

Data Mining

(Mining Knowledge from Data)

Data Preprocessing

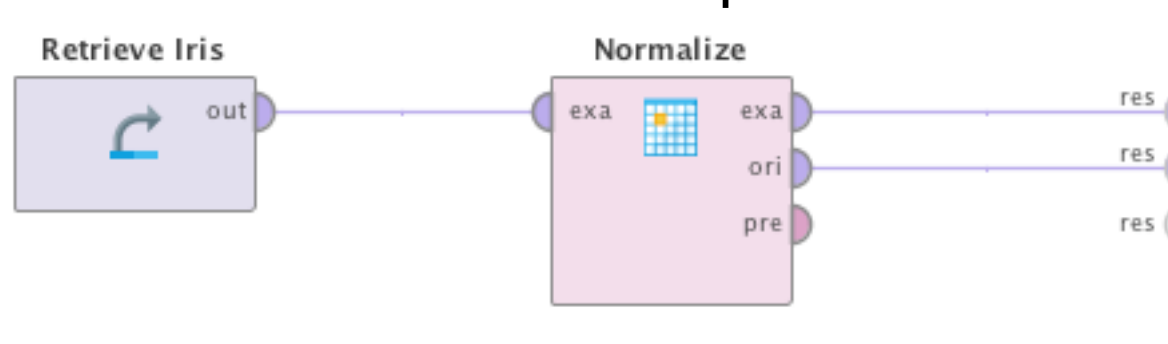
Magda Friedjungová

Exercise Outline

- Normalization
- Types of attributes
- Transformation of attributes
- Selection of attributes

Normalization

- Import Iris dataset from the repository.
- Add the Normalize operator.



- Try the Range transformation.
- Try the Z-transformation.

The screenshot shows the 'Parameters' window for the 'Normalize' operator. The window has a title bar with a close button. The main area is titled 'Normalize' and contains several settings:

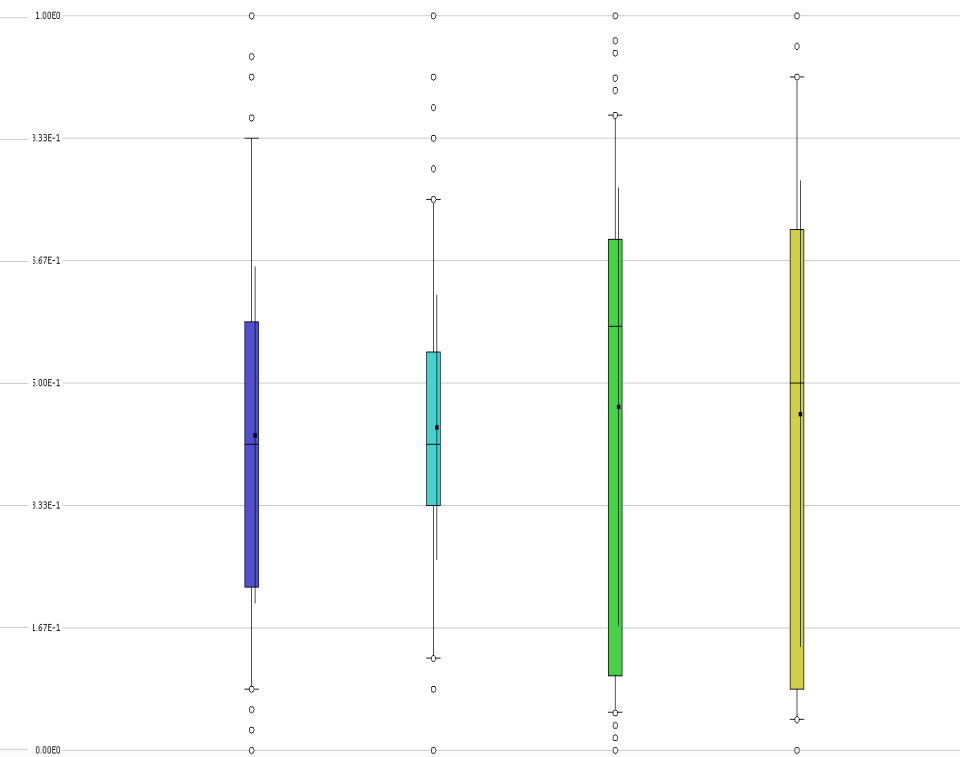
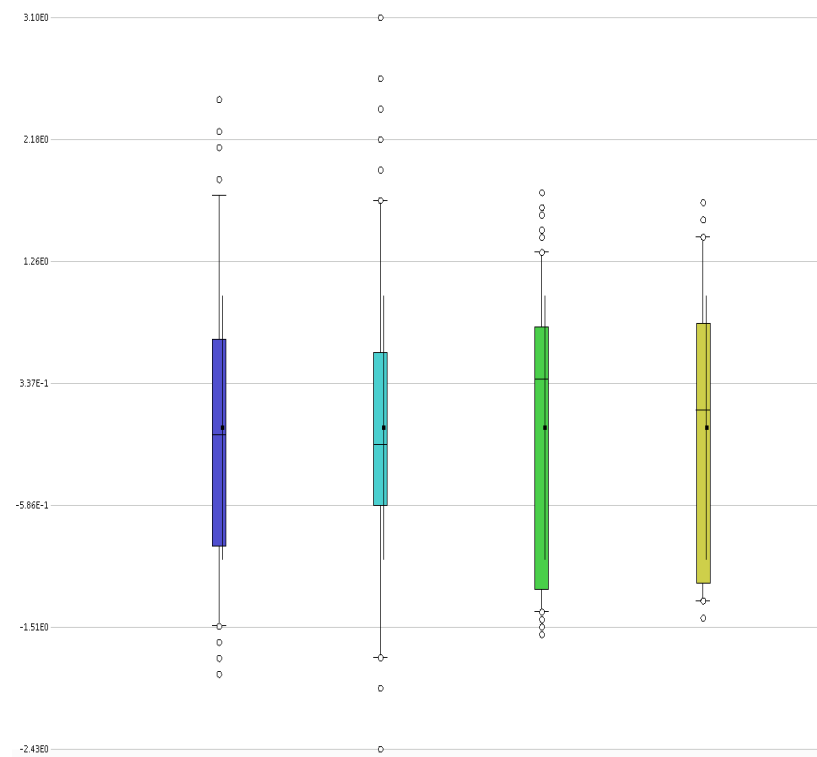
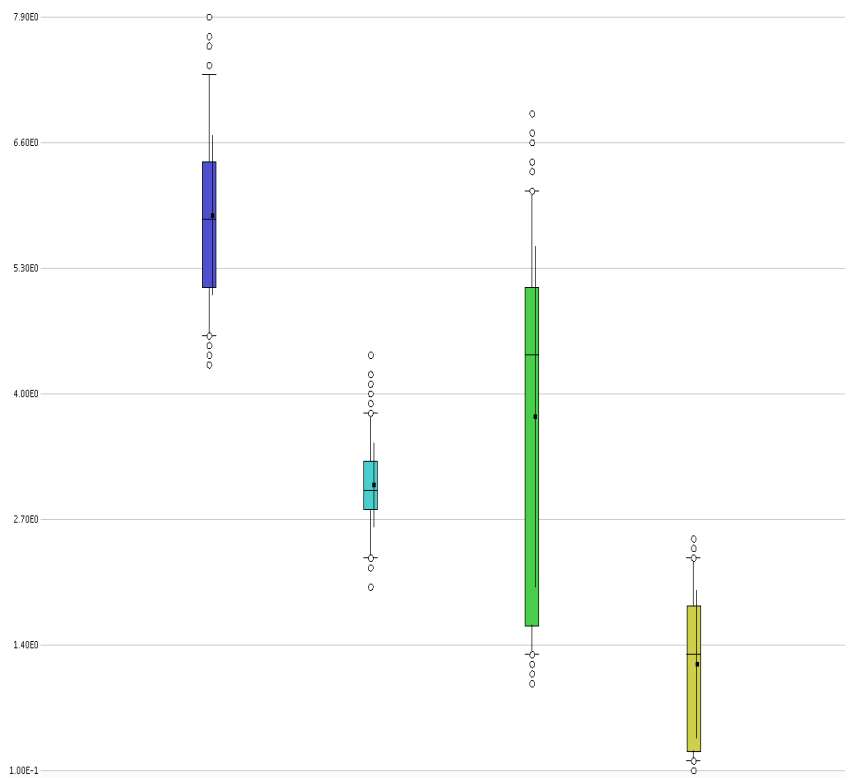
- ☐ create view
- attribute filter type: ☒ all
- ☐ invert selection
- ☐ include special attributes
- method: ☒ range transformati...
- min: 0.0
- max: 1.0

Results

- Original

- Z-transform

- 0 - 1

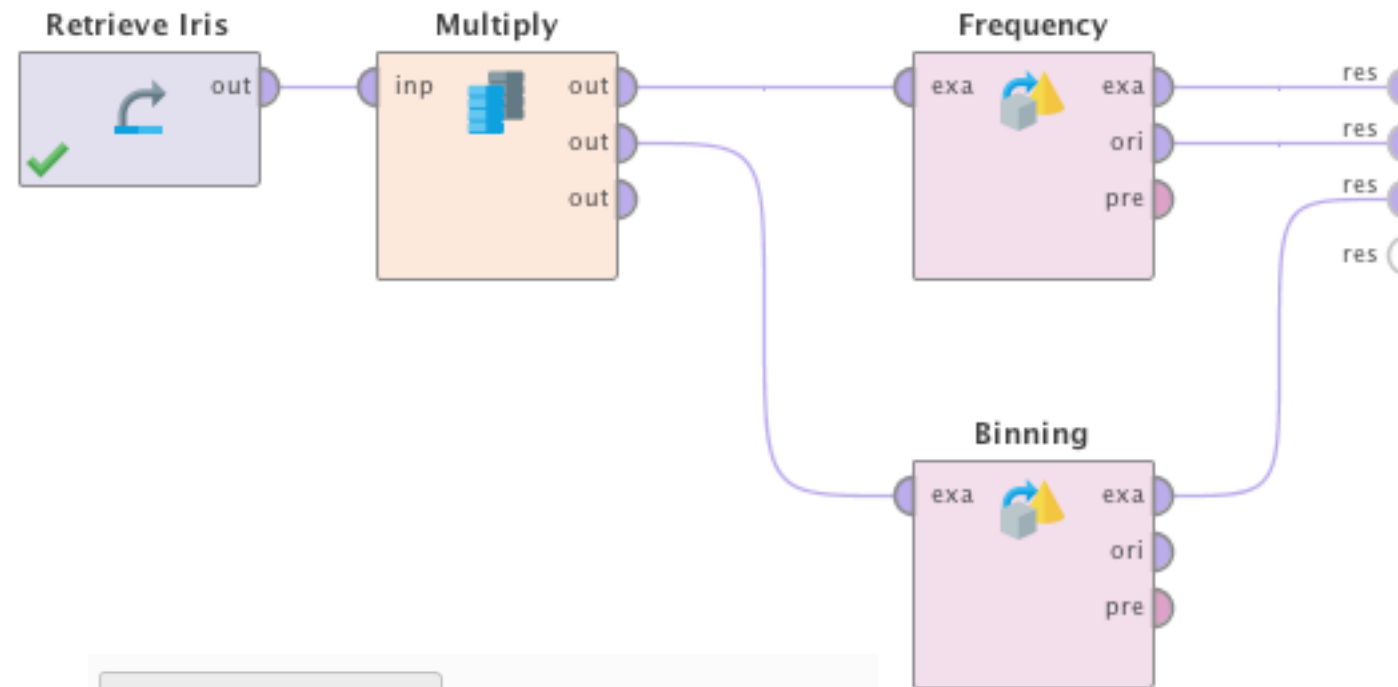


Types of Attributes

- Some models work only with certain types of attributes:
 - Neural Network - numerical
 - Decision Tree - nominal
 - Association Rules - binominal
 - etc.

Discretization

- Discretize by
 - frequency
 - binning



- Enter the number of bins

Parameters

Frequency (Discretize by Frequency)

☐ create view

attribute filter type ☒ all

☐ invert selection

☐ include special attributes

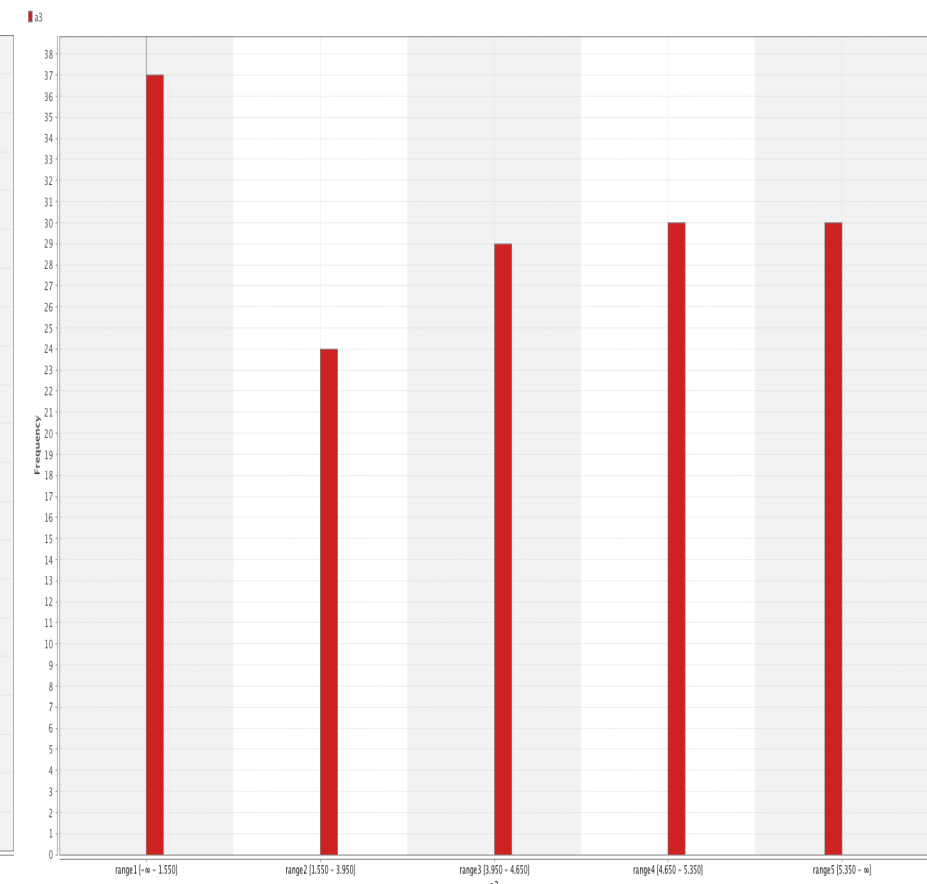
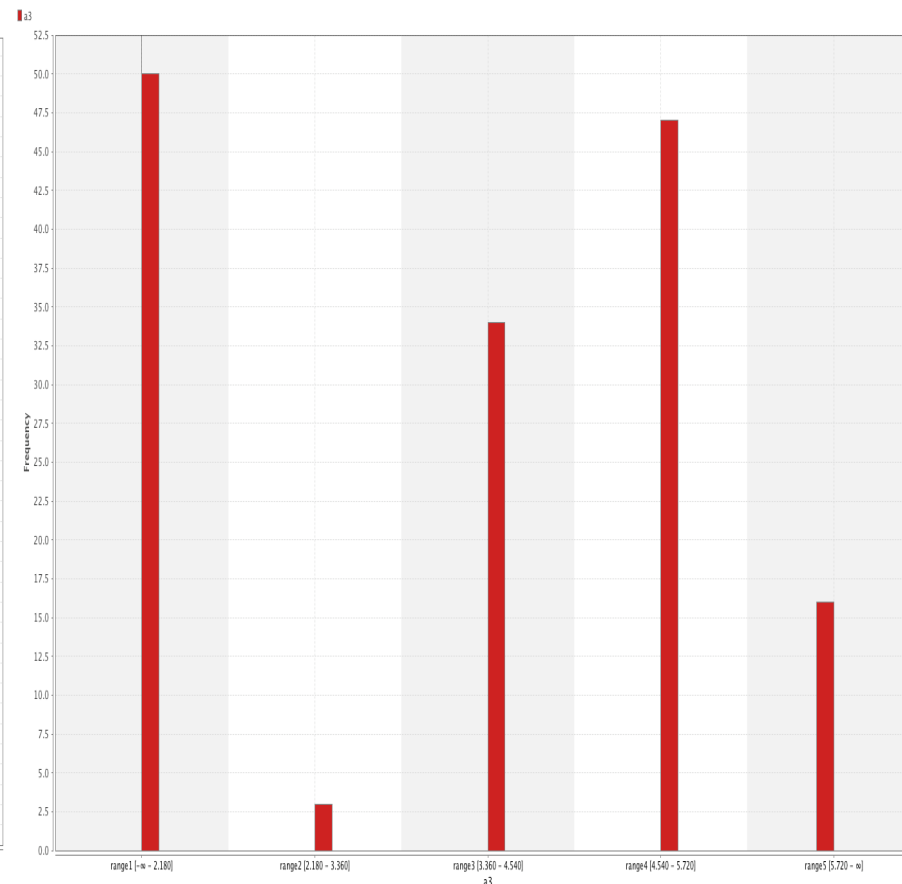
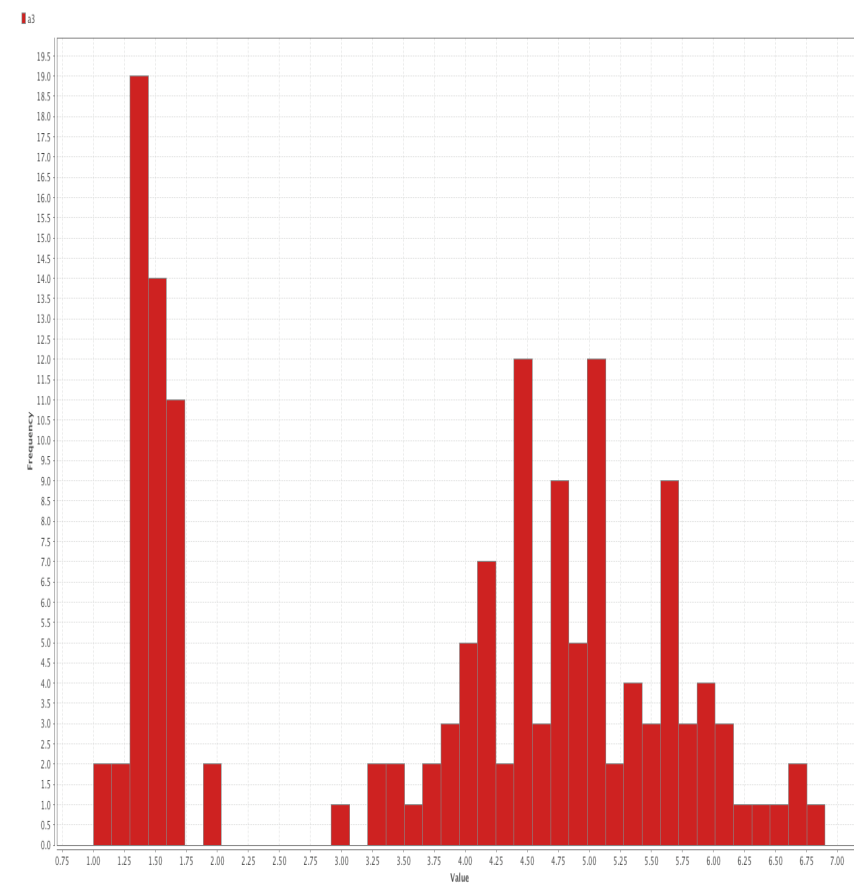
☐ use sqrt of examples

number of bins ☒ 5

range name type ☒ long

Visualization of the Results

- Original
- Binning
- Frequency



- Note that in the case of the Discretization by frequency the intervals are of different lengths.

Other Transformations

- Try other transformations (operators):
 - Nominal to Binominal
 - Nominal to Numeric
 - Numerical to Binominal
 - ...

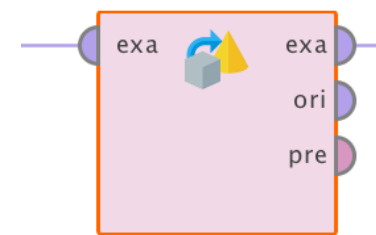
Selection of Attributes

- Manually
- Balancing of attributes
- Automatic selection of attributes

Manual Selection

- Transfer Iris data to Nominal
- Indetify the Setosa class
- Set Role operator

Nominal to Binominal



Nominal to Binominal

☐ create view

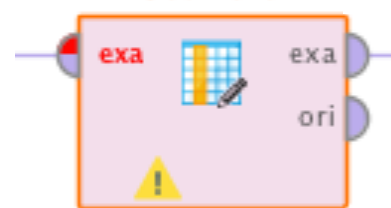
attribute filter type ☒ single

attribute label

☐ invert selection

☒ include special attributes

Set Role



Set Role

attribute name label = Iris-setosa

target role label

set additional roles [Edit List \(0\)...](#)

Select Attributes: **attributes**
The attribute which should be chosen.

Attributes

[Filter] [X]

id
label = Iris-versicolor
label = Iris-virginica

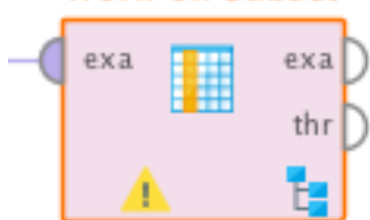
Selected Attributes

[Filter] [X]

a1
a2
a3
a4
label = Iris-setosa

- Work on Subset operator

Work on Subset



Parameters

Work on Subset

attribute filter type ☒ subset

attributes [Select Attributes...](#)

☐ invert selection

☒ include special attributes

name conflict handling error

role conflict handling error

☒ keep subset only

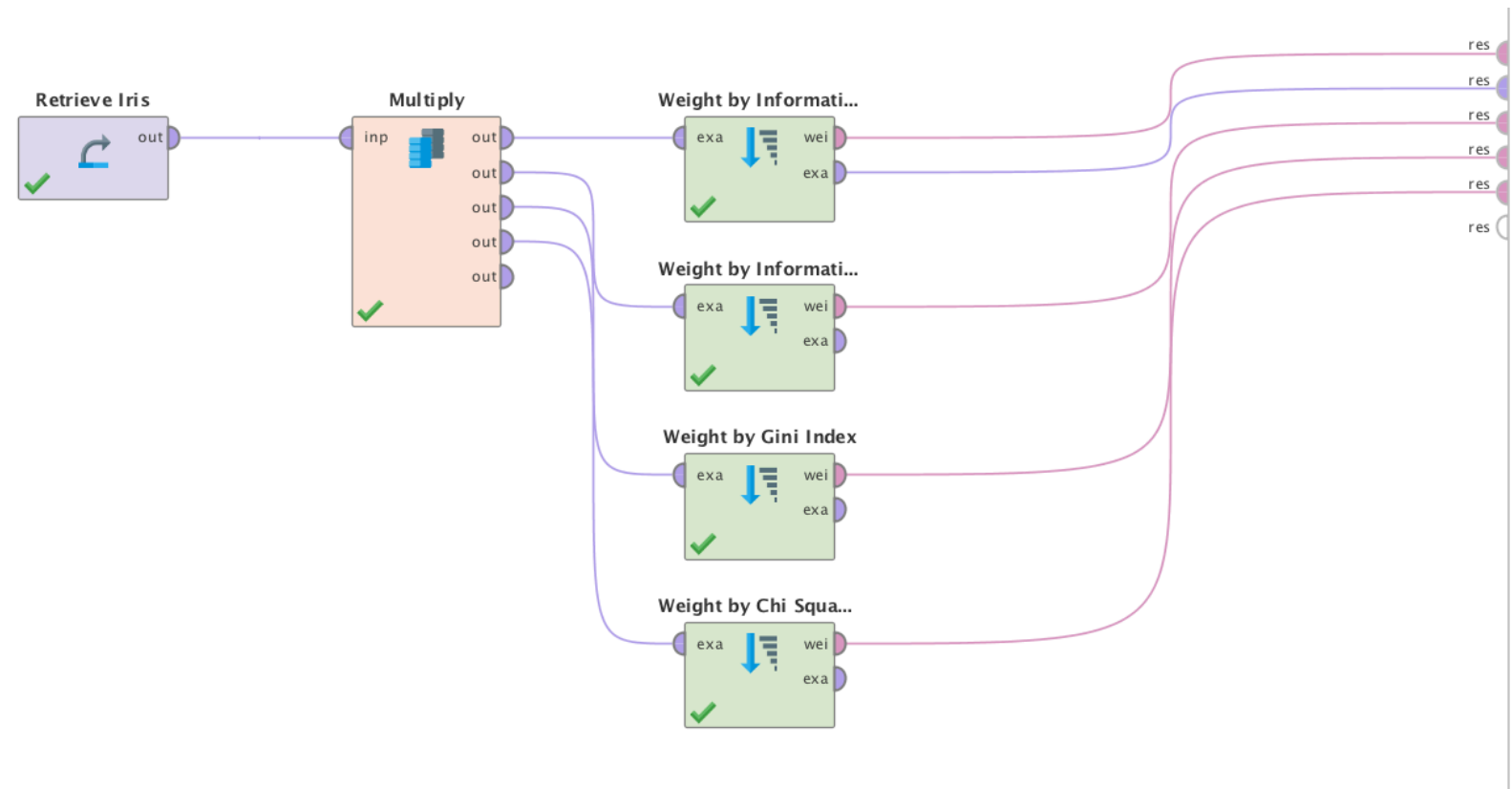
☐ deliver inner results

☐ remove roles



Evaluation of the Contribution of Attributes

- Find the contribution of each attribute on the Iris data using various methods of evaluation
- Use different methods from the group of Attribute weighting operators, e.g.
 - Information Gain
 - Information Gain Ratio
 - Gini Index
 - Chi Squared Statistic



- Compare methods among themselves.

AttributeWeights (Weight by Chi Squared Statistic)

☒ Table View ☐ Plot View ☐ Annotations

attribute	weight
a2	0
a1	0.286
a3	0.897
a4	1

AttributeWeights (Weight by Information Gain)

☒ Table View ☐ Plot View ☐ Annotations

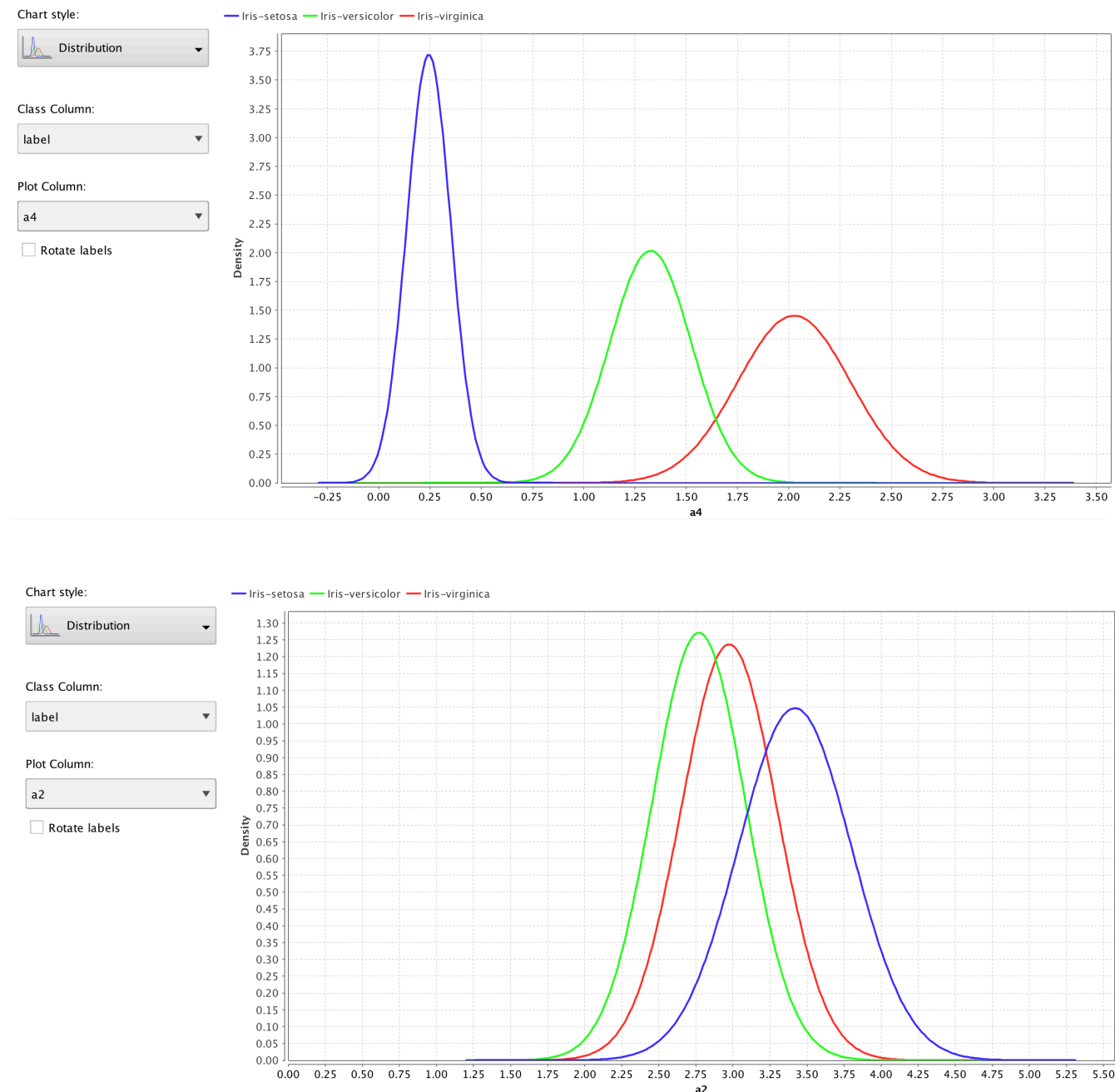
attribute	weight
a2	0
a1	0.498
a3	1
a4	1

AttributeWeights (Weight by Gini Index)

☒ Table View ☐ Plot View ☐ Annotations

attribute	weight
a2	0
a1	0.557
a3	1
a4	1

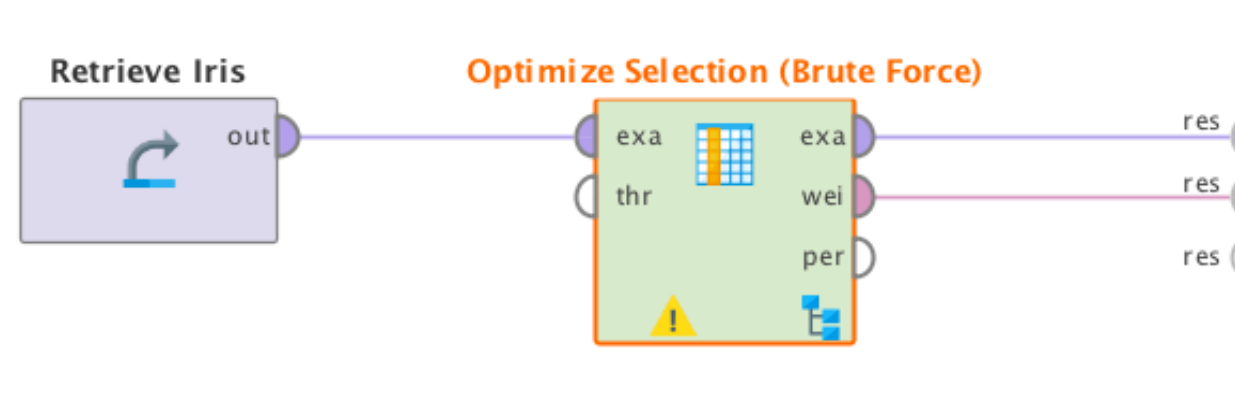
- Let's see the original Iris data, e.g. by using the Distribution plot.



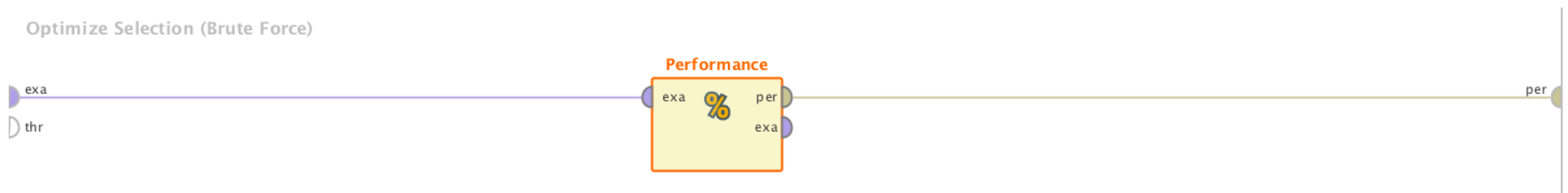
- We see that the attribute a4 separated the classes well.
- On the contrary, attribute a2 does not help us with the classification.

Selection of the Attributes

- Let's try some methods of choosing the best subset of attributes.
- Load the Iris dataset.
- Add Optimize Selection operator for searching the feature space.



- Add a method for evaluating a set of attributes. For example Performance (CFS) operator.



- Inspect the selected attributes.

attribute	weight
a1	0
a2	0
a3	1
a4	1

Wrapper Methods

- The Wrapper methods require a model.
- This rating of a set of attributes uses the misclassification error of any classifier calculated by cross-validation (more details in lecture 5).

