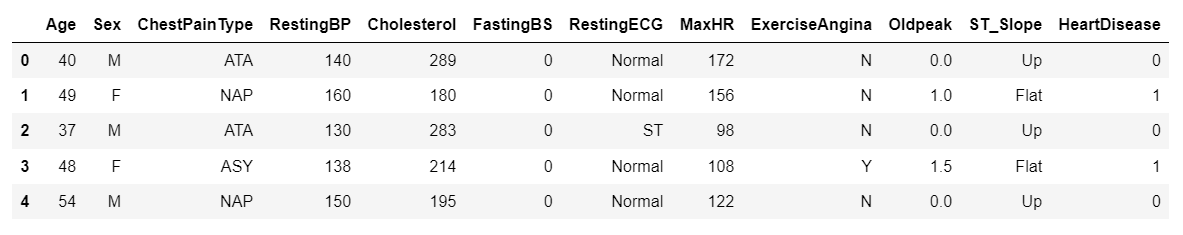
數資系 110813037 余珮綺

機器學習HW2\_Classification

程式碼語言:

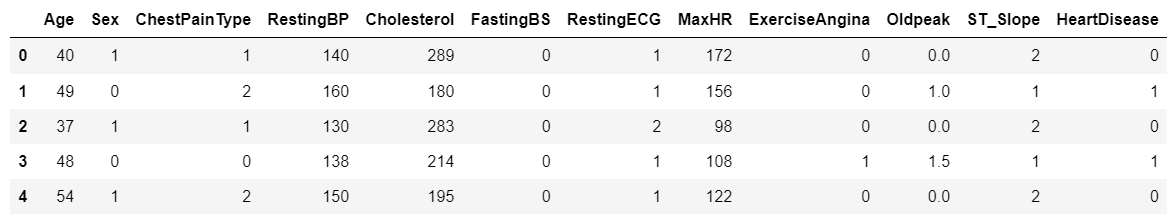
內容:

1. 資料清理與視覺化圖表



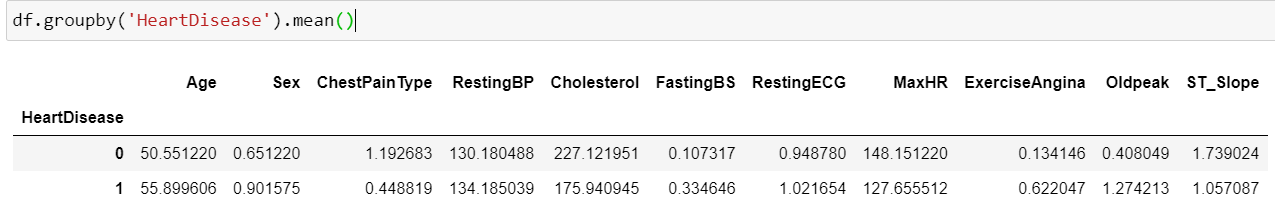


首先，先查看資料的狀況，可看見有許多欄位為類別型資料，所以後續還需轉換為數值型，此外並無看到有缺失值。

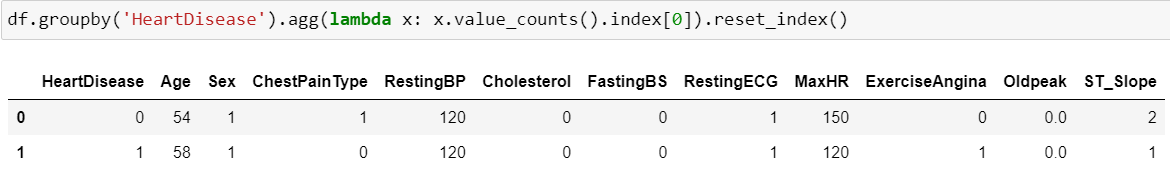


考量到接下來透過圖表理解資料狀況會更直觀，因此先將類別型資料先用label encoder而非one hot encoder。

1. 敘述性統計分析



將有無心臟疾病的人分類後，查看個欄位的資料平均值狀況，進行大致上的推測：有疾病的人資料分布大概都靠近哪個數值。



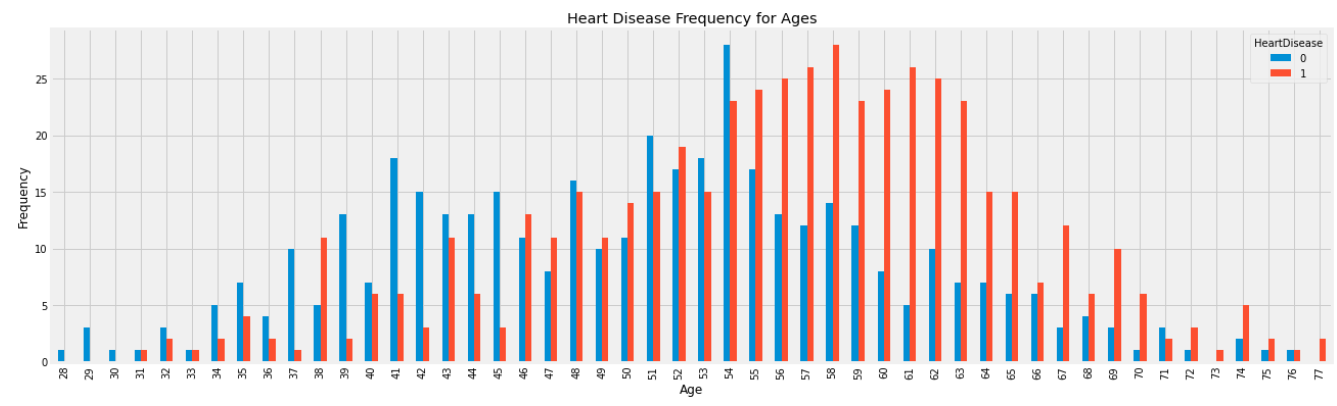
此外也看看有無心臟疾病的人，各欄位最多的人診斷出的結果為何。

綜合以上，可以推測有心臟疾病的人通常年紀集中在55~58歲，男生的人數也比較多，通常Chest Pain Type的數值通常比較小，最多的是0這個類別，然後Max HR有明顯比無疾病的人小，數值大約落在120~127，而Exercise Angina和ST Slope通常也是類別1這種狀態。

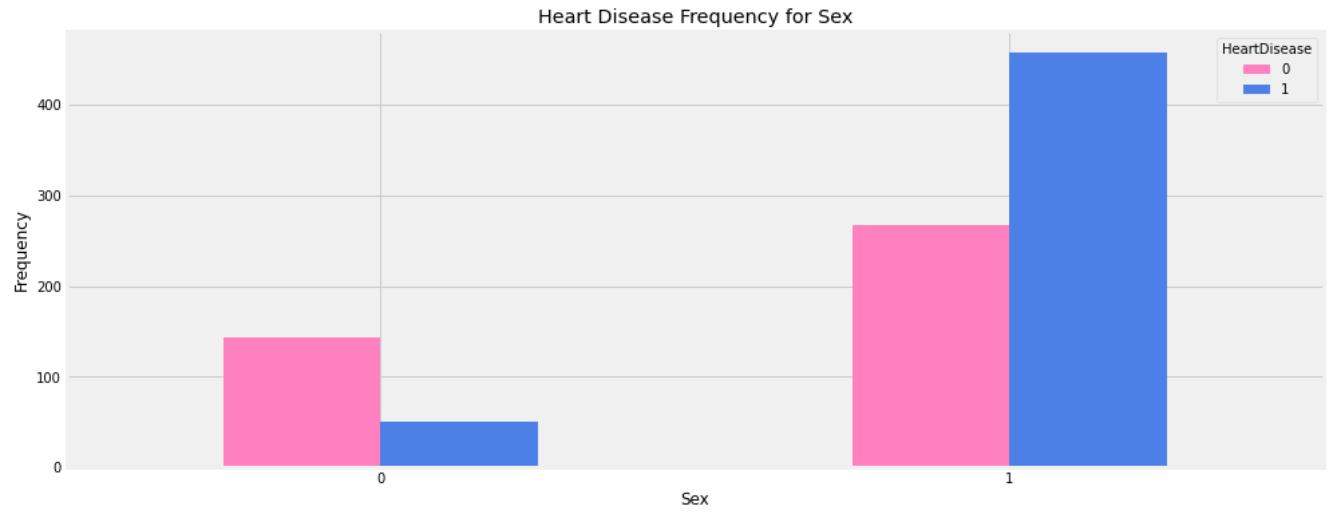


然後也看一下整體數值的分布狀況。

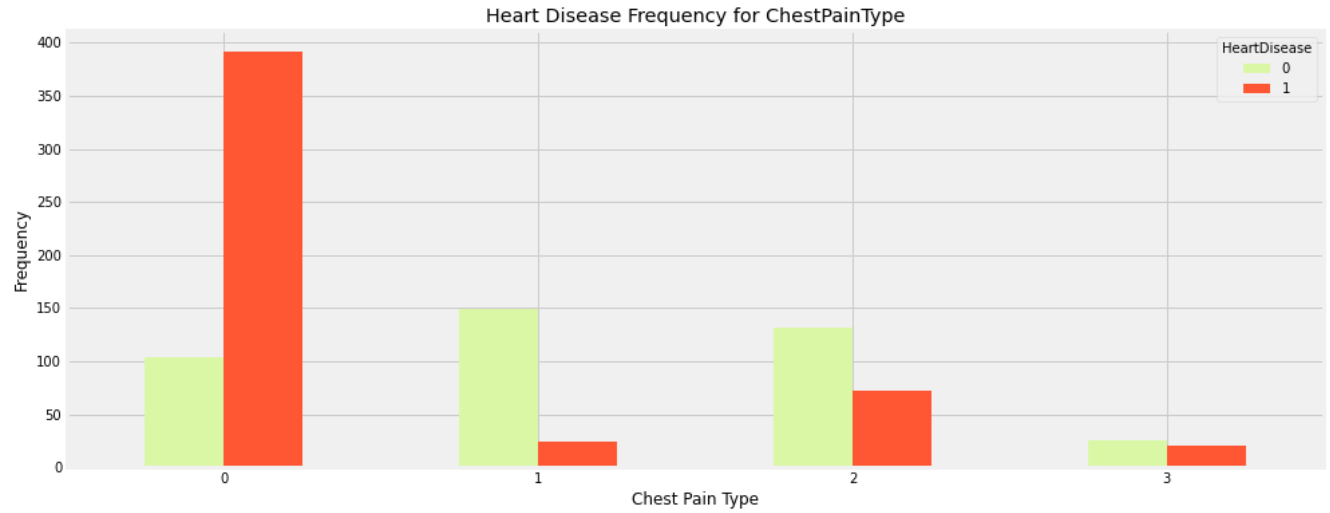
1. 特徵相關性分析



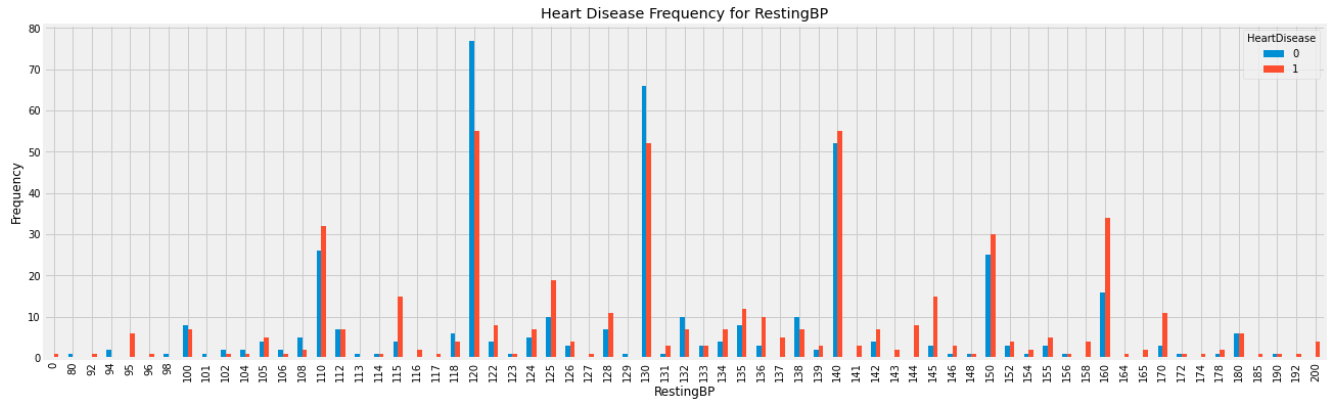
用圖表可更明顯看出有疾病的人歲數的集中範圍在一定區域。



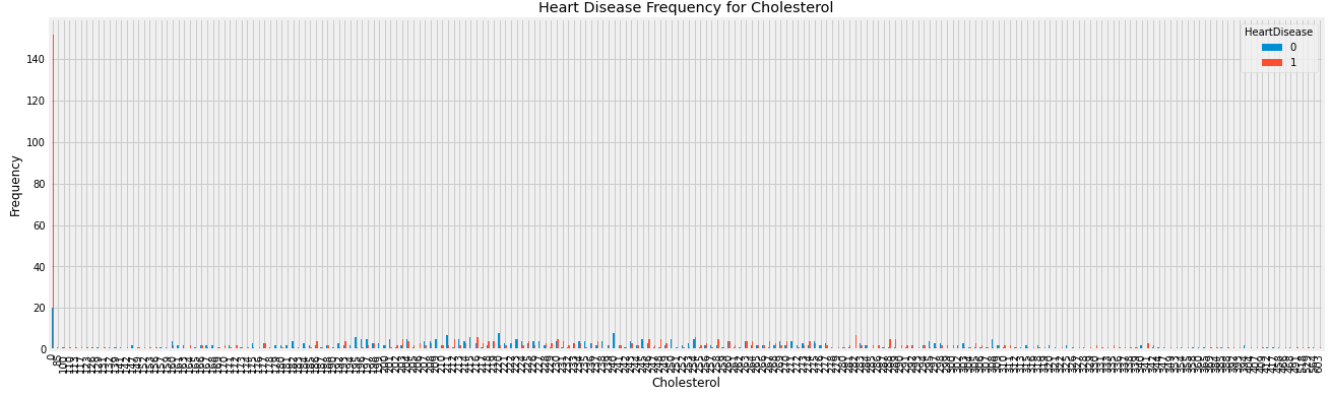
再來是性別和有無心臟疾病的圖表，可看出男生總人數比女生多，以及有心臟疾病的人大多都為男性。



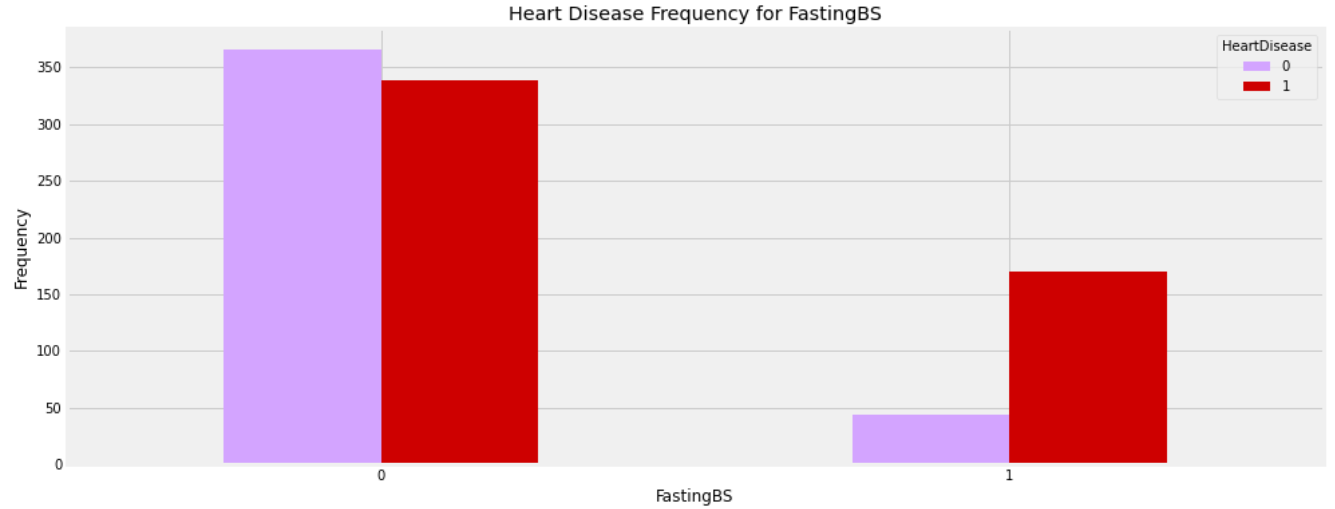
接下來看到的是Chest Pain Type很明顯的有疾病的人與0這個類別的型態有關聯。



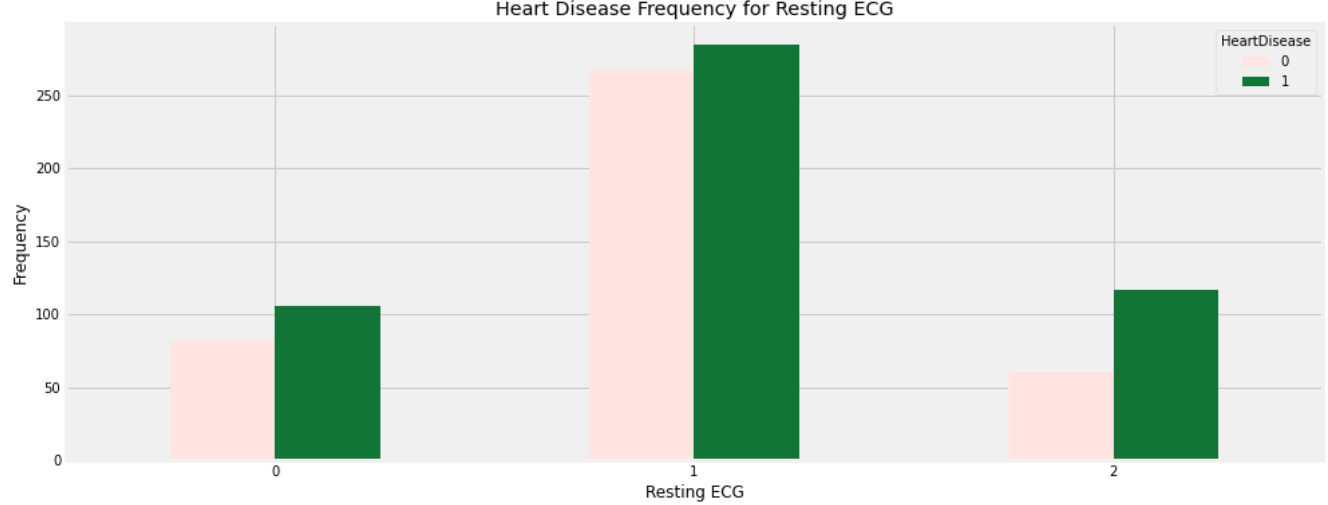
至於Resting BP好像不管有無心臟疾病，大部分的人都落在特定的些數字。



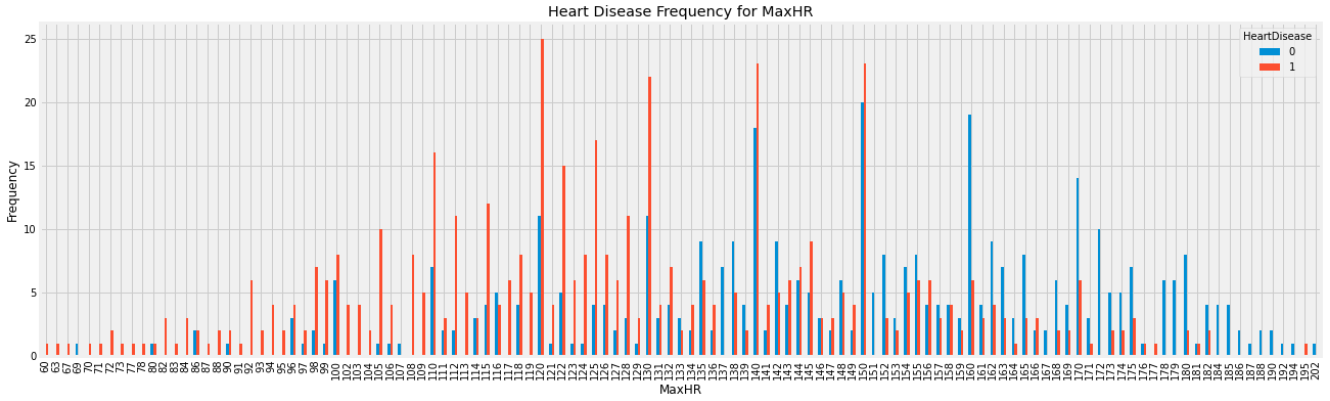
雖然這張圖表的分布因刻度太多太密集，所以不是很清楚，但還是可以見到有心臟疾病的人Cholesterol都落在某個數值。



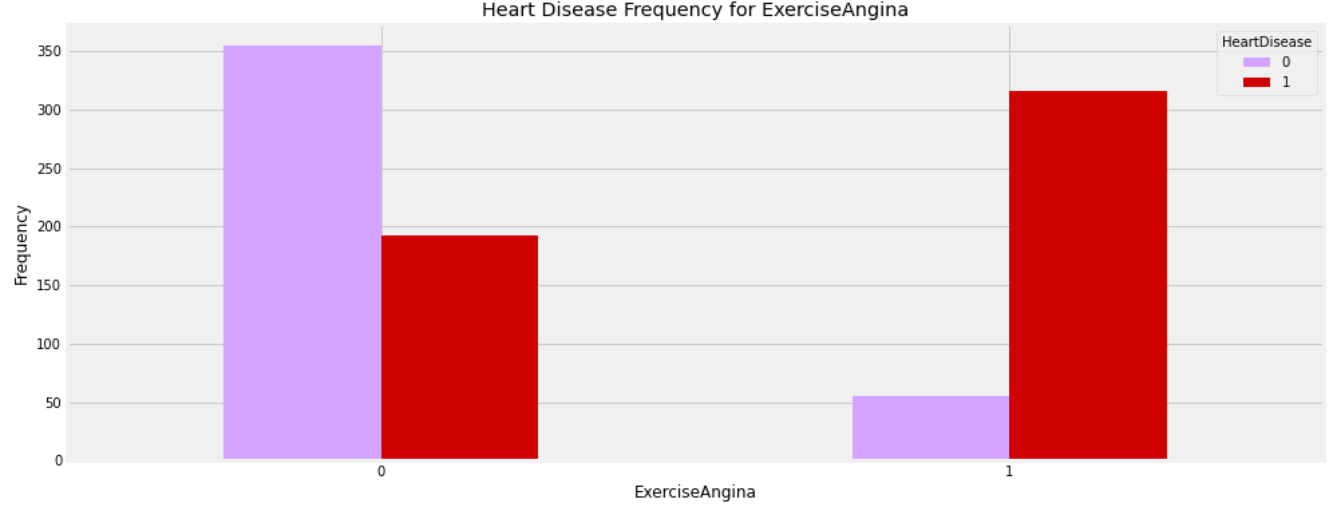
從這張圖表可見到FastingBS與心臟疾病有一定關聯，多數人患者還是落在0的類別。



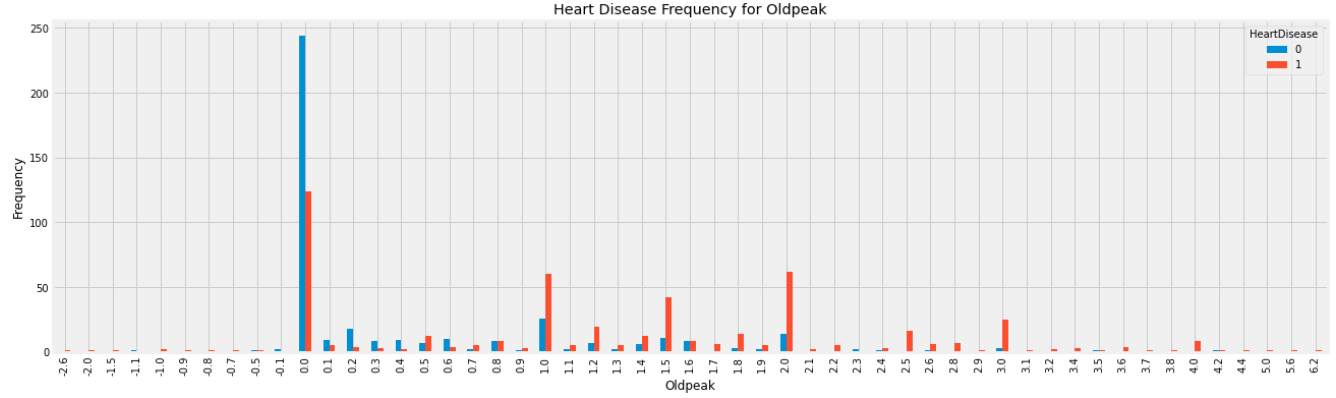
Resting ECG的圖表可顯示多數心臟疾病患者為1這個類別的身體數值狀態。



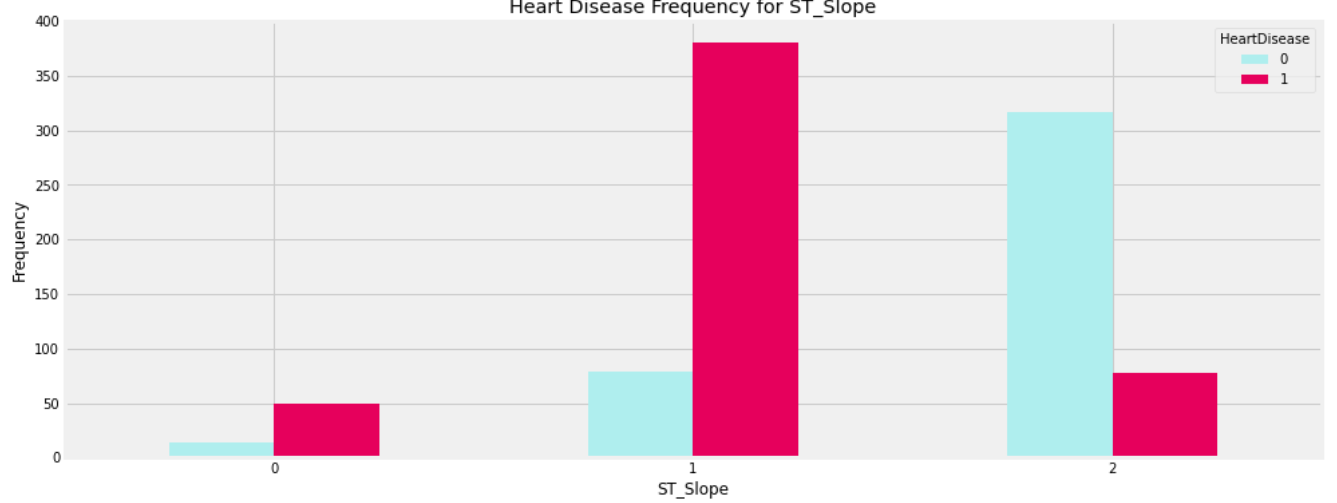
在Max HR圖表中，可應證剛剛從平均值與眾數的分析中看到的，患者多數的數值比健康的人還要小。



Exercise Angina圖表中可看出患者多數為1類別的身體狀態。



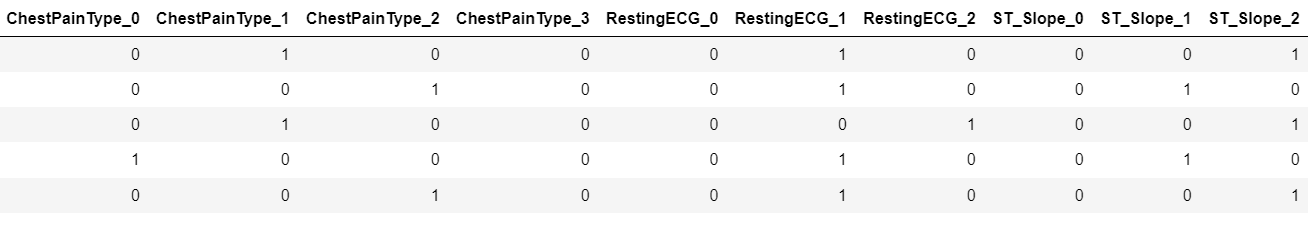
然而Oldpeak中發現似乎不管健康的人還是患者，多數都是數值0.0，所以這項類別與診斷的關聯度持保留態度。



從ST\_Slope很明顯可看到患者多數為1類別的狀態，健康的人則是2這個類別。

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

從以上分析，雖然無法得出RestingBP, Oldpeak這兩個類別與診斷結果的是否具高度相關，但還是先不將之排除納入模型的建置。

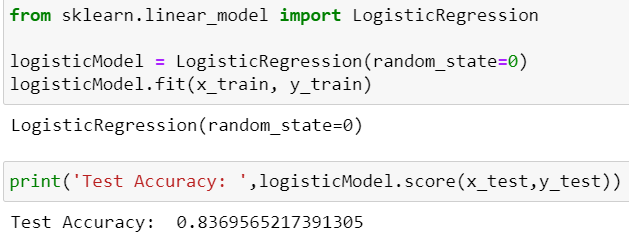


首先先將前面的label encoder後的欄位變成one hot encoder，讓類別型的資料再判斷的過程中不會產生有大小的順序的誤會。



再來就是先將數值標準化後進行訓練和測試集的資料分割。

* Logistic regression

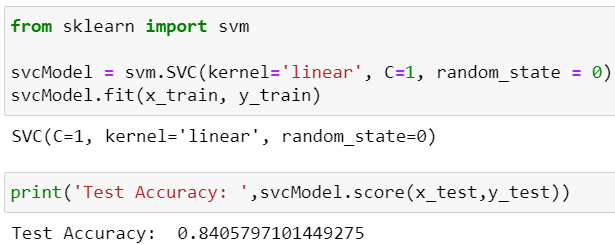


首先是Logistic regression，其準確度約為83.69%

* SVM

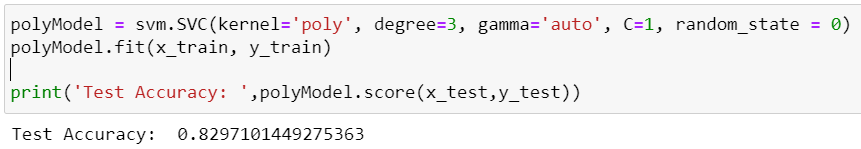
因SVM有三種kernel，所以都試試看哪個準確度最高。

* Linear



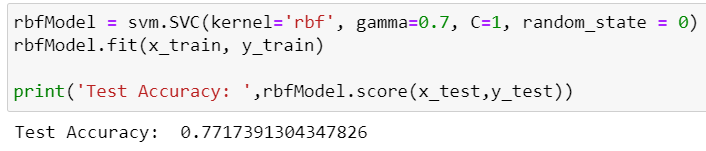
首先，linear的準確度達84.05%。

* Poly



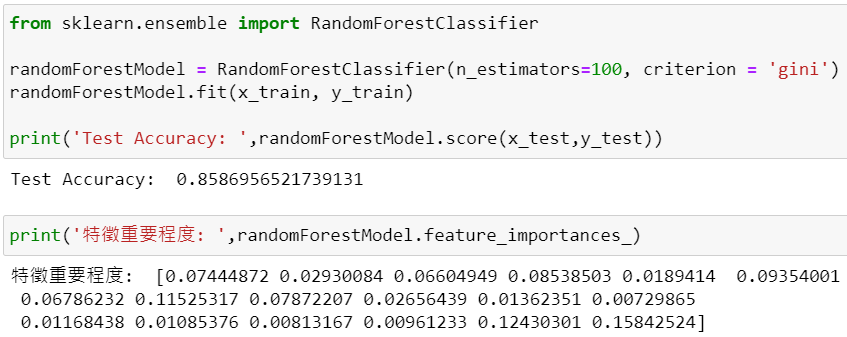
再來poly的準確度則是82.97%。

* Rbf



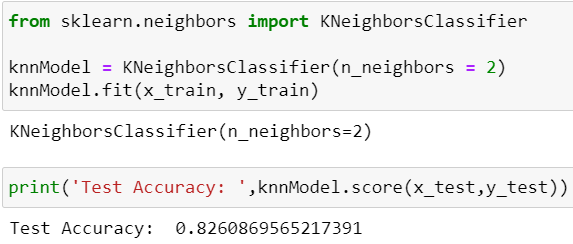
最後，預設的RBF準確度最差為77.17%。

* Random forest

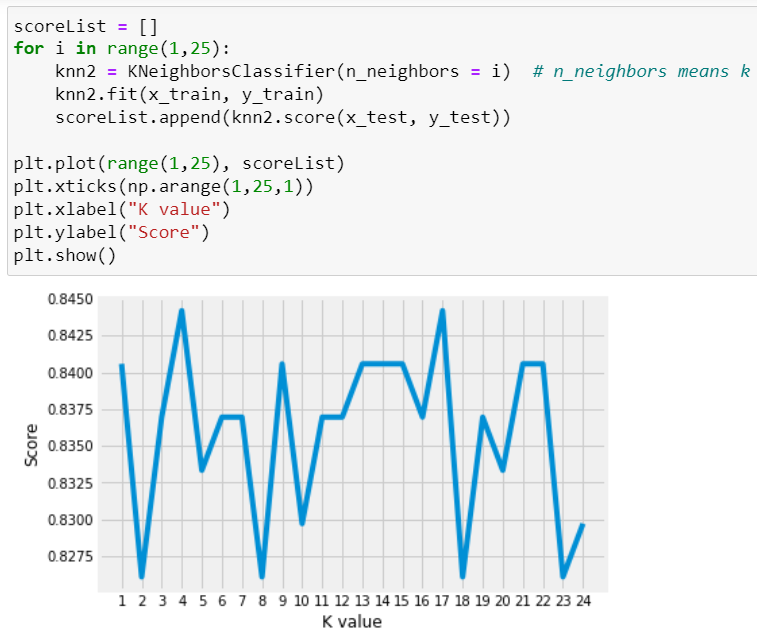


接著看到Random Forest的準確度在85.86%，另外也察看一下其判斷哪些欄位是重要的特徵值，可看到超過0.1的Oldpeak, ST\_Slope\_1, ST\_Slope\_2是最重要的，再來是超過0.07的Sex, Cholesterol, MaxHR, ChestPainType\_0。

* KNN



最後則是KNN，可看出當n=2時，準確度為82.60%。因此近一步看看當n值等於多少時準確度是最高的，發現當n=4或17時，準確度接近84.50%。



1. 綜合比較4個模型的分類結果與分析討論

* Logistic regression: 83.69%
* SVM
* Linear: 84.05%
* Poly: 82.97%
* RBF: 77.17%
* Random forest: 85.86%
* KNN: 最高接近84.50%

綜合以上，可以看出在這樣的情況下最推薦使用Random forest的模型進行預測，最不建議SVM的RBF kernel。可能是因為Random forest基於Decision Tree if-then-else的特性

，可以判斷出不同特徵之間的相互影響，更好的得出結果，而且也可以得出哪個欄位是重要的。