

66

(0)line群裡的evernote連結

<https://www.evernote.com/shard/s379/sh/5b5a1fdb-f5e4-40ee-bd8e-666c3f6ae3de/PDiHx0j2eqGqNBpwTjX7x1dLILlctS34A3K1Z1PvdQKRZDvlxy9KmNkTQ>

(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](#)

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](#)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

6/14

昇豐: readme

1. 刪除此部分

Quick start

If our topic interests you, please follow the instructions below to replicate our experimental results.

```
Rscript code/main.R --data_source 1 --model 1
```

此區域改放置指示我們文件擺放的git目錄位置。正晏:檔案將放在docs中。

文件是:3個pdf檔案, 來自

「簡報正式區、簡報草稿、此份google doc」下載的pdf檔案。

2. 在此部分, 幫忙加上3個模型的執行語法與預估執行時間 決策樹(預估3~6分鐘)

```
Rscript code/main.R --d 1 --m 1
```

隨機森林(預估2~5分鐘)

Rscript code/main.R --d 2 --m 2

XGBoost(預估1~2分鐘)

Rscript code/main.R --d 2 --m 3

- 使用不同模型執行預測，產生各模型預測結果

```
Rscript code/main.R --data_source 1 --model 1
```

參數:

- --data_source or --d: 選擇使用的資料集(1: 原始資料, 2: 已處理後資料)
- --model or --m: 選擇使用的模型(1: Decision tree, 2: Random forest, 3: XGBoost)

提醒

- 已處理後資料: 表示先經過Data process&smote切割後的資料集
- Decision tree: 不支援使用已處理後資料，且耗時較長

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)開立編輯權限。
[Austin] Done, 6/12.

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



原始資料集的分析問題：NA、重複、特徵類型、不平衡

```
> str(dataset)
'data.frame': 319,795 obs. of 18 variables:
 $ HeartDisease : chr "No" "No" "No" "No" ...
 $ BMI          : num 16.6 20.3 26.6 24.2 23.7 ...
 $ Smoking      : chr "Yes" "No" "Yes" "No" ...
 $ 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" ...
 $ Race         : chr "White" "White" "White" "White" ...
 $ 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" ...
```

summary(duplicated(dataset))

Mode	FALSE	TRUE
logical	301,717	18,078

table(dataset\$HeartDisease)

	No	Yes
	292,422	27,373

沒有NA; 有18078列重複; 沒病: 有病約10: 1(不平衡資料集)

(3) 各模型的執行情況, 第54、55頁, 示意圖____20230612完成

54



55



56



57



58



各模型執行情況_1

model	決策樹	隨機森林	XGBoost
君主	真壁	元亨/唐璋	宗佑/正英
k-fold/ training+test	k-fold, k = 3	training+test 70%+30%	training+test 70%+30%
smote	smote(trainset=100,000)	smote(trainset=210,000)	smote(trainset=210,000)
threshold(閾值)	--	--	基於F1得出 29 -> 0.4 13 -> 0.35
model 參數	minsplit = 5, minbucket = 5, cp = 0.001	ntree=100, mtry=3	objective = binary:logistic, eval_metric = error, max_depth = 6, eta = 0.3
feature selection	前處理後，產生29 feature。 基於glm，得出13 feature。		

(4) 介紹最終建議方案：

(4-1) 特徵工程，第58~59頁，示意圖

56

模型執行時間比較

Model	Training Time (s)	Validation Time (s)	Total Time (s)
Logistic Regression	1.2	0.5	1.7
Decision Tree	0.8	0.3	1.1
Random Forest	2.5	1.0	3.5
Support Vector Machine	3.0	1.2	4.2
Naive Bayes	0.5	0.2	0.7
K-Nearest Neighbors	1.5	0.8	2.3
Gradient Boosting	4.0	1.5	5.5
Neural Network	5.0	2.0	7.0

57

1. 模型執行時間比較 (續)

Model	Training Time (s)	Validation Time (s)	Total Time (s)
Logistic Regression	1.2	0.5	1.7
Decision Tree	0.8	0.3	1.1
Random Forest	2.5	1.0	3.5
Support Vector Machine	3.0	1.2	4.2
Naive Bayes	0.5	0.2	0.7
K-Nearest Neighbors	1.5	0.8	2.3
Gradient Boosting	4.0	1.5	5.5
Neural Network	5.0	2.0	7.0

58

X-squared檢視欄位跟target HeartDisease的顯著性

col_order	col_name	X-squared	df	p-value
2	BMI	6724.613578	3603	2.49E-192
3	Smoking	3713.815575	1	0
4	AlcoholDrinking	329.1041963	1	1.51E-73
5	Stroke	12390.18061	1	0
6	PhysicalHealth	9735.616088	30	0
7	MentalHealth	971.4189962	30	5.50E-185
8	DiffWalking	12953.23319	1	0
9	Sex	1568.808198	1	0
10	AgeCategory	19299.92039	12	0
11	Race	844.314886	5	2.99E-180
12	Diabetic	10959.86128	3	0
13	PhysicalActivity	3199.864826	1	0
14	GenHealth	21542.17736	4	0
15	SleepTime	2303.946242	23	0
16	Asthma	549.2855397	1	1.80E-121
17	KidneyDisease	6741.981347	1	0
18	SkinCancer	2784.787544	1	0

59

X-squared檢視欄位跟target HeartDisease的顯著性 (續)

col_order	col_name	X-squared	df	p-value
2	BMI	6724.613578	3603	2.49E-192
3	Smoking	3713.815575	1	0
4	AlcoholDrinking	329.1041963	1	1.51E-73
5	Stroke	12390.18061	1	0
6	PhysicalHealth	9735.616088	30	0
7	MentalHealth	971.4189962	30	5.50E-185
8	DiffWalking	12953.23319	1	0
9	Sex	1568.808198	1	0
10	AgeCategory	19299.92039	12	0
11	Race	844.314886	5	2.99E-180
12	Diabetic	10959.86128	3	0
13	PhysicalActivity	3199.864826	1	0
14	GenHealth	21542.17736	4	0
15	SleepTime	2303.946242	23	0
16	Asthma	549.2855397	1	1.80E-121
17	KidneyDisease	6741.981347	1	0
18	SkinCancer	2784.787544	1	0

60

X-squared檢視欄位跟target HeartDisease的顯著性 (續)

col_order	col_name	X-squared	df	p-value
2	BMI	6724.613578	3603	2.49E-192
3	Smoking	3713.815575	1	0
4	AlcoholDrinking	329.1041963	1	1.51E-73
5	Stroke	12390.18061	1	0
6	PhysicalHealth	9735.616088	30	0
7	MentalHealth	971.4189962	30	5.50E-185
8	DiffWalking	12953.23319	1	0
9	Sex	1568.808198	1	0
10	AgeCategory	19299.92039	12	0
11	Race	844.314886	5	2.99E-180
12	Diabetic	10959.86128	3	0
13	PhysicalActivity	3199.864826	1	0
14	GenHealth	21542.17736	4	0
15	SleepTime	2303.946242	23	0
16	Asthma	549.2855397	1	1.80E-121
17	KidneyDisease	6741.981347	1	0
18	SkinCancer	2784.787544	1	0

chi-square檢視欄位跟target HeartDisease的顯著性：

未處理，17個欄位

前處理後，29個欄位

col_order	col_name	X-squared	df	p-value
2	BMI	6724.613578	3603	2.49E-192
3	Smoking	3713.815575	1	0
4	AlcoholDrinking	329.1041963	1	1.51E-73
5	Stroke	12390.18061	1	0
6	PhysicalHealth	9735.616088	30	0
7	MentalHealth	971.4189962	30	5.50E-185
8	DiffWalking	12953.23319	1	0
9	Sex	1568.808198	1	0
10	AgeCategory	19299.92039	12	0
11	Race	844.314886	5	2.99E-180
12	Diabetic	10959.86128	3	0
13	PhysicalActivity	3199.864826	1	0
14	GenHealth	21542.17736	4	0
15	SleepTime	2303.946242	23	0
16	Asthma	549.2855397	1	1.80E-121
17	KidneyDisease	6741.981347	1	0
18	SkinCancer	2784.787544	1	0

col_order	col_name	X-squared	df	p-value
2	BMI	6220.709	2603	7.42E-143
3	Smoking	3296.117	1	0
4	AlcoholDrinking	387.1195	1	2.11E-68
5	Stroke	15433.4	1	0
6	PhysicalHealth	8597.502	30	0
7	MentalHealth	1005.589	30	7.19E-205
8	DiffWalking	11040.48	1	0
9	Sex	10271.667	1	0
10	AgeCategory	18912.37	12	0
11	PhysicalActivity	2583.152	1	0
12	SleepTime	3705.084	23	0
13	Asthma	286.2422	1	5.18E-86
14	KidneyDisease	4341.505	1	0
15	SkinCancer	2479.035	1	0
16	Race_American_Indian_Alaskan_Native	12.66774	1	0.000272039
17	Race_Asian	325.688	1	9.42E-74
18	Race_Black	63.58335	1	1.54E-15
19	Race_Hispanic	499.2888	1	1.36E-110
20	Race_White	11.14858	1	0.00084855
21	Race_White	721.2468	1	7.19E-159
22	Diabetic_No	8310.73	1	0
23	Diabetic_No_borderline_diabetes	57.20982	1	2.56E-14
24	Diabetic_Yes	9658.298	1	0
25	Diabetic_Yes_during_pregnancy	72.56168	1	1.62E-17
26	GenHealth_Excellent	3867.477	1	0
27	GenHealth_Fair	6192.675	1	0
28	GenHealth_Good	338.1005	1	4.19E-68
29	GenHealth_Poor	8971.352	1	0
30	GenHealth_Very_Good	2349.822	1	0

- pseudo code : `chisq.test(dataset$HeartDisease, dataset[,i], correct=FALSE)`
- **p-value < 0.05, reject H0**，所有欄位都與HeartDisease有相關(dependent)

(4-2) 介紹此方案，第60~68頁，示意圖

成

(以random forest、xgboost為主；若篇幅不夠、才加入決策樹)

59

60

61

62

63

XGBoost_1_參數與演算法選擇：

(1) 參數說明

- max_depth**：樹的最大深度，使用max_depth=6。
目的：數值愈大模型擬合度越高。
- eta**：又稱為learning_rate，使用預設值0.3。
目的：此參數用於防止over fitting。
- eval_metric = error**：此為二進制分類錯誤率，預設使用0.5來判斷。
目的：評估每回迭代的分類效果，公式「#(wrong cases) 除以 #(all cases)」

(2) 演算法選擇

- binary:logistic**：羅吉斯回歸，model對每一筆預測資料輸出機率
考量實務面，選用**logistic**。理由：使用閾值(threshod)、產生ROC圖。
- binary:hinge**：基於loss function是hinge的SVM，model對每一筆預測資料輸出2元結果(1, 0)

(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

69

70

71

72

再次重申評估與結論：

透過兩者的比對，得出分析價值是：(1) 確定**無相關**的特徵其實是重要的；
(2) 得出不具有共線性效果的特徵，既維持效能又提升建立模型的效率。

原始資料集觀察				特徵工程		
統計 直觀	Feature	No, 有病%	Yes, 有病%	Feature	caret::varImp (glm_model)	car::vif (glm_model)
有相關	Smoking	7.47	36.4	Race_Hispanic BMI DiffWalking GenHealth_Excellent Asthma	4.0550250	1.024754
	Stroke	7.47	36.4		4.4294187	1.108250
	AgeCategory (50-54)5.45	22.6(80 up)	29.3		6.5815495	1.297409
	KidneyDisease	7.77	29.3		7.0720475	1.229081
	GenHeath (Excellent)2.24	34.1(Poor)	6.3		7.2590399	1.053452
無相關	DiffWalking	6.3	22.6	KidneyDisease Smoking GenHealth_Good GenHealth_Poor GenHealth_Fair Stroke Sex(Gender) AgeCategory	9.7363807	1.036308
	BMI	27.26	28.34		13.0590904	1.031146
	Sex(Gender)	10.6(Male)	6.69(Female)		14.6508521	1.585377
	Asthma	8.1	11.5		20.6333198	1.406402
	Race	3.3(Asian)	10.4(Indian)		21.7952554	1.623939
					23.3059047	1.021945
					25.4474306	1.059703
					47.6255661	1.104021

備註：glm的權位重要性，取>4，13個權位；VIF值皆 < 2

(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。

69

(以下為 DPM 模型預測結果)

1. 預測值與實際值之差異 (RMSE) 為 0.012

2. 預測值與實際值之差異 (RMSE) 為 0.012

70

DEMO 內容

71

1. 模型評估

(以下為 DPM 模型預測結果)

1. 預測值與實際值之差異 (RMSE) 為 0.012

2. 預測值與實際值之差異 (RMSE) 為 0.012

72

模型評估

(以下為 DPM 模型預測結果)

1. 預測值與實際值之差異 (RMSE) 為 0.012

2. 預測值與實際值之差異 (RMSE) 為 0.012

73

模型評估

(以下為 DPM 模型預測結果)

1. 預測值與實際值之差異 (RMSE) 為 0.012

2. 預測值與實際值之差異 (RMSE) 為 0.012

專案艱辛之處

昇豐：從「思想的巨人，行動的低端」中脫離，進行行動力變革。讀過不少 ML、DS、統計的書，但工作既不是此領域、大學的統計也沒學好，因此這段期末報告的準備過程真的是練習對抗的過程；所幸考試後的課程會提點分析工具運用的訣竅外，老師也願意給予意見，組員也願意與信任地按照我仿照課堂教科書訂定的報告綱要準備。

書瑋：第一次接觸資料分析，也是第一次接觸 R，都是透過上課及作業學到的東西，加上不斷地 Google 及詢問 ChatGPT，慢慢拼湊出最後的結果。

元亨：(程式的互相合作)我們總共採用了 3 種(決策樹、隨機森林、Xgboost)模型來分析資料，實際寫了 5 份建模程式及 1 份視覺化程式，在這些 Code 中，每個人的 Code pattern、前處理、套件選用等都不盡相同，讓我們在溝通、整合程式時費了一番功夫。

正晏：(對於不熟悉領域的探索)平時在於工作中並未使用到與 Data Science 相關的技能，本次專案算是一個新領域的探索，在專案建立的過程中，除了需要時不時複習老師上課的內容之外，還需要不斷地去學習，在資料集確定後就遇到了第一個難關，在不同的模型上，會針對其所需要的輸入去調整 Attribute，在建模的過程中，發現若不了解模型的原理，在參數的值上也難以去調整。

宗佑：(資料集不平衡)本主題資料集相當不平衡，患者與非患者為 1 比 13，基本上使用 null model 全部預測患者無病，準確性即高達九成多，然而這樣的預測沒有太大意義，因此我們針對不平衡的資料瀏覽許多資料，最後透過 smote，降低樣本比例至 1 比 4，提高 sensitivity。

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

69

70

71

72

73

DEMO說明

(1) Rscript語法

- 產生資料分析圖示
`Rscript data_plot.R`
- 執行主程式
`Rscript main.R --d 2 --m 2`
`Rscript main.R --d 2 --m 3`

(2)參數說明

--data_source or --d

- 讀取原資料，會處理dummy和smote
- 讀取已處理資料

--model or --m

- decision tree (不支援data_source = 2，且會耗時3分鐘以上)
- random forest
- xgboost

Feature	Accuracy	Precision	Sensitivity	Specificity	F1 Score	AUC
29Features	0.9	0.41	0.24	0.97	0.31	0.8
13Features	0.9	0.41	0.28	0.96	0.33	0.82

隨機森林、XGBoost產生檔案交集：

- result.csv：模型預測評估結果。
- roc-13、roc-29：不同Feature的AUC表現。
- ImportantFeature：兩個模型產生的變量重要性。

(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

1

2

3

4

(1) 範例：專案甘苦談與專案引用的套件與外部資料，階層式綱要如下：

2. 引用套件、外部資料

套件

```
#packages
#1. rpart / rpart.plot : decision tree
#2. corplot : correlation among cols
#3. performanceEstimation : smote to tackle unbalanced dataset
#4. dplyr : tackle dataframe
#5. ROCR : roc curve and auc
#6. caret::varImp : the important order of cols for target y
#7. car::vif : check collinearity in cols
```

外部資料：

kaggle dataset

[\(example\)Hux DA5030 Project | Kaggle](#)
[\(example\)STAT 45T Project | Kaggle](#)
[\(example_explaindata\)Heart Disease Scoring | Wh](#)
[\(example_explaindata\)Heart Disease Prediction | k](#)

smote

[\(performanceEstimation\)imbalanced data - packag](#)
VIF

[\(glm_Variable Importance_VIF\)How to Perform Logistic](#)
[\(multicollinearity\)How to Fix in R: there are aliased](#)
AUC

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)

```
Accuracy:0.82
Precision:0.27
sensitivity(Recall):0.66
Specificity:0.83
F1 Score:0.38
Setting levels: control = 0, case = 1
Setting direction: controls < cases
AUC:0.74
```

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-square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

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) :

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

	MeanDecreaseAccuracy	MeanDecreaseGini
AgeCategory	66.4559862	5102.20891
Sex	37.6880821	881.84802
BMI	37.1622488	10321.42351
PhysicalHealth	30.9170674	2258.00245
GenHealth_Good	30.1246583	404.42644
MentalHealth	29.1394247	2048.08265
Stroke	28.415993	1133.09927
Smoking	22.5007621	847.41524
SleepTime	20.7433561	3126.5215
DiffWalking	19.5596559	1277.51228
Diabetic_Yes	18.6405389	460.41111
SkinCancer	16.6379981	726.61206
Diabetic_No	15.5995019	347.23006
KidneyDisease	15.4077994	483.8291
Race_White	13.744645	447.52379
GenHealth_Fair	13.5128123	476.07484
Asthma	13.0573223	707.78092
PhysicalActivity	12.1913921	896.01873
GenHealth_Poor	11.0608412	519.43874
AlcoholDrinking	8.7604312	381.97716
Race_American_Indian_Alaskan_Native	7.9394651	132.96757
GenHealth_Excellent	7.2776512	301.70271
GenHealth_Very_Good	5.6273062	384.66317
Race_Hispanic	5.3366421	296.68296
Race_Black	4.8108654	297.79088
Race_Asian	3.3330838	100.19364
Race_Other	3.0147184	189.18748
Diabetic_No_borderline_diabetes	2.8732792	185.05434
Diabetic_Yes_during_pregnancy	-0.7281113	39.80911

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://rdr.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?

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

3. 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

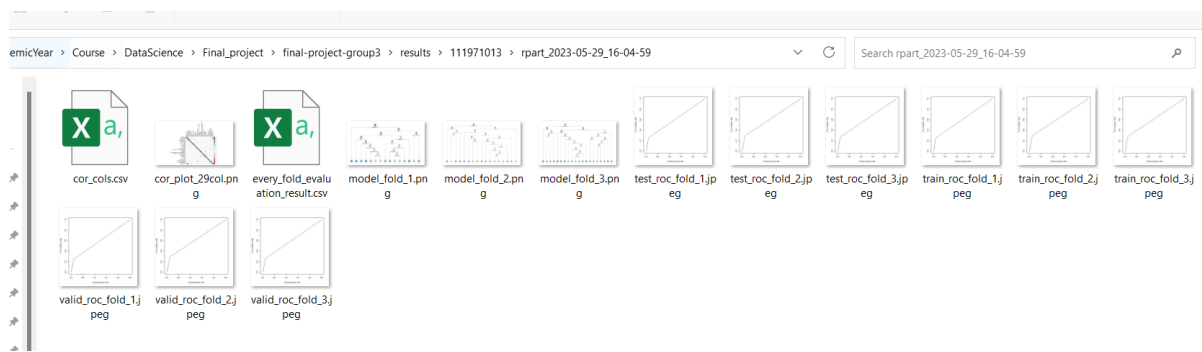
(2-2) Rscript Carter_111971013.R --input

../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)

```
Accuracy:0.82
Precision:0.27
sensitivity(Recall):0.66
Specificity:0.83
F1 Score:0.38
Setting levels: control = 0, case = 1
Setting direction: controls < cases
AUC:0.74
```

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、書暉、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. 書暉：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](#)

[信用卡流失客戶預測.ppt](#)


[我們的報告.ppt](#)

5/18

今晚討論

1. github 期末專案組別名稱：group3

06月 12日 - 06月 18日

 finalProject

 Github final projct (Click to join your group)

請透過連結加入各組 team，團隊名稱按造分組名單上的組別 group1, group2,，小組中第一個人需要先創立自己的 team，其他人則可以加入已有的 team。

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-square 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

```

> summary(ir.pca)
Importance of components:
      PC1      PC2      PC3      PC4      PC5      PC6      PC7
Standard deviation 1.7676 1.42874 1.3224 1.20312 1.16703 1.1283 1.08300
Proportion of Variance 0.1077 0.07039 0.0603 0.04991 0.04696 0.0439 0.04044
Cumulative Proportion 0.1077 0.17813 0.2384 0.28834 0.33531 0.3792 0.41965
      PC8      PC9      PC10      PC11      PC12      PC13      PC14
Standard deviation 1.06651 1.04730 1.0415 1.02274 1.01799 1.01088 1.00556
Proportion of Variance 0.03922 0.03782 0.0374 0.03607 0.03573 0.03524 0.03487
Cumulative Proportion 0.45887 0.49670 0.5341 0.57017 0.60590 0.64114 0.67601
      PC15      PC16      PC17      PC18      PC19      PC20      PC21
Standard deviation 0.98333 0.97343 0.95997 0.95500 0.9466 0.93095 0.91594
Proportion of Variance 0.03334 0.03267 0.03178 0.03145 0.0309 0.02989 0.02893
Cumulative Proportion 0.70935 0.74203 0.77380 0.80525 0.8361 0.86603 0.89496
      PC22      PC23      PC24      PC25      PC26      PC27
Standard deviation 0.86941 0.8583 0.76051 0.7225 0.67324 3.608e-14
Proportion of Variance 0.02606 0.0254 0.01994 0.0180 0.01563 0.000e+00
Cumulative Proportion 0.92103 0.9464 0.96637 0.9844 1.00000 1.000e+00
      PC28      PC29
Standard deviation 2.169e-14 1.636e-14
Proportion of Variance 0.000e+00 0.000e+00
Cumulative Proportion 1.000e+00 1.000e+00
> plot(ir.pca)

```

(2) 使用FactoMineR method

同樣得到Dim.18(=PC18)能解釋0.80525的資料。

```

> log.ir <- log1p(f_in_csv[,2:30])
> ir.species <- f_in_csv[,1]
> ir.pca <- PCA(log.ir, graph = FALSE)
> get_eig(ir.pca)

```

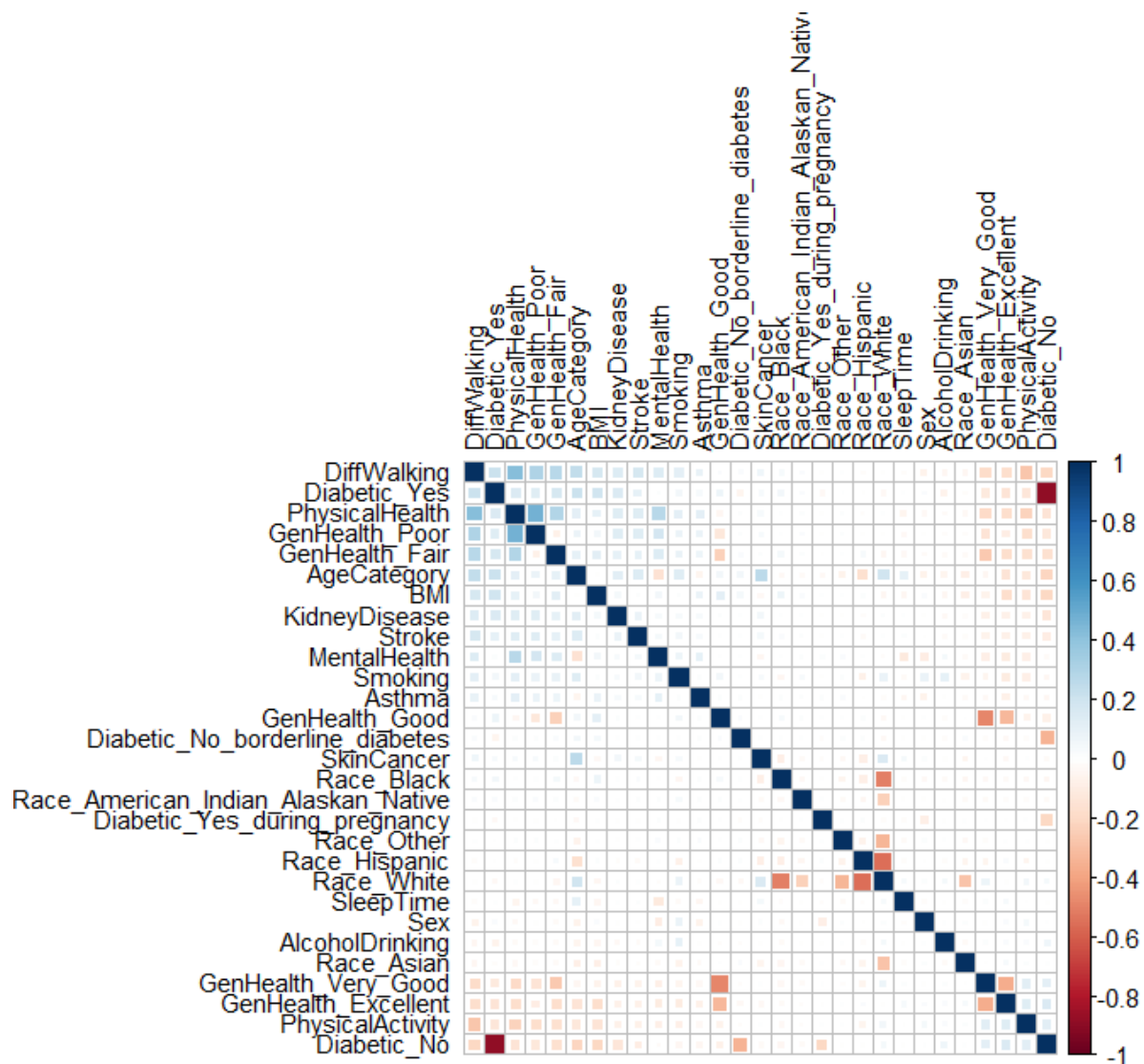
	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	3.124429e+00	1.077389e+01	10.77389
Dim.2	2.041302e+00	7.038972e+00	17.81286
Dim.3	1.748716e+00	6.030056e+00	23.84292
Dim.4	1.447496e+00	4.991366e+00	28.83429
Dim.5	1.361959e+00	4.696411e+00	33.53070
Dim.6	1.273111e+00	4.390038e+00	37.92073
Dim.7	1.172884e+00	4.044426e+00	41.96516
Dim.8	1.137454e+00	3.922255e+00	45.88742
Dim.9	1.096842e+00	3.782215e+00	49.66963
Dim.10	1.084690e+00	3.740309e+00	53.40994
Dim.11	1.046007e+00	3.606921e+00	57.01686
Dim.12	1.036307e+00	3.573473e+00	60.59033
Dim.13	1.021884e+00	3.523737e+00	64.11407
Dim.14	1.011152e+00	3.486731e+00	67.60080
Dim.15	9.669340e-01	3.334255e+00	70.93506
Dim.16	9.475727e-01	3.267492e+00	74.20255
Dim.17	9.215427e-01	3.177733e+00	77.38028
Dim.18	9.120333e-01	3.144942e+00	80.52522
Dim.19	8.959607e-01	3.089519e+00	83.61474
Dim.20	8.666751e-01	2.988535e+00	86.60328
Dim.21	8.389470e-01	2.892921e+00	89.49620
Dim.22	7.558719e-01	2.606455e+00	92.10265
Dim.23	7.366622e-01	2.540215e+00	94.64287
Dim.24	5.783772e-01	1.994404e+00	96.63727
Dim.25	5.219414e-01	1.799798e+00	98.43707
Dim.26	4.532496e-01	1.562930e+00	100.00000
Dim.27	4.171453e-23	1.438432e-22	100.00000
Dim.28	5.300077e-28	1.827613e-27	100.00000
Dim.29	3.586657e-28	1.236778e-27	100.00000

```

>

```

(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"
```

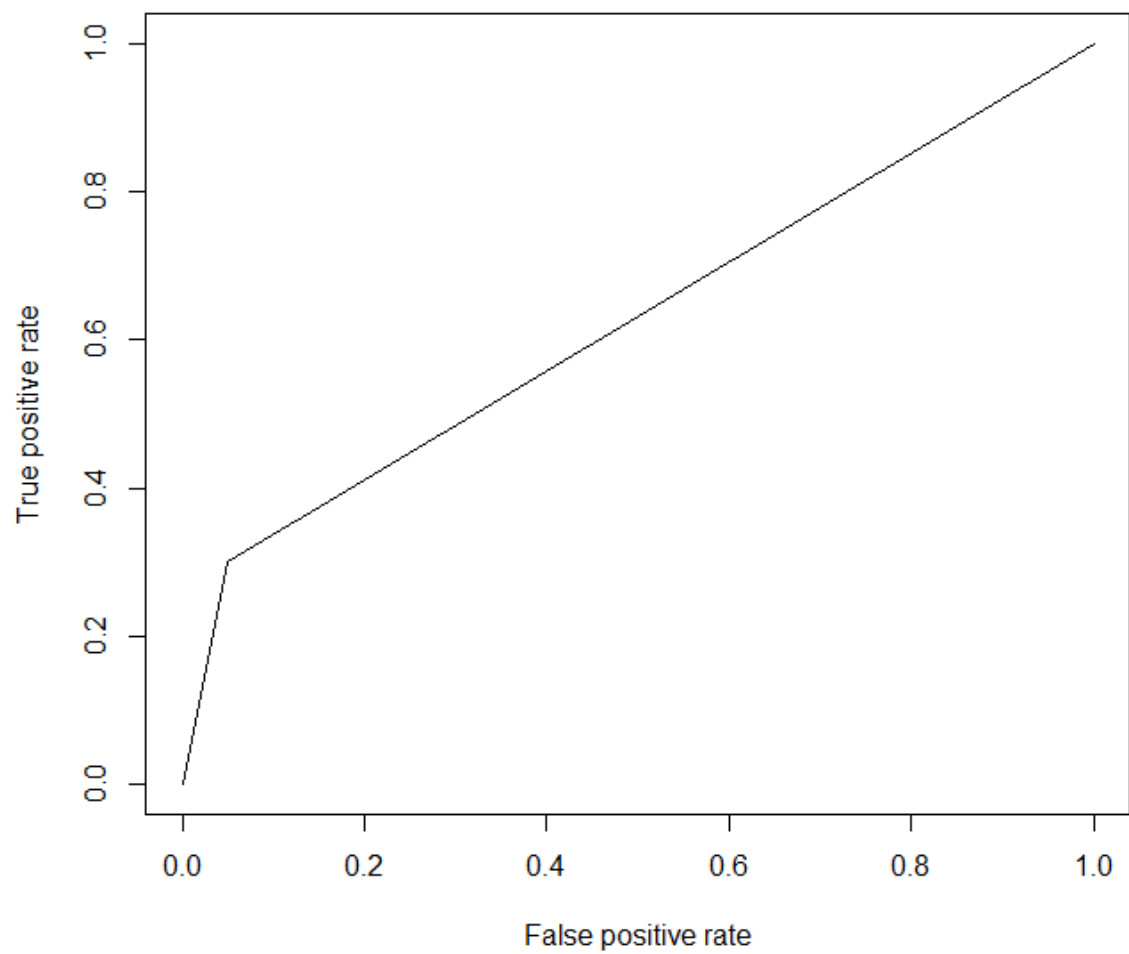
程式碼：

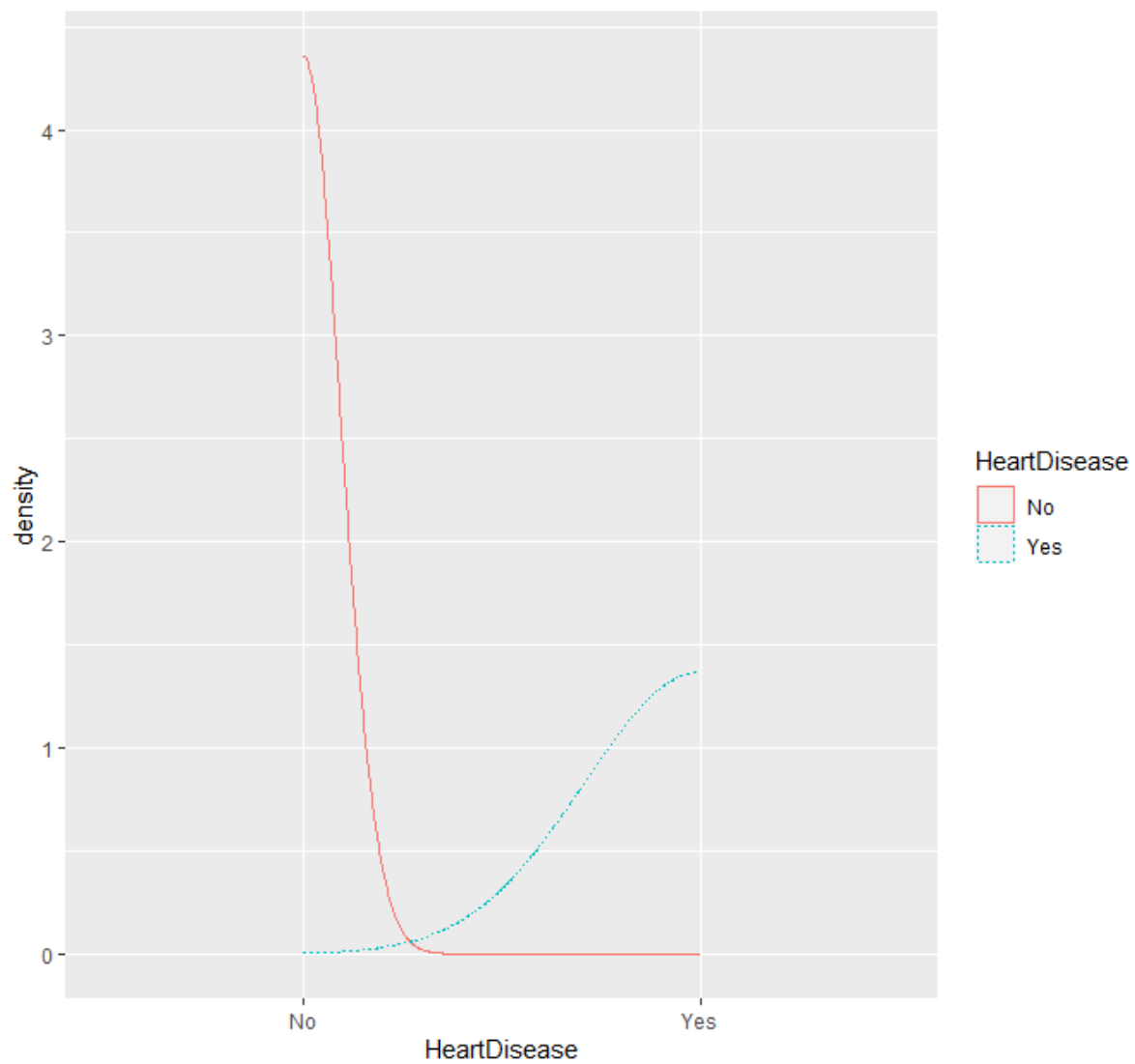
```
#df_cor_f_in_csv <- data.frame(cor(f_in_csv[, -1]))
#cor_cols <- colnames(df_cor_f_in_csv)[df_cor_f_in_csv[1]>0]
```

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





```
> print(attributes(performance(b,'auc'))$y.values[[1]])  
[1] 0.6257551
```

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

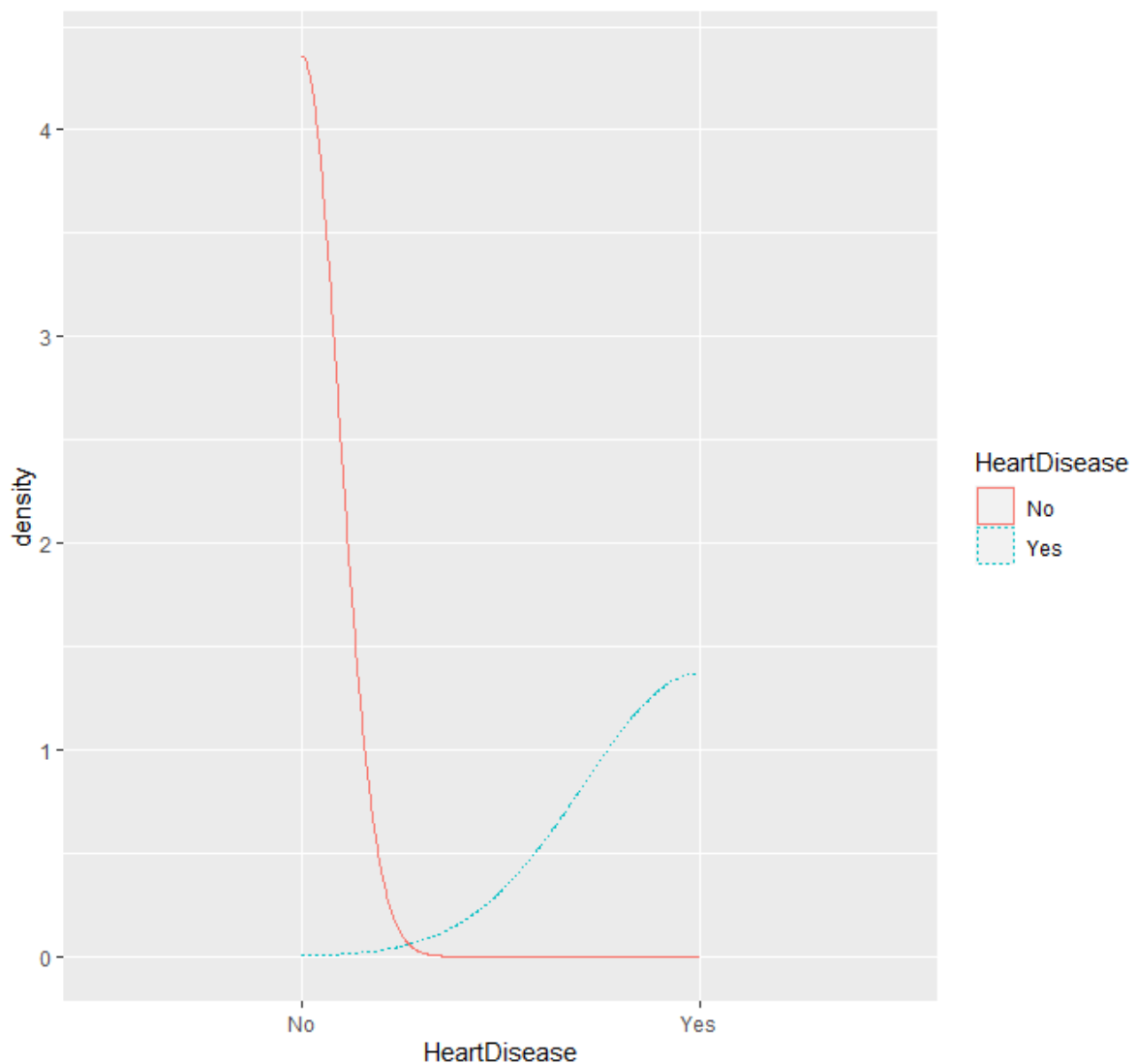
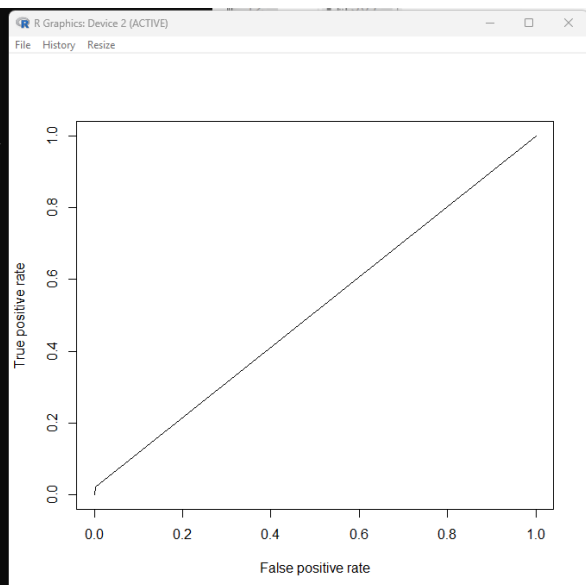

```

No 91078 408
Yes 8602 484
fold2's train_correct_percent is 0.91
fold2's test_correct_percent is 0.91
fold2's validate_correct_percent is 0.91

The 3-Fold: each fold has 100572 rows
df_test_set is from 201145 to 301716
df_validate_set is from 301717 and backward to 100571
df_train_set isn't within 201145 to 301716 and backward to 100571 , balanced

No Yes
91397 9176
[1] "-----"
pred
truth No Yes
No 91259 138
Yes 8960 216
pred
truth No Yes
No 91325 161
Yes 8894 192
pred
truth No Yes
No 91422 151
Yes 8807 192
fold3's train_correct_percent is 0.91
fold3's test_correct_percent is 0.91
fold3's validate_correct_percent is 0.91
>
> #roc
> a <- predict(model, newdata=df_validate_set, type="class")
> b <- prediction(as.numeric(a), df_validate_set$HeartDisease)
>
> ggplot(data=df_validate_set) +
+ geom_density(aes(x=HeartDisease,color=HeartDisease,linetype=HeartDisease))
+ plot(performance(b,"tpr","fpr"))
>
> #auc
> print(attributes(performance(b,"auc"))$y.values[[1]])
[1] 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_Accur	test_Accur	valid_Accur	train_Preci	test_Preci	valid_Preci	train_Sensiti	test_Sensiti	valid_Sensiti	train_Specifi	test_Specifi	valid_Specifi	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	0.6138	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_Accur	test_Accur	valid_Accur	train_Preci	test_Preci	valid_Preci	train_Sensiti	test_Sensiti	valid_Sensiti	train_Specifi	test_Specifi	valid_Specifi	train_F1	test_F1	valid_F1	train_auc	test_auc	valid_auc
-----	-------------	------------	-------------	-------------	------------	-------------	---------------	--------------	---------------	---------------	--------------	---------------	----------	---------	----------	-----------	----------	-----------

1	fold 1	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	0.6060 6	0.599 4	0.6014 6
2	fold 2	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	0.6100 6	0.604 2	0.6028 2
3	fold 3	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	0.6247 3	0.613 84	0.6124 9
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_Accur acy	test_Accur acy	valid_Accur acy	train_Preci sion	test_Preci sion	valid_Preci sion	train_Sensiti vity	test_Sensiti vity	valid_Sensiti vity	train_Specifi city	test_Specifi city	valid_Specifi city	train_ F1	test_ F1	valid_ F1	train_a uc	test_a uc	valid_a uc
1	fold 1	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	0.6184 3	0.613 42	0.61
2	fold 2	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	0.6130 4	0.606 72	0.6048 5
3	fold 3	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	0.6108 9	0.601 18	0.6033 5
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_Accur	test_Accur	valid_Accur	train_Preci	test_Preci	valid_Preci	train_Sensiti	test_Sensiti	valid_Sensiti	train_Specifi	test_Specifi	valid_Specifi	train_ test_ valid_ train_a test_a valid_a
-----	-------------	------------	-------------	-------------	------------	-------------	---------------	--------------	---------------	---------------	--------------	---------------	---

		acy	acy	acy	sion	sion	sion	vity	vity	vity	city	city	city	F1	F1	F1	uc	uc	uc
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-square: 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
```

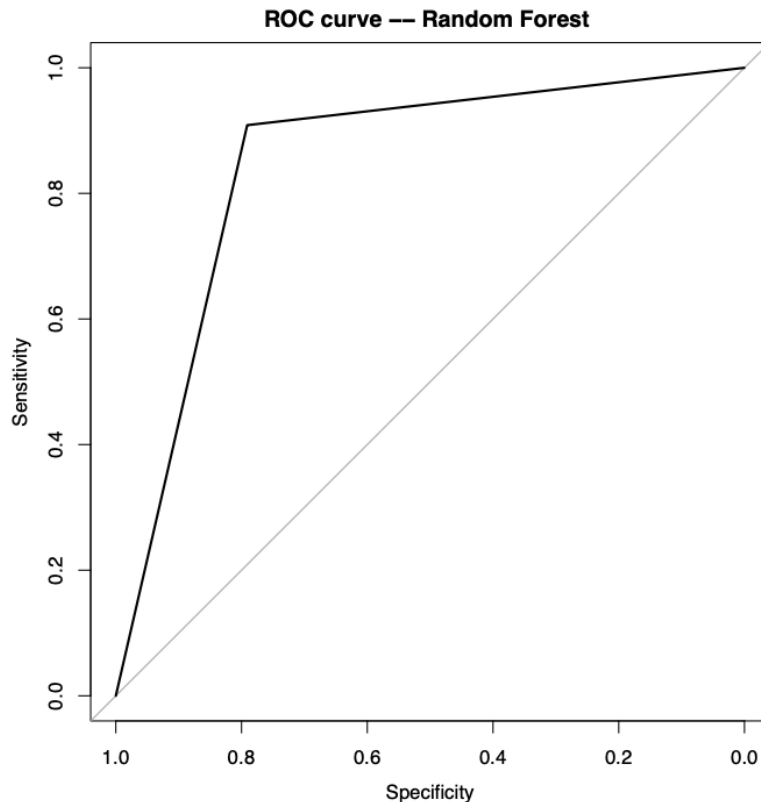
```
      0      1  
292422 27373  
[1] 319795  
  
      0      1  
292422 292422  
[1] 584844
```

使用Random Forest 產生的 important 如表

<https://docs.google.com/spreadsheets/d/1IM2-4hOuFyP3FJt-Ylv7QdDRmdFKmODZmsDmDpPo974/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),

```
[1] "sensitivity is 0.049523110785033"  
[1] "specificity is 0.99555490781899"  
[1] "precision is 0.525291828793774"  
[1] "F1 is 0.0905129064699966"
```

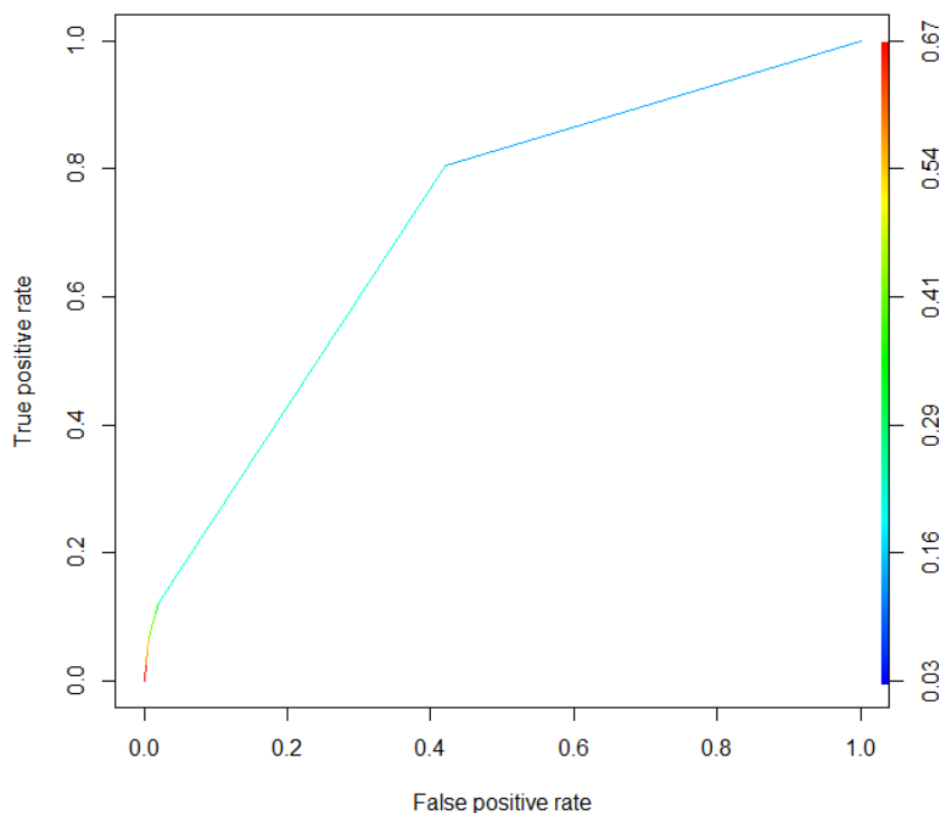
(3)尚待完成

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)

Before smote:

No	Yes
292422	27373

After smote:

No	Yes
292422	82119

資料建模：隨機森林

混淆矩陣：

```
Confusion Matrix:
      val_pred
      0      1
0  86994  732
1  10336 14299
```

Result:

```
Accuracy:0.9
Precision:0.95
Recall:0.58
F1 Score:0.72
```

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:

```
Confusion Matrix:
      val_pred
      0      1
0  87593  133
1   7995  216
Accuracy:0.92
Precision:0.62
Recall:0.03
F1 Score:0.05
```

書瑋提供 - 變數間的相關性(使用R Cor()):

<https://docs.google.com/spreadsheets/d/1IM2-4hOuFyP3FJt-Ylv7QdDRmdFKmODZmsDmDpPo974/edit#gid=0>

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-square 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/1NKB5WjAGIUhrjHC9ABC6dFSyDZwdGpSI?usp=share_linkhttps://docs.google.com/spreadsheets/d/1IM2-4hOuFyP3FJt-Ylv7QdDRmdFKmODZmsDmDpPo974/edit#gid=0

5. 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

```

> summary(duplicate)
  Mode   FALSE   TRUE
logical 301717 18078
> head(f_in_csv[duplicate>0,])
  HeartDisease BMI Smoking AlcoholDrinking Stroke PhysicalHealth
112226      No 23.49      No              No      No             0
138352      No 30.13      No              No      No             0
291748      No 28.89      No              No      No             0
171901      No 25.68     Yes              No      No             0
255300      No 24.03      No              No      No             0
237263      No 24.41      No              No      No             0
  MentalHealth DiffWalking Sex AgeCategory Race Diabetic
112226        0           No Female      50-54 White      No
138352        0           No  Male      50-54 White      No
291748        0           No  Male      55-59 White      No
171901        0           No  Male      35-39 White      No
255300        0           No Female      60-64 White      No
237263        0           No  Male      18-24 White      No
  PhysicalActivity GenHealth SleepTime Asthma KidneyDisease SkinCancer
112226      Yes Excellent         7      No              No          No
138352      Yes Excellent         7      No              No          No
291748      Yes Very good         8      No              No          No
171901      Yes      Good         7      No              No          No
255300      Yes Excellent         8      No              No          No
237263      Yes Very good         8      No              No          No
>

```

```
> 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

```

> head(dummy_cols(f_in_csv, select_columns = c('Race', 'Diabetic', 'GenHealth'$
HeartDisease BMI Smoking AlcoholDrinking Stroke PhysicalHealth MentalHealth
1 No 16.60 Yes No No 3 30
2 No 20.34 No No Yes 0 0
3 No 26.58 Yes No No 20 30
4 No 24.21 No No No 0 0
5 No 23.71 No No No 28 0
6 Yes 28.87 Yes No No 6 0
DiffWalking Sex AgeCategory Race Diabetic PhysicalActivity GenHealth
1 No Female 57 White Yes Yes Very good
2 No Female 80 White No Yes Very good
3 No Male 67 White Yes Yes Fair
4 No Female 77 White No No Good
5 Yes Female 42 White No Yes Very good
6 Yes Female 77 Black No No Fair
SleepTime Asthma KidneyDisease SkinCancer Race_American Indian/Alaskan Native
1 5 Yes No Yes 0
2 7 No No No 0
3 8 Yes No No 0
4 6 No No Yes 0
5 8 No No No 0
6 12 No No No 0
Race_Asian Race_Black Race_Hispanic Race_Other Race_White Diabetic_No
1 0 0 0 0 1 0
2 0 0 0 0 1 1
3 0 0 0 0 1 0
4 0 0 0 0 1 1
5 0 0 0 0 1 1
6 0 1 0 0 0 1
Diabetic_No, borderline diabetes Diabetic_Yes Diabetic_Yes (during pregnancy)
1 0 1 0
2 0 0 0
3 0 1 0
4 0 0 0
5 0 0 0
6 0 0 0
GenHealth_Excellent GenHealth_Fair GenHealth_Good GenHealth_Poor
1 0 0 0 0
2 0 0 0 0
3 0 1 0 0
4 0 0 1 0
5 0 0 0 0
6 0 1 0 0
GenHealth_Very good
1 1
2 1
3 0
4 0
5 1
6 0

```

```

> head(f_in_csv_dummy)
HeartDisease BMI Smoking AlcoholDrinking Stroke PhysicalHealth MentalHealth
1 No 16.60 1 0 0 3 30
2 No 20.34 0 0 1 0 0
3 No 26.58 1 0 0 20 30
4 No 24.21 0 0 0 0 0
5 No 23.71 0 0 0 28 0
6 Yes 28.87 1 0 0 6 0
DiffWalking Sex AgeCategory PhysicalActivity SleepTime Asthma KidneyDisease
1 0 0 57 1 5 1 0
2 0 0 80 1 7 0 0
3 0 1 67 1 8 1 0
4 0 0 77 0 6 0 0
5 1 0 42 1 8 0 0
6 1 0 77 0 12 0 0
SkinCancer Race_American Indian/Alaskan Native Race_Asian Race_Black
1 1 0 0 0
2 0 0 0 0
3 0 0 0 0
4 1 0 0 0
5 0 0 0 0
6 0 0 0 1
Race_Hispanic Race_Other Race_White Diabetic_No
1 0 0 1 0
2 0 0 1 1
3 0 0 1 0
4 0 0 1 1
5 0 0 1 1
6 0 0 0 1
Diabetic_No, borderline diabetes Diabetic_Yes Diabetic_Yes (during pregnancy)
1 0 1 0
2 0 0 0
3 0 1 0
4 0 0 0
5 0 0 0
6 0 0 0
GenHealth_Excellent GenHealth_Fair GenHealth_Good GenHealth_Poor
1 0 0 0 0
2 0 0 0 0
3 0 1 0 0
4 0 0 1 0
5 0 0 0 0
6 0 1 0 0
GenHealth_Very good
1 1
2 1
3 0
4 0
5 1
6 0
> ncol(f_in_csv_dummy)
[1] 30
> ncol(f_in_csv)
[1] 18

```

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))
```



```

The 1-Fold: each fold has 100572 rows
df_test_set is from 1 to 100572
df_validate_set is from 100573 to 201144
df_train_set isn't within 1 to 100572 and 100573 to 201144 , balanced num is 63133

No    Yes
36076 9019
[1] "-----"
      pred
truth  No   Yes
No    34850 1226
Yes   6885  2134
      pred
truth  No   Yes
No    88105 3306
Yes   7022  2139
      pred
truth  No   Yes
No    88193 3298
Yes   6993  2088
fold1's train_correct_percent is 0.59
fold1's test_correct_percent is 0.90
fold1's validate_correct_percent is 0.90

The 2-Fold: each fold has 100572 rows
df_test_set is from 100573 to 201144
df_validate_set is from 201145 to 301716
df_train_set isn't within 100573 to 201144 and 201145 to 301716 , balanced num is 64127

No    Yes
36644 9161
[1] "-----"
      pred
truth  No   Yes
No    34897 1747
Yes   6332  2829
      pred
truth  No   Yes
No    86881 4610
Yes   6413  2668
      pred
truth  No   Yes
No    87107 4446
Yes   6373  2646
fold2's train_correct_percent is 0.59
fold2's test_correct_percent is 0.89
fold2's validate_correct_percent is 0.89

```


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
```

```

3 # 【1】 18 cols
4 # [1] "HeartDisease"      "BMI"      "Smoking"      "AlcoholDrinking"
5 # [5] "Stroke"            "PhysicalHealth" "MentalHealth" "DiffWalking"
6 # [9] "Sex"              "AgeCategory" "Race"          "Diabetic"
7 # [13] "PhysicalActivity" "GenHealth"  "SleepTime"     "Asthma"
8 # [17] "KidneyDisease"    "SkinCancer"
9
10 # 【1-1】 no na value
11 # 【2】 head(csv)
12 # HeartDisease BMI Smoking AlcoholDrinking Stroke PhysicalHealth MentalHealth
13 #1 No 16.60 Yes No No 3 30
14 #2 No 20.34 No No Yes 0 0
15 #3 No 26.58 Yes No No 20 30
16 #4 No 24.21 No No No 0 0
17 #5 No 23.71 No No No 28 0
18 #6 Yes 28.87 Yes No No 6 0
19 # DiffWalking Sex AgeCategory Race Diabetic PhysicalActivity GenHealth
20 #1 No Female 55-59 White Yes Yes Very good
21 #2 No Female 80 or older White No Yes Very good
22 #3 No Male 65-69 White Yes Yes Fair
23 #4 No Female 75-79 White No No Good
24 #5 Yes Female 40-44 White No Yes Very good
25 #6 Yes Female 75-79 Black No No Fair
26 # SleepTime Asthma KidneyDisease SkinCancer
27 #1 5 Yes No Yes
28 #2 7 No No No
29 #3 8 Yes No No
30 #4 6 No No Yes
31 #5 8 No No No
32 #6 12 No No No
33
34 # 【3】 summary
35 #HeartDisease BMI Smoking AlcoholDrinking
36 #Length:319795 Min. :12.02 Length:319795 Length:319795
37 #Class :character 1st Qu.:24.03 Class :character Class :character
38 #Mode :character Median :27.34 Mode :character Mode :character
39 # Mean :28.33
40 # 3rd Qu.:31.42
41 # Max. :94.85
42 # Stroke PhysicalHealth MentalHealth DiffWalking
43 #Length:319795 Min. : 0.000 Min. : 0.000 Length:319795
44 #Class :character 1st Qu.: 0.000 1st Qu.: 0.000 Class :character
45 #Mode :character Median : 0.000 Median : 0.000 Mode :character
46 # Mean : 3.372 Mean : 3.898
47 # 3rd Qu.: 2.000 3rd Qu.: 3.000
48 # Max. :30.000 Max. :30.000
49 # Sex AgeCategory Race Diabetic
50 #Length:319795 Length:319795 Length:319795 Length:319795
51 #Class :character Class :character Class :character Class :character
52 #Mode :character Mode :character Mode :character Mode :character
53 #
54 #
55 #
56 #PhysicalActivity GenHealth SleepTime Asthma
57 #Length:319795 Length:319795 Min. : 1.000 Length:319795
58 #Class :character Class :character 1st Qu.: 6.000 Class :character
59 #Mode :character Mode :character Median : 7.000 Mode :character
60 # Mean : 7.097
61 # 3rd Qu.: 8.000
62 # Max. :24.000
63 #KidneyDisease SkinCancer
64 #Length:319795 Length:319795
65 #Class :character Class :character
66 #Mode :character Mode :character
67

```

```

69 # [4] no col if its data is numeric
70 # [1] "col_name : HeartDisease"
71 # [1] "No" "Yes"
72 # [1] "col_name : Smoking"
73 # [1] "Yes" "No"
74 # [1] "col_name : AlcoholDrinking"
75 # [1] "No" "Yes"
76 # [1] "col_name : Stroke"
77 # [1] "No" "Yes"
78 # [1] "col_name : DiffWalking"
79 # [1] "No" "Yes"
80 # [1] "col_name : Sex"
81 # [1] "Female" "Male"
82 # [1] "col_name : AgeCategory"
83 # [1] "55-59" "80 or older" "65-69" "75-79" "40-44"
84 # [6] "70-74" "60-64" "50-54" "45-49" "18-24"
85 # [11] "35-39" "30-34" "25-29"
86 # [1] "col_name : Race"
87 # [1] "White" "Black"
88 # [3] "Asian" "American Indian/Alaskan Native"
89 # [5] "Other" "Hispanic"
90 # [1] "col_name : Diabetic"
91 # [1] "Yes" "No"
92 # [3] "No, borderline diabetes" "Yes (during pregnancy)"
93 # [1] "col_name : PhysicalActivity"
94 # [1] "Yes" "No"
95 # [1] "col_name : GenHealth"
96 # [1] "Very good" "Fair" "Good" "Poor" "Excellent"
97 # [1] "col_name : SleepTime"
98 # [1] 5 7 8 6 12 4 9 10 15 3 2 1 16 18 14 20 11 13 17 24 19 21 22 23
99 # [1] "col_name : Asthma"
100 # [1] "Yes" "No"
101 # [1] "col_name : KidneyDisease"
102 # [1] "No" "Yes"
103 # [1] "col_name : SkinCancer"
104 # [1] "Yes" "No"

```

based on hw3's rpart and K-fold is 5:

```

1 rpart result
2 The 1-Fold: each fold has 63959 rows
3 df_test_set is from 1 to 63959
4 df_validate_set is from 63960 to 127918
5 df_train_set isn't within 1 to 63959 and 63960 to 127918
6   pred
7 truth   No   Yes
8   No 175133     0
9   Yes 16744     0
10  pred
11 truth   No   Yes
12   No 58687     0
13   Yes 5272     0
14  pred
15 truth   No   Yes
16   No 58602     0
17   Yes 5357     0
18 fold1's CP_train_correct_percent is 0.91
19 fold1's CP_test_correct_percent is 0.92
20 fold1's CP_validate_correct_percent is 0.92
21
22 The 2-Fold: each fold has 63959 rows
23 df_test_set is from 63960 to 127918
24 df_validate_set is from 127919 to 191877
25 df_train_set isn't within 63960 to 127918 and 127919 to 191877
26   pred
27 truth   No   Yes
28   No 175363     0
29   Yes 16514     0
30  pred
31 truth   No   Yes
32   No 58602     0
33   Yes 5357     0
34  pred
35 truth   No   Yes
36   No 58457     0
37   Yes 5502     0
38 fold2's CP_train_correct_percent is 0.91
39 fold2's CP_test_correct_percent is 0.92
40 fold2's CP_validate_correct_percent is 0.91
41

```

```

42 The 3-Fold: each fold has 63959 rows
43 df_test_set is from 127919 to 191877
44 df_validate_set is from 191878 to 255836
45 df_train_set isn't within 127919 to 191877 and 191878 to
46   pred
47 truth   No   Yes
48   No 175607     0
49   Yes 16270     0
50  pred
51 truth   No   Yes
52   No 58457     0
53   Yes 5502     0
54  pred
55 truth   No   Yes
56   No 58358     0
57   Yes 5601     0
58 fold3's CP_train_correct_percent is 0.92
59 fold3's CP_test_correct_percent is 0.91
60 fold3's CP_validate_correct_percent is 0.91
61
62 The 4-Fold: each fold has 63959 rows
63 df_test_set is from 191878 to 255836
64 df_validate_set is from 255837 to 319795
65 df_train_set isn't within 191878 to 255836 and 255837 to 319795
66   pred
67 truth   No   Yes
68   No 175746     0
69   Yes 16131     0
70  pred
71 truth   No   Yes
72   No 58358     0
73   Yes 5601     0
74  pred
75 truth   No   Yes
76   No 58318     0
77   Yes 5641     0
78 fold4's CP_train_correct_percent is 0.92
79 fold4's CP_test_correct_percent is 0.91
80 fold4's CP_validate_correct_percent is 0.91
81

```

```

82 The 5-Fold: each fold has 63959 rows
83 df_test_set is from 255837 to 319795
84 df_validate_set is from 1 to 63959
85 df_train_set isn't within 255837 to 319795 and 1 to 63959
86   pred
87 truth   No   Yes
88   No 175417     0
89   Yes 16460     0
90  pred
91 truth   No   Yes
92   No 58318     0
93   Yes 5641     0
94  pred
95 truth   No   Yes
96   No 58687     0
97   Yes 5272     0
98 fold5's CP_train_correct_percent is 0.91
99 fold5's CP_test_correct_percent is 0.91
100 fold5's CP_validate_correct_percent is 0.92

```

0. 選組長。
5~6人

人数 2 optional+1 CARTER BUSKY JUDE *正晏	人数 1 optional+1 書瑋	人数 1 正晏 AUSTIN	人数 1	人数 optional+1
---	---------------------------------	-----------------------------	---------	------------------

<p>2.2. data(input):</p> <p>a. 資料前處理:有無發現資料清洗的難處？</p> <p>2.3. model:</p> <p>a. 選哪個模型進行ML？例如決策數、XGBoost。</p> <p>b. 選用哪個特徵工程來優化模型？</p> <p>c. null model定義與建立。</p>	<p>2.4. evaluation(ouput):</p> <p>a. 用哪個方法來評估模型好壞？</p> <p>例如Hypothesis Test(包含Confusion matrix、precesion and recall)、R-squre</p>	<p>2.5. present:</p> <p>a. 從第2~4階段選擇資料視覺作用之處：</p> <p>例如</p> <ul style="list-style-type: none">- 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。 <p>可能得遵照老師的建議，一併做on-line visualization。</p>	<p>2.6. deploy:除了製作還原專案的步驟，在報告當下demo。</p>	<p>統整ReadMe.md</p>
---	---	---	--	---------------------------

- 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。

1	HeartDise	BMI	Smoking	AlcoholDri	Stroke	PhysicalHe	MentalHe	Hei Diff	Walkin	Sex	AgeCatego	Race	Diabetic	PhysicalAc	GenHealth	SleepTime	Asthma	KidneyDise	SkinC
2	No	16.6	Yes	No	No	3	30	No	Female	55-59	White	Yes	Yes	Very good	5	Yes	No	Yes	
3	No	20.34	No	No	Yes	0	0	No	Female	80 or older	White	No	Yes	Very good	7	No	No	No	
4	No	26.58	Yes	No	No	20	30	No	Male	65-69	White	Yes	Yes	Fair	8	Yes	No	No	
5	No	24.21	No	No	No	0	0	No	Female	75-79	White	No	No	Good	6	No	No	Yes	
6	No	23.71	No	No	No	28	0	Yes	Female	40-44	White	No	Yes	Very good	8	No	No	No	
7	Yes	28.87	Yes	No	No	6	0	Yes	Female	75-79	Black	No	No	Fair	12	No	No	No	
8	No	21.63	No	No	No	15	0	No	Female	70-74	White	No	Yes	Fair	4	Yes	No	Yes	
9	No	31.64	Yes	No	No	5	0	Yes	Female	80 or older	White	Yes	No	Good	9	Yes	No	No	
10	No	26.45	No	No	No	0	0	No	Female	80 or older	White	No, border	No	Fair	5	No	Yes	No	
11	No	40.69	No	No	No	0	0	Yes	Male	65-69	White	No	Yes	Good	10	No	No	No	
12	Yes	34.3	Yes	No	No	30	0	Yes	Male	60-64	White	Yes	No	Poor	15	Yes	No	No	
13	No	28.71	Yes	No	No	0	0	No	Female	55-59	White	No	Yes	Very good	5	No	No	No	
14	No	28.37	Yes	No	No	0	0	Yes	Male	75-79	White	Yes	Yes	Very good	8	No	No	No	
15	No	28.15	No	No	No	7	0	Yes	Female	80 or older	White	No	No	Good	7	No	No	No	
16	No	29.29	Yes	No	No	0	30	Yes	Female	60-64	White	No	No	Good	5	No	No	No	
17	No	29.18	No	No	No	1	0	No	Female	50-54	White	No	Yes	Very good	6	No	No	No	
18	No	26.26	No	No	No	5	2	No	Female	70-74	White	No	No	Very good	10	No	No	No	
19	No	22.59	Yes	No	No	0	30	Yes	Male	70-74	White	No, border	Yes	Good	8	No	No	No	

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-square

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。

15min+5min Q&A

finalproject_group1 、

all report putted in github is ok; **no more ppt to deliver**

- goal
- input
 - source 、 preprocess
- modeling
 - method 、 null model for comparison 、
- output
 - performance: precision 、 recall 、 R-square ; improvement is significant
- demo
 - on-line visualization 、 **reproduce** your result by other team
 - challenges in project
- references
 - indicate in your presentation if you use code for others.
 - package used
 - related publications