南　華　大　學

資訊管理學系碩士班

碩士論文

應用資料探勘技術於慢性腎臟病之預測

Application of Data Mining Techniques for Detection of chronic kidney disease

研 究 生：黃原博

指導教授：邱宏彬

中華民國 99 年 6 月 25 日

南　華　大　學

資訊管理學系資訊管理研究所

碩 士 學 位 論 文

應用資料探勘技術於慢性腎臟病之預測

研究生：黃厚博

經考試合格特此證明

口試委員：謝碧霞

李翔詔

邱宏彬

指導教授：邱宏彬

系主任(所長)：鍾國忠

口試日期：中華民國 99 年 06 月 25 日

南華大學資訊管理學系碩士論文著作財產權同意書

立書人：黃原博 之碩士畢業論文

中文題目：應用資料探勘技術於慢性腎臟病之預測

英文題目：Application of Data Mining Techniques for Detection of chronic kidney disease.

指導教授：邱宏彬 博士

學生與指導老師就本篇論文內容及資料其著作財產權歸屬如下：

☑ 共同享有著作權

☐ 共同享有著作權，學生願「拋棄」著作財產權

☐ 學生獨自享有著作財產權

學　　生：黃原博 (請親自簽名)

指導老師：邱宏彬 (請親自簽名)

中　華　民　國　90　年　5　月　20　月

南 華 大 學 碩 士 班 研 究 生

論 文 指 導 教 授 推 薦 函

資訊管理 系碩士班 黃原博 君所提之論文

應用資料探勘技術於慢性腎臟病之預測

係由本人指導撰述，同意提付審查。

指導教授 邱宏彬

99 年 05 月 20 日

誌　　謝

本論文之完成，首先要感謝的是我的指導教授邱宏彬老師，謝謝老師從論文開始的題目訂定、整體架構的建立及最後的細部修改，皆在有限的時間內耐心的協助我們完成，衷心感謝老師在此期間的鼓勵及教導。另承蒙口試委員謝昆霖老師及李翔詣老師在口試時不吝指正本論文之缺失並提供寶貴的意見及指正，在此致上最誠摯的感謝。

感謝在學期間校內師長及同學們熱心的協助及幫忙，讓我能在無後顧之憂的情形下，完成這二年愉快的求學生涯。

謹以此論文獻給我親愛的家人、阿信班代、科富、宜玲、學姐、美玲、抒帆、因子、德林、幼幼、伊文、王Sir等，感謝各位在此期間不論關心我為我加油打氣或是對論文完成有著直接的幫助，於此皆表達我深切的謝意。另外更要祝福你們，並將此完成學業的喜悅與大家分享。

原博

謹誌於南華大學資訊管理學系碩士班

2010/06

應用資料探勘技術於慢性腎臟病之預測

學生:黃原博　　指導教授:邱宏彬 博士

南　華　大　學　資訊管理學系碩士班

摘　　要

慢性腎臟疾病是目前全世界主要的公共健康問題之一。根據臺灣腎臟醫學會資料顯示，臺灣洗腎發生率高居世界第一位，平均每 1.2 小時，就有一名新增洗腎患者，推估慢性腎臟病第三期至第五期病人，更高達 115 萬人，超越糖尿病的百萬人口，儼然成為新國病(邱鼎鈺,健保雙月刊第 71 期)。我們如果能即早發現達到早期治療之效果，將可以改善末期腎臟病變不斷升高的情形，進而改善末期腎病變的早發性更可達到減少健保資源之耗費。本研究以南部某區域醫院之慢性腎臟個案管理病患資料為例，期望藉由資料探勘技術研究慢性腎臟病各分期演變及針對各資料數據做統計分析，找出變數之影響性及各統計數據之意義。針對本研究的發現提出相關之建議，提供臨床醫護人員協助慢性腎臟病患者各 Stage 適切之醫療診治、護理衛教等服

務。

關鍵詞:資料採礦、慢性腎臟病、決策樹、類神經網路

Application of Data Mining Techniques for Detection of chronic kidney disease

Student：Huang,Yuan-Po　　Advisors：Dr. Chiu,Hung-Pin

Department of Information Management
The Graduated Program
Nan-Hua University

ABSTRACT

Chronic renal disease is one of the most important public health problems all over the world. According to the Taiwan Society of Nephrology, the incident of dialysis in Taiwan ranked first in the world's new patient of dialysis was added every 1.2 hours; the number of chronic renal disease patients from stage III to stage V is estimated to reach up to 1.15 million, exceeding the number of million of diabetes patients. As if the new disease of Taiwan (Chiu Ting Yu, the author of Bi-monthly health insurance 71). End stage renal disease is usually slowly transformed from the chronic renal disease stage I to stage V. If we discover the symptom and put it into remedy early, we can stop the end stage renal disease from rising, and then improve the early onset of end stage renal disease so as to reduce the waste of health care recourses. This study refers to the cases of chronic renal disease patients managed by a southern regional hospital, expecting by data mining techniques for all stages of chronic kidney disease and the evolution of information and data for the statistical analysis to identify variables affecting the meaning of various statistical data. Against the findings of this study are some suggestions to provide clinical staff with appropriate medical treatment, care and health education services to chronic renal disease patients of all stages.

keywords：data mining、chronic renal disease、decision trees、natural networks

目　錄

表目錄

圖目錄

第一章 緒論

本章主要說明本研究之研究背景、研究動機、研究目的、研究範圍與限制、研究步驟以及論文架構。首先透過了解研究背景與動機，說明國內慢性腎臟病患之現況，進而歸納出研究目的，然後再針對研究範圍與限制做出說明，最後對於研究步驟及論文研究架構做概括之描述。

第一節 研究背景與動機

台灣慢性腎臟病的好發率達11.9%，平均每10人中，至少有1 人罹患，腎臟就像是人體任勞任怨的「阿信」，即使受傷也不會喊疼，「**腎臟疾病初期最常見的症狀就是沒有症狀**」，等到腎功能失去6 成以上才會出現警訊，許多人罹患慢性腎臟病卻完全不知(方德昭,2010)。

根據近期統計臺灣洗腎人口比例是全世界第1，主因國人習慣的賣藥買藥文化存在已久，國人大多數都習慣在未經醫師評估下自行至藥房買成藥服用，其中最普遍的尤是感冒、頭痛、生理痛及其它疼痛問題，民眾很容易上藥房亂買成藥解決。

另一部份，中南部老年人容易受到地下電臺吹噓偏方、不實藥品廣告後，聽信其效果並自行訂購。但是電台所販賣之健康食品、偏方、減肥藥品等藥品大部份皆為偽藥。長期服用這些藥品，可能增加

腎臟的負荷，如果再加上保腎觀念不足，進而產生臉腫、腳腫等症狀，就診驗血後才發現自己的腎臟早已出問題，最後往往落得必須終身洗腎的命運。而這一結果造成了健保成本龐大的負擔，經行政院衛生署統計如表 1-1 所示各縣市民眾因腎臟相關病症就診人數皆不在少數，另表 1-2 中所列國人從 75 年至 96 年間因腎炎、腎徵候群及腎性病變等因素死亡者明顯的逐年增加中。

表 1-1 95 年度各縣市門、住診(包括急診)人數統計表(行政院衛生署)

縣市別	門、住診合計(包括急診)人數		
	總計	男	女
總計 Total	341,042	184,486	156,556
台北市 Taipei City	40,247	22,144	18,103
台北縣 Taipei County	45,229	24,086	21,143
宜蘭縣 Ilan County	7,762	4,081	3,681
基隆市 Keelung City	5,680	2,961	2,719
金門縣 Kinmen County	831	479	352
連江縣 Lienchiang County	132	78	54
桃園縣 Taoyuan County	22,712	12,728	9,984
新竹縣 Hsinchu County	5,699	3,155	2,544
苗栗縣 Miaoli County	8,098	4,353	3,745
新竹市 Hsinchu City	4,492	2,545	1,947
台中縣 Taichung County	21,359	11,376	9,983
彰化縣 Changhua County	19,084	9,961	9,123
南投縣 Nantou County	9,609	4,895	4,714
台中市 Taichung City	13,545	7,212	6,333
雲林縣 Yunlin County	13,423	7,121	6,302
嘉義縣 Chiayi County	9,862	5,317	4,545
台南縣 Tainan County	20,533	11,090	9,443
嘉義市 Chiayi City	4,345	2,274	2,071
台南市 Tainan City	12,655	6,874	5,781
高雄市 Kaohsiung City	26,659	14,518	12,141
高雄縣 Kaohsiung County	23,199	12,821	10,378

縣市別	門、住診合計(包括急診)人數		
	總計	男	女
屏東縣 Pingtung County	15,441	8,493	6,948
澎湖縣 Penghu County	2,266	1,294	972
台東縣 Taitung County	3,483	1,960	1,523
花蓮縣 Hualien County	4,623	2,632	1,991

表 1-2 歷年死亡人數統計表(行政院衛生署)

民國 75 年至 96 年			
年別 Year		腎炎、腎徵候群及腎性病變	順位
75 年	1986	2,134	11
76 年	1987	2,113	10
77 年	1988	2,064	10
78 年	1989	2,094	10
79 年	1990	2,304	10
80 年	1991	2,527	8
81 年	1992	2,935	7
82 年	1993	2,859	7
83 年	1994	3,211	7
84 年	1995	3,519	7
85 年	1996	3,547	7
86 年	1997	3,504	8
87 年	1998	3,435	8
88 年	1999	3,474	8
89 年	2000	3,872	7
90 年	2001	4,056	7
91 年	2002	4,168	8
92 年	2003	4,306	8
93 年	2004	4,680	8
94 年	2005	4,822	8
95 年	2006	4,712	8
96 年	2007	5,099	8

另外依據中央健保局資料顯示在健保的醫療支出中，以重傷病為最大宗，健保局民國九十一年的統計資料，指出台灣目前約有五十六萬多人擁有重大傷病卡，而這些重大傷病患者一年醫療費用約 769 億元，佔當年度健保總醫療費用支出 22.64%，在 31 項公布的重大傷病項目中，其中**洗腎經費以 245 億元佔支出的第一位**，洗腎經費佔總醫療支出 **7.21%**，醫療費用支出均呈現成長趨勢（如圖 1-1、圖-2）（洪嘉鴻, 2004）。

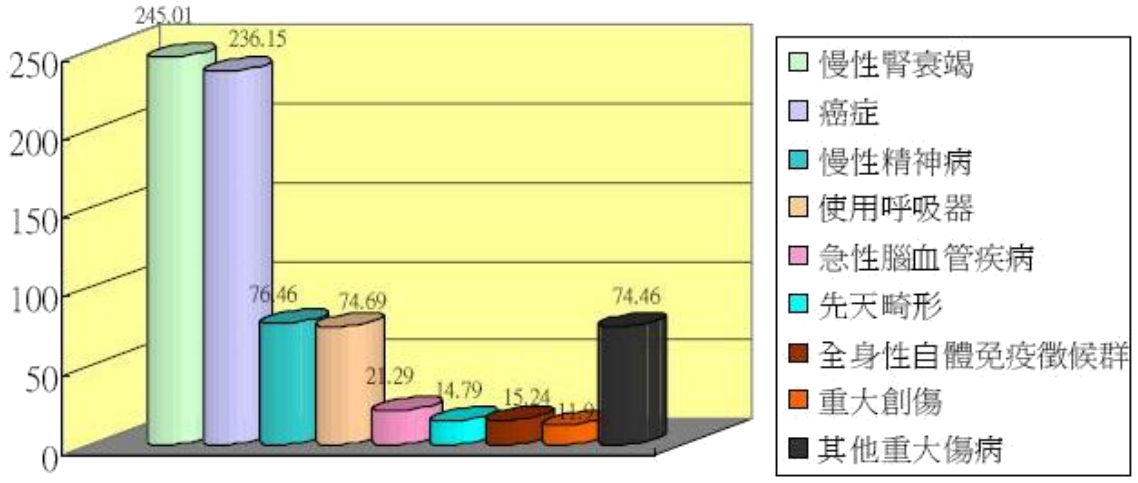


圖 1_1 重大傷病醫療費用申報狀況

資料來源:中央健康保險局，91年全民健康保險統計

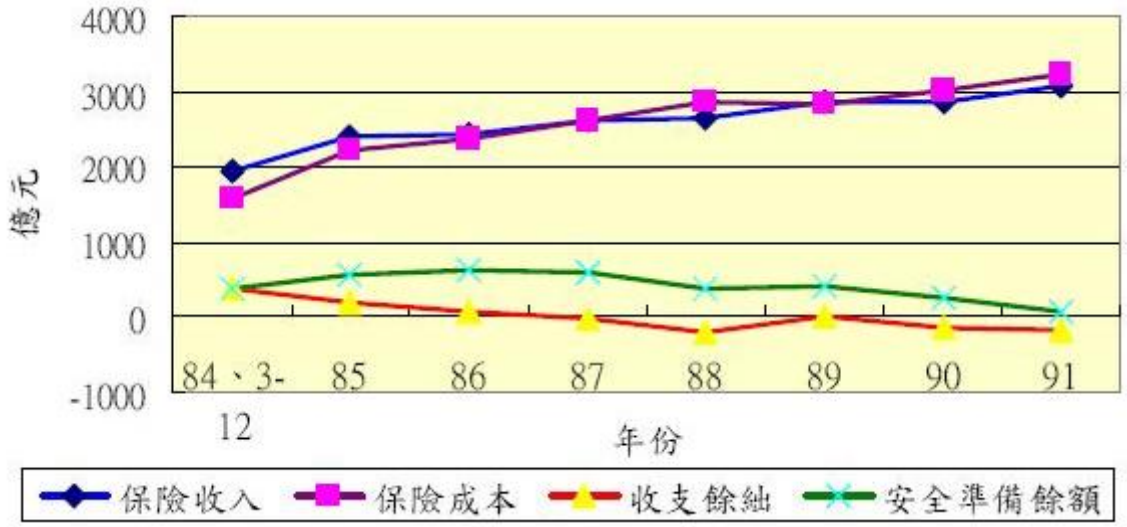


圖 1_2 全民健保歷年財務收支

資料來源：行政院衛生署，中華民國九十一年版公共衛生年報

中央健保局針對 2008 年門診統計前五大疾病排行，其中洗腎醫療支出高達 338 億元，首度超越牙科費用的 337 億元支出，躍居門診費用第一名。再者由於國內人口老化，罹患慢性病的機率增加，因此包括慢性腎衰竭、高血壓和糖尿病的病人從健保開始施行以來不斷的持續攀升，另一方面國內洗腎技術好，病患存活長，洗腎人口持續成長，費用也跟著上升。

台灣洗腎人口逐年成長，從九十六年的五六〇九〇人，成長到九十八年的的五八七二六人，每位洗腎病患，一周洗腎三次，一年約花費六十萬元（李丞華，2008）。

第二節 研究目的

因應台灣腎炎、腎徵候群及腎性病變等病患有日益增加之趨勢，一線從業醫護人員對於“早期發現早期治療”此一要點，如果希望能透過合適之診治、照護及衛教，讓病患在獲得妥善之醫療行為進而減緩慢性腎臟病各 Stage(分期)之快速惡化確是不易。

再者由於腎炎、腎徵候群及腎性病變等病症早期症狀較不易發現，就算能夠早期發現病變，通常病患因自覺身體並無嚴重之不適情形，往往不會特別去注意及自我照護進而導致病變情形快速惡化。

就以上所述不論從醫護人員或腎臟病患的角度來看，欲切實達到穩定控制病情不致使其嚴重惡化，皆因病情變化不易發現而有落實執行之難度。本研究擬以應用資料探勘技術分析現有病患個案之資料，找出影響病患各 Stage 變化顯著之要項，作為醫護人員做為診療、照護行為時，針對資料探勘之結果獲得參考之相關數據以利預測病患各 Stage 的演變時程。

第三節 研究範圍與限制

本研究所取得之資料為南部某區域醫院之慢性腎臟個案管理病患資料為例，因為只以單一地區病患之資料來源為例，故結果恐難以延伸至全國各地區病患統一適用。但就腎炎、腎徵候群及腎性病變等病患，各數據仍有共通之相似度及對應性，因此期許本研究之結果仍能作為其他各醫療院所做為控制病患 Stage 演變之參考依據。

第四節 研究步驟

本研究將探討影響腎炎、腎徵候群及腎性病變等病患各 Stage 演變之相關要項，將現有取得之病患相關數值，作為探勘研究之資料庫，以資料深勘中之決策樹採礦工具，來針對慢性腎臟病患 Stage 變化各因素彼此間相互關係的探討，以期許能找出有明顯影響之項目，再依據取得之變數進行類神經網路模式分析探討出變數於各Stage間之重要性及關連性。本研究依照如下步驟來訂定研究之方向。研究步驟說明如下：

一、 研究動機及目的：確定本研究相關動機、目的，以確認後續資料蒐集及研究之進行方向。

二、 資料蒐集與文獻查看：蒐集本研究相關資料及文獻，藉此尋求適合本研究內容之探勘技術及工具。

三、 訂定研究方法及架構：由上一節所得結果針對本研究之問題及運用技術、工具，訂定研究之方法及架構。

四、 前置處理及分析取得之資料：將所取得之資料先行檢視資料庫中各個欄位之意義及內容價值，並確實掌握各資料之原始意義及原使用特性，最後再行整理、刪除及合併欄位，使各欄位資料適當的轉換成具有分析價值之資料，此一過程可減少不正確之資料，因為不適切之資料欄位極可能影響最終分析之整體結果。

五、 資料探勘結果分析：以運用決策樹及類神經網路資料探勘工具，進行相關資料之分析，取得相關結果進而分析慢性腎臟病患，病程 Stage 間演進之變數相關性的預測模型。

六、 結論：找出顯著及潛在之影響因素，作為慢性腎臟病 Stage 變化之預測模型，以此預測結果協助醫護人員針對各 Stage 病人提供適切之診療、照護及衛教行為。

第五節 論文研究架構

本研究共分為五章：依序為緒論、文獻探討、研究方法、資料分析、結論與建議，各章節主要內容分述如下：

第一章 緒論：本章主要說明研究動機、背景、目的、限制、方法、流程，並說明本論文之整體研究架構。

第二章 文獻探討：就相關文獻介紹慢性腎臟病（腎炎、腎徵候群及腎性病變等）、資料探勘之定義，並進一步針對本研究所使用資料採礦中之決策樹及類神經網路工具加以深入介紹及探討。

第三章 研究方法：本章節針對研究資料來源、研究對象、研究架構以及資料探勘工具作一系列之介紹。

第四章 資料分析：針對資料加以整理後，進行決策樹及類神經網路資料探勘分析。最後對於分析結果加以比較、了解及作一合理解釋及探討。

第五章 結論與建議：總結本研究所達成之貢獻，並說明系統未來可進一步之研究方向。

第二章 文獻探討

本章在於探討慢性腎臟病（腎炎、腎徵候群及腎性病變等）的定義，以及資料探勘之定義、資料探勘運用之範圍，並介紹本研究所採用之資料探勘技術－決策樹、類神經網路，針對以上各點提出相關文獻之探討。

第一節 慢性腎臟病之定義

隨著人們生活物質條件不斷提高、醫療技術之進步，隨之而來的即是人口的老年化，在人口老化的過程中和一些慢性疾病的發展就會有著非常密切的關係，接下來將針對本研究之議題”**慢性腎臟病**”做一概括之介紹。

壹、認識腎臟病

腎臟位於我們的後腰部脊椎兩側，也就是在最後一根肋骨與脊椎相接的夾角地區。外形如蠶豆般，大小大概跟拳頭差不多。腎臟的基本組成單位我們稱之為「腎元」。每個腎臟約由一百萬個腎元所組成，每個腎元則包括了腎絲球及腎小管。當身體中的血液經過腎臟時，腎元就會過濾身體的廢物、水分以及電解質，成為尿液。腎臟除了製造

尿液機能外，另外還能製造紅血球生成素、活化維他命的 D3 以維持血中鈣磷平衡，以及生成腎素及血管張力素來調整血壓。(台灣腎臟醫學會，2006)

腎臟病的種類繁多，較常見的有免疫傷害引起的**腎絲球腎炎**及細菌感染有關的**腎盂腎炎**等，另外**糖尿病、高血壓**及**全身性紅斑性狼瘡**等病人也常併發腎臟病變。慢性腎功能障礙(統稱慢性腎衰竭)是所有腎臟病的共同結果，這個病是一種所謂的「**進行性**」疾病，也就是說一旦診斷確定以後，這個病只會惡化而不可能會痊癒。影響腎功能的危險因子中，藥物佔有極重要角色，其中較主要的是一些**止痛劑及抗生素**。(陳鴻鈞，2001)

腎臟受損超過三個月，導致其結構或功能無法恢復正常，稱為慢性腎臟病。慢性腎臟病分為五個階段，這個過程可能非常長久；也有可能很快地進入了**第五階段—末期腎衰竭**，甚至要**進行透析或換腎的治療**。（台灣腎臟醫學會，2006）

（表 2-1 台灣腎臟醫學會，腎臟指標）

慢性腎臟病的五個階段：			
第一期	腎功能正常但併有蛋白尿、血尿等腎臟損傷狀況	腎絲球濾過率GFR 90~100 ml/min/1.73 m²	腎臟功能約正常人60%以上，注意是否有糖尿病及高血壓，需要控制血糖、血壓與飲食，每半年作腎功能檢查，一般皆能穩住腎功能。但若有腎絲球腎炎之病患必需接受治療。
第二期	輕度慢性腎衰竭，但併有蛋白尿、血尿等	腎絲球濾過率GFR 60~89 ml/min/1.73 m²	
第三期	中度慢性腎衰竭	腎絲球濾過率GFR 30~59 ml/min/1.73 m²	腎臟功能約正常人15~59%，積極配合醫師治療，減緩進入第五期腎臟病變
第四期	重度慢性腎衰竭	腎絲球濾過率GFR 15~29 ml/min/1.73 m²	
第五期	末期腎臟病變	腎絲球濾過率GFR <15 ml/min/1.73 m²	腎臟功能剩正常人15%以下，若逐漸無法排除體內代謝廢物和水分，則必需準備與接受透析治療及腎臟移植

如果慢性腎臟病之 Stage 進入第五期，會出現尿毒症狀:有噁心、嘔吐、呼吸困難、肺水腫、心衰竭，這時病人即需要進一步之治療：血液透析、腹膜透析、腎臟移植（台灣腎臟醫學會，2006）。目前台灣病患大都採用血液透析(約 80%)：此治療須先以外科手術建立動靜脈瘻管，通常在手臂上開口，將手臂的動脈與靜脈以管相接，約 3 到 4 周後，就可到醫院做血液透析治療，治療方式為由該瘻管處將血液抽出約 200CC/每分鐘進入人造半透膜(人工腎臟)，待血液淨化後再由機器將乾淨的血液送回體內。血液透析的患者必須每週至醫療院所接受 2-3 次，每次 4-5 小時的透析。血液透析優點是由專業醫療人員執行治療、治療後體內積存的廢物較少。而每月的血液透析費用每人約為五萬元皆為健保給付，也因此才會出現全國一年洗腎醫療支出高達 338 億元的健保給付費用。

第二節 資料探勘之定義

資料探勘我們可分為兩種角度去定義：

壹、學者定義：

所謂的資料探勘是一大量自動化的過程,其運用統計分析從大量資料庫中挖掘出潛在、非顯然的、未知的、潛在的「可能」有用資訊之過程(Frawley et al.,1991)。而Grupe & Owrang (1995)學者則認為資料探勘是指從已經存在的資料庫當中挖掘出專家仍未知的新事實。Fayyad(1996)則定義知識發掘(knowledge discovery)為從大量資料中選取合適的資料，進行資料處理、轉換等工作，再進行資料探勘與結果評估的一系列過程，也就是說資料探勘只是知識發掘過程當中的一個步驟。定義資料探勘為使用自動或半自動的方法，對大量資料作分析，找出有意義的關係或法則（Berry & Linoff，1997）。Akaka(2004)提出資料探勘是一種應用資料庫的技術，像是統計分析與建立模型，用以發現資料中隱藏的模式與隱約的關係，並進行推論以預測未來結果。AAAI(American Association for Artificial Intelligence，2006)近期指出資料探勘是一種很強大的人工智慧工具，它可以自資料庫中發現有用的資訊，並可用來改善行為。

貳、廣義定義：

資料探勘我們可解釋為資料庫之知識發掘(Knowledge Discovery in Databases，簡稱 KDD）。也就是說可以從一個大型資料庫裡頭所儲存的大量資料當中萃取出有趣知識，這個大型資料庫有可能是線上作業的資料庫，也有可能是資料倉儲。

資料探勘之處理流程步驟：

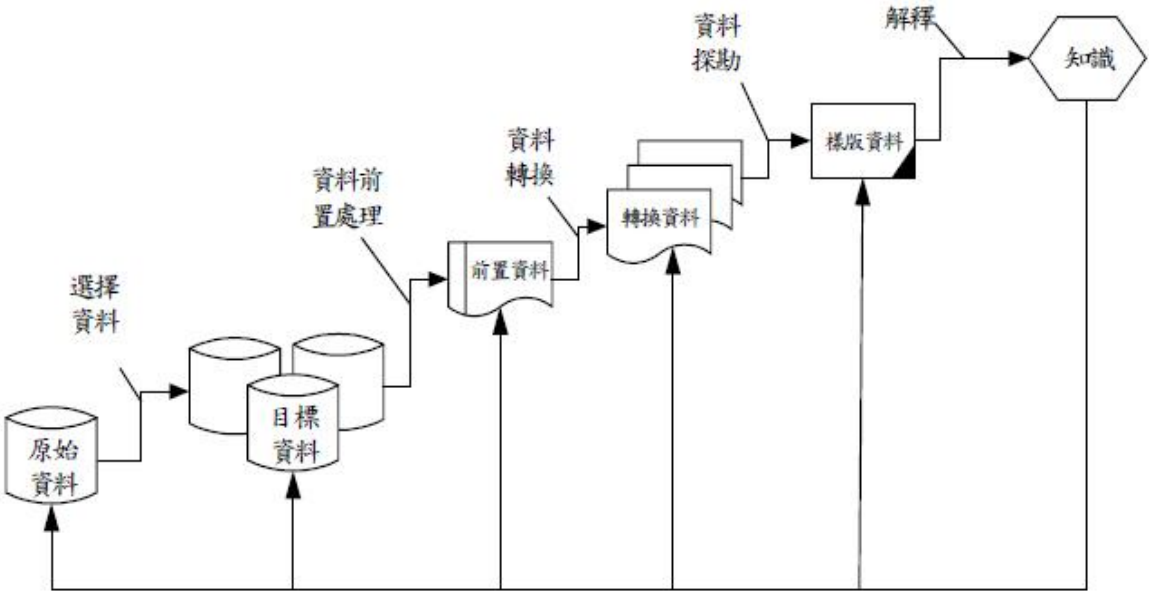


圖 2-1、KDD 之處理流程步驟

(Piatetsky-Shapiroet al.,1991;Fayyad et,al.,1996；姚吉峰，2002)

參、 資料探勘主要功能共分為六大類：

1. 分類(Classification)：按照分析資料已知的事實及其屬性加以定義，來建立類組。在分類問題中，除了提供預測的分類結果之外，亦可提供發生這個分類結果的可能預測機率。使用的技巧有『決策樹』、『記憶基礎推理』等。例如：信用卡申請者之風險性程度分類。
2. 推估(Estimation)：依據目前已有連續性數值之相關屬性資料，以獲致某一屬性未知之值。使用之法有『迴歸分析』及『類神經網路方法』，例如：顧客價值預測。
3. 群集化(Clustering)：面對大量的資訊，我們將相似的事物分群，如此可以使得複雜的資訊變得大幅簡化。群集化於商業上最常見的能用即是市場區隔。例如：顧客分群，以依顧客屬性分類(根據看電影的品味將觀眾分組)。
4. 同質分組(Affinity Group)：將一個異質母體，分隔為一些具相同性質的群體，即是從所有物件決定那些相關物件應該放在一起。例如：型錄的編排方式、貨架的擺置方式；例如大賣場相關之電器用品（電話、傳真機、電話線），放在同一個貨架上。
5. 序列(Sequential)：在同質分組中找出哪些事物會相伴發生，透過序列找出事物『先後』發生的順序，我們有時稱這樣的規則為時序規則。例如：網頁瀏覽序列分析。(尹相志，2007)

6. 描述(Description)：概略性的描述在複雜的資料庫中，到底發生了甚麼?透過這樣的方式可以決定那些相關物件應該放在一起。例如：30-40歲的男性喜歡看那種類型的書籍。

第三節 資料探勘運用範圍

依資料探勘各歷程之演進，目前它的運用範圍幾乎已經無所不在了，不論是從大量歷史資料中找出一定的規律、對於未來結果進行預測，以上皆可以運用資料探勘的技術來達成。如犯罪防治(犯罪行為與特質的相關研究)、命理、行銷、工業、體育、財務、製造廠、電信業、網路相關行業(網頁上瀏覽路徑分析 ，提供網頁或網站瀏覽建議)、零售商、製造業、醫療保健、製藥業等等。如下針對幾項運用範圍作介紹。

壹、 醫療生技業：預防醫學分析、院內感染分析、臨床病徵分析、基因圖譜比對、基因定序、演化分析。

貳、 金融保險業：信用評等、客製化金融服務、客戶資產管理、呆帳分析、保險潛在客戶名單分析、直效行銷、分析購買行為、偵測信用卡詐騙行為、股匯市行情預測。

參、 教育業：學生來源分析、學習成績評量、課程規劃

肆、 零售業：店家設點分析、銷售產品組合、促銷商品組合、

庫存管理、DM 內容。

伍、 航空業者：預測售票但乘客並未搭機，依此即可作為機票超賣之決策依據。

陸、 保險業者：分析理賠模式用以找出浮濫之申請理賠模式，防止詐領保險金之情形發生。

第四節 決策樹

決策樹是一個預測模型；含有根部 root、子節點 child node、葉部節點 leaf node，**根部**：資料從根部的節點進入決策樹；**子節點**：每一個節點代表「**是**」或「**否**」的問題點，答案代表前往下一個問題的前進路徑；**葉部節點**：決策過程一再重複，直到資料到達葉部節點為止。樹中每個節點表示某個對象，而每個分叉路徑則代表的某個可能的屬性值，而每個葉節點則對應從根節點到該葉節點所經歷的路徑所表示的對象的值。決策樹僅有單一輸出，若欲有複數輸出，可以建立獨立的決策樹以處理不同輸出。

決策樹是資料探勘中一個普通的方法。每個決策樹可以依靠資料庫的分割進行數據測試。這個過程可以遞歸式的方式對樹進行修剪。當不能再進行分割或一個單獨的類應用於某一分支時，遞歸過程就完成了。

壹、決策樹有幾種產生方法:

分類樹:是當預計結果可能為兩種類型(例如是否、輸贏等)的概念。

回歸樹:是當結果可能為實數(例如房價,患者住院時間等)的概念。

CART:是結合了上述二者的一個概念。CART 是 Classification And Regression Trees

CHAID:(Chi-Square Automatic Interaction Detector)

決策樹是功能強大且相當受歡迎的分類和預測工具。這項以樹狀圖為基礎的方法,其吸引人之處在於決策樹具有規則,和類神經網路不同。規則可以用文字來表達,讓人類了解,或是轉化為 SQL 之類的資料庫語言,讓落在特定類別的資料紀錄可以被搜尋(許依宸,2009)。

第五節 類神經網路

類神經網路的相關研究與其應用範圍在近年來發展極為迅速，其應用之領域包括工業工程、商業與金融、社會科會及科學技術等。其最大優點除了在於可應用於建構非線性模式外，對於傳統統計方法在建構模式時所要求的許多假設條件亦可予以彌補。類神經網路的原始想法與基本構造皆與神經生物學中的神經元相似如圖 2-2 所示(謝邦昌，2005)。

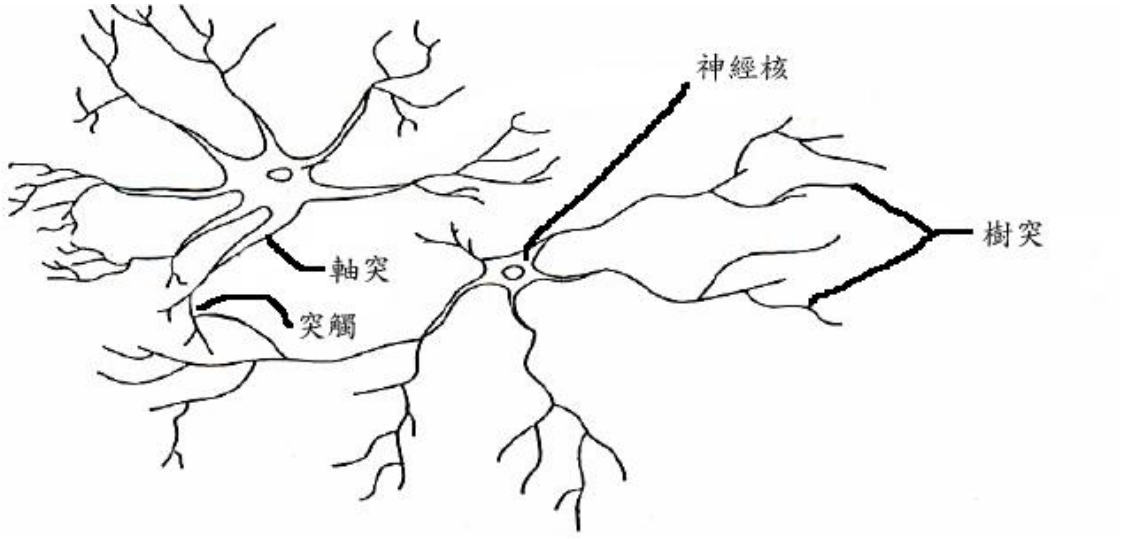


圖 2-2:人類神經元結構

類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 2-3，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 就是外部輸入的神經脈衝，但在通過樹突時，神經脈衝必須大於門檻值，才能夠傳遞至神經元(尹相志，2007)。

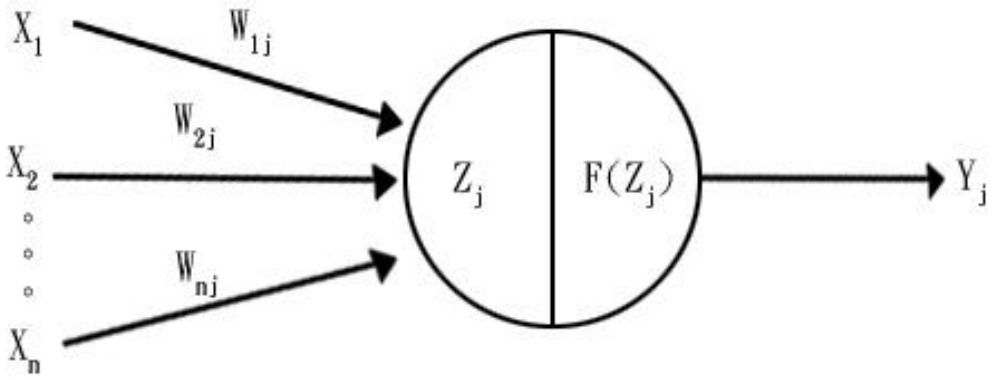


圖 2-3：神經元結構

- Xn：神經元的輸入脈衝。
- Wnj：連結權重值，類神經網路的訓練就是在調整連結權重值，使其變得更大或更小，這部份是由隨機的方式產生介於 +1 至-1 之間的初始值。其值越大，則代表連結的神經元更容易被激發，對類神經網路的影響也更大；反之，則代表對類神經網路並無太大的影響。
- Zj：加法單元，此部分將每一個輸入脈衝與連結權重值相乘後做一加總的動作。
- f(Zj)：為之為轉換函數，通常是非線性函數，有數種不同的型式，其目的是將接收的輸入脈衝總和，轉換成輸出脈衝。
- Yj：為輸出脈衝，亦即我們所需要的結果。

在類神經網路的訓練過程中，根據理論其學習率是一個非常重要的參數，它影響著類神經網路收斂的速度，若學習率較大，則類神經網路收斂的速度將變得較快，相反的較小的學習率會使得類神經網路的收斂速度變慢。太大或太小的學習率對類神經網路的訓練都有不良的影響，類神經網路的學習重點在於如何自動的、有效率的調整權重大小及學習率。

當類神經網路經由訓練組樣本訓練完成後，類神經網路的輸出已經與我們所要求的數值接近，但對於不是由訓練組樣本的輸入，我們並不知道會得到什麼樣的輸出。在此我們另外準備一組測試組樣本輸入到類神經網路中，測試其正確性及測試其結果是否與所要求的值相近。當類神經網路訓練完成後，對於與訓練組樣本相近之輸入，類神經網路就能給予一個合理的輸出，但是如果測試組樣本與訓練組樣本的差異過大，類神經網路仍是無法給予正確的數值。

第三章 研究方法

本章共分五節，第一節：本研究來源概況介紹，以作為本研究之研究基礎；第二節：本研究之研究架構；第三節：說明研究資料處理方式；第四節：針對本研究使用之探勘工具介紹。

第一節 研究來源概況介紹

本研究設計之目的在探討慢性腎臟疾病患者從開始被診斷出相關症狀後，病患 Stage 演變之"**重要影響因素**"，從中再進一步找出可預測之變項。再以相關變項建立一預測模式，以提供醫護人員診療、照護、衛教等醫療時有建設性之參考，以達到延緩各病患之 Stage 的演變。

本研究收集之研究對象之主要資料來源以南部某區域醫院之慢性腎臟個案管理病患資料為研究對象，樣本取樣時間由 96 年 7 月至 99 年 1 月。總收案人數共計 893 位，其中二位(99 年新收個案)因為只有基本資料但無血液生化檢驗值故先將之刪除，保留分析之個案數為 891 筆。96 年 7 月至 12 月收案人數為 429 位；97 年 1 月至 12 月收案人數為 218 位;98 年 1 月至 12 月收案人數為 238 位;99 年 1 月收案人數為 6 位。因為本研究資料來源之醫院於 96 年下半年度開始辦

理慢性腎臟病患之個案管理業務，故可發現 96 年下半年度收案之人數幾乎佔了總收案人數的一半，而後續 97、98 這二年收案人數即維持穩定及遞增。

由於病患收案時之病情並不代表為初期之症狀，故收案病患之慢性腎臟病 Stage 由第 1 期至第 5 期皆有可能是收案當下之症狀 Stage，如上所述不易了解影響病患病症 Stage 變化之原因，故應用資料探勘技術擬尋求病患病症 Stage 變化之潛在影響因素，希望能找出病症 Stage 間變化之各階段較明顯之影響因素，藉此使病患各 Stage 之演變能獲得控制將各時程之間距延長，最終延緩病患進入透析（血液透析、腹膜透析）之時間點。

第二節 研究架構

以南部某區域醫院之慢性腎臟個案管理中心 96 年下半年度起收案之病患個案**基本資料**及**血液生化檢驗資料**二個資料表為基礎。基本資料表包含個案姓名、身份證字號、性別、居住地、連絡電話、收案日期、收案醫院等;血液生化檢驗資料表紀錄了個案不定期回診所做血液生化之數據檢驗值(共有 19 個紀錄欄位)。其中個案性別、居住地區、血液生化檢驗主要預測欄位（與個案管理衛教護理師溝通後取

出數個特定之欄位)及病症 Stage 等將為本研究之主要研究變數。

第三節 資料處理方式

本研究所取得之資料庫初始為衛教師所使用之個案管理系統相關之內容，因此相對而言資料表之建置欄位較多也較複雜。在經過與衛教師討論後初步將一些不具分析價值之欄位刪除，以求進行探勘時不因為多餘之資料欄位而影響到最終結果之準確性。個案基本資料表欄位取病人 ID、性別、地址（轉換成縣市）此三個欄位，其餘如接案醫師、個人電話、學歷、婚姻等欄位因不會影響病症 Stage 之變化故將之刪除。另就血液生化檢驗資料表因個案衛教師所需，故紀錄了大量檢驗資料值，在紀錄中之 19 個欄位依回診情況不同醫師會開立不同的檢驗項目，故有些欄位會出現遺漏值之情形發生此一現象於後面章節再行討論。

經以上二個資料表欄位簡化修正後，再將二資料表做 join 連結的最終處理以產生含有基本資料及血液生化檢驗資料之新資料表。

以上過程概述如下：

1. 資料淨化：依取得之資料表欄位修正、刪除，留下正確可供分析之資料，期待探勘過程減少不必要欄位之值影響探勘之

結果。

2. 資料轉換：為使欄位更符合探勘之格式，將幾個原始欄位轉換，如生日將其轉換為年齡、地址轉換為居住縣市及血液生化資料表內檢驗值原始資料皆為文字格式先將其轉換為數字欄位格式另 Stage 欄位值只取出其數字部份(Stage5->5)。
3. 資料整合：最後將上述二步驟處理、淨化後之資料整合出另一個新的資料表。

第四節 本研究使用之探勘工具介紹

本研究採用 SQL Server2005 為資料採礦工具。在 SQL Server2005 中，只要是跟商業智慧有關的所有開發(包括多維度分析、資料採礦、資料轉換、報表服務)都能統一透過 BIDS(Business Intelligence Development Stuido)專案開發。SQL Server 2005 於系統內提供了決策樹、類神經網路、線性迴歸、羅吉斯迴歸、貝氏機率分類、關聯規則、時序群集、時間序列及群集演算等九種演算法，更透過豐富的視覺化呈現(12 種模型瀏覽器)，能讓分析者更易深入瞭解內容。除此之外，系統更提供了資料模型評估工具：分類矩陣、增益圖、利潤圖、散布圖等。

使用者可以利用此平台設計各種資料結構、定義資料關聯性、設

定索引鍵，再由此產生計算性欄位，還能根據既有欄位產生暫存多維度分析來幫助使用者檢視變數的分布、交叉分析。如上所述使得資訊人員能得到眾多方便又強大的功能，以達到降低系統建置、部署、分析、管理等工作之複雜性。透過 SQL Server 2005 的各種功能企業幾乎能獲得所有完整的解決方案。

SQL Server 2005 可幫助組織執行下列工作：

壹:建置、部署和管理更安全又可靠的企業應用程式。

貳:簡化開發及支援資料庫應用程式的工作,以達到最高的 IT 生產力。

參:跨多個平台、應用程式和裝置共用資料，以連接內、外部系統。

肆:控制成本同時兼顧效能、可用性、延展性和安全性。(許依宸,2009)

第四章 資料分析

第一節說明研究資料特性分析；第二節研究資料分組；第三節建立決策樹模型，針對取得之資料進行決策樹建置；第四節建立類神經網路，依決策樹模式找出之輸入變數，建立類神經網路模式；第五節成效分析，針對演算模型進行分類矩陣與增益圖評估，並找出『輸出變數 Stage』間各別之相關性分析。

第一節 研究資料特性分析

本研究中如前面章節所述將取 96 年度下半年至 99 年 1 月份之資料紀錄共有收案人數 891 筆，血液生化檢驗資料共有 5698 筆資料。依據資料來源分析資料具有明顯之區域性，在此因素影響下使得資料來源之年齡、居住地等項目會有集中性之情形產生。如下將以圖表方式分別來呈現及探討本研究資料各別項目之分佈情形。

由圖 4_1 中個案居住地分佈以雲林縣及嘉義縣這二個縣市最多，其次是南投縣另外彰化縣及台南縣等雲嘉地區附近之病人佔了大多數資料來源。

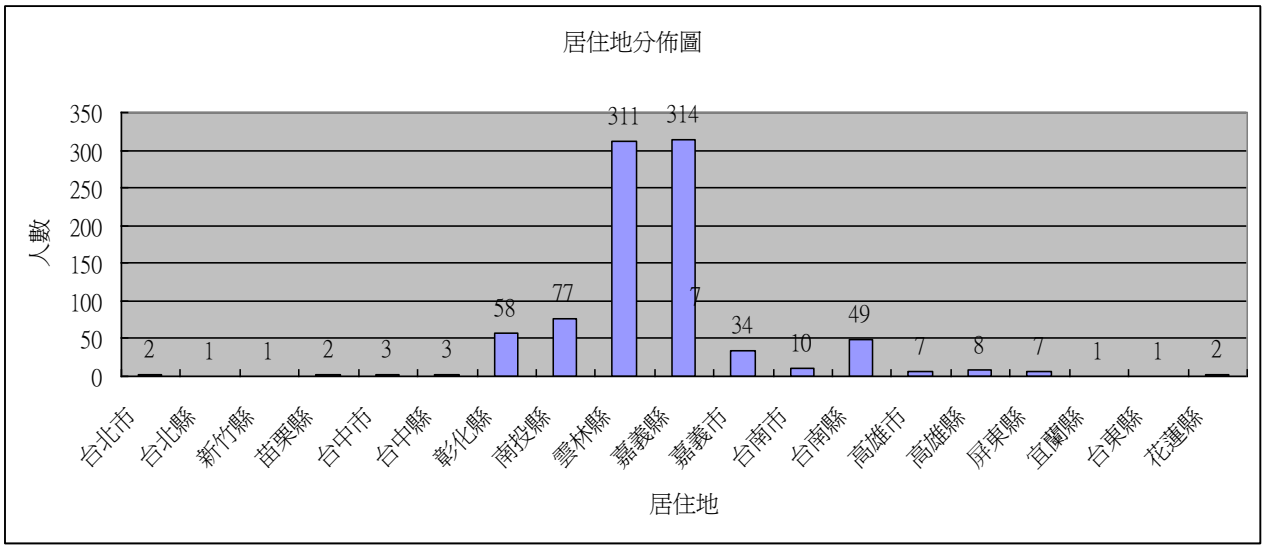


圖 4_1 個案居住地分佈圖

在「圖 4-2 個案性別分佈圖」中可看出，所取得之樣本資料，女生佔 389 位男生佔 502 位。由此發現個案研究樣本性別比例分布並不平均之狀況。

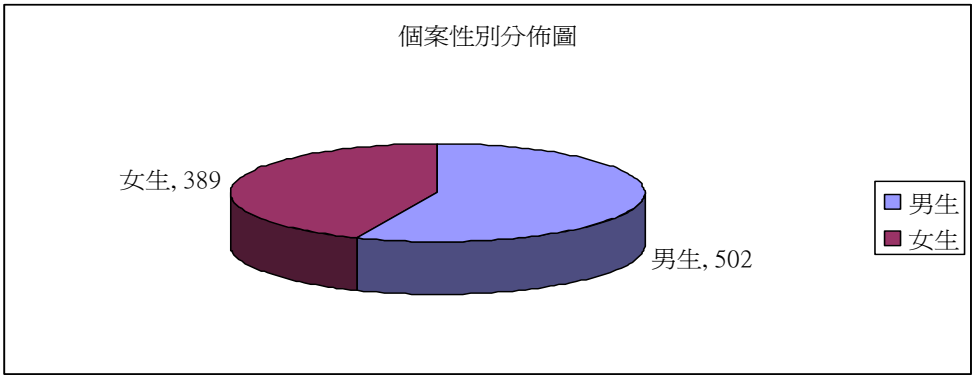


圖 4_2 個案性別分佈圖

從圖 4_3 中可看出各年齡層之病患分布比例，其中 20 歲以下鮮少有收案之病人，相對的年長者之罹病病患佔了決大多數。

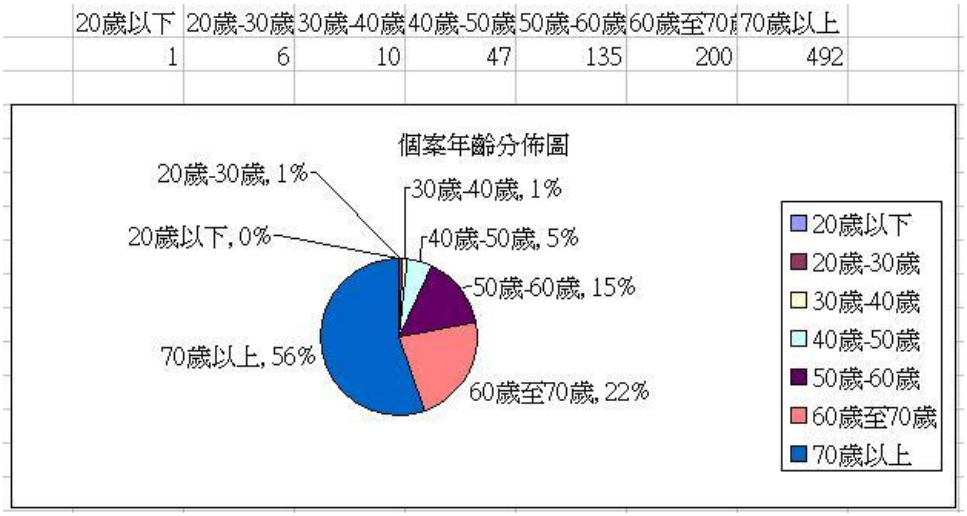


圖 4_3 個案年齡分佈圖

在圖 4_4 中我們可以發現各年度收案 Stage 主要還是晚期收案病患，雖然由 96 年至 98 年可看出晚期收案(Stage4.5)之比例有下降之趨勢，但依整年度為主體統計下 Stage3.4.5 所佔比例還是偏高。

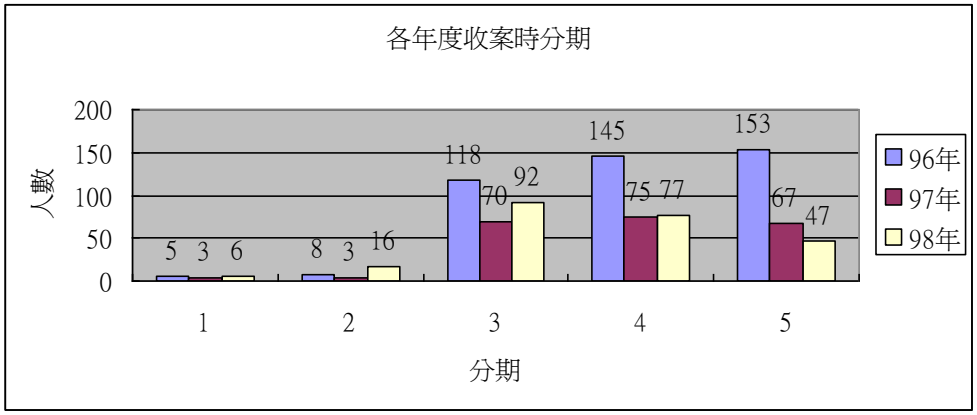


圖 4_4 個案各年度收案時 Stage 分佈圖

另依收案個案之血液生化檢驗資料 5698 筆紀錄資料分析，從中發現了 Stage1、Stage2 這二個 Stage 合計之筆數只佔了總筆數 3.31%，如表 4_1 Stage1、Stage2 筆數之百分比所示。

上述之情形通常會在資料探勘過程中遇到的抽樣問題，稱為『稀有事件』，基本上資料探勘的本質就是透過模型行大量資料、案例中找出有參考價值之『稀有事件』，但大多數的演算法卻會把稀有事件當成了雜訊，進而忽略了這些資訊的重要性。所以在探勘的過程就需要利用『誤差抽樣』的技巧把稀有事件的重要性凸顯出來。

誤差抽樣，可以分為二種模式，一種是將少的抽多(Boost)，另一種則是多的抽少(Reduce)，通常在資料探勘專案中使用『多的抽少』，主要原因在於『少的抽多』是透過抽出後放回再抽的模式，讓案例反覆被選取，這模式比較適合用於資料量過少的時候(尹相志，2007)。

而就本研究之 Stage1、Stage2 資料所佔極低之百分比，如果我們採用『少的抽多』規則做資料重新分配之過程中容易產生『偽規則』現象，另與臨床醫護人員討論後，發現前二個 Stage 之病患往往較難即時發現即時收案(主因如第一章第一節所提前期無特別明顯之症狀)。所以在本研究將 Stage1、Stage2 二階段之資料獨立取出做統計

數據概述，而後期 Stage3、Stage4、Stage5 再以探勘工具做分析，找出 Stage 間變數的相關性。

表 4_1 Stage1、Stage2 筆數之百分比

Stage	筆數	百分比
1	43	0.76%
2	145	2.55%
3	1272	22.36%
4	1461	25.68%
5	2768	48.66%

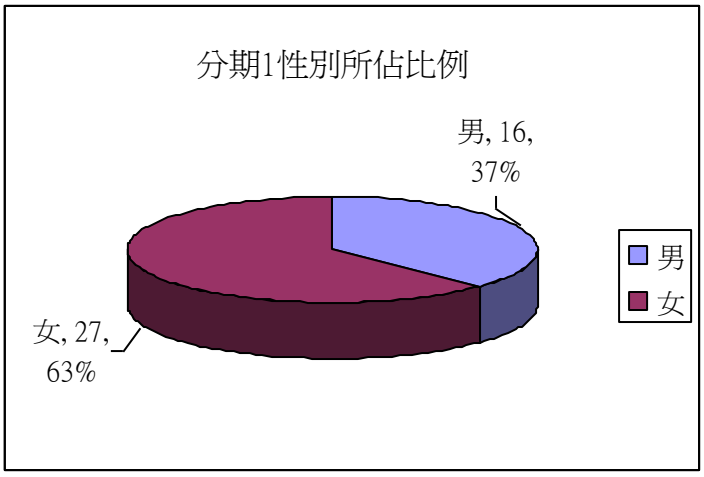


圖 4_5 Stage1 性別所佔比例

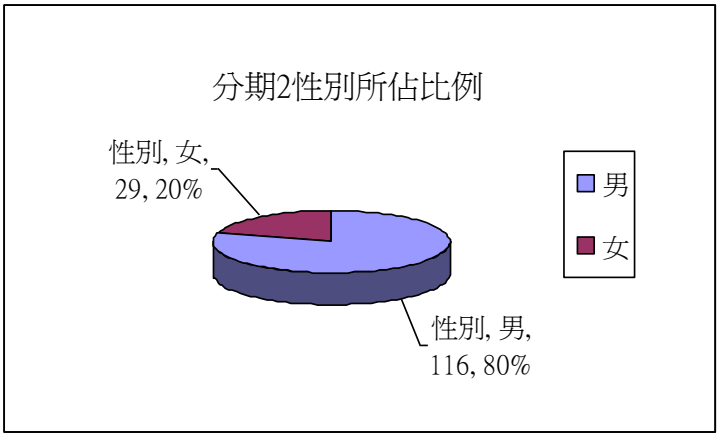


圖 4_6 Stage2 性別所佔比例

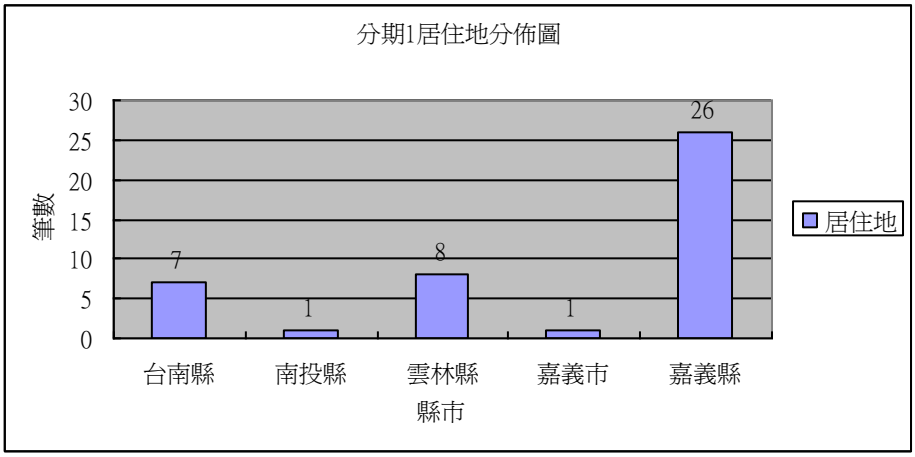


圖 4_7 Stage1 居住地分佈圖

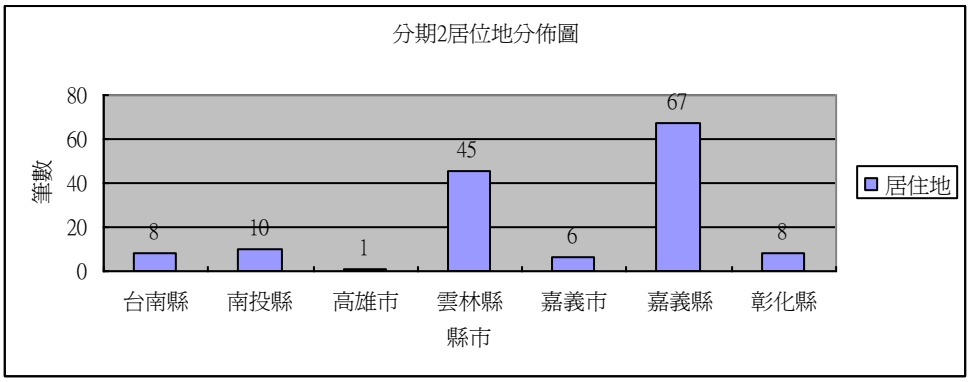


圖 4_8 Stage2 居住地分佈圖

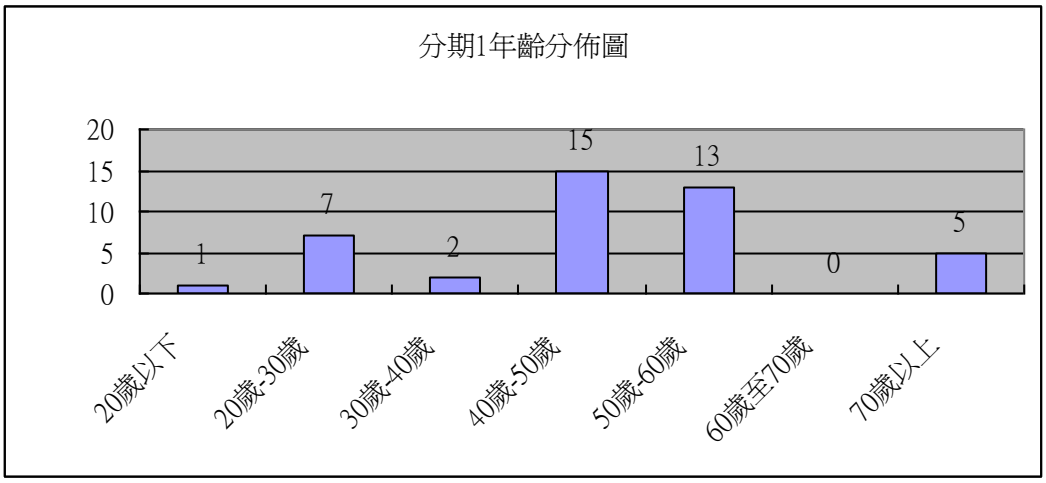


圖 4_9 Stage1 年齡分佈圖

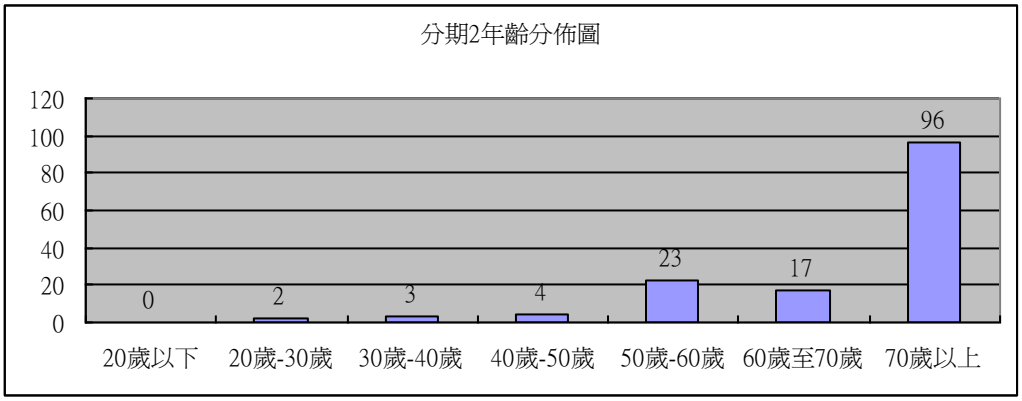


圖 4_10 Stage2 年齡分佈圖

由圖 4_5 至圖 4_10 中可看出因為受限個案筆數少，首先除了居住地受到資料來源地域性之關係，Stage1 及 Stage2 個案病患主要分佈在嘉義縣及雲林縣二個縣市；另外個案病患年齡層在 Stage1 之中主要集中在 40-60 歲間，而 Stage2 的個案病患絕大多數個案是 70 歲以上的年長者，就年齡層的分佈同樣可看出有受到地域性老年化之因素影響。

第二節 研究資料分組

將本研究資料樣本排除 Stage1、Stage2 後，利用 SQL Server 2005 Integration Services (SSIS)依 70%及 30%分為訓練組及測試組二組資料。分組流程如下

Step1:新增專案

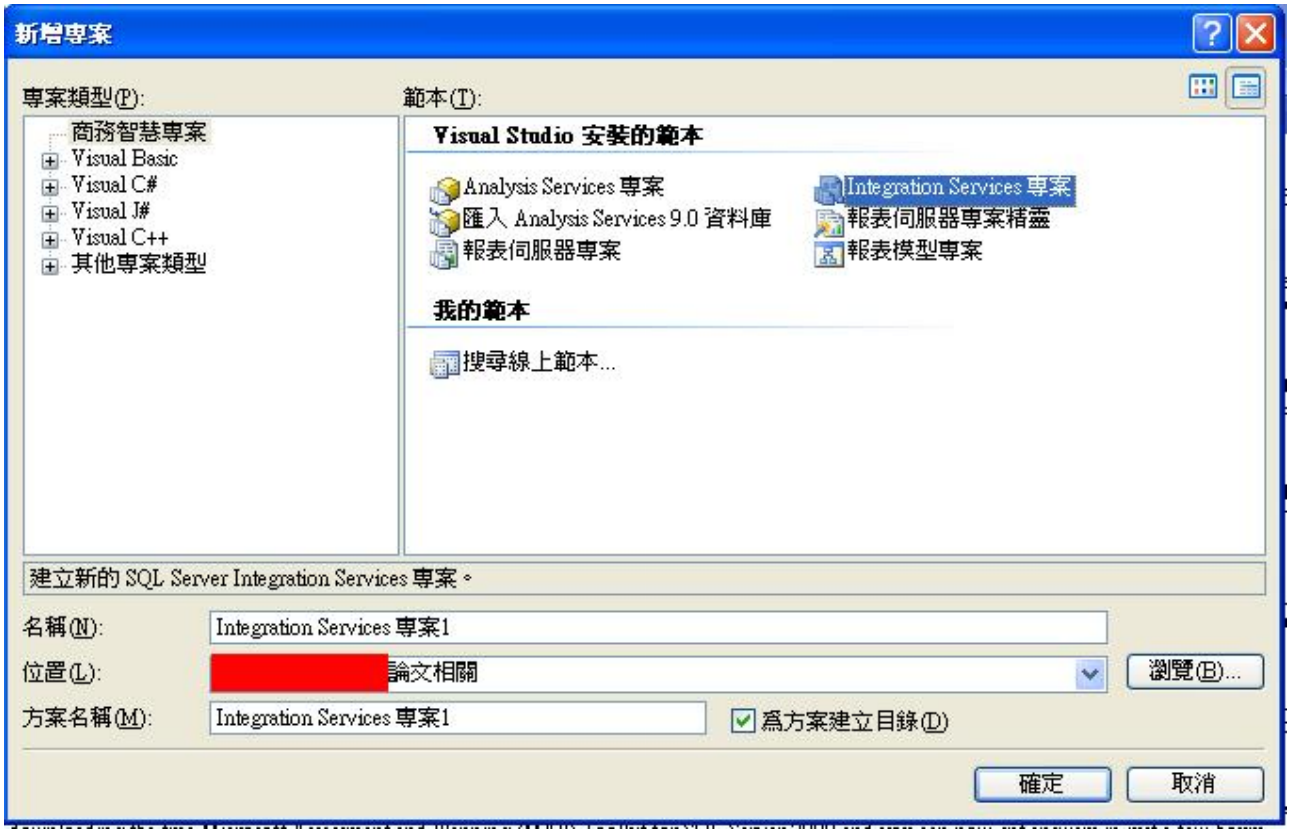


圖 4_11 新增一個 Integration Services 專案

Step2:資料來源、資料檢視設定

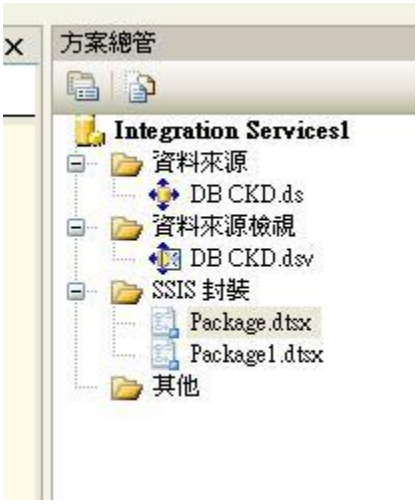


圖 4_12 將資料來源、檢視先設定好

Step3:新增資料流程工作控制項

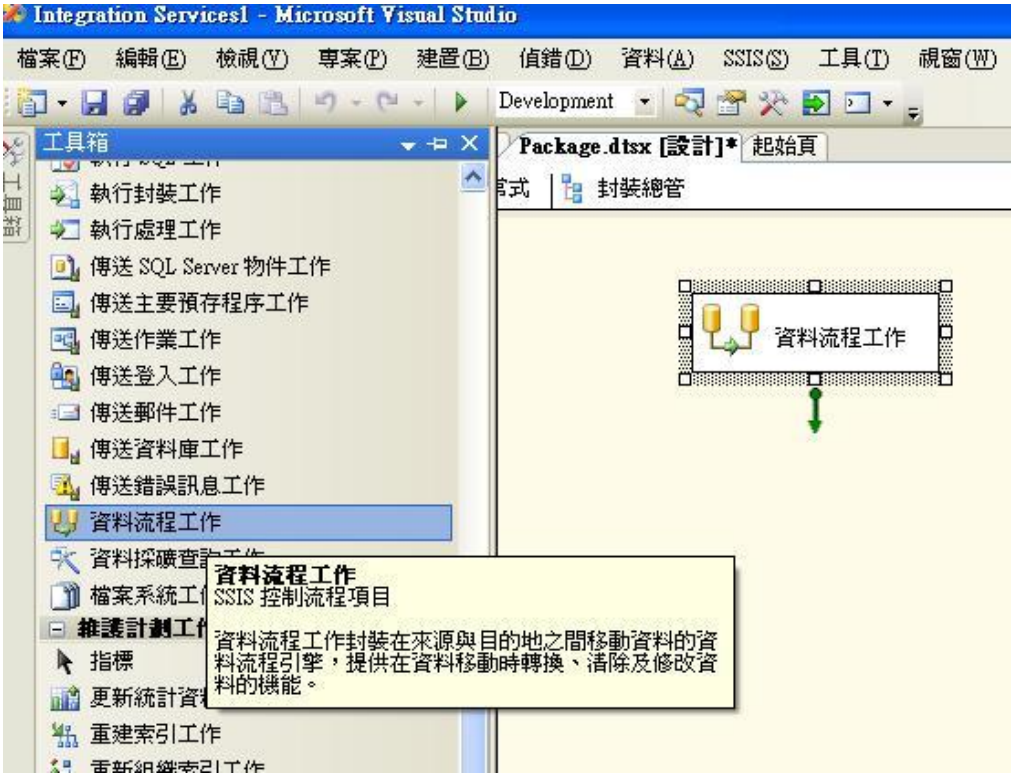


圖 4_13 於控制流程畫面由工具箱拖拉出資料流程工作控制項

Step4：資料流程畫面將各控制項佈置完成

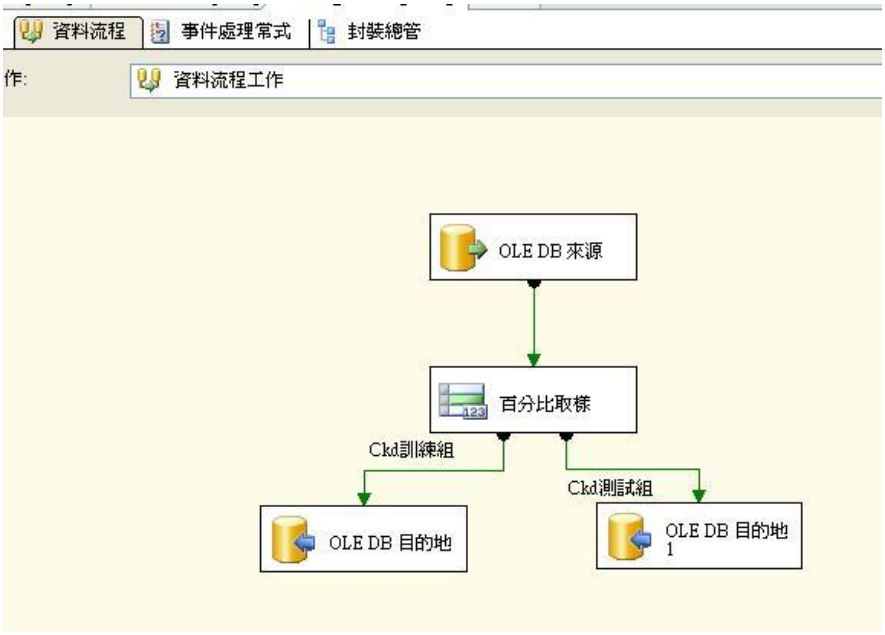


圖 4_14 切換至資料流程畫面將各控制項佈置完成

Step5：百分比條件設定(總資料 70%)

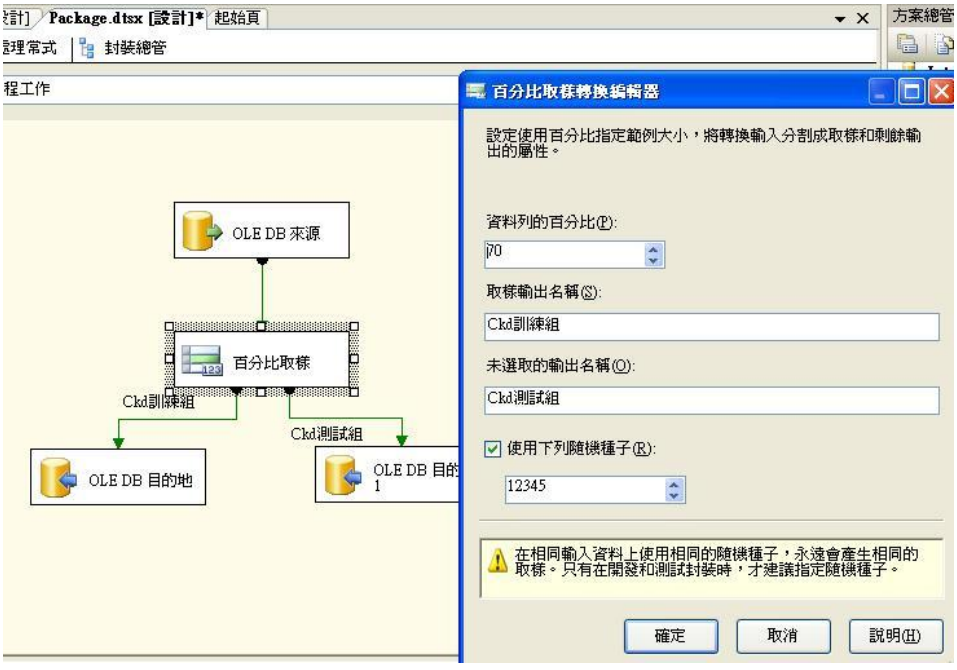


圖 4_15 設定我們所需要的訓練組百分比(總資料 70%)

Step6：執行封裝完成分組

圖 4_16 設定要回存資料之目的地，最後執行封裝即完成分組。

第三節 建立決策樹模型

透過資料前處理，將所得到的資料進行整理後，挑選出進行決策樹分析所需要的輸入欄位和預測欄位，詳細的欄位說明如「表 4-2 決策樹分析輸入欄位、預測欄位」所示：

表 4-2 決策樹分析輸入欄位、預測欄位

欄位名稱	欄位說明	資料類型	內容類型
自動編號	Key 值	Long	Key
Albumin	血液生化檢驗中之「白蛋白」值	Double	Continuous
BUN	血液生化檢驗中之「尿素氮」值	Long	Discrete
Ca	血液生化檢驗中之「鈣」值	Double	Continuous
chSex	病患性別	Text	Discrete
Cholesterol	血液生化檢驗中之「總膽固醇」值	Long	Discrete
Hb	血液生化檢驗中之「血紅素」值	Double	Continuous

HbA1c	血液生化檢驗中之「糖化血紅蛋白」值	Double	Continuous
Hct	血液生化檢驗中之「血比容」值	Double	Continuous
K	血液生化檢驗中之「鉀」值	Double	Continuous
Na	血液生化檢驗中之「鈉」值	Long	Discrete
P	血液生化檢驗中之「磷」值	Double	Continuous
Urin Creatinine	血液生化檢驗中之「尿液肌酸肝」值	Double	Continuous
Stage	預測欄位:病人病情 Stage1-5	Text	Discrete

我們將前一節所取得之訓練組資料進行建立決策樹模型。其步驟如下：

一、 於 MS SQL Server Management Studio 中先新增一個專案，再選取資料來源，按右鍵選擇「新增資料來源」，然後選擇前面已處理完成分組資料之資料庫。

二、 接著在專案中選擇資料來源檢視，將資料來源設定為步驟一新增之資料庫，並將訓練組及測試組之資料表加入檢視中，最後於採礦結構中新增一決策樹模型。

三、 依分組後之訓練組資料建置決策樹模型，在此過程中將演算法參數做調整並建立三組模型，參數調整如下

模型 1. COMPLEXITY_PENALTY=0.5，MINIMUM_SUPPORT=5
模型 2. COMPLEXITY_PENALTY=0.5，MINIMUM_SUPPORT=10
模型 3. COMPLEXITY_PENALTY=0.1，MINIMUM_SUPPORT=10

其中 COMPLEXITY_PENALTY 表示複雜性懲處其值愈接近 1，則決策樹的成長就受到較多的抑制，而產生分岔較少樹狀規則；

MINIMUM_SUPPORT 則代表每個規則節點所需最小案例數此值大於

1 代表最小案例數目為指定的絕對數目，模型建立後經比對三組模型之精確率、反查及正確率後留下較優之模型 Ckd E，如表 4_3 所示，另其模型圖示如圖 4_17 所示。

表 4_3 三模型之正確率、反查及精確率比較表

評估方式 / 模型	正確率Accuracy	反查Recall	精確率Precision
模型1	59.97%	52.62%	52.26%
模型2	58.18%	51.54%	50.08%
模型3	54.50%	49.48%	48.68%

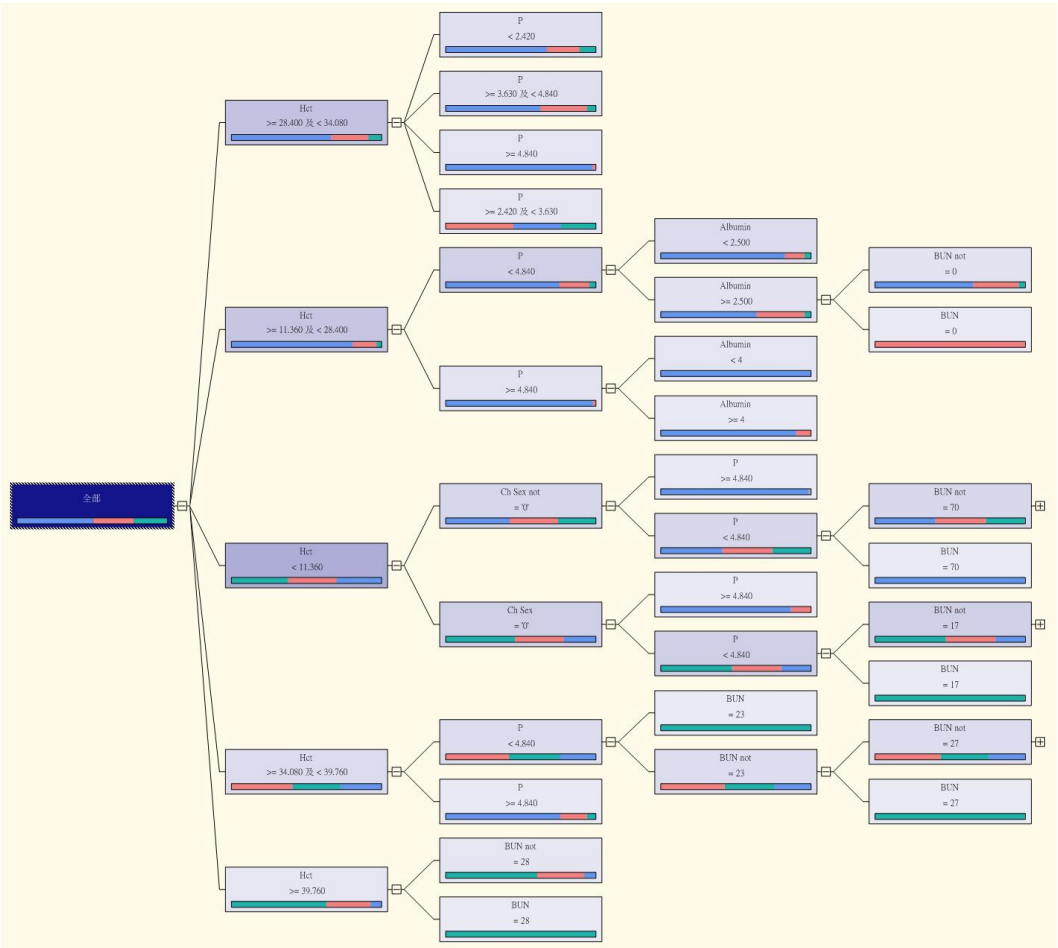


圖4_17 訓練組資料，COMPLEXITY_PENALTY=0.5，MINIMUM_SUPPORT=5 之決策樹模型

在決策樹模型中可找出各變數與預測變數之關聯性，再從模型的

相依性網路了解各輸入變數與預測變數之關聯性。根據每一個箭頭連接之強弱來了解變數之間預測關聯性之強度。此強弱是依據樹的層級數來決定其強弱，即被放在決策樹最上層之變數就是預測能力最強的變數。

經相依性網路所示可得與輸出變數 Stage 關連性較強的變數為 Hct、P、chSex、BUN、Albumin。其中又以 Hct 變數的關連性最強如圖 4_18 所示，再者依序為 P 如圖 4_19、chSex 如圖 4_20、BUN 如圖 4_21 及、Albumin 如圖 4_22 所示。

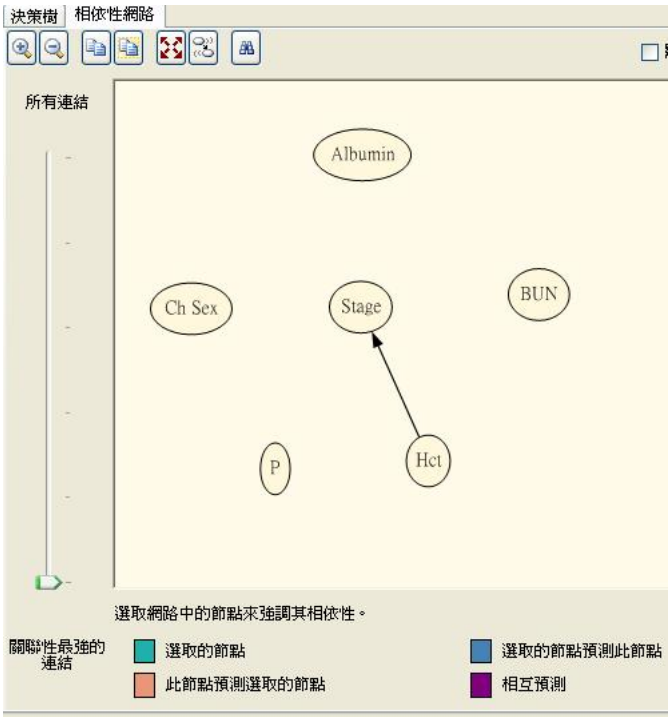


圖4_18決策樹模型之變數**Hct**相依性網路圖示

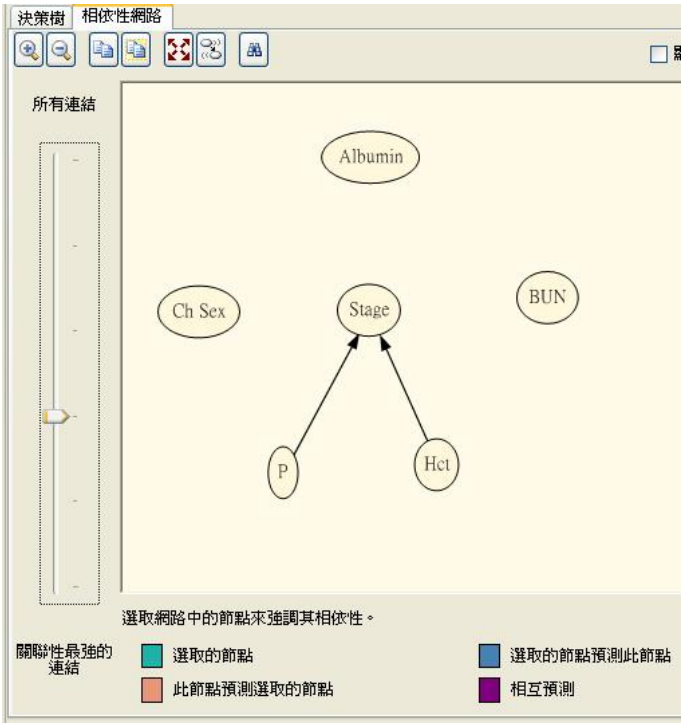


圖4_19決策樹模型之變數P相依性網路圖示

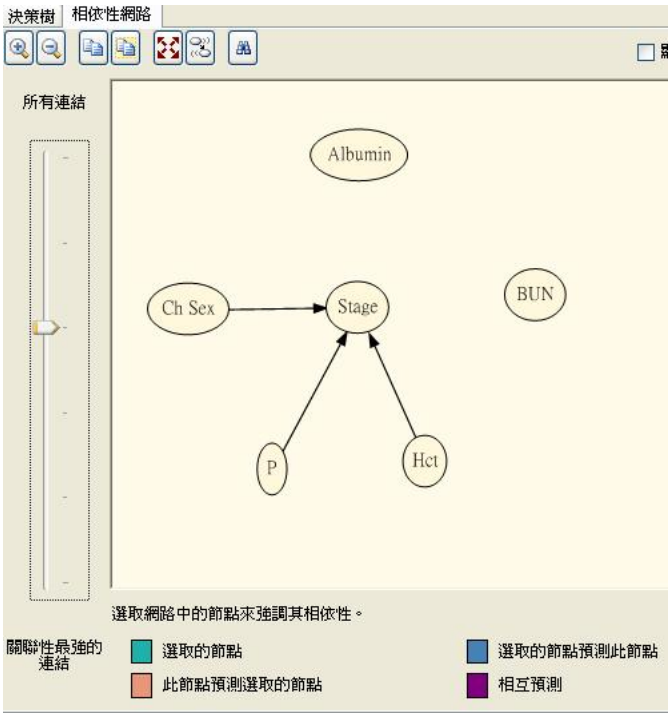


圖4_20決策樹模型之變數**chSex**相依性網路圖示

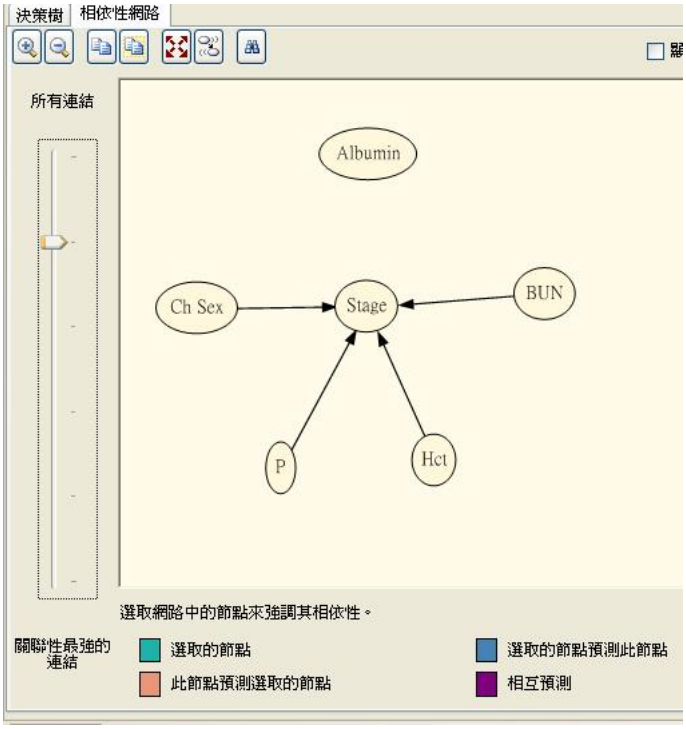


圖4_21決策樹模型之變數BUN相依性網路圖示

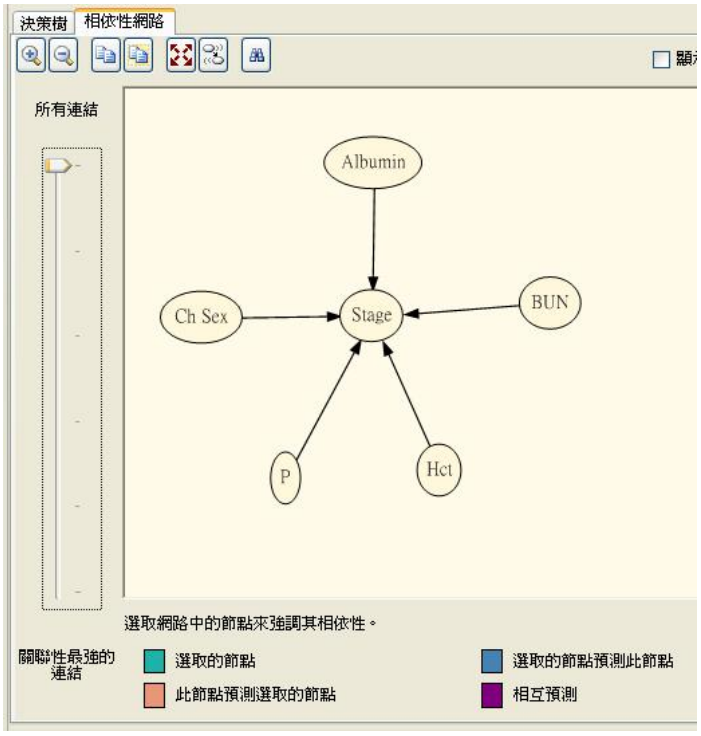


圖4_22決策樹模型之變數**Albumin**相依性網路圖示

第四節 建立類神網路模型

經上一節流程於決策樹模型建置完成後，依決策樹模型所設定之條件建立一類神經網路模型，主因類神經網路本身「**並不具有變數篩選之功能**」，故本研究先行建立決策樹演算法模型再依決策樹篩選出之關聯變數建置類神經網路模型 Ckd N，如『圖 4_23 依決策樹模式變數建立類神經網路』。

結構	Ckd E	Ckd_N
	Microsoft_Decision_Trees	Microsoft_Neural_Network
Albumin	Input	Input
Bno	Key	Key
BUN	Input	Input
Ca	Input	忽略
Ch Sex	Input	Input
Cholesterol	Input	忽略
Hb	Input	忽略
Hb A1c	Input	忽略
Hct	Input	Input
K	Input	忽略
Na	Input	忽略
P	Input	Input
Stage	PredictOnly	PredictOnly
Urin Creatinine	Input	忽略

圖4_23 依決策樹模式變數類神經網路

第五節 成效評估

類神經網路模型建置完成後進行分類矩陣與增益圖之模型評估。

分類矩陣：以分類矩陣評估效能，主要用來檢視錯誤的分佈狀態，即模型建立後以測試組資料進行測試，驗證模型之預測結果的分佈狀態；除了正確率(Accuracy)外，一般建議的評估方式另有精確率(Precision)及反查 (Recall)。

正確率(Accuracy):一般無法真正了解預測隱含的另一層意，只是直覺能知道一個大概的結果。而關鍵性用來補償系統的測量方式即須要下列二項指標。

精確率(Precision):精確率指的是全部資料有多少筆正確資料是正確分類的比例。

反查(Recall):所有預測出來的資料佔總體資料的比例。

表 4_4 驗証範例

實際結果 / 預測結果	是	否
是	10(A)	5(C)
否	6(B)	4(D)

如表 4_4 驗証範例所示，每一個結果類別一般成效評分方式對應為四種狀況，系統正確預測分為該類的有 10 筆；系統没有預測至該類的有 6 筆；實際結果不屬於該類資料但被系統分至該類的有 5 筆；實際結果不屬於該類的資料系統亦未分至該類的有 4 筆。而只要針對研究的各類別都做以上之統計後，就可將正確率(Accuracy)、精確率(Precision)、反查(Recall)等值計算出，其運算式如下：

正確率:(10+4)/(10+5+6+4)=56%，正確率越高理論上是較好的，但此一部份如果 D 值過高，反而造成不管有没有正確分類率，其「正確率」都會越趨近於 1 此一不合理現像。

精確率: 10/(10+5)=66.66%

反　查: 10/(10+6)=62.5%

增益圖：增益圖可說是最常使用的一種資料探勘評估模型，它的橫軸跟縱軸是由百分比所構成。橫軸百分比代表探勘模型根據機率由高至低排序後的名單佔總體百分比。縱軸則是整批資料的準確率。在圖中可看到一條 45 度之斜線，它代表的是隨機的狀態，以增益圖的定義來說，代表隨機篩選整體一半的名單去檢視預測結果，若增益圖

曲線越向上彎曲，即表示模型效果越好。在理想模型下，如果橫軸是全體 50%的案例,此時的正確分類比率應該為 50%;如橫軸是全體 100%的案例，此時的正確分類比率應該為 100%。因此我們可以發現此時理想模型變成 45 度直線。而其他探勘模型斜率必須低於 45 度斜線，圖形越接近理想模型 45 度斜線表示效率越好。(王宗屏,2009)

壹、倒傳遞網路驗証

驗證資料採礦模型不能直接用訓練組資料，因為這個模型是透過訓練組資料計出來的，所以使用訓練組資料驗證準確並不代表全體資料都會準確，因此必須使用先前所準備之測試組資料進行測試。另一方面，測試組資料的分佈比率需接近於真實狀態下的分佈，如此驗證的結果才會接近真實（王宗屏，2009）。從表 4_4 也可看出測試組資料與訓練組透過 SSIS 分組後各 Stage 之分佈的百分比是一致的。

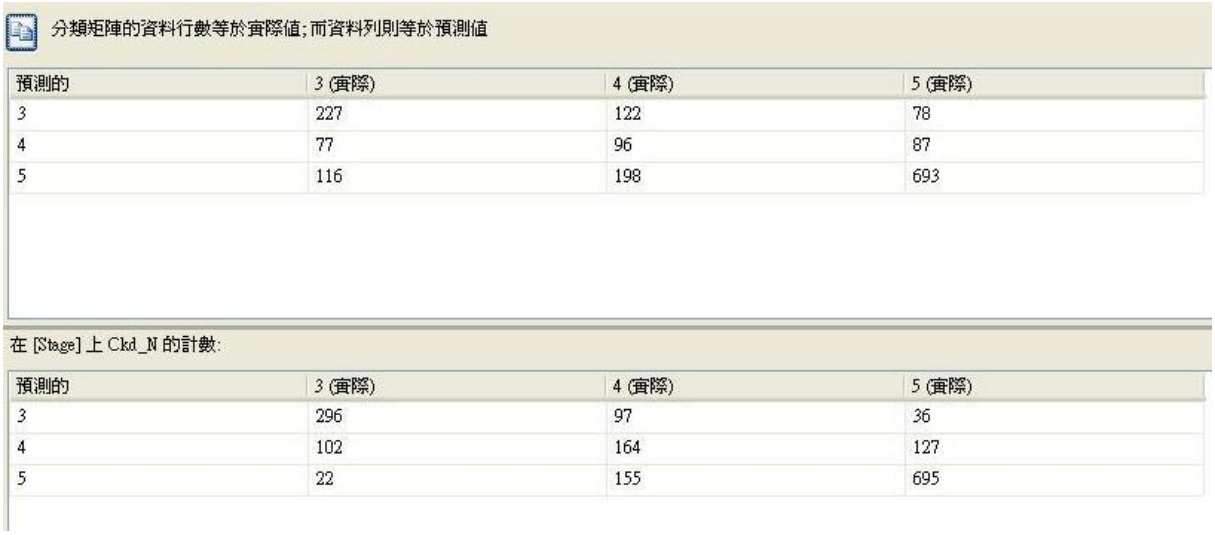
訓練組				測試組		
Stage	筆數			Stage	筆數	
3	852	22.38%		3	420	22.38%
4	1045	27.45%		4	416	27.45%
5	1910	50.17%		5	858	50.17%

表4-5 訓練樣本與測試樣本結構表

建立分類預測模型，將畫面切換至「模型精確度圖表」分頁，接著選擇「選取輸入資料表」方塊點選「選取案例資料表」，此時選擇加入之前準備好的測試組資料，因為測試組資料的欄位名稱與訓練組欄位名稱一樣時所以自動產生連結，如果沒有一樣時我們還可以自行透過選取對應欄位方式建立連結，如圖 4-24 所示。

圖4_24 建立分類預測模型

設定完之後，切換至「分類矩陣」分頁，可得到分類矩陣之各類值對照表，如圖 4-25 所示



分類矩陣的資料行數等於實際值;而資料列則等於預測值

預測的	3 (實際)	4 (實際)	5 (實際)
3	227	122	78
4	77	96	87
5	116	198	693

在 [Stage] 上 Ckd_N 的計數:

預測的	3 (實際)	4 (實際)	5 (實際)
3	296	97	36
4	102	164	127
5	22	155	695

圖4_25 分類矩陣之各類值

貳、實驗結果及效能評估-分類矩陣

表 4_5、表 4_6 分別為決策樹模型 CKd_E、類神經網路模型 CKd_N 之測試組實驗結果，表中直向欄位值代表實際值；橫向欄位值代表模型之預測值。以表 4_5 之 Stage3 為例，表中實際 Stage3 的測試組資料筆數共有 420 筆(=227+77+116)，其中實際為 Stage3 預測結果也 Stage3 的有 227 筆。而實際為 Stage3 預測為 Stage4 的有 77 筆、預測為 Stage5 的有 116 筆。整體上來看，表中由左至右的對角線之值即代表模型預測出之結果與實際結果相同。

Stage	3(實際)	4(實際)	5(實際)
3(預測)	227	122	78
4(預測)	77	96	87
5(預測)	116	198	693

表 4_6 CKd_E 決策樹模型效能評估

Stage3 Recall：227/420=54.04%　Precision：227/429=52.91%

Stage4 Recall：96/416=23.07%　Precision：96/393=24.42%

Stage5 Recall：693/858=80.76%　Precision：693/872=79.47%

平均 Recall：52.62%　平均 Precision：52.26%

正確率 Accuracy：1016/1694=59.97%

Stage	3(實際)	4(實際)	5(實際)
3(預測)	296	97	36
4(預測)	102	164	127
5(預測)	22	155	695

表 4_7 CKd_N 類神經網路模型效能評估

Stage3 Recall：296/420=70.47%　Precision：296/429=68.99%

Stage4 Recall：164/416=39.42%　Precision：164/393=41.73%

Stage5 Recall：695/858=81.00%　Precision：695/872=79.70%

平均 Recall：63.63%　平均 Precision：63.47%

正確率 Accuracy：1155/1694=68.18%

評估方式 / 模型	正確率Accuracy	反查Recall	精確率Precision
Ckd E	59.97%	52.62%	52.26%
Ckd N	68.18%	63.63%	63.47%

表4_8 實驗效果比較表

參、實驗結果及效能評估-增益圖

圖 4-26 採礦精確度設定預測值之數值；圖 4-27 預測值 Stage=3 之增益圖，圖中紫色曲線代表 Ckd N 模型，綠色曲線代表 Ckd E 模型；藍色斜線則代表隨機猜測模型；紅色線為理想模型。

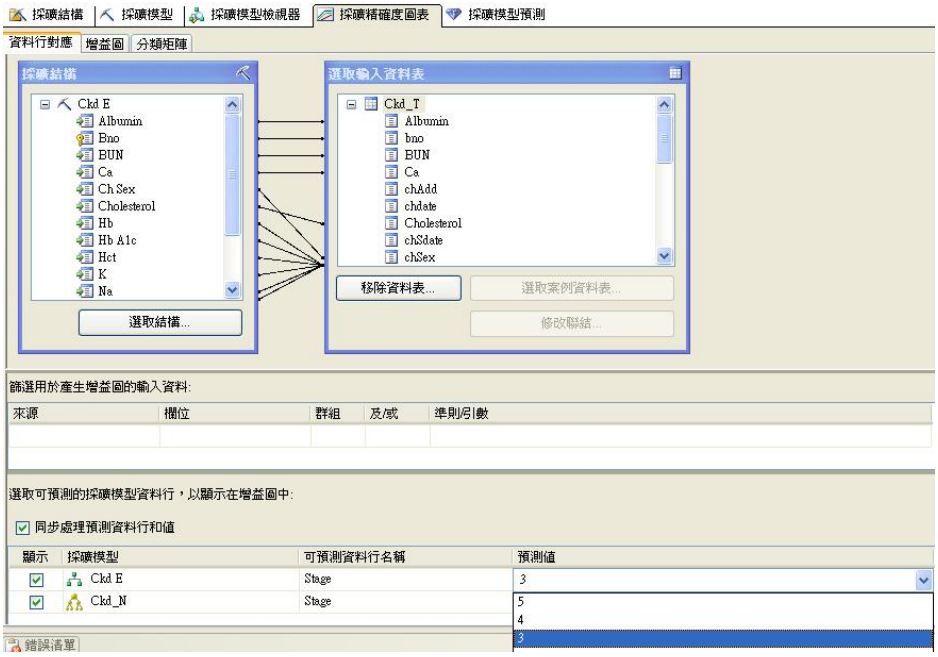


圖4_26 採礦精確度設定預測值之值

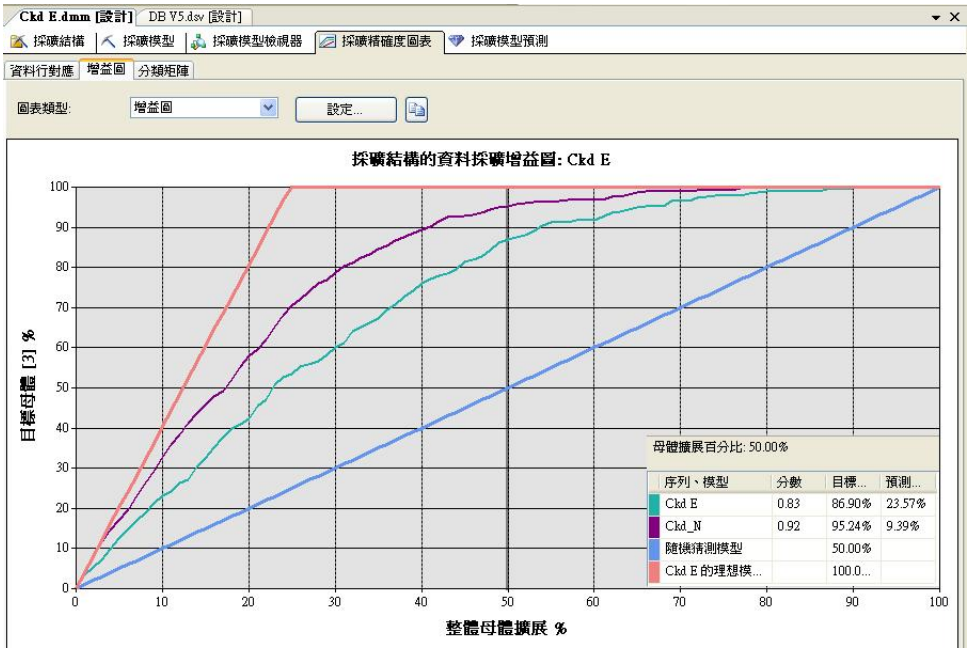


圖4_27 預測值Stage=3之增益圖

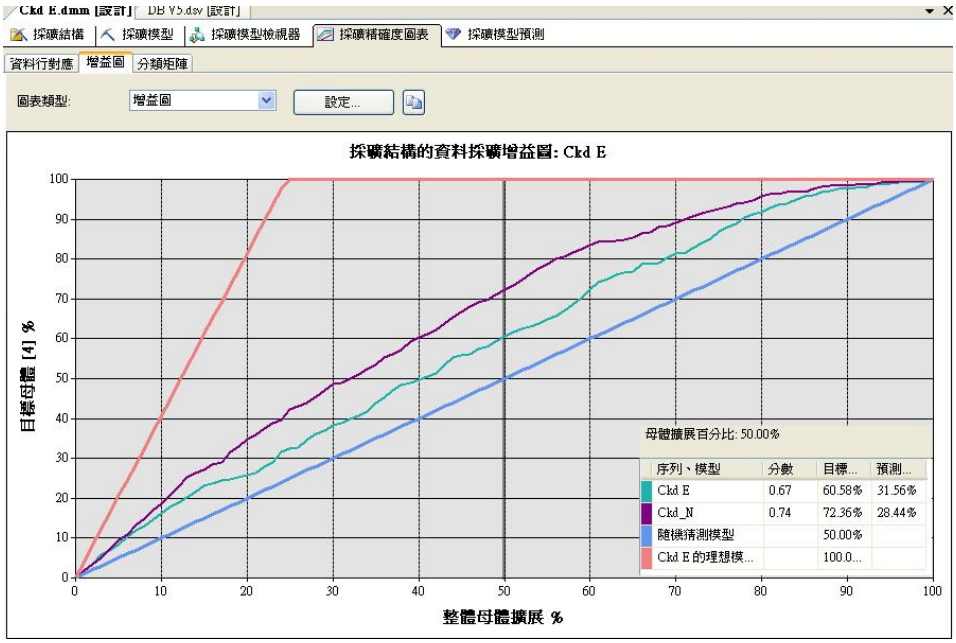


圖4_28 預測值Stage=4之增益圖

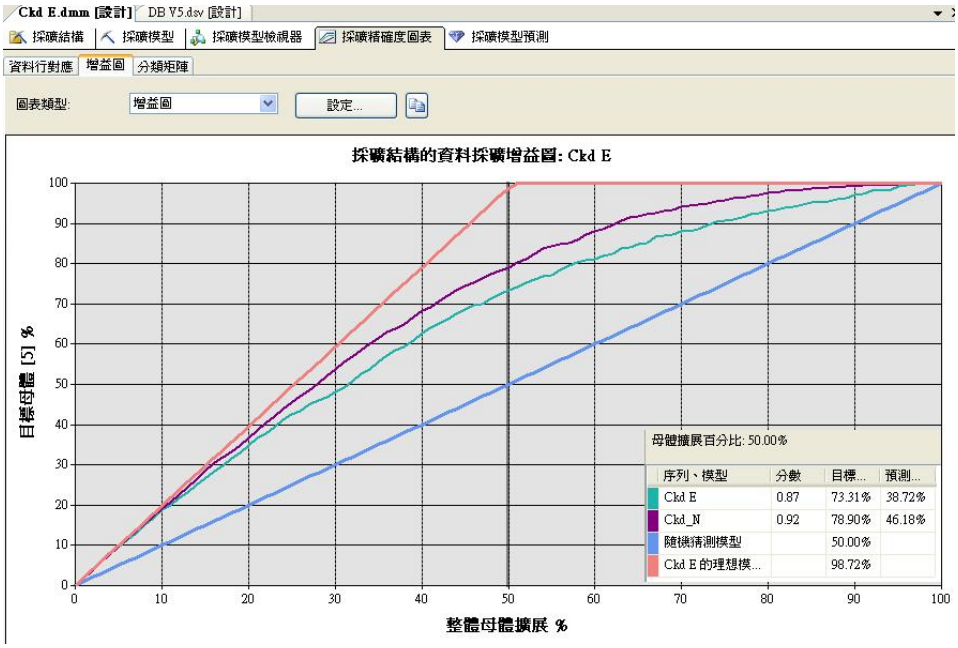


圖4_29 預測值Stage=5之增益圖

由圖4-27、圖4-28及圖4-29比較下發現，不論Stage預測值設定為何圖示之精確度皆介於理相模型及隨機猜測模型之間。而當Stage預測值=5之圖精確度優於預測值=3；預測值=3之圖精確度又優於預測值=4。

肆、效能比較與分析

圖 4-27、圖 4-28 及圖 4-29 皆為以增益圖方式評估 Ckd N 與 Ckd E 之精確度。我們可以發現模型中之紫色曲線比綠色曲線在任一張圖中都更接近理想模型之紅色線，這代表 Ckd N 模型之效能比 Ckd E 模型之效能較好。

從分類矩陣可獲得評估效能之三種資訊，即 Accuracy、Recall 與 Precision。整體精確度整理如表 4-6 實驗結果效能比較表。由表 4-6 上所列出的數據我們可以看出 Ckd N 模型(68.18%)之效能比 Ckd E 模型(59.97%)之效能較好。

另外由分類矩陣及增益圖中觀察可發現二者 Stage4 預測值的精確率及正確率相對於 Stage3、Stage5 明顯偏低。我們試著針對各變數彼此間的分佈性與 Stage 間之關係，進行統計分析並建置如下各圖示，從中進一步分析判斷 Stage3、4、5 之重疊性，以探討 Stage4 預測值偏低是否受重疊性高低之影響。

	chSex	BUN	Albumin	P	Hct
SQL 欄位型態	varchar(1)	varchar(4)	varchar(4)	varchar(4)	varchar(4)
模型內容類型	Discrete	Discrete	Continuous	Continuous	Continuous
模型資料類型	TEXT	Long	Double	Double	Double
資料範圍	0 或 1	10-637	1.4-5.1	0.2-12.2	9.8-56.9

表 4_9 各變數資料型態、資料範圍

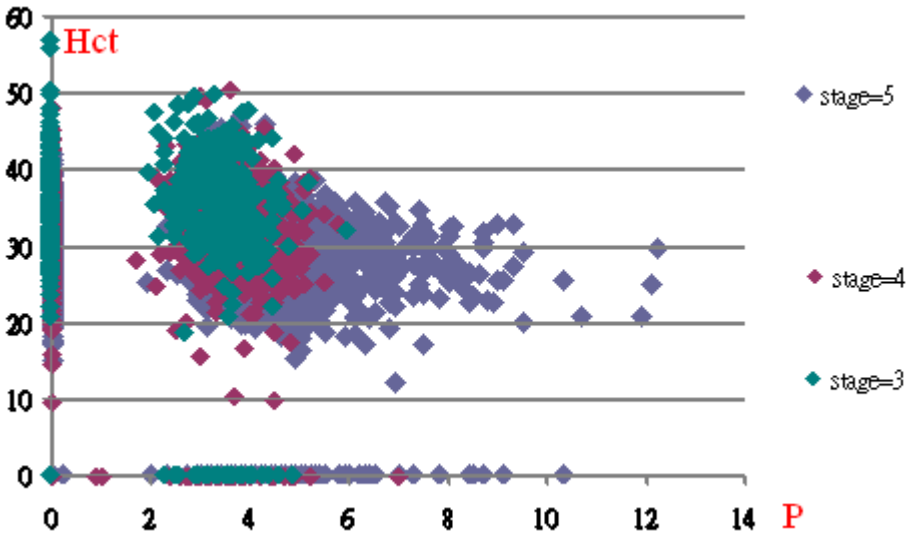


圖4_30 P & Hct 變數分佈及Stage重疊性分析

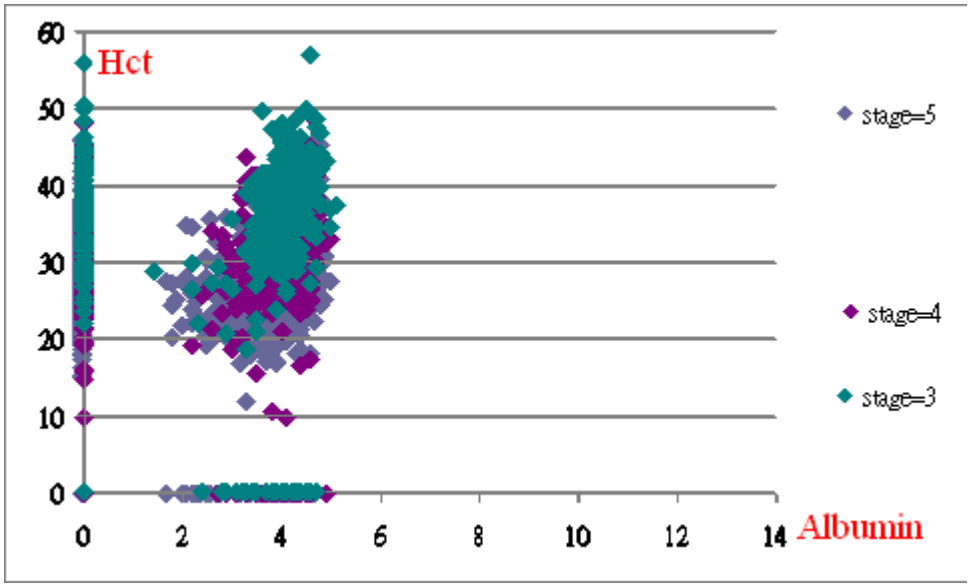


圖4_31 Albumin & Hct 變數分佈及Stage重疊性分析

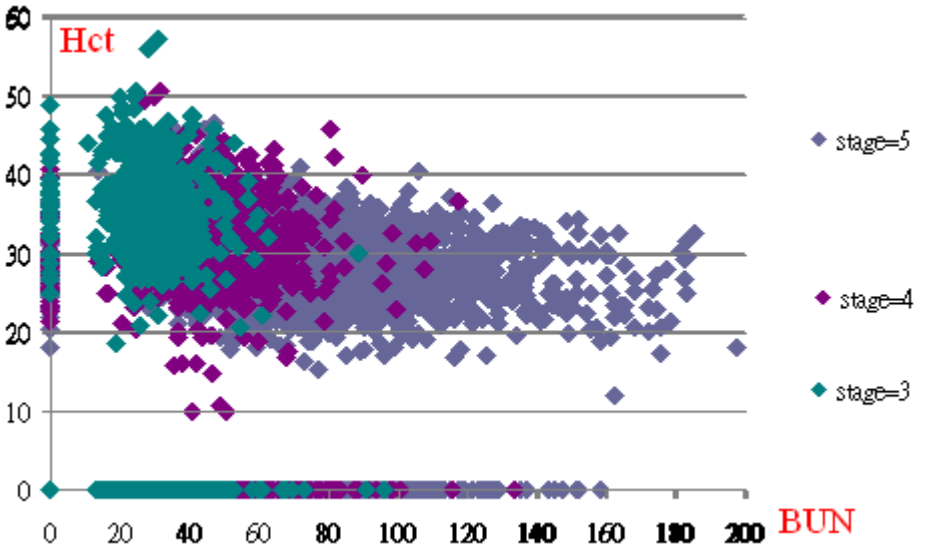


圖4_32 BUN & Hct 變數分佈及Stage重疊性分析

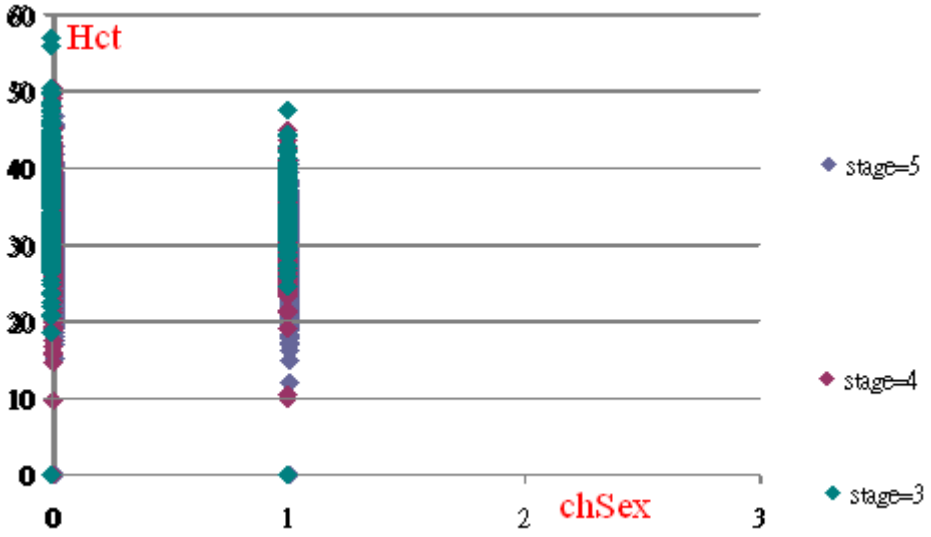


圖4_33 chSex & Hct 變數分佈及Stage重疊性分析

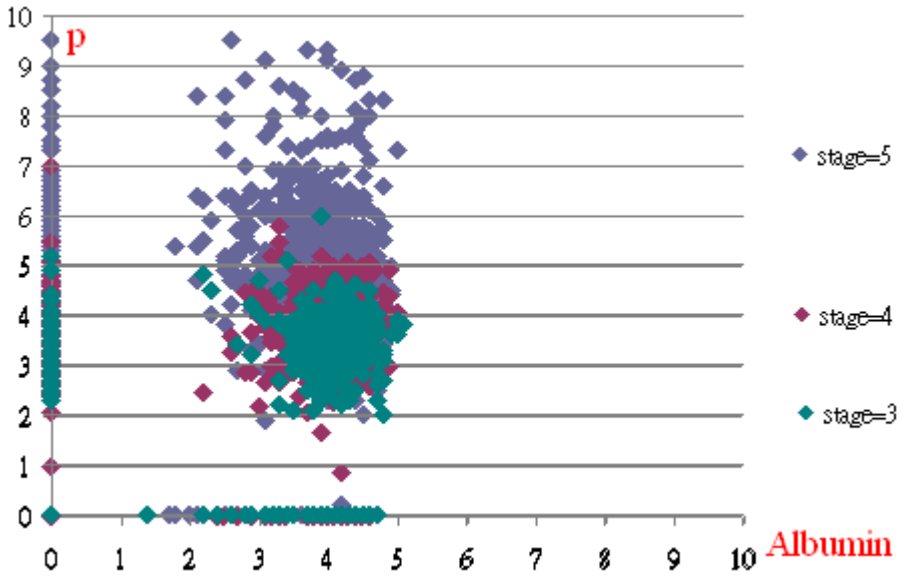


圖4_34 Albumin & P 變數分佈及Stage重疊性分析

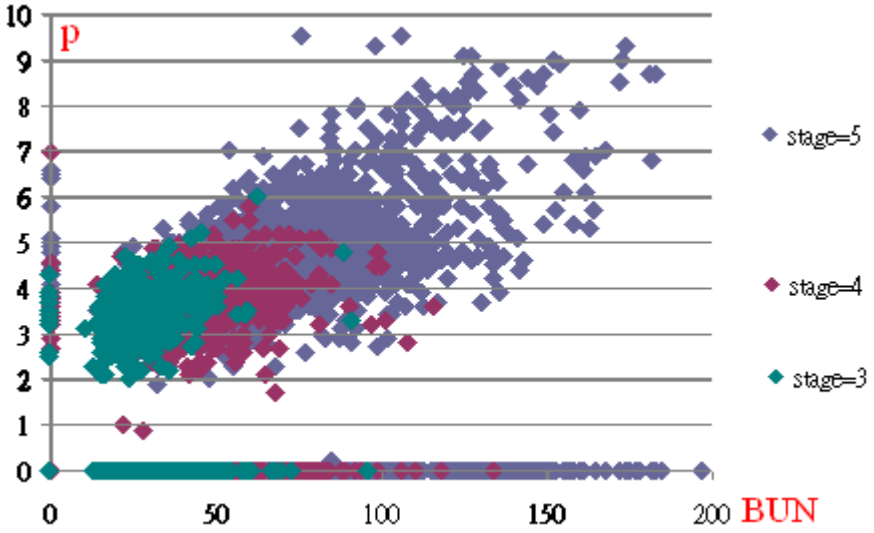


圖4_35 BUN & P 變數分佈及Stage重疊性分析

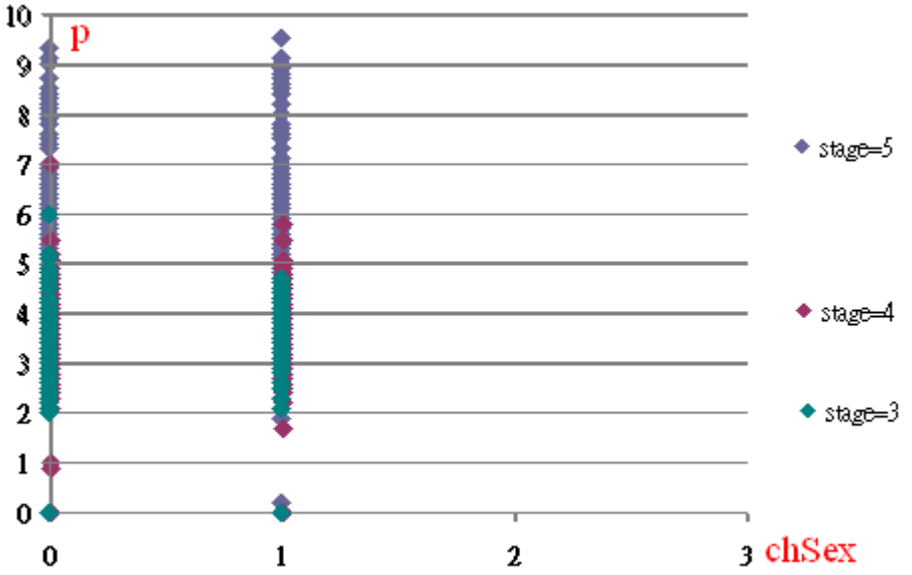


圖4_36 chSex & P 變數分佈及Stage重疊性分析

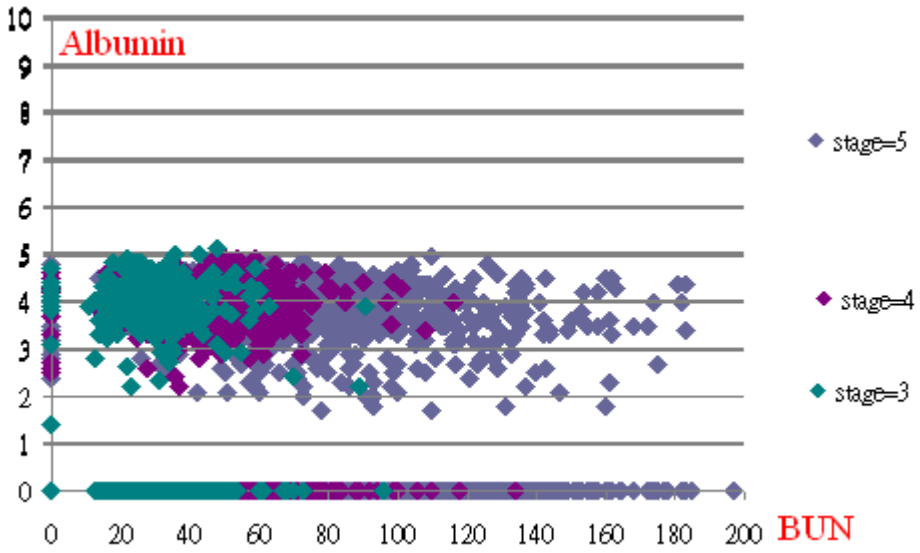


圖4_37 BUN & Albumin 變數分佈及Stage重疊性分析

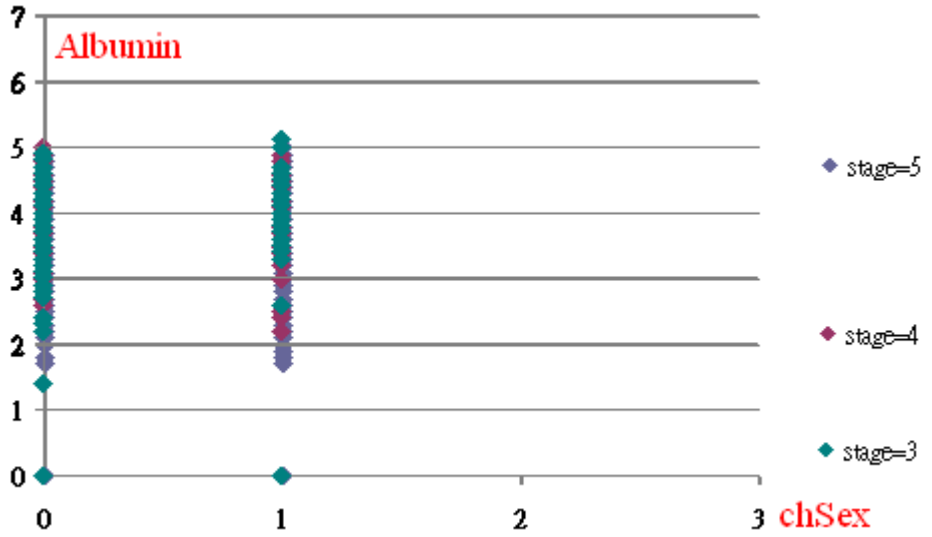


圖4_38 chSex & Albumin 變數分佈及Stage重疊性分析

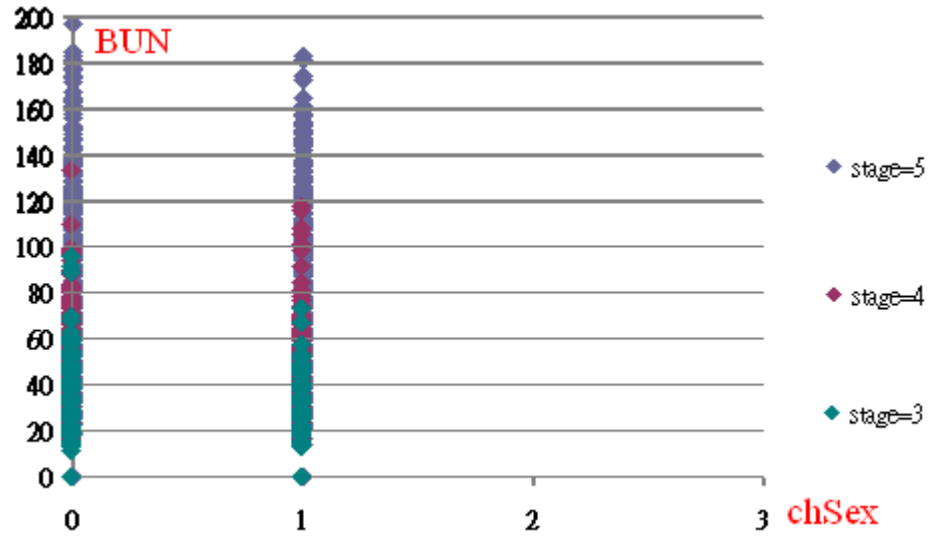


圖4_39 chSex & BUN 變數分佈及Stage重疊性分析

由圖 4_30 至圖 4_39 各別針對變數 Hct、P、Albumin、BUN、chSex 彼此間之關係與 Stage3.4.5 進行交叉比對、觀察後發現，Stage4 於各個圖表中與 Stage3.5 之重疊性呈現較高比例之狀態，我們由此可以驗証出 Stage4 因與 Stage3.5 重疊性較高因而產生預測值明顯偏低之情形。

伍、類神經網路輸出檢視

類神經網路一直讓人詬病之處，在於類神經網路是由複雜的權重與轉換函數所構成，因此，通常被視作為黑箱規則。但在資料採礦的過程中，除了獲得一個準確的模型之外，更重要的是希望能夠從規則內容中找出有意義的內容，因此 SQL Server 2005 類神經網路演算法的檢視器中，主要是呈現資料內的機率分布架構(尹相志，2007)。

在 SQL Server 2005 類神經網路檢視器中分三個區塊:輸入、輸出、變數。其中輸入條件只限於縮小母體範圍，我們可以透過改變輸入區內之屬性選擇所要的變數及改變該變數值之範圍。輸出區域中之輸出項目即是我們建立模型時之「可預測變數」。而變數區域所呈現之值依據輸入變數及輸出變數之屬性作為過濾條件依照顯著性由高至低排列出所有變數選項組合，此處需注意的是，透過類神經網路檢視器可以協助我們了解變數之間交互作用的統計，而非權重高低，我們只能使用檢視器瞭解變數重要性，不過仍無法了解類神經網路結構之全貌。

如下各圖將依類神經網路模型之檢視器來分析病患各 Stage 彼此間變數的相關性。

由「圖 4_40 Stage3 與 Stage4 相關變數分析」中可以看出以 BUN 值之影響顯著性較高，前幾項介於(16.17.18.19.21.22.23....)之值預測性傾向於 Stage3，其次為 BUN 之值介於(74.56.64.58.49.45...)其傾向於 Stage4。由於此表顯示全部變數之顯著性傾向於 Stage3 或 Stage4，單項細部之數據需進一另外擷取。特別是 BUN 此變數屬性為離散值故會產生較多之項目，所以經「表 4_9 BUN 變數值範圍統計圖」取出數據、整理後取得喜好值分數大於 30 對應之 BUN 值變數範圍。

	BUN 值(分數大於 30)
喜好值 Stage3	22,19,27,17,23,16,18,21,25,94,14,26,24,20,28,29,102,34,30
喜好值 Stage4	74,56,64,58,49,45,54,51,59,65,42,67,48,55,73,78

表 4_10Stage3 與 Stage4 傾向 BUN 變數值範圍統計圖

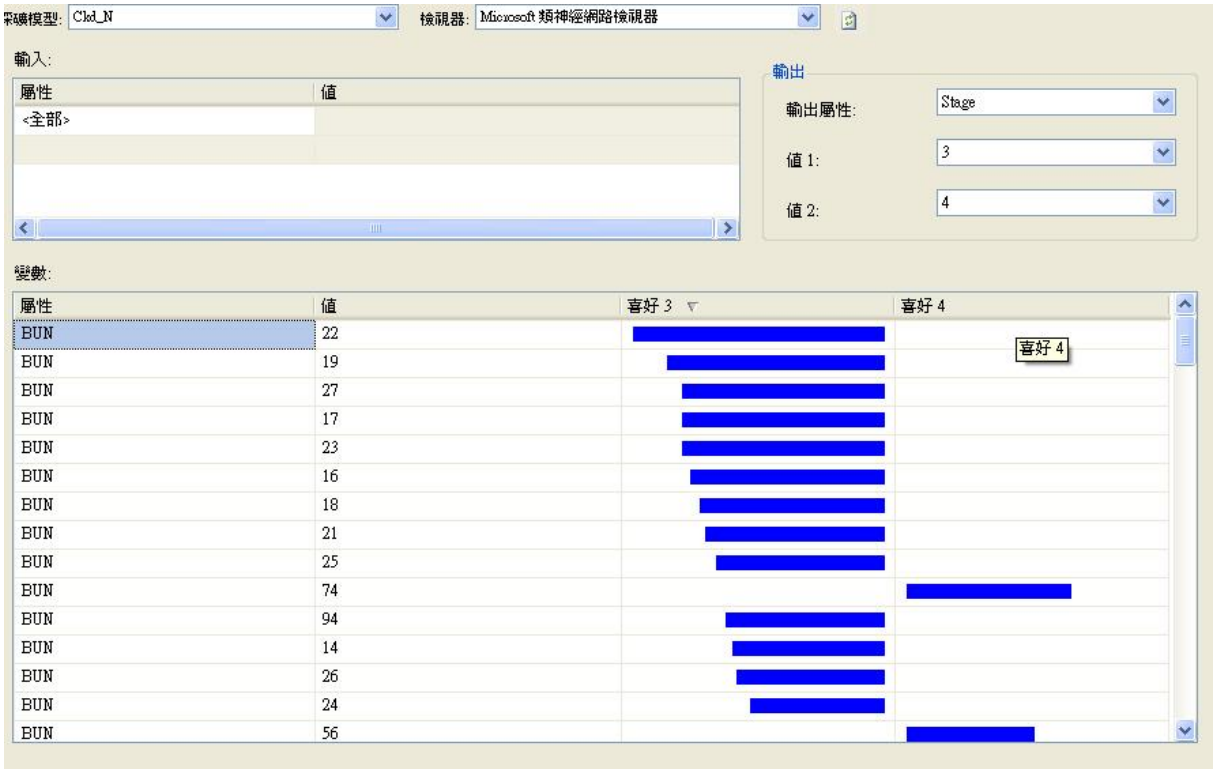


圖4_40 Stage3與Stage4相關變數分析

如「表 4_9 Stage3 與 Stage4 傾向 BUN 變數值範圍統計圖」所擷取出分數大於 30 之數值所示，喜好值 Stage3 對應之 BUN 變數值其範圍主要為 14-34 之間的數值，在此範圍之外另有 94、102 二數值零星分布。而喜好值 Stage4 對應之 BUN 變數值其範圍主要為 42-78 之間的數值，由以上統計我們可整理出 BUN 變數值如果是介於 14-34 間即傾向於 Stage3；而 BUN 變數值如果是介於 42-78 間即傾向於 Stage4。

另外「表 4_10 變數輸出喜好值擷取」為擷取各變數前幾項顯著性較高之值，其中第一項 BUN 值 22 之分數最高亦代表其顯著性最高，另系統依發生機率自動帶出增益(Lift)值決定傾向喜好之判斷，依此類推即可看出各變數之顯著性高低及 Stage 之傾向。

屬性	值	喜好 3 (Stage3)	喜好 4 (Stage4)
BUN	22	分數：100 值 1 的機率：89.52% 值 2 的機率：7.65% 值 1 的增益：7.63 值 2 的增益：0.27	
BUN	94		分數：65.63 值 1 的機率：0.58% 值 2 的機率：12.71% 值 1 的增益：0.05 值 2 的增益：0.45
P	3.633-8.954		分數：27.34 值 1 的機率：2.77.% 值 2 的機率：16.85% 值 1 的增益：0.24 值 2 的增益：0.59

chSex	1		分數：14.44 值 1 的機率：5.20% 值 2 的機率：20.53% 值 1 的增益：0.44 值 2 的增益：0.72
chSex	0	分數：13.62 值 1 的機率：22.47 值 2 的機率：34.68 值 1 的增益：1.92 值 2 的增益：1.21	
P	0.000-0.546	分數：12.58 值 1 的機率：20.39 值 2 的機率：32.58 值 1 的增益：1.74 值 2 的增益：1.14	

表4_11　Stage3與Stage4變數輸出喜好值擷取

同上所述下列圖表呈現了 Stage4 與 Stage5 之間各變數顯著性高低及 Stage 傾向。

	BUN 值(分數大於 30)
喜好值 Stage4	30,20,31,35,16,21,41,29,38,28,36,40,33,25,24,42,108,32,23,43,22,26,49,37,44,22,26,49,37,44,17,39
喜好值 Stage5	82,83,95,103,105,115,87,94,80,112,86,76,127,101,91,84,102,93,90,70,104,78,106,72,92,65,109,77,98,88,85,75,97,110,99,96,73,71,60

表4_12 Stage4與Stage5 傾向BUN變數值範圍統計圖

至於上列「表 4_11 Stage4 與 Stage5 傾向 BUN 變數值範圍統計圖」所擷取出分數大於 30 之數值所示，喜好值 Stage4 對應之 BUN 變數值其範圍主要為 16-49 之間的數值，在此範圍之外另有 108 之數值。而喜好值 Stage5 對應之 BUN 變數值其範圍主要為 60-127 之間的

數值，由上所述我們可整理出 BUN 變數值如果是介於 16-49 之間即傾向於 Stage4；而 BUN 變數值如果是介於 60-127 之間即傾向於 Stage5。

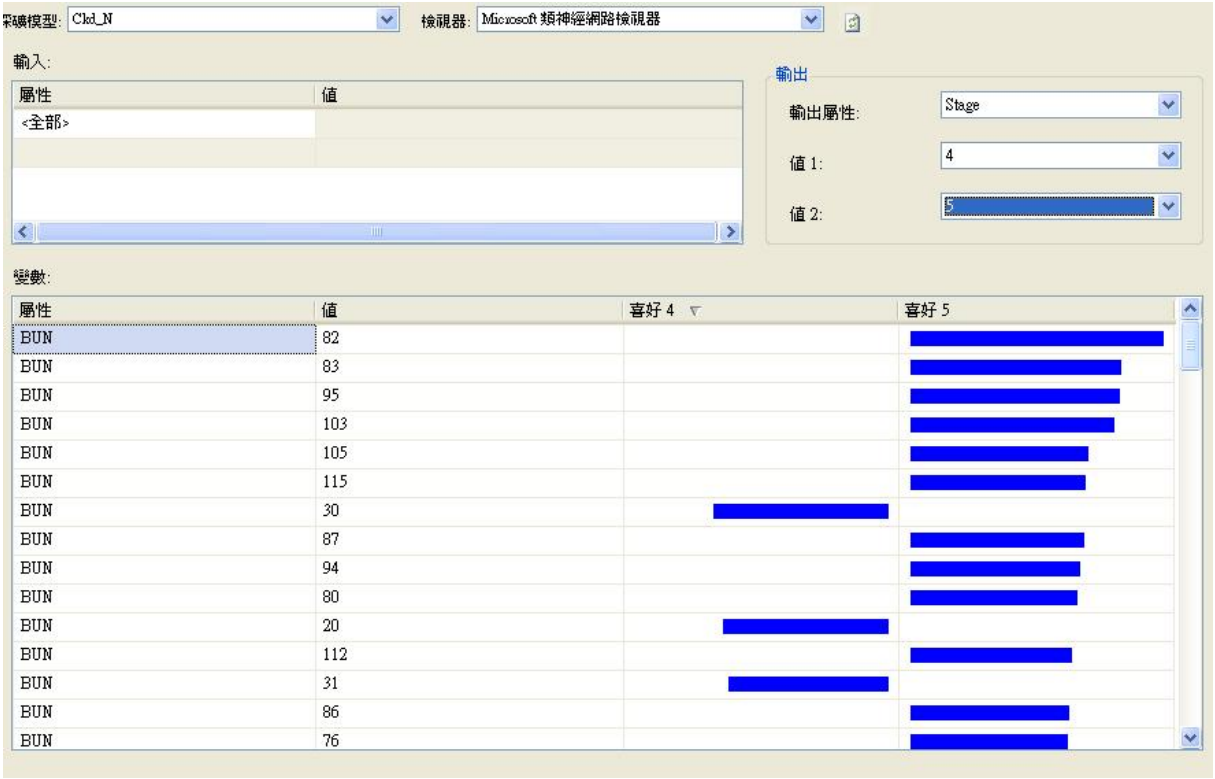


圖4_41 Stage4與Stage5相關變數分析

屬性	值	喜好 4 (Stage4)	喜好 5 (Stage5)
BUN	82		分數：100 值 1 的機率：0.72% 值 2 的機率：98.61% 值 1 的增益：0.03 值 2 的增益：1.65
BUN	30	分數：69.26 值 1 的機率：44.08% 值 2 的機率：5.07% 值 1 的增益：1.54 值 2 的增益：0.08	
P	3.633 - 8.954		分數：19.72 值 1 的機率：16.8% 值 2 的機率：80.34%

			值 1 的增益：0.59 值 2 的增益：1.35
Albumin	2.817 – 5.000	分數：15.31 值 1 的機率：39.80% 值 2 的機率：43.77% 值 1 的增益：1.39 值 2 的增益：0.73	
Hct	30.244 – 56.800		分數：13.69 值 1 的機率：19.65% 值 2 的機率：72.77% 值 1 的增益：0.69 值 2 的增益：1.22
chSex	1		分數：13.12 值 1 的機率：20.53% 值 2 的機率：74.24% 值 1 的增益：0.72 值 2 的增益：1.24

表4_13 Stage4與Stage5變數輸出喜好值擷取

透過類神經網路檢視器能協助我們瞭解變數之間的交互作用，因此我們利用此檢視器瞭解變數之重要性。根據上列表格及圖所示可發現，在 Stage 間變數 BUN 值幾乎是各個階段最顯著之影響變數，我們再搭配其他變數之影響性可做為預測該病患 Stage 之傾向。

陸、Stage 間變數之影響顯著性高低概略整理：

Stage3 與 Satge4：BUN(14-34) > BUN (42-78) > P(3.633-8.954) > chSex (1)> chSex (0) >P(0.000-0.546)

Stage4 與 Satge5：BUN(60-127) > BUN(16-49) > P(3.633 - 8.954) > Albumin(2.817–5.000)> Hct(30.244 – 56.800) > chSex(1)

第五章 結論與建議

本章共分為兩節，第一節為研究結論，針對前列資料分析之發現提出總結及相關建議，第二節為研究限制及後續研究建議，主要針對未來研究提出相關建議。

第一節 研究結論

本研究期待找出慢性腎臟病患 Stage 間，所受變數影響顯著之關係，並提供給醫師、護理人員等線上專業從事人員作為參考，以協助其在進行診治、照護及個案管理等醫療行為時，能有所依據並針對各別 Stage 之病患達到個別適切之醫療、照護行為達到控制延緩 Stage 演進之時程。

所以本研究透過建置決策樹模型找出關聯變數，再利用上述之關聯變數，建立倒傳遞類神經網路演算模型來預測 Stage3.4.5 彼此間變數影響的相關性，期待透過這樣的方式將現有病患資料轉換成有用的資訊。

本研究對於資料分析中發現：

1. 資料來源已收個案之病患主要分佈於 Stage3.4.5，反觀 Stage1.2 收案個案，由於受限於早期症狀不明顯所以不易即時發現，所以佔比例極低。
2. 現有資料統計顯示 Stage1.2 階段之病患年齡分佈大多是年長者，可能原因為本研究資料來源有地區性之故，建議醫護人員能提高對年長病患的細微病徵加強注意，並進一步做血液生化檢驗分析，達到早期發現早期治療之立基。
3. 而資料來源中之血液生化的紀錄欄位多達 19 個，本研究透過結策樹模型及類神經網路模型測試後，針對 Stage3.4.5 找出血液生化欄位 Hct、Albumin、BUN、chSex 及 P 等五個變數對於 Stage3.4.5 有顯著的影響性(其中又以 BUN 有較高的增益值)，因此若病患於上述變數之病徵有明顯變化，建議需要特別注意並對病患做出適切之醫護行為，以防病程分期之變化。
4. 研究樣本居住地主要分布於雲林縣及嘉義縣，建議可針對此地區就診病患適度加強慢性腎藏疾病之檢驗及護理衛教。另外也可針對此地區民眾進行巡迴義診及座談會等活動以發現潛在病患及達到防治之目的。

第二節 研究限制及後續研究建議

研究限制:

如果能有好的資料樣本才能訓練出好的預測模型，而有好的模型才能探勘出精確的結果。在本研究過程中所遇到最大的問題就是原始資料在 Stage1.2 的數據量所佔比例太過稀少，而 Stage4 與 Stage3.5 變數資料間有較高之重疊性。另外在血液生化檢驗值資料表中多個欄位的遺漏值也佔了一定之比例，所以一直無法順利處理模型，在樣本資料的評估及整理就花了許多的時間，並反覆的利用不同演算法模型做測試，最後再針對決策樹及類神經網路模型做多次訓練與檢視結果不合理之因素進而修正輸入變數，始得到最終之結果。

研究建議:

Stage1.2 的數據量所佔比例太過稀少、血液生化檢驗資料表欄位有遺漏值及資料樣本有強烈地域性之部份，皆會影響研究之資料探勘結果，並且未能充分展現探勘工具之功能。建議後續研究可先就上述部份之樣本資料做補充、收集。這樣我們或許可以找出其他更具影響性之變數。

另外導致病患病徵的變化不只侷限於血液生化檢驗值之變化，臨床上還有其他之因素，如：飲品習慣、抽煙、嚼檳榔、家族病史、用藥史、危險因子及其他疾病等因素，在本研究並未能取得相關資料一併進行資料探勘，建議後續研究可將相關因素納入分析之範圍。

參　考　文　獻

一、中文部份

1. 邱鼎鈺，健保雙月刊第 71 期(2010 年 1 月 13 日)
   http://www.nhinb.gov.tw/chinese/14_epaper/infor_epaper1.asp?appID=19&webtype=1&pid=113
2. 方德昭，中央網路報(2010 年 3 月 15 日)
   http://www.cdnews.com.tw
3. 洪嘉鴻，建立末期腎臟疾病患者資源耗用關聯群之初探，高雄醫學大學健康科學院公共衛生研究所，2004
4. 李丞華，中時電子報，(2009 年 4 月 25 日)
   http://tw.news.yahoo.com/article/url/d/a/090425/4/1ieht.html
5. 台灣腎臟醫學會，2006，何謂腎臟，2010 年 3 月 1 日
   http://kidney.tsn.org.tw/1_1.html
6. 陳鴻鈞 教授，高醫醫訊月刊第二十一卷第五期，2001 年 10 月
7. 台灣腎臟醫學會，2006，慢性腎臟病，2010 年 3 月 3 日
   http://kidney.tsn.org.tw/1_7.html
8. 台灣腎臟醫學會，2006，末期腎衰竭的治療方法，2010 年 3 月 6 日
   http://kidney.tsn.org.tw/pup/p1_07_04.htm
9. 姚吉峰，以關聯分析及模糊分割法建構分類規則應用於 CRM 資料分類，國立成功大學資訊管理研究所，2002
10. 尹相志，SQL Server 2005DataMining 資料採礦與 Office 2007 資料採礦增益集，悅知文化，台北，2007
11. 許依宸，資料採礦在學生流失偵測上之應用，南華大學資訊管理學系，2009
12. 王宗屏，智慧型系統於應用資訊系統使用者問題分類之研究，大同大學資訊工程研究，2009

二、西文部份

1. Frawley, W. J., Paitetsky-Shapiro, G. and Matheus, C.J., 1991, Knowledge Discovery in Databases: An Overview Knowledge Discovery in Database, AAAI/MIT Press, California, 1-30.
2. Grupe, F. H., & Owrang, M. M. (1995). Database mining discovering new knowledge and cooperative advantage. Information Systems Management, 12(4), 26-31.
3. U. M. Fayyad, Data Mining and Knowledge Discovery: Making Sense out of Data, IEEE Expert, Vol.11, No.5, October 1996, pp.20-25
4. M. J. A. Berry and G. Linoff, Data mining Technique For Marketing, Sale, And Customer Support, Wiley Computer, 1997.
5. Akaka, D. K. (2004). Data mining: federal efforts cover a wide range of uses. Washington Federal Government General Accounting Office (GAO) Report(GAO-04-548).
6. American Association for Artificial Intelligence (AAAI). (2006). Data mining and discovery, Retrieved from http://www.aaai.org/AITopics/html/mining.html