資科系 111016011 陳奕

機器學習HW2\_Classification

程式碼語言:

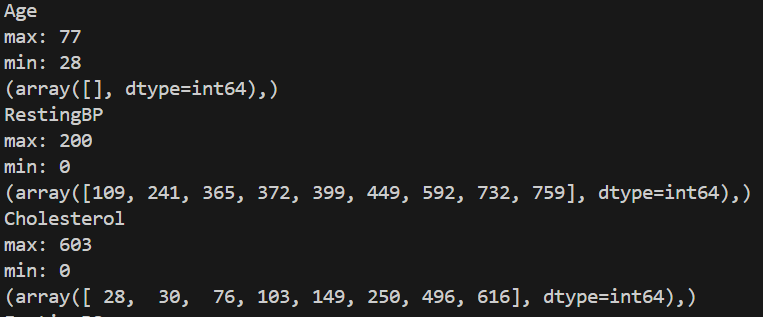
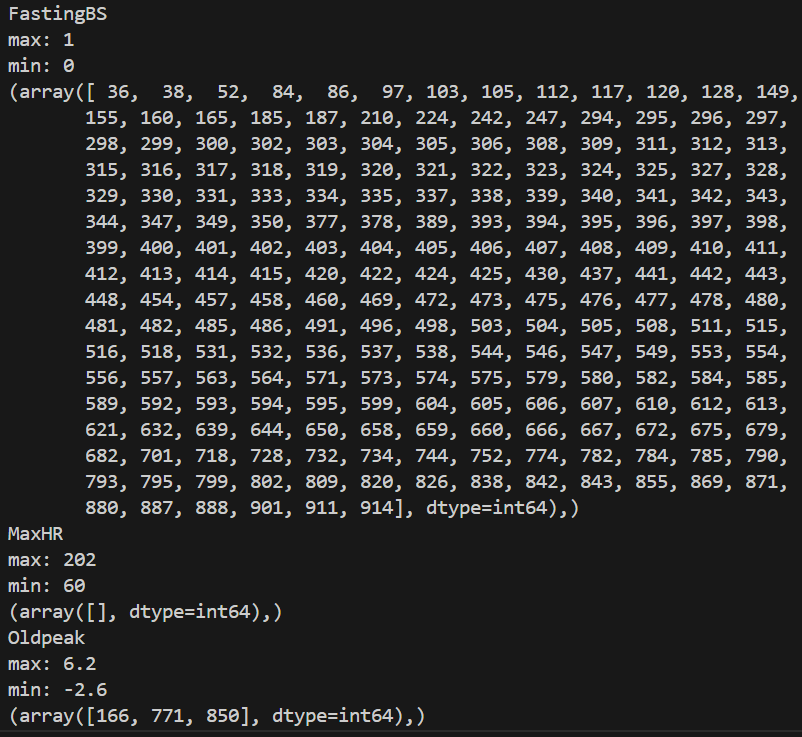
內容:

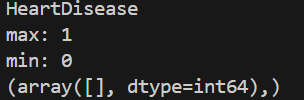
1. 資料清理與視覺化圖表 (參考程式碼Part 1)
2. 檢測遺失值

使用function 1-1檢查資料內是否有遺失值，可以發現每個特徵都是918個，並沒有資料空缺或遺失，其中有5項特徵值為object。

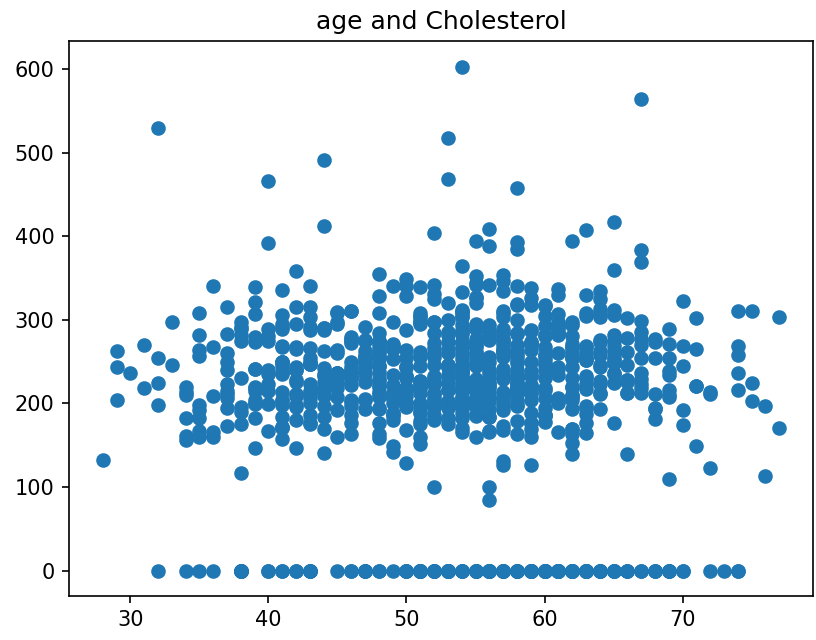
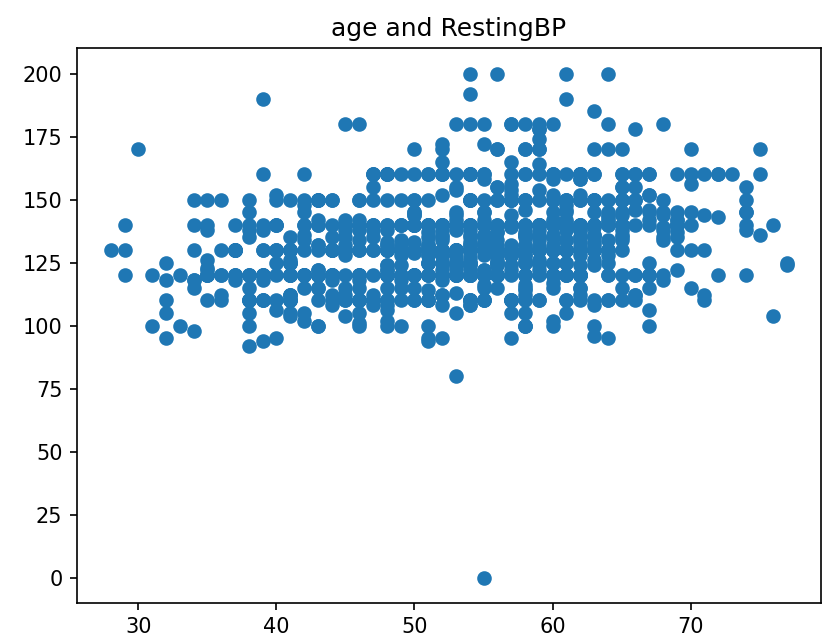


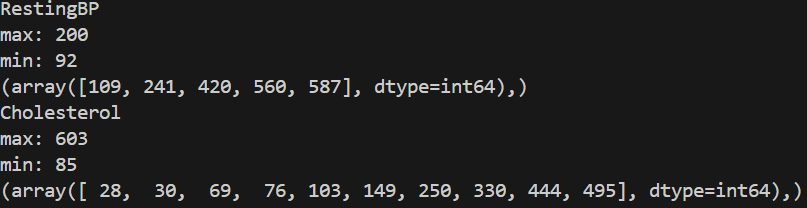
1. 清理資料

使用function1-2-2 找出數值型資料的最大值與最小值以及離群值，而離群值用function 1-2-1 利用四分位數區間尋找離群值。 



根據以上結果我認為，age的最大值與最小值皆正常，不用進行處理，resting BP 的最大值，經過google搜尋可以找到血壓超過200 mm/Hg 的案例，因此對於最大值不處理，最小值為0明顯是不正常的數據， Cholesterol同理，若患有高膽固醇血症則可能高於600 mm/dl，因此也不做處理，最小值為0明顯是不正常的數據，Fasting BS的值只包含1與0，HeartDisease同理，MaxHR在欄位說明.doc檔裡面已經說明，也確實符合60到202的範圍，oldpeak的部分，老師對不起，但我真的不知道那是什麼，所以不對其進行處理。總而言之，需要處理的資料包含 過小的resting BP與過小的Cholesterol，經過function 1-3 畫出RestingBP與Cholesterol的散布圖後，意外發現，兩者的異常值都只有為0的值，因此用function 1-4直接刪除為0的數據即可。

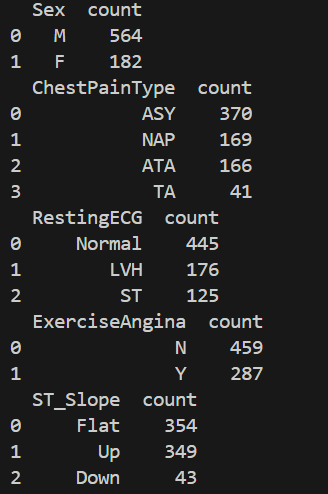


再用function 1-2-2檢查一次。

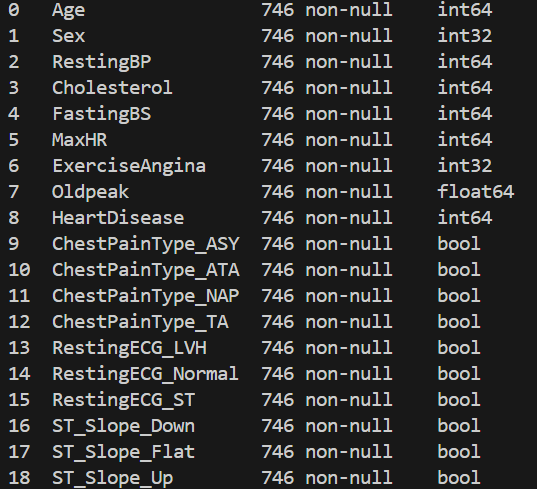
可以發現數值都在正常範圍了。

1. 整理資料

使用function 1-5，把每個型別為object的資料內的所有情形都抓出來並統計，可以發現，SEX與ExerciseAngina分別只有M與F以及Y與N，因此兩者可以直接更改為0與1，至於另外三個，則對其進行one-hot encoding處理。



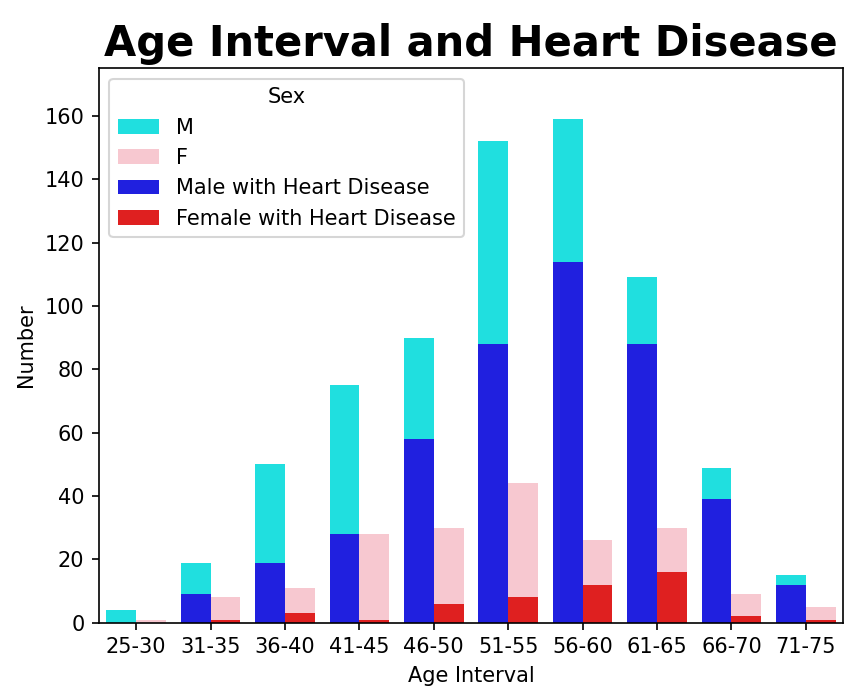
根據上述對於object資料型別的處理，使用function 1-6對其進行處理並存於HW2\_modified\_heart.csv，再使用function 1-1對其進行檢查，確認其資料型態，可以發現object型態的不存在了，方便後續運算。



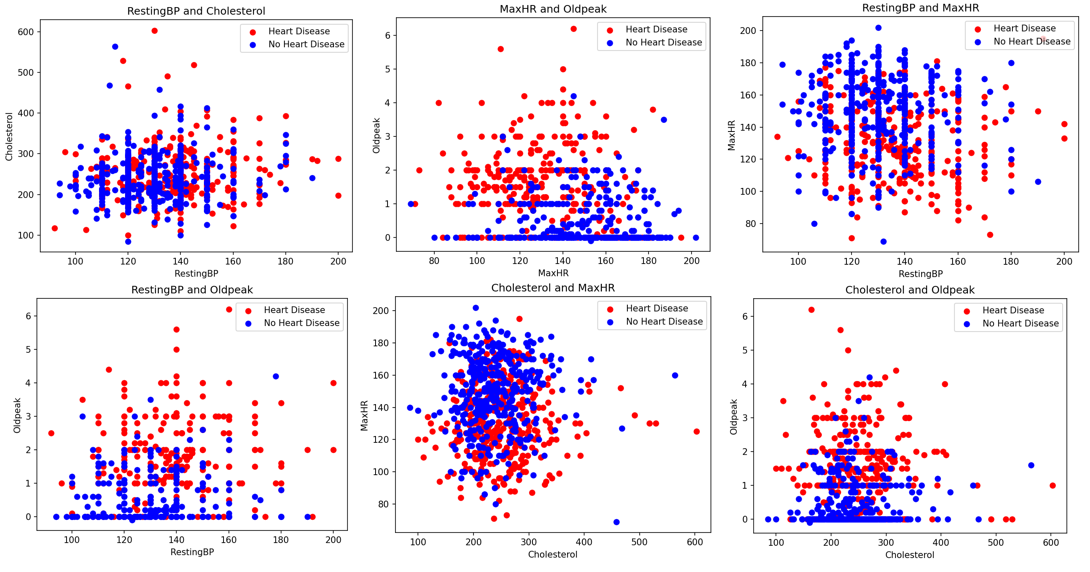
1. 圖表化

使用function 1-7-1與1-7-2畫出下列圖表，代表心臟病在各年齡層的狀況，function 1-4-1將Age切割成不同年齡層，並畫出資料內的全部人數與得病人數，function 1-7-2則用於輔助function 1-7-1。

可以發現，本筆資料中，男性的樣本多於女性，且以55~65歲的年齡層資料最多，且隨著年齡增長，得到心臟病的比例會上升，而男性又較女性容易獲得。



接下來剩下RestingBP、Cholesterol、MaxHR、Oldpeak四項數值型資料，使用function 1-8繪製下列圖表



1. 敘述性統計分析

使用function 1-9 ，可以得出下列表格，存於HW2\_heart\_descriptive\_statistics，值得注意的是，此處使用的資料只包含在原始資料中就已經為數值型的資料，而表格代表的含意由上到下代表：

count（計數）：非缺失值的數量。

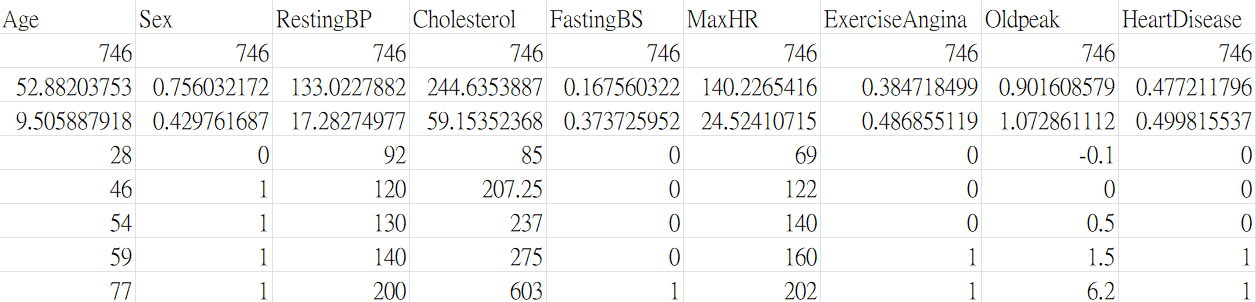
mean（平均值）：數值欄位的平均值。

std（標準差）：數值欄位的標準差。

min（最小值）：數值欄位的最小值。

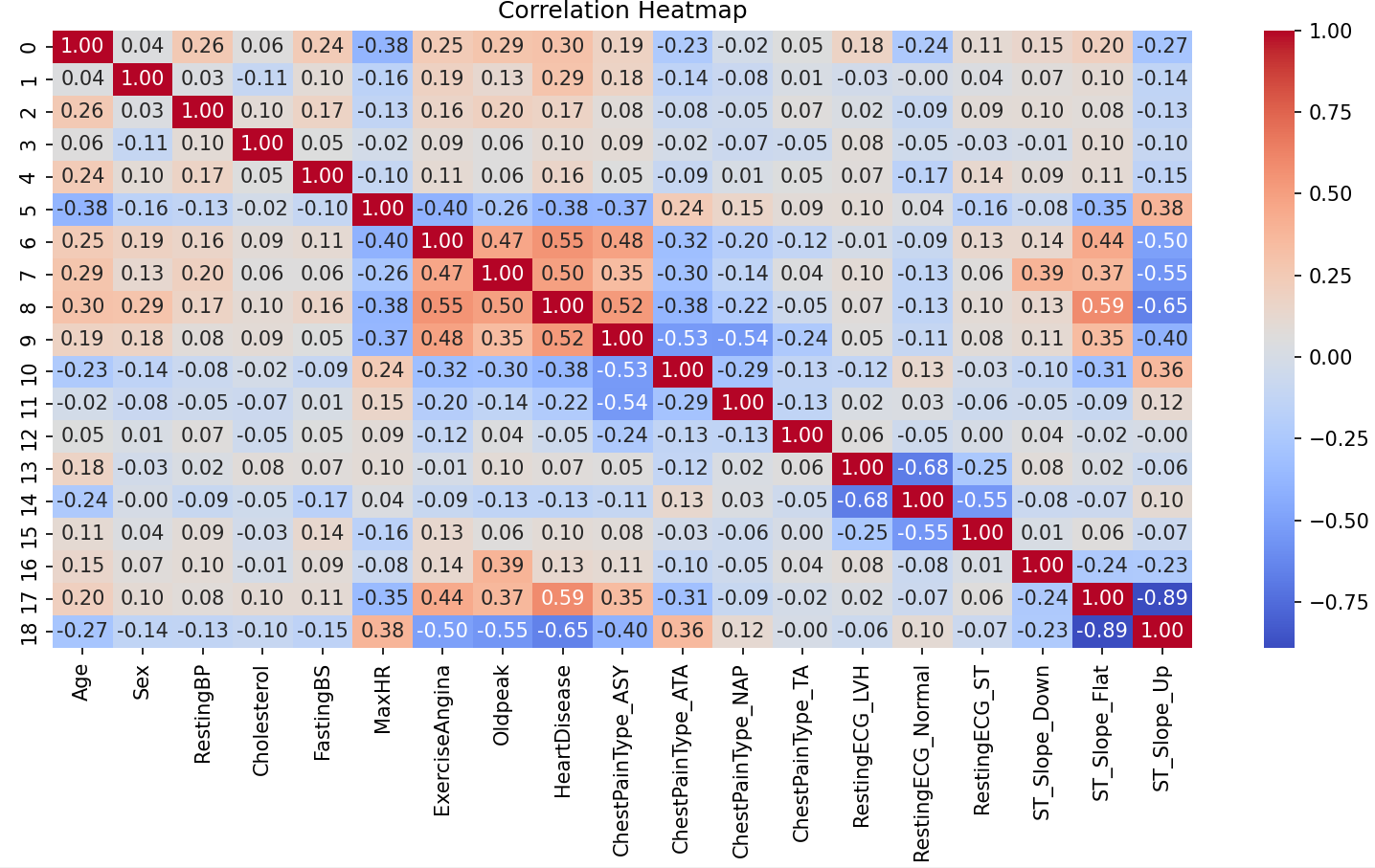
25%、50%、75%（四分位數）：四分之一、中位數和四分之三的百分位數。

max（最大值）：數值欄位的最大值。



1. 特徵相關性分析

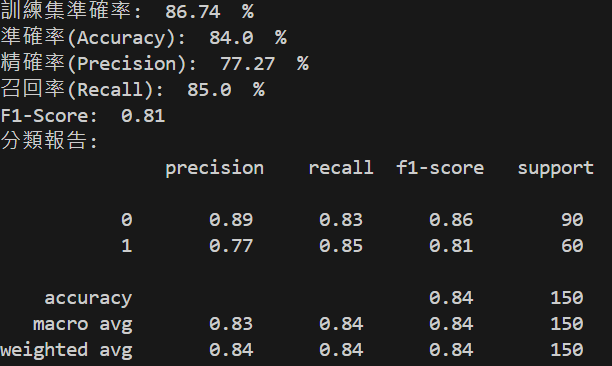
使用function 1-10-1計算HW2\_modified\_heart.csv的皮爾森相關係數，並用function 1-10-2畫成熱度圖，如下

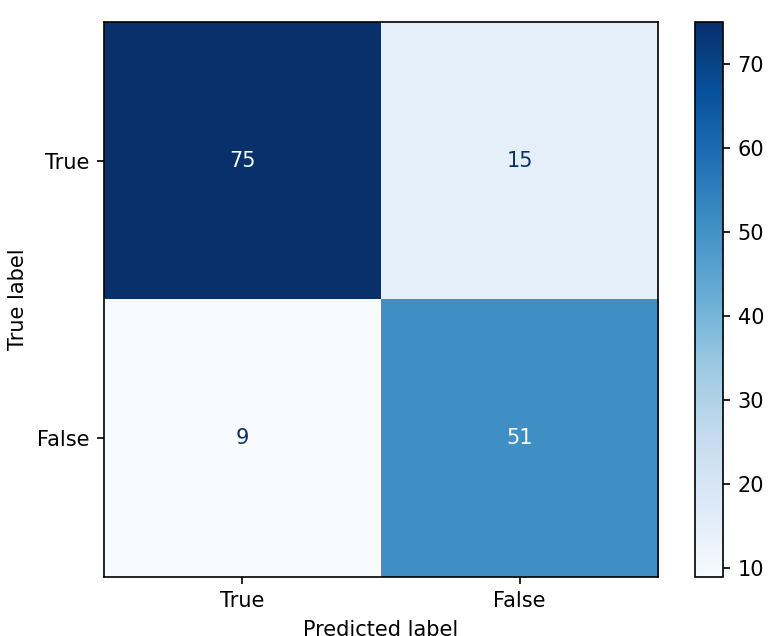


1. 資料分割與建置4個分類模型(1. Logistic regression、2. SVM、3. Random forest、4. KNN)
2. 資料分割

使用function 2-1 將資料分割成訓練集：測試集 為8：2，並固定random\_state 為 10，以便後續模型建置的進行。

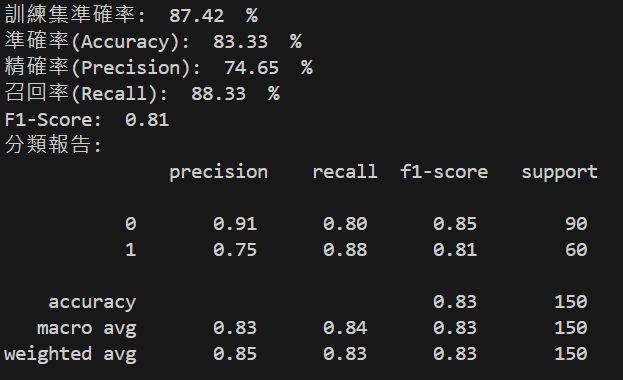
1. Logistic regression - ovr

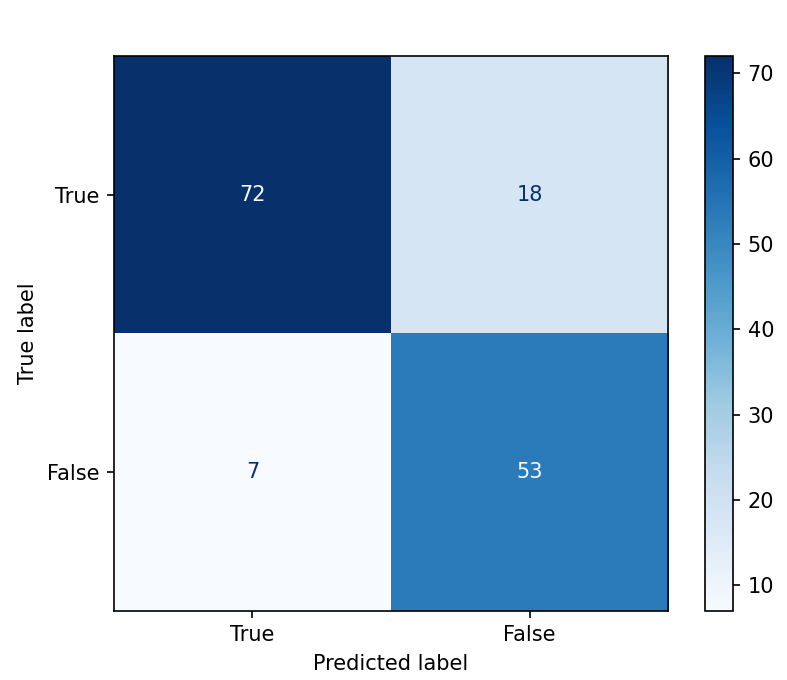
使用function 2-3-1建置Logistic regression的模型，這裡採用ovr策略，並使用function 2-2-1評估模型，以及function 2-2-2建立並繪製混淆矩陣，結果如下。



從以上結果可以發現，準確率(Accuracy)為84.0%,精確率(Precision)為77.27%,召回率(Recall)為85.0%,F1分數為0.81。混淆矩陣的結果，60位沒得病的人有9位被誤判，比例為0.15；90位病患有15位被誤判，比例為0.166…，從這兩張圖片可以看出，這個模型對於得病與沒得病的人的預測能力相差不大，但還是在預測已經患病的病患上能力較強。

1. Logistic regression – ovo

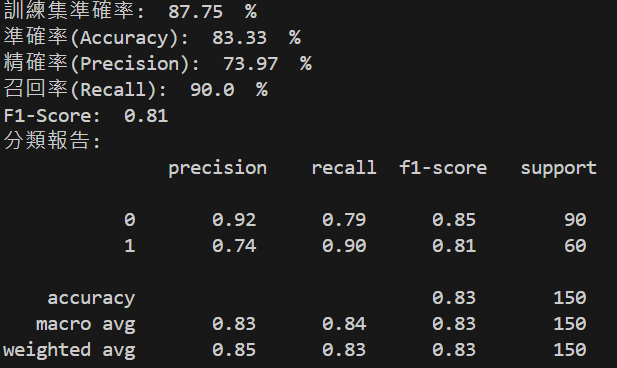
和上一項相同，一樣是Logistic regression，不過function 2-3-2這次採用ovo 策略，檢測結果的方式和上面相同，結果如下

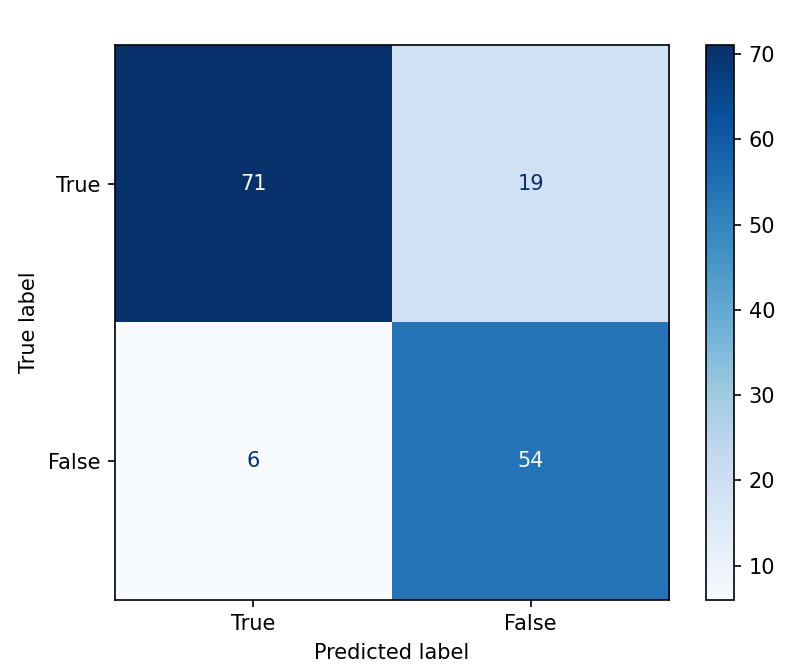


可以發現，ovo則稍微準確，總之，與ovr的結果相比，ovo在測試集的結果表現較好，但測試集的結果稍差，或許有些過擬和的情形發生。

1. SVM-linear

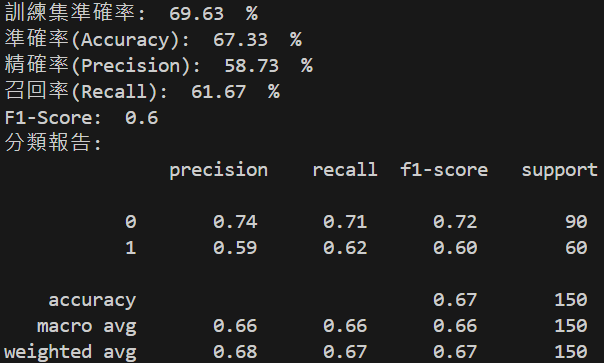
使用function 2-4-1和2-4-2建置SVM的模型，兩者皆使用線性核函數，前者採用ovr策略，後者則為ovo，並一樣使用function 2-2-1評估模型，以及function 2-2-2建立並繪製混淆矩陣，兩者結果完全相同，結果皆如下

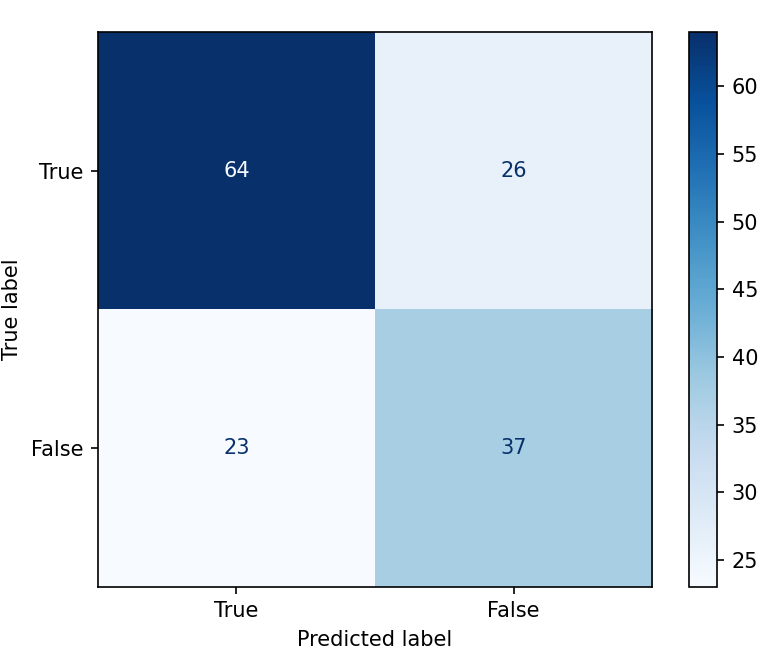




1. SVM-rbf

同上，不過這次將核函數改為逕向核函數，建立的函數為function 2-5-1(ovr)、function 2-5-2(ovo)，採用ovr與ovo的策略結果也相同，結果都如下

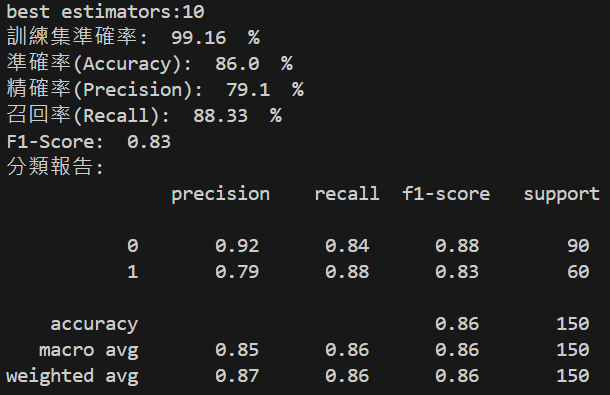


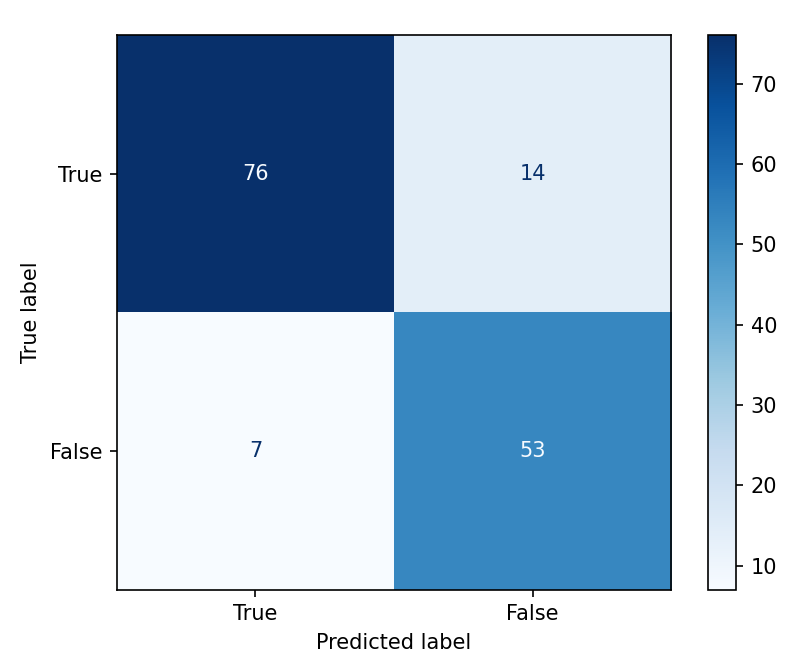


可以明顯的發現，採用線性核函數的模型表現明顯較優秀，故在下一部分比較4個模型時，皆以線性核函數的模型為主。

1. Random forest

先使用function 2-6-1 ，找到在1~200之間準確率(Accuracy)最高的n\_estimators參數，再使用function 2-6-2建立random forest 模型，評估模型方式同上，結果如下

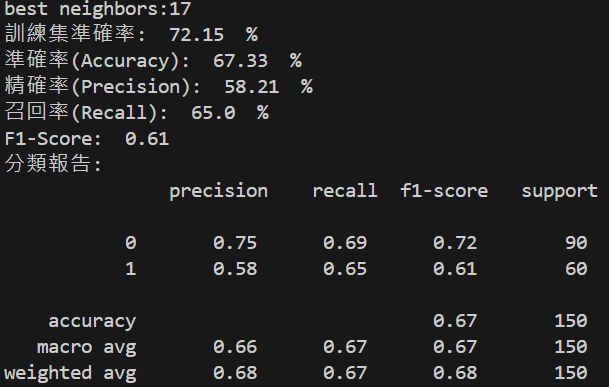


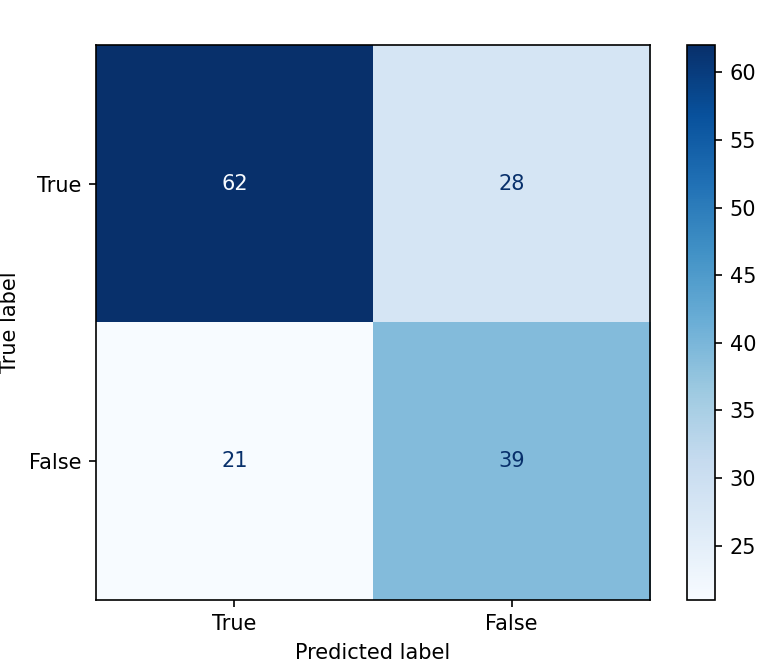


可以發現在訓練集的表現十分接近100%，而在實際上的表現就回歸正常，可能有過擬合的情形發生。

1. KNN

和random forset 的過程類似，先使用function 2-7-1找出最適合的neighbor數量，再利用function 2-7-2建置模型，評估模型的方式同上，結果如下





在neighbor = 17 時會有最高的Accuracy，但也仍然只有67.33%，整體的數據也不及之前建置的模型。

1. 綜合比較4種模型的分類結果與分析討論

以下為五個模型的分類結果，為採用ovr的Logistic regression、採用ovo的Logistic regression、SVM、random forest、KNN

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 模型 | Logistic-ovr | Logistic-ovo | SVM | random forest | KNN |
| 訓練集準確率 | 86.74% | 87.42% | 87.75% | 99.16% | 72.15% |
| Accuracy | 84% | 83.33% | 83.33% | 86% | 67.33% |
| Precision | 77.27% | 74.65% | 73.97% | 79.1% | 58.21% |
| Recall | 85% | 88.33% | 90% | 88.33 | 65% |
| F1-Score | 0.81 | 0.81 | 0.81 | 0.83 | 0.61 |

可以明顯的看出，random forest的綜合表現是全部當中最優秀的，僅只有召回率略低於SVM，而KNN是表現最差的，這次的資料有一大部分是類別型的，因此我原本就猜測隨機森林的表現會是最好的，但令我感到意外的是，隨機樹在訓練集準確率已經高達99.16%，看起來就是完完全全的過擬合，結果綜合評分看下來，隨機森林仍然是表現最佳的。