



# DỰ ĐOÁN HIỆU XUẤT HỌC

# TẬP CỦA SINH VIÊN

Giảng viên hướng dẫn:

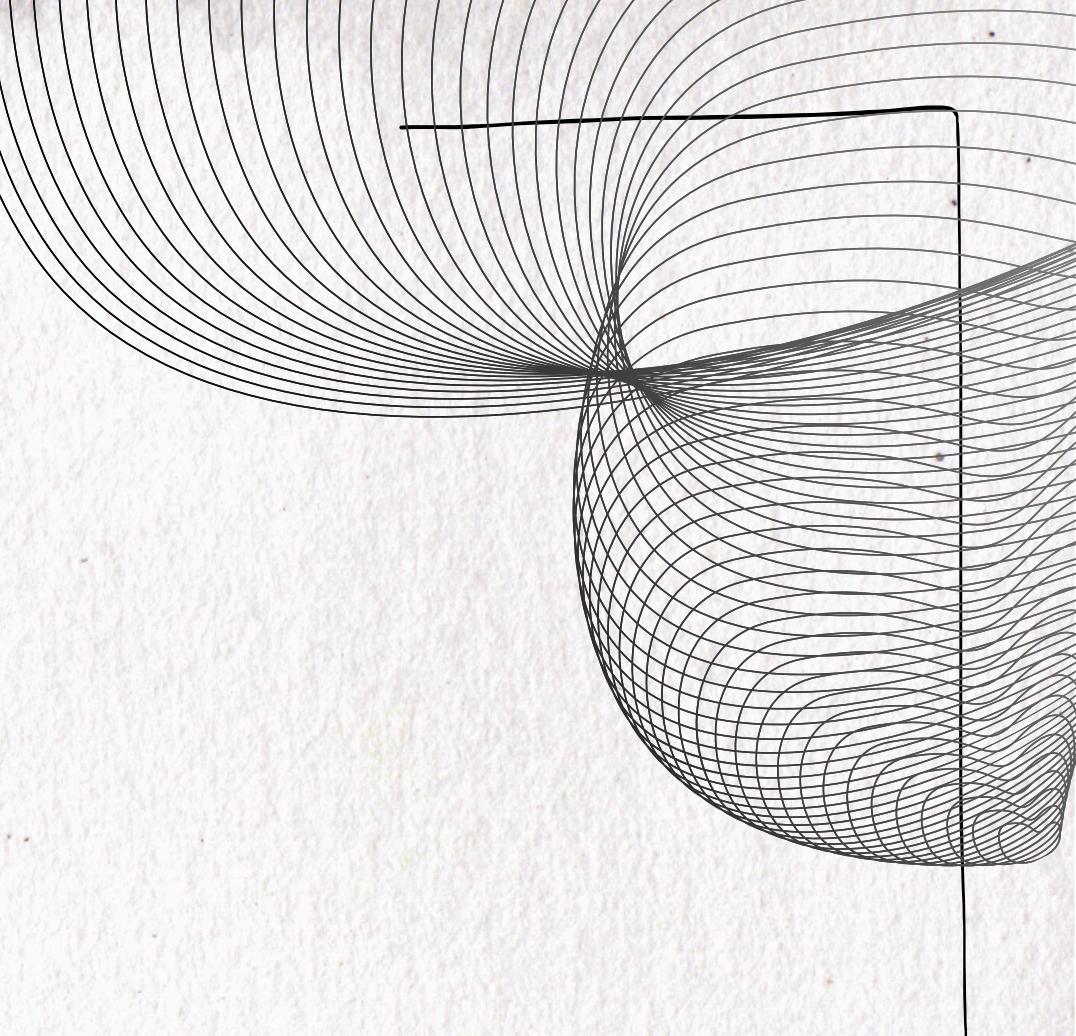
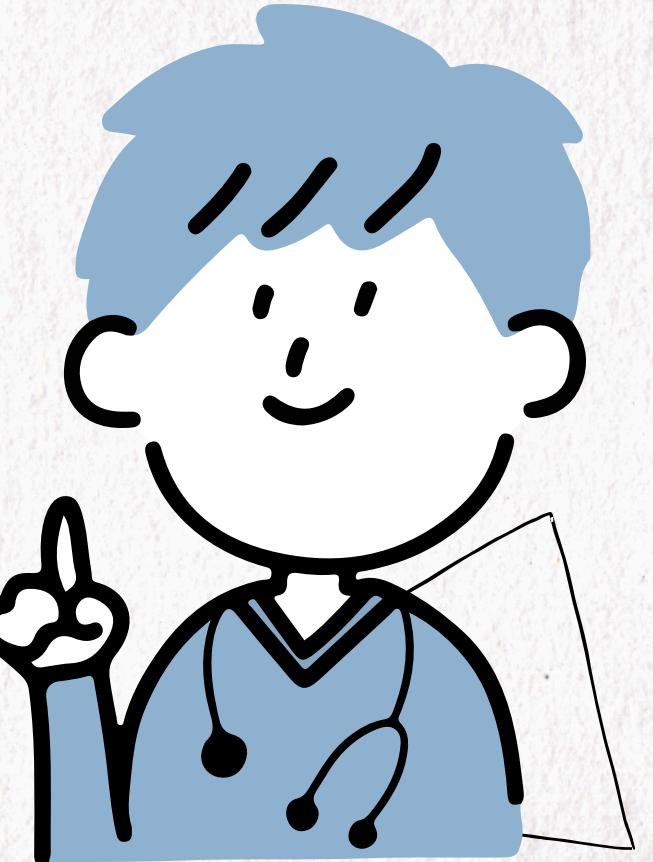
Lê Anh Nhã Uyên

Sinh viên thực hiện:

Huỳnh Chí Phi Thuận – MSSV: KHDL2211038

# Mục Lục

1. Mở đầu – Giới thiệu tổng quan
2. Cơ sở lý thuyết
3. Quy trình xây dựng mô hình CATBOOST
4. Kết luận và hướng phát triển



# 1. Mở đầu – Giới thiệu tổng quan

## Lý do chọn đề tài:

- Có nhiều sinh viên gặp khó khăn về điểm số, lối sống sinh hoạt nhưng chưa nhận ra hậu quả trong thời gian gần
- tạo mô hình nhằm mục đích dự đoán điểm số ở kì tiếp theo của sinh viên
- từ đó hỗ trợ kịp thời và nâng cao chất lượng đào tạo

# 1. Mở đầu – Giới thiệu tổng quan

- **Mục tiêu:** Xây dựng hệ thống dự đoán kết quả học tập kì tiếp theo của sinh viên (Trích xuất đặc trưng: 23 đặc trưng về học tập, lối sống của sinh viên.)
- Đối tượng áp dụng: Sinh viên, giảng viên, cố vấn học tập nhằm dự đoán hiệu suất và cải thiện điểm yếu.



## 2. Cơ sở lý thuyết

### GIỚI HIỆU MÔ HÌNH: CATBOOST

- CatBoost là một bộ công cụ mã nguồn mở được phát triển bởi Yandex, rất phổ biến cho kỹ thuật gradient boosting trên cây quyết định
- CatBoost có thể được áp dụng để giải quyết nhiều bài toán học máy khác nhau, bao gồm:
  - Phân loại (Classification) – Dự đoán nhãn hoặc nhóm của dữ liệu.
  - Hồi quy (Regression) – Dự đoán giá trị liên tục.
  - Xếp hạng (Ranking) – Sắp xếp các đối tượng theo mức độ phù hợp hoặc ưu tiên.

# 2. Cơ sở lý thuyết

## vì sao chọn: CATBOOST

bảng so sánh với các mô hình khác:

Thuộc tính	CatBoost	LightGBM	XGBoost
Xử lý đặc trưng phân loại	Tự động	Hỗ trợ cơ bản	Cần tiền xử lý
Chiến lược chia cây	Đối xứng	Theo lá	Theo độ sâu
Tốc độ & hiệu suất	Ôn định và tối ưu với dữ liệu vừa & nhỏ, đặc biệt khi có nhiều đặc trưng phân loại	Rất nhanh với tập dữ liệu lớn, tối ưu cho hàng triệu bản ghi	Nhanh, linh hoạt, mở rộng tốt trên CPU & GPU

### lý do chính chọn mô hình này"

- Xử lý đặc trưng phân loại tốt mà không cần One-Hot Encoding → tiết kiệm thời gian tiền xử lý.
- Hiệu suất ổn định ngay cả khi dữ liệu không quá lớn hoặc không cân bằng.
- Chống overfitting tốt nhờ kỹ thuật Ordered Boosting độc quyền.
- dữ liệu chỉ 5000 mẫu, nhưng vẫn đạt được kết quả tốt ở thuật toán CatBoost.

### 3. Quy trình xây dựng mô hình CATBOOST

#### 3.1. Quá trình chuẩn bị dữ liệu

- Nguồn dữ liệu: Kaggle – 5.000 sinh viên, 23 thuộc tính về học tập, xã hội, cá nhân. (Student\_ID,First\_Name,Last\_Name,Email,Gender,Age,Department,Attendance (%),Midterm\_Score,Final\_Score,Assignments\_Avg,Quizzes\_Avg,Participation\_Score...)
- **Tiền xử lý:**
- Làm sạch dữ liệu, xử lý giá trị thiếu (Parent\_Education\_Level → “Unknown”).
- Loại bỏ thông tin nhận dạng (ID, tên, email).
- Lý do chọn thuộc tính: Dùng các yếu tố gián tiếp (thói quen học, điều kiện sống) để dự đoán trước khi có điểm thi.
- **Kết quả:** Bộ dữ liệu hoàn chỉnh, đồng nhất, sẵn sàng cho huấn luyện mô hình.

# 3. Quy trình xây dựng mô hình CATBOOST

## 3.2. cách mô hình mã hóa dữ liệu

- CatBoost nhận diện các cột dữ liệu phân loại (ví dụ: giới tính, ngành học) và áp dụng kỹ thuật mã hóa đặc biệt để tránh rò rỉ dữ liệu (data leakage), một vấn đề thường gặp khi sử dụng trung bình giá trị mục tiêu (target mean) truyền thống.

### Vấn đề Rò rỉ Dữ liệu

Nếu dùng trung bình giá trị mục tiêu cho mỗi category, mô hình có thể vô tình "**nhìn thấy**" thông tin của chính dòng dữ liệu đang được xử lý, dẫn đến overfitting.

### Giải pháp của CatBoost: Expanding Mean Target Encoding

CatBoost sắp xếp dữ liệu theo một thứ tự ngẫu nhiên (permutation). Với mỗi dòng, giá trị mã hóa được tính bằng trung bình giá trị mục tiêu chỉ dựa trên các dòng **trước đó** trong thứ tự này. Điều này đảm bảo tính "tương lai" của dữ liệu không bị rò rỉ vào quá trình mã hóa.

# 3. Quy trình xây dựng mô hình CATBOOST

## 3.3. Mô hình CatBoost chưa tinh chỉnh (Baseline)

- **Phiên bản baseline** của mô hình CatBoost được huấn luyện với cấu hình mặc định để thiết lập điểm chuẩn ban đầu. Đây là nền tảng để đánh giá hiệu suất trước khi tiến hành tối ưu hóa chuyên sâu.



# 3. Quy trình xây dựng mô hình CATBOOST

## 3.4. Mô hình CatBoost tinh chỉnh tham số với RandomizedSearchCV

- Để tối ưu hóa hiệu suất của mô hình CatBoost, chúng tôi đã áp dụng phương pháp RandomizedSearchCV.
- Phương pháp này được ưu tiên hơn GridSearchCV do khả năng giảm đáng kể thời gian huấn luyện bằng cách chỉ thử ngẫu nhiên, đặc biệt phù hợp với mô hình CatBoost có nhiều siêu tham số cần điều chỉnh.

### • Phương pháp lựa chọn

Chọn RandomizedSearchCV thay vì GridSearchCV để tối ưu hóa thời gian và hiệu quả tìm kiếm.

### • Tiêu chí đánh giá

Sử dụng f1\_weighted thay vì accuracy để cân bằng ảnh hưởng giữa các lớp, đặc biệt quan trọng với dữ liệu không cân bằng.

### • Kiểm định chéo

Áp dụng cv = 3 cho kiểm định chéo, đảm bảo mô hình có khả năng tổng quát hóa tốt trên các tập dữ liệu khác nhau.

Các tham số tinh chỉnh được lựa chọn kỹ lưỡng để khai thác tối đa tiềm năng của mô hình:

#### Iterations

[100, 300, 500]: Số vòng lặp, ảnh hưởng trực tiếp đến độ sâu quá trình học của mô hình.

#### Depth

[4, 6, 8, 10]: Độ sâu tối đa của cây quyết định, kiểm soát khả năng học mồi quan hệ phức tạp.

#### Border Count

[32, 64, 128]: Số lượng ngưỡng phân chia giá trị liên tục, ảnh hưởng đến khả năng phân loại tinh vi.

#### Learning Rate

[0.01, 0.05, 0.1, 0.2]: Tốc độ học, quyết định mức độ điều chỉnh trọng số sau mỗi bước.

#### L2 Leaf Reg

[1, 3, 5, 7]: Tham số regularization, giúp mô hình tránh hiện tượng overfitting.

#### Class Weights

[5.0, 1.0, 1.0, 10.0]: Duy trì để ưu tiên các lớp thiểu số như A (học sinh giỏi) và D (học sinh yếu).

# 4. Kết luận và hướng phát triển

## 4.1. KẾT QUẢ ĐẠT ĐƯỢC

- Mô hình CatBoost được huấn luyện để dự đoán Grade của sinh viên (A/B/C/D) dựa trên các thông tin cá nhân (giới tính, tuổi, khoa, trình độ học vấn của cha mẹ, mức thu nhập gia đình, tỷ lệ tham dự, điểm giữa kỳ, điểm bài tập)
- mô hình đưa ra dự đoán điểm ở học kì tiếp theo và các lời khuyên bổ ích phụ thuộc vào các thuộc tính như số giờ ngủ mỗi ngày, số giờ học mỗi tuần để cải thiện hiệu suất học tập.
- Giao diện streamlit giúp người dùng dễ dàng sử dụng

# Kết luận và hướng phát triển

## 4.2. Hướng phát triển

- Mở rộng dữ liệu: Thu thập & huấn luyện trên tập lớn hơn, đa dạng trường/khu vực → tăng khả năng khái quát.
- Dashboard tương tác: Tích hợp Power BI/Tableau → trực quan hóa kết quả & xu hướng, hỗ trợ ra quyết định nhanh.
- Chatbot hỗ trợ học tập: Tư vấn cá nhân hóa, gợi ý tài nguyên & hướng cải thiện điểm số dựa trên dự đoán.

## 4.3. Hạn chế

- Dữ liệu hạn chế: Chưa phản ánh đầy đủ đa dạng vùng miền & điều kiện học tập.
- Thiếu yếu tố: Chưa thu thập động lực cá nhân & môi trường gia đình sâu.
- Chưa kiểm nghiệm thực tế: Độ chính xác cần kiểm chứng trên dữ liệu thời gian thực.
- Chưa giải thích sâu: Chưa áp dụng SHAP values ở mức từng học sinh.

Đ E M O C O D E

L à S ả M

P h â M

**Cảm ơn cô và các  
bạn đã lắng nghe**