66

(0)line群裡的evernote連結

<https://www.evernote.com/shard/s379/sh/5b5a1fdb-f5e4-40ee-bd8e-666c3f6ae3de/PDiHx0j2eqGqNBpwTjX7x1dLILlcgtS34A3K1Z1PvdQKRZDvIxy9KmNkTQ>

(1)google ppt

<https://docs.google.com/presentation/d/1spk0yaP0GFeoothcGUzVKUge3KSPXeAvDKnCPbTLzgg/edit?usp=sharing>

(2)google meet

如要加入這場視訊會議，請按一下這個連結：https://meet.google.com/gvf-qahv-ikd

你也可以透過電話加入通話，只要撥打 +1 385-645-8135，然後輸入以下 PIN 碼即可：618 812 916#

如要查看其他電話號碼，請按一下這個連結：https://tel.meet/gvf-qahv-ikd?hs=5

(3) 簡報正式區

[我們的報告.ppt](https://docs.google.com/presentation/d/1IMDAgaJOZQNvSJzoZPM6oZ8ZEV54DA-i/edit?usp=share_link&ouid=101209653342697512251&rtpof=true&sd=trueD3UCYrtsHnDT7d9Ylz0kqbjAxf/edit?usp=share_link&ouid=101209653342697512251&rtpof=true&sd=trueothcGUzVKUge3KSPXeAvDKnCPbTLzgg/edit?usp=sharing)

| 謝弘軒 | 資科專一 | 111971022 |
| --- | --- | --- |
| 郭書瑋 | 資科專一 | 111971023 |
| 周正晏 | 資科專一 | 111971003 |
| 胡元亨 | 資科專一 | 111971024 |
| 施宗佑 | 資科專一 | 111971005 |
| 楊昇豐 | 資科專一 | 111971013 |

目標：下週四5/18，能夠讓B組的接手做視覺。

google ppt ，[分享連結](https://docs.google.com/presentation/d/1spk0yaP0GFeoothcGUzVKUge3KSPXeAvDKnCPbTLzgg/edit?usp=sharing)出來，有結果就貼上去。

5/18

5/25 ←– 下課討論

6/1 ←– 下課討論 (30分鐘以內)

6/8 —>~~version1~~

6/12(一) version1

6/14(三) version2

6/15 —> version3

**6/12**

昇豐：

一、test set的沒病、有病的比例？我自己試仍然約是10:1

二、在google doc 6/8的【2. 簡報已完成，可以放到正式區以黑體字標示：】，能看到今日完成的項目，有以下：

(3) 各模型的執行情況、(4-2) 介紹此方案、(4-3) 再次重申評估與結論

、(5) 專案艱辛之處、(6) demo、(7) 專案引用的套件與外部資料、(8) 其他附件。

今晚不用急著美化統整，我在草稿區講可以。  
—-------------------------------------  
三、簡報正式區(Austin整理)第9頁，我理解此頁要表示綜合VIF，你挑選出4個有效性高的欄位。但這並不是我們討論要在介紹資料集得出的結論

| 資料集介紹 | 特徵選擇後 |
| --- | --- |
| 依照統計分析、直覺看來，得出哪幾個欄位**可能跟**心臟病有關。  **9個欄位** | 挑選出13個欄位**確實跟**心臟病有關。 |
| 在(4-3) 再次重申評估與結論，比較 | |

四、簡報第51頁的表格幫忙加至正式區。

五、正式區請為我([111971013@g.nccu.edu.tw](mailto:111971013@g.nccu.edu.tw))開立編輯權限。

[Austin] Done, 6/12.

**6/8**

書瑋：

資料分析圖示

Rscript data\_plot.R

主程式

Rscript main.R --d 2 --m 1

參數

--data\_source or --d

1:讀取原資料，會處理dummy和smote

2:讀取已處理資料

--model or --m

1:decision tree (不支援data\_source = 2)

2:random forest

3:xgboost

昇豐：

- p10 沒有表達完全

- 綱要順序：(1)about dataset、(2)資料前處理、(3)各模型的執行情況、(4)介紹最終建議方案、(5)專案艱辛之處、(6)demo、

(7)專案引用的套件與外部資料(不會講)、其他附件

- about dataset的小結(覺得那些欄位是有關的)

1. 簡報尚未完成的部分：

(1) 專案艱辛之處：**各人至少生一則**。我的想法是統合1~2頁放在一起講。

都寫了

(2) 專案引用的套件與外部資料。我的想法是統合放在，但*毋須講*。

我、busky、austin；  
 元亨、書瑋會補

(3) 介紹最終建議方案：  
 我不確定 元亨/書瑋、正晏/Busky簡報綱要與草稿 是否已經完成；今晚下課後討論看

這週日晚上7~9點線上見。

*虛擬碼、執行流程圖*

(4) demo，今晚下課後看看執行效果。

2. 簡報**已完成**，可以放到正式區以黑體字標示：

(以下6大點就是正式報告的順序要點)

(1) 資料集介紹。Austin已放在正式區。

(2) 資料前處理，第50~52頁，示意圖

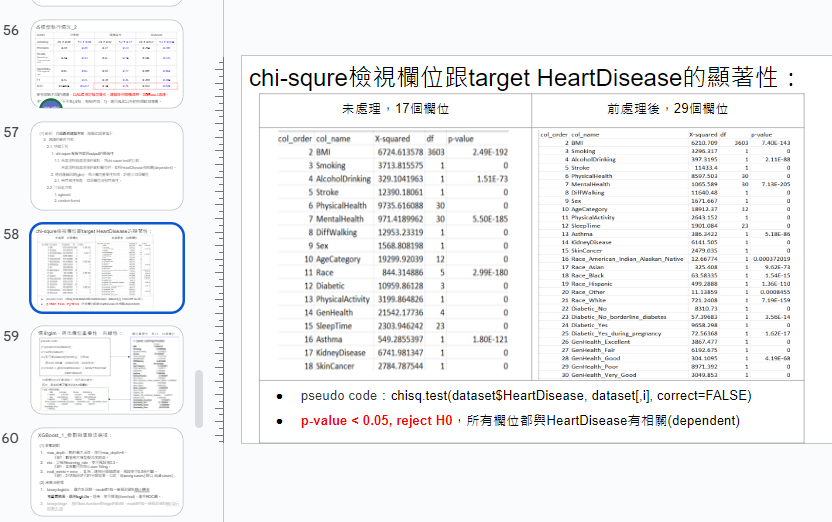


(3) 各模型的執行情況，第54、55頁，示意圖\_\_\_20230612完成



(4) 介紹最終建議方案：

(4-1) 特徵工程，第58~59頁，示意圖



(4-2) 介紹此方案，第60~68頁，示意圖\_\_\_20230612完成  
 (以random forest、xgboost為主；若篇幅不夠、才加入決策樹)



(4-3) 再次重申評估與結論

第69頁，配合Austin正式區的第6~9頁，示意圖



(5) 專案艱辛之處\_\_\_20230612完成

第73頁，整合昇豐、書瑋、宗佑、正晏、元亨，示意圖；請Austin最想要講的內容加在空白處可。



以上來自

昇豐：第71頁的第1點

元亨：第11頁的第2點

書瑋：第10頁的第1點

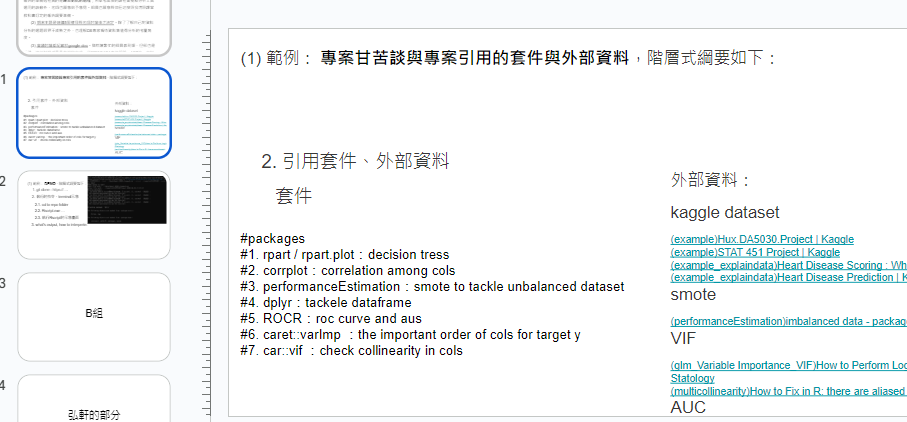
正晏：第24頁的第2點

宗佑：第47頁的第1點

(6) demo，第70頁\_\_\_20230612完成



(7) 專案引用的套件與外部資料(不會講)

昇豐：第72頁的第2點，示意圖  
 

元亨：第11頁的引用資料

書瑋：第10頁的引用資料

正晏：暫無

宗佑：暫無

(8) 其他附件(不會講)

1. 將大家所有的專案艱辛之處放此：

昇豐：第71頁

元亨：第11頁

書瑋：第10頁

正晏：第24頁

宗佑：第47頁

2.

**6/5**

昇豐：

1. 6/12 21:30 線上簡報試講

2. 簡報尚缺，請各自補齊  
 建議的最終方案

(4) 專案艱辛之處(各自都要寫)

(5) 專案引用的套件與外部資料

[Ausitn]

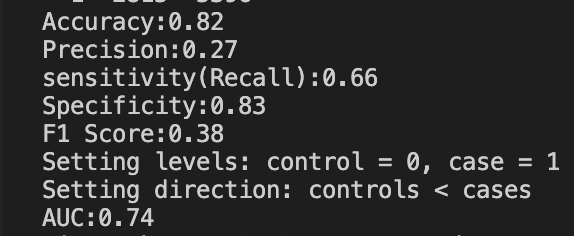
library(dplyr) -> bar chart plot

library(ggplot2) -> box chart plot

元亨：

Roc, importance用圖

其餘用數值(csv)



Austin：

1. 請大家去readme分支看，還需要增減再說；也能告知簡報哪頁可以ˊ直接用。  
     
   昇豐：我認為分兩部分，第一部分說明如何執行專案，產生什麼用途的檔案。第二部分按照6/8的「6大點就是正式報告的順序要點」，依序貼圖就好。

**6/1**

昇豐：

*0. 我們沒有計算model的R-square(correlation^2)、RMSE。*

決策樹沒法計算*R-square*

<https://stats.stackexchange.com/questions/171762/explanatory-power-of-a-decision-tree>

<https://stackoverflow.com/questions/40901445/function-to-calculate-r2-r-squared-in-r>

1. 能確定A組夥伴會提供那些簡報(請補充階層式綱要)：

資料集介紹會有小結論，分析也會有小結論，最後一起參照。

週一根據欄位重要性的集合，Austin就可以更改資料集介紹的投影片。

(1)前處理：昇豐

(2)建模

特徵工程

A-2. PCA：昇豐(prcomp)

A-3. cols correlation：昇豐、書瑋  
 沒有辦法效率挑出欄位，因此不放col correlation。

A-4. chi-squre看看特徵跟output的顯著性 昇豐

A-1. 欄位重要性：昇豐(glm)、書瑋(random)、正晏+busky(xgboost)

建立  
 B-1. f-fold 昇豐

B-2. train/test 元亨、書瑋、正晏、busky  
 B-3. null model busky

(3) 評估  
 C-1. confusion matrix相關的計算(precision、recall、specificity、F1、sensitivity)  
 C-2. 得出結論

(4) 專案艱辛之處(各自都要寫)

(5) 專案引用的套件與外部資料

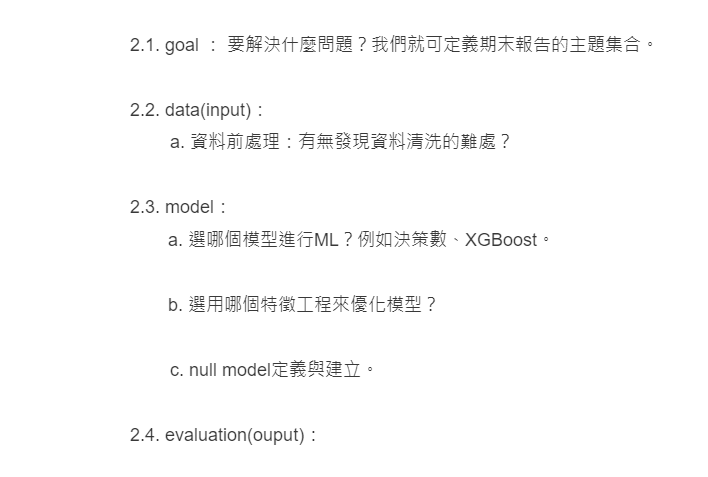
(6) demo的簡報(請補充階層式綱要)

生豐

2. 6/5 專題研討下課後討論

1. 確認**簡報草稿**哪一部分可以給Austin處理。

2. 討論Readme要寫啥？我建議就照此份google doc最末的要點按順序撰寫。需要切專屬的分支

Austin、正晏  


**5/29**

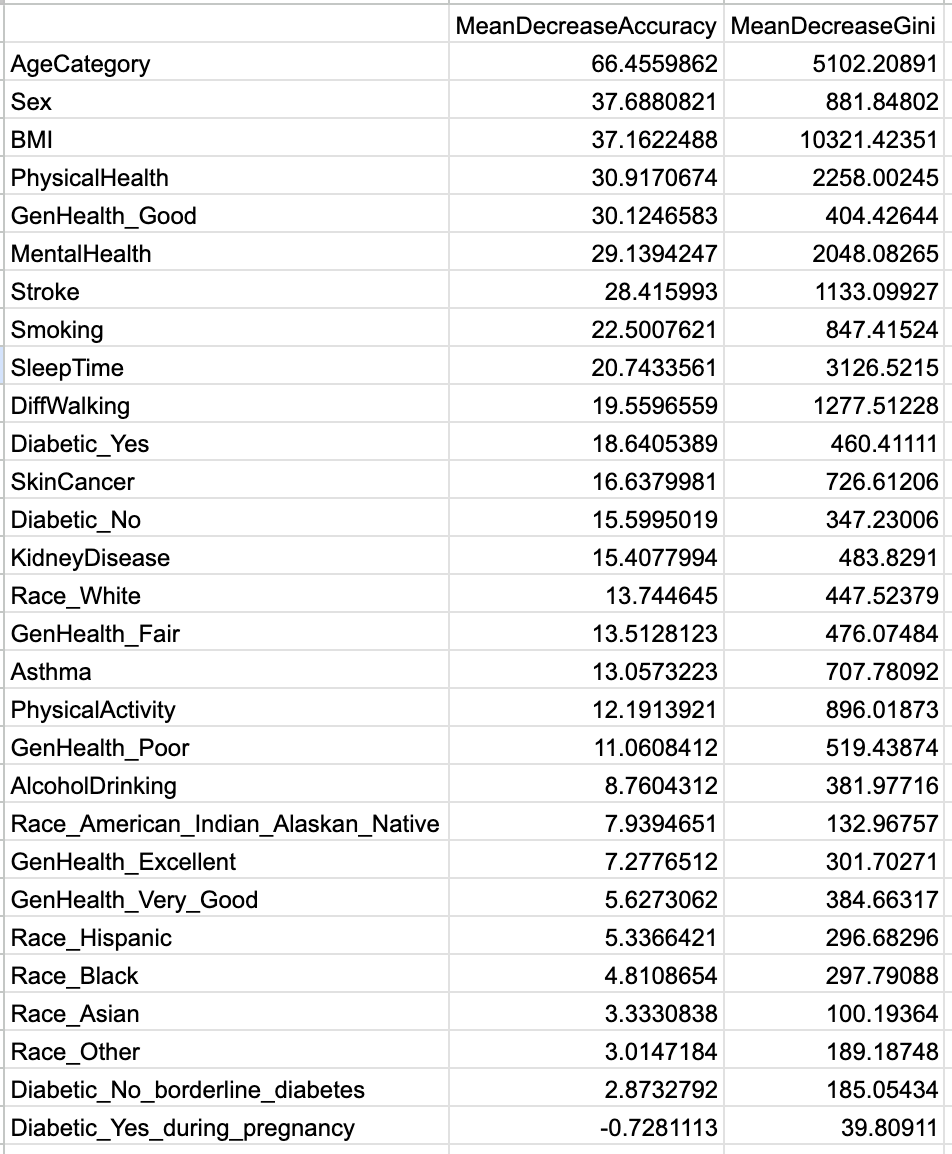
書瑋：

程式執行指令：Rscript main.R

fold 跑 mtry = 26，確認是否overfitting

important 特徵數值

* MeanDecreaseAccuracy：Accuracy的差異，越大表示該參數越重要
* MeanDecreaseGini：Gini的差異，越大表示原參數資料較純，表示參數越重要



AgeCategory+Sex+BMI+PhysicalHealth+GenHealth\_Good+MentalHealth+Stroke+Smoking+SleepTime+DiffWalking+SkinCancer+KidneyDisease

昇豐：

1. smote為何會產生NA？

* 參數k的分群  
  <https://rdrr.io/cran/performanceEstimation/man/smote.html>  
  <https://stackoverflow.com/questions/62871492/create-balanced-dataset-11-using-smote-without-modifying-the-observations-of-th>  
  <https://stackoverflow.com/questions/73574827/smote-r-outputting-rows-of-nas>
* 我選用smote\_performanceEstimation 的根據https://stackoverflow.com/questions/67085791/package-to-do-smote-in-r

2. 等待random forest、xgboost的欄位重要性數值，看看跟邏輯回歸跑出來的差多少？

2.1. 這週四問家銘不同模型跑出的欄位重要性，孰優孰劣？ ←– teacher說，把欄位組合拿去跑model最準。沒有標準答案。

能否用A模型產出的欄位重要性，運用到B模型<---teacher說可以

3. 書瑋提供的繪圖：<https://r-graph-gallery.com/index.html>

4. 確認元亨的訓練集smote後，比例是否真為1:1.33。(應該要是1:4。) 餵入訓練集的資料量造成差異。21萬筆的確如此

元亨:

1. 程式執行指令：Rscript Jude\_111971024.R --input ../data/heart\_2020\_cleaned.csv
2. pre-processing模組化是否和書瑋重工

宗佑:

1.程式執行請看Readme第一行，下面的使用現成smote後資料，速度較快。

Rscript code/xgboost\_111971005.R --input data/heart\_2020\_cleaned.csv

run smote with raw data "heart\_2020\_cleaned.csv"

Rscript code/xgboost\_111971005.R --input data/heart\_2020\_smote.csv

use already smote data "heart\_2020\_smote.csv"

**5/25**

今晚討論

昇豐：

1. 5/29 晚上9:30線上討論，可以？

1. 看各模型tune的結果(precision recall f1 auc roc)、圖片。

2. 看看欄位共線性、解釋density graph的優化成果。

還沒研究：解釋density graph

3. 各自程式推到gitrepo；試試看Rscript.exe能不能過。

3.1. 昇豐的branch：111971013\_Carter

(1)cd to path : \final-project-group3\code

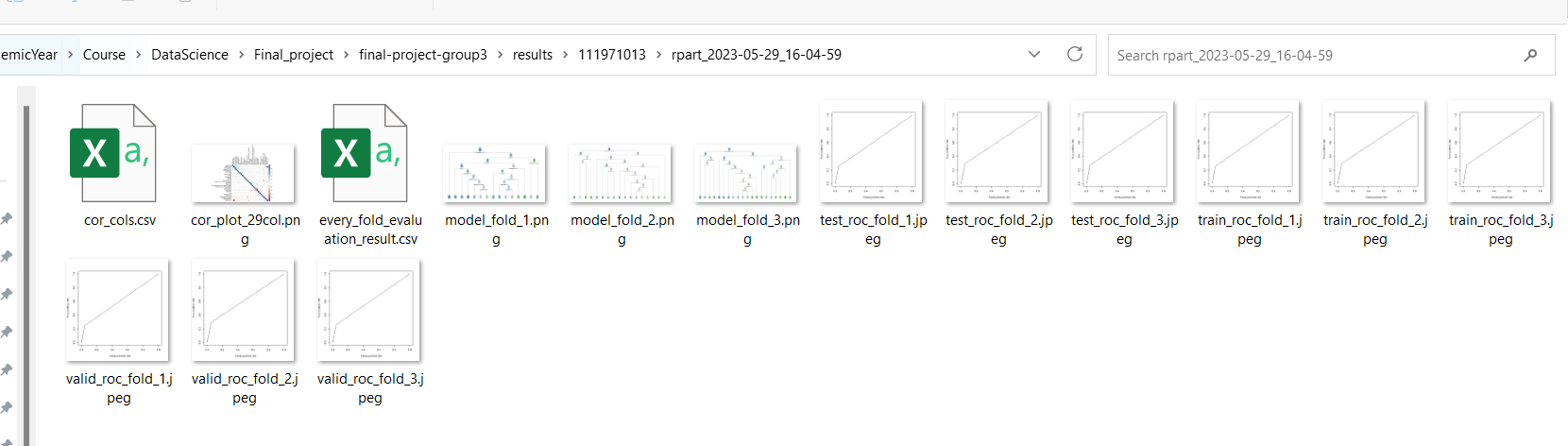
Windows執行方式

(2-1) & 'C:\Program Files\R\R-4.3.0\bin\Rscript.exe' Carter\_111971013.R --input ../data/heart\_2020\_cleaned.csv --output ../results/111971013 --model\_name rpart

Mac 、 Linux執行方式

(2-2) Rscript Carter\_111971013.R --input ../data/heart\_2020\_cleaned.csv --output ../results/111971013 --model\_name rpart

(3) results/111971013 產生 「model\_name\_時間戳記」的目錄



2. GitRepo目錄結構、分支名稱。

1. 目錄：

code/EnglishName\_studentId.R

results/ 111971013/csv、json、xlsn、圖片

data/原始資料

docs/ 會議討論【最後上傳】

2. 請用 學號+英文姓名開分支，例如111971013\_Carter

3. 簡報

2.1. 原始資料的視覺

2.2. 昇豐處理

2.3. 各自用模型的處理

2.4. 各自用模型的處理

2.5. 各自用模型的產圖出來，匯集起來看夠不夠

**4. 請大家用我的smote(套件performanceEstimation)。此份google doc有範例程式，Ctrl +f可找到。**

5. 我會再作欄位共線性、解釋density graph。  
 5.1. 透過glm()邏輯回歸能計算共線性或者欄位重要性：

sample code

set.seed(55)

f\_in\_nums <- nrow(f\_in\_csv)

fold\_interval <- round( f\_in\_nums / k\_fold\_num)

f\_in\_csv <- f\_in\_csv[sample(1:f\_in\_nums), ]

smote\_f\_in\_csv <- smote(HeartDisease ~ ., data =f\_in\_csv[c(-(100573:201144), -(201145:301716)),])

smote\_f\_in\_csv <- smote\_f\_in\_csv[!is.na(smote\_f\_in\_csv$HeartDisease),]

smote\_f\_in\_csv$HeartDisease<- ifelse(smote\_f\_in\_csv$HeartDisease=='Yes',1,0)

table(smote\_f\_in\_csv$HeartDisease)

base\_cols <- paste(colnames(f\_in\_csv)[-1], collapse="+")

model <- glm(paste("HeartDisease", base\_cols, sep="~"), family="binomial", data=smote\_f\_in\_csv)

A. 計算共線性：

> car::vif(model)

Error in vif.default(model) : there are aliased coefficients in the model

若用29欄位進行邏輯回歸，會因為multicollinearity的關係而錯誤；換句話說就是有的欄位彼此間相關性過高!

google search：R car::vif(model) Error in vif.default(model) : there are aliased coefficients in the model

<https://www.statology.org/r-aliased-coefficients-in-the-model/>

—------------------------------------------------------------------------------

採取13個欄位，就可以得到VIF。這些值都很小！13個欄位並沒有共線性

> car::vif(model)

BMI Smoking Stroke DiffWalking

1.108250 1.031146 1.021945 1.297409

Sex AgeCategory Asthma KidneyDisease

1.059703 1.104021 1.053452 1.036308

Race\_Hispanic GenHealth\_Excellent GenHealth\_Fair GenHealth\_Good

1.024754 1.229081 1.623939 1.585377

GenHealth\_Poor

1.406402

B. 欄位重要性()：

> caret::varImp(model)

Overall

**BMI 4.4294187**

**Smoking 13.0590904**

AlcoholDrinking 1.9100625

**Stroke 23.3059047**

PhysicalHealth 1.5127184

MentalHealth 2.1510181

**DiffWalking 6.5815495**

**Sex 25.4474306**

**AgeCategory 47.6255661**

PhysicalActivity 1.3458511

SleepTime 2.1629561

**Asthma 7.2590399**

**KidneyDisease 9.7363807**

SkinCancer 1.7886560

Race\_American\_Indian\_Alaskan\_Native 0.4889626

Race\_Asian 3.5278293

Race\_Black 3.8891932

**Race\_Hispanic 4.0550250**

Race\_Other 1.3882326

Diabetic\_No 1.3356281

Diabetic\_No\_borderline\_diabetes 1.0085721

Diabetic\_Yes 1.2487449

**GenHealth\_Excellent 7.0720475**

**GenHealth\_Fair 21.7952554**

**GenHealth\_Good 14.6508521**

**GenHealth\_Poor 20.6333198**

> >4=有13個欄位

base\_cols <- "BMI+Smoking+Stroke+DiffWalking+Sex+AgeCategory+Asthma+KidneyDisease+Race\_Hispanic+GenHealth\_Excellent+GenHealth\_Fair+GenHealth\_Good+GenHealth\_Poor"

<https://www.statology.org/logistic-regression-in-r/>

欄位重要性(RandomForest):

0 1 MeanDecreaseAccuracy

BMI 12.963202 29.949534 37.004319

Smoking 10.607566 17.773425 18.861257

AlcoholDrinking 5.811193 20.960525 20.037187

Stroke 18.836935 17.685078 20.910553

PhysicalHealth 10.843591 18.358841 18.118608

MentalHealth 7.589369 12.348783 15.724788

DiffWalking 9.000426 14.566796 15.664644

Sex 11.618129 23.187484 20.606749

AgeCategory 14.414068 31.190656 23.834968

PhysicalActivity 7.437506 11.081956 10.504036

SleepTime 10.567065 18.226266 14.619381

Asthma 5.924029 16.011459 13.044312

KidneyDisease 10.382129 16.648331 13.511642

SkinCancer 5.782972 9.182070 9.055548

Race\_American\_Indian\_Alaskan\_Native 2.796220 8.695581 8.008274

Race\_Asian 2.257975 9.966916 10.673360

Race\_Black 5.107070 12.991173 16.684118

Race\_Hispanic 1.053984 7.433698 9.833050

Race\_Other 3.996109 13.471029 15.059658

Race\_White 2.085537 14.525425 14.785692

Diabetic\_No 5.642429 8.530390 9.900404

Diabetic\_No\_borderline\_diabetes 9.002068 6.038766 11.811424

Diabetic\_Yes 5.471792 15.280870 11.537246

Diabetic\_Yes\_during\_pregnancy 3.494072 6.042743 7.076193

GenHealth\_Excellent 6.896207 7.159250 8.749628

GenHealth\_Fair 6.092068 9.868491 10.277100

GenHealth\_Good 5.306516 4.668623 8.729043

GenHealth\_Poor 5.548391 12.826259 11.103652

GenHealth\_Very\_Good 7.032531 4.814157 9.173067

MeanDecreaseGini

BMI 1846.32562

Smoking 784.15760

AlcoholDrinking 251.22305

Stroke 1721.21914

PhysicalHealth 2598.42676

MentalHealth 1299.69673

DiffWalking 1905.30743

Sex 1198.62882

AgeCategory 9029.17869

PhysicalActivity 548.96284

SleepTime 2534.90342

Asthma 355.36682

KidneyDisease 528.96042

SkinCancer 522.39528

Race\_American\_Indian\_Alaskan\_Native 96.04153

Race\_Asian 81.16871

Race\_Black 178.81046

Race\_Hispanic 195.88362

Race\_Other 124.15698

Race\_White 339.69815

Diabetic\_No 830.56240

Diabetic\_No\_borderline\_diabetes 100.95766

Diabetic\_Yes 1794.79180

Diabetic\_Yes\_during\_pregnancy 49.00148

GenHealth\_Excellent 1090.18637

GenHealth\_Fair 1007.33542

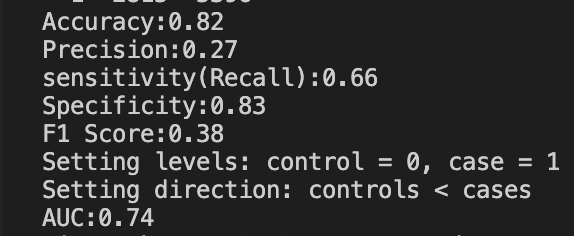
GenHealth\_Good 558.47994

GenHealth\_Poor 938.38240

GenHealth\_Very\_Good 864.57548

Roc, importance用圖

其餘用數值(csv)



5.2 PCA的decision tree圖，我用PC1~PC18；根據此作法去看看評估效果

xgboost正晏、busky；random forest元亨、書暐、busky；決策樹昇豐

正晏：

1. 昇豐建議：資料前處理方法用一致，最後產出30個欄位的資料。

2. 書暐：pred=0.5~0.8 (0.1)，用xgboost做看看效果，下周再來問。

Austin : 期末報告老師範例。

[信用卡高風險交易預測.pdf](https://www.dropbox.com/sh/3bqjakmpbw765fl/AAAPmXHG_EtDIBBmpZAFhV8ka?dl=0&preview=1101_group8_%E8%B3%87%E6%96%99%E7%A7%91%E5%AD%B8%E6%9C%9F%E6%9C%AB%E5%A0%B1%E5%91%8A(%E6%96%87%E6%9C%AC).pdf)

[信用卡流失客戶預測.ppt](https://www.dropbox.com/sh/3bqjakmpbw765fl/AAAPmXHG_EtDIBBmpZAFhV8ka?dl=0&preview=1101_group6_datascience_FP.pdf)

[我們的報告.ppt](https://docs.google.com/presentation/d/1IMDAgaJOZQNvSJzoZPM6oZ8ZEV54DA-i/edit?usp=share_link&ouid=101209653342697512251&rtpof=true&sd=trueD3UCYrtsHnDT7d9Ylz0kqbjAxf/edit?usp=share_link&ouid=101209653342697512251&rtpof=true&sd=trueothcGUzVKUge3KSPXeAvDKnCPbTLzgg/edit?usp=sharing)

**5/18**

今晚討論

1. github 期末專案組別名稱：group3



2.

1. 正晏：會嘗試xgboost(下週四上課給)；視覺化先做dataset。

2. 元亨：會嘗試隨機森林(這週四上課給)。

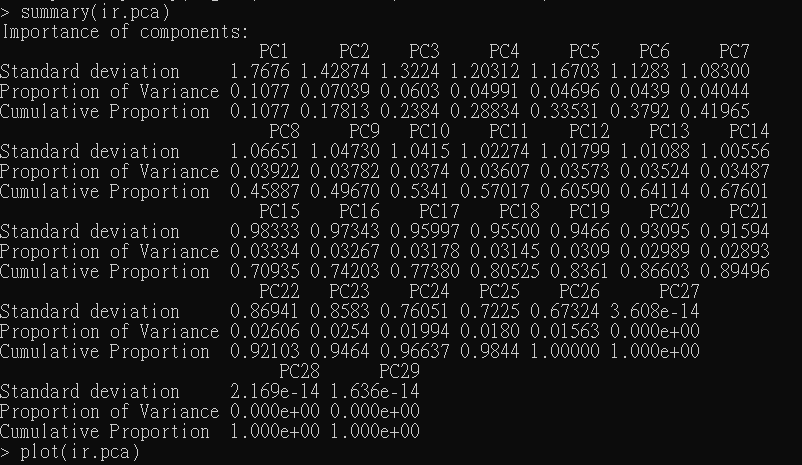
3. 昇豐：我會做特徵工程(PCA、ROC、AUC)、Hypothesis Test(chi-squre 完成的話，pass 給宗佑。)

進行中PCA：

>>>需要使用「log1p」取log，避免0成為infinite。  
 *>>>Teacher：AUC、ROC會比準確率重要；看看balanced 後的train model的 sensitivity、specificity表現怎麼樣。*

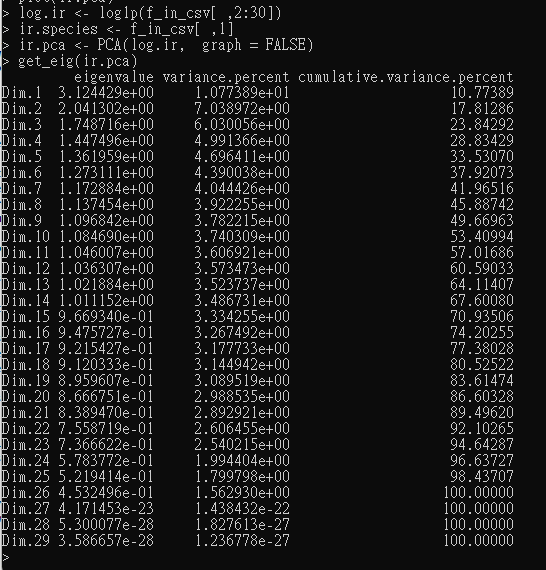
(1) 使用prcomp：

觀察到PC18時，可以解釋0.80525的資料；換句話說，取18個欄位可以

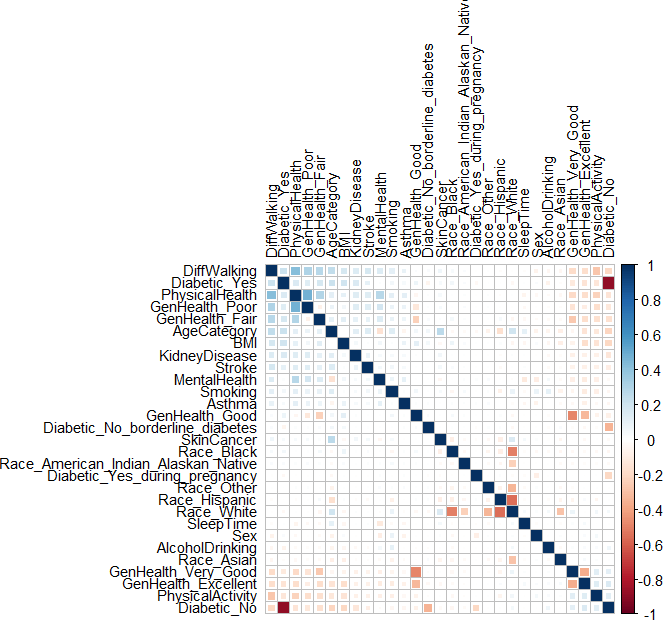


(2) 使用FactoMineR method

同樣得到Dim.18(=PC18)能解釋0.80525的資料。



(3)cor

  
取19個cor互相為正相關的欄位：  
> cor\_cols

[1] "BMI" "Smoking"

[3] "Stroke" "PhysicalHealth"

[5] "MentalHealth" "DiffWalking"

[7] "Sex" "Asthma"

[9] "KidneyDisease" "Race\_American\_Indian\_Alaskan\_Native"

[11] "Race\_Black" "Race\_Hispanic"

[13] "Race\_Other" "Diabetic\_No\_borderline\_diabetes"

[15] "Diabetic\_Yes" "Diabetic\_Yes\_during\_pregnancy"

[17] "GenHealth\_Fair" "GenHealth\_Good"

[19] "GenHealth\_Poor"

程式碼：

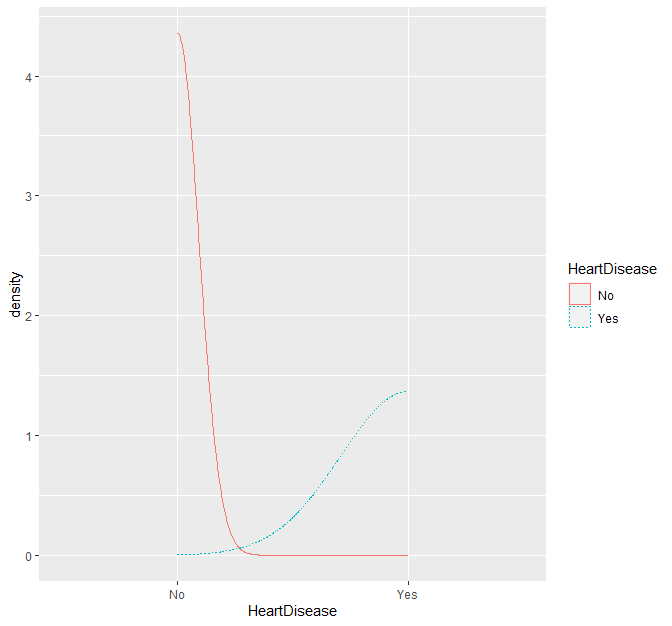
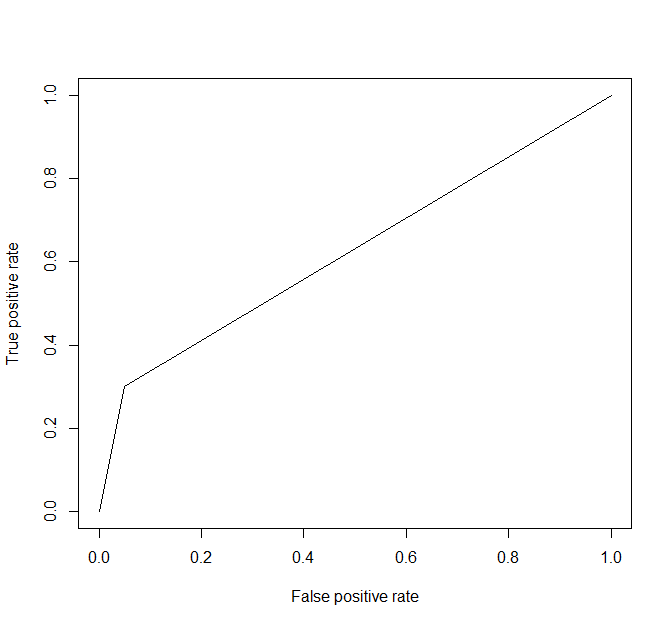
#df\_cor\_f\_in\_csv <- data.frame(cor(f\_in\_csv[, -1]))

#cor\_cols <- colnames(df\_cor\_f\_in\_csv)[df\_cor\_f\_in\_csv[1]>0]

都是正相關的反而不能全取，因為有共線性的問題。

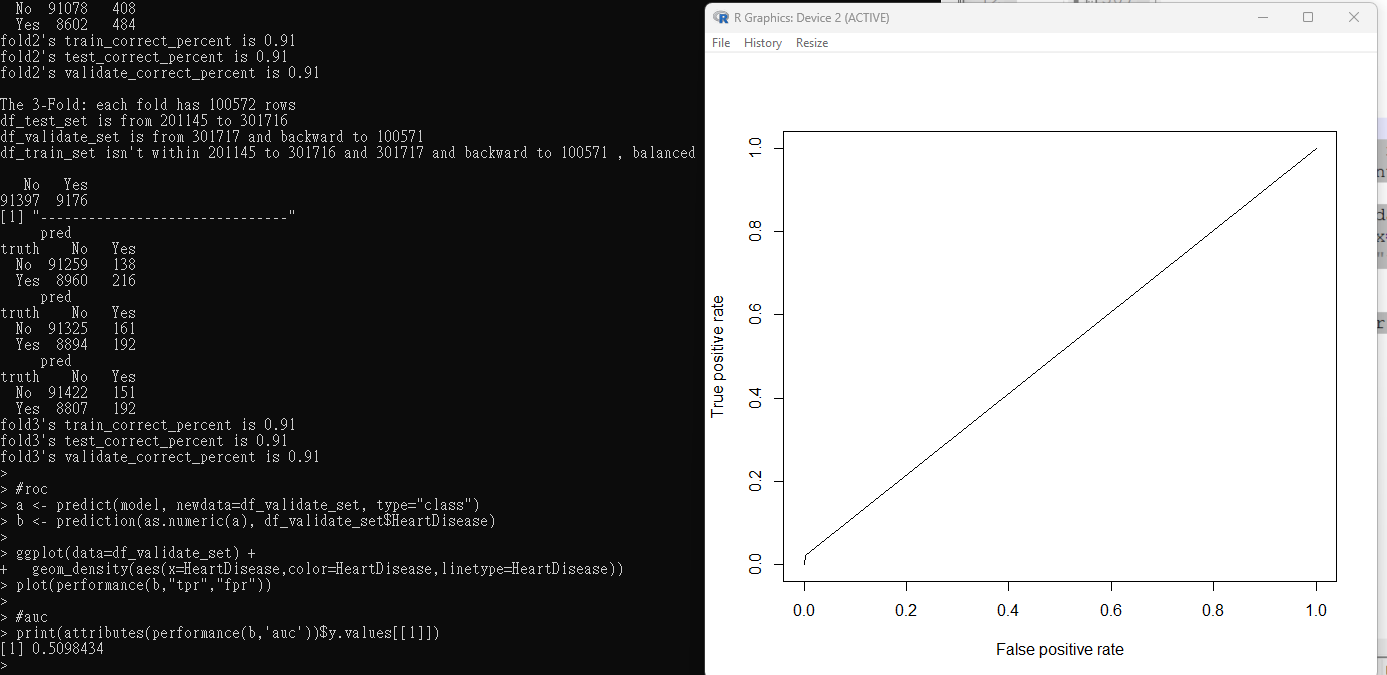
(4)AUC ROC

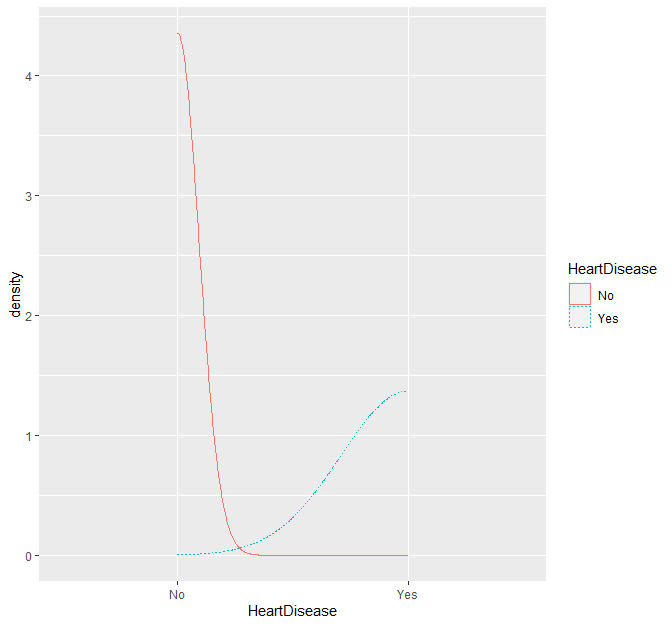
(4-1)用29個欄位；model用smote去train；**AUC=0.6257551**—-------較好





(4-2)用29個欄位；model沒用smote去train；AUC=0.5098434





(5)sensitivity(=recall=True positive rate)、specificity(=True negative rate)

**- 由以下得知，train\_set 一定要處理unbalanced，模型的AUC才能提高、sensitivity and specificity才會正常一點。**

**- 選用**欄位重要性13個欄位，或者選用PC1~PC18，得出的結果跟29個欄位差不多。

\*選用PC1~PC22的效果沒有PC1~PC18好。

\*觀察到PC18時，可以解釋0.80525的資料。

切3個fold，每個fold約10萬。train\_set有取smote、也取29個欄位且排除NA後，建立決策樹模型

|  | set | train\_Accuracy | test\_Accuracy | valid\_Accuracy | train\_Precision | test\_Precision | valid\_Precision | train\_Sensitivity | test\_Sensitivity | valid\_Sensitivity | train\_Specificity | test\_Specificity | valid\_Specificity | train\_F1 | test\_F1 | valid\_F1 | train\_auc | test\_auc | valid\_auc |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | fold1 | 0.82 | 0.9 | 0.9 | 0.63 | 0.4 | 0.4 | 0.25 | 0.25 | 0.24 | 0.96 | 0.96 | 0.96 | 0.36 | 0.31 | 0.3 | 0.60721 | 0.60576 | 0.60336 |
| 2 | fold2 | 0.82 | 0.89 | 0.89 | 0.61 | 0.37 | 0.37 | 0.29 | 0.29 | 0.29 | 0.95 | 0.95 | 0.95 | 0.4 | 0.33 | 0.32 | 0.62367 | 0.62095 | 0.61994 |
| 3 | fold3 | 0.82 | 0.9 | 0.9 | 0.64 | 0.39 | 0.38 | 0.27 | 0.27 | 0.25 | 0.96 | 0.96 | 0.96 | 0.38 | 0.32 | 0.31 | 0.61751 | 0.613 | 0.60708 |
| 4 | best | fold1 | fold1 | fold1 | fold3 | fold1 | fold1 | fold2 | fold2 | fold2 | fold1 | fold1 | fold1 | fold2 | fold2 | fold2 | fold2 | fold2 | fold2 |

切3個fold，每個fold約10萬。train\_set有取smote、也取PC1~PC18且排除NA後，建立決策樹模型

|  | set | train\_Accuracy | test\_Accuracy | valid\_Accuracy | train\_Precision | test\_Precision | valid\_Precision | train\_Sensitivity | test\_Sensitivity | valid\_Sensitivity | train\_Specificity | test\_Specificity | valid\_Specificity | train\_F1 | test\_F1 | valid\_F1 | train\_auc | test\_auc | valid\_auc |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | fold1 | 0.82 | 0.9 | 0.9 | 0.64 | 0.39 | 0.4 | 0.25 | 0.23 | 0.24 | 0.97 | 0.96 | 0.96 | 0.36 | 0.29 | 0.3 | 0.60606 | 0.5994 | 0.60146 |
| 2 | fold2 | 0.82 | 0.89 | 0.9 | 0.63 | 0.38 | 0.38 | 0.26 | 0.25 | 0.25 | 0.96 | 0.96 | 0.96 | 0.37 | 0.3 | 0.3 | 0.61006 | 0.6042 | 0.60282 |
| 3 | fold3 | 0.83 | 0.89 | 0.89 | 0.65 | 0.37 | 0.37 | 0.29 | 0.27 | 0.27 | 0.96 | 0.95 | 0.95 | 0.4 | 0.31 | 0.31 | 0.62473 | 0.61384 | 0.61249 |
| 4 | best | fold3 | fold1 | fold1 | fold3 | fold1 | fold1 | fold3 | fold3 | fold3 | fold1 | fold1 | fold1 | fold3 | fold3 | fold3 | fold3 | fold3 | fold3 |

切3個fold，每個fold約10萬。train\_set有取smote、也取PC1~PC22且排除NA後，建立決策樹模型\_\_效果更差

|  | set | train\_Accuracy | test\_Accuracy | valid\_Accuracy | train\_Precision | test\_Precision | valid\_Precision | train\_Sensitivity | test\_Sensitivity | valid\_Sensitivity | train\_Specificity | test\_Specificity | valid\_Specificity | train\_F1 | test\_F1 | valid\_F1 | train\_auc | test\_auc | valid\_auc |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | fold1 | 0.82 | 0.9 | 0.89 | 0.63 | 0.38 | 0.38 | 0.28 | 0.27 | 0.26 | 0.96 | 0.96 | 0.96 | 0.39 | 0.31 | 0.31 | 0.61843 | 0.61342 | 0.61 |
| 2 | fold2 | 0.82 | 0.89 | 0.89 | 0.63 | 0.38 | 0.37 | 0.27 | 0.26 | 0.25 | 0.96 | 0.96 | 0.96 | 0.37 | 0.3 | 0.3 | 0.61304 | 0.60672 | 0.60485 |
| 3 | fold3 | 0.82 | 0.9 | 0.9 | 0.66 | 0.39 | 0.38 | 0.25 | 0.24 | 0.24 | 0.97 | 0.96 | 0.96 | 0.37 | 0.3 | 0.3 | 0.61089 | 0.60118 | 0.60335 |
| 4 | best | fold1 | fold1 | fold3 | fold3 | fold3 | fold1 | fold1 | fold1 | fold1 | fold3 | fold1 | fold1 | fold1 | fold1 | fold1 | fold1 | fold1 | fold1 |

切3個fold，每個fold約10萬。train\_set有取smote、也取13個欄位(根據caret::varImp)且排除NA後，建立決策樹模型

|  | set | train\_Accuracy | test\_Accuracy | valid\_Accuracy | train\_Precision | test\_Precision | valid\_Precision | train\_Sensitivity | test\_Sensitivity | valid\_Sensitivity | train\_Specificity | test\_Specificity | valid\_Specificity | train\_F1 | test\_F1 | valid\_F1 | train\_auc | test\_auc | valid\_auc |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | fold1 | 0.82 | 0.89 | 0.89 | 0.61 | 0.37 | 0.37 | 0.27 | 0.27 | 0.27 | 0.96 | 0.95 | 0.95 | 0.38 | 0.32 | 0.31 | 0.61475 | 0.61418 | 0.61345 |
| 2 | fold2 | 0.82 | 0.9 | 0.9 | 0.62 | 0.39 | 0.38 | 0.26 | 0.26 | 0.25 | 0.96 | 0.96 | 0.96 | 0.36 | 0.31 | 0.3 | 0.60907 | 0.61005 | 0.60629 |
| 3 | fold3 | 0.82 | 0.89 | 0.89 | 0.61 | 0.37 | 0.37 | 0.28 | 0.28 | 0.27 | 0.96 | 0.95 | 0.95 | 0.39 | 0.32 | 0.31 | 0.61992 | 0.61614 | 0.61337 |
| 4 | best | fold1 | fold2 | fold2 | fold2 | fold2 | fold2 | fold3 | fold3 | fold1 | fold1 | fold2 | fold2 | fold3 | fold1 | fold1 | fold3 | fold3 | fold1 |

切3個fold，每個fold約10萬。train\_set沒取smote、也取29個欄位，建立決策樹模型

|  | set | train\_Accuracy | test\_Accuracy | valid\_Accuracy | train\_Precision | test\_Precision | valid\_Precision | train\_Sensitivity | test\_Sensitivity | valid\_Sensitivity | train\_Specificity | test\_Specificity | valid\_Specificity | train\_auc | test\_auc | valid\_auc |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | fold1 | 0.91 | 0.91 | 0.91 | 0.62 | 0.56 | 0.55 | 0.03 | 0.03 | 0.03 | 1 | 1 | 1 | 0.51355 | 0.5137 | 0.51344 |
| 2 | fold2 | 0.91 | 0.91 | 0.91 | 0.61 | NaN | NaN | 0 | 0 | 0 | 1 | 1 | 1 | 0.5 | 0.5 | 0.5 |
| 3 | fold3 | 0.91 | 0.91 | 0.91 | 0.62 | 0.59 | 0.54 | 0.03 | 0.03 | 0.03 | 1 | 1 | 1 | 0.51418 | 0.51228 | 0.51193 |

(5)chi-squre：檢驗各欄位對output有無顯著

根據鄉民的成果顯示，所有欄位(17個)對於HeartDisease都是顯著的。

4. 書瑋：特徵工程(PCA)【cor 計算欄位相關、ROC、AUC】。

Random Forest importance method

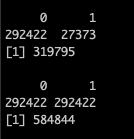
將參數的值隨機化，觀察變化

* MeanDecreaseAccuracy：Accuracy的差異，越大表示該參數越重要
* MeanDecreaseGini：Gini的差異，越大表示原參數資料較純，表示參數越重要

測試正晏建議copy YES的資料，讓它跟NO的依樣多

=> 使用ROSE的 over-sampling

d <- ovun.sample(HeartDisease ~ ., data = d, method = "over",N = 292422\*2)$data

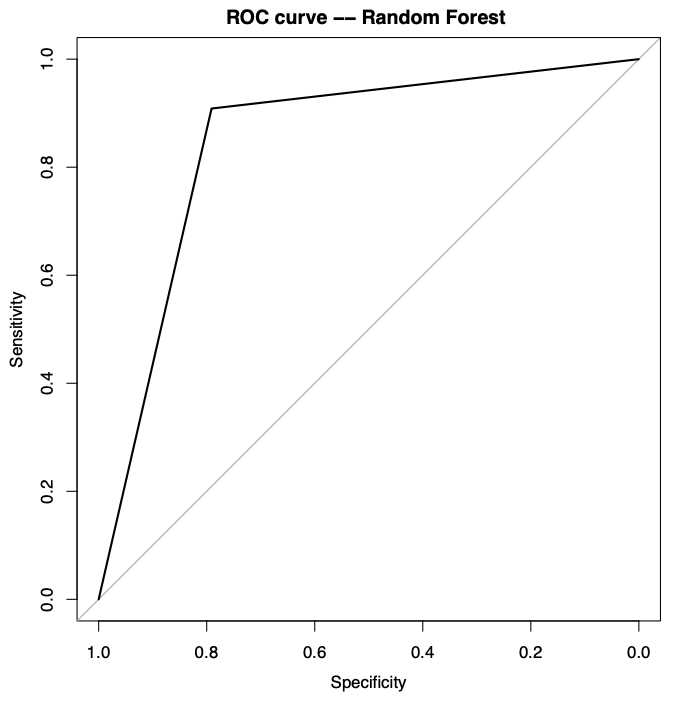


使用Random Forest 產生的 important 如表

<https://docs.google.com/spreadsheets/d/1IM2-4hOuFyP3FJt-YIv7QdDRmdFKmODZmsDmDpPo974/edit#gid=2056993645>

Accuracy = 0.85

Area under the curve: 0.8498



5. 宗佑：HW2(null model、以及其他的confusion matrix的計算結果precision recall)

分享我的程式碼

(1)用list存要測試的model，然後用for迴圈直接把要測的一起測試(使用train這個function，前面list要放method這個參數有支援的model)。

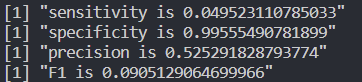
(2)撰寫兩個function去處理confusion matrix，

甲.使用現有套件-confusionMatrix (取名get\_model\_result)

但是和我使用作業2的方式算出來的sen，spe有落差，該套件好像是取一個平均值

乙.使用作業2人工計算(取名hw2\_evaluate)

目前完成tp, tn, fp, fn,f1,null-model,sensity, specificity,ROC(TPR, FPR), ROC(Sensity, Specificity)，



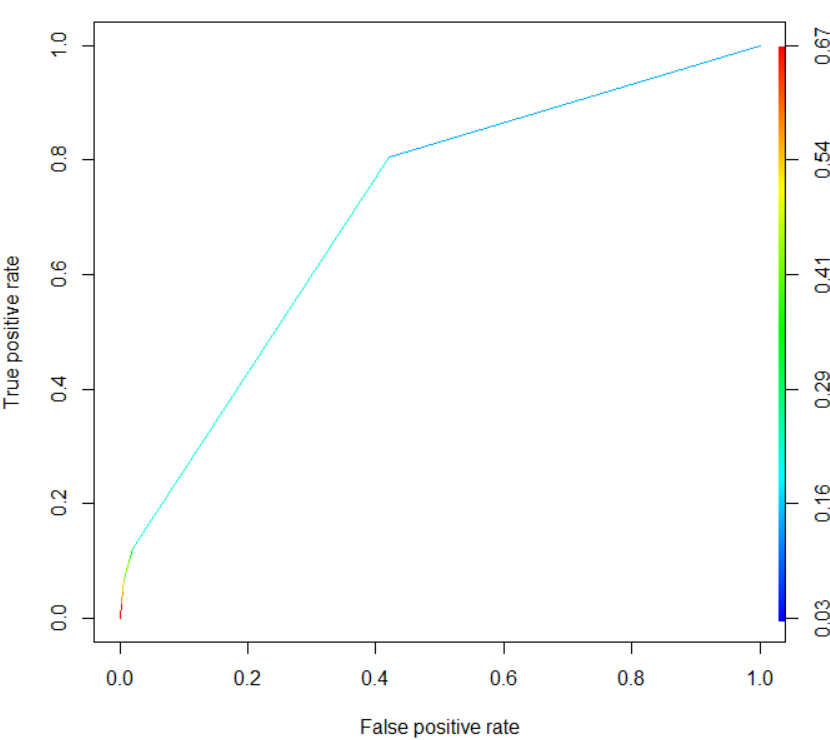
(3)尚待完成

計算loglikelihood,

(4)bug

無法一次跑完script，一次跑兩張圖片時候，圖片儲存有問題

備註:目前尚未篩濾任何特徵，採全部一起挑選進行訓練，只有使用昇豐的方法針對一些類別的資料進行處理(年紀，篩濾重複，Yes/No,Male/Femal等轉換為0/1)，目前切成train test valid但是valid還沒用到

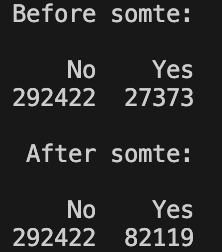


這周六晚上12點回報

元亨：

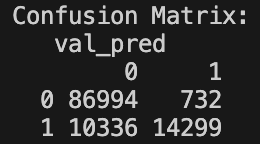
資料前處理：

1. 將資料轉為Numeric
2. Train和Valid比率切7:3
3. 使用smote，將患病資料和未患病資料轉為約1:4的比率(原先約1:10)

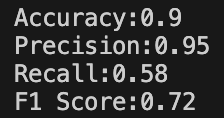


資料建模： 隨機森林

混淆矩陣:



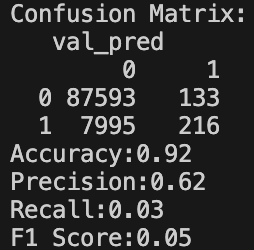
Result:



在是否患有心臟病的分析中，Recall應比Precision重要

相比於原先患病與未患病的比率(1:10), 雖然準確率差不多, 但Recall有顯著提高

原資料的:



書瑋提供 - 變數間的相關性(使用R Cor()):

<https://docs.google.com/spreadsheets/d/1IM2-4hOuFyP3FJt-YIv7QdDRmdFKmODZmsDmDpPo974/edit#gid=0>

**5/15**

今晚討論紀錄：

1. 正晏：會嘗試xgboost(下週四上課給)；視覺化先做dataset。
2. 元亨：會嘗試隨機森林(這週四上課給)。
3. 昇豐：我會做特徵工程(PCA、ROC、AUC)、Hypothesis Test(chi-squre 完成的話，pass 給宗佑。)
4. 書瑋：特徵工程(PCA)【cor 計算欄位相關、ROC、AUC】。  
   cor 計算欄位相關；最後採用7個欄位也能得到0.9的準確率。  
   書瑋提供 - 變數間的相關性(使用R Cor()):

<https://drive.google.com/drive/folders/1NKB5WjAGIUhrjHC9ABC6dFSyDZwdGpSl?usp=share_link><https://docs.google.com/spreadsheets/d/1IM2-4hOuFyP3FJt-YIv7QdDRmdFKmODZmsDmDpPo974/edit#gid=0>

1. 宗佑：HW2(null model、以及其他的confusion matrix的計算結果precision recall)

5/15 21:30~待討論

**1. 決策數的層數加多。正晏可以看每一層用那些屬性。**

>>昇：的確要調整rpart.control，但不能只定義maxdepth，就能生成決策樹，且會預測YES了。

**2. 書瑋建議資料平均，讓YES、NO的資料一樣。**

>> 昇：我用下列方式處理資料不平均

library('performanceEstimation') #instead of library('smotefamily')

smote\_f\_in\_csv <- smote(HeartDisease ~ ., data =f\_in\_csv[1:100000,])

for(i in colnames(smote\_f\_in\_csv))

{

na\_nums <- 0

na\_nums <- sum(is.na(smote\_f\_in\_csv[[i]]))

if (na\_nums > 0)

{

print(i)

print(na\_nums)

}

}

smote後，會有NA

> nrow(smote\_f\_in\_csv)

[1] 62937

> nrow(smote\_f\_in\_csv[!is.na(smote\_f\_in\_csv$HeartDisease),])

[1] 44955

> nrow(smote\_f\_in\_csv[is.na(smote\_f\_in\_csv$HeartDisease),])

[1] 17982

**3. 正晏建議copy YES的資料，讓它跟NO的依樣多。**

>> 昇：沒有嘗試

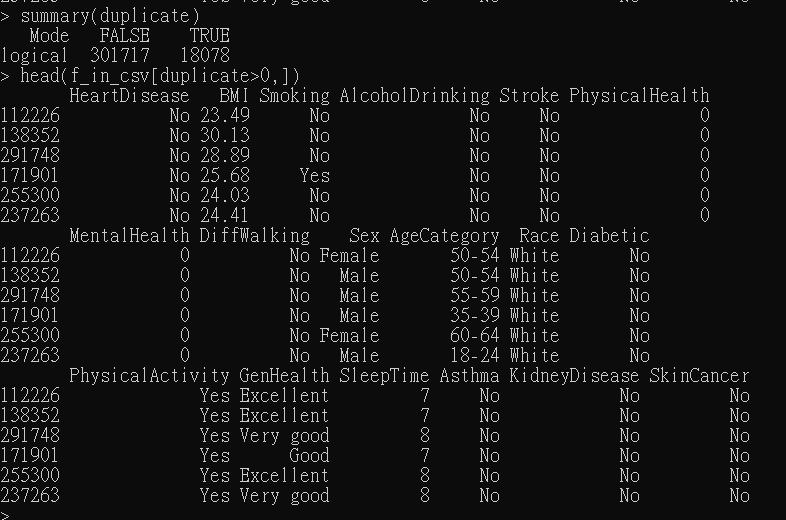
**4. PCA方法，或者暴力破解，挑出能夠預測YES的屬性集合。**

>> 昇：沒有嘗試

昇：參考鄉民：<https://www.kaggle.com/code/karenhu8/hux-da5030-project>

參考資料欄位說明 <https://www.kaggle.com/code/andls555/heart-disease-prediction>

1. 有18078列重複



> nrow(f\_in\_csv)

[1] 319795

> f\_in\_csv <- distinct(f\_in\_csv)

> nrow(f\_in\_csv)

[1] 301717

**2.1. 資料轉型**

> str(f\_in\_csv)

'data.frame': 301717 obs. of 18 variables:

$ HeartDisease : chr "No" "No" "No" "No" ... <--- target column

---------------------------------------------------------------------------------------------------

$ **BMI : num** 16.6 20.3 26.6 24.2 23.7 ...

$ Smoking : chr "Yes" "No" "Yes" "No" ... <---- 轉成1,0

$ AlcoholDrinking : chr "No" "No" "No" "No" ...

$ Stroke : chr "No" "Yes" "No" "No" ...

$ **PhysicalHealth : num** 3 0 20 0 28 6 15 5 0 0 ...

$ **MentalHealth : num** 30 0 30 0 0 0 0 0 0 0 ...

$ DiffWalking : chr "No" "No" "No" "No" ...

$ Sex : chr "Female" "Female" "Male" "Female" ...

$ **AgeCategory : chr** "55-59" "80 or older" "65-69" "75-79" ...  **<-- 這個要轉成num**

$ Race : chr "White" "White" "White" "White" ... ←–dummy

$ Diabetic : chr "Yes" "No" "Yes" "No" ...

$ PhysicalActivity: chr "Yes" "Yes" "Yes" "No" ...

$ GenHealth : chr "Very good" "Very good" "Fair" "Good" ...

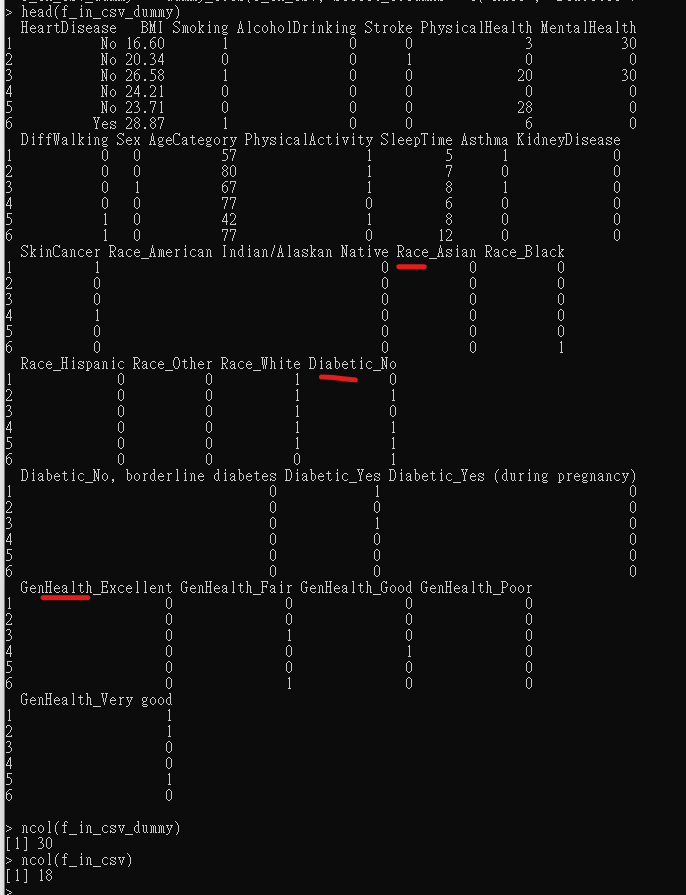
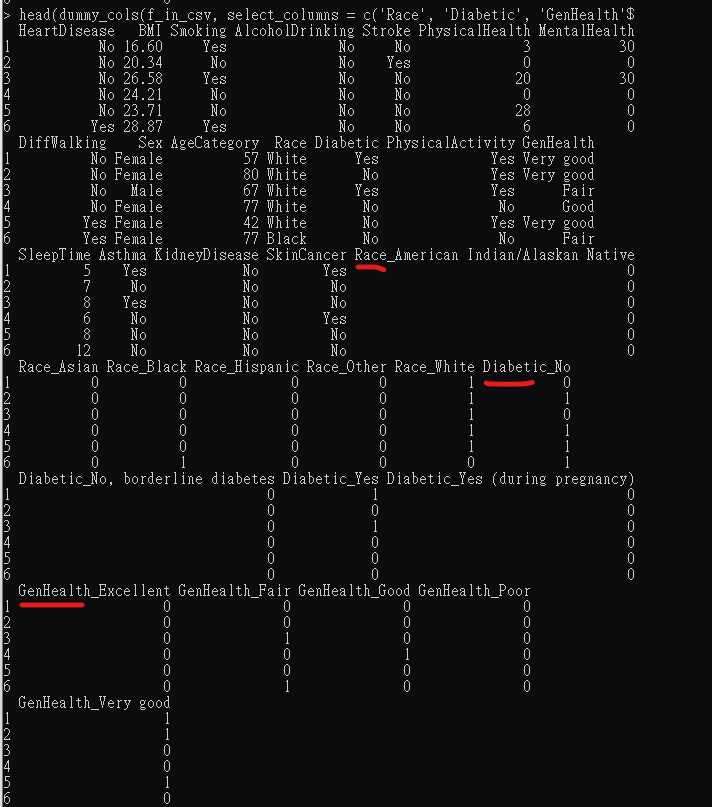
$ **SleepTime : num** 5 7 8 6 8 12 4 9 5 10 ...

$ Asthma : chr "Yes" "No" "Yes" "No" ...

$ KidneyDisease : chr "No" "No" "No" "No" ...

$ SkinCancer : chr "Yes" "No" "No" "Yes" ...

上面紅字欄位資料含有多種類，因此用dummy把之展開



3. unblanced data

#> table(f\_in\_csv$HeartDisease)

#

# No Yes

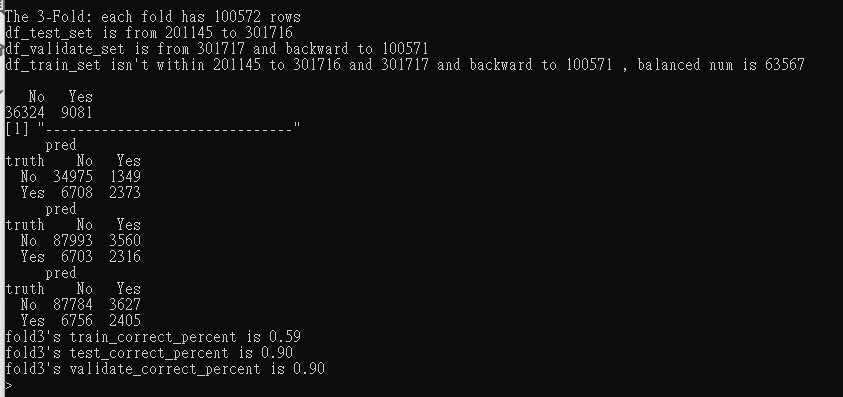
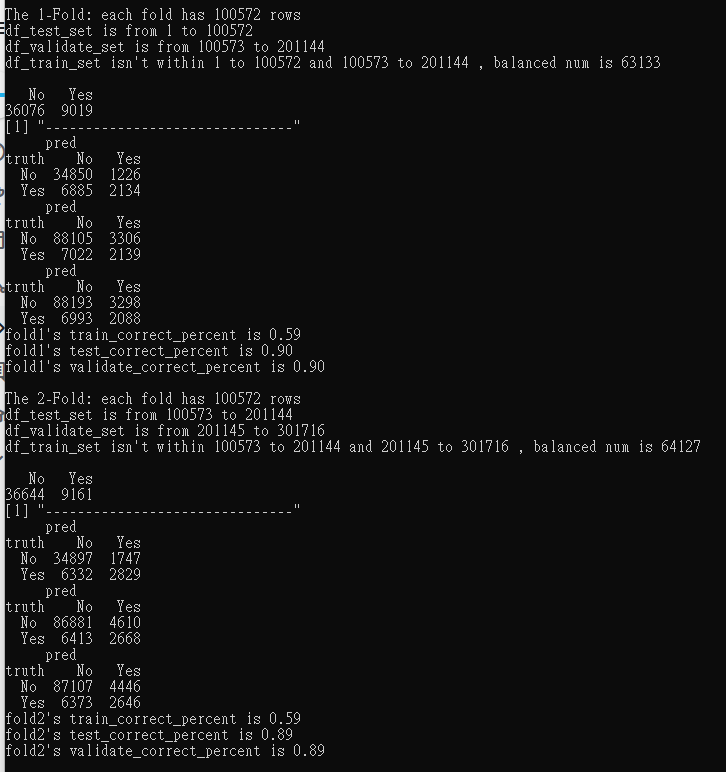
#274456 27261

3.1.

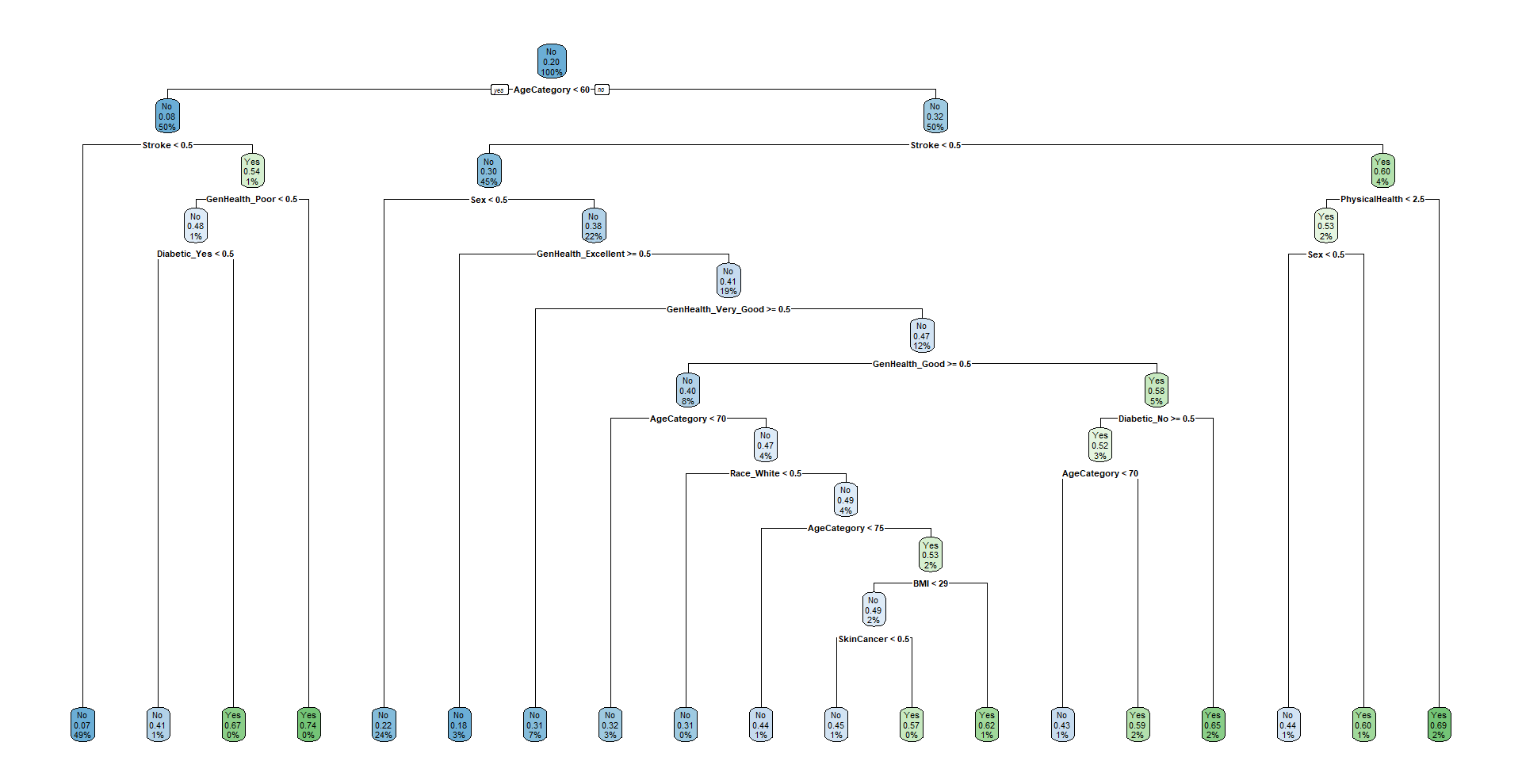
切3個fold

confusion matrix 依序是train, test, valid

* train\_correct\_percent太低的原因是，smote後的train\_set有NA(約1萬5千筆)，應排除而未排除導致。  
  training\_ls <- c(training\_ls, format(round(CP\_train\_correct\_num/**nrow(df\_train\_set)**, 2), nsmall = 2))



第三個fold的train的決策樹



**5/11**

2. 5/15晚上，A組能討論的時間。

5/15 21:30~

**1. 決策數的層數加多。正晏可以看每一層用那些屬性。**

**2. 書瑋建議資料平均，讓YES、NO的資料一樣。**

**3. 正晏建議copy YES的資料，讓它跟NO的依樣多。**

**4. PCA方法，或者暴力破解，挑出能夠預測YES的屬性集合。**

下週四5/18，能夠讓B組的接手做視覺。

3. 了解資料

> nrow(f\_in\_csv)

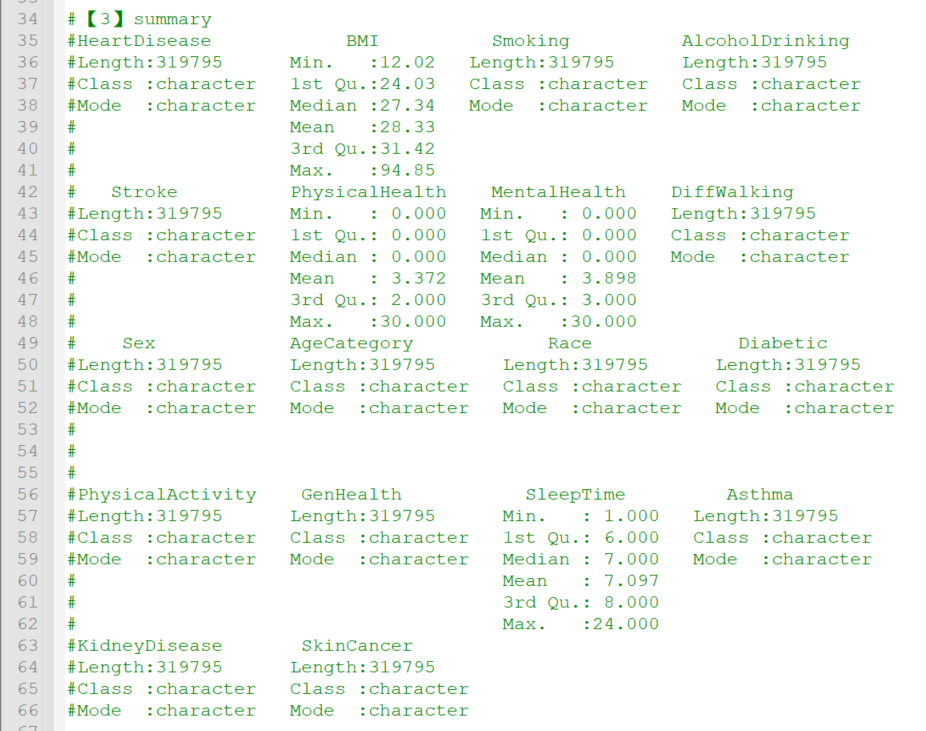
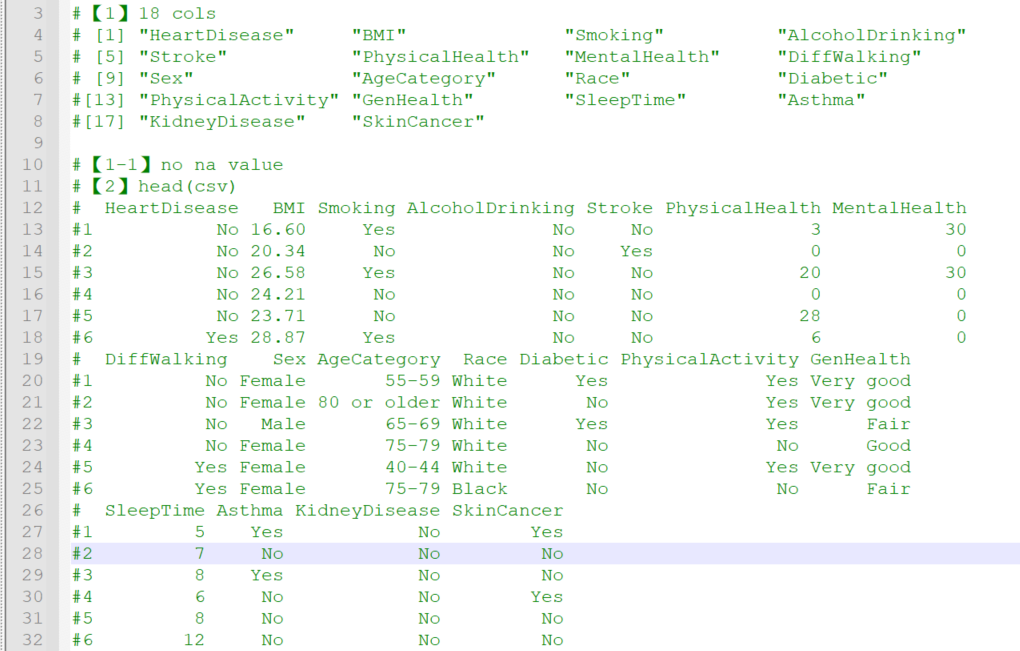
[1] 319795

> sum(ifelse(f\_in\_csv$HeartDisease == "Yes", 1, 0))

[1] 27373

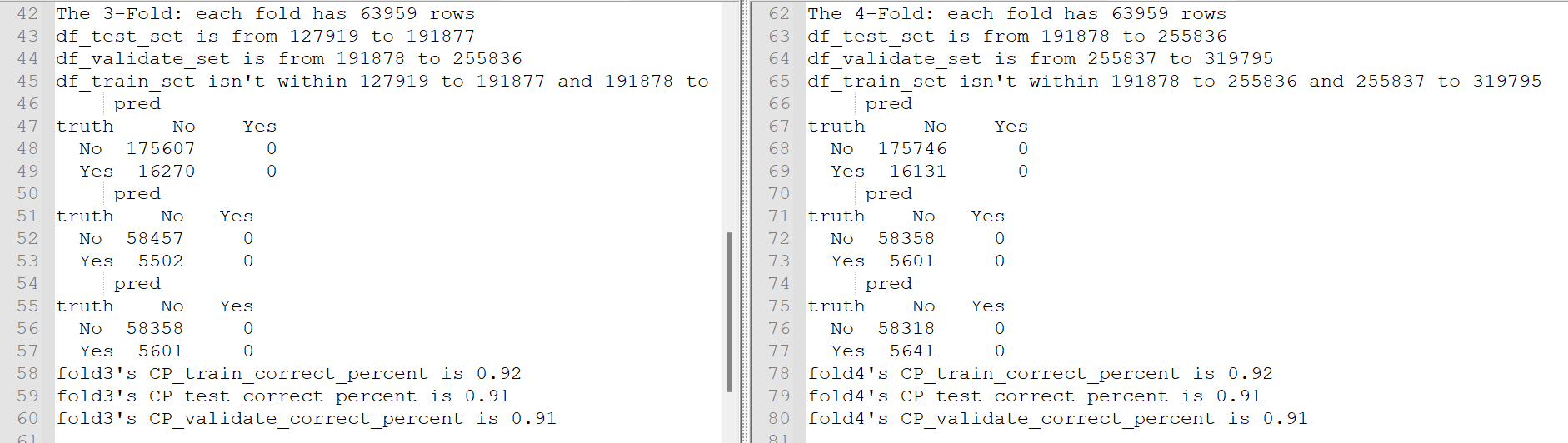
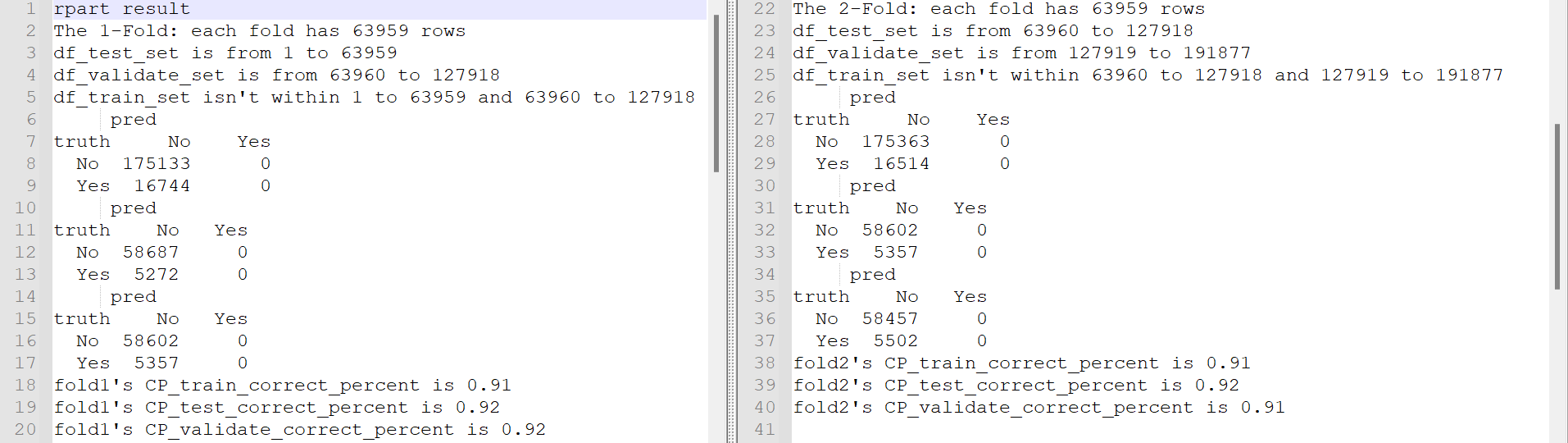
> sum(ifelse(f\_in\_csv$HeartDisease == "No", 1, 0))

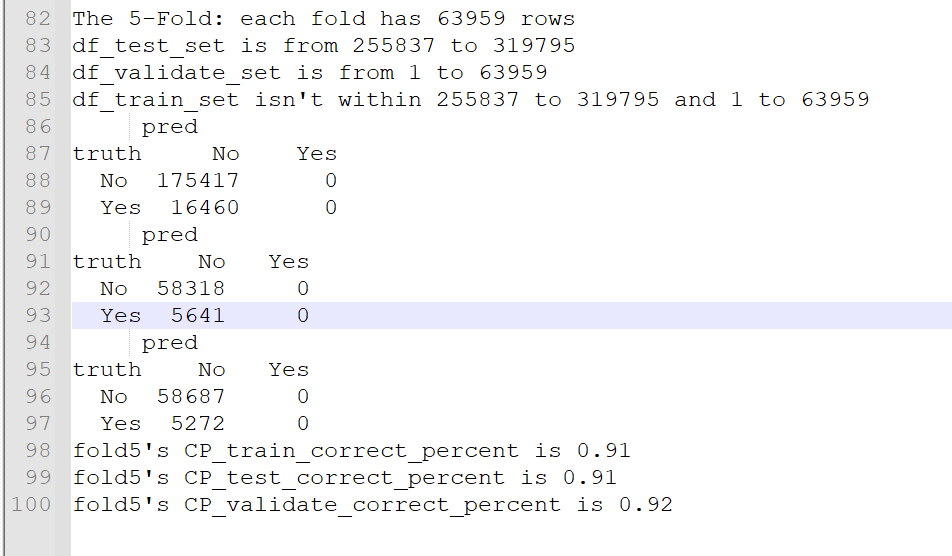
[1] 292422





based on hw3's rpart and K-fold is 5：





**4/27**

0. 選組長。

5~6人

| 人數  2  optional+1  CARTER  BUSKY  JUDE  \*正晏 | 人數  1  optional+1  書瑋 | 人數  1  正晏  AUSTIN | 人數  1 | 人數  optional+1 |
| --- | --- | --- | --- | --- |
| **2.2. data(input)：**  a. 資料前處理：有無發現資料清洗的難處？  **2.3. model：**  a. 選哪個模型進行ML？例如決策數、XGBoost。  b. 選用哪個特徵工程來優化模型？  c. null model定義與建立。 | **2.4. evaluation(ouput)：**  a. 用哪個方法來評估模型好壞？  例如Hypothesis Test(包含Confusion matrix、precesion and recall)、R-squre | **2.5. present：**  a. 從第2~4階段選擇資料視覺作用之處：  例如  - 清洗前、後的資料樣貌。  - 最優模型的評估結果。  - 以及特徵工程過程上的評估結果。  可能得遵照老師的建議，一併做on-line visualization。 | **2.6. deploy：**除了製作還原專案的步驟，在報告當下demo。 | **統整ReadMe.md** |

共同做到：

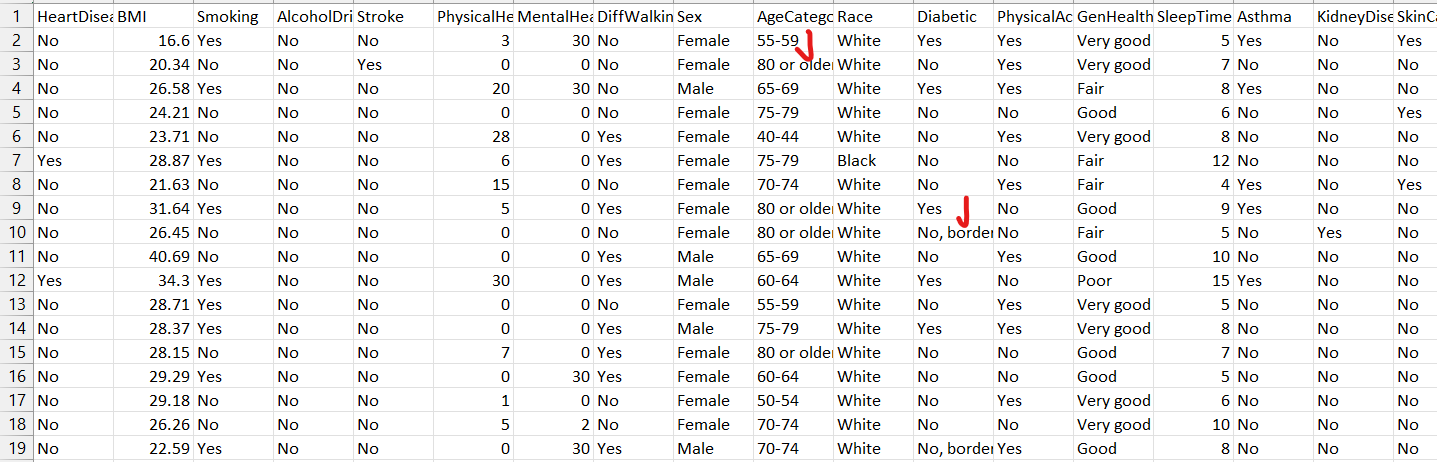
b. 用ReadMe.md做課堂報告即可。 因此不熟者要練習。

>可以作圖放在ReadMe，屆時瀏覽器打開圖片講解也可。

c. 專案艱辛之處。

d. 專案引用的套件與外部資料。

3. 會按git flow預設各分支的定義來做程式合作，因此不熟者要練習。待老師公布期末repo的目錄結構後，再來討論各分支能做什麼事、不能做什麼事。



2. 下週4/27下課後可討論幾點：

0. 選組長。

1. 是否對主題有興趣？若否，可直接去問其他隊，沒問題的。

**2. 從practice\_R教科書的資料科學流程討論：(如有未健全處，再補充可)**

2.1. goal ： 要解決什麼問題？我們就可定義期末報告的主題集合。

2.2. data(input)：

a. 資料前處理：有無發現資料清洗的難處？

2.3. model：

a. 選哪個模型進行ML？例如決策數、XGBoost。

b. 選用哪個特徵工程來優化模型？

c. null model定義與建立。

2.4. evaluation(ouput)：

a. 用哪個方法來評估模型好壞？

例如Hypothesis Test(包含Confusion matrix、precesion and recall)、R-squre

2.5. present：

a. 從第2~4階段選擇資料視覺作用之處：

例如

- 清洗前、後的資料樣貌。

- 最優模型的評估結果。

- 以及特徵工程過程上的評估結果。

可能得遵照老師的建議，一併做on-line visualization。

b. 用ReadMe.md做課堂報告即可。 因此不熟者要練習。

c. 專案艱辛之處。

d. 專案引用的套件與外部資料。

2.6. deploy：除了製作還原專案的步驟，在報告當下demo。

