Programming Assignment 3
Deadline: 5/14

此次作業主要目的在讓同學學習運用 Python 從二手車市場的歷史資料中，建立二手車的車況分類模型。請先將給定的資料集進行資料前處理，參考課堂範例，使用 One-Hot Encoding 將資料進行編碼後再使用 DecisionTreeClassifier 套件，利用 Decision Tree 分類演算法進行建模與分析，回答以下問題。

- 作業給定的 Vehicle Condition Data 已經附在 Moodle 平台上。
- 作業每人繳交一份報告，檔案類型以 **pdf** 為限。
  - 上傳檔名格式為**學號_P3**，例如：**112753XXX_P3.pdf**。
- 此次作業可以使用現有套件執行運算。

1. 給定 training_data.csv，請先針對資料進行資料前處理，參考課堂範例，經由 One-Hot Encoding 後，除了標籤欄位，(1)用於訓練 Decision Tree 模型的欄位共有幾個？(2)請依序列出各欄位名稱與該欄位說明（依照欄位名稱字母順序升冪排序)。(5%)

2. 承 1，本題任務使用 **Entropy** 作為 Impurity Metric，請在不使用 Early-Stop Rules 的情況下，**使用全部 300 筆資料**生成 **The Fully-Grown Decision Tree** (即不限制 Decision Tree 的 Max. Depth、Max. Number of Leaf Nodes、Min. Number of Instances 等)，請列出此 Decision Tree 的 (1) Max Depth 和(2) Leaf Nodes 總數。[注意：使用 DecisionTreeClassifier 時僅設定 criterion=**'entropy'**，其餘使用 DecisionTreeClassifier 的預設參數。] (10%)

3. 承 1，為了有效建立分類模型，以及評估模型分類的效果，我們採用 Holdout 策略，練習使用 sklearn.model_selection 的 train_test_split 將已有的 300 筆資料分成 70%為訓練集和 30%為測試集，再進行模型訓練，使用 train_test_split 時，僅指定 test_size=0.3、random_state=42，其餘使用 train_test_split 的預設參數，本題任務請使用 **Entropy** 作為 Impurity

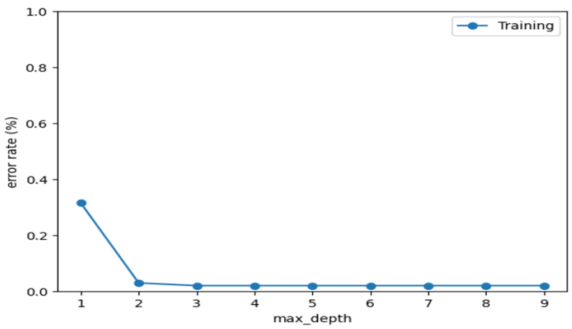
Metric，在不使用 Early-Stop Rules 的情況下，**使用訓練集 210 筆資料**生成 **The Fully-Grown Decision Tree**，請列出此 Decision Tree 的 (1) Max Depth、(2) Leaf Nodes 總數、(3) 所有的 Internal Nodes 的 Index、Attribute/Feature Name、Split Threshold (輸出請依照 Node Index 升冪排序，參考圖一)。[注意：使用 DecisionTreeClassifier 時僅指定 criterion=**'entropy'**，其餘使用 DecisionTreeClassifier 的預設參數。] (15%)

```
Internal Node Index: 0
  feature_name: petal width (cm)
  split threshold: 0.800000011920929
--------------------
Internal Node Index: 2
  feature_name: petal width (cm)
  split threshold: 1.75
--------------------
Internal Node Index: 3
  feature_name: petal length (cm)
  split threshold: 4.950000047683716
```

圖一

4. 承 3，利用已建立的 Decision Tree 模型，計算 (1) Training Error 為？(即訓練集 210 筆資料的錯誤率) (2) Test Error 為？(即測試集 90 筆資料的錯誤率) (10%)

5. 承 4，本題任務練習使用 Early-Stop Rules，使用 DecisionTreeClassifier 時僅指定 criterion=**'entropy'** 和 max_depth 數值為 1~10 的整數，其餘使用 DecisionTreeClassifier 的預設參數，並觀察 Training Error 的變化。(1)請提供 max_depth 和 Training Error 的關係曲線圖，如圖二，(2) 說明 max_depth 和 Training Error 的關係變化，並解釋為何有此現象。(15%)

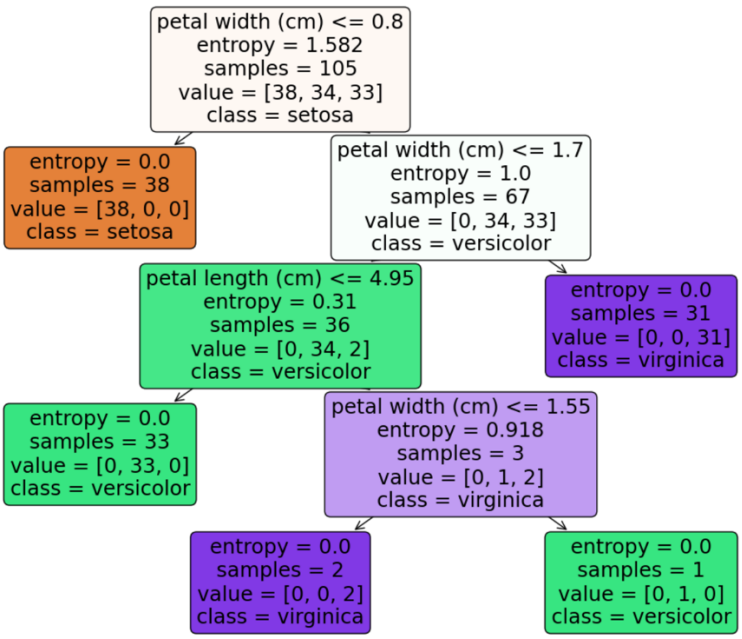


圖二

6. 承 5，請利用 Nested Cross-Validation，觀察不同 max_depth 值和 $\text{Error}_{\text{val}}$ 的變化，當 max_depth 的數值為 1~10 的整數時，請試著從中挑選 max_depth 值應該設為多少？(5%)

7. 承 6，請針對所選擇的 max_depth，使用訓練集 210 筆資料，生成 Decision Tree，在使用 DecisionTreeClassifier 時僅指定 criterion='**entropy**' 和 max_depth 數值，請列出此 Decision Tree 的 (1) Leaf Nodes 總數，(2) Training Error(即訓練集 210 筆資料的錯誤率)，(3)Test Error(即測試集 90 筆資料的錯誤率)。(15%)

8. 利用給定的 training_data.csv，自行訓練 Decision Tree 的二元分類模型，並提供此 Decision Tree 的 (1) Max Depth、(2) Leaf Nodes 總數、(3) Decision Tree 視覺化圖，如圖三。請利用此 Decision Tree 模型，針對 P3_test.csv 的測試資料，依序預測每一筆的標籤(bad/acc)，**並產生 submission.csv**，且將此檔案**提交至 Moodle**，輸出格式請參考 submission_template.csv。(25%)



圖三