Data Mining 期末報告

Heart Disease 分析

資工 3A 00557019 何寬宥

資工 2A 00657049 黃姿涵

資工 3A 00557039 王皓宇（已退選）

指導老師：蔡宇軒 助理教授

目錄

圖目錄

表目錄

壹、 動機

根據世界衛生組織的調查顯示，心臟方面的疾病在近十年以來都位居全球十大死因的第一名。在 2016 年的調查中，因全球心臟疾病死亡的人更是佔了死亡人口的近 27%。而在臺灣，自民國 96 年起，心臟疾病也位居國人十大死因的第二名，僅次於癌症。2017 年衛服部的統計資料中，因心臟疾病死亡的人佔了全台死亡人口近 12%。由此可以推估，心臟疾病與人的死因息息相關，故我們希望透過分析心臟方面的疾病，藉此推估具有哪些特質的人較易因心臟疾病致死，並探討透過哪種分析方式，具有較高的準確率。

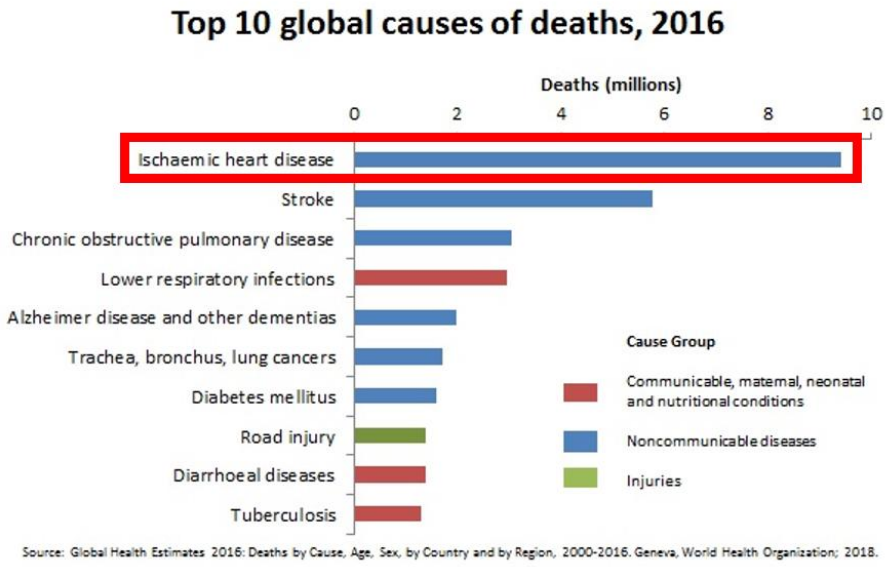


圖 1-1:2016 年全球十大死因統計

表 1 十大死因死亡人數及死亡率

	死亡人數(人)		死亡率（每十萬人口）				標準化死亡率（每十萬人口）		
	106年	較上年增減%	105年順位	106年順位	106年	較上年增減%	順位	106年	較上年增減%
所有死亡原因	171,857	-0.3			729.6	-0.5		424.3	-3.4
癌症	48,037	0.6	1	1	203.9	0.4	1	123.4	-2.7
心臟疾病（高血壓性疾病除外）	20,644	-0.8	2	2	87.6	-1.0	2	48.5	-3.6
肺炎	12,480	2.2	3	3	53.0	2.1	4	26.5	-1.5
腦血管疾病	11,755	-0.8	4	4	49.9	-1.0	3	27.5	-3.8
糖尿病	9,845	-1.2	5	5	41.8	-1.4	5	23.5	-4.1
事故傷害	6,965	-3.3	6	6	29.6	-3.3	6	21.9	-5.2
慢性下呼吸道疾病	6,260	-7.8	7	7	26.6	-8.0	7	13.3	-11.9
高血壓性疾病	6,072	3.2	8	8	25.8	3.2	8	13.3	-1.5
腎炎、腎病症候群及腎病變	5,381	3.0	9	9	22.8	2.7	10	12.4	0.0
慢性肝病及肝硬化	4,554	-3.9	10	10	19.3	-4.0	9	12.6	-6.0

圖 1-2:民國 106 年臺灣十大死因統計

貳、 資料

一、 資料來源

加州大學灣爾分校 於 1988 年公開在網路上的 Data Set。[1]

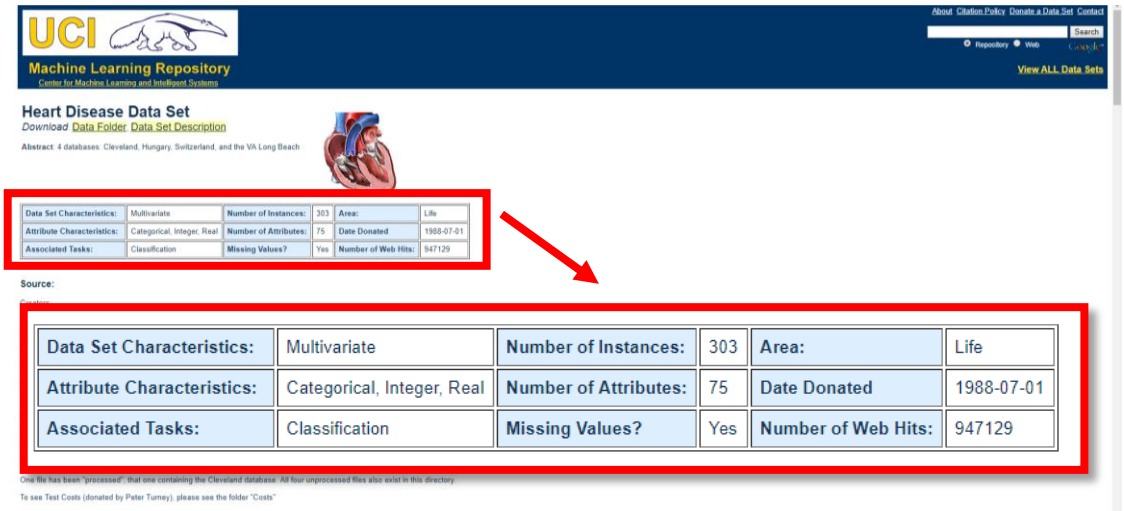


圖 2-1:資料來源圖

二、 資料特性

屬性	描述
age	年齡
sex	性別（0=女性，1=男性）
cp	胸痛經歷（0=典型心絞痛；1=非典型心絞痛；2=非心絞痛；3=無症狀）
trestbps	靜態血壓（mm Hg）
chol	膽固醇（mg / dl）
fbs	空腹血糖（ > 120 mg / dl，1=是；0=否）
restecg	靜態心電圖測量（0=正常；1=有 ST-T 波異常；2=依 Estes 標準顯示有可能或明確的左心室肥厚）
thalach	最大心率
exang	運動誘發心絞痛（0=否；1=是）
oldpeak	運動相對於休息引起的 ST 段壓低
slope	運動時 ST 段峰值的斜率（0=上升；1=平坦；2=下降）
ca	主要血管數量(0-3)
thal	地中海貧血（1=正常；2=固定缺陷；3=可逆缺陷）
target	心臟病（0=否；1=是）

表 2-1:資料重要屬性表

圖 2-2:相關係數圖

三、 資料前處理

1. 虛擬變數(dummy variable)

在這筆資料當中，其中有幾個欄位是虛擬變數，由於這些值大小是不可量化的，若丟入迴歸分析(線性、羅吉斯……)中，將會產生出數值的錯誤，導致降低預測效果，所以必須將這些屬性拆解成不同的屬性，並以二元變數(0 or 1)來表示，以避開數值預測錯誤。

cp	restecg	slope	thal
cp_typical_angina	restecg_normal	slope_upsloping	thal_normal
cp_atypical_angina	restecg_ST-T_wave_abnormality	slope_flat	thal_fixed_defect
cp_non-anginal_pain	restecg_left_ventricular_hypertrophy	slope_downsloping	thal_reversable_defect
cp_asymptomatic			

表 2-1:虛擬變數拆解後的變數名稱

2. 離散化

由於在做像是關聯法則這種的學習模型當中，不得使用連續變數來當作參數，於是我們在正規化之後的檔案當中，將[0, 1]區間中拆分成五等分，而屬於這區間中的資料分別標記上 0 ~ 4，以利於未來需要用到離散數值的資料。

3. 資料正規化

3.1 提升模型的收斂速度

在構機器學習模型時，通常都會利用梯度下降法來計算損失函數的最佳解，假設現在有兩個特徵值 X1 於[0, 10000]與 X2 於[0, 1]會導致損失函數的等高線呈現窄長型(如圖 2-2)，導致需要更多的迭代步驟，也可能導致無法收斂，因此若將資料標準化，則能減少梯度下降法的收斂時間。

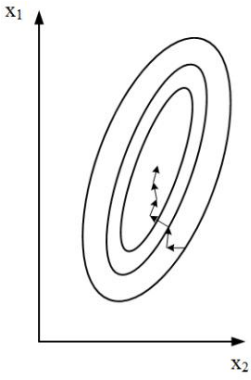


圖 2-2:未標準化梯度下降過程

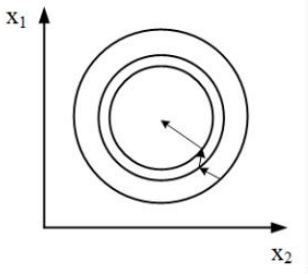


圖 2-3:標準化梯度下降過程

3.2 提高模型的精確度

將特徵值 x1 及 x2 丟入一些需計算樣本間距離的機器學習模型(K-means、K Nearest Neighbor)當中，則 x2 的影響可能遠大於 x1，而實際上 x1 的指標意義以及重要性高於 x2，這將導致分析結果失真，因此若將資料標準化，可讓每個特徵值對結果做出相近程度的貢獻。

3.3 方法

在這份資料當中，我們使用的是「Min-Max Normalization」，是將資料等比例縮放到[0, 1]區間中，可利用下列公式進行轉換：

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad \epsilon [0,1]$$

其中 X_{max} 及 X_{min} 分別為資料中的最小值與最大值。

參、 監督式學習

一、 LibSVM

1. 方法簡介

LibSVM 是我們從 weka 當中額外安裝的插件，因為我認為這也是平常在做機器學習當中經常被使用到的演算法，在這個算法當中，無法在原本平面當中做線性切割的點轉換到更高的維度中(如圖 3-1)，並找出 Hyperplane 將類切分成兩類，而此超平面轉換回原本空間可如圖 3-2 所示。此外，這個演算法當中最重要的參數有 3 個，C（cost，分割平面的錯誤容忍度，數值越大容忍度越低)、gamma（轉換後的空間分布，數值越大分布越小)及 kernel(線性轉換方式)，而前二者我們分別使用 C=100，gamma=0.01(數字太大或太小皆會造成 overfitting 或 underfitting)，kernel 則使用的是 RBF(Radial Basis Function) kernal，為一種非線性的轉換方式。

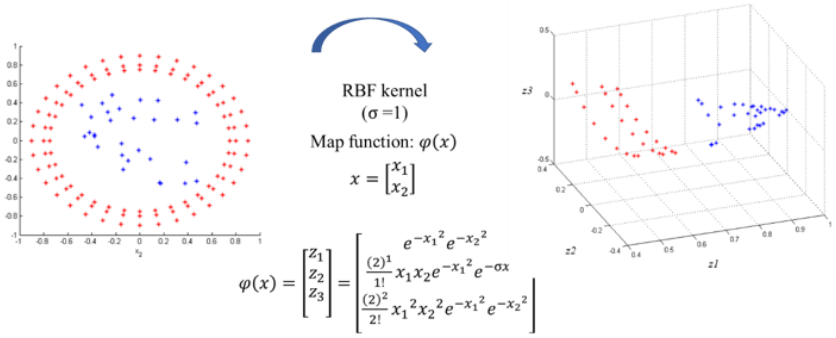


圖 3-1:將點轉換到更高維度

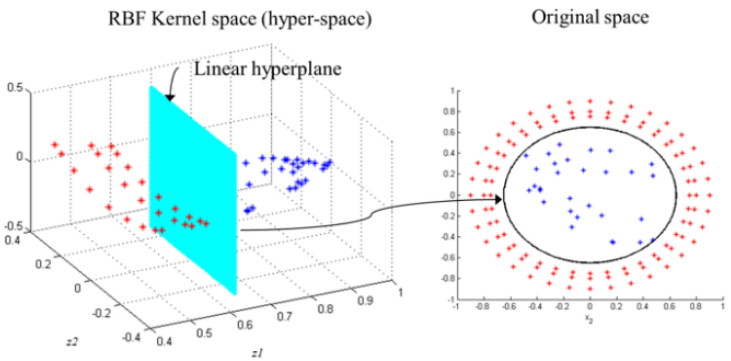


圖 3-2:在高維度中的超平面轉換回原本的空間

2. 分析結果

```
=== Classifier model (full training set) ===

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         256               84.4884 %
Incorrectly Classified Instances        47               15.5116 %
Kappa statistic                          0.6849
Mean absolute error                      0.1551
Root mean squared error                  0.3938
Relative absolute error                 31.2673 %
Root relative squared error             79.078  %
Total Number of Instances              303

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.783    0.103    0.864      0.783    0.821      0.687    0.840     0.775     0
                 0.897    0.217    0.831      0.897    0.863      0.687    0.840     0.802     1
Weighted Avg.    0.845    0.165    0.846      0.845    0.844      0.687    0.840     0.790

=== Confusion Matrix ===

   a   b   <-- classified as
 108  30 |   a = 0
  17 148 |   b = 1
```

圖 3-3:LibSVM 的分析結果

二、 IBK

1. 方法簡介

在 weka 中 IBk 是我們平常所稱的 K-nearest neighbors 算法，這種算法是一種非參數和惰性學習的方法，意味著模型結構完全以數據集決定，不須用任何數學理論假設來生成模型，可以快速的創建模型，但是測試階段會非常的緩慢，且昂貴的測試階段需要花費更多的時間和記憶，而最壞的情況下，可能需要掃描所有數據點。而這個算法的運作方式是藉由調整 k 來找到最近的 k 個鄰居，並讓這些鄰居投票給標籤，進而達成分類。為了找到最近的相似點，可以使用距離的測量公式找到點之間的距離，例如歐基里德距離、曼哈頓距離……，而我們在這次的探勘中使用的 k 值及距離公式分別為 k = 3 及歐基里德距離。

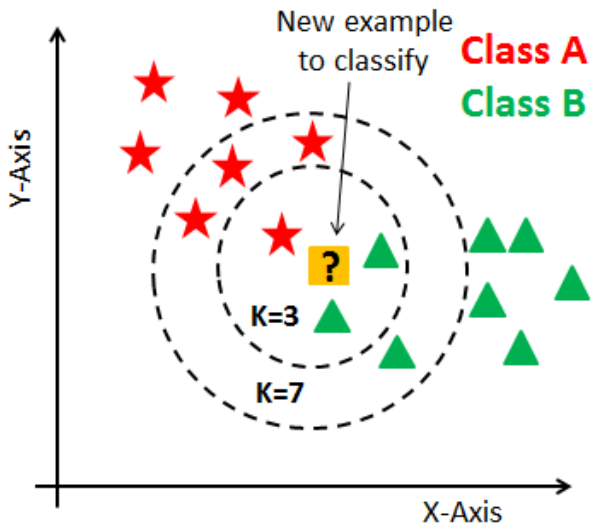


圖 3-4:KNN 算法示意圖

2. 分析結果

```
IB1 instance-based classifier
using 3 nearest neighbour(s) for classification


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         252               83.1683 %
Incorrectly Classified Instances        51               16.8317 %
Kappa statistic                          0.6613
Mean absolute error                      0.2448
Root mean squared error                  0.3923
Relative absolute error                 49.3551 %
Root relative squared error             78.7575 %
Total Number of Instances              303

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.826    0.164    0.809      0.826    0.817      0.661    0.846     0.808     0
                 0.836    0.174    0.852      0.836    0.844      0.661    0.846     0.813     1
Weighted Avg.    0.832    0.169    0.832      0.832    0.832      0.661    0.846     0.811

=== Confusion Matrix ===

   a   b   <-- classified as
 114  24 |   a = 0
  27 138 |   b = 1
```

圖 3-5:IBk 的分析結果

三、 Logistic

1. 方法簡介

Logistic 也是一種利用回歸來分類的算法但是不同於線性回歸的點在於，Logistic 在參數型學習當中通常都是用最大概似函數估計法(Maximum Likelihood Estimation, MLE)以及伯努利分布的機率密度函數做參數估計，並利用 $Pr[1|a_1,a_2,...,a_k]=1/(1+exp(-w_0-w_1a_1-...-w_ka_k))$輸出於(0,1)之間，中間值為 0.5 使小於 0.5 的值分為 A 類，其餘分為 B 類。

2. 分析結果

Odds Ratios...

Variable	Class 0
age	1.0279
sex	4.5492
trestbps	6.1088
chol	6.6677
fbs	0.8383
thalach	0.106
exang	2.1449
oldpeak	20.7722
ca	28.0143
cp_asymptomatic	0.3694
cp_atypical_angina	1.0374
cp_non-anginal_pain	0.3964
cp_typical_angina	2.7731
thal_fixed_defect	0.1567
thal_normal	0.1629
thal_reversable_defect	0.623
slope_downsloping	0.6502
slope_flat	1.6335
slope_upsloping	0.7954
restecg_ST-T_wave_abnormality	0.7414
restecg_left_ventricular_hypertrophy	1.7292
restecg_normal	1.3112

圖 3-6:勝算比

Coefficients...

Variable	Class 0
age	0.0275
sex	1.5149
trestbps	1.8097
chol	1.8973
fbs	-0.1764
thalach	-2.2442
exang	0.7631
oldpeak	3.0336
ca	3.3327
cp_asymptomatic	-0.9959
cp_atypical_angina	0.0367
cp_non-anginal_pain	-0.9253
cp_typical_angina	1.02
thal_fixed_defect	-1.8533
thal_normal	-1.8147
thal_reversable_defect	-0.4733
slope_downsloping	-0.4305
slope_flat	0.4907
slope_upsloping	-0.2289
restecg_ST-T_wave_abnormality	-0.2992
restecg_left_ventricular_hypertrophy	0.5477
restecg_normal	0.2709
Intercept	-1.2116

圖 3-7:係數

```
Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         253               83.4983 %
Incorrectly Classified Instances        50               16.5017 %
Kappa statistic                          0.6657
Mean absolute error                      0.2344
Root mean squared error                  0.3652
Relative absolute error                 47.246  %
Root relative squared error             73.3347 %
Total Number of Instances              303

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.790    0.127    0.838      0.790    0.813      0.667    0.880     0.828     0
               0.873    0.210    0.832      0.873    0.852      0.667    0.880     0.872     1
Weighted Avg.  0.835    0.172    0.835      0.835    0.834      0.667    0.880     0.852

=== Confusion Matrix ===

   a   b   <-- classified as
 109  29 |   a = 0
  21 144 |   b = 1
```

圖 3-8:Logistic 的分析結果

四、 MultilayerPerceptron

1. 方法簡介

假如數據是線性可分得，而存在一個或多個超平面可將數據分離，如果加權和大於 0，分為第一類，否則分為第二類。在 weka 當中，多層感知器更新權重的方式為利用反向傳播，對神經網路中的所有權重計算 loss function 的梯度，而這個梯度會反饋回更新權值以最小化 loss function，此外更新權重的速度取決於學習率，學習率過大可能會導致震盪，過小可能會導致收斂過慢，還有神經元的各數若過多，可能會導致 overfitting。而學習率我們使用預設的 0.3，神經元的各數也是使用預設的 a（attributes + classes）/ 2)。

2. 分析結果

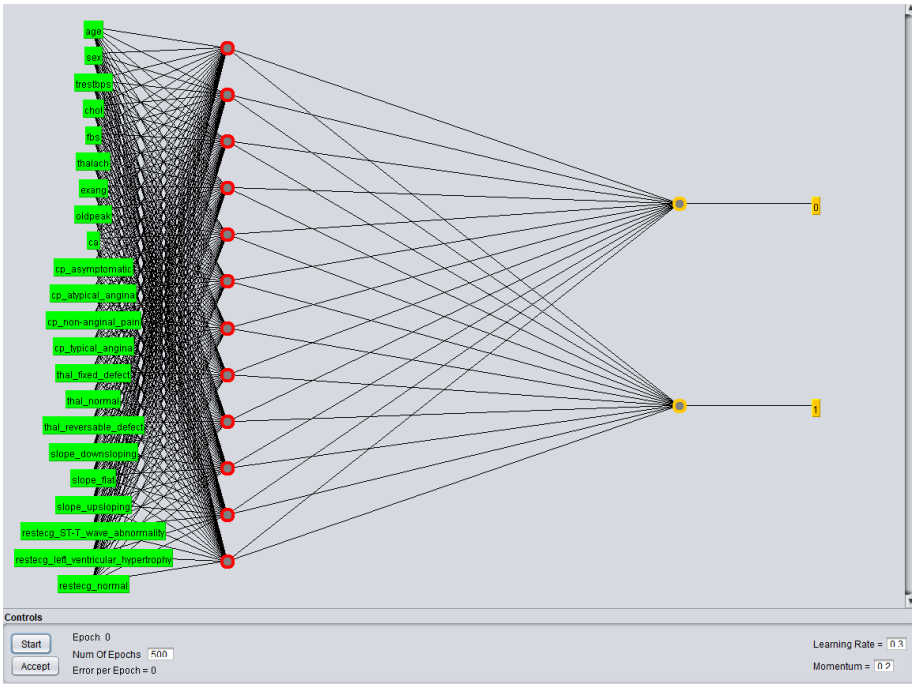


圖 3-9:多層感知器神經元圖

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         243               80.198  %
Incorrectly Classified Instances        60               19.802  %
Kappa statistic                          0.6017
Mean absolute error                      0.2014
Root mean squared error                  0.4119
Relative absolute error                 40.5987 %
Root relative squared error             82.6995 %
Total Number of Instances              303

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.797    0.194      0.775      0.797    0.786      0.602    0.869     0.843     0
                 0.806    0.203      0.826      0.806    0.816      0.602    0.869     0.874     1
Weighted Avg.    0.802    0.199      0.803      0.802    0.802      0.602    0.869     0.860

=== Confusion Matrix ===

   a   b   <-- classified as
 110  28 |   a = 0
  32 133 |   b = 1
```

圖 3-10:MultilayerPerceptron 的分析結果

五、 J48

1. 方法簡介

凡是萬物都會傾向往自由度較大的地方走，就如同將某個只裝滿特定氣體的瓶子，將它瓶蓋打開便會散發到世界各地，亂度(熵)將會變大，而決策樹的根就如同此述，亂度非常的大。而要如何重新將這些資料放回原本正確的資料裡，通常都會計算在各點中的資訊量(Information gain)，而常用的方法為熵，我們要找出當下亂度最小的地方，當作樹的下個節點，最後循環做此方法。

2. 分析結果

```
thal = 0: 0 (2.0/1.0)
thal = 1
|   ca = 0
|   |   exang = 0: 1 (5.0)
|   |   exang = 1: 0 (3.0/1.0)
|   ca = 1: 0 (4.0)
|   ca = 2: 0 (4.0)
|   ca = 3: 0 (2.0)
|   ca = 4: 0 (0.0)
thal = 2
|   ca = 0: 1 (114.0/12.0)
|   ca = 1
|   |   sex = 0: 1 (13.0/1.0)
|   |   sex = 1
|   |   |   cp = 0: 0 (9.0)
|   |   |   cp = 1: 0 (2.0/1.0)
|   |   |   cp = 2: 1 (2.0)
|   |   |   cp = 3: 1 (3.0/1.0)
|   ca = 2
|   |   exang = 0
|   |   |   oldpeak = 1: 1 (7.0/2.0)
|   |   |   oldpeak = 2: 1 (2.0)
|   |   |   oldpeak = 3: 0 (2.0)
|   |   |   oldpeak = 4: 1 (0.0)
|   |   |   oldpeak = 5: 1 (0.0)
|   |   exang = 1: 0 (3.0)
|   ca = 3: 0 (6.0/1.0)
|   ca = 4: 1 (3.0)
thal = 3
|   cp = 0: 0 (78.0/7.0)
|   cp = 1
|   |   ca = 0: 1 (6.0/2.0)
|   |   ca = 1: 0 (2.0)
|   |   ca = 2: 1 (0.0)
|   |   ca = 3: 1 (0.0)
|   |   ca = 4: 1 (1.0)
|   cp = 2
|   |   slope = 0: 1 (1.0)
|   |   slope = 1: 0 (14.0/4.0)
|   |   slope = 2: 1 (7.0/1.0)
|   cp = 3: 1 (8.0/3.0)

Number of Leaves  :     31

Size of the tree :      42
```

圖 3-11:決策樹(文字版)

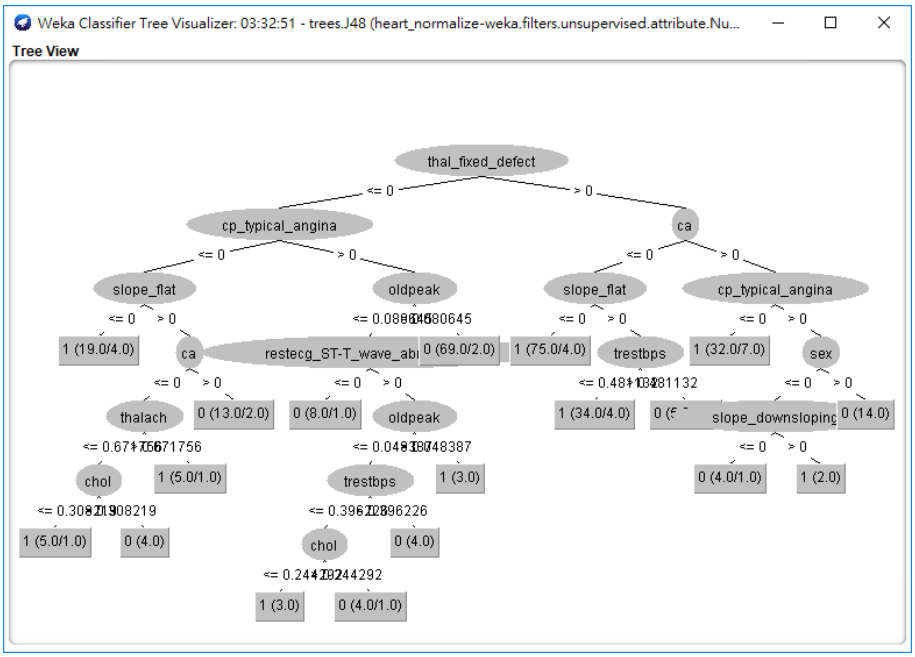


圖 3-12:決策樹

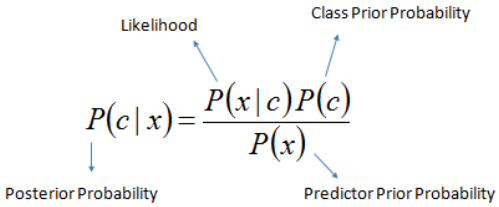
```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         227               74.9175 %
Incorrectly Classified Instances        76               25.0825 %
Kappa statistic                          0.4925
Mean absolute error                      0.2817
Root mean squared error                  0.4721
Relative absolute error                 56.7864 %
Root relative squared error             94.7897 %
Total Number of Instances              303

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.703     0.212     0.735      0.703     0.719      0.493    0.727     0.651     0
               0.788     0.297     0.760      0.788     0.774      0.493    0.727     0.692     1
Weighted Avg.  0.749     0.258     0.749      0.749     0.749      0.493    0.727     0.674

=== Confusion Matrix ===

   a   b   <-- classified as
  97  41 |   a = 0
  35 130 |   b = 1
```

圖 3-13:J48 的分析結果

六、 Naïve Bayes

1. 方法簡介

簡單貝氏分類器為直接假設所有的隨機變數之間具有條件獨立的情況，因此直接利用條件機率相乘的方式，計算出聯合機率分布（如圖 3-14），進而去比較在測試資料中分類於各類的聯合機率為何，選出較大的機率，並將測試資料標籤於該類。此外，若屬性為離散的，將直接計算該類的機率為何，而屬性為連續變數的話，將以高斯(常態)機率密度函式 $\{P(X = x \mid C = c) = g(x; \mu_c, \sigma_c),$ $where\ g(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\}$ 帶入平均值以及標準差，得出該點的機率密度。



$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

圖 3-14:簡單貝氏分類公式

2. 分析結果

Attribute	Class 0 (0.46)	1 (0.54)
age		
mean	0.5748	0.4885
std. dev.	0.1659	0.1985
weight sum	138	165
precision	0.025	0.025
sex		
0	25.0	73.0
1	115.0	94.0
[total]	140.0	167.0
trestbps		
mean	0.3798	0.3328
std. dev.	0.1759	0.1516
weight sum	138	165
precision	0.0208	0.0208
chol		
mean	0.2858	0.2652
std. dev.	0.1124	0.1218
weight sum	138	165
precision	0.0066	0.0066
fbs		
0	117.0	143.0
1	23.0	24.0
[total]	140.0	167.0
thalach		
mean	0.5198	0.6676
std. dev.	0.1723	0.1462
weight sum	138	165
precision	0.0111	0.0111

圖 3-15:機率分布(1)

exang		
0	63.0	143.0
1	77.0	24.0
[total]	140.0	167.0
oldpeak		
mean	0.257	0.0946
std. dev.	0.209	0.1249
weight sum	138	165
precision	0.0256	0.0256
ca		
0	46.0	131.0
0.25	45.0	22.0
0.5	32.0	8.0
0.75	18.0	4.0
1	2.0	5.0
[total]	143.0	170.0
cp_asymptomatic		
0	132.0	150.0
1	8.0	17.0
[total]	140.0	167.0
cp_atypical_angina		
0	130.0	125.0
1	10.0	42.0
[total]	140.0	167.0
cp_non-anginal_pain		
0	121.0	97.0
1	19.0	70.0
[total]	140.0	167.0

圖 3-16:機率分布(2)

```
cp_typical_angina
  0                                35.0   127.0
  1                               105.0    40.0
  [total]                         140.0   167.0

thal_fixed_defect
  0                               103.0    36.0
  1                                37.0   131.0
  [total]                         140.0   167.0

thal_normal
  0                               127.0   160.0
  1                                13.0     7.0
  [total]                         140.0   167.0

thal_reversable_defect
  0                                50.0   138.0
  1                                90.0    29.0
  [total]                         140.0   167.0

slope_downsloping
  0                               104.0    59.0
  1                                36.0   108.0
  [total]                         140.0   167.0

slope_flat
  0                                48.0   117.0
  1                                92.0    50.0
  [total]                         140.0   167.0
```

圖 3-17:機率分布(3)

```
slope_upsloping
  0                               127.0   157.0
  1                                13.0    10.0
  [total]                         140.0   167.0

restecg_ST-T_wave_abnormality
  0                                83.0    70.0
  1                                57.0    97.0
  [total]                         140.0   167.0

restecg_left_ventricular_hypertrophy
  0                               136.0   165.0
  1                                 4.0     2.0
  [total]                         140.0   167.0

restecg_normal
  0                                60.0    98.0
  1                                80.0    69.0
  [total]                         140.0   167.0
```

圖 3-18:機率分布(4)

```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         259               85.4785 %
Incorrectly Classified Instances        44               14.5215 %
Kappa statistic                          0.7079
Mean absolute error                      0.1715
Root mean squared error                  0.3607
Relative absolute error                 34.5689 %
Root relative squared error             72.4254 %
Total Number of Instances              303

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.855     0.145     0.831     0.855     0.843      0.708    0.911     0.890     0
               0.855     0.145     0.876     0.855     0.865      0.708    0.911     0.920     1
Weighted Avg.  0.855     0.145     0.855     0.855     0.855      0.708    0.911     0.906

=== Confusion Matrix ===

   a   b   <-- classified as
 118  20 |   a = 0
  24 141 |   b = 1
```

圖 3-19:Naïve Bayes 的分析結果

七、 結果探討

在這次的資料分析當中，我們使用了 5 種 weka 原本就有的分類器以及 1 種而外安裝的分類器，此外訓練及測試的資料皆是使用 K-fold (K=10)來拆分，而以此次學習的結果當中來看，不管是利用線性回歸(Logistic、MultilayerPerceptron)、SVM、機率模型(Naïve Bayes)來創建得模型分類正確率都有 80%以上，都還不錯，而只有使用 J48 (Decision Tree)時準確率未達 80%，且以樹的分配感覺來看，有點 overfitting 的感覺，不過可能與原本資料屬性有所關係，因為此次的資料有不少的連續變數，我們是將它以某種區間的值拆分成 5 個離散數值後將他丟進分類器當中，這已經比原本使用連續變數直接丟進去計算高出了 4-5%了，不過也許有可能是因為這些數值拆分的不夠細膩，使之無法學習出更好的節點。

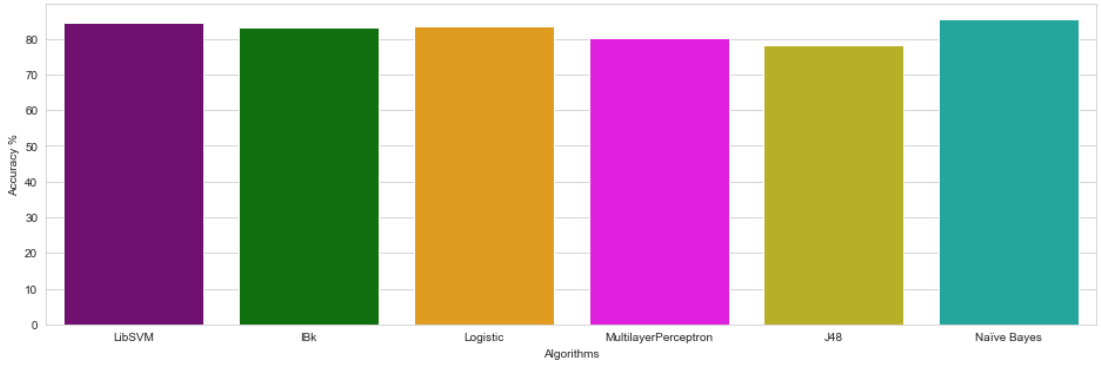


圖 3-20:分析結果正確率柱狀圖

肆、 非監督式學習

一、 Simple K-Means

1. 方法簡介

在此種集群方式通常有 6 個步驟：

(1) 先決定要分成多少(k)群

(2) 在特徵空間中隨機給出 k 個群心

(3) 對於每筆數據計算對於這 k 個群心的歐基里德距離(當然可以換成其他距離公式，例如曼哈頓距離……，不過我們在分析中使用歐基里德距離)

(4) 將每筆資料標記成距離最近的那個群心

(5) 將所有位於這個群心內的資料進行平均，重新派分群心

(6) 重複步驟 3-5，直到所有群心不再變動

2. 分析結果

```
Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       152 ( 50%)
1       151 ( 50%)


Class attribute: target
Classes to Clusters:

   0   1  <-- assigned to cluster
  23 115 | 0
 129  36 | 1

Cluster 0 <-- 1
Cluster 1 <-- 0

Incorrectly clustered instances :       59.0     19.4719 %
```

圖 4-1:Simple K-Means 的分析結果(標記)

Final cluster centroids:

Attribute	Full Data (303.0)	Cluster# 0 (152.0)	1 (151.0)
age	0.5285	0.487	0.5702
sex	0.6832	0.5658	0.8013
trestbps	0.3549	0.3408	0.3692
chol	0.2746	0.2674	0.2818
fbs	0.1485	0.1447	0.1523
thalach	0.6004	0.6868	0.5134
exang	0.3267	0.0987	0.5563
oldpeak	0.1677	0.0866	0.2493
ca	0.1823	0.1168	0.2483
cp_asymptomatic	0.0759	0.0789	0.0728
cp_atypical_angina	0.165	0.2961	0.0331
cp_non-anginal_pain	0.2871	0.4671	0.106
cp_typical_angina	0.4719	0.1579	0.7881
thal_fixed_defect	0.5479	0.9013	0.1921
thal_normal	0.0594	0	0.1192
thal_reversable_defect	0.3861	0.0921	0.6821
slope_downsloping	0.4686	0.7632	0.1722
slope_flat	0.462	0.1842	0.7417
slope_upsloping	0.0693	0.0526	0.0861
restecg_ST-T_wave_abnormality	0.5017	0.5658	0.4371
restecg_left_ventricular_hypertrophy	0.0132	0.0066	0.0199
restecg_normal	0.4851	0.4276	0.543

圖 4-2:Simple K-Means 的分析結果(群心)

二、 結果探討

在集群分析當中，我們可以藉由分析結果的群心來推得有哪些屬性更傾向於有心臟病的人，而我們把這些具有較大差異的相關係數(如圖 4-3)以及分群結果(如圖 4-4)。由此得知，具有「心率較大、運動較不會誘發心絞痛、比較不因為運動而引發的 ST 段壓低、主血管數量較少、有典型或非典型心絞痛、有固定缺陷的地中海貧血、運動時 ST 段峰值斜率下降、心電圖測量中具有 ST-T 波異常」狀況的人會比較傾向於得到心臟病。

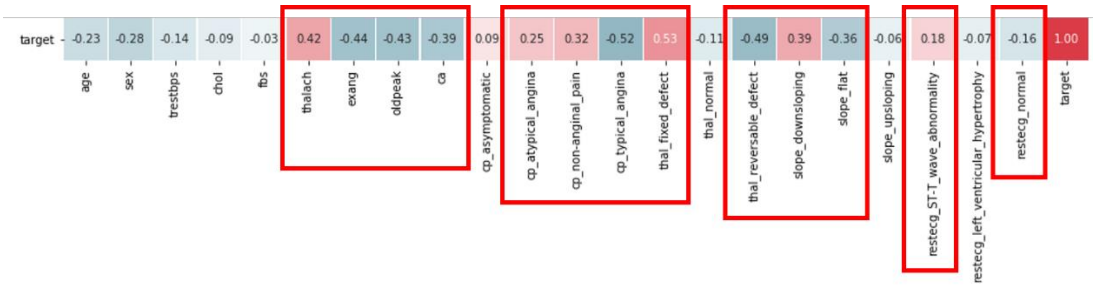


圖 4-3:相關係數中與目標有較大的正負相關係數值

thalach	0.6004	0.6868	0.5134
exang	0.3267	0.0987	0.5563
oldpeak	0.1677	0.0866	0.2493
ca	0.1823	0.1168	0.2483
cp_atypical_angina	0.165	0.2961	0.0331
cp_non-anginal_pain	0.2871	0.4671	0.106
cp_typical_angina	0.4719	0.1579	0.7881
thal_fixed_defect	0.5479	0.9013	0.1921
thal_reversable_defect	0.3861	0.0921	0.6821
slope_downsloping	0.4686	0.7632	0.1722
slope_flat	0.462	0.1842	0.7417
restecg_ST-T_wave_abnormality	0.5017	0.5658	0.4371
restecg_normal	0.4851	0.4276	0.543

圖 4-4:在集群分析中與群心有較大差異屬性

伍、 關聯法則

一、 方法簡介

關聯規則採用先驗演算法(Apriori Algorithm)，以廣度優先進行搜尋並採用迭代的方式找出各候選項集的支持度，利用支持度對候選項集進行剪枝，降低產生頻繁項集的計算複雜度。篩選出符合最小支持度和最小置信度的集合。

Apriori Algorithm 有兩大定律：

定律 1: 假設一個集合{A,B}≧最小支持度(Min_Support)，則他的子集{A},{B}出現次數必定≧最小支持度。

定律 2: 假設集合{A}出現次數≦最小支持度，則他的任何集合如{A,B}出現的次數必定≦最小支持度

評估指標為:

1. 支持度(Support): 表示為 item-set 在整個資料中出現的頻率
   Support(X) = number(X) / number(AllSamples)
   Support(X,Y) = number(XT) / number(AllSamples)
2. 置信度(Confidence): 表示當事件 X 發生的情況下，同時發生 Y 的可能性
   Confidence(X→Y) = P(Y|X) = P(X∩Y) / P(X)

二、 分析結果

當 Min_Support = 0.2 時，confidence 最高的前 30 個關聯：

```
Best rules found:

 1. sex=0 target=1 72 ==> thal=2 69    <conf:(0.96)> lift:(1.75) lev:(0.1) [29] conv:(8.14)
 2. sex=0 fbs=0 target=1 66 ==> thal=2 63    <conf:(0.95)> lift:(1.74) lev:(0.09) [26] conv:(7.46)
 3. sex=0 exang=0 target=1 64 ==> thal=2 61    <conf:(0.95)> lift:(1.74) lev:(0.09) [25] conv:(7.23)
 4. trestbps=2 ca=0 target=1 64 ==> exang=0 61    <conf:(0.95)> lift:(1.42) lev:(0.06) [17] conv:(5.23)
 5. chol=2 oldpeak=1 ca=0 72 ==> fbs=0 68    <conf:(0.94)> lift:(1.11) lev:(0.02) [6] conv:(2.14)
 6. slope=2 ca=0 thal=2 70 ==> target=1 66    <conf:(0.94)> lift:(1.73) lev:(0.09) [27] conv:(6.38)
 7. cp=0 ca=0 65 ==> fbs=0 61    <conf:(0.94)> lift:(1.1) lev:(0.02) [5] conv:(1.93)
 8. oldpeak=1 slope=2 ca=0 thal=2 65 ==> target=1 61    <conf:(0.94)> lift:(1.72) lev:(0.08) [25] conv:(5.92)
 9. exang=0 oldpeak=1 ca=0 thal=2 79 ==> target=1 74    <conf:(0.94)> lift:(1.72) lev:(0.1) [30] conv:(6)
10. chol=2 exang=0 ca=0 74 ==> fbs=0 69    <conf:(0.93)> lift:(1.1) lev:(0.02) [5] conv:(1.83)
11. fbs=0 exang=0 oldpeak=1 ca=0 thal=2 72 ==> target=1 67    <conf:(0.93)> lift:(1.71) lev:(0.09) [27] conv:(5.47)
12. slope=2 ca=0 thal=2 70 ==> oldpeak=1 65    <conf:(0.93)> lift:(1.41) lev:(0.06) [19] conv:(4)
13. restecg=1 oldpeak=1 thal=2 69 ==> fbs=0 64    <conf:(0.93)> lift:(1.09) lev:(0.02) [5] conv:(1.71)
14. restecg=1 exang=0 thal=2 target=1 68 ==> fbs=0 63    <conf:(0.93)> lift:(1.09) lev:(0.02) [5] conv:(1.68)
15. fbs=0 restecg=1 thal=2 target=1 68 ==> exang=0 63    <conf:(0.93)> lift:(1.38) lev:(0.06) [17] conv:(3.7)
16. sex=0 exang=0 thal=2 66 ==> target=1 61    <conf:(0.92)> lift:(1.7) lev:(0.08) [25] conv:(5.01)
17. slope=2 ca=0 thal=2 target=1 66 ==> oldpeak=1 61    <conf:(0.92)> lift:(1.41) lev:(0.06) [17] conv:(3.78)
18. oldpeak=1 ca=0 thal=2 91 ==> target=1 84    <conf:(0.92)> lift:(1.7) lev:(0.11) [34] conv:(5.18)
19. exang=1 target=0 76 ==> cp=0 70    <conf:(0.92)> lift:(1.95) lev:(0.11) [34] conv:(5.73)
20. restecg=1 exang=0 thal=2 76 ==> fbs=0 70    <conf:(0.92)> lift:(1.08) lev:(0.02) [5] conv:(1.61)
21. chol=2 thal=3 74 ==> sex=1 68    <conf:(0.92)> lift:(1.35) lev:(0.06) [17] conv:(3.35)
22. restecg=1 thal=2 target=1 74 ==> fbs=0 68    <conf:(0.92)> lift:(1.08) lev:(0.02) [4] conv:(1.57)
23. restecg=1 thal=2 target=1 74 ==> exang=0 68    <conf:(0.92)> lift:(1.36) lev:(0.06) [18] conv:(3.45)
24. restecg=1 thal=2 86 ==> fbs=0 79    <conf:(0.92)> lift:(1.08) lev:(0.02) [5] conv:(1.6)
25. exang=0 ca=0 thal=2 97 ==> target=1 89    <conf:(0.92)> lift:(1.68) lev:(0.12) [36] conv:(4.91)
26. sex=0 target=1 72 ==> fbs=0 66    <conf:(0.92)> lift:(1.08) lev:(0.02) [4] conv:(1.53)
27. fbs=0 oldpeak=1 ca=0 thal=2 83 ==> target=1 76    <conf:(0.92)> lift:(1.68) lev:(0.1) [30] conv:(4.73)
28. age=2 71 ==> fbs=0 65    <conf:(0.92)> lift:(1.08) lev:(0.01) [4] conv:(1.51)
29. restecg=1 exang=0 slope=2 71 ==> oldpeak=1 65    <conf:(0.92)> lift:(1.39) lev:(0.06) [18] conv:(3.48)
30. chol=2 ca=0 94 ==> fbs=0 86    <conf:(0.91)> lift:(1.07) lev:(0.02) [5] conv:(1.55)
```

圖 5-1:關聯法則圖(Min_Support = 0.2)

三、 結果探討

根據前三十筆關聯法則的分析結果，由第 6.8.9.11.16.18.25.27 筆資料，我們可以看出具有地中海貧血固定缺陷、主要血管數量為 0、運動相對於休息引起的 ST 段壓低範圍 1.3~2.4 之間的人較容易罹患心臟病。由第 1.2.3 筆資料可以看出地中海貧血固定缺陷較容易發生在女性身上。由第 12.17.29 筆資料可以看出運動時 ST 段峰值的斜率為下降的人常伴隨著運動相對於休息引起的 ST 段壓低範圍介於 1.3~2.4 的情形發生。另外，比較有趣的是，由這三十筆關聯法可以發現，患有心臟病的患者運動時並不會誘發心絞痛。

陸、 參考文獻

1. UCI Machine Learning Repository Heart Disease Data Set
2. C.L. Blake and C.J. Merz. UCI repository of machine learning databases, 1998.
3. George H. John, Pat Langley: Estimating Continuous Distributions in Bayesian Classifiers. In: Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, 338-345, 1995.
4. 風傳媒 2017國人十大死因出爐！癌症高居死因榜首36年
5. World Health Organization The top 10 causes of death

柒、 組員分工

組員	分工
00557019 何寬宥	上台報告、資料分析
00657049 黃姿涵	蒐集資料、資料前處理
00557039 王晧宇(已退選)	資料分析(K means)

表 7-1:組員分工表