HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

LÊ MINH TUẤN

NGHIÊN CỨU XÂY DỰNG HỆ THỐNG HỖ TRỢ TUYỂN SINH TẠI TRƯỜNG CAO ĐẲNG HÀ NỘI

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

HÀ NỘI – 2026

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

LÊ MINH TUẤN

NGHIÊN CỨU XÂY DỰNG HỆ THỐNG HỖ TRỢ TUYỂN SINH TẠI TRƯỜNG CAO ĐẲNG HÀ NỘI

Chuyên ngành: Hệ thống thông tin
Mã số: 8.48.01.04

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

NGƯỜI HƯỚNG DẪN KHOA HỌC: TS. VŨ VĂN THỎA

HÀ NỘI – 2026

LỜI CAM ĐOAN

Tôi cam đoan rằng công trình nghiên cứu này là thành quả của công sức cá nhân của tôi và không sao chép từ bất kỳ nguồn nào khác. Tất cả thông tin được trình bày trong đề án này đều là sản phẩm của công việc cá nhân hoặc được tổng hợp từ nhiều nguồn tài liệu khác nhau. Mọi tài liệu tham khảo đều được trích dẫn một cách hợp pháp và có nguồn gốc rõ ràng. Các dữ liệu và kết quả được trình bày trong đề án đều là trung thực và chưa từng được công bố trong bất kỳ công trình nghiên cứu nào khác.

Tác giả đề án

Tuấn

Lê Minh Tuấn

LỜI CẢM ƠN

Lời đầu tiên, học viên xin gửi lời cảm ơn sâu sắc tới các thầy, cô giáo giảng viên khoa Công nghệ thông tin 1, khoa Đào tạo Sau đại học nói riêng và các thầy, cô giáo giảng viên Học viện Công nghệ Bưu chính Viễn thông nói chung. Trong suốt quá trình học tập tại Học viện, các thầy cô đã chỉ bảo, giảng dạy cho em biết bao kiến thức, kinh nghiệm quý báu để em có hành trang vững bước trong tương lai.

Em cũng xin được gửi lời cảm ơn tới thầy hướng dẫn TS. Vũ Văn Thỏa, cảm ơn thầy đã luôn hướng dẫn chỉ bảo tận tình em trong suốt quá trình học tập, nghiên cứu và thực hiện đề án này. Những lời khuyên, sự chỉ bảo của các thầy đã giúp em hoàn thành đề án tốt nghiệp này cũng như có thêm rất nhiều kiến thức, kinh nghiệm trong việc học tập và nghiên cứu.

Dù đã nỗ lực hoàn thành đề án, em hiểu rằng có thể không tránh khỏi những sai sót. Kính mong được thầy cô và các bạn thông cảm và đóng góp ý kiến.

Em xin trân trọng cảm ơn.

MỤC LỤC

LỜI CAM ĐOAN i
LỜI CẢM ƠN ii
MỤC LỤC iii
DANH MỤC TỪ VIẾT TẮT v
DANH SÁCH CÁC BẢNG vi
DANH MỤC CÁC HÌNH vii
MỞ ĐẦU 1
CHƯƠNG 1: TỔNG QUAN VỀ HỆ THỐNG HỖ TRỢ QUẢN LÝ TUYỂN SINH VÀ CÁC VẤN ĐỀ LIÊN QUAN 3
1.1. Tổng quan về tuyển sinh tại các trường cao đẳng 3
1.1.1. Quy định của các cơ quan quản lý nhà nước về tuyển sinh 3
1.1.2. Hiện trạng tuyển sinh tại các trường cao đẳng 4
1.2. Quy trình và nghiệp vụ tuyển sinh 5
1.2.1. Quy chế tuyển sinh 5
1.2.2. Quy trình tuyển sinh 6
1.3. Bài toán hỗ trợ tư vấn tuyển sinh 9
1.3.1. Mô tả bài toán 9
1.3.2. Ý nghĩa bài toán 11
1.4. Khảo sát các mô hình giải quyết bài toán 12
1.4.1. Các mô hình dựa trên học máy truyền thống 12
1.4.2. Các mô hình dựa trên học sâu 15
1.5. Kết luận chương 1 17
CHƯƠNG 2: PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG 19
2.1. Thực trạng công tác tuyển sinh tại Trường Cao đẳng Hà Nội 19
2.2. Nghiên cứu mô hình hỗ trợ thí sinh chọn ngành phù hợp 20
2.2.1. Mô hình hồi quy Logistic 20
2.2.2. Mô hình rừng ngẫu nhiên 24
2.2.3. Mô hình LSTM 27

2.2.4. Tiêu chí lựa chọn mô hình 31
2.3. Đề xuất giải pháp triển khai hệ thống hỗ trợ tuyển sinh cho Trường Cao đẳng Hà Nội 31
2.4. Phân tích yêu cầu hệ thống 32
2.4.1. Yêu cầu chức năng 32
2.4.2. Yêu cầu phi chức năng 35
2.5. Phân tích thiết kế cơ sở dữ liệu 37
2.5.1. Nguồn dữ liệu 38
2.5.2. Tiền xử lý dữ liệu 40
2.5.3. Tổ chức và chuẩn hóa dữ liệu phục vụ hệ thống hỗ trợ tuyển sinh 42
2.5.4. Liên kết dữ liệu với mô hình dự đoán tuyển sinh 44
2.6. Kết luận chương 2 45
CHƯƠNG 3: THỬ NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ 47
3.1. Kịch bản thử nghiệm 47
3.1.1. Nội dung thử nghiệm 47
3.1.2. Công cụ và môi trường thử nghiệm 51
3.2. Triển khai thử nghiệm và kết quả 51
3.2.1. Triển khai thử nghiệm và đánh giá riêng từng mô hình 51
3.2.2. Tổng hợp kết quả và so sánh giữa các mô hình 62
3.3. Nhận xét và đánh giá 65
3.4. Kết luận chương 3 67
KẾT LUẬN 68
DANH MỤC TÀI LIỆU THAM KHẢO 70

DANH MỤC TỪ VIẾT TẮT

STT	Từ viết tắt	Tiếng Anh	Tiếng Việt/Giải thích
1	AI	Artificial Intelligence	Trí tuệ nhân tạo
2	ML	Machine Learning	Học máy
3	DL	Deep Learning	Học sâu
4	CNN	Convolutional Neural Network	Mạng nơ-ron tích chập
5	KNN	K-Nearest Neighbors	Thuật toán K láng giềng gần nhất
6	LSTM	Long Short-Term Memory	Mạng bộ nhớ ngắn hạn – dài hạn
7	RIASEC	Realistic – Investigative – Artistic – Social – Enterprising – Conventional	Mô hình định hướng nghề nghiệp Holland
8	RNN	Recurrent Neural Network	Mạng nơ-ron hồi tiếp

DANH SÁCH CÁC BẢNG

Bảng 2.1. Minh họa quá trình tổ chức và tinh gọn dữ liệu tuyển sinh 44
Bảng 3.1. Tổng quan các mô hình thử nghiệm và phần mềm sử dụng 49
Bảng 3.2. Bảng mô tả đặc trưng đầu vào 49
Bảng 3.3. Cách tổ chức dữ liệu đầu vào cho các mô hình thử nghiệm 50
Bảng 3.4. Phần mềm và môi trường thực nghiệm 51
Bảng 3.5. Thông số cấu hình mô hình hồi quy tuyến tính 52
Bảng 3.6. Hệ số hồi quy theo ngành học - Phần 1 (Intercept, Toán, Văn, Anh, Lý, Hoá, Sinh, Sử) 53
Bảng 3.7. Hệ số hồi quy theo ngành học - Phần 2 (Địa, GDCD, R, I, A, S, E, C) ... 54
Bảng 3.8. Mức độ ảnh hưởng trung bình của từng đặc trưng (mô hình hồi quy tuyến tính) 55
Bảng 3.9. Kết quả thử nghiệm mô hình hồi quy tuyến tính 55
Bảng 3.10. Cấu hình mô hình hồi quy phi tuyến (Random Forest) 57
Bảng 3.11. Mức độ quan trọng đặc trưng - Feature Importance (mô hình hồi quy phi tuyến) 57
Bảng 3.12. Kết quả thực nghiệm mô hình hồi quy phi tuyến 58
Bảng 3.13. Kiến trúc mô hình LSTM_Net 59
Bảng 3.14. Bảng tham số mô hình LSTM_Net 60
Bảng 3.15. Cấu hình huấn luyện LSTM_Net 60
Bảng 3.16. Kết quả thử nghiệm mô hình LSTM_Net 62
Bảng 3.17. Kết quả Accuracy của ba mô hình trên tập huấn luyện 62
và tập kiểm định 62

DANH MỤC CÁC HÌNH

Hình 1.1: Các bước của quy trình tuyển sinh....8
Hình 2.1. Giải thuật rừng ngẫu nhiên....26
Hình 2.2. Cấu trúc của LSTM [2]....28
Hình 2.3. Sơ đồ luồng dữ liệu chức năng nhập thông tin thí sinh....33
Hình 2.4. Sơ đồ luồng dữ liệu chức năng dự đoán ngành phù hợp....34
Hình 2.5. Sơ đồ luồng dữ liệu chức năng quản lý dữ liệu tuyển sinh....35
Hình 3.1. Quá trình huấn luyện mô hình LSTM qua 30 epoch - Training Loss, Top-1 và Top-3 Accuracy và lịch trình tốc độ học dạng Cosine....61
Hình 3.2. So sánh Top-1 và Top-3 Accuracy giữa ba mô hình....63
Hình 3.3. So sánh Accuracy huấn luyện và Accuracy kiểm định giữa ba mô hình..63
Hình 3.4. Precision, Recall và F1-Score theo từng ngành học của ba mô hình trên tập kiểm định....64
Hình 3.5. So sánh F1-Score theo từng ngành học giữa ba mô hình....65
Hình 3.6. Precision so với Recall theo từng ngành học của ba mô hình trên tập kiểm định....66

MỞ ĐẦU

Trong bối cảnh chuyển đổi số đang diễn ra mạnh mẽ trên nhiều lĩnh vực của đời sống xã hội, trí tuệ nhân tạo (Artificial Intelligence - AI) và học máy (Machine Learning) ngày càng được ứng dụng rộng rãi nhằm nâng cao hiệu quả ra quyết định và hỗ trợ con người trong nhiều hoạt động khác nhau. Đối với lĩnh vực giáo dục và tuyển sinh, việc khai thác dữ liệu học tập của người học kết hợp với các mô hình phân tích thông minh không chỉ giúp nâng cao chất lượng tư vấn hướng nghiệp mà còn góp phần lựa chọn ngành học phù hợp với năng lực, sở trường và định hướng nghề nghiệp của từng thí sinh.

Thực tế cho thấy, việc lựa chọn ngành học của học sinh hiện nay vẫn chịu ảnh hưởng bởi nhiều yếu tố như xu hướng xã hội, ý kiến gia đình hoặc cảm tính cá nhân, trong khi chưa khai thác hiệu quả dữ liệu về kết quả học tập và đặc điểm nghề nghiệp của người học. Điều này dẫn đến tình trạng lựa chọn ngành học chưa thực sự phù hợp, ảnh hưởng đến quá trình học tập, khả năng phát triển nghề nghiệp cũng như chất lượng nguồn nhân lực sau đào tạo. Vì vậy, việc nghiên cứu và xây dựng một hệ thống hỗ trợ tư vấn tuyển sinh dựa trên trí tuệ nhân tạo có ý nghĩa cả về mặt khoa học và thực tiễn.

Xuất phát từ yêu cầu đó, đề án **"Nghiên cứu xây dựng hệ thống hỗ trợ tuyển sinh tại Trường Cao đẳng Hà Nội"** được thực hiện nhằm nghiên cứu, xây dựng và đánh giá một mô hình dự đoán ngành học phù hợp cho thí sinh. Đề án sử dụng dữ liệu điểm học tập của học sinh trong ba năm trung học phổ thông kết hợp với sáu nhóm đặc trưng định hướng nghề nghiệp theo mô hình Holland (RIASEC). Trên cơ sở đó, mô hình học sâu LSTM được xây dựng và so sánh với các mô hình học máy truyền thống như Logistic Regression và Random Forest nhằm đánh giá hiệu quả dự đoán cũng như khả năng ứng dụng trong thực tế.

Đề án được bố cục thành **03 chương**, bao gồm:

Chương 1: Tổng quan về hệ thống hỗ trợ quản lý tuyển sinh và các vấn đề liên quan

Chương này trình bày cơ sở hình thành đề án, tổng quan về công tác tư vấn tuyển sinh, các nghiên cứu trong và ngoài nước liên quan đến ứng dụng trí tuệ nhân tạo trong giáo dục, đồng thời xác định mục tiêu, phạm vi nghiên cứu và hướng tiếp cận của đề án.

Chương 2: Phân tích và thiết kế hệ thống

Chương này giới thiệu các cơ sở lý thuyết về học máy, học sâu, mạng nơ-ron LSTM, mô hình định hướng nghề nghiệp RIASEC và quy trình tiền xử lý dữ liệu. Trên cơ sở đó, xây dựng mô hình dự đoán ngành học, thiết kế kiến trúc hệ thống, mô tả tập dữ liệu, quy trình huấn luyện và phương pháp đánh giá mô hình.

Chương 3: Thử nghiệm và đánh giá kết quả

Chương này trình bày quá trình triển khai thực nghiệm trên tập dữ liệu, cấu hình môi trường thử nghiệm, kết quả đánh giá của các mô hình Logistic Regression, Random Forest và LSTM thông qua các chỉ số Top-1 Accuracy, Top-3 Accuracy, Precision, Recall và Macro F1-score. Từ các kết quả thu được, đề án phân tích ưu, nhược điểm của từng mô hình, lựa chọn mô hình phù hợp nhất để tích hợp vào hệ thống hỗ trợ tư vấn tuyển sinh và đề xuất hướng phát triển trong thời gian tới.

CHƯƠNG 1: TỔNG QUAN VỀ HỆ THỐNG HỖ TRỢ QUẢN LÝ TUYỂN SINH VÀ CÁC VẤN ĐỀ LIÊN QUAN

1.1. Tổng quan về tuyển sinh tại các trường cao đẳng

Công tác tuyển sinh đóng vai trò là hoạt động then chốt, mang tính quyết định đến sự tồn tại và phát triển bền vững của các cơ sở giáo dục nghề nghiệp trong hệ thống giáo dục quốc dân. Trong bối cảnh hiện nay, dưới tác động mạnh mẽ của cuộc cách mạng công nghiệp 4.0 và sự dịch chuyển không ngừng của thị trường lao động, hoạt động tuyển sinh không chỉ dừng lại ở các nghiệp vụ hành chính đơn thuần mà còn đòi hỏi sự chuyển mình mạnh mẽ về mặt công nghệ để đáp ứng nhu cầu ngày càng cao của người học và yêu cầu quản lý từ phía Nhà nước.

1.1.1. Quy định của các cơ quan quản lý nhà nước về tuyển sinh

Hoạt động tuyển sinh tại các cơ sở giáo dục nghề nghiệp hiện nay được vận hành dựa trên một hành lang pháp lý chặt chẽ và thống nhất, nhằm bảo đảm quyền lợi của người học cũng như trách nhiệm giải trình của nhà trường trước xã hội. Các quy định này không chỉ thiết lập tiêu chuẩn về quy trình nghiệp vụ mà còn định hướng chiến lược cho việc hiện đại hóa công tác quản lý giáo dục trong kỷ nguyên số. Căn cứ vào các văn bản quy phạm pháp luật hiện hành, trọng tâm là Thông tư số 05/2021/TT-BLĐTBXH của Bộ Lao động - Thương binh và Xã hội [17], quy chế tuyển sinh yêu cầu các cơ sở đào tạo phải xây dựng và công bố đề án tuyển sinh một cách công khai, chi tiết. Việc này bao gồm các cam kết về chỉ tiêu đào tạo, điều kiện cơ sở vật chất, đội ngũ giảng viên và đặc biệt là tỷ lệ sinh viên có việc làm sau tốt nghiệp. Tính minh bạch này đóng vai trò là cơ chế kiểm soát chất lượng, giúp thí sinh và phụ huynh có đầy đủ cơ sở thông tin để đưa ra lựa chọn giáo dục phù hợp.

Theo các quy định hiện hành, các trường cao đẳng được trao quyền tự chủ cao trong việc lựa chọn và kết hợp linh hoạt các phương thức xét tuyển. Nhà trường có thể căn cứ vào kết quả học tập bậc trung học phổ thông, sử dụng điểm thi tốt nghiệp trung học phổ thông quốc gia, hoặc triển khai các kỳ thi đánh giá năng lực và phỏng vấn chuyên biệt tùy theo đặc thù của từng ngành nghề đào tạo như Công nghệ ô tô, Dược hay Điều dưỡng. Quyền tự chủ này cho phép các đơn vị chủ động trong việc

sàng lọc và tuyển chọn những cá nhân có năng lực, tố chất tương thích nhất với yêu cầu nghiệp vụ chuyên môn của từng ngành nghề.

Đi kèm với quyền tự chủ là trách nhiệm hiện đại hóa hạ tầng công nghệ theo chủ trương chuyển đổi số toàn diện của ngành giáo dục. Bộ Giáo dục và Đào tạo cùng Tổng cục Giáo dục nghề nghiệp đặc biệt nhấn mạnh yêu cầu bắt buộc đối với các trường trong việc thực hiện số hóa toàn diện dữ liệu tuyển sinh. Việc chuyển đổi từ phương thức quản lý hồ sơ giấy truyền thống sang hệ thống dữ liệu số không chỉ giúp tối ưu hóa công tác quản trị nội bộ mà còn đảm bảo tính đồng bộ, liên thông với hệ thống Cơ sở dữ liệu Giáo dục nghề nghiệp quốc gia. Đây là nền tảng kỹ thuật quan trọng phục vụ công tác thanh tra, giám sát trực tuyến và là tiền đề để triển khai các mô hình phân tích dữ liệu thông minh trong việc tư vấn hướng nghiệp cho người học. Sự tổng hòa của các quy định trên đã tạo ra một khung khổ pháp lý vừa có tính kỷ cương, vừa có tính mở, tạo điều kiện thuận lợi để các trường triển khai các hệ thống hỗ trợ tuyển sinh tiên tiến nhằm nâng cao hiệu quả quản lý và chất lượng phục vụ thí sinh trong bối cảnh cạnh tranh giáo dục ngày càng gia tăng.

1.1.2. Hiện trạng tuyển sinh tại các trường cao đẳng

Thực tiễn công tác tuyển sinh trong giai đoạn 2024–2026 đang bộc lộ nhiều biến động phức tạp và thách thức mang tính hệ thống, ảnh hưởng trực tiếp đến chiến lược phát triển của các cơ sở giáo dục nghề nghiệp. Sự mất cân đối trong xu hướng đăng ký ngành nghề ngày càng trở nên rõ rệt khi thí sinh tập trung quá lớn vào các lĩnh vực mang tính thời đại như công nghệ thông tin, kỹ thuật ô tô và chăm sóc sức khỏe, điều này vô hình trung tạo ra sự suy giảm nguồn tuyển nghiêm trọng tại các ngành nghề kỹ thuật truyền thống khác. Bên cạnh đó, tâm lý ưu tiên giáo dục đại học vẫn còn phổ biến và ăn sâu vào tiềm thức của cộng đồng phụ huynh cũng như học sinh, tạo ra áp lực cạnh tranh khốc liệt và khiến khối giáo dục nghề nghiệp luôn ở thế yếu trong việc thu hút nhân lực chất lượng cao.

Một trong những vấn đề nan giải nhất hiện nay chính là sự hạn chế trong công tác tư vấn hướng nghiệp, khi phần lớn thí sinh vẫn thực hiện lựa chọn ngành học dựa trên cảm tính cá nhân hoặc theo xu hướng đám đông mà thiếu đi sự phân tích khoa

học về sự tương quan giữa năng lực học tập thực tế và đặc điểm tính cách cá nhân. Việc thiếu hụt các công cụ định hướng chuyên sâu dẫn đến tình trạng nhiều sinh viên sau khi nhập học cảm thấy không phù hợp với chương trình đào tạo, gây lãng phí thời gian và nguồn lực của cả gia đình và xã hội. Thực trạng này đặt ra yêu cầu cấp thiết về việc nghiên cứu và phát triển các công cụ hỗ trợ ra quyết định thông minh, ứng dụng các thuật toán hiện đại để cá nhân hóa lộ trình chọn ngành cho thí sinh dựa trên các bằng chứng dữ liệu xác thực thay vì các phương thức tư vấn truyền thống vốn đã bộc lộ nhiều khiếm khuyết. Sự dịch chuyển này không chỉ giúp tối ưu hóa kết quả tuyển sinh mà còn là nền tảng để nâng cao chất lượng nguồn nhân lực sau đào tạo, đáp ứng chính xác nhu cầu khắt khe của thị trường lao động trong bối cảnh hội nhập quốc tế.

1.2. Quy trình và nghiệp vụ tuyển sinh

Nghiên cứu về quy trình và nghiệp vụ tuyển sinh là bước đi tiên quyết nhằm chuẩn hóa các luồng dữ liệu và xác định các điểm tiếp xúc quan trọng giữa nhà trường và thí sinh. Việc hiểu rõ cơ chế vận hành của quy chế và quy trình thực tiễn sẽ tạo nền tảng vững chắc cho việc thiết kế và triển khai các mô hình hỗ trợ phù hợp trong các chương tiếp theo.

1.2.1. Quy chế tuyển sinh

Hoạt động tuyển sinh tại các cơ sở giáo dục nghề nghiệp hiện nay được vận hành và điều chỉnh bởi hệ thống các văn bản pháp lý hiện hành, trong đó văn bản mang tính định hướng cốt lõi là Thông tư số 05/2021/TT-BLĐTBXH do Bộ Lao động - Thương binh và Xã hội ban hành [17]. Quy chế này không chỉ thiết lập các tiêu chuẩn khắt khe về mặt pháp lý mà còn tạo ra khung khổ nghiệp vụ thống nhất, đảm bảo tính công bằng và minh bạch trong việc tiếp cận giáo dục nghề nghiệp của mọi đối tượng thí sinh.

Xét về điều kiện xét tuyển, quy chế yêu cầu thí sinh đăng ký dự tuyển trình độ cao đẳng phải đảm bảo đã hoàn thành chương trình giáo dục trung học phổ thông hoặc có bằng tốt nghiệp trình độ tương đương theo quy định của pháp luật. Đây là ngưỡng đảm bảo chất lượng đầu vào quan trọng, giúp nhà trường sàng lọc được những

người học có nền tảng tri thức cơ bản đủ để tiếp thu các chương trình đào tạo chuyên sâu về kỹ thuật và nghiệp vụ tại trường. Bên cạnh các điều kiện về văn bằng, quy chế còn quy định cụ thể về nguyên tắc phân ngành và xét trúng tuyển dựa trên chỉ tiêu đã được phê duyệt trong đề án tuyển sinh hằng năm. Theo đó, việc xét tuyển phải tuân thủ nguyên tắc ưu tiên các thí sinh có kết quả học tập hoặc điểm thi từ cao xuống thấp cho đến khi đủ định mức chỉ tiêu, đảm bảo tính khách quan tối đa trong việc lựa chọn người học.

Đáng chú ý, quy chế tuyển sinh hiện đại cũng cho phép các cơ sở đào tạo linh hoạt trong việc áp dụng các tiêu chí phụ hoặc đa dạng hóa các tổ hợp môn xét tuyển tùy thuộc vào đặc thù của từng nhóm ngành đào tạo. Đối với các ngành trọng điểm như Công nghệ ô tô, Dược hay Điều dưỡng, việc thiết lập các tiêu chuẩn xét tuyển riêng biệt giúp nhà trường tuyển chọn được những cá nhân có năng khiếu, tố chất và năng lực học thuật tương thích nhất với yêu cầu khắt khe của thị trường lao động thực tế. Sự kết hợp giữa các quy định chung của Nhà nước và tính chủ động trong quy chế nội bộ của nhà trường đã tạo ra một nền tảng pháp lý vững chắc, không chỉ phục vụ công tác quản lý hành chính mà còn là tiền đề để triển khai các hệ thống hỗ trợ tư vấn thông minh dựa trên dữ liệu học thuật của thí sinh.

1.2.2. Quy trình tuyển sinh

Quy trình tuyển sinh của Nhà trường được chuẩn hóa thành một chu trình nghiệp vụ khép kín gồm mười bước phối hợp chặt chẽ, đảm bảo tính khoa học từ khâu hoạch định chiến lược đến khâu tiếp nhận sinh viên chính thức.

Bước 1: Tiếp nhận quyết định giao chỉ tiêu tuyển sinh

Giai đoạn khởi đầu mang tính pháp lý then chốt, trong đó Nhà trường tiếp nhận văn bản phê duyệt chỉ tiêu từ các cơ quan quản lý nhà nước có thẩm quyền. Quyết định này thiết lập hạn mức đào tạo tối đa cho từng trình độ và ngành nghề, đóng vai trò là cơ sở định lượng để Nhà trường xây dựng các kế hoạch triển khai cụ thể.

Bước 2: Rà soát năng lực đào tạo nội bộ

Sau khi xác lập chỉ tiêu, Nhà trường tiến hành thực hiện quy trình đánh giá toàn diện năng lực đào tạo. Công tác rà soát tập trung vào việc kiểm chứng khả năng

đáp ứng của hệ thống cơ sở vật chất, trang thiết bị kỹ thuật, trình độ chuyên môn của đội ngũ giảng viên và tính cập nhật của chương trình đào tạo nhằm đảm bảo chất lượng giảng dạy tốt nhất cho người học.

Bước 3: Xây dựng kế hoạch và thành lập Hội đồng tuyển sinh

Trên cơ sở dữ liệu rà soát, Nhà trường tiến hành xây dựng kế hoạch tuyển sinh chi tiết và ban hành quyết định thành lập Hội đồng tuyển sinh. Hội đồng này giữ vai trò điều hành, giám sát và chịu trách nhiệm cao nhất về tính khách quan, công bằng trong toàn bộ các công đoạn của kỳ tuyển sinh.

Bước 4: Thiết lập và công bố Đề án tuyển sinh

Nhà trường tiến hành pháp lý hóa các thông tin tuyển sinh thông qua việc xây dựng Đề án tuyển sinh chi tiết. Văn bản này quy định cụ thể về chỉ tiêu phân bổ cho từng chuyên ngành, các phương thức xét tuyển áp dụng, ngưỡng đảm bảo chất lượng và các điều kiện dự tuyển, sau đó được công bố rộng rãi trên các phương tiện thông tin đại chúng.

Bước 5: Tổ chức triển khai truyền thông tuyển sinh

Hoạt động truyền thông được thực hiện đa kênh nhằm tối ưu hóa khả năng tiếp cận đối tượng thí sinh mục tiêu. Nhà trường sử dụng kết hợp các nền tảng số như website, mạng xã hội cùng các hoạt động tư vấn trực tiếp tại hệ thống các trường Trung học phổ thông để truyền tải thông tin định hướng nghề nghiệp và quảng bá thương hiệu đào tạo.

Bước 6: Tiếp nhận hồ sơ đăng ký và tư vấn định hướng ngành nghề

Giai đoạn này ghi nhận sự tương tác trực tiếp khi Nhà trường tiếp nhận hồ sơ đăng ký dự tuyển qua cả hai phương thức trực tiếp và trực tuyến. Song song với việc kiểm tra tính hợp lệ và minh chứng hành chính, đội ngũ chuyên gia sẽ thực hiện tư vấn chuyên sâu, hỗ trợ thí sinh đối chiếu năng lực cá nhân với đặc thù từng ngành nghề để đưa ra lựa chọn phù hợp nhất.

Bước 7: Tổ chức quy trình xét tuyển chuyên môn

Hội đồng tuyển sinh thực hiện nhiệm vụ xét tuyển căn cứ theo các phương thức đã được xác lập trong đề án, bao gồm việc tổng hợp điểm học bạ, điểm thi tốt

nghiệp và áp dụng các chính sách ưu tiên theo quy định. Quá trình này đòi hỏi sự chính xác tuyệt đối trong khâu xử lý dữ liệu để xác định danh sách thí sinh đủ điều kiện trúng tuyển.

Bước 8: Phê duyệt và công bố kết quả trúng tuyển chính thức

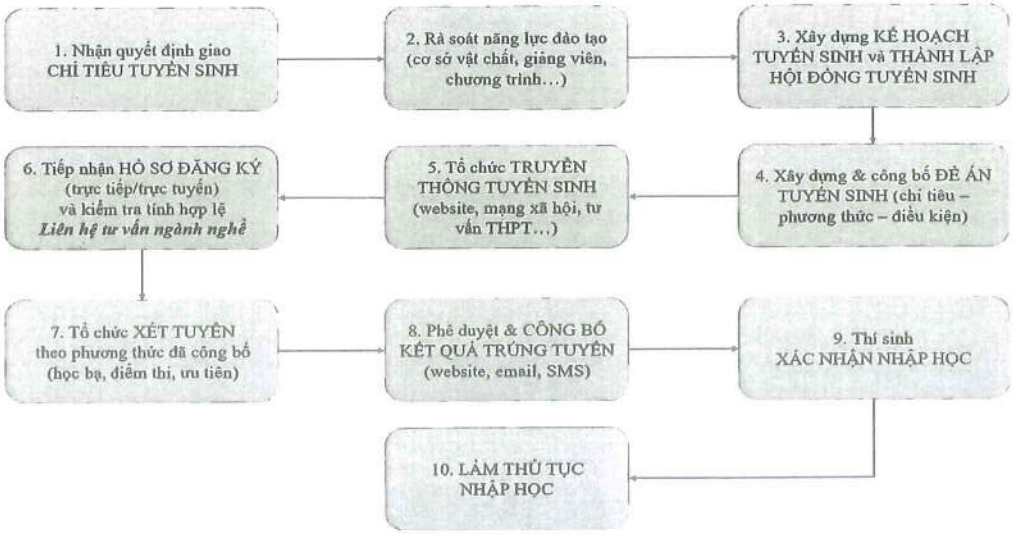
Kết quả sau khi được Hội đồng tuyển sinh phê duyệt sẽ được công bố công khai qua các phương thức tiếp nhận thông tin hiện đại như cổng tra cứu website, hệ thống thư điện tử và tin nhắn SMS. Việc này đảm bảo thông tin được truyền tải nhanh chóng và chính xác tới từng thí sinh.

Bước 9: Thí sinh thực hiện xác nhận nhập học

Căn cứ trên kết quả trúng tuyển đã công bố, thí sinh có trách nhiệm thực hiện các thủ tục xác nhận nhập học trên hệ thống theo đúng khung thời gian quy định. Đây là bước khẳng định nguyện vọng chính thức của người học, làm cơ sở để Nhà trường chuẩn bị các điều kiện hậu cần tiếp đón tân sinh viên.

Bước 10: Hoàn tất thủ tục nhập học và khởi tạo hồ sơ sinh viên

Tại bước cuối cùng, thí sinh đến trực tiếp cơ sở đào tạo để đối chiếu hồ sơ gốc và hoàn thiện các thủ tục hành chính cuối cùng. Toàn bộ dữ liệu tuyển sinh sau đó sẽ được đồng bộ hóa và chuyển giao cho bộ phận quản lý đào tạo để chính thức khởi tạo vòng đời học tập của sinh viên tại Trường.



Hình 1.1: Các bước của quy trình tuyển sinh.

Việc chuẩn hóa quy trình tuyển sinh thành mười bước nghiệp vụ chặt chẽ không chỉ phản ánh tính chuyên nghiệp trong công tác quản trị mà còn đảm bảo sự tuân thủ tuyệt đối các quy định pháp lý từ Bộ Giáo dục và Đào tạo cũng như Tổng cục Giáo dục nghề nghiệp. Chu trình khép kín này thiết lập một hệ thống kiểm soát chất lượng xuyên suốt, từ giai đoạn hoạch định chiến lược tại các bước khởi đầu cho đến khi hoàn tất thủ tục lưu trữ hồ sơ sinh viên ở giai đoạn cuối. Đặc biệt, sự xuất hiện của các khâu tương tác như tư vấn định hướng ngành nghề và xét tuyển đa phương thức tại các bước trung gian chính là "điểm chạm" quan trọng để Nhà trường thấu hiểu năng lực và nguyện vọng của thí sinh.

Dưới góc độ hệ thống thông tin, mười bước này hình thành nên một luồng dữ liệu liên tục, trong đó mỗi giai đoạn vừa kế thừa kết quả từ bước trước đó, vừa tạo ra nguồn tài nguyên số quý báu cho các hoạt động phân tích và dự báo tiếp theo. Việc chuẩn hóa và minh bạch hóa quy trình này không chỉ giúp tối ưu hóa nguồn lực nhân sự, giảm thiểu các rủi ro vận hành trong giai đoạn cao điểm mà còn góp phần xây dựng niềm tin và khẳng định uy tín thương hiệu của Nhà trường đối với thí sinh và xã hội. Đây chính là nền tảng nghiệp vụ cốt yếu để Nhà trường tiếp tục nghiên cứu và tích hợp các giải pháp công nghệ thông minh, hướng tới mục tiêu tự động hóa và cá nhân hóa sâu sắc công tác hỗ trợ tuyển sinh trong tương lai.

1.3. Bài toán hỗ trợ tư vấn tuyển sinh

Trong kỷ nguyên của dữ liệu lớn và trí tuệ nhân tạo, việc chuyển đổi từ mô hình tư vấn truyền thống sang mô hình tư vấn dựa trên dữ liệu là một bước tiến tất yếu nhằm nâng cao hiệu quả công tác tuyển sinh. Bài toán hỗ trợ tư vấn tuyển sinh được đặt ra không chỉ đơn thuần là việc cung cấp thông tin ngành học, mà còn là một quy trình kỹ thuật phức tạp nhằm tìm ra sự tương quan tối ưu giữa các đặc điểm nội tại của thí sinh và yêu cầu đặc thù của các chương trình đào tạo tại nhà trường.

1.3.1. Mô tả bài toán

Về mặt bản chất kỹ thuật, bài toán hỗ trợ tư vấn tuyển sinh được xác lập như một bài toán dự báo và gợi ý dựa trên sự phân tích đa biến dữ liệu. Mục tiêu cốt lõi của mô hình là thiết lập một hàm ánh xạ có khả năng chuyển đổi các đặc trưng nội tại

của thí sinh thành những chỉ dẫn nghề nghiệp cụ thể, đảm bảo sự tương quan tối ưu giữa năng lực cá nhân và yêu cầu đặc thù của các chương trình đào tạo. Cấu trúc của bài toán được định nghĩa qua hai thành phần dữ liệu trọng tâm sau:

Về dữ liệu đầu vào (Input data):

Thành phần đầu vào của bài toán là một tập hợp các đặc trưng đa chiều, bao gồm cả dữ liệu định lượng và dữ liệu định tính. Nhóm dữ liệu định lượng tập trung vào chuỗi kết quả học tập của chín môn văn hóa cốt lõi, được theo dõi xuyên suốt từ năm lớp 10 đến hết học kỳ 1 năm lớp 12. Việc sử dụng dữ liệu chuỗi thời gian này giúp mô hình không chỉ ghi nhận điểm số tĩnh tại một thời điểm mà còn học được xu hướng phát triển năng lực và sự ổn định trong tư duy của thí sinh. Song song với đó, nhóm dữ liệu định tính được tích hợp thông qua chỉ số từ bài trắc nghiệm tâm lý nghề nghiệp RIASEC, cung cấp các bằng chứng về sở thích, thiên hướng cá nhân và đặc điểm tính cách nghề nghiệp của người học. Ngoài ra, các yếu tố về dữ liệu lịch sử liên quan đến tỷ lệ thành công của sinh viên các khóa trước cũng được đưa vào làm biến tham chiếu để tăng cường độ tin cậy cho quá trình tính toán.

Về kết quả đầu ra (Output data):

Mục tiêu đầu ra của bài toán là xác lập một danh sách phân hạng gồm ba ngành học có mức độ tương thích cao nhất với hồ sơ tổng thể của thí sinh. Kết quả này không đơn thuần dựa trên việc đối chiếu điểm sàn xét tuyển mà là sản phẩm của quá trình phân tích sự phù hợp dựa trên năng khiếu tiềm ẩn và khả năng đáp ứng chương trình học của thí sinh đối với 12 ngành đào tạo trọng điểm tại Nhà trường. Việc cung cấp danh sách "Top 3" ngành nghề ưu tiên giúp thu hẹp phạm vi lựa chọn, giảm thiểu sự nhiễu loạn thông tin và đóng vai trò như một chỉ dẫn khoa học quan trọng hỗ trợ thí sinh trong quá trình ra quyết định cuối cùng.

Sự kết hợp chặt chẽ giữa các biến đầu vào mang tính lịch sử và mục tiêu đầu ra mang tính dự báo đã biến bài toán tư vấn tuyển sinh từ một hoạt động mang tính cảm tính trở thành một quy trình kỹ thuật chuẩn xác. Đây là nền tảng quan trọng để triển khai các thuật toán học sâu, đặc biệt là mạng nơ-ron LSTM, nhằm tìm ra những

quy luật tiềm ẩn trong dữ liệu thí sinh, từ đó nâng cao hiệu quả công tác hướng nghiệp và tuyển sinh.

Trong đề án này, bài toán hỗ trợ tư vấn tuyển sinh sẽ được nghiên cứu và triển khai thử nghiệm tại Trường Cao đẳng Hà Nội.

1.3.2. Ý nghĩa bài toán

Việc giải quyết thành công bài toán hỗ trợ tư vấn tuyển sinh không chỉ mang tính chất kỹ thuật thuần túy mà còn hàm chứa những giá trị thực tiễn sâu sắc, đóng góp trực tiếp vào sự phát triển bền vững của cả người học và cơ sở đào tạo. Ý nghĩa của bài toán được thể hiện rõ nét qua hai khía cạnh chiến lược sau:

Tối ưu hóa định hướng nghề nghiệp cho thí sinh: Trong bối cảnh thí sinh thường xuyên đối mặt với sự nhiễu loạn thông tin và áp lực từ các xu hướng đám đông, hệ thống đóng vai trò như một công cụ định hướng khách quan dựa trên bằng chứng dữ liệu xác thực. Bằng cách phân tích sự tương quan giữa năng lực học thuật thực tế và đặc điểm tính cách RIASEC, hệ thống giúp thí sinh nhận diện được những ngành nghề thực sự phù hợp với sở trường cá nhân thay vì lựa chọn theo cảm tính. Điều này có ý nghĩa quan trọng trong việc gia tăng động lực học tập, giảm thiểu tối đa tỷ lệ sinh viên chán học, bỏ học giữa chừng hoặc có nhu cầu chuyển ngành sau khi nhập học, từ đó giúp tiết kiệm đáng kể thời gian và nguồn lực kinh tế cho gia đình và xã hội.

Nâng cao hiệu quả quản trị và chất lượng đào tạo của Nhà trường: Việc ứng dụng hệ thống tư vấn thông minh giúp hiện đại hóa quy trình tuyển sinh, chuyển từ mô hình tiếp nhận hồ sơ thụ động sang mô hình chủ động hỗ trợ người học. Hệ thống cung cấp cơ sở khoa học để Nhà trường phân bổ chỉ tiêu tuyển sinh một cách hợp lý, đảm bảo sinh viên đầu vào có nền tảng năng lực tương thích với chuẩn đầu ra của chương trình đào tạo. Khi người học được xếp vào đúng chuyên môn phù hợp với năng khiếu, khả năng tiếp thu kiến thức và rèn luyện kỹ năng nghề nghiệp sẽ được phát huy tối đa, trực tiếp nâng cao chất lượng nguồn nhân lực sau tốt nghiệp. Bên cạnh đó, việc tiên phong ứng dụng trí tuệ nhân tạo vào công tác tuyển sinh còn giúp

khẳng định vị thế chuyên nghiệp và uy tín của Nhà trường trong hệ thống giáo dục nghề nghiệp, tạo dựng niềm tin vững chắc đối với phụ huynh và cộng đồng xã hội.

Sự kết hợp giữa lợi ích cá nhân của thí sinh và lợi ích chiến lược của Nhà trường đã khẳng định tính cấp thiết của việc phát triển hệ thống hỗ trợ tư vấn. Đây chính là động lực cốt lõi để nghiên cứu chuyển sang các bước phân tích và lựa chọn mô hình học sâu tối ưu nhằm giải quyết bài toán một cách hiệu quả nhất.

1.4. Khảo sát các mô hình giải quyết bài toán

Bài toán hỗ trợ tư vấn tuyển sinh nêu trên thực chất là tư vấn giúp người học lựa chọn ngành học phù hợp dựa trên năng lực và nguyện vọng cá nhân. Bài toán hỗ trợ tư vấn tuyển sinh có thể được giải quyết khi sử dụng các mô hình phân loại dữ liệu. Do đó, một số mô hình học máy [11] và học sâu [2] thường được nghiên cứu ứng dụng để giải quyết bài toán đặt ra.

Trong mục này, đề án khảo sát tổng quan một số mô hình học máy truyền thống và học sâu có thể sử dụng để giải quyết bài toán hỗ trợ tư vấn tuyển sinh. Nội dung chi tiết các mô hình được lựa chọn sẽ trình bày chi tiết trong Chương 2,

1.4.1. Các mô hình dựa trên học máy truyền thống

Các mô hình hồi quy

Hồi quy là nhóm phương pháp học máy dùng để dự đoán một giá trị số liên tục dựa trên các giá trị biến đầu vào. Trong bài toán tuyển sinh nói chung, hồi quy thường được dùng để dự đoán và tư vấn các yêu cầu sau:

(i) Mức độ phù hợp của thí sinh với ngành học.

(ii) Điểm xét tuyển của thí sinh.

(iii) Điểm chuẩn ngành học.

(iv) Xác suất trúng tuyển.

(v) Số lượng hồ sơ đăng ký.

(vi) Xu hướng tuyển sinh các năm tiếp theo.

Hai mô hình hồi quy thường được sử dụng là hồi quy tuyến tính và hồi quy phi tuyến.

a. Hồi quy tuyến tính đa biến

Mô hình hồi quy tuyến tính đa biến được xác lập là một trong những phương pháp thống kê cơ bản và kinh điển nhất trong việc phân tích mối quan hệ giữa một biến phụ thuộc với nhiều biến độc lập khác nhau. Về mặt toán học, nguyên lý của mô hình dựa trên việc tìm kiếm một tổ hợp tuyến tính các trọng số tối ưu sao cho tổng bình phương sai lệch giữa giá trị dự báo và giá trị thực tế là nhỏ nhất. Hệ thống sẽ thực hiện gán các trọng số khác nhau cho từng môn học thành phần dựa trên tính đặc thù của ngành đào tạo; ví dụ, các ngành kỹ thuật sẽ có trọng số cao hơn ở các môn Toán học và Vật lý, trong khi các ngành kinh tế có thể ưu tiên môn Ngoại ngữ. Kết quả dự báo cuối cùng là một chỉ số định lượng phản ánh khả năng thành công của thí sinh, giúp đưa ra cái nhìn sơ bộ về sự phù hợp dựa trên năng lực học thuật hiện tại.

Ưu điểm nổi bật nhất của hồi quy tuyến tính đa biến nằm ở cấu trúc toán học đơn giản, cho phép hệ thống thực hiện các phép tính với tốc độ phản hồi cực nhanh ngay cả trên các tập dữ liệu lớn. Bên cạnh đó, tính minh bạch và khả năng diễn giải cao giúp người dùng và các nhà quản lý giáo dục dễ dàng hiểu được tầm quan trọng của từng môn học đối với kết quả dự báo. Tuy nhiên, hạn chế cốt lõi của phương pháp này là tính cứng nhắc khi chỉ mô tả được các mối quan hệ tuyến tính đơn thuần. Thực tế dữ liệu tuyển sinh cho thấy năng lực của thí sinh thường có sự tương tác phức tạp và phi tuyến giữa các nhóm môn học khác nhau, đồng thời mô hình này hoàn toàn không có khả năng xử lý các dữ liệu mang tính chuỗi thời gian hay xu hướng tiến bộ học tập. Do đó, mặc dù mang lại hiệu quả về mặt chi phí tính toán, hồi quy tuyến tính đa biến thường chỉ đóng vai trò là phương pháp tham chiếu nền tảng để đối chiếu hiệu năng với các kiến trúc học sâu phức tạp hơn trong các giai đoạn thực nghiệm tiếp theo.

b. Hồi quy phi tuyến

Để khắc phục những hạn chế về tính đơn giản của phương pháp tuyến tính, mô hình hồi quy phi tuyến được đưa vào nghiên cứu như một công cụ mạnh mẽ hơn nhằm mô tả các mối quan hệ phức tạp và đa chiều giữa các biến số dữ liệu tuyển sinh. Trong thực tế giáo dục, mối tương quan giữa năng lực học thuật và sự phù hợp nghề nghiệp thường không vận hành theo một tỷ lệ thuận cố định, mà chịu sự chi phối của

các yếu tố tương tác có tính biến thiên lớn.

Mô hình hồi quy phi tuyến sử dụng các hàm toán học có cấu trúc linh hoạt hơn như hàm đa thức bậc cao, hàm mũ, hàm logarit hoặc hàm sigmoid để khớp với tập dữ liệu thực tế. Điều này cho phép hệ thống ghi nhận những biến động không đồng nhất trong hồ sơ thí sinh, ví dụ như việc một sự thay đổi nhỏ trong điểm số của nhóm môn khoa học tự nhiên có thể dẫn đến sự thay đổi mang tính đột biến trong xác suất phù hợp với các ngành kỹ thuật. Bằng cách thiết lập các đường cong dự báo thay vì các đường thẳng, mô hình có thể bám sát hơn vào các đặc trưng đặc thù của dữ liệu đầu vào, từ đó nâng cao độ chính xác trong việc phân tích các chỉ số năng lực và tâm lý RIASEC.

Mặc dù mang lại hiệu quả dự báo ưu việt hơn so với mô hình tuyến tính, hồi quy phi tuyến vẫn đối mặt với những thách thức lớn về mặt kỹ thuật, đặc biệt là hiện tượng quá khớp (overfitting). Khi mô hình quá tập trung vào việc khớp với những biến động nhỏ hoặc độ nhiễu của dữ liệu huấn luyện, khả năng tổng quát hóa trên dữ liệu mới của thí sinh sẽ bị suy giảm đáng kể. Ngoài ra, việc xác định cấu trúc hàm tối ưu và điều chỉnh các tham số phi tuyến đòi hỏi quy trình chuẩn hóa dữ liệu khắt khe cùng thời gian tính toán lớn hơn.

Mô hình Cây quyết định (Decision Tree)

Nguyên lý cơ bản của mô hình cây quyết định là chia dữ liệu thành các nhánh dựa trên thuộc tính. Trong đó, mỗi nút của cây biểu diễn một điều kiện thuộc tính. Nút lá sẽ xác định nhãn phân loại dữ liệu.

Trong bài toán tuyển sinh nói chung, mô hình cây quyết định thường được dùng để dự đoán và tư vấn các yêu cầu sau:

(i) Mức độ phù hợp của thí sinh với ngành học.

(ii) Tư vấn chọn ngành phù hợp.

(iii) Xác định khả năng trúng tuyển.

Mô hình rừng ngẫu nhiên (Random Forests - RF)

Rừng ngẫu nhiên (Random Forest) là một thuật toán học máy có giám sát (Supervised Machine Learning) thuộc nhóm học tổ hợp (Ensemble Learning), được

đề xuất bởi Leo Breiman vào năm 2001. Thuật toán xây dựng nhiều cây quyết định (Decision Trees) từ các tập dữ liệu và tập đặc trưng được chọn ngẫu nhiên, sau đó kết hợp kết quả của các cây để đưa ra dự đoán cuối cùng.

Ý tưởng cốt lõi của Random Forest là: thay vì dựa vào một cây quyết định đơn lẻ dễ bị quá khớp (overfitting), kết hợp nhiều cây quyết định độc lập sẽ tạo ra một mô hình có khả năng dự đoán chính xác và ổn định hơn.

Trong bài toán tuyển sinh nói chung, mô hình rừng ngẫu nhiên thường được dùng để dự đoán và tư vấn các yêu cầu sau:

(i) Mức độ phù hợp của thí sinh với ngành học.

(ii) Tư vấn chọn ngành phù hợp.

(iii) Xác định khả năng trúng tuyển.

Mô hình Máy véc tơ hỗ trợ (Support Vector Machine - SVM)

Nguyên lý cơ bản của mô hình máy vector hỗ trợ là xác định một siêu phẳng tối ưu để phân loại dữ liệu. Siêu phẳng tối ưu được lựa chọn sao cho khoảng cách của các điểm dữ liệu thuộc mỗi loại đến siêu phẳng là lớn nhất có thể.

Trong bài toán tuyển sinh nói chung, mô hình máy vector hỗ trợ thường được dùng để phân loại thí sinh theo nhóm ngành phù hợp.

1.4.2. Các mô hình dựa trên học sâu

Mô hình Mạng nơ-ron tích chập (Convolutional Neural Network - CNN)

Mạng nơ-ron tích chập là một mạng nơ-ron nhiều lớp nhằm tự động học các đặc trưng (feature extraction) từ dữ liệu đầu vào mà không cần con người thiết kế thủ công các đặc trưng như trong các phương pháp học máy truyền thống. Trong đó, lớp tích chập (Convolution Layer) sử dụng bộ lọc (Filter hoặc Kernel) để tìm các đặc trưng của dữ liệu. Trên cơ sở đó, mô hình CNN sẽ xác định nhãn phân loại dữ liệu.

Giai đoạn đầu, mô hình CNN được sử dụng để nhận dạng ảnh. Hiện nay, mô hình CNN cũng được sử dụng cho xử lý văn bản. Trong bài toán tuyển sinh nói chung, mô hình CNN thường được dùng để hỗ trợ các yêu cầu sau:

(i) Phân loại các câu hỏi tuyển sinh.

(ii) Phân tích phản hồi của thí sinh.

Mô hình Mạng nơ-ron hồi quy (Recurrent Neural Network - RNN)

Mạng nơ-ron hồi quy là một mô hình học sâu được thiết kế để xử lý dữ liệu có tính tuần tự (sequential data) hoặc phụ thuộc theo thời gian. Khác với mạng nơ-ron truyền thẳng (Feedforward Neural Network), RNN có khả năng ghi nhớ thông tin từ các bước trước và sử dụng thông tin đó để xử lý dữ liệu ở bước hiện tại. Ý tưởng cốt lõi của RNN là sử dụng một trạng thái ẩn (Hidden State) để lưu giữ thông tin từ quá khứ.

Trong bài toán tuyển sinh nói chung, mô hình RNN thường được dùng để hỗ trợ các yêu cầu sau:

(i) Tư vấn chọn ngành học phù hợp.

(ii) Phân tích dữ liệu tuyển sinh theo thời gian.

Mô hình Mạng bộ nhớ ngắn hạn – dài hạn (Long Short-Term Memory – LSTM)

Mô hình Mạng bộ nhớ ngắn hạn - dài hạn (LSTM) có thể xem như mô hình cải tiến của mô hình RNN. Mô hình mạng nơ-ron học sâu LSTM đại diện cho một bước tiến quan trọng trong việc xử lý các dạng dữ liệu có tính chuỗi và phụ thuộc thời gian, khắc phục triệt để những hạn chế của các phương pháp hồi quy truyền thống. Khác với các mạng nơ-ron hồi quy (RNN) thông thường hay gặp phải hiện tượng triệt tiêu đạo hàm, LSTM được thiết kế với cấu trúc các cổng (gate) đặc biệt, cho phép mô hình chủ động ghi nhớ hoặc loại bỏ thông tin trong khoảng thời gian dài.

Trong bài toán hỗ trợ tư vấn tuyển sinh, LSTM thể hiện ưu thế vượt trội khi có khả năng khai thác tối đa giá trị của chuỗi điểm số học tập xuyên suốt các năm học bậc trung học phổ thông của thí sinh. Thay vì chỉ đánh giá dựa trên điểm trung bình cuối cùng một cách tĩnh tại, mô hình có thể học được "vết" năng lực, tức là xu hướng tiến bộ hoặc sự ổn định của thí sinh qua từng học kỳ. Điều này cho phép hệ thống đưa ra những nhận định sâu sắc hơn về tiềm năng phát triển của người học trong một chuyên ngành cụ thể, ví dụ như nhận diện một thí sinh có sự bứt phá ở các môn kỹ thuật vào năm lớp 12.

Mô hình LSTM không chỉ xử lý hiệu quả các dữ liệu định lượng mà còn thể hiện sự ưu việt trong việc kết hợp đồng thời với các dữ liệu định tính như sở thích và

đặc điểm tính cách RIASEC. Cấu trúc ô nhớ (cell state) của LSTM đóng vai trò như một bộ lọc thông minh, giúp lưu giữ những đặc trưng quan trọng nhất từ hồ sơ lịch sử của thí sinh để phục vụ cho việc dự báo. Nhờ khả năng mô hình hóa các mối quan hệ phi tuyến phức tạp và tính phụ thuộc xa giữa các biến số, LSTM cho phép xây dựng một hệ thống tư vấn có tính cá nhân hóa cao, đảm bảo kết quả gợi ý ngành nghề đạt độ chính xác và tin cậy trong môi trường thực tế.

Mô hình Transformer

Mô hình Transformer là một kiến trúc học sâu sử dụng cơ chế tự chú ý (Self-Attention) để mô hình hóa các phụ thuộc trong dữ liệu tuần tự mà không cần sử dụng các kết nối hồi quy như RNN hay LSTM. Trong đó, cơ chế tự chú ý là nguyên lý hoạt động cốt lõi của mô hình Transformer và là thành phần quan trọng nhất của Transformer.

Đây là hướng nghiên cứu nổi bật từ năm 2018 đến nay.

Trong bài toán tuyển sinh nói chung, mô hình Transformer thường được dùng để hỗ trợ các yêu cầu sau:

(i) Tư vấn ngành nghề phù hợp cho thí sinh.

(ii) Xây dựng Chatbot tư vấn tuyển sinh thông minh.

Trong ngữ cảnh của bài toán hỗ trợ tuyển sinh tại Trường Cao đẳng Hà Nội, mô hình này được vận hành dựa trên giả định rằng mức độ tương thích của một thí sinh đối với một chuyên ngành cụ thể có thể được biểu diễn dưới dạng một hàm bậc nhất của các thành phần điểm số môn học và các chỉ số đặc trưng cá nhân.

Tại Trường Cao đẳng Hà Nội, với đặc thù dữ liệu thí sinh đa dạng về vùng miền và trình độ học vấn, việc áp dụng mô hình phi tuyến cần được cân nhắc kỹ lưỡng để đảm bảo tính ổn định và tránh sai số hệ thống khi triển khai thực tế.

1.5. Kết luận chương 1

Chương 1 của đề án đã hoàn thành khảo sát tổng quan vấn đề tuyển sinh tại các trường cao đẳng, trong đó trình bày chi tiết quy chế tuyển sinh và hiện trạng tuyển sinh trong thời gian gần đây. Từ đó, đề án đã khảo sát quy trình và nghiệp vụ tuyển sinh tại các trường cao đẳng.

Trong chương 1, đề án đã mô tả bài toán hỗ trợ tư vấn tuyển sinh và khảo sát tổng quan các mô hình học máy và học sâu để giải quyết bài toán này. Với mục tiêu triển khai thử nghiệm hệ thống hỗ trợ quản lý tuyển sinh tại Trường Cao đẳng Hà Nội, trong Chương 2 tiếp theo, đề án sẽ nghiên cứu chi tiết các mô hình hồi quy Logistic, mô hình rừng ngẫu nhiên và mô hình học sâu LSTM cho bài toán hỗ trợ tư vấn tuyển sinh và một số vấn đề liên quan.

CHƯƠNG 2: PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG

2.1. Thực trạng công tác tuyển sinh tại Trường Cao đẳng Hà Nội

Trường Cao đẳng Hà Nội (HNC), tiền thân là Trường Trung cấp Y Dược Hà Nội được thành lập ngày 15/08/2007, là cơ sở giáo dục nghề nghiệp tư thục vận hành dưới sự quản lý nhà nước của Bộ Lao động – Thương binh và Xã hội. Trải qua gần hai thập kỷ xây dựng và phát triển, nhà trường đã thực hiện bước chuyển đổi chiến lược từ mô hình chuyên ngành y dược sang đa ngành, đa lĩnh vực, vận hành nhất quán theo triết lý cốt lõi "Thực học – Thực hành – Thực nghiệp – Thực tài". Hiện nay, trường tọa lạc tại trụ sở chính ở Km3+350 đường Phan Trọng Tuệ, xã Tam Hiệp, huyện Thanh Trì, thành phố Hà Nội, với mục tiêu chiến lược là trở thành trường cao đẳng trọng điểm quốc gia trước năm 2030 và từng bước chuyển dịch thành cơ sở giáo dục đại học định hướng ứng dụng trong giai đoạn tiếp theo.

Về năng lực đào tạo và vị thế thực tiễn, HNC hiện đang tổ chức giảng dạy các khối ngành mũi nhọn có nhu cầu nhân lực cao như Khoa học Sức khỏe, Ngôn ngữ & Văn hóa, Kỹ thuật - Công nghệ, và Kinh tế - Du lịch, bên cạnh hệ Phổ thông Cao đẳng (Mô hình 9+). Nhà trường sở hữu hệ thống hạ tầng kỹ thuật, phòng lý thuyết và cơ sở thực hành chuyên sâu đạt chuẩn, đồng thời thiết lập mạng lưới hợp tác chiến lược sâu rộng với các doanh nghiệp, tập đoàn lớn trong và ngoài nước (như Nhật Bản, Hàn Quốc, CHLB Đức). Mô hình liên kết chặt chẽ này không chỉ tối ưu hóa tỷ lệ giải quyết việc làm cho người học mà còn khẳng định vai trò của nhà trường trong việc cung ứng nguồn lao động kỹ thuật chất lượng cao cho khu vực kinh tế phía Bắc, biến nơi đây thành một địa bàn khảo sát mang tính đại diện và có giá trị thực tiễn cao cho đề tài nghiên cứu.

Tại Trường Cao đẳng Hà Nội, công tác tuyển sinh đang bộc lộ những vấn đề đặc thù cần được giải quyết cấp thiết nhằm nâng cao vị thế cạnh tranh của Nhà trường trong hệ thống giáo dục nghề nghiệp. Với quy mô tuyển sinh hằng năm tương đối lớn, dao động từ 2.000 đến 3.000 chỉ tiêu, Nhà trường hiện đang phải đối mặt với việc quản lý một khối lượng dữ liệu hồ sơ khổng lồ và đa dạng về phương thức xét tuyển. Tuy nhiên, kết quả khảo sát thực tế cho thấy quy trình tiếp nhận và xử lý thông

tin hiện tại vẫn còn mang nặng tính thủ công hoặc bán thủ công, dẫn đến tình trạng tiêu tốn đáng kể nguồn nhân lực và thời gian, đặc biệt là trong những giai đoạn cao điểm của mùa tuyển sinh.

Điểm yếu cốt lõi và mang tính chiến lược hiện nay là Nhà trường vẫn chưa xây dựng được một hệ thống tư vấn thông minh có khả năng khai thác sâu nguồn dữ liệu lịch sử phong phú để hỗ trợ người học trong quá trình chọn ngành. Các thông tin quan trọng về điểm số học tập và nguyện vọng của thí sinh hiện mới chỉ phục vụ cho mục đích lưu trữ hành chính đơn thuần mà chưa được đưa vào các mô hình học máy để phân tích xu hướng hoặc đánh giá mức độ phù hợp nghề nghiệp một cách khoa học. Việc thiếu vắng các công cụ hỗ trợ tư vấn dựa trên nền tảng dữ liệu xác thực không chỉ làm giảm hiệu suất vận hành của bộ phận tuyển sinh mà còn gián tiếp ảnh hưởng đến chất lượng đào tạo đầu ra. Khi thí sinh lựa chọn ngành học không tương thích với sở trường và năng lực cá nhân, rủi ro về việc sinh viên thiếu động lực học tập hoặc không đáp ứng được yêu cầu chuyên môn sau khi tốt nghiệp là rất lớn, gây ảnh hưởng trực tiếp đến uy tín và thương hiệu đào tạo của Nhà trường trên thị trường lao động.

2.2. Nghiên cứu mô hình hỗ trợ thí sinh chọn ngành phù hợp

Trong lĩnh vực học máy và học sâu, việc lựa chọn mô hình phù hợp đóng vai trò quyết định đến độ chính xác, khả năng giải thích và tính khả thi của hệ thống tư vấn. Đối với bài toán hỗ trợ tuyển sinh, các mô hình được khảo sát cần xử lý được đồng thời dữ liệu điểm học tập, đặc điểm định hướng nghề nghiệp của thí sinh và bảo đảm khả năng vận hành ổn định khi triển khai trong thực tế.

2.2.1. Mô hình hồi quy Logistic

Logistic Regression [6] là mô hình được sử dụng trong đề án để xây dựng mô hình phân loại ngành học cho thí sinh. Do bài toán tư vấn tuyển sinh là bài toán phân loại nhiều lớp, mô hình Logistic Regression được mở rộng theo dạng đa lớp (Multinomial Logistic Regression) sử dụng hàm Softmax để tính xác suất thuộc về từng ngành học.

Với bài toán phân loại hai lớp, công thức Logistic Regression trong tài liệu tham khảo được biểu diễn như sau:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \quad (2.1)$$

Trong đó:

- $p(X)$ là xác suất mẫu dữ liệu có đặc trưng X thuộc về lớp cần dự đoán;
- β_0 là hệ số chặn;
- β_1 là hệ số hồi quy của biến đầu vào X.

Công thức trên có thể viết lại dưới dạng odds:

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 X} \quad (2.2)$$

Lấy logarit hai vế, ta thu được hàm logit:

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X \quad (2.3)$$

Điểm quan trọng của Logistic Regression là mặc dù mô hình dùng để dự đoán xác suất phân loại, phần logit vẫn là một tổ hợp tuyến tính của các biến đầu vào. Đây là lý do Logistic Regression được xếp vào nhóm mô hình phân loại tuyến tính.

Khi có nhiều biến đầu vào, mô hình Logistic Regression có thể được mở rộng như sau:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n}} \quad (2.4)$$

Trong đó:

- $X_1, X_2, \ldots, X_n$ là các đặc trưng đầu vào;
- $\beta_1, \beta_2, \ldots, \beta_n$ là các hệ số tương ứng với từng đặc trưng;
- $p(X)$ là xác suất dự đoán của mô hình.

Trong đề án này, bài toán hỗ trợ tư vấn tuyển sinh không phải là bài toán phân loại hai lớp, mà là bài toán phân loại nhiều lớp, trong đó mỗi lớp tương ứng với một ngành học. Cụ thể, hệ thống cần dự đoán và gợi ý ngành học phù hợp cho thí sinh trong số 12 ngành đào tạo. Vì vậy, công thức Logistic Regression hai lớp trong tài liệu tham khảo được sử dụng làm cơ sở lý thuyết, sau đó được mở rộng

sang Logistic Regression đa lớp thông qua hàm Softmax để phù hợp với yêu cầu thực nghiệm tại Chương 3.

Giả sử mỗi thí sinh được biểu diễn bởi vector đặc trưng:

$$x = (x_1, x_2, \dots, x_n)^T \tag{2.5}$$

Trong đó, $x_1, x_2, \dots, x_n$ là các đặc trưng đầu vào của thí sinh. Trong phạm vi đề án, các đặc trưng này bao gồm nhóm điểm học tập của các môn văn hóa và nhóm đặc trưng định hướng nghề nghiệp RIASEC. Gọi K là số ngành học cần phân loại, mỗi ngành học tương ứng với một lớp đầu ra k, với $k = 1,2, \dots, K$.

Với mỗi ngành học k, mô hình tính một điểm số tuyến tính:

$$z_k = \beta_{0k} + \beta_{1k}x_1 + \beta_{2k}x_2 + \cdots + \beta_{nk}x_n \tag{2.6}$$

Trong đó:

- z_k là điểm số tuyến tính của thí sinh đối với ngành học k;
- β_{0k} là hệ số chặn của lớp k;
- $\beta_{1k}, \beta_{2k}, \dots, \beta_{nk}$ là các hệ số trọng số của các đặc trưng đầu vào đối với ngành học k;
- $x_1, x_2, \dots, x_n$ là các đặc trưng mô tả hồ sơ của thí sinh.

Do có nhiều ngành học cần phân loại, các điểm số tuyến tính z_k được chuyển thành xác suất bằng hàm Softmax:

$$P(y = k \mid x) = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}} \tag{2.7}$$

Trong đó:

- $P(y = k \mid x)$ là xác suất thí sinh có hồ sơ x phù hợp với ngành học k;
- e^{z_k} là giá trị mũ của điểm số tuyến tính đối với ngành k;
- $\sum_{j=1}^{K} e^{z_j}$ là tổng giá trị mũ của toàn bộ K ngành học.

Ngành học được dự đoán là ngành có xác suất lớn nhất:

$$\hat{y} = \arg\max_{k} P(y = k \mid x) \tag{2.8}$$

Trong đó, $\hat{y}$ là nhãn ngành học dự đoán cho thí sinh.

Ngoài việc xác định một ngành học có xác suất cao nhất, mô hình còn có thể

sắp xếp các ngành học theo xác suất giảm dần để đưa ra danh sách Top-3 ngành học phù hợp nhất:

$$\text{Top3}(x) = \text{ba ngành học có giá trị } P(y = k \mid x) \text{ lớn nhất}$$

Trong quá trình huấn luyện, mô hình Logistic Regression đa lớp tìm các hệ số β sao cho xác suất dự đoán cho ngành đúng là lớn nhất. Hàm mất mát thường được sử dụng là Cross-Entropy đa lớp:

$$L(\beta) = -\sum_{i=1}^{N}\sum_{k=1}^{K} I\,(y_i = k)\log P(y_i = k \mid x_i) \qquad (2.9)$$

Trong đó:

- N là số lượng mẫu dữ liệu huấn luyện;
- K là số ngành học;
- $I(y_i = k)$ là hàm chỉ báo, nhận giá trị 1 nếu mẫu thứ i thuộc ngành k, ngược lại nhận giá trị 0;
- $P(y_i = k \mid x_i)$ là xác suất mô hình dự đoán mẫu thứ i thuộc ngành k.

Để hạn chế hiện tượng quá khớp và giúp mô hình ổn định hơn khi áp dụng trên dữ liệu mới, có thể bổ sung thành phần điều chuẩn L2 vào hàm mất mát:

$$L(\beta) = -\sum_{i=1}^{N}\sum_{k=1}^{K} I\,(y_i = k)\log P(y_i = k \mid x_i) + \lambda \sum_{k=1}^{K} \| \beta_k \|_2^2 \qquad (2.10)$$

Trong đó, λ là hệ số điều chuẩn, kiểm soát mức độ phạt đối với các hệ số có giá trị lớn.

Như vậy, công thức Logistic Regression hai lớp trong tài liệu tham khảo được sử dụng làm nền tảng lý thuyết cho hướng tiếp cận tuyến tính. Do bài toán của đề án là phân loại 12 ngành học và cần xuất xác suất phù hợp cho từng ngành, mô hình được chuyển đổi sang dạng Logistic Regression đa lớp với hàm Softmax. Việc chuyển đổi này là cần thiết để bảo đảm mô hình phù hợp với cấu trúc dữ liệu và mục tiêu thực nghiệm tại Chương 3.

Trong đề án này, Logistic Regression đa lớp được lựa chọn làm mô hình cơ sở tuyến tính vì ba lý do chính. Thứ nhất, mô hình phù hợp với bản chất bài toán phân loại

nhiều ngành học. Thứ hai, mô hình có khả năng xuất ra xác suất cho từng ngành học, từ đó hỗ trợ trực tiếp mục tiêu gợi ý Top-3 ngành học phù hợp cho thí sinh. Thứ ba, mô hình có cấu trúc tương đối đơn giản, dễ triển khai và phù hợp để làm mốc đối sánh với mô hình phi tuyến Random Forest và mô hình học sâu LSTM trong Chương 3.

Tuy nhiên, hạn chế của Logistic Regression đa lớp là giả định ranh giới phân loại giữa các ngành học có thể được biểu diễn bằng các tổ hợp tuyến tính của đặc trưng đầu vào. Trong thực tế, sự phù hợp giữa thí sinh và ngành học có thể phụ thuộc vào các quan hệ phức tạp hơn, chẳng hạn sự tương tác giữa điểm học tập, xu hướng nghề nghiệp RIASEC và đặc thù từng ngành đào tạo. Vì vậy, bên cạnh mô hình tuyến tính, đề án tiếp tục nghiên cứu mô hình phi tuyến nhằm đánh giá khả năng khai thác các quan hệ phức tạp hơn trong dữ liệu tuyển sinh.

2.2.2. Mô hình rừng ngẫu nhiên

Random Forest [11] là mô hình học máy được lựa chọn trong đề án để xây dựng mô hình phân loại ngành học cho thí sinh. Mô hình được xây dựng từ nhiều cây quyết định và sử dụng cơ chế kết hợp kết quả dự đoán của các cây nhằm nâng cao độ chính xác phân loại.

Về nguyên lý, Random Forest xây dựng B cây quyết định, ký hiệu là:

$$h_1(x), h_2(x), \dots, h_B(x) \tag{2.11}$$

Trong đó:

- B là số lượng cây quyết định trong rừng;
- $h_b(x)$ là kết quả dự đoán của cây quyết định thứ b đối với hồ sơ thí sinh x.

Mỗi cây quyết định được huấn luyện trên một tập dữ liệu con được lấy mẫu có hoàn lại từ tập dữ liệu huấn luyện ban đầu. Phương pháp này được gọi là Bootstrap. Nhờ Bootstrap, mỗi cây trong rừng được huấn luyện trên một biến thể khác nhau của dữ liệu, làm tăng tính đa dạng giữa các cây.

Tại mỗi nút phân tách của cây quyết định, thuật toán không xét toàn bộ các đặc trưng đầu vào mà chỉ chọn ngẫu nhiên một tập con các đặc trưng để tìm điều kiện phân tách tốt nhất. Với bài toán phân loại, tài liệu tham khảo nêu cách chọn

thường dùng là lấy khoảng $\sqrt{p}$ đặc trưng trong tổng số p đặc trưng tại mỗi lần phân tách. Cách chọn ngẫu nhiên này giúp giảm tương quan giữa các cây, từ đó làm cho kết quả tổng hợp của Random Forest ổn định hơn so với một cây quyết định đơn lẻ.

Đối với bài toán phân loại, một tiêu chí phổ biến để lựa chọn điểm phân tách tại mỗi nút là chỉ số Gini. Chỉ số Gini tại một nút dữ liệu m được tính như sau:

$$Gini(m) = 1 - \sum_{k=1}^{K} p_{mk}^2 \tag{2.12}$$

Trong đó:

- K là số ngành học cần phân loại;
- p_{mk} là tỷ lệ mẫu thuộc ngành học k tại nút m.

Giá trị Gini càng nhỏ thì các mẫu trong nút càng thuần nhất, nghĩa là phần lớn mẫu tại nút đó thuộc về cùng một ngành học. Khi xây dựng cây quyết định, thuật toán sẽ lựa chọn phép phân tách giúp làm giảm chỉ số Gini nhiều nhất.

Sau khi huấn luyện B cây quyết định, Random Forest đưa ra nhãn dự đoán cuối cùng bằng cơ chế bỏ phiếu đa số:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \dots, h_B(x)\} \tag{2.13}$$

Trong đó, $\hat{y}$ là ngành học được dự đoán cuối cùng, còn $mode$ là nhãn ngành học xuất hiện nhiều nhất trong kết quả dự đoán của các cây.

Bên cạnh nhãn dự đoán cuối cùng, Random Forest cũng có thể ước lượng xác suất thí sinh thuộc về từng ngành học dựa trên tỷ lệ số cây bỏ phiếu cho ngành đó:

$$P(y = k \mid x) = \frac{1}{B} \sum_{b=1}^{B} I\left(h_b(x) = k\right) \tag{2.14}$$

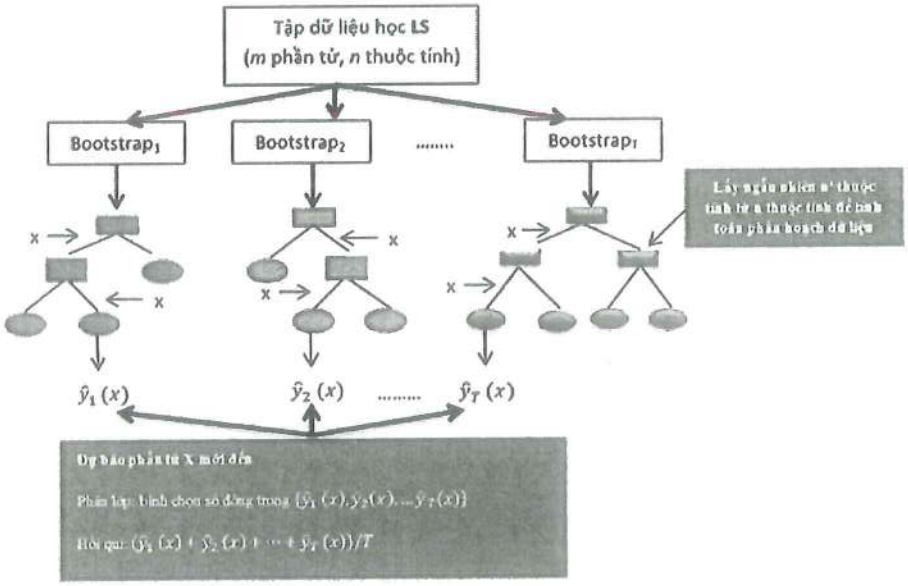
Trong đó:

- $P(y = k \mid x)$ là xác suất thí sinh có hồ sơ x phù hợp với ngành học k;
- B là tổng số cây quyết định;
- $I(h_b(x) = k)$ là hàm chỉ báo, nhận giá trị 1 nếu cây thứ b dự đoán ngành k, ngược lại nhận giá trị 0.

Từ các xác suất này, hệ thống có thể sắp xếp các ngành học theo mức độ phù hợp và đưa ra danh sách Top-3 ngành học có xác suất cao nhất cho thí sinh:

$$\text{Top3}(x) = \text{ba ngành học có giá trị } P(y = k \mid x) \text{ lớn nhất}$$

Các công thức biểu diễn cơ chế dự đoán của Random Forest trong đề án được xây dựng trên cơ sở nguyên lý đã trình bày trong tài liệu tham khảo. Tài liệu mô tả Random Forest là phương pháp kết hợp nhiều cây quyết định, trong đó mỗi cây được huấn luyện trên một mẫu Bootstrap và tại mỗi nút phân tách chỉ xem xét ngẫu nhiên một tập con các đặc trưng đầu vào. Trên cơ sở đó, đề án biểu diễn kết quả dự đoán cuối cùng bằng cơ chế bỏ phiếu đa số giữa các cây quyết định. Đồng thời, để phù hợp với mục tiêu gợi ý Top-3 ngành học, xác suất thí sinh thuộc về
từng ngành được ước lượng thông qua tỷ lệ số cây dự đoán ngành đó.



Hình 2.1. Giải thuật rừng ngẫu nhiên

Random Forest Classifier được lựa chọn trong đề án này làm mô hình cơ sở phi tuyến vì một số lý do. Thứ nhất, mô hình phù hợp với bài toán phân loại ngành học. Thứ hai, mô hình có khả năng biểu diễn quan hệ phi tuyến và tương tác giữa các đặc trưng đầu vào, phù hợp với dữ liệu tuyển sinh có sự kết hợp giữa điểm học tập và RIASEC. Thứ ba, mô hình có khả năng xuất ra xác suất dự đoán cho từng

ngành học, đáp ứng mục tiêu gợi ý Top-3 ngành phù hợp. Thứ tư, Random Forest có độ ổn định tương đối cao nhờ cơ chế kết hợp nhiều cây quyết định.

Ngoài ra, Random Forest còn có thể hỗ trợ đánh giá mức độ quan trọng của các đặc trưng đầu vào thông qua mức giảm độ hỗn tạp tại các nút phân tách. Điều này có ý nghĩa thực tiễn trong bài toán tuyển sinh, vì nhà trường có thể phân tích được nhóm yếu tố nào có ảnh hưởng lớn hơn đến kết quả tư vấn ngành học, chẳng hạn điểm học tập, nhóm môn học hoặc đặc điểm RIASEC.

Tuy nhiên, Random Forest vẫn có một số hạn chế. So với Logistic Regression, mô hình khó diễn giải trực tiếp hơn do kết quả được tổng hợp từ nhiều cây quyết định. Bên cạnh đó, Random Forest xử lý dữ liệu chủ yếu dưới dạng vector đặc trưng tĩnh, chưa khai thác sâu được quan hệ phụ thuộc theo thời gian của quá trình học tập như mô hình LSTM. Vì vậy, trong Chương 3, Random Forest được sử dụng như mô hình cơ sở phi tuyến để so sánh với Logistic Regression đa lớp và mô hình LSTM_Net, qua đó đánh giá mức độ hiệu quả của từng hướng tiếp cận đối với bài toán gợi ý ngành học phù hợp.

2.2.3. Mô hình LSTM

LSTM (Long Short-Term Memory) là một dạng mạng nơ-ron hồi tiếp (Recurrent Neural Network - RNN) được đề xuất nhằm khắc phục hiện tượng mất mát hoặc bùng nổ gradient trong quá trình huấn luyện RNN truyền thống. Đặc điểm nổi bật của LSTM là khả năng ghi nhớ thông tin trong khoảng thời gian dài thông qua cơ chế cổng, nhờ đó mô hình có thể học được các quan hệ phụ thuộc theo thời gian trong dữ liệu.

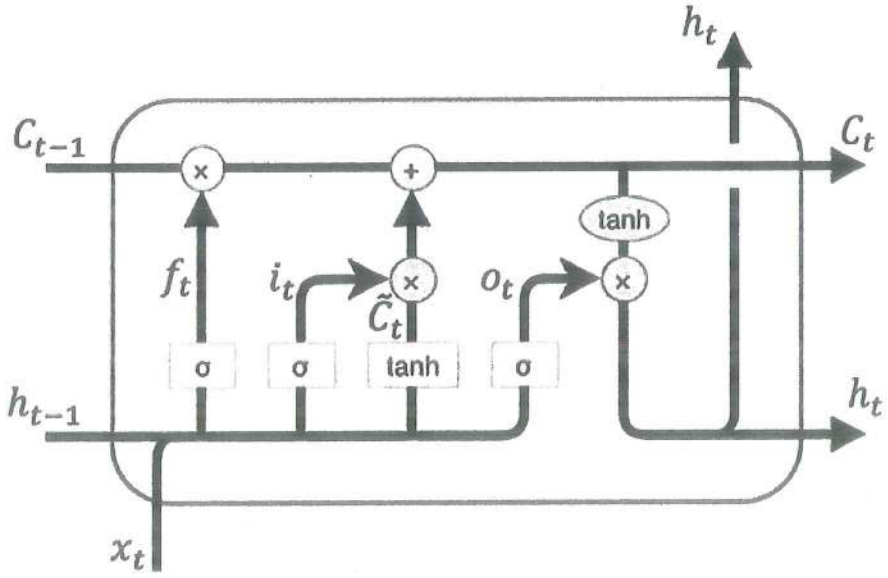
Một đơn vị LSTM bao gồm ba cổng chính: cổng quên (forget gate), cổng vào (input gate) và cổng ra (output gate). Các cổng này cho phép mô hình quyết định thông tin nào cần giữ lại, thông tin nào cần loại bỏ và thông tin nào được sử dụng để tạo ra đầu ra tại từng bước thời gian. Vì vậy, LSTM phù hợp với các bài toán có dữ liệu tuần tự và phụ thuộc theo thời gian.

Về mặt toán học, LSTM thực hiện các phép biến đổi tuyến tính kết hợp với các hàm kích hoạt phi tuyến để cập nhật trạng thái ô nhớ và trạng thái ẩn tại mỗi bước

thời gian. Cơ chế này giúp mô hình duy trì, chọn lọc và khai thác thông tin lịch sử trong quá trình dự đoán.

Cấu trúc tổng quát của mô hình Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) là một biến thể cải tiến của RNN. Mô hình này sử dụng ô nhớ (cell state) và hệ thống các cổng điều khiển để lưu giữ thông tin quan trọng trong chuỗi dữ liệu, đồng thời loại bỏ những thông tin ít liên quan đến nhiệm vụ dự đoán.



Hình 2.2. Cấu trúc của LSTM [2]

Cấu trúc của LSTM

Một đơn vị LSTM bao gồm các thành phần chính sau:

- **Cell State (C_t):**
  Là bộ nhớ dài hạn, cho phép thông tin được truyền xuyên suốt các bước thời gian.
- **Forget Gate (Cổng quên):**
  Quyết định thông tin nào từ trạng thái trước đó cần được giữ lại hoặc loại bỏ.
- **Input Gate (Cổng vào):**
  Xác định thông tin mới nào sẽ được thêm vào ô nhớ.

- **Output Gate (Cổng ra)**:

  Quyết định phần thông tin nào từ ô nhớ sẽ được xuất ra làm đầu ra.

Nhờ cơ chế này, LSTM có khả năng học được các mối quan hệ phức tạp và dài hạn trong dữ liệu chuỗi.

Mô hình toán học của LSTM

Mô hình Long Short-Term Memory (LSTM) là một dạng mạng nơ-ron hồi tiếp đặc biệt, trong đó trạng thái tại mỗi thời điểm được điều khiển thông qua các cổng logic. Tại thời điểm t, LSTM nhận đầu vào x_t, trạng thái ẩn trước đó h_(t-1) và trạng thái ô nhớ trước đó C_(t-1).

a . Cổng quên (Forget Gate)

Cổng quên quyết định mức độ giữ lại thông tin từ trạng thái ô nhớ trước đó. Công thức của cổng quên được biểu diễn như sau:

$$ft=\sigma(Wf\cdot[ht-1,xt]+bf) \tag{2.15}$$

Trong đó:

• f_t là giá trị cổng quên tại thời điểm t;

• W_f và b_f lần lượt là ma trận trọng số và vector độ lệch của cổng quên;

□ σ là hàm sigmoid, cho giá trị trong khoảng [0,1]

b. Cổng vào (Input Gate)

Cổng vào xác định thông tin mới sẽ được bổ sung vào ô nhớ. Cổng này bao gồm hai thành phần:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2.16}$$

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{2.17}$$

Trong đó:

- i_t là cổng vào
- $\tilde{C}_t$ là giá trị ô nhớ tạm thời
- $\tanh$ là hàm hyperbolic tangent, giới hạn giá trị trong $[-1, 1]$

c. Cập nhật trạng thái ô nhớ (Cell State)

Trạng thái ô nhớ tại thời điểm t được cập nhật dựa trên trạng thái ô nhớ trước đó và thông tin mới được lựa chọn thông qua cổng vào:

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \tag{2.18}$$

Trong đó:

- $\odot$ là phép nhân theo từng phần tử (element-wise multiplication)

Công thức này cho phép LSTM vừa **ghi nhớ thông tin cũ**, vừa **tiếp nhận thông tin mới** một cách có chọn lọc.

d. Cổng ra (Output Gate)

Cổng ra quyết định phần thông tin nào từ ô nhớ sẽ được xuất ra làm trạng thái ẩn:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{2.19}$$

$$h_t = o_t \odot \tanh(C_t) \tag{2.20}$$

Trong đó:

- h_t là trạng thái ẩn tại thời điểm t, được sử dụng cho dự đoán đầu ra hoặc truyền sang bước thời gian tiếp theo.

e. Ý nghĩa trong bài toán tư vấn ngành học

Trong hệ thống tư vấn ngành học cao đẳng:

- x_t có thể đại diện cho thông tin học tập của học sinh tại thời điểm hoặc giai đoạn học tập t;
- C_t lưu giữ xu hướng học tập và định hướng nghề nghiệp dài hạn;
- h_t được sử dụng để dự đoán ngành học phù hợp.

Nhờ hệ thống các cổng điều khiển và trạng thái ô nhớ, mô hình LSTM có khả năng khai thác mối quan hệ dài hạn trong dữ liệu học sinh, từ đó nâng cao độ chính xác và tính cá nhân hóa của kết quả tư vấn.

Nền tảng lý thuyết của hệ thống dựa trên sự kết hợp giữa các phương pháp học máy truyền thống và kỹ thuật học sâu hiện đại. Mặc dù các mô hình tuyến tính và phi tuyến cung cấp những cách tiếp cận quan trọng để mô tả mối liên hệ giữa các biến số, nhưng đối với dữ liệu tuyển sinh có yếu tố biến động theo thời gian, các mô hình này còn hạn chế trong việc ghi nhớ đặc trưng chuỗi. Do đó, nghiên cứu tập trung vào

mạng nơ-ron hồi tiếp và đặc biệt là biến thể LSTM, cho phép mô hình lưu giữ thông tin quan trọng từ các giai đoạn học tập trước đến giai đoạn hiện tại, qua đó hình thành cơ sở đánh giá năng lực học tập mang tính xuyên suốt.

2.2.4. Tiêu chí lựa chọn mô hình

Để xác định kiến trúc thuật toán phù hợp nhằm tích hợp vào hệ thống hỗ trợ quản lý tuyển sinh và tư vấn hướng nghiệp, nghiên cứu thiết lập các tiêu chí đánh giá mang tính hệ thống. Các tiêu chí này là cơ sở để sàng lọc, so sánh khách quan giữa các mô hình học máy truyền thống và mô hình học sâu, đồng thời bảo đảm giải pháp đề xuất có khả năng triển khai trong thực tiễn. Quá trình lựa chọn mô hình dựa trên hai nhóm tiêu chí cốt lõi sau:

–Độ chính xác và chất lượng dự đoán của mô hình (Model Performance): Đây là tiêu chí kỹ thuật quan trọng nhất đối với hệ thống tư vấn và hỗ trợ ra quyết định. Do đầu ra của bài toán là ngành học phù hợp cho thí sinh, mô hình cần được đánh giá bằng các chỉ số của bài toán phân loại như Accuracy, Precision, Recall, F1-score, Top-1 Accuracy và Top-3 Accuracy. Trong đó, Top-3 Accuracy có ý nghĩa đặc biệt quan trọng vì phù hợp với mục tiêu thực tế của hệ thống là gợi ý danh sách ba ngành học phù hợp nhất để thí sinh và cán bộ tư vấn tham khảo.

– Khả năng xử lý dữ liệu và tính khả thi triển khai (Data Processing Capability): Tiêu chí này đánh giá năng lực của thuật toán trong việc tiếp nhận, khai thác và biểu diễn các tập dữ liệu đầu vào mang đặc tính đa chiều, gồm điểm học tập, nhóm đặc trưng RIASEC và dữ liệu có yếu tố thời gian. Mô hình được lựa chọn cần có khả năng tổng quát hóa tốt, hạn chế hiện tượng quá khớp, đồng thời bảo đảm thời gian huấn luyện, tốc độ dự đoán và chi phí tính toán phù hợp với hạ tầng công nghệ thông tin của nhà trường.

2.3. Đề xuất giải pháp triển khai hệ thống hỗ trợ tuyển sinh cho Trường Cao đẳng Hà Nội

Trên cơ sở phân tích thực trạng công tác hướng nghiệp và quy mô tuyển sinh tại Trường Cao đẳng Hà Nội, kết hợp với việc nhận diện các hạn chế cốt lõi như chưa có hệ thống tư vấn thông minh và nguồn dữ liệu lịch sử tuyển sinh chưa được khai

thác hiệu quả, nghiên cứu đề xuất giải pháp triển khai "Hệ thống hỗ trợ quản lý tuyển sinh và tư vấn chọn ngành thông minh".

Giải pháp được thiết kế tổng thể dưới dạng một hệ thống phần mềm hướng kiến trúc module, tích hợp thuật toán học máy và học sâu nhằm tối ưu hóa quy trình tiếp nhận thông tin, phân tích hành vi và ra quyết định tư vấn cho thí sinh. Để làm rõ cơ chế vận hành nội tại của giải pháp đề xuất, mô hình luồng dữ liệu tổng quát của hệ thống được cấu trúc khép kín theo mô hình toán học: Input – Xử lý (Processing) – Output.

2.4. Phân tích yêu cầu hệ thống

Trong phạm vi nghiên cứu và xây dựng hệ thống hỗ trợ tuyển sinh và định hướng ngành học tại Trường Cao đẳng Hà Nội, việc phân tích yêu cầu hệ thống đóng vai trò nền tảng quan trọng nhằm xác định rõ các chức năng cần thiết, các ràng buộc hệ thống cũng như các tiêu chí đảm bảo chất lượng trong quá trình vận hành. Kết quả phân tích là cơ sở cho việc thiết kế kiến trúc, xây dựng cơ sở dữ liệu và triển khai hệ thống trong các giai đoạn tiếp theo.

2.4.1. Yêu cầu chức năng

Yêu cầu chức năng mô tả các dịch vụ và hành vi mà hệ thống phải cung cấp nhằm đáp ứng mục tiêu nghiệp vụ. Hệ thống hỗ trợ tuyển sinh và định hướng ngành học được xây dựng với các chức năng chính sau:

2.4.1.1. Nhập thông tin thí sinh

Chức năng nhập thông tin thí sinh là chức năng đầu vào của hệ thống, đóng vai trò thu thập dữ liệu phục vụ cho quá trình xử lý và phân tích.

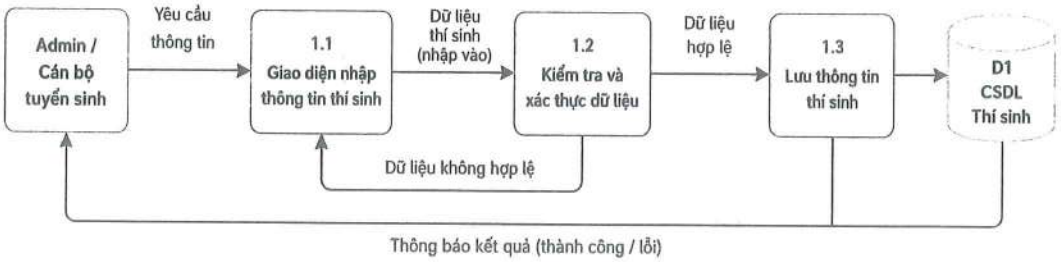
a. Mô tả chức năng

- Hệ thống cho phép người dùng nhập thông tin thí sinh bao gồm:
- Thông tin cá nhân (họ tên, ngày sinh, giới tính).
- Thông tin học tập (điểm trung bình, học lực, hạnh kiểm).
- Thông tin định hướng nghề nghiệp (sở thích, nhóm ngành quan tâm).
- Thông tin bổ trợ (kỹ năng, kết quả khảo sát).
-

b. Ràng buộc nghiệp vụ

- Dữ liệu phải đầy đủ theo trường bắt buộc
- Định dạng dữ liệu phải hợp lệ
- Không cho phép trùng lặp thí sinh
- Dữ liệu phải được kiểm tra trước khi lưu

c. Luồng xử lý chính



Hình 2.3. Sơ đồ luồng dữ liệu chức năng nhập thông tin thí sinh

2.4.1.2. Dự đoán ngành phù hợp

Chức năng dự đoán ngành phù hợp là chức năng lõi của hệ thống, thực hiện phân tích dữ liệu thí sinh và đưa ra gợi ý ngành học tối ưu.

a. Mô tả chức năng

Hệ thống sử dụng dữ liệu đầu vào để:

- Phân tích đặc điểm thí sinh
- So sánh với tiêu chí từng ngành đào tạo
- Tính toán mức độ phù hợp

b. Quy trình xử lý

Quy trình dự đoán bao gồm:

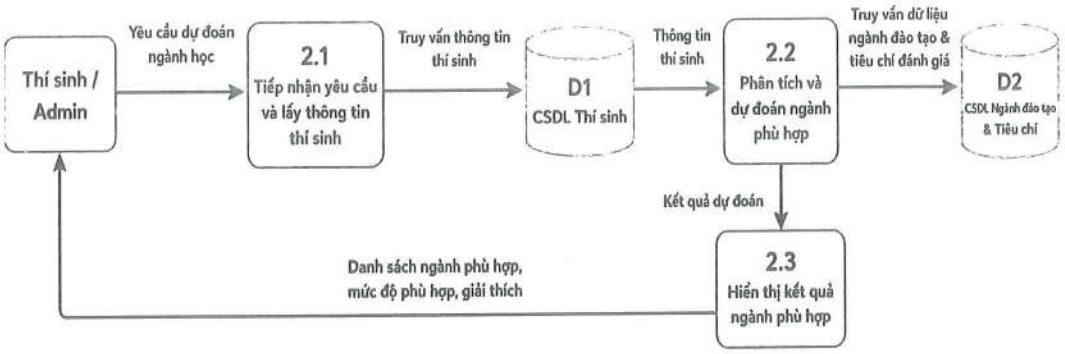
- Thu thập dữ liệu thí sinh
- Tiền xử lý dữ liệu
- Áp dụng thuật toán/mô hình dự đoán
- Xếp hạng ngành học theo mức độ phù hợp

c. Kết quả đầu ra

- Danh sách ngành học phù hợp

- Mức độ tương thích (% hoặc thang điểm)
- Gợi ý có giải thích logic

d. Luồng xử lý chính



Hình 2.4. Sơ đồ luồng dữ liệu chức năng dự đoán ngành phù hợp

2.4.1.3. Quản lý dữ liệu tuyển sinh

Chức năng quản lý dữ liệu tuyển sinh hỗ trợ cán bộ nhà trường trong việc theo dõi, cập nhật và khai thác thông tin thí sinh.

a. Mô tả chức năng

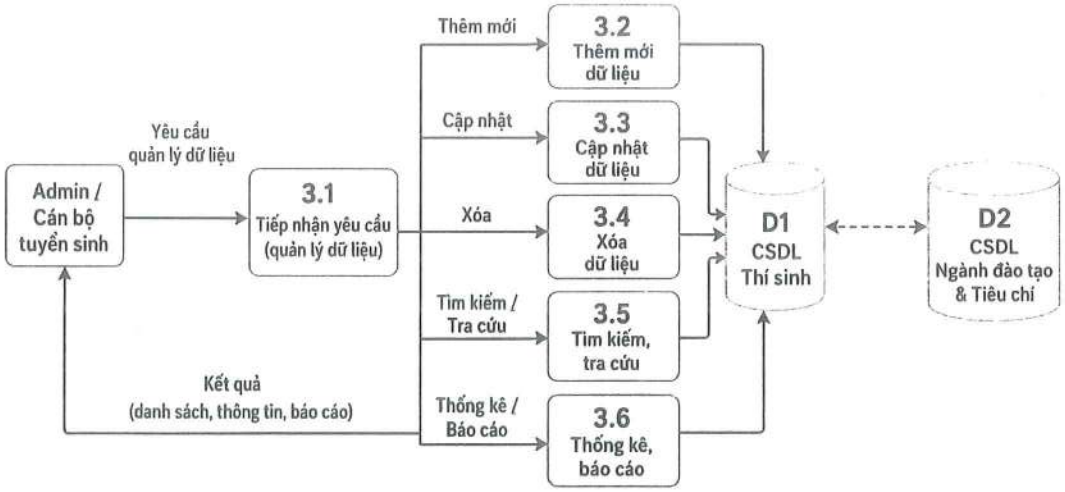
Hệ thống cung cấp các thao tác quản lý dữ liệu gồm:

- Thêm mới thí sinh
- Tìm kiếm và tra cứu thông tin
- Cập nhật dữ liệu
- Xóa dữ liệu không hợp lệ
- Thống kê và báo cáo

b. Yêu cầu nghiệp vụ

- Dữ liệu phải đảm bảo tính toàn vẹn
- Có cơ chế kiểm soát truy cập
- Hỗ trợ tìm kiếm đa tiêu chí
- Có khả năng tổng hợp báo cáo

c. Luồng xử lý chính



Hình 2.5. Sơ đồ luồng dữ liệu chức năng quản lý dữ liệu tuyển sinh

2.4.2. Yêu cầu phi chức năng

Bên cạnh các yêu cầu chức năng, yêu cầu phi chức năng đóng vai trò quan trọng trong việc đảm bảo hệ thống hỗ trợ tuyển sinh vận hành ổn định, an toàn và hiệu quả trong môi trường thực tế. Các yêu cầu phi chức năng không trực tiếp tạo ra chức năng nghiệp vụ, nhưng quyết định chất lượng, độ tin cậy và khả năng mở rộng của hệ thống.

2.4.2.1. Yêu cầu về hiệu năng (Performance)

Hệ thống hỗ trợ tuyển sinh Trường Cao đẳng Hà Nội phải đảm bảo hiệu năng ổn định, đáp ứng tốt nhu cầu sử dụng trong các giai đoạn cao điểm tuyển sinh. Cụ thể:

– Thời gian phản hồi cho các chức năng nhập, tra cứu và cập nhật thông tin thí sinh không vượt quá 2 giây trong điều kiện tải bình thường.

– Chức năng dự đoán ngành học phù hợp phải cho kết quả trong 3 giây, tối đa 5 giây khi hệ thống chịu tải cao.

– Hệ thống phải hỗ trợ tối thiểu 300 người dùng đồng thời mà không xảy ra treo hệ thống hoặc mất dữ liệu.

– Các truy vấn thống kê và báo cáo tuyển sinh phải hoàn thành trong thời gian chấp nhận được, đảm bảo không ảnh hưởng đến các chức năng chính.

– Hiệu năng hệ thống phải được duy trì ổn định khi quy mô dữ liệu tuyển sinh tăng lên theo từng năm.

2.4.2.2. Yêu cầu về bảo mật (Security)

Dữ liệu tuyển sinh chứa nhiều thông tin cá nhân nhạy cảm của thí sinh, do đó hệ thống cần đáp ứng các yêu cầu bảo mật nghiêm ngặt:

– Hệ thống phải có cơ chế xác thực người dùng (tài khoản, mật khẩu) và phân quyền rõ ràng theo vai trò: thí sinh, cán bộ tuyển sinh, quản trị hệ thống.

– Thông tin cá nhân của thí sinh phải được bảo vệ khỏi truy cập trái phép, đảm bảo tính riêng tư và tuân thủ các quy định về bảo vệ dữ liệu.

– Dữ liệu trao đổi giữa người dùng và hệ thống cần được bảo mật, tránh nguy cơ bị đánh cắp hoặc sửa đổi.

– Hệ thống phải ghi nhận lịch sử thao tác (log) để phục vụ kiểm tra, giám sát và truy vết khi xảy ra sự cố.

Yêu cầu bảo mật giúp đảm bảo an toàn thông tin, nâng cao mức độ tin cậy của hệ thống đối với người dùng và nhà trường.

2.4.2.3. Yêu cầu về độ tin cậy và tính sẵn sàng (Reliability & Availability)

Hệ thống cần đảm bảo hoạt động ổn định và liên tục trong suốt thời gian tuyển sinh:

– Hệ thống phải hạn chế tối đa lỗi trong quá trình vận hành.

– Có cơ chế sao lưu dữ liệu định kỳ nhằm phòng ngừa mất mát dữ liệu.

– Trong trường hợp xảy ra sự cố, hệ thống cần có khả năng khôi phục dữ liệu và tiếp tục hoạt động trong thời gian ngắn nhất.

– Các lỗi phát sinh phải được thông báo rõ ràng, không gây mất dữ liệu hoặc gián đoạn nghiêm trọng.

Độ tin cậy cao giúp hệ thống trở thành công cụ hỗ trợ đắc lực cho công tác tuyển sinh của nhà trường.

2.4.2.4. Yêu cầu về khả năng mở rộng (Scalability)

Hệ thống hỗ trợ tuyển sinh cần được thiết kế với khả năng mở rộng trong tương lai:

– Cho phép mở rộng số lượng thí sinh, ngành đào tạo và dữ liệu tuyển sinh qua các năm mà không ảnh hưởng đến hiệu năng.

– Dễ dàng tích hợp thêm các mô-đun mới như tư vấn trực tuyến, phân tích xu hướng tuyển sinh hoặc kết nối với các hệ thống quản lý đào tạo khác.

– Cấu trúc hệ thống và cơ sở dữ liệu cần linh hoạt, thuận lợi cho việc nâng cấp và mở rộng.

– Khả năng mở rộng đảm bảo hệ thống đáp ứng được nhu cầu phát triển lâu dài của nhà trường.

2.4.2.5. Yêu cầu về tính dễ sử dụng (Usability)

Hệ thống cần thân thiện với người dùng, phù hợp với nhiều đối tượng sử dụng:

– Giao diện rõ ràng, dễ hiểu đối với cả thí sinh và cán bộ tuyển sinh.

– Các chức năng được bố trí hợp lý, giảm thiểu thao tác phức tạp.

– Có hướng dẫn sử dụng hoặc thông báo hỗ trợ trong quá trình thao tác.

Tính dễ sử dụng giúp nâng cao hiệu quả khai thác hệ thống và giảm chi phí đào tạo người dùng.

2.4.2.6. Yêu cầu về khả năng bảo trì và vận hành (Maintainability)

Hệ thống cần thuận lợi cho việc vận hành và bảo trì:

– Cấu trúc phần mềm rõ ràng, dễ kiểm tra và sửa lỗi.

– Cho phép cập nhật dữ liệu tuyển sinh và điều chỉnh quy định tuyển sinh theo từng năm.

– Dễ dàng sao lưu, phục hồi và nâng cấp hệ thống khi cần thiết.

Yêu cầu này giúp hệ thống duy trì ổn định trong quá trình sử dụng lâu dài.

2.5. Phân tích thiết kế cơ sở dữ liệu

Trong hệ thống hỗ trợ tuyển sinh, cơ sở dữ liệu không chỉ đóng vai trò lưu trữ thông tin hành chính mà còn là nền tảng cho các hoạt động phân tích, khai phá dữ liệu và hỗ trợ ra quyết định. Do đó, việc phân tích thiết kế cơ sở dữ liệu cần đảm bảo đồng thời ba mục tiêu: đáp ứng nghiệp vụ tuyển sinh, đảm bảo chất lượng dữ liệu và phù hợp cho các mô hình dự đoán.

2.5.1. Nguồn dữ liệu

Nguồn dữ liệu của hệ thống được hình thành trong suốt quá trình tuyển sinh và có đặc điểm đa dạng, không đồng nhất và thay đổi theo thời gian. Việc phân tích rõ nguồn dữ liệu giúp định hướng đúng cấu trúc cơ sở dữ liệu và quy trình xử lý dữ liệu.

Trong phạm vi đề án, dữ liệu phục vụ thực nghiệm được xây dựng dựa trên cấu trúc dữ liệu tuyển sinh của Trường Cao đẳng Hà Nội, bao gồm thông tin điểm học tập, kết quả định hướng nghề nghiệp RIASEC và ngành học tương ứng. Để bảo đảm yêu cầu bảo mật thông tin cá nhân của thí sinh, dữ liệu sử dụng trong thực nghiệm được ẩn danh, chuẩn hóa và mô phỏng có kiểm soát theo cấu trúc nghiệp vụ tuyển sinh của Nhà trường. Phần mô tả chi tiết về quy mô, cấu trúc đặc trưng, nhãn dữ liệu và cách phân chia tập huấn luyện – kiểm định được trình bày tại mục 3.1.1 của Chương 3.

a) Dữ liệu tuyển sinh

Dữ liệu tuyển sinh là nhóm dữ liệu mang tính tổng hợp và lịch sử, phản ánh chiến lược đào tạo và kết quả tuyển sinh của nhà trường qua từng năm. Nhóm dữ liệu này có các đặc điểm sau:

- ***Nội dung dữ liệu***

  + Thông tin ngành đào tạo: mã ngành, tên ngành, khoa quản lý, thời gian đào tạo, chỉ tiêu theo từng năm.

  + Thông tin đợt tuyển sinh: năm tuyển sinh, đợt xét tuyển, thời gian bắt đầu – kết thúc, hình thức xét tuyển.

  + Kết quả tuyển sinh lịch sử: điểm chuẩn từng ngành, số lượng hồ sơ đăng ký, số lượng trúng tuyển, tỷ lệ nhập học.

  + Các quy định tuyển sinh: điều kiện xét tuyển, ngưỡng đảm bảo chất lượng, ưu tiên khu vực/đối tượng.

- ***Đặc điểm dữ liệu***

  + Dữ liệu mang tính chu kỳ theo năm học.

  + Có mối quan hệ chặt chẽ với dữ liệu thí sinh thông qua ngành đăng ký và kết quả xét tuyển.

+ Là dữ liệu đầu vào quan trọng cho các mô hình phân tích xu hướng và dự đoán ngành phù hợp.

– ***Vai trò trong hệ thống***

+ Làm cơ sở xây dựng các bảng tham chiếu (reference tables) trong cơ sở dữ liệu.

+ Cung cấp dữ liệu huấn luyện cho mô hình dự đoán khả năng trúng tuyển.

+ Hỗ trợ công tác thống kê, báo cáo và ra quyết định của lãnh đạo nhà trường.

b) Dữ liệu thí sinh

Dữ liệu thí sinh là nhóm dữ liệu chi tiết, mang tính cá nhân, phát sinh liên tục trong quá trình tuyển sinh. Đây là nhóm dữ liệu có khối lượng lớn và ảnh hưởng trực tiếp đến chất lượng kết quả dự đoán.

– ***Nội dung dữ liệu***

+ Thông tin cá nhân: họ tên, ngày sinh, giới tính, khu vực tuyển sinh, thông tin liên hệ.

+ Thông tin học tập: trường THPT, năm tốt nghiệp, điểm trung bình, điểm từng môn trong tổ hợp xét tuyển.

+ Thông tin nguyện vọng: danh sách ngành đăng ký, thứ tự ưu tiên, hình thức đào tạo.

+ Trạng thái xử lý hồ sơ: hợp lệ, trúng tuyển, nhập học, hủy hồ sơ.

– ***Đặc điểm dữ liệu***

+ Dữ liệu có thể thiếu, sai lệch hoặc không đồng nhất do nhập từ nhiều nguồn.

+ Có sự thay đổi trạng thái theo từng giai đoạn tuyển sinh.

+ Có mối quan hệ nhiều-nhiều giữa thí sinh và ngành đào tạo.

– ***Vai trò trong hệ thống***

+ Là dữ liệu đầu vào trực tiếp cho chức năng dự đoán ngành phù hợp.

+ Phục vụ quản lý hồ sơ, theo dõi kết quả tuyển sinh.

+ Là cơ sở để phân tích hành vi và xu hướng lựa chọn ngành của thí sinh.

2.5.2. Tiền xử lý dữ liệu

Do đặc thù dữ liệu tuyển sinh được thu thập qua nhiều năm, từ nhiều nguồn khác nhau và chịu ảnh hưởng của yếu tố nhập liệu thủ công, dữ liệu thường tồn tại nhiều sai lệch và không đồng nhất. Vì vậy, tiền xử lý dữ liệu là bước bắt buộc nhằm đảm bảo tính chính xác, tính nhất quán và khả năng khai thác hiệu quả của dữ liệu trong hệ thống hỗ trợ tuyển sinh. Quá trình tiền xử lý không chỉ góp phần nâng cao chất lượng dữ liệu đầu vào mà còn đóng vai trò nền tảng cho việc tổ chức dữ liệu lưu trữ và xây dựng các mô hình phân tích, dự đoán trong hệ thống.

a) Làm sạch dữ liệu

Làm sạch dữ liệu tập trung xử lý các vấn đề phát sinh trong quá trình thu thập dữ liệu:

+ Phát hiện và loại bỏ các hồ sơ thí sinh trùng lặp dựa trên các trường định danh như số căn cước công dân (CCCD), số điện thoại hoặc địa chỉ thư điện tử.
+ Xử lý các trường dữ liệu bị thiếu:
    - Bổ sung dữ liệu từ nguồn khác nếu có thể đối chiếu và xác thực.
    - Loại bỏ các bản ghi không đủ thông tin quan trọng, không đáp ứng yêu cầu cho phân tích và dự đoán.
+ Kiểm tra và hiệu chỉnh dữ liệu bất thường:
    - Điểm học tập vượt ngoài thang điểm cho phép.
    - Ngày tháng không hợp lệ.
+ Chuẩn hóa quy ước nhập liệu giữa các năm tuyển sinh khác nhau nhằm đảm bảo tính thống nhất của dữ liệu lịch sử.

Thông qua bước làm sạch dữ liệu, các yếu tố gây nhiễu được loại bỏ, độ tin cậy của dữ liệu được nâng cao, đồng thời hạn chế nguy cơ sai lệch trong quá trình phân tích, dự đoán và lưu trữ dữ liệu trong cơ sở dữ liệu của hệ thống.

b) Chuẩn hóa dữ liệu

Chuẩn hóa dữ liệu nhằm đảm bảo tính thống nhất về mặt biểu diễn trong toàn bộ hệ thống:

+ Chuẩn hóa thang điểm học tập của thí sinh về cùng một thang chuẩn, tạo điều kiện thuận lợi cho việc so sánh và phân tích.

+ Chuẩn hóa dữ liệu văn bản:

  o Thống nhất cách ghi tên trường, tên ngành đào tạo và khu vực tuyển sinh.

+ Chuẩn hóa dữ liệu phân loại:

  o Mã hóa giới tính, khu vực, hình thức đào tạo thành các giá trị chuẩn.

+ Đảm bảo việc xác định và sử dụng khóa chính, khóa ngoại được thực hiện nhất quán, phản ánh đúng mối quan hệ giữa các nhóm dữ liệu.

Chuẩn hóa dữ liệu giúp đơn giản hóa thiết kế cơ sở dữ liệu và tăng hiệu quả xử lý của các thuật toán dự đoán.

c) Chuyển đổi dữ liệu

Chuyển đổi dữ liệu nhằm biến dữ liệu thô sau khi được làm sạch và chuẩn hóa thành dữ liệu có giá trị phân tích, sẵn sàng cho việc lưu trữ và khai thác trong hệ thống hỗ trợ tuyển sinh. Các hoạt động chuyển đổi dữ liệu bao gồm:

+ Chuyển đổi dữ liệu định tính sang định lượng để phù hợp với mô hình dự đoán.

+ Tạo thuộc tính dẫn xuất (feature engineering):

  o Tính toán điểm trung bình tổng hợp theo các tổ hợp môn xét tuyển.

  o Xây dựng chỉ số phản ánh mức độ phù hợp giữa năng lực học tập của thí sinh và yêu cầu của từng ngành đào tạo.

+ Tổng hợp dữ liệu theo nhiều chiều khác nhau như theo ngành đào tạo, theo năm tuyển sinh và theo khu vực tuyển sinh nhằm phục vụ cho công tác thống kê và phân tích

+ Chuẩn bị dữ liệu cho việc lưu trữ có cấu trúc và huấn luyện mô hình dự đoán ngành học phù hợp.

Bước chuyển đổi dữ liệu giúp tối đa hóa giá trị thông tin của dữ liệu tuyển sinh, đồng thời góp phần nâng cao độ chính xác và hiệu quả của hệ thống hỗ trợ ra quyết định trong công tác tuyển sinh.

2.5.3. Tổ chức và chuẩn hóa dữ liệu phục vụ hệ thống hỗ trợ tuyển sinh

Trên cơ sở dữ liệu đã được thu thập và tiền xử lý ở mục 2.5.2, việc tổ chức và chuẩn hóa dữ liệu là bước cần thiết nhằm đảm bảo dữ liệu được lưu trữ một cách khoa học, nhất quán và có khả năng khai thác lâu dài cho cả hoạt động nghiệp vụ và phân tích. Nội dung này tập trung vào cách sắp xếp, phân tách và chuẩn hóa dữ liệu theo các thực thể thông tin độc lập, phù hợp với đặc thù tuyển sinh của Trường Cao đẳng Hà Nội.

2.5.3.1. Tổ chức dữ liệu theo các thực thể thông tin

Dữ liệu tuyển sinh được tổ chức xoay quanh các nhóm thông tin chính phản ánh đầy đủ các đối tượng và mối quan hệ trong quá trình tuyển sinh. Mỗi nhóm thông tin được xem như một thực thể dữ liệu độc lập, bao gồm: thông tin thí sinh, thông tin ngành đào tạo, thông tin đăng ký xét tuyển, kết quả tuyển sinh và các dữ liệu danh mục dùng chung.

Việc phân tách dữ liệu theo các thực thể giúp hạn chế lặp lại thông tin, tăng khả năng quản lý dữ liệu theo từng đối tượng cụ thể và tạo điều kiện thuận lợi cho việc mở rộng hệ thống trong các năm tuyển sinh tiếp theo. Đồng thời, cách tổ chức này cũng hỗ trợ hiệu quả cho việc tổng hợp, thống kê và so sánh dữ liệu giữa các kỳ tuyển sinh khác nhau.

2.5.3.2. Xác định mối quan hệ và ràng buộc dữ liệu

Giữa các thực thể dữ liệu tồn tại những mối quan hệ logic phản ánh đúng nghiệp vụ tuyển sinh, chẳng hạn như mối quan hệ giữa thí sinh và hồ sơ đăng ký, giữa hồ sơ và ngành đào tạo, hoặc giữa kết quả xét tuyển và đợt tuyển sinh. Các mối quan hệ này được xác định rõ ràng nhằm đảm bảo dữ liệu được liên kết chặt chẽ, tránh mâu thuẫn và sai lệch trong quá trình khai thác.

Bên cạnh đó, các ràng buộc dữ liệu được đặt ra nhằm đảm bảo tính toàn vẹn thông tin, giúp kiểm soát tốt hơn các thao tác cập nhật, chỉnh sửa và truy xuất dữ liệu trong hệ thống.

2.5.3.3. Chuẩn hóa dữ liệu nhằm giảm dư thừa và tăng tính nhất quán

Dữ liệu sau khi được tổ chức theo các thực thể tiếp tục được chuẩn hóa nhằm loại bỏ sự dư thừa và phụ thuộc không cần thiết giữa các thuộc tính. Quá trình chuẩn hóa được thực hiện từng bước, đảm bảo mỗi nhóm dữ liệu chỉ lưu trữ các thông tin có quan hệ trực tiếp và phụ thuộc đầy đủ vào định danh của thực thể đó.

Việc chuẩn hóa đến mức hợp lý giúp:

- Tránh trùng lặp dữ liệu giữa các nhóm thông tin;
- Dễ dàng cập nhật và bảo trì dữ liệu khi có thay đổi;
- Đảm bảo tính nhất quán của thông tin trên toàn hệ thống.

Cách tiếp cận này không chỉ góp phần nâng cao chất lượng dữ liệu mà còn tạo nền tảng vững chắc cho việc khai thác dữ liệu phục vụ thống kê, báo cáo và các hoạt động phân tích chuyên sâu.

Giai đoạn tổ chức dữ liệu	Mô tả dữ liệu ban đầu / sau điều chỉnh	Nhận xét về ý nghĩa
Dữ liệu ban đầu (chưa tinh gọn)	Hồ sơ thí sinh lưu chung: mã thí sinh, họ tên, CCCD, giới tính, tên trường THPT, khu vực, tên ngành đăng ký, tên khoa, chỉ tiêu ngành, điểm xét tuyển	Thông tin lặp lại nhiều lần giữa các hồ sơ, khó cập nhật khi có thay đổi về ngành hoặc trường
Tách nhóm thông tin thí sinh	Thông tin cá nhân thí sinh được quản lý riêng: mã thí sinh, họ tên, CCCD, giới tính, ngày sinh, khu vực	Mỗi thí sinh chỉ lưu một lần, tránh trùng lặp và sai lệch thông tin cá nhân
Tách nhóm thông tin ngành đào tạo	Thông tin ngành được quản lý độc lập: mã ngành, tên ngành, khoa phụ trách, chỉ tiêu tuyển sinh	Dễ dàng cập nhật chỉ tiêu hoặc tên ngành mà không ảnh hưởng đến hồ sơ thí sinh
Tách thông tin đăng ký xét tuyển	Quan hệ giữa thí sinh và ngành được lưu riêng: mã thí sinh, mã ngành, đợt tuyển sinh, hình thức xét tuyển	Phản ánh đúng nghiệp vụ: một thí sinh có thể đăng ký nhiều ngành

Loại bỏ phụ thuộc gián tiếp	Không lưu tên khoa trong hồ sơ thí sinh, chỉ liên kết thông qua thông tin ngành	Giảm dư thừa, đảm bảo mỗi thông tin chỉ phụ thuộc trực tiếp vào đối tượng quản lý
Cấu trúc dữ liệu hoàn thiện	Các nhóm dữ liệu được liên kết thông qua định danh, không trùng lặp thuộc tính mô tả	Dữ liệu nhất quán, dễ bảo trì, sẵn sàng cho phân tích và dự báo

Bảng 2.1. Minh họa quá trình tổ chức và tinh gọn dữ liệu tuyển sinh

2.5.3.4. Ý nghĩa của việc tổ chức và chuẩn hóa dữ liệu

Việc tổ chức và chuẩn hóa dữ liệu đóng vai trò trung tâm trong việc đảm bảo hệ thống hỗ trợ tuyển sinh hoạt động ổn định và hiệu quả. Dữ liệu được quản lý khoa học giúp giảm thiểu sai sót trong xử lý nghiệp vụ, đồng thời nâng cao độ tin cậy của các thông tin đầu vào cho các mô hình phân tích và dự báo ở các chương tiếp theo.

2.5.4. Liên kết dữ liệu với mô hình dự đoán tuyển sinh

Sau khi dữ liệu tuyển sinh đã được tổ chức và tinh gọn theo hướng nhất quán ở mục 2.5.3, bước tiếp theo là liên kết dữ liệu này với mô hình dự đoán nhằm hỗ trợ ra quyết định trong công tác tuyển sinh. Nội dung của mục này tập trung vào cách thức chuẩn bị, lựa chọn và ánh xạ dữ liệu từ hệ thống quản lý sang môi trường phân tích và dự báo.

2.5.4.1. Xác định tập dữ liệu đầu vào cho mô hình dự đoán

Dữ liệu được lựa chọn cho mô hình dự đoán chủ yếu là dữ liệu lịch sử tuyển sinh của nhà trường qua các năm, bao gồm thông tin đặc trưng của thí sinh, ngành đăng ký, hình thức xét tuyển và kết quả tuyển sinh thực tế. Các dữ liệu này phản ánh mối quan hệ giữa năng lực học tập của thí sinh và khả năng phù hợp với từng ngành đào tạo.

Việc lựa chọn dữ liệu đầu vào được thực hiện có chọn lọc, ưu tiên các thuộc tính có ý nghĩa phân biệt cao và có tác động trực tiếp đến kết quả dự đoán, đồng thời loại bỏ các thông tin không cần thiết hoặc không ổn định theo thời gian.

2.5.4.2. Ánh xạ dữ liệu từ hệ thống quản lý sang không gian phân tích

Dữ liệu trong hệ thống quản lý tuyển sinh được ánh xạ sang các tập dữ liệu phân tích thông qua các định danh thống nhất giữa các nhóm thông tin. Quá trình ánh xạ này đảm bảo rằng mỗi bản ghi dữ liệu phân tích đều có thể truy vết ngược lại dữ liệu gốc trong hệ thống khi cần thiết.

Cách tiếp cận này giúp duy trì tính liên tục và nhất quán giữa dữ liệu nghiệp vụ và dữ liệu phục vụ mô hình, đồng thời tạo điều kiện thuận lợi cho việc cập nhật dữ liệu mới theo từng đợt tuyển sinh mà không làm gián đoạn hoạt động của mô hình dự đoán.

2.5.4.3. Chuẩn bị dữ liệu cho huấn luyện và kiểm thử mô hình

Tập dữ liệu sau khi được ánh xạ tiếp tục được chia thành các tập con phục vụ cho huấn luyện và đánh giá mô hình dự đoán. Việc phân chia này dựa trên dữ liệu theo thời gian hoặc theo tỷ lệ phù hợp, nhằm đảm bảo mô hình có khả năng tổng quát hóa tốt và hạn chế hiện tượng học lệch.

Bên cạnh đó, các thuộc tính đầu vào được kiểm tra lại về mức độ phân bố và khả năng phản ánh đúng đặc điểm của dữ liệu tuyển sinh thực tế, từ đó nâng cao độ tin cậy của kết quả dự đoán.

2.5.4.4. Vai trò của liên kết dữ liệu trong hệ thống hỗ trợ ra quyết định

Việc liên kết chặt chẽ giữa dữ liệu tuyển sinh và mô hình dự đoán giúp hệ thống không chỉ dừng lại ở việc quản lý thông tin mà còn hỗ trợ hiệu quả cho công tác phân tích và ra quyết định. Kết quả dự đoán được tạo ra trên cơ sở dữ liệu lịch sử đáng tin cậy, góp phần cung cấp thông tin tham khảo cho cán bộ tuyển sinh trong việc tư vấn ngành học phù hợp và hoạch định chiến lược tuyển sinh trong tương lai.

2.6. Kết luận chương 2

Chương 2 đã trình bày quá trình phân tích yêu cầu và tổ chức dữ liệu cho hệ thống hỗ trợ tuyển sinh Trường Cao đẳng Hà Nội. Nội dung chương tập trung làm rõ các yêu cầu chức năng và phi chức năng của hệ thống, đồng thời phân tích đặc thù dữ liệu tuyển sinh và vai trò của dữ liệu trong việc hỗ trợ ra quyết định. Trên cơ sở đó, các vấn đề liên quan đến tiền xử lý dữ liệu, tổ chức dữ liệu, hoàn thiện cấu trúc dữ

liệu theo nguyên tắc chuẩn hóa và liên kết dữ liệu với mô hình dự đoán đã được phân tích một cách hệ thống.

Kết quả của chương này là việc xác định được tập dữ liệu đầu vào có chất lượng, được tổ chức hợp lý và sẵn sàng cho các hoạt động phân tích, dự đoán. Đây là nền tảng quan trọng để đảm bảo tính chính xác, ổn định và hiệu quả của hệ thống hỗ trợ tuyển sinh, đồng thời tạo điều kiện thuận lợi cho việc áp dụng các mô hình dự đoán trong các bước tiếp theo của nghiên cứu.

Trên cơ sở các kết quả phân tích và tổ chức dữ liệu đã trình bày ở Chương 2, Chương 3 sẽ tập trung vào việc xây dựng và đánh giá mô hình dự đoán phục vụ hệ thống hỗ trợ tuyển sinh. Nội dung chương này trình bày lựa chọn mô hình dự đoán phù hợp, quy trình huấn luyện và kiểm thử mô hình trên dữ liệu tuyển sinh thực tế, cũng như đánh giá hiệu quả của mô hình thông qua các chỉ tiêu cụ thể. Qua đó, Chương 3 nhằm chứng minh khả năng ứng dụng của mô hình dự đoán trong việc hỗ trợ ra quyết định và nâng cao chất lượng công tác tuyển sinh tại Trường Cao đẳng Hà Nội.

CHƯƠNG 3: THỬ NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

3.1. Kịch bản thử nghiệm

Trên cơ sở tập dữ liệu tuyển sinh đã được làm sạch, chuẩn hóa và tổ chức theo cấu trúc trình bày tại Chương 2, Chương 3 triển khai thực nghiệm nhằm huấn luyện, kiểm thử và đánh giá ba mô hình dự đoán ngành học đã được lựa chọn tại mục 2.2. Ba mô hình bao gồm: Logistic Regression đa lớp đại diện cho hướng tiếp cận tuyến tính, Random Forest Classifier đại diện cho hướng tiếp cận phi tuyến và LSTM_Net đại diện cho hướng tiếp cận học sâu xử lý dữ liệu chuỗi thời gian.

Việc triển khai đồng thời ba mô hình cho phép đánh giá toàn diện khả năng giải quyết bài toán tư vấn tuyển sinh đã đặt ra tại mục 1.3.1. Về bản chất, bài toán nghiên cứu là bài toán phân loại đa lớp, trong đó mỗi lớp tương ứng với một ngành đào tạo của Nhà trường; đầu ra ứng dụng hướng tới là danh sách Top-3 ngành học phù hợp nhất cho mỗi thí sinh.

Để tránh trùng lặp với Chương 2, các công thức và nguyên lý lý thuyết của từng mô hình không được trình bày lại chi tiết trong Chương 3. Chương này tập trung vào cách tổ chức dữ liệu đầu vào, cấu hình thực nghiệm, quá trình huấn luyện - kiểm thử, kết quả đánh giá riêng của từng mô hình, sau đó tổng hợp so sánh và rút ra nhận xét chung.

3.1.1. Nội dung thử nghiệm

Ba mô hình được thử nghiệm trên cùng một tập dữ liệu đầu vào và cùng quy trình tiền xử lý đã trình bày tại mục 2.5. Việc thống nhất tập dữ liệu, hệ đặc trưng và cách chia tập huấn luyện - kiểm định giúp bảo đảm tính khách quan khi so sánh kết quả giữa các mô hình.

Dữ liệu và đặc trưng sử dụng trong thử nghiệm

Bộ dữ liệu thực nghiệm được xây dựng nhằm phục vụ việc huấn luyện, kiểm định và so sánh hiệu quả của các mô hình Logistic Regression, Random Forest và LSTM_Net trong bài toán hỗ trợ tư vấn tuyển sinh tại Trường Cao đẳng Hà Nội. Do dữ liệu tuyển sinh thực tế có chứa thông tin cá nhân của thí sinh, trong phạm vi đề án, bộ dữ liệu sử dụng cho thực nghiệm được tổ chức theo hướng ẩn danh, chuẩn hóa

và mô phỏng có kiểm soát dựa trên cấu trúc dữ liệu tuyển sinh của Nhà trường. Cách xây dựng này nhằm bảo đảm phù hợp với mục tiêu nghiên cứu, đồng thời hạn chế việc sử dụng trực tiếp các thông tin cá nhân nhạy cảm của thí sinh.

Bộ dữ liệu gồm 150.000 mẫu, mỗi mẫu tương ứng với một hồ sơ thí sinh. Mỗi hồ sơ được biểu diễn bởi các nhóm thông tin chính gồm: nhóm đặc trưng học tập, nhóm đặc trưng định hướng nghề nghiệp RIASEC và nhãn ngành học phù hợp. Nhóm đặc trưng học tập gồm điểm của chín môn học: Toán, Ngữ văn, Tiếng Anh, Vật lý, Hóa học, Sinh học, Lịch sử, Địa lý và Giáo dục công dân. Các giá trị điểm được quy đổi về thang điểm từ 0 đến 10. Nhóm đặc trưng RIASEC gồm sáu chỉ số: Realistic, Investigative, Artistic, Social, Enterprising và Conventional, phản ánh xu hướng sở thích, năng lực cá nhân và định hướng nghề nghiệp của thí sinh.

Nhãn đầu ra của mỗi mẫu là ngành học phù hợp thuộc tập 12 ngành đào tạo được xem xét trong phạm vi đề án. Trong thực nghiệm, nhãn này được sử dụng như một nhãn đại diện cho ngành học mà thí sinh có khả năng phù hợp, phục vụ cho bài toán phân loại nhiều lớp. Kết quả dự đoán của mô hình không chỉ là một ngành có xác suất cao nhất mà còn là danh sách Top-3 ngành học phù hợp, giúp cán bộ tuyển sinh có thêm cơ sở tư vấn cho thí sinh.

Bộ dữ liệu được chia thành hai phần theo tỷ lệ 80/20. Trong đó, 120.000 mẫu được sử dụng để huấn luyện mô hình và 30.000 mẫu được sử dụng để kiểm định, đánh giá kết quả. Tập kiểm định được tổ chức tương đối cân bằng giữa 12 ngành đào tạo, trung bình khoảng 2.500 mẫu cho mỗi ngành. Việc chia dữ liệu theo tỷ lệ này giúp bảo đảm mô hình được học trên tập dữ liệu đủ lớn, đồng thời vẫn có tập kiểm định độc lập để đánh giá khách quan khả năng dự đoán trên dữ liệu chưa được sử dụng trong quá trình huấn luyện.

STT	Tên mô hình	Loại	Thuật toán cơ sở	Phần mềm
1	**Mô hình Logistic Regression đa lớp**	Baseline tuyến tính	Logistic Regression (Multinomial – lbfgs)	scikit-learn
2	**Mô hình Random Forest Classifier**	Baseline phi tuyến	Random Forest Classifier (100 cây)	scikit-learn
3	**Mô hình LSTM_Net**	Học sâu	Vanilla LSTM 2 lớp (PyTorch)	PyTorch

Bảng 3.1. Tổng quan các mô hình thử nghiệm và phần mềm sử dụng

STT	Ký hiệu	Tên đặc trưng	Đơn vị	Mô tả
1	x_1	Toán	Điểm (0–10)	Điểm môn Toán trung bình 3 năm
2	x_2	Văn	Điểm (0–10)	Điểm môn Ngữ văn
3	x_3	Anh	Điểm (0–10)	Điểm môn Tiếng Anh
4	x_4	Lý	Điểm (0–10)	Điểm môn Vật lý
5	x_5	Hoá	Điểm (0–10)	Điểm môn Hoá học
6	x_6	Sinh	Điểm (0–10)	Điểm môn Sinh học
7	x_7	Sử	Điểm (0–10)	Điểm môn Lịch sử
8	x_8	Địa	Điểm (0–10)	Điểm môn Địa lý
9	x_9	GDCD	Điểm (0–10)	Điểm môn GDCD
10	x_{10}	RIASEC-R	Điểm (0–10)	Nhóm Thực tế (Realistic)
11	x_{11}	RIASEC-I	Điểm (0–10)	Nhóm Nghiên cứu (Investigative)
12	x_{12}	RIASEC-A	Điểm (0–10)	Nhóm Nghệ thuật (Artistic)
13	x_{13}	RIASEC-S	Điểm (0–10)	Nhóm Xã hội (Social)
14	x_{14}	RIASEC-E	Điểm (0–10)	Nhóm Doanh nhân (Enterprising)
15	x_{15}	RIASEC-C	Điểm (0–10)	Nhóm Quy ước (Conventional)

Bảng 3.2. Bảng mô tả đặc trưng đầu vào

Đối với Logistic Regression đa lớp và Random Forest Classifier, mỗi học sinh được biểu diễn bằng một vector đặc trưng một chiều có 15 thành phần. Toàn bộ tập dữ liệu tạo thành ma trận đầu vào kích thước N × 15, trong đó N là số lượng học sinh. Đối với LSTM_Net, các đặc trưng được tổ chức lại theo dạng chuỗi thời gian để mô hình có thể học xu hướng thay đổi kết quả học tập qua các năm học.

$$X_ML = [x_1, x_2, \ldots, x_{15}] \in \mathbb{R}^{(N\times 15)}$$

$$X_LSTM \in \mathbb{R}^{(N\times T\times F)}$$

Trong đó, X_ML là ma trận đầu vào của các mô hình học máy truyền thống; X_LSTM là tensor đầu vào của mô hình LSTM_Net; N là số lượng học sinh; T là số bước thời gian; F là số đặc trưng tại mỗi bước thời gian.

Mô hình	**Dạng dữ liệu đầu vào**	**Cách đưa dữ liệu vào mô hình**	**Đầu ra dự đoán**
Logistic Regression đa lớp	Vector 15 đặc trưng: $X \in \mathbb{R}^{(N\times 15)}$	Tính điểm tuyến tính cho từng ngành, sau đó chuyển thành xác suất bằng Softmax	Xác suất thuộc 12 ngành; chọn Top-1 và Top-3
Random Forest Classifier	Vector 15 đặc trưng: $X \in \mathbb{R}^{(N\times 15)}$	Đưa cùng vector đặc trưng vào 100 cây quyết định; tổng hợp kết quả dự đoán của các cây	Nhãn dự đoán cuối cùng và xác suất theo tỷ lệ cây bỏ phiếu
LSTM_Net	Tensor chuỗi: $X \in \mathbb{R}^{(N\times 3\times 15)}$	Đưa chuỗi đặc trưng qua các lớp LSTM để học xu hướng biến đổi theo thời gian	Xác suất thuộc 12 ngành; đánh giá Top-1 và Top-3

Bảng 3.3. Cách tổ chức dữ liệu đầu vào cho các mô hình thử nghiệm

3.1.2. Công cụ và môi trường thử nghiệm

Để triển khai, huấn luyện và đánh giá các mô hình, nghiên cứu sử dụng Python làm ngôn ngữ lập trình chính. Hai mô hình Logistic Regression đa lớp và Random Forest Classifier được triển khai bằng thư viện scikit-learn; mô hình LSTM_Net được xây dựng và huấn luyện bằng PyTorch. Các thư viện NumPy, pandas và Matplotlib được sử dụng để xử lý dữ liệu, tính toán chỉ số đánh giá và trực quan hóa kết quả.

Hạng mục	Thông tin
Ngôn ngữ lập trình	Python 3.10
Framework học sâu	PyTorch 2.x
Thư viện học máy	Scikit-learn 1.x
Thư viện xử lý số liệu	NumPy, pandas
Thư viện trực quan hóa	Matplotlib
Môi trường huấn luyện	CPU (Intel/AMD) hoặc GPU (CUDA – Google Colab)
Hệ điều hành	Windows 10/11
Quản lý mã nguồn	Git + GitHub

Bảng 3.4. Phần mềm và môi trường thực nghiệm

3.2. Triển khai thử nghiệm và kết quả

Mục này trình bày quá trình triển khai thử nghiệm và kết quả đánh giá riêng của từng mô hình. Đối với mỗi mô hình, nội dung được trình bày theo cùng một logic: mục tiêu thử nghiệm, cách đưa dữ liệu đầu vào vào mô hình, cấu hình tham số, quy trình huấn luyện - kiểm thử và kết quả đánh giá riêng. Cách trình bày thống nhất này giúp kết quả của ba mô hình có thể được tổng hợp và so sánh trực tiếp ở mục 3.2.2.

3.2.1. Triển khai thử nghiệm và đánh giá riêng từng mô hình

3.2.1.1. Triển khai thử nghiệm mô hình hồi quy tuyến tính

Trên cơ sở lý thuyết đã trình bày tại mục 2.2.1, mô hình tuyến tính trong phần thực nghiệm được triển khai dưới dạng Logistic Regression đa lớp. Đây là mô hình cơ sở tuyến tính, được sử dụng nhằm đánh giá mức hiệu quả ban đầu của hướng tiếp cận tuyến tính đối với bài toán dự đoán ngành học phù hợp cho thí sinh.

Dữ liệu đầu vào của mô hình là ma trận đặc trưng X_ML gồm 15 đặc trưng đã được tiền xử lý. Mỗi dòng trong ma trận tương ứng với một học sinh, mỗi cột tương ứng với một đặc trưng điểm học tập hoặc đặc trưng RIASEC. Khi áp dụng công thức tại Chương 2, vector đặc trưng của một học sinh được đưa vào mô hình để tính điểm số tuyến tính cho từng ngành học.

$$z_k = \beta_{0k} + \beta_{1k} x_1 + \beta_{2k} x_2 + \ldots + \beta_{15k} x_{15}$$

Trong đó, z_k là điểm số tuyến tính của học sinh đối với ngành học thứ k; x_1, x_2, ..., x_{15} là 15 đặc trưng đầu vào; β_{0k}, β_{1k}, ..., β_{15k} là các hệ số của mô hình đối với ngành học k. Sau đó, các giá trị z_k được chuyển thành xác suất phù hợp với từng ngành học bằng hàm Softmax.

$$P(y = k|x) = \exp(z_k) / \Sigma_{j=1}^{K} \exp(z_j)$$

Nhãn dự đoán Top-1 là ngành học có xác suất lớn nhất. Danh sách Top-3 được xác định bằng cách sắp xếp xác suất dự đoán của 12 ngành theo thứ tự giảm dần và lấy ba ngành có xác suất cao nhất. Cách đưa dữ liệu đầu vào vào mô hình được tóm tắt như sau: điểm học tập và điểm RIASEC của từng học sinh tạo thành vector x; mô hình tính xác suất phù hợp của x với từng ngành; kết quả dự đoán được dùng để đánh giá trên tập kiểm định.

Tham số	Giá trị	Ý nghĩa
Thuật toán	Logistic Regression	Hồi quy logistic đa lớp (softmax)
Solver	lbfgs	Tối ưu hoá L-BFGS (Limited-memory BFGS)
Multi-class	auto (multinomial)	Phân loại đồng thời 12 lớp
Regularization	L2 (mặc định C=1.0)	Chuẩn hoá tránh overfitting
max_iter	1000	Số vòng lặp tối đa
n_jobs	-1	Sử dụng toàn bộ CPU
Phần mềm	scikit-learn (Python)	Thư viện ML chuẩn

Bảng 3.5. Thông số cấu hình mô hình hồi quy tuyến tính

Mô hình được huấn luyện trên tập huấn luyện và đánh giá trên tập kiểm định. Kết quả đánh giá riêng của mô hình được xem xét theo hai nhóm: (i) mức độ ảnh hưởng của các đặc trưng đầu vào thông qua hệ số mô hình; (ii) hiệu năng phân loại theo từng ngành học thông qua Precision, Recall và F1-score.

Ngành	Intercept	Toán	Văn	Anh	Lý	Hoá	Sinh	Sử
Chăm sóc sắc đẹp	+7,8444	-0,8434	-0,3968	+1,1838	+0,9619	+0,0748	+0,0529	+0,0258
Công nghệ ô tô	-6,3753	+1,5086	+2,7429	-0,8696	-1,0251	-0,0027	+0,2634	-0,0776
Dược	-7,3184	-0,8684	-0,5491	+2,1543	+1,8271	+0,1593	-0,0129	+0,0896
Kỹ thuật chế biến món ăn	+11,1588	-0,8158	-0,3937	+1,0802	+0,7337	+0,0282	+0,0347	+0,0278
Kỹ thuật phục hồi chức năng	+11,2655	-0,6235	-0,2274	-0,6254	-0,6277	-0,1112	+0,0429	+0,0406
Ngôn ngữ và văn hóa Hàn Quốc	+7,9531	-0,4878	-0,1095	-0,6660	-0,6552	-0,2783	-0,0587	+0,0380
Ngôn ngữ và văn hóa Trung Quốc	+10,6299	-0,4978	-0,1361	-0,6617	-0,6249	-0,2656	-0,0323	+0,0250
Quản trị khách sạn	-12,2995	-0,4203	-0,0125	-0,4399	-0,4977	-0,6138	-0,1274	-0,0218
Quản trị kinh doanh	-4,1037	+1,2881	-0,2815	-0,7435	-0,7205	-0,5080	-0,1076	+0,0440
Thiết kế đồ họa	-9,5557	-0,2595	+0,1390	-0,3842	-0,4736	+1,6590	-0,2059	-0,0936
Điều dưỡng	-4,6208	-0,8599	-0,5016	+1,0458	+1,8622	+0,0652	+0,0212	+0,0491
Ứng dụng phần mềm	-4,5784	+2,8795	-0,2736	-1,0737	-0,7602	-0,2069	+0,1296	-0,1470

Bảng 3.6. Hệ số hồi quy theo ngành học - Phần 1 (Intercept, Toán, Văn, Anh, Lý, Hoá, Sinh, Sử)

Ngành	Địa	GDCD	R (Thực tế)	I (Nghiên cứu)	A (Nghệ thuật)	S (Xã hội)	E (Doanh nhân)	C (Quy ước)
Chăm sóc sắc đẹp	-0,8146	+0,0772	+0,0897	-1,1745	+1,1713	+0,2833	-0,8649	-0,2435
Công nghệ ô tô	-0,6831	-0,0119	+1,3779	-0,0793	-1,0846	-1,4989	-1,4819	+0,1999
Dược	-0,8285	-0,0601	-0,4531	+3,6085	-1,0547	-1,0540	-1,3046	+0,2280
Kỹ thuật chế biến món ăn	-0,8200	+0,0533	+0,5791	-1,1509	+0,1039	-0,6197	-0,8644	-0,2480
Kỹ thuật phục hồi chức năng	-0,9576	-0,0241	+0,1766	+0,2892	-1,1159	+0,2232	-1,4525	-0,3236
Ngôn ngữ và văn hóa Hàn Quốc	-1,0334	+0,0106	-0,3013	-0,7768	+1,0027	+0,2195	+0,3668	-0,3105
Ngôn ngữ và văn hóa Trung Quốc	-1,0367	-0,0102	-0,3055	-0,7765	-0,0974	+0,2141	+0,4075	-0,3178
Quản trị khách sạn	+3,5021	+0,0283	-0,2786	-0,6964	-0,8456	+0,0307	+5,2532	+0,2227
Quản trị kinh doanh	+1,8901	+0,0538	-0,3460	-0,8887	-1,2832	-0,7853	+3,2831	+0,8752
Thiết kế đồ họa	+0,6827	-0,0942	-0,2047	-0,6199	+5,4500	-1,2714	-1,2155	-0,1801
Điều dưỡng	-0,8815	+0,0283	-0,3741	-0,8670	-1,0071	+5,7817	-0,8583	-0,2307
Ứng dụng phần mềm	+0,9805	-0,0508	+0,0399	+3,1323	-1,2393	-1,5233	-1,2685	+0,3284

Bảng 3.7. Hệ số hồi quy theo ngành học - Phần 2 (Địa, GDCD, R, I, A, S, E, C)

Xếp hạng	Đặc trưng	Trị tuyệt đối hệ số trung bình
1	E (Doanh nhân)	1,5518
2	A (Nghệ thuật)	1,2880
3	Địa	1,1759
4	I (Nghiên cứu)	1,1717
5	S (Xã hội)	1,1254
6	Toán	0,9461
7	Anh	0,9107

Xếp hạng	Đặc trưng	Trị tuyệt đối hệ số trung bình
8	Lý	0,8975
9	Văn	0,4803
10	R (Thực tế)	0,3772
11	Hoá	0,3311
12	C (Quy ước)	0,3090
13	Sinh	0,0908
14	Sử	0,0567
15	GDCD	0,0419

Bảng 3.8. Mức độ ảnh hưởng trung bình của từng đặc trưng (mô hình hồi quy tuyến tính)

Ngành	Precision	Recall	F1-Score	Support
Chăm sóc sắc đẹp	75.86%	79.44%	77.61%	2,500
Công nghệ ô tô	99.36%	99.12%	99.24%	2,500
Dược	97.51%	96.96%	97.23%	2,500
Kỹ thuật chế biến món ăn	80.56%	77.56%	79.03%	2,500
Kỹ thuật phục hồi chức năng	96.44%	96.48%	96.46%	2,500
Ngôn ngữ và văn hóa Hàn Quốc	63.15%	81.64%	71.21%	2,500
Ngôn ngữ và văn hóa Trung Quốc	77.25%	51.60%	61.87%	2,500
Quản trị khách sạn	88.58%	91.84%	90.18%	2,500
Quản trị kinh doanh	93.41%	92.92%	93.16%	2,500
Thiết kế đồ họa	95.32%	96.12%	95.72%	2,500
Điều dưỡng	93.09%	92.64%	92.86%	2,500
Ứng dụng phần mềm	98.80%	98.96%	98.88%	2,500

Bảng 3.9. Kết quả thử nghiệm mô hình hồi quy tuyến tính

Kết quả tại Bảng 3.9 cho thấy Logistic Regression đa lớp đạt hiệu năng tốt ở nhiều ngành có đặc trưng học tập và RIASEC rõ ràng như Công nghệ ô tô, Ứng dụng phần mềm, Dược, Kỹ thuật phục hồi chức năng và Thiết kế đồ họa. Tuy nhiên, mô hình còn gặp khó khăn ở nhóm ngành Ngôn ngữ và văn hóa Hàn Quốc, Ngôn ngữ và

văn hóa Trung Quốc do hai nhóm ngành này có đặc trưng đầu vào chồng lấp cao. Kết quả này cho thấy mô hình tuyến tính là mốc đối sánh phù hợp, nhưng vẫn cần so sánh với mô hình phi tuyến và mô hình học sâu để đánh giá khả năng cải thiện.

3.2.1.2. Triển khai thử nghiệm mô hình hồi quy phi tuyến

Trên cơ sở lý thuyết đã trình bày tại mục 2.2.2, mô hình phi tuyến trong phần thực nghiệm được triển khai bằng thuật toán Random Forest Classifier. Đây là mô hình cơ sở phi tuyến, được sử dụng nhằm đánh giá khả năng khai thác các mối quan hệ phức tạp giữa điểm học tập, đặc điểm RIASEC và ngành học phù hợp.

Dữ liệu đầu vào của Random Forest sử dụng cùng ma trận X_ML gồm 15 đặc trưng như mô hình Logistic Regression. Mỗi vector đặc trưng của học sinh được đưa vào toàn bộ các cây quyết định trong rừng. Mỗi cây tạo ra một dự đoán ngành học, sau đó mô hình tổng hợp kết quả để xác định nhãn cuối cùng và xác suất tương ứng cho từng ngành.

$$h_1(x), h_2(x), \ldots, h_B(x), \text{ với } B = 100$$

$$P(y = k|x) = (1/B) \Sigma_{\beta=1}^{B} I(h_b(x) = k)$$

Trong đó, h_b(x) là kết quả dự đoán của cây thứ b, B là số cây quyết định trong rừng và I(h_b(x)=k) là hàm chỉ báo nhận giá trị 1 nếu cây thứ b dự đoán ngành k. Với cấu hình 100 cây quyết định, xác suất của một ngành được hiểu là tỷ lệ số cây trong rừng bỏ phiếu cho ngành đó. Nhãn Top-1 là ngành có tỷ lệ bỏ phiếu cao nhất; Top-3 là ba ngành có tỷ lệ bỏ phiếu cao nhất.

Tham số	**Giá trị**	**Ý nghĩa**
Thuật toán	Random Forest	Tập hợp (ensemble) cây quyết định
n_estimators	100	Số cây quyết định trong rừng
max_features	sqrt(n_features)	Số đặc trưng xét tại mỗi nút (mặc định)
criterion	gini	Tiêu chí phân tách nút
max_depth	None	Không giới hạn độ sâu cây
min_samples_split	2	Số mẫu tối thiểu để tách nút
min_samples_leaf	1	Số mẫu tối thiểu tại nút lá
bootstrap	True	Lấy mẫu có hoàn lại (Bagging)

Tham số	Giá trị	Ý nghĩa
random_state	42	Seed để tái tạo kết quả
n_jobs	-1	Sử dụng toàn bộ CPU
Phần mềm	Scikit-learn (Python)	Thư viện học máy chuẩn

Bảng 3.10. Cấu hình mô hình hồi quy phi tuyến (Random Forest)

Quy trình thử nghiệm của Random Forest được thực hiện thống nhất với Logistic Regression: huấn luyện trên tập huấn luyện, dự đoán trên tập kiểm định, tính toán các chỉ số đánh giá và phân tích mức độ quan trọng của từng đặc trưng. Việc sử dụng cùng tập dữ liệu và cùng hệ đặc trưng giúp kết quả so sánh giữa mô hình tuyến tính và phi tuyến có tính khách quan.

Xếp hạng	Đặc trưng	Feature Importance	Tỷ lệ (%)
1	S (Xã hội)	0,042400	4,24%
2	A (Nghệ thuật)	0,041397	4,14%
3	E (Doanh nhân)	0,033797	3,38%
4	Lý	0,030030	3,00%
5	Toán	0,028469	2,85%
6	Địa	0,027482	2,75%
7	I (Nghiên cứu)	0,025751	2,58%
8	Anh	0,023186	2,32%
9	Văn	0,020840	2,08%
10	R (Thực tế)	0,020214	2,02%
11	Hoá	0,011669	1,17%
12	C (Quy ước)	0,011512	1,15%
13	Sinh	0,007131	0,71%
14	GDCD	0,007109	0,71%
15	Sử	0,007000	0,70%

Bảng 3.11. Mức độ quan trọng đặc trưng - Feature Importance (mô hình hồi quy phi tuyến)

Ngành	Precision	Recall	F1-Score	Support
Chăm sóc sắc đẹp	74.49%	84.92%	79.36%	2,500
Công nghệ ô tô	99.47%	98.20%	98.83%	2,500
Dược	97.71%	95.40%	96.54%	2,500
Kỹ thuật chế biến món ăn	82.88%	71.48%	76.76%	2,500
Kỹ thuật phục hồi chức năng	95.69%	95.88%	95.78%	2,500
Ngôn ngữ và văn hóa Hàn Quốc	61.95%	83.24%	71.04%	2,500
Ngôn ngữ và văn hóa Trung Quốc	75.52%	51.32%	61.11%	2,500
Quản trị khách sạn	90.26%	92.28%	91.26%	2,500
Quản trị kinh doanh	94.10%	93.16%	93.63%	2,500
Thiết kế đồ họa	95.53%	94.88%	95.20%	2,500
Điều dưỡng	92.28%	92.32%	92.30%	2,500
Ứng dụng phần mềm	98.21%	98.48%	98.34%	2,500

Bảng 3.12. Kết quả thực nghiệm mô hình hồi quy phi tuyến

Kết quả tại Bảng 3.12 cho thấy Random Forest đạt hiệu năng cao ở các ngành có đặc trưng rõ ràng như Công nghệ ô tô, Dược, Ứng dụng phần mềm, Thiết kế đồ họa, Quản trị kinh doanh và Quản trị khách sạn. Tuy nhiên, tương tự mô hình tuyến tính, mô hình vẫn gặp khó khăn ở nhóm ngành ngôn ngữ - văn hóa, đặc biệt là Ngôn ngữ và văn hóa Trung Quốc. Điều này cho thấy nguyên nhân sai lệch không chỉ xuất phát từ giới hạn tuyến tính của mô hình, mà còn liên quan đến mức độ chồng lấp đặc trưng giữa các ngành học.

3.2.1.3. Triển khai thử nghiệm mô hình LSTM_Net

Trên cơ sở lý thuyết về LSTM đã trình bày tại mục 2.2.3, đề án triển khai mô hình học sâu LSTM_Net nhằm khai thác đặc điểm dữ liệu học tập có tính chuỗi thời gian. Khác với Logistic Regression đa lớp và Random Forest Classifier xử lý dữ liệu

chủ yếu dưới dạng vector đặc trưng tĩnh, LSTM_Net được sử dụng để học xu hướng biến đổi kết quả học tập qua các năm học.

Dữ liệu đầu vào của LSTM_Net được tổ chức lại thành tensor ba chiều. Với mỗi học sinh, dữ liệu được biểu diễn theo trật tự thời gian của ba giai đoạn học tập. Tại mỗi bước thời gian, mô hình tiếp nhận 15 đặc trưng gồm điểm học tập và các đặc trưng RIASEC đã được chuẩn hóa. Trường hợp nhóm RIASEC được thu thập một lần, các giá trị này được kết hợp cùng đặc trưng học tập tại từng bước thời gian để bảo đảm cấu trúc đầu vào thống nhất.

$$X_LSTM \in \mathbb{R}^{(N\times3\times15)}$$

$$o = W\, h_T + b$$

$$P(y = k|X) = \text{Softmax}(o)_k$$

Trong đó, N là số lượng học sinh, 3 là số bước thời gian, 15 là số đặc trưng tại mỗi bước thời gian, h_T là trạng thái ẩn cuối cùng của LSTM và o là vector logits trước khi chuyển thành xác suất theo 12 ngành học. Mô hình sử dụng phân phối xác suất đầu ra để xác định Top-1 và Top-3 ngành học phù hợp.

Lớp	Loại	Thông số	Chi tiết
Input	-	(N, 3, 15)	Chuỗi 3 năm học, 15 đặc trưng/năm
LSTM Layer 1	LSTM	input=15, hidden=128	Đơn hướng (Unidirectional)
LSTM Layer 2	LSTM	hidden=128→128	Dropout=0.3 giữa các lớp
FC Layer 1	Linear	128 → 128	Fully Connected
Activation	ReLU	-	Hàm kích hoạt phi tuyến
Regularization	Dropout	p=0.4	Chống overfitting
FC Layer 2 (Output)	Linear	128 → 12	Logits cho 12 ngành
Output	Softmax	12 xác suất	Tổng = 1.0

Bảng 3.13. Kiến trúc mô hình LSTM_Net

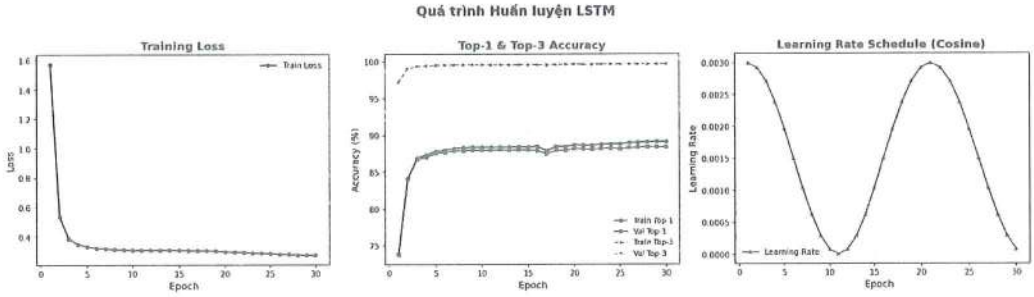
Tên tham số	Số lượng
lstm.weight_ih_l0	7,680
lstm.weight_hh_l0	65,536
lstm.bias_ih_l0	512
lstm.bias_hh_l0	512
lstm.weight_ih_l1	65,536
lstm.weight_hh_l1	65,536
lstm.bias_ih_l1	512
lstm.bias_hh_l1	512
fc.0.weight	16,384
fc.0.bias	128
fc.3.weight	1,536
fc.3.bias	12
TỔNG THAM SỐ	224,396
Tham số có thể huấn luyện	224,396

Bảng 3.14. Bảng tham số mô hình LSTM_Net

Tham số	Giá trị	Ý nghĩa
Optimizer	AdamW	Adaptive Moment Estimation + Weight Decay
Learning Rate	3×10^{-3}	Tốc độ học ban đầu
Weight Decay	1×10^{-4}	L2 Regularization
Batch Size	4,096	Số mẫu mỗi mini-batch
Max Epochs	30	Số epoch tối đa
LR Scheduler	CosineAnnealingLR	T_max=10, η_min=10^{-5}
Early Stopping	Patience=5	Dừng sớm nếu loss không giảm
Gradient Clipping	max_norm=1.0	Tránh gradient bùng nổ
Phần mềm	PyTorch 2.x (Python)	Framework học sâu

Bảng 3.15. Cấu hình huấn luyện LSTM_Net

Quá trình huấn luyện LSTM_Net được thực hiện bằng PyTorch. Mô hình sử dụng hàm mất mát Cross-Entropy phù hợp với bài toán phân loại đa lớp, bộ tối ưu AdamW, lịch trình tốc độ học CosineAnnealingLR, cơ chế Early Stopping và Gradient Clipping nhằm ổn định quá trình huấn luyện. Trong quá trình huấn luyện, các chỉ số loss, Top-1 Accuracy và Top-3 Accuracy được theo dõi qua từng epoch.



Hình 3.1. Quá trình huấn luyện mô hình LSTM qua 30 epoch - Training Loss, Top-1 và Top-3 Accuracy và lịch trình tốc độ học dạng Cosine

Quan sát Hình 3.1 cho thấy giá trị Train Loss giảm nhanh trong những epoch đầu tiên, sau đó giảm chậm và ổn định, phản ánh mô hình đã hội tụ tốt. Top-1 Accuracy duy trì ở mức ổn định, trong khi Top-3 Accuracy đạt mức rất cao, phù hợp với mục tiêu tư vấn đồng thời nhiều ngành học cho thí sinh.

Ngành	Precision	Recall	F1-Score	Support
Chăm sóc sắc đẹp	76.17%	83.76%	79.79%	2,500
Công nghệ ô tô	99.16%	99.08%	99.12%	2,500
Dược	97.31%	96.88%	97.09%	2,500
Kỹ thuật chế biến món ăn	82.64%	75.76%	79.05%	2,500
Kỹ thuật phục hồi chức năng	96.71%	96.32%	96.51%	2,500
Ngôn ngữ và văn hóa Hàn Quốc	63.93%	84.08%	72.63%	2,500
Ngôn ngữ và văn hóa Trung Quốc	79.11%	50.76%	61.84%	2,500
Quản trị khách sạn	88.96%	92.84%	90.86%	2,500
Quản trị kinh doanh	94.02%	93.08%	93.55%	2,500

Ngành	Precision	Recall	F1-Score	Support
Thiết kế đồ họa	95.83%	97.32%	96.57%	2,500
Điều dưỡng	93.61%	92.52%	93.06%	2,500
Ứng dụng phần mềm	98.72%	98.56%	98.64%	2,500

Bảng 3.16. Kết quả thử nghiệm mô hình LSTM_Net

Kết quả tại Bảng 3.16 cho thấy mô hình LSTM_Net đạt hiệu quả cao đối với các ngành có đặc trưng học tập và định hướng nghề nghiệp rõ ràng. Các ngành Công nghệ ô tô, Ứng dụng phần mềm, Dược, Thiết kế đồ họa và Kỹ thuật phục hồi chức năng đều đạt F1-score cao. Mô hình vẫn còn hạn chế đối với nhóm ngành ngôn ngữ - văn hóa; tuy nhiên, xét tổng thể, LSTM_Net cho thấy khả năng khai thác tốt hơn dữ liệu học tập theo trình tự thời gian so với hai mô hình cơ sở.

3.2.2. Tổng hợp kết quả và so sánh giữa các mô hình

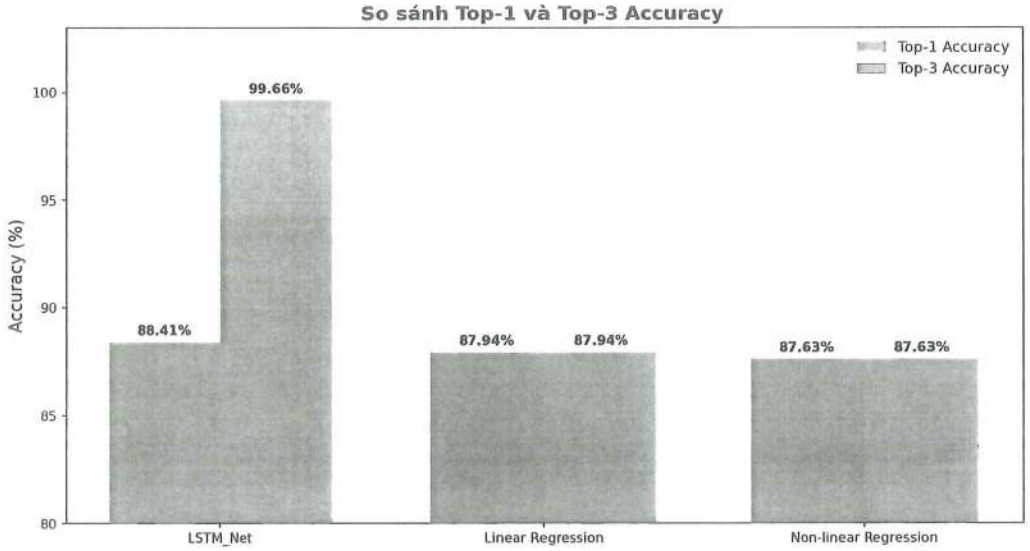
Sau khi đánh giá riêng từng mô hình tại mục 3.2.1, kết quả tổng hợp của ba mô hình được trình bày nhằm so sánh trực tiếp hiệu quả dự đoán trên cùng tập dữ liệu kiểm định. Các chỉ số được sử dụng gồm Accuracy huấn luyện, Accuracy kiểm định, Top-1 Accuracy và Top-3 Accuracy; đồng thời, các chỉ số Precision, Recall và F1-score theo từng ngành được sử dụng để phân tích sâu hơn mức độ ổn định của mô hình.

Mô hình	Accuracy huấn luyện (%)	Accuracy kiểm định (%)	Top-1 Accuracy (%)	Top-3 Accuracy (%)
LSTM_Net	88,4	88,4	88,41	99,66
Hồi quy tuyến tính	87,9	87,9	87,94	87,94
Hồi quy phi tuyến	87,6	87,6	87,63	87,63

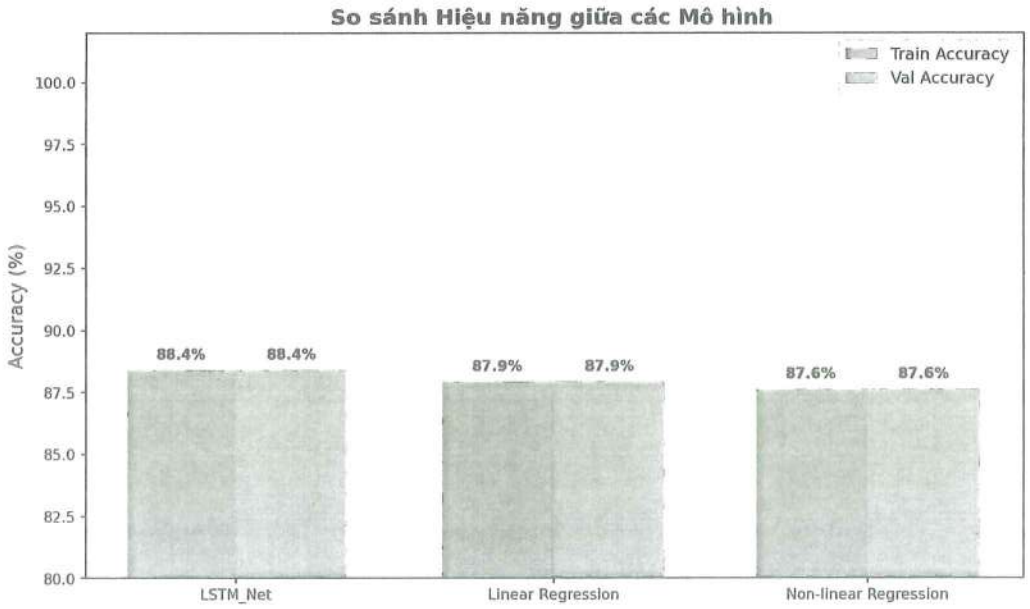
Bảng 3.17. Kết quả Accuracy của ba mô hình trên tập huấn luyện và tập kiểm định

Kết quả tại Bảng 3.17 cho thấy độ chính xác trên tập huấn luyện và tập kiểm định của cả ba mô hình có chênh lệch nhỏ, phản ánh khả năng tổng quát hóa tương

đối tốt và không có dấu hiệu quá khớp rõ rệt. Về Top-1 Accuracy, LSTM_Net đạt kết quả cao nhất với 88,41%, tiếp theo là Logistic Regression với 87,94% và Random Forest với 87,63%.

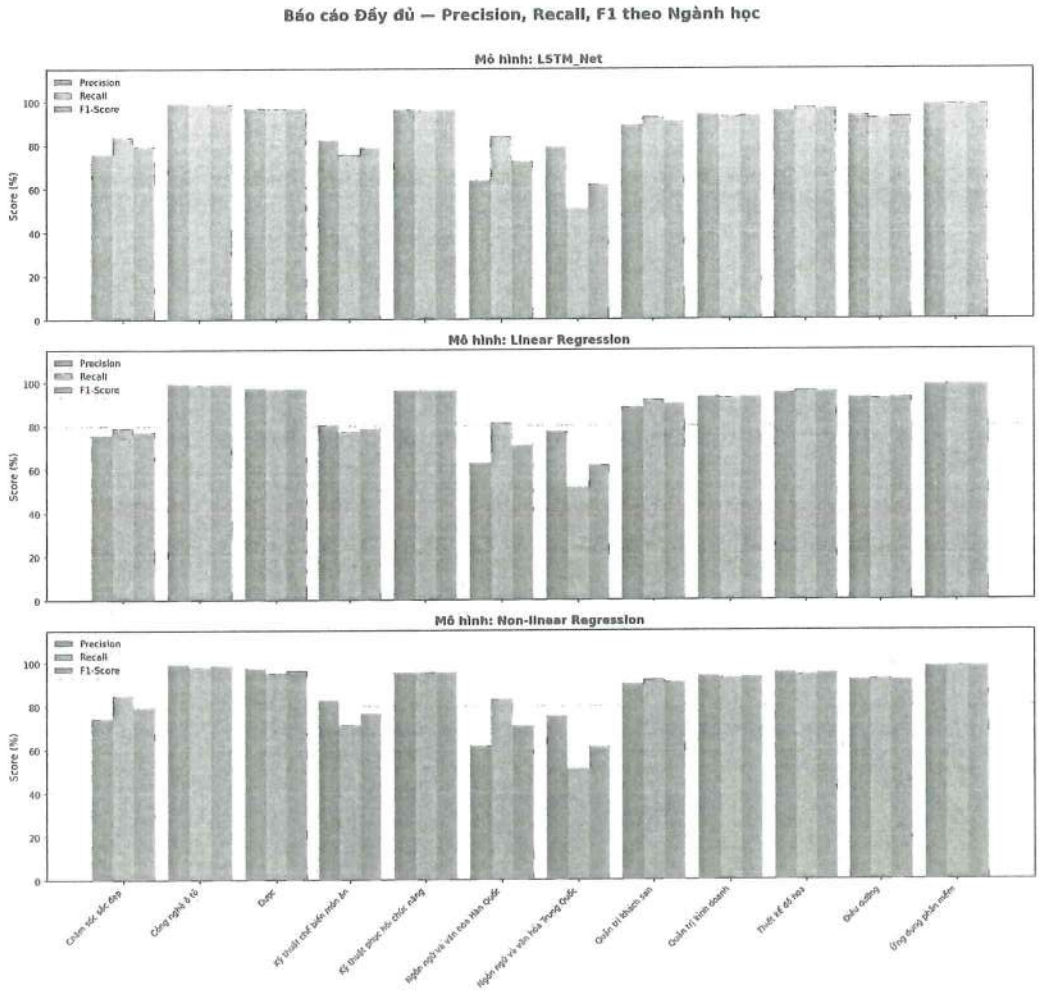


Hình 3.2. So sánh Top-1 và Top-3 Accuracy giữa ba mô hình



Hình 3.3. So sánh Accuracy huấn luyện và Accuracy kiểm định giữa ba mô hình

Hình 3.2 và Hình 3.3 minh họa trực quan kết quả so sánh giữa ba mô hình. Trong đó, ưu thế nổi bật nhất của LSTM_Net nằm ở Top-3 Accuracy, đạt 99,66%, phù hợp với mục tiêu thực tế của bài toán tư vấn tuyển sinh là gợi ý ba ngành học phù hợp nhất thay vì chỉ đưa ra một nhãn dự đoán duy nhất.



Hình 3.4. Precision, Recall và F1-Score theo từng ngành học của ba mô hình trên tập kiểm định

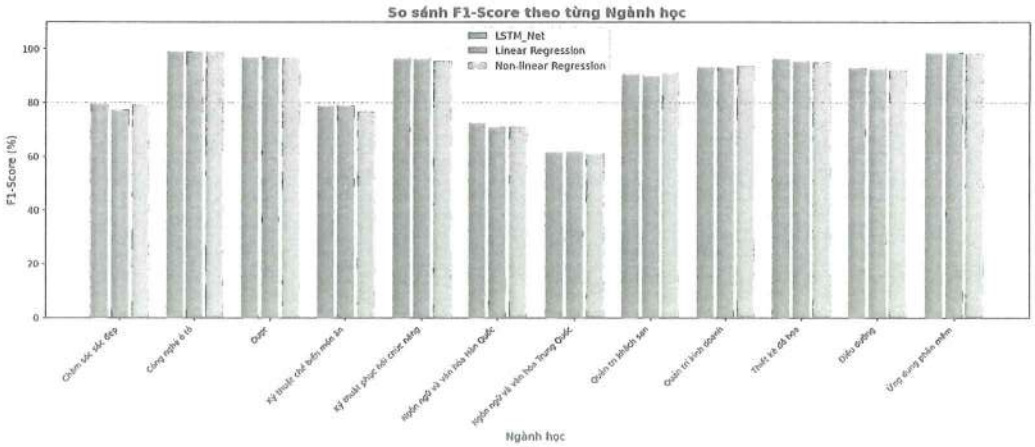
Kết quả tại Hình 3.4 cho thấy hiệu năng dự đoán không đồng đều giữa các ngành học. Nhóm ngành kỹ thuật, công nghệ và sức khỏe đạt kết quả cao ở cả ba mô hình, trong khi nhóm ngành ngôn ngữ - văn hóa có F1-score thấp hơn. Điều này cho

thấy độ khó của bài toán không chỉ phụ thuộc vào thuật toán, mà còn phụ thuộc vào mức độ phân biệt của các đặc trưng đầu vào giữa các ngành học.

3.3. Nhận xét và đánh giá

Tổng hợp các kết quả thử nghiệm cho thấy cả ba mô hình đều có khả năng dự đoán ngành học phù hợp ở mức khá cao. Logistic Regression đa lớp là mô hình cơ sở tuyến tính đơn giản, dễ triển khai và có khả năng giải thích tương đối tốt thông qua hệ số mô hình. Random Forest Classifier cho phép khai thác các quan hệ phi tuyến và đánh giá mức độ quan trọng của đặc trưng, tuy nhiên trong thử nghiệm này chưa tạo ra cải thiện rõ rệt so với Logistic Regression.

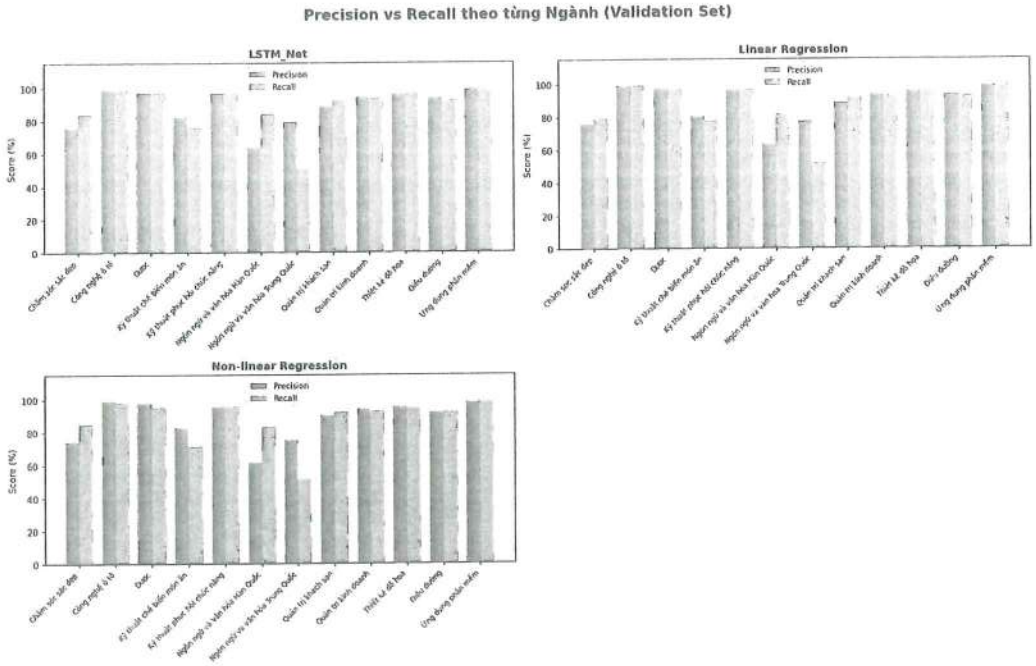
Mô hình LSTM_Net đạt kết quả tổng thể tốt nhất, đặc biệt ở chỉ số Top-3 Accuracy. Ưu thế này phù hợp với bản chất bài toán tư vấn tuyển sinh, bởi hệ thống không chỉ cần dự đoán một ngành duy nhất mà cần gợi ý một danh sách ngành học phù hợp để thí sinh tham khảo. Việc khai thác dữ liệu học tập theo trình tự thời gian giúp LSTM_Net nắm bắt xu hướng tiến bộ, ổn định hoặc thay đổi năng lực học tập của học sinh qua các năm.



Hình 3.5. So sánh F1-Score theo từng ngành học giữa ba mô hình

Hình 3.5 cho thấy xu hướng kết quả theo từng ngành của ba mô hình tương đối giống nhau. Những ngành dễ phân biệt như Công nghệ ô tô, Ứng dụng phần mềm, Dược và Thiết kế đồ họa đều đạt F1-score cao. Ngược lại, nhóm ngành Ngôn ngữ và

văn hóa Hàn Quốc, Ngôn ngữ và văn hóa Trung Quốc là những ngành khó dự đoán hơn do có đặc trưng đầu vào chồng lấp.



Hình 3.6. Precision so với Recall theo từng ngành học của ba mô hình trên tập kiểm định

Hình 3.6 tiếp tục cho thấy sự cân bằng giữa Precision và Recall của các mô hình ở phần lớn các ngành đào tạo. Tuy nhiên, ở nhóm ngành ngôn ngữ - văn hóa, khoảng cách giữa Precision và Recall còn lớn, thể hiện hiện tượng nhầm lẫn giữa các ngành có đặc điểm gần nhau. Đây là điểm cần được cải thiện trong các nghiên cứu tiếp theo bằng cách bổ sung thêm đặc trưng chuyên biệt hoặc dữ liệu về hành vi, nguyện vọng và kết quả tư vấn thực tế của thí sinh.

Xét trên mục tiêu triển khai hệ thống thực tế, LSTM_Net được đánh giá là mô hình phù hợp nhất để tích hợp vào chức năng tư vấn ngành học. Tuy nhiên, việc lựa chọn mô hình triển khai cũng cần cân nhắc chi phí tính toán, thời gian huấn luyện, khả năng giải thích và khả năng bảo trì hệ thống. Trong giai đoạn đầu triển khai, có thể sử dụng LSTM_Net làm mô hình dự đoán chính, đồng thời duy trì Logistic

Regression và Random Forest làm mô hình đối chứng phục vụ kiểm tra, giải thích và đánh giá bổ sung.

3.4. Kết luận chương 3

Chương 3 đã trình bày quá trình triển khai thử nghiệm và đánh giá ba mô hình dự đoán ngành học phù hợp cho thí sinh, gồm Logistic Regression đa lớp, Random Forest Classifier và LSTM_Net. Các mô hình được thử nghiệm trên cùng tập dữ liệu đã được tiền xử lý, cùng hệ đặc trưng đầu vào và cùng quy trình đánh giá nhằm bảo đảm tính khách quan trong so sánh.

Kết quả thử nghiệm cho thấy Logistic Regression đa lớp và Random Forest Classifier đều đạt độ chính xác tương đối cao, có giá trị như các mô hình cơ sở tuyến tính và phi tuyến. Tuy nhiên, LSTM_Net cho kết quả tốt nhất, đặc biệt ở chỉ số Top-3 Accuracy, qua đó phù hợp hơn với mục tiêu tư vấn thực tế của hệ thống là gợi ý đồng thời ba ngành học phù hợp nhất cho thí sinh.

Trên cơ sở các kết quả thử nghiệm, mô hình LSTM_Net được lựa chọn là mô hình ưu tiên để triển khai trong hệ thống hỗ trợ tư vấn tuyển sinh tại Trường Cao đẳng Hà Nội. Trong các bước tiếp theo, hệ thống cần tiếp tục được kiểm thử trên dữ liệu tuyển sinh thực tế của các kỳ tuyển sinh mới, bổ sung thêm đặc trưng đầu vào và đánh giá phản hồi của cán bộ tư vấn cũng như thí sinh để hoàn thiện chất lượng dự đoán.

KẾT LUẬN

Kết quả đạt được của đề án

Đề án đã hoàn thành các mục tiêu nghiên cứu và xây dựng hệ thống hỗ trợ dự đoán ngành học phù hợp cho học sinh dựa trên kết quả học tập trung học phổ thông kết hợp với đặc điểm định hướng nghề nghiệp theo mô hình RIASEC.

Về mặt lý thuyết, đề án đã nghiên cứu tổng quan về trí tuệ nhân tạo, học máy, học sâu và mạng nơ-ron hồi tiếp LSTM; đồng thời phân tích các phương pháp tiền xử lý dữ liệu, chuẩn hóa dữ liệu và các chỉ số đánh giá mô hình phân loại đa lớp. Đây là cơ sở để xây dựng quy trình xử lý dữ liệu và lựa chọn mô hình phù hợp với bài toán dự đoán ngành học.

Về mặt thực nghiệm, đề án đã xây dựng bộ dữ liệu đầu vào gồm điểm học tập của ba năm học THPT kết hợp với sáu chỉ số RIASEC; thiết kế quy trình tiền xử lý dữ liệu và triển khai thử nghiệm ba mô hình gồm **Logistic Regression**, **Random Forest** và **LSTM_Net**. Sau khi hoàn thiện và hiệu chỉnh quy trình xử lý dữ liệu, các mô hình được đánh giá trên cùng tập kiểm định bằng các chỉ số Top-1 Accuracy, Top-3 Accuracy, Macro Precision, Macro Recall và Macro F1.

Kết quả thực nghiệm cho thấy cả ba mô hình đều đạt hiệu quả dự đoán cao với độ chính xác cao và trong đó **LSTM_Net** đạt kết quả tốt nhất, cho thấy ưu thế của mô hình trong việc khai thác dữ liệu học tập theo chuỗi thời gian kết hợp với đặc trưng định hướng nghề nghiệp.

Bên cạnh đó, đề án cũng đã xây dựng được quy trình đánh giá thống nhất, khắc phục các hạn chế trong quá trình thử nghiệm như chuẩn hóa dữ liệu trước khi chia tập dữ liệu, tính toán chưa chính xác chỉ số Top-3 Accuracy và chưa thống nhất dữ liệu đầu vào khi dự đoán. Việc hiệu chỉnh này giúp kết quả đánh giá phản ánh khách quan hơn năng lực của từng mô hình và nâng cao độ tin cậy của hệ thống.

Những kết quả đạt được của đề án là cơ sở để triển khai hệ thống hỗ trợ tư vấn tuyển sinh theo hướng ứng dụng trí tuệ nhân tạo, góp phần nâng cao chất lượng tư vấn lựa chọn ngành học tại Trường Cao đẳng Hà Nội.

Hướng phát triển tiếp theo

Mặc dù đã đạt được các mục tiêu nghiên cứu đề ra, hệ thống vẫn còn nhiều tiềm năng để tiếp tục hoàn thiện và phát triển.

Trong thời gian tới, đề án sẽ tập trung mở rộng nguồn dữ liệu bằng cách thu thập dữ liệu tuyển sinh thực tế qua nhiều năm của Nhà trường nhằm nâng cao khả năng tổng quát hóa và độ tin cậy của mô hình. Đồng thời, ngoài điểm học tập và đặc điểm RIASEC, hệ thống có thể bổ sung thêm các đặc trưng như nguyện vọng đăng ký, năng lực ngoại ngữ, kỹ năng mềm, điều kiện kinh tế gia đình, sở thích cá nhân và nhu cầu nhân lực của thị trường lao động để nâng cao chất lượng dự đoán.

Về mặt thuật toán, có thể nghiên cứu và thử nghiệm thêm các mô hình học sâu hiện đại như **GRU, Transformer, Attention hoặc TabTransformer**, đồng thời áp dụng các kỹ thuật tối ưu siêu tham số và học tăng cường nhằm tiếp tục cải thiện độ chính xác của hệ thống. Ngoài việc dự đoán một ngành học duy nhất, hệ thống có thể phát triển theo hướng tư vấn đa tiêu chí, giải thích nguyên nhân lựa chọn ngành học và đưa ra các khuyến nghị phù hợp với năng lực của từng học sinh.

Bên cạnh đó, hệ thống cần được tích hợp vào cổng thông tin tuyển sinh hoặc phần mềm quản lý tuyển sinh của Trường Cao đẳng Hà Nội để phục vụ công tác tư vấn trực tuyến. Việc xây dựng giao diện thân thiện, cho phép người học nhập thông tin, xem kết quả dự đoán và tham khảo các ngành học phù hợp sẽ góp phần nâng cao trải nghiệm người dùng và hỗ trợ hiệu quả hơn cho hoạt động tuyển sinh của Nhà trường.

Trong tương lai, khi được bổ sung dữ liệu thực tế và tiếp tục hoàn thiện mô hình, hệ thống không chỉ hỗ trợ công tác tuyển sinh tại Trường Cao đẳng Hà Nội mà còn có thể mở rộng áp dụng cho các cơ sở giáo dục nghề nghiệp khác, góp phần thúc đẩy quá trình chuyển đổi số trong lĩnh vực giáo dục và đào tạo.

DANH MỤC TÀI LIỆU THAM KHẢO

[1]. Lê Văn Phùng (2018) – *"Cơ sở dữ liệu quan hệ và công nghệ phân tích - thiết kế"*, Nhà xuất bản Thông tin và truyền thông.

[2]. Nguyễn Thanh Tuấn (2020) – "Deep Learning cơ bản" - Nhà xuất bản Tp. Hồ Chí Minh.

[3]. Nguyễn Văn Tuấn (2020) - *"Mô hình hồi quy và khám phá khoa học"*, Nhà xuất bản Tp. Hồ Chí Minh.

[4]. Azzouni A., Pujolle G. (2017) – "A Long Short-Term Memory Recurrent Neural Network Framework for Network Traffic Matrix Prediction", arXiv preprint arXiv:1705.05690.

[5]. Cheng J., Dong L., Lapata M. (2016) – "Long Short-Term Memory Networks for Machine Reading", Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 551–561.

[6]. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (2013) – "An Introduction to Statistical Learning with Applications in R", Springer New York Heidelberg Dordrecht London.

[7]. Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2017) – "LSTM: A Search Space Odyssey", IEEE Transactions on Neural Networks and Learning Systems, 28(10), 2222–2232.

[8]. He T., Droppo J. (2016) – "Exploiting LSTM Structure in Deep Neural Networks for Speech Recognition", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5440–5444.

[9]. Karim F., Majumdar S., Darabi H., Chen, S. (2019) – "Insights into LSTM Fully Convolutional Networks for Time Series Classification", IEEE Access, 7, 67718–67725.

[10]. Krichen M., Mihoub A. (2025) – "Long Short-Term Memory Networks: A Comprehensive Survey", AI, 6(9), 215.

[11]. Shai Shalev-Shwartz and Shai Ben-David (2014) – "Understanding Machine Learning: From Theory to Algorithms", Cambridge University Press.

[12]. Staudemeyer R. C., Morris E. R. (2019) – "Understanding LSTM—A Tutorial into Long Short-Term Memory Recurrent Neural Networks", arXiv preprint arXiv:1909.09586.

[13]. Zeyer A., Doetsch P., Voigtlaender P., Schlüter, R., Ney H. (2016) – "A Comprehensive Study of Deep Bidirectional LSTM RNNs for Acoustic Modeling in Speech Recognition", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2452–2456.

[14]. Yan K., Ji Z., Shen W., Wang Y., Huang H. (2019) – "A Hybrid LSTM Neural Network for Energy Consumption Forecasting of Individual Households" - IEEE Access, 7, 157633–157642.

[15]. Zed A. Shaw (2016) – " *Learn Python 3 the Hard Way*", Addison-Wesley.

[16]. Yu Y., Si X., Hu C., Zhang J. (2019) – "A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures", Neural Computation, 31(7), 1235–1270.

[17]. https://moet.gov.vn/

[18]. https://www.python.org

[19]. https://www.tensorflow.org/

KiểmTraTàiLiệu

BÁO CÁO KIỂM TRA TRÙNG LẶP

Thông tin tài liệu

Tên tài liệu:	Lê Minh Tuấn - B23CHIS028 - Đề án tốt nghiệp thạc sĩ
Tác giả:	Lê Minh Tuấn - B23CHIS028
Điểm trùng lặp:	11
Thời gian tải lên:	10:21 09/07/2026
Thời gian sinh báo cáo:	10:24 09/07/2026
Các trang kiểm tra:	80/80 trang

Kết quả kiểm tra trùng lặp



Có 11% nội dung trùng lặp

Có 89% nội dung không trùng lặp

Có 0% nội dung người dùng loại trừ

Có 0% nội dung hệ thống bỏ qua

Học viên

Tuấn

Lê Minh Tuấn

Người hướng dẫn khoa học

TS. Vũ Văn Thỏa

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: Lê Minh Tuấn
Chuyên ngành: Hệ thống thông tin
Khóa: 2023 đợt 1
Tên đề tài: Nghiên cứu xây dựng hệ thống hỗ trợ tuyển sinh tại Trường cao đẳng Hà Nội
Người hướng dẫn khoa học: TS. Vũ Văn Thỏa
Ngày bảo vệ: 08/07/2026

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	Đánh số công thức	Học viên đã rà soát, bổ sung đánh số các công thức tại Chương 2
2	Bổ sung mô tả dữ liệu	Tiếp thu góp ý của Hội đồng, tác giả đã bổ sung thêm Mô tả dữ liệu tại mục 2.5.1 trong Chương 2 và 3.1.1 trong Chương 3

Hà Nội, ngày 10 tháng 07 năm 2026

Ký xác nhận của

CHỦ TỊCH HỘI ĐỒNG CHẤM ĐỀ ÁN	THƯ KÝ HỘI ĐỒNG	NGƯỜI HƯỚNG DẪN KHOA HỌC	HỌC VIÊN
PGS. TSKH. Hoàng Đăng Hải	TS. Đỗ Tiến Dũng	TS. Vũ Văn Thỏa	Lê Minh Tuấn

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 1895/QĐ-HV ngày 29 tháng 06 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..10...giờ..15..phút, ngày 08 tháng 07 năm 2026 tại Học viện Công nghệ Bưu chính Viễn thông để chấm đề án tốt nghiệp thạc sĩ cho:

Học viên: **Lê Minh Tuấn**

Tên đề án tốt nghiệp: Nghiên cứu xây dựng hệ thống hỗ trợ tuyển sinh tại trường Cao Đẳng Hà Nội

Chuyên ngành: Hệ thống thông tin Mã số: **8.48.01.04**

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	**PGS. TSKH. Hoàng Đăng Hải**	Chủ tịch	
2	**TS. Đỗ Tiến Dũng**	Thư ký	
3	**TS. Đặng Hoàng Long**	Phản biện 1	
4	**TS. Bùi Hải Phong**	Phản biện 2	
5	**TS. Nguyễn Xuân Thắng**	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:

Mô tả dữ liệu ?
Ngành phù hợp được đánh giá như thế nào ?
Cô có lựa chọn LSTM Bởi chi phí lớn và độ chính xác không chắc chắn ?

7. Trả lời của học viên:

Đã liên hệ phòng
Nhận thuộc ngành của trường
Cân nhắc khi đưa vào thực tế

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: PGS.TS Hoàng Thùy Hải
  2. Uỷ viên Ban kiểm phiếu: TS. Đặng Hoàng Long
  3. Uỷ viên Ban kiểm phiếu: TS. Đỗ Tiến Dũng
- 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:** 7,5

Kết luận:

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

Đánh số công thức
Bổ sung mô tả dữ liệu

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

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

Buổi làm việc kết thúc vào...10:40... cùng ngày.

Chủ tịch

PGS.TS. Hoàng Thùy Hải

Thư ký

Đỗ Tiến Dũng

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: Nghiên cứu xây dựng hệ thống hỗ trợ tuyển sinh tại trường cao đẳng Hà Nội

Chuyên ngành: Hệ thống thông tin

Mã chuyên ngành: 8.48.01.04

Họ và tên học viên: Lê Minh Tuấn

Họ và tên người nhận xét: Bùi Hải Phong

Học hàm, học vị: Tiến sĩ

Chuyên ngành: Khoa học máy tính

Cơ quan công tác: Trường ĐH Kiến Trúc Hà Nội

Số điện thoại:.0915594033. E-mail:..phongbh@hau.edu.vn

NỘI DUNG NHẬN XÉT

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

Đề tài áp dụng các giải thuật học máy, học sâu trong phân tích dữ liệu, hỗ trợ tuyển sinh tại trường cao đẳng Hà Nội. Đề tài có tính cấp thiết và khoa học.

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

Đề tài gồm 69 trang 3 chương. Đề tài áp dụng các thuật toán Random Forest, Logistic Regression và LSTM trong phân tích dữ liệu hỗ trợ tuyển sinh tại trường Cao Đẳng Hà Nội. Cấu trúc rõ

quy trình thu thập và xử lý dữ liệu
Cần đánh chỉ số các công thức toán học

III/ Những vấn đề cần giải thích thêm:

1. Đánh giá mô hình và đưa ra khuyến nghị sử dụng mô hình học máy, học sâu cho dữ liệu tuyển sinh?
2. Khó khăn trong việc xử lý dữ liệu cho các ngành có ít dữ liệu?

IV/ Kết luận:

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

Đồng ý (hoặc không đồng ý) cho phép học viên bảo vệ đề án tốt nghiệp.

Ngày 8 tháng 7 năm 2026

NGƯỜI NHẬN XÉT

Bùi Hải Phong

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: Nghiên cứu xây dựng hệ thống hỗ trợ tuyển sinh tại Trường Cao đẳng Hà Nội

Chuyên ngành: Hệ thống thông tin

Mã chuyên ngành: 8.48.01.04

Họ và tên học viên: Lê Minh Tuấn

Họ và tên người nhận xét: Đặng Hoàng Long

Học hàm, học vị: Tiến sĩ

Chuyên ngành: Khoa học máy tính

Cơ quan công tác: Học viện Công nghệ Bưu chính Viễn thông

Số điện thoại: 0335229386 E-mail: longdh@ptit.edu.vn

NỘI DUNG NHẬN XÉT

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

Đề tài có cơ sở khoa học và thực tiễn. Về khoa học, đề án dựa trên các mô hình học máy và học sâu phổ biến (Logistic Regression đa lớp, Random Forest, LSTM) kết hợp với mô hình định hướng nghề nghiệp RIASEC để giải quyết bài toán gợi ý ngành học phù hợp cho thí sinh. Về thực tiễn, công tác tư vấn tuyển sinh tại các trường cao đẳng, đặc biệt là Trường Cao đẳng Hà Nội với quy mô 2.000–3.000 chỉ tiêu mỗi năm, hiện vẫn mang tính thủ công và chưa khai thác hiệu quả dữ liệu điểm học tập cũng như nguyện vọng của thí sinh. Vì vậy, việc nghiên cứu và xây dựng một hệ thống hỗ trợ tư vấn chọn ngành dựa trên dữ liệu là hướng đi phù hợp với định hướng ứng dụng của đề án và có ý nghĩa thực tiễn thiết thực.

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

- Khảo sát tổng quan về công tác tuyển sinh, quy chế – quy trình tuyển sinh và các mô hình học máy, học sâu (Logistic Regression, Random Forest, LSTM) áp dụng cho bài toán tư vấn chọn ngành.
- Phân tích thực trạng tuyển sinh tại Trường Cao đẳng Hà Nội, xây dựng bài toán gợi ý ngành học, thiết kế các yêu cầu chức năng, phi chức năng và luồng dữ liệu của hệ thống.
- Thiết kế quy trình thu thập, làm sạch, chuẩn hóa và tổ chức dữ liệu đầu vào gồm 15 đặc trưng (9 điểm môn học văn hóa và 6 nhóm định hướng nghề nghiệp RIASEC) phục vụ huấn luyện các mô hình dự đoán.

- Triển khai thử nghiệm và so sánh ba mô hình theo các chỉ số Top-1/Top-3 Accuracy, Precision, Recall và Macro F1; kết quả cho thấy LSTM_Net đạt hiệu quả tốt nhất (Top-1 88,41%, Top-3 99,66%) và được lựa chọn để tích hợp vào hệ thống.

III/ Những vấn đề cần giải thích thêm:

- Bộ dữ liệu thực nghiệm có kích thước bao nhiêu mẫu, được thu thập từ dữ liệu tuyển sinh thực tế của Nhà trường hay dữ liệu mô phỏng? Cách phân chia tập huấn luyện – kiểm định cụ thể như thế nào?
- Nhãn "ngành phù hợp" dùng để huấn luyện và đánh giá được xác định như thế nào: dựa trên ngành thí sinh thực tế trúng tuyển/nhập học hay theo một tiêu chí đánh giá độc lập?
- Đối với LSTM_Net, dữ liệu 15 đặc trưng được tổ chức thành chuỗi thời gian 3 bước (N×3×15) như thế nào, trong khi điểm các môn được mô tả là điểm trung bình 3 năm và các chỉ số RIASEC không thay đổi theo thời gian?
- LSTM_Net chỉ cải thiện khoảng 0,5% Top-1 Accuracy so với Logistic Regression nhưng chi phí huấn luyện, tính toán và bảo trì cao hơn. Việc lựa chọn LSTM_Net để triển khai chính thức được cân nhắc trên cơ sở nào?

IV/ Kết luận:

Đề án đáp ứng yêu cầu của một đề án tốt nghiệp thạc sĩ theo định hướng ứng dụng. Đồng ý cho phép học viên bảo vệ đề án tốt nghiệp.

Hà Nội, ngày ..8.. tháng .4 ... năm 2026

NGƯỜI NHẬN XÉT

Lay

Đặng Hoàng Lay