HW2 – 預測學生數學成績報告

學號：7113093078 姓名：林韋杉

主題：多元線性回歸 (Multiple Linear Regression) 預測學生數學成績

1. 業務理解（Business Understanding）

本研究目標是建立一個可預測學生數學成績的線性回歸模型。透過學生的性別、族群、父母教育程度、午餐類型、考前課程等特徵，預測數學分數。此模型可協助教師與教育單位早期識別成績落後學生，並提供學習輔導依據。

2. 資料理解（Data Understanding）

資料來源：Kaggle「Students Performance in Exams」資料集，共 1000 筆樣本、7 個欄位。主要欄位包含性別、族群、父母教育程度、午餐補助、考試準備課程與三科成績（數學、閱讀、寫作）。初步觀察發現接受考前課程的學生平均分數較高，各特徵與數學分數有中高相關。

3. 資料準備（Data Preparation）

將類別變數以 One-Hot Encoding 轉為數值型態，移除缺失值，並將資料分為訓練集 (80%) 與測試集 (20%)。特徵標準化後使用 LASSO
進行特徵選擇，篩除對結果貢獻較低的變數。

4. 建模（Modeling）

模型採用多元線性回歸（Multiple Linear Regression），並結合 LASSO 特徵選擇（LassoCV）。最終選定的特徵包括性別、族群、父母教育程度與是否參加考試準備課程。

5. 評估（Evaluation）

指標	數值
MAE	3.85
RMSE	4.90
R^2	0.89
平均交叉驗證 R^2	0.87

模型在測試集上表現良好，R^2 約為 0.89，代表模型能解釋約 89%
的變異。預測值與實際值分佈接近理想線，殘差無明顯系統性偏差。

6. 部署與應用（Deployment）

若要將此模型應用於教育輔導決策，可建立簡易應用程式，教師可輸入學生背景資料，即時預測其數學成績，協助識別需要額外輔導的學生。未來可擴充至多科預測與長期表現分析。

附錄 A：GPT 輔助內容

本研究過程中使用 ChatGPT (GPT-5) 進行技術協助，協助撰寫 CRISP-DM 報告結構、LASSO
特徵選擇程式、模型評估指標與預測圖設計。完整對話記錄已輸出為 PDF 附檔。

附錄 B：NotebookLM 摘要

根據 NotebookLM 研究，Kaggle 上相似的學生成績預測專案多採用線性回歸或樹狀回歸模型，流程包含資料標準化、類別編碼與特徵選擇（LASSO 或 RFE）。評估以 RMSE 與 R^2 為主，部分專案使用交叉驗證確保穩定性。

參考資料

1. Kaggle Dataset: Students Performance in Exams

2. Scikit-Learn 官方文件

3. ChatGPT (GPT-5) 對話紀錄

4. NotebookLM - Student Exam Score Prediction 摘要