Data Mining Homework 2

- 作業描述：本次作業共有兩個任務，請各位預測與分析以下兩個資料集。
- Task1：Multi-class classification
  - Dataset
    - # of Training data：1109
    - # of Testing data：276
    - # of Attribute：18
    - # of class：4
  - Rule
    - 可以使用上課提過的 Tree-based 分類演算法
    - 可以對 Data 做任何處理
    - 請試著在不知道 testing accuracy 的情況下，使用 training data 建立決策樹，並且找到你認為合適的決策樹對 testing data 進行分類
    - Baseline
      - accuracy：0.22
- Task2：Clustering
  - Dataset
    - # of data：600
    - # of Attribute：60
    - # of cluster：6
  - Rule
    - 可以使用所有上課提過的分群演算法
    - 除了原始資料外，有提供透過 LSA 降維後的資料集可以使用(5 維、10 維、20 維及 40 維)。若想試其他不同的前處理方法，可以對原始 Data 做任何處理
    - 對 Data 進行分群找到自己認為表現最好的結果
    - Data is balanced
    - Baseline
      - Silhouette Coefficient：0.4
      - Normalized Mutual Information score (Private)
- 繳交作業
  - 輸出結果請交 csv 檔，並各自命名如下檔名 classification.csv， clustering.csv，內容格式請依照 prediction.csv 為範例
  - 結果過 baseline 可得基本分，其餘須看報告內容
  - 2~4 頁的報告，開頭請先告知你是用那些 Tool 完成，並說明調整參數過程
  - 若是用 WEKA 完成的同學，請附上自己進行完資料前處理後的檔案
  - 若是用其他 Tool 包含自行實作的同學請附上 Code
  - 最後所有項目包裝成壓縮檔上傳 moodle，檔名為學號加上_DM_HW2
    - EX: M12345678_DM_HW2