ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA KHOA HỌC MÁY TÍNH

BÁO CÁO ĐỀ TÀI

Dự đoán mức độ hoàn thành khóa học

Môn học: Khai thác dữ liệu và ứng dụng

Lớp: CS313.P22

Giảng viên hướng dẫn: ThS. Nguyễn Thị Anh Thư

Nhóm sinh viên thực hiện: Nhóm 2

Võ Đình Trung - MSSV: 22521571

Trương Phúc Trường - MSSV: 22521587

Mai Dương - MSSV: 22520302

Trần Qui Linh - MSSV: 22520779

Triệu Tấn Huy - MSSV: 22520581

Trần Nguyễn Anh Phong - MSSV: 22521092

Nguyễn Tấn Lợi - MSSV: 22520801

TP. Hồ Chí Minh, ngày 2 tháng 6 năm 2025

MỤC LỤC

1. Tổng quan

1.1. Giới thiệu

Giáo dục trực tuyến qua các Khóa học Trực tuyến Mở Quy mô lớn (MOOCs) đã trở thành một phần không thể thiếu của nền giáo dục hiện đại, mang lại cơ hội học tập linh hoạt cho hàng triệu người. Tuy nhiên, một trong những thách thức lớn nhất của MOOCs là tỷ lệ hoàn thành khóa học thấp.

Việc hiểu rõ các yếu tố ảnh hưởng đến sự tham gia và hoàn thành khóa học của học viên, cũng như khả năng dự đoán sớm các học viên có nguy cơ bỏ học, là vô cùng quan trọng để các nhà giáo dục và nền tảng MOOC có thể đưa ra các biện pháp can thiệp kịp thời, nâng cao hiệu quả giảng dạy và trải nghiệm học tập.

1.2. Định nghĩa bài toán

1.2.1. Đầu vào bài toán

- Dữ liệu người dùng: Thông tin cá nhân (năm sinh, giới tính, trường học - đã mã hóa), lịch sử đăng ký khóa học.
- Dữ liệu khóa học: Thông tin về khóa học (lĩnh vực, điều kiện tiên quyết - đã mã hóa), số lượng video, bài tập, giảng viên, trường liên kết, thời lượng video mặc định.
- Dữ liệu tương tác hàng tuần của người dùng với khóa học (time-series trong 4 tuần):
  - Hoạt động xem video: tổng thời gian xem, tốc độ xem, số lượng video đã xem.
  - Hoạt động làm bài tập/problem: số lần thử, điểm số, số bài tập hoàn thành, tổng số câu trả lời đúng.
  - Hoạt động xã hội: số lượng bình luận, trả lời.
  - Các đặc trưng tổng hợp về thời gian kể từ khi đăng ký đến khi thực hiện các hoạt động.

1.2.2. Đầu ra bài toán

- Biến mục tiêu chính: completion (Chỉ số hoàn thành tổng hợp, kết hợp tỷ lệ hoàn thành video và bài tập, được tính toán dựa trên tham số alpha để tính completion tùy theo cấu trúc khóa học). Giá trị này nằm trong khoảng [0, 1], thể hiện mức độ hoàn thành khóa học của học viên.

1.3. Ứng dụng

Hệ thống dự đoán này có nhiều ứng dụng thực tiễn trong lĩnh vực giáo dục trực tuyến:

- **Can thiệp sớm:** Giúp giảng viên và quản trị viên xác định những học viên có nguy cơ bỏ học hoặc hoàn thành khóa học ở mức độ thấp để có biện pháp hỗ trợ kịp thời (gửi thông báo, cung cấp tài liệu bổ sung, hỗ trợ cá nhân).
- **Cá nhân hóa lộ trình học tập:** Đề xuất nội dung hoặc phương pháp học tập phù hợp dựa trên hành vi và khả năng hoàn thành dự kiến của học viên.
- **Cải thiện chất lượng khóa học:** Phân tích các yếu tố ảnh hưởng đến việc hoàn thành khóa học giúp các nhà thiết kế khóa học tối ưu hóa nội dung và cấu trúc.
- **Tối ưu hóa nguồn lực:** Phân bổ nguồn lực hỗ trợ hiệu quả hơn cho các nhóm học viên cần thiết.

1.4. Khó khăn và thách thức

- **Khối lượng dữ liệu lớn (Big Data):** Bộ dữ liệu MOOCCubeX rất lớn, đòi hỏi kỹ thuật xử lý và lưu trữ hiệu quả.
- **Chất lượng dữ liệu:** Dữ liệu MOOC thường không đồng nhất, có thể thiếu sót, nhiễu, hoặc không nhất quán, cần quy trình đảm bảo chất lượng dữ liệu nghiêm ngặt.
- **Tính đa dạng của dữ liệu:** Dữ liệu bao gồm nhiều loại (số, văn bản, thời gian) và cấu trúc khác nhau.
- **Kỹ thuật đặc trưng (Feature Engineering):** Việc trích xuất và tạo ra các đặc trưng có ý nghĩa từ dữ liệu thô là một thách thức lớn, ảnh hưởng trực tiếp đến hiệu suất mô hình.

- **Tính động của hành vi học viên:** Hành vi học tập có thể thay đổi theo thời gian, đòi hỏi mô hình có khả năng thích ứng.
- **Tính diễn giải của mô hình:** Mô hình cần cung cấp thông tin dễ hiểu để giảng viên có thể đưa ra quyết định.

1.5. Mục tiêu

- Xây dựng và triển khai một pipeline xử lý dữ liệu MOOCs toàn diện, bao gồm các bước thu thập, tiền xử lý, và đảm bảo chất lượng dữ liệu.
- Phân tích sâu bộ dữ liệu MOOCCubeX để khám phá các yếu tố ảnh hưởng đến mức độ hoàn thành khóa học.
- Xây dựng và đánh giá các mô hình học máy có khả năng dự đoán chính xác mức độ hoàn thành khóa học (completion) của học viên.
- Xác định các đặc trưng quan trọng nhất ảnh hưởng đến dự đoán.
- (Nếu có) Phát triển một giao diện người dùng (BI tool/Dashboard) trực quan hóa kết quả phân tích và dự đoán, hỗ trợ quyết định cho giảng viên và quản trị viên.

1.6. Phạm vi thực hiện

- Tập trung vào bộ dữ liệu MOOCCubeX.
- Sử dụng các kỹ thuật khai phá dữ liệu và học máy để giải quyết bài toán hồi quy dự đoán mức độ hoàn thành khóa học.
- Đánh giá chất lượng dữ liệu theo các chiều chính: Độ đầy đủ, Tính nhất quán, Tính kịp thời, Độ chính xác (nếu có ground truth gián tiếp).
- Triển khai và so sánh hiệu suất của ít nhất 3-5 thuật toán hồi quy khác nhau.

2. Các công trình nghiên cứu liên quan

2.1. Kết quả khảo sát

Qua khảo sát các tài liệu và dự án liên quan đến phân tích dữ liệu MOOCs, một số hướng tiếp cận và kết quả nổi bật bao gồm:

- **Dự đoán bỏ học sớm (Early Dropout Prediction):** Nhiều nghiên cứu tập trung vào việc dự đoán khả năng bỏ học của sinh viên ngay từ những tuần đầu tiên. Ví dụ, nghiên cứu "Predicting MOOCs Dropout Using Only Two Easily Obtainable Features from the First Week's Activities" (2019) cho thấy độ chính xác cao (82%-94%) khi chỉ sử dụng 2 đặc trưng từ tuần đầu. Random Forest, XGBoost, Gradient Boosted Classifiers thường được sử dụng.
- **So sánh hiệu quả thuật toán:** Nghiên cứu "Predicting Student Dropout based on Machine Learning and Deep Learning: A Systematic Review" (2023) chỉ ra Random Forest là thuật toán được sử dụng phổ biến nhất với tỷ lệ chính xác cao (lên đến 99%) trong dự đoán bỏ học.
- **Hiểu rõ hành vi bỏ học:** "Understanding Dropouts in MOOCs" (2019) sử dụng dữ liệu từ XuetangX và đề xuất Mạng tương tác tính năng nhận biết bối cảnh (CFIN) để mô hình hóa và dự đoán hành vi bỏ học.
- **Ứng dụng Deep Learning:** "Deep Learning for Dropout Prediction in MOOCs" (2019) xây dựng mô hình RNN và lớp nhúng URL để dự đoán tỷ lệ hoàn thành nội dung, cho thấy hiệu quả cao hơn các mô hình học máy truyền thống.

  Các nghiên cứu này chủ yếu tập trung vào việc dự đoán "bỏ học" (dropout) như một biến nhị phân. Đề tài này hướng tới một góc nhìn chi tiết hơn bằng cách dự đoán "mức độ hoàn thành" (completion) là một giá trị liên tục, cung cấp thông tin đa dạng hơn về sự tiến bộ của học viên.

2.2. Hướng triển khai đề tài

Dựa trên các nghiên cứu liên quan và mục tiêu của đề tài, định hướng triển khai bao gồm:

- **Tập trung vào biến mục tiêu completion:** Thay vì chỉ dự đoán bỏ học, đề tài sẽ dự đoán một chỉ số hoàn thành tổng hợp, phản ánh đầy đủ hơn sự tham gia của học viên.
- **Áp dụng quy trình chất lượng dữ liệu nghiêm ngặt:** Do tính chất phức tạp và tiềm ẩn các vấn đề về chất lượng của dữ liệu MOOCs.
- **Sử dụng đa dạng các thuật toán hồi quy:** Bao gồm các thuật toán truyền thống (Linear Regression, Decision Trees), ensemble (Random Forest, Gradient Boosting), và các thuật toán khác (SVR, KNN) để so sánh và chọn ra mô hình tốt nhất.
- **Kỹ thuật đặc trưng nâng cao:** Trích xuất các đặc trưng từ dữ liệu tuần tự (time-series) và các tương tác phức tạp.
- **Hướng tới ứng dụng thực tế:** (Nếu có) Xây dựng một hệ thống BI/Dashboard để trực quan hóa kết quả, giúp người dùng cuối (giảng viên, quản trị viên) dễ dàng tiếp cận và sử dụng.

3. Cơ sở lý thuyết

3.1. **Khai phá dữ liệu (Data Mining) và Học máy (Machine Learning)**

Khai phá dữ liệu là quá trình khám phá các mẫu hữu ích và tri thức tiềm ẩn từ các tập dữ liệu lớn. Học máy, một nhánh của trí tuệ nhân tạo, cung cấp các thuật toán cho phép máy tính học hỏi từ dữ liệu mà không cần lập trình tường minh. Trong đề tài này, các kỹ thuật học máy có giám sát (supervised learning), cụ thể là các thuật toán hồi quy, sẽ được sử dụng để dự đoán biến mục tiêu liên tục (completion).

3.2. **Chất lượng dữ liệu (Data Quality - DQ)**

Chất lượng dữ liệu mô tả mức độ mà dữ liệu đáp ứng các yêu cầu và kỳ vọng của người dùng. Dữ liệu chất lượng cao cần chính xác, đầy đủ, nhất quán, kịp thời, hợp lệ và đáng tin cậy. Đề tài sẽ tập trung vào các khía cạnh DQ sau:

- **Hard Dimensions (Khía cạnh cứng - đo lường khách quan):**

    - **Accuracy (Độ chính xác):** Mức độ dữ liệu phản ánh đúng thực tế. (Đo lường gián tiếp qua Reliability và Relevance nếu không có ground truth).
    - **Completeness (Độ đầy đủ):** Tỷ lệ giữa số lượng giá trị hợp lệ và tổng số giá trị kỳ vọng.
    - **Consistency (Tính nhất quán):** Dữ liệu đồng nhất giữa các nguồn và hệ thống.
    - **Timeliness (Tính kịp thời):** Dữ liệu được cập nhật và phản ánh tình hình hiện tại.
- **Soft Dimensions (Khía cạnh mềm - đo lường chủ quan):**
    - **Validity (Tính hợp lệ), Uniqueness (Tính duy nhất), Reliability (Độ tin cậy), Relevance (Tính liên quan):** Sẽ được đánh giá qua quá trình phân tích và tiền xử lý dữ liệu.

3.3. **Các thuật toán hồi quy**

- **Random Forest Regressor:** Thuật toán ensemble dựa trên nhiều cây quyết định, giảm overfitting và tăng độ chính xác.
- **Gradient Boosting Regressor:** Thuật toán boosting xây dựng các cây tuần tự, mỗi cây sửa lỗi của cây trước đó, thường cho hiệu suất cao.
- **Support Vector Regressor (SVR):** Tìm một siêu phẳng tối ưu để hồi quy, hiệu quả với dữ liệu nhiều chiều và phi tuyến tính (khi dùng kernel).
- **K-Nearest Neighbors (KNN) Regressor:** Dự đoán giá trị của một điểm dữ liệu dựa trên giá trị trung bình của k điểm lân cận gần nhất.
- **Decision Tree Regressor:** Xây dựng mô hình dạng cây để đưa ra dự đoán, dễ diễn giải.

3.4. **Các độ đo đánh giá mô hình hồi quy**

- **RMSE (Root Mean Squared Error):** Căn bậc hai của trung bình bình phương sai số. Nhạy cảm với các lỗi lớn.

- **MAE (Mean Absolute Error):** Trung bình của giá trị tuyệt đối của sai số. Ít nhạy cảm với outliers hơn RMSE.
- **R^2 (R-squared - Hệ số xác định):** Tỷ lệ phương sai của biến phụ thuộc được giải thích bởi các biến độc lập. Giá trị càng gần 1 càng tốt.
- **MAPE (Mean Absolute Percentage Error):** Trung bình của phần trăm sai số tuyệt đối. Hữu ích khi so sánh độ chính xác trên các tập dữ liệu có thang đo khác nhau.

4. Phân tích bộ dữ liệu

4.1. Tìm hiểu dữ liệu

4.1.1. **Giới thiệu bộ dữ liệu MOOCCubeX**

Bộ dữ liệu MOOCCubeX được thu thập từ nền tảng XuetangX, một trong những nền tảng MOOC lớn nhất Trung Quốc. Bộ dữ liệu này chứa thông tin chi tiết về 4.216 khóa học, 3.330.294 sinh viên, và khoảng 296 triệu bản ghi hành vi. Dữ liệu được tổ chức thành các thực thể (entities)
như course, user, video, problem, và các quan hệ (relations) như user-video, user-problem.

4.1.2. **Tìm hiểu dữ liệu**

- **Cấu trúc dữ liệu:** Dữ liệu được cung cấp dưới dạng các file JSON. Các trường chính bao gồm ID người dùng, ID khóa học, ID video, ID problem, thời gian tương tác, nội dung bình luận, điểm số, v.v.
- **Đặc điểm dữ liệu:** Dữ liệu có tính đa dạng cao, bao gồm cả dữ liệu có cấu trúc (thông tin người dùng, khóa học) và bán cấu trúc (dữ liệu tương tác). Dữ liệu hành vi có tính thời gian (time-series).

4.2. Chuẩn bị dữ liệu

Quá trình này bao gồm nhiều bước chi tiết đã được thực hiện trong các bài tập trước:

- **Hợp nhất dữ liệu:**

- Đọc và trích xuất các cặp (user_id, course_id) từ user.json (trường course_order).
- Lọc bỏ người dùng tham gia quá nhiều khóa học (>20) và các khóa học có quá ít người đăng ký (<2) để giảm nhiễu và tập trung vào các tương tác có ý nghĩa.
- Kết hợp thông tin từ các file entities/*.json (user, course, video, problem, comment, reply) và relations/*.json (user-video, user-problem) để tạo ra một DataFrame đầu vào duy nhất cho mỗi cặp (user_id, course_id) với các đặc trưng hành vi được tổng hợp theo tuần (tuần 1, 2, 3, 4).

- **Tạo biến mục tiêu (completion):**
  - Tính video_completion = (Tổng số lượng video mà user đã xem) / (Tổng số lượng video của khóa học).
  - Tính problem_completion = (Số problem user đã làm) / (Tổng số problem trong khóa học).
  - Xác định trọng số alpha dựa trên tỷ lệ num_problems / (num_videos + epsilon) để cân bằng ảnh hưởng của video và bài tập.
  - completion = (alpha * problem_completion) + ((1 - alpha) * video_completion).
- **Xử lý giá trị thiếu:** Sử dụng các phương pháp như điền giá trị trung bình, trung vị, hoặc giá trị 0/hằng số tùy theo ngữ cảnh của đặc trưng.
- **Mã hóa đặc trưng:**
  - Label Encoding hoặc One-Hot Encoding cho các đặc trưng phân loại (ví dụ: field_encoded, prerequisites_encoded).
- **Chuẩn hóa/Scaling dữ liệu:** (Nếu cần thiết cho thuật toán cụ thể) Sử dụng StandardScaler hoặc MinMaxScaler để đưa các đặc trưng về cùng một thang đo.
- **Tạo đặc trưng mới (Feature Engineering):**

- Các đặc trưng về thời gian: days_since_enroll cho các hoạt động.
- Các đặc trưng tương tác tổng hợp/trung bình: avg_comments_per_student, avg_video_watch_time_per_student…
- Đặc trưng về sự tiến bộ theo tuần: total_correct_answer_weekN, total_video_watching_weekN…

4.3. Phân tích vấn đề

- **Thống kê mô tả:** Tính toán các giá trị trung bình, trung vị, độ lệch chuẩn, min, max cho các đặc trưng số để hiểu phân phối của chúng.
- **Phân tích phân phối:** Sử dụng histogram, boxplot để trực quan hóa sự phân phối của các đặc trưng quan trọng và biến mục tiêu.
- **Phân tích tương quan:** Sử dụng ma trận tương quan (heatmap) để xác định mối quan hệ tuyến tính giữa các đặc trưng và giữa các đặc trưng với biến mục tiêu.
- **Trực quan hóa mối quan hệ:** Sử dụng scatter plot để xem xét mối quan hệ giữa các cặp đặc trưng.
- **Phân tích theo cụm (nếu có):** Dựa trên kết quả phân cụm người học (ví dụ, thành 3 cụm: người học chăm chỉ nhưng chưa hiệu quả, người học gần như bỏ cuộc, người học xuất sắc) để hiểu các nhóm hành vi khác nhau.
- **Kiểm định thống kê:** Sử dụng t-test, ANOVA để so sánh sự khác biệt giữa các nhóm (ví dụ: so sánh composite_completion giữa nhóm có video_completion_ratio == 0 và != 0).

5. Phương pháp đề xuất

5.1. **Quy trình tổng quan**

Quy trình tổng thể của hệ thống được đề xuất bao gồm các giai đoạn chính sau:

1. **Thu thập và Tiền xử lý Dữ liệu:** Trích xuất, hợp nhất, làm sạch dữ liệu từ MOOCCubeX.

2. **Đảm bảo Chất lượng Dữ liệu:** Đánh giá và cải thiện chất lượng dữ liệu theo các chiều đã định.
3. **Kỹ thuật Đặc trưng:** Tạo các đặc trưng có ý nghĩa, đặc biệt là các đặc trưng theo tuần.
4. **Huấn luyện Mô hình:** Lựa chọn, huấn luyện và tinh chỉnh các thuật toán hồi quy.
5. **Đánh giá Mô hình:** Đánh giá hiệu suất mô hình bằng các độ đo phù hợp.
6. **Triển khai (Giao diện BI):** Trực quan hóa kết quả dự đoán và các thông tin hỗ trợ.

5.2. **Kiến trúc dữ liệu**

Dữ liệu từ các file JSON của MOOCCubeX được xử lý và hợp nhất thành một DataFrame chính cho thông tin 4 tuần học (w1234.csv), trong đó mỗi dòng đại diện cho một cặp (user_id, course_id) với các đặc trưng hành vi được tổng hợp theo 4 tuần đầu tiên và biến mục tiêu là completion.

5.3. **Đánh giá và đảm bảo chất lượng dữ liệu**

- **Completeness:** Tính toán tỷ lệ thiếu của từng cột. Các cột có tỷ lệ thiếu cao sẽ được xem xét loại bỏ hoặc xử lý cẩn thận (ví dụ: year_of_birth). Dữ liệu thiếu trong các đặc trưng hành vi có thể được điền bằng 0 (nếu có nghĩa là không có hoạt động) hoặc giá trị trung bình/trung vị.
- **Consistency:** Kiểm tra tính nhất quán của kiểu dữ liệu và định dạng (ví dụ: ID khóa học luôn theo mẫu C_xxxx).
- **Timeliness:** Phân tích thời gian đăng ký và hoạt động để đảm bảo dữ liệu phản ánh đúng tiến trình học tập.
- **Accuracy (gián tiếp):** Thông qua việc xử lý outlier, kiểm tra logic (ví dụ: thời gian xem video không thể âm).

5.4. **Xây dựng mô hình dự đoán**

- **Lựa chọn thuật toán:** Tập trung vào các thuật toán hồi quy mạnh mẽ như Random Forest Regressor và Gradient Boosting Regressor do hiệu suất cao

đã được chứng minh trong các thử nghiệm trước. Các thuật toán khác như SVR, KNN, Decision Tree Regressor cũng được xem xét để so sánh.

- **Phân chia dữ liệu:** Dữ liệu được chia thành tập huấn luyện (train), tập kiểm định (validation), và tập kiểm thử (test) theo tỷ lệ phổ biến (ví dụ: 60%/20%/20% hoặc 70%/15%/15%).
- **Huấn luyện và Tinh chỉnh:** Sử dụng GridSearchCV hoặc RandomizedSearchCV để tìm bộ siêu tham số tối ưu cho từng mô hình trên tập kiểm định.
- **Đánh giá:** Đánh giá mô hình cuối cùng trên tập kiểm thử bằng các độ đo RMSE, MAE, R^2, MAPE.
- **Phân tích đặc trưng quan trọng:** Sử dụng feature_importances_ (đối với mô hình cây) hoặc Permutation Importance để xác định các yếu tố ảnh hưởng lớn nhất đến mức độ hoàn thành.

5.5. **Hệ thống thông minh hỗ trợ người dùng (Giao diện BI)**

(Phần này dựa trên tài liệu Nội dung web và README.md của web app)
Nếu triển khai, một ứng dụng web (ví dụ, sử dụng React, Material-UI) sẽ được phát triển để:

- **Dashboard Quản trị viên:** Hiển thị KPIs tổng quan (tổng số học viên, khóa học, tỷ lệ hoàn thành trung bình, số học viên nguy cơ cao), phân phối tỷ lệ hoàn thành, top/bottom khóa học, xu hướng học viên nguy cơ theo thời gian.
- **Dashboard Giảng viên:** Hiển thị KPIs cho từng khóa học cụ thể, danh sách học viên kèm điểm dự đoán và mức độ rủi ro, phân phối điểm hoàn thành trong khóa, các yếu tố ảnh hưởng hàng đầu đến nguy cơ bỏ học trong khóa.
- **Tính năng Dự đoán Thử nghiệm:** Cho phép người dùng nhập các giá trị đặc trưng giả định để xem kết quả dự đoán của mô hình.
- **Trung tâm Mô hình Dự đoán:** Hiển thị thông tin về các mô hình đã huấn luyện, hiệu suất (R^2, RMSE, MAE, MAPE), các đặc trưng quan trọng nhất. Cho phép quản trị viên chọn mô hình hoạt động chính, hoặc huấn luyện lại mô hình.

6. Thực nghiệm

6.1. Miêu tả bộ dữ liệu

Bộ dữ liệu cuối cùng được sử dụng cho huấn luyện mô hình (ví dụ: w1234.csv) chứa khoảng 410,440 dòng sau khi lọc và hợp nhất ban đầu, và sau đó được xử lý tiếp để tạo ra các đặc trưng theo tuần. Số lượng đặc trưng cuối cùng có thể lên đến vài chục hoặc hàng trăm tùy thuộc vào cách tổng hợp dữ liệu tuần. Biến mục tiêu là completion.

6.2. Phương pháp tổ chức dữ liệu thực nghiệm

Dữ liệu được chia ngẫu nhiên thành tập huấn luyện, tập kiểm định và tập kiểm thử theo tỷ lệ 60:20:20. Cross-validation (ví dụ: KFold với n_splits=5) được sử dụng trong quá trình tinh chỉnh siêu tham số trên tập huấn luyện + kiểm định gộp lại.

6.3. Độ đo đánh giá

Các độ đo chính được sử dụng là: RMSE, MAE, R^2 Score, và MAPE.

6.4. Kịch bản thực nghiệm

- **Kịch bản 1: Huấn luyện và đánh giá các mô hình cơ sở.**
    - Các mô hình (Random Forest, Gradient Boosting, SVR, KNN, Decision Tree) được huấn luyện với các siêu tham số mặc định hoặc một bộ siêu tham số cơ bản.
- **Kịch bản 2: Tinh chỉnh siêu tham số và lựa chọn mô hình tốt nhất.**
    - Sử dụng GridSearchCV để tìm bộ siêu tham số tối ưu cho các mô hình tiềm năng nhất (ví dụ: Random Forest, Gradient Boosting).
    - So sánh hiệu suất trên tập kiểm thử.
- **Kết quả (dựa trên tài liệu BÀI TẬP PHÂN TÍCH DỮ LIỆU QUA HUẤN LUYỆN MÔ HÌNH):**

- **Random Forest Regressor:**
  - RSME (Test): 0.0029
  - MAE (Test): 0.0004
  - R^2 Score (Test): 0.9997
  - MAPE (Test): 0.3584%
  - Đặc trưng quan trọng nhất: total_video_watch_time_1 (~45%), alpha (~35%).
- **Gradient Boosting Regressor:**
  - Test RMSE: 0.0031, Test MAE: 0.0014
  - R^2 Score: 0.9996
  - MAPE: 3.83%
  - Đặc trưng quan trọng nhất: problem_ratio (45.0%), alpha (41.3%).
- **Support Vector Regressor (SVR):** (Sau tinh chỉnh)
  - RMSE (Test): 0.0068, MAE (Test): 0.0050
  - R^2 (Test): 0.9983
  - Đặc trưng quan trọng nhất (Permutation Importance): problem_ratio, alpha.
- **KNN Regressor:** (Sau tinh chỉnh)
  - RMSE (Test): 0.0137, MAE (Test): 0.0038
  - R^2 (Test): 0.9929
- Các mô hình ensemble như Random Forest và Gradient Boosting cho thấy hiệu suất vượt trội.

6.5. Đánh giá kết quả thực nghiệm

- Các mô hình ensemble, đặc biệt là Random Forest và Gradient Boosting, đạt được độ chính xác rất cao ($R^2 > 0.999$) trong việc dự đoán mức độ hoàn thành khóa học trên bộ dữ liệu MOOCCubeX đã qua xử lý.
- Các đặc trưng liên quan đến tỷ lệ hoàn thành bài tập (problem_ratio), tỷ lệ hoàn thành video (thông qua alpha và total_video_watch_time_1), và trọng số alpha (phản ánh cấu trúc khóa học) là những yếu tố quan trọng nhất.
- Kết quả ổn định qua cross-validation cho thấy các mô hình không bị overfitting nghiêm trọng.

7. Kết luận và hướng phát triển

Đề tài đã thành công trong việc xây dựng một quy trình phân tích và dự đoán mức độ hoàn thành khóa học MOOCs. Bằng cách áp dụng các kỹ thuật tiền xử lý dữ liệu, đảm bảo chất lượng dữ liệu, và huấn luyện các mô hình hồi quy tiên tiến, hệ thống đã đạt được khả năng dự đoán với độ chính xác cao. Các yếu tố chính ảnh hưởng đến sự hoàn thành khóa học đã được xác định, cung cấp những hiểu biết giá trị cho việc cải thiện trải nghiệm học tập và hỗ trợ sinh viên.

7.1. Ưu nhược điểm của phương pháp đề xuất

- **Ưu điểm:**
    - Quy trình xử lý dữ liệu toàn diện và bài bản.
    - Đạt được hiệu suất dự đoán rất cao với các mô hình ensemble.
    - Xác định được các đặc trưng có ảnh hưởng lớn, mang tính gợi ý cho các biện pháp can thiệp.
    - Biến mục tiêu completion cung cấp cái nhìn chi tiết hơn so với việc chỉ dự đoán bỏ học.
- **Nhược điểm:**

- Độ phức tạp của bộ dữ liệu MOOCCubeX đòi hỏi nhiều nỗ lực trong tiền xử lý.
- Tính diễn giải của các mô hình ensemble (Random Forest, Gradient Boosting) có thể hạn chế hơn so với các mô hình đơn giản như Decision Tree.
- Dữ liệu chỉ giới hạn trong 4 tuần đầu, có thể chưa phản ánh hết toàn bộ quá trình học tập.
- Mô hình có thể nhạy cảm với sự thay đổi trong cấu trúc khóa học hoặc hành vi người dùng theo thời gian (data drift).

7.2. Hướng phát triển tiềm năng

- **Mở rộng khung thời gian phân tích:** Phân tích dữ liệu từ nhiều tuần hơn để nắm bắt sự thay đổi hành vi dài hạn.
- **Phát triển mô hình dự đoán theo thời gian thực (Real-time Prediction):** Cảnh báo sớm hơn về nguy cơ của học viên.
- **Tích hợp các nguồn dữ liệu đa dạng hơn:** Ví dụ, dữ liệu về tương tác trên diễn đàn, kết quả học tập từ các khóa học trước.
- **Ứng dụng các thuật toán Deep Learning:** Khám phá tiềm năng của LSTM, Transformer cho dữ liệu chuỗi thời gian hành vi.
- **Cá nhân hóa can thiệp:** Phát triển các gợi ý và biện pháp hỗ trợ tự động, cá nhân hóa dựa trên kết quả dự đoán và các đặc trưng của học viên.
- **Hoàn thiện và triển khai giao diện BI:** Xây dựng một công cụ trực quan, thân thiện cho giảng viên và quản trị viên để theo dõi, phân tích và đưa ra quyết định dựa trên dữ liệu.
- **Nghiên cứu sâu hơn về tính diễn giải (Interpretability):** Áp dụng các kỹ thuật như SHAP, LIME để hiểu rõ hơn quyết định của các mô hình phức tạp.