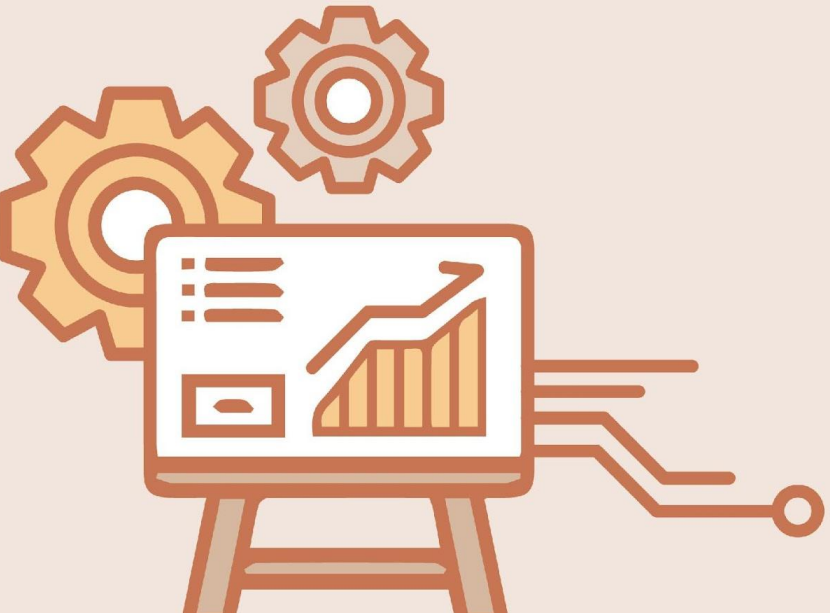
主題 3

講到大數據，

一定要分類分析呀！



- ✔ 分類
- ✔ 決策樹
- ✔ 實作
- ✔ 評估指標

問答

1 大數據分類是將數據按照特定標準劃分為多個類別

是　　非

問答

1 大數據分類是將數據按照特定標準劃分為多個類別

問答

2 大數據分類的目的是為了減少數據儲存成本

是　　非

問答

2 大數據分類的目的是為了減少數據儲存成本

問答

3 大數據分類只能使用結構化數據進行

是　　非

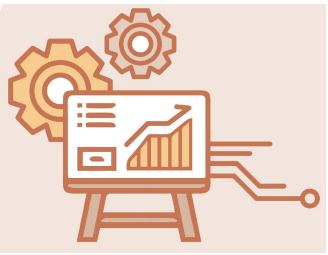
問答

3 大數據分類只能使用結構化數據進行

何謂分類？

- 分類是一種監督式學習
- 帶有類標籤的樣本訓練集
- 根據訓練集對新數據進行分類

- 目標：預測新樣本的分類標籤
- 輸入：一組訓練樣本，每個樣本都要貼類標籤
- 輸出：基於訓練集和分類標籤的模型（分類器）



分類演算法

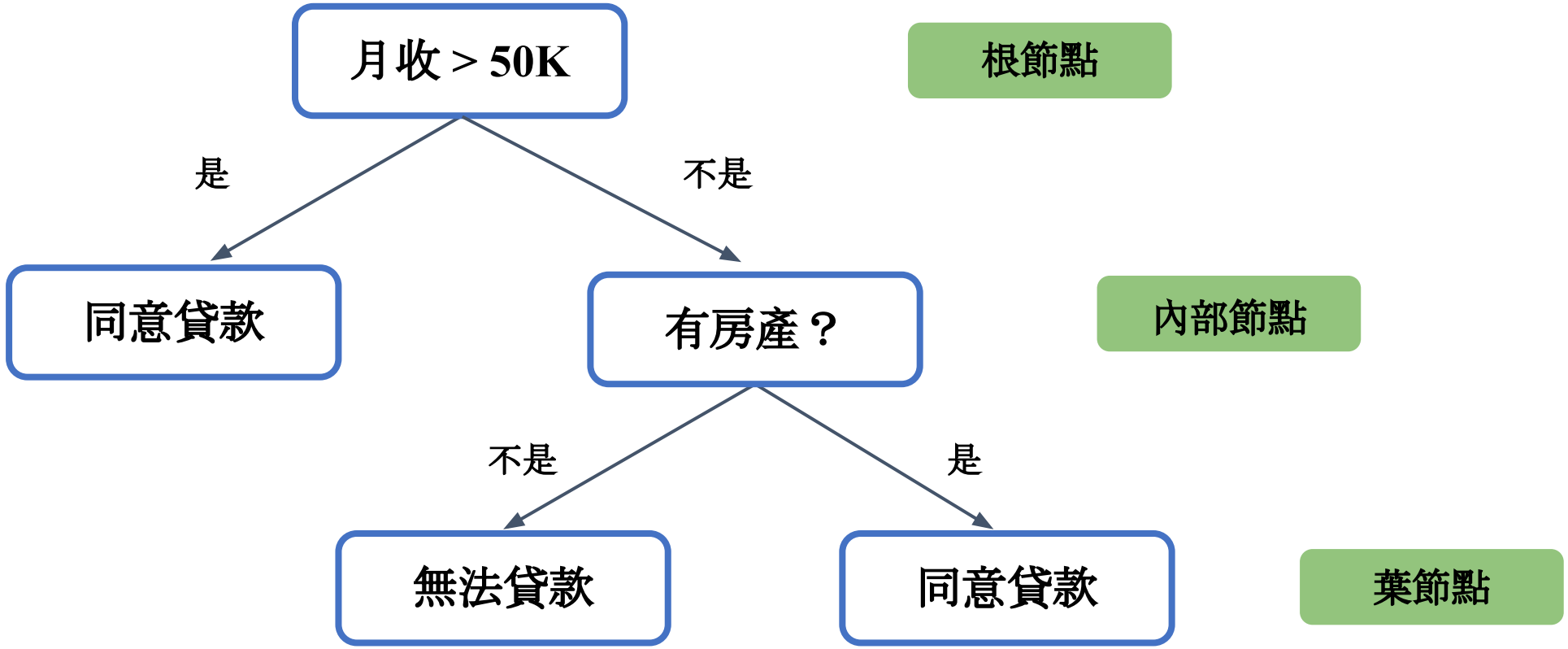
- 單純貝式 (Naive Bayes)
- 邏輯斯回歸 (Logistic Regression)
- 決策樹 (Decision Tree)
- 支援向量機 (Support Vector Machine, SVM)
- K-近鄰演算法 (K Nearest Neighbor, KNN)

分類常見應用

預測**類別**

- 貸款者是否會違約
- 顧客是否會消費
- 就醫者是否患病
- 員工是否傾向跳槽

決策樹



◎ 資料來源:大數據驅動商業決策:13 個RapidMiner 商業預測操作實務

決策樹過程

1-特徵選擇

亂度(Entropy，熵)

- ID3演算法:資訊增益(Information gain, IG)
- C4.5演算法:資訊增益率 (Gain Ratio, GR)
- CART演算法:吉尼不純度 (Gini Impurity)

2-產生決策樹

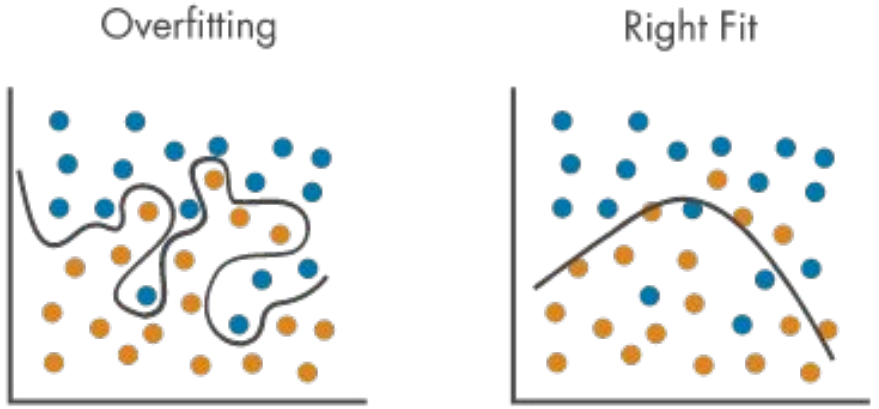
決策樹過度擬合 (Overfitting)

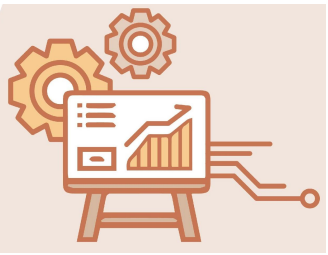
- 最小樣本數 (Minimum Samples (Size) Split)

  內部節點資料筆數最小值

- 最大深度 (Maximum depth)

  決定建立多少層的決策樹



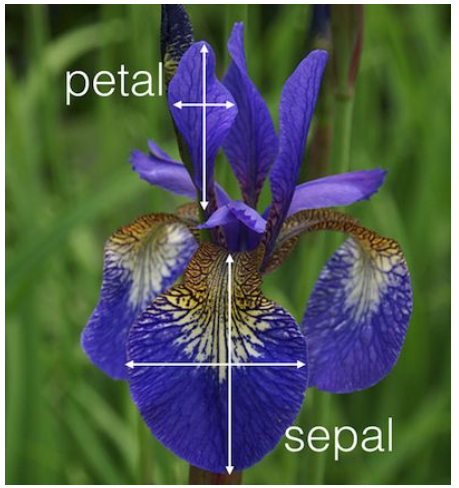


實作 - 資料集

Iris dataset

- 由英國統計學家 Ronald Fisher 爵士 在 1936 年
- 依照山鳶尾、變色鳶尾、維吉尼亞鳶尾三類進行標示
- 特徵選取:花瓣花萼的長寬數據資料
- 採集地點:加斯帕半島上的鳶尾屬花朵

鳶尾花介紹



圖片	名稱	特徵
	Setosa 山鳶尾	1. 耐寒能力很強，喜歡在潮濕的土壤中生長。 2. 極度瀕危植物。 3. 從其根莖中萃取出香水的原料。
	Versicolor 變色鳶尾	1. 美國田納西州的官方州花 2. 被認為其根莖能帶來財運，因此許多人都愛把它放在收銀機中來提高營業額。
	Virginica 維吉尼亞鳶尾	1. 花朵散發淡淡香氣，可以用來製作香水。 2. 觀賞性植物 3. 帶有輕微毒性

講到大數據，一定要分類分析呀！

資料集 - 敘述統計

欄位名稱	資料型態	欄位名稱(中文)
sepal length	Real	花萼長度 (cm)
sepal width	Real	花萼寬度 (cm)
petal length	Real	花瓣長度 (cm)
petal width	Real	花瓣寬度 (cm)
species	Nominal	花卉名稱

▲ 欄位資料型態表

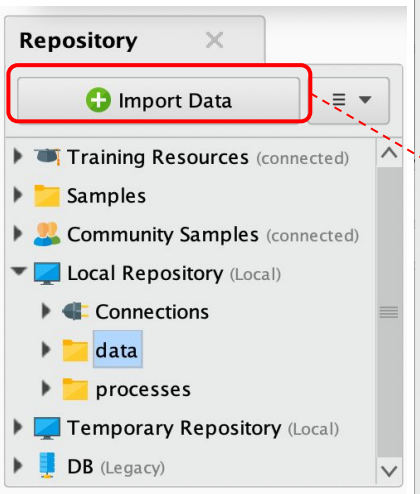
- 資料來源:Kaggle-Iris Species
- 資料筆數:150 筆

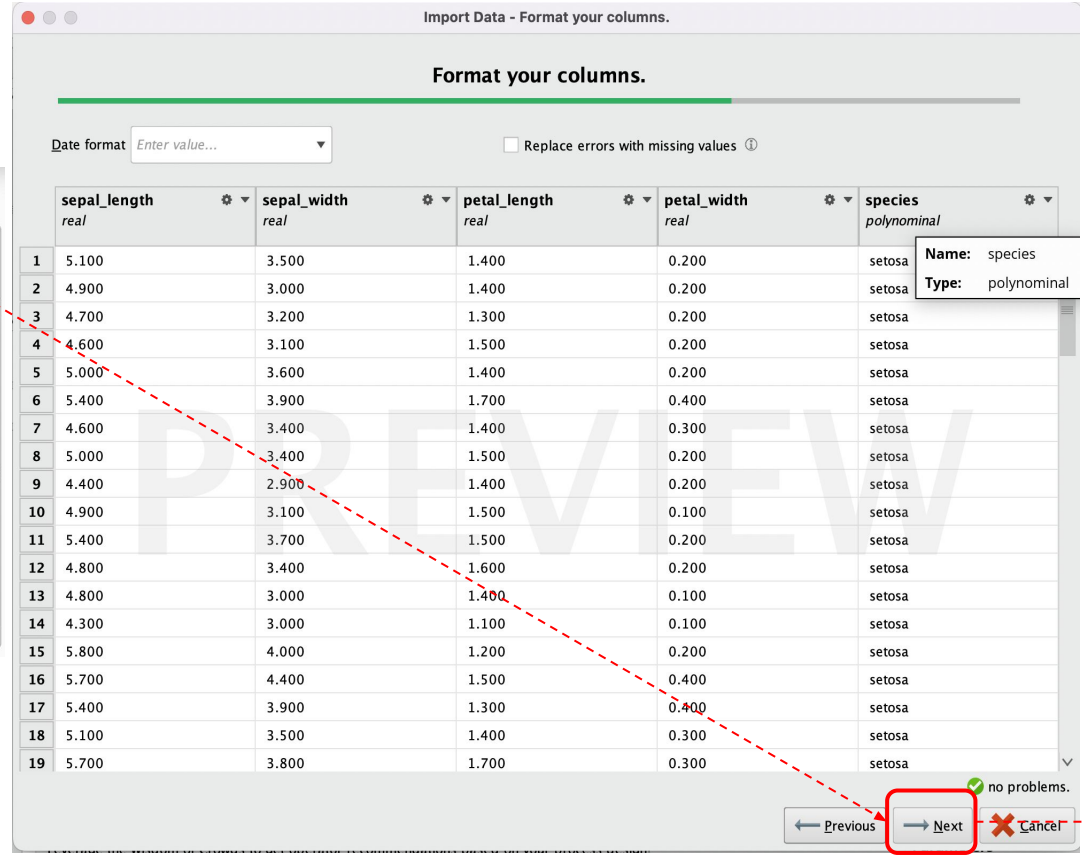
  每種花卉皆為 50 筆
- 欄位數量:5 個
- 預測欄位:species

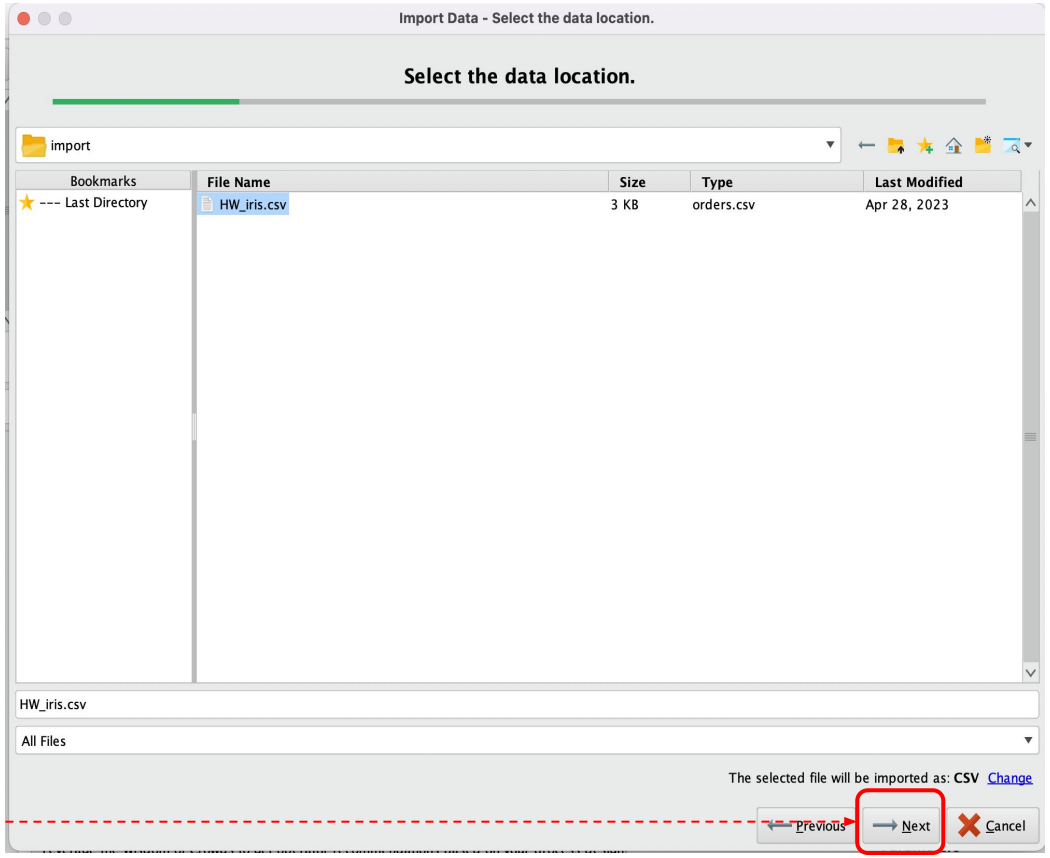
◎ 資料來源:https://www.kaggle.com/datasets/uciml/iris

講到大數據, 一定要分類分析呀！

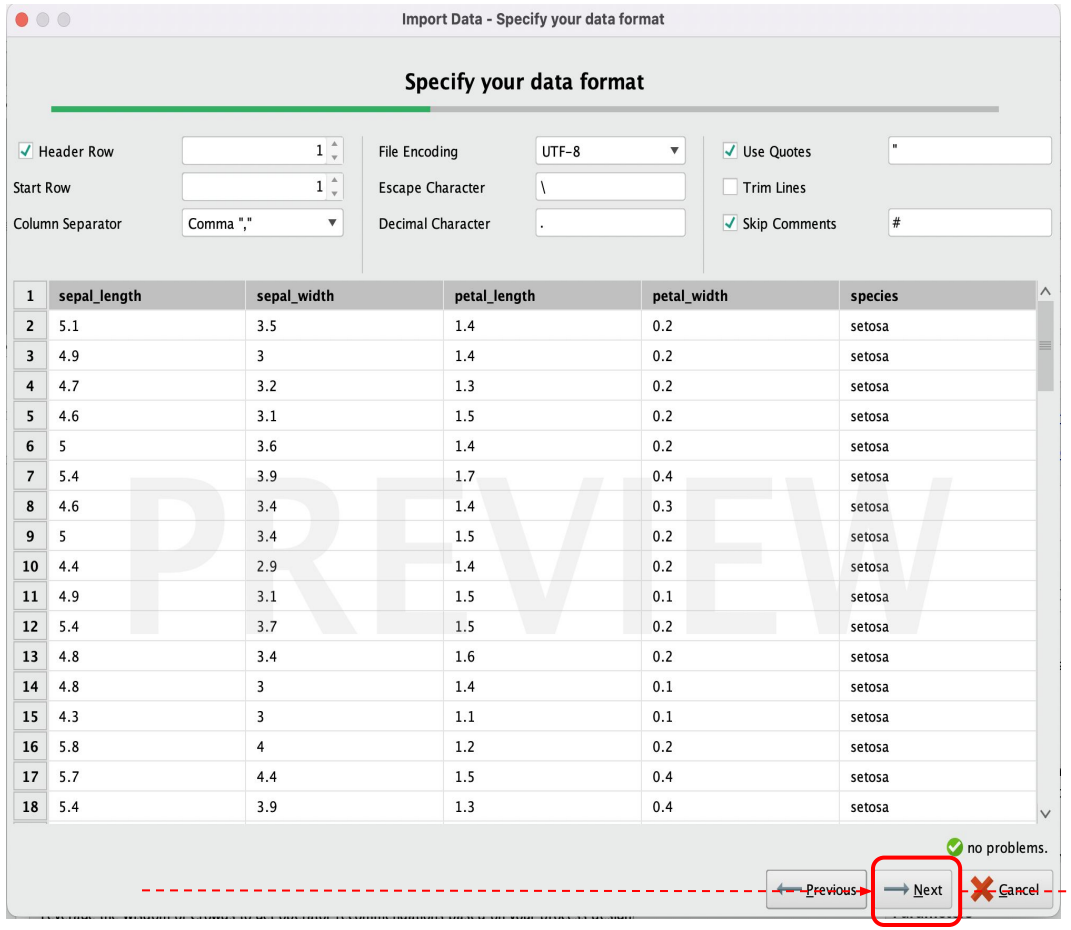
導入資料集

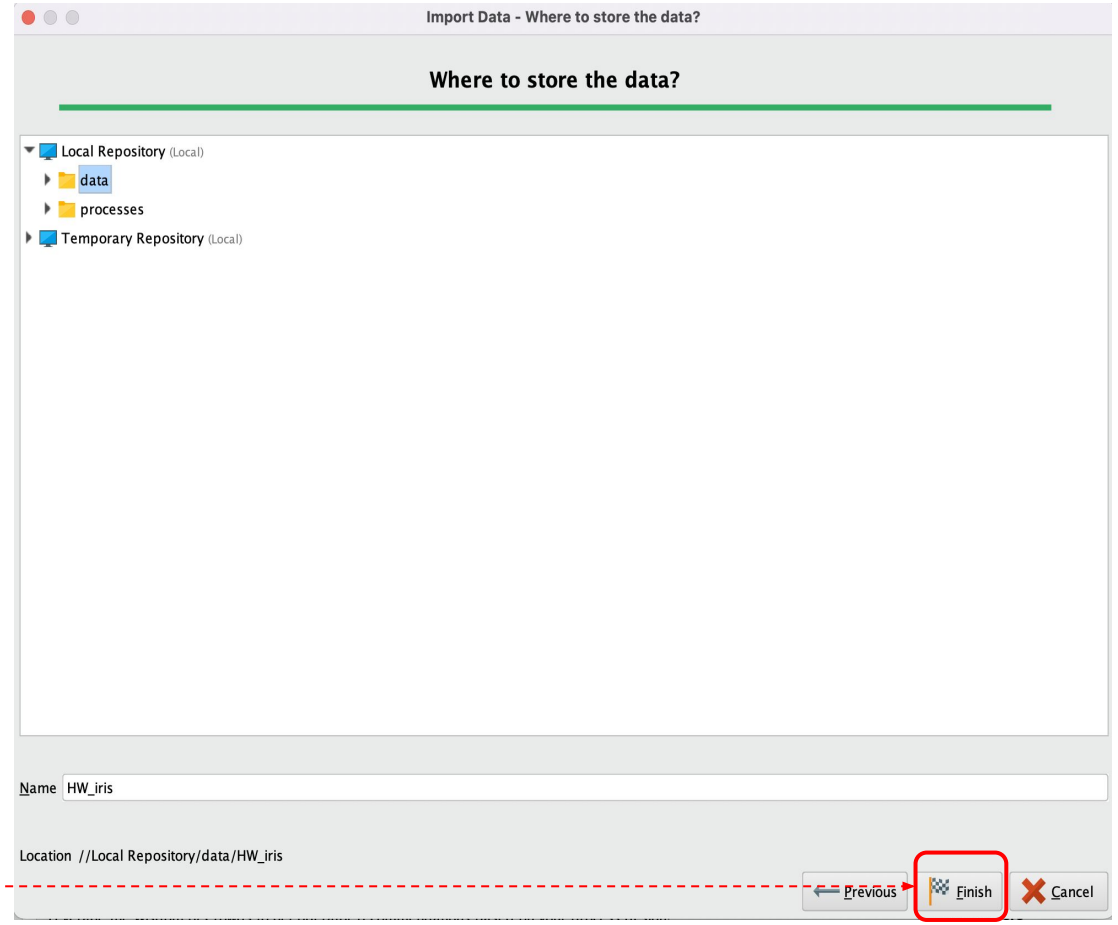

Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
processes
Temporary Repository (Local)
DB (Legacy)


Import Data - Format your columns.
Format your columns.
Date format
Enter value...
Replace errors with missing values
sepal_length real
sepal_width real
petal_length real
petal_width real
species polynominal
Name: species
Type: polynominal
1 5.100 3.500 1.400 0.200 setosa
2 4.900 3.000 1.400 0.200 setosa
3 4.700 3.200 1.300 0.200 setosa
4 4.600 3.100 1.500 0.200 setosa
5 5.000 3.600 1.400 0.200 setosa
6 5.400 3.900 1.700 0.400 setosa
7 4.600 3.400 1.400 0.300 setosa
8 5.000 3.400 1.500 0.200 setosa
9 4.400 2.900 1.400 0.200 setosa
10 4.900 3.100 1.500 0.100 setosa
11 5.400 3.700 1.500 0.200 setosa
12 4.800 3.400 1.600 0.200 setosa
13 4.800 3.000 1.400 0.100 setosa
14 4.300 3.000 1.100 0.100 setosa
15 5.800 4.000 1.200 0.200 setosa
16 5.700 4.400 1.500 0.400 setosa
17 5.400 3.900 1.300 0.400 setosa
18 5.100 3.500 1.400 0.300 setosa
19 5.700 3.800 1.700 0.300 setosa
no problems.
Previous
Next
Cancel
PREVIEW

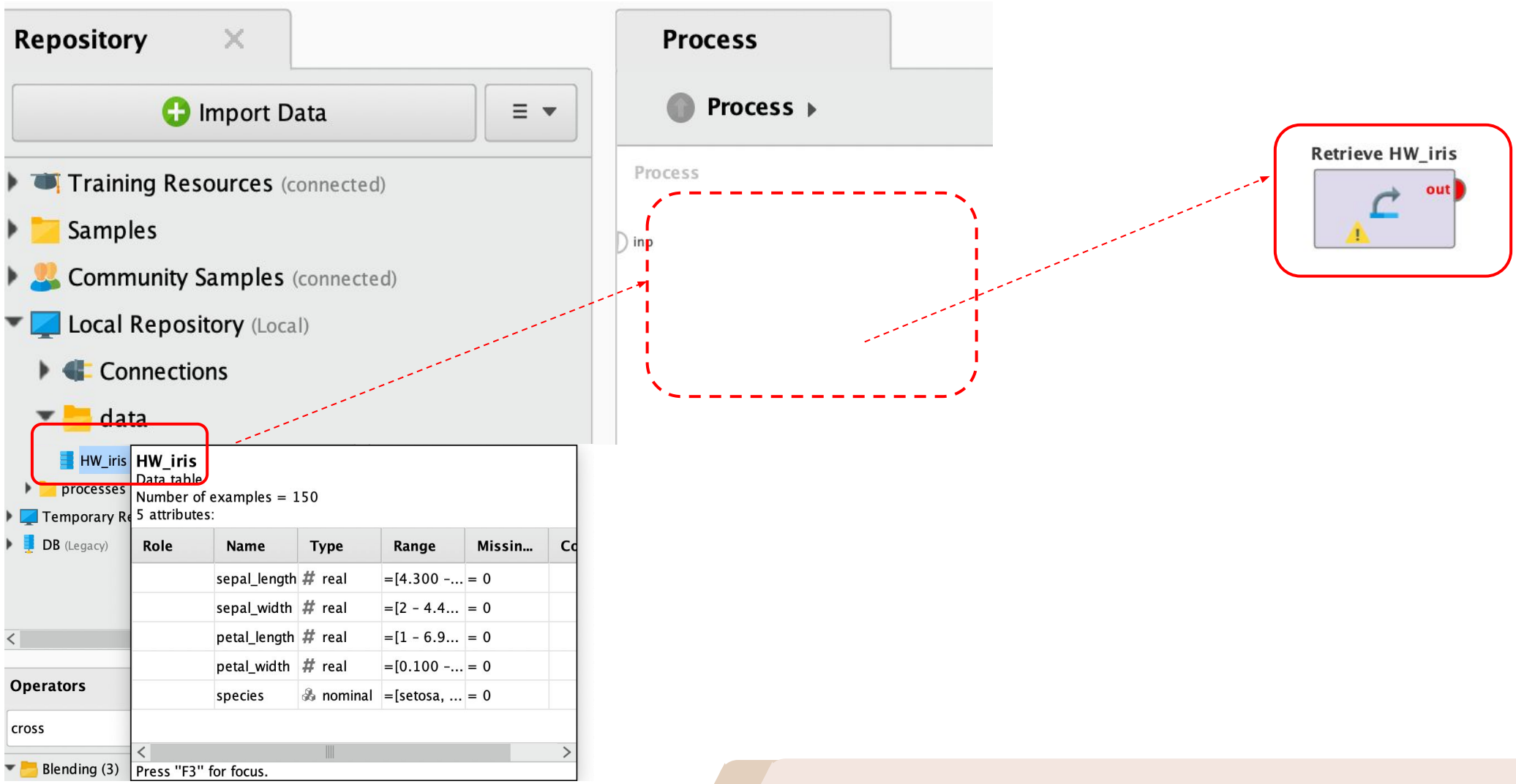

Import Data - Select the data location.
Select the data location.
import
Bookmarks
--- Last Directory
File Name
Size
Type
Last Modified
HW_iris.csv
3 KB
orders.csv
Apr 28, 2023
HW_iris.csv
All Files
The selected file will be imported as: CSV Change
Previous
Next
Cancel

導入資料集

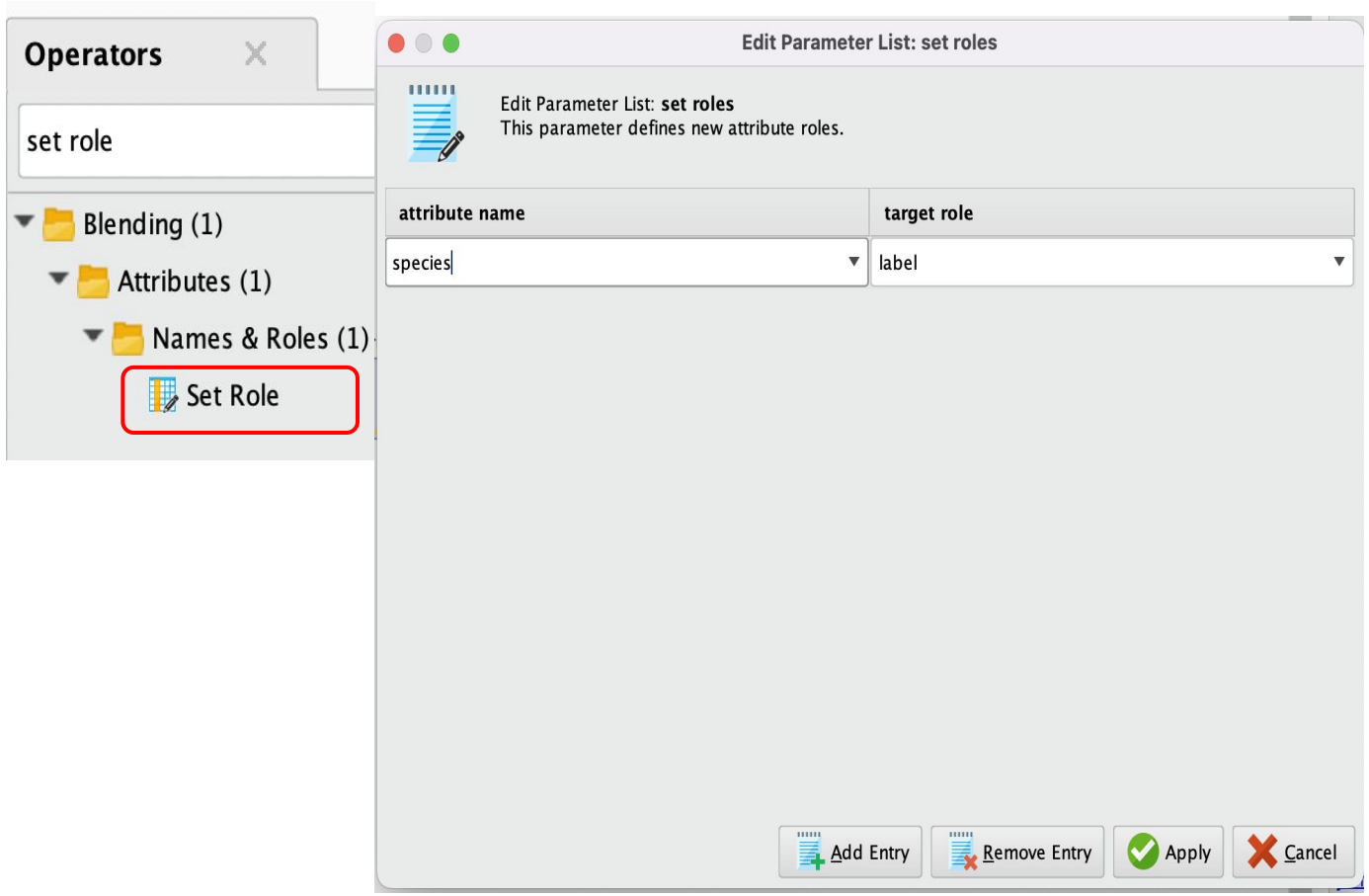

Import Data - Specify your data format
Specify your data format
Header Row
1
File Encoding
UTF-8
Use Quotes
"
Start Row
1
Escape Character
\
Trim Lines
Column Separator
Comma ","
Decimal Character
.
Skip Comments
#
1 sepal_length sepal_width petal_length petal_width species
2 5.1 3.5 1.4 0.2 setosa
3 4.9 3 1.4 0.2 setosa
4 4.7 3.2 1.3 0.2 setosa
5 4.6 3.1 1.5 0.2 setosa
6 5 3.6 1.4 0.2 setosa
7 5.4 3.9 1.7 0.4 setosa
8 4.6 3.4 1.4 0.3 setosa
9 5 3.4 1.5 0.2 setosa
10 4.4 2.9 1.4 0.2 setosa
11 4.9 3.1 1.5 0.1 setosa
12 5.4 3.7 1.5 0.2 setosa
13 4.8 3.4 1.6 0.2 setosa
14 4.8 3 1.4 0.1 setosa
15 4.3 3 1.1 0.1 setosa
16 5.8 4 1.2 0.2 setosa
17 5.7 4.4 1.5 0.4 setosa
18 5.4 3.9 1.3 0.4 setosa
PREVIEW
no problems.
Previous
Next
Cancel

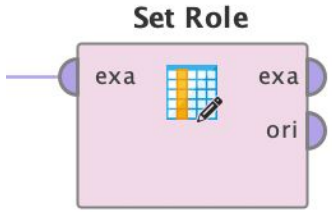

Import Data - Where to store the data?
Where to store the data?
Local Repository (Local)
data
processes
Temporary Repository (Local)
Name HW_iris
Location //Local Repository/data/HW_iris
Previous
Finish
Cancel

導入資料集


Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Connections
data
HW_iris
processes
Temporary Re
DB (Legacy)
Operators
cross
Blending (3)
HW_iris
Data table
Number of examples = 150
5 attributes:
Role | Name | Type | Range | Missin... | Co
sepal_length | real | =[4.300 -... | = 0
sepal_width | real | =[2 - 4.4... | = 0
petal_length | real | =[1 - 6.9... | = 0
petal_width | real | =[0.100 -... | = 0
species | nominal | =[setosa, ... | = 0
Press "F3" for focus.
Process
Process
Process
inp
Retrieve HW_iris
out

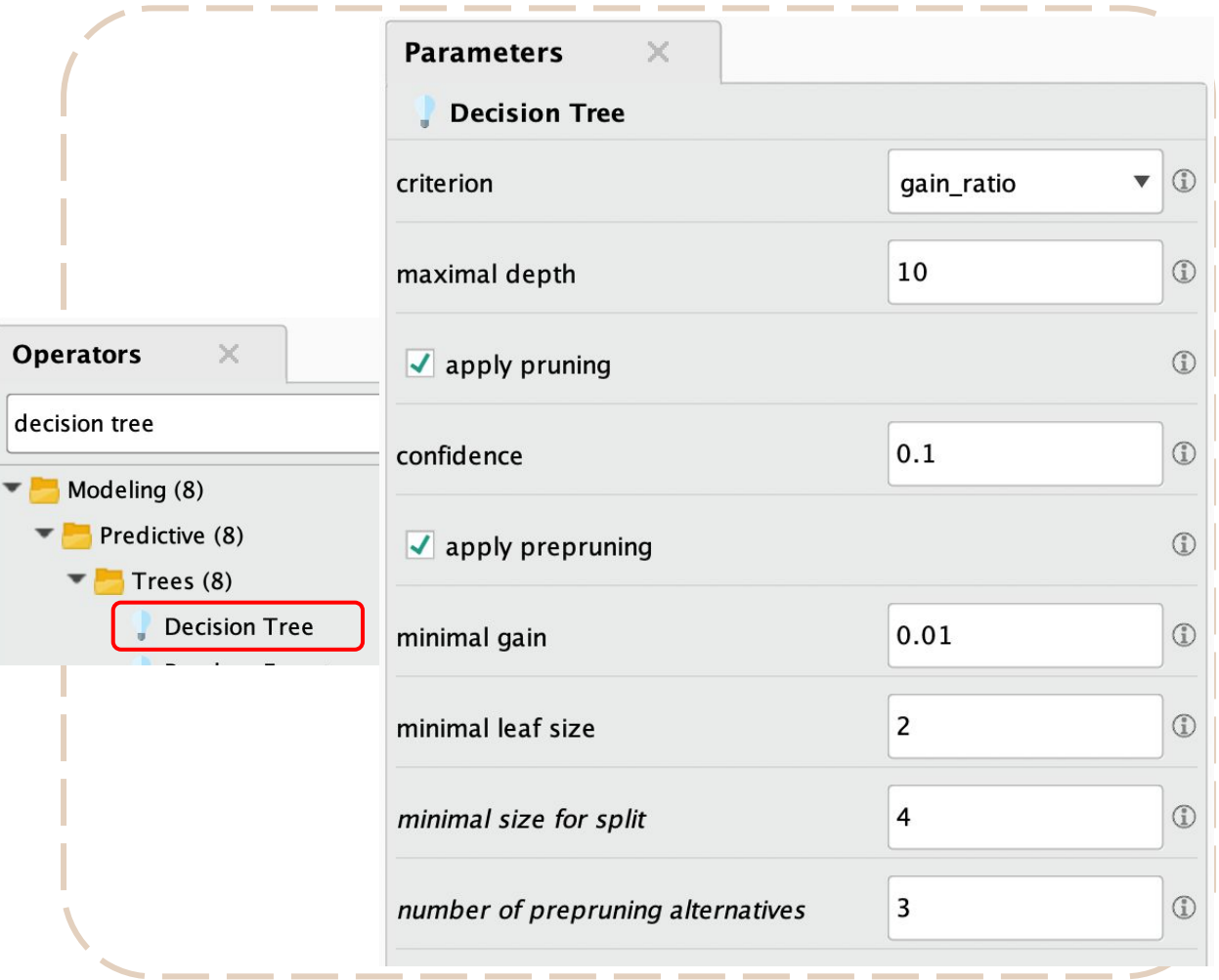
設定目標欄位



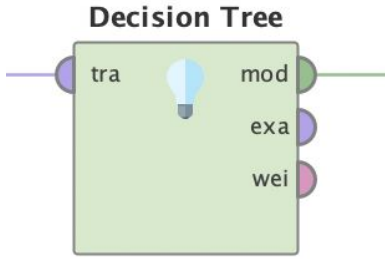


- **Operator**：Set Role
- **位置**：讀取資料之後
- **參數設置**：將「species」設為 label (Label 值為預測欄位)

建立模型 - 決策樹

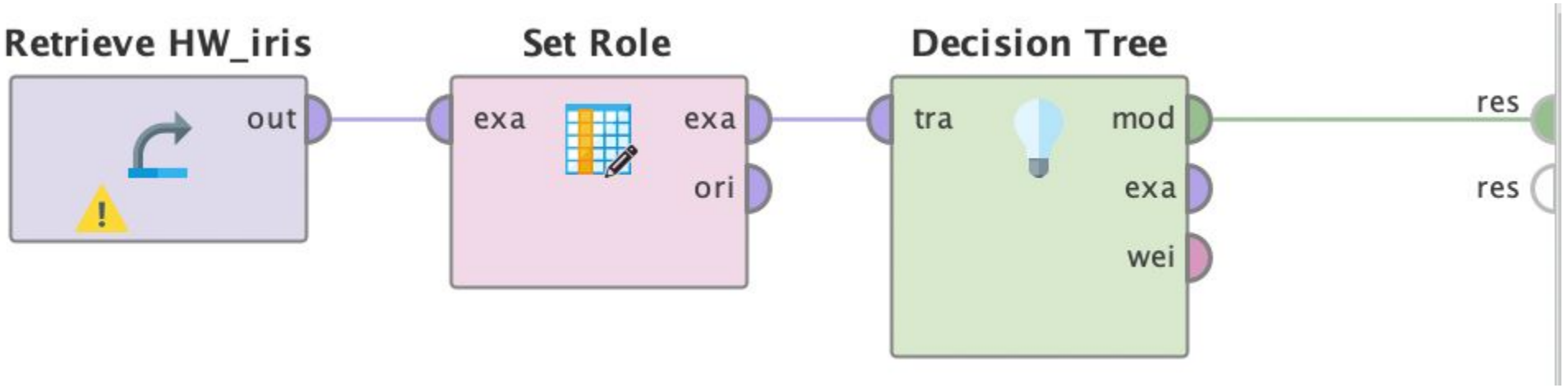


▲ Decision Tree



- **Operator**：Decision Tree
- **位置**：Set Role 後面
- **參數設置**：
  - 決策樹深度：10
  - Confidence：0.1
  - Minimal leaf size：2

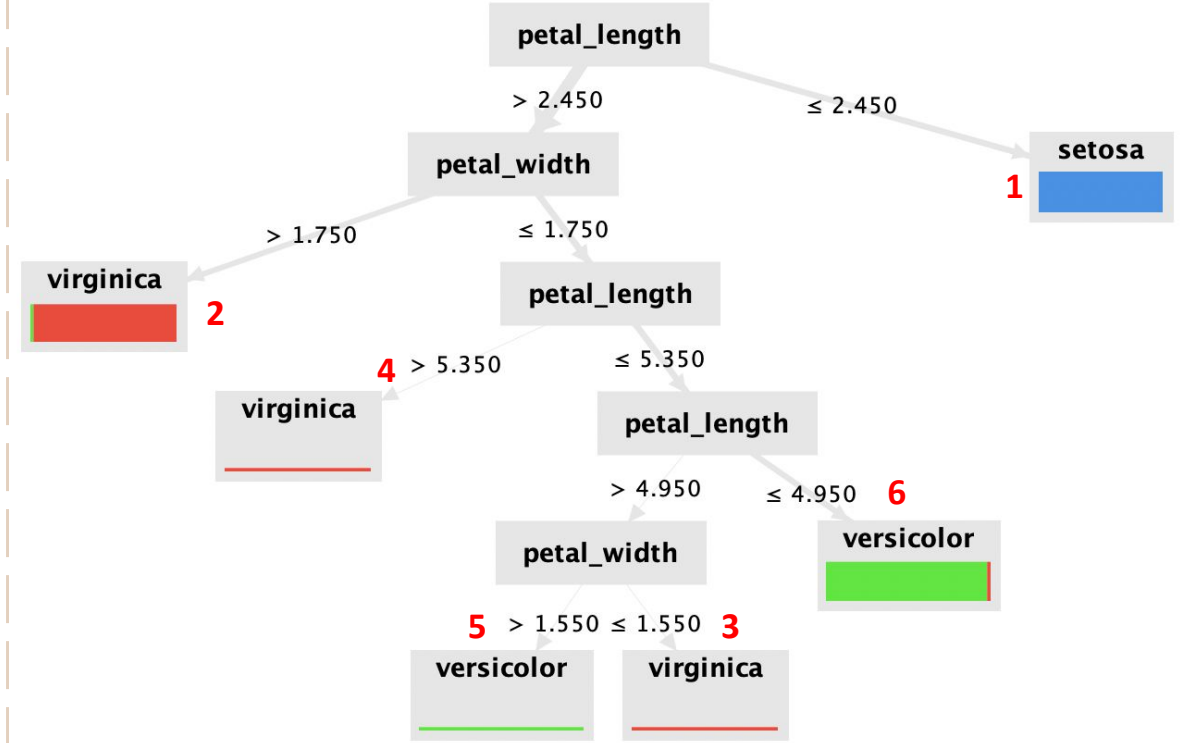
講到大數據，一定要分類分析呀！


Retrieve HW_iris
out
Set Role
exa
exa
ori
Decision Tree
tra
mod
exa
wei
res
res

▲ 決策樹流程圖

連結流程

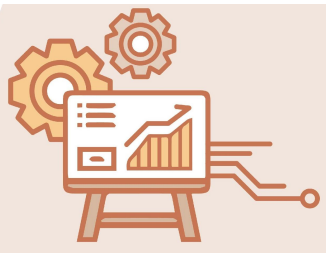
執行流程



▲ (Tree)Decision Tree

- **setosa 山鳶尾**
  - 1. 花瓣長度 0-2.45，花瓣寬度 0-∞
- **virginica 維吉尼亞鳶尾**
  - 2. 花瓣長度2.46-∞，花瓣寬度 1.76-∞
  - 3. 花瓣長度4.96-5.35，花瓣寬度 0-1.55
  - 4. 花瓣長度5.36-∞，花瓣寬度 0-1.75
- **versicolor 變色鳶尾**
  - 5. 花瓣長度4.96-5.35，花瓣寬度 1.56-1.75
  - 6. 花瓣長度2.46-4.95，花瓣寬度 0-1.75

講到大數據，一定要分類分析呀！



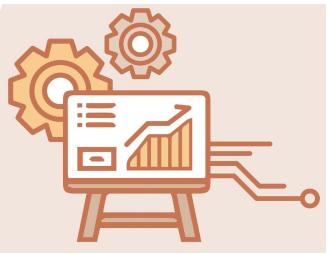
評估指標

混餚矩陣	Positive	Negative
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

準確率 (Accuracy) $= \frac{TP+TN}{Total}$

精確率 (Precision) $= \frac{TP}{TP+FP}$

召回率 (Recall) $= \frac{TP}{TP+FN}$



評估指標

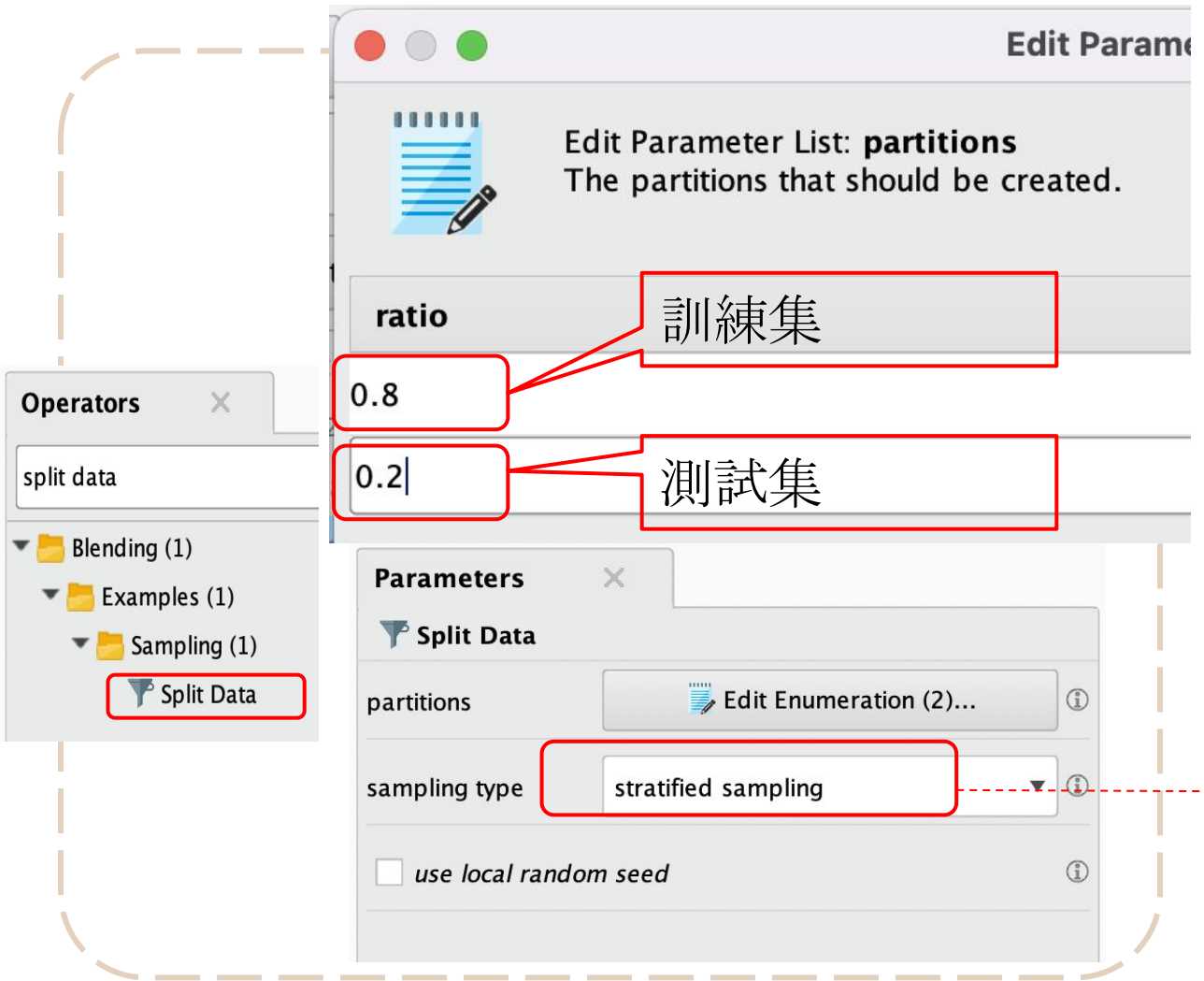
混餚矩陣		預測的類別	
		貓	狗
實際的類別	貓	5	3
	狗	2	3

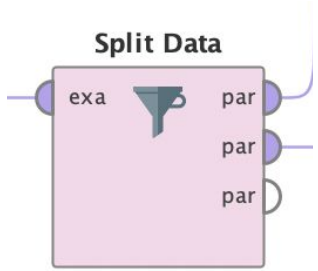
準確率 (Accuracy) $\frac{5+3}{13} = 61.5\%$

精確率 (Precision) $\frac{5}{5+3} = 62.5\%$

召回率 (Recall) $\frac{5}{5+2} = 71.4\%$

切割資料



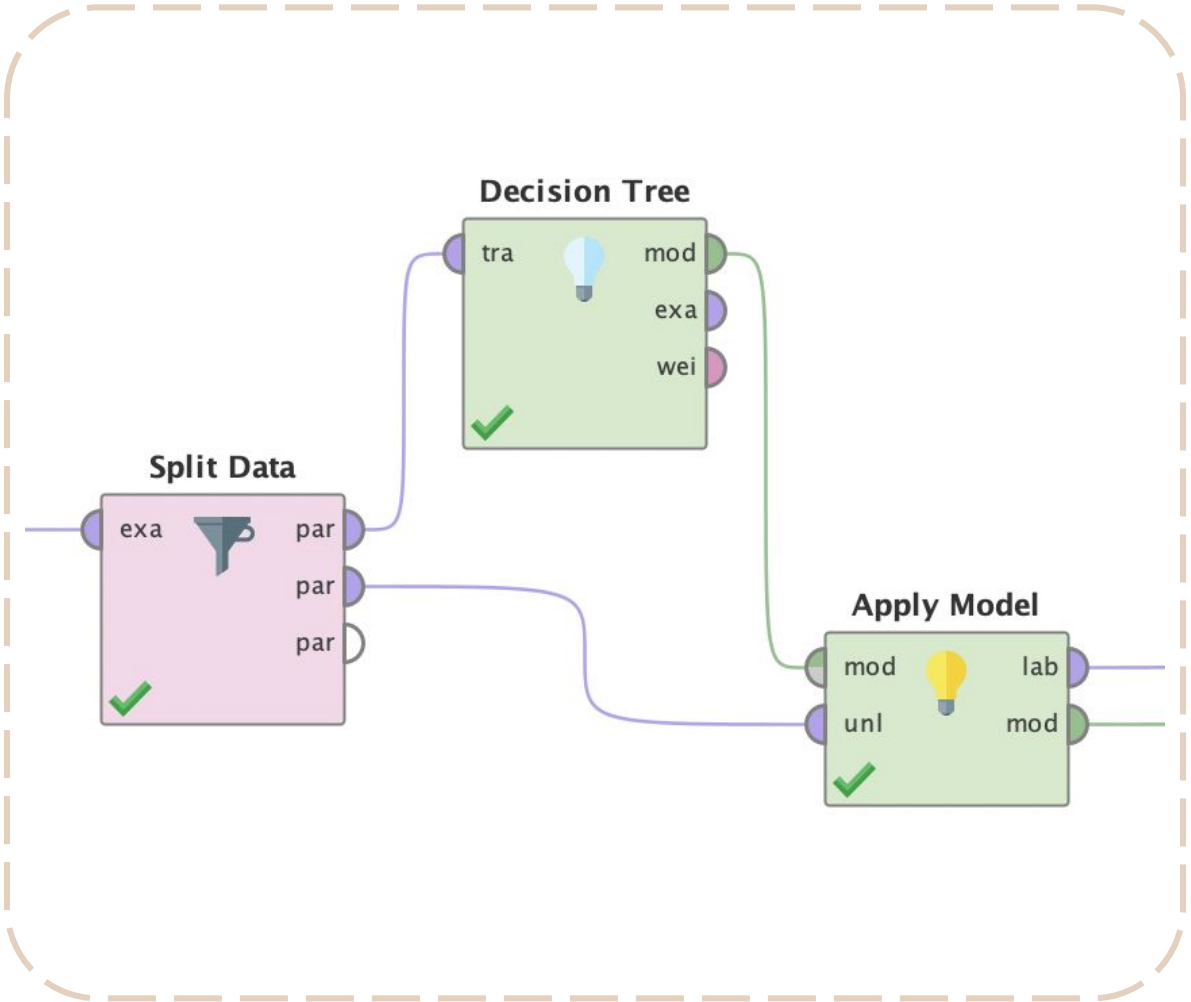


- **Operator**：Split Data
- **位置**：Set Role 與 Decision Tree 之間
- **參數設置**：
  - partitions：包括訓練集和測試集
  - sampling type：stratified sampling

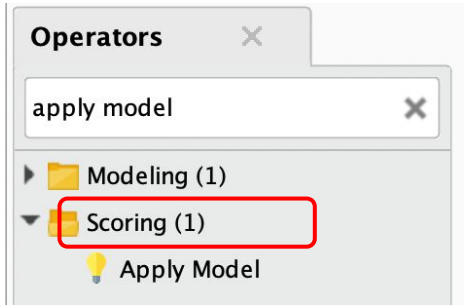
▲ Split Data

講到大數據，一定要分類分析呀！

建立應用模型分支



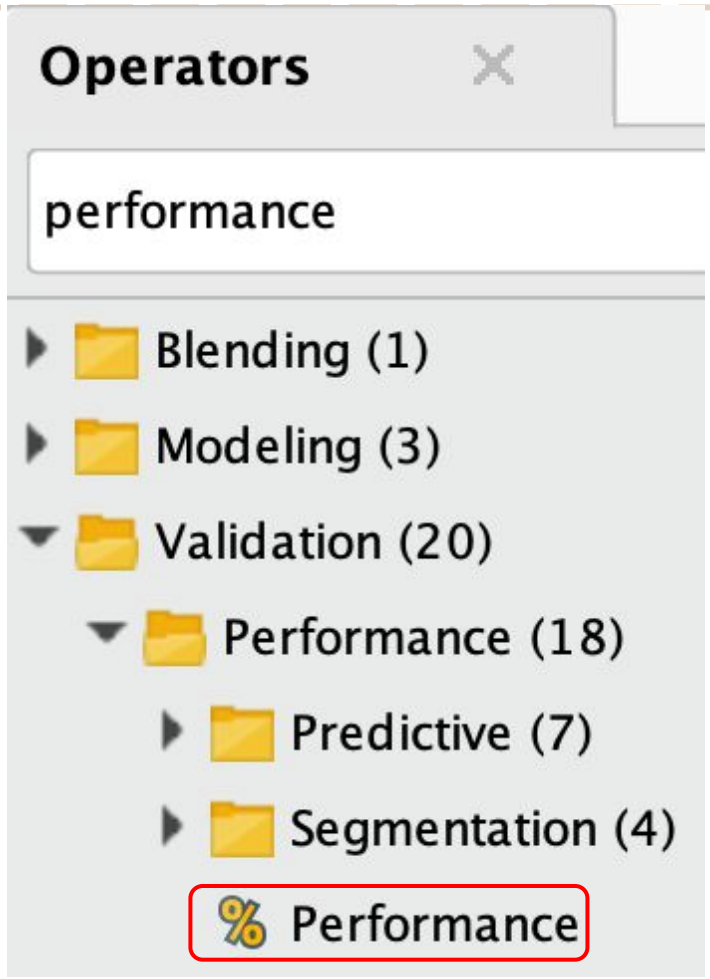
▲ Apply Model



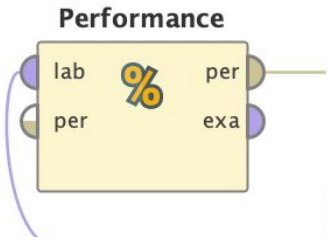
- **Operator**：Apply Model
- **連接方式**：
  - mod：Decision Tree 的訓練集
  - unl：Split Data 的測試集

講到大數據，一定要分類分析呀！

建立評估模型

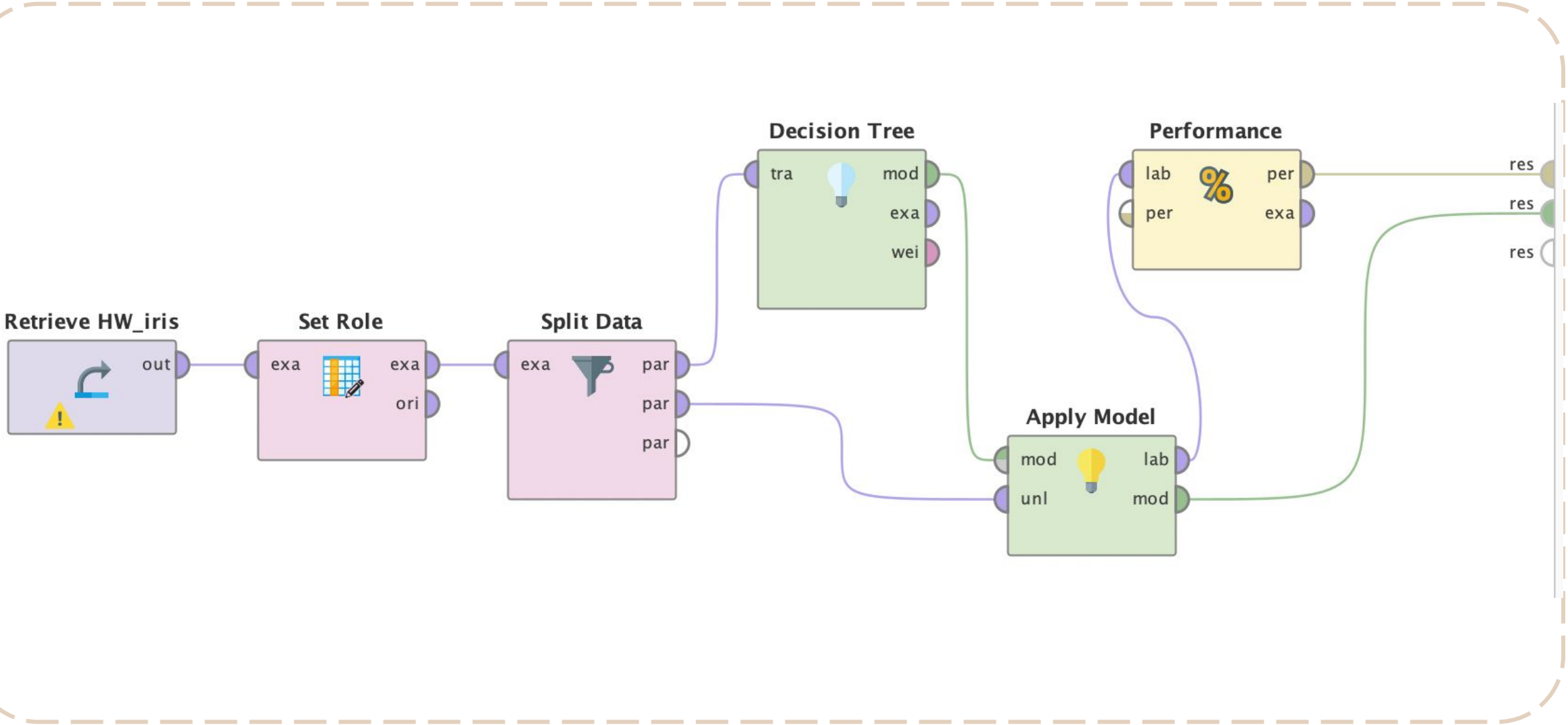


▲ Performance



- **Operator**：Performance
- **位置**：Apply Model 的lab連結lab

講到大數據，一定要分類分析呀！

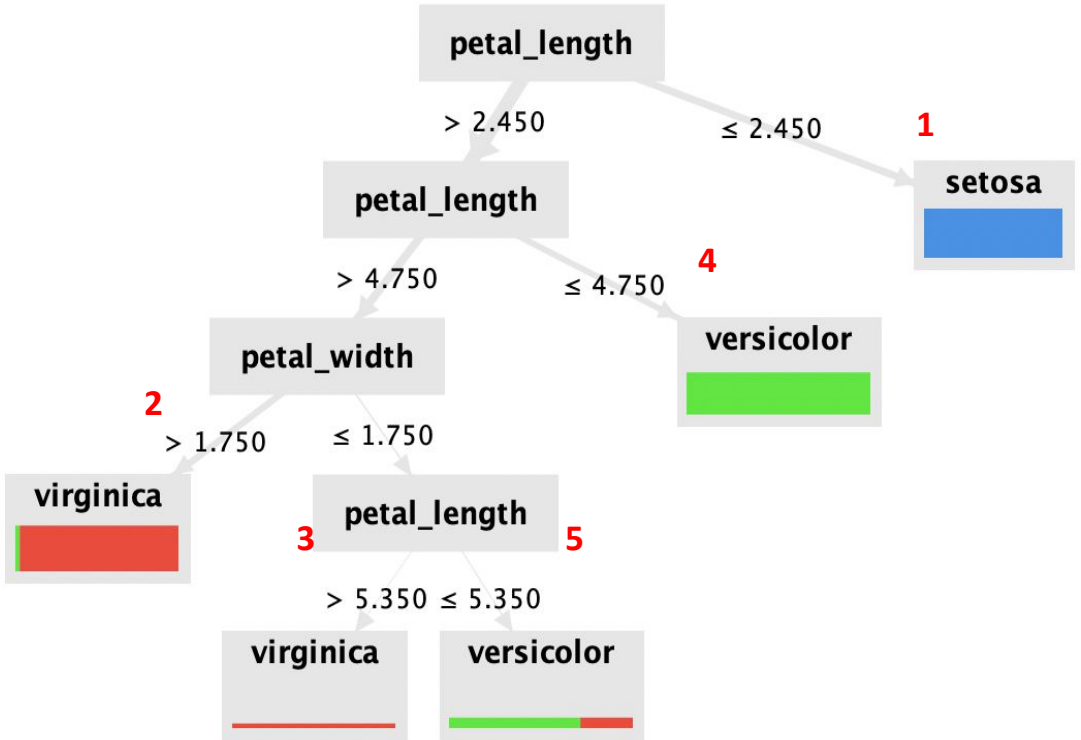


▲ 決策樹流程圖

連結流程

講到大數據，一定要分類分析呀！

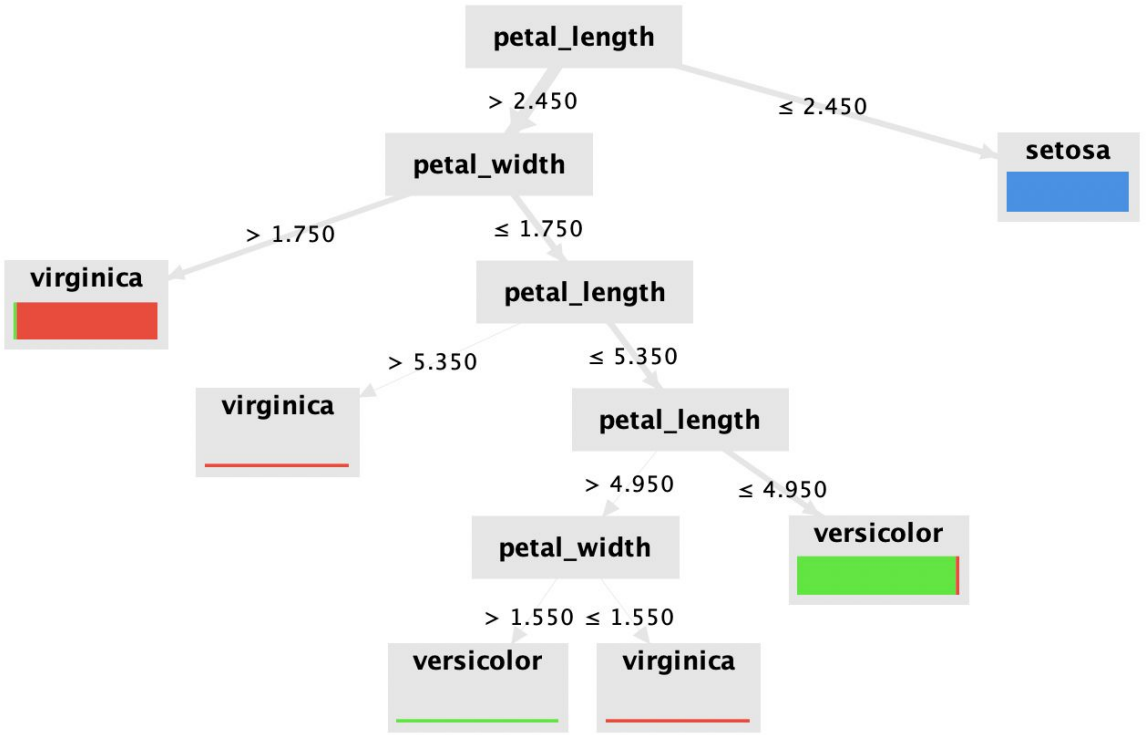
執行流程



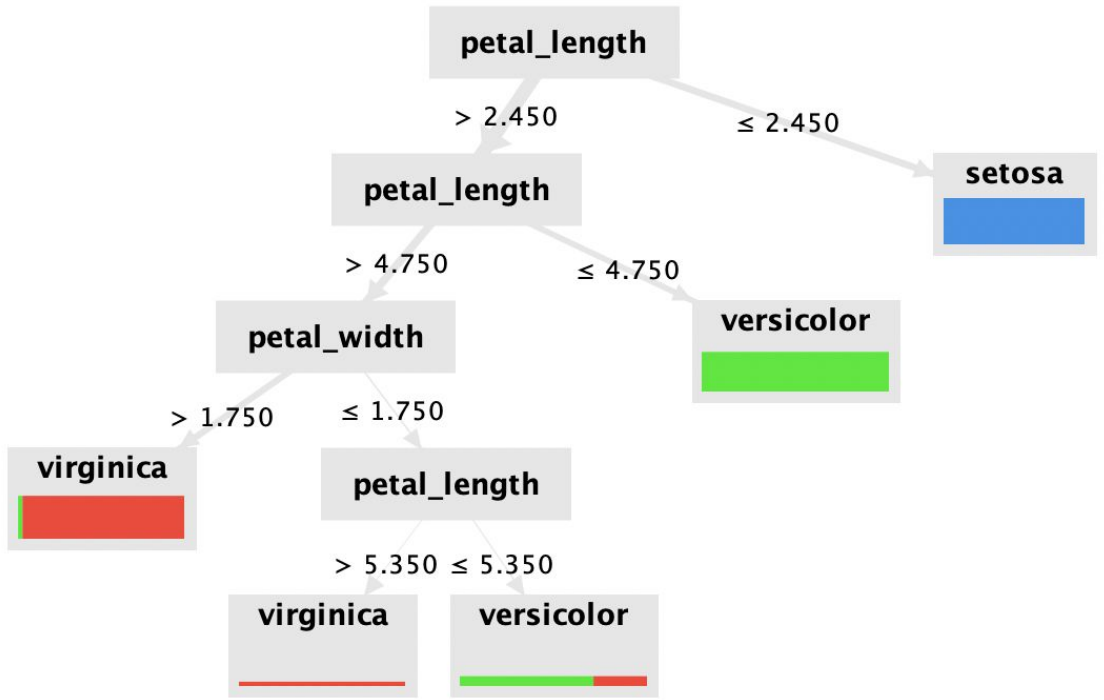
▲ Decision Tree

- **setosa 山鳶尾**
  - 1. 花瓣長度 0-2.45, 花瓣寬度 0-∞
- **virginica 維吉尼亞鳶尾**
  - 2. 花瓣長度 4.76-∞, 花瓣寬度 1.76-∞
  - 3. 花瓣長度 5.36-∞, 花瓣寬度 0-1.75
- **versicolor 變色鳶尾**
  - 4. 花瓣長度 2.46-4.75, 花瓣寬度 0-∞
  - 5. 花瓣長度 4.76-5.35, 花瓣寬度 0-1.75

講到大數據, 一定要分類分析呀！



▲ 100%數據



▲ 80% Stratified Sampling數據

決策樹分別

講到大數據，一定要分類分析呀！

評估模型

accuracy: 96.67%

	true setosa	true versicolor	true virginica	class precision
pred. setosa	10	0	0	100.00%
pred. versicolor	0	10	1	90.91%
pred. virginica	0	0	9	100.00%
class recall	100.00%	100.00%	90.00%	

- 混淆矩陣3x3：用來表示分類器的分類結果和真實標籤之間的關係
- 模型的準確度為 96.67%

▲ Correlation Matrix

回顧

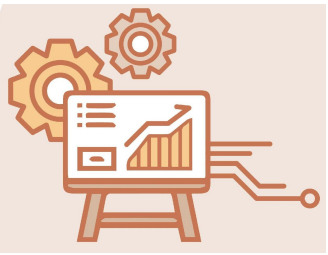
1 大數據分類是將數據. . . ？

A 按照特定標準劃分為多個類別

B 儲存在多個資料庫中

C 用於製作大型統計報告

D 轉換為有意義的信息



回顧

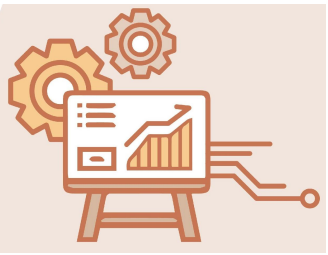
1 大數據分類是將數據. . . ？

A 按照特定標準劃分為多個類別

回顧

2 大數據分類的目的是什麼 ？

- **A** 減少數據儲存成本
- **B** 清理數據
- **C** 為未來的決策提供準確的數據
- **D** 創建數據



回顧

2 大數據分類的目的是什麼 ？

C 為未來的決策提供準確的數據

回顧

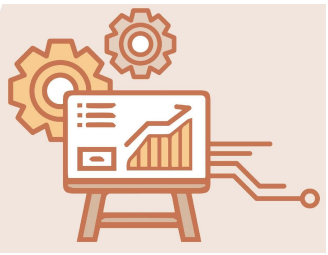
3 決策樹是一種什麼樣的演算法？

A 監督式學習

B 非監督式學習

C 強化學習

D 前饋神經網路



回顧

3 決策樹是一種什麼樣的演算法 ？

A 監督式學習