Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Bacharelado em Ciência da Computação

Análise das Disparidades Socioeconômicas e Geográficas

Impacto no desempenho dos Candidatos do ENEM

André Luiz dos Reis

JUIZ DE FORA
SETEMBRO, 2024

Análise das Disparidades Socioeconômicas e Geográficas
Impacto no desempenho dos Candidatos do ENEM

André Luiz dos Reis

Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Departamento de Ciência da Computação

Bacharelado em Ciência da Computação

Orientador: Victor Ströele de Andrade Menezes

JUIZ DE FORA

SETEMBRO, 2024

ANÁLISE DAS DISPARIDADES SOCIOECONÔMICAS E GEOGRÁFICAS

Impacto no desempenho dos Candidatos do ENEM

André Luiz dos Reis

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Victor Ströele de Andrade Menezes
Doutor em Engenharia de Sistemas e Computação pela UFRJ

Igor de Oliveira Knop
Doutor em Modelagem Computacional pela UFJF

Regina Maria Maciel Braga
Doutora em Engenharia de Sistemas e Computação pela UFRJ

JUIZ DE FORA

25 DE SETEMBRO, 2024

À minha família, que, por terem desertado de seus sonhos e instrução em prol da sobrevivência, possibilitaram a realização de parte deles através de mim.

Resumo

Atualmente, o Exame Nacional do Ensino Médio (ENEM) é o principal meio de acesso à educação superior no país, bem como auxilia na avaliação da educação nacional, fornecendo uma base de dados rica para ser explorada. Em particular, o estudo e pesquisa responsável pela aplicação de técnicas diversas com o intuito de obter conhecimento baseado em dados de ambientes educacionais é nomeado por Mineração de Dados Educacionais. O presente trabalho disserta, em particular, sobre os dados de aplicações do ENEM que ocorreram entre 2015 e 2022, buscando identificar tendências de impactos em diferentes variáveis, como a presença dos candidatos no exame e desempenho, causados por fatores socioeconômicos, demográficos ou quaisquer outros que possam ser observados. Para isso, é necessária a análise da base dados, a normalização dos dados, a construção de um banco de dados capaz de gerir os dados e a conexão com ferramentas de visualização e análise de dados, bem como a construção dos gráficos desejados. O trabalho ainda contribui com a análise das visualizações geradas, bem como questionamentos sobre justificativas para os resultados encontrados.

Palavras-chave: Análise de Dados Educacionais, Desigualdades socioeconômicas, ENEM.

Abstract

Currently, the National Secondary Education Examination (ENEM) is the main means of accessing higher education in the country, as well as helping to evaluate national education, providing a rich database to be explored. In particular, the study and research responsible for applying various techniques with the aim of obtaining knowledge based on data from educational environments is called Educational Data Mining. The present work study, in particular, data from applications of the National Secondary Education Examination (ENEM) that took place between 2015 and 2022, seeking to identify trends in impacts on different variables, such as the presence of candidates, performance, caused by socioeconomic and demographic factors. or any others that may be observed. To do this, it will be necessary to analyze the database, normalize the data, build a database capable of managing the data and connect with data visualization and analysis tools, as well as the construction of the desired graphics. The work also contributed to the analysis of the constructed visualizations, as well as questions about justifications for the results found.

Keywords: Educational Data Analysis, Socioeconomic inequalities, ENEM.

Agradecimentos

À minha família, meu alicerce, da qual nem a distância pode afastar o amor, carinho, presença, e preocupações diárias.

Aos meus amigos, inclusive aqueles que se fazem presentes mesmo estando distantes fisicamente, por todo companheirismo, apoio e presença. Nossas risadas, aventuras e incentivos marcam nossa história e contribuíram para superar mais esse desafio.

À minha psicológica que me ensinou a ter a leveza da vida como uma escolha, mesmo frente as tempestades da vida.

Ao professor Victor pela orientação, paciência e companheirismo que foram pilares para possibilitar a realização desse trabalho.

Aos funcionários dos diversos departamentos da UFJF, que durante esses anos, contribuíram de algum modo para que essa conquista se tornasse realidade.

"Quando a educação não é libertadora, o sonho do oprimido é ser o opressor".

Paulo Freire

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

CPTM	Colégio Tiradentes da Polícia Militar
DCC	Departamento de Ciência da Computação
Enade	Exame Nacional de Desempenho de Estudantes
Enem	Exame Nacional do Ensino Médio
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
MEC	Ministério da Educação
PROUNI	Programa Universidade para Todos
SISU	Sistema de Seleção Unificada
UFJF	Universidade Federal de Juiz de Fora

1 Introdução

Com a 4ª Revolução Industrial e o aumento significativo do acesso à informação, a análise de dados tornou-se essencial na era digital. Com o aumento exponencial no número de dispositivos conectados e a internet acessível a um número cada vez maior de pessoas, o volume de dados gerados diariamente alcançou níveis sem precedentes (KIM; LEE, 2021).

Nesse sentido, a análise de dados vem como ferramenta para a obtenção de informação a partir dos dados disponíveis. Enquanto dados são registros brutos, como texto ou números que podem ser coletados de diversas fontes, a informação é o resultado da análise e interpretação dos dados, possibilitando assim a extração de conhecimentos, correlações e/ou *insights* não vistos claramente apenas com a base de dados (SETZER, 2001).

A análise de dados pode ser aplicada em diversos setores, tais como indústria, trânsito (CYGANCZUK; PINTO; BASTOS, 2023), saúde (Maximiano, Caio Fernandes Chaves, 2023) e educação. Como resultado, permite as organizações tomar decisões baseado em evidências, melhorando seus processos e estratégias com mais eficiência.

Nos últimos anos, o processo de digitalização das informações avançou consideravelmente em todos os setores, e em particular na educação. Os registros e documentos acadêmicos, bem como a implantação da educação à distância e o apoio de plataformas digitais no processo de ensino e aprendizagem são exemplos que marcam a transformação digital em diversos setores educacionais (SHENKOYA; KIM, 2023).

Neste caminho, com a disponibilização de dados públicos pelos órgãos competentes, e a capacidade de coleta de dados internos, diversas oportunidades para avaliação dos dados florescem pelas instituições. A descoberta de conhecimento pela análise de dados começa a ser uma oportunidade de contribuição para melhorias constante nos processos de ensino e aprendizagem em várias frentes e nível, macro ou em contextos particulares de uma instituição, ou até mesmo de uma turma ou disciplina isolada (MASCHIO et al., 2018).

Atualmente, o Exame Nacional do Ensino Médio (Enem) é o principal meio de acesso à educação superior no país, podendo ser considerado como medida de padrão de

desempenho acadêmico dos estudantes. Desde seu início, o Enem é reflexo dos desafios como desigualdades sociais e socioeconômicos no país.

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), responsável pela realização do exame, disponibiliza uma base de dados contendo informações socioeconômicas e resultados dos participantes. Contudo, nem sempre a disponibilização de dados implica na análise ou que seja possível extrair alguma informação de padrões. Na maioria das vezes, é preciso que seja aplicado um processo análise de dados responsável pelo tratamento e transformação dos dados em conhecimento, auxiliando na identificação de padrões, tendências e outras correlações desejadas.

São diversos os desafios para uma educação pública de qualidade (FRANÇA; ALVES; DUARTE, 2022). A análise de dados educacionais pode ajudar a evidenciar a falta de equidade no acesso à educação, identificando os fatores que contribuem para essa desigualdade, bem como destacar as falhas nas políticas públicas.

Nesta direção, esse trabalho visa analisar os microdados da realização do Enem dos anos de 2015 a 2022, buscando identificar tendências de impactos em diferentes variáveis, como a presença dos candidatos nos dias do exame e seus desempenhos, analisando essas variáveis em relação aos fatores socioeconômicos, demográficos ou quaisquer outros que possam ser observados.

Nesse sentido, a pesquisa desenvolvida pode avançar além de uma crítica a realidade educacional nacional, mas também fornecer conhecimento para a tomada de decisões futuras para a construção de políticas educacionais mais resilientes, direcionadas e eficazes, contribuindo assim para a melhoria na equidade do acesso à educação.

1.1 Objetivo

Esse trabalho possui como objetivo geral analisar os dados do Enem no período de 2015 a 2022, visando identificar e avaliar tendências em diversas variáveis associadas aos dados do exame por meio de uma abordagem quantitativa e qualitativa.

Como objetivos específicos, esse trabalho propõe:

- Implementação de *scripts* que possibilitem a extração, tratamento e transformação

da base de dados disponibilizados;

- Definição de relatórios e gráficos com visões ao longo dos anos;
- Análise de tendência e variação ao longo dos anos, considerando:
  - Número de inscritos, taxas de presença e ausência;
  - Desempenho geral (médias) dos participantes, bem como por componente curricular: Linguagens, Matemática, Ciências da Natureza, Ciências Humanas e Redação;
  - Impactos das divergências demográficas no desempenho dos candidatos;
  - Impactos das divergências socioeconômicas no desempenho dos candidatos;
  - Explorar as diferenças entre treineiros (estudantes que ainda não estão no último ano do ensino médio) e não treineiros em relação aos pontos anteriores.

1.2 Organização

Esse trabalho apresenta a sua fundamentação teórica no Capítulo 2, apresentando os principais conceitos necessários para o entendimento do tema. Já o Capítulo 3 apresenta uma revisão da literatura, evidenciando trabalhos relacionados ao tema, destacando os principais pontos e considerações sobre cada estudo. Na sequência, o Capítulo 4 apresenta a estruturação e execução do trabalho, em termos técnicos e práticos, tendo os resultados obtidos discutidos no Capítulo 5. Por fim, a conclusão desse estudo e as possibilidades de trabalhos futuros é destacado no Capítulo 6.

2 Fundamentação Teórica

O Exame Nacional do Ensino Médio ocupa um papel fundamental na educação brasileira, estando presente na vida de vários candidatos que desejam acessar o ensino superior no pais. Desde sua criação, o processo sofreu diversos ajustes a fim de se adequar as novas realidades, bem como ao seu papel na sociedade.

Este capítulo apresenta uma visão da história do Enem, explorando sua origem, objetivos e as inúmeras mudanças que ocorreram com o passar dos anos. Junto a isso, também destaca a importância do exame para o acesso ao ensino superior, abordando os programas como o Sistema de Seleção Unificada (SISU) e o Programa Universidade para Todos (PROUNI). Por fim, são apresentados conceitos de análise e mineração de dados, com destaque para a aplicação na área educacional.

2.1 Enem

2.1.1 Origem e objetivos do exame

O Exame Nacional do Ensino Médio foi instituído pela Portaria nº 438/1998 do Ministério da Educação (MEC) com o objetivo inicial de avaliar o desempenho dos alunos na conclusão do ensino médio (BRASIL. Ministério da Educação, 1998). Com o passar dos anos, o Enem se mostrou um grande indicador da eficiência das políticas de educação, fornecendo dados e diagnósticos com seus resultados, além de ser um caráter seletivo para o acesso ao ensino superior.

Com o passar dos anos, algumas mudanças ocorreram na estrutura do exame:

- **Quantidade de questões:** Inicialmente o exame era composto por 63 questões, em 2009 foi ampliado para 180 questões, dividido em quatro áreas: Ciências Humanas, Ciências da Natureza, Matemática, e Linguagens e Códigos, além da redação;
- **Dias de aplicação:** Até 2016 o exame era aplicado em 2 dias (sábado e domingo) de um único final de semana. De 2017 em diante, foi dividido em dois domingos

consecutivos para reduzir a exaustão dos candiados;

- **Conteúdo Programático:** O conteúdo programático do Enem é baseado nas diretrizes de ensino e busca aplicar uma avaliação interdisciplinar do conhecimento;

- **Prova:** as provas do exame buscam avaliar diversas competências, tais como interpretação, análise crítica e construção de resolução de problemas, não focando exclusivamente na memorização de conteúdos.

2.1.2 Competências da Redação

A prova do Enem exige a produção de um texto dissertativo argumentativo que versa sobre alguma temática de ordem social, científica, cultural ou política. O texto entregue é avaliado por, ao menos, dois corretores de maneira independente.

Cada avaliador avalia o desempenho em cinco competências distintas, atribuindo um valor entre 0 e 200 pontos para cada uma. A nota final do participante é a média aritmética dos dois avaliadores. Já a nota final da redação é o somatório da nota em cada competência, podendo chegar em até 1000 pontos (BRASIL. Ministério da Educação, 2018). A seguir, podemos ver a descrição de cada competência avaliada:

- **Competência 1:** Demonstrar domínio da modalidade escrita formal da língua portuguesa;

- **Competência 2:** Compreender a proposta de redação e aplicar conceitos das várias áreas de conhecimento para desenvolver o tema, dentro dos limites estruturais do texto dissertativo-argumentativo em prosa;

- **Competência 3:** Selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista;

- **Competência 4:** Demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação;

- **Competência 5:** Elaborar proposta de intervenção para o problema abordado que respeite os direitos humanos.

2.1.3 Importância para o acesso ao ensino superior

A partir de 2009, com a criação do Sistema de Seleção Unificada, o exame começa a ser tornar a maior porta de entrada para o ensino superior no Brasil, substituindo ou complementando os vestibulares tradicionais de cada universidade. A utilização de um sistema centralizado de nota e avaliação colabora para a democratização ao acesso ao Ensino superior (NETO et al., 2014).

"O Sistema de Seleção Unificada (Sisu) é um sistema informatizado gerenciado pelo MEC, que seleciona candidatos a vagas em cursos de graduação ofertadas pelas instituições públicas de educação superior" (BRASIL. Governo Federal, 2023). Toda a classificação é baseado nos resultados obtidos no Enem do ano anterior a edição do SISU.

2.2 Análise de dados

Podemos definir a análise de dados como o processo responsável por trabalhar em cima de uma ou mais bases de dados realizando a limpeza e ajustes nas informações brutas, aplicando as transformações necessárias, bem como a modelagem de dados necessárias para a descoberta de informações úteis para a tomada de decisão em um contexto específico (ISLAM, 2020).

Já a mineração de dados é o processo responsável pela descoberta de conhecimento útil a partir de grandes quantidades de dados presente em banco de dados ou outros repositórios de dados, podendo ser dados advindo da análise de dados (KIRKPATRICK, 2019).

2.2.1 Mineração de Dados Educacionais

Em particular, o campo de pesquisa responsável por aplicar técnicas de mineração de dados no contexto da educação, explorando dados de ambientes educacionais, como sistemas de gerenciamento de aprendizagem, plataformas online e dados administrativos, para gerar melhorias na educação é denominado mineração de dados educacionais (EDM - do inglês *Educational data mining*) (ROMERO; VENTURA, 2013). A EDM pode ser aplicada em diferentes abordagens para a descoberta de informações, como método esta-

tísticos, aprendizado de máquina e inteligência artificial.

Entre as técnicas utilizadas, pode-se destacar as análises por agrupamento (*clustering*), possibilitando a identificação de grupos com características semelhantes, a análise de predição, que permite prever o desempenho dos alunos baseado em dados históricos, e a busca por regras de associação, que releva relações no formato de relação se-então (KOEDINGER et al., 2015).

Essas abordagens permites que a EDM seja útil para a descoberta de novos conhecimentos (RODRIGUES et al., 2014), bem como relevar *insights* sobre a aprendizagem, motivação e demais fenômenos do processo educacional, contribuindo para a melhoria do resultados (ROMERO; VENTURA, 2013).

As técnicas associadas ao processo de EDM tem evoluídos constantemente nos últimos anos, bem como sua presença na literatura (ROMERO; VENTURA, 2020), e, quando bem aplicada, por auxiliar na construção de sistemas educacionais e políticas educacionais (ISLAM, 2021).

2.3 Considerações parciais

A trajetória do Enem é marcada por constantes evolução e adaptações frente as políticas de educação e as necessidades dos estudantes brasileiros. Seu atual e fundamental papel como mecanismo de acesso à educação superior confirmam sua importância no cenário educacional nacional. Nesse conjunção, a análise de dados e a mineração de dados educacionais proporcionam técnicas para identificação de melhorias no exame e nas políticas públicas associadas.

Dessa maneira, compreender os dados das aplicações do Enem e as técnicas de análises de dados é ponto fundamental para a descoberta de falhas nas políticas públicas vigentes e a melhoria contínua no sistema educacional brasileiro, melhorando a equidade no acesso ao ensino superior.

3 Trabalhos Relacionados

Esse capítulo dispõe-se a apresentar e discorrer sobre pesquisa e estudos que também investigam os dados do Enem para análises educacionais, sejam elas por uma visão nacional ou não. Junto a isso, também será apresentado uma breve visão sobre uma revisão sistemática da literatura sobre o tema. Para essa busca, foram utilizados buscadores como Google Acadêmico[1] e *Schielo*[2].

Para cada trabalho apresentado nesse capítulo, serão destacados metodologia, resultados, conclusões. Os diversos trabalhos apresentados utilizam-se desde abordagens mais comuns, como estática descritiva, bem como técnicas mais avançadas de análises, tais como modelos estáticos e técnicas de mineração de dados. Além disso, serão abordados os principais destaques positivos e de lacunas identificados em cada estudo, a fim de compreender melhor os desafios associados a esse tema de pesquisa.

3.1 Análise de dados do Enade e Enem: uma revisão sistemática da literatura

O estudo de Lima et al. (2019) apresenta uma Revisão Sistemática da Literatura que espera identificar os objetivos e principais tipos de análises que são realizadas utilizando os dados do Enem e foram publicados entre 2005 e 2016. A base foi levantada pela plataforma *Google Scholar*, e após a busca, análise e leitura, foram mantidos 54 trabalhos.

A revisão sistemática mostrou que as analises realizadas são, em sua maioria, focadas nas estatísticas descritivas e dados socioeconômicos dos participantes, com a motivação dos trabalhos focada na melhoria da qualidade da educação. Além disso, o estudo destaca que existem mais pesquisas relacionados ao ENADE do que ao Enem, e pouquíssimas estudam as duas bases. Por fim, também foi pontuado a necessidade de realização de estudo utilizando outras abordagens de análise, como a mineração de dados.

[1] https://scholar.google.com/
[2] https://www.scielo.br/

Como ponto positivo, destaca-se a amplitude do período de pesquisa, bem como a identificação da falta de estudos que utilizavam análises mais sofisticadas, como estatística inferencial e mineração de dados. Por outro lado, o estudo apresenta uma limitação no intervalo de busca até 2016, o que pode excluir trabalhos mais recentes que utilizem as técnicas de análises não encontradas pelo trabalho.

3.2 Desigualdades Educacionais e na População Brasileira Pré-Universitária: Uma Visão a Partir da Análise de Dados do Enem

O artigo apresentado por Travitzki, Ferrão e Couto (2016) aborda, à partir de dos dados do Enem entre 2009 e 2012, por meio de uma análise intergeracional das desigualdades educacionais e sua relação com o desempenho e sua relação com os desempenhos dos candidatos e seus atributos sociodemográficos. A análise foi realizada utilizando o coeficiente de Gini, a curva de Lorenz e a modelagem multinível, concluindo que houve uma atenuação das desigualdades, embora o desempenho dos candiados ainda seja sensível as variáveis variáveis socioespaciais e raciais. O estudo revelou que há mais variação em níveis intramunicipais do que intermunicipais, pontuando a importância de escolas e municípios como a base para a políticas educacionais e meio para reduzir desigualdades socioeconômicas.

O estudo se destaca pela apresentação de uma análise detalhada, utilizando técnicas robustas de modelagem multinível, oferecendo uma visão detalhada das variações da desigualdade considerando diferentes contexto. Além disso, conseguiu identificar potenciais escolas e municípios como pontos de destaques para servirem como base para as ações de redução à desigualdade. Por outro lado, o estudo também indica a necessidade de mais estudos para determinação dos efeitos das escolas, infraestrutura disponível e investimento e sua correlação com os dados descobertos.

3.3 Análise de dados históricos do Enem entre 2015 à 2019

A pesquisa realizado por Nakazone e Bortolotti (2021) visa a realização de um estudo comparativo e qualitativo de dados do Enem entre 2015 à 2019, analisando os resultados dos alunos de diferentes regiões: a cidade de Mococa, a região de Ribeirão Preto e a média nacional. Foi utilizado a linguagem de programação *Python* e suas bibliotecas *Pandas*, *Matplotlib* e *Seaborn*, para analisar as notas obtida em cada critério de avaliação da redação do Enem, bem como a avaliação de desempenho segregado por sexo e por renda familiar.

O estudo mostra que a região de Ribeirão preto teve um desempenho superior à média nacional em todas as áreas avaliadas, destacando que a competência de elaboração de intervenção da Redação apresente a menor médias comparadas a demais competências. Também foi pontuado o aumento significativo na ausências dos inscritos no segundo dia de prova após a mudança do exame para o formato de dois domingo consecutivos.

Como destaque positivos, destaca-se a análise e comparação regional do estudo, principalmente nesse caso de destaque positivo, indicando possíveis boas práticas educacionais que podem servir de modelos para políticas públicas de outras regiões, bem como auxiliar na identificação na disparidade de outras regiões, e identificação de áreas que precisam de maior intervenção educacional. Entretanto, o estudo carece de análises ou destaque para a motivação dos resultados positivos encontrados na região. Além disso, a análise nos resultados das notas obtidas na redação pode sofrer variações causada pelos diferentes temas que são abordados em cada realização do exame, contexto não abordado no trabalho.

3.4 Análise de dados: um estudo do perfil dos participantes do Enem 2019

A monografia apresentada por Thiago de Oliveira Souza (2021) destaca os impactos diversas variáveis, como idade, sexo, região geográfica, tipo de escola frequentada, raça

autodeclarada, escolaridade dos responsáveis e renda familiar no desempenho dos candiados do Enem em 2019. O estudo utilizou-se de técnicas de Ciência de Dados e Estatística, e contou com o apoio da linguagem de programação *Python*, bem como bibliotecas específicas como *Pandas* e *Matplotlib* para o desenvolvimento do trabalho e a identificação das influências na nota dos candidatos.

O estudo mostrou que tanto a idade, como o sexo dos participantes não foram fatores determinísticos para o desempenho dos estudantes. Por outro lado, o estudo destaca que a Região Centro-Oeste, região com menor densidade populacional do país, possui os melhores resultados, enquanto as regiões nordeste e norte estão com o penúltimo e último lugar, respectivamente, em todos os conteúdos avaliados pelo exame. Junto a isso, o estudo também destacou que escolas particulares possuem resultados superiores em todas as competências do exame, com destaque para a redação. Por fim, destaca-se que a diferença em relação a raça autodeclarada dos participantes existes, mas não é uma diferença tão expressiva.

O presente estudo destaca-se por sua metodologia rigorosa, utilizando-se de ferramentas computacionais para tratamento de dados, construção de gráficos e aplicação de técnicas de *Machine Learning*, além da utilização de uma abordagem estatística concreta, que permite uma análise detalhada e confiável dos dados, possibilitando visualizações claras dos dados obtidos através de gráficos e tabelas, junto com uma ampla quantidade de variáveis analisadas. Em contrapartida, o estudo se limita a avaliação de um ano específico, o que limitada a conclusão de tendências a longo prazo.

3.5 Desempenho dos estudantes ao final do ensino médio: Mensurando a influência direta e indireta da educação dos pais

O trabalho apresentado em Feijó, FRANÇA e PINHO (2022) busca investigar a relação entre o desempenho no Enem e o nível educacional familiar, analisando quatro principais variáveis: tamanho de família; renda; infraestrutura domiciliar; e escola. Além disso, o

estudo examina se há impactos causados pela diferença de gênero dos inscritos.

O presente estudo utiliza-se de um modelo econométrico (regressão linear) com quatro grupos de efeitos fixos, podendo ser analisados juntos ou separados. Mesmo realizando o controle simultâneo por rendas, tamanho de família, estrutura domiciliar e escola. O estudo mostrou que a escolaridade dos pais exerce influência significativa sobre a nota dos estudantes.

Como principais resultados, notou-se que filhos de pais com nível superior possuem desempenho superior no Enem, em destaque na prova de Redação, com pontuação com diferença superior a 50%. Nos demais componentes houve variação, mas menor: Linguagens e Códigos (13,75%), Matemática (21,72%), Ciências da Natureza (15,21%) e Ciências Humanas (16,57%). Controlando-se os efeitos fixos, a influência do ensino paternal apresentou uma queda, mas ainda se mostrou relevante para as análises.

Como pontos positivos, o estudo utiliza-se de uma abordagem de análise robusta, com técnicas suficientes para isolar efeitos diretos e indiretos. Junto a isso, destaca em momentos possíveis explicações para os efeitos observados. No entanto, por tratar-se de correlações, os resultados não permitem concluir relações de causa e efeito.

3.6 Desempenho de estudantes de Minas Gerais no Enem considerando variáveis socioeconômicas

O trabalho apresentado em Lima e Brighenti (2023) visa à análise do desempenho dos alunos do Estado de Minas no Enem de 2019 e identificação de padrão de desempenho associados as distintas variáveis, tais como dependência administrativa das escolas (pública ou privada), cor,raça e sexo dos alunos, renda familiar e escolaridade do pai e da mãe. O estudo considerou uma base restrita dos dados disponibilizados, considerando:

- Alunos que apenas eram concluintes do ensino médio em 2019;
- Retirado alunos com nota zerada ou inexistente na base em duas ou mais pontuações, em geral, alunos que faltaram 1 ou 2 dias ou zeraram a redação;
- Segmentação dos dados pela dependência administrativa escolar;

- Subdivisão dos alunos das escolas estaduais em duas categorias: alunos pertencentes ao Colégio Tiradentes da Polícia Militar (CPTM) e as demais escolas estaduais.

À partir das notas dos participantes, o estudo apresentou uma curva normal para cada dependência administrativa presentes, mas, ao considerar as notas, as curvas foram deslocadas para a esquerda, indicativo que esses dados devem ser analisados separadamente.

Como principais resultados, os autores destacam que 75% dos estudantes mineiros são de escolas públicas, seguido por estudantes de escolas particulares. Além disso, foi notado uma alta presença de estudantes brancos em escolas privadas, e um alto número de estudantes pretos e partos em escolas estaduais e no CTPM.

Nas escolas privadas e federais, onde estão presentes as famílias com maiores concentrações de renda, notou-se que a variação entre o número de alunos do sexo masculino e do feminino é baixa. Já nas escolas estaduais, onde estão presentes mais estudantes de baixa renda matriculados, o maior número de alunos é do sexo feminino. Por fim, notou-se que a dependência administrativa foi a variável com maior peso dentre todas, deixando evidência a disparidade entre ensino privado e público.

No que tange aos pontos positivos, o estudo se destaca por sua análise detalhada acerca das variáveis socieconômicas aplicadas à uma realidade em particular, bem como por suas várias análises sobre as motivações dos resultados encontrados, possibilitando o formulação de políticas educacionais direcionadas e eficientes, destacando a necessidade de ações para reduzir a disparidade no ambiente educacional do estado.

Já em seus pontos negativos, podemos destacar a exclusão dos dados de alunos com nota zero, que embora justificável, pode limitar ou influir na análise de todo o cenário de desempenho. Junto a isso, a limitação do estudo à um ano específico, não podendo capturar tendências de longo prazo.

3.7 Considerações parciais

É crescente o interessante de estudos de análise de dados na área da educação, em particular, a literatura apresenta diversos trabalhos relacionados aos dados dos participantes

do Enem. Tais dados podem ser utilizados para diferentes metodologias de análises, bem como utilizados para trazer luz às diversas problemáticas acerca da edução do país, tais como a análise de desempenho por estado, por variáveis socioeconômicas, por estrutura das moradias, bem como diversas outras variáveis disponíveis.

Diante dessa revisão, encontramos caminhos que podem ser utilizados de base para novas pesquisas, bem como as lacunas e oportunidades para novas investigações. Em particular, essa trabalho se assemelha com as pesquisas apresentadas por aplicar o processo e técnicas de análise de dados, a estruturação das informações obtidas em gráficos, e a análise dos impactos socioeconômicos e demográficos em desempenhos dos candidados no exame do Enem.

4 Material e Métodos

Neste capítulo, são descritos detalhadamente os materiais e métodos utilizados para a análise dos dados do Enem utilizando técnicas de *Data Warehouse*. Primeiramente, são apresentados os dados coletados, incluindo as fontes e os critérios de seleção.

Em seguida, são explicados os procedimentos de extração, transformação e carregamento dos dados, bem como as ferramentas e tecnologias empregadas, como bancos de dados, softwares de análise e linguagens de programação. A metodologia foi planejada para garantir a integridade e a precisão dos dados, seguindo as melhores práticas da área de Data Warehouse.

A Figura 4.1 ilustra um resumo dessas etapas que são detalhadas nas próximas seções. Em resumo, o processo de descoberta de conhecimento ocorre nas etapas sequencias:

1. **Obtenção dos dados**: identificação das bases a serem estudadas;
2. **Análise dos dados**: análise dos dados presentes nas bases, identificação das variáveis de interesse e normalizações necessárias;
3. **Desenvolvimento técnico**: elaboração do modelo e desenvolvimento do banco de dados a ser utilizado, bem como os *scripts* ou programas necessários para a extração dos dados da base e carga no banco de dados;
4. **Criação de *dashboard***: uso de ferramentas para visualização de dados, definindo a conexão do banco de dados com essas ferramentas que dão suporte a tomada de decisão.

4.1 A base de dados

Esse trabalho trata sobre a análise dos microdados do Enem, que reúne informações detalhadas relacionadas ao exame numa base de dados anonimizados, isto é, não é possível

Figura 4.1: Processo de descoberta da informação.

identificar de qual candidato o dado unitário analisado pertence (INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA, 2023).

Foram analisadas as bases de dados dos exames aplicados entre 2015 e 2022, inclusive, que possuem acesso público no site do INEP[3]. As bases são compostas por documentos técnicos e documentos referentes a aplicação do exame (provas, editais, entre outros), dicionários dos dados, os dados em si e modelos de *scripts* para diversas tecnologias.

Após a análise dos dicionário de dados de todos os anos, foram escolhidas as variáveis em potencial a serem exploradas nesse estudo, conforme listado na Tabela 4.1. A escolha se justifica pela correlação das variáveis com o temática socieconomia e demográfica a ser estudado nessa trabalho. Nas próximas subseções serão descritas as variáveis apresentadas na tabela: a Seção 4.1.1 destaca como a redação dos estudantes é avaliada e como a nota final é obtida. Já a Seção 4.1.2 destacadas os grupos de escolares disponíveis na base, e, por fim, a Seção 4.1.3 destaca o agrupamento de profissões em grupos.

Tabela 4.1: Descrição das variáveis da base dos microdados do Enem.

Campo	Descrição	Variáveis Categóricas	
		Valor	Descrição
NU_INSCRICAO	Número de inscrição	-	-
NU_ANO	Ano do Enem	-	-
		1	Menor de 17 anos
		2	17 anos
		3	18 anos
		4	19 anos
		5	20 anos
		6	21 anos
		7	22 anos
		8	23 anos
		9	24 anos
TP_FAIXA_ETARIA	Faixa etária		

Continua na próxima página

[3]Base dos microdados do Enem - <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/Enem>

Tabela 4.1 – Continuação da página anterior

Campo	Descrição	Variáveis Categóricas	
		Valor	Descrição
		10	25 anos
		11	Entre 26 e 30 anos
		12	Entre 31 e 35 anos
		13	Entre 36 e 40 anos
		14	Entre 41 e 45 anos
		15	Entre 46 e 50 anos
		16	Entre 51 e 55 anos
		17	Entre 56 e 60 anos
		18	Entre 61 e 65 anos
		19	Entre 66 e 70 anos
		20	Maior de 70 anos
TP_SEXO	Sexo	M	Masculino
		F	Feminino
TP_ESTADO_CIVIL	Estado Civil	0	Não informado
		1	Solteiro(a)
		2	Casado(a)/Mora com companheiro(a)
		3	Divorciado(a)/Desquitado(a)/Separado(a)
		4	Viúvo(a)
TP_COR_RACA	Cor/raça	0	Não declarado
		1	Branca
		2	Preta
		3	Parda
		4	Amarela
		5	Indígena
		6	Não dispõe da informação
TP_NACIONALIDADE	Nacionalidade	0	Não informado
		1	Brasileiro(a)
		2	Brasileiro(a) Naturalizado(a)
		3	Estrangeiro(a)
		4	Brasileiro(a) Nato(a)
TP_ST_CONCLUSAO	Situação de conclusão do EM	1	Já concluí o EM
		2	Estou cursando e concluirei o EM no ano
		3	Estou cursando e concluirei o EM após o ano
		4	Não concluí e não estou cursando o EM
TP_ANO_CONCLUIU	Ano de Conclusão do EM	0	Não informado
		1	Ano do exame
		2	Ano do exame - 1
		3	Ano do exame -2
		4	Ano do exame - 3
		5	Ano do exame - 4
		6	Ano do exame - 5
		7	Ano do exame - 6
		8	Ano do exame - 7
		9	Ano do exame - 8
		10	Ano do exame - 9
		11	Ano do exame - 10
		12	Ano do exame - 11
		13	Ano do exame - 12
		14	Ano do exame - 13
		15	Ano do exame - 14
		16	Ano do exame - 15 ou mais
TP_ESCOLA	Tipo de escola do EM	1	Não Respondeu
		2	Pública
		3	Privada
TP_ENSINO	Tipo de instituição que concluiu ou concluirá o EM	1	Ensino Regular
		2	Educação Especial
IN_TREINEIRO	Indica se o inscrito fez a prova com intuito de apenas treinar	1	Sim
		0	Não

Continua na próxima página

Tabela 4.1 – Continuação da página anterior

Campo	Descrição	Variáveis Categóricas	
		Valor	Descrição
SG_UF_ESC	Sigla da Unidade da Federação da escola	-	-
TP_DEPENDENCIA_ADM_ESC	Dependência administrativa (Escola)	1	Federal
		2	Estadual
		3	Municipal
		3	Privada
TP_LOCALIZACAO_ESC	Localização (Escola)	1	Urbana
		2	Rural
NU_NOTA_CN	Nota da prova de Ciências da Natureza	-	-
NU_NOTA_CH	Nota da prova de Ciências Humanas	-	-
NU_NOTA_LC	Nota da prova de Linguagens e Códigos	-	-
NU_NOTA_MT	Nota da prova de Matemática	-	-
TP_LINGUA	Língua Estrangeira	1	Inglês
		2	Espanhol
TP_STATUS_REDACAO	Situação da redação do participante	1	Sem problemas
		2	Anulada
		3	Cópia Texto Motivador
		4	Em Branco
		6	Fuga ao tema
		7	Não atendimento ao tipo textual
		8	Texto insuficiente
		9	Parte desconectada
NU_NOTA_COMP1	Nota da competência 1	-	-
NU_NOTA_COMP2	Nota da competência 2	-	-
NU_NOTA_COMP3	Nota da competência 3	-	-
NU_NOTA_COMP4	Nota da competência 4	-	-
NU_NOTA_COMP5	Nota da competência 5	-	-
NU_NOTA_REDACAO	Nota da prova de redação	-	-
Q001	Até que série seu pai, ou O homem responsável por você, estudou?	A	Nunca estudou
		B	Nível B
		C	Nível C
		D	Nível D
		E	Nível E
		F	Nível F
		G	Nível G
		H	Não sei
Q002	Até que série sua mãe, ou a mulher responsável por você, estudou?	A	Nunca estudou.
		B	Nível B
		C	Nível C
		D	Nível D
		E	Nível E
		F	Nível F
		G	Nível G
		G	Não sei
Q003	Ocupação do seu pai ou do homem responsável por você?	A	Grupo 1
		B	Grupo 2
		C	Grupo 3
		D	Grupo 4
		E	Grupo 5
		F	Não sei
Q004	Ocupação da sua mãe Ou da mulher responsável por você.	A	Grupo 1
		B	Grupo 2
		C	Grupo 3
		D	Grupo 4
		E	Grupo 5
		F	Não sei
Q006	Qual é a renda mensal de sua família em salários mínimos?	A	Nenhuma Renda
		B	$(0 - 1.0]$

Continua na próxima página

Tabela 4.1 – Continuação da página anterior

Campo	Descrição	Variáveis Categóricas	
		Valor	Descrição
		C	$(1.0 - 1.5]$
		D	$(1.5 - 2.0]$
		E	$(2.0 - 2.5]$
		F	$(2.5 - 3.0]$
		G	$(3.0 - 4.0]$
		H	$(4.0 - 5.0]$
		I	$(5.0 - 6.0]$
		J	$(6.0 - 7.0]$
		K	$(7.0 - 8.0]$
		L	$(8.0 - 9.0]$
		M	$(9.0 - 10.0]$
		N	$(10.0 - 12.0]$
		O	$(12.0 - 15.0]$
		P	$(15.0 - 20.0]$
		Q	$(20.0 - \infty)$

4.1.1 Matriz de Referência da Redação

Conforme a base de microdados (INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA, 2023), a redação de cada participante é avaliada em cinco competências distintas. Abaixo temos os fatores considerados em cada uma:

- **Competência 1**: Demonstrar domínio da modalidade escrita formal da Língua Portuguesa;
- **Competência 2**: Compreender a proposta de redação e aplicar conceitos das várias áreas de conhecimento para desenvolver o tema, dentro dos limites estruturais do texto dissertativo-argumentativo em prosa;
- **Competência 3**: Selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista;
- **Competência 4**: Demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação;
- **Competência 5**: Elaborar proposta de intervenção para o problema abordado, respeitando os direitos humanos.

4.1.2 Nível de educação

Para responder as questões do questionários socieconômico referentes a escolaridade dos pais, os inscritos devem seguir de acordo com os níveis descritos a seguir:

- **Nível B**: Não completou a 4ª série/5º ano do Ensino Fundamental;
- **Nível C**: Completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental;
- **Nível D**: Completou a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio;
- **Nível E**: Completou o Ensino Médio, mas não completou a Faculdade;
- **Nível F**: Completou a Faculdade, mas não completou a Pós-graduação;
- **Nível G**: Completou a Pós-graduação.

4.1.3 Grupos de profissões

Já em relação à profissão de seus genitores, os inscritos do Enem devem escolher a opção que contém a profissão exata ou mais próxima dentre as opções de grupos disponíveis:

- **Grupo 1** - Lavrador, agricultor sem empregados, bóia fria, criador de animais (gado, porcos, galinhas, ovelhas, cavalos etc.), apicultor, pescador, lenhador, seringueiro, extrativista.
- **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.
- **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.

- **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.

- **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 empresas com mais de 10 empregados.

4.2 Construção do modelo de dados

Após a conclusão do entendimento da base e seleção das potenciais variáveis, foi necessária a criação do modelo de dados a ser implementado no Banco de dados. A construção do modelo de dados para um *data warehouse* é uma etapa essencial que envolve a definição de uma estrutura de dados voltada para consultas analíticas. Este modelo organiza os dados em uma tabela fato central, que armazena as métricas de negócios, e tabelas dimensão, que descrevem os atributos contextuais das métricas. A simplicidade do modelo de dados com essa estrutura de tabelas facilita a execução de consultas e a geração de relatórios, garantindo desempenho e uma fácil compreensão dos dados. Assim, o modelo de dados projetado visa dar suporte a análise dos dados do Enem de forma eficaz. Nesse sentido, foi decidido pela criação de três tabelas:

- **escola**: contém os dados de todas as escolas nas quais os candidatos estudaram, independente do ano de execução da prova;

- **aluno**: contém os dados de todos os inscritos no Enem por ano;

- **prova_executada**: contém os dados referentes a prova executada por um aluno inscrito em um determinado ano.

As relações e colunas mapeadas e seus detalhes são apresentadas no modelo de entidade relacionamento ilustrado na Figura 4.2, sendo *prova_ executada* a tabela fato e *escola* e *aluno* as tabelas de dimensão. O modelo de dados foi definido com uma granularidade anual, refletindo a periodicidade do Enem, que é aplicado uma vez por ano. Essa escolha de granularidade permite uma organização eficiente dos dados, facilitando a análise de tendências e comparações anuais. Com a granularidade anual, é possível consolidar informações detalhadas de cada edição do exame, proporcionando uma visão abrangente e histórica do desempenho dos candidatos ao longo dos anos. Essa abordagem também simplifica a manutenção e atualização do data warehouse, garantindo que os dados estejam sempre alinhados com o ciclo anual do Enem.

Nesta pesquisa, optou-se tela utilização do banco de dados MySQL na versão 8.0.7 [4], dada a familiaridade do autor com esta tecnologia.

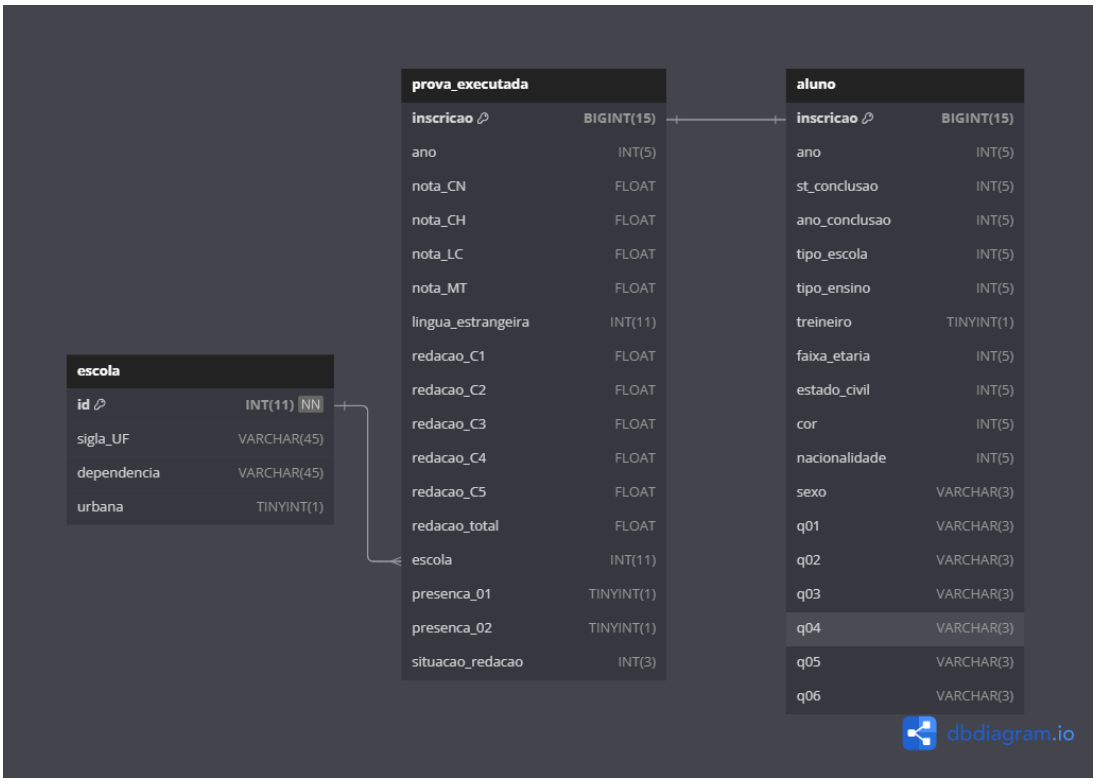

Figura 4.2: Arquitetura do modelo de dados desenvolvido.

[4] <https://www.mysql.com/>

4.3 Carga no Banco de Dados

Com o banco de dados devidamente criado, foi necessário desenvolver um programa para automatizar o processo de carga dos dados anuais do Enem. Este programa tem a responsabilidade de ler os arquivos de dados de cada ano de aplicação do exame, realizar a validação de integridade dos dados para garantir que não haja inconsistências ou erros, e, em seguida, alimentar o banco de dados com os dados validados. Optou-se pela utilização da linguagem de programação Java na versão 8.301 [5] que permitiu a criação de um programa que não apenas executa a carga de dados de forma eficiente, mas também pode ser facilmente mantido e atualizado conforme necessário, assegurando a integridade e a confiabilidade do data warehouse ao longo do tempo.

Para uma melhor organização e futuras manutenções, o programa foi desenvolvido em módulos, que podem ser acionados individualmente:

1. **Módulo de carregamento de escolas**: Como o conjunto de dados possui um número limitado de escolas, optou-se pela implementação de um módulo capaz de percorrer todos os arquivos de uma pasta e realizar a identificação apenas dos dados de escolas para inserção no BD. O módulo utiliza um *hash* criado em tempo de execução com as informações das escolas já presentes no banco na tabela *escola*. Assim, para cada linha do arquivo, verifica se a escola já está presente no *hash*, caso não esteja, adiciona a escola ao *hash* e faz a inserção dos dados no BD.

2. **Módulo de criação de *Hash* de escolas**: É responsável por consultar todas as escolas presentes na tabela *escola* do BD e criar um *hash* em memória mapeando uma *string* resultaante da concatenação dos dados de estado, localização e dependência administrativa da escola e construindo um mapeamento para a chave primária do registro no BD.

3. **Módulo de inserção de dados das provas executadas:** Módulo responsável pela leitura de todos os arquivos de uma pasta e, para cada linha do arquivo, identificar o identificador da escola através do *hash* em memória e inserir os dados da linha nas tabelas de *aluno* e *prova_ executada*.

[5] <https://www.java.com/>

4.4 Apresentação dos dados

Com o modelo de dados devidamente criado e populado, torna-se indispensável que esses dados sejam apresentados de maneira a facilitar sua análise e visualização, especialmente para usuários não técnicos. Para alcançar esse objetivo, é fundamental utilizar ferramentas de visualização de dados que ofereçam uma interface intuitiva e recursos avançados de análise. Nesse contexto, optou-se pela utilização do Power BI na ver sao 2.136.1202.0 64-bit [6] em sua versão gratuita. O Power BI é uma ferramenta que permite a criação de *dashboards* interativos e relatórios detalhados, proporcionando uma visão clara e compreensível dos dados. A escolha do Power BI se deve à sua capacidade de integrar-se facilmente com diversas fontes de dados, sua facilidade de uso e a ampla gama de visualizações disponíveis, que permitem transformar dados complexos em *insights*. Além disso, a versão gratuita do Power BI oferece funcionalidades suficientes para atender às necessidades iniciais do projeto, garantindo que os dados estejam acessíveis.

4.4.1 Conexão com o modelo e transformação de dados

O primeiro passo foi estabelecer a conexão do Power BI com o banco de dados criado, garantindo que a ferramenta pudesse acessar e manipular os dados armazenados. Após a conexão, o Power BI oferece prévias das visões de dados, permitindo aos usuários configurar relacionamentos entre as tabelas, ajustar os tipos de dados e personalizar a visualização de cada coluna. Essas funcionalidades são essenciais para garantir que os dados sejam apresentados de forma clara e precisa.

Durante essa etapa, foram realizadas diversas configurações necessárias, como a formatação de valores decimais para exibição com duas casas decimais e a definição de relacionamentos entre tabelas para facilitar a busca de informações correlacionadas. Além disso, optou-se pela criação de novas colunas para enriquecer as informações das provas, proporcionando *insights* adicionais. Entre essas novas colunas, destacam-se:

- **Média**: Calculada como a média aritmética simples das notas dos quatro componentes curriculares, somada à nota final da redação do candidato. Essa métrica

[6] Apresentação da ferramenta Power BI - <https://www.microsoft.com/pt-br/power-platform/products/power-bi>

oferece uma visão consolidada do desempenho geral do candidato.

- **Presença**: Indicador binário que mostra se o candidato esteve presente nos dois dias de aplicação da prova. Essa informação é crucial para análises de participação e desempenho.

Essas transformações e configurações no Power BI não apenas facilitam a análise dos dados, mas também tornam as informações mais acessíveis e compreensíveis.

4.4.2 Criação de relatórios

Por fim, foram desenvolvidos relatórios detalhados que permitem uma análise abrangente dos dados do Enem. Para as análises de notas, foram considerados apenas os candidatos que estiveram presentes nos dois dias de prova e que não apresentaram problemas na redação.

As visualizações criadas possibilitam a avaliação das médias das notas, das notas por componente curricular, da nota final da redação e das ausências. Para cada um desses dados destacados, foram criadas segmentações específicas, como por ano, por ano e região geográfica do país, e por ano e faixa de renda. Nas segmentações por região geográfica e faixa de renda, foram excluídos os dados que não tinham essas informações preenchidas, garantindo a precisão das análises.

Além disso, foi desenvolvida uma visão específica para identificar problemas nas redações, segmentada por faixa de renda, permitindo uma análise mais detalhada das dificuldades enfrentadas por candidatos de baixa renda.

Todas as visualizações permitem segmentar e avaliar os dados por ano e por *status* de candidato (treineiro ou não). Isso proporciona uma flexibilidade significativa na análise, permitindo que os usuários explorem os dados de diferentes perspectivas. Este trabalho não fez o uso de técnicas de *Machine Learning*, porém essa e outras abordagem também podem ser utilizadas para análise dos dados.

Por fim, destaca-se que a criação de novas visualizações é facilitada pela própria ferramenta Power BI, que oferece uma interface intuitiva e descomplicada para a personalização e expansão dos relatórios conforme necessário.

5 Resultados

Este capítulo é dedicado à discussão dos resultados obtidos ao longo do estudo, organizados de acordo com cada variável analisada. A análise detalhada de cada variável permitirá uma compreensão aprofundada dos dados e das tendências observadas. Inicialmente, apresentamos a volumetria dos dados de inscritos utilizados neste estudo, segmentados por ano de aplicação do Enem, conforme ilustrado na Tabela 5.1. Essa tabela fornece uma visão geral da quantidade de inscritos ao longo dos anos, servindo como base para as análises subsequentes.

Tabela 5.1: Quantidade de inscritos no Enem por ano

Ano	**Tamanho**	**Porcentagem**
2015	7.746.428	$16,71\%$
2016	8.627.180	$18,61\%$
2017	6.731.279	$14,52\%$
2018	5.513.734	$11,89\%$
2019	5.095.172	$10,99\%$
2020	5.783.110	$12,47\%$
2021	3.389.833	$7,31\%$
2022	3.476.106	$7,50\%$
Total	46.362.842	100%

Podemos observar uma tendência de diminuição na quantidade de inscritos ao longo dos anos. Essa redução pode ser atribuída a diversos fatores, como mudanças nas políticas educacionais, variações na percepção da importância do exame entre os estudantes, e possíveis alterações demográficas. A análise detalhada dos dados ao longo dos anos revela não apenas uma queda no número total de inscritos, mas também possíveis variações regionais e socioeconômicas que podem estar influenciando essa tendência. Compreender essas dinâmicas é crucial para identificar os desafios e oportunidades no contexto educacional e para desenvolver estratégias que possam incentivar uma maior participação no exame.

5.1 Análise de Ausência

A Figura 5.1 ilustra a porcentagem de candidatos ausentes e presentes por ano. Observa-se que, em geral, a porcentagem de ausência se mantém relativamente constante ao longo dos anos. No entanto, uma exceção notável é o ano de 2020, onde o número de ausentes ultrapassa o número de presentes. Este ano foi atípico devido ao início da pandemia de COVID-19, que impactou drasticamente a participação dos candidatos no exame. A pandemia trouxe desafios como o medo de contágio, restrições de mobilidade e dificuldades econômicas, que contribuíram para o aumento da taxa de ausência.

A Figura 5.2 apresenta a porcentagem de ausência segregada por faixa de renda. A análise revela que as faixas de renda B e C são responsáveis pelos maiores índices de ausência no exame. Isso pode ser atribuído a fatores socioeconômicos que afetam a capacidade desses candidatos de comparecerem ao exame, como falta de recursos para transporte, necessidade de trabalhar para complementar a renda familiar, ou até mesmo falta de acesso a informações e suporte educacional adequado. Em contraste, as demais faixas de renda apresentam índices de ausência significativamente menores, com cada uma contribuindo com menos de 0,05% para o total de ausências. Essa disparidade destaca a importância de políticas públicas e iniciativas que visem reduzir as barreiras enfrentadas pelos candidatos de faixas de renda mais baixas, promovendo uma maior equidade no acesso à educação.

5.2 Análise de Desempenho

O desempenho dos candidatos foi avaliado de forma abrangente, considerando a média geral, as notas de cada componente curricular e a redação. A análise foi realizada ao longo dos anos, permitindo identificar tendências e variações no desempenho dos candidatos. Além disso, os dados foram segregados por faixa de renda e por região, proporcionando uma visão detalhada das diferenças de desempenho entre diferentes grupos socioeconômicos e geográficos.

Para facilitar a compreensão das faixas de renda, a Tabela 4.1 apresenta a descrição de cada faixa. É importante notar que, quanto mais distante do início do alfabeto,

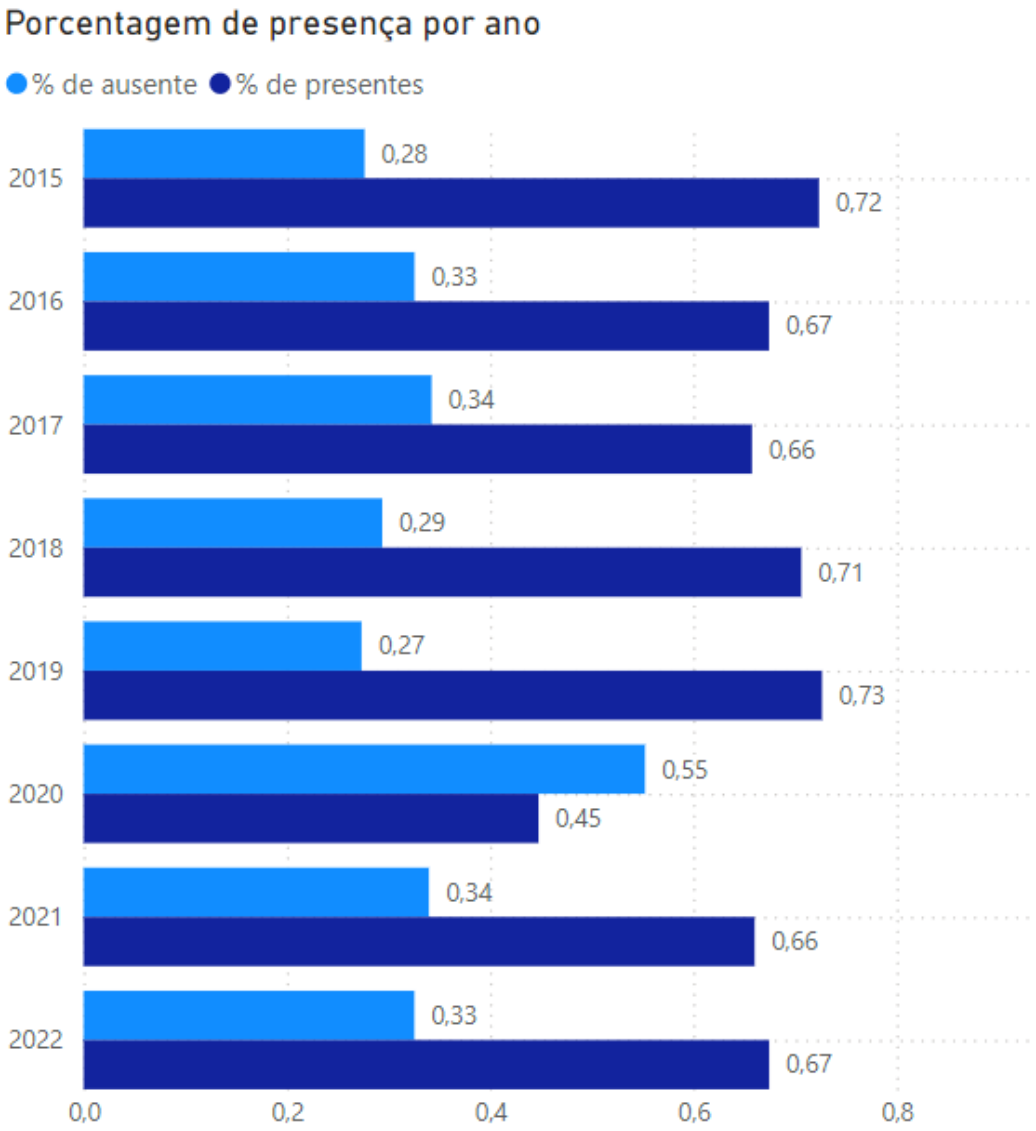

Figura 5.1: Porcentagem de inscritos presentes e ausentes por ano.

maior é o poder aquisitivo do candidato. Essa classificação permite uma análise mais precisa das influências socioeconômicas no desempenho dos candidatos.

A análise de desempenho incluirá:

- Média Geral: Avaliação da média aritmética das notas dos candidatos, proporcionando uma visão consolidada do desempenho geral ao longo dos anos.
- Componentes Curriculares: Análise detalhada das notas em cada componente curricular, como Matemática, Ciências da Natureza, Ciências Humanas e Linguagens. Isso permitirá identificar áreas de força e fraqueza entre os candidatos.

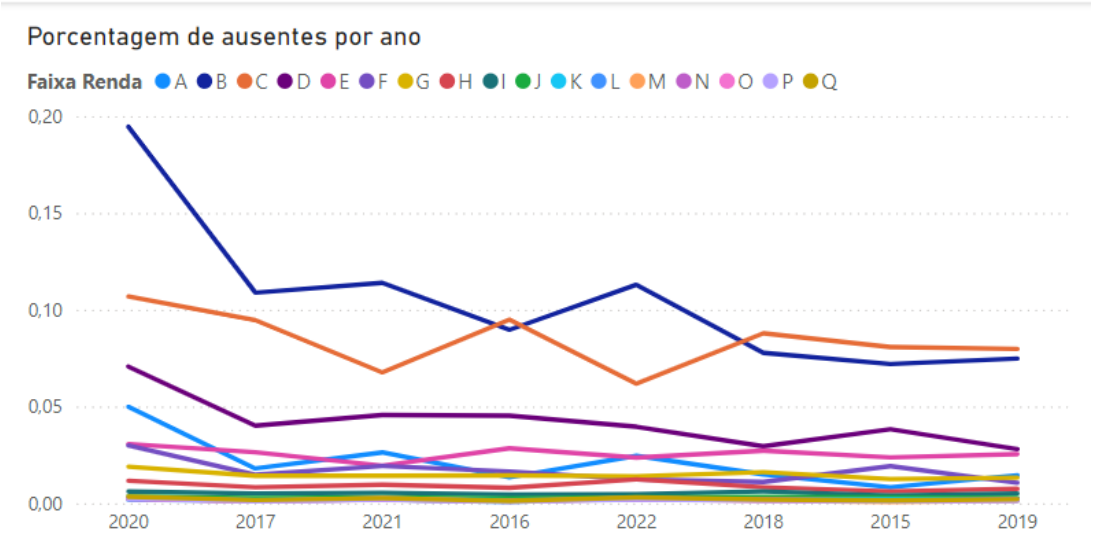

Figura 5.2: Porcentagem de inscritos ausentes por faixa de renda.

- Redação: Avaliação das notas de redação, considerando critérios como coerência, coesão, argumentação e domínio da norma culta. A análise da redação é crucial para entender as habilidades de comunicação escrita dos candidatos.

Além disso, a análise foi segmentada por:

- Faixa de Renda: Comparação do desempenho entre diferentes faixas de renda, destacando como o poder aquisitivo influencia os resultados dos candidatos. Isso ajudará a identificar desigualdades e a necessidade de políticas educacionais mais inclusivas.
- Região: Avaliação do desempenho por região geográfica, permitindo identificar disparidades regionais e direcionar esforços para melhorar a educação em áreas menos favorecidas.

Essa abordagem detalhada e segmentada permitiu uma compreensão profunda dos fatores que influenciam o desempenho dos candidatos no Enem, fornecendo *insights* valiosos para a formulação de políticas educacionais e estratégias de melhoria contínua.

5.2.1 Média geral

A Figura 5.3 evidencia uma correlação direta entre o poder aquisitivo dos candidatos e seu desempenho no exame. Observa-se que candidatos pertencentes a faixas de renda mais

altas tendem a obter notas significativamente superiores em comparação com aqueles de faixas de renda mais baixas. Essa tendência sugere que fatores socioeconômicos desempenham um papel crucial na preparação e no desempenho dos candidatos, possivelmente devido ao acesso diferenciado a recursos educacionais, como aulas particulares, material de estudo de qualidade e ambientes de aprendizagem mais favoráveis.

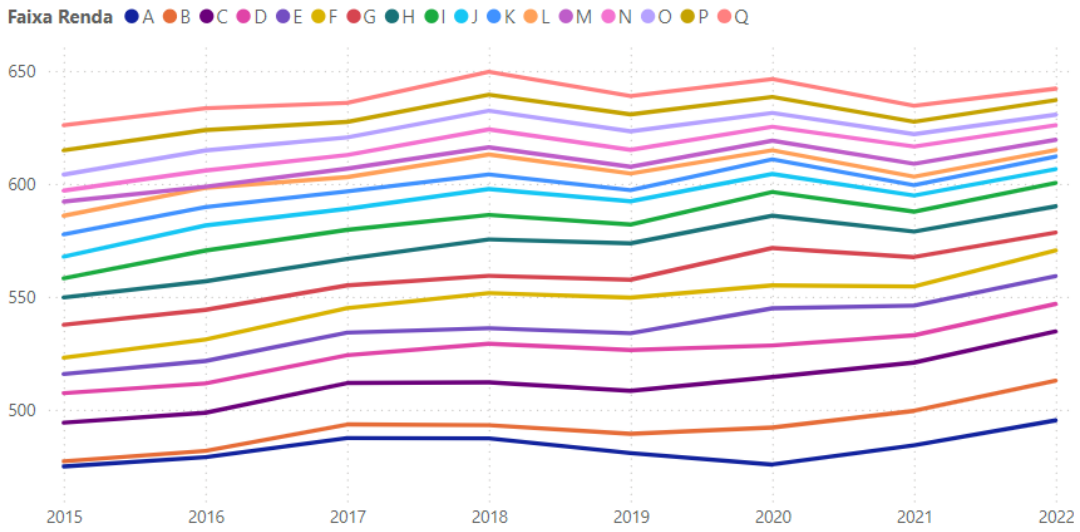

Figura 5.3: Média geral dos inscritos por faixa de renda.

Além disso, a Figura 5.4 destaca as disparidades regionais no desempenho dos candidatos. As regiões Sul e Sudeste apresentam os maiores desempenhos médios, indicando que os candidatos dessas áreas tendem a obter notas mais altas no exame. Em contraste, a região Norte registra os piores resultados.

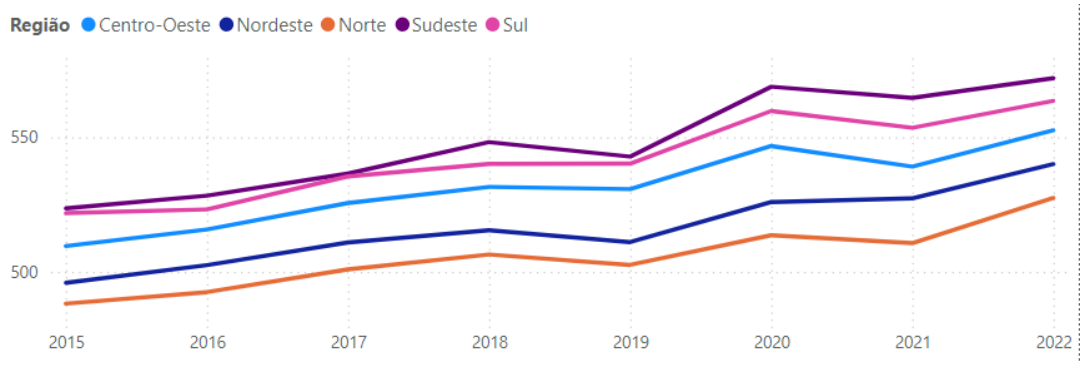

Figura 5.4: Média geral dos inscritos por região.

Essas análises revelam a importância de considerar tanto o contexto socioeconômico quanto as disparidades regionais ao avaliar o desempenho dos candidatos no Enem.

Elas também ressaltam a necessidade de políticas públicas direcionadas que visem reduzir essas desigualdades, proporcionando oportunidades mais equitativas para todos os estudantes, independentemente de sua renda ou localização geográfica.

5.2.2 Ciências Humanas

A Figura 5.5 apresenta as notas dos candidatos no componente de Ciências Humanas, revelando um padrão de desempenho que é diretamente proporcional à faixa de renda, similar ao observado na análise da média geral. Candidatos pertencentes a faixas de renda mais altas tendem a obter notas superiores.

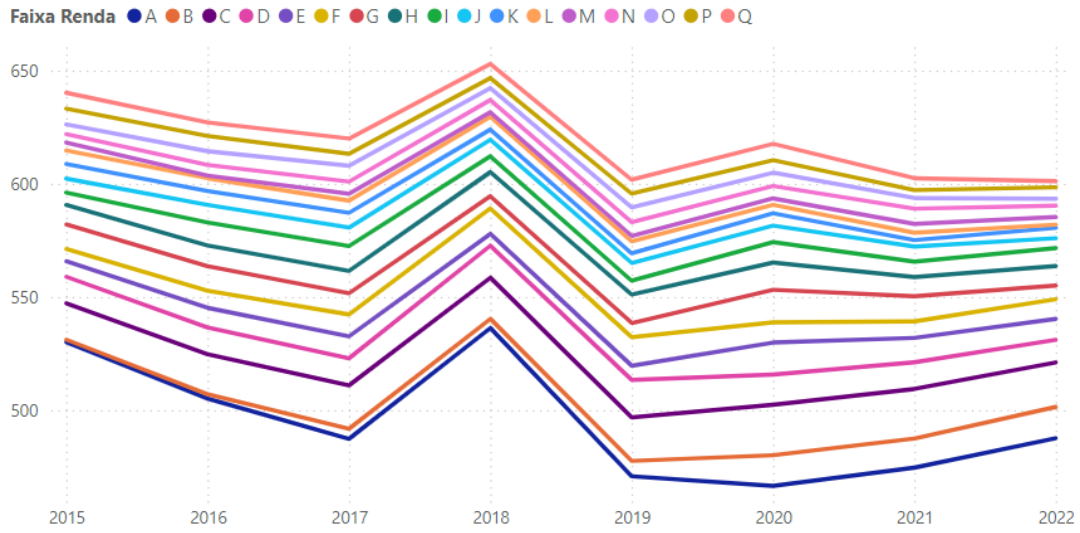

Figura 5.5: Nota dos candidatos em Ciências Humanas por faixa de renda.

Além disso, a Figura 5.6 destaca as diferenças regionais nas notas de Ciências Humanas. As regiões Sul e Sudeste apresentam desempenhos bastante próximos, indicando uma homogeneidade maior na qualidade da educação e nos recursos disponíveis nessas áreas. Em contraste, os demais estados seguem o padrão observado na média geral, com notas inferiores, especialmente nas regiões Norte e Nordeste.

5.2.3 Ciências da Natureza

No que diz respeito ao componente de Ciências da Natureza, a Figura 5.7 demonstra um comportamento similar ao observado na média geral, com uma correlação direta entre o poder aquisitivo dos candidatos e suas notas. No entanto, nota-se uma aproximação maior

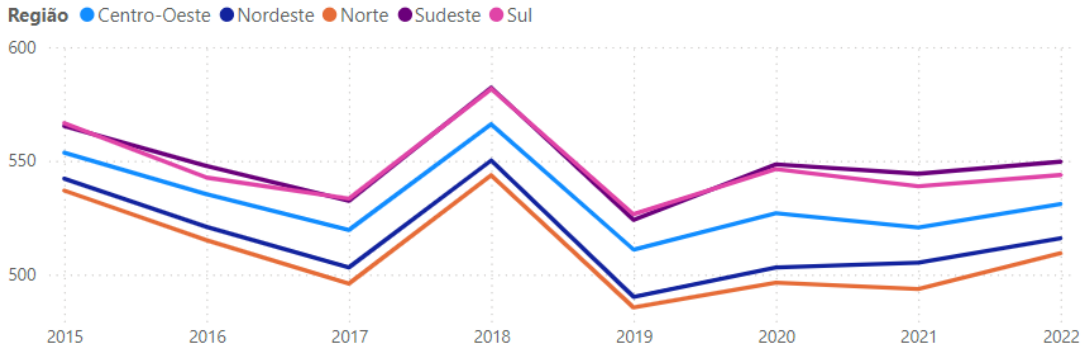

Figura 5.6: Nota dos candidatos em Ciências Humanas por região.

entre as notas das faixas de renda A e B, indicando que, embora ainda haja uma diferença de desempenho, ela é menos acentuada entre essas duas faixas em comparação com outras áreas de estudo. Esse comportamento pode ser influenciado por fatores específicos relacionados ao ensino de Ciências da Natureza, como a disponibilidade de laboratórios e recursos didáticos em escolas de diferentes faixas de renda.

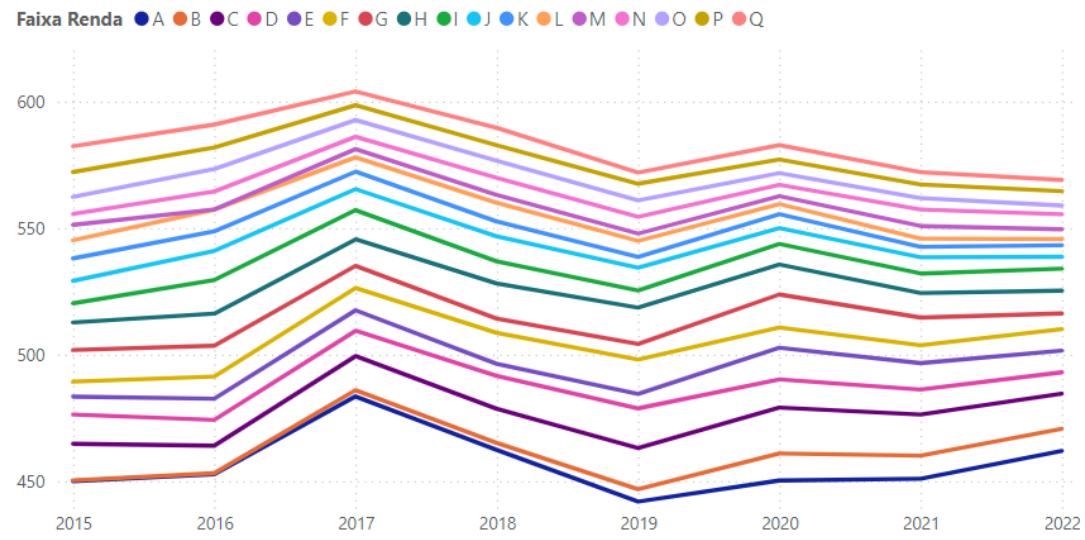

Figura 5.7: Nota dos candidatos em Ciências da Natureza por faixa de renda.

A Figura 5.8 destaca as notas dos candidatos nas regiões Sudeste e Sul, mostrando uma aproximação significativa entre os desempenhos dessas regiões. Isso sugere uma maior homogeneidade na qualidade do ensino de Ciências da Natureza nessas áreas, possivelmente devido a investimentos mais consistentes em infraestrutura educacional e recursos de ensino. Em contraste, as demais regiões seguem o padrão observado na média geral, com notas inferiores, especialmente nas regiões Norte e Nordeste.

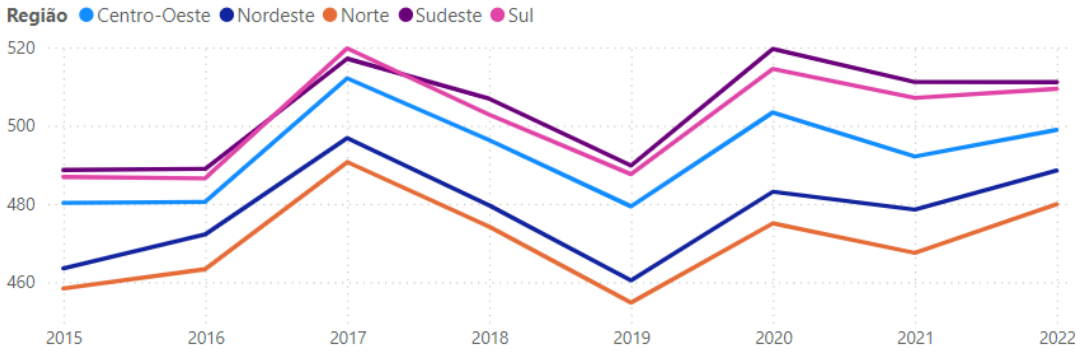

Figura 5.8: Nota dos candidatos em Ciências da Natureza por região.

5.2.4 Linguagens e Códigos

No que diz respeito ao componente de Linguagens e Códigos, a Figura 5.9 revela uma aproximação das notas de algumas faixas de renda, como L e M, e P e Q, indicando que candidatos dessas faixas apresentam desempenhos semelhantes. Além disso, observa-se uma aproximação das notas das faixas A e B entre os anos de 2015 e 2019, com um distanciamento posterior. Esse comportamento pode refletir mudanças nas políticas educacionais, variações no acesso a recursos de ensino ou outras influências socioeconômicas que impactaram o desempenho dos candidatos ao longo do tempo.

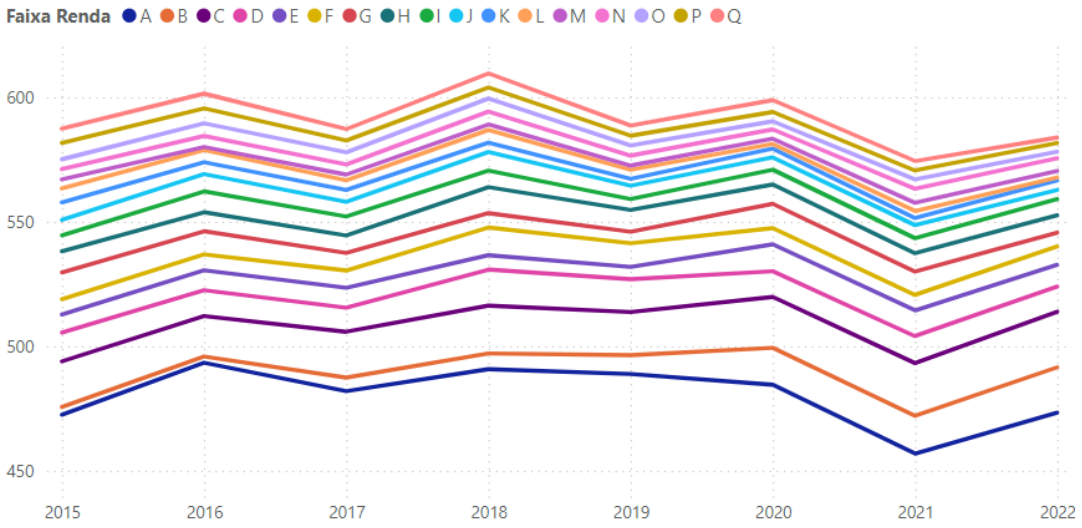

Figura 5.9: Nota dos candidatos em Linguagens e Códigos por faixa de renda.

Assim como foi observada em Ciências da Natureza, a Figura 5.10 também destaca uma aproximação entre as notas dos candidatos das regiões Sul e Sudeste, sugerindo uma homogeneidade maior na qualidade do ensino de Linguagens e Códigos nessas áreas.

Em contraste, as demais regiões seguem o padrão observado na média geral, com desempenhos inferiores, especialmente nas regiões Norte e Nordeste.

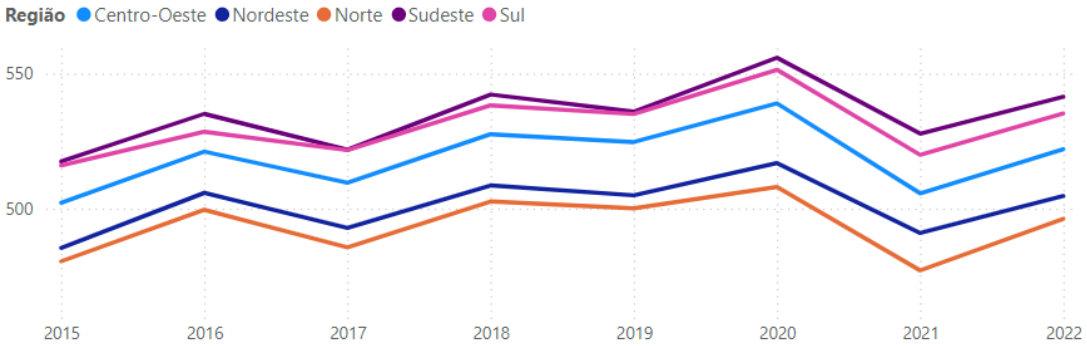

Figura 5.10: Nota dos candidatos em Linguagens e Códigos por região.

5.2.5 Matemática

No que diz respeito ao componente de Matemática, a Figura 5.11 mostra que as notas dos candidatos tendem a estar acima da média geral, com um destaque significativo para as faixas de renda mais alta. Candidatos dessas faixas apresentam um desempenho superior, o que pode ser atribuído ao acesso a melhores recursos educacionais e ambientes de aprendizagem mais apropriados e de melhor qualidade de ensino. Além disso, observa-se uma menor diferença entre as notas das faixas L e M, indicando uma certa homogeneidade no desempenho desses grupos específicos.

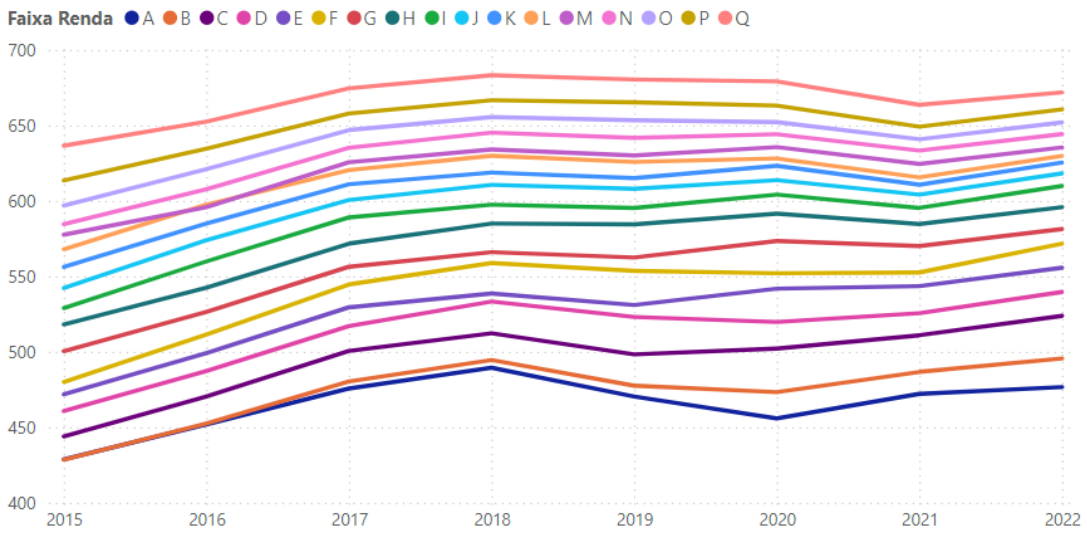

Figura 5.11: Nota dos candidatos em Matemática por faixa de renda.

A Figura 5.12 destaca que as regiões Sudeste e Sul possuem notas médias similares no componente de Matemática, apresentando a média mais próxima entre todas as regiões, seguindo o padrão identificado nas outras componentes. Essa proximidade sugere uma maior uniformidade na qualidade do ensino de Matemática nessas áreas. Em contraste, as demais regiões seguem o padrão observado na média geral e nas demais componentes, com desempenhos inferiores, particularmente nas regiões Norte e Nordeste.

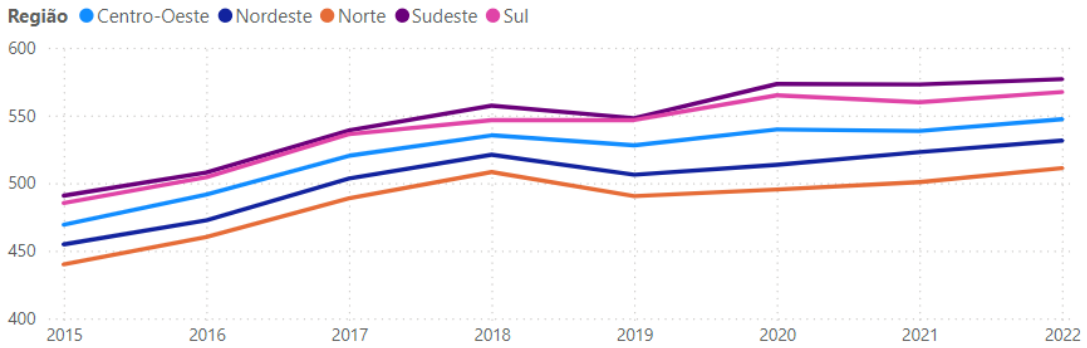

Figura 5.12: Nota dos candidatos em Matemática por região.

5.2.6 Redação

Nesta seção, vamos destacar as análises relacionadas à redação dos candidatos, abordando tanto a avaliação das notas quanto a situação das redações. A redação é um componente crucial do Enem, pois avalia a capacidade dos candidatos de expressar suas ideias de forma clara, coerente e argumentativa. Analisaremos as notas obtidas pelos candidatos ao longo dos anos, identificando padrões e tendências de desempenho. Além disso, examinaremos as situações das redações, como problemas identificados e a presença de discrepâncias, para entender melhor os desafios enfrentados pelos candidatos e as áreas que necessitam de melhorias. Essa análise detalhada pode fornecer *insights* valiosos sobre a proficiência dos candidatos na produção textual e pode contribuir para a formulação de estratégias educacionais mais eficazes.

Notas

A Figura 5.13 destaca uma significativa disparidade nas notas de redação entre os candidatos, com uma diferença média de 200 pontos entre aqueles das faixas de renda A e

Q. Essa variação é particularmente relevante para os candidatos que almejam cursos mais concorridos, onde cada ponto pode ser decisivo para a classificação no SISU. A análise revela que candidatos de faixas de renda mais alta, assim como nas outras componentes, tendem a obter notas superiores, possivelmente devido a preparação específica para a redação.

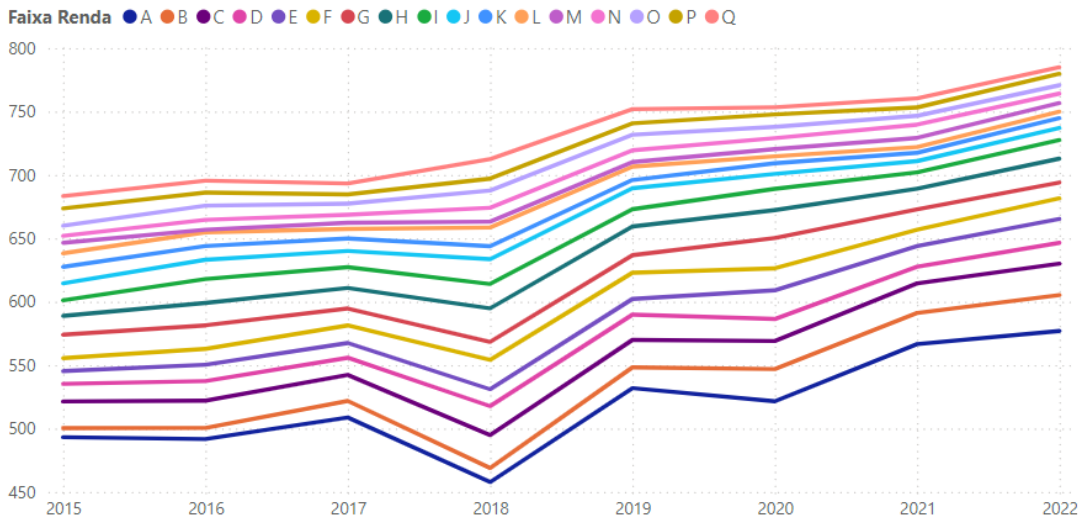

Figura 5.13: Nota dos candidatos em Redação por faixa de renda.

Além disso, a Figura 5.14 mostra que a região Sudeste possui a melhor média de notas de redação, refletindo uma qualidade de ensino mais homogênea. Notavelmente, essa figura também ilustra uma aproximação nas notas de redação entre as regiões Sul e Centro-Oeste, indicando uma melhoria na qualidade do ensino de redação nessas áreas.

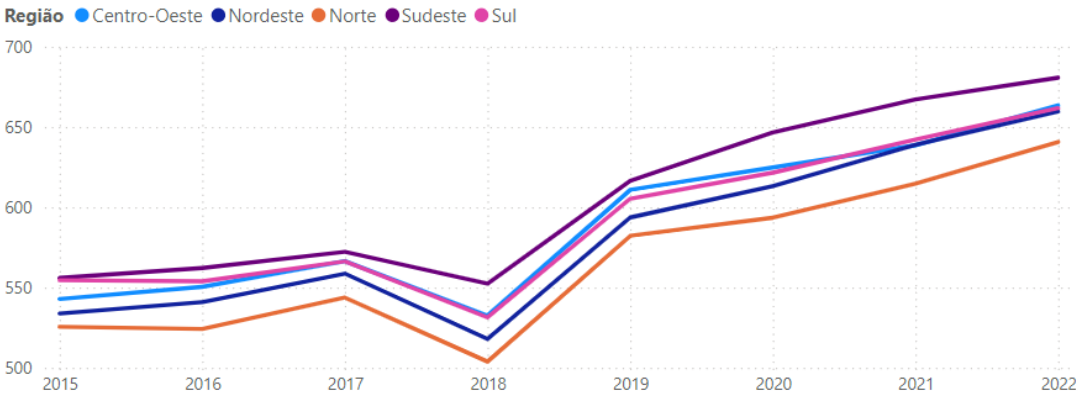

Figura 5.14: Nota dos candidatos em Redação por região.

Situação da redação

A Figura 5.15 ilustra a quantidade de redações que apresentaram algum problema, ocasionando em notas zeradas ou anuladas, o que resultou na eliminação dos candidatos do exame. A análise revela que a maioria dos erros ocorre entre os candidatos das faixas de renda mais baixas, especialmente nas faixas B e C, evidenciando uma diferença significativa em relação às demais faixas de renda. Esse padrão sugere que candidatos de faixas de renda mais baixa enfrentam maiores desafios na produção textual, possivelmente devido a falta de orientação adequada e ambientes de estudo menos favoráveis. Em contraste, as faixas de renda mais alta apresentam poucas ocorrências de redações problemáticas, indicando que esses candidatos estão melhor preparados para atender aos critérios exigidos na redação do Enem. Essas disparidades sublinham a necessidade de intervenções educacionais direcionadas que possam fornecer suporte adicional aos candidatos de faixas de renda mais baixa, ajudando a reduzir a incidência de problemas nas redações e promovendo uma maior equidade no desempenho dos candidatos.

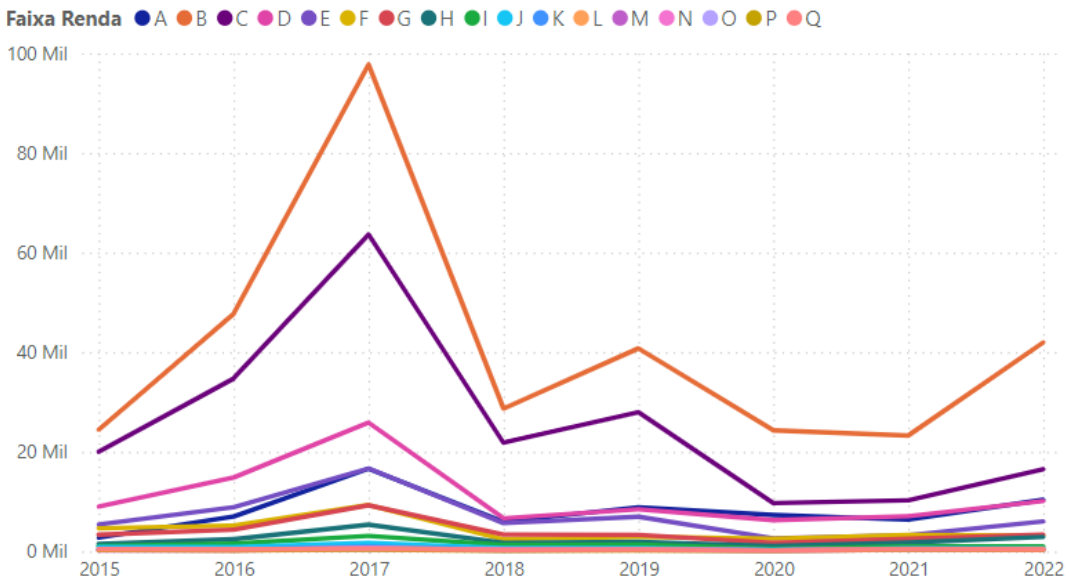

Figura 5.15: Quantidade de redações com problema por faixa de renda.

5.3 Considerações do Capítulo

Neste capítulo, foi apresentado uma análise detalhada dos diversos componentes do Exame Nacional do Ensino Médio (Enem), abrangendo as áreas de Ciências Humanas, Ciências da Natureza, Linguagens e Códigos, Matemática e Redação. As análises revelaram padrões importantes e destacaram as influências socioeconômicas e regionais no desempenho dos candidatos.

Observamos que o desempenho dos candidatos é diretamente proporcional à sua faixa de renda, com candidatos de faixas mais altas obtendo notas superiores em todos os componentes. Essa tendência é particularmente evidente nas análises de Ciências Humanas e Matemática, onde a diferença de desempenho entre as faixas de renda é mais acentuada. As regiões Sul e Sudeste, de maneira geral, apresentaram os melhores desempenhos, refletindo uma maior homogeneidade na qualidade do ensino e acesso a recursos educacionais.

A análise das notas de redação destacou uma disparidade significativa entre as faixas de renda, com candidatos de faixas mais altas obtendo notas superiores. Além disso, identificamos uma alta incidência de problemas nas redações entre candidatos de faixas de renda mais baixas, especialmente nas faixas B e C, o que resultou em notas zeradas ou anuladas.

6 Conclusão e Trabalhos Futuros

O ENEM tem se tornado cada vez mais importante para o acesso à educação superior no Brasil. Apesar das adaptações contínuas para melhorar o exame e ajustá-lo às novas realidades, inúmeros desafios ainda precisam ser superados. Este trabalho propôs o uso de um data *warehouse* para a análise dos microdados do ENEM, permitindo uma visão mais detalhada e estruturada das informações.

Através desta abordagem, foi possível observar as disparidades socioeconômicas e geográficas que afetam os candidatos. Um dos principais destaques foi a associação entre a faixa de renda e o desempenho dos candidatos, tanto na média geral quanto na análise dos componentes específicos. Observou-se que, em média, quanto maior a faixa de renda do candidato, melhor é seu desempenho no exame. Essa correlação destaca a influência significativa do contexto socioeconômico na preparação e nos resultados dos estudantes.

Além disso, a análise revelou uma maior incidência de redações anuladas ou zeradas entre os candidatos das faixas de renda mais baixas, até 1,5 salário mínimo. Esse problema pode estar relacionado à qualidade do ensino acessível a esses candidatos ou à necessidade de conciliar trabalho e estudo para contribuir com a renda familiar, o que pode levar até mesmo à evasão escolar.

No que diz respeito à ausência dos candidatos, notou-se que o percentual de ausência se mantém relativamente constante ao longo dos anos, com exceção do ano de início da pandemia de COVID-19 (2020). As faixas de renda B e C foram as que mais contribuíram para o percentual de ausência no exame, indicando que fatores socioeconômicos também influenciam a participação dos candidatos.

A utilização de um *data warehouse* para a análise dos microdados do Enem mostrou-se uma ferramenta poderosa para identificar e compreender essas disparidades. A capacidade de organizar, consolidar e analisar grandes volumes de dados de forma eficiente permite uma visão mais clara e detalhada dos desafios enfrentados pelos candidatos. Com esses *insights*, é possível desenvolver políticas públicas mais eficazes e direcionadas, visando reduzir as desigualdades e promover uma educação mais equitativa e acessível

para todos.

Em suma, este trabalho demonstrou que a análise dos microdados do ENEM através de um *data warehouse* não só facilita a identificação de padrões e tendências, mas também fornece uma base sólida para a formulação de estratégias que possam melhorar o acesso e a qualidade da educação no Brasil. A continuidade desse tipo de análise é essencial para monitorar o impacto das políticas educacionais e garantir que todos os estudantes tenham as mesmas oportunidades de sucesso.

Como trabalhos futuros, pretende-se expandir a base de dados para incluir os microdados das demais edições do ENEM que estiverem disponíveis. Essa ampliação permitirá uma análise ainda mais abrangente e detalhada, possibilitando a identificação de tendências e padrões ao longo de um período maior. Além disso, planeja-se a aplicação de algoritmos de mineração de dados para extrair *insights* mais profundos e identificar correlações ocultas nos dados. Técnicas avançadas de mineração de dados, como aprendizado de máquina e análise preditiva, poderão ser utilizadas para prever o desempenho dos candidatos com base em variáveis socioeconômicas e educacionais, bem como para identificar fatores que contribuem para a evasão escolar e a ausência no exame. Essas abordagens não apenas enriquecerão a análise dos dados, mas também fornecerão subsídios valiosos para a formulação de políticas educacionais mais eficazes e direcionadas, visando reduzir as desigualdades e promover uma educação mais equitativa e acessível para todos os estudantes.

Referências Bibliográficas

BRASIL. Governo Federal. *Sistema de Seleção Unificada (Sisu)*. 2023. Acesso em: 31 jul. 2024. Disponível em: <https://www.gov.br/mec/pt-br/areas-de-atuacao/es/sisu>.

BRASIL. Ministério da Educação. *Portaria nº 438, de 28 de maio de 1998*. 1998. Disponível em: <http://www.crmariocovas.sp.gov.br/pdf/diretrizes_p0178-0181_c.pdf>. Acesso em: 10 jun. 2024.

BRASIL. Ministério da Educação. *Cartilha do Participante*. 2018. Acesso em: 31 jul. 2024. Disponível em: <https://download.inep.gov.br/educacao_basica/enem/guia_participante/2018/manual_de_redacao_do_enem_2018.pdf>.

CYGANCZUK, M. de S.; PINTO, J. S. de P.; BASTOS, J. T. Aplicação da mineração de dados na análise de sinistros de trânsito envolvendo colisões no transporte rodoviário de cargas no paraná. *Revista Contemporânea*, v. 3, n. 11, p. 20915–20936, 2023.

FEIJÓ, J. R.; FRANÇA, J. M. S. D.; PINHO, V. R. D. Desempenho dos estudantes ao final do ensino médio: Mensurando a influência direta e indireta da educação dos pais. *Revista Brasileira de Economia*, SciELO Brasil, v. 76, n. 1, p. 30–56, 2022.

FRANÇA, S. de O.; ALVES, K. K.; DUARTE, A. L. C. A utilização do índice de desenvolvimento da educação básica (ideb) pelos gestores escolares: Desafios da qualidade da educação. *Revista Ibero-Americana de Estudos em Educação*, p. 2706–2722, 2022.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. *Microdados do Enem 2022*. Brasília: Inep: [s.n.], 2023. Disponível em: <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem>. Acesso em: 10 jun. 2024.

ISLAM, K. N. e Boishakhe Islam Shova e Tahmina Ria e Humayara Binte Rashid e A. Mineração de dados educacionais para prever o desempenho dos alunos. *Education and Information Technologies*, v. 26, p. 6051 – 6067, 2021.

ISLAM, M. Data analysis: Types, process, methods, techniques and tools. *International Journal on Data Science and Technology*, 2020.

KIM, J.; LEE, R. *Data science and digital transformation in the fourth industrial revolution*. [S.l.]: Springer Nature, 2021. v. 929.

KIRKPATRICK, K. Wha is data mining. *Data Mining for the Social Sciences*, 2019.

KOEDINGER, K.; D'MELLO, S.; MCLAUGHLIN, E.; PARDOS, Z.; ROSé, C. Data mining and education. *Wiley interdisciplinary reviews. Cognitive science*, v. 6 4, p. 333–353, 2015.

LIMA, C. C. V. d.; BRIGHENTI, C. R. G. Desempenho de estudantes de minas gerais no exame nacional do ensino médio considerando variáveis socioeconômicas. *Educação e Pesquisa*, SciELO Brasil, v. 49, p. e253303, 2023.

LIMA, P. d. S. N.; AMBRÓSIO, A. P. L.; FERREIRA, D. J.; BRANCHER, J. D. Análise de dados do enade e enem: uma revisão sistemática da literatura. *Avaliação: Revista da Avaliação da Educação Superior (Campinas)*, SciELO Brasil, v. 24, p. 89–107, 2019.

MASCHIO, P.; VIEIRA, M. A.; COSTA, N.; MELO, S. de; JÚNIOR, C. P. Um panorama acerca da mineração de dados educacionais no brasil. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2018. v. 29, n. 1, p. 1936.

Maximiano, Caio Fernandes Chaves. Monografia (Graduação), *Uso de tecnologias de big data para processamento e análise de dados da área da saúde do estado de São Paulo*. Sorocaba : [s.n.], 2023.

NAKAZONE, E.; BORTOLOTTI, L. M. Análise de dados históricos do enem entre 2015 à 2019. In: *Congresso de Tecnologia-Fatec Mococa*. [S.l.: s.n.], 2021. v. 4, n. 1.

NETO, R. de D. M.; MEDEIROS, H. A. V.; PAIVA, F. da S.; SIMõES, J. L. O impacto do enem nas políticas de democratização do acesso ao ensino superior brasileiro. In: *Comunicaçõe*. [S.l.]: Especial, 2014. v. 21, p. 109–129.

RODRIGUES, R. L.; RAMOS, J. L. C.; SILVA, J. C. S.; GOME, A. S. A literatura brasileira sobre mineração de dados educacionais. *Anais do Congresso Brasileiro de Informática na Educação*, v. 3, 2014.

ROMERO, C.; VENTURA, S. Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, v. 3, 2013.

ROMERO, C.; VENTURA, S. Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, v. 10, 2020.

SETZER, V. W. Os meios eletrônicos e a educação: Uma visão alternativa. In: *Coleção Ensaios Transversais*. São Paulo, Brasil: Editora Escrituras, 2001. v. 10.

SHENKOYA, T.; KIM, E. Sustainability in higher education: digital transformation of the fourth industrial revolution and its impact on open knowledge. *Sustainability*, MDPI, v. 15, n. 3, p. 2473, 2023.

Thiago de Oliveira Souza. Monografia (Graduação), *ANÁLISE DE DADOS: UM ESTUDO DO PERFIL DOS PARTICIPANTES DO ENEM 2019*. Mossoró: [s.n.], 2021.

TRAVITZKI, R.; FERRÃO, M. E.; COUTO, A. P. Desigualdades educacionais e socioeconómicas na população brasileira pré-universitária: uma visão a partir da análise de dados do enem. *Education Policy Analysis Archives*, v. 24, p. 74–74, 2016.