機器學習期末報告

應用機器學習建立預測模式：以乳癌資料集為例

班級：碩管一甲
學號：M09218001
學生：周彥廷
授課教師 ：陳昆皇 老師

目錄

目錄

問題發想-國人十大死因

108 年國人十大死因 VS 十大癌症		
NO	十大死因	十大癌症
1	惡性腫瘤（癌症）	氣管、支氣管和肺癌
2	心臟疾病	肝和肝內膽管癌
3	肺炎	結腸、直腸和肛門癌
4	腦血管疾病	女性乳癌
5	糖尿病	口腔癌
6	事故傷害	前列腺（攝護腺）癌
7	慢性下呼吸道疾病	胰臟癌
8	高血壓性疾病	胃癌
9	腎炎、腎病症候群及腎病變	食道癌
10	慢性肝炎及肝硬化	卵巢癌

資料來源：衛生福利部統計處；製表：洪毓琪

1.國人十大死因中，癌症連續38年居首位，其中仍以氣管、支氣管肺癌與肝和肝內膽管癌的死亡率較高，分居十大癌症前二名。且就年齡觀察，癌症多集中於55歲以上之族群 [1]。

2.國人十大死因中多數與肥胖有密切關聯，包括心臟疾病、腦血管疾病、糖尿病、慢性下呼吸道疾病、慢性肝病及肝硬化、高血壓性疾病、腎臟病等 [2]。

3.就年齡觀察，癌症多集中於55歲以上之族群，108年占8成5；65歲以上癌症死亡人數較上年增899人；0-64歲則較上年增7人 [3]。

參考來源：
1. 華人健康網 記者洪毓琪台北報導 https://www.top1health.com/Article/82117
2. 藥師週刊／台大醫院雲林分院藥劑部藥師 https://www.taiwan-pharma.org.tw/weekly/2022/2022-3-5.htm
3. Heho健康／林以璿 https://heho.com.tw/archives/87020

問題發想-乳癌

圖來源：Heho / 林以璿

1.根據世界衛生組織國際癌症研究機構(IARC)發布的2020年全球最新癌症負擔數據，乳癌首度超過肺癌，成為全球最常見的癌症 [4]。

2.乳癌雖然不是台灣發生案例最多的癌症，但如果將男女性分開看，乳癌絕對是台灣女性的頭號殺手 [4]。

3.「乳癌」在女性癌症發生率排行第一位、女性主要癌症死因排名第三位，北投健康管理醫院主任沈彥君表示，歐美國家大多為停經後罹患乳癌的機率較高，但國人罹患乳癌的平均年齡相較於歐美國家提早了約十年。
Ex：朱俐靜，阿桑，薇薇安 …。

綜上所述：台灣女性罹患乳癌的平均年齡相較於歐美國家有年輕化之趨勢，因此作為探討議題依據。

參考來源：
4. Heho健康／林以璿 https://heho.com.tw/archives/153696

問體發想-思路與研究目的

[1]

[2]

[3]

問題發想

動機

目的

健康議題

國人十大死因

主要病因

癌症(惡性腫瘤)

題目方向

乳癌議題

參考來源:
[1]漫漫健康 / 108年國人十大死因 https://havemary.com/article.php?id=5701
[2]漫漫健康 / 癌症的危險因子 https://www.havemary.com/article.php?id=5162
[3]ET today 新聞雲 https://health.ettoday.net/news/1868153

目錄

文獻探討

目錄

- 認識乳癌
- 機器學習策略思想

認識乳癌

乳癌形成：

乳癌是從乳腺的上皮細胞或小葉生長出來的惡性瘤。癌細胞是由正常細胞變異而來，如細胞出現病變時，就可能演變為癌細胞，進而出現過度繁殖的現象。當癌細胞積聚在某個組織或器官，如乳腺管或乳小葉，就會形成腫瘤 [1]。

症狀：

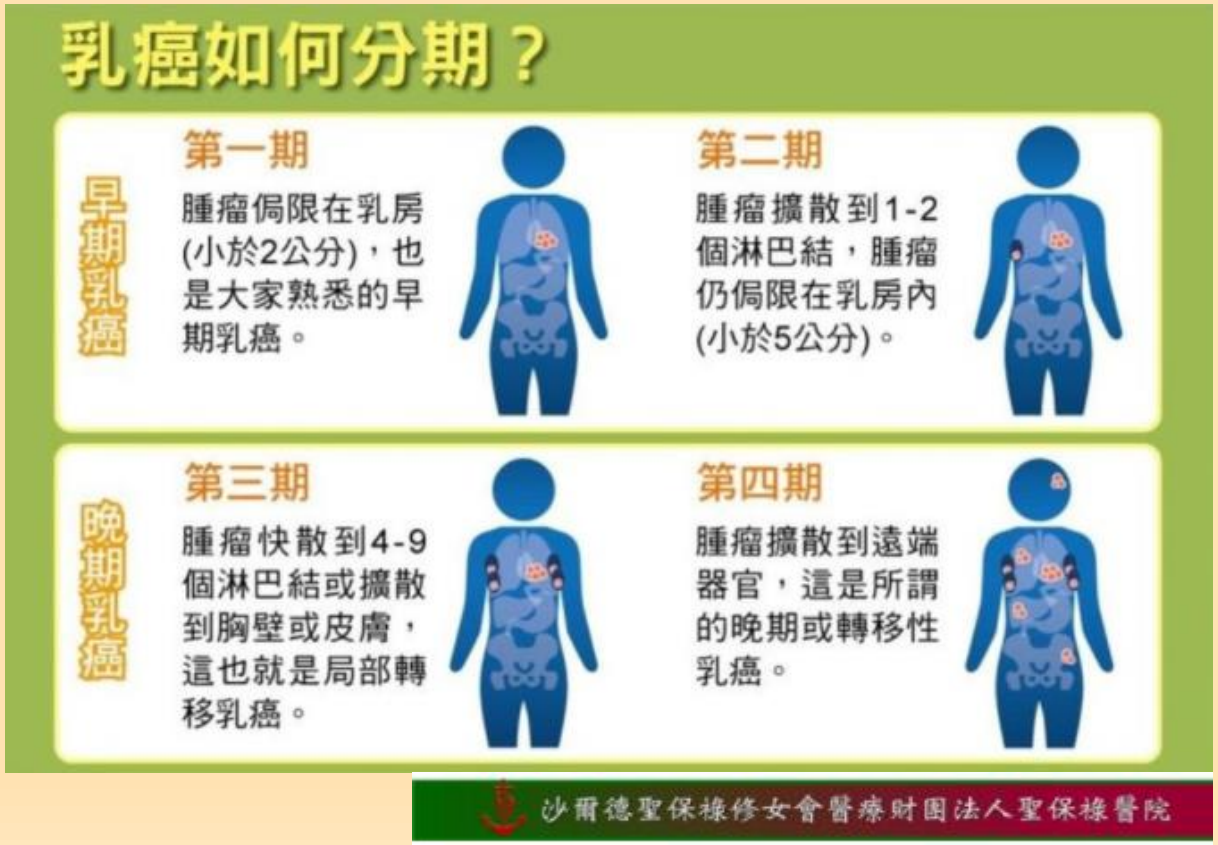
乳房出現腫塊是乳癌的重要徵兆，如出現不痛的腫塊、乳房局部變硬。除此之外，還可能有下列幾種症狀[1]：

1.出現偏硬的腫塊、形狀不規則
2.腫塊固定在皮膚或胸壁上，且邊緣不清楚
3.皮膚出現凹陷、橘子皮變化
4.乳頭凹陷且有不正常分泌物

https://frenchbebe.cyberbiz.co/blogs/%E4%BF%9D%E9%A4%8A%E5%B0%88%E6%AC%84/36768

[1]認識乳癌(一) 乳癌病程與常見症狀｜家天使居家照護 https://ghsha.com/articles/215

乳癌病程與存活率



民國 91 年新診斷癌症期別之個案數、死亡數與存活率－乳癌

	期別*			存活率 (%) **				
	第五版	第六版						
	個案數	個案數	死亡數	第一年	第二年	第三年	第四年	第五年***
Stage 0	173	12	0	100.0	100.0	100.0	100.0	100.0
Stage I	515	68	6	100.0	97.1	97.1	92.6	91.2
Stage II	882	128	17	99.2	96.1	93.0	89.8	86.7
Stage III	265	59	24	96.6	91.5	78.0	71.2	59.3
Stage IV	83	12	9	66.7	50.0	33.3	25.0	25.0
Overall	1918	279	56	97.5	93.5	88.5	84.2	79.9

* 根據 AJCC 分期
**根據 AJCC 第六版
***個案追蹤至民國 96 年 12 月 31 日

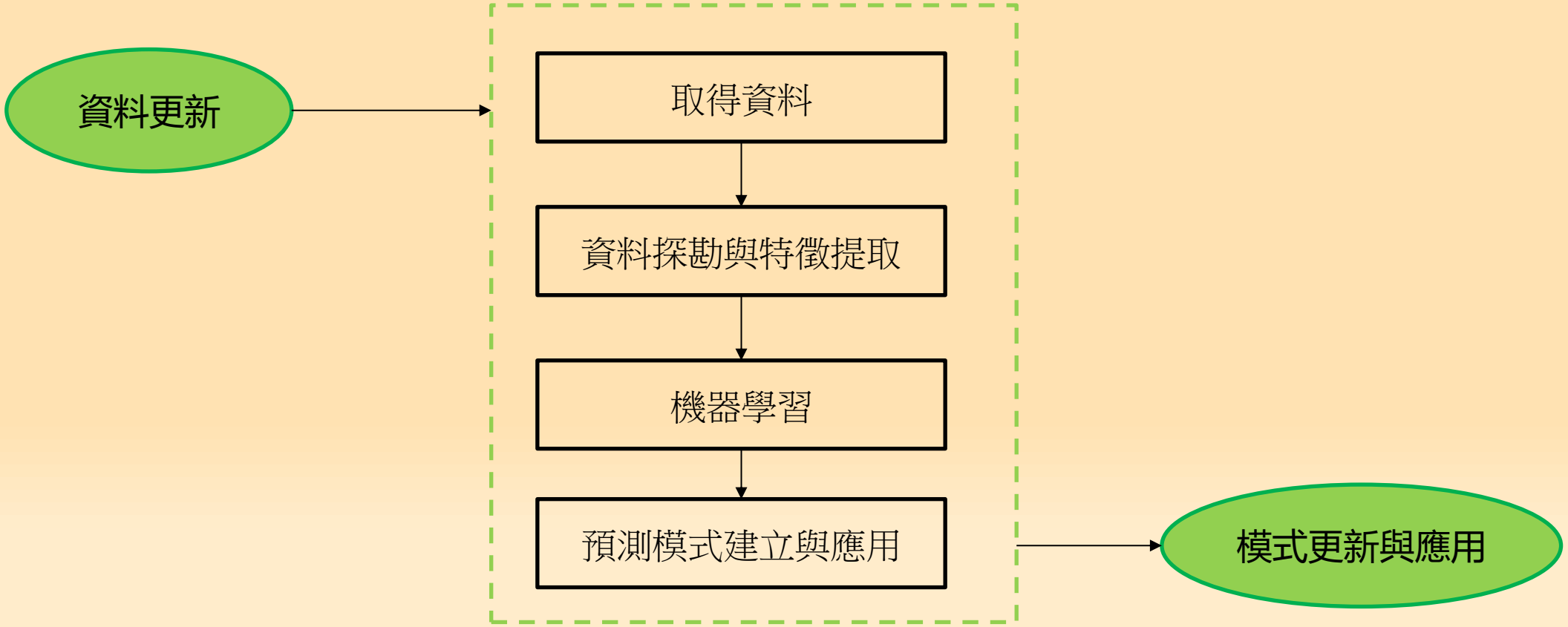
資料來源：行政院衛生福利部癌症登記小組，癌症治療期別與治療分析年度報告（六癌）

衛生福利部台灣癌症登記小組 http://crs.cph.ntu.edu.tw/main.php?Page = A1

結語：乳癌如在前二期被診斷出來與及早接受治療可以有八成以上存活率。

使用機器學習策略思想

0.假設現在有一筆乳癌相關資料(具備：乳癌病患特徵)
1.透過資料探勘→ 特徵擷取，了解各欄位變數(相關性，散佈狀況...等)。
2.使用機器學習演算法根據擷取特徵進行分類學習判斷。
3.建立一能有效判斷是否罹患乳癌模式。



目錄

收集所需資料

資料來源: UCI Machine learning 公開資料集
https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra

資料包含10項特徵116筆資料，如下表：

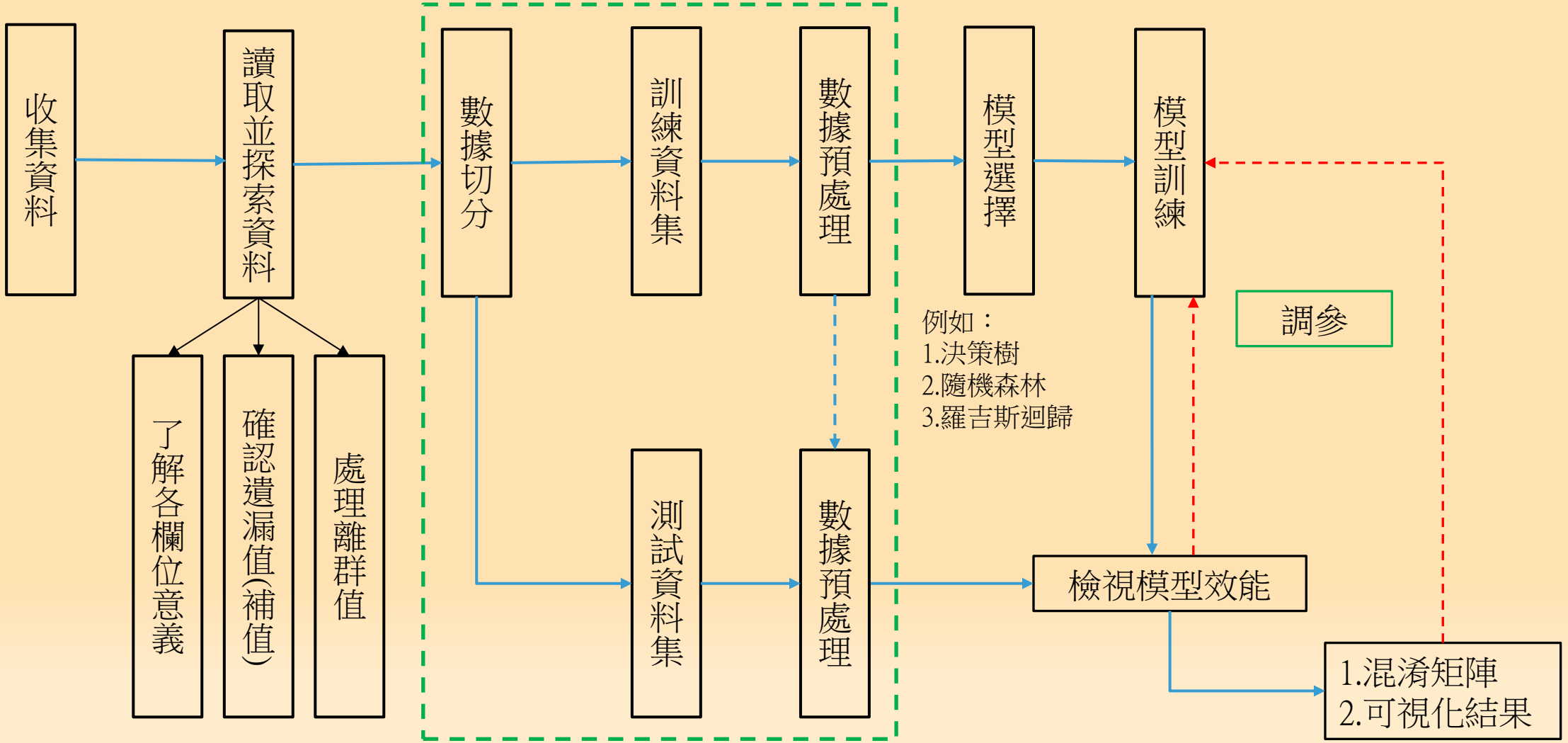
欄位	Age (years)	BMI (kg/m2)	Glucose (mg/dL)	Insulin (µU/mL)	HOMA	Leptin (ng/mL)	Adiponectin (µg/mL)	Resistin (ng/mL)	MCP-1(pg/dL)	Classification
中文名稱	年齡	身體質量指數	葡萄糖	胰島素	胰島素阻抗數值	瘦素	脂締素	阻抗素	單核球趨化蛋白	1 = 健康人 2 = 乳癌病患

預覽資料

初步將data.csv檔案載入並預覽

Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Classification
48	23.5	70	2.707	0.467409	8.8071	9.7024	7.99585	417.114	1
83	20.6905	92	3.115	0.706897	8.8438	5.42929	4.06405	468.786	1
82	23.1247	91	4.498	1.00965	17.9393	22.432	9.27715	554.697	1
68	21.3675	77	3.226	0.612725	9.8827	7.16956	12.766	928.22	1
86	21.1111	92	3.549	0.805386	6.6994	4.81924	10.5763	773.92	1
49	22.8545	92	3.226	0.732087	6.8317	13.6798	10.3176	530.41	1
89	22.7	77	4.69	0.890787	6.964	5.58987	12.9361	1256.08	1
76	23.8	118	6.47	1.8832	4.311	13.2513	5.1042	280.694	1
73	22	97	3.35	0.801543	4.47	10.3587	6.28445	136.855	1
75	23	83	4.952	1.01384	17.127	11.579	7.0913	318.302	1
34	21.47	78	3.469	0.667436	14.57	13.11	6.92	354.6	1

資料處理與進行機器學習

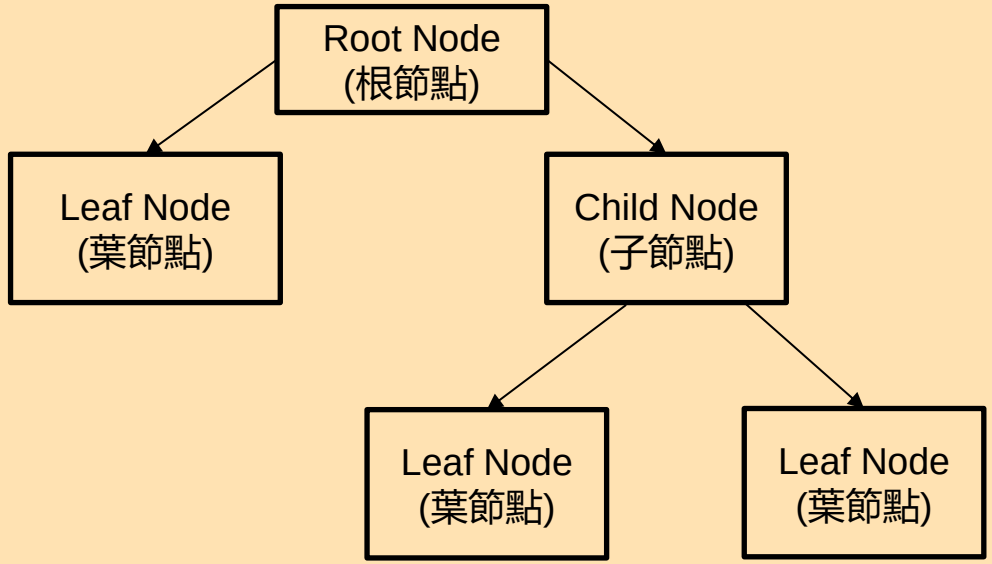


透過課堂所學, 初步繪製資料處理流程圖

決策樹(Decision Tree)

透過一系列的是非問題，幫助我們將資料進行切分
可視覺化每個決策的過程，是個具有非常高解釋性的模型
從訓練資料中找出規則，讓每一次決策能使訊息增益(Information Gain) ，訊息增益越大代表切分後的兩群資料，群內相似程度越高

Information Gain：
決策樹模型會用 features 切分資料，該選用哪個 feature 來切分則是由訊息增益的大小決定的。希望切分後的資料相似程度很高，通常使用吉尼係數(Gini)來衡量相似程度。



Breiman, L. I et al. (1984)

目錄

預期結果

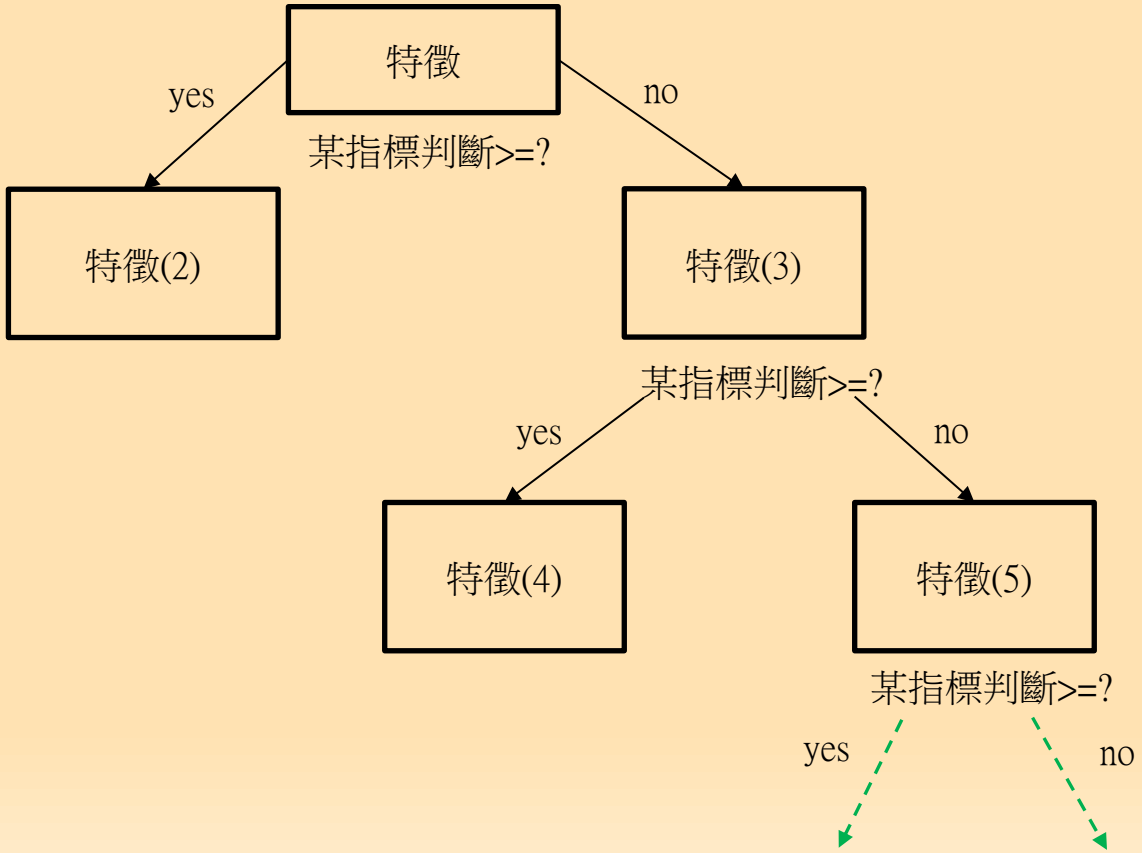
1.使用決策樹進行資料探索與預測並可視化結果。

2.比較不同演算法分類效能。

- 決策樹
- 隨機森林
- 羅吉斯迴歸

3.調整超參數與優化演算法。

4.列出重要變數列表。



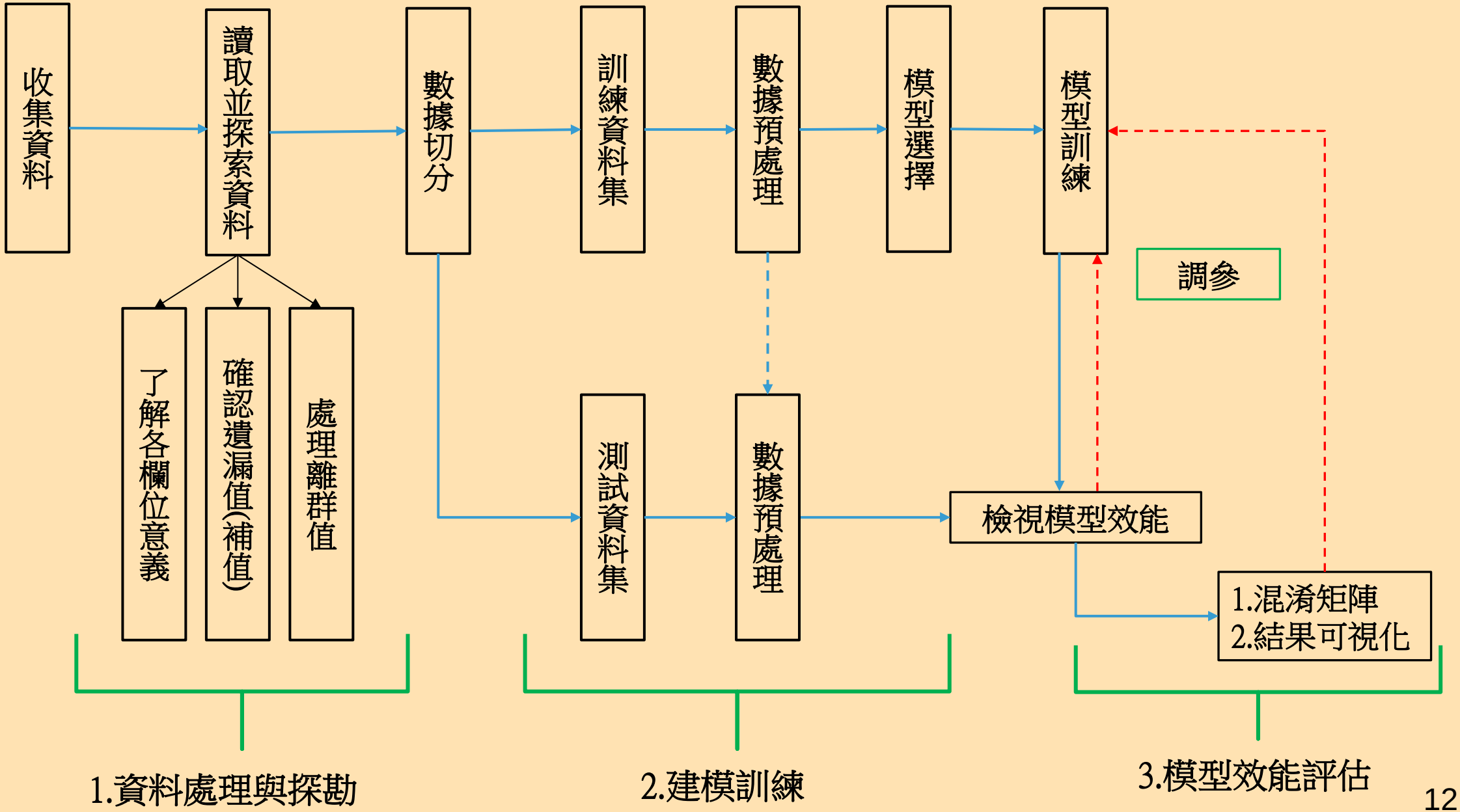
規則可視化/樹狀圖

Thanks for listening

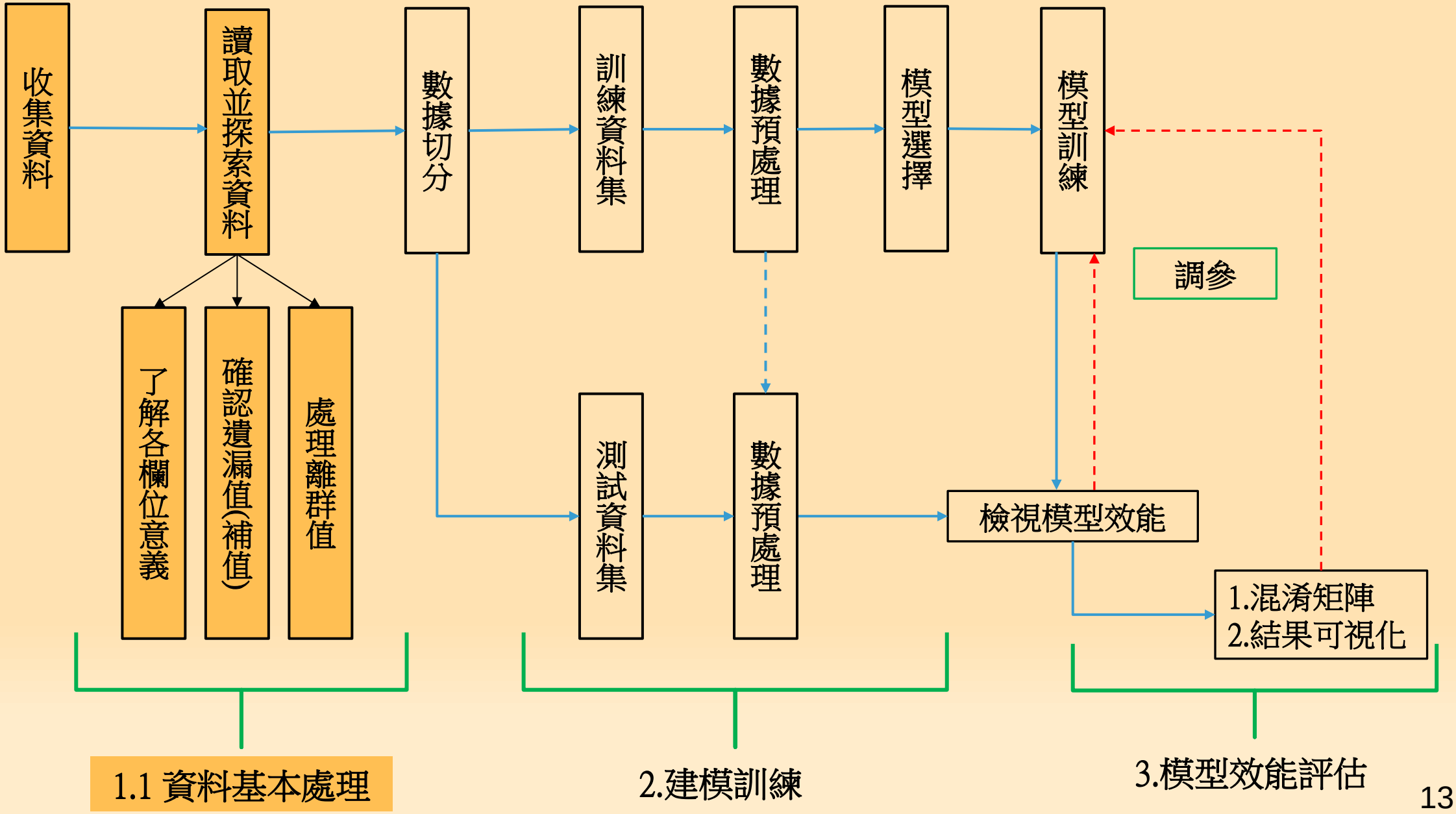
機器學習期末報告

班級：碩管一甲
學號：M09218001
學生：周彥廷
授課教師 ：陳昆皇 老師

機器學習_流程圖


收集資料
讀取並探索資料
數據切分
訓練資料集
數據預處理
模型選擇
模型訓練
調參
了解各欄位意義
確認遺漏值(補值)
處理離群值
測試資料集
數據預處理
檢視模型效能
1.混淆矩陣
2.結果可視化
1.資料處理與探勘
2.建模訓練
3.模型效能評估

機器學習_流程圖


收集資料
讀取並探索資料
了解各欄位意義
確認遺漏值(補值)
處理離群值
數據切分
訓練資料集
測試資料集
數據預處理
數據預處理
模型選擇
模型訓練
調參
檢視模型效能
1.混淆矩陣
2.結果可視化
1.1 資料基本處理
2.建模訓練
3.模型效能評估

資料基本處理(1/2)

#讀取資料集
df = pd.read_csv('ML_data.csv') → 116x10

#顯示資料型態
type(data) → pandas.core.frame.DataFrame

#顯示前5筆資料
df.head(5)

#顯示資料各項型態→1-9欄位為數量，第10欄為類別
data.info()

#計算資料缺失值
data.isnull().sum() → 每個特徵欄位缺失皆為0

#敘述統計
data.describe()

```
In [149]: df.head(5)
Out[149]:
   Age        BMI  Glucose  ...  Resistin    MCP.1  Classification
0   48  23.500000       70  ...   7.99585  417.114               0
1   83  20.690495       92  ...   4.06405  468.786               0
2   82  23.124670       91  ...   9.27715  554.697               0
3   68  21.367521       77  ...  12.76600  928.220               0
4   86  21.111111       92  ...  10.57635  773.920               0
```

```
In [147]: df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 116 entries, 0 to 115
Data columns (total 10 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Age             116 non-null    int64
 1   BMI             116 non-null    float64
 2   Glucose         116 non-null    int64
 3   Insulin         116 non-null    float64
 4   HOMA            116 non-null    float64
 5   Leptin          116 non-null    float64
 6   Adiponectin     116 non-null    float64
 7   Resistin        116 non-null    float64
 8   MCP.1           116 non-null    float64
 9   Classification  116 non-null    int32
dtypes: float64(7), int32(1), int64(2)
```

```
In [148]: df.describe()
Out[148]:
              Age         BMI  ...        MCP.1  Classification
count  116.000000  116.000000  ...   116.000000      116.000000
mean    57.301724   27.582111  ...   534.647000        0.551724
std     16.112766    5.020136  ...   345.912663        0.499475
min     24.000000   18.370000  ...    45.843000        0.000000
25%     45.000000   22.973205  ...   269.978250        0.000000
50%     56.000000   27.662416  ...   471.322500        1.000000
75%     71.000000   31.241442  ...   700.085000        1.000000
max     89.000000   38.578759  ...  1698.440000        1.000000
```

資料基本處理(2/2)

#發現在'Classification' 欄位健康者為1，乳癌患者為2，為方便後續演算法編碼能統一，
因此轉換為 0 和 1 。

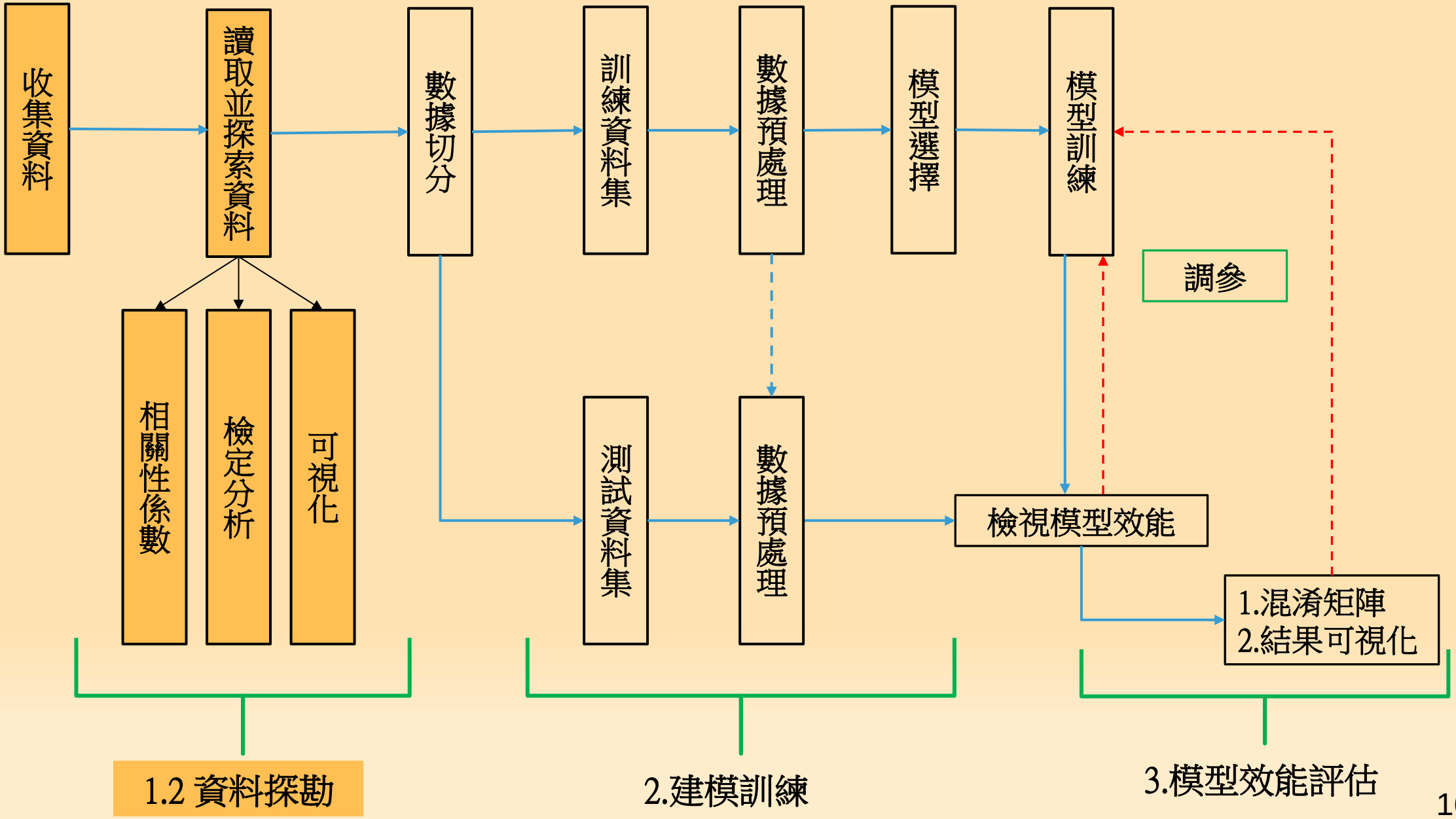
Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Classification
85	26.6	96	4.462	1.0566	7.85	7.9317	9.6135	232.006	1
76	27.1	110	26.211	7.11192	21.778	4.93563	8.49395	45.843	1
77	25.9	85	4.58	0.960273	13.74	9.75326	11.774	488.829	1
45	21.3039	102	13.852	3.48516	7.6476	21.0566	23.0341	552.444	2
45	20.83	74	4.56	0.832352	7.7529	8.23741	28.0323	382.955	2

將 'Classification' 1 換成 0 & 2 換成 1

```
df['Classification'] =
np.where(df['Classification']==1,0,1)
df.groupby(df['Classification'])
```

Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Classification
85	26.6	96	4.462	1.0566	7.85	7.9317	9.6135	232.006	0
76	27.1	110	26.211	7.11192	21.778	4.93563	8.49395	45.843	0
77	25.9	85	4.58	0.960273	13.74	9.75326	11.774	488.829	0
45	21.3039	102	13.852	3.48516	7.6476	21.0566	23.0341	552.444	1
45	20.83	74	4.56	0.832352	7.7529	8.23741	28.0323	382.955	1

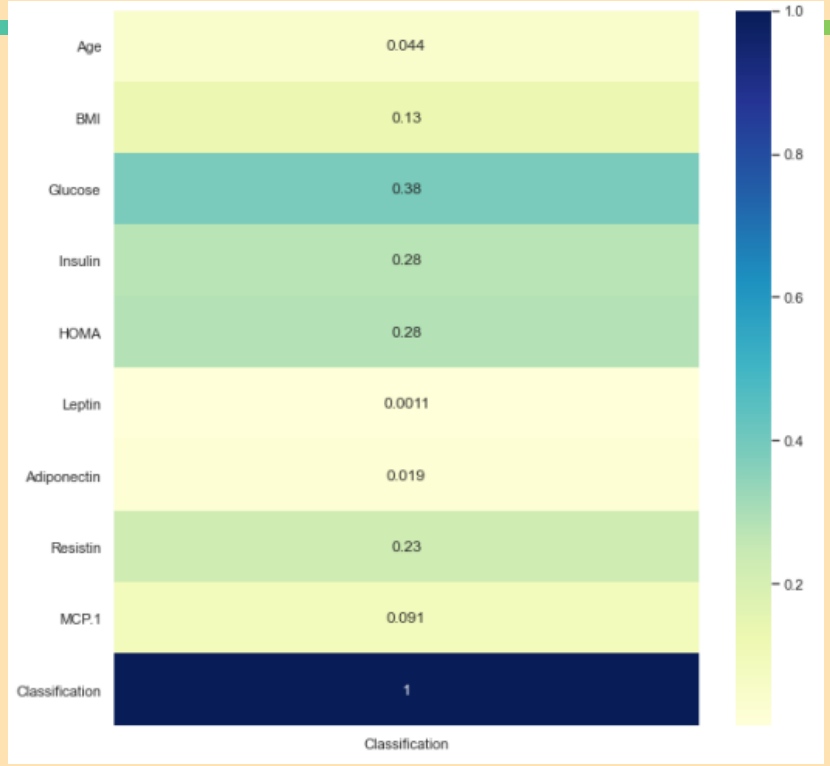
機器學習_流程圖


收集資料
讀取並探索資料
數據切分
訓練資料集
數據預處理
模型選擇
模型訓練
調參
相關性係數
檢定分析
可視化
測試資料集
數據預處理
檢視模型效能
1.混淆矩陣
2.結果可視化
1.2 資料探勘
2.建模訓練
3.模型效能評估

資料探勘(1/5)

```
#關聯係數矩陣
corr = df.corr()
corr_result=pd.DataFrame(corr['Classification'])
corr_result.sort_index(ascending=False)
corr_result.sort_values(by='Classification')
abs(corr_result.sort_values(by='Classification'))

# 繪製 以 'Classification' 為 的 Corr 關聯係數矩陣
sns.set(rc={'figure.figsize':(10,10)})
correlation_matrix =
df.corr().round(4).loc[:,['Classification']].abs()
sns.heatmap(data=correlation_matrix, annot = True,
cmap='YlGnBu')
```



	Classification
BMI	-0.132586
Age	-0.043555
Adiponectin	-0.019490
Leptin	-0.001078
MCP.1	0.091381
Resistin	0.227310
Insulin	0.276804
HOMA	0.284012
Glucose	0.384315
Classification	1.000000

顯示正負相關

	Classification
BMI	0.132586
Age	0.043555
Adiponectin	0.019490
Leptin	0.001078
MCP.1	0.091381
Resistin	0.227310
Insulin	0.276804
HOMA	0.284012
Glucose	0.384315
Classification	1.000000

取abs

資料探勘(2/5)

```
#多量對類T檢定 對 'Classification' 欄位
t_test=['Age', 'BMI', 'Glucose', 'Insulin', 'HOMA', 'Leptin', 'Adiponectin','Resistin', 'MCP.1']
alist=[]
pvlist=[]
For loop

#皮爾森相關分析 對 Age 欄位
#取 qu 變數與 Age 做皮爾森相關分析，a為相關係數，pv為P值
qu = t_test
from scipy import stats
name=[]
alist=[]
pvlist=[]
for i in range(len(qu)):
  (a,pv)=stats.pearsonr(df['Age'],df[qu[i]])
  name.append(qu[i])
  alist.append(a)
  pvlist.append(pv)
```

Index	t	pv
Age	0.465478	0.642477
BMI	1.42824	0.155957
Glucose	-4.44471	2.05222e-05
Insulin	-3.07563	0.00262986
HOMA	-3.16266	0.00200384
Leptin	0.0115148	0.990833
Adiponectin	0.208139	0.835492
Resistin	-2.49225	0.0141314
MCP.1	-0.979776	0.329271

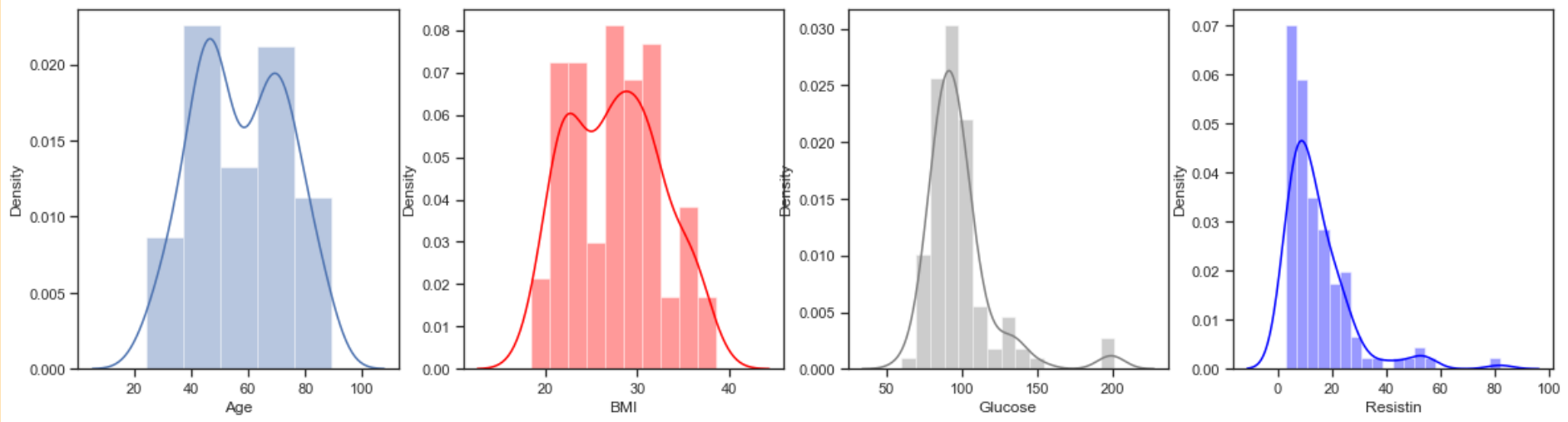
T檢定

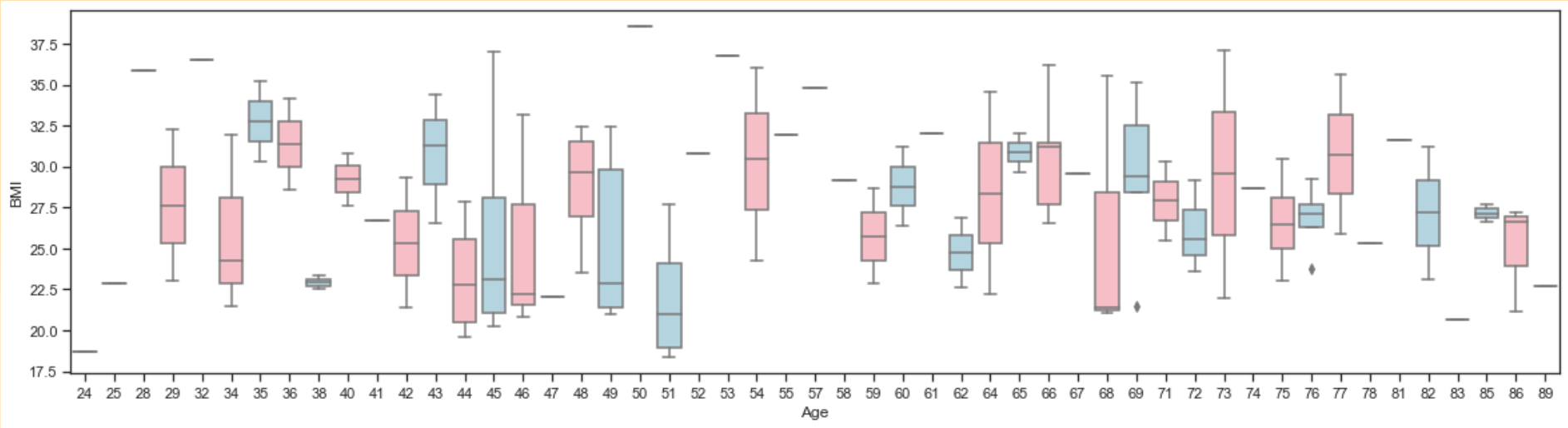
Index	a	pv
Age	1	0
BMI	0.00852986	0.927591
Glucose	0.230106	0.0129583
Insulin	0.0324954	0.729127
HOMA	0.127033	0.174185
Leptin	0.102626	0.272972
Adiponectin	-0.219813	0.0177452
Resistin	0.00274171	0.976698
MCP.1	0.0134617	0.885956

皮爾森相關分析

資料探勘(3/5)

#根據相關文獻 [1] 對 Age、BMI 、Glucose和 Resistin特徵進行可視化。

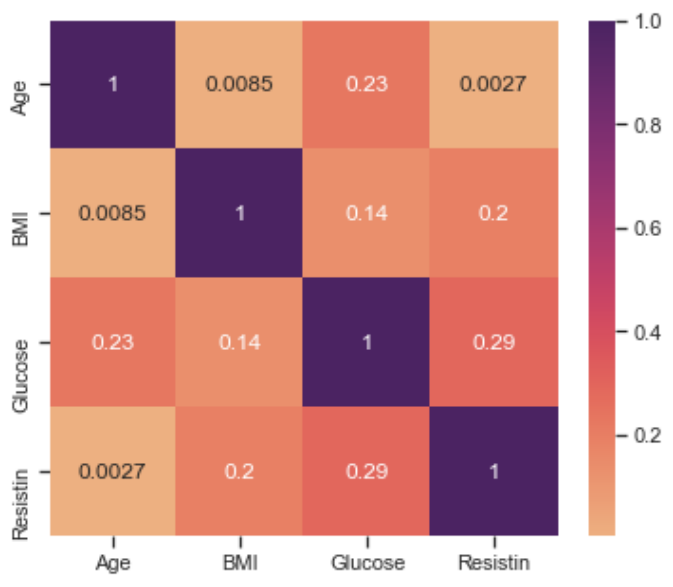




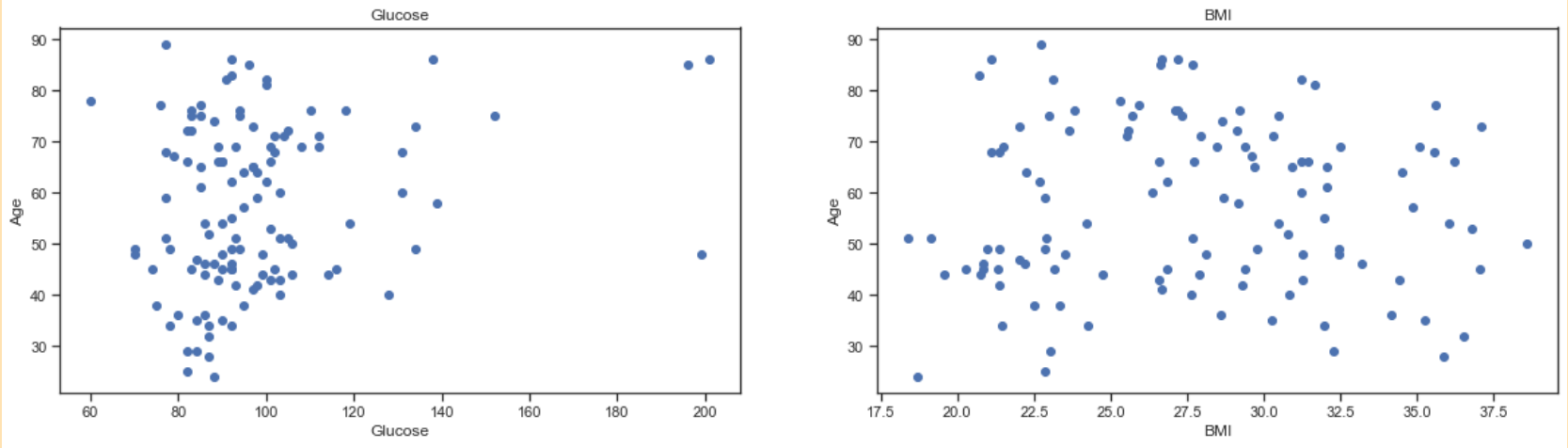
盒形圖(x = Age, y = BMI)

相關文獻：
1. M Patrício et al. (2018)

資料探勘(4/5)



熱圖(heatmap)

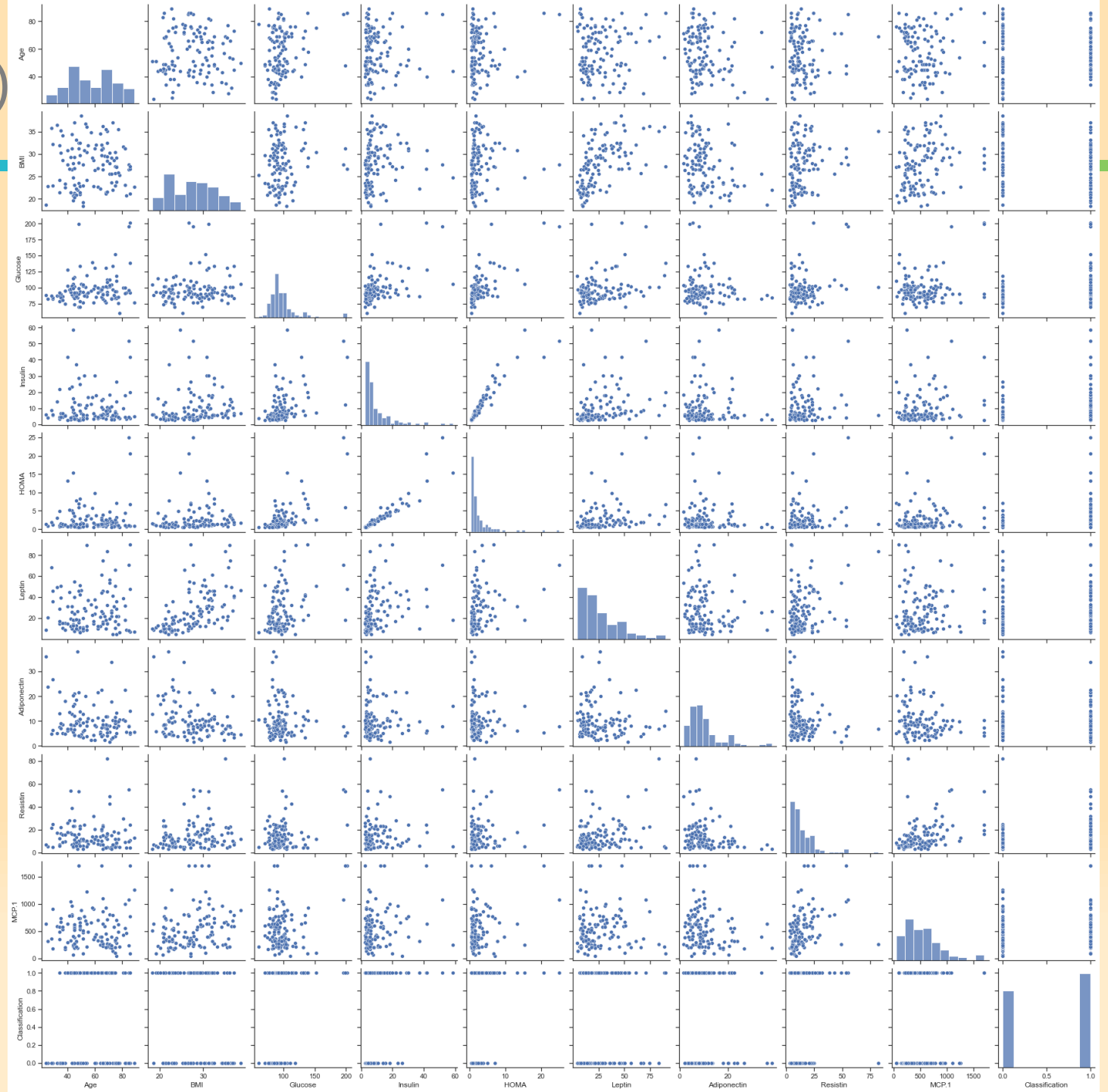


Age/Glucose 和 Age/BMI 散佈圖

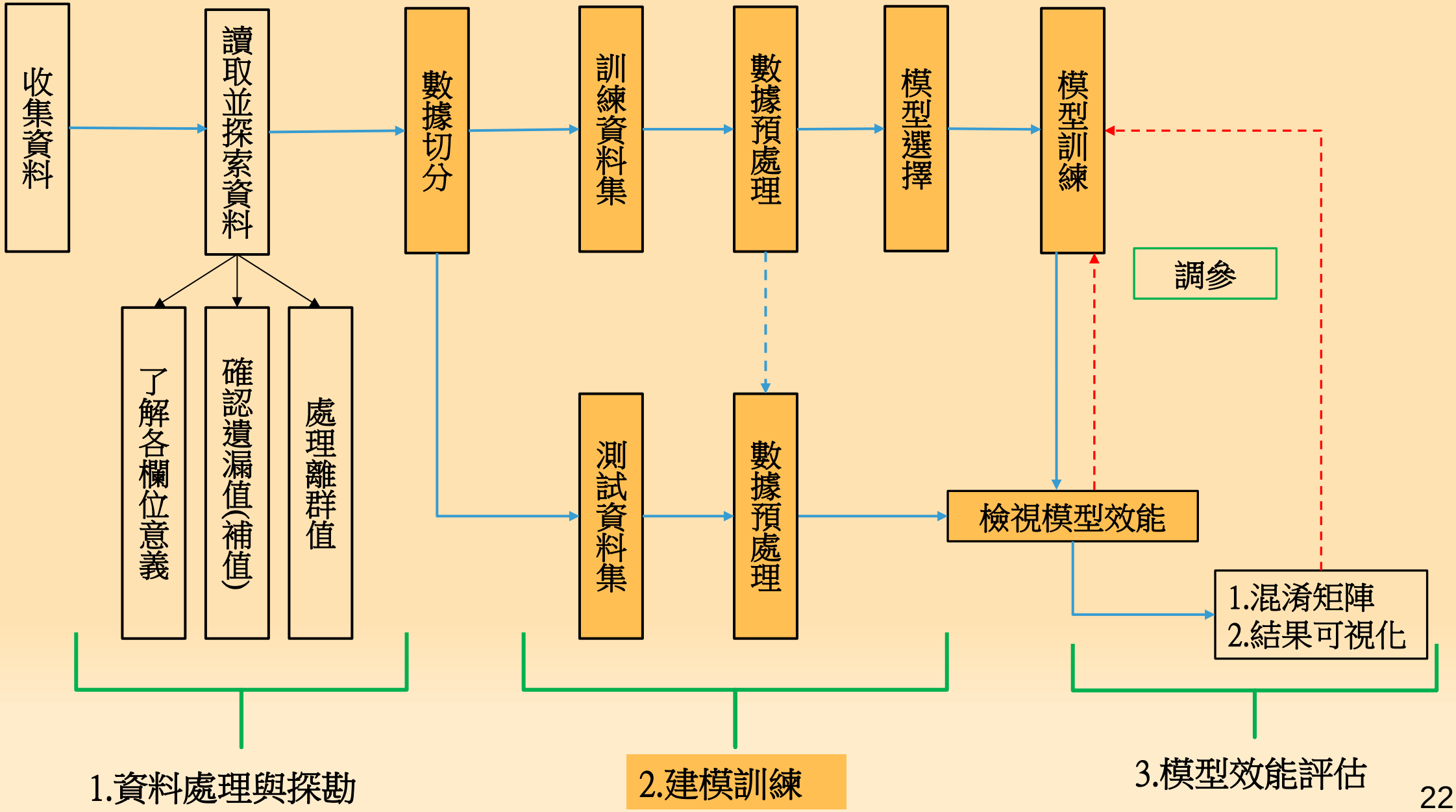
相關文獻：
1. M Patrício et al. (2018)

資料探勘(5/5)

配對圖型(整個資料集)

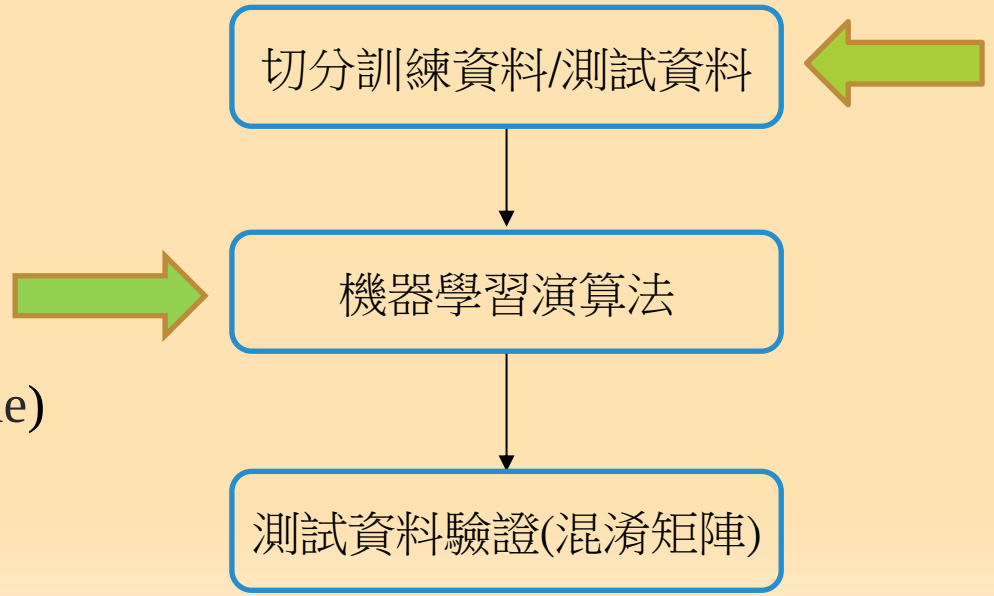


機器學習_流程圖


收集資料
讀取並探索資料
了解各欄位意義
確認遺漏值(補值)
處理離群值
數據切分
訓練資料集
數據預處理
模型選擇
模型訓練
調參
測試資料集
數據預處理
檢視模型效能
1.混淆矩陣
2.結果可視化
1.資料處理與探勘
2.建模訓練
3.模型效能評估

建模訓練

機器學習演算法:
1.決策樹(Decision Tree)
2.隨機森林(Random Forest)
3.羅吉斯迴歸(Logistic Regression)
4.支持向量機(Support Vector Machine)



train	test
6	4
7	3
8	2

資料集切分比例結果

嘗試不同資料切分比例進行訓練並比較測試資料正確率。

發現 Support Vector Machine(SVM)於下表實驗皆保持有最高測試正確率，
因此後續超參數調整將以SVM演算法作為依據。

accuracy / train： test	Decision Tree	Random Forest	Logistic Regression	Support Vector Machine
6 ： 4	0.617	0.532	0.617	0.640
7 ： 3	0.600	0.514	0.685	0.710
8 ： 2	0.541	0.625	0.708	0.710

初步模型效能與混淆矩陣

訓練：測試 = 7：3
4種模型混淆矩陣
如右圖所示：

		True Condition	
		Positive	Negative
Predicted outcome	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

```
array([[ 0, 14],
       [ 0, 21]], dtype=int64)
```

	precision	recall	f1-score	support
0	0.00	0.00	0.00	14
1	0.60	1.00	0.75	21
accuracy			0.60	35
macro avg	0.30	0.50	0.37	35
weighted avg	0.36	0.60	0.45	35

1. Decision Tree

```
array([[ 6,  8],
       [ 9, 12]], dtype=int64)
```

	precision	recall	f1-score	support
0	0.40	0.43	0.41	14
1	0.60	0.57	0.59	21
accuracy			0.51	35
macro avg	0.50	0.50	0.50	35
weighted avg	0.52	0.51	0.52	35

2.Random Forest

```
array([[10,  4],
       [ 7, 14]], dtype=int64)
```

	precision	recall	f1-score	support
0	0.59	0.71	0.65	14
1	0.78	0.67	0.72	21
accuracy			0.69	35
macro avg	0.68	0.69	0.68	35
weighted avg	0.70	0.69	0.69	35

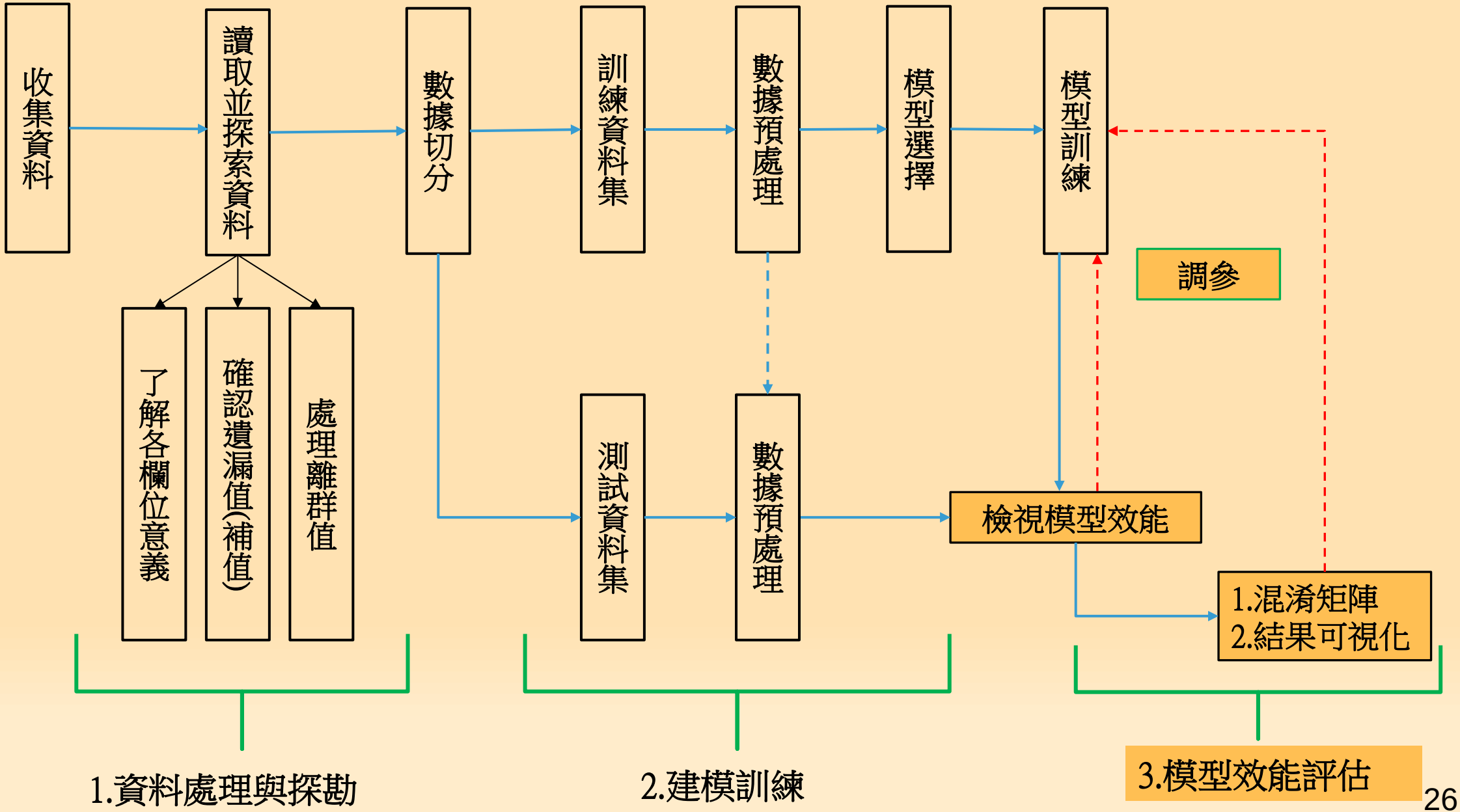
3.Logistic Regression

```
array([[11,  3],
       [ 7, 14]], dtype=int64)
```

	precision	recall	f1-score	support
0	0.61	0.79	0.69	14
1	0.82	0.67	0.74	21
accuracy			0.71	35
macro avg	0.72	0.73	0.71	35
weighted avg	0.74	0.71	0.72	35

4. Support Vector Machine

機器學習_流程圖


收集資料
讀取並探索資料
了解各欄位意義
確認遺漏值(補值)
處理離群值
數據切分
訓練資料集
測試資料集
數據預處理
數據預處理
模型選擇
模型訓練
調參
檢視模型效能
1.混淆矩陣
2.結果可視化
1.資料處理與探勘
2.建模訓練
3.模型效能評估

GridSearchCV 網格調參結果

資料切分 7：3, cross-validation為5, 使用GridSearchCV調整超參數，
搭配SVM_kernel為poly並得到最高正確率有0.8。

kernel	C	gamma	accuracy
linear	10	0.01	0.71
sigmoid	10	0.10	0.74
rbf	10	0.10	0.77
poly	1	0.10	0.80

```
array([[ 9,  5],
       [ 2, 19]], dtype=int64)
```

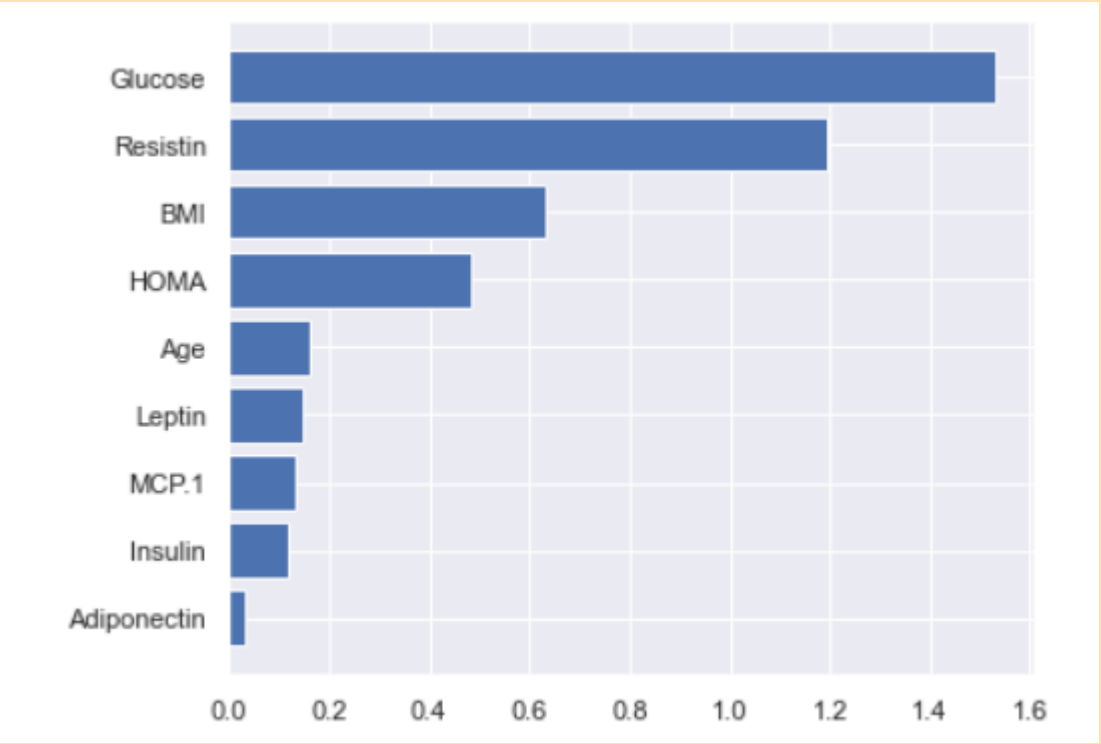
	precision	recall	f1-score	support
0	0.82	0.64	0.72	14
1	0.79	0.90	0.84	21
accuracy			0.80	35
macro avg	0.80	0.77	0.78	35
weighted avg	0.80	0.80	0.79	35

kernel = poly, C=1, gamma = 0.1

目錄

結論

結論



模型重要變數列表

1.經過超參數調整後得到最佳模型正確率80%

2.重要變數列表顯示罹患乳癌重要特徵前三位為：
Glucose(葡萄糖)、Resistin(阻抗素)與 BMI(身體質量指數)。

3.資料分析印證：相關係數分析顯示Glucose(最高正相關)和BMI(最高負相關)、T檢定(Glucose顯著)與重要變數列表趨勢一致。

3.文獻M. Patrício et al.(2018)結論指出，基於Resistin 、 Glucose BMI和Age，可以在測試數據集上預測女性乳腺癌的存在[1]。

Thanks for listening