Data Mining

(Mining Knowledge from Data)

Decision Trees

Magda Friedjungová, Marcel Jiřina

Decision Tree Construction

- Top-down approach
  - Go through the training data and find the attribute which best divides the data into classes.
  - Divide the data by the values of the attribute.
  - Treat each group recursively until it is composed of one class only.
- Bottom-up approach is also possible.

Titanic

- Download data about Titanic passengers from EDUX.
- Explore the data:
  - Is it a man? / Is it a woman?
  - Adult? / Child?
  - Belongs the person to the crew? / Is it a first-class passenger? …
- The task is to guess/estimate with the highest accuracy whether the person survived or not.

Decision Tree

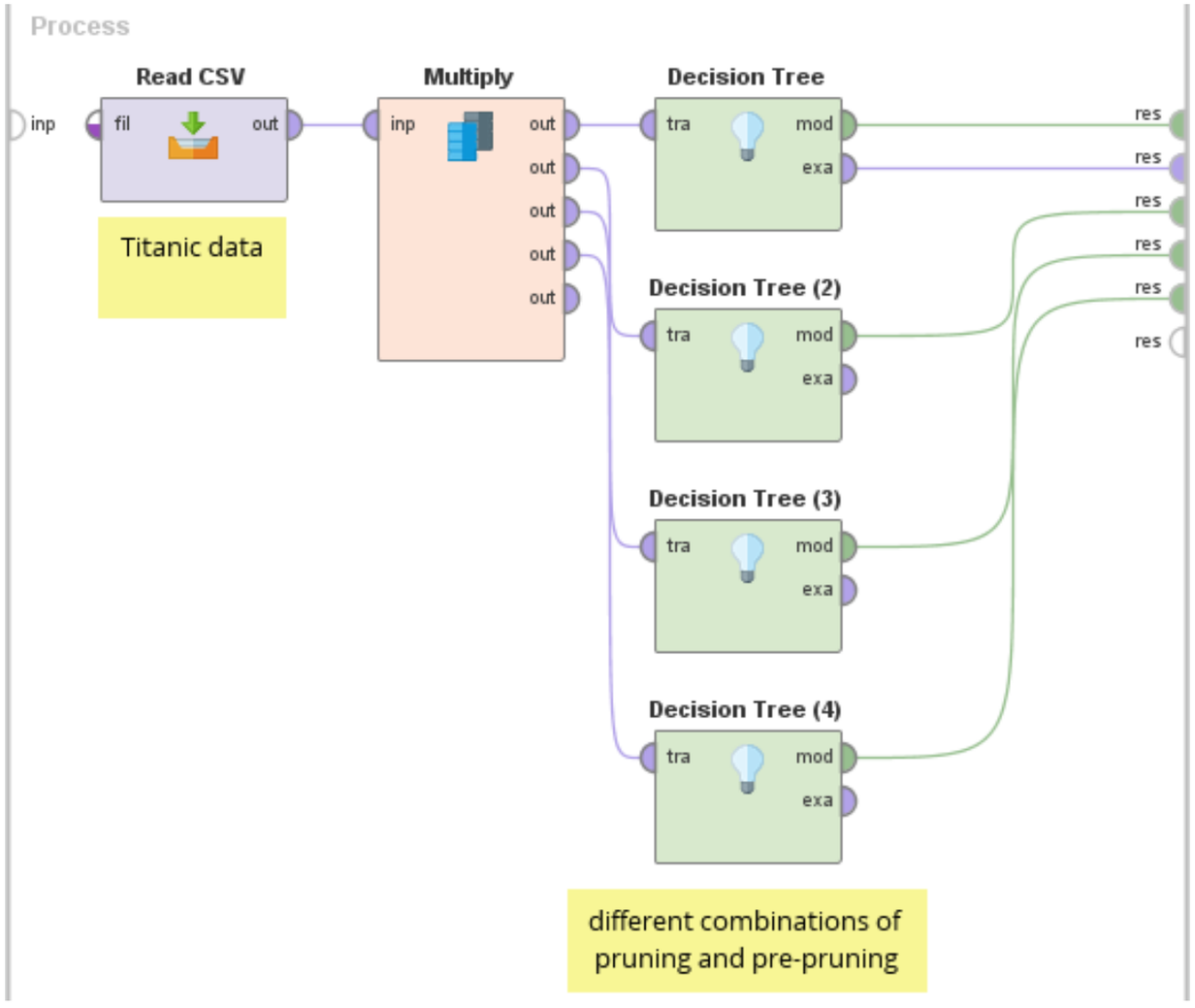
- Import the dataset using the Import Configuration Wizard.
- Set the "Survived“ attribute on the “label” flag.

ExampleSet (2201 examples, 1 special attribute, 3 regular attributes)

Row No.	Survived	Class	Age	Sex
1	yes	3rd	child	female
2	yes	crew	adult	male
3	no	3rd	adult	male
4	yes	crew	adult	male
5	no	3rd	adult	female
6	yes	2nd	adult	female
7	no	crew	adult	male
8	no	3rd	adult	male

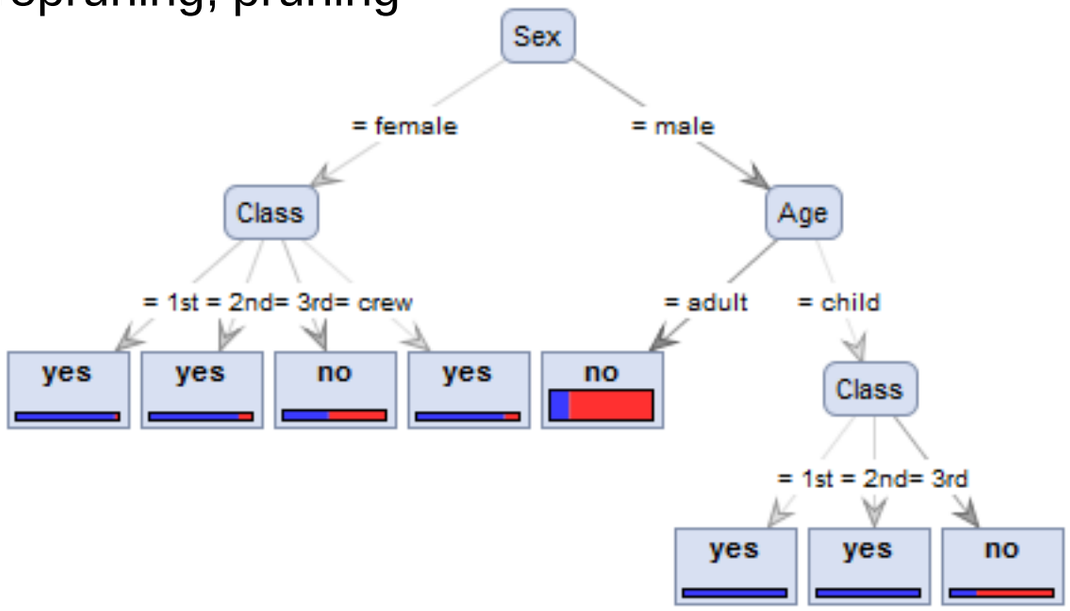
Decision Tree

- Build workflow according to the diagram.
  Use different combinations of pruning.

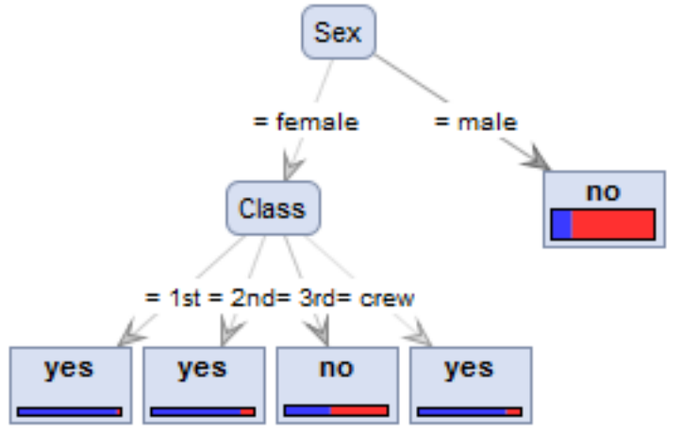


Decision Tree

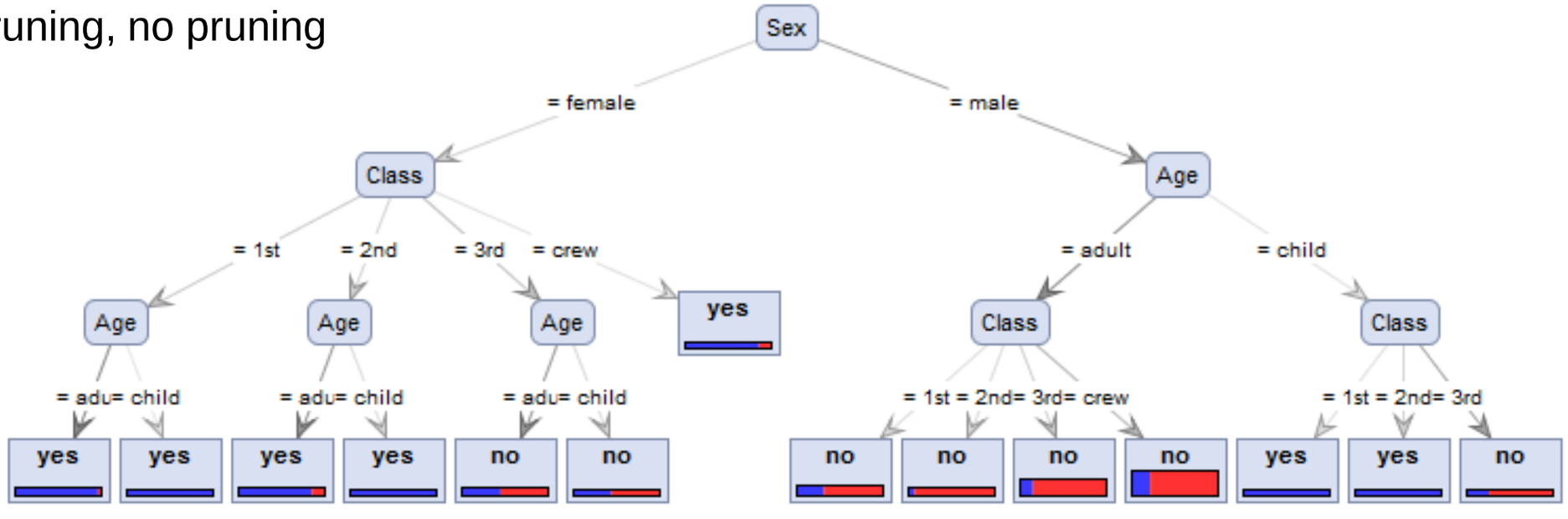
no prepruning, pruning



prepruning, pruning

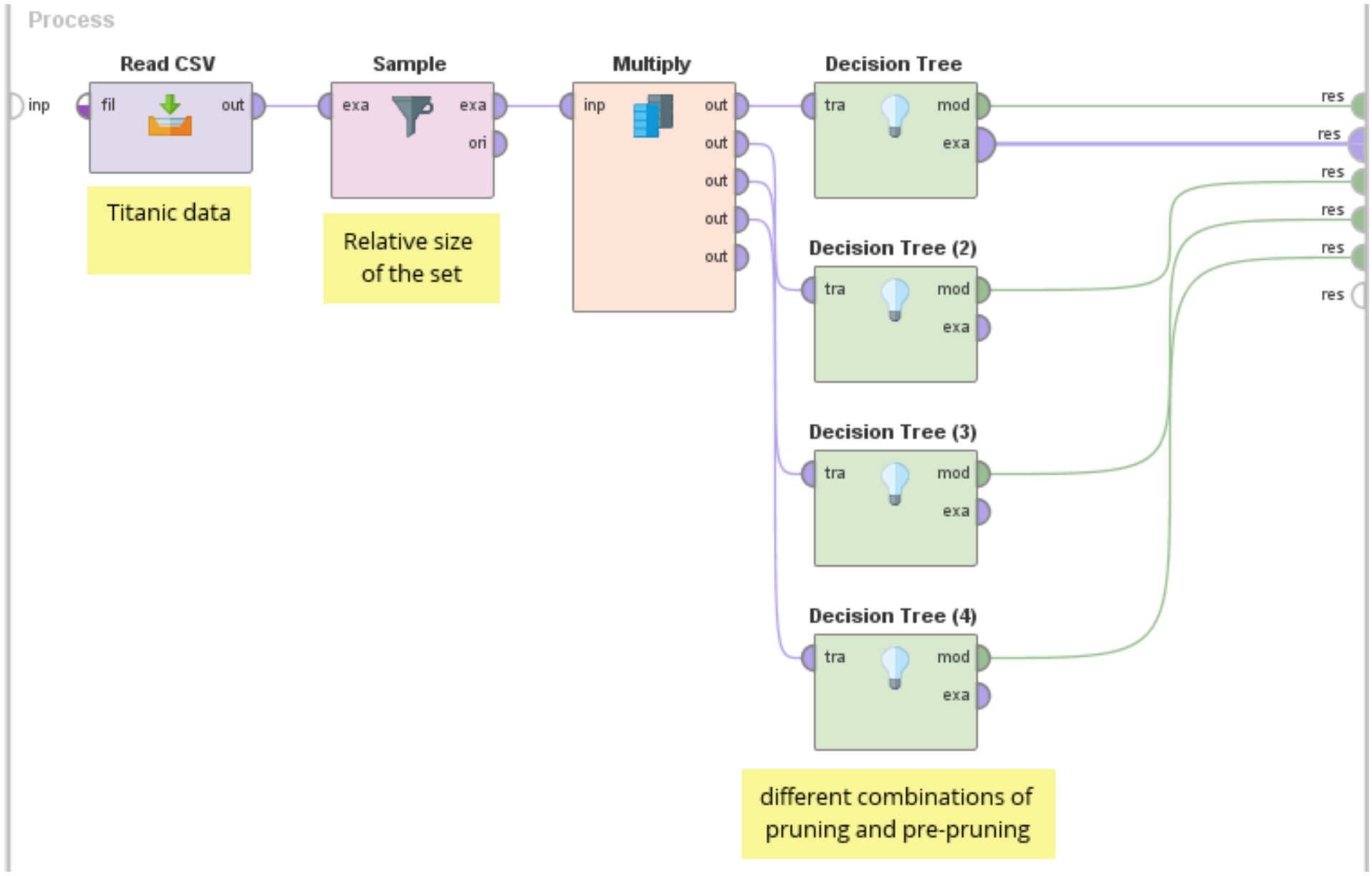


no prepruning, no pruning



The influence of the size of learning sample

- Add the “Sample” block and set it to the relative size.



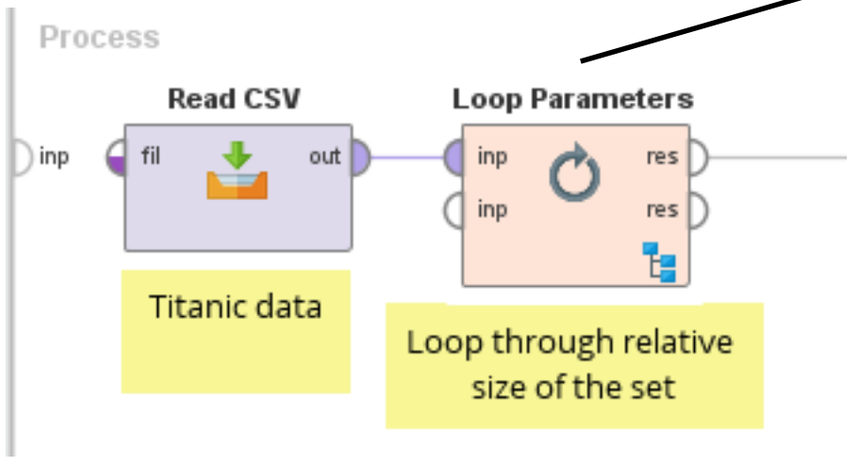
The influence of the size of learning sample

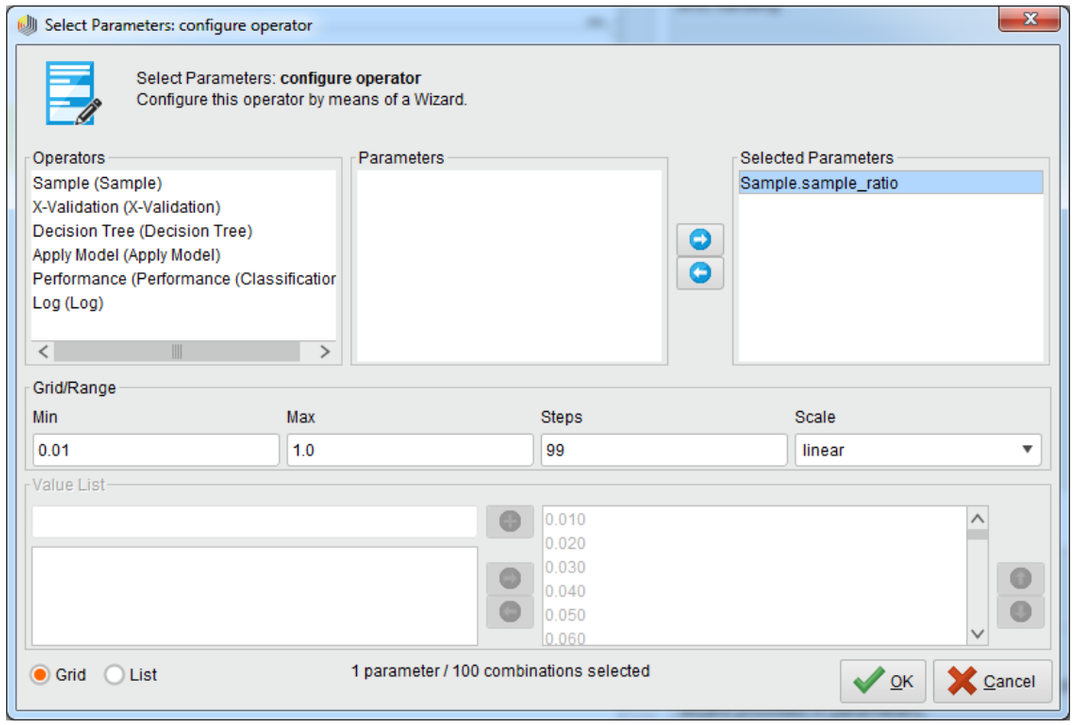
- Follow the effect of the sample size on the size of the tree.
  - Try different *split ratios* and record numbers of nodes and leaves of the created trees for different decision tree settings into a spredsheet (Excel) table.
  - Visualize the recorded data.

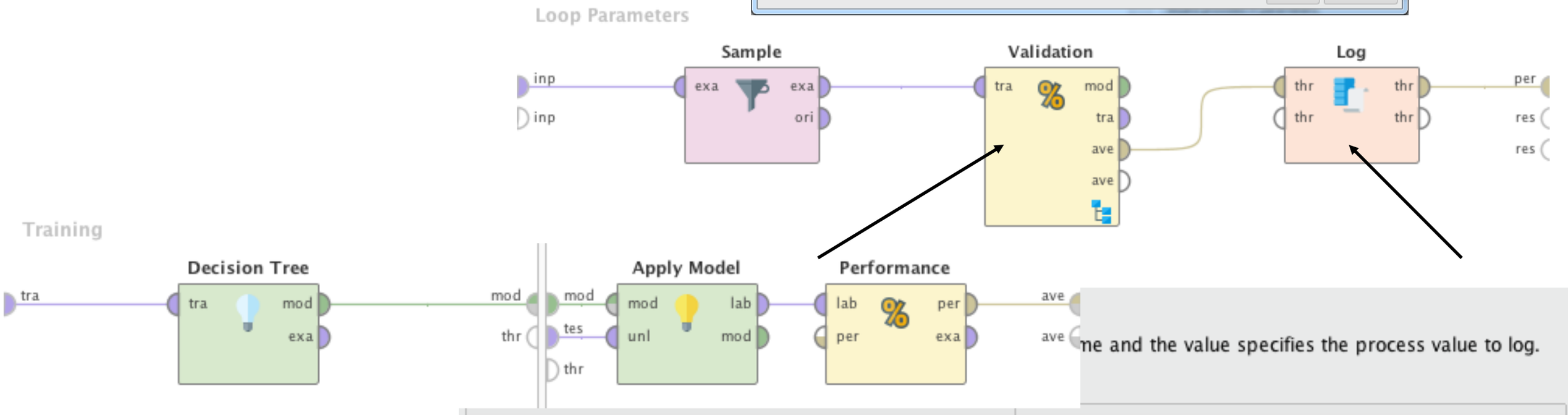
The influence of the size of learning sample

- Now, using the “Loop Parameters” and “X-Validation” determine the classification accuracy when sample ratio is between 0 and 1.

The influence of the size of learning sample



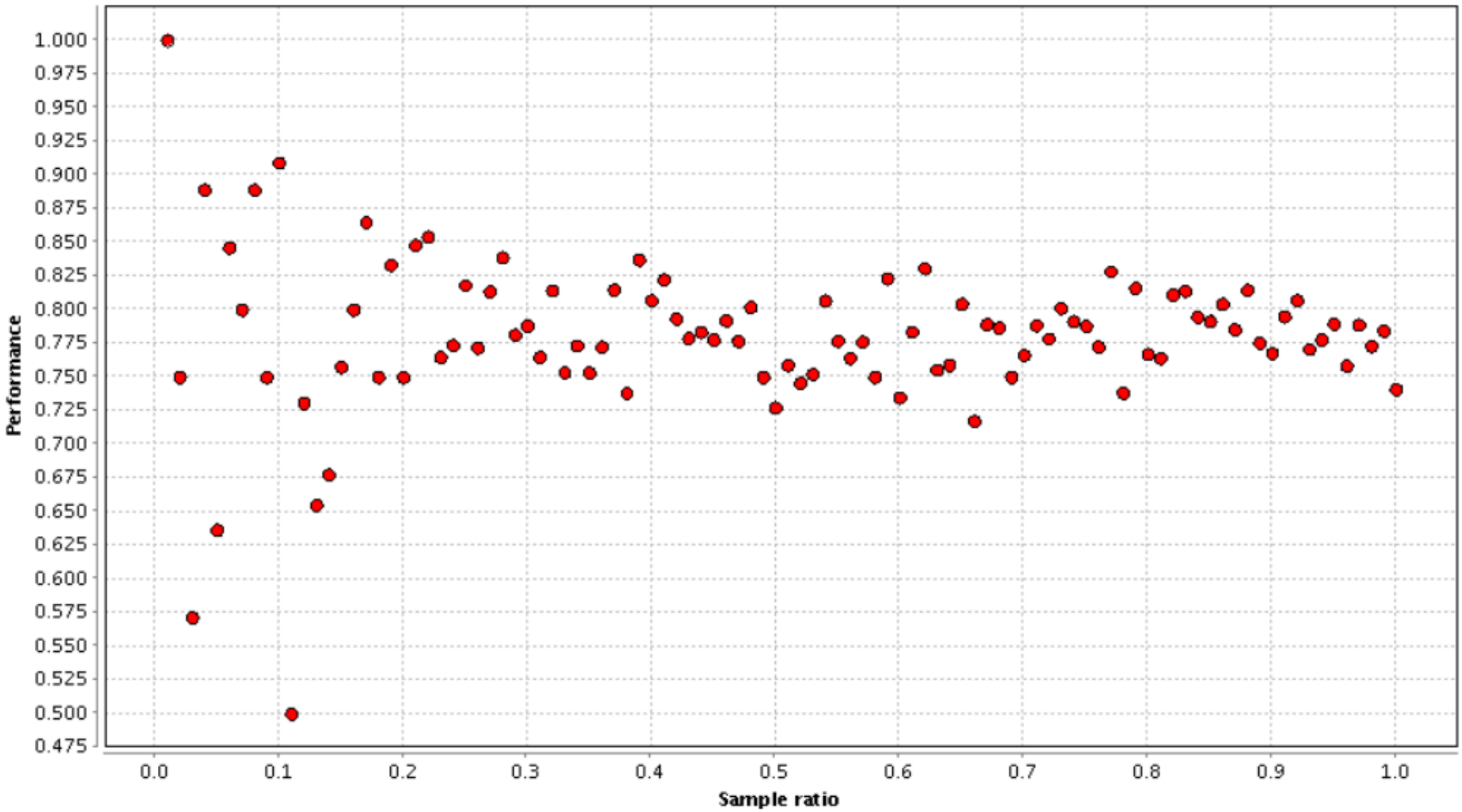




me and the value specifies the process value to log.

column name	value		
sample ratio	Sample	parameter	sample_ratio
performance	Performance	value	performance

The influence of the size of learning sample



The influence of the size of learning sample

- At the beginning of the classification the accuracy appears with high variance because easily or hardly classifiable samples can be chosen easily.

- However since some value the variance is lower and the accuracy is not increasing - the decision tree reached its limits.

Mushrooms

- Download from EDUX the dataset “mushrooms.zip”.

  The task is to determine whether a mushroom is **edible** or **poisonous**.

class	cap-shape	cap-surface	cap-color	bruises?	odor	gill-atachme	gill-spacing	gill-size	gill-color	stalk-shape	stalk-root	stalk-
polyno...	polyno...	polyno...	polyno...	polyno...	polyno...	polyno...	polyno...	polyno...	polyno...	polyno...	polyno...	polyn
label	attribute	attribute	attribute	attribute	attribute	attribute	attribute	attribute	attribute	attribute	attribute	attribu
p	x	s	n	t	p	f	c	n	k	e	e	s
e	x	s	y	t	a	f	c	b	k	e	c	s
e	b	s	w	t	l	f	c	b	n	e	c	s
p	x	y	w	t	p	f	c	n	n	e	e	s
e	x	s	g	f	n	f	w	b	k	t	e	s
e	x	y	y	t	a	f	c	b	n	e	c	s
e	b	s	w	t	a	f	c	b	g	e	c	s
e	b	y	w	t	l	f	c	b	n	e	c	s
p	x	y	w	t	p	f	c	n	p	e	e	s
e	b	s	y	t	a	f	c	b	g	e	c	s
e	x	y	y	t	l	f	c	b	g	e	c	s

Mushrooms - features

. 1. cap-shape: bell=b, conical=c, convex=x, flat=f, knobbed=k, sunken=s

. 2. cap-surface: fibrous=f, grooves=g, scaly=y, smooth=s

. 3. cap-color: brown=n, buff=b, cinnamon=c, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y

. 4. bruises?: bruises=t, no=f

. 5. odor: almond=a, anise=l, creosote=c, fishy=y, foul=f, musty=m, none=n, pungent=p, spicy=s

. 6. spore-print-color: black=k, brown=n, buff=b, chocolate=h, green=r, orange=o ,purple=u, white=w, yellow=y

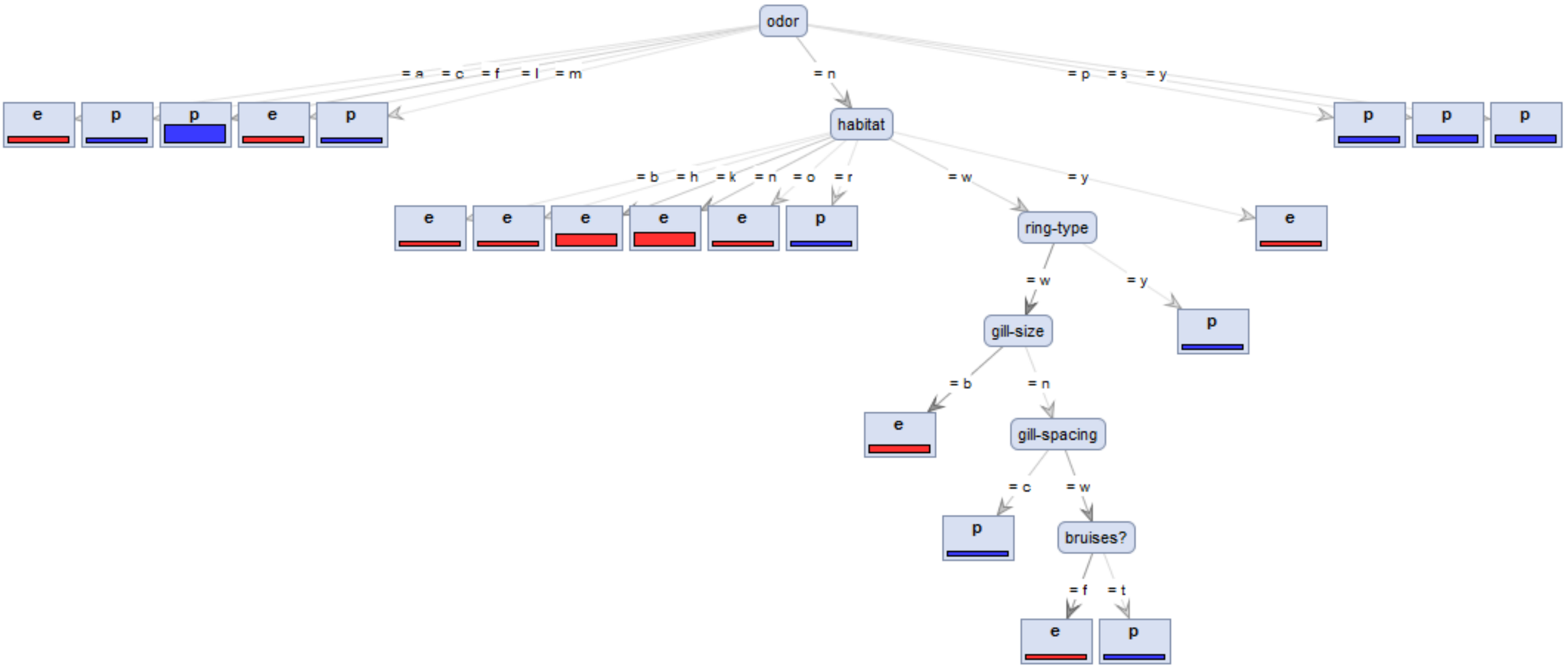
. 7. ...

Mushrooms

- Train a decision tree.
- Visualize its structure.
- Try to keep the tree for a man as comprehensible as possible while retaining a high success rate of classification.
- Use the x-validation to determine the success of the classification for the given settings of parameters.

Mushrooms

no prepruning, no pruning



Mushrooms

Accuracy

accuracy: 98.52% +/- 0.23% (mikro: 98.52%)

	true p	true e	class precision
pred. p	3796	0	100.00%
pred. e	120	4208	97.23%
class recall	96.94%	100.00%	

Mushrooms

- Use the additional measures for tree splitting (information gain, Gini index, accuracy).
- How differs the classification accuracies?
- How differs the trees?

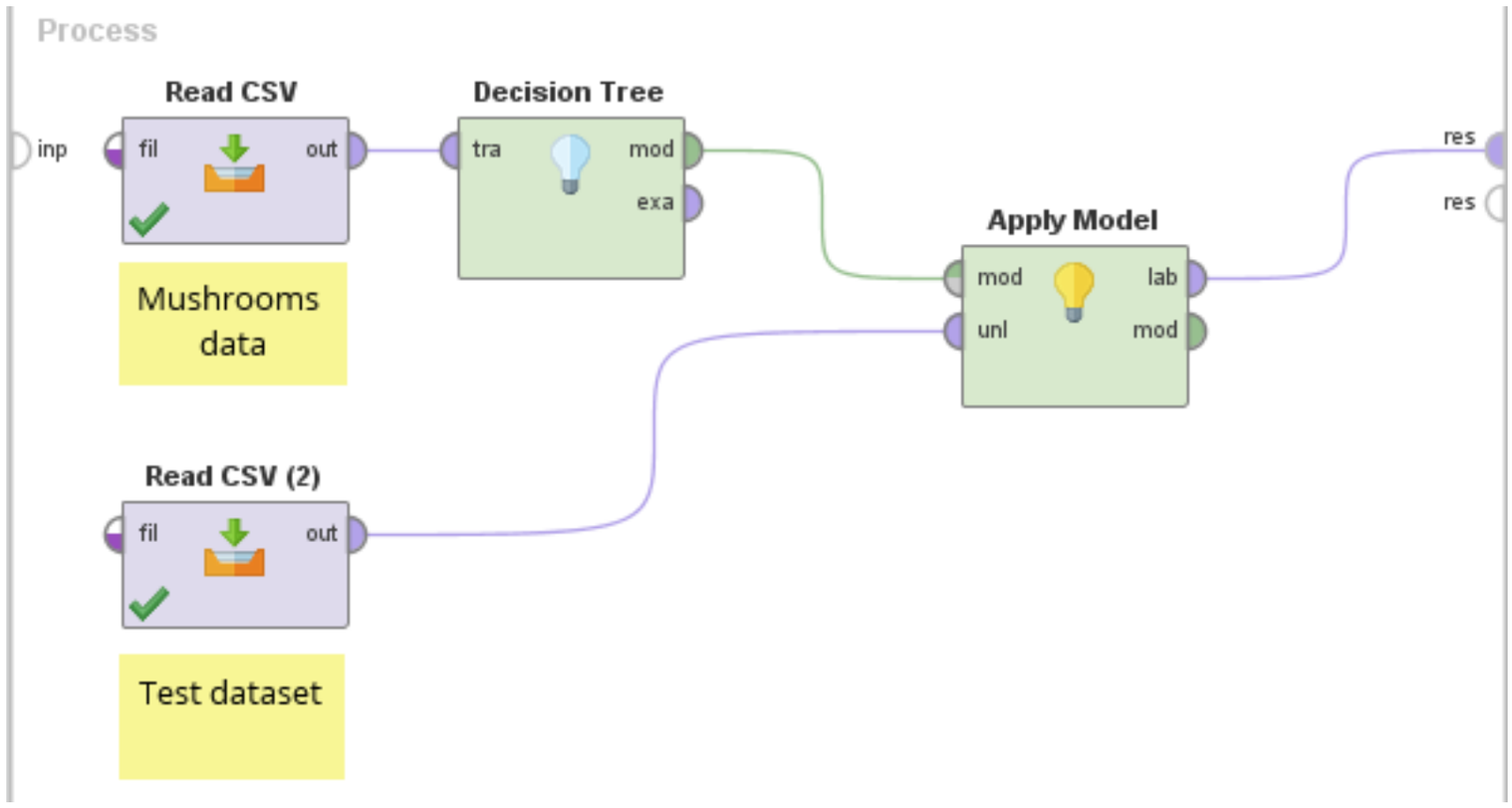
Mushrooms

- Use the best model for the classification of unknown mushrooms in the “agaricus-lepiota.test” dataset.

- Are all mushrooms edible?

Mushrooms

- Test



Mushrooms

- Mushrooms – results on the test dataset

ExampleSet (3 examples, 4 special attributes, 20 regular attributes)

Row No.	class	prediction(cl...	confidence(p)	confidence(e)	cap-shape	cap-surface	cap-color	bruises?
1	?	e	0	1	f	y	n	t
2	?	e	0	1	x	s	n	f
3	?	p	1	0	k	s	b	t

Cars

- Use “cars.csv” dataset from the last exercise
- Experiment with different algorithms for creating decision trees, with their parameters and with different techniques of data discretization.
- Generate a decision tree, which is the most comprehensible and simultaneously has a high success rate on test data.