RELATÓRIO DE COMPREENSÃO E PREPARAÇÃO DE DADOS DO ENADE 2021

Nome dos integrantes: Rennan Wictor Carpes, Filipe Roman, Lorenzo Eulalio de Souza, Alexya Ungaratti

ANÁLISE DE MICRODADOS DE CURSOS E INSTITUIÇÕES DE ENSINO SUPERIOR

Seção 1: Introdução

Este relatório tem como objetivo principal apresentar a **compreensão e o processo de preparação de um conjunto de dados complexo**, focado nos microdados do Exame Nacional de Desempenho dos Estudantes (ENADE) de 2021. Além dos dados brutos do ENADE, a análise integra informações complementares de outras bases de dados do Ministério da Educação (MEC) referentes a instituições de ensino superior e seus respectivos cursos.

Seção 2: Compreensão dos Dados e Fontes

2.1 Fontes de Dados:

O conjunto de dados para esta análise foi composto por múltiplas fontes, visando uma visão abrangente do cenário avaliado. As fontes utilizadas são:

- **Microdados do ENADE 2021**: Estes consistem em 43 arquivos de texto (microdados2021_arq1.txt a microdados2021_arq43.txt). Cada arquivo contém um subconjunto de variáveis relacionadas à avaliação do ENADE para o ano de 2021.
- **Microdados de Cadastro de Cursos**: O arquivo MICRODADOS_CADASTRO_CURSOS_2023.xlsx contém informações detalhadas sobre os cursos de Ensino Superior, como seus nomes, códigos, modalidades e outras características administrativas.
- **Microdados de Instituições de Ensino Superior**: O arquivo MICRODADOS_ED_SUP_IES_2023.xlsx oferece dados sobre as Instituições de Ensino Superior (IES) brasileiras, incluindo seus nomes, códigos identificadores, categorias administrativas e organização acadêmica.

2.2: Descrição do Dataset Unificado

O processo de construção do dataset para análise envolveu a integração de três fontes de dados distintas, visando criar uma base consolidada e enriquecida para a exploração. As etapas de unificação foram:

1. **Consolidação dos Microdados do ENADE:** Inicialmente, os 43 arquivos de texto (.txt) referentes aos microdados do ENADE 2021 foram processados. Cada arquivo foi lido, e as informações pertinentes foram extraídas utilizando as colunas NU_ANO e CO_CURSO como chaves de identificação. Através de merges sucessivos, esses arquivos foram combinados em um único DataFrame, df_enade_completo, contendo **7.997 observações** (representando cursos/participações no ENADE) e **76 variáveis** originais do exame.
2. **Integração com Dados de Cadastro de Cursos:** O DataFrame do ENADE foi, então, enriquecido com informações do arquivo MICRODADOS_CADASTRO_CURSOS_2023.CSV. Este arquivo, contendo 671.610 registros e 202 variáveis sobre os cursos de educação superior, foi unido ao df_enade_completo através de um merge do tipo 'left'. As chaves para esta unificação foram CO_CURSO e, quando presente em ambos os datasets, NU_ANO. Notavelmente, esta etapa resultou em uma expansão do número de linhas para **112.061 observações** e um total de **277 variáveis**, indicando que alguns cursos presentes nos dados do ENADE possuíam múltiplas correspondências no cadastro de cursos (possivelmente devido a diferentes entradas ou atualizações ao longo do tempo no arquivo de cadastro).
3. **Integração com Dados de Instituições de Ensino Superior (IES):** Finalmente, o dataset resultante da etapa anterior foi combinado com o arquivo MICRODADOS_ED_SUP_IES_2023.CSV, que continha 2.580 registros e 84 variáveis sobre as IES. Este merge também foi do tipo 'left', utilizando a coluna CO_IES_enade (proveniente do ENADE e mantida após o primeiro merge) como chave do lado esquerdo e a coluna CO_IES do arquivo de IES como chave do lado direito

O dataset final consolidado, df_final_completo, após todas as etapas de integração, apresenta as dimensões de **112.061 observações (linhas)** e **361**

variáveis (colunas). Este é o conjunto de dados que servirá de base para as subsequentes etapas de análise exploratória, avaliação detalhada da qualidade dos dados e o processo de pré-processamento descrito neste relatório.

2.3: Caracterização Inicial do Dataset Unificado

O dataset final unificado, df_final_completo, abrange uma vasta gama de informações, incluindo dados de desempenho e perfil dos participantes do ENADE, características detalhadas dos cursos ofertados e atributos específicos das instituições de ensino.
Uma análise inicial dos tipos de dados presentes no DataFrame, obtida através do método .info(), revela a seguinte distribuição:

- **object**: 85 colunas, geralmente representando dados textuais ou categóricos (como nomes, códigos descritivos, ou respostas a questões qualitativas).
- **float64**: 262 colunas, predominantemente representando variáveis numéricas contínuas ou quantitativas (como notas, indicadores e diversas quantidades dos censos de cursos e IES).
- **float32**: 10 colunas (originárias principalmente das notas do ENADE, otimizadas para menor uso de memória).
- **int32**: 3 colunas (códigos numéricos como CO_CURSO, CO_IES_enade, CO_MUNIC_CURSO).
- **int16**: 1 coluna (NU_ANO).

O uso de memória estimado para este DataFrame em sua forma carregada é de aproximadamente **302.4+ MB**.
Esta caracterização inicial sublinha a complexidade e a riqueza do dataset compilado, que servirá como base para as análises subsequentes. A diversidade de tipos de dados também indica a necessidade de etapas de pré-processamento específicas, como o mapeamento de códigos e o tratamento de valores ausentes, que serão detalhadas nas próximas seções."

2.4: Qualidade dos Dados e justificativa para Pré-Processamento:

Apesar da riqueza do dataset unificado, uma avaliação inicial da qualidade dos dados brutos (antes das etapas de limpeza e transformação mais

profundas) revelou desafios significativos. A análise preliminar de valores ausentes indicou que 351 das 361 colunas continham algum nível de dados faltantes, algumas com uma porcentagem considerável de ausência.
Adicionalmente, observou-se a necessidade de padronizar representações de dados categóricos (que estavam em formato de código numérico ou textual) para facilitar a interpretação. Também foram identificados outliers pontuais, como em idades de participantes, e colunas específicas de questionários aplicados apenas a um subconjunto de cursos (licenciaturas) que resultavam em alta incidência de valores não preenchidos para os demais. Essas constatações justificaram a aplicação de um conjunto de etapas de pré-processamento, detalhadas a seguir, com o objetivo de preparar um dataset mais robusto e confiável para a análise exploratória.

2.5: Etapas de pré-processamento realizadas

Com base na avaliação da qualidade dos dados, um conjunto de etapas de pré-processamento foi aplicado ao dataset unificado para aprimorar sua consistência e adequação para análise. As principais ações realizadas incluem:

1. **Mapeamento de Códigos Categóricos:** Variáveis categóricas relevantes, originalmente representadas por códigos numéricos ou alfabéticos (como CO_CATEGAD - Categoria Administrativa da IES, QE_I08 - Renda Familiar, e questões selecionadas do questionário do estudante como QE_I89 a QE_I92), foram mapeadas para seus respectivos valores descritivos textuais, facilitando a interpretação. Novas colunas com o sufixo _DESC foram criadas para armazenar essas descrições.
2. **Tratamento de Outliers:** Valores discrepantes foram tratados. Especificamente, na coluna NU_IDADE, idades inferiores a 15 anos foram consideradas implausíveis e convertidas para valores ausentes, sendo posteriormente imputadas.
3. **Remoção de Colunas Irrelevantes ou Problemáticas:** Um conjunto de 13 colunas do questionário ENADE (QE_I69 a QE_I81), específicas para cursos de licenciatura e com alta incidência de valores nulos para os demais cursos, foi removido para evitar distorções e reduzir a dimensionalidade desnecessária. Após esta etapa, o dataset passou de 361 para 354 colunas (considerando a adição prévia das colunas _DESC).

4. **Imputação de Dados Faltantes:** Foi adotada uma estratégia multifacetada para o tratamento de valores ausentes restantes:
    - Colunas numéricas específicas (majoritariamente notas e dados de perfil do ENADE) tiveram seus valores faltantes preenchidos pela **mediana** da respectiva coluna.
    - Colunas categóricas específicas (como TP_SEXO) foram imputadas utilizando a **moda**.
    - Todas as demais colunas do tipo 'object' (texto/categóricas) que ainda continham valores ausentes foram preenchidas com a string **"Não Informado"**.
    - Finalmente, todas as colunas numéricas restantes (muitas provenientes dos datasets de Cursos e IES) que ainda apresentavam valores ausentes foram preenchidas com **zero**.

Ao final dessas etapas, o dataset foi considerado limpo, sem valores ausentes, e pronto para as análises exploratórias.

2.6: Conclusão da Preparação dos Dados e Próximos Passos

Ao término das etapas de coleta, integração, limpeza e pré-processamento, obteve-se um dataset final consolidado contendo **112.061 observações e 354 colunas**. Este conjunto de dados, agora livre de valores ausentes e com suas variáveis categóricas relevantes devidamente mapeadas, encontra-se pronto para subsidiar a fase subsequente de análise estatística e exploratória dos microdados do ENADE e informações associadas de cursos e IES, conforme proposto no escopo do projeto.