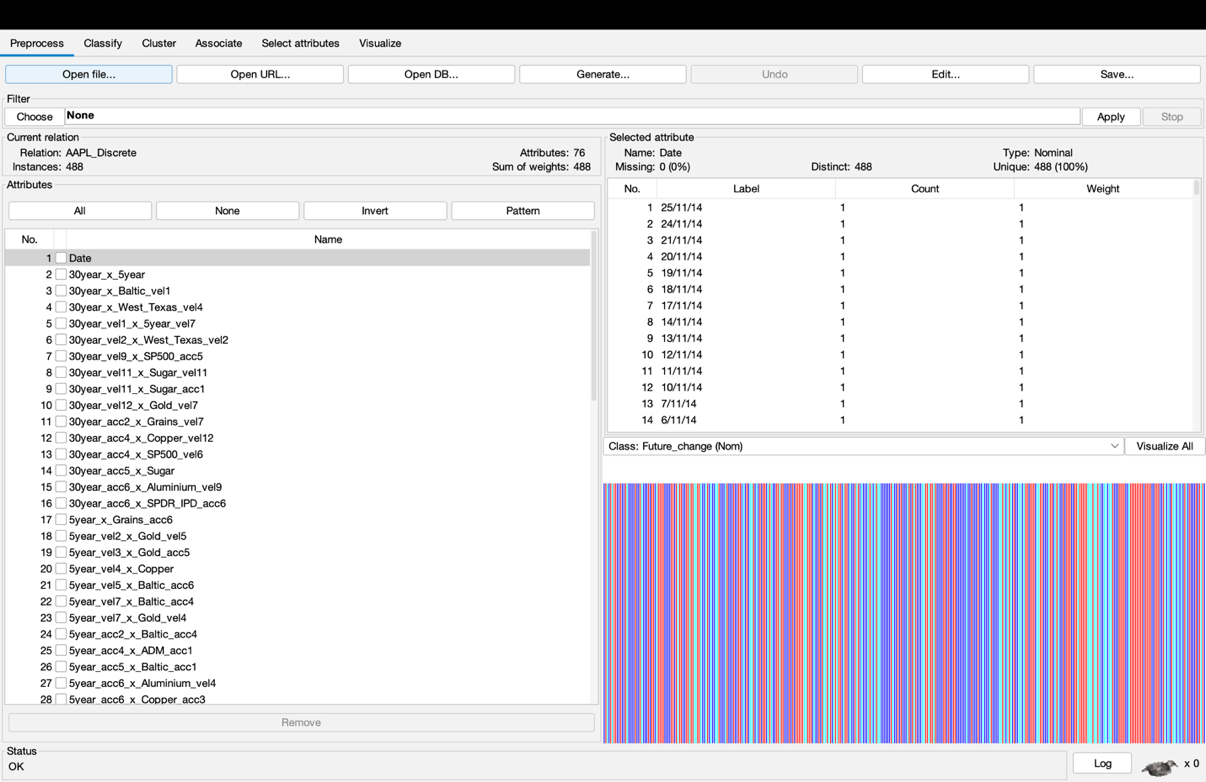
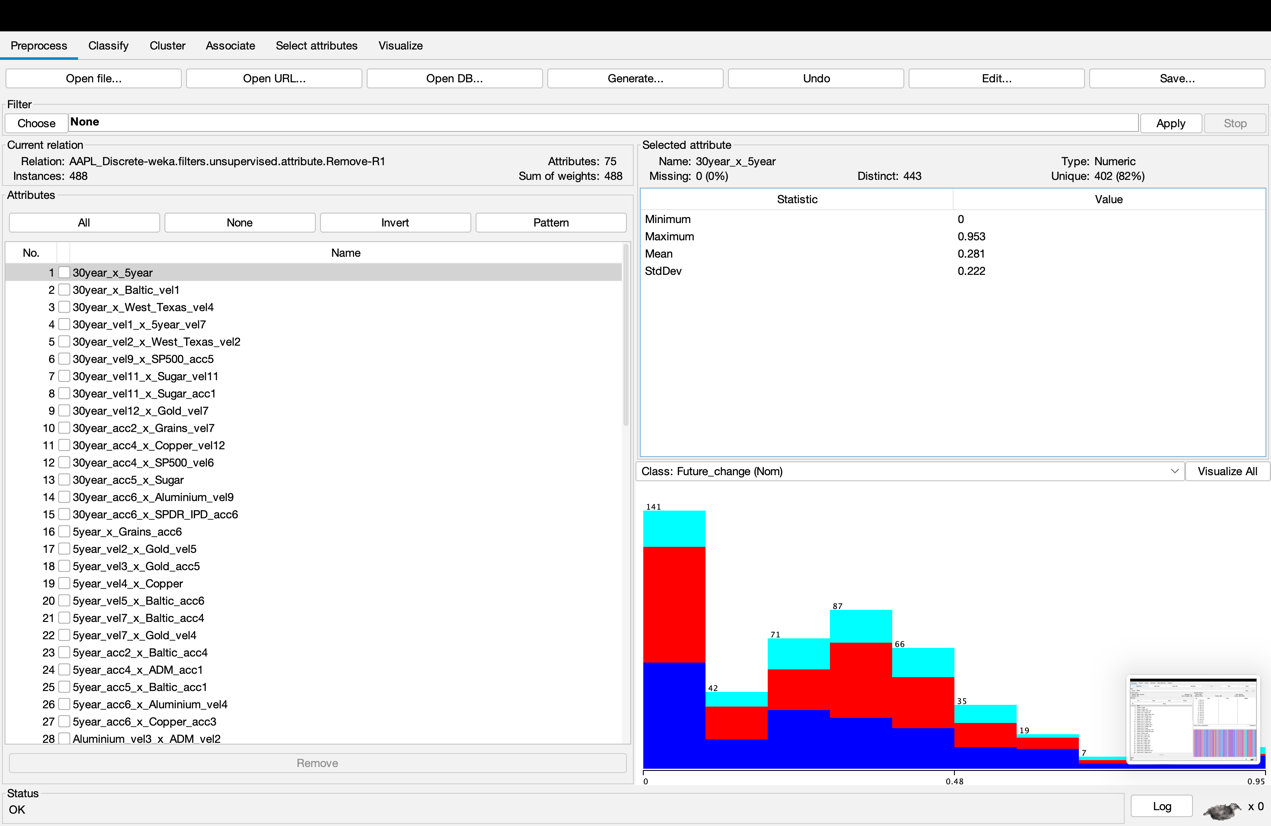
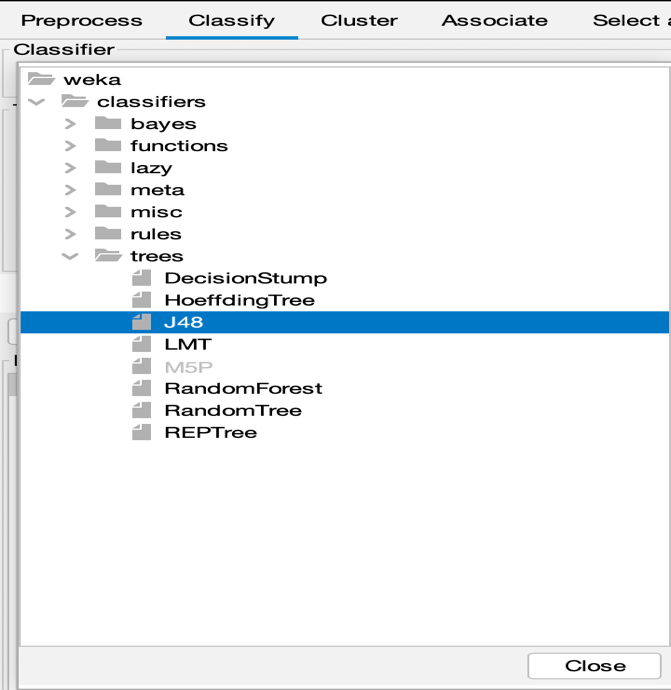
1. Start up Weka on the discrete file, AAPL\_discrete.arff



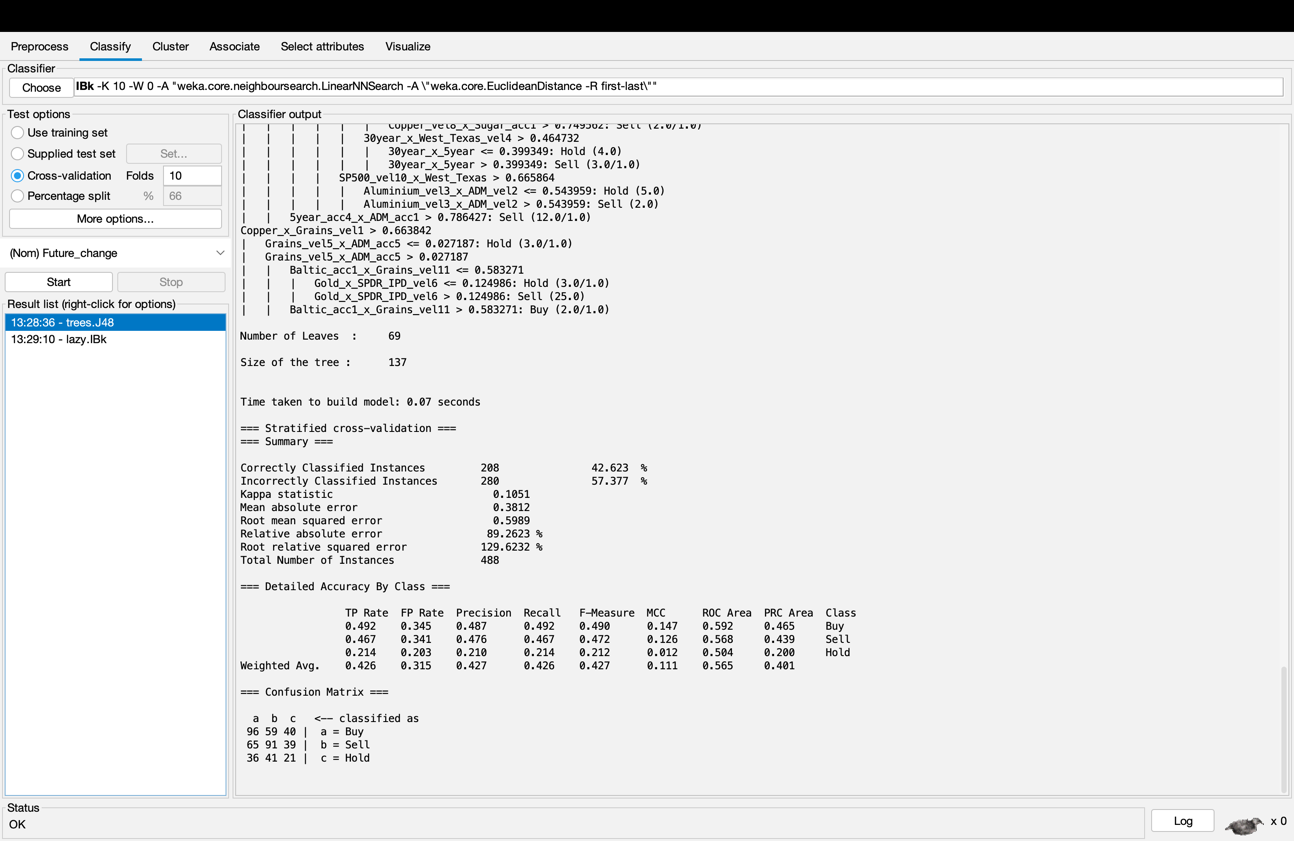
1. First, the dates don’t tell us anything, so we can delete that whole column. Using the Preprocess tab, click on the box that says “Date”, and then click at the bottom on “Remove”.



1. Then, click on the “Classify” tab, so we can choose classifier algorithms. The first one to try would be J48, a decision tree algorithm.



1. Scroll down to the bottom of the output, where it talks about the True Positive rate (TP rate) and all those other measures of accuracy from lecture in Week 4.



1. Pick a measure that you think you need to maximize or minimize:

If you want to predict a share price rise, maximize TP rate for Buy.

Logistic algorithm from functions has the highest TP rate for Buy

If you want to predict a share price fall, maximize TP rate for Sell.

Logistic algorithm from functions has the highest TP rate for Sell.

If you want to avoid errors in predicting a rise, minimize FP rate for Buy.

Logistic algorithm from functions has the lowest FP rate for Buy.

If you want general accuracy, maximize ROC Area for Weighted Avg.

Logistic algorithm from functions has the highest ROC Area for Weighted Avg.

If it’s okay to miss some buying opportunities, but the ones you choose are more likely to be correct, then maximize Precision for Buy.

Logistic algorithm from functions has the highest Precision for Buy.

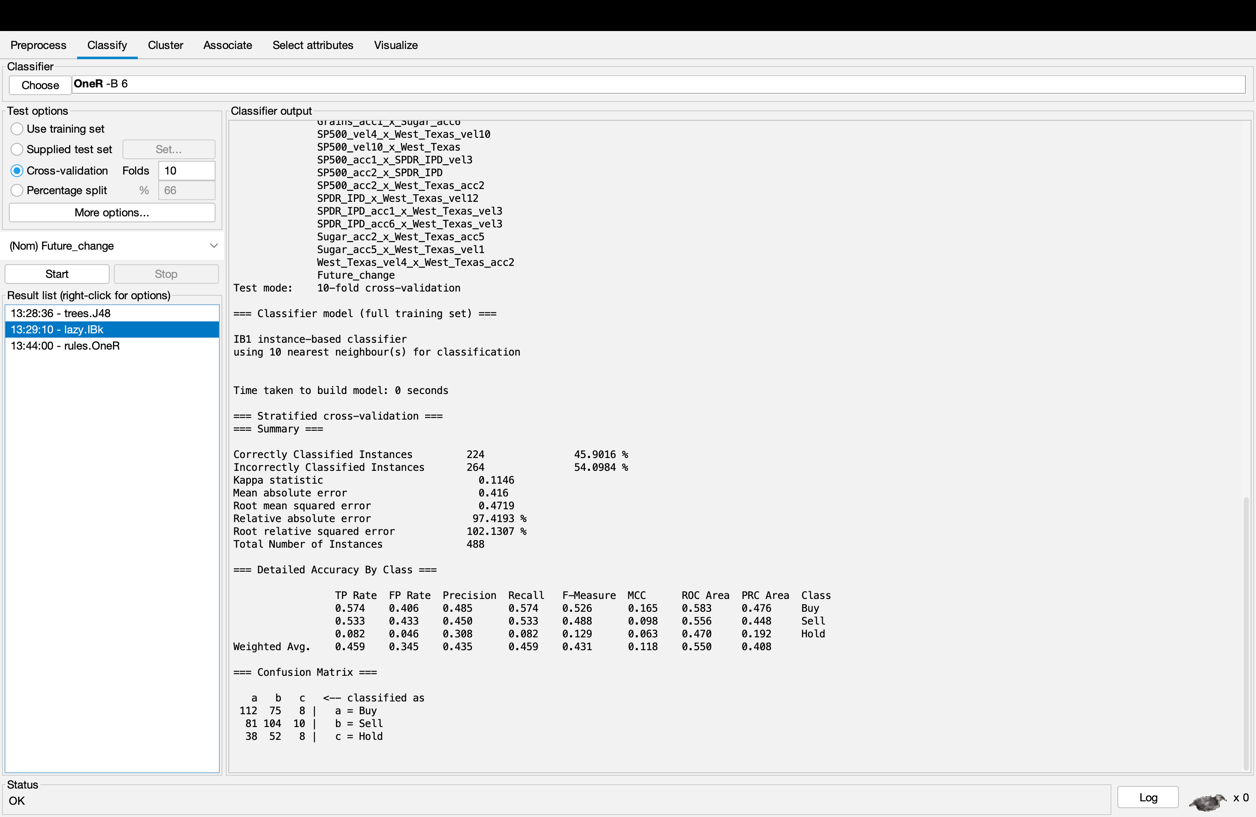
Therefore, for this data, I would recommend Logistic algorithm to Buy or Sell Apple’ shares.

1. Now try lots of classification algorithms. 1R is fast, but how accurate is it?



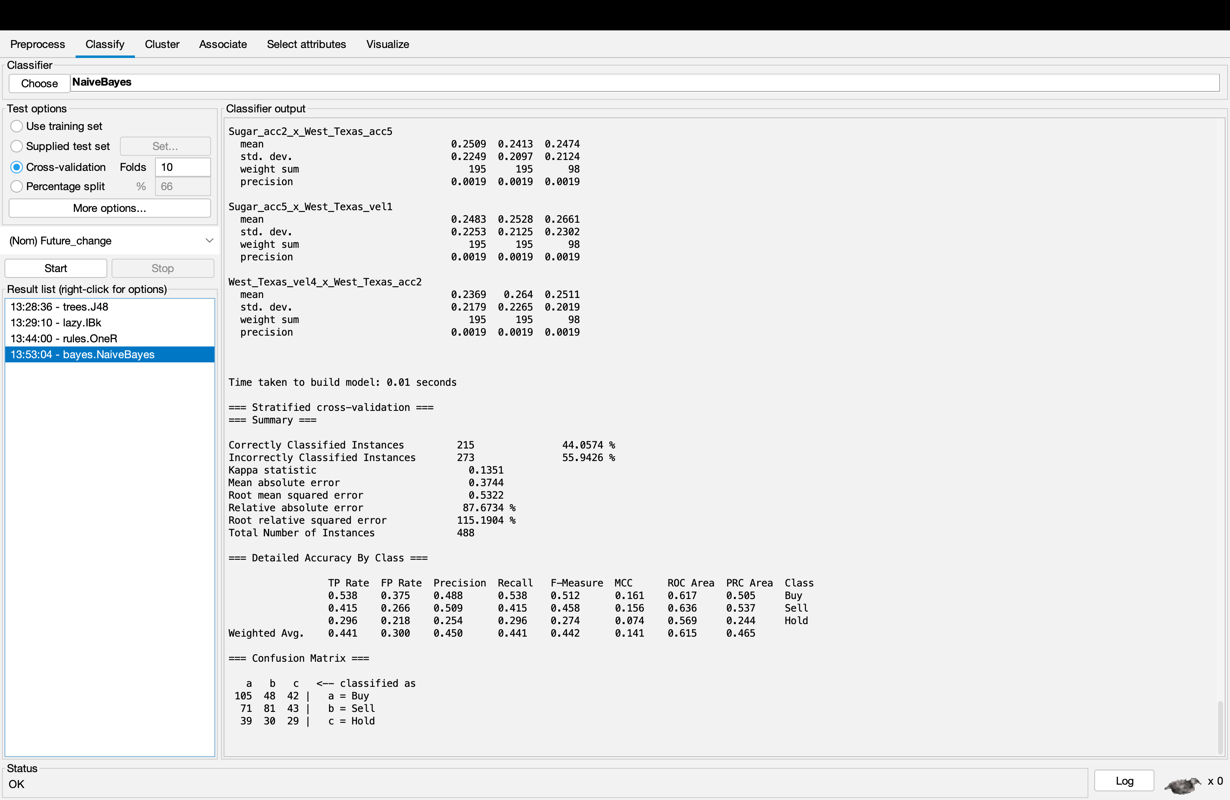
It is not that accurate and the accuracy of 1R is less than that of J48.

1. Another fast algorithm you should try is K Nearest Neighbour, one version is under the group **lazy** called **IBk** and you should change the value for K to 10.

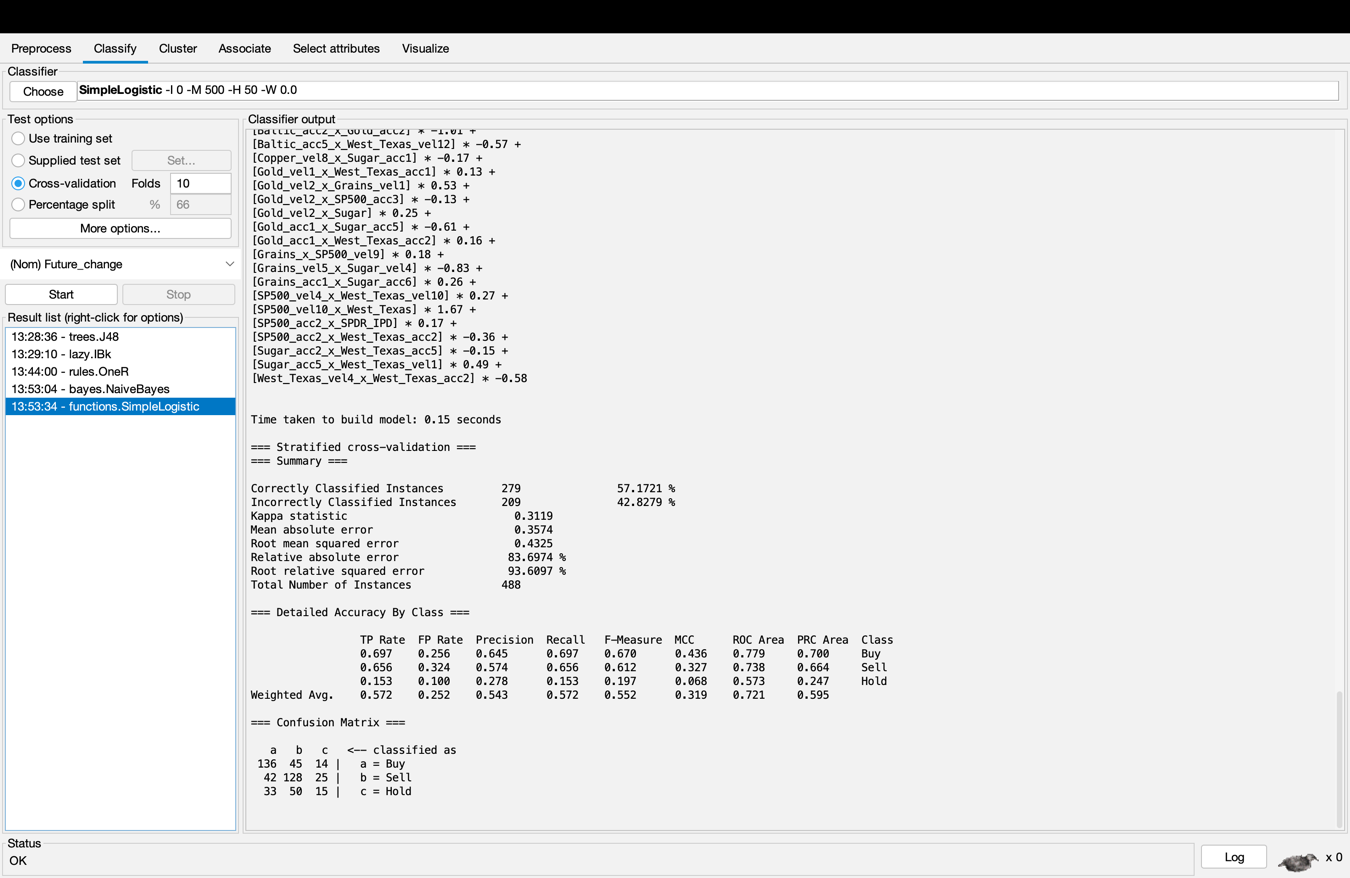


1. Some more fast algorithms you should try include:

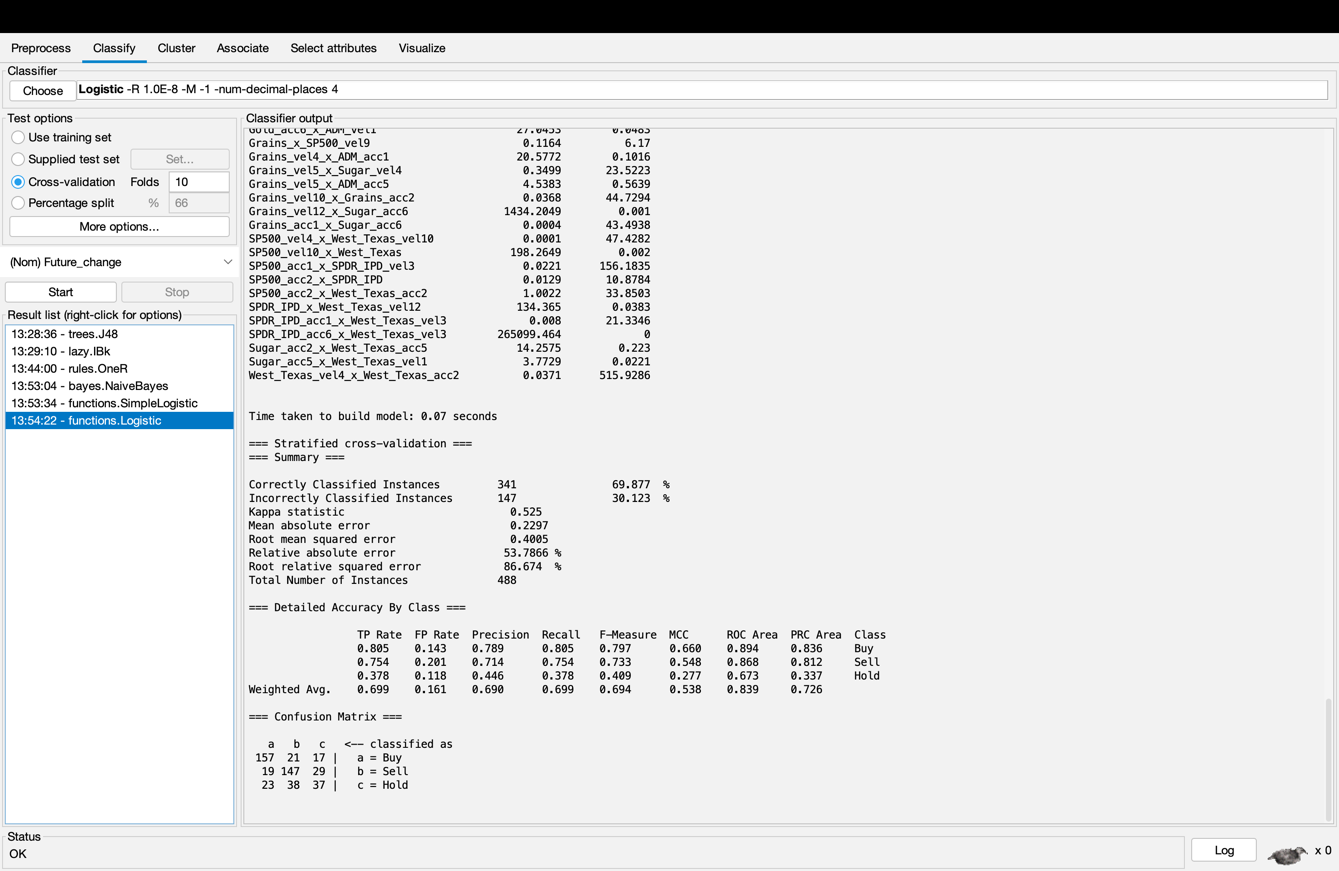
• bayes  Naïve Bayes Simple



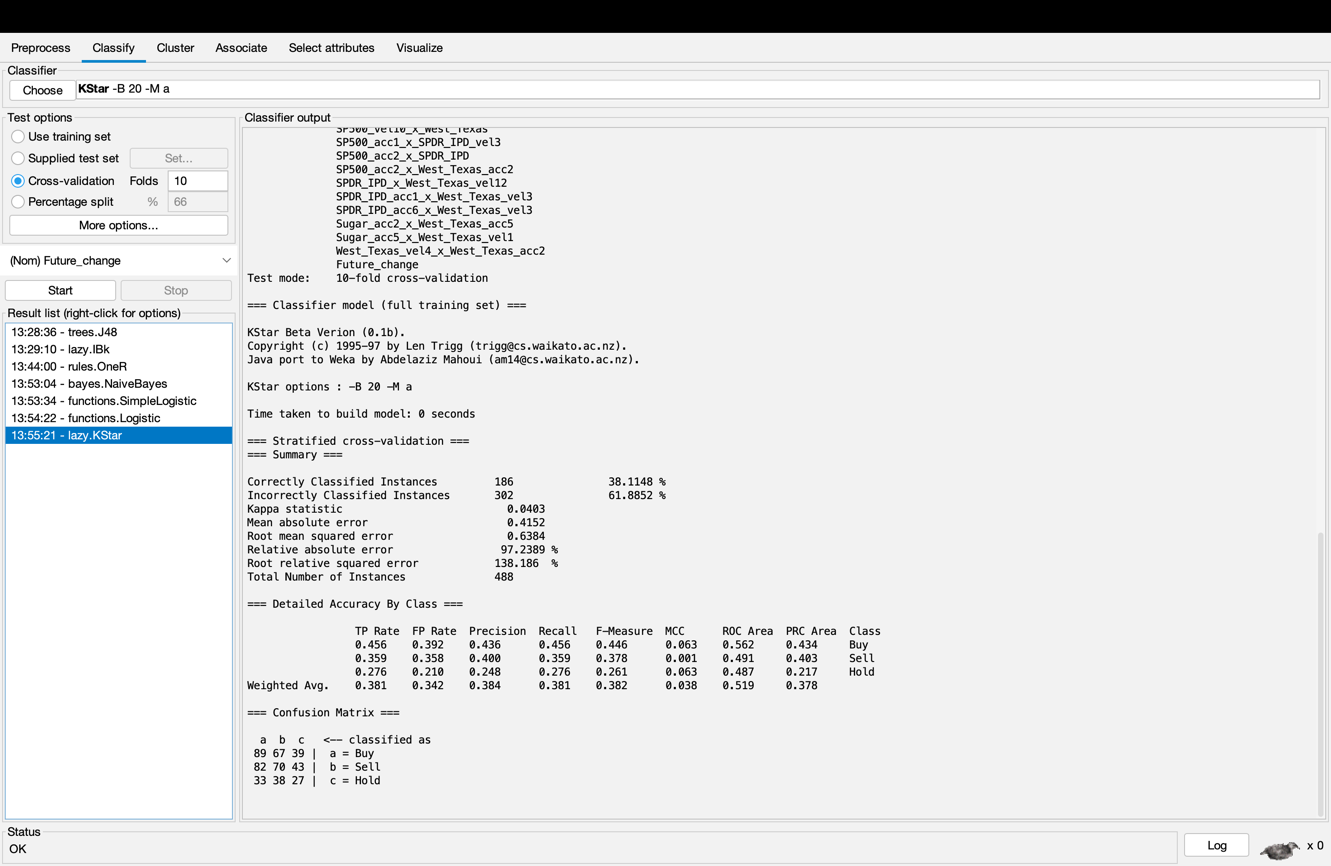
• functions  Simple Logistic



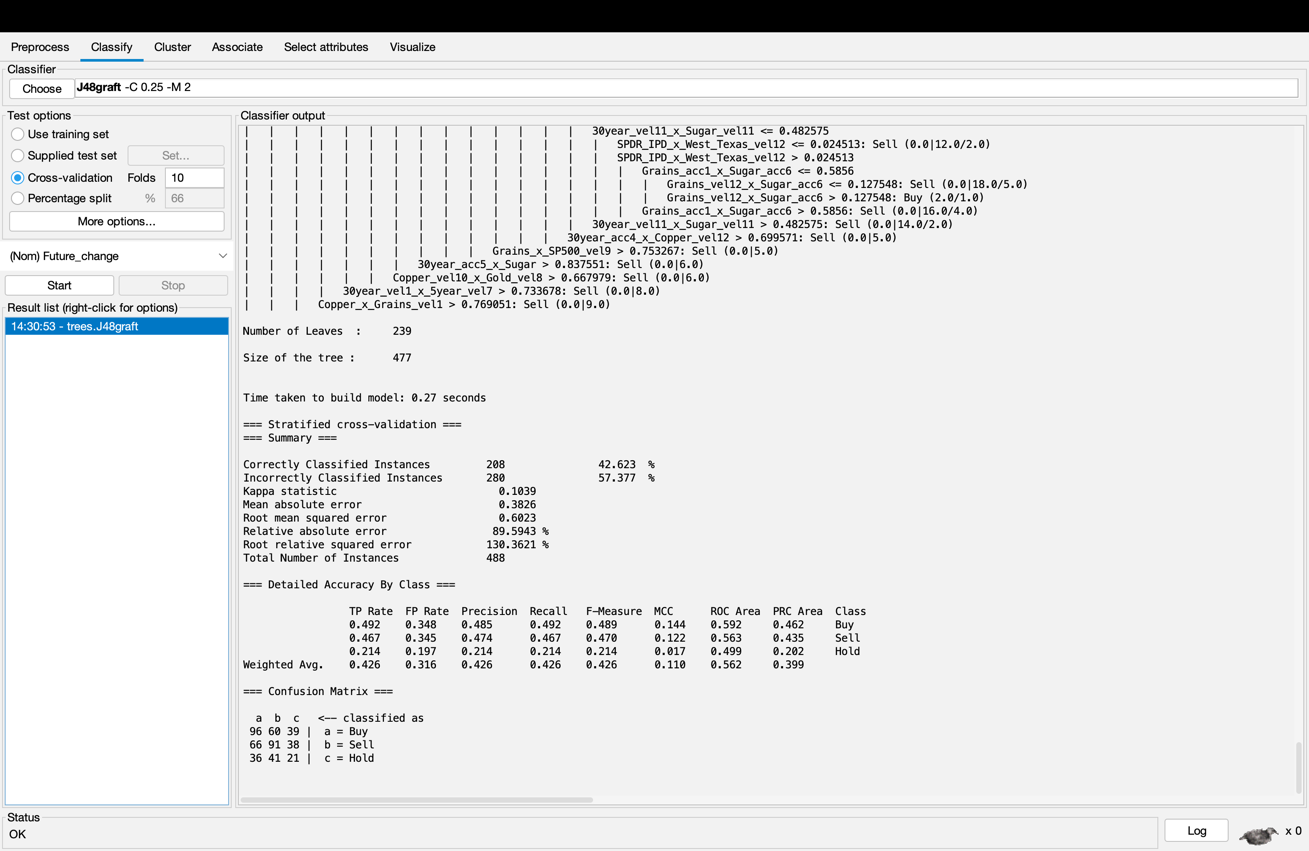
• functions  Logistic



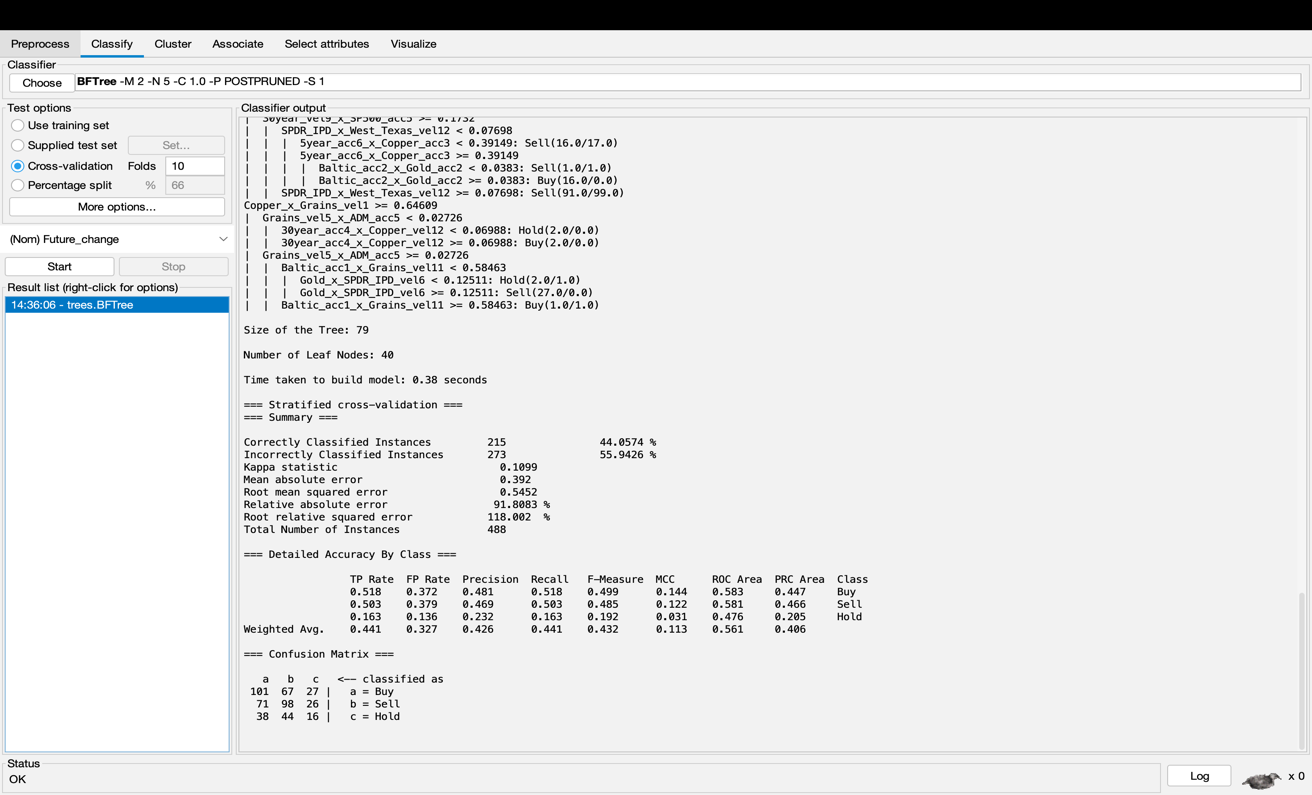
• lazy  Kstar (a more sophisticated K Nearest Neighbour algorithm)



• trees  J48graft (any better to J48?)



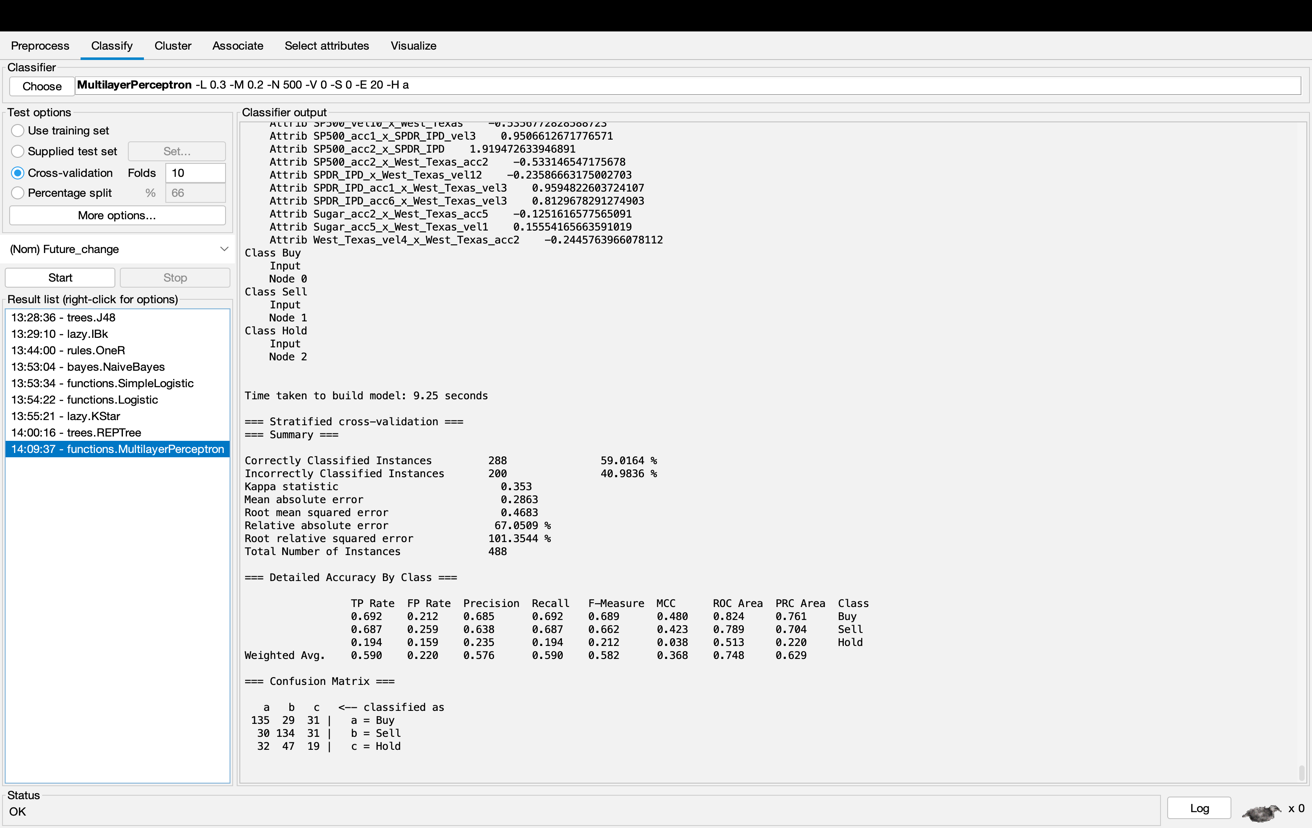
• trees  Bftree



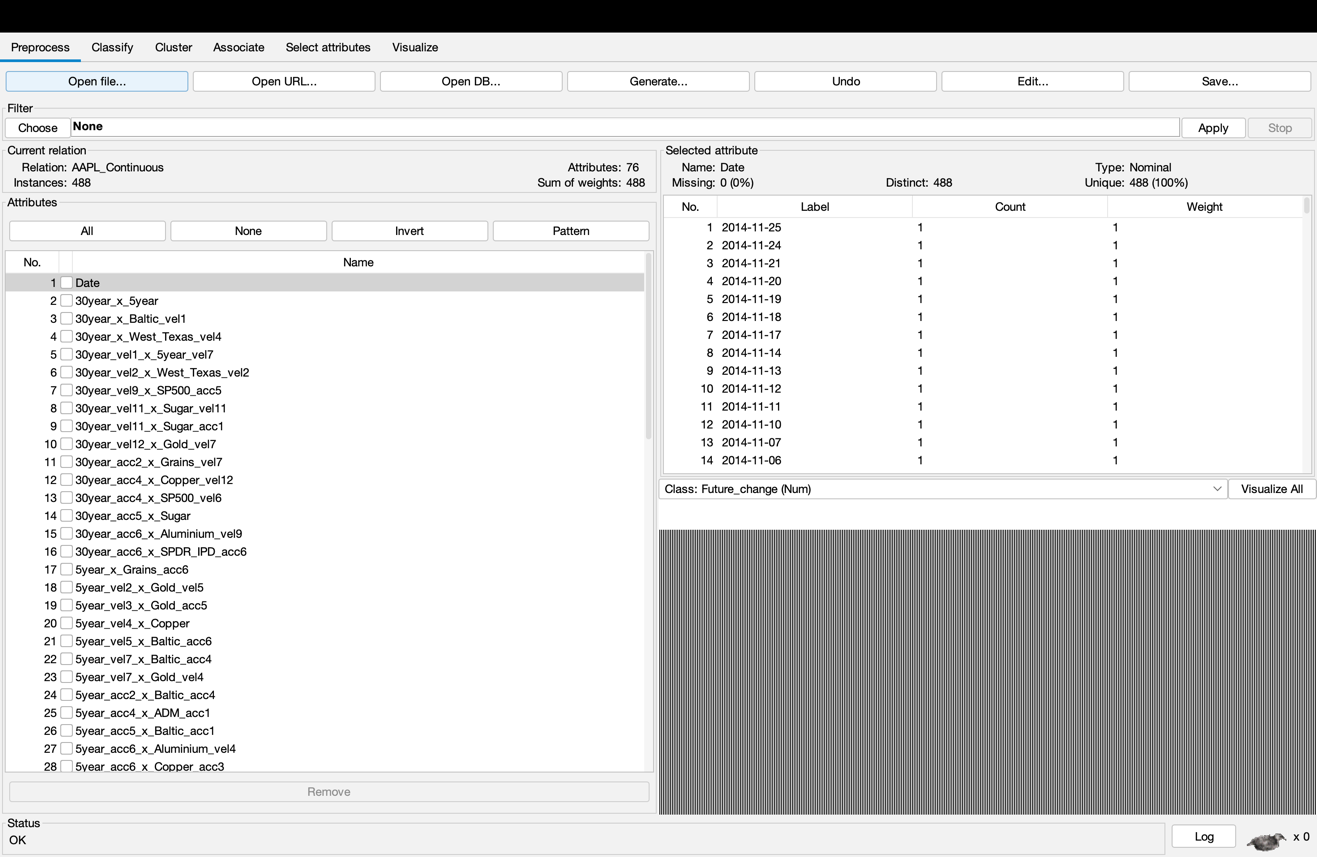
For some tree algorithms, right-click on the output, select “Visualize Tree”, then right-click to “Fit to screen”.

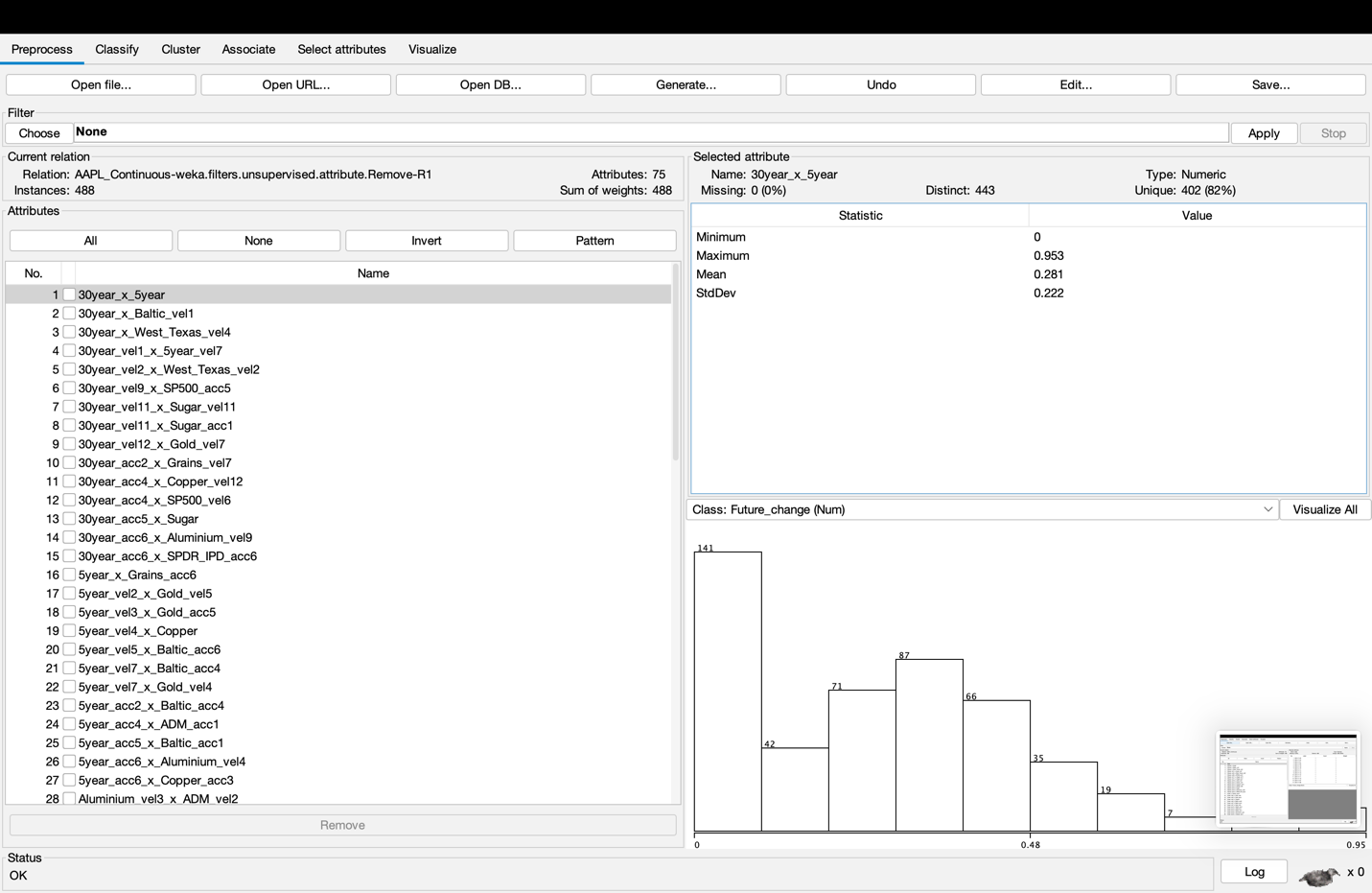


1. A slow algorithm is a neural network, try functions  Multilayer Perceptron.



1. Now load the file with continuous numbers as outputs, AAPL\_continuous.arff and delete the dates column as in step 2 above.





11. As there are no discrete outputs, there is no TP rate or any of those measures. Instead, there are some statistical measures such as:

• correlation (bigger is better)

Pace Regression algorithm has the biggest correlation coefficient.

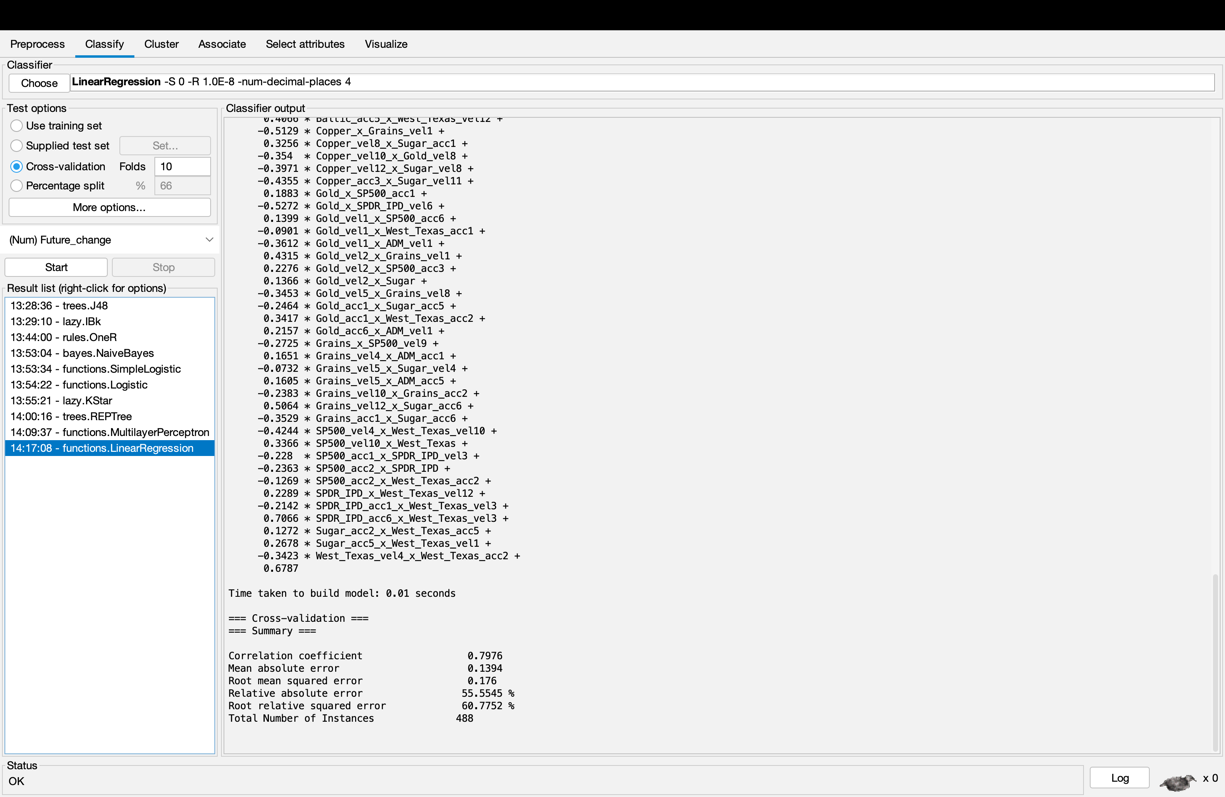
• relative absolute error (smaller is better).

Pace Regression algorithm has the smallest relative absolute error.

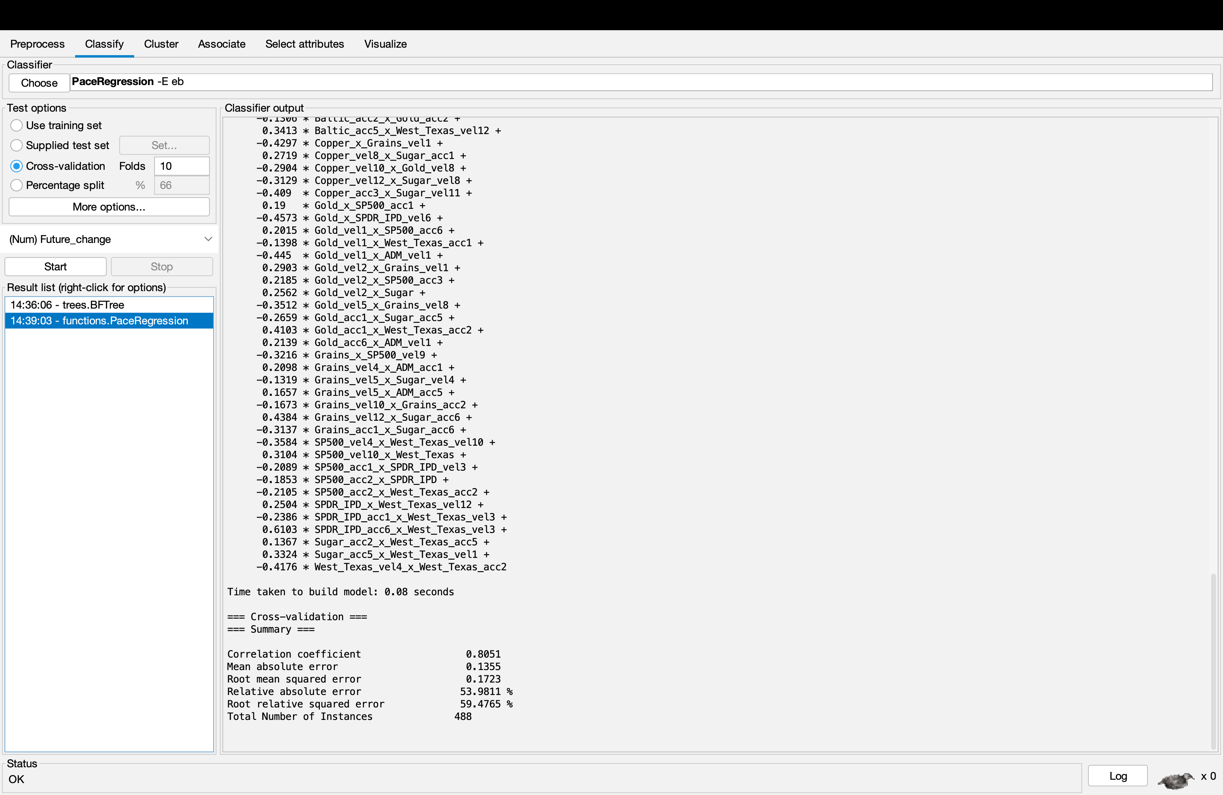
Therefore, I would recommend Pace Regression algorithm to calculate AAPL continuous.

Some algorithms only work with continuous output. You could try:

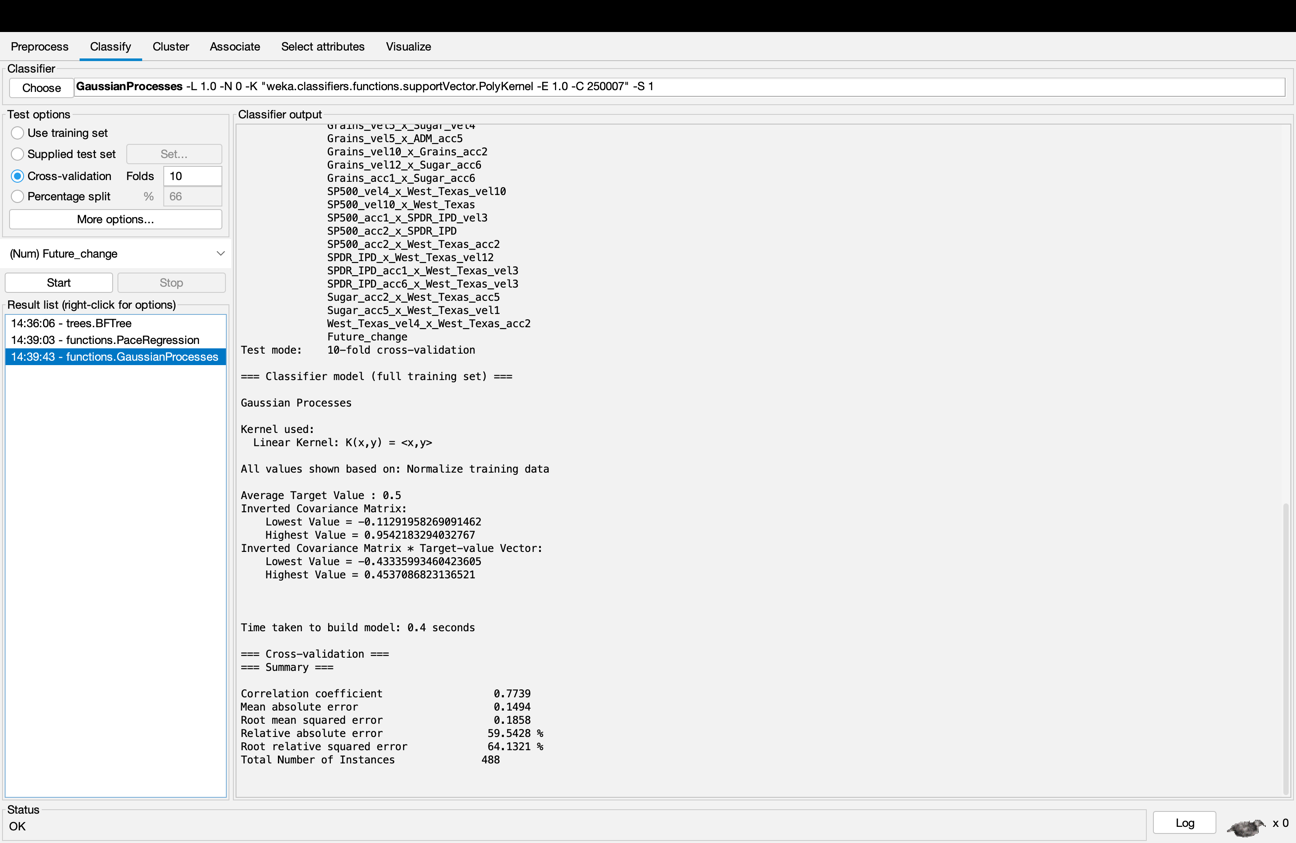
• functions  Linear Regression



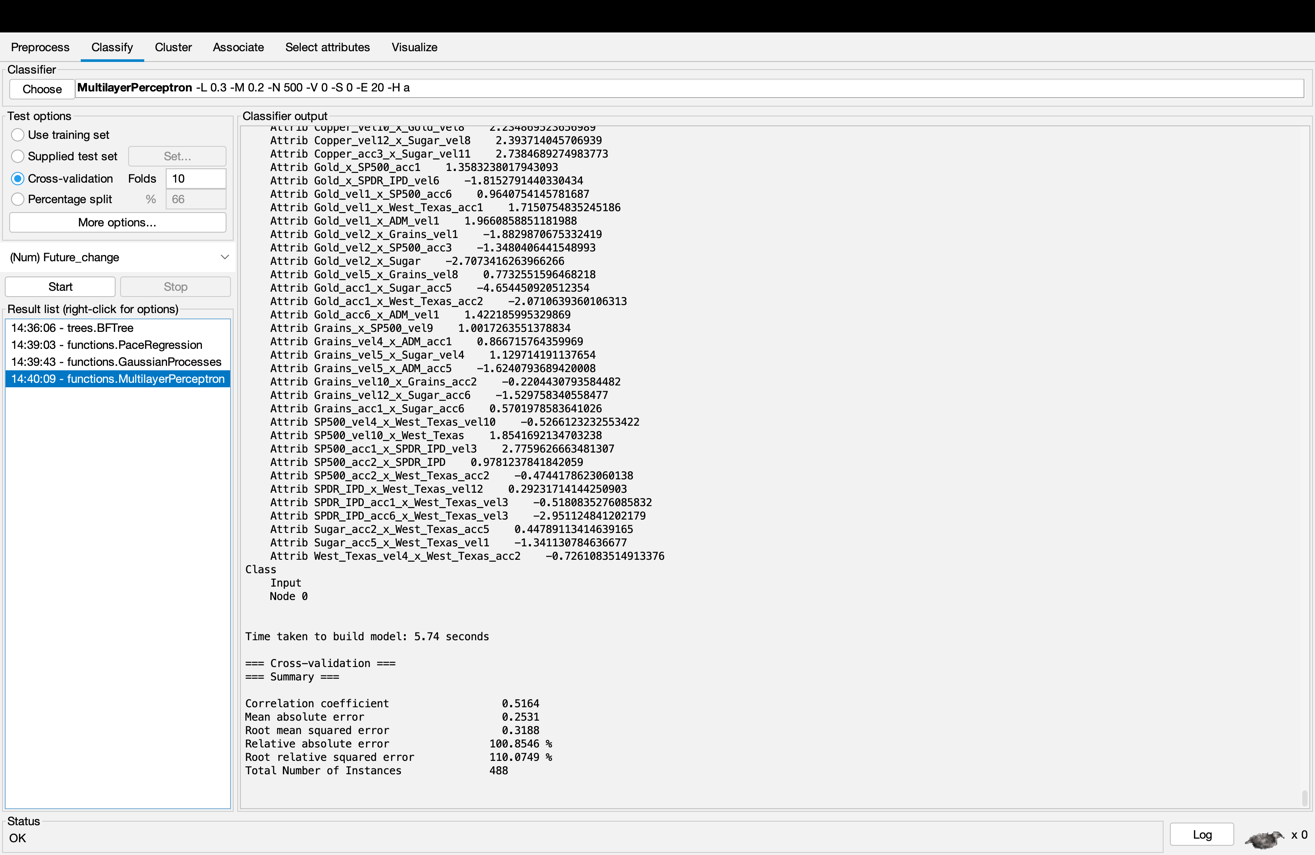
• functions  Pace Regression



• functions  Gaussian Processes



• functions  Multi Layer Perceptron (again, but output is continuous)



• lazy  IBk (k nearest neighbour again, but output is continuous)



1. So what algorithm(s) give the best results, on your chosen measure of accuracy? Try running them again, with different parameters.

When it comes to the measure of accuracy in discrete data, the logistic algorithm shows the best accuracy and when it comes to the measure of accuracy in continuous data, the pace regression algorithm shows the best accuracy.

For example, does K Nearest Neighbour work best with knn = 1 or 5 or 10 or 20 nearest neighbours?

When trying out IBK algorithm, K Nearest Neighbour work best with knn = 20. The error rate increase as the knn value increase. But for TP rate, knn = 5 has the highest rate for Buy and knn = 10 has the highest rate for Sell. I think knn = 20 works best due to having highest ROC Area which indicate high accuracy.