大數據分析與資料探勘-第五次作業

繳交日期: 2022/5/31

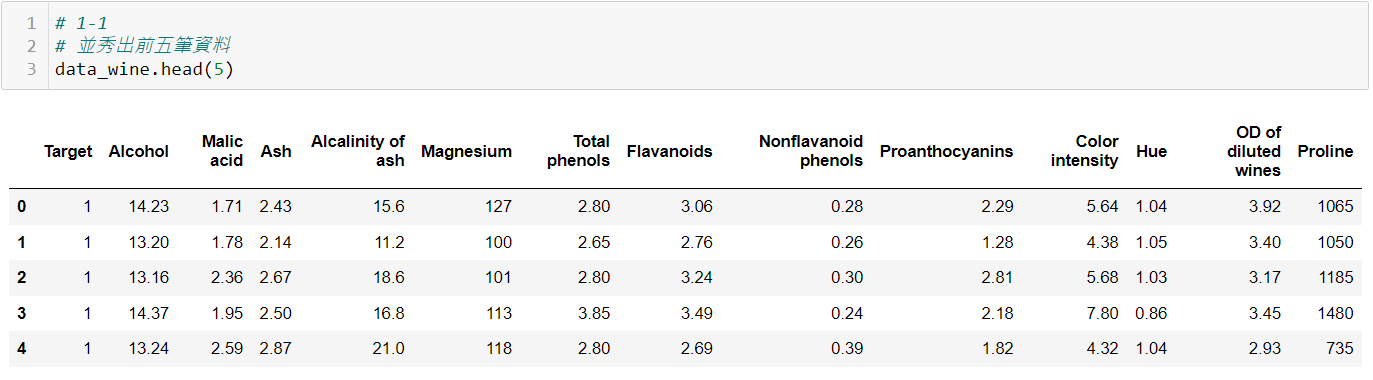
使用Python程式進行分析並回答以下所有問題。請直接在word檔每個問題的下面，貼上python程式與執行結果，最後將檔案上傳到MOODLE，包含此word檔及python程式檔(佐證用)。兩個檔案麻煩請壓縮成一個檔在上傳，檔名請命名為Homework5\_姓名.zip，另程式碼中請詳細標明對應的題數，若看不懂的話，一律不給分。上傳期限為5/31晚上11點59分前，超過繳交期限就無法上傳。

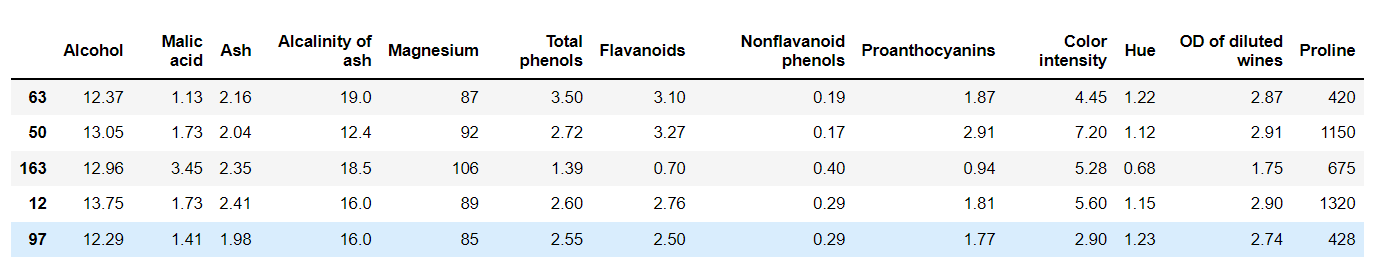
1.

1. 首先，請使用資料讀取方式，將附件中的csv檔(wine.csv)讀取到python的DataFrame，並秀出前五筆資料。此檔案為紅酒資料集，各欄位代表的意義，分別簡述如下：Target為紅酒的分類(總共分為3類，分別為1~3)。其他資料為紅酒中各項化學成分檢驗結果，包含：酒精(Alcohol)、蘋果酸(Malic acid)、灰(Ash)、灰的鹼度(Alcalinity of ash)、鎂(Magnesium)、總酚(Total phenols)、類黃酮(Flavanoids)、黃酮類酚(Nonflavanoid phenols)、原花青素(Proanthocyanins)、色彩強度(Color intensity)、色調(Hue)、稀釋酒的OD、脯氨酸(Proline)，共13個檢驗項目。

一張含有 文字 的圖片

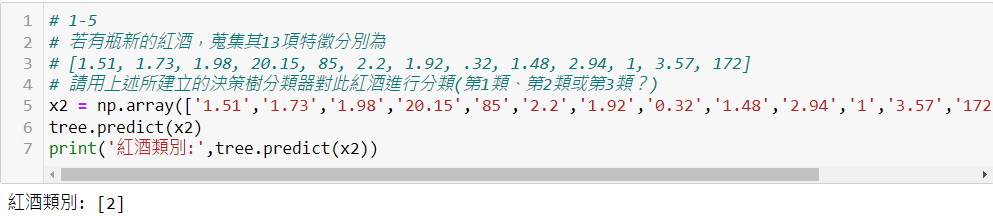
自動產生的描述



1. 一張含有 文字 的圖片

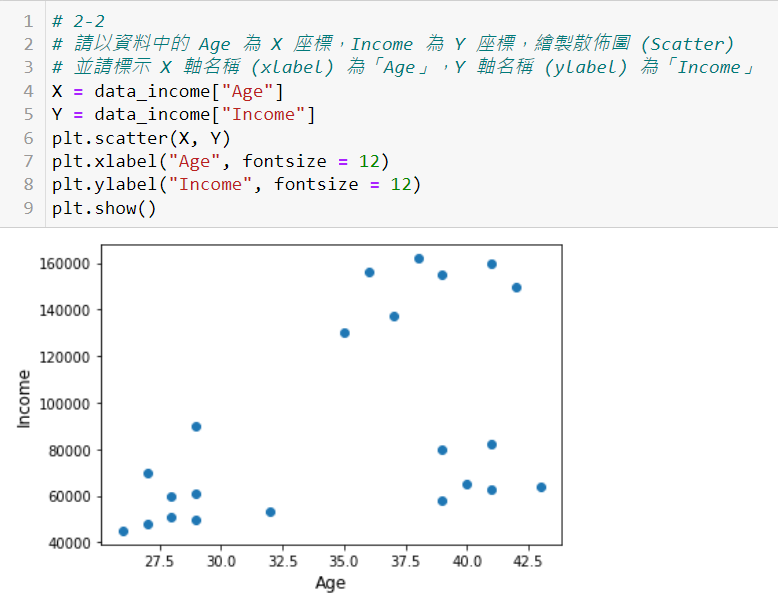
   自動產生的描述接著將13個檢驗項目同時放入X變數，而Y變數為紅酒的分類(Target)，請使用sklearn中的函式把X、Y變數分為training及test datasets。其中，training的資料占75%、test 25%，並設定Random\_state = 5，並秀出 X\_train的前五筆資料。
2. 一張含有 文字 的圖片

   自動產生的描述請載入決策樹函式，criterion使用Gini impurity (其他參數為預設值)，使用train dataset訓練此模型。再利用所建構的決策樹模型針對test dataset進行分類，並與實際分類結果作比對，印出比對後的錯誤個數。
3. 一張含有 文字 的圖片

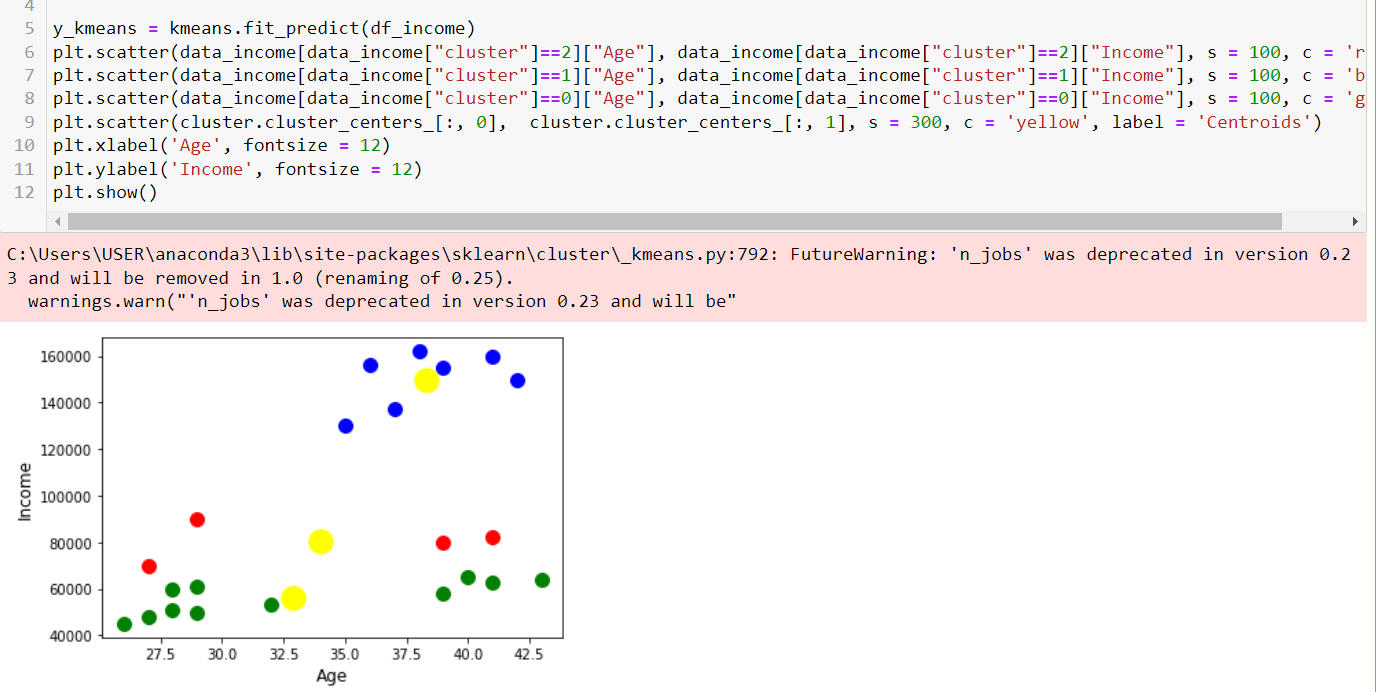
   自動產生的描述承上題，請計算針對test dataset分類結果的正確率(Accuracy)，請顯示至小數點第四位。
4. 若有瓶新的紅酒，蒐集其13項特徵分別為[1.51, 1.73, 1.98, 20.15, 85, 2.2, 1.92, .32, 1.48, 2.94, 1, 3.57, 172]，請用上述所建立的決策樹分類器對此紅酒進行分類(第1類、第2類或第3類？)

2.

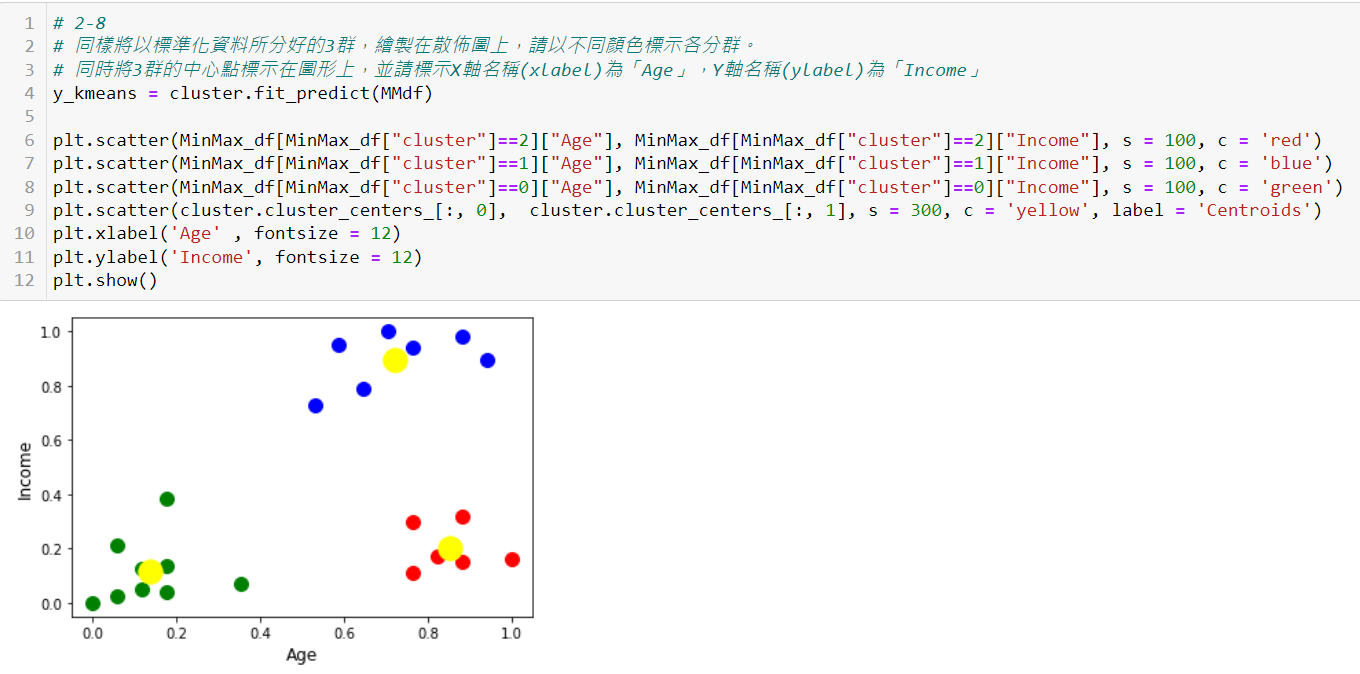
1. 一張含有 文字 的圖片

   自動產生的描述首先，請使用資料讀取方式，將附件中的Excel檔(income.xlsx)讀取到python的DataFrame，並秀出前五筆資料。
2. 請以資料中的Age為X座標，Income為Y座標，繪製散佈圖(Scatter)，並請標示X軸名稱(xlabel)為「Age」，Y軸名稱(ylabel)為「Income」。
3. 一張含有 文字 的圖片

   自動產生的描述使用KMean法針對原始資料(Age與Income)進行分群，請設定群集數目為3、random\_state為0 (其他參數不須設定)。
4. 一張含有 文字 的圖片

   自動產生的描述請在原匯入的DataFrame之最後加上一欄(column)呈現各資料點的分群結果，同時設定column名稱為「cluster」，並秀出前五筆資料。
5. 請將分好的3群，繪製在散佈圖上，請以不同顏色標示各分群。同時將3群的中心點標示在圖形上，並請標示X軸名稱(xlabel)為「Age」，Y軸名稱(ylabel)為「Income」。
6. 一張含有 文字 的圖片

   自動產生的描述請以最小-最大的標準化方法，分別針對Age與Income進行標準化。並秀出標準化的前五筆資料。
7. 一張含有 文字 的圖片

   自動產生的描述請使用KMean法針對最小-最大的標準化後的資料(標準化後的Age與Income)進行分群，同樣設定群集數目為3、random\_state為0 (其他參數不須設定)。
8. 同樣將以標準化資料所分好的3群，繪製在散佈圖上，請以不同顏色標示各分群。同時將3群的中心點標示在圖形上，並請標示X軸名稱(xlabel)為「Age」，Y軸名稱(ylabel)為「Income」。
9. 請以圖形或分群結果判斷在此案例下，「未經標準化」與「經過標準化後」的分群結果哪一個比較好？
10. 一張含有 文字 的圖片

    自動產生的描述請以上課所教的Elbow Method，透過迴圈方式，將標準化後的資料設定不同群集(請設定1~10群)，畫出群集數與wcss的圖形，請以轉折點判斷上述分群數設定為3是否合理？

