Phân tích dữ liệu sinh viên và nhận diện nguy cơ bỏ học

Nguyễn Thành Công Lợi

***Tóm tắt nội dung*—Trong nghiên cứu này, chúng tôi sử dụng bộ dữ liệu "Dropout Graduate Analysis" từ Kaggle, phản ánh quá trình học tập và tình trạng bỏ học hoặc tốt nghiệp của sinh viên đại học, nhằm khám phá các yếu tố ảnh hưởng đến khả năng duy trì việc học. Quy trình phân tích được triển khai theo hướng tiếp cận khoa học dữ liệu gồm tiền xử lý, trực quan hóa bằng biểu đồ và mô hình hóa. Để ta có những cái nhìn trực quan nhất về bộ dữ liệu cũng như nhưng mô hình học máy để giải quyết vấn đề**

I. Giới thiệu

Trong bối cảnh giáo dục đại học, tỷ lệ sinh viên bỏ học là vấn đề quan trọng ảnh hưởng tới chất lượng đào tạo và nguồn lực nhà trường. Phân tích dữ liệu có thể giúp phát hiện sớm những nhóm có nguy cơ, từ đó tổ chức hỗ trợ phù hợp. Bài báo này trình bày một chuỗi EDA và kể chuyện dữ liệu nhằm giải thích vì sao một phần sinh viên rời bỏ trước khi tốt nghiệp, đồng thời gợi ý các hành động dựa trên dữ liệu.

II. Tải và Khám phá Dữ liệu

A. Mô tả dữ liệu

Tập dữ liệu bao gồm 4.424 bản ghi với 35 biến đầu vào và một biến mục tiêu `Target`. Cụ thể:

Marital status – tình trạng hôn nhân, dạng phân loại.

Application mode – hình thức đăng ký, phân loại.

Application order – thứ tự nguyện vọng, giá trị số.

Course – ngành học đăng ký, phân loại.

Daytime/evening attendance – hệ học ban ngày hay buổi tối, nhị phân.

***Previous qualification** – loại bằng cấp trước đó, phân loại.

Previous qualification grade – điểm trung bình bằng cấp trước, giá trị số.

Nationality – quốc tịch, phân loại.

Mother's qualification – trình độ học vấn của mẹ, phân loại.

Father's qualification – trình độ học vấn của cha, phân loại.

Mother's occupation – nghề nghiệp của mẹ, phân loại.

Father's occupation – nghề nghiệp của cha, phân loại.

Displaced – tình trạng di dời, nhị phân.

Debtor – tình trạng nợ học phí, nhị phân.

Tuition fees up to date – học phí đã đóng đầy đủ, nhị phân.

Scholarship holder – có học bổng hay không, nhị phân.

Age at enrollment – tuổi khi nhập học, giá trị số.

Curricular units 1st sem credited – số tín chỉ kỳ 1 được công nhận.

Curricular units 1st sem enrolled – số môn đăng ký trong kỳ 1.

Curricular units 1st sem evaluated – số môn được đánh giá trong kỳ 1.

Curricular units 1st sem approved – số môn qua trong kỳ 1.

Curricular units 1st sem grade – điểm trung bình kỳ 1.

Curricular units 1st sem without evaluations – số môn không được đánh giá trong kỳ 1.

***Curricular units 2nd sem credited** – số tín chỉ kỳ 2 được công nhận.

Curricular units 2nd sem enrolled – số môn đăng ký trong kỳ 2.

Curricular units 2nd sem evaluated – số môn được đánh giá trong kỳ 2.

Curricular units 2nd sem approved – số môn qua trong kỳ 2.

Curricular units 2nd sem grade – điểm trung bình kỳ 2.

Curricular units 2nd sem without evaluations – số môn không được đánh giá trong kỳ 2.

Target – biến mục tiêu với ba giá trị: *Graduate* (tốt nghiệp), *Dropout* (bỏ học), *Enrolled*

III. Phân tích biểu đồ

Trong nghiên cứu này, chúng ta sử dụng bộ dữ liệu *Dropout Graduate Analysis* trên Kaggle, bao

gồm thông tin học tập, nhân khẩu học và các yếu tố liên quan đến tình trạng tốt nghiệp hay bỏ học của sinh viên. Việc trực quan hóa và phân tích dữ liệu giúp nhận diện các xu hướng tiềm ẩn, từ đó hỗ trợ việc dự đoán và xây dựng mô hình phù hợp.

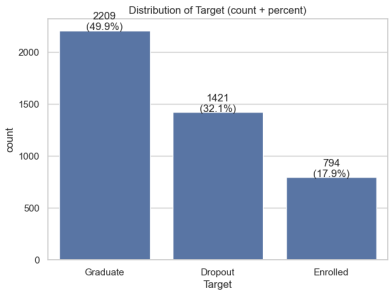
A. 1) Kiểm tra dữ liệu thiếu

("Trước khi trực quan và đánh giá, ta kiểm tra giá trị thiếu của dữ liệu. Qua đây ta cũng sẽ chuẩn bị trước những hướng xử lý khi bắt đầu làm")

```
RangeIndex: 4424 entries, 0 to 4423
Data columns (total 35 columns):
 #   Column                                          Non-Null Count  Dtype
---  ------                                          --------------  -----
 0   Marital status                                  4424 non-null   int64
 1   Application mode                                4424 non-null   int64
 2   Application order                               4424 non-null   int64
 3   Course                                          4424 non-null   int64
 4   Daytime/evening attendance                      4424 non-null   int64
 5   Previous qualification                          4424 non-null   int64
 6   Nacionality                                     4424 non-null   int64
 7   Mother's qualification                          4424 non-null   int64
 8   Father's qualification                          4424 non-null   int64
 9   Mother's occupation                             4424 non-null   int64
 10  Father's occupation                             4424 non-null   int64
 11  Displaced                                       4424 non-null   int64
 12  Educational special needs                       4424 non-null   int64
 13  Debtor                                          4424 non-null   int64
 14  Tuition fees up to date                         4424 non-null   int64
 15  Gender                                          4424 non-null   int64
 16  Scholarship holder                              4424 non-null   int64
 17  Age at enrollment                               4424 non-null   int64
 18  International                                   4424 non-null   int64
 19  Curricular units 1st sem (credited)             4424 non-null   int64
...
 33  GDP                                             4424 non-null   float64
 34  Target                                          4424 non-null   object
dtypes: float64(5), int64(29), object(1)
```

Hình 1: Kiểm tra dữ liệu thiếu trong tập dữ liệu.

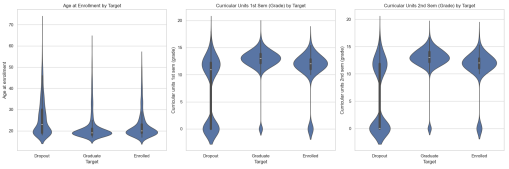
Hình 1 cho thấy dữ liệu có 4424 dòng và 35 cột, không cột nào bị thiếu. Đây là dấu hiệu đáng mừng cho một bộ dữ liệu đầy đủ. Tuy nhiên, dữ liệu đầy đủ không đồng nghĩa với việc đã phù hợp với mô hình, do đó vẫn cần kiểm tra và xử lý kỹ lưỡng hơn trước khi áp dụng vào phân tích hoặc xây dựng mô hình dự đoán. Ta sẽ phân tích những yếu tố khác để có thể xử lý dư liệu sao cho phù hợp với các mô hình sau này



Hình 2: Tỷ lệ phần trăm Target: Graduate, Dropout, Enrolled.

Biểu đồ này phản ánh một thực trạng khá phổ biến trong các hệ thống giáo dục: số lượng học viên tốt nghiệp (50%) thường cao hơn số học viên bỏ học (32.1%), nhưng số học viên bỏ học lại chiếm tỷ lệ đáng kể. Tỷ lệ học viên hiện tại đang học (17.9%) ít hơn so với các nhóm còn lại, có thể là do một phần lớn học viên đã hoàn thành chương trình học hoặc đã bỏ học trước khi hoàn tất.

Trong bối cảnh hiện nay, tỷ lệ bỏ học có thể xuất phát từ nhiều yếu tố như áp lực học tập, tài chính, và môi trường học tập không thuận lợi. Điều này cũng phản ánh nhu cầu cần cải thiện chất lượng giáo dục, hỗ trợ học viên tốt hơn để giảm tỷ lệ bỏ học và tăng tỷ lệ tốt nghiệp.

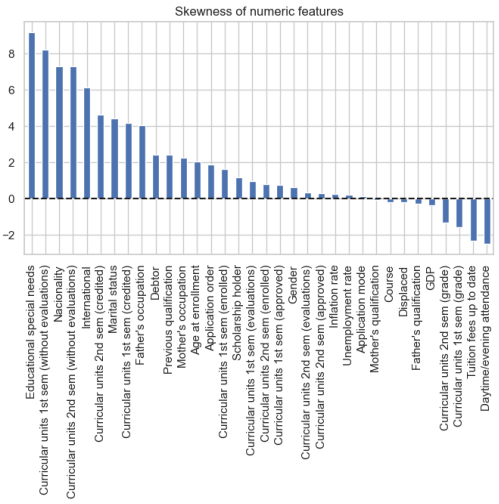


Hình 3: Mối quan hệ giữa tuổi khi nhập học, số tín chỉ tích lũy và trạng thái học tập.

Hình 3 cho thấy cái nhìn rõ ràng hơn về mối liện hệ giữa tuổi nhập học, số tín kỳ 1 và số tín kỳ 2 với 3 biến target, ta có thể thấy tuổi của nhóm bỏ học chủ yếu quanh khu vực từ 17 đến 25 tuổi, có những ngoại lai từ 40+ nhưng phân phối hẹp hơn, ở 2 trường còn lại sẽ thấy tuổi nhập học sẽ rộng hơn ở 18-25 tuổi và cũng có ngoại lai ở các tuổi nhập học lớn nhưng ko rộng bằng trường dropout. Có thể thấy tuổi nhập học là 1 đối tượng cần để ý khi những sinh viên nhập học muộn dễ có nguy cơ bỏ học hơn

Về Điểm trung bình học kỳ 1 theo trạng thái: Dropout: Có hai cực rõ: một nhóm đạt điểm gần 0 (không hoàn thành hoặc trượt toàn bộ), và một nhóm nhỏ đạt 10–13 nhưng hiếm khi cao hơn. Cho thấy điểm học kỳ 1 thấp là yếu tố cảnh báo nguy cơ bỏ học. Graduate và Enrolled: Phân phối tập trung quanh 12–14 điểm (trên thang 0–20), ít có điểm quá thấp. Điều này chứng tỏ thành công trong học kỳ 1 là tiền đề cho việc tiếp tục học hoặc tốt nghiệp.

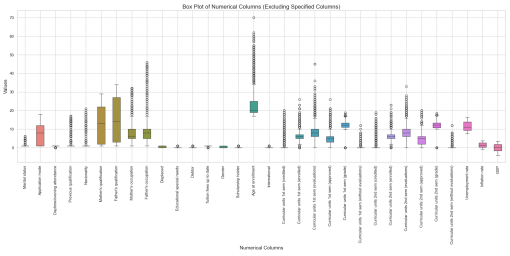
Cuối cùng Điểm trung bình học kỳ 2 theo trạng thái: Dropout: Giống học kỳ 1, nhiều sinh viên có điểm 0 hoặc cực thấp, phân phối trải dài nhưng thiên về thất bại. Graduate: Trung bình cao hơn, tập trung quanh 13–14, cho thấy kết quả học tập ổn định qua nhiều học kỳ. Enrolled: Giống nhóm Graduate, nhưng phân phối rộng hơn một chút, có cả sinh viên điểm thấp lẫn cao. Nhóm này còn đang trong quá trình học, nên kết quả chưa ổn định bằng Graduate.



Hình 4: Skewness các biến chính trong bộ dữ liệu.

Một cái nhìn tổng quan cho thấy sự mất cân bằng trong nhiều biến Một số biến có độ lệch (skewness) rất cao như: Educational special needs, Nationality, International, Curricular units without evaluations Chứng tỏ phần lớn giá trị tập trung vào một nhóm (ví dụ: đa số sinh viên không có nhu cầu đặc biệt, chủ yếu cùng một quốc tịch, hầu hết đều có đánh giá trong môn học), chỉ có một tỷ lệ nhỏ rơi vào nhóm khác. Điều này tạo ra phân phối không đối xứng, dễ gây ảnh hưởng đến mô hình nếu không xử lý, Các biến học tập (curricular units),các biến như without evaluations và credited có skew cao, cho thấy chỉ một số ít sinh viên có tình huống đặc biệt (bỏ qua đánh giá, hoặc được miễn/nhận tín chỉ). Trong khi đó, các biến như approved, enrolled, evaluations có skew thấp hơn, nghĩa là phân phối cân bằng hơn và phản ánh tình hình học tập chung. Điểm số (grades), curricular units 1st sem (grade) và 2nd sem (grade) có skew âm (-1.5 và -1.3). Điều này có nghĩa là phân phối điểm hơi lệch về phía điểm cao, tức phần lớn sinh viên đạt mức điểm khá trở lên, còn nhóm điểm thấp thì ít hơn. Thông tin nhân khẩu và gia đình Các biến như Father's occupation, Mother's occupation, Marital status có skew khá cao, cho thấy dữ liệu tập trung vào một số nhóm chính (ví dụ nhiều cha mẹ làm nghề phổ biến, đa số sinh viên độc thân). Các biến khác Một số biến có skew âm mạnh như Daytime/evening attendance và Tuition fees up to date Có thể do hầu hết sinh viên đóng học phí đầy đủ hoặc tham gia theo một hình thức học chủ yếu, còn hình thức/nhóm khác ít gặp.

Bộ dữ liệu này khá mất cân bằng ở nhiều biến, đặc biệt là các biến nhân khẩu học và một số đặc trưng học tập hiếm. Trong khi đó, điểm số và các biến về tiến trình học tập lại có phân phối tương đối hợp lý hơn. Việc tiền xử lý sẽ rất quan trọng trước khi đưa vào mô hình dự báo.



Hình 5: Boxplot cho các biến

Hình 5 cho thấy rõ hơn sự tồn tại của nhiều ngoại lệ ở các nhóm biến:

Biến nhân khẩu học – xã hội: *Age at enrollment* phân bố rộng (trung bình ∼20), nhiều ngoại lệ ở nhóm tuổi cao (30–40+), phản ánh người học muộn hoặc học lại. Các biến *Marital status, Application mode, Attendance, Gender, Debtor, Displaced, International, Scholarship holder* tập trung ở mức thấp (0–2), song vẫn có ngoại lệ ở *Application mode* và *Marital status*. *Mother's/Father's qualification và occupation* trải dài, nhiều ngoại lệ, thể hiện sự đa dạng nền tảng gia đình.

Biến học tập: *Enrolled, Evaluations, Approved, Grade, Credited* có phân bố rộng và nhiều ngoại lệ ở cả hai học kỳ. Các biến *without evaluations* nhiều giá trị 0 và vài giá trị cực cao, phản ánh sinh viên bỏ môn hoặc đăng ký nhưng không học. Điểm số (0–20) xuất hiện cực trị ở cả hai phía (0 và cao), phản ánh rõ sự phân loại trượt/đạt.

Biến kinh tế – xã hội vĩ mô: *GDP, Inflation, Unemployment* phân bố hẹp, ít ngoại lệ, do đặc thù dữ liệu vĩ mô. placeins

Biểu đồ 6 làm rõ đặc điểm phân phối:

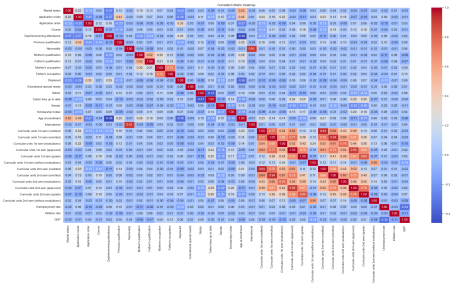
Biến nhân khẩu học – xã hội: *Marital status* lệch phải mạnh, đa số độc thân. *Application order* đa đỉnh, tập trung ở mức thấp. *Attendance* phân cực rõ, đa số học ban ngày. *Previous qualification* lệch phải, chủ yếu bằng phổ thông. *Qualification/Occupation của phụ huynh* đa đỉnh, lệch nhẹ đến vừa. *Debtor, Displaced, International, Scholarship holder, Special needs* lệch phải mạnh, đa số ở mức 0 (không có). *Gender* gần cân bằng. *Age at enrollment* lệch phải, chủ yếu 18–20 tuổi.

Biến học tập: Các biến *enrolled, credited, approved, grade* lệch phải nhẹ, có nhiều ngoại lệ.

Without evaluations lệch cực mạnh, phản ánh nhóm sinh viên đặc biệt (nguy cơ bỏ học). *Evaluations* tập trung ở mức trung bình. Điểm số học kỳ lệch trái, đa số sinh viên đạt điểm cao, ít sinh viên điểm thấp.

Biến kinh tế – xã hội vĩ mô: *Inflation, Unemployment, GDP* nhiều đỉnh nhưng hẹp, ít biến động.

Hình 6: Phân phối các biến số



Hình 7: Ma trận tương quan.

Ma Trận Tương quan cho thấy các biến liên quan đến học tập (enrolled, approved, evaluations, grade) có tương quan rất cao với nhau, phản ánh cùng một khía cạnh về kết quả học tập. Điểm trung bình học kỳ 1 và học kỳ 2 cũng có mối quan hệ chặt chẽ, cho thấy kết quả học tập thường duy trì ổn định qua các kỳ.

Nhóm biến tài chính như nợ học phí và học bổng có mối liên hệ vừa phải với kết quả học tập, trong khi các yếu tố nhân khẩu học và kinh tế vĩ mô hầu như không có ảnh hưởng đáng kể. Có thể xem xét loại bỏ khỏi mô hình dự đoán.

IV. XỬ LÝ DỮ LIỆU

Sau khi phân tích biểu đồ, ta nhận thấy dữ liệu tuy không thiếu giá trị (missing values) nhưng tồn tại một số vấn đề cần xử lý trước khi xây dựng mô hình đó chính là outlier, dữ liệ mất cân bằng ở một số biến hay độ lệch ở phân phối skewness ta cần xử lý tất cả các vấn đề đó trước khi tiến hành xây dựng mô hình dự đoán

A. Xử lý ngoại lệ

Các biến liên quan đến học tập và độ tuổi nhập học có nhiều giá trị ngoại lệ, ví dụ sinh viên đăng ký quá nhiều môn hoặc nhập học muộn. Thay vì loại bỏ, áp dụng RobustScaler để giảm ảnh hưởng.

B. Xử lý phân phối lệch

Nhiều biến có phân phối lệch phải, trong khi điểm số thường lệch trái. Để khắc phục, áp dụng Power Transformation để đưa phân phối về gần chuẩn. Với điểm số, giữ nguyên nhưng chuẩn hóa khi huấn luyện mô hình.

C. Xử lý mất cân bằng lớp

Biến mục tiêu có tỷ lệ Graduate khoảng 50%, Dropout 32% và Enrolled 18%. Khi gộp Graduate và Enrolled thành Success, dữ liệu vẫn mất cân bằng. Có thể dùng SMOTE để tạo thêm dữ liệu cho lớp thiểu số hoặc thiết lập trọng số lớp trong mô hình.

D. Mã hóa biến phân loại

Một số biến dạng số thực chất là phân loại, ví dụ tình trạng hôn nhân, ngành học, nghề nghiệp cha mẹ. Các biến này được chuyển sang dạng nhị phân bằng One-hot Encoding. Các biến nhị phân như Debtor, Scholarship holder, Gender giữ nguyên.

E. Giảm đa cộng tuyến

Các biến học tập có tương quan cao với nhau. Có thể loại bỏ biến dư thừa hoặc áp dụng PCA để giảm chiều dữ liệu .

F. Chuẩn hóa dữ liệu

một số biến đã được chuẩn hoá ngay từ đầu nên không cần chuẩn hoá lại

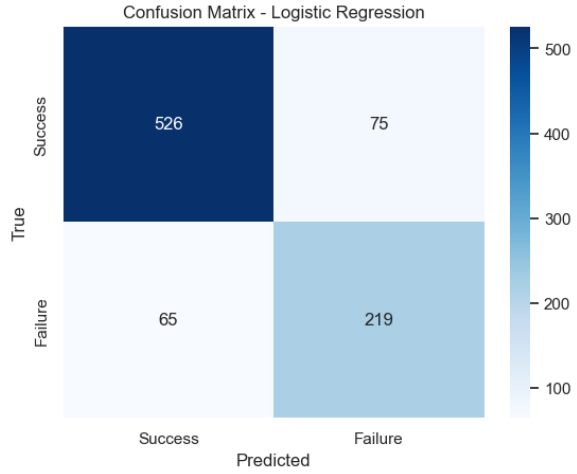
V. Kết quả mô hình

A. Ý nghĩa biến mục tiêu

Trong bộ dữ liệu, biến mục tiêu ban đầu có ba trạng thái: *Graduate* (sinh viên tốt nghiệp), *Enrolled* (sinh viên còn đang theo học), và *Dropout* (sinh viên bỏ học). Để đơn giản hóa bài toán, chúng tôi gộp *Graduate* và *Enrolled* thành một nhóm gọi là **Success (Thành công)**, và giữ *Dropout* làm nhóm **Failure (Thất bại)**. Như vậy:

- **Success**: sinh viên tiếp tục gắn bó với chương trình học (còn học hoặc đã tốt nghiệp).
- **Failure**: sinh viên rời bỏ chương trình học (bỏ học).

B. Đánh giá mô hình Decision Tree

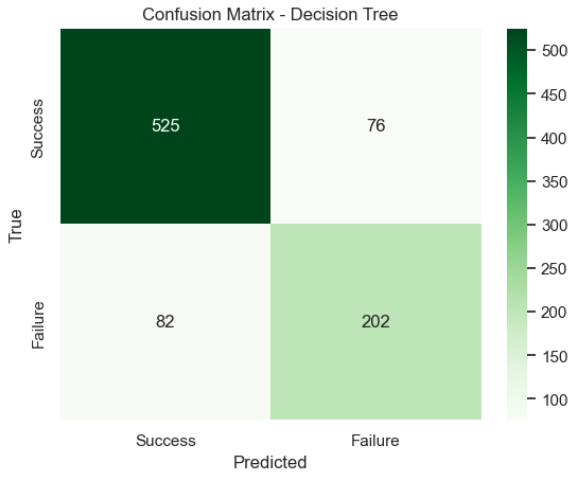


Hình 8: Confusion Matrix - Decision Tree

Hình 8 thể hiện ma trận nhầm lẫn của Decision Tree. Mô hình đạt độ chính xác tổng thể (*Accuracy*) khoảng 82%. Kết quả cho thấy lớp *Success* được dự đoán khá tốt (Recall = 0.87), trong khi lớp *Failure* có Recall thấp hơn (0.71), tức là vẫn bỏ sót một số sinh viên có nguy cơ bỏ học. Ưu điểm chính của Decision Tree là dễ diễn giải, có thể xác định rõ yếu tố nào dẫn đến kết quả dự đoán.

C. Đánh giá mô hình Logistic Regression

Hình 9 cho thấy ma trận nhầm lẫn của Logistic Regression. Mô hình đạt độ chính xác tổng thể khoảng 84%, cao hơn Decision Tree. Precision và Recall của cả hai lớp đều cân bằng hơn (Failure: Recall = 0.77, Success: Recall = 0.88). Điều này cho thấy Logistic Regression có khả năng phân biệt hai nhóm tốt hơn, đặc biệt là trong việc nhận diện sinh viên bỏ học.



Hình 9: Confusion Matrix - Logistic Regression

D. So sánh và nhận xét

- **Decision Tree** cho phép trực quan hóa và diễn giải mô hình, nhưng hiệu suất thấp hơn một chút.
- **Logistic Regression** có độ chính xác cao hơn, cân bằng giữa hai lớp, phù hợp khi ưu tiên hiệu quả dự đoán.

VI. Kết luận

Báo cáo đã thực hiện phân tích, tiền xử lý và xây dựng mô hình dự đoán khả năng duy trì học tập của sinh viên.

Kết quả cho thấy dữ liệu không có giá trị thiếu nhưng tồn tại nhiều ngoại lệ, phân phối lệch và mất cân bằng lớp, đã được xử lý bằng RobustScaler, Power Transformation, SMOTE và One-hot Encoding.

Phân tích thăm dò cho thấy các yếu tố học tập (điểm số, số môn học) và tài chính (học bổng, nợ học phí) có ảnh hưởng rõ rệt đến khả năng bỏ học, trong khi yếu tố nhân khẩu học và kinh tế vĩ mô ít tác động hơn.

Hai mô hình Logistic Regression và Decision Tree đã được triển khai. Logistic Regression đạt độ chính xác cao và cân bằng giữa hai lớp, trong khi Decision Tree dễ diễn giải và giúp hiểu rõ nguyên nhân. Nhìn chung, các mô hình có khả năng nhận diện sinh viên có nguy cơ bỏ học, qua đó hỗ trợ nhà trường đưa ra biện pháp can thiệp sớm.

Trong tương lai, có thể mở rộng bằng cách áp dụng các mô hình mạnh hơn như Random Forest hoặc XGBoost để nâng cao hiệu quả dự đoán.