

國立臺東大學資訊管理學系
環境經濟資訊管理碩士在職專班

碩士論文



In-Service Master Degree Program in Environmental Economics and
Information Management

Department of Information Science and Management Systems

National Taitung University

Master Thesis

應用資料採礦模型於新抗凝血劑用藥
出血預測之研究

A Study of NOAC Bleeding Event Prediction
using Data Mining Models

研究生：莊翔甯

Graduate Student: Hsiang-Ning Chuang

指導教授：謝明哲博士

Advisor: Ming-Che Hsieh, Ph.D.

中華民國 108 年 7 月

July, 2019

國立臺東大學資訊管理學系
環境經濟資訊管理碩士在職專班

碩士論文



In-Service Master Degree Program in Environmental Economics and
Information Management

Department of Information Science and Management Systems

National Taitung University

Master Thesis

應用資料採礦模型於新抗凝血劑用藥
出血預測之研究

A Study of NOAC Bleeding Event Prediction
using Data Mining Models

研究生：莊翔甯

Graduate Student: Hsiang-Ning Chuang

指導教授：謝明哲博士

Advisor: Ming-Che Hsieh, Ph.D.

中華民國 108 年 7 月

July, 2019

國立臺東大學 學位論文考試委員審定書

系所班：環境經濟資訊管理碩士在職專班

本班 莊翔甯 君所提之論文應用資料採礦模型於新抗凝血劑
用藥出血預測之研究

業經本委員會通過合於

☒ 碩士學位論文

☐ 博士學位論文

條件

論文學位考試委員會：

張永明

(學位考試委員會召集人)

張永明

張永明

(指導教授)

論文學位考試日期：108 年 7 月 16 日

國立臺東大學

附註：本表一式二份經學位考試委員會簽名後，分別送交系所辦公室及註冊組存查。

國立臺東大學博、碩士學位論文授權書

本授權書所授權之論文為本人在 國立臺東大學 資訊管理系(所)

環境經濟資訊管理碩士在職專班 107 學年度第 2 學期取得 碩 士學位之論文。

論文名稱：應用資料採礦模型於新抗凝血劑用藥出血預測之研究

本人具有著作財產權之論文全文資料，授權予下列單位：

同意	不同意	單 位
☒	☐	國家圖書館
☒	☐	本人畢業學校圖書館
☒	☐	與本人畢業學校圖書館簽訂合作協議之資料庫業者

得不限地域、時間與次數以微縮、光碟或其他各種數位化方式重製後散布發行或上載網站，藉由網路傳輸，提供讀者基於個人非營利性質之線上檢索、閱覽、下載或列印。

☒同意 ☐不同意 本人畢業學校圖書館基於學術傳播之目的，在上述範圍內得再授權第三人進行資料重製。

電子檔公開時程

立即公開	一年後公開	二年後公開	三年後公開
			☒

本論文因申請專利或發表(未申請者本條款請不予理會)需延後公開，依「教育部100年7月1日臺高(二)字第1000108377號函」，延後公開合理期限最多不超過五年。

☐ 本論文已申請專利並檢附證明，專利申請案號：_____；

☐ 本論文準備以上列論文投稿期刊。

延後公開日期：民國_____年_____月_____日。

上述授權內容均無須訂立讓與及授權契約書。依本授權之發行權為非專屬性發行權利。依本授權所為之收錄、重製、發行及學術研發利用均為無償。上述同意與不同意之欄位若未勾選，本人同意視同同意授權。

指導教授姓名：謝明哲 (親筆簽名)

研究生簽名：莊翔宇 (親筆正楷)

學 號： 4310611 (務必填寫)

日 期：中華民國 108 年 7 月 29 日

1. 本授權書(得自 http://portal.lib.nttu.edu.tw/lp.asp?CtNode=714&CtUnit=148&BaseDSD=52&mp=1&xq_xCat=10 下載)請以黑筆撰寫並影印裝訂於書名頁之次頁。

2. 依據 91 學年度第一學期第一次教務會議決議：研究生畢業論文「至少需授權學校圖書館數位化，並遲於三年後上載網路供各界使用及校內瀏覽。」

授權書版本:2018/03/16

謝辭

二年的碩班學習時間竟然如此的飛逝，從剛入學的懵懂到順利寫出自己的論文，首先感謝指導教授謝明哲博士的引領、耐心的教導及在學術實務經驗上的傾囊教授，對研究內容指正與教導督促，以及開拓了我在國際學術交流的視野。也要感激口試委員吳仁和教授、張永明副校長，對論文提供專業上精闢意見，使得論文得以更為嚴謹及完整。

在這兩年周末學生生活中，與班上同學共同學習彼此鼓勵及相持，班導王老師及各科老師的耐心教導中飛逝。感謝同學們及老師們讓我在這段學習過程充滿歡樂中順利完成學業，心中充滿感激及不捨之情。

感謝我的好友及家人，支持並協助我完成學業。

最後謹以此篇論文感謝曾經幫助、鼓勵過我的朋友們。

莊翔甯 謹誌於

國立臺東大學環境經濟資訊管理碩士在職專班

民國 108 年 7 月

摘 要

因台灣以人口老化指數逐年增加，正式進入「高齡社會」。對增加社會照護成本及醫療支出最直接的是人口年齡結構的改變。在高齡人口中常出現心房顫動等心臟疾病問題若控制不良，就容易引發腦中風，腦中風存活後多數導致程度不一的肢體障礙使其失去自行處理日常生活的自我照顧能力。由於 Af 容易產生血栓，而這些血栓如果被發生在腦部就可能引起腦部中風，因此抗血栓治療在 Af 的病人在預防中風和栓塞上是非常重要的。

本研究主要提出一預測模型，供臨床醫師在開立抗凝血藥物前，提供評估病患個人化之出血風險機率，預先提供警訊，提供臨床醫師進行合適的病症處置減少出血事件所造成的健康損害，提高用藥的安全性。

本研究收集某地區醫院收集使用 Rivaroxian 之患有 AF 病人相關資料，利用資料探勘建立 4 種預測模型加以比較後，選取最佳性模型為決策樹預測出血模型。決策樹預測出血模型顯示，整體分類正確率為 77%，AUC 分數為 0.84，顯示該模型區辨能力良好。從決策樹模型路徑規則可得知，影響預測出血事件的因子的重要性即其相依性。

本研究主要貢獻為應用資料採礦技術提供臨床醫師進行抗凝治療前之決策支援，提供用藥前之個人化安全性預測，提高抗凝治療之安全性及準確率，降低中風和栓塞發生風險。

關鍵字：資料探勘，決策樹，心房顫動，新型口服抗凝血劑

Abstract

As Taiwan's population aging index increases year by year, it formally enters the "aging society". The most direct way to increase the cost of social care and medical expenditure is to change the age structure of the population. AF (Atrial Fibrillation) and other heart disease problems often occur in the elderly population. If they are not well controlled, they are prone to stroke, and most of them lose their places due to limb disorders of varying degrees after stroke survives. Self-care ability in daily life. Because AF is prone to thrombosis, which may cause stroke if it occurs in the brain, antithrombotic therapy is very important in preventing stroke and embolism in AF patients.

In this study, we propose a prediction model for clinicians to evaluate the individual risk of bleeding, provide warning in advance, and provide clinicians with appropriate disease management to reduce the health damage caused by bleeding events and improve the safety of medication before the establishment of anticoagulant drugs.

The prediction bleeding model of the study showed that the overall classification accuracy was 77% and the AUC score was 0.84, which showed that the model area was good in discernment. From the decision tree model path rules, we can see that the importance of the factors that influence the prediction of bleeding events is their dependence.

Keywords: atrial fibrillation, data mining, decision tree, new oral anticoagulants, rivaroxban

目次

摘 要	i
目次	iii
表次	v
圖次	vii
第一章 緒論	1
1.1 研究動機	1
1.2 研究目的	5
1.3 論文架構	5
第二章 文獻探討	6
2.1 心房顫動疾病簡介	6
2.1.1 流行病學	6
2.1.2 心房顫動分類	6
2.1.3 臨床表現	8
2.1.4 治療原則	8
2.2 心房顫動與血栓性栓塞	9
2.2.1 抗血栓劑在心房顫動患者的應用	9
2.2.2 心房顫動患者抗血栓劑國際準則建議	9
2.2.3 RIVA 出血危險性	10
2.2.4 抗血栓劑 RIVA 臨床應用現況	11
2.3 資料探勘	12
2.3.1 資料探勘的定義	12
2.3.2 資料採礦的方法	14
第三章 研究方法	22
3.1 研究方法與步驟	22
3.2 資料來源及研究變項	23

3.2.1 資料來源.....	23
3.2.2 研究變項.....	25
3.3 資料前處理.....	25
3.3.1 資料清理.....	25
3.3.3 資料整合轉換.....	25
3.4 軟硬體與採礦工具說明.....	26
3.5 模型評估.....	27
第四章 研究結果	31
4.1 資料淨化、整理.....	31
4.2 資料探勘模型建置與執行.....	40
4.2.1 建立資料來源.....	40
4.2.2 新增採礦結構及設定參數.....	41
4.2.3 結果分析.....	51
4.3 評估比較分析模型.....	61
4.3.1 資料採礦增益圖比較.....	62
4.3.2 資料採礦評估比較.....	66
4.4.3 資料採礦結果與 HAS-BLED Score 比較.....	68
第五章 結論與建議	70
5.1 研究結論與貢獻.....	70
5.2 建議.....	71
參考文獻	73
中文.....	73
英文.....	75

表次

表 2.1 出血風險預測	10
表 3.1 研究資料分布情形	錯誤! 尚未定義書籤。
表 3.2 Microsoft Visual Studio 2017 所提供之資料採礦演算法比較.....	27
表 3.3 AUC 值與模型區辨能力關係表.....	28
表 3.4 錯差矩陣表	28
表 4.1 原始資料欄位說明	31
表 4.2 淨化、處理後資料欄位說明	35
表 4.3 EXCEL 淨化 整理後的資料表	40
表 4.4 SQL Server Business Intelligence Developmnt Studio 提供資料探 勘演算法	42
表 4.5 變數資料類型定義	44
表 4.6 SQL Server Business Intelligence Developmnt Studio 定義之資料類 型	45
表 4.7 研究中使用關聯分析演算法之參數設定說明	48
表 4.8 研究中使用決策樹演算法之參數設定	49
表 4.9 研究中使用貝氏機率分類演算法之參數設定	50
表 4.10 研究中使用類神經網路演算法之參數設定	51
表 4.11 關聯分析預測模型之輸入變數預測能力強弱	52
表 4.12 決策樹模型路徑規則表	55
表 4.13 決策樹預測模型各輸入變數預測能力強弱	57
表 4.14 貝氏機率分類預測模型各輸入變數預測能力強弱	58
表 4.15 四種預測模型之整體分類正確率表	65
表 4.16 四種預測模型之遺漏性誤差與過度預測誤差表	65
表 4.17 四種模型的預測比較表	66

表 4.18 影響目標變數較顯著變數表.....	67
表 4.19 HAS-BLED Score 出血機率	68



圖次

圖 1.1 人口三段年齡組人數.....	1
圖 1.2 老化指數.....	2
圖 1.3 十大死因死亡率-106 年 vs. 96 年.....	2
圖 1.4 歷年 65 歲以上三大主要死因死亡率分.....	3
圖 2.1 KDD 形成過程.....	13
圖 2.2 神經元構造.....	16
圖 2.3 生物神經元示意圖.....	17
圖 2.4 類神經網路模型.....	18
圖 2.5 決策樹模型.....	21
圖 3.1 研究流程.....	22
圖 3.2 增益圖範例.....	27
圖 3.3 預測分類矩陣.....	30
圖 4.1 模型訓練資料表匯進資料庫資料.....	41
圖 4.2 模型測試用資料表匯進資料庫資料.....	41
圖 4.3 指定分析所使用的資料行.....	43
圖 4.4 選擇資料採礦技術.....	45
圖 4.5 本研究所使用變數類型設定.....	46
圖 4.6 設定保留測試資料為 0%.....	47
圖 4.7 本研究所建置模型設定.....	47
圖 4.8 關聯分機預測模型之相依性網路圖.....	52
圖 4.9 項目檢視器檢視項目集.....	53
圖 4.10 規則檢視器檢視規則.....	54
圖 4.11 決策樹輸出圖形.....	55
圖 4.12 決策樹相依性輸出圖形.....	55

圖 4.13 貝氏機率分類預測模型之相依性網路圖	57
圖 4.14 貝氏機率分類預測模型之屬性設定檔	59
圖 4.15 貝氏機率分類預測模型之屬性特性圖	59
圖 4.16 貝氏機率分類預測模型之屬性辨識圖	60
圖 4.17 類神經網路預測模型之屬性預測機率檢視	61
圖 4.18 四種資料探勘模型之姿亦圖與理想模型之比較	62
圖 4.19 四種資料探勘模型之姿亦圖與理想模型之比較	63
圖 4.20 四種預測模型之預測結果分類矩陣	64



第一章 緒論

1.1 研究動機

據內政部統計處統計，自 106 年 2 月起台灣老年人口(65 歲以上)首度超越幼年人口(0-14 歲)，其中 65 歲以上人口更高達 321 萬人，如圖 1.1。去年 4 月老年人口比例達 14%，人口老化指數也逐年攀升，如圖 1.2，正式進入「高齡社會」。對增加社會照護成本及醫療支出最直接的是人口年齡結構的改變。

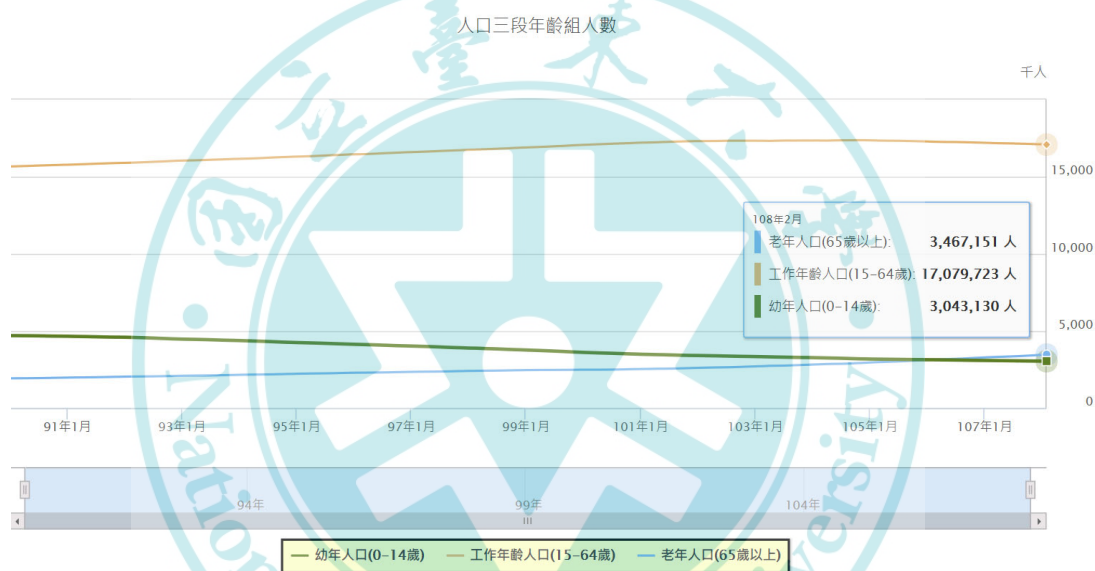


圖 1.1 人口三段年齡組人數

(資料來源：內政部統計處)

根據世界衛生組織 2016 年統計，全球每年約有 1790 萬人死於心血管疾病，占全球總死亡數的 31%。其中，死於心臟病和中風約佔 85%。非傳染性疾病所導致的 1700 萬七十歲以下死亡中，由心血管疾病所造成的約有 37%。另在我國衛生福利部死因統計中，雖然癌症為排名臺灣十大死因的第一位，但心血管的相關疾病已經連續 25 年(2016 年止)佔據十大死因第二至四名，如圖 1.3。潛在生命年數損失中，心臟疾病居平均生命年數損失前三位，如圖 1.4。



圖 1.2 老化指數

(資料來源：內政部統計處)

在高齡人口中常出現心房顫動等心臟疾病問題若控制不良，就容易引發腦中風而導致日常生活的失能狀況，腦中風是造成成人殘障的第一要因。

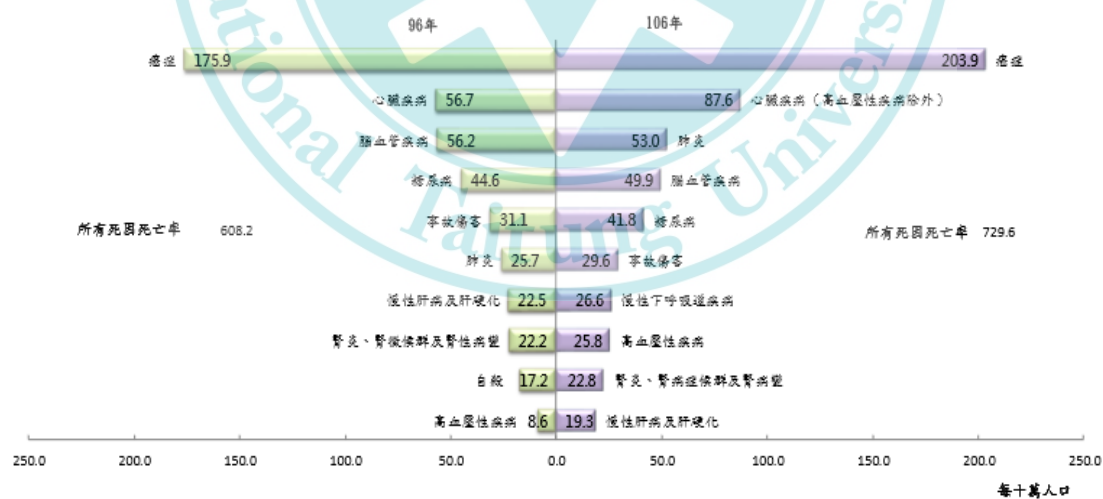


圖 1.3 十大死因死亡率-106 年 vs. 96 年

(資料來源：衛生福利部)

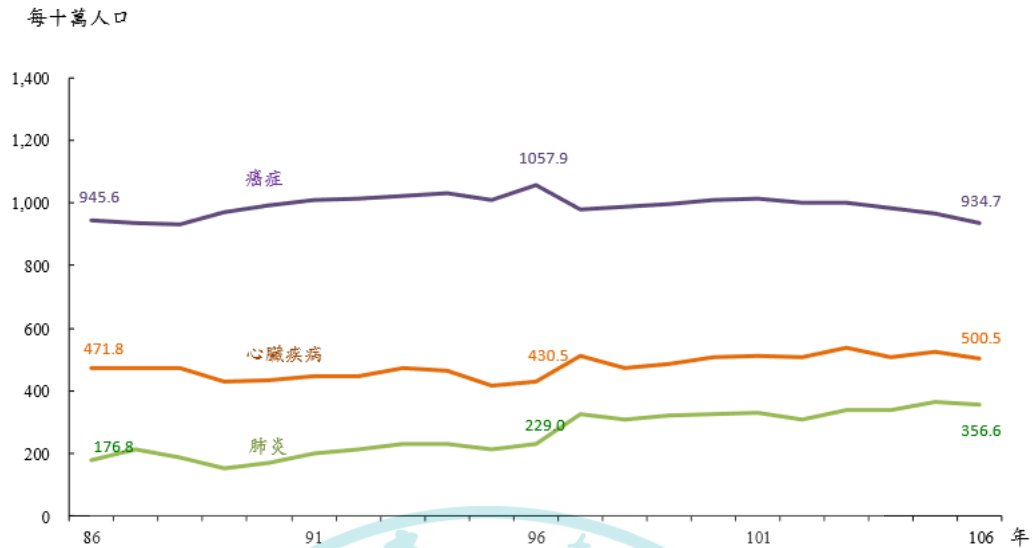


圖 1.4 歷年 65 歲以上三大主要死因死亡率分
(資料來源：衛生福利部)

腦中風及其併發症的醫療費用佔健保支出第 2 名，腦中風(stroke)影響除了病患本身面臨肢體癱瘓、語言功能及生活自理能力喪失外，家人除投入體力與心力面對漫長的失能狀況與長期照護的需求外，因長期照護需支出龐大的社會經濟負擔將加重照護壓力。

心房顫動 (Atrial fibrillation，簡稱 AF) 是目前世界上最常見的心律不整之一(Chugh, 2014)，其盛行率會隨年紀而增加(Heeringa, 2006)，約每 4 人就會有 1 人在其一生中會發生(Lloyd-Jones, 2004)此狀況。由於 AF 容易產生血栓，而這些血栓如果被發生在腦部就可能引起腦部中風，因此抗血栓治療在 AF 的病人預防中風和栓塞上是非常重要的。

在老年人的抗血栓治療上其風險及好處是 AF 中風預防的重要考慮因素(Feinberg, 1995)。高齡患者是高危險群，發病率從年齡 30 歲至 39 歲的每 1000 人中的 0.2 人穩步新增到 80 至 89 歲的每 1000 人中有 39 人。隨著年齡 50 至 59 歲的 6.7%，上升至年齡 80 至 89 歲達到 36.2%(Wolf, 1987)。心房顫動所發生的中風雖然只佔所有缺血性中風的 15%(Feinberg, 1995)，但相較於其他原因所

以起的中風，留院時間較長，失能狀況更嚴重，死亡率更高(一年內 50%)，而且更容易再發作 (Lin, 1996; Steger, 2004)。因此預防中風是臨床醫師面對心房顫動病患治療的首要考慮。

世界衛生組織 (WHO) 在 2013 年發布了《2013-2020 年預防控制非傳染性疾病全球行動計劃》，希望將癌症、心血管疾病、糖尿病和慢性呼吸道疾病等引起的過早死亡率降低 25%，其中並推薦中高風險非瓣膜性心房顫動的抗凝治療。口服抗凝血藥物 (OAC) 用於預防患者中風和全身性栓塞的功用已確定清楚。對高齡化社會的台灣而言，龐大的醫療及照護費用都逐漸增加當中。使用抗凝血藥物都存在著出血的危險性。若是中風風險會大於服用抗凝血藥物所增加的出血風險，臨床醫師們須慎重考慮讓病患接受口服抗凝血藥物(OAC)來預防中風。但因臨床症狀會依每個病人的病況都不盡相同，預防中風方式及藥物選擇也會因此而不同。醫師必須全面考量病人的狀況，經充分解說溝通讓病患了解用藥的目的及必要的注意事項，來使用選擇不同治療方式來預防心房顫動相關腦中風的發生(Fontana, 2016)。在病患開始服用抗凝血劑之前，出血風險評估相當重要，目前針對心房纖維顫動病人的使用抗凝血劑引起之出血風險評估可利用 HAS-BLED 分數表，分數越高表示發生出血機會越高(Pisters, 2010)。2010 年歐洲心臟學會 (ESC guidelines)改採用 HAS-BLED score 來預測患者出血的風險(Camm, 2010)，HAS-BLED score 評估項目較為精簡，不但可信度已由許多世代研究所驗證(Olesen, 2011; Pisters, 2010)，在臨床上目前被廣泛運用。但要留意的是，若 HAS-BLED score 的計分較高並非表示須停用或無法使用口服抗凝血藥，而是提供警訊讓臨床醫師能更謹慎頻繁的追蹤出血風險並積極預防出血事件發生。目前臨床上較常使用 HAS-BLED score 來評估病患出血的風險機率，無法提供病患個人化之預測，故本研究藉由使用新抗凝血劑 Rivaroxaban 用藥病人之用藥前後資料分析、資料探勘之利用來發展可行性

之預測模型，提供臨床醫師在開立口服抗凝血劑時用藥之預警支援系統，提高預防出血事件之發生。

1.2 研究目的

由於電腦運算能力增加、資料探勘方法的發展及資料收及能力的進步，已讓資料探勘已成為從大量資料中發出有用知識的重要工具。本研究擬以患有 AF 及使用 rivaroxaban 病人之 HAS-BLED score 相關易致病因子為變項來預測 AF 病患在用藥期間可能發生出血事件的機率，預先提供警訊，提供臨床醫師進行合適的病症處置減少出血事件所造成的健康損害。

1.3 論文架構

本研究共分有五章，各章節說明分別如下：

第一章 緒論：說明本研究主題之背景環境，進而產生探之就動機、目的。

第二章 文獻探討：本研究擬以變數建立預測，發展最佳化模型，故收集 AF、資料採礦的相關文獻做深入研究。

第三章 研究方法：

1. 解說本研究之研究流程。
2. 說明資料來源及相關變數。
3. 資料整理、轉化的方式。
4. 本研究軟硬體工具的說明使用。
5. 評估模型的方式。

第四章 研究結果：綜整資料採礦模型的建置方式與結果比較，尋求最佳化模型，發展預測評估及成果展示。

第五章 結論與建議：統整研究成果未來研究的方向及目前研究不足之處。

第二章 文獻探討

2.1 心房顫動疾病簡介

2.1.1 流行病學

在臨床上最常見的一種心律不整疾病為心房顫動(Atrial fibrillation)。其盛行率約占全球人口的 1%，心房顫動與年齡最具關連性，小於 60 歲盛行率僅有 0.5%，依序遞增 60-69 歲的 2-3%；70-79 歲 5-6%；大於 80 歲就爬升到 8-10%(張朝欽、雷孟桓，2012)。心房顫動發生的比率隨著年紀的增加也隨之增加。引起心房顫動的機制是心房組織內有很多及很快的不正常放電，產生心房無法正常的收縮導致成心跳數不規律的忽快忽慢。因心房顫動病患發生中風的機會比常人多五倍，心臟衰竭多三倍，死亡率多二倍(Feinberg, 1995)，所以心房顫動的病人，可使用抗凝血劑藥物來預防腦中風的發生。

2.1.2 心房顫動分類

依 ACC/AHA/ESC 的分類標準，當病患發生兩此以上的心房顫動(AF)事件時，可視為心房顫動(AF)復發，這時可依心房顫動(AF)發作的頻率分成以下三類：

1. 陣發性心房顫動(paroxysmal atrial fibrillation)- 心房顫動可自發性恢復，持續期不超過七天。
2. 持續性心房顫動(Persistent atrial fibrillation)- 持續期超過七天，還可以使用藥物或電擊整流回復正常心律者。
3. 永久性心房顫動 (Permanent atrial fibrillation)- 持續期超過一年以上，無法使用藥物或電擊整流恢復正常心律者。

當心房顫動發生時，病人便會覺得會產生出現有心悸、胸悶、呼吸困難、氣喘、頭暈等症狀。當心跳數突然變得很緩慢或停頓許久時，病人便可能會發生昏厥的現象；反之，當心跳數變得很快時，心臟血液輸出量除減少外，

若長時間的發作也會造成心臟功能惡化，進而造成心臟衰竭。因此當有心房顫動發生的病人，需要服用抗心律不整藥物，用以預防心房顫動發生，或是服用降低心跳藥物，用以防止心房顫動發生時發生心搏過速的情形；但當有心搏過慢且有昏厥現象發生時，便需要裝置心律調整器來監測心跳，當心跳過慢或休止時可以適時發出電流刺激心臟跳動，維持心跳的穩定。第二，心房顫動發生時，因心房組織快速且不正常放電，心房便無法有效的正常收縮，取而代之的是快速地顫動，導致血液無法正常的有效流動，當血液因此停滯時間就較長進而容易於左心房形成血栓，當血栓隨著血流流出心臟分佈到人體的各個器官，便容易引起各個器官的栓塞，如栓塞於腦部便就是腦中風。

包括高血壓、心臟衰竭、瓣膜性心臟病（30%AF 患者合併瓣膜性心臟病，尤其是二尖瓣及主動脈瓣狹窄或逆流）、缺血性心臟病、糖尿病、甲狀腺功能疾患、先天性心臟病（心房間隔缺損）、肥胖、慢性阻塞性肺病、睡眠呼吸中止症候群、慢性腎病等等是導致AF的病因。罹患心房顫動的一大原因是高齡，心房顫動的病人年齡大於65歲占了70%以上，男性多於女性。若有腦中風病史、心臟衰竭、僧帽瓣疾病、高血壓、左心房擴大、左心室肥厚、肥胖、甲狀腺疾病或者罹患睡眠呼吸中止症的病患也比較容易發生心房顫動。電氣傳導迴路的形成與自律神經失調與因局部心房內膜維持著快速且不正常的放電所引起的心房顫動也佔有重要的關係。一些研究指出，事實上在各個單一區域的電氣活動，空間及時間上的關係是相當穩定的，雖然看似不規則。心房顫動的來源可能來自最高放電頻率的心房區域，通常是位於左心房後壁或肺靜脈一帶，而且從高密度心房的定位結果得知，單一高速的迴路的區域會合併週邊心房組織的顫動傳導，這是目前認為心房顫動最重要的機制之一。

2.1.3 臨床表現

當病患有任一型態的心房顫動時，不一定會出現任何不適的症狀，但也可能會出現心悸、頭暈、疲倦、胸悶或呼吸困難等臨床症狀，常與心跳速率過快有關。若能藉藥物控制心跳節律則症狀即可獲得改善。但並非所有心房顫動事件在發作或復發時都有臨床症狀，病患也可能因栓塞或心衰竭惡化的臨床表現進而發現有心房顫動。

2.1.4 治療原則

心房顫動的節律調整(rhythm control)治療主要可分為積極性的維持正常竇性脈(心律或節律控制)及消極性的心室心速控制(心速或速率控制)。

(1) 心律控制 (rhythm control) -是使用藥物或電擊患者恢復到正常的竇性節律(sinus rhythm)，以抑制心房顫動的持續發生並使其保持正常節律(normal rhythm)。

(2) 速率控制 (rate control) 是使用藥物控制心室速率，使心室有足夠時間回充血液避免心跳過快造成缺血，以維持正常的心房及心室配合(synchronization)的功能，達到最佳的心輸出量(cardiac output)。

心房顫動的治療除了上述以心律控制 (rhythm control) 及速率控制 (rate control) 之外，還可選擇以抗血栓藥劑預防栓塞。因心臟不正常的節律引起心臟流動量(flow)及速度(velocity)降低，特別容易在心臟左心耳(left atrial appendage)形成血栓。當血栓掉落再經由血液循環容易造成動脈或腦血管栓塞事件。如果使用藥物或電擊使心房顫動回復正常節律時會產生心房頓抑(atrial stunning)，因無法立即回復心房收縮功能，造成血液容易滯留心臟而提高栓塞的危險性。

2.2 心房顫動與血栓性栓塞

2.2.1 抗血栓劑在心房顫動患者的應用

當患者開始使用抗凝血劑之前，評估出血風險顯得相當審慎。出血風險的高低與患者使用的藥物種類及患者本身病況有著密切的關係。在心房顫動與缺血性中風的關係尚未明瞭時期，只有併有風濕性瓣膜的病患會因預防中風而長期使用抗凝血劑。在 1970-1980 年間研究發現：沒有風溼性瓣膜疾病的心房顫動患者，中風的危險性更於一般人口。自 1989 年開始，許多大型臨床試驗開始評估不同抗血栓劑在心房顫動族群中預防中風的效果。目前研究多指向包括 warfarin、rivaroxaban、apixaban 或 dabigatran 治療的病患，相較其有效性及安全性。

2.2.2 心房顫動患者抗血栓劑國際準則建議

2016 歐洲心臟學會 (European Society of Cardiology, ESC) 在義大利羅馬歐洲心臟學會 (ESC) 最新公布新治療準則是關於心房顫動患者使用抗凝血藥物，其中明確指出雖然不管傳統抗凝血藥物或新型口服抗凝血藥物皆能降低心房顫動患者的死亡率，但建議患者應優先使用新型口服抗凝血藥物降低出血風險。同時提出了最新的試驗結果，一項大規模追蹤非瓣膜性心房顫動患者的回溯性分析發現，使用新型口服抗凝血藥物，包括 rivaroxaban、apixaban 或 dabigatran 治療的病患，相較於傳統上使用抗凝血藥物藥 warfarin，同樣有效但更安全。

其中針對美國心房顫動患者的追蹤顯示，在使用能預防動脈和靜脈栓塞的其中一項口服抗凝血新藥後，缺血性中風和顱內出血風險分別比使用 warfarin 降低達 29%、47%。會議中也針對人口占全球一半以上的亞洲提出警語，預估由於高齡化和人口快速增加，心血管疾病已成為亞太地區的主要死

因之一，光 2015 年就有 820 萬人因此喪命，比 2005 年的 680 萬人增加超過 20%，如何預防中風、甚至因此失能的重大醫療支出，急迫性和重要性越來越高。

2.2.3 RIVA 出血危險性

表 2.1 出血風險預測

(Clinical characteristics composing the HAS-BLED bleeding risk score)

出血危險因子	分數
Hypertension	1
Abnormal renal and liver function (1 point each)	1 or 2
Stroke	1
Bleeding history or predisposition	1
Labile INRs	1
Elderly (> 65 years)	1
Drugs or alcohol (1 point each)	1 or 2

(資料來源：Pisters et al., 2010)

2010 年歐洲 ESC guidelines 改採用 HAS-BLED score 來預測患者出血的風險(Association, 2010)，Pisters et al.(Pisters, 2010)從 The Euro Heart Survey 當中選出 3978 名有 AF 患者，所發展出的一個簡單量表 HAS-BLED score，如表 2.1，包括： 1. 高血壓 1 分。未控制的高血壓，SBP>160mmHg。 2. 異常的肝功能或腎功能，肝腎功能各 1 分。腎功能異常包括了慢性洗腎、腎臟移植或肌酐酸 >200mmol/L(約 2.26mg/dL)；肝功能異常包括了慢性肝臟疾病或黃疸高於正常值 2 倍以上或肝指數高於正常值 3 倍以上。 3. 過去有中風或出血的病史，中風及出血各 1 分，尤其注意小

間隙梗塞。 4. 不穩定的 INR, TTR<60%, 1 分。 5. 年齡>65 歲, 1 分。 6. 同時合併使用可能容易導致出血的藥物或過量飲酒, 飲酒或使用藥物各 1 分。藥物如抗血小板或非類固醇消炎等。若分數 ≥ 3 分則屬於高危險族群, 不管使用抗血小板或是抗凝血劑都要特別小心。

2.2.4 抗血栓劑 RIVA 臨床應用現況

Rivaroxaban 是一種直接作用抑制活化態第十凝血因子(Factor Xa inhibitor)的口服抗凝血藥物。其優點主要為藥物交互作用少、飲食限制少、不用監測 INR、顱內出血併發症少、預防中風和產生血栓的效果好。在中風預防的研究中關於 rivaroxaban 的使用, 有一雙盲、隨機的大型跨國研究 ROCKET AF study(Patel, 2011), 收案 14264 名 AF 患者, 結果顯示 rivaroxaban 在非瓣膜性心房顫動(Nonvalvular atrial fibrillation, NVAF)患者的中風預防效果不比 warfarin 差, 重大出血比率兩者分別為每年 3.4%(warfarin) vs 3.6%(rivaroxaban); 其中 rivaroxaban 有較多的消化道出血; 而在總出血率上兩者無太大差別, 但在腦出血及致命性出血等嚴重併發症方面, rivaroxaban 明顯低於 warfarin (Patel, 2011)。藥物不良反應發生比率分別為 81.4% vs 83.4%; 大部分是輕微流鼻血及血尿。在 ROCKET AF 試驗的次族群分析中, rivaroxaban 在中風的初級與次級預防有相同的效果。另一個則是日本人使用較低劑量 rivaroxaban(15mg)的研究, 同為 phase III 隨機雙盲試驗(J-ROCKET AF study), 在中風發生率與出血率沒有統計上差異, 但 rivaroxaban 有較少的腦出血機率的趨勢(The Bayer press server, 2011)。而在東亞族群(含中國、香港、南韓及台灣)共 942 位非瓣膜性心房顫動(Nonvalvular atrial fibrillation, NVAF)患者的次族群分析中發現, rivaroxaban 比起 warfarin 可減少 22%相對風險, 並大幅減少腦出血及致命性出血的比例 (Lip, 2015)。由此可知 rivaroxaban 可以成為非瓣膜性心房顫動

(Nonvalvular atrial fibrillation, NVAF) 患者預防中風的另外一個選擇。口服抗凝血藥物 rivaroxaban 可預防深層靜脈血栓發生，出血風險及藥費則和其它抗血栓藥物相當(毛雲川, 2014)。臺灣機械性心臟瓣膜置換者服用口服抗凝血劑的比例接近百分之百(王郁文, 2011)。經成本效益研究的公共衛生政策可讓血栓溶解劑治療缺血性中風效益最大化(周秉箴, 2013)。除此之外，檢測凝血酶原時間 (prothrombin time, PT) 可以約略了解其抗凝血活性，且國際標準化比值 (international normalized ratio, INR) 值如果升高可能意味著嚴重出血的風險增加，但 INR 值並不與其活性成正比。口服抗凝血藥物 rivaroxaban 可預防深層靜脈血栓的發生，出血風險及藥費則和其它抗血栓藥物相當。

2.3 資料探勘

2.3.1 資料探勘的定義

資料探勘(Data mining)是利用大量的資料，應用不同的資料處理方式，從資料中萃取出隱含、未知的知識及可用的模式，用以預測未來發生的事情。在跟隨著科技進步、電腦運算能力增強、電腦儲存容量的提升及網路的發達，對資料的獲取應用方式產生很大的變化。資料採礦是由資料庫系統、知識庫系統、機器學習、人工智慧、統計學及不確定推論等知識中整合出可供應用的定理、演算法或方法。Fayyad 等人認為(Fayyad, U&Stolorz, P, 1997)資料探勘是依照使用者之需求，從資料庫中選擇合適資料，再處理、轉換、探勘評估的過程。Berry&Linoff(1997)認為資料探勘是利用大量資料，以自動或半自動方式找出有意義的規則或模型。Cabena 等人(Cabena et al., 1998)定義資料探勘是從資料庫中未取出未知也有效的資訊過程，且這些資訊將提供決策管理者做決策參考。Kleissner (1998)認為資料探勘是種不斷循環的決策分析過程，從資料中找出隱藏的知識以提供企業人員參考。Han and Kamber (2011)認為，

資料探勘從資料庫之萃取有興趣資料或模型的過程。綜合以上學者對資料探勘的定義，建構資料探勘模式時，需先清楚定義問題、考慮解決方式、資料的選擇、之後進入建立模型與資料準備的相互循環階段，並將結果分析、評估及轉化為可用的知識與應用的過程。

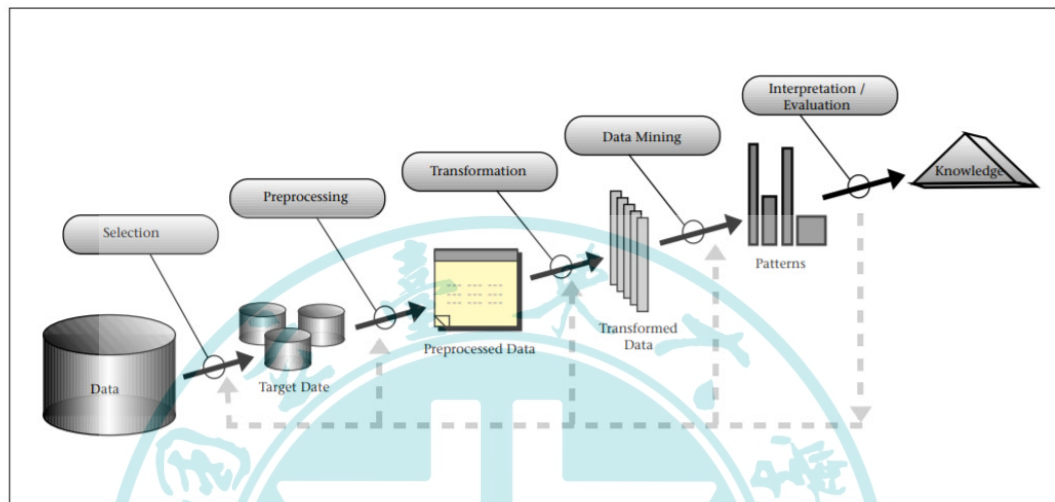


圖 2.1 KDD 形成過程

(資料來源：Faddy et al., 1996)

在資料庫知識發掘(Knowledge Discovery In Database, KDD)的形成過程中資料採礦為必要的核心步驟之一(Faddy et al., 1996)，KDD 的形成步驟如圖 2.1 所示，說明如下：

1. 資料選擇(Selection):選擇適合資料或定義目標變數。
2. 資料整理(Preprocessing):整理基本補值及空資料或遺漏值處理。
3. 資料轉化(Transformation):依分析目標所需將資料做適當的轉換化約，刪除重複或不需要的資料。
4. 資料採礦(Data Mining): 依分析目標選擇採礦演算方式進行挖掘發現有意義或有用之樣型。
5. 分析評估(Interpretation/Evaluation):將所挖掘之資料樣型加以評估，並以理解、確認、觀察及再利用之方式呈現。

6.知識形成(Knowledge):將有價值之資訊應用並驗證。

2.3.2 資料採礦的方法

依資料採礦的方式可分為已知問題結果所建立模型以及分析的監督學習(Supervised Learning)與未知問題結果而利用現有資料建立模型的非監督學習(Unsupervised Learning)。監督學習(Supervised Learning)為描述性之目的，希望使用較清楚的方式來描述隱藏在大量且複雜資料，分析資料間的關聯找出可能有用的相關趨勢、規則樣型。非監督學習(Unsupervised Learning)為預測式之目的，希望使用歷史資料間的關聯規律來建立模型，作為預測辨別未來為之的。

根據目的及需求用途有各種種類不同類型的演算法，如探索資料之間隱含關係的關聯規則(association rules)，找出目標變數與各個變數之間層級關係的決策樹分析(decision tree)，模仿生物神經網路資訊系統的類神經網路(artificial neural networks)，分析資料屬性與反應變數間機率模型的貝氏分類(Bayesian classifier)等等。因為資料探勘能夠在有目標性的方向中，從大量龐雜看似毫無相關的資料中扒梳出從所未知的規則或關聯性，進而應用在各行各業上。有一個迫切需要新一代的計算理論和工具，以協助人類從迅速增長的數位資料量中提取有用的資訊(知識)。

2.3.2.1 關聯規則

關聯規則主要是從龐大看似無相關聯的資料中、探索資料欄位之間的可能存在而未知的關聯性。關聯規則能以容易理解的解釋其產生的規則並完整呈現變數間的影響。Agrawal et al.(1993a；1993b)最早從龐大的交易資料中發掘商品間隱含的關聯規則，以了解消費者的購買行為與產品銷售的關係。根據韓家煒等(Han and Kamber,2000)將關聯規則定義為假設 $I = \{ I_1, I_2, \dots, I_m \}$ (Pisters et al.)是項目的集合。給一個交易資料庫 $D = \{ t_1, t_2, \dots, t_n \}$ ，

其中每個交易 (Transaction) t 是 I 的非空子集，即 $t \subseteq I$ ，每一個交易都與一個唯一的識別元 TID (Transaction ID) 對應。關聯規則形如 $X \Rightarrow Y$ 的關聯式，其中 $X, Y \subseteq I$ 且 $X \cap Y = \emptyset$ ， X 和 Y 分別稱為關聯規則的先導 (antecedent 或 left-hand-side, LHS) 和後繼 (consequent 或 right-hand-side, RHS)。關聯規則 $X \Rightarrow Y$ 在 D 中的支持度 (support) 是 D 中事務包含 $X \cup Y$ 的百分比，即機率 $P(X \cup Y)$ ；信賴度 (confidence) 是包含 X 的事務中同時包含 Y 的百分比，即條件機率 $P(Y | X)$ 。如果同時滿足最小支援度閾值和最小信賴度閾值，則認為關聯規則是有相關聯的。

關聯規則使用支持度、信賴度和增益來衡量顯著性、正確性及價值。透過給定最小支持度與最小信賴度作為支持度及信賴度門檻值，評估資訊規則的價值及增益。若支持度及信賴度大於或等於門檻值，表示該規則有效的。關聯規則可分為三種類型 (Han and Kamber, 2006)：

1. 以規則中屬性值的類型為基礎：若以關聯規則中的資料項是否出現，為布林關聯規則 (Boolean association rule)；若關聯規則中的資料項是數量型，為量化的關聯規則 (quantitative association rule)。

2. 以規則中資料維度為基礎：若關聯規則的項目只涉及一個維度，則它是單一維度關聯規則 (single-dimensional association rules)；若關聯規則涉及兩個或兩個以上維度，則它是複合維度關聯規則 (multi-dimensional association rules)。

3. 以規則中所涵蓋的抽象層級為基礎：若不涉及不同層次的資料項，得到的是單一層級關聯規則 (single-level association rules)；若不同階層層次中挖掘出的關聯規則稱為多階層關聯規則 (multilevel association rules)。

2.3.2.2 類神經網路

類神經網路的概念是來自於生物學上的神經元構造，模仿生物神經網路的資訊處理系統，以大量簡單的相連人工神經元模仿生物神經網路，使電腦

能夠模擬人類神經系統結構與資料處理的方式。自 1940 年起便有科學家試圖模仿簡單的神經元模式，建立原始的類神經網路(Artificial Neural Network,ANN)(謝邦昌、鄭宇庭、蘇志雄,2011)。目前類神經網路是人工智慧的核心。類神經網路是利用大量簡單的相連人工神經元來模仿生物神經網路學習能力的資訊處理系統。根據 Freeman and Skapura(1992)(Freeman, 1991)的定義，類神經網路是模仿生物神經網路的資訊處理系統，使用了大量簡單的相連人工神經元來模仿生物神經網路的能力。而在一個網路模型當中，一個人工神經元將從外界環境或其他人工神經元取得資訊，依據資訊的相對重要程度給予不同的權重，並予以加總後再經由人工神經元中的數學函數轉換，輸出其結果到外界環境或其他人工神經元當中。其運作概念可整理如圖 2.2。

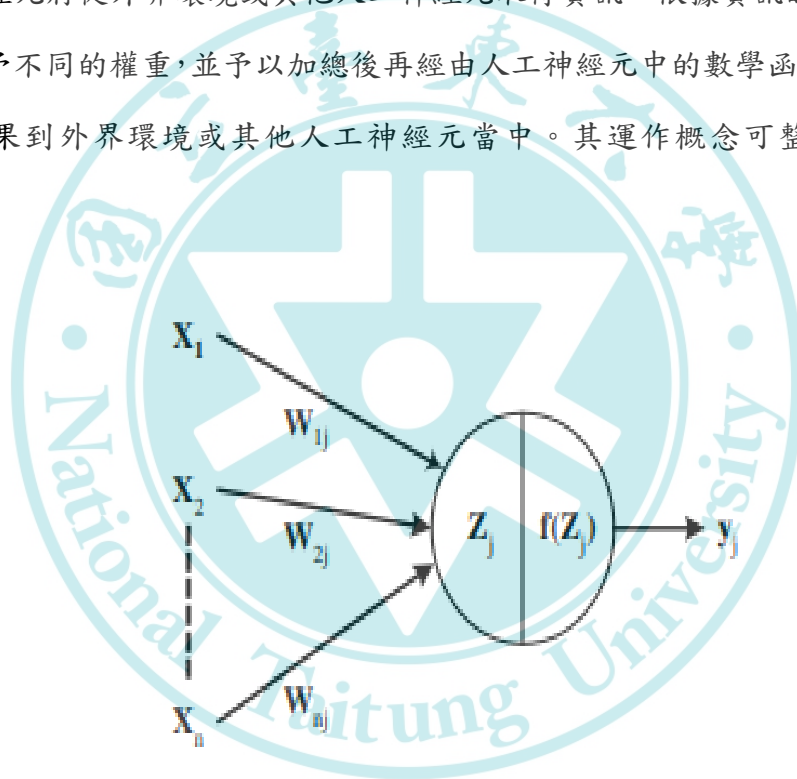


圖 2.2 神經元構造

(資料來源：謝邦昌，2017)

其中， X_1 、 X_2 、...、 X_n 代表輸入值， W_{ij} 代表連接鍵的權數， $Z_j = \sum W_{ij}$ X_i 代表加權和， $f(Z_j)$ 則是代表一轉換函數，最後 y_j 則是神經元的輸出值。以定義來說，「類神經網路是一種計算系統，包括軟體及硬體，它使用大量簡單的人工神經原來模仿生物神經元網路的能力。人工神經元是生物神經元的

簡單模擬，從外界環境或其他人工神經元的資訊，並與以非常簡單的運算，並輸出其結果到外界環境或其他人工神經元」(葉怡成，1999)。生物的神經元包含了神經細胞核、神經軸、輸出神經數及神經節，如圖 2.3 所示。

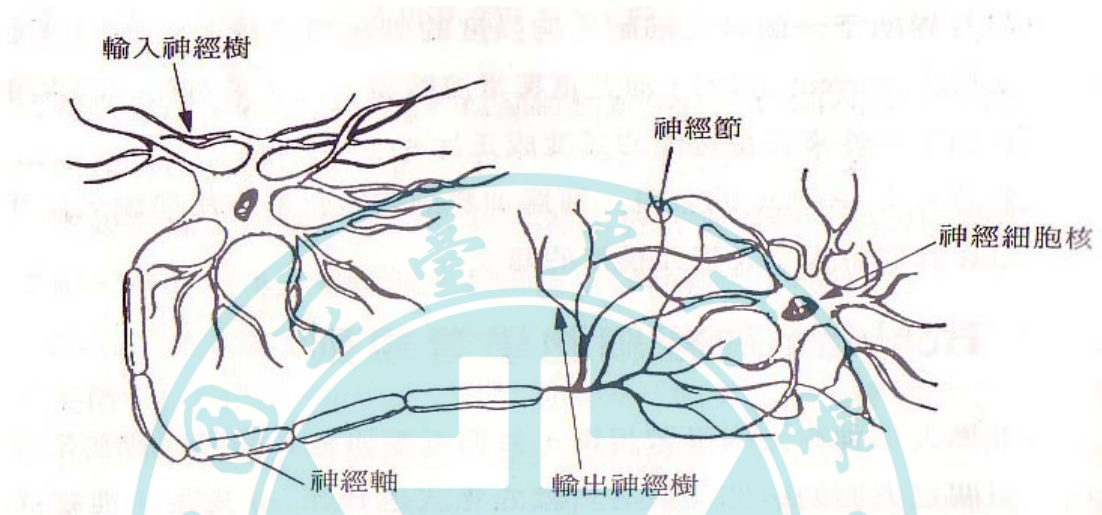


圖 2.3 生物神經元示意圖

(資料來源：王進德、蕭大全，1994)

Microsoft 類神經網路使用「多層認知」(Multilayer Perceptron) 網路，亦稱為「倒傳播差異規則」(Back-Propagated Delta Rule) 網路，包含最多 3 層神經 (Neuron) 或「認知器」(Perceptron)。類神經網路架構包含輸入層、輸出層、及多個隱藏層，如圖 2.4 所示；每一個輸入及輸出神經元都與資料集合中的變數相聯結並與鄰近的所有神經元連接。每一個隱藏神經元的反應函數決定訊號如何從輸入神經元傳遞到輸出神經元。利用大量訓練資料可以調整整個神經元間鏈結的權重，並訓練網路減少輸出的錯誤，降低預測值與目標值之間的差異，經此訓練網路可做為預測新資料的模型。多層的網路克服了單層神經網路被僅具一層限制無法從輸入中學習，因為它可以創建內部在

每一層學習不同的特徵(Rumelhart, 1986)。第一層可能負責輸入學習線條的走向。第二層可能結合第一層所學並學習識別。每提升一層就學習越多的抽象特徵。每一層都是從它下方的層中找到模式，創建了獨立於為多層網路提供的外界輸入的內部表達形式。反向傳播算法的發展目標和動機是找到一種訓練的多層神經網路的方法，可以學習合適的內部表達來讓它學習任意的輸入到輸出的映射(Rumelhart, 1986)。

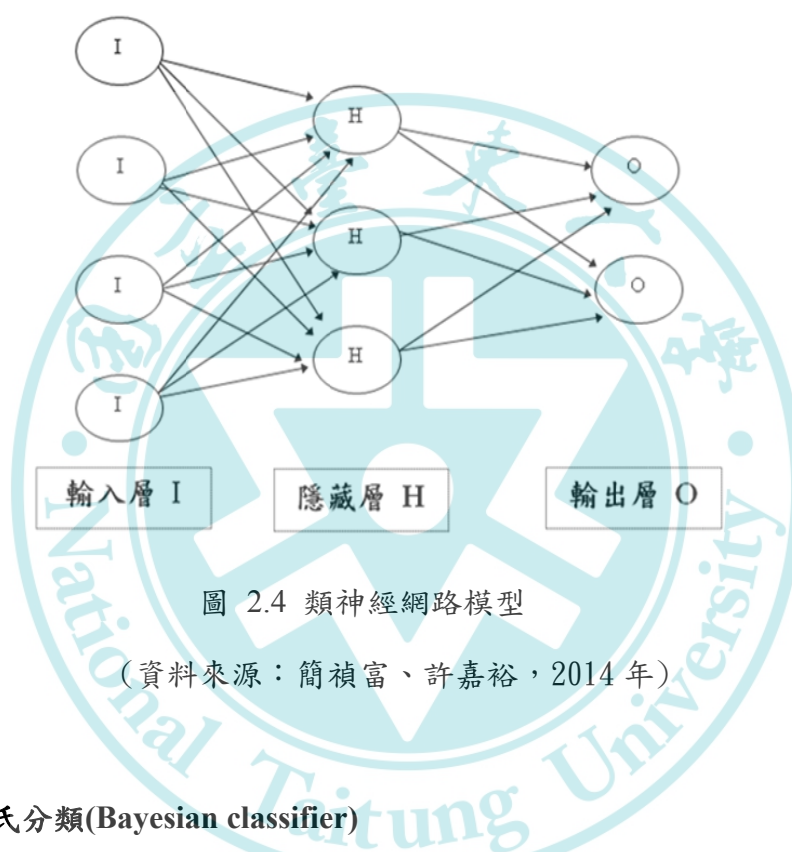


圖 2.4 類神經網路模型

(資料來源：簡禎富、許嘉裕，2014 年)

2.3.2.3 貝氏分類(Bayesian classifier)

依據貝氏定理(Bayesian theorem)來更新資料以推斷資料歸屬類別，作為分類和推論的依據，常用的有單純貝氏分類法(naïve Bayesian classifier)及貝氏網路分類法(Bayesian network classifier, 簡稱貝氏網路)。其運作原理為監督式的學習方式，分類前須事先知道分類資料型態，透過訓練樣本的學習訓練，有效的處理分類的資料(謝邦昌、鄭宇庭、蘇志雄, 2011)。單純貝氏分類器是：當樣本中有 n 項特徵(feature)： $F_1, F_2, \dots, F_n$ ，並且可以根據這些特徵被歸類為 m 個類別 $C_1, C_2, \dots, C_m$ 中的一類時，若新給定一個資料的 n 項特徵，

要如何判斷新資料屬於哪個類別。所以很直觀的可以考慮，在給定的特徵下，哪個類別的可能機率比較高，見方程式 2.1：

$$\max P(C_i|F_1, F_2, \dots, F_n) \dots\dots\dots (2.1)$$

當我們將所有類別的機率都求出後，可以選擇機率最高的類別當作我們的推測，根據貝氏定理，見方程式 2.2：

$$P(C_i|F_1, F_2, \dots, F_n) = \frac{P(C_i) \times P(F_1, F_2, \dots, F_n|C_i)}{P(F_1, F_2, \dots, F_n)} \dots\dots\dots (2.2)$$

單純貝氏分類器更進一步假設各個特徵彼此獨立(雖然在現實中不太可能成立，但是它可以大大簡化計算，而且有研究表明對分類結果的準確性影響不大，見方程式 2.3：

$$\Rightarrow P(C_i|F_1, F_2, \dots, F_n) = \frac{P(C_i) \prod_{k=1}^n P(F_k|C_i)}{P(F_1, F_2, \dots, F_n)} \dots\dots\dots (2.3)$$

因為分母部分在同一組特徵下會一樣，我們只需考慮分子部分即可，見方程式 2.4。

$$(F_1, F_2, \dots, F_n) \in C_i \Leftrightarrow i = \arg \max_{1 \leq i \leq m} P(C_i) \prod_{k=1}^n P(F_k|C_i) \dots\dots\dots (2.4)$$

Microsoft 貝氏機率分類演算法是根據貝氏的定理的分類演算法，並可用於探勘和預測模型。貝氏 (Naïve Bayes) 名稱中的 naïve 一字源自此演算法使用 Bayesian 技術但卻沒有考量可能存在的相依性。

貝氏機率分類模型的需求如下：

(一)單一索引鍵資料行：每個模型都必須包含一個能唯一識別的值或文字資料行。不允許複合的索引鍵。

- (二)輸入資料行：所有的資料行都必須是離散的資料行，或已離散化的值。
- (三)變數必須是獨立的。確保輸入屬性彼此無關也很重要。
- (四)至少有一個可預期的資料行：可預期的屬性必須包含離散或離散化的值。

2.3.2.4 決策樹(Decision Tree)

決策樹是資料探勘演算法中分類演算的一種，依用途可為分類變數及預測變數。決策樹是將輸入變數依據某種規則或方法對資料進行分類，並以束狀方式呈現表示類別之間由輸入變數所造成的區別，故可藉由決策樹的分析規則對資料進行層及架構的分類，進而挖掘出對結果有顯著影響的因素。其優點在推理過程易理解，推理過程完全依據變數的特性。決策樹通常用於監督式特徵萃取與描述，以解決資料探勘領域中分類型態的問題。決策樹依演算法的差異可採用不同分支規則和判斷方式，以選擇分支變數、分支方式以及層及架構，進行一系列的資料分類。決策樹演算法都是在分支時使類別間的不純度(impurity)最小化，「不純度」的衡量方式為區分各種決策樹演算法的重要因素之一。CART 代表分類系統或預測模型，有一個反應變數和一個以上的解釋變數，是由 Brieman 和其同事於 1984 年提出(Rumelhart, 1986)。CART 使用時需定義分支節點和葉部節點，每個分枝節點均為一個二元試驗(Bernoulli Trail,即可能為「成功」或「失敗」、「年齡大於 65 歲」或「年齡小於 65 歲」等)，決定樣本如何被送到下一層節點；當分支節點為樣本的最後分類則稱為葉部節點。因為分類最後的結果有如一棵樹，有時也稱 CART 稱為決策樹(Decision Trees)。CART 的分枝大多以 Gini 係數(Gini Index)為標準，如何將支點從根部開始延展，或以何變數的何條件作分支，都會影響此分類與迴歸樹模型分類的好壞，以 Gini index 來做分支節點的選擇，找出最小的 $gini_{split}(S)$ ，從 split 的屬性做分支。樹(模型)建構好時，會進行修

剪(pruning)的工作，修剪的依據為其整體誤差率(entire error rate)來進行修剪，使樹不僅預測能力更好且分支最少。決策樹建立的方式為：

(一)以資料母體為根結點。

(二)使用單因子變異數分析等方式找出變異量作最大的變項作為分割準則，每個葉節點為一連串法則的分類。

(三)假如判斷結果的正確率未滿足條件，則再依最大變異量長出葉節點。

決策樹模型的演算法主要包含下列四種：(1) CART；(2) C5.0；(3) CHAID；及(4) Quest。

選擇決策樹模型是依據目標變數及預測變數型態不同而有所不同，如圖 2.5 所示。



圖 2.5 決策樹模型

(資料來源：本研究整理)

第三章 研究方法

3.1 研究方法與步驟

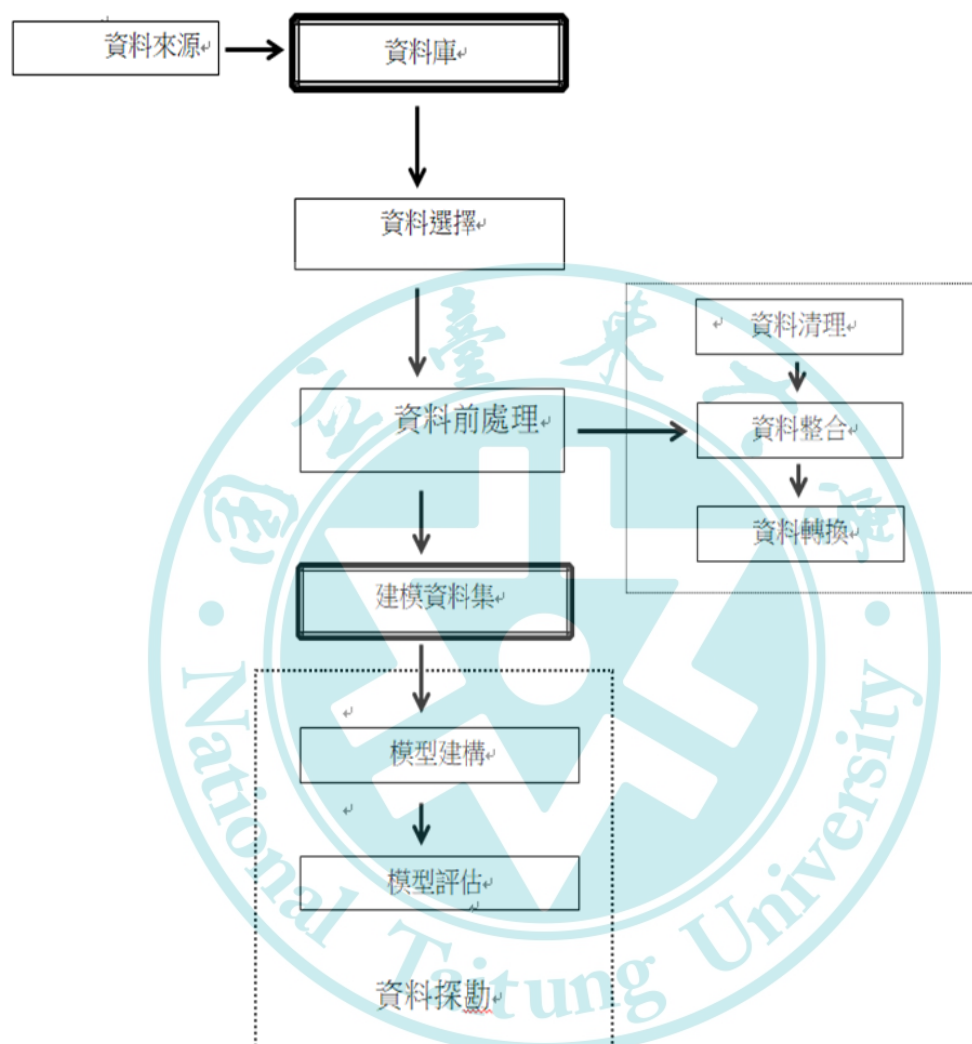


圖 3.1 研究架構

(資料來源： Hui, 2000)

本研究以 Microsoft SQL Server2017 為資料採礦軟體工具應用作為本研究資料分析處理。在資料探勘中，依分析資料及尋找樣式相關性來選擇合適演算法來建立資料探勘模型，故於本研究對資料的歸納分類依資料探勘的分類技術而定。依 Microsoft SQL Server2017 軟體工具所運用的分類演算法而採用對於

預測具代表性的決策樹、關聯分析、貝氏分類、類神經網路之不同演算法及分類技術來建立不同預測模型來為研究案例資料來建立資料探勘模型。進行模型之參數調整至最佳化、取得預測模型並加以比較選取可行模型之最佳性，利用預測系統進行實測與驗證，最後做結論及研究建議。研究架構採用(Hui, 2000)所提出之資料探勘研究步驟，研究架構如圖 3.1 所示。

3.2 資料來源及研究變項

3.2.1 資料來源

本資料來源為某地區醫院收集使用 Rivaroxian 之患有 AF 病人相關資料，如表 3.1。記錄期間為 2016/1/1 至 2018/4/30 止。資料紀錄內容包括病歷號、性別、第一次用藥時年齡、第一次用藥日期、一開始併用最久的抗心律不整藥物(AAD)、AF 種類、第一次發現患有 AF 方式、心律調節器(PPM)種類、服用抗凝血劑(NOAC)原因、因心臟衰竭住院(CHF Admission)病史、高血壓病史、糖尿病史、肝腎病史、先前大出血或易出血病史(Prior major bleeding)、抗凝血指數(Labile INR)、出血性中風(Ischemic stroke)病史、出血性中風(ICH) 病史、硬腦膜下出血(SDH)病史、蜘蛛膜下腔出血(SAH)病史、Any stroke(包括出血和梗塞) 病史、短暫性腦缺血(TIA) 病史、血栓栓塞(Thromboembolism) 病史、飲酒行為、其他併用藥物種類及使用藥物後相關事件發生紀錄:CV death 原因、是否發生 non-fatal MI、non fatal-MI 發生日期、是否發生 New Ischemic Stroke、是否發生 All cause mortality、再一次 HF 住院、是否發生 New Embolism、是否發生 ICH、發生 ICH 的類別、 Other / SDH 或 SAH (0/1)、是否發生 Hb drop > 2 or 輸血 4U、是否發生 Critical sites bleeding(眼底，後腹腔，大關節，流血致死) 、是否發生 Any other bleeding、發生 Any other bleeding 原因、是否發生 GI bleeding、是否發生 Hold AAD as bradycardia 、是否發生 ARF、是否發

生 Hypothyroidism、是否發生 Hyperthyroidism、是否發生 Lung fibrosis、HAS-BLED 計算出血風險數值。

表 3.1 研究資料分布情形

Baseline Characteristics of Rivaroxaban Users				
Characteristic			n =3158	%
Age, y				
	<65		608	19.3
	≥65		2,550	80.7
Sex				
	Male		1570	49.7
	Female		1588	50.3
Uncontrol HTN			232	7.3
DM			1071	33.9
Hepatic disease			114	3.6
Renal disease			83	2.6
Prior major bleeding			278	8.8
Labile INR			177	5.6
Ischemic stroke			775	24.5
ICH/SDH/SAH			76	2.4
Any stroke			828	26.2
TIA Thromboembolism			904	28.6
HASBLED Score				
	0		321	10.5
	1		1468	46.5
	2		921	29.2
	3		330	10.4
	4		94	3
	5		22	0.7
	6		2	0.1

(資料來源:本研究整理)

3.2.2 研究變項

本研究採用評估出血的 Hasbled Score 之計分項目作為研究變數(Camm, 2010)。HAS-BLED score 計算出血風險是依照病患是否有 H:高血壓、A:異常的肝&腎功能、S:中風病史、B :出血病史、L :INR 不穩定、E:年齡>65 歲、D:使用特定藥物(增加出血風險的藥物)&飲酒分別來做積分。

3.3 資料前處理

在做資料前處理時須將資料經資料清理、資料整合、資料轉換等資料前處理過程。此步驟是為求避免將資料中不正確、不完整或錯誤的資料影響後續資料探勘的結果。

3.3.1 資料清理

將資料做清理的目的是將所收集到的資料進行整理以及篩選的動作，用已確定資料的品質是後續進行資料探勘所需。本研究將所收集的資料逐一檢查各欄位是否有不屬於或超出該欄位的資料，在資料清洗過程中，檢視每筆資料的完整性並移除有不正確、不完整或錯誤的資料。依取得資料表欄位修正、刪除，留下研究所須知分析資料，期待探勘過程中減少不必要欄位值影響探勘結果如下：

1. 檢查第一次用藥日期是否大於出生年月日。
2. 檢查性別是否有不為「男性」或「女性」。
3. 檢查各個欄位值是否有空白。
4. 刪除無須使用的欄位資料，留下本研究中所須資料欄位。

3.3.3 資料整合轉換

將來自不同資料來源之資料表做合併處理，整合出另一個新資料表。資料轉換是為使欄位的格式或形態符合資料探勘所需，為符合本研究，產生衍生因子變項。因研究所需將原始欄位值離散化，避免數值型態之資料在進

行決策樹分析時造成分析結果過度遷就及分類效率不佳的問題。再將原始資料檔匯入資料庫，產生符合研究目的可供分析探索之資料集。其中為處理方式：

1. 病患年齡:利用病患之出生年月日與第一次用藥日期換算出病患用藥時年齡，並將年齡大於 65 者轉換為「1」，小於者為「0」。
2. 異常的肝功能或腎功能:腎功能異常包括了慢性洗腎、腎臟移植或肌酐酸 >200mmol/L(約 2.26mg/dL)；肝功能異常包括了慢性肝臟疾病或黃疸高於正常值 2 倍以上或肝指數高於正常值 3 倍以上。將肝腎功能欄位合併為異常的肝功能或腎功能欄位，其中肝功能或腎功能若有異常者轉換為「1」，無異常者為「0」。
3. 使用特定藥物(增加出血風險的藥物)&飲酒:將特定藥物欄位，包括 Aspirin、預防血栓用藥、改善血液循環用藥、抗血小板用藥、非類固醇消炎止痛藥及飲酒量等欄位合併為使用特定藥物(增加出血風險的藥物)&飲酒欄位，將各欄位中凡其中任一欄位為有使用者轉換為「1」，所有欄位皆為無使用者為「0」。
4. Hasbled 計分欄位:將高血壓、異常的肝&腎功能、中風病史、出血病史、INR 不穩定、年齡>65 歲、使用特定藥物(增加出血風險的藥物)&飲酒分別來做計分加總，將計分數轉換為 Hasbled 計分欄位。
5. 總出血事件:將新發生腦內出血(New ICH)、其他非 ICH 腦出血(SDH/SAH/other)、血色素下降(Hb drop)、重要部位出血(Critical sites bleeding 包括眼底，後腹腔，大關節，流血致死等)、其他部位出血(other bleeding)、腸胃道出血(GI bleeding)等欄位，將各欄位中凡其中任一欄位為有發生者轉換為「1」，所有欄位皆為無發生者為「0」。

3.4 軟硬體與採礦工具說明

本研究使用硬體環境: Intel Core i7- 7500U CPU, 8G RAM。軟體: Microsoft Office Professional 2010 Excel、Microsoft SQL Server Management Studio 2017、Microsoft Visual Studio 2017。

以 Microsoft Visual Studio 2017 所提供之資料採礦演算法如表 3，本研究建立模型：關聯分析、決策樹、貝氏機率分類、類神經網路。

表 2.2 Microsoft Visual Studio 2017 所提供之資料採礦演算法比較

模型	主要使用目的
決策樹	由訓練集中的連續及離散資料計算輸出的機率
關聯分析	識別各種變數間的關係
naïve Bayes	針對不同資料特定變數中的差異
時序集群	以時間的順序分組或群集化資料並可提供預測
時間序列	分析及預測基於時間連續性數值的資料
類神經網路	發現資料中模擬生物神經元預測出線機率
文字探勘	分析非結構化文字資料

(資料來源：本研究整理)

3.5 模型評估

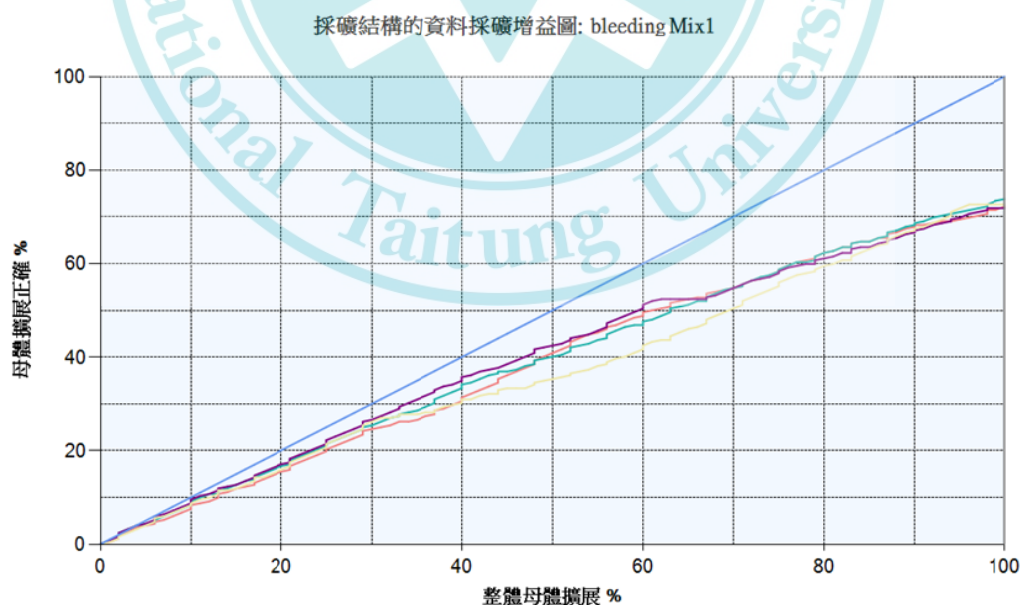


圖 3.2 增益圖範例

(資料來源：本研究整理)

為分析決策樹演算法模型，本研究使用增益圖(Lift Chart)來評估最佳模型。在 Microsoft Visual Studio 2017 建立模型後，分別帶入 30%的測試資料，Microsoft Visual Studio 2017 可產出增益圖提供分析。以圖 3.2 為例，X 軸代表資料採礦模型依據機率由高到低排序後占整體測試資料的百分比，Y 軸代表目標母體(本研究目標母體為有發生出血事件)的百分比，圖中藍色 45 度對角線代表理想模型預測結果，其他預測模型為由其他各色線段表示。

在模型帶入測試資料後，產出增益圖同時，也會產生另一組模型分數(AUC，Area under curve):「曲線下面積」與「理想模型下面積」的比值，比值越接近 1 表示模型的預測能力越好，當比值為 0.5 表示等同於隨機猜測。AUC 數值的範圍從 0 到 1，數值愈大愈好。以下表 3.3 為 AUC 數值一般的判別規則 (Hosmer Jr, 2013)：

表 3.3 AUC 值與模型區辨能力關係表

AUC值	區辨能力
<0.5	無
0.5~0.7	普通
0.7~0.8	尚可
0.8~0.9	良好
>0.9	極高

(資料來源：本研究整理)

表 3.4 錯差矩陣表

預測結果 實際發生	未發生出血事件	有發生出血事件
	未發生出血事件	有發生出血事件
未發生出血事件	TP (True positive)	FN (False Negative)
有發生出血事件	FP (False positive)	TN (True Negative)

(資料來源：由本研究整理)

為評估資料採礦所得的模型是否有真實且有效，可將資料集區分為訓練集及測試集，訓練集用來訓練模型，測試集用來測試模型結果。模型結果的

評估，本研究以錯差矩陣(Confusion Matrix)來顯示模型效能，如圖 3.4，由錯差矩陣(Confusion Matrix)可知模型預測為正確的會有多少，錯誤的會有多少。錯差矩陣表將資料分為四型：

1. 真陽(True Positive, TP)：模型預測為未出血(Positive)，實際資料為未發生出血事件(True)，顯示預測準確。
2. 真陰(True Negative, TN)：模型預測為有發生出血事件(Negative)，實際資料為未發生出血事件(True)，顯示預測準確。
3. 偽陽(Flase Positive, FP)：模型預測為未發生出血事件(Negative)，實際資料為有發生出血事件(True)，顯示預測錯誤。
4. 偽陰(Flase Negative, FN)：模型預測為有發生出血事件(Negative)，實際資料為未發生出血事件(True)，顯示預測錯誤。

Microsoft Visual Studio 2017 可產出分類矩陣(Classification Matrix)，並使用分類矩陣(Classification Matrix)進行評估工作。透過檢視此矩陣之每個資料格中的數量和百分比，查看模型正確預測的頻率，此為容易了解及說明錯誤預測的影響。以圖 3.3 為例，圖中橫軸代表預測的計數，縱軸則是實際的計數。第一個結果資料格包含 64 的值，表示值 0 的「真肯定」數目。由於 0 表示未發生出血事件，所以這個資料表示此模型在 64 個案例中預測出未發生出血事件正確值。該資料格正下方的資料格包含 50 的值，也就是「誤判」的數目，是此模型預測未發生出血事件，但實際上卻沒有的次數。在 20 的值的資料格，表示值 1 的「誤判」數目。由於 1 表示未發生出血事件，所以表示此模型在 20 個案例中預測出未發生出血事件，但實際上卻有發生出血事件的次數。最後，包含 118 值的資料格，表示目標值 1 的「真肯定」數值。表示此模型在 118 個案例中，正確預測未發生出血事件。藉由彙總對角線上相鄰之資料格中的值，就可以判斷此模型的整體精確度。一條對角線為正確預測的總數，而其他對角線則是錯誤預測的總數。

在 Bleeding Event 上 bleeding Mix1 的計數:		
預測的	0 (實際)	1 (實際)
0	64	20
1	50	118

圖 3.3 預測分類矩陣
(資料來源：由本研究整理)

本研究使用的評估指標如下：

- (1) 整體分類正確率(overall accuracy) = $(TP + TN) / (TP + FN + FP + TN)$ 為模型整體分類發生的正確率，即 TP(模型預測為未出血(Positive)，實際資料為未發生出血事件(True))與 TN(模型預測為有發生出血事件(Negative)，實際資料為未發生出血事件(True))之間的比率。
- (2) 遺漏性誤差(omission error) = $FN / (TP + FN)$ 為分類錯誤的比率，即模型預測為有發生出血事件(Negative)，實際資料為未發生出血事件(True)。
- (3) 過度預測誤差(commission error) = $FP / (FP + TN)$ 為分類錯誤的比率，模型預測為未發生出血事件(Negative)，實際資料為有發生出血事件(True)。

第四章 研究結果

本章共分五節，主要說明本研究建置模型的過程與結果。前四節說明建置模型的方式，第五節為模型評估與說明。

4.1 資料淨化、整理

由某地區醫院收集記錄自 2013.01.01-2018.04.30 期間內(共 3158 筆)使用 Rivaroxban 的 AF 病患用藥前後之狀態，資料欄位如表 4.1 所示。

表 4.1 原始資料欄位說明

項次	生日	出生年月日
1	Gender	性別 (0=女 1=男)
2	收案第一次用藥日期	收案第一次用藥日期
3	第一次用藥時年齡	第一次用藥時年齡(0 為<65 1 為≥65)
4	第一次用藥時 Age	第一次用藥時 Age (0 為<65；1 為 65~74；2 為≥75)
5	一開始併用最久的 AAD 抗心律不整藥物	一開始併用最久的藥物(0=無；1=Drone；2=Propa；3=Amio；4=Sotalol)
6	AF 種類	AF 心房纖維性顫動種類 (0=無；1=Paroxysmal 突發性；2=Persistent 堅持的；3=Permanent 永久性)
7	第一次發現患有 Af 方式	第一次發現有 Af(0=查不到；1=ECG 心電圖；2=VVIR；3=PPM 心律調節器)
8	PPM 種類 (0/AAIR:1/VVIR:2/DDD:3/CRT:4)	心律調節器種類 (0=無；1=AAIR；2=Holter 24 小時心電圖；3=DDD；4=CRT)
9	服用 NOAC 原因	因 DVT 深部靜脈栓塞(1=PE 肺栓塞；2=AF 心房纖維顫動；3=其他或找無原因；4=吃 NOAC(藥物))
10	因心臟衰竭住院 (CHF Admission) 病史	之前因心臟衰竭住院(CHF Admission)病史 (0=無；1=有)

11	HTN History or >140/90	HTN History or >140/90 有高血壓病史 或血壓高於140/90(0=無；1=有)
12	Uncontrol HTN (SBP >160)	Uncontrol HTN (SBP >160)無控制高血壓或舒張壓>160 (0=無；1=有)
13	DM History	糖尿病史 (0=無；1=有)
14	Prior major bleeding or predisposition to bleeding	先前大出血或易出血 (0=無；1=有)
15	Labile INR 曾經 >3	抗凝血指數曾經大於 3 (0=無；1=有)
16	Ischemic stroke (0	出血性中風 (0=無；1=有)
17	ICH/SDH/SAH	ICH(出血性中風)/SAH(蜘蛛膜下腔出血)/SDH(硬腦膜下出血) (0=無；1=有)
18	Any Stroke(包括出血和梗塞)/TIA/Thromboembolism	Any Stroke(包括出血和梗塞)/TIA(短暫性腦缺血)/Thromboembolism(血栓栓塞) (0=無；1=有)
19	Any stroke 包括出血和梗塞	排除 Any stroke 包括出血和梗塞 (0=無；1=有)
20	除了 stroke 之外的 Vascular disease	除了 stroke 中風之外的 Vascular disease 血管疾病(0=無；1=ACS 急性冠狀動脈症候群；2=CAD 冠狀動脈疾病；3=PAOD 周邊動脈阻塞；4=Aortic plaque 主動脈斑塊)
21	ACEI/ARB	降血壓藥 (0=無使用；1=ACE；2=ARB 乙型阻斷劑)
22	BB	使用 B-blocker(阻斷劑) (0=無；1=有)
23	Statin	使用 Statin(降膽固醇藥) (0=無；1=有)
24	alcohol	飲酒習慣(每周喝 8 瓶以上) (0=無；1=有)
25	Aspirin	使用 Aspirin(阿斯匹靈) (0=無；1=有)
26	Plavix/Brilinta	使用 Plavix(藥品名) Brilinta(藥品名)(0=無使用；1=Plavix；2=Brilinta)

27	Ticlopidine	使用 Ticlopidine(藥品名) (0=無；1=有)
28	Cilostazol(Pletaal)	使用 Cilostazol(Pletaal)(藥品名) (0=無；1=有)
29	NSAID	使用 NSAID(非類固醇類止痛藥) (0=無；1=有)
30	Anti-Platelet or NSAID	使用 Anti-Platelet(抗血小板藥物)or NSAID(非類固醇類止痛藥) (0=無；1=有)
31	是否有腎病 Renal disease	Renal disease (Cr>2.26 or H/D(洗腎) or 移植) (0=無；1=有)
32	曾經出現過肝臟疾病 Liver disease	曾經出現過 Liver disease(LC or bili>2x+ALT/AST>3x) (0=無；1=有)
33	Liver cirrhosis	有無 Liver cirrhosis(肝硬化) (0=無；1=有)
34	CV death 原因	發生 CV death(心血管疾病死亡)原因 (1=SCD 猝死)；2=MI 心肌梗塞)；3=HF 心臟衰竭)；4=others 其他)
35	non-fatal MI	發生 non-fatal MI(非致命性心肌梗塞) (0=無；1=有)
36	non fatal-MI date	發生 non-fatal MI(非致命性心肌梗塞) 日期
37	New Ischemic Stroke	是否發生 New Ischemic Stroke(新發生缺血性中風) (0=無；1=有)
38	All cause mortality	All cause mortality(所有死亡原因) (0=無；1=有)
39	再一次 HF 住院	再一次 HF(心臟衰竭)住院(0=無；1=有)
40	New Embolism	New Embolism(新的血栓) (0=無；1=DVT(深部靜脈栓塞)；2=PE(肺栓塞)；3=renal infarction(腎梗塞)；4=ischemia bowel(缺血腸)；5=週邊動脈梗塞；6=其他)
41	ICH	有無發生 ICH(顱內出血) (0=無；1=有)
42	ICH 總類 Other / SDH 或 SAH (0/1)	ICH (顱內出血)總類 (0= Other；1=SDH(硬膜下腔出血)或 SAH (蜘蛛膜下腔出血))
43	Hb drop > 2 or 輸血 4U	Hb drop(血色素數值) > 2 or 輸血 4U(單位)

44	Critical sites bleeding(眼底，後腹腔，大關節，流血致死)	有無發生 Critical sites bleeding(眼底，後腹腔，大關節，流血致死) (0=無；1=有)
45	Any other bleeding	有無發生 Any other bleeding (非 ICH, critical sites, Hb drop >2, or 輸血>4U) (0=無；1=有)
46	Any other bleeding 原因	發生 Any other bleeding 原因
47	GI bleeding	有無發生 GI bleeding(腸胃道出血) (0=無；1=有)
48	Hold AAD bradycardia	有無發生 Hold AAD as bradycardia(心跳過緩) (0=無；1=有)
49	ARF	發生 ARF(急性腎功能衰竭)(Cr 上升>3X or H/D(洗腎)) (0=無；1=有)
50	Hypothyroidism	是否發生 Hypothyroidism(甲狀腺功能低下) (0=無；1=有)
51	Hyperthyroidism	是否發生 Hyperthyroidism(甲狀腺功能亢進) (0=無；1=有)
52	Lung fibrosis	是否發生 Lung fibrosis(肺纖維化) (0=無；1=有)
53	HAS-BLED 計算出血風險數值	出血風險的 HAS-BLED score，依照病患是否有 H 高血壓、A 肝&腎功能異常、S 中風病史、B 出血病史、L INR 不穩定、E 年齡>65 歲、D 使用特定藥物(增加出血風險的藥物)&飲酒分別來給分，其中肝功能異常和腎功能異常各佔 1 分，飲酒和特定藥物也是各佔 1 分。加總總分，如果小於 3 分為低出血風險，較可以安心用藥，但如果大於 3 分，病患為高出血風險(出血機率>3.74%)，必須謹慎小心、密切監測是否有出血的表徵和症狀(sign&symptom)。

(資料來源：本研究整理)

為符合本研究之變數使用，刪除不須使用之欄位，並加入「總出血事件」欄位及 HAS-BLED 計算用欄位，將各欄位數值轉換成為是可運算數值，如表 4.2。

表 4.2 淨化、處理後資料欄位說明

項次	欄位名稱	值	說明
1	流水號	不改變 作索引	僅為資料出時的流水號
2	生日	數值	本研究不需要
3	Gender	轉換 0 代表女 1 代表 男	
4	第一次用藥時年齡	數值	0 代表年齡 <65 1 代表年齡 >65
5	第一次用藥時 Age <65(0)/65~74(1)/>=75(2)	轉換 0 代表年齡 <65；1 代表 齡>65~74；2 代表 齡 >=75	本研究不需要
6	併用最久的 AAD 抗心律不整藥	併用最久的藥物(0 = 無；1=Drone2=Propa；3=Amio；4=Sotalol)	併用最久的抗心律不整藥物
7	AF 種類	轉換 AF 心房纖維性顫動種類 (0 = 無；1 = Paroxysmal 突發性；2 = Persistent 堅持的；3 = Permanent 永久性)	Af 心房纖維性顫動分類
8	NOAC 使用原因	因 DVT 深部靜脈栓塞(1=PE 肺栓塞；2=Af 心房纖維顫動；3=其他或找無原因；4=吃 NOAC(藥物))	NOAC 用藥原因
9	CHF Admission	轉換 之前因心臟衰竭住院(CHF Admission)病史(0=無；1=有)	是否有因心臟衰竭住院病史
10	HTN History 有高血壓病史 or >140/90	轉換 HTN History 有高血壓病 or >140/90	本研究不使用

		(0=無；1=有)	
11	Uncontrol HTN (SBP >160)	轉換 0=無；1=有	Uncontrol HTN (SBP >160) 無控制高血壓或 舒張壓>160
12	DM History	轉換 0=無；1=有	糖尿病史
13	Prior major bleeding or predisposition to bleeding	轉換 0=無；1=有	先前大出血或易 出血病史
14	Labile INR 曾經>3	轉換 0=無；1=有	抗凝血指數曾經 大於 3
15	Ischemic stroke(0/1)	轉換 0=無；1=有	出血性中風病史
16	ICH/SDH/SAH	轉換 0=無；1=有	ICH(出血性中 風)/SAH(蜘蛛膜 下腔出血) /SDH(硬腦膜下 出血)病史
17	Any Stroke(包括出血 和梗塞)/TIA/ Thromboembolism	轉換 0=無；1=有	Any Stroke(包括 出血和梗塞)/ TIA(短暫性腦缺 /Thromboemboli sm(血栓栓塞)病 史
18	Any stroke 包括出血 和梗塞	轉換 0=無；1=有	排除 Any stroke 包括出血和梗塞 病史
19	除了 stroke 之外 Vascular disease	轉換 0=無；1=ACS 急性冠狀動脈症候 群； 2=CAD 冠狀動 脈疾病；3=PAOD 周 邊動脈阻塞 4=Aortic plaque 主動脈斑塊	除 stroke 中風之 外的 Vascular disease 血管疾病 病史
20	ACEI/ARB	轉換 0=無；1=有	是否使用降血壓 藥
21	BB	轉換 0=無；1=有	是否使用

			B-blocker(阻斷劑)
22	Statin	轉換 0=無；1=有	有否使 Statin(降膽固醇藥)
23	alcohol(≥ 8 drinks/week)	轉換 0=無；1=有	飲酒量有 ≥ 8 drinks/week
24	Aspirin	轉換 0=無；1=有	HAS-BLED 計算用
25	Plavix/Brilinta	轉換 0=無；1=有	HAS-BLED 計算用
26	Ticlopidine	轉換 0=無；1=有	HAS-BLED 計算用
27	Cilostazol(Pletaal)	轉換 0=無；1=有	HAS-BLED 計算用
28	NSAID	轉換 0=無；1=有	HAS-BLED 計算用
29	Anti-Platelet or NSAID	轉換 0=無；1=有	HAS-BLED 計算用
30	HAS-BLED 計算用	轉換 0=無；1=有	
31	Renal disease (Cr >2.26 or H/D or 移植)	轉換 0=無；1=有	Renal disease (Cr >2.26 or H/D(洗腎) or 移植) 病史
32	曾經出現過 Liver disease(LC or bili $>2x$ +ALT/AST $>3x$)	轉換 0=無；1=有	曾經出現過 Liver disease(LC or bili $>2x$ +ALT/AST $>3x$)病史
33	Liver cirrhosis	轉換 0=無；1=有	Liver cirrhosis(肝硬化) 病史
34	發生 CV death(心血管疾病死亡)原因	轉換 1=SCD 猝死； 2=MI 心肌梗塞； 3=HF 心臟衰竭； 4=others 其他	用藥後發生 CV death(心血管疾病死亡)原因
35	non-fatal MI	轉換 0=無；1=有	用藥後發生 non-fatal MI(非

			致命性心肌梗塞) 病史
36	New Ischemic Stroke	轉換 0=無 ; 1=有 0	用藥後發生 New IschemicStroke(新發生缺血性中風) 病史
37	All cause mortality	轉換 0=無 ; 1=有	用藥後 All cause mortality(所有死亡原因)
38	再一次 HF(心臟衰竭) 住院	轉換 0=無 ; 1=有	用藥後再一次 HF(心臟衰竭) 住院病史
39	New Embolism	轉換 0=無 1=DVT(深部靜脈栓塞) ; 2=PE(肺栓塞) ; 3=renal infarction(腎梗塞) ; 4=ischemia bowel(缺血腸) ; 5=週邊動脈梗塞 ; 6=其他	用藥後 New Embolism(新的血栓) 病史
40	ICH	轉換 0=無 ; 1=有	用藥發生 ICH(顱內出血) 病史
41	ICH 總類	轉換 0=無 ; 1=有	用藥後 ICH (顱內出血) 病史
42	Hb drop > 2 or 輸血 4U	轉換 0=無 ; 1=有	用藥後 Hb drop(血色素數值) > 2 or 輸血 4U(單位) 病史
43	Critical site bleeding(眼底，後腹腔，大關節，流血致死)	轉換 0=無 ; 1=有	用藥後發生 Critical sites bleeding(眼底，後腹腔，大關節，流血致死) 病史
44	Anyotherbleeding (非 ICH, critical sites, Hb drop >2, or 輸血>4U)	轉換 0=無 ; 1=有	用藥後發生 Any other bleeding (非 ICH, critical sites, Hbdrop >2, or 輸血>4U)

45	Any other bleeding 原因	轉換 0=無；1=有	用藥後發生 Any other bleeding 病史
46	GI bleeding	轉換 0=無；1=有	用藥後發生 GI bleeding(腸胃道出血)病史
47	Hold AAD as bradycardia(心跳過緩)	轉換 0=無；1=有	本研究不需要
48	ARF(Cr 上升>3X or H/D)	轉換 0=無；1=有	本研究不需要
49	Hypothyroidism	轉換 0=無；1=有	本研究不需要
50	Hyperthyroidism	轉換 0=無；1=有	本研究不需要
51	Lung fibrosis	轉換 0=無；1=有	本研究不需要
52	HAS-BLE 計算出血風險數值	數值	出血風險的 HAS-BLED score，依照病患是否有:H 高血壓、A 肝&腎功能異常、S 中風病史、B 出血病史、L INR 不穩定、E 年齡>65 歲、D 使用特定藥物(增加出血風險的藥物)& 飲酒分別來給分，其中肝功能異常和腎功能異常各佔 1 分，飲酒和特定藥物也是各佔 1 分。加總總分。

(資料來源：本研究整理)

將總樣本數量共 3158 筆資料(如表 4.3)，區分為有發生出血事件集及無發生出血事件集，以 EXCEL 亂數功能自無發生出血事件集分別取出與發生出血事件筆數相等分別為 1:1 之資料集，再與發生出血事件 436 筆資料，合併為新資料表。再將合併後新資料表使用 Microsoft Visual Studio 2017 中 Integration Services 功能將合併後新資料表分為兩部分，一為 70%做為模型訓練使用，另一 30%做為模型測試時使用。再將模型訓練用資料表及模型測試用資料表使用 Microsoft SQL Server Management Studio 2017 匯入資料庫中供資料採礦分析用。

表 4.3 EXCEL 淨化 整理後的資料表

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S						
No	Gender	第一次用	第一次用	第一次用	Combined	AAD	prot	SBP	DBP	AF group	PPM	grou	Cause	Cause1	pre CHF	HTN	histc	Uncontrol	DM	histor	Prior	max	Labile	INIs
1	1	81	1	1	1	0	0	110	56	2	0	3	3	0	1	0	0	0	1	1				
2	0	78	1	1	0	0	0	125	59	1	0	3	3	1	1	0	1	1	1	1				
3	1	76	1	1	3	0	0	135	73	1	3	3	3	0	1	1	1	1	1	1				
4	1	82	1	1	2	0	0	181	81	1	0	3	3	0	1	1	1	0	1	1				
5	1	79	1	1	1	0	0	166	100	1	0	3	3	0	1	1	1	1	1	1				
6	1	76	1	1	0	0	0	117	66	1	3	3	3	1	1	0	0	1	1	0				
7	1	82	1	1	0	0	0	118	80	3	0	2	2	1	1	1	1	0	0	1				
8	0	80	1	1	0	0	0	110	73	3	0	3	3	1	1	0	0	1	1					
9	1	66	1	1	0	0	0	142	95	3	0	3	3	0	1	1	1	1	1	1				
10	0	81	1	1	0	0	0	100	62	0	0	3	3	0	1	0	0	0	1	1				
11	0	69	1	1	0	0	0	110	75	3	0	3	3	0	1	0	0	0	0	1				
12	0	76	1	1	0	0	0	140	85	1	0	3	3	0	1	0	0	0	0	1				
13	1	86	1	1	1	0	0	120	62	1	3	3	3	1	1	1	1	1	1	1				
14	1	73	1	1	0	0	0	113	64	1	0	3	3	1	1	1	1	1	1	0				
15	1	92	1	1	0	0	0	140	90	3	0	3	3	1	1	1	1	0	1	1				
16	0	87	1	1	0	0	0	96	55	3	0	3	3	1	1	0	1	1	1	0				
17	1	88	1	1	0	0	0	133	74	1	0	3	3	0	1	0	0	0	0	1				
18	0	76	1	1	0	0	0	198	88	1	0	3	3	1	0	1	0	0	0	1				
19	1	70	1	1	3	0	0	166	93	1	0	3	3	1	1	0	1	0	1	0				
20	1	91	1	1	0	0	0	132	65	1	0	3	3	1	1	1	1	1	1	1				
21	1	85	1	1	3	1	0	108	100	3	2	3	3	1	1	0	0	0	0	1				
22	1	84	1	1	0	0	0	119	67	3	0	3	3	0	0	0	0	0	0	1				
23	0	84	1	1	0	0	0	116	72	3	0	3	3	1	1	0	0	1	1	0				

(資料來源：：本研究整理)

4.2 資料探勘模型建置與執行

本章節說明本研究欲使用資料庫建置分析後，便是資料探勘模型的建立與訓練。將以 Microsoft Visual Studio 2017 軟體建置資料探勘模型說明方法與過程如下：

4.2.1 建立資料來源

首先使用 Microsoft SQL Server Management Studio 2017 將已經資料淨化及轉換後匯入資料庫中，供本研究資料採礦分析用。如圖 4.1、圖 4.2。

The screenshot shows the SQL Server Enterprise Manager on the left with a tree view of databases. The main window displays a SQL query and its results. The query is a SELECT TOP (1000) statement with various columns. The results table has 10 columns: No, Gender, 第一次用藥時年齡, 第一次用藥時年齡<65yo(0/1), 第一次用藥時年齡<65(0)/65-74(1)/>=75(2), Combined AAD group, AAD problem, and AF group.

No	Gender	第一次用藥時年齡	第一次用藥時年齡<65yo(0/1)	第一次用藥時年齡<65(0)/65-74(1)/>=75(2)	Combined AAD group	AAD problem	AF group	
1	2	0	78	1	2	0	0	1
2	1	1	81	1	2	1	0	2
3	6	1	76	1	2	0	0	1
4	3	1	82	1	2	0	0	3
5	7	0	80	1	2	0	0	3
6	10	0	81	1	2	0	0	0
7	11	0	69	1	1	0	0	3
8	25	0	76	1	2	0	0	1

圖 4.1 模型訓練資料表匯進資料庫資料

(資料來源：本研究整理)

The screenshot shows the SQL Server Enterprise Manager on the left with a tree view of databases. The main window displays a SQL query and its results. The query is a SELECT TOP (1000) statement with various columns. The results table has 10 columns: No, Gender, 第一次用藥時年齡, 第一次用藥時年齡<65yo(0/1), 第一次用藥時Age <65(0)/65-74(1)/>=75(2), Combined AAD group, AAD problem, and AF group.

No	Gender	第一次用藥時年齡	第一次用藥時年齡<65yo(0/1)	第一次用藥時Age <65(0)/65-74(1)/>=75(2)	Combined AAD group	AAD problem	AF group
1	8	1	76	1	2	0	1
2	4	1	82	1	2	0	1
3	9	1	79	1	2	0	1
4	5	1	66	1	1	0	3
5	18	0	87	1	2	0	3
6	14	1	70	1	1	0	1
7	12	1	91	1	2	0	1
8	13	1	85	1	2	1	3
9	16	1	81	1	2	0	3
10	19	1	77	1	2	1	1
11	22	1	81	1	2	0	3
12	61	1	71	1	1	0	0

圖 4.2 模型測試用資料表匯進資料庫資料

(資料來源：本研究整理)

4.2.2 新增採礦結構及設定參數

開啟 SQL Server Business Intelligence Development Studio 建立專案，此專案對應單一資料來源檢視，每一單一資料來源檢視又對應單一資料來源，設定資料來源為待探勘的資料於本研究即為前一節所建立的 Rivaroxban 資料庫。

在 SQL Server Business Intelligence Developmnt Studio 軟體中提供了不同演算法共七種，內容如表 4.4 所示。本研究主要目的為建立預測模型，故在選擇訓練資料探勘模型之主要演算法選取具有預測功能之相關演算法，只採用關聯規則、決策樹、類神經網路及貝氏機率分類來尋求最佳預測模型。

表 4.4 SQL Server Business Intelligence Developmnt Studio 提供資料探勘演算法

演算法名稱	使用目的	本研究是否使用
關聯規則 Ass0ciation Rules	變數之間的關聯性	是
決策樹 Decicion Tree	預測分類出現機率 以樹狀分類方式	是
類神經網路 Neural Network	預測出現機率 具有學習功能	是
貝氏機率分類 Bayesian Classifier	預測分類與出現機率	是
集群分析 Clutering	分類	否
線性迴歸 Linear Regression	自變數與應變數之 線性關係	否
羅吉斯迴歸 Logistic Regression	二分類、次序變數的 迴歸模式	否二分類、次序變數 的迴歸模式

資料來源：本研究整理

在選擇資料探勘演算法之後，選擇資料探勘預測模型將使用的各個輸入及輸出變數。在本研究中，將流水號設為索引鍵，作為分辨每筆資料的關鍵值。在資料採礦結構中指定總出血事件是演算法的預測值，設為可預測，所有預測模型將輸入變數進行不同演算法的分析處理後，將會產生「總出血事件」的新值，此一新值將與資料庫中的值加以比較後即可知預測模型知準確度。其餘變數皆可做為預測模型輸入值，挑選適當的變數作為輸入值可提高預測

模型知準確度病減少演算法執行時間。本研究採用評估出血的 Hasbled Score 之計分項目作為研究變數，將其欄位設為輸入，如圖 16 所示。



圖 4.3 指定分析所使用的資料行
(資料來源：本研究整理)

在完成資料探勘預測模型變數設定後，各變數之資料類型也必須被定義，本研究之各變數資料類型定義於表 4.5。SQL Server Business Intelligence Development Studio 軟體對內建變數資料類型設定如表 4.6。經各步驟確認後，如圖 4.4。將資料探勘所使用變數類型設定如圖 4.5 所示。最後設定測試資料及的數量與採礦結構及採礦命名，即新增採礦結構的設定完成。

表 4.5 變數資料類型定義

項次	欄位名稱	值	說明
1	流水號	不改變 作為索引	僅為資料匯出時的流水號
2	第一次用藥時年齡	數值	0 代表年齡 <65 1 代表年齡 >65
3	Uncontrol HTN (SBP >160)	轉換 0=無； 1=有	Uncontrol HTN (SBP >160) 無控制高血壓或舒張壓>160
4	Prior major bleeding or predisposition to bleeding	轉換 0=無； 1=有	先前大出血或易出血病史
5	Labile INR 曾經>3	轉換 0=無； 1=有	抗凝血指數曾經大於 3
6	Any Stroke(包括出血和梗塞) /TIA/Thromboembolism	轉換 0=無； 1=有	Any Stroke(包括出血和梗塞) /TIA(短暫性腦血)/Thromboembolism(血栓栓塞) 病史
7	alcohol(≥ 8 drinks/week)	轉換 0=無； 1=有	飲酒量有 ≥ 8 drinks/week
8	HAS-BLED 計算用	轉換 0=無； 1=有	
9	Renal disease (Cr>2.26 or H/D or 移植)	轉換 0=無； 1=有	Renal disease (Cr>2.26 or H/D(洗腎) or 移植) 病史
10	曾經出現過 Liver disease(LC or bili>2x+ALT/AST>3x)	轉換 0=無； 1=有	曾經出現過 Liver disease(LC or bili>2x+ALT/AST>3x) 病史
11	總出血試驗	轉換 0=無； 1=有	

資料來源：本研究自行整理

表 4.6 SQL Server Business Intelligence Development Studio 定義之資料類型

演算法使用知資料類型	說明
Continuous	連續型資料
Cyclical	循環或週期性資料
Discrete	離散性資料
Discretized	被離散化資料
Ordered	有順序性資料

資料來源：本研究自行整理

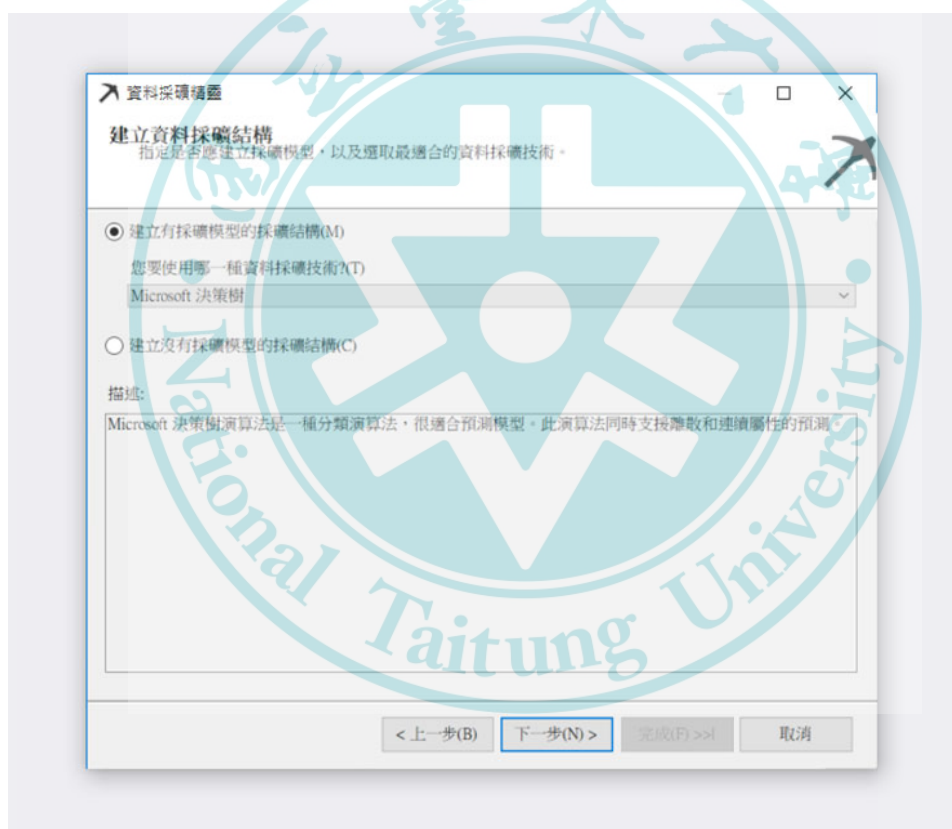


圖 4.4 選擇資料採礦技術

資料來源：本研究整理

資料採礦精靈

指定資料行的內容和資料類型
指定採礦結構資料行的內容和資料類型。

採礦模型結構(S):

資料行	內容類型	資料類型
☒ Alcohol	Discrete	Text
☒ Any Stoke	Discrete	Text
☒ Bleeding Event	Discrete	Text
☒ HAS-BLED	Discrete	Text
☒ Labile INR 曾經3	Discrete	Text
☒ Liver Dis History	Discrete	Text
☒ No	Key	Text
☒ Prior Major Bleeding	Discrete	Text
☒ Renal Dis History	Discrete	Text
☒ Uncontrol HTN SBP 160	Discrete	Text
☒ 第一次用藥時年齡65yo01	Discrete	Text

偵測數值資料行是連續或離散:

< 上一步(B) 下一步(N) > 完成(F) >>I 取消

圖 4.5 本研究所使用變數類型設定

資料來源：本研究整理

為了確保測試資料的一致性及可用性，本研究另外準備 30%測試資料，故不使用系統內的測試資料集，將保留測試資料設定為 0%，如圖 4.6。

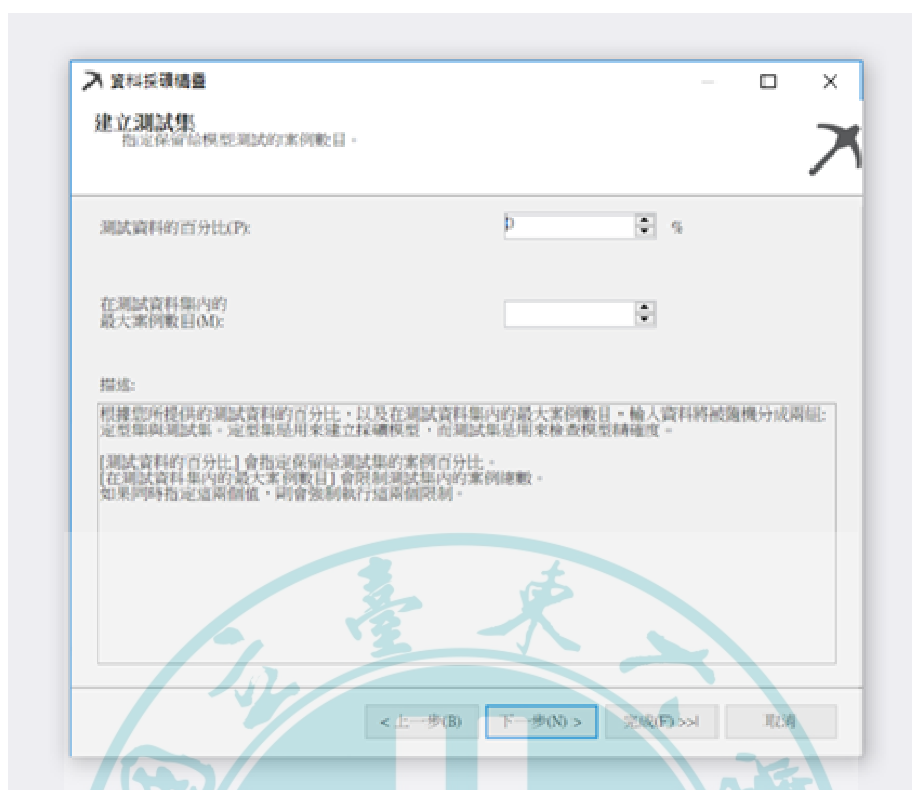


圖 4.6 設定保留測試資料為 0%

資料來源：本研究整理

參數是資料探勘演算法中影響預測模型的重要因素之一，不同演算法又各自特定的參數，需經過反覆嘗試即修正後才能找出最適用於分析資料的特定值才能做出最佳預測模型，本研究所建置模型設定如圖 4.7。

	Microsoft_Associati...	Microsoft_Decision_...	Microsoft_Naive_Ba...	Microsoft_Neural_N...
Alcohol	Input	Input	Input	Input
Any Stoke	Input	Input	Input	Input
Bleeding Event	PredictOnly	PredictOnly	PredictOnly	PredictOnly
HAS-BLED 計算用bd	Input	Input	Input	Input
Labile INR 曾經3	Input	Input	Input	Input
Liver Dis History	Input	Input	Input	Input
No	Key	Key	Key	Key
Prior Major Bleeding	Input	Input	Input	Input
Renal Dis History	Input	Input	Input	Input
Uncontrol HTN SBP 160	Input	Input	Input	Input
第一次用藥時年齡65yo01	Input	Input	Input	Input

圖 4.7 本研究所建置模型設定

資料來源：本研究整理

資料探勘之前置作業完成之後，便可執行採礦模型布署，讓資料探勘程式言連結資料庫，搭配選定之演算法即參數值對資料庫內資料訊組進行資料勘工作，最後所得到的結果一為研究所需之總出血事件測模型。經反覆調整與測試各採礦模型的演算法參數後，指出最符合本研究目的的採礦模型演算法參數，以下將四種演算法所設定參數說明如表 4.7、表 4.8、表 4.9、表 4.10。

表 4.7 研究中使用關聯分析演算法之參數設定說明

參數	說明	參數設定值
MAXIMUN_ITEMSET_COUNT	指定要產生的最大項目集數目	不指定
MAXIMUN_ITEMSET_SIZE	指定一個項目集內所允許的最大項目數目	3
MAXIMUM_SUPPORT	指定一個項目集可以支援的最大案例數目	1
MINIMUM_IMPORTANCE	指定關聯規則的重要性臨界值	-
MINIMUM_ITEMSET_SIZE	指定項目集所允許的最小項目數目	1
MINIMUM_PROBABILITY	指定規則為 True 的最小機率。	0.1
MINIMUM_SUPPORT	指定在產生規則之前，必須包含該項目集之案例的最小數目	0.01

資料來源：本研究自行整理

表 4.8 研究中使用決策樹演算法之參數設定

參數	本研究設定值	說明
COMPLEXITY_PENALTY	0.99	禁止決策樹成長。此值減少會增加拆解的可能性，此值增加則會減少拆解的可能性。
FORCE_REGRESSOR		強制演算法使用指定的資料行作為迴歸公式的迴歸輸入變數，而不考慮其在演算法計算中的重要性。
MAXIMUM_INPUT_ATTRIBUTES	255	指定在叫用特徵選取之前，演算法可處理的最大輸入屬性數目。
MAXIMUM_OUTPUT_ATTRIBUTES	255	指定在叫用特徵選取之前，演算法可處理的最大輸出屬性數目。
MAXIMUM_SUPPORT	10	指定分葉節點必須包含的最小案例數目。
SCORE_METHOD	10	指定用來計算分岔準則的方法。方法有： (1)Entropy;(3)Bayesian with K2 Prior;(4)Bayesian Dirichlet Equivalent with Uniform prior
SPLIT_METHOD	2	指定用來分岔節點的方法。方法有： (1)二元分岔 (2)完整分岔 (3)根據演算法判斷。

資料來源：本研究自行整理

表 4.9 研究中使用貝氏機率分類演算法之參數設定

參數	參數 設定 值	說明
MAXIMUM_INPUT_ATTRIBUTES	255	指定在叫 用特徵選 取之前，演 算法可處 理的 最 大 輸 入 屬 性 數 目
MAXIMUM_OUTPUT_ATTRIBUTES	255	指定在叫 用特徵選 取之前，演 算法可處 理的 最 大 輸 出 屬 性 數
MAXIMUM_STATES	100	指定演算 法所支援 屬性狀態 的 最 大 數 目
MINIMUM_DEPENDENCY_PROBABILITY	0.5	指定介於 輸入和輸 出 屬 性 (Attribute) 之 間 的 最 小 相 依 機 率

資料來源：本研究自行整理

表 4.10 研究中使用類神經網路演算法之參數設定

參數	參數設定值	說明
HIDDEN_NODE_RATIO	4	指定用於判斷隱藏層中之節點數目的數字
HOLDOUT_PERCENTAGE	30	指定定型資料中用於計算此演算法之鑑效組錯誤的案例百分比
HOLEOUT_SEED	0	指定隨機決定這個演算法的鑑效組資料時，用來植入虛擬隨機產生器的數字
MAXIMUM_INPUT_ATTRIBUTES	255	指定在叫用特徵選取之前，演算法可處理的最大輸入屬性數目
MAXIMUM_OUTPUT_ATTRIBUTES	255	指定在叫用特徵選取之前，演算法可處理的最大輸出屬性數目
MAIMUM_STATS	100	指定演算法所支援屬性狀態的最大數目
SAMPLE_SIZE	10000	指定用來定型模型的案例數目

資料來源：本研究自行整理

4.2.3 結果分析

經由上述類型與參數建置，本研究資料為無發生出血事件集與發生出血事件筆數為 1:1 之資料集，共 874 筆，建置關聯規則、決策樹、類神經網路

及貝氏機率分類模型，以目標母體為發生出血事件等於「Yes」，匯入另行保留的 30%測試資料得模型預測結果。

4.2.3.1 關聯規則預測模型

關聯規則演算法是分辨各變數之間的關係。關聯分析演算法在資料集中找出一起出現的項目。接著演算法會將一起出現的相關聯項目分組為項目集，接下來，演算法會從項目集產生規則。再依據演算法識別為是否在重要的其他特定項目中出現，這些規則將用來預測項目在資料庫出現與否。關聯規則演算法之預測模型建置及訓練完成後，可產生相一性網路、項目集、規則。從圖 4.8 可得知，共得 32 種輸入變數與預測值之間相關性。

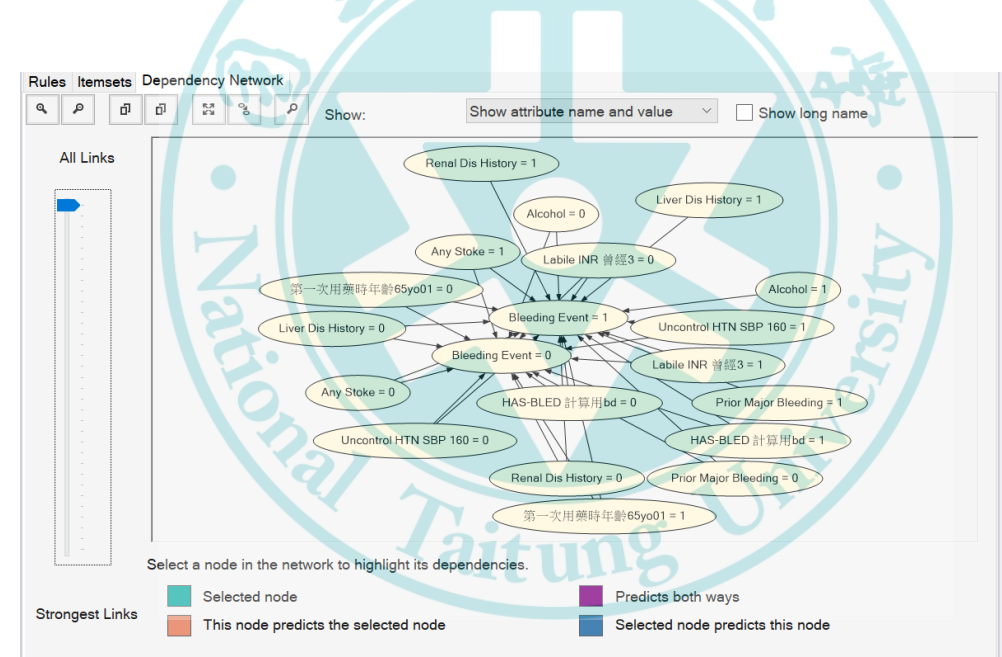


圖 4.8 關聯分析預測模型之相依性網路圖

資料來源：本研究自行整理

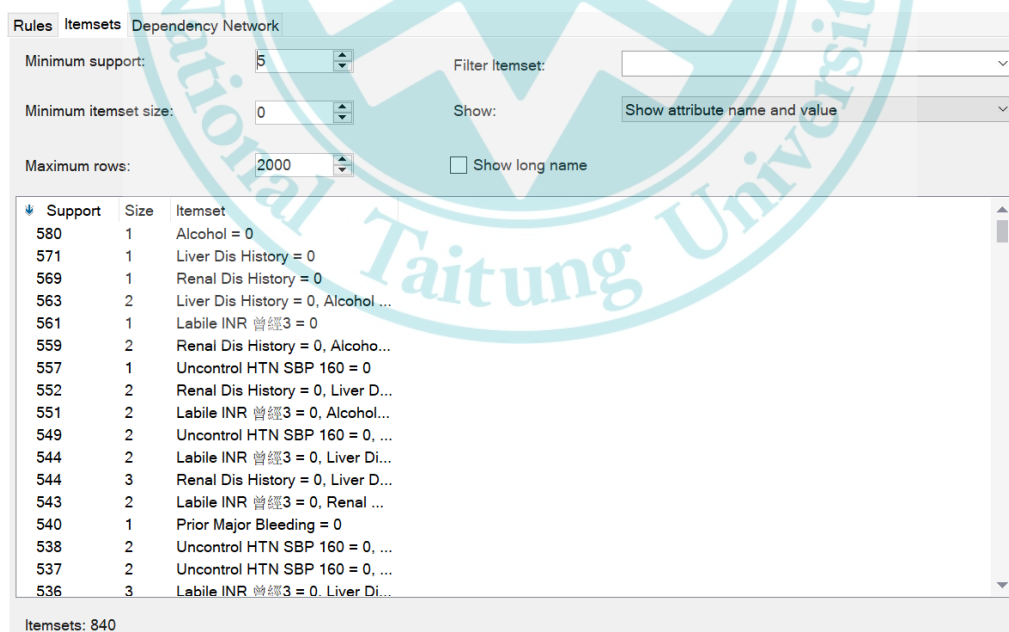
表 4.11 關聯分析預測模型之輸入變數預測能力強弱

輸入變數名稱	預設能力強弱說明(數字越小表示越強)
第一次用藥年齡大於 65 歲	1

先前大出血或易出血病史	2
有無 Renal disease (Cr>2.26 or H/D(洗腎)病史	3
抗凝血指數曾經大於 3	4
肝病史	5
Any Stroke(包括出血和梗塞)病史	6
用藥病史	7
飲酒習慣	8
無控制高血壓或舒張壓>160 病史	9

資料來源：本研究自行整理

資相依性網路圖可得知關聯分析預測模型之輸入變數預測能力強弱，如表 4.11。項目集可透過項目檢視器檢視演算法中產生的物件組及各物件組內容及支援，如圖 4.9。



Support	Size	Itemset
580	1	Alcohol = 0
571	1	Liver Dis History = 0
569	1	Renal Dis History = 0
563	2	Liver Dis History = 0, Alcohol ...
561	1	Labile INR 曾經3 = 0
559	2	Renal Dis History = 0, Alcoho...
557	1	Uncontrol HTN SBP 160 = 0
552	2	Renal Dis History = 0, Liver D...
551	2	Labile INR 曾經3 = 0, Alcohol...
549	2	Uncontrol HTN SBP 160 = 0, ...
544	2	Labile INR 曾經3 = 0, Liver Di...
544	3	Renal Dis History = 0, Liver D...
543	2	Labile INR 曾經3 = 0, Renal ...
540	1	Prior Major Bleeding = 0
538	2	Uncontrol HTN SBP 160 = 0, ...
537	2	Uncontrol HTN SBP 160 = 0, ...
536	3	Labile INR 曾經3 = 0, Liver Di...

Itemsets: 840

圖 4.9 項目檢視器檢視項目集

資料來源：本研究自行整理

規則可由規則檢視器了解關聯規則內容及其信心水準與支援，如圖 4.10。

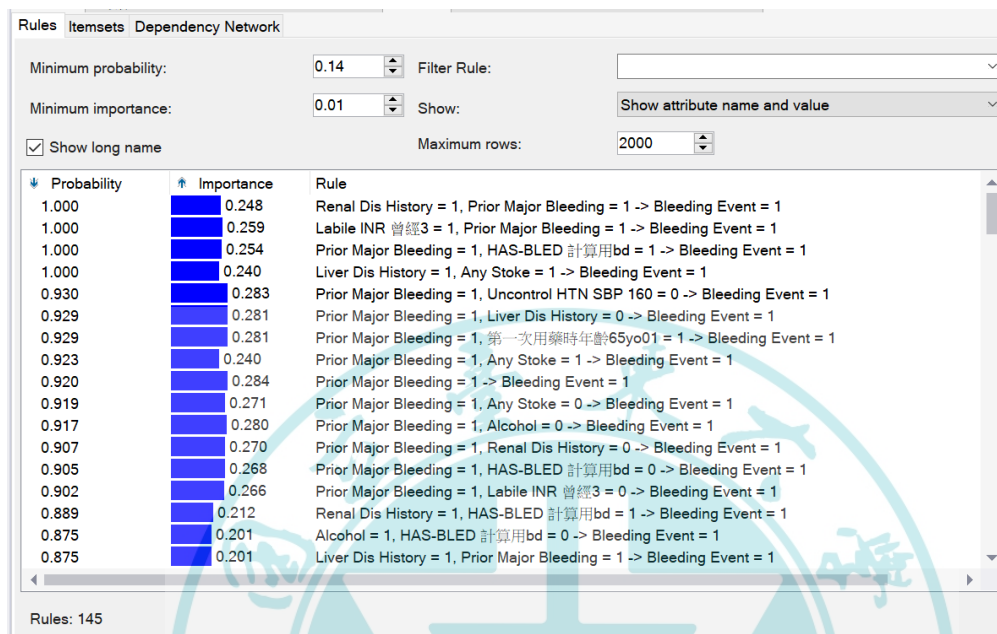


圖 4.10 規則檢視器檢視規則

資料來源：本研究自行整理

4.2.3.2 決策樹預測模型

決策樹預測模型除可用樹狀展示預測規則，也能使用相依性網路檢視那些變數對預測結果是具影響力的。決策樹輸出圖形如圖 4.11 所示。圖中共有三個葉節點，每一葉節點代表一種預測的判斷規則。由決策樹模型檢視可得規則及路徑含意如下表 4.11。以規則三為例，當發生出血事件為 1 及第一次用藥年齡大於 65 為 1 及先前是有大出血或易出血病史時，本模型所預測該事件不發生 9.09%，發生出血事件機率為 90.9%。

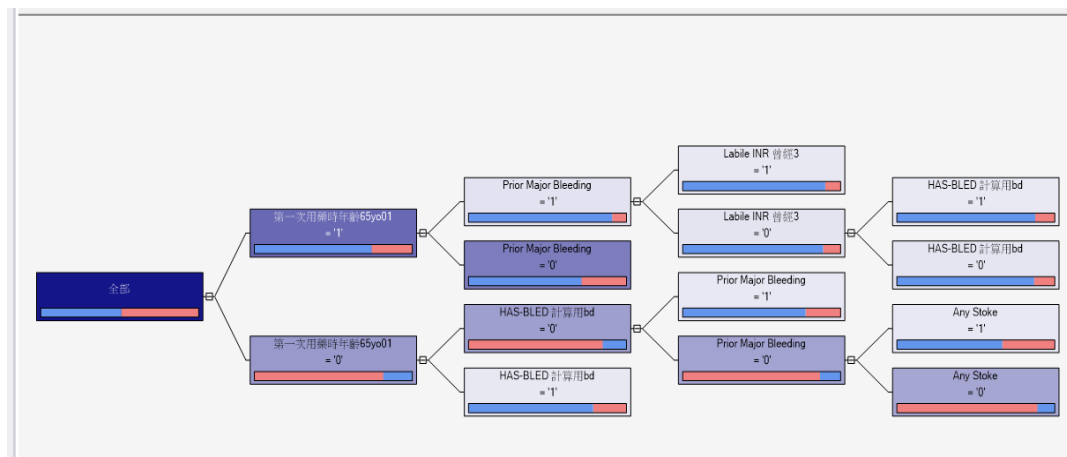


圖 4.11 決策樹輸出圖形

資料來源：本研究整理

表 4.12 決策樹模型路徑規則表

編號	路徑	值	案例	機率
規則一	發生出血事件	0	290	49.1%
		1	300	50.8%
規則二	發生出血事件及第一次用藥年齡大於 65	0	89	26%
		1	255	73.9%
規則三	發生出血事件及第一次用藥年齡大於 65 及先前是有大出血或易出血病史	0	3	9.09%
		1	39	90.9%
規則四	發生出血事件及第一次用藥年齡大於 65 及先前沒有大出血或易出血病史	0	86	28.6%
		1	216	71.3%
規則五	發生出血事件及第一次用藥年齡小於 65	0	201	81.4%
		1	45	18.5%
規則七	發生出血事件及第一次用藥年齡小於 65 及 Has-bleed 計算用(併用藥物)為無	0	2	21.4%
		1	10	78.5%

規則八	發生出血事件及第一次用藥年齡小於 65 及 Has-bleed 計算用(併用藥物)為有	0	199	84.7%
		1	35	15.2%

資料來源：本研究自行整理

相依性網路檢視那些變數對預測結果是具影響力的。決策樹相依性輸出圖形如圖 4.12 所示，從圖中可得知有 5 個輸入變數與預測發生出血事件有關，每一個帶建頭的連線表示預測關係，連線的起點為輸入變數，箭頭所指為終點亦是輸出變數。模型預測能力強弱是一具決策樹的層級數來定，也就是在決策樹第一層的輸入變數即為預測能力最強的輸入變數。本研究決策樹預測模型各輸入變數預測能力強弱如表 4.13 所示。

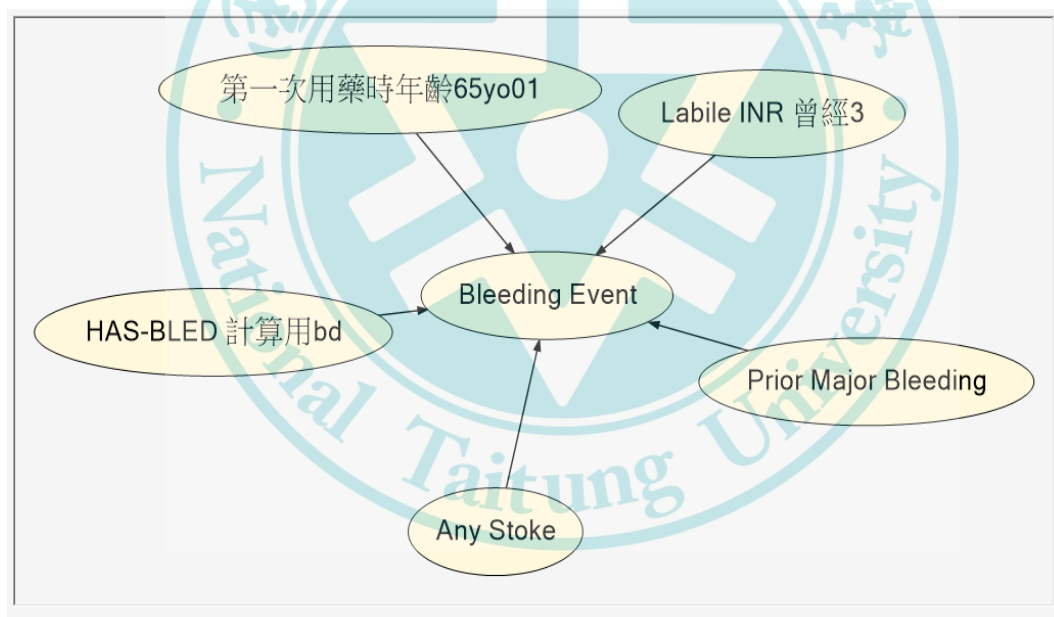


圖 5 決策樹相依性輸出圖形

資料來源：本研究自行整理

表 4.13 決策樹預測模型各輸入變數預測能力強弱

輸入變數名稱	預設能力強弱說明(數字越小表示越強)
第一次用藥年齡大於 65 歲	1
用藥病史	2
先前沒有大出血或易出血病史	3

(資料來源：本研究自行整理)

4.2.3.3 貝氏機率分類預測模型

貝氏機率分類演算法會計算輸入資料行和可預測資料行之間的條件式機率，並假設資料行是獨立的。在預測模型建置與訓練完成後，可提供使用者檢視與判讀的有相依性網路、屬性設定檔、屬性特性、屬性辨識等。由圖 4.13 可知，共有 7 種輸入變數與預測值有相關，以相依的機率決定預測能力的強研究貝氏機率分類預測模型各輸入變數預測能力強弱表 4.14 所示。

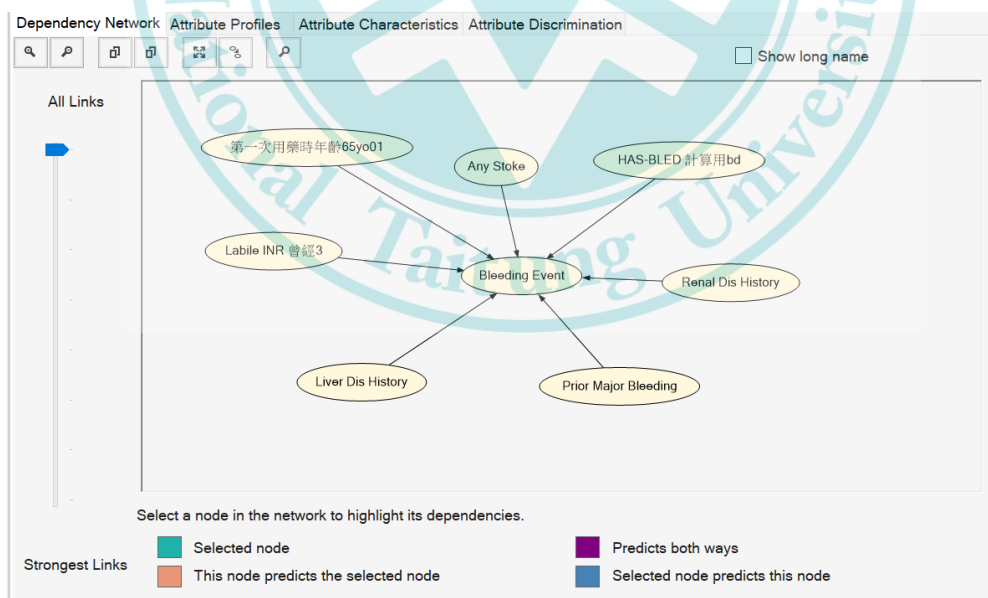


圖 4.13 貝氏機率分類預測模型之相依性網路圖

資料來源：本研究自行整理

表 4.14 貝氏機率分類預測模型各輸入變數預測能力強弱

輸入變數名稱	預設能力強弱說明(數字 越小表示越強)
第一次用藥年齡大於 65 歲	1
先前大出血或易出血病史	2
Any Stroke(包括出血和梗塞) 病史	3
用藥病史	4
抗凝血指數曾經大於 3	5
有無 Renal disease ($Cr > 2.26$ or H/D(洗腎)病史	6
肝病史	7

資料來源：本研究整理

從貝氏機率分類預測模型之屬性設定檔可了解各輸入參數在各預測值的機率分布，如圖 4.14。圖中每列為各輸入變數的資料分布情形並以不同顏色代表不同參數值範圍。



圖 4.14 貝氏機率分類預測模型之屬性設定檔

資料來源：本研究整理

在屬性特性圖中，可檢視預測結果與輸入變數之間的預測機率值。如圖 4.15 所示，以發生出血事件為預測值，將各變數的值與其預測機率由高到低依序排列。

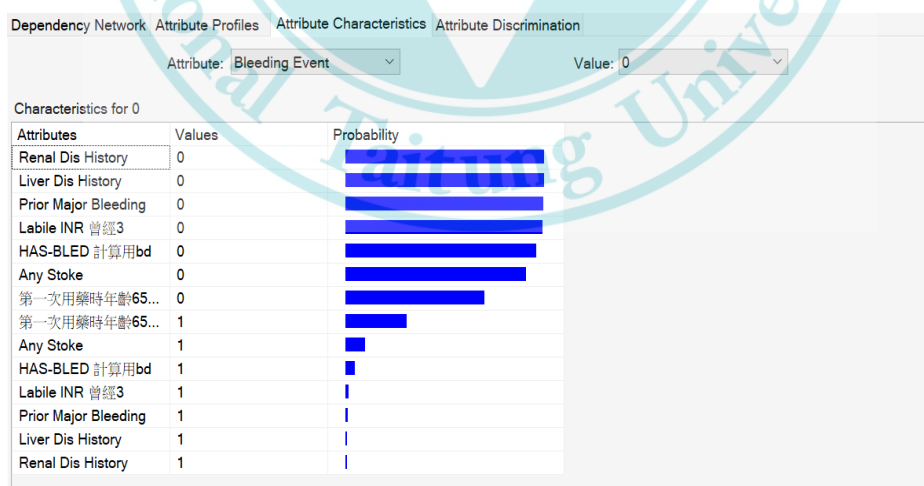


圖 4.15 貝氏機率分類預測模型之屬性特性圖

資料來源：本研究整理

比較具有比較目的分析圖是屬性辨識圖，因可將不同值的預測結果比較之下，有較高預測機率的輸入變數是哪一種，如圖 4.16。

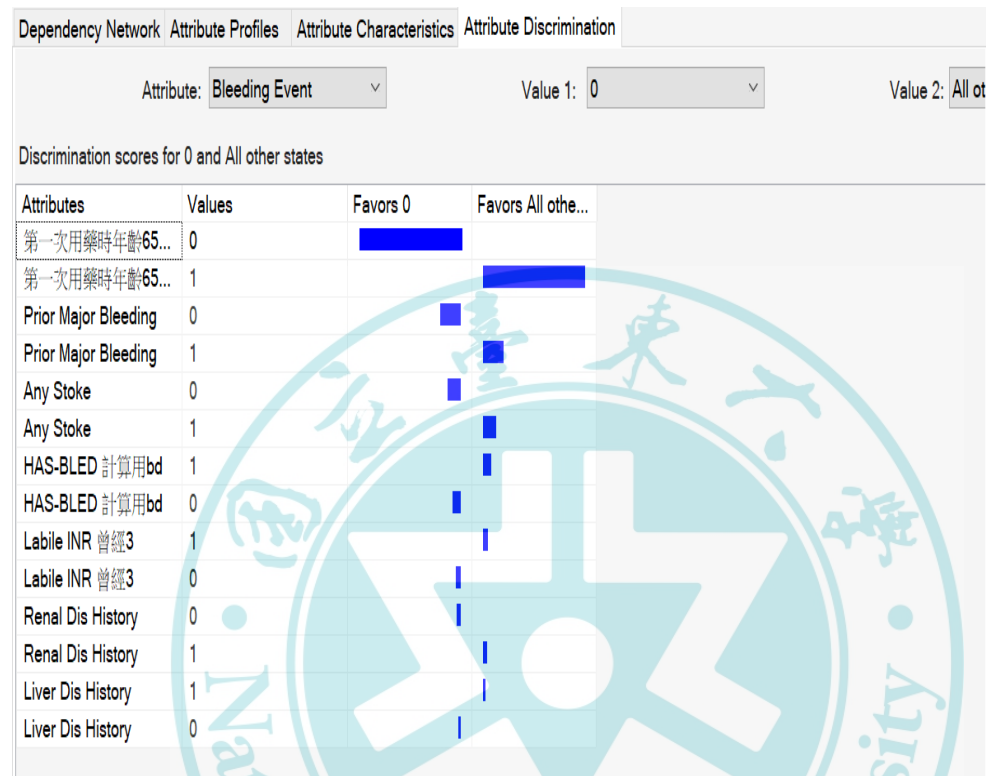


圖 4.16 貝氏機率分類預測模型之屬性辨識圖

資料來源：本研究整理

4.2.3.4 類神經網路預測模型

類神經網路為仿造人類最簡單的神經網路的結構即功能的運算模型，具有平行處理、容錯、結合性記憶、解決最佳化等特性。本研究將變數輸入模型並加以訓練，在完成模型建置訓練後，其預測結果如圖 4.17 所示，由圖中顯示採礦模型檢視器呈現何種情況透過機率的方式，對預測變數機率比重。

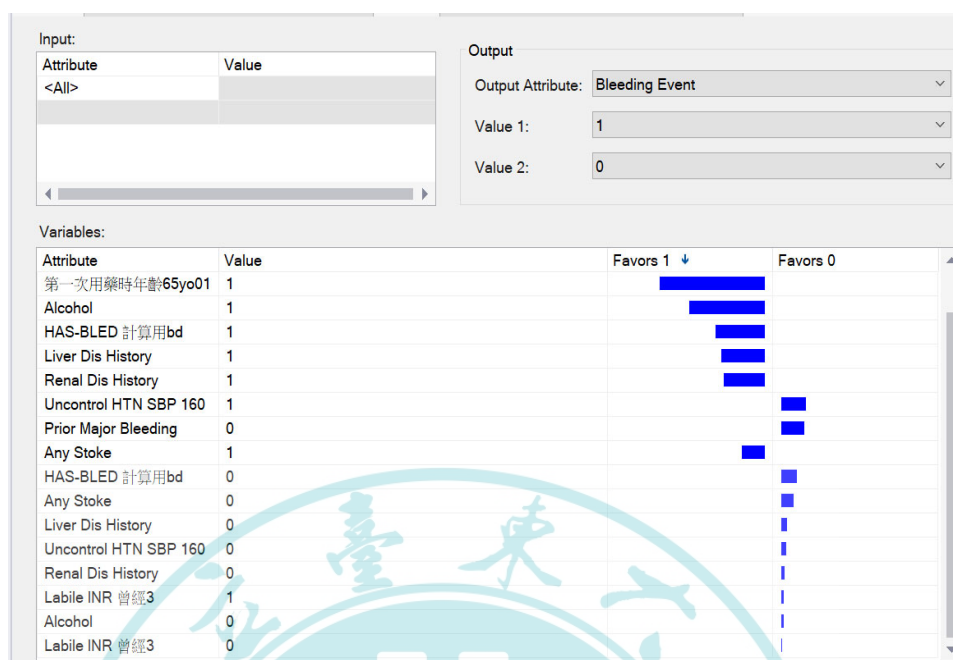


圖 6 類神經網路預測模型之屬性預測機率檢視

資料來源：本研究整理

4.3 評估比較分析模型

在資料探勘中，依分析資料及尋找樣式相關性來選擇合適演算法來建立資料探勘模型，所建立的預測模型是否有真實的價值及其意義，須對採礦結果進行評估。本研究依 Microsoft SQL Server2017 軟體工具所運用的分類演算法而採用對於預測具代表性的決策樹、關聯分析、貝氏分類、類神經網路來為研究案例資料來建立資料探勘預測模型。依不同演算法建置決策樹、關聯分析、貝氏分類、類神經網路等預測模型，以目標母體為有發生出血事件，匯入另行保留的 30%測試資料進行驗證分析，比較哪種模型具有最大評估效益，提供臨床醫學上參考。

4.3.1 資料採礦增益圖比較

常用的正確率評估為 holdout validation、cross-validation 與 bootstrapping(Letourneau et al.1999)。本研究採用以 holdout validation 的方式，將資料分為訓練集和測試集，訓練集用來訓練預測模型，測試集則用於測試預測結果，避免評估結果過於保守。本研究依 Microsoft SQL Server2017 軟體工具所運用的分類演算法而採用對於預測具代表性的決策樹、關聯分析、貝氏分類、類神經網路來為研究案例資料來建立資料探勘預測模型。本研究資料為無發生出血事件集與發生出血事件集筆數為 1:1 之資料集，共 872 筆，依不同演算法建置決策樹、關聯分析、貝氏分類、類神經網路預測模型，以目標母體為發生出血事件等於「Yes」，匯入另行保留的 30%測試資料得增益圖型，如圖 4.18。

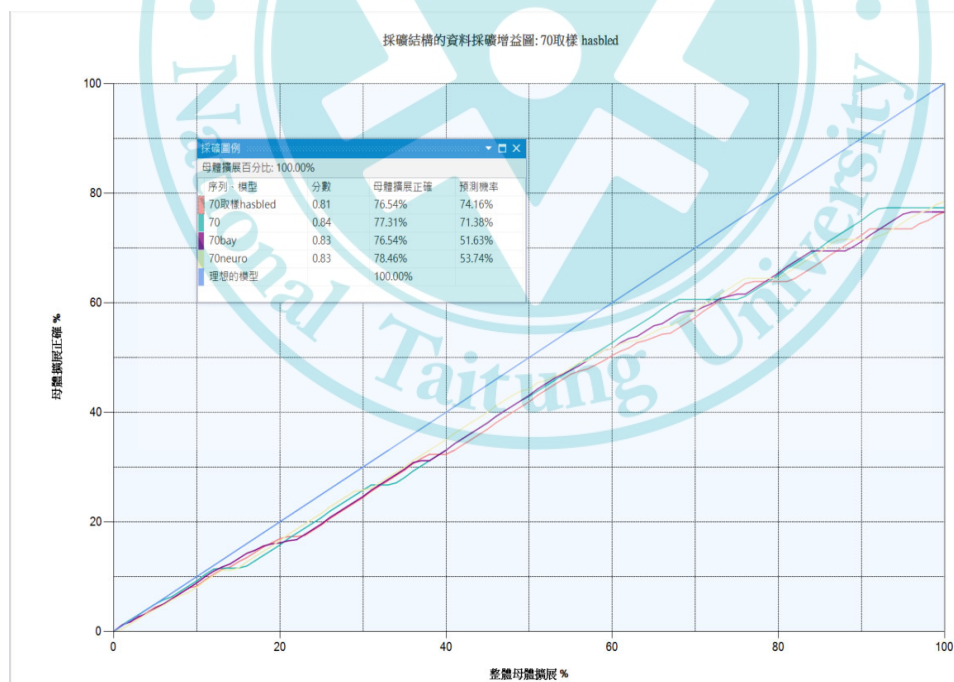
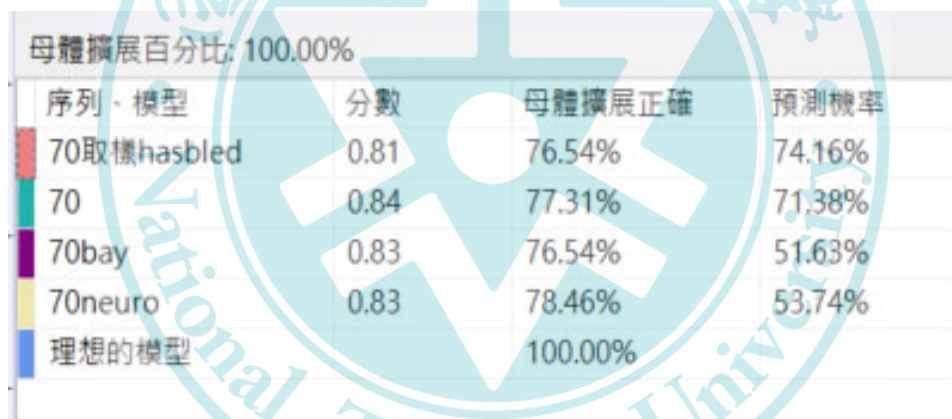


圖 7 四種資料探勘模型之增益圖與理想模型之比較

(資料來源：本研究整理)

在圖中縱軸為整體預測準確率的百分比，橫軸為整體母體擴展率的百分比。圖中一條自左下角延伸至右上角 45 度的斜線是參考標準線，代表在理想模型時，當橫軸為全體測試資料 50% 案例數時，與其對應的準確率應為 50%；若橫軸為全體測試資料 100% 案例數時，與其對應的準確率應為 100%。故預測模型所產生的線條越接近參考標準線代表該模型在預測準確率越高。

由圖 4-18 中顯示決策樹與類神經網路預測模型的預設準確率表現較佳。在此軟體中也同時提供增益圖之採礦圖例視窗，如圖 4-19。此圖例視窗中分數一欄為提供各模型之 AUC 值，當預測模型的 AUC 值越接近 1 代表該模型預測能力越好。由此可知以決策樹預測模型分數最高，為 0.84，表示在四種預測模型中以決策樹的表現最好。



母體擴展百分比: 100.00%			
序列、模型	分數	母體擴展正確	預測機率
70取樣hasbled	0.81	76.54%	74.16%
70	0.84	77.31%	71.38%
70bay	0.83	76.54%	51.63%
70neuro	0.83	78.46%	53.74%
理想的模型		100.00%	

圖 4.19 四種資料探勘模型之姿亦圖與理想模型之比較

(資料來源：本研究整理)

依各模型測試得分類矩陣表，如圖 4-20。圖中縱軸代表預測的計數，橫軸代表實際的計數。

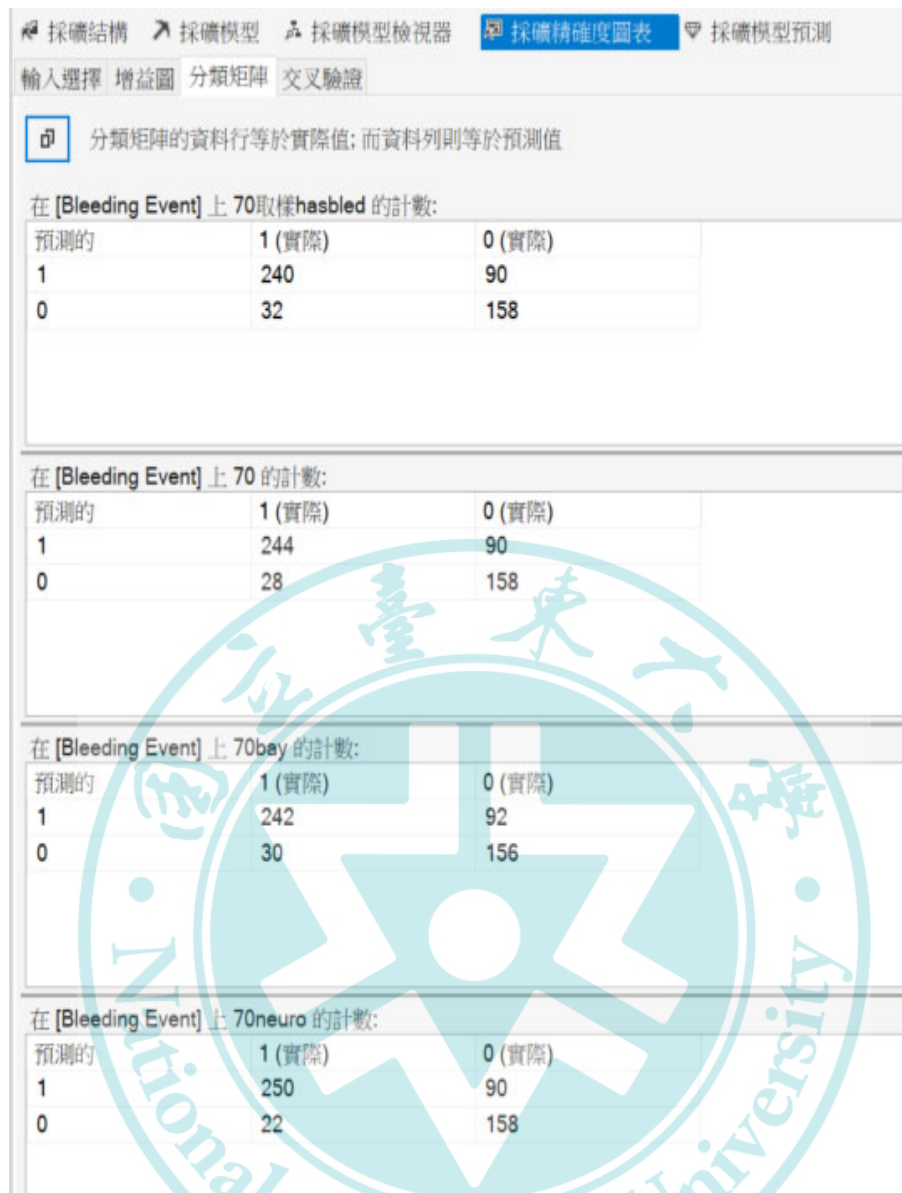


圖 4-20 四種預測模型之預測結果分類矩陣

(資料來源：本研究整理)

分類效能得評估使用分類矩陣來進行模型評估。在模型建構的整體分類正確率(overall accuracy)均大於機會正確率(Chance Accuracy)50%，如表 4.15 所示。整體分類正確率大於機會正確率，表示本研究中四種預測模型皆對發生或為未發生出血事件有良好的區別。

表 4.15 四種預測模型之整體分類正確率表

預測模型種類	整體分類正確率 (overall accuracy)	機會正確率 (Chance Accuracy)
關聯分析	76%	50%
決策樹	77%	50%
貝氏分類	76.5%	50%
類神經網路	78%	50%

(資料來源：本研究整理)

在不確定性降低方面，未發生出血事件模型預測誤差分為遺漏性誤差(omission error)，即模型預測為有發生出血事件，實際資料為未發生出血事件及過度預測誤差(commission error) 模型預測為未發生出血事件，實際資料為有發生出血事件兩種。本研究各模型之遺漏性誤差(omission error)與過度預測誤差(commission error)如表 4.16 所示。其中分別以類神經網路預測模型與決策樹預測模型有較低的遺漏性誤差值及過度預測誤差值。

表 4.16 四種預測模型之遺漏性誤差與過度預測誤差表

預測模型種類	遺漏性誤差 (omission error)	過度預測誤差 (commission error)
關聯分析	16.8%	27.2%
決策樹	15%	21.9%
貝氏分類	16.1%	27.5%
類神經網路	12%	26.4%

(資料來源：本研究整理)

4.3.2 資料採礦評估比較

經由預先保留的 30%測試資料分別匯入各模型之中進行驗證分析，AUC 分數以決策樹預測模型為 0.84 最高，顯示決策樹預測模型在提供評估及建立預測方面較優於貝氏分類、類神經網路、關聯分析等預測模型。

另比較模型預測能力優劣，如表 4.17，其中類神經網路預測模型找出實際發生出血事件數分別為 250，整體預測正確率為 78.2%，由實際測試後所得數據為決策樹模型整體預測正確率最高，為本研究最適模型，各模型影響目標變數依模型實驗結果得較顯著變數列如表 4.18。

表 4.17 四種模型的預測比較表

模型名稱	預測出血事件	實際未發生出血事件	實際有發生出血事件	AUC 分數	過度預測差 (commission error)	整體分類正確率 (overall accuracy)	優劣排序
關聯分析	未發生	158	32	0.81	27.20%	76.50%	2
	有發生	90	240				
決策樹	未發生	158	28	0.84	21.90%	77%	1
	有發生	90	244				
貝氏分類	未發生	156	30	0.83	27.50%	76.50%	4
	有發生	92	242				
類神經網路	未發生	158	22	0.83	26.40%	78.20%	3
	有發生	90	250				

(資料來源：本研究整理)

表 4.18 影響目標變數較顯著變數表

模型	變數
關聯分析	第一次用藥年齡是否小於 65 用藥前大出血或易出血病史 是否有腎病 抗凝血指數曾經大於 3 用藥前 Any Stroke(包括出血和梗塞) /TIA(短暫性腦缺血)/ /Thromboembolism(血栓栓塞) 病史
決策樹	第一次用藥年齡是否小於 65 Has-bleed 計算用(併用藥物) 用藥前大出血或易出血病史 用藥前 Any Stroke(包括出血和梗塞) /TIA(短暫性腦缺血)/ /Thromboembolism(血栓栓塞) 病史 抗凝血指數曾經大於 3
貝氏分類	第一次用藥年齡是否小於 65 用藥前大出血或易出血病史 用藥前 Any Stroke(包括出血和梗塞) /TIA(短暫性腦缺血)/ /Thromboembolism(血栓栓塞) 病史 Has-bleed 計算用(併用藥物) 抗凝血指數曾經大於 3 是否有腎病 曾經出現過肝臟疾病

類神經網路	用藥前大出血或易出血病史 第一次用藥年齡是否小於 65 飲酒習慣 Has-bleed 計算用(併用藥物) 曾經出現過肝臟疾病 是否有腎病
-------	---

(資料來源：本研究整理)

4.4.3 資料採礦結果與 HAS-BLED Score 比較

出血風險評分 (HAS-BLED) 提供了一個實用的、簡單的、方便的工具來評估臨床上的房顫患者的出血風險，從而能支持房顫患者抗血栓治療的臨床決策。HAS-BLED Score 所評估的出血風險是由各項因子所加總積分而來，如表 4.19 所示。因每個因子所貢獻的分數皆相同並無先後重要之分，再由總積分來判斷出血風險，由此可知凡總積分相同即出血風險便相同，即使計算積分的因子並不盡相同。

表 4.19 HAS-BLED Score 出血機率

Risk factor	Score	總分 (Risk Factors/Score)	出血機率 (Bleeds Per 100 Patient-Years)
高血壓 Hypertension	1	0	1.13
肝&腎功能異常 Abnormal renal and liver function (1 point each)	1	1	1.02
中風病史 (Stroke)	1	2	1.88
出血病史 (Bleeding history or predisposition)	1	3	3.74

INR 不穩定 (Labile INRs)	1	4	8.7
年齡大於 65 歲 (Elderly>65year)	1	5	12.5
使用特定藥物& 飲酒 Drugs or alcohol (1 point each)	1		

(資料來源：本研究整理)

在臨床上，各個房顫患者的計算出血風險因子並不盡相同。本研究決策樹預測出血模型顯示，整體分類正確率為 77%，AUC 分數為 0.84，顯示該模型區辨能力良好。從決策樹模型路徑規則可得知，影響預測出血事件的因子的重要性即其相依性。因此本研究希望能將出血風險評估個人化，提高準確度。亦提供評估臨床上的房顫患者的出血風險，並能在房顫患者抗血栓治療的臨床決策上提供另一個別化之預測工具。

第五章 結論與建議

本章共分為兩部分，第一部分將研究與實驗結果歸納整理，其本研究之成果以及未來後續研究之建議。

5.1 研究結論與貢獻

本研究資料來源為某地區醫院收集記錄期間自 2016/1/1 至 2018/4/30 使用 Rivaroxian 之患有 AF 病人相關資料為研究探討母體，從文獻中尋求可能導致用藥後發生之出血事件之因素，使用經歐洲心臟學會 (ESC guidelines) 推薦並目前臨床上普遍採用的 HAS-BLED score 的危險因子計分項來建構模型用以預測患者出血的風險。研究方法為使用資料採礦中決策樹演算法來進行，並事先保留 30%測試資料，其餘 70%資料作為訓練模型使用，進行分析及評估，得出結論如下：

一、自決策樹模型中取得可能為發生出血事件的主要變數為以下四項，說明

第一次用藥年齡是否小於 65、用藥前大出血或易出血病史、併用藥物、是否有腎病，為預測是否發生出血事件中要依據，整理如下：

1. 如果第一次用藥年齡大於 65 者，預測度為 50.8%。
2. 用藥前有大出血或易出血病史者，預測度為 90.91%。
3. 有併用藥物者，預測度為 87.50%。
4. 有腎病者，預測度為 90.0%。

二、自決策樹模型中取得可能為不發生出血事件的主要變數為以下四項，說

明第一次用藥年齡是否小於 65 歲、用藥前大出血或易出血病史、併用藥物、用藥前 Any Stroke(包括出血和梗塞) /TIA(短暫性腦缺血)，為預測是否發生出血事件中要依據，整理如下：

1. 第一次用藥年齡小於 65 者，預測度為 49.16%。

2. 無併用藥物者，預測度為 84.75%。
3. 用藥前無大出血或易出血病史者，預測度為 86.90%。
4. 用藥前無 Any Stroke(包括出血和梗塞) /TIA(短暫性腦缺血)者，預測度為 88.74%。

三、本論文主要貢獻為利用資料採礦(Data Mining)建立決策樹預測模型，將目前臨床較常使用評估 AF 病患發生出血機率之 HAS-BLED score 的危險因子計分項來建構模型用以預測患者出血的風險，建立預測模型。由決策樹模型得知不同致病因子對預測出血事件各有不同之預測正確機率，由不同的致病因子對預測出血事件預測正確機率之不同，可將易致病因子之計分項區分出重要性，將可提供在 HAS-BLED score 上總計分相同但致病因子計分項不同之病患較準確及個人化之出血機率評估，提供臨床醫師在開立抗凝血劑 rivaroxaban 時之決策參考，提高用藥之安全性。

5.2 建議

- 一、在罹患 AF 的病患上使用抗凝血藥物為預防性治療，若能將用藥前的風險評估準確度提高，可降低使用抗凝血藥物的出血風險，對於病患的用藥安全將提供助益。
- 二、本研究系以單一地區醫院之 AF 病患資料為母體樣本，發生出血事件之案例數偏少。要建立模型預測發生出血事件，希望能得到最多之案例數來進行分析及模型訓練，提高預測之準確率。建議建立相關資料庫用以後續者繼續探究。

三、紀錄資料欄位因依目前各種各式目的評估計分項所設計，建議可增加其他欄位，例如生活作息習慣、更詳細之生理數據等，供資料採礦使用資料採礦分析時能發掘新的、隱藏的、未知的相關致病因子，建立更精準的預測模型，提供臨床醫師在用藥前之參考。



參考文獻

中文

1. 世界衛生組織 · <http://www.who.int> 。
2. 葉怡成(1999) · 類神經網路模式與應用實作 · 台北市:儒林書局。
3. 王進德、蕭大全(1994) · 類神經網路與模糊控制理論控制入門 · 台北市:全華科技圖書股份有限公司 。
4. 毛雲川 (2014) · 抗血栓藥物 rivaroxaban 使用於接受人工膝關節及髖關節置換的病人深部靜脈栓塞預防的研究 · 碩士論文 · 高雄醫學大學。
5. 王郁文(2011) · 臺灣機械性心臟瓣膜置換病患之抗血栓劑使用評估 · 碩士論文 · 國立成功大學。
6. 王敏如 (2008) · 心房顫動病患使用抗血栓劑符合準則與否和心血管事件發生關聯性之探討 · 碩士論文 · 國立成功大學。
7. 余秀芹(2010) · 心房纖維顫動病人的缺血性腦中風發生率及預防性抗血栓治療之探討 · 碩士論文 · 長庚大學。
8. 周秉箴 (2013) · 以血栓溶解劑治療缺血性中風成本效益分析 · 碩士論文 · 國立交通大學。
9. 洪明潔 (2011) · 腦心血管疾病之抗血栓凝集用藥選擇系統離型 · 碩士論文 · 國立陽明大學。
10. 洪智中 (2015) · 心房顫動合併冠狀動脈支架置放術後病人使用口服三合一抗血栓藥物治療效果及安全性之相關性研究 · 碩士論文 · 高雄醫學大學。
11. 張簡千郁、許秀卿 (2018) · 心房顫動最新藥物治療: 新型口服抗血栓藥物 · 家庭醫學與基層醫療 , 33(1) , 23-28.

12. 衛生福利部 · <https://www.mohw.gov.tw>。
12. 張家嫻 (2013) · 臺灣心房顫動病人預防中風之抗血栓劑療效與安全性評估 · 國立成功大學臨床藥學與藥物科技研究所博士論文 · 台南市。
14. 趙雅琴(2013) · 新型口服抗凝血劑— 心房顫動中風預防的新曙光。內科學誌， 33(7)，21。
15. 黃淑萍 (2005) · 某醫學中心心房纖維顫動病患特性與其抗血栓劑處方型態之研究 · 碩士論文 · 國立臺灣大學。
16. 楊書瑜 (2012) · 評估心房顫動患者缺血性中風後使用抗血栓劑之持續比例、療效及安全性 · 碩士論文 · 臺北醫學大學。
17. 張朝欽、雷孟桓(2012) · 新型口服抗凝血劑 - 心房顫動中風預防的新曙光 · 內科學誌，23，77-97。
18. 劉家昌、林益卿、黃馨蓀 (2017) · 新型口服抗凝血藥物在預防心房纖維顫動患者中風預防的使用 · 家庭醫學與基層醫療，32(11)，314-321。
19. 鄭名惠 (2007) · 心房顫動病患使用抗血栓劑之處方型態分析 · 碩士論文 · 國立成功大學。
20. 賴怡珊(2014) · 抗血栓劑於合併心房顫動與急性冠心症的病患之療效與安全性評估 · 碩士論文 · 國立成功大學。
21. 謝明儒(2012) · 急性腦中風急診病患縮短血栓溶解治療時間介入方案的評估研究 · 碩士論文 · 國立臺灣大學。
22. 內政部統計處 · <https://www.moi.gov.tw/stat/index.aspx>。
23. 中華民國心律醫學會心房顫動衛教資訊網 · www.afhealthcare.org.tw。
24. 衛生福利部 · 105 年身心障礙者生活狀況及需求調查報告 · <https://www.mohw.gov.tw>。

25. 謝邦昌(2014)·*SQL Server 第二版 資料採礦及商業智慧*·臺北市:碁鋒資訊。
26. 簡禎富(2014)·*資料挖礦與大數據分析*·臺北市:前程。
27. 陳會安(2016)·*SQL Server 2016/2017 資料庫設計及開發實務*·臺北市:碁鋒資訊。
28. 簡禎富、許嘉裕(2014)·*資料挖礦與大數據分析*·台北:前程文化。

英文

1. Agrawal, R., Imieli, T., ski, & Swami, A. (1993). *Mining association rules between sets of items in large databases* .SIGMOD Rec. 22(2), 207-216.
2. Breiman, L. (2017). *Classification and regression trees*: Routledge.
3. Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., Zanasi, A., ., I. B. M. C., & ., I. T. S. O. (1997). *Discovering data mining: from concept to implementation*: Prentice Hall PTR New Jersey.
4. Camm, A. J., Kirchhof, P., Lip, G. Y., Schotten, U., Savelieva, I., Ernst, S., & Van Gelder, I. C. J. E. h. j. (2010). *Guidelines for the management of atrial fibrillation: the Task Force for the Management of Atrial Fibrillation of the European Society of Cardiology (ESC)*. 31(19), 2369-2429.
5. Chugh, S. S., Havmoeller, R., Narayanan, K., Singh, D., Rienstra, M., Benjamin, E. J., . . . Zheng, Z.-J. J. C. (2014). *Worldwide epidemiology of atrial fibrillation: a Global Burden of Disease 2010 Study*. 129(8), 837-847.
6. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. J. C. o. t. A. (1996). *The KDD process for extracting useful knowledge from volumes of data*. 39(11), 27-34.
7. Fayyad, U., & Stolorz, P. (1997). *Data mining and KDD: Promise and*

- challenges. *Future Generation Computer Systems*, 13(2), 99-115.
8. Feinberg, W., Blackshear, J., Laupacis, A., Kronmal, R., & Hart, R. J. A. I. M. (1995). *The prevalence of atrial fibrillation: analysis and implications*. 155, 469-473.
 9. Fontana, P., Robert-Ebadi, H., Bounameaux, H., Boehlen, F., & Righini, M. P. J. S. m. w. (2016). *Direct oral anticoagulants: a guide for daily practice*. 146, w14286.
 10. Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*: Elsevier.
 11. Heeringa, J., van der Kuip, D. A., Hofman, A., Kors, J. A., van Herpen, G., Stricker, B. H. C., . . . Witteman, J. C. J. E. h. j. (2006). *Prevalence, incidence and lifetime risk of atrial fibrillation: the Rotterdam study*. 27(8), 949-953.
 12. Hui, S. C., & Jha, G. (2000). *Data mining for customer service support*. *Information & Management*, 38(1), 1-13.
doi:[https://doi.org/10.1016/S0378-7206\(00\)00051-3](https://doi.org/10.1016/S0378-7206(00)00051-3)
 13. Kleissner, C. (1998). Data mining for the enterprise. Paper presented at the Proceedings of the Thirty-First Hawaii International Conference on System Sciences.
 14. Lin, H.-J., Wolf, P. A., Kelly-Hayes, M., Beiser, A. S., Kase, C. S., Benjamin, E. J., & D'agostino, R. B. J. S. (1996). *Stroke severity in atrial fibrillation: the Framingham Study*. 27(10), 1760-1764.
 15. Lloyd-Jones, D. M., Wang, T. J., Leip, E. P., Larson, M. G., Levy, D., Vasan, R. S., . . . Benjamin, E. J. (2004). *Lifetime Risk for Development of Atrial Fibrillation*. 110(9), 1042-1046.
 16. Olesen, J. B., Lip, G. Y. H., Hansen, M. L., Hansen, P. R., Tolstrup, J. S.,

- Lindhardsen, J., . . . Torp-Pedersen, C. (2011). *Validation of risk stratification schemes for predicting stroke and thromboembolism in patients with atrial fibrillation: nationwide cohort study*. 342, d124.
17. Pisters, R., Lane, D. A., Nieuwlaat, R., De Vos, C. B., Crijns, H. J., & Lip, G. Y. J. C. (2010). *A novel user-friendly score (HAS-BLED) to assess 1-year risk of major bleeding in patients with atrial fibrillation: the Euro Heart Survey*. 138(5), 1093-1100.
 18. Steger, C., Pratter, A., Martinek-Bregel, M., Avanzini, M., Valentin, A., Slany, J., & Stöllberger, C. J. E. h. j. (2004). *Stroke patients with atrial fibrillation have a worse prognosis than patients without: data from the Austrian Stroke registry*. 25(19), 1734-1740.
 19. Tamayo, S. G., Simeone, J. C., Nordstrom, B. L., Patel, M. R., Yuan, Z., Sicignano, N. M., & Peacock, W. F. J. J. o. t. A. C. o. C. (2016). *Risk factors for major bleeding in rivaroxaban users with atrial fibrillation*. 68(10), 1144-1146.
 20. Wolf, P. A., Abbott, R. D., & Kannel, W. B. J. A. o. i. m. (1987). *Atrial fibrillation: a major contributor to stroke in the elderly: the Framingham Study*. 147(9), 1561-1564.
 21. Pisters, R., Lane, D. A., Nieuwlaat, R., De Vos, C. B., Crijns, H. J., & Lip, G. Y. J. C. (2010). *A novel user-friendly score (HAS-BLED) to assess 1-year risk of major bleeding in patients with atrial fibrillation: the Euro Heart Survey*. 138(5), 1093-1100.