國立雲林科技大學資訊管理系

資料探勘-作業一

Department of Information Management

National Yunlin University of Science & Technology

Assignment

# 使用決策樹演算法分析成人資料集

# Analyzing the Adult Dataset by the Decision Tree Algorithm

楊欣蓓、陳怡君、鄭皓名

指導老師：許中川 博士

Advisor: Chung-Chian Hsu, Ph.D.

中華民國 112 年 10 月

October 2023

# 摘要

從過往的數據顯示，隨著不同的環境以及個人背景下，會影響到年薪(income)的多寡，而年薪的高低會跟著影響到生活品質，基於以上考量，本研究選擇 UCI Machine Learning Repository 所提供的成人資料集(Adult data)來做預研究，希望透過不同的個人背景，例如：年齡(age)、教育(education)、婚姻狀況(marital-status)、職業(occupation)…等特徵，來進行決策樹 (decision tree)分析，並預測年薪是否會達到 50K 以上。此外，本研究也會評估不同種類的決策樹，其中包括：ID3、C4.5、C5.0 和 CART(Classification and Regression tree)，分析出各個決策樹的績效後再做評比。最後本實驗找到最佳的決策樹模型為 CART，並延伸下去做探討，去研究 CART 在使用不同超參數組合對模型進行訓練後，會對測試資料的績效造成多少影響，研究後得出未修剪的決策樹與剪枝前差異不大。

針對上述修剪問題，本研究推估目前決策樹的訓練還時間不夠久，若延長決策樹的訓練時間，再進行修剪的話，也許模型的績效會表現的更好。

關鍵字： CART(Classification and Regression tree) 、資料探勘 (Data Mining)、決策樹 (Decision tree)

# 一、 緒論

## 1.1 動機

根據過往的研究顯示，個人學歷、工作類型、家庭狀況...等個人背景會影響到年薪的獲得，而年薪的多寡會反應在各個領域，例如：高年薪者可以選擇購買更豪華的房屋或是提升生活品質、高年薪者可以選擇更多種類的保險，包括醫療保險、汽車保險...等、高年薪者參加更高程度的教育和培訓，這有助於提高專業技能和事業發展機會。

透過上述的舉例，可顯現出年薪對於生活的重要性，本研究選擇成人資料集(Adult data)作為實驗的應用領域，藉由資料集中的教育程度、婚姻狀況...等個人背景，來預測年薪高於 50K 或低於 50K，藉此幫助個人和政策制定者做出明智的決策，以改善生活品質。

## 1.2 目的

本研究採用成人資料集，此資料集涵蓋多種特徵，例如：年齡(Age)、婚姻狀況(Marital-status)、職業(Occupation)、教育程度(Education-num)…等特徵，而每種特徵的預測績效會因為不同種類的決策樹而有所差異，所以需評估不同決策樹種類的性能，其中比較的種類包括:ID3、C4.5、C5.0、CART，以確定何種決策樹演算法能預測最精確，以達到本研究的目標。

最終目標是提供有關在不同個人背景條件下年收入目標的有用的資訊，以提供更有價值的參考依據，針對年薪未達 50K 的族群，希望政策制定者能提供相對應的社會福利或給予生活中的協助，以改善生活品質。透過本次研究，希望能更深入地了解決策樹的技術在預測年收入方面的效用，以提供更多明智的決策。

# 二、 方法

## 2.1 實作說明

本研究的目標是使用不同的決策樹種類，包括 ID3、C4.5、C5.0 和 CART，來分析 UCI Machine Learning Repository 提供的成人數據集。以下將依序介紹資料前處理、模型構建、不同種決策樹的績效評估…等步驟。

首先，成人數據集包含如年齡、婚姻狀況、職業、教育程度等各種特徵。在進行分析前，為了確保資料的一致性，先針對資料進行預處理，其中包含補齊缺失值、One hot encoding、正規化、特殊字元調整、重複資料整理。

接著，使用不同的決策樹種類來構建模型。這些算法將根據資料集的特徵和目標分類（是否年收入超過 5 萬美金）來生成決策樹，並按順序使用 ID3、C4.5、C5.0 和 CART 去評估各自的績效。

## 2.2 操作說明

本研究執行環境採用 Python 3.11.6 與 Python 3.7.16 (由於 Python 調用 R 語言時會發生錯誤，因此降低版本至 3.7)，以 Visual Studio Code 作為開發工具，使用 Pandas、Numpy、sklearn、Matplotlib…等，函式資料庫讀取資料、資料前處理以及將模型訓練的績效以視覺化的形式呈現，透過將資料以獨熱編碼(One-hot encoding)、正規化技術來降低模型過度擬合，以利模型訓練時可以得到較佳的泛化程度(Generalization)，再使用前述四種決策樹演算法進行訓練，最後觀察模型泛化程度。

# 三、 實驗

## 3.1 資料集

- 資料集名稱：成人資料集
- 原始資料筆數：48842
- 前處理後的資料筆數： 48813

**表 1** 成人資料集欄位介紹

| 欄位 | 屬性 | 內容 |
|---|---|---|
| 0 | age | continuous |
| 1 | workclass | Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked. |
| 2 | fnlwgt | continuous |
| 3 | education | Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool. |
| 4 | education-num | continuous |
| 5 | marital-status | Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse |
| 6 | occupation | Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces |
| 7 | relationship | Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried |
| 8 | race | White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black |
| 9 | sex | Female, Male |
| 10 | capital-gain | continuous |
| 11 | capital-loss | continuous |
| 12 | hours-per-week | continuous |
| 13 | native-country | United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands |

## 3.2 資料前處理

- 成人資料集：
  - 將資料欄位為「?」的部分，替代成「Nan」，判斷「Nan」的欄位為名目資料或數值資料，前者取眾數做替補，而後者則使用平均值做填補。
  - 刪除意思相近的特徵欄位，如：'education'、'education-num' 取其一。
  - 刪除對於預測年收入較無高度關聯性的欄位 ' fnlwgt '。
  - 刪除資料集空白的筆數。
  - 名目資料的欄位，如：'workclass','education','maritalstatus','occupation', 'relationship','race','sex','native-country','class'，透過 One hot encoding 轉成數值資料。
  - 將名目資料採獨熱編碼方式處理，經 One-hot encoding 後，訓練集的欄位比測試集多出了'native-country_Holand-Netherlands'欄位，故在測試集新增該欄位，並將其值都設為 0，讓兩個資料集欄位數相同。
  - 將有順序性的欄位資料採用 Label Encoding 技術，本次資料針對欄位 Income 做處理，將>50K 設為 1；<=50K 設為 0。
  - 刪除資料集中資料重複的筆數。
  - 數值資料使用 Normalization 技術(z-score)，將欄位 'age','education-num',capital-gain','capital-loss','hours-per-week' 做處理，以降低模型發生 Overfitting 的狀況。

**表 2** 部分訓練資料進行前處理前的結果

| 資料 / 特徵 | No.0 | No.1 | No.2 | No.3 | No.4 | No.5 |
|---|---|---|---|---|---|---|
| Age | 39 | 50 | 38 | 53 | 28 | 37 |
| education-num | 13 | 13 | 9 | 7 | 13 | 14 |
| capital-gain | 2174 | 0 | 0 | 0 | 0 | 0 |
| . . . . | . . . . | | | | | |
| capital-loss | 0 | 0 | 0 | 0 | 0 | 0 |
| hours-per-week | 40 | 13 | 40 | 40 | 40 | 40 |
| class | <=50K | <=50K | <=50K | <=50K | <=50K | >50K |

**表 3** 部分訓練資料經前處理後的結果

| 資料 / 特徵 | No.0 | No.1 | No.2 | No.3 | No.4 | No.5 |
|---|---|---|---|---|---|---|
| Age | 0.03039 | 0.836973 | -0.04294 | 1.05695 | -0.77619 | -0.11626 |
| education-num | 1.134777 | 1.134777 | 0.42067 | -1.19841 | 1.134777 | 1.523641 |
| capital-gain | 0.148292 | -0.14598 | -0.14598 | -0.14598 | -0.14598 | -0.14598 |
| . . . . | . . . . | | | | | |
| sex_Female | 0 | 0 | 0 | 0 | 1 | 1 |
| sex_Male | 1 | 1 | 1 | 1 | 0 | 0 |
| class | 0 | 0 | 0 | 0 | 0 | 1 |

## 3.3 實驗設計

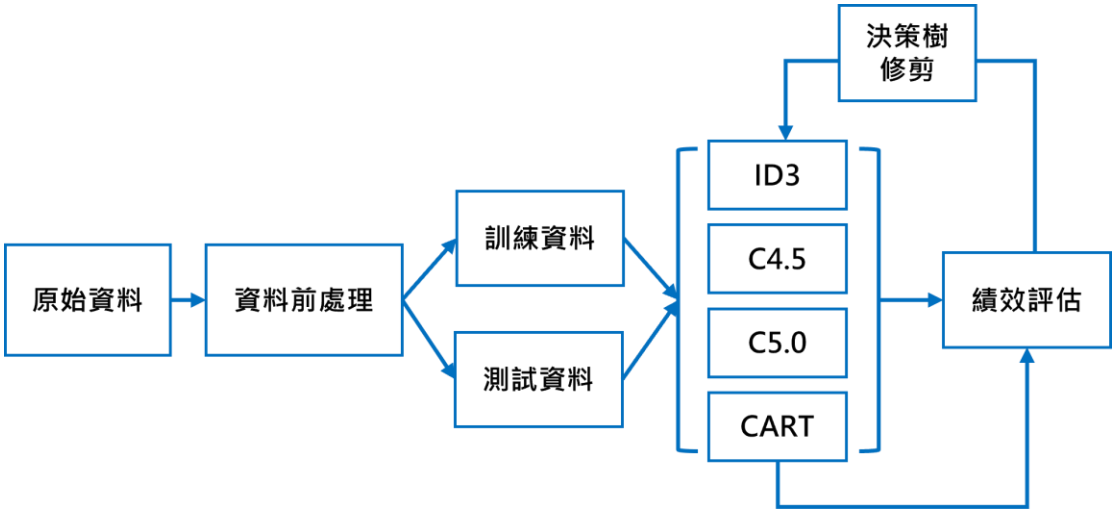



**圖 1** 實驗設計流程

本研究的實驗設計如圖 1，首先將 UCI 的成人資料集做資料前處理，得到乾淨的資料集後，再分成訓練資料和測試資料，分別使用四種決策樹的演算法，包括：ID3、C4.5、C5.0、CART 進行訓練，接著藉由測試資料來進行績效評估，最後針對績效的結果，對決策樹修剪後，再進行一次績效評估。

## 3.4 實驗結果

本研究的實驗分為兩個部分，第一部分：使用四種決策樹模型分別預測成人資料集的年收入，並針對各自的模型測試績效比較，第二部分：使用第一部份模型比較中最佳的決策樹演算法(CART)，嘗試使用不同的超參數組合比較其結果並觀察績效的變化，最後使用決策樹後剪枝技術(Post pruning decision trees)比較修剪後與未修剪的決策樹績效。

### 3.4.1 四種決策樹模型的績效比較

實驗一使用四種決策樹模型：ID3、CART、C4.5、C5.0 對成人資料集進行年收入的類別預測。模型超參數設定為 max_depth 為 10、不純度(impurity)採用 Gini 與 Entropy 作為計算函式、分岔點以模型學習的最佳分割方式(splitter=best)。下表 4 為四種決策樹模型使用測試資料集進行測試後的結果，績效的評估指標使用準確度(Accuracy)與透過混淆矩陣(Confusion Matrix)得出的精確度(Precision)、F1-score，其中 Precision 主要是從預測年收入<=50K 的資料中，找到實際年收入也是<=50K 的資料，而 F1-score 結合了精確度(Precision)與召回率

(Recall)，Recall 為實際年收入為<=50K 的資料中觀察預測年收入也是<=50K 的資料有多少。透過三種不同的績效評估指標，能更全面地觀察出四種決策樹模型的泛化程度。

透過表 4 的績效比較可看出，以 Accuracy 作為評估指標時，CART 與 C5.0 兩種演算法作為預測成人資料集的年收入的績效是最佳的，由於 C5.0 未使用到 Precision 與 F1-score 兩種評估指標，故比較的決策樹演算法以 ID3、CART、C4.5 為主，可以得出 ID3、CART 在使用前述兩種評估指標時，作為預測成人資料集的年收入的績效是最佳的。另外，該實驗中也有試圖找出哪些成人資料集的欄位屬性對於決策樹模型在訓練時發揮較大的影響，如下圖 2 所示，可以發現婚姻狀況對於決策樹模型在訓練時發揮很大的影響。

**表 4** 四種決策樹模型的績效比較

| 模型 | Criterion | Accuracy | Precision | F1-score |
|---|---|---|---|---|
| **ID3** | Entropy | 0.84 | 0.85 | 0.85 |
| **CART** | Gini | 0.86 | 0.85 | 0.85 |
| **C4.5** | Entropy | 0.83 | 0.81 | 0.81 |
| **C5.0** | Entropy | 0.86 | — | — |

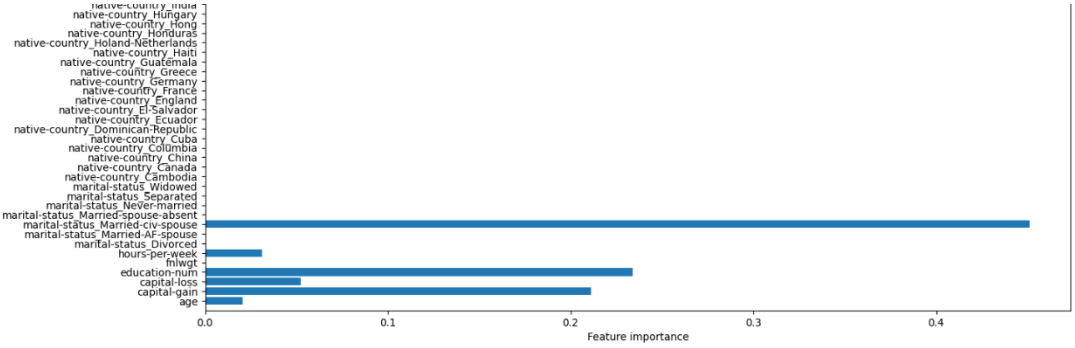



**圖 2** 成人資料集的重要特徵

### 3.4.2 決策樹模型—CART 不同超參數組合比較與事後修剪

實驗第二部分，首先透過決策樹模型—CART 在訓練與測試過程中，加入不同程度的 alpha 值，觀察模型訓練與測試階段準確度的變化，如下圖 3 所示，此處的 alpha 值可視為懲罰項，作為決策樹修剪以避免決策樹模型產生過擬合(Overfitting)的狀況。當決策樹的樹葉節點越多或深度越深時，alpha 值會隨之變大，如下圖 4(a)所示，其目的與前述決策樹修剪的原因相同，為了使決策樹模型能夠達到較好的泛化程度。下圖 4(b)也觀察了所有樹葉節點的不純度與不同的 Alpha 值之間變化的關係。

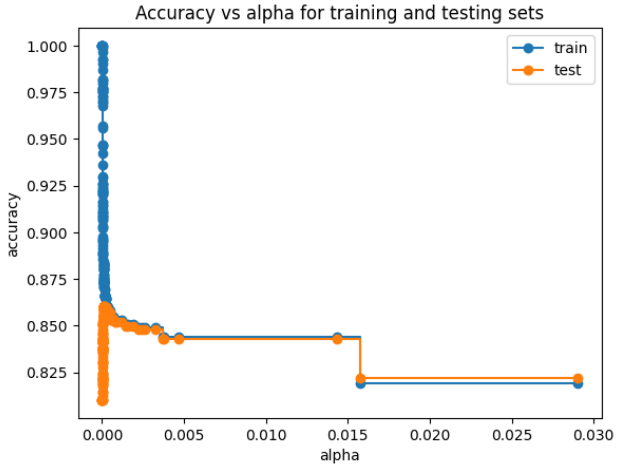
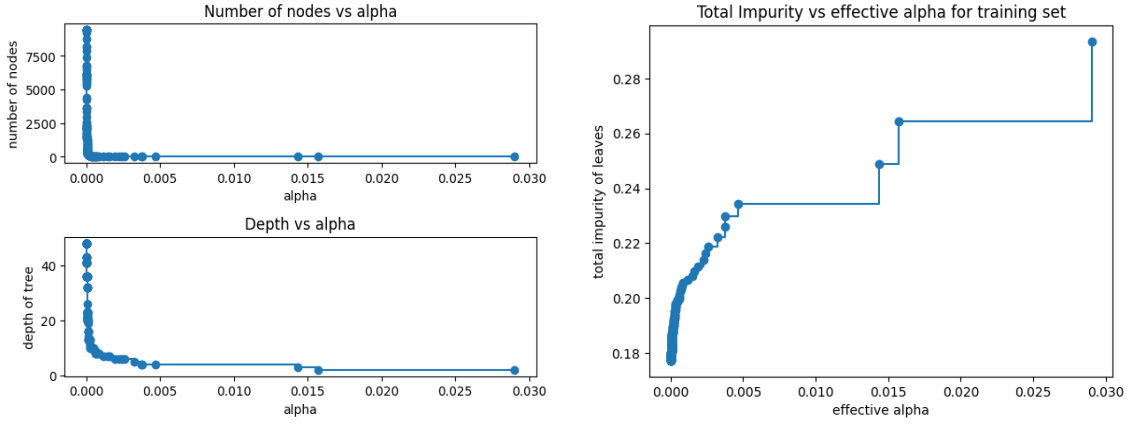



**圖 3** 不同的 Alpha 值與模型績效的關係




(a) (b)

**圖 4** (a)樹葉節點及深度與不同的 Alpha 值的關係 (b) 所有樹葉節點的不純度與不同的 Alpha 值的關係。

接著，依據圖 3 與圖 4 的實驗結果，設定決策樹模型—CART 的不同超參數組合來比較各組合的模型績效。該實驗使用參數網格搜尋結合 K-fold 交叉驗證(K-fold cross validation)方式，來找到排名前三的超參數組合。K-fold 交叉驗證主

要透過設定 k 值來切分訓練資料與測試資料，k 值為訓練的回合數，每次訓練皆會留 1 份的資料作為測試資料，其餘的 k-1 份則作為訓練資料。參數網格搜尋方法須先設定 max_depth、min_samples_leaf、min_samples_split 三種超參數的範圍，如下表 5 所示，再加上 k 值的設定，該實驗的 k 值設定為 7，由於圖 3 所呈現當 alpha 值越小時，決策樹模型的訓練與測試績效皆為最佳狀態，故網格搜尋方法未考慮 alpha 值的設定。

藉由以上設定提供超參數網格搜尋函式能找到績效排名前三的超參數組合，結果如下表 6 所示，可以發現深度在 2 階時績效最佳，min_samples_leaf、min_samples_split 變動較大，這是因為網格搜尋過程中，會計算不純度並對決策樹做剪枝，以達到更好的分類預測績效，min_samples_leaf 主要目的為防止決策樹過度擬合訓練數據，使模型更具泛化能力，而觀察 min_samples_split 的目的即防止決策樹在過小的節點上進行進一步的劃分。此三種超參數組合套入決策樹模型中的績效皆相近。

**表 5** 網格搜尋法的參數設定

| **GridSearchCV** | |
|---|---|
| **max_depth** | 2,4,6,8,10,12,14,16,18,20 |
| **min_samples_leaf** | 2,4,6,8,10,12,14,16,18,20 |
| **min_samples_split** | 2,4,6,8,10,12,14,16,18,20 |

**表 6** CART 模型—不同參數組合結果

| **Assemble** | **Criterion** | **max_depth** | **min_samples_leaf** | **min_samples_split** |
|---|---|---|---|---|
| **A1** | Gini | 2 | 2 | 2 |
| **A2** | Gini | 14 | 12 | 18 |
| **A3** | Gini | 14 | 12 | 20 |

下表 7 為進行決策樹修剪前與經過修剪後，決策樹模型績效的比較。

**表 7**

| **CART** | **Accuracy** | **Precision** | **F1-score** |
|---|---|---|---|
| **Before pruning** | 0.86 | 0.85 | 0.86 |
| **After pruning** | 0.86 | 0.84 | 0.86 |

# 四、 結論

在成人資料集中我們主要目標是預測年收入，透過實驗一依序使用四種決策樹模型：ID3、CART、C4.5、C5.0 對成人資料集進行年收入的類別預測，最後綜合所有評估指標，找到分類預測準確度最佳的決策樹模型為 CART。接著，以實驗一所找到之最佳決策樹模型，繼續往下延伸探討 CART 在使用不同超參數組合對模型進行訓練後，對測試資料預測的績效的影響，為實驗二的部分。最後發現，決策樹修剪後的績效與未修剪的績效差異不大，本研究團隊推估可能是因為決策樹訓練時間不夠久，倘若訓練時間再延長，使決策樹模型更深更大，再進行樹葉修剪，或許模型的績效相對來說會相對於修剪前還要好。

# 參考文獻

Ronny Kohavi、Barry Becker (1996)。成人數據集。
https://archive.ics.uci.edu/ml/datasets/adult

Ryan Lu (2018)。Preprocessing Data : 類別型特徵_OneHotEncoder & LabelEncoder 介紹與實作。
https://medium.com/ai%E5%8F%8D%E6%96%97%E5%9F%8E/preprocessing-data-onehotencoder-labelencoder-%E5%AF%A6%E4%BD%9C-968936124d59