Tópicos Especiais em Sistemas de Informação

Tipos de variáveis, Importação de dados, dataframes

Ely – elydasilvamiranda@gmail.com

Biblioteca e arquivos necessários

- Além da Scikit Learn, Numpy e Scipy, precisaremos do Pandas

  > pip install pandas

  – Documentação:

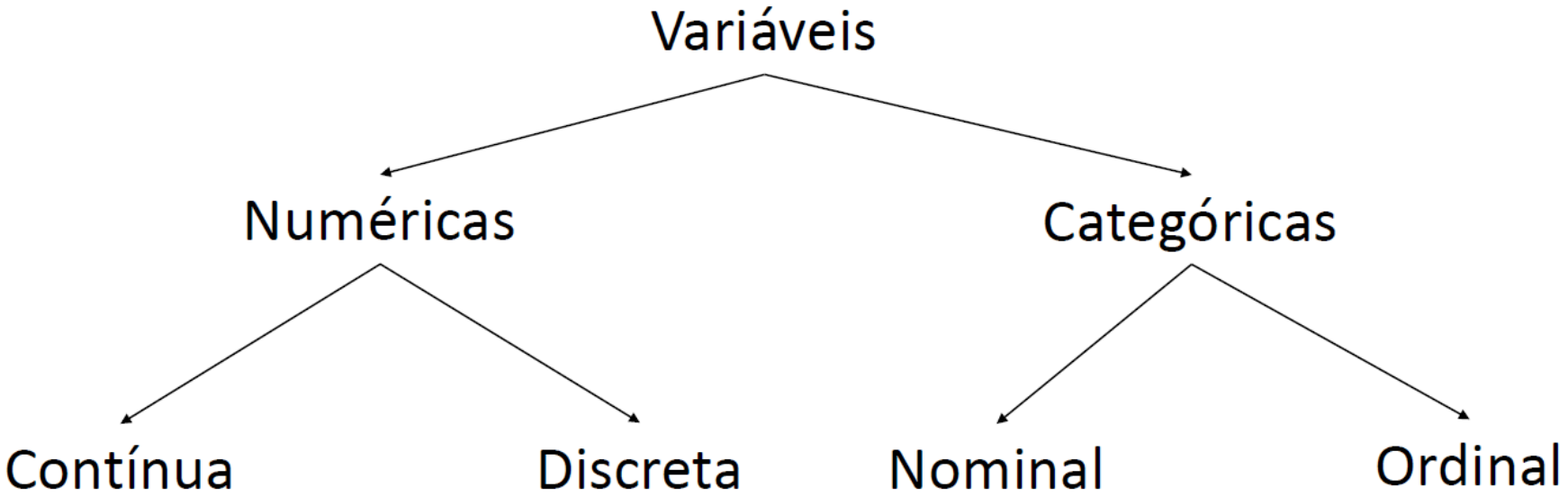
    https://pandas.pydata.org

- Arquivos disponíveis no drive da disciplina:

  – acessos.csv;

  – acessos_buscas.csv.

Tipos de variáveis



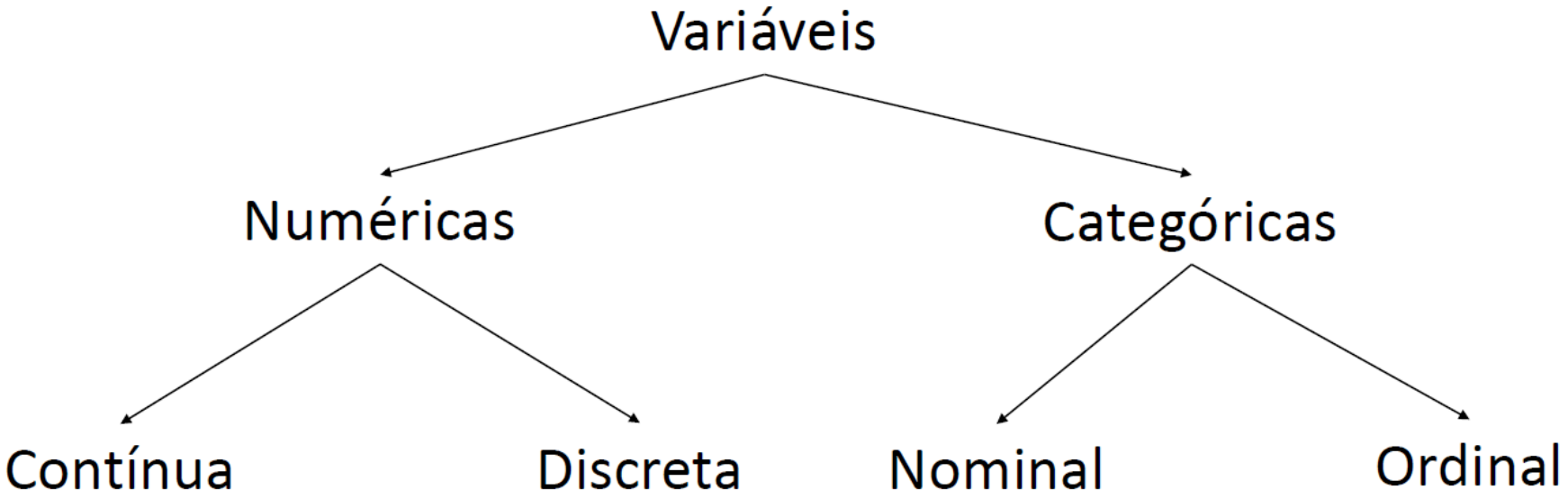
Tipos de Variáveis

- Numéricas: Representadas por números;
- Subclassificação:
  - Contínuas: Números reais.
    - Peso: 50,00, 80,00, 60,00;
    - Salário: 3.500,00;
    - Temperatura: 35,5;
  - Discretas: números inteiros, representando contagens;
    - quantidade de filhos: 3;
    - vínculos empregatícios: 1

Tipos de variáveis

- Categóricas: representadas por strings;
- Subclassificação:
  - Nominais: representam dados não mensuráveis.
    - Cor dos olhos: castanhos, azuis, verdes;
    - Gênero: masculino, feminino;
    - Etinia: Branco, negro, pardo etc;
    - Nomes e Ids
  - Ordinais: categorizáveis por ordenação.
    - Tamanho: P, M, G;
    - Cargo: Júnior, Pleno, Senior;
    - Patente: Cabo, Soldado, Sargento, Tenente.

Tipos de variáveis



Datasets em CSV

- As bases de dados que utilizaremos, os nossos **datasets**, tipicamente estão em formato CSV;
- Os arquivos CSV são formados por linhas contendo “registros”;
- Cada registro possui colunas separadas por vírgulas. Ex:

```
home,como_funciona,contato,comprou
1,1,0,0
1,1,0,0
1,1,0,0
1,1,0,0
1,1,0,0
1,0,1,1
```

```
clientid,income,age,loan,default
1,66155.9250950813,59.017015066929204,8106.53213128514,0
2,34415.1539658196,48.11715310486029,6564.745017677379,0
3,57317.1700630337,63.10804949188599,8020.953296386469,0
4,42709.534200839706,45.751972352154596,6103.642260140699,0
5,66952.68884534019,18.5843359269202,8770.09923520439,1
6,24904.064140282597,57.4716071025468,15.498598437827198,0
7,48430.3596126847,26.809132419060898,5722.58198121271,0
8,24500.1419843175,32.8975483207032,2971.00330971188,1
9,40654.8925372772,55.496852539479704,4755.8252798016,0
10,25075.872770976297,39.7763780555688,1409.23037111453,0
```

Biblioteca PANDAS e arquivos CSV

- PANDAS (*Python Data Analysis Library);*
- Possui muitas funções pra nos auxiliar com datasets sem codificação extenuante;
- Uma das muitas funções é ler/carregar arquivos CSV;
- O resultado da leitura é uma estrutura de dados semelhante a uma lista em python;
- Essa lista é mais especificamente chamada de como **dataframe.**
  - *https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.html*

Importando um arquivo CSV

```
import pandas as pd
df = pd.read_csv('acessos.csv')
print(df)
```

home	como_funciona	contato	comprou
1	1	0	0
1	1	0	0
1	1	0	0
1	1	0	0
1	1	0	0
1	0	1	1

Este exemplo considera que o arquivo CSV está na pasta de onde o código ou notebook esteja rodando

Operações corriqueiras

Exibir registros

- dataframe.head(n): imprime as n primeiras linhas de dados:

```
import pandas as pd

df = pd.read_csv('acessos.csv')

df.head(3)
```

	home	como_funciona	contato	comprou
0	1	1	0	0
1	1	1	0	0
2	1	1	0	0

Exibir colunas

- Exibir apenas uma coluna específica:

```
import pandas as pd

df = pd.read_csv('credit-data.csv')

df['age']  # ou df.age
```

- Nota, algumas operações podem ser combinadas:

```
df.age.head(3)
0    59.017015
1    48.117153
2    63.108049
Name: age, dtype: float64
```

Descrever dataframe

- describe(): mostra dados estatísticos das colunas de um dataframe

```
import pandas as pd

df = pd.read_csv('credit-data.csv')

df.describe()
```

Descrever dataframe

- describe(): mostra dados estatísticos das colunas de um dataframe

	clientid	income	age	loan	default
count	2000.000000	2000.0000	1997.000000	2000.000000	2000.000000
mean	1000.500000	45331.600018	40.807559	4444.369695	0.141500
std	577.494589	14326.327119	13.624469	3045.410024	0.348624
min	1.000000	20014.489470	-52.423280	1.377630	0.000000
25%	500.750000	32796.459717	28.990415	1939.708847	0.000000
50%	1000.500000	45789.117313	41.317159	3974.719419	0.000000
75%	1500.250000	57791.281668	52.587040	6432.410625	0.000000
max	2000.000000	69995.685578	63.971796	13766.051239	1.000000

Retirar linhas ou colunas

- drop(coluna, eixo, inplace): exclui uma linha ou coluna;

```
import pandas as pd

df = pd.read_csv('credit-data.csv')

df.drop("clientid", axis=1, inplace=True)

df.head(3)
```

	income	age	loan	default
0	66155.925095	59.017015	8106.532131	0
1	34415.153966	48.117153	6564.745018	0
2	57317.170063	63.108049	8020.953296	0

Valores únicos

- unique(): aplicada a uma coluna de um dataframe, exibe quantos valores distintos existem;

```
df = pd.read_csv('census.csv')

df['relationship'].unique()
```

```
array([' Not-in-family', ' Husband', ' Wife', ' Own-child',
' Unmarried’, ' Other-relative'], dtype=object)
```

Filtragens

- Filtragem de colunas: aplicam-se operadores às colunas e apenas resultados verdadeiros à expressão são retornados;

```
import pandas as pd

df = pd.read_csv('credit-data.csv')

df[df['age'] > 63.9]
```

	clientid	income	age	loan	default
404	405	62553.668411	63.924976	4641.704785	0
942	943	29178.977588	63.930735	1664.386062	0
1998	1999	43756.056605	63.971796	1622.722598	0

Estatísticas de colunas

- Estatísticas de colunas: há muitas funções úteis para estatística, como média, mínimo e máximo valores etc;

```
import pandas as pd

df = pd.read_csv('credit-data.csv')

print(df.age.mean())

print(df.age.max())

print(df.age.min())

40.80755937840458

63.97179584112021

-52.423279919661596
```

Valores nulos

- isnull(): retorna True ou False;

```
df.age.isnull().sum()
```

- Ou ainda:

```
import pandas as pd
df = pd.read_csv('credit-data.csv')

for field in df.columns:
    print(field, 'NaN:', df[field].isnull().sum())
```

Tópicos Especiais em Sistemas de Informação

Tipos de variáveis, Importação de dados, dataframes

Ely – elydasilvamiranda@gmail.com