Introduction to Machine Learning – Final Project

此次 Final Project 為團隊專題，旨在模擬一次 Data Mining / Machine Learning Workflow，自「資料收集」、「資料前處理」、「資料統計與視覺化」、「模型建立」至「模型應用」和「上台發表」。

I. Workflow

1. 資料導向：**各組須針對一個問題收集資料進行分析**

2. 問題需要有**實際用途**，或**趣味性**，範例：
   i. 預測天氣（O）
   ii. 判斷一漫畫出自誰手（O）
   iii. 從電廠發電量預測空汙資料以證實關聯（O）
   iv. 以點閱數預測來自哪個 youtuber（X）
   v. 以是否有喉結預測性別（X）
   vi. 占星術（X）

3. 資料前處理 —— **資料清理**：無論是自網路上收集來的資料，抑或是問卷收集得來的資料，將會有許多錯誤或缺失，因此必須要考量資料清理問題，常用的策略有：
   i. 移除有缺失的資料
   ii. 將缺失的部分以平均值或中位數補齊
   iii. 有條件的移除或補齊有缺失資料

4. **資料統計與視覺化**：為了了解資料分布，將會進行資料統計以及視覺化，其中可能會發現資料間的關聯性，可在資料前處理與建立模型的部分針對處理。

5. **資料前處理**：為了讓資料能夠符合模型所需的格式和條件，因此需要有資料前處理的步驟，常用的策略有：
   i. 型態轉換
   ii. Normalization
   iii. Label Encoding
   iv. One-hot Encoding

6. 模型建立：各組需要針對問題與資料設計**至少三種** Machine

Learning 模型，對問題進行訓練及預測。

i. 其**不限於上課有提及的模型**，亦可以試著使用 Neuron Network 等新的技術，其可以很好的針對影像進行處理。

ii. 對於每個使用到的模型，需要**提供最後的效果數值**，可以 Accuracy、Confusion Matrix、Recall、Precision、L1 平均誤差等方法表示，需注意的是此數值需要讓人有「感覺」(如單純的 R2 Score 數值就難以了解)。

iii. **需注意 Training set /Testing set 須明確分好，以防止 overfitting 的問題**

7. **模型應用**：對於所選問題，如何用團隊所建立的模型進行解決(預測或分析)，可以於報告時提供使用方法示範，將可能得到較好的成績，例如：
   i. 實際應用資料預測明天天氣
   ii. 實際預測下場比賽的勝負
   iii. 應用預測的數字(模擬)買股票

8. **上台發表**：預計於 **6/5(三)、6/12(三)** 兩天將要求各組進行 15 分鐘的 presentation + Q&A，內容需涵蓋以上內容。

9. **最後需要將 presentation 所使用的投影片(轉成 PDF)以及 source code 上傳至 E3**，一組以一人為代表上傳即可。

II. 資料要求

1. **資料不可直接自 Kaggle 中取得** —— 因 Kaggle 中將包含已知的解法
2. 自行搜集的資料可獲得較高的分數

III. 評分標準 —— 原始分數

1. **資料 —— 40%**
   i. 依原創性、獨特性、實際性評分
   ii. 直接使用 Kaggle 取得資料 -> 0%
   iii. 使用他人整理好的資料 ≤ 20%
   iv. 自行收集資料 ≥ 20%

2. **方法 —— 40%**

   i. 包含 workflow 中的內容，以及其他使用到的操作

3. **Presentation —— 20%**

IV. 最終成績

1. 你的調整分數 = (你的原始分數 – 全部人原始分數平均) x (15. / 全部人原始分數標準差) + 80.
2. 如果 presentation 中包含每個人的分工比例 (**需明確以一數字表示**)，調整分數公式為：

$$YourScore = (0.5 \times \text{ProjectScore}_{adjusted}) \times N_{people} \times ratio + (0.5 \times \text{ProjectScore}_{adjusted})$$

若超過 100 分以 100 分計。