

Minerando o ENEM: análise dos microdados do Exame Nacional do Ensino Médio a partir de técnicas de KDD

Guilherme Brizzi¹, Mathias Eckert Recktenvald¹, Ana Lilian Alfonso Toledo¹

¹Ciência da Computação – Universidade Federal de Santa Maria (UFSM)
Santa Maria – RS – Brazil

gbrizzi@inf.ufsm.br, merecktenvald@inf.ufsm.br, altoledo@inf.ufsm.br

Resumo. *O presente trabalho realiza uma análise exploratória e descritiva dos microdados do Exame Nacional do Ensino Médio (ENEM) por meio de técnicas de Knowledge Discovery from Data – KDD. Foram utilizados dados de diferentes edições do exame, abrangendo informações de desempenho, contexto socioeconômico e origem escolar dos participantes. As etapas envolveram o pré-processamento e tratamento das variáveis, aplicação de regressões lineares, algoritmos de agrupamento (clustering) e algoritmos de associação, além da elaboração de visualizações geográficas e estatísticas. Os resultados revelaram diversas informações relativas ao exame em si e também à realidade da sociedade brasileira e seu sistema educacional. Concluiu-se que os microdados do ENEM constituem uma fonte robusta para estudos quantitativos sobre educação, possibilitando a formulação de políticas públicas mais embasadas.*

1. Introdução

Com milhões de participantes toda edição, o Exame Nacional do Ensino Médio (ENEM) é uma das maiores e mais importantes provas do mundo. Sua nota é uma meta para os estudantes, visto que é usada para diversos programas de acesso ao ensino superior, como o SiSU, Prouni, FIES e outros.

O ENEM é tradicionalmente dividido em cinco provas, correspondentes a cada área do conhecimento: Linguagens, Códigos e suas Tecnologias, Ciências Humanas e suas Tecnologias, Ciências da Natureza e suas Tecnologias, Matemática e suas Tecnologias e Redação.¹

Todos os anos, após a realização e correção das provas e entrega dos boletins de desempenho, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) – organizador do ENEM – divulga os chamados “microdados”. Esse conjunto contém os dados referentes a cada candidato que prestou a prova, conforme descrito na Seção 2. Dessa forma, é inegável que os microdados do ENEM se apresentam como uma fonte rica de dados dos quais se pode extrair conhecimento útil acerca da educação brasileira.

A partir desse paradigma, o presente trabalho objetiva realizar uma análise dos microdados do ENEM a partir do processo de *Knowledge Discovery from Data* (KDD). Foram propostos questionamentos sobre o comportamento do desempenho dos candidatos e a estruturação da prova e, com as técnicas de mineração de dados, buscou-se respondê-los de forma embasada.

¹Para fins de abreviação, essas áreas serão referenciadas, respectivamente, por: Linguagens, Ciências Humanas, Ciências da Natureza, Matemática e Redação.



Figura 1. Cadernos de prova do ENEM

2. Descrição dos dados

Os microdados do Exame Nacional do Ensino Médio (ENEM) são disponibilizados anualmente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [INEP 2024]. Trata-se de uma das bases públicas mais extensas e detalhadas sobre o desempenho educacional e o perfil socioeconômico dos estudantes brasileiros. O conjunto reúne informações individuais de todos os participantes do exame em formato tabular, distribuído em arquivos CSV, acompanhados por um dicionário de variáveis que descreve o significado e a codificação de cada campo.

Cada registro da base corresponde a um participante, identificado de forma anônima, e cada coluna representa uma variável relacionada às provas, ao desempenho, ao perfil socioeconômico ou às características escolares. Entre as principais dimensões contempladas destacam-se:

- **Identificação e localização:** unidade da federação e o município de realização da prova, permitindo análises regionais e geográficas;
- **Desempenho nas provas:** notas individuais obtidas em cada uma das cinco provas – Linguagens, Ciências Humanas, Ciências da Natureza, Matemática e Redação. As notas objetivas são calculadas com base na Teoria de Resposta ao Item (TRI), o que possibilita comparabilidade entre edições do exame;
- **Contexto socioeconômico:** respostas ao Questionário Socioeconômico (QSE), que abrange informações sobre renda familiar, escolaridade dos pais ou responsáveis, posse de bens e serviços, condições de moradia, hábitos de estudo e acesso a recursos tecnológicos [INEP 2024];
- **Informações escolares:** tipo de escola (pública ou privada), rede administrativa (estadual, municipal ou federal), modalidade de ensino e participação em políticas afirmativas, como cotas e programas de acesso ao ensino superior;
- **Aspectos operacionais da aplicação:** presença ou ausência em cada dia de exame, uso de atendimento especializado, idioma escolhido na prova de língua estrangeira e situação de conclusão do ensino médio.

A amplitude e o nível de detalhamento dos microdados – frequentemente ultrapassando milhões de registros e centenas de variáveis por edição – tornam essa base um recurso fundamental para estudos estatísticos e aplicações em aprendizado de máquina voltadas à educação. No entanto, sua utilização requer etapas rigorosas de pré-processamento, dada a presença de valores ausentes, codificações heterogêneas e inconsistências entre diferentes anos. Ademais, a estrutura das variáveis sofre alterações a cada edição, o que demanda padronização prévia para estudos comparativos.

Grande parte das variáveis é categórica ou ordinal, sobretudo no questionário socioeconômico, e são representadas por códigos literais, que correspondem a faixas ou níveis. Assim, a interpretação adequada depende do uso conjunto com o dicionário de microdados disponibilizado pelo INEP. Após o tratamento, essas variáveis podem ser transformadas em representações numéricas ou binárias para utilização em modelos estatísticos e preditivos.

Devido à sua natureza pública, granular e abrangente, os microdados do ENEM podem ser empregados em pesquisas sobre desigualdade educacional, efetividade escolar, modelagem de desempenho e análises de políticas públicas [Oliveira 2021]. Entretanto, a manipulação desse volume de informações gera desafios adicionais, e, portanto, em todas implementações, os autores necessitaram realizar otimizações, sobretudo de uso de memória, nos *scripts* desenvolvidos.

Os microdados de todas edições estão disponíveis publicamente para *download* em <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem>.

3. Problemas propostos

Foram propostos cinco problemas para análise:

- Correlação de desempenho entre áreas do conhecimento
- Impacto de cada fator socioeconômico no desempenho
- Agrupamento e comparação com classificações conhecidas
- Análise geográfica da desigualdade escolar
- Regras de associação entre questões

3.1. Correlação de desempenho entre áreas do conhecimento

É comumente atribuída a Albert Einstein a fala “Todo mundo é um gênio. Mas se você julgar um peixe por sua capacidade de subir em uma árvore, ele vai gastar toda a vida acreditando que é estúpido.”

Contudo, não há evidências de que Einstein a proferiu. [Quote Investigator 2013]

Independentemente disso, a fala nos induz a um questionamento interessante sobre a setorização do aprendizado. Há aqueles que nasceram para a matemática e outros para a história? Há os que serão grandes cientistas, falhando miseravelmente na escrita?

O problema proposto trata exatamente isso: bom desempenho em uma área no ENEM tem correlação com bom desempenho em outra?

3.1.1. Metodologia

Para esta subseção, foram utilizados os microdados das edições de 2020 até 2024. Foram removidos os participantes que não haviam participado de alguma prova ou que haviam obtido nota zero, a fim de remover inconsistências que distorceriam os resultados.

Então, criou-se um *script* que realiza a regressão linear usando mínimos quadrados. Calculou-se os valores de r , r^2 e a função que define a regressão. Devido ao grande volume de dados, optou-se por construir visualizações baseadas em mapas de calor em vez de dispersão. Nos gráficos, a cor amarela representa uma frequência mais alta, e roxo uma frequência zero ou muito baixa.

Para executar o *script* em questão, basta usar o comando `python area_regression.py`.

3.1.2. Resultados

A Figura 2 mostra mapas de calor representando a frequência das combinações de notas entre diferentes áreas, par a par. A linha branca representa a modelagem da regressão linear. Abaixo de cada gráfico, estão dispostos os valores r e r^2 , além da função $f(x) = ax + b$ resultante da regressão.

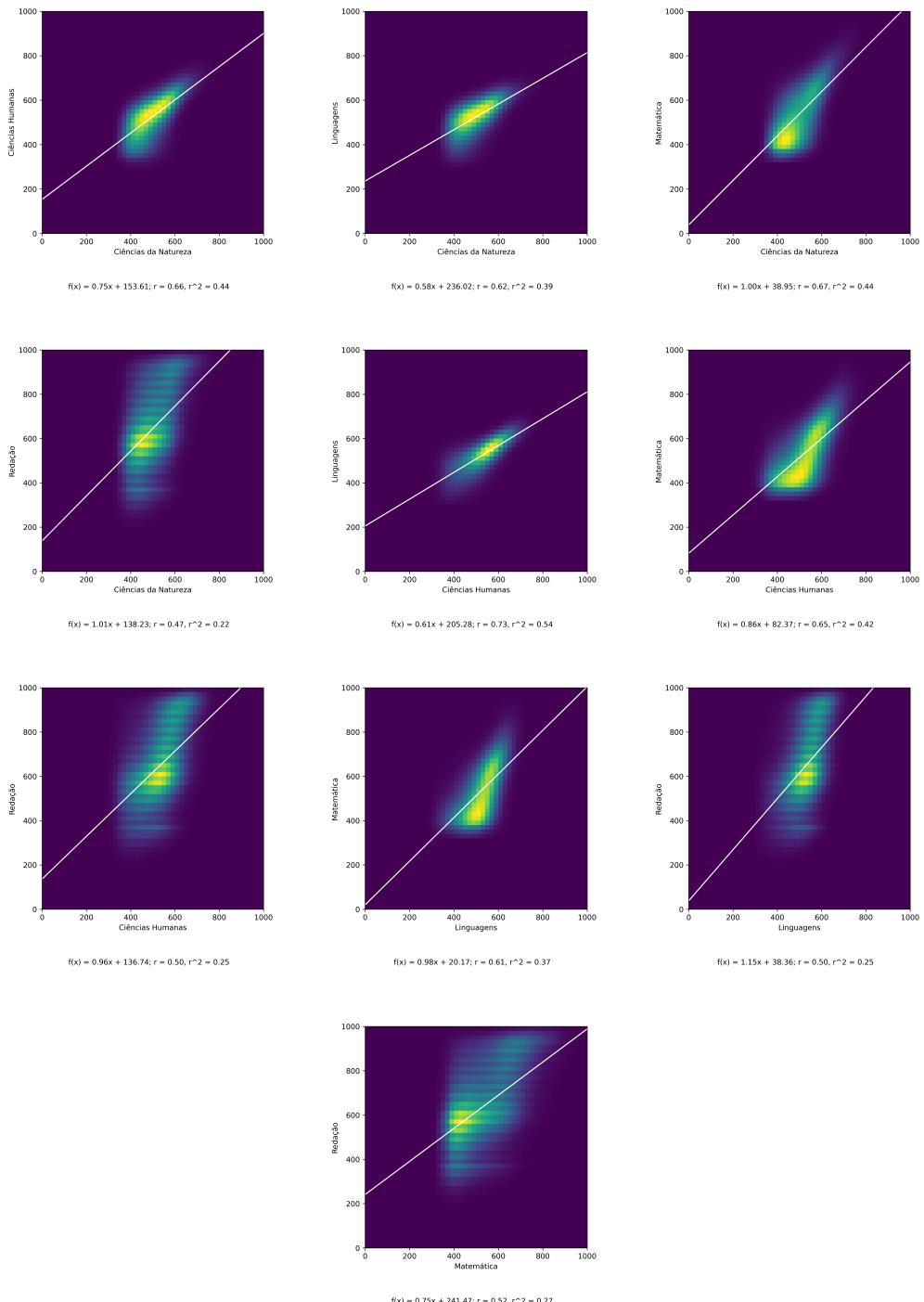


Figura 2. Regressão e mapa de calor par a par entre as provas do ENEM.

Tomando um valor de $r > 0.65$ como relevante, tem-se uma correlação relevante entre os seguintes pares, do maior valor de r ao menor:

- $r = 0.73$ – Ciências Humanas e Linguagens
- $r = 0.67$ – Ciências da Natureza e Matemática
- $r = 0.66$ – Ciências da Natureza e Ciências Humanas

Para aqueles que já realizaram a prova, a correlação forte entre Ciência Humanas e Linguagens é de trivial explicação. Ambas as provas são estruturadas de maneira similar, baseadas em leitura e interpretação. As correlação entre Ciências da Natureza e Matemática também é esperada, ao passo que a área de Natureza inclui química e física, que, muitas vezes, fazem uso da matemática. Já a correlação entre Ciências da Natureza e Ciências Humanas é mais inusitada. À primeira vista, essas são áreas bastante distintas. Contudo, é possível teorizar que a correlação deriva do fato de a prova de Ciências da Natureza também se engajar em questões sociais, como a ecologia, e de a prova de Ciências Humanas também tratar da natureza, como em tópicos de geografia física.

As correlações mais fracas foram todas da prova de Redação com as demais provas:

- $r = 0.47$ – Ciências da Natureza e Redação
- $r = 0.50$ – Linguagens e Redação
- $r = 0.50$ – Ciências Humanas e Redação
- $r = 0.52$ – Matemática e Redação

Esse resultado pode ser explicado pelo fato de a prova de redação possuir uma estrutura muito distinta das demais. Ela é a única prova dissertativa do ENEM, sendo corrigida por corretores humanos, de acordo com diferentes competências pré-determinadas. Assim, quanto os conhecimentos de Linguagens, por exemplo, possam ser aplicados na Redação, os resultados obtidos parecem mostrar que a estrutura dela é tão distinta, que um bom desempenho nas provas objetivas não se traduz necessariamente em um bom desempenho na redação.

De um modo geral, os resultados obtidos descreditam parcialmente o ditado atribuído ao renomado cientista. Nota-se que, no contexto do ENEM, não há grande especialização dos conhecimentos e há correlação clara e positiva entre uma boa performance em certa área com boa performance em outras. Sob essa ótica, a análise feita nesta subseção demonstra que há uma certa uniformidade no desempenho escolar: estudantes que tem bom aprendizado tendem a ter bom aprendizado em tudo e estudantes que tem mau aprendizado tendem a ter mau aprendizado em tudo. Esse conhecimento é de suma importância, pois a compreensão sobre o aprendizado molda os métodos de ensino que são usados pelos educadores, tendo grande impacto na educação.

3.2. Impacto de cada fator socioeconômico no desempenho

Durante a inscrição, os candidatos do ENEM são convidados a responder o questionário socioeconômico. Similar a perguntas de um censo, o questionário busca colher dados para entender a realidade dos estudantes.

3.2.1. Metodologia

Até a edição de 2023, os microdados traziam as respostas de cada estudante juntamente ao seu desempenho. Em 2022, o questionário foi composto com questionamentos diferentes, referentes às práticas de estudos dos participantes durante a pandemia de COVID-19. Já na edição 2024, os microdados foram divulgados de forma mais anonimizada, em que as respostas do questionário não estavam associadas a um estudante, mas apenas à sua escola. Dessa forma, para esta subseção, foram utilizados os microdados das edições de 2020, 2021 e 2023, os quais são consistentes.

Ademais, foram removidos os participantes que não haviam participado de alguma prova ou que haviam obtido nota zero, a fim de remover inconsistências que distorceriam os resultados.

Como os valores das respostas do questionário socioeconômico são valores textuais, discretos e qualitativos, atribuiu-se valores numéricos a cada classe de cada pergunta. Isso é possível pelo fato de que há uma clara relação hierárquica entre as classes. Veja-se as possibilidades de escolaridade como exemplo na Tabela 1.

Tabela 1. Escolaridade do pai ou responsável

Pergunta	Até que série seu pai, ou o homem responsável por você, estudou?
Opções	<ol style="list-style-type: none">1. Nunca estudou.2. Não completou a 4ª série/5º ano do Ensino Fundamental.3. Completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental.4. Completou a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio.5. Completou o Ensino Médio, mas não completou a Faculdade.6. Completou a Faculdade, mas não completou a Pós-graduação.7. Completou a Pós-graduação.8. Não sei.

Excetuando-se a resposta “*Não sei.*”, é possível realizar uma ordenação clara, indo do menor grau de estudo até o maior. Em outras perguntas do questionário, há grande similaridade quanto à hierarquia das possibilidades de resposta. Tendo em vista que respostas “*Não sei.*” e similares não se encaixam em uma ordenação, elas foram excluídas.

Ademais, algumas perguntas do questionário estão claramente anacrônicas, como “Na sua residência tem aparelho de DVD?” ou “Na sua residência tem telefone fixo?”. Devido à mudanças tecnológicas, tais equipamentos já se tornaram, em sua maioria, superados. Por isso, essas perguntas foram excluídas.

A partir disso, de forma análoga a subseção anterior, criou-se um *script* que realiza a regressão linear usando mínimos quadrados. Calculou-se os valores de r , r^2 e a função que define a regressão.

Para executar o *script* em questão, basta usar o comando `python socioeconomic_regression.py`.

Os gráficos foram plotados usando *boxplots*, em que a variação de dados pode ser representada usando quartis. As linhas das bordas representam os limites superior e inferior. As bordas do retângulo representam o primeiro e o terceiro quartis. A linha dentro do retângulo representa a mediana.

3.2.2. Resultados

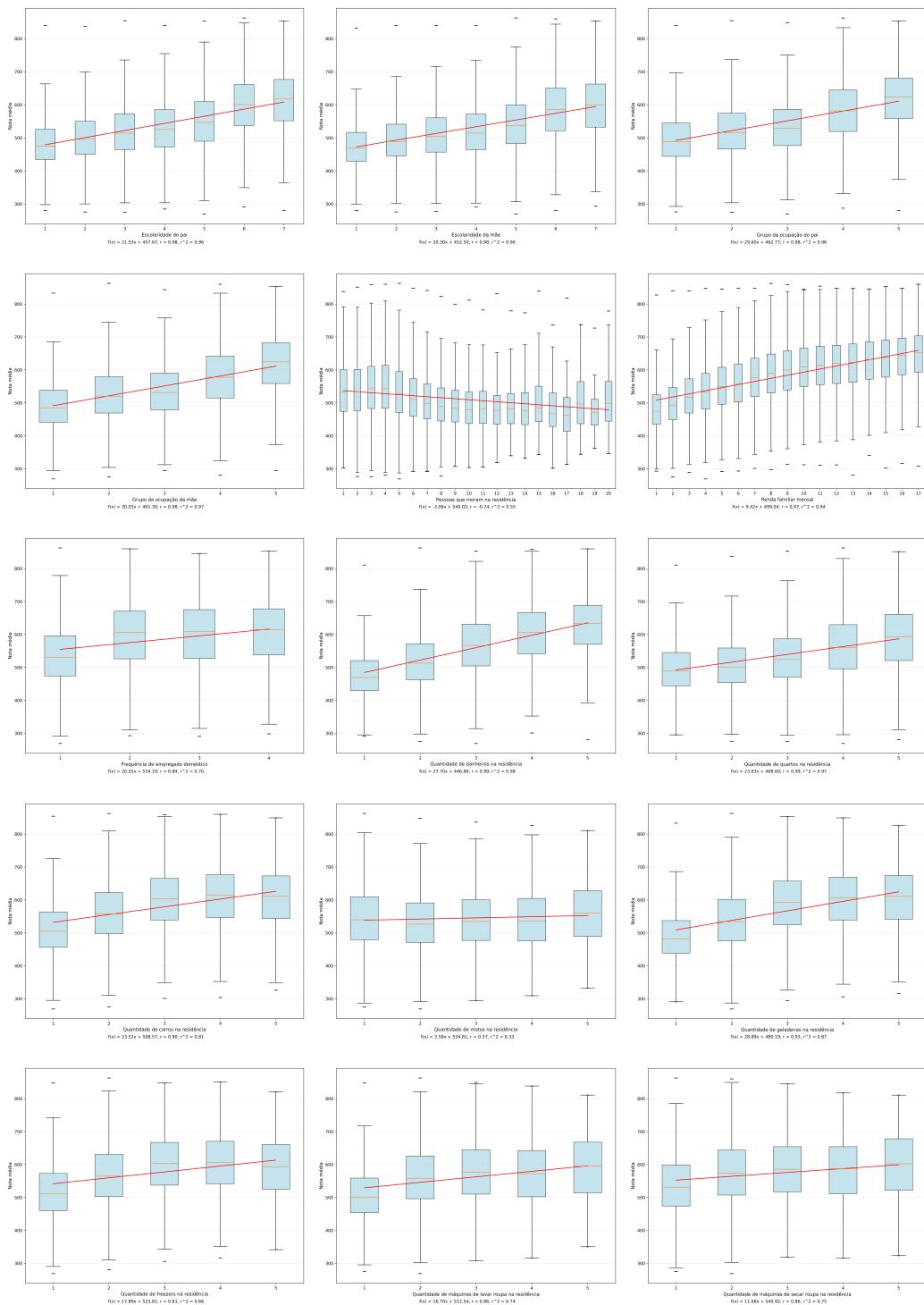


Figura 3. Box plot e regressão entre questionário socioeconômico e nota

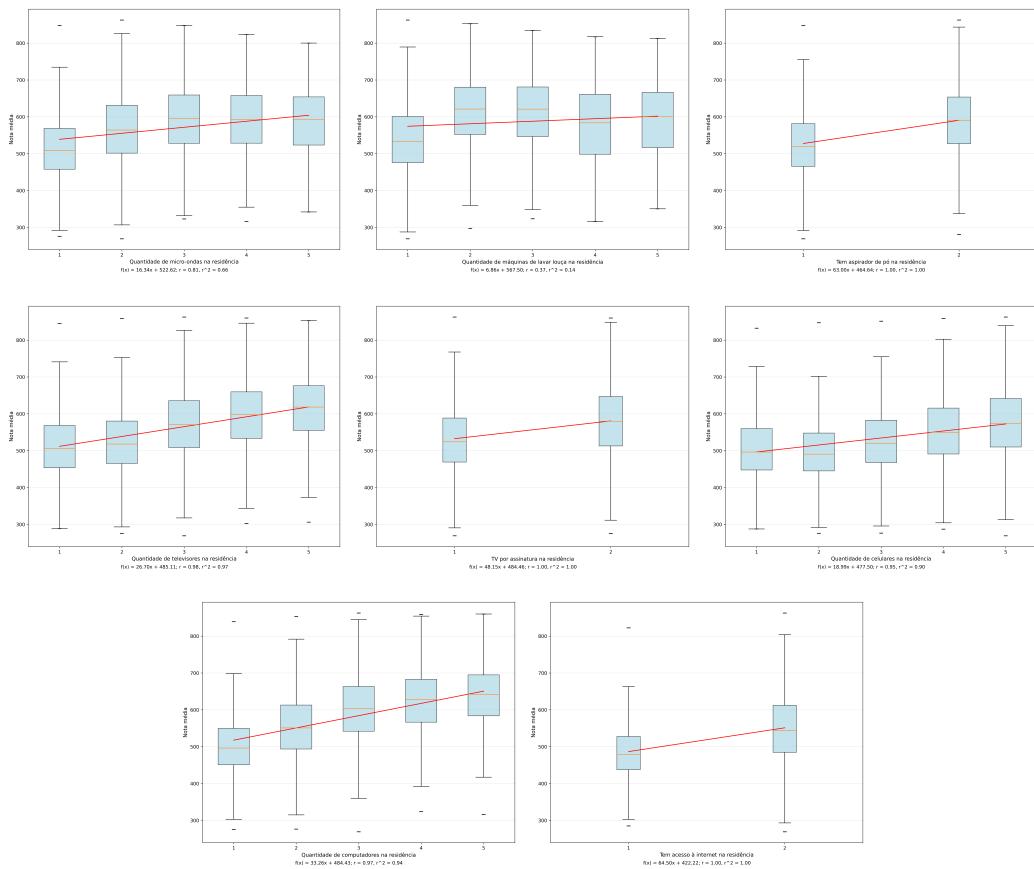


Figura 4. Box plot e regressão entre questionário socioeconômico e nota (continuação)

Nota-se que há forte e visível correlação entre fatores socioeconômicos e a nota do ENEM dos participantes. Todas as diferentes perguntas, na verdade, são reveladoras da mesma realidade subjacente: o impacto da desigualdade social no acesso à educação. Por mais que existam *outliers* no dataset, é evidente que a condição socioeconômica é um fator com forte correlação positiva com a performance nas provas.

Tal realidade instiga reflexões necessárias: a prova do ENEM é um *proxy* para uma prova de renda e patrimônio? Quais abordagens podem ser tomadas para corrigir e equalizar a distribuição? Conquanto imprescindíveis, essas perguntas se encontram além do escopo do presente trabalho – e da presente área do conhecimento.

3.3. Agrupamento e comparação com classificações conhecidas

Agrupamento e classificação são vistos como tarefas distintas, mas que compartilham de similaridades. Uma classificação de um conjunto, assim como um agrupamento, resultam em grupos que compartilham de similaridades. [Assunção 2021]

Nesta seção, buscou-se aplicar algoritmos de *clustering* aos microdados do ENEM e comparar seus resultados a classificações já conhecidas, trazidas pelos dados.

Foi explorada a seguinte relação:

Tabela 2. Relação entre *features* de *clustering* e classificação comparada

Features de clustering	↔	Classificação comparada
Perfil socioeconômico aluno	↔	Tipo de escola

3.3.1. Metodologia

O pré-processamento usado foi o mesmo descrito previamente na subseção 3.2.

Para aplicação do *clustering*, primeiro foi padronizada a escala dos dados, a fim de mitigar distorções causadas por magnitudes numéricas distintas. Após, utilizou-se o *KMeans* como algoritmo de agrupamento para a geração de *clusters* baseados puramente nas respostas do questionário socioeconômico. Com isso, obteve-se dois grupos.

Como a multidimensionalidade do espaço vetorial impossibilitava a visualização prática dos *clusters*, foi utilizado PCA para redução de dimensionalidade e o resultado foi plotado em gráfico de dispersão. Após, verificou-se para cada item de cada *cluster* se a escola era de tipo pública ou privada.

Para executar o *script* em questão, basta usar o comando `python clustering.py`.

3.3.2. Resultados

A Figura 5 mostra os *clusters* gerados, no espaço R^2 , reduzido por PCA.

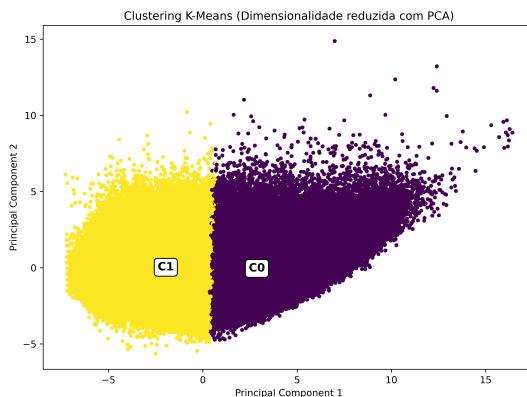


Figura 5. Clustering por fatores socioeconômicos

A Tabela 3 mostra a composição de cada *cluster* por tipo de escola.

Tabela 3. Distribuição de escolas públicas e privadas por cluster socioconômico.

Cluster	% Públcas	% Privadas
C0	34.45%	65.55%
C1	85.75%	14.25%

Nota-se claramente que um agrupamento não-supervisionado, usando apenas dados socioeconômicos dos candidatos, conseguiu gerar *clusters* que distinguem visivelmente o tipo de escola que os candidatos estudaram. Tal relação demonstra a correlação clara entre o acesso ao estudo e as condições socioeconômicas – a qual, por mais que não seja surpreendente, fica ainda mais evidente com a abordagem utilizada.

3.4. Análise geográfica da desigualdade escolar

A presente subseção visará a explorar questões, a nível de unidade federativa (UF), relativas ao desempenho no ENEM e seu retrato da desigualdade escolar no Brasil. Delinear-se-ão as seguintes análises:

- Média por estado
- Média por estado por tipo de escola
- Desigualdade relativa entre médias das escolas públicas e privadas por estado

3.4.1. Metodologia

Para esta seção, foram utilizados todos os dados do período 2020 a 2024.

Criou-se um *script* que computa a média por estado, por tipo por estado e a diferença relativa entre as públicas e privadas. A diferença foi computada por $100 \times (nota_{privada}/nota_{pública} - 1)$, representando o percentual em que a média das privadas é maior que a média das públicas.

A fim de melhorar a visualização, os resultados obtidos foram usados para a construção de mapas usando a biblioteca *geopandas* e as cartas disponibilizadas pelo IBGE. Para os dados completos, o formato tabular também está disponível.

Para executar o *script* em questão, basta usar o comando `python mapping.py`.

3.4.2. Resultados

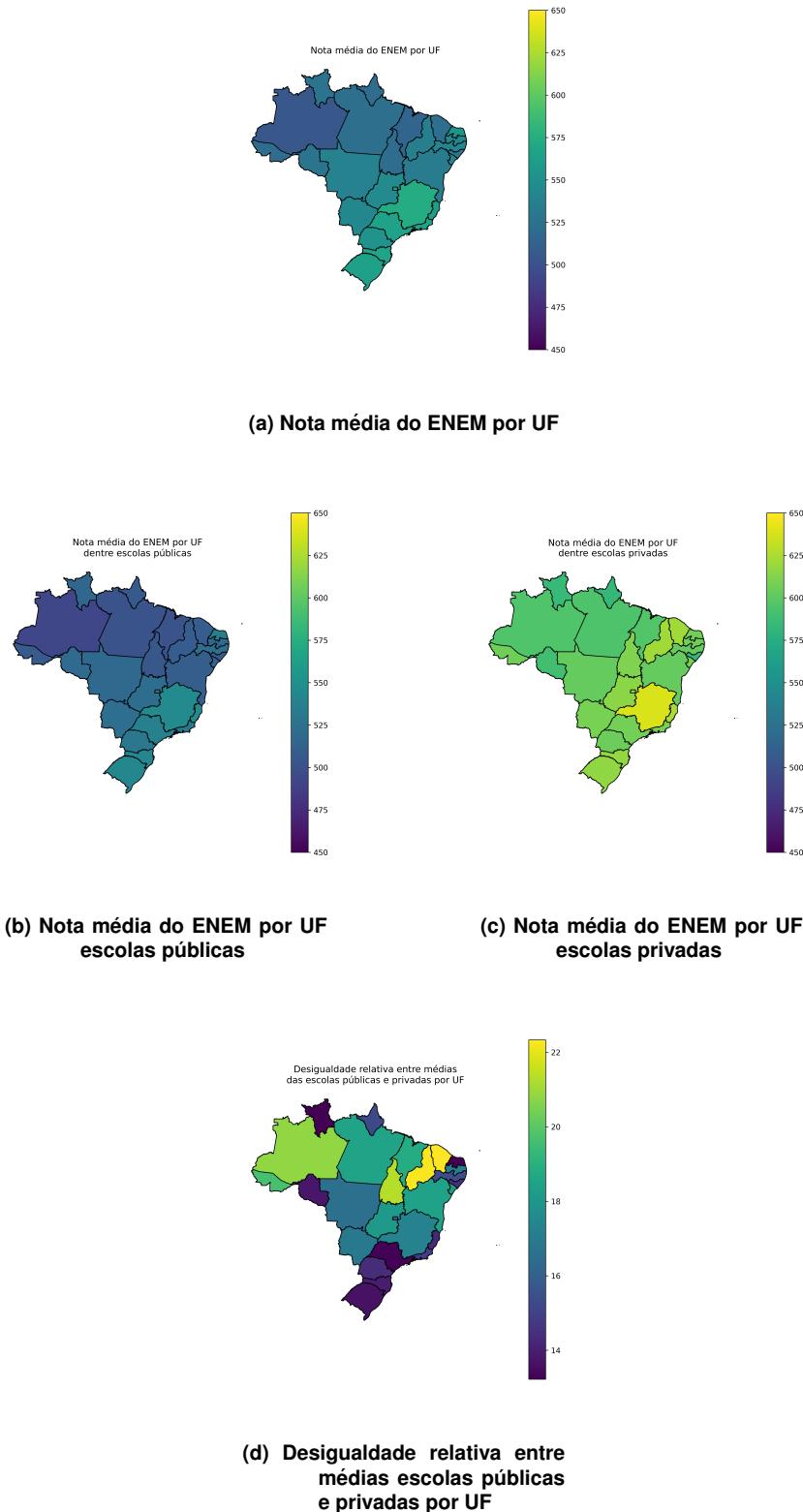


Figura 6. Distribuição de notas médias do ENEM por estado

A informação completa dos mapas também foi disposta em formato tabular. A Tabela 4 mostra as médias gerais por UF e a Tabela 5 mostra demais indicadores relativos à discrepância entre escolas públicas e privadas, por UF.

Tabela 4. Média por UF

Estado	Média
Minas Gerais	573.40
Rio de Janeiro	566.48
Santa Catarina	565.93
São Paulo	565.24
Rio Grande do Sul	563.04
Distrito Federal	562.92
Espírito Santo	562.39
Rio Grande do Norte	554.99
Paraná	552.81
Goiás	544.91
Mato Grosso do Sul	543.50
Sergipe	543.27
Paraíba	541.82
Pernambuco	541.28
Mato Grosso	538.47
Piauí	535.86
Bahia	533.47
Alagoas	527.35
Rondônia	526.63
Roraima	526.22
Pará	522.50
Ceará	521.11
Tocantins	520.43
Amapá	520.19
Acre	519.57
Maranhão	514.73
Amazonas	504.23

Tabela 5. Demais indicadores por UF

UF	Diferença privada/pública (em % maior das privadas)	Média públicas	Média privadas
Ceará	22.33	508.10	621.58
Piauí	22.26	508.18	621.29
Tocantins	21.30	504.25	611.66
Amazonas	20.81	492.70	595.25
Acre	19.65	506.57	606.10
Maranhão	18.84	501.56	596.07
Pará	18.51	501.18	593.94
Bahia	18.48	508.92	602.97
Sergipe	18.20	514.58	608.20
Goiás	18.13	520.59	614.99
Minas Gerais	17.22	544.63	638.43
Paraíba	17.12	517.93	606.58
Mato Grosso do Sul	16.90	521.12	609.16
Mato Grosso	16.55	517.22	602.81
Distrito Federal	15.76	531.26	615.01
Pernambuco	15.37	522.65	602.98
Amapá	15.22	506.06	583.09
Rio de Janeiro	14.80	531.11	609.71
Alagoas	14.48	508.09	581.68
Paraná	14.32	528.94	604.71
Espírito Santo	14.24	546.71	624.54
Santa Catarina	13.95	541.77	617.32
Rondônia	13.76	518.18	589.46
Rio Grande do Sul	13.64	542.66	616.67
São Paulo	13.27	536.22	607.39
Rio Grande do Norte	13.22	535.72	606.56
Roraima	13.22	516.18	584.43

Os resultados desta subseção trazem diversas informações relevantes.

Quanto às médias por estado, que já são divulgadas por diversos veículos da mídia, não há novidade: há ainda desigualdade regional na educação brasileira, com os estados mais ricos, da região concentrada, se sobressaindo sobre os demais.

Entretanto, os dados revelados pela análise de desigualdade relativa entre médias das escolas públicas e privadas por estado trazem resultados mais profundos. Nota-se grande variação entre estados quanto à desigualdade públicas/privadas. Na primeira colocação do *ranking*, está o Ceará – estado conhecido por suas grandes escolas que lideram o *ranking* do ENEM [Lucca 2025], mas que, aparentemente, tem enorme abismo entre sua rede pública e privada. Isso retrata um fator com o qual se deve ter cuidado: desempenhos exemplares de certas escolas e alunos fazem as manchetes, mas não revelam toda a história da educação. No caso do Ceará, o motivo de orgulho por seus ótimos colégios esconde o retrato da desigualdade de 22% entre sua rede de educação pública e sua rede privada.

De modo geral, todos os estados apresentaram discrepância maior que 10% entre essas médias, o que é um valor bastante significativo, demonstrando que a educação pública ainda tem dificuldade em alcançar os níveis de desempenho obtidos pela rede privada.

3.5. Regras de associação entre questões

Nesta subseção, buscou-se identificar se errar uma determinada questão está estatisticamente associado a errar outra. O foco é extrair regras de associação (ex: se um estudante erra a Questão A, qual a probabilidade de ele também errar a Questão B?) que revelem dependências entre questões, possíveis blocos de habilidades ou competências transversais que impactam o desempenho em diferentes áreas.

3.5.1. Metodologia

1. **Preparação dos Dados:** Os microdados foram transformados em uma matriz boleana em que cada linha representa um estudante, e cada coluna, uma questão, com o valor *True* indicando um erro naquela questão, e *False*, um acerto.
2. **Mineração de Regras:** Utilizou-se o algoritmo Apriori para identificar “*itemsets* frequentes”.
3. **Geração e Filtragem de Regras:** A partir desses conjuntos, foram geradas regras de associação. Essas regras foram então filtradas e avaliadas usando métricas estatísticas para garantir que apenas os padrões mais significativos fossem retidos:
 - **Confiança:** A métrica principal. Indica a força preditiva da regra (ex: a probabilidade de errar B, dado que já errou A).
 - **Suporte:** A frequência com que o padrão (erro em A e B) aparece no conjunto total de dados.
 - **Lift:** Mede o quão mais provável é a coocorrência dos erros em comparação com a probabilidade de eles ocorrerem de forma independente. Um *Lift* > 1 indica uma correlação positiva.

Os resultados pré-calculados estão anexos. Para executar o *script* em questão, basta usar o comando `python error_cooccurrence.py`.

3.5.2. Resultados

Foram analisadas manualmente as principais regras de cada ano, das quais surgiram dois padrões passíveis de destaque.

Padrão 1: “Blocos de Dificuldade” (Intra-área)

Em todos os anos, é visível a formação de blocos de questões, geralmente dentro da mesma área do conhecimento.

Um bloco recorrente é o de questões de língua estrangeira. Na maioria das provas do ENEM analisadas, essas regras figuram dentre as mais fortes, provavelmente devido ao fato de esse conhecimento estar dentre os mais isolados do ENEM.

De modo geral, as regras sugerem que questões envolvidas em regras de associação medem, na prática, a mesma habilidade ou competência subjacente. Errar uma é fator de risco para errar as outras, indicação de que uma habilidade-chave que é testada de formas muito similares, ou que o grau de dificuldade pode ser similarmente alto.

Padrão 2: “Questões-Hub” (Inter-áreas)

Esta descoberta é mais inusitada. Em vários anos, uma única questão (geralmente de Natureza ou Humanas) atua como um “hub” de dificuldade, aparecendo como o consequente de regras cujos antecedentes vêm de múltiplas áreas do conhecimento.

4. Conclusão

Em síntese, o presente trabalho explora os microdados do ENEM a partir da técnica de KDD para afirmar novas ideias, comprovar as que já sabemos e também contestar as que tomamos como verdade.

Há ainda, sem dúvida, uma quantidade enorme de análises que podem ser feitas, algoritmos que podem ser aplicados e visualizações que podem ser construídas a partir dos microdados do ENEM. Em nossas análises, buscamos elucidar certos questionamentos os quais se apresentaram como interessantes e passíveis de investigação.

Em trabalhos futuros, a título de exemplo, poderia explorar-se os dados demográficos (como gênero, idade, etc), a nível de município, relativo à região urbana/rural, a nível semântico de questões, dentre outros. Ademais, com o fluxo constante de dados chegando a cada edição do ENEM, reaplicação dos métodos usados e comparações históricas podem ser pertinentes.

Dessa forma, os cinco problemas delineados aqui são um convite para maior investigação, não uma lista exaustiva do que os microdados podem revelar.

Em suma, a análise trabalhada gerou informações as quais, interpretadas, são conhecimento útil na caminhada para a melhoria da educação brasileira.

5. Agradecimentos

Agradecemos ao Prof. Dr. Joaquim Vinicius Carvalho Assunção pela oportunidade de realizar este trabalho em tema de nossa escolha e ao INEP pela disponibilização de dados com tamanho detalhe.

Referências

- Assunção, J. V. C. (2021). *Uma breve introdução à mineração de dados: bases para a ciência de dados, com exemplos em R*. Novatec Editora, São Paulo, Brasil.
- INEP (2020, 2021, 2022, 2023, 2024). Microdados do ENEM. <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem>.
- Lucca, Bruno e Leite, L. (2025). Escolas privadas do Ceará se destacam nas notas do ENEM 2024.
- Oliveira, Maria Souza e Costa, L. F. (2021). Análise de desempenho escolar e desigualdade educacional a partir dos microdados do enem. *Revista Brasileira de Estudos Educacionais*, 18(2):55–72.
- Quote Investigator (2013). Quote Origin: Everybody is a Genius. But If You Judge a Fish by Its Ability to Climb a Tree, It Will Live Its Whole Life Believing That It Is Stupid. Quote Investigator (online).