

TRABALHO 4:

Conjuntos de dados (datasets) utilizados

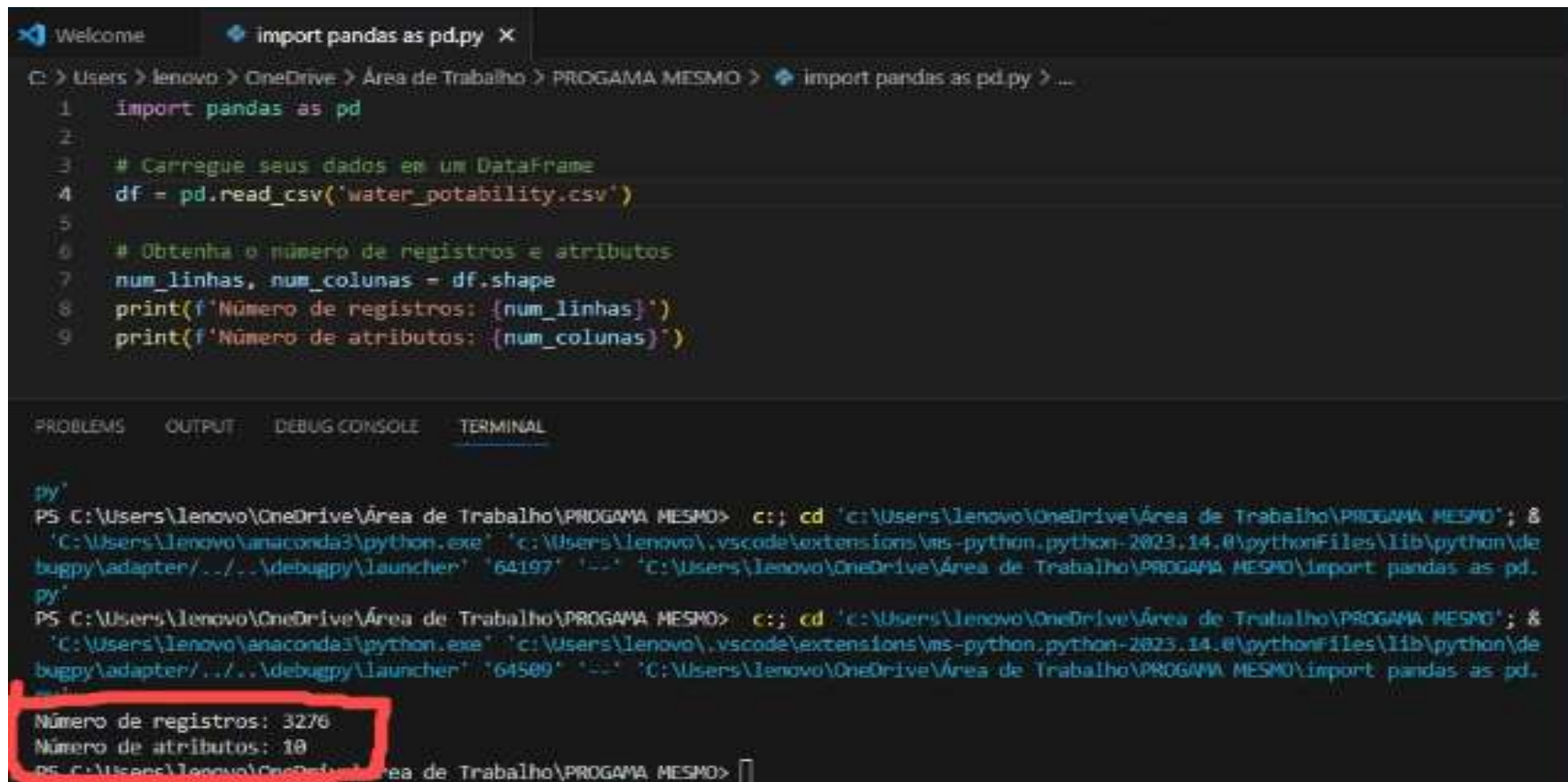
Diogo Campos
Vinícius Troiano

ESCOLHA DO DATASET

- A princípio, foi decidido que iremos trabalhar com algo relacionado a qualidade de água.
- Após isso, procuramos em sites como o Kaggle e o GitHub datasets que tinham relação com esse tema.
- Assim, foi encontrado no site Kaggle, um conjunto de dados que relaciona vários atributos específicos de determinada amostra de água com sua potabilidade.
- A partir do site, era possível ver que esse dataset possuía um bom numero de dados e que existiam vários trabalhos utilizando ele.
- Assim, decidimos utilizar esse dataset para o nosso trabalho também.

CARACTERÍSTICAS DO DATASET

- Para descobrir o número de atributos e de registros desse dataset, utilizamos o seguinte código em Python



The image shows a screenshot of a Visual Studio Code editor window. The top part displays a Python script in a file named 'import pandas as pd.py'. The script imports pandas, reads a CSV file named 'water_potability.csv', and prints the number of rows and columns. The bottom part shows the terminal output of the script, which confirms the dataset has 3276 records and 10 attributes. The output is highlighted with a red box.

```
1 import pandas as pd
2
3 # Carregue seus dados em um DataFrame
4 df = pd.read_csv('water_potability.csv')
5
6 # Obtenha o número de registros e atributos
7 num_linhas, num_colunas = df.shape
8 print(f'Número de registros: {num_linhas}')
9 print(f'Número de atributos: {num_colunas}')
```

py
PS C:\Users\lenovo\OneDrive\Área de Trabalho\PROGAMA MESMO> c:: cd 'c:\Users\lenovo\OneDrive\Área de Trabalho\PROGAMA MESMO'; & 'C:\Users\lenovo\anaconda3\python.exe' 'c:\Users\lenovo\.vscode\extensions\ms-python.python-2023.14.0\pythonFiles\lib\python\debugpy\adapter\..\..\debugpy\launcher' '64197' '--' 'C:\Users\lenovo\OneDrive\Área de Trabalho\PROGAMA MESMO\import pandas as pd.py'
PS C:\Users\lenovo\OneDrive\Área de Trabalho\PROGAMA MESMO> c:: cd 'c:\Users\lenovo\OneDrive\Área de Trabalho\PROGAMA MESMO'; & 'C:\Users\lenovo\anaconda3\python.exe' 'c:\Users\lenovo\.vscode\extensions\ms-python.python-2023.14.0\pythonFiles\lib\python\debugpy\adapter\..\..\debugpy\launcher' '64509' '--' 'C:\Users\lenovo\OneDrive\Área de Trabalho\PROGAMA MESMO\import pandas as pd.py'
Número de registros: 3276
Número de atributos: 10
PS C:\Users\lenovo\OneDrive\Área de Trabalho\PROGAMA MESMO> █

CARACTERISTICAS DO DATASET

Os 10 atributos desse dataset são:

- pH: O nível de pH da água.
- Dureza: Dureza da água, uma medida do conteúdo mineral.
- Sólidos: Total de sólidos dissolvidos na água.
- Cloraminas: Concentração de cloraminas na água.
- Sulfato: Concentração de sulfato na água.
- Condutividade: Condutividade elétrica da água.
- Organic_carbon: Conteúdo de carbono orgânico na água.
- Trihalometanos: Concentração de Trihalometanos na água.
- Turbidez: Nível de turbidez, uma medida da clareza da água.
- Potabilidade: Indica se a água é potável ou não potável.

CARACTERISTICAS DO DATASET

- Os nove primeiros atributos são do tipo contínuo, enquanto o último atributo é do tipo discreto.
- O décimo atributo “potabilidade” é o atributo classe, nele existem duas respostas possíveis 1 (potável) e 0 (não potável).

Ou seja, esse dataset possui apenas duas classes:

- Classe 1: Água potável;
- Classe 2: Água não potável.

CARACTERÍSTICAS DO DATASET

- Para descobrir a proporção “classes vs. Registros” desse dataset, utilizamos o seguinte código em Python

```
C: > Users > lenovo > OneDrive > Área de Trabalho > PROGAMA MESMO > import pandas as pd.py > ...
1  import pandas as pd
2
3  # Carregar o conjunto de dados
4  df = pd.read_csv('water_potability.csv')
5
6  # Calcular a contagem de cada resultado
7  contagem_resultados = df['Potability'].value_counts()
8
9  # Calcular a porcentagem de cada resultado
10 porcentagem_resultados = (contagem_resultados / len(df)) * 100
11
12 # Exibir o resultado
13 print(porcentagem_resultados)
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
PS C:\Users\lenovo\OneDrive\Área de Trabalho\PROGAMA MESMO> c:; cd 'c:\Users\lenovo\OneDrive\Área de Trabalho\PROGAMA MESMO'; &
'C:\Users\lenovo\anaconda3\python.exe' 'c:\Users\lenovo\.vscode\extensions\ms-python.python-2023.16.8\pythonFiles\lib\python\de
bugpy\adapter/.../..\.debugpy\launcher' '60340' '--' 'C:\Users\lenovo\OneDrive\Área de Trabalho\PROGAMA MESMO\import pandas as pd.
py'
0    60.989011
1    39.010989
Name: Potability, dtype: float64
PS C:\Users\lenovo\OneDrive\Área de Trabalho\PROGAMA MESMO> 
```

CONCLUSÃO

- Podemos concluir que esse dataset possui uma boa proporção de registros, atributos e classes.
- Podemos dizer também que ele possui classes razoavelmente equilibradas.
- Essas características possivelmente facilitaram nas próximas etapas desse projeto.
- O dataset utilizado no nosso trabalho pode ser obtido a partir do seguinte link:
<https://www.kaggle.com/datasets/uom190346a/water-quality-and-potability?resource=download>;