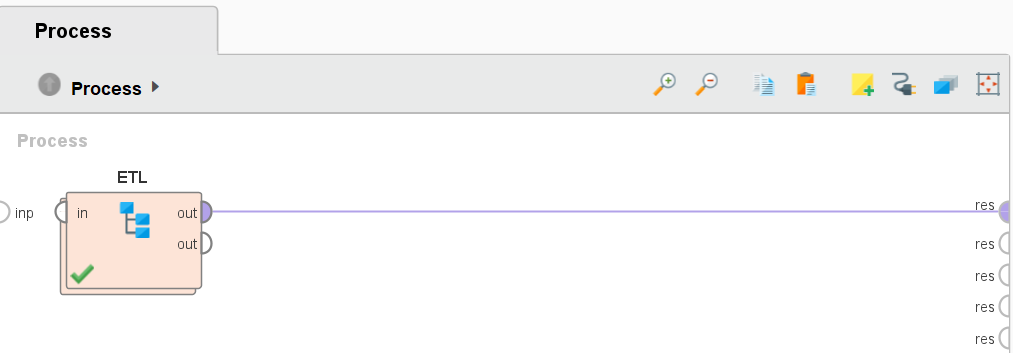
作業2: 看我用方程式算命

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：迴歸方程式.

以下作業請先讀入baston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(10%),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%)**



**說明資料:**

總共有14個欄位，分別為

※欄位名稱標示為紅色表示為常態分佈

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 欄位名稱 | 描述 | 最小值 | 最大值 | 平均數 |
| CRIM | 每個城鎮的人均犯罪率 | 0.006 | 88.976 | 3.614 |
| ZN | 被規劃為超過 25,000 平方英尺地塊的住宅用地比例 | 0 | 100 | 11.364 |
| INDUS | 每個城鎮的非零售業務佔地比例【城鎮中非商業用地佔比例】 | 0.460 | 27.740 | 11.137 |
| CHAS | 查爾斯河虛擬變量（如果是河道則為1；否則為0） | 0 | 1 | 0.069 |
| NOX | 一氧化氮濃度（每千萬份）【環保指標】 | 0.385 | 0.871 | 0.555 |
| RM | 每個住宅的平均房間數【每棟住宅房間數】 | 3.561 | 8.780 | 6.285 |
| AGE | 自 1940 年以前建造的自住單位比例 | 2.9 | 100 | 68.575 |
| DIS | 到波士頓五個就業中心的加權距離 | 1.13 | 12.127 | 3.795 |
| RAD | 距離高速公路的便利指數 | 1 | 24 | 9.549 |
| TAX | 每 $10,000 的全額房產稅稅率【每一萬美元的不動產稅率】 | 187 | 711 | 408.237 |
| PTRATIO | 每個城鎮的師生比例 | 12.6 | 22 | 18.456 |
| B | 1000（Bk - 0.63）^2，其中 Bk 是每個城鎮黑人比例 | 0.32 | 396.9 | 356.674 |
| LSTAT | 人口中低收入人群的百分比 | 1.73 | 37.970 | 12.653 |
| MEDV | 自住房屋的中位數價值，以千美元為單位 | 5 | 50 | 22.533 |

**數據統計**:

這一個波士頓房價的資料並沒有任何缺漏值，從資料的內容可以看出ZN和CHAS的值絕大多數都是0，RM(每棟住宅房間數)則呈現常態分佈，此次資料中黑人比例也都相當高，年齡分布也比較高齡。

**說明進行「必要」之資料前處理**

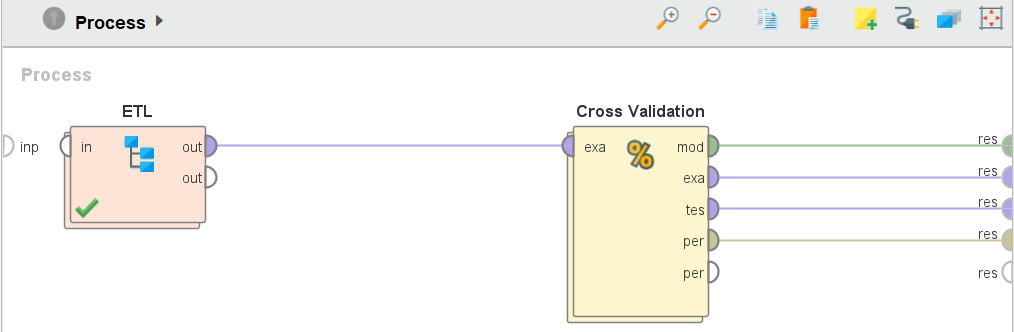
(一)先利用Selet Attributes 的operator，這個部份我將ZN和CHAS 先刪除，因ZN的資料和CHAS的資料大多數都是0，因此我認為這兩欄的資料是比較沒有參考價值的。

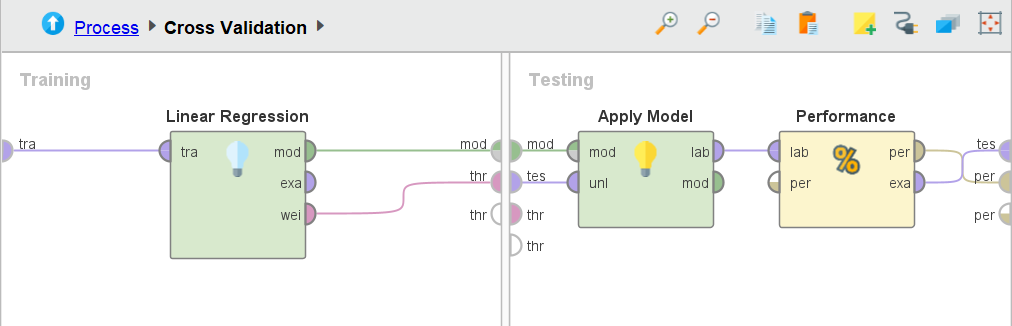
(二)再利用Set Role的operator，將要預測的MEDV設定成label

**2. 請找出cross validation, 並利用它進行模型的建置.何謂cross validation? (10%)**

cross validation是一種統計學上將樣本切割成較小子集的方法。先把資料切成k等分，第一等分設為測試或驗證集，而剩下k-1可以當作訓練集的資料，第二輪則是將第二等分當成測試或驗證集，其他k-1的資料當作訓練集，總共做k輪，再將這K次計算的準確率取平均，最後算出的這個平均值則比較沒有誤差。

**3. 請在cross validation中進行回歸模型的建構，並請貼出您的流程圖，並請說明相關之參數設定 (10%)**

****

****

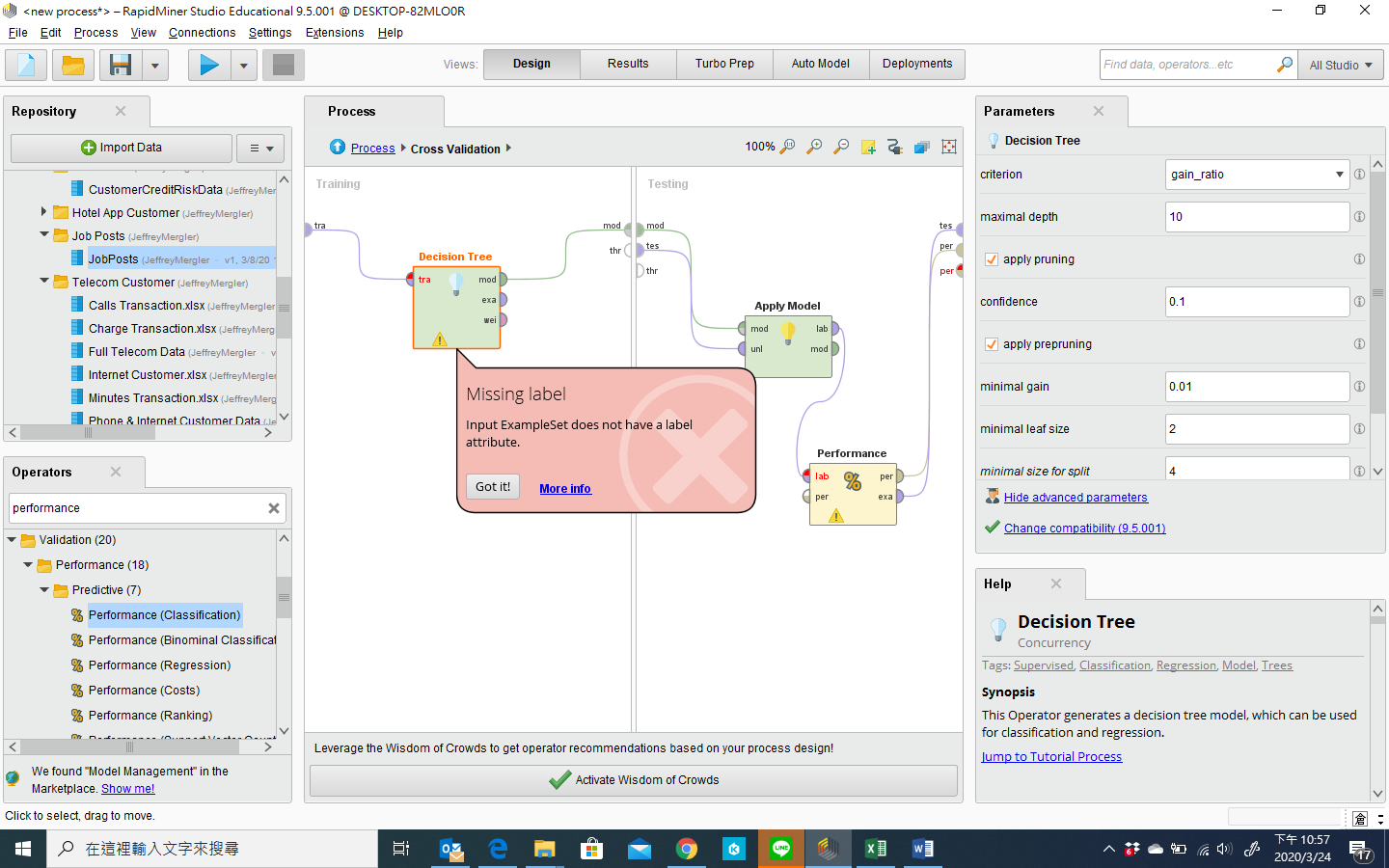
(一)在Cross Validation中的parameters有number of folds和 sampling type，分別設定如下:

**number of folds:**500，這邊的數值我設為500，因為如果可以讓模型在不overfitting的情況下，可以做多次的fold-Cross Validation以增加模型的robust，因為資料有506筆，在這邊就設定number of folds為500。

**sampling type:** automatic，會設定automatic是因為這個方法算計算出來的RMSE最低。

(二)使用linear regression 來進行回歸預測，透過Performance operator輸出RMSE和absolute error

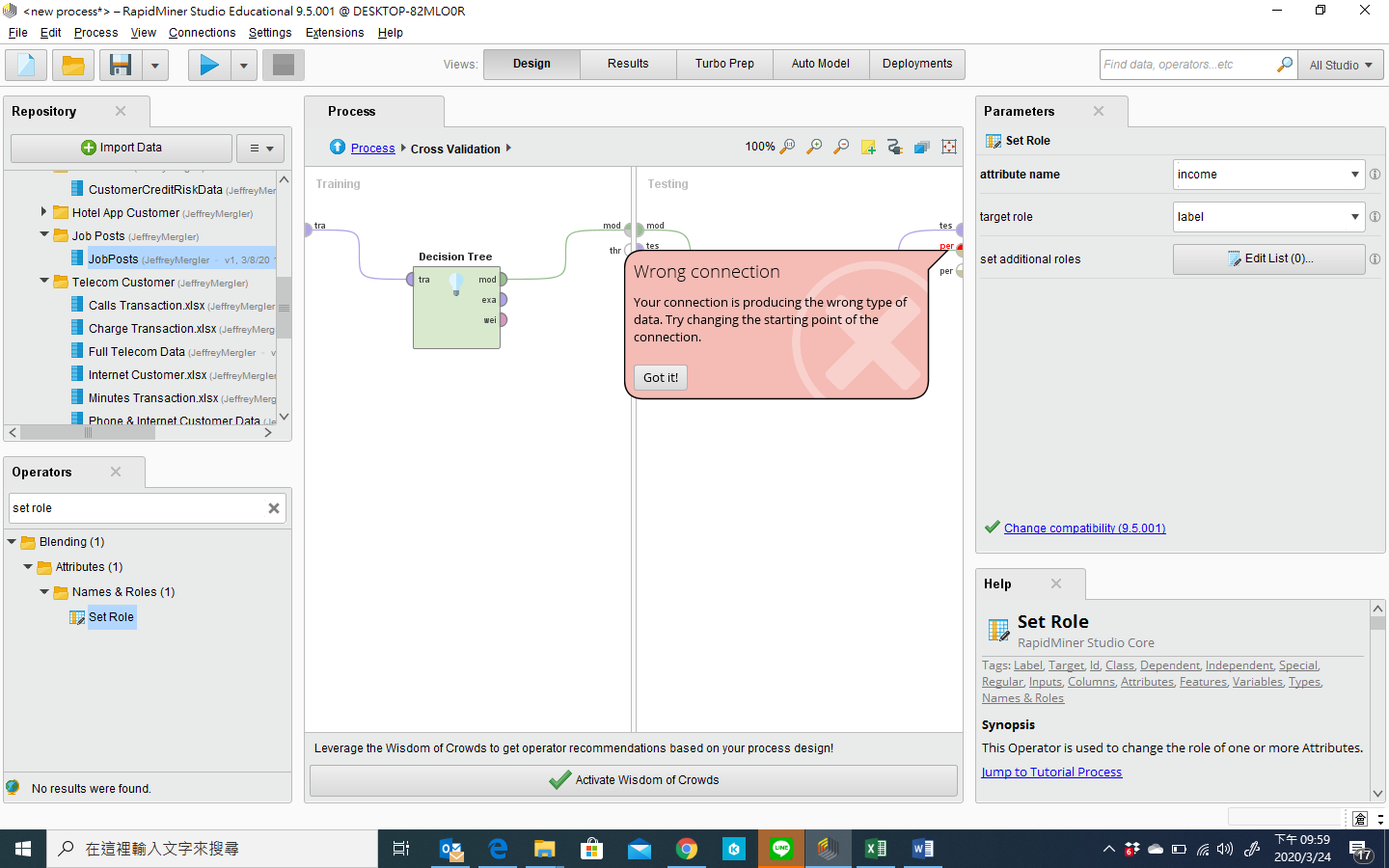
**4. 請問以下錯誤訊息是指? 該如何解決(5%)**



少了一個標籤，只有將資料檔案讀進去，沒有告訴電腦要預測哪一個欄位

解決方法:增加Set Role的operator將MEDV添加標籤(label)

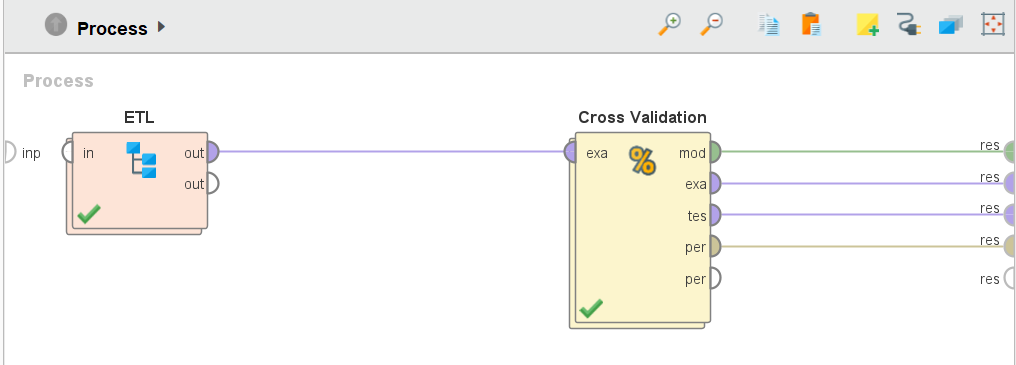
**5. 請問以下錯誤訊息是指? 該如何解決(5%)**



Wrong Connection的錯誤是指說連線的資料屬性不同，以上圖為例，最後的輸出要是performance的連線，但可能連到test的輸出線，就會造成資料的類別不同造成的錯誤。

解決方法:將performace的per輸出連到per，exa連到tes

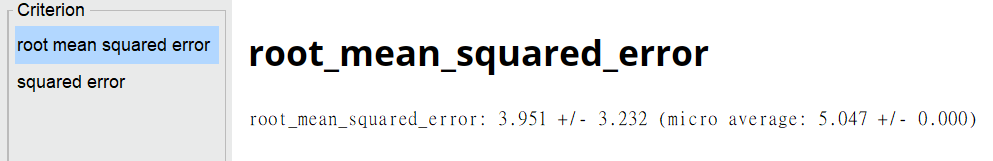
**6. 若您的模型在建構過程中有上述問題，請試著解決並以MEDV為預測標的，貼出您的決策樹,並加以說明您的方程式意涵(10%)**

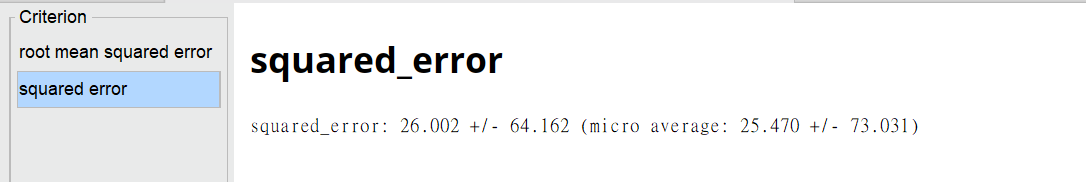
****

先進行ETL處理，這一個部分我先將資料匯入，並取出需要分析的欄位，並把要預測的欄位進行標記，就可以將預處理後的結果當作訓練模型的資料，在訓練模型的部分是使用Cross Validation，本次透過Linear Regression來訓練模型，因為資料筆數為506，所以這邊設定numbers of folds 500來增強訓練出來的模型，最後輸出模型的預測結果。

**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(10%)**







1. **root mean squared error:** 是一種誤差(Loss function)**，**觀測值(y)與真實值()之間距離的平方平均的平方根，越小(誤差越小)表示模型越正確，取根號的目的是讓單位變得一致，所以會比較直觀

公式:

1. **squared error:** 是一種誤差(Loss function)，預測值(y)與真實值()之間距離的平方和平均，越小(誤差越小)表示模型越正確，主要以平方來避免誤差正負的互相抵銷，但也因為平方的特性，所以當單一bias大的時候會有懲罰作用，對於極值會相對敏感。

公式:

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5個變數進行預測,您會選擇(請列出變數名稱)? (5%)**

1.B

2.DIS

3.LSTAT

4.PTRATIO

5.RM

MEDV

**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)? (5%)**

1.AGE

2.PTRATIO

3.RM

MEDV

**8-3您的選擇依據是? (5%) 變數變少後,正確率的變化是? (5%)**

5個變數: 3.504

3個變數: 3.860

正確率變化: 0.356

**9.最後，在５個變數的限制下，請調整參數，找出一條正確率最高的迴歸方程式，並請說明您試過那些參數的調整？ (10%)**

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

1. 五個變數使用:B、DIS、LSTAT、PTRATIO、RM，這五個變亮則是每一個變數慢慢調整，找出一個最佳的結果，在調整的過程中，又以PTRATIO、RM這兩個變數最為重要，如果移除這兩個參數，對正確率有比較大的影響。
2. number of folds:500 ，我從100,200,300,400,500慢慢調整，每次以100為單位
3. sampling type: automatic，這個部分則是找出一個使RMSE最低的類型，最終則選擇automatic