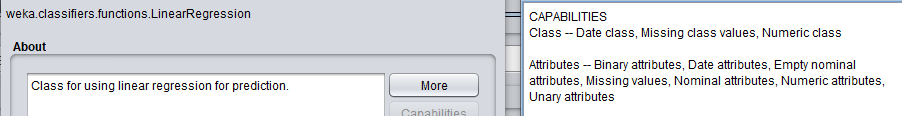
ECT HW3

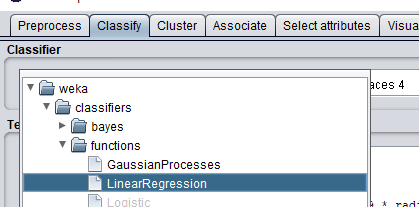
**Q1**

LinearRegression:

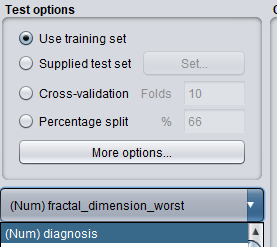


首先，先確認LinearRegression的使用條件，可以看出他的Output分類需要Numeric，因此我們在做LinearRegression時用NUM的檔案

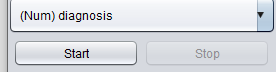
* 



選好檔案後，在Classify的function中找到LinearRegression

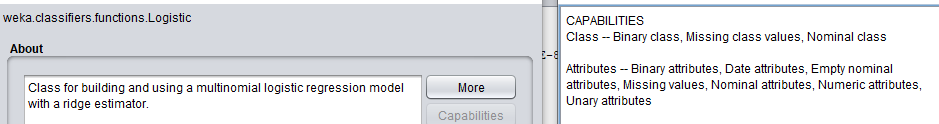


依題目要求，設定Use training set和output為diagnosis



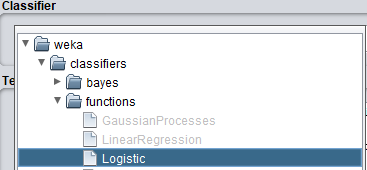
點選start則開始分析。

Logistic:

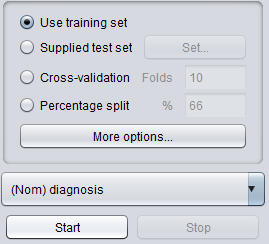


同樣查看使用條件，可以發現他的分類為Numinal，因此選用NOM檔案

* 



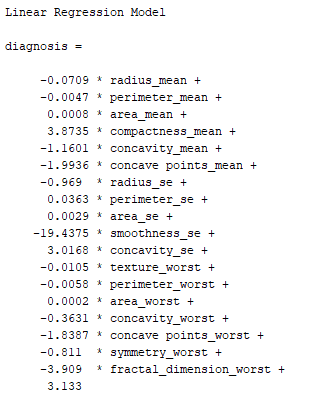
在Classifier中選取Logistic



依題目要求，設定Use training set和output為diagnosis。

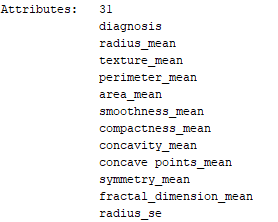
注意!這裡output已經變為Nominal。點選start即可開始分析。

**(a)**

****

由LinearRegression跑出來的結果如上所示，正相關代表參數值上升，會使output值也跟著上升，反之則是負相關，因此可得知上圖中「+」的參數為正相關、「-」的為負相關。整理得知有6個因素正相關、12個因素負相關。

**(b)**

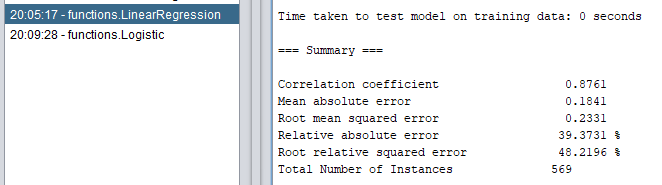


由此圖可知，input + output的屬性有31種，因此是使用30種屬性來做預測。由題(a)可得知用有(6 + 12)個因素會影響output，因此有30-18 = 12個因素不會是無相關的。例如 :

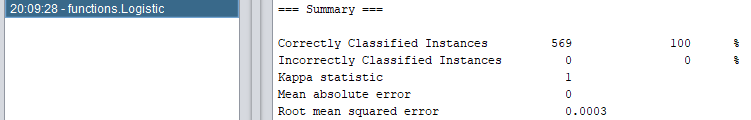
1.texture\_mean 2.smoothness\_mean 3. symmetry\_mean

**(c)**

For LinearRegression:

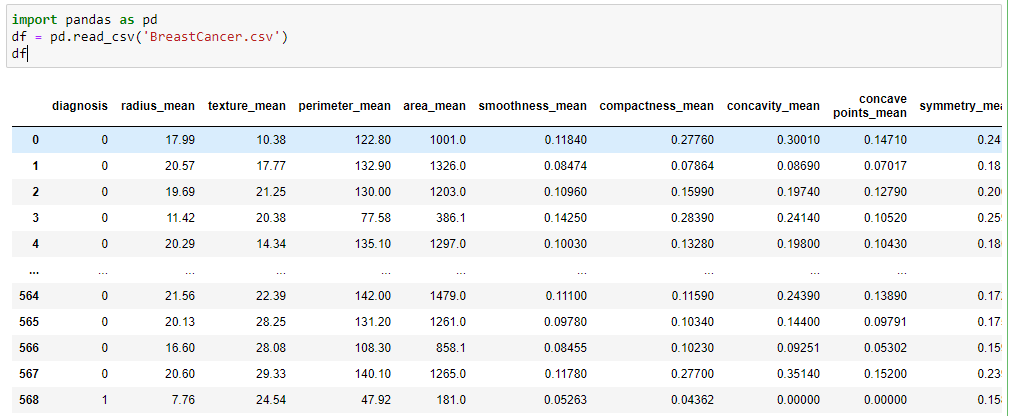


For Logistic:

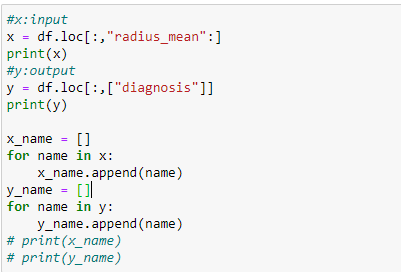


可以看出LinearRegression的均方根誤差 = 0.2331、Logistic的均方根誤差 = 0.0003，因此從誤差越小預測越準確的觀點來評價，Logistic的model表現得較好。但我們全部都是使用training set的資料來訓練、預測，因此無法保證在實際testing時會有相同的結果，雖然應該會差不多。

**Q2**



* 先讀取csv檔，並查看資料大致樣貌。

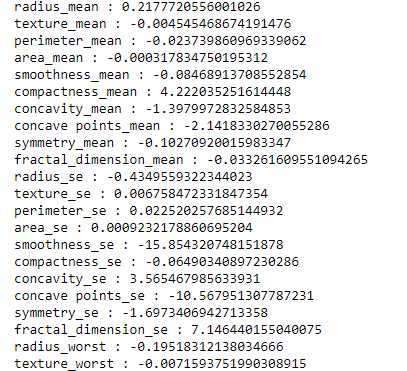


* 用loc函數把資料切成input、output，並且把每個attribute記錄下來，方便之後比對各attribute的係數。

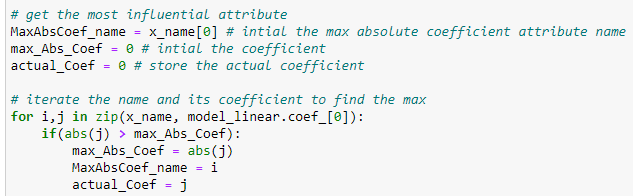


* 用LinearRegression初始化model，再用fit函數把input、output放進去訓練





* 用之前建立好的attribute\_name列表，加上coef函數取得每個attribute的係數，來印出attribute – coefficient的關係。



* 此段程式碼是為了找出「影響程度最大」的參數，也就是「係數的絕對值」最大的參數。前3行為初始化參數，用以記錄目前的最大值，並記錄實際的係數為何；後段的for迴圈用以找出絕對值後最大的係數。



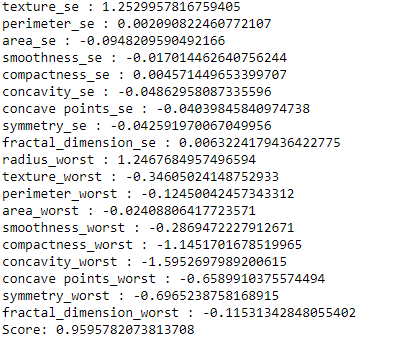


* 純粹用以觀察結果，可以看出最有影響力的係數為何、係數值為多少，並為此模型進行評分。



* 進行logistic training時也是先用LogisticRegression()函數初始化model，再用fit把input、output放進去。注意這裡有使用ravel()函數用來把數據拉成1維。雖然不使用這個函數仍然可以運行正確，但會跑出一個警告，告訴我們她期望的數據形式是1維陣列，因此做此修正。

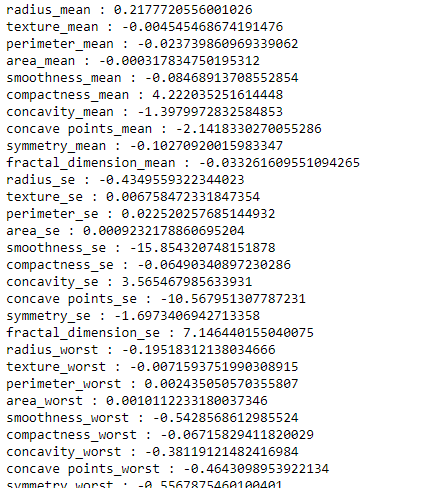




* 一樣印出結果來觀察各個係數、model的分數

**(a)**



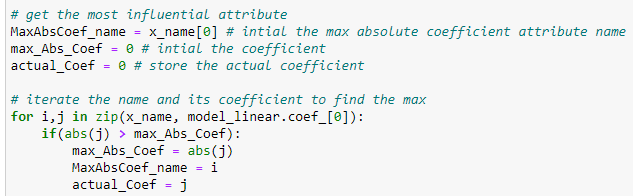


* 由上圖可知LinearRegression model中

「area\_mean」的係數 = -0.000317834750195312

**(b)**

題目要求為「影響最大的因素」，此影響應該無關「正」、「負」，因此在此定義影響最大的因素為「attribute的係數取絕對值後的最大值」，也就是哪一個維度的屬性變化後，能最有效的影響output。



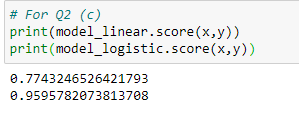




* 如上圖所示，初始化3個參數來，分別記錄當前取絕對值後的最大值、屬性名稱、實際係數。用for迴圈來做比較並更新參數。最後把它印出來。

由此可知影響最大的因素為「smoothness\_se」，它的影響程度約有-15.85左右的係數大小，只要這個增加一個小單位，就會使output下降許多。

**(c)**



* 由此圖可看出，logictic model的正確率較高