大數據分析與資料探勘-第四次作業

繳交日期: 2023/5/10

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Homework4\_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5/10晚上11點59分前，超過繳交期限就無法上傳。

1. 一張含有 文字, 螢幕擷取畫面, 軟體, 數字 的圖片

   自動產生的描述首先，請匯入iris資料(鳶尾花資料集)。
2. 由於iris資料集以dictionary資料表示，請用DataFrame方式將iris的data讀入X變數，同時設定column的名稱為feature\_names，並秀出 X的前5筆資料。

一張含有 文字, 螢幕擷取畫面, 數字, 字型 的圖片

自動產生的描述

1. 一張含有 文字, 螢幕擷取畫面, 軟體, 電腦圖示 的圖片

   自動產生的描述請用DataFrame方式將iris的target讀入Y變數，同時設定column的名稱為「target」，並秀出 Y的前5筆資料。
2. 一張含有 文字, 螢幕擷取畫面, 數字, 字型 的圖片

   自動產生的描述請將X、Y合併在同一個DataFrame顯示，同樣秀出此DataFrame的前5筆資料。
3. 一張含有 文字, 螢幕擷取畫面, 數字, 軟體 的圖片

   自動產生的描述承上題，將此DataFrame的所有features全部保留，但在target部分請留下0, 1的部份，請秀出更新後DataFrame的最後5筆資料。
4. 一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

   自動產生的描述一張含有 文字, 字型, 螢幕擷取畫面, 白色 的圖片

   自動產生的描述承上題，為了使用感知器演算法，請新增一個target\_class欄位，將target 0的部份對應成1，target 1的部份對應成-1，並把原來的target那一欄刪除，請秀出更新後DataFrame的最後5筆資料。
5. 請使用感知器演算法針對這兩種鳶尾花進行分類，其中起始w0設定為0向量(亦即數值全部設定為0)，請秀出演算法iteration為幾次，最後所算出的w為多少。
6. 一張含有 文字, 螢幕擷取畫面, 陳列, 軟體 的圖片

   自動產生的描述一張含有 文字, 螢幕擷取畫面, 陳列, 字型 的圖片

   自動產生的描述請再次使用感知器演算法針對這兩種鳶尾花進行分類，但將起始w0數值全部設定為8，請再次秀出演算法iteration為幾次，最後所算出的w為多少。
7. 假設現有個新的鳶尾花，四種feature分別為[3.0, 2.1, 4.6, 3.3]，請利用第(8)題最後所計算出的w，判斷這朵鳶尾花是target 0 (Setosa)或target 1 (Versicolor)。
8. 一張含有 文字, 字型, 螢幕擷取畫面 的圖片

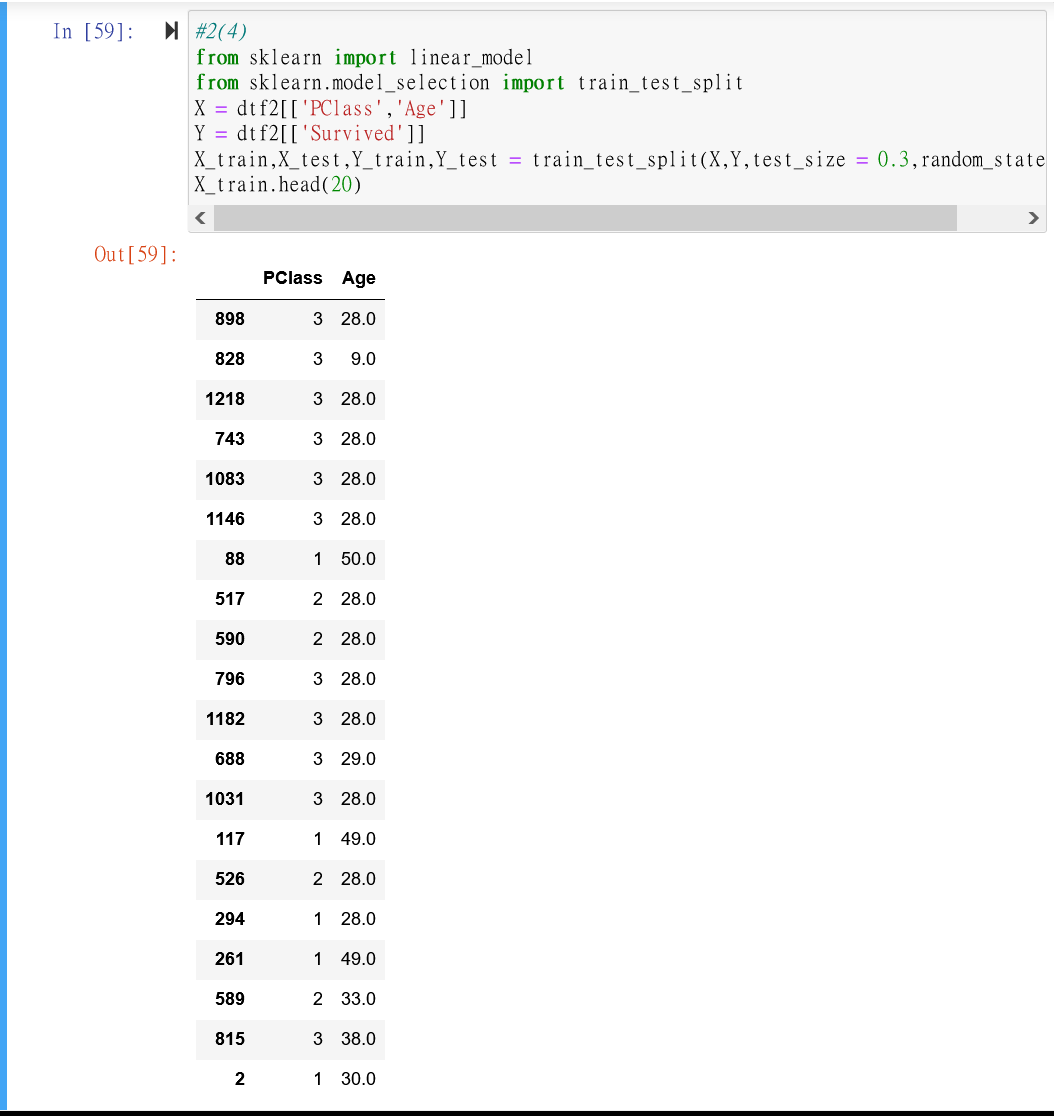
   自動產生的描述
9. 一張含有 文字, 螢幕擷取畫面, 數字, 軟體 的圖片

   自動產生的描述鐵達尼號(titanic)為資料學習中常被使用的資料庫，請匯入titanic資料檔至DataFrame，並顯示前20筆資料。其中，各欄位代表的意義，分別簡述如下：PassengerId (乘客編號)、Name (姓名)、PClass (乘坐艙等：1st, 2nd, 3rd)、Age(年齡)、Sex(性別)、Survived (是否存活，0 = No; 1 = Yes)、SexCode (性別編號，0 =male; 1 = female)。
10. 資料表中僅有Age欄位有空缺值，請將Age欄位的空缺值資料用「中位數」填補，並秀出填補後DataFrame之前20筆資料。

一張含有 文字, 螢幕擷取畫面, 數字, 軟體 的圖片

自動產生的描述

1. 一張含有 文字, 螢幕擷取畫面, 數字, 軟體 的圖片

   自動產生的描述由於PClass (乘坐艙等)顯示的並非數值資料，請將資料表中的PClass從文字直接替換成數值，如：1st =1、2nd =2、3rd =3，並加顯示新DataFrame之前20筆資料。
2. 乘上題，將新DataFrame之PClass與Age欄位設定為X變數，Y變數則設定為Survived。接著使用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 = 0，請秀出 X\_train的前20筆資料。
3. 請載入Logistic Regression model (參數皆為預設值)，使用上題(4)分割後的training資料訓練此模型，訓練完成後，請以test資料做測試，顯示test資料中各資料點分到不存活或存活兩個類別之機率。

一張含有 文字, 螢幕擷取畫面, 字型, 數字 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, 行 的圖片

   自動產生的描述承上題，同樣以test資料做測試，將模型預測的是否存活與實際是否存活進行比對，顯示出比對後的錯誤個數。
2. 一張含有 文字, 螢幕擷取畫面, 字型, 軟體 的圖片

   自動產生的描述承上題，請進一步顯示測試集資料分類結果的混淆矩陣。
3. 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，請計算測試集資料分類結果的正確率(Accuracy)，請顯示至小數點第四位。
4. 一張含有 文字, 電子產品, 螢幕擷取畫面, 字型 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。
5. 寶可夢(pokemon)為資料學習中常被使用的資料庫，首先請匯入pokemon資料檔至DataFrame，並加以顯示。各欄位代表的意義，分別簡述如下：Number (編號)、Name (名稱)、Type(屬性)、HP(血量)、Attack (攻擊力)、Defense (防禦力)、SpecialAtk (特殊攻擊)、SpecialDef (特殊防禦)、Speed (速度)。

一張含有 文字, 螢幕擷取畫面, 數字, 字型 的圖片

自動產生的描述

1. 寶可夢的Attack (攻擊力)、Defense(防禦力)兩欄位可能有遺漏值，請直接刪除這兩個欄位有遺漏值之寶可夢，並秀出刪除遺漏值後的資料表。

一張含有 文字, 螢幕擷取畫面, 數字, 字型 的圖片

自動產生的描述

1. 一張含有 文字, 螢幕擷取畫面, 數字, 字型 的圖片

   自動產生的描述承上題，Type(屬性)中包含許多屬性的寶可夢，請僅留下屬性為Fighting、Ghost與Normal的寶可夢，並秀出更新後的資料表。
2. 承上題，請將Attack (攻擊力)與Defense(防禦力)這兩個欄位設定為X變數，並進行變異數標準化(StandardScaler)後加以顯示。

一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述

1. 承上題，因Type (屬性)為類別欄位，請將屬性值進行編碼，Fighting對應成0、Ghost對應成1、Normal對應成2，編碼完成後設定為Y變數，並加以顯示Y變數。

一張含有 文字, 螢幕擷取畫面, 數字, 軟體 的圖片

自動產生的描述

1. 請載入SVM中的SVC，並將kernel設為RBF (其他參數為預設值)，使用上題(4)與(5)的X(標準化後的)、Y變數訓練此模型，訓練完成後請將模型的預測屬性值與實際屬性值結果作比對，顯示出比對後的錯誤個數。

一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述

1. 承上題，請進一步顯示出分類結果的混淆矩陣。

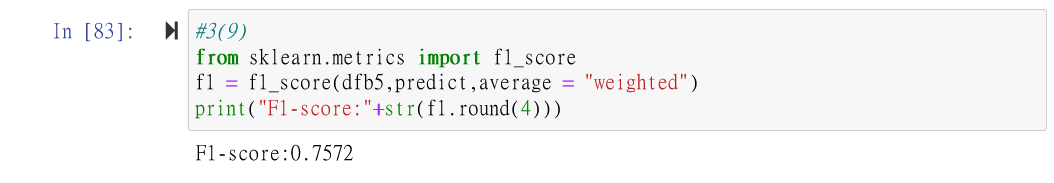
一張含有 文字, 字型, 行, 螢幕擷取畫面 的圖片

自動產生的描述

1. 承上題，請計算分類結果的正確率(Accuracy)，請顯示至小數點第四位。

一張含有 文字, 字型, 行, 白色 的圖片

自動產生的描述

1. 承上題，請計算分類結果的加權平均F1-score (請將average參數設定為weighted)，請顯示至小數點第四位。
2. 最後利用繪圖語法，將三種屬性值的寶可夢在二維平面上的座落位置及支撐向量機的切分線，在圖形中顯示出來，並設定橫軸座標名稱為Attack、縱軸座標為Defense。

一張含有 文字, 圖表, 螢幕擷取畫面 的圖片

自動產生的描述一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述