國立政治大學統計系碩士班

碩士論文

指導教授：鄭宇庭、蔡紋琦、謝邦昌博士

導入雲端運算概念
於資料採礦之分類系統

Implement the Concept of the Cloud Computing into the Classification System of Data Mining

研究生：林盈方　撰

中華民國一〇〇年五月

謝　辭

本論文終於完成，首先要感謝的人就是我的指導教授－鄭宇庭老師，不論在論文上、生活上以及學業上等方面，老師都給予大大的幫助，還有爹爹團隊的婉婷、建佑、雨慈和詠翔，同甘共苦了研究所的兩年，也在最需要的時候給予我鼓勵與支持。

再來我要感謝的是我班上的同學們，從研究所開學在九樓的奮鬥，到在研究室中撰寫論文，其中的酸甜苦辣，都深植我心，讓我在政大留下許多回憶與懷念，萬分感激這段期間協助過我的你們。

最後，要感謝的是我的爸爸、媽媽、大姊、二姊以及弟弟，因為你們的包容與支持，才能促使我進步和成長，還有我親愛的朋友們，謝謝你們的傾聽與照顧，我以此謝誌表達我最深的謝意，並與你們分享這份喜悅。

林盈方　謹致

民國一〇〇年六月

摘　要

近幾年來資料採礦及雲端運算的興起，導致許多公司企業紛紛推出有關雲端運算的服務，或利用資料採礦的技術以助於了解客戶行為。而資料採礦的技術不僅是企業所獨享的一個工具，一般非企業的使用者也常常會面臨到決策問題，為了讓一般使用者能夠方便取得軟體工具以及節省時間成本，本研究以雲端運算為概念，利用 RExcel 軟體和 Excel VBA 程式語言為研究工具，發展出一個資料採礦分類雲端運算系統。

本研究將欲分類的目標變數分為三種型態：數字連續型、數字類別型以及文字類別型，此分類系統會依照目標變數型態的不同，而採取不同的分類模型來分析使用者之資料，並分別以三個資料檔為例，上傳至此資料採礦之分類系統進行分析後，其分析結果報表將以網頁預覽的方式呈現給使用者，使用者可以針對連續型目標變數的資料分析結果，利用 MAPE 值評估分類模型之優劣，而類別型目標變數的資料分析結果，則可以正確率來評估分類模型之優劣。

使用者可透過簡易步驟來操作此系統，並選擇可解釋資料之最佳模型，也可從結果報表中獲取資料之特性，更進一步地可以進行所需的決策。

關鍵字：雲端運算、資料採礦、分類模型

Abstract

In resent years, the rise of data mining and cloud computing has led many enterprises have been offering services related to cloud computing, or using data mining techniques to understand customer behaviors. Data mining is a tool not only for enterprises, but also for general non-business users who often face making decisions. In order to enable general users to easily assess the software and save time and costs, this study proposes a classification system of data mining constructed by RExcel and Excel VBA, which is based on cloud computing.

In this study, the target variable is divided into three types: digital continuous, digital categorical and literal categorical. The classification system is in accordance with the different types of target variables, taking different classification models to analyze user's data. Taking three data as examples, respectively, uploading them to the system, then the analysis results will be present to the user in the way of page preview. The user can use MAPE values to evaluate classification models with regard to the results of the data for the continuous target variable, and use correct rate to evaluate classification models with regard to the results of the data for the categorical target variable.

Users can take simple steps to operate the system, select the best model which can explain the data, and obtain the characteristics of the data from the result reports, further to the necessary decision-making.

Keyword: cloud computing, data mining, classification models

目　錄

表　次

圖　次

第壹章　緒論

第一節 研究背景與動機

現今科技的進步日新月異，資訊發達得相當快速，網路已不再只是應用於訊息的傳遞和資料的傳播而已，資源的共享才是現在大家在網路上追求的目標，懂得利用資源的人就更能贏在起跑點，對於工具要求更快的速度、更簡易的操作、更貼近人們的需求是促使科技不斷進步的一股強大動力，就像是某手機品牌所講的：科技始終來自於人性。現今的公司企業擁有越來越多客戶的個人資料和相關資訊，以及公司內部的資料，形成規模相當龐大的資料庫，這造成公司企業面臨如何降低資料遺失風險和減少資料傳輸成本的問題，亦面臨了如何處理這些大量的資料，為了解決以上問題，公司企業漸漸傾向採用集中化的方式來管理龐大的資料庫，且需要借助平行運算、分散式運算以及多核心程式功能來處理資料，使得高速度計算成為企業界的新寵，因此雲端運算越來越受到公司企業的重視，各大企業也紛紛投入相當多的資源開發出有利自家企業的雲端技術，目的就是為了在這場「雲」的戰爭中搶得先機，以便在未來市場中佔領一席之地。隨著網際網路的發展，網頁的標準越來越開放，且不易受到阻擋，瀏覽器亦成為跨平台的載具，於是許多開發業者將網頁變成開發平台，網路服務供應商因此順勢搭上了雲端運算的列車，推出雲端運算服務，架設雲端服務的平台，如全球 CRM（客戶關係管理）軟體公司 Salesforce 透過網際網路，架設了一個提供按需定製客戶關係管理服務的網站 Slesforce.com（http://www.salesforce.com/tw/），提供了銷售管理應用程式以及建構可自行開發應用程式的平台，以供公司企業使用。

而近幾來年，公司企業漸漸趨向以客戶為導向的形態，了解客戶行為則成為一個公司企業相當重視的課題，因為我們必須從客戶的行為中去找尋線索，找出客戶的需求，因應客戶的需求提供服務或商品，因此了解客戶行為後才能進行必要且正確的決策，就是所謂的客製化服務。但當公司企業面對資料爆炸但是資訊

貧乏時，卻不知道如何從其中找尋出有用的潛在的訊息，以助於了解客戶行為，而如何從中找出這些隱藏的資訊則有賴於資料採礦的技術。資料採礦簡單來說就是理解資料與進行的工作來獲取相關知識與技術（Acquisition），以整合與查核資料（Integration and Checking），再去除錯誤或不一致的資料（Data Cleaning）後發展模式與假設（Model and Hypothesis Development），最後測試與檢驗其資料（Testing and Verification），即可解釋與使用資料（Interpretation and Use）。但資料採礦的軟體工具有眾多選擇，這些軟體需要公司企業花費昂貴的成本購買，且需要人員進行軟體的管理和維護工作，這些也需要人力和時間成本，若能將這些軟體透過網路，建立在雲端平台上，不但能使得公司企業方便取得，也能節省各種成本。

資料採礦的技術不僅是企業所獨享的一個工具，一般非企業的使用者也常常會面臨到決策問題，需要在自身擁有的資料中，利用資料採礦的技術取得一些知識，而在面對需要分析龐大的資料時，使用者可能會遇到的問題有：一般使用者無法負荷昂貴的軟體費用，因而無法取得工具來分析資料，又或者使用者需要尋找可下載軟體的載點，下載之後又必需花費時間安裝在自己的個人電腦上；由於資料量過大，若在個人電腦上操作，可能會因運算速度慢，操作時間冗長，而導效分析效率差。基於以上兩點原因，為了讓一般使用者能夠方便取得軟體工具以及節省時間成本，因而促使了本研究動機的形成。

第二節　研究目的

本研究以 RExcel 軟體和 Excel VBA 程式語言為研究工具，發展出一個資料採礦分類系統，提供一般使用者一個方便且良好的環境，以進行使用者所需要的資料採礦 。本研究以資料採礦的功能之一－分類為主要的研究目的，其中分類

的模型裡，置入了決策樹（Decision Tree，DT）、支持向量機（Support Vector Machine，SVM）、判別分析（Discriminant Analysis）及隨機森林（Random Forest）等分類模型。

本研究以雲端運算為概念，期許未來網路以及強大的伺服器，將此資料採礦之分類系統架設成一雲端系統平台（圖 1-1），透過此雲端系統平台，使用者不需自行購買、下載或安裝軟體工具，也不需要自行攜帶軟體，只要有可聯絡網際網路的介面，將要分類的資料上傳至此平台上，透過虛擬化的軟體網路介面操作此系統，選擇想要進行分類的目標變數（Target Variable）及解釋變數（Explanatory Variable），經由這些簡單的操作，使用者即使不熟悉某一特定的程式語言，便可得到想要的分析模型，更進一步地可以進行所需的決策。

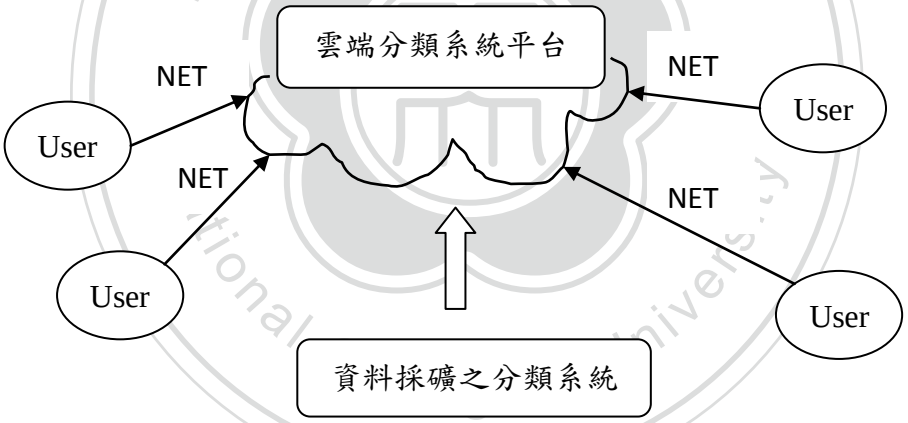


圖1-1　雲端分類系統平台之概念

第三節　研究架構

本研究的架構分成以下五個章節，其安排如下：第一章為緒論，描述本研究的背景與研究動機，經由研究動機所引發本研究之目的；第二章為文獻探討，對雲端運算、資料採礦以及雲端運算在資料採礦上之應用的相關文獻進行探討；第

三章為研究方法，介紹本研究所使用的工具以及預測模型方法；第四章為實證分析，利用資料進行示範，並呈現其分類結果報表；第五章為結論與建議。

本研究的架構如圖 1-2 所示：

圖1-2　論文架構圖

第貳章　文獻探討

本章共分為三節，第一節為「雲端運算概述」，將說明何謂雲端運算、雲端運算的演化過程、服務的產業類型及其應用；第二節為「資料採礦概述」，將概述資料採礦的定義、功能、模型建構之流程及其應用；第三節為「雲端運算在資料採礦上的相關應用」，將整理介紹相關的應用。

第一節 雲端運算的概述

一、　雲端運算的定義

雲端運算（Cloud Computing）其實代表的是一種概念，基於透過網際網路的運算方式，為企業或者個人使用者提供所需的服務。簡單來說，雲端運算即為透過網路，讓眾多不同的電腦同時為使用者處理問題，大幅提昇了處理速度和效率，而所有資源都來自於雲端，使用者只需一個可連端雲端設備和界面即可使用此服務。而雲（Cloud）即代表了網際網路，擁有規模龐大的運算功能，服務供應商將各種資源和軟體傳送到遠端伺服器上，供給使用者使用，而使用者透過網路可隨時取得資源，卻不知資源確切的所在，就像雲一樣看得見卻摸不到。

雲的類型有三種：公有雲（Public Cloud）、私有雲（Private Cloud）和混合雲（Hybrid Cloud）。公有雲亦稱外部雲（External Cloud），指具有公用服務（如水、電、天然氣、瓦斯等服務）的雲技術，是由第三方在網際網路上所提供的一項服務，公開服務所需要的使用者，並按照使用者的選擇來計費；私有雲又稱內部雲（Internal Cloud），為單一客戶單獨或是一個企業組織內部自行取用的雲端運算技術，由該使用者或企業組織自行購買、擁有、維護與管理，提供對資料、安全性及隱私性和服務品質的最有效控制;混合雲為使用者或企業組織同時使用公有雲和私有雲混合組成的雲端運算技術，因此，利用此技術可以同時擁有二種技術的優點。

美國家標準與技術局（National Institute of Standards and Technology, NIST）定義雲端運算是一種無所不在、隨需供給且方便的網路，擁有廣泛的運算資源，如網絡、伺服器、儲存、應用程式、服務等，這些資源可透過最少量的管理工作及不需與服務供應商的互動，即可快速提供各項服務給使用者，另外，NIST 亦定義了雲端運算的基本特性有隨需應變自助服務（On-demand Self-service）、廣泛網路使用（Broad Network Access）、資源彙整（Resource Pooling）、高度彈性（Rapid Elasticity）和計量服務（Measured Service）等五項，分別簡述如下：

1、 隨需應變自助服務（On-demand Self-service）：消費者在其需要時可自行使用雲端服務，如網路存取，而不需要與雲端服務供應商互動。

2、 廣泛網路使用（Broad Network Access）：由於網路使用無所不在，雲端服務供應商的服務可隨時在網路取用，且使用者所使用的平台無論為何（如手機或PDA），均可透過標準機制使用網路。

3、 資源彙整（Resource Pooling）：依據消費者的需求，雲端服務供應商透過多重租賃模式服務消費者，指派或重新指派實體及虛擬資源，而消費者通常不知道雲端服務供應商提供的所有資源之確切位置，只可能掌握國家、州或資料中心等大範圍的區域地點。這些資源包含如存貯、處理、記憶體、網路頻寬和虛擬機器等。

4、 高度彈性（Rapid Elasticity）：運算能力可以迅速且具有高度彈性的提供給消費者，彈性亦能因應要求調整資源規模大小，對消費者而言，雲端似乎無窮無盡，且能依據其需求增減運算能力採購額。

5、 計量服務（Measured Service）：雲端服務各層次均由雲端服務供應商掌控與監管，這對於計費、存取控制、資源優化、處理能力規劃及其他工作相當重要。

二、　雲端運算的演化

雲端運算並不是一蹴發展而成的，而是經由超級電腦（Super Computer），漸漸發展至各種運算而促使今日的雲端運算的崛起，以下為雲端運算的演化過程：

（一）超級電腦（Super Computer）

超級電腦為一種主機電腦，擁有最快的速度且儲存力最強，進行的運算速度最高可達一般個人電腦的十萬倍。超級電腦的機身，往往不是一個，而是由一群電腦所組成。超級電腦可利用來開發新產品和檢驗產品，亦可用來進行大規模的試驗，計算及研究。各國甚至各大型企業例如：Google、IBM 等都在積極研發或添購更強大運算速度更快的超級電腦。

美國 Discovery 頻道公布了 2010 年世界之最，結果中國的超級電腦「天河一號」（Tianhe-1A）被評為是去年速度最快的超級電腦，超越了美國的美洲豹（Jaguar）。

TOP 500 SUPERCOMPUTER SITES（http://www.top500.org/）是一個定期會公布目前名列全世界前 500 名的超級電腦排名的網站。以下為此網站於 2010 年 11 月所公佈的前十名（見表 2-1）：

表2-1　2010 年 11 月超級電腦前十名

Rank	Nation	Site	Computer
1	China	National Supercomputing Center in Tianjin	Tinahe-1A-NUDT MPP, X5670 2.93Ghz 6C, NVIDIA GPU, FT-1000 8C(NUDT)
2	United States	DOE/SC/Oak Ridge National Laboratory	Jaguar-Cray XT5-HE Opteron 6-core 2.6 GHz(Cray Inc.)
3	China	National Supercomputing Center in Shenzhen(NSCS)	X5650, NVidia Tesla C2050 GPU (Dawning)
4	Japan	GSIC Center, Tokyo Institute of Technology	TSUBAME 2.0-HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows(NEC/HP)
5	United States	DOE/SC/LBNL/NERSC	Hopper-Cray XE6 12-core 2.1 GHz (Cray Inc.)
6	France	Commissariat a 1'Energie Atomique(CEA)	Tera-100-Bull bullx super-node S6010/S6030 (Bull SA)
7	United States	DOE/NNSA/LANL	Roadrunner-Blade Center QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz/Opteron DC 1.8 GHz, Voltaire Infiniband(IBM)
8	United States	National Institute for Computational Sciences/University of Tennessee	Kraken XT5-Cray XT5-HE Opteron 6-core 2.6 GHz(Cray Inc.)
9	Germany	Forschungszentrum Juelich(FZJ)	JUGENE-Blue Gene/P Solution(IBM)
10	United States	DOE/NNSA/LANL/SNL	Cielo-Cray XE6 8-core 2.4 GHz(Cray Inc.)

（二）叢集運算（Cluster Computing）

叢集運算是將多部個人電腦以高速的區域網路連結起來，使其可達到超級電腦的高效能及共同處理程序的運算。叢集運算的優點在於具有高效能運算，可降低運算成本，且其擴充性佳，而缺點在於管理困難，要有良好的演算法，才能將工作妥散的分配到各個電腦上運作，且在運作時，需要完全相同規格的硬體及環

境，否則會很難進行（李開文，2008），以及若當叢集運算資料的結構損毀時，可能造成整個系統的癱瘓。

（三）分散式運算（Distributed Computing）

分散式運算由網路連結個人電腦所形成的運算，且在網絡中的任一部電腦都可以在同一個時間，將程式在任何另一部電腦上運算。此運算先將大型工作區分成小型工作區後，再分別由眾多電腦各自進行運算之後再彙整結果，來完成單一電腦無法勝任的工作。分散式運算的優點在於擴充性佳，任何使用者加入此系統都可享有此系統的資源。

（四）格網運算（Grid Computing）

在 1998 年，Foster 與 Kesselman 發展了「格網」的全新概念，指以公開的基準處理分散在各處的資料，亦將其形容為像電力或水力一樣，想要用的時候打開即可得到。格網運算為分散式運算的延伸，也是一種擴充叢集運算的技術，將各種不同平台，不同架構，不同等級的獨立電腦，透過分散式平行處理的方式，做整合的運用。

（五）公用運算（Utility Computing）

公用運算主要是提倡一種理想的資訊架構，把風險從使用者本身轉移至服務供應商。在此架構下，公用運算的訂價模式為採取「用多少付多少」的方式，將運算功能視為如水、電、天然氣等公用設施（Utility）一樣可隨時供使用者需求來提供此服務（On Demand Services），這些服務包含自動提供可計算、度量的 IT 資源，如服務器、存儲容量、商應用程序及網源等，且依照使用者的使用量來計算費用。

（六）雲端運算（Cloud Computing）

透過網路將龐大的運算處理程序拆成無數個較小的子程序，再交由數個伺服器所組成的龐大系統，經由搜尋、運算分析後，再將處理結果回傳給使用者，提昇網路服務處理能力，且可以更有效地共享資料。

三、 雲端運算的服務產業類型

雲端運算的服務產業類型有三種，分別為 SaaS（Software as a Service）、PaaS（Platform as a Service）及 IaaS（Infrastructure as a Service）三種，說明如下：

（一）SaaS（Software as a Service）

SaaS 為一種服務型的軟體。使用者在需要的時候，下載所需功能且安裝在電腦裡使用，或直接在網路上使用線上的軟體。有了服務型軟體，使用者可不用事先購買軟體，只在需要時付費使用，且不需要管理及維護軟體，操作也很簡單，但使用者無法對其軟體進行任何的調整，只能在外觀或者作業的設定做些微的改變，且並非所有的應用軟體都適合透過此服務來提供使用者使用。

現有的一些服務型軟體應用如 Google Map, Yahoo Mail Service 等網路信箱都是 SaaS 的產品，目前提供服務型軟體的代表企業有美國的 Salesforce 公司，SAP, ORACLE 等大軟體廠商也陸續開始提供 SaaS 型的服務。

（二）PaaS（Platform as a Service）

PaaS 為一種服務型的虛擬主機平台。是服務型軟體（SaaS）衍生出來的一種服務型態，為提供平台給系統管理人員和開發人員，用來設計、開發、測試、代管及部署制定應用程式，使用者不需自行建置行軟體主機等平台，可直接透過網路，利用提供 PaaS 服務業者的平台，能夠降低主機的維護和管理系統的成本。透過此服務，開發者可以開發新軟體並快速部署上線的時間，而現有的服務型主

機平台有 Google App Engine, AWS S3, Microsoft Azure, Yahoo Application Platform 等。

（三）IaaS（Infrastructure as a Service）

IaaS 為一種服務型的基礎設施。一開始被稱為 Haas（Hardware as a Service），後來為了作明確的區分而改稱為 IaaS。IaaS 提供了核心計算資源和網路架構的服務，亦提供了伺服器，網路設施，記憶體，儲存硬體，CPU 和資料中心設施等 IT 硬體環境，解決了傳統機房需要的硬體、軟體、儲存、電力及頻寬成本，可使公司企業更用效率的取得資源。而目前存有 IBM Blue Cloud, HP Flexible Computing Services 等服務型基礎設施產品。

雲端運算具有超大規模、高通用性、虛擬化、使用者付費、成本低、高可靠度等優點，基於擁有虛擬化技術可快速部署資源或獲得服務，且擴充套件具有相當大的彈性，又為使用者透過網際網路按需要提供資源，可高速度地處理大量資訊，使得使用者可以方便地參與其中，利用這些良好的特性及功能，雲端運算不僅是為使用者來提供其所需的服務而已，對企業來說，它能有效地降低風險及成本，以增加企業商機。於是，許多企業系統供應業紛紛投入雲端服務領域中，如 Google 應用服務引擎 Google App Engine（GAE）在 2008 年問世，是 Google 的應用程式開發與代管平台，可讓開發者提供 Python 程式碼，自行在平台上建構高流量的網路應用程式，不需管理高流量的基礎架構。而 GAE 也成功了擄獲了全球各行各業各種規模的公司企業的心，它不僅使公司企業的時間與成本降低了許多，並也改善了企業間合作的方式。另外，連相當耗費 CPU 運算的影像編輯軟體，也有服務供應商嘗試將其做成雲端服務，如 Adobe　Photoshop Express。表 2-2 為各家公司所提供的雲端運算服務的比較。

表2-2　各家公司所提供的雲端運算服務的比較

	微軟	Google	Yahoo	Amazon
平台	Windows Azure	Google App Engine	Yahoo Application Platform	Amazon EC2
技術特性	整合不同裝置與網路服務	儲存與運算的水平擴充能力	儲存與運算的水平擴充能力	可彈性配置的通用虛擬機器
核心技術	Window Server 2008 與 Hypervisor 虛擬化技術	平行分散技術 MapReduce、BigTable 資料庫系統、GFS 檔案系統	平行分散技術 Hadoop、MapReduce、Hbase 資料庫、HDFS 檔案系統	Xen 虛擬化技術
企業服務	Azure（pre-beta）Live Mesh	應用代管服務 GoogleAppEngine，每月低於 500 萬瀏覽次的網站可免費代管，可使用 500MB 儲存空間。	YAP、SearchMonke，使用 Y!OS API 的應用程式，可免費代管。	EC2，提供不同規格的虛擬機器供企業租用，但有規格上限。可動態新增多個虛擬機器分擔服務。
已支援的開發語言	.NET 語言（IIS 7 支援語言）	Web Python，未來會支援更多語言	PHP	企業可自行建置不同作業系統和平台的執行環境
已支援的資料庫系統	SQL Service，如資料表、檔案等。	BigTable 資料庫系統	HBase 資料庫系統	提供 S3 儲存服務，企業可自行建置所需資料庫系統
開源程度	開放 API	公開設計架構，程式碼未開源	完全開源	完全開源
計價方式	將按資源與服務等級（SLA）計價，細節未公布	按使用的處理器時間、儲存空間與網路流量計價	尚未公布	按使用的處理器時間、儲存空間與網路流量計價，也新增服務等級計價方式
資料來源：iThome 網站				

雲端運算雖然處處充滿商機，但仍然面臨到一些疑慮，由於使用者所需的資料可以從雲端上獲得，但能夠在需要時，是否能保證一定得到的問題，以及使用者本身的資訊會不會經由網路而洩露出去等等的個資安全性問題，另外駭客入侵

問題也是值得考量的，一些惡意攻擊都已顯示網頁及應用程式安全的重要性，且如何讓資源能夠互通有無等等，亦為雲端運算平台未來發展需要考量到的問題。

第二節 資料採礦概述

由於網際網路資訊發達、關聯式資料庫的廣泛應用和資料整合的技術越來越成熟，以及統計學、人工智慧和機器學習等理論的發展，因而促成資料採礦領域的蓬勃發展。然而在現今的社會裡，電腦的普及化，使得企業漸漸也產生了電腦化的現象，利用電腦紀錄每個顧客的消費行為，因此也累積了大量的交易資料，形成一個規模相當大的資料庫。而企業從原本以產品為導向的觀念，轉變成以顧客為導向，所以了解顧客以往的消費行為因此而變得相當的重要，且需要一種可建立起企業與顧客關係的技術，而這技術則有賴於資料採礦。

一、 資料採礦的定義

Frawley（1991）等人認為資料採礦是從資料庫中挖掘出不明確、前所未知以及潛在有用的資訊過程。Fayyad（1996）等人認為資料採礦是指由已存在的資料中挖掘出新的事實及發現專家尚且不知的新關係。雖然資料採礦的定義眾說紛云，但大致上資料採礦就是指找尋資料中所隱藏的資訊，如趨勢（Trend）、特徵（Pattern）及相關性（Relationship）的過程，亦視為資料庫知識發掘（Knowledge Discovery in Database，KDD）其中的一部分，其為在資料採礦上的應用極為重要的影響，只有資料庫知識發掘才能確保資料採礦得到有意義的結果。根據Fayyad（1996）等人對資料庫知識發掘的定義為：「KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data」，他們認為在得到知識之前，原始資料必須經過五個步驟的處理，其流程圖（見圖 2-1）及步驟如下：

1、 Selection：了解工作並選擇所需的資料

2、 Pre-processing：將所需要的資料做前置作業，刪減不必要的資料

3、 Transformation：資料轉換或簡化工作

4、 Data Mining：利用資料的趨式，採取模型進行預測、分類或推估。

5、 Interpretation/Evaluation：解釋與評估資料

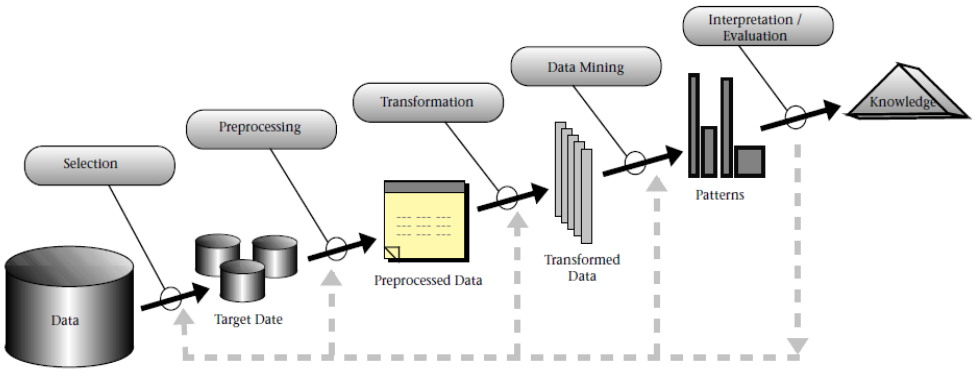


圖2-1　資料庫知識發掘之流程圖

二、　資料採礦的步驟

資料採礦是在資料庫知識發掘流程的其中一個步驟，卻也是相當重要的一個步驟，隨著不同領域的不同問題需求，資料採礦的過程也會不同，分析人員所採用的資料採礦技術也會因資料特性而有所差異，而資料採礦完整的步驟如下：

1、 理解資料與進行的工作

2、 獲取相關知識與技術（Acquisition）

3、 整合與查核資料（Integration and Checking）

4、 去除錯誤或不一致的資料（Data Cleaning）

5、 發展模式與假設（Model and Hypothesis Development）

6、 實際資料採礦工作

7、 測試與檢驗其資料（Testing and Verification）

8、 解釋與使用資料（Interpretation and Use）

因此，資料採礦涉及了大量的準備工作和複雜的過程，而資料採礦的流程有許多種，使用者最常使用的流程為 CRISP-DM （CRoss-Industry Standard Process for Data Mining)，此流程是 SPSS 和 NCR 在 1996 年時訂出的一套資料採礦標準程序，CRISP-DM 模型建構步驟及流程圖如圖 2-2：

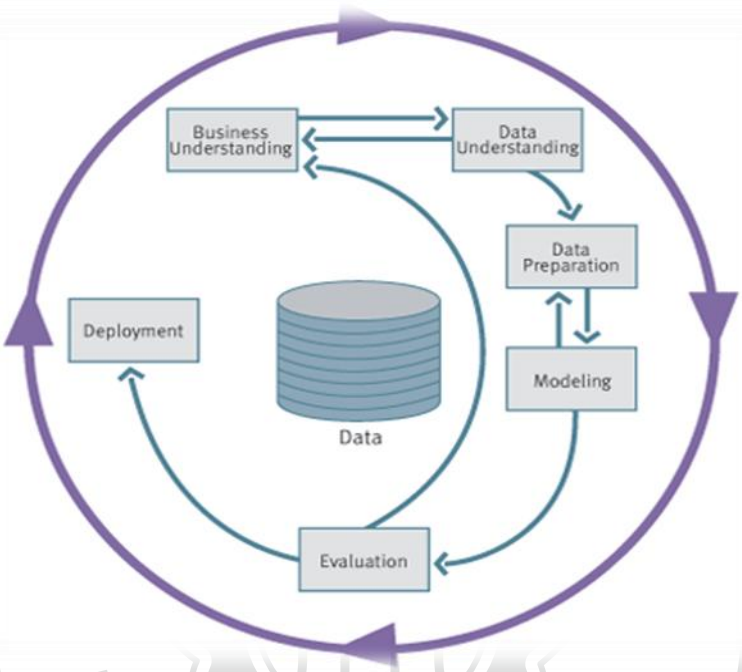


圖2-2　CRISP-DM 模型建構流程圖

（資料來源：http://www.crisp-dm.org/）

CRISP-DM 將進行一個資料採礦專案分為六大階段，簡述如下：

1、 定義商業問題（Business Understanding）

要想充分發揮資料採礦發現的知識價值，必須要先對問題有一個清晰明確的定義，有了明確的問題定義，最後才能得到結果進行衡量的標準。因此，在初始階段著重於了解商業問題，從商業角度化的問題轉化為資料採礦問題，以符合資料採礦流程，並初步計劃目標。

2、 定義分析資料（Data Understanding）

收集可用分析的完整資料，熟悉資料，並利用簡單的統計分析方法及統計軟

體作初步的分析，包括驗證此資料邏輯及資料品質，以及對商業問題設立前提假設。

3、 資料準備（Data Preparation）

這是建立模型之前的最後一步資料準備工作。將原始資料加工成最後的資料，以用在資料採礦過程的資料表，準備工作可能要重複執行數次，是最耗時又費力的部分。此階段包含了資料選擇（Data Selection）、資料清理（Data Cleaning）、資料擴充（Enrichment）及資料編碼（Data Coding）。

4、 建立模型（Modeling）

建立模型是一個反覆的過程。由重複檢測的過程中，可找出模型的最佳設定參數值，以達最佳預測效果。通常模型的設定對輸入資料的格式或型態有特定的限制，因此，回到資料準備階段往往是必要的。

5、 評估（Evaluation）

評估模型可能帶來的效益並解釋此模型的價值，確定結果是否有與商業目標結合，以達到資料採礦的效益，並謹慎的檢查執行的過程，確認是否在過程中有遺漏的部分，若沒有，才能進入最後的應用。

6、 部署（Deployment）

建構模型的結果經驗證後，得到的是此模型所帶來的專業知識，將這些知識應用到其它資料上，使一般使用者可得以運用。而事物的變遷相當的快速，隨著時間的增加，資料可能須要作更新的動作，且此模型可能只適用於一段時期，因此，要不斷測試模型，做必要的修正或重建。

三、 資料採礦的功能

資料採礦的功能可包含分類（Classification）、推估（Estimation）、預測（Prediction）、關聯分組（Affinity Grouping）及同質分組（Clustering）等五項功能，簡述如下：

1、 分類（Classification）

分類是找出新事物特性，然後判斷該事物與現存集群何者比較類似，再將其歸類到該集群；分類的主要工作就是對現有集群的特性加以定義，並利用一些統計分析技巧來建立判別的準則，並利用該準則將尚未瞭解的資料加以分類。常用的方法有決策樹（Decision Tree）以及記憶基礎推理（Memory-Based Reasoning）等。分類問題的應用如顧客的信用風測預測。

2、 推估（Estimation）

推估是依據現有的連續性資料，來估計未知屬性；在實務上的運用大多與分類功能結合運用；常用的方法有相關分析、迴歸分析（Regression Analysis）與類神經網路（Neural Network）等。推估問題的應用如商品價格的趨勢變化。

3、 預測（Prediction）

預測依據現有資料進行推測，估計未來的趨勢及數據，不論是分類、推估或預測都是利用現有資料來推測分類，而現有資料則是很好的資料來源，我們利用過去的數值來建立估計未來數值的模型。常用的方法有迴歸分析（Regression Analysis）、時間數列分析（Time Series Analysis）與類神經網路（Neural Network）等。

4、 關聯分組（Affinity Grouping）

找出彼此之間有相關聯的產品，將這些相關聯的物件放在一起。其應用如由消費者的購買行為特性，利用產品交叉銷售的方法，分析此消費者與產品關連性的強弱，藉此設計出吸引消費族群的產品組合。

5、 同質分組（Clustering）

將同質性較高的物件歸為同一類，並依照其特性加以分類，其目的是找出組間的差異，對各個組內的物件再進行挑選，此功能有助於判斷單一事物有所改變時帶來的影響。若要歸類的群並沒有加以定義，是根據資料的特性自動區隔，再由專業人員判斷分群，因同質分組相當於行銷術語中的區隔化（Segmentation），

常用的方法有 K-means Method、集群分析與判別分析等。其應用如產品自動化推薦

資料採礦結合了統計學、人工智慧、資料庫及領域知識（Domain Know-how）等技術，研究人員認為運用這些技術可以結合機器學習和演算法提昇資料庫的使用，許多產業界人士則將此領域視為可增加企業潛能的重要指標，如客戶、市場、未來趨勢等，以提供企業進行決策，提昇決策品質，進而有效地增加企業競爭優勢。資料採礦已在各個行業都有廣泛的應用，如生物科技、服務業、製造業、金融保險業、資訊電子業、醫療業等，也有不少公司企業利用此技術獲得成功控管內部的案例，如席內銀行（Signet Banking Corporation）從不同來源獲取顧客的行為資料並建立預測模型，利用預測模型的結果來推展轉帳卡業務，並獲得極大的成功。1994 年，雖然席內銀行的發卡部門被全球十大發卡公司之一的美國第一資本金融公司 Capital One 所併購，但 Capital One 同樣透過資料採礦的技術，利用顧客智慧（Consumer Intelligence）及 CRM 系統來進行顧客分群，辨別出履約風險程度高低的顧客，若是信用優良的顧客，將會給予較優惠的循環利息，但若是會帶來負面價值的顧客，公司則會採取婉轉的方式，請顧客轉移到別的公司，這樣的做法，使得 Capital One 快速發展成為擁有全方位金融服務供應商以及成功地控制貸款流失率。

由於現今的企業越來越需要了解顧客的需求和顧客的行為，以便於制定良好的決策方案，增進企業的商機，因而將資料採礦的技術導入企業。而企業將此技術運用到各個不同的領域，如行銷、市場行銷、客戶行為分析、金融市場分析等。資料採礦在各領域的應用整理如表 2-3：

表2-3　資料採礦在各領域的應用

領域類別	應　　用
金融業	直效行銷、股匯市行情預測
醫療業	預防醫學分析、臨床病徵分析、院內感染分析
教育業	學生來源分析、課程規劃、學習評量、適性化教學
通訊業	通訊品質偵測、定位應用服務、顧客購買傾向分析、顧客價值分析
生物科技	、基因圖譜比對、基因定序、演化分析
保險業	保險潛在客戶名單分析、偵測保險詐欺
信用卡公司	分析持信用卡者的購買行為、信用評等、偵測信用卡詐騙行為
航空業	顧客需求行為
零售業	購物籃分析、偵測收銀員詐騙行為、顧客購買行為

第三節 相關應用之文獻探討

國內公司企業或組織於雲端運算或資料採礦的相關文獻整理如下：

一、　中華電信

台灣電信龍頭中華電信近年來一直積極跨足雲端市場，除了在雲端運算與雲端儲存兩大部分均已投入高資本的完整建置，提供客戶最強大、高容量的雲端儲存服務，以及最完整的網路與資安雲端防護外，在 2010 年 7 月 5 日時，與全球最大筆記型電腦製造代工業者廣達電腦簽立雲端運算合作備忘錄（MOU），結合各自的資源和技術，共同創造雲端服務商機。中華電信數據通信分公司總經理陳祥義（2010）表示，今年中華電信將積極投入「智慧造雲計畫」，在生活方面打造智慧家庭雲，在商用方面打造智慧商店雲，喊出今年智慧家庭用戶 1,200 戶、智慧商店用戶 6,000 戶的目標。

所以中華電信於今年初時推出「智慧造雲計畫」，即推出家庭生活用的「智慧家庭雲」及商業用的「智慧商店雲」二大智慧雲，希望藉由這二種智慧雲端服務，讓客戶在生活上就像在雲上面一樣地無憂無慮。「智慧家庭雲」為創新應用服務，應用智慧科技結合光纖與感知網路，建構全新的智慧家庭系統平台連結社區與居家之管理，提供通訊整合、門禁對講、中央監控，以及居家安防與通報、智能控制、生活資訊及節能統計等智慧生活服務，透過網際網路將社區內住戶的環境控制系統整合，其中包含門禁、監控、家電控制、燈光控制、瓦斯防災及緊急警告防護等項目。而「智慧商店雲」透過雲端技術針對連鎖業及實體店家推出智慧化管理方案，讓店家全方位整合實體與線上通路，經營商店更簡單、方便，並開創商店智慧化管理新時代。以企業上網為基礎，可提供店家即時影像監控功能，讓店主隨時掌握店中實況；為店家打造合法、安心、專業的音樂公播環境；協助商店智慧化管理用電，降低支出並響應節能環保，為店家量身打造最舒適的店面環境。中華電信數據通信分公司副總經理鍾福貴（2010）表示：「雲端服務已經不是潮流趨勢，而是現在進行式。HiNet 一直以提供客戶最領先、最優質的服務為己任，在雲端服務這塊，是今年度首要業務重點，希望能夠在家庭、商用等全方位提供最完整的智慧雲端服務。」

二、 B&Q 特力屋

B&Q 特力屋是國內大型連鎖居家修繕賣場，屬於零售業領域。為了提昇內部員工對公司資訊的了解，B&Q 特力屋於企業入口網站設立「企業員工資源網」，是公司的每一位員工都可登入的虛擬化平台，透過這個平台，員工可以從中取得公司工作流程以及相關資訊，可如資料訊息公告、採購流程表單、電子郵件、資料庫資料管理以及商業分析報表等等訊息，讓員工在管理和處理上能夠更有效率，B&Q 特力屋亦建置了型錄回應預測模型、產品交叉銷售名單篩選模型以及客戶消費模式區隔模型來了解顧客喜好購買之產品，同時也能夠發現顧客的

消費模式，以推出良好的行銷策略，成功的搶佔商機，因此也登上居家修繕品牌龍頭的寶座。

三、　台灣雲端產業協會（Taiwan Cloud Computation Consortium，TCCC）

台灣雲端產業協會成立於 2010 年 4 月，由工研院、資策會、中華電信、台灣區電機電子公會與中華資訊軟體協會聯合規劃籌組而成，其成立宗旨為推動台灣雲端運算服務產業類型，即 IaaS、PaaS 和 SaaS 三種類型，並結合資訊科技及能源科技發展高度軟硬體整合的雲端系統平台，以及協助台灣產業朝系統解決方案及軟體服務的結構轉型。台灣雲端產業協會分為四個工作小組：

1、 雲端系統組：研擬並開放雲端系統平台的技術規格，強化 IT 與 ET 的結合，並建構國內環保節能的雲端設備產業鏈。
2、 雲端服務組：推動國內雲端應用服務的平台與環境，並發展雲端服務產業類型。
3、 法規標準組：協助國內產業在雲端作業系統、雲端資安、雲端資料中心之節能、標準、安全規格與專利等雲端標準議題，共同探討雲端相關技術專利地圖與法規標準。
4、 合作推廣組：促進國際及海峽兩岸產業合作交流，推動端產品技術標準與互通性驗證。

台灣雲端產業協會首屆理事長為中華電信董事長呂學錦（2010）表示 2011 年是「雲端運算元年」，各國都在急起直追，而台灣擁有穩固的硬體基礎，雲端能促成企業轉型，此協會將扮演政府與企業之間的橋樑，亦將促進產業合作的發展，將台灣的雲端服務輸出到國際市場，使台灣成為全球雲端設備研發製造的重鎮。

第參章　研究方法

本章共分為三節，第一節為「研究工具介紹」，將介紹 RExcel 和 Excel VBA 二種工具；第二節為「研究流程」，敘述如何以研究工具達到研究目的的流程；第三節為「研究方法」，將介紹本研究在研究過程中所採用的分類方法。

第一節　研究工具介紹

本研究以 RExcel 以及 Excel VBA 為研究工具，簡介如下：

一、　RExcel 簡介

由 Thomas Baier 和 Erich Neuwirth 二位專家所創建的，可自行下載 RAndFriends 件的壓縮檔後，即可安裝在個人電腦上。RExcel 適用於 R 軟體版本為 2.9.0 以上以及 Excel 版本為 2003 或 2007。

R 軟體是一種免費的統計軟體，為一群跨國際的志工人員組成的 R 核心發展組織（R Core-development Team）所創造而成的，並由這群志工人員持續維持且更新。由於 R 是免費的軟體，使用者在 The R project（http://www.r-project.org/）網站上便可自行下載安裝，且可於 Windows, Mac, Unix, Linux 等不同平台上執行，而 R 目前的最新版本為 2.12.1。R 擁有數百個擴充套件（Packages）可以安裝使用，能處理統計資料、統計運算與分析、統計模擬與統計繪圖功能，其特色在於以物件導向為主的程式語言，操作者必需熟悉 R 語法，才透過互動的方式與 R 進行統計運算及繪圖功能。圖 3-1 為 R 軟體的使用介面.。

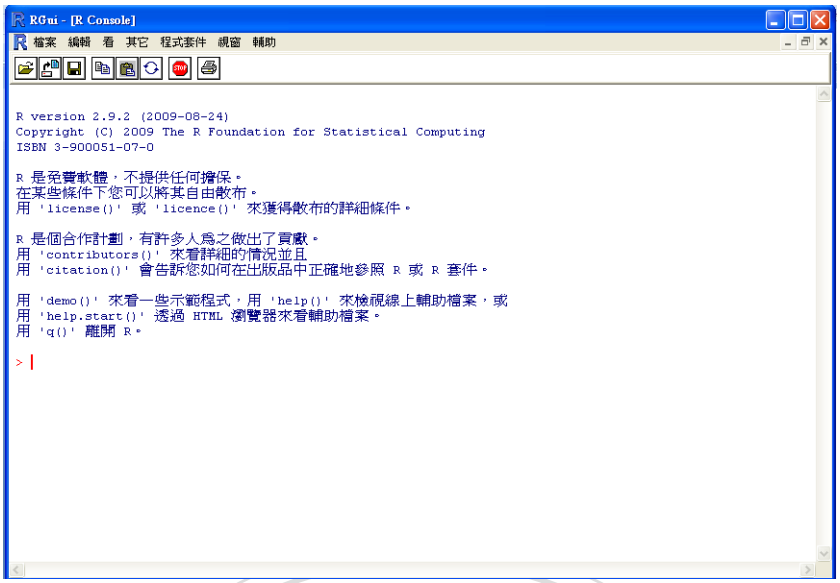


圖3-1　R 軟體介面

REXcel 是 R 軟體和 Excel 的結合，可將 R 的統計方法、圖形和結果導入 Excel，同樣地也可將在 Excel 上的資料輸送到 R 裡進行運算，這二種軟體的結合可互通有無。REXcel 安裝完畢後，會建立在 Excel 的增益集內，開啟時會同時開啟 REXcel 和 R Commander，要進行 R 程序時，可從 Excel 的選單裡選擇所要的統計方法，其介面如圖 3-2：

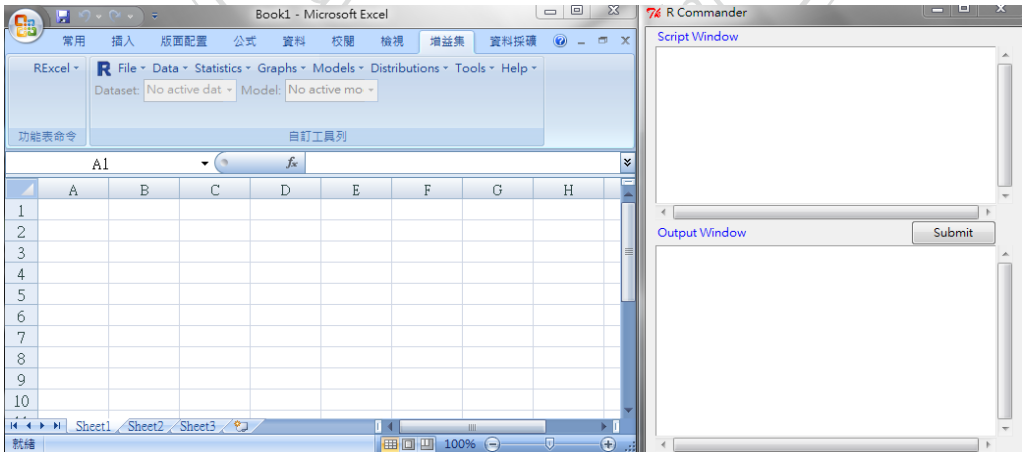


圖3-2　REXcel 和 R Commander 介面

二、 Excel VBA

VBA（Visual Basic for Application）是一種 VB（Visual Basic）的巨集語言，延伸 V B 對應用程式的存取，為 MicroSoft 微軟公司發行，將 VB 應用於其 Office 系列軟體中，如在 Office Word、Excel、Power Point 和 AutoCAD 等軟體中都有支援 VBA，使得 VBA 為共通的巨集語言，且在這些軟體中都有建立錄製巨集的功能，使用者可自行錄巨集來進行所需要的控制指令，而不必撰寫程式。由於 VBA 是一種完全面向物件體系結構的編程語言，可設計自動化的指令，且擁有在開發方面容易操作性質和強大的功能，因此許多應用程式都嵌入 VBA 作為開發工具。而 VBA 程式碼只可以在副檔名為 DOC、MDB、XLS、PPT 等檔案內執行。

VBA 最大特色就是提供了多種物件，這些物件就是各種軟體檔案格式的內容，例如在 Excel 的 VBA 內有 Workbook（活頁簿）、Worksheet（工作表），可供開發者動態地更改或控制檔案；定義 Excel 的界面、選單工具及欄位，可以簡化選項版面的使用； 在整理資料時，可對資料進行複雜的操作和分析，設計重複動作的指令，在創建報表時可以提高使用者的工作效率；可提供開發者開發複雜性及重複性計算的程式，使動作自動化，不需花費時間在制式的操作上，節省成本；以 VBA 為開發工具也具備了許多優點，如與 Office 軟體緊密結合、開發速度快、容易再進行修改動作及開發工作、使用者互動窗口效率高等。而 Excel 為大部分一般使用者較常使用的 Office 軟體，熟悉的界面，操作方便，內建大量函數，亦可連接多種資料庫，所以本研究選擇利用 Excel VBA 來開發系統。

以 Excel 作為 Visual Basic 編輯器的開發環境，其介面如圖 3-3，而圖右邊的空白處即為 VBA 編寫程式處。

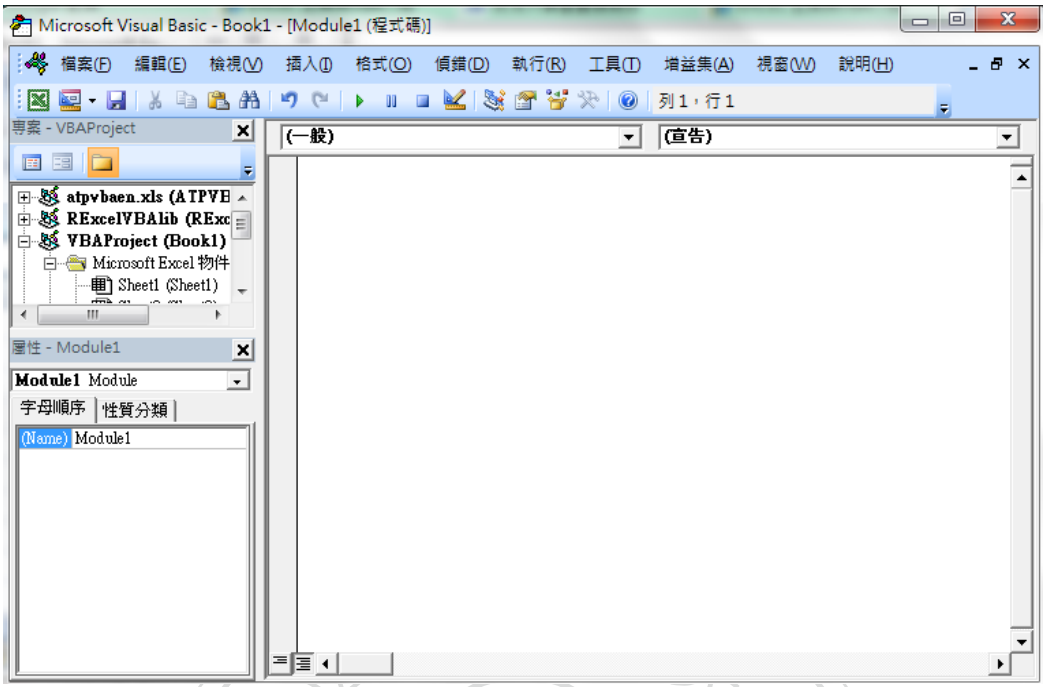


圖3-3　Excel VBA 編輯器的開發環境

第二節　研究流程

本研究主要以建立在 RExcel 中的 R 軟體發展而成的雲端系統平台，使用者將欲分析的資料上傳後，資料會透過 VBA 傳到 RExcel 中的 R 軟體中進行分類模型分析，R 軟體分析的結果再由 VBA 呼叫並傳回到 RExcel 中的 Excel 中呈現給使用者做為決策的參考，資料分析與工具流程如圖 3-4 所示。

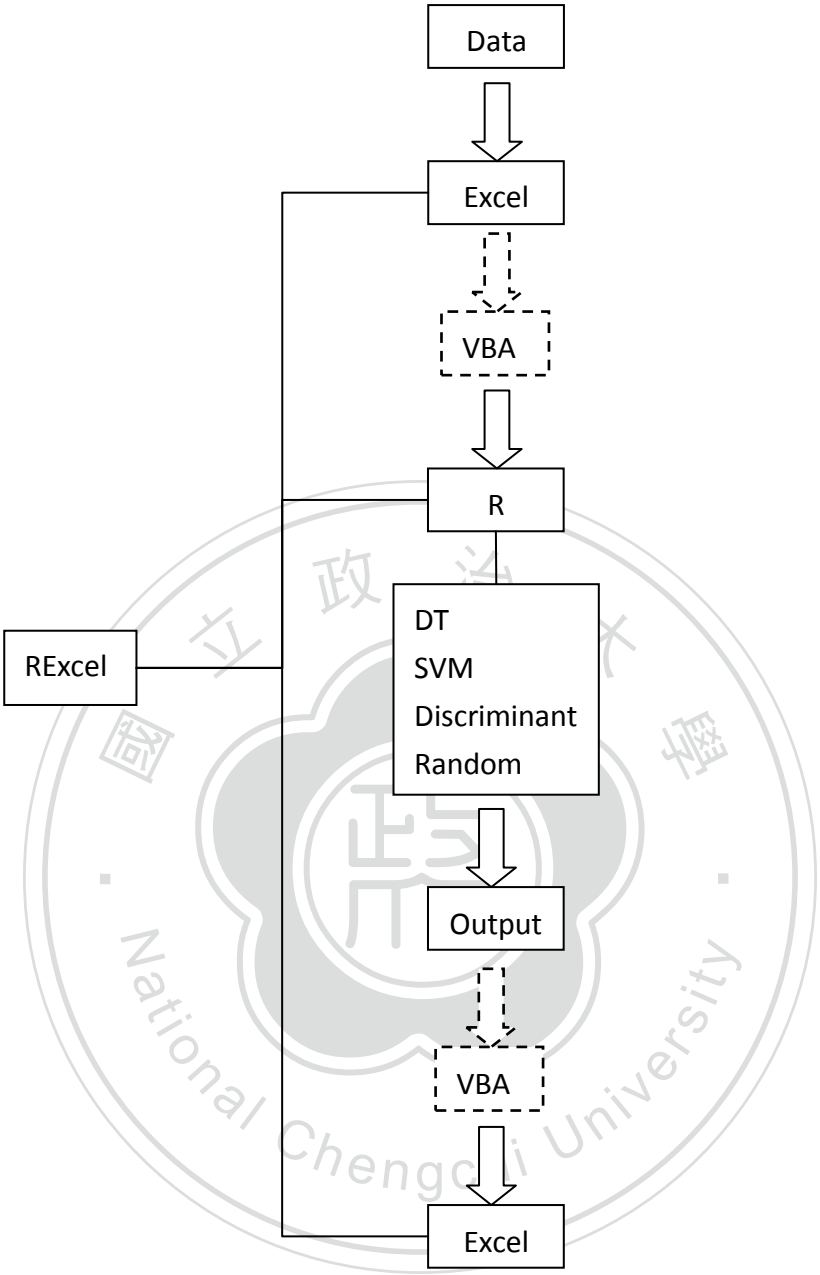


圖3-4　資料分析與工具流程

使用者在進入此雲端系統平台後，所呈現的畫面為 Excel，在進行分析時，所有的表單、對話框以及按鈕等等設計，皆為利用 VBA 撰寫出來的，使用者將透過簡單的點選動作，即可進行分析，操作步驟如下：

步驟一：使用者將所想要分析的資料讀取至 Excel 中。

步驟二：使用者檢視資料並選擇分類功能。

步驟三：使用者自行選擇欲進行分類的解釋變數及目標變數，待選擇好目標變數後，按下執行按鈕，開始分析，並輸出分析結果至 Excel 工作表。

當使用者在操作此系統時，同時 R 軟體和 VBA 也在系統背後交互運作，所有使用者接觸到的視窗皆是由 VBA 所設計出來的，而 R 軟體主要在分析並建立模型，其分析流程如下：

1、 資料傳至 Excel 並檢視資料：當使用者進行第一步驟時，將所想要分析的資料從使用者電腦瀏覽並選擇後，傳送至 Excel，按下一步，即操作第二步驟，使用者便可以檢視資料，確認資料後，便可再進行下一步。

2、 選擇目標變數並判別其型態：當使用者要進行第三步驟時，VBA 會將資料的所有變數傳送至視窗中，由使用者自行選擇解釋變數及目標變數，選擇後按下執行，系統會開始進行分析，首先利用 R 來判別定義目標變數的型態，目標變數可能有兩種型態，即數字型態和文字型態，其中若型態為數字的話，有二種可能表示方式，一為目標變數本身為連續型變數，一為目標變數以數字型態表示類別型變數，例如用數字 1 和 0 分別表示男生和女生，為了區別這二種可能，本研究自訂了一個判別方法，即設定若目標變數的種類大於 5，則判定目標變數為連續型變數；若目標變數的種類小於或等於 5，則判定為類別型變數。

3、 分析並建模：目標變數定義完後，利用 VBA 將資料傳送至 R，再將資料分為訓練資料集（Training Data）與測試資料集（Testing Data），在此預設分別為資料的 90%與 10%，然後利用訓練資料集開始建構模型，若目標變數為連續型變數，則建立決策樹、支持向量機及隨機森林等三種模型；若目

標變數為類別型變數，則建立決策樹、支持向量機、判別分析及隨機森林等四種模型。當模型建構後，不同模型會在不同的工作表中分別呈現出來，而本研究為了比較不同模型的優劣，依目標變數的型態不同來計算出特定值以便於使用者比較：若目標變數為連續型態，會計算出 MAPE(Mean Absolute Percentage Error）值，在不同的模型下，可比較 MAPE 的大小來判定何種模型較佳，若 MAPE 越小表示模型越佳，而 MAPE 計算方式為（Y_i 為實際值，$\hat{Y}_i$ 為預測值）：

$$\text{MAPE} = \frac{1}{n}\sum\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \times 100\%$$

若目標變數為類別型態，其輸出結果有訓練資料集與測試資料集的錯誤分類表（Confusion Table），並利用錯誤分類表計算出不同模型之正確率（Correct Rate），使用者可利用正確率的大小來判定何種模型較佳，另外，若目標變數為二元時，使用者可自行計算精確度（Precision Rate）以及回應率（Recall Rate），此二種值可判定模型的優劣。

以下以目標變數型態為 T 和 F 為例，其錯誤分類表如表 3-1，而正確率、精確度以及回應率計算方式為：

表3-1　二元目標變數之錯誤分類表

預測值 / 實際值	F	T
F	True Negative	False Positive
T	False Negative	True Positive

$$正確率=\frac{|TruePositvie|+|TrueNegatvie|}{|TruePositvie|+|FalsePositvie|+|TrueNegatvie|+|FalseNegatvie|}\times 100\%$$

$$精確度=\frac{|TruePositvie|}{|TruePositvie|+|FalsePositvie|}\times 100\%$$

$$回應率=\frac{|TruePositvie|}{|TruePositvie|+|FalseNegatvie|}\times 100\%$$

4、 輸出報表：經由 VBA 指令，將在 R 中分析後的結果呼叫並傳回 Excel 的工作表，再將整個工作簿以網路預覽方式呈現。

工具的運用、使用者的步驟以及分析流程可利用圖 3-5 的研究流程圖來說明：

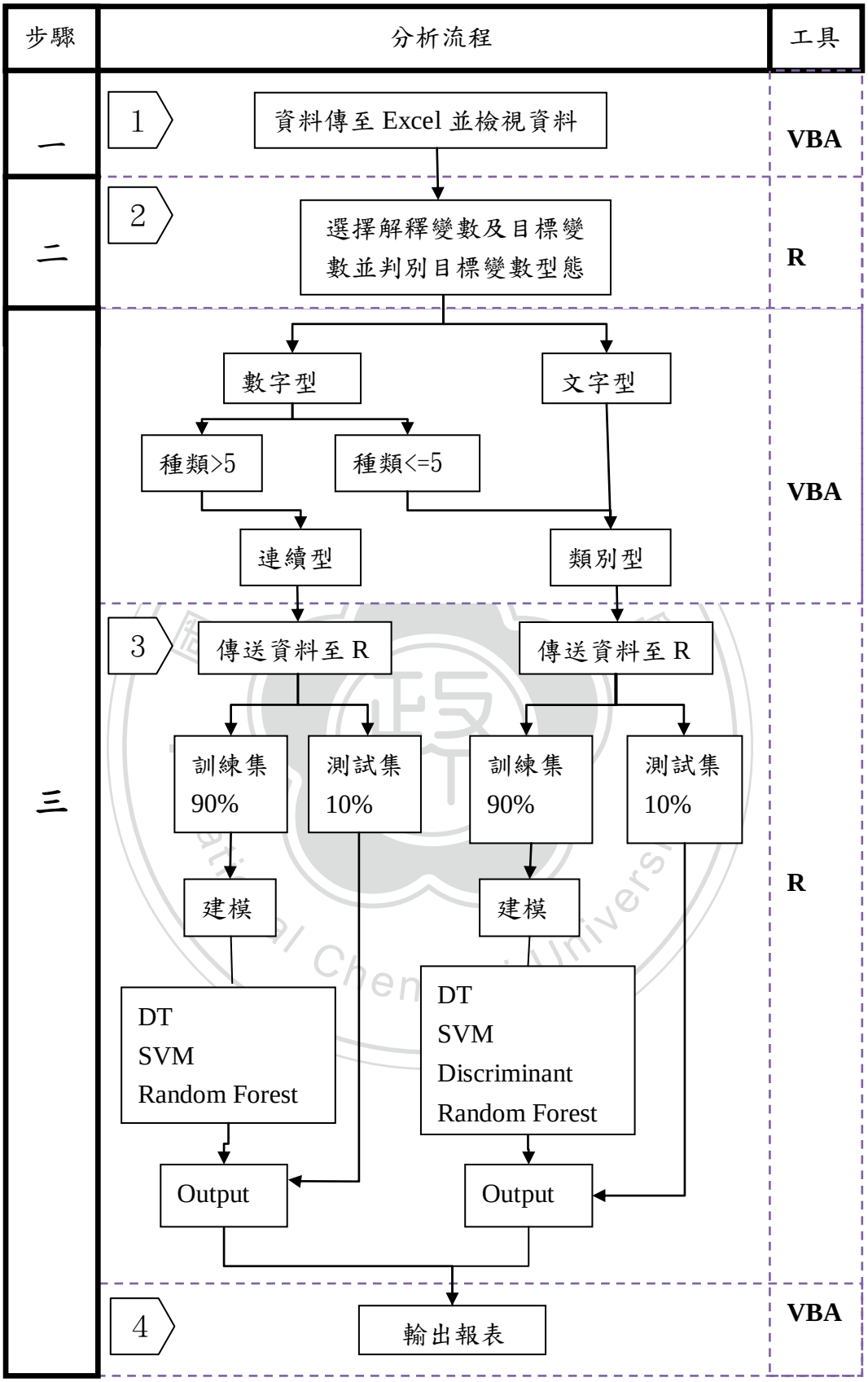

步驟
分析流程
工具
一
二
三
1
資料傳至 Excel 並檢視資料
VBA
2
選擇解釋變數及目標變數並判別目標變數型態
R
數字型
文字型
種類>5
種類<=5
連續型
類別型
VBA
3
傳送資料至 R
傳送資料至 R
訓練集 90%
測試集 10%
訓練集 90%
測試集 10%
建模
建模
DT
SVM
Random Forest
DT
SVM
Discriminant
Random Forest
Output
Output
R
4
輸出報表
VBA

圖3-5　研究流程

第三節 分類模型方法

分類是針對欲分析的資料，根據其屬性的不同，而分成不同類的過程，在此過程中，找尋出分類依據的準則，並利用該準則將尚未瞭解的資料加以分類，判斷其歸屬或做出決策。以下介紹本研究所使用的分類模型：

一、 決策樹（Decision Tree，DT）

決策樹又稱規則推理模型，藉由已知的資料建立樹狀結構，利用歸納方法找出其分類的規則，再依據此規則，對新資料進行分類，是很常用的分類工具之一，且較為其他統計分類模型較易理解，其樹狀結構圖如圖 3-6，共分為根部節點（Root Node）、中間節點（Non-leaf Node）、分支（Branches）以及葉節點（Leaf Node）四個部分。

決策樹主要的演算法包含 C&R Tree、C5.0、CHAID 以及 QUEST 四種，以下將介紹此四種演算法。

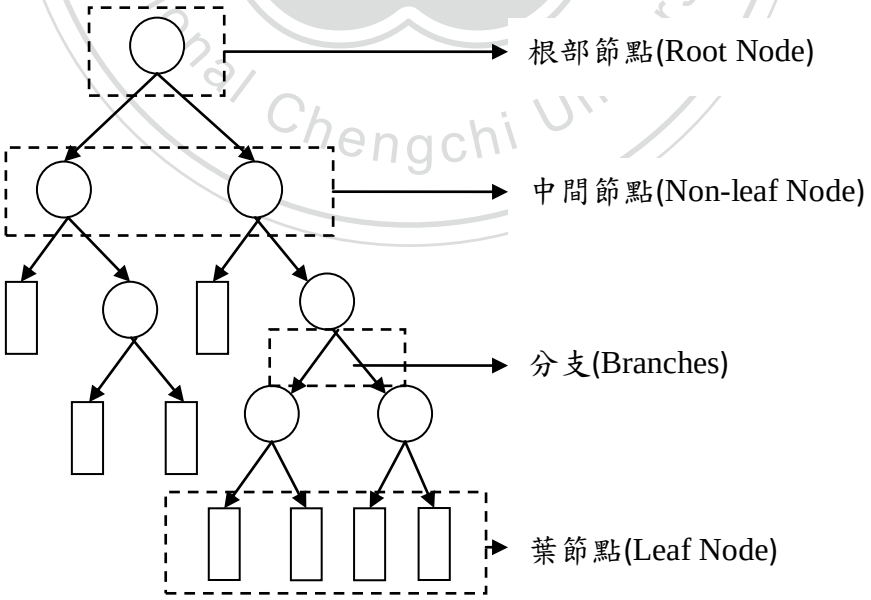


圖3-6 決策樹

(一) C&R Tree（Classification and Regression Tree）

C&R Tree 稱為分類迴歸樹，即 CART 演算法，由 Breiman 在 1984 年提出，Ripley 在 1996 年進行了修改。若當使用者設定的目標變數是類別型變數，為分類樹；若目標變數是連續型變數，則為迴歸樹。C&R Tree 是以遞迴的方法，在每個節點建立二元(Binary)分支決策樹，常用的分支節點的準則為 Gini 係數(Gini Index)，從根部節點到每一個葉節點，同一個屬性都可以被重複檢驗，直到節點完全達到純性或節點內只剩下一個值，則決策樹便停止成長，於是產生了最後的葉節點，為分類後所獲得的分類標記，透過不斷分割的方式，來提高分類的準確率。但完整的決策樹可能受到雜訊資料的影響，而對訓練樣本特徵的描述產生過於精確的現象，缺少一般代表性而無法對新資料做最佳的分類預測，出現過度配適（Over Fitting）的情況，於是需要對決策樹進行修剪的動作，而修剪的依據為決策樹整體誤差率，使修剪過後的決策樹的分支最少且具有更佳的預測能力。

(二) C5.0

C5.0 是由 ID3（Iterative Dichotomiser 3）和 C4.5 改進而來的，而 ID3 是以 Shannon 在 1949 年的資訊理論（Information Theory）為依據，由 Quinlan 於 1979 年提出。C5.0 與 C4.5 相異之處為 C5.0 利用 Boosting 的方法，按序建立多重模型，以提高其精確度，首先找出能帶來最大資訊增益（Information Gain）的輸入變數，建立第一個模型，再利用此變數將資料進行最佳分割，建立第二個模型，重複此分割方式，直到無法再被分割為止，即成為葉節點，最後，重新檢驗葉節點，將無顯著貢獻的資料形成的子樹加以修剪或刪除。

(三) CHAID（Chi-squared Automatic Interaction Detection）

CHAID 稱為卡方自動互動偵測法，由 Hartigan 在 1975 年提出的演算法，主要是以卡方檢定來選擇具有統計顯著性的輸入變數做為最佳分割的變數，此演算

法是利用輸入變數，將資料分割成兩個或兩個以上的節點，將無顯著性差異的變數合併，重新分割，直到分割後產生顯著性差異，則被保留，若保留下來的變數產生最大差異性，則被選擇為當前的分割節點。重複以上程序，直到無任何變數達到顯著性差異，則停止決策樹的成長。CHAID 演算法與 C&R T 和 C5.0 的最大差異是 CHAID 在過度配適資料的情況發生之前就將決策樹停止成長，而後兩者演算法則是先過度配適，再加以修剪配適出來的決策樹，另一差異為 CHAID 只適用於類別型資料，若為連續型變數，則必須先區隔成幾個區段範圍。

(四) QUEST（Quick Unbiased Efficient Statistic Tree）

QUEST 稱為快速、不偏且有效的統計樹，由 Loh 和 Shih 在 1997 年提出，其分類準則是利用顯著性檢定，選擇 p-value 最小且小於顯著水準的輸入變數做為當前的最佳分枝變數，若目標變數為連續型變數，則使用統計上的 ANOVA-F 的檢定；若目標變數為類別型變數，則使用統計上的卡方檢定。

二、 支持向量機 (Support Vector Machine，SVM)

支持向量機是利用最優化方法來解決機器學習問題的新工具，亦是資料採礦中的一項新技術，它能處理許多在現實生活中所遇到的問題，如時間序列分析、生物序列分析、手寫字元識別、圖像分類及判別分析等問題，其性能勝於其它大部分的學習系統。因為一般使用的資料大部分為高維度的資料，SVM 的分類概念為在資料空間裡，找出一個能將資料切割成兩類別的超平面（Hyper-plan），使屬於類別一的資料均落在超平面的同側，而屬於類別二的資料則落在超平面的另一側，其分類步驟如圖 3-7 所示，先將原始資料映成至一個高維度空間，使非線性的資料也可以被分類成不同的集合，找出分類線（如圖 3-7 的曲線）後再將原始資料做轉換，轉換過後的資料，不同類別的資料即可由超平面區隔開來，如圖 3-7 中的斜直線即為超平面。

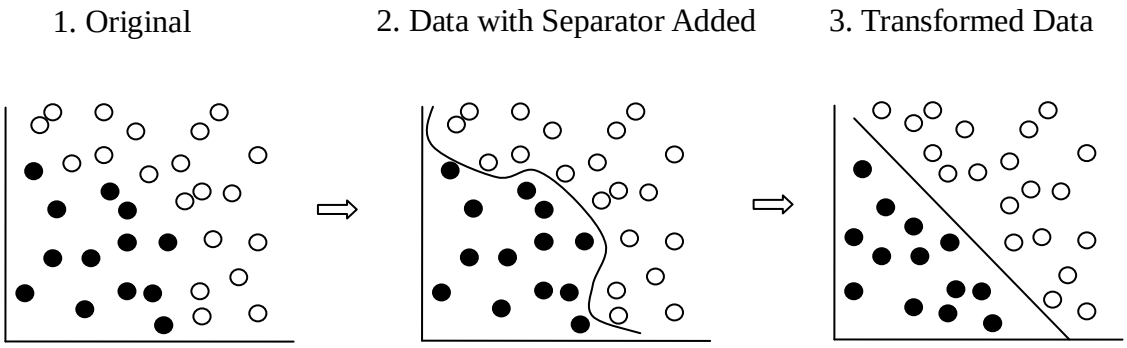


圖3-7　SVM 分類步驟

SVM 除了利用超平面區分類別外，也利用邊際線（Margin）加以定義，如圖 3-8 所示，圖中的實線為超平面，二條虛線為邊際線，邊際線差距越廣，表示其模型的預測能力越佳，而有時為了使邊際線較廣，在調校的過程中，有可能會產生少部分的錯誤分類，在此情況下，核函數（Kernel Function）中有一調校參數 C，可以在邊際線廣度以及錯誤分類間取得最適當的平衡。

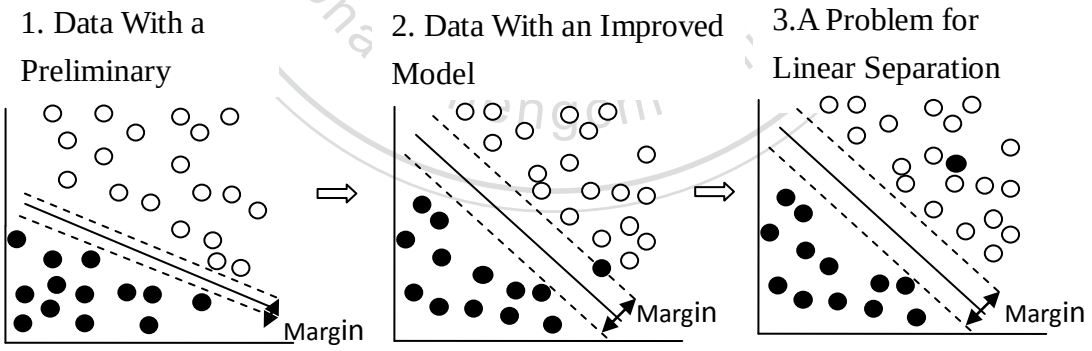


圖3-8　SVM 調校過程

三、 判別分析（Discriminant Analysis）

判別分析是一種相依的方法，其主要目的是將資料中已分類的數個群體，利用判別變數（Discriminant Variable）建立一個判別準則，而此判別準則即是判別函數（Discriminant Function），再由此函數對新個體進行分類歸屬。例如信用卡公司會依照客戶的收入、年齡、教育程度等基本資料，利用一指標將客戶區分為是違約戶與非違約戶兩種群體，當有新客戶欲申辦信用卡時，可利用此準則來判別此客戶在未來是否為違約戶。判別分析適用於當目標變數為類別型，而自變數為連續型時使用。

常用的線性判別分析有線性判別函數（或稱分類函數，Classification Function）和典型判別函數（費雪判別函數）兩種：

1、 線性判別函數（Linear Discriminant Function，LDF）：建立判別準則最常使用的原理是依據各群體會發生此組資料的機率，再將此個體判別在發生機率最大的群體。

2、 典型判別函數：由 Fisher 於 1936 年所創，尋找判別變數的線性組合之最佳權重，使其組間變異數對組內變異數的比值最大。

四、 隨機森林（Random Forest）

隨機森林是 Breiman 在 2001 年提出的一種分類方法，以 C&R Tree、拔靴集合（Bootstrap Aggregation）以及隨機子空間（Random Subspace Method）等為基礎理論所發展出來的。隨機森林是由多個決策樹子集合所構成的大型決策樹，而與一般決策樹相異之處就是隨機森林要對每個決策樹子集合進行判斷，當目標變數分別為連續型與類別型時，則分別透過簡單多數表決（Simple Majority Vote）與單棵樹輸出結果的平均，來決定最後分類的結果，也會依大數法則（Law of Large Numbers）對決策樹進行收斂，因此隨機森林不會有過度配適的情形發生。

隨機森林可以處理相當龐大且不同型態的資料，對於資料的遺漏值處理，亦有良好的方法可以估計，若有一部分的資料遺失，它仍然可以維持分類的正確率，隨機森林依照以下演算法來建構每顆決策樹：

1、 抽取出訓練資料，以 N 表示訓練資料的個數，以 M 表示自變數的個數。

2、 在 M 個自變數中，選出 m 個子集合（m<M），以決定當在一個節點上做分割時，會使用到多少個變數。

3、 從 N 個訓練資料中以 Bootstrap 抽樣，重複抽樣 N 次，形成一組訓練集。

4、 對於每一個子集合，隨機選擇 m 個子集合中的變數，再根據此 m 個變數，計算其最佳子集合，形成最佳分割方式。

5、 每棵決策樹都會完整成長而不會進行修剪。

第肆章　實證分析

本章共分為四節，第一節為「研究限制」，條列出使用者欲上傳分析之資料限制；第二節為「數字連續型目標變數」，以數字連續型目標變數為例，說明使用者操作步驟及輸出報表內容，並評估模型優劣；第三節為「數字類別型目標變數」，以數字類別型目標變數為例，說明使用者操作步驟及輸出報表內容，並評估模型優劣；第四節為「文字類別型目標變數」，以文字類別型目標變數為例，說明使用者操作步驟及輸出報表內容，並評估模型優劣。

第一節 研究限制

使用者欲分析之資料在上傳至此分類系統時，有一些資料的限制，使用者必須在上傳資料前先確認好資料是否符合資料限制，若不符合，使用者需自行將資料做適當的調整或更改，而這些資料限制如下：

1、 資料上傳的檔案類型只能是 Microsoft Office Excel 逗點分隔值檔案(.csv)、Microsoft Office Excel 工作表 (.xls)及 Microsoft Office Excel 工作表(.xlsx)三種，若資料檔案類型是文字文件(.txt)，必須轉換成上述之其中一種檔案才能在此分類系統中讀取。

2、 此系統不支援有中文字之資料檔案，若資料內容有中文字，有可能在傳送資料中會出現問題，或者在 R 軟體裡讀取資料時會出現讀取不完全之問題，因此有中文字的欄位需改成以英文表示。

3、 資料欄位中，不需要包括索引（Index）欄位，如資料的第一行為資料筆數 1、2、3、…。

4、 資料的所有欄位名稱，必須放在資料的第一列，欄位內容從第二列開始放置，資料格式如圖 4-1 所示。

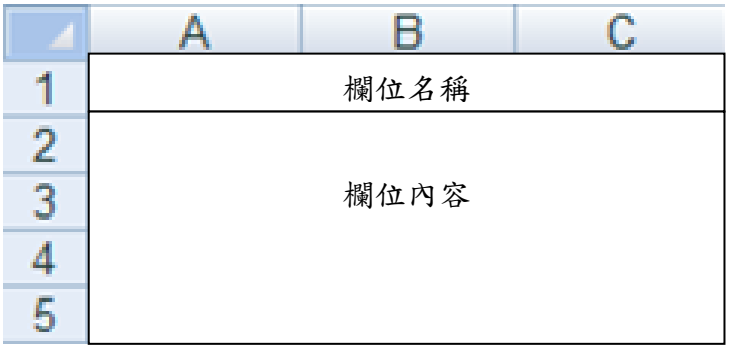


圖4-1　資料格式

5、 若資料中有遺漏值，必須以”空格”或”NA”表示，在此注意，”NA”要以半形大寫英文字表示，其餘表現方式皆不可行。

6、 由於數字型目標變數是以其種類個數大於5及小於等於5來分成連續型及類別型，所以此系統只適用於目標變數為類別型且種類個數小於等於5，若目標變數為類別型，但種類個數大於5，則會被歸類為連續型變數。

第二節 數字連續型目標變數

本研究所設定的連續型目標變數為數字型態，且目標變數種類大於5，此章節依此種類型的目標變數舉例說明。

一、　資料說明

利用「Babies」資料檔為例，此資料為懷孕母親的各項資料以及新生嬰兒體重，資料筆數共有 1,236 筆，其中有遺漏值為”NA”的有 52 筆，共 7 個欄位，資料欄位名稱說明如表 4-1 所示。

表4-1　Babies 資料說明

欄位名稱	欄位說明	欄位內容
bwt	新生嬰兒體重（盎斯）	數值連續型
gestation	母親懷孕日數（天）	數值連續型
parity	母親有無生產經驗	數值類別型
		0：有經驗；1：無經驗
age	母親年齡（歲）	數值連續型
height	母親身高（英吋）	數值連續型
weight	母親體重（磅）	數值連續型
smoke	母親有無吸菸	數值類別型
		0：無吸菸 ；1：有吸菸

資料來源：TKU Netstat 網站

二、　使用者操作說明

依照第三章的研究流程中所設計的使用者操作步驟來說明。

步驟一：

進入此系統後，會先跳出「上傳欲分析資料」之視窗，使用者可上傳欲分析的資料，操作介面如圖 4-2。

點選「上傳資料 (Upload Data)」之按鈕，即跳出「請瀏覽並選取欲載入的檔案」之視窗，如圖 4-3 所示，可瀏覽使用者電腦中的資料，使用者可選取欲分析資料，在此選擇桌面上的「Babies」檔案，其檔案類型為 Microsoft Office Excel 工作表(.xlsx)，選取完成後按下「開啟」按鈕。

圖4-2　使用者上傳欲分析資料之視窗

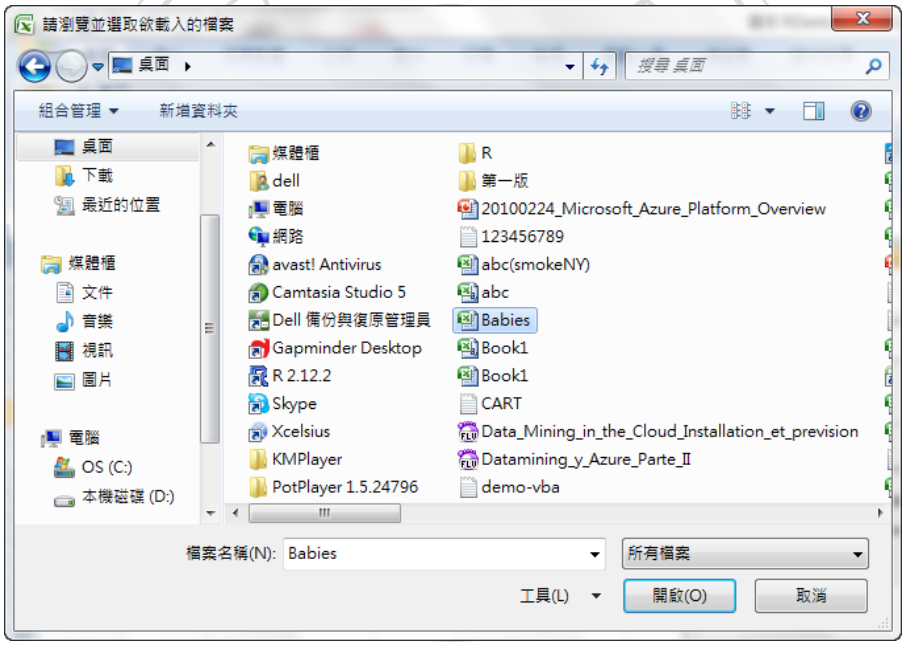


圖4-3　瀏覽並選取欲載入的檔案之視窗

步驟二：

按下開啟按鈕後，此系統會讀取使用者欲分析之資料，並跳出「檢視上傳資料」之視窗，如圖 4-4 所示，使用者可檢視資料是否上傳成功，並再次確定此資料是否為自己欲分析之資料。

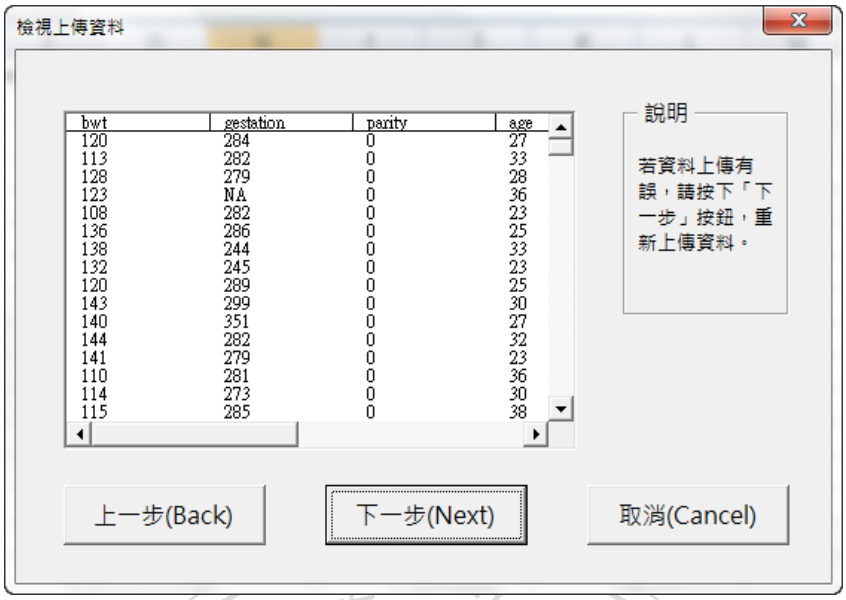


圖4-4　檢視上傳資料之視窗

使用者檢視並確定資料後，按下「下一步」按鈕，此按鈕置入了可清除資料遺漏值之功能，可清除 Babies 檔案中的 52 筆遺漏值資料，剩餘 1,184 筆資料，在清除完遺漏值部分資料後，接著會跳出「選擇欲分析之資料採礦功能」之視窗，如圖 4-5 所示，此視窗包含了四種不同功能按鈕，分別為「預測(Forecasting)」、「分類(Classification)」、「關聯規則(Association Rule)」以及「集群分析(Clustering)」，使用者可依欲分析之功能來選擇其中一種功能，本研究為研究分類之功能，因此點選「分類(Classification)」按鈕，進入下一個步驟。

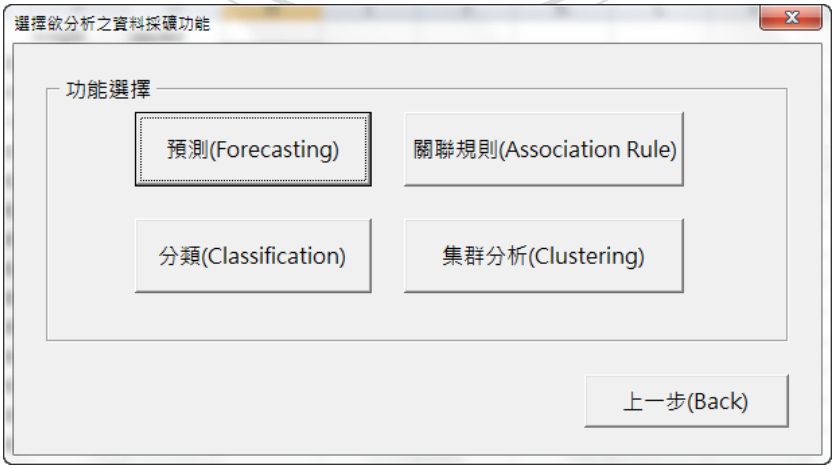


圖4-5　選擇欲分析之資料採礦功能之視窗

步驟三：

按下「分類(Classification)」按鈕後，即跳出「資料採礦之分類功能」之視窗，如圖 4-6 所示，視窗中的二個清單裡分別顯示所有資料欄位名稱，使用者可就欲分析的目的來選擇解釋變數及目標變數，在此示範的目標變數為數字連續型，於是點選「bwt」，而解釋變數為其它六個變數，選擇完畢後，按下「執行」按鈕，系統將已去除遺漏值之資料以及目標變數「bwt」傳送到 R 軟體開始進行分析，並將分析結果依不同模型分別以各別的工作表呈現。

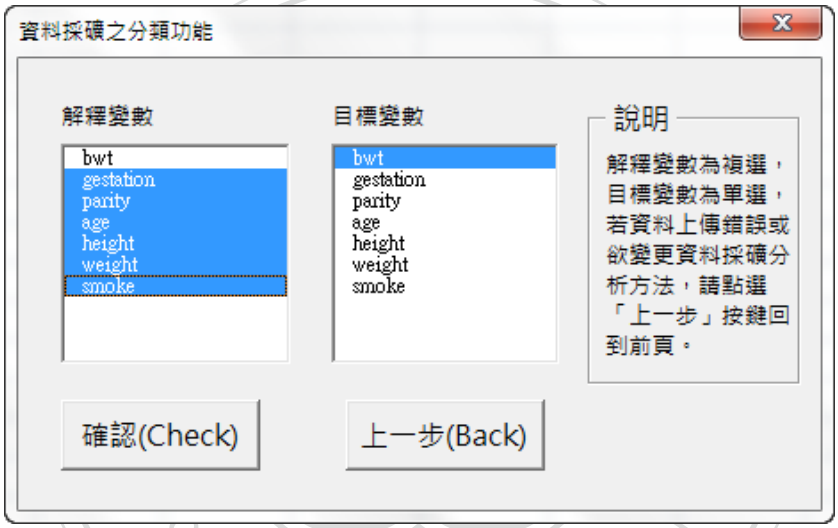


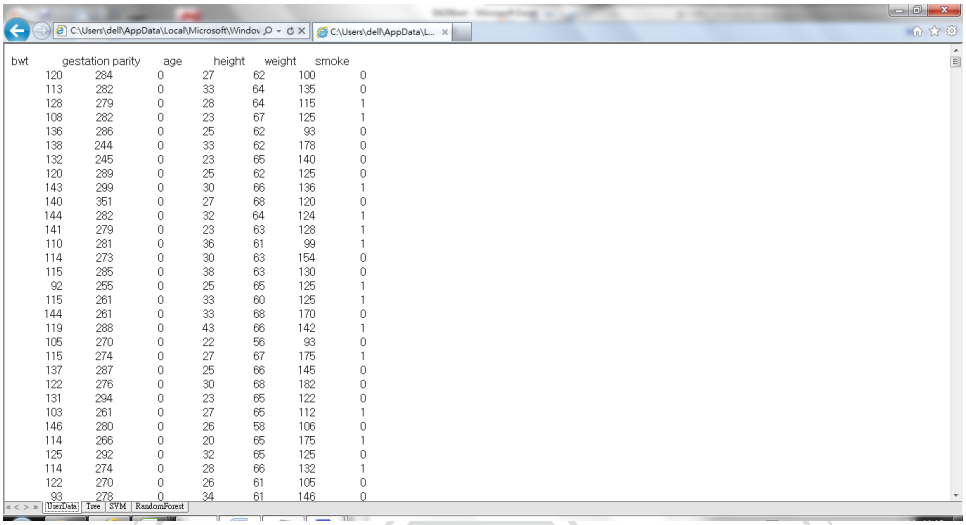
圖4-6　資料採礦之分類功能之視窗

三、　報表輸出

選擇數字連續型目標變數會產生三種模型，分別是決策樹、支持向量機以及隨機森林，分析結果會以網頁預覽方式呈現給使用者，報表中共有四個工作表，工作表名稱分別為「UserData」、「Tree」、「SVM」以及「RandomForest」，以下將以各個工作表內容做說明。

1、「UserData」工作表

此工作表內容為使用者上傳欲分析之資料且已清除遺漏值，共有 1,184 筆資料，如圖 4-7 所示。



bwt	gestation	parity	age	height	weight	smoke
120	284	0	27	62	100	0
113	282	0	33	64	135	0
128	279	0	28	64	115	1
108	282	0	23	67	125	1
136	286	0	25	62	93	0
138	244	0	33	62	178	0
132	245	0	23	65	140	0
120	289	0	25	62	125	0
143	299	0	30	66	136	1
140	351	0	27	68	120	0
144	282	0	32	64	124	1
141	279	0	23	63	128	1
110	281	0	36	61	99	1
114	273	0	30	63	154	0
115	285	0	38	63	130	0
92	255	0	25	65	125	1
115	261	0	33	60	125	1
144	261	0	33	68	170	0
119	288	0	43	66	142	1
105	270	0	22	56	93	0
115	274	0	27	67	175	1
137	287	0	25	66	145	0
122	276	0	30	68	182	0
131	294	0	23	65	122	0
103	261	0	27	65	112	1
146	280	0	26	58	106	0
114	266	0	20	65	175	1
125	292	0	32	65	125	0
114	274	0	28	66	132	1
122	270	0	26	61	105	0
93	278	0	34	61	146	0

圖4-7　「UserData」工作表

2、「Tree」工作表

此工作表為利用決策樹分析之結果，網頁預覽如圖 4-8 所示。

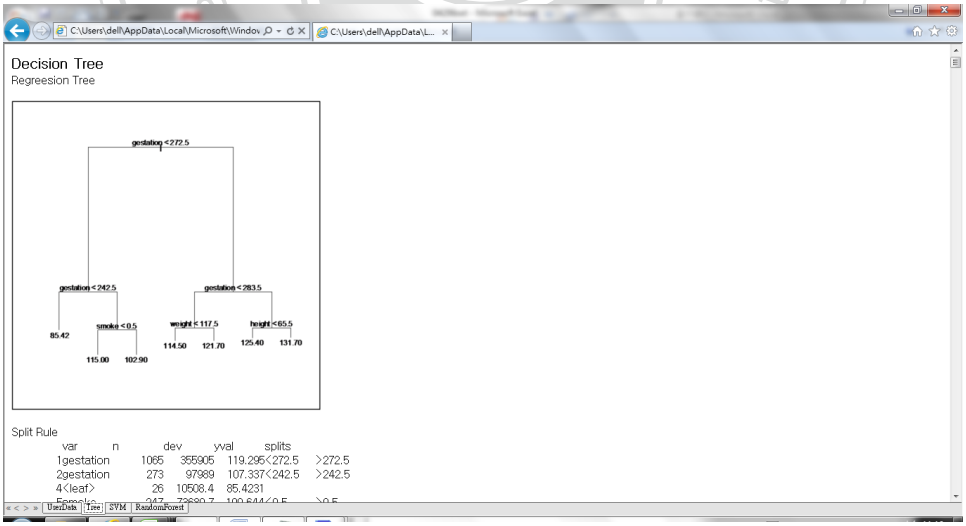


圖4-8　「Tree」工作表

決策樹分析結果報表如圖 4-9 所示，其中有四個部分，分別為：

(1) Regression Tree

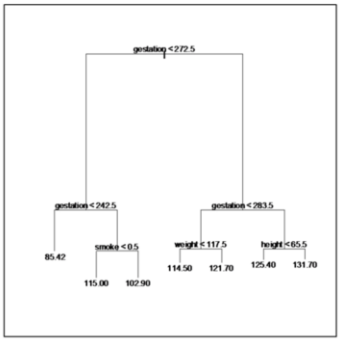
(2) Split Rule

(3) MAPE

(4) Testing Data Detail

Decision Tree

Regreesion Tree



Split Rule

var	n	dev	yval	splits	
1gestation	1065	355905	119.295	<272.5	>272.5
2gestation	273	97989	107.337	<242.5	>242.5
4<leaf>	26	10508.4	85.4231		
5smoke	247	73680.7	109.644	<0.5	>0.5
10<leaf>	137	37957.9	115.029		
11<leaf>	110	26800.6	102.936		
3gestation	792	205425	123.417	<283.5	>283.5
6weight	395	88720.9	119.342	<117.5	>117.5
12<leaf>	127	27265.5	114.457		
13<leaf>	268	56988.4	121.657		
7height	397	103619	127.471	<65.5	>65.5
14<leaf>	266	64687.6	125.365		
15<leaf>	131	35354.7	131.748		

Testing Data Detail

bwt	gestation	parity	age	height	weight	smoke	node	predicted
75	232	0	33	61	110	0	3	85.42308
132	225	0	28	67	148	0	3	85.42308
75	239	0	26	63	124	1	3	85.42308
85	234	0	33	67	130	0	3	85.42308
105	233	0	34	61	130	0	3	85.42308
116	148	0	28	66	135	0	3	85.42308
71	234	0	32	64	110	1	3	85.42308
129	235	0	24	66	135	0	3	85.42308
68	223	0	32	66	149	1	3	85.42308
96	241	0	23	64	130	1	3	85.42308
69	232	0	31	59	103	1	3	85.42308
103	240	0	26	65	140	0	3	85.42308
78	237	1	23	63	144	0	3	85.42308
87	229	0	27	62	138	0	3	85.42308
86	242	0	20	64	110	1	3	85.42308
77	238	1	23	63	103	1	3	85.42308
62	228	0	24	61	107	0	3	85.42308
92	224	0	19	63	134	1	3	85.42308
110	181	0	27	64	133	0	3	85.42308
77	238	0	38	67	135	1	3	85.42308
55	204	0	35	65	140	0	3	85.42308

MAPE

Percentage	12.18%

圖4-9　決策樹分析結果報表

3、「SVM」工作表

此工作表為利用支持向量機分析之結果，網頁預覽如圖 4-10 所示。

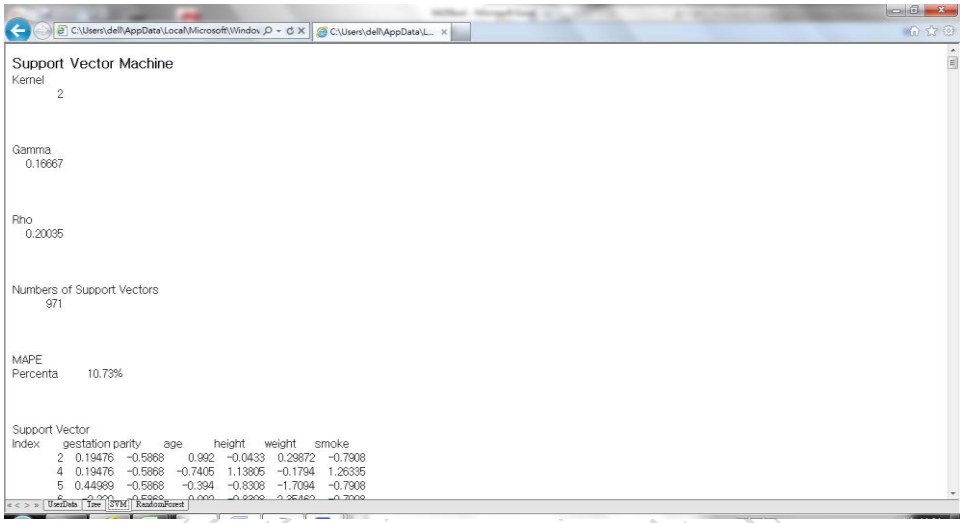


圖4-10 「SVM」工作表

支持向量機分析結果報表如圖 4-11 所示，其中有六個部分，分別為：

(1) Kernel

(2) Gamma

(3) Rho

(4) Number of Support Vector

(5) Support Vector

(6) MAPE

Support Vector Machine

Kernel

2

Gamma

0.16667

Rho

0.20035

Numbers of Support Vectors

971

Support Vector

Index	gestation	parity	age	height	weight	smoke
2	0.19476	-0.5868	0.992	-0.0433	0.29872	-0.7908
4	0.19476	-0.5868	-0.7405	1.13805	-0.1794	1.26335
5	0.44989	-0.5868	-0.394	-0.8308	-1.7094	-0.7908
6	-2.229	-0.5868	0.992	-0.8308	2.35462	-0.7908
7	-2.1652	-0.5868	-0.7405	0.35051	0.53778	-0.7908
8	0.64124	-0.5868	-0.394	-0.8308	-0.1794	-0.7908
9	1.27907	-0.5868	0.47225	0.74428	0.34653	1.26335
11	0.19476	-0.5868	0.81875	-0.0433	-0.2272	1.26335
12	0.00341	-0.5868	-0.7405	-0.437	-0.036	1.26335
13	0.13098	-0.5868	1.51176	-1.2246	-1.4225	1.26335
14	-0.3793	-0.5868	0.47225	-0.437	1.20714	-0.7908
15	0.38611	-0.5868	1.85826	-0.437	0.05966	-0.7908
16	-1.5274	-0.5868	-0.394	0.35051	-0.1794	1.26335
17	-1.1447	-0.5868	0.992	-1.6183	-0.1794	1.26335
18	-1.1447	-0.5868	0.992	1.53181	1.97213	-0.7908
19	0.57746	-0.5868	2.72451	0.74428	0.6334	1.26335
20	-0.5706	-0.5868	-0.9138	-3.1934	-1.7094	-0.7908

MAPE

Percentage 10.73%

圖4-11　支持向量機分析結果報表

4、「RandomForest」工作表

此工作表為利用隨機森林分析之結果，網頁預覽如圖 4-12 所示。

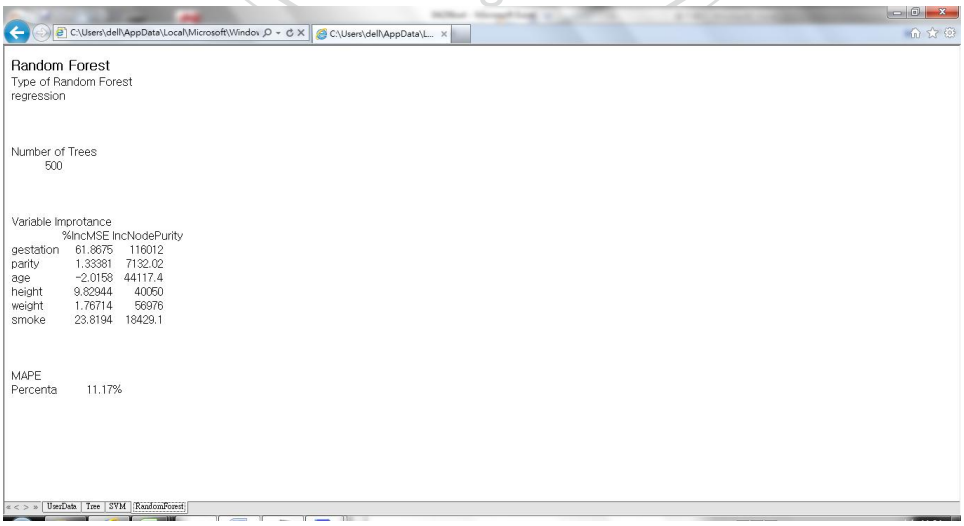


圖4-12　「RandomForest」工作表

隨機森林分析結果報表如圖 4-13 所示，其中有四個部分，分別為：

(1) Type of Random Forest

(2) Number of Trees

(3) Variable Importance

(4) MAPE

Random Forest
Type of Random Forest
regression

Number of Trees
500

Variable Importance

	%IncMSE	IncNodePurity
gestation	61.8675	116012
parity	1.33381	7132.02
age	-2.0158	44117.4
height	9.82944	40050
weight	1.76714	56976
smoke	23.8194	18429.1

MAPE
Percenta 11.17%

圖4-13　隨機森林分析結果報表

四、　評估模型優劣

選擇數字連續型目標變數所分析出來的結果中，會計算出各個模型的 MAPE 值，並呈現在報表中，使用者可以比較 MAPE 之大小來評估模型的優劣，表 4-2 為決策樹、支持向量機及隨機森林的 MAPE，其 MAPE 值的大小順序為決策樹＞隨機森林＞支持向量機，因此，可判定支持向量機所分類出來的結果最佳。

表4-2　分類模型之 MAPE 比較

模型	決策樹	支持向量機	隨機森林
MAPE	12.18%	10.73%	11.17%

第三節 數字類別型目標變數

一、 資料說明

利用 The Data and Story Library（DASL）網站中的「Egyptian Skulls」資料檔為例，此資料測量來自西元前及西元後的埃及男性頭骨之最大頭骨寬度、Basibregmatic 骨頭高度、Basialveolar 骨頭長度以及鼻骨高度，資料筆數共有 150 筆，5 個欄位，資料欄位名稱說明如表 4-3 所示。

表4-3 Egyptian Skulls 資料說明

欄位名稱	欄位說明	欄位內容
MB	最大頭骨寬度	數字連續型
BH	Basibregmatic 骨頭高度	數字連續型
BL	Basialveolar 骨頭長度	數字連續型
NH	鼻骨高度	數字連續型
Year	西元前後	文字類別型 1：西元前；0：西元後

二、 使用者操作說明

依照第三章的研究流程中所設計的使用者操作步驟來說明。

步驟一：

進入此系統後，會先跳出「上傳欲分析資料」之視窗，使用者可上傳欲分析的資料，操作介面如圖 4-14。

圖4-14　使用者上傳欲分析資料之視窗

點選「上傳資料 (Upload Data)」之按鈕，即跳出「請瀏覽並選取欲載入的檔案」之視窗，如圖 4-15 所示，可瀏覽使用者電腦中的資料，使用者可選取欲分析資料，在此選擇桌面上的「Egyptian Skulls」檔案，其檔案類型為 Microsoft Office Excel 工作表(.xlsx)，選取完成後按下「開啟」按鈕。

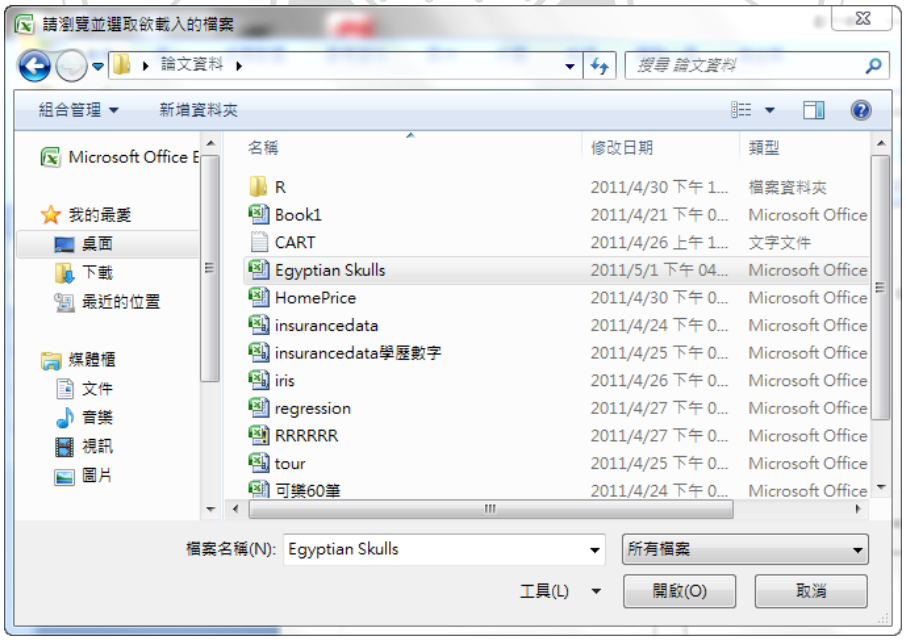


圖4-15　瀏覽並選取欲載入的檔案之視窗

步驟二：

按下開啟按鈕後，此系統會讀取使用者欲分析之資料，並跳出「檢視上傳資料」之視窗，如圖 4-16 所示，使用者可檢視資料是否上傳成功，並再次確定此資料是否為自己欲分析之資料。

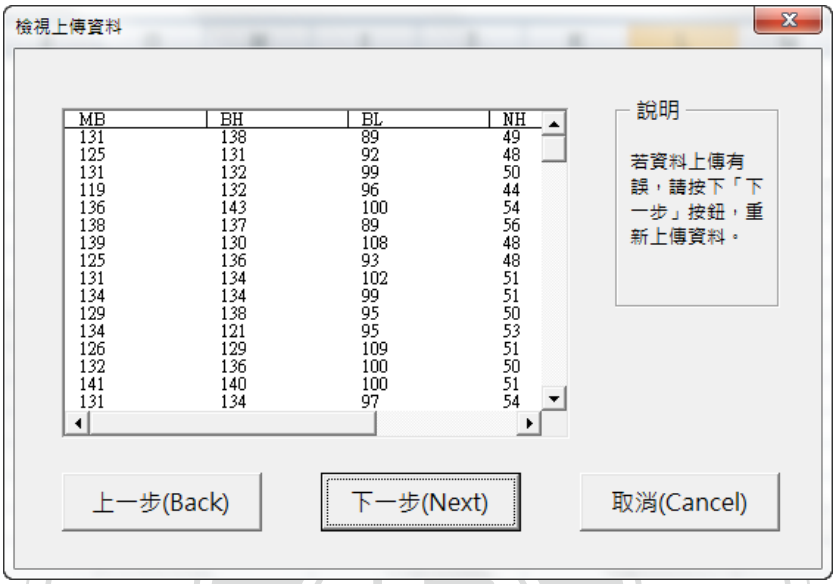


圖4-16　檢視上傳資料之視窗

使用者檢視並確定資料後，按下「下一步」按鈕，接著會跳出「選擇欲分析之資料採礦功能」之視窗，如圖 4-17 所示，點選「分類(Classification)」按鈕，進入下一個步驟。

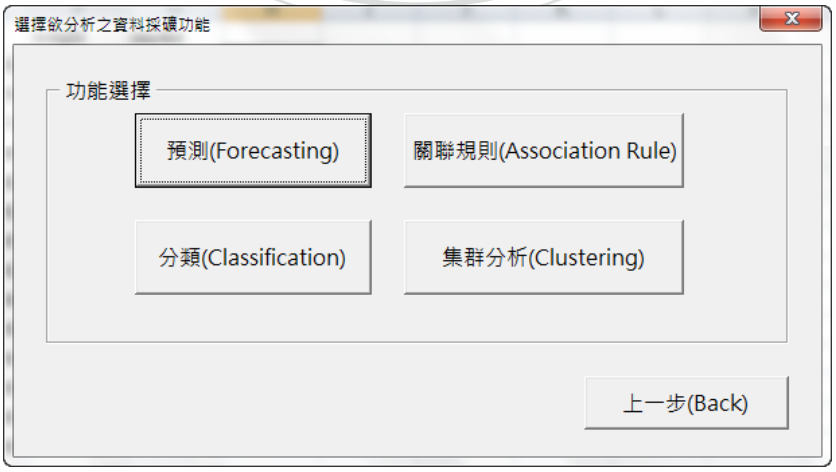


圖4-17　選擇欲分析之資料採礦功能之視窗

步驟三：

按下「分類(Classification)」按鈕後，即跳出「資料採礦之分類功能」之視窗，如圖 4-18 所示，在視窗中的二個清單裡會顯示「Egyptian Skulls」資料檔所有資料欄位名稱，在此示範的目標變數為數字類別型，於是點選「Year」，而解釋變數在此選擇其它四個變數，選擇完畢後，按下「執行」按鈕，系統將資料以及目標變數「Year」傳送到 R 軟體開始進行分析，並將分析結果依不同模型分別以各別的工作表呈現。

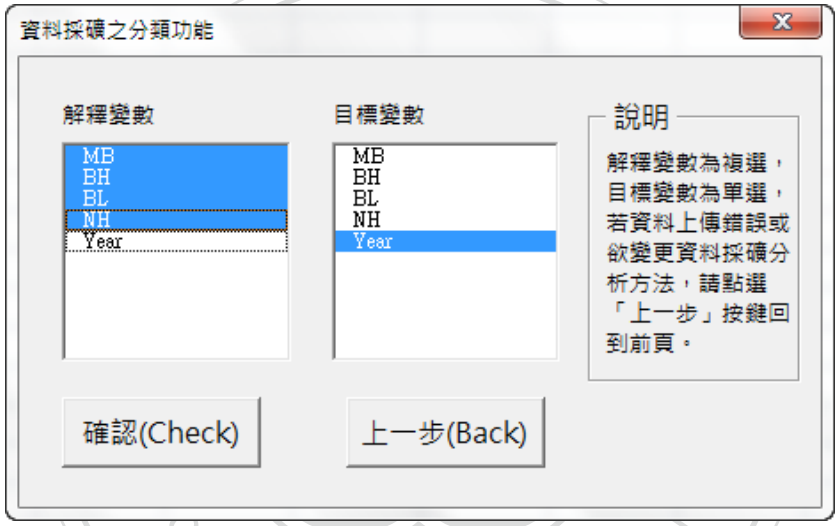


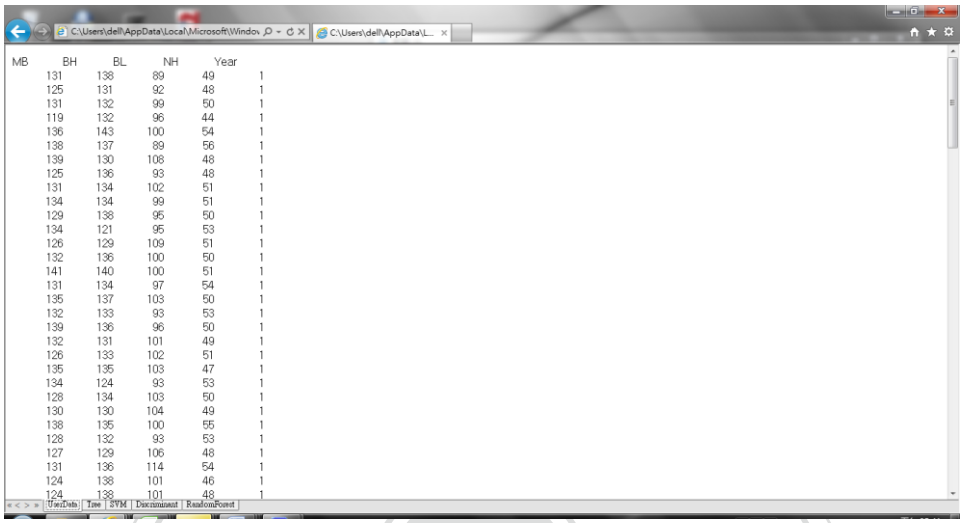
圖4-18　資料採礦之分類功能之視窗

三、　報表輸出

選擇數字類別型目標變數會產生四種模型，分別是決策樹、支持向量機、判別分析以及隨機森林，分析結果會以網頁預覽方式呈現給使用者，報表中共有四個工作表，工作表名稱，分別為「UserData」、「Tree」、「SVM」、「Discriminant」以及「RandomForest」，以下將以各個工作表內容做說明。

1、「UserData」工作表

此工作表內容為使用者上傳欲分析之資料，共有 150 筆資料，如圖 4-19 所示。

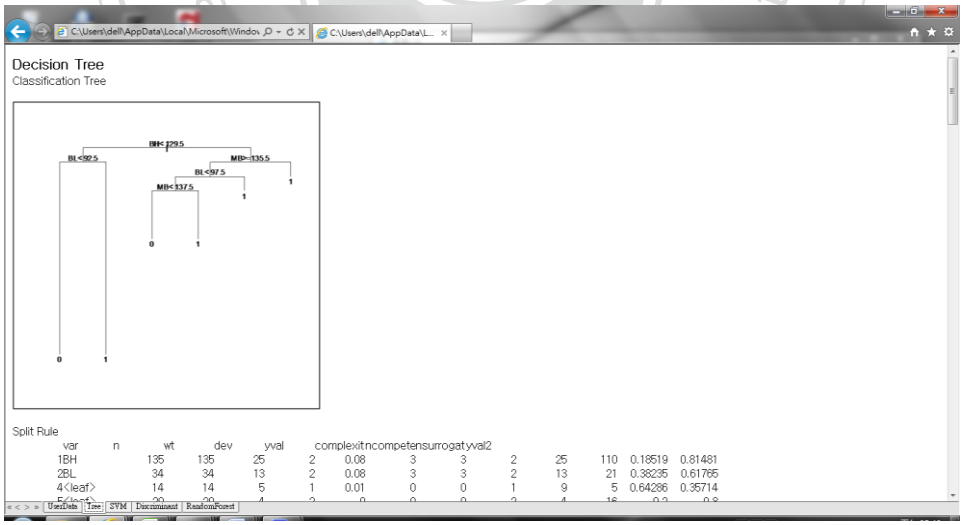


MB	BH	BL	NH	Year
	131	138	89	49	1
	125	131	92	48	1
	131	132	99	50	1
	119	132	96	44	1
	136	143	100	54	1
	138	137	89	56	1
	139	130	108	48	1
	125	136	93	48	1
	131	134	102	51	1
	134	134	99	51	1
	129	138	95	50	1
	134	121	95	53	1
	126	129	109	51	1
	132	136	100	50	1
	141	140	100	51	1
	131	134	97	54	1
	135	137	103	50	1
	132	133	93	53	1
	139	136	96	50	1
	132	131	101	49	1
	126	133	102	51	1
	135	135	103	47	1
	134	124	93	53	1
	128	134	103	50	1
	130	130	104	49	1
	138	135	100	55	1
	128	132	93	53	1
	127	129	106	48	1
	131	136	114	54	1
	124	138	101	46	1
	124	138	101	48	1

圖4-19　「UserData」工作表

2、「Tree」工作表

此工作表為利用決策樹分析之結果，網頁預覽如圖 4-20 所示。



Split Rule

	var	n	wt	dev	yval	complexit	ncompete	nsurrogat	yval2				
1	BH		135	135	25	2	0.08	3	3	2	25	110	0.18519	0.81481
2	BL		34	34	13	2	0.08	3	3	2	13	21	0.38235	0.61765
4	<leaf>		14	14	5	1	0.01	0	0	1	9	5	0.64286	0.35714

圖4-20　「Tree」工作表

決策樹分析結果報表如圖 4-21 所示，其中有七個部分，分別為：

(1)　Classification Tree

(2) Split Rule

(3) Confusion Table for Training Data

(4) Correct Rate for Training Data

(5) Confusion Table for Testing Data

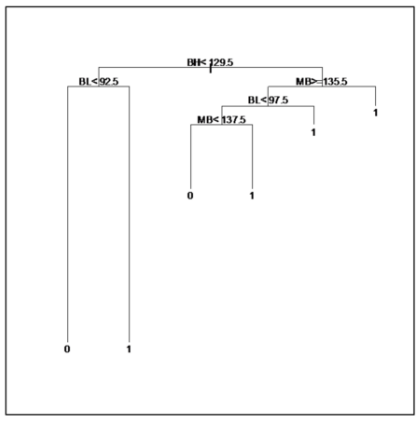
(6) Correct Rate for Testing Data

(7) Testing Data Detail

圖 4-21 第 7 個部分的 Testing Data Detail 因為資料筆數較多，此僅展示其中一部分。

Decision Tree

Classification Tree



Split Rule

var	n	wt	dev	yval	complexit	ncompe
1BH		135	135	25	2	0.08
2BL		34	34	13	2	0.08
4<leaf>		14	14	5	1	0.01
5<leaf>		20	20	4	2	0
3MB		101	101	12	2	0.01333
6BL		37	37	9	2	0.01333
12MB		24	24	8	2	0.01333
24<leaf>		9	9	4	1	0.01
25<leaf>		15	15	3	2	0.01
13<leaf>		13	13	1	2	0.01
7<leaf>		64	64	3	2	0

Confusion Table for Training Data

	0	1
0	14	11
1	9	101

Correct Rate for Training Data

Percenta 85.19%

Confusion Table for Testing Data

	0	1
0	2	3
1	4	6

Correct Rate for Testing Data

Percenta 53.33%

Testing Data Detail

Year	MB	BH	BL	NH	node	predicted.	predicted.1
1	129	126	91	50	3	0.64286	0.35714
1	134	125	90	60	3	0.64286	0.35714
1	133	120	91	46	3	0.64286	0.35714
1	131	125	88	48	3	0.64286	0.35714
1	144	124	86	50	3	0.64286	0.35714
0	137	123	91	50	3	0.64286	0.35714
0	128	126	91	57	3	0.64286	0.35714
0	126	126	92	45	3	0.64286	0.35714
0	145	129	89	47	3	0.64286	0.35714
0	143	126	88	54	3	0.64286	0.35714
0	134	124	91	55	3	0.64286	0.35714
0	137	125	85	57	3	0.64286	0.35714
0	129	128	81	52	3	0.64286	0.35714
0	147	129	87	48	3	0.64286	0.35714
1	134	121	95	53	4	0.2	0.8
1	126	129	109	51	4	0.2	0.8

圖4-21　決策樹分析結果報表

3、「SVM」工作表

此工作表為利用支持向量機分析之結果，網頁預覽如圖 4-22 所示。

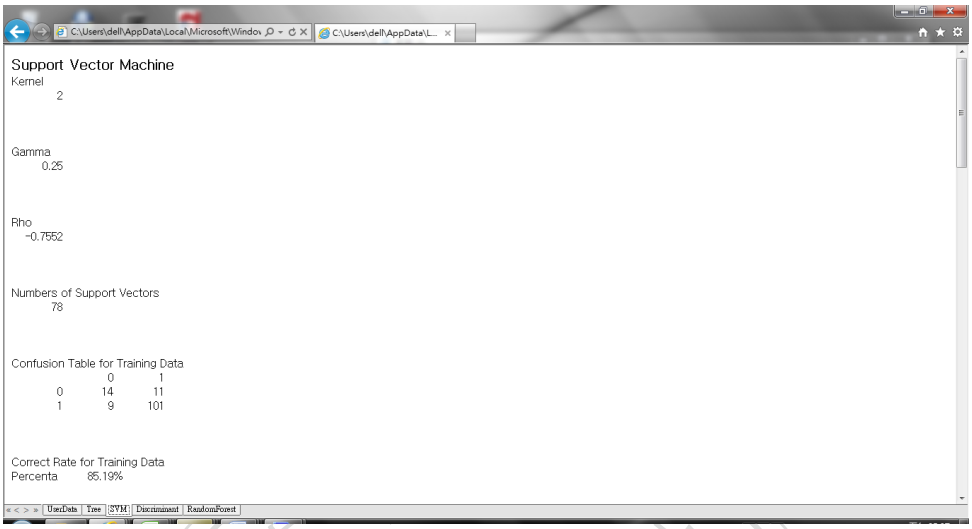


圖4-22 「SVM」工作表

支持向量機分析結果報表如圖 4-23 所示，其中有九個部分，分別為：

(1) Kernel

(2) Gamma

(3) Rho

(4) Numbers of Support Vectors

(5) Confusion Table for Training Data

(6) Correct Rate for Training Data

(7) Confusion Table for Testing Data

(8) Correct Rate for Testing Data

(9) Support Vector

圖 4-23 第 9 個部分的 Support Vector 因為資料筆數較多，此僅展示其中一部分。

Support Vector Machine

Kernel

2

Gamma

0.25

Rho

-0.7552

Numbers of Support Vectors

78

Confusion Table for Training Data

	0	1
0	14	11
1	9	101

Correct Rate for Training Data

Percenta 85.19%

Confusion Table for Testing Data

	0	1
0	0	5
1	0	10

Correct Rate for Testing Data

Percenta 66.67%

Support Vector

Index	MB	BH	BL	NH
2	-1.7673	-0.3096	-0.8418	-0.8908
4	-2.9784	-0.1067	-0.1039	-2.1242
6	0.85673	0.90769	-1.3952	1.57602
7	1.05858	-0.5125	2.10995	-0.8908
11	-0.9599	1.11056	-0.2883	-0.2741
12	0.04934	-2.3383	-0.2883	0.65096
13	-1.5654	-0.7153	2.29443	0.03426
15	1.46228	1.51632	0.63408	0.03426
19	1.05858	0.70481	-0.1039	-0.2741
21	-1.5654	0.09618	1.00304	0.03426
23	0.04934	-1.7297	-0.6573	0.65096
27	-1.1617	-0.1067	-0.6573	0.65096
29	-0.5562	0.70481	3.21685	0.95931
31	-1.9691	1.11056	0.81856	-0.8908
33	0.85673	0.29906	0.26511	-1.8158
34	2.87522	-0.7153	1.37201	0.03426
36	0.25119	0.70481	0.26511	0.34261

圖4-23　支持向量機分析結果報表

4、「Discriminant」工作表

此工作表為利用判別分析之結果，網頁預覽如圖 4-24 所示。

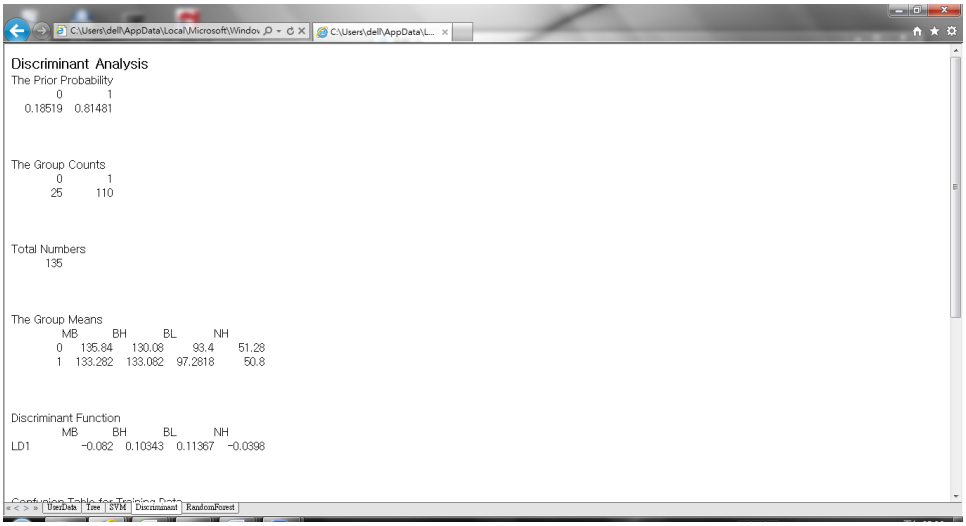


圖4-24 「Discriminant」工作表

判別分析結果報表如圖 4-25 所示，其中有九個部分，分別為：

(1) The Prior Probability

(2) The Group Counts

(3) Total Numbers

(4) The Group Means

(5) Discriminant Function

(6) Confusion Table for Training Data

(7) Correct Rate for Training Data

(8) Confusion Table for Testing Data

(9) Correct Rate for Testing Data

Discriminant Analysis

The Prior Probability

0	1
0.18519	0.81481

The Group Counts

0	1
25	110

Total Numbers

135

The Group Means

	MB	BH	BL	NH
0	135.84	130.08	93.4	51.28
1	133.282	133.082	97.2818	50.8

Discriminant Function

	MB	BH	BL	NH
LD1	-0.082	0.10343	0.11367	-0.0398

Confusion Table for Training Data

	0	1
0	14	11
1	9	101

Correct Rate for Training Data

Percenta 85.19%

Confusion Table for Testing Data

	0	1
0	1	4
1	0	10

Correct Rate for Testing Data

Percenta 73.33%

圖4-25　判別分析結果報表

5、「RandomForest」工作表

此工作表為利用隨機森林分析之結果，網頁預覽如圖 4-26 所示。

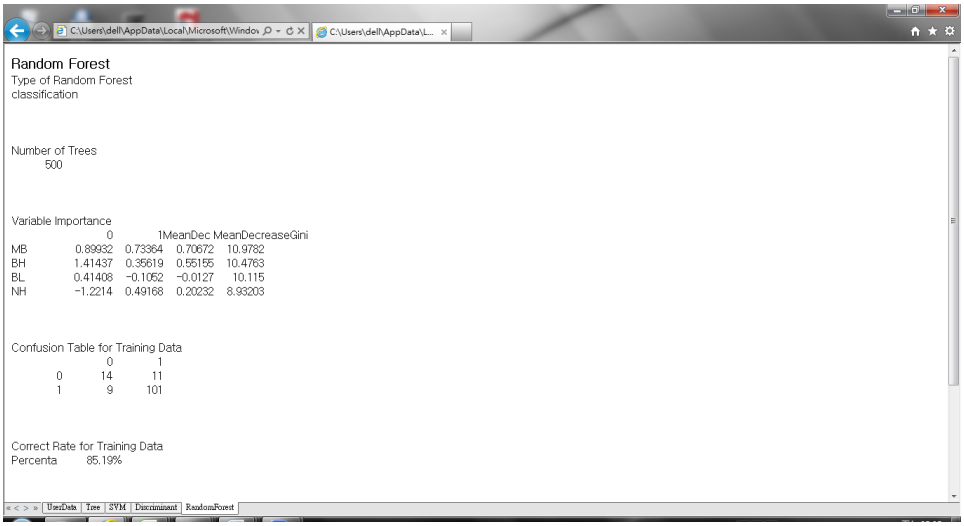


圖4-26 「RandomForest」工作表

隨機森林分析結果報表如圖 4-27 所示，其中有七個部分，分別為：

(1) Type of Random Forest

(2) Number of Trees

(3) Variable Importance

(4) Confusion Table for Training Data

(5) Correct Rate for Training Data

(6) Confusion Table for Testing Data

(7) Correct Rate for Testing Data

Random Forest
Type of Random Forest
classification

Number of Trees
500

Variable Importance

	0		1MeanDec	MeanDecreaseGini
MB	0.89932	0.73364	0.70672	10.9782
BH	1.41437	0.35619	0.55155	10.4763
BL	0.41408	-0.1052	-0.0127	10.115
NH	-1.2214	0.49168	0.20232	8.93203

Confusion Table for Training Data

	0	1
0	14	11
1	9	101

Correct Rate for Training Data
Percenta 85.19%

Confusion Table for Testing Data

	0	1
0	0	5
1	0	10

Correct Rate for Testing Data
Percenta 66.67%

圖4-27　隨機森林分析結果報表

四、　評估模型優劣

選擇數字類別型目標變數所分析出來的結果中，可計算出各個模型中訓練資料集和測試資料集的正確率，並呈現在報表中，使用者可以比較測試資料集的正確率之大小來評估模型的優劣，此目標變數為二元，使用者可自行計算出各個模型的精確度及回應率，表 4-4 為決策樹、支持向量機、判別分析及隨機森林的訓練資料集和測試資料集的正確率、精確度及回應率。分析結果顯示這四種模型在訓練資料集之表現都相同，而在測試資料集之表現，只有決策樹之回應度為 60%，其它模型皆高達 100%，然而，就模型正確率或者精確度論，都以判別分析為最高之百分比，故使用者可選擇判別分析為最佳解釋模型。

表4-4　分類模型之正確率、精確度及回應率比較

模型		決策樹	支持向量機	判別分析	隨機森林
訓練集	正確率	85.19%	85.19%	85.19%	85.19%
	精確度	83.57%	83.57%	83.57%	83.57%
	回應率	91.82%	91.82%	91.82%	91.82%
測試集	正確率	53.33%	66.67%	73.33%	66.67%
	精確度	66.67%	66.67%	71.83%	66.67%
	回應率	60%	100%	100%	100%

第四節 文字類別型目標變數

一、　資料說明

以 R 軟體資料集裡內建的「iris」資料檔為例，此資料為測量鳶尾花之三種不同種類（setosa、versicolor 及 virginica）的花萼長度、花萼寬度、花瓣長度以及花瓣寬度，每個種類分別測量 50 朵，資料筆數共有 150 筆，5 個欄位，資料欄位名稱說明如表 4-5 所示。

表4-5　iris 資料說明

欄位名稱	欄位說明	欄位內容
Sepal.Length	花萼長度（公分）	數值連續型
Sepal.Width	花萼寬度（公分）	數值連續型
Petal.Length	花瓣長度（公分）	數值連續型
Petal.Width	花瓣寬度（公分）	數值連續型
Species	花的種類	文字類別型（setosa、versicolor 及 virginica）

二、　使用者操作說明

依照第三章的研究流程中所設計的使用者操作步驟來說明。

步驟一：

進入此系統後，會先跳出「上傳欲分析資料」之視窗，使用者可上傳欲分析的資料，操作介面如圖 4-28。

圖4-28　使用者上傳欲分析資料之視窗

點選「上傳資料 (Upload Data)」之按鈕，即跳出「請瀏覽並選取欲載入的檔案」之視窗，如圖 4-29 所示，可瀏覽使用者電腦中的資料，使用者可選取欲分析資料，在此選擇「iris」檔案，其檔案類型為 Microsoft Office Excel 逗點分隔值檔案(.csv)，選取完成後按下「開啟」按鈕。

步驟二：

按下開啟按鈕後，此系統會讀取使用者欲分析之資料，並跳出「檢視上傳資料」之視窗，如圖 4-30 所示，使用者可檢視資料是否上傳成功，並再次確定此資料是否為自己欲分析之資料。

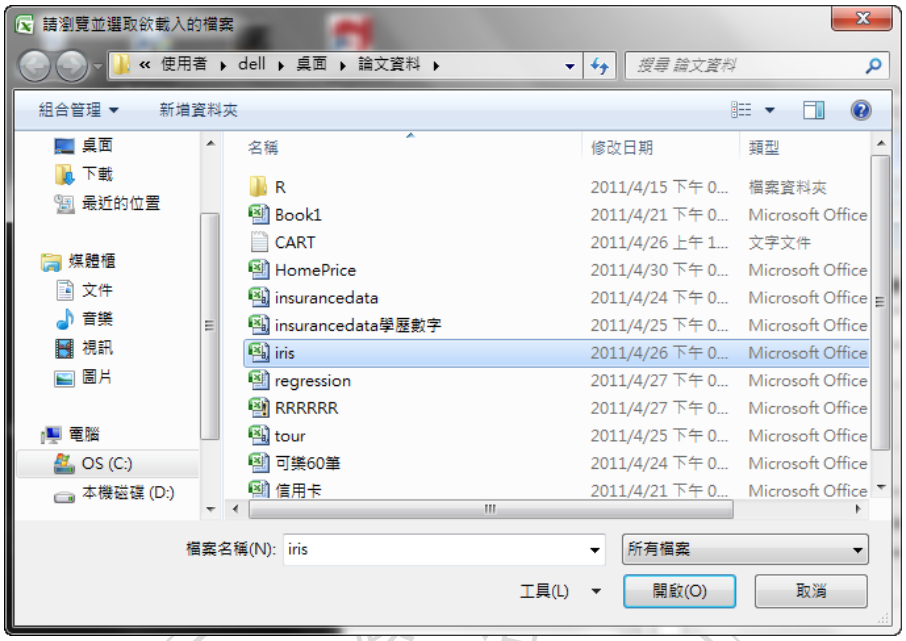


圖4-29　瀏覽並選取欲載入的檔案之視窗

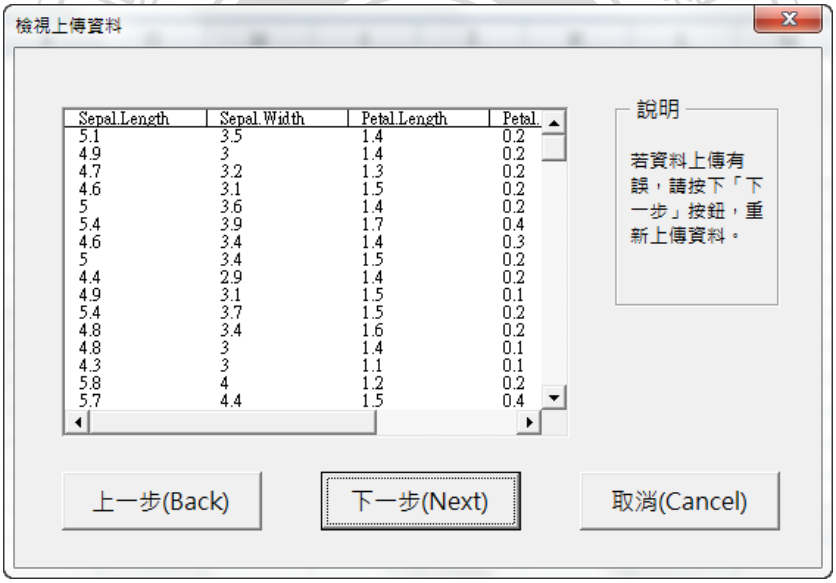


圖4-30　檢視上傳資料之視窗

使用者檢視並確定資料後，按下「下一步」按鈕，接著會跳出「選擇欲分析之資料採礦功能」之視窗，如圖 4-31 所示，點選「分類(Classification)」按鈕，進入下一個步驟。

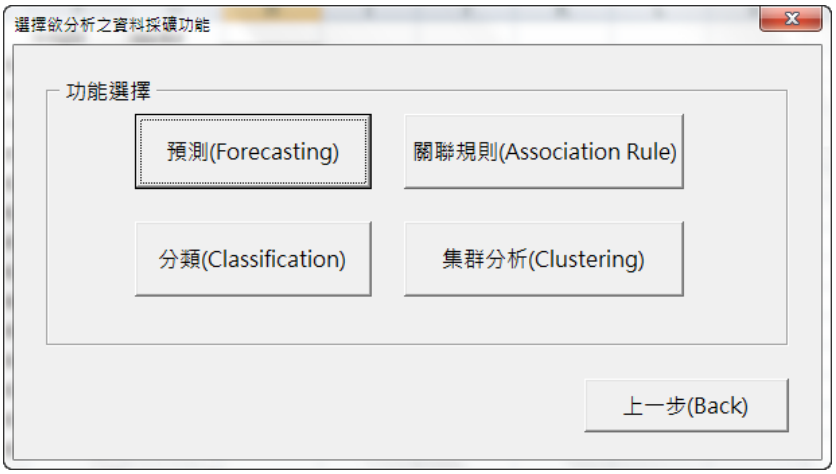


圖4-31　選擇欲分析之資料採礦功能之視窗

步驟三：

按下「分類(Classification)」按鈕後，即跳出「資料採礦之分類功能」之視窗，如圖 4-32 所示，在視窗中的二個清單裡會顯示「iris」資料檔所有資料欄位名稱，在此示範的目標變數為文字類別型，於是點選「Species」，而解釋變數則選擇其它四個變數，選擇完畢後，按下「執行」按鈕，系統便將資料以及目標變數「Species」傳送到 R 軟體開始進行分析，並將分析結果依不同模型分別以各別的工作表呈現。

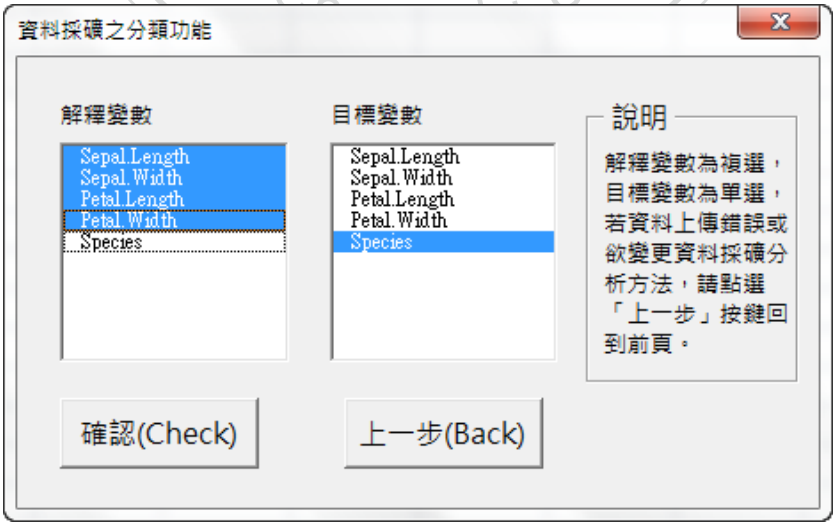


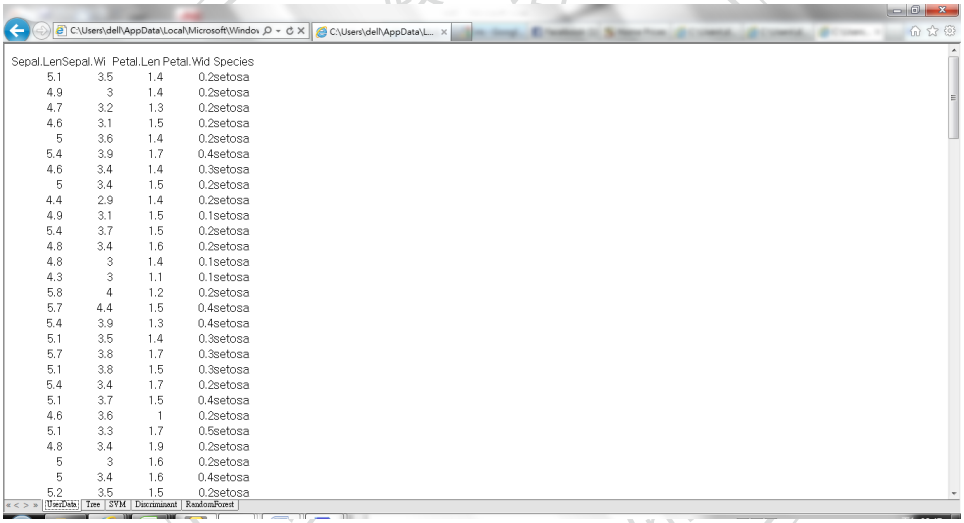
圖4-32　資料採礦之分類功能之視窗

三、　報表輸出

選擇文字類別型目標變數會產生四種模型，分別是決策樹、支持向量機、判別分析以及隨機森林，分析結果會以網頁預覽方式呈現給使用者，報表中共有四個工作表，工作表名稱分別為「UserData」、「Tree」、「SVM」、「Discriminant」以及「RandomForest」，以下將以各個工作表內容做說明。

1、「UserData」工作表

此工作表內容為使用者上傳欲分析之資料，共有 150 筆資料，如圖 4-33 所示。



Sepal.Len	Sepal.Wi	Petal.Len	Petal.Wid	Species
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3	1.4	0.1	setosa
4.3	3	1.1	0.1	setosa
5.8	4	1.2	0.2	setosa
5.7	4.4	1.5	0.4	setosa
5.4	3.9	1.3	0.4	setosa
5.1	3.5	1.4	0.3	setosa
5.7	3.8	1.7	0.3	setosa
5.1	3.8	1.5	0.3	setosa
5.4	3.4	1.7	0.2	setosa
5.1	3.7	1.5	0.4	setosa
4.6	3.6	1	0.2	setosa
5.1	3.3	1.7	0.5	setosa
4.8	3.4	1.9	0.2	setosa
5	3	1.6	0.2	setosa
5	3.4	1.6	0.4	setosa
5.2	3.5	1.5	0.2	setosa

UserData | Tree | SVM | Discriminant | RandomForest

圖4-33 「UserData」工作表

2、「Tree」工作表

此工作表為利用決策樹分析之結果，網頁預覽如圖 4-34 所示。

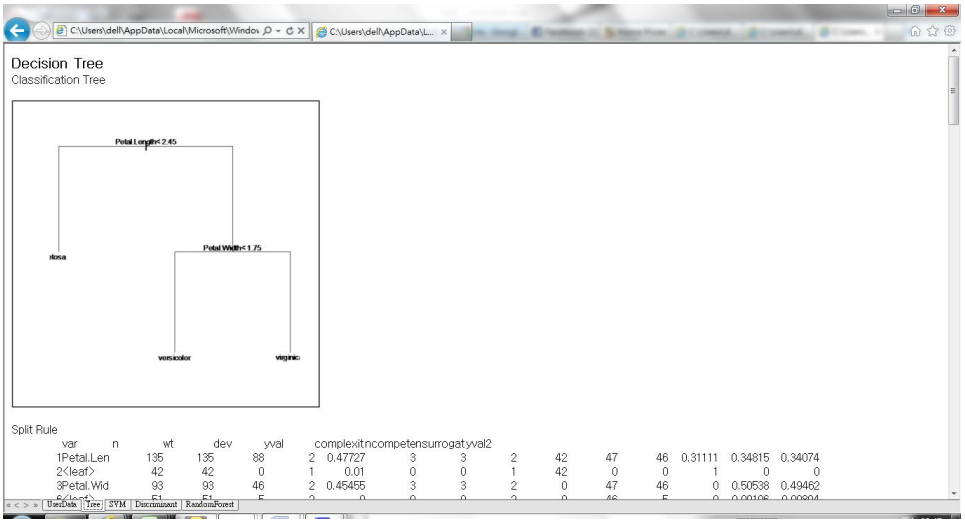


圖4-34 「Tree」工作表

決策樹分析結果報表如圖 4-35 所示，其中有七個部分，分別為：

(1) Classification Tree

(2) Split Rule

(3) Confusion Table for Training Data

(4) Correct Rate for Training Data

(5) Confusion Table for Testing Data

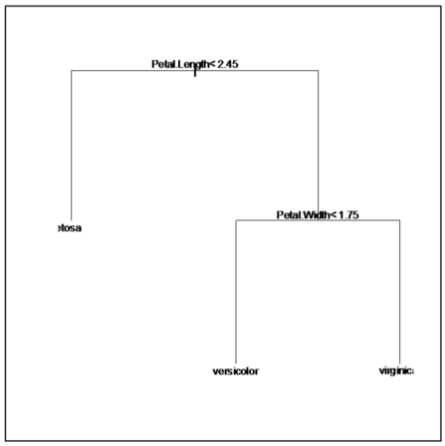
(6) Correct Rate for Testing Data

(7) Testing Data Detail

圖 4-35 第 7 個部分的 Testing Data Detail 因為資料筆數較多，此僅展示其中一部分。

Decision Tree

Classification Tree



Split Rule

	var	n	wt	dev	yval	complexit	ncompeten	surrogat	yval2					
1	Petal.Len	135	135	88	2	0.47727	3	3	2	42	47	46	0.31111	0.34815	0.34074
2	<leaf>	42	42	0	1	0.01	0	0	1	42	0	0	1	0	0
3	Petal.Wid	93	93	46	2	0.45455	3	3	2	0	47	46	0	0.50538	0.49462
6	<leaf>	51	51	5	2	0	0	0	2	0	46	5	0	0.90196	0.09804
7	<leaf>	42	42	1	3	0	0	0	3	0	1	41	0	0.02381	0.97619

Confusion Table for Training Data

	setosa	versicolor	virginica
setosa	42	0	0
versicolor	0	46	1
virginica	0	5	41

Correct Rate for Training Data

Percenta 95.56%

Confusion Table for Testing Data

	setosa	versicolor	virginica
setosa	8	0	0
versicolor	0	3	0
virginica	0	0	4

Correct Rate for Testing Data

Percenta 100%

Testing Data Detail

Species	Sepal.Len	Sepal.Wi	Petal.Len	Petal.Wid	node	predicted.	predicted.	predicted.virginica
setosa	5.1	3.5	1.4	0.2	2	1	0	0
setosa	4.9	3	1.4	0.2	2	1	0	0
setosa	4.7	3.2	1.3	0.2	2	1	0	0
setosa	4.6	3.1	1.5	0.2	2	1	0	0
setosa	5	3.6	1.4	0.2	2	1	0	0
setosa	4.6	3.4	1.4	0.3	2	1	0	0
setosa	5	3.4	1.5	0.2	2	1	0	0
setosa	4.4	2.9	1.4	0.2	2	1	0	0
setosa	4.9	3.1	1.5	0.1	2	1	0	0
setosa	5.4	3.7	1.5	0.2	2	1	0	0
setosa	4.8	3	1.4	0.1	2	1	0	0
setosa	4.3	3	1.1	0.1	2	1	0	0
setosa	5.8	4	1.2	0.2	2	1	0	0
setosa	5.7	4.4	1.5	0.4	2	1	0	0
setosa	5.1	3.5	1.4	0.3	2	1	0	0
setosa	5.7	3.8	1.7	0.3	2	1	0	0
setosa	5.1	3.8	1.5	0.3	2	1	0	0
setosa	5.4	3.4	1.7	0.2	2	1	0	0
setosa	5.1	3.7	1.5	0.4	2	1	0	0

圖4-35　決策樹分析結果報表

3、「SVM」工作表

此工作表為利用支持向量機分析之結果，網頁預覽如圖 4-36 所示。

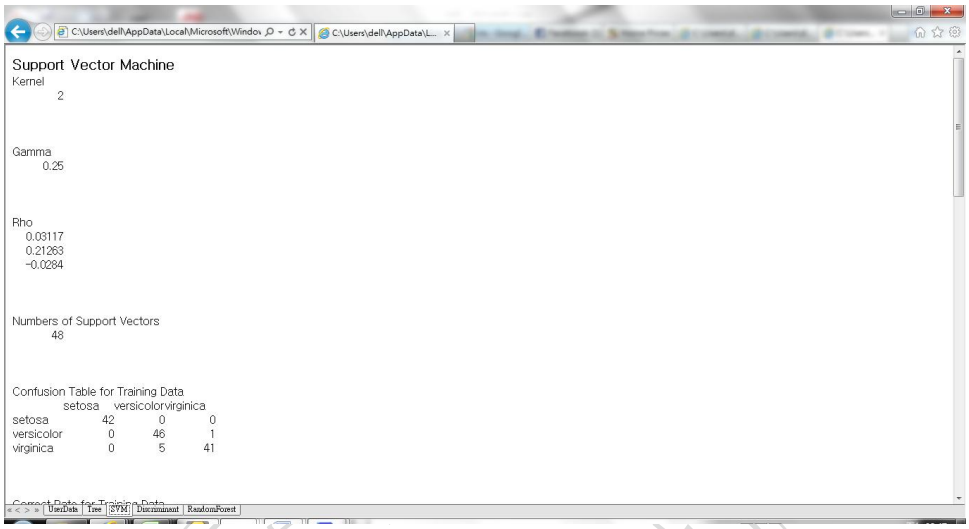


圖4-36 「SVM」工作表

支持向量機分析結果報表如圖 4-37 所示，其中有九個部分，分別為：

(1) Kernel

(2) Gamma

(3) Rho

(4) Numbers of Support Vectors

(5) Confusion Table for Training Data

(6) Correct Rate for Training Data

(7) Confusion Table for Testing Data

(8) Correct Rate for Testing Data

(9) Support Vector

圖 4-37 第 9 個部分的 Support Vector 因為資料筆數較多，此僅展示其中一部分。

Support Vector Machine

Kernel

2

Gamma

0.25

Rho

0.03117
0.21263
-0.0284

Numbers of Support Vectors

48

Confusion Table for Training Data

	setosa	versicolor	virginica
setosa	42	0	0
versicolor	0	46	1
virginica	0	5	41

Correct Rate for Training Data

Percenta 95.56%

Confusion Table for Testing Data

	setosa	versicolor	virginica
setosa	8	0	0
versicolor	0	3	0
virginica	0	0	4

Correct Rate for Testing Data

Percenta 100%

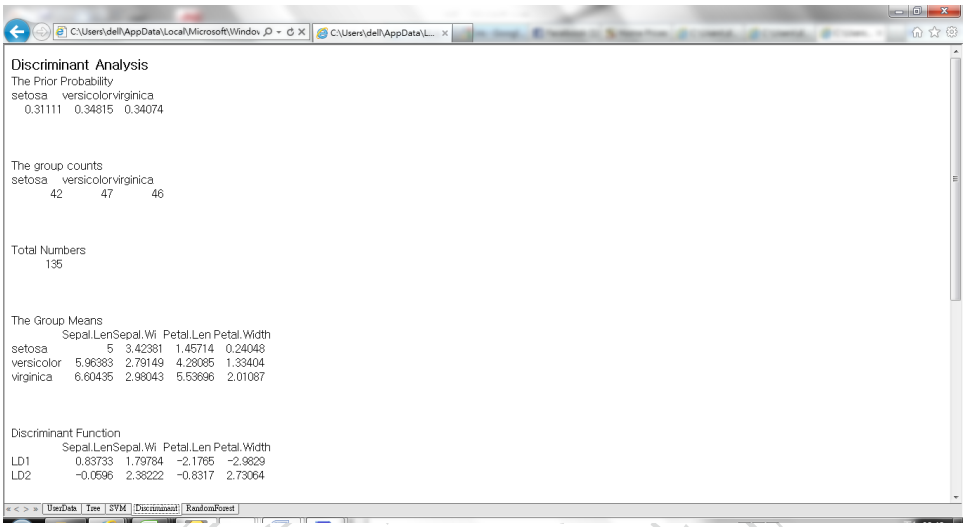
Support Vector

Index	Sepal.Len	Sepal.Wi	Petal.Len	Petal.Width
9	-1.7955	-0.3729	-1.3992	-1.3676
16	-0.2207	3.29286	-1.3416	-1.1006
19	-0.2207	1.82655	-1.2265	-1.2341
23	-1.5532	1.33778	-1.6294	-1.3676
24	-0.9476	0.60463	-1.2265	-0.9671
26	-1.0687	-0.1285	-1.284	-1.3676
32	-0.5841	0.84901	-1.3416	-1.1006
51	1.35404	0.36024	0.50065	0.23436
53	1.2329	0.11586	0.61579	0.36786
54	-0.463	-1.8392	0.09766	0.10087
55	0.74836	-0.6173	0.44308	0.36786
57	0.50608	0.60463	0.50065	0.50136
58	-1.1898	-1.5948	-0.3053	-0.2996
60	-0.8264	-0.8617	0.04009	0.23436
64	0.26381	-0.3729	0.50065	0.23436
67	-0.3419	-0.1285	0.38551	0.36786
69	0.38495	-2.0836	0.38551	0.36786
71	0.02154	0.36024	0.55822	0.76835
73	0.50608	-1.3505	0.61579	0.36786
77	1.11177	-0.6173	0.55822	0.23436
78	0.99063	-0.1285	0.67336	0.63486
84	0.14267	-0.8617	0.73093	0.50136
85	-0.5841	-0.1285	0.38551	0.36786

圖4-37　支持向量機分析結果報表

4、「Discriminant」工作表

此工作表為利用判別分析之結果，網頁預覽如圖 4-38 所示。



Discriminant Analysis
The Prior Probability
setosa versicolor virginica
0.31111 0.34815 0.34074

The group counts
setosa versicolor virginica
42 47 46

Total Numbers
135

The Group Means

	Sepal.Len	Sepal.Wi	Petal.Len	Petal.Width
setosa	5	3.42381	1.45714	0.24048
versicolor	5.96383	2.79149	4.28085	1.33404
virginica	6.60435	2.98043	5.53696	2.01087

Discriminant Function

	Sepal.Len	Sepal.Wi	Petal.Len	Petal.Width
LD1	0.83733	1.79784	-2.1765	-2.9829
LD2	-0.0596	2.38222	-0.8317	2.73064

UserData | Tree | SVM | Discriminant | RandomForest

圖4-38 「Discriminant」工作表

判別分析結果報表如圖 4-39 所示，其中有九個部分，分別為：

(1) The Prior Probability

(2) The Group Counts

(3) Total Numbers

(4) The Group Means

(5) Discriminant Function

(6) Confusion Table for Training Data

(7) Correct Rate for Training Data

(8) Confusion Table for Testing Data

(9) Correct Rate for Testing Data

Discriminant Analysis

The Prior Probability

setosa	versicolor	virginica
0.31111	0.34815	0.34074

The Group Counts

setosa	versicolor	virginica
42	47	46

Total Numbers

135

The Group Means

	Sepal.Len	Sepal.Wi	Petal.Len	Petal.Width
setosa	5	3.42381	1.45714	0.24048
versicolor	5.96383	2.79149	4.28085	1.33404
virginica	6.60435	2.98043	5.53696	2.01087

Discriminant Function

	Sepal.Len	Sepal.Wi	Petal.Len	Petal.Width
LD1	0.83733	1.79784	-2.1765	-2.9829
LD2	-0.0596	2.38222	-0.8317	2.73064

Confusion Table for Training Data

	setosa	versicolor	virginica
setosa	42	0	0
versicolor	0	46	1
virginica	0	5	41

Correct Rate for Training Data

Percenta	95.56%

Confusion Table for Testing Data

	setosa	versicolor	virginica
setosa	8	0	0
versicolor	0	3	0
virginica	0	0	4

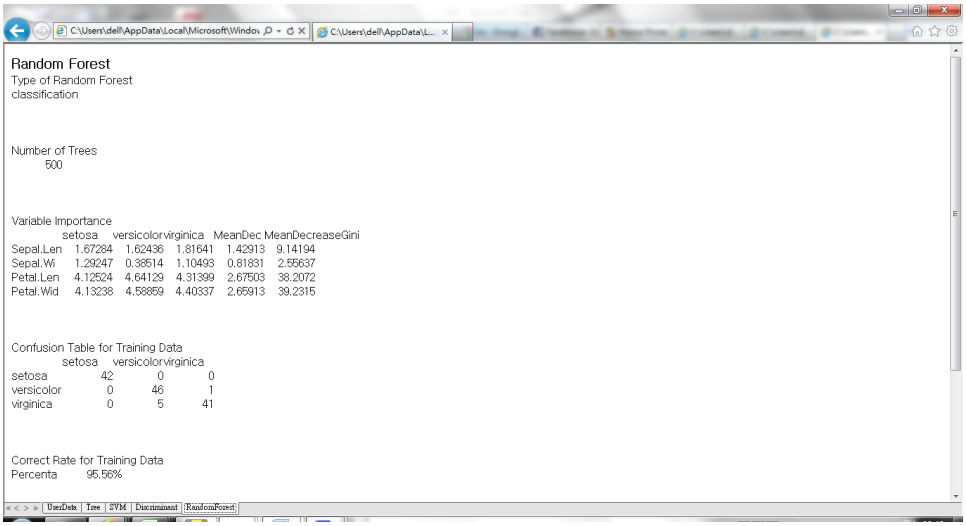
Correct Rate for Testing Data

Percenta	100%

圖4-39　判別分析結果報表

5、「RandomForest」工作表

此工作表為利用隨機森林分析之結果，網頁預覽如圖 4-40 所示。



Random Forest
Type of Random Forest
classification

Number of Trees
500

Variable Importance

	setosa	versicolor	virginica	MeanDec	MeanDecreaseGini
Sepal.Len	1.67284	1.62436	1.81641	1.42913	9.14194
Sepal.Wi	1.29247	0.38514	1.10493	0.81831	2.55637
Petal.Len	4.12524	4.64129	4.31399	2.67503	38.2072
Petal.Wid	4.13238	4.58859	4.40337	2.65913	39.2315

Confusion Table for Training Data

	setosa	versicolor	virginica
setosa	42	0	0
versicolor	0	46	1
virginica	0	5	41

Correct Rate for Training Data
Percenta 95.56%

UserData | Tree | SVM | Discriminant | RandomForest

圖4-40 「RandomForest」工作表

隨機森林分析結果報表如圖 4-41 所示，其中有七個部分，分別為：

(1) Type of Random Forest

(2) Number of Trees

(3) Variable Importance

(4) Confusion Table for Training Data

(5) Correct Rate for Training Data

(6) Confusion Table for Testing Data

(7) Correct Rate for Testing Data

Random Forest
Type of Random Forest
classification

Number of Trees
500

Variable Importance

	setosa	versicolor	virginica	MeanDec	MeanDecreaseG
Sepal.Len	1.67284	1.62436	1.81641	1.42913	9.14194
Sepal.Wi	1.29247	0.38514	1.10493	0.81831	2.55637
Petal.Len	4.12524	4.64129	4.31399	2.67503	38.2072
Petal.Wid	4.13238	4.58859	4.40337	2.65913	39.2315

Confusion Table for Training Data

	setosa	versicolor	virginica
setosa	42	0	0
versicolor	0	46	1
virginica	0	5	41

Correct Rate for Training Data
Percenta 95.56%

Confusion Table for Testing Data

	setosa	versicolor	virginica
setosa	8	0	0
versicolor	0	3	0
virginica	0	0	4

Correct Rate for Testing Data
Percenta 100%

圖4-41　隨機森林分析結果報表

四、　評估模型優劣

選擇文字類別型目標變數所分析出來的結果中，可計算出各個模型中訓練資料集和測試資料集的正確率，並呈現在報表中，使用者可以比較測試資料集的正確率之大小來評估模型的優劣，表 4-6 為決策樹、支持向量機、判別分析及隨機森林的訓練資料集和測試資料集的正確率。分析結果顯示這四種模型各別的訓練資料集和測試資料集的正確率的大小皆相同，分別為 95.56%以及 100%， 表示配適出的決策樹、支持向量機、判別分析及隨機森林都相當的良好。

表4-6　分類模型之正確率比較

模型	決策樹	支持向量機	判別分析	隨機森林
訓練集正確率	95.56%	95.56%	95.56%	95.56%
測試集正確率	100%	100%	100%	100%

第伍章　結論與建議

本章節共分為二節，第一節為「結論」，將總結本研究之內容；第二節為「建議與未來研究方向」，將針對研究期間所發現之問題，提出建議及未來研究方向。

第一節 結論

本研究主要動機為使一般使用者能夠方便應用資料採礦的技術，以及節省時間成本，達到使用者的目的，並以雲端運算為概念，建構一個資料採礦之分類系統。由於欲分類的目標變數有不同型態，本研究將其分為三種：數字連續型、數字類別型以及文字類別型，此分類系統會依照目標變數型態的不同，而採取不同的分類模型來分析使用者之資料，數字連續型的目標變數將利用決策樹、支持向量機及隨機森林等三種模型分析；而數字類別型和文字類別型則利用決策樹、支持向量機、判別分析及隨機森林等四種模型分析，經系統判斷目標變數型態後，會將資料隨機抽取 90%為訓練集資料，10%為測試集資料，再進行模型分析，其分析結果報表將以網頁預覽的方式呈現給使用者。

以「Babies」資料檔、「Egyptian Skulls」資料檔及「iris」資料檔為例，分別利用數字連續型、數字類別型以及文字類別型三種不同的目標變數型態，上傳至此資料採礦之分類系統進行分析，使用者可以針對連續型目標變數的資料分析結果，利用 MAPE 值評估決策樹、支持向量機及隨機森林模型之優劣，而類別型目標變數的資料分析結果，則可以正確率來評估決策樹、支持向量機、判別分析及隨機森林模型之優劣。

經由實證分析結果，使用者確實能透過簡易的步驟操作此資料採礦之分類系統，使用者可從各別的分析報表中得知各個模型分類的結果，並選擇可解釋資料之最佳分類模型，也可從結果報表中獲取資料之特性，更進一步地可以進行所需

的決策。

第二節 建議與未來研究方向

此資料採礦之分類系統目前只能以網頁預覽分析結果報表的方式呈現給使用者，尚未將系統雲端化，且對於使用者上傳資料仍有諸多限制存在，若欲將實現雲端系統平台之概念，則有賴於強大的程式能力之撰寫者，以下建議可供後續研究者研究參考：

1、 將此資料採礦之分類系統包裝成一個軟體套件，使其線上化成一雲端系統，供使用者透過網路即可取得，實現雲端運算之概念。

2、 目前使用者的資料只能上傳 Microsoft Office Excel 逗點分隔值檔案(.csv)、Microsoft Office Excel 工作表 (.xls)及 Microsoft Office Excel 工作表(.xlsx)三種資料檔至系統，未來可在系統中增設轉檔功能，當使用者上傳文字文件(.txt) 資料檔，系統可自動將資料檔類型轉成 Microsoft Office Excel 逗點分隔值檔案(.csv)、Microsoft Office Excel 工作表 (.xls)或 Microsoft Office Excel 工作表(.xlsx)。

3、 此系統對於目標變數為數字類別型的資料，目前只適用於種類為 5 以下，未來可將此限制放寬或解決此限制。

4、 此分類系統目前只置入了四種模型，未來可再增設分類模型，以提供使用者更多模型的選擇。

5、 目前此分類系統只能對於現有資料做分類，尚無法對新進資料做分類，未來可增設此功能。

參考文獻

一、 中文文獻

1、 尹相志（2006），*Microsoft SQL 2005 資料採礦聖經*，台北市：學貫行銷股份有限公司。

2、 何東隆、李美真（2002），*Excel 2002 VBA 與進階應用*，台北市：文魁資訊。

3、 李至和（2006），爭搶 e 商機　特力屋　虛擬商城開張，*經濟日報*，A11 版企業要聞。

4、 李開文（2008），雲端運算、格網運算與 P2P 運算(中)〔電子版〕，2011 年 2 月 28 日，取自台灣中等學校資訊管理人學會 https://sites.google.com/a/tsima.org.tw/www/wiki/dian-zi-qi-kan/2008nian10yue-hao/yun-duan-yun-suan-ge-wang-yun-suan-yup2p-yun-suan-zhong#TOC-2。

5、 林東清（2003），*資訊管理：e 化企業的核心競爭能力*（第四版），台北市：智勝文化事業有限公司。

6、 洪士吉（1999），*操作輕鬆巨集通：Excel VBA 巨集逐步操作指引*，台北市：旗標出版股份有限公司。

7、 洪嘉興（2003），資料挖掘(Data Mining)淺析，*華控月刊*，2003011，17-25。

8、 桂思強（1995），*深造 Excel VBA*，台北市：博碩顧問。

9、 曾憲雄、蔡秀滿、蘇東興、曾秋蓉、王慶堯（2007），*資料探勘 Data Mining*，台北市：旗標出版股份有限公司。

10、 雲端運算使用案例討論小組（2010），*雲端運算使用案例白皮書*（第三版）。

11、 鄒淑文（2010，3 月 19 日），雲端平台：整合看不到的價值，*聯合報經濟日報*，電子通訊。

12、 蔡士源（2004），*Excel VBA 語法字典*，台北市：文魁資訊。

13、 謝至恩（2008），服務在雲端，*NMT* 第 201 期。

14、 謝邦昌、鄭宇庭、蘇志雄（2009），*Data Mining 概述－以 Clementine12.0 為例*，新北市：中華資料採礦協會。

15、 曠文琪（2006），平價超級電腦正接手我們的生活，*商業周刊*第 972 期。

16、 瀨戶遙（2002），*嗯!Excel VBA 我也會 PRO. 2000/2002 對應*，新北市：博碩文化。

17、 蘇芷萱（2011），中華電佈局智慧雲端產業 喊出今年通信業務成長 10 億元目標，2011 年 2 月 27 日，取自華視新聞網網址。http://news.cts.com.tw/cnyes/money/201101/201101110650346.html。

二、 英文文獻

1、 Breiman, L. (1984), *Classification and regression trees*, Wadsworth Internatonal Group.

2、 Breiman, L. (2001), Random Forests. *Machine Learning*, 45 (1): 5-32.

3、 Crawley, M. J. (2007), *The R book*. Chichester, West Sussex, England ; Hoboken, N.J.: Wiley.

4、 Fayyad, U., G. P. Shapiro and P. Smyth (1996), From Data Mining to Knowledge Discovery in Database. *AI Magazine, 17*, 37-54.

5、 Foster, I. and C. Kesselman (2004), *The grid : blueprint for a new computing infrastructure*. Elsevier Inc.

6、 Frawley, W. J., G. Gregory, P. S. Matheus and C. J. Matheus (1991), *Knowledge Discovery in Databases: an Overview in Knowledge Discovery in Databases.* Cambridge, MA: AAAI/MIT, 213-228.

7、 Frawley, W. J., S. G. Paitetsky and C. J. Matheus (1996), Knowledge Discovery in Databases: An Overview. *Communications of the ACM, 39*, 1-34.

8、 Heiberger, R. M. and N. Erich (2009), *R Through Excel (1st ed.)*. New York : Springer.

9、 Han, J. and M. Kamber (2001), *Data Mining: Concepts and Techniques.* Simon Fraser University, Morgan Kaufmann Publishers.

10、 Hartigan, J. A. (1975), *Clustering algorithms*, New York: Wiley.

11、 Loh, W.Y. and Y. S. Shih (1997), Split selection methods for classification trees, *Statistica Sinica*, 7, 815-840.

12、 Maindonald, J. H. (2007), *Data analysis and graphics using R : an example-based approach* (2nd ed.). Cambridge, U.K. ; New York : Cambridge University Press.

13、 Mell, P. and T. Grance (2009), The NIST Definition of Cloud Computing [Electronic version]. *National Institute of Standards and Technology, 53,* 2-3.

14、 Quinlan, J. R. (1979), *Discovering rules from large collections of examples: a case study. In Michie*. Expert Systems in the Microelectronic Age. Edinburgh, Scotland: Edinburgh University Press.

15、 Ripley, B. D. (1996), *Pattern recognition and neural networks*, Cambridge University Press.

16、 Shannon, C. E. and W. Weaver (1949), *The Mathematical Theory of Communication.* University Illinois Press, Urbana.

17、 Usama, F., G. Gregory, P. S. Matheus and P. Smyth (1996), The KDD Process for Extracting Useful Knowledge from Volumes of Data, *Communications of the ACM*, 39(11), 27-34.

18、 Westphal, R. C. (1998), *Data mining solutions : methods and tools for solving real-world problems,* New York : John Wiley & Sons.

19、 Witten, I. H. and F. Eibe (2005), *Data mining : practical machine learning tools and techniques*, Amsterdam, The Netherlands ; Boston, Mass.

20、 Zhang, C. Q. and S. C. Zhang(2002), *Association rule mining : models and algorithms*, Berlin; New York : Springer.

三、 相關網站

1、 台灣雲端運算產業協會 http://www.twcloud.org.tw/Cloud/introduce4.do/。

2、 發展雲端技術應用 中華攜手趨勢領航新時代 http://news.networkmagazine.com.tw/Construction/2010/04/01/18518/。

3、 Slesforce.com http://www.salesforce.com/tw/。

4、 TOP 500 SUPERCOMPUTER SITES http://www.top500.org/。

5、 iThome Online http://www.ithome.com.tw/itadm/channel.php?tab=1/。

6、 Cross Industry Standard Process for Data Mining http://www.crisp-dm.org/。

7、 The Data and Story Library http://lib.stat.cmu.edu/DASL/DataArchive.html/。