**GABRIEL MASCARENHAS DE SOUZA**

**JOSÉ AUGUSTO**

**PEDRO ZERBINI**

**Análise e Extração de dados de participantes do ENEM no ano de 2022**

Sorocaba-SP

2024

**GABRIEL MASCARENHAS DE SOUZA**

**JOSÉ AUGUSTO**

**PEDRO ZERBINI**

**Análise e Extração de dados de participantes do ENEM no ano de 2022**

Trabalho de Engenharia de Dados e Análise

exploratória de Dados que será apresentado

um algoritmo para extração e analise de

dados dos participantes do ENEM no ano de

2022.

Orientador: Prof. César Candido Xavier

Sorocaba-SP

2024

**RESUMO**

O trabalho focou na extração e análise de dados de uma tabela contendo informações dos participantes do Exame Nacional do Ensino Médio (ENEM) de 2022. Utilizando técnicas de mineração de dados, os pesquisadores coletaram informações demográficas, como idade, gênero, raça/etnia, além de dados educacionais, como escolaridade dos pais, tipo de escola e localização geográfica dos participantes.

A análise desses dados permitiu identificar padrões e tendências entre os participantes do ENEM, incluindo disparidades regionais no desempenho dos alunos, diferenças socioeconômicas no acesso à educação e desafios enfrentados por grupos minoritários. Além disso, o estudo explorou a relação entre o desempenho dos alunos no ENEM e fatores como o tipo de escola frequentada e o nível de escolaridade dos pais.

Os resultados deste trabalho forneceram insights valiosos para formuladores de políticas educacionais, instituições de ensino e pesquisadores interessados em entender melhor as dinâmicas sociais e educacionais dos participantes do ENEM de 2022, contribuindo assim para o desenvolvimento de estratégias mais eficazes de inclusão e melhoria da qualidade da educação no Brasil.

**SUMÁRIO**

Página 3- Resumo.............................................................................................01

Página 5- Extração de Dados............................................................................01

Página 6- Análise de Dados..............................................................................01

Página 7- Aplicação e Exemplos ......................................................................01

Página 8- Conclusão..........................................................................................01

Página 9- Perguntas..........................................................................................01

Página 10- Código.............................................................................................01

Página 11- Biblíografia.......................................................................................01

**EXTRAÇÃO DE DADOS**

Extração de dados é o processo de coleta de informações relevantes de diferentes fontes de dados, sejam elas bancos de dados, planilhas, documentos de texto, páginas da web ou qualquer outro tipo de arquivo digital. Esse procedimento é fundamental para análise, interpretação e utilização desses dados em diversos contextos, como pesquisa acadêmica, tomada de decisões empresariais e desenvolvimento de inteligência artificial.

Para realizar a extração de dados, é necessário primeiro identificar as fontes de onde os dados serão coletados e definir os objetivos da análise. Em seguida, são utilizadas ferramentas e técnicas específicas para acessar e extrair os dados de forma automatizada ou manual, dependendo da complexidade e da estrutura dos dados.

Entre as técnicas mais comuns de extração de dados estão o web scraping, que consiste na extração de informações de páginas da web por meio de programas ou scripts, e o uso de consultas SQL para extrair dados de bancos de dados relacionais.

Além disso, é importante considerar questões de segurança e privacidade durante o processo de extração de dados, garantindo que as informações coletadas estejam em conformidade com as regulamentações e políticas de proteção de dados vigentes.

Uma vez extraídos os dados, é comum realizar etapas adicionais de limpeza, transformação e análise para preparar os dados para uso posterior. Isso inclui remover dados duplicados ou inconsistentes, padronizar formatos e realizar cálculos ou agregações, conforme necessário.

Em resumo, a extração de dados é um passo fundamental para aproveitar o potencial das informações disponíveis, permitindo insights valiosos e embasando decisões fundamentadas em dados. Através de técnicas adequadas e boas práticas, é possível extrair valor dos dados e impulsionar o progresso em diversos campos.

**ANÁLISE DE DADOS**

A análise de dados é um processo crucial para extrair insights significativos e tomar decisões informadas em uma variedade de campos, desde a pesquisa científica até o planejamento estratégico empresarial. Consiste em examinar, interpretar e visualizar conjuntos de dados para identificar padrões, tendências e relações que podem estar subjacentes aos dados brutos.

Para realizar uma análise de dados eficaz, é fundamental seguir uma metodologia sólida. Primeiramente, é necessário definir claramente os objetivos da análise e as perguntas que se deseja responder. Em seguida, os dados são preparados por meio de limpeza, transformação e integração para garantir sua qualidade e consistência. Existem várias abordagens e métodos para a análise de dados, cada um adequado para diferentes tipos de dados e questões de pesquisa. Alguns dos métodos mais comuns incluem:

* Análise estatística descritiva: Este método envolve a descrição e sumarização dos dados por meio de medidas como média, mediana, desvio padrão, e gráficos como histogramas e box plots.
* Análise exploratória de dados (AED): A AED é uma abordagem mais ampla que envolve a exploração visual e estatística dos dados para identificar padrões, anomalias e relações inesperadas.
* Análise inferencial: Este método visa fazer inferências sobre uma população com base em uma amostra de dados, utilizando técnicas como testes de hipóteses e intervalos de confiança.
* Análise de regressão: Utilizada para entender a relação entre uma variável dependente e uma ou mais variáveis independentes, a análise de regressão é útil para prever valores futuros e identificar influências significativas.
* Análise de séries temporais: Este método é utilizado quando os dados estão organizados em sequências temporais e visa identificar padrões sazonais, tendências e ciclos ao longo do tempo.

Além desses métodos, técnicas avançadas como aprendizado de máquina e mineração de dados podem ser aplicadas para análises mais complexas e preditivas.

**APLICAÇÃO E EXEMPLOS**

A análise e extração de dados das tabelas de participantes do ENEM de 2022 têm uma série de aplicações cruciais na compreensão do panorama educacional brasileiro e na formulação de políticas educacionais mais eficazes. Por exemplo, ao extrair e analisar esses dados, os formuladores de políticas podem identificar disparidades regionais no desempenho dos alunos, destacando áreas que requerem mais investimentos em infraestrutura educacional.

Além disso, a análise desses dados pode revelar tendências demográficas entre os participantes do ENEM, como diferenças no desempenho entre gêneros ou grupos étnicos, fornecendo insights para a criação de programas de apoio específicos.

Empresas e instituições de ensino também podem se beneficiar da análise desses dados para entender melhor o perfil dos estudantes que ingressam no ensino superior, orientando a elaboração de estratégias de recrutamento e retenção. Um exemplo prático seria:

* **Análise descritiva:** Inclui a utilização de estatísticas básicas, como média, mediana, moda e desvio padrão, para descrever características demográficas dos participantes, como idade, gênero, raça/etnia e localização geográfica.
* **Análise comparativa:** Permite comparar diferentes grupos de participantes do ENEM com base em variáveis como tipo de escola (pública ou privada), região geográfica, e nível socioeconômico, identificando disparidades e desigualdades no desempenho e acesso à educação.
* **Análise de correlação:** Busca identificar relações entre variáveis, como a relação entre o desempenho no ENEM e o nível de escolaridade dos pais, o tipo de escola frequentada ou a disponibilidade de recursos educacionais na região.
* **Análise de tendências ao longo do tempo:** Permite identificar mudanças e padrões no perfil dos participantes do ENEM ao longo dos anos, destacando áreas de melhoria ou sucesso nas políticas educacionais.
* **Análise geoespacial:** Utiliza técnicas de visualização de dados para mapear a distribuição geográfica dos participantes do ENEM e identificar padrões regionais de desempenho e acesso à educação.

Em resumo, a análise e extração de dados das tabelas de participantes do ENEM de 2022 têm o potencial de fornecer insights valiosos para melhorar a educação no Brasil, desde a formulação de políticas públicas até a implementação de iniciativas práticas em instituições de ensino e empresas do setor educacional.

**CONCLUSÃO**

A análise e extração de dados das tabelas de participantes do ENEM de 2022 representam uma ferramenta poderosa para compreender as complexidades do sistema educacional brasileiro. Ao examinar os dados demográficos, socioeconômicos e educacionais dos candidatos, podemos identificar padrões, tendências e desigualdades que influenciam o acesso à educação e o desempenho dos alunos.

Essa análise oferece insights valiosos para formuladores de políticas, educadores, instituições de ensino e empresas, permitindo a criação de estratégias mais eficazes para promover a equidade e a qualidade da educação em todo o país. Desde a identificação de áreas carentes de investimento até o desenvolvimento de programas de suporte para grupos marginalizados, os dados extraídos das tabelas do ENEM fornecem uma base sólida para a tomada de decisões informadas.

No entanto, é importante reconhecer que a análise de dados por si só não é suficiente para resolver todos os desafios educacionais. É essencial complementar os insights obtidos com ações concretas e políticas públicas bem planejadas, visando promover uma educação inclusiva e de qualidade para todos os brasileiros.

Portanto, ao utilizar a análise e extração de dados das tabelas do ENEM de 2022, devemos buscar não apenas compreender o cenário atual da educação no Brasil, mas também agir de forma proativa para construir um futuro mais justo e promissor para as próximas gerações.

**BIBLÍOGRAFIA**

Marwell, M. P., & Santos, V. C. B. Extração de dados em artigos científicos utilizando web scraping na base de dados do sciencedirect.

Olivieri, A. G., & Castro, G. J. (2021). A sociedade digital de extração de dados e os desafios para a democracia. *Revista Processus de Políticas Públicas e Desenvolvimento Social*, *3*(6), 19-40.

Teixeira, M. A., Nitschke, R. G., & Paiva, M. S. (2008). Análise dos dados em pesquisa qualitativa: um olhar para a proposta de Morse e Field. *Revista da Rede de Enfermagem do Nordeste*, *9*(3), 135-142.

de Sousa, J. R., & dos Santos, S. C. M. (2020). Análise de conteúdo em pesquisa qualitativa: modo de pensar e de fazer. *Pesquisa e debate em Educação*, *10*(2), 1396-1416.

Pereira, M. G., & Galvão, T. F. (2014). Extração, avaliação da qualidade e síntese dos dados para revisão sistemática. *Epidemiologia e Serviços de Saúde*, *23*, 577-578.

Spalenza, M. A., Nogueira, M. A., de Andrade, L. B., & de Oliveira, E. (2018). Uma Ferramenta para Mineração de Dados Educacionais: Extração de Informação em Ambientes Virtuais de Aprendizagem. *Anais do Computer on the Beach*, 741-750.

**PERGUNTAS**

1 – Relação das notas de ciências da natureza com ciências humanas.

2 - Relação das notas de ciências da natureza com linguagens e códigos.

3 - Relação das notas de ciências da natureza com matemática.

4 - Relação das notas de ciências da natureza com redação.

5 - Relação das notas de ciências humanas com linguagens e códigos.

6 - Relação das notas de ciências humanas com matemática.

7 - Relação das notas de ciências humanas com redação.

8 - Relação das notas de linguagens e códigos com matemática.

9 - Relação das notas de linguagens e códigos com redação.

10 - Relação das notas de matemática com redação.

11 – Relação de faixa etária com as notas de ciências humanas.

12 - Relação de faixa etária com as notas de ciências da natureza.

13 - Relação de faixa etária com as notas de linguagens e códigos.

14 - Relação de faixa etária com as notas de matemática.

15 – Relação de faixa etária com as notas de redação.

16 – Relação de sexo com as notas de ciências humanas.

17 - Relação de sexo etária com as notas de ciências da natureza.

18 - Relação de sexo com as notas de linguagens e códigos.

19 - Relação de sexo com as notas de matemática.

20 – Relação de sexo com as notas de redação.

21 – Relação de cor ou raça com as notas de ciências humanas.

22 - Relação de cor ou raça com as notas de ciências da natureza.

23 - Relação de cor ou raça com as notas de linguagens e códigos.

24 - Relação de cor ou raça com as notas de matemática.

25 – Relação de cor ou raça com as notas de redação.

26 – Relação de situação de conclusão com as notas de ciências humanas.

27 - Relação de situação de conclusão com as notas de ciências da natureza.

28 - Relação de situação de conclusão com as notas de linguagens e códigos.

29 - Relação de situação de conclusão com as notas de matemática.

30 – Relação situação de conclusão com as notas de redação.

**CÓDIGO DESENVOLVIDO E COMENTANDO**

# Importando bibliotecas

import streamlit as st

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

# Título

st.title('Análise de dados do ENEM de 2022')

# Subtítulo

st.markdown("Utilize este App e construa suas próprias análises de dados!")

st.subheader('Gráfico de dispersão')

selecao\_opcao = st.sidebar.multiselect('Selecione os estados', ['AC','CE','DF','ES','GO','MA','MG','MS','MT','PA','PB','PE','PI','PR','RJ','RN','RO','RR','RS','SC','SE','SP'], default='AC')

print(type(selecao\_opcao), selecao\_opcao)

selecao\_var\_x = st.selectbox('Selecione a variável para o eixo X', ['NOTA CIÊNCIAS DA NATUREZA','NOTA CIÊNCIAS HUMANAS','NOTA LINGUAGENS E CÓDIGOS','NOTA MATEMÁTICA', 'NOTA REDACÃO', 'NOTA TOTAL'])

selecao\_var\_y = st.selectbox('Selecione a variável para o eixo Y', ['NOTA CIÊNCIAS DA NATUREZA','NOTA CIÊNCIAS HUMANAS','NOTA LINGUAGENS E CÓDIGOS','NOTA MATEMÁTICA', 'NOTA REDACÃO', 'NOTA TOTAL'])

visualizar\_toda\_escala = st.checkbox('Visualizar toda a escala do gráfico')

# lendo o arquivo de dados penguins.csv

df = pd.read\_csv('PlanilhaEnemCorreta.csv')

# A função plt.subplots () sem parâmetros retorna uma figura e um eixo. A fig e o ax retornados são atribuídos a variáveis fig e ax, respectivamente.

# A variável fig é um objeto de figura. A figura é a imagem final que pode conter um ou mais eixos. Você pode pensar na figura como um contêiner que contém esses eixos.

# A variável ax é um objeto de eixos. Um objeto de eixos é um gráfico individual com um eixo x, um eixo y e o espaço onde você pode desenhar linhas ou pontos.

# Após essa linha, você pode usar o objeto ax para desenhar gráficos, definir rótulos, etc. e usar o objeto fig para salvar a figura em um arquivo, alterar o tamanho da figura e assim por diante.

# criando um gráfico de dispersão

fig, ax = plt.subplots()

for sel\_epc in selecao\_opcao:

    df1 = df[df.SG\_UF\_PROVA == sel\_epc]

    ax.scatter(df1[selecao\_var\_x], df1[selecao\_var\_y], alpha=0.8)

if visualizar\_toda\_escala:

    ax.set\_xlim(0, 1.1\*df[selecao\_var\_x].max())

    ax.set\_ylim(0, 1.1\*df[selecao\_var\_y].max())

ax.set\_xlabel(selecao\_var\_x)

ax.set\_ylabel(selecao\_var\_y)

ax.legend(selecao\_opcao)

ax.grid(True)

ax.set\_title(f'Dispersão {selecao\_var\_x} x {selecao\_var\_y} para os estados:'+' '.join(selecao\_opcao))

st.pyplot(fig)

st.header('Gráfico de barras')

selecao\_var\_barra\_x = st.selectbox('Selecione a variável para o eixo X', ['FAIXA ETÁRIA','SEXO','COR OU RACA','SITUACÃO DE CONCLUSÃO','ANO QUE CONCLUIU','ESCOLA,ENSINO','ESTADO DA ESCOLA','DEPENDÊNCIA ADM DA ESCOLA',

'LOCALIZACAO DA ESCOLA', 'PRESENCA CIÊNCIAS DA NATUREZA','PRESENCA CIÊNCIAS HUMANAS','PRESENCA LINGUAGENS E CÓDIGOS','PRESENCA MATEMÁTICA','CÓDIGO DA PROVA CIÊNCIAS DA NATUREZA',

'CÓDIGO DA PROVA CIÊNCIAS HUMANAS','CÓDIGO DA PROVA LINGUAGENS E CÓDIGOS','CÓDIGO DA PROVA MATEMÁTICA','LÍNGUA ESTRANGEIRA','STATUS DA REDACÃO', 'FAIXA SALARIAL', 'POSSUI INTERNET EM CASA'])

selecao\_var\_barra\_y = st.selectbox('Selecione a variável para o eixo Y', ['NOTA CIÊNCIAS DA NATUREZA','NOTA CIÊNCIAS HUMANAS','NOTA LINGUAGENS E CÓDIGOS','NOTA MATEMÁTICA','NOTA TOTAL', 'NOTA REDACÃO'])

# Filtrando o DataFrame com base na seleção em selecao\_opcao

df\_filtrado = df[df['SG\_UF\_PROVA'].isin(selecao\_opcao)]

# Criando um gráfico de barras

fig, ax = plt.subplots()

# Agrupando os dados filtrados pelo selecao\_var\_barra\_x e calculando a média do selecao\_var\_barra\_y

grouped\_data = df\_filtrado.groupby(selecao\_var\_barra\_x)[selecao\_var\_barra\_y].mean()

# Plotando o gráfico de barras

grouped\_data.plot(kind='bar', ax=ax)

# Adicionando rótulos e título

ax.set\_xlabel(selecao\_var\_barra\_x)

ax.set\_ylabel(selecao\_var\_barra\_y)

ax.set\_title(f'Média de {selecao\_var\_barra\_y} por {selecao\_var\_barra\_x} pelo(s) estado(s) '+' '.join(selecao\_opcao))

# Mostrando o gráfico

st.pyplot(fig)