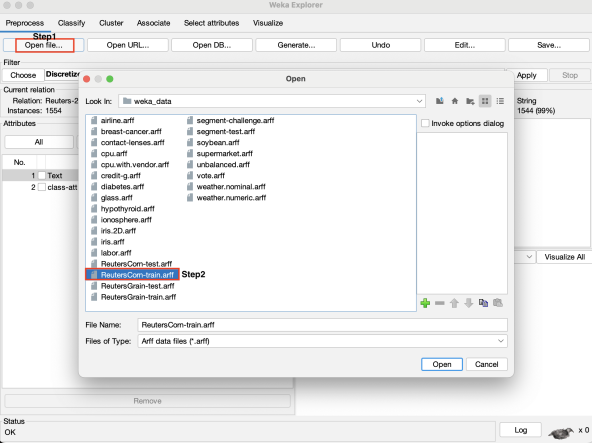
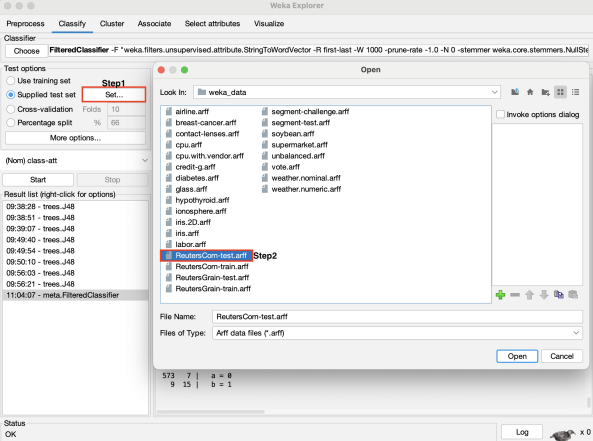
**資工三 黃偉祥 111010550**

1. **文本分類**

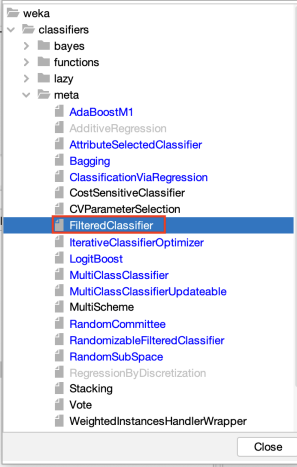
- 打開weka的Explorer 然後選擇 Step2 的檔案

****

- 接下來到Classify 視窗中選擇Supplied test set 並選擇Step2 檔案

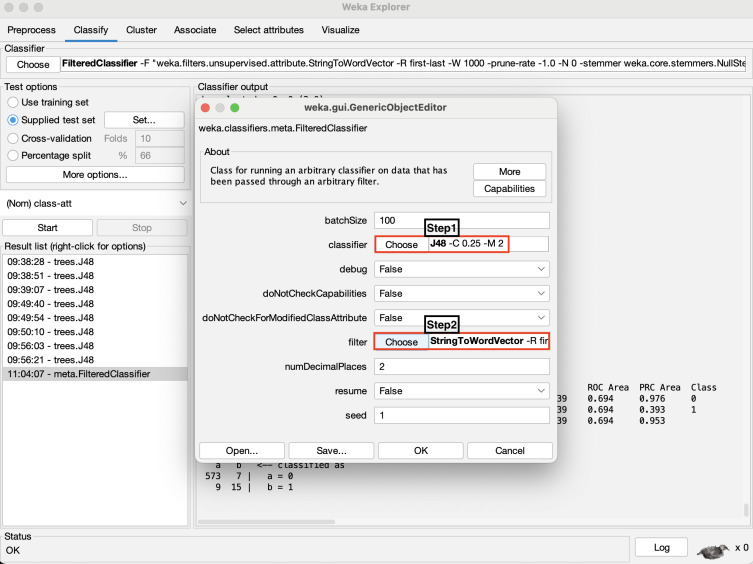


- 接著選擇 FliteredClassifier 這個分類器

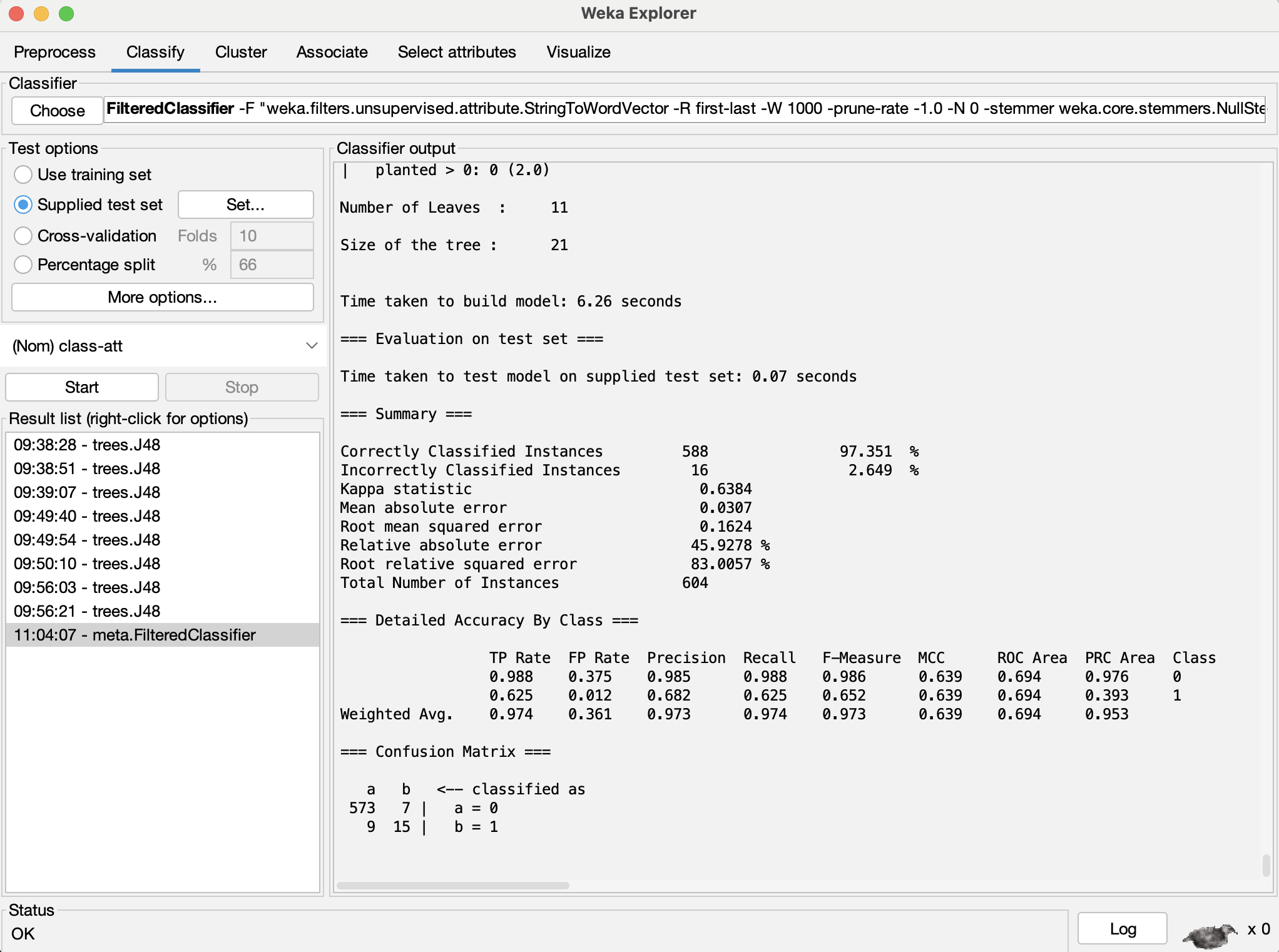


- 接下來我們需要調整FliteredClassifier 的參數

- 先選擇欲使用的分類器，然後選擇StringToWordVector 的過濾器

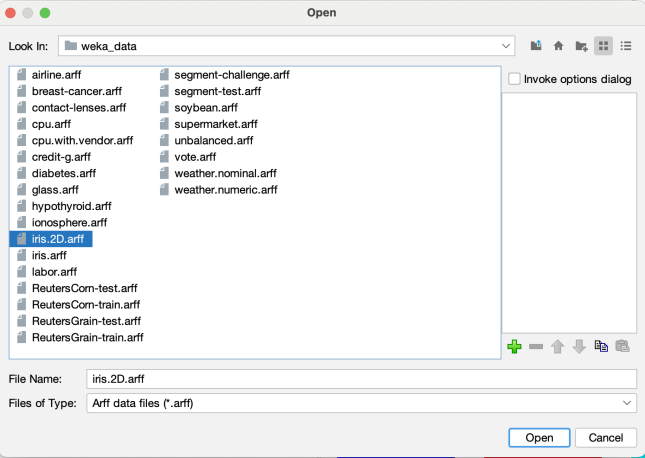


- 最後按下 Start 就能得到 97.351% 的準確率了



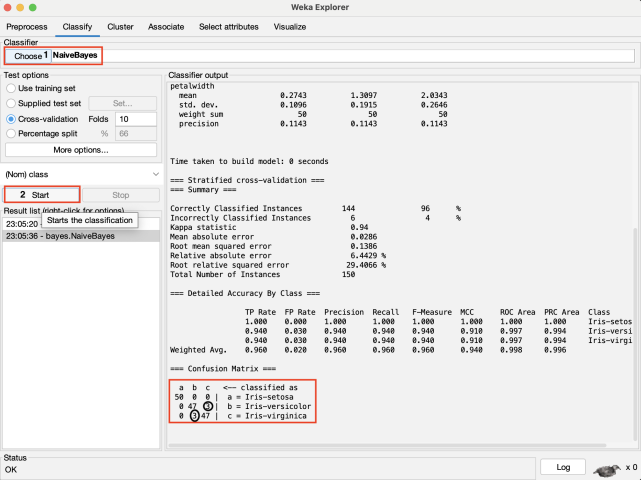
1. **評估分類器**

- 首先選擇資料集



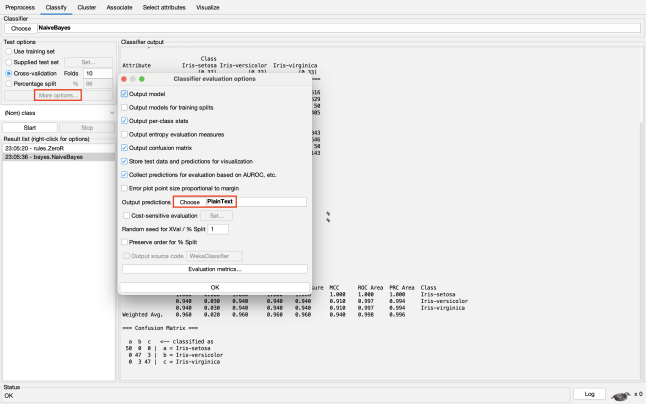
- 接下來選擇分類器我先用NaiveBayes，準確率為96%

- 可以看到Confusion Matrix 那裡被圈起來的是分類錯誤的



- 接下來我們可以將輸出的結果顯示在Weka 中

- 只要在Output predictions 選擇輸出結果到PlainText 格式

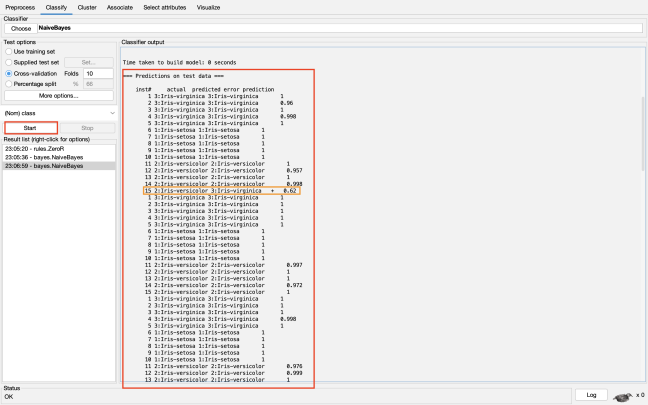


- 上一步設定好後再點擊 Start 就可以看到輸出的結果如下

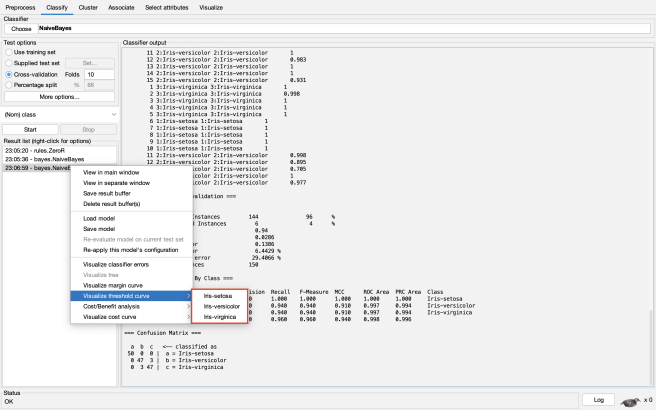
- 像第15個實例有 + 號代表預測錯誤

- 0.62 代表有62%的機率預測為Iris-virginica

- 但正確的是Iris-versicolor 而不是Iris-virginica

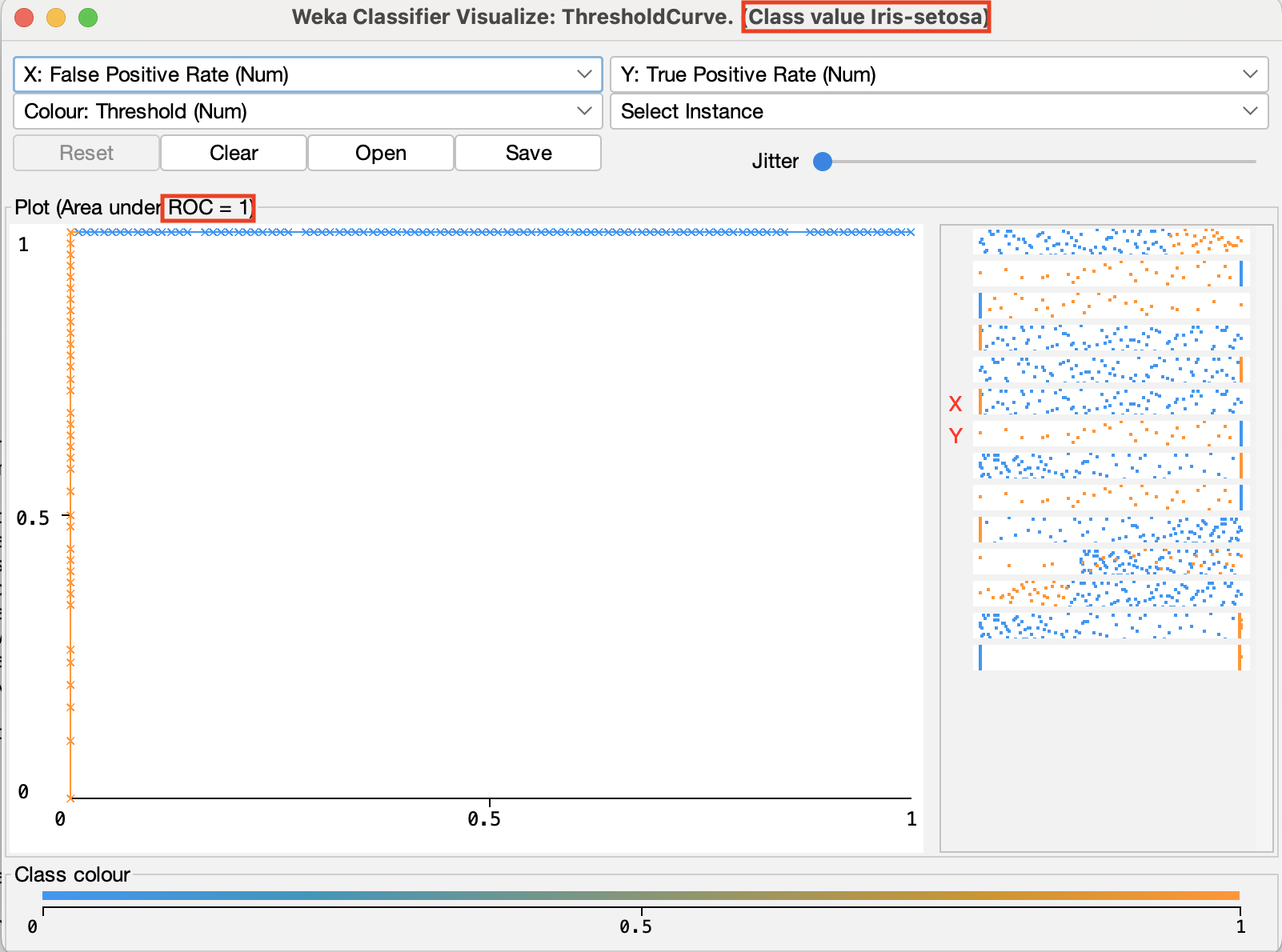


- 接著我們可以右鍵點擊剛剛的模型然後像下圖那樣就可以看到這三種的ROC 曲線圖



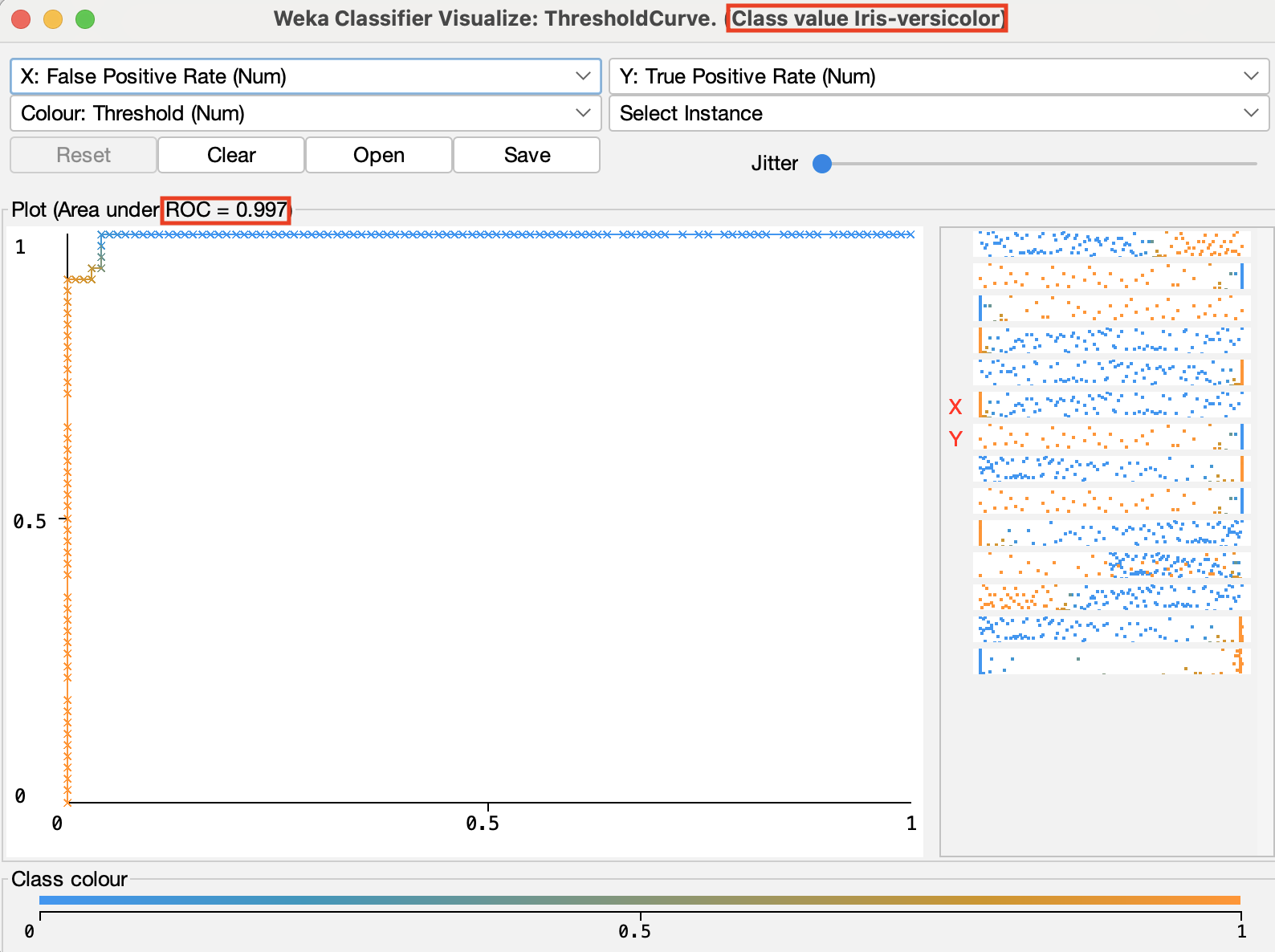
- Iris-setosa 的ROC曲線圖

- 這是理想中最完美的曲線，ROC=1



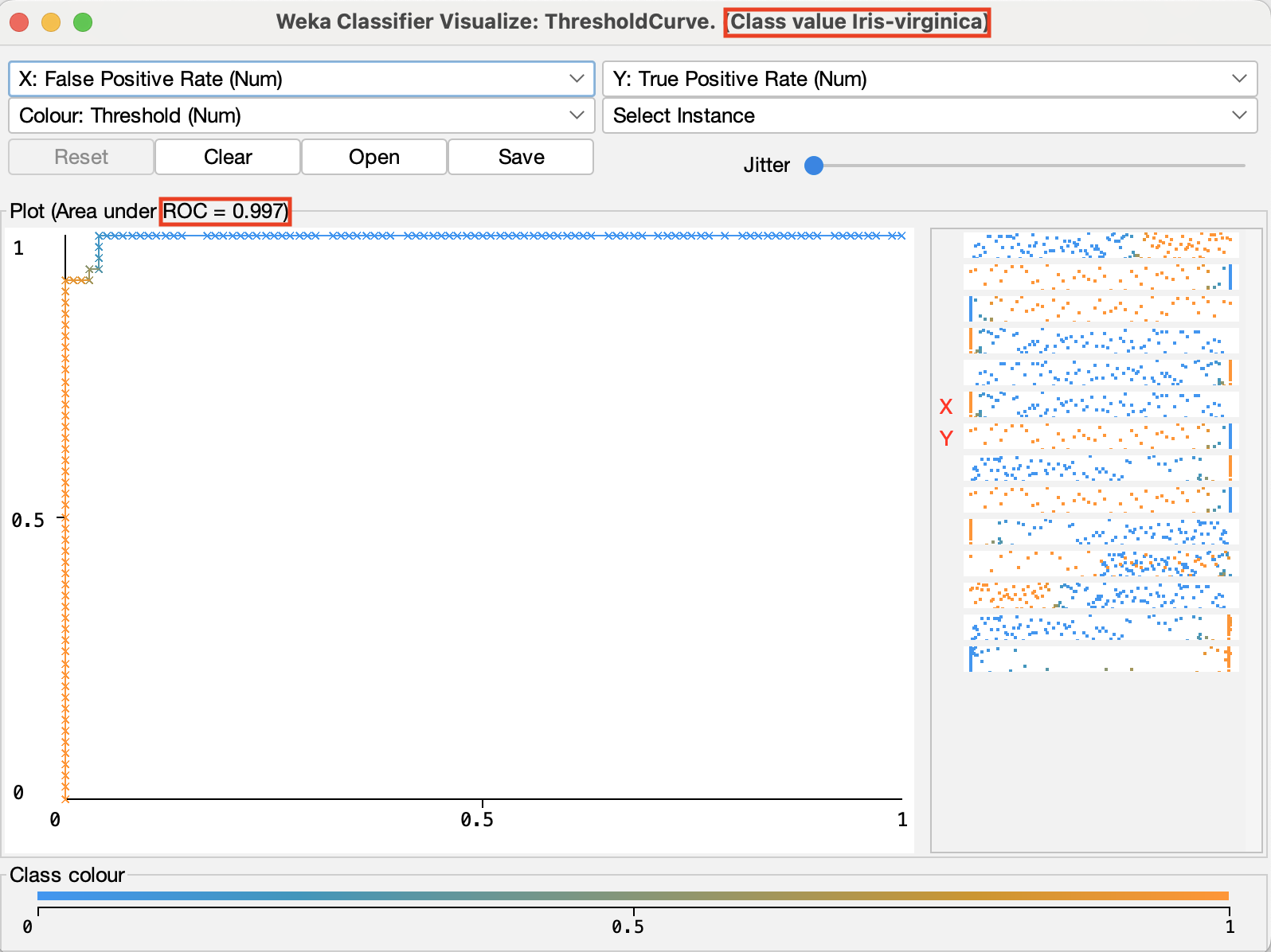
- Iris-versicolor 的ROC 曲線圖

- ROC=0.997，雖然不是理想但已經很不錯了



- Iris-virginica 的ROC 曲線圖

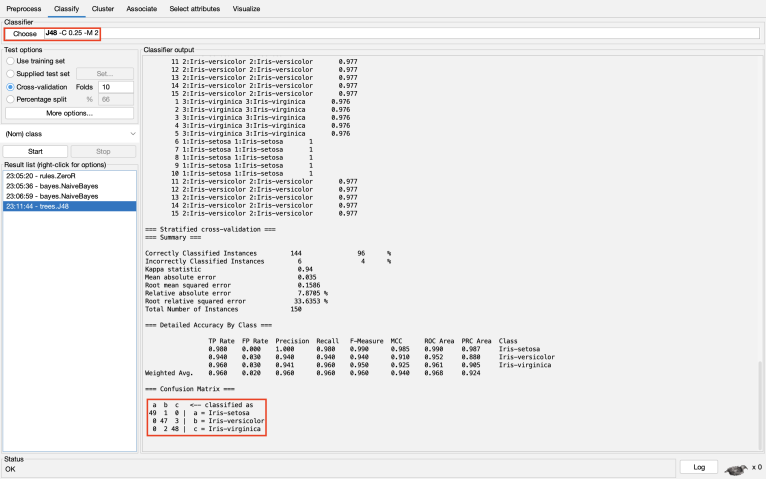
- ROC=0.997



- 接下來我們可以換一個分類器來看看有什麼差別

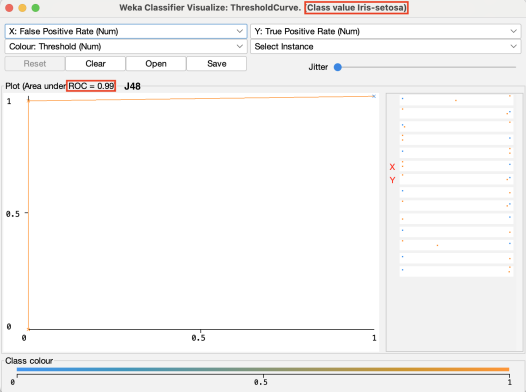
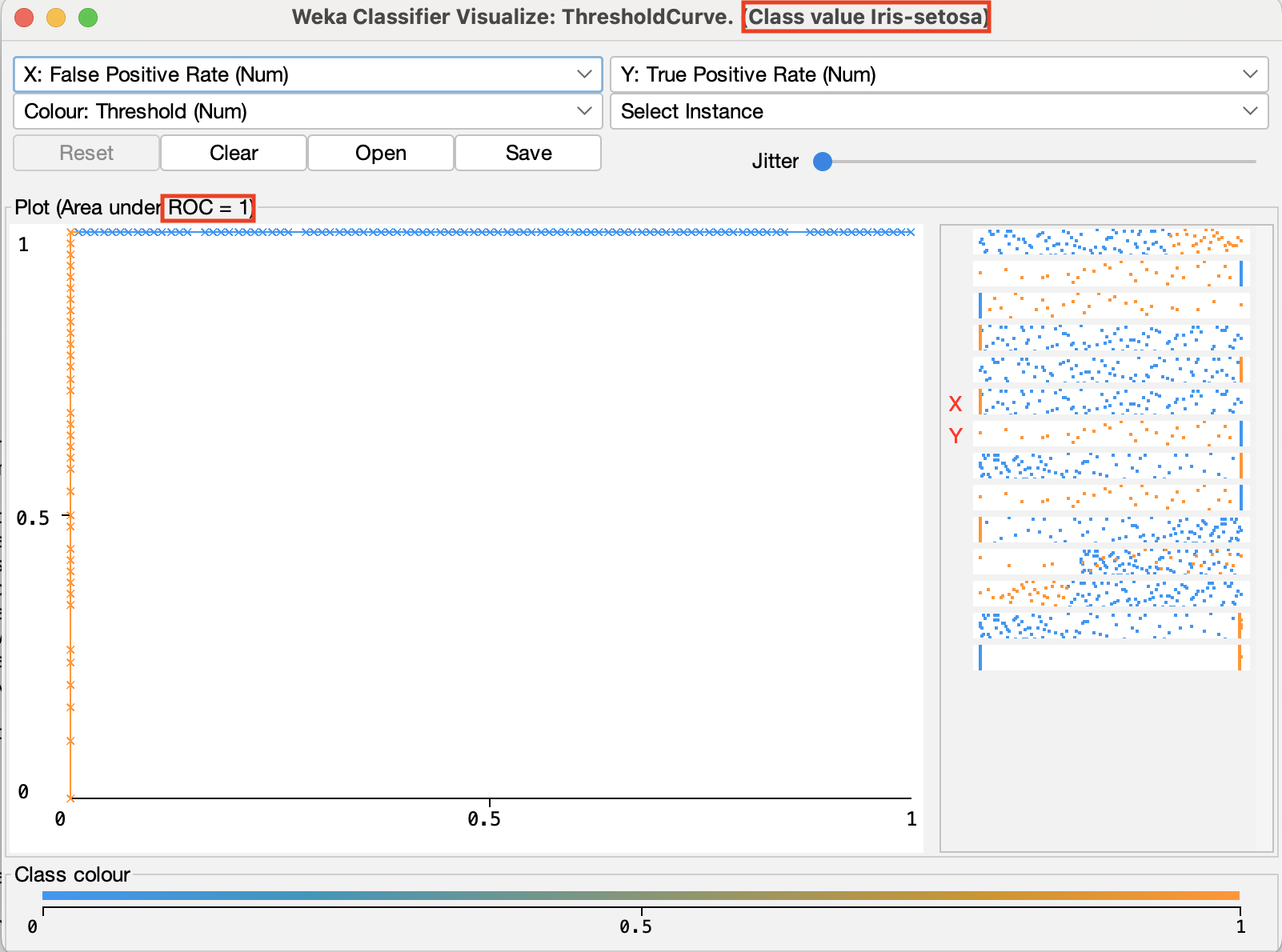
- J48 ：準確率與NaiveBayes一樣為96%

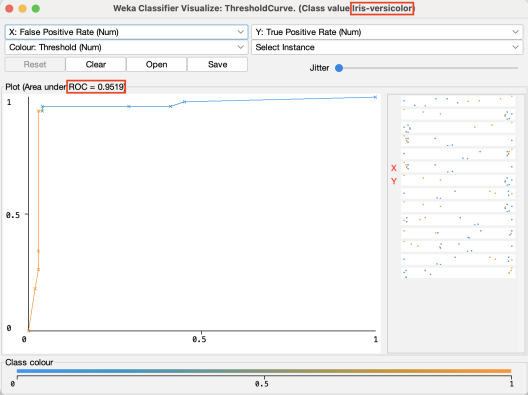
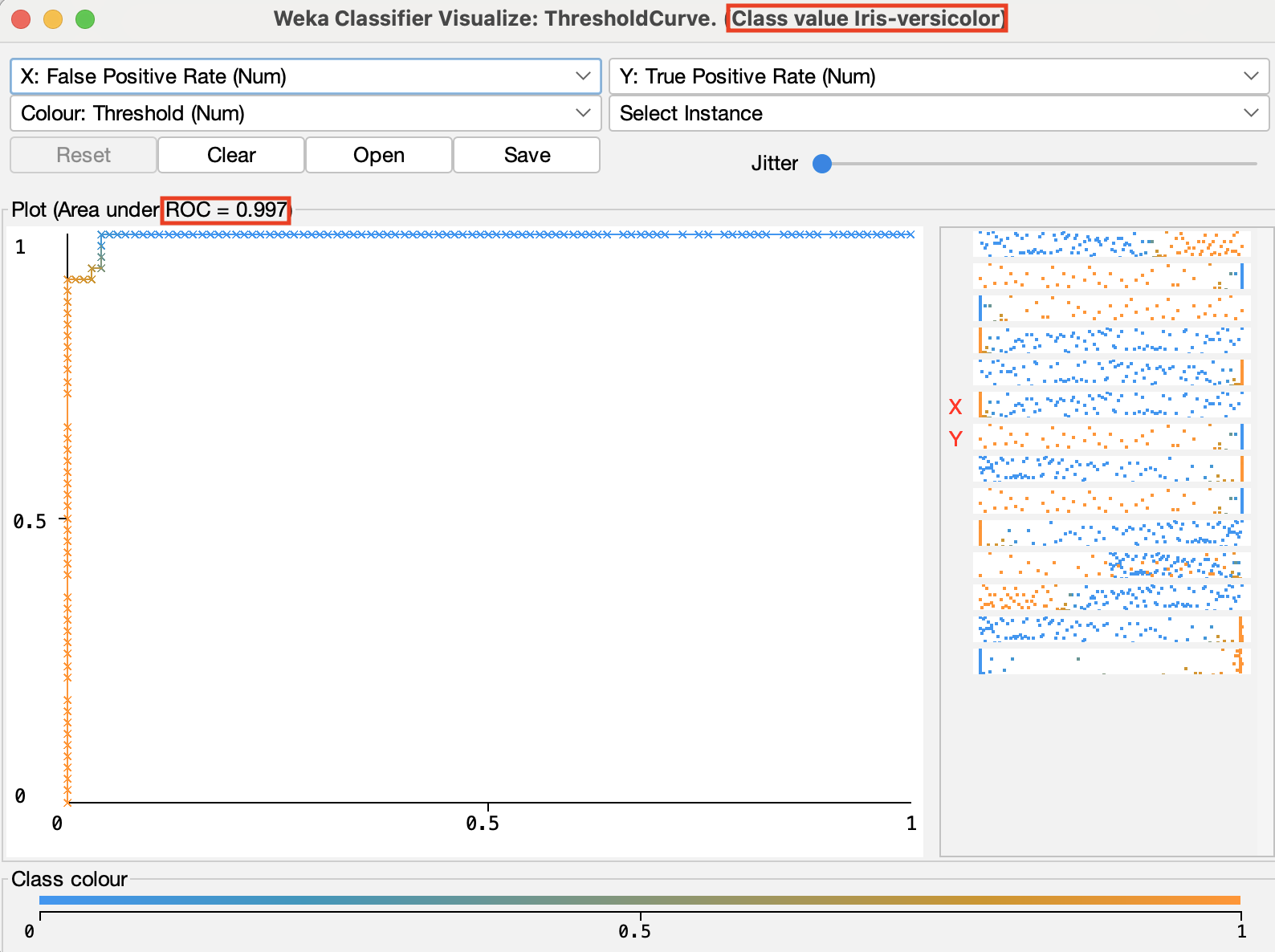
- 但在Confusion Matrix 那裡卻不一樣

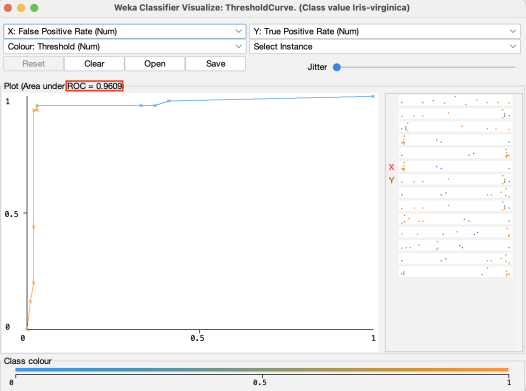
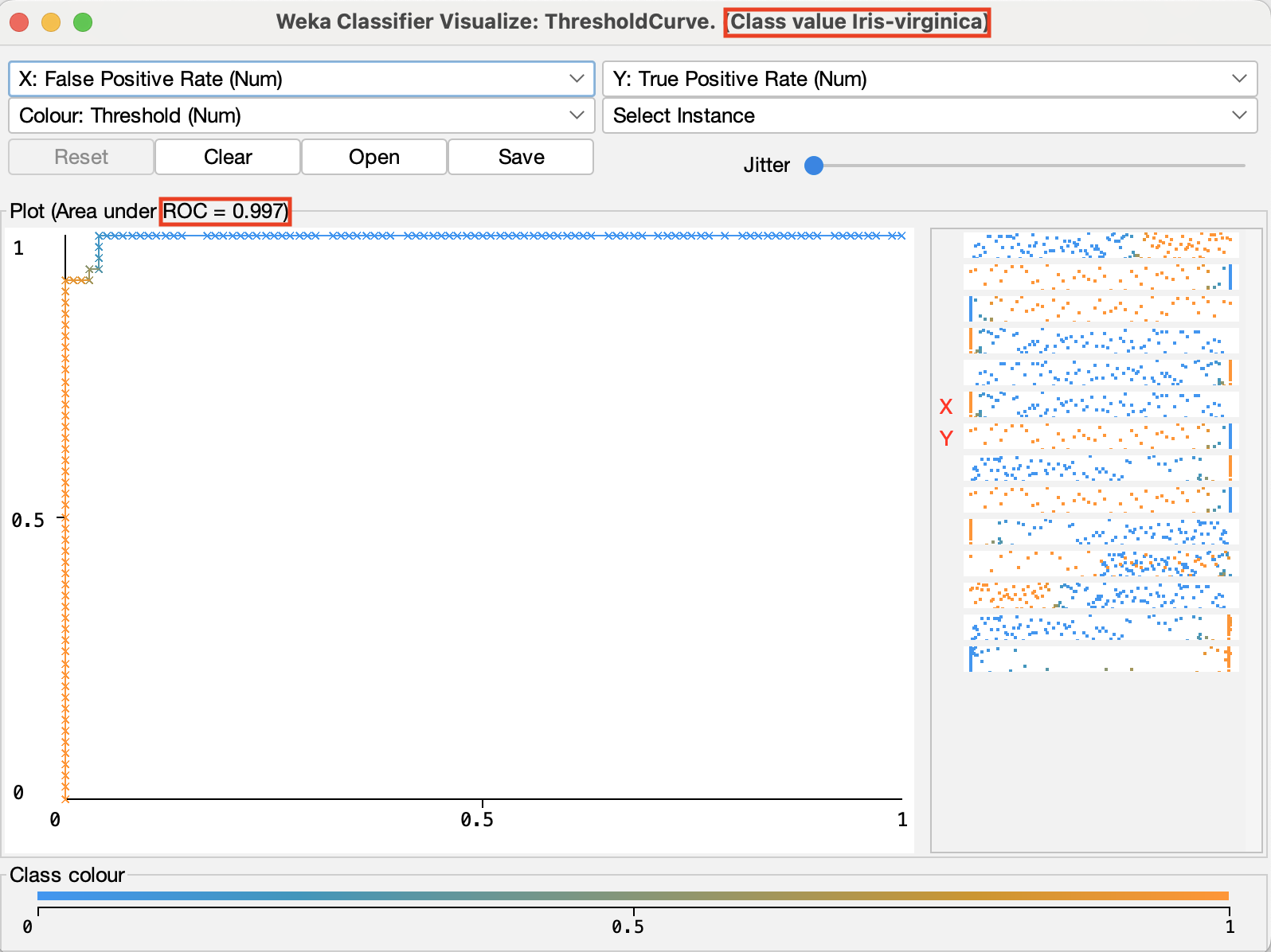


- 下列圖片為J48 與NaiveBayes 分別在三種Iris 的ROC曲線圖

- 左邊為J48 右邊為 NaiveBayes

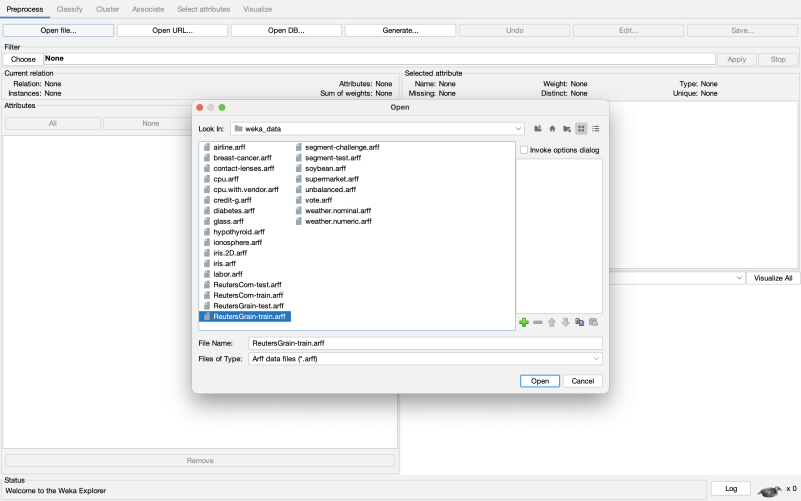
\* 從這些曲線圖我們可以發現不同的分類器對於同一個資料集的會有不同的結果

\* 還有除了準確率外我們也可以從其他指標來看出不同分類器對於同一個資料集的差別

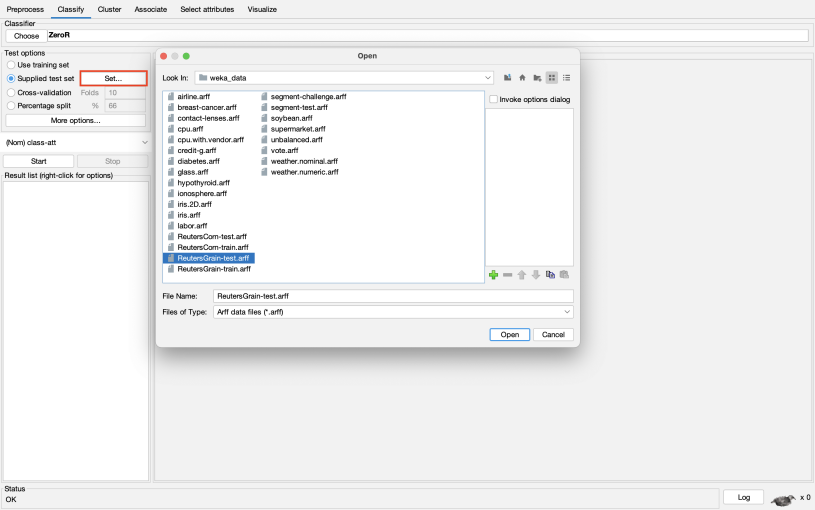
- 像是J48與NaiveBayes的準確率都是96%，但它們在對Iris-setosa的分類時，NaiveBayes 是完全正確，而J48卻將某個setosa分類成versicolor

1. **Multinominal Naive Bayes**

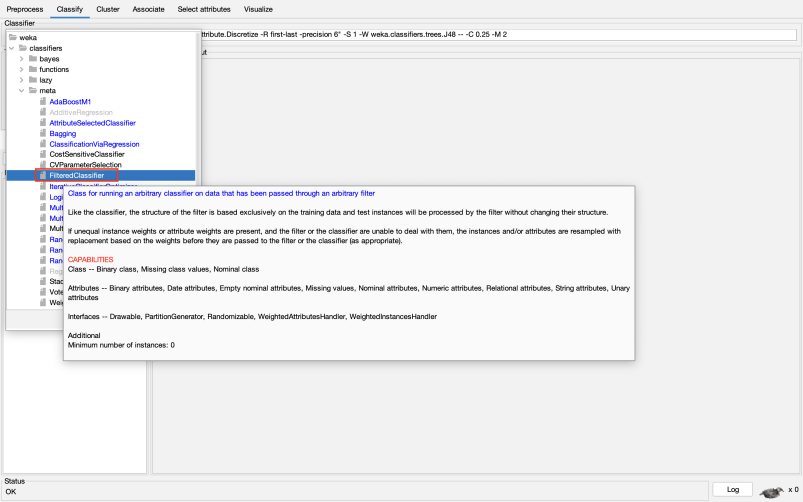
- 首先選擇資料集ReutersGrain-train.arff



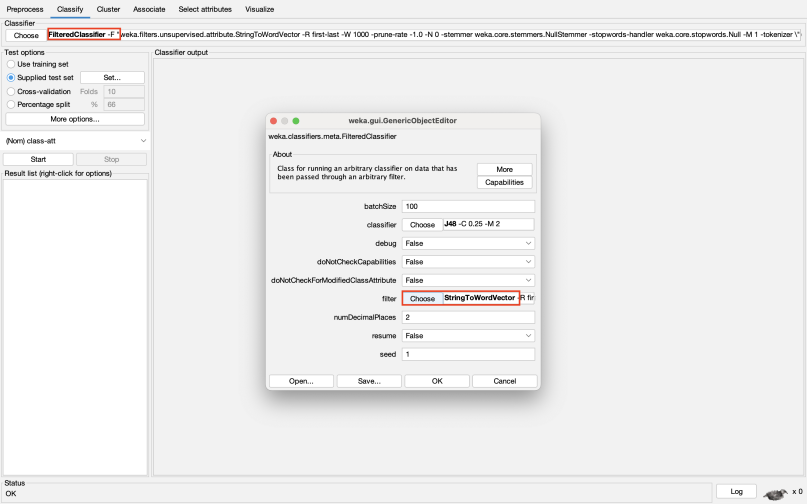
- 首先我們使用weka 內建的測試集來做測試



- 然後選擇 FilteredClassifier 分類器



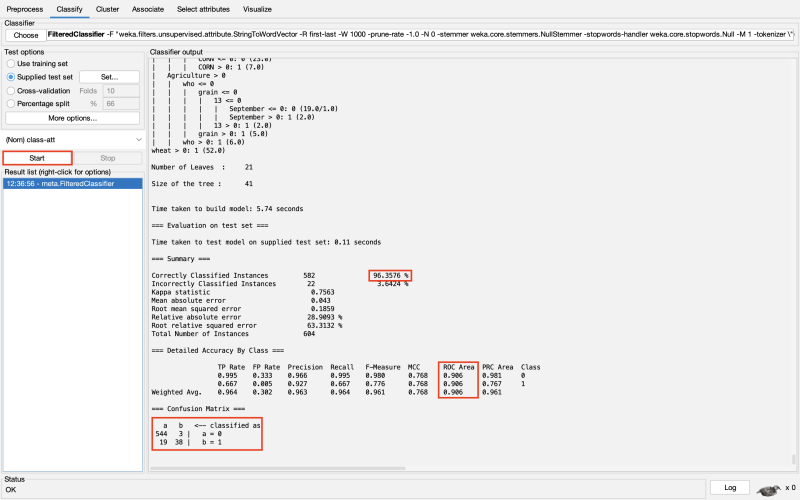
- 點擊FilteredClassifier 然後選擇filter並設定為StringToWordVector



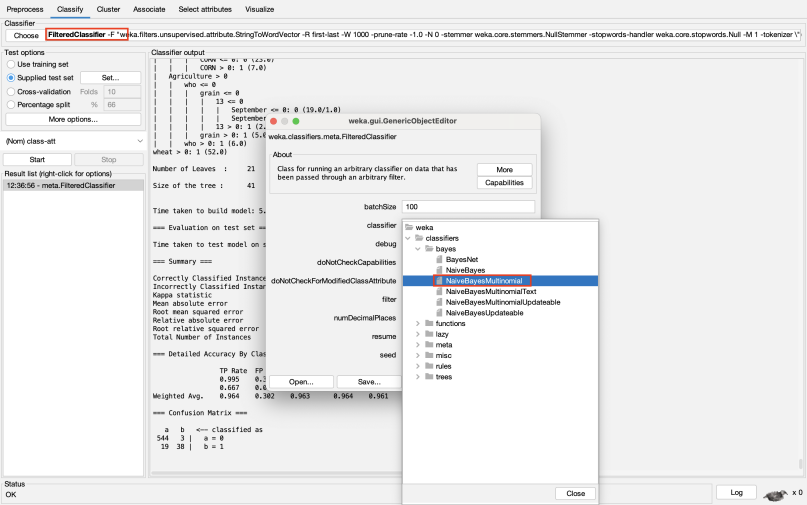
- 設定完成後點擊Start

- J48準確率為96%，ROC=0.906

- 準確率很高但在grain的分類並不是很好，57個裡面只有38個正確

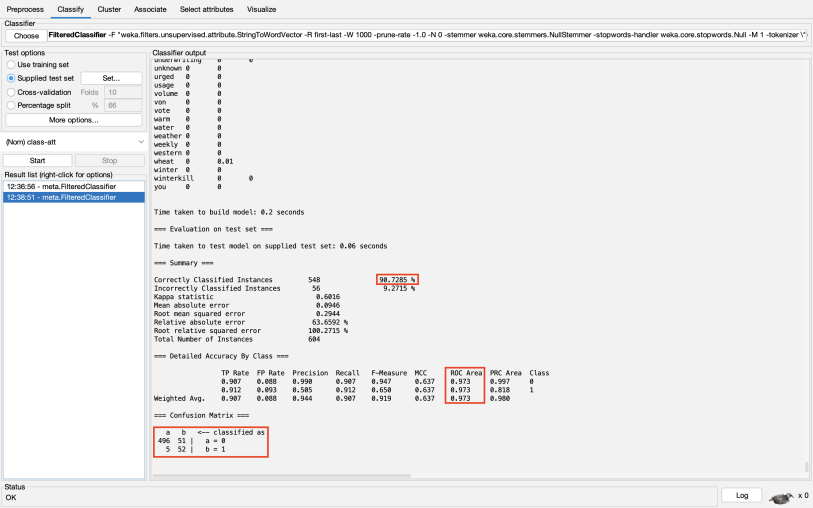


- 接下來我們換個classifier，選擇NaiveBayesMultinominal

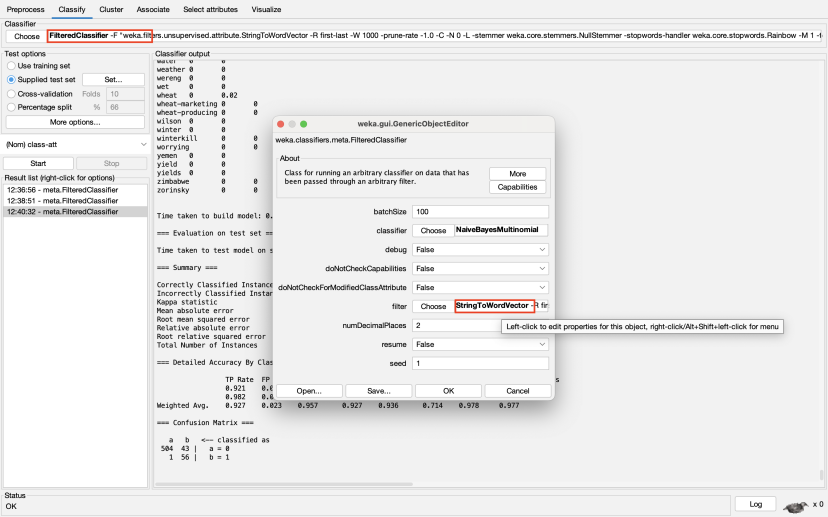


- 準確率為90.7285%，ROC=0.973

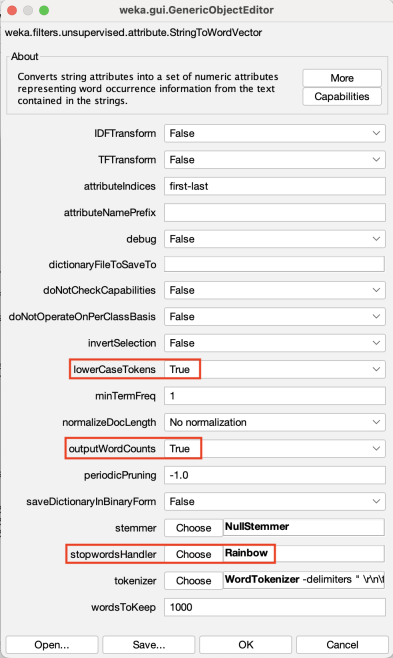
- 雖然準確率沒有J48好但在grain的分類上卻遙遙領先J48，ROC也更好



- 接下來我們可以試試看調整filter的參數



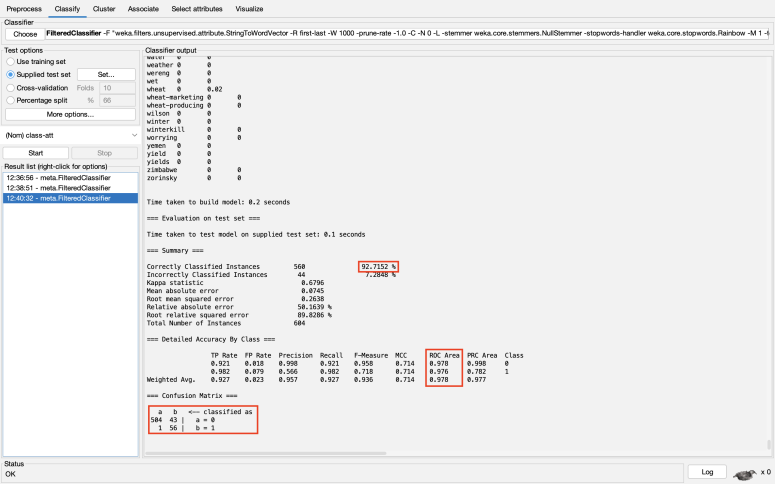
- 點擊StringToWordVector然後調整下列參數



- 完成後點擊Start

- 可以發現準確率提高到92.7152%，ROC也上升到0.978

- 在grain的分類也只有一個是錯誤的



\*我們可以發現到過濾器對於我們預測也是十分有幫助的

\*準確率雖然很重要但並不是唯一的指標