Getting Started with RapidMiner

Download tool and datasets:
http://rapidminer.com/

Agathe Merceron
Berliner University of Applied Sciences
Germany

Outline

Design and result perspectives in RapidMiner.

The Point Dataset Point1.csv:

- Reading the data
- Visual exploration to understand the data and guide the work
- Clustering and the Loop operator: searching for the right k for k-means.

Design and Result Perspectives

Organize your space to store processes, files and results.

Point1:

8 points defined by their x-, y-coordinates.

Reading the Point1-Dataset

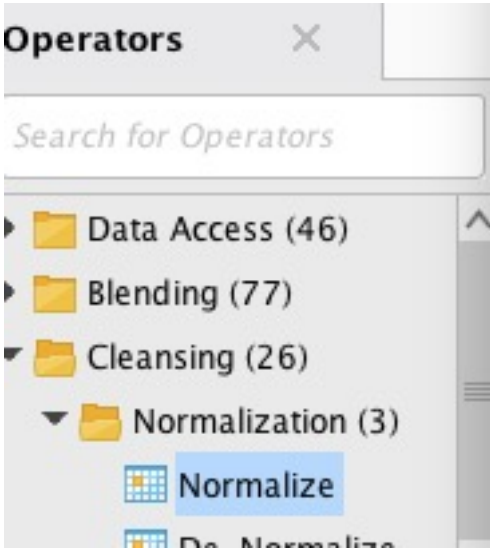
Warning Read Wizard: RapidMiner guesses the type of attributes looking at the first 100 by default.

Metadata:

- Check the order of magnitudes of attributes
- Handy for classical transformations such as:

$$\frac{x - average}{std\ deviation} \qquad \frac{x - min}{range}$$

- Transformation can be done inside RM:



Exploring the Point1-Dataset

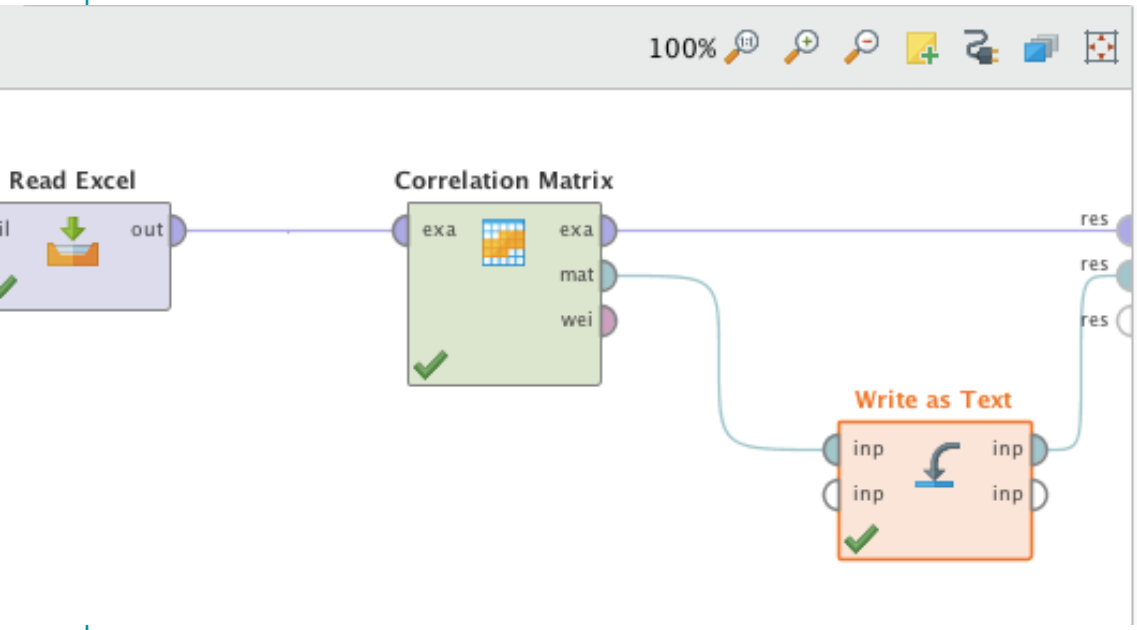
Plots:

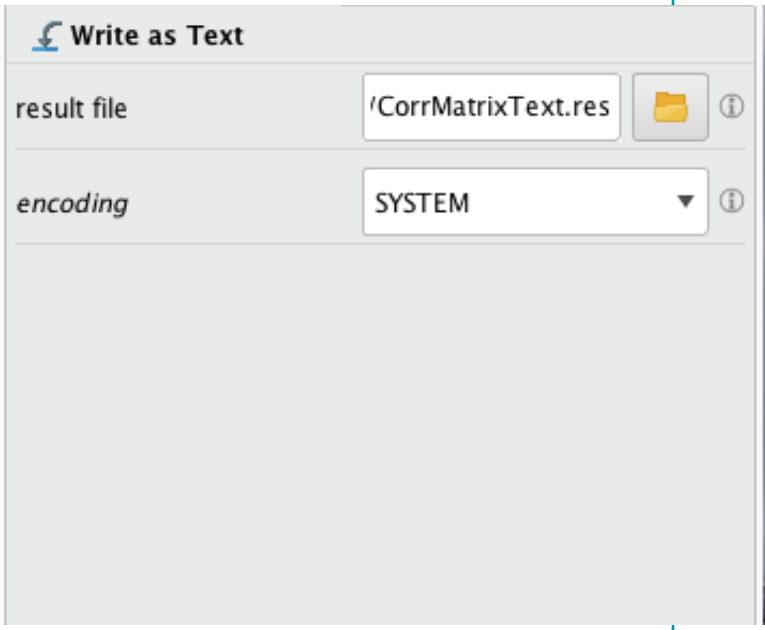
- Histogram / color
- Box plots (quartile)
- Scatter
- and much more...

Correlation Operator:

- Handy: store the matrix with the Write as text Operator

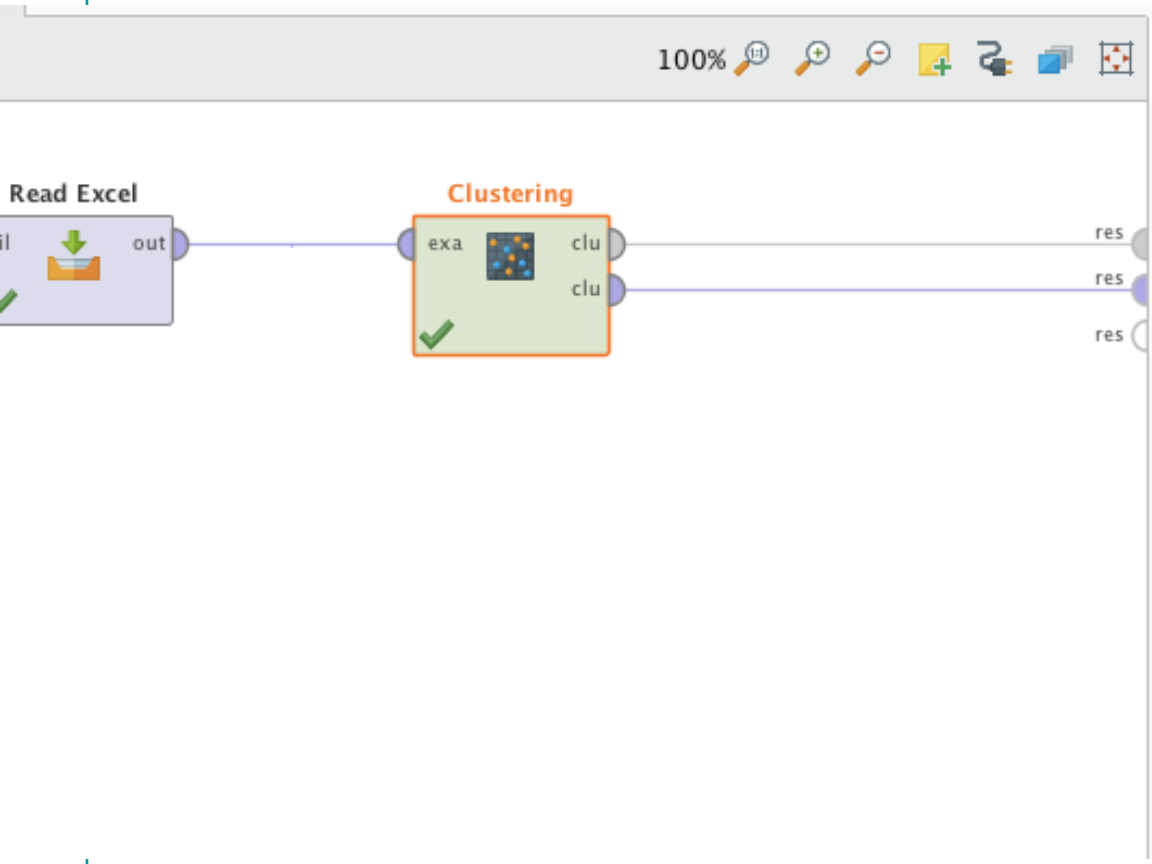
Exploring the Point1-Dataset

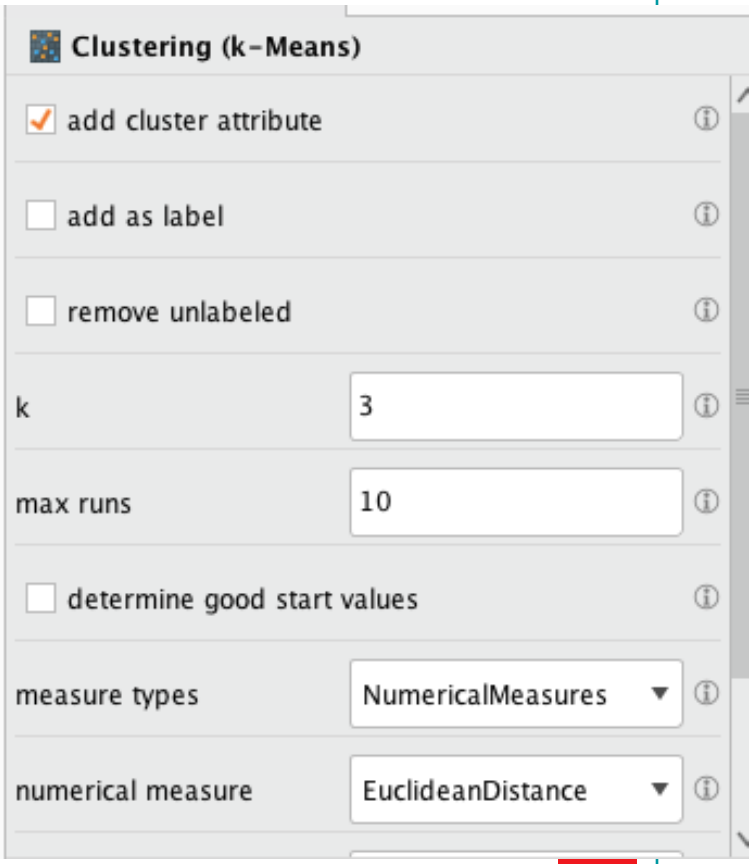




Looking for groups of points with Clustering

K-means / Scatter Plot with Color Column cluster

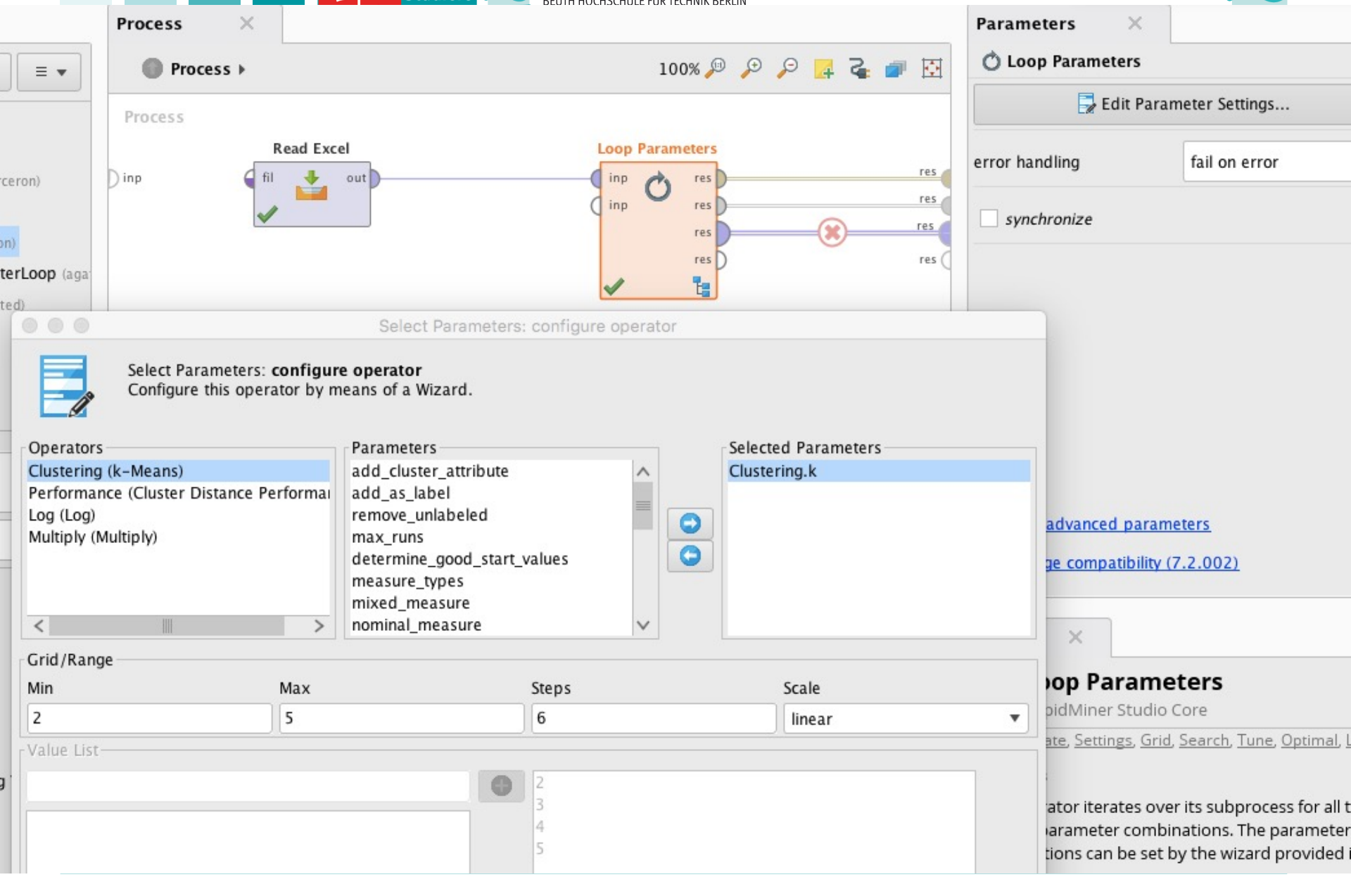




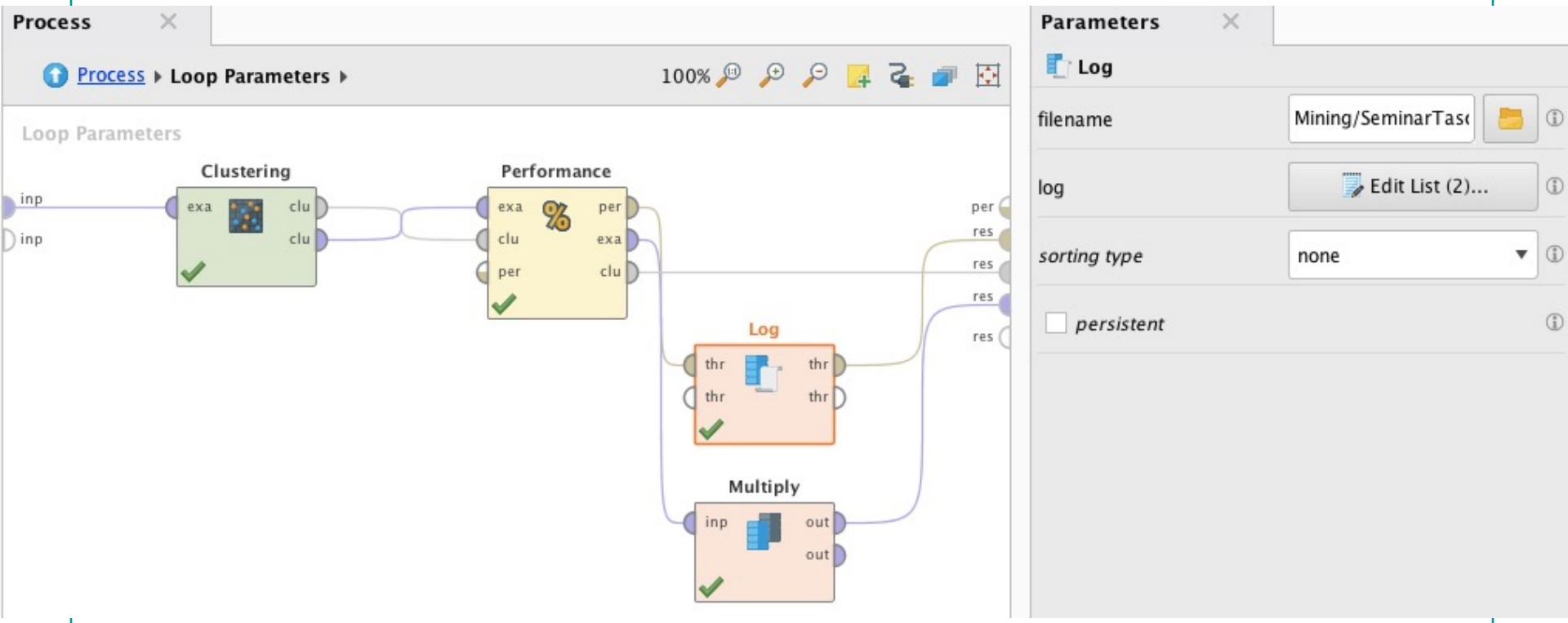
Looking for groups of points with Clustering

Loop Operator to discover the right k:

- Warning: skip the first port per inside the loop operator to get the results
- Handy: Log Operator to plot average within distance cluster or Davies Bouldin against k. Edit the parameters you want to plot, k and avg_within_distance for instance.

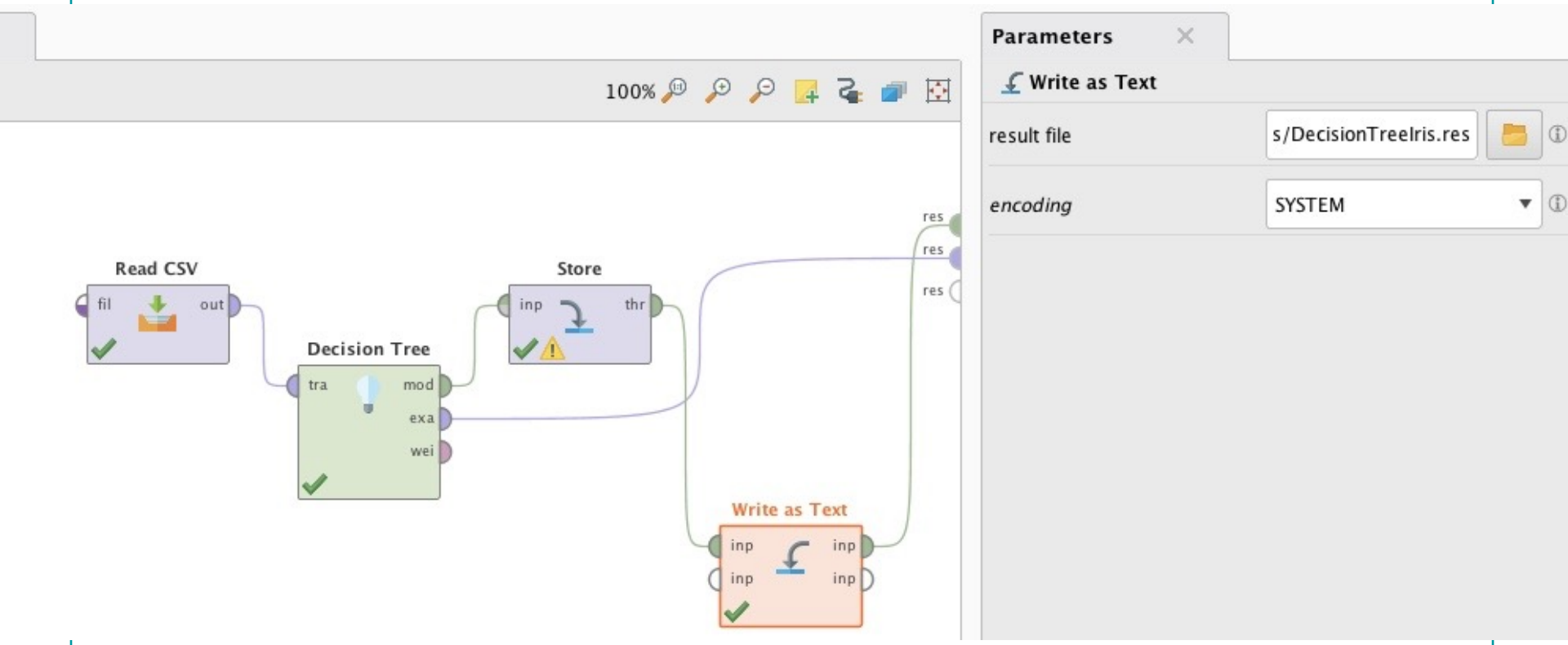

Process
Process
100%
Process
Read Excel
inp
fil
out
Loop Parameters
inp
inp
res
res
res
res
res
res
res
Parameters
Loop Parameters
Edit Parameter Settings...
error handling
fail on error
synchronize
Select Parameters: configure operator
Select Parameters: configure operator
Configure this operator by means of a Wizard.
Operators
Clustering (k-Means)
Performance (Cluster Distance Performa
Log (Log)
Multiply (Multiply)
Parameters
add_cluster_attribute
add_as_label
remove_unlabeled
max_runs
determine_good_start_values
measure_types
mixed_measure
nominal_measure
Selected Parameters
Clustering.k
Grid/Range
Min
Max
Steps
Scale
2
5
6
linear
Value List
2
3
4
5
advanced parameters
ge compatibility (7.2.002)
oop Parameters
oidMiner Studio Core
ate, Settings, Grid, Search, Tune, Optimal,
ator iterates over its subprocess for all t
arameter combinations. The parameter
tions can be set by the wizard provided

Looking for groups of points with Clustering



Classification with the iris dataset

Class should be of type Label



Classification

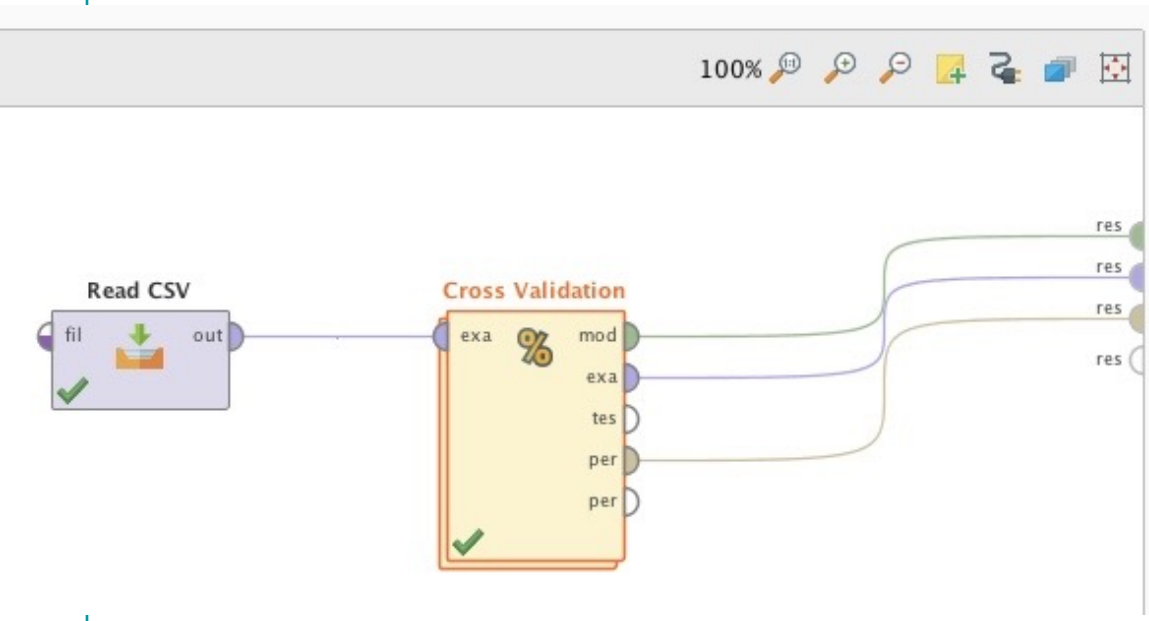
Decision tree:

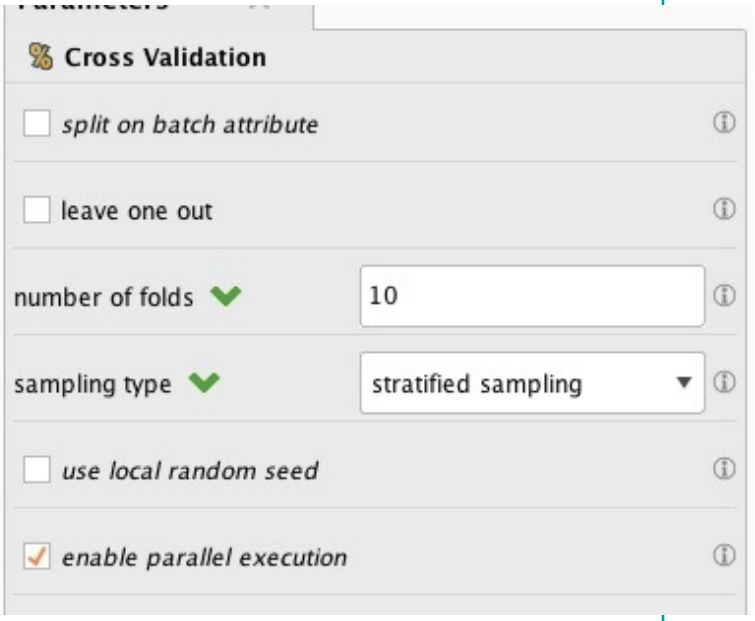
- label to set the attribute to predict

Validation:

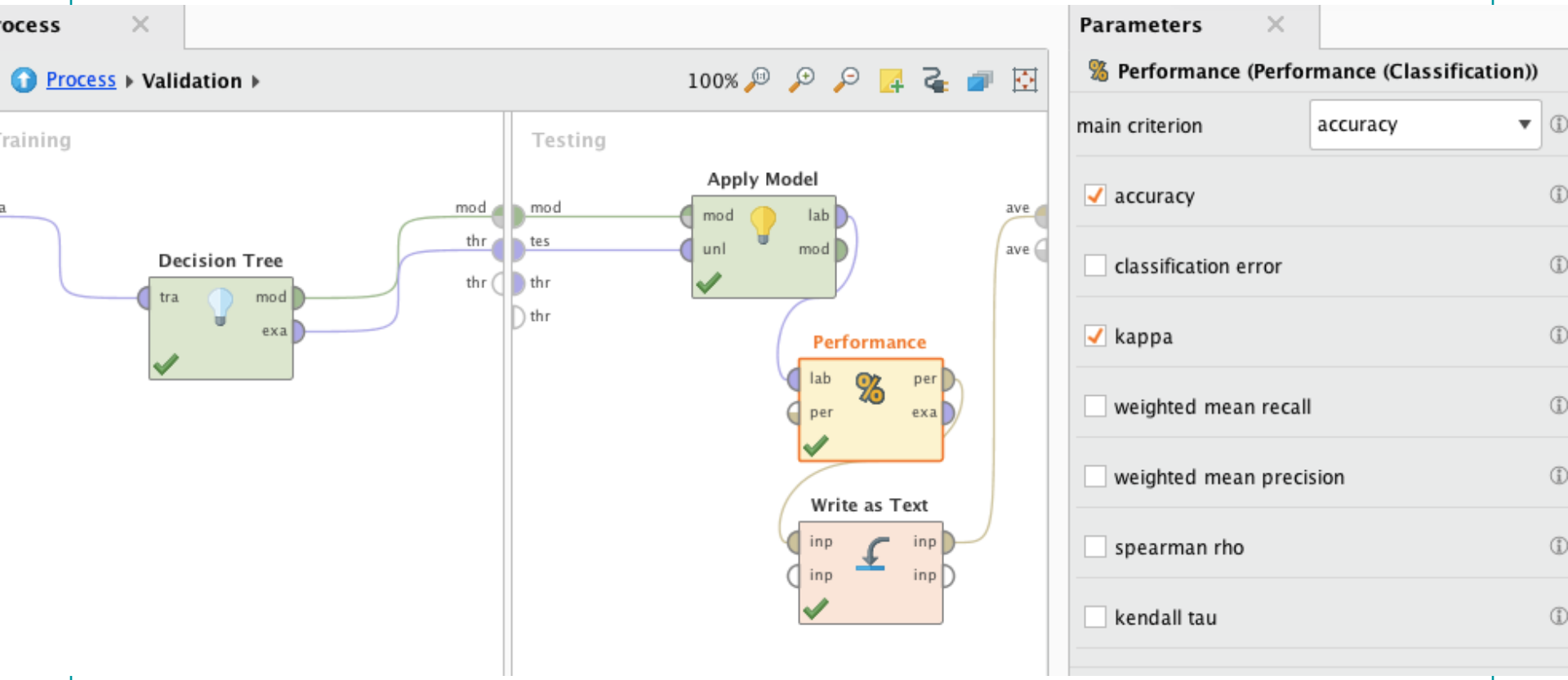
- X Validation Operator
- Handy: Write as Text Operator the output of Performance Operator to check the result step by step.

Classification





Classification



References

Rapid Miner website: rapidminer.com with good tutorials.

- https://community.rapidminer.com/
- https://docs.rapidminer.com/

Many videos on youtub, see for example:

https://www.youtube.com/watch?v=C8Ko3-2f-pA&list=PLssWC2d9JhOZLbQNZ80uOxLypglgWqbJA&index=16

References

Thank your for your attention!

Schloß Charlottenburg, Berlin