PREDIÇÃO DE DIFICULDADE DAS QUESTÕES DO ENEM

Gabriel Leão, Hudo Leonardo, Tamires Araújo

[1]Departamento de Computação – Universidade Federal Rural de Pernambuco (UFRPE)

Resumo. *Este trabalho tem como objetivo analisar e modelar a dificuldade das questões do Exame Nacional do Ensino Médio (ENEM) a partir do parâmetro de dificuldade (b), utilizando microdados oficiais fornecidos pelo INEP, referentes aos anos de 2019, 2020, 2022 e 2023. Além da análise estatística, propõe-se a construção de um modelo preditivo baseado em aprendizado de máquina capaz de estimar a dificuldade das questões com base em características textuais e visuais extraídas através de técnicas de Processamento de Linguagem Natural (PLN). Contrariando a hipótese inicial, os resultados demonstram que as métricas tradicionais de PLN possuem uma correlação muito fraca com o parâmetro de dificuldade, resultando em modelos com baixa capacidade preditiva. A pesquisa contribui ao estabelecer um importante ponto de partida, evidenciando que a dificuldade de um item do ENEM é um construto complexo que transcende a análise linguística superficial e exige abordagens que incorporem aspectos semânticos e cognitivos mais profundos..*

1. Introdução

No cenário educacional brasileiro, o Exame Nacional do Ensino Médio (ENEM) se consolidou como o principal processo seletivo para o ingresso em universidades públicas e privadas. Aplicado anualmente a milhões de candidatos, o exame tem como uma de suas premissas garantir equidade e comparabilidade nas notas atribuídas aos participantes. Para isso, o ENEM adota a **Teoria de Resposta ao Item** (TRI), um modelo estatístico que permite avaliar não apenas o número de acertos, mas a coerência das respostas com o nível de proficiência do estudante.

A TRI opera por meio de um modelo probabilístico baseado em três parâmetros: discriminação (a), dificuldade (b) e acerto casual (c). A habilidade do item refere-se à competência específica que está sendo avaliada; o parâmetro de dificuldade indica o nível de habilidade necessário para que um candidato tenha 50% de chance de acertar uma questão; a discriminação representa a capacidade do item de diferenciar entre candidatos com diferentes níveis de proficiência; e o acerto casual corrige a chance de acerto por chute. Esses parâmetros possibilitam que as notas sejam comparáveis entre diferentes edições do exame, mesmo que os cadernos de prova contenham questões distintas.

Apesar de os parâmetros da TRI serem tradicionalmente estimados após a aplicação da prova, a partir da análise das respostas de um grande volume de participantes, este projeto propõe uma abordagem complementar, a predição do parâmetro de dificuldade, utilizando apenas o conteúdo textual das questões e informações sobre a presença de elementos visuais. Tal abordagem, se eficaz, representaria uma ferramenta estratégica para a elaboração de provas mais equilibradas e para a preparação de estudantes, permitindo a classificação prévia das questões por nível de dificuldade.

Para alcançar esse objetivo, foram utilizados os microdados oficiais do ENEM dos anos de 2019, 2020, 2022 e 2023, disponibilizados pelo INEP e por plataformas como Kaggle. A partir desses dados, o parâmetro de dificuldade (b) foi adotado como variável-alvo na construção de um modelo de regressão, com o intuito de avaliar a capacidade preditiva do conteúdo da questão sobre seu nível de dificuldade.

Ao analisar a consistência do parâmetro de dificuldade ao longo de diferentes edições da prova e explorar o potencial da inteligência artificial aplicada à educação, o trabalho busca contribuir tanto para a área de mineração de dados educacionais quanto para o aprimoramento dos processos avaliativos no Brasil.

2. Metodologia

O desenvolvimento deste projeto foi estruturado em três grandes etapas: (1) Coleta e Pré-processamento dos Dados; (2) Extração de Características Linguísticas (Feature Engineering); e (3) Modelagem Preditiva. Cada uma dessas etapas foi cuidadosamente planejada para garantir a integridade, consistência e representatividade dos dados utilizados na análise. A escolha metodológica se baseou em investigar a relação entre o conteúdo textual das questões do ENEM e o parâmetro de dificuldade (b) da Teoria de Resposta ao Item (TRI), visando predizer esse parâmetro antes da aplicação do exame.

2.1. Coleta e Pré-processamento dos Dados

Os dados utilizados neste estudo foram provenientes de duas fontes principais: os arquivos estruturados em JSON contendo os textos das provas do ENEM e os microdados oficiais do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), disponibilizados em formato CSV. Foram analisadas as edições do ENEM dos anos de 2019, 2020, 2022 e 2023. As três primeiras compuseram o conjunto de treino, enquanto o ENEM 2023 foi reservado exclusivamente para teste e validação do modelo.

Os arquivos JSON foram obtidos por meio do processo de conversão automatizada das provas originais em PDF utilizando a ferramenta Enem Extractor. Essa ferramenta permite extrair, de forma estruturada, o enunciado das questões e a presença de imagem (indicador binário). Informações preservadas com o intuito de enriquecer a análise e possivelmente contribuir para a performance do modelo.

Os microdados do INEP, por sua vez, forneceram os parâmetros da TRI estimados para cada item, sendo o parâmetro de dificuldade (b) o único considerado neste projeto. Este parâmetro indica o nível de proficiência que um participante precisa ter para possuir 50% de chance de acerto em determinada questão. Os campos CO_POSICAO (nos microdados) e number (nos arquivos JSON) foram utilizados como chaves de ligação entre os dois conjuntos.

O processo de preparação dos dados foi realizado em Python, utilizando a biblioteca pandas. Inicialmente, os dados dos dois dias de prova de cada edição foram carregados e concatenados para formar um único DataFrame por ano. Em seguida, foi feita a junção entre os dados textuais e os microdados do INEP, resultando em um dataset contendo, para cada questão, o texto completo, o parâmetro de dificuldade (b) e a presença de imagem. Foram mantidas apenas as colunas relevantes para a análise, e todas as linhas com dados faltantes em text ou no parâmetro de dificuldade foram removidas.

Durante a coleta, constatou-se que os arquivos referentes à prova de 2021 apresentavam sérios problemas de codificação, com caracteres ilegíveis que comprometiam a extração correta do conteúdo textual. Dado esse comprometimento na qualidade dos dados, optou-se por excluir o ano de 2021 do escopo da pesquisa.

Além disso, a análise exploratória revelou que a quantidade de questões válidas por ano variou entre 173 e 178 itens, sendo inferior ao número esperado de 180. Essa variação se deve à anulação de questões, procedimento comum nas edições do ENEM. Essa observação foi considerada no controle de qualidade dos dados e na definição da base final.

2.2. Extração de Características Linguísticas (Feature Engineering)

A segunda etapa do projeto consistiu na transformação dos textos das questões em vetores numéricos que representassem características linguísticas significativas para o modelo de regressão. Para isso, foi utilizada a biblioteca aibox.nlp, desenvolvida para análise linguística de textos acadêmicos e educacionais em língua portuguesa. Essa biblioteca oferece um conjunto robusto de extratores baseados em estudos linguísticos, psicolinguísticos e cognitivos.

Cada questão foi processada individualmente, e foram extraídas as seguintes dez dimensões de atributos:

- **readabilityBR**: Conjunto de índices de legibilidade como Flesch e Gunning-Fog, que medem a facilidade de compreensão do texto com base na extensão de frases e na complexidade das palavras.
- **textualSimplicityBR**: Métricas que avaliam a simplicidade lexical e estrutural do texto, fundamentais para determinar a acessibilidade da linguagem.
- **syntacticComplexityBR**: Indicadores que medem a complexidade sintática do texto, como o número médio de cláusulas por frase e a profundidade da estrutura frasal.
- **lexicalDiversityBR**: Medidas de diversidade vocabular, incluindo razões tipo-token e outras métricas como o índice de Honoré.
- **connectivesV2BR**: Frequência e variedade de conectivos usados no texto, como adversativos, aditivos, causais e consecutivos, importantes para a coesão discursiva.
- **referentialCohesionBR**: Métricas de coesão referencial, que observam a repetição de termos e a continuidade temática ao longo do texto.
- **sequentialCohesionBR**: Indicadores de coesão sequencial, que avaliam como as ideias estão logicamente conectadas de uma frase para outra.
- **regencyBR**: Detecção de possíveis desvios de regência verbal e nominal, com impacto direto na clareza gramatical do item.
- **conjugationBR**: Análise de erros de conjugação verbal, aspecto relevante para avaliar a correção linguística.
- **liwcBR**: Análise de categorias psicolinguísticas e semânticas com base no LIWC (Linguistic Inquiry and Word Count), que fornece insights sobre o tom emocional e o conteúdo do discurso.

A função de extração retornou, para cada questão, um dicionário de características linguísticas, que foi posteriormente achatado (flattened) e unido ao DataFrame original.

As variáveis categóricas estruturais também foram mantidas e tratadas como potenciais preditoras. O resultado foi um conjunto final de dados composto por dezenas de atributos numéricos prontos para a etapa de modelagem.

2.3. Análise Exploratória e Considerações Iniciais

Antes da modelagem preditiva, uma análise exploratória foi realizada com foco na distribuição do parâmetro de dificuldade (b). Observou-se que a média geral de dificuldade entre os anos analisados foi de aproximadamente 1,36, com valores extremos variando entre -0,74 e 12,41. Questões com valores muito elevados indicam itens extremamente difíceis, enquanto valores negativos estão associados a itens considerados simples dentro da escala da TRI.

3. Resultado

Após a etapa de pré-processamento e extração de características, um conjunto de dados consolidado, abrangendo os anos de 2019, 2020 e 2022, foi utilizado para treinar seis modelos de regressão distintos. O objetivo era avaliar a capacidade das features linguísticas em predizer o parâmetro de dificuldade (b). O conjunto de dados foi dividido em 80% para treino e 20% para teste.

O modelo com melhor desempenho foi o Support Vector Regressor (SVR), que, apesar de superior aos demais, alcançou um R^2 de apenas 0.0566. Este valor, muito próximo de zero, indica que o modelo possui uma capacidade preditiva extremamente baixa, explicando menos de 6% da variância no parâmetro de dificuldade.

Para validar a capacidade de generalização do melhor modelo (SVR), ele foi aplicado ao conjunto de dados da prova de 2023, que não foi utilizado durante o treinamento. Nesta avaliação, o modelo obteve um R^2 de 0.0787, um resultado marginalmente superior, mas que confirma a fraca correlação entre as features extraídas e a dificuldade real dos itens.

Com o intuito de investigar a relação entre as variáveis, foi realizada uma análise de correlação entre cada característica e o parâmetro de dificuldade. As características mais correlacionadas positivamente foram as de coesão referencial com aproximadamente 0.33, enquanto as de diversidade lexical apresentaram as maiores correlações negativas -0.26

Um segundo experimento foi conduzido, treinando os mesmos modelos utilizando apenas as 10 características com maior correlação. Contudo, esta abordagem não resultou em melhoria; pelo contrário, o desempenho geral diminuiu, com o melhor modelo (Lasso) atingindo um R^2 de apenas 0.0254 no conjunto de teste e 0.0198 na validação com os dados de 2023.

4. Conclusão

Este trabalho teve como objetivo desenvolver um sistema capaz de prever a dificuldade de questões do ENEM com base em suas características textuais. No entanto, os resultados indicaram que as métricas tradicionais de Processamento de Linguagem Natural (como legibilidade, complexidade sintática, coesão e diversidade lexical) não foram

suficientes para essa tarefa. Os modelos de regressão treinados apresentaram baixo desempenho, com coeficientes de determinação (R^2) próximos de zero, evidenciando pouca ou nenhuma capacidade preditiva.

A principal conclusão é que a dificuldade de uma questão, representada pelo parâmetro de dificuldade da TRI, envolve fatores mais complexos do que os capturados por análises linguísticas superficiais. Elementos como raciocínio lógico, interpretação de enunciados, conhecimento contextual e a complexidade semântica das alternativas parecem ter um peso maior na percepção e na resolução dos itens.

Embora a hipótese inicial não tenha se confirmado, este estudo contribui ao estabelecer um ponto de partida importante para pesquisas futuras. Os resultados reforçam que a avaliação da dificuldade de questões exige abordagens mais completas, que considerem não apenas o texto, mas também aspectos cognitivos, semânticos e pedagógicos mais profundos

Referências

AIBOX Lab (2024). aibox-nlp: Biblioteca de processamento de linguagem natural para o português brasileiro. https://github.com/aiboxlab/nlp. Acessado em: 29 de julho de 2025.

Dias, L. (2022). enem-extractor: Ferramenta para extração de dados de provas do enem em pdf. https://github.com/diaslui/enem-extractor. Acessado em: 29 de julho de 2025.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) (2025). Microdados do exame nacional do ensino médio (enem). https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem. Acessado em: 29 de julho de 2025.

[AIBOX Lab 2024] [Dias 2022] [Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (I