國立雲林科技大學

資訊管理研究所

資料探勘專案作業四

指導教授：　許中川　教授

學生：M11223038 陳品佑

M11223032 張祥恩

M11223033 鍾季衡

M11223036 魏冠宇

目錄

摘要.....1

Abstract.....2

一、緒論.....3

1.1 研究動機.....3

1.2 研究目的.....3

二、研究方法.....3

三、研究實驗.....4

3.1 資料集簡介.....4

3.2 前置處理.....5

3.3 實驗設計.....5

3.4 實驗結果.....5

四、結論.....10

五、參考文獻.....13

表目錄

表 1　交易資料集屬性簡介.....4
表 2 交易資料集部份內容.....4
表 3 演算法耗時.....5
表 4 FP-Growth 參數設定與推薦產品數量之關係表.....5
表 5 Apriori 參數設定與推薦產品數量之關係表.....6
表 6 兩種演算法推薦之產品項目.....6
表 7 演算法記憶體佔存.....7

圖目錄

圖 1　Apriori 提升度分析圖 7
圖 2 Apriori 支持度與信心度之點散圖 7
圖 3　FP-Growth 支持度與信心度之點散圖 8
圖 4　FP-Growth 提升度分析圖 8
圖 5 Apriori 之頻繁項目集熱圖 9
圖 6 FP-Growth 之頻繁項目集熱圖 9
圖 7 兩種演算法之性能比較圖 10
圖 8 Apriori 演算法產品共購圖 10
圖 9 FP-Growth 演算法產品共購圖 11
圖 10 Apriori 和 FP-Growth 散點圖 11

摘要

本次專案作業主要目標是透過 Python 進行交易資料的關聯規則分析。資料集包含交易資料，而其中相同的 INVOICE_NO 表示相同的交易紀錄，而 ITEM_ID 和 ITEM_NO 則用於識別交易品項。而在前置處理中需要排除數量為零或含有負值的交易，因為這代表退貨或註銷。在分析過程中，需要設定不同的支持度(s)和信心度(c)，並記錄相應的規則數量和執行時間。而冗餘規則也需要剔除，用於減少規則數量。

此外，本專案還需要做到接受多項產品作為輸入，並利用關聯規則推薦出產品。推薦的產品在同一次推薦中不重複出現。本研究也涵蓋不同參數值設定與推薦產品數量之間的關係。最後，本研究分別使用 Apriori 演算法和 FP-Growth 演算法，並比較兩者在執行時間方面的效能差異。

關鍵字：支持度、信心度、關聯規則、Apriori 演算法、FP-Growth 演算法

Abstract

The main objective of this project is to perform association rule analysis on transaction data using Python. The dataset comprises transaction records, where the same INVOICE_NO denotes identical transaction entries, and ITEM_ID and ITEM_NO are utilized for identifying transaction items. Preprocessing involves excluding transactions with zero quantity or negative values, as these represent returns or cancellations. During the analysis, it is necessary to set different support (s) and confidence (c) levels, recording the corresponding number of rules and execution times. Redundant rules are also eliminated to reduce the total number of rules.

Additionally, the project involves accepting multiple products as input and using association rules to recommend products. Recommended products should not be repeated within the same recommendation session. The study also explores the relationship between various parameter settings and the quantity of recommended products. Finally, the research employs both the Apriori algorithm and the FP-Growth algorithm, comparing their performance differences in terms of execution time.

Keywords: Support, Confidence, Association Rules, Apriori, FP-Growth

一、緒論

1.1 研究動機

本研究因現代商業環境中有著大量且複雜的交易資料，在這些龐大的資料中蘊含著豐富的資訊，然而要從中萃取出有價值的規則卻是相當具有挑戰性的。而在資料探勘中的課程或是介紹上，不免皆會提到關於「啤酒尿布」這樣的案例，而這也的確是一個經典的資料探勘演算法稱為**關聯規則**(Association rules)。那麼本研究藉由關聯規則分析，深入了解不同產品間的潛在關聯性，進而提供企業制定更有效的營銷策略和管理方針。

1.2 研究目的

本研究在於利用 Python 實現 Apriori 演算法和 FP-Growth 演算法，進行交易資料的關聯規則分析。透過不同的支持度和信心度設定，本組將深入探討其中所產生出的規則數量與品質之間的關聯性。同時，剔除掉冗餘規則將會有助於簡化規則之集合，使其更具解釋性和應用價值。此外，本研究也聚焦於推薦系統的應用，透過關聯規則推薦產品，並在推薦中避免重複產品的產生。這些在未來的實際應用場景上將有助於提高企業對交易資料的深入理解，進而優化並有效調整行銷方案，提升營運效能。

二、研究方法

本研究利用課程中提供的交易資料集進行資料分析。前置處理部份包括將資料集中的相同 INVOICE_NO 視為同一筆交易，使用 ITEM_ID 與 ITEM_NO 識別交易項目以及排除數量為零或負值之交易。前置處理結束後，本研究套入 **Apriori** 演算法以及 **FP-Growth** 演算法，實作關聯規則的挖掘，並且透過調整不同**支持度**和**信心度**來分析規則數量、品質及演算法之執行時間。最後透過本組所撰寫的推薦程式，基於關聯規則來推薦其產品並且分析兩個演算法在不同設定下的效能表現及推薦系統的結果。

三、研究實驗

3.1 資料集簡介

本研究使用課程專案作業提供的交易資料集做為本次研究主要資料集。這是一個包含多筆交易記錄的資料集合，每筆交易都由獨特的 INVOICE_NO 識別。INVOICE_NO 可視為交易的唯一編號，代表一次完整的交易。每項產品都有其專屬的 ITEM_ID 和 ITEM_NO，用於確認交易中的商品種類。在進行資料前處理時，所有數量為零或負值的交易會被視為退貨或註銷，因此在分析階段將被排除。下表 1, 2 為本次實驗用資料集的屬性簡介以及部份內容。

- 資料集名稱：交易資料集
- 資料筆數：157396 筆資料
- 屬性數量：7 種屬性

表 1 交易資料集屬性簡介

屬性名稱	型態	尺度
INVOICE_NO	String	Nominal
CUST_ID	Int	Nominal
ITEM_ID	Int	Nominal
ITEM_NO	Int	Nominal
PRODUCT_TYPE	String	Nominal
TRX_DATE	Date	Interval
QUANTITY	Int	Ratio

表 2 交易資料集部份內容

INVOICE_NO	CUST_ID	ITEM_ID	ITEM_NO	PRODUCT_TYPE	TRX_DATE	QUAN-TITY
CX473482-03	3218	3217532	M25P40--VMN6TPB	MEMORY_E-MBEDED	2016/7/26	2500
CX473465-22	2470	3326781	AU80610-006237AASLBX9	CPU / MPU	2016/7/11	50
CX473485-34	16135	740487	MMBD28-37LT1G	DISCRETE	2016/7/27	3000

3.2 前置處理

本研究在進行分析前，首先載入需要的演算法套件，以及檢查資料中是否含有零值以及負值，若含有其中之資料，本組則將視為退貨或註銷並將其從資料中排除。

3.3 實驗設計

透過 Python 實現 Apriori 演算法和 FP-Growth 演算法，以進行關聯規則分析。調整不同支持度和信心度的設定，本研究將挖掘出不同的交易模式和相關性規則。同時，為了優化分析結果，將冗餘規則剔除，以減少規則集合的複雜度。$\exists X' \subset X \ \ conf(X' \to Y) > conf(X \to Y)$，而$X' \to Y$為冗餘規則。最後做兩種演算法之效能表現及推薦系統的結果。

3.4 實驗結果

以下在經過了 Apriori 演算法和 FP-Growth 演算法分別套進此交易資料集當中並進行前置處理後，開始讀入各別的關聯規則進行推薦系統的測試，下表 3 之結果顯示兩個演算法在耗時上的表現。而本組在分析兩種不同演算法的不同支持度和信心度的設定下與推薦產品數量多寡之間的關係呈現如下表 4 及表 5。

表 3 演算法耗時

演算法	耗時(second)
Apriori Algorithm	2.8036
FP-Growth Algorithm	17.2681

表 4 FP-Growth 參數設定與推薦產品數量之關係表

	支持度(s)	信心度(c)	推薦產品數量
FP-Growth	**0.001**	**0.3**	**10**
		0.4	**10**
		0.5	**10**
	0.005	**0.3**	**10**
		0.4	**10**
		0.5	**10**
	0.007	**0.3**	**10**
		0.4	**10**
		0.5	**10**

表 5 Apriori 參數設定與推薦產品數量之關係表

	支持度(s)	信心度(c)	推薦產品數量
Apriori	**0.001**	**0.3**	**10**
		0.4	**10**
		0.5	**10**
	0.005	**0.3**	**10**
		0.4	**10**
		0.5	**10**
	0.007	**0.3**	**10**
		0.4	**10**
		0.5	**10**

此外本組認為在支持度設定的影響下，較高的支持度將產生較為一般性的規則，而較低的支持度可能導致產生過多且特化的規則。調整支持度可以影響挖掘到的關聯規則的數量和特定性；而信心度的部份則是較高的信心度要求規則更具有可信度，但可能減少挖掘到的規則數量。較低的信心度則可能導致挖掘到的規則較不可信。在下表 6 則為 Apriori 演算法與 FP-Growth 演算法各別推薦出來的產品項目。下圖 1, 3 為兩種演算法在支持度與信心度之點散圖，主要將兩者之間的關係可視化。接著，本組也在利用了一種在評估關聯規則效果指標上的提升度分析圖(Lift Chart)來呈現如下圖 2, 4。提升度曲線描述模型在不同信心度閾值下的提升度變化，通常隨著信心度的增加，提升度也會跟著增加，提升度曲線斜率越大，代表該模型在高信心水準上的效能較好。

表 6 兩種演算法推薦之產品項目

	ITEM_ID		ITEM_ID
Apriori Algorithm	1697	FP-Growth Algorithm	23004
	3336149		15192100
	14675955		14671860
	70509		14481833
	88411		14980086
	135493		14870535
	88444		88622
	15192100		14752201
	14876857		88411
	14481833		70509

表 7 演算法記憶體佔存

演算法	記憶體佔存(MB)
Apriori Algorithm	20684
FP-Growth Algorithm	3.56

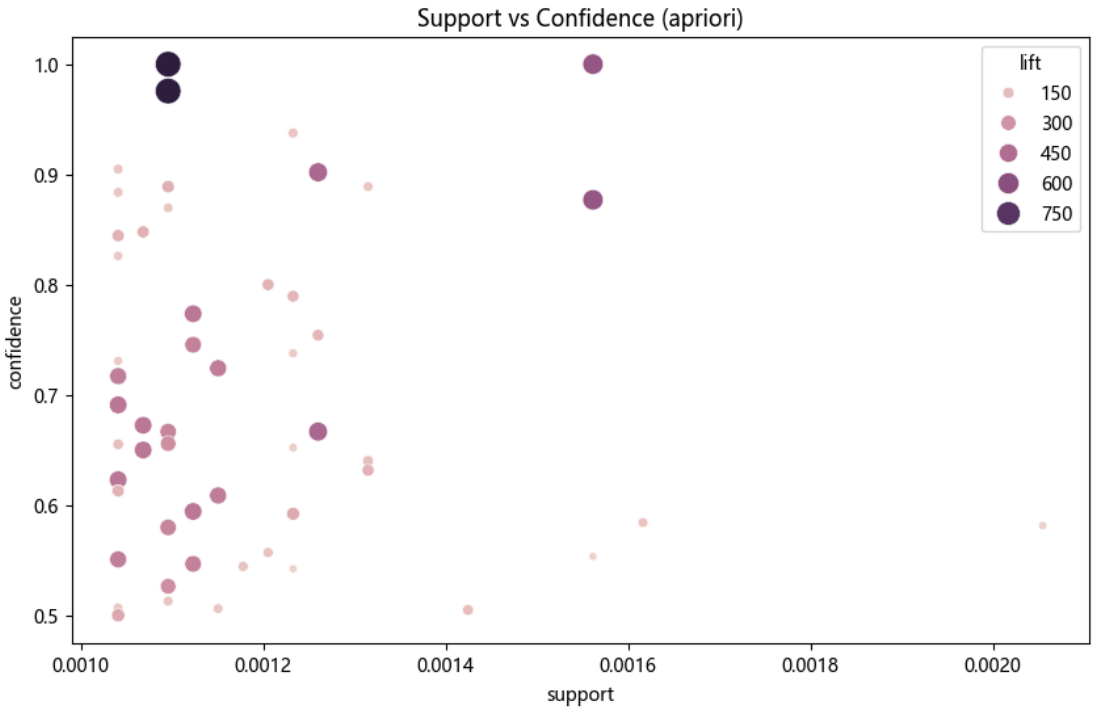

圖 2 Apriori 支持度與信心度之點散圖

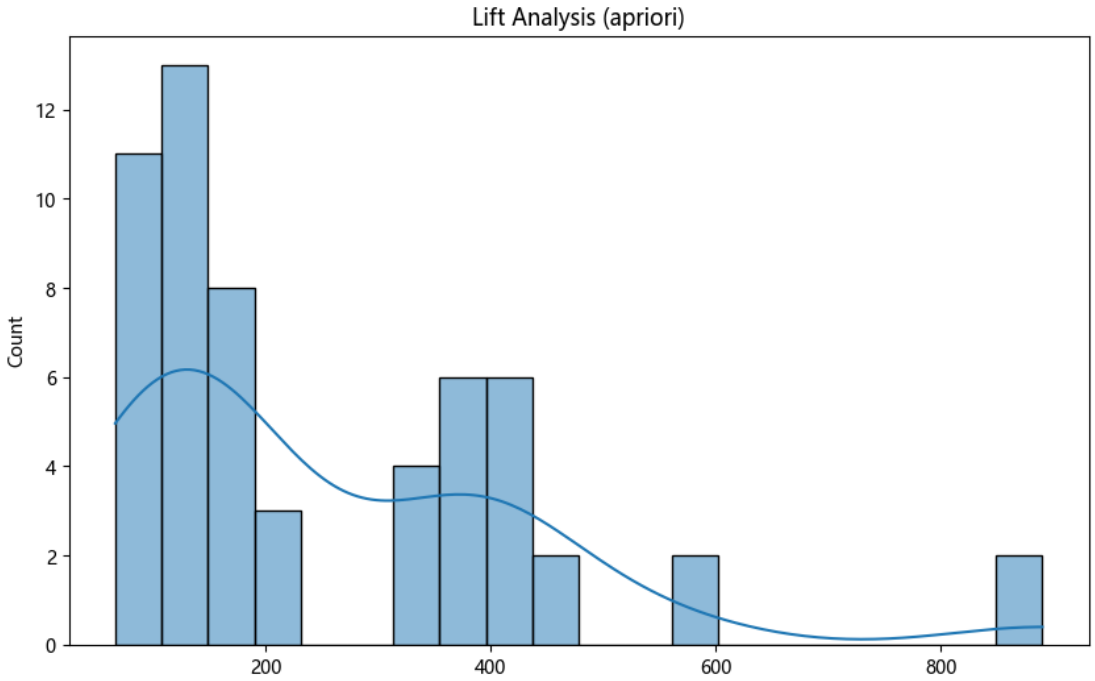

圖 1 Apriori 提升度分析圖

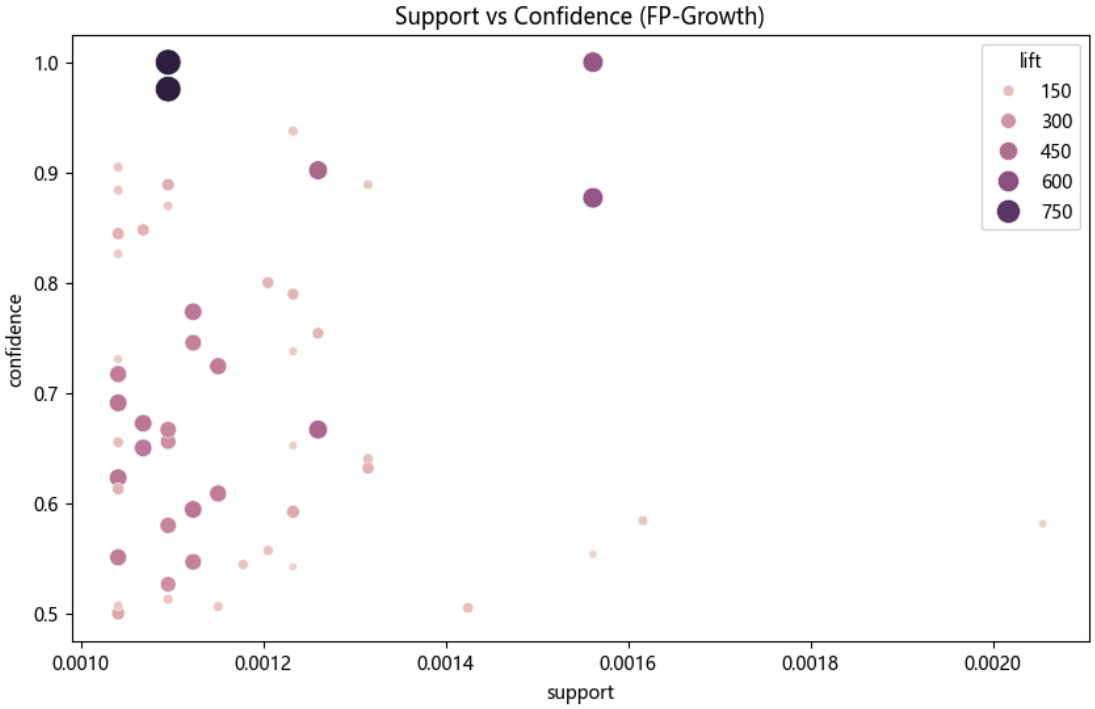

圖 3　FP-Growth 支持度與信心度之點散圖

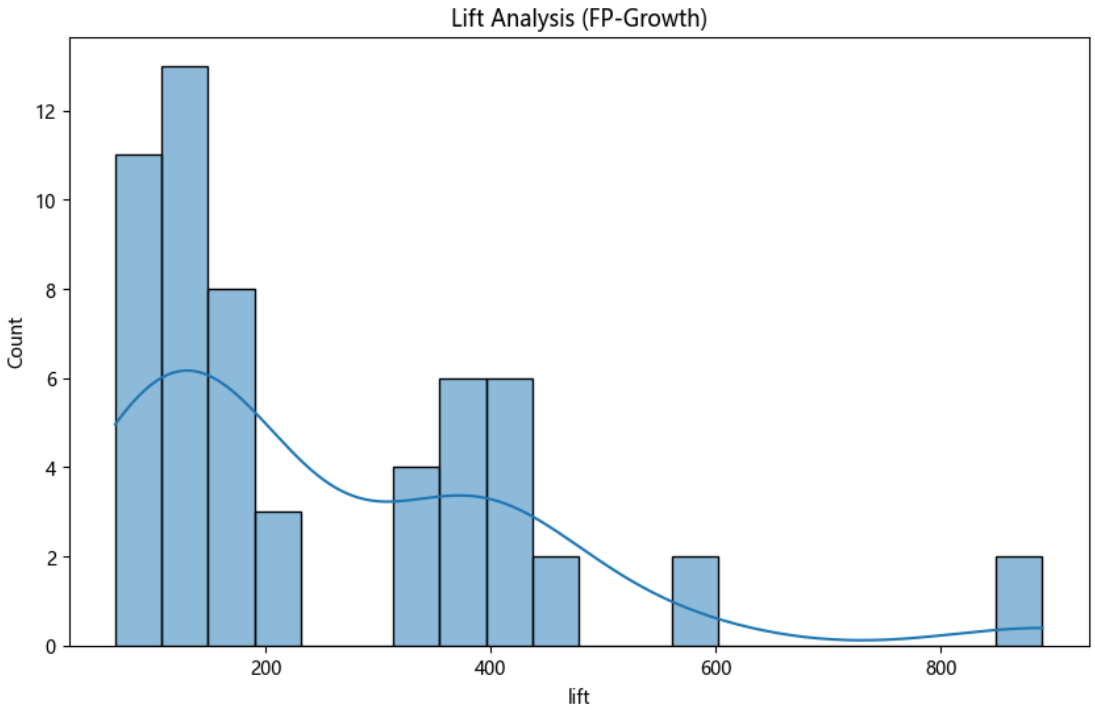

圖 4　FP-Growth 提升度分析圖

接下來，本組透過頻繁項目集熱圖，分別產生 Apriori 演算法與 FP-Growth 演算法兩種集熱圖來呈現資料集中的頻繁項目集之間的關係和出現的頻率在關聯規則分析中，頻繁項目集意旨經常一同出現的一組項目，而集熱圖可以用視覺化的方式展示這些項目集的關聯程度。如下圖 5, 6 所示。

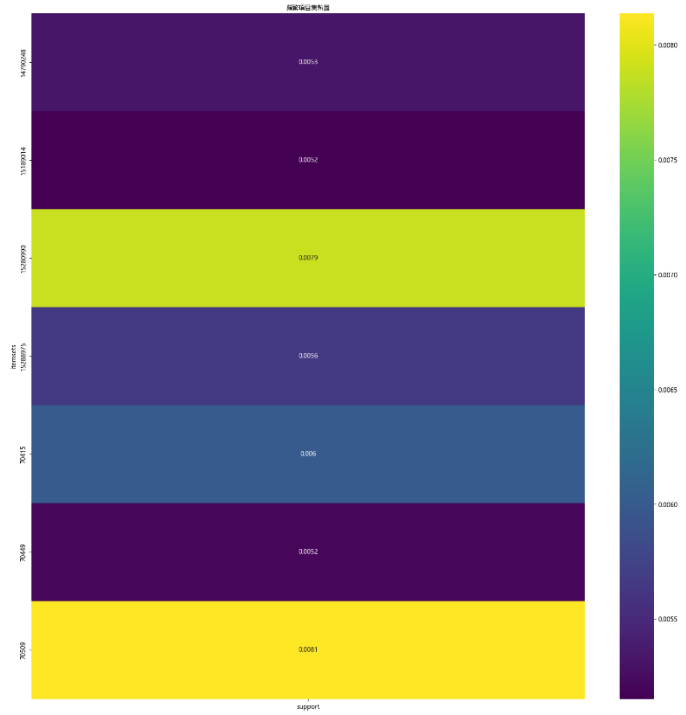

圖 5 Apriori 之頻繁項目集熱圖

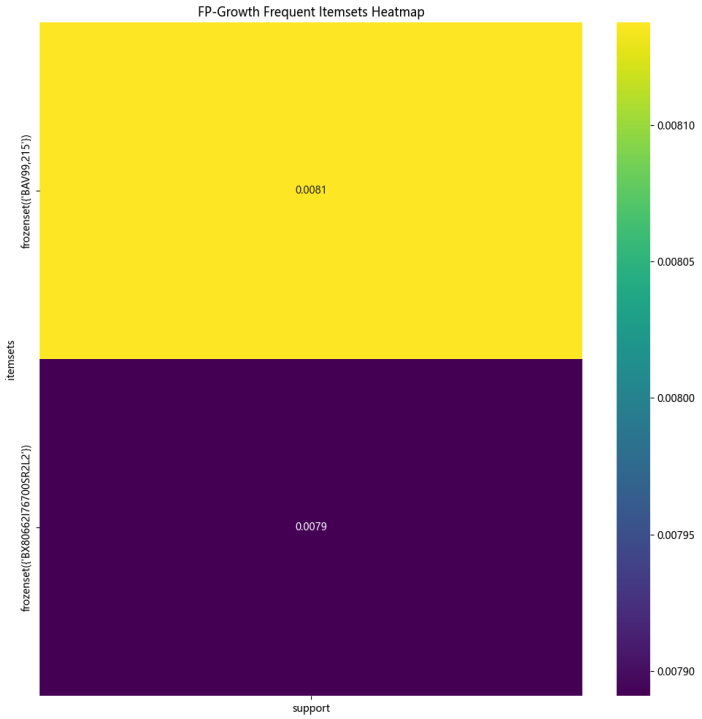

圖 6 FP-Growth 之頻繁項目集熱圖

本組對 Apriori 演算法和 FP-Growth 演算法做兩者之間的性能比較如下圖 7 所示。

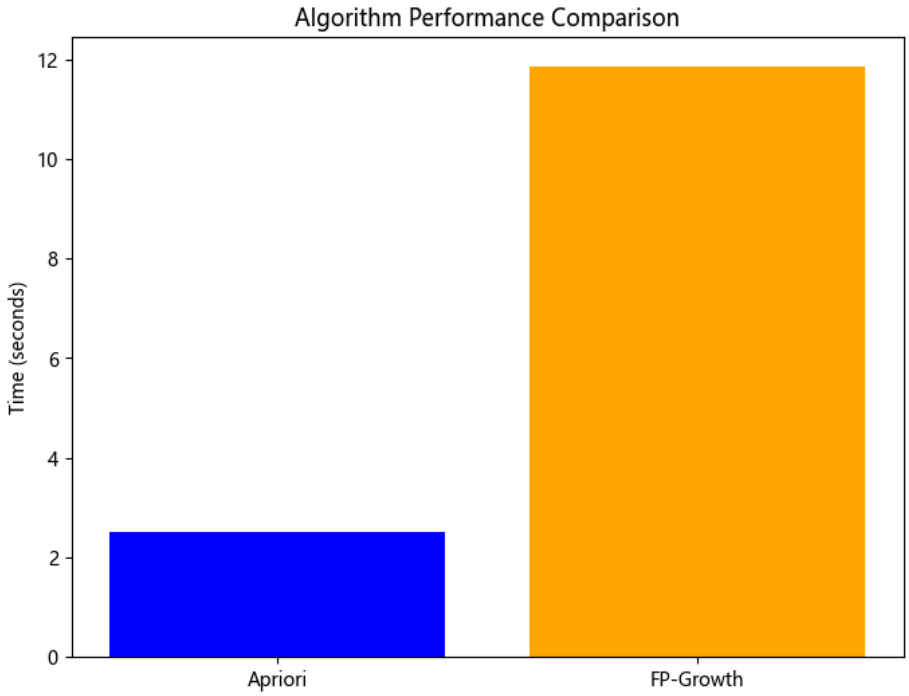

圖 7 兩種演算法之性能比較圖

本組對 Apriori 演算法取出 10 個推薦產品的共購網路圖如下圖 8 所示。

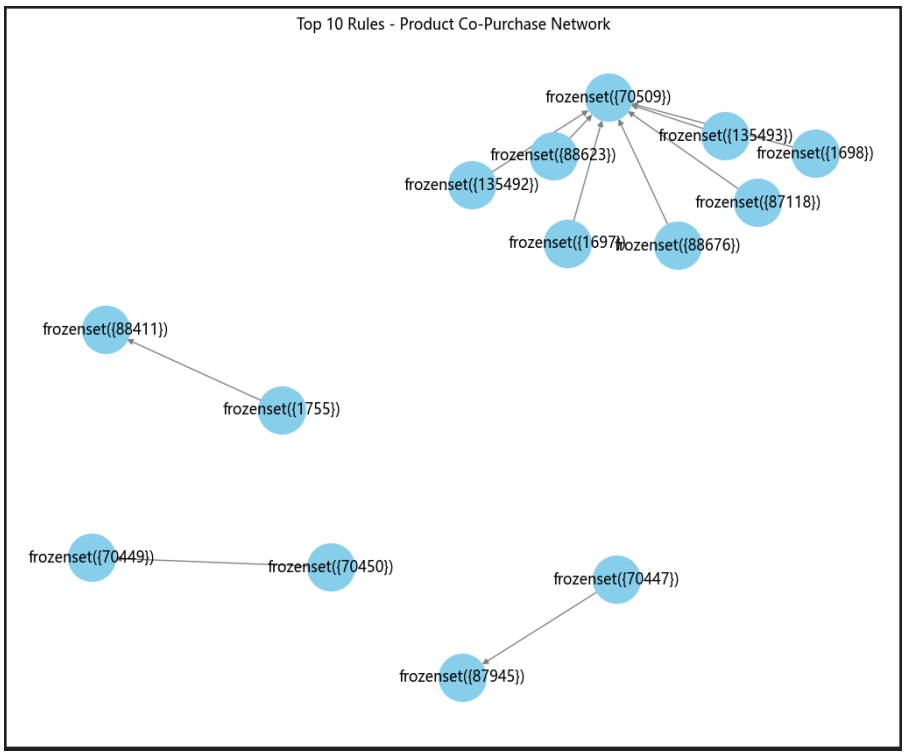

圖 8 Apriori 演算法產品共購圖

本組對 FP-Growth 演算法取出 10 個推薦產品的共購網路圖如下圖 9 所示。

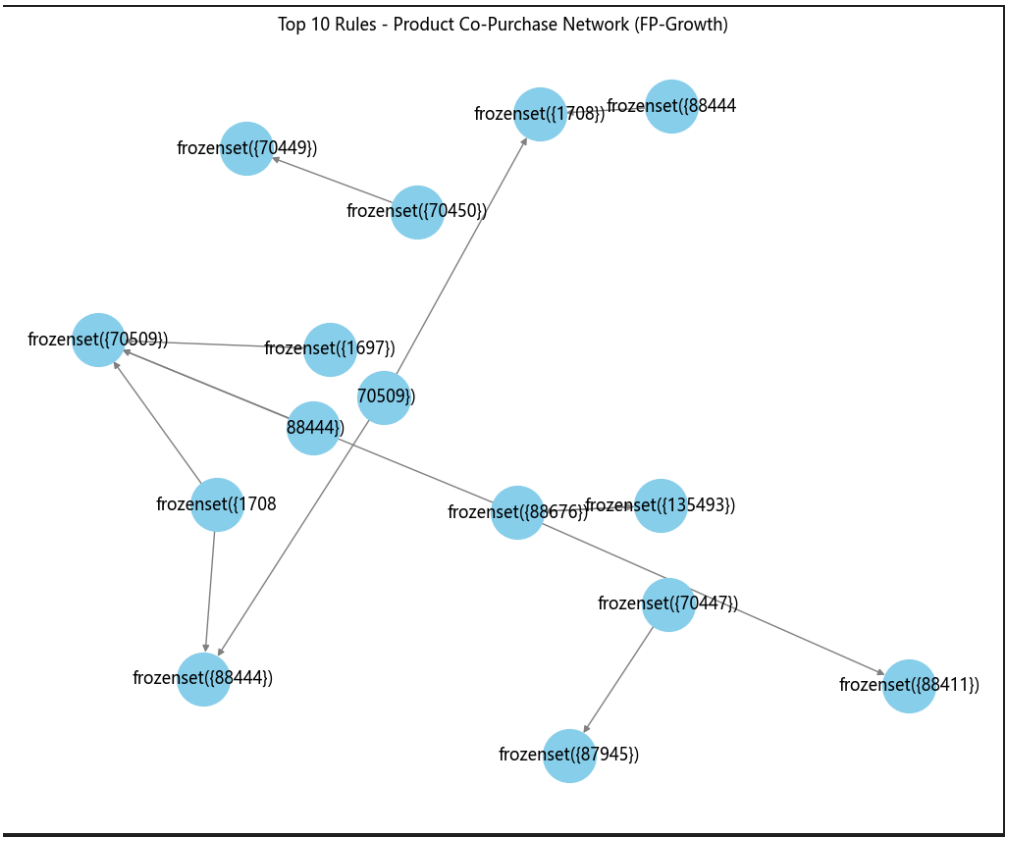

圖 9 FP-Growth 演算法產品共購圖

最後本組對 Apriori 演算法和 FP-Growth 演算法對於支持度與信心度的散點圖如下圖 10 所示。

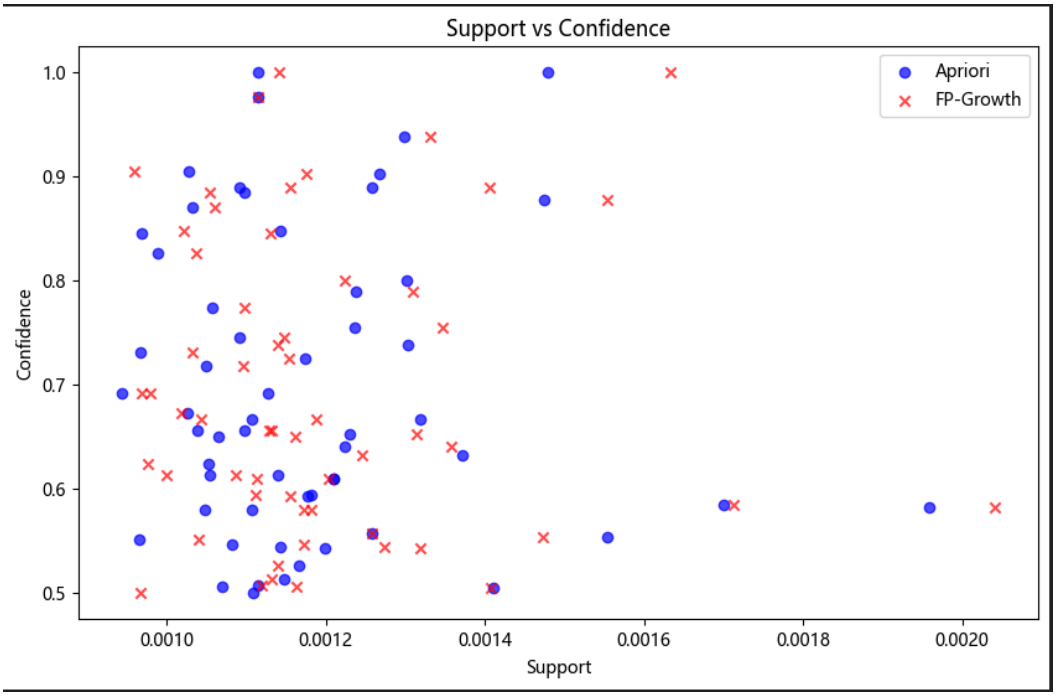

圖 10 Apriori 和 FP-Growth 散點圖

四、結論

在本次的研究中，我們著重於使用 Python 進行交易資料的關聯規則分析，採用 Apriori 演算法和 FP-Growth 演算法進行探勘。其主要目的是深入了解交易資料中的關聯模式以及學習兩個演算法的實際操作，便於未來在企業中透過關聯規則分析，其可以更好了解產品之間的關係，提供企業一個完整的分析來達到更好的營銷策略。然而，推薦系統的建立能夠更進一步豐富了客戶體驗，使企業能夠更智慧地做出決策，迎接競爭激烈的商業環境。本組也相信這項研究是在促進資料科學領域於企業行銷中的應用和發展。

五、參考文獻

Anıl Çoğalan (2023/5/22). FP-Growth Algorithm: How to Analyze User Behavior and Outrank Your Competitors. https://medium.com/@anilcogalan/fp-growth-algorithm-how-to-analyze-user-behavior-and-outrank-your-competitors-c39af08879db

Agnes. Python 實戰篇：Apriori Algorithm (Mlxtend library)。https://artsdatascience.wordpress.com/2019/12/10/python-%E5%AF%A6%E6%88%B0%E7%AF%87%EF%BC%9Aapriori-algorithm/

Apriori algorithm. https://en.wikipedia.org/wiki/Apriori_algorithm

FP Growth Algorithm in Data Mining. https://www.javatpoint.com/fp-growth-algorithm-in-data-mining

JackHCC. (2021). Apriori-and-FP_Growth. https://github.com/JackHCC/Apriori-and-FP_Growth

Max. (2020/4/10). [關聯分析] Apriori 演算法介紹 (附 Python 程式碼)。https://www.maxlist.xyz/2018/11/03/python_apriori/

Walter Chiu. (2020/1/9). 手把手程式實作分享系列：先驗演算法（Apriori Algorithm）關聯規則分析。https://reurl.cc/kr7QMx

赵赵赵颖(2020/1/6)。機器學習筆記(11)－ 關聯分析之 Apriori 演算法原理和 Python 實現。https://blog.csdn.net/leaf_zizi/article/details/103804737