Análise de Padrões e Tendências Socioeconômicas no ENEM (2019-2023) por Meio de Clusterização

Rafael Victor Araujo Bernardes[1], **Renato Miranda Filho** [1]

[1] Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais (IFMG)
Sabará, MG — Brasil

rafaelvictor.bernardes@gmail.com, renato.miranda@ifmg.edu.br

Abstract. *This study aimed to characterize the profiles of ENEM candidates from 2019 to 2023 using clustering techniques, analyzing socioeconomic patterns and their relationship with exam performance. The k-means clustering method and the SelectKBest feature selection technique were applied to identify the most relevant factors in group composition. The results highlighted the persistent influence of socioeconomic conditions on score distribution, revealing structural disparities over the years. Additionally, historical trends in candidate profiles were observed, reinforcing the importance of public policies aimed at promoting educational equity.*

Resumo. *Este estudo teve como objetivo caracterizar os perfis de candidatos do ENEM entre 2019 e 2023 por meio de técnicas de clusterização, analisando padrões socioeconômicos e sua relação com o desempenho no exame. Foram aplicados os métodos de agrupamento k-means e de seleção de características SelectKBest para identificar os fatores mais relevantes na composição dos grupos. Os resultados evidenciaram a persistente influência de condições socioeconômicas na distribuição das notas, com disparidades estruturais ao longo dos anos. Além disso, foram observadas tendências históricas nos perfis dos candidatos, reforçando a importância de políticas públicas voltadas à promoção da equidade educacional.*

1. Introdução

O Exame Nacional do Ensino Médio (ENEM) foi criado durante o governo do ex-presidente Fernando Henrique Cardoso no ano de 1998. O Ministério da Educação (MEC) o concebeu originalmente com o propósito de avaliar o desempenho dos estudantes ao fim da escolaridade básica [MEC 2023]. No entanto, ao longo dos anos, o ENEM passou por várias transformações e foi progressivamente moldado para se tornar um instrumento mais versátil e acessível. Hoje, como consequência dos diversos ajustes, o ENEM desempenha um papel central no funcionamento do sistema educacional do país, ao passo em que assumiu novas responsabilidades e constitui o principal pilar de oportunidades estudantis para os candidatos terem acesso a instituições nacionais e internacionais de nível superior.

Essas oportunidades são segmentadas por categorias, como o Sistema de Seleção Unificada (SISU) e o Programa Universidade para Todos (PROUNI), que visam atingir diferentes públicos-alvo. No entanto, ambos os programas convergem no mesmo propósito: construir vias de acesso para estudantes que desejam ingressar no ensino superior nacional, seja ele público ou privado. Vale destacar que os estudantes também

podem utilizar seus resultados como requisito parcial no ingresso a diversas instituições de educação superior internacionais que são parceiras do exame.

Diante das diferentes possibilidades de utilização dos resultados do exame e das diversas campanhas de incentivo e facilitação para a realização da prova, promovidas pelo Ministério da Educação (MEC), o ENEM tornou-se amplamente popular no Brasil, passando de 157.221 inscritos em sua primeira edição para um ápice de 9.519.827 inscritos em 2014 [MEC 2014].

Concomitantemente à crescente popularidade do exame, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) investiu esforços no aperfeiçoamento da coleta de dados referente às características da prova e dos estudantes. Os dados coletados pelo INEP, conhecidos como "Microdados ENEM" [INEP 2023], são de acesso público e abrangem uma ampla gama de informações, contemplando desde detalhes sobre as provas e gabaritos até dados socioeconômicos dos candidatos.

O presente trabalho utiliza os dados referentes às edições de 2019 a 2023 para realizar uma análise experimental, baseada no KDD (*Knowledge Discovery in Databases*). O objetivo deste estudo é caracterizar *clusters* de candidatos em múltiplos anos. Esta proposta faz-se relevante porque, apesar de sua finalidade original, uma análise adequada destas informações pode revelar vulnerabilidades e desigualdades sociais persistentes ao longo dos anos que devem ser sanadas para garantir que a população brasileira continue a se beneficiar do acesso facilitado ao ensino superior de maneira justa e equitativa.

Os resultados obtidos evidenciaram a influência de fatores socioeconômicos na distribuição das notas do ENEM ao longo dos anos analisados. A clusterização permitiu identificar típicos perfis de candidatos, cujas diferenças estavam fortemente associadas a variáveis como renda familiar, escolaridade dos pais, infraestrutura doméstica e localização geográfica. Observou-se que candidatos pertencentes a grupos com melhores condições socioeconômicas apresentaram, de forma consistente, desempenhos superiores no exame. Além disso, a análise temporal evidenciou a persistência dessas desigualdades no período analisado, indicando que, apesar dos avanços no acesso à educação, desafios estruturais ainda impactam significativamente os resultados dos participantes.

O restante deste trabalho está organizado da seguinte maneira: a seção 2 faz a revisão da literatura, com o intuito de identificar trabalhos relacionados ao tema; a seção 3 apresenta a metodologia de pesquisa aplicada no estudo; a seção 4, por sua vez, discute os resultados obtidos e, por fim, a seção 5 apresenta as considerações finais do trabalho.

2. Revisão da Literatura

O Exame Nacional do Ensino Médio se consolidou como um dos principais instrumentos de acesso à educação superior no Brasil, impactando significativamente a vida de milhões de estudantes. Dada a sua importância, diversos estudos foram conduzidos ao longo das últimas décadas para investigar as características do exame e o perfil dos seus candidatos. Neste âmbito, os trabalhos de [Lima et al. 2019] e [Dutra et al. 2023] destacam-se por terem realizado revisões sistemáticas da literatura (RSLs), identificando os objetivos e tipos de análises presentes nas pesquisas que utilizam os dados do ENEM.

Ambas as RSLs convergem em suas conclusões, apontando que fatores socioeconômicos, como renda familiar, idade, sexo, raça e escolaridade dos pais, estão fre-

quentemente associados ao desempenho dos candidatos. Além disso, esses estudos sugerem diversas oportunidades para continuidade de exploração do tema, seja por meio da análise de públicos-alvo específicos (como idosos, presidiários e pessoas com deficiência), ou correlacionando os microdados do ENEM com outras bases de dados. [Lima et al. 2019], por exemplo, propõe a investigação das correlações entre características socioeconômicas e o desempenho dos candidatos em diferentes momentos temporais, uma vez que identificou-se uma carência de estudos com esta abordagem. Esta foi uma sugestão incorporada no presente estudo, que visa preencher este aspecto da literatura.

Sob uma perspectiva quantitativa, [Lima et al. 2019] analisaram 17 publicações entre 2005 e 2016 diretamente relacionadas ao ENEM, enquanto [Dutra et al. 2023] identificaram 19 trabalhos que atenderam aos seus critérios de seleção. Diante do expressivo volume de estudos com temáticas semelhantes, este texto mencionará apenas as principais produções cujos objetivos investigativos dialogam diretamente com a pesquisa atual.

[Franco et al. 2020] investigaram a relação entre características socioeconômicas e o desempenho dos candidatos no ENEM ao longo do tempo. Para tanto, os autores utilizaram dados das edições compreendidas entre 1998 e 2019, bem como os algoritmos: *XGBoost*, *LightGBM*, *ExtraTreesClassifier*, *PCA* e *SFS* para identificar os 10 fatores socioeconômicos mais relevantes de cada ano. Ao fim do trabalho, os autores geraram um *ranking* global aglutinando os 20 fatores socioeconômicos mais relevantes para o desempenho dos candidatos. Os resultados indicaram que características como: posse de computador em casa, tipo de escola e renda familiar estão historicamente correlacionadas com o desempenho dos estudantes.

O estudo de [Silva et al. 2020], por sua vez, possui uma abordagem mais enxuta no que se refere ao volume de dados processados. Os autores valem-se das técnicas de clusterização (*k-means*) e mineração de regras de associação (*apriori*) para identificar quais variáveis socioeconômicas apresentam correlação com o desempenho de alunos do Estado de Minas Gerais, concluintes do ensino médio, na prova do ENEM de 2019. Ao todo, foram analisados 88.659 candidatos separados em dois grupos distintos, um composto por estudantes com notas geralmente mais altas e outro por estudantes com notas geralmente mais baixas. O estudo evidencia que respostas afirmativas às perguntas do questionário socioeconômico, indicando renda familiar elevada, maior nível de escolaridade declarada para a mãe do participante e autodeclaração racial como "Branco", são mais frequentes no grupo de estudantes com notas mais elevadas.

Outro estudo com recorte regionalizado foi realizado por [Carmo et al. 2021]. Neste artigo, os autores analisam dados dos estudantes do Rio Grande do Sul no ENEM de 2019 com o objetivo de identificar a distribuição das características dos candidatos pertencentes aos grupos de melhores e piores médias. Para tanto, os autores valem-se de estáticas descritivas que demonstram a forte relação do perfil socioeconômico do candidato com o seu resultado. As características mais exploradas pelos autores foram aquelas relacionadas ao acesso a recursos digitais, renda familiar, sexo, tipo de escola, cor/raça e grau de estudo dos pais. O trabalho conclui que o grupo composto por estudantes com as melhores médias apresenta as maiores concentrações de candidatos brancos, favorecidos por renda e acesso aos recursos digitais, enquanto o grupo composto por estudantes com as piores médias demonstra possuir características que se distanciam destas mencionadas. [Carmo et al. 2021] destacam ainda que a proporção de estudantes sem acesso à inter-

net no grupo de piores médias ultrapassa 10% e que trabalhos futuros poderiam explorar como estas características têm variado ao longo do tempo.

De modo semelhante, os trabalhos de [Maia et al. 2021] e [Banni et al. 2021] utilizaram a base de dados do ENEM referente ao ano de 2018 para desenvolver suas pesquisas. Em [Maia et al. 2021] os autores desenvolvem a sua própria aplicação do algoritmo *k-means* e realizam análises estatística descritivas acerca das características diretamente relacionadas aos grupos normativos pré-estabelecidos pela Lei n° 12.711/2012, conhecida como a "Lei de Cotas". Já em [Banni et al. 2021] o escopo de análise das características se dá de modo mais abrangente e com ênfase nas estratégias de visualização de dados univariados e bivariados, bem como na utilização de modelos preditivos. [Maia et al. 2021] evidenciam que as *features* determinantes para a alocação dos candidatos em algumas das subdivisões da Lei de Cotas [1]. são estatisticamente mais presentes no *cluster* formado pelos candidatos que geralmente obtêm as notas mais baixas. Já em [Banni et al. 2021] são destacadas as informações de renda familiar per capita, raça e nível de escolaridade dos pais como correlacionadas ao resultado obtido no ENEM.

Outro trabalho de grande relevância para a comunidade científica foi conduzido por [Silva et al. 2014]. Neste estudo, os autores descrevem a aplicação das etapas do KDD para o pré-processamento dos dados do ENEM de 2010 e utilização do algoritmo *apriori* como método de seleção de características. Os resultados apontaram que o nível de escolaridade, a renda familiar e o número de moradores na mesma residência são fatores importantes no desempenho dos estudantes.

Neste contexto, a Tabela 1 relaciona os trabalhos mencionados e as características socioeconômicas consideradas como influentes no desempenho dos candidatos. Por meio dela, observa-se que os atributos relacionados à cor e raça dos candidatos, infraestrutura doméstica, renda familiar e acesso à tecnologia foram frequentemente correlacionados positivamente ao resultado dos estudantes no exame.

[1] https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2012/lei/l12711.htm

Autores	Anos de interesse	Características associadas ao desempenho (não ordenadas e excluindo-se repetições)
[Franco et al. 2020]	1998; 1999; 2000; 2001; 2002; 2003; 2004; 2005; 2006; 2007; 2008; 2009; 2010; 2011; 2012; 2013; 2014; 2015; 2016; 2017; 2018; 2019	(1) Língua estrangeira; (2) Motivos que levaram a participar do ENEM; (3) Interesse por política internacional; (4) Você tem em sua casa? Microcomputador; (5) Fez curso de língua estrangeira; (6) Indique os cursos que você frequenta ou frequentou: Curso superior; (7) Tipo de escola do ensino médio; (8) Conhecimento sobre a atividade de trabalho escolhida; (9) Renda familiar mensal; (10) Sexo; (11) Acesso à internet; (12) Tipo de escola do ensino fundamental; (13) Abandono e/ou reprovação no ensino fundamental; (14) Ano de conclusão do ensino médio; (15) Pretensão de realizar curso profissionalizante após a conclusão do Ensino Médio; (16) Você tem em sua casa? Banheiro (17) Se indicou indígena, qual(is) língua(s) você domina
[Silva et al. 2020]	2019	(1) Cor / Raça autodeclarada (2) Tipo administrativo da escola (3) Existência de computador pessoal no domicílio (4) Nível de escolaridade da mãe (5) Renda média familiar
[Carmo et al. 2021]	2019	(1) Acesso à internet (2) Tipo administrativo da escola (3) Sexo (4) Cor / Raça autodeclarada (5) Renda média familiar (6) Escolaridade dos pais
[Maia et al. 2021]	2018	(1) Tipo administrativo da escola (2) Cor / Raça autodeclarada (3) Renda média familiar
[Banni et al. 2021]	2018	(1) Escolaridade dos pais (2) Renda média familiar (3) Estudante recém-formado (4) Cor / Raça autodeclarada (5) Faixa etária
[Silva et al. 2014]	2010	(1) Quantas pessoas moram com você (2) Nível de escolaridade da mãe (3) Renda média familiar (4) Tipo administrativo da escola

Table 1. Características associadas ao desempenho do candidato por trabalhos relacionados

Conforme proposto por [Lima et al. 2019] e por [Carmo et al. 2021], a proposta apresentada neste trabalho visa explorar como as características de diferentes grupos de candidatos têm variado ao longo do tempo. Como mencionado, existem exemplos de estudo na literatura focados exclusivamente na análise histórica de características selecionadas. Além disto, existem outros exemplos de estudo que aplicaram técnicas de clusterização e seleção de atributos. No entanto, não foram encontrados estudos que unificassem estes métodos com o propósito de compreender como os diferentes grupos de candidatos têm variado ao longo do tempo, bem como o presente trabalho.

Esta nova abordagem do tema se faz especialmente relevante porque, ao identificar padrões recorrentes na relação entre condições socioeconômicas e desempenho acadêmico, este trabalho fornece subsídios para a formulação de políticas públicas mais eficazes, voltadas à redução das disparidades no acesso e na qualidade da educação. Além disso, a abordagem baseada em clusterização permite uma visão mais granular das diferenças entre os candidatos, possibilitando estratégias educacionais mais direcionadas.

Por fim, destaca-se a importância da reavaliação deste tema frente à publicação anual de novos dados. Dito isto, o presente trabalho também visa o preenchimento da lacuna literária identificada para a exploração dos anos posteriores a 2019.

3. Metodologia de Pesquisa

Do ponto de vista procedimental, este trabalho seguiu um plano baseado no KDD de [Fayyad et al. 1996], reconhecendo a robustez desse modelo em processos de descoberta de conhecimento. Vale ressaltar que a escolha do modelo KDD é justificada pela ampla aceitação deste método em trabalhos relacionados à mineração de dados, incluindo as contribuições significativas de [Silva et al. 2014] e [Silva et al. 2020] mencionadas anteriormente.

Desta forma, o plano de desenvolvimento desta pesquisa seguiu as etapas de: (1) seleção; (2) pré-processamento; e (3) transformação dos dados. Adicionalmente, foram incorporadas estratégias específicas, não originalmente previstas em [Fayyad et al. 1996], devido ao objetivo de clusterização dos dados. Estas etapas incluem: (4) definição do algoritmo de clusterização a ser utilizado, com a consequente identificação do número adequado de *clusters* e aplicação da clusterização; e (5) seleção das principais características descritivas de cada *cluster* ao longo dos anos.

Destaca-se, por fim, que o código desenvolvido neste trabalho está publicamente disponível no GitHub [2]. Já o conjunto de dados que foi utilizado pode ser obtido no portal do INEP [3].

3.1. Seleção dos Dados

Após investigar os *datasets* disponibilizadas pelo INEP [2] (1998 a 2023), observou-se que nos anos anteriores a 2019 ocorriam frequentes variações na composição dos dados, por exemplo: ausência de colunas socioeconômicas, que foram padronizadas nos anos posteriores a 2011, ausência de padronização nos conjuntos de opções das colunas quando comparados aos anos posteriores a 2018, quebra de dados em múltiplas colunas que deixaram de existir após 2014, dentre outros.

Deste modo, optou-se pela utilização das bases de dados referentes às edições de 2019 a 2023, visto que estas são as mais recentes no momento de desenvolvimento deste trabalho e elas estão mais coesas e padronizadas entre si, não sendo necessário realizar adaptações bruscas de colunas ou criação de conversores de dados para garantir compatibilidade entre os anos.

3.2. Pré-Processamento

A fase do pré-processamento consistiu na avaliação e no refinamento dos dados coletados. Para isso, utilizou-se a linguagem de programação *Python*, em sua versão 3.10, juntamente com as bibliotecas *Pandas, Scikit-Learn, Category Encoders* e a ferramenta *Jupyter Notebook* como principais recursos para leitura e tratamento das bases de dados.

Inicialmente, as colunas foram categorizadas e selecionadas com base em sua relevância para os objetivos da análise. Este processo foi feito por meio de uma avaliação investigativa das bases de dados e do estudo do dicionário de dados. O dicionário de dados, por sua vez, é a documentação oficial fornecida pelo INEP - em conjunto com os *downloads* dos microdados - que descreve o significado de cada coluna e as opções de preenchimento para cada uma delas [MEC 2021].

[2]https://github.com/rafaelvictor01/IFMG_TCC_ENEM

[3]https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem

As colunas listadas na Tabela 2 foram descartadas uma vez que, por meio da interpretação do dicionário de dados, sabe-se que estas colunas possuem pouca relevância para o objetivo da análise. Foram descartadas características altamente individuais dos candidatos, o que dificultaria a identificação de tendências temporais (como as colunas com prefixo "TX_", que contemplam os gabaritos das provas de cada candidato). Além disso, colunas inexpressivas, como as que tratam da cor da prova do participante, também foram eliminadas. A coluna 'NU_ANO', por seu turno, embora seja fundamental em análises longitudinais, optou-se pelo descarte dela, pois cada ano foi analisado isoladamente, dispensando a necessidade desta informação no contexto específico das análises realizadas.

Nome da coluna	Descrição	Motivo do descarte
NU_INSCRICAO	Número de inscrição	Característica individual
NU_ANO	Ano do Enem	Característica com variância zero
TX_RESPOSTAS_CN	Vetor com as respostas da parte objetiva da prova de Ciências da Natureza	Característica individual
TX_RESPOSTAS_CH	Vetor com as respostas da parte objetiva da prova de Ciências Humanas	Característica individual
TX_RESPOSTAS_LC	Vetor com as respostas da parte objetiva da prova de Linguagens e Códigos	Característica individual
TX_RESPOSTAS_MT	Vetor com as respostas da parte objetiva da prova de Matemática	Característica individual
TX_GABARITO_CN	Vetor com o gabarito da parte objetiva da prova de Ciências da Natureza	Característica individual
TX_GABARITO_CH	Vetor com o gabarito da parte objetiva da prova de Ciências Humanas	Característica individual
TX_GABARITO_LC	Vetor com o gabarito da parte objetiva da prova de Linguagens e Códigos	Característica individual
TX_GABARITO_MT	Vetor com o gabarito da parte objetiva da prova de Matemática	Característica individual
CO_PROVA_CN	Cor da prova de Ciências da Natureza	Característica irrelevante
CO_PROVA_CH	Cor da prova de Ciências Humanas	Característica irrelevante
CO_PROVA_LC	Cor da prova de Linguagens e Códigos	Característica irrelevante
CO_PROVA_MT	Cor da prova de Matemática	Característica irrelevante

Table 2. Colunas previamente descartadas

Ressalta-se que os conjuntos de dados referentes a cada ano de prova serão manipulados de forma isolada, garantindo a preservação das informações específicas de cada edição. Dessa forma, o descarte da coluna "NU_ANO" não compromete a integridade das análises.

Em seguida, a Tabela 3 lista um novo conjunto de colunas descartadas. Neste caso, as colunas foram consideradas redundantes porque, caso fosse necessário, seria possível resgatar o conteúdo destas colunas por meio da decomposição de outras colunas.

Nome da coluna	Descrição	Motivo do descarte
NO_MUNICIPIO_ESC	Nome do município da escola	Se necessário, as informações referentes a estas colunas podem ser extraídas de: CO_MUNICIPIO_ESC
CO_UF_ESC	Código da Unidade da Federação da escola	
SG_UF_ESC	Sigla da Unidade da Federação da escola	
NO_MUNICIPIO_PROVA	Nome do município da aplicação da prova	Se necessário, as informações referentes a estas colunas podem ser extraídas de: CO_MUNICIPIO_PROVA
CO_UF_PROVA	Código da Unidade da Federação da aplicação da prova	
SG_UF_PROVA	Sigla da Unidade da Federação da aplicação da prova	
TP_ANO_CONCLUIU	Ano de Conclusão do Ensino Médio	Essa informação pode ser mais facilmente extraída da coluna: TP_ST_CONCLUSAO

Table 3. Colunas consideradas redundantes

Para lidar com os dados ausentes nos *datasets* avaliados, uma possível abordagem seria a remoção de todas as linhas que contenham quaisquer valores nulos para os atributos. Ao final deste processo, o conjunto de dados seria composto apenas por linhas com todas as colunas preenchidas, o que seria ideal para a execução de algoritmos de clusterização e seleção de atributos. No entanto, observou-se que essa técnica resultaria na perda média de 70% dos registros das bases de dados. Dessa forma, optou-se por avaliar quais

colunas das tabelas possuíam as maiores concentrações de dados nulos e descartá-las ou tratá-las, visando preservar o maior volume possível de instâncias nas tabelas.

A princípio, observou-se que, para todos os anos selecionados, as colunas que consistentemente apresentaram maior número de concentrações de dados nulos foram:

- "CO_MUNICIPIO_ESC" - Código do município da escola;
- "TP_DEPENDENCIA_ADM_ESC" - Dependência administrativa da escola;
- "TP_LOCALIZACAO_ESC" - Tipo de localização da escola;
- "TP_SIT_FUNC_ESC" - Situação de funcionamento da escola;
- "TP_ENSINO" - Tipo de instituição de ensino;

Tomando como exemplo a edição do ano de 2023, foram identificados 2.594.874 registros com a coluna "TP_ENSINO" nula e 2.975.449 registros com as demais colunas nulas. Acredita-se que estas informações relacionadas à escola do candidato não eram de preenchimento obrigatório no momento do cadastro do participante no exame. Isto justificaria a grande ausência de informações. Neste caso, optou-se por remover essas colunas do escopo de análise.

Além disso, foi identificado que outra fonte expressiva de dados nulos nas bases eram as colunas relacionadas às notas dos candidatos para as provas objetivas. Após um processo investigativo, constatou-se que este fenômeno ocorria devido à ausência do candidato no dia de aplicação do exame. Assim, optou-se por considerar apenas os participantes que estiveram presentes nos dois dias de prova. Essa escolha possibilita o descarte das quatro colunas relacionadas à presença dos candidatos nos dias de exame, pois todas assumiram variância igual a zero após a aplicação do filtro. As colunas descartadas foram: "TP_PRESENCA_CN", "TP_PRESENCA_CH", "TP_PRESENCA_LC" e "TP_PRESENCA_MT".

Por fim, observou-se uma peculiaridade nos conjuntos de dados referentes aos anos de 2020 e 2021. Mesmo após a aplicação das etapas descritas anteriormente, ainda havia registros residuais nesses anos com campos do questionário socioeconômico não preenchidos. Foram identificados 27.377 registros residuais para o ano de 2020 e 1 registro residual para o ano de 2021. Optou-se pelo mero descarte desses registros, visto que o volume de dados eliminados é insignificante em relação ao total preservado.

3.3. Transformação dos Dados

A etapa de transformação dos dados teve o objetivo de simplificar e reestruturar as informações para otimizar o subsequente processo de clusterização, reduzindo a granularidade excessiva e agrupando características de forma coerente com os objetivos da análise.

A primeira ação foi a definição de uma métrica de desempenho única, baseada na média simples das notas em todas as áreas do conhecimento. Com isso, as colunas originais — "NU_NOTA_CN" (Ciências da Natureza), "NU_NOTA_CH" (Ciências Humanas), "NU_NOTA_LC" (Linguagens e Códigos), "NU_NOTA_MT" (Matemática) e "NU_NOTA_REDACAO" (Redação) — foram consolidadas em uma única coluna denominada "MEDIA_NOTAS". Esse processo permitiu uma visualização mais direta do desempenho geral de cada participante, simplificando a análise subsequente.

Após a criação da coluna "MEDIA_NOTAS", todas as colunas de notas mencionadas foram descartadas, assim como as colunas complementares relacionadas

à redação: "NU_NOTA_COMP1", "NU_NOTA_COMP2", "NU_NOTA_COMP3", "NU_NOTA_COMP4", "NU_NOTA_COMP5" e "TP_STATUS_REDACAO". Essas colunas detalhavam o desempenho dos candidatos nas competências específicas da redação e a situação final da prova (anulada, em branco, fuga ao tema, etc.), e sua exclusão visou simplificar o conjunto de dados sem perda de informações relevantes para o objetivo da análise.

Em seguida, identificou-se a possibilidade de simplificar variáveis categóricas que continham muitas opções, agrupando respostas que representavam características similares. Tal medida foi essencial para evitar o aumento desnecessário da dimensionalidade da base final e garantir uma análise mais robusta e resistente à existência de *outliers*. As variáveis que passaram por esse processo foram:

- "Q005" - Incluindo você, quantas pessoas moram atualmente em sua residência?;
- "TP_FAIXA_ETARIA" - Faixa etária;
- "Q006" - Qual é a renda mensal de sua família?;

A variável "Q005" originalmente continha 20 opções de respostas, partindo da opção 1 (moro sozinho) até 20 (moro com vinte pessoas). Neste caso, as opções que indicavam de 6 a 20 pessoas na residência foram agrupadas na categoria "5 ou mais pessoas". De maneira similar, a variável "TP_FAIXA_ETARIA" também apresentava 20 opções de respostas, variando de "Menor de 17 anos" a "Maior de 70 anos". As faixas etárias mais elevadas (opções de 14 a 20) foram agrupadas na categoria "Maiores de 40 anos", o que simplificou a análise sem sacrificar a representatividade das informações. Por fim, na variável "Q006", que descrevia a renda familiar, as respostas foram agrupadas de modo que todas as faixas de renda acima de 9 salários mínimos (opções "M", "N", "O", "P" e "Q") fossem consolidadas na categoria "L", que representa as rendas mais elevadas.

Estas decisões evitaram a granularidade excessiva das informações, previniram problemas com *outliers* e foram essenciais para evitar a criação de diversas colunas na subsequente etapa de discretização dos dados categóricos com a aplicação do método *get_dummies*, mantendo sob controle o problema da dimensionalidade das bases.

Além dessas simplificações, a variável "CO_MUNICIPIO_PROVA" foi transformada para representar a macrorregião do Brasil. O código original incluía sete dígitos, dos quais o primeiro indicava a macrorregião, e foi utilizado para criar a nova coluna chamada "MACRO_REGIAO", com apenas cinco possibilidades: Norte, Nordeste, Sudeste, Sul e Centro-Oeste. Essa transformação preservou a informação de localização geográfica dos candidatos sem a necessidade de processar os 5.570 códigos originais de municípios. Para esta conversão utilizou-se a tabela de dígitos disponível no portal do Instituto Brasileiro de Geografia e Estatística (IBGE) [4].

Finalmente, todas as variáveis categóricas restantes foram convertidas para forma numérica utilizando o método *get_dummies*, da biblioteca *Pandas*. Este processo foi necessário, pois variáveis categóricas não podem ser diretamente interpretadas por alguns algoritmos de aprendizado de máquina. Com isso, as colunas categóricas foram transformadas em variáveis binárias, garantindo que o conjunto de dados estivesse devidamente preparado para a clusterização. Outra técnica amplamente utilizada para a transformação

[4]https://www.ibge.gov.br/geociencias/cartas-e-mapas/redes-geograficas/15778-divisoes-regionais-do-brasil.html

das variáveis categóricas em binárias é a aplicação do método *OneHotEncoder*. No entanto, este método se demonstrou menos performático. Após uma tentativa de utilização dele que foi executada ininterruptamente por cerca de 24 horas sem obtenção de resultados, optou-se pelo seu descarte [5].

Após a aplicação do *get_dummies*, toda a tabela foi convertida em variáveis binárias, com exceção da coluna "MEDIA_NOTAS", que não foi incluída no processo de transformação, uma vez que ela não será incluída no processo de clusterização. A coluna referente à média simples das notas dos candidatos será utilizada apenas para traçar futuras comparações entre os grupos naturalmente formados. Vale destacar que não foi necessário aplicar técnicas adicionais de normalização e padronização dos dados. As colunas transformadas já possuem a mesma escala, magnitude e proporcionalidade.

Ao final do processo de transformação, foram geradas, em média, 168 colunas para cada ano sob análise. Para o ano de 2021, especificamente, foram geradas 169 colunas devido à presença da resposta "Não dispõe da informação" para o item de "Cor e Raça" na base de dados. No entanto, essa diferença é irrelevante para os propósitos da análise, uma vez que a estrutura dos dados e a magnitude das variáveis permanecem consistentes entre todos os anos.

Ano de Interesse	Nº de Instâncias Antes do Pré-Processamento	Nº de Instâncias Após o Pré-Processamento	Nº de Colunas Após o Pré-Processamento
2019	5.095.171	3.701.909	168
2020	5.783.109	2.561.304	168
2021	3.389.832	2.238.106	169
2022	3.476.105	2.344.823	168
2023	3.933.955	2.678.264	168

Table 4. Caracterização das bases de dados pré-processadas para cada edição

A Tabela 4 demonstra a quantidade de colunas geradas e de registros preservados após a etapa de pré-processamento. Em média, foi possível preservar 63% do volume total de registros das bases e, para cada ano, serão processados em média 2,7 milhões de perfis de candidatos.

3.4. Clusterização

A análise de agrupamentos, também conhecida como *cluster analysis*, é um conjunto de técnicas multivariadas cuja finalidade principal é agrupar objetos com base nas características que eles possuem [Hair et al. 2009]. A ideia central da clusterização é identificar grupos de objetos que sejam mais similares entre si do que com objetos de outros grupos. Desta forma, dado um conjunto de dados C contendo n objetos, o problema da *k-clusterização* consiste em dividir este conjunto em k subconjuntos disjuntos, chamados de *clusters* [Silva et al. 2020], de modo que a similaridade entre os objetos dentro de cada *cluster* seja maximizada, e a similaridade entre os *clusters* seja minimizada.

[5]Configurações da máquina utilizada descritas no repositório do trabalho.

3.4.1. Definição do Algoritmo de Clusterização

O algoritmo *k-means*, implementado por meio da biblioteca *Scikit-learn*, foi escolhido como a técnica de agrupamento para a presente análise, devido à sua simplicidade e eficiência ao lidar com grandes volumes de dados. O *k-means* demonstrou um melhor balanceamento entre custo computacional e qualidade de agrupamento em comparação com outras técnicas avaliadas.

Em essência, o *k-means* é um algoritmo de clusterização não-hierárquico que busca minimizar a variância interna dos *clusters*, utilizando a distância euclidiana como critério de similaridade entre os pontos. Sua eficiência computacional é um dos principais motivos de sua popularidade, sendo particularmente útil para dados de alta dimensionalidade e já transformados. No caso desta análise, o algoritmo se mostrou adequado para os dados categóricos binários gerados pelo método *get_dummies*, uma vez que o *k-means* é menos sensível à dimensionalidade, o que o torna capaz de operar eficientemente mesmo após a transformação e simplificação das variáveis.

É importante observar que, embora o *k-means* tenha sido o método selecionado, outras abordagens de clusterização foram consideradas ao longo do desenvolvimento do trabalho. Entre elas, destacam-se o *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN), o *Aglomerative Clustering* e o *Balanced Iterative Reducing and Clustering using Hierarchies* (BIRCH), todos amplamente reconhecidos por suas qualidades específicas. Contudo, esses algoritmos se mostraram inviáveis para o processamento dos microdados ENEM, considerando os recursos computacionais disponíveis e ao elevado tempo de processamento que demandariam.

3.4.2. Identificação do Número Adequado de *Clusters* e Aplicação da Clusterização

Um dos desafios centrais na aplicação do algoritmo *k-means* é a determinação do número adequado de *clusters* (*k*), já que o método exige a definição prévia desse parâmetro. Definir um valor apropriado de *k* é fundamental para garantir que os grupos formados sejam semanticamente coerentes com a estrutura subjacente dos dados, evitando tanto a sub-segmentação quanto a super-segmentação dos agrupamentos.

Existem várias abordagens para determinar o número ideal de *clusters*, sendo algumas das mais comuns o *Silhouette Score*, o método de análise de *gap* estatístico e o *Elbow Method*. Cada uma dessas técnicas busca balancear a complexidade do modelo (número de *clusters*) com a variabilidade explicada dentro de cada *cluster*.

Para este trabalho, foi adotado o *Elbow Method*, por ser amplamente utilizado em problemas de mineração de dados educacionais (MDLs) e por fornecer uma interpretação visual clara da relação entre o número de *clusters* e a soma dos erros quadráticos dentro dos *clusters* (*within-cluster sum of squares* - WCSS). O método *k-means* consiste em calcular o WCSS para diferentes valores de *k* e, em seguida, plotar o resultado em um gráfico de linha. O ponto onde houver uma diminuição mais suave na curva — o "cotovelo" — é considerado o valor ideal de *k*, pois representa o melhor balanceamento entre a quantidade de grupos e a simplicidade do modelo.

Desta forma, para cada conjunto de dados foi gerado um gráfico "*Elbow Method*"

ilustrando a relação de diminuição dos erros quadráticos com o crescimento do número de *clusters*. Os dados dos gráficos foram obtidos por meio da aplicação do método *KMeans* da biblioteca *scikit-learn* considerando-se iterações para *k* entre 1 e 9. Os demais parâmetros do método *KMeans* foram configurados da seguinte forma: método de inicialização (*init*) com a função *k-means++*, sementes de centroide (*n_init*) com valor *"auto"*, número máximo de iterações do algoritmo (*max_iter*) com o valor 10 e *random_state* igual a 72769 para reprodução dos mesmos resultados.

Ao utilizar a função de inicialização como *k-means++*, o *KMeans* irá selecionar os centroides iniciais do cluster usando amostragem com base em uma distribuição de probabilidade empírica da contribuição dos pontos para a inércia geral. Esta técnica acelera a convergência e otimiza o tempo de processamento para grandes volumes de dados. Vale destacar que o algoritmo efetivamente implementado é o *"greedy k-means++"*. Ele se difere do *"vanilla k-means++"* por fazer várias tentativas em cada etapa de amostragem e escolher o melhor centroide entre elas [scikit-learn developers 2025]. A configuração de *n_init* com valor *"auto"* permite a utilização do melhor centroide calculado pelo *"greedy k-means++"*. Já a limitação do número máximo de iterações com o valor 10 foi necessário para lidar com o grande volume de registros.

Após uma avaliação dos gráficos gerados para todos os anos, observou-se que as curvas plotadas nos gráficos possuíam aspecto semelhante a curvas exponenciais decrescentes, sem a visualização nítida do "cotovelo", conforme ilustrado pela Figura 1. Neste caso, foram considerados como ideais os pontos onde se observou que a inclusão de novos *clusters* não traria ganho significativo para a redução do WCSS. Este ponto foi encontrado com valor de *k* igual a 4 para os anos de 2019 e 2020, com o valor de *k* igual a 5 para os anos de 2021 e 2022 e para *k* igual a 7 para a edição de 2023.

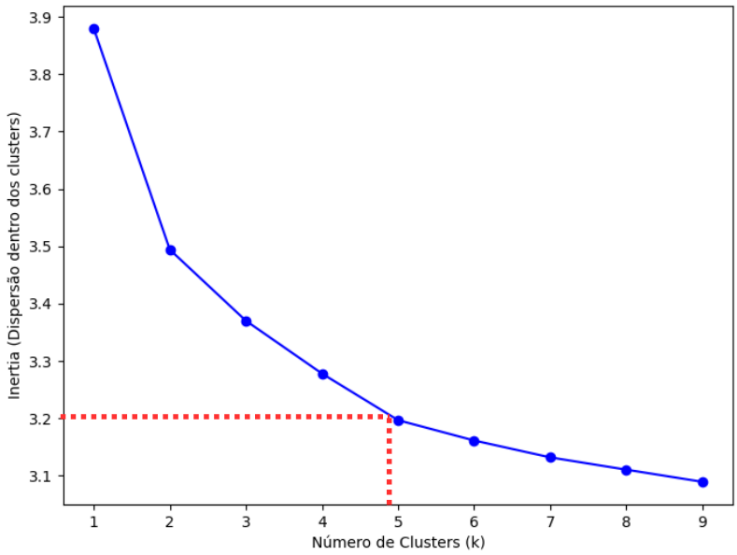

Figure 1. Gráfico *Elbow Method* - Número de *K* vs WCSS - 2022

Realizando a média simples entre os valores encontrados para *k*, chegamos ao resultado que foi adotado de *k* igual a 5. A escolha deste número médio de *clusters* para todos os anos, em detrimento dos valores específicos mencionados anteriormente, teve

como objetivo manter a consistência dos resultados ao longo das diferentes edições do ENEM, facilitando a comparação entre os agrupamentos e assegurando que o modelo utilizado fosse igualmente eficiente em todos os conjuntos de dados analisados, mas mantendo a simplicidade do modelo.

Por fim, o processo de clusterização seguiu a mesma aplicação do método *KMeans* da biblioteca *scikit-learn* com a mesma atribuição de parâmetros descrita anteriormente, mas com a utilização fixa do valor *k* igual a 5. Neste ponto, uma nova coluna chamada `K_Cluster` foi criada visando designar a qual *cluster* cada registro pertence, ação necessária para viabilização das análises por agrupamentos.

3.5. Seleção das Principais Características Descritivas de Cada *Cluster*

A etapa final da metodologia consistiu na seleção das características mais relevantes para a descrição de cada *cluster*. Para essa tarefa, foram considerados diferentes algoritmos de *feature selection* incluindo: *RandomForestClassifier*, *Lasso*, *Principal Component Analysis (PCA)* e *SelectKBest*.

Optou-se pela utilização do algoritmo *SelectKBest*, em conjunto com a função de pontuação *chi2* pois a aplicação do *RandomForestClassifier* mostrou-se inviável devido ao excessivo tempo de processamento exigido. O *PCA*, embora eficiente na redução da dimensionalidade, transforma as variáveis originais em componentes principais (*principal components* – PCs), tornando mais complexa a identificação direta das características mais relevantes em cada cluster, pois exigiria uma etapa adicional de decomposição dos PCs. O *Lasso*, por sua vez, apesar de eficiente na seleção de atributos em problemas de regressão linear, não captura de maneira adequada relações não lineares nos dados [Li et al. 2005], o que poderia comprometer a representatividade das variáveis selecionadas.

A escolha da função *chi2* foi motivada por sua capacidade de medir a dependência entre variáveis categóricas e a variável-alvo, identificando aquelas com maior poder de discriminação entre os *clusters*. Segue abaixo a fórmula matemática do *chi2*, onde podemos observar como a estatística é calculada:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

A utilização do método *SelectKBest* com a função *chi2* demonstrou-se eficiente na identificação das variáveis mais representativas em cada *cluster*, mesmo diante do grande volume de dados processados.

Uma premissa fundamental para a aplicação do *SelectKBest* é a definição do número de atributos a serem selecionados (*k*). Com base na abordagem proposta por [Franco et al. 2020], foram escolhidos os 10 atributos mais relevantes para cada grupo de candidatos em cada edição. Esse número foi considerado adequado para os objetivos da análise, pois garante um equilíbrio entre a representatividade das características e a simplicidade do modelo.

Vale ressaltar que a quantidade de atributos selecionados neste estudo é superior à média observada na literatura relacionada, conforme evidenciado na Tabela 1. Isso abre

precedentes para a exploração de novas features e potenciais refinamentos na análise. Embora a inclusão de um número maior de atributos tenha sido considerada, essa abordagem aumentaria o tempo de processamento e poderia elevar a redundância entre os atributos selecionados para diferentes *clusters*, comprometendo a interpretabilidade dos resultados e a eficácia da segmentação.

4. Análise e Discussão de Resultados

A fim de sistematizar a análise e viabilizar a comparação dos resultados ao longo dos anos, adotou-se um critério de classificação para os *clusters*, fundamentado na média geral das notas dos candidatos. Observe que os *clusters* foram construídos sem observar as notas dos candidatos. Assim, os *clusters* foram classificados de "A" a "E", em ordem crescente de desempenho. A classe "A" referenciará sempre os grupos com as menores médias gerais, seguida pela classe "B", que compreende aqueles com o segundo menor desempenho, até a classe "E", que corresponde aos *clusters* cujos candidatos possuem as maiores médias em cada ano analisado.

A Tabela 5 sintetiza essa classificação ao longo dos anos, apresentando a média das notas dos candidatos para cada edição do exame e para cada classe definida. Adicionalmente, a tabela inclui o número de candidatos em cada classe, fornecendo uma visão complementar à Tabela 4, ao evidenciar a segmentação do conjunto de dados analisado.

Ano do Exame	**Média da Pontuação no Exame**	**Classe do *Cluster***	**Média da Pontuação no *Cluster***	**Número de Candidatos no *Cluster***
2019	522,62	A	481,69	823.212
		B	486,01	544.437
		C	527,35	1.121.353
		D	528,58	563.775
		E	591,91	649.132
2020	526,60	A	469,97	471.799
		B	501,84	652.313
		C	524,99	458.071
		D	548,08	593.876
		E	606,68	385.245
2021	535,54	A	485,72	576.558
		B	528,16	473.155
		C	529,95	430.964
		D	562,41	418.162
		E	604,49	339.267
2022	543,48	A	496,47	366.881
		B	510,72	605.064
		C	546,73	456.878
		D	563,98	579.864
		E	614,02	336.136
2023	540,61	A	494,29	383.750
		B	507,33	827.801
		C	542,10	499.204
		D	569,10	613.796
		E	617,21	353.713

Table 5. Ordenação e classificação dos *clusters* por meio das médias

A partir dos dados apresentados, observa-se a progressão de notas em três frentes

principais. A princípio, é possível verificar o aumento na média geral de pontuação do exame, indicando um melhor preparo dos candidatos a cada edição. Adiante, os diferentes tons de verde evidenciam a progressão das notas de cada grupo que designou a atribuição da sua classe. Por fim, outro aspecto que contribui para a noção geral de crescimento das médias das notas é a comparação entre classes semelhantes de diferentes edições. Por meio desta observação, verifica-se que, apesar das oscilações no curto prazo, todos os grupos apresentaram um aumento expressivo em suas médias individuais ao longo do período. O grupo "D", em particular, destacou-se com um aumento de 7,67% em sua média de notas no intervalo analisado.

A Figura 2 complementa a Tabela 5 ao ilustrar a evolução do número de candidatos em cada uma das classes entre 2019 e 2023.

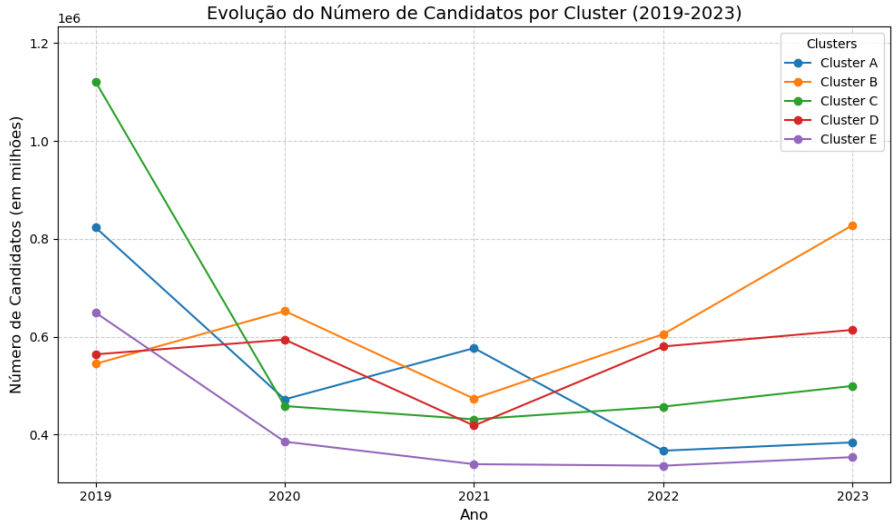

Figure 2. Evolução do Número de Candidatos por Classe (2019-2023)

O gráfico evidencia a dinâmica de alternância na concentração de candidatos nos *clusters* no período. A princípio, observa-se que as classes “A” e “B” alternaram-se como mais populosas em quatro dos cinco anos analisados. De modo semelhante, a classe “E” esteve consolidada como a menos representativa desde 2020. Além disso, outro aspecto de destaque é a diminuição geral no número de inscritos para o intervalo compreendido entre 2020 e 2022. Essa diminuição pode ser associada, em parte, aos impactos da pandemia de *COVID-19*, que, além de gerar receio quanto à participação presencial no exame, também resultou em dificuldades quanto à transição para o ensino remoto. Como reportado pela *British Broadcasting Corporation* (BBC) em [Idoeta 2021], a crise sanitária gerou uma grave desconexão dos jovens com a escola e com o ENEM.

As análises específicas dos resultados obtidos a partir dos dados do ENEM entre 2019 e 2023 estão estruturadas em três frentes principais: (i) análise descritiva das características selecionadas para os *clusters* obtidos, (ii) avaliação das tendências temporais observáveis ao longo do período analisado e, por fim, (iii) a caracterização dos grupos. Cada uma dessas etapas foi aplicada sistematicamente a todas as classes de "A" a "E".

4.1. Classes de *Clusters* "A" e "B": Menores Desempenhos

Conforme mencionado, os grupos denominados "A" e "B" são aqueles associados aos *clusters* com as menores pontuações médias em cada edição. Optou-se por realizar a análise conjunta deles devido à similaridade das *features* selecionadas para estas classes.

A Figura 3 relaciona o significado de cada variável selecionada pelo método *SelectKBest* com a resposta à qual ela faz referência. Além disso, a tabela indica em qual grupo cada variável foi selecionada: se para “A”, se para “B” ou se para ambos.

Feature Selecionada	Pergunta do Questionário	Resposta do Candidato	Grupo A	Grupo B
Q003_A	(...) grupo que contempla a ocupação mais próxima do seu pai ou do homem responsável por você.	Lavrador, agricultor sem empregados, bóia fria, criador de animais (...), lenhador, seringueiro, extrativista.	SIM	NÃO
Q004_A	(...) grupo que contempla a ocupação mais próxima da sua mãe ou da mulher responsável por você.	Lavradora, agricultora sem empregados, bóia fria, criadora de animais (...), lenhadora, seringueira, extrativista.	SIM	NÃO
Q006_B	Qual é a renda mensal de sua família?	Até 1 salário mínimo	SIM	SIM
Q010_A	Na sua residência tem carro?	Não	SIM	SIM
Q010_B		Sim, um	NÃO	SIM
Q014_A	Na sua residência tem máquina de lavar roupa?	Não	SIM	SIM
Q014_B		Sim, uma	SIM	SIM
Q016_A	Na sua residência tem forno micro-ondas?	Não	SIM	SIM
Q016_B		Sim, um	SIM	SIM
Q018_B	Na sua residência tem aspirador de pó?	Sim	SIM	SIM
Q022_B	Na sua residência tem telefone celular?	Sim, um	SIM	NÃO
Q024_A	Na sua residência tem computador?	Não	SIM	SIM
Q025_A	Na sua residência tem acesso à Internet?	Não	SIM	NÃO
TP_ST_CONCLUSAO_1	Situação de conclusão do Ensino Médio	Já concluí o Ensino Médio	SIM	SIM
TP_ST_CONCLUSAO_2		Estou cursando e concluirei o Ensino Médio neste ano	SIM	SIM
TP_ESCOLA_1	Tipo de escola do Ensino Médio	Não Respondeu	SIM	SIM
TP_ESCOLA_2		Pública	SIM	SIM
TP_FAIXA_ETARIA_1	Faixa etária	Menor de 17 anos	NÃO	SIM
TP_FAIXA_ETARIA_2		17 anos	NÃO	SIM
TP_FAIXA_ETARIA_3		18 anos	NÃO	SIM

Figure 3. Características selecionadas para as classes "A" e "B"

Observa-se que aspectos relacionados à renda familiar, infraestrutura doméstica e inclusão no ensino médio foram indicados como relevantes para a caracterização dos candidatos de ambos os grupos. Entretanto, presume-se que a classe "A" possua condições socioeconômicas mais deficitárias, uma vez que as respostas afirmativas ao acesso limitado à internet, smartphones e veículos estão associadas aos grupos da classificação "A". O grupo "B", por seu turno, se diferencia pela indicação de posse de automóveis e pela diversidade de respostas relacionadas à ocupação dos pais.

Para seguimento com a segunda frente da análise, deve-se considerar o modo no qual as características detalhadas pela Figura 3 aparecem distribuídas ao longo dos anos. Esta análise permite identificar padrões longitudinais na segmentação dos candidatos. Tendo isto em vista, a Tabela 6 abaixo complementa as informações abordadas para o grupo "A" ao passo em que apresenta a recorrência de seleção das *features* nos anos analisados.

Ano do Exame	***Features* Selecionadas Pelo *SelectKBest***									
2019	Q006_B	Q014_A	Q016_A	Q024_A	Q016_B	Q022_B	Q014_B	Q025_A	Q003_A	Q004_A
2020	Q006_B	Q014_A	Q016_A	Q024_A	Q016_B	Q022_B	Q014_B	Q025_A	Q003_A	Q004_A
2021	Q006_B	Q014_A	Q016_A	Q024_A	Q016_B	Q010_A	Q014_B	Q025_A	Q003_A	Q004_A
2022	Q006_B	Q014_A	Q016_A	Q024_A	Q018_B	Q010_A	CONCLUSAO_1	CONCLUSAO_2	ESCOLA_1	ESCOLA_2
2023	Q006_B	Q014_A	Q016_A	Q024_A	Q016_B	Q010_A	CONCLUSAO_1	CONCLUSAO_2	ESCOLA_1	ESCOLA_2

Table 6. *Features* selecionadas pelo *SelectKBest* para o grupo "A"[6]

Observa-se que os *clusters* do grupo "A" apresentaram uma distribuição relativamente estável das variáveis ao longo dos anos, com mudanças mais expressivas a partir de 2022 - momento que coincide com a retomada da ascensão do número de inscritos, conforme ilustrado pela Figura 2. A partir de 2022, observa-se a introdução de indicadores ligados à situação escolar dos candidatos em detrimento de variáveis relacionadas à ocupação profissional dos pais, acesso à internet e infraestrutura doméstica. A emergência dessas variáveis nos últimos anos pode indicar uma mudança no perfil dos participantes dos *clusters*, possivelmente relacionada a transformações socioeconômicas dos inscritos ou a modificações na formulação do exame.

Outro aspecto de mudança nos *clusters* envolve atributo `Q025` ("Na sua residência tem acesso à internet?") resposta "A" ("Não"). Observa-se que esta era uma característica determinante e consistente dentro do grupo até 2021, mas deixou de estar presente em 2022 e 2023, indicando uma nova tendência. Associa-se a isto a possibilidade de que políticas de inclusão digital realizadas no período da pandemia COVID-19 tenham surtido um efeito significativo na composição deste grupo. Conforme noticiado em [Oliveira 2021], apesar dessas políticas terem passado por diversos desafios até o início da sua vigência, a pandemia acelerou um processo de conectividade nas escolas que começou em 1997.

Apesar das recentes mudanças observadas, as *features* mais recorrentemente selecionadas para discriminação destes grupos foram: `Q006_B`, `Q014_A`, `Q016_A`, `Q024_A`, `Q016_B`, `Q010_A`, `Q014_B`, `Q025_A`, `Q003_A` e `Q004_A`, estando presentes em pelo menos três dos cinco anos analisados. Estas características são especial-

[6] *Labels* das *Features* adaptados para visualização na tabela. Remoção dos sufixos e alteração de "Faixa Etária" para "Idade"

mente relevantes porque a alta recorrência com que elas foram selecionadas sugere que o perfil dos candidatos dos *clusters* desta classe tende a ser mais bem representado por elas.

Dado o contexto, entende-se que os *clusters* do grupo "A" poderiam ser descritos, sob linhas gerais, por meio das seguintes afirmações:

- Os candidatos destes grupos apresentam baixa renda familiar, predominantemente proveniente de atividades sazonais (ou manuais), conforme indicado pela ocupação dos pais;
- Os grupos demonstram acesso limitado a carros, eletrodomésticos e outros aparatos que poderiam otimizar o tempo de estudo dos candidatos;
- Os candidatos destes grupos frequentemente relatam falta de acesso à internet e aos dispositivos necessários para inclusão digital;
- A partir de 2022, observa-se uma mudança no perfil do grupo, com a remoção de características anteriormente relevantes, dando lugar a variáveis relacionadas ao Ensino Médio. Isso pode indicar uma leve melhora no acesso à internet e na infraestrutura doméstica.

No caso do grupo "B", a distribuição das *features* pode ser observada na Tabela 7.

Ano do Exame	***Features* Selecionadas Pelo SelectKBest**									
2019	CONCLUSAO_1	CONCLUSAO_2	ESCOLA_1	ESCOLA_2	Q014_A	Q006_B	Q016_B	Q016_A	Q014_B	Q024_A
2020	CONCLUSAO_1	CONCLUSAO_2	ESCOLA_1	ESCOLA_2	Q014_A	IDADE_2	Q010_B	Q016_A	Q010_A	Q018_B
2021	CONCLUSAO_1	CONCLUSAO_2	ESCOLA_1	ESCOLA_2	Q014_A	IDADE_2	Q010_B	IDADE_3	IDADE_3	IDADE_1
2022	CONCLUSAO_1	CONCLUSAO_2	ESCOLA_1	ESCOLA_2	Q014_A	Q006_B	Q016_B	Q016_A	Q010_A	Q018_B
2023	CONCLUSAO_1	CONCLUSAO_2	ESCOLA_1	ESCOLA_2	Q014_A	Q006_B	Q016_B	Q016_A	Q010_A	Q018_B

Table 7. *Features* selecionadas pelo *SelectKBest* para o grupo "B" [6]

Para este grupo, observa-se um padrão de recorrência das características ainda mais estável do que no grupo "A". Enquanto o primeiro apresentava mudanças expressivas a partir de 2022, o grupo "B" manteve um perfil socioeconômico relativamente homogêneo em todo o período analisado.

Como mencionado, as classes "A" e "B" se assemelham por serem conjuntos intersectantes, mas diferem na especificidade dos indicadores socioeconômicos. O conjunto "A" destaca-se por atributos como "atividade profissional dos pais sazonal" e "ausência de internet em casa", enquanto o conjunto "B" apresenta uma composição mais ampla, incluindo variáveis como faixa etária, situação de conclusão do ensino médio e tipo administrativo da escola frequentada.

Um dos pontos de distinção entre os grupos pode ser observado pela presença da *feature* `Q010_B` (indicação de um carro na família) em 2020 e 2021, sugerindo que o grupo "B" apresenta um menor nível de vulnerabilidade social que o grupo "A". Por fim, destaca-se que a estabilidade das *features* selecionadas, especialmente nos anos mais recentes, sugere que o perfil dos candidatos com médias baixas se mantém o mesmo desde 2019.

Semelhante ao grupo "A", o grupo "B" poderia ser descrito pelas seguintes afirmações:

- O grupo é majoritariamente composto por candidatos de escolas públicas (ou não informado) que estão próximos à data de conclusão do Ensino Médio, possuindo pouca diversidade na faixa etária;

- Embora a renda familiar ainda seja baixa, observa-se uma menor dependência de atividades sazonais ou informais, sugerindo uma fonte de renda mais estável.
- A presença de bens como carros e eletrodomésticos sugere um nível socioeconômico relativamente mais estável do que no grupo "A";
- O grupo apresenta melhor acesso à internet e dispositivos tecnológicos, reduzindo barreiras de inclusão digital;

4.2. Classe de *Clusters* "C": Desempenho intermediário

Dando continuidade à análise dos perfis de candidatos, esta seção explora os *clusters* da classe "C", que se diferenciam dos grupos "A" e "B" por apresentarem desempenho intermediário nas provas em suas respectivas edições. Frente a isto, a análise deste grupo também busca identificar se a progressão de média das notas está correlacionada à seleção de variáveis que indiquem um perfil socioeconômico menos fragilizado.

A Figura 4 relaciona o significado de cada uma das variáveis selecionadas para os *clusters* deste grupo com as respostas às quais elas fazem referência.

Feature Selecionada	Pergunta do Questionário	Resposta do Candidato
Q004_B	(...) grupo que contempla a ocupação mais próxima da sua mãe ou da mulher responsável por você.	Grupo 2: Diarista, babá, cozinheira, motorista particular, etc.
Q006_B	Qual é a renda mensal de sua família?	Até 1 salário mínimo
Q006_C		Entre 1 e 2 salários mínimos
Q008_B	Na sua residência tem banheiro?	Sim, um
Q008_C		Sim, dois
Q009_D	Na sua residência tem quartos para dormir?	Sim, três
Q010_A	Na sua residência tem carro?	Não
Q010_B		Sim, um
Q014_A	Na sua residência tem máquina de lavar roupa?	Não
Q018_B	Na sua residência tem aspirador de pó?	Sim
Q024_A	Na sua residência tem computador?	Não
Q024_B		Sim, um
Q025_A	Na sua residência tem acesso à Internet?	Não
TP_ST_CONCLUSAO_1	Situação de conclusão do Ensino Médio	Já concluí o Ensino Médio
TP_ST_CONCLUSAO_2		Estou cursando e concluirei o Ensino Médio neste ano
TP_ST_CONCLUSAO_3		Estou cursando e concluirei o Ensino Médio após este ano
TP_ESCOLA_1	Tipo de escola do Ensino Médio	Não Respondeu
TP_ESCOLA_2		Pública
TP_FAIXA_ETARIA_2	Faixa etária	17 anos
TP_FAIXA_ETARIA_3		18 anos

Figure 4. Características selecionadas para a classe "C"

Enquanto o grupo "B" exibia diferenças sutis em relação ao grupo "A", o grupo "C" apresenta um distanciamento mais evidente da classe "A". Esse afastamento pode ser observado pela mudança nos padrões de ocupação das mães dos candidatos, pela presença de respostas que indicam uma renda familiar mais elevada e por uma infraestrutura doméstica relativamente mais robusta.

Contudo, esse grupo ainda revela um contingente significativo de candidatos com indicadores socioeconômicos deficitários, visto que algumas características selecionadas evidenciam acesso limitado à internet, a computadores, a máquinas de lavar e outros bens domésticos. Dessa forma, o grupo "C" parece representar um ponto de equilíbrio entre vulnerabilidade e acesso a recursos, diferenciando-se dos grupos "A" e "B" por exibir traços de melhoria econômica sem uma consolidação plena das condições favoráveis observadas em grupos de desempenho superior.

A Tabela 8 descreve a evolução dos *clusters* da classe "C" ao longo dos anos.

Ano do Exame	Features Selecionadas Pelo SelectKBest									
2019	ESCOLA_1	ESCOLA_2	IDADE_2	CONCLUSAO_1	Q024_A	Q024_B	Q014_A	Q025_A	Q006_B	CONCLUSAO_2
2020	ESCOLA_1	ESCOLA_2	IDADE_2	CONCLUSAO_1	Q025_A	Q010_B	Q014_A	CONCLUSAO_3	IDADE_3	CONCLUSAO_2
2021	ESCOLA_1	ESCOLA_2	IDADE_2	CONCLUSAO_1	Q004_B	Q018_B	Q008_B	Q009_D	Q006_C	CONCLUSAO_2
2022	ESCOLA_1	ESCOLA_2	IDADE_2	CONCLUSAO_1	Q010_A	Q010_B	Q014_A	CONCLUSAO_3	IDADE_3	CONCLUSAO_2
2023	ESCOLA_1	ESCOLA_2	Q024_A	Q024_B	Q010_A	Q010_B	Q014_A	Q008_C	Q006_B	CONCLUSAO_2

Table 8. *Features* selecionadas pelo *SelectKBest* para o grupo "C" [6]

A evolução dos *clusters* do grupo "C" ao longo dos anos revela que, semelhante ao grupo "A", a distribuição das características se mostrou sensível a mudanças temporais. Embora algumas características socioeconômicas tenham permanecido constantes, como a predominância de candidatos oriundos da rede pública de ensino e a distribuição etária entre 17 e 18 anos, observa-se uma variação nos indicadores de infraestrutura doméstica.

Entre os anos de 2019 e 2021, destaca-se a ausência de computador e internet para uma parte dos candidatos, enquanto nos anos seguintes há um aumento na presença desses recursos. A partir de 2022, a recorrência de variáveis como "`Q010_B`" (indicação de pelo menos um carro na família) e "`Q024_B`" (presença de computador) pode sugerir uma leve melhora no acesso a bens materiais. Essa transição também pode estar relacionada a fatores como programas governamentais de inclusão digital e mudanças no perfil de renda das famílias ao longo do período analisado.

Por meio das características descritas pela Figura 4 e pela Tabela 8, entende-se que os *clusters* do grupo "C" podem ser descritos, sob linhas gerais, por meio das seguintes afirmações:

- A maioria dos candidatos estudou em escolas públicas e já concluiu ou está prestes a concluir o Ensino Médio, sendo predominantemente estudantes entre 17 e 18 anos;
- Os candidatos deste grupo apresentam renda familiar variando entre um e dois salários mínimos, com responsáveis frequentemente empregados em atividades de prestação de serviços como diaristas e motoristas particulares;
- O acesso a bens de consumo se demonstra instável dentro do grupo, com uma parcela significativa possuindo itens como aspirador de pó, mas ainda havendo um percentual considerável sem máquina de lavar e carro próprio;

- O acesso à internet e a computadores ainda é uma limitação para parte dos candidatos, com uma parcela significativa reportando ausência de computador em casa e dificuldades de acesso à internet.

Dessa forma, o grupo "C" distingue-se dos anteriores por apresentar um equilíbrio entre restrições e acessos. Embora ainda possuam desafios estruturais, os candidatos desse grupo demonstram ter mais acesso a recursos que poderiam garantir mais tempo, qualidade de vida e de estudo.

4.3. Classes de *Clusters* "D" e "E": Maiores Desempenhos

Nesta seção, analisam-se conjuntamente os grupos "D" e "E", assim como foi feito anteriormente para os grupos "A" e "B" em 4.1. Essa abordagem se justifica pelo fato de ambas as classes agregarem os *clusters* com maior desempenho no ENEM em todas as edições analisadas, além de uma inversão atípica de tendência observada em 2023 que se relaciona com *clusters* destes dois grupos. Neste ano, um *cluster* com características socioeconômicas compatíveis com a classe "D" obteve uma média de pontuação superior a um *cluster* alinhado à classe "E", contrariando os padrões históricos. A análise conjunta permite compreender de maneira mais intuitiva as diferenças e semelhanças entre esses dois grupos e avaliar com maior nitidez a mudança de padrão observado neste período.

Como mencionado, a edição do ano 2023 contou com um evento atípico de inversão de pontuação entre os *clusters*. Isto porque um *cluster* altamente compatível com a classe "D" apresentou uma pontuação mais elevada do que o *cluster* compatível com a classe "E" deste mesmo ano. Desta forma, as Tabelas 9 e 10 adiante ilustram, de antemão, a distribuição das *features* selecionadas para estas classes no período analisado. Por meio delas, pode-se observar o fenômeno descrito com a abrupta ruptura do padrão seguido até 2023.

Ano do Exame	Features Selecionadas Pelo SelectKBest									
2019	CONCLUSAO_2	ESCOLA_1	ESCOLA_2	Q010_B	Q014_A	CONCLUSAO_3	CONCLUSAO_1	Q014_B	IDADE_2	IDADE_3
2020	CONCLUSAO_2	ESCOLA_1	ESCOLA_2	Q010_B	Q014_A	Q016_A	Q016_B	Q014_B	Q024_A	Q024_B
2021	CONCLUSAO_2	ESCOLA_1	ESCOLA_2	Q010_B	Q014_A	CONCLUSAO_3	IDADE_1	Q008_B	Q008_C	Q010_A
2022	CONCLUSAO_2	ESCOLA_1	ESCOLA_2	Q010_B	Q014_A	CONCLUSAO_3	CONCLUSAO_1	Q006_B	Q024_A	Q024_B
2023	Q001_G	Q003_E	Q004_E	Q006_L	Q008_D	Q008_E	Q010_C	Q018_B	Q019_E	Q021_B

Table 9. *Features* selecionadas pelo *SelectKBest* para o grupo "D" [6]

Ano do Exame	Features Selecionadas Pelo SelectKBest									
2019	Q023_B	Q003_E	Q024_C	Q006_L	Q008_D	Q008_E	Q010_C	Q018_B	Q019_D	Q021_B
2020	Q001_G	Q003_E	Q004_E	Q006_L	Q008_D	Q008_E	Q010_C	Q018_B	Q019_E	Q021_B
2021	Q001_G	Q003_E	Q004_E	Q006_L	Q008_D	Q008_E	Q010_C	Q018_B	Q019_E	Q021_B
2022	Q001_G	Q003_E	Q004_E	Q006_L	Q008_D	Q008_E	Q010_C	Q018_B	Q019_E	Q021_B
2023	CONCLUSAO_2	ESCOLA_1	ESCOLA_2	Q010_B	Q014_A	CONCLUSAO_3	CONCLUSAO_1	Q014_B	IDADE_2	IDADE_3

Table 10. *Features* selecionadas pelo *SelectKBest* para o grupo "E" [6]

A distribuição das *features* ao longo dos anos demonstra uma tendência bem definida dentro dos grupos até 2023. Enquanto os *clusters* da classe "E" caracterizavam-se fortemente pela presença de variáveis relacionadas à escolaridade e à ocupação dos responsáveis (`Q001_G`, `Q003_E`, `Q004_E`), renda familiar elevada (`Q006_L`), e presença de diversos indicadores de infraestrutura doméstica, como múltiplos banheiros (`Q008_D`,

Q008_E) e outros bens de consumo, a classe "D" era caracterizada pelo envolvimento das características de situação de conclusão do ensino médio (TP_ST_CONCLUSAO_1, (TP_ST_CONCLUSAO_2, (TP_ST_CONCLUSAO_3), tipo de escola (TP_ESCOLA_1, TP_ESCOLA_2) e alguns outros indicadores acerca da infraestrutura doméstica.

Vale destacar que a classe "E", em especial, apresentou os *clusters* mais homogêneos dentro do período analisado, mantendo as mesmas características desde 2020 (com variações pontuais em 2019). No entanto, em 2023, um *cluster* que dispunha dos mesmos padrões observados desde 2020 para a classe "E", obteve uma pontuação média menor do que um *cluster* tipicamente caracterizado como pertencente à classe "D". As Tabelas 11 e 12 ilustram como a inversão desses *clusters* poderia ter mantido a coerência dos padrões anteriores.

Ano do Exame	Features Selecionadas Pelo SelectKBest									
2019	CONCLUSAO_2	ESCOLA_1	ESCOLA_2	Q010_B	Q014_A	CONCLUSAO_3	CONCLUSAO_1	Q014_B	IDADE_2	IDADE_3
2020	CONCLUSAO_2	ESCOLA_1	ESCOLA_2	Q010_B	Q014_A	Q016_A	Q016_B	Q014_B	Q024_A	Q024_B
2021	CONCLUSAO_2	ESCOLA_1	ESCOLA_2	Q010_B	Q014_A	CONCLUSAO_3	CONCLUSAO_1	Q008_B	Q008_C	Q010_A
2022	CONCLUSAO_2	ESCOLA_1	ESCOLA_2	Q010_B	Q014_A	CONCLUSAO_3	CONCLUSAO_1	Q006_B	Q024_A	Q024_B
2023 *	CONCLUSAO_2	ESCOLA_1	ESCOLA_2	Q010_B	Q014_A	CONCLUSAO_3	CONCLUSAO_1	Q014_B	IDADE_2	IDADE_3

Table 11. *Cluster* 2023 invertido com a classe "E" [6]

Ano do Exame	Features Selecionadas Pelo SelectKBest									
2019	Q023_B	Q003_E	Q024_C	Q006_L	Q008_D	Q008_E	Q010_C	Q018_B	Q019_D	Q021_B
2020	Q001_G	Q003_E	Q004_E	Q006_L	Q008_D	Q008_E	Q010_C	Q018_B	Q019_E	Q021_B
2021	Q001_G	Q003_E	Q004_E	Q006_L	Q008_D	Q008_E	Q010_C	Q018_B	Q019_E	Q021_B
2022	Q001_G	Q003_E	Q004_E	Q006_L	Q008_D	Q008_E	Q010_C	Q018_B	Q019_E	Q021_B
2023 *	Q001_G	Q003_E	Q004_E	Q006_L	Q008_D	Q008_E	Q010_C	Q018_B	Q019_E	Q021_B

Table 12. *Cluster* 2023 invertido com a classe "D" [6]

Esse evento singular representa um ponto de inflexão na análise, pois se contrapõe ao padrão historicamente observado entre o desempenho dos candidatos e seus indicadores socioeconômicos. O fenômeno sugere que, dadas as devidas condições, candidatos com perfis socioeconômicos menos favorecidos podem atingir desempenhos equivalentes ou superiores aos de candidatos com melhores condições. No entanto, para compreender os fatores que efetivamente promoveram essa inversão, fazem-se necessárias análises complementares envolvendo múltiplas áreas de conhecimento. Levanta-se a hipótese de que algumas possíveis explicações para o fenômeno seriam: mudanças no perfil dos candidatos após a pandemia, impactos de políticas públicas que podem ter afetado a distribuição das oportunidades educacionais, necessidade de mais *clusters* para realizar uma melhor segmentação entre perfis que são distintos, conforme discutido na Seção 3.4.2, dentre outros.

Seguindo com o propósito investigativo da análise descritiva das características selecionadas para os *clusters* e com caracterização dos grupos. As Figuras 5 e 6 relacionam as variáveis apresentadas anteriormente com seus respectivos significados.

Feature Selecionada	Pergunta do Questionário	Resposta do Candidato
Q006_B	Qual é a renda mensal de sua família?	Até um salário minimo
Q008_B	Na sua residência tem banheiro?	Sim, um.
Q008_C		Sim, dois.
Q010_A	Na sua residência tem carro?	Não.
Q010_B		Sim, um.
Q014_A	Na sua residência tem máquina de lavar roupa?	Não.
Q014_B		Sim, uma.
Q016_A	Na sua residência tem forno micro-ondas?	Não.
Q016_B		Sim, um.
Q024_A	Na sua residência tem computador?	Não.
Q024_B		Sim, um.
TP_ST_CONCLUSAO_1	Situação de conclusão do Ensino Médio	Já concluí o Ensino Médio
TP_ST_CONCLUSAO_2		Estou cursando e concluirei o Ensino Médio neste ano
TP_ST_CONCLUSAO_3		Estou cursando e concluirei o Ensino Médio após este ano
TP_ESCOLA_1	Tipo de escola do Ensino Médio	Não Respondeu
TP_ESCOLA_2		Pública
TP_FAIXA_ETARIA_1	Faixa etária	Menor de 17 anos
TP_FAIXA_ETARIA_2		17 anos
TP_FAIXA_ETARIA_3		18 anos

Figure 5. Características tipicamente selecionadas para a classe "D" até 2022 e para a classe "E" em 2023

Feature Selecionada	Pergunta do Questionário	Resposta do Candidato
Q001_G	Até que série seu pai, ou o homem responsável por você, estudou?	Completou a Pós-graduação.
Q003_E	Ocupação do seu pai ou do homem responsável por você	Médico, engenheiro, dentista, psicólogo, economista, advogado, etc.
Q004_E	Ocupação da sua mãe ou da mulher responsável por você	Médica, engenheira, dentista, psicóloga, economista, advogada, etc.
Q006_L	Qual é a renda mensal de sua família?	Nove ou mais salários mínimos
Q008_D	Na sua residência tem banheiro?	Sim, três.
Q008_E		Sim, quatro ou mais.
Q010_C	Na sua residência tem carro?	Sim, dois.
Q018_B	Na sua residência tem aspirador de pó?	Sim.
Q019_D	Na sua residência tem televisão em cores?	Sim, três.
Q019_E		Sim, quatro ou mais.
Q021_B	Na sua residência tem TV por assinatura?	Sim.
Q023_B	Na sua residência tem telefone fixo?	Sim.
Q024_C	Na sua residência tem computador?	Sim, dois.

Figure 6. Características tipicamente selecionadas para a classe "E" até 2022 e para a classe "D" em 2023

A análise das classes de clusters "D" e "E" revela um grupo de candidatos com os perfis socioeconômicos mais favorecidos dentro dos conjuntos de dados. Em particular, o grupo "E" apresenta características que o diferenciam significativamente dos demais, indicando um nível de acesso a recursos, renda e infraestrutura consideravelmente superior. O grupo "D", por sua vez, representa uma evolução típica do grupo "C", apresentando melhorias mais discretas em seus indicadores socioeconômicos (desconsiderando-se o evento atípico de 2023).

Dado o contexto, considera-se que os *clusters* da classe "D" poderiam ser descritos, em linhas gerais, por meio das seguintes afirmações:

- A maioria dos candidatos estudou em escolas públicas e já concluiu ou está prestes a concluir o Ensino Médio, sendo predominantemente estudantes entre 17 e 18 anos;
- Os candidatos deste grupo ainda não apresentam rendas familiares muito elevadas, mas observa-se uma constância maior em respostas afirmativas a posse de bens de consumo;
- O acesso a computadores ainda é uma instabilidade dentro do grupo.

De modo geral, a classe "E" diverge quase integralmente dos padrões observados nos grupamentos "A" e "B". Esta classe pode ser descrita, de modo generalista, por meio das seguintes afirmações que se relacionam aos aspectos mais marcantes deste grupo:

- Indicação de um alto nível de escolaridade paterna;
- Os candidatos deste grupo apresentam renda familiar superior a nove salários mínimos, com responsáveis frequentemente empregados em atividades profissionais que demandam de graduação no Ensino Superior;
- Os candidatos deste grupo demonstram possuir um amplo acesso a bens de consumo, de tecnologia e de infraestrutura domestica;
- Historicamente, este grupo apresenta frequentemente as maiores médias de pontuação entre os demais grupos analisados e as características que definem esta classe são bastante padronizadas. Em 2023 houve o primeiro caso identificado na análise onde o *cluster* compatível com as características da classe "E" não apresentou a maior pontuação média do exame;

4.4. Análises Complementares

Além das principais investigações realizadas neste estudo, algumas análises complementares foram conduzidas com o intuito de fornecer uma visão mais abrangente sobre as classes de *clusters*. Nesta seção, exploramos quantitativamente aspectos adicionais que, embora não estejam diretamente relacionados ao objetivo central da pesquisa, contribuem para uma caracterização mais detalhada dos grupos formados. Essas análises permitem identificar discrepâncias relevantes entre as classes, examinar variáveis ainda não abordadas e aprofundar a compreensão de fenômenos já discutidos na literatura.

Conforme ilustrado pela Tabela 1, algumas *features* são recorrentemente selecionadas na literatura para evidenciar a correlação entre características socioeconômicas e o desempenho dos candidatos. O atributo cor e raça, por exemplo, é amplamente discutido em estudos como [Silva et al. 2020], [Carmo et al. 2021], [Maia et al. 2021], [Banni et al. 2021] e [Silva et al. 2014]. No presente contexto, observa-se que este atributo não foi selecionado pelo método *SelectKBest*, possivelmente devido à limitação de k em 10. No entanto, ainda que não tenha sido incluído entre as variáveis mais relevantes segundo esse critério, as respostas autodeclaradas pelos candidatos revelam distribuições discrepantes entre as classes. A Figura 7 ilustra a distribuição dessas respostas para as classes "A" e "E" ao longo do período analisado.

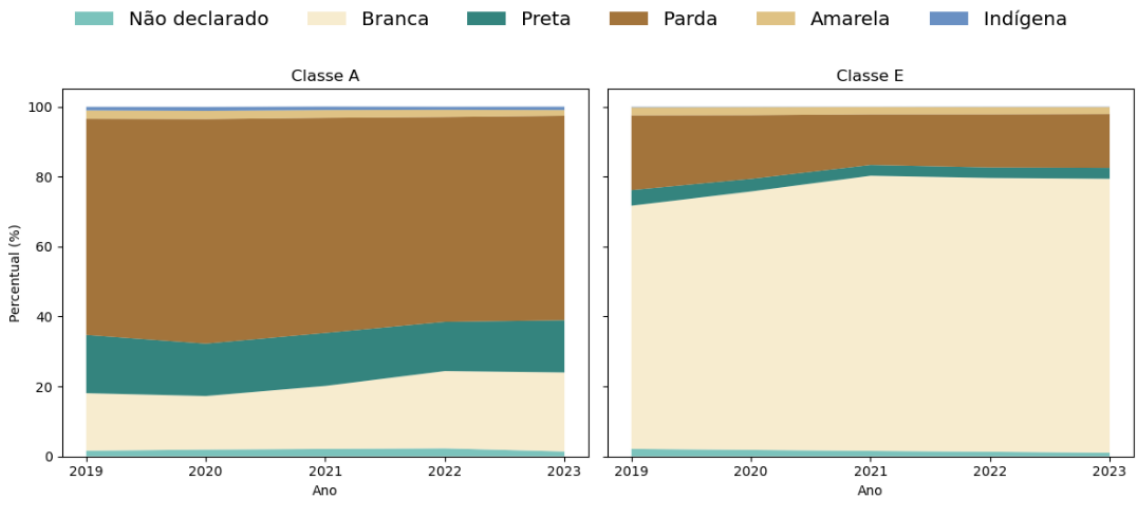

Figure 7. Composição das classes "A" e "E" por autodeclaração de cor/raça (2019-2023)

Os gráficos evidenciam que, para os *clusters* mais discrepantes em termos de características e pontuações, há também uma diferenciação significativa nas autodeclarações de cor e raça. Enquanto os *clusters* da classe "A" (menores pontuações) são predominantemente compostos por candidatos pardos e com outra parcela expressiva composta por candidatos pretos, os *clusters* da classe "E" (maiores pontuações) são formados majoritariamente por candidatos brancos ao longo do período analisado, com uma representatividade bastante discreta de candidatos pretos. Essa observação complementa os achados da literatura ao reafirmar a mesma tendência para o ano de 2023, período que ainda não havia sido explorado nos estudos anteriores.

Além da análise de cor e raça, a transformação dos dados descrita na Seção 3.3 permitiu investigar a distribuição geográfica dos candidatos nos *clusters*. Esse aspecto não foi abordado nos trabalhos relacionados e traz uma nova perspectiva sobre a composição das classes "A" e "E". A Figura 8 apresenta a distribuição da origem dos candidatos dessas classes em relação às macrorregiões brasileiras.

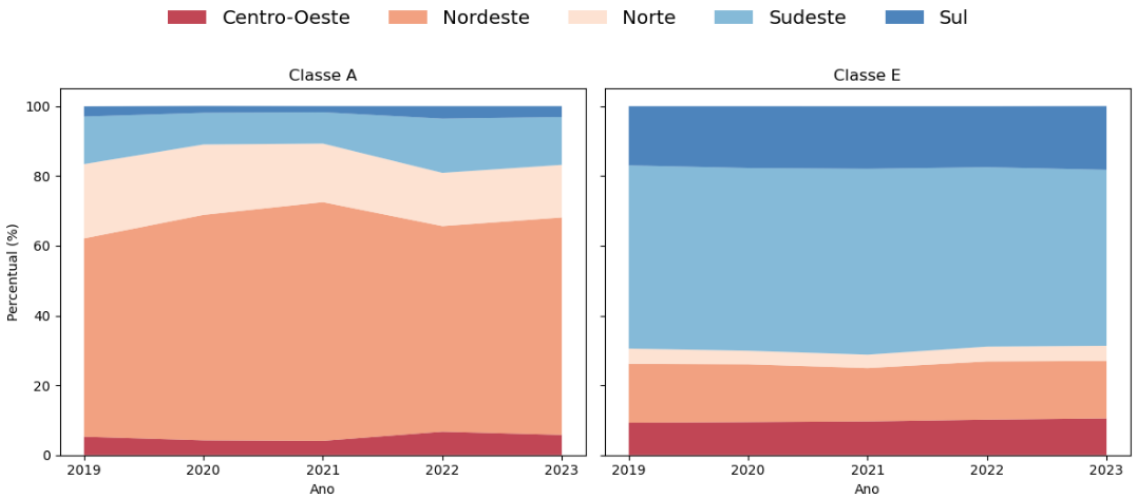

Figure 8. Composição das classes "A" e "E" por macrorregiões brasileiras (2019-2023)

Observa-se que a classe "A" é majoritariamente composta por candidatos oriundos das regiões Nordeste, Norte e Sudeste, enquanto a classe "E" exibe uma composição mais estável, concentrando candidatos das regiões Sudeste, Sul e Nordeste. Um aspecto relevante identificado na Figura 8 é a alternância entre as regiões Norte e Sul nas classes analisadas. A macrorregião Norte aparece de forma expressiva na classe "A" e de maneira discreta na classe "E", enquanto a macrorregião Sul exibe o comportamento oposto, com baixa representatividade na classe "A" e predominância na classe "E".

Outro fator frequentemente associado ao desempenho no ENEM é a escolaridade dos pais. Estudos como [Silva et al. 2014] e [Silva et al. 2020] sugerem que níveis mais elevados de escolaridade dos pais estão positivamente correlacionados com maiores pontuações no exame. Nesse contexto, as Figuras 9 e 10 ilustram a distribuição das classes "A" e "E" em relação à escolaridade paterna e materna, respectivamente.

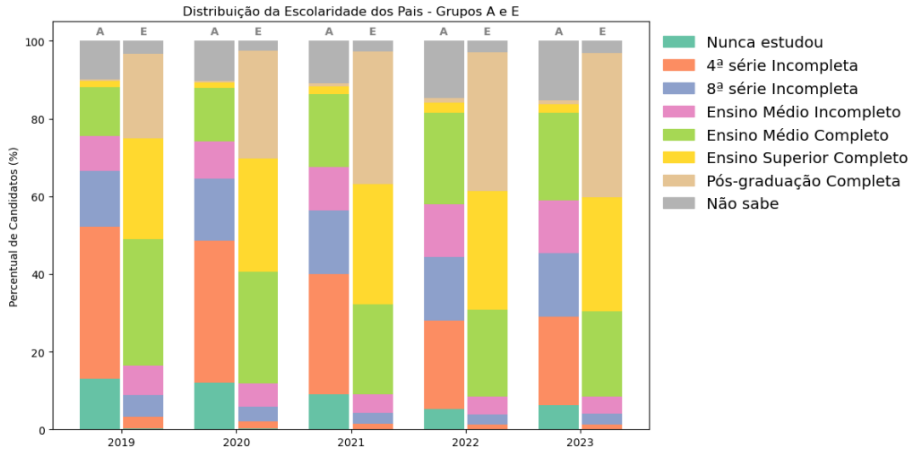

Figure 9. Composição das classes "A" e "E" quanto a escolaridade paterna

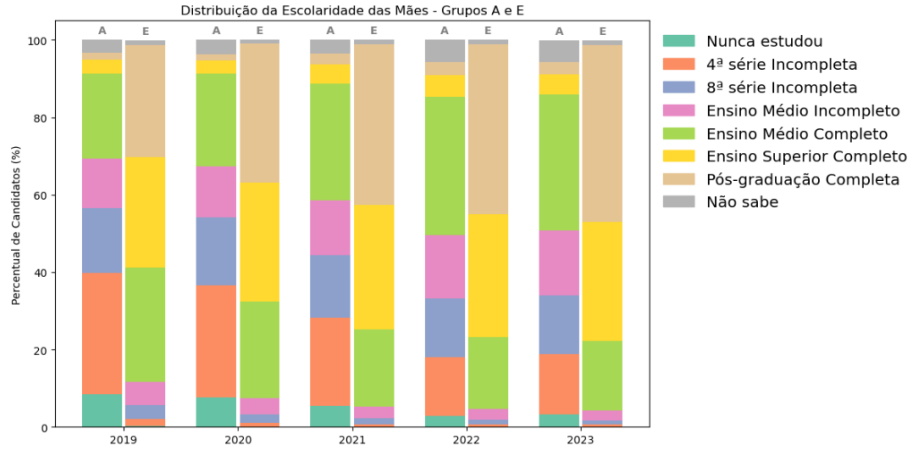

Figure 10. Composição das classes "A" e "E" quanto a escolaridade materna

A análise do período revela uma elevação progressiva nos níveis de escolaridade tanto dos pais quanto das mães dos candidatos. Na classe "E", por exemplo, a proporção de candidatos cuja mãe declarou ter concluído a pós-graduação aumentou de 29% em 2019 para mais de 45% em 2023. Esse aumento pode estar relacionado às próprias oportunidades de ingresso no Ensino Superior proporcionadas pelo ENEM, conforme discutido na Seção 1. Além disso, pode estar associado à melhora gradual na pontuação média dos candidatos, conforme evidenciado na Tabela 5.

A comparação entre as classes "A" e "E" quanto à escolaridade dos pais também revela novas diferenças expressivas entre estes grupos. Enquanto na classe "A" há um percentual significativo de candidatos cujos pais estudaram apenas até a quarta série do ensino fundamental, na classe "E" essa resposta possui magnitude irrelevante. O fenômeno oposto é observado para o nível de escolaridade mais elevado, onde a conclusão da pós-graduação é significativamente mais presente na classe "E" e praticamente inexistente na classe "A". Outro aspecto notável é a maior frequência de respostas indicando desconhecimento da escolaridade paterna em comparação à escolaridade materna, sugerindo que a ausência do pai no convívio familiar pode ser mais comum do que a ausência materna.

Em síntese, as análises complementares realizadas permitiram a identificação de outras tendências temporais não identificadas anteriormente além de uma exploração mais aprofundada das diferenças entre as classes "A" e "E", evidenciando a influência de fatores socioeconômicos, regionais e educacionais na composição dos *clusters*, ampliando o escopo das discussões já presentes na literatura.

5. Considerações Finais e Trabalhos Futuros

O presente estudo teve como objetivo central caracterizar os *clusters* de candidatos do ENEM ao longo dos anos de 2019 a 2023, com foco na identificação de padrões socioeconômicos e demográficos que contribuam para o entendimento das tendências evolutivas dos perfis típicos dos participantes.

A clusterização permitiu identificar cinco grupos bem definidos quanto aos seus aspectos socioeconômicos. Embora o agrupamento não tenha considerado as notas dos candidatos como critério de segmentação, utilizamos as médias de pontuação dos participantes para estabelecer comparações entre os grupos. As análises evidenciaram uma correlação positiva entre o aprimoramento dos indicadores socioeconômicos e o desempenho no exame.

Aspectos como renda familiar, origem da renda, infraestrutura doméstica e acesso à tecnologia foram recorrentes na caracterização dos grupos, tanto neste estudo quanto em pesquisas relacionadas. Além disso, análises quantitativas complementares demonstraram a influência de fatores como região de origem, cor/raça e escolaridade dos pais na composição dos grupos com maiores e menores desempenhos no exame.

As comparações entre os clusters "A" e "E" corroboraram evidências da literatura sobre a influência do contexto socioeconômico na distribuição das notas, ao mesmo tempo em que exploraram novas perspectivas, como a composição regional dos grupos. Dentre as principais contribuições deste trabalho, destaca-se a abordagem investigativa voltada para a evolução histórica dos perfis identificados.

Os resultados indicam que, apesar dos avanços no acesso ao ensino superior e da melhoria nos índices de escolarização da população, a relação entre condições socioeconômicas favoráveis e melhores resultados no exame persiste ao longo do período analisado. Foram observados padrões estruturais de desigualdade, evidenciados pela distribuição dos candidatos entre os diferentes grupos de pontuação, com poucas exceções pontuais.

Dada a riqueza dos dados disponibilizados pelo INEP, há diversas oportunidades de exploração para trabalhos futuros. Primeiramente, recomenda-se a continuidade da análise dos *clusters* nos anos subsequentes, a fim de monitorar as tendências e possíveis mudanças nos padrões observados.

Além disso, sugere-se a realização de estudos interdisciplinares voltados à avaliação da eficácia das políticas públicas educacionais vigentes. A continuidade desse tipo de investigação pode subsidiar o desenvolvimento de estratégias mais inclusivas e efetivas, ampliando a compreensão das condições que impactam o desempenho acadêmico dos estudantes brasileiros.

Outro aspecto relevante para pesquisas futuras é a exploração de subgrupos menos representativos dentro do contexto global, como candidatos em faixas etárias elevadas,

indígenas e estudantes de áreas rurais, cujo acesso ao ensino superior ainda apresenta desafios significativos.

Por fim, destaca-se a necessidade de aprimoramento das metodologias de clusterização, visando o aumento da coesão interna e da separação entre os grupos, o que pode contribuir para análises ainda mais precisas sobre os perfis dos candidatos.

A continuidade dessas investigações é essencial para fornecer subsídios atualizados aos debates sobre equidade educacional, auxiliando na formulação de iniciativas voltadas à redução das disparidades observadas.

Referências

[Banni et al. 2021] Banni, M., Oliveira, M., and Bernardini, F. (2021). Uma análise experimental usando mineração de dados educacionais sobre os dados do enem para identificação de causas do desempenho dos estudantes. In *Anais do II Workshop sobre as Implicações da Computação na Sociedade*, pages 57–66, Porto Alegre, RS, Brasil. SBC.

[Carmo et al. 2021] Carmo, Vinicios do, R., Felipe Heckler, W., and Varella de Carvalho, J. (2021). Uma análise do desempenho dos estudantes do rio grande do sul no enem 2019. *Revista Novas Tecnologias na Educação*, 18(2):378–387.

[Dutra et al. 2023] Dutra, J. F., Firmino Júnior, J. B., and Fernandes, D. Y. d. S. (2023). Fatores que podem interferir no desempenho de estudantes no enem: uma revisão sistemática da literatura. *Revista Brasileira de Informática na Educação*, 31:323–351.

[Fayyad et al. 1996] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37.

[Franco et al. 2020] Franco, J., Miranda, F., Stiegler, D., Dantas, F., Brancher, J., and Nogueira, T. (2020). Usando mineração de dados para identificar fatores mais importantes do enem dos Últimos 22 anos. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 1112–1121, Porto Alegre, RS, Brasil. SBC.

[Hair et al. 2009] Hair, J., Black, W., Babin, B., Anderson, R., and Tatham, R. (2009). *Análise multivariada de dados - 6ed*. Bookman.

[Idoeta 2021] Idoeta, P. A. (2021). Enem: o que explica menor número de inscritos na prova em mais de uma década. Disponível em: `https://www.bbc.com/portuguese/brasil-58021267`. Acesso em: 21 fev 2025.

[INEP 2023] INEP (2023). Exame nacional do ensino médio (enem). Disponível em: `https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem`. Acesso em: 12 set 2023.

[Li et al. 2005] Li, F., Yang, Y., and Xing, E. (2005). From lasso regression to feature vector machine. *Advances in neural information processing systems*, 18.

[Lima et al. 2019] Lima, P. d. S. N., Ambrósio, A. P. L., Ferreira, D. J., and Brancher, J. D. (2019). Análise de dados do enade e enem: uma revisão sistemática da literatura. *Avaliação: Revista da Avaliação da Educação Superior (Campinas)*, 24(1):89–107.

[Maia et al. 2021] Maia, M. M., de Andrade, L. H. F., and Fernandes, S. (2021). K-means na análise de características socioeconômicas de candidatos ao ensino superior. In *Anais do Encontro de Computação do Oeste Potiguar ECOP/UFERSA*.

[MEC 2014] MEC (2014). Exame bate recorde de inscritos e chega a 9,5 milhões de candidatos na edição 2014. Disponível em: http://portal.mec.gov.br/ultimas-noticias/418-enem-946573306/20456-exame-bate-recorde-de-inscritos-e-chega-a-95-milhoes-de-candidatos-na-edicao-2014. Acesso em: 20 set 2024.

[MEC 2021] MEC (2021). Dicionário de Dados. Disponível em: https://dadosabertos.mec.gov.br/sisu/item/133-dicionario-de-dados. Acesso em: 07 mai 2024.

[MEC 2023] MEC (2023). 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: 12 set 2023.

[Oliveira 2021] Oliveira, E. (2021). Governo promulga lei que garante internet gratuita a alunos e professores de escola pública. Disponível em: https://g1.globo.com/educacao/noticia/2021/06/11/governo-promulga-lei-que-garante-internet-gratuita-a-alunos-e-professores-de-escola-publica.ghtml. Acesso em: 09 mar 2025.

[scikit-learn developers 2025] scikit-learn developers (2025). KMeans. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html#kmeans. Acesso em: 06 mar 2025.

[Silva et al. 2014] Silva, L. A., Morino, A. H., and Sato, T. M. C. (2014). Prática de mineração de dados no exame nacional do ensino médio. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, pages 651–660.

[Silva et al. 2020] Silva, V., Moreno, L., Gonçalves, L., Soares, S., and Júnior, R. S. (2020). Identificação de desigualdades sociais a partir do desempenho dos alunos do ensino médio no enem 2019 utilizando mineração de dados. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 72–81, Porto Alegre, RS, Brasil. SBC.