

## 作業二

壹、請使用 weka 完成以下題目，並截圖結果附上適當說明，以PDF文件呈現：

- 1.1 載入 train.arff，將 PassengerId、Name、Ticket 欄位刪除 (5%)
- 1.2 將 Cabin 的非空值以 1 替代，空值以 0 填入（先用MergeManyValues 取代非空值，再用ReplaceMissingWithUserConstant 取代空值） (5%)
- 1.3 使用 ReplaceMissingValues 將 Age 的空值以 Age 平均數填入 (5%)
- 1.4 將 Survived 與 Pclass 轉為 Nominal，並說明為何 Numeric 無法使用在 Decision tree (5%)
- 1.5 以 70% 切割訓練資料，使用 J48 對 Survived 進行分類，並截圖分類準確率、混淆矩陣及視覺化的Decision tree (10%)

貳、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：

- 2.1 以 DataFrame 格式載入 train.csv (5%)
- 2.2 請檢查並列出 train.csv 中每個欄位的空值個數 (5%)
- 2.3 將 Age 欄位空值以該性別平均值填入。(10%)
- 2.4 將 Cabin 欄位重製為 Pclass \* Fare (5%)
- 2.5 將 Survived 欄位重製為 0=Alive 1=Dead (5%)
- 2.6 將 Sex 與 Embarked 欄位轉為數字型態（例如：男性=0，女性=1） (5%)
- 2.7 請以 PassengerId、Survived、Name、Ticket、Pclass 以外的欄位作為訓練資料，建立 Decision tree 來預測 Survived，將訓練資料比例設為 50%，random\_state 設為 12，stratify = y，並繪出 Decision tree 的樹狀圖 (10%)
- 2.8 計算出在 2.7 測試資料上的平均準確率 (5%)

2.9 請用 2.7 的結果評估決策樹好壞(使用 classification\_report)產生類似以下結果(5%)

|              | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| Alive        | 0.78      | 0.89   | 0.83     | 55      |
| Dead         | 0.78      | 0.60   | 0.68     | 35      |
| accuracy     |           |        | 0.78     | 90      |
| macro avg    | 0.78      | 0.75   | 0.75     | 90      |
| weighted avg | 0.78      | 0.78   | 0.77     | 90      |

2.10 請分別以訓練資料比例 60%、70%、80%、90% 建立 Decision tree，random\_state 皆設為 12，並將不同資料比例與平均準確率的比較結果以 DataFrame 呈現，如右圖所示。(10%)

|   | split_proportion | score    |
|---|------------------|----------|
| 0 | 50/50            | 0.777778 |
| 1 | 60/40            | 0.815126 |
| 2 | 70/30            | 0.813433 |
| 3 | 80/20            | 0.804469 |
| 4 | 90/10            | 0.777778 |

2.11 呈上題，將此比較結果以折線圖呈現，如下圖所示：(5%)



繳交期限：3/16 中午 12 點

第一題請繳交.PDF檔，檔名為ECT\_HW2\_學號.pdf，請適當附文字說明。

第二題請繳交.ipynb檔與整理後的csv檔，檔名為ECT\_HW2\_學號.ipynb與

ECT\_HW2\_學號.csv，程式中請適當附上註解

遲交一天扣該次作業5%（最多扣50%）