**UNIVERSIDADE ESTÁCIO DE SÁ**

**DESENVOLVIMENTO FULL STACK**

# Mundo 05 - Nível 03

## Missão Prática

## Tratando a imensidão dos dados

Leonardo Schaffer Mota

Matrícula - 202205090981

SANTO ANDRÉ – SP

Agosto de 2024

## Introdução

A missão prática envolveu o tratamento e limpeza de um conjunto de dados utilizando a biblioteca Pandas da linguagem Python. O objetivo principal foi preparar os dados para futuras análises, garantindo que eles estivessem em um formato adequado para mineração e interpretação de dados. O conjunto de dados fornecido continha informações sobre atividades físicas, incluindo colunas como ID, Duration, Date, Pulse, Maxpulse e Calories.

**Importação de Dados:** O primeiro passo foi importar o conjunto de dados a partir de um arquivo CSV. Para isso, utilizei a função “read_csv” da biblioteca Pandas, especificando o separador de colunas (';'), a engine de leitura e o encoding dos dados.

```python
import pandas as pd

# Nome do arquivo CSV
csv_file = 'dados1.csv'

# Ler o arquivo CSV
df = pd.read_csv(csv_file, sep=';', engine='python', encoding='utf-8')

# Verificar importação
print("Informações Gerais do DataFrame:")
print(df.info())

print("\nPrimeiras 5 Linhas:")
print(df.head())

print("\nÚltimas 5 Linhas:")
print(df.tail())
```

**Criação de uma Cópia dos Dados:** Para evitar alterações no conjunto de dados original, criei uma cópia dos dados importados.

```python
# Criar uma cópia dos dados
df_copy = df.copy()
```

**Tratamento de Valores Nulos:** Substituí os valores nulos na coluna Calories por 0 e na coluna Date por uma data padrão ('1900/01/01').

```
# Substituir valores nulos na coluna 'Calories' por 0
df_copy['Calories'].fillna(0, inplace=True)
print("\nDataFrame após substituir valores nulos na coluna 'Calories':")
print(df_copy)

# Substituir valores nulos na coluna 'Date' por '1900/01/01'
df_copy['Date'].fillna('1900/01/01', inplace=True)
print("\nDataFrame após substituir valores nulos na coluna 'Date':")
print(df_copy)
```

**Correção de Formato de Datas:** Para corrigir datas no formato YYYYMMDD, utilizei uma combinação dos métodos “replace” e “to_datetime”.

```
# Corrigir formato específico de datas
df_copy['Date'] = df_copy['Date'].str.strip("'")
df_copy['Date'] = df_copy['Date'].astype(str).replace({'20201226': '2020/12/26'})
df_copy['Date'] = pd.to_datetime(df_copy['Date'], format='%Y/%m/%d', errors='coerce')
print("\nDataFrame após corrigir datas no formato '20201226':")
print(df_copy)
```

**Transformação de Datas:** Convertendo a coluna Date para o tipo datetime e substituindo datas inválidas ('1900/01/01') por “NaN”.

```
# Transformar a coluna 'Date' em datetime
df_copy['Date'] = pd.to_datetime(df_copy['Date'], format='%Y/%m/%d', errors='coerce')
print("\nDataFrame após transformar a coluna 'Date' em datetime:")
print(df_copy)

# Transformar na coluna Date o valor '1900/01/01' por 'NaN'
df_copy['Date'].replace(pd.Timestamp('1900-01-01'), pd.NaT, inplace=True)
print(df_copy)
```

**Remoção de Registros com Valores Nulos:** Remover registros que ainda continham valores nulos na coluna Date.

```
# Remover registros com valores nulos na coluna 'Date'
df_clean = df_copy.dropna(subset=['Date'])
print("\nDataFrame após remover registros com valores nulos na coluna 'Date':")
print(df_clean)
```

**Resultados:**

O resultado foi um DataFrame limpo e pronto para análise, onde:

- Valores nulos na coluna “Calories” foram substituídos por 0.

- Valores nulos na coluna “Date” foram inicialmente substituídos por '1900/01/01', corrigidos para o formato datetime, e posteriormente substituídos por “NaN”.
- Registros com datas inválidas foram removidos.

## Conclusão

A prática de tratamento e limpeza de dados é crucial para garantir a qualidade das análises subsequentes. Esta atividade proporcionou uma compreensão prática das operações de limpeza de dados, desde a substituição de valores nulos até a correção de formatos de datas. Utilizando a biblioteca Pandas, foi possível preparar um conjunto de dados adequadamente estruturado e livre de inconsistências, pronto para ser utilizado em tarefas de mineração e análise de dados.