人工智慧導論　程式作業 3 – Machine Learning Report

第 17 組
組員: 403210028 吳士瑋
403410035 李政勳
403530029 謝季霖
404410068 鄭諺輿
404410088 佘承澔

一、程式執行方式說明

原始程式碼包含 tmpLabelGenerate.py、labelEncode.py、rfTrain.py、rfTest.py 共四份 python 檔案，執行環境為 python 3.6，使用前先確認有安裝 sklearn、pandas 等 python 套件，接著將要 train 和 test 的資料分別命名為 trainData.csv 和 testData.csv，並放在同一個目錄下，依序執行: tmpLabelGenerate.py、labelEncode.py、rfTrain.py、rfTest.py 四份 python 檔案，執行完即可得知 train 完和 test 完的結果為何，並 output 一份預測出來的結果 output.csv。

二、方法說明

- 在 **tmpLabelGenerate.py** 中將 train 和 test 資料中所有的 label 合在一起，生成另一個檔案: tmpLabel.csv，讓 labelEncode.py 來使用，以避免在 testData.csv 中出現 new label 的問題。

- 在 **labelEncode.py** 中將 tmpLabel.csv 各欄位(共 12 columns)所出現的 label 去讓 LabelEncoder 做 fit 的動作(不同 unique 的 label 各給一個獨立的 number)，並一一將各欄位 fit 完的結果 dump 下來成 12 個檔案，方便之後做 transform 將 data 轉為數字編碼。

- 在 **rfTrain.py** 中將 trainData.csv 的資料讀進來，並且將剛剛的 12 個 lebel fit 完的檔案一一開啟，分別對 12 個欄位做 transform，將原本的字串 label 轉換為數字編碼的形式。接著將這編碼完的這 12 個欄位和最後是否有 click 的欄位，分別丟進 sklearn 的 RandomForestClassifier 分類器 (隨機森林，random forest 方法) 中去 fit 訓練，並將訓練好的分類器 dump 下來，方便 rfTest.py 去做 predict，最後試著 predict train data，看看 Training accuracy 等數值如何。

- 在 **rfTest.py** 中分別讀入剛剛 dump 下來訓練好的分類器和 testData.csv，之後一樣去做各欄位編碼，接著讓分類去去 predict 這個 test data，並將

結果輸出為 output.csv，若是這個 test data 已經有最後一個 click 這個欄位，則可以順便算出剛剛分類器 predict 的結果和實際答案之間的 accuracy 和 F-measure 等值，並顯示出來。

- **使用方法: random forest(隨機森林)說明**

  Random forest 是種用來做機器學習的方法，我們在使用這個方法時，設定參數讓它產生了 100 顆的 decision tree 去 ensemble，要談 random forest 是怎麼做的則必須先簡單說明 decision tree 是在做什麼。Decision tree(決策樹)，是一個樹狀結構，對於某個 case 從 root 開始，給它一連串的測試，依據其 feature 來決定下一個測試應該是往哪個節點走，最後決定出這個 case 的結果為何，在機器學習上我們就是想辦法給足夠量的資料，去找出一個合適的 tree，讓它能正確的分類各式各樣的 case，這就是訓練完所要的分類器。而 random forest(隨機森林)則是一個包含很多個 decision tree 的分類器，藉由產生多個 decision tree，讓其 ensemble 起來，多個 tree 的結果去做投票來決定產生 RF 的結果，可以避免只單獨訓練一個 decision tree 時可能造成過大的偏差。

三、組員互評