Desafio de Ciência de Dados com Microdados do ENEM 2023

Professora Carolina Ribeiro Xavier

Introdução

O Exame Nacional do Ensino Médio (ENEM) é uma das maiores avaliações do Brasil, abrangendo milhões de candidatos em todo o país. Os microdados do ENEM 2023 oferecem um vasto conjunto de informações que permitem uma análise aprofundada dos resultados e das características dos participantes.

Neste desafio de análise de dados, vocês utilizarão esses microdados para desenvolver uma análise completa, desde o tratamento inicial da base para descrição básica da base até a obtenção de insights estatísticos e visuais, incluindo mapas.

Objetivos do Desafio

Os objetivos principais deste desafio incluem:

- Realizar o tratamento e a organização dos microdados do ENEM 2023.
- Desenvolver análises descritivas e exploratórias com foco em variáveis como região, gênero, tipo de escola e notas por disciplina.
- Utilizar visualizações para facilitar a interpretação dos dados e identificar padrões relevantes.
- Explorar as correlações entre as notas nas diferentes disciplinas, visando descobrir insights que relacionem o desempenho dos alunos.
- Utilizar uma base adicional para correlacionar com as informação obtidas por essa base, como por exemplo, o IDH da região ou proporção de pessoas inscritas no CadÚnico na região, ou número de pessoas matriculadas nas escolas por gênero e etc.

Tarefas

O desafio está dividido em cinco etapas principais:

1. Tratamento e Organização dos Dados

- Carregar e inspecionar o conjunto de dados para identificar colunas relevantes (como região, gênero, tipo de escola e notas).
- Realizar tratamento de valores ausentes e dados inconsistentes.
- Criar novas colunas categóricas que ajudem a agrupar os dados por região, gênero e tipo de escola.

2. Análise Estatística Descritiva

- Calcular medidas de tendência central e dispersão, como média, mediana, desvio padrão e quartis, das notas para cada disciplina.
- Analisar as estatísticas por gênero, região e tipo de escola.

3. Visualização dos Dados

- Criar gráficos de distribuição das notas segmentados por região, gênero e tipo de escola.
- Comparar as médias das notas por região e tipo de escola, destacando as diferenças entre gêneros.

4. Análise de Correlação entre Notas

- Calcular a correlação entre as notas de cada disciplina.
- Visualizar a matriz de correlação para identificar possíveis relações entre as disciplinas.

5. Descoberta de Informações e Insights

- Analisar padrões e tendências, como disparidades de desempenho entre regiões e tipos de escola.
- Levantar hipóteses sobre o desempenho por disciplina e o impacto de diferentes fatores no desempenho geral.

Entregáveis

- **Código Python:** Código organizado e comentado, preferencialmente em formato de Jupyter Notebook.
- **Visualizações:** Gráficos que demonstrem as distribuições de notas e as correlações entre disciplinas.
- **Relatório:** Resumo dos principais insights obtidos, destacando as descobertas mais relevantes.

Recursos

Os dados para o desafio estão disponíveis em: `https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem`. Recomenda-se utilizar bibliotecas como Pandas, Streamlit, Matplotlib e Seaborn para a análise e visualização dos dados.

Prazo e Avaliação

O prazo para a entrega do desafio é de 02/12 (nos dias 25 e 27 estarei a disposição na sala para tirar dúvidas, mas não terá chamada, se organizem para não perder o prazo.). A avaliação será baseada na qualidade do tratamento dos dados, nas visualizações produzidas e na profundidade dos insights obtidos.

Grupo

O trabalho deve ser feito em grupos de até três pessoas e será discutido em sala no dia 02, quando todos deverão compartilhar os seus achados com a turma (apresentação breve de cada grupo dos achados).