TRUSTEE: Tree-based Regression for Undergraduate Student Tracking and Educational Explainability

Trần Nam Hải, Nguyễn Vũ Trung Kiên, Phạm Tiến Dũng
Trường Công nghệ Thông tin và Truyền thông, Đại học Bách khoa Hà Nội, Việt Nam
Email: {trannamhai.5d, ngkienn89, tiendungcbh1801}@gmail.com

Đội thi HD4K - Vòng loại cuộc thi DataFlow 2026

***Tóm tắt nội dung*—Dự báo sớm kết quả học tập là một bài toán thách thức trong khai phá dữ liệu giáo dục do tính chất đa chiều của hồ sơ sinh viên. Để giải quyết vấn đề này, chúng tôi đề xuất TRUSTEE – một hệ thống dự báo số tín chỉ hoàn thành của sinh viên trong kỳ học. Điểm đột phá trong phương pháp của chúng tôi nằm ở chiến lược phân tầng dữ liệu theo thâm niên của học sinh (Fresher và Senior) để nắm bắt chính xác các động lực học tập riêng biệt. Đồng thời, phương pháp của chúng tôi giải quyết triệt để bài toán biến thiên độ khó đề thi tuyển sinh qua các năm bằng cách tích hợp và chuẩn hóa dữ liệu phổ điểm thi THPT Quốc gia. Thông qua việc thực nghiệm so sánh ba phương pháp tiếp cận (trực tiếp, phần dư, tỷ lệ) trên các kiến trúc Ensemble Learning, kết quả cho thấy mô hình LightGBM với hàm mất mát Tweedie đạt hiệu năng vượt trội, ghi nhận chỉ số RMSE đạt 3.5040 trên tập kiểm thử. Đặc biệt, chúng tôi tích hợp khung giải thích đa tầng (SHAP, LIME, DiCE) cho phép chuyển đổi các kết quả dự báo thành các khuyến nghị hành động cụ thể, cung cấp công cụ hỗ trợ ra quyết định minh bạch và hiệu quả cho công tác cố vấn học tập. Mã nguồn khả dụng tại: https://github.com/CryAndRRich/trustee.**

***Từ khoá*—Khai phá dữ liệu giáo dục, Dự báo kết quả học tập, Học kết hợp, LightGBM, xAI.**

I. GIỚI THIỆU BÀI TOÁN

Trong kỷ nguyên số hóa giáo dục, dữ liệu chính là chìa khóa để giải quyết các vấn đề thực tiễn trong quản lý đào tạo. Tại nhiều trường đại học lớn, tình trạng sinh viên đăng ký khối lượng tín chỉ lớn nhưng không hoàn thành, dẫn đến trượt môn, chậm tiến độ, hay nghiêm trọng hơn là bị cảnh cáo và buộc thôi học đang là một thách thức nhức nhối.

Đứng trước bài toán này, phương pháp quản lý truyền thống (chỉ xử lý khi sinh viên đã trượt) là chưa đủ và sẽ không giải quyết được triệt để vấn đề. Nhu cầu cấp thiết đặt ra là phải xây dựng được một công cụ có khả năng dự báo sớm kết quả học tập ngay từ đầu kỳ. Điều này sẽ vừa giúp nhà trường tối ưu hóa nguồn lực hỗ trợ, vừa giúp chính sinh viên điều chỉnh lộ trình học tập phù hợp, giảm thiểu rủi ro trong học tập.

Trong khuôn khổ của vòng loại cuộc thi "*DataFlow 2026: The Alchemy of Minds*", đề bài yêu cầu các đội thi dự báo số tín chỉ hoàn thành của mỗi sinh viên sau khi kết thúc một kỳ học. Dựa trên bản chất biến mục tiêu là một giá trị số thực, chúng tôi định nghĩa đây là một bài toán hồi quy (Regression) trong lĩnh vực khai phá dữ liệu giáo dục.

Để giải quyết bài toán này, chúng tôi đề xuất **TRUSTEE** (**T**ree-based **R**egression for **U**ndergraduate **S**tudent **T**racking and **E**ducational **E**xplainability) với mục tiêu là dự báo **số tín chỉ thực tế sinh viên hoàn thành** (`TC_HOANTHANH`) sau khi kết thúc kỳ học.

Tập dữ liệu được cung cấp là sự tổng hợp đa chiều các nguồn thông tin về các sinh viên, với ba nhóm thông tin chính bao gồm:

- Năng lực nền tảng: Thông qua dữ liệu tuyển sinh (điểm thi, khối thi, phương thức xét tuyển).
- Hành vi quá khứ: Lịch sử học tập, điểm GPA/CPA và thói quen tích lũy tín chỉ của các kỳ trước.
- Kế hoạch hiện tại: Số tín chỉ sinh viên đăng ký trong học kỳ cần dự báo (`TC_DANGKY`).

Với bài toán cụ thể này, chúng tôi xác định chiến lược giải quyết vấn đề như sau:

- Tận dụng dữ liệu toàn diện: Bên cạnh hai tập dữ liệu chính là `admission.csv` và `academic_records.csv`, chúng tôi sẽ khai thác thêm các nguồn dữ liệu công khai (như phổ điểm thi THPT Quốc gia qua các năm) để chuẩn hóa và làm giàu đặc trưng, giúp mô hình hiểu rõ hơn về năng lực tương đối của sinh viên.
- Mô hình hóa mạnh mẽ và minh bạch: Không chỉ dừng lại ở việc tối ưu các chỉ số sai số (R^2, $RMSE$), chúng tôi hướng tới việc xây dựng các mô hình dựa trên cây quyết định (Tree-based models [1]) có khả năng giải thích cao (bằng cách tận dụng thêm Explainable AI [2]). Điều này nhằm đáp ứng một trong những tiêu chí quan trọng của đề bài: giải thích được yếu tố nào ảnh hưởng lớn nhất đến khả năng hoàn thành tín chỉ của sinh viên.
- Định hướng ứng dụng: Kết quả dự báo trên tập kiểm thử cho học kỳ 1 năm học 2024-2025 sẽ là cơ sở để chúng tôi đề xuất các nhóm giải pháp phân loại và hỗ trợ sinh viên, biến kết quả từ mô hình thuật toán thành giá trị thực tiễn cho công tác cố vấn học tập.

II. XỬ LÝ DỮ LIỆU

Trước hết, chúng tôi tiến hành phân tích, làm sạch và trích xuất đặc trưng từ bộ dữ liệu gốc, đồng thời tích hợp các nguồn dữ liệu bổ sung để tối ưu hóa độ chính xác của mô hình.

A. Dữ liệu tổng quan

Bộ dữ liệu nguyên bản bao gồm hai bảng dữ liệu huấn luyện và một bảng kiểm thử.

1) Cấu trúc dữ liệu:

- `admission.csv` **(Thông tin tuyển sinh):** Bao gồm dữ liệu của 30.217 sinh viên nhập học trong giai đoạn 2018-2025, phản ánh nền tảng năng lực đầu vào. Các trường thông tin trọng yếu bao gồm:
  - Năm tuyển sinh - Năm sinh viên bắt đầu nhập học
  - Phương thức xét tuyển
  - Tổ hợp môn xét tuyển
  - Tổng điểm trúng tuyển của sinh viên
  - Điểm chuẩn đầu vào của ngành học tương ứng
- `academic_records.csv` **(Lịch sử học tập):** Lưu trữ kết quả học tập của 20.381 sinh viên từ niên khóa 2020-2021 đến 2023-2024, thể hiện phong độ học tập theo thời gian. Các biến số quan trọng gồm:
  - Điểm GPA và CPA
  - Số tín chỉ sinh viên đăng ký vào đầu kỳ
  - Số tín chỉ thực tế sinh viên hoàn thành đến cuối kỳ
- `test.csv` **(Tập kiểm thử):** Bao gồm danh sách của 16.502 sinh viên và số tín chỉ đăng ký cho Học kỳ 1 năm học 2024-2025. Mục tiêu là dự báo tổng số tín chỉ hoàn thành của mỗi sinh viên tại thời điểm kết thúc học kỳ.

2) Các vấn đề khi xử lý dữ liệu:

Thông qua quá trình phân tích dữ liệu (EDA), chúng tôi ghi nhận một số vấn đề ảnh hưởng trực tiếp đến tính nhất quán của mô hình huấn luyện:

- **Dữ liệu trùng lặp:** Tồn tại các bản ghi trùng lặp khóa của cùng một sinh viên trong cùng một học kỳ (trùng `MA_SO_SV, HOC_KY`) trong bảng dữ liệu `academic_records.csv` nhưng thông tin thuộc tính (`GPA, TC_HOANTHANH, TC_DANGKY`) lại không đồng nhất. Tổng cộng có 460 sinh viên sở hữu hai kết quả học tập khác nhau trong cùng một học kỳ (phần lớn tập trung ở học kỳ 1 năm học 2023-2024 với 428 trường hợp). Nguyên nhân có thể xuất phát từ việc học cải thiện hoặc lỗi hệ thống. Để đảm bảo tính nhất quán, chúng tôi ưu tiên giữ lại bản ghi phản ánh kết quả học tập tốt hơn (tỷ lệ hoàn thành cao, GPA cao).
- **Nhãn chưa xuất hiện:** Tập kiểm thử `test.csv` xuất hiện các nhãn dữ liệu chưa từng có trong tập huấn luyện. Cụ thể, trường `PTXT` (Phương thức xét tuyển) xuất hiện thêm mã `303` và trường `TOHOP_XT` (Tổ hợp môn xét tuyển) ghi nhận thêm các tổ hợp `D24, H01, TT`. Việc áp dụng One-Hot Encoding truyền thống có thể gây lỗi chiều dữ liệu hoặc mất mát thông tin. Do đó, chúng tôi áp dụng kỹ thuật gộp nhóm các giá trị hiếm kết hợp với CatBoost Encoder [3].
- **Dữ liệu chuỗi thời gian bị ngắt quãng:** Ghi nhận 650 sinh viên có lịch sử học tập không liên tục (ví dụ: sinh viên `001565088153` có dữ liệu đến học kỳ 1 năm học 2021-2022, sau đó gián đoạn và xuất hiện lại ở học kỳ 2 năm học 2023-2024). Nguyên nhân có thể do bảo lưu kết quả học tập. Sự gián đoạn này gây nhiễu nghiêm trọng cho việc tính toán các đặc trưng chuỗi thời gian. Giải pháp chúng tôi sử dụng để giải quyết vấn đề này là thiết lập thuật toán xác định chỉ số học kỳ thực tế (`SEMESTER_INDEX`) thay vì dựa thuần túy vào niên khóa.

B. Dữ liệu bổ sung

Đáp ứng khuyến nghị về việc tận dụng dữ liệu mở, chúng tôi nhận định rằng điểm trúng tuyển đơn thuần chưa phản ánh chính xác năng lực tương đối của sinh viên do biến động độ khó của đề thi THPT Quốc gia qua các năm.

Dữ liệu **Phổ điểm thi THPT Quốc gia giai đoạn 2020 - 2024** (Phụ lục A) đã được thu thập và tích hợp nhằm chuẩn hóa điểm số về một mặt bằng chung dựa trên phân phối điểm toàn quốc của từng năm. Điều này giúp mô hình nhận diện sự khác biệt về "trọng số" của cùng một mức điểm (ví dụ: điểm 25.0) giữa các năm khác nhau.

Trên cơ sở đó, quy trình trích xuất đặc trưng được thực hiện nhằm cung cấp góc nhìn toàn diện về hồ sơ sinh viên. Các biến số được phân loại thành ba nhóm chính:

1) Đặc trưng Tuyển sinh: Nhóm đặc trưng này đánh giá năng lực đầu vào so với mặt bằng chung:

- **Z-Score đầu vào** (`Z_SCORE`): Định lượng độ lệch chuẩn của điểm sinh viên so với trung bình của năm tuyển sinh tương ứng.
- **Khoảng cách điểm** (`SCORE_GAP, GAP_RATIO`): Chênh lệch tuyệt đối và tương đối giữa điểm trúng tuyển của sinh viên so với điểm chuẩn của tổ hợp môn xét tuyển.
- **Xếp hạng tương đối** (`ENTRY_RANK`): Thứ hạng phần trăm của sinh viên trong cùng khóa tuyển sinh.

2) Đặc trưng Lịch sử học tập: Đóng vai trò cốt lõi, nhóm đặc trưng này ứng dụng kỹ thuật cửa sổ trượt để nắm bắt xu hướng:

- **Kết quả học tập tích lũy:**
  - `HIST_AVG_GPA, HIST_STD_GPA`: Thể hiện trung bình và độ lệch chuẩn GPA tích lũy, phản ánh năng lực nền tảng và sự ổn định phong độ của sinh viên.
  - `HIST_MAX_PASSED, HIST_MAX_GPA`: Ghi nhận kỷ lục cá nhân về số tín chỉ và điểm số đạt được trong quá khứ.
  - `OVERLOAD_VS_MAX`: Đánh giá mức độ rủi ro bằng cách so sánh số tín chỉ đăng ký hiện tại với kỷ lục hoàn thành trong quá khứ.
- **Kết quả kỳ liền trước:**
  - `LAST_GPA, LAST_FAIL`: Kết quả học tập và số tín chỉ trượt của kỳ gần nhất.
  - `LAST_PASS_RATIO`: Tỷ lệ hoàn thành tín chỉ của kỳ liền trước.
- **Kết quả hai, ba kỳ gần nhất:**
  - `R2_AVG_GPA, R3_AVG_GPA`: Phong độ học tập trung bình trong hai, ba kỳ gần nhất.
  - `FAIL_TREND_R2`: Xu hướng trượt môn (so sánh số tín chỉ trượt kỳ trước với trung bình trượt của hai kỳ gần nhất).
  - `PRESSURE_VS_R2`: Áp lực học tập, được tính bằng tỷ lệ giữa số tín chỉ đăng ký hiện tại so với năng lực hoàn thành trung bình gần đây.

3) Đặc trưng Niên khóa: Biến số này xác định năm học hiện tại của sinh viên (Năm nhất, Năm hai,...) dựa trên chênh lệch giữa thời gian thực và năm tuyển sinh.

Đối với đối tượng sinh viên năm nhất thiếu hụt lịch sử học tập, kỹ thuật điền khuyết (Imputation) được áp dụng với các giá trị mặc định hợp lý (ví dụ: `LAST_FAIL` = 0,

LAST_GPA = trung bình toàn trường) nhằm đảm bảo tính liên tục trong tính toán của mô hình.

C. Phân chia dữ liệu

Để đảm bảo tính khách quan, khả năng tổng quát hóa của mô hình cũng như yêu cầu đề bài, chúng tôi áp dụng chiến lược phân chia dữ liệu dựa trên hai chiều: phân nhóm theo thâm niên sinh viên để tối ưu hóa đặc trưng và phân chia theo thời gian để kiểm định tính ổn định.

1) Chiến lược phân nhóm mô hình: Phân tích tương quan cho thấy sự thay đổi rõ rệt về mức độ ảnh hưởng của các đặc trưng theo thâm niên học tập. Các đặc trưng tuyển sinh có tác động lớn đến sinh viên năm nhất (Fresher), trong khi đối với sinh viên năm hai trở đi (Senior), lịch sử học tập thực tế mới là yếu tố quyết định.

Do đó, thay vì sử dụng một mô hình tổng quát, chúng tôi đề xuất chiến lược chia tách dữ liệu huấn luyện thành hai tập con:

- **Nhóm Fresher (Sinh viên năm nhất):** Nhóm chưa tích lũy dữ liệu lịch sử. Mô hình sẽ tập trung trọng số vào các đặc trưng tuyển sinh và dữ liệu bổ sung từ phổ điểm THPT, tổng cộng là 15 đặc trưng.
- **Nhóm Senior (Sinh viên năm hai trở đi):** Nhóm đã có lịch sử học tập dày dặn. Mô hình tập trung khai thác các chuỗi đặc trưng lịch sử học tập với tổng cộng 22 đặc trưng.

Chiến lược này cho phép xây dựng hai mô hình chuyên biệt, tối ưu hóa hiệu suất dự báo so với phương pháp tiếp cận đơn mô hình truyền thống. Chi tiết về các đặc trưng được trình bày tại Phụ lục B.

2) Phân chia tập Huấn luyện - Kiểm định - Kiểm thử: Theo như quy định phân chia tập dữ liệu của cuộc thi, chúng tôi phân chia theo mốc thời gian như sau:

- **Tập Huấn luyện (Training Set):** Sử dụng dữ liệu lịch sử từ năm học 2020-2021 đến hết học kỳ 1 năm học 2023-2024. Đây là tập dữ liệu nền tảng để mô hình học các quy luật chung.
- **Tập Kiểm định (Validation Set):** Sử dụng dữ liệu học kỳ 2 năm học 2023-2024. Tập này dùng để tinh chỉnh tham số (Hyperparameter Tuning) và đánh giá khả năng dự báo của mô hình.
- **Tập Kiểm thử (Test Set):** Sử dụng dữ liệu học kỳ 1 năm học 2024-2025. Đây là tập dữ liệu hoàn toàn mới, dùng để đánh giá hiệu suất thực tế cuối cùng của mô hình.

III. PHƯƠNG PHÁP ĐỀ XUẤT

A. Hướng tiếp cận

Để giải quyết bài toán dự báo số tín chỉ hoàn thành (TC_HOANTHANH) - một biến số thực không âm và bị giới hạn bởi số tín chỉ đăng ký, ngoài việc dự báo trực tiếp, chúng tôi cũng đề xuất các phương pháp tiếp cận gián tiếp nhằm khai thác tối đa mối tương quan dữ liệu. Cụ thể, chúng tôi triển khai ba chiến lược mô hình hóa song song:

1) Chiến lược 1: Dự báo trực tiếp (Direct Credits prediction): Đây là hướng tiếp cận cơ bản nhất, trong đó mô hình học máy được huấn luyện để ánh xạ trực tiếp các đặc trưng đầu vào (tuyển sinh, lịch sử học tập) sang biến mục tiêu TC_HOANTHANH.

$$\hat{y}_{credits} = f_{credits}(X)$$

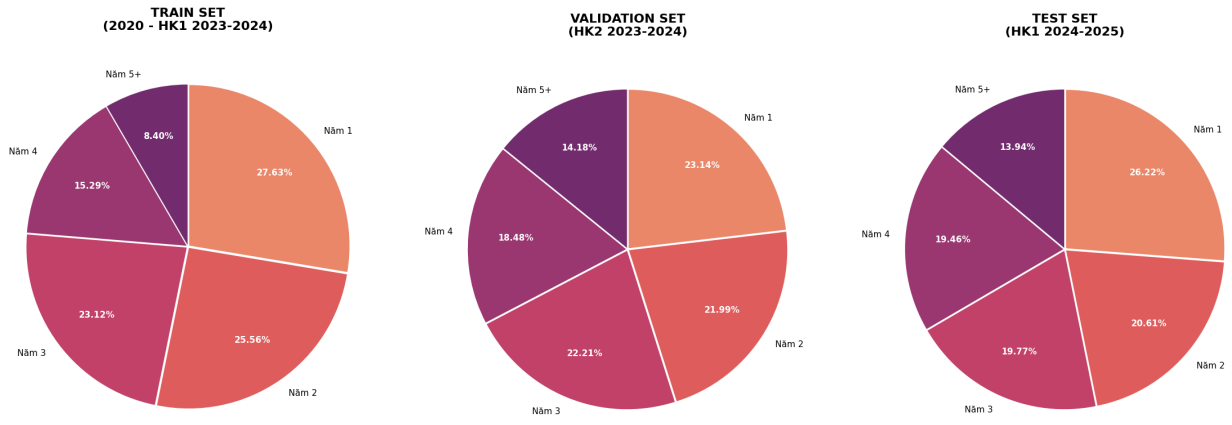
2) Chiến lược 2: Dự báo phần dư (Gap prediction): Nhận thấy rằng việc dự báo trực tiếp số tín chỉ hoàn thành thường gặp khó khăn do miền giá trị của biến mục tiêu trải rất rộng (dao động từ 0 đến hơn 40 tín chỉ) với độ phương sai lớn. Trong khi đó, số tín chỉ trượt/rút lại có đặc điểm phân phối hội tụ hơn rất nhiều.

Đặc biệt, đối với nhóm sinh viên có kết quả học tập tốt, số tín chỉ trượt thường đồng nhất bằng 0. Ngay cả đối với nhóm sinh viên có học lực yếu hơn, giá trị này cũng chỉ dao động trong một khoảng hẹp (thường nhỏ hơn 20 tín chỉ), thấp hơn đáng kể so với biên độ của tổng tín chỉ đăng ký. Do đó, chúng tôi đề xuất chiến lược mô hình hóa khoảng chênh lệch (GAP) thay vì dự báo trực tiếp:

$$\text{GAP} = \text{TC\_DANGKY} - \text{TC\_HOANTHANH}$$

Kết quả dự báo cuối cùng sẽ được suy diễn ngược lại:

$$\hat{y}_{gap} = \text{TC\_DANGKY} - \hat{f}_{gap}(X)$$



Hình 1: Phân bố sinh viên theo năm học trong các tập dữ liệu.

3) Chiến lược 3: Dự báo tỷ lệ (Ratio prediction): Một nhược điểm của hai chiến lược trước (Direct credits và Gap prediction) là biến mục tiêu phụ thuộc tuyến tính vào quy mô của TC_DANGKY. Miền giá trị của chúng không cố định (có thể từ 0 đến 40+), dẫn đến việc mô hình có thể bị bias bởi những sinh viên đăng ký quá ít hoặc quá nhiều tín chỉ.

Để khắc phục, hướng tiếp cận này chuẩn hóa biến mục tiêu về miền giá trị đóng $[0, 1]$, bất kể sinh viên đăng ký bao nhiêu tín chỉ. Điều này giúp chuyển bài toán từ dự báo "số lượng" sang dự báo "xác suất hoàn thành", giúp mô hình học được bản chất năng lực sinh viên ổn định hơn.

$$\texttt{RATIO} = \frac{\texttt{TC\_HOANTHANH}}{\texttt{TC\_DANGKY}}$$

Giá trị dự báo sau đó được chuyển đổi về đơn vị tín chỉ:

$$\hat{y}_{ratio} = \texttt{TC\_DANGKY} \times \hat{f}_{ratio}(X)$$

B. Mô hình đề xuất

Đặc thù của dữ liệu giáo dục nói chung và bài toán dự báo tiến độ học tập nói riêng nằm ở tính hỗn hợp cao (bao gồm cả biến định lượng và định tính) cùng sự tồn tại của các mối tương quan phi tuyến phức tạp. Để giải quyết bài toán này một cách toàn diện, chúng tôi không giới hạn trong một thuật toán duy nhất mà thiết lập một quy trình thực nghiệm trên bốn kiến trúc mô hình khác nhau. Các mô hình này đại diện cho sự phát triển từ cơ bản (Single Tree) đến các kỹ thuật tổ hợp tiên tiến (Bagging, Boosting), nhằm tìm ra sự cân bằng tối ưu giữa độ chính xác (Bias) và độ ổn định (Variance).

1) Decision Tree Regressor:

Chúng tôi sử dụng Decision Tree [4] làm mô hình cơ sở (baseline) để thiết lập mức chuẩn đánh giá hiệu năng. Mặc dù khả năng tổng quát hóa thường thấp hơn các phương pháp Ensemble, Decision Tree có giá trị đặc biệt trong giai đoạn đầu của quá trình mô hình hóa nhờ đặc tính "hộp trắng" cho phép minh bạch hóa cơ chế ra quyết định.

Cơ chế phân chia không gian dữ liệu dựa trên các luật "Nếu - Thì" của mô hình cho phép chúng tôi trực quan hóa các ngưỡng quan trọng của các biến đầu vào, ví dụ: xác định điểm cắt của GPA mà tại đó xu hướng hoàn thành tín chỉ thay đổi đột ngột. Tuy nhiên, nhược điểm cố hữu của mô hình này là độ nhạy cao với nhiễu, dễ dẫn đến hiện tượng overfitting khi cây phát triển quá sâu, do đó chỉ đóng vai trò tham chiếu so sánh.

2) Random Forest Regressor:

Để khắc phục tính bất ổn định của cây quyết định đơn lẻ, chúng tôi áp dụng Random Forest [5], một thuật toán thuộc họ Bagging. Về mặt kỹ thuật, mô hình này xây dựng song song hàng loạt cây quyết định trên các tập con dữ liệu được lấy mẫu ngẫu nhiên có hoàn lại (Bootstrap samples).

Sự ưu việt của Random Forest trong bài toán này nằm ở cơ chế trung bình hóa các dự báo từ các cây thành phần. Đối với dữ liệu lịch sử học tập, vốn thường chứa nhiễu do các yếu tố ngoại cảnh tác động đến sinh viên, việc kết hợp kết quả từ nhiều cây độc lập giúp giảm thiểu phương sai một cách hiệu quả. Điều này mang lại độ ổn định cao hơn đáng kể, đảm bảo mô hình không bị chi phối bởi các điểm dữ liệu bất thường (outliers).

3) XGBoost Regressor:

Tiếp cận bài toán từ góc độ giảm độ chệch (Bias Reduction), chúng tôi sử dụng XGBoost (Extreme Gradient Boosting) [6], một phiên bản tối ưu hóa cao của kỹ thuật Gradient Boosting [7]. Khác với cơ chế xây dựng song song của Random Forest, XGBoost hoạt động theo cơ chế tuần tự, trong đó mỗi cây mới được sinh ra nhằm mục đích sửa lỗi (tối ưu hoá phần dư - residuals) của các cây trước đó.

Điểm mạnh kỹ thuật khiến chúng tôi lựa chọn XGBoost là khả năng tối ưu hóa hàm mất mát thông qua khai triển Taylor bậc hai, giúp mô hình hội tụ nhanh và chính xác hơn. Bên cạnh đó, thuật toán này sở hữu cơ chế "Sparsity-aware split finding" [8], cho phép xử lý tự động và hiệu quả các giá trị khuyết. Đây là tính năng quan trọng khi xử lý dữ liệu của nhóm sinh viên năm nhất hoặc sinh viên quay lại sau bảo lưu, nơi lịch sử học tập thường không liên tục.

4) LightGBM Regressor:

Đóng vai trò là mô hình chủ lực trong giải pháp đề xuất, LightGBM (Light Gradient Boosting Machine) [9] được lựa chọn nhờ khả năng xử lý vượt trội trên tập dữ liệu lớn với tốc độ huấn luyện cao. Sự khác biệt cốt lõi của LightGBM nằm ở chiến lược phát triển cây theo lá (Leaf-wise growth) thay vì theo tầng (Level-wise) như các thuật toán truyền thống. Chiến lược này cho phép mô hình giảm thiểu hàm mất mát nhanh hơn, mặc dù đòi hỏi sự kiểm soát chặt chẽ các tham số điều chuẩn để tránh overfitting.

Đặc biệt, để phù hợp với bản chất của biến mục tiêu TC_HOANTHANH (số thực không âm, thường có phân phối lệch phải và tập trung nhiều giá trị 0), chúng tôi thay thế hàm mất mát MSE truyền thống bằng hàm mục tiêu phân phối **Tweedie** [10]. Phân phối Tweedie (với tham số phương sai $1 < p < 2$) cho phép mô hình học tốt hơn cấu trúc của dữ liệu, qua đó cải thiện đáng kể độ chính xác tại các điểm dữ liệu có giá trị bằng 0 hoặc giá trị nhỏ, cũng chính là nhóm đối tượng rủi ro cao mà bài toán hướng tới giải quyết.

Cả bốn mô hình đều được tối ưu hóa các siêu tham số sử dụng thư viện Optuna [11] (Chi tiết hơn trình bày tại Phụ lục C).

C. Khung tích hợp giải thích mô hình (xAI)

Việc ứng dụng các mô hình học máy dạng "hộp đen" như LightGBM hay XGBoost thường gặp rào cản lớn trong công tác tư vấn giáo dục do thiếu tính minh bạch. Để khắc phục điều này, chúng tôi xây dựng một khung giải thích đa tầng, kết hợp ba kỹ thuật SHAP, LIME và DiCE nhằm giải quyết ba bài toán cốt lõi: định lượng yếu tố ảnh hưởng, kiểm chứng logic dự báo và đề xuất hành động cụ thể.

1) SHAP:

Chúng tôi sử dụng SHAP (Shapley Additive Explanations) [12] làm công cụ định lượng mức độ đóng góp của các đặc trưng dựa trên lý thuyết trò chơi. Một số vai trò của SHAP bao gồm:

- Phân tích toàn cục: Thông qua biểu đồ Beeswarm, hệ thống xác định các biến số có tác động lớn nhất đến kết quả dự báo trên toàn bộ tập dữ liệu.
- Phân tích đối chiếu: Đây là kỹ thuật nâng cao được chúng tôi triển khai trong mã nguồn nhằm so sánh trực tiếp giữa một sinh viên có dự báo rủi ro cao và một sinh viên có hồ sơ tương đồng nhưng đạt kết quả tốt. Việc tính toán hiệu số giá trị SHAP giúp chỉ ra chính xác khoảng cách năng lực thực tế giữa hai đối tượng.

2) LIME:

Trong khi SHAP cung cấp cái nhìn tổng quan và nhất quán, LIME (Local Interpretable Model-agnostic Explanations) [13] được sử dụng để giải thích hành vi của mô hình tại các điểm dữ liệu cụ thể. Bằng cách xấp xỉ mô hình gốc bằng một mô hình tuyến tính đơn giản xung quanh điểm cần giải thích, LIME cho phép kiểm tra độ tin cậy của các dự báo đối với các trường hợp cá biệt hoặc dữ liệu ngoại lai (outliers). Điều này giúp người dùng có thể xác nhận lại lý do mô hình đưa ra cảnh báo (ví dụ: do trượt môn kỳ trước hay do đăng ký quá tải) có phù hợp với nghiệp vụ sư phạm hay không.

3) DiCE:

Khác với hai kỹ thuật trên tập trung vào câu hỏi "tại sao?", DiCE (Diverse Counterfactual Explanations) [14] được tích hợp để giải quyết bài toán "làm thế nào?". Chúng tôi sử dụng DiCE để sinh ra các phản ví dụ (counterfactuals), những kịch bản giả định cho thấy sự thay đổi nhỏ nhất cần thiết để đảo ngược kết quả dự báo.

Cụ thể, thuật toán sẽ tìm kiếm các thay đổi trên những biến có thể can thiệp được như số tín chỉ đăng ký, nhằm đưa dự báo tỷ lệ hoàn thành của sinh viên từ mức rủi ro về ngưỡng an toàn (ví dụ: từ dưới 50% lên trên 70%). Kết quả này là cơ sở trực tiếp để hệ thống đưa ra các khuyến nghị điều chỉnh lộ trình học tập khả thi cho từng sinh viên.

IV. KẾT QUẢ THỰC NGHIỆM

A. Kết quả dự đoán

Để đánh giá toàn diện hiệu năng của các mô hình hồi quy, chúng tôi sử dụng bộ bốn chỉ số đo lường tiêu chuẩn được đề xuất bởi ban tổ chức, bao gồm: MSE, $RMSE$, R^2 và $wMAPE$ [15]. Gọi y_i là giá trị thực tế (số tín chỉ hoàn thành), $\hat{y}_i$ là giá trị dự báo, $\bar{y}$ là trung bình của các giá trị thực tế, và n là tổng số mẫu dữ liệu. Các chỉ số được định nghĩa cụ thể như sau:

- **Mean Squared Error (MSE) - Sai số bình phương trung bình:** Đo lường trung bình bình phương độ chênh lệch giữa giá trị dự báo và giá trị thực tế. MSE phạt rất nặng các sai số lớn (outliers), giúp mô hình chú trọng vào việc giảm thiểu các sai lệch nghiêm trọng.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1}$$

- **Root Mean Squared Error ($RMSE$) - Căn bậc hai của sai số bình phương trung bình:** Là căn bậc hai của MSE. Ưu điểm của $RMSE$ là đưa đơn vị đo lường sai số về cùng đơn vị với biến mục tiêu (số tín chỉ), giúp kết quả dễ diễn giải hơn về mặt thực tiễn.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

- **R-squared (R^2) - Hệ số xác định:** Biểu thị tỷ lệ phương sai của biến mục tiêu có thể được giải thích bởi các biến độc lập trong mô hình. Chỉ số này cho biết mức độ phù hợp của mô hình so với một mô hình tham chiếu đơn giản (trung bình).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3}$$

- **Weighted Mean Absolute Percentage Error ($wMAPE$):** Thông thường, chỉ số $MAPE$ được ưa chuộng để đo sai số phần trăm. Tuy nhiên, trong tập dữ liệu thực tế của bài toán này, tồn tại nhiều trường hợp sinh viên có số tín chỉ hoàn thành bằng 0 ($y_i = 0$), dẫn đến việc tính toán $MAPE$ bị lỗi (chia cho 0) hoặc tạo ra các giá trị vô cùng. Do đó, chúng tôi đề xuất sử dụng $wMAPE$ để khắc phục nhược điểm này bằng cách tính tổng sai số tuyệt đối chia cho tổng giá trị thực tế.

$$wMAPE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{\sum_{i=1}^{n}|y_i|} \tag{4}$$

Mặc dù cả bốn chỉ số trên đều cung cấp những góc nhìn quan trọng về hiệu suất mô hình, $RMSE$ sẽ được chúng tôi ưu tiên xem xét là **chỉ số đánh giá chính** (Primary Metric). Lý do là bài toán này được cấu trúc dưới dạng một cuộc thi dữ liệu, trong đó Ban tổ chức đã thiết lập $RMSE$ làm tiêu chí xếp hạng chính thức cho tập kiểm thử (Test Set). Việc tối ưu hóa $RMSE$ không chỉ đảm bảo độ chính xác tổng thể mà còn giúp giải pháp bám sát yêu cầu thực tiễn của cuộc thi.

B. Đánh giá kết quả

Dựa trên Bảng I, chúng tôi thực hiện đánh giá hiệu năng của các mô hình dựa trên hai tiêu chí chính: độ ổn định trên tập Validation và khả năng tổng quát hóa trên tập Public Test.

1) So sánh giữa các nhóm thuật toán:

Nhìn chung, các mô hình tập hợp (Ensemble Learning) bao gồm LightGBM, XGBoost và Random Forest đều cho kết quả vượt trội hoàn toàn so với Decision Tree đơn lẻ. Decision Tree ghi nhận sai số $RMSE$ cao nhất ở cả 3 chiến lược tiếp cận (dao động từ 3.87 đến 3.97 trên tập Validation), cho thấy mô hình này dễ bị quá khớp (overfitting) và thiếu tính ổn định khi xử lý dữ liệu sinh viên có độ nhiễu cao.

2) So sánh chiến lược tiếp cận:

- **Về độ chính xác trên tập Validation:** Chiến lược *Gap Prediction* (dự báo phần dư) kết hợp với LightGBM thể hiện sự thống trị tuyệt đối khi đạt $RMSE$ thấp nhất (3.7295) và R^2 cao nhất (0.7178). Điều này chứng minh rằng việc chuyển đổi bài toán sang dự báo khoảng chênh lệch giúp mô hình học được các mẫu hình dữ liệu tốt hơn trong giai đoạn huấn luyện.
- **Về chiến lược Ratio Prediction:** Mặc dù từng được kỳ vọng cao, nhưng trên tập dữ liệu này, chiến lược dự báo tỷ lệ chỉ đứng thứ hai về độ chính xác trên tập Validation (với LightGBM đạt $RMSE = 3.7375$) và không giữ được phong độ tốt nhất khi kiểm thử thực tế.

3) Hiệu năng trên tập Public Test:

Một hiện tượng thú vị được quan sát thấy là sự "đảo chiều"về thứ hạng giữa tập Validation và Public Test.

Mặc dù *Gap Prediction* dẫn đầu trong quá trình huấn luyện, nhưng chiến lược **Direct Credits Prediction** (dự báo trực tiếp) sử dụng mô hình **LightGBM** lại chứng minh khả năng tổng quát hóa tốt nhất trên dữ liệu chưa từng thấy (Unseen Data), đạt Hạng 1 với $RMSE = 3.5040$.

Đứng ở vị trí thứ hai là *Gap Prediction* sử dụng **Random Forest** ($RMSE = 3.5161$), vượt qua cả XGBoost và

		Validation Set				Public Test Set
Approach	**Model**	R^2 (↑)	$RMSE$ (↓)	MSE (↓)	$wMAPE$ (↓)	$RMSE$ (↓)
Direct Credits Prediction	Decision Tree	0.6794	3.9749	15.8002	0.1928	3.7589
	Random Forest	0.7069	3.7831	14.3115	0.1832	3.5332
	XGBoost	0.7149	3.7488	14.0534	0.1786	3.5431
	LightGBM	0.7147	3.7499	14.0620	**0.1777**	**3.5040**
Gap Prediction	Decision Tree	0.6948	3.8785	15.0427	0.1911	3.6859
	Random Forest	0.7148	3.7491	14.0556	0.1809	**3.5161**
	XGBoost	0.7152	3.7469	14.0395	**0.1778**	3.5203
	LightGBM	**0.7178**	**3.7295**	**13.9091**	**0.1781**	**3.5171**
Ratio Prediction	Decision Tree	0.6961	3.8703	14.9789	0.1870	3.6492
	Random Forest	0.6446	4.1857	17.5198	0.1985	3.5758
	XGBoost	**0.7173**	**3.7328**	**13.9340**	0.1792	3.5371
	LightGBM	**0.7166**	**3.7375**	**13.9687**	0.1800	3.5205

So sánh toàn cục: ***Hạng 1*** , ***Hạng 2*** , ***Hạng 3***

Bảng I: Bảng so sánh hiệu năng các mô hình.

LightGBM cùng chiến lược. Sự chênh lệch giữa vị trí nhất và nhì là rất nhỏ (khoảng 0.012 đơn vị RMSE), cho thấy tính cạnh tranh cao giữa các giải pháp.

Như vậy mô hình **LightGBM** với chiến lược **Direct Credits Prediction** được lựa chọn là mô hình cuối cùng để đệ trình. Mặc dù không đạt điểm số cao nhất trên tập Validation, nhưng nó thể hiện sự ổn định và hiệu quả thực tế tốt nhất, tránh được hiện tượng học vẹt (overfitting) mà các chiến lược phức tạp hơn (như Gap hay Ratio) có thể mắc phải.

C. Đánh giá mô hình

Trong bối cảnh ứng dụng vào môi trường giáo dục, việc đạt được độ chính xác cao ($RMSE$ thấp) chỉ là điều kiện cần. Để hệ thống thực sự trở thành một công cụ hỗ trợ ra quyết định đáng tin cậy cho cố vấn học tập và sinh viên, cơ chế hoạt động của mô hình cần phải minh bạch và dễ hiểu. Do đó, chúng tôi không chỉ dừng lại ở việc so sánh các chỉ số hiệu năng định lượng, mà còn đi sâu vào phân tích bản chất hành vi của mô hình thông qua các kỹ thuật **Explainable AI (xAI)**.

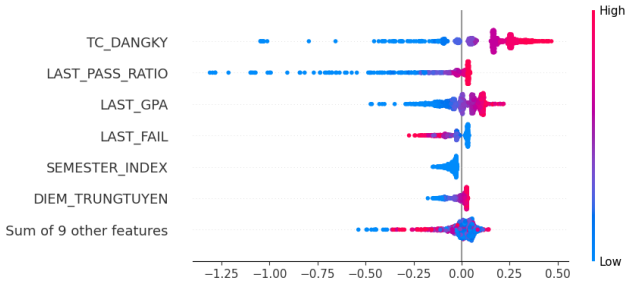
1) Phân tích toàn cục:

Biểu đồ SHAP Summary (dạng Beeswarm) minh họa sự phân bố tác động của các đặc trưng hàng đầu đối với hai loại mô hình ứng với hai nhóm sinh viên là *Fresher* và *Senior*, trong đó:

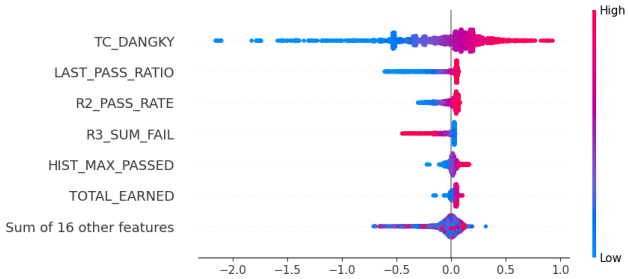
- **Trục tung:** Các đặc trưng được sắp xếp theo mức độ quan trọng giảm dần từ trên xuống dưới.
- **Trục hoành (SHAP value):** Tác động đến kết quả dự báo. Giá trị dương (bên phải) làm tăng số tín chỉ dự báo, giá trị âm (bên trái) làm giảm số tín chỉ.
- **Màu sắc:** Thể hiện giá trị thực tế của đặc trưng (Đỏ: Cao, Xanh: Thấp).

Dựa trên Hình 2, chúng tôi rút ra các nhận định quan trọng sau:

- **Khối lượng đăng ký** (`TC_DANGKY`): Là yếu tố quan trọng nhất. Mối quan hệ là tuyến tính thuận: đăng ký nhiều tín chỉ (màu đỏ) dẫn đến dự báo hoàn thành cao (SHAP dương).
- **Đà học tập ngắn hạn:** Các biến đặc biệt quan trọng với sinh viên năm nhất kỳ hai là `LAST_PASS_RATIO` và `LAST_GPA`. Đáng chú ý, các điểm màu xanh dương (tỷ lệ qua môn thấp) của `LAST_PASS_RATIO` kéo dài về phía bên trái, cho thấy việc trượt môn ở kỳ trước là một "hình phạt"rất nặng, kéo tụt kết quả dự báo xuống đáng kể.
- **Vai trò của điểm đầu vào:** Đặc trưng điểm thi đầu vào của sinh viên (`DIEM_TRUNGTUYEN`) xuất hiện trong top các yếu tố ảnh hưởng của nhóm Fresher, nhưng không nằm trong top các đặc trưng quan trọng của nhóm Senior. Điều này chứng minh giả thuyết của nhóm: sau năm thứ nhất, năng lực thực tế tại đại học lấn át hoàn toàn kết quả thi THPT.
- **Tầm quan trọng của lịch sử dài hạn:** Các đặc trưng mang tính tích lũy như `HIST_MAX_PASSED` (Kỷ lục số tín chỉ từng qua) hay `TOTAL_EARNED` đóng vai trò trọng yếu khi mô hình dự đoán. Những sinh viên có "kỷ lục"cao trong quá khứ (màu đỏ) có xu hướng duy



(a) Mô hình Fresher (Sinh viên năm nhất)



(b) Mô hình Senior (Sinh viên năm hai trở đi)

Hình 2: Mức độ đóng góp của các đặc trưng đối với từng nhóm sinh viên.

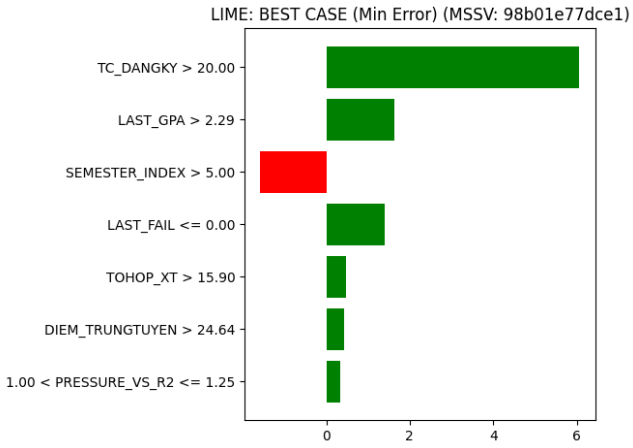
trì phong độ tốt.

- **Xu hướng rủi ro:** Đặc trưng `R3_SUM_FAIL` (Tổng số tín chỉ trượt ba kỳ gần nhất) có tác động tiêu cực rõ rệt. Những điểm màu đỏ (trượt nhiều) nằm hoàn toàn về phía giá trị SHAP âm.

Như vậy có thể thấy việc phân tách dữ liệu thành hai nhóm và dùng hai mô hình riêng biệt giúp chúng tôi nắm bắt chính xác các động lực học tập khác nhau: Nhóm Fresher dựa vào *năng lực đầu vào* và *sự thích nghi ban đầu*, trong khi nhóm Senior dựa vào *sự bền bỉ* và *lịch sử tích lũy* (Chi tiết hơn trình bày tại Phụ lục D).

2) Phân tích cục bộ:

Bên cạnh phân tích toàn cục với SHAP, chúng tôi sử dụng kỹ thuật LIME để giải thích hành vi của mô hình trên từng cá thể sinh viên, cho từng trường hợp dự báo tiêu biểu cả hai loại mô hình Fresher và Senior (Chi tiết được trình bày tại Phụ lục E).



Hình 3: Phân tích LIME cho trường hợp mô hình dự báo chính xác cao.

Quan sát Hình 3, một trường hợp dự báo tiêu biểu thuộc nhóm *Best Case* của mô hình Fresher, chúng ta có thể thấy rõ cơ chế ra quyết định của mô hình hoàn toàn phù hợp với logic giáo dục:

- **Tác động tích cực (Thanh màu xanh):** Yếu tố đóng góp lớn nhất vào kết quả dự báo là số tín chỉ đăng ký cao (`TC_DANGKY > 20.00`), cộng thêm xấp xỉ 6 tín chỉ vào kết quả dự báo. Bên cạnh đó, lịch sử học tập tốt thể hiện qua `LAST_GPA > 2.29` và không nợ môn (`LAST_FAIL <= 0.00`) cũng đóng vai trò là các yếu tố "thưởng", giúp nâng cao mức dự báo hoàn thành.
- **Tác động tiêu cực (Thanh màu đỏ):** Yếu tố `SEMESTER_INDEX > 5.00` đóng vai trò điều chỉnh giảm. Điều này phản ánh thực tế rằng khi sinh viên bước vào các kỳ học sau (hoặc học lại, học cải thiện cùng khóa sau), khối lượng kiến thức chuyên ngành khó hơn hoặc sự phân tán sự tập trung có thể làm giảm nhẹ khả năng hoàn thành tín chỉ tối đa.

Tóm lại phân tích cục bộ khẳng định rằng mô hình không chỉ đạt độ chính xác cao về mặt thống kê mà còn học được các quy luật nhân quả đúng đắn từ dữ liệu thực tế.

3) Phân tích phản chứng:

Chúng tôi tiếp tục đi thêm một bước nữa để đưa ra các khuyến nghị hành động cụ thể cho sinh viên có nguy cơ chậm tiến độ. Bảng II minh họa phân tích cho sinh viên mã số `bc6bd14ea87e` (thuộc nhóm Senior - Kỳ 7). Mô hình hiện tại dự báo tỷ lệ tín chỉ hoàn thành ở mức thấp (**0.4846**). DiCE đề xuất các kịch bản thay thế để nâng mức dự báo lên nhóm an toàn (> 0.7).

Đặc trưng	**Trạng thái hiện tại**	**Kịch bản 1**	**Kịch bản 2**
Tỉ lệ hoàn thành	0.48	0.78	0.85
LAST_PASS_RATIO	0.46	0.65	0.80
R2_PASS_RATE	0.16	0.82	0.73
FAIL_TREND_R2	-3.00	-	-
TOTAL_EARNED	100.0	38.8	58.2

Bảng II: Gợi ý cải thiện cho sinh viên bc6bd14ea87e

- **Vấn đề cốt lõi:** Sinh viên đang chịu ảnh hưởng tiêu cực từ phong độ ngắn hạn. Tỷ lệ qua môn kỳ liền trước (`LAST_PASS_RATIO`) chỉ đạt 46%, và trung bình 2 kỳ gần nhất (`R2_PASS_RATE`) rất thấp (16%), cho thấy sinh viên đang trong đà trượt dốc (Fail Trend).
- **Khuyến nghị hành động:** Các kịch bản phản chứng chỉ ra rằng, để quay lại vùng an toàn, sinh viên không nhất thiết phải đăng ký ít tín chỉ đi, mà bắt buộc phải cải thiện **hiệu suất qua môn**. Cụ thể, sinh viên cần đảm bảo tỷ lệ qua môn ở kỳ tiếp theo đạt ít nhất **60% - 80%** để khôi phục niềm tin của mô hình vào khả năng hoàn thành tiến độ ở các kỳ tiếp theo.
- **Ý nghĩa:** Kết quả từ DiCE cung cấp cơ sở định lượng để Cố vấn học tập đặt mục tiêu cụ thể cho sinh viên (ví dụ: "Em cần phải qua được 60% số môn kỳ này"), thay vì chỉ đưa ra lời khuyên chung chung.

V. KẾT LUẬN

Trong khuôn khổ của vòng loại cuộc thi "*DataFlow 2026: The Alchemy of Minds*", chúng tôi đề xuất hệ thống **TRUSTEE** giải quyết bài toán dự báo tiến độ học tập thông qua chiến lược mô hình hóa phân tầng (Fresher/Senior) và chuẩn hóa dữ liệu phổ điểm tuyển sinh. Kết quả thực nghiệm cho thấy mô hình LightGBM sử dụng hàm mất mát Tweedie đạt hiệu năng vượt trội ($RMSE = 3.5040$). Đặc biệt, giá trị thực tiễn của giải pháp nằm ở việc tích hợp khung giải thích đa tầng (SHAP, LIME, DiCE) giúp chuyển hóa các dự báo kỹ thuật thành các khuyến nghị hành động cụ thể, biến thuật toán blackbox thành công cụ tư vấn học tập minh bạch.

Tuy nhiên, giới hạn của mô hình hiện tại là sự phụ thuộc vào dữ liệu kết quả học tập mang tính thời điểm, chưa phản ánh được các biến động hành vi liên tục trong quá trình học. Kế hoạch phát triển tiếp theo sẽ tập trung mở rộng nguồn dữ liệu đa chiều (như mức độ chuyên cần, hoạt động tương tác) và thử nghiệm các kiến trúc Deep learning (như LSTM, Transformers) để nắm bắt tốt hơn các phụ thuộc chuỗi thời gian. Về mặt thực tiễn, chúng tôi hướng đến việc xây dựng một Dashboard tương tác tích cho phép sinh viên chủ động mô phỏng và tối ưu hóa lộ trình học tập của chính mình.

TÀI LIỆU THAM KHẢO

[1] Loh, W. Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1), 14-23.

[2] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82-115.

[3] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 31.

[4] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA.

[5] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

[6] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).

[7] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189-1232.

[8] He, X., Pan, J., Jin, O., Xu, T., Liu, B., Tao, T., ... & Candela, J. Q. (2014). Practical lessons from predicting clicks on ads at Facebook. In *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising* (pp. 1-9).

[9] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 3146-3154.

[10] Tweedie, M. C. K. (1984). An index which distinguishes between some important exponential families. In *Statistics: Applications and New Directions: Proceedings of the Indian Statistical Institute Golden Jubilee International Conference* (pp. 579-604).

[11] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 2623-2631).

[12] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

[13] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135-1144).

[14] Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 607-617).

[15] Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679-688.

[16] Alyahyan, E., & Düştegör, D. (2020). Predicting academic performance in higher education: a systematic review of the past decade. *IEEE Access*, 8, 79694-79717.

[17] Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, 113, 177-194.

[18] Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., & Durán-Domínguez, A. (2020). Analyzing and predicting students' performance by means of machine learning: A review. *Applied Sciences*, 10(3), 1042.

PHỤ LỤC

A. Dữ liệu Phổ điểm thi THPT Quốc gia giai đoạn 2020 - 2024

Dữ liệu tổng hợp kết quả thi của thí sinh trong kỳ thi Tốt nghiệp Trung học Phổ thông Quốc gia (THPTQG) tại Việt Nam giai đoạn 2020 - 2024. Dữ liệu bao gồm 5 tập tin định dạng CSV, được thu thập và tổng hợp từ các nguồn công khai sau đây:

- Năm 2020: Kaggle
- Năm 2021: Kaggle
- Năm 2022: Github
- Năm 2023: Github
- Năm 2024: Github

B. Các đặc trưng sử dụng

Quá trình trích chọn đặc trưng (Feature Selection) đóng vai trò then chốt trong việc nâng cao hiệu suất mô hình. Ban đầu, chúng tôi tạo ra một tập hợp lớn các đặc trưng bao gồm thông tin tuyển sinh, lịch sử học tập tích lũy và các chỉ số xu hướng.

Sau đó, chúng tôi sử dụng kỹ thuật đánh giá tầm quan trọng của đặc trưng từ mô hình *Decision Tree* và *XGBoost* để loại bỏ các biến nhiễu hoặc ít mang lại giá trị thông tin. Kết quả cuối cùng là hai bộ đặc trưng tối ưu riêng biệt cho mô hình Fresher và Senior như trình bày dưới đây.

Bảng III: Bộ đặc trưng cho mô hình Fresher

Tên đặc trưng	**Ý nghĩa / Mô tả**
TC_DANGKY	Số tín chỉ đăng ký
SEMESTER_INDEX	Chỉ số thứ tự học kỳ
PTXT	Mã phương thức xét tuyển
TOHOP_XT	Mã tổ hợp môn xét tuyển
DIEM_TRUNGTUYEN	Tổng điểm trúng tuyển đầu vào
DIEM_CHUAN	Điểm chuẩn đầu vào năm đó
SCORE_GAP	Chênh lệch điểm trúng tuyển và điểm chuẩn
Z_SCORE	Điểm chuẩn hóa theo phổ điểm năm
ENTRY_RANK	Thứ hạng phần trăm của sinh viên trong khóa
BENCHMARK_TIER	Phân loại xếp hạng của điểm tổ hợp xét tuyển
GAP_RATIO	Tỷ lệ vượt so với điểm chuẩn
LAST_GPA	GPA kỳ trước (cho sinh viên kỳ 2 năm nhất)
LAST_FAIL	Số tín chỉ trượt kỳ trước (cho sinh viên kỳ 2 năm nhất)
LAST_PASS_RATIO	Tỷ lệ qua môn kỳ trước (cho sinh viên kỳ 2 năm nhất)
PRESSURE_VS_R2	Áp lực học tập (cho sinh viên kỳ 2 năm nhất)

Bảng IV: Bộ đặc trưng cho mô hình Senior

Tên đặc trưng	**Ý nghĩa / Mô tả**
TC_DANGKY	Số tín chỉ đăng ký
SEMESTER_INDEX	Chỉ số thứ tự học kỳ
SV_NAM_THU	Sinh viên năm thứ mấy
LAST_GPA	GPA kỳ trước
LAST_FAIL	Số tín chỉ trượt kỳ trước
LAST_PASS_RATIO	Tỷ lệ tín chỉ hoàn thành kỳ trước
R2_AVG_GPA	Trung bình GPA của 2 kỳ gần nhất
R2_SUM_FAIL	Tổng tín chỉ trượt trong 2 kỳ gần nhất
R2_PASS_RATE	Tỷ lệ hoàn thành tín chỉ trong 2 kỳ gần nhất
FAIL_TREND_R2	Xu hướng trượt
GPA_TREND_R2	Xu hướng điểm số
R3_AVG_GPA	Trung bình GPA của 3 kỳ gần nhất
R3_SUM_FAIL	Tổng tín chỉ trượt trong 3 kỳ gần nhất
PRESSURE_VS_R2	Áp lực so với 2 kỳ gần nhất
PRESSURE_VS_R3	Áp lực so với 3 kỳ gần nhất
OVERLOAD_R3	Mức độ quá tải so với trung bình 3 kỳ gần nhất
TOTAL_EARNED	Tổng số tín chỉ tích lũy từ đầu khóa
OVERLOAD_VS_MAX	So sánh đăng ký với kỷ lục cao nhất quá khứ
HIST_AVG_GPA	Điểm GPA trung bình tích lũy toàn khóa
HIST_MAX_PASSED	Kỷ lục số tín chỉ hoàn thành trong 1 kỳ
HIST_MAX_GPA	Kỷ lục GPA
HIST_STD_GPA	Độ lệch chuẩn GPA

C. Tối ưu Siêu Tham số (Hyperparameters)

Thay vì áp dụng các phương pháp truyền thống như *Grid Search* (tìm kiếm vét cạn, tốn kém tài nguyên tính toán) hay *Random Search* (tìm kiếm ngẫu nhiên, thiếu tính định hướng), chúng tôi quyết định chọn thư viện **Optuna**, một khung phần mềm tối ưu hóa tự động thế hệ mới với hiệu suất vượt trội, sử dụng thuật toán *Tree-structured Parzen Estimator*

(TPE). Bên cạnh đó, chúng tôi cũng tận dụng cơ chế *Pruning* (cắt tỉa) của Optuna để tự động dừng sớm các thử nghiệm (trials) cho thấy hiệu quả kém ngay từ những vòng lặp đầu tiên, giúp tiết kiệm tối đa thời gian huấn luyện.

Hàm mục tiêu (Objective Function) của quá trình tối ưu được thiết lập để tối thiểu hóa chỉ số *Root Mean Squared Error* ($RMSE$) trên tập kiểm định (Validation Set). Để đảm bảo tính minh bạch và khả năng tái lập kết quả thực nghiệm (reproducibility), toàn bộ quá trình tìm kiếm không gian tham số đều được cố định với `random_seed=42`.

Chi tiết về thời gian tối ưu và kết quả tham số của mỗi mô hình nhận được có thể xem chi tiết tại đây.

1) Decision Tree Regressor: Quá trình dò tìm tham số cho mô hình Cây quyết định được thực hiện với 300 trials trong khoảng thời gian trung bình 4.76 phút.

Tham số	**Vùng tìm kiếm**
criterion	{squared_error, friedman_mse, poisson}
splitter	{best, random}
max_depth	[3, 50]
max_leaf_nodes	[10, 500]
min_samples_split	[2, 40]
min_samples_leaf	[1, 20]
max_features	[0.1, 1.0]
ccp_alpha	[0.0, 0.05]
min_impurity_decrease	[0.0, 0.1]

Bảng V: Không gian tham số Decision Tree

2) Random Forest Regressor: Mô hình Rừng ngẫu nhiên được tối ưu hóa tập trung vào cấu trúc cây và phương pháp Bagging. Thực hiện dò tìm với 150 trials trong trung bình 359.33 phút.

Tham số	**Vùng tìm kiếm**
n_estimators	1000
criterion	{squared_error, friedman_mse, poisson}
max_depth	[5, 50]
max_features	[0.1, 1.0]
min_samples_split	[2, 40]
min_samples_leaf	[1, 20]
bootstrap	{True, False}
ccp_alpha	[0.0, 0.05]
min_impurity_decrease	[0.0, 0.1]
max_samples	[0.5, 0.99]

Bảng VI: Không gian tham số Random Forest

3) XGBoost Regressor: Đối với XGBoost, chúng tôi sử dụng phân phối Tweedie và tối ưu hóa sâu các tham số điều chuẩn (regularization). Thực hiện dò tìm với 300 trials trong trung bình 125.33 phút.

Tham số	**Vùng tìm kiếm**
objective	reg:tweedie
tree_method	hist
n_estimators	4000
early_stopping_rounds	100
learning_rate	[1e-4, 0.3]
tweedie_variance_power	[1.01, 1.99]
max_delta_step	[0.0, 10.0]
max_depth	[3, 12]
min_child_weight	[1, 100]
gamma	[1e-8, 10.0]
grow_policy	{depthwise, lossguide}
max_bin	{256, 512}
subsample	[0.5, 1.0]
colsample_bytree	[0.5, 1.0]
reg_alpha	[1e-8, 100.0]
reg_lambda	[1e-8, 100.0]
max_leaves	[16, 256]

Bảng VII: Không gian tham số cho XGBoost

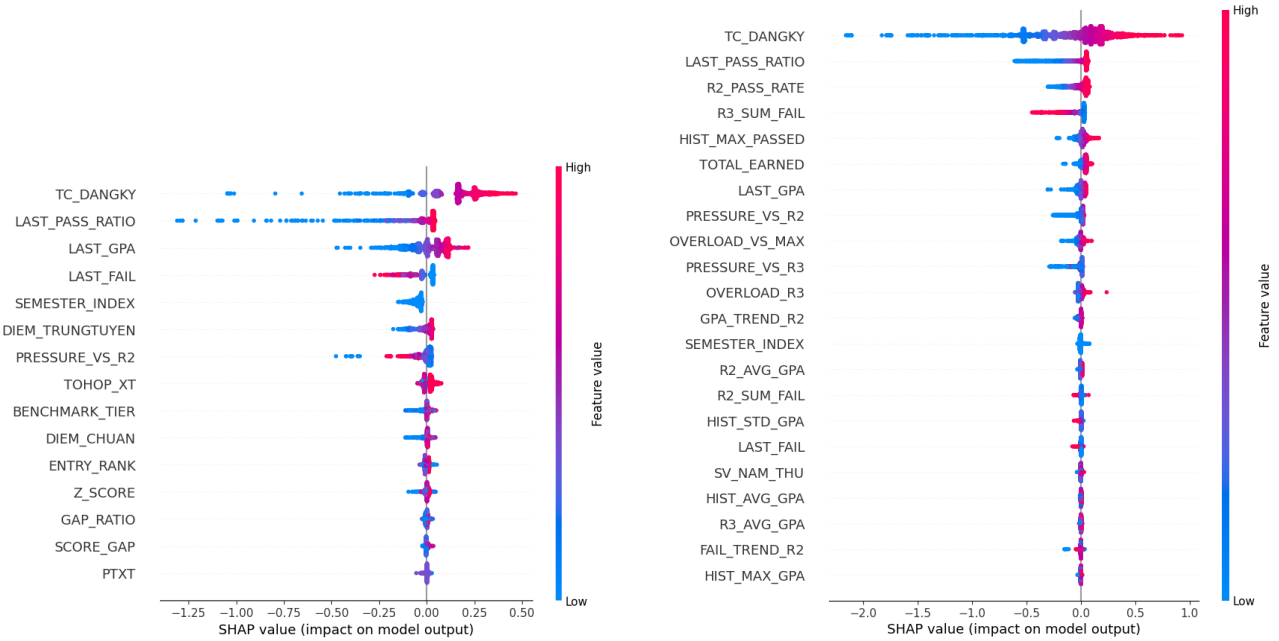
4) LightGBM Regressor: Tương tự XGBoost, LightGBM được cấu hình sử dụng thuật toán GBDT với phân phối Tweedie. Thực hiện dò tìm với 300 trials trong trung bình 68.84 phút.

Tham số	Vùng tìm kiếm
objective	tweedie
boosting_type	gbdt
boost_from_average	True
metric	rmse
n_estimators	4000
learning_rate	[1e-4, 0.1]
num_leaves	[10, 200]
max_depth	[3, 20]
min_data_in_leaf	[10, 100]
feature_fraction	[0.5, 1.0]
bagging_fraction	[0.5, 1.0]
lambda_l1	[1e-8, 10.0]
lambda_l2	[1e-8, 10.0]

Bảng VIII: Không gian tham số cho LightGBM

D. Chi tiết kết quả phân tích SHAP

Hình 4 cung cấp cái nhìn toàn cảnh về mức độ đóng góp của **toàn bộ bộ đặc trưng** được sử dụng trong quá trình huấn luyện.



(a) Mô hình Fresher

(b) Mô hình Senior

Hình 4: So sánh mức độ quan trọng của toàn bộ đặc trưng.

Phân tích mở rộng:

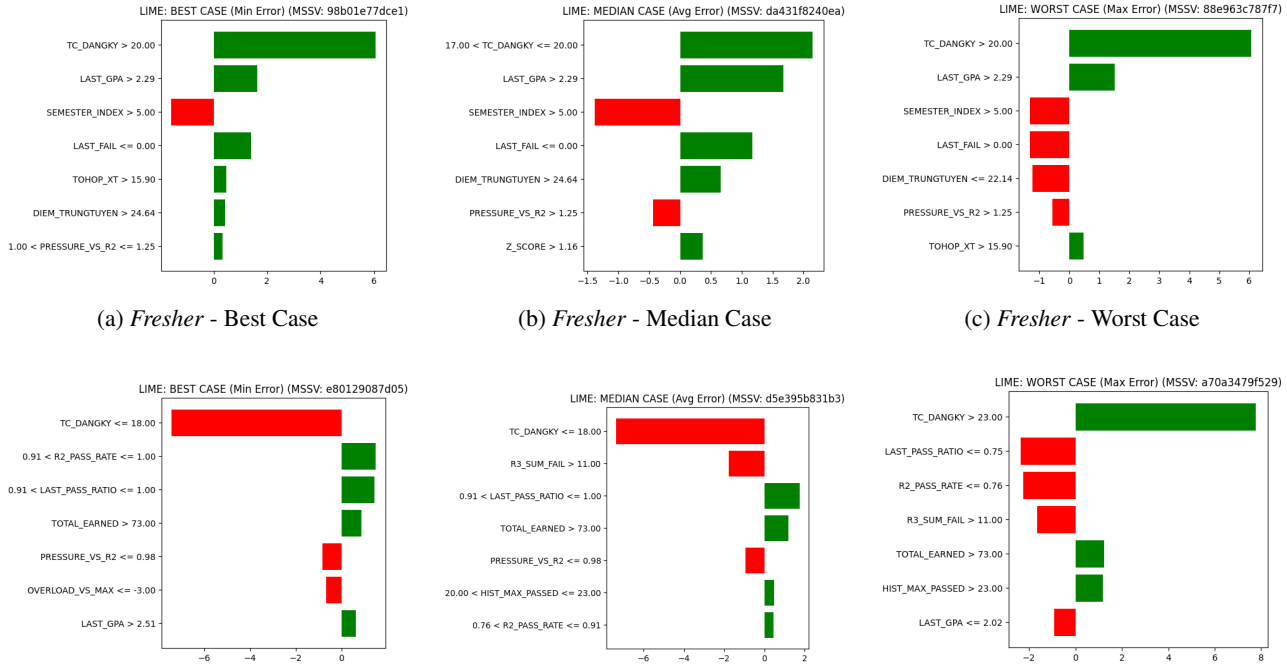
- **Đối với mô hình Fresher** (Hình 4a): Một phát hiện quan trọng là sự phân hóa rõ rệt trong nhóm đặc trưng tuyển sinh. Trong khi điểm số thô (`DIEM_TRUNGTUYEN`) vẫn giữ vai trò nhất định (Top 6), thì toàn bộ các *chỉ số phái sinh phức tạp* như `Z_SCORE`, `ENTRY_RANK`, hay `GAP_RATIO` đều nằm ở đáy bảng xếp hạng với giá trị SHAP xấp xỉ 0.
  → *Kết luận:* Đối với dự báo năm nhất, "độ khó" của việc trúng tuyển hay thứ hạng tương đối không quan trọng bằng hành vi đăng ký tín chỉ thực tế (`TC_DANGKY`) và năng lực thích nghi ban đầu.
- **Đối với mô hình Senior** (Hình 4b): Trái ngược với giả định thông thường rằng "GPA tích lũy càng cao thì kết quả càng tốt", biểu đồ cho thấy `HIST_AVG_GPA` và `HIST_MAX_GPA` lại nằm trong nhóm các đặc trưng **ít quan trọng nhất** (đáy biểu đồ). Thay vào đó, mô hình đặt trọng số cực lớn vào **phong độ ngắn hạn** (`LAST_PASS_RATIO`, `R2_PASS_RATE`) và **rủi ro gần đây** (`R3_SUM_FAIL`).
  → *Kết luận:* Mô hình đã học được tính chất "Momentum" (Đà học tập): Những gì sinh viên thể hiện trong các kỳ gần nhất có giá trị dự báo cao hơn nhiều so với lịch sử hào hùng hoặc bảng điểm đẹp trong quá khứ xa.

E. Chi tiết kết quả phân tích LIME

Để có cái nhìn đa chiều về độ tin cậy của mô hình, chúng tôi thực hiện phân tích LIME trên ba nhóm đối tượng đại diện cho các mức độ sai số dự báo khác nhau:

- **Best Case (Min Error):** Những trường hợp mô hình dự báo chính xác nhất.
- **Median Case (Avg Error):** Những trường hợp có sai số nằm ở mức trung bình.
- **Worst Case (Max Error):** Những trường hợp mô hình dự báo sai lệch nhiều nhất (outliers).

Hình 5 dưới đây trình bày chi tiết sáu kịch bản phân tích cho hai mô hình Fresher và Senior.



(a) *Fresher* - Best Case (b) *Fresher* - Median Case (c) *Fresher* - Worst Case

(d) *Senior* - Best Case (e) *Senior* - Median Case (f) *Senior* - Worst Case

Hình 5: Tổng hợp phân tích cục bộ LIME.