期末專案計畫書

111304019 統計三 林承佑
111304011 統計三 方品皓
111304015 統計三 陳則銓

一、資料來源

使用之資料來源為「數位發展部」主辦、由「聯合行銷研究股份有限公司」執行之《111 年網路沉迷研究調查》，目前由中央研究院社會科學資料庫彙整與發布（計畫編號 AE20004 https://srda.sinica.edu.tw/search/metadata/detail/AE120004）。

二、資料概述

1. 資料收集方式

該調查採用電話訪問法，以加權方式進行抽樣以確保樣本與台灣人口母體在性別、年齡、居住地區的結構一致。調查後以多變項的反覆加權方式進行資料後處理。

2. 何時何地收集

- 於 111 年 12 月 1 日至 12 月 8 日進行
- 利用電腦輔助電話訪問系統（Computer Assisted Telephone Interviewing，簡稱 CATI 系統）以電話訪問的方式蒐集調查資料。

3. 樣本特性

- 年齡範圍：12 歲以上，全齡樣本（含學生、上班族、退休人士等）
- 全國住宅及手機雙電話作為底冊。

4. 資料變項

（1）目標變項（Y）

- **CIAS-10 總分是否 ≥ 28**：作為分類任務的 target 變項。定義為 binary label：1 表示有網路沉迷傾向，0 表示無。(CIAS 是由台大心理系陳淑惠 教授所編製的「陳氏網路沉迷量表」的所寫（Chen Internet Addiction Scale))

（2）自變項（X) （這裡根據問卷的題目分類，以滿足我們探討問題之期待）

- 心理健康因子 Q28 – Q30
  - 憂鬱：以「做事時難以集中注意力」作為指標

- 無聊感：以「覺得生活總是千篇一律或無聊」作為指標
- 課業壓力：以「搞不懂課業或工作要求」為指標

- 疫情影響因素 Q21 – Q27
  - 是否確診或有症狀
  - 是否因疫情失去工作 / 減少工作量
  - 是否有居家上課 / 上班經驗
  - 疫情後上網時間是否增加
- 人口背景 Q31 – Q36
  - 年齡
  - 性別
  - 教育程度（國小以下至研究所以上）
  - 就業狀況（學生、就業者、非就業者）
  - 居住地區（北中南東與離島）

三、探討的問題

「**預測網路成癮傾向（分類問題）**」。我們希望透過統計與機器學習的方法建立模型，預測某一位受訪者是否有網路沉迷的傾向。（不需要做量表即可得知）換言之，「我們能否透過一個人的心理狀態、疫情經歷與人口背景特徵，準確預測其是否有網路沉迷傾向？」

即可能包含以下子問題：

- 哪些心理健康因素（如憂鬱、無聊感、課業壓力）與網路沉迷風險相關？
- 疫情造成的生活改變是否加重網路使用與沉迷風險？
- 是否存在特定職業或年齡族群更容易沉迷？
- 可否藉由人口特徵與心理指標預測網路沉迷高風險族群，協助早期識別與介入？

四、暫定分析計劃

1. 建模架構

我們將研究的問題設定為二元分類的問題，依據是否「網路沉迷」進行預測。最後再對於得出的結果進行討論並找尋更多的改進空間。

2. 資料前處理與 EDA

- 資料清洗與缺值處理：無效回應和處理拒答的部分
- 類別變數：根據使用的模型分別轉為 One-hot encoding 以及 dummy encoding（如性別、教育、就業等），雖然兩種編碼方式導致資料略有不同，但這種設計可讓每種模型發揮其最適特性。
- 數值變數標準化（如心理健康的分數）
- 目標變項建立：依據 CIAS-10 的總分是否 ≥ 28，建立二元分類變項

3. 特徵處理和使用相關模型

- 使用 Ridge 回歸 找出重要性高的變數，並進行特徵選擇
- 邏輯斯迴歸（Logistic Regression）
    - 可以轉換成 Odds Ratio，具有高的可解釋性。
- 隨機森林（Random Forest）
    - 能夠解釋變數重要性，也有決策樹的結構，可以幫助我們處理分類問題。

4. 評估方法

（1）交叉驗證（Cross-Validation）

- 採用 **5-fold 交叉驗證**，將資料隨機分為五等分，輪流進行訓練與測試，以降低單次資料切分所造成的偏誤。有助於穩定模型在不同樣本下的能力。

（2）模型評估指標

- Accuracy：整體正確率，作為初步評估
- Precision：預測為「沉迷」中實際沉迷的比例，能反映 false positive 的情況
- Recall：實際沉迷者中被正確預測的比例
- F1-score：綜合 precision 與 recall 的調和平均
- ROC-AUC：衡量模型對不同閾值下分類能力的整體表現

（3）混淆矩陣（Confusion Matrix）分析

可以快速看出真陽性、假陽性、真陰性、假陰性的比例，有助於理解錯誤型態與模型偏誤。同時，我們也可以針對誤分類造成的潛在風險進行下一步的討論。