
Mineração de Dados

Transformação de Dados

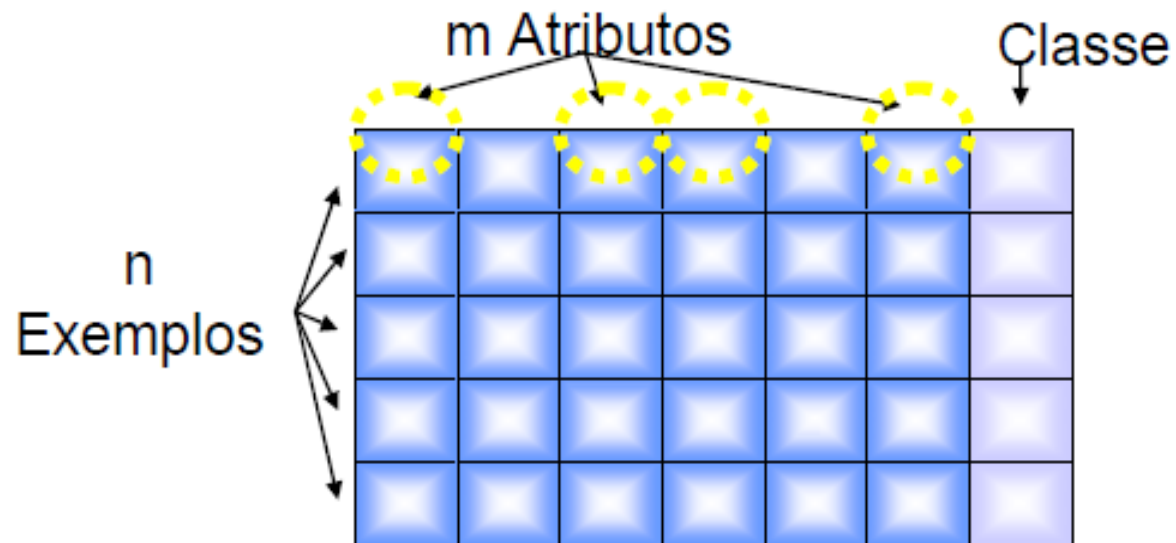
João Carlos Xavier Júnior

jcxavier@imd.ufn.br

Redução

❑ Redução de dimensões:

- ❖ A supressão de uma coluna (**atributo**) é muito mais **delicada** do que a supressão de uma linha (**padrão**).

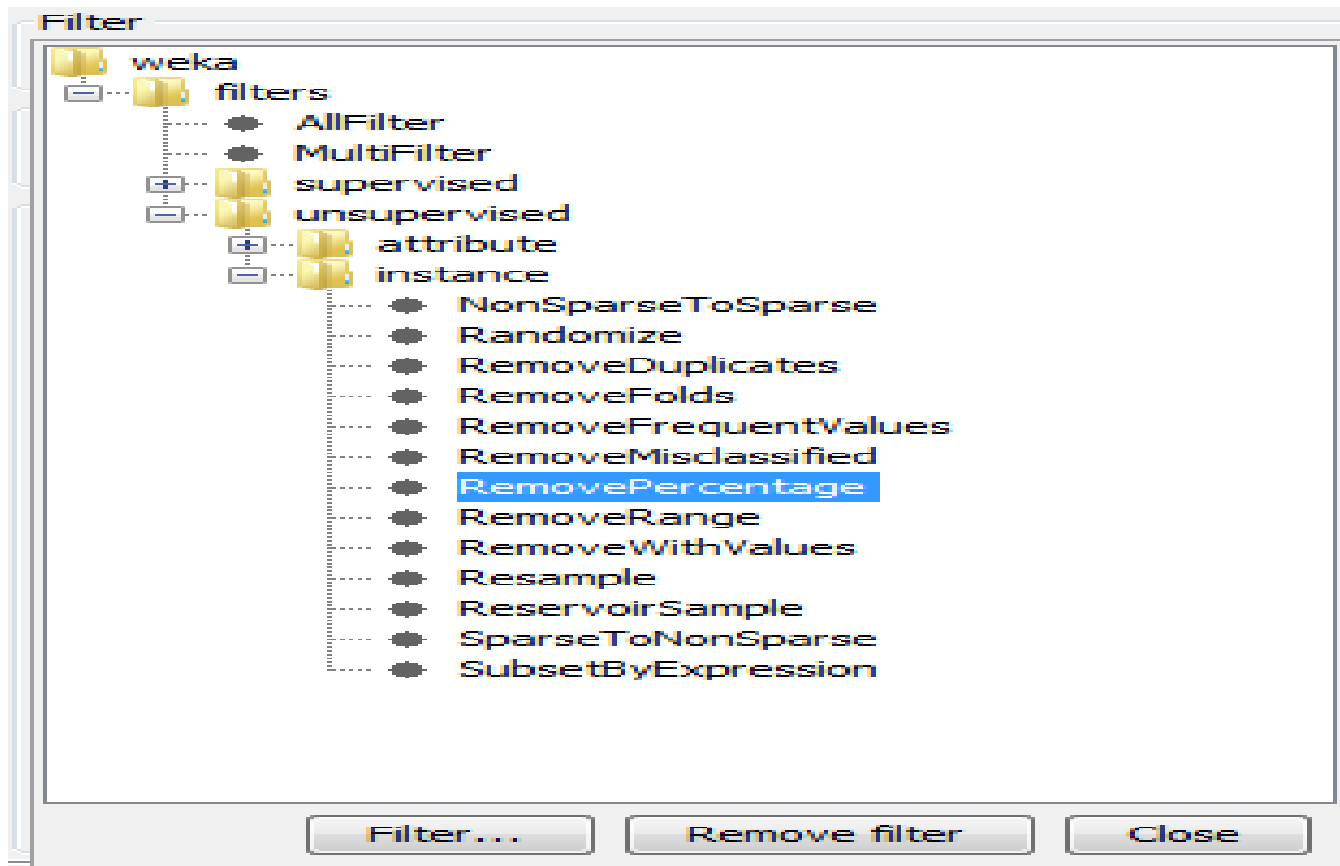


Redução de Instâncias

- ❑ Amostragem Incremental:
 - ❖ Permite a redução do número de instâncias.
 - ❖ Expressa em percentual:
 - 10%
 - 30%;
 - 50%;
 - 70%;
 -

Amostragem Incremental

□ Exemplo - WEKA:



Redução de Atributos

- ❑ Redução de dimensão:
 - Significa retirar atributos.
 - Selecionar atributos.
- ❑ Quais ou como?



Seleção de Atributos

❑ Manual:

- Melhor método se for baseado em um entendimento profundo sobre:
 - O problema de aprendizado;
 - O significado de cada atributo (**Correlação de Pearson**).

❑ Automático:

- Métodos de Seleção de atributos:
 - Não-supervisionada;
 - Supervisionada.

Correlação de Pearson

- O coeficiente de correlação de Pearson (r):
 - Mede o grau da correlação linear entre duas variáveis quantitativas.
 - É um índice com valores situados entre -1,0 e 1,0.
 - $r = 1$, significa uma correlação perfeita positiva entre as duas variáveis;
 - $r = -1$, significa uma correlação negativa perfeita entre as duas variáveis, se uma aumenta, a outra sempre diminui.
 - $r = 0$, significa que as **duas variáveis não dependem linearmente uma da outra.**

Correlação de Pearson

- O coeficiente de **correlação de Pearson** (**r**) é uma medida de associação linear entre variáveis. Sua fórmula é a seguinte:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

where:

- n is the sample size
- x_i, y_i are the individual sample points indexed with i
- $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ (the sample mean); and analogously for \bar{y}

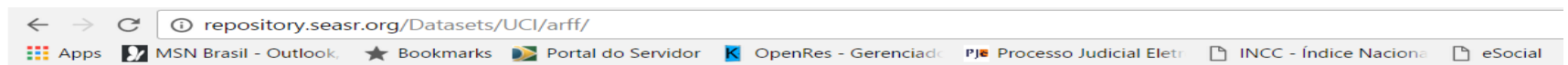
Correlação de Pearson

- Como interpretar a Correlação de Pearson:
 - ❖ $r = 0,10$ até $0,30$ (fraco);
 - ❖ $r = 0,40$ até $0,6$ (moderado);
 - ❖ $r = 0,70$ até 1 (forte).

Correlação de Pearson

Dataset:

❖ <https://github.com/renatopp/arff-datasets/tree/master/classification>



Index of /Datasets/UCI/arff/

../		
anneal.ORIG.arff	15-Jul-2008 15:27	84017
anneal.arff	15-Jul-2008 15:27	143336
arrhythmia.arff	15-Jul-2008 15:27	418221
audiology.arff	15-Jul-2008 15:27	45903
autos.arff	15-Jul-2008 15:27	30676
balance-scale.arff	15-Jul-2008 15:27	8714
breast-cancer.arff	15-Jul-2008 15:27	29418
breast-w.arff	15-Jul-2008 15:27	19167
bridges_version1.arff	15-Jul-2008 15:27	11911
bridges_version2.arff	15-Jul-2008 15:27	12313
car.arff	15-Jul-2008 15:27	55474
cmc.arff	15-Jul-2008 15:27	33589
colic.ORIG.arff	15-Jul-2008 15:27	42294
colic.arff	15-Jul-2008 15:27	63983
credit-a.arff	15-Jul-2008 15:27	34315
credit-g.arff	15-Jul-2008 15:27	162249
cylinder-bands.arff	15-Jul-2008 15:27	113893
dermatology.arff	15-Jul-2008 15:27	32417
diabetes.arff	15-Jul-2008 15:27	37419
ecoli.arff	15-Jul-2008 15:27	15714
flags.arff	15-Jul-2008 15:27	21255

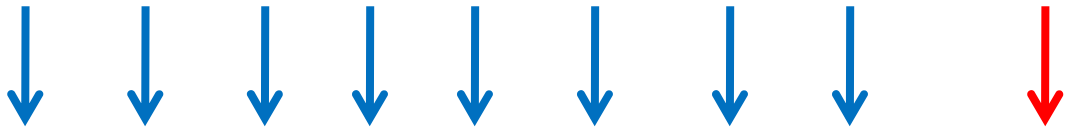
Correlação de Pearson

□ Dataset (atributos):

- ❖ 1. Número de gravidez (preg);
- ❖ 2. Concentração de glicose no plasma (plas);
- ❖ 3. Pressão sanguínea (pres);
- ❖ 4. Tamanho das dobras cutâneas no tríceps (skin);
- ❖ 5. Insulina sérica (insu);
- ❖ 6. Índice de massa corporal - IMC (mass);
- ❖ 7. Casos de Diabets na família (pedi);
- ❖ 8. Idade em anos (age);
- ❖ 9. Variável classe (0 – negativo ou 1 – positivo).

Correlação de Pearson

Dataset:



preg	plas	pres	skin	insu	mass	pedi	age	class
6	148	72	35	0	33,6	0,627	50	tested_positive
1	85	66	29	0	26,6	0,351	31	tested_negative
8	183	64	0	0	23,3	0,672	32	tested_positive
1	89	66	23	94	28,1	0,167	21	tested_negative
0	137	40	35	168	43,1	2,288	33	tested_positive
5	116	74	0	0	25,6	0,201	30	tested_negative
3	78	50	32	88	31	0,248	26	tested_positive
10	115	0	0	0	35,3	0,134	29	tested_negative
2	197	70	45	543	30,5	0,158	53	tested_positive
8	125	96	0	0	0	0,232	54	tested_positive
4	110	92	0	0	37,6	0,191	30	tested_negative
10	168	74	0	0	38	0,537	34	tested_positive
10	139	80	0	0	27,1	1,441	57	tested_negative
1	189	60	23	846	30,1	0,398	59	tested_positive
5	166	72	19	175	25,8	0,587	51	tested_positive
7	100	0	0	0	30	0,484	32	tested_positive
0	118	84	47	230	45,8	0,551	31	tested_positive
7	107	74	0	0	29,6	0,254	31	tested_positive
1	103	30	38	83	43,3	0,183	33	tested_negative
1	115	70	30	96	34,6	0,529	32	tested_positive
3	126	88	41	235	39,3	0,704	27	tested_negative
8	99	84	0	0	35,4	0,388	50	tested_negative
7	196	90	0	0	39,8	0,451	41	tested_positive
9	119	80	35	0	29	0,263	29	tested_positive
11	143	94	33	146	36,6	0,254	51	tested_positive
10	125	70	26	115	31,1	0,205	41	tested_positive

Correlação de Pearson

□ Exemplo:

Correl[1-2] 0,12946	Correl[1-3] 0,14128	Correl[1-4] -0,08167	Correl[1-5] -0,07353	Correl[1-6] 0,01768	Correl[1-7] -0,03352	Correl[1-8] 0,54434
Correl[2-3] 0,15259	Correl[2-4] 0,05733	Correl[2-5] 0,33136	Correl[2-6] 0,22107	Correl[2-7] 0,13734	Correl[2-8] 0,26351	
Correl[3-4] 0,20737	Correl[3-5] 0,08893	Correl[3-6] 0,28181	Correl[3-7] 0,04126	Correl[3-8] 0,23953		
Correl[4-5] 0,43678	Correl[4-6] 0,39257	Correl[4-7] 0,18393	Correl[4-8] -0,11397			
Correl[5-6] 0,19786	Correl[5-7] 0,18507	Correl[5-8] -0,04216				
Correl[6-7] 0,14065	Correl[6-8] 0,03624					
Correl[7-8] 0,03356						

Correlação de Pearson

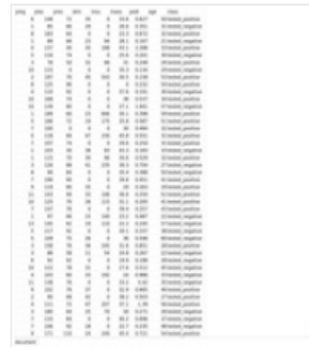
❏ Baixando os arquivos:

MPES0009_Material

Classificado por nome

A screenshot of a CSV file named 'Arrhythmia.csv'. It contains a large number of rows, each with 28 columns of numerical data. The first few rows show values like 0.0, 0.0, 0.0, etc., followed by various decimal values.

Arrhythmia.csv

A screenshot of a CSV file named 'diabetes.csv'. It contains a large number of rows, each with 10 columns of numerical data. The first few rows show values like 0.0, 0.0, 0.0, etc., followed by various decimal values.

diabetes.csv



Pearson.py

A screenshot of a CSV file named 'Pessoa.csv'. It contains a large number of rows, each with 10 columns of numerical data. The first few rows show values like 0.0, 0.0, 0.0, etc., followed by various decimal values.

Pessoa.csv

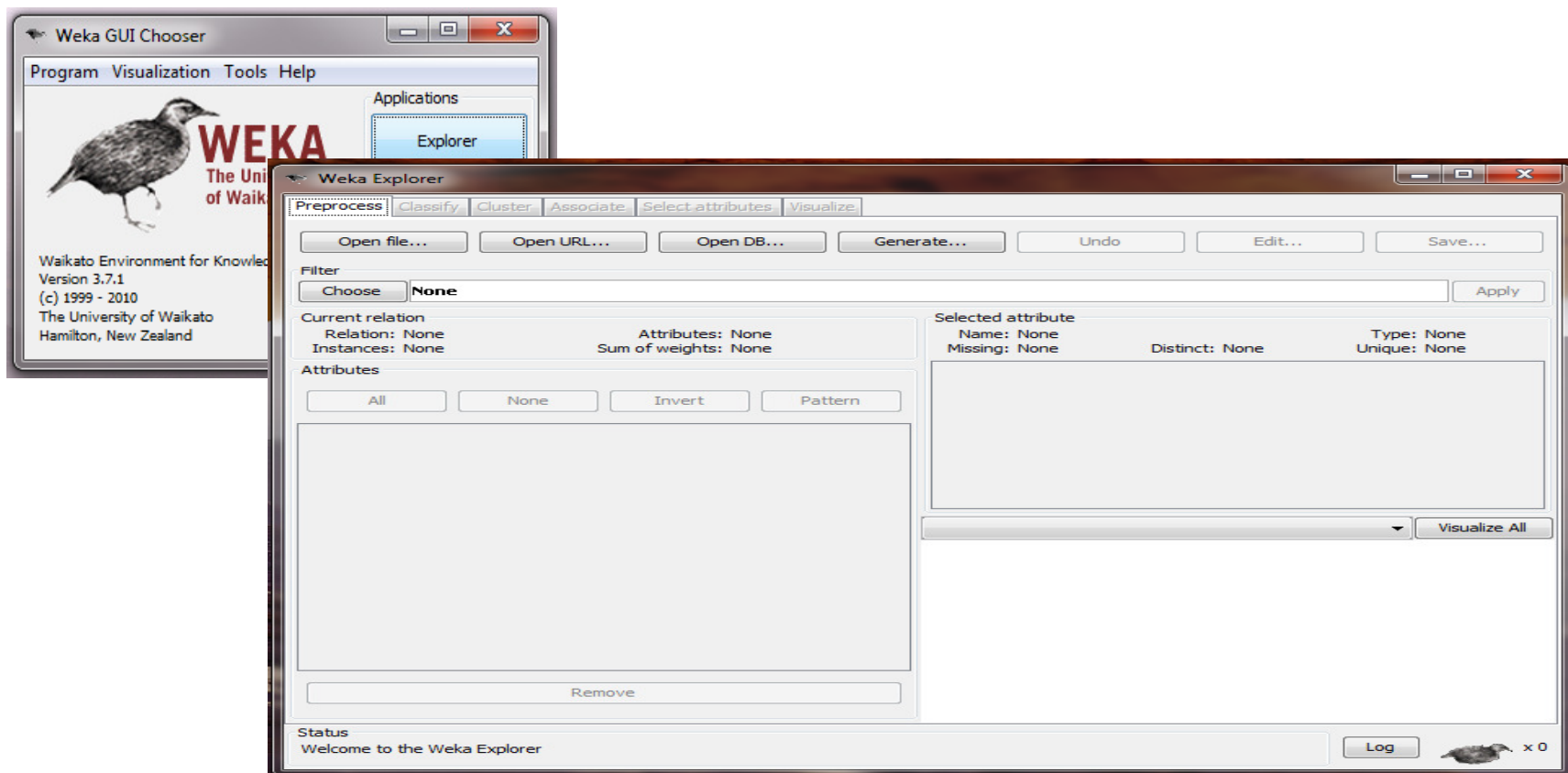


weka-src.jar

❖ <https://www.dropbox.com/sh/b03djhggsziugl7h/AAARsUXaqAA7TYSsdk1-ZALKa?dl=0>

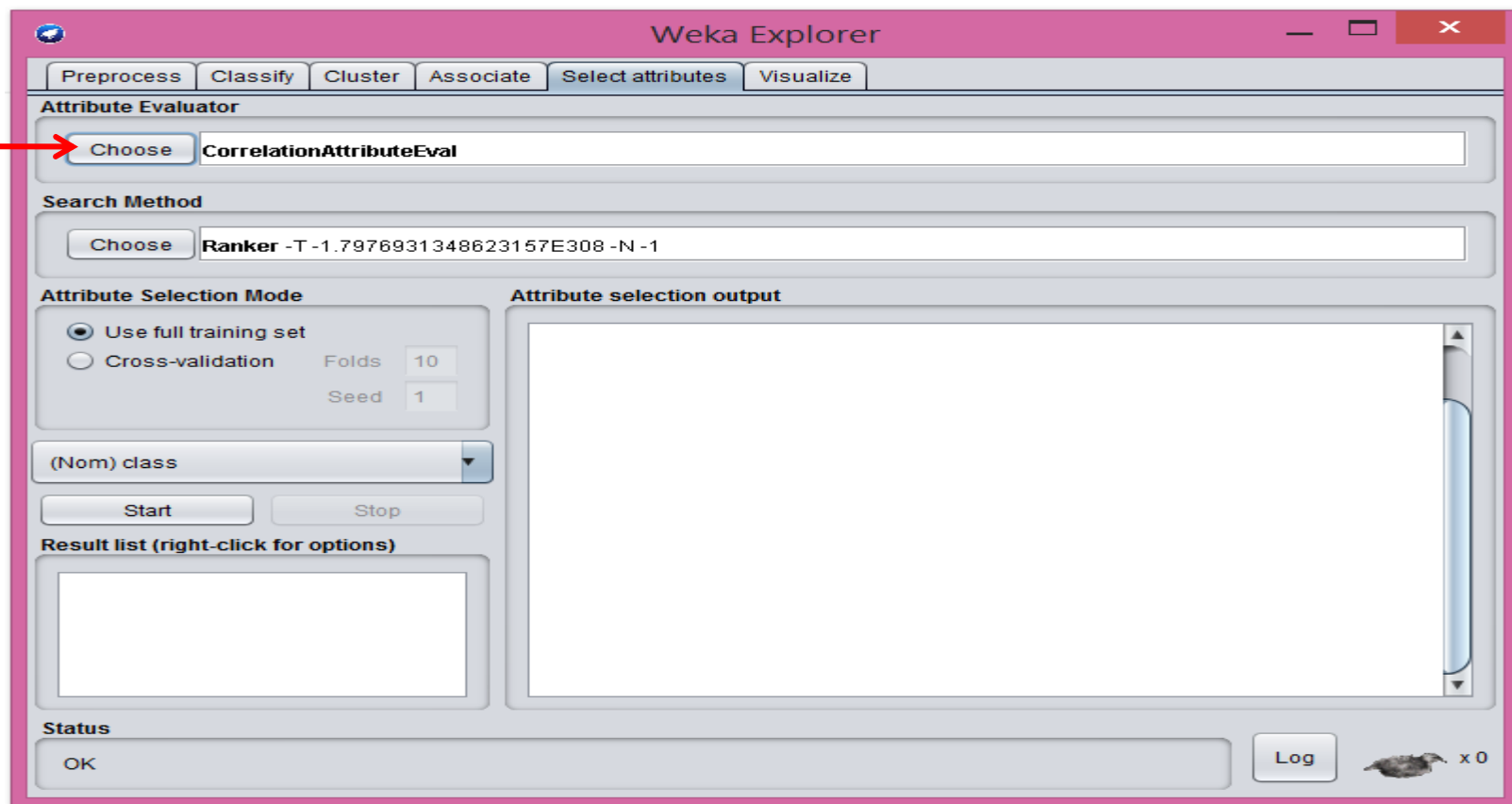
Correlação de Pearson

❑ Executando Weka:



Correlação de Pearson

❑ Seleção de atributos:



Correlação de Pearson

❑ Seleção de atributos:

The screenshot displays the WEKA Attribute Selection interface. On the left, the 'Attribute Selection Mode' section has 'Use full training set' selected. Below it, a dropdown menu shows '(Nom) class'. The 'Start' and 'Stop' buttons are visible. The 'Result list (right-click for options)' section shows a single entry: '11:21:49 - Ranker + CorrelationAttribut'. On the right, the 'Attribute selection output' window shows the results of the 'Correlation Ranking Filter'. It lists 'Ranked attributes' with their correlation coefficients and names. A red arrow points to the 'Selected attributes' line at the bottom of the output window.

Attribute Selection Mode

- ☒ Use full training set
- ☐ Cross-validation

Folds: 10
Seed: 1

(Nom) class

Start Stop

Result list (right-click for options)

11:21:49 - Ranker + CorrelationAttribut

Attribute selection output

Correlation Ranking Filter

Ranked attributes:

0.4666	2 plas
0.2927	6 mass
0.2384	8 age
0.2219	1 preg
0.1738	7 pedi
0.1305	5 insu
0.0748	4 skin
0.0651	3 pres

Selected attributes: 2,6,8,1,7,5,4,3 : 8

Correlação de Pearson

- ❑ Utilizando Panda (biblioteca para Análise de Dados em Python):

```
pandas.DataFrame.corr
```

```
DataFrame.corr(method='pearson', min_periods=1)
```

[source]

Compute pairwise correlation of columns, excluding NA/null values

Parameters:

method : {'pearson', 'kendall', 'spearman'}

- pearson : standard correlation coefficient
- kendall : Kendall Tau correlation coefficient
- spearman : Spearman rank correlation

min_periods : int, optional

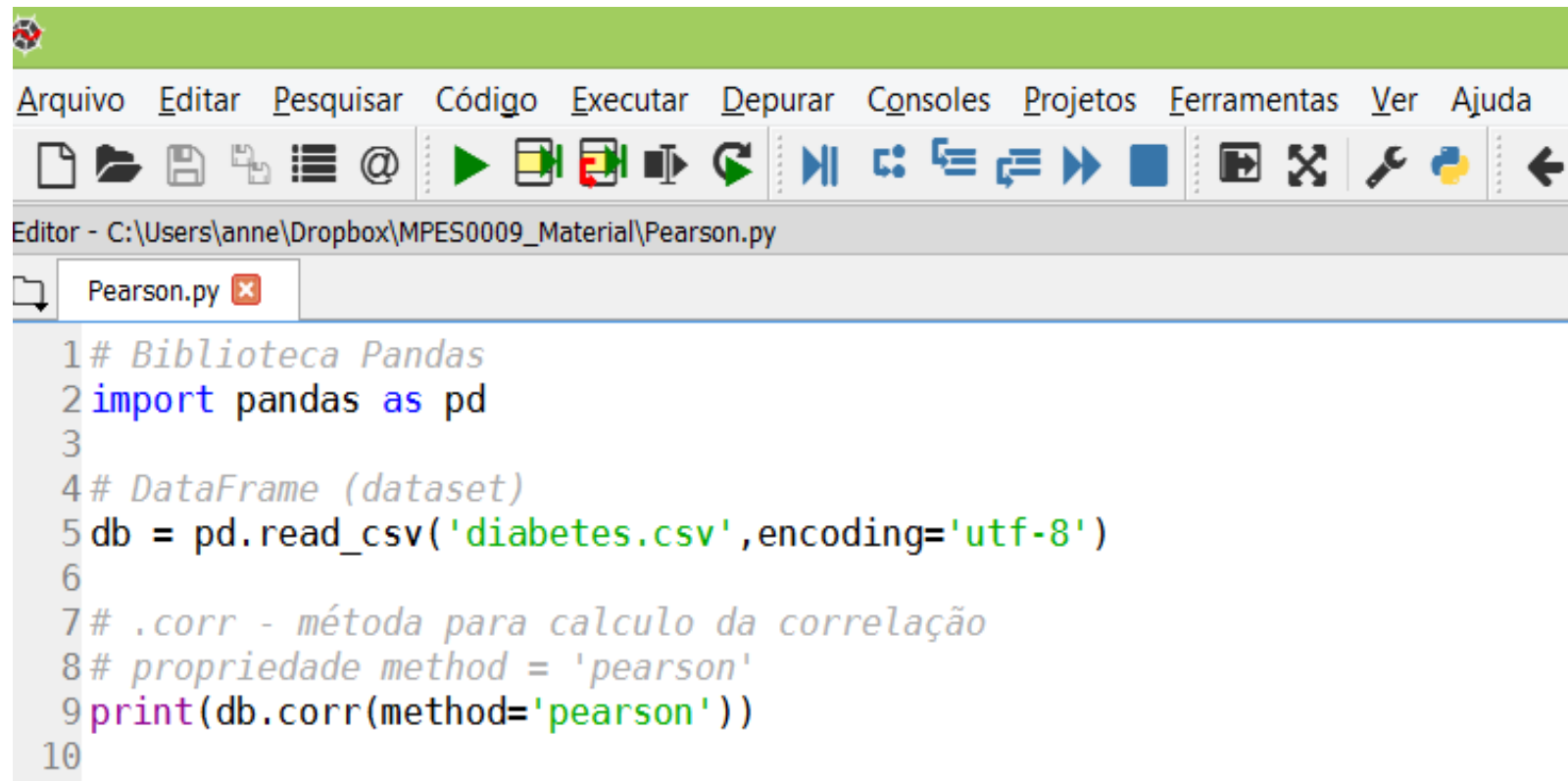
Minimum number of observations required per pair of columns to have a valid result.
Currently only available for pearson and spearman correlation

Returns:

y : DataFrame

Correlação de Pearson

❑ Código em Python:

A screenshot of a Python IDE window. The title bar is green. The menu bar includes 'Arquivo', 'Editar', 'Pesquisar', 'Código', 'Executar', 'Depurar', 'Consoles', 'Projetos', 'Ferramentas', 'Ver', and 'Ajuda'. The toolbar contains various icons for file operations, execution, and debugging. The editor window shows a file named 'Pearson.py' with the following Python code:

```
1 # Biblioteca Pandas
2 import pandas as pd
3
4 # DataFrame (dataset)
5 db = pd.read_csv('diabetes.csv', encoding='utf-8')
6
7 # .corr - método para calculo da correlação
8 # propriedade method = 'pearson'
9 print(db.corr(method='pearson'))
10
```

Correlação de Pearson

❑ Correlação calculada:

```
Console IPython
Console 1/A ✖
In [1]:
In [1]: runfile('C:/Users/anne/Dropbox/MPES0009_Material/Pearson.py', wdir='C:/
Users/anne/Dropbox/MPES0009_Material')
      preg      plas      pres      skin      insu      mass      pedi  \
preg  1.000000  0.129459  0.141282 -0.081672 -0.073535  0.017683 -0.033523
plas  0.129459  1.000000  0.152590  0.057328  0.331357  0.221071  0.137337
pres  0.141282  0.152590  1.000000  0.207371  0.088933  0.281805  0.041265
skin -0.081672  0.057328  0.207371  1.000000  0.436783  0.392573  0.183928
insu -0.073535  0.331357  0.088933  0.436783  1.000000  0.197859  0.185071
mass  0.017683  0.221071  0.281805  0.392573  0.197859  1.000000  0.140647
pedi -0.033523  0.137337  0.041265  0.183928  0.185071  0.140647  1.000000
age  0.544341  0.263514  0.239528 -0.113970 -0.042163  0.036242  0.033561

      age
preg  0.544341
plas  0.263514
pres  0.239528
skin -0.113970
insu -0.042163
mass  0.036242
pedi  0.033561
age   1.000000
```

Python Data Analysis Library

□ Pandas (referências):

- ❖ <http://minerandodados.com.br/index.php/2017/09/26/python-para-analise-de-dados/>
- ❖ <https://pandas.pydata.org/>

Dúvidas ...

