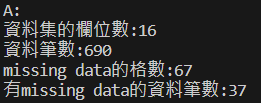
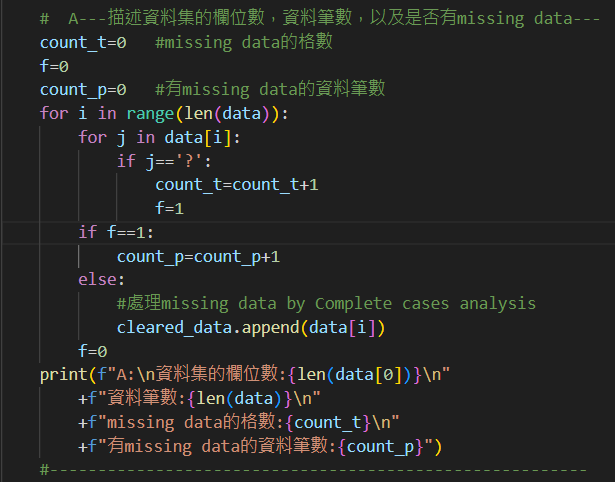
根據你的學號最後一個尾數除以6的餘數，解開所附的hw03\_data.zip檔案，選擇對應的資料夾下的資料做以下的事情：

A. 描述資料集的欄位數，資料筆數，以及是否有missing data。

A: 資料集的欄位數:16，資料筆數:690，missing data的格數:67，有missing data的資料筆數:37。





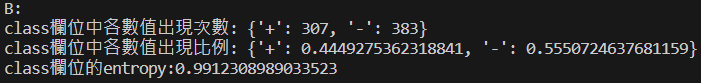
B. 分析資料集中最一個欄位(除資料集4為RiskLevel外,其他為class)的分佈，包括

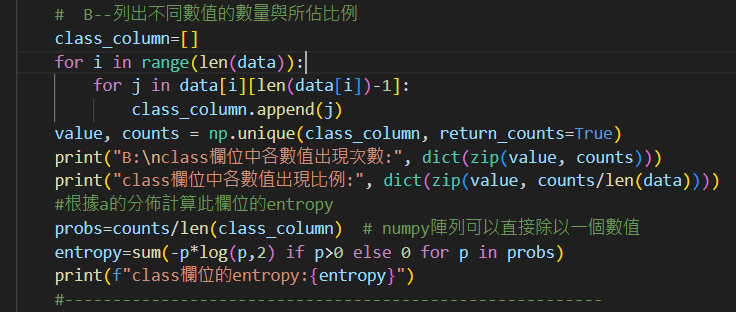
* 列出不同數值的數量與所佔比例
* 根據a的分佈計算此欄位的entropy

A: class欄位中各數值出現次數: {'+': 307, '-': 383}，

class欄位中各數值出現比例: {'+': 0.4449275362318841, '-': 0.5550724637681159}。

class欄位的entropy:0.9912308989033523。





C. 從KNN,Decision Tree,或是Naive Bayes中三選一，建立模型並回報模型的準確度。注意

* 以0.8/0.2的比例將資料集分成訓練與測試用資料
* 明確敘述你在建立模型過程中手動設定的所有參數(例如train\_test\_split中的random\_state,KNN的K值等)

A: 建立模型過程中手動設定的所有參數有: 在函式train\_test\_split中test\_size=0.2,random\_state=50。DecisionTreeClassifier(criterion="entropy")是指使用entropy作為節點分裂的指標。

Accuracy=0.8091603053435115





繳交內容：

1. 你所用的資料檔案
2. 可以重現你的結果的程式碼，並在以下幾個部份以加上註解的方式標注出來(在python請在註解前加上井字號#為開頭)
   * 將資料切成訓練/測試的地方
   * 產生模型與訓練模型的的地方
   * 計算(手動或呼叫套件提供方法)準確度的地方
3. 一個pdf檔，內容是上述A,B,以及C的部份

將所有內容以zip格式壓縮成一個檔案之後上傳。

其他事項：

* 使用Decision Tree的同學如果將你所建立出來的樹轉成png圖檔輸出並附上來，會有額外加分。

