UNIVERSIDADE FEDERAL DO CEARÁ
CURSO DE CIÊNCIA DA COMPUTAÇÃO

CIÊNCIA DE DADOS

T1: Definição e Manipulação de Dados

LUIZ GUSTTAVO MACEDO MAGALHÃES - 556217

Crateús
2025

1. Introdução

Este relatório apresenta os resultados do Trabalho T1, cujo objetivo foi selecionar, explorar e documentar três conjuntos de dados abertos. O processo envolveu a manipulação inicial dos dados utilizando a biblioteca Pandas em ambiente Jupyter Notebook, a análise de seus atributos e a estruturação de um projeto de ciência de dados organizado.

2. Conjuntos de Dados e Fontes

Abaixo estão os links de onde os dados foram extraídos, acompanhados de um breve resumo de cada um.

2.1. Iris Dataset

Um conjunto de dados clássico da área de aprendizado de máquina, contendo 150 amostras de 3 espécies de flores Iris. O objetivo é classificar a espécie com base nas medidas de suas pétalas e sépalas.

- **Fonte:** UCI Machine Learning Repository
- **Link:** https://archive.ics.uci.edu/dataset/53/iris

2.2. Heart Disease Dataset

Este dataset contém 303 amostras e 14 atributos clínicos de pacientes para o diagnóstico de doenças cardíacas. Os dados são provenientes da Cleveland Clinic Foundation. Na análise, foi utilizada a versão "processada", onde os dados faltantes, marcados com ?, foram tratados.

- **Fonte:** UCI Machine learning Repository
- **Link:** https://archive.ics.uci.edu/dataset/45/heart+disease

2.3. COVID-19 Brasil Dataset

Um conjunto de dados grande com o histórico de casos e óbitos por COVID-19 no Brasil, detalhado por data e localidade. Os dados foram obtidos através da iniciativa Brasil.IO, que consolida boletins das secretarias de saúde. Devido ao grande volume do arquivo (mais de 100 MB), a solução adotada foi carregá-lo diretamente de sua fonte online, garantindo a reprodutibilidade da análise sem sobrecarregar o git.

- **Fonte:** Brasil.IO
- **Link:** https://data.brasil.io/dataset/covid19/caso_full.csv.gz

3. Ferramentas e Metodologia

- **Linguagem:** Python 3
- **Bibliotecas Principais:** Pandas para manipulação e análise dos dados.
- **Ambiente:** Visual Studio Code com a extensão Jupyter para a criação dos notebooks de exploração.
- **Controle de Versão:** Git e GitHub.

O trabalho seguiu a estrutura de pastas padrão para projetos de ciência de dados, com a documentação dos atributos na pasta references e os notebooks de exploração na pasta notebooks.