

Universidade do Minho
Escola de Engenharia

K nowledge D iscovery

José Machado
Cristiana Neto



K-MEANS CLUSTERING WITH RAPIDMINER

CONTEXT E PRESPECTIVE



Sonia is a program director for a major health insurance provider.

Recently she has been reading in medical journals and other articles, and found a strong emphasis on the influence of weight, gender and cholesterol on the development of coronary heart disease.

She begins brainstorming ideas for her company to offer weight and cholesterol management programs to individuals who receive health insurance through her employer.

As she considers where her efforts might be most effective, she finds herself wondering if there are natural groups of individuals who are most at risk for high weight and high cholesterol, and if there are such groups, where the natural dividing lines between the groups occur.

Data Mining can help her understand these groups.

BUSINESS UNDERSTANDING



Sonia's goal is to identify and then try to reach out to individuals insured by her employer who are at high risk for coronary heart disease because of their weight and/or high cholesterol. She understands that those at low risk, that is, those with low weight and cholesterol, are unlikely to participate in the programs she will offer.

She also understands that there are probably policy holders with high weight and low cholesterol, those with high weight and high cholesterol, and those with low weight and high cholesterol. She further recognizes there are likely to be a lot of people somewhere in between.

In order to accomplish her goal, she needs to search among the thousands of policy holders to find groups of people with similar characteristics and craft programs and communications that will be relevant and appealing to people in these different groups.

DATA UNDERSTANDING



Using the insurance company's claims database, Sonia extracts three attributes for 547 randomly selected individuals.

The three attributes are the insured's **weight** in pounds as recorded on the person's most recent medical examination, their last **cholesterol** level determined by blood work in their doctor's lab, and their **gender**. As is typical in many data sets, the gender attribute uses 0 to indicate Female and 1 to indicate Male.

We will use this sample data to build a cluster model to help Sonia understand how her company's clients, the health insurance policy holders, appear to group together on the basis of their weights, genders and cholesterol levels.

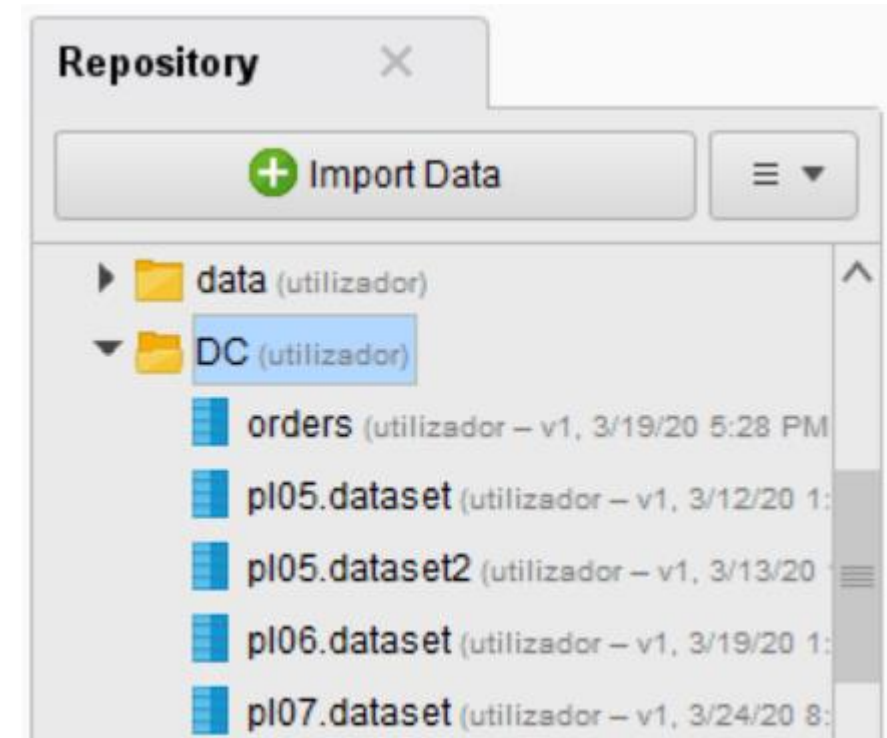
We should remember as we do this that means are particularly susceptible to undue influence by extreme outliers, so watching for inconsistent data when using the k-Means clustering data mining methodology is very important.

DATA PREPARATION



Download dataset: pl07.dataset.csv

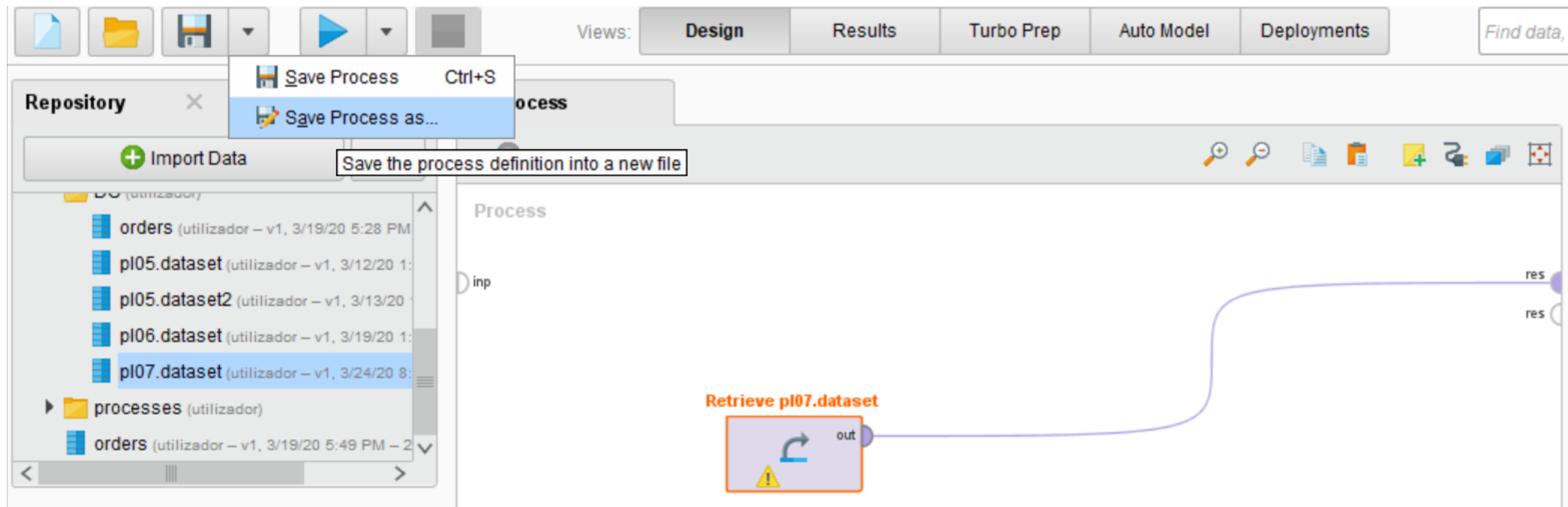
1. Import the CSV into the RapidMiner repository (Import Data -> My Computer)
2. Check the results view and inspect the imported CSV data (Data, Statistics)



DATA PREPARATION



3. Drag the **pl07.dataset** dataset to a new process window in RapidMiner
4. Run the model to inspect the data and save the process as **pl07_processo**.



DATA PREPARATION



5. Select the “Results” view and choose the “Statistics” option. Note that:

- There is no missing value for any of the 12 attributes.
- None of the values appear to be inconsistent (remember the comments from the previous lesson on using standard deviations to find statistical discrepancies).

MODELING



The 'k' in k-means clustering stands for some number of groups (clusters). The aim of this data mining methodology is to look at each observation's individual attribute values and compare them to the means of potential groups of other observations in order to find natural groups that are similar to one another.

The k-means algorithm accomplishes this by sampling some set of observations in the data set, calculating the averages, or means, for each attribute for the observations in that sample, and then comparing the other attributes in the data set to that sample's means.

The system does this repetitively in order to 'circle-in' on the best matches and then to formulate groups of observations which become the clusters. As the means calculated become more and more similar, clusters are formed, and each observation whose attributes values are most like the means of a cluster become members of that cluster.

MODELING



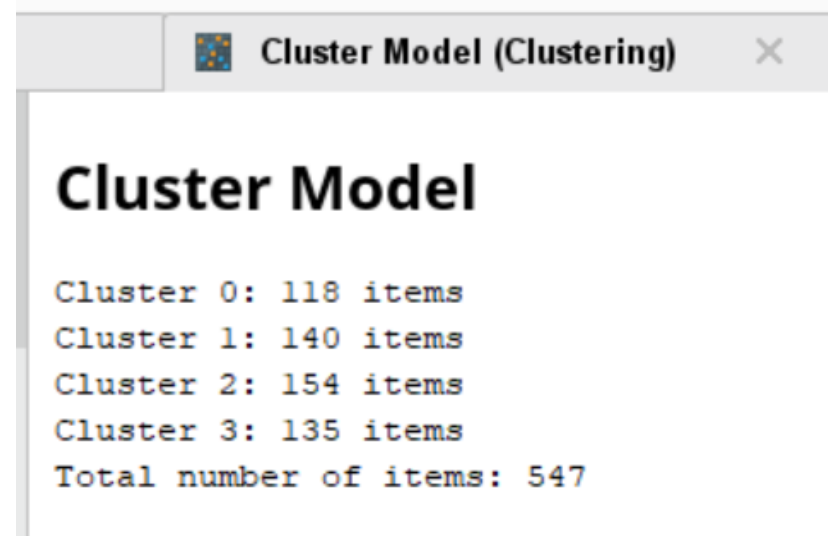
1. Search and drag k-means operator into the process. Regarding the value of k (in the parameters on the right side), as there are likely to be at least four potentially different groups, we will change the value of k to 4. Click Run.

The screenshot displays the Orange3 data mining software interface. The top bar includes tabs for 'Design', 'Results', 'Turbo Prep', 'Auto Model', and 'Deployments'. The 'Design' tab is active, showing a workflow canvas with two operators: 'Retrieve pl07.dataset' and 'Clustering'. The 'Clustering' operator is highlighted with an orange border. To the left, the 'Repository' pane shows a list of datasets, including 'pl07.dataset'. Below it, the 'Operators' pane shows a search for 'k-means' with several results listed under the 'Modeling' category. On the right, the 'Parameters' pane for the 'Clustering (k-Means)' operator is open. It shows various settings: 'add cluster attribute' is checked, 'add as label' is unchecked, and 'remove unlabeled' is unchecked. The 'k' parameter is set to 4, with an orange arrow pointing to its input field. Other parameters include 'max runs' (10), 'determine good start values' (checked), 'measure types' (BregmanDivergences), 'divergence' (SquaredEuclideanDist...), and 'max optimization steps' (100).

MODELING



2. When the model is executed, we find an initial report on the number of items that remained in each of our four clusters. In this particular model, our clusters are reasonably well balanced.



```
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

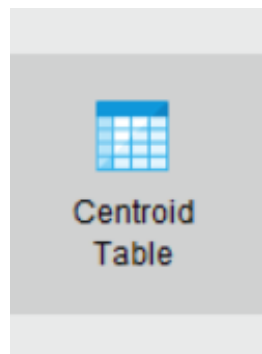
At this point, we could go back and adjust our number of clusters, our 'max-runs' value or even try other parameters presented by the k-Means operator.

EVALUATION



Recall that Sonia's major objective in the hypothetical scenario posed at the beginning of the chapter was to try to find natural breaks between different types of heart disease risk groups. Using the k-Means operator in RapidMiner, we identified four groups, and we can now assess their usefulness.

1. Select the "Centroid Table" option. This window contains the averages for each attribute in each of the four clusters created.



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



- Cluster 2 has the highest average “Weight” and “Cholesterol”;
- With 0 representing Female and 1 representing Male, an average of 0.591 indicates that we have more men than women in this cluster.

EVALUATION



High cholesterol and weight are two key indicators of the risk of heart disease that policy holders can do something about.

What does this mean?



Sonia should start with members of cluster 2 when promoting its new programs and then extend it to members of clusters 0 and 3, who are, respectively, the members with the highest averages for these two key risk factor attributes.

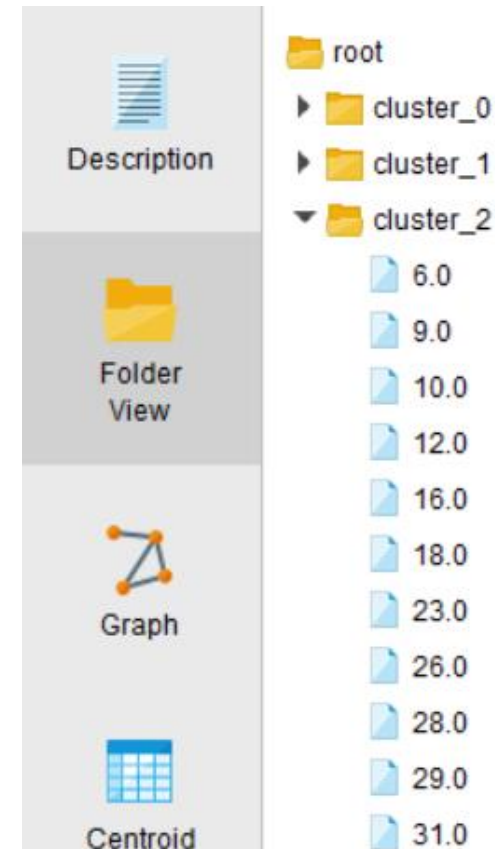
EVALUATION



Sonia knows that cluster 2 is where she will concentrate her first efforts, but how does she know who to contact? Who are the members of this highest risk group?



2. Select the “Folder View” option to access this type of information.



EVALUATION




3. Click on an observation to see its details.

The means for cluster 2 were just over 184 for weight and just under 219 for cholesterol. The person represented in observation 6 is heavier and has a higher cholesterol than the average for this highest risk group.




This is a person that Sonia can help!



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




We know from the description of the Cluster Model that there are 154 members in the dataset that fit this group.




Clicking on each of them is a time-consuming and inefficient process.



We can help Sonia to extract the observations from cluster 2 very quickly and easily.


Description


Folder View

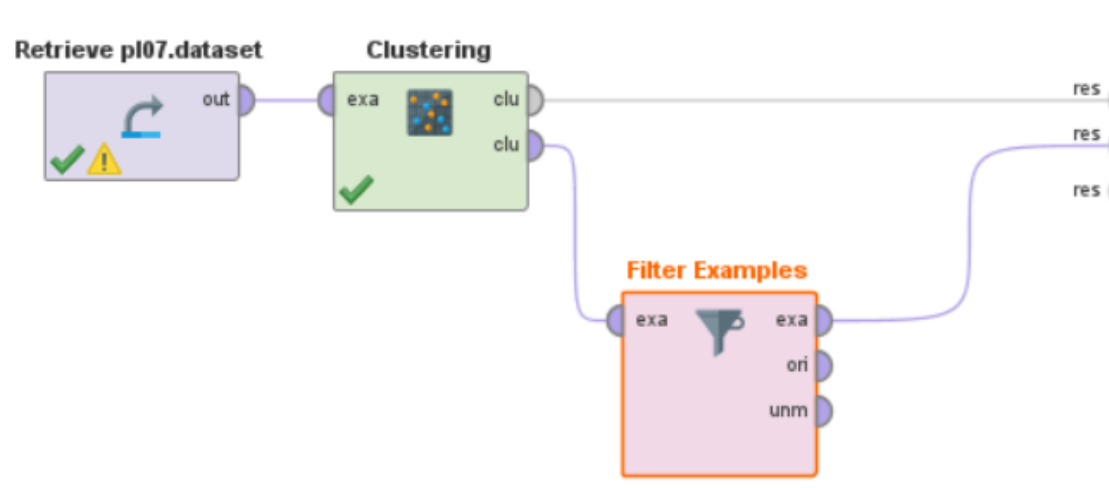
Cluster Model

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

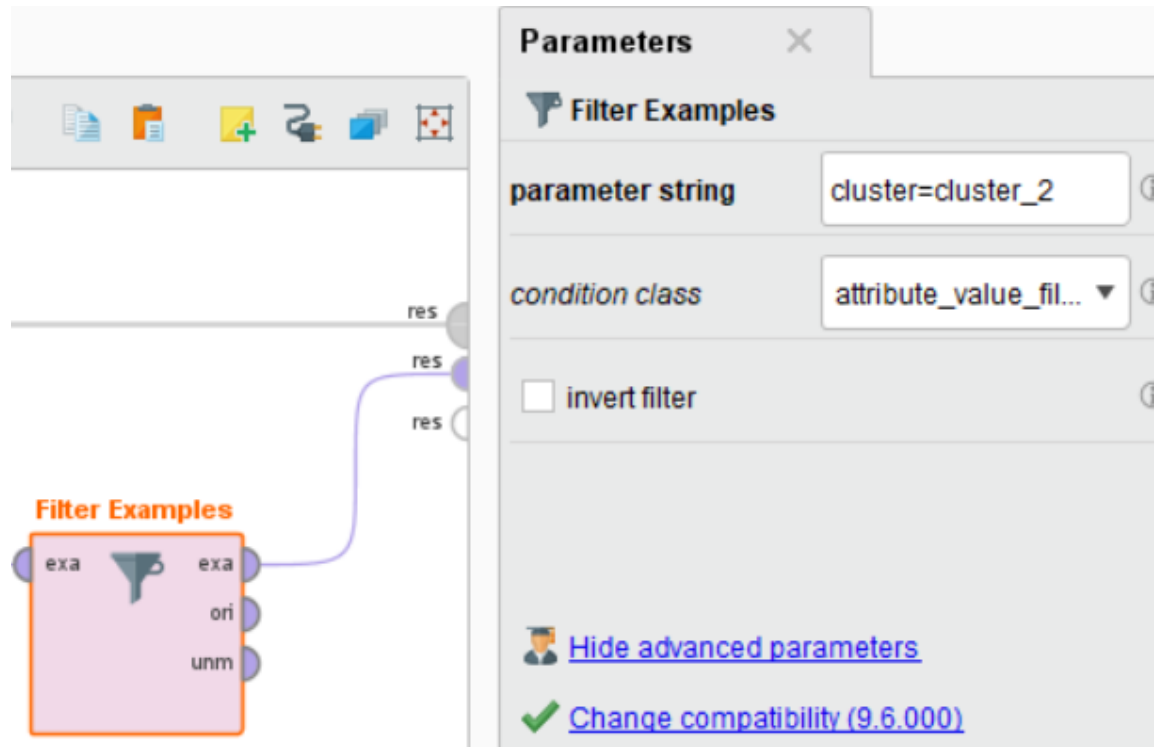
DEPLOYMENT



1. Get back to the Design perspective in RapidMiner.
2. Find and drag the “Filter Examples” operator and connect it to the k-Means Clustering operator. Connect the second ‘clu’ (cluster) port to the ‘exa’ port of the “Filter Examples” operator, and connect the ‘ex’ port of the ‘Filter Examples’ to the final ‘res’ port.



DEPLOYMENT



3. In the “condition class” field, select the ‘attribute_value_filter’ option, and for the “parameter string” field, type the following: cluster = cluster_2

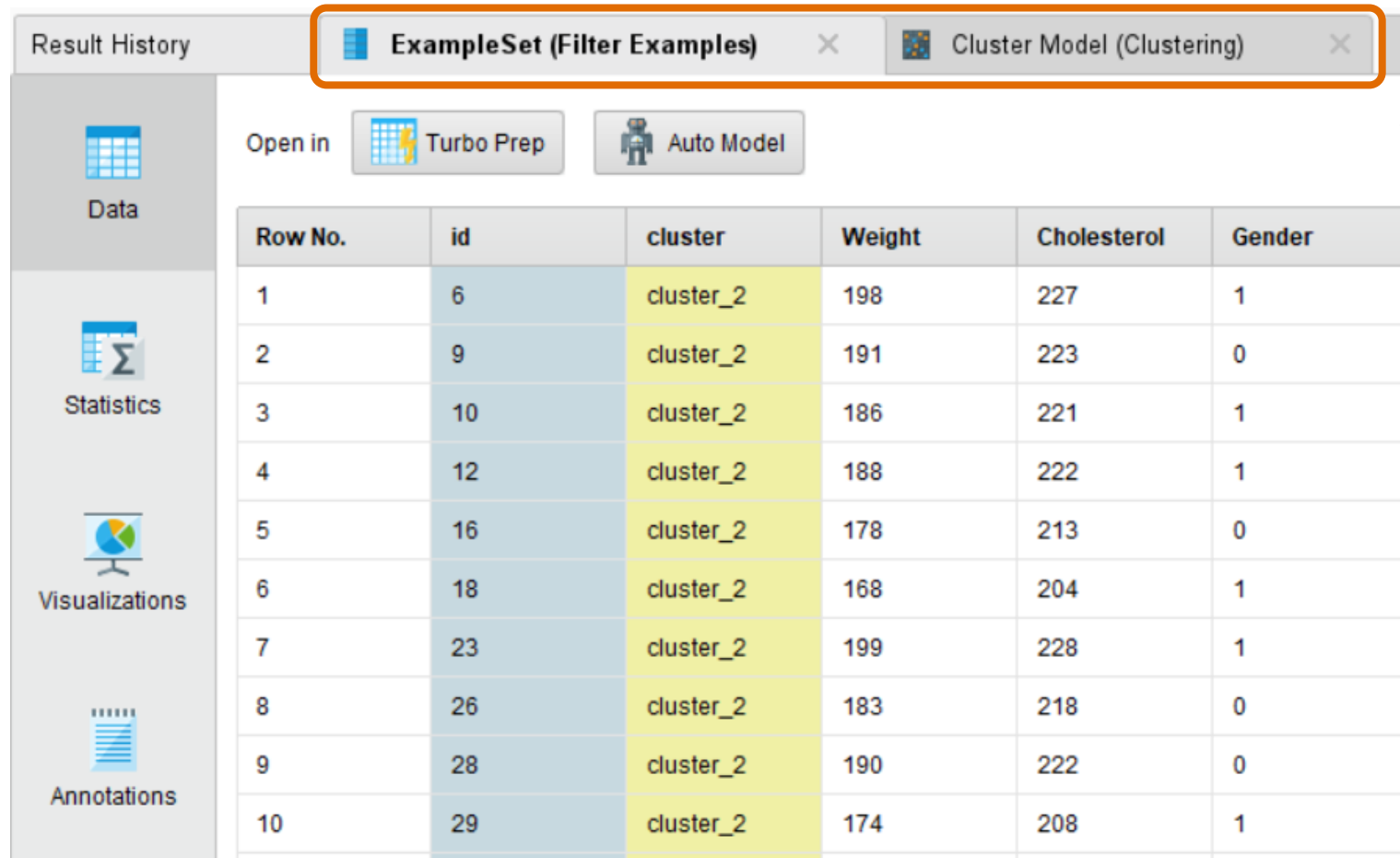


This parameter refers to the “cluster” attribute and tells RapidMiner to filter all observations where the value of that attribute is cluster_2. This means that only observations from the dataset that are classified as cluster_2 will be kept.

DEPLOYMENT



4. Execute the model.



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

In addition to the “Cluster Model” tab, there is the “ExampleSet” tab, which contains only the 154 observations that belong to cluster 2.

DEPLOYMENT



The high-risk group has weights between 167 and 203, and cholesterol levels between 204 and 235

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

DEPLOYMENT



Sonia can use these numbers to start contacting potential participants. For that, she must access her company's database and perform an SQL query like this:

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



Through this query, Sonia is able to obtain the contact list of each person who falls into the group at greatest risk (cluster 2) in the hope of raising awareness, educating policy holders, and modifying behaviors that will lead to lower incidence of heart disease among her employer's clients.

SUMMARY



k-Means clustering is a data mining model that falls primarily on the side of Classification.

For this example, it does not necessarily predict which insurance policy holders *will* or *will not* develop heart disease. It simply takes known indicators from the attributes in a data set, and groups them together based on those attributes' similarity to group averages.

Because any attributes that can be quantified can also have means calculated, k-means clustering provides an effective way of grouping observations together based on what is typical or normal for that group. It also helps us understand where one group begins and the other ends, or in other words, where the natural breaks occur between groups in a data set. k-Means clustering is very flexible in its ability to group observations together.

While fairly simple in its set-up and definition, k-Means clustering is a powerful method for finding natural groups of observations in a data set.