

PROJETO DE LABORATÓRIO DE DESENVOLVIMENTO DE BANCO DE DADOS V

ETAPA 2 – Coleta, Limpeza e Preparação dos Dados (ETL)

Data de Entrega: 14/10/2025

Membros do Grupo: Carlos Henrique Hideki Koti da Silva

Gabriel da Silva Pereira

Lucas Soares Fabricio

1. Coleta dos Dados

A aquisição dos dados foi o passo inicial do projeto, executado através do download direto do portal do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), a fonte oficial para dados educacionais no Brasil. Foi selecionada a base de "Microdados do Censo Escolar", que abrange informações detalhadas em nível de escola sobre a educação básica em todo o território nacional. Além do arquivo de dados principal, foram baixados os respectivos "Dicionários de Dados" e "Layout", documentação essencial que descreve o significado e a estrutura de cada variável presente na base.

2. Análise Inicial

Após a coleta, foi realizada uma análise exploratória da base de dados bruta para identificar desafios e planejar o processo de tratamento. As principais características identificadas foram:

- **Formatos Inconsistentes:** O arquivo original utilizava a codificação de caracteres latin-1, um padrão que gera erros na renderização de acentos e caracteres especiais em sistemas modernos que operam com UTF-8.
- **Dados Faltantes:** Constatou-se a presença de valores nulos (ausentes) em diversas colunas, especialmente em campos numéricos que representam quantidades, como número de equipamentos ou salas.
- **Dados Codificados:** Verificou-se que a maioria das variáveis categóricas, como dependência administrativa (TP_DEPENDENCIA), era representada por códigos numéricos (ex: 1 para "Federal", 2 para "Estadual"), tornando a interpretação direta dos dados complexa e dependente da consulta ao dicionário de dados.
- **Alta Dimensionalidade:** A base de dados completa possuía um volume excessivo de variáveis (426 colunas), muitas das quais estavam fora do escopo definido para o projeto.

3. Transformação (ETL)

Para solucionar os problemas identificados e estruturar os dados para a modelagem, foi implementado um processo de **Extração, Transformação e Carga (ETL)**.

- **Extração (Extract):** Os dados foram extraídos do arquivo .csv original e carregados em um DataFrame utilizando a biblioteca Pandas da linguagem Python.
- **Transformação (Transform):** Esta fase concentrou as atividades de limpeza, normalização e padronização. As principais ações foram:
 - **Limpeza:** Tratamento de valores nulos, substituindo-os por 0 em campos numéricos para viabilizar operações matemáticas.
 - **Normalização:** Conversão dos dados categóricos de seus códigos numéricos para seus respectivos valores textuais (ex: o código 1 na coluna TP_LOCALIZACAO foi transformado para "Urbana").
 - **Padronização:** Renomeação de todas as colunas selecionadas para nomes claros e padronizados em português (ex: NO_ENTIDADE para nome_escola), facilitando a compreensão e o uso futuro da base.
- **Carga (Load):** O conjunto de dados final, já limpo e transformado, foi carregado (salvo) em uma nova estrutura temporária, um novo arquivo .csv (dados_censo_escolar_tratados.csv), com a codificação universal **UTF-8**, garantindo a integridade dos dados e a correta exibição de todos os caracteres.

4. Documentação dos Procedimentos

O processo de ETL foi automatizado através de um script desenvolvido em Python, com o uso da biblioteca Pandas. O script segue uma lógica sequencial clara, projetada para garantir a reprodutibilidade e a transparência das decisões tomadas.

A execução inicia-se com a **fase de extração**, onde o script lê o arquivo .csv bruto. Foi implementada uma rotina de tratamento de exceções para a codificação de caracteres: o script primeiro tenta ler o arquivo com o padrão UTF-8; caso ocorra um erro de decodificação (UnicodeDecodeError), ele automaticamente executa uma segunda tentativa com a codificação latin-1, garantindo que os dados sejam carregados corretamente.

Na **fase de transformação**, a primeira decisão foi a **seleção de variáveis**. Com base no escopo do projeto, foi criada uma lista contendo apenas as 17 colunas de interesse, resultando na criação de um novo DataFrame, mais enxuto e focado. Em seguida, foi aplicado um processo de **renomeação**, utilizando um dicionário Python para mapear os nomes técnicos originais para nomes intuitivos e padronizados.

A etapa seguinte foi o **tratamento dos dados**. Foram criados dicionários de mapeamento específicos para cada variável categórica a ser decodificada. A função .map() do Pandas foi utilizada para aplicar essas transformações, substituindo os códigos numéricos pelos seus correspondentes textuais. Para o tratamento de dados faltantes, a

decisão foi utilizar o método `.fillna()` para substituir valores nulos por 0 em todas as colunas quantitativas. Adicionalmente, o script enriqueceu a base ao criar uma nova coluna, `qtd_total_computadores_alunos`, que agrega a soma de três outras colunas.

Finalmente, na **fase de carga**, o script salva o DataFrame transformado em um novo arquivo `.csv`, especificando o uso da codificação UTF-8 para assegurar a compatibilidade e a correta exibição dos dados em qualquer sistema. O processo foi documentado para ser executado de forma autônoma, gerando um resultado final limpo e consistente a partir do arquivo bruto.