

RapidMiner walkthrough

1. Install RapidMiner 7.3 from

<https://my.rapidminer.com/nexus/account/index.html#downloads>

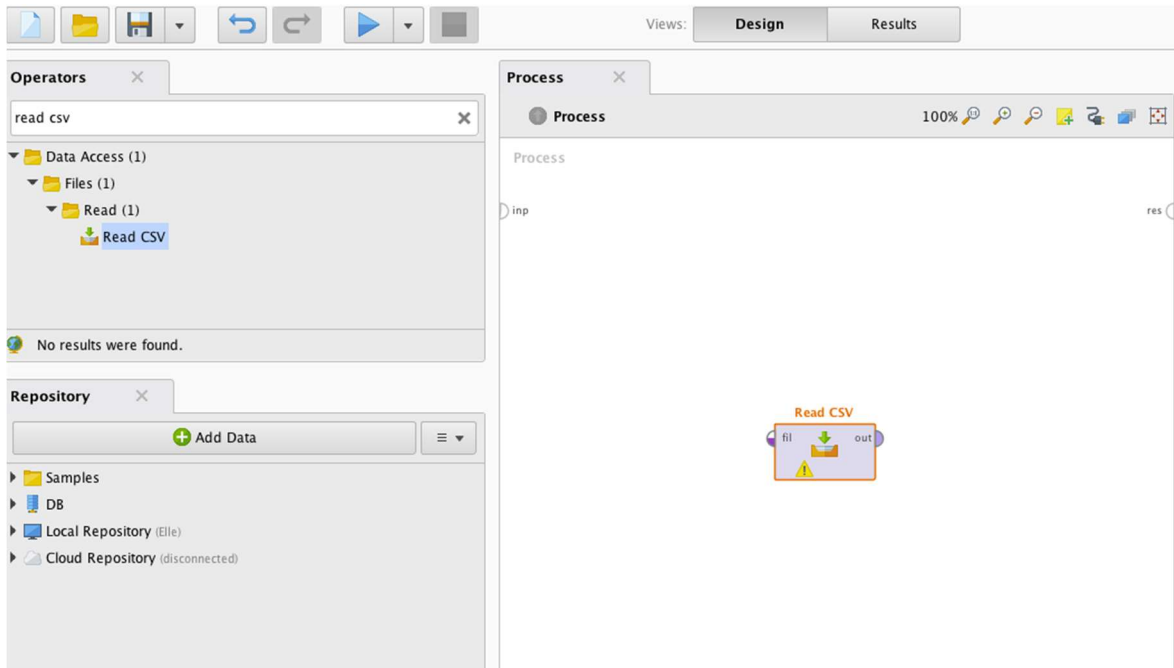
Please also remember to apply for an Educational license now or after this walkthrough practice so that unlimited data rows are allowed. (The default version only allows up to 10,000 rows). You can do so here: <https://my.rapidminer.com/nexus/account/index.html#licenses/request>

When successfully installed, see the next step.

2. Open RapidMiner 7.3 and open a new process

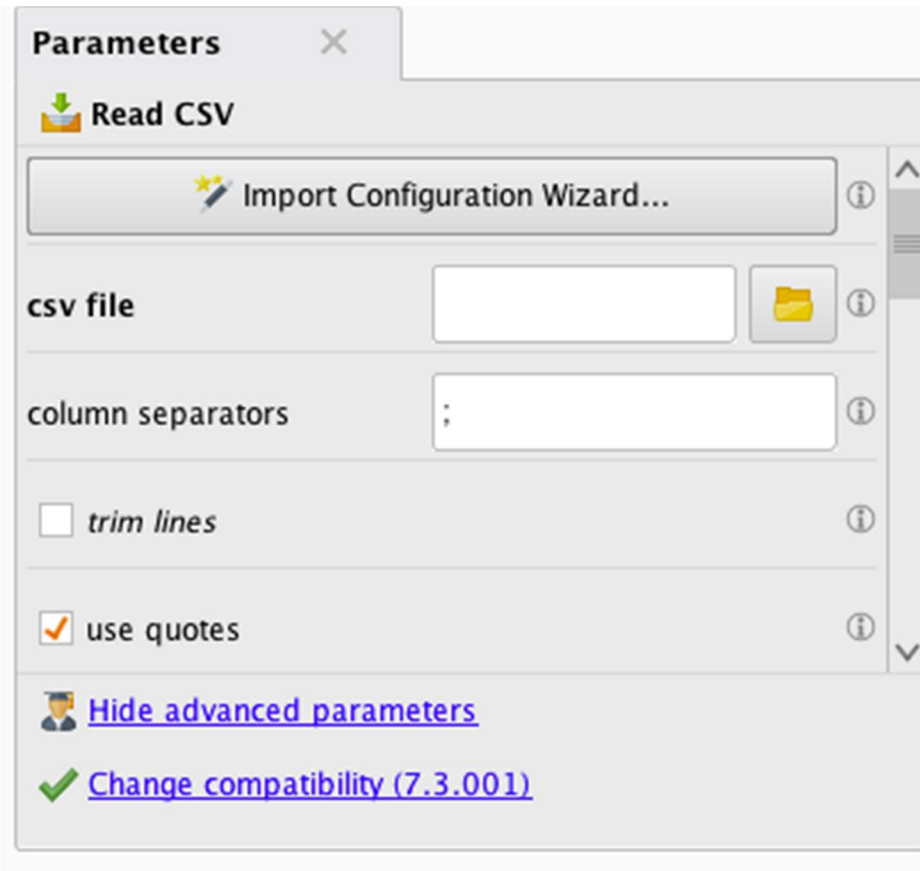
When done, see the next step.

3. Type Read CSV into the operator box to create a new “Read CSV” Operator



When done, see the next step

4. Click on the Import Configuration Wizard on the right side of the interface



When done, see the next step

5. Select file

“SaoPedroetal(2013)_UMUAI_DesigningControlledExperiments_cummandlocalfeatures.csv”

You will have to download it from the course webpage

When done, see the next step

6. This is a "csv" file, so select "Comma Delimited"

Data import wizard - Step 2 of 4

This wizard guides you to import your data.
Step 2: Please specify how the file should be parsed and how columns are separated.

File Reading

File Encoding: UTF-8

Trim Lines

Skip Comments: #

Column Separation

Comma ","

Space

Semicolon ";"

Tab

Regular Expression: .\s*|;\s*

Escape Character: \

Use Quotes: "

Designi...	Group	StateCh...	All t cnt	All t sum	All t mean	All t std...	All t min	All t max	All t med	Run cnt	Run t sum	Run t m...
N	2	1	2	18	9	9.8994...	2	16	9	1	2	2
N	2	2	5	13	2.6	1.5165...	1	5	2	2	3	1.5
Y	3	1	0	0	0	0	0	0	0	0	0	0
N	2	1	13	100	7.6923...	7.7069...	1	27	4	3	8	2.6666...
N	1	1	9	293	32.555...	69.125...	2	216	11	3	223	74.333...
Y	3	2	11	162	14.727...	32.875...	1	113	3	2	6	3
N	6	1	1	151	151	0	151	151	151	0	0	0
N	2	4	7	192	27.428...	58.965...	2	161	6	1	10	10
N	5	1	6	16	2.6666...	1.9663...	1	6	2	1	1	1
N	2	1	0	0	0	0	0	0	0	0	0	0
Y	3	1	11	678	61.636...	177.28...	2	595	4	2	10	5
N	6	2	0	0	0	0	0	0	0	0	0	0
N	2	1	2	150	75	81.750...	0	142	75	1	142	142

Row, Column	Error	Original value	Message
-------------	-------	----------------	---------

← Previous → Next Finish Cancel

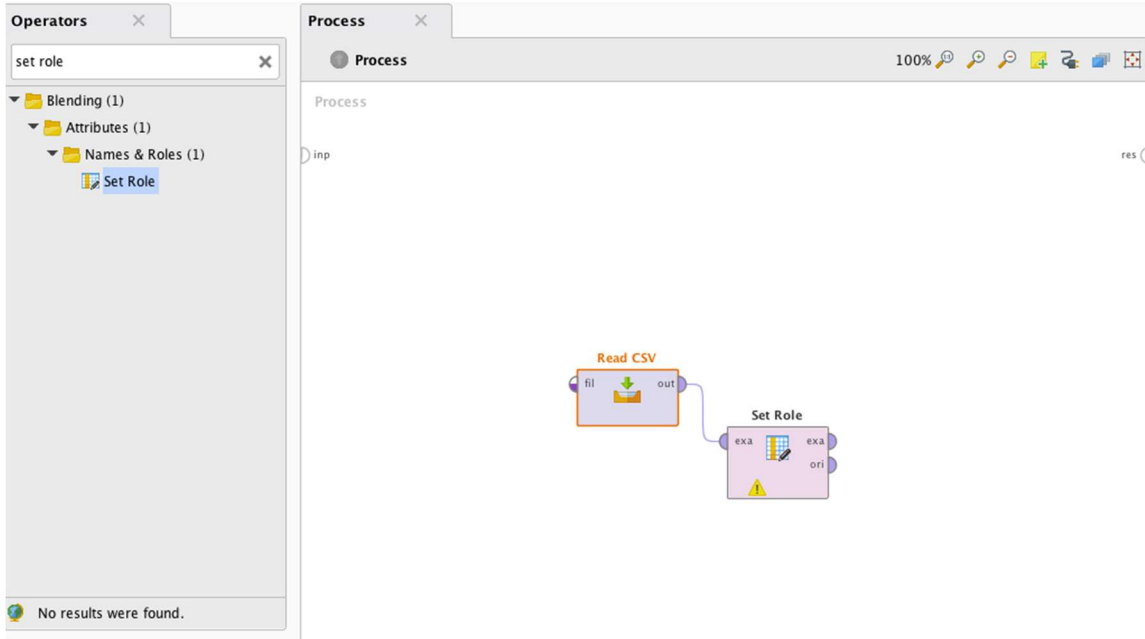
When done, click [HERE](#)

7. Click Next until the system does not let you click Next anymore. Then click Finish.

When done, see the next step

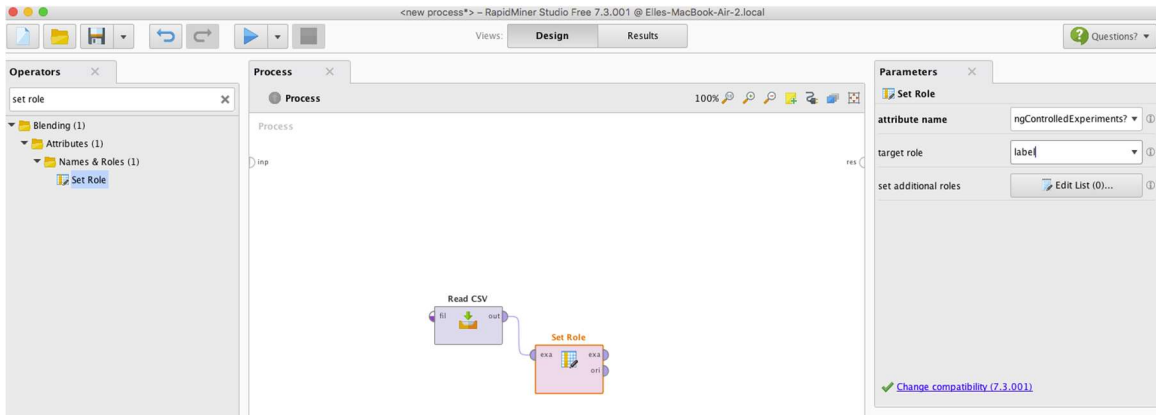
8. Create a “Set Role” operator in the operator box at the top-left.

Then connect the output bubble on the right side of “Read CSV” to the input bubble on the left side of “Set Role” by clicking on the output bubble and then clicking on the input bubble. Your screen should look like this.



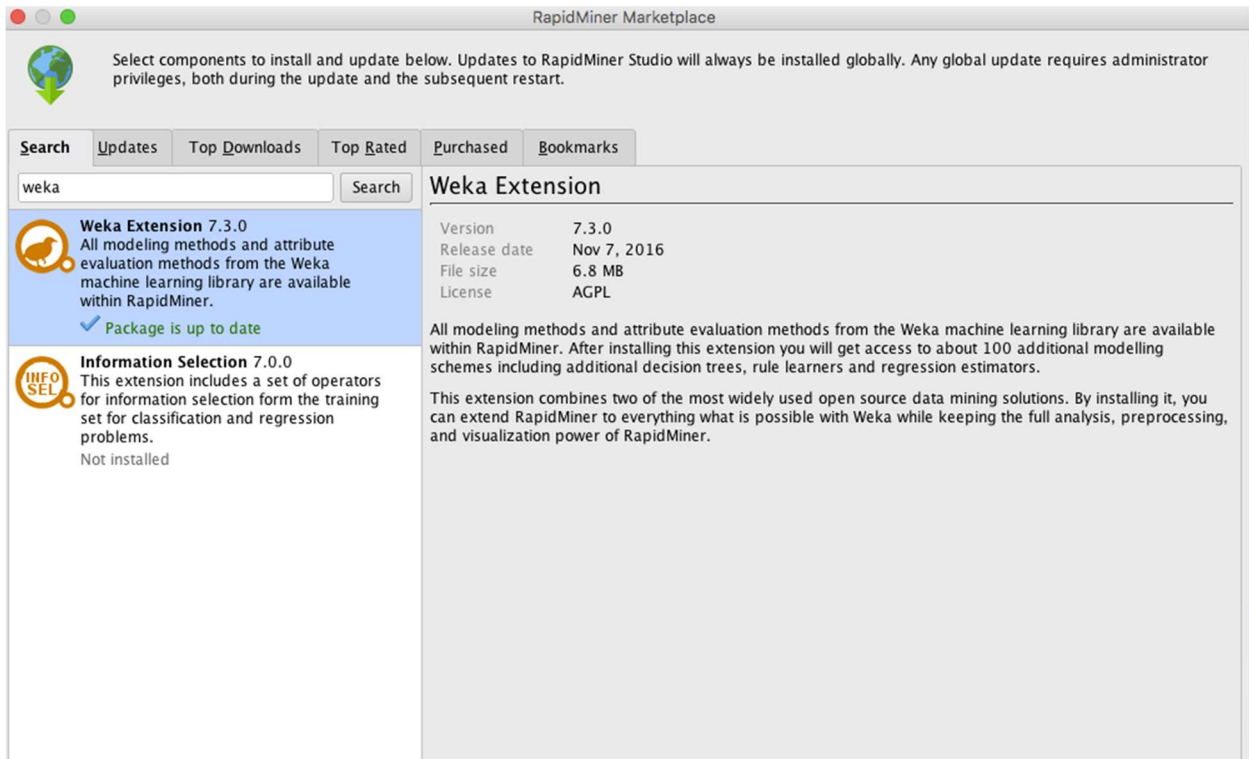
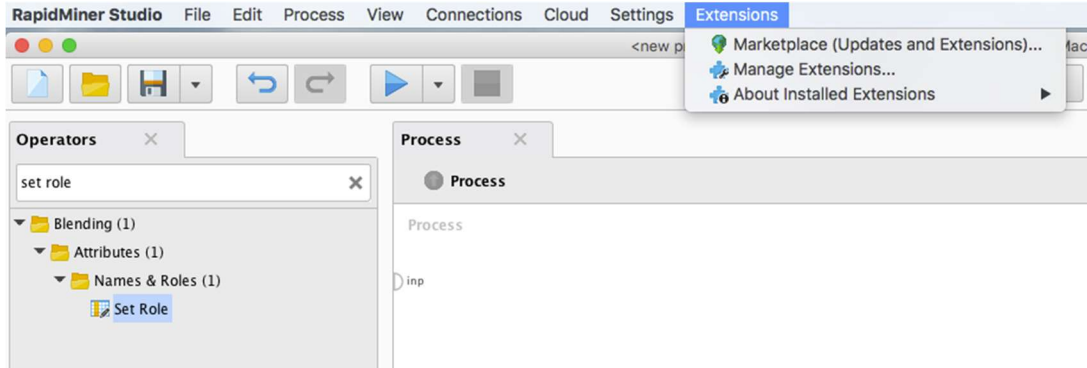
When done, see the next step.

- Now go over to the right side and select `DesigningControlledExperiments` as the variable you want to change, and set it to be a “label” in the target role box.



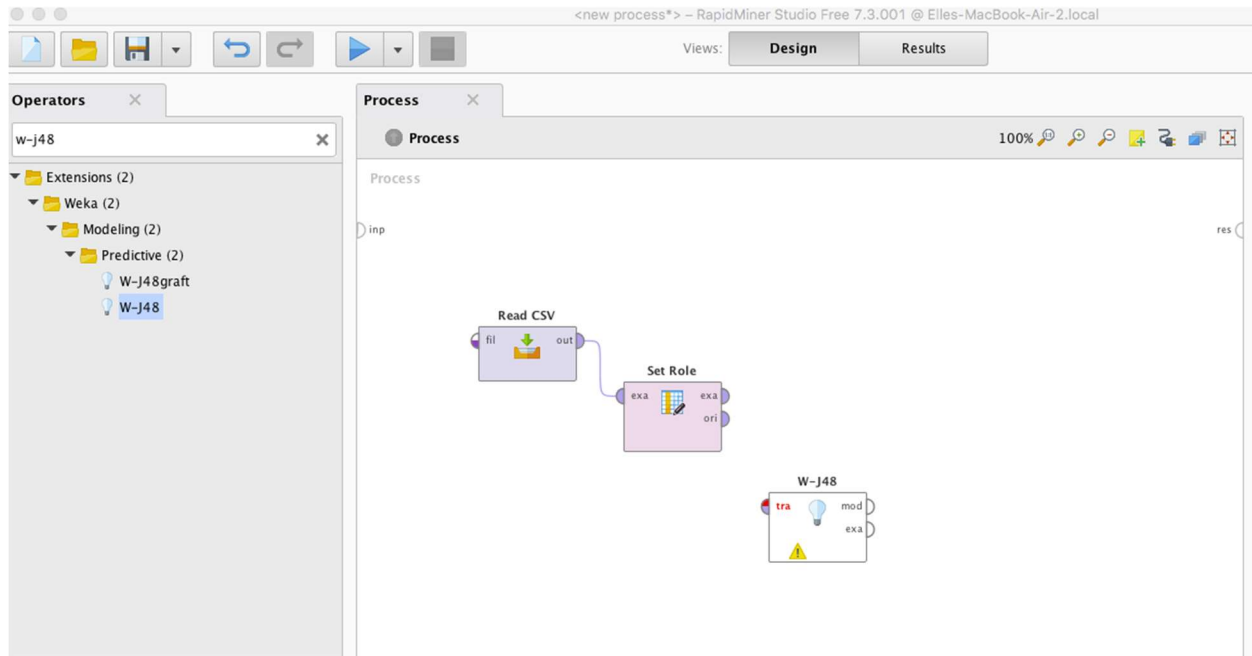
When done, see the next step.

10. Install the WEKA Expansion Pack. To do this go to the Extensions menu, and select Marketplace (Updates and Extensions). Search for Weka, and install the Weka Expansion Pack.



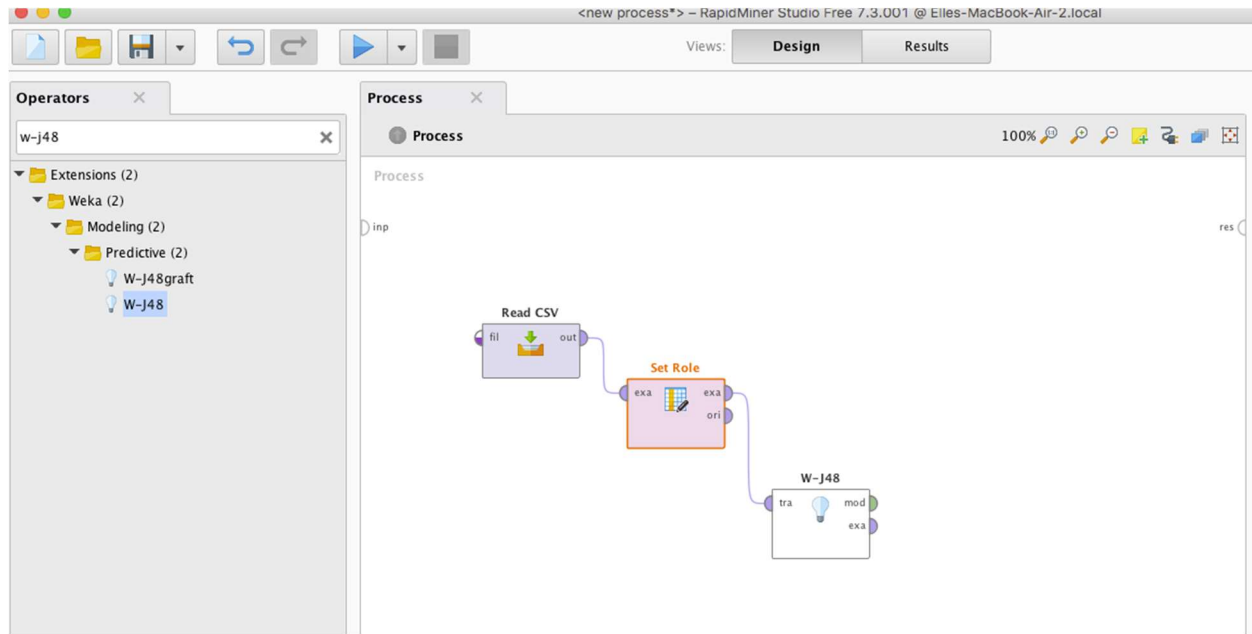
When done, see the next step.

11. Type w-j48 into the operators window, and create the w-j48 operator



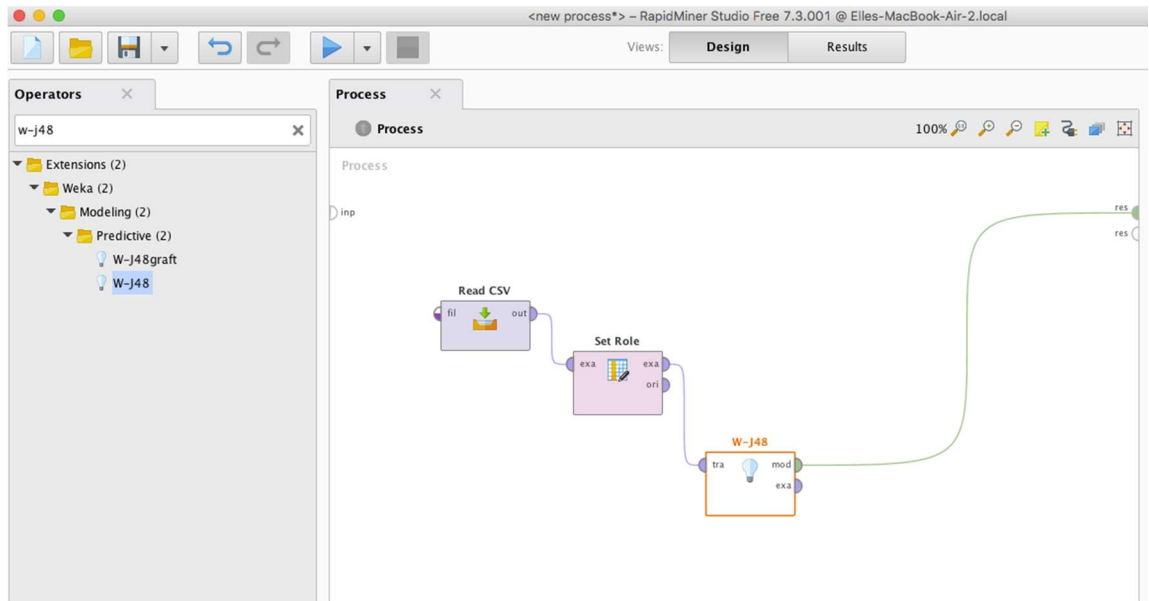
When done, see the next step.

12. Now connect the output bubble from Set Role (exa for example set) to the input bubble from J48 (tra for training set)



When done, see the next step.

13. Then connect the output bubble from W-J48 (model) to the res (result) bubble on the far right



When done, see the next step.

14. Then press play at the top of the screen. After a minute or so (possibly longer for slower computers), you should see your model

The screenshot shows the RapidMiner Studio interface. At the top, there's a title bar with the text "<new process*> - RapidMiner Studio Free 7.3.001 @ Elles-MacBook-Air-2.local". Below the title bar is a toolbar with icons for file operations and a play button. The main window is titled "W-J48 (W-J48)" and shows a "Weka result" window. The window displays a decision tree for "J48 pruned tree". The tree structure is as follows:

```
Cm CVS cnt <= 0: N (271.0/2.0)
Cm CVS cnt > 0
  | CVS cnt <= 0
  |   | Run t sum <= 11
  |   |   | Hyp table show t sum <= 1
  |   |   |   | Cm Pause cnt <= 2
  |   |   |   |   | Data table show cnt <= 0
  |   |   |   |   |   | Hyp var change cnt <= 4
  |   |   |   |   |   |   | Cm Hyp var change cnt <= 12
  |   |   |   |   |   |   |   | Mw iv change t med <= 6.5
  |   |   |   |   |   |   |   |   | Cm Run cnt <= 4
  |   |   |   |   |   |   |   |   |   | All t min <= 1
  |   |   |   |   |   |   |   |   |   |   | Cm Incmplt run t min <= 2
  |   |   |   |   |   |   |   |   |   |   |   | Cm Cmplt run t sum <= 2: N (10.0/1.0)
  |   |   |   |   |   |   |   |   |   |   |   |   | Cm Cmplt run t sum > 2
  |   |   |   |   |   |   |   |   |   |   |   |   |   | Hyp var change t min <= 1
  |   |   |   |   |   |   |   |   |   |   |   |   |   |   | Hyp make t sum <= 112
  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | Cm Mw iv change t max <= 17
  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | Cm Rept cnt <= 0
  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | Cm Hyp make t stddev <= 33.234019
  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | Cm Data table show cnt <= 2
  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | Cm All t min <= 1
  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | Cm All t cnt <= 12: N (4.0/1.0)
  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | Cm All t cnt > 12
  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | Cm Data table show t sum <= 12
  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | Cm Hyp var change t min <= 3: Y (15.0/1.0)
  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | Cm Hyp var change t min > 3: N (4.0/1.0)
  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | Cm Data table show t sum > 12: N (3.0)
  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | Cm All t min > 1: Y (2.0)
  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | Cm Data table show cnt > 2: N (2.0)
  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | Cm Hyp make t stddev > 33.234019: N (2.0)
  |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   | Cm Rept cnt > 0: Y (9.0)
```

When done, see the next step.

15. This representation shows how the model makes decisions. You can read it as follows:

If the variable CM cvs cnt is less than or equal to zero, then the model predicts No.

In the original data set, there were 271 cases where this prediction was correct, and 2 cases where it was wrong. So the confidence of this prediction is $(271)/(271+2) = 271/273 = 99.27\%$.

If the variable CM cvs cnt is greater than zero, then the model goes to the next variable.

If the variable CVS ct is less than or equal to zero, then

If the variable Run T Sum is less than or equal to 11, then
about 11 other things,

to finally get to a prediction of No with $10/11 = 90.9\%$ confidence

(Note that you have to scroll down to see the case where CVS ct is greater than zero).

When done, see the next step.

16. Note that J48 decision trees are extremely complicated to think through all at once.

And they are one of the simpler algorithms to interpret!

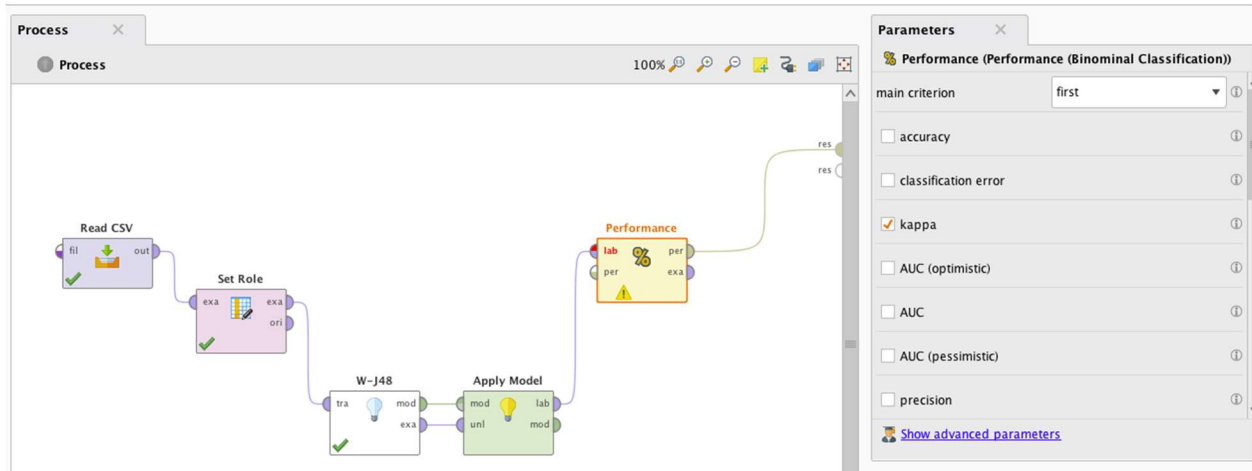
When done, see the next step.

17. Click on the Design button at the top to go back to the main screen.

The screenshot shows the Weka software interface. At the top, there is a toolbar with icons for file operations and navigation, and a 'Views:' section with 'Design' and 'Results' buttons. Below the toolbar, a tab labeled 'W-J48 (W-J48)' is active. The main area displays the 'Description' of the result, titled 'W-J48' and 'J48 pruned tree'. The result is a decision tree structure represented by a series of vertical lines and text. The tree starts with a root node 'Cm CVS cnt <= 0: N (271.0/2.0)'. The left branch leads to 'Cm CVS cnt > 0', which further branches into 'CVS cnt <= 0' and 'Run t sum <= 11'. The right branch leads to 'Hyp table show t sum <= 1', which branches into 'Cm Pause cnt <= 2' and 'Data table show cnt <= 0'. The 'Data table show cnt <= 0' node branches into 'Hyp var change cnt <= 4' and 'Cm Hyp var change cnt <= 12'. The 'Hyp var change cnt <= 4' node branches into 'Mw iv change t med <= 6.5' and 'Cm Run cnt <= 4'. The 'Mw iv change t med <= 6.5' node branches into 'All t min <= 1' and 'Cm Incmplt run t min <= 2'. The 'All t min <= 1' node branches into 'Cm Cmpl run t sum <= 2: N (10.0/1.0)' and 'Cm Cmpl run t sum > 2'. The 'Cm Cmpl run t sum <= 2: N (10.0/1.0)' node branches into 'Hyp var change t min <= 1'.

When done, see the next step.

18. Now add two more operators to the right of W-J48. First, an Apply Model, and second, a Performance (Binomial Classification). Choose kappa in the window to the right. Make sure that you link the operators as shown here. You can delete a link by right-clicking on it and selecting delete, or you can click on it and press the delete button. Then press run.



When done, see the next step.

19. You should see this screen. This shows you the model's Kappa and confusion matrix. The kappa is excellent, in fact too good. Keep in mind we did not use cross-validation, so this model is being trained and tested on the same data set.

Here's how to read the confusion matrix. There are 165 cases where the model says "Y" and the data says "Y". There are 383 cases where the model says "N" and the data says "N". There are 11 cases where the model says "N" and the data says "Y". There are 5 cases where the model says "Y" and the data says "N".

Table View Plot View

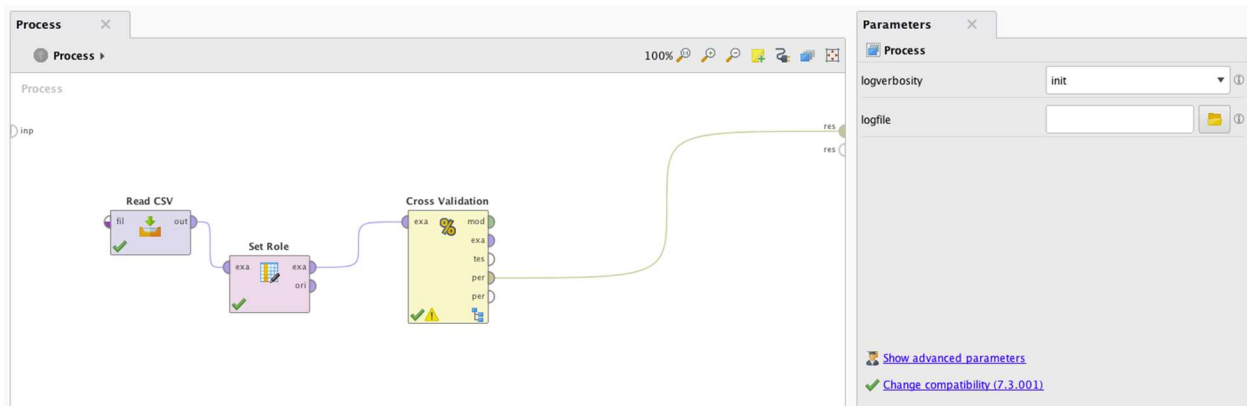
kappa: 0.933

	true N	true Y	class precision
pred. N	383	11	97.21%
pred. Y	5	165	97.06%
class recall	98.71%	93.75%	

When done see the next step.

20. Now go back to the main screen, and create what you see here. You should **delete W-J48, Apply Model, and Performance**, and **add Cross Validation**. You will get some error messages. Don't worry about those for now. In many cases, you'll want to do Batch X-Validation instead of X-Validation. Batch-X-Validation allows you to do student-level cross-validation, or item-level cross-validation, or population-level cross-validation. Regular X-validation supports flat cross-validation, as talked about in the video lecture.

Note the options over to the right, which allow you to do k-fold cross-validation (currently set up to do 10-fold cross-validation), or to do leave-one-out cross-validation.



When done see the next step.

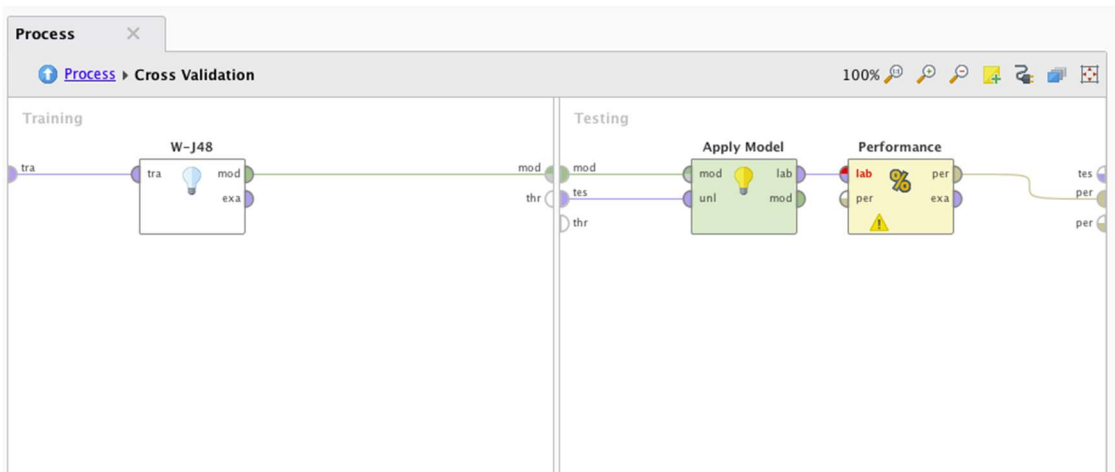
21. Now double click on the validation box (the tall yellow one).

It will bring you to another screen. Add operators as shown here – the same ones you just deleted.

The left box represents what you do with the training folds – build a model.

And the right box represents what you do with the test folds – apply the model, and see how well it does.

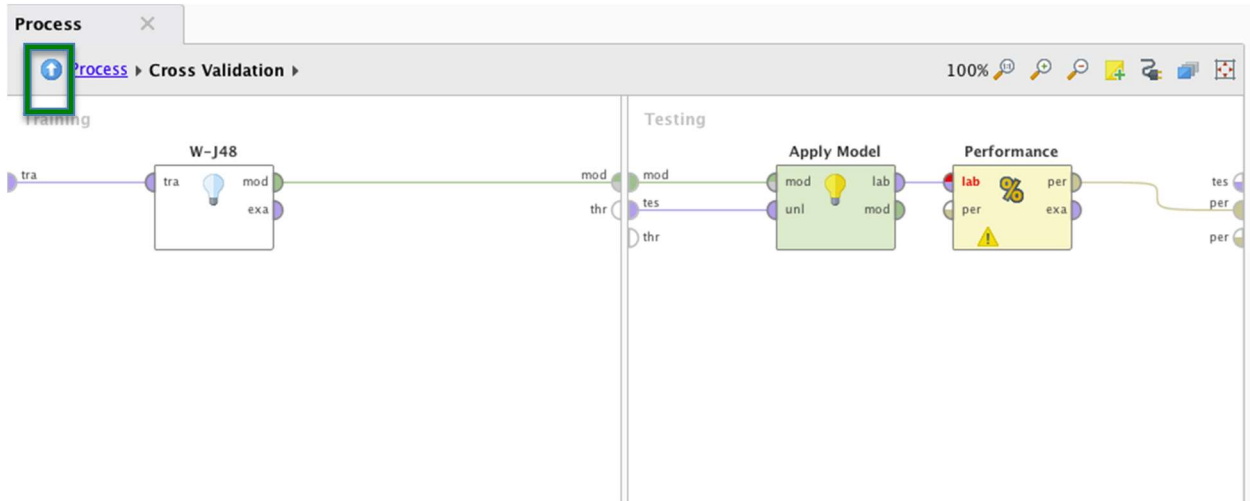
Set up everything the same way you did before, e.g. with Performance (Binomial Classification) and the kappa statistic.



When done see the next step.

22. You can click the blue up arrow to go back to the main screen

When done see the next step.



23. Click to run the model. You should get this. Note that kappa is a lot lower once we're cross-validating.

Table View Plot View

kappa: 0.442 +/- 0.153 (mikro: 0.445)

	true N	true Y	class precision
pred. N	325	70	82.28%
pred. Y	63	106	62.72%
class recall	83.76%	60.23%	

When done, see the next step.

24. So now you've built a model and validated it. There's a lot more things you could do.

You could

- Use student-level cross-validation (you would have to add the variable student back in)
- Try different algorithms, such as W-Jrip, W-KStar, KNN, Logistic Regression, Linear Regression (which gives you Step Regression for binomial data)
- Try creating new features (try Generate Attributes) or removing features (try Remove Correlated Attributes)

Have fun!