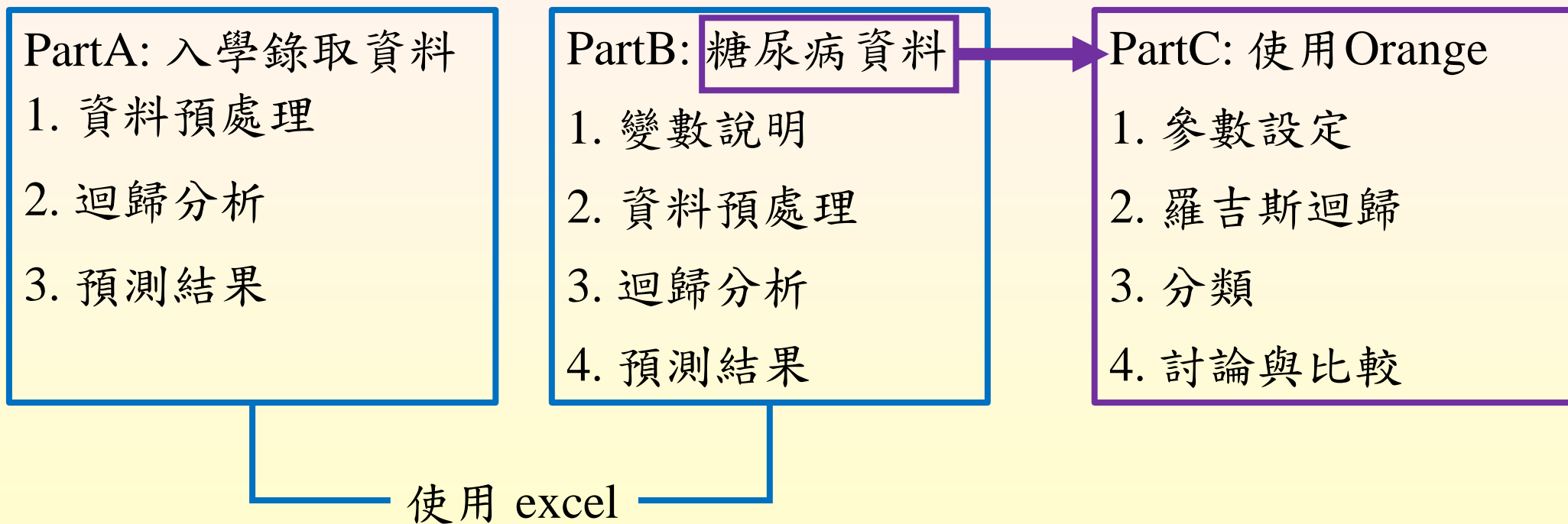


羅吉斯回歸與分類

—以入學錄取與糖尿病資料集為例

大綱



PartA 1.1 檔案轉換

入學錄取資料

B6-1 入學錄取資料.txt				
檔案 編輯 檢視				
錄取	GRE 分數	GPA 分數	畢業學校排名	
0	380	3.61	3	
1	660	3.67	3	
1	800	4	1	
1	640	3.19	4	
0	520	2.93	4	
1	760	3	2	
1	560	2.98	1	
0	400	3.08	2	
1	540	3.39	3	
0	700	3.92	2	

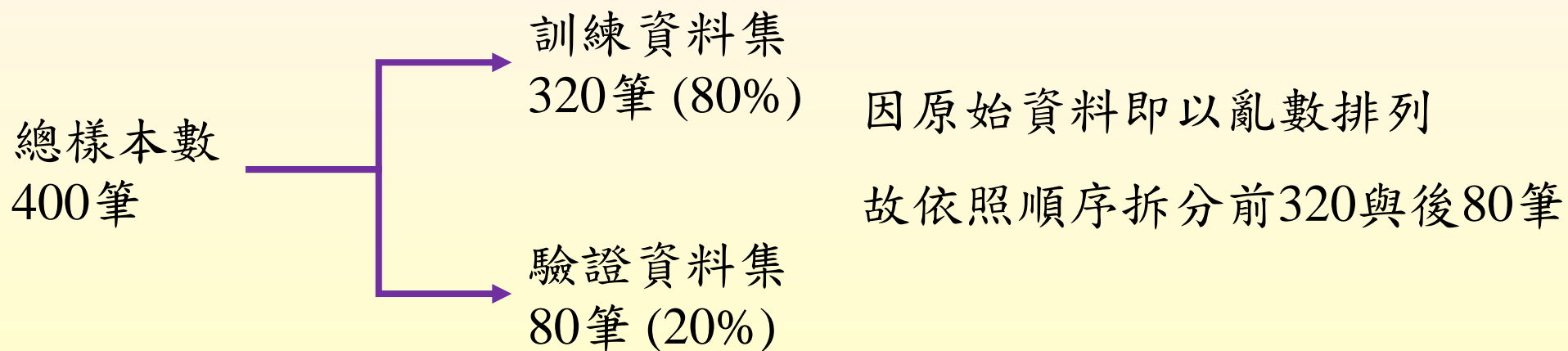
用excel開啟txt.

以空白分隔

	A	B	C	D
1	錄取	GRE 分數	GPA 分數	畢業學校排名
2	0	380	3.61	3
3	1	660	3.67	3
4	1	800	4	1
5	1	640	3.19	4
6	0	520	2.93	4
7	1	760	3	2
8	1	560	2.98	1
9	0	400	3.08	2
10	1	540	3.39	3

PartA 1.2 拆分資料集

入學錄取資料



PartA 2.1 迴歸試算一

excel (線性) 迴歸試算如下

$$\begin{aligned} y = & -0.21 \\ & + 0.0004 \text{ GRE分數} \\ & + 0.17 \text{ GPA分數} \\ & - 0.11 \text{ 畢業學校排名} \end{aligned}$$

截距與GRE分數變項

未通過95%顯著性檢定

迴歸統計	
R 的倍數	0.32
R 平方	0.11
調整的 R 平方	0.10
標準誤	0.44
觀察值個數	320

ANOVA

	自由度	SS	MS	F	顯著值
迴歸	3	7.29	2.43	12.42	< 0.001
殘差	316	61.83	0.20		
總和	319	69.12			

	係數	標準誤	t 統計	P-值	下限 95%	上限 95%
截距	-0.21	0.24	-0.86	0.39	-0.69	0.27
GRE 分數	0.0004	0.00	1.89	0.06	0.00	0.00
GPA 分數	0.17	0.07	2.29	0.02	0.02	0.31
畢業學校排名	-0.11	0.03	-4.38	0.00	-0.16	-0.06

PartA 2.2 迴歸試算二

扣除GRE分數變項後試算如下

$$y = -0.13 + 0.22 \text{ GPA分數} - 0.12 \text{ 畢業學校排名}$$

僅截距未通過95%顯著性檢定

迴歸統計	
R 的倍數	0.31
R 平方	0.10
調整的 R 平方	0.09
標準誤	0.44
觀察值個數	320

ANOVA

	自由度	SS	MS	F	顯著值
迴歸	2	6.59	3.30	16.71	< 0.001
殘差	317	62.53	0.20		
總和	319	69.12			

	係數	標準誤	t 統計	P-值	下限 95%	上限 95%
截距	-0.13	0.24	-0.53	0.59	-0.60	0.34
GPA 分數	0.22	0.07	3.29	0.001	0.09	0.35
畢業學校排名	-0.12	0.03	-4.58	< 0.001	-0.17	-0.07

PartA 2.3 迴歸試算整理

由於屆時可設定決定是否錄取的y臨界值，因此截距是否通過顯著檢定可忽略
預測將由以下兩模型分別討論結果：

試算一：加入所有變項

$$\begin{aligned} y = & -0.21 \\ & + 0.0004 \text{ GRE分數} \\ & + 0.17 \text{ GPA分數} \\ & - 0.11 \text{ 畢業學校排名} \end{aligned}$$

試算二：扣除 GRE 分數

$$\begin{aligned} y = & -0.13 \\ & + 0.22 \text{ GPA分數} \\ & - 0.12 \text{ 畢業學校排名} \end{aligned}$$

PartA 3.1 預測結果評估指標

混淆矩陣	預測 錄取	預測 不錄取
實際 錄取	TP	FP
實際 不錄取	FN	TN

$$\text{準確率} = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy



實際不錄取 (54) 較錄取 (36) 偏多，
準確率可能無法準確反映預測情形


$$\text{精確度} = \frac{TP}{TP + FP}$$


Precision



若預測正確但實際錯誤，此數值低
此失誤的代價僅為教育資源消耗

PartA 3.1 預測結果評估指標

召回率
$$\text{Recall} = \frac{TP}{TP + FN}$$
  若實際正確但預測錯誤，此數值低
此失誤可能造成錯失人才

F1 分數
$$= \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}}$$
  同時考慮精確度與召回率

由於錯失人才的代價巨大，因此本題預測指標將採用 **召回率 (Recall)**

由於 y 臨界值設定越低召回率勢必越高，因此統一設為 **0.5**

PartA 3.2 兩種試算預測結果

試算一：加入所有變項

		預測		
		不錄取	錄取	小計
實際	不錄取	49	5	54
	錄取	20	6	26
	小計	69	11	80
y 臨界	0.5			
準確率	Accuracy	0.69		
精確度	Precision	0.55		
召回率	Recall	0.23		
F1 分數	F1 score	0.32		

試算二：扣除 GRE 分數

		預測		
		不錄取	錄取	小計
實際	不錄取	50	4	54
	錄取	21	5	26
	小計	71	9	80
y 臨界	0.5			
準確率	Accuracy	0.69		
精確度	Precision	0.56		
召回率	Recall	0.19		
F1 分數	F1 score	0.29		

PartA 3.3 降低 y 臨界值

試算一：加入所有變項

		預測		
		不錄取	錄取	小計
實際	不錄取	21	33	54
	錄取	5	21	26
	小計	26	54	80
y 臨界	0.3			
準確率	Accuracy	0.53		
精確度	Precision	0.39		
召回率	Recall	0.81		
F1 分數	F1 score	0.53		

試算二：扣除 GRE 分數

		預測		
		不錄取	錄取	小計
實際	不錄取	21	33	54
	錄取	6	20	26
	小計	27	53	80
y 臨界	0.3			
準確率	Accuracy	0.51		
精確度	Precision	0.38		
召回率	Recall	0.77		
F1 分數	F1 score	0.51		

PartA 3.4 結果討論

- 兩種試算結果相似，各種指標差異皆小於 0.04
- 機率臨界值由 0.5 降至 0.3 有以下變化
 - ✓ 召回率顯著提升 0.58
 - ✓ 但準確率、精確度降低 0.16 ~ 0.18
 - ✓ 經過測試，再將臨界值降至 0.2，準確率將低於 0.5
- 召回率 (留住更多人才) 與精確度 (節省教育開支) 必須權衡
- 將 y 臨界值設為 **0.3** 似乎為佳，但樣本數有過少之嫌

PartB 1. 糖尿病資料集變數說明

身體質量指數(BMI)：身高(公尺) / 體重(公斤)²

糖化血色素：血中葡萄糖和紅血球中血色素結合的比例

(資料來源：亞州大學附屬醫院)

吸菸史	no info (無資訊) never (從不) not current former current (目前有) ever	資料集並未給出明確定義，kaggle討論區亦眾說紛紜，無法排序
-----	---	---------------------------------

→ 保留各頻率之英文，並開設虛擬變項以 no info 為基準，開設五個吸菸史相關變數

PartB 2.1 吸菸史資料轉換

吸菸史	虛擬變項轉換				
no info	吸菸史 current	吸菸史 ever	吸菸史 former	吸菸史 not current	吸菸史 never
no info	0	0	0	0	0
never	0	0	0	0	1
not current	0	0	0	1	0
former	0	0	1	0	0
current	1	0	0	0	0
ever	0	1	0	0	0

參考 商管實務的資料分析 解釋型迴歸分析 類別自變數 (孔令傑副教授)

<https://youtu.be/0IaZsp025pY?feature=shared>

PartB 2.2 性別資料轉換

事實上有18筆資料性別為other，但考慮樣本數100,000，18筆影響甚小，因此設為女生

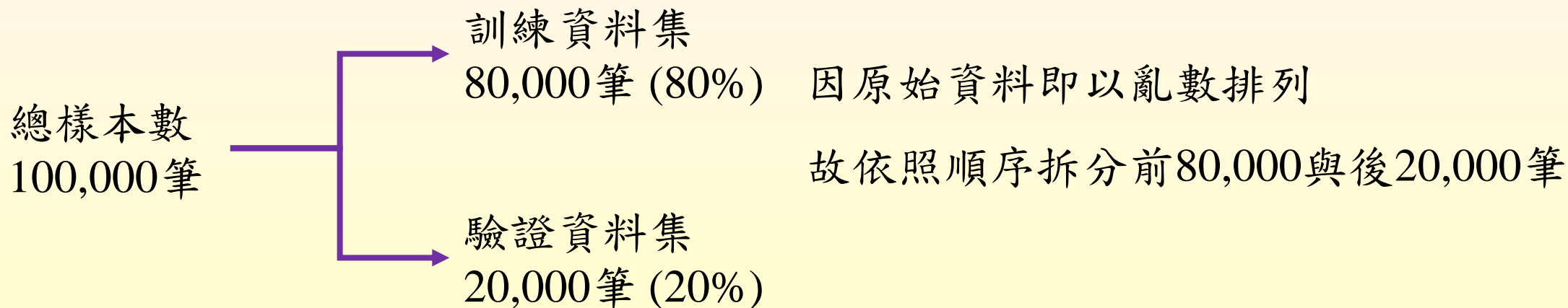
性別
女
女
男
女
男
女

虛擬變項轉換

性別
0 女
1 男
0
0
1
0
1
0

PartB 2.3 拆分資料集

糖尿病資料



PartB 3.1 線性回歸報表

迴歸統計					
R 的倍數	0.59				
R 平方	0.35				
調整的 R 平方	0.35				
標準誤	0.23				
觀察值個數	80000				
ANOVA					
	自由度	SS	MS	F	顯著值
迴歸	12	2181.85	181.821	3582.04	0
殘差	79987	4060.06	0.05076		
總和	79999	6241.91			

	係數	標準誤	t 統計	P-值	下限 95%	上限 95%
截距	-0.87	0.0055	-157.69	< 0.001	-0.88	-0.86
性別	0.01	0.0016	7.72	< 0.001	0.01	0.02
年齡	0.001	0.0000	32.27	< 0.001	0.00	0.00
高血壓	0.09	0.0032	29.68	< 0.001	0.09	0.10
心臟病	0.12	0.0043	28.08	< 0.001	0.11	0.13
吸菸史 current	0.01	0.0030	3.30	< 0.001	0.00	0.02
吸菸史 ever	0.01	0.0043	3.09	< 0.001	0.00	0.02
吸菸史 former	0.03	0.0031	8.76	< 0.001	0.02	0.03
吸菸史 not current	0.01	0.0035	3.29	< 0.001	0.00	0.02
吸菸史 never	0.01	0.0020	5.43	< 0.001	0.01	0.01
BMI	0.004	0.0001	32.20	< 0.001	0.00	0.00
糖化血色素	0.08	0.0008	107.56	< 0.001	0.08	0.08
血糖濃度	0.002	0.0000	113.31	< 0.001	0.00	0.00

各斜率與截距皆通過95%顯著性檢定

PartB 3.2 迴歸模型函數

$$\begin{aligned} y = & -0.87 + 0.01 \text{ 性別 (0 女; 1 男)} + 0.001 \text{ 年齡 (歲)} \\ & + 0.09 \text{ 高血壓 (是 1; 否 0)} + 0.12 \text{ 心臟病 (是 1; 否 0)} \\ & + 0.01 \text{ 吸菸史 (current 1; 否 0)} + 0.01 \text{ 吸菸史 (ever 1; 否 0)} \\ & + 0.03 \text{ 吸菸史 (former 1; 否 0)} \\ & + 0.01 \text{ 吸菸史 (not current 1; 否 0)} + 0.01 \text{ 吸菸史 (never 1; 否 0)} \\ & + 0.004 \text{ BMI} + 0.08 \text{ 糖化血色素 (\%)} + 0.002 \text{ 血糖濃度 (mg/dL)} \end{aligned}$$

PartB 4.1 預測結果評估指標

準確率
Accuracy → 實際無患病 (18324) 較患病 (1676) 偏多，
準確率可能無法準確反映預測情形

精確度
Precision → 若預測正確但實際錯誤，此數值低
此失誤的代價僅為更多額外檢查

由於人命攸關，因此本題預測
指標將採用 **召回率 (Recall)**

召回率
Recall → 若實際正確但預測錯誤，此數值低
此失誤可能造成患者得不到及時治療失去性命

PartB 4.2 預測結果

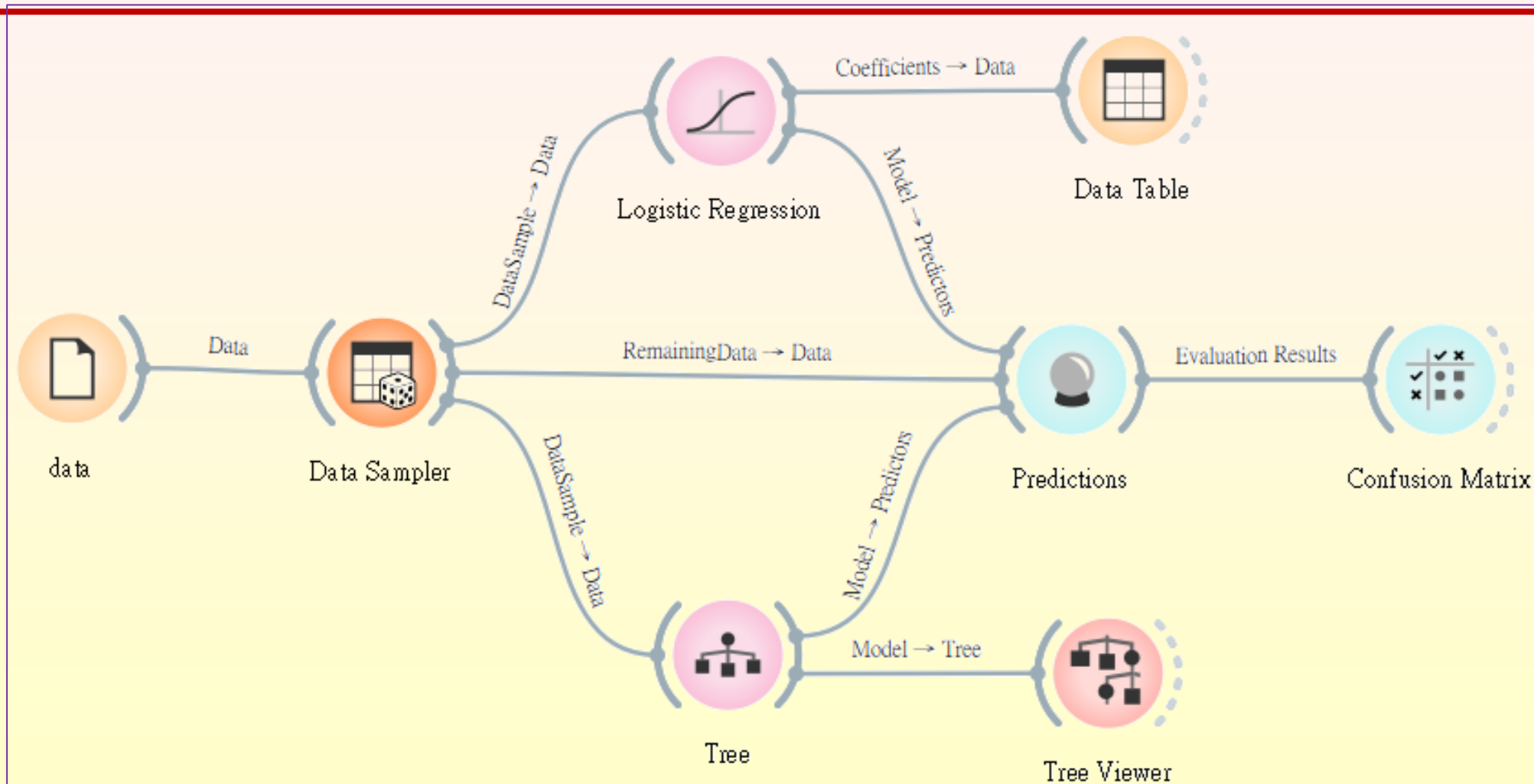
		預測		
		無患病	有患病	小計
實際	無患病	17830	494	18324
	有患病	517	1159	1676
	小計	18347	1653	20000
機率臨界	0.3			
Accuracy	0.95			
Precision	0.70			
Recall	0.69			
F1 score	0.70			

		預測		
		無患病	有患病	小計
實際	無患病	15986	2338	18324
	有患病	221	1455	1676
	小計	16207	3793	20000
機率臨界	0.2			
Accuracy	0.87			
Precision	0.38			
Recall	0.87			
F1 score	0.53			

PartB 4.3 結果討論

- 機率臨界值由 0.3 降至 0.2 有以下變化
 - ✓ 召回率提升 0.18
 - ✓ 準確率降低 0.08
 - ✓ 精確度顯著降低 0.32
- 召回率 (挽救更多生命) 與精確度 (節省醫療開支) 必須權衡
- 雖將 y 臨界值設為 0.2 有更高召回率，但精確度下降過多，使醫療開支遽增
因此模型將 y 臨界值設為 **0.3** 為佳 (召回率亦有 0.69 已足夠)

PartC 1.1 orange介面



PartC 1.2 變數設定與拆分資料集

	Name	Type	Role	Values
1	性別	C categorical	feature	Other, 女, 男
2	年齡 (歲)	N numeric	feature	
3	高血壓	C categorical	feature	0, 1
4	心臟病	C categorical	feature	0, 1
5	吸菸史	C categorical	feature	current, ever, former, never, no info, not current
6	BMI	N numeric	feature	
7	糖化血色素 (%)	N numeric	feature	
8	血糖濃度 (%)	N numeric	feature	
9	糖尿病	C categorical	target	0, 1

Sampling Type

☒ Fixed proportion of data:

80 %

註:

請老師打開 ows file 時，點選 data 後按 reload，使用的 excel 資料為作業檔案，工作表為 - 糖尿病_orange 用，並點選 9 糖尿病為 target (預設為 feature)，謝謝

PartC 2. 羅吉斯回歸

intercept	-10.6366
性別=Other	-3.86464
性別=女	-3.52362
性別=男	-3.24836
年齡 (歲)	0.0454289
高血壓=0	-5.7128
高血壓=1	-4.92381
心臟病=0	-5.70497
心臟病=1	-4.93164
吸菸史=current	-1.58455
吸菸史=ever	-1.63892
吸菸史=former	-1.6562
吸菸史=never	-1.73049
吸菸史=no info	-2.29726
吸菸史=not cu...	-1.7292
BMI	0.0906472
糖化血色素 (%)	2.34392
血糖濃度 (%)	0.0334363

		Predicted		Σ
		0	1	
Actual	0	18148	158	18306
	1	637	1057	1694
Σ		18785	1215	20000

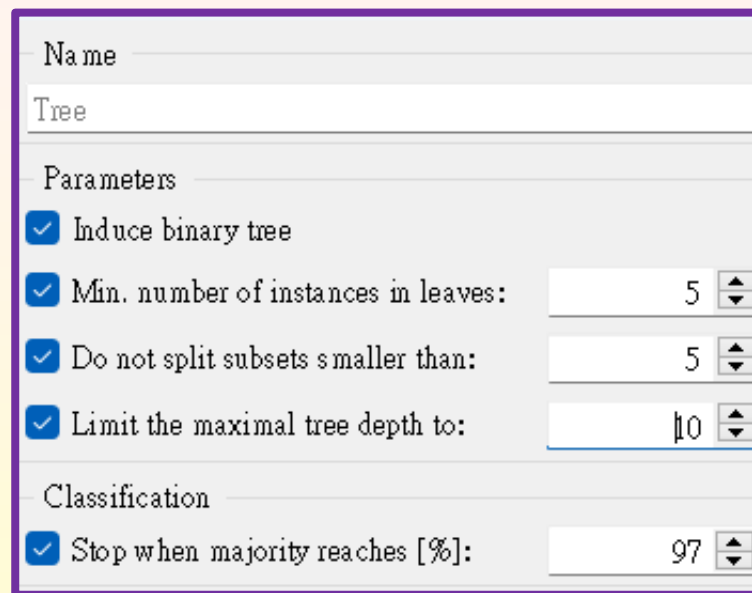
準確率 Accuracy = 0.96

精確度 Pprecision = 0.87

召回率 Recall = 0.62

F1 score = 0.73

PartC 3.1 分類試算一



Name

Tree

Parameters

- ☒ Induce binary tree
- ☒ Min. number of instances in leaves: 5
- ☒ Do not split subsets smaller than: 5
- ☒ Limit the maximal tree depth to: 10

Classification

- ☒ Stop when majority reaches [%]: 97

條件設定

- 樹的最大深度 = 10
- 當每群中糖尿病患者達 97% 停止

		Predicted		
		0	1	Σ
Actual	0	18306	0	18306
	1	573	1121	1694
Σ		18879	1121	20000

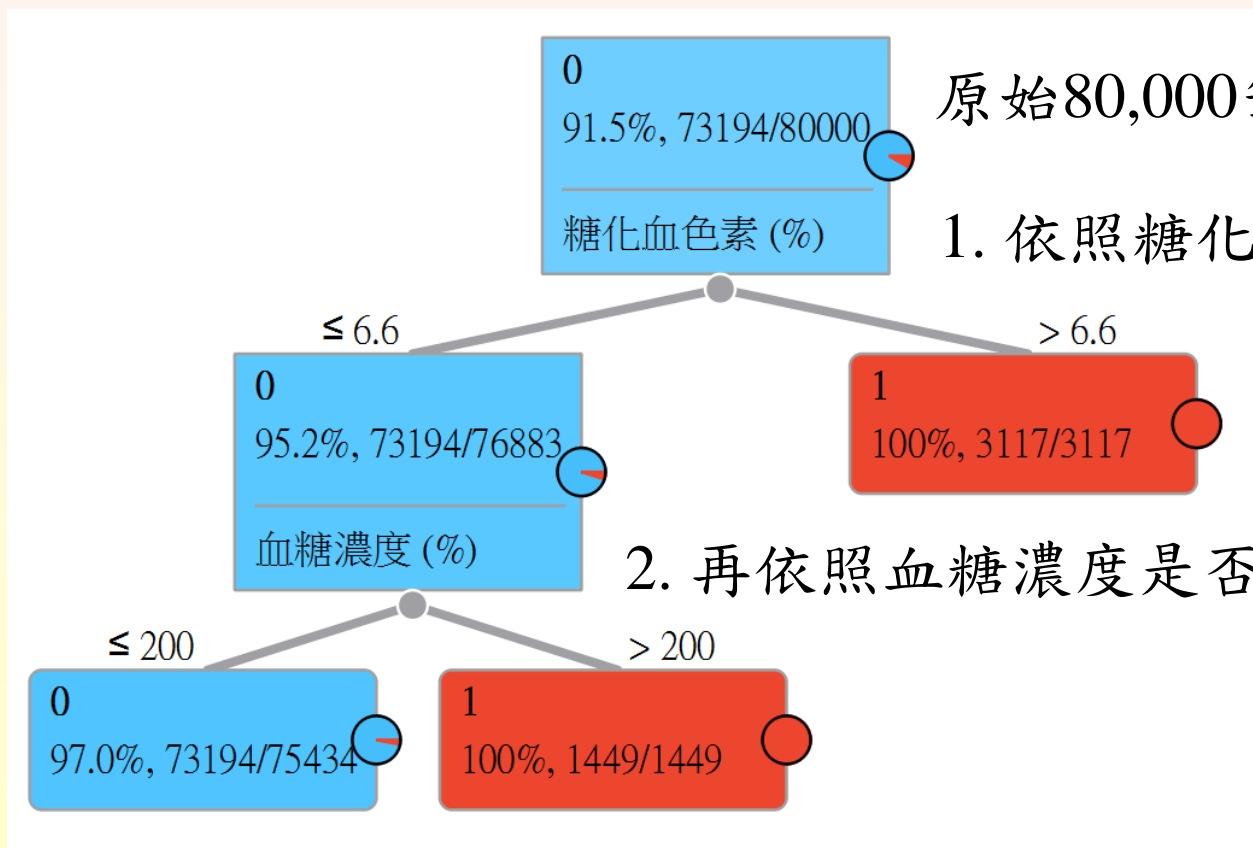
準確率 Accuracy = 0.83

精確度 Pprecision = 1.00

召回率 Recall = 0.66

F1 score = 0.80

PartC 3. 分類試算一



原始80,000筆尋練資料中有91.5%為有糖尿病

1. 依照糖化血色素是否 $> 6.6\%$ 區分

2. 再依照血糖濃度是否 $> 200\text{mg/dL}$ 區分

3. 已分至97%，停止

PartC 3. 分類試算二

Name
Tree

Parameters

- ☒ Induce binary tree
- ☒ Min. number of instances in leaves: 5
- ☒ Do not split subsets smaller than: 5
- ☒ Limit the maximal tree depth to: 15

Classification

- ☒ Stop when majority reaches [%]: 98

條件設定

- 樹的最大深度 = 15
- 當每群中糖尿病患者達 98% 停止

		Predicted		Σ
		0	1	
Actual	0	18200	106	18306
	1	530	1164	1694
Σ		18730	1270	20000

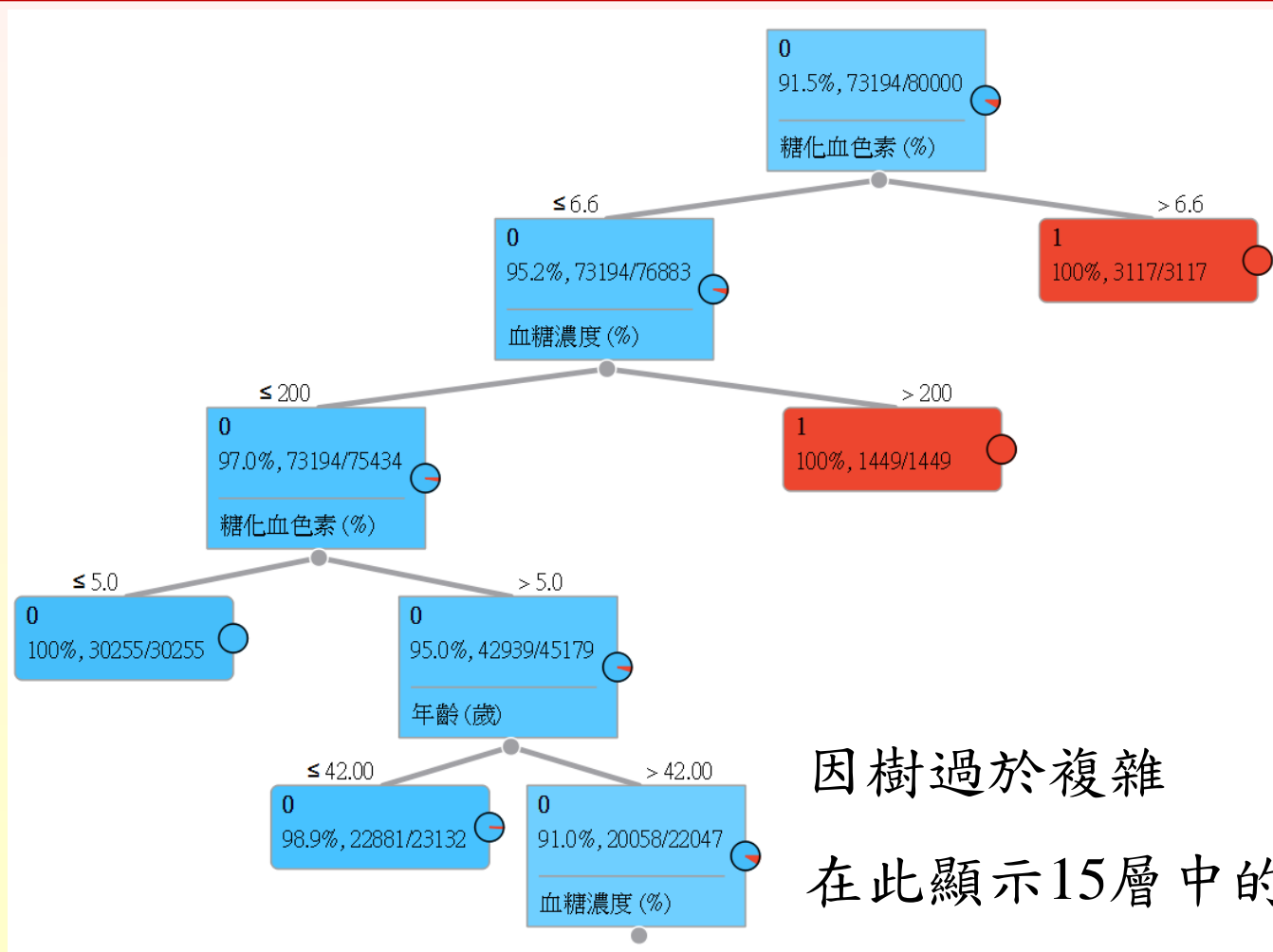
準確率 Accuracy = 0.95

精確度 Pprecision = 0.92

召回率 Recall = 0.69

F1 score = 0.79

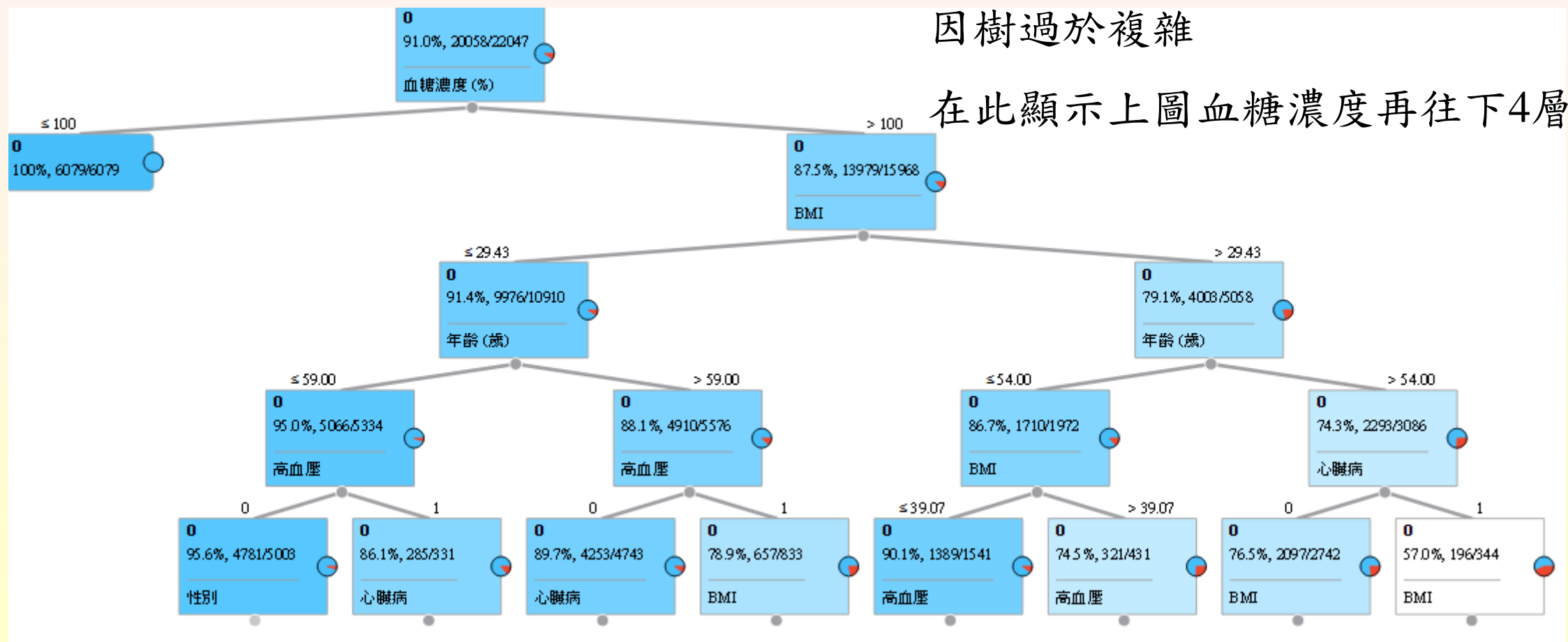
PartC 3. 分類試算二



因樹過於複雜

在此顯示15層中的前5層

PartC 3. 分類試算二



PartC 3.1 excel迴歸 y臨界值 0.5

補上excel迴歸將 y 臨界值設為 0.5 方便與 orange 結果比較

		預測		小計
		無患病	有患病	
實際	無患病	18315	9	18324
	有患病	1179	497	1676
	小計	19494	506	20000
機率臨界	0.5			
Accuracy	0.94			
Precision	0.98			
Recall	0.30			
F1 score	0.46			

PartC 3.2 excel 與 orange 比較

模型	準確率 Accuracy	精確度 Precision	召回率 Recall	F1 score
Excel 線性回歸	0.94	0.98	0.30	0.46
Orange 羅吉斯回歸	0.96	0.87	0.62	0.73
Orange 分類 試算一 • 最大深度 = 10 • 每群糖尿病患者達 97% 停止	0.83	1.00	0.66	0.80
Orange 分類 試算二 • 最大深度 = 15 • 每群糖尿病患者達 98% 停止	0.95	0.92	0.69	0.79

PartC 3.1 excel 與 orange 比較

- 整體而言，orange 分類設定樹最大深度 = 15、每群糖尿病患者達 98% 停止，召回率最佳，達 0.69 接近7成
- 且準確率、精確度亦高皆達9成，可減少醫療資源浪費情形
- excel 迴歸雖可設定y臨界值至 0.3、0.2，有較高召回率可達0.87，但精確度只有0.38，十分浪費醫療資源
- 權衡生命挽救與珍惜醫療資源下，orange分類仍為最佳模型