**LLC-Kaggle 平台實戰(111AIGO)**

# 心臟病預測及資料視覺化

作者：謝昊廷

**目錄：**

## 1、 前言

根據世界衛生組織（WHO）於 2020 年 12 月 9 日所發布《2019 年全球衛生評估報告》中[1]，自 2000 年至 2019 年之統計數據，全球十大死因中，七項為「非傳染性疾病」，而心臟病是所有疾病中第一名。世界衛生組織估計全世界每年有 1200 萬人死於心臟病。在美國和其他發達國家，一半的死亡是由於心血管疾病。心臟病發作，是許多心血管疾病患者的夢魘，歐洲心臟科醫學會(European Society of Cardiology, ESC)在 2019 年發表了一篇論文，顯示機器學習(machine learning)正取代人類，預測死亡或心臟病的發生[2]。根據研究，心臟病的發生機率，與我們的年齡、性別、家族病史、生活習慣息息相關。衛生福利部國民健康署也列舉出影響心臟病預後的因子如下表[3]：

| 心血管疾病的危險因子 | 標的器官損傷 | 伴隨有高血壓後遺症的臨床症狀 |
|---|---|---|
| Ⅰ.用於區分危險性<br>血壓高的級別(1~3 級)<br>55 歲以上的男性<br>65 歲以上的女性<br>抽菸<br>血中總膽固醇濃度超過<br>6.5mmol/l(250mg/dl)<br>糖尿病<br>家族中早發性心血管疾病<br><br>Ⅱ.其他影響預後的不良因子<br>高密度脂蛋白膽固醇降低<br>低密度脂蛋白膽固醇增加<br>糖尿病併有蛋白尿<br>葡萄糖耐受性降低<br>肥胖<br>久坐式的生活型態<br>纖維蛋白原增加 | (target-organ damage)<br>左心室肥大<br>蛋白尿或是輕度腎功能障礙(血漿中肌酸酐濃度介 1.2~2.0mg/dl)<br>由超音波或 X 光發現動脈粥狀硬化(頸動脈、腸骨動脈及股動脈、或主動脈)<br>全面性或部分的視網膜病變 | 腦血管疾病<br>缺血性腦中風<br>腦出血<br>暫時性腦缺血<br>心臟疾病<br>心肌梗塞<br>心絞痛<br>冠狀動脈繞道手術<br>充血性心衰竭<br>腎臟疾病<br>糖尿病腎病變<br>腎衰竭(血漿中肌酐酸濃度大於 2.0mg/dl)<br>血管疾病<br>主動脈瘤<br>有癥候之動脈疾病，如頸動脈或冠狀動脈狹窄<br>續發性之高血壓性視網膜病變<br>網膜出血<br>視神經乳突水腫 |

表一、影響心臟病預後的因子

由此可知在年齡、性別、所患血中膽固醇（主要為低密度膽固醇）、血管寬度、運動等因素，皆與心臟病有所相關。同時也列舉出量測血壓與各項因子，在患有部分相關疾病的因素下，對心臟病發生的危險性分類（表二）。

| | 第一級(輕度高血壓)<br>毫米汞柱(mmHg) | 第二級(中度高血壓)<br>毫米汞柱(mmHg) | 第三級(嚴重高血壓<br>毫米汞柱(mmHg) |
|---|---|---|---|
| 心血管疾病危險因子及疾病史 | 140≦收縮壓≦159 或<br>90≦舒張壓≦99 | 160≦收縮壓≦179 或<br>100≦舒張壓≦109 | 收縮壓>180 或<br>舒張壓>100 |
| Ⅰ.沒有其他危險因子 | 低危險 | 中度危險 | 高危險 |
| Ⅱ.有 1 至 2 個危險因子 | 中度危險 | 中度危險 | 極高度危險 |
| Ⅲ.有 3 個以上的危險因子或有標的器官受損或糖尿病患者 | 高危險 | 高危險 | 極高度危險 |
| Ⅳ.曾經伴隨有高血壓後遺症的臨床症狀 | 極高度危險 | 極高度危險 | 極高度危險 |

表二、不同血壓高合併其他危險因子者發生心臟血管疾病危險性分類

心血管疾病的早期預後可以幫助決定改變高危患者的生活方式，從而減少併發症。本研究旨在查明心臟病最相關/風險因素，並使用機器學習預測總體風險。藉由 Kaggle 上心臟資料庫裡所記錄的相關生理特徵與檢驗數值[(4)]，建立預測模型分類導致心臟疾病的相關因子，並使用機器學習預測總體罹患心臟病／復發之風險。

## 2、 研究目標

1. 特徵化工程

本研究採用 kaggle 上公開之心臟病患預測資料庫，有別於作者對資料較無多餘處理，本組以此為延伸，期望經由特徵化工程，提煉各特徵之特點，及除去資料雜質，以期在最終準確率上勝過原作者。

2. 資料視覺化

原作者對各特徵僅單純地統計柱狀圖，與類別特徵中，各分類比例介紹，沒有各特徵對罹患心臟病之關係，本組希望能加以完整之。

## 3、 資料來源

此次以 Kaggle 中的 Heart Disease Prediction (95% Accuracy & Recall)資料集為訓練資料集，有 198 比病患資料，12 項特徵。該資料集中的資料特徵共計 12 項如下表三：

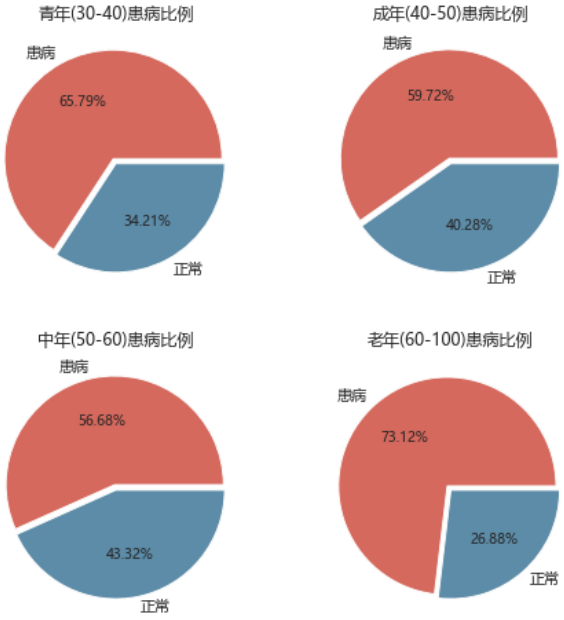
| 特徵名稱 | 描述 |
|---|---|
| Age | 年齡 |
| Sex | 性別 |
| ChestPainType | 胸絞痛類型。<br>ASY:無症狀；NAP:非心絞痛；<br>ATA:非典型心絞痛；TA:典型心絞痛。 |
| RestingBP | 靜息收縮壓 |
| Cholesterol | 血清膽固醇 |
| FastingBS | 空腹血糖>120mg/dl。True=1, False=0 |
| RestingECG | 靜息心電圖。<br>Normal:正常；LVH: left ventricular hypertrophy 左心室肥大；<br>ST-T wave abnormality ST-T 波異常。 |
| MaxHR | 最大心跳速率 |
| ExerciseAngina | 運動誘發的心絞痛。Y=會；N=不會。 |
| Oldpeak | 運動高峰心電圖。相對於休息的運動引起的 ST 值(ST 值與心電圖上的位置有關) |
| ST_Slope | 高峰 ST 段。Up:向上傾斜；Flat:持平；Down:向下傾斜。 |
| HeartDisease | 是否患有心臟病 0:無；1:有。 |

表三、Kaggle 心臟病資料集欄位說明

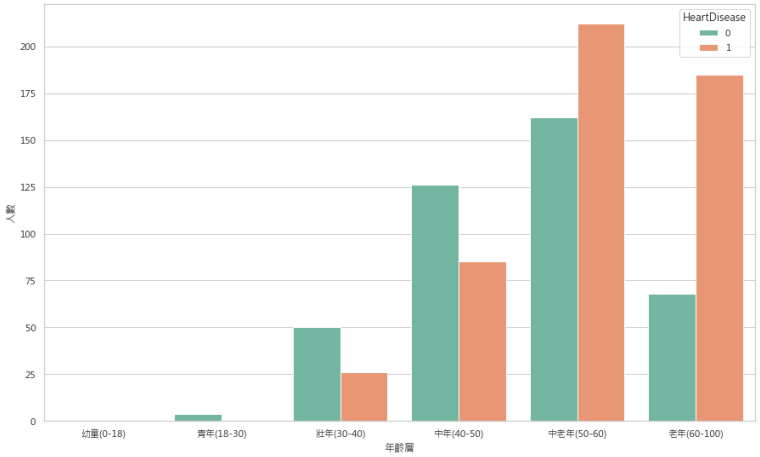
## 4、 資料比較

我們首先將 11 項不同欄位進行與心臟病之關係比較如下：

1. 年齡



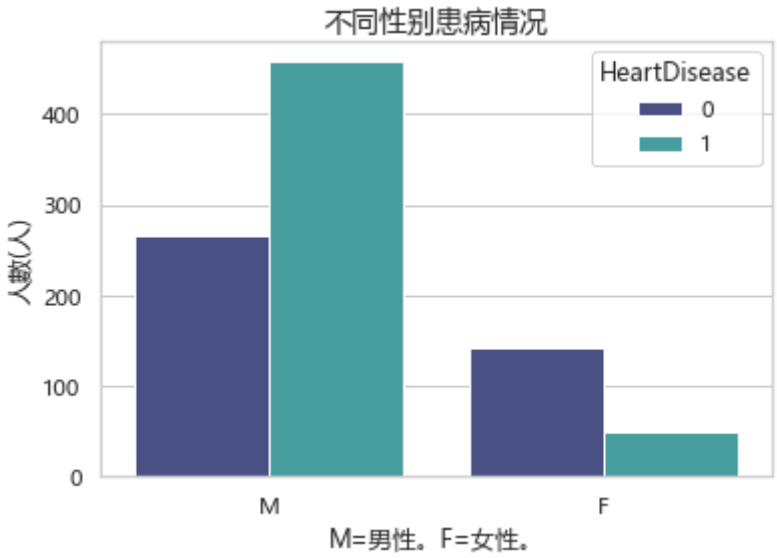

圖一、各年齡層與心臟病患病比例




圖二、各年齡層患有心臟病之人數

在臨床生理上，因年歲增長及生活作息的關係，會造成血管本身彈性降低、硬化，甚或有害物質堆積於血管之中，因而造成血管阻塞，進而提高心臟病的機率。在圖二中，很明顯的顯示出自 50 歲後，患有心臟病之人數大幅上升，故可推斷年齡與心臟病有著正相關的關係。

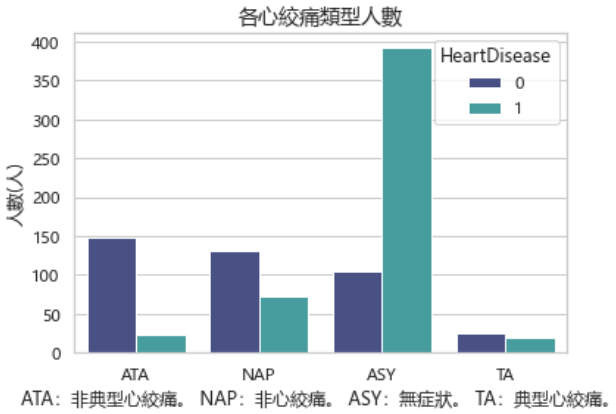
2. 性別




圖三、性別與心臟病病患之關係圖

於此圖中，明顯男女比例患病差異大，但根據近年歐洲心臟科醫學會((European Society of Cardiology)研究抽菸為一項影響心臟病之重要因子，且男性抽菸者比例上較女性多，是否此為背後主要因素，後續可做進一步探討。

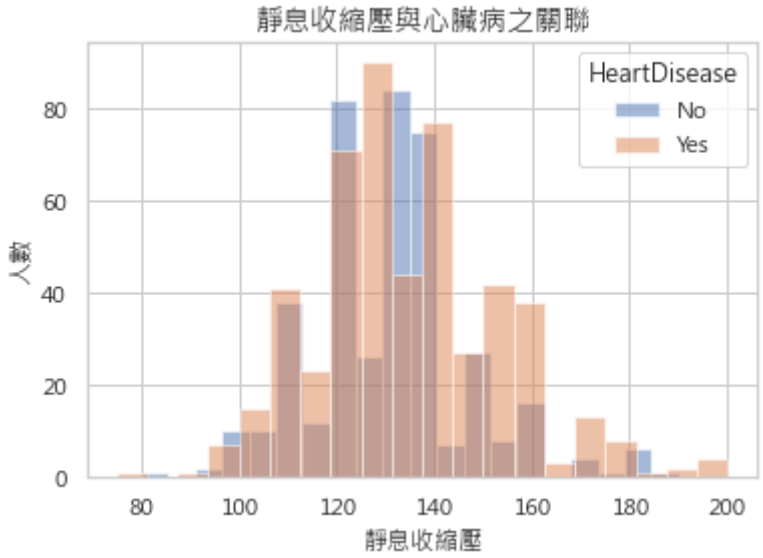
3. 心絞痛




圖四、心絞痛種類與心臟病病患之關係圖

心絞痛在醫學上為冠狀動脈心臟病臨床表現之一，但在此資料集中，心絞痛的人數較一般民眾少，或許此為資料建置時可以考慮是否採用。

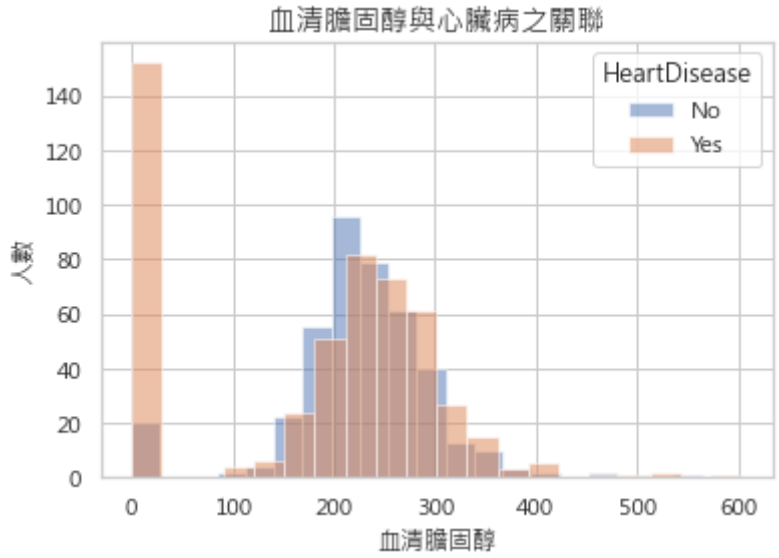
4. 靜息收縮壓




圖五、靜息收縮壓與心臟病病患之關係圖

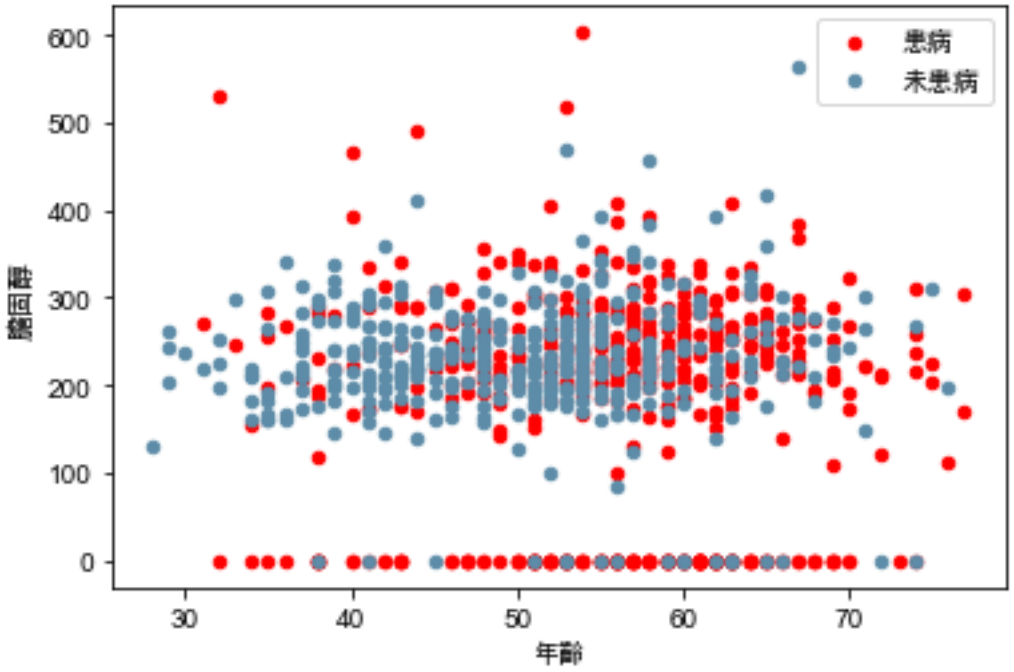
血壓通常為生理量測一項重要指標。根據美國心臟協會[6]於 2017 年 11 月 13 日發布有關高血壓標準值的最新治療指引，重新定義血壓達 130/80 毫米汞柱（收縮壓 130 毫米汞柱、舒張壓 80 毫米汞柱）以上即為高血壓。於臨床醫學上，高血壓與心臟病有著非常大的關係，因此在圖六中，也可明顯的看出，以 140mmHg 為分界，高於 140mmHg 之血壓，患有心臟病之人數皆超過無心臟病之人數，故此可推斷此項目與心臟病為正相關。

5. 血清膽固醇



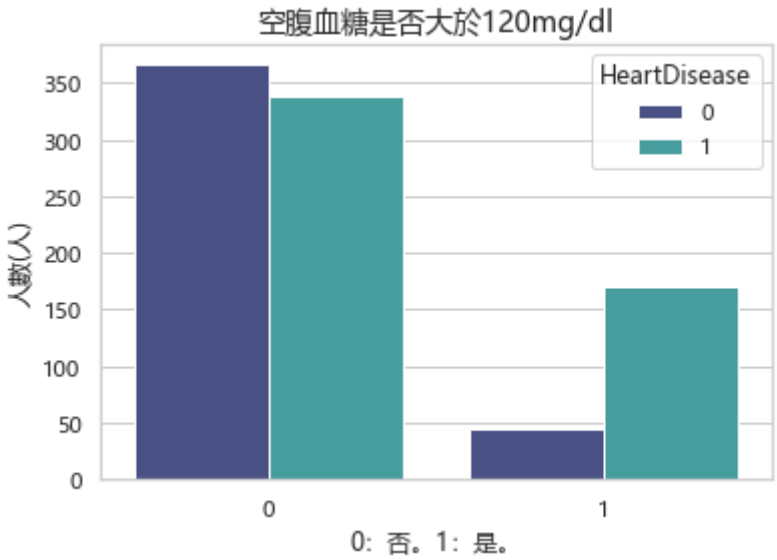

圖六、血清膽固醇與心臟病病患之關係圖

在臨床上，膽固醇升高會促進動脈硬化，繼而增加心臟病的風險。故在圖七中，高於 240mg/dl 所患有心臟病之人數，明顯皆高於未有心臟病之人數，故此可確立血清膽固醇與心臟病為正相關。於此同時，本組也探討了加入年齡之後的結果如下圖八。同樣的，年紀越大與膽固醇越高，則患有心臟病的人也越多。




圖七、年齡、血清膽固醇與心臟病病患之關係圖

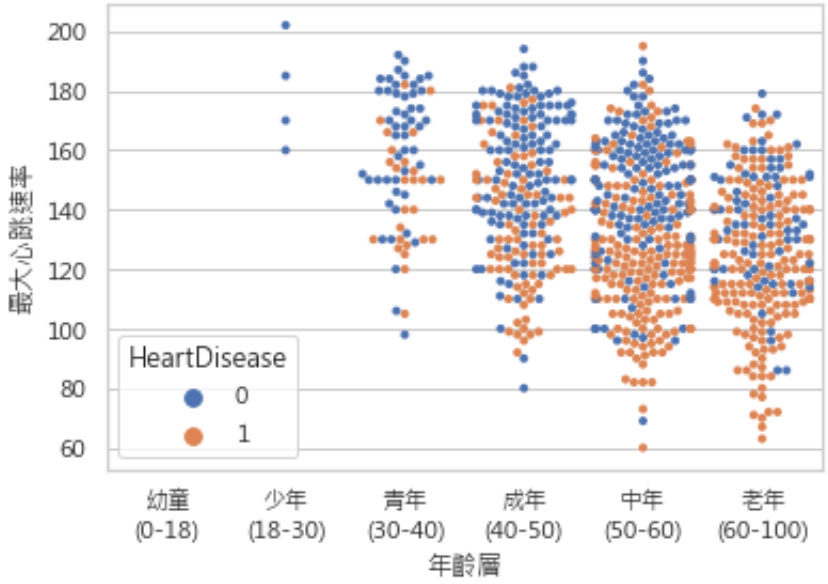
6. 空腹血糖




圖八、空腹血糖值與心臟病病患之關係圖

空腹血糖值為糖尿病的重點檢驗數值之一。在臨床統計上，糖尿病患者罹患心臟病的風險是一般人的兩倍，也較容易在年輕時罹患心血管併發症，常導致提早死亡。高血糖會使得動脈壁較易增厚硬化，導致負責輸送血液到心臟肌肉的血管管腔變得狹窄，且心肌因此無法獲得充足的血液與氧氣，造成心臟病。在此數據集中，因糖尿病僅為提高觸發心臟病之因素之一，而非重點成因之一，因此患有心臟之病人也不一定會有高血糖之情形。但可觀察到右半部高血糖值且同時有心臟病的人數，比未有心臟病的人數明顯數量高很多，因此此項可作為參考數值。

7. 最大心跳速率

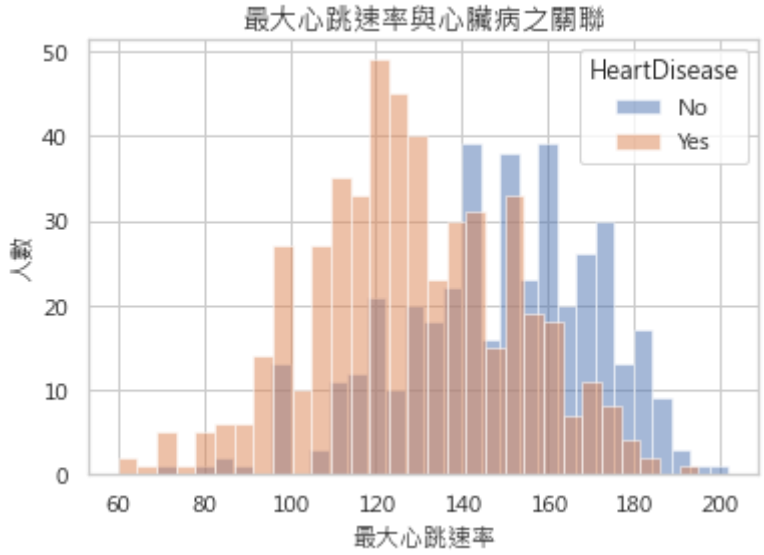



圖九、最大心率與年齡層關係圖

根據近年國際期刊上提出[5]，最大心跳速率計算方式為：

**最大心跳速率 = 206.9-(0.67x 年齡)**

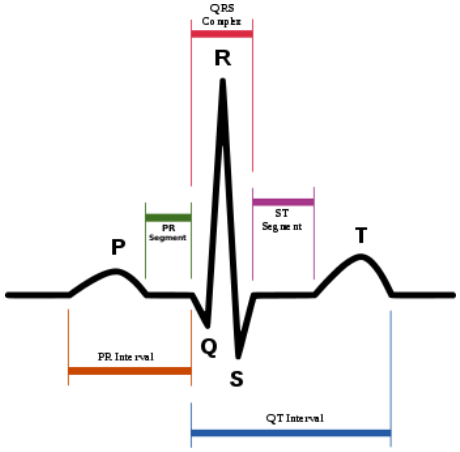
故最大心跳速率則與年齡為負相關，年齡越大，最大心跳速率則越小；由此進一步推斷，最大心跳率與心臟病將呈現負相關之情形。為證明此想法，本組便試著做關係圖，如(下圖十)。於此圖上，最大心跳速率越小，則患有心臟病之人數越高，反之則較無心臟病患者之出現。



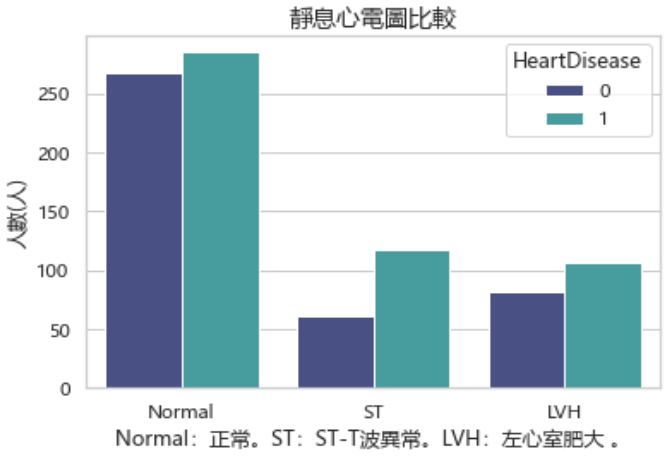

圖十、最大心跳速率與心臟病病患之關係圖

8. 心電圖與運動

在心電圖中，可將一個波分為四個部分：P 波、QRT 波與 T 波（如下圖十一）。就臨床上來說，心電圖分為三種：靜態、運動和持續性心電圖，心臟病患者或潛在患者，在靜態心電圖有可能檢測不出來。因此有時在量測時會輔以運動或持續行心電圖來檢查是否有心臟病的問題。

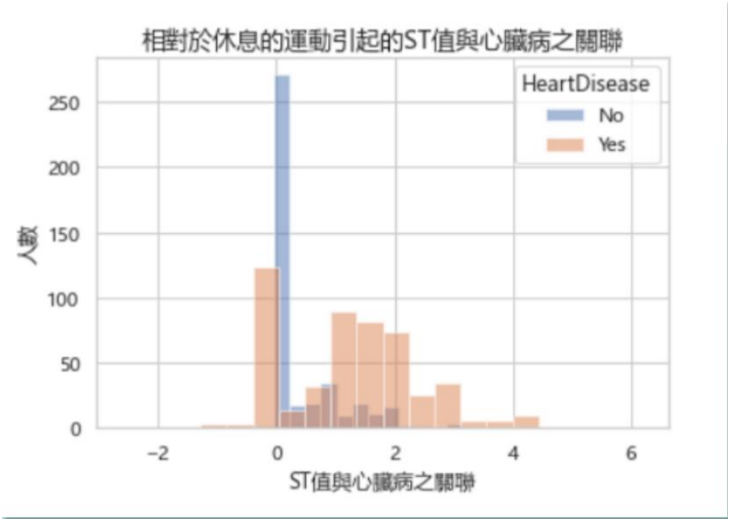



圖十一、正常狀態下的 EKG 圖




圖十二、靜息心電圖

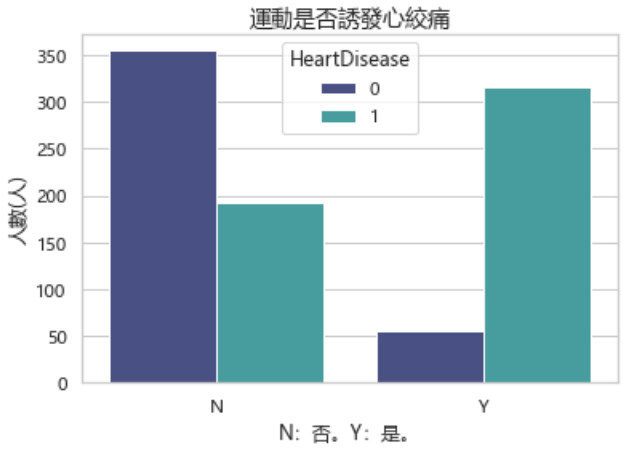
而心電圖出現 ST 段或者 T 波的變化，應該特別留意是持續性變化還是動態（一過性）變化，如果持續存在多數不是心肌缺血或者冠心病所致；如果 ST 段（T 波）變化與胸痛相關，則極有可能是不穩定性心絞痛或者心肌梗塞。因此在此次資料集中，我們找出於運動時的 ST 值與心臟病之關聯性（如下圖十四）。明顯運動狀態下，心臟病患者之 ST 值的確較容易出現異常情形。



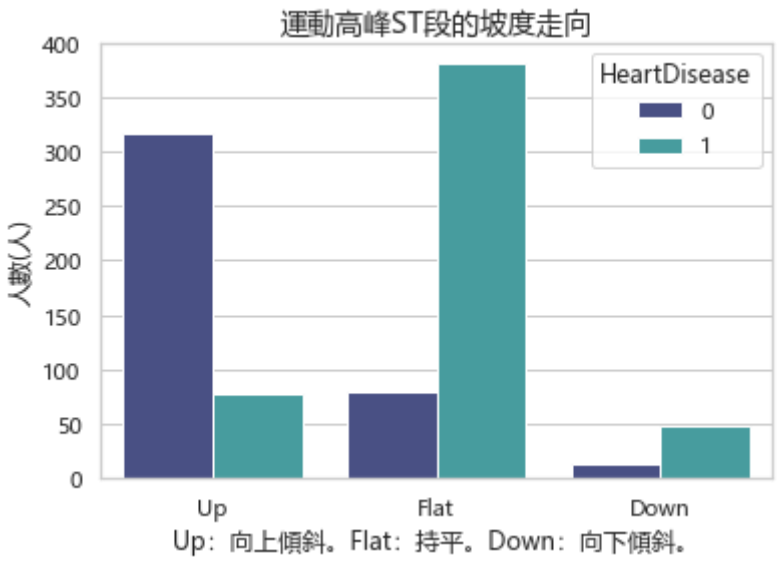

圖十三、ST 值與心臟病之關聯

9. 心絞痛

心臟病患者通常不建議進行過於劇烈運動，因運動時會增加心肌負荷及心肌耗氧量，尤其是有血管硬化病史者，容易影響心肌的血液供應，導致胸悶、胸骨後疼痛、心前區不適等心絞痛症狀。在此資料集中也有明顯呈現了運動誘發心絞痛的心臟病患者人數，高出非心臟病患者之人數（如下圖十四）。但同時此圖另有一點可值得注意後續追蹤的，是非心臟病患者卻出現心絞痛的群眾。該群體未來也有心臟病之比例，可能較未有心絞痛的一般民眾來的高。




圖十四、運動是否誘發心絞痛




圖十五、運動高峰 ST 段的坡度走向

10. 運動高峰 ST 段的坡度走向

除此之外，在運動心電圖中，ST 波段的走向，也是判別是否有心臟病的情形（如圖十五）。正常人在運動過程中，ST 區段會線上傾斜，但很明顯地在心臟病的患者上，該區段則呈現持平或下降的情形產生。因此該特徵也可做為是否患有心臟病之參考特徵。

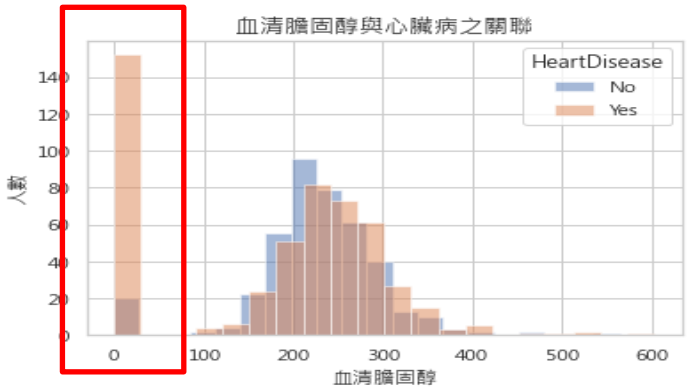
## 5、 特徵工程

### 1. 缺失值

```
#檢視是否有缺失值
df.isnull().sum()
```

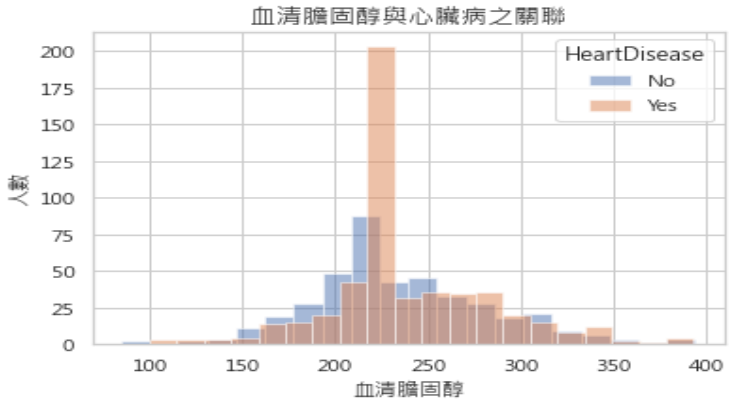
```
Age                0
Sex                0
ChestPainType      0
RestingBP          0
Cholesterol        0
FastingBS          0
RestingECG         0
MaxHR              0
ExerciseAngina     0
Oldpeak            0
ST_Slope           0
HeartDisease       0
dtype: int64
```

圖十六、檢查缺失值




圖十七、血清膽固醇數值分析圖

在以程式檢查缺失值過程中，發現未有缺失值(如上圖十六)，其原因在於很多某些地方在資料建置時，缺失值便以 0 取代（如上圖十七）。但在醫學資料上，缺失值通常會因不同檢查而有不同替代方案，而非完全以 0 取代。若以 0 取代，將會造成嚴重的誤判。因此在此次練習中，本組試著將缺失值以平均數取代，得出下列結果（如圖十八），該結果則較接近臨床上之情形。




圖十八、血清膽固醇以平均值取代缺失值

2. 離群值

在資料庫處理上，難免會有些記錄錯誤。除缺失值外，離群值也是其中一種錯誤，因此本組在用程式檢查後(平均數+-3 個標準差)，發現在 RestingBP 也就是「靜息收縮壓」有 8 組離群值，並將其以上界與下界取代。至於其他特徵的離群組，在經過偵測後，並未有過於不實的數據，因本資料有許多心臟病患者資料，其數值本來就比一般人較為異常。

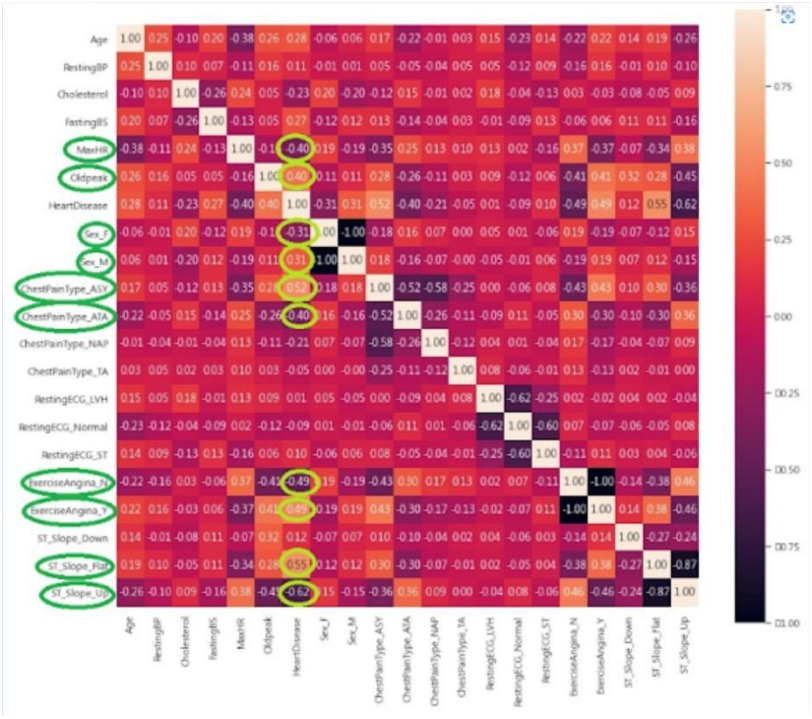
```python
#使用超過上界之離群值，用上界值代替，超過下界亦如此之方法，取代離群值
def windsorization(data,col,para,strategy='both'):
    """
    top-coding & bottom coding (capping the maximum of a distribution at an arbitrarily set value,vice versa)
    """

    data_copy = data.copy(deep=True)
    if strategy == 'both':
        data_copy.loc[data_copy[col]>para[0],col] = para[0]
        data_copy.loc[data_copy[col]<para[1],col] = para[1]
    elif strategy == 'top':
        data_copy.loc[data_copy[col]>para[0],col] = para[0]
    elif strategy == 'bottom':
        data_copy.loc[data_copy[col]<para[1],col] = para[1]
    return data_copy

#設定離群組為上界200下界77
#因為只有第449筆為0其他都有測到，故上界以最大值200，下界以平均減3個標準差內的77
#呼叫涵式
index,para = outlier_detect_arbitrary(data,'RestingBP',200,77)
print('Upper bound:',para[0],'\nLower bound:',para[1])
#印出來測試有無改到
print('Upper bound:',para[0],'\nLower bound:',para[1])
print('before handling outlier:')
print(data.loc[449]['RestingBP'])
print('after handling outlier:')
# see index 258,263,271 have been replaced with top/bottom coding
data3 = windsorization(data=data,col='RestingBP',para=para,strategy='both')
print(data3.loc[449]['RestingBP'])
```

```
Num of outlier detected: 8
Proportion of outlier detected 0.008714596949891068
Upper bound: 187.93897652094347
Lower bound: 76.85405180149661
449      0
109    190
592    190
759    192
241    200
365    200
399    200
732    200
Name: RestingBP, dtype: int64
Upper bound: 200
Lower bound: 77
Upper bound: 200
Lower bound: 77
before handling outlier:
0.0
after handling outlier:
77.0
```

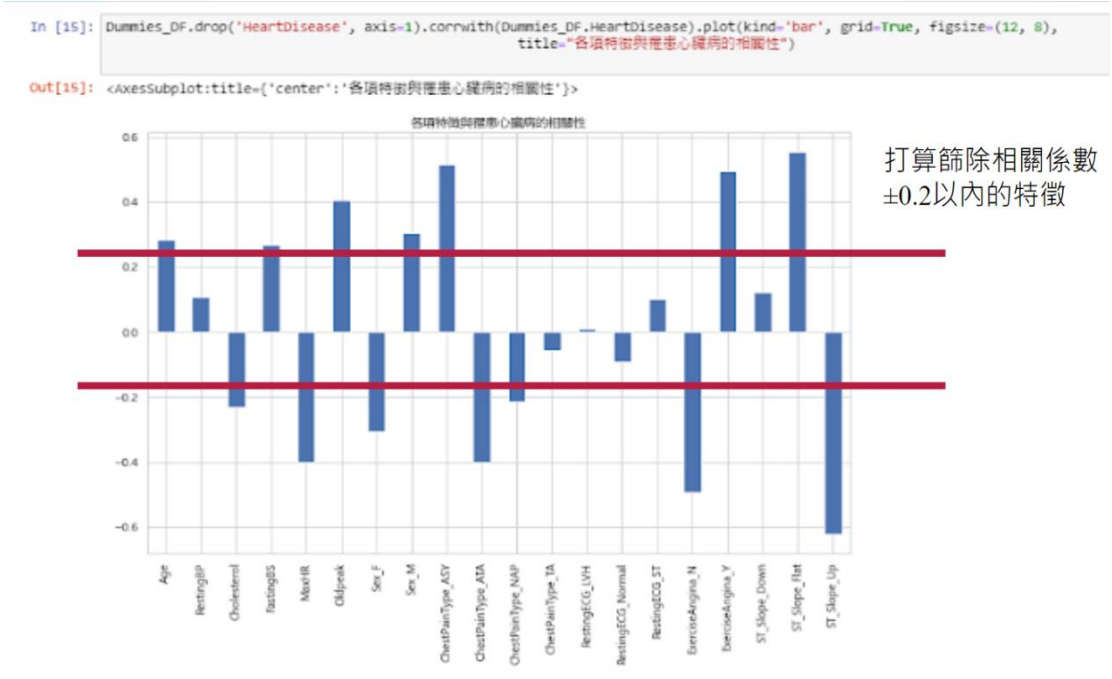
圖十九、離群值處理

3. 特徵選擇

本組分別做了三次特徵選擇，分別是：特徵全選擇、手動選擇（如圖二十）、篩除選擇（如圖二十一）




圖二十、手動選擇關聯係數較大的特徵




圖二十一、篩除選擇特徵

## 6、 模型建立

1. 訓練集與測試集

此次本組以 80%之資料作為訓練集，20%做為測試集（如圖二十二）並做不同特徵下的訓練模型

```
#選取全特徵

X = Dummies_DF.drop(columns=['HeartDisease'])

y = Dummies_DF['HeartDisease'].values
X_train, X_test, y_train, y_test =  train_test_split(X, y,
                                    test_size=0.2,random_state = 1)


#0.3
'''
X = Dummies_DF[['MaxHR','Oldpeak','Sex_F','Sex_M','ChestPainType_ASY','ChestPainType_ATA','ChestPainType_NAP',
                'ExerciseAngina_N','ExerciseAngina_Y','ST_Slope_Up','ST_Slope_Flat']].values
y = Dummies_DF['HeartDisease'].values
X_train, X_test, y_train, y_test =  train_test_split(X, y,
                                    test_size=0.2,random_state = 1)
'''
#0.2
'''
X = Dummies_DF[['Age','Cholesterol','FastingBS','MaxHR','Oldpeak','Sex_F','Sex_M','ChestPainType_ASY','ChestPainType_ATA',
                'ExerciseAngina_N','ExerciseAngina_Y','ST_Slope_Up','ST_Slope_Flat']].values
y = Dummies_DF['HeartDisease'].values
X_train, X_test, y_train, y_test =  train_test_split(X, y,
                                    test_size=0.2,random_state = 1)
'''

"\nX = Dummies_DF[['Age','Cholesterol','FastingBS','MaxHR','Oldpeak','Sex_F','Sex_M','ChestPainType_ASY','ChestPainType_ATA',\n                'ExerciseAngina_N','ExerciseAngina_Y','ST_Slope_Up','ST_Slope_Flat']].values  \ny = Dummies_DF['HeartDisease'].values  \nX_train, X_test, y_train, y_test =  train_test_split(X, y,\n                                   test_size=0.2,random_state = 1)\n"
```

圖二十二、選取不同特徵方法訓練模型

2. 模型預測：SVM、決策樹、隨機森林

此次預測訓練 SVC、決策樹、隨機森林等三種工具建立模型。經過測試之後，以篩除相關係數±0.2 內之特徵，三個模型準確度將近 91%。

**SVC**

```
scaler = preprocessing.StandardScaler().fit(X_train)
X_train = scaler.transform(X_train)

model = SVC(kernel='rbf', C=1)
model.fit(X_train, y_train)

X_test = scaler.transform(X_test)
y_pred = model.predict(X_test)


accuracy = accuracy_score(y_test, y_pred)
num_correct_samples = accuracy_score(y_test, y_pred, normalize=False)
con_matrix = confusion_matrix(y_test, y_pred)

print('number of correct sample: {}'.format(num_correct_samples))
print('accuracy: {}'.format(accuracy))
print('con_matrix: {}'.format(con_matrix))
```

```
number of correct sample: 167
accuracy: 0.907608695652174
con_matrix: [[ 64  10]
 [  7 103]]
```

圖二十三、SVC

## 決策樹

```
model = DecisionTreeClassifier(max_depth=5)
model.fit(X_train, y_train)

X_test = scaler.transform(X_test)
y_pred = model.predict(X_test)


accuracy = accuracy_score(y_test, y_pred)
num_correct_samples = accuracy_score(y_test, y_pred, normalize=False)
con_matrix = confusion_matrix(y_test, y_pred)

print('number of correct sample: {}'.format(num_correct_samples))
print('accuracy: {}'.format(accuracy))
print('con_matrix: {}'.format(con_matrix))
```

```
number of correct sample: 167
accuracy: 0.907608695652174
con_matrix: [[ 65   9]
 [  8 102]]
```

圖二十四、決策樹

## 隨機森林

```
regressor = RandomForestRegressor(n_estimators = 75,max_depth=4)
regressor.fit(X_train,y_train)
y_pred=regressor.predict(X_test)
y_pred=(y_pred>0.5)

y_test=list(y_test)
y_pred=list(y_pred)


pred_train=regressor.predict(X_train)
pred_train=(pred_train>0.5)

print("The accuracy for Random Forest Regressor= "+str(accuracy_score(y_train,pred_train)*100))

print("The Test accuracy for Random Forest Regressor= "+str(accuracy_score(y_test,y_pred)*100))
```

```
The accuracy for Random Forest Regressor= 90.19073569482289
The Test accuracy for Random Forest Regressor= 90.76086956521739
```

圖二十五、隨機森林

## 7、 結語與後續討論

1. 此資料集經分析後，部分資料與臨床上有所差異，或許是導致準確度無法更精確之原因。於此部分後續可在思考討論如何處理。如：不納入該項特徵，或以 one hot encoding 或其他方式將資料轉換，使該資料之影響變小，或許能讓準確度提升。
2. 此次嘗試以抓取不同特徵之方式建立模型，但仍有其他方式，例如如何填補缺失值、極端值之調整、訓練集與測試集之比例、工具之選擇...這些後續都仍可繼續討論研究。
3. 雖此次執行準確度未能比原作者更佳，但是對整個機器學習的操作流程更加深刻，同時也更加熟悉資料視覺化及特徵化工程的執行。
4. 該資料集因為主要提供練習使用，故在資料類型上較少，僅部分與心臟病有關之資料。未來除了樣本增加測試，可加入 BMI、抽菸習慣、家族病史等更多特徵，作更多此資料處理後測試。
5. 針對此次題目，也得知中國醫學大學及長庚大學在心臟病預測模型以有些實質應用之成果，或許可作為後續此研究之參考。

## 8、 參考資料


1. WHO: Global Health Estimates https://www.who.int/data/global-health-estimates/
2. Machine learning overtakes humans in predicting death or heart attack. European Society of Cardiology https://www.escardio.org/The-ESC/Press-Office/Press-releases/machine-learning-overtakes-humans-in-predicting-death-or-heart-attack
3. 衛生福利部國民健康署：
   https://www.hpa.gov.tw/Pages/Detail.aspx?nodeid=632&pid=1188
4. Kaggle: Heart Disease Prediction (95% Accuracy & Recall)
   https://www.kaggle.com/code/fearsomejockey/heart-disease-prediction-95-accuracy-recall/notebook
5. 美國心臟協會 The American Heart Association. https://www.heart.org/