ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

BÁO CÁO TỔNG QUAN ĐỀ TÀI

DỰ ĐOÁN MỨC ĐỘ HÀI LÒNG CỦA HỌC VIÊN ĐỐI VỚI KHÓA HỌC

GVHD: ThS. Nguyễn Thị Anh Thư

Nhóm thực hiện: Nhóm 8

Thành phố Hồ Chí Minh, 9/2025

1. Giới thiệu

1.1. Thực trạng

Trong bối cảnh giáo dục trực tuyến ngày càng phát triển, các nền tảng học tập online dần thu hút được nhiều học viên. Tuy nhiên, chất lượng khóa học và mức độ hài lòng của học viên là những yếu tố quan trọng quyết định sự thành công và khả năng duy trì lâu dài của các nền tảng này.

Đề tài ***Dự đoán mức độ hài lòng của học viên đối với khóa học*** nhằm mục tiêu xây dựng mô hình phân tích dữ liệu học tập từ bộ dữ liệu MOOC-CubeX, qua đó dự đoán được mức độ hài lòng (phân lớp với 5 mức độ đánh giá), hỗ trợ cho việc cải tiến thiết kế khóa học, tăng tỷ lệ duy trì học viên.

1.2. Ứng dụng

Trong bối cảnh có sự chuyển mình mạnh mẽ về khoa học kỹ thuật, đặc biệt là công nghệ thông tin thì việc chuyển đổi từ giáo dục truyền thống (giảng dạy trực tiếp) có sự chuyển dịch mạnh mẽ sang đào tạo giáo dục trực tuyến

1.3. Khó khăn và thách thức

Bộ dữ liệu MOOC-CubeX cung cấp lượng lớn dữ liệu từ người học, từ đó cho ra nhiều kết quả tích cực. Song, bộ dữ liệu này còn có nhiều hạn chế chẳng hạn như:

- **Từ phía học viên:** Nhiều người học có những trải nghiệm khác nhau về việc học trực tuyến từ đó đưa ra các mức độ hài lòng riêng. Tuy vậy, bộ dữ liệu chỉ đưa ra các thông tin chung về người học, số lượng khóa học mà người đó tham gia hay thời lượng giờ học mà không có các đánh giá mức

độ hài lòng từ người học gây ảnh hưởng lớn đến chất lượng khóa học và không phản ánh đúng thực tế về tình trạng người học.

- **Từ phía nền tảng:** Nền tảng cho ra các kết quả cơ bản về thông tin người học, khóa học hay các thông tin liên quan giữa người học và khóa học. Điều này không phản ảnh đúng chất lượng người học hay toàn bộ khóa học.

2. Bộ dữ liệu sử dụng

2.1. Giới thiệu tổng quan về bộ dữ liệu

Bộ dữ liệu MOOC-CubeX được duy trì bởi Nhóm Knowledge Engineering thuộc Đại học Tsinghua và được hỗ trợ bởi XuetangX, một trong những trang web MOOC lớn nhất ở Trung Quốc. Đây là một kho dữ liệu khổng lồ và đa dạng, được xây dựng để hỗ trợ các nghiên cứu về học tập thích ứng trong môi trường MOOC.

Bộ dữ liệu này bao gồm:

- 4.216 khóa học
- 230.263 video
- 358.265 bài tập
- 637.572 khái niệm chi tiết (fine-grained concepts)
- Hơn 296 triệu bản ghi hành vi thô của 3.330.294 sinh viên

Với quy mô và độ bao phủ cao, MOOC-CubeX cung cấp một nền tảng vững chắc cho các nghiên cứu chuyên sâu. Bộ dữ liệu này nổi bật nhờ các đặc điểm chính sau:

- **Tính bao phủ cao:** MOOC-CubeX tổng hợp nhiều nguồn tài nguyên học thuật đa dạng, từ tài nguyên khóa học đến các bản ghi về hành vi học tập, làm bài tập và thảo luận của sinh viên.

- **Quy mô lớn:** So với các kho dữ liệu giáo dục mở khác, MOOC-CubeX có quy mô lớn hơn đáng kể, phù hợp cho việc khám phá và xây dựng các mô hình học sâu với yêu cầu dữ liệu cao.
- **Tập trung vào khái niệm**: Dữ liệu không đồng nhất được tổ chức theo các khái niệm chi tiết, giúp các tài nguyên trở nên liên quan, dễ dàng được biểu diễn, tìm kiếm và mô hình hóa hơn.

2.2. Mô tả chi tiết về tập dữ liệu

Dữ liệu trong MOOC-CubeX được tổ chức xoay quanh một đồ thị khái niệm chi tiết, cho phép liên kết các thành phần khác nhau của hệ sinh thái MOOC. Các thành phần chính của bộ dữ liệu bao gồm:

- **Dữ liệu về Khóa học:**
    - Thông tin cơ bản về các khóa học.
    - Cấu trúc khóa học, bao gồm các chương, bài giảng và video.
- **Dữ liệu về Người dùng (Sinh viên):**
    - Thông tin nhân khẩu học ẩn danh.
    - Hồ sơ học tập và tiến trình trong các khóa học.
- **Dữ liệu Hành vi:**
    - **Tương tác với video:** Dữ liệu chi tiết về việc xem video, chẳng hạn như thời gian bắt đầu, thời gian kết thúc và các sự kiện tạm dừng.
    - **Tương tác với bài tập:** Lịch sử nộp bài, kết quả và thời gian làm bài.
    - **Tương tác trên diễn đàn:** Dữ liệu về các bài đăng, bình luận và thảo luận.
- **Đồ thị Tri thức (Knowledge Graph):**
    - Một mạng lưới các khái niệm học thuật được liên kết với nhau.

- Các mối quan hệ giữa các khái niệm, chẳng hạn như mối quan hệ tiên quyết (ví dụ: khái niệm A là điều kiện tiên quyết để học khái niệm B).
- Liên kết các khái niệm này với các tài nguyên trong khóa học như video và bài tập.

2.3. Đánh giá bộ dữ liệu

Ưu điểm:

Các bộ dữ liệu MOOC sở hữu nhiều đặc tính quan trọng, có thể khai thác hiệu quả để suy luận về mức độ hài lòng của người học:

- **Khai thác dữ liệu hành vi ngầm**: Điểm mạnh cốt lõi của các bộ dữ liệu này nằm ở khả năng phân tích các chỉ số hành vi ngầm. Từ đó, có thể xây dựng mô hình dự đoán sự hài lòng dựa trên:

-
    - *Mức độ hoàn thành*: Tỷ lệ hoàn thành học phần hoặc toàn bộ khóa học phản ánh rõ ràng cam kết và mức độ hài lòng.

    - *Mức độ tương tác*: Những hành vi như thời lượng xem video, tần suất xem lại các bài giảng khó, số lần thử làm bài kiểm tra, hay sự tham gia trên diễn đàn đều cho thấy mức độ đầu tư thực sự của người học.

- **Khai thác tiềm năng từ dữ liệu văn bản**: Các bài đăng, bình luận trên diễn đàn và phần đánh giá khóa học là nguồn thông tin giá trị để áp dụng kỹ thuật xử lý ngôn ngữ tự nhiên , đặc biệt là phân tích cảm xúc

(Sentiment Analysis). Điều này giúp bổ sung bằng chứng trực tiếp hơn về cảm nhận và sự hài lòng của người học.

- **Quy mô dữ liệu lớn**: Với số lượng người dùng khổng lồ và hàng triệu bản ghi hành vi, các bộ dữ liệu MOOC cho phép huấn luyện những mô hình học máy phức tạp, đủ khả năng nhận diện các mẫu hành vi tinh vi mà những bộ dữ liệu nhỏ hơn khó phát hiện.

2.4. Phân tích thách thức và hạn chế

Bên cạnh tiềm năng, việc khai thác dữ liệu MOOC cho bài toán dự đoán sự hài lòng cũng tồn tại nhiều thách thức cố hữu:

- **Thiếu nhãn dữ liệu trực tiếp (Absence of Direct Labels)**: Đây là trở ngại lớn nhất. Thông thường, các bộ dữ liệu không có trường thông tin rõ ràng phản ánh "sự hài lòng". Buộc phải tạo ra "nhãn đại diện" (proxy labels) dựa trên hành vi, chẳng hạn coi "hoàn thành > 80%" là "hài lòng". Tuy nhiên, cách tiếp cận này mang tính giả định và dễ đưa sai số vào mô hình.

- **Nhiễu trong dữ liệu hành vi (Noise in Behavioral Data)**: Dữ liệu thô chứa nhiều yếu tố gây nhiễu. Ví dụ, một người học bỏ dở khóa học không hẳn vì thiếu hài lòng mà có thể do bận rộn hoặc thay đổi mục tiêu cá nhân. Ngược lại, thời lượng xem video dài cũng không chắc phản ánh mức độ học tập tích cực.

- **Thiên lệch trong mẫu dữ liệu (Data Sampling Bias)**: Phản hồi thường chịu ảnh hưởng bởi thiên lệch, khi chỉ những người học rất hài lòng hoặc rất không hài lòng mới để lại đánh giá. Ngoài ra, dữ liệu thu thập từ một nền tảng duy nhất dễ bị thiên lệch về nhân khẩu học và văn hóa, làm giảm khả năng khái quát của các mô hình được xây dựng.

3. Hướng nghiên cứu triển khai

3.1. Định nghĩa bài toán

Đầu vào: dữ liệu học viên, dữ liệu khóa học, hành vi học tập của học viên và các phản hồi từ học viên.

Đầu ra: mức độ hài lòng của học viên, theo thang từ 1-5:

- 1: Rất không hài lòng
- 2: Không hài lòng
- 3: Bình thường
- 4: Hài lòng
- 5: Rất hài lòng

Mục tiêu: xử lý dữ liệu để làm đầu vào cho các mô hình và tìm ra mô hình tối ưu dự đoán mức độ hài lòng dựa trên đặc trưng từ dữ liệu MOOC-CubeX, sau đó xây dựng ứng dụng để dự đoán trực tuyến và trực quan dữ liệu (nếu có thể).

3.2. Phân tích dữ liệu

Mục tiêu của giai đoạn này là đi sâu vào bản chất của dữ liệu, không chỉ mô tả bề mặt mà còn khám phá các mối quan hệ ẩn, kiểm định các giả thuyết và rút ra

những hiểu biết chiến lược. Những phát hiện này sẽ là nền tảng vững chắc để định hướng cho việc xây dựng đặc trưng và lựa chọn mô hình.

a. Phân tích Đơn biến (Univariate Analysis) – Hiểu rõ từng thuộc tính

Bước đầu tiên là "giải phẫu" từng đặc trưng quan trọng để hiểu phân phối và các đặc tính thống kê của chúng.

- **Phân tích các đặc trưng hành vi (số lượng):**
    - **Mục tiêu:** Xác định xu hướng tương tác chung của người học.
    - **Phương pháp:** Sử dụng **biểu đồ Histogram và biểu đồ mật độ (Density Plot)** cho các biến như số lần **xem video**, **thời lượng xem**, **số lần nộp bài tập.**
    - **Câu hỏi cần trả lời:**
        - Phân phối của các hành vi này có bị lệch không? (Ví dụ: Đa số học viên chỉ xem một vài video rồi nghỉ, hay phân phối đều hơn?)
        - Sử dụng **biểu đồ Boxplot** để xác định các giá trị ngoại lệ (outliers). Những học viên có hành vi cực đoan (xem video nhiều bất thường hoặc làm bài tập hàng trăm lần) là ai và họ có đặc điểm gì?
- **Phân tích các đặc trưng kết quả (tỷ lệ và điểm số):**
    - **Mục tiêu:** Đánh giá hiệu suất học tập chung trên toàn bộ nền tảng.
    - **Phương pháp:** Trực quan hóa phân phối của **tỷ lệ hoàn thành khóa học** và **điểm số trung bình**.
    - **Câu hỏi cần trả lời:**
        - Tỷ lệ hoàn thành trung bình là bao nhiêu? Có bao nhiêu phần trăm học viên hoàn thành trên 80% khóa học? Điều này giúp hình dung về mức độ cam kết chung.

        - Phổ điểm của các bài tập tập trung ở đâu? Điểm số có phân phối chuẩn không hay tập trung ở mức cao/thấp?

b. Phân tích Đa biến (Bivariate & Multivariate Analysis) – Tìm kiếm mối liên kết

Đây là bước cốt lõi để tìm ra những yếu tố thực sự ảnh hưởng đến kết quả và sự hài lòng của học viên.

- **Tương quan giữa Hành vi và Kết quả:**
    - **Mục tiêu:** Định lượng mối quan hệ giữa nỗ lực của học viên và thành quả họ đạt được.
    - **Phương pháp:**
        - Xây dựng **Ma trận tương quan (Correlation Matrix)** sử dụng biểu đồ nhiệt (Heatmap) để xem xét mối quan hệ tuyến tính giữa tất cả các cặp đặc trưng số (ví dụ: **thời lượng xem** với **điểm số**, **số bài đăng trên diễn đàn** với **tỷ lệ hoàn thành**).
        - Sử dụng **Biểu đồ phân tán (Scatter Plot)** để khám phá các mối quan hệ phi tuyến tiềm ẩn.
    - **Câu hỏi cần trả lời:**
        - Việc xem nhiều video hơn có thực sự dẫn đến điểm số cao hơn không? Mối quan hệ này mạnh đến mức nào?
        - Sự tích cực trên diễn đàn có phải là một chỉ báo tốt cho việc hoàn thành khóa học không?
- **Phân tích theo Phân khúc Người dùng (User Segmentation):**
    - **Mục tiêu:** Xác định các "chân dung" (persona) người học điển hình dựa trên hành vi của họ.
    - **Phương pháp:** Sử dụng các thuật toán gom cụm như **K-Means** trên các đặc trưng hành vi chính để phân nhóm người học.

    - **Các chân dung có thể phát hiện:**
        - **"Người học toàn diện":** Tích cực ở mọi mặt (xem video, làm bài, thảo luận).
        - **"Người học thụ động":** Chỉ xem video nhưng ít tương tác.
        - **"Người học tập trung vào mục tiêu":** Bỏ qua video, chỉ tập trung làm bài tập để lấy chứng chỉ.
        - **"Người học bỏ cuộc sớm":** Chỉ hoạt động trong tuần đầu tiên rồi biến mất.
    - **Insight:** Sau khi có các nhóm này, chúng ta có thể phân tích tỷ lệ hoàn thành và (giả định) mức độ hài lòng cho từng nhóm, từ đó tìm ra mô hình hành vi nào dẫn đến thành công.
- **Phân tích theo Dòng thời gian (Temporal Analysis):**
    - **Mục tiêu:** Xác định các "điểm rơi" (dropout points) và các giai đoạn quan trọng trong vòng đời của một khóa học.
    - **Phương pháp:** Sử dụng **biểu đồ đường (Line Chart)** để vẽ số lượng học viên hoạt động hoặc số lượng tương tác theo từng tuần của khóa học.
    - **Câu hỏi cần trả lời:**
        - Học viên thường bỏ học ở tuần thứ mấy? Giai đoạn đó có bài giảng hoặc bài tập đặc biệt khó không?
        - Mức độ tương tác trên diễn đàn tăng hay giảm theo thời gian? Nó có tăng đột biến quanh các kỳ thi không?

c. Xây dựng và Kiểm định Giả thuyết (Hypothesis Formulation & Testing)

Dựa trên các phân tích trên, chúng ta sẽ định hình các giả thuyết rõ ràng và sử dụng các phương pháp thống kê để kiểm chứng, thay vì chỉ quan sát bằng mắt.

- **Giả thuyết 1:** *Mức độ tương tác xã hội (thảo luận trên diễn đàn) có tương quan dương mạnh mẽ với tỷ lệ hoàn thành khóa học.*
    - **Kiểm định:** So sánh tỷ lệ hoàn thành của nhóm tích cực thảo luận và nhóm không thảo luận bằng kiểm định T-test (T-test for independent samples).
- **Giả thuyết 2:** *Việc xem lại các video bài giảng nhiều lần là một chỉ báo cho thấy học viên gặp khó khăn với khái niệm đó, và thường dẫn đến điểm số bài tập liên quan thấp hơn ở lần thử đầu tiên.*
    - **Kiểm định:** Phân tích tương quan giữa **số lần xem lại** một video và **điểm số lần đầu** của bài tập tương ứng.
- **Giả thuyết 3:** *Các "chân dung" người học được xác định từ bước phân khúc có tỷ lệ hoàn thành khóa học khác biệt một cách có ý nghĩa thống kê.*
    - **Kiểm định:** Sử dụng phân tích phương sai **ANOVA** để so sánh giá trị trung bình của **tỷ lệ hoàn thành** giữa các nhóm chân dung đã được gom cụm.

Những phân tích chuyên sâu này không chỉ cung cấp cái nhìn toàn diện về dữ liệu mà còn trực tiếp gợi ý các đặc trưng (features) chất lượng cao cho mô hình dự đoán.

3.3. Chuẩn bị dữ liệu

Thu thập và lưu trữ dữ liệu:

- Dữ liệu được lấy từ bộ MoocCubeX và sắp xếp thành các tập dữ liệu nhỏ: thông tin khóa học, thông tin người học, hành vi học tập (video, bài tập, diễn đàn), và phản hồi.

Làm sạch dữ liệu (Data cleaning):

- Loại bỏ các trường không đáng tin cậy hoặc trùng lặp.
- Chuẩn hóa kiểu dữ liệu (ngày tháng, định dạng số, mã khóa học) để đảm bảo tính thống nhất.

Xử lý dữ liệu văn bản:

- Làm sạch các trường văn bản (mô tả khóa học, phản hồi,...).
- Mã hoá văn bản thành dạng số để phục vụ phân tích.

Tlền xử lý dữ liệu số và phân loại:

- Chuẩn hoá các đặc trưng số (số lần xem video, số bài tập hoàn thành,...).
- Mã hoá các đặc trưng phân loại (giới tính, chuyên ngành).

Xây dựng đặc trưng:

- Tạo thêm các đặc trưng phản ánh hành vi học tập:
    - Tỷ lệ hoàn thành khóa học.
    - Số lần đăng nhập.
    - Số lần tham gia thảo luận.
- Các đặc trưng này kết hợp cùng thông tin nhân khẩu học và thông tin khóa học tạo thành bộ dữ liệu đầu vào cho mô hình.

3.4. Đánh giá chất lượng dữ liệu

Xử lý giá trị thiếu:

- **Phương án:** Xác định trường dữ liệu bị thiếu, lựa chọn cách điền khuyết phù hợp để tránh làm sai lệch phân tích. (vd: điền khuyết bằng giá trị thống kê trung bình/ trung vị).

Xử lý dữ liệu ngoại lệ:

- **Phương án:** Kiểm tra các giá trị bất thường hoặc không hợp lệ, điều chỉnh dựa trên quy tắc thống kê và nghiệp vụ. (vd: điểm số không thể âm).

Xử lý mất cân bằng lớp:

- **Phương án:** Xem xét sự phân bố không đồng đều giữa các nhãn dữ liệu, áp dụng giải pháp phù hợp để hạn chế thiên lệch trong mô hình. (vd: sử dụng class_weights trong quá trình huấn luyện, kết hợp undersampling hoặc SMOTE để tạo cân bằng).

Kiểm tra trùng lặp:

- **Phương án:** Xoá bỏ bản ghi trùng ID hoặc bản ghi hành vi giống nhau trong cùng một thời điểm để tránh làm sai lệch kết quả thống kê.

Đánh giá phân phối dữ liệu:

- **Phương án:** Phân tích và trực quan hoá phân phối của các đặc trưng chính, so sánh trước và sau xử lý để đảm bảo tính hợp lý.

3.5. Xây dựng mô hình

Lựa chọn thuật toán:

- Mô hình chính: Random Forest. Đây là mô hình Ensemble cho thấy hiệu suất vượt trội và ổn định trong các báo cáo tương tự.
- Mô hình cơ sở: Decision Tree để làm thước đo so sánh.

Phương pháp đánh giá và tối ưu:

- Độ đo hiệu năng chính: Weighted F1-Score để đánh giá cân bằng giữa Precision và Recall, đặc biệt quan trọng khi dữ liệu mất cân bằng.
- Kỹ thuật kiểm định: K-Fold Cross-Validation để đảm bảo kết quả đánh giá khách quan.
- Tối ưu siêu tham số: RandomizedSearchCV để tìm ra bộ tham số tốt nhất một cách hiệu quả về thời gian.

3.6. Nền tảng triển khai (dự kiến)

- **Xử lý dữ liệu lớn**: PySpark để làm sạch và feature engineering dữ liệu lớn
- **Huấn luyện mô hình**: Spark MLlib, TensorFlow/PyTorch
- **Hạ tầng triển khai**:
    - **FE:** Reactjs (UI dự đoán, dashboard, …)
    - **BE:** Flask (api prediction từ mô hình ML/DL)

4. Hướng phát triển đề tài

***Triển khai hoàn chỉnh trên kiến trúc dữ liệu lớn với Kafka/Spark**

- **Pipeline:** MoocCubeX → Kafka (streaming) → Spark (process) → Model → Web
- **Nguồn dữ liệu:** dữ liệu MoocCubeX, đọc từng dòng vào Kafka topic
- **Streaming dữ liệu:** Kafka nhận dữ liệu streaming (consumer/producer)
- **Xử lý dữ liệu:** dùng Spark đọc dữ liệu từ Kafka → Xử lí/chuẩn hóa → Cập nhật/ghi vào db **(**VD: đếm số sự kiện xem video theo học viên mỗi 5 phút, tính tỷ lệ hoàn thành, …)
- **Truyền dữ liệu vào mô hình:** dùng mô hình đã train từ trước
- **Triển khai:** triển khai web react/flask