使用機器學習預測患者是否為潛在心臟病患者

指導教授：王昭能

學生：謝季恩

生醫系3A 107025005

**摘要**

心臟疾病長期位居國人十大死因的前3名，其中「心肌梗塞」的死亡率更是高居不下。心肌梗塞發生率正逐年攀升，且罹病者也有年輕化的趨勢。因此本研究使用邏輯回歸（分類）算法對心臟病數據集做機器學習，並預測患者是否患心臟病。本次數據使用的數據共303筆，首先進行數據預處理，將Label與特徵進行分離，避免後續機器學習在讀取上出現錯誤，演算法使用隨機森林(RandomForest)進行模型訓練，為了方便訓練，將數據切分成8比2的訓練集與測試集，並根據預測後的混淆矩陣來判斷模型是否準確。

**前言**

根據衛生福利部107年死因統計，心臟病是國人第二號殺手，造成21,569人死亡，平均每24分鐘就有1人死於心臟病，如再加上腦中風、高血壓、糖尿病及腎臟病等血管性疾病，每年造成53,977人死亡，遠超過頭號殺手癌症奪走的48,784條人命。美國2010年統計[1]，美國估計有200,070人死於心臟病、中風和高血壓疾病，其中56%發生在 65 歲以下的人群中。總體年齡標準化死亡率為每100,000人中60.7人。

為了預防心臟病，本研究收集了各年齡層的數據共303筆，希望透過機器學習預測出心臟病的潛在患者，讓患者能夠及早接受治療。

**方法**

1. 隨機森林：

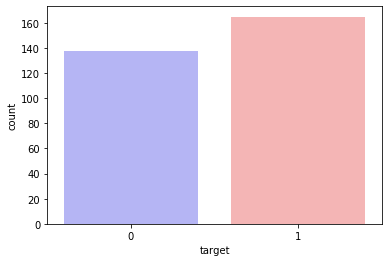
本研究這次使用隨機森林(RandomForest)[2]進行模型訓練與混淆矩陣[3]進行評估。Random Forest的基本原理是結合多顆CART樹（Classification And Regression Tree，CART）是使用GINI算法的決策樹，並加入隨機分配的訓練資料，以大幅增進最終的運算結果。

1. 流程：

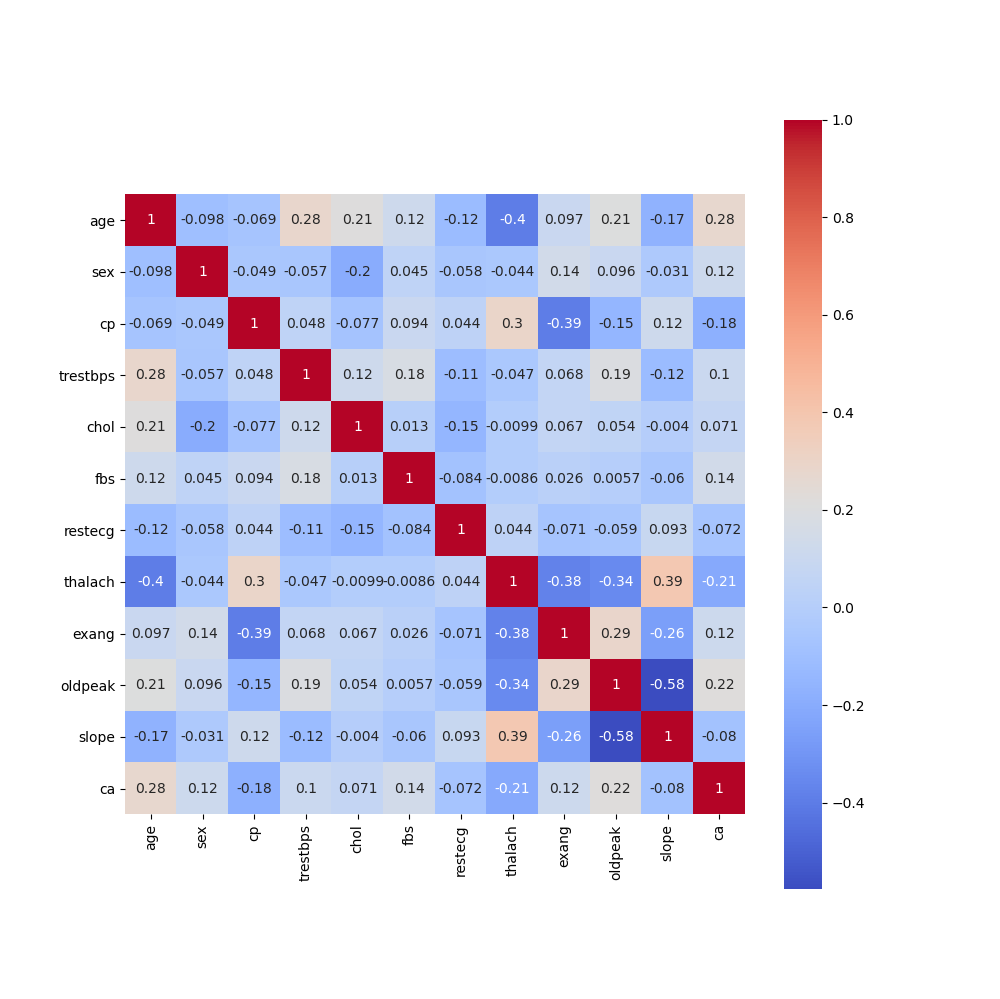
本研究首先對數據進行前處理，將label與特徵分離後，在將前處理後的數據拆分成8比2的數據量進行訓練與預測，參數的部分經過最佳化後，參數的設定為n\_estimators = 801,max\_depth=13, min\_samples\_leaf=2, min\_samples\_split=130。

**數據說明**

本研究根據數據將數據分為兩類，一類為沒有心臟病，一類為有心臟病，並統計患有心臟病與沒有心臟病的人數(圖一)，再繪製所有特徵的相關係數圖 (圖二)，比對其關聯性。



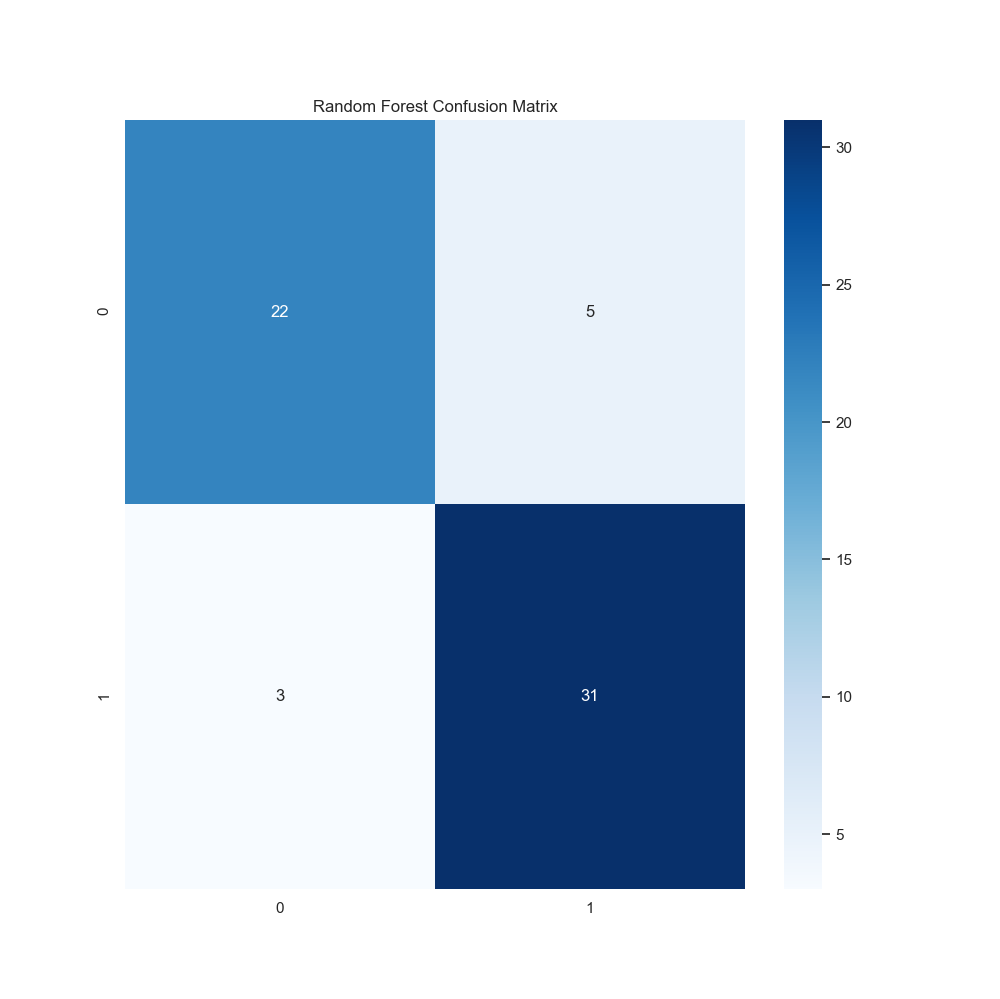
圖一、Count為人數，0為沒有患有心臟病，1為患有心臟病



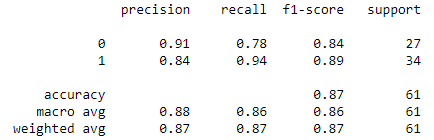
圖二、個特徵的相關係數圖

**結果分析**

本研究在使用RandomForest進行模型訓練後，最終經過混淆矩陣(圖三)可視化後可以看到在預測61筆數據中，只有8筆數據分配錯誤，並計算其召回率(Recall)(圖四)，從圖三中可以看到兩類個別的召回率分別為0.78與0.94，代表模型在0類的預測中不能很正確的預測。



圖三、模型預測結果使用混淆矩陣可視化



圖四、各類別的準確度與趙召回率

**結論**

本研究使用隨機森林進行心臟病預測，根據以上的結果顯示，模型在預測上依舊有不足的地方，例如：在沒有心臟病的預測中，模型容易將其分為有心臟病，為此需要再新增新的數據量來看模型是否能夠更準確的預測。

參考文獻

1. Control, C.f.D. and Prevention, *Vital signs: avoidable deaths from heart disease, stroke, and hypertensive disease-United States, 2001-2010.* MMWR. Morbidity and mortality weekly report, 2013. **62**(35): p. 721-727.

2. Liaw, A. and M. Wiener, *Classification and regression by randomForest.* R news, 2002. **2**(3): p. 18-22.

3. Story, M. and R.G. Congalton, *Accuracy assessment: a user’s perspective.* Photogrammetric Engineering and remote sensing, 1986. **52**(3): p. 397-399.