作業 4: 關連分析

1. 在關連分析中最為著名的例子不外乎就是啤酒與尿步的例子,您可以簡單說明這是什麼樣的一個故事嗎? (5%)

→ *查一下這故事，大概描述一下。*

Ans:

這個故事產生於 20 世紀 90 年代的美國沃爾瑪超市中，沃爾瑪的超市管理人員分析銷售數據時發現了一個令人難於理解的現象：在某些特定的情況下，「啤酒」與「尿布」兩件看上去毫無關係的商品會經常出現在同一個購物籃中。

- 尿布和啤酒的銷售量有正向關聯性！？

  原來美國的婦女通常在家照顧孩子，所以她們經常會囑咐丈夫在下班回家的路上為孩子買尿布，而丈夫在買尿布的同時又會順手購買自己愛喝的啤酒。

- WalMart 銷售策略

  後來沃爾瑪採取合購策略，啤酒和尿布擺設放在同一區域，意外讓這兩項產品的銷售量提升 30%。

2. 而除了啤酒與尿布的例子外，您還能舉例說明一個關連分析的例子嗎? (10%)

→ *查一下相關故事，大概描述一下。*

- 蕎麥冷麵與納豆

購買蕎麥冷麵的顧客也會同時購買納豆，而這種情況在夏季尤其明顯。據此，超市調整了這兩種商品的陳列位置，將它們放在一起，以提高銷售量。

- 魚肉香腸與麵包

當店裡出售魚肉香腸時，麵包的銷售量也同時增加了。原來顧客喜歡把魚肉香腸放在麵包裡一起食用，因此，麵包店調整了魚肉香腸和麵包的陳列位置，把它們放在一起，提高了銷售量。

- 酸奶與盒飯

當店裡出售盒飯時，酸奶的銷售量也同時增加了。經過調查和分析，原來顧客通常會在吃盒飯時喝酸奶，而這種現象尤其在午餐時間和健身人士中更為明顯。因此，這家餐飲店調整了酸奶和盒飯的陳列位置，把它們放在一起，提高了銷售量。

- 零食與汽水

在一個便利店的銷售數據中，發現購買零食的顧客也會同時購買汽水，因此將這些商品放在一起陳列，提高了銷售量。

太棒了,這樣您就知道至少 2 個例子囉.接下來我們要開始來實做!

在開始之前,我們要先了解一下,其實關連分析是一項　監督式　非監督式　學習（請圈選）　(5%)

所以.這與您之前所使用的流程不太一樣唷!

Ans：非監督式

因為 Association 為非監督式，所以不會有正確答案。

目標是從大數據中，找出那些經常一起出現的東西，EX.商品。然後靠這些結果，總結出關聯規則，以用於後續的商業目的。

3.請開啟 gamestudy.csv. 這是一個關於遊戲研究的資料集, 如同我們在先前課程所述，在這個作業裡我們希望您可以使用 FP-Growth Algorithm 進行關連分析。**只使用 90%的資料(只使用 90%的資料, 只使用 90%的資料)**進行關連分析模型建構，並請將該流程存檔為學號-1.rmp 檔 (15%)

〔首先補充 ETL 資料處理〕

- Part1：填補缺失值

在這個作業，ETL - 填補缺失值的部份我採用跟之前作業不同的方法。我參考網路上的教學資源，利用 Python & Colab 填補缺失值，再用 RapidMiner 讀入檔案。因此來源檔請改用我這邊提供的檔案來讀取。

- 教學資源：

  https://www.kaggle.com/code/eyadamin1233/classify-gamers-mentality/notebook

- 我的程式碼：

  來源壹（.ipynb）:

  國立臺北科技大學_金融大數據_作業4_關聯分析(20230519).ipynb

  來源貳（我的 github）:

  https://github.com/AcerPing/NTUT_FinancialBigData/blob/main/%E5%9C%8B%E7%AB%8B%E8%87%BA%E5%8C%97%E7%A7%91%E6%8A%80%E5%A4%A7%E5%AD%B8_%E9%87%91%E8%9E%8D%E5%A

4%A7%E6%95%B8%E6%93%9A_%E4%BD%9C%E6%A5%AD4_%E9%97%9C%E8%81%AF%E5%88%86%E6%9E%90(20230519).ipynb

來源參：直接開啟資料夾的 ipynb 檔案

gamestudy.csv

- 輸出檔案（csv）：

- Part2：數字資料分群

  因為只能處理文字類的資料，不能處理數字類的資料。必須要先轉成 Binominal。

- 方法一：

  我參考上次教案的投影片，發現 Operator 可以用 Numerical to Binominal + Discretize（離散化），但因為我不會使用，且執行上有問題（會再請教老師及同學），因此作業改用方法二。

  投影片：

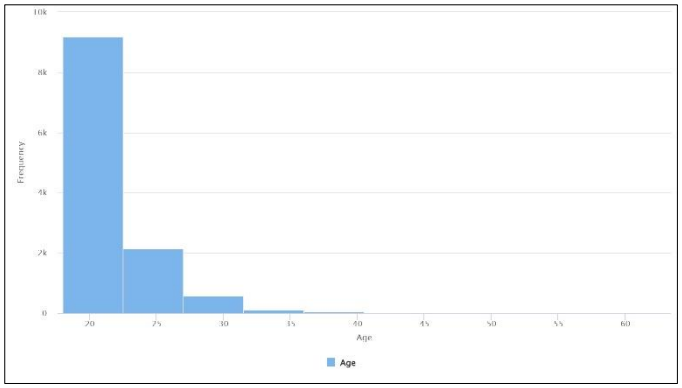
  https://docs.google.com/presentation/d/1gnJ0GO5QJB1ShNLLBWWfjvMgQmroefew/edit#slide=id.p13

- 方法二：

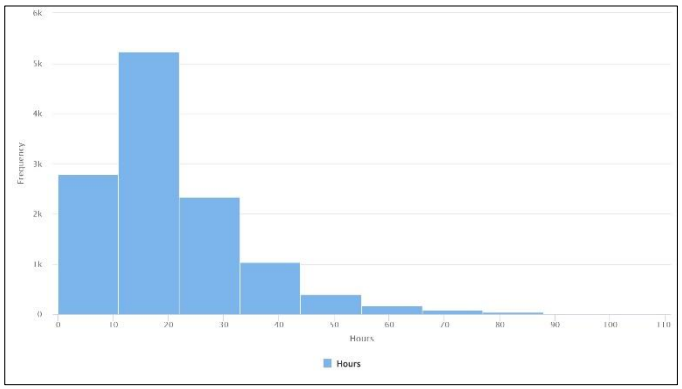
  老師上課有提到，如果數字太多，可以分成級距，比方說 Age → 0~10 歲、10~20 歲...，因此我參考 RapidMiner 的長條圖及數據，再利用

Excel 的 IFS 函數完成，並轉成 xlsx 檔案。

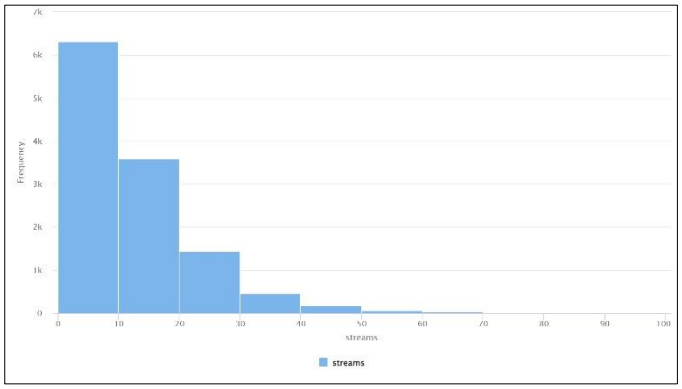
- age



- hours

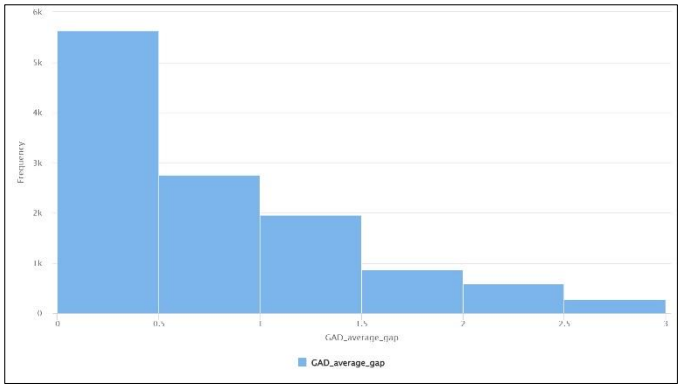


- streams

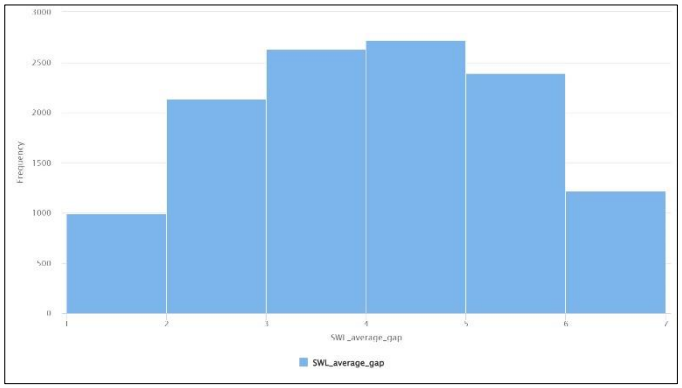


※ 問卷的分數，需要取平均！

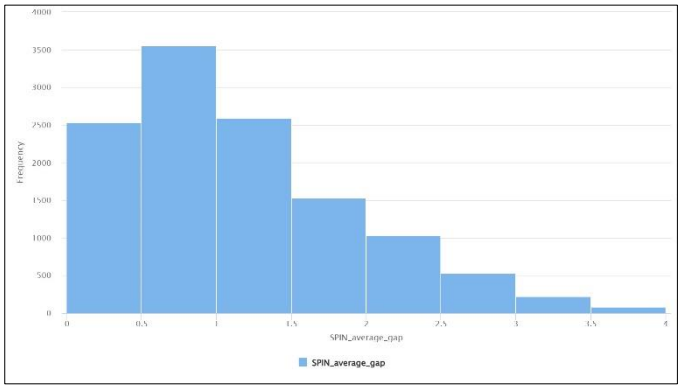
- GAD_average_gap



- SWL_average_gap



- SPIN_average_gap

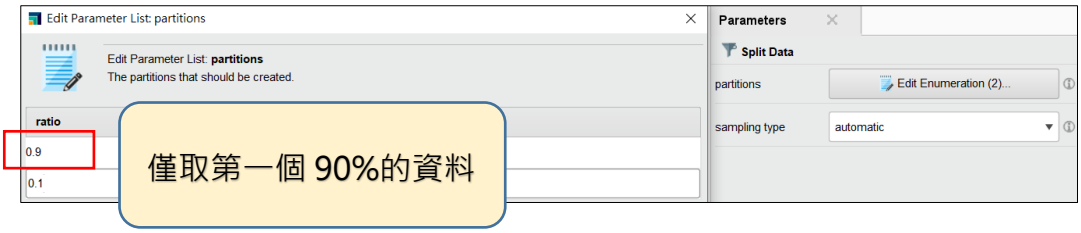


- Narcissism 我在讀取的時候，會先設定成 polynominal。

gamestudy.xlsx

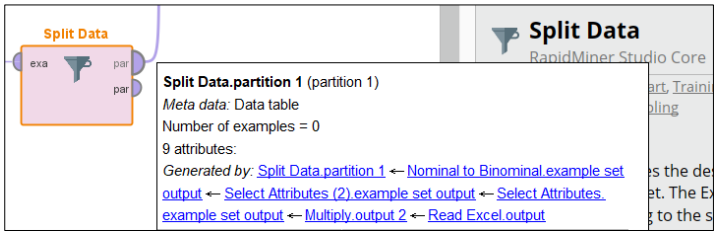
- 輸出檔案 (xlsx) :

- Split Data 已經設定為 0.9



〔留意〕**只使用 90%的資料**

雖然 Split Data Ratio 設定 0.9、0.1，但僅取 Partition 1 的資料進行關連分析模型建構。

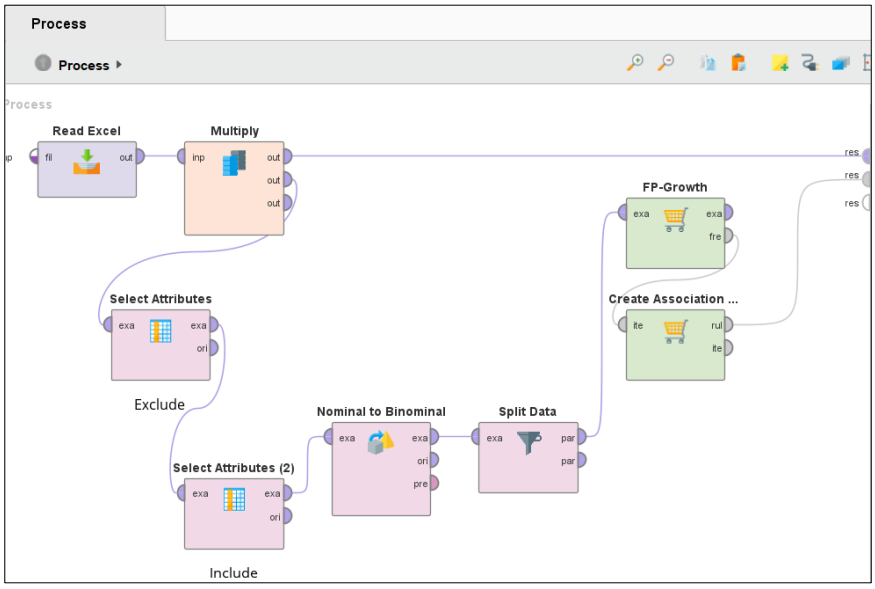


訓練的資料量會是 10,873 筆

（原始資料量 12081 筆× 90%=10,872.9 取整數後為 10,873 筆）

ExampleSet (10,873 examples, 0 special attributes, 58 regular attributes)

- ANS: **關連分析模型建構**



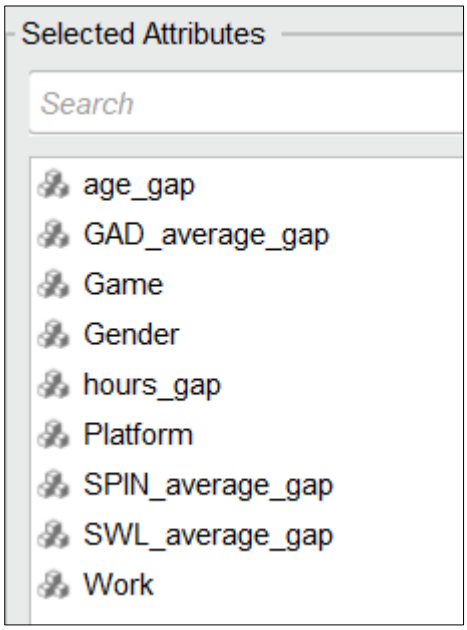
存檔名稱：111C71008_1.rmp

4. 請說明您模型的發現(包括:你留下那些欄位,您期待找到些什麼**關連規則**;您所尋找到的規則是否符合您的預期，如果可以請在 doc 檔中貼上您的規則，以方便作業批改) (15%)

- 留下那些欄位

  age_gap、hours_gap、Game、Gender、Work

  再加上問卷的平均分數級距 GAD_average_gap、SWL_average_gap、SPIN_average_gap



  ※ 要放入 Include 問卷的分數

- 您期待找到些什麼**關連規則**

  Confidence→要超過多少數值才願意推薦

  假設我今天是電玩店家，會先瞭解客人是什麼身分，再推薦這位客人玩哪一種遊戲會比較合適。(重點擺在 Confidence)

關連規則

Premises	Conclusion	Support	Confidence
Platform = PC, Gender = Male, GAD_average_gap = GAD_average為0-0.5	Game = League of Legends	0.379	0.851
Gender = Male, hours_gap = Hours為11-22	Platform = PC, Game = League of Legends	0.350	0.851
Platform = PC, GAD_average_gap = GAD_average為0-0.5	Game = League of Legends	0.392	0.852
Platform = PC, Gender = Male	Game = League of Legends	0.791	0.852
Platform = PC, Gender = Male, SPIN_average_gap = SPIN_average為0.5-1	Game = League of Legends	0.239	0.852
hours_gap = Hours為11-22	Platform = PC, Game = League of Legends	0.369	0.852
Work = Student at college / university, GAD_average_gap = GAD_average為0-0.5	age_gap = age為18-22.5	0.208	0.853
Platform = PC, SPIN_average_gap = SPIN_average為0.5-1	Game = League of Legends	0.247	0.853
Platform = PC	Game = League of Legends	0.838	0.853
Platform = PC, Work = Student at college / university, GAD_average_gap = GAD_averag	age_gap = age為18-22.5	0.205	0.853
Gender = Male, age_gap = age為18-22.5, GAD_average_gap = GAD_average為0-0.5	Game = League of Legends	0.288	0.854
age_gap = age為18-22.5, GAD_average_gap = GAD_average為0-0.5	Game = League of Legends	0.295	0.854
Gender = Male, age_gap = age為18-22.5	Game = League of Legends	0.616	0.854
Work = Student at college / university	age_gap = age為18-22.5	0.456	0.854
Gender = Male, Work = Student at college / university	Game = League of Legends	0.430	0.854
Work = Student at college / university, GAD_average_gap = GAD_average為0-0.5	Platform = PC, Game = League of Legends	0.209	0.855
Work = Student at college / university	Game = League of Legends	0.457	0.855
age_gap = age為18-22.5	Game = League of Legends	0.649	0.855

符合我的預期。

根據結果，假設今天有位男性，有在使用電腦平台玩遊戲（Platform = PC），GAD 問卷平均分數落於 0-0.5，他有在玩《英雄聯盟》的可能性是 0.8512181513855309~0.8517452320978769（可信度 Confidence 為 0.85）。

※ 關聯分析**不是找因果關係**，只是找兩者之間**同時發生**的機率。

5. 最後,我們來進行相關參數調整,請說明 support 值和 confidence 調高及調低的影響對於分析結果有何影響(8%)

Ans:

設定太低的話，會導致關聯分析的結果出現太多的關聯規則，太高的話，關聯規則太少, 都不利參考分析結果做決策。

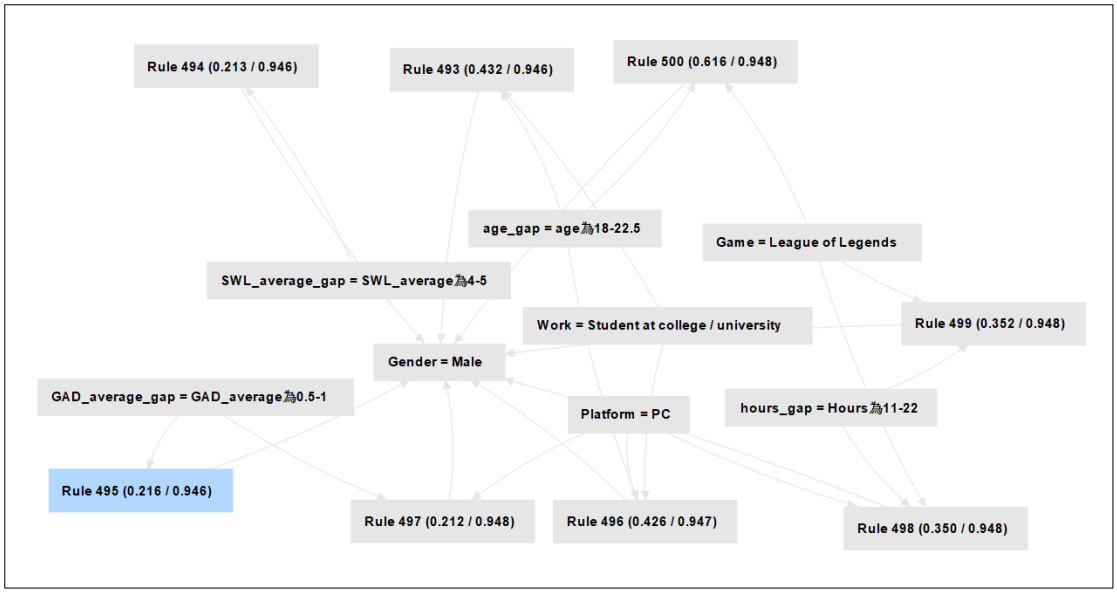
- 支持度（Support）
  - 意義：表示物品集在數據庫中出現的次數比例
  - 參數 min_support 最小支持度：要超過多少數值才會放進來分析，數值介於 0~1 之間，因為不知道實際多少數值，所以用比例%。
  - 因此當我將 Support 調低，則項目越多，規則也就越多；反之當我將 Support 調高，則項目越少，規則也就越少。
  - 如果最小支持度或最小頻率的值設置得太高，算法可能會找到零項集；若支持度太小，說明相應規則可能只是偶然發生，在商業環境中覆蓋太少案例的規則，通常沒有價值。
- 信心水準（Confidence）:
  - 意義：表示兩物品同時出現的條件機率，簡單來說就是在已經出現商品 A 的情況下，出現商品 B 的機率。
  - 參數 min_confidence 最小可信度：要超過多少數值才願意推薦。
  - 當我將 Confidence 調低，則規則越多；

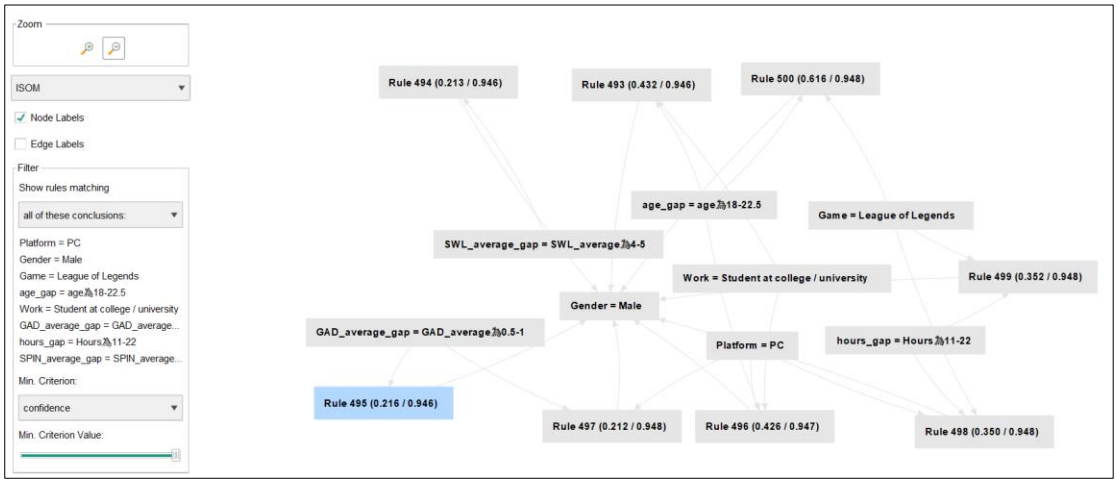
當我將 Confidence 調高，則規則越少。

- 可信度則決定規則的可預測度，若規則的置信度太低，很能從 X 判斷出 Y，在應用上也沒有太多用處。

※ 先解決 min_support 再解決 Confidence

6. 最後,請將拓樸圖畫出 (7%)

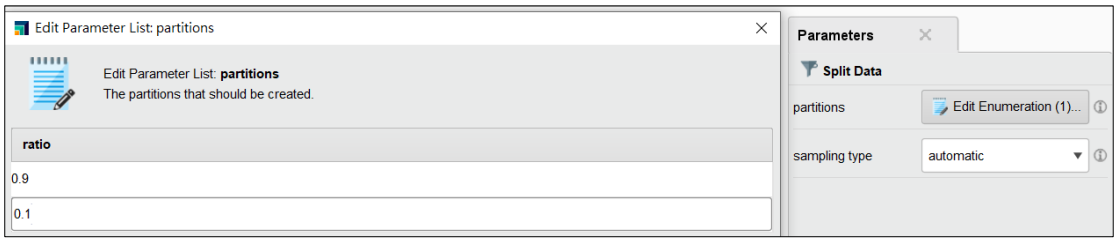




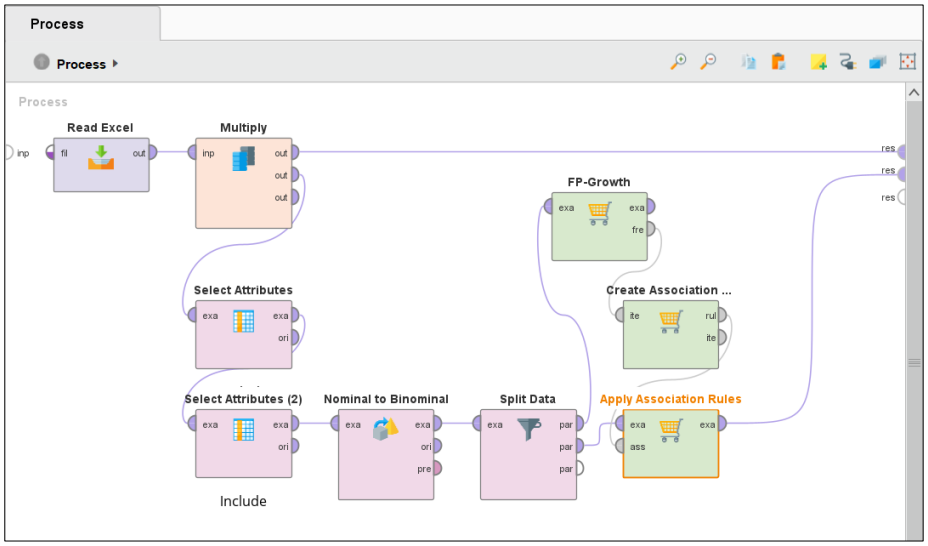
※ 節點放中間，Rules 放四周。

7.請將您模型存成學號後,用剩下的 10%的資料進行驗證,並加以說明您可能的應用結果(15%),並將相關流程存為學號_2.rmp。

- Split Data → 利用剩下 10%資料驗證



- 模型建構



存檔名稱：111C71008_2.rmp

- 結果應用

➢ premises 為 Platform = PC, Gender = Male, GAD_average_gap = GAD_average 為 0-0.5

➢ conclusion 為 Game = League of Legends

➢ support 為 0.3788593659465276

➢ confidence 為 0.8512181513855309

根據關聯規則的結果，可以得出在滿足以下條件時，玩家玩 "League of Legends" 的可能性（信心指數為 0.85）較高：

- 使用的遊戲平台是 PC
- 玩家的性別是男性
- GAD 平均分數在 0-0.5 之間

因此，在進行市場營銷時，可以有針對性地向滿足以上條件的人群進行推廣和廣告宣傳，以吸引更多的潛在顧客。在遊戲的改進和推廣方面，可以針對這個人群的特點進行改進和推廣，以提高遊戲的吸引力和受歡迎程度。

Reference 參考資料

- 讓你的組織注入 AI 魂：關聯分析

https://docs.google.com/presentation/d/1gnJ0GO5QJB1ShNLLBWWfjvMgQmroefew/edit#slide=id.p1

https://docs.google.com/presentation/d/1m8P8aM-YnYnTFc5KxXc7W_arylnBcUVl/edit#slide=id.p2

- 數據挖掘相關聯繫著名案例——啤酒與尿布

https://cloud.tencent.com/developer/article/1105331

- Day 06：購物籃分析背後的演算法 – Apriori

https://ithelp.ithome.com.tw/articles/10218530