

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

<new process*> – RapidMiner Studio Free 9.6.000 @ Agathes-Air.home

Views: Design Results Turbo Prep Auto Model Deployments

Find data, operators...etc All Studio

Repository

Import Data

- Training Resources (connected)
- Samples
- Community Samples (connected)
- DB (Legacy)
- Local Repository (agathemerceron)

Process

Process

```

    graph LR
      Inp((inp)) --> ReadCSV[Read CSV]
      ReadCSV --> NN[Neural Net]
      NN --> Res1((res))
      NN --> Res2((res))
      NN --> Res3((res))
  
```

Parameters

Read CSV

Import Configuration Wizard...

csv file: ?0/MIB-Students_en.csv

column separators: ;

trim lines

[Hide advanced parameters](#)

[Change compatibility \(9.6.000\)](#)

Operators

Search for Operators

- Predictive (62)
 - Lazy (2)
 - Bayesian (2)
 - Trees (9)
 - Rules (5)
 - Neural Nets (4)
 - Deep Learning
 - Neural Net
 - AutoMLP
 - Perceptron

[Get more operators from the Marketplace](#)

Help

Read CSV

RapidMiner Studio Core

Tags: Load, Import, Read, Data, Files, Text, Commas, Spreadsheets, Excel, Datasets, Tsv

Synopsis

This Operator reads an ExampleSet from the specified CSV file.

[Jump to Tutorial Process](#)

Description

CSV is an abbreviation for Comma-Separated Values. The CSV files store data (both numerical and text) in plain-text form. All values corresponding to an Example are stored as one line in the CSV file. Values for different Attributes are separated by a

Recommended Operators

- Retrieve 72%
- Apply Model 56%
- Set Role 40%

Operator Neural Net – Data Access

<new process*> – RapidMiner Studio Free 9.6.000 @ Agathes-Air.home

Views: Design Results Turbo Prep Auto Model Deployments

Find data, operators...etc All Studio

Repository

- Import Data
- Training Resources (connected)
- Samples
- Community Samples (connected)
- DB (Legacy)
- Local Repository (agathemerceron)

Process

Process

Process

Read CSV

Neural Net

Parameters

Read CSV

Import Configuration Wizard...

csv file ?0/MIB-Students_en.csv

column separators ;

Import Data - Select the data location.

Select the data location.

agathemerceron

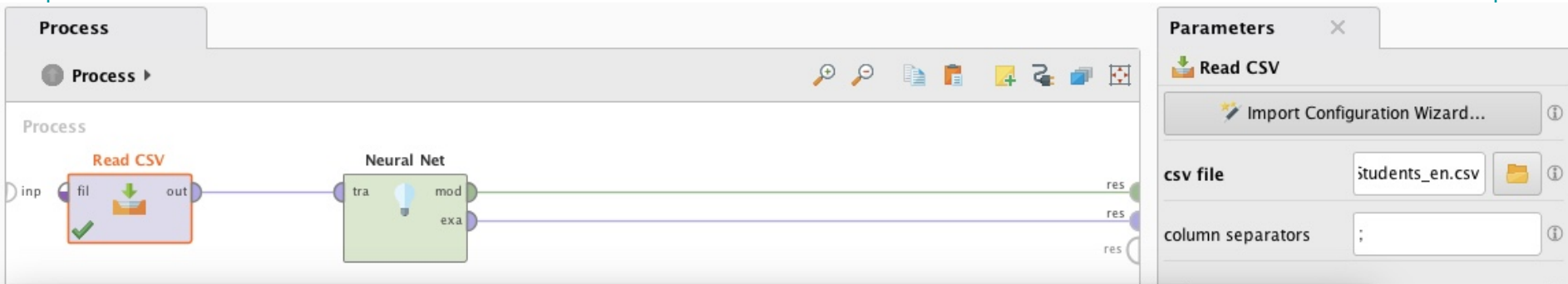
Bookmarks	File Name	Size	Type	Last Modified
★ --- Last Directory				
	anaconda3		File Folder	Mar 24, 2020
	Applications		File Folder	Apr 9, 2020
	Creative Cloud Files		File Folder	May 25, 2020
	Desktop		File Folder	Dec 27, 2019
	Documents		File Folder	May 14, 2020
	Downloads		File Folder	May 27, 2020
	Dropbox		File Folder	Feb 19, 2020
	iCloud Drive (Archive)		File Folder	Aug 24, 2017
	Library		File Folder	Apr 22, 2020
	Movies		File Folder	May 28, 2017
	Music		File Folder	Nov 22, 2019
	ownCloud		File Folder	May 25, 2020
	Pictures		File Folder	May 11, 2020
	Public		File Folder	May 25, 2018

mmas, Spreadsheet

he specified CSV file.

ted Values. The CSV
n plain-text form. All
tored as one line in

Operator Neural Net – Data Access



Import Data - Format your columns.

Format your columns.

Date format Replace errors with missing values ⓘ

	Number_Enrollments_1 <i>integer</i>	Number_Courses_Passed_1 <i>integer</i>	Average_Mark_1 <i>real</i>	class <i>binominal</i>
1	3	3	3.570	dropout
2	5	5	3.160	dropout

A red arrow points to the 'class' column header, which is currently set to 'binominal'.

Operator Neural Net – Role class

Process

Process

Read CSV

Neural Net

Parameters

Read CSV

Import Configuration Wizard...

csv file: students_en.csv

column separators: ;

Import Data - Format your columns.

Format your columns.

Date format: Enter value... Replace errors with missing values

	Number_Enrollments_1 <i>integer</i>	Number_Courses_Passed_1 <i>integer</i>	Average_Mark_1 <i>real</i>	class <i>binominal</i>
1	3			dropout
2	5			dropout
3	4			dropout
4	5			dropout
5	5			dropout
6	5			dropout
7	4			dropout
8	4			dropout

Change role

Please enter the new role:

label

OK Cancel

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.

Result History | ExampleSet (Read CSV) | ImprovedNeuralNet (Neural Net)

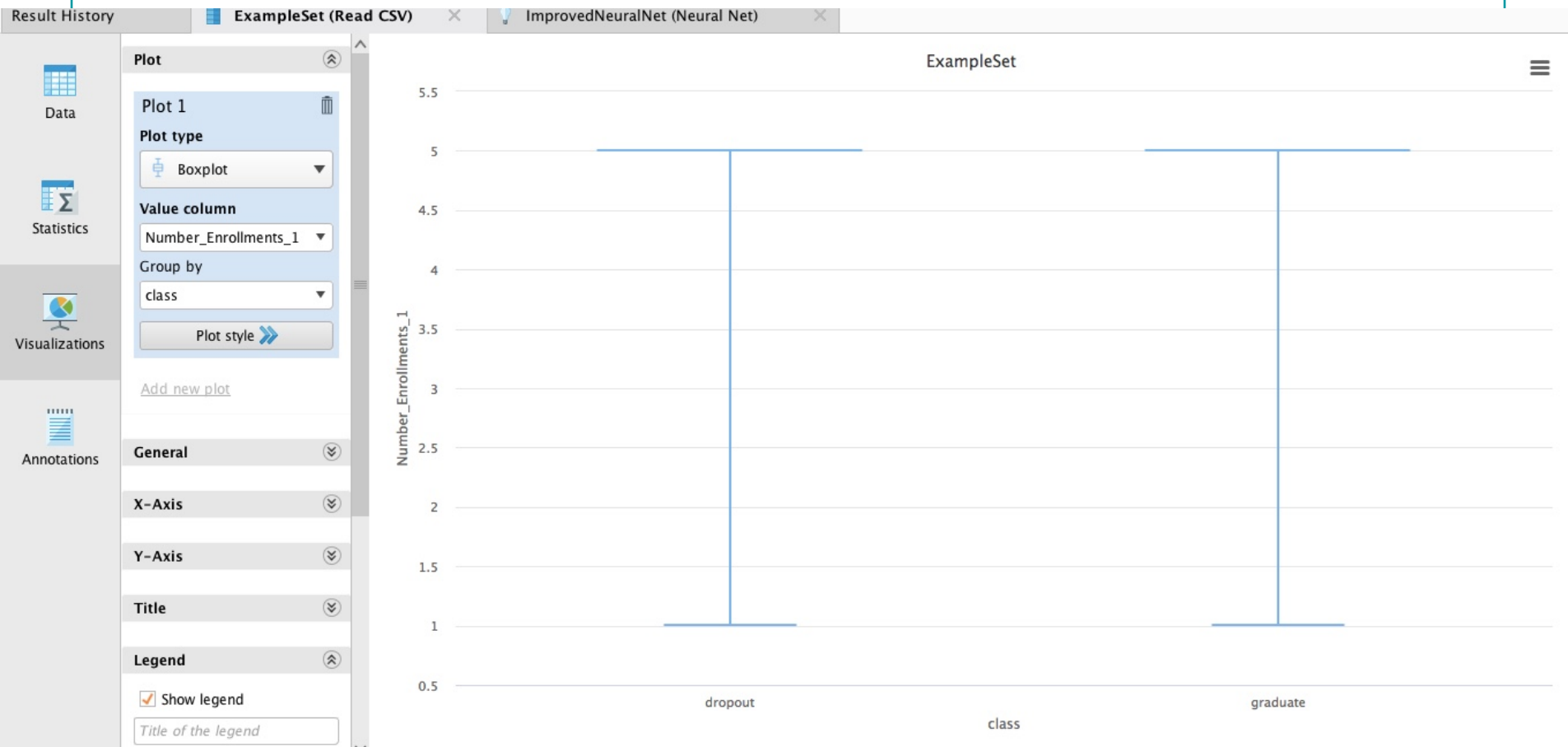
Name	Type	Missing	Statistics	Filter (4 / 4 attributes):
Label class	Polynomial	0	<p>Least graduate (844) Most dropout (877)</p> <p>Values: dropout (877), graduate (844) Details...</p>	Search for Attribute. <input type="text"/>
Number_Enrollments_1	Integer	0	<p>Min 1 Max 5 Average 4.707 Deviation 0.745</p>	
Number_Courses_Passed_1	Integer	0	<p>Min 0 Max 5 Average 3.180 Deviation 1.907</p>	
Average_Mark_1	Real	0	<p>Min 0 Max 4.700 Average 3.056 Deviation 1.447</p>	

Showing attributes 1 - 4

Examples: 1,721 Special Attributes: 1 Regular Attributes: 3

Operator Neural Net – Exploring the Data

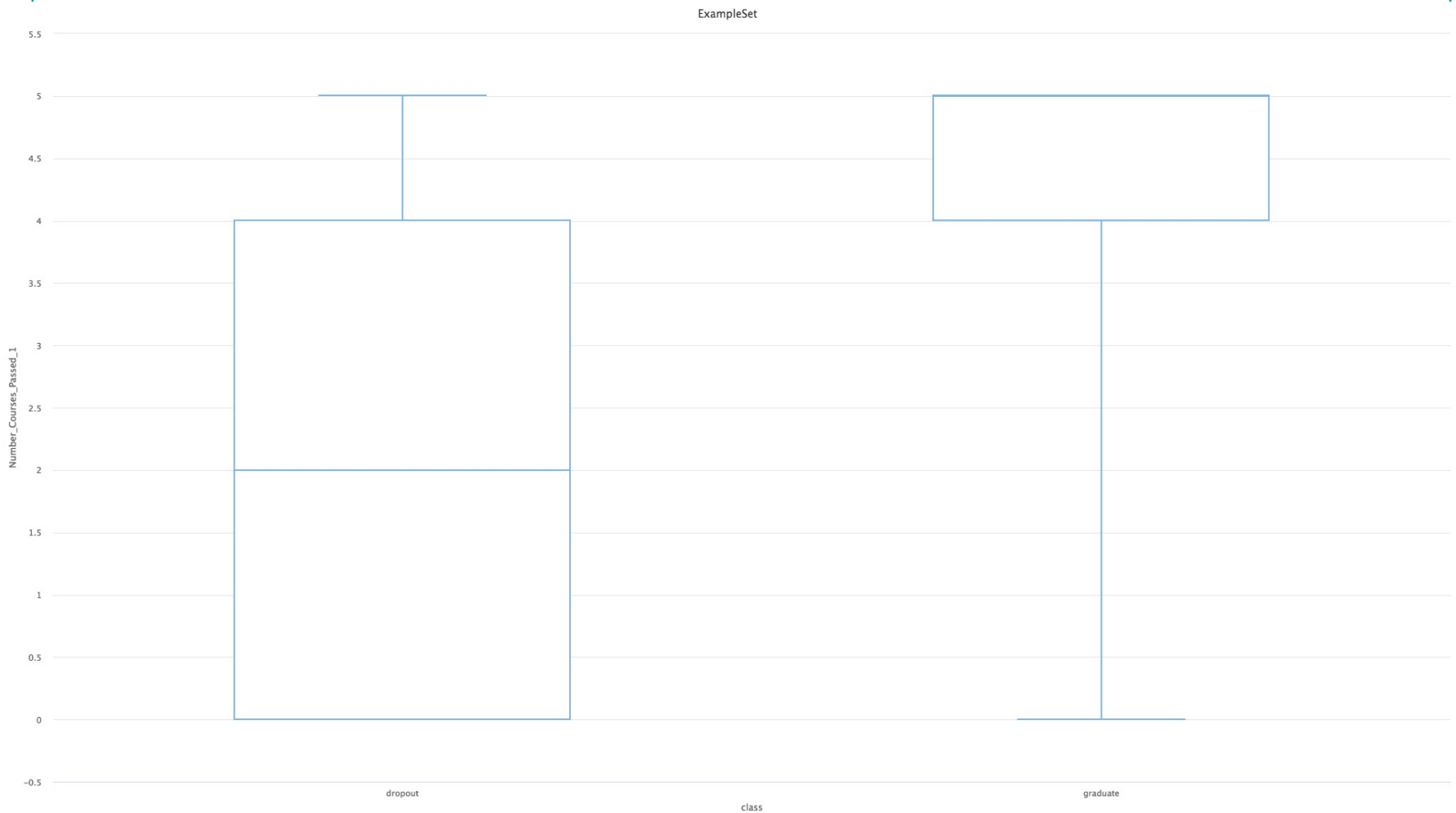
■ Box Plot Number_Enrollments





Operator Neural Net – Exploring the Data

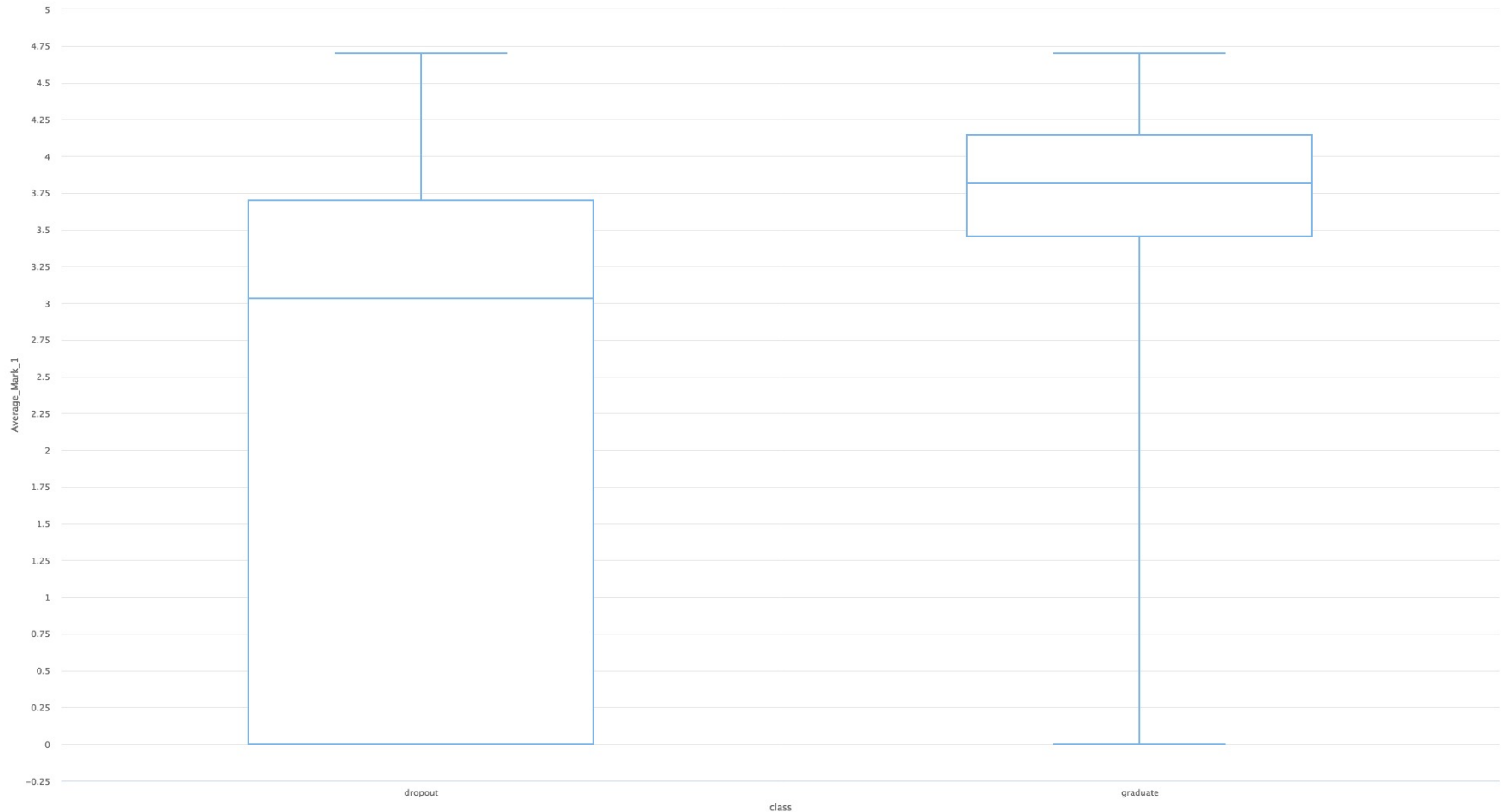
■ Box Plot Number_Courses_Passed



Operator Neural Net – Exploring the Data

■ Box Plot Average_Mark

ExampleSet

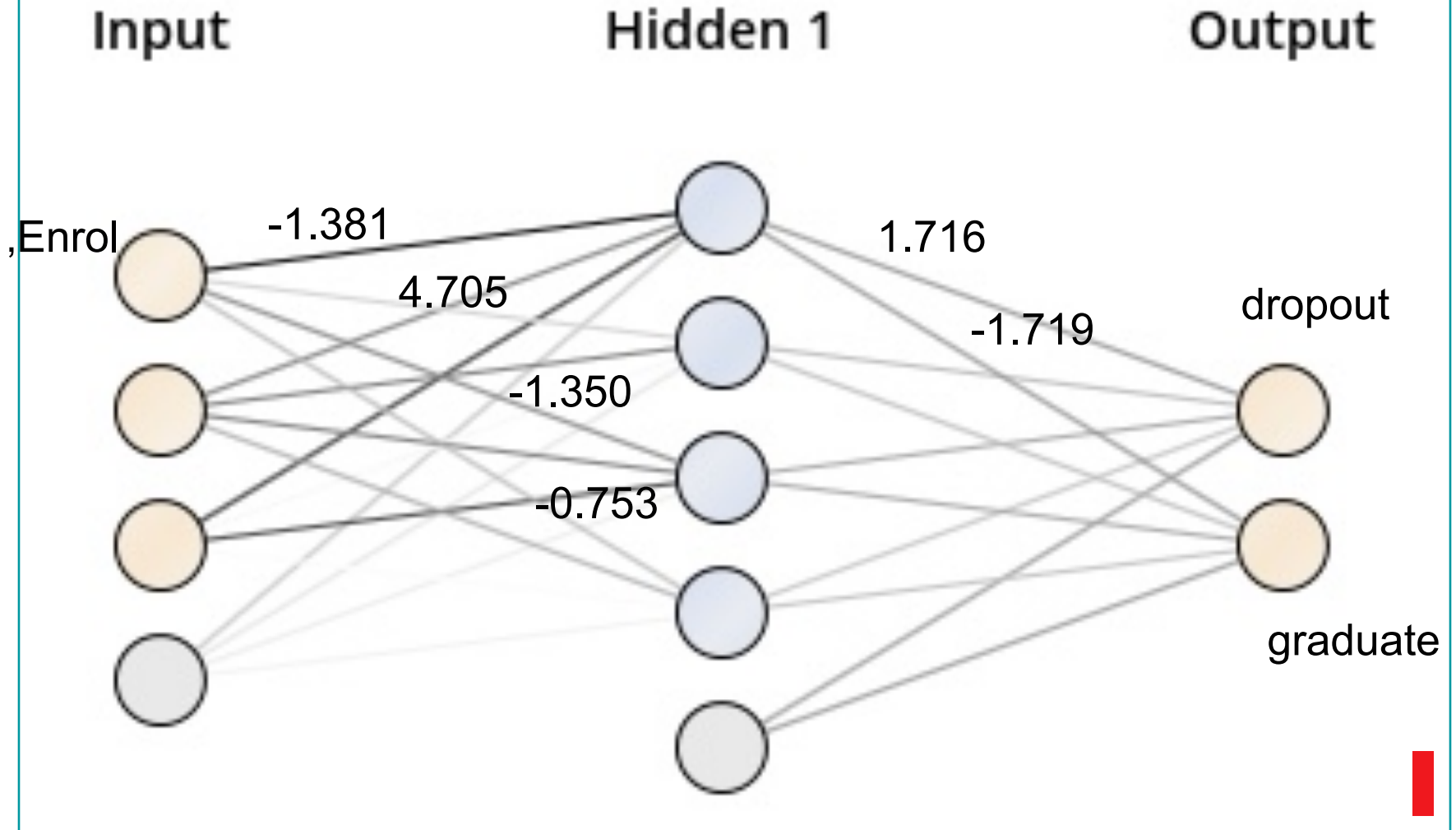


Operator Neural Net – Exploring the Data

- 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.

The screenshot displays a workflow editor interface. On the left, a 'Process' tab is active, showing a workflow with two main components: 'Read CSV' and 'Cross Validation'. The 'Read CSV' process has an input port 'inp' and an output port 'out'. The 'Cross Validation' process has an input port 'exa' and multiple output ports labeled 'res'. The 'Cross Validation' process is highlighted with an orange border. On the right, a 'Parameters' panel is open, showing the configuration for the 'Cross Validation' process. The parameters include:

- split on batch attribute
- leave one out
- number of fol...
- sampling type
- use local random seed
- local random se...
- [Hide advanced parameters](#)
- [Change compatibility \(9.6.000\)](#)

At the bottom of the parameters panel, there is a 'Help' button.

Cross Validation

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

The screenshot displays the Orange3 software interface for a cross-validation workflow. The workflow is divided into two main sections: Training and Testing.

Training Phase:

- An input port labeled 'tra' feeds into the **Neural Net** operator.
- The **Neural Net** operator has two output ports: 'mod' (model) and 'exa' (examples).
- The 'mod' output of the Neural Net operator connects to the 'mod' input of the **Apply Model** operator.
- The 'exa' output of the Neural Net operator connects to the 'exa' input of the **Apply Model** operator.
- An **AutoMLP** operator is also present in the training phase, with its 'tra' input connected to the 'tra' input of the Neural Net operator.

Testing Phase:

- The 'mod' output of the **Apply Model** operator connects to the 'mod' input of the **Performance** operator.
- The 'exa' output of the **Apply Model** operator connects to the 'exa' input of the **Performance** operator.
- The **Performance** operator has two output ports: 'lab' (label) and 'per' (performance).
- The 'per' output of the **Performance** operator connects to the 'thr' input of the **Log** operator.
- The 'lab' output of the **Performance** operator connects to the 'thr' input of the **Log** operator.

Parameters Panel (Performance Operator):

- Performance (Performance (Binomi...)**
- manually set positive class
- positive class:** dropout
- main criterion:** AUC
- accuracy
- classification error ✓
- kappa
- AUC (optimistic)
- [Hide advanced parameters](#)

Help Panel (Performance Operator):

- Performance Binominal Classification**

Cross Validation

- Log your preferred measures.

The screenshot shows an Orange3 workflow for cross-validation. It includes a 'Log' widget in the workflow, a 'Parameters' panel on the right, and an 'Edit Parameter List: log' dialog box in the foreground. A red arrow points from the 'Edit List (4)...' button in the Parameters panel to the dialog box.

Parameters Panel:

- Log
- filename: []
- log: [Edit List (4)...]
- sorting type: none
- persistent

Edit Parameter List: log Dialog:

List of key value pairs where the key is the column name and the value specifies the process value to log.

column name	value
Accuracy	Performance value accuracy
kappa	Performance value kappa
AUC	Performance value AUC
f_measure	Performance value f_measure

Buttons: Add Entry, Remove Entry, Apply, Cancel

Cross Validation

- Results:

	<i>true dropout</i>	<i>true graduate</i>	<i>class precision</i>
<i>pred. dropout</i>	642	115	84.81%
<i>pred. graduate</i>	235	729	75.62%
<i>class recall</i>	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.

<i>Acc.</i>	<i>Kappa</i>	<i>AUC</i>	<i>F1</i>	<i>Recall</i>
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.

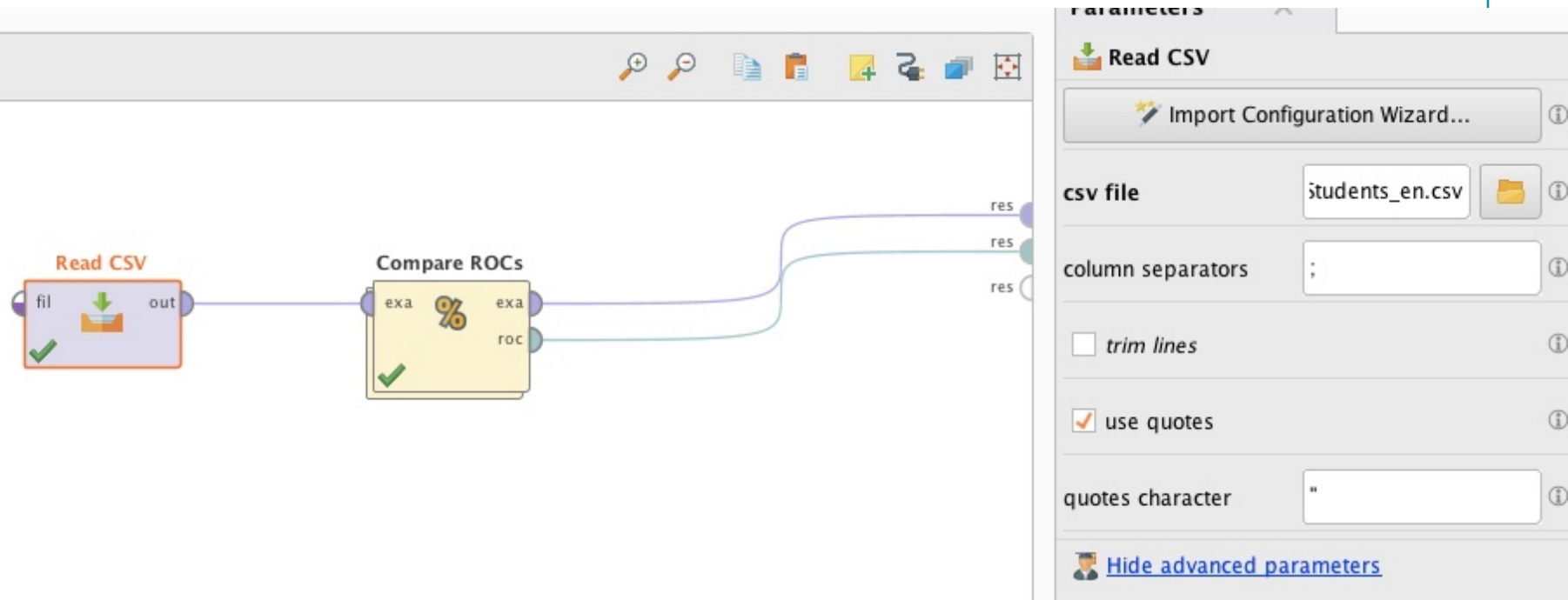


Compare ROCs Operator

- 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 classifiers you like.

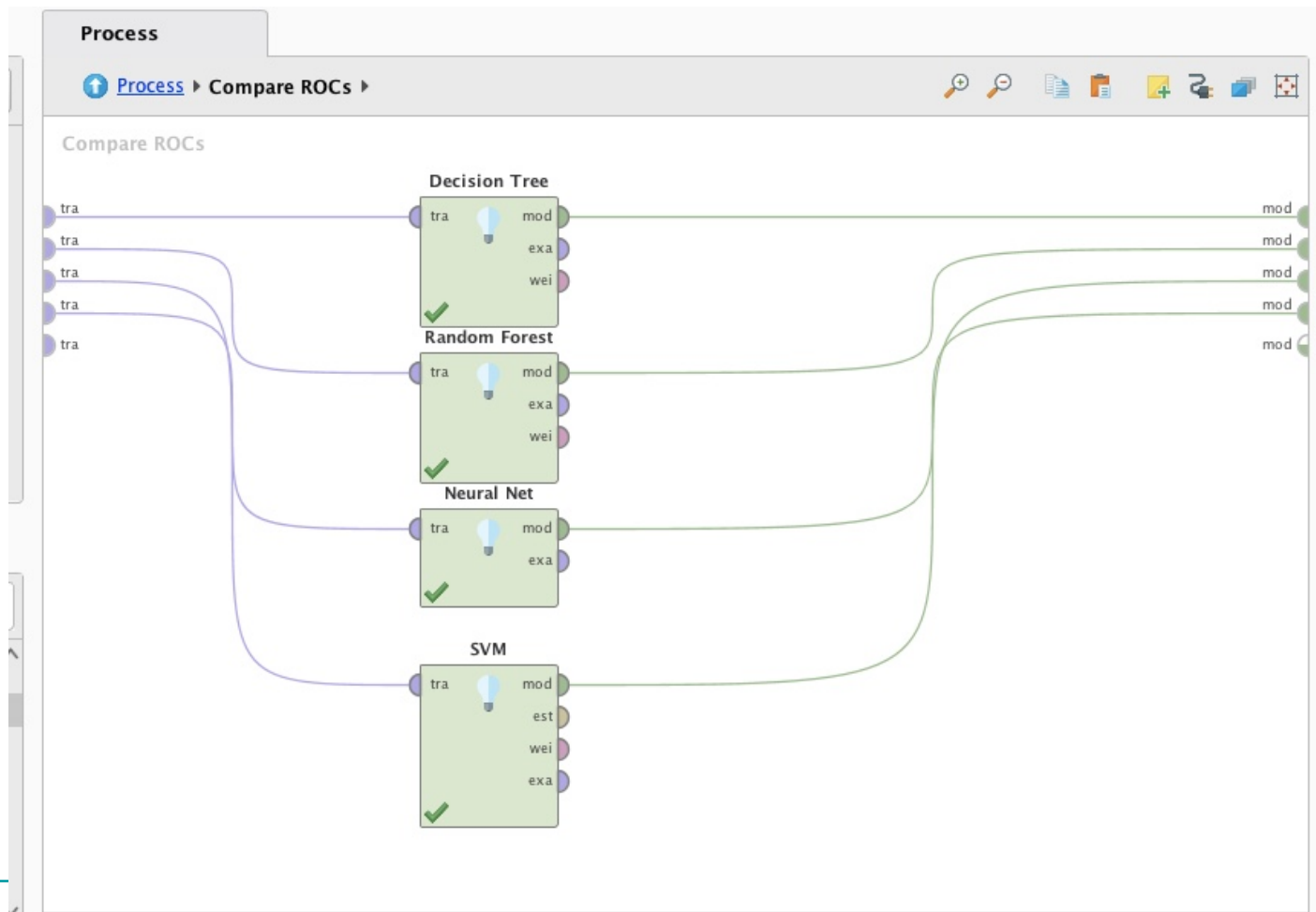
Compare ROCs Operator

- Nested Operator – Performs a Cross Validation



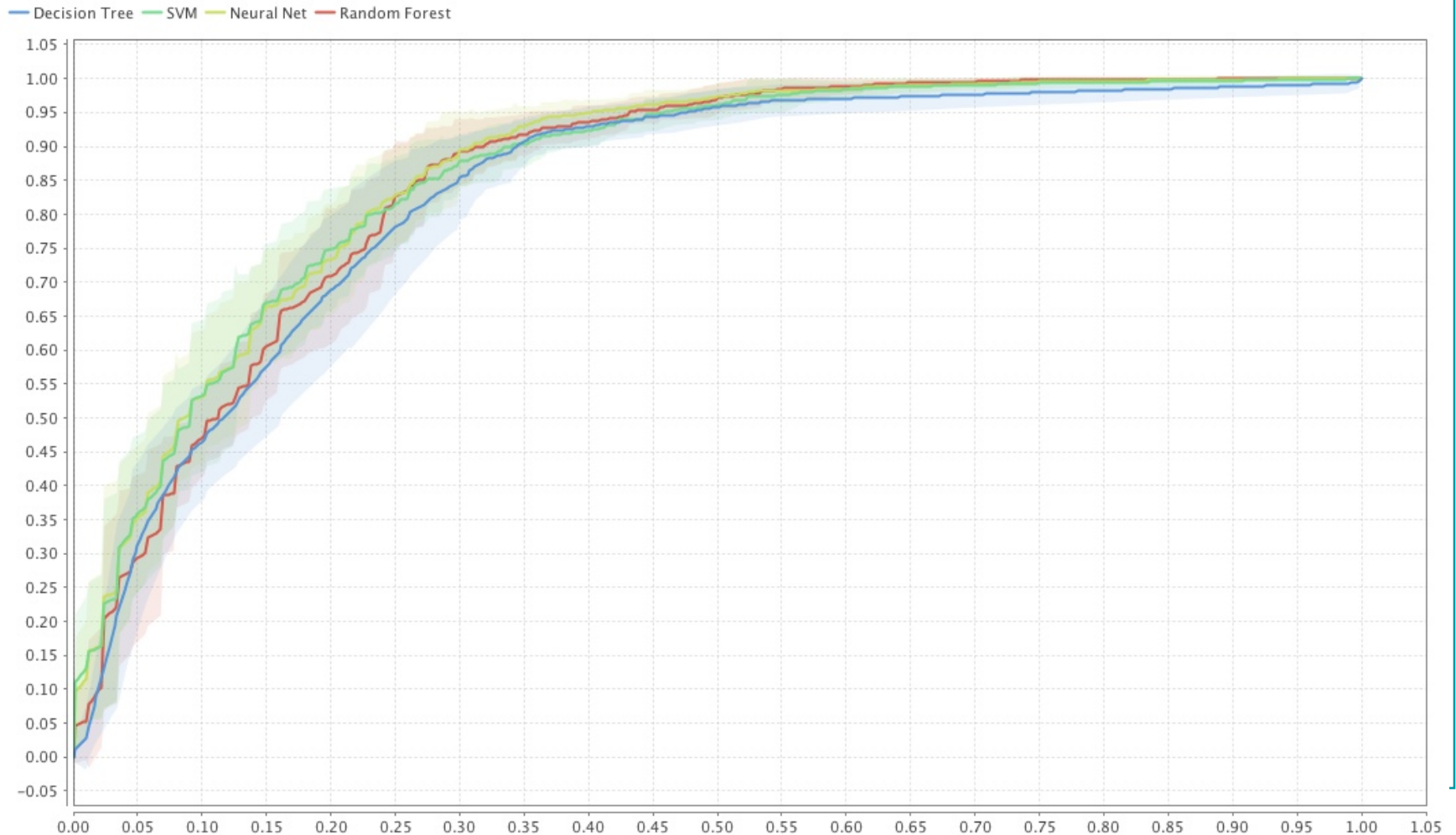
Compare ROCs Operator

- 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

The screenshot displays the Process Designer interface. The main workspace shows a process flow starting with an input 'inp' leading to a 'Read CSV' block, which then connects to an 'Optimize Parameters (Grid)' block. The 'Optimize Parameters (Grid)' block has multiple output ports labeled 'per', 'mod', 'par', and two 'out' ports, each connected to a resource 'res'. The 'Parameters' panel on the right is open, showing settings for 'Optimize Parameters (Grid)'. It includes an 'Edit Parameter Settings...' button, an 'error handling' dropdown set to 'fail on error', and several checkboxes: 'log performance', 'log all criteria', 'synchronize', and 'enable parallel execution'. There are also links for 'Hide advanced parameters' and 'Change compatibility (9.6.000)'. A 'Help' panel is partially visible at the bottom right.

Parameter Optimization

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

Select Parameters: configure operator
Configure this operator by means of a Wizard.

Operators

- Cross Validation (2) (Cross Validation)
- Neural Net (2) (Neural Net)
- AutoMLP (2) (AutoMLP)
- Apply Model (2) (Apply Model)
- Performance (2) (Performance (Binomina
- Log (2) (Log)

Parameters

Selected Parameters

- Neural Net (2).learning_rate

Grid/Range

Min	Max	Steps	Scale
0.01	0.9	20	linear

Value List

0.010
0.055
0.099
0.144
0.188
0.233
0.277
0.322

Parameters

Optimize Parameters (Grid)

Edit Parameter Settings...

error handling: fail on error

log performance

log all criteria

synchronize

enable parallel execution

[Hide advanced parameters](#)

[Change compatibility \(9.6.000\)](#)

Help

Optimize Parameters (Grid)

Concurrency

Tags: [Iterate](#), [Settings](#), [Grid](#), [Search](#), [Tune](#), [Optimal](#), [Parameters](#)

Synopsis

This Operator finds the optimal values of the selected parameters for the Operators in its subprocess.

Parameter Optimization

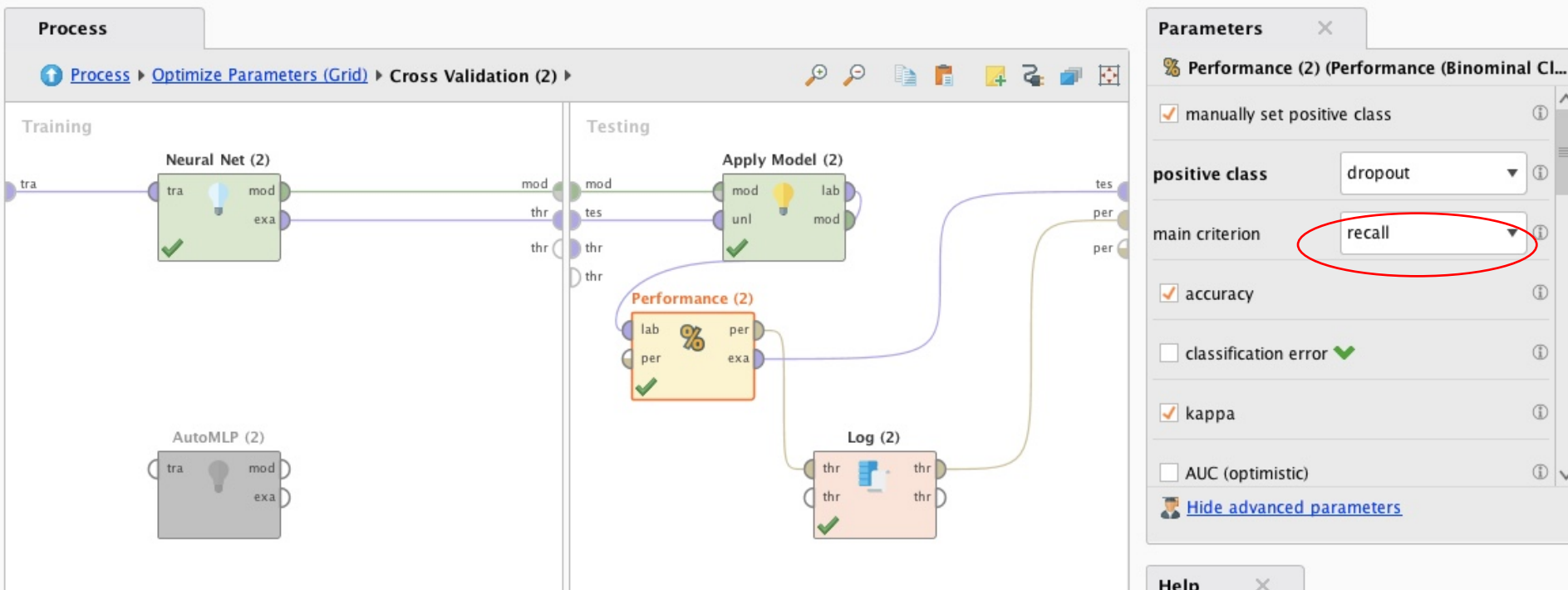
- Optimize Parameters (Grid): nested operator; performs cross-validation.

The screenshot displays a software interface for parameter optimization. The main window is titled "Process" and contains a sub-window "Optimize Parameters (Grid)". Inside this sub-window, a central component labeled "Cross Validation (2)" is connected to various input and output ports. The inputs are labeled "inp" and "exa", while the outputs are labeled "mod", "exa", "tes", "per", and "out". To the right, a "Parameters" panel is open, showing configuration options for the "Cross Validation (2) (Cross Validation)" process. The parameters include:

- split on batch attribute
- leave one out
- number of folds: 10
- sampling type: stratified sampling
- use local random seed
- local random seed: 1992
- [Hide advanced parameters](#)
- [Change compatibility \(9.6.000\)](#)

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!

