Nội dung web (Chủ yếu là sửa + bổ sung mới, cân nhắc thông tin và chức năng cũ cái nào bỏ, cái nào giữ)

Tên Đề Tài Mới (Chọn 1):

- "Nền Tảng Phân Tích Và Dự Đoán Tỷ Lệ Hoàn Thành Cùng Tối Ưu Hóa Trải Nghiệm Học Tập Dựa Trên Các Phương Pháp Máy Học"
- "Hệ Thống BI Dự Đoán Mức Độ Hoàn Thành Khóa Học Của Học Viên: Tiếp Cận Dựa Trên Dữ Liệu Lớn"
- Nền Tảng Dự Đoán Mức Độ Hoàn Thành Của Học Viên Và Hỗ Trợ Quyết Định Cho Giáo Dục Trực Tuyến

I. Thiết Kế Cơ Sở Dữ Liệu

Các Cột Cần Bổ Sung Cho Các Bảng Chính Yếu Trong Database: (chỉ liệt kê những cái cần bổ sung tương ứng trong csv)

1. **Về Users (Người dùng):**
    - user_school (Tên trường gốc, ví dụ: "Đại học X" - nếu bạn có thể ánh xạ lại từ user_school_encoded)
    - year_of_birth (Năm sinh)
    - gender_original (Giới tính gốc, ví dụ: "Nam", "Nữ" - nếu có thể ánh xạ lại)
2. **Về Courses (Khóa học):**
    - course_name (Tên khóa học - nếu có, hoặc một tên mô tả)
    - course_field_original (Lĩnh vực gốc, ví dụ: "Khoa học Máy tính" - nếu có thể ánh xạ lại từ field_encoded)
    - course_prerequisites_original (Mô tả điều kiện tiên quyết bằng chữ - nếu có, ánh xạ lại từ encoded)
3. **Về UserCourseWeeklyActivity (Hoạt động Hàng tuần của Người dùng - Khóa học):**
    - Làm sao đó để chia ra 4 file cho tho6ngt tin 4 tuần

4. **Về Models (Mô hình): Cần tự tạo thêm bảng này trong database, tham khảo thông tin trong các notebook train trong nhóm tùy tình huống để lấy thông tin làm web**
    - model_id (PK)
    - model_name (Tên mô hình, ví dụ: "Random Forest v1", "Completion Predictor LGBM")
    - model_version (Phiên bản)
    - algorithm_used (Thuật toán sử dụng, ví dụ: "RandomForestRegressor", "GradientBoostingRegressor")
    - training_date (Ngày huấn luyện)
    - performance_metrics_json (Lưu trữ RMSE, MAE, R2, MAPE dưới dạng JSON)
    - feature_importance_json (Lưu trữ các đặc trưng hàng đầu và độ quan trọng của chúng dưới dạng JSON)
    - hyperparameters_json (Lưu trữ các siêu tham số đã sử dụng để huấn luyện mô hình)
    - is_active (Boolean, đánh dấu mô hình đang được triển khai mặc định)

III. Các Trang Web và Nội Dung Chi Tiết

1. Trang Chủ (Landing Page)

- **Tiêu đề chính:** Tên đề tài mới của bạn.
- **Giới thiệu ngắn (Ngữ cảnh):**
    - "Sự bùng nổ của các khóa học trực tuyến mở mang lại cơ hội học tập to lớn cho hàng triệu người. Tuy nhiên, tỷ lệ bỏ học cao vẫn là một thách thức đáng kể. Việc hiểu rõ các yếu tố ảnh hưởng và dự đoán sớm khả năng hoàn thành khóa học của học viên là vô cùng quan trọng để các nhà giáo dục và nền tảng MOOC có thể đưa ra những can thiệp kịp thời, nâng cao hiệu quả giảng dạy và trải nghiệm học tập."
- **Giải pháp của chúng tôi:**

    - "Nền tảng [Tên đề tài] ứng dụng Trí tuệ Nhân tạo và các kỹ thuật Khai thác Dữ liệu tiên tiến trên một kiến trúc dữ liệu lớn để phân tích hành vi học tập và dự đoán mức độ hoàn thành khóa học của người dùng. Chúng tôi cung cấp một giao diện BI trực quan, cung cấp thông tin chi tiết, có thể hành động cho giảng viên, giúp họ hỗ trợ học viên tốt hơn, đồng thời mang đến những gợi ý cá nhân hóa cho người học (trong tương lai)."
- **Đầu vào chính (Đơn giản hóa cho Trang chủ):**
    - Lấy trong file thuyết minh đề tài
- **Đầu ra chính (Đơn giản hóa cho Trang chủ):**
    - Lấy trong file thuyết minh đề tài
- **Lợi ích:**
    - **Đối với Giảng viên:** Chủ động xác định và hỗ trợ học viên gặp khó khăn, tối ưu hóa nội dung và phương pháp giảng dạy.
    - **Đối với Học viên:** Nhận thức rõ hơn về tiến trình học tập của mình, nhận được gợi ý để cải thiện (tính năng tương lai).
    - **Đối với Tổ chức/Nền tảng MOOC:** Cải thiện tỷ lệ hoàn thành khóa học, nâng cao chất lượng giáo dục và sự hài lòng của người học.
- **Nút:** "Khám phá Dashboards", "Tìm hiểu Phương pháp Luận", "Xem Chi tiết Dữ liệu"

2. Trang Phân Tích & Chất Lượng Dữ Liệu (tách riêng ra khỏi trang home)

- **Tiêu đề:** Phân Tích và Đảm Bảo Chất Lượng Dữ Liệu
- **Phần 1: Tổng quan Bộ dữ liệu** (Có thể bảo thư ký làm phần này hoặc lấy thông tin từ các file bài tập quá trình đầu tiên)
    - Nguồn dữ liệu: MOOCCubeX (mô tả ngắn gọn về bộ dữ liệu, quy mô).
    - Các file chính: Liệt kê các file và nội dung chứa của MOOC (readme của github của MOOC)

- Thống kê cấp cao: Tổng số người dùng, khóa học, bản ghi tương tác đã được xử lý,…

- **Phần 2: Đánh giá Chất lượng Dữ liệu (Khía cạnh Cứng & Mềm - Hard & Soft Dimensions) (Phần này làm tạm, nhớ nhắn t để t bổ sung sau)**
    - **Giải thích Hard vs Soft Dimensions:**
        - **Hard Dimensions (Khía cạnh Cứng):** Là các yếu tố có thể đo lường khách quan, thường liên quan đến cấu trúc và tính toàn vẹn của dữ liệu. Ví dụ:
            - **Tính đầy đủ (Completeness):**
                - Hiển thị biểu đồ "Completeness theo từng cột dữ liệu" .
                - Nêu rõ tỷ lệ phần trăm giá trị thiếu cho các cột quan trọng (trước và sau khi xử lý). *Ví dụ: "Cột year_of_birth ban đầu thiếu 20% dữ liệu, sau khi xử lý bằng [phương pháp], tỷ lệ thiếu giảm còn 5%."*
            - **Tính duy nhất (Uniqueness):** Tỷ lệ phần trăm bản ghi trùng lặp (ví dụ: user_id và course_id trùng lặp) được xác định và cách xử lý.
            - **Tính nhất quán (Consistency):** Đảm bảo định dạng dữ liệu đồng nhất (ví dụ: course_id luôn theo mẫu C_xxxx, ngày tháng theo định dạng chuẩn).
            - **(Có thể có) Tính hợp lệ (Validity):** Dữ liệu có nằm trong phạm vi cho phép không (ví dụ: year_of_birth không thể là năm tương lai).
        - **Soft Dimensions (Khía cạnh Mềm):** Là các yếu tố mang tính chủ quan hơn, liên quan đến mức độ phù hợp và hữu ích của dữ liệu đối với mục đích sử dụng. Ví dụ:
            - **Tính liên quan (Relevance):** *Ví dụ: "Các đặc trưng như video_completion_ratio và problem_ratio được đánh*

giá là có tính liên quan cao đến việc dự đoán khả năng hoàn thành khóa học, dựa trên phân tích ban đầu và kết quả từ mô hình."

- **Khả năng diễn giải (Interpretability):** Dữ liệu và kết quả có dễ hiểu đối với người dùng cuối (giảng viên, học viên) không? *Ví dụ: "Chúng tôi đã giải mã hóa các trường như user_school_encoded thành tên trường cụ thể để tăng khả năng diễn giải trên dashboard."*
- **Tính chính xác (Accuracy):** (Liên quan đến ground truth, nếu có) Mức độ chính xác của dữ liệu đầu vào. *Ví dụ: "Dữ liệu về số lượt xem video được coi là chính xác vì được thu thập trực tiếp từ hệ thống."*
- **Tính kịp thời (Timeliness):** (Quan trọng cho hệ thống real-time) Dữ liệu có được cập nhật đủ nhanh để phản ánh tình hình hiện tại không? (Đối với bộ dữ liệu tĩnh, có thể nói về thời điểm thu thập dữ liệu).

- **Hành động đã thực hiện để cải thiện chất lượng dữ liệu:** Mô tả chi tiết các bước làm sạch, điền khuyết (imputation), chuẩn hóa, biến đổi dữ liệu của bạn. *Ví dụ: "Giá trị thiếu trong total_score_week1 được điền bằng giá trị trung bình của cột đó."*

- **Phần 3: Phân tích Khám phá Dữ liệu (EDA)**
  - Nhúng các biểu đồ trực quan hóa dữ liệu quan trọng:
    - Biểu đồ "Phân phối mức độ hoàn thành tổng hợp" (Histogram).
    - Biểu đồ "Tỷ lệ hoàn thành trung bình" cho Video và Bài tập (Bar chart hoặc Pie chart).
    - Các biểu đồ scatter plot "Mức độ hoàn thành vs Số lượng video/bài tập/học viên".
  - Ma trận tương quan (Heatmap) thể hiện mối quan hệ giữa các đặc trưng số quan trọng và biến mục tiêu.

    - Phân phối của các đặc trưng số quan trọng khác (ví dụ: phân phối số lượng khóa học mỗi sinh viên đăng ký).
- **Phần 4: Làm giàu/Tăng cường Dữ liệu (Chức năng: Cải thiện Chất lượng Dữ liệu) (Phần này cũng để tạm, tùy thời bỏ hoặc t bổ sung sau)**
    - **Khái niệm:** Giải thích ngắn gọn về làm giàu dữ liệu (Data Enrichment - thêm các thuộc tính mới từ nguồn bên ngoài hoặc tính toán từ dữ liệu hiện có) và tăng cường dữ liệu (Data Augmentation - tạo thêm dữ liệu mẫu, thường dùng cho tập dữ liệu nhỏ).
    - **Những gì đã làm:** *Ví dụ: "Nhóm đã thực hiện làm giàu dữ liệu bằng cách tính toán các đặc trưng mới như composite_completion (kết hợp tỷ lệ hoàn thành video và bài tập), days_since_enroll (nếu có dữ liệu ngày), hoặc các tỷ lệ tương tác theo tuần."*
    - **Tiềm năng tương lai:** *Ví dụ: "Trong tương lai, hệ thống có thể được mở rộng để tích hợp dữ liệu từ các hệ thống quản lý học tập (LMS) khác, hoặc dữ liệu về thành tích học tập trước đó của sinh viên để làm giàu thêm thông tin đầu vào."*

3. Trung Tâm Mô Hình Dự Đoán (Prediction Model Hub) - Trang Riêng Cho Dự Đoán

- **Tiêu đề:** Thông Tin Mô Hình Dự Đoán
- **Phần 1: Lựa chọn Mô Hình (Cho phép người dùng chọn model - nếu có nhiều model)**
    - Dropdown hoặc danh sách các mô hình đã huấn luyện (lấy từ bảng Models có is_active hoặc tất cả).
    - Hiển thị thông tin cơ bản của mô hình được chọn: Tên mô hình, thuật toán, ngày huấn luyện.
- **Phần 2: Hiệu Suất Mô Hình (Cập nhật dựa trên model được chọn)**
    - Hiển thị rõ các chỉ số đánh giá chính: R^2, RMSE, MAE, MAPE.
    - Biểu đồ so sánh hiệu suất giữa các mô hình (nếu có nhiều).

    - Giải thích ý nghĩa của các chỉ số: *Ví dụ: "$R^2 = 0.946$ cho thấy mô hình giải thích được 94.6% sự biến thiên trong dữ liệu mức độ hoàn thành."*

- **Phần 3: Các Đặc Trưng Quan Trọng Nhất (Cập nhật dựa trên model được chọn)**
    - Hiển thị biểu đồ cột (Bar chart) về Độ quan trọng của Đặc trưng (Feature Importance) cho mô hình đang chọn.
    - Liệt kê 5-7 đặc trưng hàng đầu và giải thích ngắn gọn *ý nghĩa nghiệp vụ* của chúng. *Ví dụ: "Đặc trưng problem_ratio (tỷ lệ hoàn thành bài tập) có độ quan trọng cao nhất, cho thấy việc hoàn thành bài tập là một chỉ báo mạnh mẽ về khả năng hoàn thành toàn bộ khóa học."*
- **Phần 4: Chỉnh sửa & Huấn luyện lại Mô hình (Chức năng: Modify model - Dành cho Quản trị viên)**
    - **Khái niệm:** "Hiệu suất của mô hình có thể suy giảm theo thời gian do sự thay đổi trong hành vi người dùng hoặc nội dung khóa học. Việc huấn luyện lại và cập nhật mô hình là cần thiết để duy trì độ chính xác."
    - **Giao diện (Ý tưởng):**
        - Nút "Huấn luyện lại Mô hình" (chỉ hiển thị cho vai trò admin).
        - Có thể cho phép admin chọn một tập dữ liệu mới (upload file CSV) hoặc chỉ định sử dụng dữ liệu mới nhất trong CSDL.
        - (Nâng cao) Cho phép điều chỉnh một vài siêu tham số cơ bản trước khi huấn luyện lại.
        - Hiển thị log quá trình huấn luyện (đơn giản) và kết quả so sánh với mô hình cũ.
        - Nút "Kích hoạt Mô hình này" để chọn mô hình mới làm mô hình hoạt động chính.
- **Phần 5: (Tùy chọn) Dự đoán Thử Nghiệm**

- Cho phép người dùng (có thể là giảng viên) nhập thủ công các giá trị đặc trưng quan trọng cho một sinh viên giả định để xem kết quả dự đoán. Điều này giúp họ hiểu rõ hơn về cách mô hình hoạt động.

4. Bảng Điều Khiển (Dashboards)

A. Dashboard Chung/Quản trị viên (Admin):
*** KPIs Tổng quan (Hiển thị nổi bật):**
* Tổng số Học viên Đang hoạt động.
* Tổng số Khóa học Hiện có.
* Tỷ lệ Hoàn thành Dự đoán Trung bình (toàn hệ thống).
* Số lượng/Tỷ lệ Học viên Có nguy cơ Cao.
* Số lượng Khóa học có tỷ lệ hoàn thành dự đoán dưới ngưỡng X%.
*** Biểu đồ và Bảng:**
*** Phân phối Tỷ lệ Hoàn thành Dự đoán Toàn Hệ thống:** Biểu đồ Histogram (có thể so sánh với phân phối thực tế nếu có).
*** Top 5 & Cuối 5 Khóa học theo Tỷ lệ Hoàn thành Dự đoán Trung bình:** Biểu đồ cột. (Cho phép click vào để xem chi tiết khóa học đó).
*** Xu hướng Tỷ lệ Học viên Có nguy cơ Theo Thời gian (Tuần/Tháng):** Biểu đồ đường.
*** Bảng: Tóm tắt Toàn bộ Khóa học**
* Cột: Tên Khóa học, Lĩnh vực, Số HV Ghi danh, Tỷ lệ Hoàn thành Dự đoán TB, % HV Có nguy cơ Cao.
* Tính năng: Sắp xếp, tìm kiếm, lọc theo Lĩnh vực.
*** Bản đồ nhiệt (Heatmap):** Mức độ hoàn thành trung bình theo sự kết hợp của Lĩnh vực khóa học và Số tuần học (ví dụ).
*** Bộ lọc:** Khoảng thời gian, Lĩnh vực khóa học, Mức độ rủi ro.

B. Dashboard Giảng viên (Đặc điểm riêng - Giảng viên chọn khóa học mình dạy từ danh sách):
*** KPIs của Khóa học [Tên Khóa học Được chọn]:**
* Tổng số Học viên trong khóa.
* Tỷ lệ Hoàn thành Dự đoán Trung bình của khóa.
* Số lượng/Tỷ lệ Học viên Có nguy cơ Cao trong khóa.
* Tỷ lệ tương tác trung bình (ví dụ: % video đã xem, % bài tập đã nộp).
*** Biểu đồ và Bảng:**

* **Bảng: Danh sách Học viên của Khóa học [Tên Khóa học]**
* Cột: Mã HV (hoặc Tên HV), Điểm Hoàn thành Dự đoán, Mức độ Rủi ro (Cao/TB/Thấp - có màu sắc cảnh báo), Ngày hoạt động cuối, Các yếu tố chính ảnh hưởng (ví dụ: "Ít xem video tuần 2", "Điểm bài tập thấp").
* Tính năng: Sắp xếp theo Điểm dự đoán/Mức độ rủi ro, tìm kiếm HV.
* *Hành động tiềm năng:* Icon "Gửi thông báo/Hỗ trợ" (mở ra một form email/tin nhắn mẫu).
* **Phân phối Điểm Hoàn thành Dự đoán của Khóa học:** Biểu đồ Histogram.
* **So sánh Tiến độ Trung bình của Khóa với Toàn Hệ thống (cho các chỉ số tương tự):** Biểu đồ cột.
* **Các Yếu tố Ảnh hưởng Hàng đầu đến Nguy cơ Bỏ học trong Khóa:** Biểu đồ cột (tổng hợp từ các đặc trưng quan trọng của những sinh viên có nguy cơ).
* **Bộ lọc:** Mức độ rủi ro (chỉ hiển thị HV nguy cơ cao), Tuần học (để xem hoạt động theo tuần).

C. Dashboard Học viên (Đặc điểm riêng - Cá nhân hóa, yêu cầu đăng nhập):
* **Tiêu đề:** Tiến trình Học tập của Bạn
* **Phần "Các Khóa học của Tôi":**
* Hiển thị dạng thẻ (card) cho mỗi khóa học đang tham gia.
* Mỗi thẻ có: Tên Khóa học, Điểm Hoàn thành Dự đoán của bạn, Thanh tiến độ trực quan.
* (Nâng cao) So sánh tiến độ của bạn với trung bình của khóa học.
* Gợi ý hành động: "Bạn nên tập trung vào các video của Tuần 3", "Xem lại bài tập chương X".
* **Biểu đồ Tiến độ Theo Thời gian (cho một khóa học được chọn):** Biểu đồ đường thể hiện điểm số hoặc mức độ hoàn thành dự đoán của bạn qua các tuần.
* **(Tùy chọn) So sánh ẩn danh với các bạn học khác (ví dụ: "Bạn đang ở trong top 20% học viên có tiến độ nhanh nhất").**

5. Chức năng In Báo Cáo BI (Xuất file PDF/Excel)

- Đây là một tính năng có trên các trang Dashboard (đặc biệt là của Admin và Giảng viên).
- **Nút:** "Tải Báo cáo PDF" hoặc "Xuất ra Excel".
- **Nội dung Báo cáo (Tùy chỉnh theo vai trò và bộ lọc đang áp dụng):**

- **Báo cáo Admin:** KPIs tổng quan, các biểu đồ chính từ dashboard admin, bảng tóm tắt các khóa học (có thể là top/bottom N khóa học).
- **Báo cáo Giảng viên (cho một khóa học cụ thể):** KPIs của khóa học, danh sách chi tiết học viên (đặc biệt là nhóm có nguy cơ cao) kèm theo điểm dự đoán và các yếu tố ảnh hưởng, các biểu đồ phân tích của khóa học đó.
- **Định dạng:** Báo cáo cần được trình bày rõ ràng, chuyên nghiệp, có logo (nếu có), ngày giờ xuất báo cáo, các bộ lọc đã áp dụng.

6. Trang Phương Pháp Luận / Giới Thiệu (nhảy từ trang home qua, này sẽ đi sâu hơn về các thông tin thay vì tổng quan như trang home, sẽ bổ sung thông tin trang này sau khi cả nhóm làm xong các phần khác)

- **Quy trình Khai thác Dữ liệu:** Sơ đồ hoặc mô tả các bước từ thu thập dữ liệu (MOOCCubeX), tiền xử lý, xây dựng đặc trưng, huấn luyện mô hình, đánh giá.
- **Mô hình Dự đoán Chi tiết:** Giải thích sâu hơn về thuật toán đã chọn (ví dụ: Random Forest), tại sao nó phù hợp, các siêu tham số quan trọng.
- **Thông tin Nhóm Phát triển.**
- **….**

7. Bổ sung các chức năng (điểm cộng): Cải thiện Data Quality (Làm giàu/ tăng cường), Modify model, cloud (Azure, AWS):

- **Cải thiện Data Quality:** Đã tích hợp vào trang "Phân Tích & Chất Lượng Dữ Liệu" và phần mô tả cách bạn đã làm.
- **Modify model:** Tích hợp vào "Trung Tâm Mô Hình Dự Đoán" như đã mô tả ở trên, tập trung vào khả năng huấn luyện lại và chọn mô hình.
- **Cloud (Azure, AWS):** Tích hợp vào trang "Phương Pháp Luận / Giới Thiệu" với sơ đồ kiến trúc và mô tả.

Kế hoạch từng bước: (Tóm tắt những gì cần làm, 1-6 là bắt buộc, 6-10: bổ sung hết mức có thể tùy tình huống)

1. **Chốt Tên Đề Tài.**

2. **Thiết lập Cơ sở Dữ liệu.**
3. **Phát triển Backend.**
4. **Phát triển Frontend.**
5. **Tích hợp "Hard/Soft Dimensions"** vào nội dung và hình ảnh của trang "Phân Tích & Chất Lượng Dữ Liệu".
6. **Triển khai Hiển thị Dự đoán Cốt lõi:** Hiển thị điểm dự đoán trên các dashboard.
7. **Triển khai các phần "Cải thiện DQ", "Modify Model", "Cloud" (Trước tiên là về mặt khái niệm, sau đó triển khai những gì khả thi):**
    - **Cải thiện DQ:** Tập trung giải thích các quy trình bạn *đã thực hiện* (làm sạch, xây dựng đặc trưng).
    - **Modify Model:** Giải thích khái niệm. Nếu có thời gian, quản trị viên có thể có giao diện để upload siêu tham số mới và huấn luyện lại mô hình (có thể offline), sau đó cập nhật bảng Models.
    - **Cloud:** Mô tả kiến trúc khái niệm trên trang "Giới Thiệu".
8. **Xây dựng Chức năng Xuất Báo cáo BI:** Sử dụng thư viện phù hợp (ví dụ: ReportLab, WeasyPrint cho Python backend; jsPDF cho frontend) để tạo file PDF từ dữ liệu dashboard.
9. **Kiểm thử (Testing):** Kiểm tra kỹ lưỡng việc hiển thị dữ liệu, kết quả dự đoán, các tương tác người dùng.
10. **Triển khai (Deployment - Khái niệm hoặc Thực tế):** Nếu có thời gian, thử triển khai một phiên bản cơ bản lên một dịch vụ đám mây đơn giản (ví dụ: Heroku, PythonAnywhere, Azure App Service free tier).