**使用 Seq2Seq LSTM 與 Seq2Seq VAE 生成「未來學習行為序列」數據**

## **作業目標**

* 理解 **序列生成模型**（LSTM 與 VAE）的基本設計理念
* 在相同教育資料集上實作並訓練兩種模型
* 比較它們在「未來學習行為序列」上的  
  + 單一路徑預測能力
  + 多樣化生成能力
  + 評估指標（MSE、Best-of-N、Coverage、Diversity）

## **1) 資料**

* **Open University Learning Analytics Dataset (OULAD)**
* [**https://www.kaggle.com/datasets/anlgrbz/student-demographics-online-education-dataoulad?resource=download**](https://www.kaggle.com/datasets/anlgrbz/student-demographics-online-education-dataoulad?resource=download)
* 必須至少包含：  
  + studentInfo.csv
  + studentVle.csv
  + studentAssessment.csv
* 單位：以 **學生 ID** 為基準，建立「每週序列」
* 輸入特徵：  
  + clicks：每週點擊數
  + submit\_cnt / has\_submit：是否提交作業
  + avg\_score\_sofar：作業分數累積平均
  + clicks\_diff1：點擊數差分
* **資料切分**：以學生 ID 分割 train/valid/test，避免資料洩漏

## **2) 模型設計**

### **Seq2Seq LSTM**

* Encoder：讀取過去 4 週特徵序列
* Decoder：輸出未來 2 週的 clicks
* 單一路徑輸出

### **Seq2Seq VAE**

* Encoder：輸入過去序列 → 潛在空間 z，輸出均值 μ 與對數方差 logσ²
* Reparameterization trick：從 z 抽樣
* Decoder：輸入 z，生成未來 2 週序列
* 可生成多條可能的未來（多樣性）

## **3) 訓練設定**

* **固定**
  + Batch size：128
  + Optimizer：Adam（lr=1e-3）
  + Loss：  
    - LSTM：MSE
    - VAE：重建 MSE + β·KLD
* **隨機種子**：seed=42
* **可彈性調整**
  + Epoch（建議至少 5，可增加至 30+）
  + 隱藏層大小（建議 64）
  + Latent dimension（建議 16）
  + 激活函數

## **4) 輸出結果**

* **LSTM**：輸出單一路徑預測
* **VAE**：生成 N 條序列樣本（例如 20 條）
* **可視化**：  
  + Ground Truth（真實未來 clicks）
  + LSTM 單一路徑
  + VAE 多條樣本（半透明曲線）
  + 比較差異（單一 vs 多樣）
* **數值比較**：  
  + LSTM：MSE（單一路徑）
  + VAE：Best-of-N MSE、Diversity（std）、Coverage（比例）
* **分析項目**：  
  + 單一路徑準確度（LSTM vs VAE）
  + 多樣化生成能力（VAE 特有）
  + 覆蓋率（Coverage）
  + 訓練/推理效率

## **作業繳交格式**

* **程式檔** – 包含 LSTM 與 VAE 模型，以 Colab 方式繳交
* **文字分析檔** – Word 檔，需包含：  
  + 模型比較（MSE、Best-of-N、Coverage、Diversity）
  + 視覺化結果
  + 優缺點討論
* **繳交方式**
  + 以 Colab 環境撰寫
  + 上傳到 GitHub
  + 繳交 GitHub 連結  
    



