南台科技大學

資訊管理研究所

碩士學位論文

應用決策樹與關聯規則於學生成績之分析-以台南市某職業學校為例

Application of Decision Tree and Association Rules for the Analysis of Student Achievement - A Case Study of a Vocational School in Tainan

研究生：吳靜薇

指導教授：黃仁鵬

中華民國一〇三年一月

南台科技大學

資訊管理研究所

碩士學位論文

應用決策樹與關聯規則於學生成績之分析-以台南市某職業學校為例

Application of Decision Tree and Association Rules for the Analysis of Student Achievement - A Case Study of a Vocational School in Tainan

研究生：吳靜薇

指導教授：黃仁鵬

中華民國一〇三年一月

南臺科技大學 碩士論文

應用決策樹與關聯規則於學生成績之分析-以台南市某職業學校為例

研究生：吳靜薇

本論文業經審查及口試合格特此證明

論文考試委員

郭煌政 郭煌政　　張雲龍 張雲龍

黃仁鵬 黃仁鵬

指導教授　黃仁鵬 黃仁鵬

所　　長　黃仁鵬 黃仁鵬

中 華 民 國 一 〇 三 年 一 月 二 十 一 日

摘要

本研究旨在探討高職入學成績、入學方式、性別與畢業成績間之關聯性做為研究主題，研究方法使用 Data Mining 之關聯法則及決策樹規則，研究結果做為將來招生決策的參考。結論摘要如下：

一、入學成績對於畢業成績確有極大的影響，兩者呈現正相關。

二、女學生之學業成績相較於男學生之學業成績較為優異。

三、畢業國中於某特定國中之學生，學業成績較佳。

四、對於入學身份為其他的學生，畢業成績較差，校方更應關注學生的學習態度，並給予生活及精神上的關心注意學生的品格教育。

五、決策樹與關聯規則之結果大致上一致，唯類別太多且資料量小，決策樹無法顯示個別學校之差異，關聯規則較能完整的呈現各類別之影響。

關鍵詞：資料探勘、關聯式規則、決策樹規則、招生策略

Abstract

This thesis aimed to explore the relevance of vocational entrance exam grades, entrance way, gender and graduation grades as a research themes. In research methods, we use decision trees and association rules of data minining. The result is for reference of the future admissions decisions. The results are as follows：

1. Entrance exam grades is postively correlated with graduation grades.
2. Academic performance of female students compared to male students are more excellent.
3. A specific graduation junior high school students have better grades than others.
4. To those whose enrollment identity is "other" ,school should pay more attention to those students' study attitude.
5. The results of the decision tree are consistenct to those of association rules. But if the categories of decision tree are numerous with a small amount of data, decision tree can not tell the differences from individual schools. Association rules are better and have a complete view of each category of relationship.

Key Words: Data Mining, Association Rules, Decision Tree rules, Admissions Policies

誌謝

研究所一路走來的歲月裡，非常辛苦，卻也充實。或許是要感謝的人事太多了，或許離別的時刻將近，心中有著太多的感想與觸發，以致於尚無法將心中所思所想完整地整理出，但此刻的我，心中盡是滿滿的感激。

首先，感謝指導教授黃仁鵬教授，指導學生論文的撰寫，幫助學生建構整個論文組織體系，澄清學生在撰寫論文時之疑惑處，使之條理分明。更感謝老師不斷的鼓勵和開導學生，如同黑暗中的明燈，引導學生走出迷網，迎向光明，讓學生有勇氣面對種種的困難。

同時，也謝謝口試委員郭煌政教授和張雲龍教授對本文提供寶貴的建議，使論文更臻完備，幫助本研究之架構與內容更為完整與豐富。特此致謝。

對於專班的所有同學們和身邊的好友，在研究的過程中，大家的陪伴和共勉，讓我深感懷念；尤其是同學許銘源助教一路無怨無悔的陪伴。

最後，感謝家人的支持與關懷，讓我在無後顧之憂的情況下，專心於研究，順利的完成碩士論文。謹以此一成果獻給所有愛我以及我愛的人。

靜薇誌於南台

Jan. 2014

目次

表目錄

圖目錄

第一章　緒論

1.1 研究背景

提昇教育品質為教育改革的重要目的，而其中學生素質影響教育品質甚鉅，可是學校能吸引學生就讀的因素，如辦學績效、畢業生表現及學校風評等都需要中長期的努力耕耘。短期最有效的辦法就是研擬適當的招生策略。如目前每年舉辦的高中職博覽會，以各式校園活動、社團及獎學金等吸引學生就讀，或是學生參與校外競賽成績優異，就是學校發揮特色進行行銷的舞台。

以往各校的招生方式皆憑經驗來進行，但是單憑有限經驗並無法將資源做最有效的利用，尤其因應多元的入學方式，該如何運用有限的資源，招收到適合的人才成為各個學校的重大挑戰。過去職業學校招生策略問題的研究，大多是利用問卷調查，但通常會受限於問卷設計的良窳，或填答者是否據實以告等困難因素，而使得問卷的結果恐有不盡完善之虞[12]（Berry and Linoff，1997）。

資料探勘(data mining)技術是近年來被廣泛應用於各個領域的技術，其主要訴求是從大量的歷史資料中挖掘隱藏其中的有用資訊，國內外許多的文獻及商業界都存在許多資料探勘成功的案例，例如百貨業、醫學界、壽險業、銀行業

及通訊業等。近年許多學者也應用資料探勘至各種研究[23, 32, 34, 39, 44]，以持續演進資料探勘的發展與應用。

有鑒於私校招生日益困難，本研究希望了解高職入學成績、入學管道、性別、入學身份、畢業國中與高職在校成績之間的關聯，為學校招生和學生學習做出更有益的貢獻；本研究利用資料探勘技術(Data Mining)中的決策樹(Decision Tree)及關聯規則，探討分析影響事件結果發生的因素，透過決策樹及關聯規則的分析，並克服了傳統研究 (如交叉分析法) 的限制，歸納出影響學生成績的因素。由此可作為學校辦學招生的參考重點。同時讓學校了解學生實際的需求，調整未來教學方向或招生名額。

1.2 研究動機與目的

學校的學籍資料庫，也是儲存著大量的招生的資訊，然而這些大量的資料當中所隱藏的一些有用資訊，如何將這些大量的資料當中找出有用的資訊是一個非常重要的探討議題。

資料探勘的主要目的，就在於提供方法從大量的資料當中找出有用的資訊，其中包含了關聯規則（Association Rules）、分類（Classification）、分群（Clustering）以及序列型樣分析（Sequential pattern analysis）[7, 23, 43]。

本研究透過這些資料探勘的技術，可以讓人們獲得未知並且有用的知識，進而提高學校在市場上的競爭力。

由於少子化效應，私立職校招生活動已是各校兵家必爭之地，如何在學校的學籍資料庫，利用資料探勘技術，找出當中所隱藏的一些有用資訊，提出招生建議，並了解學生學習狀況，實為當務之急。

本研究之研究目的，利用學籍資料庫及學生成績資料庫挖掘出學生入學成績、學期成績、性別、入學方式、入學身份、畢業國中這此因素之間的關聯性，提供學生招生決策時採用，及教師教學成效之參考。

1.3 研究範圍與限制

1.3.1 研究範圍

本研究以台南市某私立高級工商職業學校為個案研究對象，收集該校完整且正確的學生資料，再利用資料探勘技術中決策樹規則與關聯規則分析的技術來發掘學生特性，了解其中的關聯，進而提供學校研擬入學招生策略之參考依據。

1.3.2 研究限制

一、在研究範圍方面，僅取得南部某職業學校做為研究對象，無法擴及其他學校。

二、本研究以某私立高職為研究對象，不探討社會價值觀與教育政策等中介變項之影響，而是以高職入學成績、入學管道、性別、入學身份、畢業國中與高職在校成績之間的關聯之關聯來進行探討，因此，所得之研究結果僅能代表目前該所高職所存在之情況。

三、使用資料探勘的概念是，只要有資料，就會有探勘的結果，但如果資料品質不佳，或數量過少，可能會影響探勘結果。

1.4 論文架構

本研究共分為五章，其各章節架構之詳細如下：第一章緒論，說明本研究之研究動機與背景、研究目的、研究步驟、研究範圍與限制以及本研究之論文架構。第二章文獻探討，針對高職入學現況，入學變數與在學成績之關聯性研究及資料探勘法及決策樹規則演算法相關文獻加以探討。第三章研究方法，針對研究資料來源作概略簡介，其中包含研究對象、研究架構、資料處理方法及資料採礦工作之介紹。第四章資料分析，針對研究資料先做基本分析，並且對於所需要更進一步研究之部分加以整理並進行資料採礦分析，分析後獲得之結果進行比較。第五章 結論與建議，針對第四章資料分析後所獲得之結果進行統整，並提出相關結論以及相關建議。

第二章　文獻探討

近幾年資料探勘的應用領域非常廣泛，舉凡零售業、科技業、製造業、電信業、醫療、旅遊業、教育界等等[41]，皆可看到資料探勘的蹤跡，尤其近幾年對於教育領域的研究逐漸備受重視，許多學者利用資料探勘來進行學生狀況的了解、學校招生、選課、學習成績及線上教學等相關研究。

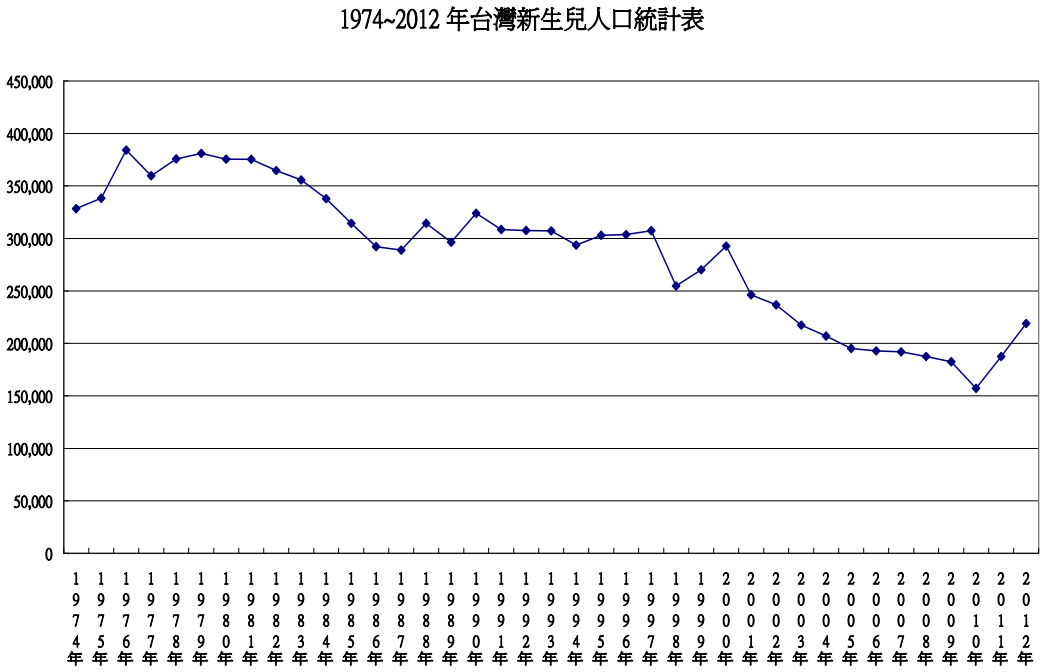

圖 2.1 1974-2012 年台灣新生兒人口統計圖

圖 2.1 為根據內政部統計處(2012)的資料顯示，我國新生兒在 1976 年為 38 萬餘人，是二次世界大戰後的高峰值，但到了 2012 年，我國新生兒人數卻下降

到 22 萬餘人，人口呈緩慢減少趨勢。少子化最先衝擊的是學前教育機構[24]，其所造成各機構普遍招生不足，其次是國小階段，開始大幅減班，此一現象未來勢必影響到高職教育。

許多專家預估大約在 2016 年左右，少子化現象全面衝擊到高中、職，進而造成高中、職班級數銳減。私立高中職是以學費為主要的辦學經費來源，當其面臨到學生人數減少之窘境時，很可能導致整併或停辦之情況產生。

登記分發入學、甄選入學、申請入學，三種入學管道都適用「國中生基本學力測驗」分數，其中「登記分發入學」的名額須佔五○% 以上，仍居多元入學方案主流，此三種入學方式仍都要採計學業成績，於是「基本學力測驗」就成了聯考的替代品。由此可見，國中基測分數仍為高中職分發篩選之重要依據。

另外，本文入學方式所提到的自行報名，乃屬其他入學方式，由各校自行研討可免試入學方式，經主管教育行政機關同意後辦理，此種自行報告方式佔私立職校相當大的比重。

2.1 我國高職入學現況

依國民教育法第 2 條「凡六歲至十五歲之國民，應受國民教育。六歲至十五歲國民之強迫入學，另以法律定之」（強迫入學條例）。我國現已實施六歲至十五歲之九年國民教育，是強迫入學、免學費、義務的教育，接受教育是國

民的權利，也是義務。另依教育基本法第 11 條「國民基本教育應視社會發展需要延長其年限；其實施另以法律定之。」因此，教育部正積極整合現行「高級中學法」及「職業學校法」，制定「高級中等教育法」，也將微調「專科學校法」相關條文，作為實施十二年國民基本教育之法律依據，同時，配合訂定及修正相關子法，以完備十二年國民基本教育法制。

2.1.1 高職多元入學方案公佈後（民國87年以後）

高職入學方式，自八十六學年度起，除原有聯合招生考試、自願就學輔導方案分發、技（藝）能優良學生甄審保送等入學方式外，各區已陸續開始試辦推薦甄選入學方式[37]；至九十學年度止，高職多元入學方案共分六種入學管道，分別是：推薦甄選入學、申請入學、基本學力測驗登記分發入學、自願就學輔導方案分發入學、特定對象學生入學、其他入學方式[37]。

一、推薦甄選入學

以現行高職聯招區劃分為原則，分別聯合區域內公私立學校辦理。甄選項目及甄選方式等，由各高職個別訂定，甄選條件可包括在校成績、競賽表現、獲獎記錄、社團及班級服務、自傳、推薦函件、指定項目測驗及面試等，以符合自主選才之原則；各國中分別成立推薦委員會，依各高職學校所定之甄選條件，推薦適合之應屆畢業生，其資格經審核符合各校科甄選標準，由各校辦理甄選入學。

二、申請入學

僅高職辦理，由各公私立高職依其意願辦理單獨招生。採行申請入學之高職，應自行訂定申請條件，提供符合申請條件之國中應屆畢業生自行衡量本身條件，向辦理申請入學之各高職提出申請。以性向測驗、口試、國中在校表現或各校自訂之其他項目成績為主要入學依據。

三、基本學力測驗登記分發入學

教育部委託國立台灣師範大學測驗中心，研發具科學性、公平性、教育性之國中基本學力測驗，以評量國民中學學生基本學力表現為目的，消除入學考試對於國民中學教育的不利影響，進而充分發展學生的潛能。九十學年度起採國中基本學力測驗分數作為替代分發入學高職之依據，各校科得視實際需要對相關科目加權計分或設定最低標準，並依學生測驗成績及志願辦理分發事宜。同時，各項多元入學方式，並得規定參採或使用國中基本學力測驗分數為入學條件之一。報名資格包括國中畢業生或具同等學力者。

四、自願就學輔導方案分發入學

依照現行各地區「試辦國民中學畢業生自願就學輔導方案」分發辦法辦理。

五、特定對象學生入學

包括下列特定招生對象：（1）技能優良學生甄審保送入學高職。（2）運動績優學生甄審保送入學。（3）國中技藝教育班畢業生分發實用技能班。（4）

國中畢業生申請登記輪調式建教合作班。（5）國中輕度智障學生甄選入學高職特殊教育實驗班。（6）五專原住民專班甄試（選）入學。（7）其他（如僑生、派外人員子女等）。特殊身分學生之入學優待參照現行升學優待辦法辦理。

六、其他入學方式

由各區或各校自行研訂可行免試入學方式，經主管教育行政機關同意後辦理。本文所提到的自行報名，即屬此種報名方式。報名資格包括國中畢業生或具同等學力者。

以下將針對修正後的三種入學方式簡表如下（高中職多元入學方案，2002/08/29 公佈）

表 2.1 甄選入學招生簡表

實施範圍	1.【得跨區】音樂、美術、舞蹈、戲劇、體育特殊才能班、各單類科高中及經教育主管機關核定之高職海事、水產、護理、藝術、農業類科之科、班、校，得跨區聯合辦理招生，學生應向單一學校或跨區聯合甄選委員會報名。 2.【不得跨區】依一般智能及學術性向所設之數理、語文資優班，同一招生區學校可聯合辦理；學生限向國三學籍所在地之招生區單一學校或聯合甄選委員會報名。

實施對象	1. 國民中學應屆畢業生取得國民中學學生基本學力測驗分數者。 2. 非應屆國民中學畢業生及同等學力取得當年度國民中學學生基本學力測驗分者。 3. 符合「資賦優異學生降低入學年齡縮短修業年限及升學辦法」之規定，取得國民中學學生基本學力測驗分數者。
實施方式	1. 以國民中學學生基本學力測驗分數為依據，不採計在校學科績。 2. 國中學生依各校或聯合甄選委員會所定條件，主動向所就讀國中提出申請，由國中彙整後辦理報名。 3. 各校應配合招生之科、班性質參採學生在校藝能表現、綜合表現或特殊事蹟等。 4. 各校應視實際需要就實驗、口試…術科等項選擇辦理，但不得加考任何學科紙筆測驗。
實施時間	1. 每年第一次國民中學學生基本學力測驗成績公布後。 2. 放榜時間應於每年第二次國民中學學生基本學力測驗報名前。

資料來源：整理自教育部中教司資料網

表 2.2 申請入學招生簡表

實施範圍	【不得跨區】 學生限於國三學籍所在地申請入學招生區提出申請，其申請方式如下： 1. 只向一所高中提出。 2. 只向一所高職提出。 3. 同時向一所高中及一所高職提出。

實施對象	1. 國民中學應屆畢業生取得國民中學學生基本學力測驗分數者。 2. 非應屆國民中學畢業生及同等學力取得當年度國民中學學生基本學力測驗分數者。
實施方式	1. 學生依學校所訂條件，自行向欲就讀之學校或聯合申請入學委員會提出申請，各國民中學應提供必要之協助。 2. 各校以國民中學學生基本學力測驗成績為申請依據，採書面審查方式，不得再辦理任何型式之測驗。 3. 各校應參採學生之在校成績（限直升入學及自學方案）、特殊才能、優良品德或綜合表現等方面之具體表現。 4. 各校得考量社區地緣因素，提供若干名額予鄰近國中，其提供名額之原則應報經主管教育行政機關核定。
實施時間	1. 每年第一次國民中學學生基本學力測驗成績公布後。 2. 放榜時間應於每年第二次國民中學學生基本學力測驗報名前。

資料來源：整理自教育部中教司資料網

表 2.3　聯合登記分發入學招生簡表

實施範圍	【得跨區】 1. 學生可選擇一登記分發區參加分發。 2. 各區公立高中職應聯合辦理登記分發。 3. 各區高中職登記分發入學委員會與各區五專登記分發入學委員會應統合辦理分發。

實施對象	1. 國民中學應屆畢業生取得國民中學學生基本學力測驗分數者。 2. 非應屆國民中學畢業生及同等學力已取得當年度國民中學學生基本學力測驗分數者。
實施方式	1. 以國民中學學生基本學力測驗分數為分發依據，不得加權計分。 2. 以當年度一次國民中學學生基本學力測驗分數完整使用，並於登記分發入學報名時選擇一個登記分發區參加分發。 3. 學生應依登記分發入學招生委員會之規定提出申請，各國民中學應提供必要之協助。
實施時間	1. 第二次國民中學學生基本學力測驗結束後。 2. 於每年七月辦理為原則。

資料來源：整理自教育部師資培育及藝術教育中教司資料網

登記分發入學、甄選入學、申請入學，三種入學管道都適用「國中生基本學力測驗」分數，其中「登記分發入學」的名額須佔五○% 以上，仍居多元入學方案主流，此三種入學方式仍都要採計學業成績，於是「基本學力測驗」就成了聯考的替代品。由此可見，國中基測分數仍為高中職分發篩選之重要依據。

2.2 資料探勘法之決策樹規則與關聯規則相關研究

2.2.1 資料探勘法簡介

"Data Mining"在各式中文期刊或文獻中又譯為數據挖掘、資料採礦、資料挖礦、資料探勘。它是資料庫知識挖掘（Knowledge-Discovery in Databases，KDD）中的一個步驟。資料探勘法一般是指從大量的資料中自動搜尋隱藏於其中的有著特殊關聯性的訊息的過程（Association rule learning）。資料探勘通常與電腦

科學有關，並通過統計、線上分析處理（On-Line Analytical Processing，OLAP）、機器學習（Machine Learning）、專家系統（Expert System）和模式識別（Pattern recognition，又稱圖形識別）等諸多方法來實現上述目標[35]。

不同的學者對資料探勘不同的定義：Michael 等學者 [12]：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。Frawley 學者於 1919 年[6]提出從資料中提取出隱含的過去未知的有價值的潛在訊息。面對資訊爆炸時代來臨，電子化資料也愈來愈多，累積形成龐大資料庫。從資料庫中利用新的技術及工具，結合智慧化及自動化處理資料，轉換成有用的資訊及知識，愈顯重要。資料採礦的定義即是從大型資料庫中，探索與分析資料，擷取出有價值之資訊及知識，也就是將資料轉換成知識的行為[22]。

Simoudis 等學者則認為 [8, 9, 15]：資料採礦為資料庫知識發現的核心，其最終目的是為提出有價值的資訊，以做為最有利決策的支援。Grupe 與 Owrang [7]認為資料採礦是從已經存在的資料庫當中挖掘出專家仍未知的新事實。

Kleissner 學者認為資料採礦是一種新的且不斷循環的決策支援分析過程，它能從資料中發現隱藏價值的知識，以提供專業企業人員參考[10]。經過資料採礦後，所挖掘到的知識，其主要目的為提供豐富及前所未知之資訊，做為使用者知識發現及決策支援之用[13] [8]。資料採礦吸引人之處主要在於他具有建立「預報」（predictive）、而不是「回顧」（retrospective）模型的能力。

Fayyad 學者定義知識發掘（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[5]。資料探勘是一個確定資料中有效、新的與可能有用，並且模式最終能被理解的重要過程。資料探勘是可萃取出資料中有效、潛在效益的一項非細瑣流程，其最終目的系瞭解資料的樣式。Fayyad 學者認為資料採礦是資料庫知識發現（Knowledge Discovery in Database，KDD）的一部分[5]。圖 2.1 所示，資料庫知識發現的整個過程包含資料選取（Data Selection）、資料前處理（Data Processing）、資料轉換（Data Transformed）、資料採礦（Data Mining）、解釋評估（Interpretation Evaluation）等階段。資料庫知識發現的整個過程說明如下：

一、資料選取：從資料庫中選取所需分析的資料，再將其整合成為目標資料（target data）。

二、資料前處理：從目標資料中，清除不一致性和不需要的資料，對遺失的、多餘的、錯誤的以及無關係之資料做刪除或修正處理。

三、資料轉換：透過轉換或合併成為適合探勘的格式。

四、資料採礦：運用演算法來挖掘並取得資料樣式(patterns）。

五、解釋評估：經評估或辯認資料樣式是否令人感到興趣或是有存在的價值，將經過評估而有意義之資料樣式，依視覺化或其他技術將知識呈現。

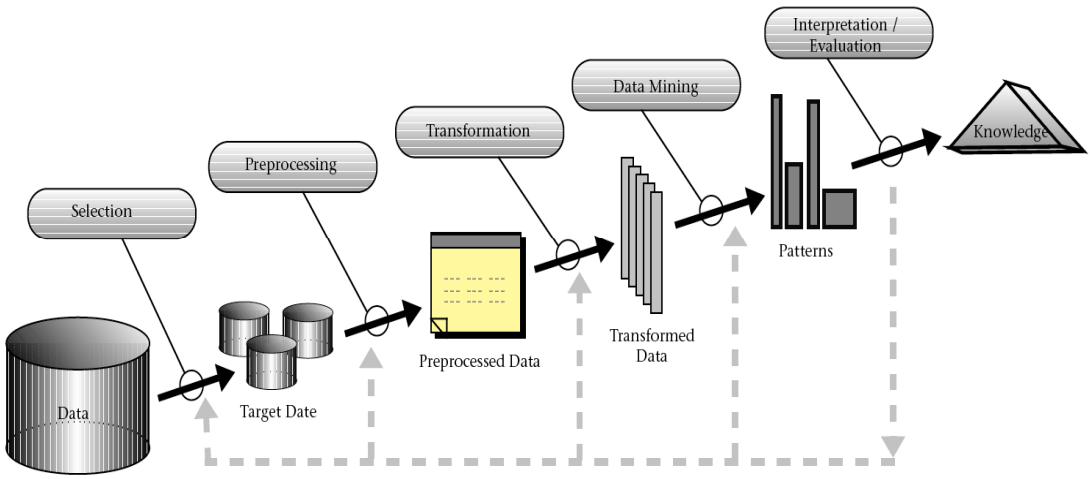

圖 2.2 KDD 演算法流程圖

資料來源：Fayyad（1996）[5]

Curt 學者指出資料探勘是一種資料轉化的過程，先由沒有組織的數字與文字所集合而成的資料，轉換成為有用的資訊，再將資訊轉換為知識，最後產生決策[4]，(如圖 2.3)。

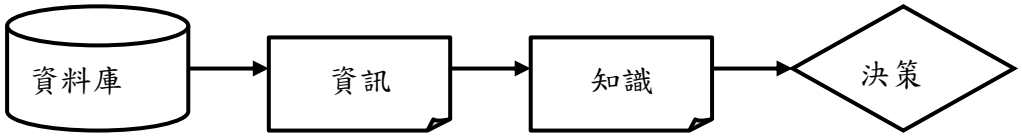

圖 2.3 資料採礦資料轉化過程

資料來源：Curt[4]

Peacock 學者指出自動發現隱藏在資料庫中有用但不明顯的模式，所謂有用的指可能會影響策略與戰略，最後影響到組織目標[14]。

一、狹義的資料採礦圍繞在機器學習的方法論上，也就是強調發現的過程。

二、廣義的資料採礦：在研究確認或測試發現過程中所揭發的關係。使用統計方法、建立假說，或研究並確認關係以支持在狹義的資料採礦中形成的模式，一般都在強調人的學習以及管理者與分析者的投入。

三、最廣義的定義則與資料庫知識發現（KDD）同義。包含內部與外部資料獲取、清理、資料轉換、格式化、分析、確認、賦予資料意義、建立與執行決策支援工具和系統，使得資料採礦的結果對決策者發生效用，並能不斷對模型加以修正與系統維護。

2.2.2、資料探勘的模型

資料探勘的模型主要有以下四種：資料分類(data classification)、資料關連(data association)、資料分群(data clustering)以及循序樣式探勘(sequetial pattern mining)。只要能建立並充分運用這些模型，即可探勘出潛藏的有用資訊[31]。

一、資料分類：

分類(classification)可按照分析對象的屬性分門別類加以定義，建立模組(class)[21, 26, 36]。例如：將學生的程度，區分為高分組、中分組和低分組三種組別。此模型可以用以對一些已經分類的資料來研究它們的特徵，在根據這些特徵對其他未分類或是新的資料庫作預測。用以找出特徵的已分類資料通常來自現有的歷史資料，或是對一個完整資料庫作部分取樣，在經由此模型作預測。例如：找一大群學生來取樣建立分類模型(classification model)，找出其特徵然後再對其他學生作預測。

二、資料關聯：

此模型是探討資料項目間的關係，找出在某一事件或是資料中會同時出現的項目，也叫做關聯規則探勘[33]。例如：如果學生具備 A 入學條件，則這個學生選擇 B 校就讀的機率是 70%。關連法則探勘的優點在於試圖找出多條規則，而且每一條規則都可以得到一個相對應的結論。然而缺點就是需要花費較多的時間，而且產生出來的規則不能夠像標準模型，好比是決策樹或是類神經網路，那麼直接可以用於預測。

三、資料分群：

此模型可以自動將資料庫區隔為幾個特性接近的資料群集，其主要的功能是將群集與群集之間的差異找出來，同時也可以將同一群中成員們的相似性找出來。群集分析(clustering)與分類(classification)不同之處在於你不曉得它會分成多少群或是根據什麼特徵來分群，所以必須分析解讀分群之後各群集所代表的意義。而群集分析會將性質相似的、特性可預測的資料組織至群集中。一般來說，這些相似的特性可能基於某些因素而被隱藏，或者是非直覺可察覺的。例如：有一群學生，他們在加法表現不好，在乘法表現不好，在分數的表現也不好。透過觀察這些資料是為何被群集在一起的，可以更了解資料間的關係，以及這些關係將會如何影響預言的結果。

四、循序樣式探勘：

循序樣式探勘主要是用在分析一些與序列相關的資料[25, 40]，這些資料也包含離散的序列。通常在序列中的屬性值都有特定的次序，例如學生學習概念的順序。在這些包含次序的資料中進行關連法則探勘，找出是件發生先後順序的關連性，這便是循序樣式探勘的目的。且循序樣式探勘所得到的結果往往可以用來作為趨勢預測的依據之一，例如：以現在學生在概念 A 的學習表現，預測在下一個概念 B 的學習表現，以便教師作補救教學的準備。

本研究主要探討入學成績與畢業成績間的關係，利用關聯法則探勘模式，找出在某一入學條件之下或是某一入學條件中會同時出現的項目。而關聯法則探勘的優點在於試圖找出多條規則，而且每一條規則都可以得到一個相對應的結論。雖然需要花費較多的時間，但幸好以現在的電腦運算速度，皆能克服此一問題。

2.2.3 決策樹分析(Decision Tree)

決策樹(Decision Tree)是指對所有已知的資料做分類或預測，並且利用資料中各個不同屬性值將資料分割成許多單一類別的子集合；是一功能強大且相當受歡迎的分類和預測工具；而且一次只觀察一個變數/屬性，以找出含最多資訊的變數。而我們唯一在意的東西，是分類或預測的正確性。

決策樹這種以樹狀圖為基礎的規則找尋法，其吸引人的地方在於它具有規則，其產生出來的結果容易讓使用者了解與明白，因為使用者可以在不具備任

何統計以及分析的知識下即可藉由決策樹來分析顧客以消費者的特質；決策樹的每一分支，就是一種對單一變數的檢驗，此檢驗會把空間分成兩塊或更多塊。一筆資料從根部的節點進入決策樹；在根部，應用一項測驗來決定這筆資料該進入下一層的哪一個子節點(Child Node)。選擇一開始的測驗有不同的演算法，但目的都是一樣的；這個過程一再重覆，直到資料到達葉部節點(Leaf Node)。所有到達某一個決策樹葉部的資料都以相同的方法來分類。從根部到每一個葉部都有一套獨特的路徑，這個路徑就是用來分類資料的規則的一種表達方式。

然而決策樹生成之後，可能會非常的複雜以及龐大，所以我們必須要對決策樹做修剪的動作；而修剪的最主要目的是為了移除葉節點與枝節的過程，其最終結果則是為了要改善決策樹的效能。

在決策樹中，比較著名的演算法有以下三種：ID3 或 C4.5、CART、CHAID，其功能我們分別敘述如下：

一、 ID3(Interactive Dichotometer 3)：

是指交互作用二分法的意思，是在 1983 年由 Quinlan 所提出的；由它的名稱，我們便可以清楚地知道 ID3 的邏輯特徵，意指如果在資料中有某項顯著特徵，它就會以此特徵將資料分成兩群，然後這兩群中如果又有一個為特殊特徵，就分為二，以此類推，反覆運作，直到所有同一特徵的資料都在一個類別裡為止。其做法是，首先選擇一個最佳的特徵作為根節點，由根節點開始，把所有

訓練資料依照此特徵分配到其所屬的分支，如果有某一個分支的訓練資料都已是同一個類別時，則此分支即成為一個樹葉，如此，這個分支的推導也就算完成，可以結束了。同時，這個分支也就是一個分類法則。否則，只要在尚未找到樹葉時，分類的動作就得繼續下去，直至所有資料皆屬同一特徵、同一類別才能宣告完成。這個系統也之稱為觀念學習系統，是各種演算法中最典型的一個方法。

C4.5 是 ID3 的改良版，其運作的原理和過程皆和 ID3 相同，只不過 C4.5 可以處理遺漏的預測子和含有連續值的預測子；此外，C4.5 更加入了決策樹的修剪功能以及法則轉換功能。

二、 CART(Classification and Regression Tree)：

分類迴歸樹；其最大的優點之一是演算法會自動檢驗模型，找出一個最佳的一般模型。這是由 Leo Breiman、Jerome Friedman、Richard Olshen 和 Charles Stone 在 1984 年所提出的演算法。CART 的方式是先建立一棵非常複雜的樹，然後再根據交互簡易或測試集檢驗的結果，將決策樹修剪成最佳的一般樹。在修剪之時，CART 是以整體錯誤率(Entire Error Rate) 做為修剪依據，期望以最小的樹，也就是最少層的樹得到最有效的分類。CART 會針對任何節點先檢查節點中的資料是否屬於同一類別，若節點中的資料只屬於同一類別，則此節點不需要再分割。若節點中仍有二個以上的類別，則 CART 會測試所有資料屬性。

非連續屬性會檢查所有合併方式後，選擇最佳方法將屬性值合併為二組；連續屬性則會選擇最佳分離值將一為二。CART 和 C4.5 不同的是，C4.5 只根據訓練資料集來決定修剪結果；而 CART 則是根據各種版本對測試集資料的效能，來決定修剪的結果。

在這，必須有一個觀念，那就是最複雜的樹不一定都是最好的樹，因為這樣有可能造成過度適應(Over fitting)的情形發生。此外 CART 在處理遺漏資料方面是相當拿手的，通常若某個屬性有遺漏值，就無法在建立樹的時候用來決定佳分割,事實上 CART 會儘量利用手上所有的資訊,將遺漏值透過替代(Surrogate)的方式處理。而替代是樹中真正分割值和投入屬性的模擬，可以在投入屬性遺漏時使用。

三、 CHAID(Chi-Square Automatic Interaction Detector)：

卡方自動互動偵測；此方法是三種決策樹演算法最古老的一種，是由 J.A. Hartigan 學者在 1975 年先提出的,而它是從 1964 年 J.A. Morgan 和 J.N.Sonquist 所提出的一套自動互動偵測系統 AID 所衍生出來的[9]。其原理是偵測變數之間的統計關係，藉此建構出一棵決策樹。CHAID 與 CART 不同的是，CHAID 是利用連續卡方來測試並決定哪一類的預測子最不受預測值影響。

決策樹的建立以及演算法雖然複雜，但是在資料挖掘中卻是最常見的，因為它可以輕易地將結果表達，轉換法則使人清楚了解。以下我們來看個例子，我們以圖 2.4 為例說明如下：

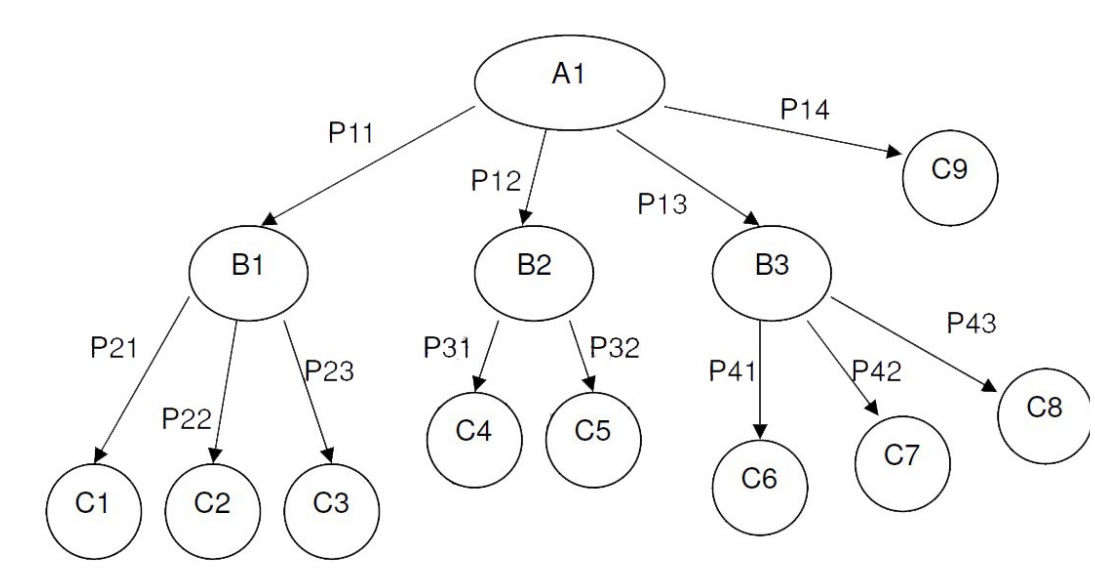

圖2.4 決策樹架構

決策樹轉換成法則的形式，只要將從根節點到樹葉所經過的路徑加上最後的決策，就可以轉換成對應的法則，圖 2.4 的決策樹經轉換形成之法則如下：

法則1：IF A1=P11 and B1=P21 THEN C1

法則2：IF A1=P11 and B1=P22 THEN C2

法則3：IF A1=P11 and B1=P23 THEN C3

法則4：IF A1=P12 and B2=P31 THEN C4

法則5：IF A1=P12 and B2=P32 THEN C5

法則6：IF A1=P13 and B3=P41 THEN C6

法則7：IF A1=P13 and B3=P42 THEN C7

法則8：IF A1=P13 and B3=P43 THEN C8

法則9：IF A1=P14 then C9

關聯法則方式則不需要依照某種固定順序[1]，因此，在某些狀況下，以關聯法則方式所推得的法則會比用決策樹方式推得的關聯法則短。值得注意的是決策樹可以很容易轉換成關聯法則的形式，但是關聯法則未必能轉成決策樹的形式。

2.2.4 關聯性法則（Association Rule）

Agrawal 等學者指出[1]，在資料探勘的領域之中，關聯性法則(Association Rule)是最常被使用的方法。關聯性法則在於找出資料庫中的資料間彼此的相關聯性，而這種方法現已經普遍運用於各領域之中；例如，80%消費者購買碳粉匣，也會購買報表紙。假設在資料庫中，L = {l1, l2, …,ln} 是所有顧客的知識與需求之集合，其中 X 及 Y 均為決策變數且是 L 的子集合 (Subset) 並互相獨立，因此關聯式法則的表示形式為：X→Y， X→L，Y→L 且 X ∩ Y=ψ。

關聯式法則的產生由兩個參數來決定：支持度 (Support) 及可靠度 (Confidence) [16]。關聯性法則的建立，按照 Agrawal& Srikant (1994) [2]兩位學者所設計的流程，有以下二個步驟:

(1) 從資料庫中找出高頻的項目集合(Large Itemsets)，亦即此集合之各個決策變數的組合，同時要大於所設定之最低支持度(Minimum Support)。

(2) 接著，用前述步驟所產生的高頻項目集合產生關聯性法則，並計算其可靠度，若高於所設定的最低可靠度 (Minimum Confidence)，則此法則確定成立。

其處理程序說明如下：

(1) 定義最低支持度 (Minimum Support) 及最低可靠度 (Minimum Confidence)。

(2) Apriori 演算法使用了候選項目集合 (Candidate Itemsets) 的觀念，若候選項目集合的支持度大於或等於最低支持度 (Minimum Support)，則該候選項目集合為高頻項目集合(Large Itemsets)。

(3) 首先由資料庫讀入所有的交易，得到第一候選項目集合(Candidate 1-Itemset)的支持度，再找出第一高頻項目的集合(Large 1-Itemset)，並利用這些高頻單項目集合的結合，產生第二候選項目集合 (Candidate 2-itemset)。

(4) 再掃描資料庫，得出第二候選項目集合的支持度以後，再找出第二高頻項目集合，並利用這些第二高頻項目集合的結合，產生第三候選項目集合。

(5) 反覆掃描整個資料庫，再與最低支持度相比較，產生高頻的項目集合，再結合產生下一層候選項目集合，直到不再結合產生出新的候選項目集合為止。

以下則利用簡單的例子，來看 Apriori 演算法的處理過程。若資料庫中有四筆交易，每筆交易都具有不同的 ID 作代表，而交易中都包含了有數種物品，如下表 2.4 所示：

表 2.4：資料庫中交易記錄

ID	Items
001	ACD
002	BCE
003	ABCE
004	BE

則 Apriori 產生候選項目集合和高頻項目集合的計算流程如下：首先在掃瞄完整個資料庫後，將所有出現商品的次數予以計數，如此即得 C1 表（第一候選項目集合），將不符合最小支持度之項目剔除後，即得 L1 表（第一高頻項目集合）。藉此反覆遞迴的過程，依次產生第二高頻項目集合與第三高頻項目集合(如表 2.5)。

表 2-5: Apriori 演算法產生的候選項目集合和高頻項目集合

Scan Database →

C1	
Itemset	Support
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

→

L1	
Itemset	Support
{A}	2
{B}	3
{C}	3
{E}	3

C2
Itemset
{AB}
{AC}
{AE}
{BC}
{BE}
{CE}

Scan Database →

C2	
Itemset	Support
{AB}	1
{AC}	2
{AE}	1
{BC}	2
{BE}	3
{CE}	2

→

L2	
Itemset	Support
{AC}	2
{BC}	2
{BE}	3
{CE}	2

C3
Itemset
{BCE}

Scan Database →

C3	
Itemset	Support
{BCE}	2

→

L3	
Itemset	Support
{BCE}	2

資料來源：Kouris, I. N., Makris, C. H., Tsakalidis, A. K. [11]

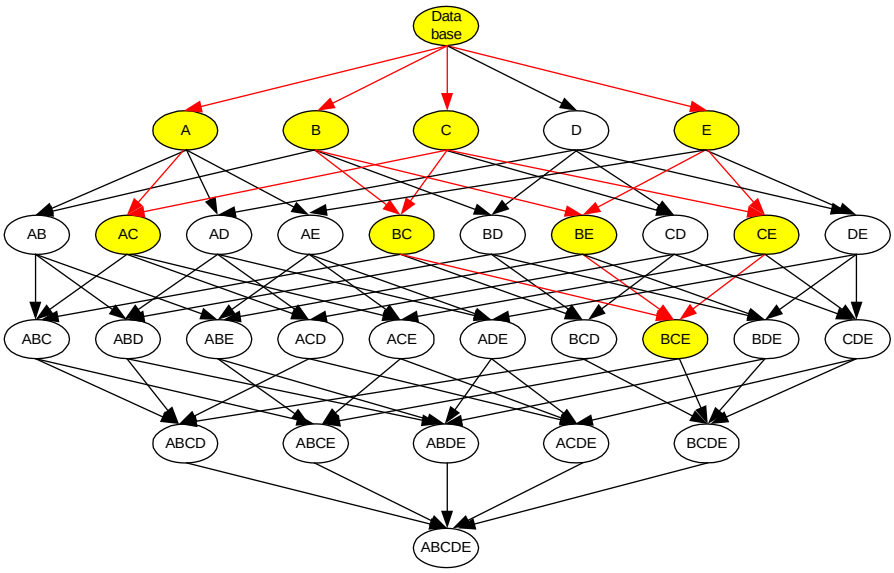

圖 2.5：五維度的子集合示意圖

資料來源：Coenen, F., Goulbourne, G., & Leng, P. [3]

當我們想要產生第三候選項目集合時，所產生的集合項目中，必須皆已產生於第二高頻項目集合中，由圖 2.5 可以很清楚的看到整個演算的路徑。因此第三候選項目僅剩 {BCE}，無法再產生 C4，所以演算法就此終止。

在關聯式法則之使用中，Apriori 是最為著名且廣泛運用的演算法。最早是由 Agrawal & Srikant 等兩位學者於 1994 年首先提出[2]，而在這之後許多應用的相關演算法，僅是修正 Apriori 中的部分概念而來，以增強 Apriori 的演算效能，例如 DHP 演算法、DLG 演算法、DIC 演算法與 FP-Tree 演算法等。

表 2.6 比較決策樹與關聯規則之優缺點。

表 2.6 決策樹與關聯規則比較表[12]

工具	決策樹	關聯規則
優點	※明白指出最佳變數的能力 ※產生易於了解的規則 ※節省分類時的計算時間 ※可處理連續與類別變數	※能完整的呈現各變數之影響，與變數組合 ※能產生簡單明瞭的結論 ※適用不同形式的原始資料 ※計算模式簡單易懂
缺點	※當類別太多時，資料量小，錯誤會增加的比較快。無法全面的看到各種變數的影響程度 ※對有時間順序的資料，需要很多欲處理的工作	※結果多且凌亂，難已決定哪一變數較重要與適當的組合 ※當問題變大時，運算量會成幾何級數增加 ※對於資料的個別資訊不甚重視 ※容易剷除罕見變數

2.3 入學成績、在學成績與畢業成績之關聯性研究

劉玉春等學者[42]在民國 78 年接受大考中心之委託，進行「高中學生在學三年成績與大學入學考試成績相關性之研究」。而其研究結果發現：(1)比較各校之大學入學成績的高低情形，可知與高中在學成績有密切的關係，即學生的學業成就及學業性向之間有密切關係存在。(2)各校若由在學成績轉換為大學入學考試成績之當量分數時，其間必然有加權係數之存在。

部份研究者提出其他變項來檢視高中在校成績與大學學測之相關性分析，如陸炳杉 [29]學者以性別、入學管道與各科表現來進行研究，結果提出一、就

在校學業成績現況而言，不同性別學生，女生平均分數高於男生；不同入學管道學生，以直升學生平均分數最高，推薦甄選學生最低。二、就在校學業成績差異而言，不同入學管道學生在校學業成績，直升學生顯著優於推薦甄選學生。三、就大學學科能力測驗成績現況而言，不同性別學生，女生在國文科、英文科、社會科三科成績平均高於男生；男生在數學科、自然科、總級分成績平均高於女生。四、就大學學科能力測驗成績現況而言，不同入學管道學生大學學測各科成績平均高低不同。五、就大學學科能力測驗成績差異而言，不同性別學生，女生在國文科、英文科、社會科三科成績顯著優於男生；男生在數學科、自然科、總級分成績顯著優於女生。六、就大學學科能力測驗成績差異而言，不同入學管道學生大學學測成績在數學科、自然科、總級分有顯著差異。七、三種不同入學管道學生在校學業成績與大學學科能力測驗英文科、總級分成績均有顯著正相關。

溫侑柯 [38]學者探討大學入學聯招成績與南華大學學生在學成績之相關性。研究結果顯示英文入學成績好的同學，在學校各科的整體表現上也較好；〝歷史入學成績好〞的幾乎所有在學科目都很差；〝歷史入學成績差〞的在學成績幾乎都好。造成此結果的主要原因，除了教師教學大都使用原文書外，另外一個原因就是，入學時總分相差不大，因而與國文英文入學成績好的同學其歷史地理相對較差，國文英文入學成績差的同學其歷史地理相對較好有關。因

此各校系可參照其研究的方法及結果，找到各校系的入學標準，進而慎選學生，達成各校系的特色與目標。

另外，陳怡靖等學者[27]以全國性大樣本資料「臺灣教育長期追蹤資料庫」（TEPS）做分析的結果顯示：相較於低社經背景的學生，高社經背景的學生在多元入學中的確占了優勢；他們有較多機會透過聯考入公立高中，與直升私立名校，而且高中成績較佳。低社經背景的學生在聯考與直升中都居劣勢，他們較多以申請或登記分發進入私立高中，成績較差，不過透過推甄進入公立高中的機會，並不低於高社經背景的學生。值得關注的是直升入學，家庭收入較高的學生，直升學費昂貴的私立名校來幫助其往後的升學，教育機會受社經背景的影響最明顯。而聯考也以高社經背景的學生占優勢，故回復聯考很可能仍難以降低教育機會不均等。

賴文漢 [43] 學者將科別，科目帶入研究變項，提出(1)統測成績與在校成績呈顯著正相關。在國文、英文、數學方面：以英文相關係數最高，國文最低。(2)專業科目方面：專業科目(一)相關係數高於專業科目(二)。另外，其還以在校成績分別預測統測成績之國、英、數、專一、專二。

宋珮怡 [19] 學者以某公立高中 94 學年的高三為樣本，以其學測成績與在學成績作為研究範例，使用學生性別，類組，在學成績與模擬考成績等變數對學測成績作預測分析，透過複迴歸分析，類神經網路及二階段整合模型等三種

方式建立預測之最適模型。李佳玲 [20] 學者在 2001 年也做過類似的研究，其以高中在校成績預測大學學測成績，以簡單迴歸來做預測模型，而彭重恩 [30] 學者則是利用複迴歸建立預測模型。

陳超 [28] 學者探討國中基測及高三學生的模擬考與大學學測成績之相關性，利用相關性分析和線性迴歸等統計方式，對資料進行分析。傅說道學者以資料採礦的群集分析針對南部某明星高中之高三學生，以其入學國中基測成績與高中在校成績對於大學學測之相關性進行分析。

呂學智 [18] 學者探討四技二專統一入學測驗數學科成績與高職在校前五學期數學平均成績兩者的關系。並分析不同背景變項的進修學校學生對其兩者差異情形。

王建華 [17] 學者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（2007）學者則是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16 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

資料探勘的方法流程為資料收集、資料前置處理、資料倉儲建立、資料探勘、樣式評估、結果展示。而本研究已擁有學籍資料庫，可省略資料收集這一步驟，從資料前置處理開始，將所要探討之因素，在前置處理部份加入，並將整個資料庫做更新，勘誤的動作，使研究能夠方便進行，並得到預期的結果。

3.1 研究架構

本研究是利用 Microsoft SQL server R2 2008 Business Intelligence Development Studio 軟體來做資料探勘的分析，此套軟體提供許多資料探勘的演算法，例如：時序群集、類神經網路、決策樹、貝式機率分類、關聯規則等等。本研究主要為探討學校入學方式、入學成績及不同學期間的成績，了解彼此間的相互關係，所以分別採用關聯規則中決策樹演算法及關聯式規則來做此分析，接下來本研究將分成五個階段來說明，第一階段為原始資料庫形式與資料格式，第二階段為前置處理，前置處理包含增加及刪除表格及欄位，並將多個學期的交易資料表獨立匯到新的資料庫；第三階段為前置處理後資料庫表單資料格式；第四階段為實驗與結果分析，最後為結論。

本研究資料之學生歷年成績資料與學生學籍資料共 1730 筆進行資料前處理，整理出符合分析範圍的資料，進行資料前處理，刪除不需要的欄位，並依需求新增欄位，資料處理後共計 762 筆，利用關聯規則找出學生特質分析的結果與招生策略上的建議，研究架構如圖 3.1。

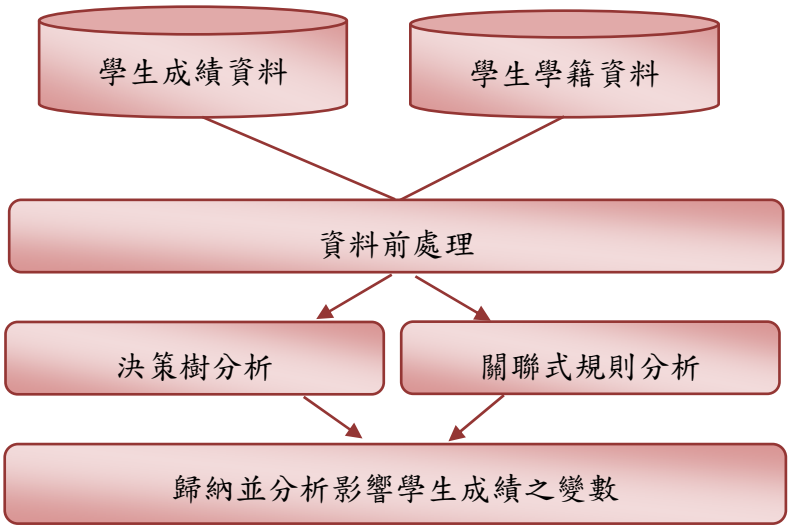

圖 3.1 研究架構圖

3.2 資料分析範圍

取得的資料範圍為 98 學年度至 100 學年度的學生歷年成績資料與學生學籍資料共計 1730 筆，學生學籍資料如表 3.2。

本研究在資料的分析過程中，將學生各學期之學期成績視為獨立互不影響的資料，而不以學生整體表現為分析主軸，因此在成績資料表中能取得學生於

高一至高三各學期的學習成效，即學期成績；而在學籍資料中可取得學生背景資料，例如畢業學校、性別、入學身份等。

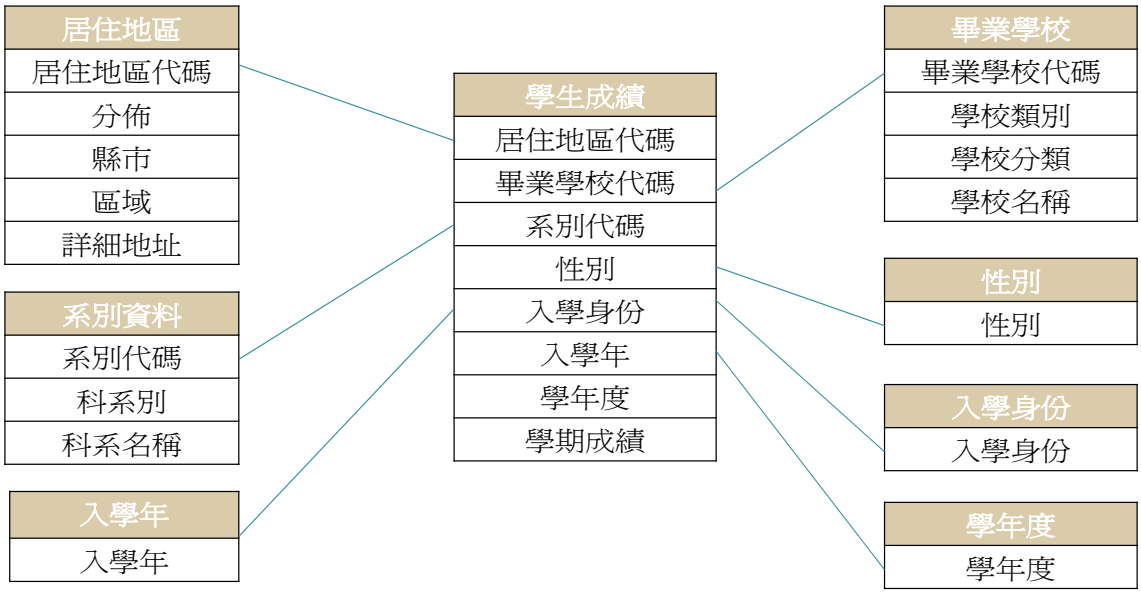

圖 3.2 學籍資料圖

本研究並將學生資料做一簡單分類，以了解該校特性。

一、性別比例

由圖 3.3 性別比例圖可看出，該校雖男女兼收，但以男性居多，男女生比例為 83：17。

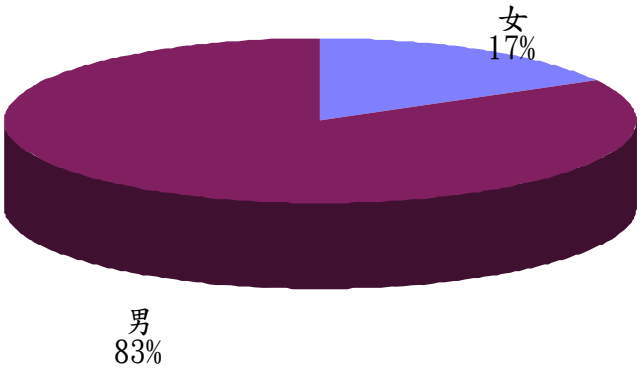

圖 3.3 性別比例圖

二、科系比例

高職以教導專業知能、培育實用技術人才，做為將來職業生涯之基礎，希望學生可以具備基礎專業知能，在畢業後直接進入就業市場，而高職類科包含了工業、商業、農業、家事、海事、水產、藝術等類科。本研究學校為工商職業學校，其學制包含：綜合高中，高職，及實用技能學程。其中高職包含汽車科，資訊科，電子科，電機科，電子商務科。實用技能學程包含汽車修護科，微電腦修護科，廣告科，烘焙食品科。由圖 3.4 科系比例圖可看出，該校以綜合高中人數最多，佔 21%，其次為汽車和資訊，分別為 15%，13%。

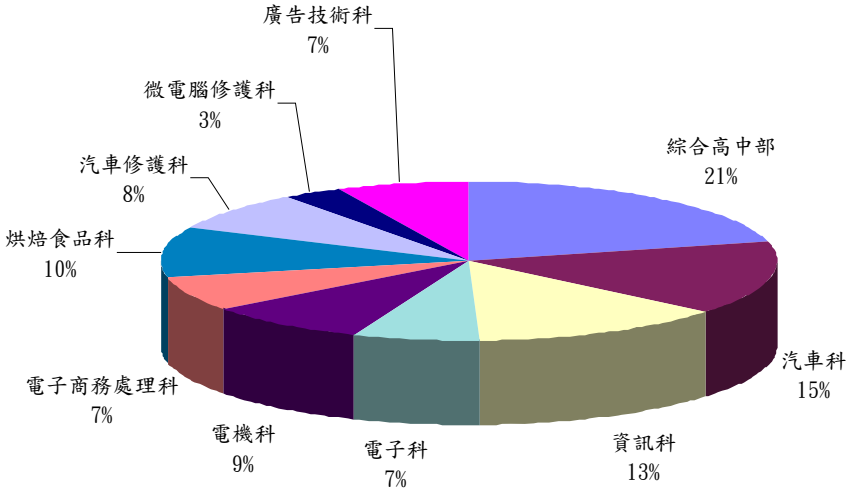

圖 3.4 科系比例圖

三、入學身份比例

入學身份包含一般生，外國學生，身心障礙生，原住民及其他；身份為其他主要是包含生活扶助戶（低收入戶），更生受保護人家庭，長期失業家庭，單薪且單親家庭或其他特殊之家庭；由圖 3.5 入學身份比例圖得知，該校學生之身份比例以一般生佔全體學生 94.18%為最高。身心障礙生，原住民，外國學生及其他身份之學生佔 5.82%。

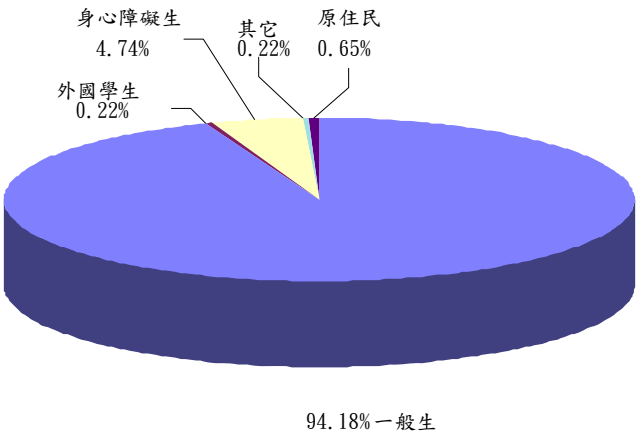

圖 3.5 入學身份比例圖

四、入學方式比例：

教育部於 98 年 9 月 4 日發布「擴大高中職及五專免試入學實施方案」，99 年起各公私立高中職及五專皆須辦理免試入學，並逐步提高免試入學名額比率。本研究取得之資料為 98，99，100 三個年度，99 年初辦免試入學，入學名額較少。由圖 3.6 入學方式比例圖，可知入學方式以自行報名最多，佔 65%。

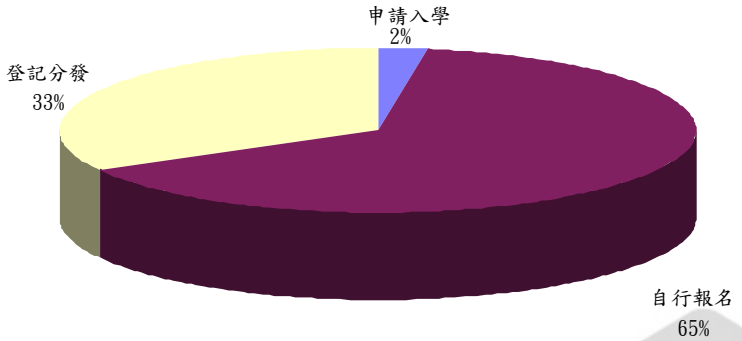

圖 3.6 入學方式比例圖

五、成績比例：

由圖 3.7 各學期成績比例圖可看出，成績比例於各學期中，以 B 為最大宗，佔大約 50%左右，成績為 A 者，各學期大約 20%。

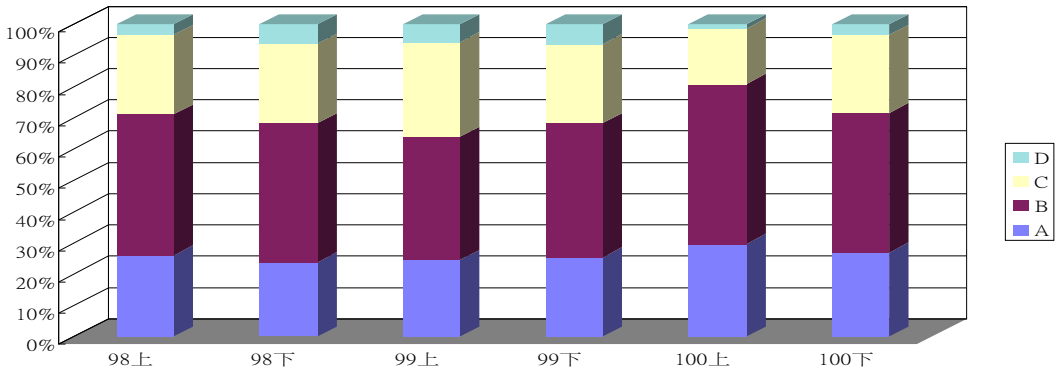

圖 3.7 各學期成績比例圖

六、畢業國中比例：

由圖 3.8 畢業國中比例圖可看出，該校以大灣高中附設國中部就讀比例最高，其次是復興國中，再來是歸仁國中。

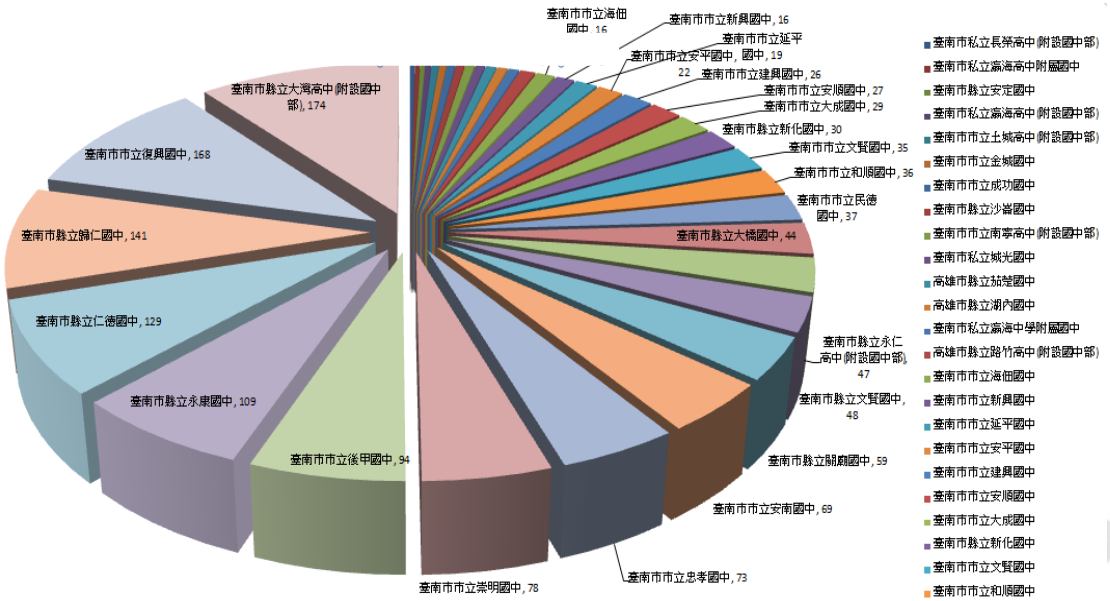

圖3.8 畢業國中比例圖

3.3 資料前處理

有好品質的資料，才有好品質的探勘結果。因此在進行資料探勘前，需進行資料前處理，以使資料更適合進行探勘的工作；而且，在整個探勘過程中，前置處理也是最耗費時間，同時也最影響探勘品質。本研究將資料前處理分成三個部分，即資料淨化（data cleaning）、資料整合（data integration）、資料轉換（data transformation），如圖 3.9 所示。以下分別就各部分說明：

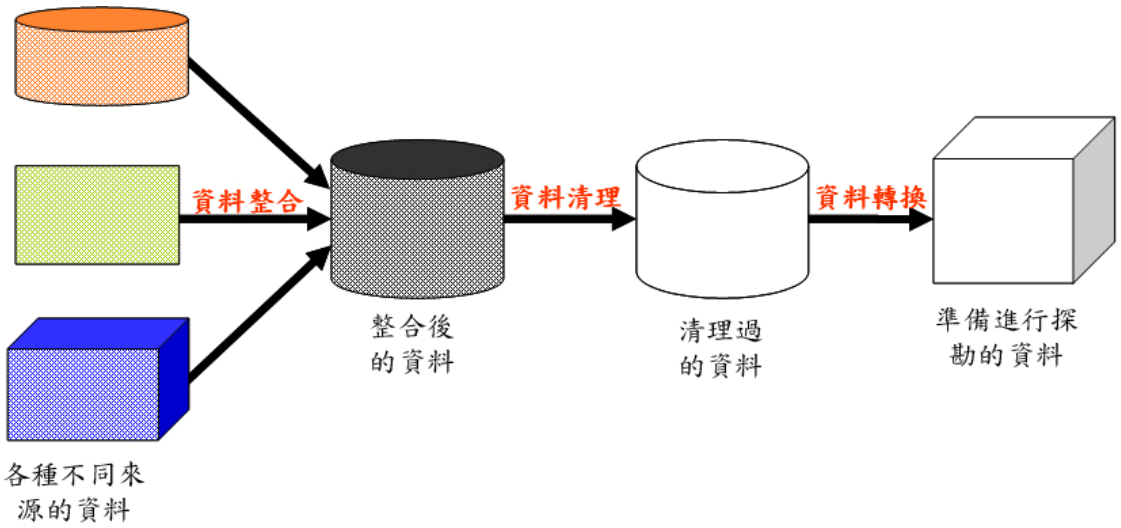

圖 3.9 資料前置處理的工作圖[31]

3.3.1 資料整合

資料整合主要是解決多重資料來源的整合問題。而資料整合的主工作有二：

(一)、消除資料不一致

- 數值不一致 (data value conflict)
- 綱目不一致 (schema conflict)

(二)、消除資料重複性

- 數值重複
- 綱目重複

學期成績資料表和學生學籍資料表將依學號欄位進行資料整合，在進行整合前，由於學期成績平均資料表為文字資料型態，學號為文字資料型態，而學生學籍資料表為微軟 Excel 格式，處理方式會先將學籍資料轉為 Access 格式，其學號資料型態與大小比照學期成績平均資料表。

在整合過程若遭遇數值不一致(如同一學號，卻性別不同)，同學號出現兩筆資料情形，則以人工判斷其正確性，若無法判斷則予以刪除；或綱目不一致(如學期平均資料內採用學號欄位，學籍資料表採用學號 ID 欄位)，名稱不同，但實際所代表的意義是一樣的，可以透過屬性更名進行統一。資料重複性的情況同樣會發生在數值和綱目兩部分，同理亦可以人工判斷將資料重複性都消除。接下來便可進行資料清理的動作，使資料更適合探勘的需求。

3.3.2 資料淨化

資料清理的步驟主要是確認資料的正確性以及完整性，使得探勘工作能夠順利進行，在資料的正確性方面，本研究進行下列資料清理的工作如表 3.1 所示。

表 3.1 資料檢查說明表

檢查內容	說明
屬性的有效值或有效範圍	例如：性別屬性的值不是男性就是女性；生日的月份應該介於 1 和 12 之間。
數值的唯一性	例如：身分證字號或是學號不可有重複。
參考完整性 (referential integrity)	例如：存在於學籍資料表中的學號必須同時存在於學期成績資料表中。
資料的合理性驗證	例如：從學籍資料表的生日計算出該學生的年齡只有 10 歲，學生就讀高職的年紀應該 15-20 歲，顯然不合理。

對於成績資料的遺漏值處理，例如學期成績，則追溯該生於當學期修課之課程學期成績，以計算學期平均，若為無法追溯部分，則比照學籍資料的遺漏值處理，即刪除該生所有資料。對於資料雜質（noise data）和不一致的資料，例如學號原為 6 碼，前一碼為入學年，若不符合該原則，為避免影響資料之處理，則一律予以刪除。另外，對於休學及退學，中途轉入或轉出之轉學生因資料不齊全，亦予以刪除。經過清理的步驟，修正原本不正確或是有缺漏的資料，將更有利於探勘工作。

3.3.3 資料轉換

本研究在學生成績資料表中，即新增欄位「學期成績等級」。等級的定義，乃企圖跨越不同入學年、科系別及不一致的評分標準，將學生學期成績以一標準值取代，如此則能以相同的標準判讀學生的學期成績表現。本文試圖將文字

資料轉換成數值，以便 SQL 程式執行。性別欄位轉換如表 3.2，科系名稱轉換如表 3.3，身份欄位轉換如表 3.4，入學欄位轉換如表 3.5，各學期成績轉換如表 3.6。

表 3.2 性別欄位轉換表

性別欄位的轉換	轉換前	男	女
	轉換後	1	2

表 3.3 科系名稱欄位轉換表

科系名稱欄位的轉換	轉換前	資訊科	汽車科	烘焙食品科	電子科	汽車修護科	電子商務處理科	電機科	廣告技術科	綜合高中部	微電腦修護科
	轉換後	305	303	K21	306	L01	425	308	M03	109	L37

表 3.4 身份名稱欄位轉換表

身份名稱欄位的轉換	轉換前	一般生	身心障礙生	原住民	其它	外國學生
	轉換後	0	4	1	12	15

表 3.5 入學類別欄位轉換表

入學類別欄位的轉換	轉換前	自行報名	登記分發	申請入學	聯招生
	轉換後	5	6	3	1

表 3.6 各學期成績欄位轉換表

各學期成績欄位的轉換	轉換前	90-100	80-89	70-79	60-69
	轉換後	A	B	C	D

無論是資料整合、資料清理還是資料轉換，前置處理的動作都是在於提升資料的品質，有好品質的資料，才有好品質的探勘結果。

3.4 實驗設計

本節主要為探討入學成績及學期成績，畢業成績之間的關係，因此選擇利用決策樹及關聯規則分別對該資料庫進行分析。3.4.1 介紹決策樹實驗設計，3.4.2 介紹關聯規則實驗設計。

3.4.1 決策樹實驗設計

在進行資料探勘中的決策樹產生的規則後，針對性質較相近的學生彼此做歸類，或探討在不同學生族群之間的流動分析。

其中，本研究之實驗設計乃利用學籍資料庫及學生成績資料庫挖掘出學生入學成績、前後學期成績、性別、入學方式、入學身份、畢業國中這此因素之間的關聯性。

各個實驗之參數設定、差異比較之流程，步驟如下：

步驟 1：選入所需之變數。

步驟 2：設定演算法參數。

步驟 3：調整最小案例數及複雜性懲處。

步驟 4：分析入學變數對入學後成績之關係。

步驟 5：結論與建議。

3.4.1.1 決策樹參數說明與設計

探勘前必須勾選所需變數來做決策樹分析，因此本研究以「學號」為索引鍵，利用「學籍資料庫」、「成績資料庫」分別以「入學成績級距」、「性別」、「入學方式」、「入學身份」、「分區」為輸入變數，如表 3.7，預測變數為「100 下學期成績(畢業成績)」，如圖 3.10，來進行資料探勘的決策樹分析。

表 3.7 決策樹輸入之變數表

索引鍵	資料庫	輸入變數	資料屬性	資料長度
學號	學籍資料庫、成績資料庫	入學成績	文字	8
		98 上學期成績	文字	8
		98 下學期成績	文字	8
		99 上學期成績	文字	8
		99 下學期成績	文字	8
		100 上學期成績	文字	8
		100 下學期成績	文字	8
		性別	文字	8
		科系	文字	10
		入學方式	文字	10
		入學身份	文字	10
		居住區域	文字	10

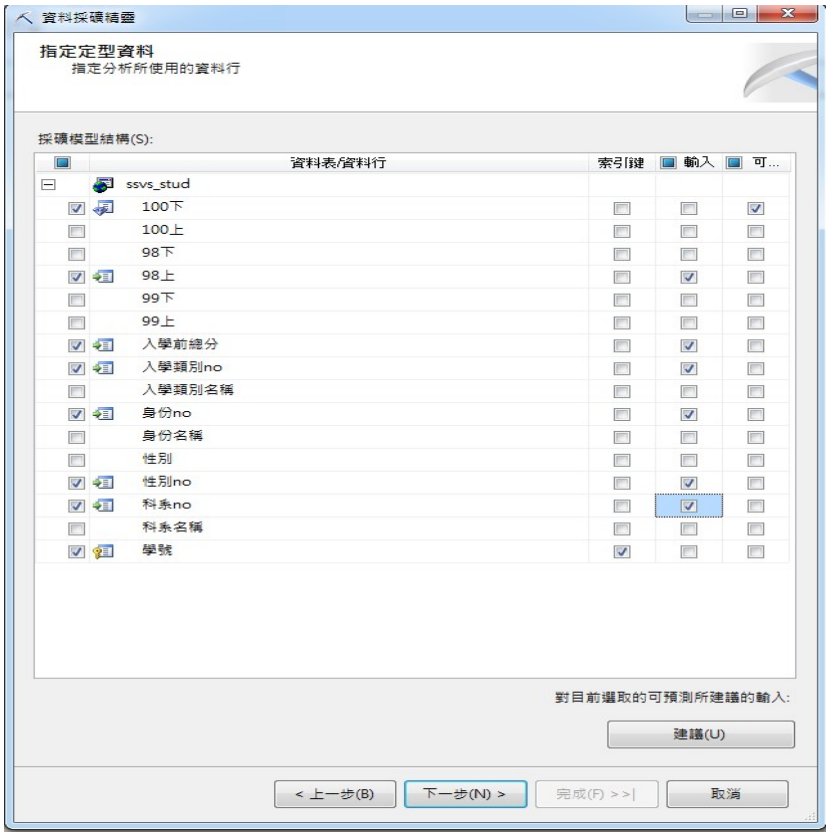

圖 3.10 決策樹指定分析資料設定圖

3.4.1.2 決策樹參數設定

決策樹參數設定之意義如表 3.8。決策樹參數設定介面如圖 3.11。

表 3.8 決策樹參數設定說明圖

參數	描述
MAXIMUM_INPUT_ATTRIBUTES	定義在叫用功能選項之前，演算法可以處理輸入屬性的數目。將此值設定為0 來關閉功能選項。 預設值為255。
MAXIMUM_OUTPUT_ATTRIBUTES	定義在叫用功能選項之前，演算法可以處理輸出屬性的數目。將此值設定為0 來關閉功能選項。 預設值為255。

SCORE_METHOD	決定用來計算分岔準則的方法。可用的選項：Entropy (1)、Bayesian with K2 Prior (2) 或Bayesian Dirichlet Equivalent (BDE) Prior (3)。 預設值為3。
SPLIT_METHOD	決定用來分岔節點的方法。可用的選項：Binary (1)、Complete (2) 或Both (3)。 預設值為3。
MINIMUM_SUPPORT	決定要在決策樹中產生分岔所需的最小分葉案例數目。預設值為10。
COMPLEXITY_PENALTY	控制決策樹的成長。低值會增加分岔數目，而高值會減少分岔數目。預設值是依據特定模型的屬性數目，如下列清單所述： 針對1 到9 個屬性，預設值為0.5。 針對10 到99 個屬性，預設值為0.9。 針對100個以上的屬性，預設值為0.99。
FORCED_REGRESSOR	強制演算法使用指定的資料行作為迴歸輸入變數，不考慮演算法計算出來之資料行的重要性。此參數只用於預測連續屬性的決策樹。

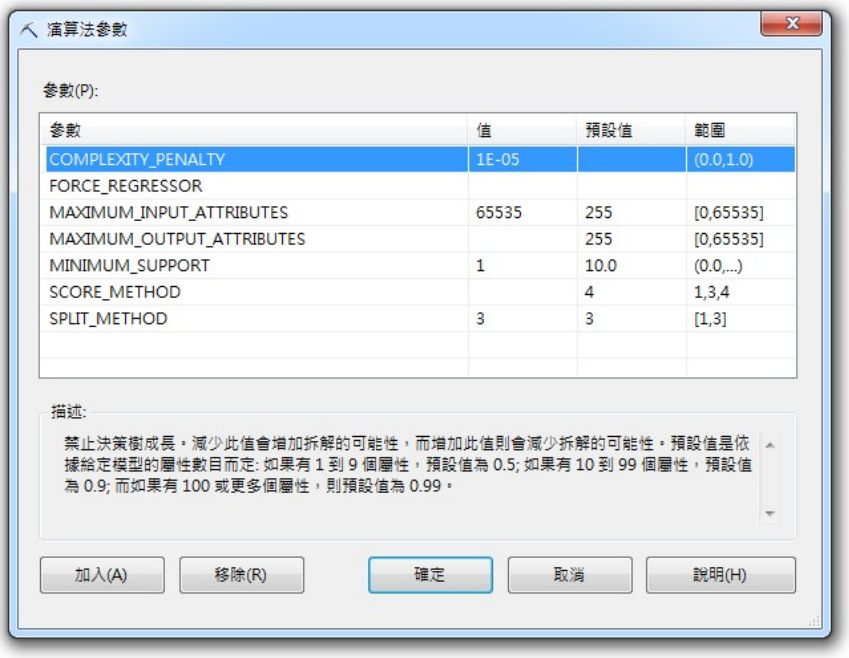

圖 3.11 決策樹演算法參數設定介面

3.4.1.3 決策樹設定測試資料及訓練資料

此設定目的在於使門檻值降低，增加可檢視的規則。在圖 3.12 的畫面中可以輸入測試資料的百分比(預設值為 30%，表示是用 70%建模)，或是可以輸入「在測試資料集內的最大案例數目」，系統會取其小者，作為測試資料。設定結束後，圖 3.13 為探勘後結果畫面

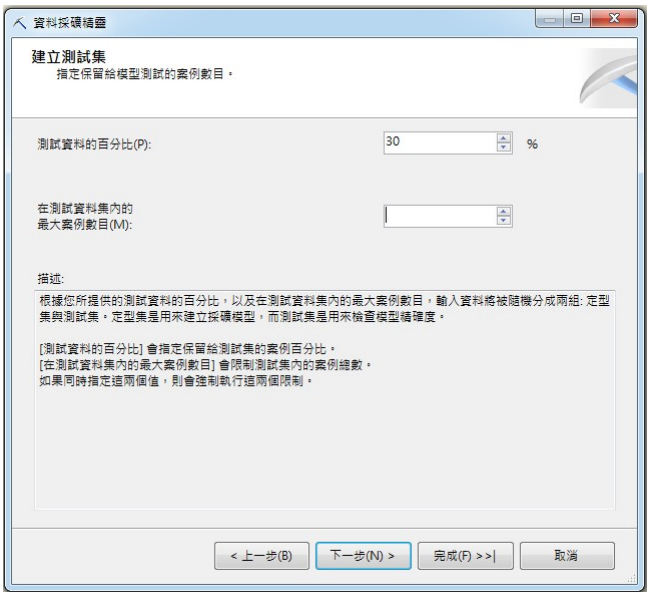

圖 3.12 決策樹建立測試資料百分比

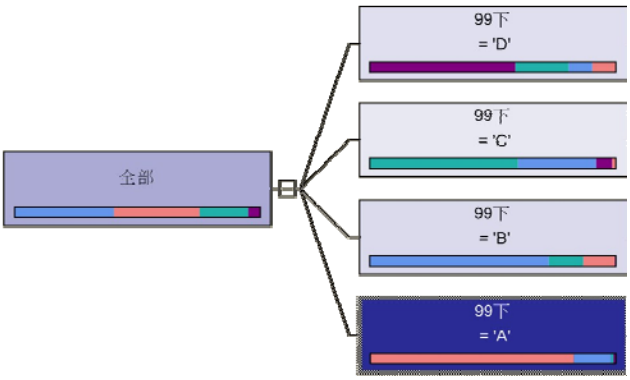

圖 3.13 決策樹分析結果畫面

3.4.2 關聯規則實驗設計

利用資料探勘中的關聯規則，針對性質較相近的學生彼此做歸類，或探討在不同學生族群之間的流動分析，預測學生成績。

其中，本研究之實驗設計乃利用學籍資料庫及學生成績資料庫挖掘出學生入學前成績、性別、入學方式、入學身份、畢業國中這此因素之間的關聯性。

各個實驗之參數設定、差異比較之流程，步驟如下：

步驟 1：選入所需之變數。

步驟 2：設定演算法參數。

步驟 3：調整最小支持度過濾冗長規則。。

步驟 4：分析入學變數對入學後成績之關係。

步驟 5：結論與建議。

3.4.2.1 關聯規則參數說明與設計

探勘前必須勾選所需變數來做決策樹分析，因此本研究以「學號」為索引鍵，利用「學籍資料庫」、「成績資料庫」分別以「入學成績級距」、「性別」、「入學方式」、「入學身份」、「分區」為輸入變數，如表 3.9，預測變數為「100下學期成績(畢業成績)」，來進行資料探勘的決策樹分析，如圖 3.14。

表 3.9 關聯規則輸入之變數表

索引鍵	資料庫	輸入變數	資料屬性	資料長度
學號	學籍資料庫、成績資料庫	入學成績	文字	8
		98 上學期成績	文字	8
		98 下學期成績	文字	8
		99 上學期成績	文字	8
		99 下學期成績	文字	8
		100 上學期成績	文字	8
		100 下學期成績	文字	8
		性別	文字	8
		科系	文字	10
		入學方式	文字	10
		入學身份	文字	10
		居住區域	文字	10

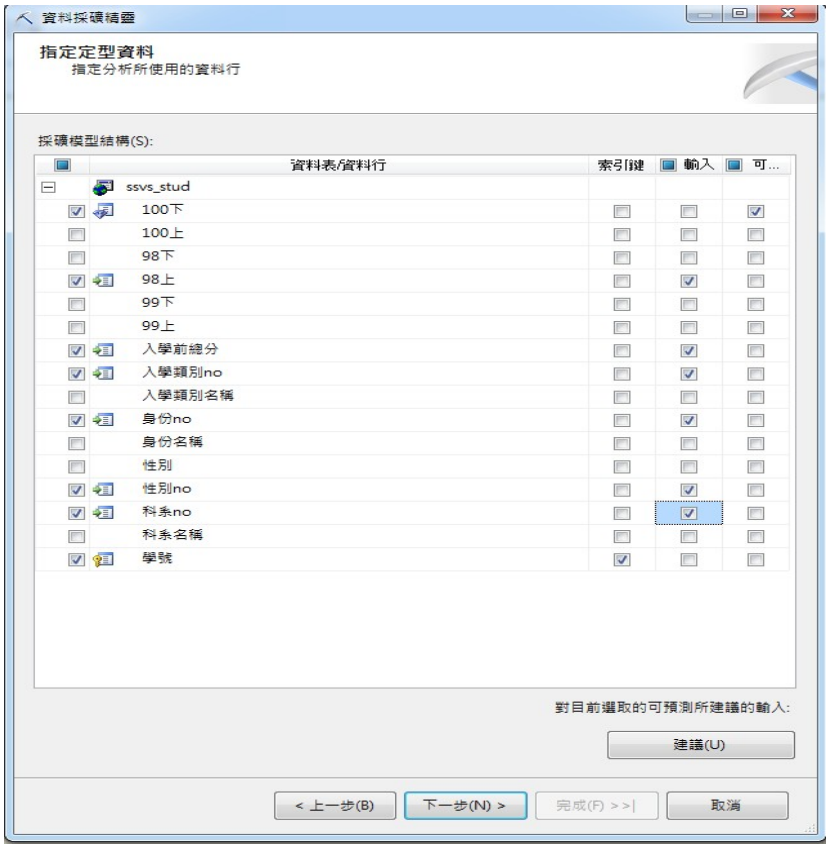

圖 3.14 關聯規則指定分析資料設定圖

3.4.2.2 關聯規則參數設定

關聯規則參數設定如表 3.10。

表 3.10 關聯規則參數設定說明圖

參數名稱	參數說明
MAXIMUM_ITEMSET_COUNT	最大物件組數量，如果未設定此參數則將產生所有可能之物件組。
MAXIMUM_ITEMSET_SIZE	一個物件組最多可包含之物件數量，如果設為0，則代表此物件組沒有物件數量之大小限制。
MINIMUM_ITEMSET_SIZE	一個物件組最少可包含之物件數量。
MAXIMUM_SUPPORT	最大支援門檻，物件組如果支援高於此參數將會被刪除。如果此參數為大於1之整數，代表絕對數量，如果是小於1之小數時，則代表佔總體交易數的百分比。
MINIMUM_SUPPORT	最小支援門檻，物件組如果支援低於此參數將會被刪除。如果此參數為大於1之整數，代表絕對數量，如果是小於1之小數時，則代表佔總體交易數的百分比。如果系統記憶體不足，演算法可能會動態調高此參數。
MINIMUM_PROBABILITY	規則最小信心水準門檻，必須填入介於1~0的小數，信心水準低於此值的規則將會被刪除。
MINIMUM_IMPORTANCE	規則最小重要性(增益值)門檻，當歸則重議性低於此值時則會被刪除。

本參數分別為支持度（support）及信度（confidence），此兩個參數可避免找到沒有意義性與有用性的規則。支持度用來限制所找到的規則必須高於一定的比例，也就是有足夠的代表性與意義性；而信度則是用來表示這兩個項目集合間交互關係的相關程度。支持度與最低信賴度都需要由資料探勘的專家來給

定，或是由管理者依其需要給定之，也就是支持度與信度必須大於最小支持度（ minimum support ）與最低信賴度（minimum confidence）的門檻值（threshold），找出的規則才具有它的意義。

但是門檻值該怎麼輸入又是一個很大的學問，門檻值設得不好，可能會導致最後探勘出來的規則數過多、過少甚至找不出規則。門檻值的設定也沒有什麼準則，就只能依照經驗法則和該資料庫的特性來做判斷。若是完全毫無頭緒，就只能以試誤法來慢慢調整門檻值，直到產生的規則數目達到預期的數量。本研究經多次試驗後，設定為最小支持度 0.01，最低信賴度為 0.05，設計方式如圖 3.15。實驗參數設定如表 3.11。在產生規則之後，規則過於冗長，因此再調高最小支持度，過濾不必要的規則，使之大約為 30 筆左右。以下為調整之說明：

表 3.11 實驗之參數設定

參數說明	設定值
最小支持度	0.01
最低信賴度	0.05
最小重要性過濾門檻	1.00
最低信賴度過濾門檻	0.20

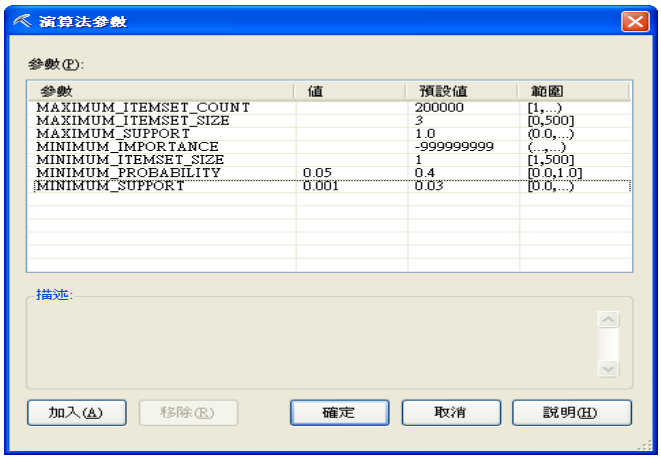

圖 3.15 關聯規則演算法參數設定介面

3.4.2.3 關聯規則設定測試資料及訓練資料

此設定目的在於使門檻值降低，增加可檢視的規則。在圖 3.16 的畫面中可以輸入測試資料的百分比(預設值為 30%，表示是用 70%建模)，或是可以輸入「在測試資料集內的最大案例數目」，系統會取其小者，作為測試資料。設定結束後，指定資料內容與資料類型如圖 3.17。圖 3.18 為探勘後結果畫面。

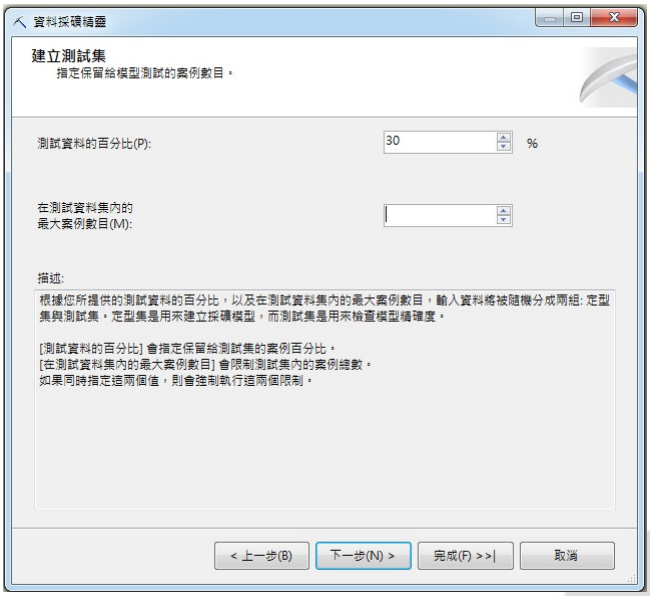

圖 3.16 關聯規則建立測試資料表

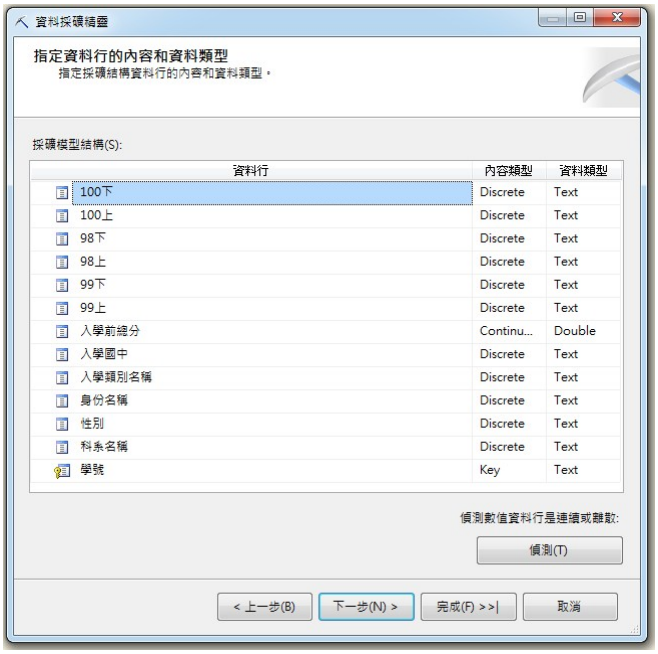

圖 3.17 關聯規則指定分析資料設定圖

機率	重要性	規則
0.673	0.513	入學前總分 >= 245.6167570944 -> 98上 = A
0.875	0.507	入學前總分 >= 245.6167570944, 性別no = 2 -> 98上 = A
0.673	0.506	入學前總分 >= 245.6167570944, 身份no = 0 -> 98上 = A
0.700	0.500	入學前總分 >= 245.6167570944, 入學類別no = 5 -> 98上 = A
1.000	0.497	科系no = 306, 入學前總分 >= 245.6167570944 -> 98上 = A
1.000	0.466	科系no = L37, 入學前總分 >= 245.6167570944 -> 98上 = A
1.000	0.466	科系no = 305, 性別no = 2 -> 98上 = A
1.000	0.466	科系no = 425, 入學前總分 >= 245.6167570944 -> 98上 = A
1.000	0.466	科系no = L37, 性別no = 2 -> 98上 = A
1.000	0.466	科系no = K21, 入學前總分 = 202.1144702464 - 245.6167570944 -> 98上 = A
0.638	0.464	入學前總分 >= 245.6167570944, 性別no = 1 -> 98上 = A
0.696	0.458	性別no = 2, 入學前總分 = 202.1144702464 - 245.6167570944 -> 98上 = A
0.800	0.449	科系no = L01, 入學前總分 = 202.1144702464 - 245.6167570944 -> 98上 = A
0.800	0.449	科系no = M03, 入學前總分 >= 245.6167570944 -> 98上 = A
0.613	0.412	入學前總分 >= 245.6167570944, 科系no = 109 -> 98上 = A
1.000	0.412	身份no = 15, 入學前總分 >= 245.6167570944 -> 98上 = A
1.000	0.412	身份no = 4, 科系no = 109 -> 98上 = A
1.000	0.412	入學類別no = 1, 性別no = 2 -> 98上 = A
1.000	0.412	入學類別no = 1, 科系no = 425 -> 98上 = A
1.000	0.412	身份no = 15, 入學類別no = 5 -> 98上 = A
1.000	0.412	身份no = 4, 入學前總分 >= 245.6167570944 -> 98上 = A
1.000	0.412	入學類別no = 1 -> 98上 = A
1.000	0.412	科系no = 308, 入學前總分 >= 245.6167570944 -> 98上 = A
1.000	0.412	入學類別no = 1, 身份no = 0 -> 98上 = A
1.000	0.412	身份no = 15, 科系no = 109 -> 98上 = A
1.000	0.412	身份no = 15 -> 98上 = A
1.000	0.412	身份no = 15, 性別no = 1 -> 98上 = A
1.000	0.412	入學類別no = 1, 入學前總分 = 202.1144702464 - 245.6167570944 -> 98上 = A
0.615	0.381	科系no = 303, 入學前總分 = 202.1144702464 - 245.6167570944 -> 98上 = A
0.507	0.375	性別no = 2, 身份no = 0 -> 98上 = A
0.625	0.375	科系no = 425, 入學前總分 = 202.1144702464 - 245.6167570944 -> 98上 = A

圖 3.18 關聯規則分析結果畫面

第四章　實驗結果說明

本章節將呈現本研究之實驗結果說明。本研究共有四項實驗，每個實驗分別進行決策樹分析與關聯規則分析。實驗一將所有入學變數代入，分別進行決策樹和關聯規則分析。唯恐決策樹過度修剪，因此本研究於實驗二至實驗四，分別探討性別、入學國中及入學方式與畢業成績之關聯關系。

實驗一：輸入入學成績級距、入學身份、性別、入學方式，預測並找出畢業成績之關聯性。

實驗二：輸入性別與畢業成績，預測並找出兩者之關聯性。

實驗三：輸入入學國中與畢業成績，預測並找出兩者之關聯性。

實驗四：輸入入學方式與畢業成績，預測並找出兩者之關聯性。

4.1 實驗一：入學成績、性別、入學國中、入學身份與畢業成績之關聯分析

一、決策樹分析：

1、「決策樹」之「樹狀檢視器」可檢視輸入變數影響的機率，以本研究為例可以得知影響學生畢業成績的首要因素為入學成績級距，因它位於結構的第二層。當入學成績級距="A"時，影響學生畢業成績機率最大，其次為性別。

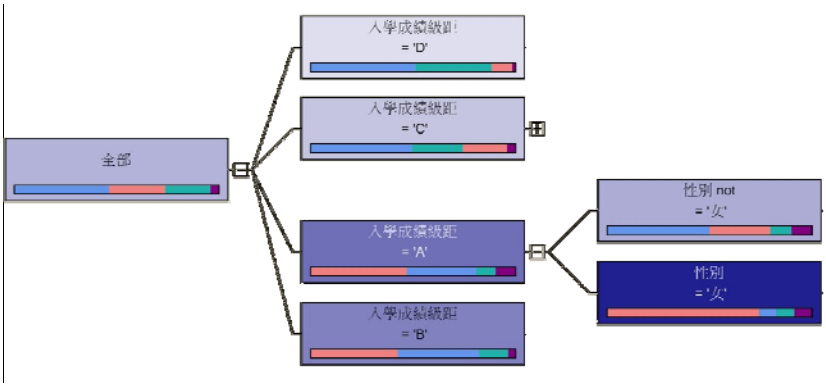

圖 4.1 入學成績級距、入學身份、性別、入學國中、入學方式與畢業成績之決策樹分析圖

2、 當入學成績級距="A"時，影響學生畢業成績機率="A"最大，且第二影響成績的因素是性別="女"。當入學成績級距="A"且性別="女"時，學生畢業成績為 A 的機率高於入學成績優異且畢業成績優異的男學生。P(入學成績=A/性別=女∩畢業成績=A)=3/4=0.75>P(入學成績=A/性別=男∩畢業成績=A)=1/4=0.25。

表 4.1 入學成績級距、性別與畢業成績之交叉分析表

		100 下畢業成績				
入學成績級距	性別	A	B	C	D	總計
A	女	3				3
	男	1	2			3
A 合計		4	2			6

B	女	18	7	2	1	28
	男	49	57	23	3	132
B 合計		67	64	25	4	160
C	女	15	21	6		42
	男	35	99	66	12	212
C 合計		50	120	72	12	254
D	女	1	4	1		6
	男	3	17	18		38
D 合計		4	21	19		44
總計		125	207	116	16	464

3、 但是，深入了解發現，入學成績級距為 A，畢業成績為 A 之人數，僅有 4 人，佔全體畢業成績人數為 A 比率過低，比率 4/125，以此驟下定論，認定入學成績與畢業成績具有極大關聯，似有失公允。以入學成績為 B，畢業成績為 A，佔 62/114；入學成績為 C，畢業成績為 A，佔 44/114；由此可知，入學成績為 B，C，雖入學成績不高，但經由三年高中教育後，畢業成績為 A，佔 106/114，92.98%，顯示教育有其成效。

4、 透過「相依性網路」檢視輸入變數與預測變數之間的關聯性強弱。「決策樹」是根據樹的層級來決定遠近，被放在越上層的變數，就是預測能力最強的變數，如入學成績級距變數與學生畢業成績變數關聯性最強，這與「樹狀檢視器」檢視結果相同，可見入學成績級距變數是影響學生畢業成績變數的主要變數。由圖中也可知道入學成績級距為 A 時，性別為女性，也會影響學生畢業成績。

由「樹狀檢視器」及「相依性網路」分析結果(如圖 4.2)，入學成績級距變數是影響學生畢業成績的主要因素，次要因素為性別，當學校要做招生行銷時，可以使用這二種因素作為行銷的策略。

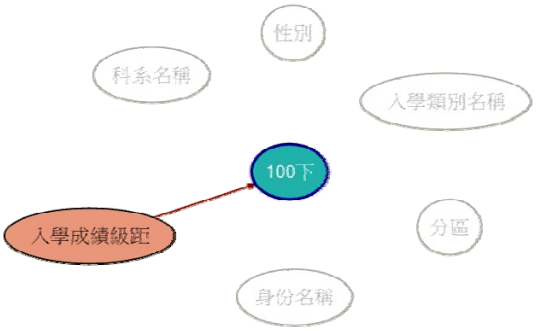

圖 4.2 入學成績級距、入學身份、性別、入學方式與畢業成績之決策樹相依性網路圖

二、關聯規則分析：

表 4.2 入學成績級距、入學身份、性別、入學方式與畢業成績之關聯規則圖

信賴度	重要性	關聯規則	
1.00	0.459	入學成績級距=A	100 下學期成績=A
1.00	0.490	入學成績級距=A，性別=女	100 下學期成績=A
1.00	0.459	身份類別為其他	100 下學期成績=C

由表 4.2 可知，在入學成績級距=A，則 100 下學期為 A 之機率愈大。在入學成績級距=A，性別=女，則 100 下學期為 A 之機率亦大。在入學身份為其他，則 98 上學期為 C 之機率愈大。

(1)、相依性網路分析

「關聯規則」之「相依性網路」是檢視成績之間之相依性，當點選之成績對外連結越多表示此節點越能夠影響其它節點，將其連結調整至最下面（即關聯性最強的連結），如入學前成績=A 變數相依性最大。表示是主要影響 100 下

學期成績=A 的主要原因。由圖 4.3 亦可驗證入學成績高，亦為影響 100 下學期成績的重要因素。

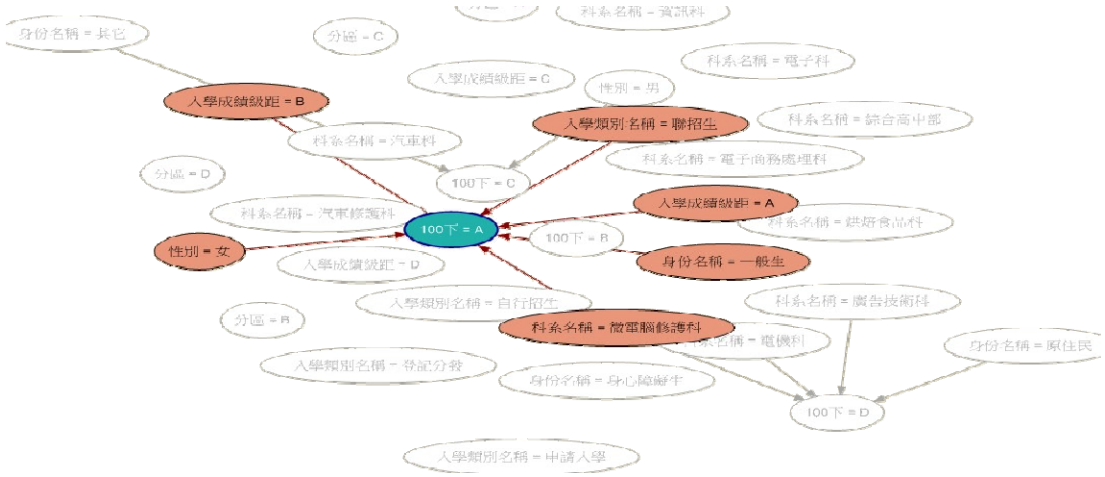

圖4.3 入學成績級距、入學身份、性別、入學國中、入學方式與畢業成績之關聯規則相依性網路圖

由實驗一可知：

入學成績對於畢業成績確有極大的影響，為求學生在校學習的表現較佳，從入學成績慎選學生是非常重要的。

女學生較男學生畢業成績優異，建議學校招生時，增加女學生入學管道。

入學身份為其他，主要是包含生活扶助戶（低收入戶），更生受保護人家庭，長期失業家庭，單薪且單親家庭；對於入學身份為其他的學生，入學後第一學期成績為 C 之機率為 100%，重要性為 0.434。

4.2 實驗二：性別與畢業成績之關聯分析

一、決策樹分析

1、 實驗二輸入性別與畢業成績，預測並找出兩者關係。圖 4.4 為性別與畢業成績之決策樹分析圖。由表 4.3 顯示女學生畢業成績優秀的比例 47.7%(P(性別=女/畢業成績=A)=37/79=46.8%)高於男學生畢業成績優秀的比例 22.86%(P(性別=女/畢業成績=A)=88/385= 22.86%，由實驗二可知，女學生畢業成績 A 比男學生畢業成績比例多，建議招生可以多吸引女生就讀。

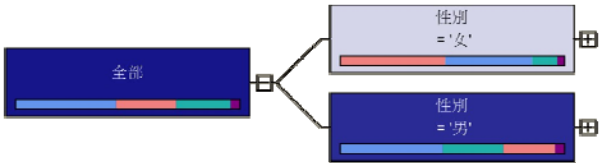

圖 4.4 性別與畢業成績之決策樹分析圖

表 4.3 性別與畢業成績之交叉分析圖

	性別		
100 下	女	男	總計
A	37	88	125
B	32	175	207
C	9	107	116
D	1	15	16
總計	79	385	464

二、關聯規則分析

表 4.4 學生性別與 100 下成績(畢業成績)之關聯規則

信賴度	重要性	關聯規則	
0.494	0.367	性別=女	100 下學期成績=A
0.460	0.04	性別=男	100 下學期成績=B

由表4.4可知，性別=女，則100下學期為A之機率愈大。性別=男，則100下學期為B之機率愈大。

(1)、相依性網路分析

性別=女變數相依性最大，表示是100下學期成績=A且性別=“女”同時出現的機率最高。

由圖4.5可知，透過「相依性網路」檢視輸入變數與預測變數之間的關聯性強弱。「決策樹」是根據樹的層級來決定遠近，被放在越上層的變數，就是預測能力最強的變數，如性別與學生畢業成績變數關聯性最強，這與「樹狀檢視器」檢視結果相同，可見性別是扣除入學成績級距後，影響學生畢業成績變數的主要變數。建議校方多吸納女學生的入學管道。

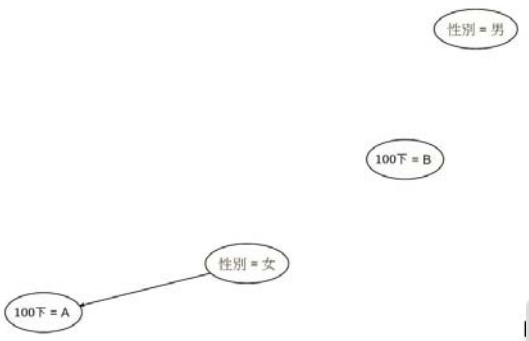

圖4.5 學生性別與畢業成績之相依性網路圖

由本實驗可知：女學生學業成績相較於男學生學業成績，較為優異，為求學生在校學習的表現較佳，招生時可增加女學生錄取名額。

4.3 實驗三：入學國中與畢業成績之關聯分析

一、決策樹分析

圖4.6 入學國中與畢業成績之決策樹分析圖

實驗三以入學國中為輸入變數，每一國中為輸入類別，當類別太多時(如表4.5)，資料量小，錯誤會增加的比較快。無法全面的看到各種變數的影響程度，因此執行結果，只顯示全部學生，如圖4.6。

表 4.5 入學國中與畢業成績之交叉分析表

入學國中	A	B	C	D	總計
屏東縣縣立X州國中		1			1
高雄市縣立X林國中		1			1
高雄市縣立X蓮國中		1			1
高雄市縣立X莊國中	2	1			3
高雄市縣立X內國中	3	4			7
高雄市縣立X竹高中(附設國中部)	2	4	2		8
雲林縣縣立X背國中	1				1
彰化縣縣立X村國中		1			1
臺南市市立X城中學附屬國中		1			1
臺南市市立X成國中	4	6			10
臺南市市立X山國中	1				1
臺南市市立X賢國中	4	4	3		11
臺南市市立X德國中	4	7	4		15
臺南市市立X平國中	2	3	2	1	8
臺南市市立X南國中	3	7	10		20
臺南市市立X順國中	3	2	2		7
臺南市市立X功國中	1				1
臺南市市立X順國中	5	2	4	1	12
臺南市市立X平國中		4			4
臺南市市立X孝國中	9	9	5	1	24
臺南市市立X城國中			1		1

臺南市市立X寧高中(附設國中部)		2	2		4
臺南市市立X興國中	1	3	2		6
臺南市市立X甲國中	4	7	5	1	17
臺南市市立X佃國中	3	5	1		9
臺南市市立X明國中	3	14	8	1	26
臺南市市立X興國中	11	21	13	4	49
臺南市市立X興國中	1	2	2		5
臺南市私立X榮高中(附設國中部)			1		1
臺南市私立X光國中		1			1
臺南市私立X山高中(附設國中部)		1			1
臺南市私立X門高中(附設國中部)				1	1
臺南市私立X海高中附屬國中	1		2		3
臺南市縣立X橋國中	3	3	2		8
臺南市縣立X灣高中(附設國中部)	13	22	11	1	47
臺南市縣立X德國中	5	13	7		25
臺南市縣立X賢國中	7	5			12
臺南市縣立X仁高中(附設國中部)	1	6	5	3	15
臺南市縣立X康國中	13	20	8		41
臺南市縣立X定國中		1			1
臺南市縣立X港國中	1		1		2
臺南市縣立X豆國中		1	1		2
臺南市縣立X化國中	4	5	2		11
臺南市縣立X仁國中	6	11	7	1	25
臺南市縣立X廟國中	4	6	3	1	14
總計	125	207	116	16	464

二、關聯規則分析

表 4.6 學生入學國中與 100 下成績(畢業成績)之關聯規則

信賴度	重要性	關聯規則	
1.00	0.394	入學國中=台南市成功國中	100 下學期成績=A
1.00	0.394	入學國中=雲林縣崙背國中	100 下學期成績=A
1.00	0.394	入學國中=台南市中山國中	100 下學期成績=A

由表 4.6 可知，入學國中為台南市成功國中，中山國中，雲林縣崙背國中，則 100 下學期成績為 A，表現較佳之機率愈大。

(1)、相依性網路分析

入學國中為台南市成功國中，中山國中，雲林縣崙背國中變數相依性最大，如圖 4.7，表現較佳。建議校方多針對這些國中(台南市成功國中，中山國中，雲林縣崙背國中)進行更多的招生宣導，以招攬學生素質較佳的學生，但由於雲林縣離本研究個案距離太遠，且就讀人數太過稀少(由表 4.5 顯示，人數僅一人)，因此不列入招生重點學校。

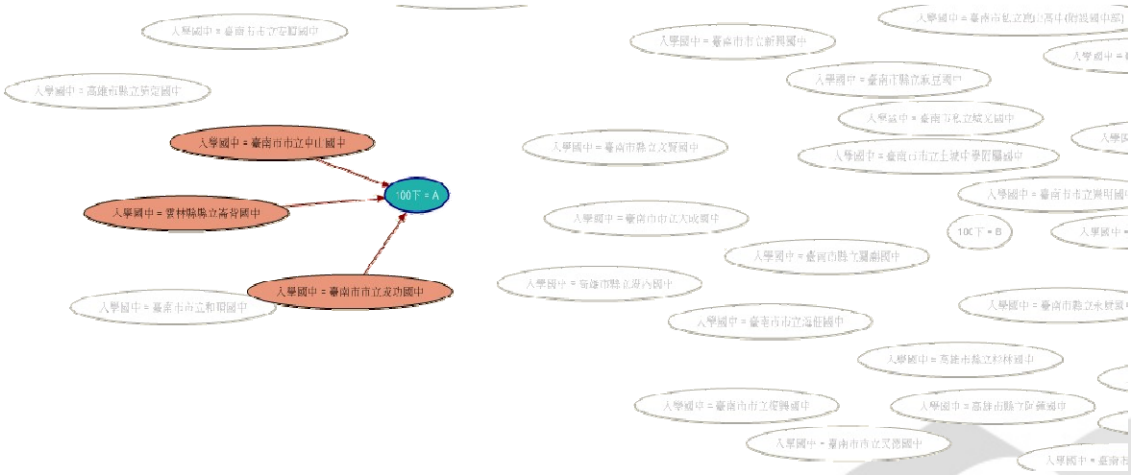

圖4.7 入學國中與畢業成績(100下學期)之相依性網路圖

以實驗三來看，決策樹由於國中類別太多時，資料量小，錯誤增加快速，經過過度修剪，無法顯示個別學校之差異，關聯規則較能完整的呈現各類別之影響，其分析結果較佳。畢業國中於某特定國中之學生，學業成績較佳，建議學校可針對此特定學校，研擬招生策略；另一方面，經項目集發現，某些國中入學人數較多，不管學生入學成績是否優異，校方應努力提升學生素質，做為將來招生時，學生返回國中原校，宣傳學校的優點。

4.4 實驗四：入學身份與畢業成績之關聯分析

一、決策樹分析

實驗四輸入入學身份與畢業成績，預測並找出兩者關係。圖 4.8 為入學身份與畢業成績之決策樹分析圖。由表 4.7 顯示一般生畢業成績優秀的比例 28.38%(P(身份=一般生/畢業成績=A)=124/437=28.38%)低於外國學生畢業成績優秀的比例 100%(P(身份=外國學生/畢業成績=A)=1/1=100%，雖然，外國學生表現較優異，但外國學生樣本僅一人，過於稀少，以此認定外國學生表現優異，恐有偏頗。

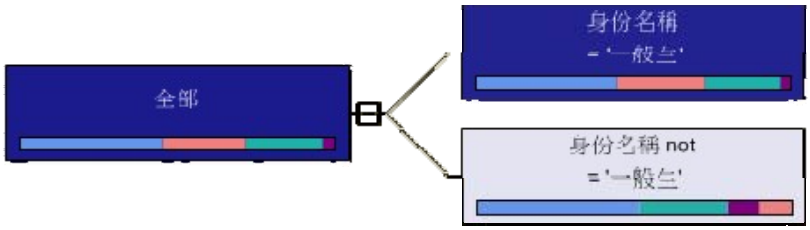

圖 4.8 入學身份與畢業成績之決策樹分析圖

表 4.7 入學身份與畢業成績之交叉分析圖

身份名稱	A	B	C	D	總計
一般生	124	190	108	15	437
外國學生	1				1
身心障礙生	1	15	6		22
其它			1		1
原住民		1	1	1	3
總計	126	206	116	16	463

二、關聯規則分析

表 4.8 學生身份與 100 下成績(畢業成績)之關聯規則

信賴度	重要性	關聯規則	
1.00	0.427	身份=其他	100 下學期成績=C
0.682	0.186	身份=身心障礙生	100 下學期成績=B

由表 4.8 可知，身份=其他，則 100 下學期為 C 之機率愈大。身份=身心障礙，則 100 下學期為 B 之機率愈大。

(1)、相依性網路分析

身份="其他"變數相依性最大，表示是 100 下學期成績=A 且身份="其他"同時出現的機率最高。由圖 4.9 可知，透過「相依性網路」檢視輸入變數與預測變數之間的關聯性強弱。「決策樹」是根據樹的層級來決定遠近，被放在越上層的變數，就是預測能力最強的變數，如學生身份與學生畢業成績變數關聯性強。建議校方應針對身份=其他，及身心障礙的學生多加注意。

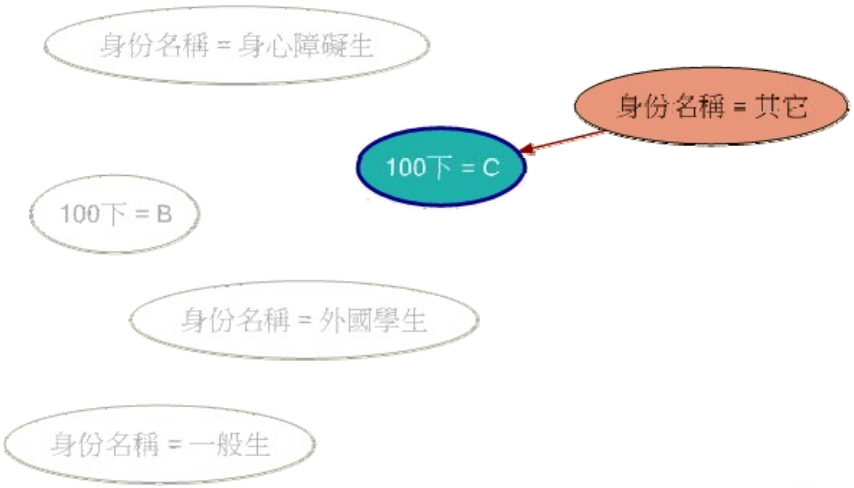

圖4.9 學生身份與100下學期(畢業成績)之相依性網路圖

由本實驗可知：身份="其他"及"身心障礙"學生，學業成績相較一般生來的低落。身份="其他"主要是包含生活扶助戶（低收入戶），更生受保護人家庭，長期失業家庭，單薪且單親家庭或其他特殊之家庭。此類學生，需要學校付予更多的關愛及注意。

第五章　結論

本研究旨在探討高職入學成績、入學方式與畢業成績間之關聯性做為研究主題，以台南市某私立高職民國 98，99，100 年三個年度的入學成績級距、入學身份、性別、畢業國中與入學方式為研究資料，在研究方法方面，使用 Data Mining 之關聯法則及決策樹規則，對 100 下學期(畢業學期)成績做關聯分析，以增進學校對學生在校成績、入學成績及入學方式之了解，做為將來招生決策的參考。同時亦可了解男女學生、不同畢業國中及不同入學方式，入學之後的學業表現。

本研究得知：

一、入學成績對於畢業成績確有極大的影響，兩者呈現正相關，為求學生在校學習的表現較佳，從入學成績慎選學生是非常重要的。

二、女學生之學業成績相較於男學生之學業成績較為優異。

三、畢業國中於某特定國中之學生，學業成績較佳。

四、對於入學身份為其他的學生，校方更應關注學生的學習態度，並給予生活及精神上的關心注意學生的品格教育，深入了解是否在學習上遇到瓶頸。

五、決策樹與關聯規則之結果大致上相同，唯決策樹由於國中類別太多時，資料量小，錯誤增加快速，經過過度修剪，無法顯示個別學校之差異，關聯規則較能完整的呈現各類別之影響，其分析結果較佳。

另外為求提高學生平均素質，亦可降低本校學生的學習能力的差異性，使學生整體程度更為均質化，而利於教師授課與各項教學提昇措施的施行，建議

學校，將入學成績較高，女學生，畢業國中為某特定國中之學生，符合以上條件之一的學生，列為重點招生對象。

後續研究建議

本研究僅針對台南市某工商職業學校學生成績研究，不宜過度推論至其他研究樣本，如對本研究想深入探討或相關類似的研究方向，建議後續研究者擴大研究對象，將樣本數擴及全台南市，甚至擴大到全國各高中，則研究將更具代表性。

參考文獻

1. Agrawal, R., Imieliński, T., and Swami, A. "Mining Association Rules between Sets of Items in Very Large Database," *Proceedings of the ACM SIGMOD Conference on Management of Data*, pp 207-216.
2. Agrawal, R., and Srikant, R. "Fast Algorithms for Mining Association Rules in Large Database," *Proceedings of the 20th International Conference on Very Large Data Bases*, pp 487-499.
3. Coenen, F., Goulbourne, G., and Leng, P. "Tree Structures for Mining Association Rules," *Data Mining and Knowledge Discovery* Vol. 8, No. 1, 2004, pp 25-51.
4. Curt, H. "The devil's in the details: techniques, tools, and application for database mining and knowledge discovery part I," *Intelligent Software Strategie* Vol. 6, No. 9, 1995, pp 1-15.
5. Fayyad, U.M., Piatetsky-Shapiro, G., and Smyth, P. *From Data Mining to Knowledge Discovery:An Overview* American Association for Artificial Intelligence Menlo Park, 1996.
6. Frawley, W.J., Piatetsky-Shapiro, G., and Matheus, C.J. "Knowledge Discovery in Databases: An Overview," *Association for the Advancement of Artificial Intelligence* Vol. 1991. pp 1-27.
7. Grupe, F.H., and Owrang, M.M. "Database Mining Discovering New Knowledge and Cooperative Advantage," *Information System Management* Vol. 12, No. 4, 1995, pp 26-30.
8. Han, J., and Fu, Y. "Discovery of Multiple-Level Association Rules from Large Databases," *VLDB '95 Proceedings of the 21th International Conference on Very Large Data Bases*, pp 420-431

9. Han, J., Kamber, M., and Pei, J. *Data Mining: Concepts and Techniques* Morgan Kaufmann, 2001.
10. Kleissner, C. " Data mining for the enterprise," *In Proceedings of the Thirty-First Hawaii International Conference*, pp 295-304.
11. Kouris, I.N., Makri, C.H., and Tsakalidis, A.K. "Using Information Retrieval techniques for supporting data mining," *Data & Knowledge Engineering* Vol. 52, No. 3, 2005, pp 353-383.
12. Linoff, G.S., and Berry, M.J.A. *Data Mining Techniques: for Marketing Sale and Customer Support* Wiley, 1997.
13. Olaru, C., and Wehenkel, L. "A complete fuzzy decision tree technique," *Fuzzy Sets and Systems* Vol. 138, 2003, pp 221-254.
14. Peacock, P.R. " Data Mining in Marketing: Part1," *Marketing Management* Vol. 16, No. 4, 1998, pp 8-18.
15. Simoudis, E. "Reality check for data mining," *IEEE Expert: Intelligent Systems and Their Application* Vol. 11, No. 5, 1996, pp 26-33.
16. Wang, Y.-F., Chuang, Y.-L., Hsu, M.-H., and Keh, H.-C. "A personalized recommender system for the cosmetic business," *Expert Systems with Applications* Vol. 26, No. 3, 2004, pp 427-343.
17. 王建華 "資料挖掘技術在技職院校中途離校生輔導之應用—以醒吾技術學院為例," in: *國防管理學院國防資訊研究所*, 2004.
18. 呂學智 "四技二專統一入學測驗成績與高職在校成績關係之研究－以彰師附工進修學校數學科目為例," in: *國立彰化師範大學工業教育與技術學院研究所碩士論文*, 2005.
19. 宋珮怡 "我國大學學科能力測驗成績之預測建模研究," in: *輔仁大學應用統計研究所碩士論文*, 2007.
20. 李佳玲 "大學入學測驗與高中生在校成績關係之研究," in: *國立台北師範學院國民教育研究所碩士論文*, 2001.
21. 李俊宏, and 古清仁 "類神經網路與資料探勘技術在醫療診斷之應用研究," *工程科技與教育學刊* Vol. 7, No. 1, 2010, pp 154-169.

22. 沈清正、陳彥良、陳仕昇、高鴻斌、張元哲、陳家仁、黃琮盛 "資料間隱含關係的挖掘與展望," *資訊管理學報* Vol. 9, 2002.
23. 林誠, and 劉福堂 "資料探勘在寬頻網路客戶目標行銷之應用研究," *電子商務學報* Vol. 7, No. 2, 2005, pp 121-138.
24. 施宏彥 "強化幼兒教育政策減緩少子化衝擊之研究," *嘉南學報（人文類）* Vol. 31, 2005, pp 476-492.
25. 洪菁憶 "循序探勘在軟體版本控制上的應用," in: *中央大學資訊管理學系學位論文*, 2008.
26. 洪嘉聲, 盧鈺欣, and 歐進士 "運用決策樹技術探討會計師選擇之關鍵決定因素," *會計與公司治理* Vol. 5, No. 2, 2008, pp 55-77.
27. 陳怡靖、陳密桃、黃毅志 "臺灣地區高中多元入學與教育機會的關聯性之實徵研究," *教育與心理研究* Vol. 29, No. 3, 2006.
28. 陳超 "國中基測與大學學測之相關分析－以旭光高中為例," in: *中華大學應用數學系碩士論文*, 2006.
29. 陸炳杉 "多元入學學生學業成就之研究－以高雄市立中正高級中學為例," in: *國立高雄師範大學工業科技教育學系碩士論文*, 2003.
30. 彭重恩 "高中學生生活科技及數理成績與大學學科能力測驗數理科成績之相關研究," in: *國立台灣師範大學工業科技教育學系碩士論文*, 2005.
31. 曾憲雄、蔡秀滿、蘇東興、曾秋蓉、王慶堯 *資料探勘 Data Mining* 旗標出版股份有限公司, 2006.
32. 黃仁鵬, and 柯柏瑄 "GSSA：以階段分組排序搜尋機制探勘關聯規則之演算法," *電子商務學報* Vol. 11, No. 3, 2009, pp 551-568.
33. 黃仁鵬, 黃南傑, 郭煌政, and 許耀文 "快速模組拆解之關聯規則探勘-QMD," *Electronic Commerce Studies* Vol. 2, No. 3, 2004, pp 333-352.
34. 黃仁鵬, and 藍國誠 "高效率探勘關聯規則之演算法—GRA," *電子商務學報* Vol. 8, No. 4, 2006, pp 469-498.
35. 黃舒郁, 謝銘智, 蕭釧瑛, and 林永青 "淺談資料倉儲," *中興工程* Vol. 109, 2010.

36. 楊乃玉, and 鄭瓊如 "藉由貝氏屬性挑選法改善簡易貝氏分類器處理連續型態資料之效能," *工程科技與教育學刊* Vol. 9, No. 2, 2012, pp 197-206.
37. 楊朝祥 "高職五專多元入學方案之規劃與實施," 立法院院聞, 2000, pp. 15-21.
38. 溫侑柯 "應用資料探勘之關聯法則探討大學入學成績對在學成績的影響－以資管系為例," in: *南華大學資訊管理研究所碩士論文*, 2006.
39. 董毅程 "以資料探勘來探討資管系學生學習成效之影響因素─以某私立大學為例," in: *長榮大學資訊管理研究所學位論文*, 2009.
40. 趙景明, and 楊慧雯 "多重資料串流環境序列樣式探勘之應用－以台灣股市為例," *資訊管理展望* Vol. 12, No. 2, 2010, pp 113-132.
41. 劉介傳 "利用資料探勘技術探討台灣壽險市場發展之研究," in: *嶺東科技大學資訊科技應用研究所學位論文*, 2013.
42. 劉玉春、王澤玲、林益三、陳清平 *高中學生在學三年成績與大學入學考試成績相關性之研究* 大學入學考試中心, 台北, 1990.
43. 賴文漢 "四技二專統一入學測驗成績與高職在校成績關係之研究-以國立二林工商為例," in: *國立彰化師範大學，工業教育與技術學系碩士論文。*, 2006.
44. 蘇中信, 劉俞志, and 劉蕙 "以顧客價值為基礎之資料庫行銷架構," *資訊管理學報* Vol. 20, No. 3, 2013, pp 341-365.