

Trabalho Prático — DGT2823

Tecnologias para Desenvolvimento de Soluções de Big Data

Aluna: Fernanda Costa

Curso: Engenharia de Software (Full Stack)

Período: 2025.2

1. Contextualização

Como analista de dados, recebi um dataset contendo informações de exercícios físicos diários. O objetivo foi limpar e preparar os dados para que possam ser utilizados futuramente em análises de padrões ou modelos de machine learning. O conjunto de dados original (dados.csv) contém as colunas: ID, Duration, Date, Pulse, Maxpulse e Calories.

2. Etapas do Desenvolvimento

Leitura do arquivo CSV

```
import pandas as pd

df = pd.read_csv('dados.csv', sep=';', engine='python')
print(df.head())
```

Visualização inicial

```
print(df.info())
print(df.head(5))
print(df.tail(5))
```

Criação de cópia do dataframe

```
df_limpo = df.copy()
```

Substituição de valores nulos na coluna 'Calories'

```
df_limpo['Calories'].fillna(0, inplace=True)
print(df_limpo)
```

Tratamento da coluna 'Date'

```
df_limpo['Date'].fillna('1900/01/01', inplace=True)
df_limpo['Date'] = df_limpo['Date'].replace('20201226', '2020/12/26')
df_limpo['Date'] = pd.to_datetime(df_limpo['Date'], format='%Y/%m/%d', errors='coerce')
```

Remoção de registros nulos restantes

```
df_limpo.dropna(subset=['Date'], inplace=True)
print(df_limpo)
```

3. Resultado Final e Conclusão

Após as transformações, o dataset ficou padronizado e pronto para uso. Todas as colunas numéricas foram tratadas e as datas convertidas para o tipo datetime. O código foi testado no ambiente Jupyter Notebook (Python 3.12) e executou sem erros.

4. Referências

• Análise de Dados em Python com Pandas — Prof. Fernando Cardoso Durier da Silva • Big Data Analytics e TensorFlow — Prof. Fernando Cardoso Durier da Silva • Princípios de Desenvolvimento de Spark com Python — Prof. Sérgio Assunção Monteiro • Roteiro do Trabalho Prático DGT2823 — Estácio