

Universidade do Minho
Escola de Engenharia

D escoberta do C onhecimento

José Machado
Cristiana Neto



K-MEANS CLUSTERING COM O RAPIDMINER

CONTEXTO E PRESPECTIVA



A Sónia é diretora de programas de um grande provedor de seguros de saúde.

Recentemente, ela tem lido revistas médicas e outros artigos, e encontrou um forte ênfase na influência do género, peso e colesterol no desenvolvimento de doença cardíaca coronária.

Ela começa a levantar ideias para que a sua empresa ofereça programas de controlo de peso e de colesterol para indivíduos que recebem seguro de saúde através da sua empresa.

Enquanto ela considera onde os seus esforços podem ser mais eficazes, ela começa a perguntar-se se existem grupos naturais de indivíduos que estão em maior de peso alto e colesterol alto e, se houver grupos, onde ocorrem as naturais linhas divisórias entre os grupos.

O Data Mining pode ajudá-la a compreender estes grupos.

BUSINESS UNDERSTANDING



O objetivo da Sónia é identificar e tentar entrar em contato com pessoas seguradas pelo seu empregador com alto risco de doença cardíaca coronária devido ao seu peso e/ou colesterol alto.

Ela entende que é improvável que as pessoas com baixo risco, ou seja, aquelas com peso baixo e colesterol baixo, participem dos programas que ela oferecerá.

Ela entende também que provavelmente existem segurados com alto peso e baixo colesterol, outros com peso alto e colesterol alto e outros com peso baixo e colesterol alto. Ela reconhece ainda que é provável que haja muitas pessoas alíngues entre estes tipos.

Para atingir o seu objetivo, ela precisa pesquisar entre os milhares de segurados para encontrar grupos de pessoas com características semelhantes e criar programas que sejam relevantes e atraentes para as pessoas desses diferentes grupos.

DATA UNDERSTANDING



Usando a base de dados de requerimentos da companhia de seguros, a Sónia extraiu três atributos para 547 indivíduos selecionados aleatoriamente.

Os três atributos são o **peso** do segurado em libras, conforme registado no exame médico mais recente da pessoa, o seu último **nível de colesterol** determinado pelo exame de sangue e seu **sexo**. Como é típico em muitos *datasets*, o atributo *gender* usa 0 para indicar Feminino e 1 para indicar Masculino.

Usaremos estes dados para criar um modelo de cluster para ajudar Sónia a entender como os clientes de sua empresa parecem agrupar-se com base em seus pesos, géneros e níveis de colesterol.

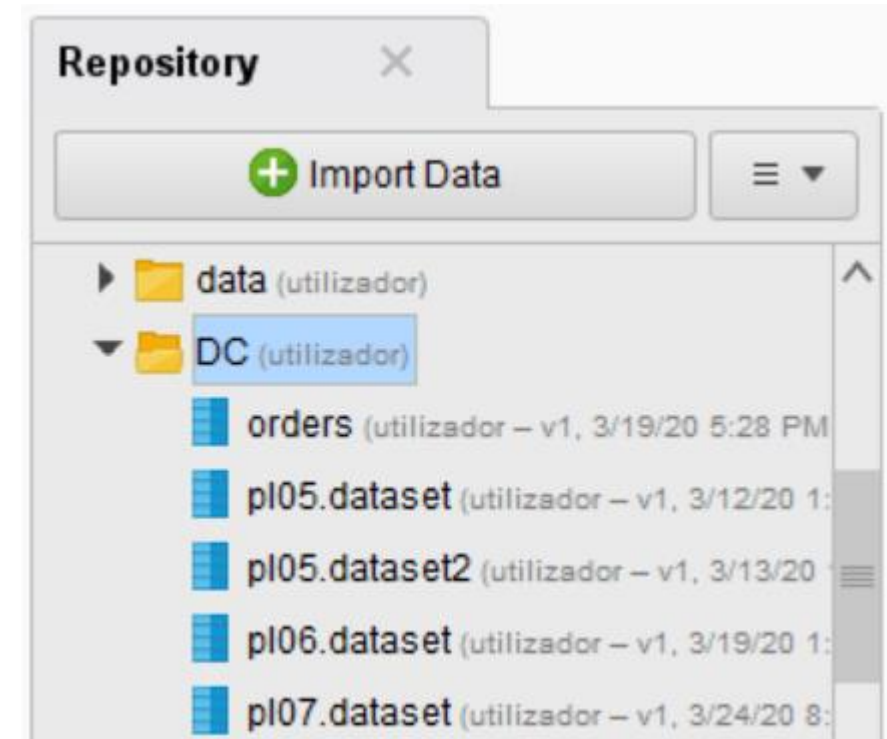
Importa lembrar que, ao fazer isso, as médias são particularmente suscetíveis à influência indevida de valores extremos, portanto, é muito importante identificar dados inconsistentes ao usar a metodologia de *data mining* de **k-Means clustering**.

DATA PREPARATION



Download do dataset: pl07.dataset.csv

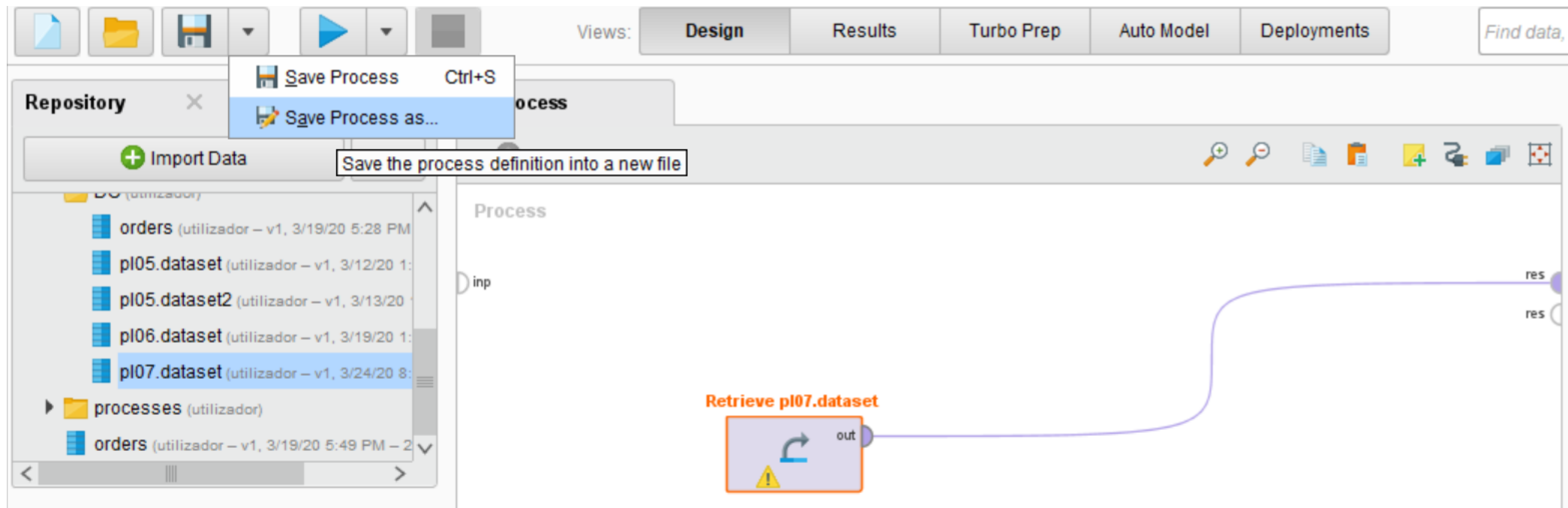
1. Importar o CSV para o repositório rapidminer (Import Data -> My Computer)
2. Verificar a *view* dos resultados e inspecionar os dados CSV importados (Data, Statistics)



DATA PREPARATION



3. Arraste o dataset **pl07.dataset** para uma nova janela de processo no RapidMiner
4. Execute o modelo para inspecionar os dados e salve o processo como **pl07_processo**.



DATA PREPARATION



5. Seleccione a *view* “Results” e escolha a opção “Statistics”. Note que:

- Não existe nenhum *missing value* para nenhum dos 12 atributos.
- Nenhum dos valores parece ser inconsistente (lembre-se dos comentários da aula anterior sobre o uso dos desvios padrão para encontrar discrepâncias estatísticas).

MODELING



O 'k' em *k-means* significa agrupar significa um número de grupos ou *clusters*. O objetivo dessa metodologia de *data mining* é examinar os valores dos atributos individuais de cada observação e compará-los com as médias de potenciais grupos de outras observações, a fim de encontrar grupos naturais que são semelhantes entre si.

O algoritmo *k-means* faz isso amostrando um conjunto de observações no *data set*, calculando as médias de cada atributo para as observações nessa amostra e comparando os outros atributos do dataset com as médias dessa amostra.

O sistema faz isso repetidamente para encontrar as melhores correspondências e depois formular grupos de observações que se tornam *clusters*. À medida que as médias calculadas se tornam cada vez mais semelhantes, clusters são formados e cada observação cujos valores dos atributos são mais parecidos com as médias de um cluster, esta torna-se membro desse cluster.

MODELING



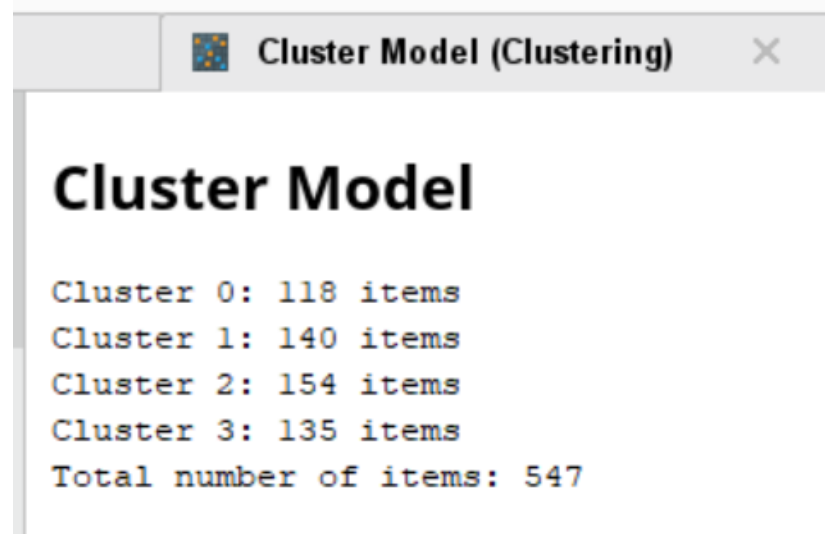
1. Procure e arraste operador *k-means* para o processo. Relativamente ao valor do *k* (nos parâmetros do lado direito), como é provável que existam pelo menos quatro grupos potencialmente diferentes, vamos alterar o valor de *k* para 4. Clique em Run.

The screenshot displays the Orange3 data mining software interface. The top toolbar includes icons for file operations and a 'Views' dropdown set to 'Design'. The 'Repository' panel on the left shows a list of datasets, with 'pl07.dataset' selected. The 'Operators' panel below it shows the 'k-means' operator under the 'Modeling' category. The central 'Process' canvas shows a workflow: 'Retrieve pl07.dataset' (purple box) connected to 'Clustering' (orange box). The 'Clustering' operator has two output ports labeled 'clu'. The 'Parameters' panel on the right is open for the 'Clustering (k-Means)' operator. It shows several settings: 'add cluster attribute' is checked, 'add as label' is unchecked, 'remove unlabeled' is unchecked, 'k' is set to 4 (highlighted with an orange arrow), 'max runs' is 10, 'determine good start values' is checked, 'measure types' is 'BregmanDivergences', 'divergence' is 'SquaredEuclideanDist...', and 'max optimization steps' is 100.

MODELING



2. Quando o modelo é executado, encontramos um relatório inicial do número de itens que ficaram em cada um de nossos quatro clusters. Neste modelo em particular, nossos clusters são razoavelmente bem equilibrados.



```
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
```

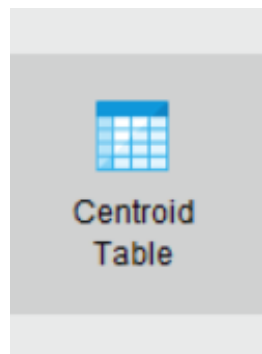
Neste ponto poderíamos voltar atrás e ajustar o nosso número de clusters, o nosso valor de 'max-runs' ou até experimentar outros parâmetros apresentados pelo operador k-Means.

EVALUATION



Lembre-se que o principal objetivo da Sónia é tentar encontrar quebras naturais entre diferentes tipos de grupos de risco de doenças cardíacas. Utilizando o operador k-Means no RapidMiner, identificámos quatro grupos, e podemos agora avaliar a sua utilidade.

1. Seleccione a opção “Centroid Table”. Esta janela contém as médias para cada atributo em cada um dos quatro clusters criados.



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

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



- o cluster 2 tem a média de “Weight” e “Cholesterol” mais elevados;
- com 0 representando Feminino e 1 representando Masculino, uma média de 0,591 indica que temos mais homens do que mulheres neste cluster.

EVALUATION



Colesterol e peso elevados são dois indicadores chave do risco de doenças cardíacas sobre os quais os detentores de apólices podem fazer algo.

O que é que isto significa?



A Sónia deve começar com os membros do cluster 2 ao promover os seus novos programas e depois estender aos membros dos clusters 0 e 3, que são, respetivamente, os membros com as médias mais elevadas para estes dois atributos-chave de fatores de risco.

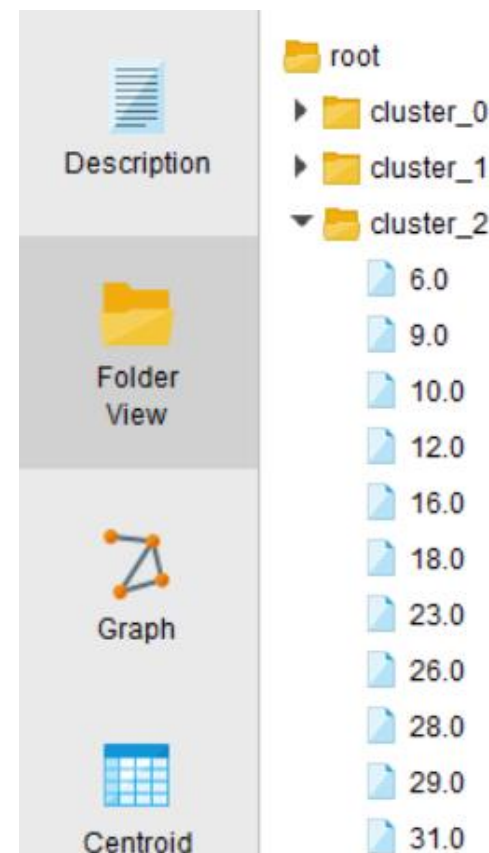
EVALUATION



A Sónia sabe que o cluster 2 é onde vai concentrar os seus primeiros esforços, mas como é que ela sabe quem deve contactar? Quem são os membros deste grupo de maior risco?



2. Seleccione a opção “Folder View” para ter acesso a este tipo de informações.



EVALUATION




3. Clique em cima de uma observação para ver os seus detalhes.


As médias para o cluster 2 eram pouco mais de 184 para o peso e pouco menos de 219 para o colesterol. A pessoa representada na observação 6 é mais pesada e tem um colesterol mais alto do que a média deste grupo de maior risco.



Esta é uma pessoa que a Sónia pode ajudar!

 This dialog shows detailed information about the example with ID 6.

Attribute	Value
Weight	198
Cholesterol	227
Gender	1
id	6
cluster	cluster_2

 Close

EVALUATION





Sabemos pela descrição do Cluster Model que existem 154 membros no *dataset* que se enquadram neste grupo.



Clicar em cada um deles é um processo moroso e pouco eficiente.




Description


Folder View

Cluster Model

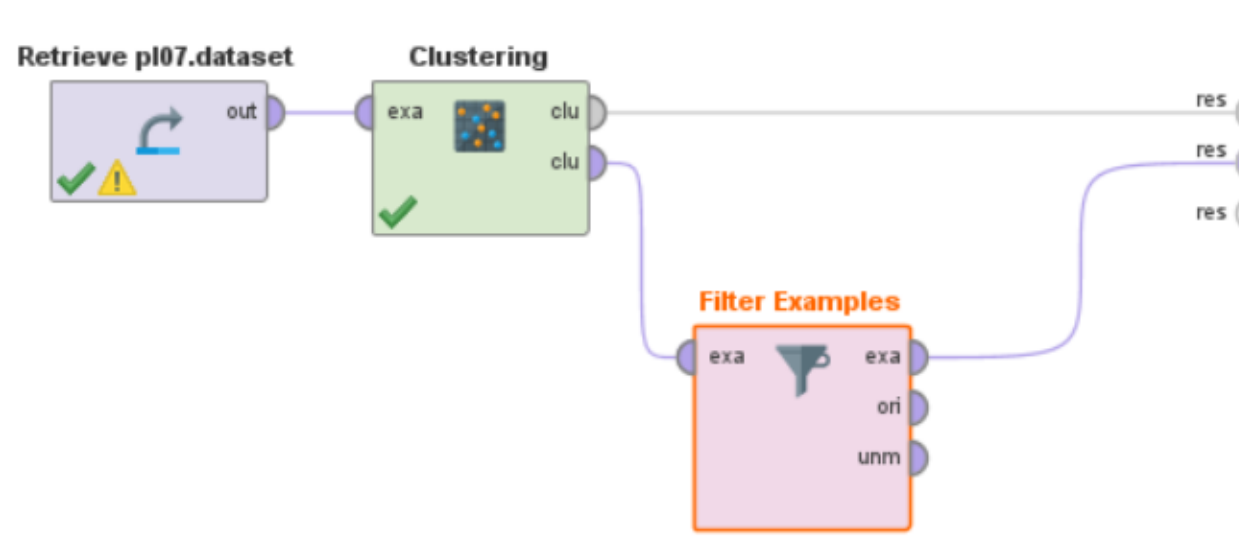
```
Cluster 0: 118 items
Cluster 1: 140 items
Cluster 2: 154 items
Cluster 3: 135 items
Total number of items: 547
```

Podemos ajudar a Sónia a extrair as observações do cluster 2 de forma bastante rápida e fácil.

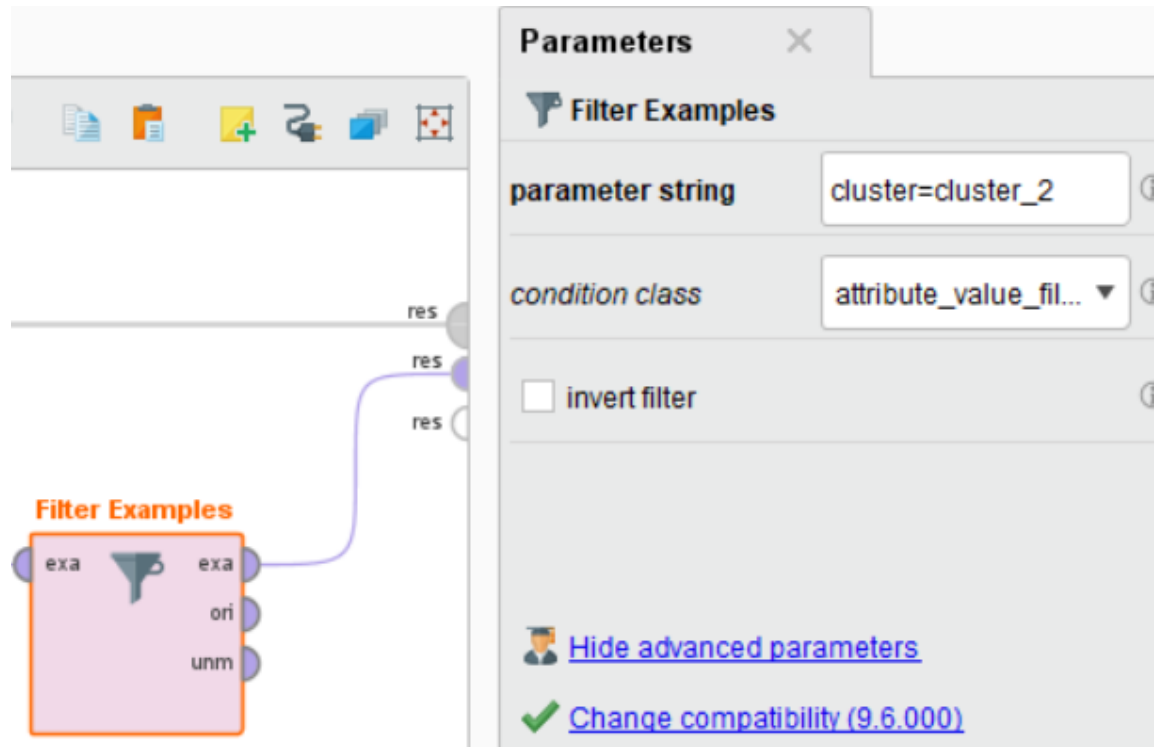
DEPLOYMENT



1. Volte à perspectiva de *Design* no RapidMiner.
2. Procure e arraste o operador “Filter Examples” e conecte-o ao operador k-Means Clustering. Conecte a segunda porta ‘clu’ (cluster) à porta ‘exa’ do operador “Filter Examples”, e conecte a porta ‘exa’ do “Filter Examples” à porta ‘res’ final.



DEPLOYMENT



3. No campo “condition class”, selecione a opção ‘attribute_value_filter’, e para o campo “parameter string”, digite o seguinte: cluster=cluster_2

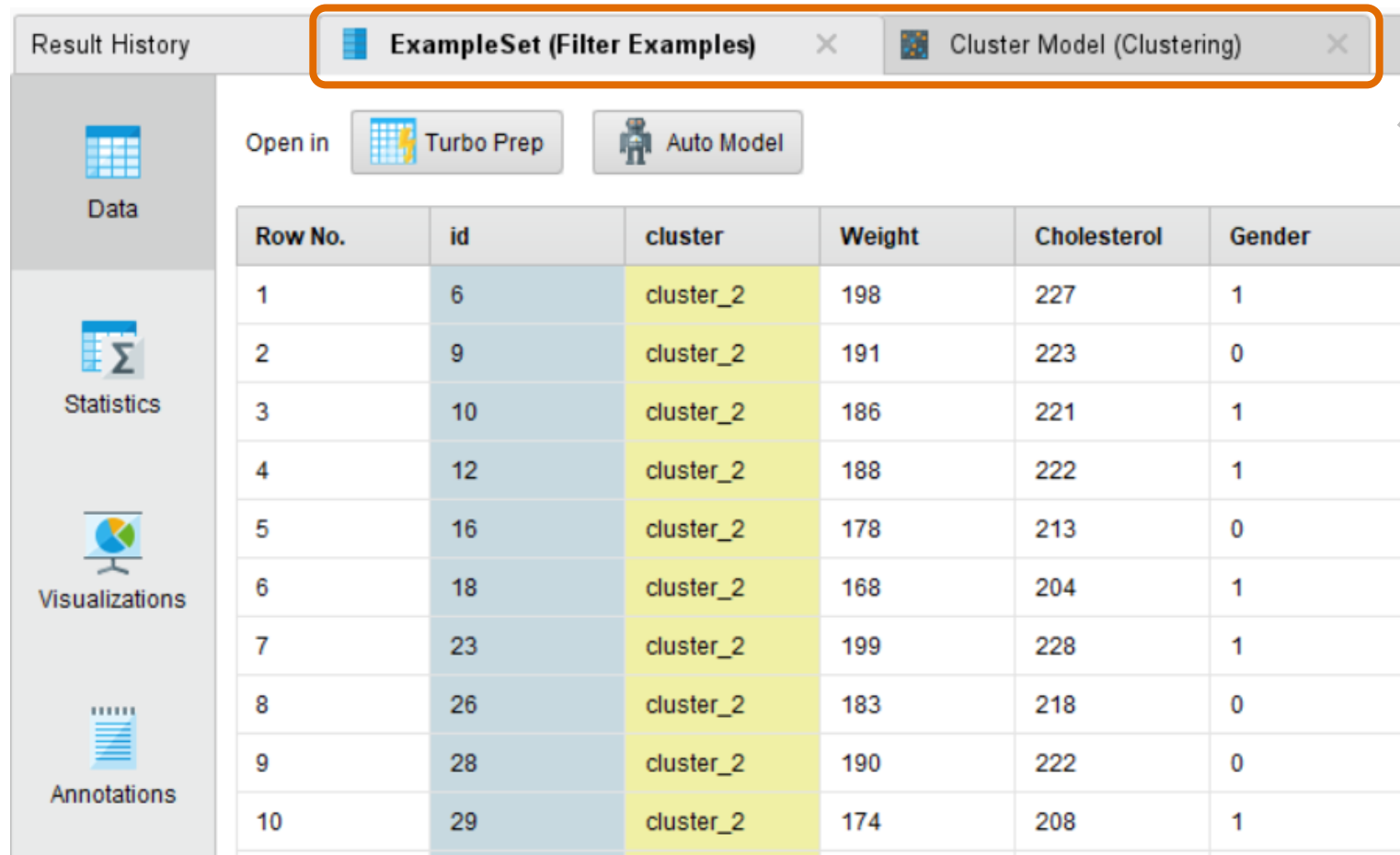


Este parâmetro refere-se ao atributo “cluster” e diz ao RapidMiner para filtrar todas as observações em que o valor desse atributo é o cluster_2. Isto significa que apenas as observações do *dataset* que estão classificadas como cluster_2 serão mantidas.

DEPLOYMENT



4. Execute o modelo.



Result History

ExampleSet (Filter Examples) × Cluster Model (Clustering) ×

Open in Turbo Prep Auto Model

Row No.	id	cluster	Weight	Cholesterol	Gender
1	6	cluster_2	198	227	1
2	9	cluster_2	191	223	0
3	10	cluster_2	186	221	1
4	12	cluster_2	188	222	1
5	16	cluster_2	178	213	0
6	18	cluster_2	168	204	1
7	23	cluster_2	199	228	1
8	26	cluster_2	183	218	0
9	28	cluster_2	190	222	0
10	29	cluster_2	174	208	1

Para além do separador “Cluster Model”, existe o separador “ExampleSet”, que contém apenas as 154 observações que pertencem ao cluster 2.

DEPLOYMENT



O grupo de alto risco tem pesos entre 167 e 203, e níveis de colesterol entre 204 e 235

	Name	Type	Missing	Statistics		Filter (5 / 5 attributes): <input type="text" value="Search for Attributes"/>
	Id	Integer	0	Min 6	Max 543	Average 271.727
	Cluster	Nominal	0	Least cluster_3 (0)	Most cluster_2 (154)	Values cluster_2 (154), cluster_0 (0), ...[2 more]
	Weight	Integer	0	Min 167	Max 203	Average 184.318
	Cholesterol	Integer	0	Min 204	Max 235	Average 218.916
	Gender	Integer	0	Min 0	Max 1	Average 0.591

DEPLOYMENT



A Sónia pode usar estes números para começar a contactar potenciais participantes. Para isso ela deve aceder à base de dados da sua empresa e efetuar uma *query* SQL como esta:

```
SELECT First_Name, Last_Name, Policy_Num, Address, Phone_Num  
FROM PolicyHolders_view  
WHERE Weight >= 167  
AND Cholesterol >= 204;
```



Através desta *query* a Sónia consegue obter a lista de contactos de cada pessoa que se enquadra no grupo de maior risco (cluster 2) na esperança de aumentar a consciencialização, educar os detentores de apólices e modificar comportamentos que levarão a uma menor incidência de doenças cardíacas.

RESUMO



O **k-Means clustering** é um modelo de *Data Mining* que se enquadra principalmente na Classificação. Neste exemplo, ele não prevê necessariamente quais segurados irão ou não desenvolver doenças cardíacas. Ao invés, ele lida com indicadores conhecidos dos atributos num *dataset* e agrupa-os com base na semelhança desses atributos com as médias do grupo.

O **k-Means** é uma forma eficaz de agrupar observações com base no que é típico ou normal para um grupo. Para além disso, ajuda a perceber onde um grupo começa e o outro termina, ou por outras palavras, onde as quebras naturais ocorrem entre grupos num determinado conjunto de dados.

Embora bastante simples na sua configuração e definição, o **k-Means clustering** é um método **poderoso** e **flexível** para encontrar grupos naturais de observações num conjunto de dados.