

# 計算思維與人工智慧

## TA Class #04

RapidMiner1

主講者: 程至榮



政大  
NATIONAL CHENGCHI UNIVERSITY



政大資訊科學系  
Department of Computer Science, National Chengchi University

# 參考書目

大數據驅動商業決策:13 個 RapidMiner 商業預測操作實務

# 基本介紹

# RapidMiner 是什麼？

RapidMiner 是一個做資料科學的平台，使用者不需要寫任何程式碼就能完成機器學習的分析預測；可透過模組化的操作進行前置資料準備、機器學習的建模、評估及驗證，有超過 1500 種功能及模型。(參考自 [RapidMiner 臺灣總代理昊青網站](#))



RAPIDMINER

# 註冊

請記得勾選 Educational purposes  
獲取教育版本一年的使用權限

RAPIDMINER [My Account](#) [Downloads](#) [Sign in](#) | [Register](#)

## Create your RapidMiner Account

This account gives you access to RapidMiner products (trials, licenses, updates, and extensions), training via the Academy, and the RapidMiner Community.

What are you using Rapidminer for?

☐ Commercial purposes (e.g., business, evaluation, not-for-profit)

☒ Educational purposes (e.g., educator, student)

First name:  
Zhi Rong

Last name:  
Cheng

Country:  
Taiwan

University:  
National Chengchi University (NCCU)

Role:  
Student

Email address:  
jordan990301@gmail.com

Create a password:

Confirm your password:


Register

Already have an account? [Sign in here.](#)

<https://my.rapidminer.com/nexus/account/index.html#signup>

# 註冊

## 註冊成功、登入後檢查使用權限

 **RAPIDMINER**

[My Account](#) [Profile](#) [Downloads](#) [Licenses](#)

[Sign out](#)

### Licenses

View license keys for your RapidMiner products.

[Studio](#) [AI Hub/Server](#) [Scoring Agent](#) [Radoop](#) [Go \(Legacy\)](#)

▼ RapidMiner Studio 7.2+

<b>RapidMiner Studio</b>	Educational	Zhi Rong Cheng	Expires <b>Fri, Aug 2nd 2024 (a year left)</b>	<a href="#">View License Key</a>
<b>RapidMiner Studio</b>	Free	Zhi Rong Cheng	Never Expires 10,000 rows limit	<a href="#">View License Key</a>

<https://my.rapidminer.com/nexus/account/index.html#licenses/rapidminer-studio>

# 安裝

上課使用的軟體版本為 9.10(與[參考書目](#)相同版本), 不同版本的程式元件會有差異。

先登入你的帳號 -> 前往[這個連結](#)-> 網頁拉到最下面 -> 找到以下圖示、點擊它



← 點擊選擇版本

或是使用[這個雲端連結](#)的檔案進行安裝

# 安裝

根據你的作業系統下載安裝檔

Version 9.10

← 顯示目前選擇的版本

The following download URLs are valid for 24 hours.

## RapidMiner Studio 9.10

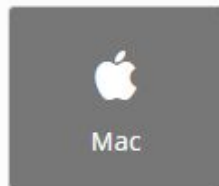
Click on your operating system to start the download:



32bit



64bit



Requires: Java 8

- [Installation Guide](#)
- [Getting Started Tutorials](#)
- [Support](#)
- [Download Source](#)



# 軟體介面

The screenshot displays the RapidMiner Studio Educational 9.10.013 interface. The main window is titled 'Welcome to RapidMiner Studio!' and features a 'Start with' dialog box. The dialog box is divided into two sections: 'Start with' and 'Choose a template to start from'.

**Start with section:**

- Blank Process:** Start a new process from scratch in the design view. (This option is highlighted with a red rectangle and the text '開啟空白 Process' in red).
- Turbo Prep:** Prepare your data interactively; transform, clean and combine data sets.
- Auto Model:** Build and optimize models using automated machine learning.

**Choose a template to start from section:**

- Churn Modeling:** Predict which of your customers will churn and why with a decision tree.
- Direct Marketing:** Predict response to campaigns and increase the conversion rate of your campaign.
- Credit Risk Modeling:** Model credit default risk by training an optimized Support Vector Machine (SVM) model.
- Market Basket Analysis:** Find products frequently purchased together and turn them into rules for recommendations.
- Predictive Maintenance:** Model equipment failures to schedule maintenance pre-emptively.
- Price Risk Clustering:** Cluster price developments using X-Means to unveil price-risk-relationships.
- Lift Chart:** Create a lift chart to visualize the improvement that a model provides compared to guessing.
- Operationalization:** Embed predictive models into business processes to trigger the right actions automatically.
- Outlier Detection:** Detect anomalies in data resulting from a chemical analysis of wines.
- Geographic Distances:**
- Medical Fraud Detection:**
- Web Analytics:**

**Repository panel (left):**

- Import Data
- Training Resources (connected)
- Community Samples (connected)
- Samples
- Local Repository (Local)
- DB (Legacy)

**Operators panel (bottom left):**

- Search for Operators
- Data Access (63)
- Blending (82)
- Cleansing (28)
- Modeling (167)
- Scoring (14)
- Validation (30)
- Utility (85)
- Extensions (2)

**Parameters panel (right):**

- logverbosity: init
- logfile: [empty]
- resultfile: [empty]
- random seed: 2001
- send mail: never
- encoding: SYSTEM

**Recommended Operators (bottom):**

- Retrieve (12%)
- Select Attributes (6%)
- Set Role (5%)
- Apply Model (4%)
- Filter Examples (4%)

**Footer:**

- Get more operators from the Marketplace
- Hide advanced parameters
- Change compatibility (9.10.013)

**執行** 在 Process 和執行結果畫面之間切換

File Edit Process View Connections Settings Extensions Help

Views Design Results Turbo Prep Auto Model Deployments

Repository

Import Data

Training Resources (connected)  
Community Samples (connected)  
Samples  
Local Repository (Local)  
DB (Legacy)

**資料元件**

Operators

Search for Operators

Data Access (63)  
Blending (82)  
Cleansing (28)  
Modeling (167)  
Scoring (14)  
Validation (30)  
Utility (85)  
Extensions (2)

**運算元件**

Process

Process

Find data, operators, etc. All Studio

Parameters

Process

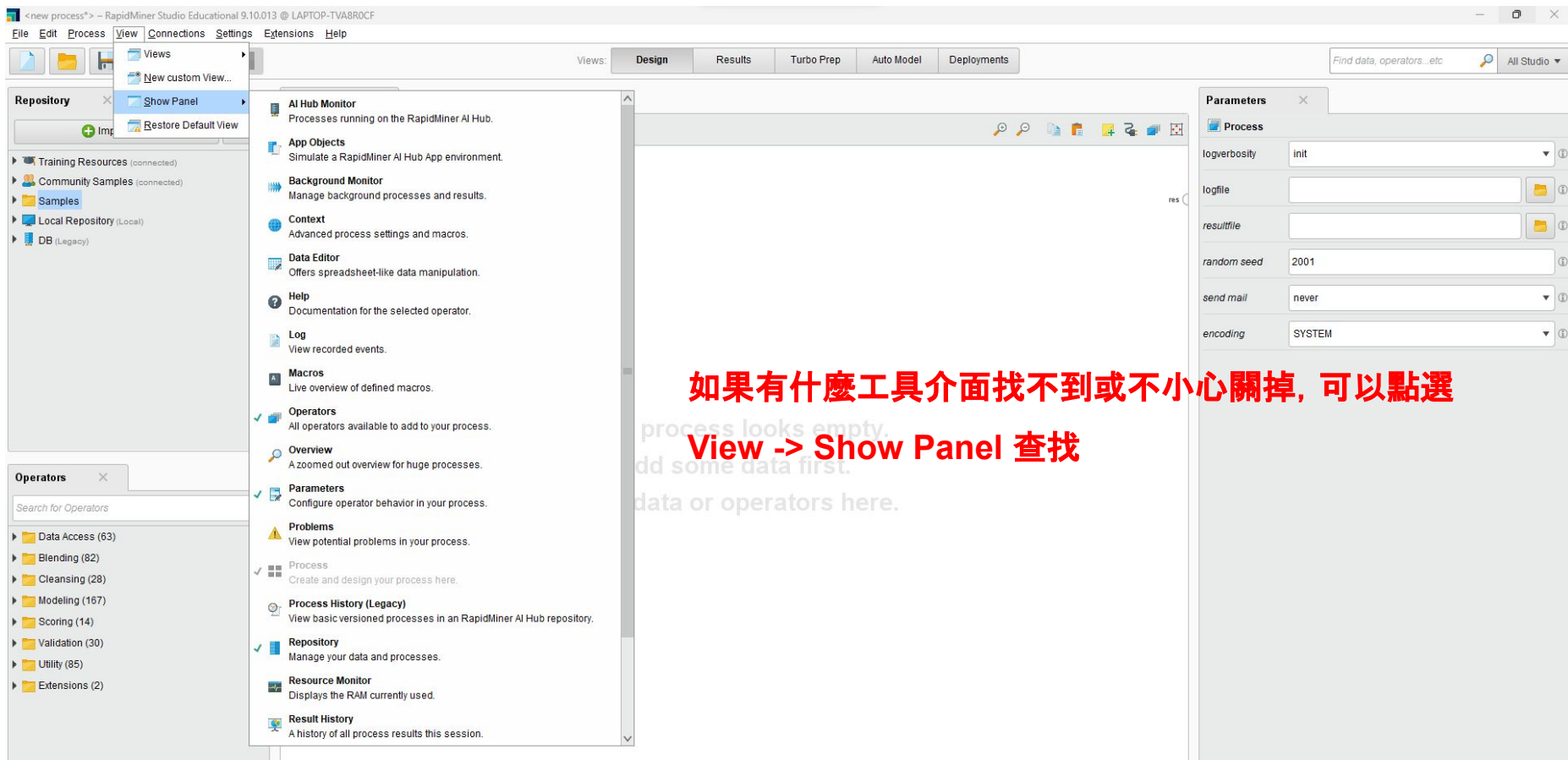
logverbosity: Init  
logfile:  
resultfile:  
random seed: 2001  
send mail: never  
encoding: SYSTEM

**點擊元件-> 調整參數**

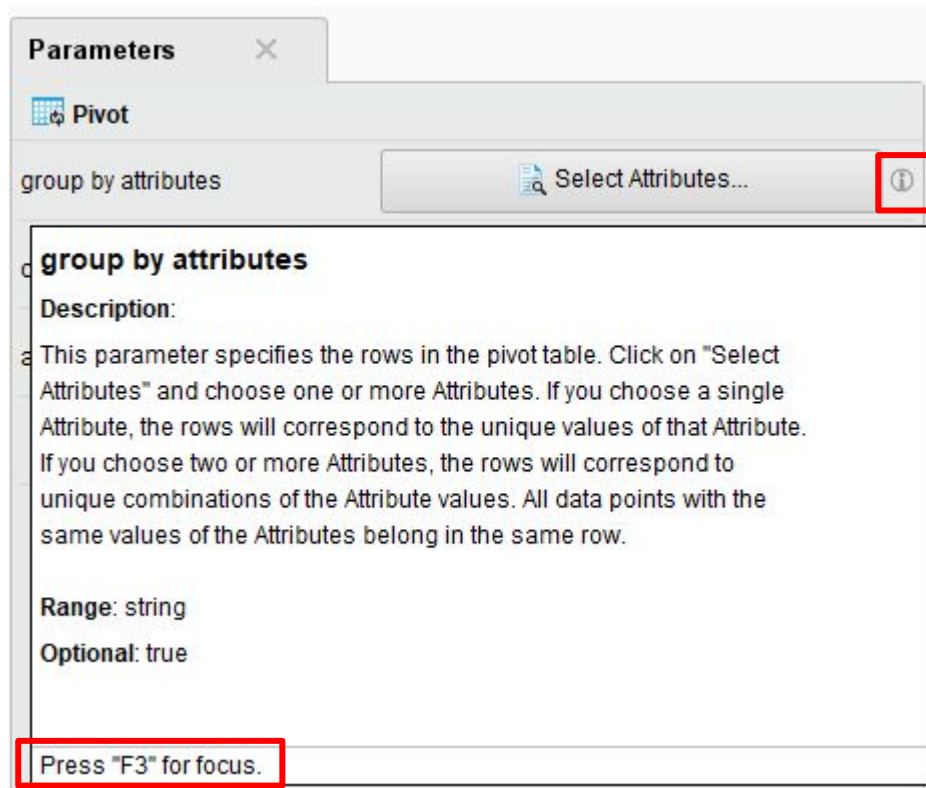
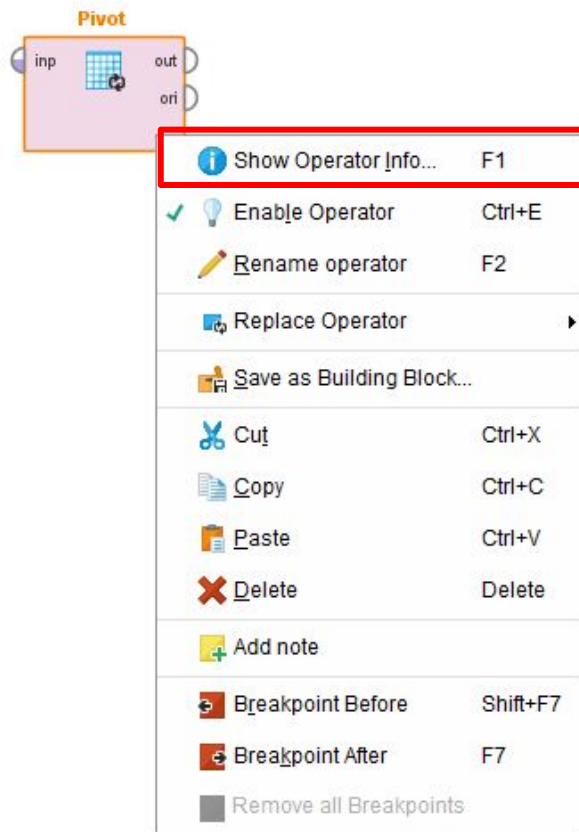
基本上都是把左邊 Repository 或 Operators 的元件拖曳到中間執行操作, 請善用搜尋功能查找元件

Recommended Operators

Retrieve 12%  
Select Attributes 6%  
Set Role 5%  
Apply Model 4%  
Filter Examples 4%



如果對 Operators 或其他功能有疑問, 可以在元件點選右鍵 -> F1;  
或是滑鼠移到 (i) 的圖示、按下 F3 查看定義



# 關於機器學習

# 關於機器學習

As managing the process that can transform **hypotheses** and **data** into actionable **predictions**

- Predicting who will win an election
- What products will sell well together
- Which loans will default
- Which advertisements will be clicked on

# 關於機器學習

The data scientist is responsible for

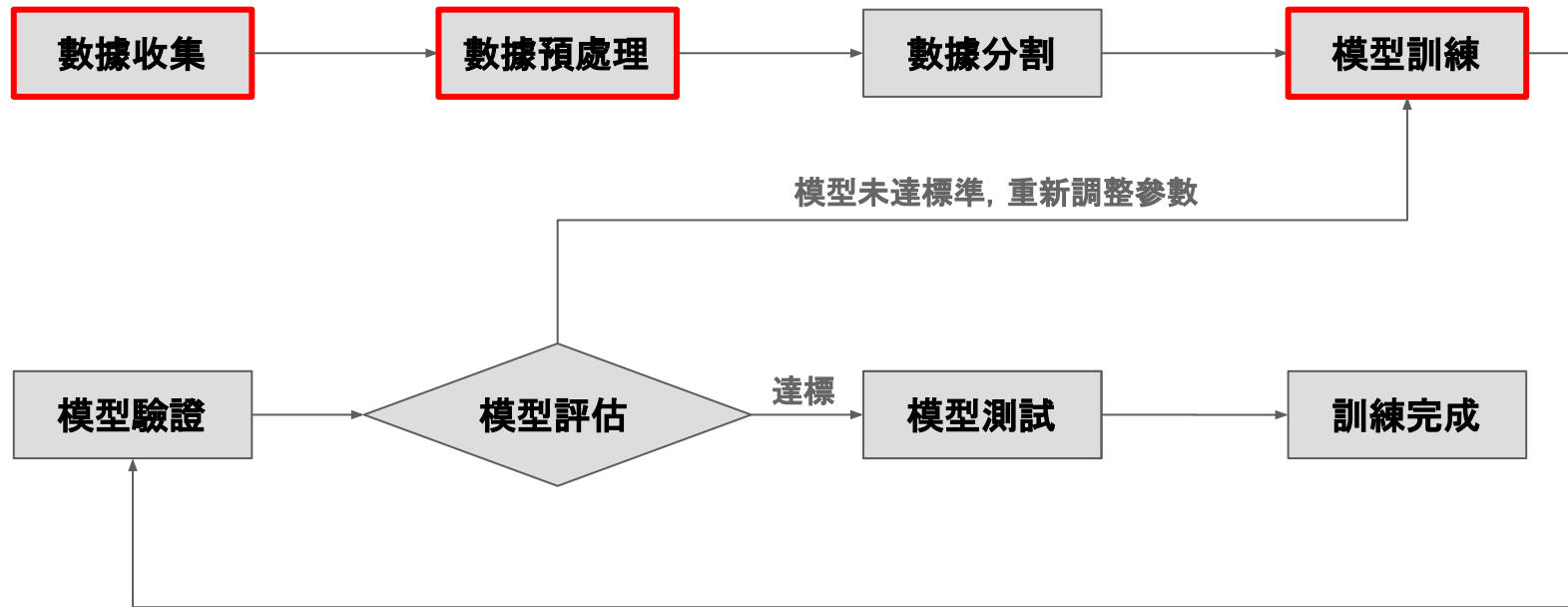
- Data : acquiring the data, managing the data
- Modeling: choosing the modeling technique, writing the code
- Evaluation: verifying the results

## supervised v.s. unsupervised

- supervised statistical learning
  - involves building a statistical model for predicting, or estimating, an output based on one or more inputs
- unsupervised statistical learning
  - there are inputs but no supervising output; nevertheless we can learn relationships and structure from such data



# 模型訓練流程



# 模型訓練流程 - 名詞

- **數據收集:**

按照決策需求與產業知識決定自變數 X、應變數 Y

- **數據預處理:**

確保後續的模型訓練順利進行，常見的手段包含數據轉換、填補缺失值 (Missing Data)、刪除離群值 (Outlier)、數據標準化(Standardization)、數據正規化 (Normalization)

- **數據分割:**

將資料分割成三個部分，分別用於**訓練 (Training)**、**驗證 (Validation)**、**測試 (Testing)**。更嚴謹的做法是執行**交叉驗證 (Cross Validation)**，將以上的訓練及驗證資料合併並切分成  $n$  等份，循環使用其中  $n - 1$  份資料訓練模型、使用剩下的 1 份資料驗證模型

# 模型訓練流程 - 名詞

- **模型訓練:**

使用恰當的機器學習演算法對訓練資料進行**擬合 (Fitting)**，進而產生出可用於後續預測分析用的**模型 (Model)**

- **模型驗證:**

將驗證數據的自變數  $X$  輸入訓練好的模型 (Model)，得到**預測值  $\hat{Y}$** ， $\hat{Y}$  又稱為**擬合值**

- **模型評估:**

主要的方式為比較 **應變數  $Y$**  和 **擬合值  $\hat{Y}$**  之間的差異性，差異愈小則模型效果愈好、反之則愈差；評估方式參考以下：

- **應變數  $Y$  是數值(量性變項):** 均方根誤差 (RMSE)、平均絕對誤差 (MAE)、判定係數 ( $R^2$ )
- **應變數  $Y$  是類別(質性變項):** 混淆矩陣 (Confusion Matrix)、準確度 (Accuracy)、ROC 曲線下方面積 (AUC)

# 模型訓練流程 - 名詞

- **模型測試:**

將來自同一個資料集且**沒有參與訓練與驗證的數據**輸入模型中進行模型評估，評估通過的模型即可應用於實際場景。

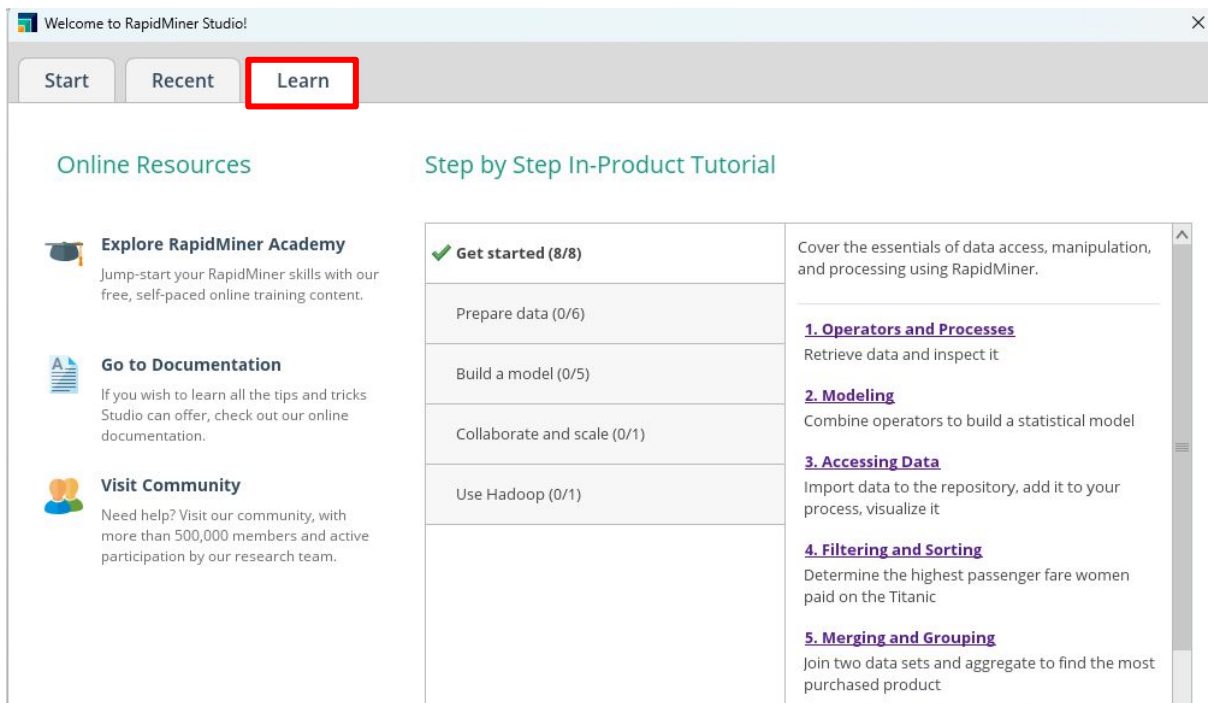
**Idea #3:** Split data into **train**, **val**, and **test**; choose hyperparameters on val and evaluate on test

**Better!**



請大家從 Help -> Tutorials 打開內建的使用教學，接下來將會帶大家拆解練習

## Get started (8/8) + Prepare data(4/6)



The screenshot shows the 'Learn' tab in the RapidMiner Studio interface. At the top, there are three buttons: 'Start', 'Recent', and 'Learn', with 'Learn' highlighted by a red rectangle. Below the buttons, the interface is divided into two main sections: 'Online Resources' on the left and 'Step by Step In-Product Tutorial' on the right.

**Online Resources**

- Explore RapidMiner Academy**  
Jump-start your RapidMiner skills with our free, self-paced online training content.
- Go to Documentation**  
If you wish to learn all the tips and tricks Studio can offer, check out our online documentation.
- Visit Community**  
Need help? Visit our community, with more than 500,000 members and active participation by our research team.

**Step by Step In-Product Tutorial**

Step	Description
✓ Get started (8/8)	Cover the essentials of data access, manipulation, and processing using RapidMiner.
Prepare data (0/6)	
Build a model (0/5)	
Collaborate and scale (0/1)	
Use Hadoop (0/1)	

**1. Operators and Processes**  
Retrieve data and inspect it

**2. Modeling**  
Combine operators to build a statistical model

**3. Accessing Data**  
Import data to the repository, add it to your process, visualize it

**4. Filtering and Sorting**  
Determine the highest passenger fare women paid on the Titanic

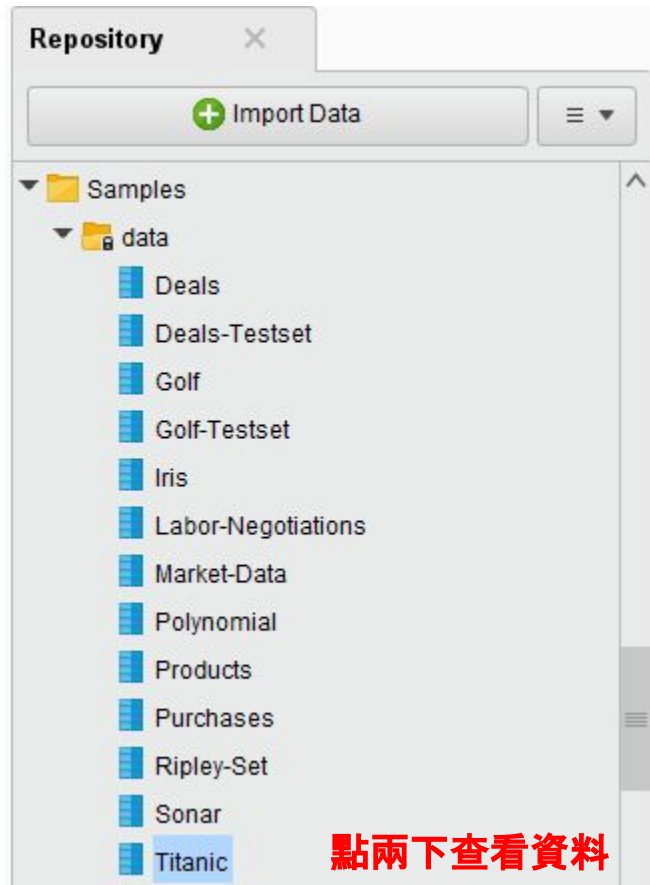
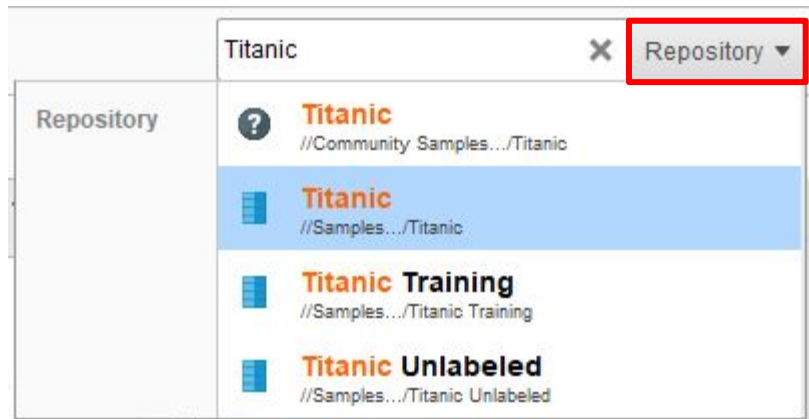
**5. Merging and Grouping**  
Join two data sets and aggregate to find the most purchased product

Skip “**Macros and Sampling**”, “**Looping, Branching, and Appending**” till the next time.

# 樞紐分析

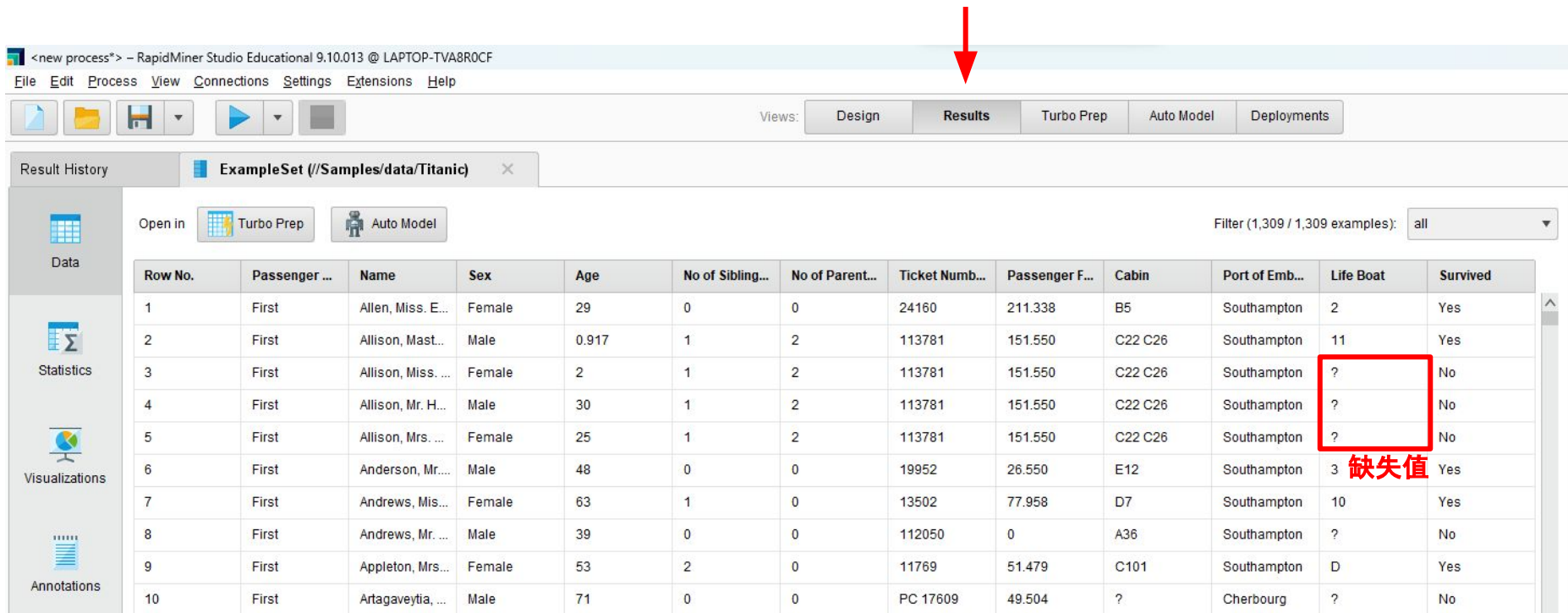
# 資料集

透過右上搜尋欄尋找 Titanic 資料集, 位於Repository/Samples/data 底下



# 資料集

共有 1309 筆資料, 分別為 1309 位登上鐵達尼號的旅客; 每一筆資料有 13 個欄位, 表示旅客的資訊。  
這份數據主要記錄了旅客是否在這次災難中存活, 可以發現資料中存在一些缺失值。



Views: Design Results Turbo Prep Auto Model Deployments

Result History: ExampleSet (/Samples/data/Titanic)

Open in: Turbo Prep Auto Model

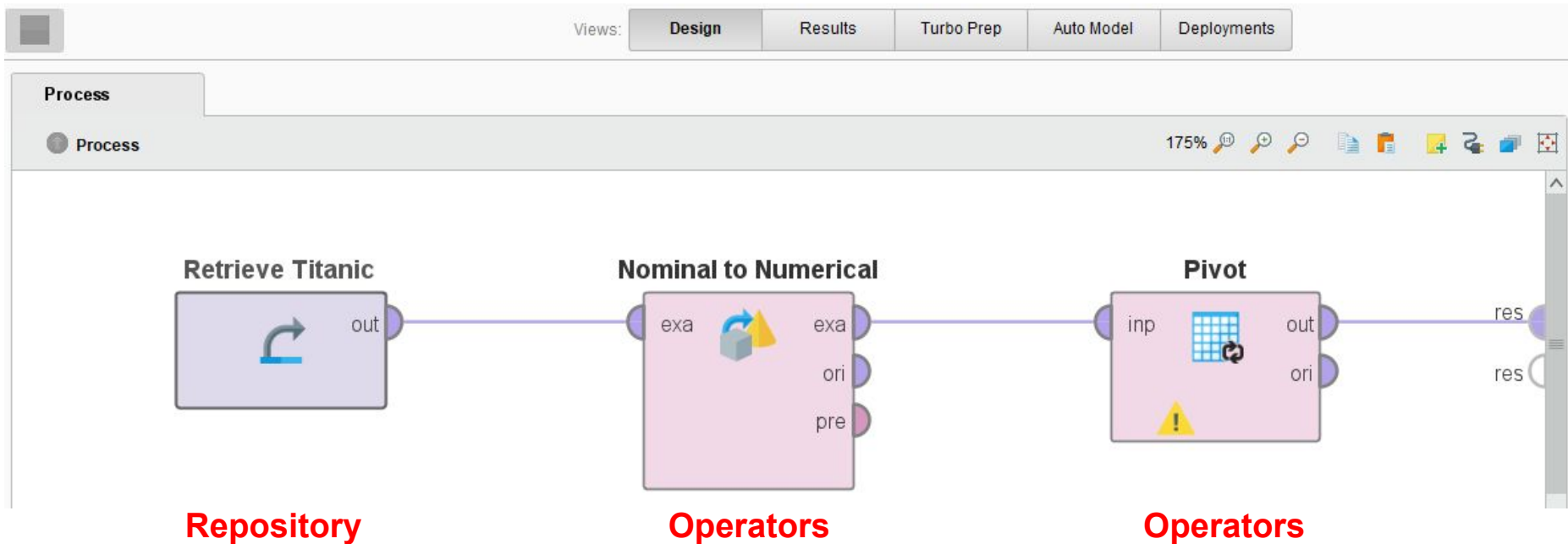
Filter (1,309 / 1,309 examples): all

Row No.	Passenger ...	Name	Sex	Age	No of Sibling...	No of Parent...	Ticket Num...	Passenger F...	Cabin	Port of Emb...	Life Boat	Survived
1	First	Allen, Miss. E...	Female	29	0	0	24160	211.338	B5	Southampton	2	Yes
2	First	Allison, Mast...	Male	0.917	1	2	113781	151.550	C22 C26	Southampton	11	Yes
3	First	Allison, Miss. ...	Female	2	1	2	113781	151.550	C22 C26	Southampton	?	No
4	First	Allison, Mr. H...	Male	30	1	2	113781	151.550	C22 C26	Southampton	?	No
5	First	Allison, Mrs. ...	Female	25	1	2	113781	151.550	C22 C26	Southampton	?	No
6	First	Anderson, Mr...	Male	48	0	0	19952	26.550	E12	Southampton	3 缺失值	Yes
7	First	Andrews, Mis...	Female	63	1	0	13502	77.958	D7	Southampton	10	Yes
8	First	Andrews, Mr. ...	Male	39	0	0	112050	0	A36	Southampton	?	No
9	First	Appleton, Mrs...	Female	53	2	0	11769	51.479	C101	Southampton	D	Yes
10	First	Artagaveytia, ...	Male	71	0	0	PC 17609	49.504	?	Cherbourg	?	No



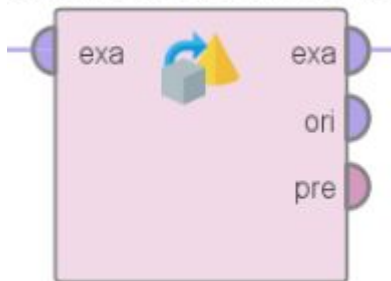
# 設計流程

請依據關鍵字，在 Repository 或 Operators 找出以下指定元件，並拖曳到 Process 區域、連結元件  
(善用搜尋功能)



# 參數設定

## Nominal to Numerical



## Operators

**Parameters** ✕

**Nominal to Numerical**

☐ *create view* ⓘ

**attribute filter type** single ▼ ⓘ

**attribute** Survived ▼ ⓘ

☐ *invert selection* ⓘ

☐ *include special attributes* ⓘ

**coding type** dummy coding ▼ ⓘ

☐ *use comparison groups* ⓘ

**unexpected value handling** all 0 and warning ▼ ⓘ

☐ *use underscore in name* ⓘ

# 參數設定

## Pivot



## Operators

Parameters

Pivot

group by attributes Select Attributes...

column grouping attribute Sex

aggregation attributes Edit List (1)...

☐ use default aggregation



Select Attributes: group by attributes

Select Attributes: **group by attributes**  
Attributes that groups the examples which form one row after pivoting.

Attributes

Search

- # Age
- Cabin
- Life Boat
- Name
- # No of Parents or Children on Board
- # No of Siblings or Spouses on Board
- Passenger Class**
- # Passenger Fare
- Port of Embarkation
- Sex
- # Survived = No
- # Survived = Yes
- Ticket Number

Selected Attributes

Search

☒ Apply ☐ Cancel

# 參數設定

## Pivot



## Operators

Parameters

Pivot

group by attributes Select Attributes...

column grouping attribute Sex

aggregation attributes Edit List (1)...

☐ use default aggregation



Edit Parameter List: aggregation attributes

Edit Parameter List: **aggregation attributes**  
The attributes which should be aggregated.

aggregation attribute	aggregation function
Survived = Yes	average

Add Entry Remove Entry Apply Cancel

# 執行

<new process\*> - RapidMiner Studio Educational 9.10.013 @ LAPTOP-TVA8R0CF

File Edit Process View Connections Settings Extensions Help



Views:

Design

Results

Turbo Prep

Auto Model

Deployments

Repository

+ Import Data

- Polynomial
- Products
- Purchases
- Ripley-Set
- Sonar
- Titanic
- Titanic Training
- Titanic Unlabeled
- Transactions
- Weighting

- processes
- Templates
- Time Series
- Tutorials

Local Repository (Local)

DB (Legacy)

Operators

Process

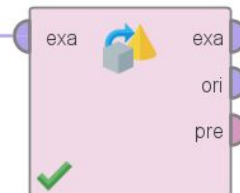
Process

175%

Retrieve Titanic



Nominal to Numerical



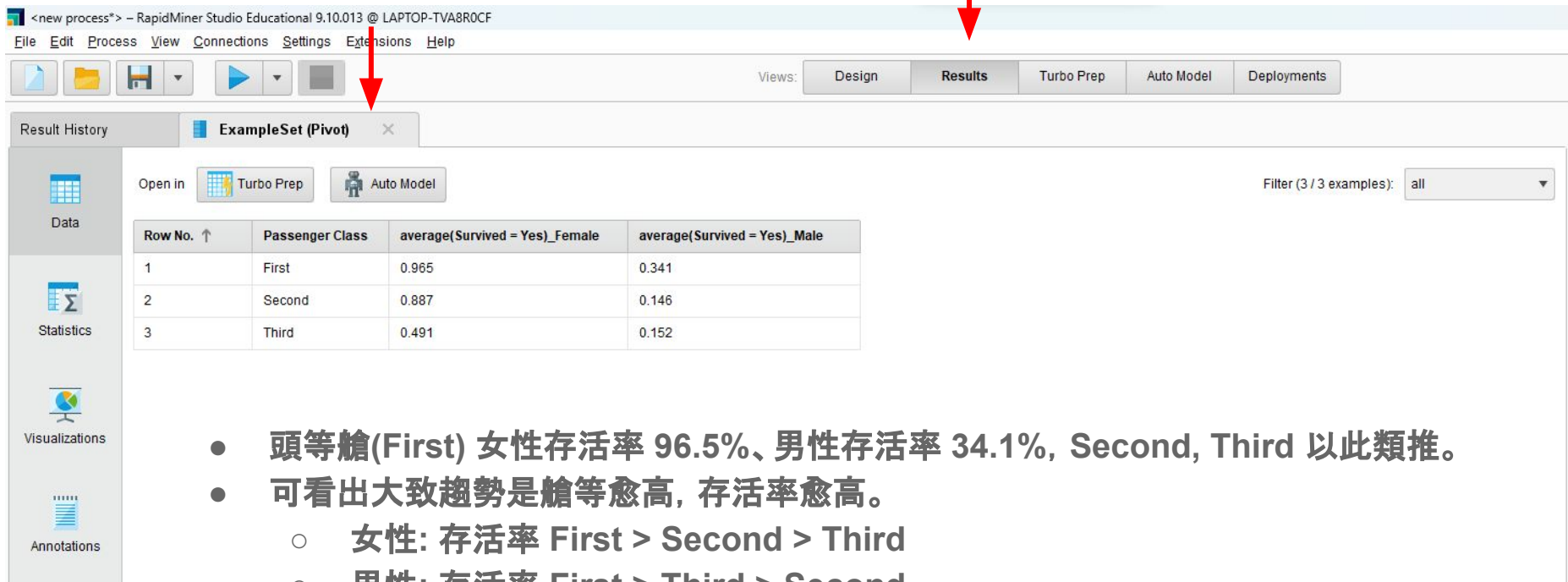
Pivot



res

res

# 查看結果



The screenshot shows the RapidMiner Studio interface. The top menu bar includes File, Edit, Process, View, Connections, Settings, Extensions, and Help. The toolbar contains icons for file operations and execution. The 'Views' section at the top right has tabs for Design, Results, Turbo Prep, Auto Model, and Deployments. The 'Results' tab is active, displaying a table of results for an 'ExampleSet (Pivot)'. The table has four columns: Row No., Passenger Class, average(Survived = Yes)\_Female, and average(Survived = Yes)\_Male. The data shows survival rates for First, Second, and Third class passengers, with First class having the highest survival rate for both genders.

Row No. ↑	Passenger Class	average(Survived = Yes)_Female	average(Survived = Yes)_Male
1	First	0.965	0.341
2	Second	0.887	0.146
3	Third	0.491	0.152

- 頭等艙(First) 女性存活率 96.5%、男性存活率 34.1%, Second, Third 以此類推。
- 可看出大致趨勢是艙等愈高, 存活率愈高。
  - 女性: 存活率 First > Second > Third
  - 男性: 存活率 First > Third > Second

# Kaggle 競賽

# Kaggle 介紹

這次的作業將進行 Titanic 生存預測 (預測目標為 Survived: 生存 1、死亡 0)

競賽連結會放在 Moodle, 請同學透過連結加入班級競賽

Community Prediction Competition · Private

## 1121 W2\_234 Computational Thinking - RapidMiner1

Titanic - Machine Learning from Disaster. Please use the software "RapidMiner" exclusively for this competition.

Host

Overview

Data

Discussion

Leaderboard

Rules

Team

Submissions

...

### Overview

This competition is based on the **builtin-datasets provided by the RapidMiner**. The target is to predict the survival of the passengers in the Titanic-Test.csv by training a Machine-Learning model from the Titanic-Train.csv (**Survived = 1 means alive, Survived = 0 means dead**).

Please follow the guidance provided in Moodle to build your own RapidMiner workflow.

#### Start

Set start date via the [launch checklist](#).

#### Close

15 days to go

#### Competition Host

cheng-zhi-rong

#### Prizes & Awards

Kudos  
Does not award Points or Medals

#### Participation

0 Competitors  
0 Teams  
0 Entries

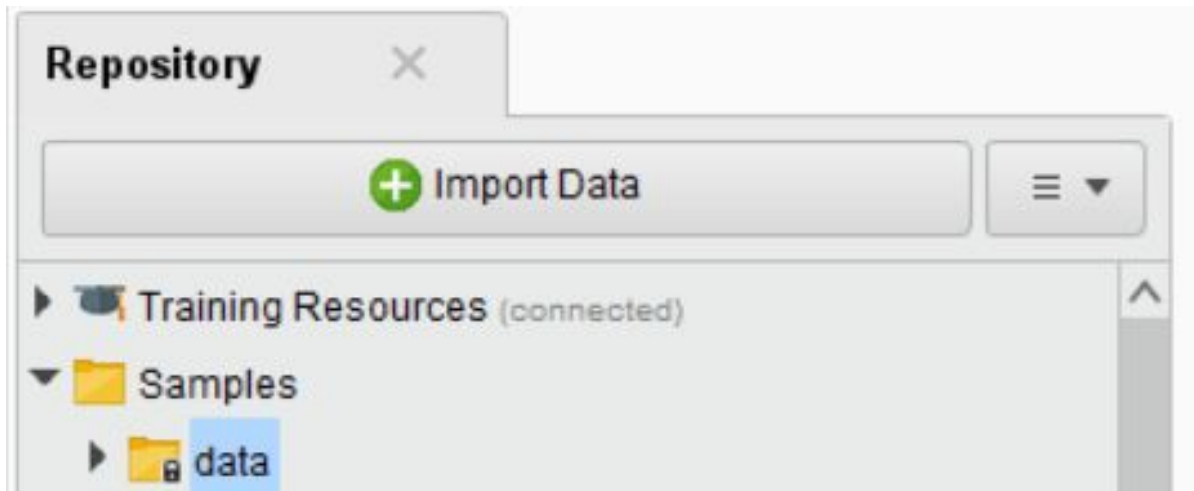
#### Tags

Add Tags



# Titanic 資料集

競賽使用的 Titanic 資料集是修改自 RapidMiner 內建的 Samples/data 資料



# 競賽資料集

請從 Kaggle 下載 Titanic-Train.csv, Titanic-Test.csv 兩個檔案, 分別為訓練資料、測試資料 (**不要用 RapidMiner 內建的 Titanic**)。另外 Sample-Submission.csv 為參考資料, 表示規定的上傳格式。

## Data Explorer

113.95 kB

Sample-Submission.csv

Titanic-Test.csv

Titanic-Train.csv

Detail Compact Column

### About this file

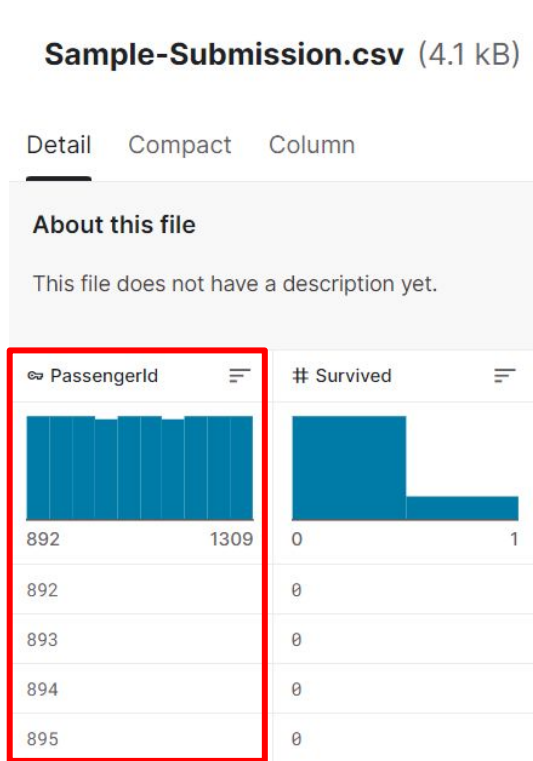
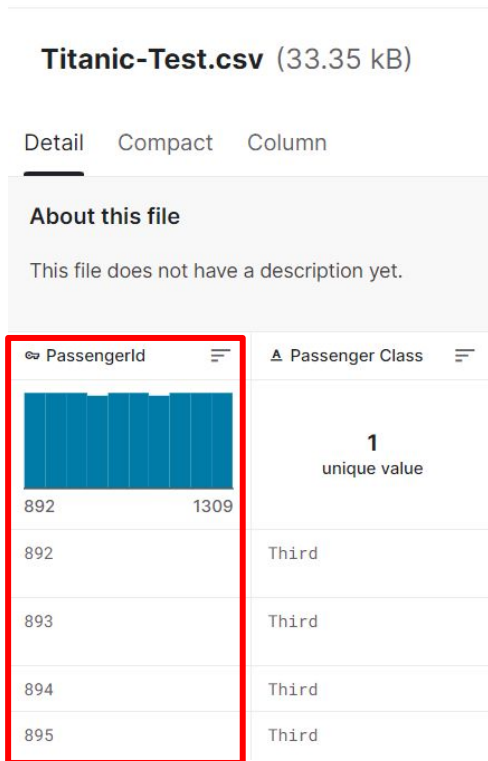
This file does not have a description yet.

上傳格式只能有這兩個欄位

PassengerId	# Survived
892	0
893	1
894	0
895	0
896	1

# 競賽目標

透過 Titanic-Train.csv 訓練模型 >> 拿去預測 Titanic-Test.csv >> 產生對應的猜測結果  
Sample\_Submission.csv (可以看到兩個資料集的 PassengerId 是相同的)



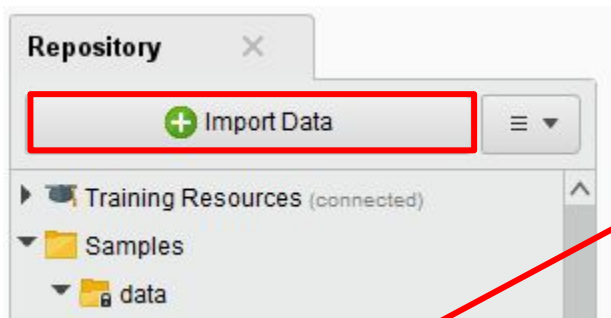
## 競賽評分方式 - Accuracy

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

只需要知道上傳的分數**愈高愈好**即可

# 匯入訓練資料

將 **Titanic-Train.csv** 匯入 RapidMiner 的 Local Repository/data 資料夾，請記得做以下設定



PassengerId 要 Change Role 為 id

Import Data - Format your columns.

Format your columns.

Date format: Enter value...

☒ Replace errors with missing values 勾選

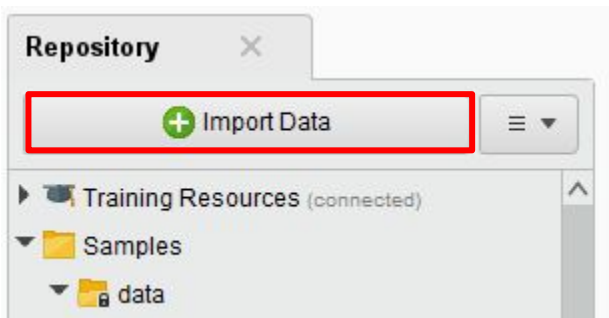
	PassengerId	Passenger ...	Name	Sex	Age	No of Sibli...	No of Pare...	Ticket Num..
	integer	polynomial	polynomial	polynomial	real	integer	integer	polynomial
1	1	First	Allen, Miss. Elisa...	Female	29.000	0	0	24160
2	2	First	Allison, Master. H...	Male	0.917	1	2	113781
3	3	First	Allison, Miss. Hel...	Female	2.000	1	2	113781
4	4	First	Allison, Mr. Huds...	Male	30.000	1	2	113781
5	5	First	Allison, Mrs. Hud...	Female	25.000	1	2	113781
6	6	First	Anderson, Mr. Ha...	Male	48.000	0	0	19952
7	7	First	Andrews, Miss. K...	Female	63.000	1	0	13502
8	8	First	Andrews, Mr. Tho...	Male	39.000	0	0	112050
9	9	First	Appleton, Mrs. Ed...	Female	53.000	2	0	11769
10	10	First	Artagaveytia, Mr. ...	Male	71.000	0	0	PC 17609
11	11	First	Astor, Col. John J...	Male	47.000	1	0	PC 17757
12	12	First	Astor, Mrs. John ...	Female	18.000	1	0	PC 17757
13	13	First	Aubart, Mme. Leo...	Female	24.000	0	0	PC 17477
14	14	First	Barber, Miss. Ell...	Female	26.000	0	0	19877
15	15	First	Barkworth, Mr. Al...	Male	80.000	0	0	27042
16	16	First	Baumann, Mr. Jo...	Male	?	0	0	PC 17318
17	17	First	Baxter, Mr. Quigg ...	Male	24.000	0	1	PC 17558

no problems.

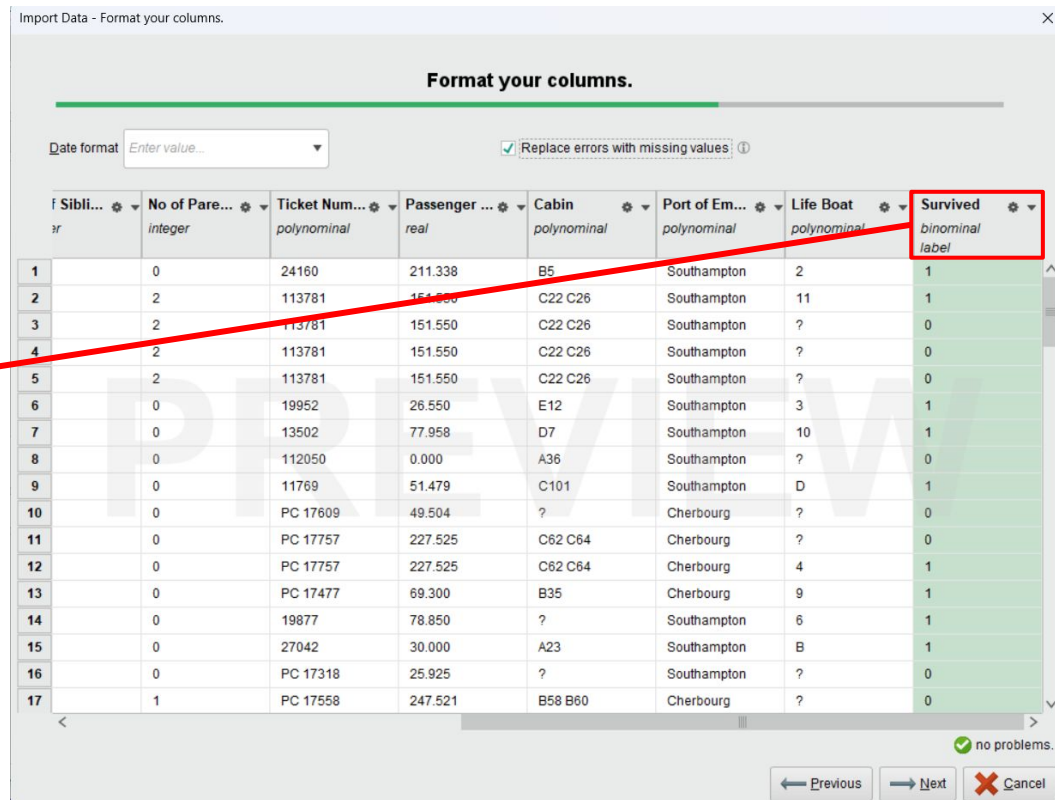
Previous Next Cancel

# 匯入訓練資料

將 **Titanic-Train.csv** 匯入 RapidMiner 的 Local Repository/data 資料夾, 請記得做以下設定

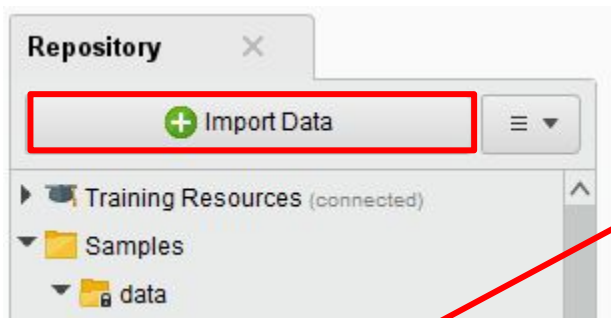


Survived 要 Change Role 為 label,  
且要 Change Type 為 binominal



# 匯入訓練資料

將 **Titanic-Test.csv** 匯入 RapidMiner 的 Local Repository/data 資料夾, 請記得做以下設定



PassengerId 要 Change Role 為 id

Import Data - Format your columns.

Format your columns.

Date format: Enter value...

☒ Replace errors with missing values 勾選

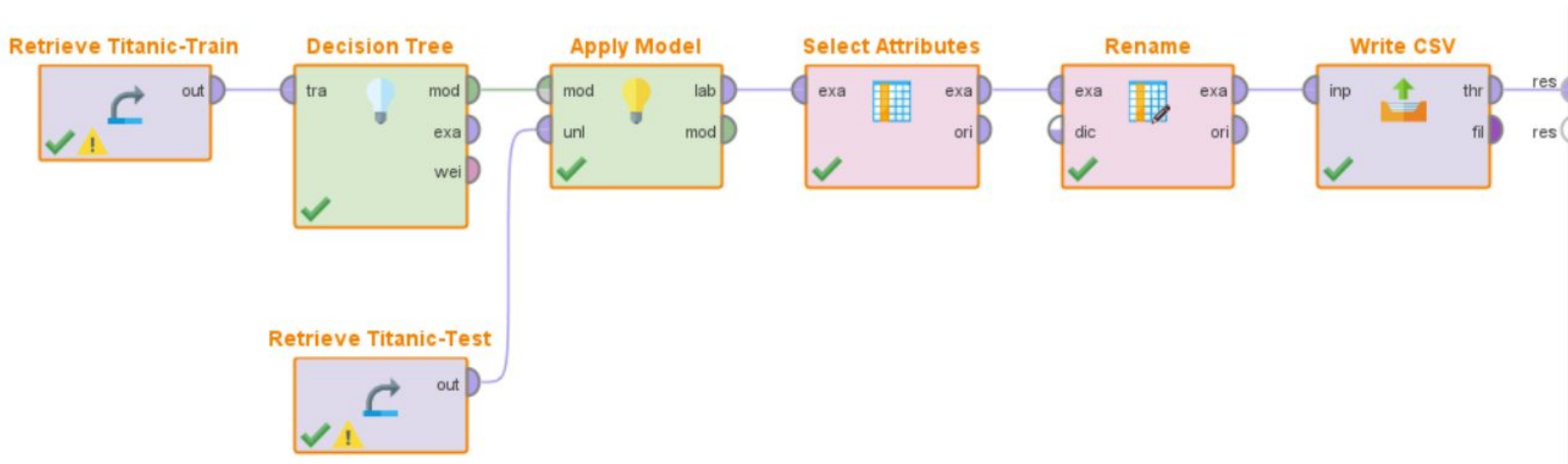
	PassengerId	Passenger ...	Name	Sex	Age	No of Sibli...	No of Pare...	Ticket Num..
	integer	polynomial	polynomial	polynomial	real	integer	integer	polynomial
1	1	First	Allen, Miss. Elisa...	Female	29.000	0	0	24160
2	2	First	Allison, Master. H...	Male	0.917	1	2	113781
3	3	First	Allison, Miss. Hel...	Female	2.000	1	2	113781
4	4	First	Allison, Mr. Huds...	Male	30.000	1	2	113781
5	5	First	Allison, Mrs. Hud...	Female	25.000	1	2	113781
6	6	First	Anderson, Mr. Ha...	Male	48.000	0	0	19952
7	7	First	Andrews, Miss. K...	Female	63.000	1	0	13502
8	8	First	Andrews, Mr. Tho...	Male	39.000	0	0	112050
9	9	First	Appleton, Mrs. Ed...	Female	53.000	2	0	11769
10	10	First	Artagaveytia, Mr. ...	Male	71.000	0	0	PC 17609
11	11	First	Astor, Col. John J...	Male	47.000	1	0	PC 17757
12	12	First	Astor, Mrs. John ...	Female	18.000	1	0	PC 17757
13	13	First	Aubart, Mme. Leo...	Female	24.000	0	0	PC 17477
14	14	First	Barber, Miss. Ell...	Female	26.000	0	0	19877
15	15	First	Barkworth, Mr. Al...	Male	80.000	0	0	27042
16	16	First	Baumann, Mr. Jo...	Male	?	0	0	PC 17318
17	17	First	Baxter, Mr. Quigg ...	Male	24.000	0	1	PC 17558

no problems.

Previous Next Cancel

# 基本設計流程

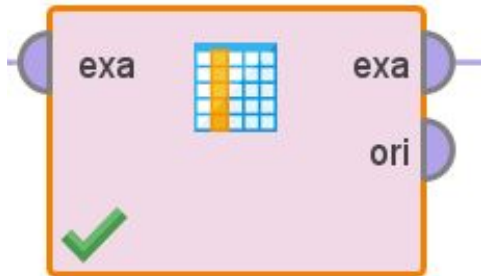
以下為基本的設計流程，請透過RapidMiner 的搜尋功能找出以下元件並排好。接著會說明參數設定，沒有特別講就是不用改設定。





# 參數設定

## Select Attributes



**Parameters**

Select Attributes

attribute filter type ☒ subset

attributes ☒ Select Attributes...

☐ invert selection

☒ include special attributes

記得勾選

Select Attributes: attributes

Select Attributes: **attributes**  
The attribute which should be chosen.

Attributes

Search

- # Age
- # Cabin
- # confidence(0)
- # confidence(1)
- # Embarked
- # Fare
- # Name
- # Parch
- # Pclass
- # Sex
- # SibSp
- # Ticket

Selected Attributes

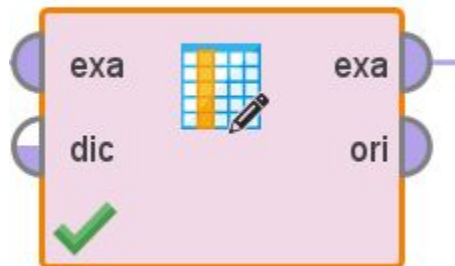
Search

- # PassengerId
- # prediction(Survived)

Apply Cancel

# 參數設定

## Rename



Edit Parameter List: rename attributes

Edit Parameter List: **rename attributes**  
Use this list to define the renaming of the attributes.

old name	new name
prediction(Survived)	Survived

Buttons: Add Entry, Remove Entry, Apply, Cancel


# 參數設定




更改成你要存放的位置，並命名為 xxx.csv

Parameters

Write CSV


csv file   ⓘ

column separator   是英文的逗號 ⓘ

☒ write attribute names ⓘ

☒ quote nominal values ⓘ



☒ format date attributes ⓘ

date format   ⓘ

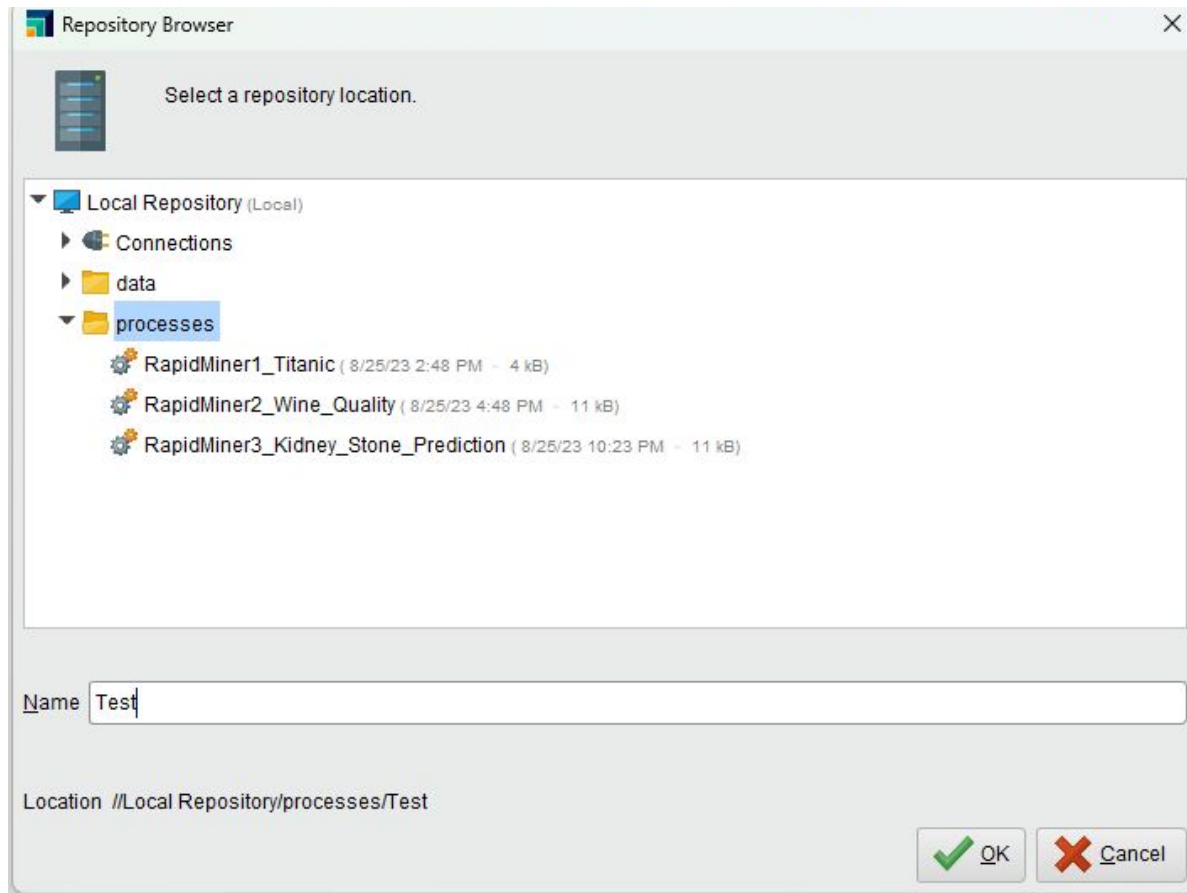
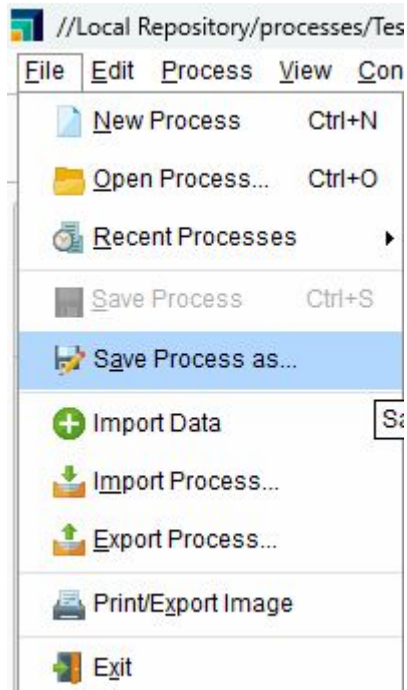
☐ append to file ⓘ

encoding  ⓘ

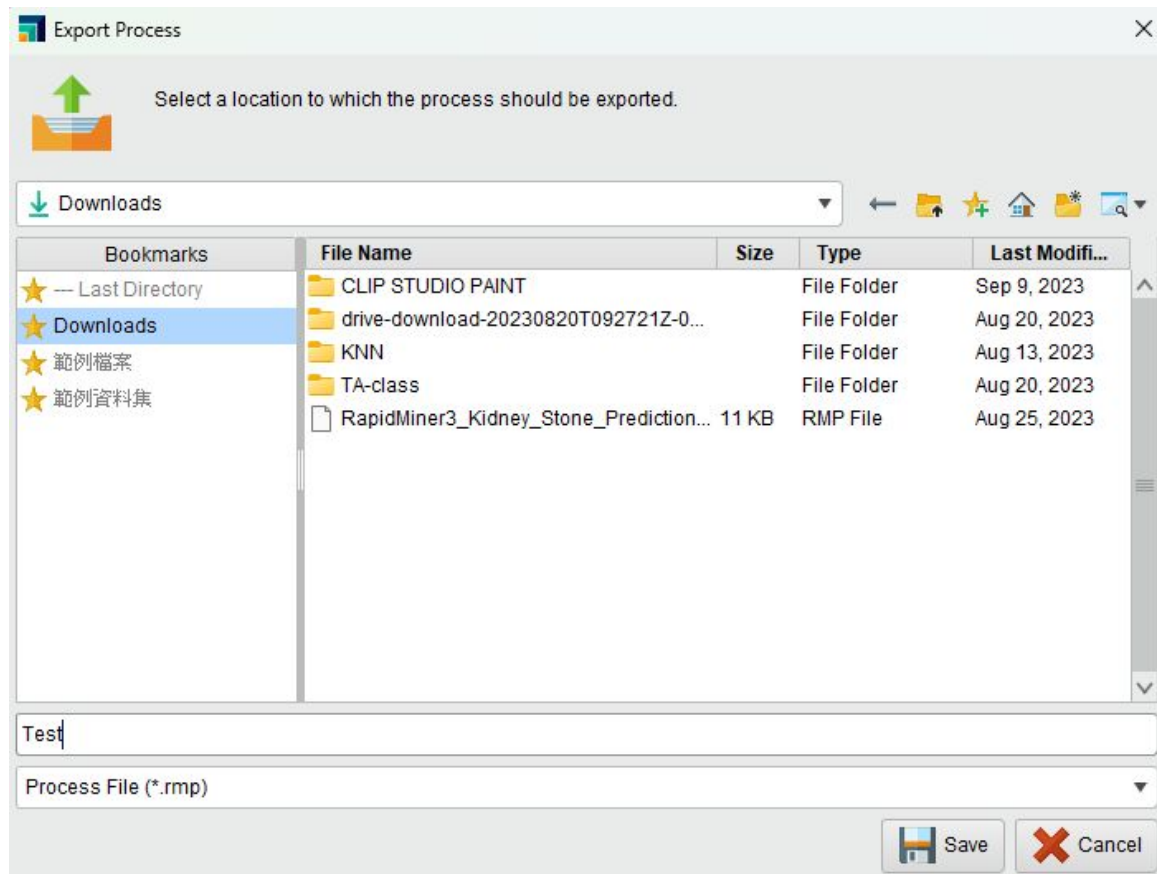
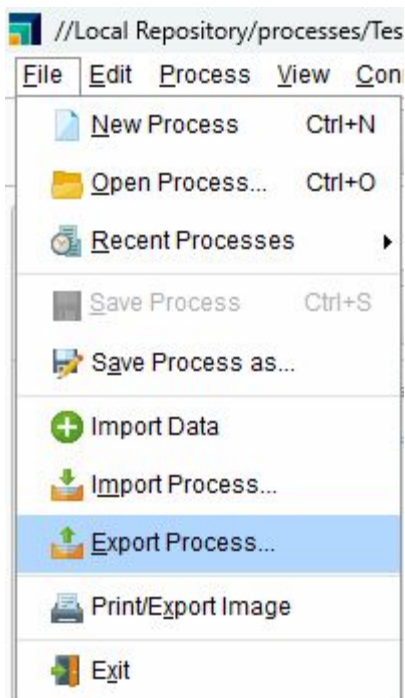
# 檢查輸出

ExampleSet (Rename) <span>×</span>		
Open in <span> Turbo Prep</span> <span> Auto Model</span>		
Row No.	PassengerId	Survived
1	892	0
2	893	0
3	894	0
4	895	0
5	896	1
6	897	0
7	898	0
8	899	0
9	900	1
10	901	0

# 儲存 Process 檔案於 Local Repository



# Export Process (副檔名 .rmp)



# 更改 Kaggle Team Name

請注意 Team name 務必改成以下格式 學號-系級-名字

Overview Data Code Discussion Leaderboard Rules Team

Submissions

Submit Predictions

...

## Your Team

Everyone that competes in a Competition does so as a team - even if you're competing by yourself. [Learn more.](#)

## General

TEAM NAME

111753151-資碩一-程至榮

This name will appear on your team's leaderboard position.

# Kaggle 上傳方式

## Step1. 點選 Submit Predictions

 Community Prediction Competition · Private

### 1121 W2\_234 Computational Thinking - RapidMiner1

Titanic - Machine Learning from Disaster. Please use the software "RapidMiner" exclusively for this competition.

15 days to go

[Overview](#) [Data](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Team](#) [Host](#) [Submissions](#) [Submit Predictions](#) [...](#)

## Step2. 檢查是否上傳成功

### Submissions

Select up to 2 submissions that will count towards your final leaderboard score. If less than 2 are selected, Kaggle will automatically select from your best scoring submissions. [Learn More](#)

0/2

☐ Auto-selection candidates [?](#)

All

Successful

Selected

Errors

Recent ▾

Submission and Description

Public Score ⓘ

Select



Sample\_Submission.csv

Complete · 30s ago

0.74242








# Leaderboard - 查看排名

點選 Leaderboard 可以查看自己的名次【請注意 Team 名稱格式, 避免影響分數統計】

Public Private

This leaderboard is calculated with approximately 32% of the test data. The final results will be based on the other 68%, so the final standings may be different.

#	Team	Members	Score	Entries	Last
	Baseline		0.74242		
1	111753151-資碩一-程至榮		0.74242	1	2m
	Your First Entry! Welcome to the leaderboard!				

# Leaderboard - Public vs Private

請注意, "比賽過程中" Leaderboard 上的排名都只是暫時的、參考用(Public score), **要到比賽結束後才會顯示你的真實排名 (Private score)**。原因是避免參賽者用不斷上傳的方式猜出比賽測試集的真實答案, 導致排名失準。

**Note:** Public score 是指用 **30%** 的測試資料算出來的分數, Private score 是指用剩下的 **70%** 測試資料算出來的分數, 最後會以 **Private score** 作為排名依據。

Public Private

---

This leaderboard is calculated with approximately 30% of the test data. The final results will be based on the other 70%, so the final standings may be different.

Public Private

---

The private leaderboard is calculated with approximately 70% of the test data.  
This competition has completed. This leaderboard reflects the final standings.

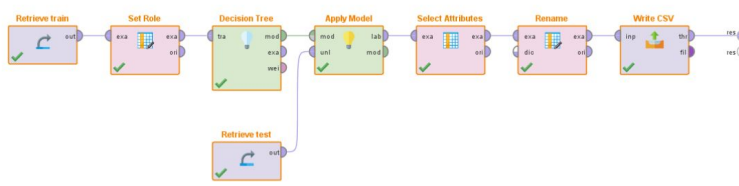
# 作業要求

# 基本題 4 分

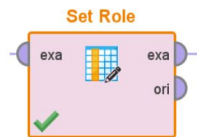
1. 不限定模型，成功上傳 kaggle(即 Leaderboard 有你的名字，且**格式正確**),
2. **Public Score** 優於 Baseline 的 **0.74242**
3. 在 Moodle **上傳至少 1 頁 PDF**，說明你的設計流程、參數設定，並附上截圖(請參考簡報前面的教學是怎麼做的)
4. **上傳你的 Process file** (檔名: 學號\_RapidMiner1.rmp, 範例: 111753151\_RapidMiner1.rmp)

## 基本設計流程

以下為基本的設計流程，請透過 RapidMiner 的搜尋功能找出以下元件並排好。接著會說明參數設定，沒有特別講就是不用改設定。



## 參數設定




Parameters	
Set Role	
attribute name	Survived
target role	label
set additional roles	<a href="#">Edit List (0)...</a>

## 加分題 1 分

1. 結算後在 **Leaderboard** 排名前 50% 且 **Private Score** 優於 **Baseline** 的同學可以得到額外的分數。

Public Private

The private leaderboard is calculated with approximately 68% of the test data.

#	△	Team	Members	Score	Entries	Last
		Baseline		0.74825		

透過上課教過的方法 + 上網查詢、調整資料前處理的方法 or 模型參數。

接下來的作業都沒有標準答案, 請大家盡可能的去嘗試!

# 作業注意事項

- 為了公平起見，且大家的**期末專題海報**預計會與 micro:bit 或 RapidMiner 有關，此作業**限定使用 RapidMiner 產生的 Submission 參賽**，請確認繳交的 Process file 可以產生正確的 Submission 檔案
- **請不要抄襲、或是直接拿別人的 Submission 檔案上傳**，助教會隨機抽查是否有排名分數與 Process file 不一致的問題。

# Welcome to RapidMiner Documentation!

Documentation, tutorials, and reference materials for the RapidMiner platform

## New to RapidMiner?

Quickly learn the basics of RapidMiner Studio – the core of the RapidMiner platform – with this tutorial:

[Getting Started with RapidMiner Studio](#)

# 線上資源 - Kaggle Learn

≡ kaggle

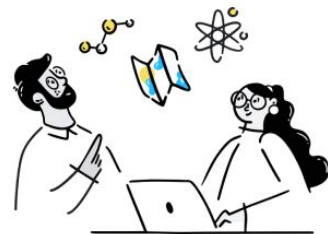
+ Create

- Home
- Competitions
- Datasets
- Models
- Code
- Discussions
- Learn**
- More
- User Rankings
- Blog
- Documentation
- Progression
- Host a Competition
- KaggleX Mentorship
- Support/Contact
- Community Guidelines

🔍 Search

## Learn

Gain the skills you need to do independent data science projects.



### Your Courses

Active



#### Intro to Machine Learning

Next up: [Exercise: Explore Your Data](#)

54%



#### Intermediate Machine Learning

Next up: [Introduction](#)

7%



#### Feature Engineering

Next up: [What Is Feature Engineering](#)

9%



#### Intro to Deep Learning

Next up: [A Single Neuron](#)

8%



#### Data Cleaning

Next up: [Handling Missing Values](#)

10%



# 線上資源 - [scikit-learn](https://scikit-learn.org/)



[Install](#) [User Guide](#) [API](#) [Examples](#) [Community](#) [More ▾](#)

## scikit-learn

*Machine Learning in Python*

[Getting Started](#)

[Release Highlights for 1.3](#)

[GitHub](#)

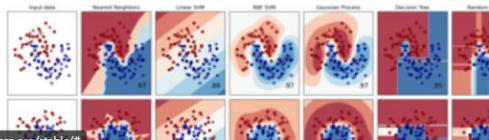
- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

### Classification

Identifying which category an object belongs to.

**Applications:** Spam detection, image recognition.

**Algorithms:** Gradient boosting, nearest neighbors, random forest, logistic regression, and more...

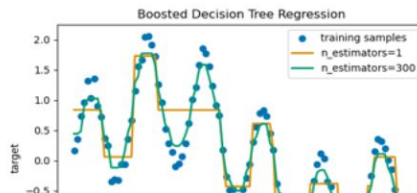


### Regression

Predicting a continuous-valued attribute associated with an object.

**Applications:** Drug response, Stock prices.

**Algorithms:** Gradient boosting, nearest neighbors, random forest, ridge, and more...

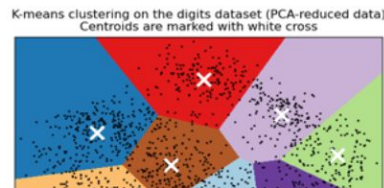


### Clustering

Automatic grouping of similar objects into sets.

**Applications:** Customer segmentation, Grouping experiment outcomes

**Algorithms:** k-Means, HDBSCAN, hierarchical clustering, and more...





**Bonus**

# Unsupervised learning (K-Means)



# 匯入訓練資料

Import Data - Format your columns. ✕

**Format your columns.**

Date format  ▼ ☒ Replace errors with missing values ℹ

	CustomerID <span>⚙</span> <span>▼</span>	Gender <span>⚙</span> <span>▼</span>	Age <span>⚙</span> <span>▼</span>	Annual Income (k\$) <span>⚙</span> <span>▼</span>	Spending Score (1-100) <span>⚙</span> <span>▼</span>
	<i>integer</i> <i>id</i>	<i>binominal</i>	<i>integer</i>	<i>integer</i>	<i>integer</i>
1	1	1	19	15	39
2	2	1	21	15	81
3	3	2	20	16	6
4	4	2	23	16	77
5	5	2	31	17	40
6	6	2	22	17	76
7	7	2	35	18	6
8	8	2	23	18	94
9	9	1	64	19	3
10	10	2	30	19	72
11	11	1	67	19	14
12	12	2	35	19	99
13	13	2	58	20	15
14	14	2	24	20	77
15	15	1	37	20	13
16	16	1	22	20	79
17	17	2	35	21	35
18	18	1	20	21	66

✓ no problems.

← Previous → Next ✕ Cancel

# 基本設計流程

The image displays a data science workflow and its configuration parameters. On the left, a workflow diagram shows two connected components: 'Retrieve C2\_MallCustomers' and 'Clustering'. The first component has an 'out' port, and the second has an 'exa' port. Both components have a green checkmark and a yellow warning icon. The 'Clustering' component has two 'clu' ports and three 'res' ports. On the right, the 'Parameters' panel for the 'Clustering (k-Means)' component is shown. It includes several checkboxes and input fields for configuring the clustering process.

**Parameters**

**Clustering (k-Means)**

- ☒ add cluster attribute ⓘ
- ☐ add as label ⓘ
- ☐ remove unlabeled ⓘ
- k ☒ 5 ⓘ
- max runs 10 ⓘ
- ☒ determine good start values ☒ ⓘ
- measure types ☒ MixedMeasures ⓘ
- mixed measure MixedEuclideanDistance ⓘ
- max optimization steps 100 ⓘ
- ☒ use local random seed ⓘ
- local random seed 1992 ⓘ

# 結果

Result History

ExampleSet (Clustering)

Cluster Model (Clustering)

Data

Statistics

Visualizations

Annotations

Open in

Turbo Prep

Auto Model

Row No.	CustomerID	cluster	Gender	Age	Annual Inco...	Spending Sc...
1	1	cluster_3	1	19	15	39
2	2	cluster_2	1	21	15	81
3	3	cluster_3	2	20	16	6
4	4	cluster_2	2	23	16	77
5	5	cluster_3	2	31	17	40
6	6	cluster_2	2	22	17	76
7	7	cluster_3	2	35	18	6
8	8	cluster_2	2	23	18	94
9	9	cluster_3	1	64	19	3
10	10	cluster_2	2	30	19	72
11	11	cluster_3	1	67	19	14

# 結果

Result History

ExampleSet (Clustering) ×

Cluster Model (Clustering) ×



Description



Folder  
View



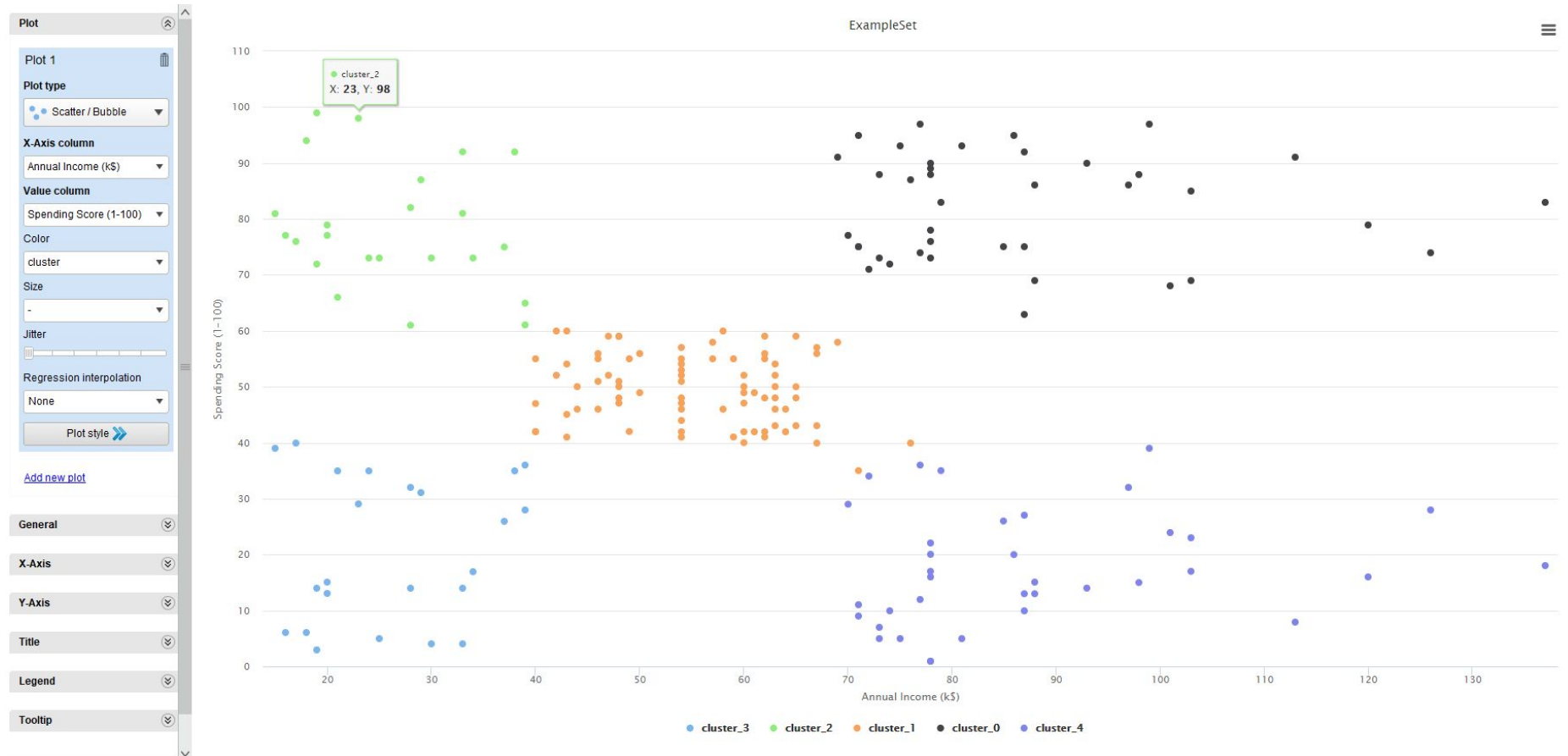
Graph



Centroid  
Table

Attribute	cluster_0	cluster_1	cluster_2	cluster_3
Gender	1.538	1.582	1.609	1.609
Age	32.692	43.089	25.522	45.217
Annual Income (k\$)	86.538	55.291	26.304	26.304
Spending Score (1-100)	82.128	49.570	78.565	20.913

# 結果





# Reference

---

- [大數據驅動商業決策: 13 個 RapidMiner 商業預測操作實務](#)
- [RapidMiner 人工智慧機器學習軟體](#)
- [Data Science course by professor Jia-Ming Chang](#)
- [基礎統計名詞介紹網頁](#)
- [2021 iThome 鐵人賽 - 全民瘋 AI 系列 2.0](#)
- [機器學習演算法 – K 近鄰\(KNN\)](#) (這個是監督式學習, 非 K-Means!)

## Tools

- [ZoomIt - Sysinternals - Microsoft Learn](#)

