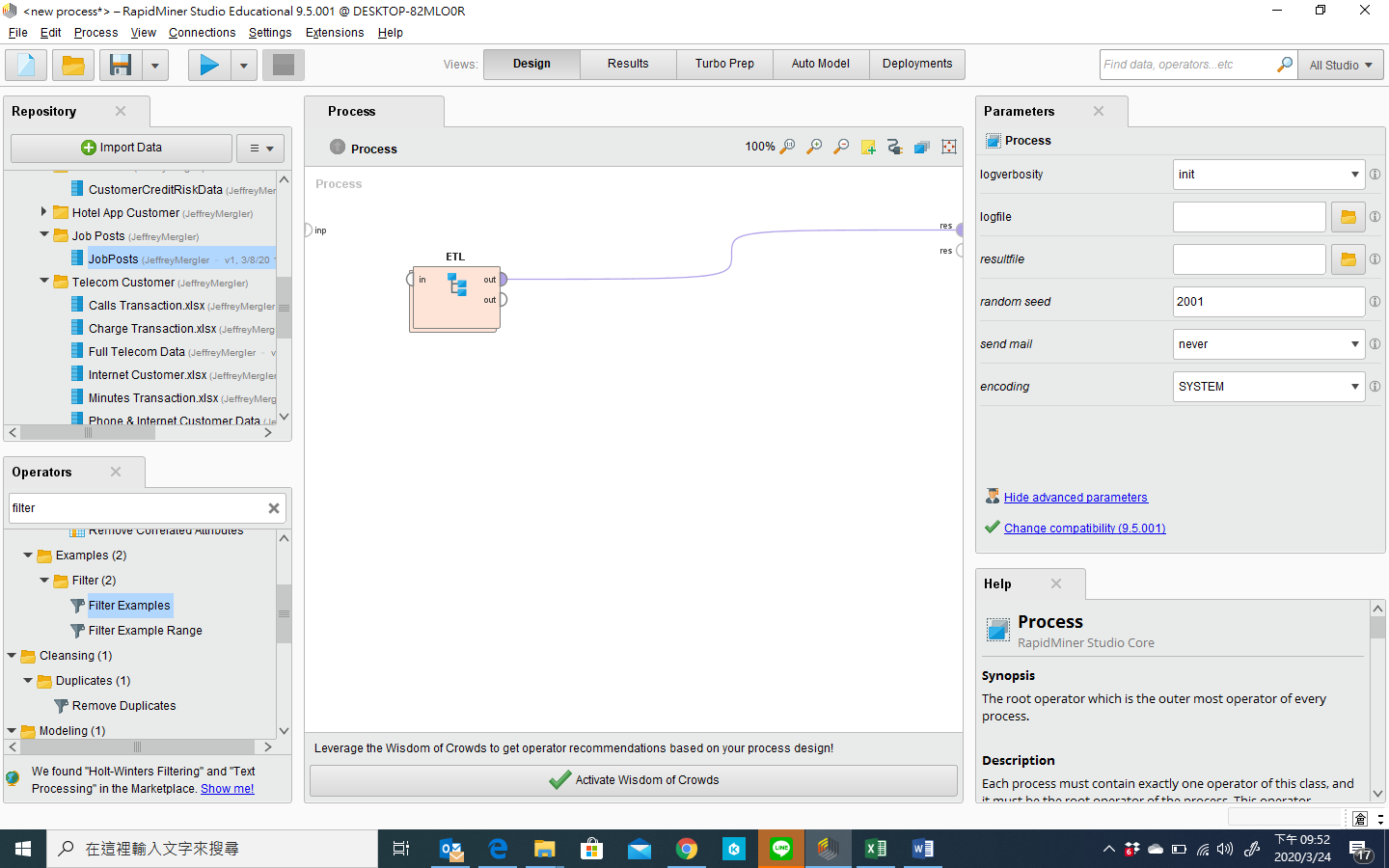
分類: 見樹不見林

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income檔案,並開始分析:

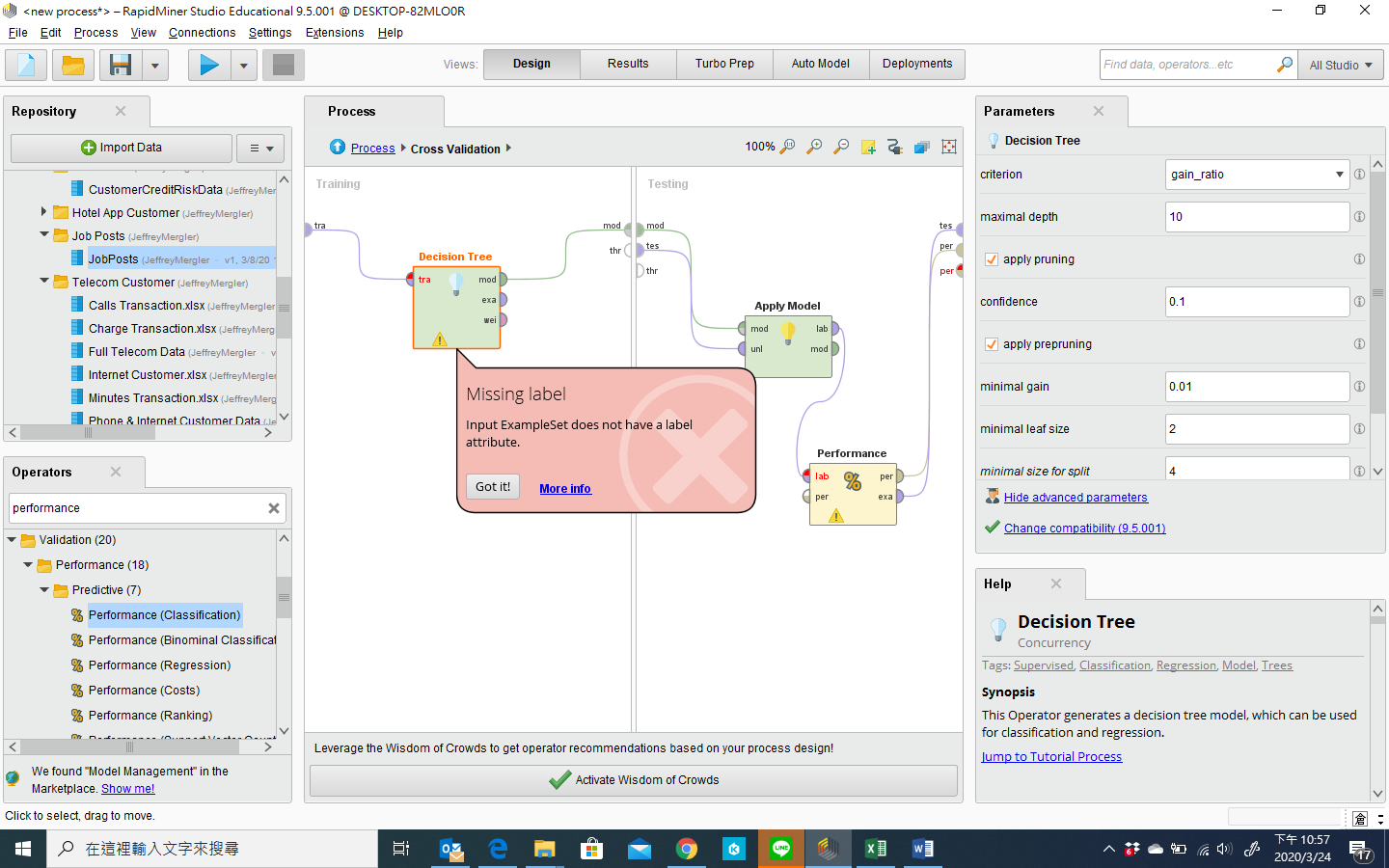
1. 讀入檔案後,請大概說明一下您對於資料的了解(5%),並進行「必要」之資料前處理,並請將相關前處理存於ETL子流程如下圖(請說明您做了什麼,為什麼) (10%)



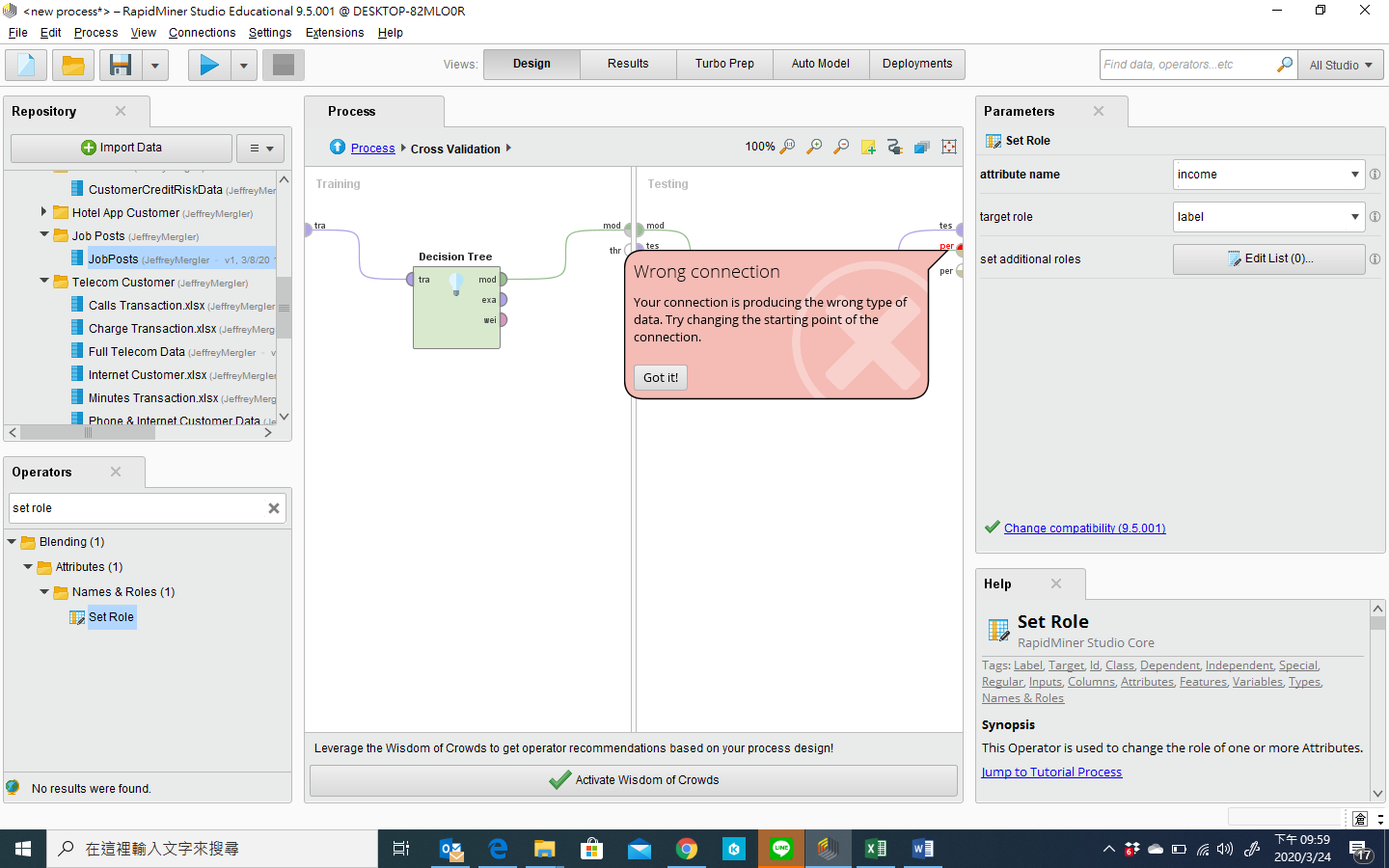
2. 請找出cross validation, 並利用它進行模型的建置.何謂cross validation? (5%)

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%)

4. 請問以下錯誤訊息是指? 該如何解決(5%)



5. 請問以下錯誤訊息是指? 該如何解決(5%)



6. 若您的模型在建構過程中有上述問題，請試著解決並以收入（income）為預測標的，貼出您的決策樹,並加以說明您的決策樹意涵(15%)

7 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(8%)



但是,您用了幾乎全部的變數.幾乎全部的變數. 幾乎全部的變數.

8.以下我們將限制您的變數使用量，

8-1請問若只讓您使用6個變數進行預測,您會選擇(請列出變數名稱)? (4%)

8-2請問若只讓您使用3個變數進行預測,您會選擇(請列出變數名稱)? (4%)

8-3您的選擇依據是? (4%) 變數變少後,正確率的變化是? (4%)

9.最後，在５個變數的限制下，請調整參數，找出一顆正確率最高的樹，並請說明您試過那些參數的調整？

請將您的process存檔為學號-1, 如: 106AB001\_1.rmp檔

10 請讀入Bike.csv檔案,這是一個共享經濟單車的資料集,

- instant: record index   
- dteday : date   
- season : season (1:winter, 2:spring, 3:summer, 4:fall)   
- yr : year (0: 2011, 1:2012)   
- mnth : month ( 1 to 12)   
- hr : hour (0 to 23)   
- holiday : weather day is holiday or not (extracted from [[Web Link]](http://dchr.dc.gov/page/holiday-schedule))   
- weekday : day of the week   
- workingday : if day is neither weekend nor holiday is 1, otherwise is 0.   
+ weathersit :   
- 1: Clear, Few clouds, Partly cloudy, Partly cloudy   
- 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist   
- 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds   
- 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog   
- temp : Normalized temperature in Celsius. The values are derived via (t-t\_min)/(t\_max-t\_min), t\_min=-8, t\_max=+39 (only in hourly scale)   
- atemp: Normalized feeling temperature in Celsius. The values are derived via (t-t\_min)/(t\_max-t\_min), t\_min=-16, t\_max=+50 (only in hourly scale)   
- hum: Normalized humidity. The values are divided to 100 (max)   
- windspeed: Normalized wind speed. The values are divided to 67 (max)   
- casual: count of casual users   
- registered: count of registered users   
- cnt: count of total rental bikes including both casual and registered

10-1以2011年做為訓練集,2012年做為測試集, 以cnt為您的預測欄位,進行回歸模型的建立。請列出您的迴歸方程式,並加以說明相關正確率等數值. (10%)

10-2請問若只讓您使用6個變數進行預測,您會選擇(請列出變數名稱)? (5%)

10-3請問若只讓您使用3個變數進行預測,您會選擇(請列出變數名稱)? (5%)

10-4 根據您的分析,您會建議UBIKE當局怎麼做(5%)