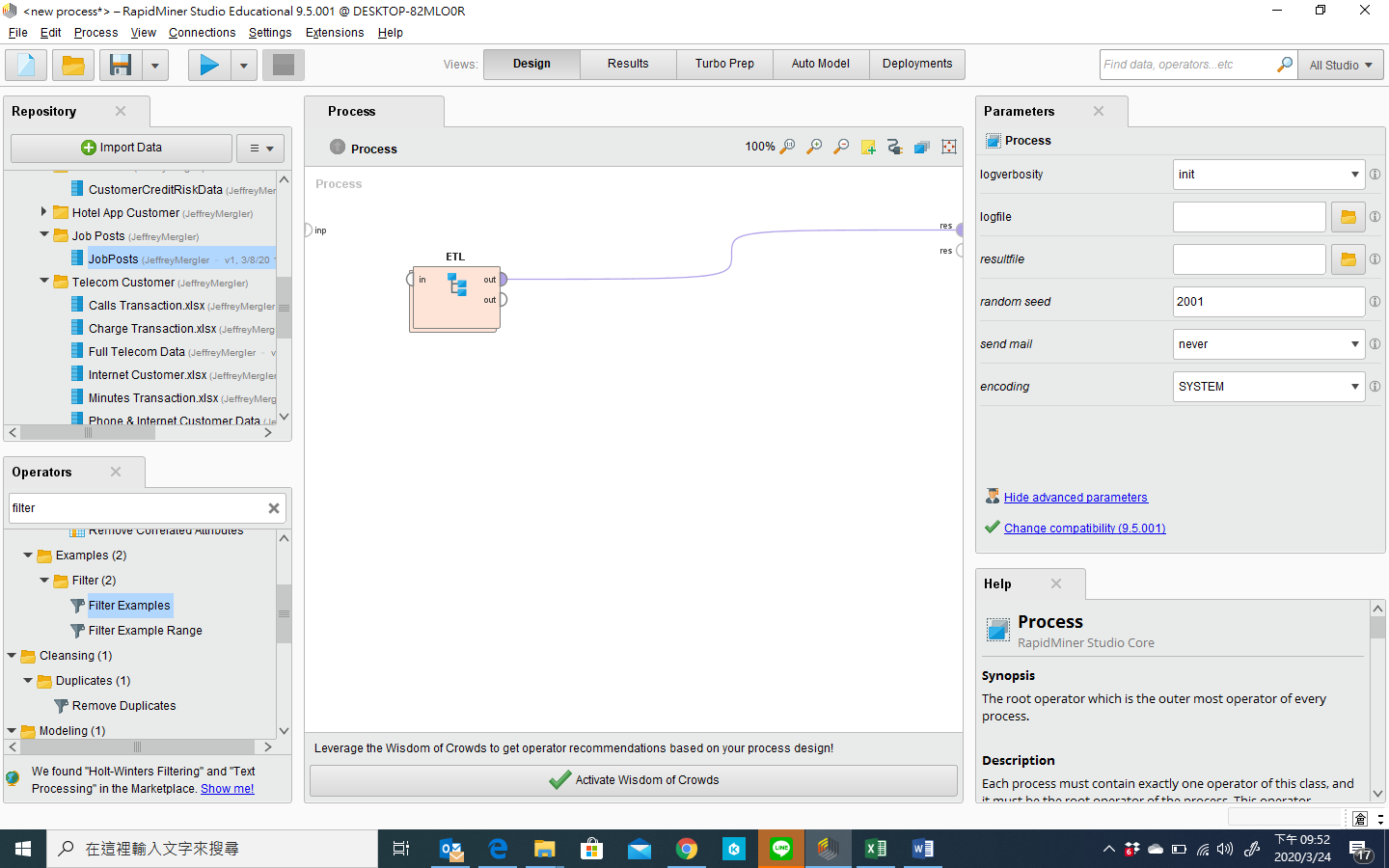
作業2: 我分類你~

是否很期待呢,我們終於要開始第一個分析專案了。

講到分析，不得不來了解一下大名鼎鼎的狠角色：決策樹(Decision Tree).

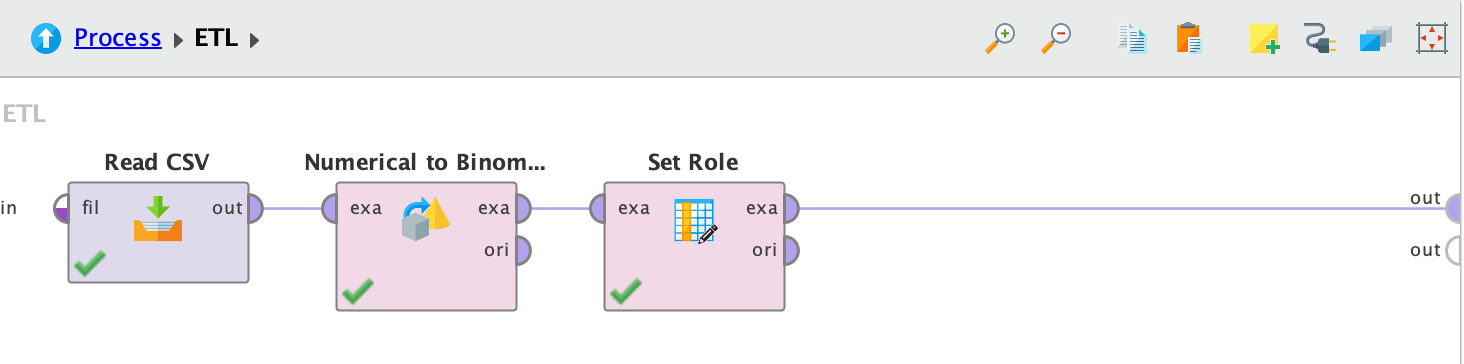
以下作業請先讀入marketing\_campaign檔案,並開始分析:

1. 讀入檔案後,請大概說明一下您對於資料的了解(5%),並進行「必要」之資料前處理,並請將相關前處理存於ETL子流程如下圖(請說明您做了什麼,為什麼) (10%)



資料內包含所有客戶的一些基本資料，包含出生年、收入、婚姻狀況、是否有小孩、購買的產品類型、及購買次數。公司可針對五個不同活動對客戶進行分析

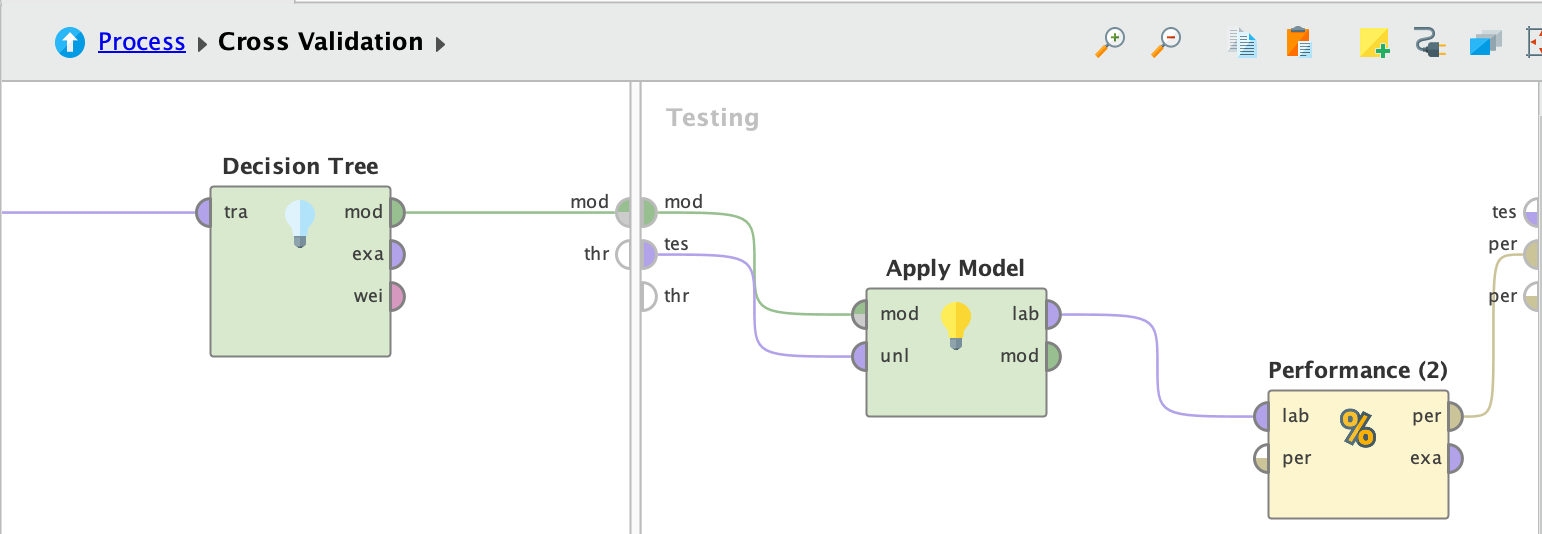
ETL流程：讀檔後，將response欄位的數值格式轉換成二分類格式，接著將response欄位設為label



2. 請找出cross validation, 並利用它進行模型的建置.何謂cross validation? (5%)

k-folds :有k筆資料，當第一筆資料用來當做驗證時，其他資料當作訓練，以此類推

3. 請在cross validation中進行決策樹的建構，並請貼出您的流程圖，並請說明相關之參數設定 (10%)

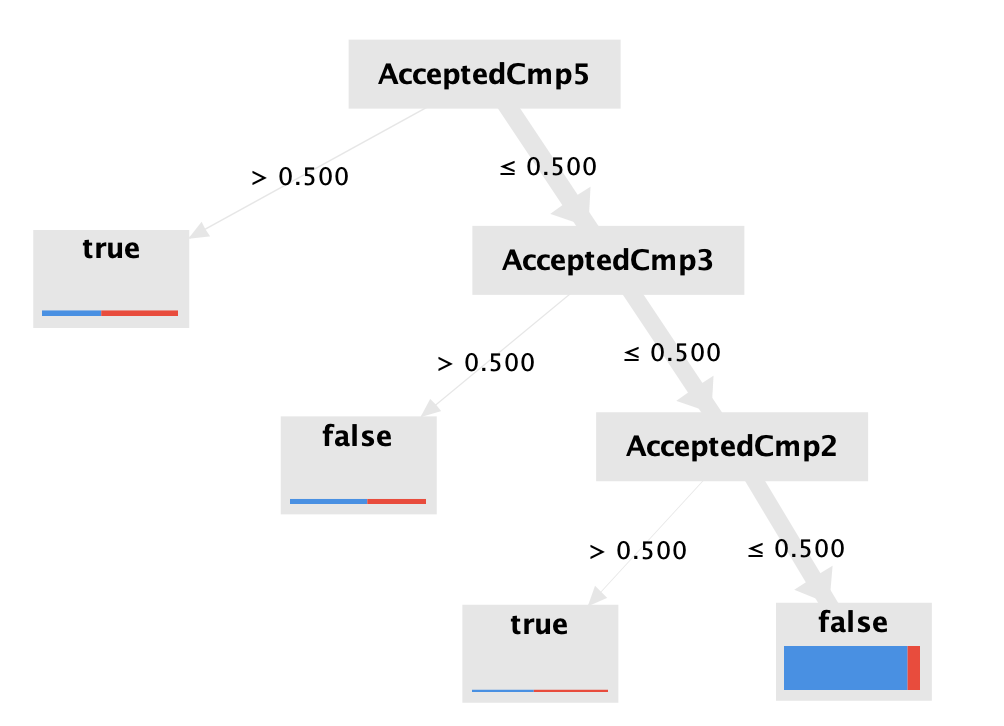


參數修改

decision tree：maximal depth=5, minimal of size=10

performance：number of folds=5, sampling type : automatic

4. 若您的模型在建構過程中有上述問題，請試著解決並以回應（Respone）為預測標的，貼出您的預測模型並加以說明 (15%)



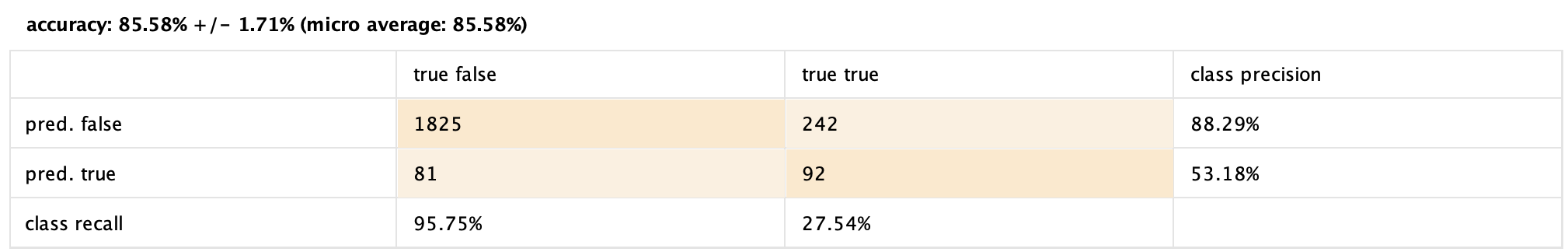
首先以是否接受活動5為分類，若活動5的接受率<=0.5，且則活動3及活動2的接受也都小於0.5，則最終response的答案為錯誤的機率很高。

6 請列出您的混肴矩陣,並請加以說明其意涵，除此之外，我還希望您能列出您模型的下述４個指標，並請簡略說明他們的計算方式及含意(5%) 請將您的流程存檔為學號\_1.rmp

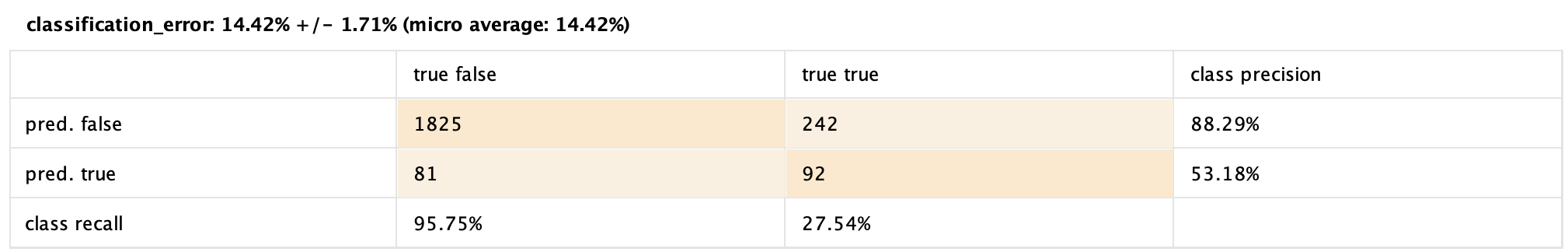


混淆矩陣是用來平衡模型好壞的方法，在原始資料中，可以發現是錯誤的而預測也為錯誤的數量達到1825筆，是正確的而預測也為正確的數量92筆。

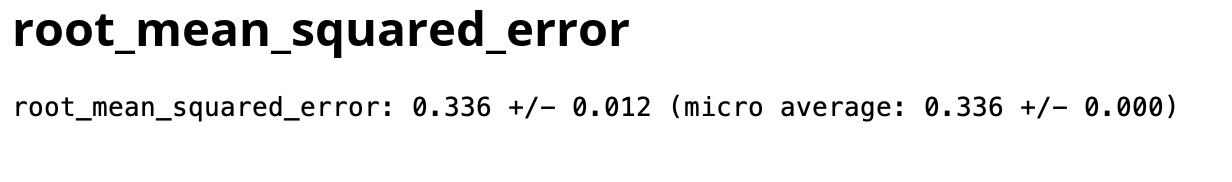
準確率反映了那些預測為“是”的比例中實際為“是”的比率，以accuracy為例，整體準確度達到85.58%。



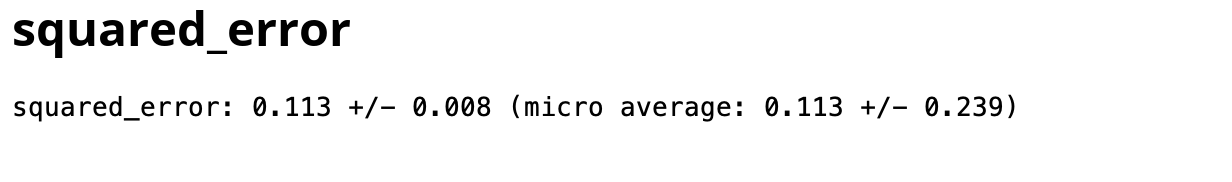
以classification\_error為例，分類錯誤率指所有分類錯誤的樣本數占所有樣本數的比例，而此模型的錯誤率為14.42% +/- 1.71%。



以root\_mean\_squared\_error為例，均方根誤差是一種常用的測量數值之間差異的量度，代表預測的值和觀察到的值之差的樣本標準差，它是「誤差」的平方的期望值。而此模型的均方根誤差為0.336 +/- 0.012。



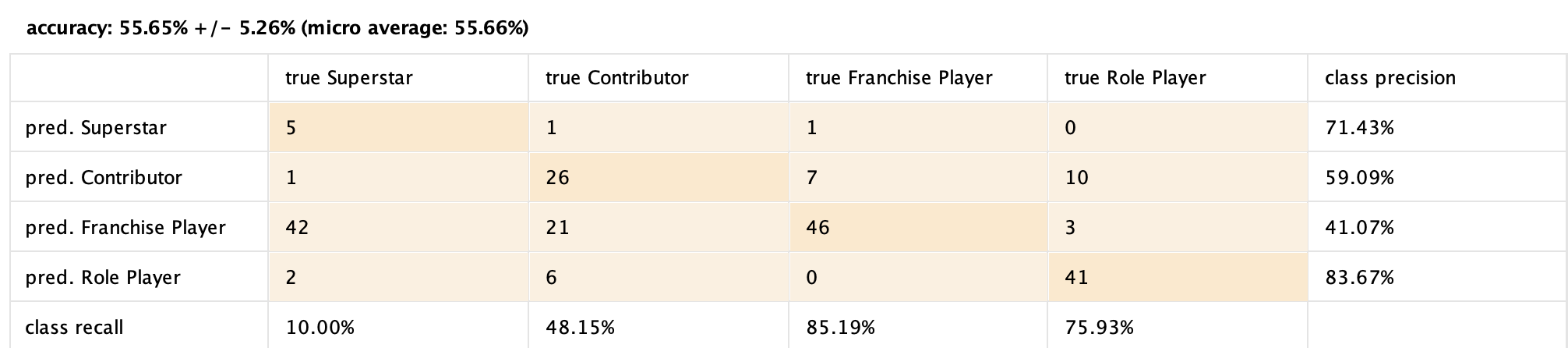
square\_error是預測值和實際觀測值間差的平方。 它只考慮誤差，不考慮其方向。 但由於經過平方，與真實值偏離較多的預測值會比偏離較少的預測值受到更為嚴重的懲罰。而此模型的平方誤差為0.113 +/- 0.008。



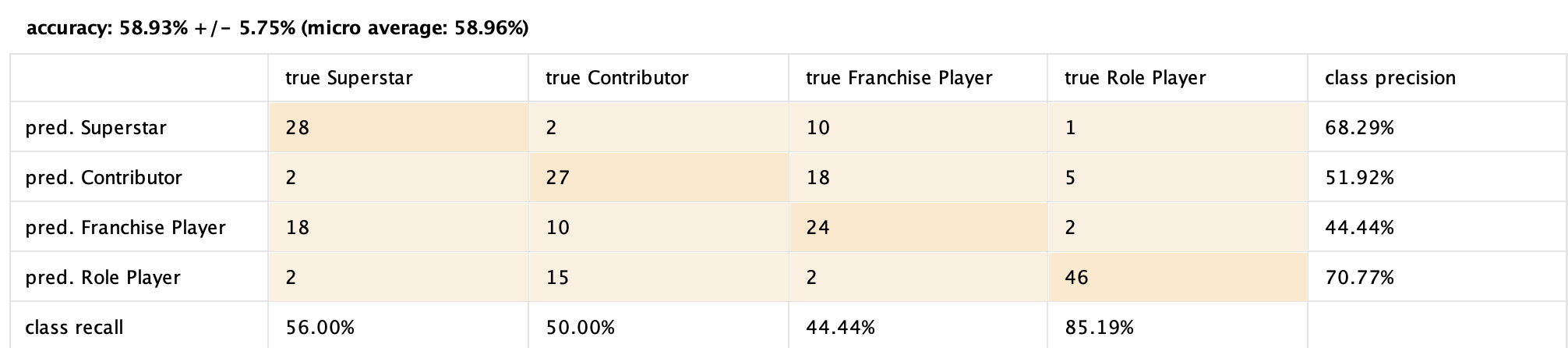
接下來就到了我們期待的神對手V.S豬隊友一題了.

俗話說的好,不怕神一般的對的隊手,就怕豬一樣的隊友,所以,跟進一組(隊)就變的相當重要囉! 在進行數據分析時,為人所津津樂道的一個例子應該就屬”魔球”(Money Ball)了。 雖然無法人人都擁有主角光環，但我們還是想辦法弄來了一個資料集，大家試試打開player\_train.csv吧.

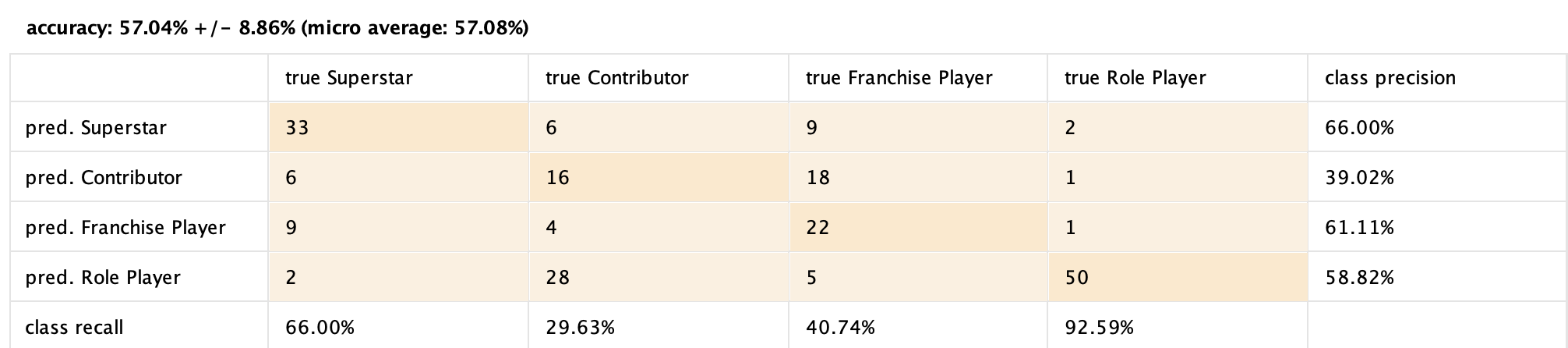
7. 請使用player\_train.csv進行模型的訓練(建議使用cross validation),並請找出正確率最高的預測演算法. (35%)，請將您的流程存檔為學號\_2.rmp

Decision Tree: 55.65%

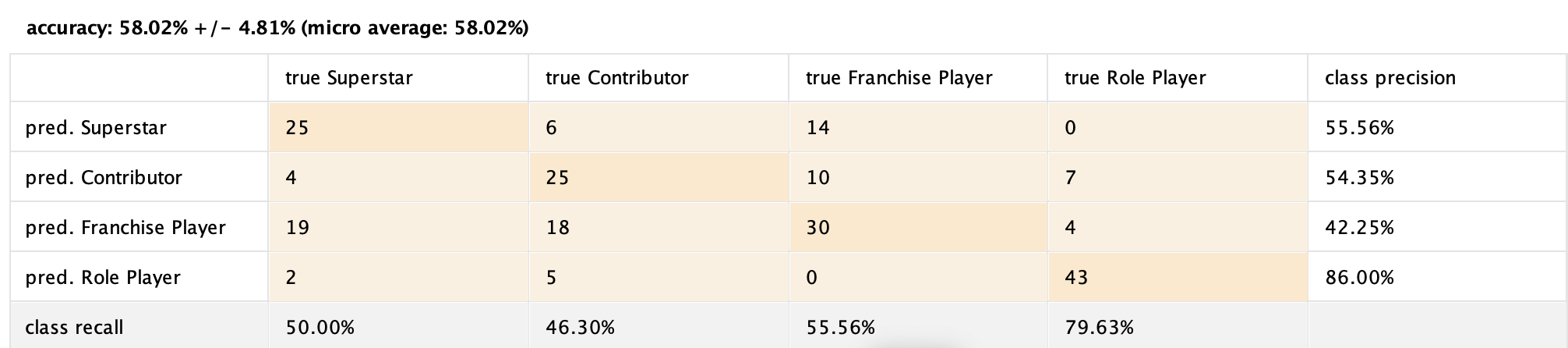
Deep Learning:58.93%



Naïve Bayes:57.04%



KNN:58.02%



四種方法中，以Deep learning 的正確率為最高。

8. 請使用player\_scoring.csv進行隊員的篩選,並請說明您的策略是? (15%)

首先先用player\_train的檔案進行deep learning的訓練模型，接著讀取player\_scoring，並將total\_point設為label，當作篩選team\_value的條件來預測。

