BÁO CÁO NGHIÊN CỨU: ỨNG DỤNG TRÍ TUỆ NHÂN TẠO TRONG PHÂN TÍCH DỮ LIỆU HỌC TẬP (LEARNING ANALYTICS)

I. THỰC HÀNH PHÂN TÍCH DỮ LIỆU HỌC TẬP (LEARNING ANALYTICS CASE STUDY)

Nghiên cứu này tiến hành phân tích trên một tập dữ liệu mô phỏng gồm 500 sinh viên tham gia khóa học trực tuyến "Cơ sở dữ liệu" trong thời gian 10 tuần trên hệ thống Quản lý học tập (LMS). Dữ liệu thu thập bao gồm: tần suất tương tác (số lần click), thời gian xem video, điểm số các bài kiểm tra ngắn (Quiz), và trạng thái hoàn thành môn học.

1. Phân tích mẫu tương tác và hành vi học tập (Behavioral Patterns)

Qua trích xuất dữ liệu log-file, hành vi của sinh viên được chia làm 3 nhóm tương tác chính:

- **Tương tác tài nguyên:** Đọc tài liệu PDF, bài giảng điện tử (chiếm 45% tổng lượng click).
- **Tương tác video:** Xem và xem lại các video bài giảng (chiếm 35%).
- **Tương tác xã hội:** Thảo luận trên Forum, đặt câu hỏi cho giảng viên (chiếm 20%).

Dữ liệu cho thấy một mô hình hành vi rõ rệt: Lượng truy cập hệ thống đạt đỉnh vào các ngày Thứ Ba và Thứ Tư (ngay sau buổi học lý thuyết) và giảm sâu vào cuối tuần.

2. Xác định các yếu tố dự đoán thành công hoặc khó khăn trong học tập

Bằng cách chạy mô hình hồi quy và ma trận tương quan, nghiên cứu xác định được các biến số có ảnh hưởng lớn nhất đến kết quả cuối kỳ:

- **Yếu tố dự đoán thành công (Chỉ số tích cực):** * *Tần suất phân bổ thời gian:* Sinh viên truy cập đều đặn (>4 ngày/tuần) có tỷ lệ đạt điểm giỏi cao hơn 68% so với sinh viên học dồn vào cuối kỳ.
    - *Sự tích cực trên Forum:* Số bài viết đóng góp trên diễn đàn tỉ lệ thuận với điểm số thực hành
- **Yếu tố dự đoán rủi ro (Chỉ số cảnh báo khó khăn):**
    - *Sự sụt giảm tương tác:* Sinh viên không đăng nhập hệ thống liên tiếp quá 10 ngày từ tuần thứ 3 đến tuần thứ 5 có nguy cơ bỏ học hoặc trượt môn lên tới 82%.

- *Thời lượng xem video ngắn:* Nhóm sinh viên có thời lượng xem video dưới 30% tổng thời lượng bài giảng thường có điểm số Quiz dưới trung bình.

3. Trực quan hóa kết quả phân tích

Dưới đây là bảng tổng hợp mối quan hệ giữa hành vi tương tác trong 4 tuần đầu tiên và kết quả phân tích dự báo:

Nhóm Sinh Viên	**Tần suất Click trung bình/tuần**	**Tỷ lệ hoàn thành Video**	**Điểm Quiz trung bình**	**Dự báo của AI (Tuần 5)**	**Kết quả thực tế cuối kỳ**
Nhóm A (Tích cực)	> 150 lần	85% - 100%	8.5 / 10	Hoàn thành xuất sắc	92% Đạt điểm A/B
Nhóm B (Thụ động)	50 - 120 lần	40% - 70%	6.0 / 10	An toàn nhưng rủi ro nhẹ	75% Đạt điểm C/D
Nhóm C (Nguy cơ)	< 20 lần	< 20%	3.2 / 10	**Cảnh báo nguy cơ trượt**	80% Bỏ học hoặc Trượt

II. CƠ SỞ LÝ LUẬN VÀ NGUYÊN LÝ HOẠT ĐỘNG CỦA AI TRONG GIÁO DỤC

1. Nguyên lý hoạt động của các thuật toán AI

Hệ thống phân tích dữ liệu học tập vận hành thông qua một chu trình khép kín: Thu thập; Xử lý; Phân tích; Dự đoán; Can thiệp.

[Dữ liệu thô từ LMS] [Tiền xử lý & Chuẩn hóa] [Mô hình AI / Thuật toán] [Báo cáo & Cảnh báo sớm]

Để xử lý khối dữ liệu hành vi phức tạp, ba nhóm thuật toán cốt lõi sau được áp dụng:

- **Thuật toán Phân cụm (Clustering - K-Means):** Thuật toán này đo lường khoảng cách hình học giữa các vector hành vi của sinh viên (chẳng hạn như: Số giờ học, Số bài tập đã nộp). Học máy sẽ tự động gom các sinh viên có hành vi tương đồng vào cùng một cụm mà không cần dán nhãn trước, giúp giảng viên nhận diện nhanh các phân khúc học viên trong lớp.
- **Thuật toán Phân lớp (Classification - Decision Tree / Random Forest):** Dựa trên dữ liệu của các khóa học trước, thuật toán xây dựng các cây quyết định dạng "Nếu... Thì...". Ví dụ: *Nếu (Thời gian học < 2 giờ/tuần) và (Điểm Quiz 1 < 5) Thì (Nhãn: Nguy cơ cao)*. Mô hình Random Forest kết hợp hàng trăm cây quyết định để đưa ra dự đoán xác suất trượt môn của một sinh viên với độ chính xác cao.
- **Khai phá luật kết hợp (Association Rule Mining - Apriori):** Thuật toán tìm kiếm các mối quan hệ ẩn giữa các chuỗi hành động. Nó giúp phát hiện ra quy luật: "Sinh viên xem tài liệu hướng dẫn trước khi làm bài Lab thường có điểm số cao hơn 40% so với sinh viên làm bài Lab ngay".

2. Cách AI giúp cải thiện thiết kế khóa học và can thiệp kịp thời

- **Tối ưu hóa cấu trúc học liệu (Data-driven Course Design):** AI phân tích bản đồ nhiệt (Heatmap) và thời gian tương tác trên từng trang tài liệu. Nếu một bài đọc có thời gian dừng lại quá ngắn nhưng điểm bài tập liên quan lại thấp, điều đó chứng tỏ tài liệu quá hàn lâm, nhàm chán hoặc khó hiểu, đòi hỏi thiết kế lại nội dung.
- **Hệ thống cảnh báo sớm (Early Warning System - EWS):** Thay vì đợi đến khi có điểm thi cuối kỳ - thời điểm mọi sự đã quá muộn - chạy ngầm và quét dữ liệu hàng ngày. Ngay khi sinh viên xuất hiện các biểu hiện của "Nhóm C (Nguy cơ)", hệ thống tự động kích hoạt lệnh gửi thông báo cho cố vấn học tập.
- **Cá nhân hóa lộ trình học tập (Adaptive Learning):** AI đóng vai trò như một trợ lý học tập 1:1. Khi sinh viên trả lời sai một câu hỏi về "Vòng lặp trong lập trình", hệ thống AI sẽ tự động điều hướng, gợi ý sinh viên xem lại video ngắn về vòng lặp và cung cấp bài tập ở cấp độ dễ hơn để củng cố kiến thức trước khi bước tiếp.

3. Các vấn đề đạo đức trong thu thập và phân tích dữ liệu học tập

Việc ứng dụng AI vào môi trường giáo dục đi kèm với những thách thức đạo đức nghiêm trọng:

- **Quyền riêng tư và Giám sát quá mức (Privacy & Surveillance):** Việc ghi lại từng click chuột, thời gian rảnh, thậm chí là quét khuôn mặt qua webcam khi thi trực tuyến dễ tạo ra tâm lý lo lắng, bị theo dõi ở sinh viên, biến môi trường giáo dục thành một "nhà tù số".
- **Định kiến thuật toán (Algorithmic Bias):** Nếu dữ liệu dùng để huấn luyện AI trong quá khứ chứa các định kiến (ví dụ: sinh viên vùng sâu vùng xa có tỷ lệ trượt cao do hạ tầng mạng kém), AI có thể tiếp tục dán nhãn "nguy cơ cao" cho các sinh viên mới thuộc khu vực này một cách bất công, làm giảm động lực phấn đấu của họ.
- **Tính minh bạch và Quyền được giải thích (Explainability):** Các mô hình Deep Learning thường là một "hộp đen" (Black Box). Nếu một sinh viên bị hệ thống từ chối cho thi tiếp vì AI đánh giá "gian lận hoặc không đủ điều kiện", nhà trường phải có khả năng giải thích rõ ràng lý do logic đằng sau quyết định đó, chứ không thể mơ hồ đổ lỗi cho "thuật toán quét ra như vậy".

III. KẾ HOẠCH ỨNG DỤNG PHÂN TÍCH DỮ LIỆU HỌC TẬP TẠI TRƯỜNG ĐẠI HỌC VIỆT NAM

Bối cảnh áp dụng: Một trường Đại học công lập tại Việt Nam đang áp dụng phương thức đào tạo tín chỉ và sử dụng hệ thống Moodle làm nền tảng LMS cốt lõi.

1. Loại dữ liệu cần thu thập

Hệ thống sẽ thu thập dữ liệu từ ba nguồn chính, được số hóa và đồng bộ theo thời gian thực:

- **Dữ liệu nền tảng (Static Data):** Điểm thi THPT Quốc gia, điểm GPA các kỳ trước, ngành học, tình trạng học bổng/học phí.
- **Dữ liệu hành vi trực tuyến (LMS Log Data):** Lịch sử đăng nhập, tổng thời gian on-page, tiến độ xem video bài giảng, số lượng câu hỏi thảo luận, thời gian nộp bài bài tập (sớm, đúng hạn, hay muộn).
- **Dữ liệu khảo sát (Perceptual Data):** Khảo sát khảo sát mức độ hài lòng của sinh viên giữa kỳ và cuối kỳ.

2. Phương pháp phân tích và mô hình AI phù hợp

- **Giai đoạn 1 (Tuần 1 - Tuần 3):** Sử dụng phương pháp **Phân tích mô tả (Descriptive Analytics)** để vẽ bức tranh tổng quan về tỷ lệ tương tác của các lớp học phần, báo cáo cho Phòng Đào tạo những lớp có tỷ lệ sinh viên "vắng mặt số" cao.
- **Giai đoạn 2 (Tuần 4 - Tuần 8):** Triển khai mô hình **Random Forest Classification**. Mô hình được huấn luyện dựa trên dữ liệu lịch sử 5 năm của nhà trường để dự đoán xác suất sinh viên có nguy cơ nhận điểm F (trượt môn) vào tuần thứ 4 của kỳ học.
- **Giai đoạn 3 (Học kỳ tiếp theo):** Ứng dụng **Hệ thống gợi ý (Recommendation System)** dựa trên Lọc cộng tác (Collaborative Filtering) để gợi ý các khóa học tự chọn hoặc tài liệu tham khảo phù hợp với năng lực và sở thích của từng sinh viên.

3. Kế hoạch sử dụng kết quả phân tích để cải thiện dạy và học

- **Đối với Giảng viên (Hỗ trợ giảng dạy):** Nhà trường cung cấp một bảng điều khiển (Teacher Dashboard) trực quan. Giảng viên nhìn vào sẽ biết ngay chương nào đang có nhiều sinh viên làm bài sai nhất để dành thời gian giảng lại trên lớp, thay vì dạy dàn trải theo giáo án cũ.
- **Đối với Cố vấn học tập (Can thiệp kịp thời):** Vào tuần thứ 5, hệ thống tự động xuất danh sách "Sinh viên cần hỗ trợ". Cố vấn học tập sẽ chủ động hẹn gặp, tìm hiểu khó khăn (về gia đình, tài chính, hay phương pháp học) để đưa ra giải pháp hỗ trợ, giảm tỷ lệ thôi học.
- **Đối với Sinh viên (Nâng cao nhận thức):** Sinh viên được cung cấp một thanh tiến trình cá nhân (Student Dashboard) so sánh hiệu suất của mình với mức trung bình của lớp. Điều này kích thích tinh thần tự giác và thúc đẩy tư duy tự điều chỉnh (Self-regulated learning).

4. Các biện pháp bảo vệ quyền riêng tư của sinh viên

Để đảm bảo tuân thủ đạo đức và quy định pháp luật về bảo vệ dữ liệu cá nhân tại Việt Nam, nhà trường áp dụng quy trình 4 bước:

- **Mã hóa và Ẩn danh (Anonymization):** Trước khi đưa dữ liệu vào đường ống huấn luyện AI, toàn bộ các thông tin định danh trực tiếp như Họ tên, Mã số sinh viên, Số điện thoại, Email sẽ bị loại bỏ và thay thế bằng một chuỗi Hash ngẫu nhiên (Ví dụ: SV_99182X).
- **Cơ chế đồng thuận chủ động (Opt-in/Opt-out):** Đầu mỗi khóa học, sinh viên sẽ ký một thỏa thuận kỹ thuật số. Sinh viên có quyền từ chối cho phép AI phân tích hành vi click chuột của mình mà không bị phân biệt đối xử hay ảnh hưởng đến quyền lợi học tập.

- **Phân quyền truy cập nghiêm ngặt (Role-based Access Control):** Giảng viên chỉ được xem dữ liệu tổng hợp của lớp mình dạy. Cố vấn học tập chỉ được tiếp cận danh tính thực của sinh viên thuộc diện "Cảnh báo đỏ" sau khi được hệ thống phê duyệt lý do chính đáng.
- **Xóa bỏ dữ liệu (Data Retention Policy):** Dữ liệu hành vi chi tiết (log-click) sẽ tự động được xóa bỏ hoặc lưu trữ vĩnh viễn dưới dạng ẩn danh hoàn toàn sau 1 năm kể từ khi sinh viên hoàn thành học phần.

IV. TÀI LIỆU THAM KHẢO (HARVARD STYLE)

- Baker, R.S. and Inventado, P.S. (2014) 'Educational Data Mining and Learning Analytics', in *Learning Analytics*, Springer, pp. 61-75.
- Conijn, R., Snijders, C., Kleingeld, A. and Matzat, U. (2017) 'Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle', *IEEE Transactions on Learning Technologies*, 10(1), pp. 17-29.
- Siemens, G. (2013) 'Learning Analytics: Envisioning a research discipline and a domain of practice', *American Behavioral Scientist*, 57(10), pp. 1380-1400.