中臺科技大學

資訊管理系

畢業專題

比較分類技術於骨髓移植之研究

學生：陳建維　　周佳德

張國強　　張曉寧

王貽郁　　吳宜嬪

范慈凌

口試老師：

指導老師：李桂春 助理教授、翁政雄 教授

系主任　：羅亦斯

中華民國 111 年 12 月 07 日(口試日期)

感謝

首先，感謝我們的專題指導老師翁政雄教授，在大學期間教導我們論文，在他的細心指導下讓我們學習到貝氏分類、決策樹、類神經網路、等機器學習演算法有基礎的概念。也學會了本次專題首要的軟體-Weka，讓我們藉由此軟體完成了大部份專題研究及實驗，以及自行使用 excel 計算各演算法，藉此加深對各個演算法的概念。

在研究專題的過程中，我們也遇到了許多問題需要去挑戰，中途曾想過放棄，最終也堅持了下來。尤其是在期間遇到疫情的影響。導致我們需要遠距線上上課及討論，無法面對面表達意見及想法，進度也遲遲落後，更導致整個專題的不順利。

也感謝期間翁老師給予我們一些鼓勵的話，在翁老師的指導下，我們的研究方向、方法都越來越明確，也讓我們越來越熟悉這個領域的知識，提升了不少在撰寫專題論文中的能力，若是碰到數學公式不熟悉以及研究方法不知如何解決時，也都會給予我們細心的講解，在專題論文及報告上也用心地幫我們檢查哪裡有錯誤，且會依我們個人的程度幫我們進行各別訓練，還會幫助我們進行口試練習，讓我們能夠在台上演講時可以更有信心。也感謝給予我們幫助的學長姐及我們辛苦的組員。沒有各位的協助，就沒有這篇論文的出現，我們也學不到這些專業的知識。謝謝你們的聆聽指教及建議，讓我們順利完成本篇專題論文。

摘要

根據統計，接受骨髓移植的病人的確比一般的健康者有著較高的罹癌機率。至於骨髓移植病人的罹癌機率為什麼會增加，就得先區分骨髓移植所增加的罹癌機率究竟是原發性或是次發性的惡性腫瘤。首先，如果病人本身在骨髓移植之前就已經罹患惡性腫瘤，而在骨髓移植之後，原來的惡性腫瘤仍無法治癒或是治癒後復發，這是第一種情況。在這種情況下，因為所罹患的腫瘤仍是原來的惡性腫瘤，因此應該視之為骨髓移植本身治療無效。(出處：癌症防治網-游介宇醫師)然而隨著資訊科技的進步，資料探勘技術的應用在近年來發展完善，而分類技術是重要的資料分析技術，用以萃取模型中的資料類別，若運用在預測骨髓移植的研究上，便能使病患盡早發現盡早治療，也能作為判斷是否需要骨髓移植。本研究運用四種分類演算法(C45 決策樹、貝氏分類法、類神經網路)於 bone-marrow(資料集中，建立不的模型預測模型，嘗試比較何種分類演算法在資料集中有較佳的預測能力。本研究結果顯示:在 Recall(或稱為 TPR)ROC AreaPRCArea 式分類表現最佳本研究運用預測模型整理出決策樹規則。期望能助醫療判斷骨髓移植發作診斷之參考。

開鍵詞:資料探勘、C4.5 決策樹、貝氏分類法、類神經網路、骨髓移植

目錄

表目錄

圖目錄

壹、緒論

有些遺傳或先天染色體異常，如先天性再生不良性貧血、或唐氏症患者，有較高機會罹患血癌。 受到輻射線的照射，例如核爆之後，或長久居住在輻射屋的影響，會增加血癌機會。暴露有毒化學物質、有機溶劑、殺蟲劑等也有影響。 此外，如果過去接受過化學治療治療其他疾病（例如：乳癌、骨肉瘤等）也可能增加血癌的風險。(出處：一本讀通血癌-唐季祿)都與骨髓移植有相當大的關聯，總而言之，生活中有天性再生不良性貧血、或唐氏症患者，需要骨髓移植的機率越大。雖然醫學科技日益的進步，但還是有很多無法克服的難題，我們希望利用資料探勘技術即早預測到需要骨髓移植的患者，實現早期治療的理念，若是能夠達成提高生存的機率之外，也可讓患者的家屬有一份安心。

根據統計，接受骨髓移植的病人的確比一般的健康者有著較高的罹癌機率。至於骨髓移植病人的罹癌機率為什麼會增加，就得先區分骨髓移植所增加的罹癌機率究竟是原發性或是次發性的惡性腫瘤。首先，如果病人本身在骨髓移植之前就已經罹患惡性腫瘤，而在骨髓移植之後，原來的惡性腫瘤仍無法治癒或是治癒後復發，這是第一種情況。在這種情況下，因為所罹患的腫瘤仍是原來的惡性腫瘤，因此應該視之為骨髓移植本身治療無效。(出處：癌症防治網-游介宇醫師)然而隨著資訊科技的進步，資料探勘技術的應用在近年來發展完善，資料探勘乃是由存在的資料中挖掘出新的事實，而 Berry＆Linoff(2000)認為「資料探勘是為要發現出有意義的樣型或規則，而必須從大量資料之中以自動或是半自動的方式來探索和分析資料」。分類技術乃是資料探勘技術中常用方法，分類(classification)乃是根據已知的資料及其類別屬性來建立資料的分類模型。分類模型的建立可以讓我們了解屬於各種類別屬性的資料具備哪些特徵，同時也可以用來預測新進資料的類別屬性。

由於 C4.5 決策樹、貝氏分類法、類神經網路在相關的應用上都有一定的準確度。因此，本研究將分別運用 C45 決策樹、貝氏分類法、類神經網路建立骨隨移植預測模型，並嘗試比較 C4.5 決策樹、貝氏分類法、類神經網路等預測模型之準確度。本研究其他章節規劃如下:第二章為文獻探討。第三章為研究方法。第四章為實驗部份，將使用 Kaggle (加州大學機器學習與智慧系統研究中心)的 Bone marrow transplant(骨髓移植數據集)資料集，比較 C4.5 決策樹、貝氏分類法、類神經網路預測模型的準確度。

貳、文獻探討

資料探勘又稱資料採礦，指從大量的資料信息中，找出脈絡關係，形成知識管理中的偽資訊，是一種已存在的資料中探討出未知的知識，為整個知識發現過程中的一個必要階段，要發現出有意義的樣本型態或必須從大量資料規則之中以自動或是半自動的方式。

以探索和分析資料有關資料探勘的其他定義有:「資料探勘是一個確定資料中有效的、新的、可能有用的，並且最終能被理解的模式的重要過程」「資料探勘是從大量資料中萃取出來的知識(Berry＆Linoff，2000;Fayyad＆Smyth，1996;Han＆Kamber，2006)，知識發現流程的主要程序(如圖 1 所示)，主程序描述如下：

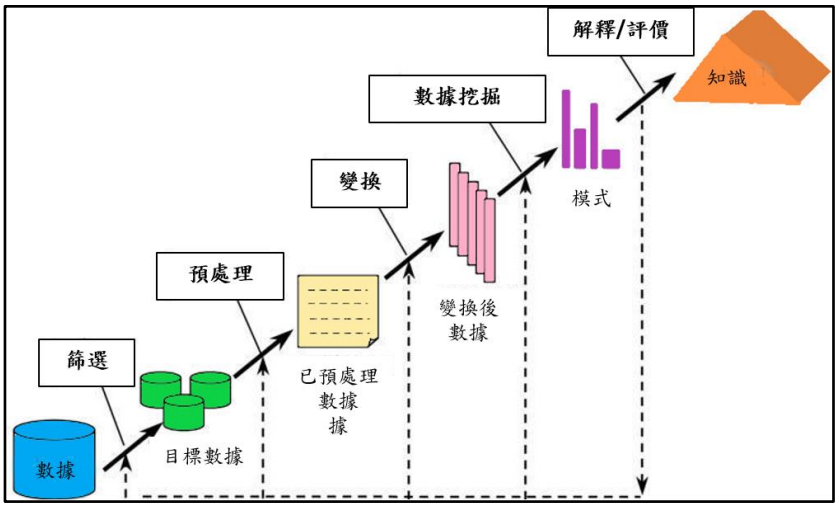

圖 1：知識發現流程圖(The KDD Process)

1. 資料選擇:依據探勘的需求找尋資料項目，建立資料集，避免探勘過程中耗費不必要的成本。
2. 資料前置處理:當資料的目標屬性或類別選取完後，接著就對資料進行後續的處理，包含資料中的雜訊、缺漏資料、資料整合等等。
3. 知識發現、演算轉換:將資料處理後，把資料根據探勘的設計，將資料轉換成統一的格式，資料才能夠被探勘工具所使用。
4. 資料探勘:將資料轉換之後，根據知識挖掘的目的進行資料的探勘，並將資料中隱藏的涵義轉換成實體的特徵或規則。
5. 解釋及評估:根據資料探勘後所得到的特徵或模型，來說明或解釋探勘的結果，同時評估結果是否符合當初的設計以及使否合適。

一、決策樹演算法及應用

決策樹是通過圖示羅列解題的有關步驟以及各步發生的條件與結果的一種方法，過程直覺單純、執行效率也相當高的監督式機器學習模型，不僅可以幫助人們理解問題，還可以幫助人們解決問題。決策樹一般是自上而下的來生成的。每個決策或事件（即自然狀態）都可能引出兩個或多個事件，導致不同的結果，把這種決策分支畫成圖形很像一棵樹的枝幹，提供指令讓我們實際的模擬並繪出從根部、各枝葉到最終節點的決策過程，故稱決策樹。此外，決策樹的特點是每個決策階段都相當的明確清楚（不是 YES 就是 NO），適用於 classification 及 regression 資料類型的預測。

Quinlan(1979)所開發的決策樹演算法 ID3(此為 C4.5 的前身)中，是使用資訊獲利(Information Gain)作為分岔準則，但是發現應用在實際案例上，資訊獲利會偏好選擇選項數較多的變數作為分岔變數。後來 Quinlan(1979)改良 ID3 變成 C4.5 決策樹方法，使其對於連續屬性資料之處理較有效率。其因決策樹的概念簡單，很容易就可以應用於資料分析的問題。

關於決策樹演算法之應用之文獻整理(如表 1 所示)，詳細的敘述如下：

林宇恆(民 105)認為利用該預測模型預測領前 1 至 7 天的氣溫與雨量，並針對預測值給予一定信賴水準的信賴區間。為了說明此預測模型的效能，該預測模型與其他時間序列的預測方法進行比較，其中時間序列的預測方法包括自迴歸、移動平均法、自迴歸差分整合移動平均法

孫苙達(民 107)認為以機器學習決策樹演算法探討總體經濟變數、籌碼面資料、系統風險相關變數以及航運類股產業相關變數對於台灣航運股股價漲跌的影響，研究期間為 2002 年至 2017 年資料，包含其七航運股與 23 種研究變數，並檢驗機器學習決策樹的預測準確度與作為交易策略的績效。

廖介銘(民 92)認為許多醫學資料與日俱增地累積下來。面對這大量未經整理的資料，我們可經由資料探勘發掘可能有意義的資訊，再經由醫師作更進一步之研究。此研究以決策樹演算法，找出糖尿病患者的決定影響因子。研究樣本為 2001 年 11 月至 2002 年 9 月之間中部某醫院就診紀錄。先以統計方式篩選有顯著差異之變數，再將其戴入至決策樹演算法歸納出糖尿病決定因子。

林育成(民 103)利用進行白血球切割，然後從切割出的區域取得特徵值之後，使用人工免疫選擇演算法做屬性篩選與參數選擇，藉由演算法中複製、變異、選擇以及保留最佳抗體等概念，得到最佳化的參數與選擇之特徵值輸入決策樹進行血球分類。

詹育佳(民 107)利用目前現有的顧客使用精油產品用於芳香療法之舒壓上，用決策樹加以歸納分析出會員分類模型，找出芳療顧客消費及使用精油的決定影響因子，再提出將其分類好的顧客特性及對應的適用之精油產品，並用建構客戶相關資訊的特徵來架構其前端作業的 CRM 系統，且能協助芳療師快速的找出對應顧客需求之精油產品，因此也能規劃最佳的行銷策略以提升其競爭力。

表 1：決策樹應用文獻表

作者(年)	應用領域	目的
林宇恆 (民 105)	科學	認為利用該預測模型預測領前 1 至 7 天的氣溫與雨量，並針對預測值給予一定信賴水準的信賴區間。
孫苙達 (民 107)	商業	以機器學習決策樹演算法探討總體經濟變數、籌碼面資料、系統風險相關變數以及航運類股產業相關變數對於台灣航運股股價漲跌的影響，並檢驗機器學習決策樹的預測準確度與作為交易策略的績效。
廖介銘 (民 92)	醫學	先以統計方式篩選有顯著差異之變數，再將其戴入至決策樹演算法歸納出糖尿病決定因子。
林育成 (民 103)	醫學	藉由演算法中複製、變異、選擇以及保留最佳抗體等概念，得到最佳化的參數與選擇之特徵值輸入決策樹進行血球分類。
詹育佳 (民 107)	商業	利用目前現有的顧客使用精油產品用於芳香療法之舒壓上，用決策樹加以歸納分析出會員分類模型，找出芳療顧客消費及使用精油的決定影響因子。

二、貝氏分類法及其應用

貝式分類法（Bayes classifier）乃是根據貝氏定理（Bayes' theorem）為基礎，是一種統計分類，用已知的事件發生之機率來推測未知資料的類別，此為貝氏分類最大的特色。當新的樣本資料加入時，只要再調整某些機率，即可以得到新的分類的模型，因此當資料不斷增加的時候，會有比較好的分類效能。用以判斷未知類別的資料應該最接近哪一個類別。整個貝式分類法的目標是希望能透過機率統計的分析，達到最小誤差的一種分類方式。

關於貝氏分類法之應用之文獻整理(如表 2 所示)，詳細的敘述如下：

翁毓謙(民 94)利用貝氏分類決策理論出發，同時考慮語音模型的不確定性與語音模型間的鑑別性，設計出以最小化貝氏期望風險為主之語音辨識演算法。在最小化貝氏風險分類決策法則中，損失函數應考慮類別鑑別性而設計，以期改進辨識效能。

曾麗文(民 100)希望能透過對臨床醫學診斷最好的貝氏分類方法。應用於不同醫療數據集之結果驗證與評估，了解何種醫療數據集在何種貝氏分類方法之預測結果較佳，如此將對醫療之診斷治療及預後預測提供相當的助益。

黃代鈞(民 93)利用在數量廣大的人類基因中，分析基因表現的資料找出遺傳疾病相關的關鍵基因，日後得以只用這些基因的表現量來識別一個病患是否患有該特定疾病。

廖育民(民 106)利用導入貝氏分類，將失智症的危險因子，如性別、年齡、教育程度、甲狀腺機能低下、高血壓、心臟病、糖尿病、腦中風等因子，並考慮醫界常用來評估失智症的簡易智能量表（Mini-mental state examination， MMSE），以此建構出可行的失智症臨床初期診斷模式。

鄭鈺傑(民 104)利用智慧型手機以車輛對稱性與貝氏分類為基礎的前車辨識演算法，以車輛對稱性辨識前方目標物是否為車輛，並使用貝氏分類做機率追蹤，並將此辨識系統移植於 Android 智慧型手機上。

表 2：貝氏分類應用文獻表

作者(年)	應用領域	目的
翁毓謙 (民 94)	科學	利用貝氏分類決策理論，設計出以最小化貝氏期望風險為主之語音辨識演算法，在最小化貝氏風險分類決策法則中，損失函數應考慮類別鑑別性而設計，以期改進辨識效能。
曾麗文 (民 100)	醫學	藉由使用數個醫學數據資料，利用多項貝氏分類法之分類方法進行實驗比較，針對預測結果，進行多元分類結果、ROC 曲線、正確率、精確度、召回率等分析。
黃代鈞 (民 93)	醫學	利用在數量廣大的人類基因中，分析基因表現的資料找出遺傳疾病相關的關鍵基因。
廖育民 (民 106)	醫學	利用導入貝氏分類，將失智症的危險因子，以此建構出可行的失智症臨床初期診斷模式。
鄭鈺傑 (民 104)	科學	根據內政部警政署 101 年及 102 年交通事故肇事統計，除了機械故障等其他外在因素之外，大部分肇事原因是駕駛者過失所造成，因此若能提供一個準確的前車偵測警示系統，預先判定可能的危險，提醒駕駛者提早做出因應措施，可有效預防車前事故的發生，減少對駕駛者危險。

三、類神經網路及其應用

人工神經網路（ANN），簡稱神經網路（NN）或類神經網路，在機器學習和認知科學領域，是一種模仿生物神經網路的結構和功能的數學模型或計算模型，用於對函式進行估計或近似。神經網路由大量的人工神經元聯結進行計算。大多數情況下人工神經網路能在外界資訊的基礎上改變內部結構，是一種自適應系統，通俗地講就是具備學習功能。

現代神經網路是一種非線性統計性資料建模工具，神經網路通常是通過一個基於數學統計學類型的學習方法（Learning Method）得以最佳化，所以也是數學統計學方法的一種實際應用，通過統計學的標準數學方法我們能夠得到大量的可以用函式來表達的局部結構空間，另一方面在人工智慧學的人工感知領域，我們通過數學統計學的應用可以來做人工感知方面的決定問題（也就是說通過統計學的方法，人工神經網路能夠類似人一樣具有簡單的決定能力和簡單的判斷能力），這種方法比起正式的邏輯學推理演算更具有優勢。

有關類神經網路之應用之文獻整理(如表 3 所示)，詳細的敘述如下：

林逸塵(民 91)調查針對高雄都會區之懸浮微粒濃度及能見度，以類神經網路加以預測，本研究之目的有二：(1)針對氣象因子及空氣污染物濃度對懸浮微粒濃度及能見度之影響加以探討與分析。(2)利用類神經網路預測懸浮微粒濃度與大氣能見度，並討論類神經網路之優缺點及可能之改善方法。

陳采蓉(民 109)使用研究方法選擇兩種具有預測能力的工具，分別是類神經網路 (Artificial neural network，ANN)及邏輯斯迴歸分析 (Logistic regression analysis)，利用兩者方法來預測膀胱鏡檢查導致泌尿道感染機率，並比較兩者方法之預測能力。 本研究首先會進行相關的文獻回顧，再敘述研究方法，最後進行資料分析， 並由結果提出結論與建議。

曾世任(民 100)利用類神經網路透過低複雜性的數學運算過程，省去與資料庫中每一筆特徵作比對的時間，所以速度比初期的辨識方法快。因此，此論文也將使用類神經網路作為辨識工具加上不同的量化方法，藉此能提高辨識率。

辜炳寰(民 91)採用了類神經網路作為分析工具來發展簡易經驗評估法來評估土壤液化問題。在論文中，首先提出一套合乎邏輯的程序來篩選土壤鑽探資料，而用來發展類神經網路系統整體結構和分析流程將於文中詳細探討。

鍾炳煌(民 91)利用汽車駕駛模擬系統為收集資料工具，構建擬真之道路與交通環境，進行駕駛實驗。並藉由類神經網路處理非線性問題的能力，彙整進口匝道區域車流特性及車輛併入行為，構建智慧型併入決策支援模式。

表 3：類神經網路應用文獻表

作者(年)	應用領域	目的
林逸塵 (民 91)	科學	調查針對高雄都會區之懸浮微粒濃度及能見度，以類神經網路加以預測，並討論類神經網路之優缺點及可能之改善方法。
陳采蓉 (民 109)	醫學	希望藉由機器學習(Machine learning)來做資料探勘，過去醫學疾病判斷多只靠醫師以往的經驗，但是現今疾病判斷指標相當多元，如果能配合資料探勘的技術則可以輔助醫生以提高診斷正確率。許多研究也證實應用機器學習技術，對於疾病診斷是有相當大的幫助。
曾世任 (民 100)	科技	研究目的有二個，一是使用非公認的人臉資料庫時，探討是否能得到近似的辨識結果，同時也觀察當背景不同時，是否會影響 SVM 的辨識效果。二是透過個人手寫的習性作為生物特徵以達手寫作者之辨識，並探討其可行性。
辜炳寰 (民 91)	科學	利用台灣最常見的 SPT 鑽探資料，配合類神經網路理論，選擇適當的參數作為輸入值，提出一套理論化的簡易經驗評估法，來研究液化問題，成為一套完整的液化防制的參考方法。
鍾炳煌 (民 91)	工程	此研究所探討範圍著重於高速公路進口匝道駕駛行為之分析。進口匝道係連接兩不同等級之道路系統，於交會處必須設置加速車道以使車輛得以順利匯流進入高速公路主線車道。

參、研究方法

一、決策樹演算法

Shannon and Weaver(1949)提出資訊理論(information theory)，的概念如下：

假設有一資料庫 D 共有|D|筆資料，這|D|筆資料可以分類為 m 種結果，這 m 種結果發生的機率分別為 p1，…，pm，這些機率都是已知，則這個事件發生後所得到的資訊量為：

$$Info(D) = \sum_{i=1}^{m} - p_i \log_2 p_i \quad (1)$$

上述公式也代表以二進位方式來表達該項事件所獲得的資訊時所需的平均位元數。由上述的公式也可以看出：各種結果發生的機率越平均，所求得的資訊量越大。因此，資訊量也可以當作亂度(Entropy)的指標，即資訊量越大亂度越大。

決策樹產生的基本演算法是利用貪婪演算法(greedy algorithm)，它是一種由上而下(top down)的方法，用遞迴(recursive)和各個擊破(divide and conquer)來建立樹狀結構。Quinlan(1979)所提出的 ID3 決策樹演算法，採用 Shannon and Weaver(1949)提出資訊理論作為選擇測試屬性的依據。因此，ID3 決策樹演算法是以選取測試後資訊量最小的屬性為主，也就是選擇資訊獲利最大的屬性。資訊獲利(Information Gains)定義如下：

$$Gain(A) = Info(D) - Info_A(D) \quad (2)$$

早期 Quinlan(1979)所開發的決策樹演算法 ID3(此為 C4.5 的前身)中，是使用資訊獲利(Information Gain)作為分岔準則，但是發現應用在實際案例上，資訊獲利會偏好選擇選項數較多的變數作為分岔變數。因此，使用該分岔準則所建立出來的決策樹規則數目偏多，較容易造成過度學習的效應。為了修正此項偏誤，Quinlan(1993)提出 C4.5 決策樹演算法，重新定義「增益比值(Gain Ratio)」的計算公式來取代原有的分岔準則。首先，考量子資料庫(或稱子資料集合)切割時所產生的切割資訊量(split information)，其定義如下：

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2 \frac{|D_j|}{|D|} \quad (3)$$

分岔準則的目的是用以檢視使用某變數作為分岔變數時，母節點與子節點總和的純淨度變化量，能使純淨度提升越多的變數就是有效變數，基於這項目的，增益比值(Gain Ratio)的公式定義如下：

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(D)} \quad (4)$$

表 4 的範例資料中，得知骨髓移植後活著共 5 人，死亡的共 5 人，根據以下公式來計算母節點(狀態)的資訊量(即亂度)為：

$$Info_{\text{母體}} = -\frac{5}{10}log_2\frac{5}{10} - \frac{5}{10}log_2\frac{5}{10} = 1$$

表 4：範例資料

ID	年齡	性別	是否鉅細胞病毒感染	是否存活
1	10	男	是	是(存活)
2	4	男	否	否(死亡)
3	7	男	是	是(存活)
4	18	男	否	是(存活)
5	1	男	是	否(死亡)
6	11	女	否	是(存活)
7	16	女	否	否(死亡)
8	4	女	否	否(死亡)
9	18	女	否	否(死亡)
10	8	女	是	是(存活)

若用「是否鉅細胞病毒感染」來做為分岔變數。利用這個變數可將母節點切成 2 部分。第一部分是「是」，共計 4 人，有存活 3 人；第二部分是「否」，共計 6 人，沒有存活 4 人，因此，「是否鉅細胞病毒感染」作為分岔變數後，其亂度為：

$$Info_{\text{是否鉅細胞病毒感染}} = \frac{4}{10}\left(-\frac{3}{4}log_2\frac{3}{4} - \frac{1}{4}log_2\frac{1}{4}\right) + \frac{6}{10}\left(-\frac{4}{6}log_2\frac{4}{6} - \frac{2}{6}log_2\frac{2}{6}\right) = 0.8464$$

由此可知，利用「是否鉅細胞病毒感染」進行變數分割資料後，整體資料亂度由 1 降低至 0.8464，由此可計算出資訊獲利為 0.1536，即 Gain(是否鉅細胞病毒感染)=1-0.846439=0.1536，如圖 2 所示：

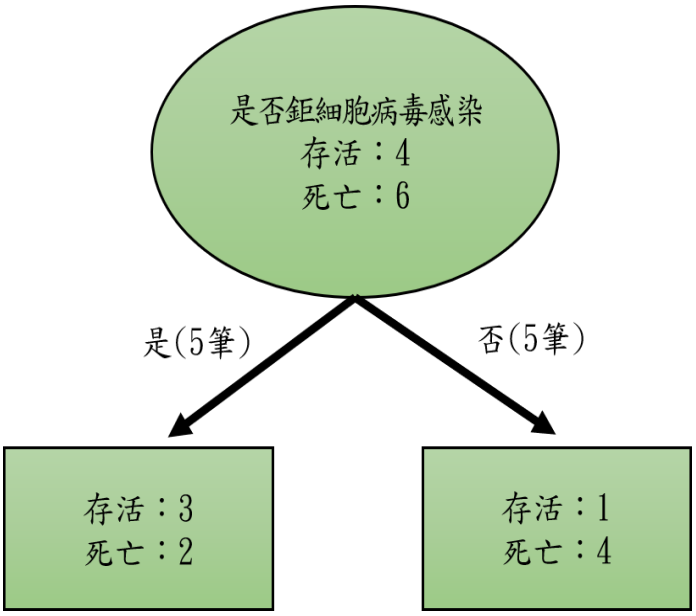

圖 2：以「是否鉅細胞病毒感染」為分岔變數

在這個案例中，透過「是否鉅細胞病毒感染」將資料由 10 筆切割為 4、6 筆兩部分，其代表「切割量」的分岔亂度為：

$$SplitInfo_{\text{是否鉅細胞病毒感染}}(D) = -\frac{4}{10}log_2\frac{4}{10} - \frac{6}{10}log_2\frac{6}{10} = 0.9710$$

因此，透過「是否鉅細胞病毒感染」作為第一層分岔變數時的增益比值為：Gain Ratio=0.153561/0.970951=0.1582

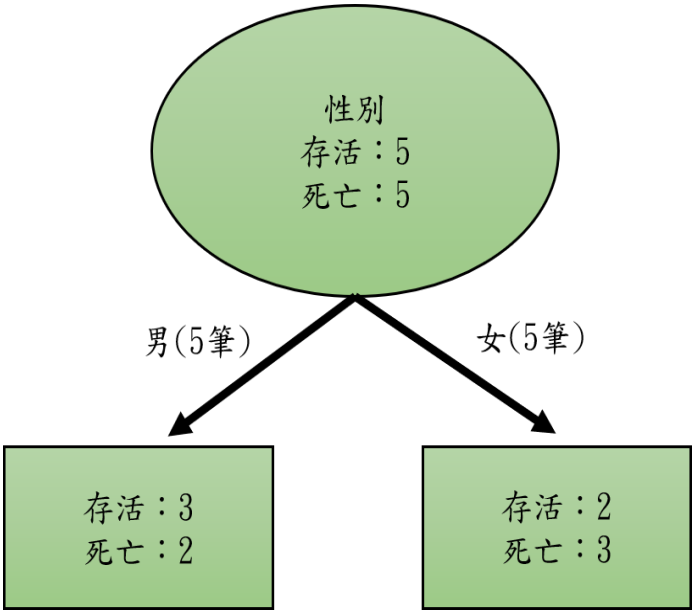

圖 3：以「性別」為分岔變數

重複上述步驟計算「性別」的分岔變數(如圖 3)，依照增益比值(GainRatio)的排序為：是否鉅細胞病毒感染(0.1582)>性別(0.0291)，由此可知「是否鉅細胞病毒感染」為最佳之第一層分岔變數，如表 5 所示。

表 5：範例屬性之增益比值計算統計表

	是否鉅細胞病毒感染	性別
母亂度	1.0000	1.0000
子亂度	0.8464	1.0000
資訊獲利	0.1582	0.0291
分岔亂度	0.9710	0.9710
增益比值	0.1581	0.0291

二、貝氏分類法

貝氏分類法(Bayesian Classifier)是以貝氏理論(Bayesian Theory)為基礎所設計出來的分類法，是屬於統計的分類法之一。貝氏理論是由 Thomas Bayes 於十八世紀初所提出。貝氏分類法主要是計算某案例屬於各類別機率。再根據此項機率資料，將此案例歸類於機率最高的類別。貝氏分類法的理論基礎乃是根據機率統計學中的貝氏定理(Bayesian Theorem)，貝氏定理定義如下：

$$\mathrm{P(A|B)} = \frac{\mathrm{P(B|A)P(A)}}{\mathrm{P(B)}} \tag{5}$$

應用此公式在分類上，*B* 代表某一個未知案例，*A* 代表某一類別，此公式的意義為：*B* 案例屬於 *A* 類別的機率=(*A* 類別中出現 *B* 案例的機率)×(*A* 類別出現的機率)/(*B* 案例出現的機率)。例如：假設要計算某一位病人得病的機率，則 *B* 案例即是這位病人，*A* 類別即是有得病的病患，根據貝氏定理：某位病人得病的機率=(有得病中出現某位病人的機率)×(得病的機率)/(某位病人出現的機率)。

然而，依據上述公式直接計算「某一位病人得病的機率」恐有困難，因為「判斷病患是否得死亡的機率」無法直接從已知的樣本中得知。為了將貝氏定理應用在分類上，令 *B=<B1，…，Bk>*，其中，*B1，…，Bk* 為案例 *B* 的 k 個屬性值。引用條件獨立的假設：人得病的機率」恐有困難，因為「判斷病患是否得死亡的機率」無法直接從已知的樣本中得知。為了將貝氏定理應用在分類上，令 *B=<b_1，…，b_k>*，其中，b_1，…，b_k 為案例 *B* 的 *k* 個屬性值。引用條件獨立的假設：

$$P\left(B =< b_1 \text{，} \dots \text{，} b_k > |A\right) \approx P(b_1|A) \dots P(b_k|A) \tag{6}$$

而公式(5)簡化為

$$P(C|X) = \frac{\mathrm{P}(X_1|\mathrm{C})\cdots\mathrm{P}(X_k)\times\mathrm{P(C)}}{\mathrm{P(X)}} \tag{7}$$

所謂的貝氏分類法即是利用公式(7)計算出未知案例屬於各個類別的機率，取機率值最大者做為該案例的類別預測。以上述「是否得病」為例，倘若該位病患「得病」大於「沒有得病」，則我們將預測該位病患有「得病」。沿用表 6 的範例資料為樣本。

表 6：範例資料

ID	年齡	性別	是否鉅細胞病毒感染	是否存活
1	5-10	男	是	是(存活)
2	1-5	男	否	否(死亡)
3	5-10	男	是	是(存活)
4	15-20	男	否	是(存活)
5	1-5	男	是	否(死亡)
6	11-15	女	否	是(存活)
7	15-20	女	否	否(死亡)
8	1-5	女	否	否(死亡)
9	15-20	女	否	否(死亡)
10	5-10	女	是	是(存活)
11	15-20	女	是	?

資料欄位有 ID、年齡、性別、是否鉅細胞病毒感染，預測是否存活。我們使用"ID 為 1"的資料來假設其基本資料為：年齡(5-10)、性別(男)、是否鉅細胞病毒感染(是)、預測是否存活(是)。本研究將以此範例說明如何運用貝氏分類預測該位新病人的診斷狀況。此範例中，"存活"的人有 5 筆，而"死亡"的人有 5 筆。因此，正常的機率為 P(存活)及死亡的機率為 P(死亡)分別計算如下：

- P(存活)=5/10=0.5
- P(死亡)=5/10=0.5

以下將針對"年齡"、"是否鉅細胞病毒感染"、"性別"等 3 種屬性，分別計算"存活"及"死亡"的條件下的各項條件機率：

- 年齡

  由於「年齡」欄位我們分成 4 個區間做離散化，分別為 1-5 歲(3 筆)、5-10 歲(4 筆)、15-20 歲(3 筆)，以下將分別計算在預測屬性「有存在鉅細胞病毒感染狀況」為"存活"及"死亡"前提下的分布機率如下：

  - P(1-5 歲｜存活)=0/3=0
  - P(1-5 歲｜死亡)=3/3=1
  - P(5-10 歲｜存活)=3/3=1
  - P(5-10 歲｜死亡)=0/3=0
  - P(15-20 歲｜存活)=1/3=0.333333
  - P(15-20 歲｜死亡)=2/3=0.666667

- 是否鉅細胞病毒感染

  由於「是否鉅細胞病毒感染」欄位中有兩種值，分別為否(6 筆)、是(4 筆)，以下將分別計算在預測屬性「有存活的狀況」為"存活"、"死亡"的前提下的分布機率如下：

  - P(感染=否｜存活)=2/6=0.333333
  - P(感染=否｜死亡)=4/6=0.666667
  - P(感染=是｜存活)=3/4=0.75
  - P(感染=是｜死亡)=1/4=0.25

- 性別

  由於「性別」欄位中有兩種值，分別為男(5 筆)、女(5 筆)，以下將分別計算在預測屬性「存活的狀況」為"存活"、"死亡"的前提下的分布機率如下：

  - P(男｜存活)=3/5=0.6
  - P(男｜死亡)=2/5=0.4
  - P(女｜存活)=2/5=0.4
  - P(女｜死亡)=3/5=0.6

- 已知有某位"新病患"的基本資料為：年齡(15-20)、是否鉅細胞病毒感染(是)、性別(女)，則預測新病人是否會得病如下：

  - P(新病患條件|存活)=P(15-20 歲|存活)×P(感染=否｜存活)×P(女｜存活)=0.333333×0.333333×0.4=0.044444
  - P(新病患條件|死亡)=P(15-20 歲|死亡)×P(感染=否｜死亡)×P(女｜死亡)=0.666667×0.666667×0.6=0.266667

  將以上結果代回到公式(7)得到：

  - P(存活|新病患條件)=P(存活)×P(新病患|存活)/P(新病患條件)=0.5×0.044444/P(新病患條件)
  - P(死亡|新病患條件)=P(死亡)×P(新病患|死亡)/P(新病患條件)=0.5×0.266667/P(新病患條件)

  為了簡化計算，可以進一步假設：P(存活|新病患條件)+P(死亡|新病患條件)=1。將此關係式加入上列公式解聯立方程式，則得到新病患死亡的機率為 86%，故預測新病患為「死亡」，其計算過程如下：

  - P(存活|新病患條件)=0.5×0.044444/(0.5×0.044444+0.5×0.266667)=14%
  - P(死亡|新病患條件)=0.5×0.266667/(0.5×0.044444+0.5×0.266667)=86%

三、類神經網路

神經網路神經元的組成是仿效人類神經元的結構，其結構如圖 4 所示。其中 X1、X2、X3 就是輸入變數值，而 W1、W2、W3 則是輸入變數的權重。X1 乘上 W1 就等於外部輸入的神經脈衝，但是在通過樹突時，神經脈衝必須大於門檻值，才能夠傳遞至神經元。所以對於神經元來說，所有的輸入訊號可以用下列公式來表示：

$$\text{Ij} = \sum_i W_{ij} O_i + \theta_j \quad (8)$$

其中 *I* 表示輸入值，*O* 表示前端神經元的輸出值，*W* 表示權重，而 *ϑ* 表示該神經原本常數項。

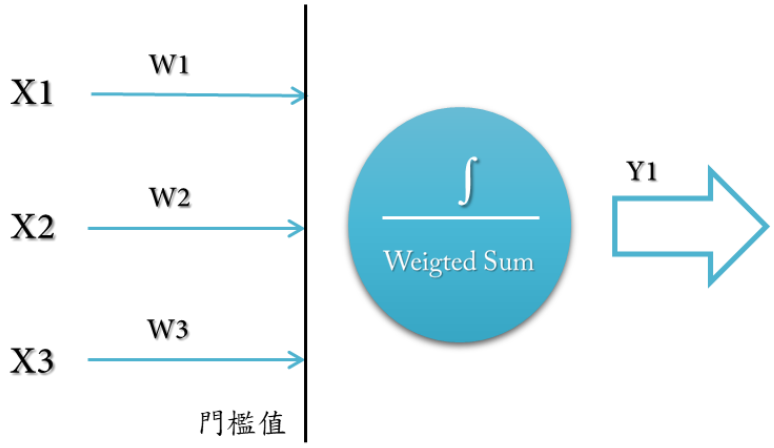

圖 4：神經元架構圖

當脈衝通過樹突進入神經元後，神經元會透過加總函數把所有的神經脈衝累加，然後透過轉換函數(Activation Function)的方式，產生新的神經脈衝(Y1)，向外傳遞。傳遞訊號的過程是使用轉換函數，以及神經元彼此連結的權重。倒傳導網路一開始權重是透過隨機的方式給予，因此一開始所產生的輸出結果與實際結果產生極大落差。而反向傳導則是當輸出值(預測結果)與實際結果有落差時，則將可以計算出單筆案例的誤差訊號，公式如下：

$$Error_j = O_j \times (1 - O_j) \times (T_j - O_j) \quad (9)$$

其中 *O* 表示輸出訊號，而 *T* 表示實際的真值。

所謂的倒傳導，指的是要將此誤差項從輸出層反饋至隱藏層，此時，輸出層誤差項會根據神經連結，將誤差依照權重分配至隱藏層神經元，公式如下：

$$Error_j = O_j \times (1 - O_j) \times \sum_k (Error_k \times W_{jk}) \quad (10)$$

演算法可以根據誤差訊號的大小，同步修正各神經連結的權重，假設修正量與輸入訊號強度以及錯誤訊號成正比，此時，修正量以及修正後權重可以透過下式表示：

$$\Delta W_{ij} = l \times Error_j \times O_i \quad (11)$$

$$\Delta\theta_j = l \times Error_j \quad (12)$$

其中上式中的 *l* 表示學習速率(Learning Rate)，通常是介於 0~1 之間，當數值越大，每次權數的修正量就會越大。因此，類神經網路的學習重心在於如何自動的、有效率的調整權重大小以及學習速率，目前最常使用的方式是最陡坡降法(The Gradient Steepest Descent Method)，就是將誤差已可微分函數表示，透過微分斜率的梯度來產生權重的修正量，誤差為正向時(輸出值大於實際值)，此時，降低神經權重，反之則增加神經權重，如此不斷反覆學習，讓誤差降低，這個過程就是所謂的「學習」(Han & Kamber，2006)。

本研究以圖 5 的相位為例，依序來說明類神經網路對於權重的計算模式的步驟。首先利用隨機的方式，產生各神經元間的權重，以及隱藏層與輸出層神經元的常數項，如表 7 所示。

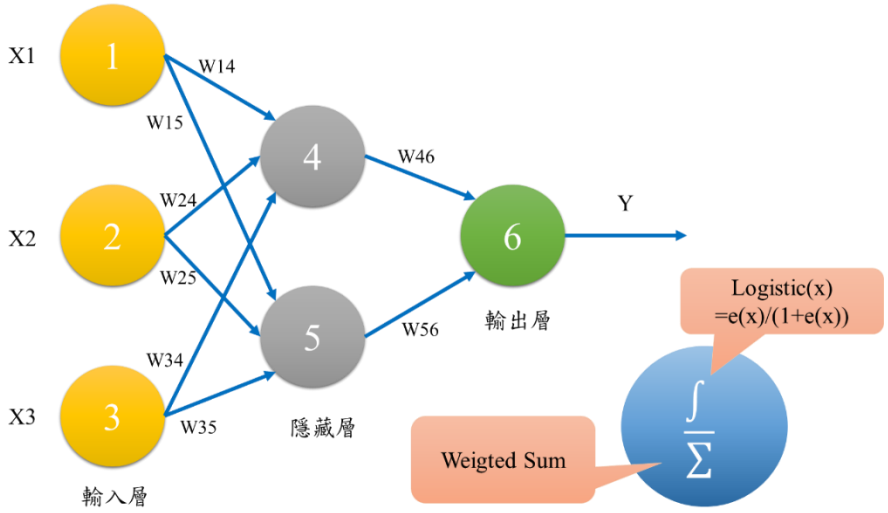

圖 5：類神經範例架構圖

表 7：隨機產生的參數(權重及常數項)

W_{14}	W_{15}	W_{24}	W_{25}	W_{34}	W_{35}	W_{46}	W_{56}	θ_4	θ_5	θ_6
0.2	-0.3	0.4	0.1	-0.5	0.2	-0.3	-0.2	-0.4	0.2	0.1

假設輸入案例為(X1,X2,X3,Y)=(1,0,1,1)，根據輸入訊號，計算各隱藏層神經元的輸出訊號。以下將依序說明神經元 4,5,6 的計算過程。

- 神經元 4：
  - 總輸入訊號：$I_j = \sum_i W_{ij} O_i + \theta_j = X1 \times W_{14} + X2 \times W_{24} + X3 \times W_{34} + \theta_4$
    $= 1 \times 0.2 + 0 \times 0.4 + 1 \times (-0.5) + (-0.4) = -0.7$
  - 輸出神經元：$O_j = \frac{1}{1+e^{-ij}} = \frac{1}{1+e^{-(-0.7)}} = 0.3318$

- 神經元 5：
  - 總輸入訊號：$I_j = \sum_i W_{ij}\, O_i + \theta_j = X1 \times W_{15} + X2 \times W_{25} + X3 \times W_{35} + \theta_5$
    $= 1 \times (-0.3) + 0 \times 0.1 + 1 \times 0.2 + 0.2 = 0.1$
  - 輸出神經元： $O_j = \frac{1}{1+e^{-ij}} = \frac{1}{1+e^{-0.1}} = 0.525$
- 神經元 6：
  - 總輸入訊號：$I_j = \sum_i W_{ij}\, O_i + \theta_j = O_4 \times W_{46} + O_5 \times W_{56} + \theta_6$
    $= 0.3318 \times (-0.3) + 0.525 \times (-0.2) + 0.1 = -0.105$
  - 輸出神經元：$O_j = \frac{1}{1+e^{-ij}} = \frac{1}{1+e^{-(-0.105)}} = 0.4739$

此時，輸出值 0.4739 與實際值(1)不一致，因此可以計算神經元 6 的誤差項。神經元 6 的誤差項：$Error_j = O_j(1 - O_j)(T - O_j) = 0.4739 \times (1 - 0.4739) \times (1 - 0.4739) = 0.1312$。將此誤差項反饋至隱藏層，可計算出隱藏層神經元誤差值。

- 神經元 4 誤差項：$Error_j = O_j(1 - O_j) \sum_k Error_k\, W_{jk} = 0.3318 \times (1 - 0.3318) \times 0.1312 \times (-0.3) = (-0.0087)$
- 神經元 5 誤差項：$Error_j = O_j(1 - O_j) \sum_k Error_k\, W_{jk} = 0.525 \times (1 - 0.525) \times 0.1312 \times (-0.2) = (-0.0065)$

最後根據神經元的誤差項，更新各神經元的權重以及常數項，假設學習速率為 0.9。如此就達成一個學習循環的類神經網路權重新修正，接下來持續此步驟，即可以使得輸出值越來越接近實際值，而達到建立模型的目的。修正後的參數(權重及常數項)如下：

$$W_{14} = 0.2 + 0.9 \times (-0.0087) \times 1 = 0.1921$$
$$W_{15} = (-0.3) + 0.9 \times (-0.0065) \times 1 = -0.3059$$
$$W_{24} = 0.4 + 0.9 \times (-0.0087) \times 0 = 0.4$$
$$W_{25} = 0.1 + 0.9 \times (-0.0065) \times 0 = 0.1$$
$$W_{34} = (-0.5) + 0.9 \times (-0.0087) \times 1 = -0.5079$$
$$W_{35} = 0.2 + 0.9 \times (-0.0065) \times 1 = 0.1941$$
$$W_{46} = (-0.3) + 0.9 \times 0.525 \times 0.1312 = -0.2608$$
$$W_{56} = (-0.2) + 0.9 \times 0.13 \times 0.1312 = -0.138$$
$$\theta_4 = (-0.4) + 0.9 \times 0.1312 = 0.2181$$
$$\theta_5 = 0.2 + 0.9 \times (-0.0065) = 0.1914$$
$$\theta_6 = 0.1 + 0.9 \times (-0.0087) = -0.4079$$

肆、實驗

一、實驗資料

本章節將使用 kaggle(加州大學機器學習與智慧系統研究中心)的 Bone marrow transplant (骨髓移植數據集)資料集，本研究將 187 筆病患資料區分成 169 筆訓練資料，則測試資料為 19 筆，資料量比例約為 9：1。其中，資料集屬性欄位包含：ID、年齡、性別、血型、人類紅血球表面有無 RhD 抗原、是否鉅細胞病毒感染、患者疾病、移植血型匹配、鉅細胞病毒狀態、人類白血球抗原匹配、抗原、對偶基因、生物風險、幹細胞來源、診斷等。資料屬性欄位(如表 8 所示)：

表 8：實驗資料(部分)

屬性	病例 1	病例 2
ID	144	9
年齡	5	18
性別	女	男
血型	B 型	A 型
人類紅血球表面有無 RhD 抗原	陽性	陽性
是否鉅細胞病毒感染	否	否
患者疾病	急性淋巴性白血病	非惡性腫瘤
移植血型匹配	不匹配	不匹配
鉅細胞病毒狀態	0	1
人類白血球抗原匹配	匹配	不匹配
抗原	2	2
對偶基因	1	3
生物風險	低	低
幹細胞來源	末梢血液	骨髓血液
診斷	存活	死亡

資料參考:
https://www.kaggle.com/adamgudys/bone-marrow-transplant-children

二、C4.5 決策樹、貝氏分類、類神經網路預測模型的評比

分類模型建構完成之後，本研究建立混亂矩陣(如表 9 所示)，用以比較各模型技術之差異。混亂矩陣中項目所代表的意義以下分別描述：

表 9：二元分類問題的混亂矩陣(confusion matrix)

		Predicted Class	
		Positive	Negative
Actual Class	Positive	True-Positive (TP)	False-Negative (FN)
	Negative	False-Positive (FP)	True-Negative (TN)

- True Positive (TP)：真陽性，實際上是陽性(Positive)，而且真的被判定為陽性(Positive)。
- False Positive (FP)：偽陽性，實際上是陰性(Negative)，但是卻被判定為陽性(Positive)。
- True Negative (TN)：真陰性，實際上是陰性(Negative)，而且真的被判定為陰性(Negative)。
- False Negative (FN)：偽陰性，實際上是陽性(Positive)，但是卻被判定為陰性(Negative)。

如何去評估一個分類系統的優劣？傳統上，我們會使用 Accuracy(準確率)、Precision(精確率)、Specificity(特異度)、Recall(召回率)以及 F-Measure 來評估一個分類系統的好壞(Tan et al.，2005)。其公式定義如下：

$$Sensitivity = Recall = TPR = \frac{TP}{TP + FN} \tag{13}$$

$$Specificity = TNR = \frac{TN}{FP + TN} \tag{14}$$

$$FPR = \frac{FP}{FP + TN} = 1 - TNR \tag{15}$$

$$FNR = \frac{FN}{TP + FN} = 1 - TPR \tag{16}$$

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \tag{18}$$

$$F_1 = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{19}$$

三、實驗結果

本研究使用 Kaggle(加州大學機器學習與智慧系統研究中心)的 bone-marrow(資料集共 188 筆病患資料中，即採用 10 折交叉驗證方法(如圖 6 所示)。最後，本研究運用 Weka 所提供的「C4.5 決策樹」、「貝氏分類」、「類神經網路」等演算法建立骨髓移植之研究模型的評比結果，如圖 7、8、9 所示。

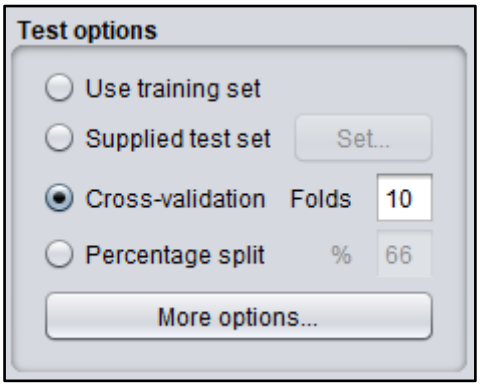

圖 6：參數設置

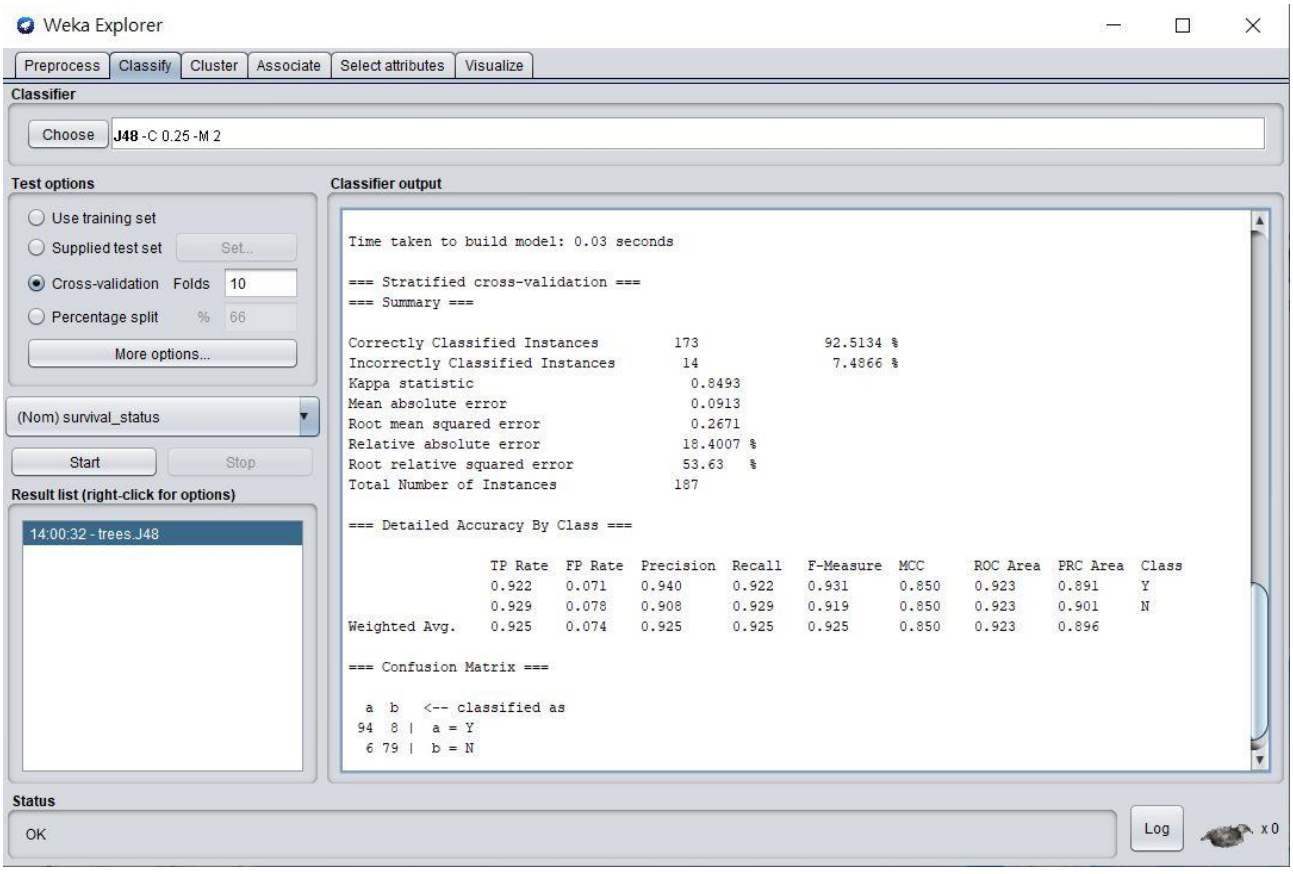

圖 7：決策樹分類模型的評比結果

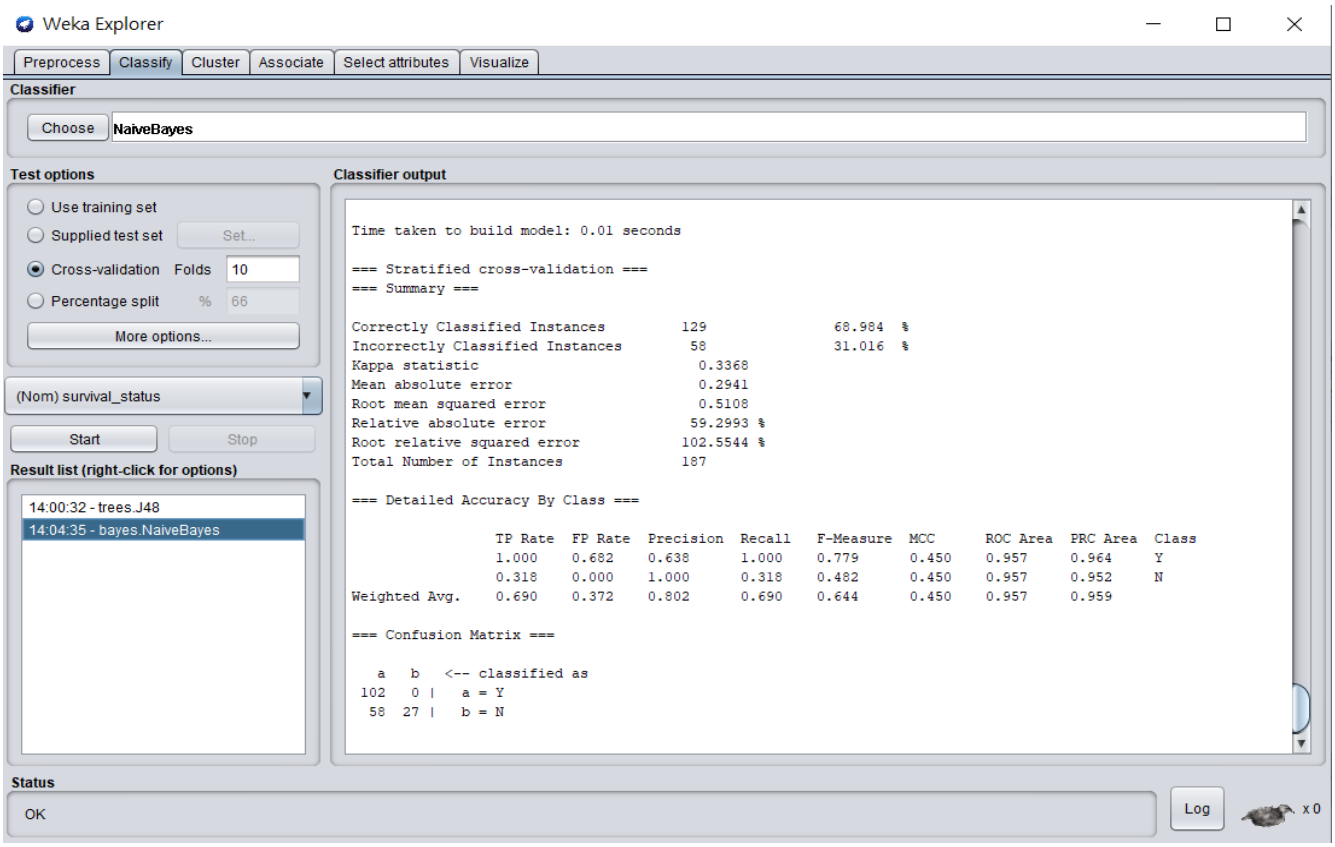

圖 8：貝氏分類模型的評比結果

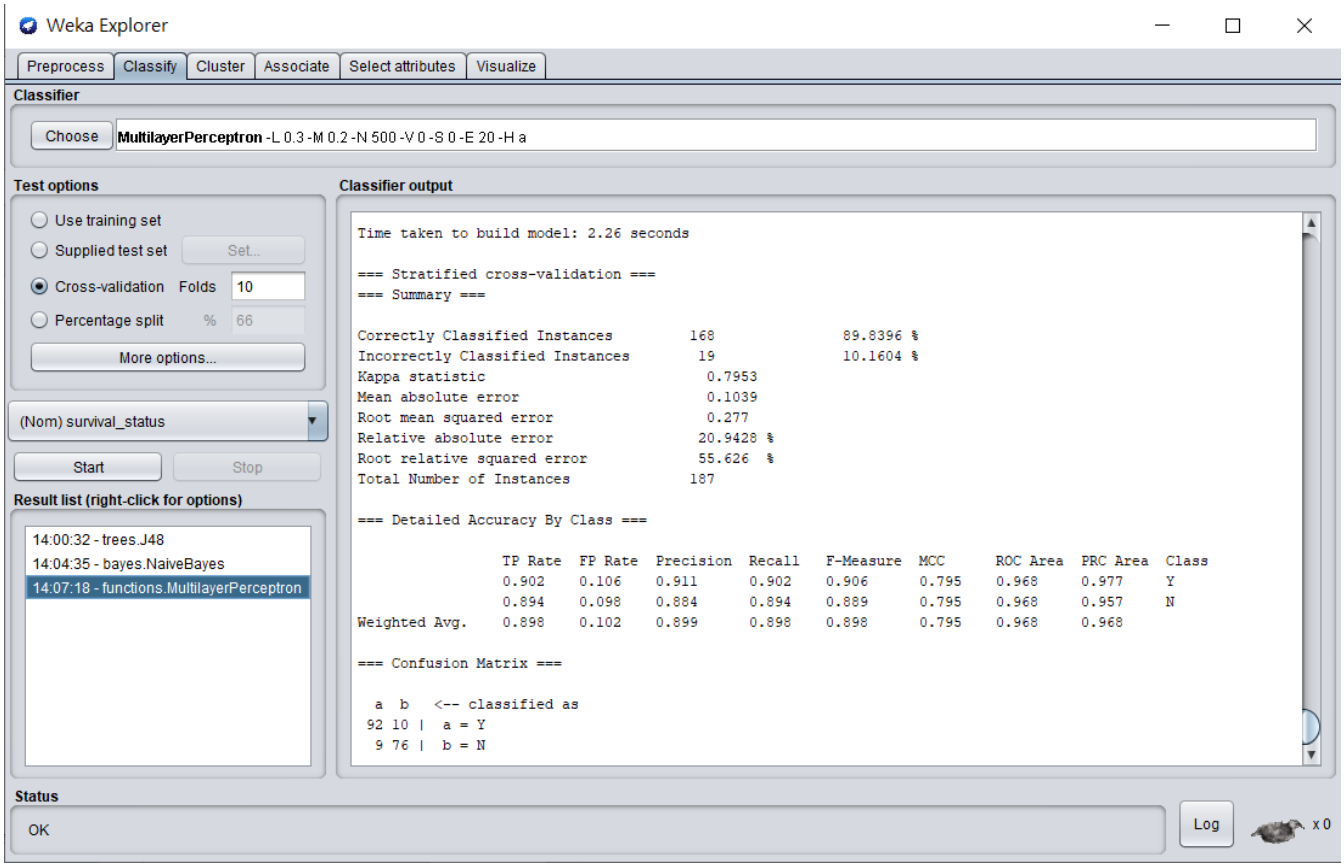

圖 9：類神經網路分類模型的評比結果

本研究使用 3 種分類技術(C4.5 決策樹、貝氏分類、類神經網路)於骨髓移植預測，分類模型效能(如表 10 所示)，

本研究發現：(1)在 Recall(或稱為 TPR)以貝氏分類表現最佳，其準確為 100%；(2)在 Specificity(特異度，即 TNR)以貝氏分類表現最佳，其準確為 68.2%；(3)Precision(精確率)以決策樹表現最佳，其準確為 94%；(4)Accuracy(準確度)以貝氏分類表現最佳，其準確為 100%；(5)F1 指標決策樹表現最佳，其準確為 93.1%。

表 10：分類技術比較表

分類技術	決策樹	貝氏分類	類神經網路
TPR	0.922	1.000	0.902
TNR	0.071	0.682	0.106
Precision	0.940	0.638	0.911
Accuracy	0.922	1.000	0.902
F-Measure	0.931	0.779	0.906
MCC	0.850	0.450	0.795
ROC Area	0.923	0.957	0.968
PRC Area	0.891	0.964	0.977

四、分類規則

本研究使用決策樹演算法所產生決策樹(如圖 10 所示)，從決策樹圖形中可以整理出 4 項關聯規則，詳細說明如下：

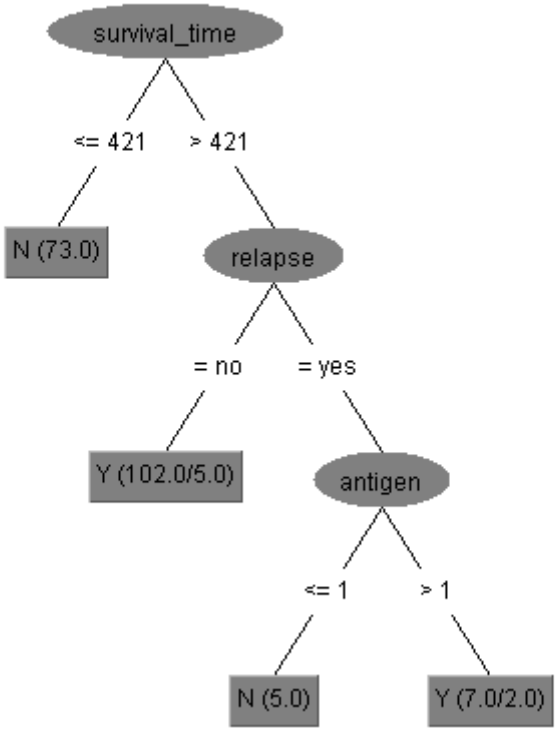

圖 10 決策樹分析結果之樹狀圖

1. 如果「存活時間<=421 天」，則「骨髓移植死亡」其機率為 39% (73/187)。
2. 如果「存活時間>421 天」，且「疾病有復發」，且「抗原<=1」，則「骨髓移植死亡」其機率為 95% (10/187)。
3. 如果「存活時間>421 天」，且「疾病沒有復發」，則「骨髓移植存活」其機率為 98%(5/187)
4. 如果「存活時間>421 天」，且「疾病沒有復發」，且「抗原>1」，則「骨髓移植死亡」其機率為 95% (5/187)。

伍、結論

近年來資料探勘的技術被廣泛應用在各領域中，其中Ｃ4.5 決策樹、貝氏分類法、類神經網路皆為重要的資料探勘技術，被應用於諸多領域而且皆有不錯的成效。

本研究以 kaggle (加州大學機器學習與智慧系統研究中心)的 Bone marrow transplant (骨髓移植)資料集，將資料探勘技術運用於骨髓移植預測，分別運用 C4.5 決策樹、貝氏分類法、類神經網路(MLP)等演算法建立骨髓移預測模型，期望能提供給醫師做為骨髓移植是否成功之用。

其實驗結果顯示：決策樹演算法於 Precision (0.940), F-Measure (0.931), MCC (0.850)等指標表現較佳；貝氏分類於 TPR (1.000), Accuracy (1.000)等指標表現較佳；類神經網路(MLP)則於 ROC Area (0.968), PRC Area (0.977)等指標表現較佳。

在未來，本研究希望能結合其他的資料探勘技術及更多的資料來預測並做評估比較，以提高預測模型的準確率。

參考文獻

中文文獻：

1. 游介宇(民 102)。接受骨髓移植的病人容易罹癌嗎？。台灣癌症防治網。取自 https://reurl.cc/eEnKaW
2. 唐季祿(民 105)。一本讀通血癌。台北市：天下生活。
3. 林宇恆(民 105) 基於決策樹的統計預報模型利用該模型預測領前 1 至 7 天的氣溫與雨量，國立臺灣師範大學電機工程學系碩士論文，未出版，台北市。
4. 孫苙達(民 107)。利用決策樹演算法探討總體經濟變數並檢驗決策樹的預測準確度與作為交易策略的績效，國立中山大學財務管理學系碩士論文，未出版，高雄市。
5. 廖介銘(民 92)。應用決策數先以統計方式篩選有顯著差異之變數，再將其戴入至決策樹演算法歸納出糖尿病決定因子，華梵大學資訊管理學系碩士論文，未出版，新北市。
6. 林育成(民 103)。藉由演算法中複製、變異、選擇以及保留最佳抗體等概念，得到最佳化的參數與選擇之特徵值輸入決策樹進行血球分類，華梵大學資訊管理學系碩士論文，未出版，新北市。
7. 詹育佳(民 107)。用決策樹加以歸納分析出會員分類模型，找出芳療顧客消費及使用精油的決定影響因子。龍華科技大學資訊管理系碩士論文，未出版，桃園市。
8. 翁毓謙(民 94)。利用貝氏分類決策理論，設計出以最小化貝氏期望風險為主之語音辨識演算法，國立成功大學資訊工程學系碩士論文，未出版，台南市。
9. 曾麗文(民 100)。採用多項貝氏分類法之分類方法進行實驗比較，進行多元分類結果、ROC 曲線、正確率、精確度、召回率等分析，國立成功大學工業與資訊管理學系碩士論文，未出版，台南市。
10. 黃代鈞(民 93)。利用貝氏分類法在數量廣大的人類基因中，分析基因表現的資料找出遺傳疾病相關的關鍵基因，長庚大學資訊管理系碩士論文，未出版，桃園市。
11. 廖育民(民 106)。利用導入貝氏分類，將失智症的危險因子，以此建構出可行的失智症臨床初期診斷模式，逢甲大學應用數學系碩士論文，未出版，台中市
12. 鄭鈺傑(民 104)。利用智慧型手機以車輛對稱性與貝氏分類為基礎的前車辨識演算法，國立東華大學電機工程學系碩士論文，未出版，花蓮縣。
13. 林逸塵(民 91)。利用類神經網路預測高雄都會區之懸浮微粒濃度及能見度，並討論類神經網路之優缺點及可能之改善方法，國立中山大學環境工程系碩士論文，未出版，高雄市。

14. 陳采蓉(民 109)。應用類神經網路和羅吉斯迴歸分析來預測膀胱鏡檢查後導致泌尿道感染的機率，國立嘉義大學生化科技學系暨研究所碩士論文，未出版，嘉義市。
15. 曾世任(民 100)。利用類神經網路透過低複雜性的數學運算過程，並將其作為辨識工具加上不同的量化方法，國立高雄海洋科技大學電訊工程所碩士論文，未出版，高雄市。
16. 辜炳寰(民 91)。使用類神經網路理論，提出一套理論化的簡易經驗評估法，來研究液化問題，國立成功大學土木工程系碩士論文，未出版，台南市。
17. 鍾炳煌(民 91)。以類神經為分析工具，應用於汽車駕照模擬系統從事高速加速車道併入行為之研究，國立成功大學交通管理科學系碩士論文，未出版，台南市。

英文文獻：

1. Berry，M.，and Linoff，G.(2000)“Mastering Data Mining:The Art and Science of Customer Relationship Management”，JohnWiley&Sons，NewYork.
2. Fayyad， U.， Piatetsky-Shapiro， G.， & Smyth， P.(1996).The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*，39(11)，27-34.
3. Han，J.，and Kamber，M.(2006).Data Mining:Concepts and Techniques(2nded.)，Morgan Kaufmann Publishers，San Francisco，CA.
4. Quinlan， J. R.(1979).Discovering rules from large collections of examples: a case study in MICHIE.Expert Systems in the Micro Electronic Age， 1-100.
5. Quinlan， J. R.(1993).C4.5:Programs for machine learning”，Morgan Kaufmann Publishers.
6. Shannon，C.E.，and Weaver，W.(1949).“The Mathematical Theory of Communication”， IL:The University of Illinois Press，Urbana，1-117.
7. Tan，P.N.，Steinbach，M.，Kumar，V.(2005).“Introduction to Data Mining，”Addison-Wesley，Indiana.