作業五

壹、請使用 python 完成以下題目，並在文字框附上適當註解，以 ipynb 檔繳交：

這是一份顧客的資料名單，請大家分析一下，並且判斷是否會繼續消費。

(Target: churn)

1 對 Churn 欄位使用 Stratified sampling 並從原本的資料集取出 60%資料。(10%)

2 列出各類別 (churn) 資料個數。(5%)

3 資料前處理 (填補空值等等)。(10%)

4 用 10 folds cross validation 建立 Logistic Regression 和 SVM 模型。(10%)

5 利用測試資料來測試兩個模型印出 Accuracy。(10%)

6 重複 1~5 題 30 次，並印出兩種模型最終的平均 Accuracy。(10%)

7 以 paired t-test 比較前面做 30 次 10 folds cross validation 的兩種模型，並說明結論。(10%)

貳、請使用 weka 完成以下題目，並截圖結果附上適當說明，在截圖上圈出能滿足題目要求的設定即使是預設值，作業以 PDF 文件呈現：

1 從原本的資料集中使用 Stratified sampling 取 60%的資料。(5%)

2 列出各類別 (churn) 資料個數。(5%)

3 資料前處理 (填補空值等等)。(5%)

4 重複 10 次 repeated 10 folds cross validation 並用 Paired t-test 比較 Logistic Regression 和 SVM。(10%)

5 說明結論並輔以 weka 的輸出。(10%)

備註：

1. 若資料量太大以致於 weka 無法運作， 在附上截圖證明後， 可降低抽樣數量。
2. weka t-test 功能在 Experimenter 中

作業繳交說明：繳交期限： 5/3 (三) 0:00

- Python 題請繳交 ipynb 檔 、 Weka 題請繳交 pdf 檔， 檔名 ECT_HW5_學號。
- 程式中請以註解或文字方塊標示題號。
- 需確保程式執行上傳至 eeclass 作業區那一版本的資料集不會出錯。
- 上傳至 eeclass 作業區，遲交一天扣該次作業得分 5 %，最多 50% 。