

應用資料探勘技術於 學生特質與學業表現之研究

蔡秀滿

明新科技大學資工系

pauray@must.edu.tw

林文宗

明新科技大學資工系

wtlin@must.edu.tw

劉紫瑩

明新科技大學資管系

lze48@hotmail.com

摘要

自 91 年起實施多元入學方案以來，學校如何以各系所的特色來吸引學生就讀已成為各校所面臨的一個重要課題。若能運用有限的資源來協助招生策略的制訂，將能有效掌握入學學生的來源，進而提升學校各系所的競爭力。本論文利用資料探勘技術中的關聯法則與決策樹分析的方法，發掘明新科技大學資訊管理系日間部二技與四技學生的基本資料與學業成績之間的關係。我們使用 C4.5 決策樹分類法發掘學生特質與各類成績之間的關係，這些結果可以提供學校在制訂招生策略時的參考。此外，我們使用關聯法則探勘，從學生各類相關科目成績所發掘出來的關聯法則，可以提供資管系的老師從中進一步探討可能影響學生成績表現的各項因素。

關鍵字：資料探勘、關聯法則、決策樹、學生特質、學業表現

Abstract

The multi-enrollment plan has been administered since 2002. For college, how to attract students has become an important topic. If we can use limited resources to assist making enrollment strategies, it will helpful to control the student source and promote college competition. This study used data mining techniques, such as association rule mining and decision tree analysis, to discover relationships between students' profile and academic scores.

The students we considered consist of two-year programs and four-year programs day school students in the Department of Information Management at Minghsin University of Science and Technology. We used C4.5 decision tree classification method to discover relationships between student characteristics and various course scores. The results can provide references for making enrollment strategies. Besides, we used association rule mining to discover relationships among various course scores. The results can provide teachers with information for further discussion about the factors of affecting academic achievement.

keyword: data mining、association rule、decision tree、student characteristics、academic achievement

1. 簡介

近年來，大學錄取率和四技錄取率年年升高，尤其自 91 年起實施多元入學方案以來，學校如何以各系所的特色來吸引學生就讀，已成為各校招生所面臨的一個重要課題。招生作業是否成功將會決定入學學生的素質，因此如果能夠運用有限的資源來協助招生作業的進行，有效掌握入學學生的來源，將能提升學校各系所的競爭力。

資料探勘(data mining)的目的是從一個大型資料庫所儲存的大量資料中萃取出一些有用的知識[2,9,10]。國內已有許多研究將資料

探勘的技術應用在學校招生策略的領域。文獻[5]使用 OLAP 的工具，了解學生的「居住地區」、「入學身份」與「畢業高中」之間的相關變化，並藉由決策樹的分析，找出「居住地區」、「入學身份」、「性別」、「入學年」和「畢業學校類別」等屬性與成績表現間的關係，探勘出有助於招生策略制訂的參考規則。

文獻[7]是利用群集分析、決策樹、以及序列型樣等不同的資料探勘方法，針對學生的特質、學業平均的表現、以及選修課程序列等方向進行探勘，發掘出影響學生學業表現的重要因素。而文獻[3]則是利用關聯法則、群集分析、決策樹等探勘技術，針對學生特質、學業表現、入學管道等資料進行探勘，發現在該校的學生，以指定考式入學的學生學業成績表現較好。

文獻[6]使用關聯法則的技術，從龐大的考生選填志願資料中找出各校系的關聯性，並且製作完成一個考生選填志願關聯性分析系統。而文獻[4]則是利用關聯法則的探勘技術，分析三屆學生的入學成績與在校成績之相關性，探討某一入學科目與多個在學科目成績的關係，以及多個入學科目與某一科在學科目成績的關係。

文獻[1]利用資料探勘工具找出中途離校生的行為特徵，藉以和「賴氏人格測驗」的分析結果相比較，發掘出兩者的差異性。文獻[8]則是利用資料探勘工具，針對學生的特性(性別、星座等)與課業表現之間的關係進行研究，並且檢驗星座特性與修課表現是否一致。

Weka[11,19](Waikato Environment for Knowledge Analysis)是一套由紐西蘭 Waikato 大學所發展出來的資料探勘軟體。Weka 包括資料前置處理(data pre-processing)、分類(classification)、迴歸分析(regression)、群集分析(clustering)、關聯法則(association rules)、視覺化(visualization)等工具。由於 Weka 是共享軟體，已廣泛為許多學者所使用，為現今學術

研究常採用的資料探勘工具之一。

本研究使用 Weka 資料探勘軟體，針對本校資管系 91 年度至 94 年度的學生學籍資料和學業成績資料進行探勘。我們利用 Apriori 關聯法則探勘的方法，發掘出各類相關科目之間學生成績表現的關聯性。這些結果可以提供給資管系的老師參考，從中進一步探討可能影響學生成績表現的各項因素。此外，我們使用 C4.5 決策樹分類法，分析「性別」、「居住地區」、「入學管道」和「畢業學校類別」與學生在共同科目、管理類科目以及技術類科目的表現之間的關係。這些結果可以提供學校在進行招生時的參考，制訂出更有效的招生策略。

2. 研究方法

資料探勘[18]是知識發現 (Knowledge Discovery in Database; KDD) 過程中的一個重要步驟。KDD的過程包括「資料收集」、「資料前置處理」、「資料倉儲建立」、「資料探勘」、「樣式評估」等階段[20]。在這一節中，我們將分別說明本研究在進行資料探勘之前所做的「資料收集」、「資料前置處理」、與「資料倉儲建立」等工作，以及介紹所使用的兩種探勘技術：「關聯法則探勘」和「決策樹」。

表 1 原始資料綱要

學生學籍資料表	學業成績資料
編號	編號
性別	入學年
學制	成績學年
入學年	成績學期
年級	已修學分
郵遞區號	通過學分
住址	學期成績平均
畢業學校	學科成績
入學管道	課名

2.1 資料收集

本研究只針對已畢業的學生資料進行分析。我們以明新科技大學 91 至 94 學年，資訊管理系日間部二技與四技的入學學生學籍資料和學業成績作為資料探勘的分析資料。原始資料的綱要(schema)如表 1 所示。

2.2 資料前置處理

我們將成績依科目分為「共同科目」、「基礎資訊技術類」和「基礎管理知識類」等三大方向來進行分析。共同科目包括四技一年級的共同必修課「英文」和「商用微積分」與二技的「英語語練」。然而，因為二技只有一門共同必修，科目太少不利於分析，故共同科目只考慮四技的資料。在專業科目方面，明新資管系四技成立較晚，已畢業的班級數較少，故專業科目的管理類和技術類的分析，目前先以二技的資料來進行。

資料前置處理主要包含下列幾項工作：資料整合、資料清理、以及資料轉換等工作。

● 資料整合(data integration)

處理的步驟如下：

- (1) 將來自註冊組的學生個人基本資料和課務組的學業成績資料統一整合到 Access 檔案內。
- (2) 將不必要的欄位刪除。被刪除的欄位包括：入學年、年級、郵遞區號、成績學年、成績學期、已修學分、通過學分。
- (3) 根據全學程開課時序手冊，將每一位學生的必修課名及學科成績逐一展開，而選修科目則先予以刪除。因為將每一屆資料的 schema 展開後，入學年度對分析沒有幫助，故將「入學年」這個時間維度刪除。
- (4) 解決綱要不一致(schema conflict)的問題。來自不同資料來源的 schema，其科目名稱可能不相同，但實際代表的意義卻是相同的。例如畢業學校名稱的問題，

私立光武技術學院的原校名是私立光武工商專科學校，之後又更名為私立北台科學技術學院，故需要將該校畢業生的畢業學校欄位改成相同的名稱，以解決綱要不一致的問題。

● 資料清理(data cleaning)

處理的步驟如下：

- (1) Null 值的處理。例如某學生某科必修科目的成績是空值，為了避免影響後續探勘結果的正確性，故將該學生的資料全數刪除。
- (2) 未完成學業學生的資料處理。例如日間部四技某學生在 91 年入學，91 年有各科成績資料，但到了 92 年卻只有幾筆成績資料，93 年後則完全無任何資料，因此判定該學生可能是休學或退學。為避免影響後續探勘結果的正確性，故我們將該學生的資料刪除，不予考慮。
- (3) 重修成績的資料處理。這是指必修科目沒通過最低標準 60 分而被當的學生，在其它學年再重修這門必修課。為了在後續的探勘中可以找出一些重要的規則，所以這類學生的科目成績，我們只取不及格的成績，而重修成績則予以刪除。

● 資料轉換(data transformation)

處理的步驟如下：

- (1) 內容轉換：根據原本的資料內容，轉換成新的欄位內容。例如從「學校名稱」，轉換成「畢業學校類別」欄位的值為「國立」或「私立」。
- (2) 課程分類：取資管系二技專業必修科目與四技共同必修科目來作分析：
 - (a) 專業必修分為兩類
 - i. 基礎資訊技術類：包括資料庫、軟體工程、資料結構、離散數學和網際網路技術等科目，這些科目的平均成績稱為「基礎資訊技術類平均」成績。

- ii. 基礎管理知識類：包括管理資訊系統和統計學等科目，這些科目的平均成績稱為「基礎管理知識類平均」成績。
- (b) 共同必修科目包括英文和商用微積分，這些科目的平均成績稱為「共同科目平均」成績。
- (3) 資料一般化：若將成績直接進行比較，可能因不同授課老師的評分標準不同而產生不正確的比較結果，所以我們對各科成績進行正規化。正規化的目的是將不同班級學生的成績以統一的標準值來呈現，讓各班的學生能夠在相同的基準下作比較。

表 2 範例成績表

	A 生	B 生	C 生
該生離散數學成績(分)	53	75	66
該科學期平均(分)	62	62	62
該科學期標準差	2	2	2

例如，91 學年上學期某班 3 位學生的離散數學成績如表 2 所示。成績正規化轉換公式為：

$$\text{成績正規化 (Z 分數)} = \frac{\text{該學生學科成績} - \text{該科學期平均}}{\text{該科學期標準差}}$$

根據上述的公式，3 位學生的成績 Z 分數分別為：A 生為 -4.5，B 生為 6.5，C 生為 2。由於各類平均成績的 Z 分數範圍差異不太，根據觀察可以以 ± 1.5 為界限來切分合適的 Z 分數範圍。本研究將 Z 值 ≥ 1.5 定義為「高」，Z 值介於 0 到 1.5 之間定義為「中上」，Z 值介於 -1.5 到 -0.0001 之間定義為「中下」，Z 值小於 -1.5 定義為「低」，如圖 1 所示。在單科成績方面，計算出來

的成績 Z 分數範圍差異大，所以不適合採用相同的切割方式，以免造成分析結果的誤差。我們以各單科的最大值和最小值的平均數當作中間的切割點 q 。將大於 q 的部份，以 q 和單科的最大值之平均數再切分為兩部份，而小於 q 的部份，以 q 和單科的最小值之平均數再切分為兩部份，如圖 2 所示。本研究將 Z 值大於 r 定義為「高」，Z 值介於 q 到 r 之間定義為「中上」，Z 值介於 p 到 q 之間定義為「中下」，Z 值小於 p 定義為「低」。

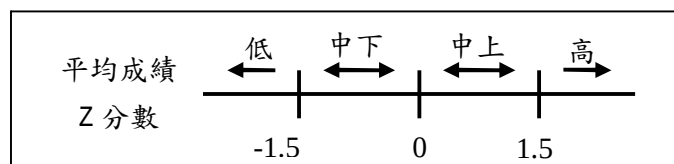


圖 1 各類平均成績的 Z 分數範圍

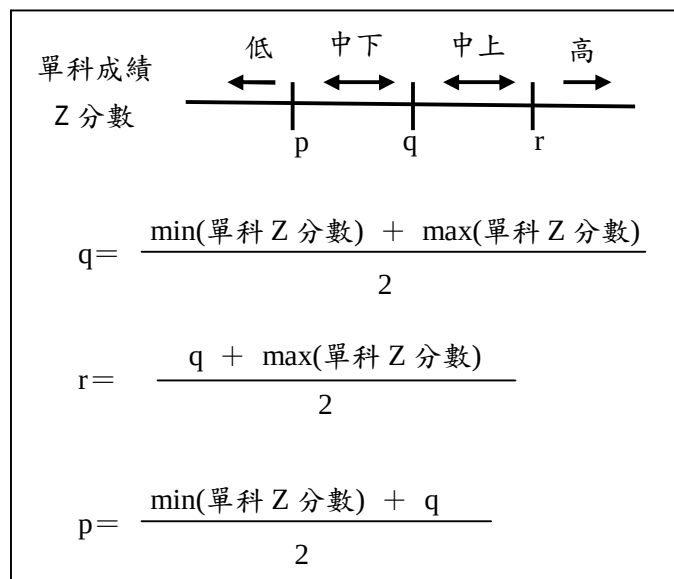


圖 2 單科成績 Z 分數範圍

2.3 資料倉儲的建置

本研究建置資料倉儲時的資料表包含了事實資料表和維度資料表。

- 事實資料表：包含各單科成績與各分類平均成績、所對應的 Z 分數、以及 Z 分數

等級等欄位。

- 維度資料表：包括「居住地區」、「入學管道」、「性別」、「畢業學校」等四個表格。其中，「居住地區」的維度分為「分佈」(包含北部、中部、南部、東部)和「區域」(包含台北縣市、基隆縣市、桃園縣市、新竹縣市等)二個層級。

我們採用的架構是星狀架構，如圖 3 和圖 4 所示。

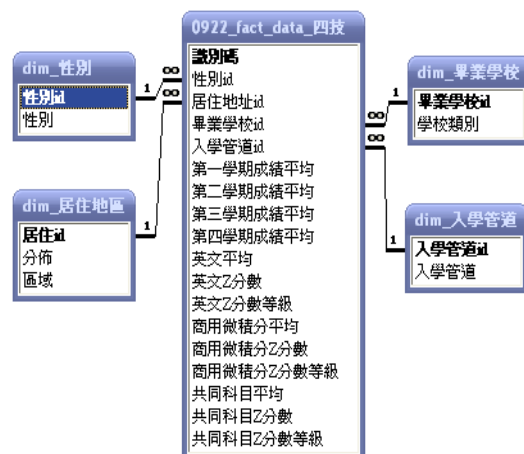


圖 4 四技資料星狀架構圖

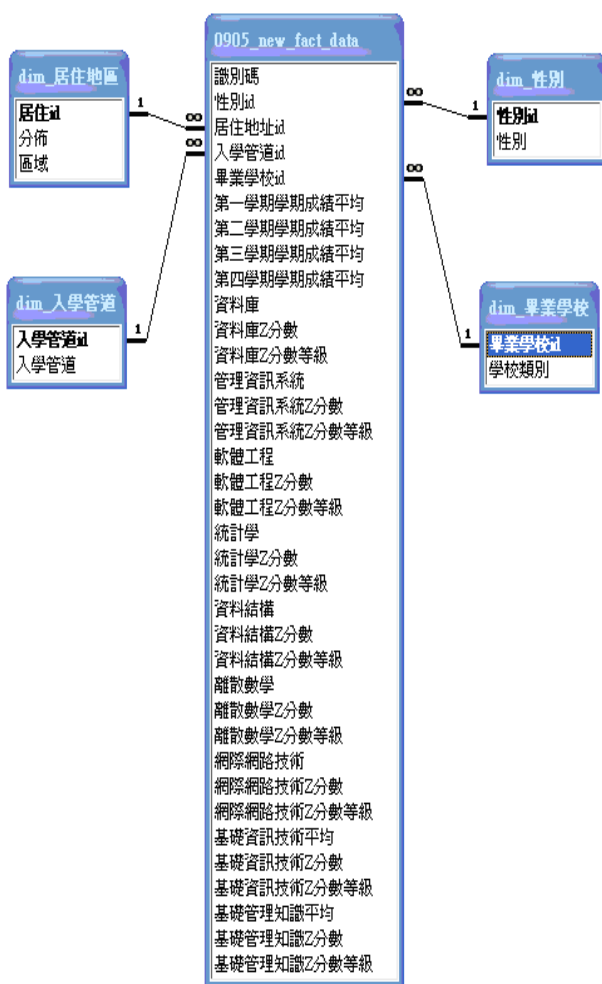


圖 3 二技資料星狀架構圖

2.4 關聯法則探勘

關聯法則探勘可以找出項目之間有趣的關聯，最著名的方法為 Apriori[14,15]演算法。本研究使用 Weka 所提供的工具，以 Apriori 的方法來分析各科目間成績的相關性。例如資料結構成績表現好的人，其離散數學成績表現又是如何？或是這兩科成績之間沒有太大的相關性？這些都是我們研究所要討論的部分。

2.5 決策樹

分類是指根據已知的資料及其類別屬性來建立資料的分類模型。利用分類模型的建立可以了解屬於各種類別屬性的資料具備哪些特徵[2]。最常用來描述分類的模型為決策樹 (decision tree)[13]，建立決策樹常見的方法包括 ID3 和 C4.5，兩者都是由同一位學者 Quinlan 所提出的。ID3 演算法[13]以資訊獲利 (information gain) 做為基準，選擇最佳的屬性當作樹的節點，造成決策樹的結果傾向於選擇屬性值較多的屬性。針對此項缺點 Quinlan 提出 C4.5 演算法[12]來做修正，以減少產生過多子樹所造成的影響。

本研究使用 Weka 工具中 C4.5 決策樹的方法來進行分類，分析「性別」、「居住分佈」、「畢業學校類別」、「入學管道」與科目成績之間的關係。另外，我們採用交叉驗證實驗法

(k-fold cross validation)來評估分類模型的準確度[16,17]。這個方法是將資料樣本分成 k 個子樣本群，然後將其中一個子樣本當作測試資料，其他 k-1 個子樣本當作訓練資料，以此方式重複 k 次的訓練與測試後，找出準確度最高的分類模型作為分類的依據[2]。

3. 研究內容

在這一節中，我們將分別說明本研究在進行資料探勘之前所做的一些統計分析，以及使用 Weka 的「關聯法則」與「決策樹」工具進行資料探勘所得到的結果。

3.1 統計分析

我們對二技與四技的基本資料與各科必修成績做統計分析，以了解其分佈的趨勢。

● 二技資料的統計分析

(1) 基本資料的統計分析

明新科技大學 91 年至 93 年資管系二技學生總人數為 405 人，男生人數略多，佔總人數的 55%(如表 3)。居住地區以北部地區居多，佔總數的 71%(如表 4)。畢業學校類別以私立學校居多，佔總數的 89%(如表 5)。入學管道以聯合登記分發的人數最多，佔總數的 57%(如表 6)。

表 3 性別統計表

	人數	百分比
男	222	55%
女	183	45%
Total	405	100%

表 4 居住地區統計表

	人數	百分比
北部	286	71%
中部	52	13%
南部	35	9%
東部	32	8%
Total	405	100%

表 5 畢業學校類別統計表

	人數	百分比
國立	44	11%
私立	361	89%
Total	405	100%

表 6 入學管道統計表

	人數	百分比
聯合登記分發	230	57%
技優甄保	20	5%
推薦甄選	155	38%
Total	405	100%

(2) 各科成績的統計分析

明新科技大學資管系二技學生資料庫成績分佈(如表 7)、管理資訊系統成績分佈(如表 8)、軟體工程成績分佈(如表 9)、資料結構成績分佈(如表 10)、統計學成績分佈(如表 11)、以及網際網路技術成績分佈(如表 12)，均以「中上」居多。而離散數學成績分佈(如表 13)以「中下」居多，佔總數的 42%，顯示學生在數學方面的學習能力較弱。

表 7 資料庫統計表

	人數	百分比
高	71	18%
中上	188	46%
中下	141	35%
低	5	1%
Total	405	100%

表 8 管理資訊系統統計表

	人數	百分比
高	34	8%
中上	172	42%
中下	155	38%
低	44	11%
Total	405	100%

表 9 軟體工程統計表

	人數	百分比
高	61	15%
中上	175	43%
中下	145	36%
低	24	6%
Total	405	100%

表 10 資料結構統計表

	人數	百分比
高	133	33%
中上	194	48%
中下	73	18%
低	5	1%
Total	405	100%

表 11 統計學統計表

	人數	百分比
高	53	13%
中上	170	42%
中下	160	40%
低	22	5%
Total	405	100%

表 12 網際網路技術統計表

	人數	百分比
高	114	28%
中上	147	36%
中下	137	34%
低	7	2%
Total	405	100%

表 13 離散數學統計表

	人數	百分比
高	4	1%
中上	72	18%
中下	170	42%
低	159	39%
Total	405	100%

● 四技資料的統計分析

(1) 基本資料的統計分析

明新科技大學 91 年至 93 年資管系四技學生總人數為 308 人，以女生人數略多，佔總人數的 58%(如表 14)。居住地區以北部地區居多，佔總數的 68%(如表 15)。畢業學校類別以國立學校居多，佔總數的 66%(如表 16)。入學管道以聯合登記分發比例最高，佔總數的 64%(如表 17)。

表 14 性別統計表

	人數	百分比
男	128	42%
女	180	58%
Total	308	100%

表 15 居住地區統計表

	人數	百分比
北部	209	68%
中部	48	16%
南部	16	5%
東部	35	11%
Total	308	100%

表 16 畢業學校類別統計表

	人數	百分比
國立	202	66%
私立	106	34%
Total	308	100%

表 17 入學管道統計表

	人數	百分比
聯合登記分發	197	64%
技優甄保	5	2%
申請入學	21	7%
推薦甄選	85	28%
Total	308	100%

(2) 各科成績的統計分析

明新科技大學四技資管系學生英文成績分佈以「高」居多，佔總數的 69%(如表 18)；商用微積分成績分佈以「中上」居多，佔總數的 51%(如表 19)。

表 18 英文統計表

	人數	百分比
高	213	69%
中上	64	21%
中下	23	7%
低	8	3%
Total	308	100%

表 19 商用微積分統計表

	人數	百分比
高	83	27%
中上	158	51%
中下	65	21%
低	2	1%
Total	308	100%

3.2 使用 Apriori 關聯法則探勘技術分析結果

針對四技共同科目及二技專業必修科目分析各類科目成績的相關性。

● 共同科目各科成績的相關性

使用 Weka 工具輸入變數：英文 Z 分數等級、商用微積分 Z 分數等級，最小支持度 (minsup) 為 0.05，信心水準(confidence) 為 0.6。結果如圖 5 所示。

Best rules found:	
1. 英文Z分數等級=高 ==> 商用微積分Z分數等級=中上	conf:(0.7)
2. 商用微積分Z分數等級=中下 ==> 英文Z分數等級=中上	conf:(0.66)
3. 商用微積分Z分數等級=高 ==> 英文Z分數等級=高	conf:(0.65)
4. 英文Z分數等級=中上 ==> 商用微積分Z分數等級=中上	conf:(0.64)

圖 5 四技共同科目成績的關聯法則

● 基礎資訊技術類各科成績的相關性

使用 Weka 工具輸入變數：資料庫 Z 分數等級、軟體工程 Z 分數等級、資料結構 Z 分

數等級、離散數學 Z 分數等級、網際網路技術 Z 分數等級，最小支持度為 0.05，信心水準為 0.6。結果如圖 6 所示。

Best rules found:	
1. 軟體工程Z分數等級=中上 資料庫Z分數等級=中上 網際網路技術Z分數等級=高 ==> 離散數學Z分數等級=中下	conf:(0.81)
2. 資料庫Z分數等級=中下 離散數學Z分數等級=中下 網際網路技術Z分數等級=中上 ==> 資料結構Z分數等級=中上	conf:(0.77)
3. 軟體工程Z分數等級=中下 資料結構Z分數等級=中上 網際網路技術Z分數等級=中下 ==> 離散數學Z分數等級=中下	conf:(0.77)
4. 資料庫Z分數等級=中上 離散數學Z分數等級=中下 網際網路技術Z分數等級=中下 ==> 資料結構Z分數等級=中上	conf:(0.76)
5. 資料庫Z分數等級=中上 網際網路技術Z分數等級=中下 ==> 資料結構Z分數等級=中上	conf:(0.75)
6. 資料庫Z分數等級=中下 網際網路技術Z分數等級=中上 ==> 資料結構Z分數等級=中上	conf:(0.75)

圖 6 二技基礎資訊技術類各科成績的關聯法則

● 基礎管理知識類各科成績的相關性

使用 Weka 工具輸入變數：統計學 Z 分數等級、管理資訊系統 Z 分數等級，最小支持度為 0.05，信心水準為 0.1。結果如圖 7 所示，我們發現統計學與管理資訊系統的規則之 confidence 偏低，代表這兩個科目之間沒有太大的關聯性。

Best rules found:	
1. 管理資訊系統Z分數等級=低 ==> 統計學Z分數等級=中下	conf:(0.59)
2. 統計學Z分數等級=高 ==> 管理資訊系統Z分數等級=中上	conf:(0.49)
3. 統計學Z分數等級=中上 ==> 管理資訊系統Z分數等級=中上	conf:(0.46)
4. 管理資訊系統Z分數等級=中上 ==> 統計學Z分數等級=中上	conf:(0.45)

圖 7 二技基礎管理知識類各科成績的關聯法則

3.3 使用 C4.5 決策樹分類探勘技術分析結果

我們使用決策數分析性別、居住分佈、畢業學校類別、入學管道與共同科目成績、基礎資訊技術類成績、基礎管理知識類成績之間的相關性。

● 性別、居住分佈、畢業學校類別、入學管道與商用微積分成績之間的相關性

使用 Weka 工具輸入變數：性別、分佈、畢業學校類別、入學管道。輸出變數：商用微積分 Z 分數等級。結果如圖 8 所示，正確率約 62%：

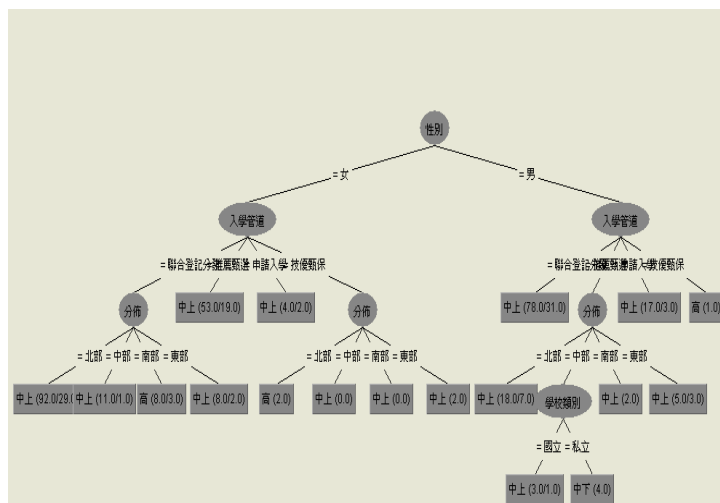


圖 8 四技商用微積分 Z 分數等級之決策樹

根據決策樹模型，我們可以得到下列 18 條可參考的規則：

規則一：如果性別為女，入學管道為聯合登記分發，且分佈為北部，則商用微積分成績表現為「中上」。

規則二：如果性別為女，入學管道為聯合登記分發，且分佈為中部，則商用微積分成績表現為「中上」。

規則三：如果性別為女，入學管道為聯合登記分發，且分佈為南部，則商用微積分成績表現為「高」。

規則四：如果性別為女，入學管道為聯合登記分發，且分佈為東部，則商用微積分成績表現為「中上」。

規則五：如果性別為女，且入學管道為推薦甄選，則商用微積分成績表現為「中上」。

規則六：如果性別為女，且入學管道為申請入學，則商用微積分成績表現為「中上」。

規則七：如果性別為女，入學管道為技優甄保，且分佈為北部，則商用微積分成績表現為「高」。

規則八：如果性別為女，入學管道為技優甄保，且分佈為中部，則商用微積分成績表現為「中上」。

績表現為「中上」。

規則九：如果性別為女，入學管道為技優甄保，且分佈為南部，則商用微積分成績表現為「中上」。

規則十：如果性別為女，入學管道為技優甄保，且分佈為東部，則商用微積分成績表現為「中上」。

規則十一：如果性別為男，且入學管道為聯合登記分發，則商用微積分成績表現為「中上」。

規則十二：如果性別為男，入學管道為推薦甄選，且分佈為北部，則商用微積分成績表現為「中上」。

規則十三：如果性別為男，入學管道為推薦甄選，分佈為中部，且畢業學校類別為國立，則四技商用微積分成績表現為「中上」。

規則十四：如果性別為男，入學管道為推薦甄選，分佈為中部，且畢業學校類別為私立，則商用微積分成績表現為「中下」。

規則十五：如果性別為男，入學管道為推薦甄選，且分佈為南部，則商用微積分成績表現為「中上」。

規則十六：如果性別為男，入學管道為推薦甄選，且分佈為東部，則商用微積分成績表現為「中上」。

規則十七：如果性別為男，且入學管道為申請入學，則商用微積分成績表現為「中上」。

規則十八：如果性別為男，且入學管道為技優甄保，則商用微積分成績表現為「高」。

● 性別、居住分佈、畢業學校類別、入學管道與英文成績之間的相關性

使用 Weka 工具輸入變數：性別、分佈、畢業學校類別、入學管道。輸出變數：英文 Z

分數等級。結果如圖 9 所示，正確率約 59%。

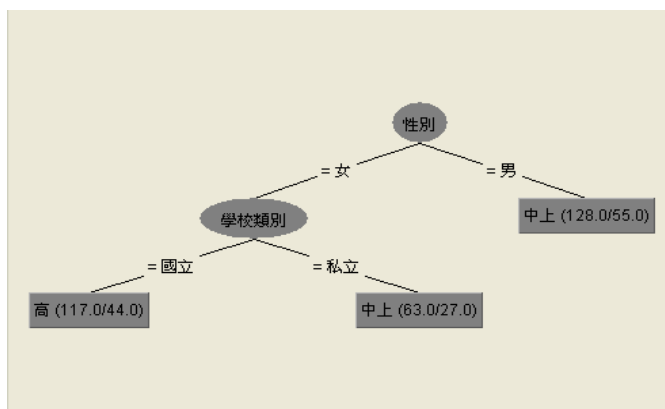


圖 9 四技英文 Z 分數等級之決策樹

根據決策樹模型，我們可以得到下列 3 條可參考的規則：

規則一：如果性別為女，且畢業學校類別為國立，則英文成績表現為「高」。

規則二：如果性別為女，且畢業學校類別為私立，則英文成績表現為「中上」。

規則三：如果性別為男，則英文成績表現為「中上」。

● 性別、居住分佈、畢業學校類別、入學管道與基礎資訊技術類成績之間的相關性

使用 Weka 工具輸入變數：性別、分佈、畢業學校類別、入學管道。輸出變數：基礎資訊技術 Z 分數等級。結果如圖 10 所示，正確率約 50%。

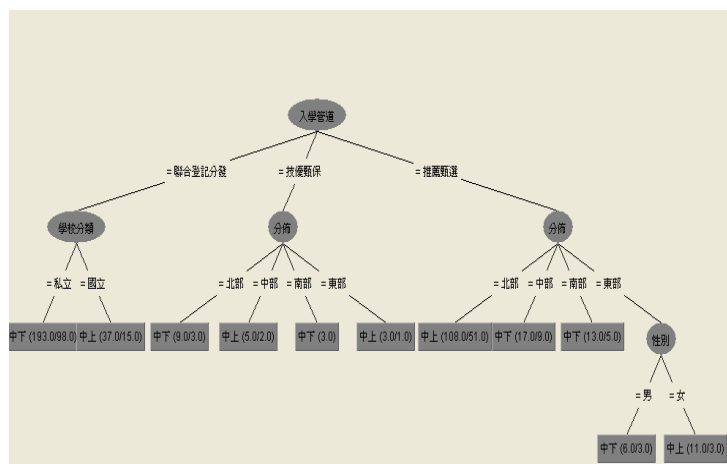


圖 10 二技基礎資訊技術 Z 分數等級之決策樹

根據決策樹模型，我們可以得到下列 11 條可參考的規則：

規則一：如果入學管道為聯合登記分發，且畢業學校類別為私立，則基礎資訊技術成績表現為「中下」。

規則二：如果入學管道為聯合登記分發，且畢業學校類別為國立，則基礎資訊技術成績表現為「中上」。

規則三：如果入學管道為技優甄保，且分佈為北部，則基礎資訊技術成績表現為「中下」。

規則四：如果入學管道為技優甄保，且分佈為中部，則基礎資訊技術成績表現為「中上」。

規則五：如果入學管道為技優甄保，且分佈為南部，則基礎資訊技術成績表現為「中下」。

規則六：如果入學管道為技優甄保，且分佈為東部，則基礎資訊技術成績表現為「中上」。

規則七：如果入學管道為推薦甄選，且分佈為北部，則基礎資訊技術成績表現為「中上」。

規則八：如果入學管道為推薦甄選，且分佈為中部，則基礎資訊技術成績表現為「中下」。

規則九：如果入學管道為推薦甄選，且分佈為南部，則基礎資訊技術成績表現為「中下」。

規則十：如果入學管道為推薦甄選，分佈為東部，且性別為男，則基礎資訊技術成績表現為「中下」。

規則十一：如果入學管道為推薦甄選，分佈為東部，且性別為女，則基礎資訊技術成績表現為「中上」。

● 性別、居住分佈、畢業學校類別、入學管道與基礎管理知識類成績之間的相關性

使用 Weka 工具輸入變數：性別、分佈、畢業學校類別、入學管道。輸出變數：基礎管理知識 Z 分數等級。結果如圖 11 所示，正確率約 49%。

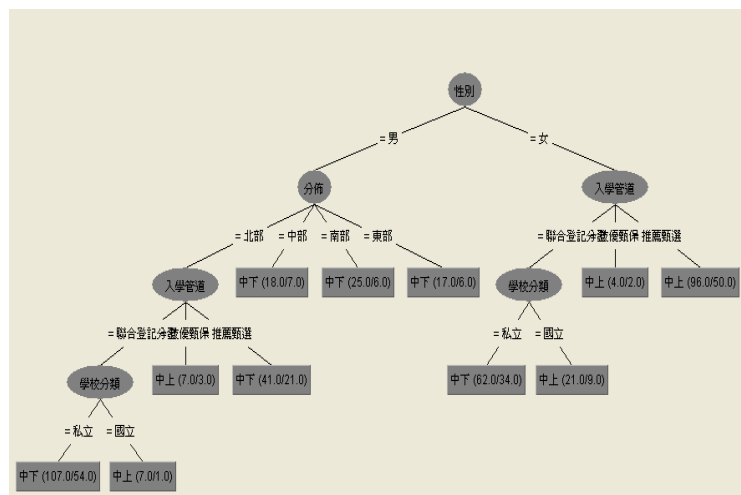


圖 11 二技基礎管理知識 Z 分數成績之決策樹

根據決策樹模型，我們可以得到下列 11 條可參考的規則：

規則一：如果性別為男，分佈為北部，入學管道為聯合登記分發，且畢業學校類別為私立，則基礎管理知識成績表現為「中下」。

規則二：如果性別為男，分佈為北部，入學管道為聯合登記分發，且畢業學校類別為國立，則基礎管理知識成績表現為「中上」。

規則三：如果性別為男，分佈為北部，且入學管道為技優甄保，則基礎管理知識成績表現為「中上」。

規則四：如果性別為男，分佈為北部，且入學管道為推薦甄選，則基礎管理知識成績表現為「中下」。

規則五：如果性別為男，且分佈為中部，則基礎管理知識成績表現為「中下」。

規則六：如果性別為男，且分佈為南部，則基礎管理知識成績表現為「中下」。

規則七：如果性別為男，且分佈為東部，則基

礎管理知識成績表現為「中下」。

規則八：如果性別為女，入學管道為聯合登記分發，且畢業學校類別為私立，則基礎管理知識成績表現為「中下」。

規則九：如果性別為女，入學管道為聯合登記分發，且畢業學校類別為國立，則基礎管理知識成績表現為「中上」。

規則十：如果性別為女，且入學管道為技優甄保，則基礎管理知識成績表現為「中上」。

規則十一：如果性別為女，且入學管道為推薦甄選，則基礎管理知識成績表現為「中上」。

4. 結論

本研究根據明新科技大學 91 年至 93 年資管系二技與四技學生的學籍資料和學業成績資料，對學生特質與學業表現進行探勘。我們利用 Apriori 關聯法則探勘方法，發掘出各類相關科目之間學生成績表現的關聯性。這些結果可以提供給資管系的老師參考，從中進一步探討可能影響學生成績表現的各項因素。此外，我們使用 C4.5 決策樹分類法，發掘出學生特質與各類成績之間的關係。這些結果可以提供學校在進行招生時的參考，制訂出更有效的招生策略。接續本研究，我們將再進行下列擴充：

- 待四技畢業學生人數達一定數目後，新增四技專業必修科目的成績，以進行相關的資料探勘。
- 考慮使用普遍等級的成績分配來作分析。所謂的普遍等級就是以一般大家熟知的分數等級來劃分，例如將成績分為 A(100~80 分)、B(79~70 分)、C(69~60 分)、D(59 分以下)等四個等級。比較使用普遍等級的成績分配與 Z 分數的成績分配之探勘差異，將是我們下一步研究的重點。

參考文獻

- [1]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」，**國防大學國防管理學院國防資訊管理研究所碩士學位論文**，93 年。
- [2]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，95 年。
- [3]彭莉櫻，「運用料探勘技術於學習成效與招生策略之研究」，**玄奘大學資訊科學學系碩士論文**，96 年。
- [4]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」，**南華大學資訊管理研究所碩士班碩士論文**，94 年。
- [5]楊琇媛，「利用資料倉儲與資料探勘技術於招生策略與學生特質分析之研究」，**中原大學資訊管理學系碩士學位論文**，92 年。
- [6]楊哲智，「以聯招考選填志願挖掘系所關聯性之研究」，**元智大學工業工程與管理研究所碩士論文**，93 年。
- [7]顏博文，「應用資料探勘技術分析學生選課特性與學業表現」，**中原大學資訊管理學系碩士學位論文**，92 年。
- [8]羅才輝，「應用資訊探勘技術分析人格特質與選課之關係－以新竹地區大學生為例」，**中華大學資訊工程學系碩士論文**，95 年。
- [9]Frawley, W. J. et al. (Eds) , Knowledge discovery in databases : an overview, **Menlo Park** , CA:AAAI Press/The MIT Press, 1991.
- [10]Ian H. Witten, Eibe Frank, Data Mining Practical Machine Learning Tools and Techniques (Second Edition), **Amsterdam ; Boston, MA : Morgan Kaufman**, 2005.
- [11]Berry, M.J.A. and Linoff, G., Data Mining Technique for Marketing, Sale, and Customer Support, **Wiley Computer**, 1997.
- [12]Quinlan, J. R., C4.5: Programs for Machine Learning, **Morgan-Kaufmann**, San Francisco. 1993.
- [13]Quinlan, J.R., 1986, Induction of Decision Trees, **Machine Learning**, Vol.1, No. 1, pp. 81-106.
- [14]R. Agrawal, T. Imielinski, and A. Swami, ;Mining Association Rules between Sets of Items in Large Databases,; **Proc. ACM SIGMOD**, pp. 207-216, May 1993.
- [15]R. Agrawal and R. Srikant, ;Fast Algorithms for Mining Association Rules in Large Databases,; **Proc. 20th Int'l Conf. Very Large Data Bases**, pp. 478-499, Sept. 1994.
- [16]Salzberg, S.L., ;On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach,; **Data Mining and Knowledge Discovery**, Kluwer Academic Publishers, Boston, Vol. 1, pp. 317-327,1997.
- [17]Tan, A.C. and Gilbert, D., ;An empirical comparison of supervised machine learning techniques in bioinformatics,; **Proceedings of the First Asia Pacific Bioinformatics Conference**, Adelaide, Australia, pp. 219-222,2003.
- [18]U. Fayyad, and P. Smyth, ;From Data Mining to Knowledge Discovery: An Overview,; **Advances in Knowledge Discovery and Data Mining**, pp. 1-36, 1996.
- [19]Weka, <http://www.cs.waikato.ac.nz/~ml/weka/index.html>.
- [20]Zhen Liu, Minyi Guo, ;A Proposal of Integrating Data Mining and On-Line Analytical Processing in Data Warehouse,; **Proceedings of 2001 International Conferences on Info-tech and Info-net**, pp.146~pp.151, October 2001.