
# 巨量資料分析：概念與實務

國立清華大學竹師教育學院

**Big Data Analytics: Concepts and Practice (Fall 2024)**

College of Education, National Tsing Hua University

教師：郭孟倫 博士 Tonny Menglun Kuo, PhD　　Email：tonny@gapp.nthu.edu.tw
助教：　　教室：GEN II 綜二 420
課程時間：　　課程網站：eeclass
Office hour：Friday 10-12:00 a.m. (By Email)　　學分數：3 學分

**先修建議**：本課程建議修讀過統計學的同學修讀（具統計基礎觀念）。即便學生過去不具備程式或資料分析背景也可以修讀本門課程，但要願意學習寫程式。我們會在課堂上逐步教授初階的程式撰寫技巧。本課程並非設計讓學生掌握資料探勘技巧的方法論，而是強調學生如何在小組實作中能夠掌握資料分析專案的特點，學會實用且帶得走的經驗。

## 課程簡介

領會過統計的魅力，卻不知如何應用在職場、研究與生活中？資料如何以視覺化方法呈現，有哪些操作秘訣？資料探勘有哪些不同種類，該如何評估預測的結果呢？掌握資料驅動/知會的決策對於教育現場、教師教學、學生學習、政策改善有哪些幫助呢？

本課程旨在讓學生運用 R 語言來解決複雜多變的研究及實務問題，讓學生理解統計及機器學習的基本概念後，讓同學能夠在課堂當中實務操作並依照分析目標及利害關係人選用正確的機器學習方法與評估模式。本課程強調解構問題、資料間的關係以及動手實作，學生會學到兩套統計軟體基本功能（R 語言與 Tableau），並可以利用這些軟體的功能來分析巨量資料軟體在教育上的應用。

本門課程共分為資料分析基礎、機器學習方法、期末專題三個部分。在資料分析基礎中，我們將會逐步透過講授與練習的方式來學習 R 語言的基礎與特性、資料特性與資料轉換、基礎畫圖與視覺化。這個部分主要是想幫同學打下資料分析概念的基礎，有助於同學運用進階方法在課程當中。本課程的第二部分介紹機器學習的基本方法與應用，包含資料探勘歷程、維度縮減、多元線性迴歸、評估預測效果、KNN、分類與迴歸樹、神經網絡等方法。同學需了解這些基本方法背後的邏輯，並且嘗試在動手實作當中找出問題並討論。第三部分則由同學組成小組來針對一套或多套資料以專題方式進行分析，同學需要運用本課程學會的方法，並依照不同利害關係人來做出相關的分析結果與行動策略。

本課程主要透過微型講授、練習、小組討論、分組報告、個別作業等多種方式進行。本課程將會使用兩套資料探勘（機器學習）軟體：R 語言與 Tableau。學生將可熟悉相關資料探勘分析技術的基礎套件（package），分析、操作並解讀資料探勘的結果，建立教育資料探勘的基礎。需特別注意的是：學生不僅需要報告資料探勘的結果，且需在本課程的要求與規範下學會將資料探勘的報告轉化為業界及公部門容易理解的語言（實務摘要與行動方案）。

學完這門課程後，同學將能夠分析研究問題所需要的統計模型、運用統計軟體呈現量化資料結果，有助於未來在畢業專題（Senior Project）、產業實習、或實際工作場中以有效率的方式分析統計資料。

**課程目標**

1. 選用正確的統計模型來描述、解釋或預測資料並回答研究問題。
2. 以 R 軟體分析資料，掌握統計資料的分佈與特性。
3. 應用 R 語言進行巨量資料探勘。
4. 與不同背景的同儕合作，在專案中學習如何與跨領域成員溝通。
5. 解讀不同利害關係人的立場跟預期，將資料探勘的結果轉譯為可供參照的簡報、實務摘要與行動方案。

**課程進度表**

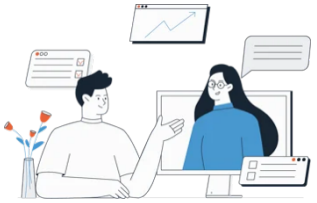
（課表為預排參考，授課教師有權依進度及學生程度滾動調整內容）

| 日期 | 週次 | 主題 | 作業 /備註 |
|---|---|---|---|
| | 1 | **課程概述**<br>基本概念、術語、R 語言的歷史、R 操作環境介紹 | 作業 1：R 環境建置 |
| | 2 | **R 語言基礎**<br>操作介面、變量、資料型態 （data type, levels, vector）<br>matrix, list, array, and data frame, factor, table, string, date, and time | 作業 2：教育大數據分組報告<br>作業 3：Swirl R programing 課程 1-5 |
| | 3 | **資料轉換**<br>dplyr (e.g., selecting, appending, sorting, sampling, filtering, pivoting)<br>Merging, binning, reshaping, date operations, apply fuction | 分組報告 1<br>作業 4：Swirl R programing 單元 10-11 |
| | 4 | **畫圖與相關**<br>types of plots, cor, cor.test<br>Basic plots: scatterplot, histograms, boxplot, density plot ggplot2<br>相關數與共變數 | 分組報告 2<br>作業 5：Swirl R programing 單元 15 |
| | 5 | **資料視覺化簡介**<br>資料類型<br>描述性分析<br>Tableau 軟體簡介與操作 | 分組報告 3<br>作業 6：Tableau 資料視覺化 |
| | 6 | **機器學習方法簡介**<br>核心概念：分類、預測、關聯與推薦、維度縮減、資料探索與視覺化<br>資料分割：檢視、分割、抽樣 | 作業 7 |
| | 7 | **資料探勘歷程**<br>研究目標與機器學習目標<br>處理極端值<br>過度擬和（overfitting） | 分組報告 4 |
| | 8 | **維度縮減**：Dimensionality, aggregation, pivot table, correlation analysis, PCA for classification and prediction)<br>主成分分析（PCA） | 作業 8 |

| 日期 | 週次 | 主題 | 作業 /備註 |
|---|---|---|---|
| | 9 | **多元線性迴歸**<br>解釋性 v.s. 預測性迴歸<br>適配度、多元共線性、非常態資料<br>Exhaustive Search v.s. Partial Search Algorithms<br>Shrinkage - Lasso | 分組報告 5 |
| | 10 | **評估預測效果**<br>測量預測偏誤 MAE、MAPE、RMSE、<br>Gains and lift charts<br>測量分類正確性 Confusion Matrix、Error Rate、<br>Accuracy、Cutoff (threshold)、ROC | 作業 9 |
| | 11 | **期中報告：**<br>小組概念發想提案<br>資料來源、機器學習目的、潛在效果<br>期末評量尺規 | |
| | 12 | **K-Nearest-Neighbor (KNN)**<br>Euclidean distance<br>Choosing k<br>Using K-NN for Prediction | 分組報告 6 |
| | 13 | **分類與迴歸樹**<br>切割方法<br>測量不純度（Impurity）: Gini Index、Entropy<br>Regression Trees<br>Random Forests and Boosted Trees | 作業 10 |
| | 14 | **神經網絡（Neural Nets）**<br>Input, hidden, and output layers<br>Weights<br>Specify Network Architecture<br>Overfitting<br>Deep Learning | 分組報告 7 |
| | 15 | **小組個別討論**<br>資料品質、回饋報告 | 組別個別討論 |
| | 16 | **期末分組口頭報告**<br>建議與修正<br>利害關係人行動策略<br>繳交期末報告 | 繳交簡報檔案，一週後繳交修正後紙本報告 |

**課程評量**

| | |
|---|---|
| 作業（繳交 10 次個人小作業，須附上程式碼） | 40% |
| 分組報告（每次一組報告） | 20% |
| 課堂討論參與+小考 | 20% |
| 期末小組專題 | 20% |



**評量注意事項：**

1. 除特殊原因或課程進度關係外，每份作業/報告都需要在期限內繳交。前期的作業以回家練習＋截圖證明為主。後期的作業主要是在課堂上給同學小任務，利用剛剛學習過的概

念或程式碼進行修改。

2. 分組報告的內容可能會是書籍、期刊文章或是老師指定的讀物。每次會安排一小段時間給同學進行分組報告。
3. 同學應積極參與課堂討論，增加課堂互動性。老師會不定時在講授期間或學習活動上增加討論的議題，讓同學能夠更加理解課程內容。
4. 課堂小考採用 Slido 測驗，主要測驗課程的重要概念內容，主要是上週講述的內容，只要掌握大概念即可作答。

**生成式 AI 倫理聲明**

基於透明與負責任的原則，本課程鼓勵學生利用 AI 進行協作或互學，以提升本門課產出品質。根據本校公布之佈的「大學教育場域 AI 協作、共學與素養培養指引」，本門課程採取：「**有條件開放，請註明如何使用生成式 AI 於課程產出**」

學生須於課堂作業或報告中的「標題頁註腳」或「引用文獻後」簡要說明如何使用生成式 AI 進行議題發想、程式修改、文句潤飾或結構參考等使用方式。若經查核使用卻無在作業或報告中標明，教師、學校或相關單位有權重新針對作業或報告重新評分或不予計分。本門課授課教材或學習資料若有引用自生成式 AI，教師也將在投影片或口頭標注。修讀本課程之學生於選課時視為同意以上倫理聲明。

**參考書目**

本課程沒有指定用書，但可參考以下書籍。


Field, A. (2022). Discovering Statistics Using R and RStudio. Sage Publications.
（Field 有很好的分析概念的理解，初學 R 的學生很適合讀這本書）

Lander, J. P. (2017). R for everyone: Advanced analytics and graphics (Second edition). Addison-Wesley.

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl, K. C. (2017). Data mining for business analytics: Concepts, techniques, and applications in R. Wiley.
（Shmueli 老師是商業分析的專家，這本入門書雖是英文但是字句淺白，容易閱讀。裡面的程式碼都有附註可以參考，是實用的商業分析入門書籍）

Verhoef, P. C., Kooge, E., Walk, N., & Wieringa, J. E. (2022). Creating value with data analytics in marketing: Mastering data science (Second edition). Routledge.
（Verhoef 列出了許多消費者相關的指標，並說明這些指標如何在數據中創造價值）

Yau, N. (2011). Visualize this: The Flowing Data guide to design, visualization, and statistics. Wiley Publications.

鄭中平、許清芳。(2015)。R 在行為科學之應用。台北：雙葉書廊。