

## 作業 3: 決策問題決定一切

### Q1. 典型的資料分析可以用以解決下面六種問題。

Data analysts typically work with six problem types

1. Making predictions 2. Categorizing things 3. Spotting something unusual 4. Identifying themes 5. Discovering connections 6. Finding patterns

Resource :

- Excerpt From Google Data Analytics.  
<https://www.linkedin.com/pulse/excerpt-from-google-data-analytics-oladapo-ishola-olayinka>
- Common Problem Types of Data Analysis with Examples from our Daily Lives  
[https://www.linkedin.com/pulse/common-problem-types-data-analysis-examples-from-our-daily-atasert?trk=public\\_profile\\_article\\_view](https://www.linkedin.com/pulse/common-problem-types-data-analysis-examples-from-our-daily-atasert?trk=public_profile_article_view)
- Problem Solving with Data Analytics | Google Data Analytics Certificate  
[https://www.youtube.com/watch?v=wSct\\_cbqzpM](https://www.youtube.com/watch?v=wSct_cbqzpM)

### Making predictions 預測結果

Using data to make informed decisions about how things may be in the future.

A company that wants to know the best advertising method to bring in new customers. Analysts with data on location, type of media, and number of new customers acquired as a result of past ads can't guarantee future results, but they can help predict the best placement of advertising to reach the target audience.

### Categorizing things 分類事物

Grouping data based on common features.

An example of a problem requiring analysts to categorize things is a company's goal to improve customer satisfaction. Analysts might classify customer service

calls based on certain keywords or scores. This could help identify top-performing customer service representatives or help correlate certain actions taken with higher customer satisfaction scores.

## **Spotting something unusual 察覺異相**

### **Identifying data that is different from the norm.**

A company that sells smart watches that help people monitor their health would be interested in designing their software to spot something unusual. Analysts who have analyzed aggregated health data can help product developers determine the right algorithms to spot and set off alarms when certain data doesn't trend normally.

## **Identifying themes 辨識主題**

### **Recognizing broader concepts and trends from categorized data.**

User experience (UX) designers might rely on analysts to analyze user interaction data. Similar to problems that require analysts to categorize things, usability improvement projects might require help prioritize the right product features for improvement. Themes are most often used to help researchers explore certain aspects of data. In a user study, user beliefs, practices, and needs are examples of themes.

By now you might be wondering if there is a difference between categorizing things and identifying themes. The best way to think about it is: categorizing things involves assigning items to categories; identifying themes takes those categories a step further by grouping them into broader themes.

## **Discovering connections**

### **Identifying similar challenges across different entities—and using data and insights to find common solutions.**







A third-party logistics company working with another company to get shipments delivered to customers on time is a problem requiring analysts to discover connections. By analyzing the wait times at shipping hubs, analysts

can determine the appropriate schedule changes to increase the number of on-time deliveries.

## Finding patterns

Using historical data about what happened in the past to understand how likely it is to happen again.

Minimizing downtime caused by machine failure is an example of a problem requiring analysts to find patterns in data. For example, by analyzing maintenance data, they might discover that most failures happen if regular maintenance is delayed by more than a 15-day window.

<b>1. Making predictions</b> 	<b>2. Categorizing things</b> 	<b>3. Spotting something unusual</b> 
<b>4. Identifying themes</b> 	<b>5. Discovering connections</b> 	<b>6. Finding patterns</b> 

接下來，請依照您抽到的牌卡順序，各給出例子，並應用可能的 open data 資料說明其輸入和輸出資料。

例如：

1. Making predictions（行）：

- 輸入：[即時交通事故資料\(AI 類\)](#) | [政府資料開放平臺 \(data.gov.tw\)](#)

- 輸出：預測交通事故風險，使用預測模型對未來可能發生事故的地點和時間進行預測。
- 決策：制定交通管制、監測和安全提示等策略，減少交通事故風險。

## 2. Categorizing things (老)：

- 輸入：[臺南市各區長照需求人口推估表 - 臺南市 111 年度各區長照需求人口推估表 - 臺南市政府資料開放平台 \(tainan.gov.tw\)](#)
- 輸出：分類區域是否具有高比例的 65 歲以上失能老人，透過將人口以 "65 歲以上失能老人" 欄位分成兩類，假設 500 人以上為高比率，500 人以下為低比率。
- 決策：如果地區 65 歲以上失能老人的比例較高，則相關機構可以考慮增加該地區的居家照護服務和社區照顧中心等措施，以支持這些老人在家中生活。

## 3. Spotting something unusual (病)：

- 輸入：[台灣 COVID-19 冠狀病毒檢測每日送驗數 | 政府資料開放平臺 \(data.gov.tw\)](#)
- 輸出：使用時間序列分析來檢測病毒數量隨著時間的變化是否有異常高峰或下降。同時，比較陽性測試結果和總測試量，以判斷陽性率是否有異常的變化，這可能意味著病毒的傳播率發生了變化。

- 決策：及早發現疫情對於採取快速有效的公共衛生措施阻止病毒傳播非常重要。這能讓公共衛生官員盡快採取隔離、接觸者追蹤、檢疫措施和增加檢測等干預措施，以打破傳播鏈，防止進一步傳播。

#### 4. Identifying themes （衣）：

〔範例一〕

- 輸入：[Women's E-Commerce Clothing Reviews](#)
- 輸出：藉由 Class Name 產品類別（已分類）並分析 Review Text 客戶正面的評論，識別出當前的潮流趨勢和受歡迎的服裝款式，並分析共同的優缺點（優點包含受女性喜愛的共同原因；缺點包含被客訴的理由）。
- 決策：以消費者為導向，幫助服裝品牌和零售商更好地了解市場需求和趨勢，並及時調整產品設計和庫存管理。

〔範例二〕

- Input: 電子發票消費熱度指標
- Output: 不同地區的消費熱度趨勢
- Usage: 政府可以透過分析電子發票消費熱度指標，制定相應的經濟、行銷等方面的政策，以推動產業發展

#### 5. Discovering connections （育）

- 輸入：各級學校分布位置、[臺北市各學校可上網載具設備配備情形](#)

- 輸出：網路設備配備情形和學校的分布位置之間的相關性

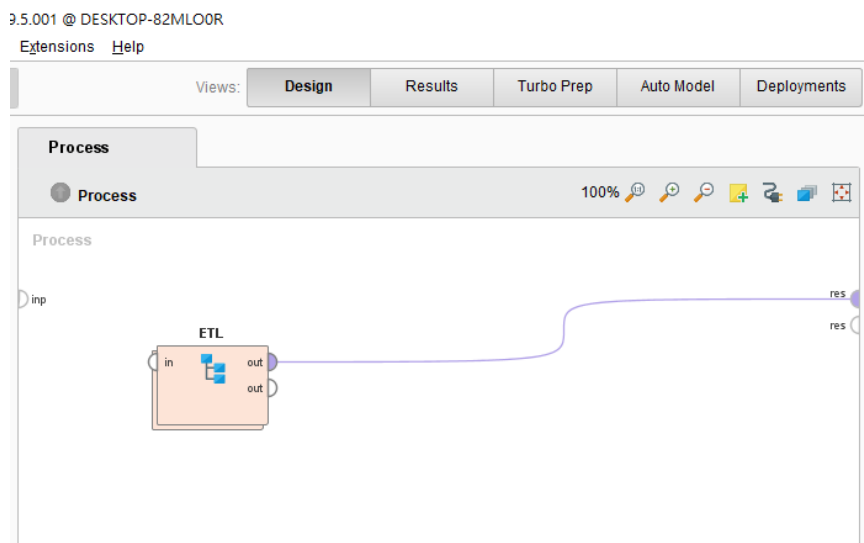
## 6. Finding patterns (樂) :

- 輸入：[Daily Box Office For 2023 - Box Office Mojo](#)
- 輸出：預測電影季節性趨勢，分析資料以確定電影收入是否存在任何季節性趨勢。
- 決策：如果分析結果顯示某些月份或季節電影收入通常更高，投資者可能會考慮投資於這些時間上映的電影。此信息還可以幫助投資者有效地分配預算並優化投資組合，以利用季節性趨勢。如果分析顯示某些季節或月份對電影業來說通常利潤較低，投資者可以調整他們的投資策略，以減少在這些時期的風險。

## Q2 薪資大不同

2-1 讀入檔案(salary.csv),請大概說明一下您對於資料的了解(5%),並進行「必要」之資料前處理,並請將相關前處理存於 ETL 子流程如下圖(請說明您做了什麼,為什麼)

(10%)



- Objective 目標：觀察各個特徵 ( age 、 workclass 、 fnlwgt 、 education 、 education-num 、 marital-status 、 occupation 、 relationship 、 race 、 sex 、 capital-gain 、 capital-loss 、 hours-per-week 、 native-country ) 是否會影響薪資 ( 以 50K 為基準 )

※ [ 思考 ] 第 2-4 的管理意涵 → 個人/人資/主管的角度去分析

- 資料欄位說明

欄位	內容	資料型態
age	年齡	整數 Integer
workclass	就業狀況、工作類型	Nominal

education	教育程度	Nominal
education-num	受了多久的教育	整數 Integer
marital-status	婚姻狀況	Nominal
occupation	職業	Nominal
relationship	家庭角色	Nominal
race	種族	Nominal
sex	性別	Nominal
capital-gain	資本收益 ( 股票、房產等 )	整數 Integer
capital-loss	資本損失 ( 房屋稅等 )	整數 Integer
hours-per-week	每週工作小時	整數 Integer
native-country	國籍	Nominal
salary	薪資 ( 年收入 ) 是否大於 50K	Nominal

- 資料前處理 ( 先處理缺失值後，再去看敘述統計 )
- 有缺失值的欄位：workclass ( 1836 筆)、occupation ( 1843 筆)、native-country ( 583 筆)
- [ Workclass ]

Workclass 欄位中的各資料筆數



workclass	
Private	22696
Self-emp-not-inc	2541
Local-gov	2093
?	1836
State-gov	1298
Self-emp-inc	1116
Federal-gov	960
Without-pay	14
Never-worked	7

以 Private 占最大宗，因此 workclass 的缺失值我會填上 Private。

### ➤ [ occupation ]

另外透過資料觀察，當 workclass 為缺失值，occupation 通常也為缺失值。

#### Result:

From the above table we can see that wherever the 'workclass' feature missing, the 'occupation' feature is also missing. Hence, 'occupation' is missing at random.

So, here is how we deal with it:

We will fill 'workclass' NaN values by its mode (most frequent), and then we will fill the 'occupation' missing values by the value which has the highest frequency with 'workclass' being the mode.

當 workclass 為 Private 時，occupation 的各資料筆數：

occupation	
Craft-repair	3195
Sales	2942
Adm-clerical	2833
Other-service	2740
Exec-managerial	2691
Prof-specialty	2313
Machine-op-inspct	1913
Handlers-cleaners	1273
Transport-moving	1266
Tech-support	736
Farming-fishing	455
Protective-serv	190
Priv-house-serv	149

以 Craft-repair 占最大宗，所以當 workclass 為缺失值且 occupation 為缺失值

時，我會在 occupation 欄位上填寫 Craft-repair ( workclass→填上 Private ；

occupation→填上 Craft-repair ) 。

occupation 第一次填完後，剩下的資料是：workclass 為 Never-worked，我會

在 occupation 填上 No Occupation 。

	A	B	C	D	E	F	G
1	age	workclass	fnlwgt	education	education-num	marital-status	occupation
5363	18	Never-worked	206359	10th	6	Never-married	?
10847	23	Never-worked	188535	7th-8th	4	Divorced	?
14774	17	Never-worked	237272	10th	6	Never-married	?
20339	18	Never-worked	157131	11th	7	Never-married	?
23234	20	Never-worked	462294	Some-college	10	Never-married	?
32306	30	Never-worked	176673	HS-grad	9	Married-civ-spouse	?
32316	18	Never-worked	153663	Some-college	10	Never-married	?

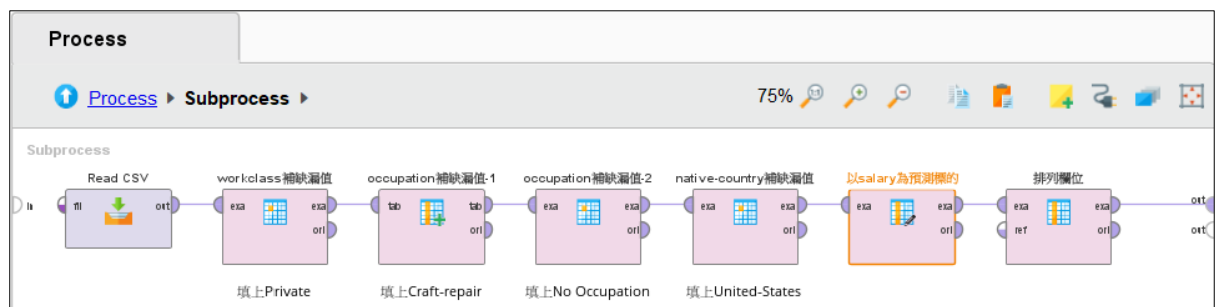
➤ [ native-country ]

native-country	
United-States	29170
Mexico	643
Philippines	198
Germany	137
Canada	121
Puerto-Rico	114
El-Salvador	106
India	100
Cuba	95
England	90
Jamaica	81
South	80
China	75
Italy	73

以 United-States 占最大宗，因此 native-country 的缺失值我會填上 United-

States 。

➤ 以上資料前處理流程



- 對於資料的了解 - 初步判斷是否呈現常態分佈（視覺化初判推測）



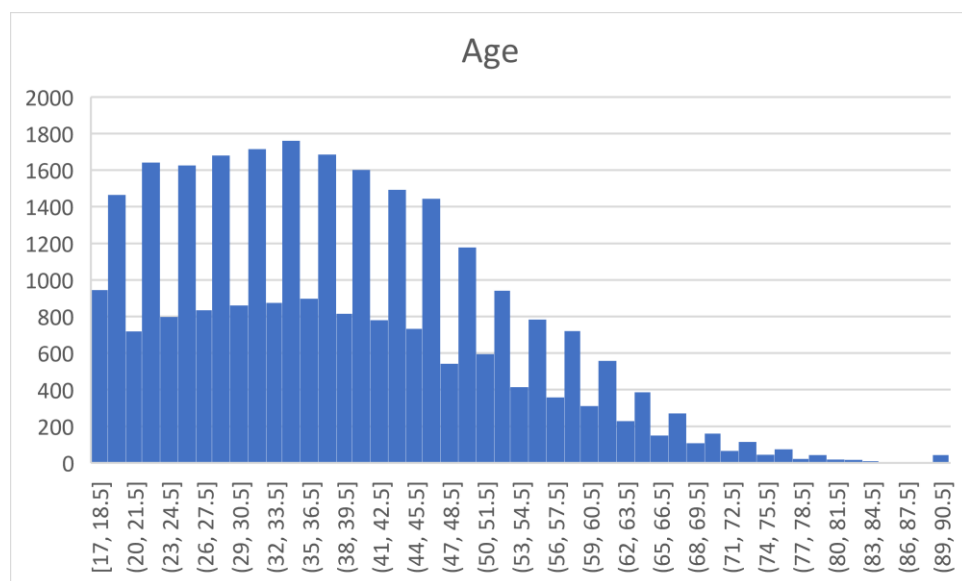
敘述統計.xlsx

〔敘述統計〕

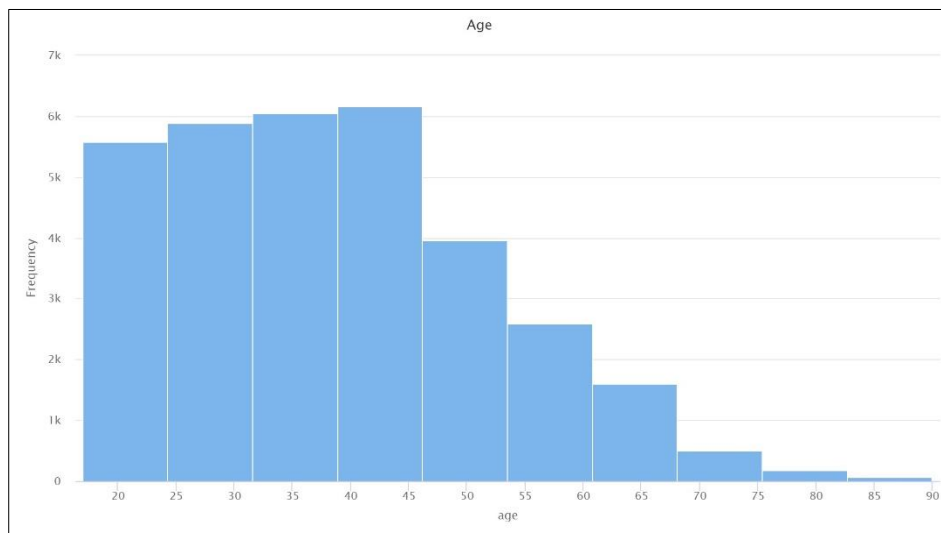
數值型欄位資料	平均數	中間值	標準差	偏度	偏態	最小值	最大值	Q1四分位數	Q3四分位數
age	38.58	37	13.640433	0.5587	右偏	17	90	28	48
education-num	10.08	10	2.5727203	-0.312	左偏	1	16	9	12
capital-gain	1077.65	0	7385.2921	11.954	右偏	0	99999	0	0
capital-loss	87.30	0	402.96022	4.5946	右偏	0	4356	0	0
hours-per-week	40.44	40	12.347429	0.2276	右偏	1	99	40	45

〔資料樣態〕

➤ Age



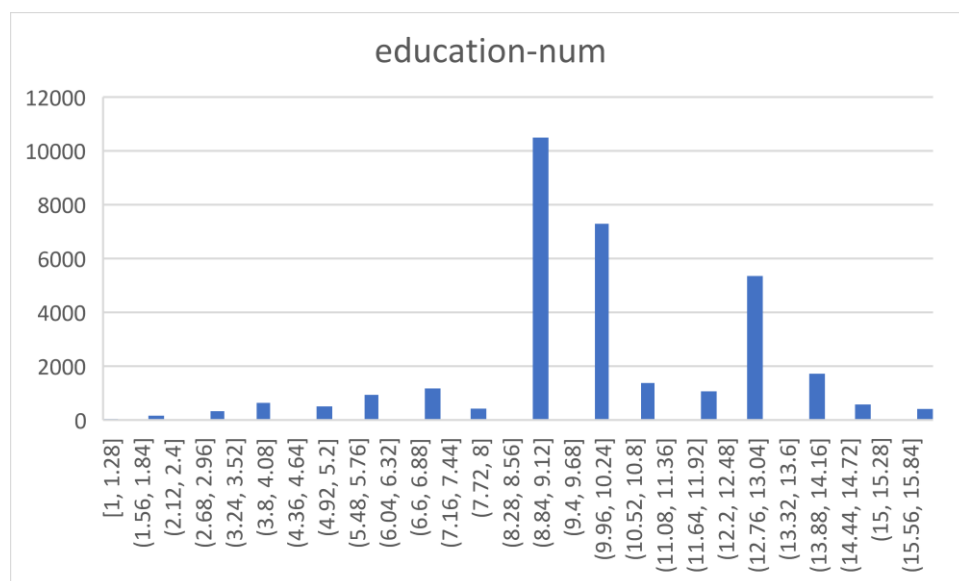
By Excel



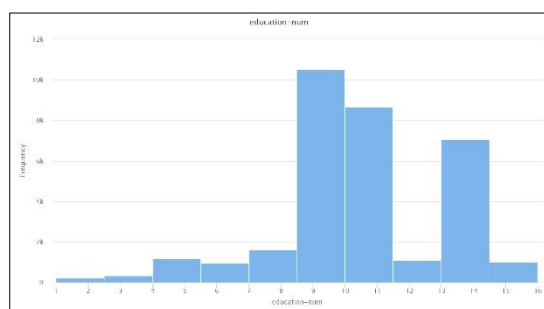
By RapidMiner

→ 非常態分佈，資料分配較多集中在低數那方。

education-num



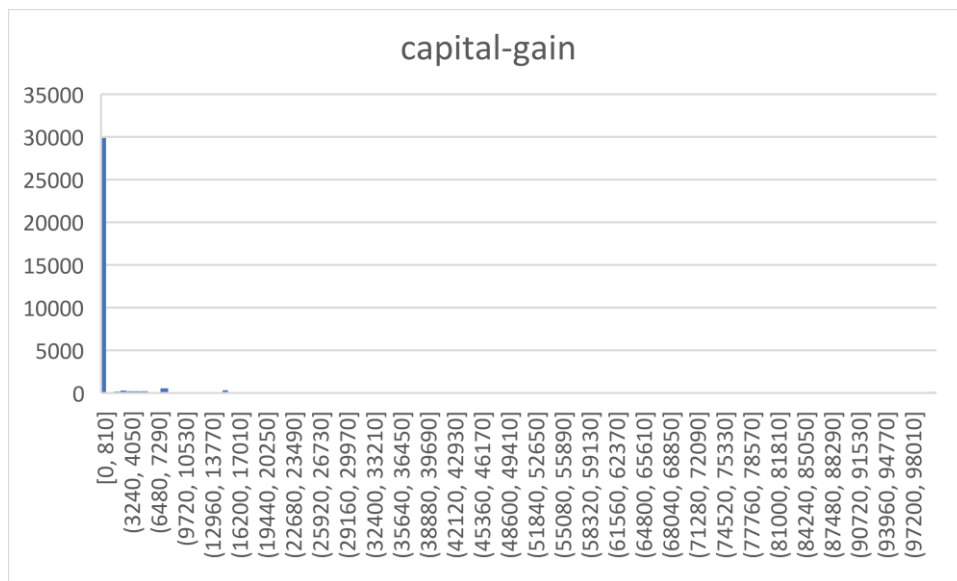
By Excel



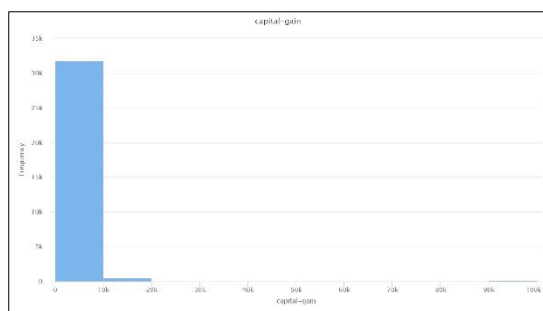
ByRapidMiner

→ 非常態分佈，資料分配較多集中在數字較大的那一方。

➤ capital-gain



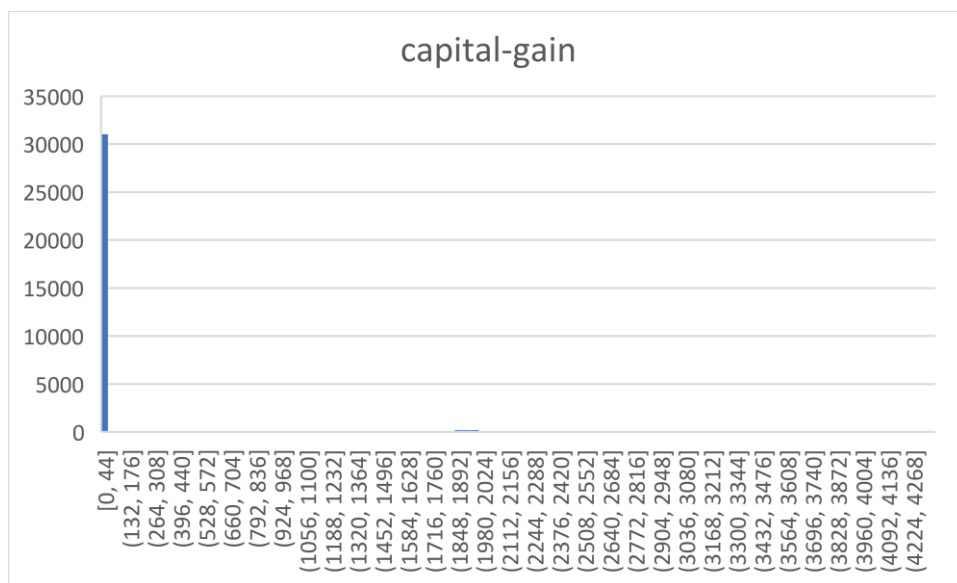
By Excel



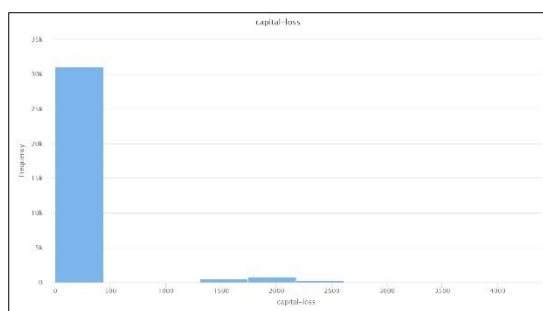
ByRapidMiner

→ 非常態分佈，資料分配較多集中在低數那方。

➤ capital-loss



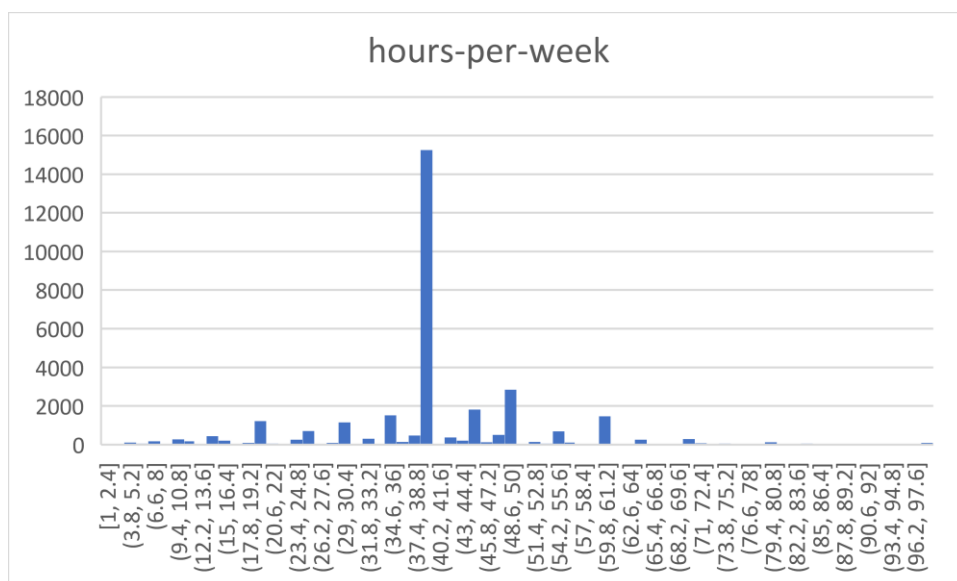
By Excel



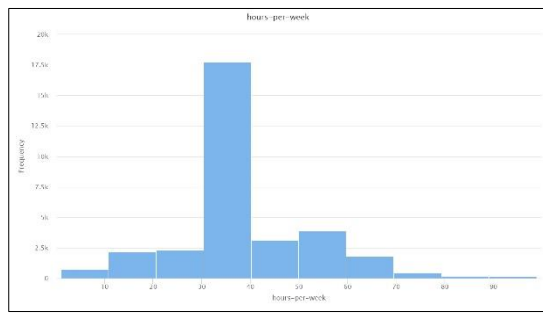
ByRapidMiner

→ 非常態分佈，資料分配較多集中在低數那方。

➤ hours-per-week



By Excel



ByRapidMiner

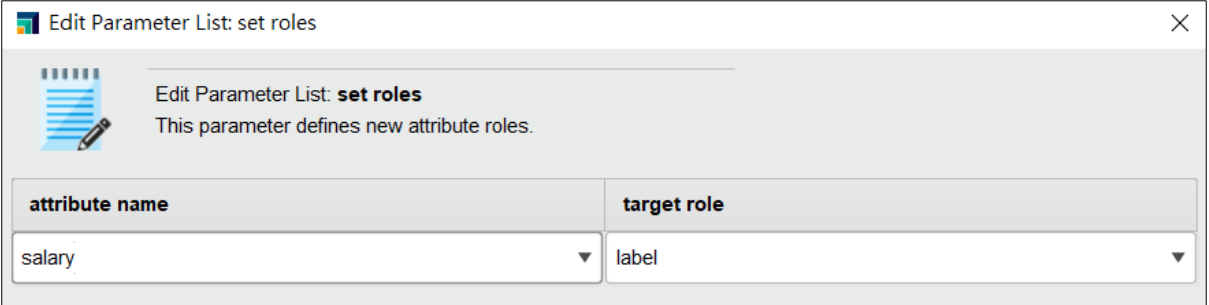
→ 非常態分佈，資料分配較多集中在低數那方。

〔提問〕關於敘述統計或觀察資料分布，應該要是在填補缺失值之前或之後？

- 我翻閱網路上的資料，大部分是在之前；
- 可是我主觀認為要是之後，資料異動過後再觀察，才有意義。

2-2 .請試著解決並以 salary 為預測標的，貼出您的決策樹,並加以說它的意涵(10%)

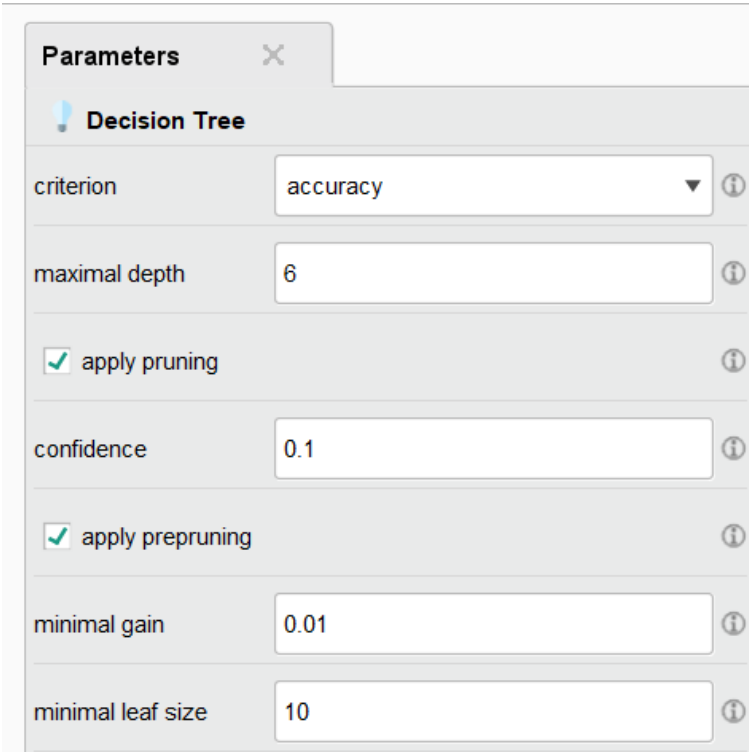
➤ 以 salary 為預測標的



attribute name	target role
salary	label

➤ Decision Tree Operator

資料集切分比例為 Train:Test = 0.7:0.3



**Parameters**

**Decision Tree**

criterion: accuracy

maximal depth: 6

☒ apply pruning

confidence: 0.1

☒ apply prepruning

minimal gain: 0.01

minimal leaf size: 10

Maximal Depth=6 ; minimal leaf size=10

最多 6 層 且 一個節點 ( Node ) 要過 10 個才會繼續往下分

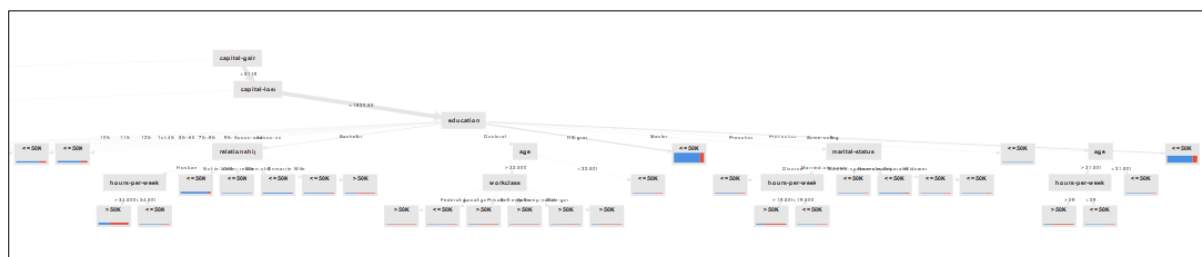
→準確率已超過 85%



## ➤ 決策樹

決策樹透過特徵與對應的值將資料切分，來找出最適合的分枝並繼續往下拓展，

並且依據訓練出來的規則來對新樣本進行預測。（就是每次選擇一個特徵，將資料分成多個部分。）



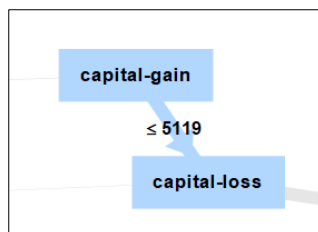
### Tree

```
capital-gain > 5119: >50K {<=50K=58, >50K=1033}
capital-gain <= 5119
| capital-loss > 1820.500
| | relationship = Husband
| | | capital-loss > 1989.500
| | | | capital-loss > 2364.500: >50K {<=50K=10, >50K=48}
| | | | capital-loss <= 2364.500: <=50K {<=50K=51, >50K=5}
| | | capital-loss <= 1989.500: >50K {<=50K=11, >50K=358}
| | relationship = Not-in-family
| | | capital-loss > 2365.500: >50K {<=50K=5, >50K=23}
| | | capital-loss <= 2365.500: <=50K {<=50K=74, >50K=6}
| | relationship = Other-relative: <=50K {<=50K=6, >50K=4}
| | relationship = Own-child: <=50K {<=50K=16, >50K=3}
| | relationship = Unmarried
| | | education-num > 9.500: >50K {<=50K=4, >50K=8}
| | | education-num <= 9.500: <=50K {<=50K=9, >50K=2}
| | relationship = Wife: >50K {<=50K=8, >50K=50}
| capital-loss <= 1820.500
| | education = 10th: <=50K {<=50K=614, >50K=24}
| | education = 11th: <=50K {<=50K=773, >50K=26}
| | education = 12th: <=50K {<=50K=285, >50K=18}
| | education = 1st-4th: <=50K {<=50K=110, >50K=5}
| | education = 5th-6th: <=50K {<=50K=221, >50K=8}
| | education = 7th-8th: <=50K {<=50K=411, >50K=17}
| | education = 9th: <=50K {<=50K=324, >50K=16}
| | education = Assoc-acdm: <=50K {<=50K=550, >50K=157}
| | education = Assoc-voc: <=50K {<=50K=712, >50K=192}
| | education = Bachelors
```

範例：總共有 22,793 筆資料作為訓練，其中 5,489 筆資料 *Salary* > 50K。

### 〔資料主要劃分流程〕

1. 第一關會以 *capital-gain* 為判斷標準，是否  $\leq 5119$

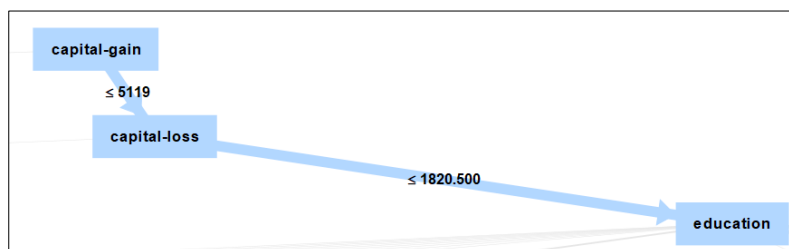


```
capital-gain > 5119: >50K {<=50K=58, >50K=1033}
capital-gain ≤ 5119
```

A. capital-gain > 5,119 的資料有 1091 筆，在這當中 *Salary* > 50K 的有 1033 筆。

B. 95.21%的資料 capital-gain ≤ 5119，繼續第二關。

2. 第二關會以 capital-loss 為判斷標準，是否 ≤ 1820.5



A. 若 capital-loss > 1820.5，則以 relationship 繼續切分。

( 因此種資料僅占 3.08%，不再贅述 )

```
capital-loss > 1820.500
| relationship = Husband
| | capital-loss > 1989.500
| | | capital-loss > 2364.500: >50K {<=50K=10, >50K=48}
| | | capital-loss ≤ 2364.500: <=50K {<=50K=51, >50K=5}
| | | capital-loss ≤ 1989.500: >50K {<=50K=11, >50K=358}
| | relationship = Not-in-family
| | | capital-loss > 2365.500: >50K {<=50K=5, >50K=23}
| | | capital-loss ≤ 2365.500: <=50K {<=50K=74, >50K=6}
| | relationship = Other-relative: <=50K {<=50K=6, >50K=4}
| | relationship = Own-child: <=50K {<=50K=16, >50K=3}
| | relationship = Unmarried
| | | education-num > 9.500: >50K {<=50K=4, >50K=8}
| | | education-num ≤ 9.500: <=50K {<=50K=9, >50K=2}
| | relationship = Wife: >50K {<=50K=8, >50K=50}
capital-loss ≤ 1820.500
```

B. 92.14%的資料  $\text{capital-loss} \leq 1820.5$ ，繼續第三關。

3. 第三關會以 Education 為判斷標準

A. 若僅受基本教育 ( 1<sup>st</sup>-10<sup>th</sup> 小學 1 年級-高一、Preschool 幼兒園、HS-grad 高中畢業或 Some-college 大學未畢業 )，則多數資料  $\text{Salary} < 50K$ 。

```
education = 10th: <=50K {<=50K=614, >50K=24}  
education = 11th: <=50K {<=50K=773, >50K=26}  
education = 12th: <=50K {<=50K=285, >50K=18}  
education = 1st-4th: <=50K {<=50K=110, >50K=5}  
education = 5th-6th: <=50K {<=50K=221, >50K=8}  
education = 7th-8th: <=50K {<=50K=411, >50K=17}  
education = 9th: <=50K {<=50K=324, >50K=16}
```

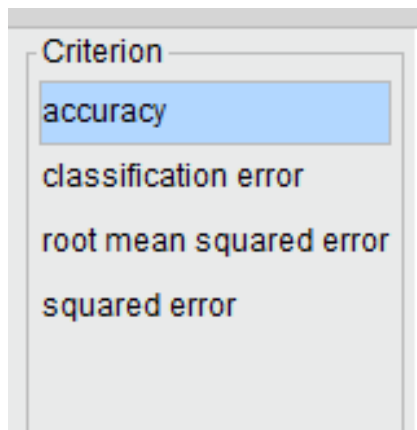
B. Education 為 Bachelors，則以 relationship 往下劃分；

Education 為 Doctorate 或 Prof-school，則以 Age 及 workclass 往下劃分；

Education 為 Masters，則以 marital-status 往下劃分。

2-3 請列出您的混淆矩陣,並請加以說明其意涵，除此之外，我還希望您能列出您模型

的下述 4 個指標，並請簡略說明他們的計算方式及含意(5%)



➤ 混淆矩陣

accuracy: 85.01%			
	true <=50K	true >50K	class precision
pred. <=50K	7025	1073	86.75%
pred. >50K	391	1279	76.59%
class recall	94.73%	54.38%	

True/False 預測正確？		Positive/Negative 預測方向	
	實際 YES	實際 NO	
預測 YES	TP (True Positive)	FP (False Positive) Type I Error	
預測 NO	FN (False Negative) Type II Error	TN (True Negative)	

- ✓ TP：實際答案為 True，而模型預測結果也為 True。預測的結果與實際情況相同。
- ✓ TF：實際答案為 False，而模型預測結果也為 False。預測的結果與實際情況相同。
- ✓ FP：實際答案為 False，而模型預測結果卻為 True。預測的結果與實際情況不同。
- ✓ FN：實際答案為 True，而模型預測結果卻為 False。預測的結果與實際情況不同。

[ 模型指標 ]

### PerformanceVector

```
PerformanceVector:
accuracy: 85.01%
ConfusionMatrix:
True:  <=50K  >50K
<=50K: 7025   1073
>50K:  391    1279
classification_error: 14.99%
ConfusionMatrix:
True:  <=50K  >50K
<=50K: 7025   1073
>50K:  391    1279
weighted_mean_recall: 74.55%, weights: 1, 1
ConfusionMatrix:
True:  <=50K  >50K
<=50K: 7025   1073
>50K:  391    1279
weighted_mean_precision: 81.67%, weights: 1, 1
ConfusionMatrix:
True:  <=50K  >50K
<=50K: 7025   1073
>50K:  391    1279
root_mean_squared_error: 0.347 +/- 0.000
squared_error: 0.121 +/- 0.237
```

※RMSE & Squared Error 已在作業二敘述，因此不再贅述。

- Accuracy 準確率：在所有情況中，正確判斷真假的比例。

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\rightarrow (7025 + 1279) / 9768 = 85.01\%$$

- Classification Error：在所有情況中，錯誤判斷真假的比例。（即 1-Accuracy Score）

$$\text{公式：} (fp + fn) / (tp + fp + fn + tn)$$

$$\rightarrow (1073 + 391) / 9768 = 1 - 85.01\% = 14.99\%$$

- 精確率 Precision：判斷為真的情況下，有多少是真的真。

$$Precision = \frac{TP}{TP + FP}$$

$$\begin{aligned} \rightarrow \text{weighted\_mean\_precision} &= ((7025 / (1073 + 7025)) + (1279 / (1279 + 391))) / 2 \\ &= 81.67\% \end{aligned}$$

- 召回率 Recall：為真的情況下，有多少被正確判斷出來。

$$Recall = \frac{TP}{TP + FN}$$

$$\begin{aligned} \rightarrow \text{weighted\_mean\_recall} &= ((1279 / (1073 + 1279)) + (7025 / (7025 + 391))) / 2 = \\ &= 74.55\% \end{aligned}$$

## 2-4 最後，請說明您的管理意涵？ (10%)

決策→先想假設我是誰？

### ➤ [ 個人人生管理 ]

- 思考：站在自身的角度，如何拿到超過 50K 的薪資？
- 假設現階段我的 Capital Gain 少於 5,119 且 Capital Loss 少於 1,820.5、  
Education 為 Master ( 前提能順利畢業 XD )，則我下一步要關注的點是  
marital-status ( 是否結婚 )，**最好要為 Married-civ-spouse** ( 已婚平民配偶 )，而且每週工作小時 hours-per-week 還得超過 19.500 小時，才有可能獲得超過 50K 的薪資。然而這還得看 Precision 精確率，模型算出這樣的結果，我們能信任多少，在這邊我們得到的 Precision 是 81.67%，認為還算高，所以未來畢業後，**最好能找個對象結婚**。
- ◆ 樹狀圖路線：capital-gain≤5119 → capital-loss≤1820.5 →  
education=Masters → marital-status=Married-civ-spouse →  
hours-per-week>19.5
- 若無法碩士畢業，僅有學士文憑，則必須成為 Husband 且每週工時增加為 34.500 小時，才或許能拿到超過 50K 的薪資。
- ◆ 樹狀圖路線：capital-gain≤5119 → capital-loss≤1820.5 →  
education=Bachelors → relationship=Husband → hours-per-week>34.5

➤ [ 站在人資角度 ]

■ 思考：公司是否應該發給這位應徵者或員工超過 50K 的薪資？

A. 若有跟我一樣資質的人，給予薪資的判斷標準依序為 Capital Gain→Capital Loss→Education→marital-status→hours-per-week。

B. 另外可依據 Education 學歷程度制定不同的工作時數， Bachelors 學士要工作時間較長 ( 34.5 小時 )、Masters 碩士工作時間 19.5 小時，才核可發給超過 50K 的薪資。

※PS. 只要在 32.5 歲前取得博士學位，不管從事什麼工作類型 ( Federal-gov 聯邦政府、Private 個體、Self-emp-inc 有限責任公司自由職業、Self-emp-not-inc 非有限責任公司自由職業、State-gov 州政府 ) 有很大的比率薪資超過 50K。

➤ [ 站在公司或主管角度 ]

■ 身為主管的我，如何帶領同仁拿到更高的薪水？

■ 如同我在 2-2 題所敘述的內容，影響薪資的第三個關鍵就是**教育程度** ( 最好要有 Bachelors 學士、Masters 碩士、Doctorate 博士 )，所以若我今天身為主管，要**帶領部門成員都拿高薪**，一定會鼓勵組員們繼續進修，充實自我專業技術，**拿到更高的文憑**，不斷學習和提高技能以保持競爭力；同時向公司申請舉辦教育訓練，制定培訓和發展計畫，及鼓勵 33 歲前的職員獲得博士學位，給予學習獎勵金，讓員工可以在不同領域和職位上發揮自己的才能和實現職業目標。



請將您的 **process** 存檔為學號-1, 如: 106AB001\_1.rmp 檔

## Reference :

### 1. Kaggle Salary Prediction Classification

- <https://www.kaggle.com/code/mahmoudftolba/salary-prediction-classification/notebook>
- <https://www.kaggle.com/datasets/ayessa/salary-prediction-classification?datasetId=2096417&sortBy=commentCount>
- 数据挖掘笔记六：KNN 分类  
<https://www.xdnote.com/data-mining-knn/>

### 2. Kaggle Salary Prediction Classification — NoteBook

- <https://www.kaggle.com/code/hikmatullahmohammadi/salary-classification-3-models-comparison>
- <https://www.kaggle.com/code/mahmoudftolba/salary-prediction-classification>
- <https://www.kaggle.com/code/artugceylan/salary-prediction-classification-87-score>

### 3. Decision Tree

- Creating a 'Decision Tree' Model  
<https://academy.rapidminer.com/learn/video/creating-a-decision-tree-model>
- Decision Tree demo  
<https://academy.rapidminer.com/learn/video/decision-tree-demo>
- Building Decision Tree Models using RapidMiner Studio  
<https://www.youtube.com/watch?v=U3FVLqV5Jzg>
- Interpret a Decision Tree in RapidMiner - Classification - Business Intelligence with Data Mining  
<https://www.youtube.com/watch?v=9ygE-CzOm7Y>