轉譯醫學期末報告 分析中心

組長：柯雅云

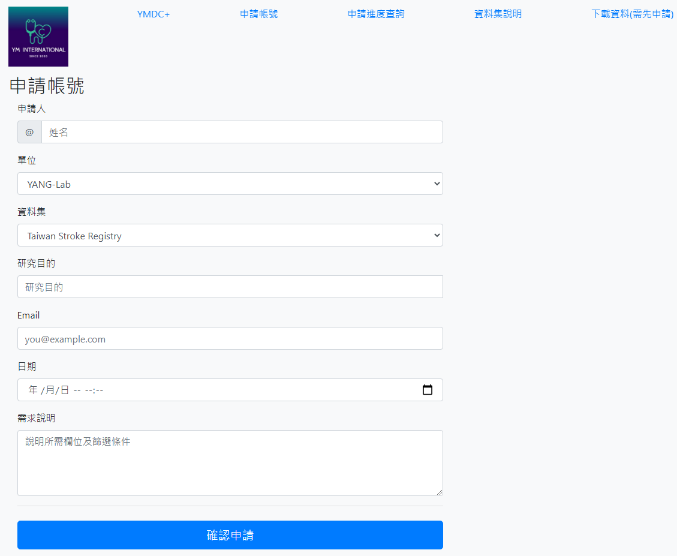
組員：盧嘉宏、朱子羽、黃千瑜、溫建泓

1. 目的
2. 資料介紹
3. 資料處理
4. 建立模型
5. 使用模型介紹
6. 中風機率公式尋找
7. 閾值訂定
8. 組間互動
9. 分工及問題討論
10. 目的 (負責組員：盧嘉宏)
    1. 建立模型
    2. 在科學研究的發展當中，模型在建立與實驗的檢驗中扮演相當重要的角色。透過模型，可以將複雜的關係歸納並有結構的同整，以方便理解或是更進一步的應用。此研究利用美兆健康資源中心所建立的中風病患資料庫進行資料探勘、分析，尋找與中風相關的危險因子，並建立中風風險的評估模型。
    3. 分析流程圖
11. 資料介紹 (負責組員：盧嘉宏)
12. 資料取得

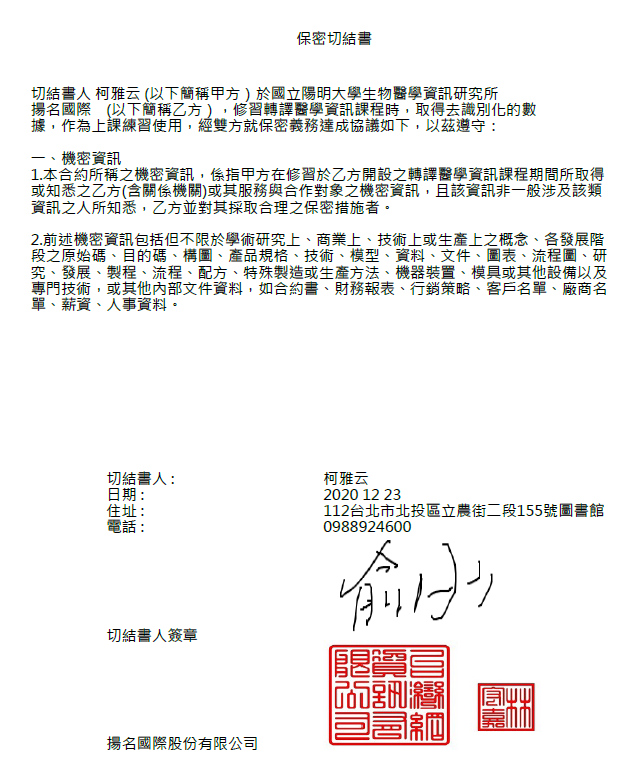
* 向數據中心取得raw data:
* 數據中心會先進行初步的資料清理，並上傳至Digital Medicine Alliance Data Center (<http://120.126.47.101:8001/>) 以供下載使用。
* 前往Digital Medicine Alliance Data Center:



* 於網站申請raw data前需註冊帳號:



* 待帳號申請結束，登入並簽署電子保密協定書:

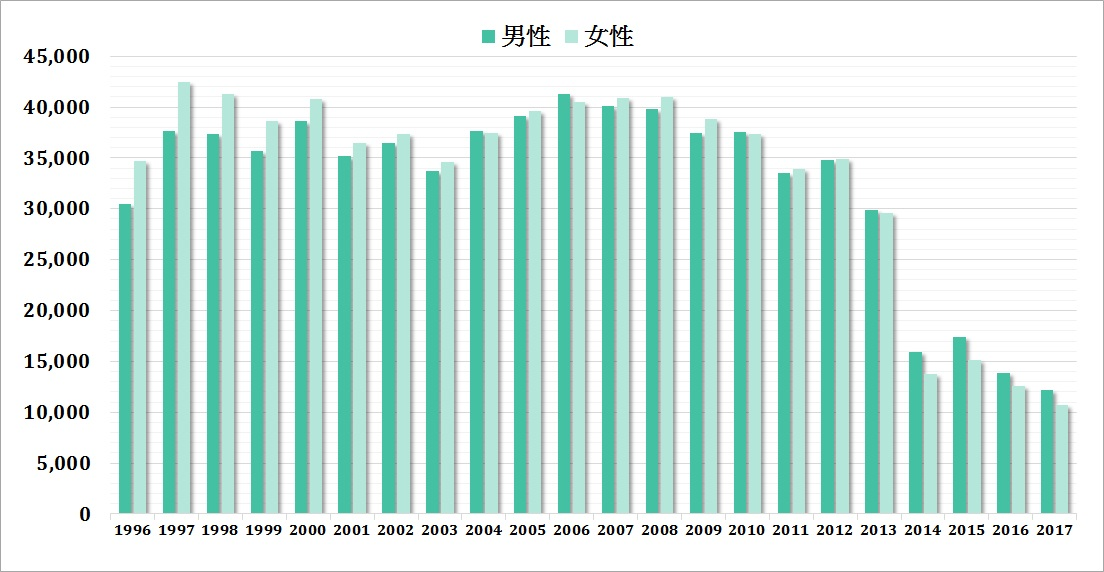


* 簽署電子保密協定書完畢，得以存取raw data:

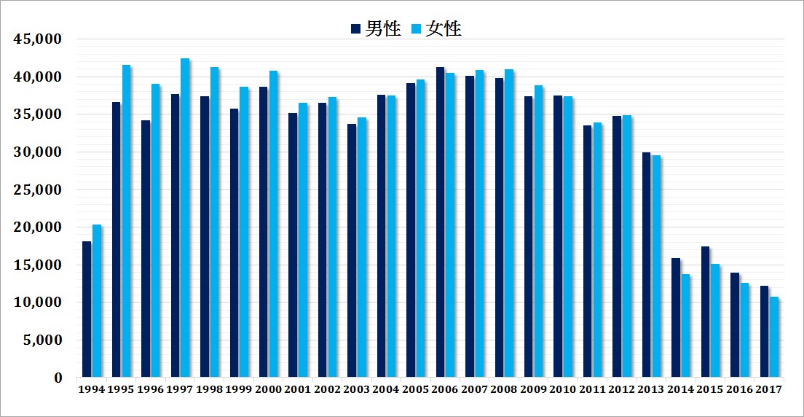


1. 資料來源介紹

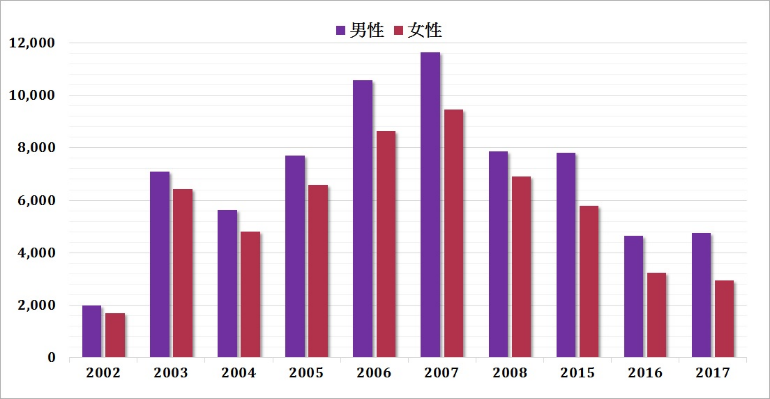
* 此研究所使用之資料來自於美兆健康資源中心的中風數據資料庫，其中資料型態包含基本人口學資料、健康問卷、人體生物以及生化數據等資料，總共約145萬筆資料以及489個欄位。每一個資料庫均由專人專職負責，以控管資料庫之品質，數據資料亦會定期進行更新，以提供最完整而豐富的醫學研究資源。
* 美兆健康問卷資料庫
* 此健康問卷，分為五大類，總共有134道題目，內容包含個人基本資料、個人及家族病史、近況調查、生活習慣及飲食習慣。自1996年至今已經有相當巨量的資料。美兆健康資源中心的資料與檢體之對象包含健康與亞健康族群；此外，其資料庫也收納以「家庭」為單位、會員定期回檢的長期健檢資料。(下圖為每年收錄之資料筆數，引用自美兆健康資源中心網站)

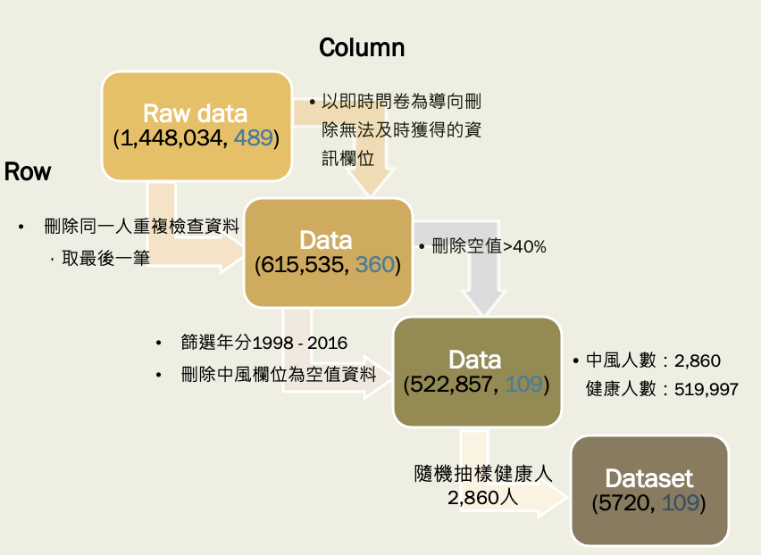


* 美兆生化數據資料庫
* 生化數據資料庫包含100餘項健康檢查項目之數據，其中包含身體基本測量、血液檢驗、血糖、血脂、肝功能、腎功能、肺功能、骨密度、腫瘤標記、甲狀腺、發炎指標、糞便、耳鼻喉、眼、心電圖、超音波、胸腹X光、婦科、上腸胃道X光等多項。並且早在1994年就已推行數據電子化。可與「美兆健康問卷資料庫」串連進行統計分析。(下圖為每年收錄之資料筆數，引用自美兆健康資源中心網站)



* 美兆人體生物資料庫
* 美兆健康資源中心設有全台灣第一所經行政院衛生福利部核准通過之民營人體生物資料庫。自2002年起開始採集血液檢體，目前已存有超過10萬人、13萬人次、總計超過100萬管的血液樣本，包含多次捐贈者之檢體。按照血清、血漿、血沉中分層三類分裝，穩定儲存於-80°C低溫冰箱。(下圖為每年收錄之資料筆數，引用自美兆健康資源中心網站)

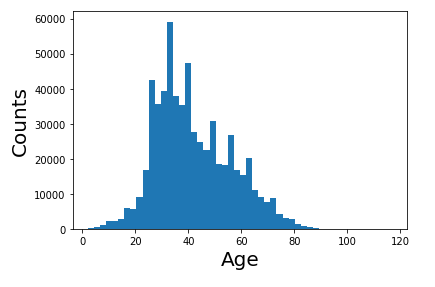


1. 資料處理 (負責組員：朱子羽)
   1. 資料處理流程：
   2. 資料篩選（空值、生理數據刪除）
2. 檢體ID：row欄位，同一人會有相同ID。
3. 篩選年份為1998-2016年
4. 資料中出現相同ID則取該ID最新的資料。
5. 中風欄位空白則刪除
6. 問卷特徵值：column欄位。
7. 刪除無法直接即時獲得的資訊欄。
8. 刪除空值 >40% 的欄位。
   1. 如何決定中風因子
9. 想法：設定目標使用族群為一般民眾，能夠觸及到APP成為使用者皆可進行此問卷式選填。
10. 過程：
11. 在經過前一點提及的資料篩選後，先進行資料搜尋方式，找尋經過實驗、研究發表與中風相關的特徵。
12. 使用不同的方法 ( ExtraTreesClassifier, RandomForest, LogisticRegression, GaussianNB, DecisionTrees  )挑選並檢測importance score 數值高者。
13. 挑選到的因子

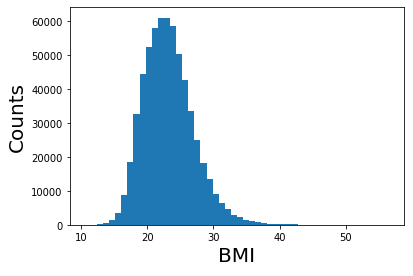
* 可變型因子：是否喝酒、BMI、右手收縮壓、糖尿病、高血脂。
* 不可變型因子：年齡、性別。

1. 參考資料：WHO官方內容及其參考論文，NIH官方內容及其參考論文, *Stroke Risk Factors, Genetics, and PreventionAmelia K.2020*
   1. 如何補缺值
2. 連續型資料：以平均值補缺失值。

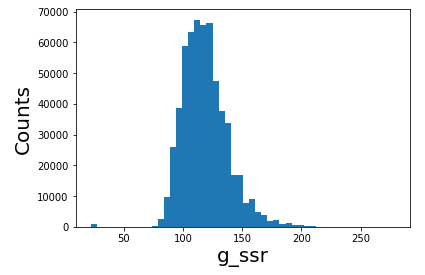
* 年齡：缺值補上所有人的平均年齡。

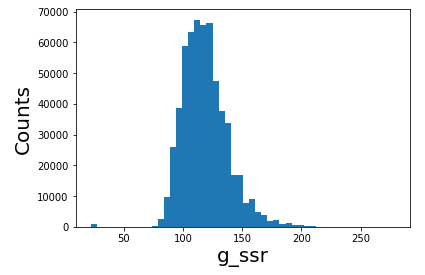


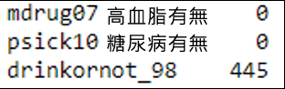
* BMI：男女分開，各自補上各性別的平均BMI。



* 右手收縮壓





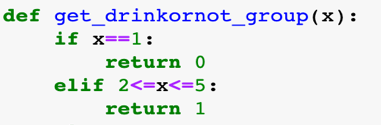
1. 類別型資料：

* 高血脂有無：無缺值
* 高血壓有無：無缺值
* 是否喝酒：共有445筆缺失值
* 原本《是否喝酒》問券內容如下：

分為(1)~(5)，我們在處理時希望能夠簡化複雜度



* 所以將其歸為0, 1( 0：不喝或每週少於一次; 1：有喝酒)，而缺失值的部分補為0, 1中數量多的那一組，這在邊我們是補0。



* 1. 進行資料集抽選：（隨機抽樣）
* 因為在我們的資料中，健康的人數為 519,997; 中風的人數為2,860，兩個族群的數量差異甚鉅，所以在這邊進行隨機抽樣。
* 在兩個族群中分別各取，2,860人次，形成5,720的資料集，再進行後續的建模分析。

1. 建立模型 (負責組員：黃千瑜)
2. 原始的資料經過上述的資料處理及分析方式後，我們便開始建立模型，以下為我們整個模型的流程圖。

Split data

**(test\_size=0.2)**

Build models

**(kfolds= 5)**

Analyze results

**(visualization)**

Validation

**(logistic regression)**

1. 第一步先將資料接下來進行切割，並把其中20 % 當作test data，另外80%去做train data。

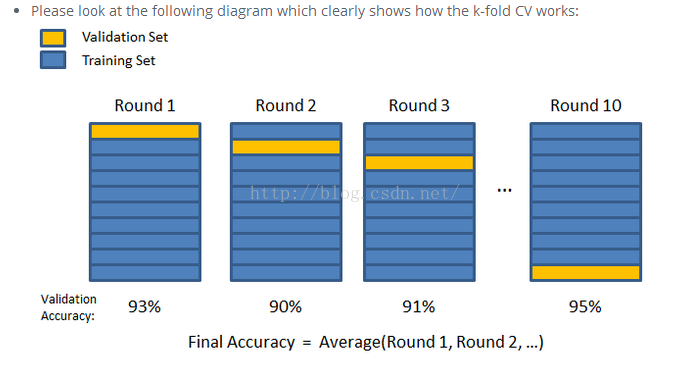
train data

test data

1. 之後分別利用5種模型做訓練 (模型有Gaussian naive Bayes, Logistic Regression, Decision Tree, Random Forest and Extremely randomized trees)，並且設定random state 以使每一次進行模型訓練時，結果會一樣。



1. 使我們可以透過訓練模型後，挑出其中最好的模型，來作為我們預測中風模型。其中我們利用k fold cross validation的方式來訓練模型。

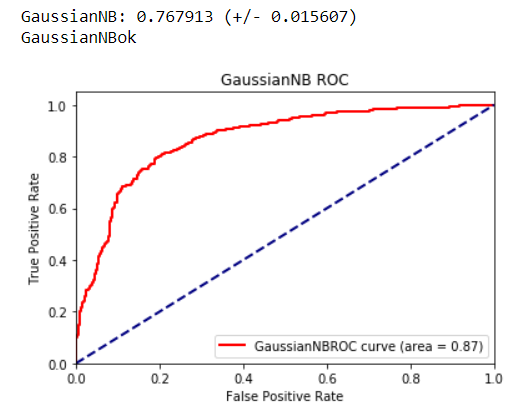


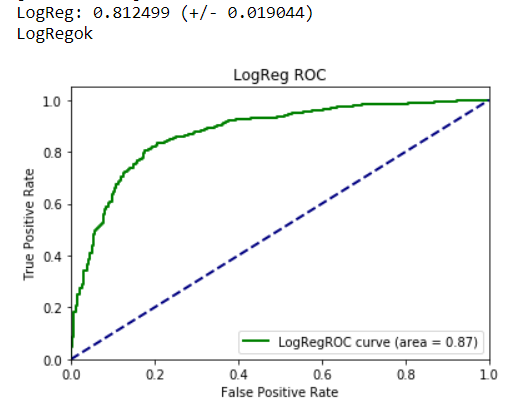
在K fold cross validation中，使用不同的資料組合來驗證你訓練的模型，如上圖(3) 所示我們將資料分成10等份，其中第1等分用來當作驗證的測試資料，其餘9份拿來訓練，下一輪我們繼續將第2等分拿來當作驗證的測試資料，其餘9份一樣拿來訓練，總共做10次。

不斷重複這個動作，這樣你可以得到不同的訓練/測試資料組合，提供更多數據去驗證。

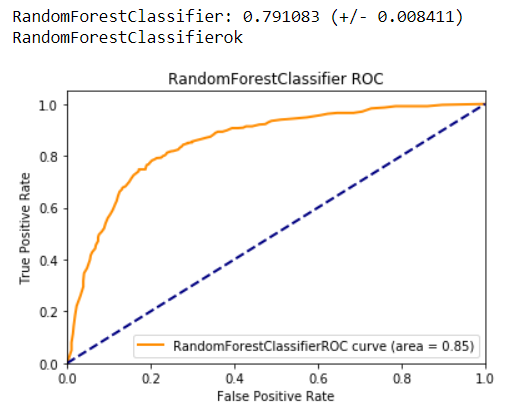
藉著將10次的準確性(Accuracy)平均，而這個得到的平均值，我們可以自信的說這個數值就是我們的準確度沒有偏差。

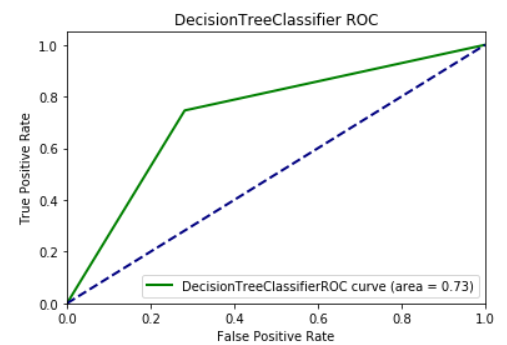
1. 之後五種建模的結果在使用ROC curve和Accuracy去評估模型的結果
2. 最後，訓練玩模型之後，匯出每個模型的ROC curve和Accuracy，去比較哪一中模型的效果最好。訓練結果如下：

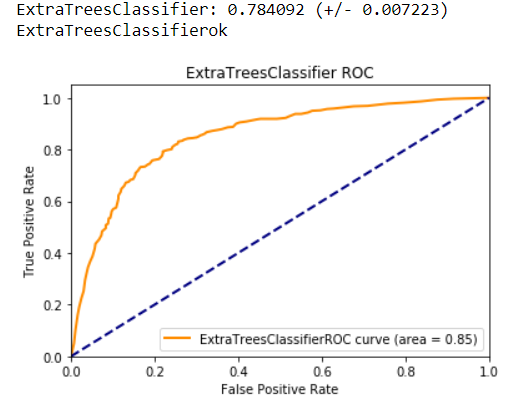
* **Gaussian naive Bayes**
* **Logistic Regression**

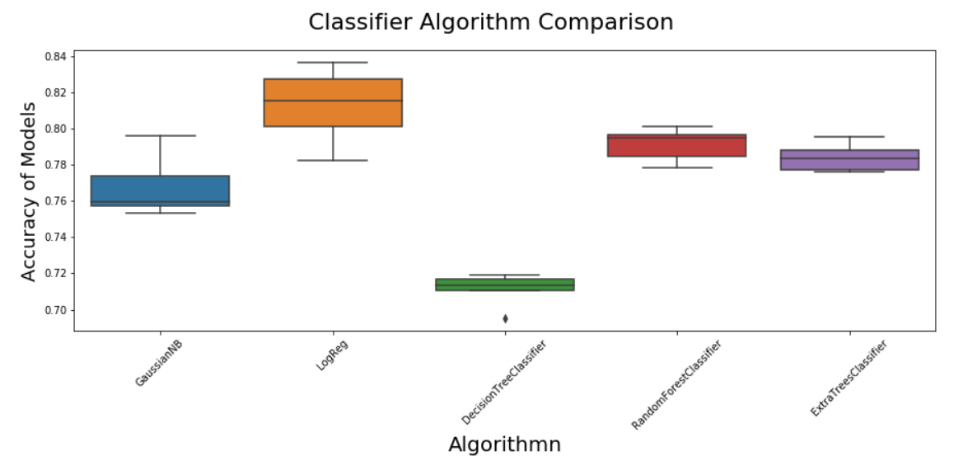


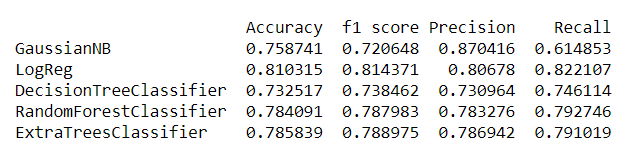
* **Decision Tree**



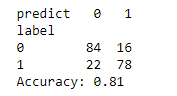


* **Random Forest**
* **Extremely randomized trees**
* 再來將每個模型的預測結果呈現在盒形圖上。





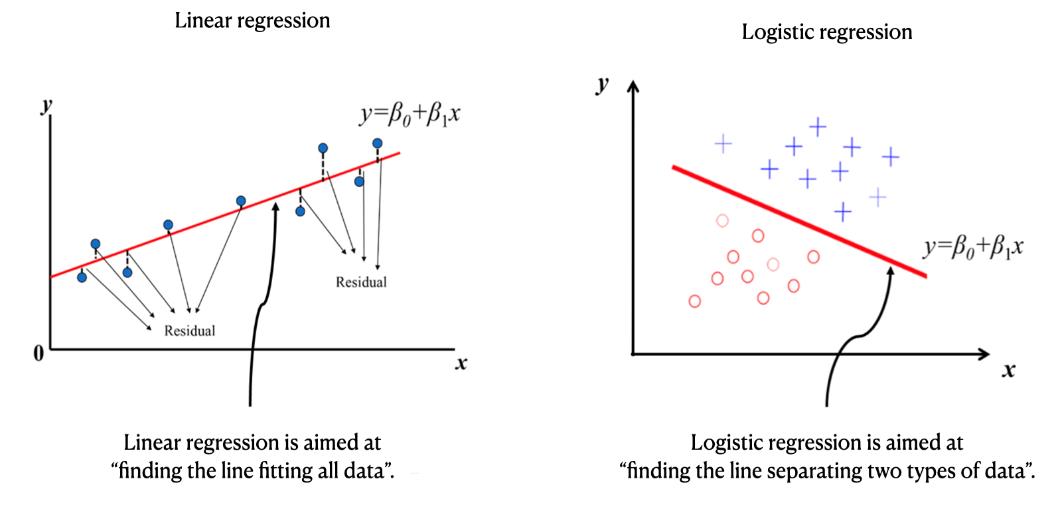
* 由上述的結果得知，Logistic Regression模型的表現最佳，所以我們最後決定用Logistic Regression來做我們的中風預測模型。
* 接下來是匯出模型並作驗證，利用剛剛一開始分的test data來做驗證。



1. 使用模型介紹 (負責組員： 溫建泓)

因為本實驗已經將Logistic regression 作為中風預測的模型，所以在此詳細介紹Logistic regression ，而其他的模型則簡述。

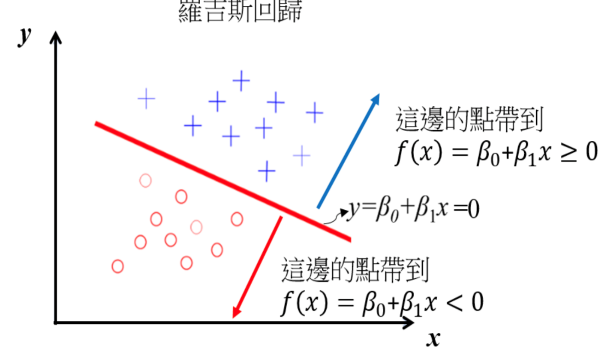
1. **Logistic Regression** 
   1. Logistic Regression 和我們最常見的 Linear regression 不同，前者是為了找出一條線區分兩種不同的資料輸出的是 1 或 0 ，而 Linear regression 是它一個變數 它會輸出預測的 ，示意圖如下：



* 1. 另外，Logistic regression 其實和 Linear discriminant analysis 有幾分神似，因為它們的目的都是分群，只不過前者只能分兩群，而後者是可以分出多群的。
  2. 下面列出的是 Logistic regression 和 Linear discriminant analysis 兩者在二分法上面的差別：

|  |  |
| --- | --- |
| **Logistic** **regression** | **Linear discriminant analysis** |
| Based on Maximum likelihood estimation | Based on Least squares estimation |
| Estimates probability immediately and conditionally | Estimates probability mediately via classificatory device which uses both conditional and marginal information |
| Not so exigent to the level of the scale and the form of the distribution in predictors | Predictirs desirably interval level with multivariate normal distribution |
| No requirements about the within-group covariance matrices of the predictors | The within-group covariance matrices should be identical in population |
| The groups may have quite different | The groups should have similar |
| Not so sensitive to outliers | Quite sensitive to outliers |
| Younger method | Older method |
| Usually preferred, because less exigent / more robust | With all its requirements met, often classifies better than Logistic regression |

* 1. 那具體來說，Logistic regression 是如何做回歸的呢？

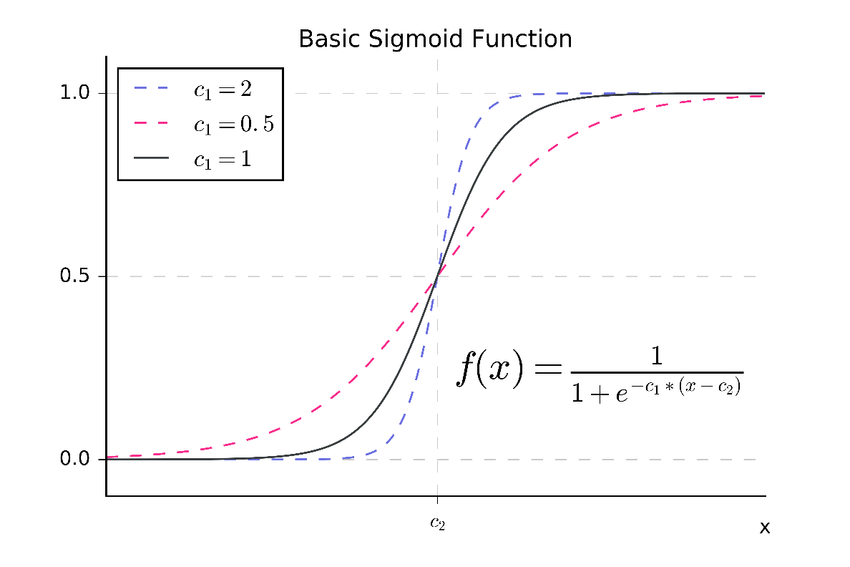


Logistic Regression

Bring these points to

Bring these points to

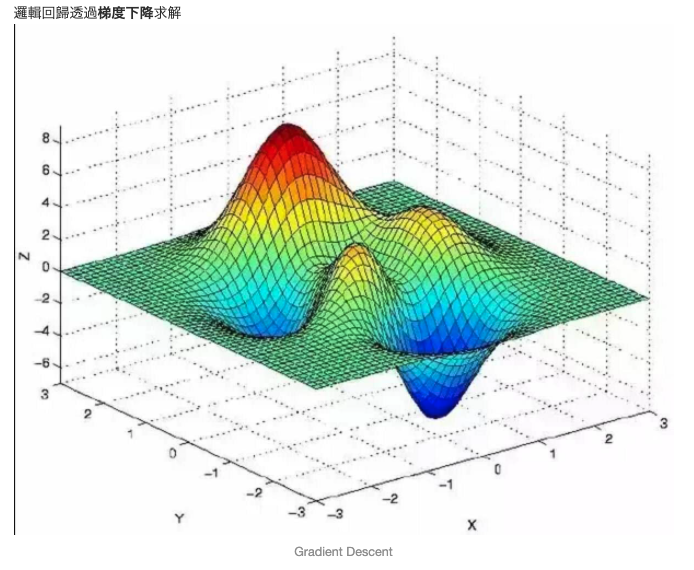
* 首先我們要知道 Logistic regression 的目的，是為了很好的區分出兩種資料，那在回歸的時候就必須將回歸線附近的點分到其中一邊，如上圖所示。
* 而 Logistic regression 用的原理有三種：Sigmoid function、Maximum likelihood estimation 和 Bernoulli distribution，Sigmoid function 是為了讓靠近 0 的資料，因為理論上越容易分類錯誤，所以這時候我們加上一個對數函數，這時候這個輸出就更有彈性。



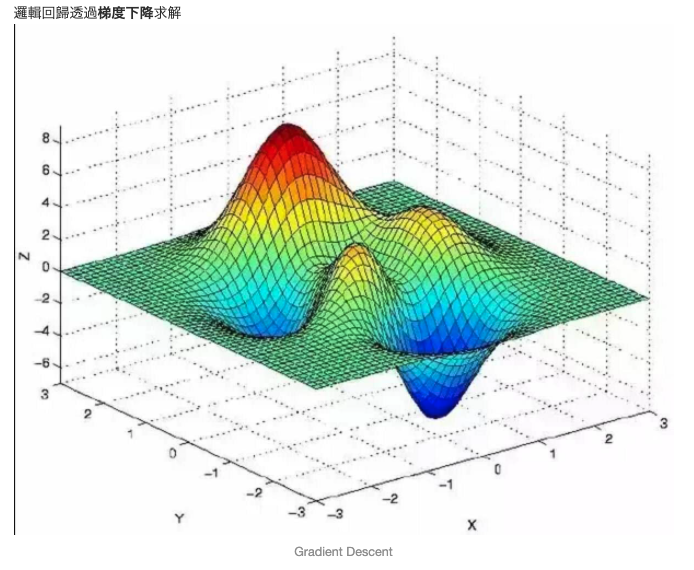
* 綜合上述，Logistic regression 的公式其實就是：
* 而之前提到的 Bernoulli distribution，和 Logistic regression 有什麼關係？因為 Logistic regression 的目標是做二分類問題，這時候的輸出有兩個，target 和 non-target 機率，但機率的輸出最後還是要判斷是不是 target ( 是 target， 是 non-target)。
* 再將它和 Maximum likelihood estimation：
* 做結合取 後化簡，因為取的是負號所以變成最小概似函數：
* 然後再透過對 的偏微分找梯度，利用梯度法找出適合的 ：

是學習率，合適的 可以更快的找到解，過大可能找不到，過小會找太久。

* 1. 但是 Logistic regression 本身是有一些問題的，如圖，假設我的起點在紅色的頂端，透過梯度法我會一路找到局部極小值，而不是絕對極小值(深藍色區域)

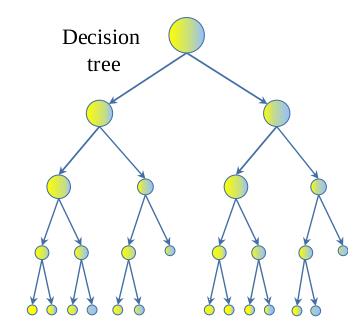


如果你的人品夠好起點設在該設的地方那你就可以得到絕對極小值，如圖。



但是，總體而言，Logistic regression 在現今的回歸法中還是有相當不錯的表現。

1. **Gaussian naive Bayes**
2. Gaussian naive Bayes 基於一個簡單的假定：給定目標值時屬性之間相互條件獨立。 Naive Bayes是生成方法，也就是直接找出特徵輸出 和特徵 的聯合分佈，然後用得出。
3. 公式：
4. 假設訓練集中包含連續值 ，我們按照類別將資料分類，並計算每個分類的均值和偏差。 是對應類別 下 值的均值，是方差。
5. 假設我們已經收集到一些觀測值 。在給定分類 下 的概率分佈 可以通過將 帶入到由 和決定的高斯分佈公式中得到。
6. **Decision Tree**
7. 其實 Decision Tree 我想大家都蠻熟悉的，在應用上，不是在原理上，他大致上的示意圖如下：



每一個點代表一個條件，符合條件者走一邊，不符合條件者走另一邊，最後走到的點就是這個模型決定的結果。

1. Decision Tree 在訓練時會從最後上方的樹根開始將資料的特徵將資料分割到不同邊，分割的原則是：這樣的分割要能得到最大的資訊增益。

資訊增益：

二元分類資訊增益：

由於我們希望獲得的資訊量要最大，因此經由分割後的資訊量要越小越好。

1. 而常見的資訊量有兩種：熵(Entropy) 以及 Gini不純度(Gini Impurity)

Entropy ：

Gini Impurity：

1. **Random Forest**
2. Random Forest的基本原理是，結合多顆CART樹 (CART樹為使用GINI算法的決策樹) ，並加入隨機分配的訓練資料，以大幅增進最終的運算結果。不過，這個方法必須基於下面的理論：**Ensemble Method**
3. Ensemble Method 的想法是，如果單個分類器表現OK，那麼將多個分類器組合起來，其表現會優於單個分類器。

* 也就是論基於「人多力量大，三個臭皮匠勝過一個諸葛亮。」，不過要滿足這理論是有條件的：

1. 各個分類器之間須具有差異性
2. 每個分類器的準確度必須大於 0.5

也就是說，可不是一群一樣笨的臭皮匠就可以勝過一個諸葛亮。

* 這些臭皮匠要符合如下的條件：
  1. 他們不能太笨 (分類器的準確度必須大於0.5)
  2. 他們彼此之間的經驗有所不同 (分類器之間具有差異性) 。

1. 必須符合這兩個條件，才能讓結合多顆CART樹的隨機森林其效力大於單一的決策樹，這種方式稱為Ensemble Method。
2. 不過由於我們的樣本只有一個，所以要形成多顆具差異性的樹以進行Ensemble Method，必須先將訓練樣本進行分化，才能產生多顆具差異性的CART樹。



1. **Extra-Trees**
2. Extremely randomized trees 是由 Pierre Geurts 等人於 2006 年提出。該算法與 Random Forest 十分相似，都是由許多決策樹構成。但該算法與隨機森林有兩點主要的區別：
3. 隨機森林應用的是 Bagging 模型，而 Extra-Trees 是使用所有的訓練樣本得到每棵決策樹，也就是每棵決策樹應用的是相同的全部訓練樣本。
4. 隨機森林是在一個隨機子集內得到最佳分叉屬性，而 Extra-Trees 是完全隨機的得到分叉值，從而實現對決策樹進行分叉的。
5. 以二叉樹為例：

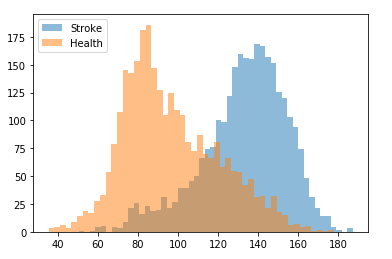
* 當特徵屬性是類別的形式時，隨機選擇具有某些類別的樣本為左分支，而把具有其他類別的樣本作為右分支。
* 當特徵屬性是數值的形式時，隨機選擇一個處於該特徵屬性的最大值和最小值之間的任意數，當樣本的該特徵屬性值大於該值時，作為左分支，當小於該值時，作為右分支。
* 這樣就實現了在該特徵屬性下把樣本隨機分配到兩個分支上的目的。然後計算此時的分叉值 (如果特徵屬性是類別的形式，可以應用基尼指數；如果特徵屬性是數值的形式，可以應用均方誤差) 。
* 遍歷節點內的所有特徵屬性，按上述方法得到所有特徵屬性的分叉值，我們選擇分叉值最大的那種形式實現對該節點的分叉。從上面的介紹可以看出，這種方法比 Random Forest 的隨機性更強。

1. 對於某棵決策樹，由於它的最佳分叉屬性是隨機選擇的，因此用它的預測結果往往是不准確的，但多棵決策樹組合在一起，就可以達到很好的預測效果。
2. 當 Extra-Trees 構建好了以後，我們也可以應用全部的訓練樣本來得到該 Extra-Trees 的預測誤差。

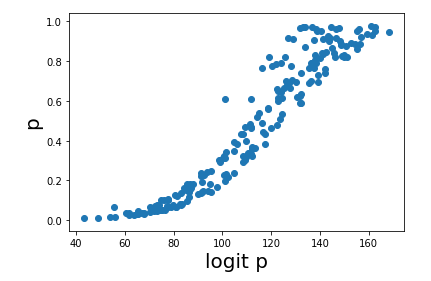
* 這是因為儘管構建決策樹和預測應用的是同一個訓練樣本集，但由於最佳分叉屬性是隨機選擇的，所以我們仍然會得到完全不同的預測結果，用該預測結果就可以與樣本的真實響應值比較，從而得到預測誤差。
* 如果與隨機森林相類比的話，在ET中，全部訓練樣本都是帶外樣本，所以計算 Extra-Trees 的預測誤差，也就是計算這個帶外的誤差。

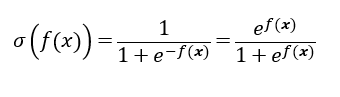
1. 中風機率公式尋找 (負責組員：柯雅云)
2. 利用logistic regression 中的權重與截距產生一個權重公式(logit p)：

1. 根據中風與未中風的人可以分別帶入公式得到一分布圖：



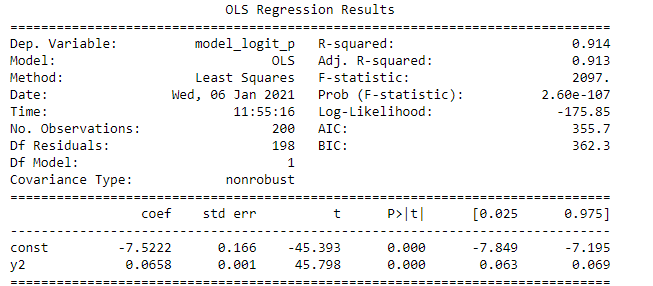
1. 利用logistic regression套件中判定為中風和未中風機率(p)與logit p 作圖可得一乙狀曲線(sigmoid curve)：



1. 根據logistic regression原理，只要將logit p算出來的值帶回以下sigmoid函數就能得出機率p
2. 但因為我們實際帶入logit p的各項數值並未經過正規化(在APP端蒐集資料時也無法及時正規化)，無法直接求出p值，我們不清楚套件中實際正規化的方式，也查詢了許多方法例如sklearn.preprocessing.normalize中的正規化方法，但都無法找出合理的機率值，於是我們使用logistic regression套件內判斷為中風和未中風的機率(p)去反推套件中的logit p2，預期logit p2 與logit p有一線性關係，由上面公式推導可得出以下公式：
3. 由以下圖得知套件內的logit p2 和我們得出的logit p 大致呈一線性關係：

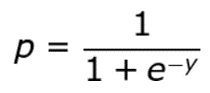


1. 利用統計方法找出此回歸線，並檢定是否顯著：



由以上回歸檢定可得出一方程式：

1. 接著由以下三步驟，即可得出此人的中風機率，以下示範使用方式：
2. **公式一：**
3. **公式二：** 



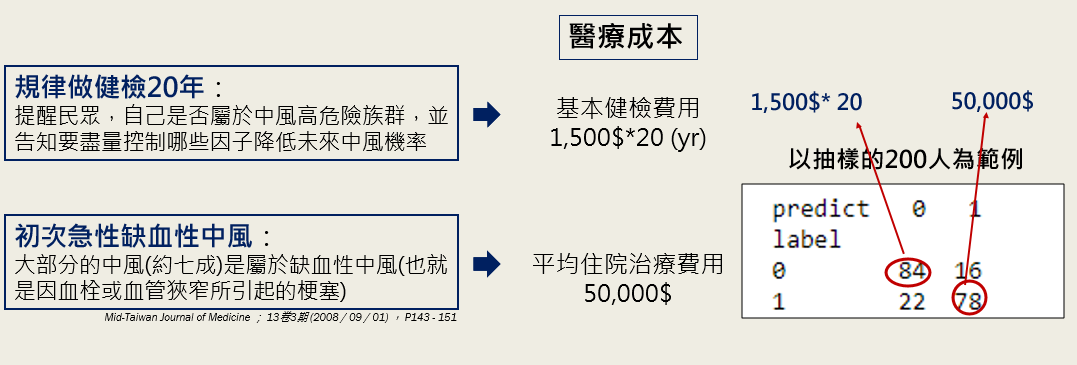
1. **公式三：**
2. 假設Table 1為app組蒐集的一筆資料：

|  |  |
| --- | --- |
| Table 1 | |
| gender | 2 |
| age | 52 |
| sbp | 118 |
| bmi | 21.09 |
| lipid | 0 |
| DM | 0 |
| drink | 1 |
| gender | 2 |

1. 代入公式一得logit p = 112.4046
2. logit p= 112.4046 代入公式二得y = -0.1257
3. 代入公式三得**中風機率p = 0.468**
4. 閾值訂定 (負責組員：柯雅云)
5. Model 預設閥值0.5有著最高的準確率，但對我們的問題來說卻未必是最佳的閥值

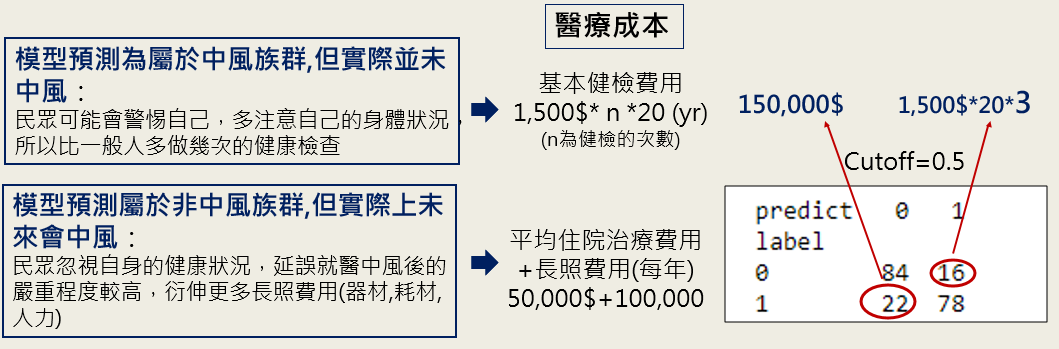
(不一定是最省成本的方式)，在篩檢的概念上，我們不希望這個人其實有病但卻篩檢不出來預測出他沒病，所以會容許偽陽性率(false positive)高一點，若我們以醫療成本為考量去訂出閥值。

1. 根據模型的Confusion\_matirx，分為四群人，可大致估算其預花費的醫療費用，目的要降低醫療成本。

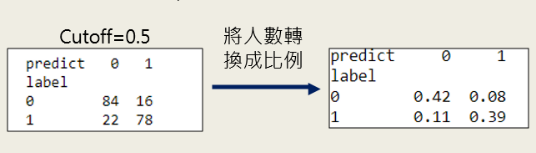


1. 這裡基本的健檢費用是以醫院平均基本健康檢查費用，預測為不會中風實際上也沒有中風的這群人，只會有基本的健檢費用，透過不斷推播中風相關知識可以提醒民眾注意自己的身體健康，降低未來的中風發生率。

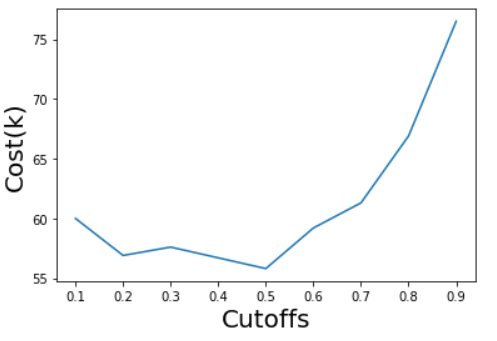
在預測會中風的病人實際上未來也會中風，中風平均治療費用是參考國內文獻[Mid-Taiwan Journal of Medicine；13卷3期2008 P143 – 151]，訂定出來的粗估費用，實際上應該更高。



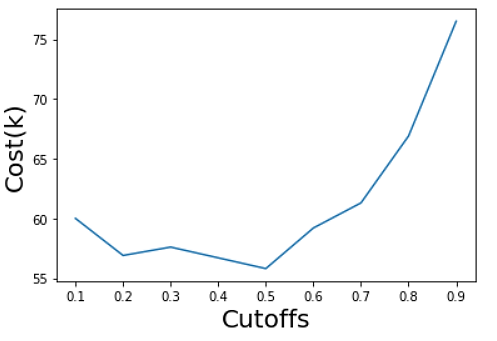
1. 模型預測為中風族群但實際上並不會中風的族群，民眾可能會更注重得自己的健康並多做幾次檢查，這裡假設一年做了三次檢查，另一群為最不樂見的情況，預測沒有中風但實際上未來會中風的族群，民眾可能會忽視自身健康狀況，延誤就醫，使得中風嚴重程度較高，因此而衍生許多長照費用。
2. 根據以上四群人的成本，每個confusion matrix都可算出一個總成本，計算cutoff 在0.1-0.9的總花費，這裡將人數轉換為比例去做計算。



0.42\* 1,500\* 20+ 0.39\* 50,000+ 0.08\* 4,500\* 20+ 0.11\* 150,000 = **55,800 $**



1. 由圖顯示cutoff在0.5時有最低的花費，但當cutoff為0.2時成本只高了一點點，卻能篩出更多的人，故我們依此兩點中風機率又分為三群分別為低、中、高危險族群

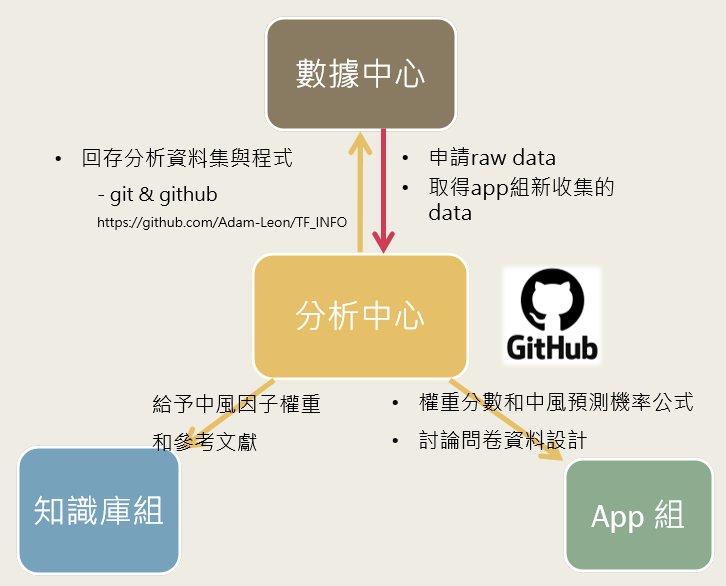


**中風中度危險族群**

**中風低危險族群**

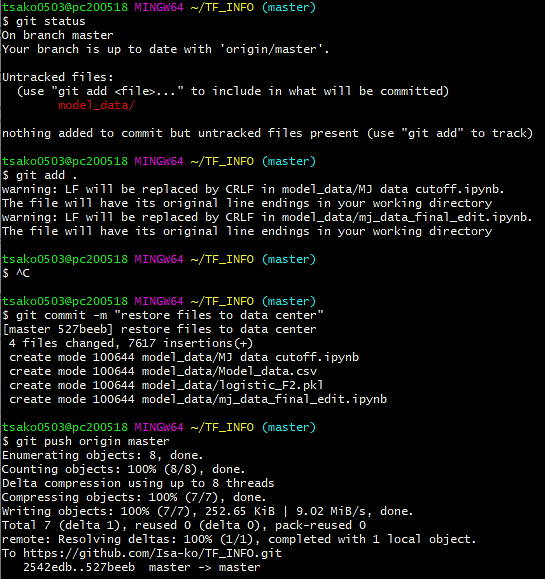
**中風高危險族群**

1. 組間互動 (負責組員：柯雅云)



1. 與數據中心互動

* 透過git hub 存取、回存資料，回存資料包含模型建立的程式檔(ipynb)、模型匯出檔、模型dataset(.csv檔)，以及release note。
* 在本機clone一個數據中心的repository，檔案的任何修改皆不影響數據中心的原檔，直到在終端機下指令將檔案使用git push的方式回存，並在git hub 上create pull request，待數據中心同意合併才可將檔案存至資料庫，此方法可以達到跨組間的版本控管

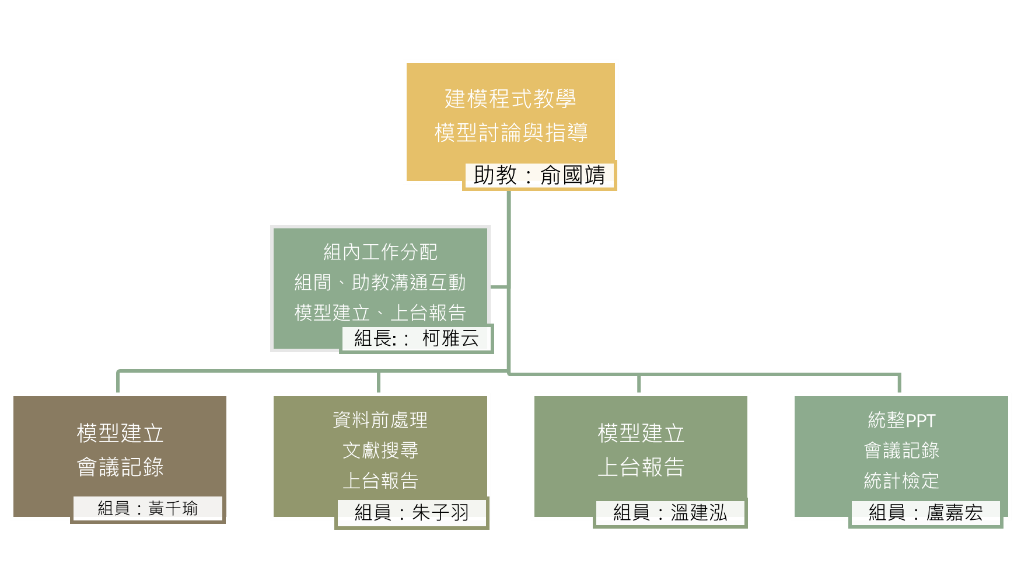




1. **與App組互動**

透過LINE傳送release note給app組，內容包含如何使用公式得出中風機率和問卷設計須包含哪些項目與限制(如Table 2)，完整檔已附在專案資料夾中。

|  |  |  |
| --- | --- | --- |
| Table 2 | | |
| feature | content | range |
| gender | 1: male 2: female |  |
| age | numeric | 6 – 95 yrs |
| SBP | numeric | 20 – 250 mmHg |
| BMI | numeric | 10 - 50 |
| Hyperlipidemia\_ornot | 0: No 1: yes |  |
| DM\_ornot | 0: No 1: yes |  |
| drink\_ornot | 0: No 1: yes |  |

1. 分工及問題討論

下圖為分工大致情形：

期末報告書寫：

1. 在每個標題後均有加上負責同學：
   * 盧嘉宏：目的、資料介紹
   * 朱子羽：資料處理
   * 黃千瑜：建立模型，統整
   * 溫建泓：使用模型介紹
   * 柯雅云：建立模型、中風機率公式尋找、閾值訂定、組間互動
2. 盧嘉宏
3. 彙整進度報告ppt、會議記錄、統計檢定
4. 問題討論:

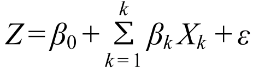
在進行Digital Medicine Alliance Data Center 的帳號註冊時，註冊完畢之後卻沒有提及密碼如何取得，經過詢問之後才得知密碼會在申請核准後的通知信當中提供。

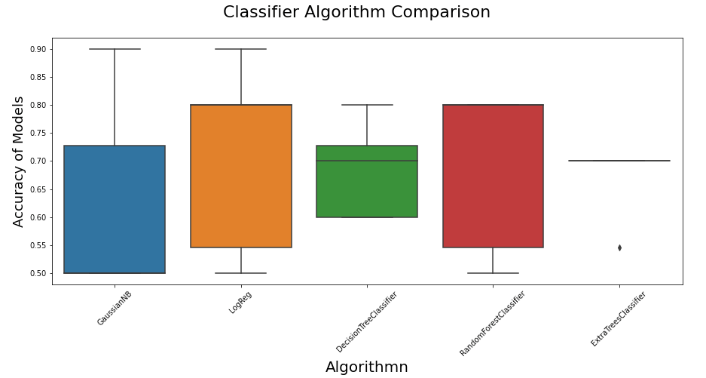
個人認為此步驟需註記在帳號申請的頁面，否則會有許多人不知道申請完帳號之後的下一步是如何。

1. 朱子羽
2. 負責部分資料前處理的整理與分析，與上台口頭報告自己所整理的部分。
3. 在最後報告完整報告後，有許多問題被提出，所以以下想討論幾點：
4. 因子的篩選：我們在篩選因子時，不是照傳統大數據分析的方式，將因子全部丟入分析，排序後挑選出重要程度高的因子;而是用資料搜尋方式，其小優點可能是因為因子經過查詢，代表在別的資料集有呈現這樣的狀況，那可能在我們的資料集如果也看到這樣訓練出來的模型，可能更有機會使用在不同群體間，但這樣的方式可能會讓我們遺失掉其實重要的因子，或者說無法新發現重要的因子。
5. 在最後算出的公式中，我們年齡的加權係數為1.5，權重較高，導致算出的結果只要年齡高，就會有比較高的分數呈現，這樣的結果可能會導致稀釋了其他因子的重要性。所以是否需要做matching的動作，切成不同年齡層做分析？

如果使用年齡做分層後就無法使用年齡當做這個方法中我們其中一個因子，但是確實年齡是會間接影響其他因子的像是BMI以及血壓。

在此提出另一的想法：將模型後續設計成兩階段，一開始先試算年齡的一個分數，後續再填寫問卷套入原本的公式中（已做年齡分層處理）結算出一個總分數，但是因為還沒有時間實際操作，或許這樣還有其他問題需要解決。

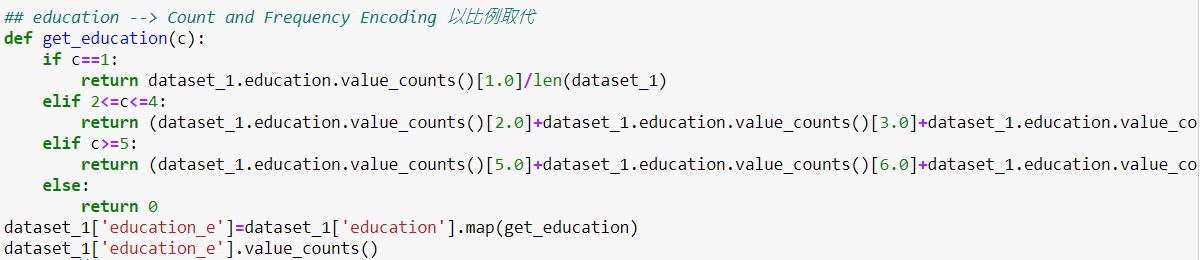
1. 檢討補缺失值的方式：我們在補連續型資料時，是使用最簡易補平均值的方式，他的缺點是會造成變異量的低估，減弱變異數之間的相關性並扭曲觀察資料群內的數值分布 ，可能還可以使用其他更好的補值方式，例如：回歸補插 ，其是利用這個公式補缺失值，是利用有遺漏資料之變數與資料檔內其他變數的關係為基礎，預測遺漏資料的數值。
2. 黃千瑜
3. 訓練簡單的中風預測模型
   1. 先利用少許的資料來訓練模型，熟悉程式碼的操作，以方便真正的大資料進來時可較上手。
   2. 先將資料進行前處理，而挑選一些可能的中風因子來訓練模型
   3. 中風因子分別有年齡、血壓坐姿右手舒張、收縮壓，是否長期服用藥物高血脂藥物、曾患糖尿病、曾患心臟血管疾病、有無中風
   4. 也是分別用5種方式來預測模型，最後結果較好的模型為Logistic Regression, Decision Tree。



1. 問題討論
2. 在補值的階段，因為資料有很多欄位都無資料，導致我們只能採用補中間值，不另一種類別來補缺值，不過平均真的能代表真實數據嗎?又或者類別變項的補值是利用將空值補第三個分類，雖然說這是最迅速的方法，如果能找到更好的補值方法，或許類別的補值會更好。
3. 在建立模型時，一開始沒有設定random state，導致每次結果模型都會有所不同，了解狀況後，有設立random state，使模型不在變動。
4. 另外，我們在挑選要採用哪些類別的資料時，僅參考一些文獻來決定中風因子，可能有一些偏頗，不過因為最終目的是要利用 app 來收集資料，導致有一些生化數值，不可來再填問卷當下就知道的資訊給先剔除了，也有可能剔除的欄位中有較重要的，不過礙於最終目的，只好在這部分妥協。
5. 關於閾值部分，可能也可因最後的目標而有所改變像這次的目標就是要讓病患的醫療成本降低，所以才利用這種方式。
6. 溫建泓
7. 分工：模型原理查詢、模型建立建議、報告
8. 問題討論：如同前面在介紹 Logistic regression 的時候提及到的，可能在開始回歸時選擇不同的起點試著回歸；亦或者在資料前處理的部份對數值做取平方、開 log 、做 up/down sampling 等等。
9. 柯雅云
10. **類別資料處理可以依照問卷內容有不同處理方式**
11. 本次美兆資料集扣除生化血液欄位，大多為問卷式的選項，先前的類別資料處理方式，是將欄位盡量變成二元化(binary)的方式，使得有些欄位有多重選項無法將這些差異放入模型中，若使用dummy variable的方式，選項類別過多會有欄位暴增的問題
12. 可以使用Count and Frequency Encoding 的方法，將選項用比例取代，不僅可以得到落在0-1之間的數字又可以將選項差異考慮進去，以下以education欄位作範例，依照codebook裡：

|  |  |  |
| --- | --- | --- |
| education | 教育程度 | (1)不識字 (2)國小 (3)國中 (4)高中(職) (5)專科 (6)大學 (7)研究所以上 |

將此欄位重新分類成不識字、國小-高中、專科以上三群，並以分布人數比例去做計****算最後取代為選項如下右圖



1. **特徵挑選不夠嚴謹，以下提出方法：**

Feature selection 有眾多的方法，以下使用Filter Methods、Embedded methods進行feature selection

* **Basic Filter Methods：**
  + 常數特徵(constant feature) 移除
  + 半常數特徵(quasi-Constant Features)移除

將特徵在各資料間變異度小於一定比例的欄位去除，表示此欄位對模型並無太大貢獻(此欄位幾乎都是一樣的值)，使用先前提及的篩選條件最後資料集剩下(5720, 109)，篩出變異量低於0.015的欄位共17項，最後dataset為(5720, 92)



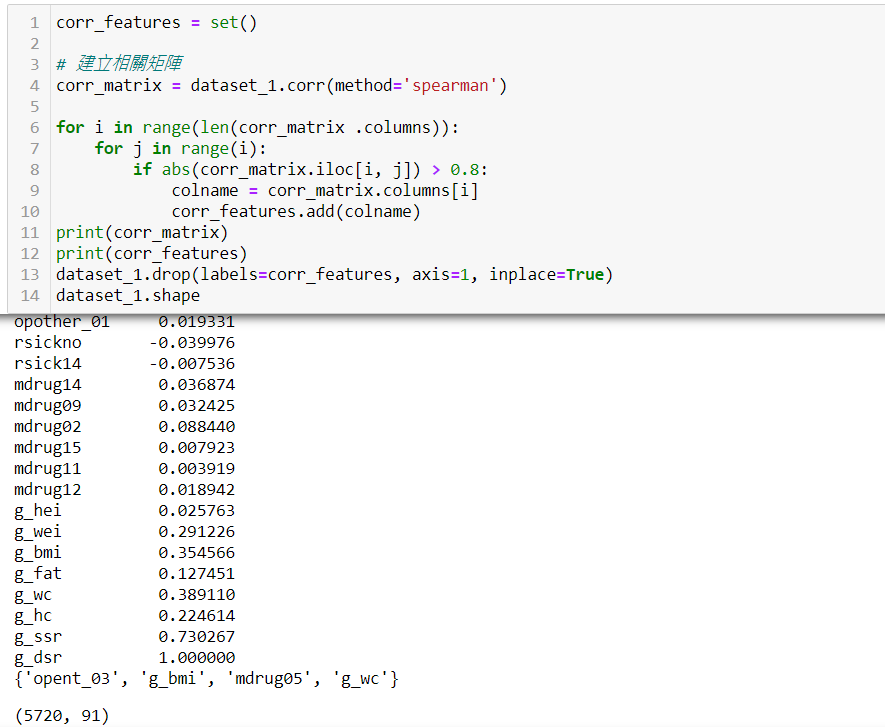
### Correlation Filter Methods

### Correlation(相關)是衡量兩個變數之間的線性關係，也就是一個變數依賴另一個變數的程度。假如兩個特徵高度相關，對於我們要預測的標的(target)來說，他們提供了重複的資訊，因為我們只取其中一個特徵就可以做出正確的預測了。所以移除其中一個特徵，不只能降低資料及維度(dimensionality)還可排除雜訊(noise)。

### 有以下幾種方式：

* + Pearson correlation coefficient(皮爾森相關係數)
  + Spearman's rank correlation coefficient(斯皮爾曼等級相關係數)
  + Kendall's rank correlation coefficient(肯德爾等級相關係數)

這裡使用Spearman's rank correlation coefficient，它是屬於非參數檢定(nonparametric test)，它用單調函數(monotonic function)來衡量兩個變數相關程度，它的值介於 1 和 －1 之間。當兩個變數完全單調相關時，斯皮爾曼等級相關係數則為1 或 －1。它類似皮爾森相關係數，但是皮爾森相關係數衡量線性關係，而斯皮爾曼等級相關係數衡量單調關係(不論線性與否)。斯皮爾曼等級相關係數適用適用於連續型(continuous)或不連續型(discrete)有序變數。它也不對變數分布做假設。將相關係數>0.8得欄位去除，得dataset(5270, 88)

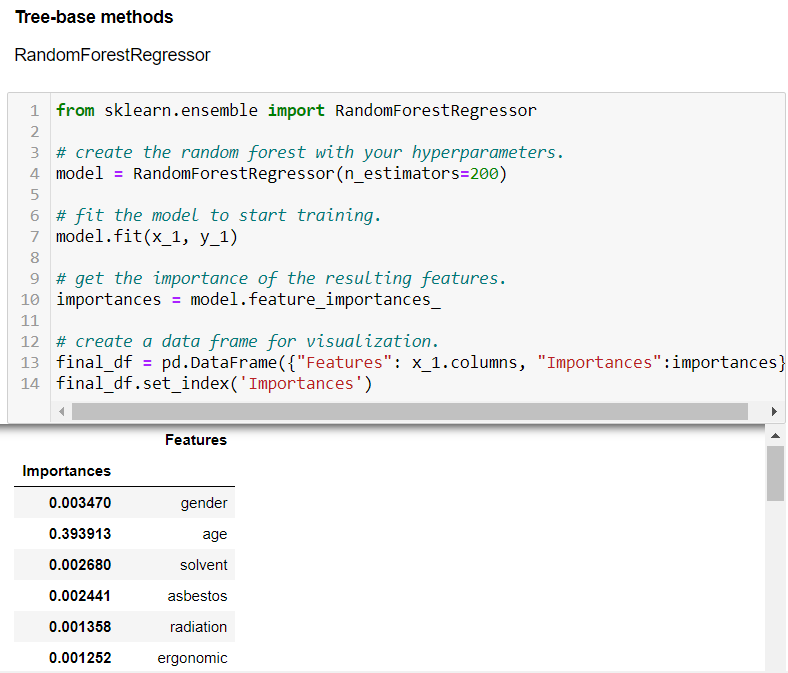
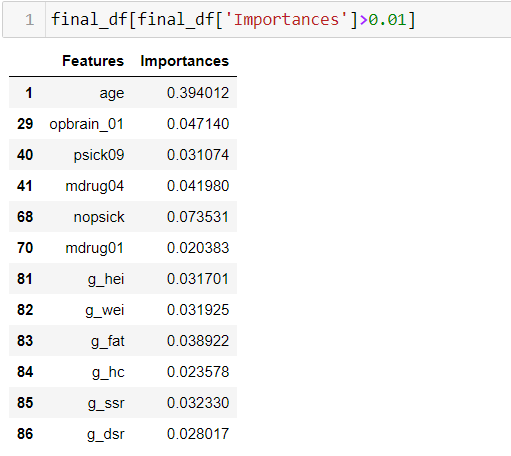


### Embedded methods(嵌入法)[¶](http://localhost:8888/notebooks/MJ%20data%20-final%20(featurce%20selection)%20.ipynb#Embedded-methods(%E5%B5%8C%E5%85%A5%E6%B3%95))

1. Regularization methods
   * + 1. Lasso(Least Absolute Shrinkage and Selection Operator) regression使用L1 regularization 讓不重要的係數值變成0，L1正則化像非正則化線性模型一樣也是不穩定的，如果特徵集合中具有相關聯的特徵，當資料發生細微變化時也有可能導致很大的模型差異。
       2. Ridge regression使用L2 regularization 讓係數值分散的更平均 並且不會讓係數縮小成0
2. Tree-base methods

Tree-base演算法和模型也提供特徵重要性(feature importance)來讓我們做特徵選擇，我們可以使用任何樹狀基礎的學習模型，gradient boosting algorithms(如：XGBoost，CatBoost等)是較好的模型，因為它提供正確的特徵重要性。

用Randonforrest模型方式，將最後的dataset(5720, 88)放入，可得到各欄位的權重，設篩選條件>0.01的欄位可得出如下圖右的12個特徵，與先前挑選的特徵有差異。



1. **在本模型中，年齡權重過高**：

雖然年齡確實是中風重要的危險因子，但在此模型中，年齡幾乎直接決定了logit p的數值，造成年齡越大中風機率會大大上升，減弱了其他因子的效應，況且年齡為一不可改變的的危險因子(nonmodifiable risk factor)，我們推測此問題與年齡資料正規化有關和資料集的年齡差距太大(6-95歲)，或許應該先以年齡分層方式分別做出不同的模型，還有需要更了解套件中正規化的方式，將年齡權重下降至一合理的範圍。