資料探勘專案作業一

決策樹預測類別數值

指導教授：

許中川 教授

成員：

M11021012 林承威

M11021028 劉軒瑋

M11021052 邱守燦

M11021059 李鴻庭

日期：

2021 年 10 月 28 日

摘要

過往都以傳統的統計模型，進行分類預測，但隨著時代的演進，機器學習技術已逐漸取代。透過不同的演算法、參數設定，從大量數據中挖掘分析，得到有用的資訊，這種技術稱為資料探勘。本次研究分別使用 Adult Dataset 及 Wine-quality Dataset，探討影響薪資高於 50,000 之屬性、影響白酒品質等級之屬性，透過資料前處理抓取特徵及屬性，將資料集分割為 80%訓練資料、20%測試資料，使用決策樹(Decision Tree)分類預測方法，最後透過訓練資料與測試資料的比對，並根據混淆矩陣之結果得知，Adult Dataset 使用 Capital-gain、Marital-status、Education-num 屬性、Wine-quality Dataset 使用 Alcohol、volatile acidity 屬性進行分類時，準確度分別為 83%及 51%。

關鍵字：決策樹(Decision Tree)、隨機森林(Random Forest)。

一、 緒論

1.1 動機

1.1.1 Adult Dataset

經濟不景氣，在民生物價持續喊漲，房價只升不跌的大環境下，如何增加收入，已是每個人必須研究的課題，據國勢處普查局統計，2020 全年工業及服務業受雇員工平均月薪資為 54,320 元。倘若有一決策樹，以年收入是否大於特定薪資做類別，輸入相關資訊後能正確分類，並準確率夠高的話，反過來說，薪資大於某個數值的人，都擁有決策樹上相對應的特質，或許就能深入探討為何這群人能達成如此高的收入，以供研究者研究。

1.1.2 Wine-quality Dataset

高檔白酒，與能開懷暢飲的普通啤酒、便宜酒類等不同，一般人僅在特定節日搭配高級料理飲用。白酒能透過口感輕盈或厚實、甜度清爽或酸度濃度來區分等級，但這終究屬於主觀感受，並不是每個人都能品嘗處個別差異。倘若能透過客觀數據來分類白酒等級，對於消費者來說，說服力較高；對於釀造的酒廠而言，也能監測相關數據，來控制產出白酒的等級，也能制定等級與相對應的價格，供消費者選擇，提升產品競爭力。

1.2 研究目的

1.2.1 Adult Dataset

為了瞭解收入大於特定值的人，大部分都擁有哪些特質，本組選取 Adult 資料集，並使用決策樹與隨機森林進行分類，投入工作類別、婚姻狀況、教育程度、職業、所屬國籍等 15 種屬性資料，但礙於資料集提供的資料限制，本次研究所分類出的類別設定為大於 50,000 元，經決策樹與隨機森林進行分類過後，將得知哪些屬性將會被分類到大於薪資 50,000 元的類別，以此來反推大於 50,000 元者擁有哪些屬性。

1.2.2 Wine-quality Dataset

隨著檢測儀器的精密度提升、量測技術的發展，釀造酒廠可以隨時監控酒精在體內包含的各種物質，根據提供的 Wine-quality 資料集，裡面包含固定酸度(Fixed acidity)、檸檬酸(Critric acid)、剩餘糖分(Residual sugar)、酒精(Alcohol)、氯化物(Chlorides)等 12 個屬性資料，為了能有效且客觀的定義酒品等級，採用決策樹與隨機森林替酒品等級做分類，切割 80%的訓練資料集與 20%的測試資料集，對品質(Quality)類別做分類預測，解決過往主觀感受評級的問題。

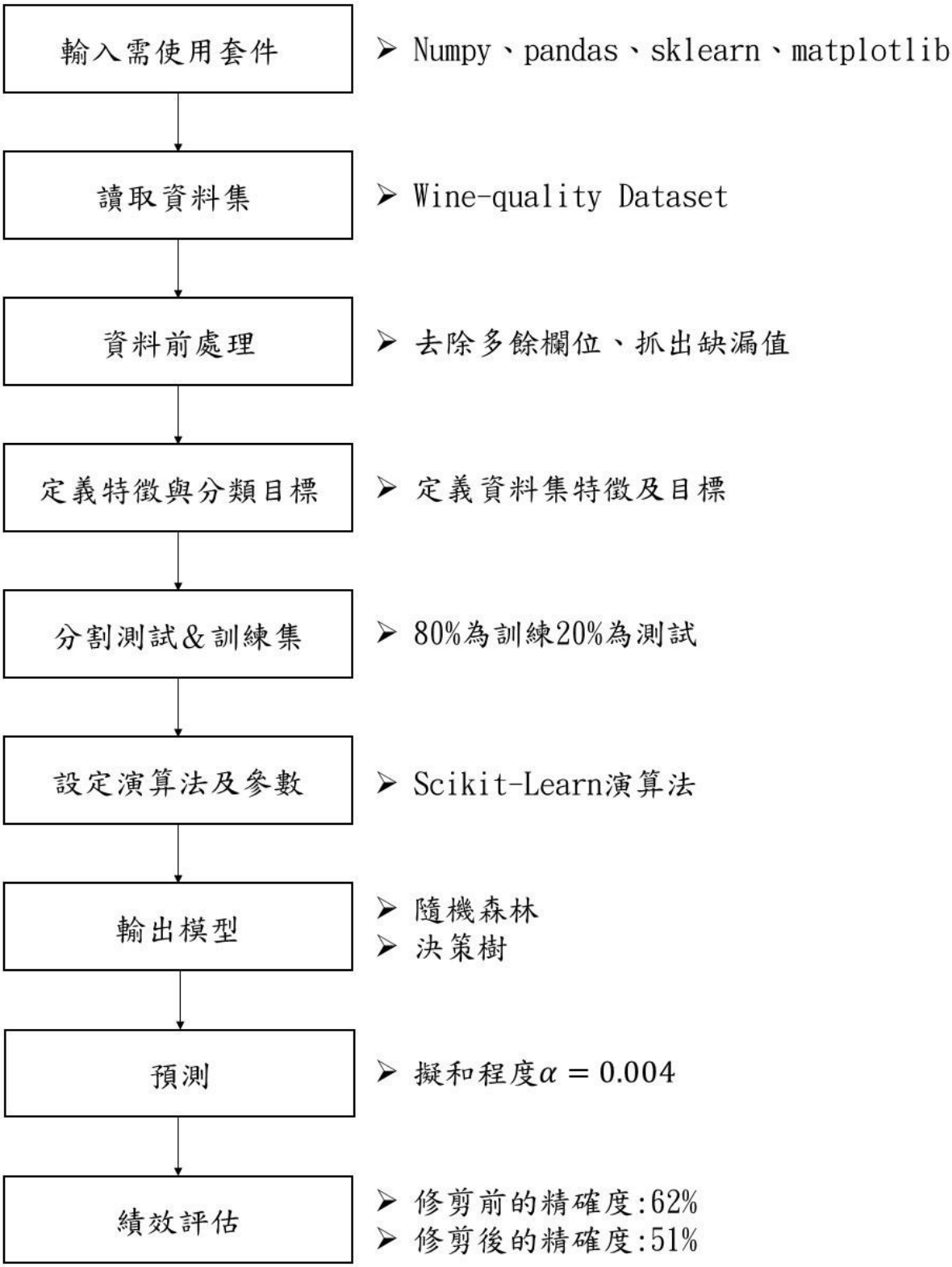
二、 方法

2.1 程式架構

2.1.1 Adult Dataset

圖一 Adult Dataset 之程式架構流程圖及說明

2.1.2 Wine-quality Dataset



圖二 Wine-quality Dataset 之程式架構流程圖及說明

2.2 程式執行方法

2.2.1 決策樹(Decision Tree)

決策樹(Decision Tree)是一種特殊的樹狀結構，透過像樹枝一樣的圖形或決策模型的支持工具，它是一個監督式學習的演算法，決策樹經常在運籌學中使用，特別是在決策分析中，它幫助決策者確定一個最可能達到目標的策略並用來輔助決策。

2.2.2 隨機森林(Random Forest)

隨機森林(Random Forest)是一個包含多個決策樹的分類器，並且其輸出的類別是由個別樹輸出的類別的眾數而定，隨機森林的重點在於抽樣，其樣本抽樣設計是採取後放回，並且在遇到過度配置時，隨機森林能夠解決，因為在大數法則下決策樹的結果會趨向一致，能夠大幅降低錯誤率。

2.2.3 KDD

KDD 表示將低層數據轉換為高層知識的完整過程，KDD 在獲取資料後，將資料處理並清洗，再將資料轉換成數學模式，進而建模，並在最後從中發掘有意義且具有價值的資訊，並評估其成效。

三、 實驗

3.1 資料集

3.1.1 Adult Dataset 說明

此資料集建立於 1996 年 5 月 1 日共有 48,842 筆，15 個欄位。

表一 Adult Dataset 欄位資料說明彙總表

欄位名稱	欄位說明
Age	年齡
Workclass	工作類別
Fnlwgt	連續數值
Education	教育程度
Education-num	教育人數(連續數值)
Marital-status	婚姻狀況
Occupation	職業
Relationship	關係
Race	種族
Sex	性別
Capital-gain	資本收益
Capital-loss	資本損失
Hours-per-week	小時/周
Native-country	所屬國家
Salary	年收入

3.1.2 Wine-quality Dataset 說明

此資料集建立於 2018 年 9 月 7 日共有 4899 筆，12 個欄位。

表二 Wine-quality Dataset 欄位資料說明彙總表

欄位名稱	欄位說明
Fixed acidity	固定酸度
Volatile acidity	揮發性酸度
Critric acid	檸檬酸
Residual sugar	剩餘糖分
Chlorides	氯化物
Free sulfur dioxide	游離二氧化硫
Total sulfur dioxide	二氧化硫總量
Density	密度
pH	酸鹼值
Sulphates	硫酸鹽含量
Alcohol	酒精
Quality	品質

3.1.3 實驗數據

(1) Adult Dataset 資料集

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	salary
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
32556	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wife	White	Female	0	0	38	United-States	<=50K
32557	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	United-States	>50K
32558	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	0	0	40	United-States	<=50K
32559	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	0	0	20	United-States	<=50K
32560	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wife	White	Female	15024	0	40	United-States	>50K

圖一 Adult Dataset 資料集

(2) Wine-quality Dataset 資料集

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.00100	3.00	0.45	8.8	6
1	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.99400	3.30	0.49	9.5	6
2	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.99510	3.26	0.44	10.1	6
3	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.99560	3.19	0.40	9.9	6
4	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.99560	3.19	0.40	9.9	6
...	...	...	...	...	...	...	...	...	...	...	...	...
4893	6.2	0.21	0.29	1.6	0.039	24.0	92.0	0.99114	3.27	0.50	11.2	6
4894	6.6	0.32	0.36	8.0	0.047	57.0	168.0	0.99490	3.15	0.46	9.6	5
4895	6.5	0.24	0.19	1.2	0.041	30.0	111.0	0.99254	2.99	0.46	9.4	6
4896	5.5	0.29	0.30	1.1	0.022	20.0	110.0	0.98869	3.34	0.38	12.8	7
4897	6.0	0.21	0.38	0.8	0.020	22.0	98.0	0.98941	3.26	0.32	11.8	6

圖二 Wine quality Dataset 資料集

(3)Adult Train 資料集

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	race	sex	capital-gain	capital-loss	hours-per-week	native-country	salary
0	39	5	77516	9	13	4	0	4	1	2174	0	40	38	0
1	50	4	83311	9	13	2	3	4	1	0	0	13	38	0
2	38	2	215646	11	9	0	5	4	1	0	0	40	38	0
3	53	2	234721	1	7	2	5	2	1	0	0	40	38	0
4	28	2	338409	9	13	2	9	2	0	0	0	40	4	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
32556	27	2	257302	7	12	2	12	4	0	0	0	38	38	0
32557	40	2	154374	11	9	2	6	4	1	0	0	40	38	1
32558	58	2	151910	11	9	6	0	4	0	0	0	40	38	0
32559	22	2	201490	11	9	4	0	4	1	0	0	20	38	0
32560	52	3	287927	11	9	2	3	4	0	15024	0	40	38	1

圖三 Adult Train 資料集

3.2 前置處理

3.2.1 Adult Dataset

將 Adult 資料集，由原來的 txt 檔案格式轉換為 csv 檔案格式，發現在些許的資料欄位中具有不必要的符號資料，進行資料的清洗，利用 lambda 語法將具有'?'資料篩選與過濾，完成資料清洗，接著將 Adult Dataset 裡面文字屬性使用 Labelencoder 進行轉碼，最後讀取印出完成資料的前置處理。

3.2.2 Wine-quality Dataset

將 Wine-quality 資料集，利用 Pandas 套件的 DataFrame 物件讀取 csv 檔，使用 isna 函數檢查資料欄位中是否有缺失值，將資料過濾，完成資料的前置處理。

3.3 實驗設計

3.3.1 Adult Dataset

(1)套件載入:由 Scikit-Learn 演算法載入決策樹的分類(Classifier)套件、決策樹建模的分離(train_test_split)套件、決策樹度量的分類報表(classification_report)與混淆矩陣(confusion_matrix)套件，最後載入 python 的繪圖語言(pydot)與建立記憶體內的 str 輸入輸出空間(StringIO)。

(2)隨機森林繪製:安裝並啟動 Graphviz 開源工具，由 StringIO 的輸入 pydot 指令繪製 pydot 語言敘述的圖形，最後透過特徵值(feature_names)、分類值(class_names)的定義，繪製出隨機森林。

(3)修剪:載入 matplotlib 套件，界定各節點的不純度的路徑，由圖表呈現訓練資料集的實際擬合程度(α)，並在訓練資料集完成葉節點與深度的擬合程度評估，再由圖表呈現訓練資料集與測試資料集各自精確度的擬合程度，觀察是否有過度訓練(overfitting)的情形，並在($\alpha = 0.028$)做決策樹修剪。

(4)決策樹繪製:由擬合程度($\alpha = 0.028$)進行修剪，並繪製出決策樹。

3.3.2 Wine-quality Dataset

(1)將資料集做分割(80%為 train，20%為 test)，載入 Scikit-Learn 演算法決策樹的分類(Classifier)套件、決策樹建模的分離(train_test_split)套件、決策樹度量的分類報表(classification_report)與混淆矩陣(confusion_matrix)套件，最後載入 python 的繪圖語言(pydot)與建立記憶體內的 str 輸入輸出空間(StringIO)

(2)隨機森林繪製:安裝並啟動 Graphviz 開源工具，由 StringIO 的輸入 pydot 指令繪製 pydot 語言敘述的圖形，最後透過特徵值(feature_names)、分類值(class_names)的定義，繪製出隨機森林。

(3)修剪:載入 matplotlib 套件，界定各節點的不純度的路徑，由圖表呈現訓練資料集的實際擬合程度(α)，並在訓練資料集完成葉節點與深度的擬合程度評估，再由圖表呈現訓練資料集與測試資料集各自精確度的擬合程度，觀察是否有過度訓練(overfitting)的情形，並在($\alpha = 0.004$)做決策樹修剪。

(4)決策樹繪製:由擬合程度($\alpha = 0.004$)進行修剪，並繪製出決策樹。

3.4 實驗結果

3.4.1 Adult Decision Tree 績效評估

	Precision	recall	f1-score	support
0[<=50K]	0.88	0.87	0.88	4495
1[>50K]	0.63	0.65	0.64	1538
Accuracy			0.82	6033
Macro avg	0.76	0.76	0.76	6033
Weighted avg	0.82	0.82	0.82	6033

表一 修剪前 Adult Decision Tree 績效評估(Accuracy=0.82)

	Precision	recall	f1-score	support
0[<=50K]	0.87	0.92	0.89	4584
1[>50K]	0.70	0.55	0.61	1449
Accuracy			0.83	6033
Macro avg	0.78	0.74	0.75	6033
Weighted avg	0.83	0.83	0.83	6033

表二 修剪後 Adult Decision Tree 績效評估(Accuracy=0.83)

3.4.2 Wine-quality Decision Tree 績效評估

	Precision	recall	f1-score	support
3	0	0	0	3
4	0.44	0.37	0.40	38
5	0.61	0.61	0.61	283
6	0.69	0.67	0.68	462
7	0.58	0.64	0.61	153
8	0.38	0.38	0.38	39
9	0	0	0	2
Accuracy			0.62	980
Macro avg	0.38	0.38	0.38	980
Weighted avg	0.62	0.62	0.62	980

表三 修剪前 Wine-quality Decision Tree 績效評估(Accuracy=0.62)

	Precision	recall	f1-score	support
3	0	0	0	1
4	0	0	0	30
5	0.52	0.66	0.59	291
6	0.50	0.62	0.56	442
7	0.51	0.21	0.30	173
8	0	0	0	40
9	0	0	0	3
Accuracy			0.51	980
Macro avg	0.22	0.21	0.21	980
Weighted avg	0.47	0.51	0.48	980

表四 修剪後 Wine-quality Decision Tree 績效評估(Accuracy=0.51)

四、 結論

4.1 Adult Dataset

本研究將 Adult Dataset 使用分為 80%的訓練資料與 20%的測試資料，並使用決策數與隨機森林對 Salary 類別做分類。

4.2.1 節點選擇

節點(1) Capital_gain

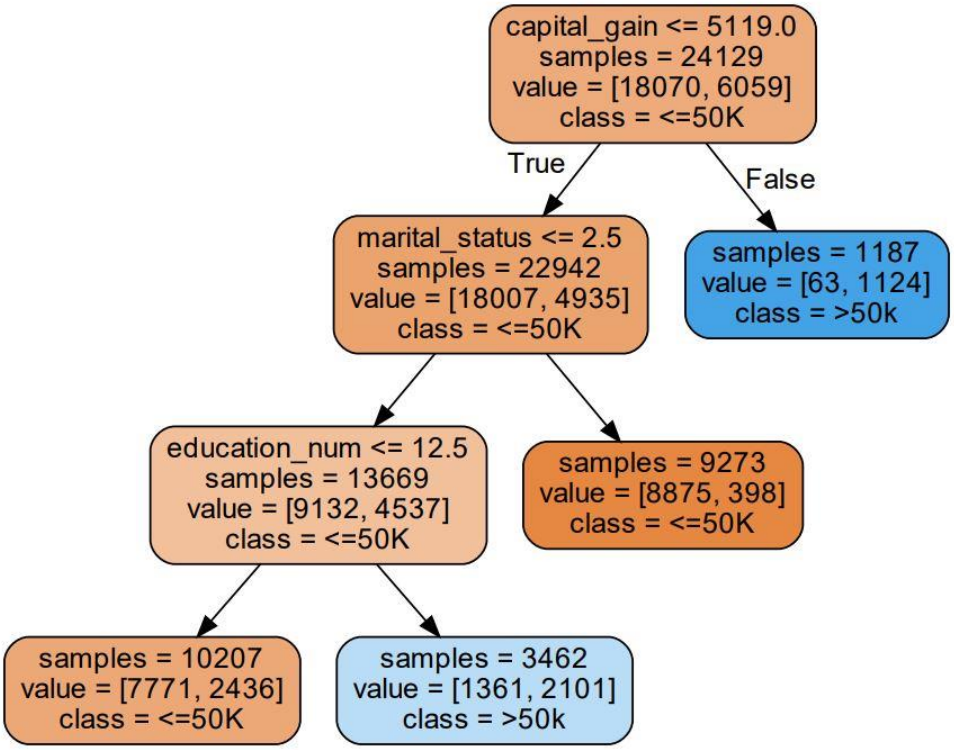
節點(2) Marital-status

節點(3) Education-num

4.2.2 整體績效

(1) 修剪前 Accuracy = 0.82

(2) 修剪後 Accuracy = 0.83



圖四 Adult Dataset 決策樹

4.2 Wine-quality Dataset

本研究將 Wine-quality Dataset 使用分為 80%的訓練資料與 20%的測試資料，並使用決策數與隨機森林對 Quality 類別做分類。

4.2.1 節點選擇

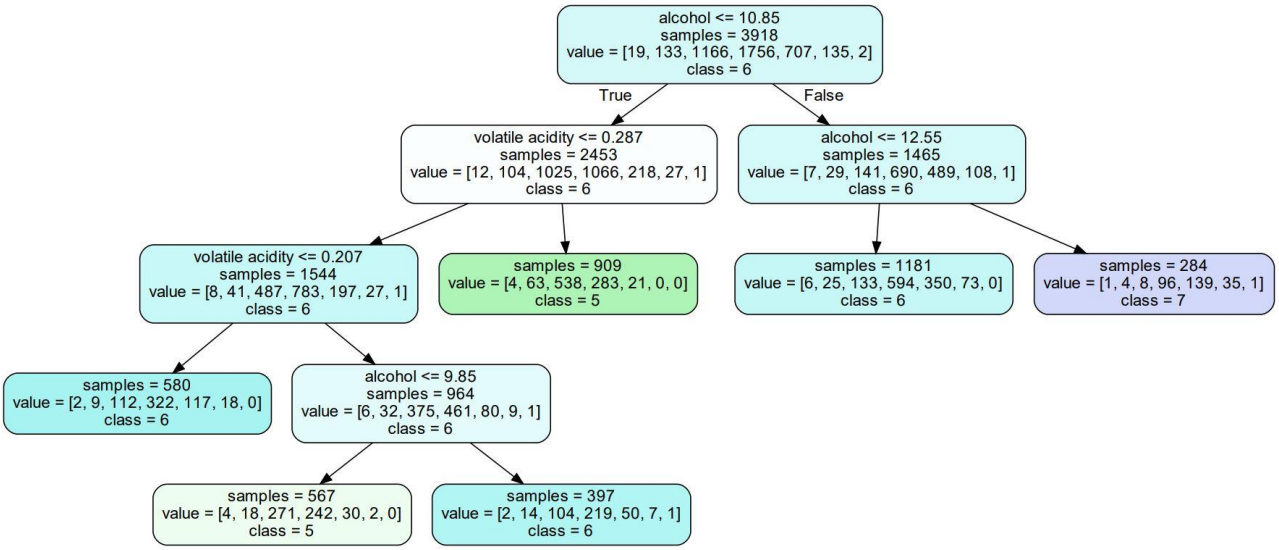
節點(1) alcohol

節點(2) volatile acidity

4.2.2 整體績效

(1) 修剪前 Accuracy = 0.62

(2) 修剪後 Accuracy = 0.51



圖五 Wine-quality Dataset 決策樹

五、 參考文獻

[1]機器學習首部曲隨機森林模型簡介 Rondom Forest
https://pyecontech.com/2019/09/17/random_forest/

[2]109 年全年工業及服務業受僱員工人數為 795 萬 5 千人，全年每人每月總薪資平均為 54,320 元
https://www.stat.gov.tw/ct.asp?xItem=46898&ctNode=527&mp=4

[3]學習機器學習必知的程序-資料庫知識探索
https://medium.com/marketingdatascience/%E5%AD%B8%E7%BF%92%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92%E5%BF%85%E7%9F%A5%E7%9A%84%E7%A8%8B%E5%BA%8F-%E8%B3%87%E6%96%99%E5%BA%AB%E7%9F%A5%E8%AD%98%E6%8E%A2%E7%B4%A2-72bd2d73781c

[4] Scikit-learn: Machine Learning in Python
https://jmlr.org/papers/v12/pedregosa11a.html
[5]用 Python 自學資料科學與機器學習入門實戰：Scikit Learn 基礎入門
https://blog.techbridge.cc/2017/11/24/python-data-science-and-machine-learning-scikit-learn-basic-tutorial/
[6] Pandas 教學資料視覺化必懂的 Pandas 套件繪製 Matplotlib 分析圖表實戰
https://www.learncodewithmike.com/2021/03/pandas-and-matplotlib.html
[7]混淆矩陣(confusion matrix)介紹
https://medium.com/nlp-tsupei/%E6%B7%B7%E6%B7%86%E7%9F%A9%E9%99%A3-confusion-matrix-%E4%BB%8B%E7%B4%B9-5eecd6da02ba
[8] StringIO Module in Python
https://www.geeksforgeeks.org/stringio-module-in-python/
[9] Simple python interface for Graphviz-GitHub
https://github.com/xflr6/graphviz
[10] python graphviz pydot 安裝配置決策樹視覺化
https://www.796t.com/article.php?id=223099
[11]機器學習 ML NOTE]Overfitting 過度學習
https://medium.com/%E9%9B%9E%E9%9B%9E%E8%88%87%E5%85%94%E5%85%94%E7%9A%84%E5%B7%A5%E7%A8%8B%E4%B8%96%E7%95%8C/%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92-ml-note-overfitting-%E9%81%8E%E5%BA%A6%E5%AD%B8%E7%BF%92-6196902481bb
[12] Predicting Wine Quality with Several Classification Techniques | by Terence Shin | Towards Data Science
https://towardsdatascience.com/predicting-wine-quality-with-several-classification-techniques-179038ea6434
[13] Estimating Wine Quality with Decision Tree
https://rstudio-pubs-static.s3.amazonaws.com/227997_869ca5f2dc144f7b85cdbc3f45a47bb6.html
[14] Decision Tree Analysis of Wine Quality Data | Kaggle
https://www.kaggle.com/rajyellow46/decision-tree-analysis-of-wine-quality-data
[15] Using Machine Learning to Classify the Quality of Wine – Leo Qin
https://www.leozqin.me/using-machine-learning-to-classify-the-quality-of-wine/