National Taipei University of Technology

Computer Science and Information Engineering

Data Science Principles with Applications on

Educational Data

Spring 2025

Semester Group Project Proposal

**教育數據中的關鍵因素探索：以大學排名為例**

**Exploring Key Factors in Educational Data:**

**A Case Study on University Rankings**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Name: | 張育丞, | 張字青, | 周姿妤, | 黃詳諺, | Duong Van Nhat Quang |
| Sid: | 113598043, | 113598032, | 113598090, | 113598091, | 113998411 |

Date: 04/16/2025

1. **Introduction**

* **About Dataset**

世界上有許多大學，對於如何進行排名是個重大的議題，因政治及多項因素進行干涉，使得目前全世界有著數百種的排名，未統一性的衡量標準，也成為一大困擾，該資料集來自於不同的三個地區或國家進行全球的大學排名，相關資料集依據將參考泰晤士高等教育世界大學排名、世界大學學術排名、世界大學排名中心等，分別英國、中國及沙烏地阿拉伯的統計，資料集擷取2014~2015年期間的排名依據，以求資料完整性。

* **Motivation**

近年來，全球大學排名的重要性日益增加，對高等教育政策、資源分配及院校策略產生了深遠影響。了解大學排名的依據，對於希望提升競爭力與社會影響力的學術機構而言至關重要。教育數據涵蓋學術產出、創新能力、校友成就等多項指標，為此類分析提供了豐富的資料基礎。本專案旨在探討不同教育指標如何共同影響大學排名，挖掘其中的關鍵模式，並提出具體可行的建議，促進未來教育發展。

* **Objectives**

本專案的主要目標包括以下內容：

1. 辨識並分析對全球大學排名影響最顯著的關鍵教育指標。
2. 整合多項特徵（如學術引用量、專利數、發表數量），運用資料科學方法探討其綜合影響。
3. 視覺化指標之間的關聯性與模式，以提升資料的可解讀性。
4. 提供基於數據的分析基礎，協助理解大學的競爭力。
5. **Project Plan and Deadlines**

* **Dataset Preview**

本團隊選擇於Kaggle平台中的「World University Ranking」專案之資料集 [1] 進行專案的使用，原資料集用於探討現今世界大學排名系統公平應問題，然而本團隊將透過以下共14個標頭（Header）進行相關的分析，進一步對於學術相關進行探討。

* + **Header**

|  |  |  |
| --- | --- | --- |
| No. | Header (English) | 標頭（中文） |
| 1 | world\_rank | 世界排名 |
| 2 | institution | 大學名稱 |
| 3 | country | 國家 |
| 4 | national\_rank | 國內排名 |
| 5 | quality\_of\_education | 教育品質 |
| 6 | alumni\_employment | 校友就業 |
| 7 | quality\_of\_faculty | 師資品質 |
| 8 | publications | 發表數量 |
| 9 | influence | 影響力 |
| 10 | citations | 論文被引用次數 |
| 11 | broad\_impact | 廣泛影響力 |
| 12 | patents | 專利數 |
| 13 | score | 綜合得分 |
| 14 | year | 年份 |

* **Related Work**
  + 預計探討議題與方法

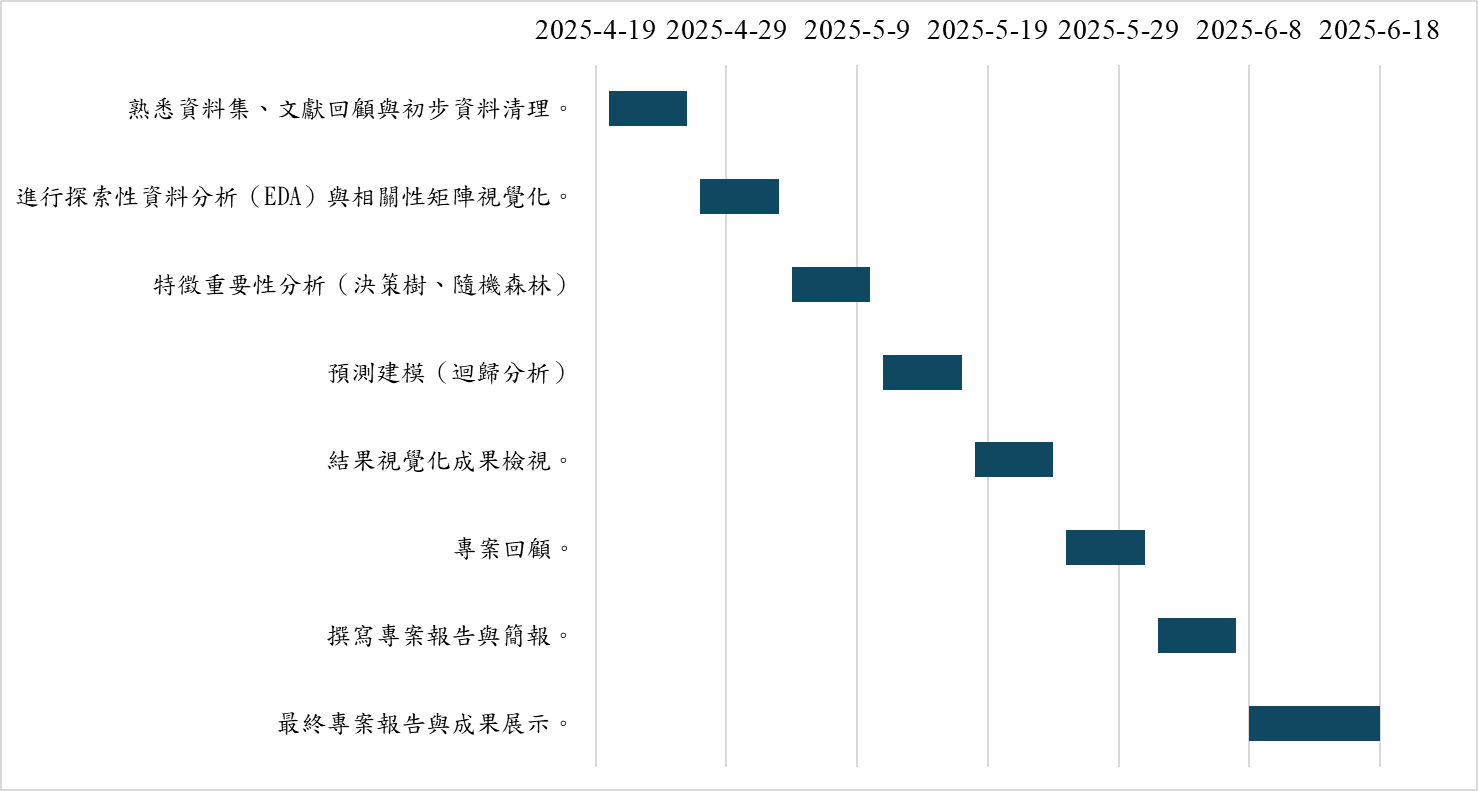
|  |  |  |
| --- | --- | --- |
| 題目 | 問題說明 | 方法 |
| 1. 論文引用量是否影響世界排名？ | 被引用越多，世界排名是否越前？ | 相關分析 + 線性迴歸 |
| 2. 校友就業力是否推動得分？ | 有高就業力的學校，是否得分也會高的？ | 相關分析 + 迴歸 |
| 3. 哪個指標最影響世界排名？ | 教育品質、師資、發表、專利，誰最影響排名？ | 關係性分析 (關係權重) |
| 4. 國家對排名的影響力？ | 特定國家是否在相同分數下排名更高？ | 群組分析 (GroupBy) |
| 5. 高專利產出 vs 高學術影響力大學比較 | 多專利的大學和多引用的大學，誰的排名高？ | 分組比較 |
| 6. 學術 vs 創新型大學的分類 | 根據論文引用、專利將大學分類 | 分群 (Clustering) |

* + Pre-process Dataset
    - 清理遺漏或不一致的資料。
    - 對特徵進行標準化，以確保可比性。
  + 探索性資料分析（EDA）
    - 生成相關係數矩陣，以辨識特徵間的關聯。
    - 視覺化特徵分佈與交互關係。
  + 特徵重要性分析
    - 應用決策樹與隨機森林，找出最具影響力的指標。
    - 使用主成分分析（PCA）掌握主要變異方向。
  + 預測建模
    - 建立迴歸模型（如線性迴歸、決策樹），以關鍵指標預測大學得分或排名。
  + 視覺化
    - 製作相關矩陣圖、特徵重要性圖與迴歸結果圖等視覺化呈現。
* **Expected Results**
  + 分析影響大學排名最重要的關鍵因素。
  + 建立能夠根據特定教育指標預測排名或得分的量化模型。
  + 呈現多個教育指標間關係的視覺化洞見。
  + 提出基於數據分析的建議，供學術機構參考。

1. **Timeline**

我們將從期中考後進行安排專案的進程，我們以每週為檢查點進行相關專案的探討與製作，此外也將其繪製成甘特圖，展現專案工作的細節。

|  |  |
| --- | --- |
| 週數 | 工作內容 |
| 第10週 | 熟悉資料集、文獻回顧與初步資料清理。 |
| 第11週 | 進行探索性資料分析（EDA）與相關性矩陣視覺化。 |
| 第12週 | 特徵重要性分析（決策樹、隨機森林） |
| 第13 週 | 預測建模（迴歸分析） |
| 第14週 | 結果視覺化成果檢視。 |
| 第15 週 | 專案回顧。 |
| 第16 週 | 撰寫專案報告與簡報。 |
| 第17 週 | 最終專案報告與成果展示。 |

****

1. **Reference**

[1] Kaggle, World University Ranking, https://www.kaggle.com/datasets/mylesoneill/world-university-rankings/data