An Introduction to Neural Networks
-
Hands-On with RapidMiner Studio

Agathe Merceron
Beuth University of Applied Sciences
Berlin, Germany

Agenda

- Data MIB-Students_en.csv
- Operator Neural Net process1.rmp
- Cross Validation process2.rmp
- Models Comparison process3.rmp
- Parameter Optimization process4.rmp

Data

- Results from 1st year Students of one degree program of a university in Germany
- Data from fall 2005 till summer 2018; stand preprocessing February 2020

Number_Enrollments_1	Number_Courses_Passed_1	Average_Mark_1	class
5	0	0	dropout
4	3	2.8	dropout
5	5	3.47	dropout
5	3	2.75	graduate
5	5	3.45	graduate
5	4	3.62	graduate

Data

- Similar (but bigger) dataset in Wagner, K., Merceron, A. & Sauer, P., (2020). *Accuracy of a Cross-Program Model for Dropout Prediction in Higher Education*.
  In Companion Proceedings of the 10th Learning Analytics and Knowledge Conference (LAK'20). Workshop on Addressing Dropout Rates in Higher Education, Frankfurt am Main, Germany, 2020, 744-749.

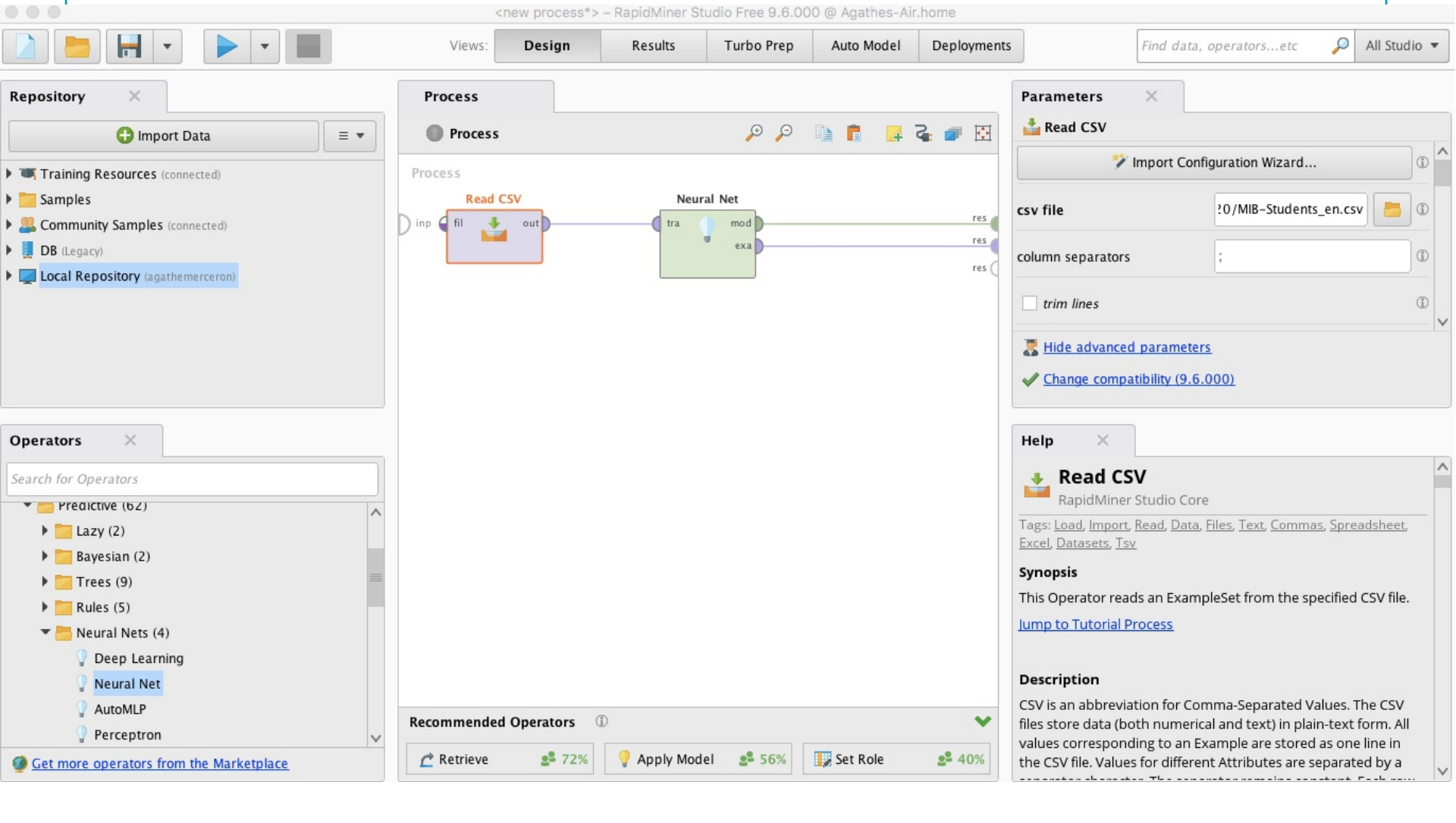
Operator Neural Net – Process Overview

- Import process1Path.rmp

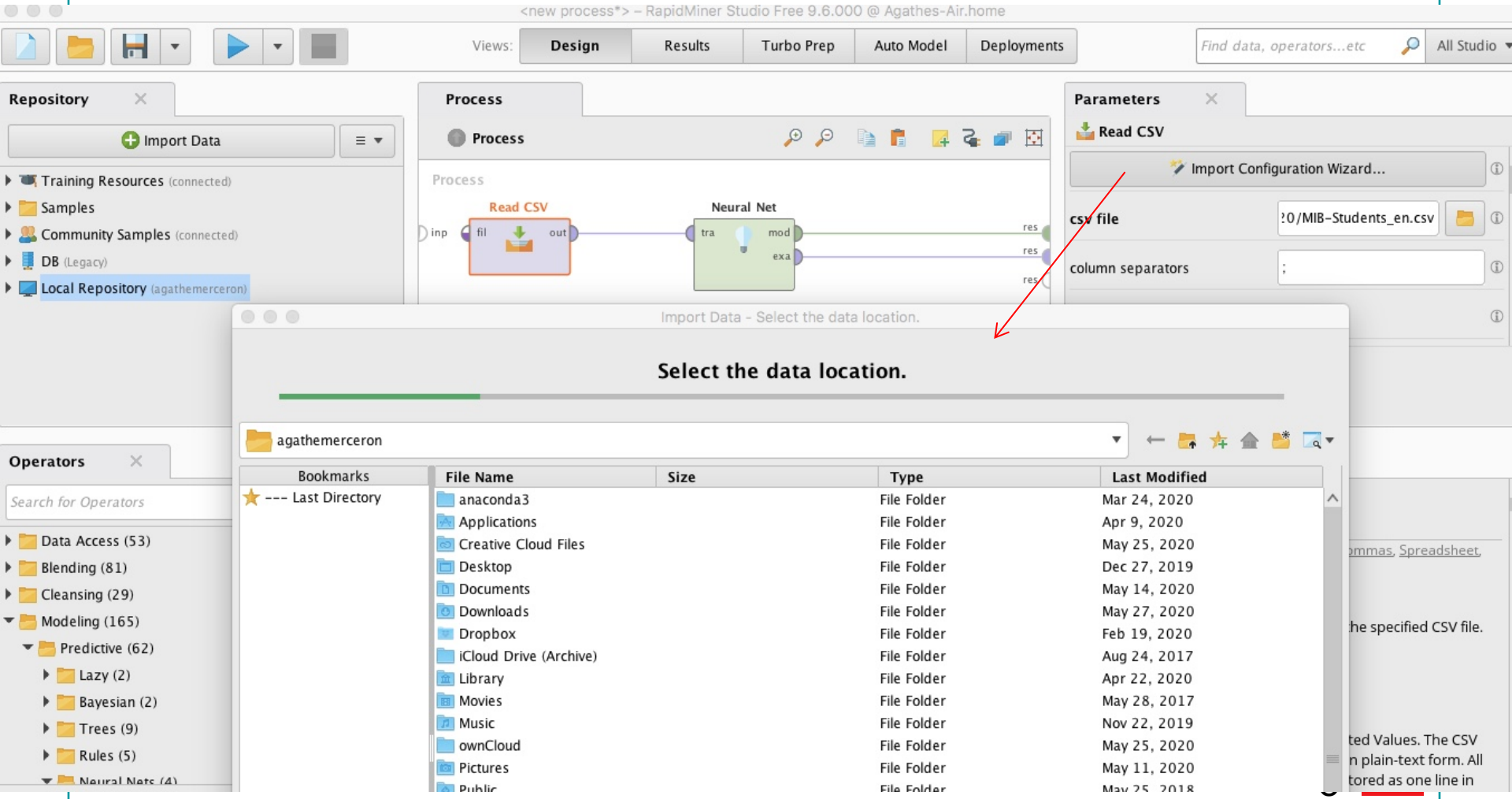
Operator Neural Net – Process Overview

- Read the data with the CSV Operator from Data Access – the type of the class attribute should be set to *binomial* and the role to *label*.
- Attach the Neural Net Operator from Modelling / Predictive / Neural Nets.
- Run the process.
- Explore visually the data.
- Inspect the neural network.

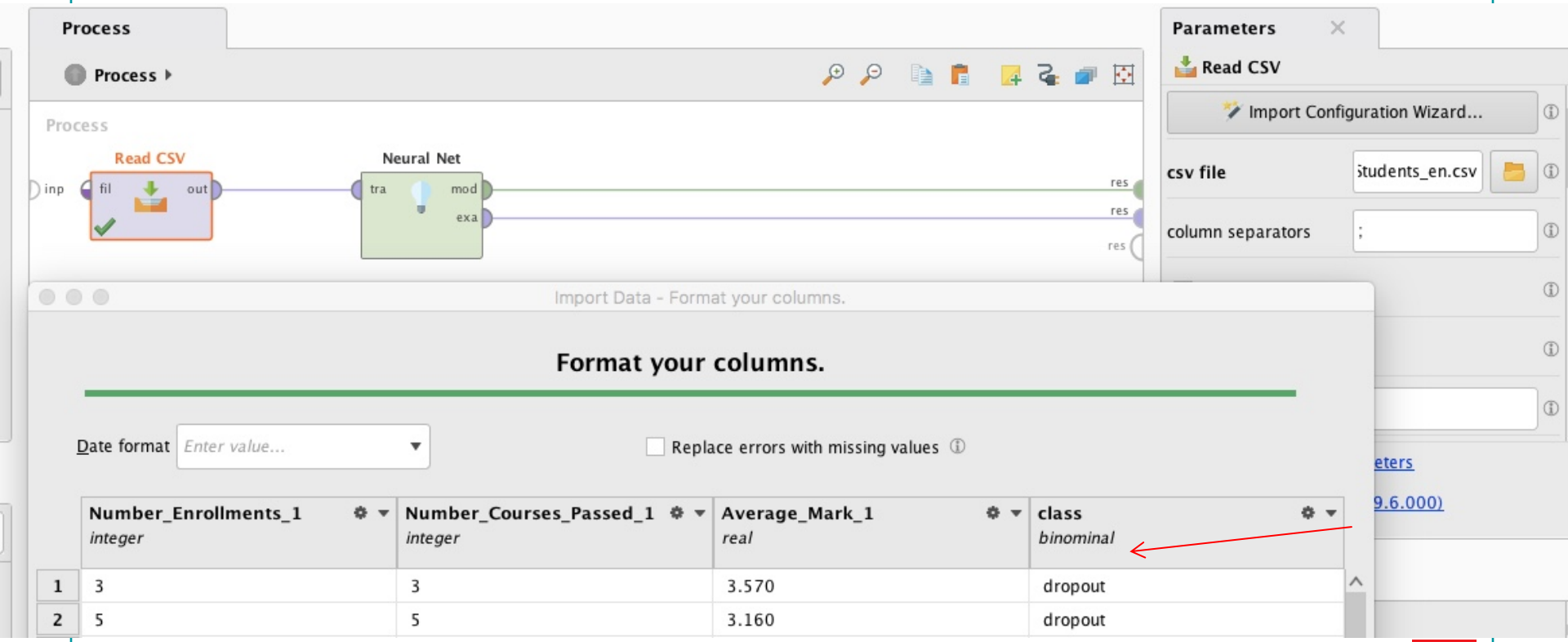
Operator Neural Net - Process Overview



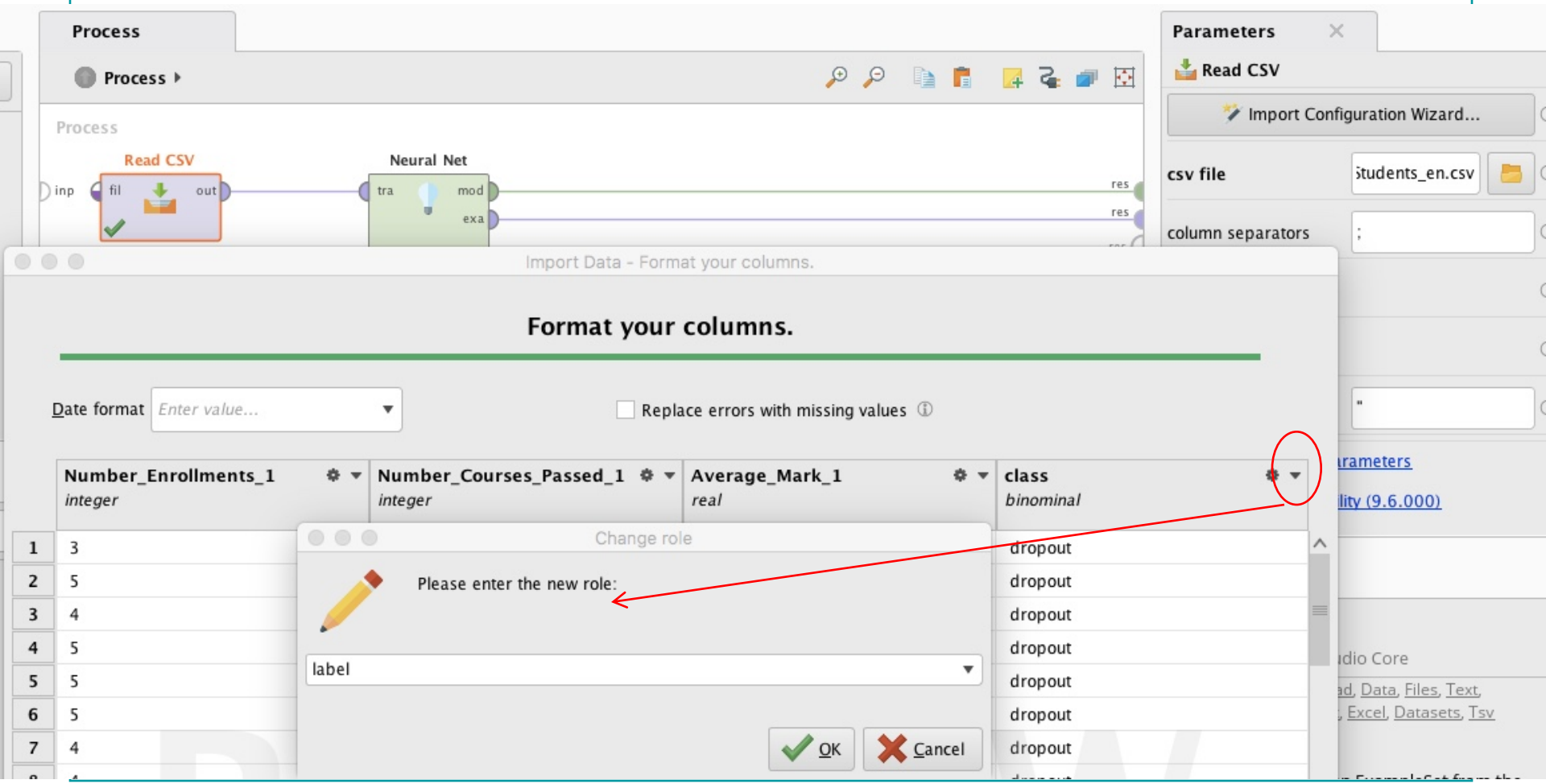
Operator Neural Net – Data Access



Operator Neural Net – Data Access



Operator Neural Net – Role class

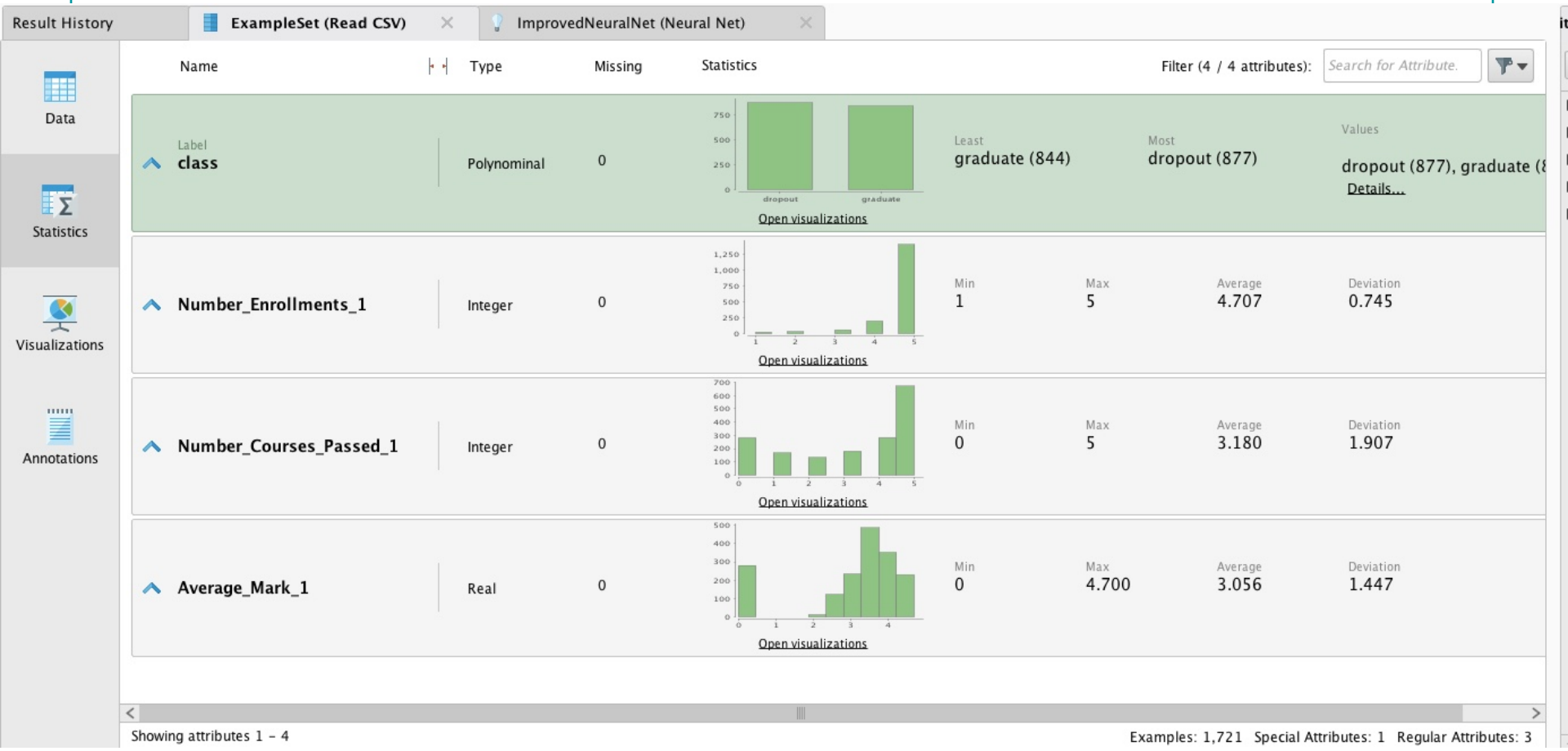


Operator Neural Net – Exploring the Data

- Hit the blue arrow at the top to execute the process.
- Press the Results-Button at the top, below select the tab ExampleSet, select Statistics on the left.
- Select Visualizations / Plot Type Box Plot.
- Try other plots as you like.

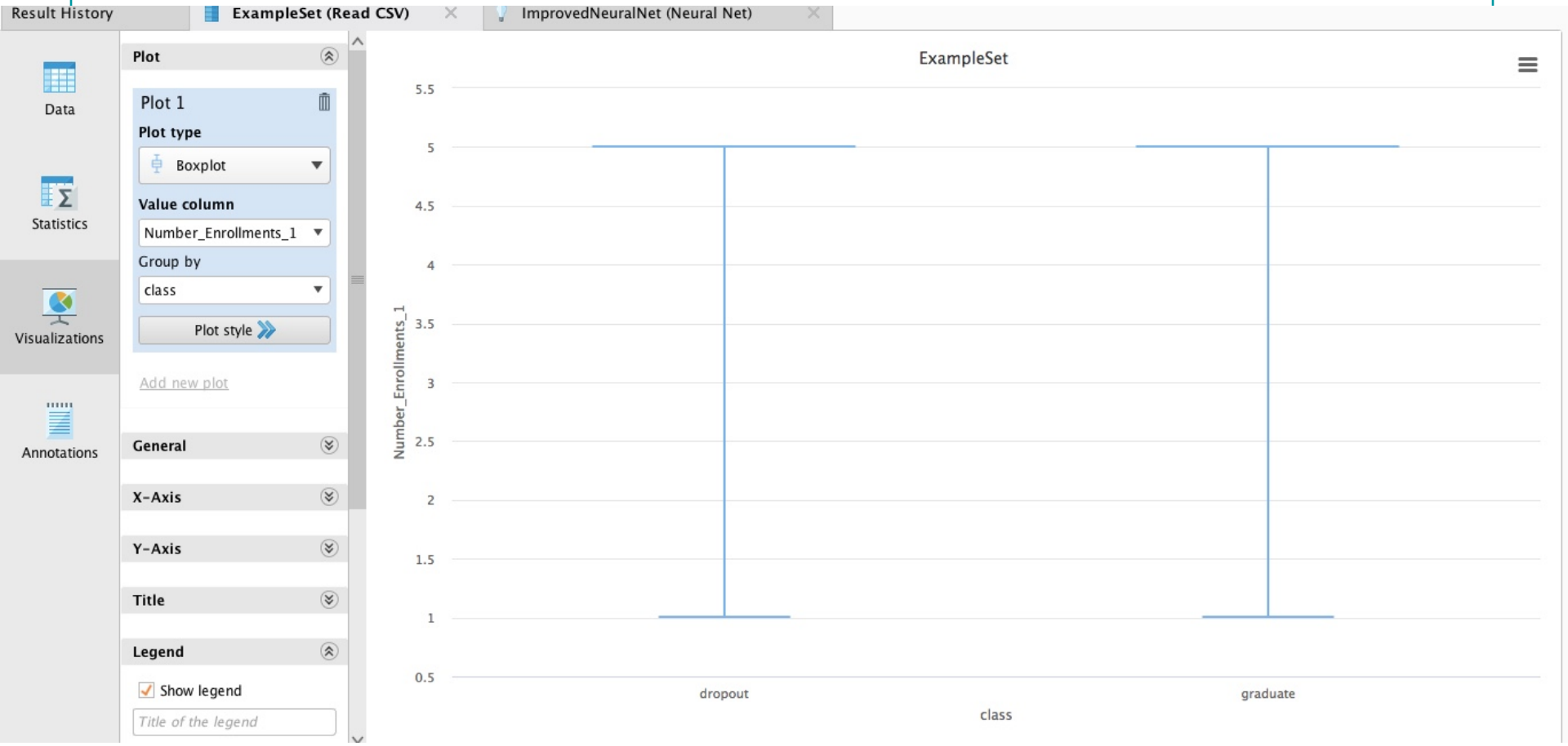
Operator Neural Net – Exploring the Data

- Large number of the class dropout due to data filtering.



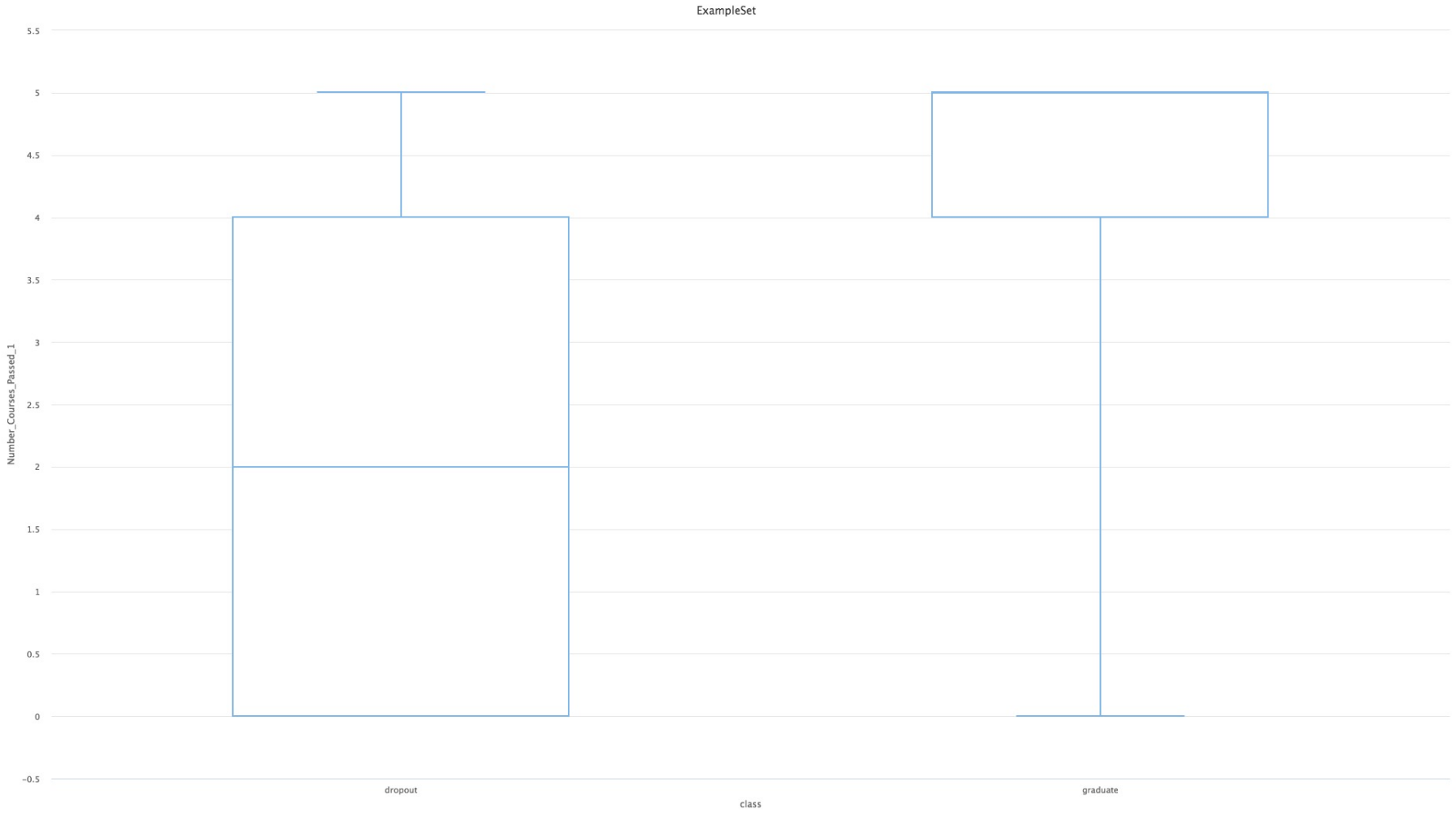
Operator Neural Net – Exploring the Data

- Box Plot Number_Enrollments



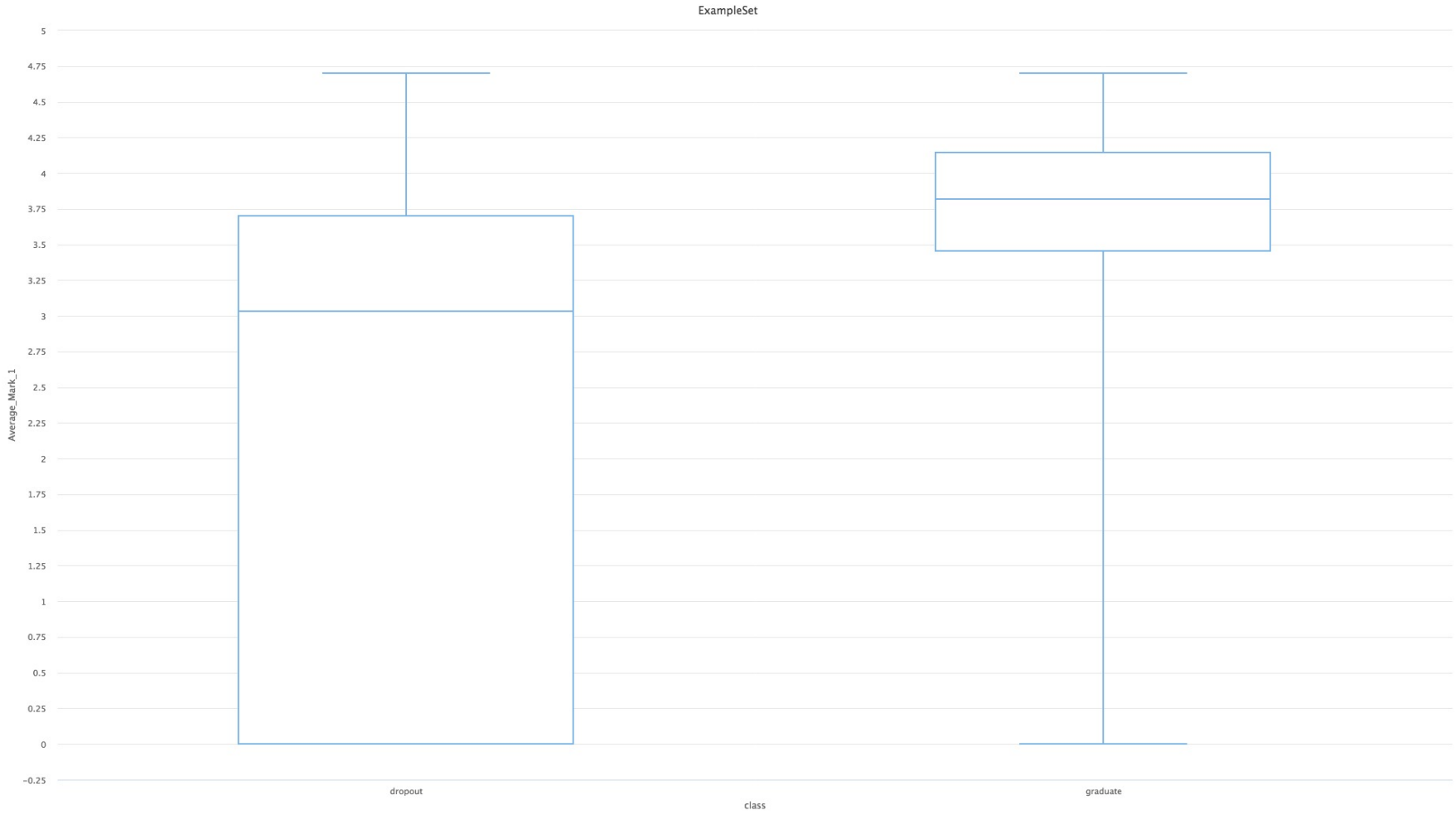
Operator Neural Net – Exploring the Data

- Box Plot Number_Courses_Passed



Operator Neural Net – Exploring the Data

- Box Plot Average_Mark

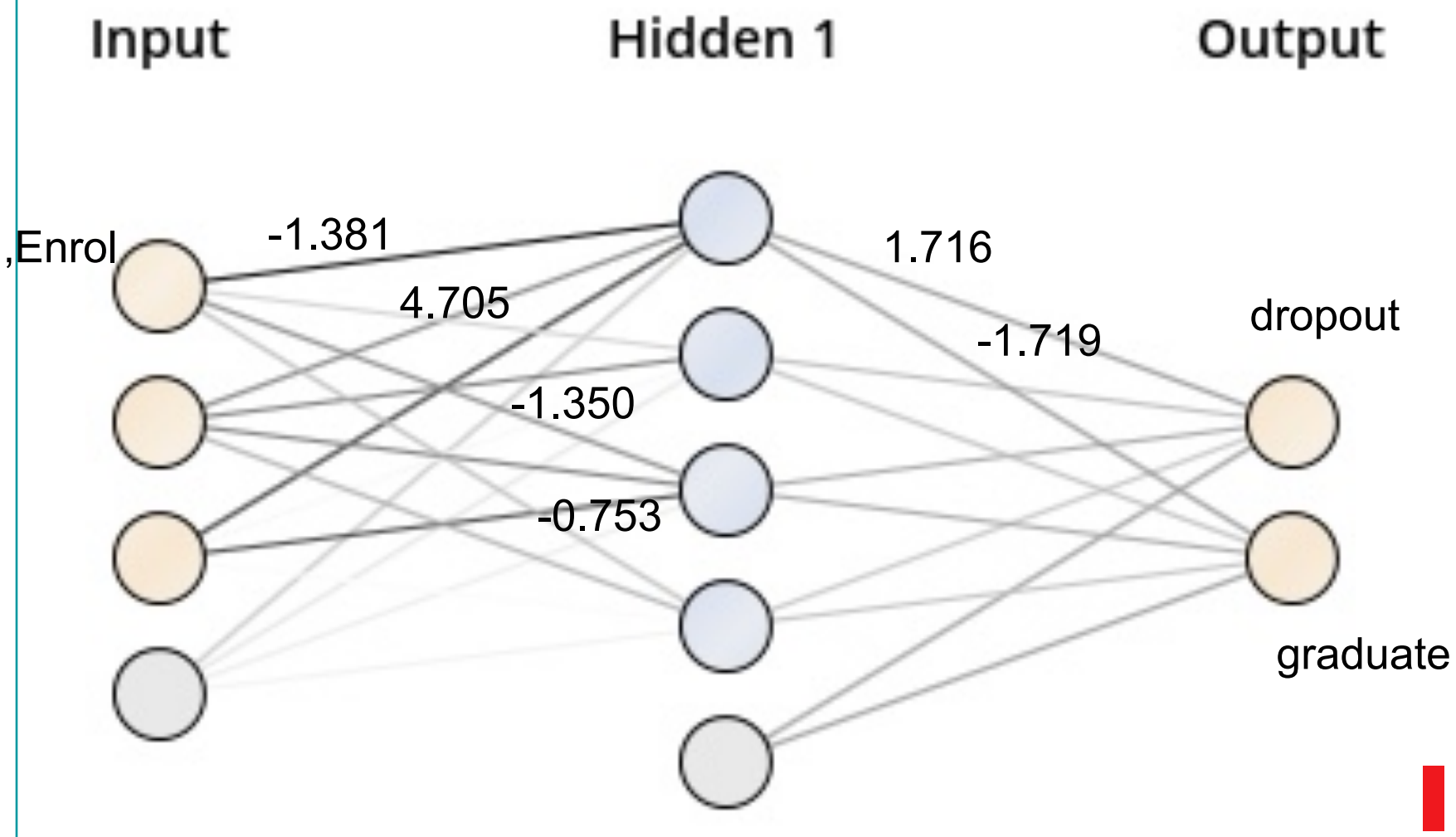


Operator Neural Net – Exploring the Data

- Students labelled “graduate” have a distinct average mark and a distinct number of passed courses from those labelled “dropout”.

- Is the feature Number_Enrollments helpful to predict graduate versus dropout?

Operator Neural Net – Neural Net



Operator neural Net – Neural Net

- All Weights are given in RapidMiner under the tab ImprovedNeuralNet and Description on the left.

- How good is this neural net?

K-Cross Validation

- The dataset is split into a training set and a testing set.
- The model is trained on the training set and evaluated with various measures on the test set.
- This is repeated k times; the average and standard deviation of the measures are returned.

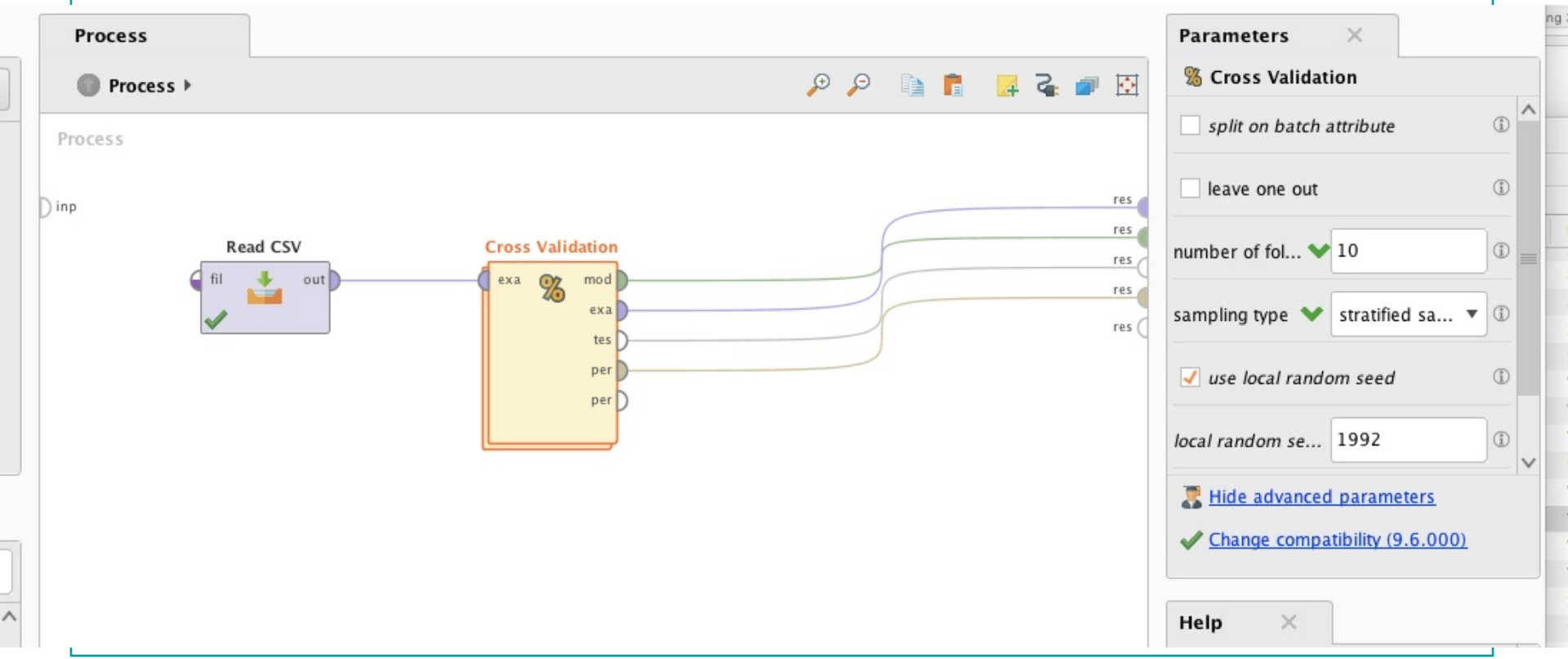
- Import process2Path.rmp

K-Cross Validation

- Import process2Path.rmp
- Read CSV: see 1st process.
- Change K, number of folds in Cross Validation, if you wish.
- Double Click Cross Validation.
- Choose your favorite measures in Performance.
  The main criterion does not play any role in the present setting.
- Choose a place and a file name in Log to store the results obtained in each fold. Edit List to choose the measures you want to store.
- Explore AutoMLP (does not perform better here).

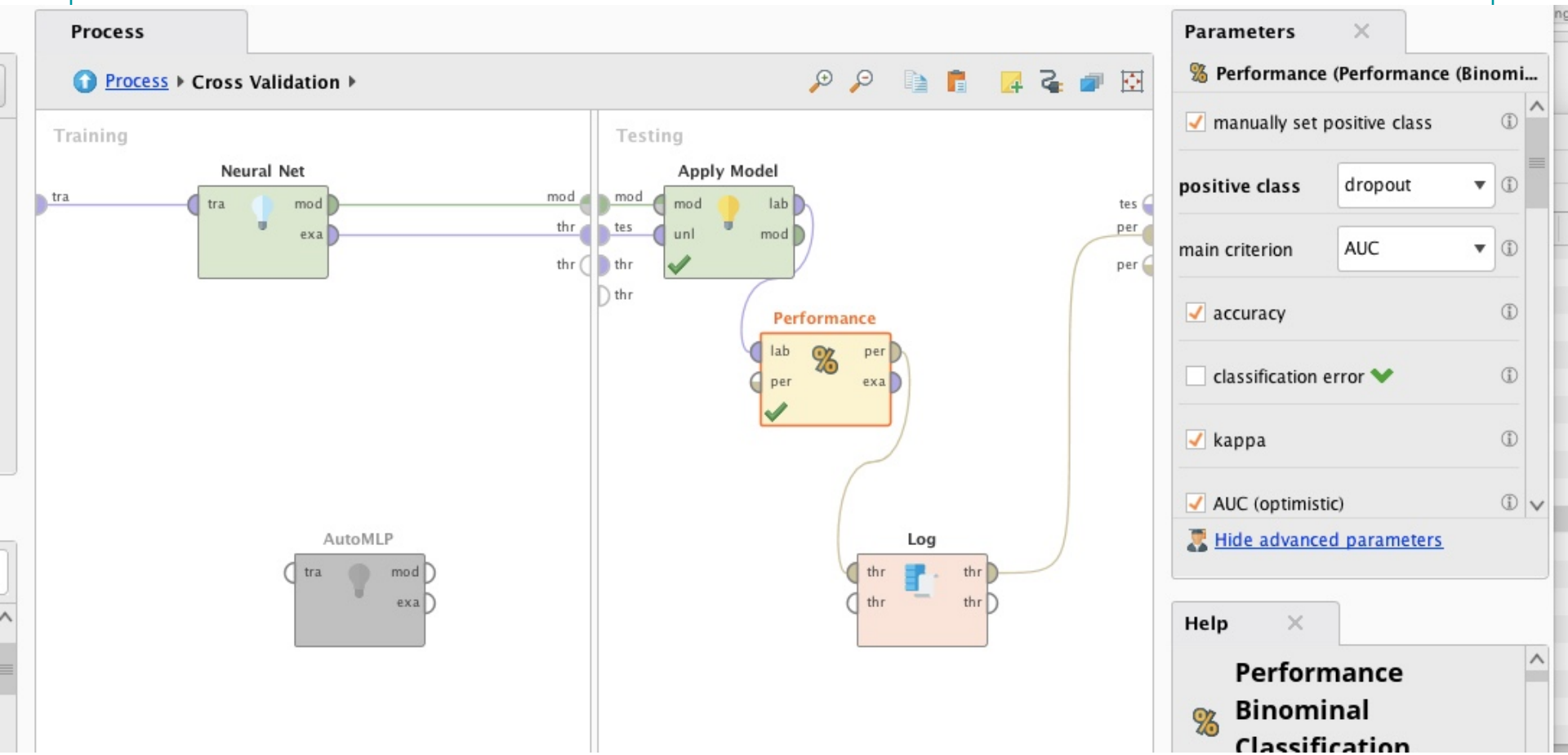
Cross Validation

- Cross Validation is a nested operator.



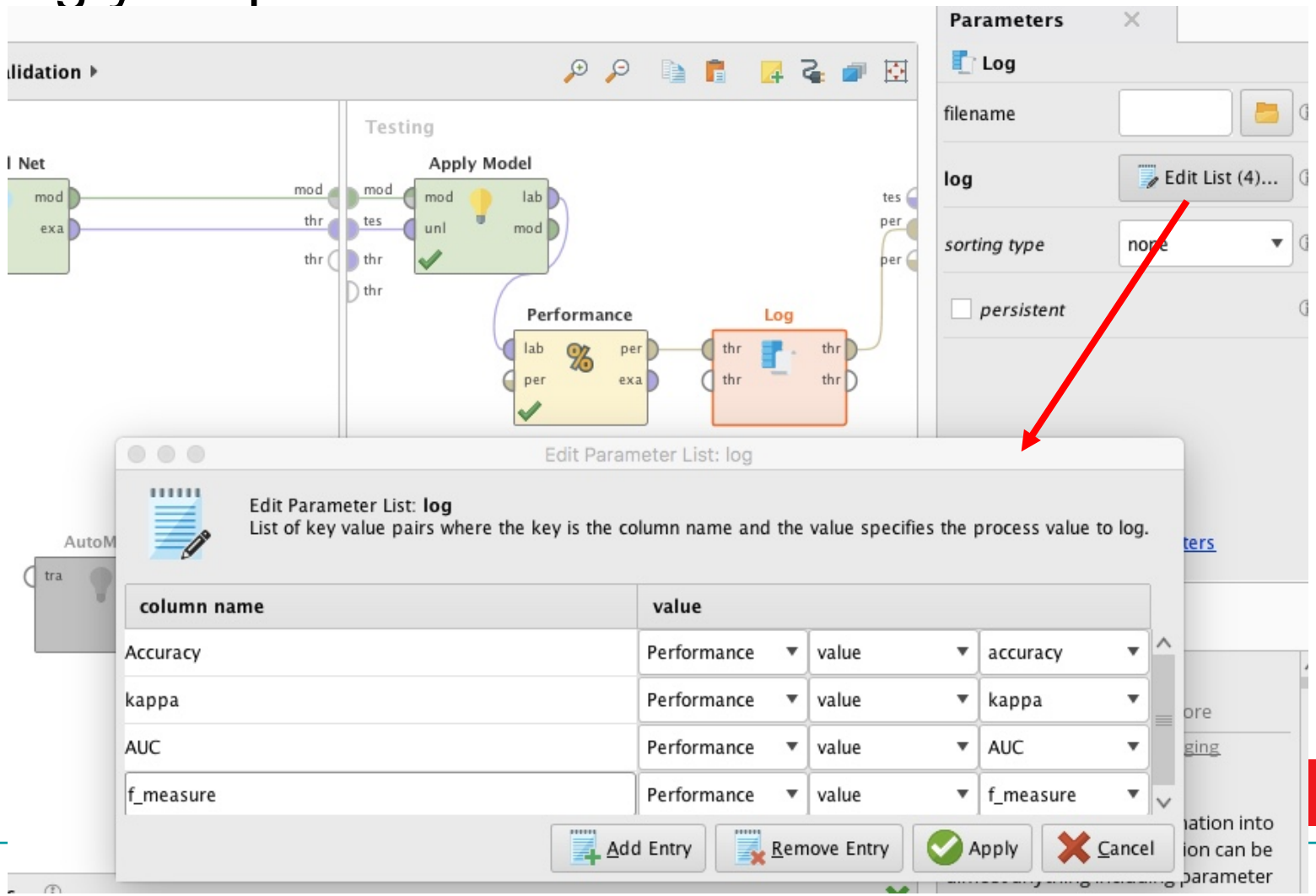
Cross Validation

- Choose Performance Operator for binomial classification and set dropout as positive class.



Cross Validation

- Log your prefered measures.



Cross Validation

- Results:

	true dropout	*true graduate*	*class precision*
pred. dropout	642	115	84.81%
pred. graduate	235	729	75.62%
class recall	73.20%	86.37%	

Cross Validation

Results:

- accuracy: 79.66% +/- 2.83%
- AUC: 0.866 +/- 0.032
- Precision: 85.15% +/- 4.61
- recall: 73.19% +/- 6.10%
- f_measure: 78.48% +/- 3.56%

Cross Validation

Results:

- Log file sorted on accuracy: except recall, all measures evolve almost the same.

Acc.	*Kappa*	*AUC*	*F1*	*Recall*
0.843	0.688	0.896	0.834	0.772
0.825	0.652	0.877	0.810	0.727
0.813	0.628	0.904	0.8	0.735
0.808	0.616	0.881	0.811	0.806

Cross Validation

- The log file gives the measures for each fold. Copy the table in a sheet to have results with more decimal numbers.

- Is a Neural Net better than other models?

Compare ROCs Operator

- Import process3.rmp

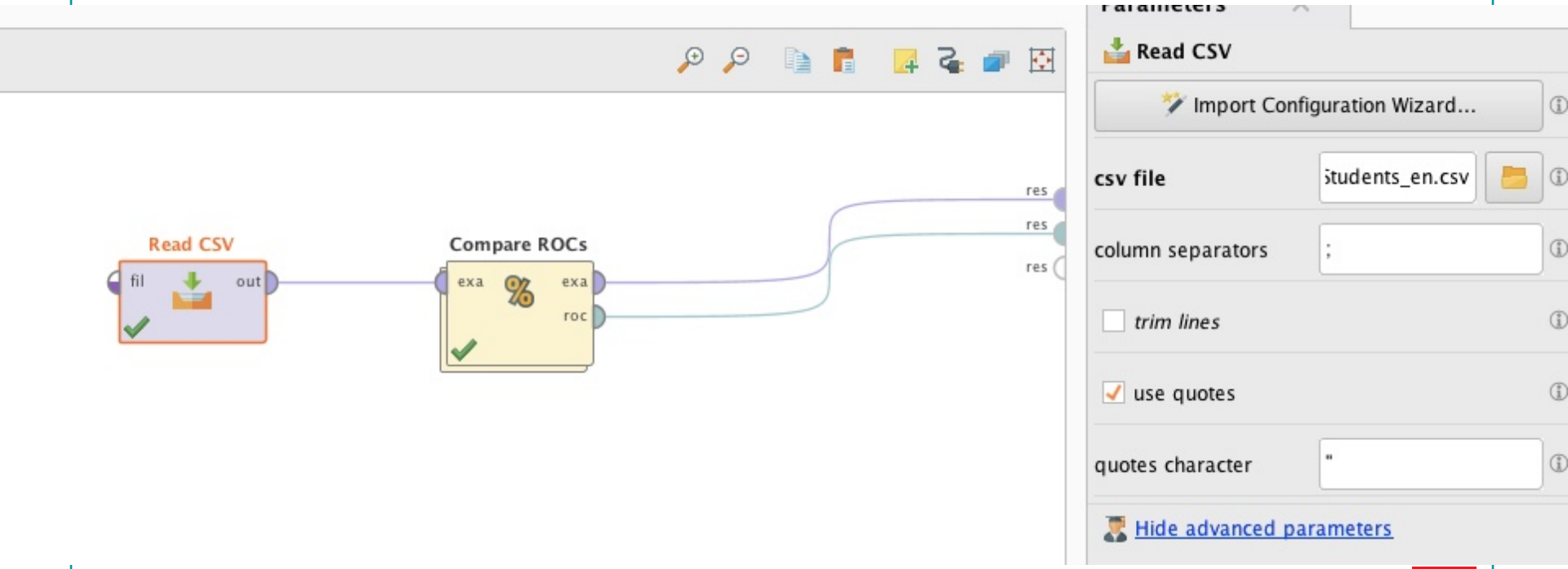
- The Compare ROCs operator performs a cross validation of the algorithms that one has selected and returns a single graph with the ROC curves of the algorithms.

Compare ROCs Operator

- Import process3Path.rmp
- Read CSV: see 1st process.
- Change the number of folds in Compare ROCs, if you wish.
- Double Click Compare ROCs.
- Feel free to change the algorithms! In Operators (left from Process) Modelling > Predictive choose the classificators you like.

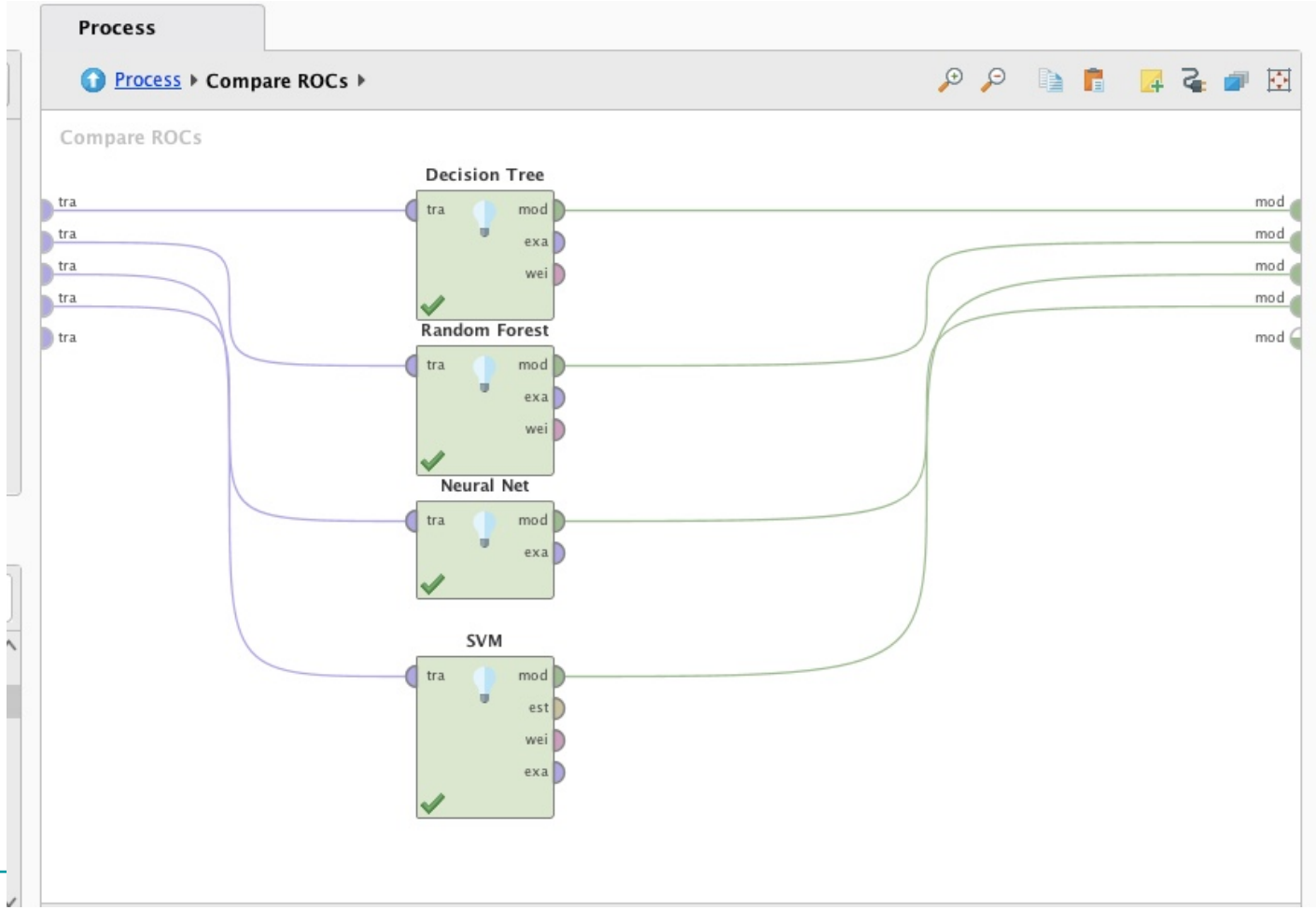
Compare ROCs Operator

- Nested Operator – Performs a Cross Validation



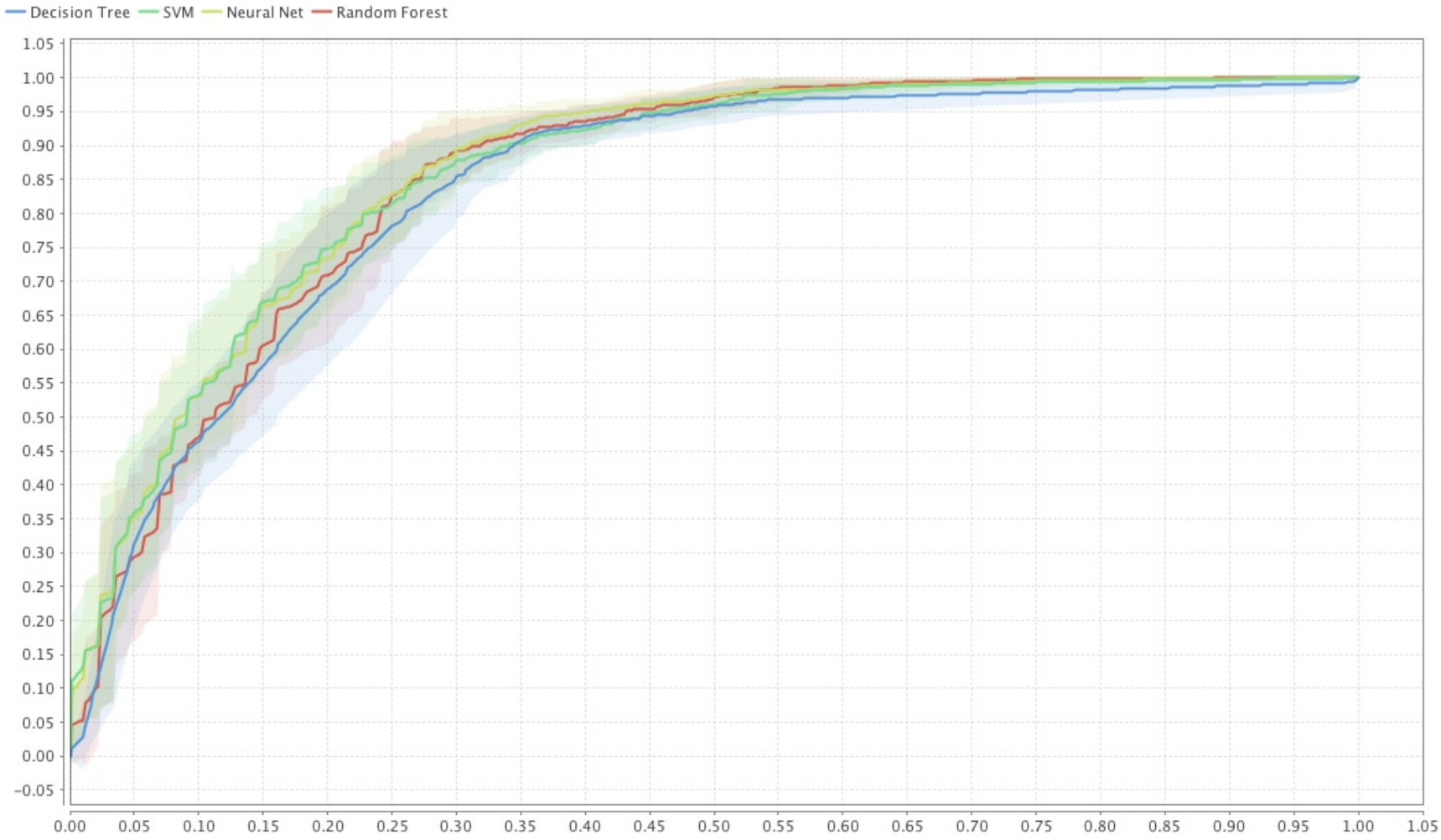
Compare ROCs Operator

- Pick the models to compare



Compare ROCs Operator

- SVM and Neural Net have the best curves.



Parameter Optimization – Process Overview

- Are the default values for learning rate, momentum, number of hidden layers, number of units etc. the best values?

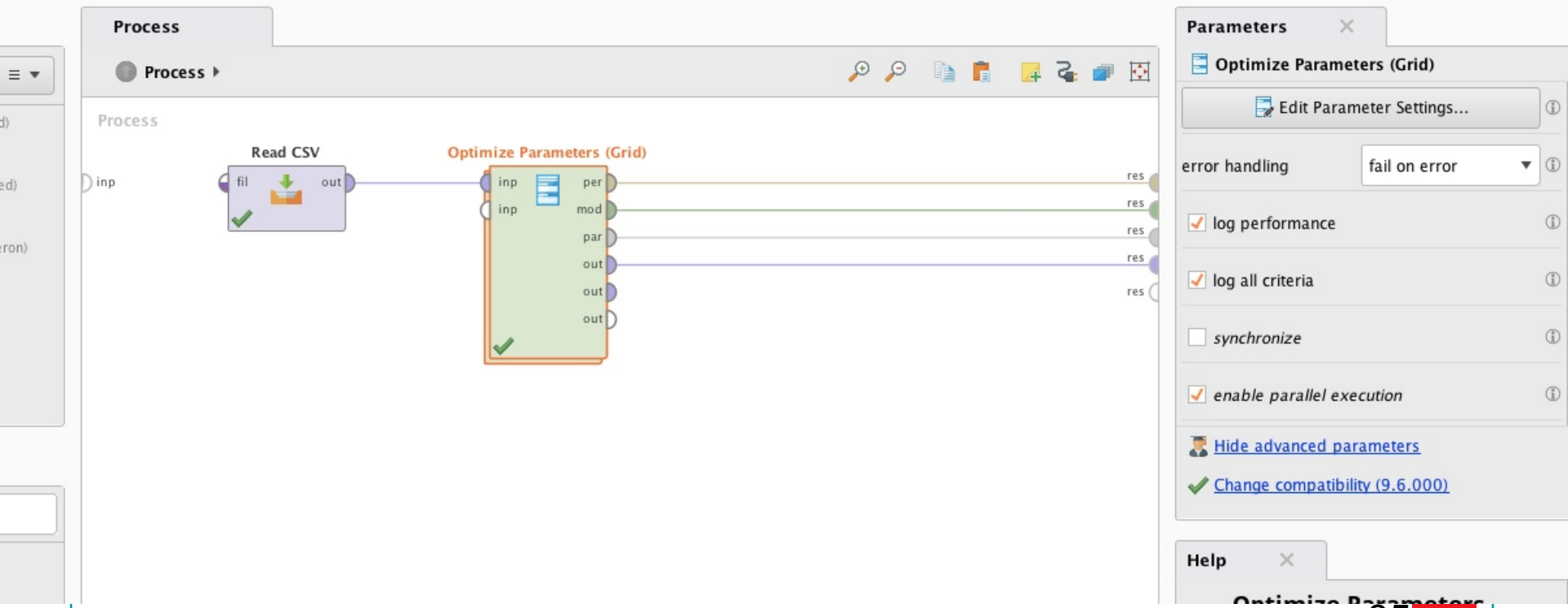
- A Grid Search varies those parameters and returns the optimal ones. Note: Computation time might be (very) high.
- Import process4.rmp

Parameter Optimization – Process Overview

- Import process4Path.rmp
- Read CSV: see 1st process.
- Edit Parameters Settings and choose the hyperparameters you want to vary (if you don‘t have much time, choose only one parameter).
- Double Clicking Optimize Parameters (Grid) leads you to Cross Validation.

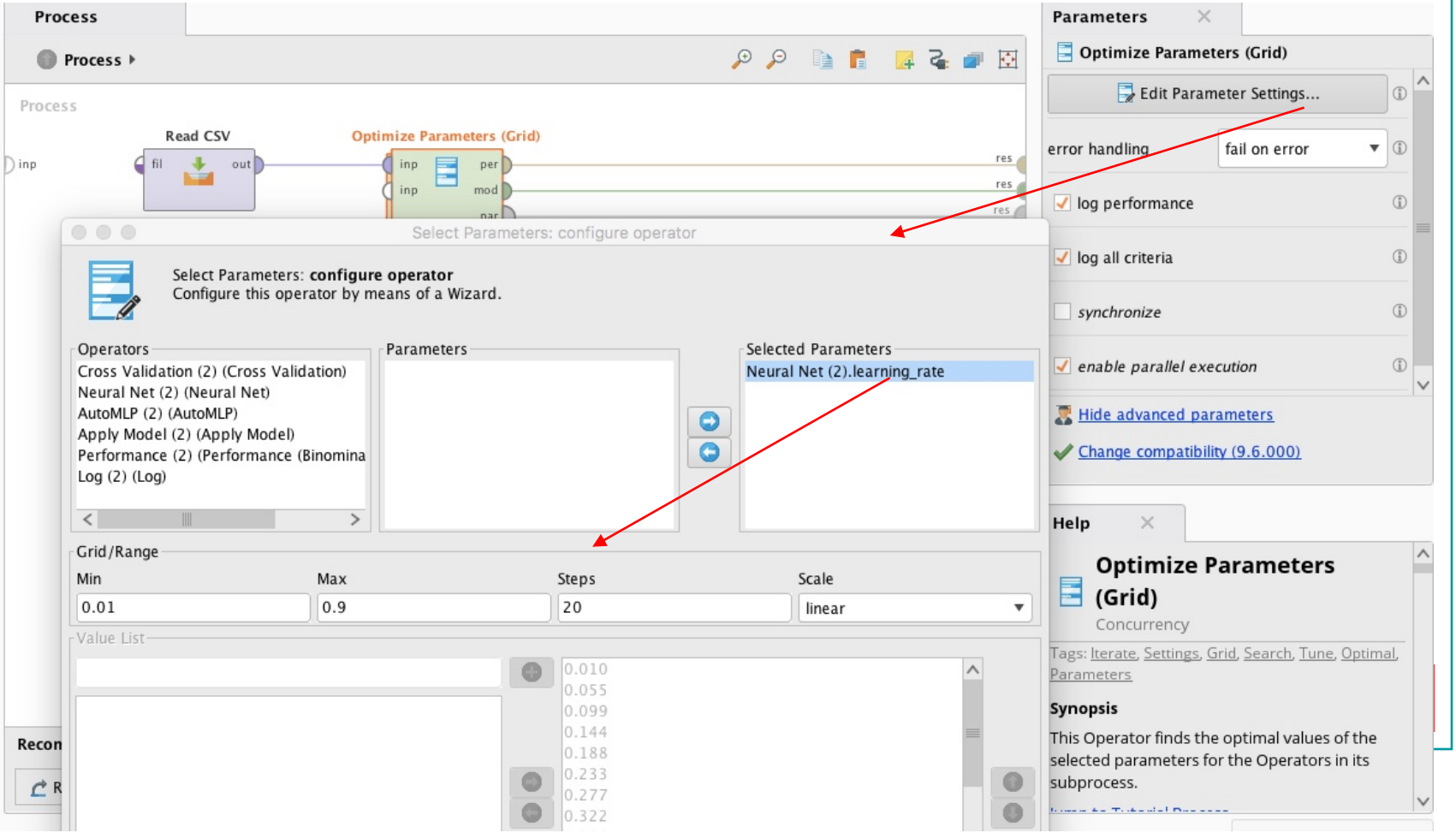
Parameter Optimization – Process Overview

- Import process4.rmp



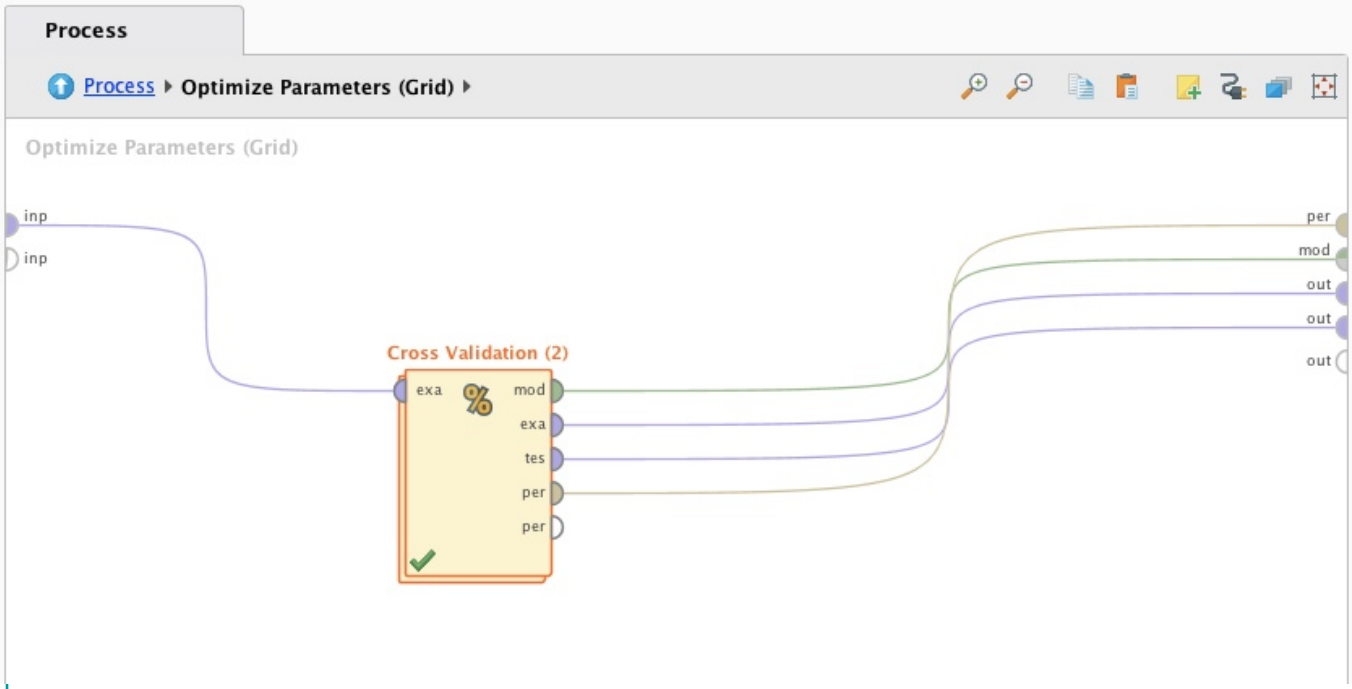
Parameter Optimization

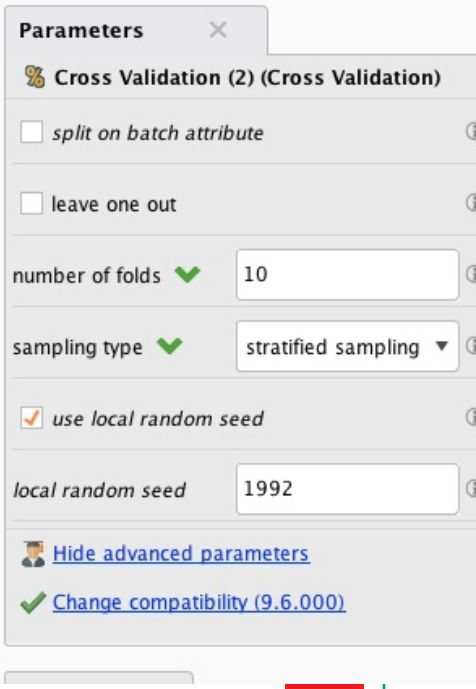
- Optimize Parameters (Grid): choose the hyperparameters you want to optimize.



Parameter Optimization

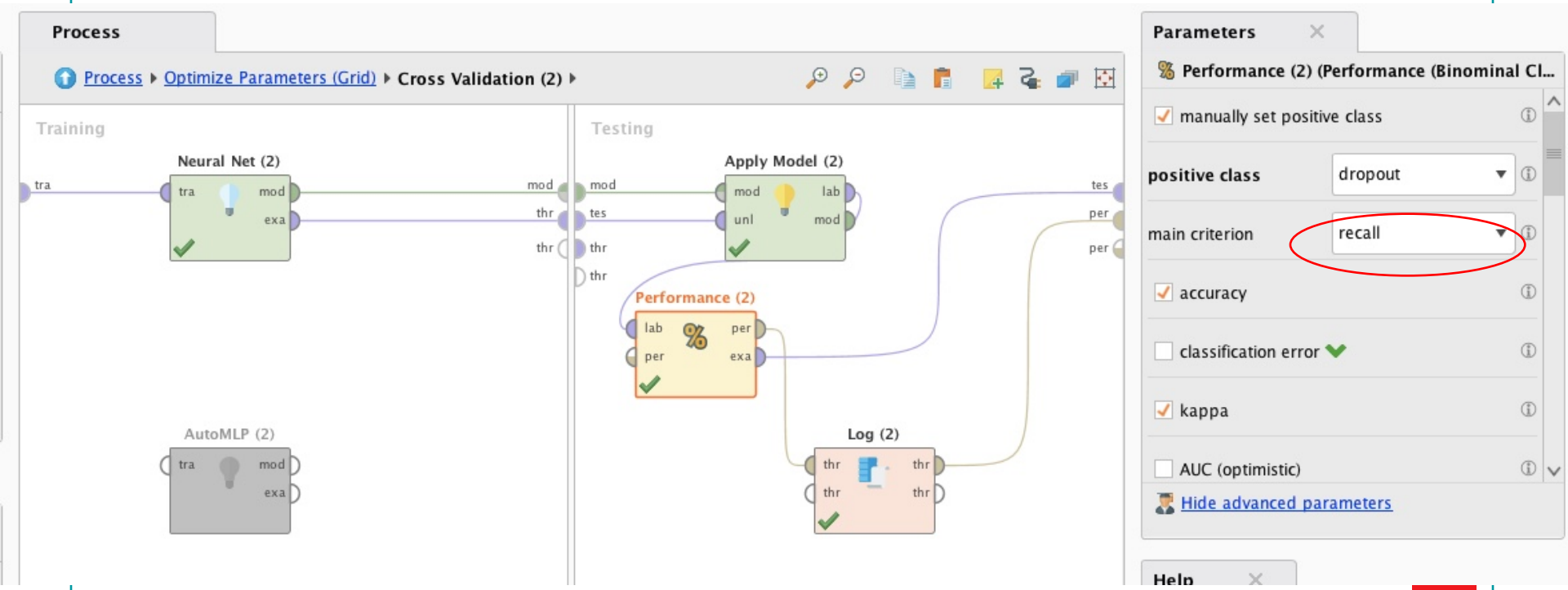
- Optimize Parameters (Grid): nested operator; performs cross-validation.





Parameter Optimization

- Main Criterion decides the optimized parameters.



Parameter Optimization

- Results:

- Look at the best learning rate in ParameterSet
- Compare the confusion matrix with the one from cross-validation.
- Explore all the results by looking at Log, sort on different columns.

References

- Good videos on you tube, for example: https://www.youtube.com/watch?v=C8Ko3-2f-pA&list=PLssWC2d9JhOZLbQNZ80uOxLypglgWqbJA&index=16
- https://community.rapidminer.com/
- https://docs.rapidminer.com/

Questions?

- Thank you for your attention!