**USING DECISION TREE TO ANALYSIS WOMEN'S CONTRACEPTIVE METHODS CHOICE AND ADULT ANNUAL INCOME**

**以決策樹分析預測婦女避孕方法以及預測成人年收入**

林郁凱，彭冠穎，彭嘉瑋，陳羿欣

Department of Information Management, National Yunlin University of Science and Technology

**摘要**

關於多屬性分類的問題現今已有許多成熟的技術可解決分類之問題，像是SVM、神經網路（Neural Networks，NN）以及決策樹（Decision tree），由於決策樹模型有可解釋性且模型建構容易，因此在本研究中使用的決策樹模型作為本研究主要之預測方法，並探討Contraceptive Method Choice Data Set，如何預測出婦女適合選擇何種避孕方法以及Adult Data Set依成人資料背景來預測其年薪是否超過5萬。

**一、緒論**

**1.1動機**

本研究使用兩種資料集，利用決策樹模型預測以及解決兩資料集所遇到之問題。

根據聯合國統計，目前全球人口數最多的兩個國家為中國以及印度，中國早在1979年已經開始實行節制生育的政策，然而印度在1952年已經提出人口控制計畫，為育齡之婦女提供避孕方法和服務，但是到了1972年印度人口居然有增無減，從3.8億人口激增到6.13億，漲幅高達61%。其可能歸咎於政策沒有強制執行。因此在避孕選擇資料集（Contraceptive Method Choice Data Set）包含了婦女以及她的丈夫的年齡、教育、職業、宗教信仰與生活水準等屬性，哪些婦女會選擇長期服用、短期服用或是不服用避孕藥，根據上述資料來預測女性當前適合的避孕方法。

我們在第二個資料集中，主要預測個人年收入是否超過5萬美金，個人的年收入與其教育程度、婚姻狀況等等屬性有關係，我們透過成人資料集（Adult Data Set）所收集之人口普查收入數據來預測何種個人背景的情況下年收入能超過5萬美金。

然而在上述的兩組資料集的問題皆為分類(Classifier)問題，關於資料分類預測的任務上，已有許多成熟的技術可解決分類之問題，像是早期的Rule-base系統、Memory base reasoning，到現在的機器學習（machine learning）演算法，目前最常用的演算法包括Bayesian Belief Networks、支援向量機(Support Vector Machines，SVM)、神經網路（Neural network）以及決策樹（Decision tree），由於決策樹可解釋性原則（understandable），且模型易於建構，因此在本研究中，我們採用決策樹學習來預測本研究之兩組資料集。

**1.2目的**

決策樹(Decision tree)可以用來分析資料輔助決策，利用樹結構來分析資料，由於樹的結構易於理解且容易實現，與神經網路之差異為我們可以透過決策樹來解釋分析後之結果所表達之意義，且在短時間內能夠對大型數據做出良好的分析結果。本研究使用之兩種資料集，皆從UCI開放資料庫所下載之，欲解決之問題分別為「婦女之避孕選擇」以及「個人年收入是否超過50K」，資料已經做過一定的整理，資料完整度甚高，因此我們利用決策樹能在短時間內得出有效之結果，並可透過決策樹的圖像清楚得知該資料集之分析流程及相關參數。

**二、方法**

**2.1 實作說明**

本研究將「避孕選擇資料集（Contraceptive Method Choice Data Set）」、「成人資料集（Adult Data Set）」做數據的前置處理，其包括資料清理、Ono-hot encoding、資料切割（即為將資料分成訓練資料(train data)以及測試資料(test data)），模型建構則利用決策樹，分別選用Gini以及Entropy衡量各個樹節點（node）之不純度（Impurity），並且控制樹的深度(depth)，最後預測「婦女會選擇何種的避孕方式」以及「何種成人的年收入會超過美金5萬元」，並輸出兩者的決策樹模型以及訓練資料與測試資料的績效（accuracy），最終匯出分析結果表格。

**2.2 操作說明**

本研究執行環境皆為Python3.6，分別使用Jupyter Notebook以及Spyder作為分析工具，利用Pandas、Numpy來讀取資料以及做資料的前處理，預測模型則利用Sklearn套件的Tree工具來建構，最後在使用Graphviz、Matplotlib套件將決策樹圖形輸出之。

**三、實驗**

**3.1資料集**

本研究使用兩組資料集做預測分析，分別為1987年印度婦女問卷調查結果與成人收入調查結果之相關數據，以下為該兩組資料集之資料名稱、資料筆數，以及資料表的欄位介紹。

**3.1.1** **避孕選擇-婦女相關背景資料**

* 名稱：Contraceptive Method Choice資料集
* 原始資料筆數：1473
* 正規化後之訓練資料筆數：1178
* 正規化後之測試資料筆數：295

表格 一：Contraceptive Method Choice資料集欄位介紹

| 欄位 | 欄位名稱 | 內容 |
| --- | --- | --- |
| 0 | age | continuous |
| 1 | Wife's education | (categorical) 1=low, 2, 3, 4=high |
| 2 | Husband's education | (categorical) 1=low, 2, 3, 4=high |
| 3 | Number of children | continuous |
| 4 | Wife's religion | (binary) 0=Non-Islam, 1=Islam |
| 5 | Wife's working | (binary) 0=Yes, 1=No |
| 6 | Husband's occupation | (categorical) 1, 2, 3, 4 |
| 7 | Standard-of-living | (categorical) 1=low, 2, 3, 4=high |
| 8 | Media exposure | (binary) 0=Good, 1=Not good |
| 9 | Contraceptive method used | class attribute  1=No-use 2=Long-term 3=Short-term |

表格 二：顯示部分Contraceptive Method Choice資料集

（欄位編號對應表一之欄位名稱）

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 欄位 | 0 | 1 | 2 | 3 | 4 |
| 0 | 24 | 45 | 43 | 42 | 36 |
| 1 | 2 | 1 | 2 | 3 | 3 |
| 2 | 3 | 3 | 3 | 2 | 3 |
| 3 | 3 | 10 | 7 | 9 | 8 |
| 4 | 1 | 1 | 1 | 1 | 1 |
| 5 | 1 | 1 | 1 | 1 | 1 |
| 6 | 2 | 3 | 3 | 3 | 3 |
| 7 | 3 | 4 | 4 | 3 | 2 |
| 8 | 0 | 0 | 0 | 0 | 0 |
| 9 | 1 | 1 | 1 | 1 | 1 |

**3.1.2 薪資調查-部分國家成人資料**

* 名稱：Adult資料集
* 原始資料筆數：48842
* 正規化後之訓練資料筆數：30162
* 正規化後之測試資料筆數：15059

表格 三 ：Adult資料集欄位介紹

| 欄位 | 屬性 | 內容 |
| --- | --- | --- |
| 0 | age | continuous |
| 1 | workplace | Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked |
| 2 | fnlwt | 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 |
| 14 | salary | <=50K,>50K |

表格 四：顯示部分Adult資料集

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 欄位 | 0 | 1 | 2 | 3 | 4 |
| 0 | 39 | 50 | 38 | 53 | 28 |
| 1 | State-gov | Self-emp-not-inc | Private | Private | Private |
| 2 | 77516 | 83311 | 215646 | 234721 | 338409 |
| 3 | Bachelors | Bachelors | HS-grad | 11th | Bachelors |
| 4 | 13 | 13 | 9 | 7 | 13 |
| 5 | Never-married | Married-civ-spouse | Divorced | Married-civ-spouse | Married-civ-spouse |
| 6 | Adm-clerical | Exec-managerial | Handlers-cleaners | Handlers-cleaners | Prof-specialty |
| 7 | Not-in-family | Husband | Not-in-family | Husband | Wife |
| 8 | White | White | White | Black | Black |
| 9 | Male | Male | Male | Male | Female |
| 10 | 2174 | 0 | 0 | 0 | 0 |
| 11 | 0 | 0 | 0 | 0 | 0 |
| 12 | 40 | 13 | 40 | 40 | 40 |
| 13 | United-States | United-States | United-States | United-States | Cuba |
| 14 | <=50K | <=50K | <=50K | <=50K | <=50K |

**3.2前置處理**

圖 1：前置處理流程圖

* 資料清理：將資料中的缺失值以及過濾與預測結果不相關資訊。
* One-Hot encoding：對非數值的類別屬性進行特徵數字化。
* 資料切割：將80%當成訓練資料，其餘20%為測試資料。

表格 五：避孕選擇資料集之資料前處理後部分資料

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| age | wife\_  edu | husband\_  edu | child\_  num | religion | worked | Husband  occupation\_1 | Husband  occupation\_2 | Husband  occupation\_3 | Husband  occupation\_4 | living\_standard | media | contraception |
| 24 | 2 | 3 | 3 | 1 | 1 | 0 | 1 | 0 | 0 | 3 | 0 | 1 |
| 45 | 1 | 3 | 10 | 1 | 1 | 0 | 0 | 1 | 0 | 4 | 0 | 1 |
| 43 | 2 | 3 | 7 | 1 | 1 | 0 | 0 | 1 | 0 | 4 | 0 | 1 |

表格 六：成人資料集之資料前處理後部分資料

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| age | education-num | hours-per-week | new\_income | workclass\_ Federal-gov | workclass\_ Local-gov | workclass\_ Private | workclass\_ Self-emp-inc | workclass\_ Self-emp-not-inc | workclass\_ State-gov | ... |
| 39 | 13 | 40 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | ... |
| 50 | 13 | 13 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | ... |
| 38 | 9 | 40 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | ... |
| 53 | 7 | 40 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | ... |
| 28 | 13 | 40 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | ... |

**3.3實驗設計**

以下為Contraceptive Method Choice資料集與Adult資料集模型建構之設計，依序的流程步驟：

圖 2：實驗流程圖

**3.3.1避孕選擇**

1. 建構樹並設定樹深度最大為**7**，分割方法為Gini。
2. 將欲處理之資料匯入樹當中並訓練
3. 將樹做視覺化並驗證樹之架構
4. 產出其正確率
5. 測試多個參數並使用最佳化之參數
6. 匯出預測結果

**3.3.2薪資調查-部分國家成人資料**

1. 建構樹並設定樹深度最大為5，分割方法為Entropy。。

2. 將欲處理之資料匯入樹當中並訓練

3. 將樹做視覺化並驗證樹之架構

4. 產出其正確率

5. 測試多個參數並使用最佳化之參數

6. 匯出預測結果

**3.4實驗結果**

以下為避孕選擇以及成人資料集之預測圖表以及決策樹圖：

**3.4.1****避孕選擇**

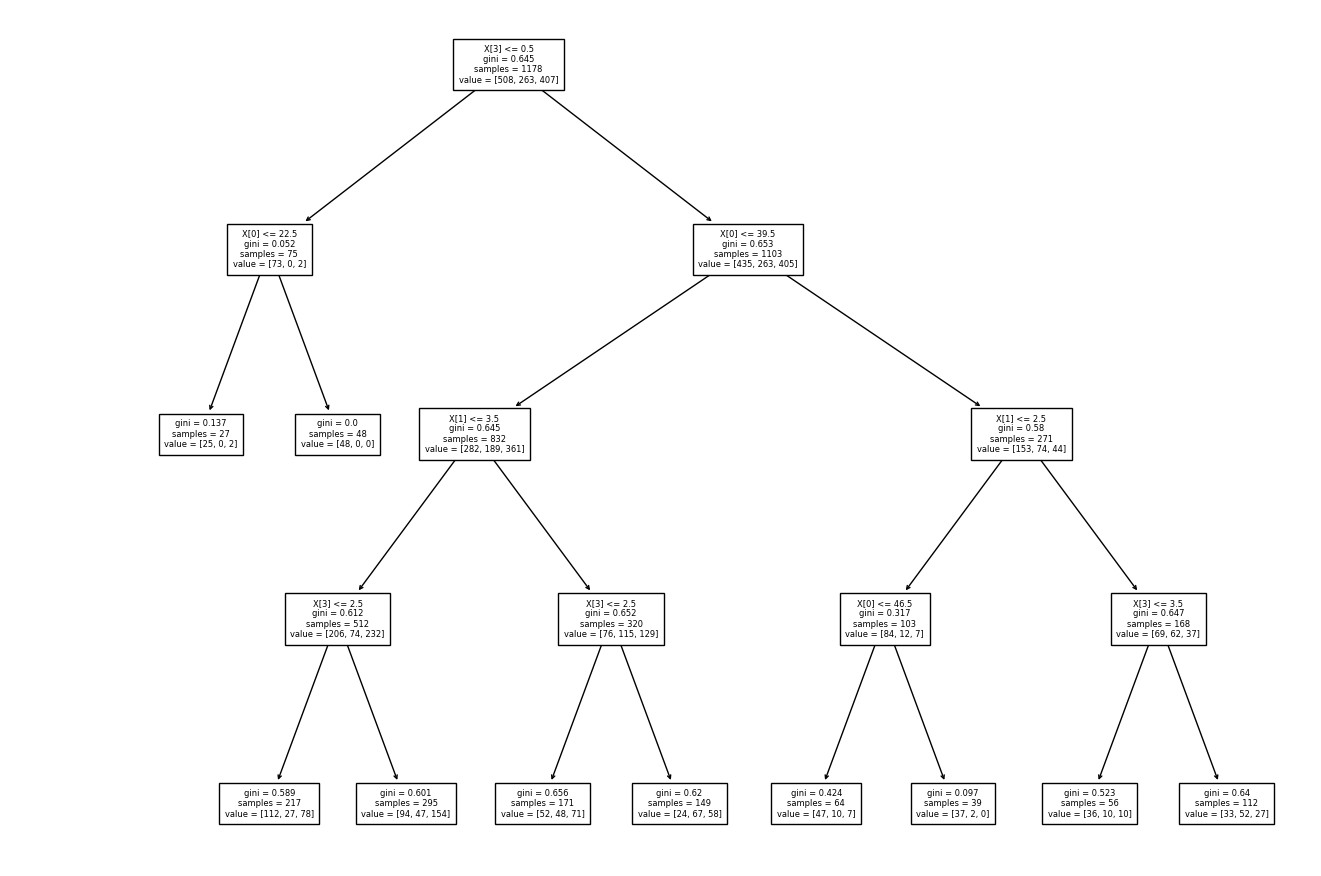


圖 3：避孕選擇資料集之決策樹圖形

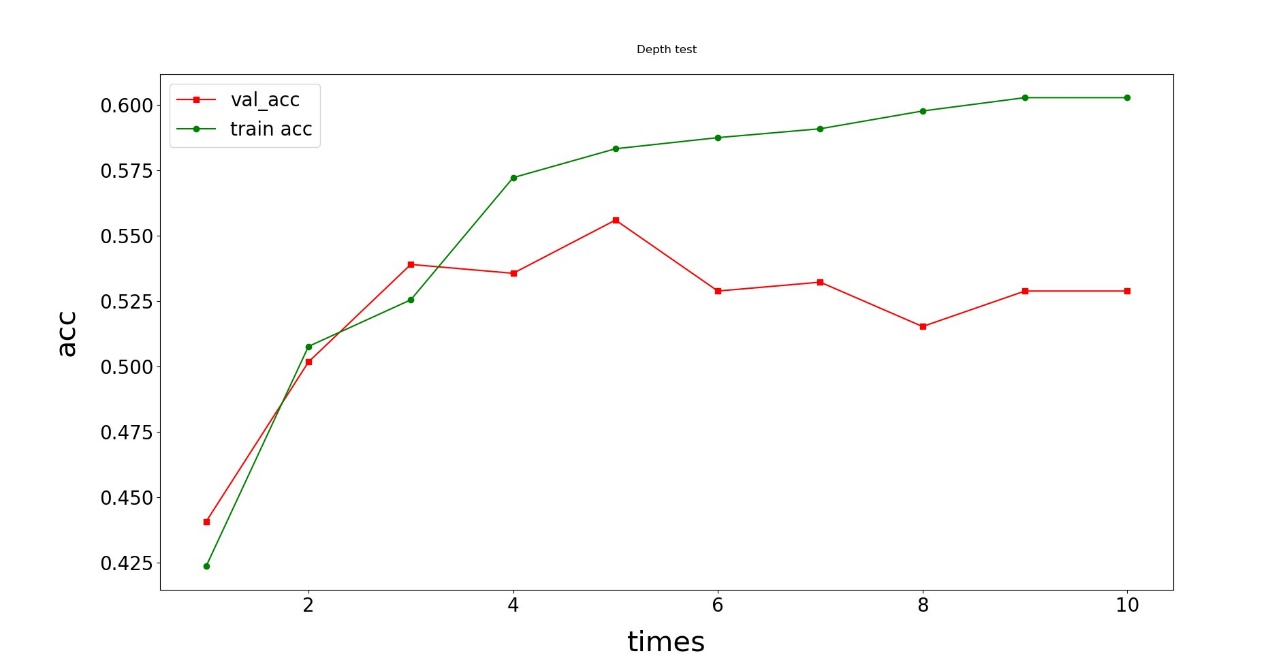


圖 4：避孕選擇資料集之訓練、測試資料集之準確度。

(X軸為樹深度，Y軸為預測之準確度)

**3.4.2薪資調查-部分國家成人資料**



圖 5：避孕選擇資料集之決策樹圖形

(該部分較清晰之圖表已放置Github)

圖 6：避孕選擇資料集之訓練、測試資料集之準確度。

(X軸為樹深度，Y軸為預測之準確度)

**四、結論**

由上述兩項實驗可以發現不同決策樹深度及寬度參數預測準確度會有所差異，另外決策樹並非為最好的預測模型，在Contraceptive Method Choice資料集中，進行10次的預測，精準度皆在0.4~0.65之間浮動，若樹的深度越深，則會產生overfitting的情形，Adult 資料集雖有較良好的預測結果，精準度約在0.80~0.85之間，同樣地若再加深樹的深度亦會產生overfitting之情形，因此決策樹方法雖然易於實現且具可解釋性，但是預測資料不一定會有良好的預期結果。

**參考文獻**

[1] Tjen-Sien Lim (1997) A subset of the 1987 National Indonesia Contraceptive Prevalence Survey[Data Set].

[2] Ronny Kohavi, Barry Becker(1996) Adult Data Set[Data Set].

[3] Post pruning decision trees with cost complexity pruning ( https://scikit-learn.org/stable/auto\_examples/tree/plot\_cost\_complexity\_pruning.html#sphx-glr-auto-examples-tree-plot-cost-complexity-pruning-py).