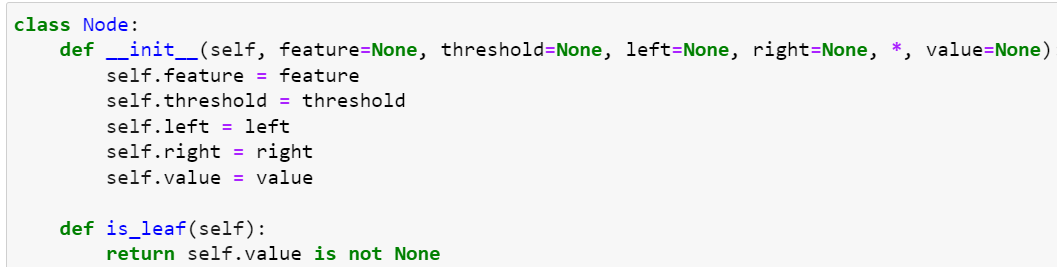
ML Term Project Report

109550178 黃昱翰

Decision Tree Algorithm:

根據information gain找到當前最好的feature去split samples，分割完後再recursively的往左右的child node去找最好切割的feature，並且設置停止條件，當深度達到max\_depth或當前node只有一種類別或當前node沒有instances。

Implementation:



is\_leaf回傳true or false根據node的value是否為None 一張含有 文字 的圖片

自動產生的描述 一張含有 文字 的圖片

自動產生的描述

呼叫\_best\_split找最好的feature, threshold，再判斷是否觸碰到停止條件，若

是則把當前node中最多instances的類別當作這個node的class，若非則切割instance並繼續往左右recursively \_build\_tree。一張含有 文字 的圖片

自動產生的描述

Loop所有features和所有features value當作threshold去計算information gain，紀錄最佳的infromation gain所用的feature和threshold

一張含有 文字 的圖片

自動產生的描述

若深度達到最大深度、當前node只剩一個類別、當前node沒有instance，或當前node instances和parent node instances一樣多就達到停止條件。

一張含有 文字 的圖片

自動產生的描述

\_create\_split根據threshold切割instances

\_split\_criterion根據self.split\_criterion判斷計算gini或entropy。

\_information\_gain計算parent node和child nodes的entropy 或gini index來計算infromation gain

一張含有 文字 的圖片

自動產生的描述

fit: 開始build tree並assign root node。

\_predict\_one, predict: 根據預測傳進來的instance feature去走一遍path得到prediction class。

getdeepest: 回傳最深走到第幾層。

* Dataset 1

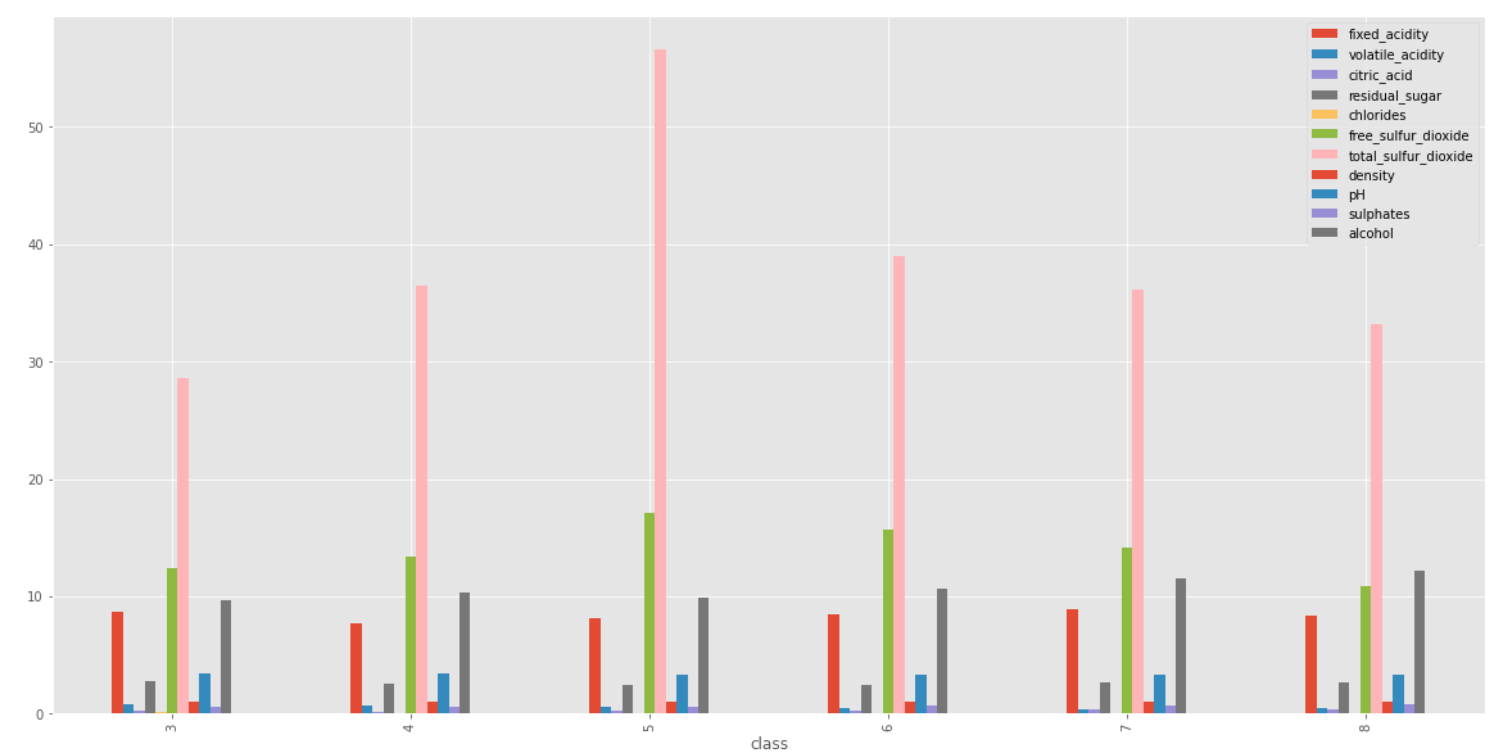
Data Preprocessing:

1. Normalize Feature & Remove Outlier

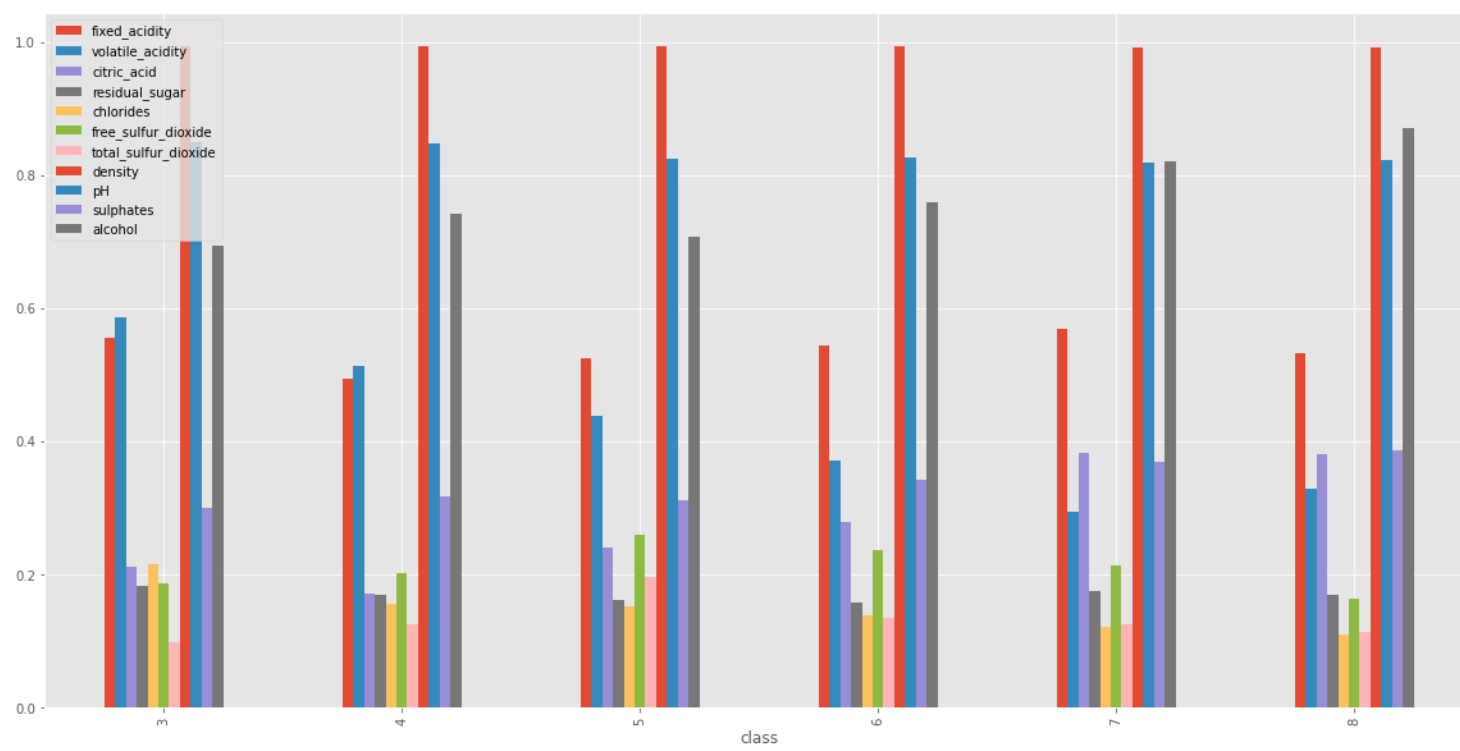
使用IQR，把落在upper bound和lower bound外的instance remove，但因為資料本來就不多而且很不平衡，所以有額外設定threshold去決定是否要remove

1. Impute Missing Value

使用KNN imputer並根據距離去找最近的5個neighbors去補值



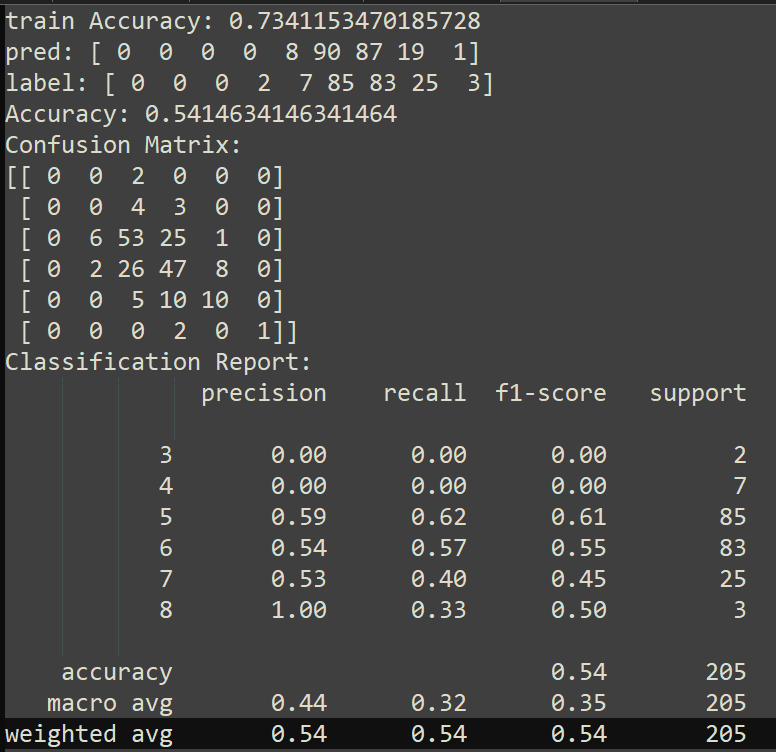
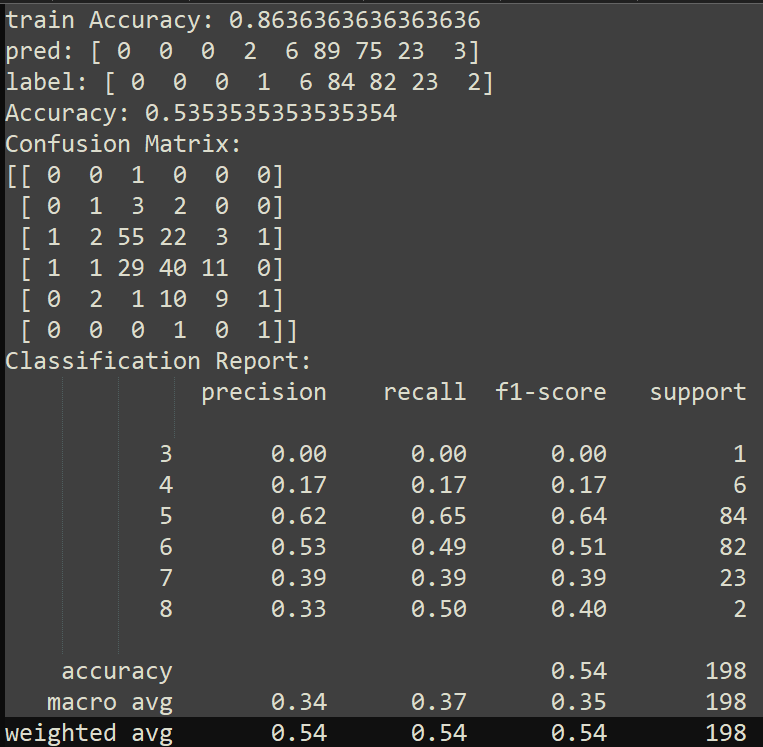
Unnormalized data



Normalized data

Experiment Result:

1. Without Data Preprocessing 2. Add Data Preprocessing

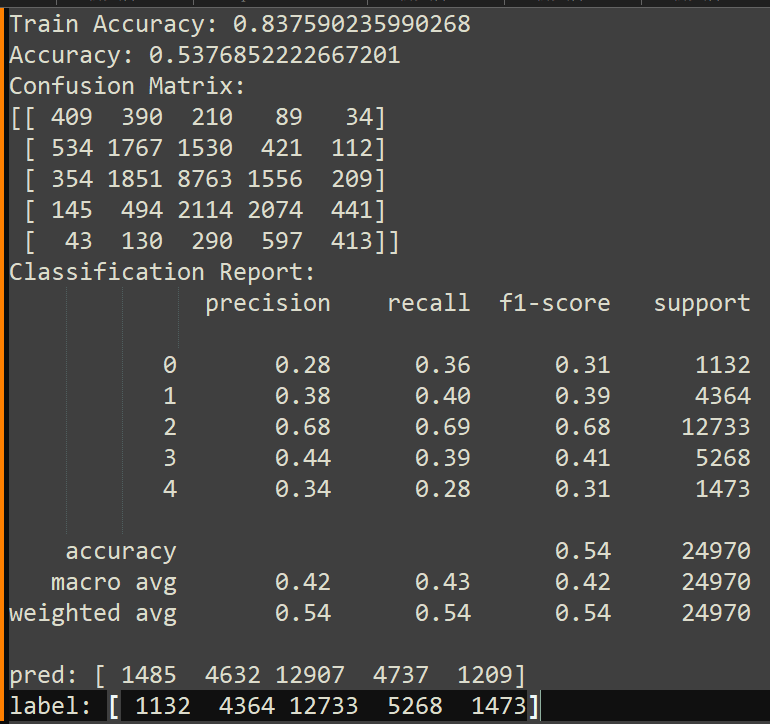
實驗發現即使做了data preprocessing，prediction也沒有取得明顯的進步，並且也和沒做data preprocessing時一樣在0.48~0.64的accuracy，推測可能是Data Preprocessing做的不夠完整，可能有irrelevent feature之類的影響。也有使用迴圈去看不同hyper parameter的performance，但基本上也只是在差不多的區間有小小的浮動。

* Dataset 2

Data Preprocessing:

基於demo的code，把phrase切成單字並移除stop words(大量出現但不助於情感分析的單詞)，再轉換成one hot encoded的list並padding成最長的phrase。有移除missing value的資料。

Experiment Result:



這個dataset資料雖然比較充足，但也是不平衡，Accuracy比上一個dataset浮動的要來的小很多，基本上都在0.52~0.54的區間。

Conclusion:

雖然dataset1有做了滿多資料前處理，但Accuracy還是這麼低，不太確定問題是出在哪裡，或是可能這就是DT的極限(?)。至於dataset2我有點不太知道該怎麼進行處理所以就只有處理missing value，原本有想要試試random forest，但一直遇到各種error，就放棄了。