RapidMiner Walkthrough

EDUC6191, Fall 2022 v2

If you haven't applied for a RapidMiner academic license, please follow the instructions below to do so.

Having an academic license enables you to conduct analyses on more than 10,000 rows of data, and unlocks features available only to licensed users.

Apply for a RapidMiner academic license via this website:
https://my.rapidminer.com/nexus/account/index.html#signup

You could try one of the following two ways for quicker approval:

1. Apply with an email address from an internationally-recognized academic institution,

2. Copy and paste the following statement in the section of the application that asks, "What will you be using the Academic License for?":

   - If you are a taking Prof. Baker's in-person class in UPenn: "**I am a student of Core Methods in Educational Data Mining by Dr. Ryan Baker at UPenn**"

   - If you are taking the MOOC: "**I am student of the edX MOOC Big Data and Education by Dr. Ryan Baker at UPenn**"

RapidMiner Walkthrough

EDUC6191, Fall 2022

1. Install the latest version of RapidMiner from
   https://my.rapidminer.com/nexus/account/index.html#downloads

Please also remember to apply for an Educational License now or after this walkthrough 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.

RapidMiner Walkthrough

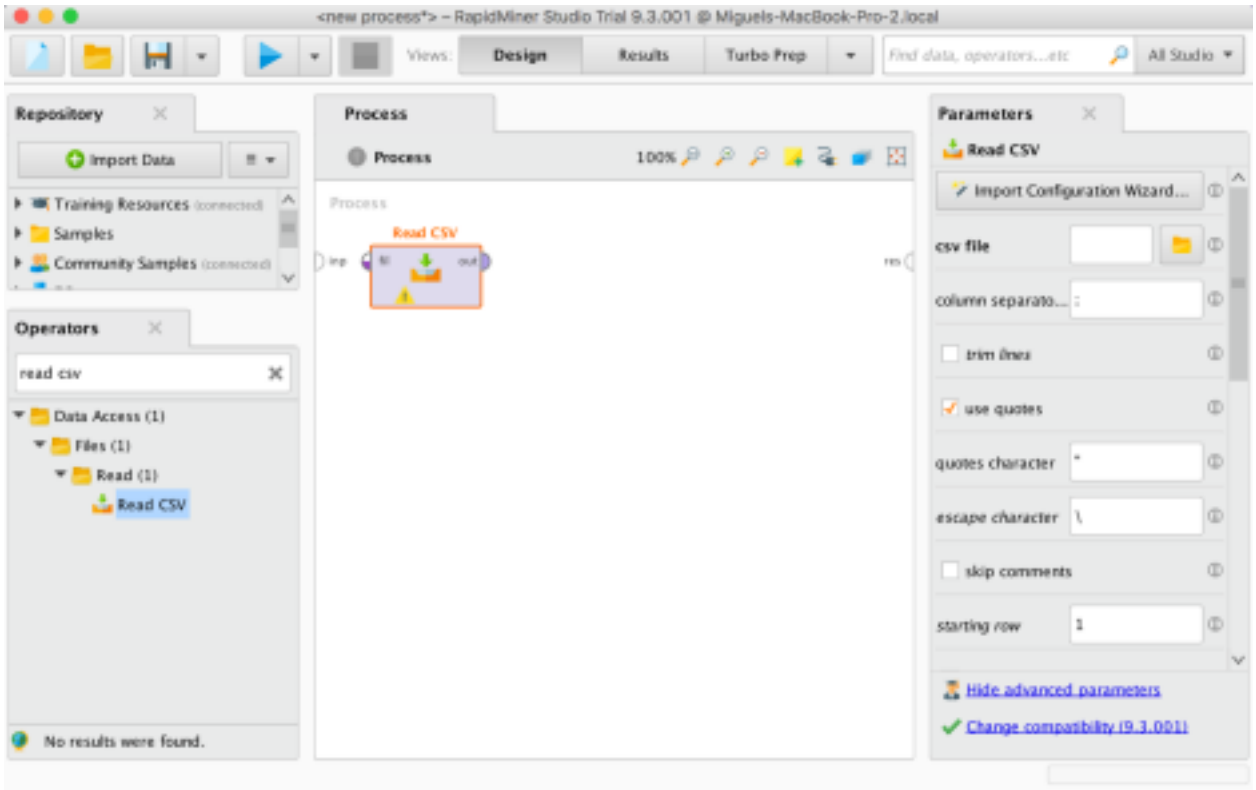
EDUC 6191, Fall 2022 2. Run RapidMiner and

start a New Process.

RapidMiner Walkthrough

EDUC 6191, Fall 2022

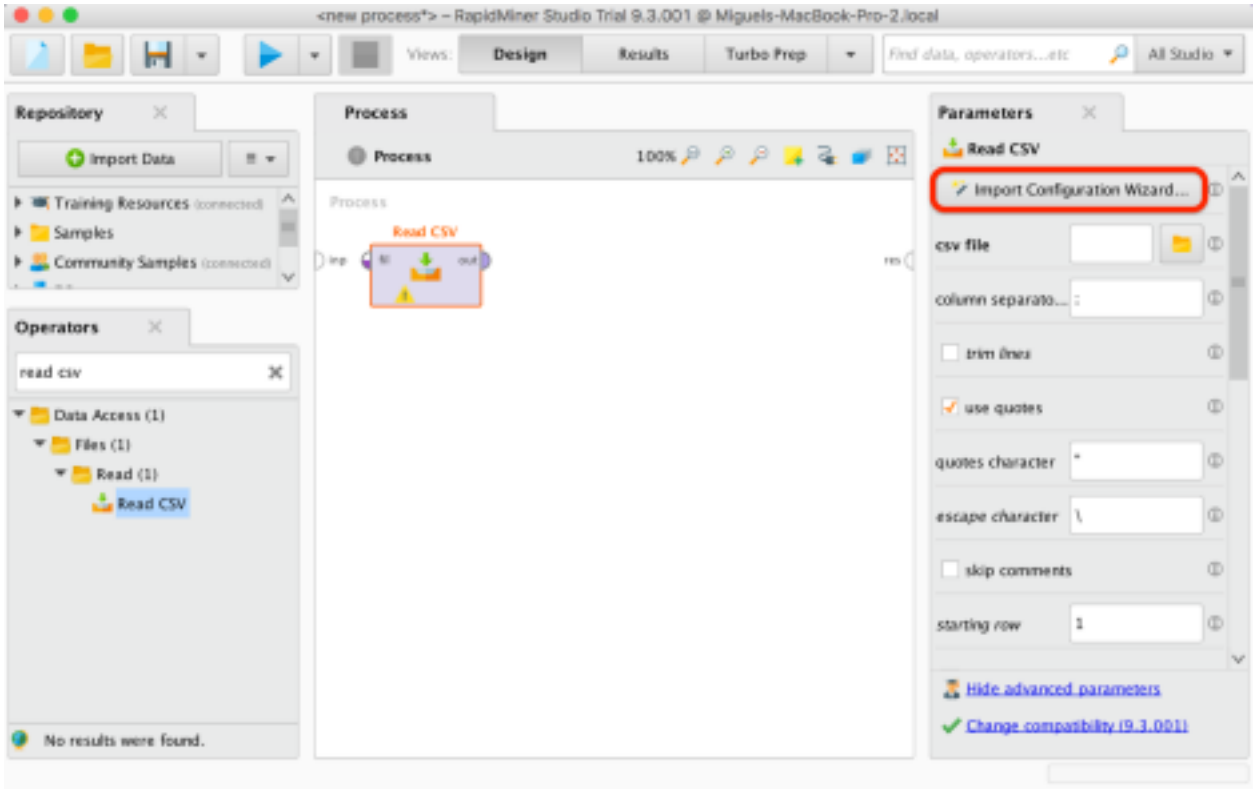
3. Type "`read csv`" in the Operators search bar and create a new "`Read CSV`" operator by dragging it into your Process window.



RapidMiner Walkthrough

EDUC 6191, Fall 2022

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



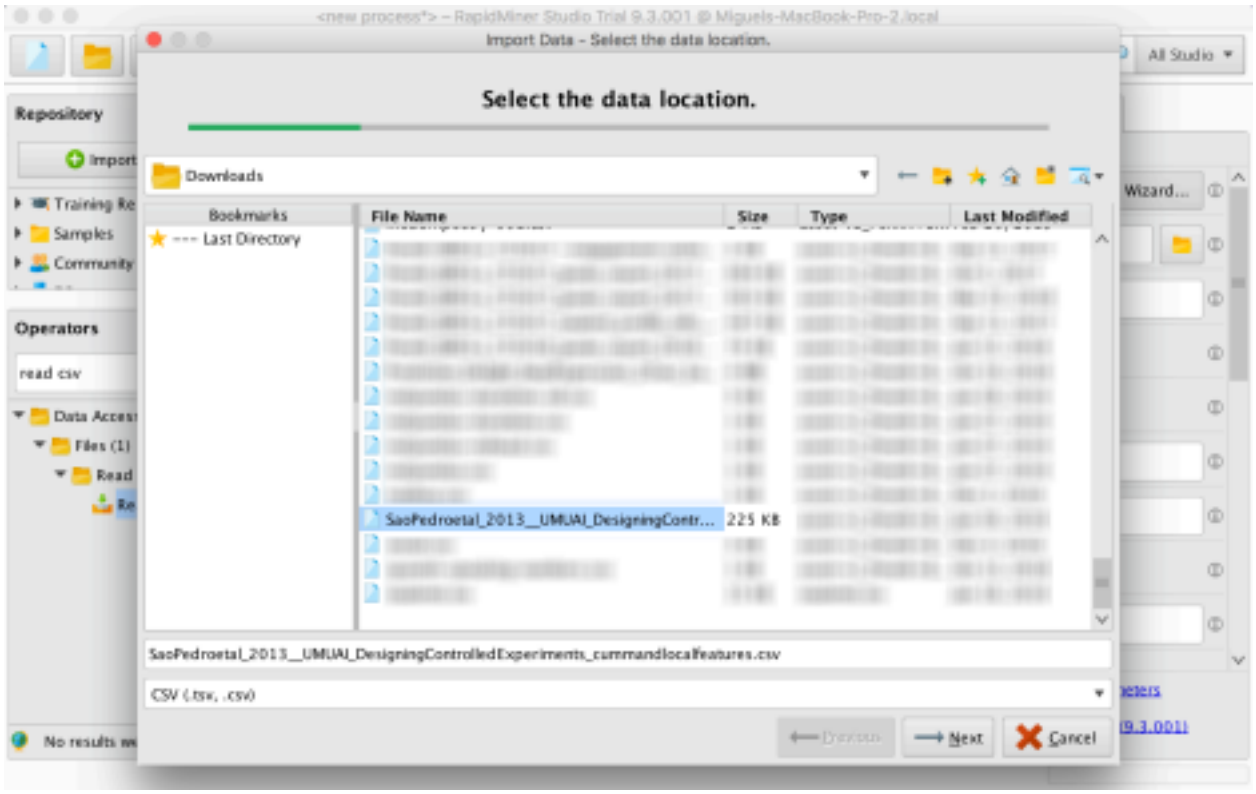
RapidMiner Walkthrough

EDUC 6191, Fall 2022

5. Browse to and select

```
SaoPedroetal(2013)_UMUAI_DesigningControlledExperiments_cummandloca
l features.csv
```

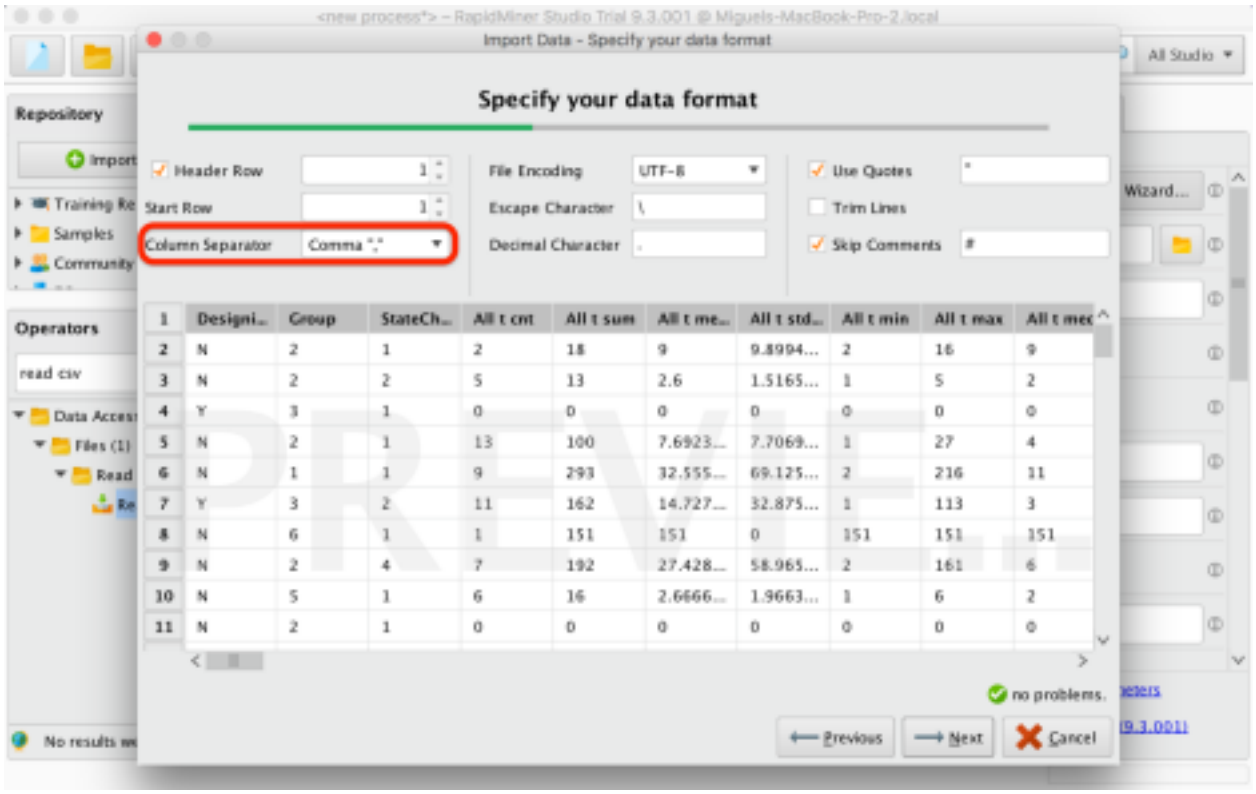
and click Next.



RapidMiner Walkthrough

EDUC 6191, Fall 2022

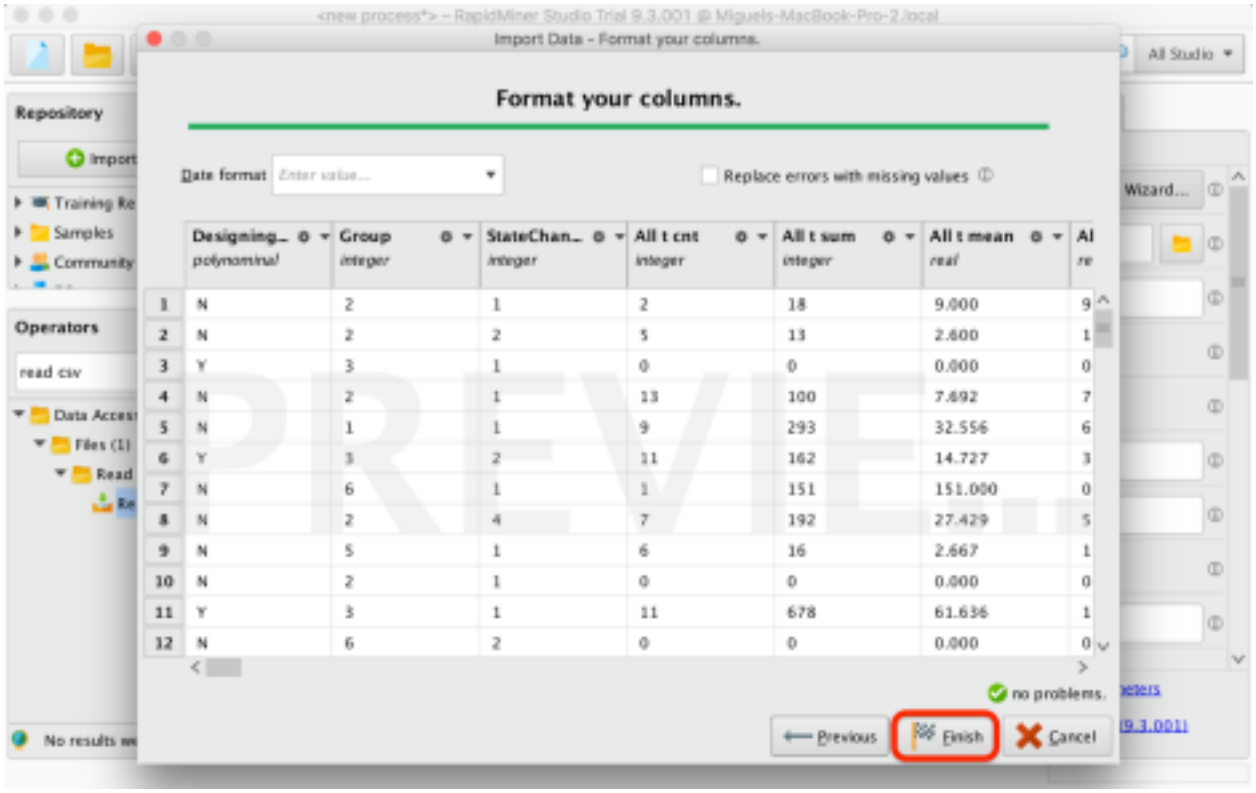
6. Since we are using a csv file, ensure that `Comma ","` is the selected Column Separator. Click Next.



RapidMiner Walkthrough

EDUC 6191, Fall 2022

7. Click Finish in the next window.

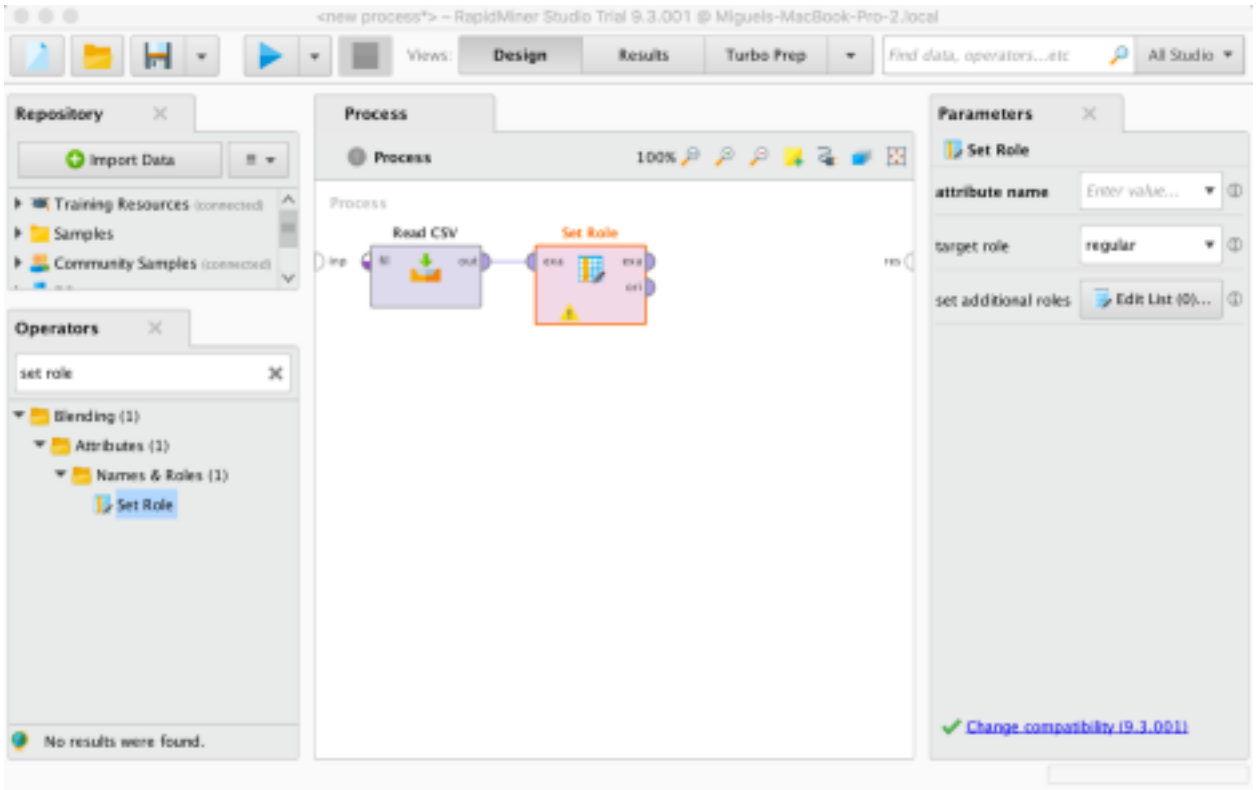


RapidMiner Walkthrough

EDUC 6191, Fall 2022

8. Type "`set role`" in the Operators search bar and create a new "`Set Role`" operator by dragging it into your Process window.

   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:

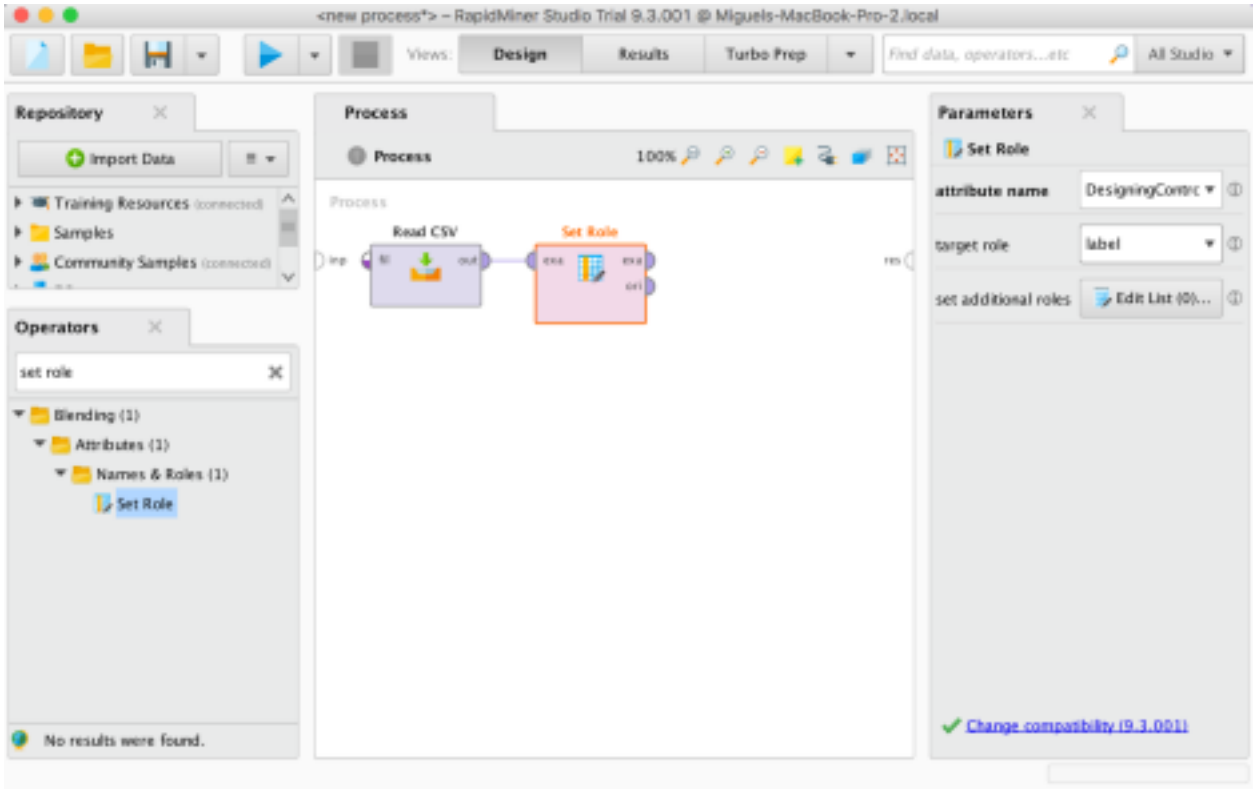


RapidMiner Walkthrough

EDUC 6191, Fall 2022

9. Now go over to the Parameters window on the right side of the interface and select `DesigningControlledExperiments?` as the variable you want to change.

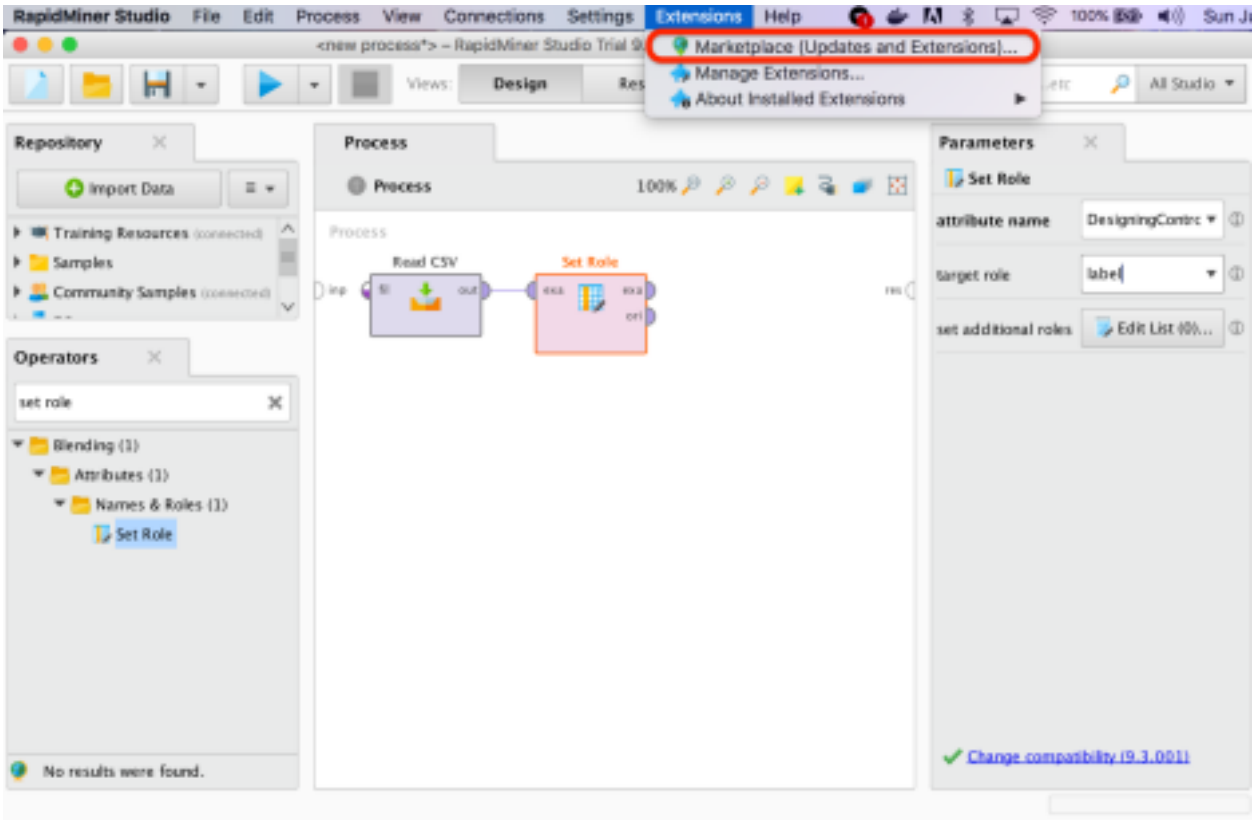
   Set it to be a "`label`" in the target role box.



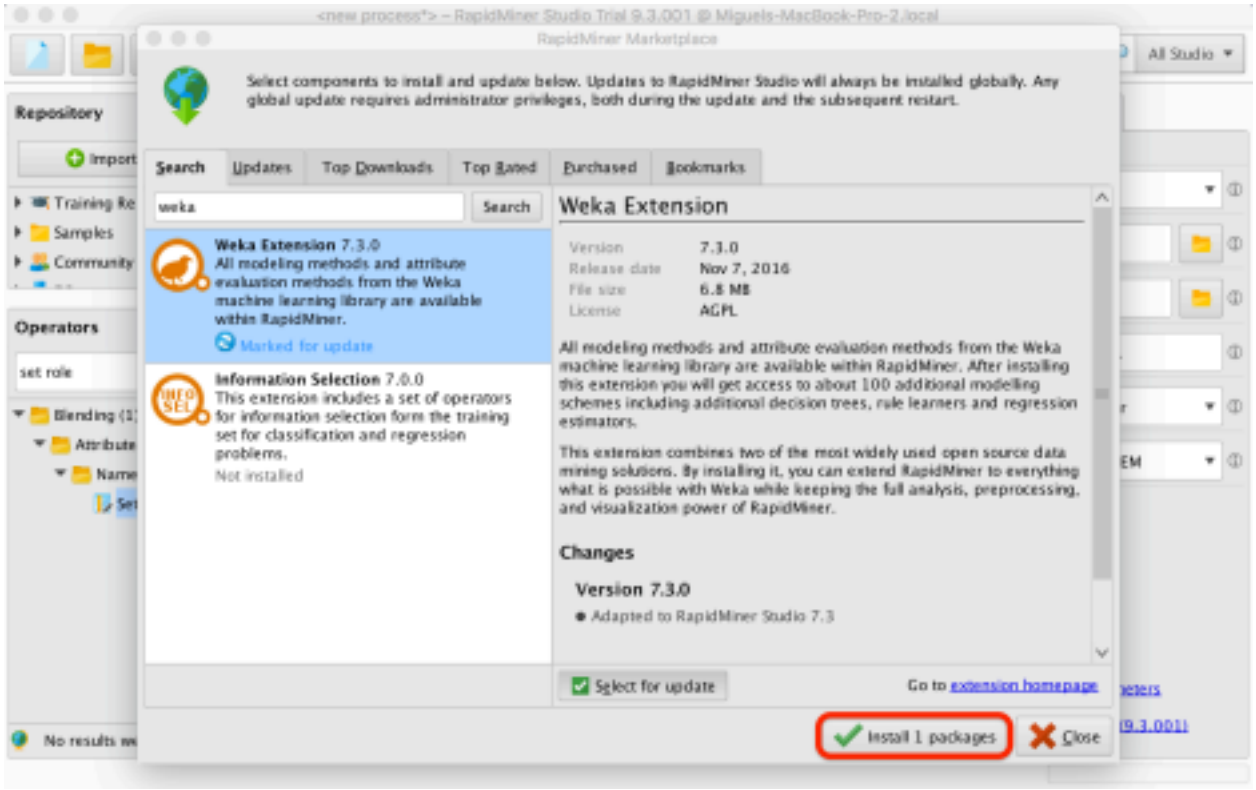
RapidMiner Walkthrough

EDUC 6191, Fall 2022

10. Install the WEKA Extension. To do this go to the `Extensions` menu, and select `Marketplace (Updates and Extensions)`.



Search for "`weka`" and install `Weka Extension`. When prompted, restart RapidMiner. Don't forget to save your work before restarting!

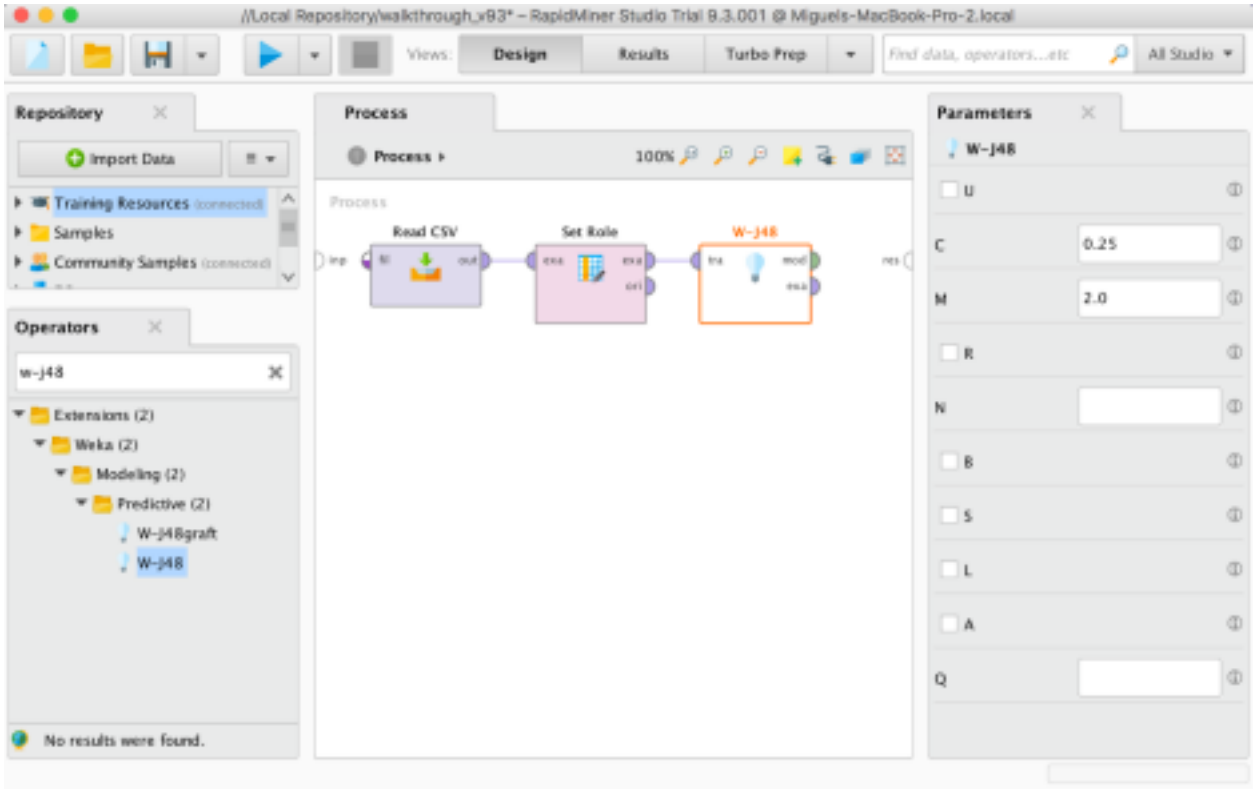

<new process*> – RapidMiner Studio Trial 9.3.001 @ Miguels-MacBook-Pro-2.local
RapidMiner Marketplace
Select components to install and update below. Updates to RapidMiner Studio will always be installed globally. Any global update requires administrator privileges, both during the update and the subsequent restart.
Search
Updates
Top Downloads
Top Rated
Purchased
Bookmarks
weka
Search
Weka Extension 7.3.0
All modeling methods and attribute evaluation methods from the Weka machine learning library are available within RapidMiner.
Marked for update
Information Selection 7.0.0
This extension includes a set of operators for information selection form the training set for classification and regression problems.
Not installed
Weka Extension
Version 7.3.0
Release date Nov 7, 2016
File size 6.8 MB
License AGPL
All modeling methods and attribute evaluation methods from the Weka machine learning library are available within RapidMiner. After installing this extension you will get access to about 100 additional modelling schemes including additional decision trees, rule learners and regression estimators.
This extension combines two of the most widely used open source data mining solutions. By installing it, you can extend RapidMiner to everything what is possible with Weka while keeping the full analysis, preprocessing, and visualization power of RapidMiner.
Changes
Version 7.3.0
Adapted to RapidMiner Studio 7.3
Select for update
Go to extension homepage
Install 1 packages
Close
Repository
Operators
All Studio

RapidMiner Walkthrough

EDUC 6191, Fall 2022

11. Type "`w-j48`" in the Operators search bar and create a new "`W-J48`" operator by dragging it into your Process window.

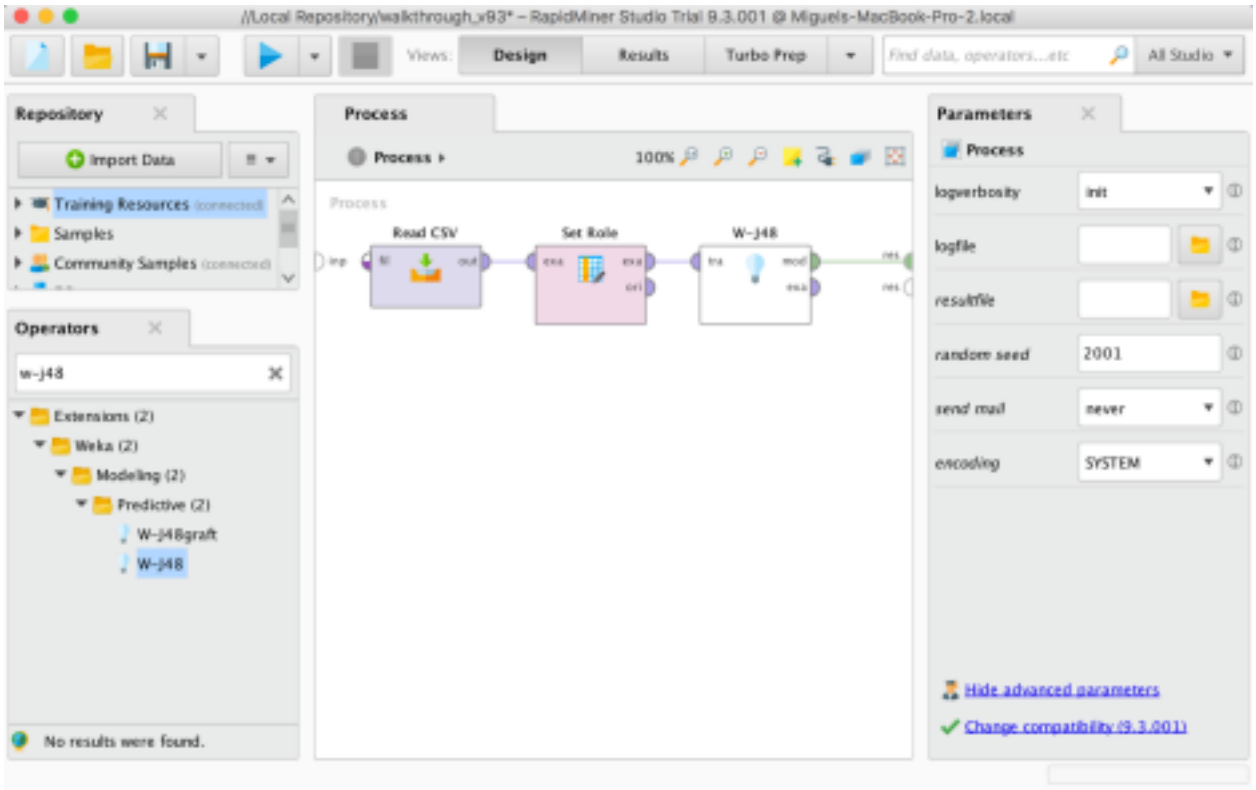
    Connect the output bubble of Set Role (`exa` for example set) to the input bubble of W-J48 (`tra` for training set).



RapidMiner Walkthrough

EDUC 6191, Fall 2022

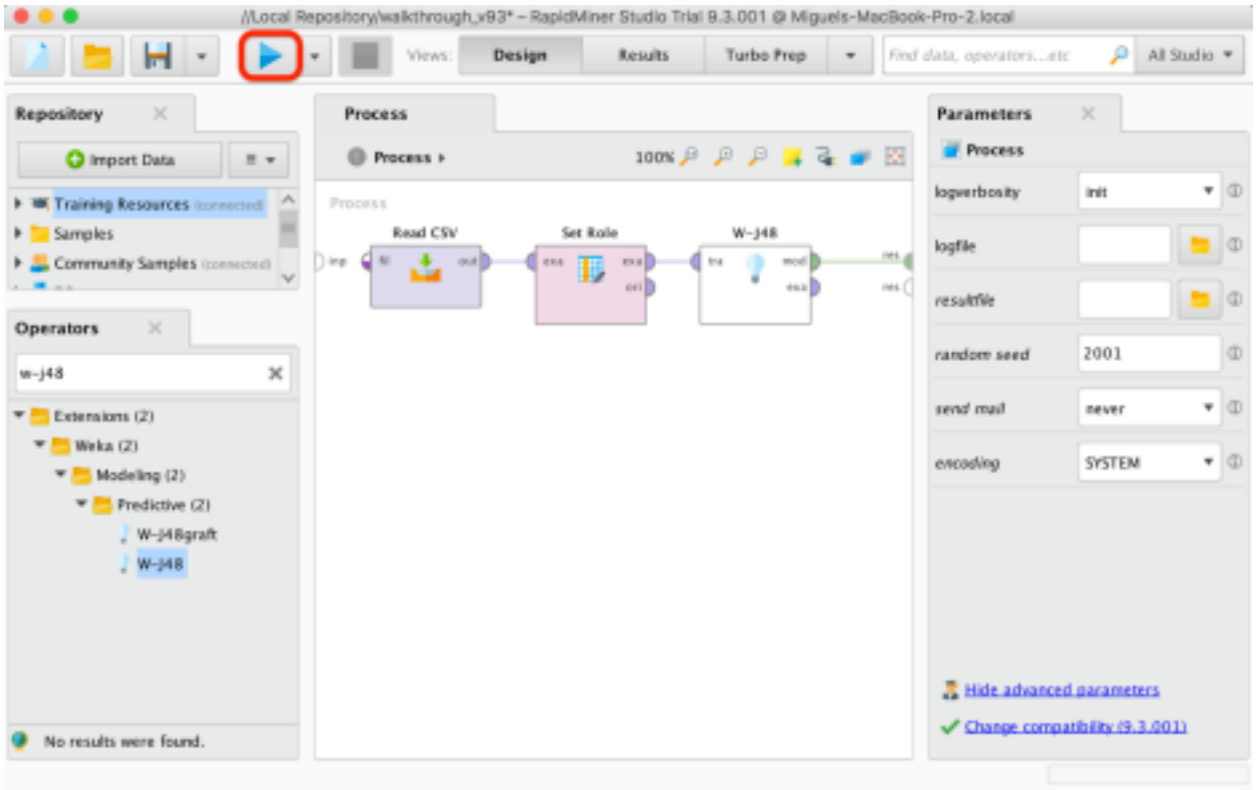
12. Connect the output bubble from W-J48 (`mod` for model) to the `res` (result) bubble on the far right.



RapidMiner Walkthrough

EDUC 6191, Fall 2022

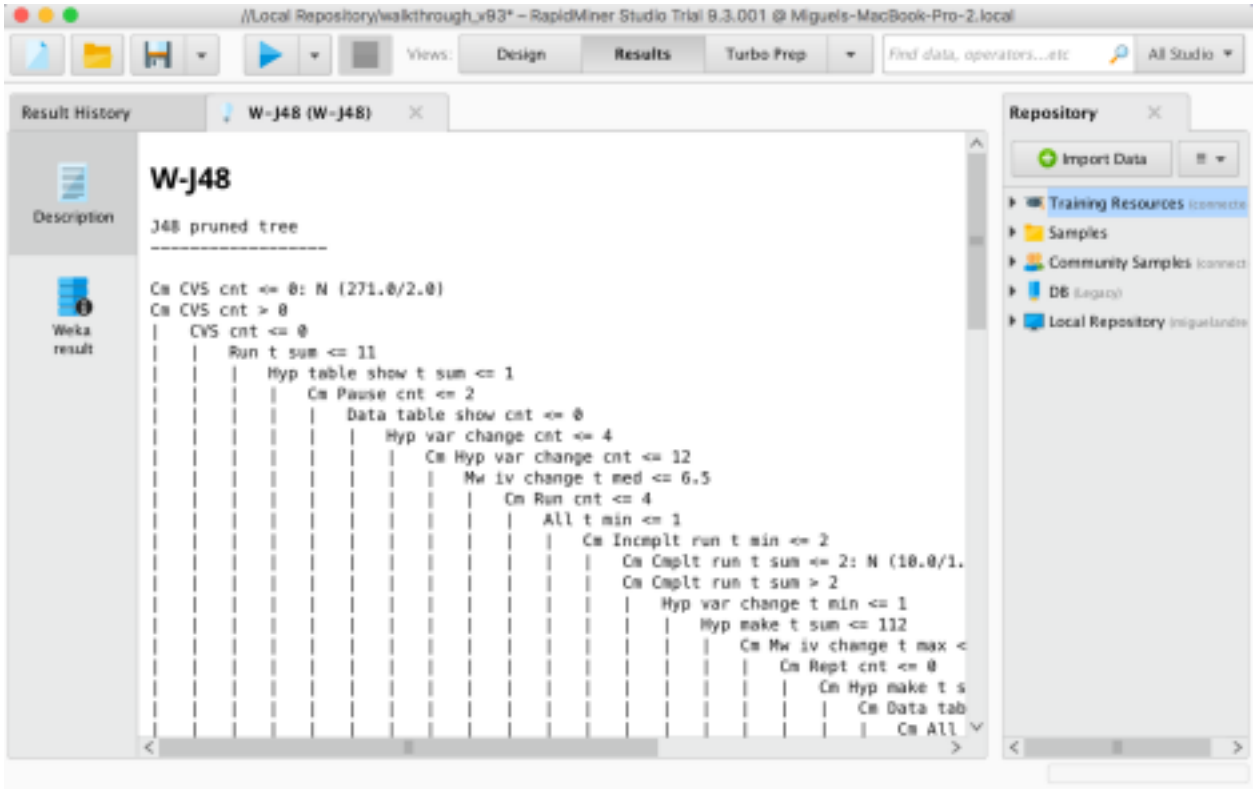
13. Then press play at the top of the screen to run your process.



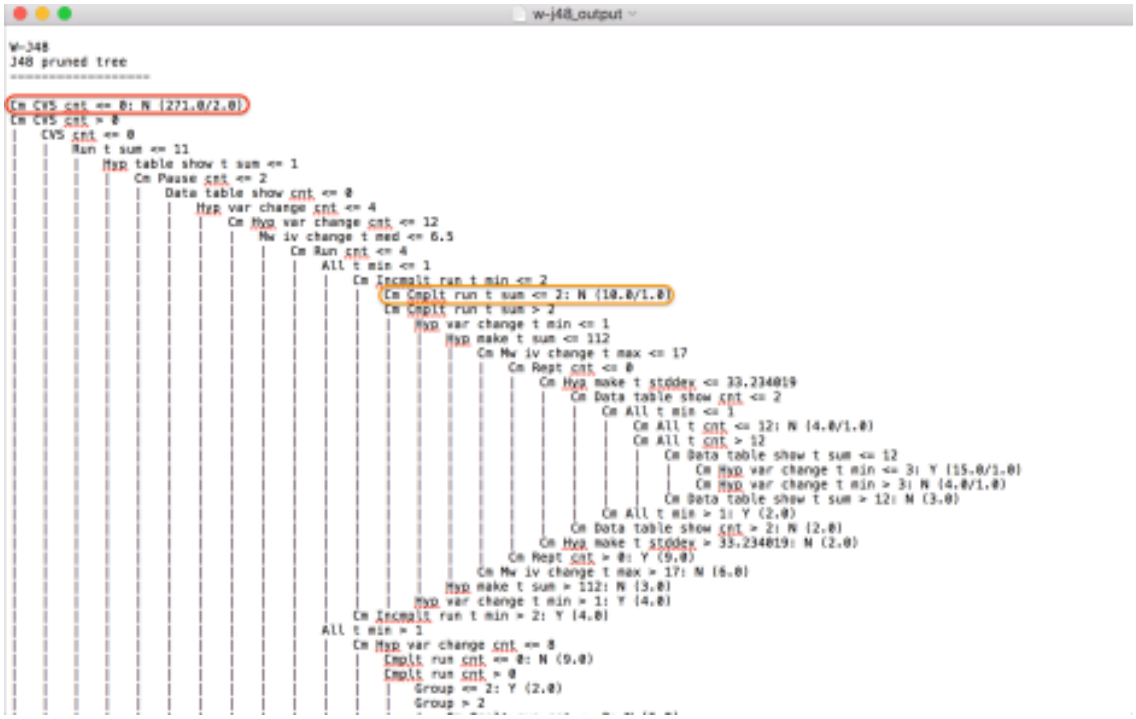
RapidMiner Walkthrough

EDUC 6191, Fall 2022

14. After a minute or so (possibly longer for slower computers), you should see your model.



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. (Encircled in red.)

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%. On the other hand, if the variable `Cm CVS cnt` is greater than zero, then the model goes to the next variable.

If the variable `CVS cnt` 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 (encircled in orange) with 10/11 = 90.9% confidence

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

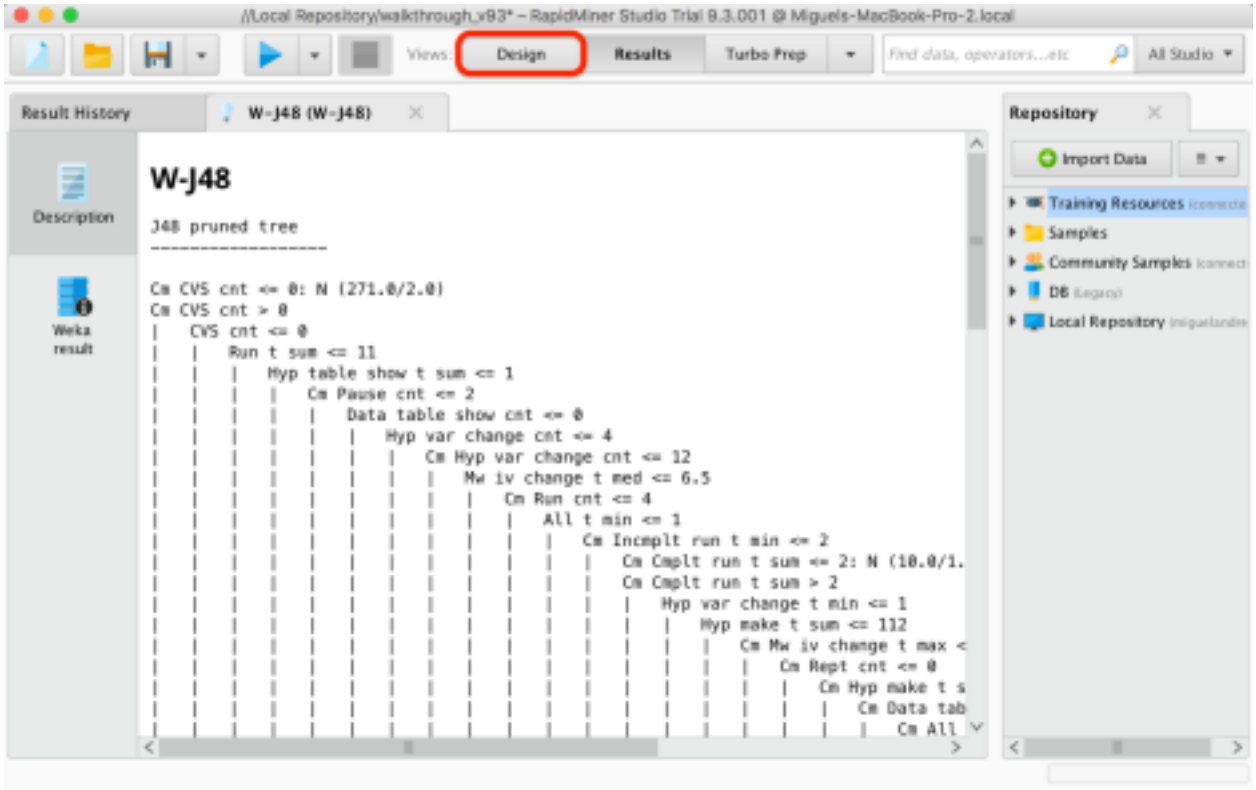
RapidMiner Walkthrough

16. Note that J-48 decision trees are extremely complicated to think through all at once. And they are one of the simpler algorithms to interpret!

RapidMiner Walkthrough

EDUC 6191, Fall 2022

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



RapidMiner Walkthrough

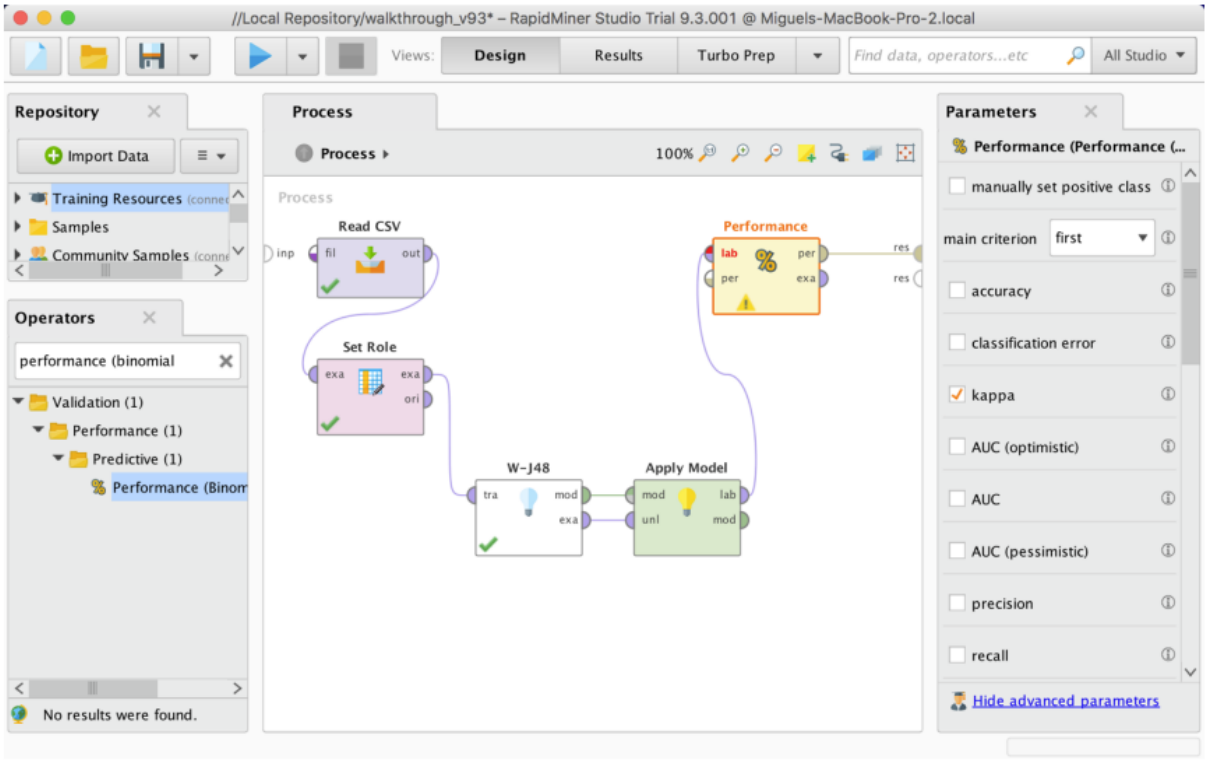
EDUC 6191, Fall 2022

18. Now add two more operators to the right of W-J48. First, an `Apply Model`, and second, a `Performance (Binomial Classification)`.

    Select the Performance operator, and choose `kappa` in the Parameters window on the right.

    Link the operators as shown below. You can delete a link by hovering over it and clicking the big red X that appears.

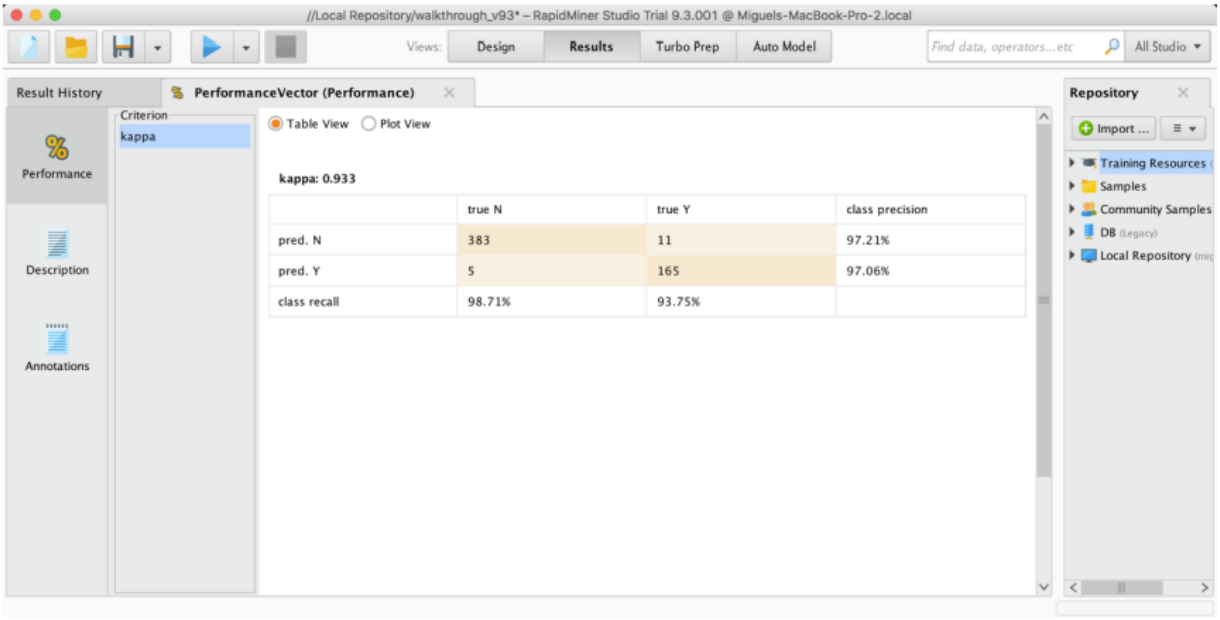
    Finally, press play to re-run your process.



RapidMiner Walkthrough

EDUC 6191, Fall 2022

19. You should see this screen:



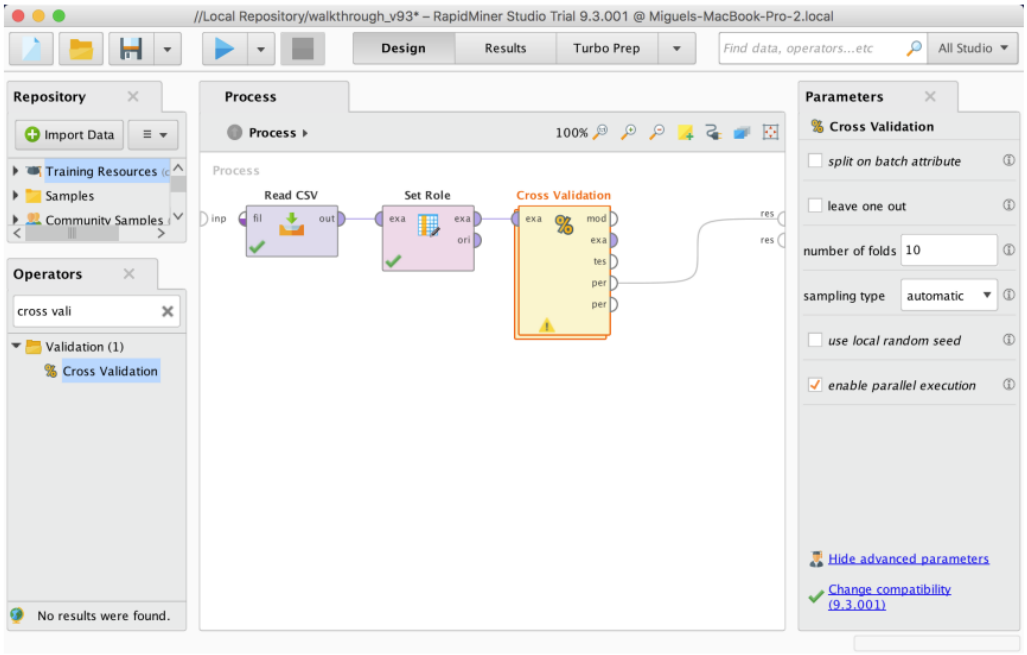
This shows you the model's kappa and confusion matrix. The kappa is excellent; too good, in fact. Keep in mind we did not use cross-validation, so this model is being trained and tested on the same data set.

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".

RapidMiner Walkthrough

EDUC 6191, Fall 2022

20. Go back to the main screen by clicking `Design`, and create the process you see below.



You should delete `W-J48`, `Apply Model`, and `Performance`.
You should add `Cross Validation`.

You will get some error messages – don't worry about them for now.

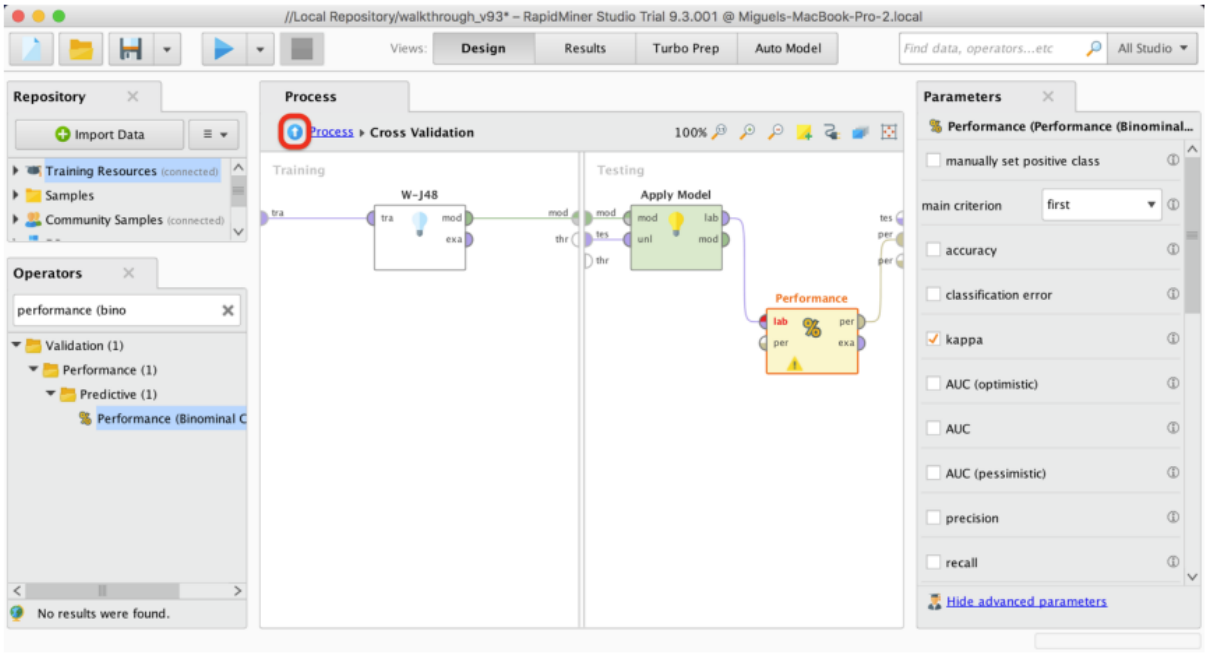
Note that in many cases, you'll want to do a *batch* cross-validation instead of a normal cross-validation. Both are done through the "`Cross Validation`" operator. Batch cross-validation allows you to do student-level cross-validation, or item-level cross validation, or population-level cross-validation. Regular cross-validation supports flat cross-validation, as talked about it the lecture video.

The parameter options on the right side of the interface, which allow you to do: 1. Flat k-fold cross-validation; currently set-up to do `10`-fold cross-validation, 2. Leave-one-out cross-validation, and
3. Batch cross-validation, through the use of a `batch` attribute.

RapidMiner Walkthrough

EDUC 6191, Fall 2022

21. Now double-click on the `Cross Validation` operator (the tall yellow one). It will bring you to another screen. Add operators as shown here – the same ones you just deleted. Make sure you select `kappa`.



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

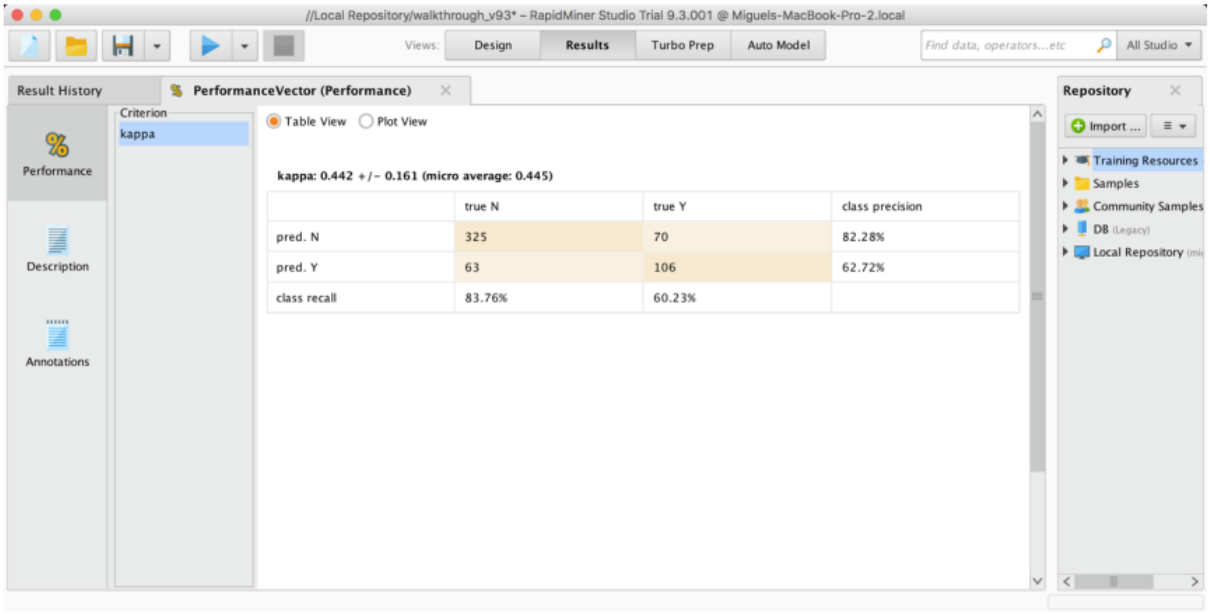
The right box represents what you do with the test folds – apply the model and see how well it does.

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

RapidMiner Walkthrough

EDUC 6191, Fall 2022

23. Click on the play button to re-run your process. You should get the results shown below. Note that kappa is a lot lower when cross-validating.



RapidMiner Walkthrough

EDUC 6191, Fall 2022

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
- 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!