大數據分析與資料探勘-第五次作業

繳交日期: 2023/5/31

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；此外，若檔案有缺少的話，亦會扣分。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. 一張含有 螢幕擷取畫面, 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csv檔(wine.csv)讀取到python的DataFrame，並秀出前十筆資料。註：此檔案為紅酒資料集，各欄位代表的意義，分別簡述如下：Target欄位為紅酒的分類(總共分為3類，分別為1~3)。其他欄位分別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。
2. 一張含有 文字, 螢幕擷取畫面, 軟體 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70%、test 30%，並設定random\_state =42，請秀出 X\_test的前十筆資料。
3. 一張含有 文字, 螢幕擷取畫面, 軟體, 陳列 的圖片

   自動產生的描述請載入決策樹函式，criterion使用entropy (random\_state 設定為42，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。
4. 一張含有 文字, 螢幕擷取畫面, 字型 的圖片

   自動產生的描述承上題，請進一步顯示測試集資料分類結果的混淆矩陣。
5. 承上題，請計算針對測試集分類結果的正確率(Accuracy)，請顯示至小數點第四位。

一張含有 文字, 螢幕擷取畫面, 字型, 行 的圖片

自動產生的描述

1. 一張含有 文字, 螢幕擷取畫面, 字型, 軟體 的圖片

   自動產生的描述承上題，請一次顯示出測試集資料分類結果的多個評估指標，並將這些評估指標的值皆顯示至小數點第四位。
2. 承上題，請使用export\_graphviz函式將上述所建構出的決策樹加以視覺化，決策樹中請顯示feature\_names與class\_numbers。

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

一張含有 文字, 行, 圖表, 設計 的圖片

自動產生的描述

1. 一張含有 文字, 螢幕擷取畫面, 字型, 多媒體軟體 的圖片

   自動產生的描述若有瓶新的紅酒，蒐集其13項特徵分別為[13.71, 5.65, 2.45, 20.5, 95, 1.68, .61, .52, 1.06, 7.7, .64, 1.74, 720]，請用上述所建立的決策樹分類器對此新的紅酒進行分類(第1類、第2類或第3類？)
2. 請載入隨機森林函式，criterion同樣使用entropy (random\_state 亦設定為42，n\_estimators設定為100，其他參數為預設值)，同樣使用小題(2)分割後的train dataset訓練模型。再利用所建構的隨機森林模型針對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, 螢幕擷取畫面, 字型 的圖片

   自動產生的描述承上題，請計算隨機森林模型針對測試集分類結果的正確率(Accuracy)，請顯示至小數點第四位。

一張含有 文字, 字型, 螢幕擷取畫面 的圖片

自動產生的描述

1. 請以任一個分類評估指標判斷在此案例下，「決策樹」與「隨機森林」哪一種的分類結果比較好？

以準確率來看，隨機森林的準確率到1.0000，也就是無任何錯誤，相較決策樹的準確率0.8519表現好上不少，因此我認為隨機森林的分類結果比較好。

2.

1. 首先，請使用資料讀取方式，將附件中的csv檔(Mall\_Customers. csv)讀取到python的DataFrame，並秀出前十筆資料。

一張含有 文字, 螢幕擷取畫面, 軟體, 多媒體軟體 的圖片

自動產生的描述

1. 一張含有 螢幕擷取畫面, 文字, 圖表 的圖片

   自動產生的描述一張含有 文字, 螢幕擷取畫面, 字型, 設計 的圖片

   自動產生的描述請以此資料集中的Annual Income (k$)欄位為X座標，Spending Score (1-100)欄位為Y座標，繪製散佈圖(Scatter)，並請標示X軸名稱(xlabel)為「Annual Income (k$)」，Y軸名稱(ylabel)為「Spending Score (1-100)」。
2. 一張含有 文字, 螢幕擷取畫面, 字型, 軟體 的圖片

   自動產生的描述由散佈圖顯示，資料約可區分為5群，請使用KMean法針對Annual Income (k$)欄位及Spending Score (1-100) 欄位進行分群，請設定群集數目為5、random\_state為0 (其他參數不須設定)，請秀出分群後每一群的中心點。
3. 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」，並秀出前十筆一張含有 文字, 螢幕擷取畫面, 軟體, 多媒體軟體 的圖片

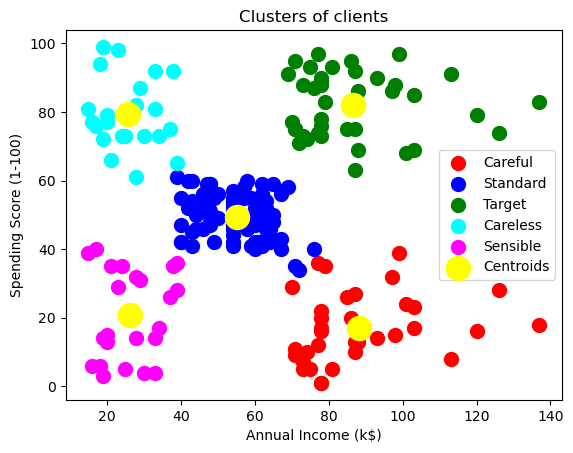
   自動產生的描述資料。
4. 一張含有 文字, 行, 螢幕擷取畫面, 圖表 的圖片

   自動產生的描述一張含有 文字, 螢幕擷取畫面, 字型, 陳列 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Annual Income (k$)欄位與Spending Score (1-100)欄位設定不同群集(請設定1~10群)，畫出群集數與WCSS的圖形。圖形中請標示X軸名稱(xlabel)為「Number of Clusters」，Y軸名稱(ylabel)為「WCSS」，圖表名稱為「The Elbow Method」。
5. 承上題，請以Elbow Method圖形的轉折點判斷上述KMean法分群數設定為5是否合理？

根據Elbow Method圖形，我們可以觀察到當群集數從1逐漸增加至5時，WCSS值急速下降。然而，從群集數為5之後，WCSS的下降趨勢趨於平緩。因此，根據Elbow Method的結果，可以認為將KMeans法的群集數設定為5是合理的，因為在群集數為5之後，WCSS的下降幅度不再明顯。這表示5個群集可以有效地區分資料，並且在保持群集數目合理的前提下，提供了足夠的區分度。

1. 最後，將小題(3)的所分好的5群，繪製散佈圖，請以不同顏色標示各分群與分群中心點。其中，cluster為0以紅色標示，標籤為Careful，cluster為1以藍色標示，標籤為Standard，cluster為2以綠色標示，標籤為Target，cluster為3以青色(cyan)標示，標籤為Careless，cluster為4以紅紫色(magenta)標示，標籤為Sensible，中心點以黃色標示，標籤為Centroids。此外，並請標示X軸名稱(xlabel)為「Annual Income (k$)」，Y軸名稱(ylabel)為「Spending Score (1-100)」，圖表名稱為「Clusters of clients」，繪製出如以下的圖形。



一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

一張含有 文字, 螢幕擷取畫面, 鮮豔, 圖表 的圖片

自動產生的描述