BỘ KHOA HỌC VÀ CÔNG NGHỆ
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



HOÀNG NGUYỄN LONG THẢO

XÂY DỰNG CÁC MÔ HÌNH HỌC SÂU DỰ ĐOÁN TỈ LỆ SINH VIÊN BỎ HỌC

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

THÀNH PHỐ HÀ NỘI – 2026

BỘ KHOA HỌC VÀ CÔNG NGHỆ
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

HOÀNG NGUYỄN LONG THẢO

XÂY DỰNG CÁC MÔ HÌNH HỌC SÂU DỰ ĐOÁN TỈ LỆ SINH VIÊN BỎ HỌC

Chuyên ngành: Khoa học máy tính

Mã số: 8.48.01.01

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

NGƯỜI HƯỚNG DẪN KHOA HỌC: TS ĐỖ TIẾN DŨNG

THÀNH PHỐ HÀ NỘI – 2026

LỜI CAM ĐOAN

Tôi cam đoan rằng đề án tốt nghiệp thạc sĩ **"XÂY DỰNG CÁC MÔ HÌNH HỌC SÂU DỰ ĐOÁN TỈ LỆ SINH VIÊN BỎ HỌC"** là công trình nghiên cứu của chính tôi.

Tôi xin cam đoan rằng tất cả các số liệu, kết quả, cũng như các thông tin được trình bày trong đề án này đều hoàn toàn trung thực và chính xác, không có bất kỳ sự can thiệp hay thay đổi nào. Các số liệu này là kết quả của quá trình nghiên cứu, khảo sát và phân tích độc lập của tôi, và chưa từng được công bố trong bất kỳ công trình khoa học hay nghiên cứu nào trước đây.

Tôi xin cam kết rằng trong quá trình thực hiện nghiên cứu, tôi đã sử dụng các nguồn tài liệu tham khảo một cách chính xác và đầy đủ, không sao chép hay sử dụng bất kỳ kết quả nghiên cứu nào của người khác mà không ghi rõ nguồn gốc, đảm bảo tính minh bạch trong nghiên cứu khoa học.

Tp.Hà Nội, ngày 29 tháng 7 năm 2026

Học viên thực hiện đề án



Hoàng Nguyễn Long Thảo

LỜI CẢM ƠN

Trước hết, tôi xin bày tỏ lòng biết ơn sâu sắc và trân trọng nhất tới Thầy TS Đỗ Tiến Dũng, người đã luôn tận tâm và hỗ trợ em trong suốt quá trình thực hiện đề án tốt nghiệp. Tôi rất biết ơn những lời khuyên và sự chỉ bảo quý báu của Thầy, chính những chỉ dẫn đó đã giúp tôi tích lũy thêm kiến thức và kinh nghiệm quý giá trong quá trình nghiên cứu.

Tôi xin gửi lời cảm ơn chân thành đến Khoa Đào tạo sau Đại học – Học viện Công nghệ Bưu chính Viễn thông đã tạo mọi điều kiện để giúp tôi hoàn thành đề án.

Cuối cùng, tôi xin gửi lời cảm ơn chân thành đến gia đình, bạn bè và đồng nghiệp trong cơ quan, những người đã luôn động viên và hỗ trợ tôi trong suốt quá trình học tập và thực hiện đề án.

Tôi xin chân thành cảm ơn!

Tp. Hà Nội, ngày 23 tháng 1 năm 2026

Học viên thực hiện đề án



Hoàng Nguyễn Long Thảo

MỤC LỤC

LỜI CAM ĐOAN i
LỜI CẢM ƠN ii
MỤC LỤC iii
DANH MỤC CÁC CHỮ VIẾT TẮT v
DANH MỤC CÁC BẢNG vi
DANH MỤC HÌNH ẢNH vii
I. MỞ ĐẦU 1
1.1. Tổng quan đề tài nghiên cứu 1
1.1.1. Tình hình sinh viên bỏ học trong các trường đại học 1
1.1.2. Lý do lựa chọn đề tài 2
1.2. Mục tiêu và phạm vi nghiên cứu 3
1.2.1. Mục tiêu nghiên cứu 4
1.2.2. Phạm vi nghiên cứu 5
1.3. Phương pháp nghiên cứu 6
1.3.1. Thu thập và xử lý dữ liệu 6
1.3.2. Các phương pháp phân tích 6
II. NỘI DUNG 8
CHƯƠNG 1: TỔNG QUAN LÝ THUYẾT ĐỀ TÀI XÂY DỰNG CÁC MÔ HÌNH HỌC SÂU DỰ ĐOÁN TỈ LỆ SINH VIÊN BỎ HỌC 9
1.1. Tỷ lệ bỏ học và nguyên nhân 9
1.2. Các mô hình học sâu phục vụ bài toàn 14
1.3. Dữ liệu và công cụ 22
1.4. Các nghiên cứu liên quan 27
1.5. Kết luận chương 28
CHƯƠNG 2: XÂY DỰNG CÁC MÔ HÌNH HỌC SÂU 30

2.1. Tiền xử lý dữ liệu 30
2.2. Xây dựng mô hình RNN 38
2.3. Xây dựng mô hình LSTM 42
2.4. Xây dựng mô hình Hybrid 46
2.5. Kết luận chương 59
CHƯƠNG 3: THỬ NGHIỆM VÀ ĐÁNH GIÁ 61
3.1. Trình bày bài toán 61
3.2. Một số bài báo nghiên cứu trước đây 63
3.3. Đánh giá kết quả thử nghiệm các mô hình 67
3.4. Phân tích các đặc trưng quan trọng của dữ liệu 71
3.5. Thảo luận 73
3.6. Kết luận chương 77
CHƯƠNG 4: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 79
4.1. Kết luận chính 79
4.2. Đóng góp khoa học 81
4.3. Hướng nghiên cứu tương lai 82
III. KẾT LUẬN 86
IV. TÀI LIỆU THAM KHẢO 88

DANH MỤC CÁC CHỮ VIẾT TẮT

Viết tắt	Tiếng anh	Tiếng việt
AI	Artificial Intelligence	Trí tuệ nhân tạo
ML	Machine learning	Học máy
DL	Deep learning	Học sâu
RNN	Recurrent neural network	Mạng nơ ron hồi tiếp
LSTM	Long Short-Term Memory	Mạng bộ nhớ dài-ngắn hạn
LMS	Learning Management System	Hệ thống quản lý học tập

DANH MỤC CÁC BẢNG

Table 3.1. Hiệu suất các mô hình.....68

Table 3.2. Các đặc trưng có mức độ ảnh hưởng cao đến kết quả dự đoán.....71

DANH MỤC HÌNH ẢNH

Figure 1.1. Sự khác biệt giữa ML và DL 14
Figure 1.2. Mô hình RNN 16
Figure 1.3. Mô hình LSTM 18
Figure 2.1. Sự phân bố của lớp mục tiêu Graduate và Dropout 31
Figure 2.2. Tỷ lệ thất nghiệp 32
Figure 2.3. Sự ảnh hưởng của StandardScaler 35
Figure 2.4. Xây dựng mô hình RNN 40
Figure 2.5. Xây dựng mô hình LSTM 44
Figure 2.6. Xây dựng mô hình Hybrid 50

I. MỞ ĐẦU

1.1. Tổng quan đề tài nghiên cứu

1.1.1. Tình hình sinh viên bỏ học trong các trường đại học

Hiện tượng sinh viên bỏ học tại các trường đại học đang trở thành một vấn đề đáng báo động không chỉ ở Việt Nam mà còn trên phạm vi toàn cầu. Ở Việt Nam, số liệu từ các trường đại học như Đại học Luật TP.HCM, Đại học Nguyễn Tất Thành hay Đại học Bách khoa Hà Nội cho thấy hàng năm có hàng nghìn sinh viên bị cảnh báo học vụ hoặc buộc thôi học. Nguyên nhân thường gặp bao gồm không đạt yêu cầu học tập, điểm trung bình học kỳ quá thấp (dưới 1,0 trên thang điểm 4), nợ tín chỉ kéo dài, hoặc không đảm bảo thời gian tối đa hoàn thành chương trình học. Đáng chú ý, tại một số trường như Đại học Nguyễn Tất Thành, đã có tới 3.000 sinh viên bị cảnh báo học vụ lần 1 chỉ trong một học kỳ, trong khi Đại học Bách khoa TP.HCM chỉ đạt tỷ lệ khoảng 70% sinh viên tốt nghiệp đúng hạn. Trên thế giới, các quốc gia phát triển cũng đối mặt với hiện tượng tương tự, khi tỷ lệ bỏ học dao động từ 20% đến 30%, tùy thuộc vào môi trường giáo dục và mức độ hỗ trợ dành cho sinh viên. Tình trạng này phản ánh không chỉ những thách thức về hệ thống giáo dục mà còn là vấn đề chung của xã hội, đòi hỏi sự quan tâm và cải cách mạnh mẽ từ nhiều phía.

Tỷ lệ bỏ học cao có tác động sâu rộng đến chất lượng giáo dục và công tác quản lý trong các trường đại học. Đầu tiên, việc nhiều sinh viên rời bỏ giảng đường không chỉ ảnh hưởng trực tiếp đến uy tín và danh tiếng của nhà trường mà còn làm giảm chất lượng đầu ra của nền giáo dục. Các trường đại học phải đối mặt với áp lực lớn trong việc giải quyết tình trạng này, từ việc xây dựng chương trình hỗ trợ học tập, cải thiện môi trường giảng dạy, đến triển khai các biện pháp giám sát và tư vấn cho sinh viên. Bên cạnh đó, sự gia tăng tỷ lệ bỏ học đồng nghĩa với việc các nguồn lực giáo dục như cơ sở vật chất, đội ngũ giảng viên và kinh phí đào tạo bị lãng phí, làm giảm hiệu quả đầu tư của nhà trường. Trên quy mô quốc gia, tỷ lệ bỏ học cao còn dẫn đến tổn thất lớn về nguồn nhân lực, khi nhiều sinh viên không thể hoàn thành chương trình đào tạo để trở thành những người lao động có trình độ cao. Điều này không chỉ ảnh hưởng đến sự phát triển kinh tế mà còn làm suy yếu vị thế cạnh tranh của quốc gia trong bối cảnh toàn cầu hóa.

Nguyên nhân dẫn đến hiện tượng sinh viên bỏ học có thể được phân tích từ nhiều góc độ, bao gồm yếu tố văn hóa, xã hội và kinh tế. Về mặt văn hóa, nhiều sinh viên thiếu động lực học tập, không xác định rõ mục tiêu nghề nghiệp, hoặc cảm thấy chương trình học không phù hợp với nhu cầu cá nhân. Mặt khác, áp lực từ gia đình trong việc phải đạt kết quả học tập cao, hoặc từ bạn bè cùng trang lứa trong việc theo đuổi những giá trị nhất thời, cũng có thể khiến sinh viên cảm thấy bế tắc và mất phương hướng. Về xã hội, các yếu tố như môi trường học tập thiếu hỗ trợ, sự cô lập trong môi trường đại học, hoặc xu hướng tìm kiếm các công việc tạm thời thay vì đầu tư vào học vấn dài hạn, đều góp phần gia tăng nguy cơ bỏ học. Đặc biệt, yếu tố kinh tế đóng vai trò quyết định trong nhiều trường hợp. Học phí tăng cao, chi phí sinh hoạt ngày càng đắt đỏ và thiếu sự hỗ trợ tài chính kịp thời khiến nhiều sinh viên, đặc biệt là những người đến từ gia đình có hoàn cảnh khó khăn, không thể tiếp tục việc học. Những yếu tố này không chỉ đòi hỏi các trường đại học phải linh hoạt hơn trong chính sách hỗ trợ mà còn cần sự phối hợp từ phía chính phủ và xã hội để tạo điều kiện học tập tốt nhất cho sinh viên.

Những số liệu và phân tích trên đây nhấn mạnh tính nghiêm trọng và phức tạp của vấn đề sinh viên bỏ học. Đây không chỉ là bài toán của riêng các trường đại học mà còn là một thách thức lớn đối với hệ thống giáo dục và toàn xã hội. Việc nghiên cứu và triển khai các giải pháp toàn diện là điều cần thiết để giảm thiểu tình trạng này, góp phần nâng cao chất lượng giáo dục và xây dựng một lực lượng lao động tri thức bền vững cho tương lai.

1.1.2. Lý do lựa chọn đề tài

Hiện tượng sinh viên bỏ học trong các trường đại học không chỉ là một vấn đề về số lượng mà còn mang theo những hệ lụy sâu rộng cho cá nhân sinh viên, nhà trường, và cả hệ thống giáo dục. Trong bối cảnh đó, việc dự đoán tỷ lệ sinh viên bỏ học đóng vai trò vô cùng quan trọng. Nếu có thể dự đoán chính xác nguy cơ bỏ học của sinh viên, các trường đại học sẽ có cơ hội can thiệp sớm, ngăn chặn tình trạng này và giảm thiểu những tác động tiêu cực. Dự đoán tỷ lệ bỏ học không chỉ mang ý nghĩa thống kê mà còn là một công cụ chiến lược, giúp nhà quản lý giáo dục đưa ra các quyết định sáng suốt, tối ưu hóa nguồn lực, và cải thiện chất lượng giảng dạy. Hơn nữa, trong thời đại dữ liệu số, các mô hình dự đoán được phát triển dựa trên trí tuệ nhân tạo (AI) và học máy (machine learning) mang lại khả năng phân tích chính

xác, nhanh chóng, và hiệu quả, tạo ra sự khác biệt rõ rệt so với các phương pháp truyền thống.

Mục tiêu của đề tài này là xây dựng một mô hình dự đoán tỷ lệ sinh viên bỏ học nhằm giảm thiểu tình trạng này trong các trường đại học. Đề tài không chỉ dừng lại ở việc tìm hiểu các yếu tố tác động đến quyết định bỏ học của sinh viên mà còn hướng tới việc phát triển một hệ thống có khả năng nhận diện các nhóm sinh viên có nguy cơ cao ngay từ những giai đoạn đầu của quá trình học tập. Điều này đặc biệt quan trọng bởi lẽ việc can thiệp sớm có thể giúp sinh viên vượt qua khó khăn, khắc phục các yếu điểm và tiếp tục theo đuổi con đường học vấn. Mô hình dự đoán cũng giúp các trường đại học cá nhân hóa phương pháp hỗ trợ, từ đó tạo ra các chính sách giáo dục linh hoạt và hiệu quả hơn. Không chỉ giúp giảm tỷ lệ bỏ học, đề tài còn nhằm nâng cao trải nghiệm học tập và gắn kết sinh viên với môi trường đại học.

Lợi ích của việc ứng dụng mô hình dự đoán trong giáo dục là rất lớn, không chỉ đối với nhà trường mà còn đối với toàn bộ hệ thống giáo dục. Trước hết, mô hình này giúp các trường đại học cải thiện chương trình đào tạo bằng cách điều chỉnh nội dung và phương pháp giảng dạy sao cho phù hợp hơn với nhu cầu của sinh viên. Đồng thời, các trường có thể thiết kế các chương trình hỗ trợ học tập, như tư vấn tâm lý, học thuật, hoặc hỗ trợ tài chính một cách nhắm đúng đối tượng. Thứ hai, việc can thiệp sớm nhờ mô hình dự đoán giúp nhà trường không chỉ giảm số lượng sinh viên bỏ học mà còn xây dựng một môi trường học tập tích cực hơn, nơi sinh viên cảm thấy được hỗ trợ và động viên. Cuối cùng, việc giảm tỷ lệ sinh viên bỏ học còn mang lại lợi ích về mặt kinh tế và xã hội, đảm bảo sự bền vững trong đầu tư giáo dục và cung cấp một lực lượng lao động chất lượng cao cho xã hội.

Với tầm quan trọng của vấn đề và những lợi ích thực tiễn mà mô hình dự đoán mang lại, việc nghiên cứu và triển khai đề tài này là cần thiết và có ý nghĩa cả về mặt học thuật lẫn ứng dụng. Đề tài không chỉ góp phần giải quyết một trong những thách thức lớn nhất của giáo dục đại học mà còn mở ra những hướng đi mới trong việc áp dụng công nghệ hiện đại vào quản lý và cải thiện chất lượng giáo dục. Đây chính là lý do cốt lõi để lựa chọn đề tài này cho nghiên cứu, với kỳ vọng rằng kết quả của nghiên cứu sẽ mang lại những giá trị thiết thực và bền vững cho ngành giáo dục.

1.2. Mục tiêu và phạm vi nghiên cứu

1.2.1. Mục tiêu nghiên cứu

Nghiên cứu này tập trung vào việc xây dựng và so sánh các mô hình học sâu để dự đoán tỷ lệ sinh viên bỏ học trong các trường đại học, với mục tiêu tìm ra phương pháp hiệu quả nhất nhằm giảm thiểu tỷ lệ bỏ học và nâng cao chất lượng giáo dục. Trọng tâm của nghiên cứu là ứng dụng các công nghệ hiện đại, bao gồm trí tuệ nhân tạo và học sâu, để phát triển các mô hình dự đoán chính xác và phân tích các yếu tố ảnh hưởng chính đến quyết định bỏ học của sinh viên.

Cụ thể, nghiên cứu tập trung vào việc thiết kế, triển khai và so sánh các mô hình học sâu: RNN (Recurrent Neural Network), LSTM (Long Short-Term Memory), Logistic Regression, đặc biệt là mô hình Hybrid – được phát triển dựa trên việc kết hợp hai mô hình RNN và LSTM để đánh giá hiệu quả trong bài toán dự đoán dữ liệu chuỗi thời gian. Các mô hình này sẽ được tối ưu hóa dựa trên các tham số đặc trưng, đảm bảo đạt hiệu suất cao nhất khi áp dụng vào dữ liệu thực tế. Quá trình đánh giá được thực hiện dựa trên các thước đo toàn diện như độ chính xác (accuracy), độ nhạy (recall), F1-score và thời gian xử lý, qua đó xác định mô hình không chỉ có hiệu quả cao mà còn có tính khả thi trong thực tiễn.

Một phần quan trọng khác của nghiên cứu là việc phân tích các yếu tố ảnh hưởng đến khả năng bỏ học của sinh viên. Các yếu tố này bao gồm kết quả học tập (GPA), số lượng tín chỉ nợ, tình trạng tài chính, hành vi học tập, và mức độ tương tác với môi trường đại học. Thông qua việc khai thác và phân tích dữ liệu, nghiên cứu sẽ xác định những đặc trưng quan trọng nhất, từ đó không chỉ tối ưu hóa dữ liệu đầu vào mà còn cung cấp thông tin giá trị để hỗ trợ các trường đại học thiết kế các chính sách can thiệp kịp thời và hiệu quả.

Ngoài ra, nghiên cứu còn chú trọng vào việc đánh giá và tối ưu hóa mô hình thông qua các kỹ thuật như lựa chọn đặc trưng (feature selection), điều chỉnh tham số (hyperparameter tuning), và thử nghiệm trên nhiều tập dữ liệu khác nhau. Những kỹ thuật này không chỉ nâng cao khả năng tổng quát hóa của mô hình mà còn đảm bảo độ tin cậy khi áp dụng vào các bối cảnh giáo dục khác nhau.

Mục tiêu tổng quát của nghiên cứu là phát triển một hệ thống dự đoán tiên tiến, vừa có khả năng ứng dụng thực tiễn trong quản lý giáo dục, vừa đóng góp ý nghĩa vào lĩnh vực nghiên cứu trí tuệ nhân tạo. Kết quả nghiên cứu không chỉ giúp so sánh hiệu quả của các mô hình học sâu mà còn mở ra những giải pháp phù hợp với đặc thù

từng cơ sở giáo dục, qua đó góp phần xây dựng một nền tảng giáo dục đại học hiện đại, bền vững và toàn diện hơn.

1.2.2. Phạm vi nghiên cứu

Nghiên cứu này được triển khai với phạm vi cụ thể, tập trung vào việc phân tích và phát triển các mô hình học sâu nhằm dự đoán tỷ lệ sinh viên bỏ học trong các trường đại học. Phạm vi nghiên cứu bao gồm việc xác định các yếu tố ảnh hưởng chính, xây dựng và thử nghiệm ba mô hình học sâu, RNN (Recurrent Neural Network), LSTM (Long Short-Term Memory) và mô hình Hybrid, đồng thời đánh giá hiệu quả của các mô hình trên dữ liệu thực tế từ một trường đại học tại Việt Nam. Nghiên cứu không chỉ dừng lại ở việc xây dựng mô hình mà còn tập trung vào việc tối ưu hóa, đánh giá hiệu suất, và phân tích ý nghĩa thực tiễn của các kết quả, nhằm cung cấp các giải pháp khả thi và giá trị cho hệ thống giáo dục đại học.

Phạm vi đầu tiên của nghiên cứu là xác định các yếu tố quan trọng ảnh hưởng đến quyết định bỏ học của sinh viên. Các yếu tố này bao gồm kết quả học tập (GPA), số lượng tín chỉ nợ, tình trạng tài chính, hành vi học tập, mức độ tương tác với môi trường đại học, và các yếu tố bên ngoài như áp lực gia đình hoặc môi trường xã hội. Thông qua việc thu thập dữ liệu từ hệ thống quản lý của nhà trường và các khảo sát liên quan, nghiên cứu sẽ xây dựng bộ dữ liệu chi tiết và đầy đủ, làm cơ sở cho việc phân tích và tối ưu hóa các mô hình dự đoán.

Phạm vi tiếp theo tập trung vào việc áp dụng và so sánh các mô hình học sâu, RNN, LSTM và Hybrid, trên dữ liệu thực tế của sinh viên. Các mô hình này được lựa chọn do khả năng xử lý dữ liệu chuỗi thời gian, phù hợp với đặc điểm của bài toán dự đoán tỷ lệ bỏ học. Dữ liệu sử dụng bao gồm thông tin cá nhân, hồ sơ học tập, và các yếu tố liên quan đến hành vi học tập của sinh viên. Các mô hình sẽ được tối ưu hóa thông qua các kỹ thuật hiện đại như điều chỉnh tham số (hyperparameter tuning) và lựa chọn đặc trưng (feature selection) để đảm bảo đạt hiệu suất cao nhất. Hiệu quả của các mô hình sẽ được đánh giá thông qua các thước đo như độ chính xác (accuracy), độ nhạy (recall), F1-score, và thời gian xử lý.

Ngoài ra, nghiên cứu còn kiểm tra tính tổng quát hóa của các mô hình bằng cách thử nghiệm trên các tập dữ liệu khác nhau, nhằm đánh giá khả năng áp dụng trong các bối cảnh đa dạng của hệ thống giáo dục đại học. Phạm vi này không chỉ giúp kiểm chứng khả năng thực tiễn của các mô hình mà còn mở ra hướng phát triển một giải

pháp linh hoạt và bền vững cho việc dự đoán và giảm thiểu tỷ lệ bỏ học trong hệ thống giáo dục.

Phạm vi nghiên cứu được thiết kế để đảm bảo kết nối chặt chẽ với các mục tiêu đề ra, vừa đi sâu vào việc phân tích dữ liệu và tối ưu hóa mô hình, vừa đảm bảo tính ứng dụng cao. Qua việc tập trung vào các mô hình học sâu và phân tích dữ liệu thực tế, nghiên cứu không chỉ mang lại các giải pháp hiệu quả để giảm tỷ lệ bỏ học mà còn cung cấp một nền tảng có thể mở rộng, góp phần cải thiện chất lượng và hiệu quả quản lý giáo dục trong dài hạn.

1.3. Phương pháp nghiên cứu

1.3.1. Thu thập và xử lý dữ liệu

Quá trình thu thập dữ liệu trong nghiên cứu được thực hiện thông qua việc sử dụng các bảng câu hỏi khảo sát, dữ liệu học tập, và thông tin cá nhân của sinh viên từ hệ thống quản lý của trường đại học. Các bảng câu hỏi được thiết kế để thu thập thông tin về hoàn cảnh kinh tế, động lực học tập, và các yếu tố xã hội ảnh hưởng đến quyết định học tập của sinh viên. Dữ liệu học tập bao gồm điểm trung bình học kỳ (GPA), số lượng tín chỉ đã hoàn thành, số tín chỉ nợ, và các cảnh báo học vụ, trong khi thông tin cá nhân liên quan đến tuổi, giới tính, hoặc tình trạng gia đình. Sau khi thu thập, dữ liệu sẽ trải qua quá trình tiền xử lý gồm làm sạch dữ liệu để loại bỏ các giá trị không hợp lệ, xử lý giá trị thiếu thông qua các kỹ thuật như thay thế bằng giá trị trung bình hoặc sử dụng mô hình dự đoán, và chuẩn hóa các đặc trưng để đảm bảo tính đồng nhất trong quá trình phân tích.

1.3.2. Các phương pháp phân tích

Trong nghiên cứu này, các phương pháp được lựa chọn nhằm đảm bảo tính hiệu quả và chính xác trong việc dự đoán tỷ lệ sinh viên bỏ học, đồng thời phân tích các yếu tố ảnh hưởng quan trọng đến quyết định này. Trọng tâm của nghiên cứu là các mô hình học sâu tiên tiến, bao gồm RNN (Recurrent Neural Network), LSTM (Long Short-Term Memory) và đặc biệt là mô hình Hybrid , được triển khai để xử lý dữ liệu chuỗi thời gian và nắm bắt các mối quan hệ theo trình tự trong hành vi học tập của sinh viên.

Mạng RNN được lựa chọn nhờ khả năng xử lý dữ liệu có tính phụ thuộc theo thời gian, trong khi LSTM, với cơ chế bộ nhớ ngắn hạn và dài hạn, khắc phục hạn

chế về vấn đề mất mát thông tin trong các chuỗi dài của RNN, qua đó mang lại hiệu quả vượt trội trong bài toán dự đoán dữ liệu chuỗi. Mô hình Hybrid phân tách dữ liệu đầu vào thành hai nhóm: đặc trưng số và đặc trưng phân loại. Các đặc trưng số được xử lý thông qua mạng nơ-ron sâu (DNN), trong khi các đặc trưng phân loại được mã hóa bằng embedding và tiếp tục được học thông qua DNN, sau đó các đặc trưng này được hợp nhất để đưa ra dự đoán cuối cùng. Những mô hình này không chỉ hỗ trợ việc nhận diện các xu hướng bỏ học mà còn giúp phát hiện các mẫu ẩn trong hành vi học tập.

Để tối ưu hóa và tăng độ chính xác của mô hình, nghiên cứu áp dụng các kỹ thuật như lựa chọn đặc trưng (feature selection) để xác định những yếu tố quan trọng nhất trong dữ liệu, và điều chỉnh tham số (hyperparameter tuning) thông qua Grid Search nhằm tìm ra cấu hình tốt nhất cho từng mô hình. Các kỹ thuật này giúp giảm độ phức tạp của mô hình, cải thiện hiệu suất và đảm bảo tính tổng quát hóa trên nhiều bộ dữ liệu khác nhau.

Ngoài ra, để đảm bảo tính khách quan và toàn diện, nghiên cứu cũng sử dụng các phương pháp đánh giá hiệu suất mô hình, bao gồm các thước đo như độ chính xác (accuracy), độ nhạy (recall), F1-score và thời gian xử lý. Những tiêu chí này giúp so sánh hiệu quả của từng mô hình, từ đó xác định phương pháp phù hợp nhất cho bài toán.

Sự kết hợp của các phương pháp trên không chỉ đảm bảo khả năng dự đoán chính xác mà còn cung cấp thông tin giá trị về các yếu tố ảnh hưởng đến tỷ lệ bỏ học. Điều này tạo nền tảng cho việc ứng dụng mô hình vào thực tiễn quản lý giáo dục, hỗ trợ các trường đại học xây dựng các chiến lược can thiệp kịp thời và hiệu quả.

II. NỘI DUNG

Với mục đích, đối tượng, phạm vi và phương pháp nghiên cứu đã được xác định, đề án này được thiết kế thành 4 chương, đi từ lý thuyết đến thực tiễn cụ thể như sau:

Chương 1: Tổng quan lý thuyết

Chương 2: Xây dựng các mô hình học sâu

Chương 3: Thử nghiệm và đánh giá

Chương 4: Kết luận và hướng phát triển

CHƯƠNG 1: TỔNG QUAN LÝ THUYẾT ĐỀ TÀI XÂY DỰNG CÁC MÔ HÌNH HỌC SÂU DỰ ĐOÁN TỈ LỆ SINH VIÊN BỎ HỌC

- Giới thiệu chương

Trong lĩnh vực dự đoán dữ liệu, việc lựa chọn mô hình phù hợp đóng vai trò then chốt trong việc đạt được hiệu quả cao và khả năng áp dụng thực tế. Chương này tập trung nghiên cứu bài toán dự đoán tỷ lệ bỏ học, một vấn đề không chỉ ảnh hưởng đến chất lượng giáo dục mà còn để lại hệ lụy sâu sắc đối với xã hội. Trong bối cảnh giáo dục hiện đại, việc sớm nhận diện nguy cơ bỏ học và đưa ra các biện pháp can thiệp kịp thời là điều cần thiết để nâng cao hiệu quả giảng dạy và hỗ trợ học sinh vượt qua khó khăn. Để tiếp cận vấn đề, chương sẽ phân tích tỷ lệ bỏ học và các nguyên nhân chủ yếu nhằm cung cấp cái nhìn tổng quan về hiện trạng. Bên cạnh đó, các mô hình học sâu như RNN và LSTM, với khả năng vượt trội trong xử lý dữ liệu chuỗi thời gian, sẽ được giới thiệu như những giải pháp kỹ thuật hứa hẹn cho bài toán này. Song song, chương sẽ trình bày các công cụ và thư viện phổ biến như TensorFlow, PyTorch và Scikit-learn, đây là những công cụ hỗ trợ quan trọng trong việc triển khai và đánh giá các mô hình học máy. Cuối cùng, chương sẽ tổng hợp các nghiên cứu liên quan, đặt nền móng cho việc phát triển các giải pháp mới nhằm dự đoán và giảm thiểu tình trạng bỏ học.

1.1. Tỷ lệ bỏ học và nguyên nhân

1.1.1. Thực trạng tỷ lệ bỏ học

Tại Việt Nam, tỷ lệ bỏ học có sự phân hóa rõ rệt theo từng cấp học, đồng thời phản ánh sự chuyển dịch mạnh mẽ trong hành vi học tập của người học theo độ tuổi. Dữ liệu thống kê mới nhất từ Tổng cục Thống kê và Bộ Giáo dục và Đào tạo (2024–2025) cho thấy, ở cấp tiểu học, tỷ lệ đi học đúng độ tuổi đạt mức rất cao, khoảng 98,7%, tương đương với tỷ lệ bỏ học chỉ khoảng 1,3%. Điều này cho thấy hệ thống giáo dục cơ bản đã đạt được mức độ bao phủ gần như toàn diện.

Tuy nhiên, khi chuyển lên các cấp học cao hơn, tỷ lệ này có xu hướng suy giảm đáng kể. Ở cấp trung học cơ sở, tỷ lệ đi học đúng độ tuổi đạt khoảng 95,6%, đồng nghĩa với khoảng 4,4% học sinh không tiếp tục học tập hoặc không theo đúng lộ trình.

Đặc biệt, tại cấp trung học phổ thông, tỷ lệ này giảm mạnh xuống còn 79,9%, cho thấy có tới 20,1% thanh thiếu niên trong độ tuổi 15–17 đã rời khỏi hệ thống giáo dục phổ thông hoặc chuyển sang các hình thức đào tạo khác như giáo dục nghề nghiệp. Đây là một con số đáng lưu ý, phản ánh giai đoạn chuyển tiếp quan trọng nơi người học bắt đầu đưa ra các quyết định mang tính định hướng tương lai.

Ở bậc giáo dục đại học và giáo dục nghề nghiệp, vấn đề bỏ học trở nên phức tạp hơn khi không chỉ bao gồm việc thôi học giữa chừng mà còn xuất hiện hiện tượng "bỏ nhập học". Cụ thể, trong kỳ tuyển sinh năm 2025, khoảng 19,1% thí sinh trúng tuyển đại học đợt 1 đã không xác nhận nhập học, tương đương hơn 147.000 người. Điều này cho thấy sự thay đổi trong hành vi lựa chọn giáo dục của người học, khi đại học không còn là con đường duy nhất.

Đối với sinh viên đang theo học, tỷ lệ thôi học vẫn là một vấn đề đáng quan tâm. Theo quy định hiện hành (Thông tư 01/2024), các trường đại học được yêu cầu kiểm soát tỷ lệ thôi học không vượt quá 10% mỗi năm (và không quá 15% đối với năm nhất). Tuy nhiên, thực tế tại nhiều cơ sở đào tạo cho thấy số lượng sinh viên bị buộc thôi học hoặc tự ý bỏ học vẫn ở mức cao. Ví dụ, Đại học Bách khoa Hà Nội ghi nhận khoảng 700–800 sinh viên thôi học mỗi năm, trong khi Đại học Sư phạm Kỹ thuật TP.HCM có tới khoảng 2.000 trường hợp bị xử lý học vụ trong năm 2024.

Đáng chú ý, trong hệ thống giáo dục nghề nghiệp, tỷ lệ bỏ học có sự dao động lớn, từ 9% đến 35% tùy thuộc vào ngành nghề và cơ sở đào tạo. Điều này phản ánh sự thiếu ổn định trong định hướng nghề nghiệp của người học, cũng như sự chênh lệch về chất lượng đào tạo và cơ hội việc làm giữa các ngành.

Xét theo yếu tố vùng miền, sự chênh lệch càng trở nên rõ nét. Khu vực Đồng bằng sông Cửu Long được ghi nhận là vùng có tỷ lệ bỏ học cao nhất cả nước, với mức từng lên tới 13,3%. Tại các vùng sâu, vùng xa và khu vực biên giới, tình trạng học sinh bỏ học sau cấp trung học cơ sở để tham gia lao động sớm được đánh giá là "đáng báo động". Những con số này không chỉ phản ánh sự bất bình đẳng trong tiếp cận giáo dục mà còn cho thấy tác động mạnh mẽ của yếu tố kinh tế – xã hội đến quyết định học tập.

Trên phạm vi toàn cầu, tỷ lệ bỏ học ở bậc đại học dao động từ 20% đến 30% tùy theo quốc gia và hệ thống giáo dục. Điều này cho thấy đây là một vấn đề mang

tính phổ quát, không chỉ riêng ở các quốc gia đang phát triển mà cả ở những nền giáo dục tiên tiến.

1.1.2. Nguyên nhân của hiện tượng bỏ học

Hiện tượng sinh viên bỏ học không phải là kết quả của một nguyên nhân đơn lẻ, mà là hệ quả của sự tương tác phức tạp giữa nhiều nhóm yếu tố khác nhau, bao gồm yếu tố học thuật, kinh tế, tâm lý cá nhân, môi trường xã hội và các yếu tố mang tính hệ thống [5]. Những yếu tố này không tồn tại độc lập mà thường xuyên tác động qua lại, tạo thành một chuỗi nguyên nhân – hệ quả tích lũy theo thời gian. Do đó, việc phân tích toàn diện các nguyên nhân không chỉ giúp nhận diện bản chất của vấn đề mà còn đóng vai trò quan trọng trong việc xây dựng các mô hình dự đoán chính xác và đề xuất các giải pháp can thiệp hiệu quả trong thực tiễn.

Nhóm nguyên nhân học thuật

Yếu tố học thuật được xem là nguyên nhân trực tiếp và phổ biến nhất dẫn đến quyết định bỏ học của sinh viên. Trong môi trường giáo dục đại học, yêu cầu về năng lực tự học, tư duy phản biện và khả năng thích ứng với khối lượng kiến thức lớn cao hơn đáng kể so với các bậc học trước. Tuy nhiên, một bộ phận sinh viên chưa được trang bị đầy đủ các kỹ năng này, dẫn đến sự "lệch pha" giữa yêu cầu đào tạo và năng lực cá nhân.

Các biểu hiện cụ thể của khó khăn học thuật thường bao gồm điểm trung bình học tập thấp, không đạt chuẩn đầu ra theo từng học kỳ, tình trạng nợ tín chỉ kéo dài hoặc bị cảnh báo học vụ nhiều lần [5] [6]. Đáng chú ý, những thất bại ban đầu trong học tập có xu hướng tạo ra hiệu ứng dây chuyền, làm suy giảm sự tự tin và động lực học tập của sinh viên. Khi áp lực học tập vượt quá khả năng thích ứng, sinh viên dễ rơi vào trạng thái căng thẳng kéo dài, từ đó dẫn đến quyết định tạm dừng hoặc chấm dứt việc học.

Bên cạnh đó, phương pháp giảng dạy và cấu trúc chương trình đào tạo cũng là những yếu tố có ảnh hưởng đáng kể. Các chương trình mang tính lý thuyết cao, thiếu tính thực tiễn hoặc không phù hợp với năng lực đầu vào của sinh viên có thể làm gia tăng khoảng cách giữa người học và nội dung đào tạo, góp phần làm tăng nguy cơ bỏ học.

Nhóm nguyên nhân kinh tế

Yếu tố kinh tế đóng vai trò then chốt, đặc biệt trong bối cảnh chi phí giáo dục ngày càng gia tăng. Đối với nhiều sinh viên, đặc biệt là những người đến từ các gia đình có hoàn cảnh khó khăn, việc duy trì học tập không chỉ là vấn đề học thuật mà còn là bài toán tài chính.

Chi phí học tập bao gồm học phí, chi phí sinh hoạt, tài liệu học tập và các khoản chi phí phát sinh khác có thể tạo ra áp lực lớn đối với sinh viên [7][8]. Khi nguồn lực tài chính không đảm bảo, sinh viên buộc phải lựa chọn giữa việc tiếp tục học tập và tham gia lao động để hỗ trợ gia đình. Trong nhiều trường hợp, việc đi làm thêm ban đầu nhằm mục đích trang trải chi phí, nhưng dần dần lại chiếm phần lớn thời gian và năng lượng, làm suy giảm kết quả học tập và gia tăng nguy cơ bỏ học.

Ngoài ra, sự thiếu hụt hoặc chậm trễ trong các chính sách hỗ trợ tài chính như học bổng, vay vốn hoặc trợ cấp cũng có thể trở thành rào cản lớn. Điều này cho thấy yếu tố kinh tế không chỉ tác động trực tiếp mà còn gián tiếp thông qua việc ảnh hưởng đến các yếu tố khác như hiệu suất học tập và trạng thái tâm lý của sinh viên.

Nhóm nguyên nhân tâm lý – động lực

Yếu tố tâm lý và động lực học tập đóng vai trò cốt lõi trong việc duy trì sự gắn bó của sinh viên với môi trường đại học. Một trong những nguyên nhân phổ biến là việc lựa chọn ngành học không phù hợp với năng lực, sở thích hoặc định hướng nghề nghiệp cá nhân [9]. Nhiều sinh viên đưa ra quyết định chọn ngành dựa trên xu hướng xã hội, áp lực gia đình hoặc thông tin thiếu chính xác, dẫn đến tình trạng "lệch hướng" ngay từ đầu.

Sự thiếu rõ ràng về mục tiêu học tập và nghề nghiệp khiến sinh viên khó duy trì động lực dài hạn. Khi không nhận thấy ý nghĩa hoặc giá trị của việc học, sinh viên dễ rơi vào trạng thái chán nản, mất phương hướng và dần tách rời khỏi quá trình học tập. Bên cạnh đó, các vấn đề tâm lý như căng thẳng, lo âu, trầm cảm hoặc cảm giác cô lập trong môi trường đại học cũng ngày càng phổ biến, đặc biệt trong bối cảnh học tập cạnh tranh và áp lực thành tích cao [10].

Đáng chú ý, các yếu tố tâm lý thường không biểu hiện rõ ràng ngay từ đầu mà tích lũy theo thời gian. Khi không được phát hiện và hỗ trợ kịp thời, chúng có thể trở

thành nguyên nhân trực tiếp dẫn đến quyết định bỏ học. Điều này đặt ra yêu cầu cần có các cơ chế theo dõi và hỗ trợ tâm lý hiệu quả trong môi trường giáo dục đại học.

Nhóm nguyên nhân xã hội và môi trường học tập

Môi trường học tập và các yếu tố xã hội xung quanh có ảnh hưởng đáng kể đến trải nghiệm học tập và quyết định duy trì học tập của sinh viên. Một môi trường học tập thiếu sự hỗ trợ, thiếu tương tác giữa giảng viên và sinh viên hoặc thiếu các hoạt động gắn kết có thể khiến sinh viên cảm thấy bị cô lập và không thuộc về cộng đồng học thuật [9].

Phương pháp giảng dạy truyền thống, thiếu tính tương tác hoặc không cá nhân hóa cũng có thể làm giảm sự hứng thú và mức độ tham gia của sinh viên. Trong khi đó, các yếu tố xã hội như ảnh hưởng từ bạn bè, xu hướng "học nhanh – làm sớm", hoặc sự hấp dẫn của thị trường lao động ngắn hạn có thể khiến sinh viên ưu tiên lợi ích trước mắt thay vì đầu tư vào giáo dục dài hạn.

Ngoài ra, sự phát triển mạnh mẽ của công nghệ và mạng xã hội cũng tạo ra những tác động hai chiều. Một mặt, chúng hỗ trợ việc học tập; mặt khác, chúng có thể làm phân tán sự tập trung, giảm hiệu quả học tập và gián tiếp làm gia tăng nguy cơ bỏ học.

Nhóm nguyên nhân hệ thống

Ở cấp độ vĩ mô, những hạn chế trong hệ thống giáo dục và chính sách quản lý cũng góp phần quan trọng vào tỷ lệ bỏ học. Chương trình đào tạo chậm cập nhật, thiếu tính liên kết với nhu cầu của thị trường lao động hoặc thiếu linh hoạt trong cấu trúc học tập có thể làm giảm giá trị cảm nhận của người học đối với bằng cấp.

Bên cạnh đó, cơ chế quản lý và hỗ trợ sinh viên tại nhiều cơ sở đào tạo còn mang tính thụ động, thiếu các hệ thống cảnh báo sớm và can thiệp kịp thời [11]. Việc chưa tận dụng hiệu quả dữ liệu học tập để theo dõi tiến trình học tập của sinh viên khiến nhiều trường hợp có nguy cơ cao không được phát hiện sớm. Việc nhận diện sớm sinh viên có nguy cơ bỏ học cũng cần xem xét các yếu tố động lực học tập và định hướng nghề nghiệp, bởi đây là những yếu tố có liên quan trực tiếp đến ý định bỏ học của sinh viên trong giai đoạn đầu đại học [12]. Bên cạnh đó, các yếu tố kinh tế - xã hội cũng được xem là nhóm nguyên nhân quan trọng ảnh hưởng đến quyết định duy trì hay rời bỏ quá trình học tập của người học [13].

Khi người học không nhìn thấy mối liên hệ rõ ràng giữa việc học và cơ hội nghề nghiệp trong tương lai, họ có xu hướng tìm kiếm các con đường thay thế mang lại lợi ích nhanh hơn. Điều này đặc biệt rõ trong bối cảnh thị trường lao động ngày càng mở rộng với nhiều cơ hội không yêu cầu bằng cấp truyền thống.

1.2. Các mô hình học sâu phục vụ bài toàn

1.2.1. Tổng quan về học máy và học sâu trong dự đoán

Trong những năm gần đây, sự phát triển mạnh mẽ của khoa học dữ liệu và trí tuệ nhân tạo đã mở ra nhiều hướng tiếp cận mới trong việc phân tích và dự đoán các hiện tượng phức tạp trong nhiều lĩnh vực, đặc biệt là giáo dục. Trong đó, học máy (Machine Learning – ML) và học sâu (Deep Learning – DL) được xem là hai nền tảng cốt lõi, cho phép khai thác hiệu quả dữ liệu lớn nhằm phát hiện các quy luật tiềm ẩn và đưa ra dự đoán có độ chính xác cao [14][15].

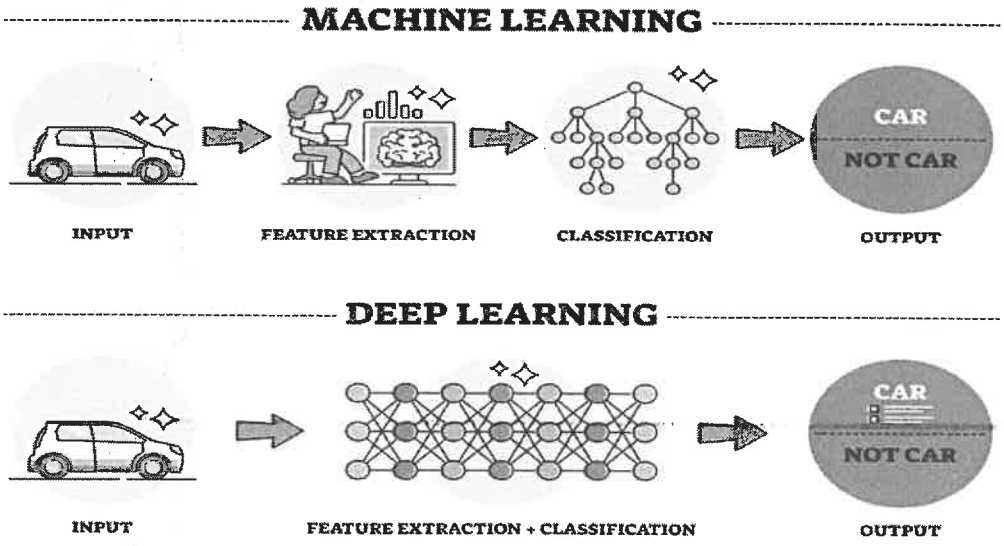


Figure 1.1. Sự khác biệt giữa ML và DL

Học máy là một nhánh của trí tuệ nhân tạo, tập trung vào việc xây dựng các thuật toán có khả năng học từ dữ liệu mà không cần lập trình tường minh cho từng trường hợp cụ thể. Các mô hình học máy truyền thống như Logistic Regression, Decision Tree, Random Forest hay Support Vector Machine đã được áp dụng rộng rãi trong các bài toán phân loại và dự đoán [16]. Những mô hình này thường yêu cầu

quá trình lựa chọn đặc trưng (feature engineering) thủ công và giả định mối quan hệ giữa các biến đầu vào và đầu ra ở mức tương đối đơn giản.

Tuy nhiên, trong các bài toán thực tế như dự đoán hành vi học tập của sinh viên, dữ liệu thường mang tính phi tuyến, đa chiều và có sự phụ thuộc theo thời gian [17][18]. Khi đó, các phương pháp học máy truyền thống gặp nhiều hạn chế trong việc nắm bắt các mối quan hệ phức tạp và động giữa các yếu tố. Một số nghiên cứu gần đây cũng cho thấy các kỹ thuật xử lý mất cân bằng dữ liệu và các mô hình như XGBoost có thể cải thiện hiệu quả dự đoán trong các bài toán giáo dục, đặc biệt đối với dữ liệu dạng bảng [19][20][21], tuy nhiên nó vẫn gặp những hạn chế.Vì vậy, đây chính là bối cảnh mà học sâu nổi lên như một giải pháp hiệu quả.

Học sâu là một nhánh nâng cao của học máy, sử dụng các mạng nơ-ron nhân tạo nhiều tầng (deep neural networks) để tự động học các đặc trưng từ dữ liệu mà không cần quá nhiều sự can thiệp thủ công. Nhờ cấu trúc nhiều lớp, các mô hình học sâu có khả năng trích xuất các biểu diễn dữ liệu ở nhiều cấp độ khác nhau, từ đơn giản đến phức tạp. Điều này giúp mô hình không chỉ học được các mối quan hệ tuyến tính mà còn có thể phát hiện các mẫu phi tuyến và tương tác phức tạp giữa các biến.

Một ưu điểm nổi bật khác của học sâu là khả năng xử lý hiệu quả dữ liệu có cấu trúc phức tạp như dữ liệu chuỗi thời gian, dữ liệu văn bản hoặc dữ liệu hành vi. Trong bối cảnh giáo dục, dữ liệu học tập của sinh viên thường được thu thập theo từng học kỳ, từng môn học hoặc theo dòng thời gian, phản ánh quá trình thay đổi liên tục của người học [22]. Các mô hình học sâu, đặc biệt là các kiến trúc như RNN và LSTM, có khả năng ghi nhớ và khai thác thông tin theo chuỗi, từ đó nắm bắt được xu hướng và sự biến động trong hành vi học tập.

Bên cạnh đó, học sâu còn thể hiện ưu thế rõ rệt khi làm việc với dữ liệu lớn (big data). Khi lượng dữ liệu ngày càng tăng về quy mô và độ đa dạng, các mô hình học sâu có khả năng tận dụng tốt nguồn dữ liệu này để cải thiện hiệu suất dự đoán. Điều này đặc biệt phù hợp với các hệ thống quản lý học tập (LMS), nơi lưu trữ lượng lớn dữ liệu liên quan đến điểm số, mức độ tham gia, tương tác học tập và nhiều yếu tố khác của sinh viên.

Đối với bài toán dự đoán tỷ lệ sinh viên bỏ học, bản chất của vấn đề là xác định các mẫu hành vi tiềm ẩn dẫn đến quyết định rời bỏ chương trình học. Đây là một bài toán phức tạp, chịu ảnh hưởng đồng thời của nhiều yếu tố như kết quả học tập, điều

kiện kinh tế, động lực cá nhân và môi trường xã hội. Các yếu tố này không chỉ tương tác theo chiều không gian (đa biến) mà còn biến đổi theo thời gian, đòi hỏi một mô hình có khả năng học được cả hai khía cạnh này.

Trong bối cảnh đó, học sâu không chỉ cung cấp công cụ mạnh mẽ để mô hình hóa các mối quan hệ phức tạp mà còn cho phép xây dựng các hệ thống dự đoán có khả năng thích ứng và mở rộng theo dữ liệu thực tế. Khả năng tự động học đặc trưng, xử lý dữ liệu chuỗi và khai thác hiệu quả dữ liệu lớn khiến học sâu trở thành lựa chọn phù hợp và có tiềm năng cao trong việc giải quyết bài toán dự đoán hành vi học tập.

Vì vậy, các mô hình học sâu được xem là đặc biệt phù hợp cho bài toán dự đoán hành vi học tập của sinh viên, trong đó bao gồm việc nhận diện sớm nguy cơ bỏ học, góp phần hỗ trợ các cơ sở giáo dục đưa ra các biện pháp can thiệp kịp thời và nâng cao chất lượng đào tạo.

1.2.2. Tổng quan về mô hình RNN

Mạng nơ-ron hồi tiếp (Recurrent Neural Network – RNN) là một lớp mô hình học sâu được thiết kế chuyên biệt để xử lý dữ liệu dạng chuỗi (sequential data), trong đó các phần tử dữ liệu có mối liên hệ phụ thuộc theo thời gian hoặc thứ tự. Khác với các mạng nơ-ron truyền thẳng, nơi mỗi đầu vào được xử lý độc lập, RNN cho phép lưu trữ và khai thác thông tin từ các bước thời gian trước đó thông qua cơ chế trạng thái ẩn (hidden state). Nhờ đó, mô hình có khả năng học được các mối quan hệ động và phụ thuộc theo chuỗi trong dữ liệu.

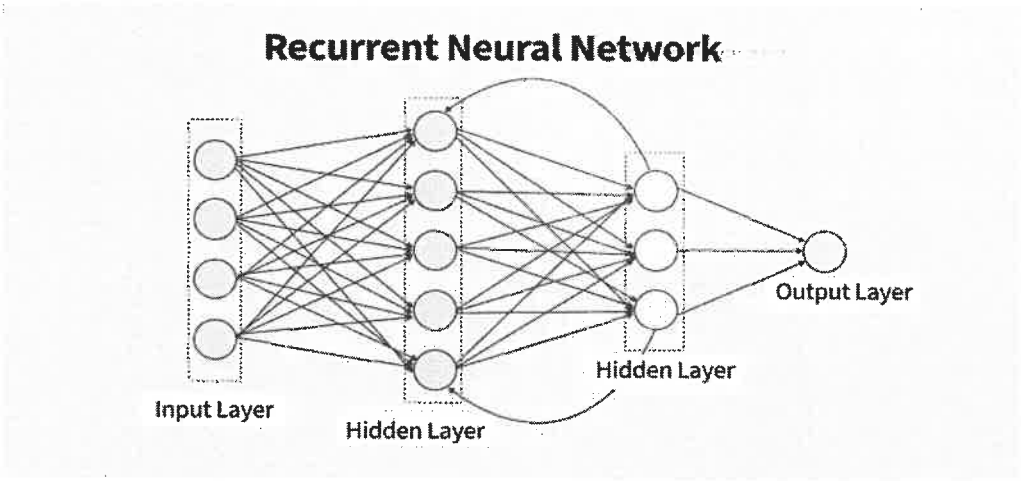


Figure 1.2. Mô hình RNN

Cụ thể, tại mỗi thời điểm t, RNN nhận đầu vào hiện tại xt và kết hợp với trạng thái ẩn của bước trước đó ht-1 để tạo ra trạng thái ẩn mới ht. Quá trình này có thể được biểu diễn dưới dạng toán học như sau:

$$h_t = f\left(W_{\{h\}h_{\{t-1\}}} + W_{\{x\}x_t} + b_h\right)$$

Trong đó Wh và Wx là các ma trận trọng số, bh là vector bias, và f là hàm kích hoạt phi tuyến, thường là hàm tanh hoặc ReLU. Trạng thái ẩn ht đóng vai trò như một "bộ nhớ" tạm thời, lưu trữ thông tin đã được tích lũy từ các bước trước đó. Dựa trên trạng thái này, đầu ra của mô hình tại thời điểm t được tính như sau:

$$y_{\{t\}} = W_{\{y\}}h_{\{t\}} + b_{\{y\}}$$

Cấu trúc hồi tiếp này cho phép RNN xử lý hiệu quả các bài toán mà kết quả phụ thuộc vào lịch sử dữ liệu, chẳng hạn như nhận dạng giọng nói, xử lý ngôn ngữ tự nhiên, dự báo chuỗi thời gian và đặc biệt là phân tích hành vi theo thời gian.

Trong bối cảnh bài toán dự đoán sinh viên bỏ học, dữ liệu học tập thường mang tính chuỗi, phản ánh quá trình tích lũy theo từng học kỳ hoặc từng giai đoạn học tập. Ví dụ, điểm số, số tín chỉ tích lũy, mức độ tham gia học tập hay các cảnh báo học vụ đều thay đổi theo thời gian và có ảnh hưởng lẫn nhau. RNN, với khả năng ghi nhớ và cập nhật trạng thái theo chuỗi, có thể khai thác hiệu quả các đặc trưng này để phát hiện các xu hướng tiềm ẩn trong hành vi học tập của sinh viên. Một trong những ưu điểm quan trọng của RNN là khả năng "ghi nhớ theo thời gian", tức là mô hình có thể duy trì thông tin từ các bước trước để hỗ trợ việc ra quyết định ở bước hiện tại. Điều này đặc biệt hữu ích trong các bài toán mà dữ liệu hiện tại không đủ để đưa ra dự đoán chính xác nếu không xét đến ngữ cảnh trước đó. Nhờ cơ chế này, RNN có thể mô hình hóa các mối quan hệ tuần tự và học được các mẫu hành vi mang tính thời gian một cách hiệu quả.

Tuy nhiên, bên cạnh những ưu điểm nổi bật, RNN cũng tồn tại một số hạn chế đáng kể. Vấn đề lớn nhất là hiện tượng "vanishing gradient" (tiêu biến gradient) và "exploding gradient" (bùng nổ gradient) trong quá trình lan truyền ngược (backpropagation through time – BPTT). Khi chuỗi dữ liệu quá dài, gradient có thể trở nên rất nhỏ hoặc rất lớn, khiến mô hình khó học được các phụ thuộc dài hạn hoặc dẫn đến quá trình huấn luyện không ổn định. Điều này làm giảm hiệu quả của RNN trong các bài toán yêu cầu ghi nhớ thông tin trong khoảng thời gian dài.

Chính vì hạn chế này, các biến thể nâng cao của RNN như Long Short-Term Memory (LSTM) và Gated Recurrent Unit (GRU) đã được phát triển nhằm cải thiện khả năng ghi nhớ dài hạn thông qua các cơ chế cổng (gating mechanisms) [23][24]. Những mô hình này cho phép kiểm soát luồng thông tin một cách linh hoạt, từ đó khắc phục hiệu quả các vấn đề của RNN truyền thống.

1.2.3. Tổng quan mô hình LSTM

Mạng Long Short-Term Memory (LSTM) là một biến thể nâng cao của mạng nơ-ron hồi tiếp (RNN), được đề xuất nhằm khắc phục những hạn chế của RNN truyền thống, đặc biệt là vấn đề tiêu biến gradient (vanishing gradient) trong quá trình huấn luyện. Trong khi RNN gặp khó khăn trong việc ghi nhớ thông tin ở các khoảng thời gian dài, LSTM được thiết kế với một kiến trúc đặc biệt cho phép duy trì và kiểm soát thông tin một cách hiệu quả xuyên suốt chuỗi dữ liệu.

Điểm cốt lõi của LSTM nằm ở việc giới thiệu khái niệm trạng thái tế bào (cell state), đóng vai trò như một "đường truyền thông tin" xuyên suốt các bước thời gian. Khác với trạng thái ẩn trong RNN chỉ mang tính tạm thời, cell state trong LSTM có khả năng lưu trữ thông tin dài hạn và hạn chế sự suy giảm của gradient. Để điều khiển dòng thông tin đi qua cell state, LSTM sử dụng một hệ thống các cổng (gates), bao gồm: cổng quên (forget gate), cổng đầu vào (input gate) và cổng đầu ra (output gate). Các cổng này hoạt động như những bộ lọc thông minh, quyết định thông tin nào cần được giữ lại, cập nhật hoặc loại bỏ.

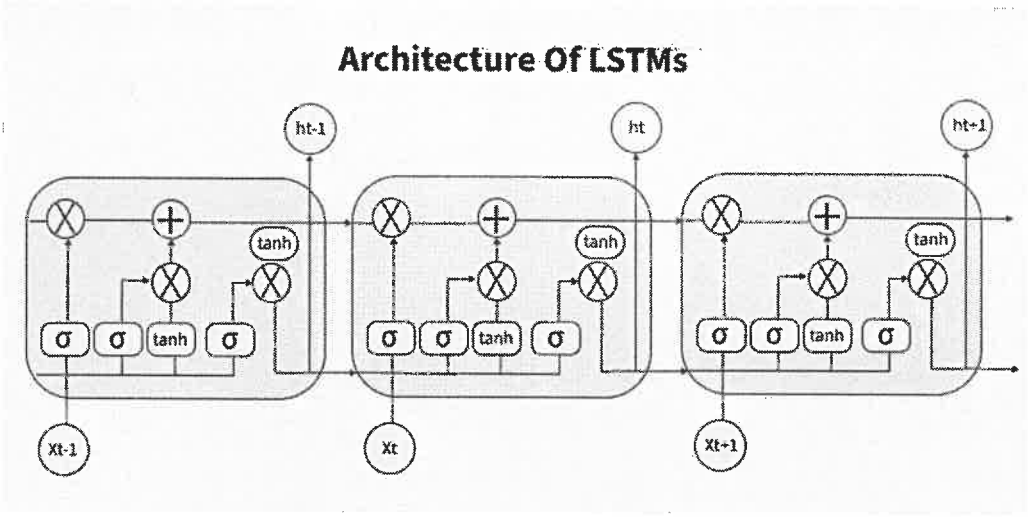


Figure 1.3. Mô hình LSTM

Cụ thể, tại mỗi thời điểm t, forget gave (cổng quên) xác định mức độ thông tin từ trạng thái trước đó cần được duy trì:

$$f_{\{t\}} = \sigma\left(W_{\{f\}}\left[h_{\{t-1\}}, x_{\{t\}}\right] + b_{\{f\}}\right)$$

Trong đó ft là vector quyết định quên hay giữ thông tin, với giá trị nằm trong khoảng [0, 1]. Giá trị càng gần 0 thì thông tin càng bị loại bỏ, ngược lại càng gần 1 thì thông tin càng được giữ lại.

Tiếp theo inpute gate (cổng đầu vào) kiểm soát lượng thông tin mới được đưa vào bộ nhớ, kết hợp với trạng thái ứng viên $\{\widetilde{c}\}_t$ để tạo ra nội dung cập nhật:

$$c_t = f_t \cdot c_{\{t-1\}} + i_t \cdot \{\widetilde{c}\}_t$$

Trong đó i_t là cổng đầu vào quyết định mức độ tiếp nhận thông tin mới, còn $\{\widetilde{c}\}_t$ là trạng thái ứng viên chứa thông tin tiềm năng được thêm vào bộ nhớ. Cơ chế này cho phép LSTM linh hoạt lựa chọn giữa việc duy trì thông tin cũ và cập nhật thông tin mới, từ đó hình thành một biểu diễn dữ liệu có tính thích nghi cao theo thời gian. Cuối cùng, output gate (cổng đầu ra) xác định phần thông tin từ cell state sẽ được xuất ra dưới dạng trạng thái ẩn h_t, phục vụ cho bước tính toán tiếp theo:

$$h_t = o_t \,.\tanh(c_t)$$

Thông qua việc điều phối ba cổng này, LSTM có khả năng kiểm soát dòng thông tin một cách chặt chẽ, giúp duy trì các phụ thuộc dài hạn mà không bị ảnh hưởng bởi vấn đề tiêu biến gradient như trong RNN truyền thống.

Nhờ kiến trúc đặc biệt này, LSTM thể hiện nhiều ưu điểm vượt trội so với RNN. Trước hết, khả năng ghi nhớ dài hạn giúp mô hình học được các mối quan hệ phức tạp trong dữ liệu chuỗi, đặc biệt là những phụ thuộc không liền kề. Thứ hai, cơ chế cổng giúp lọc bỏ thông tin không cần thiết và giữ lại những đặc trưng quan trọng, từ đó nâng cao chất lượng biểu diễn dữ liệu. Ngoài ra, LSTM cũng có độ ổn định cao hơn trong quá trình huấn luyện, giảm thiểu hiện tượng mất mát thông tin qua nhiều bước thời gian.

Trong bài toán dự đoán sinh viên bỏ học, các yếu tố như kết quả học tập, mức độ tham gia học tập hay các cảnh báo học vụ không chỉ ảnh hưởng tại một thời điểm mà còn có tác động tích lũy theo thời gian. Do đó, việc nắm bắt các xu hướng dài hạn trong hành vi học tập là rất quan trọng. LSTM, với khả năng xử lý hiệu quả các phụ

thuộc dài hạn, cho phép mô hình nhận diện sớm các dấu hiệu rủi ro dựa trên toàn bộ lịch sử học tập của sinh viên, thay vì chỉ dựa vào dữ liệu tại một thời điểm đơn lẻ.

1.2.4. Tổng quan về mô hình Hybrid

Trong các bài toán dự đoán thực tế, đặc biệt là trong lĩnh vực giáo dục, dữ liệu đầu vào thường có tính chất đa dạng, bao gồm cả dữ liệu số (numerical data) và dữ liệu phân loại (categorical data). Mỗi loại dữ liệu mang những đặc trưng riêng và yêu cầu các phương pháp xử lý khác nhau để có thể khai thác hiệu quả thông tin. Do đó, việc sử dụng một mô hình đơn lẻ có thể không tận dụng được toàn bộ tiềm năng của dữ liệu. Xuất phát từ yêu cầu đó, mô hình Hybrid được đề xuất như một hướng tiếp cận linh hoạt và hiệu quả, cho phép kết hợp nhiều kỹ thuật xử lý dữ liệu trong cùng một kiến trúc thống nhất.

Ý tưởng cốt lõi của mô hình Hybrid là tách biệt các loại đặc trưng đầu vào theo bản chất dữ liệu, từ đó áp dụng các phương pháp xử lý phù hợp cho từng nhóm, trước khi tích hợp lại để đưa ra dự đoán cuối cùng. Cụ thể, đối với các đặc trưng dạng số như điểm trung bình (GPA), số tín chỉ tích lũy, số tín chỉ nợ hoặc tần suất tham gia học tập, mô hình sử dụng mạng nơ-ron sâu (Deep Neural Network – DNN) để học các mối quan hệ phi tuyến giữa các biến. Nhờ khả năng trích xuất đặc trưng đa tầng, DNN có thể phát hiện các mẫu ẩn phức tạp trong dữ liệu số mà các phương pháp truyền thống khó có thể nhận diện.

Đối với các đặc trưng phân loại như ngành học, giới tính, khu vực địa lý hoặc trạng thái học tập, việc sử dụng phương pháp mã hóa thông thường như one-hot encoding có thể dẫn đến hiện tượng dữ liệu thưa (sparse data) và làm tăng độ phức tạp của mô hình. Thay vào đó, mô hình Hybrid áp dụng kỹ thuật embedding, cho phép biểu diễn các biến phân loại dưới dạng vector liên tục có kích thước thấp hơn nhưng vẫn giữ được mối quan hệ ngữ nghĩa giữa các giá trị. Các vector embedding này sau đó được đưa vào các lớp mạng nơ-ron để tiếp tục học các đặc trưng có ý nghĩa.

Sau khi xử lý riêng biệt từng nhóm dữ liệu, các đặc trưng đã được trích xuất sẽ được kết hợp (concatenate) thành một vector đặc trưng tổng hợp. Vector này đại diện cho toàn bộ thông tin đầu vào ở mức trừu tượng cao hơn, từ đó được đưa qua các lớp fully-connected để thực hiện nhiệm vụ dự đoán. Cách tiếp cận này cho phép mô hình

tận dụng tối đa thông tin từ nhiều nguồn dữ liệu khác nhau, đồng thời giảm thiểu mất mát thông tin trong quá trình xử lý.

Ưu điểm nổi bật của mô hình Hybrid là khả năng tích hợp linh hoạt nhiều loại dữ liệu và nhiều kỹ thuật học sâu trong cùng một kiến trúc, giúp nâng cao hiệu suất dự đoán so với các mô hình đơn lẻ. Đồng thời, việc phân tách và xử lý riêng biệt các nhóm đặc trưng cũng giúp mô hình dễ dàng mở rộng và điều chỉnh khi có thêm dữ liệu mới.

Trong bối cảnh bài toán dự đoán sinh viên bỏ học, dữ liệu thường bao gồm cả thông tin học tập (dạng số) và thông tin cá nhân, xã hội (dạng phân loại). Việc kết hợp các nguồn dữ liệu này một cách hiệu quả đóng vai trò quan trọng trong việc nâng cao độ chính xác của mô hình. Do đó, mô hình Hybrid không chỉ là một giải pháp kỹ thuật mà còn là một hướng tiếp cận phù hợp với bản chất đa chiều của dữ liệu giáo dục. Nhờ khả năng kết hợp linh hoạt và khai thác hiệu quả nhiều loại đặc trưng, mô hình Hybrid được kỳ vọng sẽ mang lại hiệu quả vượt trội trong bài toán dự đoán hành vi học tập, đặc biệt là trong việc nhận diện sớm nguy cơ sinh viên bỏ học.

1.2.5. Lý do lựa chọn mô hình cho bài toán

Việc lựa chọn mô hình phù hợp đóng vai trò then chốt trong việc xây dựng hệ thống dự đoán hiệu quả, đặc biệt đối với các bài toán phức tạp như dự đoán nguy cơ sinh viên bỏ học. Để đưa ra quyết định lựa chọn mô hình, cần xem xét kỹ lưỡng đặc điểm của dữ liệu cũng như bản chất của bài toán.

Trước hết, dữ liệu học tập của sinh viên mang tính chuỗi thời gian (time-series) rõ rệt. Các thông tin như điểm trung bình học kỳ, số tín chỉ tích lũy, số tín chỉ nợ, hay các cảnh báo học vụ đều được ghi nhận theo từng giai đoạn học tập và có sự phụ thuộc lẫn nhau theo thời gian. Quyết định bỏ học của sinh viên không phải là một sự kiện xảy ra tức thời, mà là kết quả của một quá trình tích lũy các yếu tố bất lợi trong suốt quá trình học tập. Do đó, mô hình dự đoán cần có khả năng khai thác được các mối quan hệ theo chuỗi và ghi nhớ thông tin từ các bước trước đó.

Bên cạnh yếu tố thời gian, dữ liệu sinh viên còn có tính chất đa dạng về kiểu dữ liệu, bao gồm cả đặc trưng dạng số (numerical) và đặc trưng dạng phân loại (categorical). Các đặc trưng số như điểm số, số tín chỉ hay tần suất tham gia học tập phản ánh trực tiếp hiệu suất học tập, trong khi các đặc trưng phân loại như ngành học, giới tính, khu vực địa lý hay hoàn cảnh gia đình lại cung cấp bối cảnh bổ sung quan

trọng. Việc xử lý đồng thời hai loại dữ liệu này đòi hỏi mô hình phải có khả năng biểu diễn linh hoạt và tận dụng được thông tin từ nhiều nguồn khác nhau.

Xuất phát từ những đặc điểm trên, các mô hình học sâu, đặc biệt là các kiến trúc xử lý dữ liệu chuỗi, được xem là lựa chọn phù hợp. Trong đó, RNN được sử dụng như một mô hình cơ sở nhờ khả năng xử lý dữ liệu theo chuỗi và ghi nhớ thông tin ngắn hạn. Tuy nhiên, do hạn chế trong việc học các phụ thuộc dài hạn, mô hình LSTM được lựa chọn như một giải pháp cải tiến, cho phép mô hình hóa hiệu quả hơn các xu hướng dài hạn trong hành vi học tập của sinh viên thông qua cơ chế bộ nhớ và các cổng điều khiển thông tin.

Bên cạnh đó, để tận dụng tối đa tính đa dạng của dữ liệu đầu vào, mô hình Hybrid được đề xuất nhằm kết hợp các phương pháp xử lý khác nhau cho từng loại đặc trưng. Cụ thể, các đặc trưng dạng số được xử lý thông qua mạng nơ-ron sâu (DNN), trong khi các đặc trưng phân loại được biểu diễn bằng embedding để giảm chiều dữ liệu và giữ được thông tin ngữ nghĩa. Việc tích hợp các đặc trưng này trong một kiến trúc thống nhất giúp mô hình khai thác hiệu quả cả khía cạnh thời gian và khía cạnh đa dạng dữ liệu.

Sự kết hợp giữa RNN, LSTM và Hybrid không chỉ cho phép so sánh hiệu quả giữa các mô hình mà còn giúp đánh giá mức độ đóng góp của từng hướng tiếp cận trong bài toán dự đoán. Đồng thời, cách tiếp cận này cũng tạo nền tảng cho việc xây dựng các hệ thống dự đoán linh hoạt, có khả năng mở rộng và thích ứng với các nguồn dữ liệu khác nhau trong thực tế.

1.3. Dữ liệu và công cụ

1.3.1. Dữ liệu nghiên cứu

Trong nghiên cứu này, bộ dữ liệu được sử dụng có tên "Predicting Student Dropout and Academic Success", là một tập dữ liệu thực tế được xây dựng nhằm phục vụ cho các bài toán phân tích và dự đoán kết quả học tập cũng như nguy cơ bỏ học của sinh viên. Bộ dữ liệu này cung cấp một nền tảng toàn diện, cho phép khai thác nhiều khía cạnh khác nhau liên quan đến hành vi học tập và quá trình đào tạo của người học.

Cụ thể, bộ dữ liệu bao gồm 4.424 bản ghi, trong đó mỗi bản ghi tương ứng với một sinh viên. Mỗi sinh viên được mô tả thông qua 35 thuộc tính (features), phản ánh

đa dạng các yếu tố ảnh hưởng đến quá trình học tập. Các thuộc tính này có thể được phân thành ba nhóm chính: (i) thông tin cá nhân và nhân khẩu học (demographic), (ii) thông tin học tập và kết quả đào tạo (academic), và (iii) các yếu tố kinh tế – xã hội (socio-economic). Việc kết hợp các nhóm đặc trưng này giúp xây dựng một cái nhìn toàn diện về các yếu tố tác động đến khả năng duy trì học tập của sinh viên.

Biến mục tiêu (target variable) của bài toán được định nghĩa dưới dạng phân loại, phản ánh trạng thái học tập cuối cùng của sinh viên. Trong nghiên cứu này, biến mục tiêu bao gồm hai lớp chính: "Graduate" (tốt nghiệp) và "Dropout" (bỏ học). Phân bố cụ thể của dữ liệu cho thấy có 3.003 sinh viên thuộc nhóm tốt nghiệp và 1.421 sinh viên thuộc nhóm bỏ học. Sự phân bố này cho thấy dữ liệu có tồn tại mất cân bằng nhẹ (class imbalance), khi số lượng sinh viên tốt nghiệp chiếm tỷ lệ cao hơn so với sinh viên bỏ học.

Mặc dù mức độ mất cân bằng không quá lớn, nhưng đây vẫn là một yếu tố cần được xem xét trong quá trình xây dựng và đánh giá mô hình, bởi nó có thể ảnh hưởng đến khả năng học của mô hình đối với lớp thiểu số (Dropout). Đồng thời, sự tồn tại của hai nhóm dữ liệu rõ ràng cũng tạo điều kiện thuận lợi cho việc xây dựng các mô hình phân loại nhị phân, giúp mô hình tập trung vào việc nhận diện chính xác các trường hợp có nguy cơ bỏ học.

Một điểm đáng chú ý khác của bộ dữ liệu là sự đa dạng về kiểu dữ liệu, bao gồm cả các biến dạng số (numerical) và biến phân loại (categorical). Điều này đặt ra yêu cầu cần có các phương pháp tiền xử lý và biểu diễn dữ liệu phù hợp, đồng thời cũng là cơ sở để áp dụng các mô hình học sâu có khả năng xử lý đa dạng đặc trưng, như đã trình bày trong các phần trước.Nhìn chung, với quy mô tương đối lớn, số lượng đặc trưng phong phú và phản ánh nhiều khía cạnh khác nhau của quá trình học tập, bộ dữ liệu này được đánh giá là phù hợp cho bài toán dự đoán sinh viên bỏ học. Đây cũng là nền tảng quan trọng để triển khai và đánh giá hiệu quả của các mô hình học sâu trong các chương tiếp theo.

1.3.2. Công cụ và thư viện sử dụng

TensorFlow

Trong nghiên cứu này, TensorFlow được sử dụng như một trong những nền tảng chính để xây dựng và triển khai các mô hình học sâu, đặc biệt là các kiến trúc mạng nơ-ron phục vụ cho bài toán dự đoán nguy cơ sinh viên bỏ học. TensorFlow là một

thư viện mã nguồn mở do Google phát triển, được thiết kế nhằm hỗ trợ xây dựng các hệ thống học máy có khả năng mở rộng, hiệu quả và phù hợp với các bài toán xử lý dữ liệu lớn.

Về bản chất, TensorFlow hoạt động dựa trên cơ chế đồ thị tính toán (computational graph), trong đó các phép toán được biểu diễn dưới dạng các nút (nodes) và dữ liệu được truyền qua các cạnh (edges). Cách tiếp cận này cho phép tối ưu hóa quá trình tính toán, đặc biệt khi triển khai trên các hệ thống phần cứng hiệu năng cao như GPU hoặc TPU. Nhờ đó, TensorFlow có khả năng xử lý hiệu quả các mô hình phức tạp với số lượng lớn tham số và dữ liệu đầu vào.

Một ưu điểm quan trọng của TensorFlow là khả năng hỗ trợ xây dựng đa dạng các kiến trúc mạng nơ-ron, từ các mô hình cơ bản như mạng nơ-ron truyền thẳng (DNN) đến các mô hình xử lý dữ liệu chuỗi như RNN và LSTM. Điều này đặc biệt phù hợp với bài toán trong nghiên cứu, khi dữ liệu sinh viên vừa có tính chất chuỗi thời gian, vừa bao gồm nhiều loại đặc trưng khác nhau. TensorFlow cung cấp các lớp và hàm dựng sẵn, cho phép dễ dàng triển khai các mô hình này mà không cần xây dựng từ đầu.

Bên cạnh đó, thư viện Keras, được tích hợp trực tiếp trong TensorFlow, đóng vai trò như một giao diện lập trình mức cao (high-level API), giúp đơn giản hóa quá trình thiết kế và huấn luyện mô hình. Thông qua Keras, người dùng có thể nhanh chóng xây dựng các mô hình học sâu bằng cách định nghĩa các lớp (layers), hàm kích hoạt (activation functions), hàm mất mát (loss functions) và thuật toán tối ưu (optimizers) một cách trực quan. Điều này không chỉ giúp rút ngắn thời gian phát triển mà còn tăng tính linh hoạt trong việc thử nghiệm và điều chỉnh mô hình.

Ngoài ra, TensorFlow còn cung cấp nhiều công cụ hỗ trợ trong quá trình huấn luyện và đánh giá mô hình, như cơ chế theo dõi quá trình học (training logs), lưu và tải mô hình (model checkpointing), cũng như các công cụ trực quan hóa như TensorBoard. Những tính năng này giúp người nghiên cứu dễ dàng theo dõi hiệu suất mô hình, phát hiện các vấn đề trong quá trình huấn luyện và đưa ra các điều chỉnh phù hợp.

Trong bối cảnh nghiên cứu này, TensorFlow được sử dụng chủ yếu để triển khai các mô hình học sâu như RNN, LSTM và mô hình Hybrid, cho phép tận dụng tối đa khả năng xử lý dữ liệu chuỗi và dữ liệu đa dạng. Việc lựa chọn TensorFlow không

chỉ dựa trên hiệu năng và độ ổn định, mà còn bởi tính phổ biến và cộng đồng hỗ trợ mạnh mẽ, giúp đảm bảo khả năng tái lập và mở rộng của nghiên cứu trong tương lai.

PyTorch

Bên cạnh TensorFlow, PyTorch cũng được sử dụng như một công cụ quan trọng trong nghiên cứu nhằm hỗ trợ quá trình xây dựng và thử nghiệm các mô hình học sâu. PyTorch là một thư viện mã nguồn mở được phát triển bởi Facebook (Meta AI), nổi bật với tính linh hoạt cao và khả năng hỗ trợ hiệu quả cho các bài toán nghiên cứu và phát triển mô hình trí tuệ nhân tạo.

Khác với TensorFlow – vốn sử dụng cơ chế đồ thị tính toán tĩnh (static graph) trong các phiên bản truyền thống, PyTorch được xây dựng dựa trên đồ thị tính toán động (dynamic computation graph). Điều này có nghĩa là cấu trúc mô hình được định nghĩa và thay đổi linh hoạt trong quá trình thực thi, cho phép người dùng dễ dàng điều chỉnh kiến trúc, kiểm tra từng bước tính toán và gỡ lỗi (debug) một cách trực quan. Chính đặc điểm này khiến PyTorch trở thành lựa chọn phổ biến trong các nghiên cứu học thuật, nơi yêu cầu cao về tính thử nghiệm và khả năng tùy biến mô hình.

Một ưu điểm nổi bật khác của PyTorch là cú pháp lập trình gần gũi với Python, giúp quá trình xây dựng mô hình trở nên trực quan và dễ hiểu hơn. Các thành phần của mạng nơ-ron như lớp (layers), hàm kích hoạt, hàm mất mát và thuật toán tối ưu đều được thiết kế theo hướng đơn giản hóa, nhưng vẫn đảm bảo tính linh hoạt và khả năng mở rộng. Điều này cho phép người nghiên cứu dễ dàng triển khai các kiến trúc phức tạp như RNN và LSTM, đồng thời thực hiện các thử nghiệm với nhiều cấu hình khác nhau một cách nhanh chóng.

PyTorch cũng cung cấp khả năng hỗ trợ mạnh mẽ cho việc tính toán song song trên GPU, giúp tăng tốc đáng kể quá trình huấn luyện mô hình, đặc biệt khi làm việc với dữ liệu có quy mô lớn. Thư viện này tích hợp chặt chẽ với các công cụ xử lý tensor hiệu năng cao, cho phép thực hiện các phép toán ma trận và tối ưu hóa mô hình một cách hiệu quả.

Trong nghiên cứu này, PyTorch được sử dụng chủ yếu trong các giai đoạn thử nghiệm và phát triển mô hình, nơi cần đánh giá và so sánh nhiều kiến trúc khác nhau trước khi lựa chọn mô hình tối ưu. Nhờ khả năng linh hoạt trong việc xây dựng và

điều chỉnh mô hình, PyTorch hỗ trợ hiệu quả trong việc kiểm chứng các giả thuyết và tinh chỉnh các tham số của mô hình học sâu.

Việc kết hợp sử dụng cả TensorFlow và PyTorch không chỉ giúp tận dụng ưu điểm của từng nền tảng mà còn tăng độ tin cậy của kết quả nghiên cứu thông qua việc kiểm chứng chéo. Đồng thời, điều này cũng đảm bảo rằng các mô hình được xây dựng có thể dễ dàng triển khai và mở rộng trong nhiều môi trường khác nhau.

Scikit-learn (sklearn)

Bên cạnh các thư viện học sâu như TensorFlow và PyTorch, Scikit-learn (sklearn) được sử dụng như một công cụ quan trọng trong nghiên cứu nhằm hỗ trợ quá trình tiền xử lý dữ liệu, xây dựng các mô hình cơ sở và đánh giá hiệu suất của mô hình. Scikit-learn là một thư viện học máy phổ biến trong Python, cung cấp một tập hợp phong phú các thuật toán và công cụ phục vụ cho toàn bộ quy trình phân tích dữ liệu.

Một trong những vai trò chính của Scikit-learn trong nghiên cứu này là hỗ trợ tiền xử lý dữ liệu (data preprocessing). Bộ dữ liệu sinh viên bao gồm nhiều loại đặc trưng khác nhau, do đó cần thực hiện các bước xử lý như chuẩn hóa dữ liệu (normalization), biến đổi thang đo (scaling), và mã hóa các biến phân loại (encoding). Scikit-learn cung cấp các công cụ hiệu quả như StandardScaler, MinMaxScaler và LabelEncoder, giúp đảm bảo dữ liệu đầu vào có dạng phù hợp cho quá trình huấn luyện mô hình, đồng thời cải thiện khả năng hội tụ và độ chính xác của các thuật toán.

Ngoài ra, Scikit-learn còn được sử dụng để phân chia tập dữ liệu thành các tập huấn luyện và kiểm tra (train-test split), đảm bảo việc đánh giá mô hình được thực hiện một cách khách quan. Việc phân chia dữ liệu hợp lý giúp tránh hiện tượng overfitting, đồng thời cung cấp cơ sở đáng tin cậy để so sánh hiệu quả giữa các mô hình khác nhau.

Một vai trò quan trọng khác của Scikit-learn là xây dựng các mô hình học máy cơ sở (baseline models) như Logistic Regression, Decision Tree hoặc Random Forest. Các mô hình này không chỉ giúp thiết lập một mức tham chiếu ban đầu về hiệu suất mà còn đóng vai trò so sánh với các mô hình học sâu như RNN và LSTM. Thông qua việc so sánh này, nghiên cứu có thể đánh giá rõ ràng mức độ cải thiện mà các phương pháp học sâu mang lại.

Bên cạnh đó, Scikit-learn cung cấp hệ thống các chỉ số đánh giá (evaluation metrics) toàn diện cho các bài toán phân loại, bao gồm độ chính xác (accuracy), độ chính xác dương (precision), độ bao phủ (recall) và F1-score. Những chỉ số này giúp đánh giá mô hình một cách đa chiều, đặc biệt trong bối cảnh dữ liệu có sự mất cân bằng giữa các lớp như bài toán dự đoán sinh viên bỏ học. Ngoài ra, các công cụ như confusion matrix cũng được sử dụng để phân tích chi tiết kết quả dự đoán, từ đó hiểu rõ hơn về نقاط mạnh và hạn chế của từng mô hình. Nhìn chung, Scikit-learn đóng vai trò như một nền tảng hỗ trợ quan trọng, kết nối giữa các bước tiền xử lý dữ liệu, xây dựng mô hình cơ sở và đánh giá hiệu suất. Việc sử dụng thư viện này không chỉ giúp chuẩn hóa quy trình nghiên cứu mà còn đảm bảo tính chính xác, minh bạch và khả năng tái lập của kết quả.

1.4. Các nghiên cứu liên quan

Trong những năm gần đây, bài toán dự đoán sinh viên bỏ học và kết quả học tập đã thu hút sự quan tâm rộng rãi của cộng đồng nghiên cứu, đặc biệt trong bối cảnh chuyển đổi số trong giáo dục và sự gia tăng của dữ liệu học tập. Nhiều nghiên cứu đã được thực hiện với các phương pháp khác nhau, từ các mô hình học máy truyền thống đến các mô hình học sâu tiên tiến, nhằm khai thác dữ liệu và nâng cao độ chính xác trong dự đoán.

Một trong những nghiên cứu tiêu biểu là của Realinho et al. (2022) với công trình "Predicting Student Dropout and Academic Success". Nghiên cứu này sử dụng một bộ dữ liệu toàn diện bao gồm các yếu tố học thuật, nhân khẩu học và kinh tế – xã hội để xây dựng các mô hình dự đoán trạng thái học tập của sinh viên. Kết quả cho thấy rằng việc kết hợp nhiều loại đặc trưng khác nhau giúp cải thiện đáng kể hiệu suất dự đoán. Đồng thời, nghiên cứu cũng nhấn mạnh vai trò quan trọng của dữ liệu đa chiều trong việc phản ánh toàn diện hành vi học tập. Đây chính là cơ sở dữ liệu và định hướng quan trọng được kế thừa và phát triển trong nghiên cứu này.

Bên cạnh đó, vấn đề mất cân bằng dữ liệu (data imbalance) – một đặc điểm phổ biến trong bài toán dự đoán bỏ học – cũng được nhiều nghiên cứu quan tâm. Mduma (2023) đã tập trung nghiên cứu các kỹ thuật cân bằng dữ liệu nhằm cải thiện hiệu quả của các mô hình học máy. Các phương pháp như oversampling, undersampling và SMOTE được áp dụng để xử lý sự chênh lệch giữa các lớp dữ liệu. Kết quả cho thấy việc cân bằng dữ liệu giúp cải thiện đáng kể khả năng nhận diện lớp thiểu số

(Dropout), từ đó nâng cao hiệu quả tổng thể của mô hình. Nghiên cứu này cung cấp một góc nhìn quan trọng trong việc xử lý dữ liệu đầu vào, đặc biệt khi xây dựng các mô hình phân loại trong môi trường thực tế.

Ngoài ra, các phương pháp học máy truyền thống cũng được áp dụng rộng rãi trong dự đoán kết quả học tập. Yan (2021) đã sử dụng thuật toán XGBoost để dự đoán hiệu suất học tập của sinh viên từ góc nhìn vĩ mô. XGBoost, với khả năng xử lý dữ liệu lớn và tối ưu hóa hiệu suất thông qua boosting, đã đạt được kết quả khả quan trong việc phát hiện các yếu tố ảnh hưởng đến kết quả học tập. Tuy nhiên, các mô hình dạng này thường gặp hạn chế trong việc xử lý dữ liệu chuỗi thời gian và khai thác các mối quan hệ phụ thuộc dài hạn, điều mà các mô hình học sâu có thể thực hiện tốt hơn.

Tổng hợp các nghiên cứu trên cho thấy rằng, mặc dù các phương pháp học máy truyền thống và các kỹ thuật xử lý dữ liệu đã đạt được nhiều kết quả tích cực, nhưng vẫn tồn tại một số hạn chế. Cụ thể, các mô hình truyền thống thường phụ thuộc nhiều vào quá trình lựa chọn đặc trưng thủ công và gặp khó khăn trong việc khai thác các mối quan hệ phi tuyến và phụ thuộc theo thời gian. Trong khi đó, các nghiên cứu về học sâu, đặc biệt là các mô hình xử lý dữ liệu chuỗi như RNN và LSTM, vẫn chưa được khai thác đầy đủ trong bối cảnh dữ liệu giáo dục đa chiều.

Do đó, nghiên cứu này được thực hiện nhằm kế thừa và phát triển các hướng tiếp cận hiện có, đồng thời đề xuất việc áp dụng các mô hình học sâu như RNN, LSTM và mô hình Hybrid để giải quyết bài toán dự đoán sinh viên bỏ học một cách hiệu quả hơn. Bằng cách kết hợp khả năng xử lý dữ liệu chuỗi và khai thác đa dạng đặc trưng, nghiên cứu kỳ vọng sẽ cải thiện độ chính xác dự đoán và cung cấp những hiểu biết sâu sắc hơn về hành vi học tập của sinh viên.

1.5. Kết luận chương

Chương 1 đã trình bày một cách hệ thống và toàn diện các cơ sở lý thuyết liên quan đến bài toán dự đoán tỷ lệ sinh viên bỏ học. Trước hết, chương đã làm rõ thực trạng và mức độ nghiêm trọng của hiện tượng bỏ học trong giáo dục, đồng thời phân tích các nguyên nhân chủ yếu từ nhiều góc độ khác nhau, bao gồm học thuật, kinh tế, tâm lý và môi trường xã hội. Những phân tích này không chỉ giúp nhận diện bản chất phức tạp của vấn đề mà còn khẳng định tính cần thiết của việc xây dựng các mô hình dự đoán nhằm hỗ trợ can thiệp sớm.

Tiếp theo, chương đã giới thiệu tổng quan về các phương pháp học máy và học sâu, làm rõ vai trò và ưu điểm của học sâu trong việc xử lý các bài toán dữ liệu phức tạp, đặc biệt là dữ liệu có tính chuỗi thời gian và đa dạng về đặc trưng. Các mô hình tiêu biểu như RNN, LSTM và mô hình Hybrid đã được trình bày chi tiết, từ nguyên lý hoạt động đến khả năng ứng dụng trong bài toán dự đoán hành vi học tập. Đồng thời, chương cũng đã phân tích lý do lựa chọn các mô hình này dựa trên đặc điểm của dữ liệu sinh viên, qua đó thiết lập nền tảng lý thuyết vững chắc cho các bước triển khai tiếp theo.

Bên cạnh đó, chương đã mô tả bộ dữ liệu sử dụng trong nghiên cứu cùng với các công cụ và thư viện hỗ trợ như TensorFlow, PyTorch và Scikit-learn, nhằm đảm bảo tính khả thi và hiệu quả trong quá trình xây dựng và đánh giá mô hình. Cuối cùng, việc tổng hợp và phân tích các nghiên cứu liên quan đã giúp xác định khoảng trống nghiên cứu, đồng thời làm rõ hướng tiếp cận và đóng góp của đề tài.

Nhìn chung, các nội dung trong chương này đã cung cấp nền tảng lý thuyết và thực tiễn cần thiết cho bài toán dự đoán sinh viên bỏ học. Trên cơ sở đó, Chương 2 sẽ tập trung vào việc xây dựng và triển khai cụ thể các mô hình học sâu, bao gồm các bước tiền xử lý dữ liệu, thiết kế kiến trúc mô hình và phương pháp đánh giá, nhằm kiểm chứng hiệu quả của các phương pháp đã đề xuất.

CHƯƠNG 2: XÂY DỰNG CÁC MÔ HÌNH HỌC SÂU

- Giới thiệu chương

Chương 2 trình bày quy trình xây dựng các mô hình học sâu phục vụ bài toán dự đoán nguy cơ sinh viên bỏ học, bắt đầu từ giai đoạn tiền xử lý dữ liệu nhằm đảm bảo dữ liệu đầu vào có chất lượng, nhất quán và phù hợp với yêu cầu của các mô hình. Trên cơ sở dữ liệu đã được làm sạch, chuyển đổi và phân chia, chương lần lượt trình bày quá trình xây dựng hai mô hình RNN và LSTM nhằm khai thác khả năng học các mối quan hệ tuần tự trong dữ liệu sinh viên. Bên cạnh đó, chương tập trung làm rõ mô hình Hybrid được đề xuất, từ lý do xây dựng, kiến trúc tổng thể, hàm mất mát, thuật toán tối ưu đến các ưu điểm và sự khác biệt so với các mô hình còn lại. Thông qua các nội dung này, chương cung cấp cơ sở thực nghiệm quan trọng cho việc triển khai, so sánh và đánh giá hiệu quả của các mô hình trong Chương 3.

2.1. Tiền xử lý dữ liệu

2.1.1. Mô tả dữ liệu đầu vào

Chất lượng và tính đại diện của dữ liệu đóng vai trò quyết định đối với hiệu quả của các mô hình học máy và học sâu. Trong bài toán dự đoán nguy cơ sinh viên bỏ học, dữ liệu không chỉ phản ánh kết quả học tập mà còn chứa đựng nhiều thông tin liên quan đến đặc điểm cá nhân, hoàn cảnh gia đình và môi trường kinh tế - xã hội của người học. Do đó, việc lựa chọn một bộ dữ liệu có tính toàn diện và độ tin cậy cao là điều kiện tiên quyết để xây dựng các mô hình dự đoán có khả năng khái quát hóa tốt và ứng dụng trong thực tế.

Trong nghiên cứu này, bộ dữ liệu "Predicting Student Dropout and Academic Success" được sử dụng làm nguồn dữ liệu đầu vào cho quá trình huấn luyện và đánh giá mô hình. Đây là bộ dữ liệu được công bố công khai và đã được sử dụng trong nhiều nghiên cứu về dự đoán kết quả học tập và nguy cơ bỏ học trong giáo dục đại học. Bộ dữ liệu bao gồm 4.424 bản ghi, trong đó mỗi bản ghi tương ứng với một sinh viên và được mô tả thông qua 35 thuộc tính đầu vào, phản ánh nhiều khía cạnh khác nhau của quá trình học tập và hoàn cảnh của người học.

Các thuộc tính trong bộ dữ liệu được xây dựng dựa trên nhiều nguồn thông tin khác nhau, bao gồm dữ liệu học tập, dữ liệu nhân khẩu học và dữ liệu kinh tế - xã

hội. Nhờ đó, bộ dữ liệu không chỉ cung cấp thông tin về kết quả học tập của sinh viên mà còn phản ánh các yếu tố bên ngoài có khả năng tác động đến quyết định tiếp tục hay từ bỏ chương trình đào tạo.

Biến mục tiêu (Target) của bài toán là trạng thái học tập cuối cùng của sinh viên. Trong nghiên cứu này, bài toán được xây dựng dưới dạng phân loại nhị phân với hai lớp chính là Graduate (sinh viên tốt nghiệp) và Dropout (sinh viên bỏ học). Theo thống kê của bộ dữ liệu, có 3.003 sinh viên thuộc nhóm tốt nghiệp và 1.421 sinh viên thuộc nhóm bỏ học, tương ứng với tỷ lệ khoảng 67,9% và 32,1%. Mặc dù tồn tại sự chênh lệch nhất định giữa hai lớp dữ liệu, mức độ mất cân bằng này không quá nghiêm trọng và vẫn đảm bảo điều kiện phù hợp để xây dựng các mô hình phân loại.

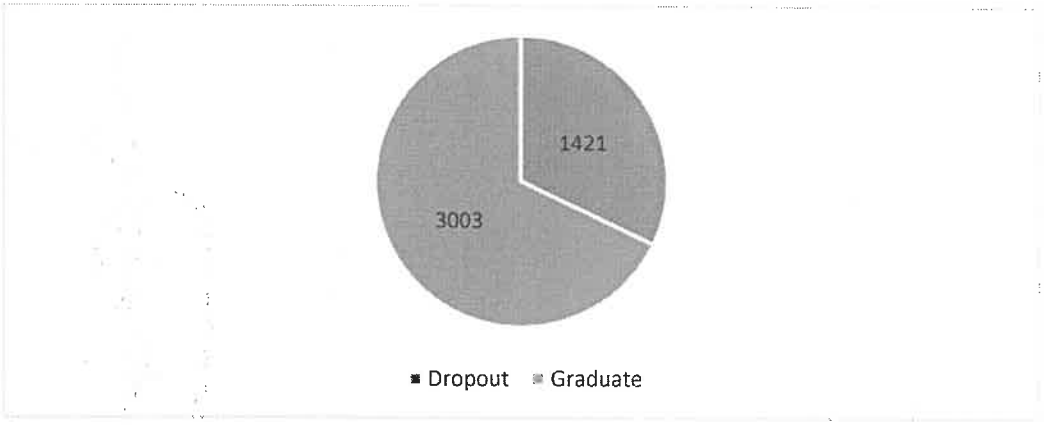


Figure 2.1. Sự phân bố của lớp mục tiêu Graduate và Dropout

Để thuận tiện cho việc phân tích và xây dựng mô hình, các thuộc tính đầu vào có thể được chia thành ba nhóm chính như sau:

Nhóm đặc trưng nhân khẩu học (Demographic Features)

Nhóm đặc trưng này phản ánh các thông tin cá nhân và hoàn cảnh gia đình của sinh viên, bao gồm các thuộc tính như tình trạng hôn nhân (Marital Status), quốc tịch (Nationality), trình độ học vấn của cha mẹ (Mother's Qualification, Father's Qualification), nghề nghiệp của cha mẹ (Mother's Occupation, Father's Occupation), giới tính (Gender), tình trạng học bổng (Scholarship Holder) và tình trạng sinh viên quốc tế (International). Những thông tin này giúp phản ánh điều kiện xã hội và nền tảng gia đình của người học, vốn được xem là những yếu tố có ảnh hưởng đáng kể đến khả năng duy trì học tập trong môi trường đại học.

Nhóm đặc trưng học thuật (Academic Features)

Đây là nhóm thuộc tính quan trọng nhất đối với bài toán dự đoán bỏ học, bởi chúng phản ánh trực tiếp quá trình học tập và kết quả đào tạo của sinh viên. Các thuộc tính tiêu biểu bao gồm điểm tuyển sinh đầu vào (Admission Grade), số tín chỉ đăng ký, số tín chỉ hoàn thành, số tín chỉ được công nhận và điểm trung bình các học phần trong từng học kỳ. Những thông tin này cho phép đánh giá mức độ tiến bộ trong học tập, khả năng thích nghi với môi trường đại học cũng như nguy cơ gặp khó khăn về học vụ của sinh viên.

Đặc biệt, các biến liên quan đến số tín chỉ tích lũy và kết quả học tập theo từng học kỳ mang tính kế thừa theo thời gian, phản ánh quá trình phát triển học tập liên tục của sinh viên. Đây cũng là cơ sở quan trọng để áp dụng các mô hình học sâu có khả năng khai thác thông tin theo chuỗi như RNN và LSTM.

Nhóm đặc trưng kinh tế - xã hội (Socio-economic Features)

Ngoài các yếu tố cá nhân và học thuật, bộ dữ liệu còn bao gồm một số chỉ số kinh tế vĩ mô như tỷ lệ thất nghiệp (Unemployment Rate), tỷ lệ lạm phát (Inflation Rate) và tổng sản phẩm quốc nội (GDP). Những biến này được bổ sung nhằm phản ánh bối cảnh kinh tế tại thời điểm sinh viên theo học. Trong nhiều nghiên cứu trước đây, các yếu tố kinh tế được chứng minh có ảnh hưởng gián tiếp đến khả năng duy trì học tập thông qua áp lực tài chính, cơ hội việc làm và điều kiện sống của sinh viên. Việc tích hợp các biến kinh tế - xã hội giúp mô hình có cái nhìn toàn diện hơn về các yếu tố ảnh hưởng đến quyết định bỏ học.

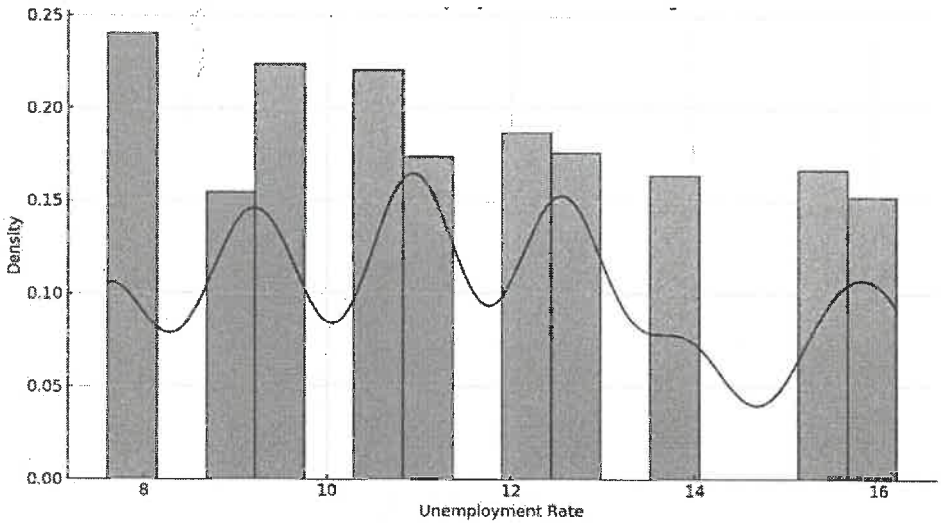


Figure 2.2. Tỷ lệ thất nghiệp

Nhìn chung, bộ dữ liệu được sử dụng trong nghiên cứu có quy mô tương đối lớn, bao phủ nhiều nhóm đặc trưng khác nhau và phản ánh tương đối đầy đủ các yếu tố có liên quan đến nguy cơ bỏ học của sinh viên. Sự kết hợp giữa các đặc trưng học thuật, nhân khẩu học và kinh tế - xã hội không chỉ giúp nâng cao khả năng dự đoán của mô hình mà còn tạo điều kiện để phân tích sâu hơn các nguyên nhân dẫn đến hiện tượng bỏ học trong giáo dục đại học. Đây cũng chính là cơ sở quan trọng cho các bước tiền xử lý dữ liệu và xây dựng mô hình học sâu được trình bày trong các phần tiếp theo của nghiên cứu.

2.1.2. Làm sạch và chuyển đổi dữ liệu

Sau khi thu thập dữ liệu, bước tiếp theo trong quá trình xây dựng mô hình là thực hiện tiền xử lý dữ liệu. Đây là một trong những giai đoạn quan trọng nhất của quy trình khai phá dữ liệu và học máy, bởi chất lượng của dữ liệu đầu vào có ảnh hưởng trực tiếp đến hiệu quả và độ chính xác của mô hình dự đoán. Trên thực tế, dữ liệu thu thập được thường tồn tại nhiều vấn đề như dữ liệu trùng lặp, dữ liệu thiếu, sự khác biệt về thang đo giữa các thuộc tính hoặc sự không đồng nhất trong cách biểu diễn dữ liệu. Nếu không được xử lý phù hợp, các vấn đề này có thể làm giảm khả năng học của mô hình và dẫn đến các kết quả dự đoán thiếu chính xác.

Trong nghiên cứu này, dữ liệu được xử lý thông qua một quy trình gồm các bước chính: loại bỏ dữ liệu trùng lặp, xử lý dữ liệu thiếu, chuyển đổi biến mục tiêu, phân loại thuộc tính và chuẩn hóa dữ liệu.

Loại bỏ dữ liệu trùng lặp

Trong các nghiên cứu về khai phá dữ liệu và học máy, dữ liệu trùng lặp được xem là một trong những nguyên nhân gây sai lệch phân bố dữ liệu và làm giảm khả năng tổng quát hóa của mô hình. Theo Han, Kamber và Pei [25], các bản ghi trùng lặp có thể tạo ra hiện tượng thiên lệch trong quá trình học, khiến mô hình đánh giá quá mức một số mẫu dữ liệu nhất định. Do đó, bước đầu tiên của quá trình tiền xử lý là kiểm tra và loại bỏ toàn bộ các bản ghi bị lặp hoàn toàn trên tất cả các thuộc tính.

Việc loại bỏ dữ liệu trùng lặp giúp đảm bảo mỗi đối tượng trong tập dữ liệu chỉ được biểu diễn duy nhất một lần, từ đó nâng cao độ tin cậy và tính nhất quán của dữ liệu đầu vào trước khi tiến hành huấn luyện mô hình.

Xử lý dữ liệu thiếu

Dữ liệu thiếu (Missing Values) là một hiện tượng phổ biến trong các hệ thống thông tin thực tế do nhiều nguyên nhân khác nhau như lỗi nhập liệu, lỗi thu thập dữ liệu hoặc thiếu thông tin từ người dùng. Theo Little và Rubin [26], dữ liệu thiếu nếu không được xử lý thích hợp có thể làm suy giảm hiệu quả dự đoán và dẫn đến kết quả phân tích sai lệch.

Trong nghiên cứu này, phương pháp Forward Fill (FFill) được áp dụng để xử lý các giá trị khuyết. Phương pháp này thay thế giá trị thiếu bằng giá trị hợp lệ gần nhất xuất hiện trước đó trong cùng thuộc tính. Đây là một kỹ thuật thường được sử dụng đối với các tập dữ liệu có tính liên tục hoặc có sự kế thừa giữa các quan sát [27].

Việc xử lý dữ liệu thiếu giúp đảm bảo tính toàn vẹn của tập dữ liệu, đồng thời tạo điều kiện thuận lợi cho quá trình huấn luyện các mô hình học sâu ở các bước tiếp theo.

Chuyển đổi biến mục tiêu

Trong các bài toán phân loại sử dụng học máy và học sâu, dữ liệu đầu ra cần được biểu diễn dưới dạng số để mô hình có thể thực hiện các phép tính toán học. Theo Géron [28], quá trình mã hóa nhãn (Label Encoding) là một bước tiền xử lý cơ bản nhằm chuyển đổi các nhãn dạng văn bản thành các giá trị số tương ứng.

Biến mục tiêu của bộ dữ liệu ban đầu bao gồm hai trạng thái là Graduate và Dropout. Để phù hợp với bài toán phân loại nhị phân, nghiên cứu thực hiện ánh xạ:

Dropout $\rightarrow$ 0

Graduate $\rightarrow$ 1

Việc chuyển đổi này không chỉ giúp mô hình xử lý dữ liệu hiệu quả hơn mà còn thuận tiện cho việc tính toán các chỉ số đánh giá như Accuracy, Precision, Recall và F1-score trong các giai đoạn tiếp theo.

Phân loại thuộc tính đầu vào

Theo Goodfellow et al. [29], việc nhận diện đúng bản chất của các thuộc tính đầu vào là điều kiện quan trọng để lựa chọn phương pháp biểu diễn dữ liệu phù hợp. Trong nghiên cứu này, các thuộc tính được chia thành hai nhóm chính:

Thuộc tính phân loại (Categorical Features)

Thuộc tính số (Numerical Features)

Việc phân loại này có ý nghĩa đặc biệt quan trọng đối với mô hình Hybrid được đề xuất trong nghiên cứu. Các đặc trưng số sẽ được xử lý trực tiếp thông qua mạng nơ-ron sâu, trong khi các đặc trưng phân loại sẽ được biểu diễn bằng kỹ thuật Embedding nhằm khai thác tốt hơn mối quan hệ ngữ nghĩa giữa các giá trị.

Chuẩn hóa dữ liệu

Các thuộc tính số trong bộ dữ liệu có đơn vị đo lường và miền giá trị khác nhau. Nếu sử dụng trực tiếp các giá trị này trong quá trình huấn luyện, các thuộc tính có giá trị lớn có thể chi phối hàm mất mát và ảnh hưởng đến quá trình tối ưu hóa mô hình.

Theo Bishop [30], chuẩn hóa dữ liệu là một bước tiền xử lý quan trọng giúp đưa các thuộc tính về cùng một thang đo, từ đó cải thiện tốc độ hội tụ và độ ổn định của các thuật toán học máy.

Trong nghiên cứu này, phương pháp Standardization được thực hiện thông qua thuật toán StandardScaler với công thức:

$$z = \frac{(x - \mu)}{\sigma}$$

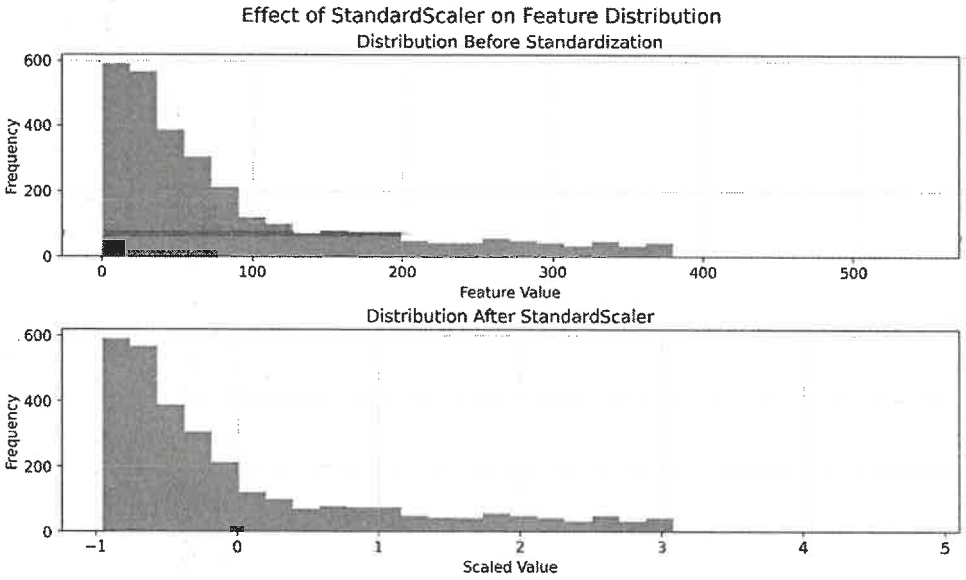


Figure 2.3. Sự ảnh hưởng của StandardScaler

Trong đó:

- x là giá trị ban đầu;

- μ là giá trị trung bình;
- σ là độ lệch chuẩn.

Sau khi chuẩn hóa, các thuộc tính có giá trị trung bình xấp xỉ 0 và độ lệch chuẩn xấp xỉ 1, giúp tăng hiệu quả học tập của các mô hình RNN, LSTM và Hybrid.

Loại bỏ thuộc tính không mang thông tin

Theo Kuhn và Johnson [31], các thuộc tính có phương sai bằng 0 hoặc không có sự biến thiên giữa các quan sát không mang lại giá trị cho quá trình học của mô hình. Những thuộc tính này không đóng góp vào khả năng phân biệt giữa các lớp dữ liệu và thậm chí có thể làm gia tăng độ phức tạp tính toán.

Do đó, nghiên cứu tiến hành loại bỏ các thuộc tính có toàn bộ giá trị bằng 0 sau quá trình chuẩn hóa. Bước xử lý này giúp giảm số chiều dữ liệu, tối ưu hóa tài nguyên tính toán và nâng cao hiệu suất của mô hình trong quá trình huấn luyện.

2.1.3. Phân chia tập dữ liệu

Sau khi hoàn thành các bước làm sạch và chuyển đổi dữ liệu, tập dữ liệu đã sẵn sàng để đưa vào quá trình huấn luyện và đánh giá mô hình. Tuy nhiên, để đảm bảo tính khách quan và khả năng tổng quát hóa của mô hình, dữ liệu không được sử dụng toàn bộ cho quá trình huấn luyện mà cần được phân chia thành các tập dữ liệu độc lập phục vụ các mục đích khác nhau.

Trong lĩnh vực học máy và học sâu, việc phân chia dữ liệu là một bước quan trọng nhằm đánh giá khả năng dự đoán của mô hình trên các dữ liệu chưa từng xuất hiện trong quá trình huấn luyện. Theo Géron [28], nếu mô hình được đánh giá trên chính tập dữ liệu đã sử dụng để huấn luyện, kết quả thu được có thể phản ánh quá mức khả năng ghi nhớ dữ liệu thay vì khả năng học được các quy luật tổng quát. Điều này dẫn đến hiện tượng quá khớp (overfitting), trong đó mô hình đạt hiệu suất rất cao trên dữ liệu huấn luyện nhưng hoạt động kém khi áp dụng trên dữ liệu mới.

Để khắc phục vấn đề này, nghiên cứu thực hiện phân chia tập dữ liệu thành hai phần riêng biệt gồm tập huấn luyện (Training Set) và tập kiểm tra (Testing Set). Trong đó, 80% số lượng mẫu dữ liệu được sử dụng cho quá trình huấn luyện mô hình và 20% còn lại được sử dụng để đánh giá hiệu suất dự đoán. Tỷ lệ phân chia 80/20 là một trong những tỷ lệ được sử dụng phổ biến trong nhiều nghiên cứu học máy do

đảm bảo được sự cân bằng giữa lượng dữ liệu phục vụ học tập và lượng dữ liệu phục vụ đánh giá mô hình [28], [30].

Việc phân chia dữ liệu được thực hiện ngẫu nhiên nhằm giảm thiểu các sai lệch có thể phát sinh từ thứ tự sắp xếp ban đầu của dữ liệu. Tuy nhiên, do bộ dữ liệu có sự chênh lệch nhất định giữa hai lớp Graduate và Dropout, nghiên cứu áp dụng phương pháp Stratified Sampling (lấy mẫu phân tầng) để duy trì tỷ lệ phân bố của các lớp trong cả tập huấn luyện và tập kiểm tra.

Cụ thể, theo thống kê ban đầu của bộ dữ liệu, nhóm sinh viên tốt nghiệp (Graduate) chiếm khoảng 67,9%, trong khi nhóm sinh viên bỏ học (Dropout) chiếm khoảng 32,1%. Nếu quá trình chia dữ liệu được thực hiện hoàn toàn ngẫu nhiên, có thể xảy ra trường hợp một trong hai tập dữ liệu chứa tỷ lệ các lớp quá khác biệt so với tập dữ liệu gốc, làm ảnh hưởng đến quá trình huấn luyện và đánh giá mô hình. Phương pháp Stratified Sampling giúp đảm bảo rằng tỷ lệ giữa các lớp trong tập huấn luyện và tập kiểm tra vẫn được duy trì gần tương đương với tỷ lệ ban đầu của toàn bộ tập dữ liệu.

Việc sử dụng lấy mẫu phân tầng đặc biệt quan trọng đối với các bài toán phân loại trong giáo dục, nơi số lượng sinh viên bỏ học thường thấp hơn số lượng sinh viên tốt nghiệp. Nhiều nghiên cứu đã chỉ ra rằng việc duy trì phân bố lớp trong quá trình chia dữ liệu giúp cải thiện độ ổn định của mô hình và tăng độ tin cậy của các chỉ số đánh giá [5], [30].

Sau quá trình phân chia, tập huấn luyện được sử dụng để xây dựng và tối ưu hóa các mô hình RNN, LSTM và Hybrid. Trong khi đó, tập kiểm tra hoàn toàn không tham gia vào quá trình học của mô hình mà chỉ được sử dụng ở giai đoạn cuối nhằm đánh giá khả năng dự đoán trên dữ liệu chưa từng được quan sát trước đó. Cách tiếp cận này giúp đảm bảo rằng các kết quả thực nghiệm phản ánh đúng năng lực tổng quát hóa của mô hình, đồng thời tạo cơ sở khách quan để so sánh hiệu quả giữa các phương pháp được đề xuất trong nghiên cứu.

Như vậy, việc phân chia dữ liệu theo tỷ lệ 80/20 kết hợp với kỹ thuật Stratified Sampling không chỉ giúp tối ưu hóa quá trình huấn luyện mà còn góp phần nâng cao độ tin cậy của kết quả đánh giá, tạo nền tảng vững chắc cho các thực nghiệm được trình bày trong các phần tiếp theo.

2.2. Xây dựng mô hình RNN

Sau khi hoàn thành quá trình tiền xử lý dữ liệu, nghiên cứu tiến hành xây dựng mô hình Recurrent Neural Network (RNN) nhằm khai thác khả năng học các mối quan hệ phụ thuộc trong dữ liệu đầu vào. RNN là một trong những kiến trúc mạng nơ-ron được thiết kế chuyên biệt cho dữ liệu có tính tuần tự, cho phép mô hình lưu giữ thông tin từ các trạng thái trước đó thông qua cơ chế trạng thái ẩn (hidden state) [15], [29]. Nhờ đặc điểm này, RNN có khả năng mô hình hóa các mối quan hệ giữa các đặc trưng không chỉ dựa trên giá trị hiện tại mà còn dựa trên thông tin đã xuất hiện trước đó trong chuỗi dữ liệu.

Trong bài toán dự đoán sinh viên bỏ học, mặc dù dữ liệu được lưu trữ dưới dạng các thuộc tính độc lập, nhưng bản chất của quá trình học tập lại mang tính liên tục và tích lũy theo thời gian. Các yếu tố như điểm số, số tín chỉ tích lũy, số tín chỉ đạt được hoặc tình trạng học vụ ở các giai đoạn khác nhau đều có mối liên hệ chặt chẽ với nhau và cùng tác động đến khả năng tiếp tục học tập của sinh viên. Do đó, việc sử dụng RNN cho phép mô hình học được các mẫu quan hệ tiềm ẩn giữa các đặc trưng này, thay vì chỉ xem xét từng biến một cách độc lập như nhiều thuật toán học máy truyền thống.

Kiến trúc mô hình

Mô hình RNN được thiết kế theo kiến trúc nhiều tầng (multi-layer architecture) nhằm tăng khả năng học biểu diễn đặc trưng ở các mức độ khác nhau. Thay vì chỉ sử dụng một lớp RNN đơn lẻ, nghiên cứu xây dựng mô hình với hai lớp RNN liên tiếp để cho phép mạng học được các mối quan hệ tuần tự từ mức tổng quát đến mức trừu tượng hơn. Cách thiết kế này phù hợp với đặc điểm của dữ liệu sinh viên, trong đó các thuộc tính đầu vào không chỉ mang thông tin riêng lẻ mà còn có thể tồn tại các mối quan hệ phụ thuộc lẫn nhau trong quá trình học tập.

Cấu trúc tổng thể của mô hình bao gồm hai lớp RNN, các lớp Dropout đóng vai trò điều chuẩn hóa và các lớp Dense phục vụ cho quá trình phân loại cuối cùng. Lớp RNN đầu tiên được xây dựng với 128 đơn vị ẩn nhằm mở rộng khả năng học biểu diễn từ dữ liệu đầu vào. Với số lượng đơn vị ẩn lớn hơn, lớp này có thể tiếp nhận và xử lý nhiều thông tin hơn từ các đặc trưng ban đầu, qua đó học được các mẫu quan hệ tổng quát giữa các yếu tố như kết quả học tập, tín chỉ, thông tin cá nhân và các yếu tố kinh tế - xã hội.

Kết quả đầu ra từ lớp RNN thứ nhất tiếp tục được truyền sang lớp RNN thứ hai. Lớp này gồm 64 đơn vị ẩn, có vai trò tinh lọc thông tin và học các biểu diễn sâu hơn từ những đặc trưng đã được trích xuất ở tầng trước. Việc giảm số lượng đơn vị ẩn từ 128 xuống 64 giúp mô hình từng bước nén thông tin, loại bỏ các nhiễu không cần thiết và tập trung vào những đặc trưng có ý nghĩa hơn đối với nhiệm vụ dự đoán. Đây là cách thiết kế thường được sử dụng trong các mạng học sâu nhằm cân bằng giữa năng lực biểu diễn và khả năng tổng quát hóa của mô hình.

Trong kiến trúc này, lớp RNN đầu tiên được thiết lập để trả về toàn bộ chuỗi đầu ra, cho phép lớp RNN tiếp theo tiếp tục khai thác thông tin tuần tự qua các bước trong chuỗi. Lớp RNN thứ hai chỉ trả về trạng thái cuối cùng, đây được xem là biểu diễn tổng hợp của toàn bộ chuỗi đầu vào. Biểu diễn này sau đó được đưa vào các lớp kết nối đầy đủ để thực hiện quá trình phân loại.

Sau các lớp RNN, nghiên cứu sử dụng một lớp Dense trung gian gồm 16 nút với hàm kích hoạt tanh. Lớp này đóng vai trò tổng hợp và chuyển đổi các biểu diễn tuần tự đã học được thành một không gian đặc trưng nhỏ gọn hơn, phù hợp cho bước dự đoán cuối cùng. Việc sử dụng lớp Dense trung gian giúp mô hình có thêm khả năng học các tương tác phi tuyến giữa các đặc trưng đã được trích xuất, thay vì đưa trực tiếp đầu ra của RNN vào lớp phân loại.

Cuối cùng, lớp đầu ra sử dụng hàm kích hoạt Softmax nhằm chuyển đổi kết quả của mô hình thành phân bố xác suất trên các lớp dự đoán. Nhờ đó, mô hình không chỉ đưa ra nhãn dự đoán cuối cùng mà còn cung cấp mức độ tin cậy tương ứng cho từng lớp. Điều này đặc biệt hữu ích trong bài toán dự đoán nguy cơ bỏ học, bởi xác suất dự đoán có thể được sử dụng như một chỉ báo hỗ trợ nhà trường nhận diện mức độ rủi ro của từng sinh viên.

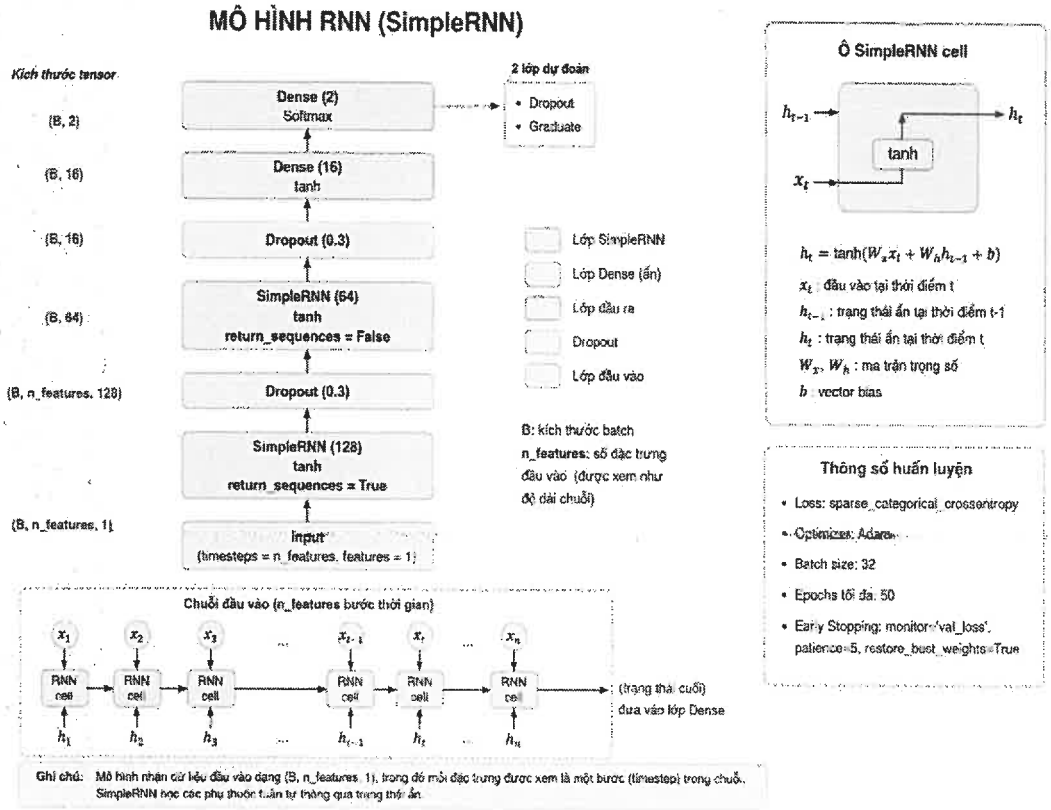


Figure 2.4. Xây dựng mô hình RNN

Lựa chọn hàm kích hoạt

Trong các lớp RNN, nghiên cứu sử dụng hàm kích hoạt (tanh). Đây là hàm kích hoạt được sử dụng phổ biến trong các kiến trúc RNN truyền thống do có khả năng chuẩn hóa đầu ra trong khoảng từ -1 đến 1, giúp quá trình lan truyền thông tin giữa các trạng thái ẩn ổn định hơn [29].

So với hàm sigmoid, hàm (tanh) có giá trị trung bình gần bằng 0, từ đó giúp quá trình tối ưu hóa diễn ra hiệu quả hơn và giảm hiện tượng dịch chuyển phân phối dữ liệu trong mạng. Chính vì vậy, (tanh) được lựa chọn thay vì các hàm kích hoạt khác cho các lớp RNN trong nghiên cứu này.

Điều chuẩn hóa mô hình bằng Dropout

Một trong những thách thức thường gặp khi xây dựng các mô hình học sâu là hiện tượng quá khớp (overfitting), xảy ra khi mô hình học quá tốt trên dữ liệu huấn luyện nhưng lại suy giảm hiệu suất trên dữ liệu mới. Để giảm thiểu vấn đề này, nghiên cứu áp dụng kỹ thuật Dropout giữa các lớp mạng.

Dropout hoạt động bằng cách vô hiệu hóa ngẫu nhiên một tỷ lệ nhất định các nơ-ron trong quá trình huấn luyện, buộc mô hình phải học các biểu diễn dữ liệu mang tính tổng quát hơn thay vì phụ thuộc vào một nhóm đặc trưng cụ thể [29]. Trong nghiên cứu này, tỷ lệ Dropout được lựa chọn là 0,3. Đây là mức giá trị được sử dụng phổ biến trong nhiều nghiên cứu học sâu và được đánh giá là phù hợp để cân bằng giữa khả năng học tập của mô hình và khả năng chống quá khớp.

Thuật toán tối ưu

Để cập nhật trọng số trong quá trình huấn luyện, nghiên cứu sử dụng thuật toán tối ưu Adam (Adaptive Moment Estimation). Adam là một trong những thuật toán tối ưu được sử dụng rộng rãi trong các mô hình học sâu nhờ khả năng tự động điều chỉnh tốc độ học cho từng tham số dựa trên các moment bậc nhất và bậc hai của gradient [28]. Cụ thể, thuật toán này không chỉ xét hướng cập nhật hiện tại của gradient mà còn tận dụng thông tin từ các lần cập nhật trước đó, từ đó giúp quá trình tối ưu diễn ra ổn định hơn.

So với các phương pháp tối ưu truyền thống như Gradient Descent hoặc Stochastic Gradient Descent (SGD), Adam thường cho tốc độ hội tụ nhanh hơn và ít nhạy cảm hơn với việc lựa chọn tốc độ học ban đầu. Điều này đặc biệt quan trọng đối với các mô hình học sâu như RNN, nơi quá trình lan truyền ngược qua thời gian có thể làm gradient thay đổi phức tạp. Nhờ khả năng thích nghi theo từng tham số, Adam giúp hạn chế dao động trong quá trình cập nhật trọng số và hỗ trợ mô hình đạt trạng thái hội tụ hiệu quả hơn.

Trong nghiên cứu này, Adam được lựa chọn làm bộ tối ưu chính cho quá trình huấn luyện mô hình RNN vì mô hình có nhiều tham số cần học và dữ liệu đầu vào có nhiều đặc trưng khác nhau. Việc sử dụng Adam giúp mô hình khai thác tốt hơn các mẫu dữ liệu tiềm ẩn, đồng thời giảm thời gian huấn luyện so với các phương pháp tối ưu cơ bản. Nhờ đó, quá trình huấn luyện mô hình RNN đạt được sự cân bằng giữa hiệu quả tính toán, tốc độ hội tụ và khả năng ổn định trong quá trình học..

Thiết lập quá trình huấn luyện

Quá trình huấn luyện được thực hiện theo phương pháp mini-batch learning với kích thước batch là 32. Việc sử dụng mini-batch giúp tận dụng ưu điểm của cả hai phương pháp học toàn bộ dữ liệu và học từng mẫu đơn lẻ, đồng thời nâng cao hiệu quả tính toán trên phần cứng hiện đại.

Số lượng epoch tối đa được thiết lập là 50. Tuy nhiên, thay vì huấn luyện đủ số epoch cố định, nghiên cứu sử dụng cơ chế Early Stopping để tự động dừng quá trình học khi hiệu suất trên tập xác thực không còn được cải thiện. Cụ thể, nếu giá trị hàm mất mát trên tập xác thực không giảm sau nhiều vòng lặp liên tiếp, quá trình huấn luyện sẽ được kết thúc và mô hình tốt nhất sẽ được giữ lại.Cách tiếp cận này giúp giảm thời gian huấn luyện, đồng thời hạn chế hiện tượng quá khớp thường gặp trong các mô hình học sâu.

Nhìn chung, mô hình RNN được xây dựng trong nghiên cứu bao gồm nhiều lớp xử lý tuần tự kết hợp với các kỹ thuật điều chuẩn hóa và tối ưu hiện đại. Việc lựa chọn kiến trúc nhiều tầng, sử dụng hàm kích hoạt (tanh), kỹ thuật Dropout, thuật toán Adam và cơ chế Early Stopping được thực hiện dựa trên các nghiên cứu trước đây cũng như đặc điểm của tập dữ liệu sinh viên [15], [28], [29].

2.3. Xây dựng mô hình LSTM

Sau khi xây dựng mô hình RNN cơ sở, nghiên cứu tiếp tục triển khai mô hình Long Short-Term Memory (LSTM) nhằm khắc phục các hạn chế của RNN truyền thống trong việc học các phụ thuộc dài hạn. LSTM là một biến thể nâng cao của RNN, được thiết kế với cơ chế bộ nhớ và các cổng điều khiển thông tin, cho phép mô hình quyết định thông tin nào cần được ghi nhớ, cập nhật hoặc loại bỏ trong quá trình xử lý chuỗi dữ liệu [23]. Nhờ đó, LSTM có khả năng duy trì thông tin quan trọng qua nhiều bước thời gian, đồng thời giảm ảnh hưởng của hiện tượng tiêu biến gradient vốn thường gặp ở RNN.

Trong bài toán dự đoán sinh viên bỏ học, các yếu tố ảnh hưởng đến quyết định rời bỏ chương trình học thường không xuất hiện một cách tức thời, mà được hình thành thông qua quá trình tích lũy trong suốt quá trình học tập. Chẳng hạn, kết quả học tập thấp trong một học kỳ, số tín chỉ không đạt, tình trạng học phí hoặc các yếu tố kinh tế - xã hội có thể chưa lập tức dẫn đến bỏ học, nhưng khi kết hợp và kéo dài theo thời gian, chúng có thể trở thành dấu hiệu cảnh báo rõ ràng. Do đó, mô hình LSTM được lựa chọn nhằm tăng khả năng học các mối quan hệ dài hạn và các xu hướng tiềm ẩn trong dữ liệu sinh viên.

Kiến trúc mô hình

Mô hình LSTM trong nghiên cứu được thiết kế theo kiến trúc nhiều tầng, bao gồm hai lớp LSTM liên tiếp, các lớp Dropout để hạn chế quá khớp, một lớp Dense

trung gian và lớp đầu ra phục vụ nhiệm vụ phân loại. Cách thiết kế này tương đồng với mô hình RNN cơ sở nhằm đảm bảo tính công bằng khi so sánh hiệu suất giữa các mô hình, đồng thời cho phép đánh giá rõ ràng mức độ cải thiện mà LSTM mang lại so với RNN truyền thống.

Lớp LSTM thứ nhất gồm 128 đơn vị ẩn, có nhiệm vụ học các biểu diễn tổng quát từ dữ liệu đầu vào. Tham số return_sequences=True được sử dụng để lớp này trả về toàn bộ chuỗi đầu ra, tạo điều kiện cho lớp LSTM tiếp theo tiếp tục khai thác các quan hệ tuần tự đã được học ở tầng trước. Lớp LSTM thứ hai gồm 64 đơn vị ẩn, đóng vai trò tinh lọc và tổng hợp thông tin chuỗi, sau đó chỉ trạng thái cuối cùng được truyền sang các lớp phân loại phía sau.

Việc lựa chọn số lượng đơn vị ẩn theo hướng giảm dần từ 128 xuống 64 nhằm giúp mô hình trước hết học được không gian biểu diễn đủ rộng ở tầng đầu, sau đó nén và chọn lọc các thông tin quan trọng ở tầng sau. Cách thiết kế này giúp giảm dần độ phức tạp của biểu diễn, hạn chế số lượng tham số không cần thiết và tăng khả năng tổng quát hóa của mô hình.

Sau hai lớp LSTM, mô hình sử dụng một lớp Dense gồm 16 nút với hàm kích hoạt (tanh). Lớp này có nhiệm vụ kết hợp các đặc trưng đã được trích xuất từ chuỗi đầu vào và chuyển chúng thành biểu diễn phù hợp cho bước phân loại cuối cùng. Lớp đầu ra sử dụng hàm kích hoạt Softmax để biểu diễn xác suất dự đoán cho từng lớp trạng thái học tập của sinh viên.

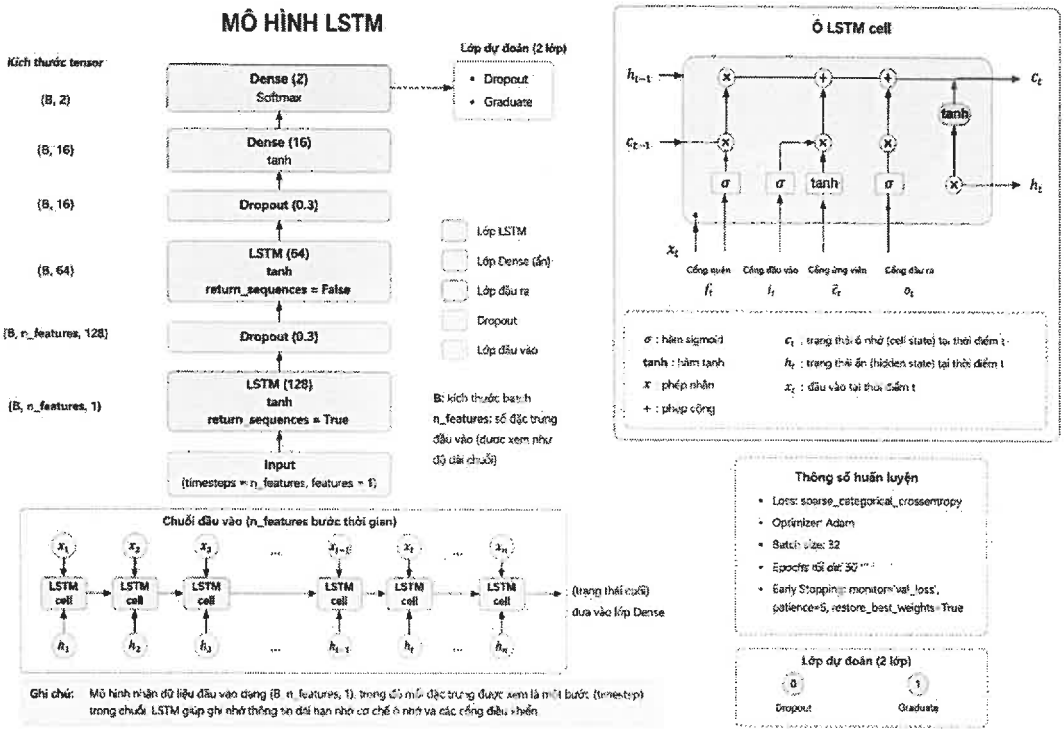


Figure 2.5. Xây dựng mô hình LSTM

Lựa chọn hàm kích hoạt

Trong các lớp LSTM, hàm kích hoạt (tanh) được sử dụng nhằm điều chỉnh giá trị đầu ra của trạng thái ẩn trong khoảng từ -1 đến 1. Đây là lựa chọn phổ biến trong các mô hình LSTM, bởi hàm (tanh) giúp duy trì sự ổn định của giá trị trạng thái khi thông tin được truyền qua nhiều bước thời gian [23], [29].

Bên cạnh đó, bản thân kiến trúc LSTM đã sử dụng các hàm sigmoid trong các cổng điều khiển như cổng quên, cổng đầu vào và cổng đầu ra. Sự kết hợp giữa sigmoid và (tanh) cho phép LSTM vừa kiểm soát lượng thông tin đi qua các cổng, vừa biểu diễn nội dung bộ nhớ dưới dạng giá trị liên tục ổn định. Nhờ vậy, mô hình có khả năng duy trì các thông tin quan trọng trong chuỗi và loại bỏ những thông tin ít liên quan.

Điều chuẩn hóa bằng Dropout

Tương tự như mô hình RNN, mô hình LSTM cũng có nguy cơ gặp hiện tượng quá khớp do số lượng tham số lớn và khả năng biểu diễn mạnh. Để hạn chế vấn đề này, nghiên cứu sử dụng kỹ thuật Dropout sau mỗi lớp LSTM với tỷ lệ 0,3. Việc áp

dụng Dropout giúp giảm sự phụ thuộc của mô hình vào một số đơn vị nơ-ron nhất định, buộc mạng phải học các biểu diễn có tính tổng quát hơn.

Tỷ lệ Dropout 0,3 được lựa chọn nhằm đạt được sự cân bằng giữa hai mục tiêu: duy trì đủ năng lực học của mô hình và giảm nguy cơ ghi nhớ quá mức dữ liệu huấn luyện. Nếu tỷ lệ Dropout quá thấp, khả năng chống quá khớp có thể không rõ rệt; ngược lại, nếu tỷ lệ quá cao, mô hình có thể mất quá nhiều thông tin trong quá trình học và làm giảm hiệu suất dự đoán. Do đó, mức 0,3 được xem là phù hợp trong bối cảnh dữ liệu có quy mô vừa và mô hình có nhiều tầng học sâu.

Hàm mất mát và thuật toán tối ưu

Mô hình được huấn luyện với hàm mất mát sparse_categorical_crossentropy, đây là hàm mất mát phù hợp cho các bài toán phân loại khi nhãn đầu ra được biểu diễn dưới dạng số nguyên. Trong bài toán dự đoán trạng thái học tập của sinh viên, mỗi mẫu dữ liệu được gán một nhãn tương ứng với lớp đầu ra, chẳng hạn như Dropout hoặc Graduate. Hàm mất mát này có nhiệm vụ đo lường mức độ sai khác giữa phân bố xác suất do mô hình dự đoán và nhãn thực tế của dữ liệu. Nói cách khác, nếu mô hình gán xác suất cao cho lớp đúng thì giá trị mất mát sẽ thấp; ngược lại, nếu mô hình dự đoán sai hoặc thiếu chắc chắn đối với lớp đúng thì giá trị mất mát sẽ tăng lên.

Việc sử dụng sparse_categorical_crossentropy giúp đơn giản hóa quá trình huấn luyện vì không yêu cầu chuyển đổi nhãn sang dạng one-hot encoding. Điều này đặc biệt hữu ích khi làm việc với tập dữ liệu có nhiều mẫu, giúp giảm chi phí bộ nhớ và giữ cho quy trình xử lý dữ liệu gọn hơn. Hàm mất mát này cũng phù hợp với lớp đầu ra sử dụng hàm kích hoạt Softmax, bởi Softmax tạo ra phân bố xác suất trên các lớp dự đoán, còn sparse_categorical_crossentropy đánh giá trực tiếp độ lệch giữa phân bố xác suất đó và nhãn thực tế.

Để tối ưu hóa trọng số của mô hình, nghiên cứu sử dụng thuật toán Adam (Adaptive Moment Estimation). Adam là một trong những thuật toán tối ưu phổ biến trong học sâu nhờ khả năng kết hợp ưu điểm của Momentum và RMSProp. Thuật toán này không sử dụng một tốc độ học cố định cho tất cả các tham số, mà tự động điều chỉnh tốc độ học riêng cho từng tham số dựa trên thống kê của gradient trong các bước cập nhật trước đó. Nhờ vậy, Adam giúp quá trình huấn luyện diễn ra ổn định hơn, đặc biệt trong các mô hình có số lượng tham số lớn và dữ liệu đầu vào có nhiều chiều [28].

Đối với mô hình LSTM, quá trình lan truyền ngược qua thời gian thường phức tạp hơn so với các mạng truyền thẳng thông thường do mô hình phải học và cập nhật thông tin qua nhiều bước trong chuỗi. Vì vậy, việc lựa chọn một thuật toán tối ưu có khả năng thích nghi tốt như Adam là cần thiết để giảm dao động trong quá trình cập nhật trọng số và cải thiện tốc độ hội tụ. Ngoài ra, Adam cũng giúp mô hình duy trì hiệu quả học tập trong trường hợp gradient có sự thay đổi không đồng đều giữa các tham số, điều thường xuất hiện trong các kiến trúc mạng hồi tiếp và mạng LSTM.

Thiết lập quá trình huấn luyện

Quá trình huấn luyện mô hình LSTM được thiết lập tương tự mô hình RNN nhằm đảm bảo điều kiện so sánh công bằng giữa các kiến trúc. Tập dữ liệu sau tiền xử lý được chia thành tập huấn luyện và tập kiểm tra theo tỷ lệ 80/20, đồng thời dữ liệu đầu vào được định dạng lại theo cấu trúc ba chiều để phù hợp với yêu cầu của LSTM.

Mô hình được huấn luyện tối đa trong 50 epoch với kích thước batch là 32. Batch size 32 được lựa chọn vì đây là cấu hình phổ biến trong các thực nghiệm học sâu, giúp cân bằng giữa độ ổn định của quá trình cập nhật gradient và hiệu quả sử dụng bộ nhớ tính toán. Trong khi đó, số epoch tối đa 50 cho phép mô hình có đủ số vòng lặp để học các mẫu dữ liệu quan trọng, nhưng vẫn được kiểm soát thông qua cơ chế dừng sớm.

Để hạn chế hiện tượng quá khớp, nghiên cứu áp dụng Early Stopping với tiêu chí theo dõi là giá trị mất mát trên tập xác thực. Nếu giá trị này không cải thiện sau 5 epoch liên tiếp, quá trình huấn luyện sẽ được dừng lại và mô hình khôi phục bộ trọng số tốt nhất đã đạt được. Cách tiếp cận này giúp mô hình không tiếp tục học khi hiệu suất trên dữ liệu xác thực đã bão hòa, qua đó cải thiện khả năng tổng quát hóa trên dữ liệu mới.

2.4. Xây dựng mô hình Hybrid

2.4.1. Lý do đề xuất mô hình Hybrid

Mặc dù các mô hình học sâu như RNN và LSTM đã chứng minh được hiệu quả trong nhiều bài toán dự đoán và phân loại, tuy nhiên các mô hình này chủ yếu được thiết kế để xử lý dữ liệu tuần tự hoặc dữ liệu có tính phụ thuộc theo thời gian. Trong

khi đó, bộ dữ liệu sử dụng trong nghiên cứu về dự đoán sinh viên bỏ học lại có đặc điểm khá đặc biệt khi đồng thời chứa nhiều loại đặc trưng khác nhau.

Cụ thể, tập dữ liệu bao gồm các thuộc tính số như điểm tuyển sinh, số tín chỉ tích lũy, số tín chỉ hoàn thành, điểm trung bình học tập và các chỉ số kinh tế vĩ mô; đồng thời cũng chứa nhiều thuộc tính phân loại như giới tính, quốc tịch, trình độ học vấn của cha mẹ, tình trạng học bổng và loại hình tuyển sinh. Hai nhóm dữ liệu này có bản chất hoàn toàn khác nhau và mang những thông tin bổ sung lẫn nhau trong quá trình mô hình hóa hành vi học tập của sinh viên.

Trong các phương pháp xử lý truyền thống, toàn bộ dữ liệu thường được chuyển đổi về cùng một định dạng và đưa vào mô hình như một khối thống nhất. Tuy nhiên, cách tiếp cận này có thể làm mất đi nhiều thông tin ngữ nghĩa quan trọng, đặc biệt đối với các thuộc tính phân loại có số lượng giá trị lớn. Chẳng hạn, việc mã hóa one-hot cho các thuộc tính như ngành học, trình độ học vấn của cha mẹ hoặc hình thức tuyển sinh có thể làm gia tăng số chiều dữ liệu nhưng không phản ánh được mối quan hệ tiềm ẩn giữa các giá trị.

Bên cạnh đó, các đặc trưng số và đặc trưng phân loại thường có mức độ ảnh hưởng khác nhau đối với nguy cơ bỏ học. Một số nghiên cứu đã chỉ ra rằng các yếu tố học thuật như điểm số và số tín chỉ tích lũy thường là chỉ báo mạnh đối với nguy cơ bỏ học, trong khi các yếu tố xã hội và nhân khẩu học đóng vai trò hỗ trợ trong việc giải thích hành vi của người học [17], [18], [21]. Điều này cho thấy việc áp dụng cùng một cơ chế xử lý cho tất cả các loại đặc trưng có thể chưa phải là lựa chọn tối ưu.

Xuất phát từ những nhận định trên, nghiên cứu đề xuất một kiến trúc Hybrid nhằm xử lý riêng biệt từng nhóm dữ liệu trước khi tiến hành hợp nhất. Ý tưởng cốt lõi của mô hình là tận dụng thế mạnh của các mạng nơ-ron truyền thẳng đối với dữ liệu số và kỹ thuật Embedding đối với dữ liệu phân loại. Cách tiếp cận này cho phép mô hình khai thác sâu hơn các đặc trưng đặc thù của từng nhóm dữ liệu, đồng thời học được các tương tác phức tạp giữa chúng trong giai đoạn hợp nhất.

Khác với RNN và LSTM, vốn tập trung vào khả năng ghi nhớ thông tin theo chuỗi thời gian, mô hình Hybrid được thiết kế nhằm tối ưu hóa quá trình học biểu diễn đặc trưng (representation learning) trên dữ liệu hỗn hợp. Đây cũng chính là đóng góp chính của nghiên cứu, khi mô hình được xây dựng dựa trên đặc điểm thực tế của dữ liệu sinh viên thay vì áp dụng trực tiếp các kiến trúc học sâu có sẵn.

2.4.2. Kiến trúc tổng thể của mô hình

Mô hình Hybrid được đề xuất trong nghiên cứu được xây dựng theo kiến trúc đa nhánh (Multi-Branch Architecture), trong đó mỗi nhánh được thiết kế chuyên biệt để xử lý một nhóm dữ liệu có bản chất riêng. Ý tưởng cốt lõi của kiến trúc này xuất phát từ thực tế rằng dữ liệu sinh viên là một dạng dữ liệu hỗn hợp (heterogeneous data), bao gồm đồng thời các thuộc tính số và các thuộc tính phân loại. Hai nhóm dữ liệu này không chỉ khác nhau về cách biểu diễn mà còn khác nhau về ý nghĩa và phương thức khai thác thông tin. Do đó, việc xử lý chúng thông qua cùng một cơ chế học có thể làm giảm hiệu quả khai thác đặc trưng của mô hình.

Khác với các kiến trúc mạng truyền thống chỉ sử dụng một luồng xử lý duy nhất cho toàn bộ dữ liệu đầu vào, mô hình Hybrid cho phép từng nhóm đặc trưng được học theo cơ chế tối ưu nhất đối với bản chất của chúng. Nhờ đó, mô hình vừa bảo toàn được thông tin riêng của từng loại dữ liệu, vừa có khả năng học được các tương tác phức tạp giữa các nhóm đặc trưng trong giai đoạn sau.

Về tổng thể, kiến trúc Hybrid được chia thành ba giai đoạn chính:

- Giai đoạn xử lý dữ liệu số (Numerical Feature Processing).
- Giai đoạn xử lý dữ liệu phân loại (Categorical Feature Processing).
- Giai đoạn hợp nhất đặc trưng và phân loại (Feature Fusion and Classification).

Ở giai đoạn đầu tiên, các thuộc tính số được đưa vào một mạng nơ-ron truyền thẳng nhiều tầng để học các mối quan hệ phi tuyến giữa các biến liên tục. Các thuộc tính này chủ yếu phản ánh kết quả học tập, mức độ hoàn thành chương trình đào tạo và các yếu tố kinh tế - xã hội của sinh viên. Do các biến số thường có mối quan hệ định lượng rõ ràng, mạng nơ-ron truyền thẳng được lựa chọn nhằm tận dụng khả năng học các quan hệ phi tuyến và tương tác bậc cao giữa các đặc trưng.

Song song với đó, các thuộc tính phân loại được xử lý trong một nhánh độc lập. Thay vì sử dụng các phương pháp mã hóa truyền thống như One-Hot Encoding, nghiên cứu áp dụng kỹ thuật Embedding để chuyển đổi dữ liệu từ không gian rời rạc sang không gian vector liên tục. Quá trình này giúp mô hình học được các mối quan hệ ngữ nghĩa tiềm ẩn giữa các giá trị phân loại, đồng thời giảm đáng kể số chiều dữ

liệu đầu vào. Đây là một ưu điểm quan trọng khi xử lý các thuộc tính có nhiều giá trị khác nhau như trình độ học vấn của cha mẹ, loại hình tuyển sinh hoặc ngành học.

Sau khi hai nhánh xử lý hoàn tất quá trình học biểu diễn đặc trưng, đầu ra từ các nhánh sẽ được hợp nhất thông qua lớp Concatenate. Quá trình hợp nhất này cho phép mô hình kết hợp đồng thời các thông tin định lượng từ dữ liệu số và các thông tin ngữ nghĩa từ dữ liệu phân loại để tạo thành một vector biểu diễn tổng hợp duy nhất.

Vector tổng hợp này có ý nghĩa đặc biệt quan trọng vì nó phản ánh đầy đủ các khía cạnh khác nhau của người học. Không giống như các mô hình chỉ dựa trên kết quả học tập hoặc chỉ dựa trên thông tin nhân khẩu học, biểu diễn tổng hợp của mô hình Hybrid chứa đồng thời các yếu tố học thuật, cá nhân, gia đình và kinh tế - xã hội. Điều này tạo điều kiện để mô hình học được những mối quan hệ phức tạp mà các phương pháp xử lý đơn lẻ khó có thể phát hiện.

Sau bước hợp nhất, vector đặc trưng tiếp tục được đưa qua nhiều lớp Dense nhằm học các tương tác bậc cao giữa các nhóm dữ liệu. Đây là giai đoạn mà mô hình bắt đầu khai thác các mối liên hệ liên ngành giữa các đặc trưng. Chẳng hạn, cùng một mức điểm trung bình học tập, nguy cơ bỏ học của sinh viên có thể khác nhau tùy thuộc vào việc sinh viên đó có học bổng hay không, đến từ khu vực nào hoặc có hoàn cảnh gia đình ra sao. Những mối quan hệ đa chiều như vậy thường rất khó biểu diễn bằng các mô hình tuyến tính hoặc các mô hình xử lý từng nhóm dữ liệu riêng biệt.

Một điểm đáng chú ý khác của kiến trúc Hybrid là khả năng mở rộng. Nhờ cấu trúc đa nhánh, mô hình có thể dễ dàng tích hợp thêm các nguồn dữ liệu mới trong tương lai mà không cần thay đổi toàn bộ kiến trúc. Ví dụ, nếu bổ sung dữ liệu từ hệ thống LMS, dữ liệu điểm danh, dữ liệu hành vi học tập trực tuyến hoặc dữ liệu tương tác trên nền tảng học tập số, các nguồn dữ liệu này có thể được xử lý bởi các nhánh chuyên biệt trước khi hợp nhất với các đặc trưng hiện có. Điều này giúp mô hình có tính linh hoạt cao và phù hợp với xu hướng phát triển của các hệ thống phân tích học tập hiện đại.

Nhìn chung, kiến trúc Hybrid được xây dựng dựa trên nguyên tắc xử lý chuyên biệt cho từng nhóm dữ liệu trước khi thực hiện học biểu diễn tổng hợp. Thiết kế này không chỉ tận dụng được thế mạnh của từng phương pháp xử lý đặc trưng mà còn cho phép mô hình khai thác các tương tác phức tạp giữa nhiều nguồn thông tin khác nhau.

Đây chính là điểm khác biệt quan trọng giữa mô hình Hybrid đề xuất trong nghiên cứu và các kiến trúc mạng đơn nhánh truyền thống như RNN hoặc LSTM.

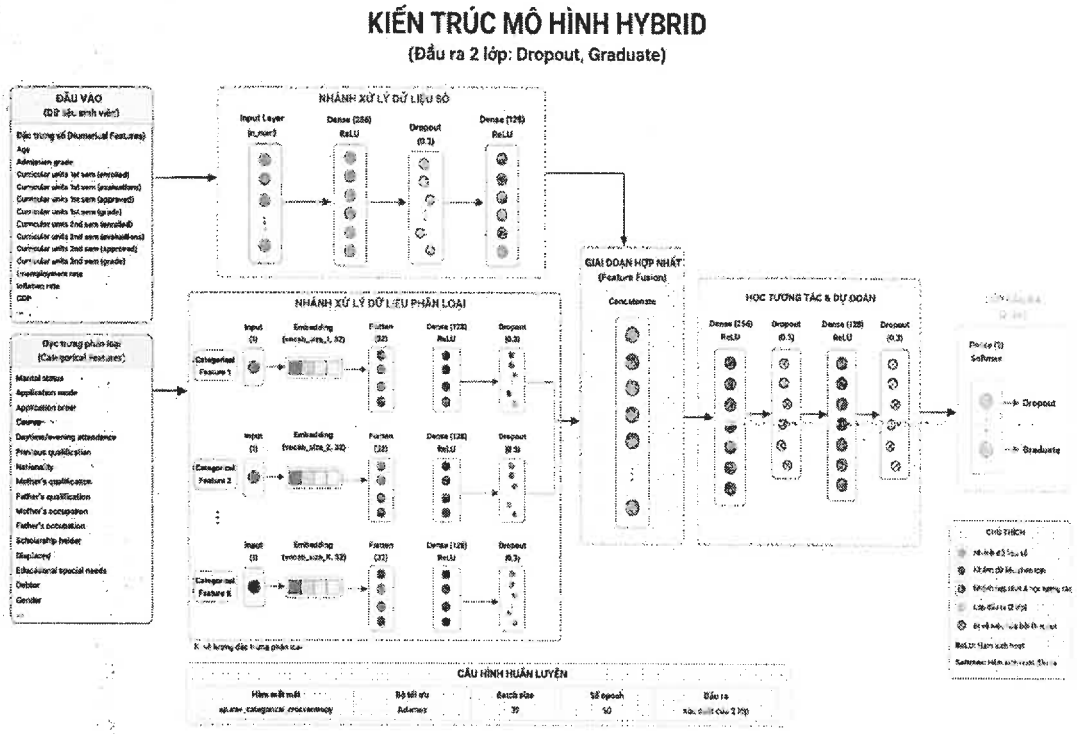


Figure 2.6. Xây dựng mô hình Hybrid

Nhánh xử lý dữ liệu số

Các thuộc tính số trong bộ dữ liệu sinh viên bao gồm nhiều thông tin quan trọng phản ánh quá trình học tập và điều kiện kinh tế - xã hội của người học, chẳng hạn như điểm tuyển sinh đầu vào, số tín chỉ đăng ký, số tín chỉ hoàn thành, điểm trung bình học tập, tỷ lệ thất nghiệp, tỷ lệ lạm phát và GDP. Đây đều là các biến liên tục, mang tính định lượng và có khả năng phản ánh trực tiếp nguy cơ bỏ học của sinh viên.

Để khai thác hiệu quả nhóm đặc trưng này, nghiên cứu sử dụng mạng nơ-ron truyền thẳng nhiều tầng (Feedforward Neural Network - FNN) làm bộ trích xuất đặc trưng. Khác với các phương pháp học máy truyền thống chỉ học các ranh giới phân loại tuyến tính hoặc bán tuyến tính, mạng nơ-ron sâu có khả năng mô hình hóa các mối quan hệ phi tuyến phức tạp giữa các biến đầu vào. Điều này đặc biệt phù hợp với dữ liệu giáo dục, nơi các yếu tố ảnh hưởng đến nguy cơ bỏ học thường không tồn tại độc lập mà có sự tương tác đa chiều.

Ví dụ, tác động của điểm trung bình học tập đối với khả năng bỏ học có thể khác nhau giữa các nhóm sinh viên nhận học bổng và không nhận học bổng. Tương tự, cùng một mức tín chỉ tích lũy nhưng nguy cơ bỏ học có thể thay đổi tùy theo độ tuổi hoặc hoàn cảnh kinh tế của sinh viên. Những mối quan hệ phi tuyến như vậy rất khó được mô hình hóa bằng các phương pháp thống kê truyền thống nhưng lại là thế mạnh của mạng nơ-ron sâu.

Trong nghiên cứu này, nhánh xử lý dữ liệu số được xây dựng với hai lớp Dense gồm 256 và 128 nơ-ron. Kiến trúc này cho phép mô hình học dần từ các biểu diễn tổng quát ở tầng đầu sang các biểu diễn chuyên biệt hơn ở các tầng tiếp theo. Số lượng nơ-ron tương đối lớn ở tầng đầu giúp mô hình có đủ năng lực biểu diễn để khai thác các mẫu dữ liệu tiềm ẩn, trong khi tầng sau thực hiện nhiệm vụ tinh lọc và cô đọng thông tin trước khi chuyển sang giai đoạn hợp nhất.

Hàm kích hoạt ReLU được sử dụng cho toàn bộ các lớp Dense trong nhánh này. Việc lựa chọn ReLU xuất phát từ khả năng học hiệu quả trên các mạng sâu, giảm hiện tượng tiêu biến gradient và tăng tốc độ hội tụ của mô hình [29]. So với các hàm kích hoạt sigmoid hoặc tanh, ReLU thường cho kết quả tốt hơn đối với các bài toán dự đoán sử dụng dữ liệu có số chiều lớn và nhiều mối quan hệ phi tuyến.

Bên cạnh đó, kỹ thuật Dropout với tỷ lệ 0,3 được áp dụng nhằm hạn chế hiện tượng quá khớp. Trong quá trình huấn luyện, một phần các nơ-ron sẽ được vô hiệu hóa ngẫu nhiên, buộc mô hình phải học các biểu diễn mang tính tổng quát hơn thay vì ghi nhớ dữ liệu huấn luyện. Điều này giúp cải thiện khả năng dự đoán khi áp dụng trên các tập dữ liệu mới.

Nhánh xử lý dữ liệu phân loại

Khác với dữ liệu số, các thuộc tính phân loại trong bộ dữ liệu không mang ý nghĩa định lượng mà chủ yếu phản ánh các đặc điểm cá nhân, gia đình và xã hội của sinh viên. Các thuộc tính tiêu biểu bao gồm giới tính, tình trạng hôn nhân, quốc tịch, trình độ học vấn của cha mẹ, nghề nghiệp của cha mẹ, loại hình tuyển sinh, tình trạng học bổng và nhiều yếu tố khác.

Một trong những thách thức lớn khi xử lý dữ liệu phân loại là làm thế nào để biểu diễn các giá trị rời rạc dưới dạng mà mạng nơ-ron có thể học được. Phương pháp phổ biến nhất là One-Hot Encoding, tuy nhiên phương pháp này tồn tại nhiều hạn chế khi số lượng giá trị của thuộc tính tăng lên. One-Hot Encoding làm gia tăng số chiều

dữ liệu, đồng thời không thể hiện được mức độ tương đồng hoặc mối quan hệ ngữ nghĩa giữa các giá trị khác nhau.

Để khắc phục hạn chế đó, nghiên cứu sử dụng kỹ thuật Embedding. Về bản chất, Embedding là quá trình ánh xạ các giá trị phân loại sang một không gian vector liên tục có số chiều thấp hơn. Trong quá trình huấn luyện, các vector này được cập nhật liên tục, giúp mô hình học được các đặc trưng ngữ nghĩa tiềm ẩn giữa các giá trị.

Ví dụ, hai nhóm sinh viên có trình độ học vấn của cha mẹ tương đối gần nhau có thể được ánh xạ tới các vị trí gần nhau trong không gian Embedding. Tương tự, các loại hình tuyển sinh có đặc điểm tương đồng cũng có thể được biểu diễn bằng các vector gần nhau. Đây là ưu điểm quan trọng mà các phương pháp mã hóa truyền thống không thể thực hiện được.

Trong nghiên cứu này, mỗi thuộc tính phân loại được đưa qua một lớp Embedding với kích thước vector là 32. Giá trị này được lựa chọn nhằm đảm bảo sự cân bằng giữa khả năng biểu diễn thông tin và chi phí tính toán. Nếu số chiều Embedding quá nhỏ, mô hình có thể không học được đầy đủ thông tin ngữ nghĩa; ngược lại, nếu quá lớn sẽ làm tăng số lượng tham số và nguy cơ quá khớp.

Sau khi Embedding, các vector được làm phẳng thông qua lớp Flatten và tiếp tục được xử lý bởi một lớp Dense gồm 128 nơ-ron. Mục tiêu của giai đoạn này là trích xuất các đặc trưng bậc cao từ các vector Embedding trước khi chuyển sang giai đoạn hợp nhất. Tương tự nhánh dữ liệu số, kỹ thuật Dropout với tỷ lệ 0,3 cũng được áp dụng nhằm tăng khả năng tổng quát hóa của mô hình.

Cơ chế hợp nhất các đặc trưng

au khi hoàn thành quá trình học biểu diễn trên từng nhóm dữ liệu, các đầu ra từ nhánh xử lý dữ liệu số và nhánh xử lý dữ liệu phân loại được đưa vào giai đoạn hợp nhất đặc trưng (Feature Fusion). Đây được xem là thành phần quan trọng nhất trong toàn bộ kiến trúc Hybrid, bởi hiệu quả dự đoán của mô hình không chỉ phụ thuộc vào khả năng học đặc trưng riêng lẻ của từng nhánh mà còn phụ thuộc vào khả năng khai thác các mối quan hệ tương tác giữa nhiều nguồn thông tin khác nhau.

Trong các nghiên cứu dự đoán bỏ học truyền thống, các đặc trưng thường được đưa trực tiếp vào cùng một mô hình sau khi trải qua quá trình mã hóa và chuẩn hóa. Cách tiếp cận này tuy đơn giản nhưng có thể làm giảm khả năng khai thác các đặc

điểm riêng biệt của từng loại dữ liệu. Các thuộc tính số và thuộc tính phân loại vốn có bản chất hoàn toàn khác nhau, do đó việc xử lý chúng trong cùng một không gian biểu diễn ngay từ đầu có thể dẫn đến hiện tượng mất mát thông tin hoặc làm suy giảm ý nghĩa ngữ nghĩa của dữ liệu.

Ngược lại, trong mô hình Hybrid được đề xuất, mỗi nhóm dữ liệu trước tiên được xử lý thông qua một nhánh chuyên biệt nhằm học ra các biểu diễn đặc trưng tối ưu nhất. Các thuộc tính số được chuyển đổi thành các vector đặc trưng phản ánh mối quan hệ định lượng giữa các biến liên tục, trong khi các thuộc tính phân loại được ánh xạ sang không gian vector thông qua Embedding để biểu diễn các quan hệ ngữ nghĩa tiềm ẩn giữa các giá trị. Chỉ sau khi hoàn thành quá trình học biểu diễn ở từng nhánh, các đặc trưng mới được đưa vào giai đoạn hợp nhất.

Quá trình hợp nhất được thực hiện thông qua lớp Concatenate, trong đó toàn bộ các vector đặc trưng từ nhánh số và nhánh phân loại được ghép lại thành một vector biểu diễn tổng hợp duy nhất. Vector này chứa đồng thời các thông tin học thuật, nhân khẩu học, điều kiện kinh tế - xã hội, hoàn cảnh gia đình và các đặc trưng ngữ nghĩa được học từ các lớp Embedding. Có thể xem đây là một biểu diễn toàn diện của người học, trong đó mỗi chiều dữ liệu phản ánh một khía cạnh khác nhau có khả năng ảnh hưởng đến nguy cơ bỏ học.

Điểm quan trọng nhất của giai đoạn này không chỉ nằm ở việc ghép nối các đặc trưng, mà còn ở khả năng tạo ra một không gian biểu diễn chung để mô hình học được các tương tác đa chiều giữa các yếu tố khác nhau. Trong thực tế, quyết định bỏ học của sinh viên hiếm khi xuất phát từ một nguyên nhân đơn lẻ. Thay vào đó, đây thường là kết quả của sự tác động đồng thời từ nhiều yếu tố học thuật, tài chính, xã hội và cá nhân.

Ví dụ, một sinh viên có điểm trung bình thấp chưa chắc đã có nguy cơ bỏ học cao nếu người học đó nhận được học bổng, có động lực học tập tốt hoặc nhận được sự hỗ trợ từ gia đình. Ngược lại, một sinh viên có kết quả học tập tương đối ổn định nhưng gặp khó khăn về tài chính hoặc phải vừa học vừa làm có thể đối mặt với nguy cơ bỏ học lớn hơn. Những mối quan hệ tương tác phức tạp như vậy rất khó được phát hiện nếu các đặc trưng được xem xét một cách độc lập. Tuy nhiên, thông qua cơ chế hợp nhất đặc trưng, mô hình Hybrid có khả năng học được các mẫu dữ liệu đa chiều và các tương tác tiềm ẩn giữa nhiều nhóm biến khác nhau.

Sau bước Concatenate, vector tổng hợp được đưa qua hai lớp Dense gồm 256 và 128 nơ-ron sử dụng hàm kích hoạt ReLU. Các lớp này đóng vai trò học các tương tác bậc cao (high-order interactions) giữa các nhóm đặc trưng đã được hợp nhất. Nếu các nhánh trước đó tập trung vào việc học đặc trưng riêng của từng loại dữ liệu, thì các lớp Dense sau hợp nhất sẽ đảm nhiệm việc khám phá các quy luật kết hợp giữa chúng. Đây chính là giai đoạn mà mô hình có thể nhận diện những mẫu hành vi phức tạp liên quan đến nguy cơ bỏ học mà các mô hình đơn nhánh hoặc các phương pháp học máy truyền thống khó có thể phát hiện.

Bên cạnh đó, để nâng cao khả năng tổng quát hóa và hạn chế hiện tượng quá khớp, kỹ thuật Dropout với tỷ lệ 0,3 được áp dụng sau mỗi lớp Dense trong giai đoạn hợp nhất. Điều này giúp mô hình tránh phụ thuộc quá mức vào một số đặc trưng hoặc tương tác cụ thể, từ đó cải thiện khả năng dự đoán trên các tập dữ liệu chưa từng xuất hiện trong quá trình huấn luyện.

Có thể thấy rằng, cơ chế hợp nhất đặc trưng không đơn thuần là bước kết nối đầu ra từ các nhánh xử lý dữ liệu, mà chính là trung tâm của toàn bộ kiến trúc Hybrid. Thông qua việc kết hợp các biểu diễn đặc trưng khác nhau vào cùng một không gian học tập, mô hình có khả năng xây dựng một cái nhìn toàn diện hơn về người học, đồng thời khai thác hiệu quả các mối quan hệ đa chiều giữa các yếu tố ảnh hưởng đến quá trình học tập. Đây cũng chính là nền tảng giúp mô hình Hybrid được kỳ vọng đạt hiệu quả dự đoán cao hơn so với các kiến trúc truyền thống như RNN và LSTM trong bài toán dự đoán sinh viên bỏ học.

2.4.3. Hàm mất mát và thuật toán tối ưu

Để huấn luyện mô hình Hybrid, nghiên cứu sử dụng hàm mất mát sparse_categorical_crossentropy. Đây là hàm mất mát được sử dụng phổ biến trong các bài toán phân loại nhiều lớp khi nhãn đầu ra được biểu diễn dưới dạng số nguyên. Trong bài toán dự đoán trạng thái học tập của sinh viên, mô hình cần xác định xác suất sinh viên thuộc từng nhóm kết quả học tập khác nhau. Do đó, việc sử dụng hàm mất mát sparse_categorical_crossentropy giúp đo lường hiệu quả mức độ phù hợp giữa xác suất dự đoán của mô hình và nhãn thực tế của dữ liệu.

Về mặt toán học, hàm mất mát này được xây dựng dựa trên nguyên lý entropy chéo (cross-entropy), cho phép đánh giá khoảng cách giữa hai phân bố xác suất. Khi xác suất dự đoán của lớp đúng càng cao thì giá trị mất mát càng nhỏ; ngược lại, nếu

mô hình dự đoán sai hoặc thiếu chắc chắn thì giá trị mất mát sẽ tăng lên đáng kể. Điều này tạo ra tín hiệu phản hồi giúp mô hình liên tục điều chỉnh trọng số trong quá trình huấn luyện.

Một trong những ưu điểm quan trọng của sparse_categorical_crossentropy là không yêu cầu chuyển đổi nhãn sang dạng one-hot encoding. Điều này giúp giảm chi phí bộ nhớ và đơn giản hóa quá trình xử lý dữ liệu, đặc biệt đối với các tập dữ liệu có số lượng mẫu lớn.

Bên cạnh hàm mất mát, việc lựa chọn thuật toán tối ưu đóng vai trò quyết định đến tốc độ hội tụ và hiệu quả học tập của mô hình. Trong nghiên cứu này, thuật toán Adamax được lựa chọn thay cho Adam truyền thống.

Adamax là một biến thể của thuật toán Adam, được đề xuất bởi Kingma và Ba trong cùng nghiên cứu giới thiệu Adam. Trong khi Adam sử dụng chuẩn Euclidean (L2 norm) để tính toán tốc độ học thích nghi, Adamax sử dụng chuẩn vô cực (infinity norm) để thực hiện quá trình cập nhật trọng số. Nhờ đó, Adamax thường hoạt động ổn định hơn trong các bài toán có không gian tham số lớn hoặc dữ liệu có tính thưa (sparse data).

Việc lựa chọn Adamax trong nghiên cứu xuất phát từ đặc điểm của kiến trúc Hybrid. Mô hình không chỉ bao gồm các lớp Dense thông thường mà còn chứa nhiều lớp Embedding dành cho dữ liệu phân loại. Các lớp Embedding tạo ra số lượng lớn tham số cần tối ưu, đồng thời nhiều tham số trong số đó chỉ được cập nhật đối với một phần nhỏ mẫu dữ liệu trong mỗi batch huấn luyện. Trong trường hợp này, Adamax thường thể hiện khả năng ổn định tốt hơn so với Adam do ít nhạy cảm hơn với các gradient có độ lớn biến động mạnh.

Ngoài ra, mô hình Hybrid được xây dựng theo cấu trúc đa nhánh (multi-branch architecture), trong đó các nhánh xử lý dữ liệu số và dữ liệu phân loại được huấn luyện đồng thời. Quá trình này tạo ra các gradient có đặc điểm khác nhau ở từng nhánh mạng. Adamax giúp duy trì tốc độ học ổn định trên toàn bộ mô hình, hạn chế hiện tượng một nhóm tham số được cập nhật quá nhanh trong khi nhóm khác hội tụ chậm.

Thông qua việc kết hợp giữa hàm mất mát sparse_categorical_crossentropy và thuật toán tối ưu Adamax, mô hình Hybrid có khả năng học hiệu quả trên tập dữ liệu

có cấu trúc hỗn hợp, đồng thời đảm bảo quá trình huấn luyện diễn ra ổn định và đạt khả năng hội tụ tốt.

2.4.4. Ưu điểm của mô hình Hybrid

So với các mô hình học sâu truyền thống như RNN và LSTM, mô hình Hybrid được đề xuất trong nghiên cứu mang lại một số ưu điểm nổi bật khi xử lý dữ liệu giáo dục có tính dị thể (heterogeneous data).

Ưu điểm đầu tiên là khả năng xử lý chuyên biệt đối với từng loại dữ liệu. Trong khi RNN và LSTM thường xử lý toàn bộ các thuộc tính đầu vào theo cùng một cơ chế, mô hình Hybrid cho phép mỗi nhóm dữ liệu được đưa qua một nhánh xử lý phù hợp với bản chất của nó. Các đặc trưng số được khai thác thông qua mạng nơ-ron truyền thẳng nhiều tầng, trong khi các đặc trưng phân loại được biểu diễn bằng Embedding trước khi học các biểu diễn ngữ nghĩa ở mức cao hơn.

Cách tiếp cận này giúp giảm hiện tượng mất mát thông tin thường gặp khi áp dụng các phương pháp mã hóa truyền thống. Đặc biệt, đối với các thuộc tính phân loại có nhiều giá trị khác nhau, Embedding cho phép mô hình học được mối quan hệ tiềm ẩn giữa các giá trị thay vì xem chúng là hoàn toàn độc lập.

Ưu điểm thứ hai là khả năng học các tương tác phức tạp giữa các nhóm đặc trưng. Trong thực tế, quyết định bỏ học của sinh viên thường không xuất phát từ một nguyên nhân đơn lẻ mà là kết quả của sự kết hợp giữa nhiều yếu tố khác nhau. Một sinh viên có kết quả học tập trung bình vẫn có thể tốt nghiệp nếu nhận được sự hỗ trợ tài chính hoặc có động lực học tập cao. Ngược lại, một sinh viên có thành tích học tập khá nhưng gặp khó khăn về kinh tế hoặc thiếu sự gắn kết với môi trường học tập vẫn có nguy cơ bỏ học.

Các mối quan hệ đa chiều như vậy rất khó được mô hình hóa bằng các thuật toán tuyến tính hoặc các mô hình chỉ tập trung vào một loại đặc trưng. Nhờ cơ chế hợp nhất đặc trưng và các tầng học sâu phía sau, mô hình Hybrid có khả năng phát hiện các tương tác bậc cao giữa các yếu tố học thuật, cá nhân và kinh tế - xã hội.

Ưu điểm thứ ba là khả năng mở rộng và thích nghi với các bộ dữ liệu giáo dục khác nhau. Kiến trúc đa nhánh cho phép dễ dàng bổ sung thêm các nhóm đặc trưng mới mà không cần thay đổi toàn bộ cấu trúc mô hình. Ví dụ, trong tương lai có thể bổ sung dữ liệu từ hệ thống LMS, dữ liệu điểm danh, dữ liệu hành vi học tập hoặc dữ

liệu tương tác trên các nền tảng trực tuyến bằng cách xây dựng thêm các nhánh xử lý chuyên biệt.

Một ưu điểm quan trọng khác là tính linh hoạt trong biểu diễn dữ liệu. Thông qua cơ chế Embedding và học biểu diễn sâu, mô hình không phụ thuộc nhiều vào việc thiết kế đặc trưng thủ công như các phương pháp học máy truyền thống. Điều này giúp giảm công sức tiền xử lý dữ liệu và tăng khả năng tự động phát hiện các đặc trưng quan trọng từ dữ liệu gốc.

Từ góc độ nghiên cứu, mô hình Hybrid cũng mang ý nghĩa thực tiễn cao. Thay vì chỉ tập trung vào kết quả dự đoán cuối cùng, mô hình cho phép phân tích mức độ đóng góp của từng nhóm đặc trưng đối với nguy cơ bỏ học. Điều này có thể hỗ trợ các nhà quản lý giáo dục xây dựng các chính sách can thiệp phù hợp, chẳng hạn như hỗ trợ học thuật, hỗ trợ tài chính hoặc tư vấn tâm lý đối với các nhóm sinh viên có nguy cơ cao.

Tóm lại, mô hình Hybrid được đề xuất không chỉ tận dụng được ưu điểm của các mạng nơ-ron sâu hiện đại mà còn được thiết kế dựa trên đặc điểm thực tế của dữ liệu giáo dục. Khả năng xử lý chuyên biệt cho từng nhóm đặc trưng, học các tương tác phức tạp và mở rộng linh hoạt giúp mô hình trở thành một giải pháp tiềm năng cho bài toán dự đoán sinh viên bỏ học. Đây cũng là mô hình được kỳ vọng đạt hiệu quả dự đoán cao nhất trong số các phương pháp được khảo sát trong nghiên cứu này.

2.4.5. So sánh lý thuyết giữa các mô hình RNN, LSTM và Hybrid

Trong nghiên cứu này, ba mô hình học sâu gồm RNN, LSTM và Hybrid được lựa chọn để xây dựng và đánh giá trên cùng một tập dữ liệu dự đoán nguy cơ sinh viên bỏ học. Mỗi mô hình đại diện cho một hướng tiếp cận khác nhau trong việc học biểu diễn dữ liệu và khai thác các mối quan hệ tiềm ẩn giữa các đặc trưng đầu vào.

RNN là mô hình cơ sở được xây dựng đầu tiên trong nghiên cứu. Với khả năng lưu giữ thông tin thông qua trạng thái ẩn, RNN có thể học được các mối quan hệ tuần tự giữa các đặc trưng trong dữ liệu đầu vào. Kiến trúc này đặc biệt phù hợp với các bài toán có tính phụ thuộc theo thời gian hoặc có cấu trúc chuỗi dữ liệu [15], [29]. Tuy nhiên, nhược điểm lớn nhất của RNN là hiện tượng tiêu biến gradient khi xử lý các chuỗi dài, làm suy giảm khả năng ghi nhớ thông tin ở các bước thời gian xa [23].

Để khắc phục hạn chế này, mô hình LSTM được đề xuất như một phiên bản cải tiến của RNN. Thông qua cơ chế bộ nhớ dài hạn (cell state) cùng các cổng điều khiển thông tin, LSTM có khả năng duy trì và cập nhật thông tin quan trọng trong thời gian dài hơn. Điều này giúp mô hình học tốt hơn các phụ thuộc dài hạn và thường đạt hiệu quả cao hơn RNN trong nhiều bài toán dự đoán chuỗi [23]. Tuy nhiên, đổi lại, LSTM có cấu trúc phức tạp hơn, số lượng tham số lớn hơn và yêu cầu thời gian huấn luyện cao hơn.

Mặc dù RNN và LSTM đều có ưu thế trong việc khai thác dữ liệu mang tính tuần tự, nhưng bộ dữ liệu sử dụng trong nghiên cứu này lại có đặc điểm khác biệt. Dữ liệu sinh viên không chỉ bao gồm các thuộc tính học thuật mà còn chứa nhiều thông tin nhân khẩu học, kinh tế - xã hội và các biến phân loại có số lượng giá trị lớn. Trong trường hợp này, việc xử lý toàn bộ dữ liệu theo cùng một cơ chế có thể chưa khai thác được đầy đủ giá trị của từng nhóm đặc trưng.

Xuất phát từ nhận định đó, mô hình Hybrid được đề xuất nhằm xử lý riêng biệt các nhóm dữ liệu có bản chất khác nhau trước khi thực hiện hợp nhất. Các đặc trưng số được học thông qua mạng nơ-ron truyền thẳng nhiều tầng, trong khi các đặc trưng phân loại được biểu diễn bằng Embedding và xử lý thông qua các nhánh chuyên biệt. Sau đó, các đặc trưng được hợp nhất để học các tương tác bậc cao giữa các nhóm dữ liệu.

Khác với RNN và LSTM, trọng tâm của Hybrid không nằm ở việc học phụ thuộc theo thời gian mà tập trung vào khả năng học biểu diễn dữ liệu dị thể. Điều này đặc biệt phù hợp với các bộ dữ liệu giáo dục hiện đại, nơi thông tin được thu thập từ nhiều nguồn khác nhau và tồn tại dưới nhiều định dạng khác nhau. Từ góc độ lý thuyết, RNN và LSTM được kỳ vọng có khả năng học tốt các mối quan hệ tuần tự trong dữ liệu, trong khi mô hình Hybrid được kỳ vọng tận dụng hiệu quả hơn các đặc trưng đa dạng của bộ dữ liệu sinh viên. Do đó, nghiên cứu tiến hành xây dựng và đánh giá cả ba mô hình nhằm xác định kiến trúc phù hợp nhất cho bài toán dự đoán nguy cơ bỏ học trong môi trường giáo dục đại học.

Các kết quả thực nghiệm và phân tích chi tiết sẽ được trình bày trong phần tiếp theo nhằm đánh giá khách quan hiệu quả của từng mô hình trên cùng một tập dữ liệu và cùng điều kiện thực nghiệm.

2.5. Kết luận chương

Chương 2 đã trình bày toàn bộ quy trình xây dựng các mô hình học sâu phục vụ bài toán dự đoán nguy cơ sinh viên bỏ học. Trước hết, nghiên cứu đã thực hiện các bước tiền xử lý dữ liệu, bao gồm mô tả dữ liệu đầu vào, làm sạch dữ liệu, chuyển đổi các thuộc tính phù hợp với yêu cầu của mô hình học sâu và phân chia tập dữ liệu phục vụ quá trình huấn luyện và đánh giá. Đây là bước quan trọng nhằm đảm bảo chất lượng dữ liệu đầu vào, đồng thời tạo điều kiện để các mô hình có thể học được những đặc trưng mang ý nghĩa dự báo.

Trên cơ sở dữ liệu đã được chuẩn bị, nghiên cứu đã xây dựng và triển khai hai mô hình học sâu phổ biến là RNN và LSTM. Các mô hình này được lựa chọn nhằm khai thác khả năng học các mối quan hệ phi tuyến và các phụ thuộc tiềm ẩn trong dữ liệu sinh viên. Trong đó, RNN đóng vai trò là mô hình cơ sở với khả năng xử lý dữ liệu tuần tự, còn LSTM được phát triển nhằm khắc phục hạn chế về khả năng ghi nhớ dài hạn của RNN thông qua cơ chế bộ nhớ và các cổng điều khiển thông tin.

Bên cạnh đó, chương đã tập trung trình bày mô hình Hybrid – mô hình được đề xuất trong nghiên cứu. Khác với các kiến trúc truyền thống, mô hình Hybrid được xây dựng theo hướng xử lý chuyên biệt từng nhóm dữ liệu trước khi thực hiện hợp nhất đặc trưng. Các thuộc tính số được khai thác thông qua mạng nơ-ron truyền thẳng nhiều tầng, trong khi các thuộc tính phân loại được biểu diễn bằng kỹ thuật Embedding nhằm học các mối quan hệ ngữ nghĩa tiềm ẩn giữa các giá trị. Sau quá trình học biểu diễn, các đặc trưng được hợp nhất để tạo thành một không gian biểu diễn tổng hợp, cho phép mô hình khai thác đồng thời các yếu tố học thuật, nhân khẩu học, gia đình và kinh tế - xã hội có liên quan đến nguy cơ bỏ học của sinh viên. Đây được xem là đóng góp chính của nghiên cứu trong việc nâng cao khả năng biểu diễn dữ liệu và cải thiện hiệu quả dự đoán.

Ngoài ra, chương cũng đã trình bày các hàm mất mát, thuật toán tối ưu và các lựa chọn kiến trúc được sử dụng trong từng mô hình, đồng thời phân tích ưu điểm, hạn chế và sự khác biệt giữa RNN, LSTM và Hybrid trên phương diện lý thuyết. Những phân tích này là cơ sở để xây dựng các giả thuyết nghiên cứu về hiệu quả của từng mô hình trước khi tiến hành thực nghiệm.

Nhìn chung, các nội dung được trình bày trong chương đã hình thành một quy trình hoàn chỉnh từ chuẩn bị dữ liệu đến thiết kế và triển khai mô hình. Đây là nền

tảng quan trọng cho việc tiến hành huấn luyện, đánh giá và so sánh hiệu suất của các mô hình trong chương tiếp theo. Kết quả thực nghiệm được trình bày ở Chương 3 sẽ giúp kiểm chứng các giả thuyết nghiên cứu, đồng thời đánh giá mức độ phù hợp của từng mô hình đối với bài toán dự đoán nguy cơ sinh viên bỏ học trong môi trường giáo dục đại học.

CHƯƠNG 3: THỬ NGHIỆM VÀ ĐÁNH GIÁ

- Giới thiệu chương

Chương này tập trung trình bày kết quả thực nghiệm và phân tích sâu về bài toán dự đoán tỷ lệ bỏ học, nhằm cung cấp cái nhìn toàn diện về hiệu quả của các mô hình học sâu đã xây dựng. Trước tiên, chương sẽ bắt đầu với việc trình bày bài toán một cách cụ thể, làm rõ cách thức đánh giá và mục tiêu của thử nghiệm. Sau đó, kết quả thử nghiệm sẽ được phân tích, với việc so sánh hiệu suất của các mô hình RNN và LSTM dựa trên các thước đo quan trọng. Tiếp theo, chương sẽ khám phá các đặc trưng quan trọng trong dữ liệu, giúp lý giải các yếu tố ảnh hưởng đến khả năng dự đoán và hiệu suất của mô hình. Cuối cùng, chương sẽ thảo luận về ý nghĩa thực tế của các kết quả đạt được, đồng thời chỉ ra những hạn chế của các mô hình, mở ra hướng cải tiến trong các nghiên cứu và ứng dụng tiếp theo. Với nội dung này, chương không chỉ cung cấp cái nhìn khách quan về hiệu quả của các mô hình mà còn đặt vấn đề trong bối cảnh thực tiễn và khoa học.

3.1. Trình bày bài toán

Bài toán dự đoán nguy cơ sinh viên bỏ học là một trong những hướng nghiên cứu quan trọng trong lĩnh vực khai phá dữ liệu giáo dục và phân tích học tập. Trong bối cảnh các cơ sở giáo dục đại học ngày càng chú trọng đến chất lượng đào tạo, tỷ lệ duy trì người học và hiệu quả quản lý sinh viên, việc nhận diện sớm những sinh viên có nguy cơ rời bỏ chương trình học có ý nghĩa đặc biệt quan trọng. Nếu nhà trường có thể phát hiện sớm các trường hợp tiềm ẩn rủi ro, các biện pháp hỗ trợ như tư vấn học tập, hỗ trợ tài chính, điều chỉnh lộ trình học hoặc can thiệp tâm lý có thể được triển khai kịp thời. Do đó, bài toán dự đoán bỏ học không chỉ mang ý nghĩa về mặt kỹ thuật dự báo, mà còn có giá trị thực tiễn trong việc nâng cao tỷ lệ tốt nghiệp, giảm lãng phí nguồn lực đào tạo và cải thiện trải nghiệm học tập của sinh viên.

Trong nghiên cứu này, bài toán được xây dựng dưới dạng bài toán phân loại nhị phân, trong đó mục tiêu là dự đoán trạng thái học tập cuối cùng của sinh viên dựa trên các thông tin đầu vào được thu thập trong quá trình học tập. Mỗi bản ghi dữ liệu tương ứng với một sinh viên và được mô tả bởi tập hợp các thuộc tính phản ánh nhiều khía cạnh khác nhau của người học, bao gồm thông tin nhân khẩu học, đặc điểm học

thuật, hoàn cảnh kinh tế - xã hội và các yếu tố liên quan đến quá trình đào tạo. Cách tiếp cận này cho phép mô hình không chỉ xem xét kết quả học tập đơn thuần mà còn khai thác đồng thời nhiều nhóm yếu tố có khả năng tác động đến quyết định tiếp tục hay rời bỏ chương trình học.

Biến mục tiêu của nghiên cứu được biểu diễn bằng hai lớp chính. Lớp Dropout được mã hóa bằng giá trị 0, đại diện cho nhóm sinh viên bỏ học hoặc không hoàn thành chương trình đào tạo. Lớp Graduate được mã hóa bằng giá trị 1, đại diện cho nhóm sinh viên hoàn thành chương trình và tốt nghiệp. Việc đưa bài toán về dạng phân loại nhị phân giúp mô hình tập trung trực tiếp vào nhiệm vụ quan trọng nhất là phân biệt giữa nhóm sinh viên có nguy cơ bỏ học và nhóm sinh viên có khả năng hoàn thành chương trình. Đây cũng là cách tiếp cận phù hợp với mục tiêu ứng dụng thực tiễn của nghiên cứu, bởi trong quản lý giáo dục, việc nhận diện chính xác nhóm sinh viên có nguy cơ bỏ học thường có ý nghĩa cấp thiết hơn so với việc phân loại chi tiết nhiều trạng thái học tập khác nhau.

Dữ liệu sử dụng trong thực nghiệm là bộ dữ liệu "Predicting Student Dropout and Academic Success", bao gồm 4.424 bản ghi và 35 thuộc tính đầu vào. Các thuộc tính này bao phủ nhiều nhóm thông tin khác nhau, phản ánh tương đối toàn diện bối cảnh học tập và điều kiện cá nhân của sinh viên. Nhóm thông tin nhân khẩu học bao gồm các thuộc tính như tình trạng hôn nhân, quốc tịch, trình độ học vấn và nghề nghiệp của cha mẹ. Nhóm thông tin học tập bao gồm điểm đầu vào, số tín chỉ đăng ký, số tín chỉ hoàn thành, số tín chỉ được phê duyệt và kết quả học tập theo từng học kỳ. Bên cạnh đó, nhóm thông tin kinh tế - xã hội bao gồm các chỉ số như tỷ lệ thất nghiệp, tỷ lệ lạm phát và GDP, giúp mô hình xem xét thêm tác động của bối cảnh kinh tế đến khả năng duy trì học tập của sinh viên.

Trên cơ sở bộ dữ liệu này, nghiên cứu tiến hành xây dựng và đánh giá ba mô hình học sâu gồm RNN, LSTM và Hybrid. Mô hình RNN được sử dụng như một mô hình cơ sở nhằm khai thác các mối quan hệ tuần tự tiềm ẩn trong dữ liệu. Mô hình LSTM được triển khai nhằm cải thiện khả năng ghi nhớ thông tin dài hạn, khắc phục một số hạn chế của RNN trong quá trình học các phụ thuộc xa. Trong khi đó, mô hình Hybrid được đề xuất nhằm tận dụng đặc điểm dị thể của dữ liệu bằng cách xử lý riêng biệt các đặc trưng số và đặc trưng phân loại trước khi hợp nhất để đưa ra dự đoán cuối cùng. Việc triển khai đồng thời ba mô hình cho phép nghiên cứu so sánh hiệu quả của các kiến trúc khác nhau trên cùng một bài toán và cùng một tập dữ liệu.

Để đảm bảo quá trình đánh giá được thực hiện một cách khách quan, các mô hình được huấn luyện và kiểm tra trên cùng một tập dữ liệu đã qua tiền xử lý và phân chia theo cùng một tỷ lệ. Hiệu năng của các mô hình được đánh giá thông qua các chỉ số phổ biến trong bài toán phân loại, bao gồm Accuracy, Precision, Recall, F1-score và ROC-AUC. Trong đó, Accuracy phản ánh tỷ lệ dự đoán đúng tổng thể; Precision cho biết mức độ chính xác khi mô hình dự đoán một sinh viên thuộc nhóm có nguy cơ; Recall thể hiện khả năng phát hiện đầy đủ các trường hợp cần quan tâm; F1-score cân bằng giữa Precision và Recall; còn ROC-AUC đánh giá khả năng phân biệt giữa các lớp ở nhiều ngưỡng quyết định khác nhau.

Như vậy, bài toán thực nghiệm trong chương này không chỉ nhằm xác định mô hình có độ chính xác cao nhất, mà còn hướng tới đánh giá toàn diện khả năng nhận diện sinh viên có nguy cơ bỏ học của từng phương pháp. Kết quả thu được sẽ là cơ sở để phân tích ưu điểm, hạn chế và mức độ phù hợp của các mô hình học sâu trong bối cảnh giáo dục đại học. Đồng thời, những kết quả này cũng cung cấp nền tảng thực tiễn cho việc đề xuất các giải pháp hỗ trợ sinh viên, hướng tới mục tiêu giảm tỷ lệ bỏ học và nâng cao chất lượng đào tạo.

3.2. Một số bài báo nghiên cứu trước đây

3.2.1. Predicting Student Dropout from Day One

Carballo-Mendívil và cộng sự (2025) [32] đề xuất một hệ thống cảnh báo sớm dựa trên thuật toán XGBoost, trong đó mô hình chỉ khai thác dữ liệu có sẵn tại thời điểm tuyển sinh. Nghiên cứu sử dụng dữ liệu của 38.932 sinh viên năm thứ nhất trong giai đoạn 2014–2024 tại một trường đại học công lập ở Mexico. Các thuộc tính đầu vào bao gồm thành tích học tập trước đại học, thông tin nhân khẩu học, điều kiện kinh tế – xã hội, hoàn cảnh gia đình, điều kiện đi lại, việc làm và một số yếu tố nhận thức của sinh viên. Cách tiếp cận này kế thừa quan điểm cho rằng nguy cơ bỏ học là kết quả của sự tương tác giữa các yếu tố học thuật, cá nhân và môi trường, phù hợp với các mô hình lý thuyết của Tinto [33] [34], Bean và Metzner [35], Bourdieu [36] và Kerby [37].

Bài toán được xây dựng dưới dạng phân loại nhị phân, gồm hai lớp Dropout và Non-dropout. Một sinh viên được xác định là bỏ học khi không đăng ký học trong bốn kỳ liên tiếp. Quy trình nghiên cứu được tổ chức theo mô hình CRISP-DM, bao gồm các bước xác định bài toán, tìm hiểu dữ liệu, chuẩn bị dữ liệu, xây dựng mô hình,

đánh giá và triển khai. Trong quá trình tiền xử lý, các biến định danh và các thuộc tính có khả năng gây data leakage được loại bỏ; dữ liệu phân loại được mã hóa bằng one-hot encoding hoặc ordinal encoding và các biến số được chuẩn hóa trước khi huấn luyện.

Do lớp Dropout chỉ chiếm khoảng 26% tập dữ liệu, nghiên cứu sử dụng random undersampling để giảm ảnh hưởng của hiện tượng mất cân bằng lớp. Dữ liệu từ các khóa 2014–2022 được sử dụng cho huấn luyện và đánh giá theo tỷ lệ 80:20 với lấy mẫu phân tầng, trong khi dữ liệu 2023–2024 được giữ lại nhằm mô phỏng quá trình triển khai mô hình trên sinh viên mới. Cách tổ chức theo thời gian này giúp đánh giá mô hình gần hơn với điều kiện sử dụng thực tế.

Về mô hình, nghiên cứu so sánh bốn thuật toán gồm Logistic Regression, Random Forest, LightGBM và XGBoost. Trong đánh giá ban đầu, AUC-ROC lần lượt đạt 0,6752; 0,6842; 0,6882 và 0,6892. Sau tối ưu, Random Forest đạt AUC-ROC cao hơn XGBoost, nhưng XGBoost đạt Recall 0,91 đối với lớp Dropout, cao hơn đáng kể so với Random Forest và LightGBM. Vì mục tiêu của hệ thống cảnh báo sớm là hạn chế bỏ sót sinh viên thực sự có nguy cơ, tác giả lựa chọn XGBoost làm mô hình cuối cùng thay vì chỉ dựa trên Accuracy hoặc AUC. Cách lựa chọn này phù hợp với đặc thù của bài toán dự đoán bỏ học, trong đó False Negative có thể làm mất cơ hội can thiệp đối với một sinh viên thực sự cần hỗ trợ.

XGBoost được tối ưu bằng GridSearchCV kết hợp five-fold stratified cross-validation, với cấu hình cuối cùng gồm 200 cây, độ sâu tối đa 7, learning rate 0,05 và tỷ lệ lấy mẫu 0,8. Ngoài ra, thay vì sử dụng ngưỡng phân loại mặc định 0,5, nghiên cứu hiệu chỉnh ngưỡng xuống 0,338 nhằm tăng khả năng phát hiện lớp Dropout. Sau hiệu chỉnh, mô hình đạt AUC-ROC = 0,6902, Recall = 88%, Precision = 57% và F1-score = 0,6946 đối với lớp Dropout. Kết quả cho thấy mặc dù khả năng phân tách tổng thể chỉ ở mức trung bình, mô hình vẫn có độ nhạy cao và đặc biệt có giá trị vì toàn bộ dự báo được thực hiện trước khi có dữ liệu kết quả học tập ở bậc đại học.

Phân tích độ quan trọng của đặc trưng cho thấy tuổi, điểm trung bình trung học phổ thông và tình trạng nhập học có điều kiện là những biến có ảnh hưởng mạnh nhất đến dự báo. Ngoài ra, các yếu tố như gánh nặng gia đình, điều kiện kinh tế, nguồn chi trả học phí, thời gian rảnh, phương tiện di chuyển và trình độ học vấn của cha mẹ

cũng có vai trò đáng kể. Kết quả này tiếp tục củng cố quan điểm của các nghiên cứu trước rằng bỏ học là một hiện tượng đa yếu tố, không thể giải thích chỉ bằng thành tích học tập [33] [37].

Để tăng khả năng giải thích mô hình, nghiên cứu sử dụng SHAP nhằm xác định mức độ và chiều tác động của từng đặc trưng đến dự đoán của từng sinh viên. Đây là một điểm có ý nghĩa thực tiễn vì hệ thống không chỉ đưa ra xác suất bỏ học mà còn hỗ trợ giải thích nguyên nhân khiến một sinh viên được xếp vào nhóm nguy cơ cao. Sau khi huấn luyện, mô hình được áp dụng cho 4.682 sinh viên thuộc các khóa 2023–2024 và khoảng 59% được phân loại vào nhóm nguy cơ. Tác giả định hướng tích hợp mô hình vào hệ thống quản lý sinh viên để hỗ trợ cố vấn học tập, mentoring, học bổng và các hình thức hỗ trợ tài chính.

Nhìn chung, nghiên cứu của Carballo-Mendívil và cộng sự cho thấy giá trị của một hệ thống cảnh báo sớm không chỉ nằm ở việc đạt Accuracy cao mà còn ở khả năng phát hiện sớm và hạn chế bỏ sót sinh viên có nguy cơ. Đây cũng là một định hướng quan trọng đối với các nghiên cứu dự đoán bỏ học, trong đó cần lựa chọn chỉ số đánh giá, ngưỡng phân loại và chiến lược huấn luyện phù hợp với mục tiêu can thiệp thực tế.

3.2.2. Predicting Student Dropout Using Machine Learning Algorithms

Sulak và Koklu (2024) nghiên cứu khả năng ứng dụng các thuật toán học máy trong dự đoán tình trạng bỏ học và kết quả học tập của sinh viên. Theo các tác giả, bỏ học là một vấn đề đa yếu tố, không chỉ liên quan đến năng lực học tập mà còn chịu ảnh hưởng của điều kiện kinh tế – xã hội, đặc điểm nhân khẩu học, hoàn cảnh gia đình, yếu tố tâm lý và động lực học tập. Các nghiên cứu trước đó cũng cho thấy khó khăn kinh tế, thất bại trong học tập và suy giảm động lực, cũng như các yếu tố cá nhân, xã hội và gia đình, có thể làm gia tăng nguy cơ sinh viên rời khỏi quá trình đào tạo. Trên cơ sở đó, học máy được xem là một công cụ phù hợp để khai thác lượng lớn dữ liệu giáo dục, nhận diện các mẫu tiềm ẩn và hỗ trợ phát hiện sớm những sinh viên có nguy cơ.

Nghiên cứu sử dụng bộ dữ liệu Predict Students' Dropout and Academic Success của Realinho và cộng sự, bao gồm 4.424 sinh viên với 36 thuộc tính đầu vào và một biến mục tiêu, tương ứng tổng cộng 37 trường dữ liệu. Bài toán được xây dựng dưới dạng phân loại đa lớp với ba trạng thái: Dropout, Enrolled và Graduate. Các thuộc

tính phản ánh khá toàn diện hồ sơ của sinh viên, bao gồm tình trạng hôn nhân, phương thức và thứ tự đăng ký, ngành học, trình độ trước khi nhập học, điểm đầu vào, trình độ và nghề nghiệp của cha mẹ, tình trạng học phí, học bổng, giới tính, tuổi nhập học, cũng như kết quả học tập trong học kỳ thứ nhất và thứ hai. Ngoài ra, tập dữ liệu còn chứa các yếu tố kinh tế vĩ mô như tỷ lệ thất nghiệp, lạm phát và GDP.

Để đánh giá khả năng dự đoán, tác giả lựa chọn ba thuật toán gồm Decision Tree (DT), Random Forest (RF) và Artificial Neural Network (ANN). Đây là những phương pháp đại diện cho ba hướng tiếp cận khác nhau: cây quyết định có ưu điểm về khả năng diễn giải; Random Forest sử dụng tập hợp nhiều cây nhằm tăng tính ổn định và giảm nguy cơ quá khớp [38], [39]; trong khi ANN có khả năng học các quan hệ phi tuyến và các mẫu phức tạp trong dữ liệu [40]. Để tăng độ tin cậy của quá trình đánh giá, nghiên cứu áp dụng 10-fold cross-validation, trong đó dữ liệu lần lượt được chia thành mười tập con để mỗi phần đều được sử dụng cho cả huấn luyện và kiểm thử. Việc sử dụng cross-validation giúp hạn chế ảnh hưởng của cách chia dữ liệu ngẫu nhiên và cung cấp đánh giá ổn định hơn về khả năng tổng quát hóa của mô hình [41], [42]. Hiệu năng được đánh giá thông qua bốn chỉ số phổ biến gồm Accuracy, Precision, Recall và F-score.

Kết quả thực nghiệm cho thấy sự khác biệt đáng kể giữa ba thuật toán. Decision Tree đạt Accuracy 70,1%, Precision 70,0%, Recall 70,1% và F-score 70,0%. Random Forest cải thiện kết quả lên Accuracy 75,5%, Precision 73,9%, Recall 75,5% và F-score 74,2%. Trong khi đó, Artificial Neural Network đạt hiệu năng cao nhất, với Accuracy 77,3%, Precision 76,0%, Recall 77,3% và F-score 76,2%. Kết quả này cho thấy ANN có khả năng mô hình hóa tốt hơn các quan hệ phức tạp giữa các thuộc tính của sinh viên so với DT và RF trên tập dữ liệu được sử dụng.

Phân tích ma trận nhầm lẫn cũng cho thấy cả ba mô hình đều nhận diện lớp Graduate tốt hơn đáng kể so với lớp Enrolled. Với ANN, mô hình dự đoán đúng 2.012 trường hợp Graduate và 1.101 trường hợp Dropout, nhưng chỉ dự đoán đúng 307 trường hợp Enrolled. Tương tự, RF và DT cũng gặp nhiều khó khăn khi phân biệt sinh viên đang tiếp tục học với hai nhóm còn lại. Điều này cho thấy chỉ sử dụng Accuracy tổng thể chưa đủ để phản ánh đầy đủ chất lượng của mô hình trong bài toán phân loại trạng thái học tập; cần xem xét đồng thời kết quả theo từng lớp, đặc biệt đối với lớp Dropout nếu mục tiêu của hệ thống là cảnh báo sớm sinh viên có nguy cơ.

Từ kết quả thực nghiệm, tác giả kết luận rằng các mô hình học máy, đặc biệt là ANN, có tiềm năng hỗ trợ phân tích và dự đoán tình trạng học tập của sinh viên. Tuy nhiên, hiệu năng mô hình vẫn còn khả năng cải thiện. Các hướng nghiên cứu tiếp theo được đề xuất bao gồm tối ưu siêu tham số bằng Grid Search hoặc Random Search, chuẩn hóa và xử lý ngoại lệ, lựa chọn hoặc giảm số chiều đặc trưng, cũng như phát triển các mô hình Hybrid kết hợp nhiều thuật toán. Đây là điểm đáng chú ý đối với các nghiên cứu tiếp theo về dự đoán bỏ học, bởi kết quả của Sulak và Koklu cho thấy ANN đã vượt qua các thuật toán dựa trên cây, đồng thời tác giả trực tiếp đề xuất khai thác các mô hình lai nhằm tiếp tục nâng cao hiệu năng dự báo.

3.3. Đánh giá kết quả thử nghiệm các mô hình

Sau khi hoàn thành quá trình huấn luyện, ba mô hình học sâu gồm RNN, LSTM và Hybrid được đánh giá trên cùng một tập dữ liệu kiểm tra nhằm đảm bảo tính khách quan trong quá trình so sánh. Mục tiêu của thực nghiệm là đánh giá khả năng dự đoán nguy cơ sinh viên bỏ học của từng mô hình, đồng thời xác định kiến trúc phù hợp nhất đối với bài toán dự đoán trong lĩnh vực giáo dục đại học.

Để đánh giá hiệu năng của các mô hình một cách toàn diện, nghiên cứu sử dụng năm chỉ số phổ biến trong các bài toán phân loại gồm Accuracy, Precision, Recall, F1-score và ROC-AUC. Việc sử dụng đồng thời nhiều chỉ số giúp phản ánh đầy đủ chất lượng dự đoán của mô hình thay vì chỉ dựa trên một thước đo duy nhất. Đặc biệt, đối với bài toán dự đoán nguy cơ bỏ học, khả năng nhận diện chính xác các sinh viên thuộc nhóm rủi ro cao có ý nghĩa thực tiễn rất lớn, bởi đây là cơ sở để nhà trường triển khai các biện pháp hỗ trợ và can thiệp kịp thời.

Kết quả đánh giá của ba mô hình được trình bày trong Bảng 3.1

Table 3.1. Hiệu suất các mô hình

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Hybrid	90	90	90,58	90,43	94,36
LSTM	88	88,44	88,52	88,28	93,26
RNN	87	87	87,34	87,03	91,44
XGBoost	60	34	91	50	70,1
ANN	77,3	76	77,3	76,2	
Logistic Regression	66.59				72.9

Kết quả cho thấy cả ba mô hình được xây dựng trong nghiên cứu đều đạt hiệu năng cao. Trong đó, RNN đạt Accuracy 87%, Precision 87%, Recall 87,34%, F1-score 87,03% và ROC-AUC 91,44%. Mặc dù là kiến trúc đơn giản nhất trong ba mô hình, RNN vẫn cho thấy khả năng học được các quan hệ phi tuyến trong dữ liệu và đạt mức độ phân biệt tương đối tốt giữa các lớp mục tiêu. Tuy nhiên, hạn chế của RNN trong việc duy trì thông tin qua các bước xử lý và nguy cơ tiêu biến gradient có thể làm giảm khả năng biểu diễn những quan hệ phức tạp giữa các đặc trưng.

LSTM cải thiện kết quả so với RNN với Accuracy đạt 88%, Precision 88,44%, Recall 88,52%, F1-score 88,28% và ROC-AUC 93,26%. Sự gia tăng đồng thời của các chỉ số cho thấy cơ chế bộ nhớ và các cổng điều khiển của LSTM giúp duy trì thông tin quan trọng tốt hơn trong quá trình học. Đáng chú ý, ROC-AUC tăng từ 91,44% ở RNN lên 93,26% ở LSTM, phản ánh khả năng phân biệt các trạng thái đầu ra tốt hơn.

Trong ba kiến trúc được đề xuất, mô hình Hybrid đạt hiệu năng cao nhất, với Accuracy 90%, Precision 90%, Recall 90,58%, F1-score 90,43% và ROC-AUC 94,36%. So với RNN, Hybrid cải thiện Accuracy 3 điểm phần trăm và ROC-AUC 2,92 điểm phần trăm; so với LSTM, mức cải thiện tương ứng là 2 và 1,10 điểm phần trăm. Đáng chú ý hơn, Hybrid duy trì tương đối cân bằng giữa Precision và Recall, cho thấy việc cải thiện khả năng phát hiện sinh viên có nguy cơ không phải đánh đổi bằng sự suy giảm lớn về độ chính xác của các dự đoán.

Một cơ sở quan trọng để đánh giá kết quả của đề tài là nghiên cứu của Sulak và Koklu (2024). Khác với các nghiên cứu sử dụng nguồn dữ liệu khác, Sulak và Koklu sử dụng chính bộ dữ liệu Predict Students' Dropout and Academic Success gồm 4.424 sinh viên được khai thác từ UCI Machine Learning Repository [21], cũng là nguồn dữ liệu được sử dụng trong nghiên cứu này. Các tác giả xây dựng và so sánh Decision Tree, Random Forest và Artificial Neural Network (ANN), trong đó ANN đạt kết quả tốt nhất với Accuracy 77,3%, Precision 76,0%, Recall 77,3% và F1-score 76,2%.

Việc sử dụng cùng nguồn dữ liệu tạo ra một mốc tham chiếu có ý nghĩa để đánh giá hiệu quả của các kiến trúc được đề xuất. So với ANN của Sulak và Koklu, ngay cả mô hình RNN có hiệu năng thấp nhất trong ba mô hình của đề tài cũng đạt Accuracy cao hơn 9,7 điểm phần trăm và F1-score cao hơn 10,83 điểm phần trăm. Với LSTM, mức chênh lệch Accuracy tăng lên 10,7 điểm phần trăm, trong khi mô hình Hybrid đạt mức cải thiện lớn nhất, với Accuracy cao hơn 12,7 điểm phần trăm và F1-score cao hơn 14,23 điểm phần trăm so với ANN. Điều này cho thấy các kiến trúc học sâu được khảo sát trong đề tài có khả năng khai thác hiệu quả hơn các cấu trúc và quan hệ tiềm ẩn trong cùng nguồn dữ liệu sinh viên.

Đặc biệt, kết quả của Hybrid có ý nghĩa khi đặt trong bối cảnh chính kết luận của Sulak và Koklu. Trong nghiên cứu của họ, ANN đã đạt Accuracy 77,3%, cao hơn Random Forest (75,5%) và Decision Tree (70,1%). Các tác giả cho rằng khả năng học các quan hệ phi tuyến của ANN giúp mô hình xử lý dữ liệu phức tạp tốt hơn các phương pháp dựa trên cây; đồng thời đề xuất các nghiên cứu tiếp theo xem xét tối ưu siêu tham số, cải thiện tiền xử lý dữ liệu, lựa chọn đặc trưng và phát triển các phương pháp Hybrid nhằm tiếp tục nâng cao khả năng dự đoán. Kết quả của đề tài này phần nào củng cố định hướng đó, khi kiến trúc Hybrid đạt hiệu năng cao nhất trong các mô hình được khảo sát

Tuy nhiên, mức chênh lệch trên cần được diễn giải thận trọng. Mặc dù hai nghiên cứu sử dụng cùng nguồn dữ liệu 4.424 sinh viên, Sulak và Koklu thực hiện bài toán phân loại ba lớp gồm Dropout, Enrolled và Graduate, trong khi thiết lập thực nghiệm của đề tài tập trung vào hai lớp Dropout và Graduate. Ngoài ra, quy trình tiền xử lý, chiến lược phân chia dữ liệu, cấu hình mô hình và tối ưu siêu tham số giữa hai nghiên cứu không hoàn toàn giống nhau. Vì vậy, các kết quả trên nên được xem là đối chiếu trên cùng nguồn dữ liệu, thay vì một phép so sánh thực nghiệm hoàn toàn đồng nhất.

Xét riêng mô hình Hybrid, hiệu năng cao có thể được lý giải bởi cơ chế xử lý chuyên biệt cho các nhóm đặc trưng khác nhau. Thay vì đưa toàn bộ thuộc tính qua một luồng biểu diễn thống nhất, mô hình phân tách các biến số và biến phân loại. Nhóm đặc trưng số được xử lý thông qua các lớp Dense nhằm khai thác quan hệ định lượng, trong khi các đặc trưng phân loại được biểu diễn thông qua Embedding để học các quan hệ tiềm ẩn giữa các giá trị. Hai không gian đặc trưng sau đó được hợp nhất trước khi thực hiện phân loại.

Cách tổ chức này phù hợp với đặc điểm dị thể của dữ liệu giáo dục, trong đó nguy cơ bỏ học thường không được quyết định bởi một biến đơn lẻ mà hình thành từ sự tương tác giữa nhiều nhóm yếu tố như kết quả học tập, đặc điểm cá nhân, hoàn cảnh kinh tế và điều kiện gia đình. Khả năng học đồng thời biểu diễn riêng của từng nhóm dữ liệu và tương tác giữa chúng có thể là một trong những nguyên nhân giúp Hybrid đạt Accuracy 90% và ROC-AUC 94,36%, cao nhất trong ba mô hình được khảo sát.

Khi mở rộng đối chiếu với các phương pháp khác, kết quả cũng cho thấy ưu thế của các mô hình được xây dựng trong đề tài. Logistic Regression trong nghiên cứu của Aulck và cộng sự đạt Accuracy khoảng 66,59% và ROC-AUC 72,9%, trong khi mô hình XGBoost được tham chiếu đạt Accuracy 60% nhưng Recall đối với lớp Dropout lên tới 91%. Trường hợp XGBoost cho thấy một điểm đáng chú ý: Accuracy thấp không đồng nghĩa mô hình không có giá trị, bởi Recall cao có thể phù hợp với hệ thống ưu tiên hạn chế bỏ sót sinh viên có nguy cơ. Do đó, việc đánh giá mô hình dự đoán bỏ học cần dựa trên nhiều chỉ số thay vì chỉ so sánh Accuracy.

Nhìn chung, kết quả thực nghiệm đã chứng minh tính hiệu quả của các mô hình học sâu trong bài toán dự đoán nguy cơ sinh viên bỏ học. Trong số các mô hình được khảo sát, mô hình Hybrid đạt hiệu năng tốt nhất trên tất cả các chỉ số đánh giá quan trọng, đồng thời thể hiện khả năng phân biệt giữa các lớp mục tiêu tốt hơn thông qua chỉ số ROC-AUC cao nhất. Những kết quả này khẳng định rằng việc kết hợp các kỹ thuật học biểu diễn chuyên biệt cho từng loại dữ liệu cùng cơ chế hợp nhất đặc trưng là một hướng tiếp cận phù hợp đối với các bộ dữ liệu giáo dục có tính dị thể. Đồng thời, kết quả nghiên cứu cũng cho thấy tiềm năng ứng dụng của mô hình Hybrid trong các hệ thống cảnh báo sớm, hỗ trợ các cơ sở giáo dục nhận diện và can thiệp kịp thời đối với những sinh viên có nguy cơ bỏ học cao.

3.4. Phân tích các đặc trưng quan trọng của dữ liệu

Bên cạnh việc đánh giá hiệu năng của các mô hình dự đoán, việc xác định các đặc trưng có ảnh hưởng lớn đến nguy cơ bỏ học của sinh viên cũng là một mục tiêu quan trọng của nghiên cứu. Kết quả này không chỉ giúp giải thích hành vi của mô hình mà còn cung cấp những thông tin có giá trị cho các nhà quản lý giáo dục trong quá trình xây dựng các chính sách hỗ trợ sinh viên.

Thông qua quá trình huấn luyện và phân tích dữ liệu, nghiên cứu nhận thấy các đặc trưng liên quan đến kết quả học tập và tình trạng tài chính là những yếu tố có ảnh hưởng mạnh nhất đến khả năng bỏ học của sinh viên. Điều này phù hợp với nhiều nghiên cứu trước đây trong lĩnh vực Educational Data Mining khi cho rằng kết quả học tập và điều kiện tài chính là hai nhóm nguyên nhân quan trọng nhất dẫn đến hiện tượng sinh viên rời bỏ chương trình đào tạo [7], [8], [17].

Table 3.2. Các đặc trưng có mức độ ảnh hưởng cao đến kết quả dự đoán

STT	Đặc trưng
1	Curricular units 2nd sem (approved)
2	Curricular units 2nd sem (grade)
3	Curricular units 1st sem (approved)
4	Curricular units 1st sem (grade)
5	Tuition fees up to date
6	Scholarship holder
7	Admission grade
8	Debtor
9	Age at enrollment

Kết quả cho thấy nhóm đặc trưng có ảnh hưởng lớn nhất thuộc về kết quả học tập trong năm học đầu tiên. Đặc biệt, số lượng học phần được hoàn thành và điểm số

của học kỳ thứ nhất và học kỳ thứ hai xuất hiện trong nhóm các đặc trưng quan trọng nhất. Điều này phản ánh rằng thành tích học tập ban đầu đóng vai trò quyết định đối với khả năng tiếp tục theo học của sinh viên.

Trong đó, biến Curricular units 2nd sem (approved) là đặc trưng có mức độ ảnh hưởng cao nhất đến kết quả dự đoán. Số lượng học phần được hoàn thành thành công trong học kỳ thứ hai phản ánh trực tiếp mức độ thích nghi của sinh viên với môi trường đại học cũng như khả năng đáp ứng yêu cầu học tập của chương trình đào tạo. Những sinh viên hoàn thành ít học phần hoặc thường xuyên trượt môn có xu hướng đối mặt với nguy cơ bỏ học cao hơn đáng kể so với các sinh viên còn lại.

Bên cạnh đó, điểm số trung bình của các học phần trong học kỳ thứ nhất và học kỳ thứ hai cũng xuất hiện trong nhóm đặc trưng quan trọng nhất. Kết quả này cho thấy chất lượng học tập không chỉ được phản ánh qua việc hoàn thành môn học mà còn thể hiện thông qua mức độ thành công trong từng học phần. Sinh viên có kết quả học tập thấp thường gặp nhiều khó khăn trong việc duy trì động lực học tập, dễ bị cảnh báo học vụ và có khả năng bỏ học cao hơn.

Ngoài các yếu tố học thuật, nhóm đặc trưng liên quan đến tình trạng tài chính cũng thể hiện mức độ ảnh hưởng đáng kể. Biến Tuition fees up to date phản ánh tình trạng hoàn thành nghĩa vụ học phí của sinh viên. Kết quả phân tích cho thấy những sinh viên gặp khó khăn trong việc đóng học phí đúng hạn thường có nguy cơ bỏ học cao hơn. Điều này phù hợp với các nghiên cứu về duy trì người học khi yếu tố tài chính được xem là một trong những nguyên nhân quan trọng dẫn đến việc gián đoạn hoặc chấm dứt quá trình học tập [8].

Tương tự, biến Debtor cũng xuất hiện trong nhóm các đặc trưng có ảnh hưởng lớn. Sự xuất hiện đồng thời của hai biến liên quan đến tài chính cho thấy hoàn cảnh kinh tế là một trong những yếu tố có khả năng tác động trực tiếp đến quyết định tiếp tục hay rời bỏ chương trình học của sinh viên.

Một kết quả đáng chú ý khác là sự xuất hiện của biến Scholarship holder trong nhóm đặc trưng quan trọng. Sinh viên nhận học bổng thường có động lực học tập cao hơn, đồng thời nhận được sự hỗ trợ tài chính nhất định từ nhà trường hoặc các tổ chức tài trợ. Điều này góp phần làm giảm nguy cơ bỏ học và tăng khả năng hoàn thành chương trình đào tạo.

Biến Admission grade cũng được xác định là một yếu tố có ảnh hưởng đáng kể đến kết quả dự đoán. Điểm đầu vào phản ánh nền tảng học tập của sinh viên trước khi bước vào môi trường đại học. Những sinh viên có điểm đầu vào cao thường có khả năng thích nghi và tiếp thu kiến thức tốt hơn, từ đó giảm nguy cơ gặp khó khăn trong quá trình học tập.

Cuối cùng, biến Age at enrollment cho thấy độ tuổi nhập học cũng có mối liên hệ nhất định với khả năng duy trì học tập. Sự khác biệt về độ tuổi có thể phản ánh những khác biệt về hoàn cảnh cá nhân, trách nhiệm gia đình, điều kiện kinh tế hoặc kinh nghiệm sống, từ đó ảnh hưởng đến quyết định tiếp tục theo học của sinh viên.

Nhìn chung, kết quả phân tích đặc trưng cho thấy các yếu tố học thuật và tài chính là hai nhóm nhân tố có ảnh hưởng mạnh nhất đến nguy cơ bỏ học. Điều này hoàn toàn phù hợp với các cơ sở lý thuyết và các nghiên cứu trước đây đã được trình bày trong Chương 1. Đồng thời, kết quả cũng chứng minh rằng các mô hình học sâu không chỉ có khả năng dự đoán chính xác mà còn có thể hỗ trợ xác định các yếu tố rủi ro quan trọng, góp phần xây dựng các hệ thống cảnh báo sớm và hỗ trợ ra quyết định trong giáo dục đại học.

3.5. Thảo luận

3.5.1. Ý nghĩa thực tế của kết quả

Kết quả thực nghiệm của nghiên cứu cho thấy các mô hình học sâu có khả năng dự đoán hiệu quả nguy cơ sinh viên bỏ học dựa trên các dữ liệu học tập và thông tin cá nhân liên quan. Đặc biệt, mô hình Hybrid đạt độ chính xác 90% và chỉ số ROC-AUC 94,36%, cho thấy khả năng phân biệt rất tốt giữa nhóm sinh viên có nguy cơ bỏ học và nhóm sinh viên có khả năng hoàn thành chương trình đào tạo. Kết quả này không chỉ khẳng định tính khả thi của việc ứng dụng các phương pháp học sâu trong lĩnh vực giáo dục mà còn mở ra nhiều cơ hội ứng dụng thực tiễn trong công tác quản lý đào tạo và hỗ trợ người học.

Một trong những giá trị quan trọng nhất của nghiên cứu là khả năng xây dựng các hệ thống cảnh báo sớm dành cho sinh viên có nguy cơ bỏ học. Trong môi trường giáo dục đại học truyền thống, việc phát hiện các trường hợp có nguy cơ thường được thực hiện thông qua các hình thức theo dõi thủ công hoặc dựa trên kết quả học tập cuối kỳ. Tuy nhiên, khi các dấu hiệu này xuất hiện rõ ràng thì sinh viên thường đã gặp nhiều khó khăn trong học tập hoặc đã có ý định rời bỏ chương trình đào tạo. Điều

này khiến các biện pháp hỗ trợ được triển khai muộn và hiệu quả can thiệp không cao.

Ngược lại, với sự hỗ trợ của các mô hình dự đoán, nguy cơ bỏ học có thể được nhận diện ngay từ những giai đoạn đầu của quá trình học tập. Thông qua việc phân tích dữ liệu học tập, thông tin nhân khẩu học và điều kiện kinh tế - xã hội của sinh viên, hệ thống có thể tự động xác định các trường hợp có khả năng gặp rủi ro cao. Trên cơ sở đó, nhà trường có thể chủ động triển khai các biện pháp hỗ trợ phù hợp như tư vấn học tập, hỗ trợ tâm lý, hỗ trợ tài chính hoặc điều chỉnh lộ trình học tập. Việc can thiệp sớm không chỉ giúp giảm tỷ lệ bỏ học mà còn góp phần nâng cao chất lượng đào tạo và cải thiện trải nghiệm học tập của sinh viên.

Kết quả nghiên cứu cũng mang ý nghĩa quan trọng đối với công tác quản lý đào tạo trong bối cảnh chuyển đổi số giáo dục hiện nay. Thay vì dựa hoàn toàn vào kinh nghiệm hoặc các đánh giá mang tính định tính, các quyết định quản lý có thể được hỗ trợ bởi các bằng chứng định lượng được khai thác từ dữ liệu. Đây là xu hướng đang được nhiều trường đại học trên thế giới áp dụng nhằm xây dựng mô hình quản trị đại học dựa trên dữ liệu. Trong mô hình này, dữ liệu không chỉ được sử dụng để lưu trữ thông tin mà còn trở thành nguồn tri thức phục vụ cho việc dự báo, phân tích và hỗ trợ ra quyết định.

Một kết quả đáng chú ý khác của nghiên cứu là việc xác định được các nhóm đặc trưng có ảnh hưởng lớn nhất đến nguy cơ bỏ học. Các yếu tố học thuật như số học phần hoàn thành, kết quả học tập trong các học kỳ đầu tiên và điểm tuyển sinh đầu vào được xác định là những đặc trưng quan trọng nhất. Điều này cho thấy quá trình thích nghi học tập trong năm đầu tiên đóng vai trò quyết định đối với khả năng duy trì học tập của sinh viên. Bên cạnh đó, các yếu tố tài chính như tình trạng đóng học phí, tình trạng công nợ học phí và học bổng cũng xuất hiện trong nhóm các đặc trưng có ảnh hưởng lớn. Kết quả này hoàn toàn phù hợp với các nghiên cứu lý thuyết đã được trình bày ở Chương 1, đồng thời cung cấp thêm cơ sở thực tiễn cho việc xây dựng các chính sách hỗ trợ sinh viên.

Từ góc độ quản lý, các kết quả trên cho phép nhà trường phân bổ nguồn lực hỗ trợ một cách hiệu quả hơn. Thay vì triển khai các chương trình hỗ trợ đại trà cho toàn bộ sinh viên, các đơn vị quản lý có thể tập trung nguồn lực vào các nhóm đối tượng có nguy cơ cao. Chẳng hạn, những sinh viên có kết quả học tập giảm sút trong các

học kỳ đầu tiên có thể được đưa vào các chương trình cố vấn học tập chuyên sâu; những sinh viên gặp khó khăn tài chính có thể được ưu tiên tiếp cận các nguồn học bổng hoặc chính sách hỗ trợ học phí. Cách tiếp cận này không chỉ nâng cao hiệu quả sử dụng nguồn lực mà còn giúp tăng khả năng thành công của các biện pháp can thiệp.

Ngoài ý nghĩa đối với công tác quản lý đào tạo, nghiên cứu còn góp phần khẳng định tiềm năng ứng dụng của trí tuệ nhân tạo trong lĩnh vực giáo dục. So với các phương pháp học máy truyền thống, các mô hình học sâu cho thấy khả năng khai thác hiệu quả các mối quan hệ phi tuyến và các tương tác phức tạp giữa nhiều nhóm đặc trưng khác nhau. Đặc biệt, mô hình Hybrid được đề xuất trong nghiên cứu đã chứng minh hiệu quả khi xử lý đồng thời dữ liệu số và dữ liệu phân loại thông qua cơ chế học biểu diễn chuyên biệt và hợp nhất đặc trưng. Đây là một hướng tiếp cận phù hợp đối với các bộ dữ liệu giáo dục hiện đại vốn có tính đa dạng và dị thể cao.

Bên cạnh đó, kiến trúc Hybrid còn cho thấy khả năng mở rộng tốt trong tương lai. Khi các trường đại học ngày càng triển khai mạnh mẽ các hệ thống quản lý học tập trực tuyến, lượng dữ liệu thu thập được từ LMS, hệ thống điểm danh điện tử, nền tảng học tập số và các hệ thống quản lý sinh viên sẽ ngày càng phong phú hơn. Những nguồn dữ liệu này hoàn toàn có thể được tích hợp vào mô hình nhằm nâng cao hơn nữa khả năng dự đoán. Điều này mở ra triển vọng xây dựng các hệ thống phân tích học tập thông minh có khả năng theo dõi liên tục tình trạng học tập của sinh viên theo thời gian thực.

Kết quả của nghiên cứu không chỉ dừng lại ở việc xây dựng một mô hình dự đoán có độ chính xác cao mà còn góp phần khẳng định vai trò của dữ liệu và trí tuệ nhân tạo trong quá trình đổi mới quản lý giáo dục đại học. Nếu được triển khai trong môi trường thực tế, mô hình có thể trở thành một công cụ hỗ trợ hữu ích cho các nhà quản lý, giảng viên và cố vấn học tập trong việc phát hiện sớm các trường hợp có nguy cơ bỏ học, từ đó nâng cao hiệu quả hỗ trợ người học, cải thiện tỷ lệ tốt nghiệp và góp phần nâng cao chất lượng đào tạo của nhà trường.

3.5.2. Những hạn chế của mô hình

Mặc dù các kết quả thực nghiệm cho thấy mô hình Hybrid đạt hiệu năng dự đoán cao và vượt trội hơn so với các mô hình RNN và LSTM, nghiên cứu vẫn nhận thấy một số hạn chế cần lưu ý khi đánh giá khả năng ứng dụng thực tế của mô hình.

Việc nhận diện các hạn chế này không chỉ giúp đánh giá khách quan kết quả nghiên cứu mà còn là cơ sở quan trọng để định hướng các nghiên cứu tiếp theo.

Hạn chế về dữ liệu: Bộ dữ liệu sử dụng trong nghiên cứu, "Predicting Student Dropout and Academic Success", bao gồm 4.424 bản ghi sinh viên cùng các đặc trưng học thuật, nhân khẩu học và kinh tế – xã hội. Mặc dù bộ dữ liệu này có quy mô tương đối lớn và được sử dụng phổ biến trong các nghiên cứu về dự đoán bỏ học, nó vẫn chỉ phản ánh đặc điểm của một môi trường giáo dục cụ thể tại một thời điểm nhất định. Do đó, khả năng tổng quát hóa của mô hình đối với các trường đại học hoặc hệ thống giáo dục khác vẫn cần được kiểm chứng thông qua các nghiên cứu mở rộng trên nhiều nguồn dữ liệu độc lập.

Hạn chế về loại dữ liệu và mối quan hệ giữa các biến: Dữ liệu đầu vào chủ yếu là tĩnh, tập trung vào các thông tin định lượng như kết quả học tập, thông tin cá nhân và điều kiện kinh tế – xã hội. Tuy nhiên, quyết định bỏ học của sinh viên thường chịu ảnh hưởng từ nhiều yếu tố phức tạp khác, bao gồm tâm lý, môi trường học tập và các mối quan hệ xã hội. Các yếu tố này chưa được phản ánh đầy đủ trong bộ dữ liệu hiện tại. Do đó, mặc dù mô hình đạt độ chính xác cao, kết quả dự đoán vẫn chưa thể bao quát toàn diện tất cả các nguyên nhân dẫn đến hiện tượng bỏ học.

Hạn chế về mô hình RNN và LSTM: Cần thừa nhận rằng bộ dữ liệu nghiên cứu không phải là dữ liệu chuỗi thời gian thuần túy. Tuy nhiên, các biến đầu vào vẫn có sự liên hệ theo tiến trình học tập của sinh viên, đặc biệt là các biến thuộc học kỳ 1 và học kỳ 2. Vì vậy, RNN và LSTM được thử nghiệm nhằm đánh giá khả năng học các mối quan hệ phụ thuộc giữa các đặc trưng học tập theo tiến trình đào tạo. Qua quá trình thử nghiệm, nghiên cứu nhận thấy RNN và LSTM có ưu thế trong việc học các mối quan hệ tuần tự, nhưng chưa khai thác tối ưu các nhóm đặc trưng không mang tính tuần tự như thông tin nhân khẩu học, tài chính hay học bổng. Đây chính là cơ sở để đề xuất mô hình Hybrid, nơi dữ liệu số và dữ liệu phân loại được xử lý riêng biệt trước khi hợp nhất, giúp khai thác triệt để các đặc trưng đa dạng của bộ dữ liệu.

Hạn chế về giải thích kết quả: Các mô hình học sâu, bao gồm RNN, LSTM và Hybrid, dựa trên hàng nghìn tham số được tối ưu hóa trong quá trình huấn luyện. Điều này giúp mô hình học các mối quan hệ phi tuyến phức tạp nhưng đồng thời giảm khả năng giải thích kết quả dự đoán, dẫn đến vấn đề "hộp đen" trong học sâu. Trong

bối cảnh giáo dục, nơi quyết định can thiệp cần minh bạch, hạn chế này ảnh hưởng đến mức độ tin cậy và khả năng áp dụng

Hạn chế về dữ liệu tĩnh: Nghiên cứu sử dụng dữ liệu thu thập tại một thời điểm nhất định, trong khi hành vi học tập của sinh viên liên tục thay đổi. Một sinh viên có thể cải thiện hoặc gặp khó khăn ở các học kỳ tiếp theo, điều này không được phản ánh đầy đủ bởi dữ liệu tĩnh. Việc tích hợp dữ liệu chuỗi thời gian hoặc dữ liệu thời gian thực sẽ giúp mô hình dự báo chính xác hơn và có thể triển khai hệ thống cảnh báo sớm.

Hạn chế về triển khai thực tế: Dữ liệu giáo dục thực tế thường gặp các vấn đề như thiếu dữ liệu, dữ liệu không đồng nhất giữa các hệ thống, hoặc dữ liệu cập nhật không liên tục. Do đó, bên cạnh cải thiện thuật toán dự đoán, việc xây dựng quy trình quản trị dữ liệu và đảm bảo chất lượng đầu vào là cần thiết cho các hệ thống phân tích học tập trong tương lai.

Mặc dù mô hình đạt hiệu năng cao và hỗ trợ nhận diện sớm sinh viên có nguy cơ bỏ học, kết quả dự đoán nên được kết hợp với đánh giá của giảng viên, cố vấn học tập và các chuyên gia giáo dục để đảm bảo quyết định can thiệp chính xác và toàn diện. Những hạn chế này đồng thời mở ra nhiều hướng nghiên cứu tiềm năng, như mở rộng dữ liệu, tích hợp dữ liệu hành vi từ LMS, áp dụng các mô hình học sâu tiên tiến, và sử dụng Explainable AI nhằm nâng cao cả hiệu năng và khả năng giải thích của hệ thống.

3.6. Kết luận chương

Chương 3 đã trình bày quá trình thực nghiệm và đánh giá hiệu quả của các mô hình học sâu trong bài toán dự đoán nguy cơ sinh viên bỏ học. Trên cơ sở bộ dữ liệu "Predicting Student Dropout and Academic Success", nghiên cứu đã tiến hành xây dựng, huấn luyện và đánh giá ba mô hình gồm RNN, LSTM và Hybrid nhằm xác định kiến trúc phù hợp nhất cho bài toán dự đoán trong lĩnh vực giáo dục.

Kết quả thực nghiệm cho thấy cả ba mô hình đều đạt được hiệu năng dự đoán tương đối cao, chứng minh khả năng ứng dụng của các phương pháp học sâu trong việc khai thác dữ liệu giáo dục. Trong đó, mô hình RNN đạt độ chính xác 87%, mô hình LSTM đạt độ chính xác 88%, trong khi mô hình Hybrid đạt kết quả tốt nhất với độ chính xác 90% và chỉ số ROC-AUC đạt 94,36%. Sự cải thiện về hiệu năng của mô hình Hybrid cho thấy việc xử lý riêng biệt các thuộc tính số và thuộc tính phân

loại trước khi thực hiện hợp nhất đặc trưng đã giúp khai thác hiệu quả hơn các nguồn thông tin khác nhau trong bộ dữ liệu.

Bên cạnh việc đánh giá hiệu năng dự đoán, nghiên cứu cũng tiến hành phân tích các đặc trưng quan trọng ảnh hưởng đến nguy cơ bỏ học của sinh viên. Kết quả cho thấy các yếu tố học thuật như số học phần hoàn thành, kết quả học tập trong các học kỳ đầu tiên và điểm tuyển sinh đầu vào đóng vai trò đặc biệt quan trọng. Ngoài ra, các yếu tố tài chính như tình trạng đóng học phí, công nợ học phí và học bổng cũng được xác định là những nhân tố có ảnh hưởng đáng kể đến khả năng duy trì học tập của sinh viên. Những kết quả này phù hợp với các cơ sở lý thuyết và các nghiên cứu trước đây, đồng thời cung cấp thêm bằng chứng thực nghiệm về vai trò của các yếu tố học thuật và kinh tế trong hiện tượng bỏ học ở bậc đại học.

Thông qua quá trình phân tích và thảo luận kết quả, nghiên cứu đã cho thấy tiềm năng ứng dụng của các mô hình học sâu trong việc xây dựng các hệ thống cảnh báo sớm dành cho sinh viên có nguy cơ bỏ học. Việc nhận diện sớm các trường hợp rủi ro có thể hỗ trợ nhà trường triển khai các biện pháp can thiệp kịp thời, góp phần nâng cao tỷ lệ duy trì học tập, cải thiện chất lượng đào tạo và tối ưu hóa công tác quản lý người học. Đồng thời, kết quả nghiên cứu cũng khẳng định tính hiệu quả của hướng tiếp cận dựa trên dữ liệu và trí tuệ nhân tạo trong bối cảnh chuyển đổi số giáo dục hiện nay.

Mặc dù còn tồn tại một số hạn chế liên quan đến dữ liệu, khả năng tổng quát hóa và tính giải thích của mô hình, các kết quả đạt được đã chứng minh tính khả thi của việc ứng dụng học sâu vào bài toán dự đoán nguy cơ sinh viên bỏ học. Đây là cơ sở quan trọng để tiếp tục mở rộng nghiên cứu theo hướng tích hợp thêm các nguồn dữ liệu hành vi học tập, dữ liệu từ hệ thống LMS và các phương pháp trí tuệ nhân tạo có khả năng giải thích nhằm nâng cao hơn nữa hiệu quả và tính ứng dụng của hệ thống trong môi trường giáo dục thực tế.

CHƯƠNG 4: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

- Giới thiệu chương

Sau khi hoàn thành quá trình nghiên cứu lý thuyết, xây dựng mô hình và đánh giá thực nghiệm, Chương 4 trình bày các kết luận chính được rút ra từ nghiên cứu về ứng dụng các mô hình học sâu trong dự đoán nguy cơ sinh viên bỏ học. Trên cơ sở các kết quả đạt được, chương tiến hành tổng hợp những đóng góp khoa học và giá trị thực tiễn của nghiên cứu đối với lĩnh vực khai phá dữ liệu giáo dục và phân tích học tập. Đồng thời, chương cũng chỉ ra một số hạn chế còn tồn tại và đề xuất các hướng nghiên cứu trong tương lai nhằm nâng cao hiệu quả dự đoán cũng như mở rộng khả năng ứng dụng của mô hình trong môi trường giáo dục thực tế. Những nội dung được trình bày trong chương không chỉ khép lại quá trình nghiên cứu mà còn là cơ sở định hướng cho các nghiên cứu tiếp theo về hệ thống cảnh báo sớm và hỗ trợ ra quyết định trong giáo dục đại học.

4.1. Kết luận chính

Nghiên cứu được thực hiện với mục tiêu xây dựng và đánh giá các mô hình học sâu nhằm dự đoán nguy cơ sinh viên bỏ học dựa trên dữ liệu giáo dục. Trong bối cảnh tỷ lệ bỏ học tại các cơ sở giáo dục đại học vẫn là một vấn đề đáng quan tâm, việc ứng dụng các kỹ thuật trí tuệ nhân tạo để nhận diện sớm sinh viên có nguy cơ bỏ học được xem là một hướng tiếp cận có ý nghĩa cả về mặt khoa học lẫn thực tiễn. Trên cơ sở đó, nghiên cứu đã tiến hành khảo sát các nguyên nhân dẫn đến hiện tượng bỏ học, tổng quan các phương pháp học máy và học sâu liên quan, đồng thời xây dựng các mô hình dự đoán dựa trên dữ liệu sinh viên.

Bộ dữ liệu được sử dụng trong nghiên cứu là bộ dữ liệu "Predicting Student Dropout and Academic Success", bao gồm 4.424 bản ghi sinh viên với 35 thuộc tính phản ánh các yếu tố học thuật, nhân khẩu học và kinh tế - xã hội. Trước khi đưa vào huấn luyện, dữ liệu đã được làm sạch, chuẩn hóa và chuyển đổi phù hợp với yêu cầu của các mô hình học sâu. Quá trình tiền xử lý giúp nâng cao chất lượng dữ liệu đầu vào, giảm nhiễu và tạo điều kiện để các mô hình học được các đặc trưng quan trọng phục vụ cho quá trình dự đoán.

Trên cơ sở dữ liệu đã được xử lý, nghiên cứu đã xây dựng và đánh giá ba mô hình gồm RNN, LSTM và Hybrid. Trong đó, RNN được sử dụng như mô hình cơ sở nhằm khai thác khả năng học các quan hệ tuần tự trong dữ liệu. LSTM được triển khai nhằm khắc phục hạn chế về khả năng ghi nhớ dài hạn của RNN thông qua cơ chế bộ nhớ và các cổng điều khiển thông tin. Bên cạnh đó, nghiên cứu đề xuất mô hình Hybrid với kiến trúc đa nhánh, cho phép xử lý riêng biệt dữ liệu số và dữ liệu phân loại trước khi thực hiện hợp nhất đặc trưng để đưa ra dự đoán cuối cùng.

Kết quả thực nghiệm cho thấy cả ba mô hình đều đạt hiệu năng dự đoán cao, khẳng định tính phù hợp của các phương pháp học sâu đối với bài toán dự đoán nguy cơ sinh viên bỏ học. Trong đó, mô hình RNN đạt độ chính xác 87%, mô hình LSTM đạt độ chính xác 88%, trong khi mô hình Hybrid đạt kết quả tốt nhất với Accuracy đạt 90%, F1-score đạt 90,43% và ROC-AUC đạt 94,36%. Kết quả này cho thấy việc xử lý chuyên biệt các nhóm dữ liệu khác nhau trước khi thực hiện hợp nhất đặc trưng giúp mô hình khai thác hiệu quả hơn các nguồn thông tin trong dữ liệu giáo dục, từ đó nâng cao khả năng dự đoán.

Ngoài việc đánh giá hiệu năng của các mô hình, nghiên cứu còn tiến hành phân tích các đặc trưng có ảnh hưởng lớn đến nguy cơ bỏ học của sinh viên. Kết quả cho thấy các yếu tố liên quan đến kết quả học tập trong các học kỳ đầu tiên, số học phần hoàn thành, tình trạng đóng học phí, học bổng và điều kiện tài chính là những đặc trưng có ảnh hưởng mạnh nhất đến kết quả dự đoán. Điều này cho thấy nguy cơ bỏ học không xuất phát từ một nguyên nhân đơn lẻ mà là kết quả của sự tác động đồng thời giữa nhiều nhóm yếu tố khác nhau, trong đó các yếu tố học thuật và tài chính đóng vai trò đặc biệt quan trọng.

Những kết quả đạt được không chỉ khẳng định hiệu quả của các mô hình học sâu trong lĩnh vực khai phá dữ liệu giáo dục mà còn cho thấy tiềm năng ứng dụng của các hệ thống dự đoán trong thực tế. Nếu được triển khai tại các cơ sở giáo dục đại học, mô hình có thể hỗ trợ xây dựng các hệ thống cảnh báo sớm, giúp nhà trường nhận diện kịp thời các sinh viên có nguy cơ bỏ học và triển khai các biện pháp hỗ trợ phù hợp. Điều này góp phần nâng cao tỷ lệ duy trì học tập, giảm tỷ lệ bỏ học và cải thiện chất lượng đào tạo.

Tóm lại, nghiên cứu đã hoàn thành các mục tiêu đề ra, từ việc xây dựng cơ sở lý thuyết, phát triển mô hình dự đoán đến đánh giá hiệu quả thực nghiệm. Kết quả

nghiên cứu cho thấy mô hình Hybrid là giải pháp phù hợp nhất trong số các mô hình được khảo sát, đồng thời khẳng định tiềm năng của việc ứng dụng học sâu trong các hệ thống hỗ trợ ra quyết định và phân tích học tập trong giáo dục đại học.

4.2. Đóng góp khoa học

Nghiên cứu đã đạt được một số đóng góp nhất định về mặt học thuật và thực tiễn trong lĩnh vực khai phá dữ liệu giáo dục và phân tích học tập. Những đóng góp này không chỉ thể hiện ở việc xây dựng thành công các mô hình dự đoán nguy cơ sinh viên bỏ học mà còn ở việc đề xuất hướng tiếp cận phù hợp đối với dữ liệu giáo dục có tính chất đa dạng và phức tạp.

Thứ nhất, nghiên cứu đã hệ thống hóa cơ sở lý thuyết liên quan đến bài toán dự đoán sinh viên bỏ học, bao gồm các nguyên nhân dẫn đến hiện tượng bỏ học, các nhóm yếu tố ảnh hưởng và các phương pháp học máy, học sâu thường được sử dụng trong lĩnh vực giáo dục. Trên cơ sở tổng hợp các nghiên cứu trước đây, luận văn đã làm rõ mối quan hệ giữa các yếu tố học thuật, tài chính, tâm lý và xã hội đối với khả năng duy trì học tập của sinh viên. Đây là cơ sở quan trọng để lựa chọn đặc trưng đầu vào và xây dựng các mô hình dự đoán phù hợp.

Thứ hai, nghiên cứu đã xây dựng và đánh giá thành công ba mô hình học sâu gồm RNN, LSTM và Hybrid trên cùng một bộ dữ liệu giáo dục. Việc triển khai đồng thời nhiều kiến trúc mạng khác nhau cho phép đánh giá khách quan hiệu quả của từng phương pháp, đồng thời cung cấp cơ sở thực nghiệm cho việc lựa chọn mô hình phù hợp đối với bài toán dự đoán nguy cơ bỏ học. Kết quả nghiên cứu cho thấy các mô hình học sâu đều đạt hiệu năng cao và vượt trội so với nhiều phương pháp học máy truyền thống đã được công bố trước đây.

Thứ ba, đóng góp quan trọng nhất của nghiên cứu là việc đề xuất và xây dựng mô hình Hybrid theo kiến trúc đa nhánh nhằm xử lý riêng biệt các nhóm dữ liệu có bản chất khác nhau. Trong mô hình này, các thuộc tính số được xử lý thông qua mạng nơ-ron truyền thẳng nhiều tầng, trong khi các thuộc tính phân loại được biểu diễn bằng kỹ thuật Embedding trước khi thực hiện hợp nhất đặc trưng. Cách tiếp cận này cho phép mô hình khai thác hiệu quả hơn các đặc điểm riêng của từng loại dữ liệu, đồng thời học được các tương tác phức tạp giữa nhiều nhóm đặc trưng khác nhau.

Kết quả thực nghiệm cho thấy mô hình Hybrid đạt độ chính xác 90% và ROC-AUC đạt 94,36%, cao hơn so với các mô hình RNN và LSTM được khảo sát trong

nghiên cứu. Điều này chứng minh tính hiệu quả của kiến trúc được đề xuất và cho thấy việc xử lý chuyên biệt từng nhóm đặc trưng trước khi hợp nhất là một hướng tiếp cận phù hợp đối với các bộ dữ liệu giáo dục có tính dị thể cao.

Thứ tư, nghiên cứu đã xác định được các đặc trưng có ảnh hưởng lớn nhất đến nguy cơ bỏ học của sinh viên. Kết quả cho thấy các yếu tố liên quan đến kết quả học tập trong các học kỳ đầu tiên, số học phần hoàn thành, tình trạng đóng học phí, học bổng và điều kiện tài chính là những đặc trưng quan trọng nhất. Những phát hiện này không chỉ giúp nâng cao khả năng giải thích của mô hình mà còn cung cấp thêm bằng chứng thực nghiệm cho các nghiên cứu về hành vi học tập và duy trì người học trong giáo dục đại học.

Thứ năm, nghiên cứu góp phần khẳng định khả năng ứng dụng của các kỹ thuật học sâu trong việc xây dựng các hệ thống cảnh báo sớm dành cho sinh viên có nguy cơ bỏ học. Thay vì chỉ dừng lại ở việc phân tích dữ liệu lịch sử, mô hình được đề xuất có thể được sử dụng như một công cụ hỗ trợ ra quyết định, giúp các cơ sở giáo dục nhận diện sớm các trường hợp có nguy cơ cao và triển khai các biện pháp hỗ trợ phù hợp. Đây là một hướng ứng dụng có ý nghĩa thực tiễn trong bối cảnh chuyển đổi số giáo dục và phát triển các hệ thống quản lý đào tạo dựa trên dữ liệu.

Nhìn chung, đóng góp nổi bật nhất của nghiên cứu là đề xuất thành công mô hình Hybrid dành cho bài toán dự đoán nguy cơ sinh viên bỏ học, đồng thời chứng minh được hiệu quả của mô hình thông qua các kết quả thực nghiệm. Những kết quả đạt được không chỉ có giá trị tham khảo đối với các nghiên cứu tiếp theo trong lĩnh vực khai phá dữ liệu giáo dục mà còn tạo nền tảng cho việc phát triển các hệ thống phân tích học tập thông minh trong môi trường giáo dục đại học.

4.3. Hướng nghiên cứu tương lai

4.3.1. Phát triển theo mô hình thời gian thực

Mặc dù mô hình được đề xuất trong nghiên cứu đã đạt được kết quả dự đoán khả quan trên bộ dữ liệu thực nghiệm, quá trình dự đoán hiện tại vẫn chủ yếu dựa trên dữ liệu tĩnh được thu thập và xử lý theo từng đợt. Trong thực tế, hành vi học tập của sinh viên là một quá trình liên tục thay đổi theo thời gian, chịu ảnh hưởng bởi nhiều yếu tố khác nhau như kết quả học tập, mức độ tham gia học tập, tương tác với giảng viên và các yếu tố cá nhân phát sinh trong quá trình đào tạo. Vì vậy, một hướng

nghiên cứu tiềm năng trong tương lai là phát triển mô hình dự đoán theo thời gian thực nhằm phản ánh chính xác hơn trạng thái học tập hiện tại của sinh viên.

Trong các hệ thống đào tạo hiện đại, dữ liệu học tập được tạo ra liên tục từ nhiều nguồn khác nhau như hệ thống quản lý học tập (Learning Management System - LMS), hệ thống điểm danh điện tử, hệ thống quản lý đào tạo, nền tảng học tập trực tuyến và các công cụ hỗ trợ học tập số. Những dữ liệu này cung cấp thông tin cập nhật về quá trình học tập của sinh viên theo từng ngày, từng tuần hoặc từng học kỳ. Việc tích hợp các nguồn dữ liệu thời gian thực vào mô hình sẽ cho phép hệ thống theo dõi sự thay đổi của người học một cách liên tục thay vì chỉ đưa ra dự đoán tại một thời điểm cố định.

Bên cạnh đó, các kiến trúc học sâu hiện đại như Transformer, Temporal Convolutional Network (TCN) hoặc các mô hình Attention-Based Network có thể được nghiên cứu và áp dụng nhằm khai thác hiệu quả hơn các dữ liệu chuỗi thời gian. Các mô hình này không chỉ giúp cải thiện độ chính xác dự đoán mà còn có khả năng phát hiện sớm những thay đổi bất thường trong hành vi học tập của sinh viên trước khi nguy cơ bỏ học thực sự xảy ra.

Một hướng phát triển quan trọng khác là xây dựng hệ thống cảnh báo sớm (Early Warning System) dựa trên mô hình dự đoán. Trong hệ thống này, kết quả dự đoán không chỉ được sử dụng cho mục đích nghiên cứu mà còn được tích hợp trực tiếp vào quy trình quản lý đào tạo. Khi một sinh viên được xác định có nguy cơ bỏ học vượt quá một ngưỡng nhất định, hệ thống có thể tự động gửi cảnh báo đến cố vấn học tập, giảng viên hoặc bộ phận hỗ trợ sinh viên để triển khai các biện pháp can thiệp phù hợp. Điều này giúp chuyển đổi mô hình từ một công cụ phân tích dữ liệu sang một hệ thống hỗ trợ ra quyết định có khả năng ứng dụng thực tế trong môi trường giáo dục đại học.

Xa hơn nữa, việc kết hợp mô hình dự đoán với các kỹ thuật Explainable AI (XAI) có thể giúp hệ thống không chỉ trả lời câu hỏi "sinh viên nào có nguy cơ bỏ học" mà còn giải thích "vì sao sinh viên đó có nguy cơ bỏ học". Đây là yếu tố quan trọng để nâng cao tính minh bạch và mức độ tin cậy của các hệ thống trí tuệ nhân tạo trong giáo dục.

4.3.2. Mở rộng ứng dụng sang bối cảnh khác

Nghiên cứu hiện tại tập trung vào bài toán dự đoán nguy cơ sinh viên bỏ học trong môi trường giáo dục đại học dựa trên các dữ liệu học thuật, nhân khẩu học và kinh tế - xã hội. Mặc dù kết quả thực nghiệm cho thấy hiệu quả cao của các mô hình học sâu được đề xuất, phạm vi ứng dụng của nghiên cứu vẫn còn nhiều tiềm năng để mở rộng trong tương lai.

Trước hết, mô hình có thể được triển khai và đánh giá trên các bộ dữ liệu được thu thập từ nhiều cơ sở giáo dục khác nhau. Mỗi trường đại học thường có chương trình đào tạo, phương thức quản lý và đặc điểm sinh viên riêng biệt. Việc mở rộng nghiên cứu trên các môi trường đào tạo khác nhau sẽ giúp đánh giá khả năng tổng quát hóa của mô hình, đồng thời tạo cơ sở để xây dựng các hệ thống dự đoán có thể áp dụng trên quy mô rộng hơn. Đây là bước quan trọng để chuyển đổi mô hình từ một giải pháp nghiên cứu thành một công cụ hỗ trợ quản lý giáo dục có khả năng ứng dụng thực tế.

Bên cạnh đó, hướng nghiên cứu có thể được mở rộng sang các cấp học khác ngoài giáo dục đại học. Các bài toán như dự đoán nguy cơ học sinh bỏ học ở bậc trung học phổ thông, dự đoán kết quả học tập ở bậc phổ thông hoặc dự đoán khả năng hoàn thành chương trình đào tạo nghề đều có nhiều điểm tương đồng về mặt dữ liệu và phương pháp tiếp cận. Việc áp dụng các mô hình học sâu trong các bối cảnh này có thể góp phần hỗ trợ các cơ sở giáo dục phát hiện sớm các đối tượng có nguy cơ và triển khai các biện pháp hỗ trợ phù hợp.

Một hướng phát triển đầy tiềm năng là mở rộng từ bài toán dự đoán bỏ học sang các bài toán phân tích học tập khác. Thay vì chỉ dự đoán trạng thái cuối cùng của sinh viên, các mô hình có thể được sử dụng để dự đoán kết quả học tập, đánh giá mức độ thành công của người học, nhận diện sinh viên có nguy cơ học lực yếu hoặc dự báo khả năng hoàn thành từng học phần cụ thể. Những ứng dụng này sẽ giúp hình thành một hệ sinh thái phân tích học tập toàn diện, hỗ trợ quá trình quản lý và nâng cao chất lượng đào tạo.

Trong bối cảnh giáo dục số ngày càng phát triển, việc tích hợp dữ liệu từ các hệ thống học tập trực tuyến cũng là một hướng nghiên cứu quan trọng. Các nguồn dữ liệu như lịch sử truy cập LMS, thời gian học tập trực tuyến, số lần tương tác với học liệu, mức độ tham gia thảo luận hoặc kết quả làm bài tập trực tuyến có thể cung cấp

những thông tin phản ánh trực tiếp hành vi học tập của sinh viên. Khi kết hợp với các dữ liệu học thuật truyền thống, mô hình có thể xây dựng được bức tranh toàn diện hơn về quá trình học tập và nâng cao khả năng dự đoán.

Ngoài ra, nghiên cứu cũng có thể mở rộng sang bài toán phân tích mức độ tham gia học tập và phát hiện tình trạng mất tập trung trong học tập. Đây là một hướng nghiên cứu đang nhận được nhiều sự quan tâm trong lĩnh vực trí tuệ nhân tạo ứng dụng trong giáo dục. Thay vì chỉ xác định nguy cơ bỏ học khi các vấn đề đã xuất hiện rõ ràng, các hệ thống phân tích hành vi học tập có thể phát hiện sớm các dấu hiệu suy giảm mức độ tham gia của sinh viên thông qua dữ liệu tương tác, dữ liệu từ hệ thống học tập hoặc dữ liệu thị giác máy tính. Việc kết hợp giữa dự đoán bỏ học và phân tích mức độ tham gia học tập có thể tạo ra các hệ thống hỗ trợ học tập thông minh với khả năng can thiệp sớm và hiệu quả hơn.

Trong tương lai, cùng với sự phát triển của trí tuệ nhân tạo, dữ liệu lớn và chuyển đổi số giáo dục, các mô hình dự đoán học tập không chỉ đóng vai trò là công cụ phân tích dữ liệu mà còn trở thành một thành phần quan trọng trong các hệ thống quản trị giáo dục thông minh. Những hệ thống này có khả năng hỗ trợ nhà trường theo dõi toàn diện quá trình học tập của sinh viên, cá nhân hóa các hoạt động hỗ trợ và nâng cao chất lượng đào tạo dựa trên dữ liệu. Đây cũng chính là định hướng phát triển dài hạn của các nghiên cứu ứng dụng trí tuệ nhân tạo trong giáo dục hiện đại.

III. KẾT LUẬN

Trong bối cảnh chuyển đổi số giáo dục và sự phát triển mạnh mẽ của các công nghệ trí tuệ nhân tạo, việc ứng dụng các phương pháp học máy và học sâu vào phân tích dữ liệu giáo dục đang trở thành một hướng nghiên cứu có ý nghĩa cả về mặt khoa học lẫn thực tiễn. Bài toán dự đoán nguy cơ sinh viên bỏ học là một trong những vấn đề quan trọng của các cơ sở giáo dục đại học, bởi hiện tượng bỏ học không chỉ ảnh hưởng đến kết quả học tập của người học mà còn tác động đến chất lượng đào tạo, hiệu quả quản lý và sự phát triển bền vững của nhà trường.

Xuất phát từ thực tiễn đó, đồ án đã tập trung nghiên cứu và xây dựng các mô hình học sâu phục vụ bài toán dự đoán nguy cơ sinh viên bỏ học dựa trên dữ liệu giáo dục. Trên cơ sở tổng quan các nghiên cứu liên quan, đồ án đã phân tích các nguyên nhân dẫn đến hiện tượng bỏ học, khảo sát các phương pháp học sâu phổ biến và lựa chọn ba mô hình gồm RNN, LSTM và Hybrid để triển khai thực nghiệm. Đồng thời, nghiên cứu cũng thực hiện các bước tiền xử lý dữ liệu, chuẩn hóa dữ liệu đầu vào và xây dựng quy trình đánh giá mô hình theo các tiêu chí phổ biến trong bài toán phân loại.

Kết quả thực nghiệm trên bộ dữ liệu "Predicting Student Dropout and Academic Success" cho thấy các mô hình học sâu đều đạt hiệu năng dự đoán cao. Trong đó, mô hình Hybrid đạt kết quả tốt nhất với độ chính xác 90%, F1-score đạt 90,43% và ROC-AUC đạt 94,36%, vượt trội hơn so với các mô hình RNN và LSTM được khảo sát. Kết quả này chứng minh rằng việc xử lý riêng biệt các đặc trưng số và đặc trưng phân loại trước khi thực hiện hợp nhất đặc trưng là một hướng tiếp cận hiệu quả đối với các bộ dữ liệu giáo dục có tính dị thể. Bên cạnh đó, nghiên cứu cũng xác định được các yếu tố học thuật và tài chính là những nhóm đặc trưng có ảnh hưởng lớn nhất đến nguy cơ bỏ học của sinh viên, qua đó cung cấp thêm cơ sở thực tiễn cho công tác quản lý và hỗ trợ người học.

Những kết quả đạt được không chỉ khẳng định khả năng ứng dụng của học sâu trong lĩnh vực khai phá dữ liệu giáo dục mà còn cho thấy tiềm năng phát triển các hệ thống cảnh báo sớm dành cho sinh viên có nguy cơ bỏ học. Nếu được triển khai trong môi trường thực tế, các mô hình dự đoán có thể hỗ trợ nhà trường nhận diện sớm các trường hợp rủi ro, từ đó đưa ra các biện pháp can thiệp phù hợp nhằm nâng cao tỷ lệ duy trì học tập và cải thiện chất lượng đào tạo.

Mặc dù vẫn còn một số hạn chế về dữ liệu, khả năng giải thích mô hình và phạm vi thực nghiệm, nghiên cứu đã hoàn thành các mục tiêu đề ra và tạo nền tảng cho các hướng nghiên cứu tiếp theo. Trong tương lai, việc tích hợp thêm dữ liệu hành vi học tập từ các hệ thống LMS, phát triển các mô hình dự đoán theo thời gian thực và ứng dụng các kỹ thuật trí tuệ nhân tạo có khả năng giải thích sẽ góp phần nâng cao hơn nữa hiệu quả và tính ứng dụng của các hệ thống hỗ trợ ra quyết định trong giáo dục đại học.

IV. TÀI LIỆU THAM KHẢO

[1] Tổng cục Thống kê, "Thông cáo báo chí kết quả sơ bộ Điều tra dân số và nhà ở giữa kỳ năm 2024," 2024.

[2] Bộ Giáo dục và Đào tạo, "Thông tư số 01/2024/TT-BGDĐT ban hành Chuẩn cơ sở giáo dục đại học," 2024.

[3] VietnamNet, "Hơn 147.000 thí sinh bỏ nhập học đại học dù trúng tuyển năm 2025," 2025.

[4] OECD, "Who is expected to complete tertiary education?," in Education at a Glance 2025, OECD Publishing, 2025.

[5] S. Kim, E. Yoo, and S. Kim, "Why Do Students Drop Out? University Dropout Prediction and Associated Factor Analysis Using Machine Learning Techniques."

[6] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers, "Predicting students drop out: A case study," in Proc. 2nd Int. Conf. Educ. Data Mining, Cordoba, Spain, 2009, pp. 41–50.

[7] J. P. Bean and B. S. Metzner, "A conceptual model of nontraditional undergraduate student attrition," Rev. Educ. Res., vol. 55, no. 4, pp. 485–540, 1985.

[8] A. F. Cabrera, A. Nora, and M. B. Castañeda, "The role of finances in the persistence process: A structural model," Res. High. Educ., vol. 33, no. 5, pp. 571–593, 1992.

[9] V. Tinto, Leaving College: Rethinking the Causes and Cures of Student Attrition, 2nd ed. Chicago, IL, USA: University of Chicago Press, 1993.

[10] E. R. Kahu and K. Nelson, "Student engagement in the educational interface: Understanding the mechanisms of student success," High. Educ. Res. Dev., vol. 37, no. 1, pp. 58–71, 2018.

[11] S. M. Jayaprakash, E. W. Moody, E. J. M. Lauría, J. R. Regan, and J. D. Baron, "Early alert of academically at-risk students: An open source analytics initiative," J. Learn. Anal., vol. 1, no. 1, pp. 6–47, 2014.

[12] C. Bargmann, L. Thiele, and S. Kauffeld, "Motivation matters: predicting students' career decidedness and intention to dropout after the first year in higher

education," Higher Education, vol. 83, no. 4, pp. 845–861, Apr. 2022, doi: 10.1007/s10734-021-00707-6.

[13] C. Aina, E. Baici, G. Casalone, and F. Pastore, "The determinants of university dropout: A review of the socio-economic literature," Socio-Economic Planning Sciences, vol. 79, p. 101102, Feb. 2022, doi: 10.1016/j.seps.2021.101102.

[14] A. Namoun and A. Alshanqiti, "Predicting student performance using data mining and learning analytics techniques: A systematic literature review," Applied Sciences, vol. 11, no. 1, p. 237, Dec. 2020, doi: 10.3390/app11010237.

[15] B. Albreiki, N. Zaki, and H. Alashwal, "A systematic literature review of students' performance prediction using machine learning techniques," Education Sciences, vol. 11, no. 9, p. 552, Sep. 2021, doi: 10.3390/educsci11090552.

[16] R. Z. Pek, S. T. Ozyer, T. Elhage, T. Ozyer, and R. Alhajj, "The role of machine learning in identifying students at-risk and minimizing failure," IEEE Access, vol. 11, pp. 1224–1243, 2023, doi: 10.1109/ACCESS.2022.3232984.

[17] V. Realinho, J. Machado, L. Baptista, and M. V. Martins, "Predicting student dropout and academic success," Data, vol. 7, no. 11, p. 146, Oct. 2022, doi: 10.3390/data7110146.

[18] H. Huo et al., "Predicting dropout for nontraditional undergraduate students: A machine learning approach," Journal of College Student Retention: Research, Theory & Practice, vol. 24, no. 4, pp. 1054–1077, Feb. 2023, doi: 10.1177/1521025120963821.

[19] N. Mduma, "Data balancing techniques for predicting student dropout using machine learning," Data, vol. 8, no. 3, p. 49, Feb. 2023, doi: 10.3390/data8030049.

[20] K. Yan, "Student performance prediction using XGBoost method from a macro perspective," in Proc. 2021 2nd International Conference on Computing and Data Science (CDS), Jan. 2021, pp. 453–459, doi: 10.1109/CDS52072.2021.00084.

[21] D. Duan, C. Dai, and R. Tu, "Research on the prediction of students' academic performance based on XGBoost," in Proc. 2021 Tenth International Conference of Educational Innovation through Technology (EITT), Dec. 2021, pp. 316–319, doi: 10.1109/EITT53287.2021.00068.

[22] M. V. Martins, D. Tolledo, J. Machado, L. M. T. Baptista, and V. Realinho, "Early prediction of student's performance in higher education: A case study," in Trends and Applications in Information Systems and Technologies, 2021, pp. 166–175, doi: 10.1007/978-3-030-72657-7_16.

[23] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[24] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," arXiv preprint arXiv:1409.1259, 2014.

[25] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Morgan Kaufmann, 2011.

[26] R. J. A. Little and D. B. Rubin, Statistical Analysis with Missing Data, 2nd ed. Wiley, 2002.

[27] W. McKinney, Python for Data Analysis, 3rd ed. O'Reilly Media, 2022.

[28] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 3rd ed. O'Reilly Media, 2022.

[29] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.

[30] C. M. Bishop, Pattern Recognition and Machine Learning. Springer, 2006.

[31] M. Kuhn and K. Johnson, Applied Predictive Modeling. Springer, 2013.

[32] B. Carballo-Mendívil, A. Arellano-González, N. J. Ríos-Vázquez, and M. del P. Lizardi-Duarte, "Predicting Student Dropout from Day One: XGBoost-Based Early Warning System Using Pre-Enrollment Data," Applied Sciences, vol. 15, no. 16, Art. no. 9202, 2025, doi: 10.3390/app15169202.

[33] Tinto, V. Dropout from Higher Education: A Theoretical Synthesis of Recent Research. Rev. Educ. Res. 1975, 45, 125–189.

[34] Tinto, V. Through the Eyes of Students. J. Coll. Stud. Retent. 2017, 19, 254–269.

[35] Bean, J.P.; Metzner, B.S. A Conceptual Model of Nontraditional Undergraduate Student Attrition. Rev. Educ. Res. 1985, 55, 485–540.

[36] Bourdieu, P. Poder, Derecho y Clases Sociales, 2nd ed.; Desclée de Brouwer: Bilbao, Spain, 2001.

[37] Kerby, M.B. Toward a New Predictive Model of Student Retention in Higher Education: An Application of Classical Sociological Theory. J. Coll. Stud. Retent. 2015, 17, 138–161.

[38] Biau, G., and Scornet, E. (2016). A random forest guided tour. Test, 25, 197-227.

[39] Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

[40] Agatonovic-Kustrin, S., and Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. Journal of pharmaceutical and biomedical analysis, 22(5), 717-727.

[41] Arlot, S., and Celisse, A. (2010). A survey of cross-validation procedures for model selection.

[42] Kaya, I. and Cinar, I. (2024). Evaluation of Machine Learning and Deep Learning Approaches for Automatic Detection of Eye Diseases. Intelligent Methods In Engineering Sciences, 3(1), 37-45.

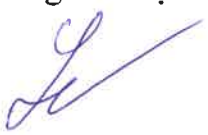
BẢN CAM ĐOAN

Tôi cam đoan đã thực hiện việc kiểm tra mức độ tương đồng nội dung đề án qua phần mềm Turnitin một cách trung thực và đạt kết quả mức độ tương đồng 09 % toàn bộ nội dung đề án. Bản đề án kiểm tra qua phần mềm là bản cứng đề án đã nộp để bảo vệ trước hội đồng. Nếu sai tôi xin chịu các hình thức kỷ luật theo quy định hiện hành của Học viện.

Hà Nội., ngày 22 tháng 6 năm 2026

HỌC VIÊN CAO HỌC/NCS

(Ký và ghi rõ họ tên)



Hoàng Nguyễn Long Thảo

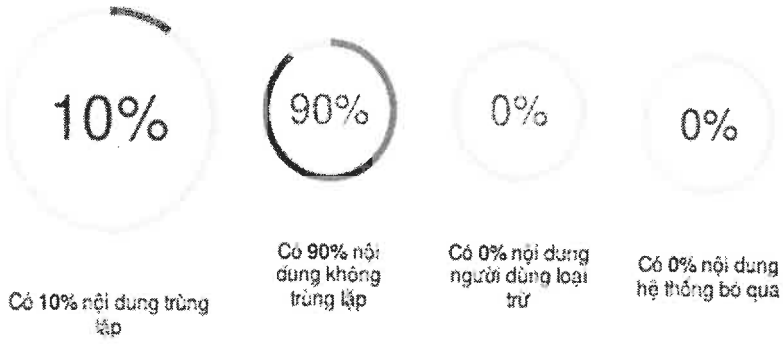
KiểmTraTàiLiệu

BÁO CÁO KIỂM TRA TRÙNG LẶP

Thông tin tài liệu

Tên tài liệu:	22.6 ĐATN_Xây dựng các mô hình học sâu dự đoán tỉ lệ sinh viên bỏ học_Hoàng Nguyễn Long Thảo (1)
Tác giả:	Shireikan Guenki
Điểm trùng lặp:	10
Thời gian tải lên:	20:58 28/07/2026
Thời gian sinh báo cáo:	21:02 28/07/2026
Các trang kiểm tra:	100/100 trang

Kết quả kiểm tra trùng lặp



Nguồn trùng lặp tiêu biểu

tailieu.vn 123docz.net gwdu.ptit.edu.vn

Học viên

Hoàng Nguyễn Long Thảo

Người hướng dẫn

TS. Đỗ Tiến Dũng

BỘ KHOA HỌC VÀ CÔNG NGHỆ
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

BÁO CÁO GIẢI TRÌNH
SỬA CHỮA, HOÀN THIỆN ĐỀ ÁN TỐT NGHIỆP

Họ và tên học viên: Hoàng Nguyễn Long Thảo
Chuyên ngành: KHMT
Khóa: 2023 đợt 2

Tên đề tài: Xây dựng các mô hình học sâu dự đoán tỉ lệ sinh viên bỏ học

Người hướng dẫn khoa học: TS. Đỗ Tiến Dũng

Ngày bảo vệ: 21/07/2026

Các nội dung học viên đã sửa chữa, bổ sung trong đề án tốt nghiệp theo ý kiến đóng góp của Hội đồng chấm đề án tốt nghiệp:

TT	Ý kiến hội đồng	Sửa chữa của học viên
1	Làm rõ bài toán cần giải quyết gồm biểu diễn dữ liệu	Học viên đã chỉnh sửa hình ảnh biểu diễn các mô hình học sâu ở trong chương 2 tại mục 2.2, 2.3, 2.4.
2	Chỉnh sửa lại tài liệu tham khảo	Tiếp thu góp ý của Hội đồng, học viên đã bổ sung, chỉnh sửa thêm 10 tài liệu tham khảo, và còn thêm nội dung tại mục 3.2 trong chương 3.
3	Thực nghiệm lại và mô tả thực nghiệm thể hiện tính đúng đắn	Tiếp thu góp ý của Hội đồng, học viên đã bổ sung thêm nội dung trong phần những hạn chế của mô hình tại mục 3.5.2 trong chương 3

Hà Nội, ngày 29 tháng 7 năm 2026

Ký xác nhận của

CHỦ TỊCH HỘI ĐỒNG CHẤM ĐỀ ÁN	THƯ KÝ HỘI ĐỒNG	NGƯỜI HƯỚNG DẪN KHOA HỌC	HỌC VIÊN
TS.Nguyễn Duy Phương	TS. Đào Thị Thúy Quỳnh	TS. Đỗ Tiến Dũng	Hoàng Nguyễn Long Thảo

BỘ KHOA HỌC VÀ CÔNG NGHỆ
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

BIÊN BẢN
HỌP HỘI ĐỒNG CHẤM ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

Căn cứ quyết định số Quyết định số 1949/QĐ-HV ngày 07 tháng 07 năm 2026 của Giám đốc Học viện Công nghệ Bưu chính Viễn thông về việc thành lập Hội đồng chấm đề án tốt nghiệp thạc sĩ. Hội đồng đã họp vào hồi....8...giờ..50..phút, ngày..21.. tháng 07 năm 2026 tại Học viện Công nghệ Bưu chính Viễn thông để chấm đề án tốt nghiệp thạc sĩ cho:

Học viên: **Hoàng Nguyễn Long Thảo**

Tên đề án tốt nghiệp: **Xây dựng các mô hình học sâu dự đoán tỉ lệ sinh viên bỏ học**

Chuyên ngành: **Khoa học máy tính** Mã số: **8480101**

Các thành viên của Hội đồng chấm đề án tốt nghiệp có mặt: ..5...../ 05

TT	HỌ VÀ TÊN	TRÁCH NHIỆM TRONG HĐ	GHI CHÚ
1	**TS. Nguyễn Duy Phương**	Chủ tịch	
2	**TS. Đào Thị Thúy Quỳnh**	Thư ký	
3	**TS. Trần Minh Tân**	Phản biện 1	
4	**TS. Đào Ngọc Phong**	Phản biện 2	
5	**PGS.TS. Đặng Văn Đức**	Uỷ viên	

Các nội dung thực hiện:

1. Chủ tịch Hội đồng điều khiển buổi họp. Công bố quyết định của Giám đốc Học viện Công nghệ Bưu chính Viễn thông về việc thành lập Hội đồng chấm đề án tốt nghiệp thạc sĩ.
2. Người hướng dẫn khoa học hoặc thư ký đọc lý lịch khoa học và các điều kiện bảo vệ đề án tốt nghiệp của học viên. (có bản lý lịch khoa học và kết quả các môn học cao học của học viên kèm theo).
3. Học viên trình bày tóm tắt đề án tốt nghiệp.
4. Phản biện 1 đọc nhận xét (có văn bản kèm theo)
5. Phản biện 2 đọc nhận xét (có văn bản kèm theo)
6. Các câu hỏi của thành viên Hội đồng:

1) Làm rõ bài toán là phân loại nhị phân hay đa phân loại?
Kiểm tra lại số lượng nhãn của tập DL?
.....
2) Việc lựa chọn mô hình BiLSTM và LSTM là dựa trên cơ sở nào? Em hãy giải thích?
.....

7. Trả lời của học viên:

1) Bài toán dự đoán sự nghỉ học của em đang thực hiện là bài toán phân loại nhị phân, em sẽ rà soát và chỉnh sửa lại cho phù hợp.
2) Học viên sẽ thực hiện rà soát và điều chỉnh lại mô hình và thử nghiệm.

8. Thư ký đọc nhận xét về quá trình thực hiện đề án tốt nghiệp của học viên *(có văn bản kèm theo).*

9. Hội đồng họp riêng:

- Bầu Ban kiểm phiếu:
  1. Trưởng Ban kiểm phiếu: TS. Nguyễn Duy Phương
  2. Uỷ viên Ban kiểm phiếu: TS. Đào Ngọc Phong
  3. Uỷ viên Ban kiểm phiếu: TS. Đào Thị Thúy Quỳnh
- Hội đồng chấm đề án tốt nghiệp bằng bỏ phiếu kín.
- Ban kiểm phiếu làm việc:
- Trưởng Ban kiểm phiếu báo cáo kết quả kiểm phiếu *(có Biên bản họp Ban kiểm phiếu kèm theo)*
- **Điểm trung bình của đề án tốt nghiệp:** 8,0

Kết luận:

1. Các nội dung cần chỉnh sửa, hoàn thiện sau bảo vệ đề án tốt nghiệp:

- Làm rõ bài toán cần giải quyết gồm biểu diễn dữ liệu.
- Chỉnh sửa lại TITK
- Thử nghiệm lại và mô tả thử nghiệm thể hiện tính đúng đắn.

2. Đề nghị Học viện công nhận (hoặc không) và cấp bằng (hoặc không) thạc sĩ cho học viên: Đề nghị công nhận

3. Đề án tốt nghiệp có thể phát triển thành đề tài nghiên cứu cho NCS.....Không.....

Buổi làm việc kết thúc vào..... cùng ngày.

Chủ tịch	**Thư ký**
Phương	
Nguyễn Duy Phương	Đào Thị Thúy Quỳnh

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

BẢN NHẬN XÉT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

(Dùng cho người phản biện)

Tên đề tài đề án tốt nghiệp: Xây dựng các mô hình học sâu dự đoán tỉ lệ sinh viên bỏ học.

Chuyên ngành: Khoa học máy tính.

Mã chuyên ngành: 8.48.01.01.

Họ và tên học viên: Hoàng Nguyễn Long Thảo.

Họ và tên người nhận xét: Trần Minh Tân.

Học hàm, học vị: Tiến sỹ.

Chuyên ngành: Công nghệ thông tin.

Cơ quan công tác: Học viện Chiến lược Khoa học và Công nghệ.

Số điện thoại: 0913275577; E-mail: tantm@mst.gov.vn

NỘI DUNG NHẬN XÉT

I/ Cơ sở khoa học và thực tiễn, tính cấp thiết của đề tài:

Đề tài lựa chọn nghiên cứu bài toán dự đoán nguy cơ sinh viên bỏ học bằng các mô hình học sâu. Cụ thể là tập trung vào xây dựng và so sánh các mô hình học sâu (RNN, LSTM, Hybrid) để dự đoán nguy cơ sinh viên bỏ học dựa trên dữ liệu học tập và các đặc trưng liên quan. Đây là hướng nghiên cứu có cơ sở khoa học, có ý nghĩa thực tiễn đối với công tác quản lý đào tạo trong các trường đại học.

Đề án đã phân tích tương đối đầy đủ thực trạng sinh viên bỏ học trong nước và quốc tế, chỉ ra các nguyên nhân học thuật, kinh tế, tâm lý và môi trường học tập; từ đó làm rõ nhu cầu xây dựng các công cụ dự báo nhằm hỗ trợ nhà trường can thiệp sớm.

II/ Nội dung của đề án tốt nghiệp, các kết quả đã đạt được:

Học viên đã thực hiện các nội dung chính:

1. Thu thập và tiền xử lý dữ liệu;

2. Xây dựng các mô hình RNN, LSTM và Hybrid;

3. Tiến hành đánh giá các mô hình theo các chỉ tiêu Accuracy, Recall, F1-score và thời gian xử lý;

4. Phân tích các đặc trưng ảnh hưởng đến khả năng bỏ học của sinh viên.

Đây là một quy trình nghiên cứu tương đối đầy đủ đối với một đề án theo định hướng ứng dụng.

Bên cạnh các kết quả đạt được, đề án vẫn còn một số điểm cần hoàn thiện:

1. Mô hình Hybrid chủ yếu là sự kết hợp các kiến trúc đã có; Đề án chưa nêu được điểm mới so với các nghiên cứu quốc tế gần đây về dự đoán sinh viên bỏ học để làm rõ, nổi bật được các đóng góp khoa học trong đề tài.

2. Lựa chọn mô hình cần được giải thích sâu hơn. Đề án sử dụng RNN và LSTM cho bài toán dự đoán trên dữ liệu sinh viên. Tuy nhiên, nếu dữ liệu chủ yếu là dữ liệu bảng, cần làm rõ cơ sở khoa học của việc lựa chọn các mô hình xử lý chuỗi thay vì các mô hình mạnh cho dữ liệu bảng như XGBoost, LightGBM hoặc CatBoost.

3. Cần bổ sung thêm thiết kế thực nghiệm; Đề án cần mô tả rõ hơn quy trình chia tập dữ liệu, xử lý mất cân bằng dữ liệu, lựa chọn siêu tham số và đánh giá độ ổn định của mô hình qua nhiều lần chạy.

4. Khả năng tổng quát hóa của mô hình chưa được kiểm chứng trên nhiều bộ dữ liệu hoặc nhiều cơ sở đào tạo khác nhau.

5. Khả năng triển khai thực tế trong hệ thống quản lý đào tạo mới được đề cập ở mức định hướng, chưa xây dựng được quy trình ứng dụng cụ thể.

Nhìn chung, học viên đã hoàn thành các mục tiêu nghiên cứu đã đề ra và triển khai được một hệ thống thử nghiệm tương đối đầy đủ.

III/ Những vấn đề cần giải thích thêm:

- Đóng góp mới của đề án là gì nếu so với các nghiên cứu đã có trước đây cũng đang sử dụng RNN, LSTM trong dự đoán sinh viên bỏ học?

- Mô hình Hybrid được đề xuất cải tiến ở điểm nào so với việc sử dụng riêng RNN hoặc LSTM?

- Trong trường hợp triển khai tại một trường đại học khác với đặc điểm sinh viên khác biệt, mô hình có cần huấn luyện lại hay có thể sử dụng trực tiếp?

IV/ Kết luận:

Đề án lựa chọn một bài toán có ý nghĩa thực tiễn cao đối với quản lý giáo dục đại học. Học viên đã nghiên cứu cơ sở lý thuyết, xây dựng các mô hình học sâu và tiến hành thực nghiệm đánh giá trên dữ liệu thực tế. Các kết quả nghiên cứu bước đầu cho thấy khả năng ứng dụng của trí tuệ nhân tạo trong dự báo nguy cơ sinh viên bỏ học.

Mặc dù còn một số hạn chế về tính mới của mô hình, thiết kế thực nghiệm và phân tích khả năng tổng quát hóa, các hạn chế này không làm ảnh hưởng đến giá trị tổng thể của đề án.

Đánh giá chung, Học viên đã hoàn thành các nội dung nghiên cứu theo đề cương đặt ra, kết quả báo cáo đạt yêu cầu.

Đồng ý cho phép học viên bảo vệ đề án tốt nghiệp./.

Ngày 20 tháng 7 năm 2026
NGƯỜI NHẬN XÉT



Trần Minh Tân

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM

Độc lập – Tự do – Hạnh phúc

BẢN NHẬN XÉT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

(Dùng cho người phản biện)

Tên đề tài đề án: Xây dựng các mô hình học sâu dự đoán tỉ lệ sinh viên bỏ học.

Chuyên ngành: Khoa học máy tính.

Mã số chuyên ngành: 8.48.01.01.

Học viên thực hiện: Hoàng Nguyễn Long Thảo.

Họ và tên người nhận xét: Đào Ngọc Phong

Học hàm, học vị: Tiến sĩ

Chuyên ngành:Hệ thống thông tin

Cơ quan công tác: Khoa Công nghệ thông tin 1

Số điện thoại: 0912239696 E-mail: phongdn@ptit.edu.vn

NỘI DUNG NHẬN XÉT

I/ Cơ sở khoa học và thực tiễn, tính cấp thiết của đề tài:

Trong bối cảnh chuyển đổi số giáo dục và sự phát triển mạnh mẽ của các công nghệ trí tuệ nhân tạo, việc ứng dụng các phương pháp học máy và học sâu vào phân tích dữ liệu giáo dục đang trở thành một hướng nghiên cứu có ý nghĩa cả về mặt khoa học lẫn thực tiễn. Bài toán dự đoán nguy cơ sinh viên bỏ học là một trong những vấn đề quan trọng của các cơ sở giáo dục đại học, bởi hiện tượng bỏ học không chỉ ảnh hưởng đến kết quả học tập của người học mà còn tác động đến chất lượng đào tạo, hiệu quả quản lý và sự phát triển bền vững của nhà trường.

Xuất phát từ thực tiễn đó, đồ án đã tập trung nghiên cứu và xây dựng các mô hình học sâu phục vụ bài toán dự đoán nguy cơ sinh viên bỏ học dựa trên dữ liệu giáo dục. Trên cơ sở tổng quan các nghiên cứu liên quan, đồ án đã phân tích các nguyên nhân dẫn đến hiện tượng bỏ học, khảo sát các phương pháp học sâu phổ biến và lựa chọn ba mô hình gồm RNN, LSTM và Hybrid để triển khai thực nghiệm. Đồng thời, nghiên cứu cũng thực hiện các bước tiền xử lý dữ liệu, chuẩn hóa dữ liệu đầu vào và xây dựng quy trình đánh giá mô hình theo các tiêu chí phổ biến trong bài toán phân loại Việc ứng dụng các kỹ thuật phân tích dữ liệu tiên tiến để xây dựng hệ thống cảnh báo sớm (Early Warning System) là vô cùng cấp thiết.

Đề tài chọn hướng tiếp cận sử dụng học sâu (Deep Learning) để dự đoán nguy cơ bỏ học nhằm hỗ trợ can thiệp kịp thời là phù hợp và có ý nghĩa ứng dụng.

II/ Nội dung của đề án tốt nghiệp, các kết quả đã đạt được:

Đề án nghiên cứu đã hệ thống hóa cơ sở lý thuyết liên quan đến bài toán dự đoán sinh viên bỏ học, bao gồm các nguyên nhân dẫn đến hiện tượng bỏ học, các nhóm yếu tố ảnh hưởng và các phương pháp học máy, học sâu thường được sử dụng trong lĩnh vực giáo dục. Trên cơ sở tổng hợp các nghiên cứu trước đây, luận văn đã làm rõ mối quan hệ giữa các yếu tố học thuật, tài chính, tâm lý và xã hội đối với khả năng duy trì học tập của sinh viên. Đây là cơ sở quan trọng để lựa chọn đặc trưng đầu vào và xây dựng các mô hình dự đoán phù hợp.

Đề án nghiên cứu đã xây dựng và đánh giá thành công ba mô hình học sâu gồm RNN, LSTM và Hybrid trên cùng một bộ dữ liệu giáo dục. Việc triển khai đồng thời nhiều kiến trúc mạng khác nhau cho phép đánh giá khách quan hiệu quả của từng phương pháp, đồng thời cung cấp cơ sở thực nghiệm cho việc lựa chọn mô hình phù hợp đối với bài toán dự đoán nguy cơ bỏ học.

Đóng góp quan trọng nhất của nghiên cứu là việc đề xuất và xây dựng mô hình Hybrid theo kiến trúc đa nhánh nhằm xử lý riêng biệt các nhóm dữ liệu có bản chất khác nhau. Trong mô hình này, các thuộc tính số được xử lý thông qua mạng nơ-ron truyền thẳng nhiều tầng, trong khi các thuộc tính phân loại được biểu diễn bằng kỹ thuật Embedding trước khi thực hiện hợp nhất đặc trưng.

Kết quả thực nghiệm cho thấy mô hình Hybrid đạt độ chính xác 90% và ROCAUC đạt 94,36%, cao hơn so với các mô hình RNN và LSTM được khảo sát trong nghiên cứu.

Xác định được các đặc trưng có ảnh hưởng lớn nhất đến nguy cơ bỏ học của sinh viên. Kết quả cho thấy các yếu tố liên quan đến kết quả học tập trong các học kỳ đầu tiên, số học phần hoàn thành, tình trạng đóng học phí, học bổng và điều kiện tài chính là những đặc trưng quan trọng.

III/ Những vấn đề cần giải thích thêm:

Một số câu hỏi:

- Trang 34, học viên trình bày giải pháp xử lý dữ liệu thiếu bằng phương pháp Forward Fill (FFill) - tức là lấy giá trị hợp lệ của dòng trước gán cho ô trống của dòng sau. Kỹ thuật này chỉ có nghĩa khi các dòng dữ liệu có tính kế thừa liên tục (như dữ liệu cảm biến hoặc chuỗi thời gian liên tục). Tuy nhiên, bộ dữ liệu của đề án khi thực nghiệm tại Chương 3 không đề cập đến việc xử lý dữ liệu thiếu.
- Tại Chương 3, học viên đã đưa ra bảng xếp hạng 9 đặc trưng có mức độ ảnh hưởng cao nhất đến kết quả dự đoán. Trong phần hạn chế ở trang 72, học viên có trình bày về việc các mô hình học sâu này (đặc biệt là Hybrid)

hoạt động như một "hộp đen" (black box) và rất khó giải thích cơ chế bên trong. Vậy học viên làm rõ về dữ liệu xếp hạng các đặc trưng về mức độ ảnh hưởng.

- Học viện đánh giá về khả năng áp dụng vào hệ thống quản lý đào tạo của 1 trường đại học (ví dụ như Học viện BCVT).

V/ Kết luận:

Đồng ý cho phép học viên bảo vệ đề án tốt nghiệp.

Ngày 20 tháng 07 năm 2026

NGƯỜI NHẬN XÉT

Đào Ngọc Phong