資料探勘 專案作業一

本作業主要目的為練習資料探勘實作，使用Python及其套件開發決策樹分類器。

分組：每組最多四人，請自行選出一人擔任組長。請將小組成員資料寄給助教蔡治紘。

安裝Anaconda裡的Spyder或Jupyter Notebook系統。有問題可請教助教。

Anaconda：https://www.anaconda.com/download/（請安裝Pyhton3.7）

Decision Tree (Python) <https://scikit-learn.org/stable/modules/tree.html>

UCI Datasets：<http://archive.ics.uci.edu/ml/datasets.php>

作業內容包括：

1. 從UCI Datasets中，自選一個資料集，資料筆數至少1000以上。分成訓練及測試資料，進行模型訓練，並分別衡量訓練資料及測試資料的分類正確率。
2. 從UCI Datasets中，下載Adult資料集，分成訓練及測試資料，進行模型訓練，並分別衡量訓練資料及測試資料的分類正確率。
3. 將測試資料結果輸出到Excel檔案，包含正確的類別以及分類器預測的類別。
4. 參考[Post pruning decision trees with cost complexity pruning](https://scikit-learn.org/stable/auto_examples/tree/plot_cost_complexity_pruning.html#sphx-glr-auto-examples-tree-plot-cost-complexity-pruning-py)章節，自行選擇三種適當的節點數及樹深度的參數，劃出訓練完的決策樹，並比較這三種參數設定之下，決策樹的分類預測正確率。

請遵循學術論文寫作格式撰寫專案報告，內容至少必須包括下列要項：

**摘要** (學術論文摘要通常不分段，內容簡要地涵蓋動機、目的、方法與實驗結果)

**一、緒論**

**1.1動機** (探勘所選用的資料集之動機)

**1.2目的** (想進行哪種探勘，期望了解甚麼？)

**二、方法**

(1. 簡單說明程式架構，2. 執行程式的方法，無法執行程式，視為不正確)

**三、實驗**

**3.1資料集** (簡介資料集：名稱、筆數、欄位、型態…、並以資料表格呈現部份資料內容)

**3.2前置處理** (如果有的話)

**3.3實驗設計** (實驗如何進行、參數如何設定等）

**3.4實驗結果** (呈現結果且附上一些畫面)

**四、結論**

**參考文獻** (參考論文、網站程式碼、資料，請註明引用文獻或來源)

* 評分標準

內容(難度、豐富性、方法正確性)：60%，論文格式：20%，寫作技巧(文筆通順)：20%

* **任何複製剪貼網站資料或學長姐以前作業，皆屬抄襲，嚴重者零分。引用網站或他人資料，應註明引用來源。**
* 此份報告最多不超過15頁，請加註頁碼。請遵循投稿論文寫作格式，使用一欄式，無需製作論文目錄。可以觀摩 ACM Transactions on Knowledge Discovery from Data 期刊論文的寫作技巧與格式。
* 紙本報告於上課前繳交或直接放入許老師的系辦信箱。專案程式直接上傳GitHub，專案名稱：hw1+組長學號，並新增助教為專案成員，以利下載評分。缺紙本或程式皆視為未完成繳交作業。
* 繳交紙本時，在左上角訂一訂書針即可，無須其他任何裝訂。
* 每遲交一天，扣該次作業成績10分。程式繳交時間以Github上的時間為準。
* 繳交日期：109/10/15（四）