# 南　華　大　學

## 資訊管理學系碩士班

## 碩士論文

# 資料採礦在學生流失偵測上之應用

# Application of Data Mining Techniques for Detection of Student Drop out

研 究 生：許依宸

指導教授：邱宏彬

中華民國　98　年　6　月　20　日

# 南　華　大　學

## 資訊管理學系碩士班

## 碩 士 學 位 論 文

資 料 採 礦 在 學 生 流 失 偵 測 上 之 應 用

研究生：許依展

經考試合格特此證明

口試委員：

李翔詣

指導教授：

系主任(所長)：

口試日期：中華民國　98　年　06　月　20　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：許依宸 之碩士畢業論文

中文題目：資料採礦在學生流失偵測上之應用

英文題目：Application of Data Mining Techniques for Detection of Student Dropout

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：許依宸 （請親自簽名）

指導老師：邱宏彬 （請親自簽名）

中　華　民　國　98　年　6　月　20　月

# 誌 謝

很快地，碩士班生涯兩年時間即將告一段落，回首這兩年，多的是學習上的喜悅，至於苦的部分，現今回想起來也不算什麼了，之所以現在能在此說這些，其實都要感謝大家。

首先，我要感謝我的指導教授邱宏彬老師，是他的鼓勵讓我覺得我應該作得到，原本考慮是否該再給自己一點時間，然而，感謝老師的不斷鼓勵，結果，我真的做到了。感謝口試委員謝昆霖老師及李翔詣老師不吝指正本論文之缺失，並惠賜寶貴建議，使論文內容更臻完備，在此重申感謝之意。

感謝我的同事怡茜每每在我請休假去上課時，在公務上的幫忙與協助。感謝我的主管的通融，讓我定期請休假去上課，同時也感謝系上老師及同學們的包容。

感謝家人的支持與體諒，並包容我這段期間幾次家庭聚會的缺席。

最後，我要感謝我親愛的老公毛俊景，他是我完成論文的最大動力與支柱。

謹以此論文獻給所有幫助過我的人，更要祝福你們，並將此完成學業的喜悅與你們分享。

依宸

謹誌於南華大學資訊管理學系碩士班

2009/06

# 資料採礦在學生流失偵測上之應用

學生：許依宸　　指導教授：邱宏彬 博士

南　華　大　學　資訊管理學系碩士班

## 摘　　要

依據教育部資料，本國大專校院數量不斷增加，大學日間部學生招生人數亦隨之增加，然而，依據內政部之統計資料，本國出生人口總數卻是不斷減少，因此招生情況是日益嚴峻。本研究以南華大學 94-96 學年度大學日間部入學學生資料為例，運用資料採礦技術，分析學生歷史學籍資料，找出流失學生之潛在共同因素，建立學生流失預測模型。針對本研究之資料採礦結果，本研究將提出相關之建議，提供學校參考以改善學生流失之情況。

關鍵詞：資料採礦、學生流失、決策樹、類神經網路

Application of Data Mining Techniques for Detection of Student Drop out

Student：Hsu,Yi-Chen Advisors：Dr. Chiu,Hung-Pin

Department of Information Management
The M.I.M. Program
Nan-Hua University

## ABSTRACT

According to the data from Ministry of Education, the number of college increase constantly. The enrollment of students in the day school also increases. However, the statistics showed from Ministry of Internal Affair that total birth rate of Taiwan is decreasing. Therefore, the enrollment status is getting tougher. This research is based on the day school student enrollment data between 2005~2007 in Nan-hua University. By using the data mining techniques and analyzing the history of student record, this research tried to find the command of potential reason and created a predict model of why students drop out. The analyzing and mining results from this research will provide relevant suggestions to help the university to improve the situation of student drop out.

Keyword: data mining, student drop out, decision tree, neural networks

# 目　錄

# 表目錄

# 圖目錄

# 第一章 緒論

本章說明本研究之研究背景與動機、研究目的、以及研究範圍與限制。第一節為研究背景與動機，說明目前各大專院校學生流失之現況；第二節為研究目的；第三節為研究範圍與限制。

## 第一節 研究背景與動機

依據教育部高教司統計處的資料，本國大學校院數量由民國七十五學年度的二十八所增加為九十七學年度的一百四十七所（參見表 1-1　75 學年度至 97 學年度大學校院數變動表），而大學日間部的總招生人數也由九十一學年度的 96,847 增加為九十六學年度的 110,099 人（表 1-2　近六年大學校院及研究所核定招生名額統計（分學制別）），而根據內政部統計，民國八十五年以前，每年出生人口都有三十萬人的水準，八十六年後出生人口越來越少（八十九年除外），九十五年出生人口更是只有二十萬人四千多人（表 1-3　歷年人口出生數及其比率），少子化呈現的並不是只有短期問題，而是一波波對未來的衝擊，對教育而言，從小學、國中、高中職到大學，都會受到影響。

表 1- 1　75 學年度至 97 學年度大學校院數變動表

| 學年度 | 大學 | | | | 學院 | | | | 總計 |
|---|---|---|---|---|---|---|---|---|---|
| | 國立 | 市立 | 私立 | 計 | 國立 | 市立 | 私立 | 計 | |
| 75 | 9 | 0 | 7 | 16 | 6 | 0 | 6 | 12 | 28 |
| 80 | 13 | 0 | 8 | 21 | 14 | 1 | 14 | 29 | 50 |
| 81 | 13 | 0 | 8 | 21 | 14 | 1 | 14 | 29 | 50 |
| 82 | 13 | 0 | 8 | 21 | 14 | 1 | 15 | 30 | 51 |
| 83 | 15 | 0 | 8 | 23 | 16 | 1 | 18 | 35 | 58 |
| 84 | 16 | 0 | 8 | 24 | 17 | 1 | 18 | 36 | 60 |
| 85 | 16 | 0 | 8 | 24 | 19 | 2 | 22 | 43 | 67 |
| 86 | 20 | 0 | 18 | 38 | 19 | 2 | 19 | 40 | 78 |
| 87 | 21 | 0 | 18 | 39 | 20 | 2 | 23 | 45 | 84 |
| 88 | 21 | 0 | 23 | 44 | 23 | 2 | 36 | 61 | 105 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 89 | 25 | 0 | 28 | 53 | 22 | 2 | 50 | 74 | 127 |
| 90 | 27 | 0 | 30 | 57 | 21 | 2 | 55 | 78 | 135 |
| 91 | 27 | 0 | 34 | 61 | 21 | 2 | 55 | 78 | 139 |
| 92 | 30 | 0 | 37 | 67 | 19 | 2 | 54 | 75 | 142 |
| 93 | 34 | 0 | 41 | 75 | 15 | 2 | 53 | 70 | 145 |
| 94 | 40 | 1 | 48 | 89 | 9 | 1 | 46 | 56 | 145 |
| 95 | 40 | 1 | 53 | 94 | 10 | 1 | 42 | 53 | 147 |
| 96 | 41 | 1 | 58 | 100 | 9 | 1 | 39 | 49 | 149 |
| 97 | 41 | 1 | 60 | 102 | 7 | 1 | 37 | 45 | 147 |

說明：1.92、95 及 96 學年度校數為第一學期資料。

2.本表資料不含專科學校（15 所）、軍事院校、警大及空大。

3.資料來源：教育部高教司統計處。

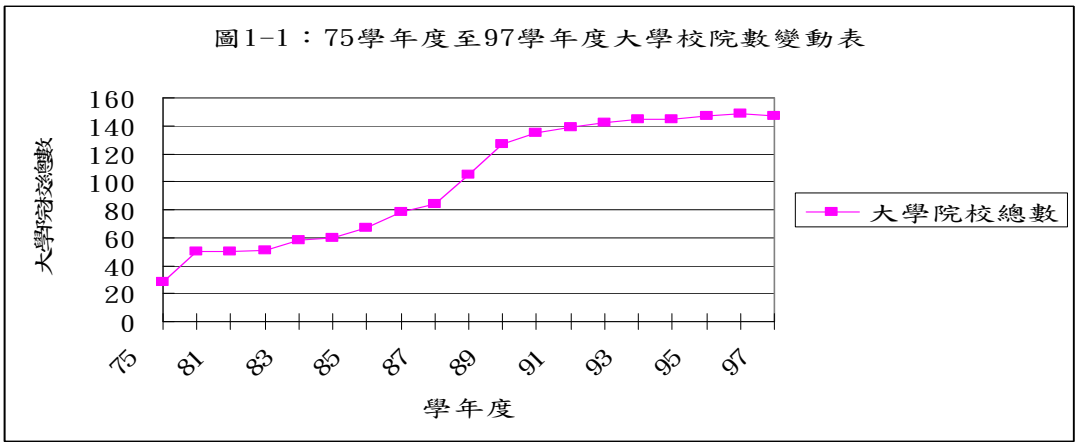


圖 1- 1　75 學年度至 97 學年度大學校院數變動表

表 1- 2　近六年大學校院及研究所核定招生名額統計（分學制別）

| | 日間學制 | | | | 夜間學制 | | | 日夜間總計 |
|---|---|---|---|---|---|---|---|---|
| | 大學部 | 碩士班 | 博士班 | 小計 | 大學部 | 碩士班 | 小計 | |
| 91 | 96,847 | 27,634 | 4,458 | 128,939 | 19,692 | 8,499 | 28,191 | 157,130 |
| 92 | 104,190 | 30,841 | 5,175 | 140,206 | 22,172 | 11,251 | 33,423 | 173,629 |
| 93 | 104,619 | 31,173 | 5,511 | 141,303 | 21,922 | 12,140 | 34,062 | 175,365 |
| 94 | 105,181 | 32,343 | 5,823 | 143,347 | 22,426 | 12,749 | 35,175 | 178,522 |
| 95 | 106,696 | 34,005 | 6,140 | 146,841 | 23,650 | 13,783 | 37,433 | 184,274 |
| 96 | 109,274 | 35,288 | 6,365 | 150,927 | 20,582 | 14,002 | 34,584 | 185,511 |
| 97 | 110,099 | 35,646 | 6,321 | 152,066 | 50,205 | 13,657 | 33,862 | 185,928 |
| 97-96 | 825 | 358 | -44 | 1,139 | -377 | -345 | -722 | 417 |

資料來源：教育部高教技職簡訓 011 期（不含師範、技職校院）

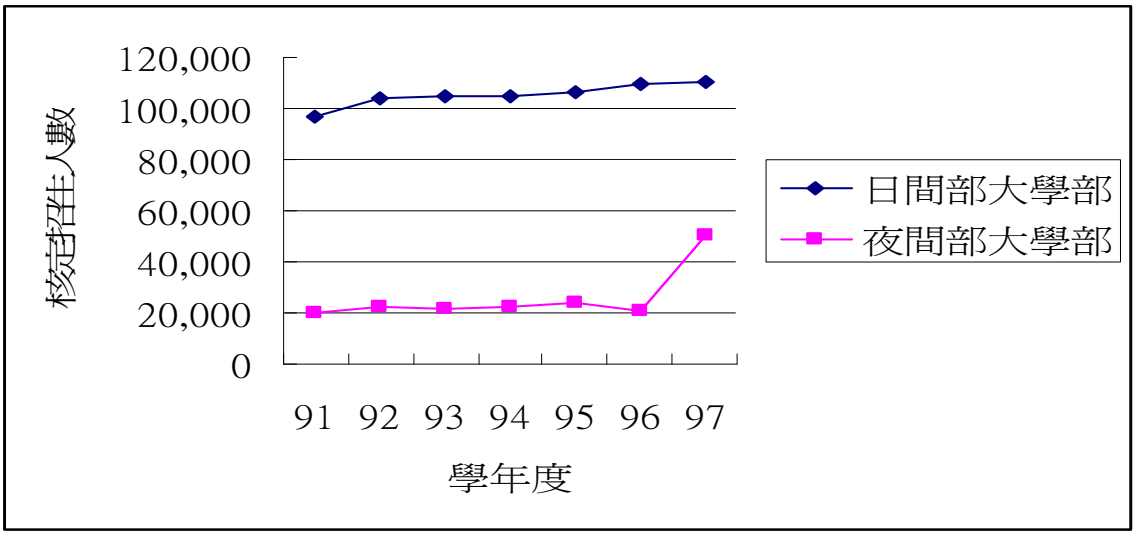


圖 1-2　近六年大學校院大學部核定招生名額統計

表 1-3　歷年人口出生數及其比率

| 年別 | | 出生 | |
|---|---|---|---|
| | | 人口數 | 粗出生率（千分率） |
| 民國 75 年 | 1986 | 309,230 | 15.9 |
| 民國 76 年 | 1987 | 314,024 | 16.0 |
| 民國 77 年 | 1988 | 342,031 | 17.2 |
| 民國 78 年 | 1989 | 315,299 | 15.7 |
| 民國 79 年 | 1990 | 335,618 | 16.6 |
| 民國 80 年 | 1991 | 321,932 | 15.7 |
| 民國 81 年 | 1992 | 321,632 | 15.5 |
| 民國 82 年 | 1993 | 325,613 | 15.6 |
| 民國 83 年 | 1994 | 322,938 | 15.3 |
| 民國 84 年 | 1995 | 329,581 | 15.5 |
| 民國 85 年 | 1996 | 325,545 | 15.2 |
| 民國 86 年 | 1997 | 326,002 | 15.1 |
| 民國 87 年 | 1998 | 271,450 | 12.4 |
| 民國 88 年 | 1999 | 283,661 | 12.9 |
| 民國 89 年 | 2000 | 305,321 | 13.8 |
| 民國 90 年 | 2001 | 260,354 | 11.7 |
| 民國 91 年 | 2002 | 247,530 | 11.0 |
| 民國 92 年 | 2003 | 227,070 | 10.1 |
| 民國 93 年 | 2004 | 216,419 | 9.6 |
| 民國 94 年 | 2005 | 205,854 | 9.1 |
| 民國 95 年 | 2006 | 204,459 | 9.0 |

資料來源：內政部戶政司人口統計資料

依據教育部高教司統計處的資料，96 學年度第 2 學期因學業成績不及格遭退學人數統計，該學期一般校院（不含技職、體院、軍警校院）退學人數合計 3,955 人，比較 96 學年度第 1 學期退學人數 3,519 人，增加 436 人。而截至 97 學年度第 2 學期為止，各公私立大學院校的休學人數統計高達 5,138 人，人數已大幅度增加。

根據聯合晚報（2009/01/03）的報導指出，少子化趨勢下，每年新生兒人數不到 20 萬，但目前大學技院招生總人數多達 24 到 25 萬人，教育部預期大學招生供過於求的問題將更嚴峻，正著手研擬逐年調降大學招生總量的公式，因此「大學逐年減招」是非走不可的路。

在面臨招生人數逐年下降，而休退學人數逐年增加的雙面夾殺之下，如何輔導學生，讓那些辛苦招收進來的學生在自身學校順利完成學業，成為各大專校院亟需探討的問題。

## 第二節 研究目的

為因應招生情況的日益嚴峻，針對招收進來的在校學生，學校應加強輔導、多加關心，對於其學習及生活之各種情形應多加觀察注意，努力讓學生在本校順利完成學業。

由於學校各人員及相關教師之日常行政工作及教學活動已相當忙碌，如何協助這些人員及早發現潛在可能流失學生，而加以關心輔導，以減少其流失率，則是一重要關鍵。

然因學生眾多，要發現其潛在可能流失學生實在相當不易，本研究擬運用資料採礦技術分析學生歷史學籍資料，找出流失學生之潛在共同因素，作為在學學生可能流失之預測，針對資料採礦之結果，提供相關之建議，以降低學生流失之情形。

## 第三節 研究範圍、限制

本研究之研究範圍以南華大學 94 學年度至 96 學年度大學日間部學生為主，因以單一學校資料庫為分析來源，結果恐難概化至他校，然仍可作為各校減少學生流失之參考；此外，本研究以資料庫資料為分析來源，對於學生之情緒性因素、家庭、經濟、交友等情形較難以分析。而因資料取得問題，本研究將研究範圍限制在三個系。

## 第四節 研究步驟

本研究將探討影響學生流失的相關因素，以「南華大學 94-96 學年度大學日間部入學新生之資料」，作為研究使用之資料庫，以資料採礦中之決策樹及類神經網路之採礦工具，來深入討探學生流失之重要關鍵因素，現依循下列步驟來完成本研究。研究步驟包含：研究動機及目的確認、文獻探討、擬定研究方法及架構、資料採礦與分析、結論與建議，其說明如下：

壹、研究動機及目的確認：確認本研究之動機與目的，以確認資料蒐集之方向及研究之進行與操作。

貳、資料蒐集與文獻回顧：蒐集相關文獻並瞭解研究之相關知識，並尋求適合本研究之技術及工具。

參、擬定研究方法及架構：針對所研究之問題及所運用之工具，擬定適合之研究方法及架構。

肆、分析資料的取得：為探討學生流失之重要關鍵因素，本研究以「南華大學 94-96 學年度大學日間部入學新生之資料」，作為研究使用之資料庫，為正確地使用資料採礦技術，在進行資料庫的分析之前，須先行檢視資料庫中各個資料欄位之意義與價值，同時完全了解及

掌握各項資料之原始意義與其使用特性和限制條件，以期在進行分析時可以正確地運用分析方法與資料性質，進行資料採礦工作。

伍、資料前置處理：將不具分析價值之資料欄位予以刪減、修改錯誤的資料格式及利用適當方法將原有資料轉換成具分析價值之資料，此動作將可避免不適合的資料對分析結果造成不利的影響。

陸、資料採礦分析：運用決策樹及類神經網路等資料採礦工具，進行相關資料之分析，取得學生流失預測模型。

柒、結論：找出流失學生之潛在共同因素，作為在學學生可能流失之預測，針對資料採礦之結果，提供可行之建議，以降低學生流失之情形。

## 第五節 論文架構

本研究之呈現共分為五章，各章節結構說明如下：

第一章　緒論

說明本研究之研究背景與動機、研究目的、研究步驟、研究範圍與限制、研究步驟以及本研究之論文架構。

第二章　文獻探討

就相關文獻分別探討顧客流失、學生流失及資料採礦之定義，並針對本研究所運用之決策樹及類神經網路等資料採礦方法加以深入探討。

第三章　研究方法

本章節針對研究資料來源作一概略介紹，包括研究對象、研究架構、資料處理以及資料採礦工具的介紹。

第四章　資料分析

針對研究資料先作一基本敘述統計，並針對所需要進一步研究的部分進行整理，之後進行資料採礦分析，並將結果加以比較。

第五章　結論與建議

針對第四章所得結果進行統整，並提出相關結論及建議。

# 第二章 文獻探討

本章共分為六節，第一節為顧客流失與學生流失定義，第二節為資料採礦定義，第三節為資料採礦運用的介紹，第四節介紹本研究所採用之資料採礦技術-決策樹，第五節介紹另一採礦技術-類神經網路，第六節針對兩篇資料採礦在學生流失上的應用作相關探討。

## 第一節 顧客流失與學生流失定義

根據各行業平均統計值顯示，在既有顧客中每年約有 85%會留下，相對地，有 15%的顧客流失(夏載，2001)；對企業而言，顧客快速流失為企業獲利下降的警訊，將顧客流失率降到最低，企業本身才達到最佳顧客保留率(Strouse, 1999)。以服務業而言，維繫舊客戶比起擴張企業競爭優勢，諸如擴大經營規模、市場佔有率以及降低單位成本等方法具有更大的獲利價值(Colgate et al., 1996)。透過成本的節省、顧客重複購買與購買量的增加與口碑效果，若能減少 5%的流失率，可使獲利率增加 25%到 85%(Reichheled and Sasser, 1990)，而企業如能有效延緩顧客流失的速度，預估可增加 15%的長期收益(Bolton, 1998)；相反地，高流失率對企業的利潤與競爭力產生極大負面影響，研究顯示美國企業的客戶流失造成其績效降低 25%到 50%(Reichheld and Teal, 1996)。

維持既有客戶的企業成本遠低於開發新客戶群的成本(Heskett et al., 1989)，而顧客本身具有的學習效果也讓企業能夠降低服務成本支出(Ganesh, et al 2000)；就利潤面考量，既有客戶對於企業獲利之貢獻更遠多過新開發客戶群，口耳相傳的正面宣傳效果替企業提高新顧客流量(Keaveney,1995)，除了增加顧客的留存率之外，亦降低企業開發新客戶

群的行銷成本(Peters 1987)；再者，忠實顧客願意在高毛利的商品上購買較多的數量(Grant and Schlesinger, 1995)，即使價格增加時，忠實的顧客依然會增加服務的使用(Bolton and Lemon, 1999)。由此可見，促使顧客流失情形的減少均能夠創造出高收益和低成本的利益(Fornell and Wernerfelt1987)。

顧客忠誠度可為企業帶來實質利益，企業一年的顧客保留率增加5%，平均可為公司帶來多達 75%的總利潤(Reichheld and Frederick, 1990)，甚或是 100%的企業利潤(Reichheld and Sasser, 1990; O'Malley, 1998)。由此可知，與客戶建立緊密且長久的關係，是企業創造利潤的重要關鍵，如何預防客戶流失將十分重要。

雖然學校性質不似企業，然隨著各大專院校轉型後，招生競爭日趨強烈，顧客關係管理（Custmer Relationship Management,CRM）亦逐漸被運用在學校方面。從 CRM 的概念來看，中途離校生即等於客戶流失，若能運用顧客關係管理的方式經營與學生的關係，從整體來看對學校經營應有所助益。

「流失」一詞普遍被運用在地質學及團體輔導的領域中，在英文中，有幾個關於成員流失的用字，例如："drop out" 、"withdraw" 、"premature termination" 及"casualty"等等，但其內涵上有些許的差異(盧梅莉，民81)。王智弘(民 77、民 83 )曾論及「流失」一詞的定義：對於團體成員的流失在英語上的用字是"drop out"也就是「中途退出」的意思，而另外常用的字還包括"premature termination"（未成熟的終結）與"casualty"（成員的傷亡）等，一般對團體成員流失的定義是：「在事先約定的團體過程結束之前，團體成員不顧諮商員或團體領導員的勸告而中途離開團體」。不過對於"casualty" 一字，Garfield & Bergin (1978) 則建議要與"premature termination" 一字加以區別，認為：「並非所有離開團體的

人都可稱之為『成員之傷亡』，有些離開的成員並非因受到團體經驗的傷害而離開，而可能只是因為對團體有不適當的期望或團體領導的不夠敏感，而感到團體不能滿足其需要而已」。也就是說"casualty"一詞更專指那些受到團體經驗傷害而中途流失的成員(引自王智弘，民 83)。

根據上述對「流失」一詞用語的相關討論，以"drop out"也就是「中途退出」的解釋，所謂「流失」(drop out) 係指團體成員在完成團體訓練或輔導的過程中，團體成員中途退出團體組織的情形，較能合乎本研究的意涵。所以在本研究中對流失的界定為：『曾經註冊繳費，擁有學籍資料之本校學生，因故無法完成學業，中途退出，未取得畢業證書之學生』為本研究對學生流失之定義。

## 第二節 資料採礦定義

Data mining 經中華資料採礦協會（Chung-Hua Data Mining Society，CDMS）譯為「資料採礦」。依中華資料採礦協會（2002）指出資料採礦最早由 Useama Fayyad(1991)提出，其目的為從龐大的維修資料中，找出規則。

資料採礦是一大量自動化的過程，其運用統計分析來從大量的資料集合中，發現有用的、不明顯的和先前未知的特徵或資料趨勢（Frawley et al., 1992)。Kleissner (1998)則表示，資料採礦是去發現公司資料中所隱含的知識並讓企業的管理者能夠瞭解，而來支援決策的分析過程。Berson et al. (2000)認為，資料採礦是指從儲存著大量資料的倉儲中進行挖掘，以發現資料間有意義的新關係、型樣和趨勢的過程。因此，資料採礦不外乎是發現型樣的過程，而這過程一定是自動的或半自動的，所發現的型樣一定是有用的並可帶來一些優勢，其所謂的優勢，通常是經濟優勢(Witten and Frank, 2005)。藉由資料採礦的技術，可以增進對顧客需求和

行為的瞭解，並有助於企業提供客製化的服務，強化與顧客之間的連結、溝通與互動(Cheng et al., 2005)，亦即可發掘大量關於顧客特徵和購買模式而有益於行銷的知識(Shaw et al., 2001)，多數公司運用資料採礦作為策略的基礎，協助其打敗競爭者、確認新顧客以及降低成本(Davis, 1999)。一般常用的資料採礦技術主要有下述六類：

壹、分類（Classification）：根據已知資料及其分類屬性，建立資料的分類模型，接著利用此分類模型預測新資料的類別。例如：顧客是否會購買筆記型電腦的分類模型。

貳、推估（Estimation）：可用來處理連續性數值的結果，給定一些輸入資料以推估未知的連續性變數的值。例如：金融商品價格之預測等。

參、群集化（Cluster）：群集即為物以類聚，資料依照本身的自我相似性(self-similarity)而群集在一起，群集(clusters)的意義要靠事後的闡釋才能得知。例如：線上購物網站的使用者族群與消費能力-具有類似基本資料的人，通常也有相近的行為模式。

肆、關聯法則（Association Rule）：就是從歷史資料中，找出哪些事件總是相伴發生。例如：產品自動化推薦等。

伍、序列（Sequential）：在同質分組中是找出哪些事物會相伴發生，但是透過序列，可以找出事物「先後」發生的順序，我們稱為時序規則（Sequential Pattern）。

陸、描述（Description）：描述在複雜的資料庫中到底發生了什麼?透過這種方式，可以讓我們對我們的客戶、產品以及流程等有更多的認識與了解。

## 第三節 資料採礦運用

資料採礦各類技術依其特性之各異而適用於不同之領域，例如：

壹、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。

貳、零售製造業-分店設點區位分析、銷售產品組合、庫存管理、即時輔助購買決策、連續銷售、促銷商品組合、DM 名單、庫存分析。

參、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。

肆、教育業-學生來源分析、課程規劃、學習評量、適性化教學。

伍、零售業者而言-瞭解顧客消費特性，發掘顧客採購模式，強化客戶關係，達到留住顧客目的。

陸、銀行業者而言-瞭解信用卡發放可能產生之弊端，找出最有利潤、忠誠度佳的顧客。

柒、保險業者而言-分析保戶要求理賠之模式，並可加強稽核，以防止詐財之發生

## 第四節 決策樹

決策樹是一項建立分類模式(classification models)的方式之一，針對給定的資料利用歸納的方式產生樹狀結構的模式。為了要將輸入的資料分類，決策樹的每一個節點即為一個判斷式，判斷式針對一個變數去判斷輸入的資料大於或等於或小於某個數值，每一個節點因而可以將輸入的資料分成若干類。

決策樹修剪技術可分為事前修剪（Prepruning），即在訓練模型的同時，將該節點設定為「葉節點」，因此該節點停止生長；另一為事後修剪(Postpruning)，即將已產生的決策樹多餘的規則修剪掉。在本研究則採

用調整 COMPLEXITY_PENALTY 參數方式，作事前修剪；採用調整 MINIMUM_SUPPORT 參數方式，作事後修剪。

決策樹不一定比其他模式建立的技術來的準確，但與其他技術相比，決策樹很容易讓人了解，因此大為有用，舉例來說，決策樹常用來找尋購買特定產品的顧客類別，由於決策樹可以讓使用者立刻得到可以理解的結果，使用者便可利用這項結果進行後續動作。

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。

## 第五節 類神經網路

類神經網路的歷史最早可追遡至 1943 年，心理學家華倫・麥克庫羅（Warren McCulloch），與邏輯數學家華特・匹茲（Walter Pits）最早提出了描述神經元運作的數學模式（MP 模式）。1949 年，Hebb 提出了著名的 Hebb 學習定律，認為如果兩個神經元同時被激發時，則它們之間的連接就會獲得加強。（尹相志，2007）

人類的大腦大約由 1011 個神經細胞（Nerve Cells）組成，它的結構包括了幾個主要的單元，如「圖 2-1 人類神經元結構」所示：

壹、 神經核 (Soma)：神經元的中央處理部位。

貳、 軸突 (Axon)：神經元中負責把神經脈衝從細胞體往外傳遞的神經纖維。

參、樹突 (Dendrites)：神經元中負責把神經脈衝傳遞至細胞體的神經纖維。

肆、突觸 (Synapse)：神經元之間的聯結機制。

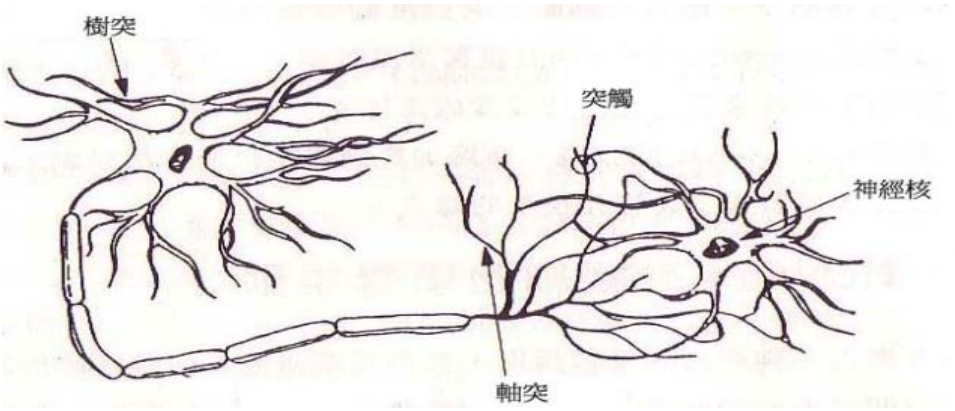


圖 2-3 人類神經元結構

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。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，如「圖 2-2 神經元架構」所示。

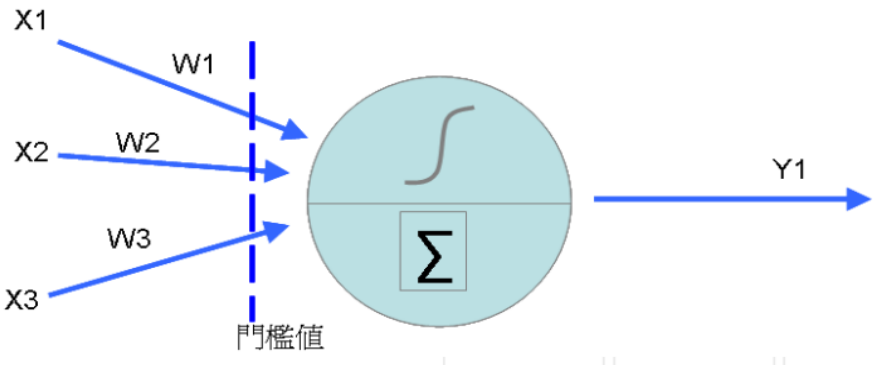


圖 2-4 神經元架構

說明：其中 I 表示輸入值，O 表示輸出值，w 表示權重，而 q 表示該神經元本身的常數項。

將神經元彼此連結，就構成了類神經網路架構，也就是一個神經元的輸出可以變成下一個類神經網路的輸出脈衝。

類神經網路可以應用在連續變數以及類別變數的預測，如果是連續變數預測，則結果就是單一輸出層神經元的輸出訊號強度。如果是類神經網路分類問題，假設我們要使用類神經網路預測這筆交易是否為信用卡盜刷，則結果只有「是」或「否」，此時類神經網路輸出層就會具有兩個神經元分別代表「是」與「否」，當代表「是」的神經元輸出訊號強過代表「否」的神經元時，則預測結果為「是」。（尹相志，2007）

## 第六節 相關研究之回顧與探討

本節介紹兩篇運用資料採礦技術在學生流失議題上之文章。

王建華（民 93）探討中途離校生之出缺席狀況及學業成績關聯，找出技職五專學生其行為特徵並比較與人格測驗之分析結果，其使用變項為德行成績（出席記錄、個人因素）、學業成績，運用決策樹 C4.5 分析缺曠課情形（包含請假原因及節次，找出請假缺課最嚴重的節次）及學業成績（分學期、分數（0 分、及格、不及格）、科系），其研究發現中途離校生缺課情形嚴重且成績較差。

趙瑞麟（民 96）探討以資料探勘技術降低學生流失，針對進修院校二技二專學生，探討科系、性別、年紀、原畢業學校分類、居住地區、學業成績等級、操行成績等級、職業等變項對學生流失之影響，使用 SPSS Clementine7.2 為主要資料探勘工具，利用決策樹 C5.0 、 CART、類神經網路等技術來建立學生流失預測模型，其研究發現一、在預測學生流失的績效上，以 C5.0 決策樹演算法表現最佳，二、學生流失的主要因素為：逾期未註冊、逾期未復學與工作因素主動辦理退學，三、透過決策樹演算法發現二技中輟學生以「操行成績」為主要分類變項，其次為「年

紀」與「學業成績」，而二專學生也以「操行成績」為主要分類變項，其次為「學業成績」與「年紀」。

# 第三章　研究方法

本研究主要針對前述研究背景與動機、研究目的及相關文獻探討，提出本研究之架構並說明資料處理之方式，以作為後續研究之基礎。首先在第一節說明研究對象之概況；第二節提出本研究之架構；在第三節說明資料處理方式；第四節針對資料採礦工具進行介紹。

## 第一節 研究對象

本研究以南華大學 94 學年度至 96 學年度入學之大學日間部學生為研究對象，由於研究樣本選取時間為 97 學年度第 2 學期，故其就學狀態為 97 學年度第 2 學期時之就學狀態。此外，因工作之便，將研究範圍限制在建築與景觀學系、資訊工程學系與資訊管理學系等三個系。

南華大學自民國 85 年成立至今，學生規模由大學日間部 77 人增加為五千多人，大學日間部學系亦由兩個系擴增為 22 個系，然而其大學日間部學生報到人數由 95 學年度的 1,059 人下降到 97 學年度的 916 人，而大學日間部學生流失人數亦由 95 學年度第 1 學期的 45 人，增加為 97 學年度第 1 學期的 97 人，每年不斷遞增。

由於南華大學對於學生離校是採填寫申請表方式，而從學生填寫離校申請單中所填寫之內容，學生多數僅填寫個人因素，故無法獲得更多更明確之訊息；而校務行政系統上登載之學生離校原因包括--休學、退學、轉學、逾期未註冊、逾期未復學、超過修業年限等。

而逾期未註冊及逾期未復學，二者應為離校之結果而非實際上離校之原因，由校務行政系統上並無法確實得知南華大學學生離校之實際原因，故本研究採用資料採礦方式，擬探求學生流失之潛在影響因素，希

望藉由發現學生流失之潛在共同因素，提供南華大學作參考，以減少學生流失之情形。

## 第二節 研究架構

以南華大學 94 學年度至 96 學年度入學之大學日間部學生學籍歷史資料為基礎，主要為校務行政系統上之學生資料，包括學生性別、科系、入學方式、學期成績、歷年學業平均成績、歷年修課學分數、歷年實得學分數、居住地區等為本研究的主要研究變數。

## 第三節 資料處理

本研究將流失界定為：『曾經註冊繳費，擁有學籍資料之本校學生，因故無法完成學業，中途退出，未取得畢業證書之學生』。依本研究所取得之資料，學生流失即就學狀態為：休學、退學、轉學之學生。

由於考慮到資料庫中有部分欄位不具分析價值，在了解每個資料欄位所代表的含意與其使用之價值後，將不具分析價值的欄位予以刪減，以避免造成分析時的負擔與影響資料採礦分析結果之準確性。如操行成績，由於研究對象之操行成績具基本設定，除少數幾位學生成績有所不同之外，彼此間差異不大，故在此予以刪除。此外，刪除就學狀態為「取消入學」者，因其原本錄取然從未至校註冊，故雖有其基本資料，但實際上並不算本校學生，故予以刪除。

本節就資料之前處理及後續分析之使用指標分別說明如下：

壹、資料前置處理主要包括：

一、 資料淨化：主要是確認資料的完整性及正確性。將用不到的資料欄位予以刪除，如學生姓名、身份證字號、電話、地址等等。

二、 資料轉換：為使資料內容更容易資料採礦之進行，將部分資料進行轉換，如居住縣市轉換為居住地區。另依據每學期修課學分及每學期實得學分計算出平均每學期不及格學分數。

三、 資料整合：將學生基本學籍資料表與成績表進行整合，以利進行相關後續分析。

本研究之欄位及其意義分別說明如下：

表 3- 1　學生資料表

| 欄位 | 屬性 | 長度 | 說明 |
|---|---|---|---|
| 學號 | 數字 | 8 | |
| 性別 | 文字 | 1 | 男、女 |
| 科系 | 文字 | 8 | 資訊管理學系<br>資訊工程學系<br>建築與景觀學系 |
| 班級 | 文字 | 1 | |
| 入學方式 | 文字 | 10 | |
| 入學年度 | 文字 | 10 | |
| 郵遞區號 | 數字 | 3 | |
| 就學狀態 | 文字 | 2 | |
| 居住縣市 | 文字 | 3 | |
| 居住地區 | 文字 | 2 | |
| 歷年學業平均成績 | 數字 | 5 | |
| 平均每學期不及格學分數 | 數字 | 5 | |

貳、增益（Lift）指標：

「增益（Lift）」指標指的是〝事件在子群的發生機率除以該事件在原始母體之發生機率〞，增益值大於 1 表示事件在子群的發生機率大於該事件在原始母體之發生機率，因此該條件下被視作傾向發生。

其公式為：

Lift 值＝事件在子群之發生機率/該事件在原始母體之發生機率

以「表 3-2　不同性別之就學狀態」中，男生之退學為例來說：

Lift 值=（男生退學人數/總退學人數）/（男生總人數/全體總人數）

＝(64/77)/(522/77)=1.12

表 3- 2　不同性別之就學狀態

| | | 就學狀態 | | | | | 總和 |
|---|---|---|---|---|---|---|---|
| | | 休學 | 在學 | 校內 | 退學 | 轉學 | |
| sex 性別 | 女 | 5 | 132 | 1 | 13 | 31 | 182 |
| | 增益值 | 1.02 | 1.00 | 1.29 | 0.65 | 1.3 | |
| | 男 | 14 | 381 | 2 | 64 | 61 | 522 |
| | 增益值 | 0.99 | 1.00 | 0.90 | 1.12 | 0.89 | |
| 總和 | | 19 | 513 | 3 | 77 | 92 | 704 |

參、 分類矩陣及 3R：

資料採礦的模型評估指標根據分類矩陣及 3R 指標，其中分類矩陣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；3R 指標則是用來判斷模型之成效，其說明如下：

一、 Response Rate 回應率：在我們的預測名單中找出多少稀有事件？

二、 Recall 反查：預測出來的稀有事件佔總體稀有事件多少比例？

三、 Range Reduce 間距縮減：透過資料採礦模型來找尋稀有事件時，名單縮小了多少？

若是套用在本研究，回應率指的是我們「預測」的學生流失名單中，有多少學生最後真的流失了，回應率愈高，則代表模型愈好，它的計算方式，以「表 3-3 模型訓練組資料 1：1 之分類矩陣」為例：

一、 Response Rate 回應率＝預測會流失實際已流失/所有預測已流失的學生＝43/（43+74）＝36.75%，至於 36.75%是高還是低，我們必須要跟總體的回應率作比較：

總體 Response Rate 回應率＝所有實際已流失/全體學生＝（43+13）/（43+13+74+80）＝26.67％，我們可以發現原始回應率為 26.67％，運用訓練組資料 1：1，COMPLEXITY_PENALTY=0.5，MINIMUM_SUPPORT=10 之決策樹模型提升為 36.75％，因此回應率提升了 1.37 倍。

二、Recall＝預測會流失實際已流失/所有實際已流失的學生＝43/（43+13）＝76.79％，完美的預測模型反查會是 100％，但是通常反查會與回應率互斥，當我希望回應率愈高，因此就必須把回應率的門檻提高，而排除回應率低的名單，相對的會造成反查的降低。（尹相志，2007）

三、Range Reduce＝所有預測已流失的學生/全體學生＝（43+74）/（43+13+74+80）＝55.71％

我們可以發現，運用這個預測模型可以讓學生名單縮減至原本的 55.71％，但是卻包含了總體 78.78％會流失的學生（反查），讓回應率提升了原先的 1.37 倍。

表 3-3　模型訓練組資料 1：1 之分類矩陣

| 模型訓練組資料 1：1 | 預測 | 實際已流失 | 實際未流失 |
|---|---|---|---|
| 訓練組資料 1：1，COMPLEXITY_PENALTY=0.5，MINIMUM_SUPPORT=10 之決策樹模型 | 已流失 | 43 | 74 |
| | 未流失 | 13 | 80 |
| | 正確總數、百分比 | 125，58.57％ | |
| | 錯誤總數、百分比 | 87，41.42％ | |

## 第四節　資料採礦工具

本研究採用 SQL Server2005 為資料採礦工具。SQL Server 2005 不僅擁有決策樹與群集演算法之外，更有類神經網路、線性迴歸、羅吉斯迴歸、貝氏機率分類、關聯規則、時序群集以及時間序列等演算法，而其豐富的視覺化呈現不僅能讓分析者深入瞭解模型規則的內容，更可透過

互動的機制，讓使用者深入瞭解潛藏在模型中的趨勢。此外，SQL Server 2005 提供分類矩陣(Classification Matrix)、增益圖(Lift Chart)、利潤圖(Profit Chart)以及散布圖(Scatter Plot)等資料模型評估工具。

SQL Server 2005 為 IT 專業人員以及資訊工作者提供了強大而又熟悉的工具，降低在行動裝置、企業資料系統或其他平台上建立、部署、管理和使用企業資料及分析應用程式的複雜性。透過豐富的功能集、與現有系統的互通性，以及例行工作的自動化，SQL Server 2005 為各種規模的企業提供完整的資料解決方案。

SQL Server 2005 是一個功能完備的資料庫平台，利用整合式商業智慧 (BI) 工具，提供企業級資料管理功能。SQL Server 2005 資料庫引擎提供更安全、可靠的儲存環境給關聯式和結構式資料，能夠建置並管理用於企業的高可用性、高效能資料應用程式。這個新一代的資料管理和分析解決方案，為企業資料和分析應用程式提供更高的安全性、延展性和可用性，同時使它們更容易建置、部署和管理。

SQL Server 2005 提供整合式資料管理和分析解決方案，可幫助任何規模的組織執行下列工作：

壹、建置、部署和管理更安全、可延展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，以便連接內部和外部系統。

肆、控制成本，同時兼顧效能、可用性、延展性和安全性。

# 第四章 資料分析

本章共有四節，第一節資料特性分析，針對所取得之資料進行一描述性統計，說明所取得資料之概況，第二節決策樹分析，針對所取得之資料進行決策樹分析，第三節為類神經網路分析。

## 第一節 資料特性分析

壹、資料說明：

由於研究樣本之各項分佈比例不甚平均，若使用百本分比之表示方式，並無法展現其在各自子群中所佔之比例，舉例來說，研究樣本中之性別比例原本分佈並不甚平均，若視其流失之百分比，則原本所佔比例較高之性別，其流失之人數與比例容易相對較高，但以增益值計算，增益值較高的性別並不見得就是流失人數最多的，流失增益值較高者，代表其傾向流失。故在此以「增益值」為主的方式表達，比較的基準不是絕對的比率，而是相對的增益。

貳、就學狀態分析：

本研究共取得分析樣本 704 筆，由「表 4-1　學生就學狀態表」，可知其中目前就學狀態為休學者有 19 筆，在學有 513 筆，校內轉系有 3 筆，退學佔 77 筆，轉學佔 92 筆。由其百分比欄位可看出研究樣本之學生就學狀態為休學、退學、轉學者，合計高達 26.7%。

表 4- 1　學生就學狀態表

| 就學狀態 | 次數 | 百分比 |
|---|---|---|
| 休學 | 19 | 2.7 |
| 在學 | 513 | 72.9 |
| 校內轉系 | 3 | .4 |
| 退學 | 77 | 10.9 |
| | 92 | 13.1 |

| | | |
|---|---|---|
| 轉學<br>總和 | 704 | 100.0 |

由「圖 4-1　學生就學狀態分佈圖」可以看出，研究樣本之學生就學狀態比例。

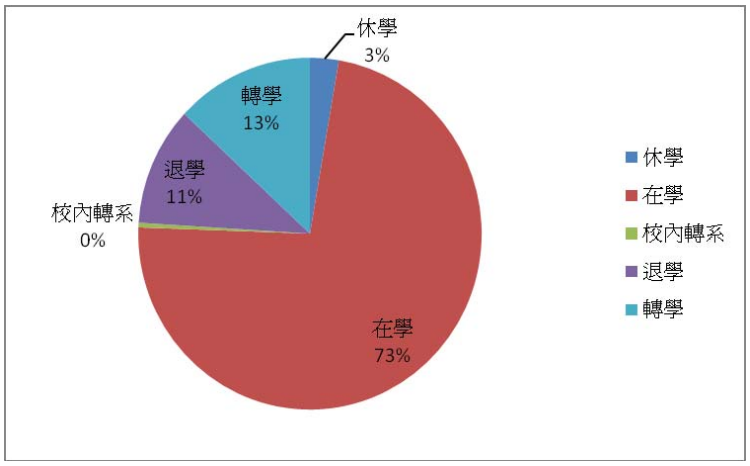


圖 4-1　學生之就學狀態分佈圖

肆、性別資料分析：

在「圖 4-2　學生性別分佈圖」中可看出，所取得之樣本資料，女生佔 182 筆，男生佔 522 筆，可看出研究樣本所佔性別比例極不平均的狀況。

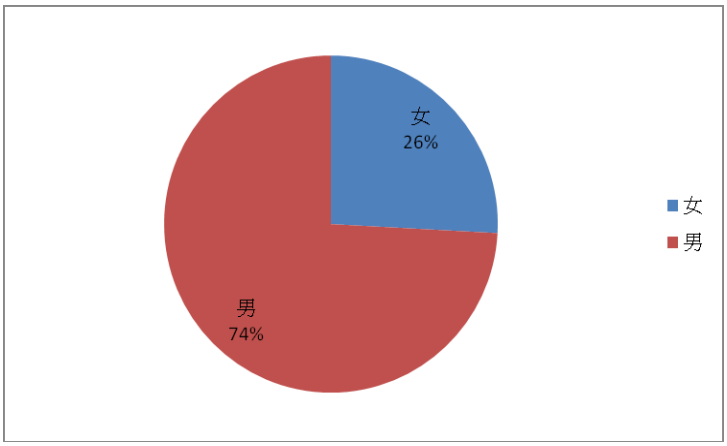


圖 4-2　學生性別比例分佈圖

由「圖 4-3　女生之就學狀態分佈」顯示，女生之流失比例以轉學為最多，其次為退學，而與「圖 4-4　男生之就學狀態分佈」比較，發現男女生之在學比例相當，而男生之流失比例以退學及轉學較高。

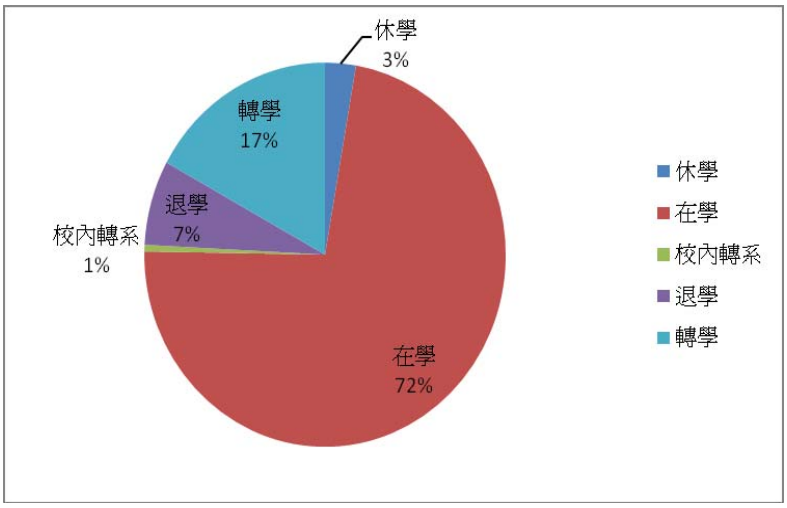


圖 4-3　女生之就學狀態分佈

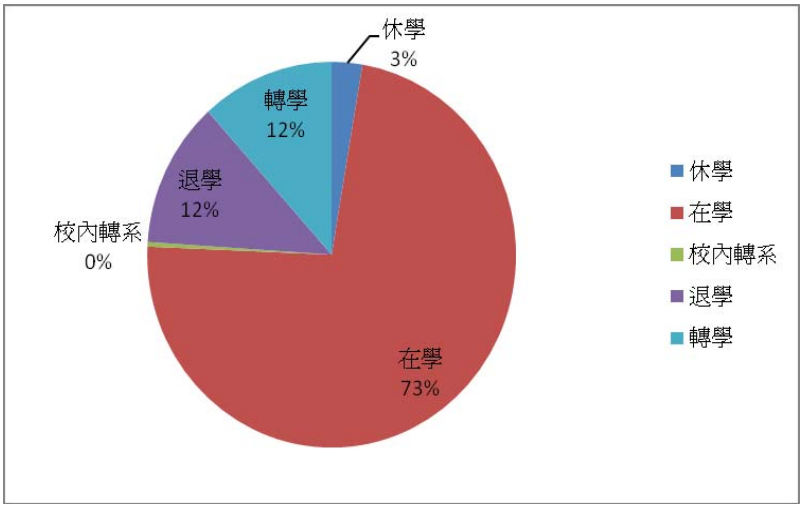


圖 4-4　男生之就學狀態分佈

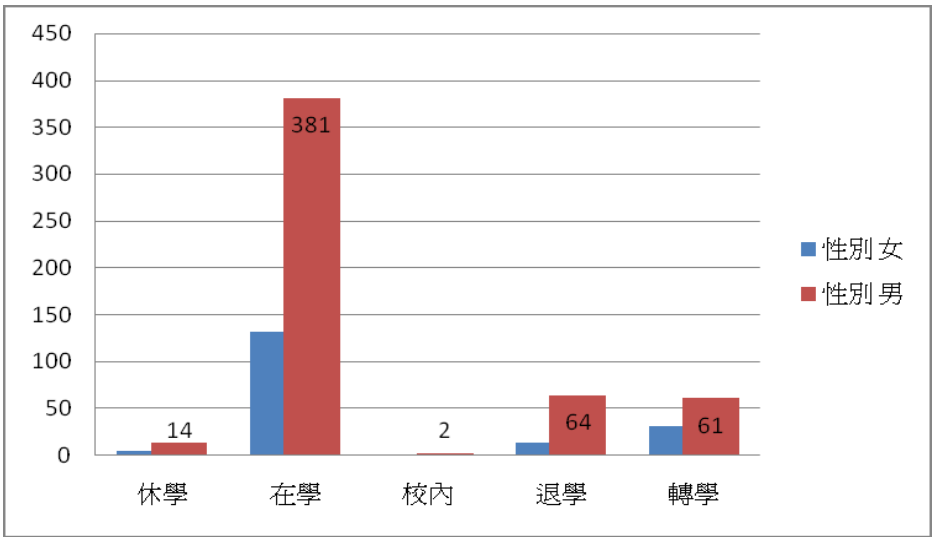


圖 4- 5　不同性別其就學狀態圖

由「圖 4-5　不同性別其就學狀態圖」可以發現，在學人數雖以男生為多，但休學、退學、轉學人數也以男生為多，男生之流失情形似乎較為嚴重，然而，詳細觀察「表 4-2　不同性別其就學狀態表」並計算其個別之「增益(Lift)值」，由於增益值愈高表示傾向發生的機率就愈大，則可以發現，就退學方面來講，以男生之增益值較高，而就轉學方面來講則是女生增益值較高。

表 4- 2　不同性別其就學狀態表

| | | 就學狀態 | | | | | 總和 |
|---|---|---|---|---|---|---|---|
| | | 休學 | 在學 | 校內轉系 | 退學 | 轉學 | |
| sex 性別 | 女 | 5 | 132 | 1 | 13 | 31 | 182 |
| | 增益值 | 1.02 | 1.00 | 1.29 | 0.65 | 1.3 | |
| | 男 | 14 | 381 | 2 | 64 | 61 | 522 |
| | 增益值 | 0.99 | 1.00 | 0.90 | 1.12 | 0.89 | |
| 總和 | | 19 | 513 | 3 | 77 | 92 | 704 |

肆、系別資料分析：

如「表 4-3　不同學系之就學狀態表」所示，研究樣本中建築與景觀學系之樣本佔 126 筆，資訊工程學系樣本佔 235 筆，資訊管理學系樣本佔 343 筆。其中資訊工程學系之轉學人數為 33 人、退學人數

為 31 人，資訊管理學系其轉學人數為 44 人，退學人數為 37 人，建築與景觀學系其轉學人數為 15 人、退學人數為 9 人。

分別比較其增益值，在就學狀態為休學時，建景系之增益值大於資管系；而就學狀態為退學時，則以資工系之增益值大於資管系大於建景系；而就學狀態為轉學時，則資工系之增益值大於資管系大於建景系。

表 4- 3　不同學系之就學狀態表

| | | | 就學狀態 | | | | | 總和 |
|---|---|---|---|---|---|---|---|---|
| | | | 休學 | 在學 | 校內轉系 | 退學 | 轉學 | 百分比 |
| 學系 | 建築與景觀學 | 個數 | 5 | 96 | 1 | 9 | 15 | 126 |
| | | 增益值 | 1.47 | 1.05 | 1.86 | 0.65 | 0.91 | 17.9% |
| | 資訊工程學系 | 個數 | 3 | 168 | 0 | 31 | 33 | 235 |
| | | 增益值 | 0.47 | 0.98 | 0 | 1.21 | 1.07 | 33.4% |
| | 資訊管理學系 | 個數 | 11 | 249 | 2 | 37 | 44 | 343 |
| | | 增益值 | 1.19 | 1.00 | 1.37 | 0.99 | 0.98 | 48.7% |
| 總和 | | 個數 | 19 | 513 | 3 | 77 | 92 | 704 |

由「圖 4-6」-「圖 4-8」可看出研究樣本中，各系學生之就學狀態分佈情形，可以發現三系中，其退學及轉學的比例都相當大。

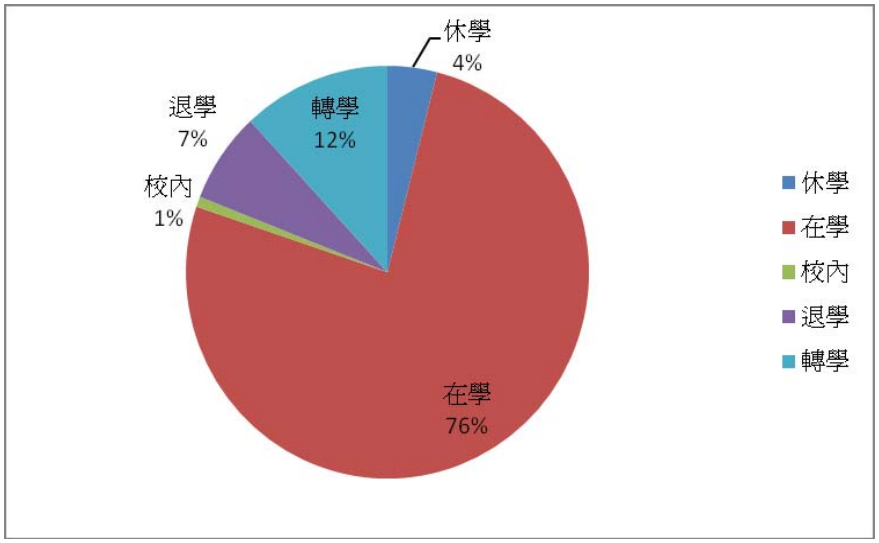


圖 4- 6　建築與景觀學系學生就學狀態分佈圖

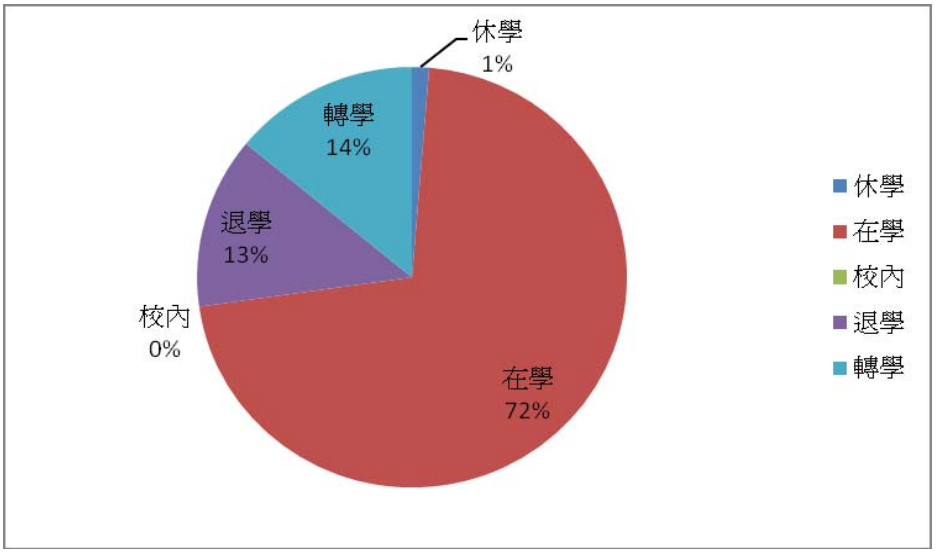


圖 4-7　資訊工程學系學生就學狀態分佈圖

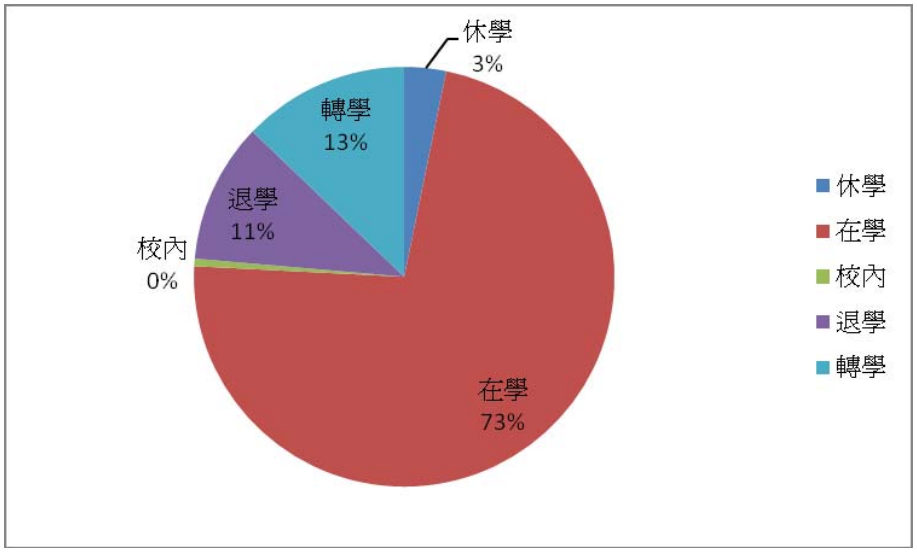


圖 4-8　資訊管理學系學生就學狀態分佈圖

伍、資料異動學年學期資料分析：

由「表 4-4　入學學年學期*異動學年學期交叉表」，其中入學學年學期表示學生的入學年度學期，多數為第一學期，僅 96 學年有第 2 學期的資料，因為自 96 學年開始開放寒假轉學，故其入學年度為第 2 學期，在此因其僅有 3 筆資料，且為新施行之入學管道，無法與

94 及 95 學年度相比較，故在此項資料異動學年學期資料分析時予以刪除。

異動學年學期表示學生資料有所異動，諸如休學、退學、轉學，皆稱為異動，以入學學年學期為 94 學年第 1 學期而言，其資料異動在 951 之筆數為 33，明顯比其餘學年學期之比數來得多，同樣的情形發生在入學學年學期為 95 學年第 1 學期，其資料異動在 961 之筆數為 32，明顯比其餘學年學期之比數來得多，入學學年學期為 96 學年第 1 學期的則為 971 資料異動筆數較多，此顯示入學新生在一年級升二年級時，比其餘時間更容易發生資料異動，所以資料異動時間多數為二年級上學期。

表 4-4　入學學年學期 *　異動學年學期 交叉表

| | | 異動學年學期 | | | | | | | | | 總和 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 在學 | 94學年度第1學期 | 94學年度第2學期 | 95學年度第1學期 | 95學年度第2學期 | 96學年度第1學期 | 96學年度第1學期 | 97學年度第1學期 | 97學年度第2學期 | |
| 入學學年學期 | 094學年第 1 學期 | 152 | 4 | 5 | 33 | 5 | 7 | 0 | 1 | 0 | 207 |
| | 095學年第 1 學期 | 157 | 0 | 0 | 3 | 11 | 32 | 9 | 17 | 2 | 231 |
| | 096學年第 1 學期 | 188 | 0 | 0 | 0 | 0 | 4 | 12 | 46 | 13 | 263 |
| | 總和 | 497 | 4 | 5 | 36 | 16 | 43 | 21 | 65 | 15 | 701 |

由「圖4-9　資料異動時間趨勢圖」亦可看出，資料異動的時間在每學年第一學期達到高峰，94學年第1學期入學的學生，資料異動在95學年第1學期達到高峰；95學年第1學期入學的學生，資料異動在96學年第1學期達到高峰；96學年第1學期入學的學生，資料異動在97

學年第1學期達到高峰，由此顯示，學生在入學後一年，升二年級時容易產生資料異動。

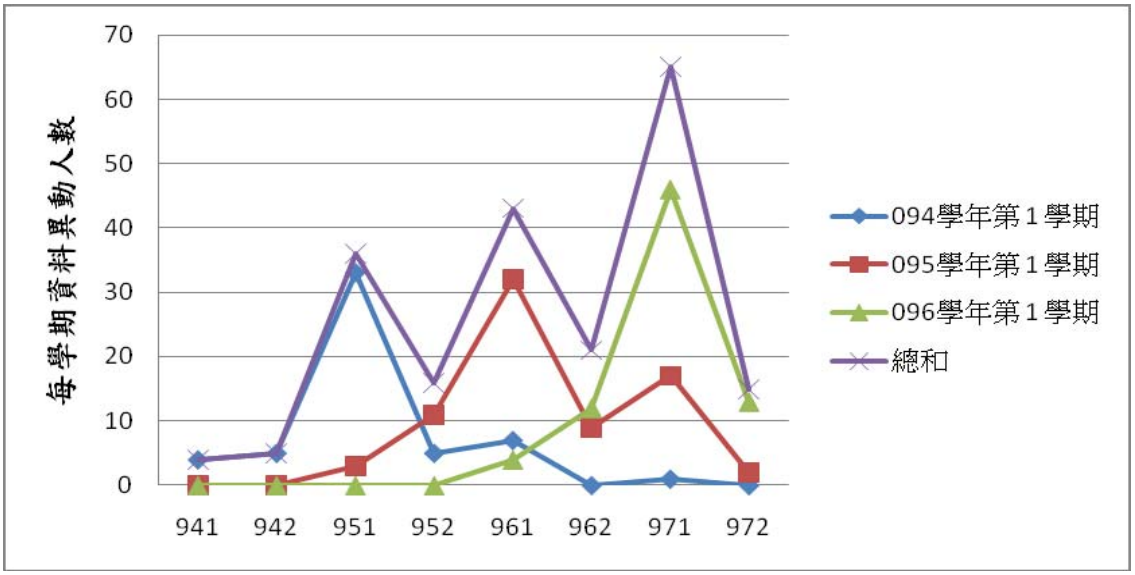


圖 4-9　資料異動時間趨勢圖

陸、入學方式資料分析：

在分析樣本中，入學方式以考試分發464筆佔最多，第二為個人申請佔118筆，第三為轉學考佔87筆，由「圖4-10　入學方式分佈圖」即可明顯看出其比例之差異。

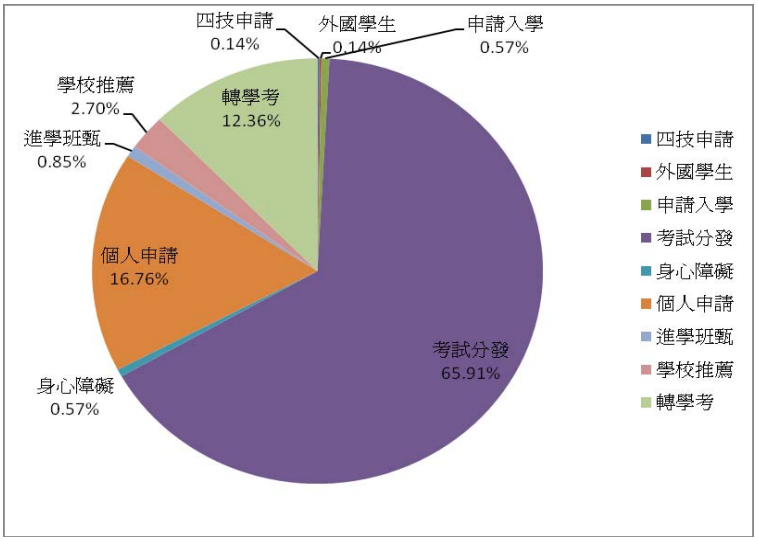


圖 4-10　入學方式分佈圖

由「表 4-5　就學狀態*入學方式交叉表」中可發現，就學狀態為休學者，其入學方式為轉學考的增益值為 2.98，大於入學方式為考試分發者，顯示轉學生傾向休學的程度很高；此外，就學狀態為退學者，其入學方式以考試分發及轉學考的增益值較高，顯示兩者都有退學的傾向；就學狀態為轉學者，其入學方式以考試分發及個人申請的增益值較高，表示兩者都傾向發生轉學。

表 4-5　就學狀態 * 入學方式 交叉表

| | | 就學狀態 | | | | | 總和 |
|---|---|---|---|---|---|---|---|
| | | 休學 | 在學 | 校內 | 退學 | 轉學 | |
| Application 入學方式 | 四技申請 | 0 | 1 | 0 | 0 | 0 | 1 |
| | 外國學生 | 0 | 1 | 0 | 0 | 0 | 1 |
| | 申請入學 | 0 | 2 | 0 | 0 | 2 | 4 |
| | 考試分發 | 8 | 338 | 3 | 50 | 65 | 464 |
| | 增益值 | 0.64 | 1.00 | | 0.99 | 1.07 | |
| | 身心障礙 | 0 | 3 | 0 | 1 | 0 | 4 |
| | 個人申請 | 2 | 94 | 0 | 5 | 17 | 118 |
| | 增益值 | | 1.09 | | | 1.10 | |
| | 進學班甄 | 2 | 4 | 0 | 0 | 0 | 6 |
| | 學校推薦 | 0 | 16 | 0 | 2 | 1 | 19 |
| | 增益值 | | 1.16 | | | | |
| | 轉學考 | 7 | 54 | 0 | 19 | 7 | 87 |
| | 增益值 | 2.98 | 0.85 | | 2.00 | 0.62 | |
| 總和 | | 19 | 513 | 3 | 77 | 92 | 704 |

由「圖 4-11 就學狀態為休學者之入學方式」可看出學生就學狀態為休學者，其入學方式以考試分發居多，其次為申請入學，之後是個人申請及進學班甄選。

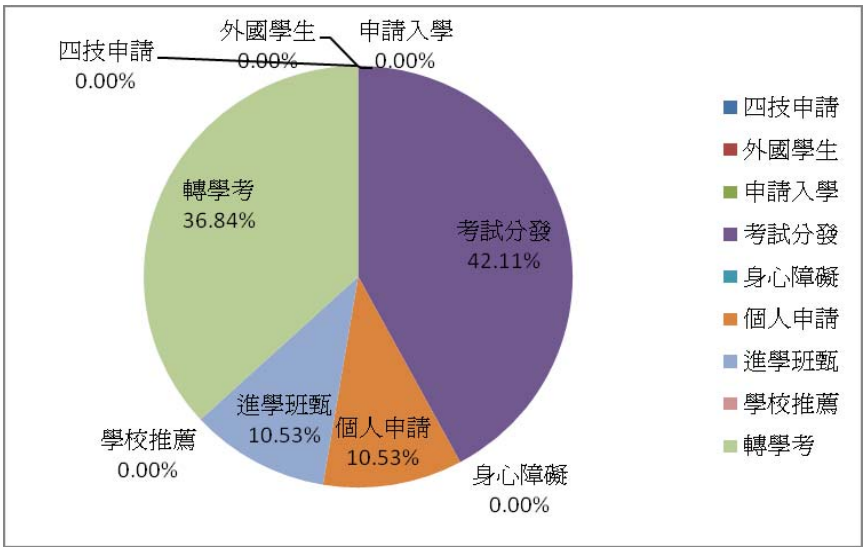


圖 4- 11　就學狀態為休學者之入學方式

由「圖 4-12 就學狀態為轉學者之入學方式」可看出，學生就學狀態為轉學者（轉出去），其入學方式以考試分發居多，其次為個人申請，之後是轉學考。

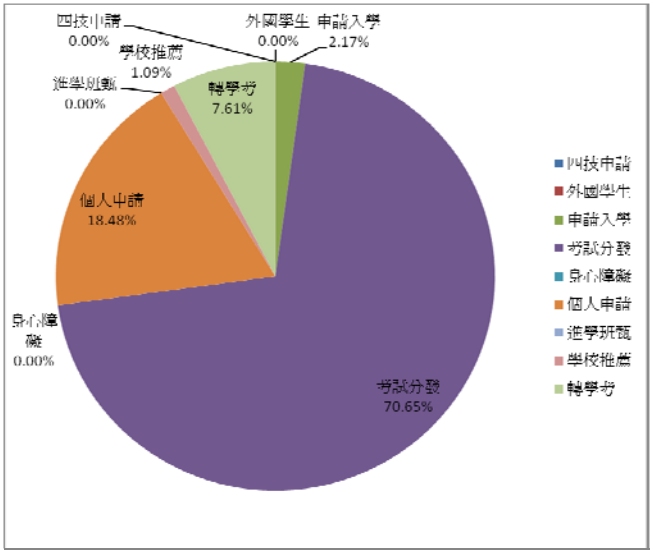


圖 4- 12　就學狀態為轉學者之入學方式

由「圖 4-13 就學狀態為退學者之入學方式」可看出，學生就學狀態為退學者，其入學方式以考試分發居多，其次為轉學考，之後是個人申請。

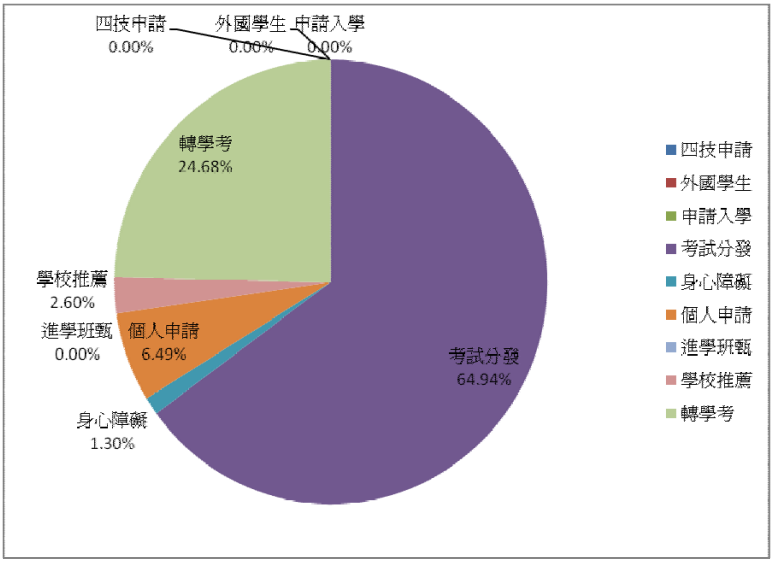


圖 4- 13　就學狀態為退學者之入學方式

由「圖 4-14 就學狀態為在學者之入學方式」可看出，學生就學狀態為在學者，其入學方式以考試分發居多，其次為個人申請，之後是轉學考。

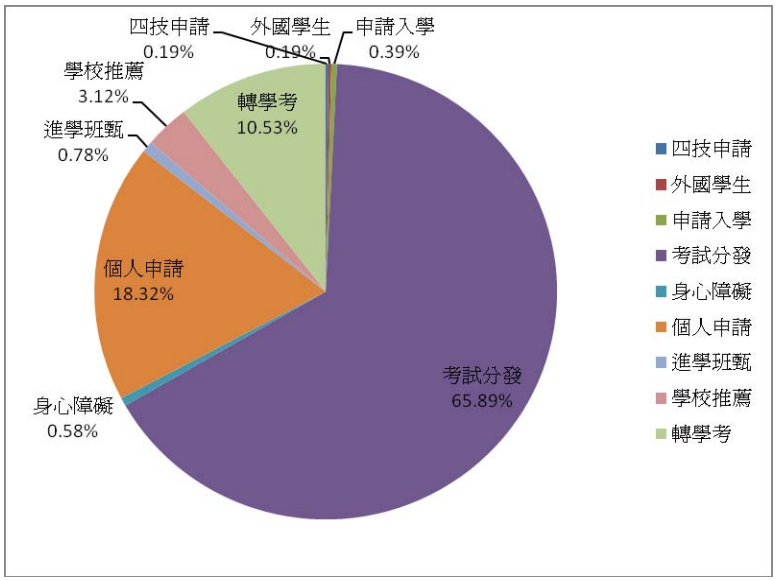


圖 4- 14　就學狀態為在學者之入學方式

柒、居住地區資料分析：

研究樣本中，居住地區以中區、北區及南區為多，其中又以南區為最多，共計有273筆，如「表4-6　居住地區次數分配表」及「圖4-15 居住地區長條圖」所示。

表 4- 6　居住地區次數分配表

| 居住地區 | 次數 |
|---|---|
| 外國 | 1 |
| 中區 | 199 |
| 北區 | 205 |
| 外島 | 6 |
| 東區 | 20 |
| 南區 | 273 |
| 總和 | 704 |

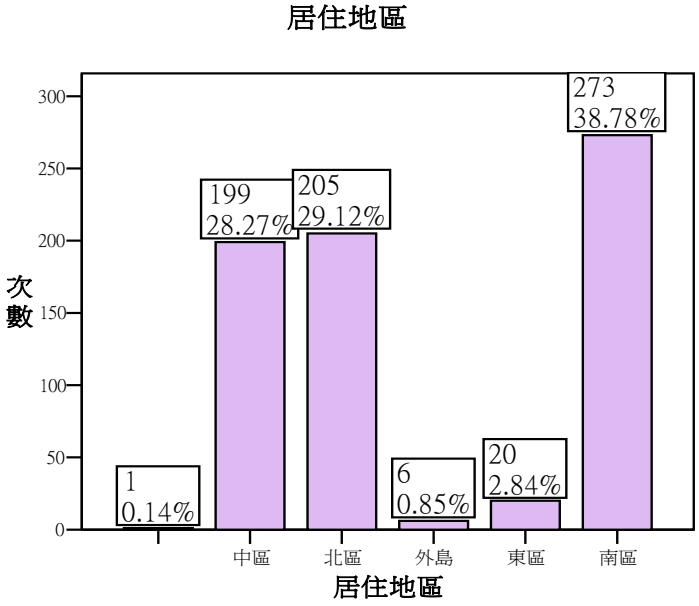


圖 4- 15　居住地區長條圖

由「圖 4-16　居住縣市長條圖」可以發現，學生以高雄市、屏東縣、雲林縣、台北縣、台北市等縣市為最多，反而嘉義縣市地區人數並沒有很多。

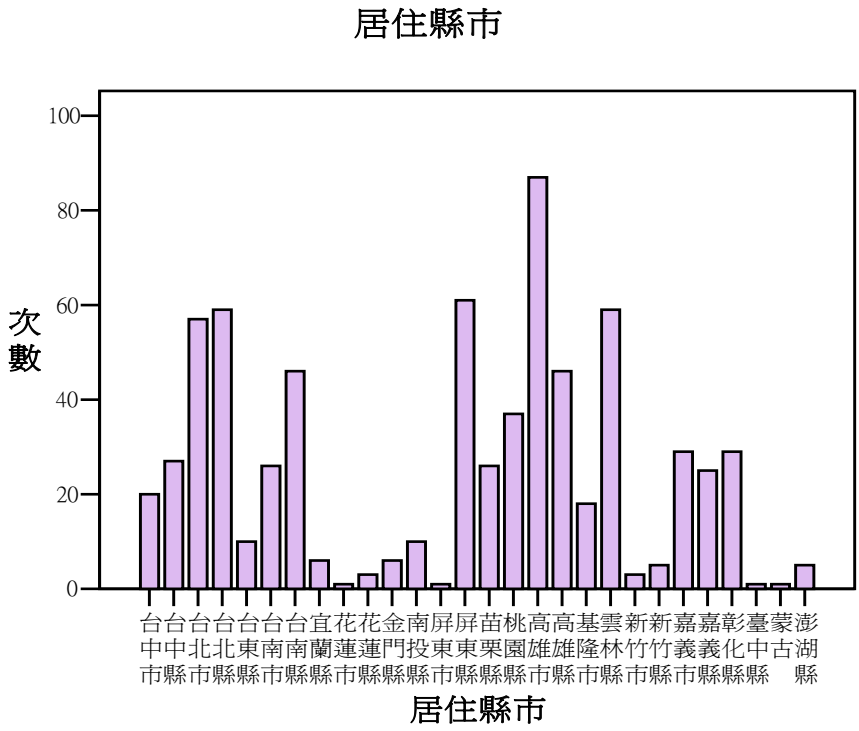


圖 4- 16　居住縣市長條圖

由「表 4-7 居住縣市增益值」可以發現，雖然學生以居住高雄市、屏東縣、雲林縣、台北縣、台北市等縣市為最多，然台北縣市及台南縣之已流失增益值也明顯大於其他縣市之增益值，表示其傾向已流失。

表 4- 7　居住縣市增益值

| | 休學 | 在學 | 校內 | 退學 | 轉學 | 就學狀態總計 | 已流失 | 未流失 | 已流失增益直 | 未流失增益值 |
|---|---|---|---|---|---|---|---|---|---|---|
| 蒙古 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0.00% | 136.43% |
| 台中市 | 0 | 15 | 0 | 1 | 4 | 20 | 5 | 15 | 93.62% | 102.33% |
| 台中縣 | 1 | 20 | 0 | 4 | 3 | 28 | 8 | 20 | 106.99% | 97.45% |
| 南投縣 | 0 | 9 | 0 | 1 | 0 | 10 | 1 | 9 | 37.45% | 122.79% |
| 雲林縣 | 2 | 45 | 0 | 5 | 7 | 59 | 14 | 45 | 88.86% | 104.06% |
| 嘉義市 | 1 | 22 | 0 | 3 | 3 | 29 | 7 | 22 | 90.39% | 103.50% |
| 嘉義縣 | 0 | 17 | 0 | 5 | 3 | 25 | 8 | 17 | 119.83% | 92.78% |
| 彰化縣 | 1 | 22 | 0 | 1 | 5 | 29 | 7 | 22 | 90.39% | 103.50% |
| 台北市 | 0 | 38 | 0 | 10 | 9 | 57 | 19 | 38 | 124.82% | 90.96% |
| 台北縣 | 3 | 41 | 0 | 8 | 7 | 59 | 18 | 41 | 114.24% | 94.81% |
| 苗栗縣 | 1 | 22 | 0 | 0 | 3 | 26 | 4 | 22 | 57.61% | 115.44% |
| 桃園縣 | 1 | 26 | 0 | 4 | 6 | 37 | 11 | 26 | 111.33% | 95.87% |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 基隆市 | 2 | 11 | 0 | 5 | 0 | 18 | 7 | 11 | 145.63% | 83.38% |
| 新竹市 | 0 | 1 | 0 | 1 | 1 | 3 | 2 | 1 | 249.65% | 45.48% |
| 新竹縣 | 0 | 1 | 0 | 0 | 4 | 5 | 4 | 1 | 299.57% | 27.29% |
| 金門縣 | 0 | 4 | 0 | 1 | 1 | 6 | 2 | 4 | 124.82% | 90.96% |
| 台東縣 | 0 | 7 | 0 | 2 | 1 | 10 | 3 | 7 | 112.34% | 95.50% |
| 宜蘭縣 | 0 | 4 | 0 | 1 | 1 | 6 | 2 | 4 | 124.82% | 90.96% |
| 花蓮市 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0.00% | 136.43% |
| 花蓮縣 | 0 | 1 | 0 | 1 | 1 | 3 | 2 | 1 | 249.65% | 45.48% |
| 台南市 | 0 | 17 | 0 | 7 | 2 | 26 | 9 | 17 | 129.62% | 89.21% |
| 台南縣 | 3 | 32 | 0 | 3 | 8 | 46 | 14 | 32 | 113.97% | 94.91% |
| 屏東市 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0.00% | 136.43% |
| 屏東縣 | 1 | 53 | 1 | 3 | 3 | 61 | 7 | 54 | 42.97% | 120.78% |
| 高雄市 | 2 | 64 | 1 | 6 | 14 | 87 | 22 | 65 | 94.69% | 101.93% |
| 高雄縣 | 1 | 35 | 1 | 4 | 5 | 46 | 10 | 36 | 81.41% | 106.77% |
| 澎湖縣 | 0 | 3 | 0 | 1 | 1 | 5 | 2 | 3 | 149.79% | 81.86% |
| 總計 | 19 | 513 | 3 | 77 | 92 | 704 | 188 | 516 | 100.00% | 100.00% |

## 第二節 決策樹分析

壹、資料說明：

將資料先進行預處理，刪除入學年度為 96 學年度第 2 學期資料 3 筆，總計本研究後續分析資料共 701 筆，增加「流失與否」欄位，在本研究中對流失的界定為：『曾經註冊繳費，擁有學籍資料之本校學生，因故無法完成學業，中途退出，未取得畢業證書之學生』，故將學生就學狀態為在學及校內轉系列為「未流失」，將學生就學狀態為休學、退學、轉學等列為「已流失」，其中「未流失」資料共計 514 筆，「已流失」資料共計 187 筆，「已流失」與「未流失」之比例約為 1：2.75，此外，將學生入學身份整合為考試分發、個人申請、學校推薦、轉學及其他等。

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 4-8　決策樹分析輸入欄位」及「表 4-9　決策樹分析預測欄位」所示：

表 4- 8　決策樹分析輸入欄位

| 欄位名稱 | 欄位說明 |
|---|---|
| 學號（ID） | |
| 性別（Sex） | 男、女 |
| 科系（DEPART） | 建景、資工、資管 |
| 入學方式（Application） | 考試分發、個人申請、學校推薦、轉學及其他 |
| 居住地區（Area） | 北部、中部、南部、東部、離島 |
| 歷年學業平均成績（Total AV#） | |
| 平均每學期不及格學分數（Lose AV#） | |

表 4- 9　決策樹分析預測欄位

| 欄位名稱 | 欄位說明 |
|---|---|
| 流失與否（dropout or not） | 已流失、未流失 |

貳、 資料採礦作業：

為使後續所建立之模型能獲得驗證，在此將資料分為訓練組及測試組，使用訓練組之資料進行模型之建立，待模型建立後使用測試組之資料進行模型之驗證。

在資料採礦過程中，因為預測欄位值的分佈不甚平均（已流失與未流失分佈不均衡），故會面臨到一個「稀有事件」的問題，在本研究中「稀有事件」指的是學生流失與否欄位為「已流失」，雖然在本研究樣本中，「已流失」資料佔整體比例不至於低到稱為「稀有事件」，然相對於「未流失」來講，仍稱得上是「稀有事件」。

「誤差抽樣（Error-Sampling）」則是處理「稀有事件」最常採用的一種技巧，其基本精神就是不按照原先值的分配等比例抽樣，而將稀有事件透過抽樣的方式將其比重提升。

首先，我們將資料分為訓練組及測試組，其中測試組佔總資料的 30%，即 210 筆資料（其中已流失為 56 筆，未流失為 154 筆，其已流失與未流失之比例為 1：2.75，與原始資料比例相當），將訓練組資料中所含之已流失資料共計 131 筆全部抽出後，再就剩餘之未流失

資料分別抽出 131 筆、262 筆及 360 筆資料，分別組合成含已流失與未流失比例為 1：1、1：2 及 1：2.75（1：2.75 即是扣除測試組資料後的全部資料）之訓練組資料。

資料準備好之後，接著我們進行資料採礦作業，其步驟如下：

一、 於 MS SQL Server Management Studio 中新增資料庫，選取新增之資料庫，按右鍵選擇 Tasks，選擇 Import Data，然後選擇我們存在電腦中的資料庫。

二、 在 SQL Server Business Intelligence Development Studio 中新增一個 Analysis Service 專案，將資料來源設定為剛剛新增的資料庫，設定資料來源檢視，採礦結構設定為決策樹。

三、 分別依訓練組資料 1：1、訓練組資料 1：2 及訓練組資料 1：2.75 建立決策樹模型，在各個模型之中，分別調整其演算法參數，經測試篩選後，刪除模型無差異之組合，以 COMPLEXITY_PENALTY=0.5，MINIMUM_SUPPORT=10、COMPLEXITY_PENALTY=0.1，MINIMUM_SUPPORT=5、COMPLEXITY_PENALTY=0.05，MINIMUM_SUPPORT=5，分別建立模型。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。如「圖 4-17」至「圖 4-25」為訓練組資料所分別建立之不同參數模型。

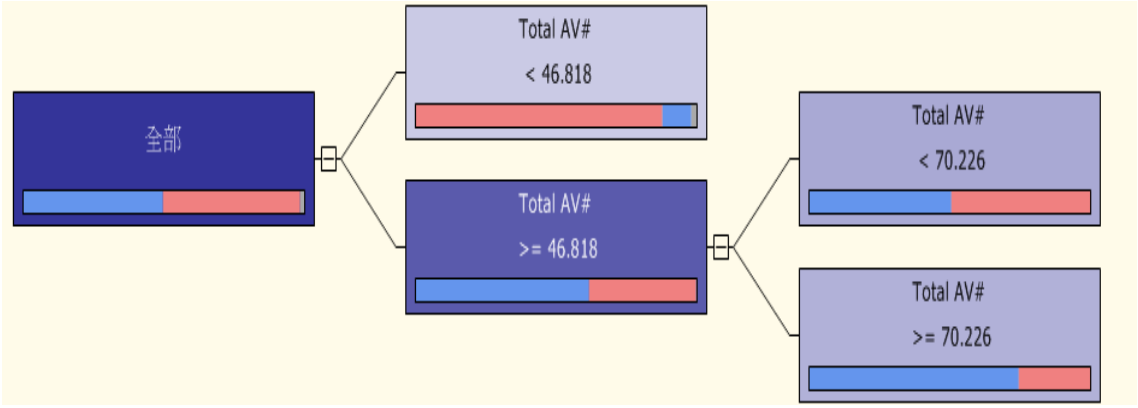


圖 4- 17　訓練組資料 1：1，COMPLEXITY_PENALTY=0.5，MINIMUM_SUPPORT=10 之決策樹模型

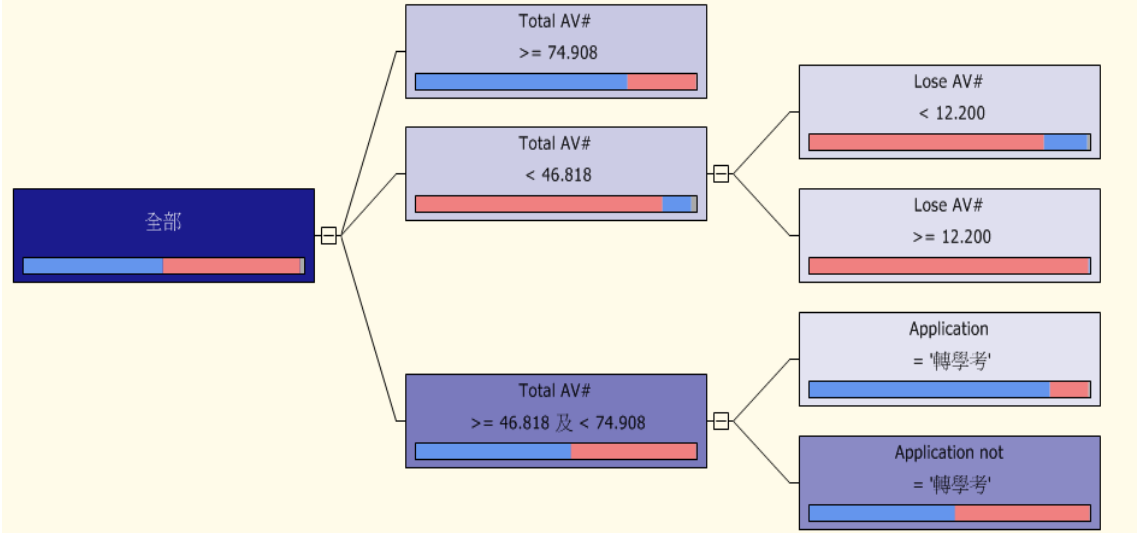


圖 4- 18　訓練組資料 1：1，COMPLEXITY_PENALTY=0.1，MINIMUM_SUPPORT=5 之決策樹模型

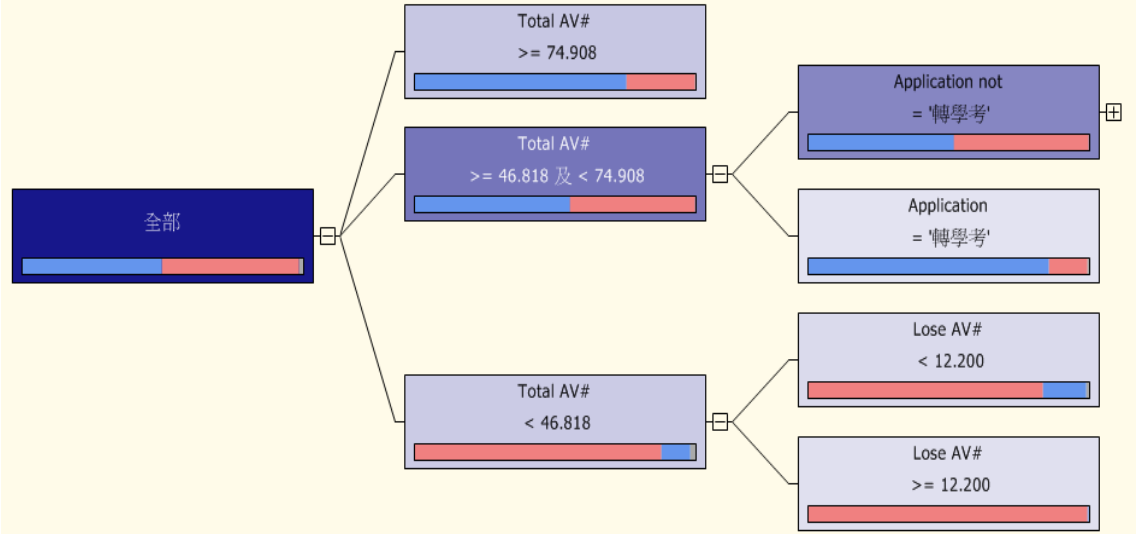


圖 4- 19　訓練組資料 1：1，COMPLEXITY_PENALTY=0.05，MINIMUM_SUPPORT=5 之決策樹模型

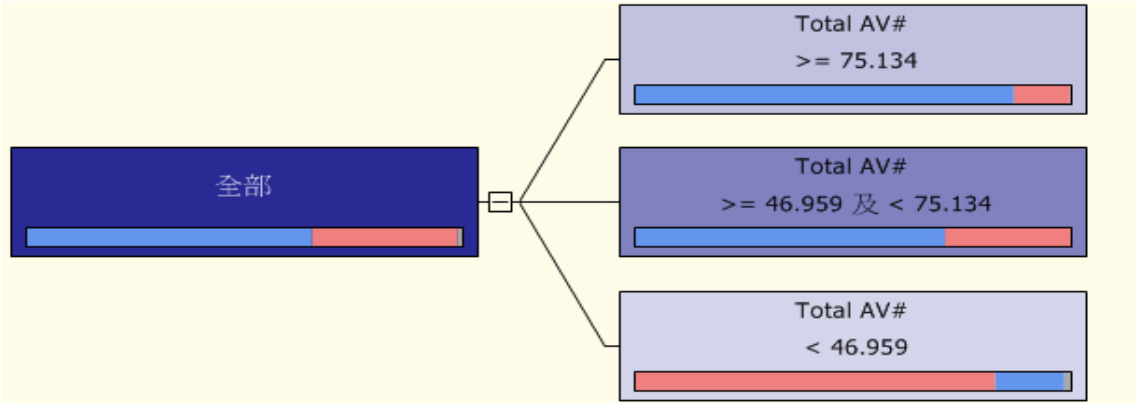


圖 4- 20　訓練組資料 1：2，COMPLEXITY_PENALTY=0.5，MINIMUM_SUPPORT=10 之決策樹模型

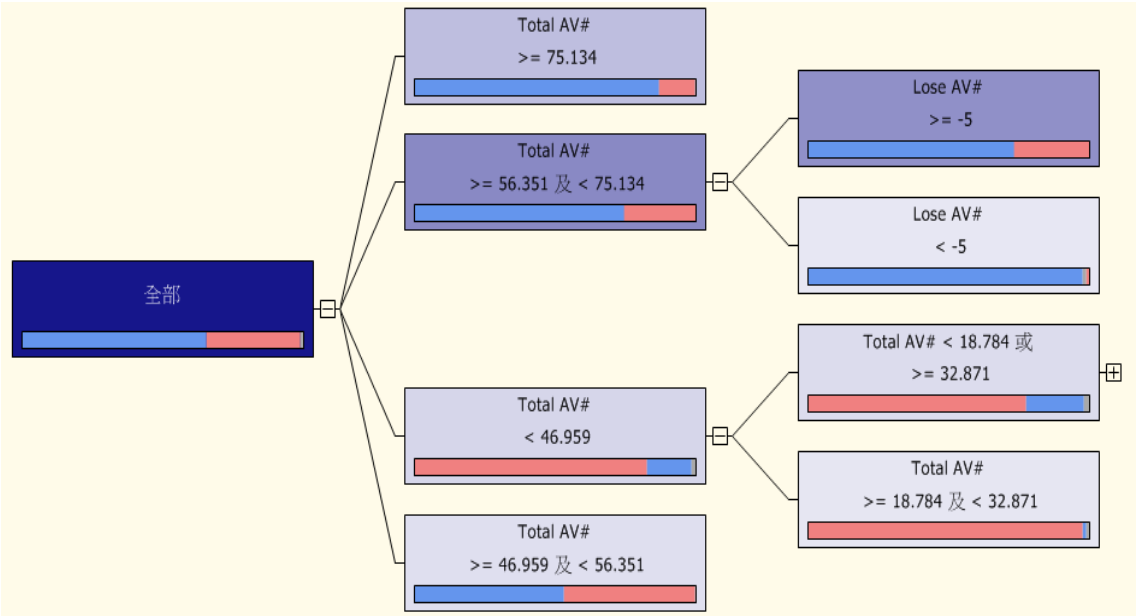


圖 4- 21　訓練組資料 1：2，COMPLEXITY_PENALTY=0.1，MINIMUM_SUPPORT=5 之決策樹模型

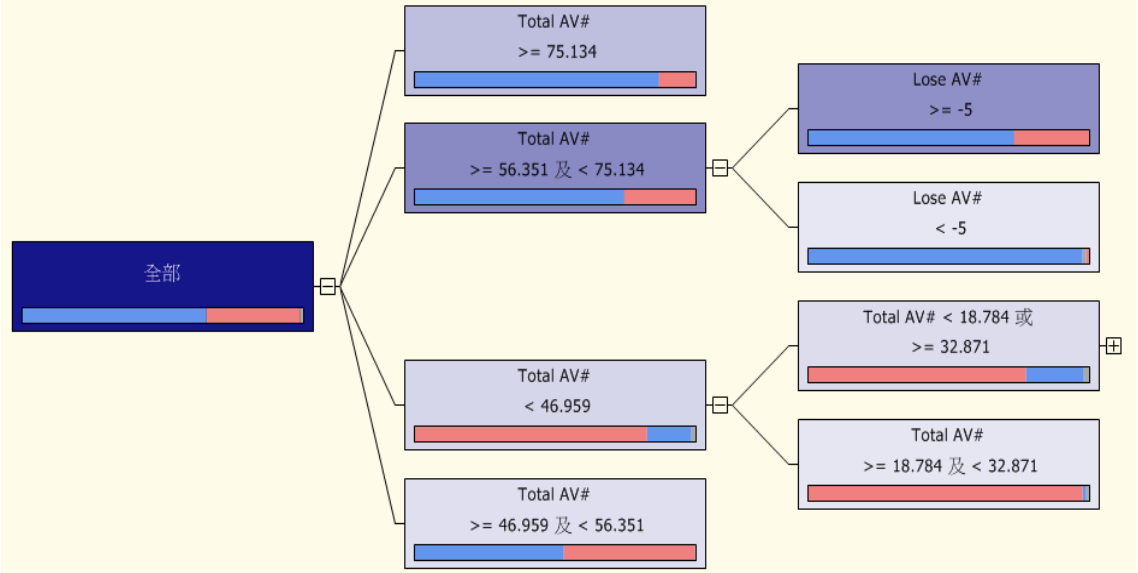


圖 4- 22　訓練組資料 1：2，COMPLEXITY_PENALTY=0.05，MINIMUM_SUPPORT=5 之決策樹模型

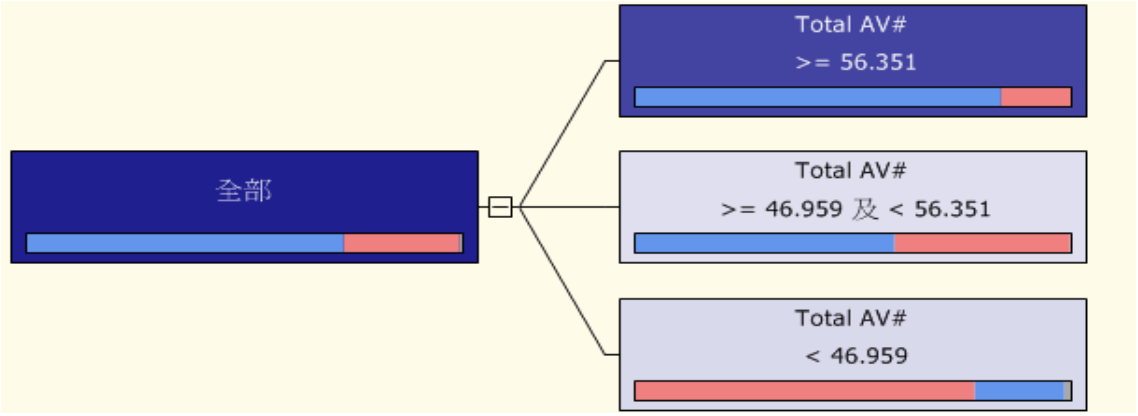


圖 4- 23　訓練組資料 1：2.75，COMPLEXITY_PENALTY=0.5，MINIMUM_SUPPORT=10 之決策樹模型

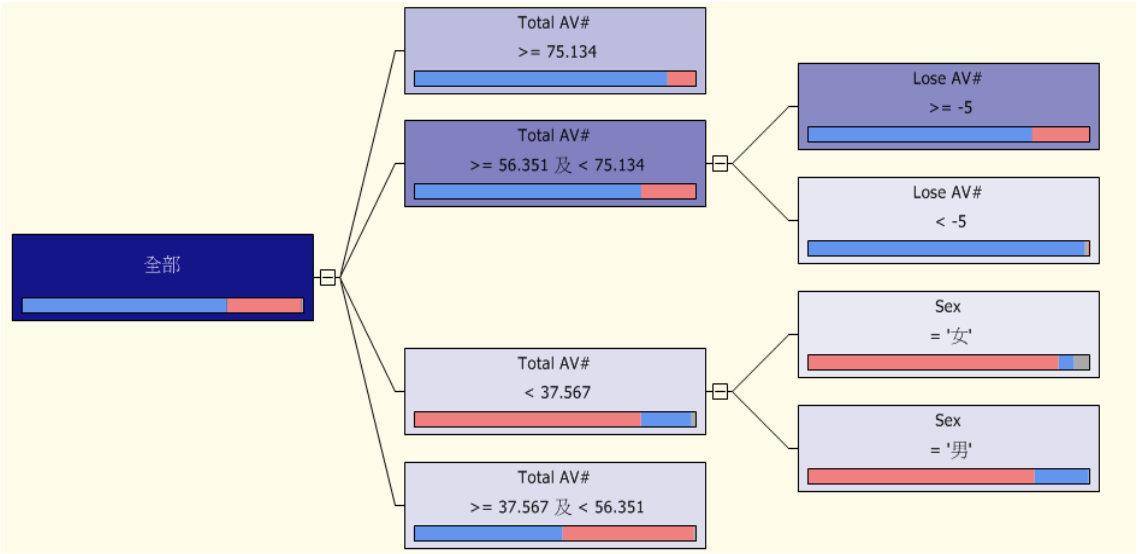


圖 4- 24　訓練組資料 1：2.75，COMPLEXITY_PENALTY=0.1，MINIMUM_SUPPORT=5 之決策樹模型

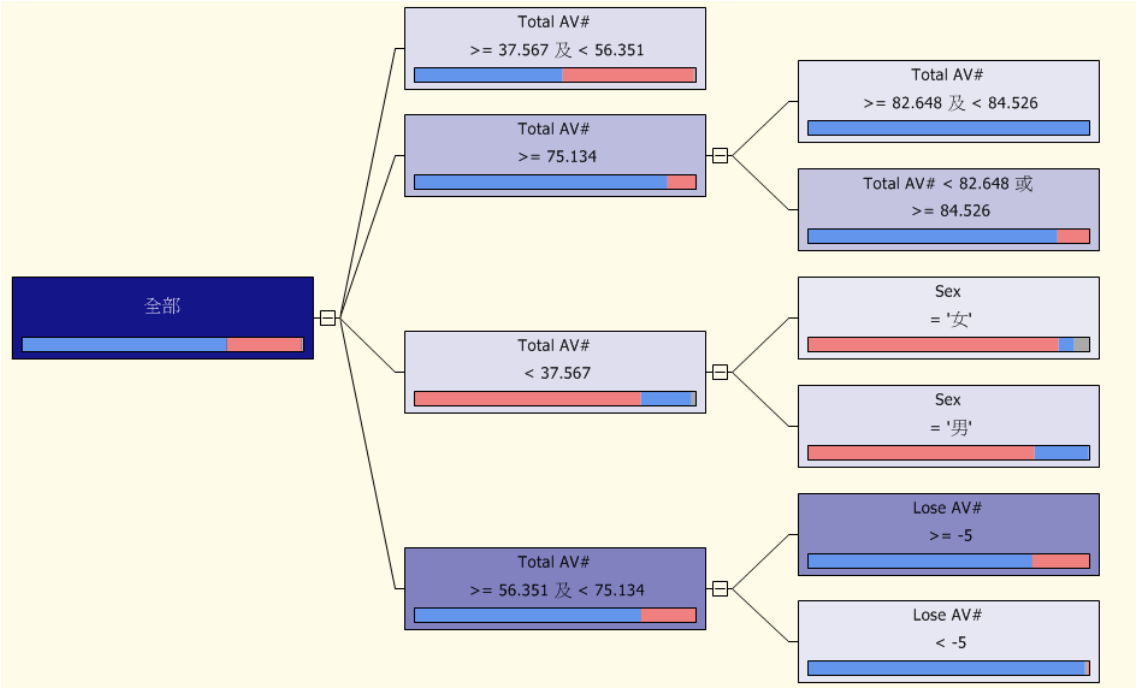


圖 4- 25　訓練組資料 1：2.75，COMPLEXITY_PENALTY=0.05，MINIMUM_SUPPORT=5 之決策樹模型

四、 模型建立後，分別代入測試組資料，進行分別測試。

五、 分別建立分類矩陣，檢視各組資料模型的錯誤分佈，如「表 4-10 各模型之分類矩陣比較」，由於是使用測試組資料進行模型測試，可能由於資料僅 210 筆的關係，雖然各個模型不盡相同，但其分類矩陣數據仍有多處相似，如表中加網底的幾個模型，然根據「錯誤不等價」的觀念，並非正確率愈高的模型愈好，故依據分類矩陣資訊進行進一步分析。

表 4- 10　各模型之分類矩陣比較

| 模型訓練組資料 1：1 | 預測 | 實際已流失 | 實際未流失 |
|---|---|---|---|
| 訓練組資料 1：1，COMPLEXITY_PENALTY=0.5，MINIMUM_SUPPORT=10 之決策樹模型 | 已流失 | 43 | 74 |
| | 未流失 | 13 | 80 |
| | 正確總數、百分比 | 125，58.57% | |
| | 錯誤總數、百分比 | 87，41.42% | |
| 訓練組資料 1：1，COMPLEXITY_PENALTY=0.1，MINIMUM_SUPPORT=5 之決策樹模型 | 已流失 | 22 | 4 |
| | 未流失 | 34 | 150 |
| | 正確總數、百分比 | 172，81.90% | |
| | 錯誤總數、百分比 | 38，18.09% | |
| 訓練組資料 1：1，COMPLEXITY_PENALTY=0.05，MINIMUM_SUPPORT=5 之決策樹模型 | 已流失 | 37 | 65 |
| | 未流失 | 19 | 89 |
| | 正確總數、百分比 | 126，60% | |
| | 錯誤總數、百分比 | 84，40% | |
| 模型訓練組資料 1：2 | 預測 | 實際已流失 | 實際未流失 |
| 訓練組資料 1：2，COMPLEXITY_PENALTY=0.5，MINIMUM_SUPPORT=10 之決策樹模型 | 已流失 | 22 | 4 |
| | 未流失 | 34 | 150 |
| | 正確總數、百分比 | 172，81.90% | |
| | 錯誤總數、百分比 | 37，18.09% | |
| 訓練組資料 1：2，COMPLEXITY_PENALTY=0.1，MINIMUM_SUPPORT=5 之決策樹模型 | 已流失 | 22 | 4 |
| | 未流失 | 34 | 150 |
| | 正確總數、百分比 | 172，81.90% | |
| | 錯誤總數、百分比 | 37，18.09% | |
| 訓練組資料 1：2，COMPLEXITY_PENALTY=0.05，MINIMUM_SUPPORT=5 之決策樹模型 | 已流失 | 22 | 4 |
| | 未流失 | 34 | 150 |
| | 正確總數、百分比 | 172，81.90% | |
| | 錯誤總數、百分比 | 37，18.09% | |
| 模型訓練組資料 1：2.75 | 預測 | 實際已流失 | 實際未流失 |

| | | | |
|---|---|---|---|
| 訓練組資料 1：2.75，COMPLEXITY_PENALTY=0.5，MINIMUM_SUPPORT=10 之決策樹模型 | 已流失 | 22 | 4 |
| | 未流失 | 34 | 150 |
| | 正確總數、百分比 | 172，81.90% | |
| | 錯誤總數、百分比 | 37，18.09% | |
| 訓練組資料 1：2.75，COMPLEXITY_PENALTY=0.1，MINIMUM_SUPPORT=5 之決策樹模型 | 已流失 | 15 | 2 |
| | 未流失 | 41 | 152 |
| | 正確總數、百分比 | 167，79.52% | |
| | 錯誤總數、百分比 | 43，20.47% | |
| 訓練組資料 1：2.75，COMPLEXITY_PENALTY=0.05，MINIMUM_SUPPORT=5 之決策樹模型 | 已流失 | 15 | 2 |
| | 未流失 | 41 | 152 |
| | 正確總數、百分比 | 167，79.52% | |
| | 錯誤總數、百分比 | 43，20.47% | |

六、 分別計算 3R：

由「表 4-11　各種訓練模型之 3R 比較表」中，似乎三種模型都各有所長，實在很難看出哪一個模型比較好，我們另外使用增益圖來進行分析判斷。

表 4- 11　各種誤差抽樣之 3R 比較表%

| | 總體回應率 | Response Rate 回應率 | Recall 反查 | Range Reduce 間距縮減 | 回應率提升 |
|---|---|---|---|---|---|
| 資料意義 | | 愈高愈好 | 愈高愈好 | 愈低愈好 | 愈高愈好 |
| 訓練組資料 1：1，COMPLEXITY_PENALTY=0.5，MINIMUM_SUPPORT=10 之決策樹模型 | 26.67 | 36.75 | 76.79 | 55.71 | 137.82 |
| 訓練組資料 1：1，COMPLEXITY_PENALTY=0.1，MINIMUM_SUPPORT=5 之決策樹模型 | 26.67 | 84.62 | 39.29 | 12.38 | 317.31 |
| 訓練組資料 1：1，COMPLEXITY_PENALTY=0.05，MINIMUM_SUPPORT=5 之決策樹模型 | 26.67 | 36.27 | 66.07 | 48.57 | 136.03 |
| 訓練組資料 1：2，COMPLEXITY_PENALTY=0.5，MINIMUM_SUPPORT=10 之決策樹模型 | 26.67 | 84.62 | 39.29 | 12.38 | 317.31 |

| | | | | | |
|---|---|---|---|---|---|
| 訓練組資料 1：2，COMPLEXITY_PENALTY=0.1，MINIMUM_SUPPORT=5 之決策樹模型 | 26.67 | 84.62 | 39.29 | 12.38 | 317.31 |
| 訓練組資料 1：2，COMPLEXITY_PENALTY=0.05，MINIMUM_SUPPORT=5 之決策樹模型 | 26.67 | 84.62 | 39.29 | 12.38 | 317.31 |
| 訓練組資料 1：2.75，COMPLEXITY_PENALTY=0.5，MINIMUM_SUPPORT=10 之決策樹模型 | 26.67 | 84.62 | 39.29 | 12.38 | 317.31 |
| 訓練組資料 1：2.75，COMPLEXITY_PENALTY=0.1，MINIMUM_SUPPORT=5 之決策樹模型 | 26.67 | 88.28 | 26.79 | 8.10 | 330.88 |
| 訓練組資料 1：2.75，COMPLEXITY_PENALTY=0.05，MINIMUM_SUPPORT=5 之決策樹模型 | 26.67 | 88.24 | 26.79 | 8.10 | 330.88 |

七、 分別建立增益圖：

增益圖是最普遍使用的一種資料採礦模型評估圖，如「圖 4-26　訓練組資料 1：1 之決策樹模型增益圖」，它的橫軸以及縱軸都是由百分比所構成。橫軸百分比代表資料採礦模型根據機率從高至低排序後的名單佔總體百分比。縱軸則是在這批名單中稀有事件的人數佔總體稀有事件人數的百分比。

增益圖中有一條 45 度的斜線，代表隨機的狀態，以增益圖的定義來說，代表當我們篩選一半的名單去檢視學生流失狀況時，就會剛好包含全體名單一半的學生流失數量（等於隨機亂猜）。

正常的模型增益圖必定要比 45 度線向第二象限彎曲，增益圖愈向上彎曲，表示模型效果愈好。「理想模型」的線代表完美

預測的結果。所有模型的曲線必須介於「隨機猜測」以及「理想模型」之間，愈接近「理想模型」愈好。（尹相志，2007）

在「表 4-12 不同比例訓練組資料模型增益圖比較」中，「分數」指的是模型曲線下面積與完美曲線下面積的比值，因此分數愈接近 1，就表示模型預測力愈高。

「目標母體」欄位會顯示對應的縱軸，這個值愈高就表示模型的預測力愈高。

橫軸表示整體母體擴展，即 overall population%，以「圖 4-26 訓練組資料 1：1 之決策樹模型增益圖」為例，綠色的線對應在橫軸 50%對應到縱軸的值為 76.79%，這代表資料採礦根據流失機率由最高至最低排序的前 50%名單中抓到會流失的學生，就佔了總體會流失的學生的 76.79%，相對的，剩下的 50%學生中，只佔了總體會流失學生 23.21%。在此，我們以母體擴展 50%為基準作各模型之比較。

由於「分數」指的是整體的面積，而「目標母體」值則僅就部分母體擴展而言，故以「分數」高低做為模型優劣的衡量。

「表 4-12　不同比例訓練組資料模型增益圖比較」中，可以看出以訓練組資料 1：2.75，COMPLEXITY_PENALTY=0.5，MINIMUM_SUPPORT=10 之決策樹模型之分數 0.96 為最高，因此判斷此為最佳之模型。

表 4- 12　不同比例訓練組資料模型增益圖比較（以母體擴展 50%為比較基準）

| 序列、模型<br>（顏色代表增益圖的線的顏色） | 分數 | 目標母體% | 預測機率% |
|---|---|---|---|
| 隨機猜測模型 | | 50.00 | |
| 理想模型 | | 100.00 | |
| 訓練組資料 1：1，<br>COMPLEXITY_PENALTY=0.5，<br>MINIMUM_SUPPORT=10 之決策樹模型 | 0.90 | 76.79 | 49.85 |

| | | | |
|---|---|---|---|
| 訓練組資料 1：1，COMPLEXITY_PENALTY=0.1，MINIMUM_SUPPORT=5 之決策樹模型 | 0.83 | 73.21 | 48.21 |
| 訓練組資料 1：1，COMPLEXITY_PENALTY=0.05，MINIMUM_SUPPORT=5 之決策樹模型 | 0.81 | 71.43 | 34.37 |
| 訓練組資料 1：2，COMPLEXITY_PENALTY=0.5，MINIMUM_SUPPORT=10 之決策樹模型 | 0.89 | 82.14 | 29.19 |
| 訓練組資料 1：2，COMPLEXITY_PENALTY=0.1，MINIMUM_SUPPORT=5 之決策樹模型 | 0.85 | 78.57 | 27.13 |
| 訓練組資料 1：2，COMPLEXITY_PENALTY=0.05，MINIMUM_SUPPORT=5 之決策樹模型 | 085 | 78.57 | 27.13 |
| 訓練組資料 1：2.75，COMPLEXITY_PENALTY=0.5，MINIMUM_SUPPORT=10 之決策樹模型 | 0.96 | 100 | 16.29 |
| 訓練組資料 1：2.75，COMPLEXITY_PENALTY=0.1，MINIMUM_SUPPORT=5 之決策樹模型 | 0.86 | 78.57 | 20.71 |
| 訓練組資料 1：2.75，COMPLEXITY_PENALTY=0.05，MINIMUM_SUPPORT=5 之決策樹模型 | 0.85 | 78.57 | 20.71 |

由「圖 4-26 訓練組資料 1:1 決策樹模型增益圖」可以看出，在已流失與未流失資料 1：1 的訓練組資料中，以調整參數的 COMPLEXITY_PENALTY=0.5，MINIMUM_SUPPORT=10 之模型曲線預測力最好，其次為訓練組資料 1：1，COMPLEXITY_PENALTY=0.1，MINIMUM_SUPPORT=5 之決策樹模型，再次之為訓練組資料 1：1，COMPLEXITY_PENALTY=0.05，MINIMUM_SUPPORT=5 之決策樹模型。

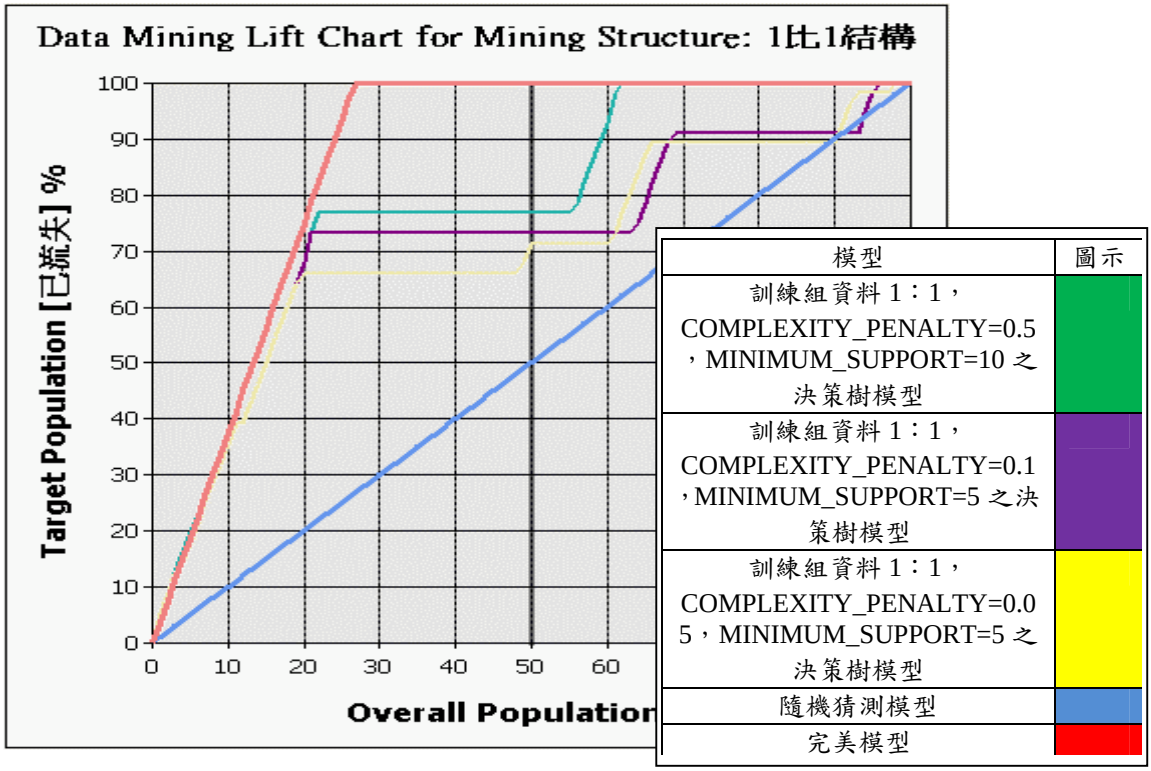


圖 4- 26　訓練組資料 1：1 之決策樹模型增益圖

由「圖 4-27 訓練組資料 1:2 決策樹模型增益圖」可以看出，在已流失與未流失資料 1：2 的訓練組資料中，以調整參數的 COMPLEXITY_PENALTY=0.5，MINIMUM_SUPPORT=10 之模型曲線預測力最好，而訓練組資料 1：2，COMPLEXITY_PENALTY=0.1，MINIMUM_SUPPORT=5 之決策樹模型與訓練組資料 1：2，COMPLEXITY_PENALTY=0.05，MINIMUM_SUPPORT=5 之決策樹模型產生重疊。

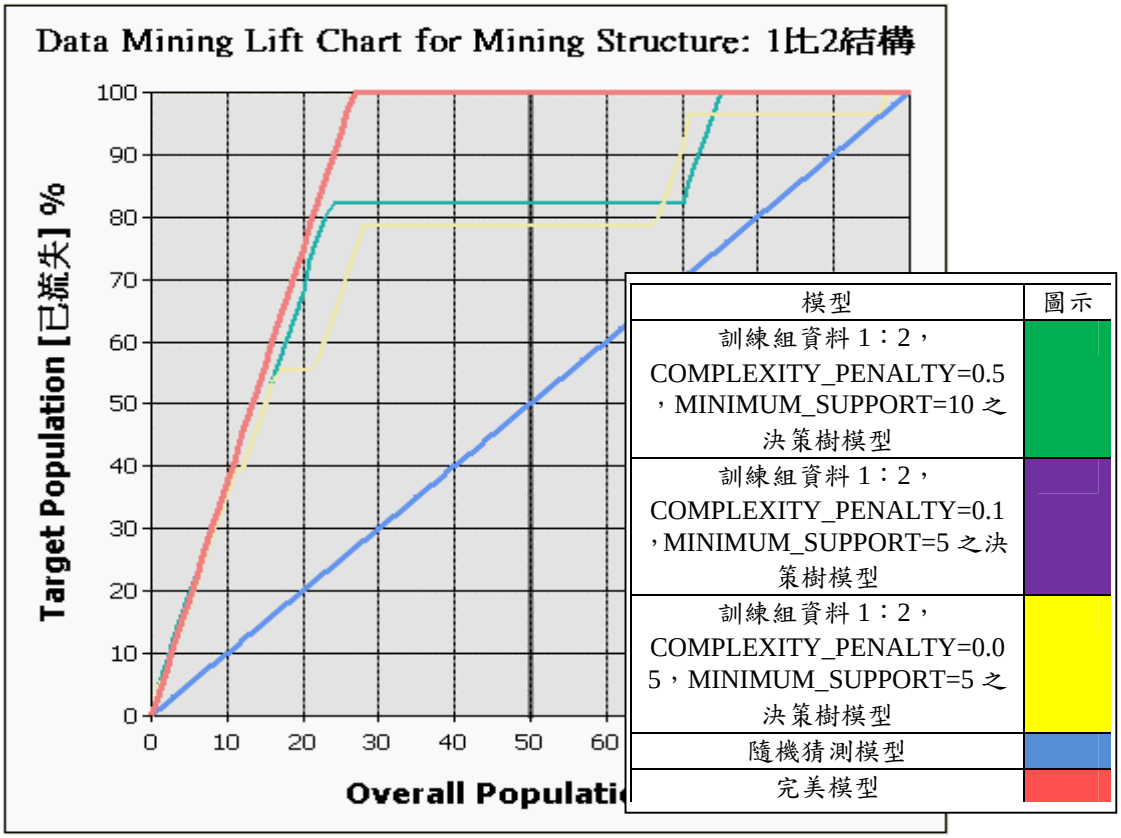


圖 4- 27　訓練組資料 1：2 之決策樹模型增益圖

由「圖 4-28 訓練組資料 1：2.75 決策樹模型增益圖」可以看出，在已流失與未流失資料 1：1 的訓練組資料中，以調整參數的 COMPLEXITY_PENALTY=0.5，MINIMUM_SUPPORT=10 之模型曲線預測力最好，甚至在母體擴展即橫軸 34%時，目標母體即高達 100 %，而訓練組資料 1：2.75，COMPLEXITY_PENALTY=0.1，MINIMUM_SUPPORT=5 之決策樹模型，與訓練組資料 1：2.75，COMPLEXITY_PENALTY=0.05，MINIMUM_SUPPORT=5 之決策樹模型則產生部分重疊。

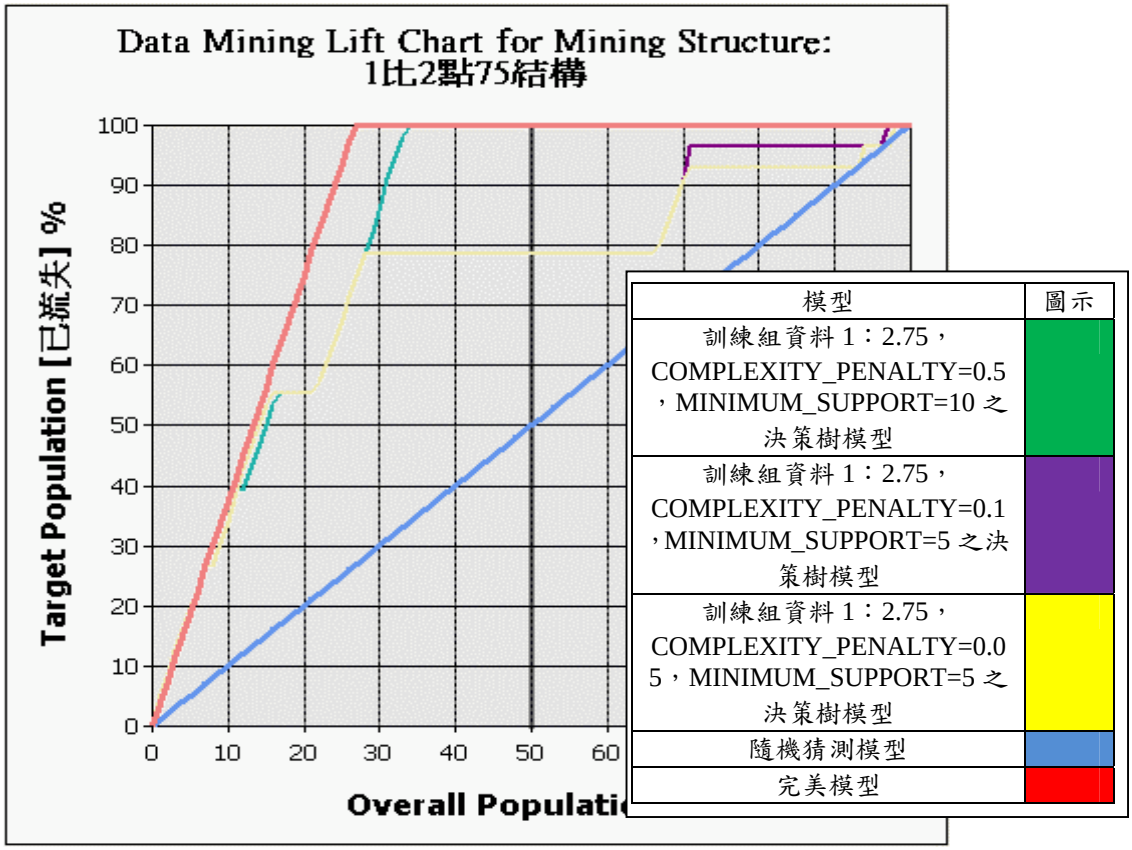


圖 4- 28　訓練組資料 1：2.75 之決策樹模型增益圖

參、 採礦結果與圖表

由訓練組資料 1：2.75，COMPLEXITY_PENALTY=0.5，MINIMUM_SUPPORT=10 之決策樹模型，所建立之決策樹模型如「圖 4-29　訓練組資料 1：2.75，COMPLEXITY_PENALTY=0.5，MINIMUM_SUPPORT=10 之決策樹模型」所示，其中「Total AV#」為歷年學業平均成績，顯示學生之學習成績仍為學生流失與否之重要預測因素。至於性別、科系、入學方式、居住地區、每學期不及格學分數等變項對學生流失與否都沒有顯著影響。

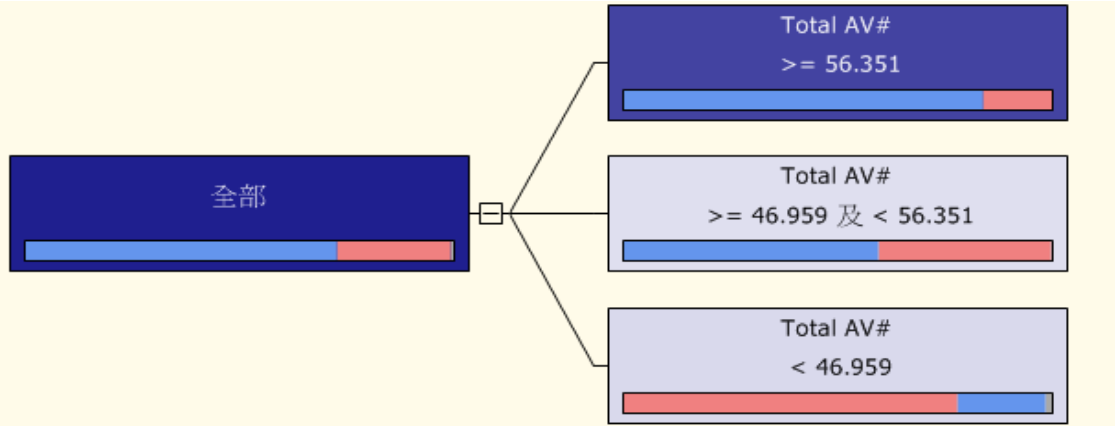


圖 4- 29　訓練組資料 1：2.75，COMPLEXITY_PENALTY=0.5，MINIMUM_SUPPORT=10 之決策樹模型

由決策樹模型獲得規則及其所代表之意涵為：

如「表 4-13 決策樹規則一」所示，Rule1：如果以歷年學業平均成績≧56.351 去預測學生流失與否，則有 16.29%機率為已流失，83.44%的機率為未流失。

表 4- 13　決策樹規則一

| 規則編號 | 節點路徑 | 值 | 案例 | 機率% |
|---|---|---|---|---|
| Rule1 | 歷年學業平均成績≧56.351 | 已流失 | 62 | 16.29 |
| | | 未流失 | 322 | 83.44 |
| | | 遺漏 | 0 | 0.28 |

如「表 4-14　決策樹規則二」所示，Rule2：如果以歷年學業平均成績≧46.959 及＜56.351 去預測學生流失與否，則有 40.34%機率為已流失，59.03%的機率為未流失。

表 4- 14　決策樹規則二

| 規則編號 | 節點路徑 | 值 | 案例 | 機率% |
|---|---|---|---|---|
| Rule2 | 46.959≦歷年學業平均成績＜56.351 | 已流失 | 17 | 40.34 |
| | | 未流失 | 25 | 59.03 |
| | | 遺漏 | 0 | 0.62 |

如「表 4-15 決策樹規則三」所示，Rule3：如果以歷年學業平均成績＜46.959 去預測學生流失與否，則有 77.29%機率為已流失，20.77%的機率為未流失。

表 4- 15　決策樹規則三

| 規則編號 | 節點路徑 | 值 | 案例 | 機率% |
|---|---|---|---|---|
| Rule3 | 歷年學業平均成績＜46.959 | 已流失 | 52 | 77.29 |
| | | 未流失 | 13 | 20.77 |
| | | 遺漏 | 0 | 1.93 |

得到上述三條規則之後，我們再進行原始資料檢視發現，在訓練組 1：2.75 資料中，歷年學業平均成績≧56.351 的學生其就學狀態，如「圖 4-30 規則一之就學狀態」，已流失之就學狀態以轉學居多，顯示規則一：歷年學業平均成績≧56.351，之學生多為自願性流失。

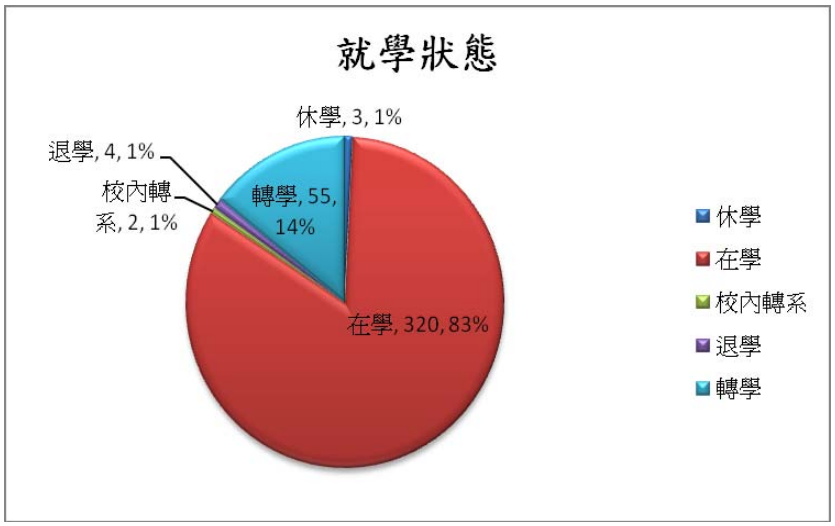


圖 4- 30　規則一之就學狀態

至於規則二，由於資料筆數較少，檢視其原始資料並未發現其他特殊狀況。而規則三之原始資料中則發現，其已流失之就學狀態多為退學，如「圖 4-31 規則三之就學狀態」所示，顯示規則三：歷年學業平均成績＜46.959，學生之流失多為非自願性流失。

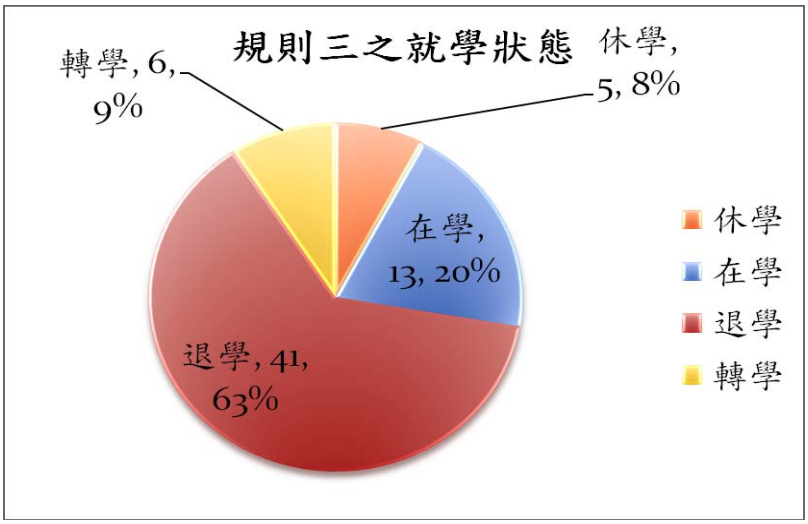


圖 4- 31　規則三之就學狀態

而訓練組資料 1：2.75，COMPLEXITY_PENALTY=0.5，MINIMUM_SUPPORT=10 所得之決策樹模型經測試組資料測試後所得分類矩陣如「表 4-16 決策樹的測試結果」所示，其正確率為 81.90%，錯誤率為 18.09%。

表 4- 16　決策樹的測試結果

| 模型訓練組資料 1：2.75 | 預測 | 實際已流失 | 實際未流失 |
|---|---|---|---|
| 訓練組資料 1：2.75，COMPLEXITY_PENALTY=0.5，MINIMUM_SUPPORT=10 之決策樹模型 | 已流失 | 22 | 4 |
| | 未流失 | 34 | 150 |
| | 正確總數、百分比 | 172，81.90% | |
| | 錯誤總數、百分比 | 37，18.09% | |

## 第三節　類神經網路分析

類神經網路模型只能處理連續數值，如果輸入變數為類別變數時，會先將它轉換為虛擬變數（也就是每個選項轉為 1 或 0 的編碼）。類神經網路是透過修正權重的模式來產生學習成果，演算法本身並不具有變數篩選功能，也就是所有的變數都會計算出屬於自己的權重。因此，在

SQL Server 2005 類神經網路演算法的檢視器中，並非以呈現權重為目的，而主要是呈現資料內部的機率分佈架構。

依據前一節所準備之訓練組及測試組資料，分別進行類神經網路模型之製作，在調整各個模型之內建參數之後，發現調整參數之後的差異變化不大，故皆不予調整，完成之後分代入測試組資料。

壹、分類矩陣：

我們可以透過「表 4-17 類神經網路分類矩陣」來檢視各組資料之類神經網路模型的錯誤分佈。而我們「預測未流失實際卻已流失」會造成比較嚴重損失，是我們所關心的部分。

表 4- 17　類神經網路分類矩陣

| 訓練組資料 | 預測 | 實際已流失 | 實際未流失 |
|---|---|---|---|
| 1：1 | 已流失 | 37 | 50 |
| | 未流失 | 19 | 104 |
| 訓練組資料 | 預測 | 實際已流失 | 實際未流失 |
| 1：2 | 已流失 | 22 | 12 |
| | 未流失 | 34 | 142 |
| 訓練組資料 | 預測 | 實際已流失 | 實際未流失 |
| 1：2.75 | 已流失 | 22 | 10 |
| | 未流失 | 34 | 144 |

貳、3R 分析：

透過分類矩陣所獲得之資訊，我們進一步來計算 3R 指標，以判斷模型之成效，所得結果如「表 4-18 類神經網路模型評估 3R%」所示，由於三組模型各有優劣，我們繼續進行增益圖之比較。

表 4- 18　類神經網路模型評估 3R%

| | 總體回應率 | Response Rate 回應率 | Recall 反查 | Range Reduce 間距縮減 | 回應率提升 |
|---|---|---|---|---|---|
| 資料意義 | | 愈高愈好 | 愈高愈好 | 愈低愈好 | 愈高愈好 |
| 類神經網路 1：1 模型 | 26.67 | 42.53 | 66.07 | 41.43 | 159.48 |
| 類神經網路 1：2 模型 | 26.67 | 64.71 | 39.29 | 16.19 | 242.65 |
| 類神經網路 1：2.75 模型 | 26.67 | 68.75 | 39.29 | 15.24 | 257.81 |

參、增益圖：

「圖 4-32　訓練資料 1：1 類神經網路模型增益圖」中，淺藍色曲線即為類神經網路之模型增益曲線，由「表 4-19　類神經網路模型增益圖比較表」中可看出類神經網路 1：2.75 模型之分數為 0.78 與其他兩者之差距甚微。

表 4- 19　類神經網路模型增益圖比較表

| 序列、模型 | 分數 | 目標母體% | 預測機率% |
|---|---|---|---|
| 類神經網路 1：1 模型 | 0.77 | 73.21 | 45.59 |
| 類神經網路 1：2 模型 | 0.75 | 69.64 | 23.03 |
| 類神經網路 1：2.75 模型 | 0.78 | 73.21 | 16.67 |
| 隨機猜測模型 | | 50.00 | |
| 理想模型 | | 100.00 | |

由「圖 4-32」至「圖 4-34」可以看出，不論哪一種訓練資料模型，決策樹模型的預測力都比類神經網路模型的預測力來得高。

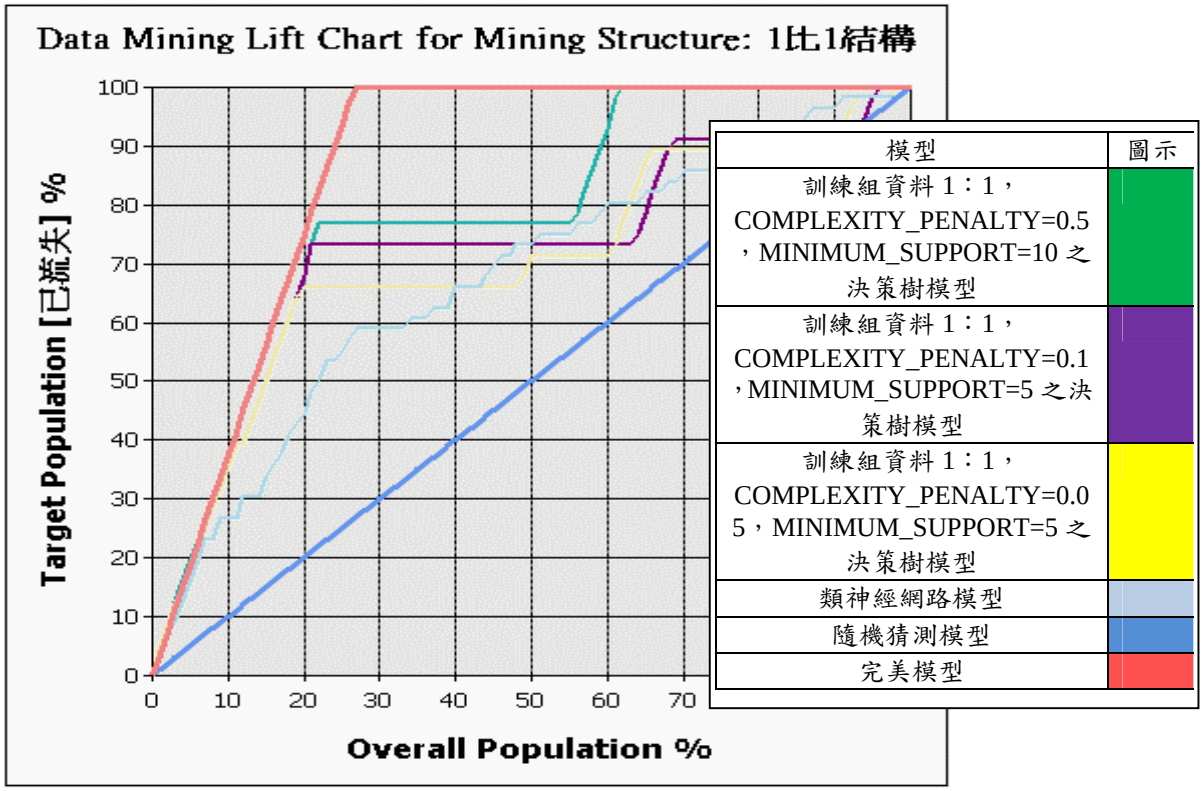


圖 4- 32　訓練資料 1：1 類神經網路模型增益圖

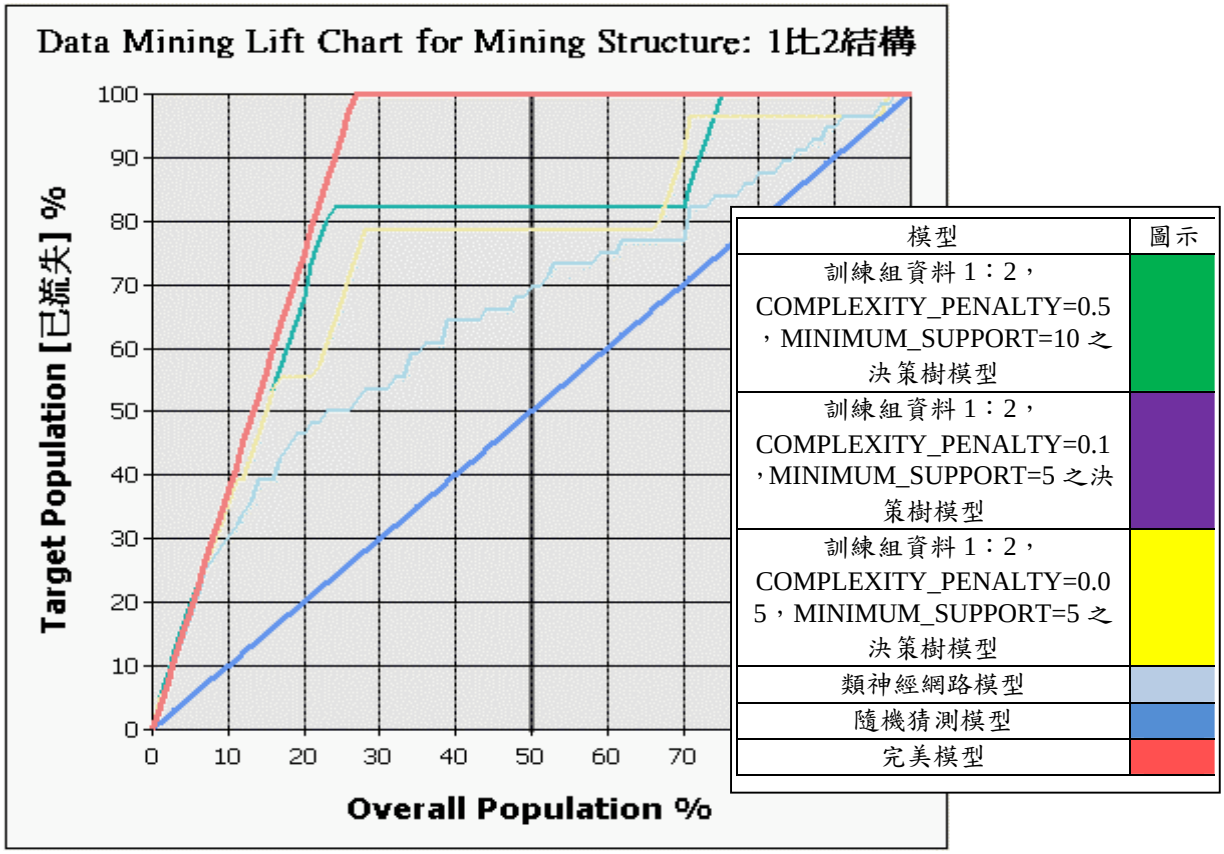


圖 4- 33　訓練資料 1：2 類神經網路模型增益圖

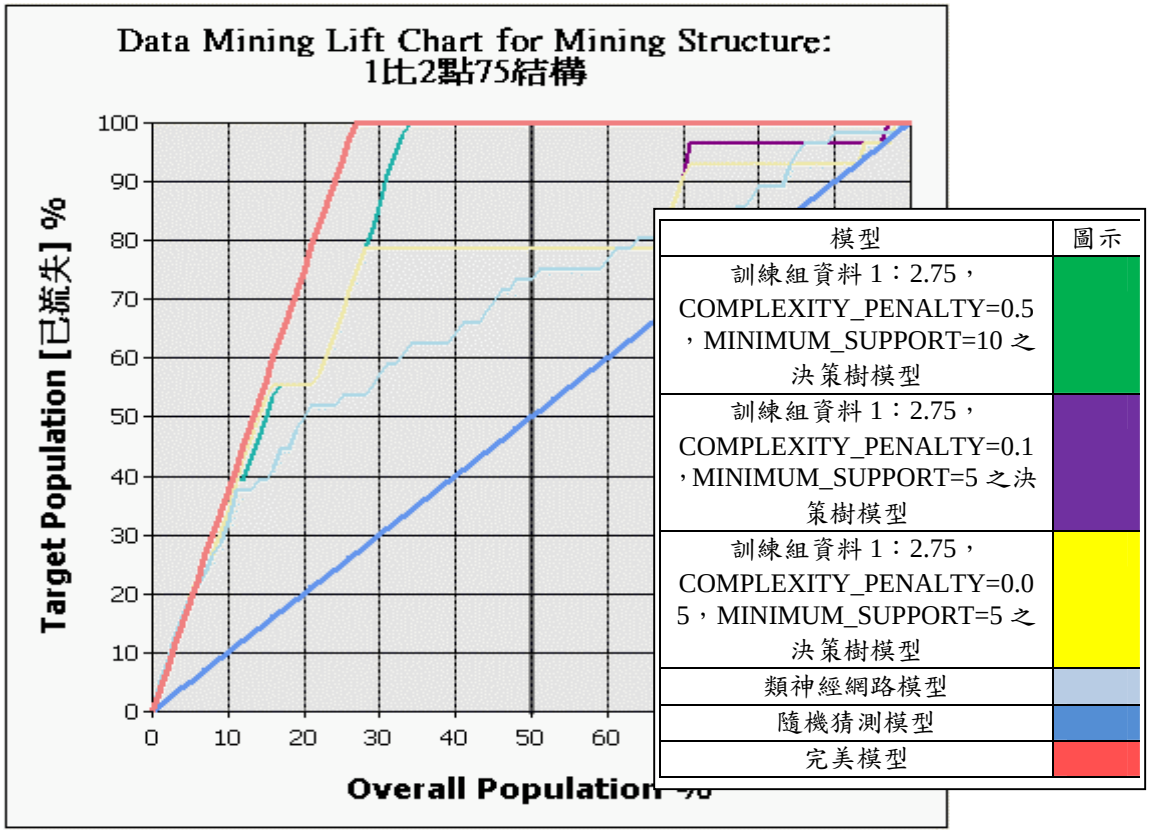


圖 4- 34　訓練資料 1：2.75 類神經網路模型增益圖

肆、採礦模型：

由於類神經網路的三個模型其模型評估指標在模型與模型之間差異很小，故在此採用決策樹中獲得較佳模型之訓練組資料 1：2.75，進行類神經網路分析，所得模型如「圖 4-35　類神經網路輸出檢視」所示。Total AV#為歷年學業平均成績，Lose AV#為平均每學期不及格學分數，Lose AV#正常值應為正數，而其中為負數者，其原因在於學生進行學分抵免，抵免時的學分列入每學期的實得學分數，故造成當學期的實得學分數大於其修課學分數的情形。

「圖 4-35　類神經網路輸出檢視」所示為訓練組全部資料的所有變數選項組合，依照顯著性由高至低排列。即 Attribute 欄位依照其顯著性由高至低排列，顯著性即系統自動計算出之分數。

Favors 則依據自動算出之 Lift 值，決定為已流失或未流失，如「表 4-20 類神經網路輸出變數喜好值」所示。

由「圖 4-35　類神經網路輸出檢視」中可以看出，以歷年學業平均成績 5.305-50.526 之顯著性為最高，其傾向已流失；其次為居住地區為離島，其傾向為已流失；接著是歷年學業平均成績 76.759-93.918，其傾向為未流失，依此類推。

由於此檢視表乃顯示全部變數之顯著性及已流失與未流失傾向，細部資料還需進一步進行相關之選取。


圖 4- 35　類神經網路輸出檢視

「表 4-20　類神經網路輸出變數喜好值」為擷取顯著性較高的前幾項變數，其中，觀察第一列歷年學業平均成績為 5.305-50.526 的變數，其 Score 最高，即顯著性最高，而由系統依發生機率幫我們

自動算出之 Lift 值決定其傾向為已流失或未流失，而計算出來之數值其呈現即為「圖 4-35　類神經網路輸出檢視」。依此類推，可看出各變數值之顯著性高低及其已流失與未流失之傾向。

表 4- 20　類神經網路輸出變數喜好值（擷取）

| Attribute | Value | 傾向已流失 | 傾向未流失 |
|---|---|---|---|
| Total AV# | 5.305 - 50.526 | Score: 100<br>Probability of Value1: 72.06%<br>Probability of Value2: 27.65%<br>Lift for Value1: 3.97<br>Lift for Value2: 0.34 | |
| Area | 離島 | Score: 64.11<br>Probability of Value1: 51.73%<br>Probability of Value2: 47.98%<br>Lift for Value1: 2.85<br>Lift for Value2: 0.59 | |
| Total AV# | 76.759 - 93.918 | | Score: 59.9<br>Probability of Value1: 4.85%<br>Probability of Value2: 94.87%<br>Lift for Value1: 0.27<br>Lift for Value2: 1.16 |
| Application | 其他 | Score: 54.74<br>Probability of Value1: 45.99%<br>Probability of Value2: 53.72%<br>Lift for Value1: 2.53<br>Lift for Value2: 0.66 | |
| Lose AV# | 7.036 - 20.981 | Score: 45.76<br>Probability of Value1: 40.59%<br>Probability of Value2: 59.12%<br>Lift for Value1: 2.23<br>Lift for Value2: 0.73 | |
| Area | 外國 | Score: 41.86<br>Probability of Value1: 38.30%<br>Probability of Value2: 61.41%<br>Lift for Value1: 2.11<br>Lift for Value2: 0.75 | |
| Lose AV# | -14.999 - -1.054 | | Score: 38.01<br>Probability of Value1: 8.02%<br>Probability of Value2: 91.69%<br>Lift for Value1: 0.44<br>Lift for Value2: 1.12 |
| Area | 東部 | Score: 29.75<br>Probability of Value1: 31.56%<br>Probability of Value2: 68.15%<br>Lift for Value1: 1.74<br>Lift for Value2: 0.84 | |

伍、輸入「過濾條件」顯示：

在類神經網路中，可由使用者自行輸入過濾條件來限縮母體的範圍，使用者可以透過輸入區域的「屬性」選擇所要的變數，不過由於輸入之過濾條件僅就原始模型之細部資料察看，與原模型之差別並不大，僅顯著性高低之不同，但仍可提供我們作參考，以「圖 4-36　類神經網路檢視輸入」為例，選擇學生入學方式為輸入條件，發現每一種入學身份的流失與否，成績仍為最重要的因素，但仍可觀察各個變數選項中可預測變數的選項分佈機率。

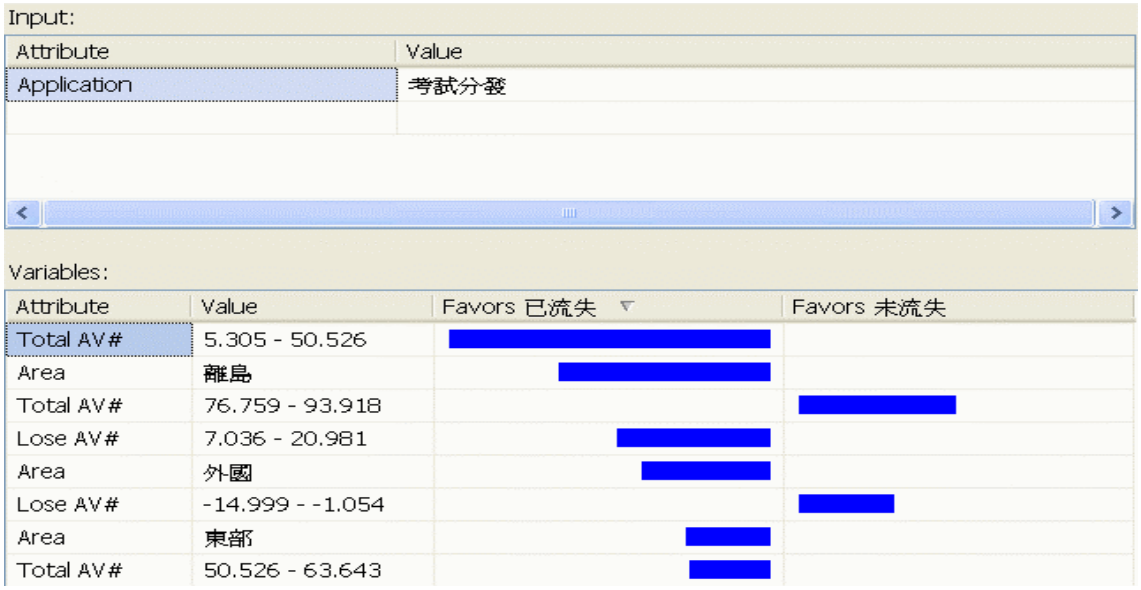


圖 4- 36　類神經網路檢視輸入

我們輸入過濾條件為歷年學業平均成績在 76.759-93.918 來進行細部檢視，其檢視圖如「圖 4-37　類神經網路模型檢視」所示。歷年學業平均成績在 76.759-93.918 者，傾向於未流失，其顯著性以 Lose AV#平均每學期不及格學分數-14.999 ~ -1.054 為最高，其傾向為未流失。其次為入學身份為個人申請、科系為資訊工程系，依此類推。

此外，細部資料顯示居住地區為離島或外國，以及入學身份為其他者，傾向已流失。檢視原始資料後發現，居住地區為外國者，僅 1

人，且為未流失，此外居住於離島或東部的資料亦算少數，因此發現資料筆數較少者，其經過 Lift 值計算之後，容易造成已流失或未流失的偏好顯示異常，此外，其 Attribute 顯著性也提高許多。

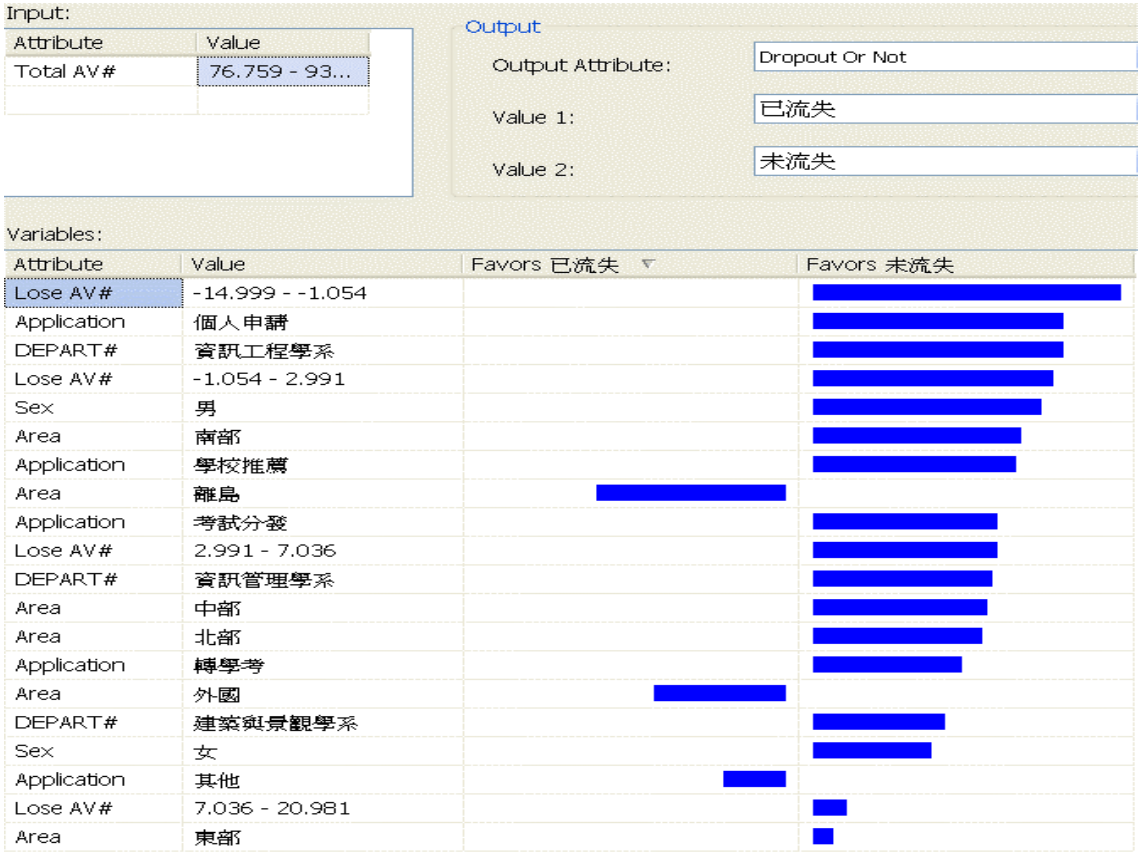


圖 4- 37　類神經網路模型檢視

由此可知透過類神經網路檢視器只能協助我們瞭解變數之間的交互作用，而非權重高低，因此我們只能使用此檢視器瞭解變數重要性。

# 第五章 結論與建議

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

## 第一節 研究結論

本研究試圖找出潛在可能流失學生，並提供校方及老師們作為參考，以協助教師輔導及關心學生時有所依據並找出重點輔導對象，以減少學生流失之情形。找出流失學生之潛在共同特質，運用於在學學生之資料分析，預測在學卻可能流失之學生，並提供可行之建議，作為學生輔導之依據，以改善學生流失的情況。

本研究在資料特性分析中發現：

壹、大學部一年級學生升二年級時的流失率較其他各年級為高。可能原因為新生剛進校門所以適應及學習方面有困難，建議於新生入學時，加強對於系所之介紹，包括課程及校園環境等等，使新生能充分瞭解並融入，並多辦理活動以增加其認同及歸屬感，進而減少其流失。

貳、性別比例當中，就退學方面來講，以男生增益值較高，而就轉學方面來講則是女生增益值較高。針對可能流失學生輔導時，可加強其對本系專業之興趣，或可建議其轉系，固然學生轉系對系上來說是學生流失，然而對學校來說校內轉系並不算學生流失。

參、以科系來說，就學狀態為休學時，建景系較資管系增益值高；而就學狀態為退學時，則以資工系增益值較高，其次為資管系，之後才

是建景系；而就學狀態為轉學時，以資工系增益值較高，其次為資管系，之後是建景系。

肆、入學方式為轉學考者，其休學及退學的增益值較高，由於轉學生需要融入一個一群本來就互相熟悉的群體，加上課業上的不熟悉，很容易就變成孤立的個體而最後選擇流失，若各系之轉學生能有學長姐制，輔導自己系上的轉學生融入本系，或可減少其流失。

伍、研究樣本居住縣市以高雄市、屏東縣、雲林縣、台北縣、台北市等縣市為最多，建議可針對未流失增益值較高的縣市加強招生及宣傳，如高雄縣市、屏東縣及雲林縣等。

本研究在資料採礦模型中發現，針對學生的已流失與未流失分類仍以歷年學業平均成績為主要分類依據，顯示學生的學習情況仍是決定學生流失與否之重要原因。

在決策樹模型中所發現，歷年學業平均成績≧56.351 之流失以轉學居多，歷年學業平均成績＜46.959 之流失以退學居多。針對自願性流失之學生，學校及老師可加強其生活輔導，增加其對本校或科系之認同度，而針對非自願性流失之學生，學校則需生活及課業輔導雙管齊下。

在決策樹中，以成績為主要分類依據，另外可以參考平均不及格學分數及入學方式是否為轉學考等兩項因素，在類神經網路分析亦以成績為主要因素，而居住地區亦可能是影響學生流失與否的潛在因素之一。

整體而言，決策樹任一模型皆比類神經網路模型的預測力來得高，兩種資料採礦方法包括決策樹及類神經網路都發現學生學習成績為分類的重要因素，顯示成績確為學生流失與否之重要因素。

## 第二節 研究建議

一、 囿於分析資料之稀少，並未充分展現資料採礦之功用，建議後續之研究可增加分析樣本之數量，或可發現更多影響學生流失之潛在因素。

二、 由於影響學生流失之其他相關因素，諸如：家庭、經濟、個人因素等等，並未納入考量，而本研究之已流失包括休學、退學、轉學，建議後續可分別進行相關之研究。

三、 此外，因取得資料之欄位，並未將休學及退學或轉學之先後順序欄位分開，以致無法分析其是否有時序關係，建議若資料庫進行修正時，可納入考量。

# 參　考　文　獻

## 一、中文部份


〔1〕 Microsoft SQL Server: SQL Server 首頁 http://www.microsoft.com/taiwan/sql/default.mspx

〔2〕 中華資料採礦協會 http://www.cdms.org.tw/xoops2/html/modules/news/

〔3〕 尹相志，SQL Server 2005 Data Mining 資料採礦與 Office 2007 資料採礦增益集，悅知文化，台北，2007。

〔4〕 內政部戶政司人口統計資料 http://sowf.moi.gov.tw/stat/month/m1-02.xls

〔5〕 王智弘，「探討團體成員流失問題之原因、影響及因應對策」，輔導月刊，24 卷 6 期，37-40 頁，民 77。

〔6〕 王智弘，「團體成員流失問題之探討」，輔導季刊，30 卷 4 期，27-36 頁，民 83。

〔7〕 夏載，”剖析資料採礦在顧客關係管理中的應用”，電子化企業經紀人報告 eBusiness Executive Report，20 期，71-75 頁，2001。

〔8〕 教育部高教司統計處 http://www.edu.tw/files/site_content/B0013/overview03.xls

〔9〕 教育部高教技職簡訓 011 期 http://www.edu.tw/statistics/content.aspx?site_content_sn=8956

〔10〕 盧梅莉，「團體成員流失之探討」，諮商與輔導，76 期，34-36 頁，民 81。

## 二、西文部份


〔1〕 Berson, A., Smith, S. and Thearling, K., ”Building Data Mining Applications for CRM”, Customer Retention, New York, McGraw-Hill, 2000.

〔2〕 Bolton,Ruth N,”A Dynamic Model of the Duration of the Customer’s Relationship with A Continuous Service Provider:The Roe of Satisfaction,” Marketing Science,Vol. 17,No. 1,pp.45~65, 1998.

〔3〕 Davids,M.,”How to avoid the 10 Biggest Mistake in CRM”,Journal of Business Strategy,Vol.4,pp.22-26,1999.

〔4〕 Frawley, W., Piatetsky-Shapiro, G. and Matheus, C., “Knowledge Discovery in Databases: An Overview”, AI Magazine, pp. 213-228 , Fall 1992.

〔5〕 Ganesh,Jaishankar Mark J. Arnold,and Kristy E. Reynolds,”Understanding of Customer Base of Service Providers:An Examination of Differences between Switchers and Stayers”,Journal of Marketing,Vol.64,pp.65~87, July 2000.

〔6〕 Ian H. Witten, Eibe Frank. ,*Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

〔7〕 Keaveney,Susan M,”Customer Switching in Service Industries:An Exploratory Study”,Journal of Marketing,Vol.59,pp.71~82. April 1995.

〔8〕 Kleissner, C., “Data mining for the enterprise”, Proceedings of the Thirty-First Hawaii International Conference, pp.295-304, 1998.

〔9〕 Shaw, M. & Subramaniam, C, “Knowledge management and data mining for marketing”, Decision Support Systems, pp. 127–137, 2001, 31.

〔10〕Strouse, Karen G,*Marketing Telecommunications Services New Approaches for A Changing Environment*,Boston:Artech House.1999