南華大學科技學院資訊管理學系

碩士論文

Department of Information Management

College of Science and Technology

Nanhua University

Master Thesis

以Python實作多元入學學生之流失率與學業表現的分析

Using Python to Implement the Data Analysis for Dropout Rates and Course Performances of Multiple-Enrolled Students

吳梓豪

Zih-Hao Wu

指導教授：邱宏彬 博士

Advisor: Hung-Pin Chiu, Ph.D.

中華民國 108 年 6 月

June 2019

南 華 大 學

資訊管理學系

碩 士 學 位 論 文

以 Python 實作多元入學學生之流失率與學業表現的分析

Using Python to Implement the Data Analysis for Dropout Rates and Course Performances of Multiple-enrolled Students

研究生：吳梓豪

經考試合格特此證明

口試委員：林延澤

陳張宗榮

邱宏林

邱宏林

指導教授：

系主任(所長)：陳信良

口試日期：中華民國 108 年 6 月 28 日

以Python實作多元入學學生之流失率與學業表現的分析

學生：吳梓豪　　指導教授：邱宏彬 博士

南 華 大 學 資訊管理學系

摘 要

有鑑於國內少子化的影響，大專校院學生數劇減在民國 95年至105年間呈現負成長，招生日趨競爭，若能減少在校學生的流失，對學校而言則是一大助力。 本研究以一大學 102 學年度入學的資管系學生學籍資料，運用Python進行資料分析，找出多元入學學生流失率與學業表現等因素，提供相關之建議以降低學生流失。

本研究共取得有效資料 76 筆，在資料特性分析發現學生學業表現以繁星推薦學生學業表現仍為各入學管道中，學業表現較為突出部分；反觀轉學考進入學校之學生學業表現仍為最弱；所以學業表現與學生入學方式有關。在學生流失率部份，與居住地區和入學方式有關。最後，以資料採礦的決策樹建立流失的預測模型，並且分析和探討學生流失的重要因素，作為改善學生流失率的參考依據。

關鍵詞：python、決策樹、多元入學、學業表現、學生流失

Using Python to Implement the Data Analysis for Dropout Rates and Course Performances of Multiple-Enrolled Students

Student：Wu, Zih-Hao Advisor：Dr. Chiu, Hung-Pin

Department of Information Management

The M.I.M. Program

Nan-Hua University

ABSTRACT

In view of the impact of the domestic minority, college students in D.C. 2006 to 2016 years showed negative growth, enrollment increasingly competition, if the loss of students in school, the school is a big help. This study uses the data mining technology to find out the factors such as the loss rate of the students and the academic performance of the students, and provide relevant suggestions to reduce the loss of students.

In this study, a total of 76 valid data were obtained. In the analysis of the data, it was found that the academic performance of the students was still recommended by the stars. The academic performance of the students was still the most prominent part of the students. The academic performance of the students was still the weakest. So academic performance and student enrollment. In the part of the student's wastage, related to the area of residence and admission. In the data mining part of the discovery, to academic total score for the most important factor, school enrollment and residential areas as the second factor, followed by gender and class.

Keywords: python, decision tree, Multiple-enroll, course performances, student dropout

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，說明目前大專校院學生人數之現況；第二節為研究目的；第三節為研究範圍與限制；第四節為研究步驟；第五節論文架構。

第一節 研究背景與動機

依據教育部統計處公告大專校院概況中顯示學生人數民國 106 年版中報告指出，大專校院學生數自 95 學年起至 105 學年度產生大幅度負成長，總計約降 4,552 人（教育部統計處 106 年大專校院概況 http://stats.moe.gov.tw/files/important/OVERVIEW_U01.pdf)在各校招生情況日趨嚴峻的情況下，如果對於已招收的在校生能夠減少學生休退學、降低在校生流失留住學生，對於學校應是一大助力。

第二節 研究目的

為因應招生情況日益嚴峻，加上教育部逐年增加學生多元入學方案，學校應針對入學後學生在校學習與在校生活各種情形多加觀察注意，加強輔導關心學生使學生能適應學校環境，避免學生的流失，如此亦可讓學生家長對學校積極輔導的作法有正向的回應。

由於導師而言，管理班上眾多學生已屬不易，且學校各人員及任課教師之日常行政工作及教學活動已相當忙碌，如何協助及早發現潛在可能流失學生，而多加關心輔導以減少其流失率，則是一重要關鍵。

另一方面可提供系所在面對各多元管道招生學生學業表現上有所助益。

本研究擬運用 Python 分析技術分析學生歷史學籍資料，找出多元入學管道之流失學生其影響因素，作為在學學生可能流失之預測，針對 Python 分析之結果， 提供相關之建議，以探討多元入學學生流失與學業表現之情形。

第三節 研究範圍、限制

本研究之研究範圍以研究學校 102 學年度入學之大學日間部資管系學生為主，因分析來源為單一學校科系，結果恐難類推至所有大專學校，但仍可作為類似各校減少學生流失與各招收管道學生學業表現之參考；此外，本研究以學校校務系統中的資料庫之資料為分析來源，對於學生之情緒性因素、家庭、經濟、交友等情形較難以取得與分析；再者因個資限制為保護學生隱私有關德行成績相關之獎懲記錄、缺曠課資料，校方無法提供書面或電子檔案，僅在工作中可取得資料。

第四節 研究步驟

本研究將以研究學校 102 學年度大學日間部資管系入學學生之個人資料，研究探討個案學校多元入學學生流失與學業表現，研究工具為 Python 中之決策樹，並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、Python 與分析、結論與建議，其說明如下：

壹、確認研究動機及目的：

確認本研究之動機與目的，以確認資料蒐集及研究進行方向。

貳、資料蒐集與回顧文獻：

蒐集相關文獻瞭解研究之相關知識，並尋求適合之技術及工具。

參、擬定研究方法及流程：

針對所研究之問題及運用工具，擬定適合的研究方法及流程，以Python應用程式下做分析。

肆、取得分析資料：

本研究以個案學校102 學年度入學資管系學生之資料，作為研究使用的資料庫。為能正確地使用Python技術，在進行資料庫的分析之前，先行檢視資料中各資料欄位之意義與價值，了解各項資料原始意義及與使用特性和限制條件，以便能正確地運用分析方法與資料性質，進行資料分析工作。

伍、資料前置處理：

將不具分析價值之資料欄位加以刪減、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，以避免不適合的資料對分析結果造成不利的影響。

陸、Python資料分析：

運用Python進行相關資料之分析，以了解影響多元入學學生的學業表現與流失率的相關因素。最後，以資料採礦的決策樹建立流失的預

測模型，並且分析和探討學生流失的重要因素，作為改善學生流失率的參考依據。

柒、結論：

找出流失學生潛在之共同因素，作為在學學生可能流失之預測，針對 Python 分析之結果，提供可行之建議，以降低學生流失之情形。

第五節 論文架構

本研究共分為五章，各章節結構說明如下：

壹、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。

貳、文獻探討

就相關文獻分別探討學生流失及資料採礦之定義，並針對本研究所運用之統計資料分析與Python分析方法決策樹的加以深入探討。

參、研究方法

本章節針對研究資料來源作概略介紹，包括研究對象、研究架構、資料處理以及Python工具的介紹。

肆、資料分析

先對研究資料作基本敘述統計，再針對所需要進一步研究的部分

進行整理後做Python分析，並比較結果。

伍、結論與未來工作

針對第四章所得結果進行整理，並提出相關結論及未來工作。

第二章、文獻探討

第一節 我國大學多元入學制度發展沿革、理念及變革原因

壹、我國大學入學制度發展沿革

我國從民國43 年起正式辦理大學聯考，並於91 學年度起正式實施多元入學方案，至今我國大學制度已創立 57 個年頭。隨著時代的變遷大學升學制度受到諸多的關注，過去大學聯招制度，可以使不同社經背景的學生都有機會出人頭地，但卻也造成學生之間激烈競爭的現象（王家通，1991）。在大學聯招的制度下升學競爭日益激烈，使得臺灣學子迷失在功利主義之中。為改善該現象，教育部於民國65 年成立「大學入學考試委員會」，下設「試務委員會」辦理聯合招生，「研究委員會」則負責研究入學考試改進方法。至民國76 年，教育部的試務委員會專案研究小組認為有必要成立研究改進大學入學考試之專責機構，因此在民國78 年又成立了大學入學考中心，成為推動改革聯招制度的第一步，於民國82 年正式廢止大學入學考試委員會，招生工作重新回歸大學辦理，民國91 年起廢除傳統大學聯考，改採大學多元入學新方案，至此揭櫫多元入學時代。

多元入學方案試行至今將屆10 年，期間可透過保送、推甄、申請、繁星及考試分發等多元方式入學。由於多元管道下，大學入學相關試務日趨龐雜，反而致使考生、家長、學校承辦單位及試務人員產生混

淆，無法跟上教育改革的腳步。自一百學年度起，化繁為簡，將多元入學管道合併成為兩大管道，分別為「甄選入學」（含繁星推薦和個人申請）以及「考試分發」，使多元入學管道能更受到肯定。

依據丘愛鈴（1998）對於我國大學入學制度辦理的時期分類，本研究參酌相關研究者的列表(大學入學考試中心,1997;秦夢群,2004；蘇裕隆、葉連祺、陳恭，2005），整理出我國大學入學制度發展沿革（見表2-1），以利更進一步瞭解大學入學制度之背景。

(表2-1) 我國大學入學制度發展沿革表

分期	年代	招生機構	考試機構	招生管道	類組	錄取依據
創立期	民國43-60年	大專聯招會（ 四所公立大學組成)大學暨專科學校		大學暨專科學校	分甲乙丙丁四組考試： 1.甲組多為理工科系； 2.乙組多為文科科系； 3.丙組多為醫學與農業科系； 4.丁組多為法商科系	聯招分數
規劃期	民國61-72年	大學考試委員會(教育部官員兼任）		大學		
新制期	民國73-81年	大學考試委員會(教育部官員兼任）		大學	分一二三四類組考試，並可跨組選考： 1.一組為文法商科系； 2.二組為理工科系； 3.三組為醫學科系； 4.四組為農學科系	
	民國82-85年	大學自辦並委託財團法人大學入學考試中心擔任總會		大學入學考試中心辦理「學科能力測驗」及「甄選入學」，與大學聯招並行。		聯招分數 兩階段分數（學科能力驗、指定科目考試）
	民國86-90年	大學招生策進會(簡稱招策會）		公告大學多元入學新方案。大考中心辦理學科能力測驗。入學管道分為：資優生保送、甄選入學、考試分發入學。		
多元期	民國91-92年	大學招生委員會聯合會	財團法人大學入學考試中心	大學多元入學改進方案將甄選入學合併成繁星推薦和考試分發	依考試階段、考科、成績採計及分發方式不同區分為甲、乙、丙，由大學各校系擇一採行。	
	民國93- 至今	大學招生委員會聯合會	財團法人大學入學考試中心	96 學年度新增繁星計畫。而於民國 100 學年度整合「甄選入學」及「個人申請」，使入學管道簡化為「甄選入學」與「考試分發」二大管道。	考試分發入學甲、乙、丙三案招生選擇方式合而為一，考生均須參加指定科目考試。	大學科系自訂指定考試目，以3~6 科為限（含術科試，不含學科能力測驗）。

貳、我國大學多元入學制度理念及其變革原因

傳統大學聯招始自民國43 年實施至90 年止，有48 年之久，此制度是過去國人所熟悉的一試定終身的傳統聯考管道入學，過去在臺灣功績主義的影響下，該制度不斷增強文憑主義宰制臺灣社會的意識形態。隨著時勢變化，大學聯招制度受到了來自各界的壓力，不得不進行教育制度變革。

高等教育負有強化人力素質，建立與國際匹敵的競爭力，其選才制度舉足輕重。秦夢群（2004）曾簡要整理美國、法國和日本的大學入學制度，說明該些國家的制度有主要幾點特色：（1）大學享有招生自主權；（2）參酌入學依據的多樣化；（3）學生有選校選系的自主權；（4）提供多樣的入學途徑。反觀我國，在國內的輿論壓力逐漸形成之下，大學聯招政策受到諸多批評，主要問題為：（1）一試定終身的升學方式；（2）聯招考科的僵化；（3）過分強調制式答案的考試方式及(4)過分重視大學科系排行,無法達成適性發展(教育部，2002)。相較於先進國家的入學制度採用彈性且多重標準衡量學生的能力，相形之下我國的聯招制度即有改善的空間。

基於大學聯招政策的缺失,大學入學考試中心(2002)在民國91 年提出的《我國大學入學制度改革建議書—大學多元入學方案》中，即提到以多元入學方式替代大學聯招，希冀能達到三大具體目標，分別為：(1)學生學習與選擇方面：重視學習歷程、顧及學生性向與興趣、

激勵向學動機、提供多元入學途徑、尊重家長教育選擇權、顧及弱勢族群教育機會；（2）學校特色與選才方面：尊重學校招生自主性、促進學校間均衡發展、輔導學校發展特色、建立學生多元價值觀念、多元評量學生學習成就、符合公平公正、公開的精神；（3）教育發展方面：促進學生五育均衡發展、提升適性教學品質、減緩過度升學競爭壓力。莊佩真（2002）也提及多元入學方案規畫精神，作
為我國大學多元入學制度的指引，分別為：（1）多元智慧：藉由多元入學方案，促進教學正常化，發展學生多元性向；（2）多元選擇：各項自行選擇多元招生方式，學生主動選擇入學方式；（3）多元特色：藉由多元入學方案，促使大學科系發展特色。

根據教育部（2010）《升學制度審議委員會總結報告》對各種升學管道的優缺點進行全面性的檢討，其中提及實施大學多元入學制度以來，學生升學壓力仍存在，主要原因有三：(1)升學管道過於繁雜，仍有諸多家長和學生不瞭解實施內容；（2）傳統價值觀念影響，無法跳脫考試引導教學的弊病；（3）多元入學方案規劃未臻周延，無法合乎時勢的需求。如同報告所提，一項升學制度必須要能從旁協助學校、學生選擇，促使學校發展本位特色、引導社會企業用人選才，應作為升學制度變革的努力方向。

在多次教育變革之下，本研究試圖歸納其背後的變革因素如下：

一、 創立期變革（民國43-60 年）：

奠定我國自行培育和遴選人才之制度，以考試公平性為主要考量。

二、 規劃期變革（民國61-72 年）：

制定更完善的考試制度，以考試公平性為主要考量。

三、新制期變革（民國73-90 年）：

追求考試制度的完善之餘，尚服膺社會的多元化及經濟發展的需求，逐步開放更多元的入學管道，發掘有潛力之學生，但仍較為保守。初期以考試公平性為著眼，隨著甄選入學、資優生保送管道的開放，將原本僅注意考試公平性的入學制度轉專注在學生選擇機會的發展。

四、多元期變革（民國91 年-迄今）：

不論是國際或國內對於人才培育的需求，皆已走向學生適性發展、大學自主招生以及發展學校特色階段。從民國91年起發展多元的管道，最早以增加學生選擇機會為主要目標，曾有學校推薦、個人申請、繁星計畫、考試分發等管道，讓學生有更多的選擇機會。隨著大學多元入學方案的實施愈臻成熟，所關注的焦點將回歸到學生的適性發展，因此在96 學年度和100 學年度所進行的教育變革，所注重的是學生能否透過大學多元入學管道依據自己的性向選擇適合

的科系，然後在進入大學後適性發展與學習，奠定個人得職涯基礎成為更卓越之人才。此也是符合本研究的卓越指標。

由上述可發現我國高等教育變革的軌跡，沿著公平性、選擇機會、卓越性等三個指標之順序調整教育政策之比重。可預期我國大學多元入學方案將本著多元化、大學自主、學生適性發展等理念下逐步革新，透過選才制度帶領學生更適性發展、大專院校更自主選才、高中、高職學校教學正常化。

第二節 資料採礦的定義

Data mining 經中華資料採礦協會（Chung-Hua Data Mining Society）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Usama 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 etal., 1992)。其實，資料採礦並非是一種技術或者是一套軟體。事實上， 它是一種結合數種專業技術的應用。並且不是無所不能，它只是從大量的資料中發掘出各種假設（Hypothesis）；但是，它並不會幫忙檢查，也不會幫忙確認假設。同時，無法協助判斷這些假

設對使用者的價值為何。（謝邦昌、鄭宇庭、蘇志雄，2011）。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。

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)。

一般來說，大致上常用的資料採礦技術主要有下述六類(許依宸，民國98)：

壹、 分類（Classification）：

預測類別變數的過程，我們稱之為「分類」。依據已知的資料及其分類的屬性，建立出資料的分類模型；接著，利用此分類模型來預測新資料的類別。例如：顧客的購物習性分類模型..等。

貳、 推估（Estimation）：

運用處理過後連續性數值的結果，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）：

群集意為物以類聚。即依資料本身的自我相似性（self-similarity）

而群集在一起。群集(Clusters)的意義則要經由事後之闡釋方能得知。

肆、關聯法則（Association Rule）：

指的是從歷史資料中，找出哪些事件總是相伴發生。

伍、序列（Sequential）：

即在同質分組中透過序列來找出事物「先後」發生的順序，這樣的規則被稱為時序規則（Sequential Pattern）。

陸、描述（Description）：

指的是在資料採礦的過程中，除了分析的預測模型以外，更重要的是在分析與處理資料的過程中，透過資料視覺化以及觀察來找出許多有意義的規則。

第三節 資料採礦的功能與應用

本研究是以 Python 應用程式軟體，什麼是資料採礦？資料採礦指的是對現有的一些資料進行相應的處理和分析，最終得到資料與資料之間深層次關係的一種技術。例如在對超市貨品進行擺放時，牛奶到底是和麵包擺放在一起銷量更高，還是和其他商品擺在一起銷量更高。資料探勘技術就可以用於解決這類問題。具體來說，超市的貨品擺放問題可以劃分為關聯分析類場景。

在日常生活中，資料採礦技術應用的非常廣泛。例如對於商戶而言，常常需要對其客戶的等級（svip、vip、普通客戶等）進

行劃分，這時候可以將一部分客戶資料作為訓練資料，另一部分客戶資料作為測試資料。然後將訓練資料輸入到模型中進行訓練，在訓練完成後，輸入另一部分資料進行測試，最終實現客戶等級的自動劃分。其他類似的應用例子還有驗證碼識別、水果品質自動篩選等。

壹、分類

資料採礦技術可以用於解決分類問題，如對客戶等級進行劃分、驗證碼識別、水果品質自動篩選等。

以驗證碼識別為例，現需要設計一種方案，用以識別由 0 到 9 的手寫體數字組成的驗證碼。有一種解決思路是，先將一些出現的 0 到 9 的手寫體數字劃分為訓練集，然後人工的對這個訓練集進行劃分，即將各個手寫體對映到其對應的數字類別下面，在建立了這些對映關係之後，就可以通過分類演算法建立相應的模型。這時候如果出現了一個新的數字手寫體，該模型可以對該手寫體代表的數字進行預測，即它到底屬於哪個數字類別。例如該模型預測某手寫體屬於數字 1 的這個類別，就可以將該手寫體自動識別為數字 1。所以驗證碼識別問題實質上就是一個分類問題。

水果品質的自動篩選問題也是一個分類問題。水果的大小、顏色等特徵也可以對映到對應的甜度類別下面，例如 1 這個類別

可以代表甜，0 這個類別代表不甜。在獲得一些訓練集的資料之後，同樣可以通過分類演算法建立模型，這時候如果出現一個新的水果，就可以通過它的大小、顏色等特徵來自動的判斷它到底是甜的還是不甜的。這樣就實現了水果品質的自動篩選。

貳、迴歸：對連續型資料進行預測、趨勢預測等

除了分類之外，資料採礦技術還有一個非常經典的場景——迴歸。在前文提到的分類的場景，其類別的數量都有一定的限制。比如數字驗證碼識別場景中，包含了 0 到 9 的數字類別；再比如字母驗證碼識別場景中，包含了 a 到 z 的有限的類別。無論是數字類別還是字母類別，其類別數量都是有限的。

現在假設存在一些資料，在對其進行對映後，最好的結果沒有落在某個 0、1 或者 2 的點上，而是連續的落在 1.2、1.3、1.4...上面。而分類演算法就無法解決這類問題，這時候就可以採用迴歸分析演算法進行解決。在實際的應用中，迴歸分析演算法可以實現對連續型資料進行預測和趨勢預測等。

參、聚類：客戶價值預測、商圈預測等

什麼是聚類？在上文中提過，要想解決分類問題，必須要有歷史資料（即人為建立的正確的訓練資料）。倘若沒有歷史資料，而需要直接將某物件的特徵劃分到其對應的類別，分類演算法和迴歸演算法無法解決這個問題。這種時候有一種解決辦

法——聚類，聚類方法直接根據物件特徵劃分出對應的類別，它是不需要經過訓練的，所以它是一種非監督的學習方法。

在什麼時候能用到聚類？假如資料庫中有一群客戶的特徵資料，現在需要根據這些客戶的特徵直接劃分出客戶的級別（如SVIP 客戶、VIP 客戶），這時候就可以使用聚類的模型去解決。另外在預測商圈的時候，也可以使用聚類的演算法。

肆、關聯分析：超市貨品擺放、個性化推薦等

關聯分析是指對物品之間的關聯性進行分析。例如，某超市記憶體放有大量的貨品，現在需要分析出這些貨品之間的關聯性，如麵包商品與牛奶商品之間的關聯性的強弱程度，這時候可以採用關聯分析演算法，藉助於使用者的購買記錄等資訊，直接分析出這些商品之間的關聯性。在瞭解了這些商品的關聯性之後，就可以將之應用於超市的商品擺放，通過將關聯性強的商品放在相近的位置上，可以有效提升該超市的商品銷量。

此外，關聯分析還可以用於個性化推薦技術。比如，藉助於使用者的瀏覽記錄，分析各個網頁之間存在的關聯性，在使用者瀏覽網頁時，可以向其推送強關聯的網頁。例如，在分析了瀏覽記錄資料後，發現網頁 A 與網頁 C 之間有很強的關聯關係，那麼在某個使用者瀏覽網頁 A 時，可以向他推送網頁 C，這樣就實現了個性化推薦。

伍、自然語言處理：文字相似度技術、聊天機器人等

除了上述的應用場景之外，資料採礦和機器學習技術也可以用於自然語言處理和語音處理等等。例如對文字相似度的計算和聊天機器人。

第四節 決策樹

壹、分類（classification）

分類其實是從特定的資料中挖掘模式，作出判斷的過程。比如 Gmail 郵箱裡有垃圾郵件分類器，一開始的時候可能什麼都不過濾，在日常使用過程中，我人工對於每一封郵件點選“垃圾”或“不是垃圾”，過一段時間，Gmail 就體現出一定的智慧，能夠自動過濾掉一些垃圾郵件了。

這是因為在點選的過程中，其實是給每一條郵件打了一個“標籤”，這個標籤只有兩個值，要麼是“垃圾”，要麼“不是垃圾”，Gmail 就會不斷研究哪些特點的郵件是垃圾，哪些特點的不是垃圾，形成一些判別的模式，這樣當一封信的郵件到來，就可以自動把郵件分到“垃圾”和“不是垃圾”這兩個我們人工設定的分類的其中一個。

分類學習主要過程如下：

一、 訓練資料集存在一個類標記號，判斷它是正向資料集（起積極作用，不垃圾郵件），還是負向資料集（起抑制作用，垃圾郵件）。

二、 然後需要對資料集進行學習訓練，並構建一個訓練的模型。

三、 通過該模型對預測資料集進預測，並計算其結果的效能

貳、決策樹（decision tree）

決策樹是用於分類和預測的主要技術之一，決策樹學習是以例項為基礎的歸納學習演算法，它著眼於從一組無次序、無規則的例項中推理出以決策樹表示的分類規則。構造決策樹的目的是找出屬性和類別間的關係，用它來預測將來未知類別的記錄的類別。它採用自頂向下的遞迴方式，在決策樹的內部節點進行屬性的比較，並根據不同屬性值判斷從該節點向下的分支，在決策樹的葉節點得到結論。

決策樹演算法根據資料的屬性採用樹狀結構建立決策模型， 決策樹模型常用來解決分類和迴歸問題。常見的演算法包括：分類及迴歸樹（Classification And Regression Tree， CART）， ID3 (Iterative Dichotomiser 3)， C4.5，Chi-squared Automatic Interaction Detection(CHAID), Decision Stump, 隨機森（Random Forest）， 多元自適應迴歸樣條（MARS）以及梯度推進機（Gradient Boosting Machine，

GBM）。

決策數有兩大優點：1）決策樹模型可以讀性好，具有描述性，有助於人工分析；2）效率高，決策樹只需要一次構建，反覆使用，每一次預測的最大計算次數不超過決策樹的深度。

第五節 Python的應用

Python（英國發音：/ˈpaɪθən/ 美國發音：/ˈpaɪθɑːn/）是一種廣泛使用的直譯式、進階編程、通用型程式語言，由吉多·范羅蘇姆創造，第一版釋出於 1991 年。可以視之為一種改良（加入一些其他程式語言的優點，如物件導向）的 LISP。Python 的設計哲學強調程式碼的可讀性和簡潔的語法（尤其是使用空格縮排劃分程式碼塊，而非使用大括號或者關鍵詞）。相比於 C++或 Java，Python 讓開發者能夠用更少的代碼表達想法。不管是小型還是大型程式，該語言都試圖讓程式的結構清晰明了。

與 Scheme、Ruby、Perl、Tcl 等動態型別程式語言一樣，Python 擁有動態型別系統和垃圾回收功能，能夠自動管理記憶體使用，並且支援多種編程範式，包括物件導向、命令式、函數式和程序式編程。其本身擁有一個巨大而廣泛的標準庫。

Python 直譯器本身幾乎可以在所有的作業系統中執行。Python 的其中一個直譯器 CPython 是用 C 語言編寫的、是一個由社群驅動的自由軟體，目前由 Python 軟體基金會管理。

Python 經常被用於 Web 開發。比如，通過 mod_wsgi 模組，Apache 可以運行用 Python 編寫的 Web 程式。使用 Python 語言編寫的 Gunicorn 作為 Web 伺服器，也能夠執行 Python 語言編寫的 Web 程式。Python 定義了 WSGI 標準應用介面來協調 Http 伺服器與基於 Python 的 Web 程式之間的溝通。一些 Web 框架，如 Django、Pyramid、TurboGears、Tornado、web2py、Zope、Flask 等，可以讓程式設計師輕鬆地開發和管理複雜的 Web 程式。

Python 對於各種網路協定的支援很完善，因此經常被用於編寫伺服器軟體、網路爬蟲。第三方函式庫 Twisted 支援非同步線上編寫程式和多數標準的網路協定（包含用戶端和伺服器），並且提供了多種工具，被廣泛用於編寫高效能的伺服器軟體。另有 gevent 這個流行的第三方庫，同樣能夠支援高效能高並行的網路開發。

很多遊戲使用 C++編寫圖形顯示等高效能模組，而使用 Python 或者 Lua 編寫遊戲的邏輯、伺服器。相較於 Python，Lua 的功能更簡單、體積更小；而 Python 則支援更多的特性和資料類型。很多遊戲，如 EVE Online 使用 Python 來處理遊戲中繁多的邏輯。

YouTube、Google、Yahoo!、NASA 都在內部大量地使用 Python。OLPC 的作業系統 Sugar 項目的大多數軟體都是使用 Python 編寫。

壹、GUI 開發

Python 本身包含的 Tkinter 庫能夠支援簡單的 GUI 開發。但是越來越多的 Python 程式設計師選擇 wxPython 或者 PyQt 等 GUI 套件來開發跨平台的桌面軟體。使用它們開發的桌面軟體執行速度快，與用戶的桌面環境相契合。通過 PyInstaller 還能將程式釋出為獨立的安裝程式包。

貳、作業系統

在很多作業系統裡，Python 是標準的系統元件。大多數 Linux 發行版和 Mac OS X 都整合了 Python，可以在終端機下直接執行 Python。有一些 Linux 發行版的安裝器使用 Python 語言編寫，比如 Ubuntu 的 Ubiquity 安裝器、Red Hat Linux 和 Fedora 的 Anaconda 安裝器。在 RPM 系列 Linux 發行版中，有一些系統元件就是用 Python 編寫的。Gentoo Linux 使用 Python 來編寫它的 Portage 軟體包管理系統。Python 標準庫包含了多個調用作業系統功能的函式庫。通過 pywin32 這個第三方軟體包，Python 能夠存取 Windows 的 COM 服務及其它 Windows API。使用 IronPython，Python 程式能夠直接調用.Net Framework。

參、科學計算的套件

NumPy、Pandas、Matplotlib 和 SciPy 是 Python 非常流行的常用套件，可以讓 Python 程式設計師編寫科學計算程式。

一、NumPy：

NumPy 是 Python 語言的一個擴充程式庫。支援高階大量的維度陣列與矩陣運算，此外也針對陣列運算提供大量的數學函式函式庫。NumPy 的前身 Numeric 最早是由 Jim Hugunin 與其它協作者共同開發，2005 年，Travis Oliphant 在 Numeric 中結合了另一個同性質的程式庫 Numarray 的特色，並加入了其它擴充功能而開發了 NumPy。NumPy 為開放原始碼並且由許多協作者共同維護開發。

NumPy 參考 CPython(一個使用位元組碼的直譯器)，而在這個 Python 實作直譯器上所寫的數學演算法程式碼通常遠比編譯過的相同程式碼要來得慢。為了解決這個難題，NumPy 引入了多維陣列以及可以直接有效率地操作多維陣列的函式與運算子。因此在 NumPy 上只要能被表示為針對陣列或矩陣運算的演算法，其執行效率幾乎都可以與編譯過的等效 C 語言程式碼一樣快。

NumPy 提供了與 MATLAB 相似的功能與操作方式，因為兩者皆為直譯語言，並且都可以讓使用者在針對陣列或矩陣運

算時提供較純量運算更快的效能。兩者相較之下，MATLAB 提供了大量的擴充工具箱(例如 Simulink)；而 NumPy 則是根基於 Python 這個更現代、完整並且開放原始碼的程式語言之上。此外 NumPy 也可以結合其它的 Python 擴充函式庫。例如 SciPy，這個函式庫提供了更多與 MATLAB 相似的功能；以及 Matplotlib，這是一個與 MATLAB 內建繪圖功能類似的函式庫。而從本質上來說，NumPy 與 MATLAB 同樣是利用 BLAS 與 LAPACK 來提供高效率的線性代數運算。

二、Pandas：

在計算機編程中，**pandas** 是為 Python 編程語言編寫的軟件庫，用於數據操作和分析。特別是，它提供了用於操作數值表和時間序列的數據結構和操作。它是根據三條款 BSD 許可證發布的免費軟件。該名稱源自術語“ 面板數據 ”，這是一個數據集的計量經濟學術語，包括對同一個體的多個時間段的觀察。

（一）DataFrame 對象，用於集成索引的數據操作。

（二）用於在內存數據結構和不同文件格式之間讀取和寫入數據的工具。

（三）數據對齊和缺失數據的集成處理。

（四）重塑和轉動數據集。

（五）基於標籤的切片，花式索引和大型數據集的子集化。

（六）數據結構列的插入和刪除。

（七）按引擎分組，允許對數據集進行拆分應用組合操作。

（八）數據集合併和加入。

（九）分層軸索引以在較低維數據結構中處理高維數據。

（十）時間序列 – 功能：日期範圍生成和頻率轉換，移動窗口統計，移動窗口線性回歸，日期轉換和滯後。

（十一）提供數據過濾。

該庫針對性能進行了高度優化，使用 Cython 或 C 編寫了關鍵代碼路徑。

三、**Matplotlib**：

matplotlib 是 Python 程式語言及其數值數學擴展包 NumPy 的可視化操作界面。它利用通用的圖形用戶界面工具包，如 Tkinter, wxPython, Qt 或 GTK+，向應用程式嵌入式繪圖提供了應用程式接口（API）。此外，matplotlib 還有一個基於圖像處理庫（如開放圖形庫 OpenGL）的 pylab 接口，其設計與 MATLAB 非常類似--儘管並不怎麼好用。SciPy 就是用 matplotlib 進行圖形繪製。

matplotlib 最初由 John D. Hunter 撰寫，它擁有一個活躍的開發社區，並且根據 BSD 樣式許可證分發。 在 John D. Hunter2012 年去世前不久，Michael Droettboom 被提名為 matplotlib 的主要開發者。

截至到 2015 年 10 月 30 日，matplotlib 1.5.x 支持 Python 2.7 到 3.5 版本。Matplotlib 1.2 是第一個支持 Python 3.x 的版本。Matplotlib 1.4 是支持 Python 2.6 的最後一個版本。

優點：

（一）帶有內置代碼的默認繪圖樣式。

（二）與 Python 的深度集成。

（三）Matlab 風格的編程接口（對一些人來說是優點，但對於其他人來說可能是缺點）。

（四）圖形繪製相較 Gnuplot 更加美觀。

（五）跨語言解決方案：可以用作通過管道或文件以不同語言編寫的應用程式（例如 GNU Octave，Maxima，JavaGnuplotHybrid）中的繪圖引擎。

（六）獨立程序：沒有外部依賴。

（七）處理大型數據集時非常快。

（八）更容易操縱繪圖細節。

缺點：

（一）高度依賴其他包，如 Numpy。

（二）只適用於 Python：很難/不可能在 Python 以外的語言中使用。（但可以從 Julia 通過 PyPlot 軟體包使用）

（三）舊的默認繪圖樣式：通常需要小的調整以產生有吸引力的圖。

（四）在開發中活躍成員的數量較少（與 Matplotlib 相比）。

四、**SciPy NumPy：**

SciPy 是一個開源的 Python 演算法庫和數學工具包。

SciPy 包含的模組有最佳化、線性代數、積分、插值、特殊函數、快速傅立葉變換、訊號處理和圖像處理、常微分方程式求解和其他科學與工程中常用的計算。與其功能相類似的軟體還有 MATLAB、GNU Octave 和 Scilab。

第三章 研究方法

本章節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流程； 第三節說明資料處理方式；第四節針對 Python 應用程式工具進行介紹。

第一節 研究對象

本研究以 102學年度入學大學日間部資管系學生為主研究對象，由於研究樣本選取時間為 105學年度第 2 學期，故其就學狀態為 102學年度第1學期至105 學年度第 2 學期時之就學狀態。 個案學校學生休學是採填寫申請表方式，而學生填寫休學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、退學及轉學等， 轉學分為轉入與轉出。由校務行政系統上並無法確實得知學生離校之實際原因，故本研究採用資料採礦方式，擬探求學生流失率，希望藉由發現各多元入學管道之學生學習表現與學生流失率是否相關，提供個案學校作參考。

第二節 研究流程

本研究的主要研究變數為入學管道、性別、居住地區、學業成績，主要為校務行政系統上之學生資料。然後用決策樹來分析資料，找出各入學管道學生流失率關聯。本研究之研究流程如圖3- 1所示，確定

研究動機、探討相關文獻、分析資料特性、資料前置處理、資料特性分析、決策樹分析、結論與未來工作。

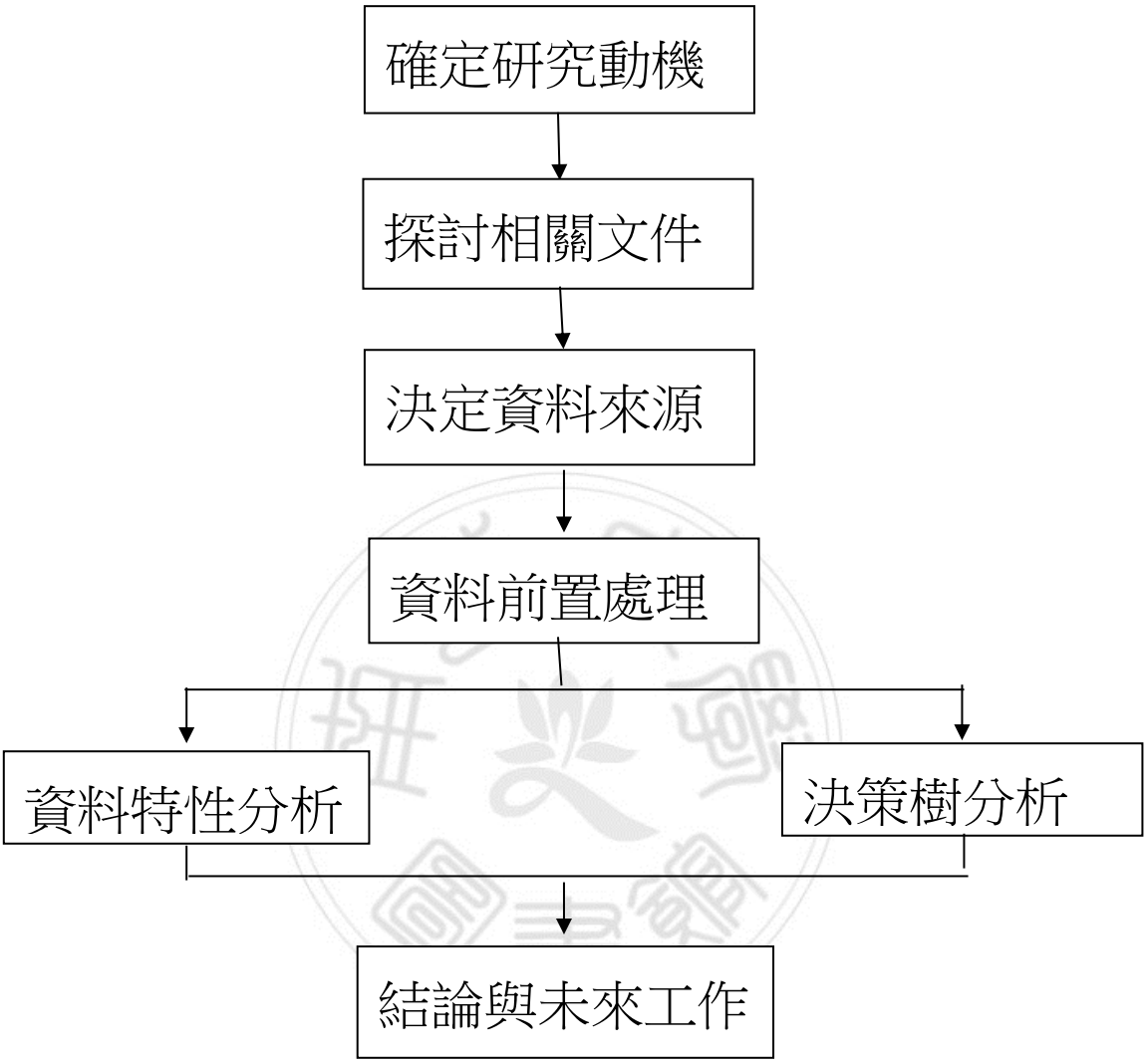

圖 3-1 研究流程圖

資料來源:本研究整理

第三節 資料前置處理

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」。依本研究流失之定義，將所取得的學生資料中異動原因為休學、退學、轉

學(轉出）之學生，其就學狀態為不在學。

了解每個資料欄位所代表的含意與其使用之價值後，由於考慮到資料庫中有部分欄位不具分析價值，如電話、身份證號碼、家長姓名等，故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。

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；同時刪除在就學狀況進進出出之狀況學生刪除。

貳、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，如居住地址轉為縣市別；地址資料中之臺與台統一改為台字；桃園縣與桃園市縣市合併後，統一更改為桃園市。另將每學期學業成績計算出學業平均。

參、資料整合：

將大學四年的8學期學業成績與最後的總成績進行整合，以利進行相關後續分析。

肆、資料處理後學生資料表

說明如表 3-1 學生資料表所示

表 3-1 學生資料表(一)

欄名	屬性	長度	說明
學號	文字	8	102XXXXX
性別	文字	1	男、女
居住地區	文字	3	XX 縣、XX 市
各學期學業成績	數字	浮點	
班級	文字	1	A、B
入學方式	文字	4	考試分發、個人申請 繁星推薦、轉學考
就學狀態	文字	2	休學、在學、退學、轉學
學期總成績	數字	浮點	

資料來源：本研究整理

第四章、資料分析

本章共有二節，第一節資料特性分析，針對所取得之資料進行描述性統計，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。

第一節 資料特性分析

壹、資料說明：

本研究所得之資料為 102 學年度第 1 學期至 105 學年度第 2 學期大學日間部資管系 102 級學生資料，共取得 76 筆，本節將先運用 Python 從許多角度來分析資料，找出影響多元入學學生流失率與學業表現等相關因素的特性。

一、 性別比例分析

首先分析資料集內學生的性別比例。Python 程式碼如下所示：

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
# 讀入 Excel 資料檔
df = pd.read_excel(r"D:\研究生\梓豪\DataSet102.xlsx", encoding="utf8")
# 畫 性別 Pie 圖
groupbySex = df.groupby('性別').size(). reset_index(name='counts')
#type(groupbySex) # DataFrame
print(groupbySex)
groupbySex.to_html(r"D:\梓豪\sex.html")
```

```
labels = groupbySex['性別']
ratings = groupbySex['counts']
patches, texts, autotexts = plt.pie(ratings, labels=labels, autopct="%1.1f%%")
plt.legend(patches, labels, loc="best")
plt.title("性別比例")
plt.axis("equal")
plt.show()
```

Python 程式碼的執行結果如下所示：

表 4-1 102 級資管系學生男、女生人數表

	性別	counts
0	女	11
1	男	65

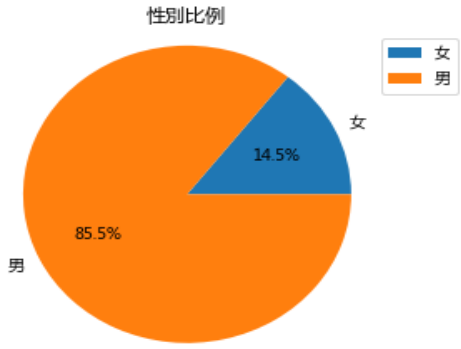

圖4-1 102級資管系男女生比例圖

資料來源：本研究整理

由表 4-1 和圖 4-1 可看出 102 級資管系男、女生的人數和比例，其中男生為 65 人、女生為 11 人，男生比例是 85.5%、女生比例為 14.5%。

二、 流失狀態分析：

接著分析學生的流失狀態比例。Python 程式碼如下所示：

```
# 流失狀態
groupbyDropout = df.groupby('流失
').size().reset_index(name='counts')
print(groupbyDropout)
groupbyDropout.to_html("D:\梓豪\Dropout.html")
labels = groupbyDropout['流失']
ratings = groupbyDropout['counts']
patches, texts, autotexts = plt.pie(ratings, labels=labels,
autopct="%1.1f%%")

plt.legend(patches, labels, loc="best")
plt.title("流失狀態")
plt.axis("equal")
plt.show()
```

Python 程式碼的執行結果如下所示：

表 4-2 流失資料表

	流失	counts
0	否	60
1	是	16

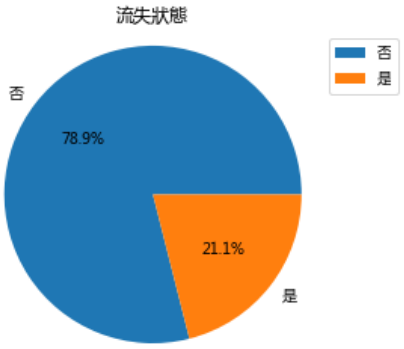

圖4-2學生流失狀態圖

資料來源：本研究整理

由表 4-2 可看出本研究資料中流失狀態未流失有 60 人、流失 16 人，由圖 4-2 學生流失狀態表可看出流失比例為 21%。

三、 就學狀態分析:

分析多元入學學生的就學狀態比例的 Python 程式碼如下所示：

```
# 就學狀態
groupbyStatus = df.groupby('就學狀態').size().reset_index(name='counts')
print(groupbyStatus)
groupbyStatus.to_html("D:\梓豪\status.html")
labels = groupbyStatus['就學狀態']
```

```
ratings = groupbyStatus['counts']
patches, texts, autotexts = plt.pie(ratings, labels=labels, autopct="%1.1f%%")
plt.legend(patches, labels, loc="best")
plt.title("就學狀態比例")
plt.axis("equal")
plt.show()
```

Python 程式碼的執行結果如下所示：

表 4-3 不同就學狀況比例表

	就學狀態	counts
0	休學	1
1	在學	60
2	轉學	9
3	退學	6

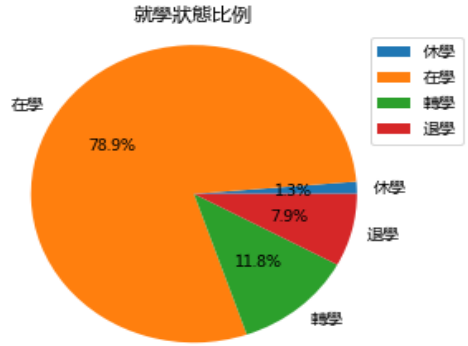

圖4-3就學狀況比例圖

資料來源：本研究整理

由圖 4-3 就學狀況比例圖中，可看出不在學的人數比例中轉學比例 12.1%最高。

四、 各入學方式與學業成績表現分析:

分析多元入學學生學業表現的 Python 程式碼如下所示：

```
# 入學方式與學業表現
df2 = pd.read_excel(r"D:\梓豪\學年成績 102_105.xlsx", encoding="utf8")
df2[df2 == 0.0] = np.nan
groupbyAdmission = df2.groupby('入學方式').mean()
print(groupbyAdmission)
groupbyAdmission.to_html("D:\梓豪\Admission.html")

labels = ["102 學年度", "103 學年度", "104 學年度", "105 學年度"]
index = np.arange(len(labels) * 4)

# 個人申請
admin_1 = groupbyAdmission.iloc[0]
#print(admin_1)
plt.bar(index[0::4], admin_1, label="個人申請")
# 繁星推薦
admin_2 = groupbyAdmission.iloc[1]
plt.bar(index[1::4], admin_2, label="繁星推薦")
# 考試分發
admin_3 = groupbyAdmission.iloc[2]
plt.bar(index[2::4], admin_3, label="考試分發")
# 轉學考
admin_4 = groupbyAdmission.iloc[3]
```

```
plt.bar(index[3::4], admin_4, label="轉學考")


plt.ylim((0, 100))
plt.legend(loc='upper center', ncol=len(groupbyAdmission.columns))
plt.xticks(index[1::4], labels)
plt.ylabel("平均成績")
plt.title("入學方式與學業表現")
plt.show()
```

Python 程式碼的執行結果如下所示：

表 4-4 入學方式學業表現一覽表

	102 成績	**103 成績**	**104 成績**	**105 成績**
入學方式				
個人申請	71.572778	70.122174	76.475000	78.655714
繁星推薦	78.435000	80.680000	81.120000	81.585000
考試分發	76.788889	76.547692	77.446857	79.666714
轉學考	NaN	73.000000	60.951667	64.971667

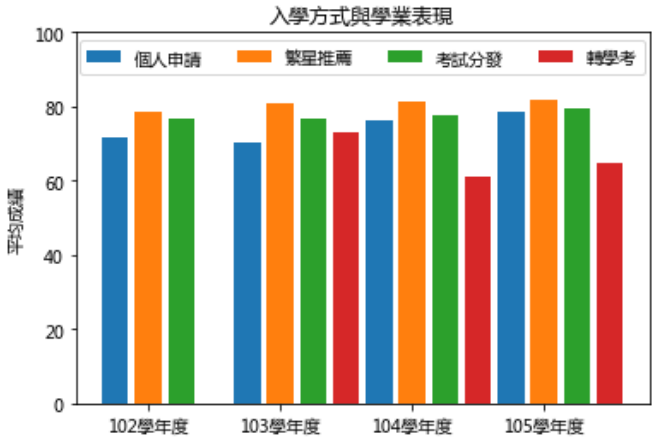

圖 4-4 入學方式與學業成績分析圖

由表 4-4 及圖 4-4 看出，繁星推薦學生學業表現仍為各入學管道中，學業表現較為突出部分；反觀轉學考進入學校之學生學業表現仍為最弱。

五、 居住地區就讀人數與流失人數資料分析:

接著分析學生在居住地區的就讀人數與流失人數。Python 程式碼如下所示：

```
s_groupbyArea = df.groupby('居住地區')
def NumDropout(s):
    ctr = 0
    for x in s:
        if x == '是':
            ctr = ctr + 1
    return ctr

groupbyArea = s_groupbyArea.agg({'學生證號':'size', '流失
':NumDropout}).rename(columns = {'學生證號’:’就讀人數’, ’流失’:’流失人數’ })
print(groupbyArea)
```

Python 程式碼的執行結果如下所示：

表 4-5 居住地區就讀人數與流失人數表

	就讀人數	流失人數
居住地區		
台中市	5	1
台北市	3	1
台南市	13	2
嘉義市	1	0
嘉義縣	5	3
基隆市	1	0
屏東縣	8	2
彰化縣	5	1
新北市	5	2
新竹市	1	0
桃園市	4	1
花蓮縣	3	1
苗栗縣	1	0
雲林縣	4	1
高雄市	17	1

資料來源：本研究整理

由表 4-5 可看出，以高雄市就讀人數 17 人最多，但流失人數僅有 1 人，推估表示於高雄市地區仍占有較大比例招生市場；反觀於嘉義縣地區雖就讀人數非為最少，但在流失比例上卻是最高，其流失人數與居住地區相關。

六、 入學方式與流失人數資料分析

分析多元入學學生的入學方式與流失人數的 Python 程式碼如下所示：

```
s_groupbyAdmin = df.groupby('入學方式')
groupbyAdmin = s_groupbyAdmin.agg({'學生證號':'size', '流失
':NumDropout}).rename(columns = {'學生證號’:’就讀人數’, ’流失’:’流失人
數'})
#print(type(groupbyAdmin))
#print(groupbyAdmin.columns)
groupbyAdmin.loc[:, "比例"] = groupbyAdmin.loc[:, "流失人數"] /
groupbyAdmin.loc[:, "就讀人數"]
print(groupbyAdmin)
```

Python 程式碼的執行結果如下所示：

表 4-6 入學方式流失人數統計與比例表

	就讀人數	流失人數	比例
入學方式			
個人申請	27	6	0.222222
繁星推薦	1	0	0.000000
考試分發	45	10	0.222222
轉學考	3	0	0.000000

由表 4-6 可看出，在繁星推薦及轉學考中，學生的流失人數為 0，基本上多為穩定；而在考試分發及個人申請上的流失比例皆為

22.22%左右，顯示出流失學生多為此兩類學生，其流失人數與入學方式相關。

七、 居住地區、入學方式、就讀人數與流失人數分析

進一步分析多元入學學生的居住地區、入學方式、就讀人數與流失人數的 Python 程式碼如下所示：

```
#  居住地區、入學方式、就讀人數與流失人數
s_groupbyAreaAdmin = df.groupby(['居住地區', '入學方式'])
groupbyAreaAdmin = s_groupbyAreaAdmin.agg({'學生證號':'size', '流失' :
NumDropout}).rename(columns = {'學生證號' :' 就讀人數', '流失' :' 流失人
數'})
print(groupbyAreaAdmin)
groupbyAreaAdmin.to_html("groupbyAreaAdmin.html")
```

下列是程式碼的解釋：

groupby()函數使用“居住地區”和 “入學方式“欄位來群組資料，agg 是針對分完組後的各個群組做聚合運算，算出各群組就讀人數和流失人數。

Python 程式碼的執行結果如下所示：

表 4-7 居住地區、入學方式、就讀人數與流失人數統計表

		就讀人數	流失人數
居住地區	入學方式		
台中市	個人申請	1	1
	考試分發	4	0
台北市	考試分發	3	1
台南市	個人申請	8	2
	考試分發	5	0
嘉義市	考試分發	1	0
嘉義縣	個人申請	3	2
	考試分發	1	1
	轉學考	1	0
基隆市	考試分發	1	0
屏東縣	個人申請	2	0
	繁星推薦	1	0
	考試分發	5	2
彰化縣	個人申請	1	0
	考試分發	4	1
新北市	考試分發	5	2
新竹市	個人申請	1	0
桃園市	個人申請	3	1
	考試分發	1	0
花蓮縣	個人申請	1	0
	考試分發	2	1
苗栗縣	轉學考	1	0
雲林縣	個人申請	1	0
	考試分發	2	1
	轉學考	1	0
高雄市	個人申請	6	0
	考試分發	11	1

由表 4-7 可看出，以高雄市地區考試分發 11 人，流失 1 人，學生就讀狀況最為穩定，推估此地區將來可作為招生安排較為頻繁且有效地區；而以就就讀人數與流失人數來看，以嘉義縣個人申請 3 人，流失 2 人，其流失比例最為嚴重；其流失人數與居住地區、入學方式相關。

第二節 決策樹分析

前一節中，我們運用 Python 對多元入學學生的流失人數和許多相關因素的關係分別進行分析和探討，很明顯可以看到流失人數與居住地區、入學方式相關。我們將進一步以資料採礦的決策樹建立流失的預測模型，綜合探討不同因素影響學生流失的重要性，並且預先預測可能流失的學生，以作為改善學生流失率的參考依據。

一、 決策樹的建構

建立決策樹的基本步驟如下：

(一)、 開始，所有記錄看作一個節點。

(二)、 遍歷每個特徵的每一種分裂方式，找到最好的分裂特徵（分裂點）。

(三)、分裂成兩個或多個節點。

(四)、對分裂後的節點分別繼續執行 2-3 步，直到每個節點足夠“純”為止。

如何評估分裂點的好壞？如果一個分裂點可以將當前的所有節點分為兩類，使得每一類都很“純”，也就是同一類的記錄較多，那麼就是一個好分裂點。具體實踐中，到底選擇哪個特徵作為當前分裂特徵，常用的有下面三種演算法：

ID3：使用資訊增益 g(D,A)進行特徵選擇

C4.5：資訊增益率 =g(D,A)/H(A)

CART：基尼係數

一個特徵的資訊增益(或資訊增益率，或基尼係數)越大，表明特徵對樣本的熵的減少能力更強，這個特徵使得資料由不確定性到確定性的能力越強。本研究將以 ID3 演算法建立學生流失的決策樹預測模型。

二、 資料預處理

先將資料先行預處理，處理後所要分析的資料欄位說明如表4-8。

表 4-8 決策樹分析之欄位

屬性	欄名	說明
Key	學號	
Input	性別	男、女
Input	班級	A、B
Input	入學方式	考試分發、個人申請、繁星推薦、轉學考
Input	居住地區	
Input	學業總成績	A、B、C、D、E
PredictOnly	流失	Yes、No

為了符合ID3演算法可直接處理離散型資料型態的特性，以及觀察不同成績等級和流失的相關性，本研究將學業總成績轉換成A到E五大類：A類成績是90分以上，B類的成績是80~89之間，C類是70~79之間，D類是60~69之間，而E類就是小於60分。表4-9是轉換後的資料集。

表 4-9 學生資料表(二)

Gender	Class	Admission	TotalScore	Grade	Dropout	Area	ChArea
Female	B	Exam	75.23	C	No	Kaoh	高雄市
Male	A	apply	63.72	D	No	Ping	屏東縣
Male	A	apply	76.67	C	No	Kaoh	高雄市
Male	A	apply	86.6	B	No	Kaoh	高雄市
Female	A	apply	85.46	B	No	Kaoh	高雄市
Female	A	apply	82.17	B	No	Kaoh	高雄市
Female	A	apply	82.25	B	No	Kaoh	高雄市

Male	A	apply	73.44	C	No	Tain	台南市
Male	A	apply	74.31	C	No	Tain	台南市
Male	A	apply	68.43	D	Yes	Tain	台南市
Male	A	apply	76.99	C	No	Tain	台南市
Male	A	apply	75.32	C	No	Tain	台南市
Male	A	apply	74.54	C	No	Tain	台南市
Male	A	apply	37.37	E	Yes	Tain	台南市
Male	A	apply	74.23	C	No	Yunl	雲林縣
Male	A	apply	79.94	C	No	Chia	嘉義縣
Male	A	apply	63.14	D	Yes	Chia	嘉義縣
Male	A	apply	76.87	C	No	Hual	花蓮縣
Male	A	apply	63.48	D	No	Taoy	桃園市
Male	B	apply	51.66	E	Yes	Chia	嘉義縣
Male	B	apply	72.18	C	Yes	Taic	台中市
Male	B	apply	74.28	C	No	Ping	屏東縣
Male	B	apply	66.08	D	No	Kaoh	高雄市
Male	B	apply	71.05	C	No	Tain	台南市
Male	B	apply	77.65	C	No	Chan	彰化縣
Male	B	apply	63.06	D	No	Hsin	新竹市
Male	B	apply	71.99	C	Yes	Taoy	桃園市
Male	B	apply	80.15	B	No	Taoy	桃園市
Male	A	star	80.31	B	No	Ping	屏東縣
Male	B	Exam	71.09	C	Yes	NewT	新北市
Female	B	Exam	87.58	B	Yes	Taip	台北市
Female	B	Exam	85.89	B	No	Keel	基隆市
Female	B	Exam	79.01	C	No	Taoy	桃園市
Male	B	Exam	64.82	D	No	Taic	台中市
Male	B	Exam	70.82	C	No	Taic	台中市
Male	B	Exam	78.72	C	No	Taic	台中市
Male	B	Exam	82.27	B	No	Taic	台中市
Female	B	Exam	78.28	C	Yes	Chan	彰化縣
Male	B	Exam	83.86	B	No	Chan	彰化縣
Female	B	Exam	78.98	C	Yes	Yunl	雲林縣
Male	B	Exam	78.52	C	No	Tain	台南市
Male	B	Exam	81.47	B	No	Tain	台南市
Male	B	Exam	78.44	C	No	Tain	台南市
Male	B	Exam	81.61	B	No	Tain	台南市
Male	B	Exam	75.03	C	No	Kaoh	高雄市

Male	B	Exam	75.82	C	No	Kaoh	高雄市
Male	B	Exam	78.11	C	Yes	Kaoh	高雄市
Male	B	Exam	67.03	D	No	Kaoh	高雄市
Female	B	Exam	80.54	B	No	Kaoh	高雄市
Male	B	Exam	79.89	C	Yes	Hual	花蓮縣
Male	B	Exam	73.58	C	Yes	NewT	新北市
Male	B	Exam	76.6	C	No	Kaoh	高雄市
Male	B	Exam	75.43	C	Yes	Ping	屏東縣
Male	A	Exam	75.97	C	No	Taip	台北市
Male	A	Exam	59.58	E	No	Chan	彰化縣
Male	A	Exam	73.65	C	No	ChiC	嘉義市
Male	A	Exam	71.95	C	No	Kaoh	高雄市
Male	A	Exam	77.81	C	No	Kaoh	高雄市
Male	A	Exam	79.97	C	No	Kaoh	高雄市
Female	A	Exam	84.29	B	No	Ping	屏東縣
Male	A	Exam	82.89	B	No	Taip	台北市
Male	A	Exam	81.36	B	No	NewT	新北市
Male	A	Exam	76.75	C	No	NewT	新北市
Male	A	Exam	72.49	C	No	NewT	新北市
Male	A	Exam	80.13	B	No	Chan	彰化縣
Male	A	Exam	81.63	B	No	Yunl	雲林縣
Male	A	Exam	89.7	B	Yes	Chia	嘉義縣
Male	A	Exam	77.66	C	No	Tain	台南市
Male	A	Exam	77.69	C	No	Kaoh	高雄市
Male	A	Exam	75.29	C	No	Ping	屏東縣
Male	A	Exam	78.45	C	No	Ping	屏東縣
Male	A	Exam	71.4	C	Yes	Ping	屏東縣
Male	A	Exam	80.97	B	No	Hual	花蓮縣
Male	B	Transfer	76.82	C	No	Miao	苗栗縣
Male	A	Transfer	45.21	E	No	Yunl	雲林縣
Male	A	Transfer	68.21	D	No	Chia	嘉義縣

三、 ID3 決策樹的流失模型與分析：

本研究以 Python 套件 Id3Estimator 來建立預測流失的決策樹

模型。我們以資料集的 80%約 60 筆，作為訓練集(training set)來建立決策樹模型;測試資料集(testing set)是原資料集的 20%約 16 筆，用來評估決策樹模型的預測效能。Python 程式碼如下所示：

```
"""
    Decision Tree: Id3Estimator
"""
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from id3 import Id3Estimator, export_graphviz
# 讀入 Excel 資料檔
df_dm = pd.read_excel(r"D:\研究生\梓豪\DataSet102_DM_e.xlsx",
encoding="utf8")
#print(df_dm.head())
X = df_dm[['Gender', 'Class', 'Admission', 'Grade', 'Area']]
y = df_dm["Dropout"]
xTrain, xTest, yTrain, yTest = train_test_split(X, y, test_size=0.2,
random_state=1)
#print(len(xTrain))
#print(len(xTest))

feature_names = xTrain.columns
input_data = np.array(xTrain.values.tolist())
target = np.array(yTrain.values.tolist())
```

```
estimator = Id3Estimator(min_samples_split=5, gain_ratio=True)
estimator.fit(input_data, target, check_input=True)
export_graphviz(estimator.tree_, "out.dot", feature_names)
estimator.predict(xTest.values.tolist())
```

在 Id3Estimator 建立流失模型之後，如果拿測試集的 16 人做輸入，結果有 14 人的預測結果是準確的，準確率是 87.5%，還算不錯，所以我們觀察其決策樹進行進一步的討論，模型對應的決策樹如圖 4-5 所示。

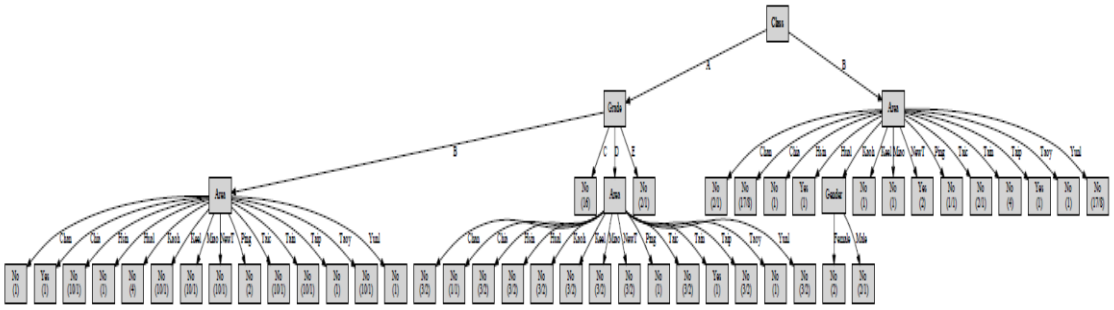

圖 4-5 決策樹分析圖

決策樹的根節點是欄位 Class，分為 A、B 兩班，可以看出 A、B 兩班的流失狀況明顯的不同，如圖 4-6 所示。

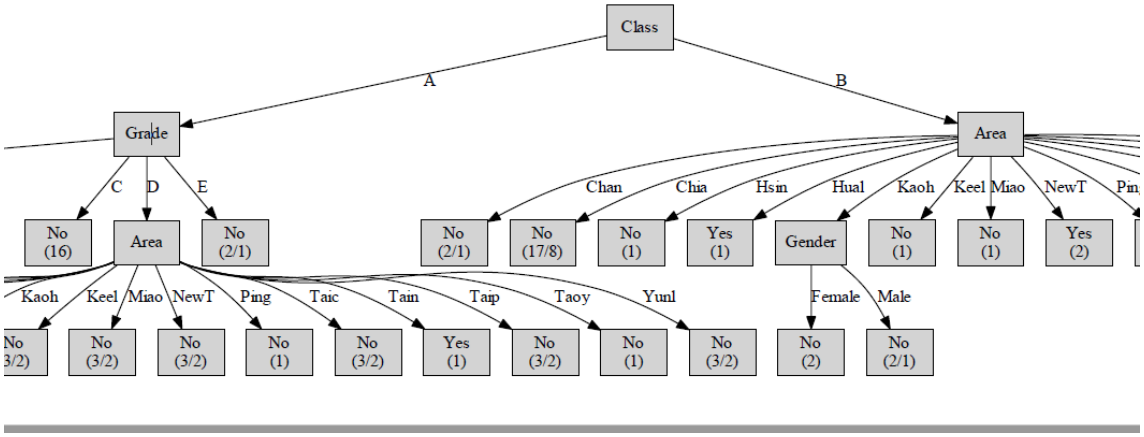

圖 4-6 決策樹根節點的部分擷取圖

由圖 4-6 的 A 班子樹，可看出影響 A 班學生流失率的主要因素是成績等級 Grade，進一步觀察，C 類成績的學生都不會流失，D 類成績學生的流失跟居住地區有關。而 B 類成績學生的流失也跟居住地區有關，但明顯傾向於不會流失，如圖 4-7 所示。

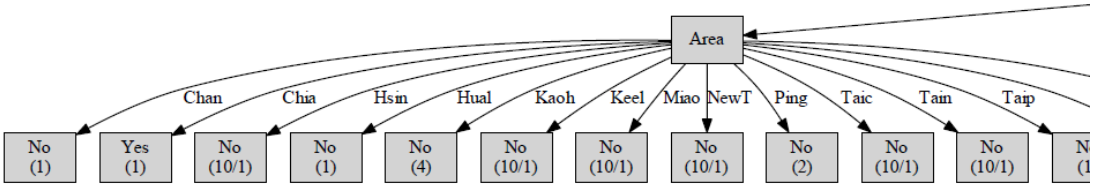

圖 4-7 A 班 B 類成績的決策樹部分擷取圖

由圖 4-6 的 B 班子樹，可看出影響 B 班學生流失率的主要因素是居住地區 Area。台北地區和花蓮的學生都已流失，而高雄市

的流失又和性別有關，女生完全沒有流失，如圖 4-8 和圖 4-9 所示。

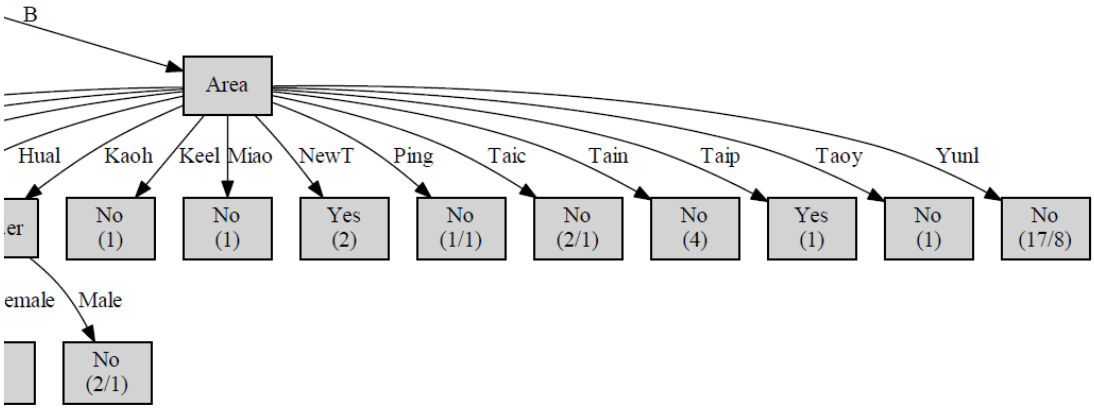

圖 4-8　B 班的決策樹部分擷取圖 (a)

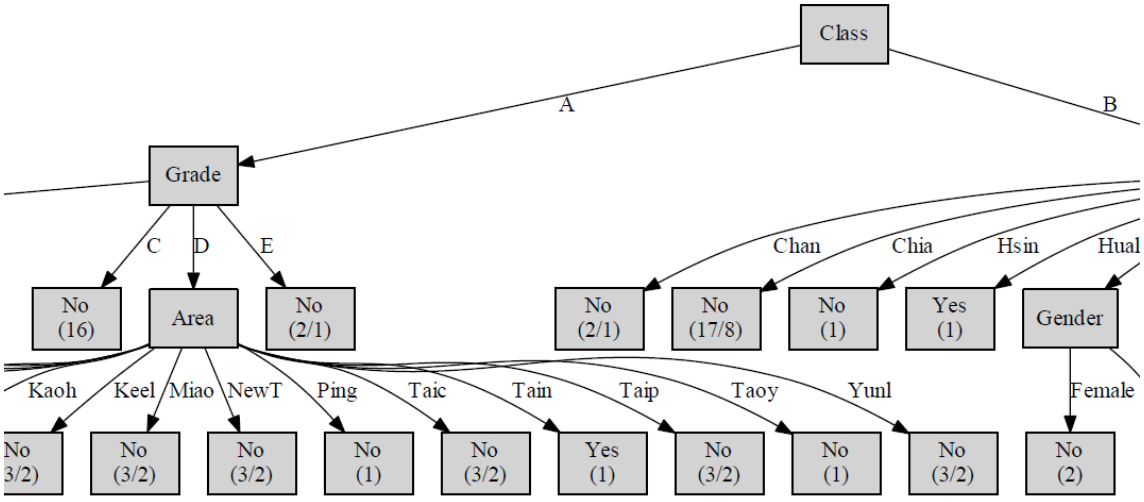

圖 4-9　B 班的決策樹部分擷取圖 (b)

第五章、結論與未來工作

第一節 結論

為了探討學生的流失率與入學方式的相關性，本研究試圖在學的學生資料，找出不同的入學方式學生其學業成績表現與學生流失的關聯，以提供給學校老師於教學過程中注意學生的學習表現；並協助學生的學習過程以預防學生流失。

一、在多元入學學生學業表現上，繁星推薦學生學業表現仍為各入學管道中，學業表現較為突出部分；反觀轉學考進入學校之學生學業表現仍為最弱。

二、在多元入學學生之流失率分析中發現以下狀況：

（一）在就學狀況中，在休學、退學及轉學方面流失率，以轉學 12.1%最高。

（二）在居住地區就讀人數與流失人數部份，以高雄市就讀人數最多，且流失人數比例偏小；反觀於嘉義縣地區雖就讀人數非為最少，但在流失比例上卻是最高，其影響流失率程度與居住地區有相關。

（三）在入學方式與流失人數部份，繁星推薦及轉學考中，學生的流失人數為 0，基本上多為穩定；而在考試分發及個人申請上則流失比例較為偏高，顯示出流失學生多為此兩類學生，其影響流失率程度與入學方式有相關。

（四）在居住地區、入學方式、就讀人數與流失人數部份，以高雄市地區考試分發最多人，流失人數偏低，學生就讀狀況最為穩定；而以嘉義縣個人申請人數及流失人數比例最高，推估其影響流失率程度與居住地區及入學方式有相關。

三、在決策樹之流失模型分析中發現以下狀況：

（一）A、B 兩班的流失狀況明顯的不同。

（二）影響 A 班學生流失率的主要因素是成績等級 Grade，C 類成績的學生都不會流失，D 類成績學生的流失跟居住地區有關。而 B 類成績學生的流失也跟居住地區有關，但明顯傾向於不會流失。

（三）影響 B 班學生流失率的主要因素是居住地區 Area。台北地區和花蓮的學生都已流失，而高雄市的流失又和性別有關，女生完全沒有流失。

第二節 未來工作與建議

一、 對於其他影響學生流失的相關因素如：家庭狀況、經濟狀況、交友情形等本研究未加入考量，建議後續可加入進行相關研究。

二、 限於本研究只取得 102 級入學資管系學生資料，資料範圍

有明顯的不足，建議後續可增加研究樣本數量以更有效的推論與預測影響學生流失率的因素。

三、 針對本研究給研究學校的建議：

（一）對於轉學生而言，剛進入學校環境比較陌生難免有適應不良的地方，校方應隨時注意學生狀況，適時給予學習輔導。

（二）善加利用課業輔導及預警制度，將學業表現較低學生列入輔導，進而提高學生學習興趣，降低學生流失率。

（三）另外有關多元入學與學生流失關係，可以作為招生時，各管道招生員額設定及重點經營地區，進而提高註冊率。

參考文獻

一、 中文部份

1. 行政院研究發展考核委員會西元 2012 年 4 月編印
2. 中華資料採礦協會http://www.cdms.org.tw/xoops2/html/modules/news/
3. 林青翰(民106年)「以資料採礦探討多元入學學生之流失率與學業表現之研究」， 南華大學資訊管理學系碩士班碩士論文。
4. ITREAD(機器學習決策樹演算法實戰) 2018 年 10 月 05 日發表
5. 台部落(一文看懂 Python 主要應用領域或應用場景 2018 年 11 月 12 日發表
6. 許依宸(民 98)，「資料採礦在學生流失偵測上之應用」，南華大學資訊管理學系碩士班碩士論文。
7. 陳芳君（民 101），「以資料採礦探討學生流失及其相關因素之研究」，南華大學資訊管理學系碩士班碩士論文。
8. 盧梅莉，「團體成員流失之探討」，諮商與輔導，76期，34-36頁，民81。
9. 謝邦昌、鄭宇庭、蘇志雄，SQL Server 2008 R2 資料採礦與商業智慧，碁峰資訊，初版，台北市，2011。
10. 吉多、范羅蘇姆「Python 軟體基金會」，維基百科(2019 年 3 月 25 日)第一版釋出於 1991 年

網址:https://zh.wikipedia.org/wiki/Python

11. Travis Oliphant Numeric, 1995 年; NumPy, 2006 年(2019 年 4 月 21 日）維基百科, 網站:www.numpy.org

12. 韋斯麥金尼，出版（2008 年 1 月 11 日）維基百科

網站:pandas.pydata.org。

13. John D.Huter，出版(2019 年 2 月 26 日)維基百科

網站:matplotlib.org。

14. 社群專案，由 Enthought 資助(2019 年 2 月 9 日)維基百科

網站:www.scipy.org

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. Fonell, Clases, and Birger Wernerfelt "Defensive Marketing Strategy by Customer Complaint Management: A Theoretical Analysis", Journal of Marketing Research, 24, pp. 337-346, November 1987.

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

6. Heskett, James L. W.Earl Sasser Jr., and Leonard A. Schlesinger, The Service Profit Chain, Free Press, New York, 1997.

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. Reichheld, Fredirick F. & Sasser, W.E. Jr., Zero Defection: Quality Comes to Services, Harvard Business Review, Vol. 68, No.5, pp. 105~111, 1990.

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

11. Strouse, Karen G, Marketing Telecommunications Services New Approaches for A Changing Environment, Boston:Artech House.1999.