



PL13 – RapidMiner: k-Means Clustering

AEC - Mestrado em Engenharia Biomédica

<https://hpeixoto.me/class/aec>

Plano de Aula – PL13



Clustering

Plano de Aula – PL13

O algoritmo K-means é um método de agrupamento (clustering) amplamente utilizado na análise de dados. O intuito é dividir um conjunto de dados em **K** grupos distintos, sendo **K** um número definido pelo utilizador.

Plano de Aula – PL13

Inicialização: Inicialmente, escolhem-se K pontos aleatórios do conjunto de dados para serem os centros iniciais dos clusters (centroides).

Atribuição a clusters: Em seguida, cada ponto do conjunto de dados é atribuído ao cluster cujo centroide está mais próximo. Esta proximidade é geralmente medida usando a distância euclidiana.

Atualização dos centroides: Após todos os pontos serem atribuídos a um cluster, recalcula-se o centro de cada cluster. Isso é feito encontrando a média de todos os pontos atribuídos a esse cluster.

Repetição: Os passos 2 e 3 são repetidos iterativamente até que os centroides não se movam significativamente entre as iterações, indicando que os clusters estão relativamente estáveis e o algoritmo atingiu a convergência.

Plano de Aula – PL13

O K-means é particularmente eficaz em grandes conjuntos de dados e é usado em uma variedade de aplicações, como segmentação de mercado, agrupamento de documentos, compressão de imagem e muito mais.

No entanto tem algumas limitações:

- Sensibilidade à escolha inicial dos centroides;
- Dificuldade em lidar com clusters de formas não esféricas ou tamanhos variados.



Clustering: Exemplo

Clustering

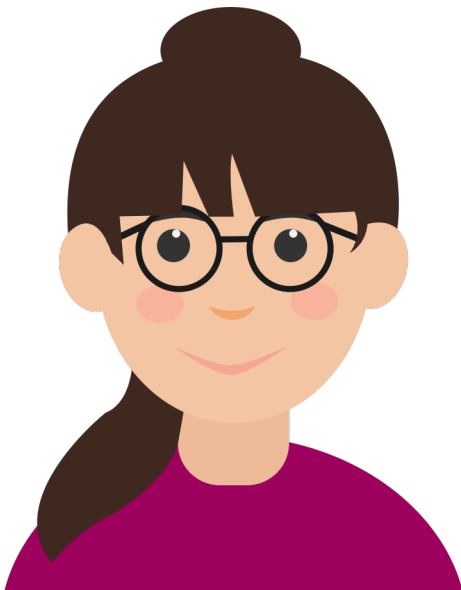
Business Understanding

A Sónia é diretora de uma grande seguradora de saúde.

Recentemente, tem estado a ler revistas médicas e outros artigos e encontrou uma forte crença influência do peso, do sexo e do colesterol no desenvolvimento de doenças coronárias.

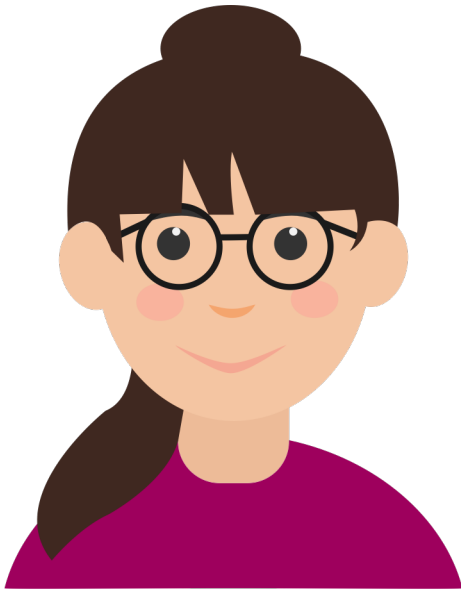
Começa a pensar em ideias para a sua empresa oferecer programas de controlo de peso e colesterol aos indivíduos que recebem um seguro de saúde através da sua entidade patronal.

Ao considerar onde os seus esforços podem ser mais eficazes, dá por si a pensar se existem grupos naturais de indivíduos com maior risco de peso e colesterol elevados e, caso existam, onde se situam as linhas divisórias naturais entre os grupos.



Clustering

Business Understanding



O objetivo da Sónia é identificar e tentar chegar aos indivíduos segurados pela sua entidade patronal que correm um risco elevado de doença coronária devido ao seu peso e/ou colesterol elevado.

Ela compreende que as pessoas com baixo risco, ou seja, com baixo peso e colesterol, dificilmente participarão nos programas que ela irá oferecer.

Também compreende que é provável que existam (i) segurados com peso elevado e colesterol baixo, (ii) segurados com peso elevado e (iii) colesterol elevado e (iv) segurados com peso baixo e colesterol elevado.

Reconhece ainda que é provável que haja muitas pessoas algures no meio.

Clustering

Data Understanding

Utilizando a base de dados de sinistros da companhia de seguros, Sónia extrai três atributos para 547 indivíduos selecionados aleatoriamente.

Os três atributos são o peso (**weight**) do segurado em libras, tal como registado no exame médico mais recente da pessoa, o seu último nível de colesterol (**cholesterol**) determinado por análises ao sangue no laboratório do médico e o seu sexo (**gender**).

O atributo sexo utiliza 0 para indicar Mulher e 1 para indicar Homem.

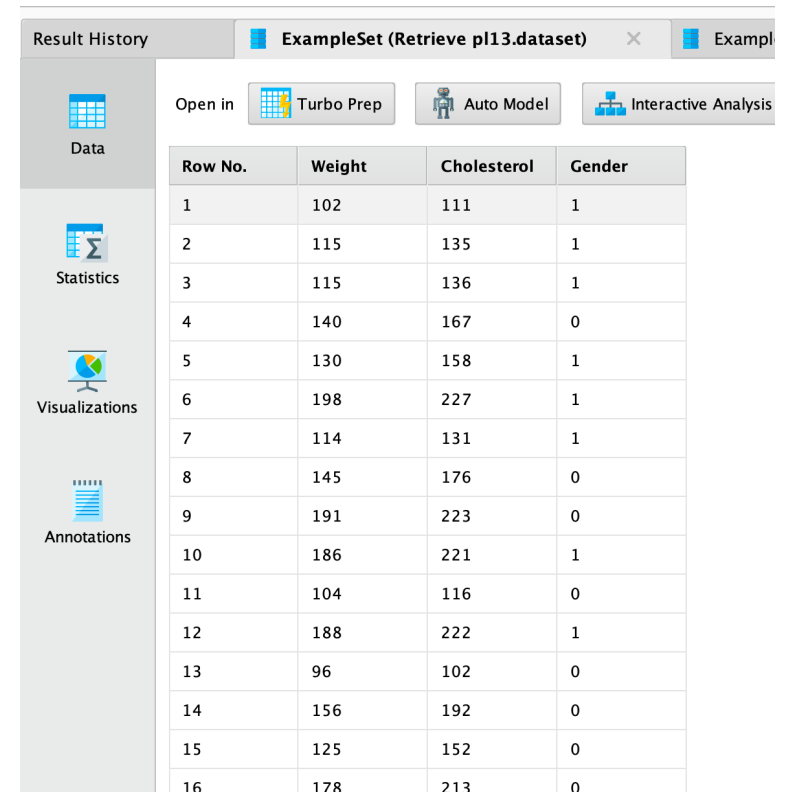
Devemos lembrar que as médias são particularmente suscetíveis à influência indevida de valores extremos, pelo que é muito importante estar atento aos dados inconsistentes quando se utiliza a metodologia de extração de dados de agrupamento k-Means.

Clustering




Data Preparation

Download do data set: pl13-dataset.csv

1. Importe o data set para o repositório RapidMiner
2. Avalie os dados importados



Result History | ExampleSet (Retrieve pl13.dataset) | Example

Open in  Turbo Prep  Auto Model  Interactive Analysis

Row No.	Weight	Cholesterol	Gender
1	102	111	1
2	115	135	1
3	115	136	1
4	140	167	0
5	130	158	1
6	198	227	1
7	114	131	1
8	145	176	0
9	191	223	0
10	186	221	1
11	104	116	0
12	188	222	1
13	96	102	0
14	156	192	0
15	125	152	0
16	178	213	0

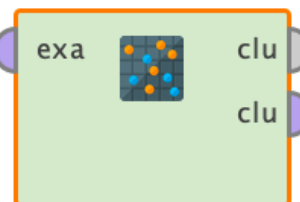
Clustering

Modelling

Retrieve pl13.dataset



Clustering



res

res

Clustering

Modelling

Como discutido anteriormente, Sónia já reconheceu que existem provavelmente vários tipos diferentes de grupos a serem considerados.

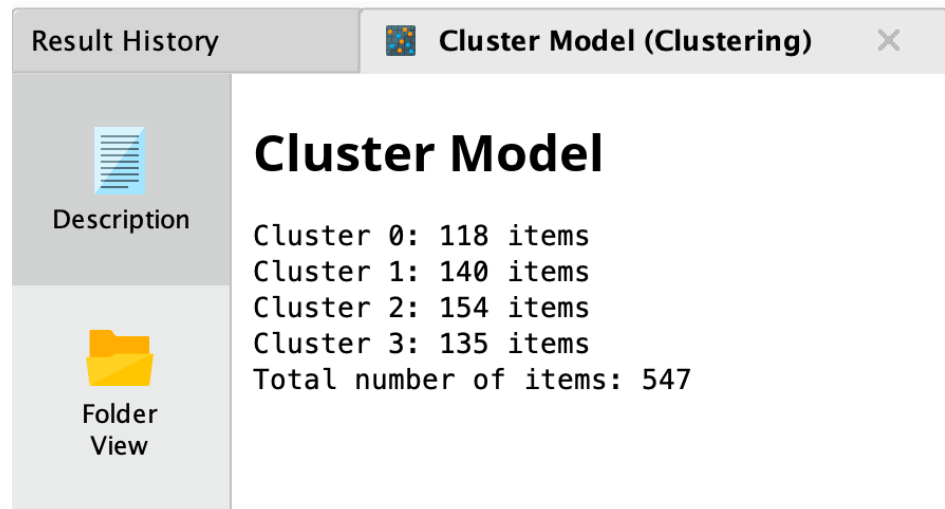
A simples divisão do conjunto de dados em dois grupos provavelmente não dará a Sónia o nível de detalhe que ela procura.

Como a Sónia sentiu que havia provavelmente pelo menos 4 grupos potencialmente diferentes, vamos alterar o valor de k para quatro.

Clustering

Modelling

Corra o modelo e vamos observar os resultados:



Result History Cluster Model (Clustering) X

Cluster Model

Cluster 0: 118 items
Cluster 1: 140 items
Cluster 2: 154 items
Cluster 3: 135 items
Total number of items: 547

Description

Folder View

Clustering

Evaluation

O principal objetivo da Sónia no cenário hipotético apresentado era tentar encontrar divisões naturais entre diferentes tipos de grupos de risco de doenças cardíacas.

Utilizando o operador k-Means no RapidMiner, identificamos quatro clusters para a Sónia, e agora podemos avaliar sua utilidade em responder à questão dela.

Vamos examinar a Tabela de Centróides. Essa visualização dos resultados, apresenta as médias de cada atributo em cada um dos quatro clusters criados.

Evaluation

Attribute	cluster_0	cluster_1	cluster_2	cluster_3
Weight	152.093	106.850	184.318	127.726
Cholesterol	185.907	119.536	218.916	154.385
Gender	0.441	0.543	0.591	0.459

Nesta visualização, observamos que o cluster_2 tem a maior média de peso e colesterol.

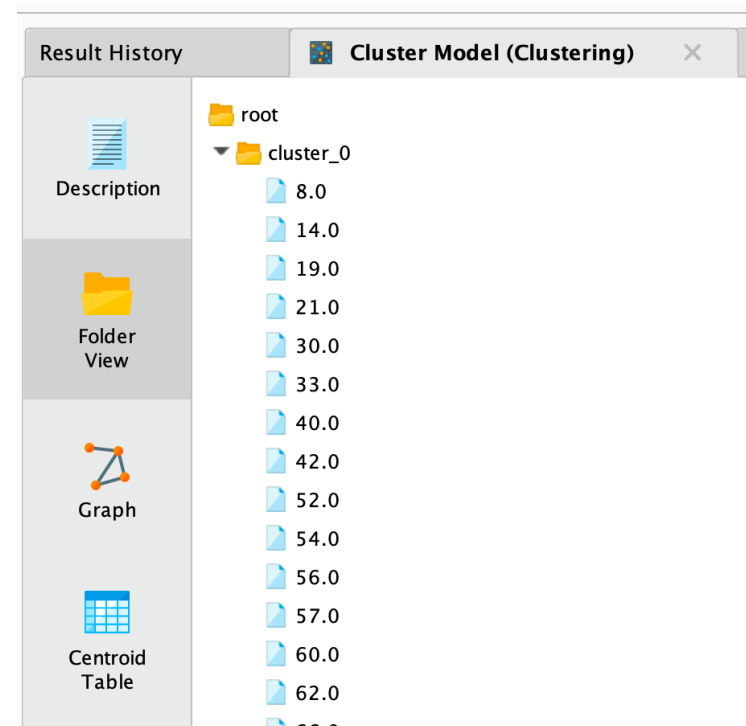
Com 0 representando mulheres e 1 representando homens, uma média de 0,591 indica que há mais homens do que mulheres representados neste cluster.

Sabendo que o colesterol alto e o peso são dois indicadores-chave de risco de doenças cardíacas que os segurados podem controlar, a Sónia provavelmente gostaria de começar com os membros do cluster_2 ao promover seus novos programas. Em seguida, ela poderia expandir sua programação para incluir as pessoas nos clusters 0 e 3, que apresentam médias incrementalmente menores para esses dois atributos de risco.

Clustering

Evaluation

Como saber os membros desse cluster?



Clustering

Evaluation

Visualizar os clusters criados:

Plot

Plot 1

Plot type
Scatter / Bubble

X-Axis column
Weight

Value column
Cholesterol

Color
cluster

Size
-

Jitter

Regression interpolation
None

Plot style >>

[Add new plot](#)

General

X-Axis

Y-Axis

Title

Legend

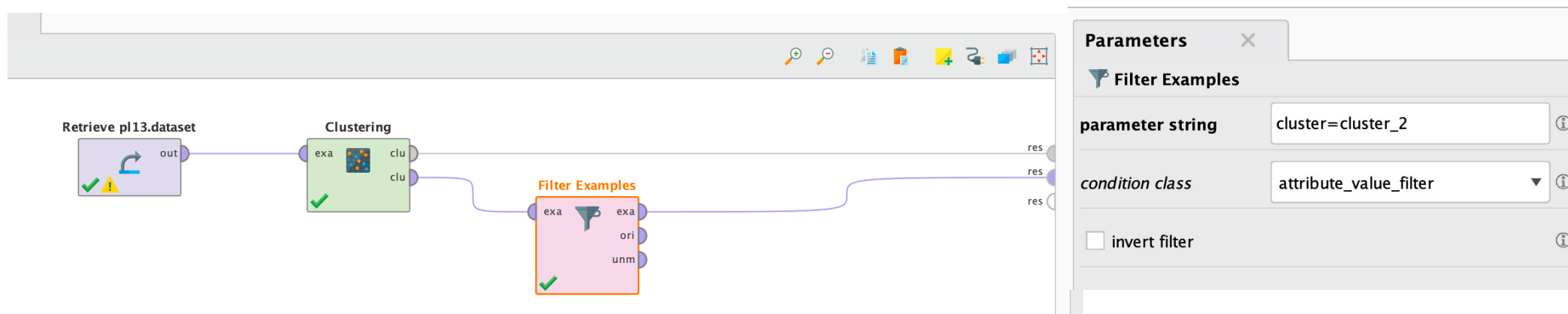
Tooltip



Clustering

Deployment

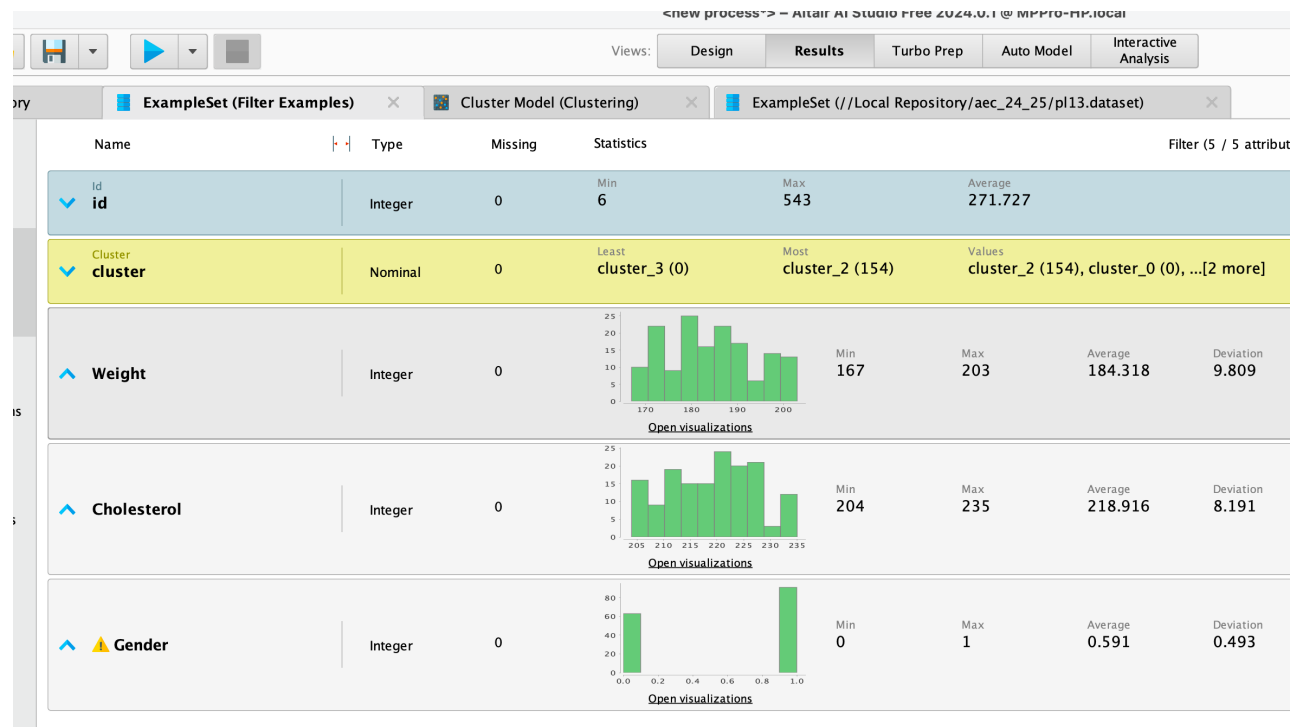
Para apoiar a Sónia na seleção pode ser implementado um filtro para apresentar apenas os membros do cluster_2.



Clustering

Deployment

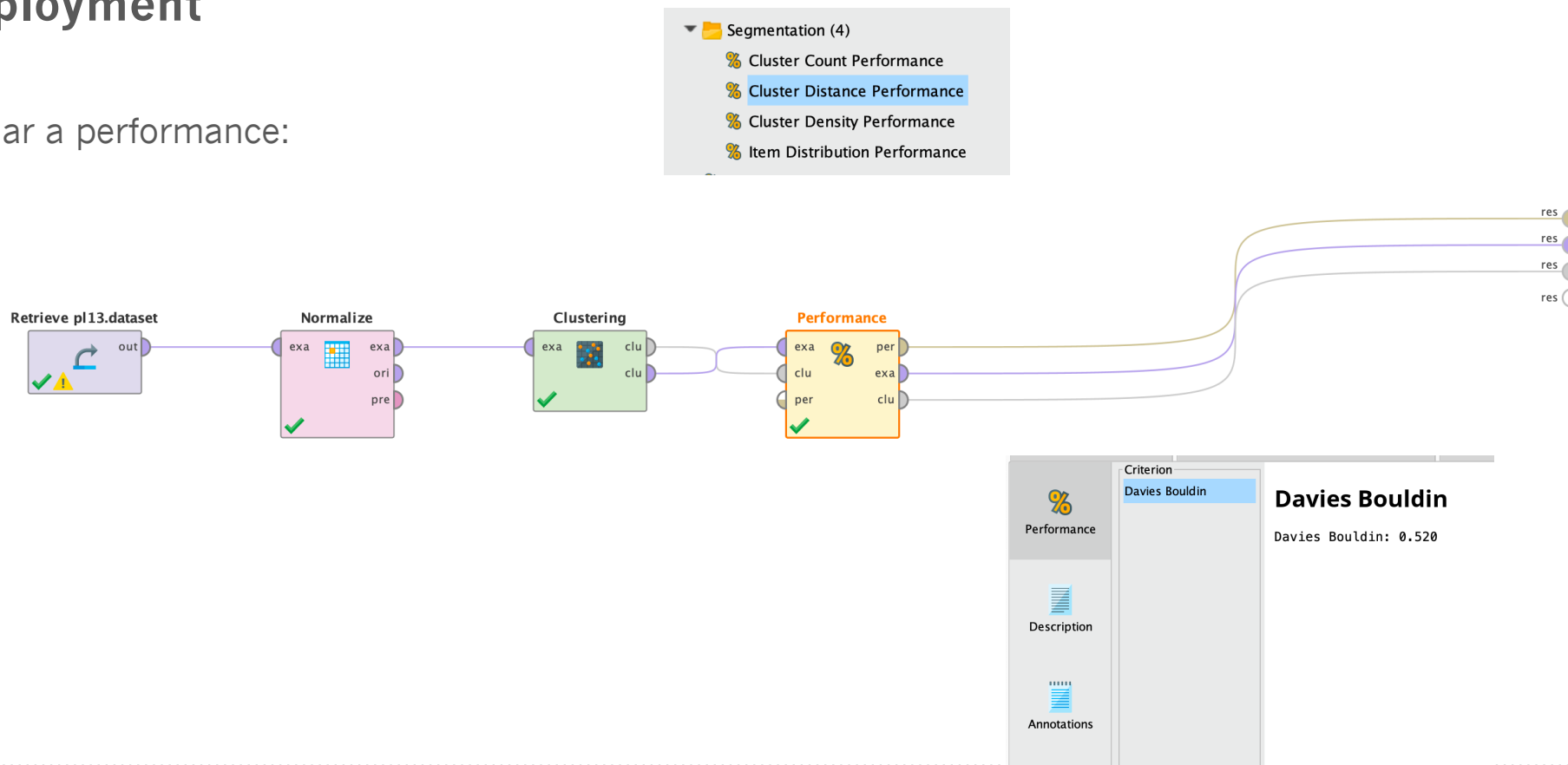
Avaliar as estatísticas do cluster:



Clustering

Deployment

Avaliar a performance:



Ficha de Exercícios 09



PL13 – RapidMiner: k-Means Clustering

AEC - Mestrado em Engenharia Biomédica

<https://hpeixoto.me/class/aec>