từ nội dung cô giao, các bạn làm step đầu tiên là **phân tích vấn đề**, xác định **input**, **output**, **các bước thực hiện**, các **kỹ thuật**, **mô hình có thể sử dụng**, **vẽ sơ đồ tổng thể**

từ bộ dataset + đề tài, e phân tích các thuộc tính, xử lý dữ liệu, trực quan

sau đó phân tích lựa chọn mô hình phù hợp để dự đoán sinh viên bỏ học/ kết quả học tập kém/ đánh giá kết quả giảng dạy của gv và đề xuất cải tiến

chủ yếu sẽ cần em có khả năng phân tích, sử dụng các kỹ thuật chuẩn hóa, tiền xử lý dữ liệu

lựa chọn huấn luyện mô hình

# 1. Vấn đề bài toán

* Trực quang hóa dữ liệu đánh giá hiệu xuất học tập của sinh viên
* Lựa chọn mô hình phù hợp dể dự đoán sinh viên có khả năng bỏ học/ kết quả học tập kém/ đánh giá kết quả giảng dại của giảng viên – từ đó đề xuất hướng cải tiến (ví dụ: gợi ý giờ học hợp lý, tư vấn đăng ký tính chỉ phù hợp, có thể tích hợp chat bot hỗ trợ học tập)

# 2. intput

<https://www.kaggle.com/datasets/mahmoudelhemaly/students-grading-dataset?select=Students_Grading_Dataset.csv>

* Bộ dữ liệu này là dữ liệu thực của 5.000 hồ sơ được thu thập từ một nhà cung cấp dịch vụ học tập tư nhân.  
  Bộ dữ liệu bao gồm các thuộc tính chính cần thiết để khám phá các mô hình, mối tương quan và thông tin chi tiết liên quan đến hiệu suất học tập.

**Các cột như sau** :

1. Student\_ID: Mã định danh duy nhất cho mỗi sinh viên.
2. First\_Name: Tên của học sinh.
3. Last\_Name: Họ của học sinh.
4. Email: Email liên hệ (có thể ẩn danh).
5. Giới tính: Nam, Nữ, Khác.
6. Tuổi: Độ tuổi của học sinh.
7. Khoa: Khoa của sinh viên (ví dụ: Khoa học máy tính, Kỹ thuật, Kinh doanh).
8. Tỷ lệ tham dự (%): Tỷ lệ tham dự (0-100%).
9. Midterm\_Score: Điểm thi giữa kỳ (trên thang điểm 100).
10. Final\_Score: Điểm thi cuối kỳ (trên thang điểm 100).
11. Assignments\_Avg: Điểm trung bình của tất cả bài tập (trên thang điểm 100).
12. Quizzes\_Avg: Điểm bài kiểm tra trung bình (trên 100).
13. Participation\_Score: Điểm dựa trên sự tham gia của lớp (0-10).
14. Projects\_Score: Điểm đánh giá dự án (trên thang điểm 100).
15. Total\_Score: Tổng điểm có trọng số của tất cả các điểm.
16. Điểm: Điểm chữ (A, B, C, D, F).
17. Study\_Hours\_per\_Week: Số giờ học trung bình mỗi tuần.
18. Hoạt động ngoại khóa: Học sinh có tham gia hoạt động ngoại khóa hay không (Có/Không).
19. Internet\_Access\_at\_Home: Học sinh có thể truy cập Internet tại nhà không? (Có/Không).
20. Parent\_Education\_Level: Trình độ học vấn cao nhất của cha mẹ (Không có, Trung học phổ thông, Cử nhân, Thạc sĩ, Tiến sĩ).
21. Mức thu nhập gia đình: Thấp, Trung bình, Cao.
22. Mức độ căng thẳng (1-10): Mức độ căng thẳng tự báo cáo (1: Thấp, 10: Cao).
23. Sleep\_Hours\_per\_Night: Số giờ ngủ trung bình mỗi đêm.

# 3. output

* **Biểu đồ** thể hiện hiệu suất học tập (điểm số, tỷ lệ tham dự, căng thẳng) theo khoa, tuổi, hoặc các yếu tố khác.(**trực quan hóa dữ liệu**)
* **Mô hình dự đoán**:
  + Xác suất sinh viên bỏ học (binary: Có/Không).
  + Xác suất kết quả học tập kém (Total\_Score < 60).
  + Đánh giá hiệu quả giảng dạy của giảng viên (dựa trên trung bình điểm lớp, tỷ lệ tham dự).
* **Đề xuất cải tiến**: Gợi ý giờ học hợp lý, số tín chỉ phù hợp, hoặc tích hợp **chatbot** hỗ trợ.

# 4. các bước thực hiện

* **Thu thập** data kaggle ta được file csv (Extract)
* **Làm sạch & chuẩn hóa dữ liệu (T**ransform**)**
* Lưu dữ liệu đã xử lý ra một cơ sở dữ liệu mới (**Load**)
* **Phân tích dữ liệu**: Xem xét các thuộc tính, kiểm tra giá trị thiếu, bất thường
* **Trực quan hóa**: Vẽ biểu đồ (scatter, bar) để phân tích mối quan hệ (ví dụ: căng thẳng vs điểm số, Tương quan giữa điểm và giờ ngủ, Tỷ lệ học sinh yếu).

\*\* phần mô hình em sẽ tìm hiểu kỉ sau do rộng quá\*\*

**dự đoán Xác suất kết quả học tập kém , Xác suất sinh viên bỏ học**

| **Tiêu chí** | **Random Forest** | **Logistic Regression** | **K-Nearest Neighbors (KNN)** | **Neural Networks (MLP)** |
| --- | --- | --- | --- | --- |
| **Loại mô hình** | Ensemble learning (tập hợp nhiều cây quyết định) | Mô hình tuyến tính | Mô hình dựa trên khoảng cách (distance-based) | Mô hình phi tuyến gồm nhiều lớp và nơ-ron nhân tạo |
| **Nguyên lý hoạt động** | Tạo nhiều cây quyết định trên các tập con dữ liệu, lấy kết quả theo đa số | Ước lượng xác suất qua hàm sigmoid và tìm ngưỡng phân loại | Tính khoảng cách (thường là Euclidean) từ điểm cần dự đoán đến các điểm gần nhất | Lan truyền dữ liệu qua nhiều lớp ẩn để học đặc trưng phi tuyến của dữ liệu |
| **Xử lý dữ liệu mất cân bằng** | Tốt (có thể dùng class\_weight hoặc sampling) | Trung bình (có thể điều chỉnh trọng số) | Kém (nhạy với dữ liệu lệch lớp) | Tốt nếu được huấn luyện đúng cách (dùng trọng số, kỹ thuật data augmentation, v.v.) |
| **Khả năng giải thích** | Tốt (có thể xem feature importance) | Rất tốt (hệ số hồi quy thể hiện rõ vai trò biến đầu vào) | Kém (khó giải thích tại sao dự đoán như vậy) | Kém (hộp đen, khó hiểu với người dùng không chuyên) |
| **Hiệu suất trên dữ liệu lớn** | Tốt, song song tốt | Rất tốt | Kém (tính khoảng cách với toàn bộ tập dữ liệu) | Trung bình đến tốt (tùy kiến trúc mạng và số lớp) |
| **Yêu cầu xử lý đặc trưng (feature)** | Không cần chuẩn hóa, xử lý rất linh hoạt | Cần chuẩn hóa nếu có nhiều tỷ lệ khác nhau | Cần chuẩn hóa dữ liệu (min-max, z-score) | Cần chuẩn hóa dữ liệu (rất quan trọng để huấn luyện hiệu quả) |
| **Tốc độ huấn luyện** | Trung bình (nhanh hơn neural net, chậm hơn logistic) | Rất nhanh | Không cần huấn luyện, nhưng dự đoán chậm khi dữ liệu lớn | Chậm nếu nhiều lớp và nhiều dữ liệu |
| **Tốc độ dự đoán** | Nhanh | Rất nhanh | Chậm nếu dữ liệu lớn (phải tính khoảng cách với tất cả điểm) | Trung bình đến chậm (tùy số lớp) |
| **Độ chính xác (Accuracy)** | Cao, ổn định trên nhiều bài toán | Trung bình đến cao (nếu dữ liệu tuyến tính) | Trung bình (bị ảnh hưởng bởi noise) | Rất cao nếu đủ dữ liệu và huấn luyện đúng cách |
| **Ứng dụng phù hợp** | Hầu hết các bài toán phân loại (đa lớp, nhị phân), đặc biệt với dữ liệu tabular | Phân loại nhị phân, bài toán tuyến tính | Phân loại đơn giản, bài toán nhỏ | Phân loại ảnh, tiếng nói, văn bản, dữ liệu phức tạp, nhận diện mẫu |
| **Khả năng overfitting** | Thấp (do sử dụng nhiều cây) | Trung bình (có thể dùng regularization) | Cao nếu K quá nhỏ | Cao nếu không regularize hoặc dữ liệu không đủ |

**xây dựng chat bot**

* Chatbot AI từ Hugging Face (trả lời tự do giống ChatGPT)
* Mô hình nội bộ (dự đoán sinh viên có nguy cơ bỏ học. tư vấn sâu…)
* Giao diện với Streamlit
* Cuối cùng Kết nối Gắn mô hình với giao diện, xử lý đầu vào & xuất kết quả

# vẽ sơ đồ tổng thể (flowchart)

<https://drive.google.com/file/d/1kC5IbvXnCHZsVtNdjv4ZG1PvBaTjzpVv/view?usp=sharing>

