



Missão Prática | Tratando a Imensidão dos Dados

Aluno: Eros Santos de Vasconcelos / 202307120545

Polo: Iputinga-PE

Disciplina: RPG0033 – Tratando a imensidão de dados

1. Objetivo do Projeto

O objetivo deste projeto é tratar um conjunto de dados utilizando Python e a biblioteca Pandas, assegurando a limpeza, padronização e preparação para futuras análises.

2. Características Principais do Projeto

- Leitura de dados de um CSV simulado via StringIO.
- Exploração inicial com `df.info()`, `head()` e `tail()`.
- Tratamento de valores ausentes (substituição e remoção).
- Conversão de strings de data para `datetime`. - Remoção de registros inconsistentes.

3. Procedimentos de Desenvolvimento

3.1 Leitura dos Dados:

```
df = pd.read_csv(StringIO(dados_csv), sep=';', quotechar="")
```

3.2 Exploração Inicial: `df.info()`, `df.head()`, `df.tail()`

3.3 Cópia dos Dados:

```
df_clean = df.copy()
```

3.4 Tratamento de NaN:

```
df_clean['Calories'] = df_clean['Calories'].fillna(0)
```

```
df_clean['Date'] = df_clean['Date'].fillna("1900/01/01")
```

3.5 Conversão para `datetime` e ajustes: substituição de '1900/01/01' por `pd.NA`, remoção de aspas, conversão específica de '20201226', e `to_datetime`.

3.6 Remoção de registros inválidos:

```
df_clean.dropna(subset=['Date'])
```

4. Resultados Obtidos

- Dataset final com 30 registros.
- Coluna Date em formato `datetime`.
- Coluna Calories sem valores nulos.
- Base pronta para análise subsequente.

5. Conclusão

O processo demonstrou as principais técnicas de limpeza e padronização de dados, essenciais para qualquer etapa de mineração e análise de dados.