109 Data Mining Project 02

I. 目的:

Project 01 的主題是「資料前處理」，Project 02 則是「模型建置」。

Project 02 的分類器只能選擇 Decision Tree 或是 Logistic Regression，請擇一實做。

請勿抄襲網路上的 code, 且禁止使用套件。

II. 程式語言:

只可以使用 Python, C, C++或 Java.

III. 資料集說明:

Project 02 將會接續 Project 01 的主題使用相同的資料型態，不同的數據集。

「Info」: 一個 row 代表一位患者。

共病症_Comorbidities:一個數字代表擁有一種病症，標示「0」則表示沒有任何共病症。

若患者此欄位內容為(1,2,4)，則代表患者同時有病症 1, 病症 2 以及病症 4。

「TPR」:「溫度、脈搏、呼吸速率、收縮壓、舒張壓」的時間序列資料，同樣的「No」代表是同一位患者的資料。

IV. 實驗流程提示:

有鑑於部分同學在 Project 01 並沒有達到自己預想的成績，本次 Project 將提供相對適當的實驗流程，確保同學的 Project 02 不會受到 Project 01 太多的影響。

當然，如果你對於 Project 01 的資料前處理、驗證方法很有自信，也相當歡迎繼續沿用。

但如果你的自我驗證與 Testing data 結果相差甚遠，請務必參考以下提示。

1. 合併 TPR 與 Info 的表單:

最終預測是以 No（患者編號）為單位，也就是跟 Info 的 size 相同，因此建議將 TPR 以患者為單位，並透過統計方法(採樣或是統計值)轉換成固定格式，解決患者測量天數不同的問題，再與 Info 依照「No」合併，讓每位患者僅有一筆資料。

2. 資料轉換:

TPR 的重要性遠遠大於 Info，建議花較多的心力在 TPR 的處理。

Info 大部分的 features 可以使用 one-hot encoding，再去除出現頻率太低的欄位。

3. 模型驗證:

好的驗證方法應該能大略估計出真實的測試分數，並讓你藉此調整 feature 以及 model。

首先，將資料整理成以患者為單位的格式，確保每一筆資料是幾乎獨立。

接著，先 shuffle，再進行 K-fold CV 進行驗證，而過大的 K 值會讓驗證集太小，過小的 K 則會讓訓練資料不足，對於本次的資料集建議 K＝5。

在此特別提醒，Project 01 中有些同學對整個 TPR sheet 做 shuffle，這個動作會讓同一位患者不

同天的測量值分散在 Training data 與 Validation data 中，進而因資料洩漏導致 overestimate，如果你 Project 01 自我評量 F1-Score > 0.8，Testing 結果卻很低，很有可能是犯了這項錯誤。

最後，由於這個資料集略為不平衡(目標比例大約 1:3)，因此只看 Accuracy 意義不大，而且題目要求使用 F1-Score 作為評分依據，因此在驗證時，至少要將 F1-Score 作為模型評估的參考值。對於本次資料集，正常的 F1-Score 應該介於 0.4~0.7 之間，如果你的驗證分數超過 0.8，很大的機率是有哪個步驟做錯了!

V. 模型實作：

請自由選擇實作 Decision Tree 或是 Logistic Regression，兩者擇一即可，若兩個都做，只會取分數高者作為此次 Project 02 的成績。

可以使用基本加減乘除的函式庫，但禁止使用套件，請務必自行實作，助教會檢查程式碼。

*以下括號內的分數皆為學期總分，7% = 學期總分的 7 分。

- Decision Tree（7%）：

基本指定功能：

1. fit (1%): 利用訓練資料生成一棵樹，不限制 impurity 使用 gini, entropy 或其它。此外，至少要有一個超參數(max_depth)去限制樹的深度。

   輸入: 訓練資料、超參數　**輸出:** 模型的參數

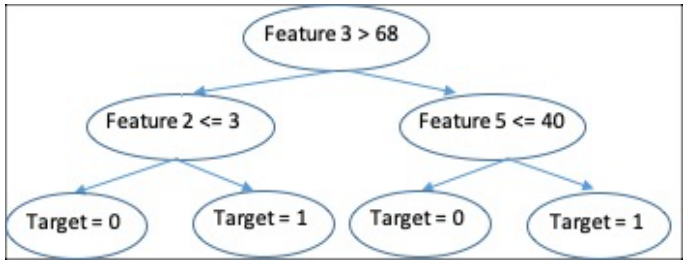
2. Predict (2%): 基本分類功能，將測試資料的 label 預測出來。

   輸入: 模型、測試資料　**輸出:** 預測的類別，只能是 1 或 0。

3. Pred_Probability (2%)：進階版的分類功能，預測資料屬於各個 class 的機率。

   輸入: 模型、測試資料　**輸出:** 預測患者是(Target=1)的機率。

4. Visualize (2%)：把你的樹印出來，須包含「該節點的 Feature 名稱、切割點、leafs」，如下圖所示，如何畫出圓形與箭頭不是 Project 重點，可先用程式產生每個節點的內容，再用 PPT, word 等輔佐工具完成繪製，並將結果附在 Report 中。



額外加分功能：

1. Pruning (1%): 如果使用 Pre-pruning，需有超參數可指定「條件」，並在 Report 中註記該超參數名稱以及剪枝前後視覺化的比對圖。如果使用 Post-pruning，請在 Report 中提供剪枝前後比對圖。
2. 處理缺失值 (1%)：請勿直接補平均或眾數，需使用 DT 的方式處理缺失值。
3. Random Forest (1%): 需有超參數指定「生成幾棵樹」並在 Report 中註記該超參數名稱。

- Logistic Regression (7%):

基本指定功能:

1. fit (1%): 利用訓練資料生模型。

   輸入: 訓練資料　**輸出:** 模型的參數

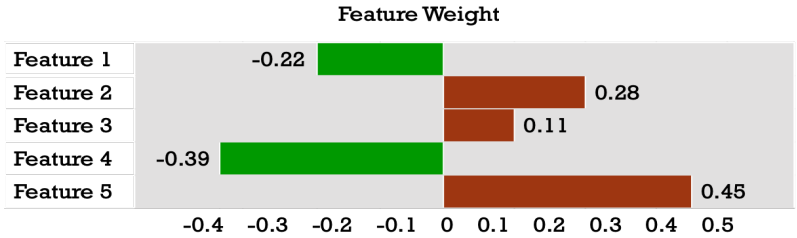
2. Predict (2%): 基本分類功能，將測試資料的 label 預測出來

   輸入: 模型、測試資料　**輸出:** 預測的類別，只能是 1 或 0。

3. Pred_Probability (2%): 進階版的分類功能，預測資料屬於各個 class 的機率

   輸入: 模型、測試資料　**輸出:** 預測患者是(Target=1)的機率。

4. Visualize (2%): 把模型印出來，Weight 相當於對 LR 找出的方程式係數取 Exponential，Feature 權重的正負值表示該 Feature 的變化會使結果趨向 P(y=1)或 P(y=0)。圖片要標明係數與 Feature 名稱，如下圖範例，但可以不上色。



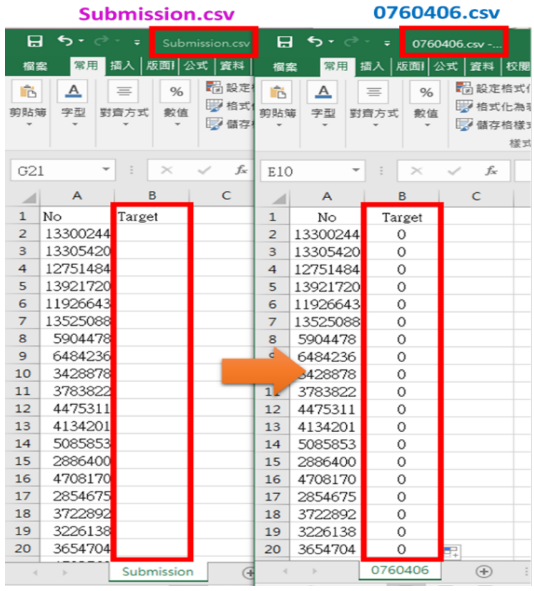
加分功能:

1. Regularization (1%): 在 cost function 上加上懲罰項，Report 要標註前後視覺化的比對圖。
2. Ensemble (1%): 需有超參數可指定「生成幾個模型」，並在 Report 中註記該超參數名稱。

VI. 繳交檔案名稱與格式:

1. <學號>.csv

   - 此份檔案是用來存放 Test 檔預測的結果
   - 「Submission.csv」於上傳前，需重新命名為「學號.csv」EX: 0760406.csv
   - 「Target」欄位只能填入 0 或是 1，沒有其他選項
   - 整份 CSV 檔只能有這兩欄，不能有其他多餘的欄位



2. <學號>_Report.pdf

- 檔名: <學號>_Report.pdf　EX: 0760406_Report.pdf
- 字體大小: 12　- 字型: 英文(Times New Roman), 中文(標楷體)
- 內文: 第一部分先寫「ReadMe」，說明程式執行方式以及參數配置。
  至少要有「資料前處理」、「模型建製」與「驗證方法」，其餘不限，若有必要可以自行增加。

3. <學號>.****

- 這個是你的程式碼。請注意，不可以是 exe，要是可以看到程式碼的檔案格式。助教執行後，需要可以看到，至少一個跟 report 中一樣的驗證結果。

VII. 上傳內容:

- 請上傳 zip 至 e3 作業繳交區

VIII. 扣分項目:

扣分項目	Project 總分
檔案名稱錯誤	-20%
答案繳交格式錯誤	-20%
遲交 1 天	-20%
使用 LR 或 DT 或加分項以外的分類器	-100%
分類器 Call 套件 或抄襲網上他人撰寫的程式	-100%

*請注意，你可以 google 到的，我們也可以。

Item	學期總分
F1-score (base on Target=1)	7%
Model	7%
Project Report	7%

$$*F1-score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} = \frac{2TP}{(2TP+FP+FN)}$$

$$TP: True\ Positive \quad TN: True\ Negative \quad FP: False\ Positive$$

IX. 截止日期:

- 2021/01/08(五) 零晨 12:00 整，助教上傳 Test 檔
- 2021/01/10(日) 晚上 11:50 作業上傳截止