Universidade Federal do ABC

Projeto de Pesquisa de Iniciação Científica

EDITAL Nro.: 04/2022

Título:

A conexão entre questões socio-econômicas e a opção pela língua espanhola no ENEM: um estudo de correlações

Palavras chave

ENEM, Análise Exploratória de Dados, Regressão Logística, Engenharia de Dados

Área de Conhecimento

Educação, Banco de Dados e Estatística

Modalidade

Bolsista

São Paulo, 30 de Maio de 2022

Resumo

O Exame Nacional do Ensino Médio (ENEM) refere-se ao instrumento de diagnóstico para balizar modificações no sistema pedagógico promovidas pela Lei de Diretrizes e Bases da Educação Nacional, 9.394/1996. No entanto, ao longo dos anos, esse objetivo primeiro foi transmutado para o enfoque atual: uso do resultado do ENEM para o ingresso em instituições público-privada do ensino superior.

Anualmente, o INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira) provê resumos descritivos que retratam o perfil dos participantes e suas respectivas escolas. Porém, esses resumos não representam análises profundas capazes de suportar o objetivo original do ENEM.

A reunião de análises sobre os resultados do ENEM possibilita a identificação de padrões ou correlações que explicitam oportunidades de aprimoramento da educação básica como um todo: o sistema pedagógico, os conteúdos e as disciplinas, a condição dos seus estudantes e professores. Portanto, estudos com esse mote são estratégicos para o desenvolvimento da educação no país.

Baseados nessa perspectiva, a literatura apresenta vários estudos sobre diferentes aspectos do ENEM, incluindo o conteúdo das questões e o desempenho dos candidatos [1]. Especificamente sobre a língua espanhola, interesse do presente trabalho, os estudos enfocam as estratégias de leitura dos participantes ou os fatores que os estudantes consideram na seleção da língua estrangeira [2]. Contudo, não identificamos estudos que analisam a influência das questões sócio-econômicas na seleção da língua espanhola.

Para contribuir com a literatura acima, o objetivo do presente trabalho recai sobre a aplicação da inferência estatística para examinar se certas questões socio-econômicas (e.g., realização de curso de língua estrangeira ou cursinho) e o tipo da escola estão correlacionadas à seleção da língua espanhola nas provas do ENEM. Este trabalho adotará uma visão temporal ao analisar as variáveis acima mencionadas para os participantes que realizaram o ENEM entre os anos de 2010 e 2020 inclusive.

Conteúdo

1 Introdução

A aprovação da Lei de Diretrizes e Bases da Educação Nacional (LDB), 9.394/1996, estabeleceu o ensino médio como a etapa final da educação básica no Brasil. Desse modo, esse deveria promover a formação para a vida cidadã do educando, bem como prepará-lo para seu ingresso ao mercado de trabalho. De maneira objetiva, a reforma promovida pela LDB enfocava o ajuste no sistema pedagógico e político das escolas e a adição de um mecanismo de diagnóstico e melhoria contínua do mesmo.

Marco da educação brasileira de 1998, o Exame Nacional do Ensino Médio (doravante ENEM) é proposto como o referido mecanismo diagnóstico. Ao longo dos anos, esse exame sofreu modificações de modo a ampliar as áreas de conhecimento cobertas (expansão conteudista), incluindo as línguas estrangeiras a partir de 2010. No formato atual, os participantes do ENEM podem optar por responder questões referentes às línguas inglesa ou espanhola que são parte do currículo da educação básica vigente.

Essas modificações decorrem do deslocamento do objetivo original do ENEM: de balizador de política de educação para instrumento do vestibular nacional. O objetivo primeiro do ENEM era oferecer uma uma prova com maior foco em raciocínio cujo resultado poderia fomentar e balizar as transformações curriculares e o desenvolvimento de políticas públicas de educação básica. Contudo, atualmente, seu resultado é amplamente utilizado como critério de seleção em instituições de ensino superior público ou privada, fato que tem incrementado ano-a-ano o número de candidatos ao exame.

Anualmente, o INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira) disponibiliza resumos baseados em estatística descritiva sobre todos os dados anonimizados dos participantes do ENEM. Trata-se de um retrato do perfil dos estudantes e suas respectivas escolas. Portanto, não é o resultado de análises mais profundas (e.g., estatística inferencial) capazes de revelar padrões e correlações que suportem o objetivo original do ENEM.

Com o objetivo de alargar as análises do INEP, a literatura dispõe de variados estudos que analisam aspectos do ENEM, tais como o conteúdo de suas questões, o desempenho dos candidatos, as mudanças culturais, a desigualdade social, dentre outros temas [1]. Interesse do presente trabalho, a literatura sobre a língua espanhola no ENEM vêm examinado aspectos relacionados ao conteúdos das questões [3], as estratégias de leitura empregada pelos candidatos [4] e o efeito do formato das questões na prática dos educadores do ensino médio [5]. Contudo, poucos trabalhos têm observado os fatores da seleção da língua espanhola por parte dos candidatos.

1.1 Caracterização do Problema

O ensino da língua espanhola pode ser justificado por ser a terceira língua mais falada no mundo, bem como pela situação geográfica do Brasil que o coloca com vizinhos que utilizam o Espanhol. Apesar disso, essa língua estrangeira moderna possui uma contradição: embora sua carga horária seja muito inferior àquela destinada ao Inglês, muitos candidatos optam pelo Espanhol [2].

Diferentes variáveis podem influenciar ou caracterizar correlações que levem a seleção do Espanhol. Os trabalhos [6] e [7] descrevem estudos de campo (sobre uma escola particular) cujas análises revelaram que a proximidade entre o Português e Espanhol ou a experiência anterior com a língua constituem fatores potenciais na escolha do Espanhol como a língua avaliada. Por sua vez, um estudo [2] envolvendo seis escolas revelou que os candidatos das escolas privadas optam pelo Espanhol por gostarem da língua, enquanto aqueles provenientes das escolas públicas selecionam pela "proximidade" com a língua portuguesa. Tal fato corrobora que o ensino do Inglês é insatisfatório nas escolas públicas [2].

1.2 Objetivo

A reunião de resultados de análises sobre o ENEM permite a identificação de padrões que podem revelar oportunidades de melhoria da educação básica ou da condição dos seus candidatos e professores. Logo, deveria ser tratada como uma área estratégica para o desenvolvimento do país. Nesse sentido, o objetivo do presente trabalho pode ser descrito como:

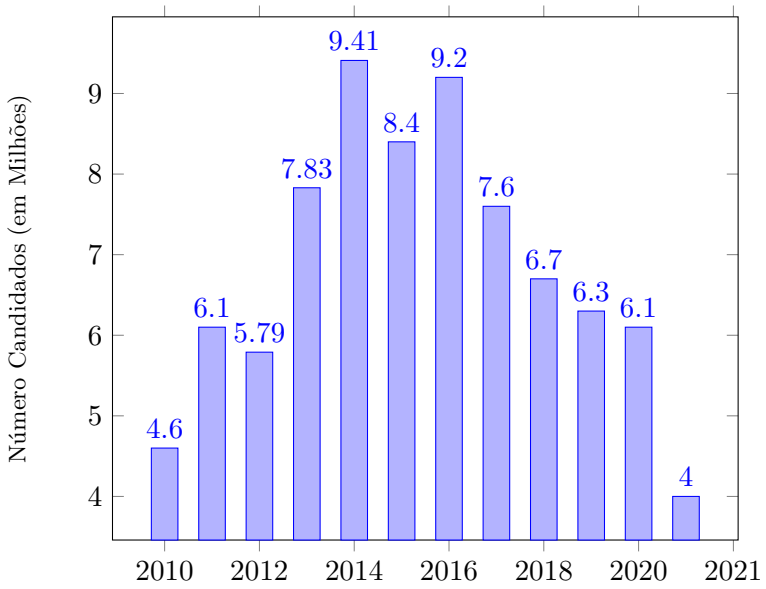
> **Aplicar a inferência estatística para revelar se o tipo da escola e certas questões socio-econômicas (participação em cursos de língua estrangeira, cursinho ou informática) dos candidatos estão correlacionadas a seleção da língua espanhola nas provas do ENEM entre 2010 e 2020.**

2 Fundamentos Teóricos

2.1 Exame Nacional do Ensino Médio

Criado em 1998, o ENEM é realizado anualmente pelo INEP e o Ministério da Educação (MEC). Seus três objetivos iniciais são: *i*) permitir ao egresso planejar para a cidadania, *ii*) servir de base para processos seletivos em organizações e *iii*) processos seletivos de instituições do ensino superior (doravente IES). Ao longo do tempo, O ENEM ampliou seus objetivos passando a atuar como base para a classificação das escolas (2008), como documento certificatório de conclusão do ensino médio (2010) ou como elemento obrigatório para a solicitação do Fundo de Financiamento Estudantil (FIES). Além disso, seu uso por IES foi alastrado por meio do SISU (2009), conforme ilustra a Figura 1.

Figura 1: *Número de inscrições no ENEM (Fonte: [8])*



Em sua história, o ENEM apresentou dois modelos. O primeiro modelo vigorou entre 1998 e 2008 e consistia de 21 habilidades articuladas em cinco competências. Logo, não havia áreas de conhecimento e conteúdos definidos. A prova correspondente apresentava 3 itens de dificuldades crescentes para cada habilidade, totalizando 63 questões. Os resultados eram analisados pelo simples números de acertos, característico da Teoria Clássica de Testes.

O modelo vigente, implantado em 2009, sofreu inúmeras modificações em relação ao anterior. A de maior impacto remete a adição de uma matriz de referência composta de 4 áreas de conhecimento, 5 eixos cognitivos e 30 competências hierarquicamente relacionadas a 120 habilidades e conteúdos (objetos e conhecimento). Cumpre também destacar que a avaliação dos resultados passou a aplicar a

Teoria de Resposta ao Item. Tais modificações ampliaram o formato (de 63 questões para 45 por área de conhecimento) e a duração da prova (de 1 dia para 2 dias).

2.1.1 Línguas Estrangeiras no ENEM

A LDB 9394/96 regulamentou o ensino de língua estrangeira de forma obrigatória para o nível de ensino fundamental (a partir do quinto ano) e médio. Para o primeiro define como obrigatório o ensino de uma língua estrangeira moderna, enquanto para o segundo abre a opção de uma segunda língua optativa.

Com a reformulação de 2010, o ENEM acrescentou cinco questões relativas às línguas estrangeiras modernas (Inglês ou Espanhol) e permitiu aos candidato optar por uma delas no momento de inscrição. Essas questões alargam a área de conhecimento de *comunicação e linguagem* com o objetivo de ampliar as possibilidades dos estudantes ao acesso de informações (e.g., cultura, tecnologia) não expressas na língua nativa.

Contudo, essas questões apresentam instruções em português e enfocam a compreensão de pequenos textos na língua estrangeira escolhida para escolha múltipla também em português. Esse formato vêm recebendo, desde a sua concepção, recorrentes críticas por reduzir o uso das línguas estrangeiras às práticas de leitura e tradução [9].

2.2 Sistema Gerenciador de Banco de Dados Relacional

Um Sistema Gerenciador de Banco de Dados (SGBD) é dito relacional quando seu modelo de dados é baseado em uma abordagem matemática na qual a estrutura dos dados e seu acesso são consistentes com a lógica de primeira ordem [10]. Tal modelo permite descrever os dados, bem como seus relacionamentos e restrições de consistência.

A estrutura de um banco de dados relacional BD consiste de um conjunto de esquemas de relações, denotado por $BD = \{R_1, R_2, R_3..., R_m\}$, $m \geq 1$. O esquema da relação (ou relação) remete a um conjunto de atributos que representam sua estrutura, denotado por $R(A) = \{a_1,..., a_k\}$, onde k é a aridade da relação. O estado da relação é conjunto de tuplas que uma relação pode apresentar em um certo instante de tempo $\mathcal{T}$, denotado por $r(R_i) = \{t_1, t_2, t_3..., t_n\}$. Por fim, uma tupla t_p, $p \in [1, n]$, é uma lista de q valores $t_p = \{v_1, v_2, ..., v_q\}$ onde cada valor v_s, $s \in [1, q]$, é um elemento do domínio de um atributo a_s, denotado por $t[a_s]$.

2.3 Projeto de Banco de Dados

O projeto de banco de dados (PDD) denota o ciclo de vida formado pelas etapas de modelagem conceitual, lógica e física. Conjuntamente, essas são responsáveis por transmutar as necessidades do universo de discurso em modelos de dados implementados em um ou mais produtos de SGBD. O conhecimento desse ciclo é amplamente reconhecido como crucial para garantir o suporte adequado desses modelos de dados às operações de uma organização, principalmente no contexto do *Big Data*.

O PDD é um trabalho intelectualmente exigente porque cada etapa requer um conjunto diferente de habilidades, conhecimento sobre o universo de discurso e técnico. A modelagem conceitual constrói abstrações de domínios de problema complexos e mal estruturado a partir do consenso entre usuários e modeladores de dados [11]. Na outra extremidade, a modelagem física do banco de dados requer um profundo conhecimento técnico e contextual (por exemplo, sistemas de informação e infra-estrutura) para decidir pelo recurso mais adequado a uma determinada necessidade de desempenho [11].

A modelagem lógica representa a transformação das abstrações do domínio do problema em uma representação que segue as estruturas e restrições do modelo de dados suportado pelo SGBD alvo

da implementação [12]. Como exemplo, abstrações de negócio são adaptados a *relações* e *conexões* (envolvendo chave primária e estrangeira) entre relações no caso da implementação em um SGBD Relacional. Vale à pena ressaltar que essa etapa de modelagem também é responsável pelo procedimento de aferição da qualidade dos modelos gerados.

2.4 Engenharia de Dados

As facilidades de geração e armazenamento de dados na Internet levaram ao aumento expressivo na quantidade de dados disponíveis para análises por parte de empresas e pessoas físicas. Contudo, esses dados (chamados de *dados brutos*) não estão pronto para consumo por não estarem organizados e avaliados (quanto a qualidade) de acordo com as necessidades dos processos analíticos.

A engenharia de dados remete a área de conhecimento responsável pelo desenvolvimento do processo de transformação dos dados brutos em dados moldados às necessidades de um processo de análise [13]. Esse processo deve ser capaz de capturar, avaliar a qualidade, integrar, ajustar e enriquecer os dados brutos.

2.5 Inferência Estatística

A inferência estatística compreende um conjunto de técnicas com o objetivo de inferir características de uma população por meio da observação de uma amostra. A inferência remete ao uso de uma forma de raciocínio denominada de indução, isto é, o raciocínio parte do específico (a amostra) para o geral (população). Contudo, esse raciocínio não é necessário quando se dispõe de toda a população.

A regressão é uma das técnicas de inferência que investiga o relacionamento entre variáveis [14]. Sua aplicação é ampla e pode ser encontra nas áreas de engenharia, química, educação, economia, dentre outras. Dentre os tipos de regressão, o presente trabalho se concentra na família das técnicas de regressão linear.

2.5.1 Regressão Linear

A regressão linear é um modelo utilizado para predizer ou avaliar se existe uma correlação linear entre variáveis. Quando o foco da análise recai sobre duas variáveis numéricas x e y, a regressão é dita simples e segue a expressão abaixo:

$$E(y|x) = \beta_0 + \beta_1 x \tag{1}$$

Essa equação define que a variável x (chamada de exploratória ou independente) é utilizada para determinar a variável y, denominada de dependente [14]. Os parâmetros β_0 e β_1 são estimados diretamente dos dados analisados. Por outro lado, quando o objeto de análise recai sobre mais do que duas variáveis exploratórias em relação a uma dependente, a regressão é dita múltipla e segue a expressão abaixo [14].

$$E(y|x) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \tag{2}$$

2.5.2 Regressão Logística

Diferentemente da linear, a variável dependente na regressão logística é dicotômica. Tal fato faz com que os princípios gerais dessa sigam àqueles da regressão linear. Porém, há diferenças na forma do modelo logístico e seus pressupostos. A expressão abaixo representa a regressão logística.

$$E(y|x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \tag{3}$$

3 Metodologia

Baseado na problemática e os objetivos discutidos respectivamentes nas Seções 1 e 1.2, a abordagem metodológica desse projeto consiste das seguintes fases:

- Consolidar os fundamentos teóricos;
- Projetar um modelo de dados temporal;
- Projetar um processo de Engenharia de Dados;
- Conduzir Análises de Correlação.

3.1 Consolidar os fundamentos teóricos

O presente projeto demanda diferentes saberes fundamentais para a consecução de seus objetivos. O primeiro deles remete compreender as características do exame e dados do ENEM e, em especial, àquelas ligadas as línguas estrangeiras avaliadas. O segundo conjunto de saberes remete a compreender o processo de análise dos dados. Aqui o discente deverá aprofundar o entendimento relativo a técnicas de preparação de dados, métodos de exploração visual dos dados e técnicas de correlação estatística.

3.2 Projetar um modelo de dados temporal

Usualmente, modelos em sistemas gerenciadores de banco de dados somente armazenam o estado corrente dos dados, isto é, seu conteúdo reflete o valor atual dos conceitos persistidos. Por exemplo, todas as instâncias de uma relação de empregados E possui o salário corrente dos empregados. Caso uma instância e da relação empregado, $e \in E$, seja atualizada, o valor do salário anterior dessa instância é perdido.

Uma vez que os dados do ENEM são sensíveis ao tempo, faz-se necessário projetar um modelo de dados capaz de acomodar todos esses dados considerando o *tempo de validade dos dados* que, no caso do ENEM, é o ano de realização do exame. Essa atividade de projeto deve considerar todas as características dos micro dados do ENEM e, então, apresentar um modelo temporal e normalizado por meio da UML.

3.3 Projetar um processo de Engenharia de Dados

Como o presente trabalho cobre um espectro temporal alargado, mudanças na formatação e codificação dos dados do ENEM são prováveis. Além disso, o volume e a variabilidade da consistência dos dados disponíveis também é outro cuidado a ser considerado. Frente a esse cenário de incertezas, o presente trabalho demanda um processo de transformação capaz de limpar e moldar os dados para a aplicação adequada das técnicas de análise (Seção 2.5).

3.4 Conduzir Análises de Correlação

A partir dos dados organizados pela etapa anterior, a etapa de análise remete a aplicação das técnicas de regressão discutidas na Seção 2.5. Para tal, faz-se necessário um processo que examine as correlações de diferentes variáveis exploratórias frente a variável dependente: “opção pela língua espanhola”. Esse processo deverá ser capaz de separar as correlações mais significativas daquelas com menor poder de correlação.

4 Cronograma Previsto

O Quadro 1 representa as principais macro etapas do presente projeto e as respectivas expectativas de conclusão. O aspecto temporal é baseado no espaço temporal *Mês de Trabalho* cuja contagem tem início a partir da data de aprovação do presente projeto.

As atividades 2 e 3 remetem ao desdobramento da etapa de "Projetar um Modelo de Dados Temporal", enquanto as atividades 4 e 5 correspondem a etapa "Projetar um processo de Engenharia de Dados". Por fim, a atividade 7 representa o esforço de produzir e revisar o relatório final requerido pelo presente edital de iniciação científica.

Atividade	Set.22	Out.22	Nov.22	Dez.22	Jan.23	Fev.23	Mar.23	Abr.23	Mai.23	Jun.23	Jul.23	Ago.23
1. Consolidar Fundamentos Teóricos	•	•	•	•	•	•						
2. Analisar os dados do ENEM		•	•									
3. Projetar um Modelos de Dados Temporal		•	•									
4. Projetar e Construir a Coleta e Análise de Qualidade			•	•								
5. Projetar e Construir a Persistência				•								
6. Conduzir as Análises de Correlação				•	•	•	•	•	•			
7. Composição do Relatório					•	•	•	•	•	•	•	•

Quadro 1: *Cronograma estimado do Projeto (Fonte: O autor)*

5 O Discente e os Recursos de Apoio

O presente projeto irá utilizar a infra-estrutura computacional regular da UFABC e a particular dos membros desse projeto. Serão utilizadas ferramentas gratuitas (ou comunitárias) fortemente estabelecidas na comunidade acadêmica ou empresarial. A priori, o presente trabalho irá utilizar as linguagens de programação Python [15], o sistema gerenciador de banco de dados PostgreSQL [16] e a ferramenta de exploração interativa de dados Tableau [17].

Por fim, cumpre destacar que o discente possui experiência anterior com iniciação científica (PDPD) na qual fez uso da estatística descritiva para descrever os resultados de suas análises.

6 Referências Bibliográficas

[1] Lima PdSN, Ambrósio APL, Ferreira DJ, Brancher JD. Análise de dados do Enade e Enem: uma revisão sistemática da literatura. Avaliação: Revista da Avaliação da Educação Superior (Campinas). 2019;24:89-107. 1

[2] Moteler LCG. As provas do ENEM e a opção de língua estrangeira. Chapecó, SC: Universidade Federal da Fronteira Sul; 2016. 1

[3] Kanashiro DSK. As linhas e as entrelinhas: um estudo das questões de língua espanhola no Enem. Universidade de São Paulo; 2012. 1

[4] da Silva Miranda M, dos Santos Rodrigues E, Preuss EO. O processo de leitura de questões de espanhol do Enem. Letrônica. 2020;13(4):e37530-0. 1

[5] Fernandes PSR. Os itens de espanhol do Enem: em busca de efeito (s) retroativo (s) na prática do professor em serviço. Universidade de Brasília; 2016. 1

[6] Mendes M, da Costa Nunes MA. INGLÊS OU ESPANHOL? QUAIS OS FATORES QUE OS ALUNOS PRIVILEGIAM NA ESCOLHA DE UMA LÍNGUA PARA O ENEM? Vivências. 2019;15(28):124-34. 1

[7] de Oliveira Santos P, Soares BEM. Crenças linguísticas que norteiam a escolha da língua espanhola para o ENEM. A Cor das Letras. 2019;20(1):245-57. 1

[8] INEP. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira;. Online; acessado 29 Fevereiro de 2022. https://www.gov.br/inep/pt-br. 2

[9] Faraco CA. Norma-padrão brasileira: desembaraçando alguns nós. Linguística da norma São Paulo: Loyola. 2002:37-61. 3

[10] Abiteboul S, Hull R, Vianu V. A Larger Perspective. In: Foundations of databases. vol. 8. Addison-Wesley Reading; 1995. p. 216-35. 3

[11] Borovina Josko JM. A Problem-based approach to teach physical database design: an experience report. In: Anais do XXVIII Workshop sobre Educação em Computação. SBC; 2020. p. 11-5. 3

[12] Heuser CA. Projeto de banco de dados: Volume 4 da Série Livros didáticos informática UFRGS. Bookman Editora; 2009. 4

[13] Widanage C, Perera N, Abeykoon V, Kamburugamuve S, Kanewala TA, Maithree H, et al. High performance data engineering everywhere. In: 2020 IEEE International Conference on Smart Data Services (SMDS). IEEE; 2020. p. 122-32. 4

[14] Montgomery DC, Peck EA, Vining GG. Introduction to linear regression analysis. John Wiley & Sons; 2021. 4

[15] Matthes E. Python crash course: a hands-on, project-based introduction to programming. No Starch Press; 2015. 6

[16] Schönig HJ. PostgreSQL Administration Essentials. Packt Publishing Ltd; 2014. 6

[17] Nandeshwar A. Tableau data visualization cookbook. Packt publishing; 2013. 6