?
Q011	Na sua residência tem motocicleta?
Q012	Na sua residência tem geladeira?
Q013	Na sua residência tem freezer (independente ou segunda porta da geladeira)?
Q014	Na sua residência tem máquina de lavar roupa? (o tanquinho NÃO deve ser considerado).
Q015	Na sua residência tem máquina de secar roupa (independente ou em conjunto com a máquina de lavar roupa)?
Q016	Na sua residência tem forno micro-ondas?
Q017	Na sua residência tem máquina de lavar louça?
Q018	Na sua residência tem aspirador de pó?
Q019	Na sua residência tem televisão em cores?
Q020	Na sua residência tem aparelho de DVD?
Q021	Na sua residência tem TV por assinatura?
Q022	Na sua residência tem telefone celular?
Q023	Na sua residência tem telefone fixo?
Q024	Na sua residência tem computador?
Q025	Na sua residência tem acesso à internet?
NOTA_CN	Nota da prova de ciências da natureza.
NOTA_CH	Nota da prova de ciências humanas.
NOTA_LC	Nota da prova de linguagens e códigos.
NOTA_MT	Nota da prova de matemática.
NOTA_REDACAO	Nota da prova de redação.
MEDIA_NOTAS	Média aritmética das notas das quatro áreas do conhecimento e da redação.

Fonte: Autoria própria (2025)

3.4 Técnicas de Análise de Dados

As subseções a seguir expõem as técnicas de análise de dados adotadas neste trabalho.

3.4.1 Classificação das Variáveis

Inicialmente, as colunas do *dataframe*, ou seja, as variáveis analisadas, foram classificadas em dois grupos: variáveis independentes e variáveis dependentes. As variáveis independentes são aquelas que exercem influência sobre outras, podendo interferir em seus

valores. Já as variáveis dependentes são aquelas cujo comportamento se busca compreender com base nos efeitos provocados pelas variáveis independentes. Neste trabalho, a média e as notas dos candidatos em cada área do conhecimento foram classificadas como variáveis dependentes, enquanto as demais variáveis foram classificadas como variáveis independentes (Melo et al., 2021).

3.4.2 Coeficiente de Correlação de Spearman

Com base nessa classificação, foi calculado o coeficiente de correlação entre as variáveis dependentes e independentes, com o objetivo de identificar associações estatísticas relevantes. Utilizou-se o coeficiente de correlação de Spearman (ρ), por se tratar de uma medida não paramétrica adequada para variáveis ordinais ou não normalmente distribuídas, como é o caso dos microdados do ENEM (Field, 2013; Andrade, 2023).

Os resultados de Spearman variam entre -1 e 1. Valores positivos indicam correlação monotônica direta e valores negativos indicam correlação monotônica inversa. A significância estatística da correlação é avaliada pelo valor-p, que, segundo Field (2013), deve ser comparado com um nível de significância $\alpha = 0,05$. Assim:

- Se $p \leq \alpha$, a correlação é estatisticamente significativa;
- Se $p > \alpha$, não há evidência estatística suficiente para afirmar que existe correlação entre as variáveis.

Como as variáveis foram convertidas em *dummies* (valores 0 ou 1), a correlação de Spearman mostrou-se especialmente adequada, pois pode ser aplicada a dados lineares ou não lineares, e não exige distribuição normal (Andrade, 2023).

Para facilitar a interpretação dos coeficientes, adotou-se a escala qualitativa sugerida por Rumsey (2023), conforme apresentada no Quadro 3.

Quadro 3 – Escala qualitativa de interpretação da magnitude do coeficiente ρ

ρ	INTERPRETAÇÃO
± 1	Correlação perfeita (positiva ou negativa)
± 0.70 a ± 0.99	Correlação forte (positiva ou negativa)
± 0.50 a ± 0.69	Correlação moderada (positiva ou negativa)
± 0.30 a ± 0.49	Correlação fraca (positiva ou negativa)
-0.29 a $+0.29$	Correlação desprezível ou ausência de correlação

Fonte: Autoria própria (2025)

3.4.3 Agrupamento com K-Means

Com o objetivo de agrupar candidatos com características socioeconômicas e pessoais semelhantes, foi aplicado o algoritmo de agrupamento K-Means. Esse método é amplamente utilizado em tarefas de segmentação de dados, por sua simplicidade, eficiência computacional e boa performance em conjuntos de dados de média a grande escala (Maia, 2020).

O K-Means organiza os dados com base na proximidade em relação a centróides, ou seja, pontos que representam o centro de cada grupo (*cluster*). Inicialmente, esses centróides são definidos de forma aleatória e os dados são iterativamente realocados em função da menor distância euclidiana aos centróides até que a configuração dos grupos se estabilize.

A escolha do K-Means se justifica por sua capacidade de identificar padrões latentes nos dados, permitindo agrupar indivíduos com perfis semelhantes de forma não supervisionada. A simplicidade interpretativa dos resultados também é um fator relevante, especialmente em estudos com aplicação prática em políticas públicas educacionais.

Com base nas variáveis selecionadas, os candidatos foram segmentados em três grupos distintos, os quais possibilitaram a categorização dos participantes em três perfis de desempenho: (i) desempenho baixo, (ii) desempenho moderado e (iii) desempenho bom.

Essa abordagem facilitou a visualização de grupos sociais com diferentes níveis de desempenho, contribuindo para a identificação de possíveis desigualdades estruturais no acesso à educação e nas condições de aprendizagem.

3.4.4 Árvore de Decisão com DecisionTreeRegressor

Com o objetivo de identificar quais variáveis estão diretamente associadas a um melhor desempenho, representado por notas mais altas, foi treinado um modelo de árvore de decisão utilizando o algoritmo DecisionTreeRegressor. Esse algoritmo é amplamente utilizado em tarefas de regressão e se mostrou adequado para o propósito deste trabalho, ao lidar com dados numéricos e permitir a visualização clara dos fatores que mais influenciam os resultados dos candidatos.

As árvores de decisão permitem representar o processo de tomada de decisão por meio de uma estrutura hierárquica, composta por nós de decisão e ramos, que organizam os dados em caminhos lógicos com base nos valores das variáveis preditoras. Segundo Silva (2021), esses caminhos são divididos em dois tipos: ramos de ações, que representam alternativas

concretas sob controle do decisor; e ramos de acontecimentos, que expressam eventos incertos ou probabilísticos, fora do controle direto.

No contexto deste trabalho, a árvore gerada com o `DecisionTreeRegressor` foi utilizada para identificar padrões socioeconômicos associados a desempenhos mais altos ou mais baixos no ENEM. A partir da estrutura da árvore, foi possível compreender quais características dos candidatos, como posse de bens, escolaridade dos pais, tipo de escola frequentada, entre outras, estão mais fortemente relacionadas à variação nas médias das notas. Esse modelo contribui, portanto, para a análise explicativa e a visualização das influências entre as variáveis e o desempenho final dos participantes.

3.4.5 Análise Exploratória

Com o objetivo de examinar variáveis que apresentaram valores discrepantes ou que se destacaram na representação dos resultados, foi realizada uma análise exploratória mais aprofundada, visando identificar padrões relevantes nos dados. Para isso, as possíveis respostas de cada variável foram segmentadas em grupos. Em seguida, calcularam-se as médias de desempenho correspondentes a cada grupo, bem como a quantidade de candidatos em cada resposta. Essa abordagem possibilitou a identificação de padrões significativos nas diferenças de desempenho entre os grupos, evidenciando possíveis associações entre determinadas características e os resultados obtidos no exame.

A análise exploratória também possibilitou aprofundar a interpretação das variáveis que apresentaram maior relevância nos métodos de análise, contribuindo para a identificação de fatores associados a desempenhos mais elevados no ENEM 2023.

Adicionalmente, a utilização de representações visuais, como gráficos de colunas, facilitou a interpretação e organização das informações, favorecendo uma comunicação mais clara e acessível dos dados analisados (Melo, 2021; Andrade, 2023).

3.5 Ambiente Computacional e Tecnologias Utilizadas

Neste trabalho, as análises foram conduzidas utilizando a linguagem de programação Python, amplamente reconhecida por sua versatilidade e popularidade em projetos de ciência de dados. Por ser uma linguagem de propósito geral e contar com uma ampla variedade de bibliotecas especializadas, o Python mostrou-se especialmente adequado às etapas de tratamento, análise e visualização dos microdados do ENEM (Souza, 2021).

Entre as bibliotecas utilizadas, destacam-se: (i) `pandas`, para manipulação e estruturação de dados tabulares; (ii) `matplotlib` e `seaborn`, para visualização gráfica dos dados; (iii) `scipy.stats`, para o cálculo do coeficiente de correlação de Spearman; (iv) `sklearn.tree`, para construção da árvore de decisão; e (v) `sklearn.cluster`, para a aplicação das técnicas de agrupamento.

O ambiente de desenvolvimento adotado foi o Jupyter Notebook, uma plataforma interativa e de código aberto que permite a combinação de código, visualizações e documentação em um mesmo ambiente. Essa escolha favoreceu a organização do fluxo de trabalho e a reprodutibilidade das análises.

A integração dessas ferramentas possibilitou uma abordagem robusta e eficiente para o tratamento de grandes volumes de dados, garantindo maior controle, clareza e confiabilidade na produção dos resultados (Maia, 2020).

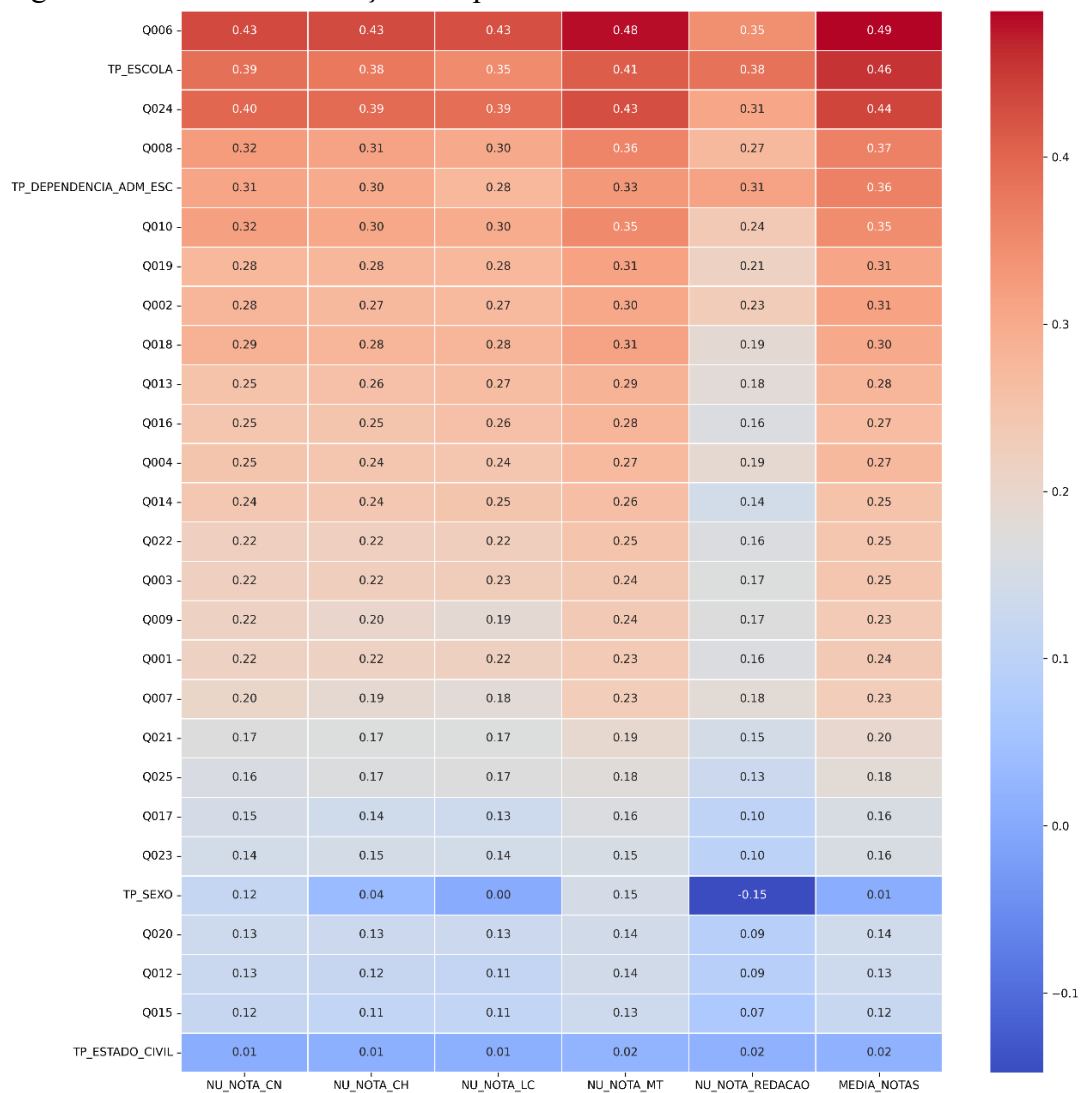
4 RESULTADOS E DISCUSSÃO

Este capítulo descreve os resultados obtidos neste trabalho, estando organizado da seguinte forma: a Seção 4.1 apresenta os resultados da análise de correlação; a Seção 4.2 descreve os resultados do agrupamento; a Seção 4.3 apresenta os resultados da árvore de decisão; e, por fim, a Seção 4.4 discute uma análise exploratória baseada nos resultados.

4.1 Análise da Correlação de Spearman

A Figura 1 apresenta a matriz de correlação de Spearman, representada por um mapa de calor, que destaca as correlações mais relevantes entre as variáveis independentes e dependentes.

Figura 1 – Matriz de correlação de Spearman



Fonte: Autoria própria (2025)

No mapa de calor, a intensidade das cores representa a força da correlação entre as variáveis. Tons quentes mais intensos indicam correlações positivas de maior magnitude, enquanto tons frios mais intensos indicam correlações negativas de maior magnitude. Tons mais claros, independentemente da cor, indicam correlações de menor magnitude e, portanto, potencialmente menos relevantes para a análise.

Para facilitar a interpretação dos resultados ilustrados na Figura 1, a análise será conduzida individualmente para cada variável dependente. Os coeficientes de correlação de Spearman (ρ) apresentados a seguir correspondem às maiores associações identificadas entre as variáveis independentes e as variáveis dependentes, situando-se na faixa de +0.30 a +0.49, o que indica correlações positivas de baixa magnitude. Ressalta-se que todas as correlações apresentadas foram estatisticamente significativas, com valor- $p < 0,05$. As demais variáveis independentes apresentaram correlações desprezíveis e, por esse motivo, foram excluídas da discussão. O Quadro 4 apresenta as variáveis independentes que apresentaram as correlações mais relevantes com a variável MEDIA_NOTAS.

Quadro 4 – Coeficientes ρ entre as variáveis independentes e MEDIA NOTAS

VARIÁVEL	ρ	INTERPRETAÇÃO
Q006	0.49	Correlação fraca positiva
TP_ESCOLA	0.46	
Q024	0.44	
Q008	0.37	
TP_DEPENDENCIA_ADM_ESC	0.36	
Q010	0.35	
Q002	0.31	
Q019	0.31	
Q018	0.30	

Fonte: Autoria própria (2025)

O Quadro 5 apresenta as variáveis independentes que apresentaram as correlações mais relevantes com a variável NU_NOTA_CN.

Quadro 5 – Coeficientes ρ entre as variáveis independentes e NU_NOTA_CN

VARIÁVEL	ρ	INTERPRETAÇÃO
Q006	0.43	Correlação fraca positiva
Q024	0.40	
TP_ESCOLA	0.39	
Q008	0.32	
Q010	0.32	
TP_DEPENDENCIA_ADM_ESC	0.31	

Fonte: Autoria própria (2025)

O Quadro 6 apresenta as variáveis independentes que apresentaram as correlações mais relevantes com a variável NU_NOTA_CH.

Quadro 6 – Coeficientes ρ entre as variáveis independentes e NU_NOTA_CH

VARIÁVEL	ρ	INTERPRETAÇÃO
Q006	0.43	Correlação fraca positiva
Q024	0.39	
TP_ESCOLA	0.38	
Q008	0.31	
Q010	0.30	
TP_DEPENDENCIA_ADM_ESC	0.30	

Fonte: Autoria própria (2025)

O Quadro 7 apresenta as variáveis independentes que apresentaram as correlações mais relevantes com a variável NU_NOTA_LC.

Quadro 7 – Coeficientes ρ entre as variáveis independentes e NU_NOTA_LC

VARIÁVEL	ρ	INTERPRETAÇÃO
Q006	0.43	Correlação fraca positiva
Q024	0.39	
TP_ESCOLA	0.35	
Q008	0.30	
Q010	0.30	

Fonte: Autoria própria (2025)

Por fim, o Quadro 8 apresenta as variáveis independentes que apresentaram as correlações mais relevantes com a variável NU_NOTA_MT.

Quadro 8 – Coeficientes ρ entre as variáveis independentes e NU_NOTA_MT

VARIÁVEL	ρ	INTERPRETAÇÃO
Q006	0.48	Correlação fraca positiva
Q024	0.43	
TP_ESCOLA	0.41	
Q008	0.36	
Q010	0.35	
TP_DEPENDENCIA_ADM_ESC	0.33	
Q018	0.31	
Q019	0.31	
Q002	0.30	

Fonte: Autoria própria (2025)

Por fim, o Quadro 9 apresenta as variáveis independentes que apresentaram as correlações mais relevantes com a variável NU_NOTA_REDACAO.

Quadro 9 – Coeficientes ρ entre as variáveis independentes e NU_NOTA_REDACAO

VARIÁVEL	ρ	INTERPRETAÇÃO
TP_ESCOLA	0.38	Correlação fraca positiva
Q006	0.35	
Q024	0.31	
TP_DEPENDENCIA_ADM_ESC	0.31	

Fonte: Autoria própria (2025)

A análise dos resultados apresentados nos Quadros 4 a 9 revela um padrão consistente de associação positiva entre determinadas variáveis socioeconômicas e o desempenho dos participantes nas diferentes áreas do ENEM 2023. Em especial, destacam-se fatores como níveis mais elevados de renda familiar, tipo de escola em que o ensino médio foi concluído, posse de computador, banheiro e carro na residência, bem como a dependência administrativa da escola frequentada. Em algumas áreas, observou-se ainda a relevância da posse de bens adicionais, como aspirador de pó e televisão, além do nível de escolaridade da mãe ou responsável feminina.

Esses fatores estiveram presentes de forma recorrente nas áreas de Ciências da Natureza, Ciências Humanas e Linguagens e Códigos. A prova de Matemática e suas Tecnologias, por sua vez, apresentou forte associação com praticamente todos os fatores mencionados, indicando sensibilidade acentuada ao perfil socioeconômico dos participantes. A redação, por outro lado, demonstrou menor amplitude nas correlações, com destaque apenas para renda, tipo de escola, posse de computador e dependência administrativa.

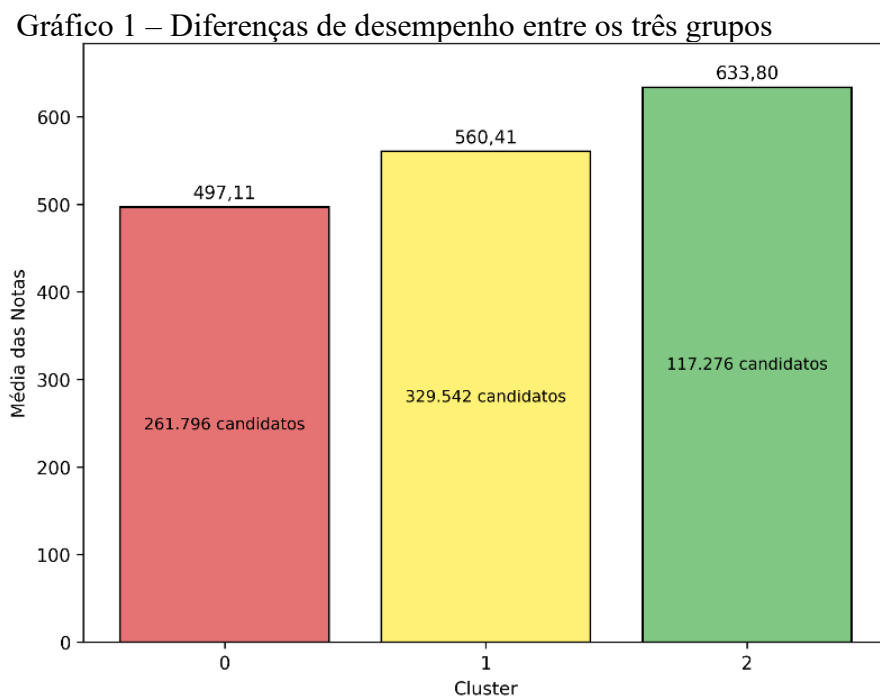
De modo geral, o desempenho médio (MEDIA_NOTAS), por representar a média aritmética das notas das quatro áreas do conhecimento e da redação, apresentou os coeficientes mais elevados e as relações mais robustas com as variáveis analisadas. Em seguida, a prova de Matemática e suas Tecnologias se destacou por apresentar correlações relevantes em diversas dimensões. Já a redação apresentou os resultados mais discretos, o que pode indicar uma menor influência direta das condições socioeconômicas nesse componente específico do exame.

Cabe destacar, ainda, que a variável TP_SEXO apresentou variações entre as diferentes áreas avaliadas. Esse resultado reforça a necessidade de análises exploratórias mais aprofundadas, a fim de compreender os fatores contextuais que podem estar associados a essas

divergências entre os sexos e suas possíveis implicações para a compreensão das desigualdades educacionais, conforme discutido na Seção 4.4.

4.2 Análise de Agrupamento com K-Means

A partir da análise dos três *clusters* formados por meio do algoritmo K-Means, foi possível identificar diferenças significativas entre os grupos. O *Cluster 0* agrupou 261.796 candidatos com características semelhantes e apresentou um desempenho médio de 497,11 pontos. O *Cluster 1* reuniu 329.542 candidatos, com desempenho médio de 560,41 pontos. Por fim, o *Cluster 2* agrupou 117.276 candidatos, que obtiveram um desempenho médio de 633,80 pontos. O Gráfico 1 ilustra as diferenças de desempenho identificadas nos três grupos.



Fonte: Autoria própria (2025)

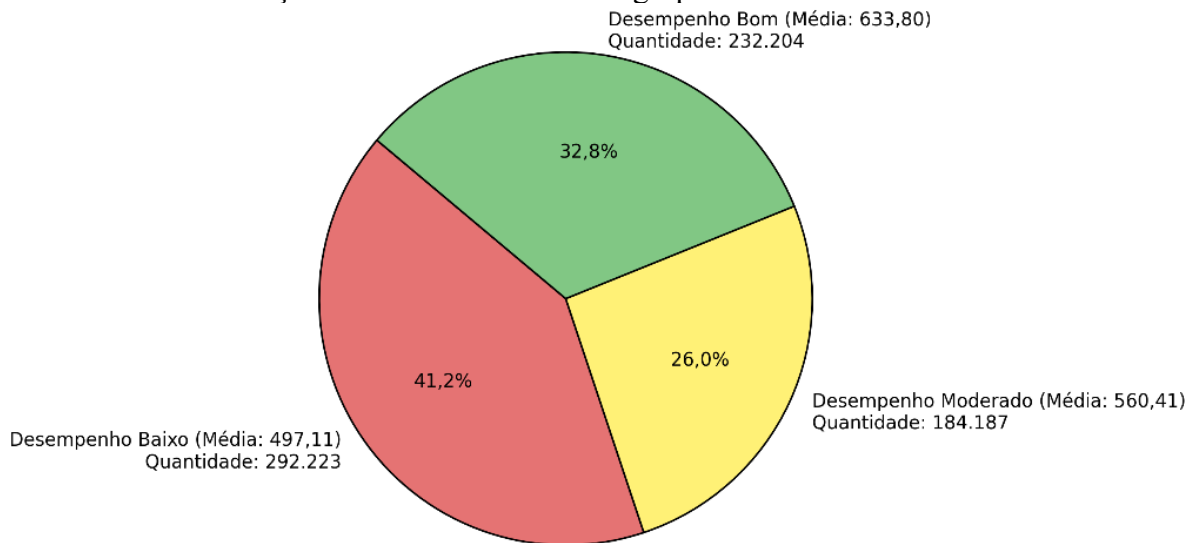
O desempenho médio dos candidatos analisados nesta amostra foi de 549,17 pontos. As notas variaram entre um mínimo de 0,0 e um máximo de 862,58. Observa-se que o *Cluster 1*, que reúne o maior número de candidatos entre os três grupos, apresenta um desempenho mais próximo da média, com uma diferença de apenas 11,24 pontos acima desta. Em contraste, o *Cluster 2*, composto pelo menor número de candidatos, registrou uma diferença de 84,63 pontos superior à média geral. Por fim, o *Cluster 0* apresentou desempenho inferior, com uma diferença de 52,06 pontos abaixo da média geral.

Dessa forma, é possível classificar os candidatos individualmente com base nas médias de desempenho de cada *cluster*. O *Cluster 0*, com média inferior à geral, foi classificado como “desempenho baixo”. O *Cluster 1*, por apresentar a menor diferença em relação à média geral, foi classificado como “desempenho moderado”. Por fim, o *Cluster 2*, cuja média é superior à geral, foi classificado como “desempenho bom”. As médias de cada grupo são:

- *Cluster 0*: desempenho baixo $\approx 497,11$
- *Cluster 1*: desempenho moderado $\approx 560,41$
- *Cluster 2*: desempenho bom $\approx 633,80$

Diante disso, ao analisar a nota média de cada candidato individualmente, é possível classificá-los nas três categorias previamente estabelecidas. O Gráfico 2 ilustra como essa separação foi realizada, evidenciando a distribuição dos candidatos entre os *clusters* de desempenho baixo, moderado e bom.

Gráfico 2 – Distribuição dos candidatos nos três grupos



Fonte: Autoria própria (2025)

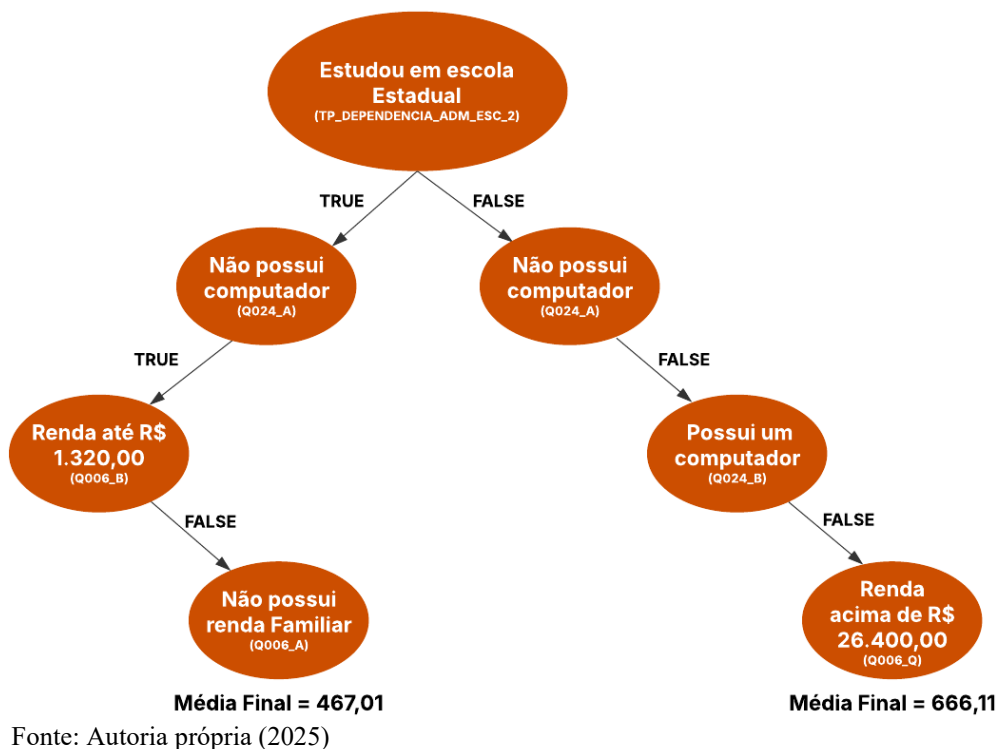
É possível observar que, de acordo com a classificação apresentada anteriormente, 292.223 candidatos, dos 708.614 analisados, foram classificados com desempenho baixo, representando 41,2% do total. Já 184.187 candidatos apresentaram desempenho moderado, o que corresponde a 26% da amostra. Por fim, 232.204 candidatos apresentaram desempenho bom, o que equivale a 32,8% da amostra do ENEM 2023.

4.3 Análise da Árvore de Decisão com DecisionTreeRegressor

Tendo como base os resultados de correlação discutidos na Seção 4.1, foi treinado um modelo de árvore de decisão utilizando o algoritmo DecisionTreeRegressor, com o objetivo de identificar padrões nos perfis de candidatos associados a desempenhos mais altos ou mais baixos no ENEM 2023. Para isso, foram utilizadas as variáveis independentes que apresentaram coeficientes de correlação iguais ou superiores a 0.30 com as variáveis dependentes, de forma a garantir maior relevância na predição.

A árvore gerada permitiu a visualização de diferentes caminhos de decisão, os quais foram analisados para a identificação de cenários extremos, tanto de baixo quanto de alto desempenho. Dentre os diversos caminhos possíveis gerados pela árvore de decisão, dois perfis extremos de candidatos foram destacados. Em ambos os casos, observaram-se três variáveis que se mostraram relevantes na definição dos desempenhos: a dependência administrativa da escola de conclusão do ensino médio, a posse de computador e a renda familiar.

Figura 2 – Árvore de decisão de dois casos extremos



A Figura 2 ilustra os dois caminhos percorridos na árvore de decisão, definidos com base na veracidade das condições presentes em cada nó: o percurso segue à esquerda quando a condição é falsa e à direita quando é verdadeira.

Os dois principais cenários identificados foram:

- Cenário 1 (perfil de menor desempenho): candidato que concluiu o ensino médio em escola pública estadual (445.748 candidatos), não possui computador em casa (315.764 candidatos) e não tem renda familiar declarada (27.929 candidatos). Esse perfil corresponde a um grupo de 22.810 candidatos, cuja média de desempenho foi de 467,01 pontos, sendo este o pior cenário identificado pela árvore.
- Cenário 2 (perfil de maior desempenho): Candidato que não estudou em escola estadual (262.866 candidatos), possui mais de um computador na residência (137.879 candidatos) e tem renda familiar superior a R\$ 26.400,00 (15.928 candidatos). Esse grupo abrange 13.234 candidatos, com uma média de desempenho de 666,11 pontos, representando o melhor desempenho identificado pela árvore.

A comparação entre os dois cenários evidencia a influência direta de fatores socioeconômicos sobre o desempenho no exame, reforçando as desigualdades estruturais presentes no acesso à educação de qualidade no Brasil. A árvore de decisão, ao sintetizar esses padrões, se mostra uma ferramenta eficaz na identificação de perfis críticos, podendo subsidiar ações voltadas à redução dessas desigualdades.

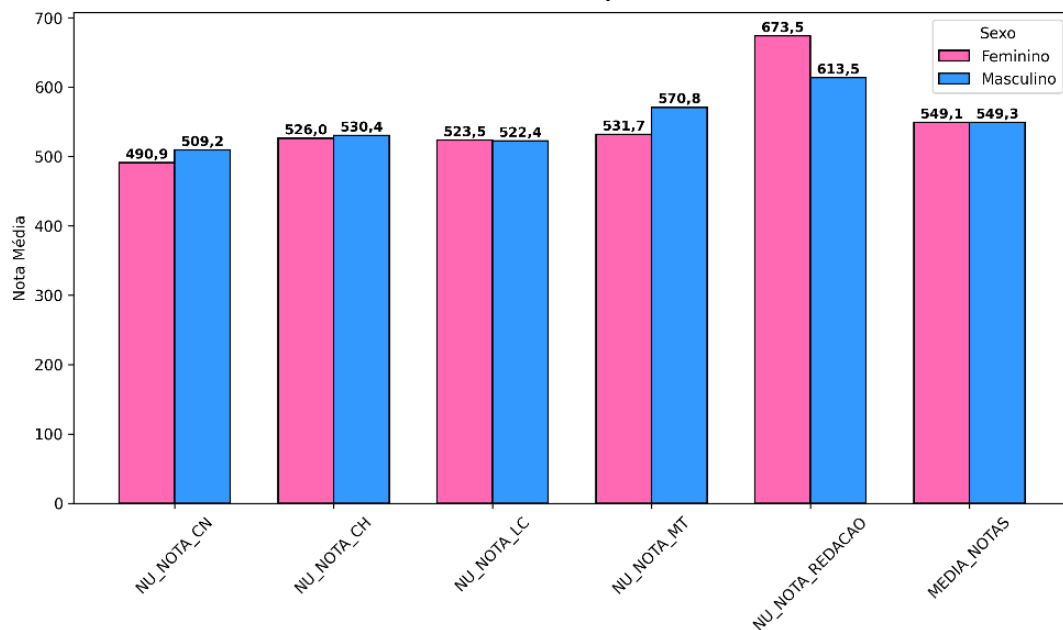
4.4 Análise Exploratória

Devido à diferença significativa observada nos resultados associados à variável TP_SEXO, foi realizada uma análise exploratória com o objetivo de investigar a elevada variância de correlação atribuída a essa variável.

Para isso, o Gráfico 3 apresenta as médias de desempenho dos candidatos nas quatro áreas do conhecimento e na redação, além da média geral, discriminadas por sexo. A amostra considerada nessa análise é composta por 392.440 participantes do sexo feminino e 316.174 do sexo masculino.

Observa-se, com base nos dados apresentados, a justificativa para a discrepância de correlação identificada na Seção 4.1 em relação à variável sexo. Em especial, destacam-se diferenças de desempenho nas provas de Matemática e suas Tecnologias e Redação. Na prova de Matemática, os participantes do sexo masculino obtiveram, em média, 39,1 pontos a mais do que os do sexo feminino. Em contrapartida, na prova de Redação, o desempenho das participantes do sexo feminino foi superior, com média 60,0 pontos acima da dos participantes do sexo masculino.

Gráfico 3 – Média das áreas do conhecimento por sexo



Fonte: Autoria própria (2025)

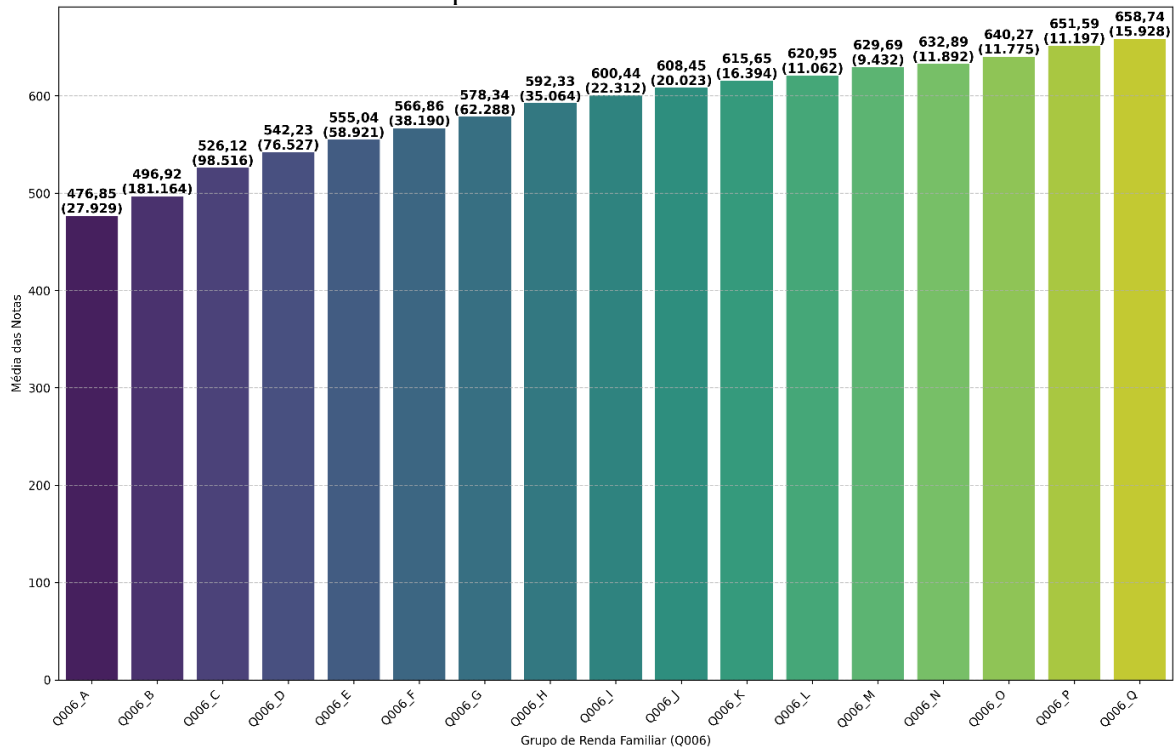
Cabe ressaltar que o tema da redação no ENEM 2023 foi “*Desafios para o enfrentamento da invisibilidade do trabalho de cuidado realizado pela mulher no Brasil*”. Por estar diretamente relacionado às vivências femininas, esse tema pode ter favorecido uma maior identificação por parte das candidatas, facilitando a compreensão e o desenvolvimento da problemática proposta, o que possivelmente se refletiu em um melhor desempenho.

Esses resultados evidenciam um contraste de desempenho entre os sexos em diferentes áreas do conhecimento, o que contribui para a variabilidade observada na correlação dessa variável.

A renda familiar mostrou ter grande influência no desempenho dos estudantes, por representar a correlação mais alta de toda a amostra, e estar presente nos dois cenários extremos criados pela árvore de decisão. Dada sua relevância, foi conduzida uma análise exploratória complementar, com o objetivo de examinar mais detalhadamente os padrões de desempenho associados às diferentes faixas de renda familiar, conforme categorização definida no questionário socioeconômico.

O Gráfico 4 apresenta a média de desempenho dos candidatos, agrupados segundo as faixas de renda familiar, conforme os intervalos definidos no dicionário de microdados do ENEM 2023. Cada coluna representa uma faixa de renda, acompanhada de sua respectiva média de desempenho e da quantidade de candidatos pertencentes àquela categoria. Os resultados evidenciam de forma clara o impacto da renda no desempenho no ENEM: à medida que a renda familiar aumenta, observa-se um crescimento progressivo na média geral dos candidatos.

Gráfico 4 – Média dos candidatos por renda



Fonte: Autoria própria (2025)

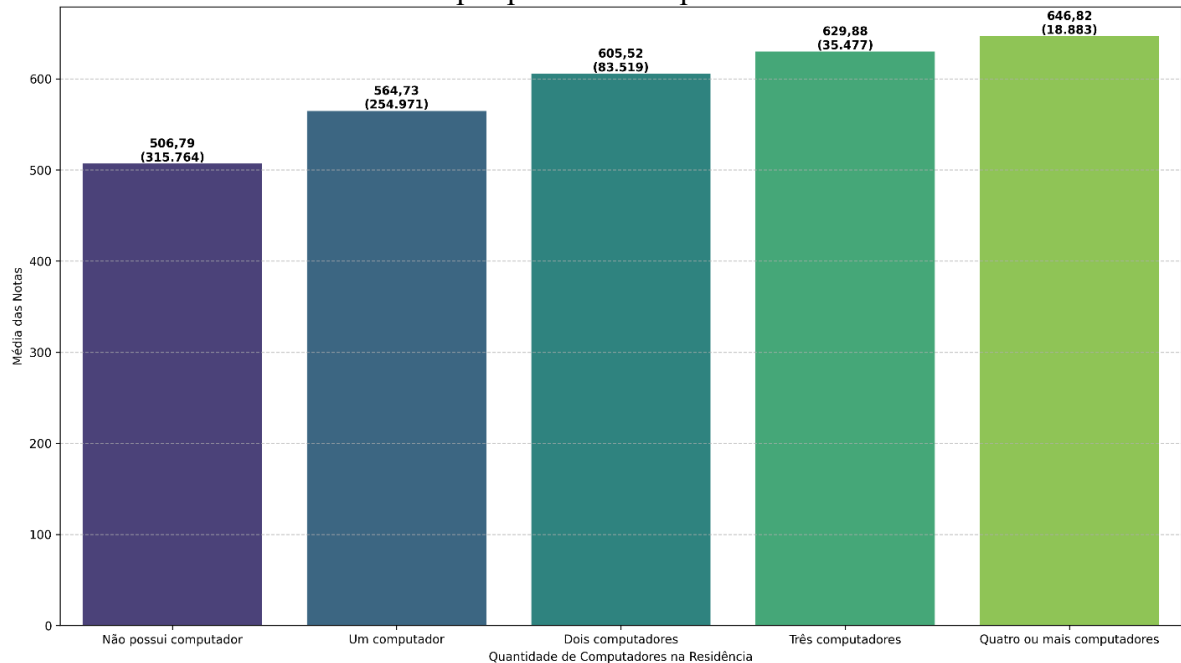
Nos extremos desse espectro, observa-se uma diferença de 181,89 pontos na média de desempenho entre os candidatos que declararam não possuir renda familiar (Q006_A) e aqueles cuja renda familiar é superior a R\$ 26.400,00 (Q006_Q). Embora a faixa de renda com o maior número de candidatos da amostra seja a de R\$ 1.320,00 (Q006_B), a diferença no total de participantes entre os dois grupos extremos é relativamente pequena, com apenas 12.001 candidatos de diferença.

De forma similar à renda, a posse de computador também demonstrou elevada relevância no desempenho, sendo a terceira variável com maior correlação positiva da amostra. O Gráfico 5 ilustra as médias obtidas pelos candidatos de acordo com a quantidade de computadores disponíveis em suas residências (Q024). Cada categoria corresponde a uma resposta do questionário socioeconômico, permitindo observar a relação direta entre o acesso a esse recurso tecnológico e o desempenho acadêmico no exame.

Observa-se um crescimento progressivo da média das notas à medida que aumenta a quantidade de computadores na residência. O grupo com maior número de participantes foi o dos candidatos que declararam não possuir computador (Q024_A), totalizando 315.764 respondentes. Por outro lado, ao comparar os casos extremos, isto é, candidatos sem computador e aqueles com quatro ou mais computadores (Q024_E), verifica-se uma diferença

de 140,03 pontos na média de desempenho, evidenciando o impacto do acesso à tecnologia nos resultados educacionais.

Gráfico 5 – Média dos candidatos por posse de computador



Fonte: Autoria própria (2025)

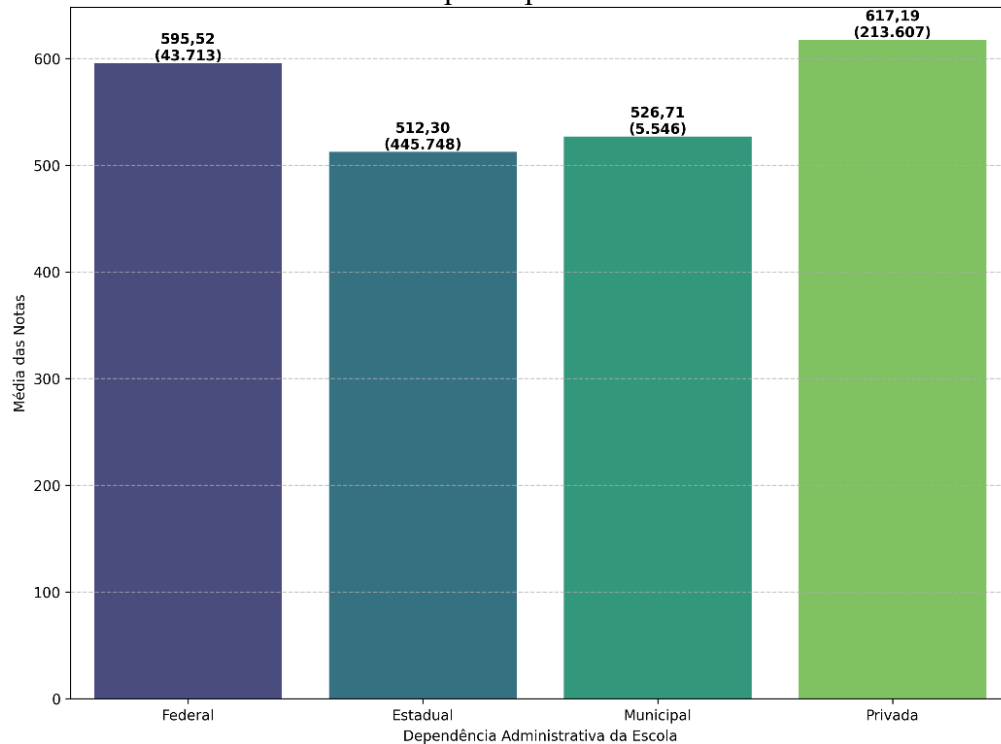
Além disso, a instituição de ensino frequentada pelos candidatos também se revelou um fator determinante. As duas variáveis associadas a essa característica, TP_ESCOLA (tipo de escola) e TP_DEPENDENCIA_ADM_ESC (dependência administrativa da escola de conclusão), figuram entre os maiores coeficientes de correlação com o desempenho. Optou-se por aprofundar a análise da variável TP_DEPENDENCIA_ADM_ESC, por oferecer maior variedade de categorias e, conseqüentemente, possibilitar uma análise mais detalhada, mesmo tendo apresentado uma correlação ligeiramente inferior à de TP_ESCOLA.

O Gráfico 6 apresenta as médias de desempenho dos candidatos de acordo com a dependência administrativa da escola em que concluíram o ensino médio. Nesse contexto, a diferença entre os estudantes provenientes de escolas estaduais e escolas privadas é particularmente expressiva, com uma média de 104,89 pontos a mais para os alunos da rede privada. Esse achado está alinhado o resultado da árvore de decisão, que destacou essa variável como a mais relevante para explicar o desempenho dos participantes.

Apesar de o número de estudantes oriundos de escolas municipais ser expressivamente inferior ao das escolas estaduais, os primeiros apresentaram um desempenho médio mais elevado. Já as escolas federais, mesmo com uma quantidade significativamente menor de

participantes, obtiveram resultados próximos aos das escolas privadas, com uma diferença média de apenas 21,67 pontos entre esses dois grupos.

Gráfico 6 – Média dos candidatos por dependência administrativa da escola



Fonte: Autoria própria (2025)

Esses resultados reforçam o impacto significativo da instituição de ensino no rendimento dos estudantes no exame, apontando para desigualdades estruturais entre as diferentes redes de ensino. Tal constatação destaca a necessidade de políticas públicas voltadas à equidade educacional, sobretudo no que se refere à qualidade da oferta nas redes estaduais e municipais.

4.5 Discussão Integrada dos Resultados

Com base na presença constante de determinadas variáveis nos diferentes métodos de análise utilizados, observa-se que certos atributos apresentam tendência de associação entre fatores socioeconômicos e o desempenho no ENEM.

Os resultados obtidos por meio das técnicas aplicadas indicam que fatores como a escola em que o candidato concluiu o ensino médio, a renda familiar e a posse de computador na residência exercem impacto direto sobre as notas dos participantes do ENEM. Essas três variáveis estiveram presentes tanto nos cenários extremos da árvore de decisão quanto entre as

correlações mais elevadas identificadas na análise estatística, figurando entre as três com maiores valores de correlação em todas as cinco áreas do conhecimento, bem como na média geral.

Embora as correlações com a média geral tenham registrado os maiores coeficientes, algumas áreas demonstraram ser mais sensíveis a esses fatores do que outras. A prova de Matemática e suas Tecnologias revelou-se a mais influenciada por variáveis socioeconômicas, enquanto a Redação apresentou menor sensibilidade a esses aspectos.

Além disso, a análise exploratória permitiu aprofundar a compreensão do impacto da posse de bens duráveis sobre o desempenho. Um exemplo expressivo foi observado entre os 544 candidatos que declararam possuir todos os itens das questões Q010 a Q024 (como geladeira, televisão, aspirador de pó, computador, entre outros). Esse grupo obteve média de 617,61 pontos, valor semelhante ao perfil de melhor desempenho identificado pela árvore de decisão e ao grupo classificado como “desempenho bom” no agrupamento com K-Means.

O Quadro 10 apresenta a frequência das variáveis nos diferentes métodos de análise. Das quatro técnicas utilizadas neste estudo, apenas três permitem a identificação direta das variáveis mais relevantes para o desempenho. O K-Means, por não considerar variáveis-alvo para a formação dos agrupamentos, possui natureza distinta e, por esse motivo, não foi contemplado na composição do quadro.

Quadro 8 – Frequência das variáveis nas técnicas empregadas

VARIÁVEL	FREQUÊNCIA	TÉCNICA
Q006	3/3	Correlação, árvore de decisão e análise exploratória
Q024	3/3	Correlação, árvore de decisão e análise exploratória
TP_ESCOLA	2/3	Correlação e análise exploratória
TP_DEPENDENCIA_ADM_ESC	2/3	Árvore de decisão e análise exploratória
TP_SEXO	1/3	Análise exploratória
Q010 a Q024	1/3	Análise exploratória

Fonte: Autoria própria (2025)

A renda familiar (Q006) demonstrou ser um fator determinante no desempenho dos participantes do ENEM, apresentando relevância em todas as técnicas de análise supervisionadas aplicadas. Esse achado corrobora os resultados de estudos anteriores, como os de Melo et al. (2021), Souza (2021), Fernandes et al. (2021) e Brito e Pedroso (2023). De forma

semelhante, a posse de computador (Q024) também mostrou impacto significativo nas três técnicas analisadas, em consonância com os achados de Souza (2021), que destacou a importância do acesso a esse recurso para o desempenho dos participantes.

As variáveis relacionadas à escola de conclusão do ensino médio (TP_ESCOLA e TP_DEPENDENCIA_ADM_ESC) também apresentaram relevância, ainda que em diferentes técnicas. Esses resultados estão alinhados com estudos de Lucena e Santos (2020), Henares et al. (2021), Souza (2021), e Brito e Pedroso (2023), os quais evidenciam a influência que a escola de conclusão do ensino médio tem na média dos participantes do ENEM.

Adicionalmente, embora algumas variáveis não tenham se destacado em todas as técnicas selecionadas, a análise exploratória permitiu identificar que a posse de determinados bens, como carro, motocicleta, smartphone, entre outros (Q010 a Q024), pode influenciar indiretamente o desempenho dos candidatos, refletindo em melhores condições de estudo e aprendizagem. Esse aspecto também é apontado por Souza (2021) e Fernandes (2021).

5 CONSIDERAÇÕES FINAIS

Este estudo teve como objetivo principal investigar a influência de fatores socioeconômicos sobre o desempenho dos participantes do ENEM 2023. Utilizando os microdados disponibilizados publicamente pelo INEP, foram analisadas variáveis pessoais, socioeconômicas e de desempenho nas provas das quatro áreas do conhecimento e na redação. A partir dessas informações, aplicaram-se diferentes técnicas de análise de dados, incluindo correlação de Spearman, agrupamento com K-Means, árvore de decisão com o algoritmo DecisionTreeRegressor e análise exploratória, com o intuito de identificar padrões e associações relevantes.

Os achados obtidos respondem diretamente à questão central da pesquisa, afirmando que a escola em que o candidato concluiu o ensino médio, a posse de computador no domicílio e a renda familiar exercem influência direta sobre o desempenho dos participantes. Ainda que algumas variáveis, de forma isolada, apresentem coeficientes de correlação baixos, a análise exploratória revelou que o contexto socioeconômico em que o estudante está inserido pode explicar diferenças significativas nos desempenhos obtidos.

Diante desse panorama, reafirma-se a importância de políticas públicas voltadas à redução das desigualdades educacionais no Brasil. Iniciativas que ampliem o acesso às tecnologias, promovam infraestrutura escolar adequada e ofereçam apoio financeiro às famílias em situação de vulnerabilidade são fundamentais para garantir que estudantes de todos os perfis tenham as mesmas oportunidades de acesso a uma educação de qualidade e de preparação para avaliações de larga escala como o ENEM.

5.1 Limitações do Estudo

Apesar das contribuições apresentadas, este estudo possui algumas limitações que devem ser consideradas. Primeiramente, a análise foi conduzida com base em dados autodeclaratórios dos participantes, os quais podem estar sujeitos a imprecisões ou omissões. Além disso, a exclusão de candidatos com informações incompletas pode ter reduzido a representatividade de certos grupos sociais.

Outro ponto importante é que os métodos estatísticos utilizados identificam associações, mas não permitem inferir causalidade. Assim, embora haja evidências de correlações entre variáveis socioeconômicas e desempenho, não é possível afirmar com certeza que uma variável cause diretamente a outra.

5.2 Trabalhos Futuros

A partir dos resultados alcançados, abrem-se diversas possibilidades para estudos futuros. Uma proposta relevante seria a realização de análises longitudinais com dados de edições anteriores do ENEM, o que permitiria observar a persistência ou a mudança no impacto dos fatores socioeconômicos ao longo do tempo. Isso poderia incluir, por exemplo, os efeitos de políticas educacionais específicas ou eventos históricos marcantes, como a pandemia de COVID-19.

Outra linha de investigação seria explorar a adesão ao ENEM, considerando número de inscrições, taxas de comparecimento e fatores relacionados à ausência dos candidatos, o que contribuiria para uma análise mais aprofundada sobre o acesso ao exame.

Também se recomenda a inclusão de dados de outros cenários educacionais, como avaliações estaduais, municipais e privadas, o que permitiria comparar redes de ensino distintas. Novas abordagens metodológicas, como modelos preditivos mais robustos ou técnicas de aprendizado de máquina, também podem ser exploradas para refinar as análises.

Além disso, a realização de um mapeamento sistemático da literatura sobre o tema contribuiria para consolidar o conhecimento já produzido, identificar lacunas de pesquisa e apoiar a construção de políticas públicas baseadas em evidências.

REFERÊNCIAS

ANDRADE, R. J. de. Impacto da pandemia da COVID-19 nos resultados do ENEM do Estado do Paraná. 2023. 55 f. Trabalho de Conclusão de Curso (Especialização em Educação: Métodos e Técnicas de Ensino) – Universidade Tecnológica Federal do Paraná, Medianeira, 2023. Disponível em: <http://repositorio.utfpr.edu.br/jspui/handle/1/35902>. Acesso em: 8 abr. 2025.

BRASIL. Constituição da República Federativa do Brasil: texto constitucional promulgado em 5 de outubro de 1988, com as alterações determinadas pelas Emendas Constitucionais de Revisão nos 1 a 6/94, pelas Emendas Constitucionais nos 1/92 a 91/2016 e pelo Decreto Legislativo no 186/2008. Brasília: Senado Federal, Coordenação de Edições Técnicas, 2016. Disponível em: chrome-extension://efaidnbmnnnibpcajpcgclefindmkaj/https://www2.senado.leg.br/bdsf/bitstream/handle/id/518231/CF88_Livro_EC91_2016.pdf. Acesso em: 6 fev. 2025.

BRASIL. Constituição da República Federativa do Brasil: texto constitucional promulgado em 5 de outubro de 1988, com as alterações determinadas pelas Emendas Constitucionais de Revisão nos 1 a 6/94, pelas Emendas Constitucionais nos 1/92 a 91/2016 e pelo Decreto Legislativo no 186/2008. Brasília: Senado Federal, Coordenação de Edições Técnicas, 2016. Disponível em: chrome-extension://efaidnbmnnnibpcajpcgclefindmkaj/https://www2.senado.leg.br/bdsf/bitstream/handle/id/518231/CF88_Livro_EC91_2016.pdf. Acesso em: 6 fev. 2025.

BRITO, W. H.; PEDROSO, F. P. *Impactos de variáveis socioeconômicas no desempenho no ENEM no primeiro biênio da pandemia de COVID-19*. **Revista de Gestão e Avaliação Educacional**, Santa Maria, v. 12, n. 21, 2023. DOI: <https://doi.org/10.5902/2318133884069>. Disponível em: <https://periodicos.ufsm.br/regae/article/view/84069/61624>. Acesso em: 13 fev. 2025.

FERNANDES, L. C. S.; PEREIRA, L. B.; MOURA, E. G.. Variáveis que influenciaram o desempenho dos candidatos no ENEM em matemática e suas tecnologias: um estudo com candidatos que realizaram a prova em Barra do Corda, Maranhão, em 2021. **Revista Ibero-Americana de Humanidades, Ciências e Educação**, v. 9, n. 9, p. 3587–3599, 2023. Disponível em: <https://periodicorease.pro.br/rease/article/view/11454>. Acesso em: 6 fev. 2025.

FIELD, A. Descobrindo a estatística usando o SPSS. 4. ed. Porto Alegre: Artmed, 2013.

HENARES, L.; MENEZES FILHO, N. A.; KOMATSU, B. K. Como as Desigualdades entre os Alunos se Refletem nas Notas dos Vários Componentes do ENEM?. **Insper**. [s.l.], 2021. Disponível em: https://www.insper.edu.br/wpcontent/uploads/2021/01/Policy_Paper_Enem_2015_v7.pdf. Acesso em: 20 fev. 2025.

IESENSEN, A. R. A. **A influência de fatores socioeconômicos sobre o desempenho individual no Exame Nacional do Ensino Médio (ENEM)**. 2022. 59 f. Monografia (Bacharelado em Economia) – Faculdade de Economia, Administração, Ciências Contábeis e Atuariais, Universidade de São Paulo, São Paulo, 2022.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP). Confira o passo a passo para pedir isenção. Disponível em: <https://www.gov.br/inep/pt-br/assuntos/noticias/enem/confira-o-passo-a-passo-para-pedir-isencao>. Acesso em: 1 fev. 2025.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP). *Divulgados os resultados do Pisa 2022*. Disponível em: <https://www.gov.br/inep/pt-br/assuntos/noticias/acoes-internacionais/divulgados-os-resultados-do-pisa-2022>. Acesso em: 11 fev. 2025.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP). Exame Nacional do Ensino Médio (ENEM). Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem>. Acesso em: 1 fev. 2025.

LUCENA, J. P. O.; SANTOS, H. N. L. A relação entre desempenho no exame nacional do ensino médio e o perfil socioeconômico: um estudo com os microdados de 2016. **Revista de Gestão e Secretariado**, São Paulo, v. 11, n. 2, p. 1-23, 2020. Disponível em: <https://ojs.revistagesec.org.br/secretariado/article/view/994>. Acesso em: 11 de fev. 2025.

MAIA, M. M.; ANDRADE, L. H. F. de; OLIVEIRA, F. F. de. Relação entre proficiência do Enem e situação socioeconômica dos candidatos ao ensino superior por meio do algoritmo k-means. 2020. Trabalho de Conclusão de Curso (Bacharelado em Ciência e Tecnologia) – Universidade Federal Rural do Semi-Árido, Mossoró, 2020. Disponível em: <https://repositorio.ufersa.edu.br/items/9f0805b2-399f-4258-a256-3b42b4c453ae>. Acesso em: 9 abr. 2025.

MELO, R. O.; FREITAS, A. C.; FRANCISCO, E. R.; MOTOKANE, M. T. Impacto das variáveis socioeconômicas no desempenho do Enem: uma análise espacial e sociológica. **Revista de Administração Pública**, Rio de Janeiro, v. 55, n. 6, p. 1271-1294, 2021. Disponível em: [DOI: 10.1590/0034-761220200843](https://doi.org/10.1590/0034-761220200843). Acesso em: 9 fev. 2025.

Microdados e sinopse estatística do ENEM 2023 disponíveis. Disponível em: <https://www.gov.br/inep/pt-br/assuntos/noticias/enem/microdados-e-sinopse-estatistica-do-enem-2023-disponiveis>. Acesso em: 1 fev. 2025.

MOUTINHO, S. O. M.; ALENCAR, D. F.; MARTINS, P. G. M.; CONEGLIAN, C. S. Ciência da Informação e Ciência de Dados: convergências interdisciplinares. *Encontros Bibli*, Florianópolis, v. 29, p. 1-26, 2024. DOI: <https://doi.org/10.5007/1518-2924.2024.e99127>. Disponível em: <https://periodicos.ufsc.br/index.php/eb/article/view/99127/57779>. Acesso em: 6 abr. 2025.

RUMSEY, D. J. What is r value correlation? *Dummies*, 6 fev. 2023. Disponível em: <https://www.dummies.com/article/academics-the-arts/math/statistics/how-to-interpret-a-correlation-coefficient-r-169792/>. Acesso em: 29 abr. 2025.

SILVA, D. N. A. da. Métodos de classificação por teoria da decisão para mensuração de dados do ENEM. 2021. 34 f. Trabalho de Conclusão de Curso (Licenciatura em Matemática) — Universidade Federal de Campina Grande, Centro de Educação e Saúde, Cuité, 2021.

Disponível em: <http://dspace.sti.ufcg.edu.br:8080/jspui/handle/riufcg/24344>. Acesso em: 11 abr. 2025.

SOUZA, T. O. de. Análise de dados: um estudo do perfil dos participantes do ENEM 2019. 2021. 44 f. Monografia (Bacharelado em Ciência da Computação) – Universidade Federal Rural do Semi-Árido, Mossoró, 2021. Disponível em: <https://repositorio.ufersa.edu.br/items/41d53ffd-2f12-4610-a169-20b8beadfdb7>. Acesso em: 6 abr. 2025.

VELOSO, F.; BARBOSA, F. F. H.; PERUCHETTI, P. Impactos da educação no mercado de trabalho. Rio de Janeiro: Instituto Brasileiro de Economia da Fundação Getúlio Vargas (FGV IBRE), 2022. Disponível em: <https://ibre.fgv.br/observatorio-produtividade/artigos/impactos-da-educacao-no-mercado-de-trabalho-0>. Acesso em: 1 mar. 2025.

APÊNDICE A – CONVERSÃO EM VARIÁVEIS *DUMMIES*

VARIÁVEL	NOME DA COLUNA	DESCRIÇÃO	MIN	MAX
TP_FAIXA_ETARIA	TP_FAIXA_ETARIA_1	Menos de 17 anos	0	1
	TP_FAIXA_ETARIA_2	17 anos	0	1
	TP_FAIXA_ETARIA_3	18 anos	0	1
	TP_FAIXA_ETARIA_4	19 anos	0	1
	TP_FAIXA_ETARIA_5	20 anos	0	1
	TP_FAIXA_ETARIA_6	21 anos	0	1
	TP_FAIXA_ETARIA_7	22 anos	0	1
	TP_FAIXA_ETARIA_8	23 anos	0	1
	TP_FAIXA_ETARIA_9	24 anos	0	1
	TP_FAIXA_ETARIA_10	25 anos	0	1
	TP_FAIXA_ETARIA_11	Entre 26 e 30 anos	0	1
	TP_FAIXA_ETARIA_12	Entre 31 e 35 anos	0	1
	TP_FAIXA_ETARIA_13	Entre 36 e 40 anos	0	1
	TP_FAIXA_ETARIA_14	Entre 41 e 45 anos	0	1
	TP_FAIXA_ETARIA_15	Entre 46 e 50 anos	0	1
	TP_FAIXA_ETARIA_16	Entre 51 e 55 anos	0	1
	TP_FAIXA_ETARIA_17	Entre 56 e 60 anos	0	1
	TP_FAIXA_ETARIA_18	Entre 61 e 65 anos	0	1
	TP_FAIXA_ETARIA_19	Entre 66 e 70 anos	0	1
	TP_FAIXA_ETARIA_20	Maior de 70 anos	0	1
TP_SEXO	TP_SEXO_M	Masculino	0	1
	TP_SEXO_F	Feminino	0	1
TP_ESTADO_CIVIL	TP_ESTADO_CIVIL_0	Não informado	0	1
	TP_ESTADO_CIVIL_1	Solteiro(a)	0	1
	TP_ESTADO_CIVIL_2	Casado(a)/Mora com companheiro(a)	0	1
	TP_ESTADO_CIVIL_3	Divorciado(a)/Desquitado(a)/ Separado(a)	0	1
	TP_ESTADO_CIVIL_4	Viúvo(a)	0	1
TP_COR_RACA	TP_COR_RACA_0	Não declarado	0	1
	TP_COR_RACA_1	Branca	0	1
	TP_COR_RACA_2	Preta	0	1
	TP_COR_RACA_3	Parda	0	1
	TP_COR_RACA_4	Amarela	0	1
	TP_COR_RACA_5	Indígena	0	1
TP_ENSINO	TP_ENSINO_1	Ensino regular	0	1
	TP_ENSINO_2	Ensino Especial - Modalidade Substitutiva	0	1
TP_ESCOLA	TP_ESCOLA_1	Pública	0	1
	TP_ESCOLA_2	Privada	0	1
TP_DEPENDENCIA_ESC	TP_DEPENDENCIA_ESC_1	Federal	0	1

(Continua)

(Continuação)

	TP_DEPENDENCIA_ESC_2	Estadual	0	1
	TP_DEPENDENCIA_ESC_3	Municipal	0	1
	TP_DEPENDENCIA_ESC_4	Privada	0	1
TP_LOCALIZACAO_ESC	TP_LOCALIZACAO_ESC_1	Urbana	0	1
	TP_LOCALIZACAO_ESC_2	Rural	0	1
Q001	Q001_A	Nunca estudou	0	1
	Q001_B	Não completou a 4ª série/5º ano do Ensino Fundamental	0	1
	Q001_C	Completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental	0	1
	Q001_D	Completou a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio	0	1
	Q001_E	Completou o Ensino Médio, mas não completou a Faculdade	0	1
	Q001_F	Completou a Faculdade, mas não completou a Pós-graduação	0	1
	Q001_G	Completou a Pós-graduação	0	1
	Q001_H	Não sei	0	1
Q002	Q002_A	Nunca estudou	0	1
	Q002_B	Não completou a 4ª série/5º ano do Ensino Fundamental	0	1
	Q002_C	Completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental	0	1
	Q002_D	Completou a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio	0	1
	Q002_E	Completou o Ensino Médio, mas não completou a Faculdade	0	1
	Q002_F	Completou a Faculdade, mas não completou a Pós-graduação	0	1
	Q002_G	Completou a Pós-graduação	0	1
	Q002_H	Não sei	0	1
Q003	Q003_A	Grupo 1: Lavrador, agricultor sem empregados, boia fria, criador de animais (gado, porcos, galinhas, ovelhas,	0	1

(Continua)

(Continuação)

		cavalos etc.), apicultor, pescador, lenhador, seringueiro, extrativista		
	Q003_B	Grupo 2: Diarista, empregado doméstico, cuidador de idosos, babá, cozinheiro (em casas particulares), motorista particular, jardineiro, faxineiro de empresas e prédios, vigilante, porteiro, carteiro, office-boy, vendedor, caixa, atendente de loja, auxiliar administrativo, recepcionista, servente de pedreiro, repositor de mercadoria	0	1
	Q003_C	Grupo 3: Padeiro, cozinheiro industrial ou em restaurantes, sapateiro, costureiro, joalheiro, torneiro mecânico, operador de máquinas, soldador, operário de fábrica, trabalhador da mineração, pedreiro, pintor, eletricista, encanador, motorista, caminhoneiro, taxista	0	1
	Q003_D	Grupo 4: Professor (de ensino fundamental ou médio, idioma, música, artes etc.), técnico (de enfermagem, contabilidade, eletrônica etc.), policial, militar de baixa patente (soldado, cabo, sargento), corretor de imóveis, supervisor, gerente, mestre de obras, pastor, microempresário (proprietário de empresa com menos de 10 empregados), pequeno comerciante, pequeno proprietário de terras, trabalhador autônomo ou por conta própria	0	1
	Q003_E	Grupo 5: Médico, engenheiro, dentista, psicólogo, economista, advogado, juiz, promotor, defensor, delegado, tenente, capitão, coronel, professor universitário, diretor em empresas públicas ou privadas, político, proprietário de	0	1

(Continua)

(Continuação)

		empresas com mais de 10 empregados.		
	Q003_F	Não sei	0	1
Q004	Q004_A	Grupo 1: Lavradora, agricultora sem empregados, boia fria, criadora de animais (gado, porcos, galinhas, ovelhas, cavalos etc.), apicultura, pescadora, lenhadora, seringueira, extrativista	0	1
	Q004_B	Grupo 2: Diarista, empregada doméstica, cuidadora de idosos, babá, cozinheira (em casas particulares), motorista particular, jardineira, faxineira de empresas e prédios, vigilante, porteira, carteira, office-boy, vendedora, caixa, atendente de loja, auxiliar administrativa, recepcionista, servente de pedreiro, repositora de mercadoria	0	1
	Q004_C	Grupo 3: Padeira, cozinheira industrial ou em restaurantes, sapateira, costureira, joalheira, torneira mecânica, operadora de máquinas, soldadora, operária de fábrica, trabalhadora da mineração, pedreira, pintora, eletricista, encanadora, motorista, caminhoneira, taxista	0	1
	Q004_D	Grupo 4: Professora (de ensino fundamental ou médio, idioma, música, artes etc.), técnica (de enfermagem, contabilidade, eletrônica etc.), policial, militar de baixa patente (soldado, cabo, sargento), corretora de imóveis, supervisora, gerente, mestre de obras, pastora, microempresária (proprietária de empresa com menos de 10 empregados), pequena comerciante, pequena proprietária de terras, trabalhadora autônoma ou por conta própria	0	1

(Continua)

(Continuação)

	Q004_E	Grupo 5: Médica, engenheira, dentista, psicóloga, economista, advogada, juíza, promotora, defensora, delegada, tenente, capitã, coronel, professora universitária, diretora em empresas públicas ou privadas, política, proprietária de empresas com mais de 10 empregados	0	1
	Q004_F	Não sei	0	1
Q005	Q005_1	1, pois moro sozinho(a)	0	1
	Q005_2	2	0	1
	Q005_3	3	0	1
	Q005_4	4	0	1
	Q005_5	5	0	1
	Q005_6	6	0	1
	Q005_7	7	0	1
	Q005_8	8	0	1
	Q005_9	9	0	1
	Q005_10	10	0	1
	Q005_11	11	0	1
	Q005_12	12	0	1
	Q005_13	13	0	1
	Q005_14	14	0	1
	Q005_15	15	0	1
	Q005_16	16	0	1
	Q005_17	17	0	1
	Q005_18	18	0	1
	Q005_19	19	0	1
	Q005_20	20	0	1
Q006	Q006_A	Nenhuma Renda	0	1
	Q006_B	Até R\$ 1.320,00	0	1
	Q006_C	De R\$ 1.320,01 até R\$ 1.980,00	0	1
	Q006_D	De R\$ 1.980,01 até R\$ 2.640,00	0	1
	Q006_E	De R\$ 2.640,01 até R\$ 3.300,00	0	1
	Q006_F	De R\$ 3.300,01 até R\$ 3.960,00	0	1
	Q006_G	De R\$ 3.960,01 até R\$ 5.280,00	0	1
	Q006_H	De R\$ 5.280,01 até R\$ 6.600,00	0	1

(Continua)

(Continuação)

	Q006_I	De R\$ 6.600,01 até R\$ 7.920,00	0	1
	Q006_J	De R\$ 7.920,01 até R\$ 9.240,00	0	1
	Q006_K	De R\$ 9.240,01 até R\$ 10.560,00	0	1
	Q006_L	De R\$ 10.560,01 até R\$ 11.880,00	0	1
	Q006_M	De R\$ 11.880,01 até R\$ 13.200,00	0	1
	Q006_N	De R\$ 13.200,01 até R\$ 15.840,00	0	1
	Q006_O	De R\$ 15.840,01 até R\$ 19.800,00	0	1
	Q006_P	De R\$ 19.800,01 até R\$ 26.400,00	0	1
	Q006_Q	Acima de R\$ 26.400,00	0	1
Q007	Q007_A	Não	0	1
	Q007_B	Sim, um ou dois dias por semana	0	1
	Q007_C	Sim, três ou quatro dias por semana	0	1
	Q007_D	Sim, pelo menos cinco dias por semana	0	1
Q008	Q008_A	Não	0	1
	Q008_B	Sim, um	0	1
	Q008_C	Sim, dois	0	1
	Q008_D	Sim, três	0	1
	Q008_E	Sim, quatro ou mais	0	1
Q009	Q009_A	Não	0	1
	Q009_B	Sim, um	0	1
	Q009_C	Sim, dois	0	1
	Q009_D	Sim, três	0	1
	Q009_E	Sim, quatro ou mais	0	1
Q010	Q010_A	Não	0	1
	Q010_B	Sim, um	0	1
	Q010_C	Sim, dois	0	1
	Q010_D	Sim, três	0	1
	Q010_E	Sim, quatro ou mais	0	1
Q011	Q011_A	Não	0	1
	Q011_B	Sim, um	0	1
	Q011_C	Sim, dois	0	1
	Q011_D	Sim, três	0	1
	Q011_E	Sim, quatro ou mais	0	1
Q012	Q012_A	Não	0	1
	Q012_B	Sim, um	0	1

(Continua)

(Continuação)

	Q012_C	Sim, dois	0	1
	Q012_D	Sim, três	0	1
	Q012_E	Sim, quatro ou mais	0	1
Q013	Q013_A	Não	0	1
	Q013_B	Sim, um	0	1
	Q013_C	Sim, dois	0	1
	Q013_D	Sim, três	0	1
	Q013_E	Sim, quatro ou mais	0	1
Q014	Q014_A	Não	0	1
	Q014_B	Sim, um	0	1
	Q014_C	Sim, dois	0	1
	Q014_D	Sim, três	0	1
	Q014_E	Sim, quatro ou mais	0	1
Q015	Q015_A	Não	0	1
	Q015_B	Sim, um	0	1
	Q015_C	Sim, dois	0	1
	Q015_D	Sim, três	0	1
	Q015_E	Sim, quatro ou mais	0	1
Q016	Q016_A	Não	0	1
	Q016_B	Sim, um	0	1
	Q016_C	Sim, dois	0	1
	Q016_D	Sim, três	0	1
	Q016_E	Sim, quatro ou mais	0	1
Q017	Q017_A	Não	0	1
	Q017_B	Sim, um	0	1
	Q017_C	Sim, dois	0	1
	Q017_D	Sim, três	0	1
	Q017_E	Sim, quatro ou mais	0	1
Q018	Q018_A	Não	0	1
	Q018_B	Sim	0	1
Q019	Q019_A	Não	0	1
	Q019_B	Sim, um	0	1
	Q019_C	Sim, dois	0	1
	Q019_D	Sim, três	0	1
	Q019_E	Sim, quatro ou mais	0	1
Q020	Q020_A	Não	0	1
	Q020_B	Sim	0	1
Q021	Q021_A	Não	0	1
	Q021_B	Sim	0	1
Q022	Q022_A	Não	0	1
	Q022_B	Sim, um	0	1
	Q022_C	Sim, dois	0	1
	Q022_D	Sim, três	0	1

(Continua)

(Conclusão)

	Q022_E	Sim, quatro ou mais	0	1
Q023	Q023_A	Não	0	1
	Q023_B	Sim	0	1
Q024	Q024_A	Não	0	1
	Q024_B	Sim, um	0	1
	Q024_C	Sim, dois	0	1
	Q024_D	Sim, três	0	1
	Q024_E	Sim, quatro ou mais	0	1
Q025	Q025_A	Não	0	1
	Q025_B	Sim	0	1

Fonte: Autoria própria (2025)