國立雲林科技大學

資訊管理研究所

資料探勘專案作業 1

指導教授:許中川教授

學生：M11223038 陳品佑

M11223032 張祥恩

M11223033 鍾季衡

M11223036 魏冠宇

Data Mining Homework: The Adult Dataset Mining

摘要

現今社會貧富差距日趨擴大，以及物價通膨嚴重，公民在生活上產生困難。人口普查收入能夠提供政府國家或地區的經濟結構，透過分析收入數據可以更全面地了解不同社會群體的經濟狀況，以便制定更精準的經濟政策照顧公民，提升社會的公平性。本研究以 Adult 資料集進行決策樹演算法 ID3、C4.5、C5.0 和 CART 的分類預測之績效，發現 CART 準確率最高。

關鍵字：人口普查收入、Adult 資料集、ID3、C4.5、C5.0、CART

1. 緒論

人口普查收入對於國家與地區提供了相當重要的資訊，藉由分析收入數據，能夠了解各個年齡層、教育程度、性別等等不同社會群體經濟狀況，以及預測出個人的年收入是否超過 50000 美元。

1.1 動機

人口普查收入的數據能夠讓國家或地區執行更具解決經濟狀況的政策，透過年齡、教育程度和年限、國家、種族、家庭角色、性別、投資收入、所在產業、婚姻狀況、職業等特徵，進行分類預測。

1.2 目的

透過 ID3、C4.5、C5.0、CART 等四種決策樹演算法進行分類預測，藉由訓練出來的模型，希望可以準確地預測出個人的年收入是否超過 50000 美元，以及觀察到各個變數影響到收入的重要性。

2. 方法

匯入資料後進行前處理，標準化與資料清理，其資料屬性分為年齡、教育程度和年限、國家、種族、家庭角色、性別、投資收入、所在產業、婚姻狀況、職業。本研究使用 ID3、C4.5、C5.0、CART 四種決策樹演算法進行分類預測，比較各個的準確率，也可以得知哪些變數對於收入的重要度較高。

3. 實驗

3.1 資料集

- 資料集名稱：Adult
- 資料筆數：32560
- 屬性數量：14

表 1 Adult 資料集屬性簡介

屬性名稱	型態	尺度
age	Int	Ordinal
workclass	String	Nominal
fnlgwt	Int	Nominal
education	String	Ordinal
educationnum	Int	Ordinal
marital-status	String	Nominal
occupation	String	Nominal
relationship	String	Nominal
race	String	Nominal
gender	String	Nominal
country	String	Nominal
capital-gain	Int	Ratio
hours-per-week	Int	Ordinal
income	String	Ordinal

3.2 前置處理

本研究前先將資料進行標準化，且因員工序號(fnlgwt)對收入的影響並無關聯，故將 fnlgwt 此屬性做資料清理進行刪除。

3.3 實驗設計

先進行前處理，將資料標準化與清理，利用 ID3、C4.5、C5.0 與 CART 決策樹演算法將資料分類預測，對四種演算法得出的準確率做比較。

3.4 實驗結果

模型初步訓練	
Accuracy	0.83

圖 1 模型初步訓練準確率

C4.5 Classification Report:	precision	recall	f1-score	support
0	0.87	0.89	0.88	6159
1	0.63	0.6	0.61	1982
accuracy			0.82	8141
macro avg	0.75	0.74	0.75	8141
weighted avg	0.81	0.82	0.81	8141

圖 2 C4.5 分類預測結果

C5.0 Classification Report:	precision	recall	f1-score	support
0	0.85	0.95	0.9	7397
1	0.76	0.47	0.58	2372
accuracy			0.83	9769
macro avg	0.8	0.71	0.74	9769
weighted avg	0.83	0.83	0.82	9769

圖 3 C5.0 分類預測結果

CART Classification Report:	precision	recall	f1-score	support
0	0.88	0.87	0.88	7418
1	0.61	0.62	0.61	2351
accuracy			0.81	9769
macro avg	0.74	0.75	0.74	9769
weighted avg	0.81	0.81	0.81	9769

圖 4 CART 分類預測結果

ID3 Classification Report:	precision
accuracy	0.77

圖 5 ID3 分類預測結果

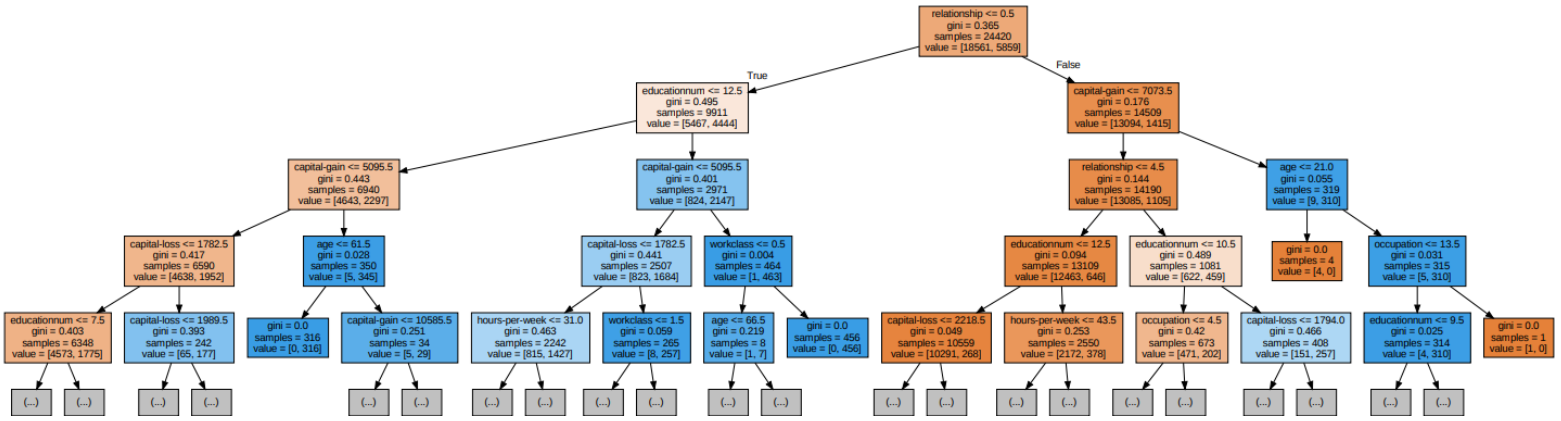

圖 6 C4.5 決策樹

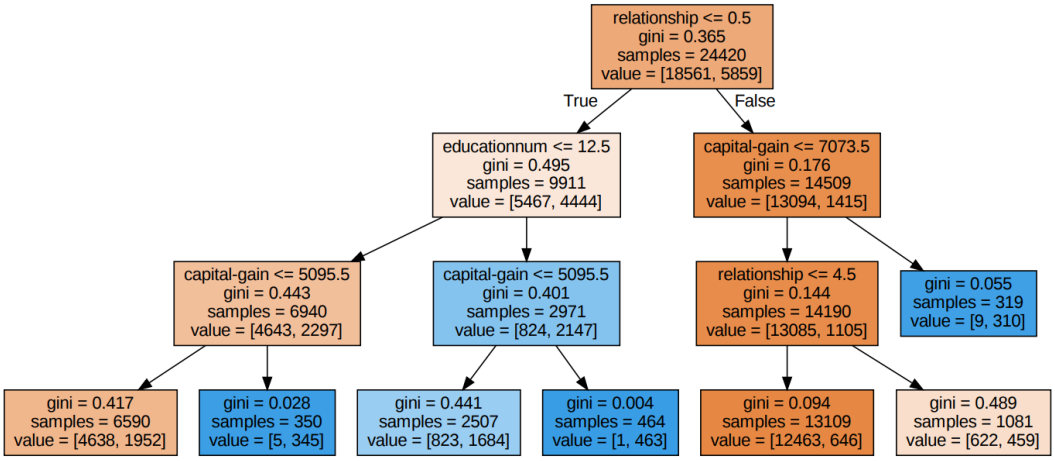

圖 7 C4.5 決策樹 (設定深度與節點數)

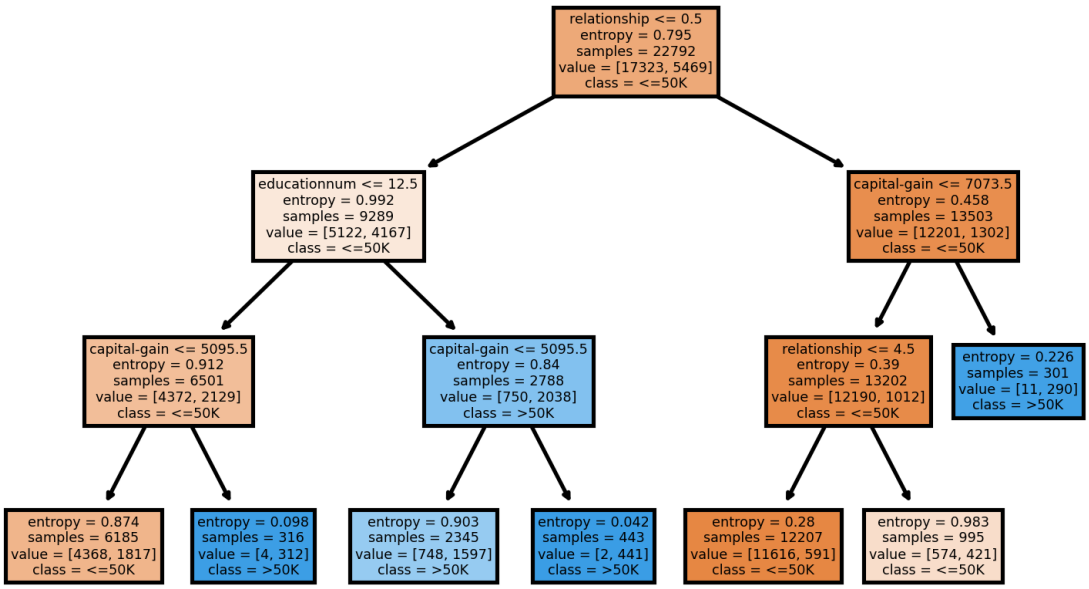

圖 8 C5.0 決策樹

C5.0 Confusion Matrix:		
	7043(TP)	354(FP)
	1268(FN)	1104(TN)

圖 9 C5.0 混淆矩陣

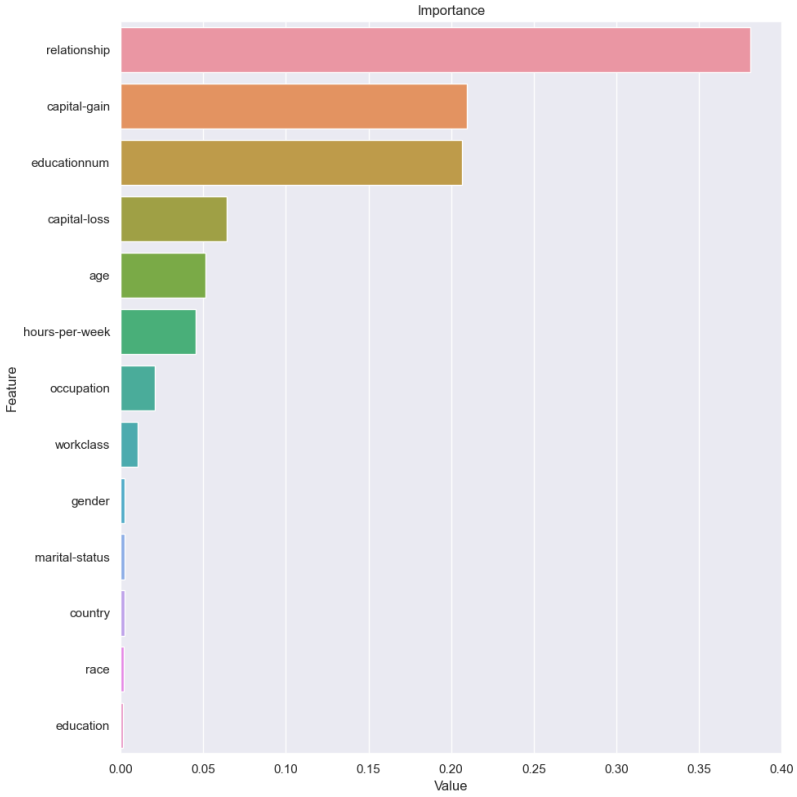

圖 10 CART 屬性之重要性

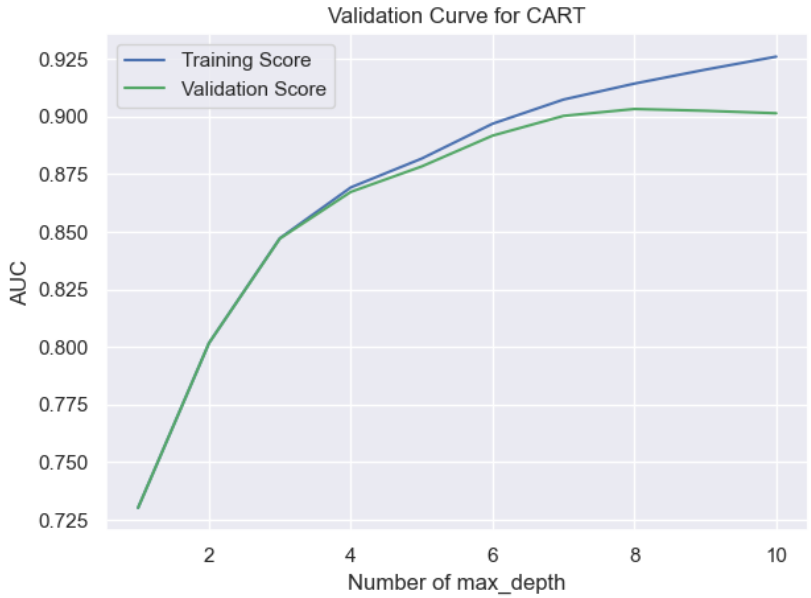

圖 11 CART AUC

CART 的 AUC 介於 0.725~0.925 代表此分類是有預測能力的，因此它在模型評估和比較中是具有價值的。

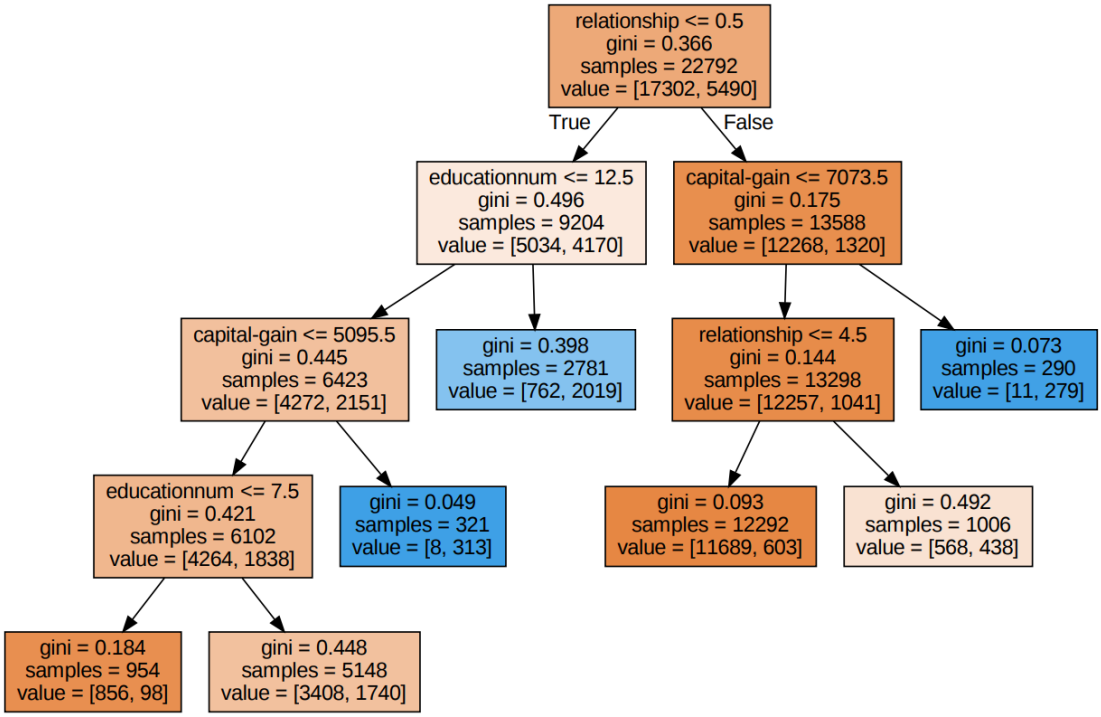

圖 12 CART 決策樹

4.結論

透過演算法結果可以得知 C5.0 相較於其他演算法，它的準確率最高且決策樹規模較小，C5.0 改進 C4.5 的性質與功能，計算速度更快，占用的記憶體更少，提供更高的效率，在處理大數據時更為出色。

參考文獻

Mohammad Ataei(2021 年 6 月 12 日)。Decision-Tree-ID3。https://github.com/mohammadataei93/Decision-Tree-ID3-#decision-tree-id3。

Buya Al-Fariz(2021 年 10 月 24 日)。BIKIN POHON PAKE PYTHON (Algoritma C4.5 Decision Tree)。https://www.youtube.com/watch?v=-PcB6l3F990。

Tanveer Khan(2021 年 7 月 8 日)。C5.0-in-Python。https://github.com/tkhan11/C5.0-in-Python。

Baha Uluğ (2021 年 10 月 20 日)。Classification and Regression Tree (CART)。https://www.kaggle.com/code/bahaulug/classification-and-regression-tree-cart。