Correlação de Pearson

- Variável aumenta e outra diminui
- Variáveis aumentam juntas
- Variáveis diminuem juntas

Exemplo:

Correlação entre salário e quantidade de bens, quanto maior o salário maior quantidade de bens, e vice-versa.

Motivo de usar correlação:

Eliminar variáveis que são muito correlacionadas, pois podem não ser efetivas quando uma variável cresce e outra também, sendo uma cópia redundante nos dados.

Problema:

Se tiverem variáveis muito correlacionadas, a característica pode receber peso dobrado ao treinar o modelo.

Cálculo da correlação (Pearson):

Varia de -1 até 1

Correlação = 1: Positiva perfeita
Correlação = -1: Negativa perfeita
Correlação = 0: As variáveis não dependem linearmente uma da outra, mas pode haver uma dependência não linear, sendo necessário investigação por outros meios.

0.9 para mais ou menos indica uma correlação muito forte
0.7 a 0.9 positivo ou negativo indica uma correlação forte
0.5 a 0.7 positivo ou negativo indica uma correlação moderada
0.3 a 0.8 positivo ou negativo indica uma correlação fraca
0 a 0.3 positivo ou negativo indica uma correlação desprezível

$$\rho = \frac{\sum_{i-1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X) \cdot \operatorname{var}(Y)}}$$

Basicamente acima temos o cáculo da correlação de pearson que nada mais é do que a covariância das duas variáveis, dividido pela raiz da multiplicação entre a variância de cada variável.

Covariância é calculada pelo somatório do valor de cada amostra menos a média da variável X, multiplicando pelo somatório do valor de cada amostra menos a média de Y

Variância é feita pelo somatório das amostras menos a média elevado ao quadrado.

```
import pandas as pd

pd.set_option('display.width', 320) // Ajustar tamanho do dataframe do
pandas

colunas = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age',
'class']

dados = pd.read_csv('pima-indians-diabetes.csv', names = colunas)

print(dados.corr(method = 'pearson'))
```

	preg	plas	pres	skin	test	mass	pedi	age	class
preg	1.000000	0.129459	0.141282	-0.081672	-0.073535	0.017683	-0.033523	0.544341	0.221898
plas	0.129459	1.000000	0.152590	0.057328	0.331357	0.221071	0.137337	0.263514	0.466581
pres	0.141282	0.152590	1.000000	0.207371	0.088933	0.281805	0.041265	0.239528	0.065068
skin	-0.081672	0.057328	0.207371	1.000000	0.436783	0.392573	0.183928	-0.113970	0.074752
test	-0.073535	0.331357	0.088933	0.436783	1.000000	0.197859	0.185071	-0.042163	0.130548
mass	0.017683	0.221071	0.281805	0.392573	0.197859	1.000000	0.140647	0.036242	0.292695
pedi	-0.033523	0.137337	0.041265	0.183928	0.185071	0.140647	1.000000	0.033561	0.173844
age	0.544341	0.263514	0.239528	-0.113970	-0.042163	0.036242	0.033561	1.000000	0.238356
class	0.221898	0.466581	0.065068	0.074752	0.130548	0.292695	0.173844	0.238356	1.000000

Plotando o mapa de calor da correlação das variáveis do Dataframe:

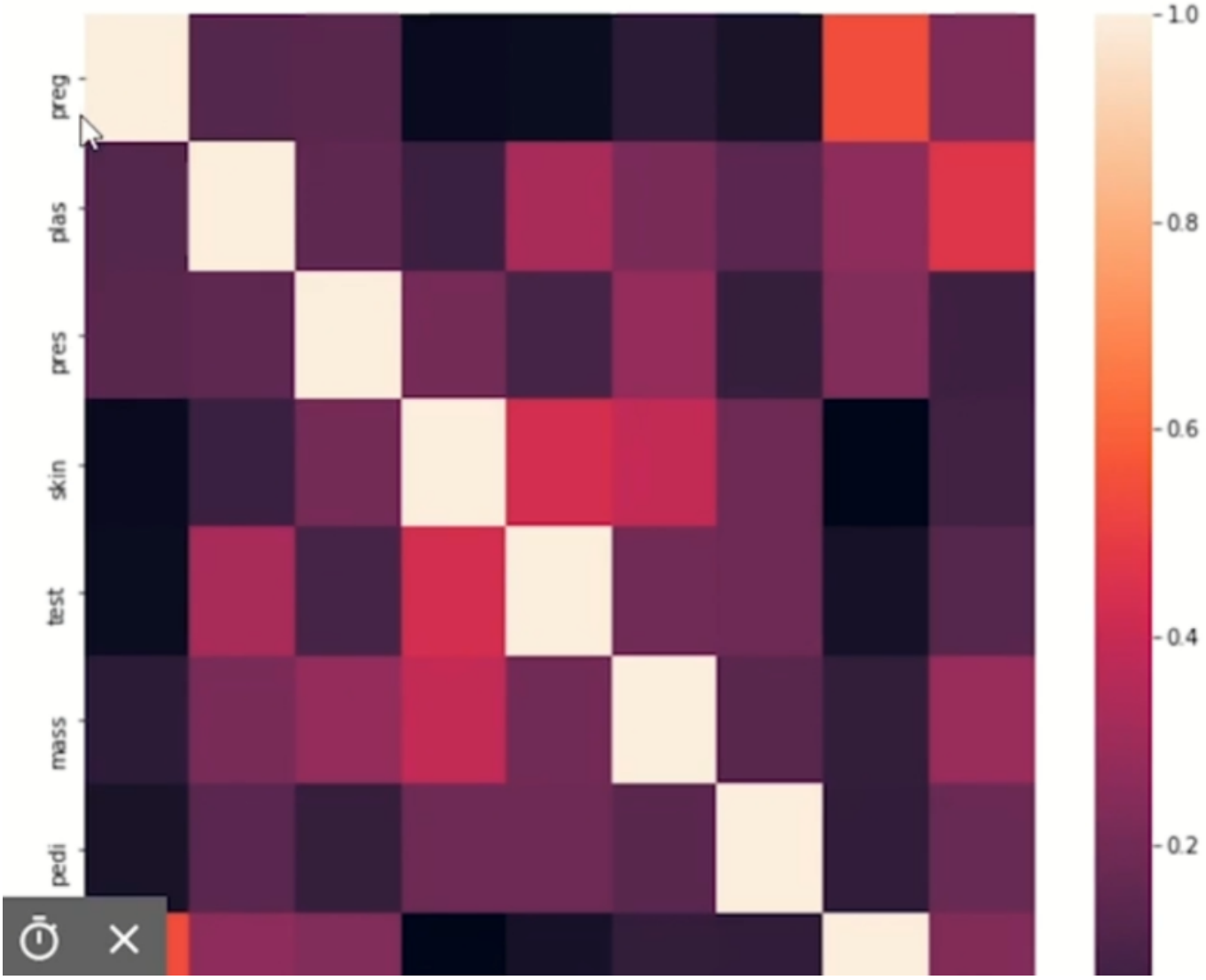
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

colunas = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age',
'class']

dados = pd.read_csv('pima-indians-diabetes.csv', names = colunas)

plt.figure(figsize=(10, 10))

sns.heatmap(dados.corr())
```


preg
plas
pres
skin
test
mass
pedi
1.0
0.8
0.6
0.4
0.2