Análise do ENEM 2022

Este notebook realiza uma análise exploratória dos dados do ENEM 2022, utilizando técnicas de processamento de dados, visualização e agrupamento. O ENEM (Exame Nacional do Ensino Médio) é uma avaliação realizada anualmente pelo governo brasileiro para medir o desempenho dos estudantes e servir como critério de acesso ao ensino superior.

Objetivo

Explorar e identificar padrões nos dados do ENEM 2022, com foco em variáveis demográficas e de desempenho dos estudantes. Isso inclui o tratamento de dados, análise descritiva e técnicas de agrupamento.

1. Introdução

Utilizamos os microdados do ENEM disponibilizados pelo INEP, que podem ser acessados aqui: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem.

2. Processamento dos Dados

As etapas de processamento envolvem limpeza de dados, engenharia de atributos e modelagem, conforme necessário ao longo da análise.

Pacotes Utilizados:

```
import pandas as pd
import numpy as np
```

```
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans
import warnings
warnings.filterwarnings("ignore")
```

Carregamento dos Dados:

Os dados são carregados diretamente do arquivo CSV fornecido pelo INEP, e configuramos o Pandas para exibir todas as colunas do dataset.

```
Enem = pd.read_csv("MICRODADOS_ENEM_2022.csv", sep=';', encoding='ISO-8859-1')
pd.set_option('display.max_columns', None)
Enem.head()
```

Limpeza e Ajustes no Dataset:

Note que a codificação original do arquivo fornecido pelo MEC (ISO-8859-1) não segue o padrão usual de UTF-8, o que exige a especificação da codificação adequada durante o carregamento dos dados.

3. Análise Descritiva

Nesta seção, realizamos uma análise descritiva das principais variáveis do ENEM, como faixa etária, sexo, raça e desempenho nas provas.

Distribuição de Faixa Etária:

```
sns.countplot(x='TP_FAIXA_ETARIA', data=Enem)
```

```
plt.title('Distribuição por Faixa Etária')
plt.show()
```

Desempenho nas Provas:

Aqui, investigamos a distribuição de notas nas diversas áreas do conhecimento (Ciências Humanas, Ciências da Natureza, Linguagens, Matemática, e Redação).

```
sns.histplot(Enem['NU_NOTA_MT'], kde=True)
plt.title('Distribuição das Notas de Matemática')
plt.show()
```

4. Agrupamento de Dados

Nesta etapa, aplicamos o algoritmo K-Means para identificar padrões de agrupamento entre os participantes.

Definição e Treinamento do Modelo:

```
kmeans = KMeans(n_clusters=5)
clusters = kmeans.fit_predict(Enem[['NU_NOTA_MT', 'NU_NOTA_CN', 'NU_NOTA_CH', 'NU_NOTA_LC']])
Enem['Cluster'] = clusters
```

Visualizamos os grupos gerados com base nas notas das diferentes áreas.

5. Conclusão

Com base na análise exploratória e nos agrupamentos, pudemos identificar alguns padrões importantes entre os participantes do ENEM 2022. Essas informações podem auxiliar em políticas educacionais e na melhora contínua do processo de avaliação.

Referências:

Microdados do ENEM: Link para download: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem

Documentação do Scikit-Learn: KMeans: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html