BỘ GIÁO DỤC ĐÀO TẠO

TRƯỜNG ĐẠI HỌC ĐẠI NAM

NGUYỄN HẢI ĐĂNG

MSV: 1571020059

PHÂN TÍCH VÀ DỰ ĐOÁN TỶ LỆ THẤT NGHIỆP KẾT HỢP DỮ LIỆU KINH TẾ VÀ XÃ HỘI

BÁO CÁO ĐỒ ÁN TỐT NGHIỆP

NGÀNH: CÔNG NGHỆ THÔNG TIN

HÀ NỘI, NĂM 2025

BỘ GIÁO DỤC ĐÀO TẠO

TRƯỜNG ĐẠI HỌC ĐẠI NAM

NGUYỄN HẢI ĐĂNG

MSV: 1571020059, KHÓA: 15

PHÂN TÍCH VÀ DỰ ĐOÁN TỶ LỆ THẤT NGHIỆP KẾT HỢP DỮ LIỆU KINH TẾ VÀ XÃ HỘI

BÁO CÁO ĐỒ ÁN TỐT NGHIỆP

NGÀNH: CÔNG NGHỆ THÔNG TIN

GIẢNG VIÊN HƯỚNG DẪN:

ThS. TRẦN THỊ THANH NHÀN

HÀ NỘI, NĂM 2025

LỜI CAM ĐOAN

Tôi xin cam đoan rằng đồ án tốt nghiệp với đề tài "Phân tích và dự đoán tỉ lệ thất nghiệp kết hợp dữ liệu kinh tế xã hội" là kết quả nghiên cứu, tìm hiểu và thực hiện của chính bản thân tôi trong suốt quá trình học tập và rèn luyện tại trường, đặc biệt là trong học kỳ thực hiện đồ án tốt nghiệp dưới sự hướng dẫn tận tình của cô Trần Thị Thanh Nhàn.

Toàn bộ nội dung trình bày trong đồ án đều do tôi tự thu thập, phân tích và biên soạn dựa trên cơ sở lý thuyết đã học, các tài liệu tham khảo có nguồn gốc rõ ràng và quá trình thực tế khảo sát, thực hiện. Tôi hoàn toàn không sao chép nội dung của bất kỳ cá nhân hay tổ chức nào khác một cách trái phép. Những trích dẫn, bảng biểu, hình ảnh, tài liệu tham khảo trong đồ án đều đã được tôi chú thích rõ ràng nguồn gốc, tuân thủ đúng quy định về đạo đức học thuật và bản quyền.

Tôi xin chịu hoàn toàn trách nhiệm trước Hội đồng chấm đồ án, Khoa, Nhà trường và pháp luật về tính trung thực, bản quyền và nội dung của đồ án này. Trong trường hợp phát hiện có bất kỳ hành vi gian lận, đạo văn hoặc vi phạm quy định, tôi xin cam kết chấp nhận mọi hình thức xử lý theo quy định hiện hành.

Tôi cũng xin cam kết rằng đồ án này chỉ được sử dụng vào mục đích học tập, nghiên cứu khoa học và phục vụ cho việc tốt nghiệp theo đúng quy định của Nhà trường. Mọi hình thức sử dụng khác như thương mại hóa, chuyển nhượng nội dung hoặc phát hành công khai mà không có sự đồng ý của tác giả và Nhà trường đều không được phép. Nếu có bất kỳ cá nhân hoặc tổ chức nào muốn sử dụng tài liệu này vào các mục đích khác, cần liên hệ và có sự đồng ý bằng văn bản của tôi cũng như sự chấp thuận của Nhà trường.

Họ và tên sinh viên

Nguyễn Hải Đăng

LỜI CẢM ƠN

Sau quá trình học tập tại trường, sinh viên được hệ thống lại toàn bộ lý thuyết chuyên ngành và có cơ hội tham gia kiến tập tại một số vị trí nghiệp vụ cơ bản liên quan đến những kiến thức đã được tiếp thu. Với sự cho phép của Khoa Công nghệ thông tin – Trường Đại học Đại Nam, cùng sự quan tâm, chỉ đạo và hướng dẫn tận tình từ các thầy cô, em đã tiến hành thực hiện đồ án tốt nghiệp của mình.

Mặc dù thời gian thực hiện đồ án không dài, nhưng quá trình này đã mang lại cho em nhiều kinh nghiệm quý báu, giúp em nâng cao kiến thức chuyên môn và kỹ năng thực tiễn. Tuy nhiên, do hạn chế về thời gian cũng như kinh nghiệm còn chưa nhiều, nội dung đồ án chắc chắn không tránh khỏi những thiếu sót và hạn chế nhất định. Đồ án này là kết quả của quá trình tổng hợp kiến thức, tìm hiểu và nghiên cứu về bài toán học máy, được xây dựng từ nền tảng học tập và rèn luyện trong suốt quá trình học. Em rất mong nhận được những ý kiến đóng góp quý báu từ quý thầy cô để có thể hoàn thiện hơn cả về nội dung đồ án lẫn bản thân trong quá trình học tập và làm việc sau này.

Em xin trân trọng gửi lời cảm ơn đến cô Trần Thị Thanh Nhàn – giảng viên Khoa Công nghệ thông tin, người đã tận tình hướng dẫn và tạo điều kiện thuận lợi để em hoàn thành tốt học phần đồ án tốt nghiệp.

Em xin chân thành cảm ơn!

LỜI NÓI ĐẦU

Trong bối cảnh toàn cầu hóa và chuyển đổi số đang diễn ra mạnh mẽ, nền kinh tế thế giới không ngừng biến động dưới tác động của nhiều yếu tố như đại dịch COVID-19, biến đổi khí hậu, xung đột chính trị - kinh tế giữa các quốc gia, và sự thay đổi không ngừng của công nghệ. Những yếu tố này đã và đang ảnh hưởng sâu sắc đến thị trường lao động, khiến tỷ lệ thất nghiệp trở thành một trong những vấn đề đáng quan tâm hàng đầu của các chính phủ và tổ chức quốc tế.

Trong thời đại công nghệ 4.0, dữ liệu không còn là yếu tố hỗ trợ mà đã trở thành trung tâm của mọi hoạt động nghiên cứu và ra quyết định. Việc áp dụng các phương pháp phân tích dữ liệu hiện đại và mô hình học máy (machine learning) giúp nâng cao độ chính xác trong việc đánh giá và dự đoán các hiện tượng kinh tế – xã hội. Trên cơ sở đó, đề tài "Phân tích và dự đoán tỷ lệ thất nghiệp kết hợp dữ liệu kinh tế và xã hội" được thực hiện với mục tiêu tận dụng sức mạnh của khoa học dữ liệu để khám phá sâu hơn về hiện trạng và xu hướng thất nghiệp.

Trong quá trình nghiên cứu, tôi đã khai thác và xử lý bộ dữ liệu toàn cầu về tỷ lệ thất nghiệp theo giới tính và nhóm tuổi, trải dài từ năm 2014 đến năm 2024. Bộ dữ liệu phản ánh thông tin chi tiết của hơn 100 quốc gia, phân chia theo độ tuổi lao động, nhóm giới tính, đồng thời bổ sung các yếu tố xã hội khác như trình độ học vấn, điều kiện kinh tế và dân số. Việc sử dụng dữ liệu thực tế và có tính cập nhật cao giúp đề tài đảm bảo tính khách quan và sát với thực tiễn.

Bên cạnh việc phân tích mô tả để tìm ra các xu hướng nổi bật và mối quan hệ giữa các yếu tố, tôi đã áp dụng một số mô hình dự báo như Hồi quy tuyến tính, Random Forest và XGBoost để dự đoán tỷ lệ thất nghiệp trong tương lai. Các kết quả thu được không chỉ giúp xác định các nhóm dân cư có nguy cơ thất nghiệp cao, mà còn hỗ trợ hoạch định chính sách lao động và giáo dục phù hợp hơn với bối cảnh kinh tế – xã hội từng quốc gia.

Đề tài là sự kết hợp giữa lý thuyết và ứng dụng thực tiễn, giữa tư duy phân tích và khả năng vận dụng công nghệ để giải quyết một vấn đề có ý nghĩa xã hội sâu sắc. Tôi hy vọng thông qua nghiên cứu này, có thể đóng góp một phần nhỏ vào việc nâng cao chất lượng dự báo và quản lý nguồn lao động, đồng thời khẳng định vai trò của dữ liệu và công nghệ trong việc giải quyết các bài toán kinh tế - xã hội hiện nay.

DANH MỤC VIẾT TẮT

STT	Từ viết tắt	Lý giải từ viết tắt
1	EVFTA	Hiệp định thương mại tự do giữa Liên minh châu Âu và Việt Nam
2	GDP	Tăng trưởng kinh tế
3	ML	Machine Learning
4	AI	Trí tuệ nhân tạo
5	ILO	Tổ chức lao động quốc tế
6	DL	Deep Learning
7	CPTPP	Hiệp định Đối tác Toàn diện và Tiến bộ xuyên Thái Bình Dương
8	FDI	Đầu tư trực tiếp từ nước ngoài
9	AR	Autoregressive
10	MA	Moving Average
11	ARIMA	Autoregressive Integrated Moving Average
12	SARIMA	Seasonal ARIMA
13	VAR	Vector Autoregression
14	VECM	Vector Error Correction Model
15	RNN	Mạng nơ-ron hồi tiếp
16	LSTM	Long Short-Term Memory
17	EDA	Explore Data Analysis
18	MAE	Mean Absolute Error
19	MSE	Mean Squared Error
20	RMSE	Root Mean Squared Error
21	R^2	Hệ số xác định
22	RNN	Recurrent Neural Network

23	MVC	(Model – View – Controller)
24	SHAP	(SHapley Additive exPlanations)

DANH MỤC HÌNH ẢNH

DANH MỤC BẢNG BIỂU

MỤC LỤC

CHƯƠNG 1. TỔNG QUAN VỀ THẤT NGHIỆP VÀ DỰ BÁO DỮ LIỆU KINH TẾ

1.1. Tổng quan về thất nghiệp

Thất nghiệp là một hiện tượng kinh tế – xã hội thể hiện tình trạng một bộ phận lực lượng lao động có khả năng và mong muốn làm việc nhưng không tìm được việc làm phù hợp trong một khoảng thời gian nhất định. Đây là một trong những chỉ số phản ánh rõ nét mức độ phát triển kinh tế và sự ổn định xã hội của một quốc gia.

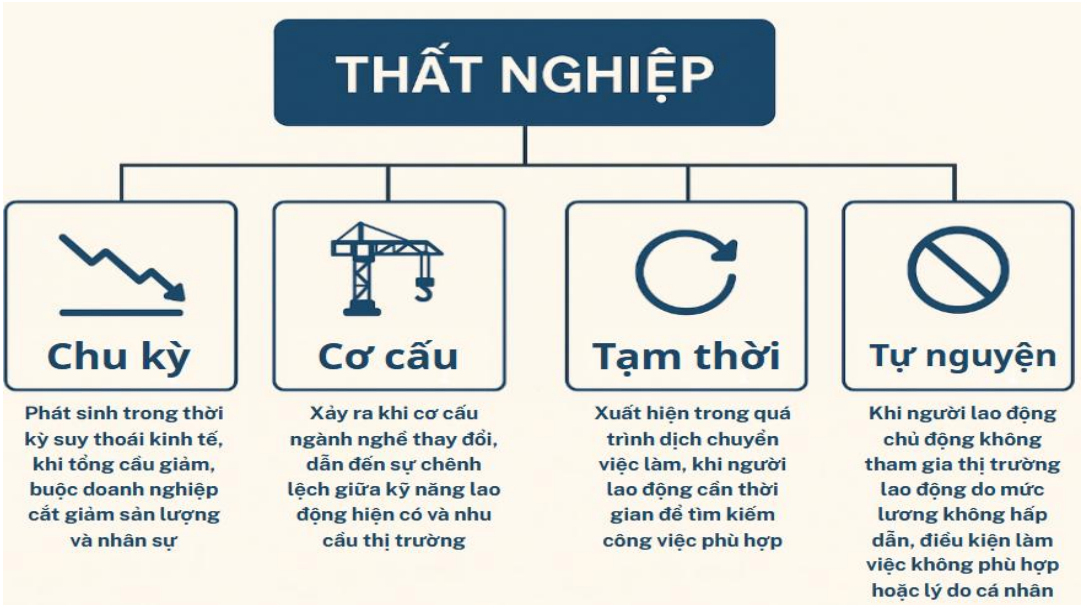
Theo Tổ chức Lao động Quốc tế (ILO, 2022), người thất nghiệp là người trong độ tuổi lao động, có khả năng làm việc, đang tích cực tìm kiếm việc làm nhưng vẫn chưa có việc làm trong thời gian khảo sát. Tỷ lệ thất nghiệp không chỉ phản ánh hiệu quả sử dụng nguồn nhân lực, mà còn liên quan chặt chẽ đến các yếu tố kinh tế vĩ mô như tăng trưởng GDP, lạm phát, thu nhập bình quân đầu người và mức sống dân cư (Okun, 1962; Phillips, 1958).

Bảng 1.1 Tỷ lệ thất nghiệp ASEAN (2023)

(Nguồn: TheGlobalEconomy.com)

Quốc gia	Tỷ lệ thất nghiệp (%)
Campuchia	0,24
Thái Lan	0,91
Lào	1,18
Việt Nam	1,6
Philippines	2,23
Singapore	3,47
Indonesia	3,42
Malaysia	3,86
Brunei	5,27

- Nguyên nhân của thất nghiệp có thể phân loại thành bốn nhóm chính:



Hình 1.1 Các nhóm nguyên nhân dẫn đến tình trang thất nghiệp

Tỷ lệ thất nghiệp cao có thể dẫn đến nhiều hệ lụy tiêu cực như: giảm thu nhập của người dân, gia tăng tệ nạn xã hội, tăng áp lực cho hệ thống an sinh xã hội và làm suy giảm niềm tin của nhà đầu tư. Vì vậy, việc nghiên cứu, phân tích và dự báo tình hình thất nghiệp đóng vai trò quan trọng trong hoạch định chính sách phát triển kinh tế – xã hội, cũng như trong quản lý lao động và điều tiết thị trường việc làm.

Trong bối cảnh kinh tế toàn cầu liên tục biến động bởi ảnh hưởng của toàn cầu hóa, cách mạng công nghiệp 4.0, dịch bệnh, và xung đột địa chính trị, việc xây dựng các mô hình dự báo thất nghiệp ngày càng trở nên cần thiết nhằm hỗ trợ chính phủ và doanh nghiệp đưa ra các quyết sách kịp thời, hiệu quả (ILO, 2022; World Bank, 2023).

1.2. Các yếu tố ảnh hưởng đến tỉ lệ thất nghiệp

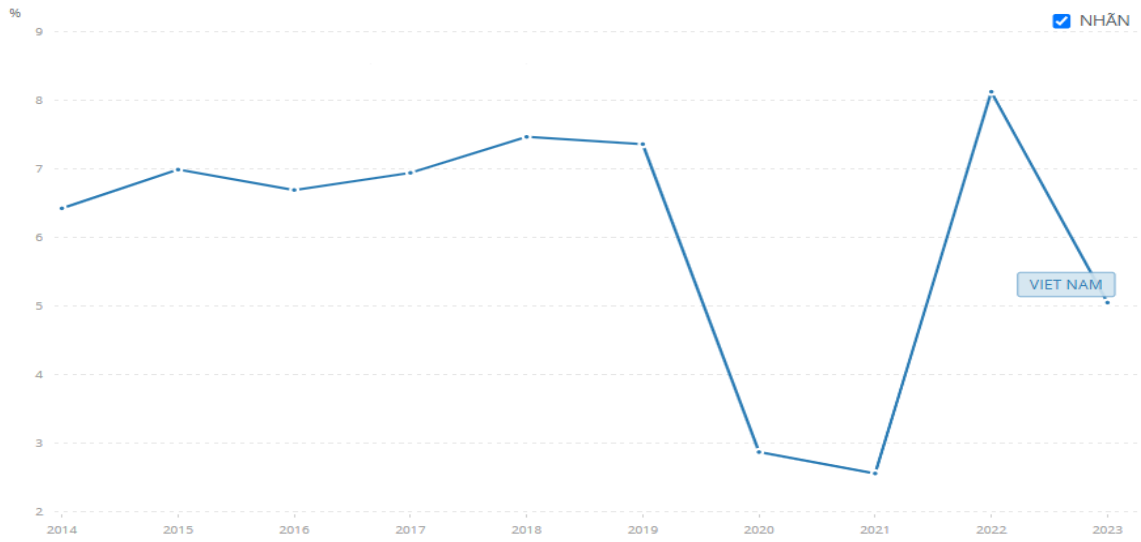
Tỷ lệ thất nghiệp là một chỉ số kinh tế tổng hợp phản ánh mức độ hiệu quả trong việc sử dụng nguồn lao động của một quốc gia. Tình trạng thất nghiệp không chỉ chịu ảnh hưởng từ các yếu tố nội tại của thị trường lao động mà còn liên quan chặt chẽ đến bối cảnh kinh tế vĩ mô, chính sách nhà nước và các biến động toàn cầu. Việc hiểu rõ các yếu tố ảnh hưởng đến tỷ lệ thất nghiệp sẽ giúp các nhà hoạch định

chính sách, doanh nghiệp và người lao động đưa ra các quyết định phù hợp trong việc quản lý và phát triển nguồn nhân lực.

1.2.1. Tăng trưởng kinh tế

Tăng trưởng kinh tế là một trong những yếu tố quan trọng nhất tác động đến tỷ lệ thất nghiệp. Khi nền kinh tế tăng trưởng, tổng cầu về hàng hóa và dịch vụ gia tăng, từ đó thúc đẩy hoạt động sản xuất và kinh doanh, kéo theo nhu cầu tuyển dụng lao động tăng lên.

Tại Việt Nam, năm 2022, GDP tăng trưởng 8,02% – mức cao nhất trong hơn một thập kỷ – giúp tỷ lệ thất nghiệp giảm xuống còn khoảng 2,3% so với 3,1% trong năm 2021 (Tổng cục Thống kê). Ngược lại, vào năm 2020 khi GDP chỉ tăng 2,91% do ảnh hưởng của đại dịch COVID-19, tỷ lệ thất nghiệp tăng đáng kể, đặc biệt ở khu vực đô thị.

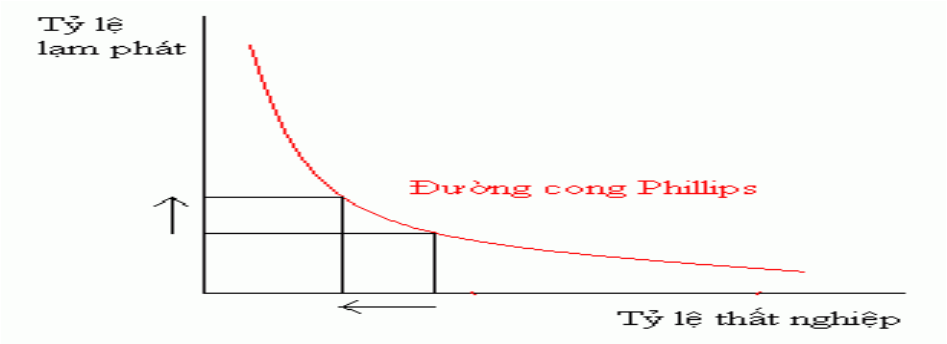


Hình 1.2 Tốc độ tăng trưởng GDP Việt Nam 2014-2023

(Nguồn: Data.WordBank)

Định luật Okun chỉ ra rằng khi GDP thực tế tăng trưởng chậm hơn mức tiềm năng, tỷ lệ thất nghiệp có xu hướng tăng. Tuy nhiên, mức độ ảnh hưởng cụ thể còn phụ thuộc vào cơ cấu nền kinh tế và khả năng hấp thụ lao động của từng lĩnh vực.

1.2.2. Lạm phát



Hình 1.3 Mối quan hệ giữa lạm phát và thất nghiệp

Nguồn: Wikipedia

Lạm phát, đặc biệt là trong ngắn hạn, có mối quan hệ ngược chiều với thất nghiệp theo lý thuyết đường cong Phillips. Khi lạm phát tăng, thu nhập danh nghĩa tăng, từ đó kích thích tiêu dùng và mở rộng sản xuất, giúp giảm thất nghiệp.

Tuy nhiên, nếu lạm phát vượt tầm kiểm soát, đặc biệt trong bối cảnh chi phí nguyên liệu đầu vào tăng mạnh, doanh nghiệp có xu hướng thắt chặt tuyển dụng để giảm chi phí. Tại Việt Nam, trong năm 2023, áp lực lạm phát tăng do giá xăng dầu và nguyên vật liệu đầu vào tăng cao, đã ảnh hưởng đến các doanh nghiệp sản xuất và gia công, dẫn đến cắt giảm lao động ở một số ngành như dệt may, da giày.

1.2.3. Chính sách tiền tệ và tài khóa

Chính sách tiền tệ và chính sách tài khóa là hai công cụ điều tiết vĩ mô quan trọng nhằm ổn định kinh tế và kiểm soát tỷ lệ thất nghiệp, tuy nhiên có phạm vi và cách thức tác động khác nhau.

Chính sách tiền tệ, do Ngân hàng Nhà nước quản lý, điều chỉnh lượng cung tiền và lãi suất. Khi nền kinh tế suy thoái, chính sách tiền tệ nới lỏng (giảm lãi suất, giảm dự trữ bắt buộc, bơm tiền) giúp doanh nghiệp tiếp cận vốn dễ dàng, mở rộng sản xuất và tạo việc làm. Ngược lại, chính sách thắt chặt làm tăng chi phí vay vốn, hạn chế đầu tư, có thể làm thất nghiệp tăng trong ngắn hạn.

Chính sách tài khóa, do Chính phủ thực hiện qua việc điều chỉnh thuế và chi tiêu công, tác động trực tiếp đến tổng cầu. Chính sách tài khóa mở rộng (tăng chi

đầu tư công, giảm thuế, hỗ trợ doanh nghiệp, người lao động) kích thích sản xuất và việc làm, trong khi chính sách thắt chặt có thể làm giảm tổng cầu và làm thất nghiệp gia tăng.

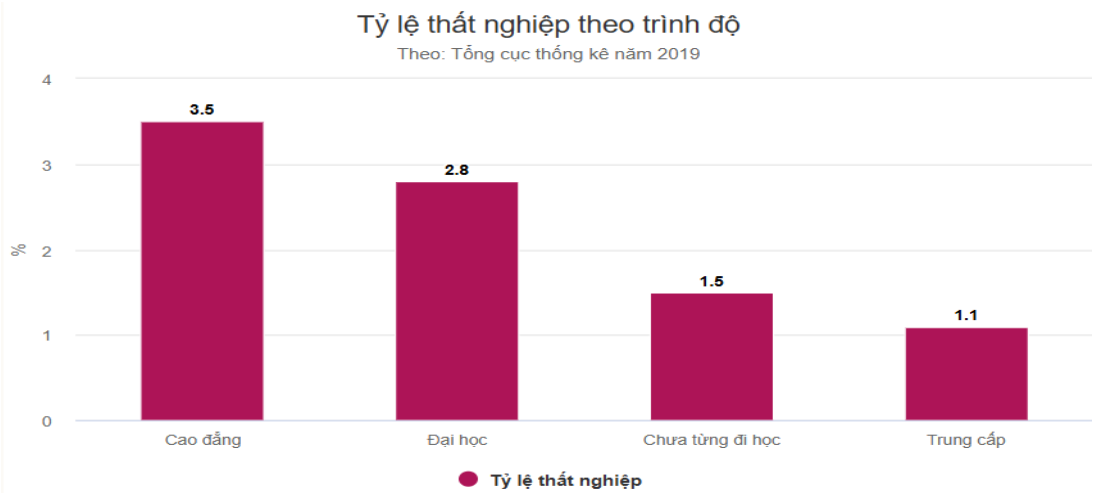
Trong giai đoạn 2020–2022, nhằm ứng phó với tác động tiêu cực của đại dịch COVID-19 đến thị trường lao động và doanh nghiệp, Việt Nam đã triển khai đồng bộ cả hai loại chính sách:

- Về tài khóa, Chính phủ đã ban hành gói hỗ trợ 62.000 tỷ đồng vào năm 2020, tập trung hỗ trợ người lao động và hộ kinh doanh bị ảnh hưởng, và gói phục hồi kinh tế 350.000 tỷ đồng vào năm 2022, bao gồm đầu tư công, hỗ trợ lãi suất và các chính sách an sinh xã hội.
- Về tiền tệ, Ngân hàng Nhà nước đã nhiều lần điều chỉnh giảm lãi suất điều hành trong giai đoạn 2020–2021 nhằm thúc đẩy tín dụng, hỗ trợ doanh nghiệp vượt qua khó khăn, duy trì sản xuất và việc làm.

Những biện pháp đồng bộ nêu trên đã góp phần hạn chế sự gia tăng tỷ lệ thất nghiệp và tạo nền tảng phục hồi cho thị trường lao động trong bối cảnh bất ổn kinh tế toàn cầu.

1.2.4. Trình độ và cơ cấu lao động

Trình độ chuyên môn và cơ cấu ngành nghề của lực lượng lao động ảnh hưởng mạnh mẽ đến khả năng tìm kiếm việc làm. Tại Việt Nam, tỷ lệ thất nghiệp ở nhóm lao động chưa qua đào tạo vẫn cao hơn đáng kể so với nhóm có trình độ đại học trở lên. Theo Tổng cục Thống kê, năm 2023, tỷ lệ thất nghiệp trong nhóm lao động không có bằng cấp và chứng chỉ lên tới 3,5%, trong khi nhóm có trình độ đại học chỉ khoảng 1,8%.



Hình 1.4 Tỉ lệ thất nghiệp theo trình độ

Bên cạnh đó, sự không đồng bộ giữa đào tạo và nhu cầu của thị trường dẫn đến tình trạng thất nghiệp cơ cấu, đặc biệt trong nhóm sinh viên mới ra trường. Nhiều sinh viên thiếu kỹ năng mềm, ngoại ngữ hoặc kinh nghiệm thực tiễn, khó đáp ứng được yêu cầu tuyển dụng.

1.2.5. Tiến bộ công nghệ và tự động hóa

Sự phát triển nhanh chóng của công nghệ, đặc biệt là trí tuệ nhân tạo, máy học, và tự động hóa robot đã và đang thay đổi mạnh mẽ cơ cấu nghề nghiệp trên toàn cầu, trong đó có Việt Nam. Nhiều doanh nghiệp sản xuất tại các khu công nghiệp đã ứng dụng các dây chuyền sản xuất tự động, robot hóa và các hệ thống điều khiển thông minh nhằm thay thế lao động phổ thông, nâng cao năng suất và giảm chi phí.

Theo Báo cáo của Tổ chức Lao động Quốc tế, khoảng 70% công việc trong ngành chế biến – chế tạo tại Việt Nam có nguy cơ bị thay thế bởi công nghệ tự động và trí tuệ nhân tạo trong vòng 15 năm tới. Điều này đặt ra yêu cầu cấp thiết cho người lao động phải liên tục học tập, nâng cao trình độ và chuyển đổi kỹ năng để thích nghi với thị trường lao động mới, hướng đến các ngành nghề đòi hỏi kỹ năng cao hơn và khả năng vận hành, quản lý các công nghệ mới.

Song song với việc tự động hóa, các công nghệ Big Data và AI đang được sử dụng rộng rãi trong phân tích thị trường lao động và dự báo tỷ lệ thất nghiệp. Các mô hình Machine Learning như hồi quy tuyến tính, mạng nơ-ron nhân tạo (Artificial

Neural Networks), cây quyết định (Decision Trees), và các thuật toán học sâu (Deep Learning) được áp dụng để phân tích dữ liệu lớn về các yếu tố kinh tế - xã hội, dữ liệu tuyển dụng, biến động thị trường lao động, cũng như tác động của tự động hóa đến từng ngành nghề cụ thể.

Việc khai thác Big Data từ các nguồn dữ liệu đa dạng như mạng xã hội, sàn giao dịch việc làm trực tuyến, báo cáo kinh tế vĩ mô và dữ liệu điều tra dân số giúp các nhà nghiên cứu xây dựng các mô hình dự báo chính xác hơn, hỗ trợ hoạch định chính sách lao động hiệu quả và phát hiện sớm các xu hướng thay đổi trong thị trường lao động.

Ngoài ra, công nghệ cũng mở ra nhiều lĩnh vực mới như thương mại điện tử, logistics, công nghệ thông tin, thiết kế số… góp phần tạo ra nhiều cơ hội việc làm chất lượng cao cho lực lượng lao động trẻ, linh hoạt và có khả năng thích ứng nhanh với sự chuyển đổi số. Việc ứng dụng AI không chỉ làm thay đổi cơ cấu việc làm mà còn thúc đẩy sự phát triển của các ngành nghề sáng tạo và dịch vụ công nghệ cao, góp phần nâng cao chất lượng nguồn nhân lực và tăng trưởng kinh tế bền vững.

1.2.6. Toàn cầu hóa và hội nhập kinh tế

Toàn cầu hóa mang lại nhiều cơ hội tiếp cận thị trường và thu hút đầu tư nước ngoài. Việt Nam là một trong những quốc gia hưởng lợi lớn từ dòng vốn FDI nhờ các hiệp định thương mại tự do như EVFTA, CPTPP. Điều này giúp tạo ra hàng triệu việc làm mới, đặc biệt trong các lĩnh vực sản xuất, dịch vụ và công nghệ cao.

Tuy nhiên, áp lực cạnh tranh quốc tế cũng khiến các doanh nghiệp phải liên tục tái cấu trúc, tinh gọn bộ máy, hoặc chuyển dịch nhà máy sang khu vực có chi phí thấp hơn. Điều này có thể làm tăng tỷ lệ thất nghiệp ở những khu vực hoặc ngành nghề không còn lợi thế cạnh tranh.

Trong năm 2023, một số doanh nghiệp gia công xuất khẩu trong ngành dệt may và điện tử ở phía Nam đã cắt giảm lao động do đơn hàng giảm mạnh từ thị trường Mỹ và EU.

1.2.7. Yếu tố bất ổn và các cú sốc bên ngoài

Các yếu tố như dịch bệnh, khủng hoảng địa chính trị, biến đổi khí hậu… đều có thể gây ra các cú sốc lớn đối với thị trường lao động.

Tại Việt Nam, đại dịch COVID-19 đã ảnh hưởng nghiêm trọng đến việc làm: hơn 32,1 triệu lao động bị ảnh hưởng (giảm thu nhập, nghỉ việc, giãn việc) trong năm 2020. Trong đó, các ngành như du lịch, dịch vụ lưu trú – ăn uống, vận tải – kho bãi bị ảnh hưởng nặng nề nhất.

Ngoài ra, biến đổi khí hậu và thiên tai như hạn hán, xâm nhập mặn ở đồng bằng sông Cửu Long cũng khiến hàng trăm nghìn lao động nông nghiệp phải chuyển đổi nghề nghiệp, trong khi các lựa chọn thay thế còn hạn chế.

1.3. Các phương pháp dự báo dữ liệu kinh tế

Dự báo dữ liệu kinh tế là một phần quan trọng trong phân tích và hoạch định chính sách vĩ mô, vi mô. Nó cho phép các nhà quản lý, nhà đầu tư và chính phủ đưa ra các quyết định chiến lược dựa trên các xu hướng dự đoán của các biến kinh tế như GDP, lạm phát, lãi suất, tỷ lệ thất nghiệp… Mỗi biến số kinh tế đều mang tính phức tạp, thường chịu ảnh hưởng bởi nhiều yếu tố cùng lúc, trong đó có cả yếu tố thời gian và phi tuyến tính. Do đó, việc lựa chọn phương pháp dự báo phù hợp là điều kiện tiên quyết để đảm bảo độ tin cậy và chính xác của kết quả.

Tùy theo mục tiêu phân tích, đặc điểm của dữ liệu, cũng như nguồn lực kỹ thuật, người ta có thể sử dụng các nhóm phương pháp sau đây:

1.3.1. Phương pháp định tính

Phương pháp định tính là những kỹ thuật dự báo không phụ thuộc vào mô hình toán học mà dựa chủ yếu vào phán đoán của chuyên gia, kinh nghiệm thực tiễn, hoặc các phân tích kịch bản. Phương pháp này thường được sử dụng trong bối cảnh thiếu dữ liệu lịch sử, hoặc khi thị trường chịu ảnh hưởng mạnh bởi các yếu tố bất ổn như biến động chính trị, đại dịch, xung đột toàn cầu…

Các kỹ thuật định tính phổ biến bao gồm:

- Phỏng vấn chuyên gia: Thu thập ý kiến từ các nhà kinh tế, lãnh đạo doanh nghiệp hoặc chuyên gia ngành để đưa ra nhận định về xu hướng sắp tới.
- Phân tích Delphi: Là một phương pháp dự báo nhóm, trong đó các chuyên gia trả lời bảng khảo sát theo nhiều vòng, và được cung cấp thông tin phản hồi trung gian nhằm đạt được sự đồng thuận.
- Phân tích kịch bản: Xây dựng các tình huống tương lai khác nhau (kịch bản lạc quan, trung bình, bi quan) dựa trên giả định về các biến động kinh tế - xã hội.

Mặc dù không mang tính định lượng chính xác, các phương pháp định tính vẫn rất hữu ích trong việc định hướng chiến lược và phân tích rủi ro ở những lĩnh vực nhiều bất định. Tuy nhiên, phương pháp định tính phụ thuộc nhiều vào ý kiến cá nhân, dễ bị thiên lệch, không đưa ra được các kết quả định lượng cụ thể. Ngoài ra khó kiểm chứng độ chính xác và tái lập kết quả và hiệu quả phụ thuộc vào trình độ và kinh nghiệm của người tham gia đánh giá.

1.3.2. Phương pháp định lượng truyền thống

Các phương pháp định lượng truyền thống sử dụng nền tảng thống kê và toán học nhằm xác định mối quan hệ giữa các biến số, từ đó đưa ra mô hình dự báo. Nhóm phương pháp này chiếm vai trò trung tâm trong kinh tế lượng và phân tích chuỗi thời gian.

- Mô hình hồi quy tuyến tính

Hồi quy tuyến tính (Linear Regression) là mô hình nền tảng trong kinh tế lượng. Nó mô tả mối quan hệ giữa một biến phụ thuộc (Y) và một hoặc nhiều biến độc lập (X). Tỷ lệ thất nghiệp có thể được mô hình hóa là hàm của GDP, CPI (chỉ số giá tiêu dùng), và lãi suất.

Công thức tổng quát:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon$$

Trong đó:

1. β_0 là hệ số chặn (intercept)

2. β_1, β_2, … là các hệ số hồi quy
3. ε là sai số ngẫu nhiên

Ưu điểm của hồi quy tuyến tính là dễ hiểu, dễ ước lượng và giải thích. Tuy nhiên, mô hình này giả định mối quan hệ tuyến tính giữa các biến – điều không phải lúc nào cũng đúng trong thực tế kinh tế.

- Mô hình chuỗi thời gian

Bảng 1.2 Mô hình chuỗi thời gian

Mô hình	Đặc điểm chính	Áp dụng cho	Loại chuỗi	Sử dụng điển hình
AR	Phụ thuộc vào chính các giá trị quá khứ của chuỗi	Dữ liệu có tính tự tương quan	Dừng	Phân tích tài chính, kinh tế
MA	Phụ thuộc vào trung bình có trọng số của các sai số trong quá khứ	Dữ liệu có nhiễu trắng rõ ràng	Dừng	Khử nhiễu, dự báo ngắn hạn
ARMA	Kết hợp AR và MA	Chuỗi dừng không có xu hướng rõ rệt	Dừng	Mô hình hóa dữ liệu kinh tế ổn định
ARIMA	Thêm phép vi phân để xử lý dữ liệu không dừng	Chuỗi có xu hướng hoặc không ổn định	Không dừng	Tỷ lệ thất nghiệp, giá tiêu dùng, ...
SARIMA	Mở rộng ARIMA với thành phần mùa vụ (theo quý, tháng, năm)	Dữ liệu có tính mùa vụ rõ rệt	Không dừng & có mùa vụ	Tỷ lệ thất nghiệp theo mùa, doanh thu quý, ...

Mô hình	Đặc điểm chính	Áp dụng cho	Loại chuỗi	Sử dụng điển hình
VAR	Mô hình đa biến, các biến phụ thuộc lẫn nhau	Tập hợp các biến kinh tế vĩ mô	Dừng	GDP, lạm phát, thất nghiệp, ...
VECM	Mô hình VAR cho chuỗi không dừng nhưng có quan hệ đồng liên kết dài hạn	Các chuỗi có mối quan hệ cân bằng dài hạn	Không dừng & đồng liên kết	Chính sách tiền tệ, mô hình kinh tế học thuật

1.3.3. Phương pháp hiện đại: Học máy và AI

Trong bối cảnh dữ liệu lớn và sự phát triển mạnh mẽ của trí tuệ nhân tạo, các mô hình học máy và học sâu đã trở thành xu hướng mới trong lĩnh vực dự báo kinh tế.

- Học máy:

Các thuật toán phổ biến bao gồm:

+ Random Forest: Dựa trên tập hợp nhiều cây quyết định, có khả năng xử lý tốt dữ liệu phi tuyến và tránh overfitting.

+ Gradient Boosting (XGBoost, LightGBM): Các mô hình boosting cải thiện hiệu suất dự báo thông qua việc học từ các lỗi trước đó.

+ Support Vector Regression: Tối ưu hóa hàm mất mát với một mức dung sai cho phép.

+ K-Nearest Neighbors: Dự báo giá trị dựa trên các điểm dữ liệu gần nhất.

Ưu điểm: có khả năng dự báo chính xác cao, đặc biệt khi dữ liệu lớn và phức tạp. Tuy nhiên, mô hình thường cần điều chỉnh tham số và đánh giá kỹ càng để tránh sai lệch.

- Học sâu

+ Mạng nơ-ron nhân tạo: Mô phỏng mạng thần kinh sinh học, giúp nhận diện các mẫu phức tạp trong dữ liệu.

+ LSTM: Một biến thể của RNN, được thiết kế đặc biệt cho dữ liệu chuỗi thời gian, có khả năng ghi nhớ thông tin dài hạn và giảm hiện tượng mất thông tin.

+ Transformer: Được ứng dụng nhiều trong xử lý ngôn ngữ tự nhiên, gần đây cũng được áp dụng vào dữ báo kinh tế nhờ khả năng ghi nhớ và học từ toàn bộ chuỗi.

Mặc dù các mô hình học sâu yêu cầu tài nguyên tính toán lớn, nhưng chúng đang dần trở thành công cụ mạnh mẽ trong các hệ thống dự báo tiên tiến.

- Lý do chọn phương pháp dự báo:

Trong quá trình xây dựng ứng dụng dự báo tỷ lệ thất nghiệp, tôi đã cân nhắc nhiều phương pháp dự báo khác nhau, từ các mô hình thống kê truyền thống đến các thuật toán học máy hiện đại. Tuy nhiên, sau khi phân tích kỹ lưỡng đặc điểm của dữ liệu và mục tiêu nghiên cứu, tôi quyết định lựa chọn mô hình học sâu LSTM làm phương pháp chính để triển khai hệ thống dự báo.

Tỷ lệ thất nghiệp là một chỉ số kinh tế có tính chất chuỗi thời gian rõ rệt, thường bị ảnh hưởng bởi nhiều yếu tố có tác động lặp lại theo chu kỳ (như tăng trưởng kinh tế, biến động thị trường lao động, tác động chính sách...). Vì vậy, mô hình dự báo cần có khả năng xử lý tốt dữ liệu thời gian, nhận biết xu hướng, chu kỳ và các mối quan hệ phụ thuộc trong dài hạn. LSTM là một biến thể đặc biệt của mạng nơ-ron hồi tiếp, được thiết kế để giải quyết các vấn đề về ghi nhớ thông tin dài hạn mà các mô hình truyền thống không xử lý tốt.

- Một số lý do lựa chọn LSTM:

+ Khả năng ghi nhớ dài hạn: Mô hình LSTM được thiết kế với các cổng nhớ, cho phép lưu trữ và khai thác thông tin từ các bước thời gian trước đó. Điều này đặc biệt phù hợp với dữ liệu kinh tế có tính phụ thuộc theo thời gian dài, như tỷ lệ thất nghiệp, lạm phát hoặc GDP.

+ Mô hình hóa quan hệ phi tuyến: LSTM có khả năng học các mối quan hệ phi tuyến giữa các biến đầu vào và đầu ra. Trong bối cảnh dữ liệu kinh tế, nơi nhiều yếu tố kinh tế - xã hội tương tác theo cách phi tuyến, khả năng này giúp mô hình nâng cao độ chính xác trong dự báo.

+ Hiệu quả vượt trội trong thực nghiệm: Nhiều nghiên cứu đã chứng minh rằng LSTM thường cho kết quả tốt hơn so với các mô hình truyền thống như ARIMA hoặc các mô hình học máy cơ bản khi áp dụng cho các bài toán dự báo biến động kinh tế như giá cả, tỷ lệ thất nghiệp hoặc doanh thu.

+ Tính linh hoạt và khả năng cập nhật: Mô hình LSTM có thể dễ dàng cập nhật khi có dữ liệu mới thông qua việc tái huấn luyện hoặc fine-tuning. Điều này rất quan trọng trong bối cảnh kinh tế liên tục biến động và yêu cầu mô hình phải thích ứng nhanh với các thay đổi.

Việc lựa chọn LSTM làm phương pháp chính không chỉ nhằm tối ưu hóa độ chính xác của dự báo mà còn mang lại tính linh hoạt và khả năng thích nghi cao cho ứng dụng. Mô hình này giúp tôi xây dựng được một hệ thống dự báo có khả năng học và cập nhật liên tục, hỗ trợ tốt cho việc đưa ra các quyết định quản lý lao động, hoạch định chính sách và cảnh báo sớm các rủi ro trong thị trường lao động.

CHƯƠNG 2. XÂY DỰNG MÔ HÌNH AI DỰ ĐOÁN TỈ LỆ THẤT NGHIỆP

2.1. Quy trình xây dựng mô hình AI

2.1.1. Tiền xử lý dữ liệu

Tiền xử lý dữ liệu là bước đầu tiên và vô cùng quan trọng trong quy trình phân tích và xây dựng mô hình học máy, nhằm đảm bảo chất lượng, tính toàn vẹn và khả năng khai thác của dữ liệu. Trong nghiên cứu này, quá trình tiền xử lý dữ liệu được thực hiện qua các bước chính như sau:

- Chuyển đổi định dạng dữ liệu (Wide → Long Format)

```
# Chuyển từ wide -> long format
df_long = df.melt(
    id_vars=["country_name", "sex", "age_group", "age_categories"],
    value_vars=[str(year) for year in range(2014, 2025)],
    var_name="year",
    value_name="unemployment_rate"
)
df_long["year"] = df_long["year"].astype(int)
```

Hình 2.1 Chuyển đổi định dạng

Dữ liệu gốc được thu thập ở định dạng "wide format", trong đó mỗi cột đại diện cho một năm từ 2014 đến 2024. Mặc dù định dạng này có thể thuận tiện để hiển thị, nhưng lại gây khó khăn khi áp dụng các mô hình phân tích theo chuỗi thời gian hoặc khai thác sự thay đổi của tỷ lệ thất nghiệp theo năm.

Do đó, dữ liệu được chuyển đổi sang "long format" bằng phương pháp melt trong thư viện pandas. Việc này giúp gom các giá trị theo năm vào một cột duy nhất year, với các giá trị tương ứng là "unemployment_rate", đồng thời giữ lại các cột định danh như "country_name", "sex", "age_group", và "age_categories". Việc chuyển đổi này giúp dữ liệu trở nên trực quan hơn và thuận tiện cho việc nhóm, lọc, và trực quan hóa trong các bước phân tích tiếp theo.

- Xử lý dữ liệu thiếu và loại bỏ giá trị ngoại lệ (Outliers)

* Lý do chọn phương pháp IQR để xử lý ngoại lệ

+ Không phụ thuộc phân phối chuẩn: IQR sử dụng các phân vị (Q1, Q3) nên hoạt động tốt với dữ liệu không phân phối chuẩn, phổ biến trong dữ liệu kinh tế như tỷ lệ thất nghiệp.

+ Ít bị ảnh hưởng bởi ngoại lệ: Không giống Z-score (dựa vào trung bình và độ lệch chuẩn), IQR không bị chi phối bởi các giá trị cực đoan, nên phát hiện ngoại lệ hiệu quả hơn.

+ Mục tiêu phù hợp: RobustScaler dùng IQR để chuẩn hóa dữ liệu, trong khi mục tiêu ở đây là phát hiện và loại bỏ ngoại lệ. Do đó, dùng trực tiếp IQR là hợp lý hơn.

+ Thực nghiệm hiệu quả: IQR cho kết quả xử lý dữ liệu tốt hơn trong thực nghiệm, giúp tăng độ chính xác của mô hình dự báo sau đó.

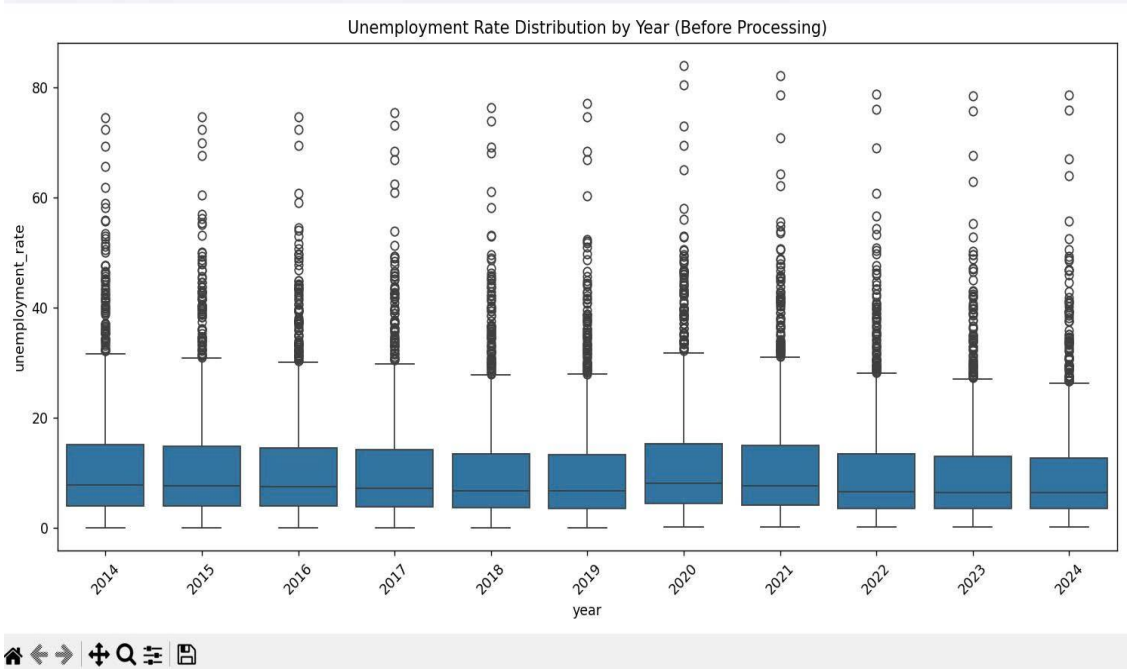
```
# Loại bỏ dữ liệu thiếu và xử lý outliers
df_model = df_long.dropna(subset=["unemployment_rate"]).copy()
Q1 = df_model["unemployment_rate"].quantile(0.25)
Q3 = df_model["unemployment_rate"].quantile(0.75)
IQR = Q3 - Q1
df_model = df_model[
    (df_model["unemployment_rate"] >= Q1 - 1.5 * IQR) &
    (df_model["unemployment_rate"] <= Q3 + 1.5 * IQR)
]
```

Hình 2.2 Xử lý giá trị thiếu và ngoại lệ

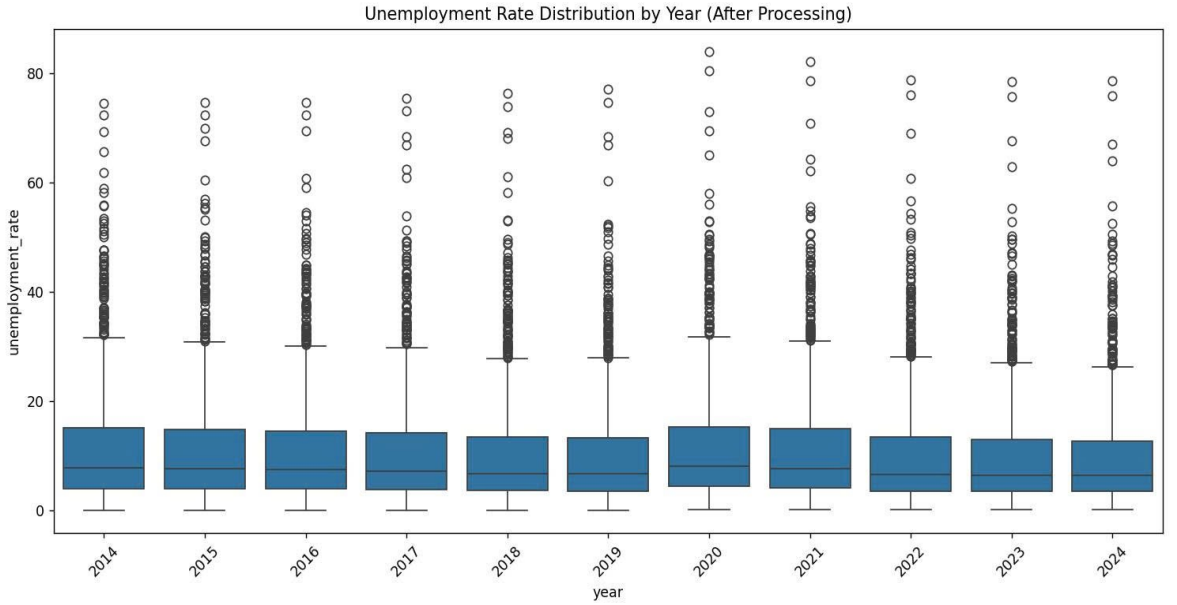
Sau bước chuyển đổi định dạng, dữ liệu tiếp tục được làm sạch bằng cách loại bỏ các dòng bị thiếu giá trị (NaN) trong cột “unemployment_rate”. Việc này đảm bảo rằng các mô hình học máy sẽ không bị ảnh hưởng bởi những thông tin không đầy đủ, vốn có thể gây sai lệch hoặc làm giảm độ chính xác của mô hình.

Tiếp theo, để xử lý các giá trị ngoại lệ – vốn có thể là kết quả của sai sót trong quá trình thu thập hoặc những biến động bất thường – phương pháp “Interquartile Range (IQR)” được sử dụng. Đây là một kỹ thuật phổ biến và hiệu quả để xác định và loại bỏ các giá trị nằm quá xa phạm vi phân bố chuẩn. Cụ thể, các giá trị tỷ lệ thất nghiệp nằm ngoài khoảng từ “Q1 - 1.5 * IQR” đến “Q3 + 1.5 * IQR” sẽ bị loại bỏ.

Việc này giúp mô hình tránh bị chi phối bởi những điểm dữ liệu không đại diện cho xu hướng chung.



Hình 2.3 Boxplot hiển thị dữ liệu trước khi xử lý



Hình 2.4 Boxplot hiển thị dữ liệu sau khi xử lý

- Sau khi xử lý dữ liệu thiếu và ngoại lệ ta có thể thấy:

Bảng 2.1 Đánh giá sau khi xử lý

Tiêu chí	**Trước xử lý**	**Sau xử lý**

Độ phân tán	Cao	Giảm đáng kể
Outliers	Nhiều	Giảm bớt
Độ ổn định giữa các năm	Kém	Ổn định hơn
Tính sẵn sàng cho phân tích	Thấp	Cao hơn (đã làm sạch dữ liệu)

- Mã hóa các biến phân loại (Label Encoding)

Trong quá trình tiền xử lý dữ liệu, các biến phân loại như “country_name”, “sex” và “age_group” được mã hóa bằng phương pháp Label Encoding. Phương pháp này chuyển các giá trị phân loại thành các số nguyên liên tiếp, "Male" → 0, "Female" → 1.

* Lý do không sử dụng One-Hot Encoding

One-Hot Encoding là một lựa chọn phổ biến, đặc biệt hiệu quả với các mô hình tuyến tính, nó không được áp dụng trong trường hợp này vì:

- Số lượng nhãn lớn: Với biến “country_name”, số lượng quốc gia có thể lên đến hàng trăm. Việc áp dụng One-Hot Encoding sẽ dẫn đến bùng nổ số chiều, gây tốn tài nguyên tính toán và có thể làm giảm hiệu quả của mô hình.
- Hiệu quả với mô hình cây: Các mô hình như Random Forest và XGBoost không bị ảnh hưởng bởi thứ tự số của nhãn được mã hóa bằng Label Encoding. Các mô hình này phân chia dữ liệu dựa trên giá trị tại từng cột mà không giả định mối quan hệ tuyến tính giữa các giá trị, do đó Label Encoding là phương án phù hợp hơn trong trường hợp này.
- Tuy nhiên Việc sử dụng Label Encoding có thể gây sai lệch khi áp dụng với các mô hình tuyến tính như Linear Regression. Nguyên nhân là do mô hình tuyến tính có thể hiểu các giá trị được mã hóa như một chuỗi có thứ tự tuyến tính, như "Male" → 0, "Female" → 1 có thể bị hiểu nhầm là "Female > Male", hoặc "15-24" → 0, "55-64" → 4 có thể bị coi là một thang đo liên tục, dẫn đến kết quả sai lệch.

- Trong trường hợp mà mô hình tuyến tính đóng vai trò quan trọng, One-Hot Encoding là lựa chọn an toàn hơn để đảm bảo rằng mô hình không bị dẫn dắt bởi các giả định sai lệch về thứ tự.

Việc lựa chọn Label Encoding trong dự án này nhằm tối ưu hóa hiệu suất tính toán và phù hợp với các mô hình như Random Forest và XGBoost. Tuy nhiên, cần lưu ý Label Encoding có thể gây ra sai lệch khi áp dụng cho các mô hình tuyến tính.

```
# Mã hóa các biến phân loại
categorical_cols = ["country_name", "sex", "age_group"]
for col in categorical_cols:
    le = LabelEncoder()
    df_long[col] = le.fit_transform(df_long[col])
    self.label_encoders[col] = le
```

Hình 2.5 Mã hóa các biến phân loại

Để các mô hình học máy có thể xử lý được dữ liệu dạng văn bản (chuỗi), các biến phân loại như “country_name”, “sex”, và “age_group” cần được chuyển đổi sang dạng số. Phương pháp “Label Encoding” được lựa chọn trong trường hợp này vì dữ liệu không có thứ tự rõ ràng và Label Encoding giúp giữ cho số chiều của dữ liệu ở mức thấp hơn so với One-Hot Encoding.

Mỗi giá trị phân loại sẽ được gán một số nguyên duy nhất. Các quốc gia sẽ được đánh số từ 0, 1, 2, … theo thứ tự xuất hiện trong dữ liệu. Quá trình mã hóa này được thực hiện đồng thời lưu lại các bộ mã để phục vụ cho bước giải mã kết quả và tái sử dụng mô hình sau này.

- Chuẩn hóa dữ liệu (Feature Scaling)

Một vấn đề thường gặp trong các mô hình học máy là sự chênh lệch về đơn vị hoặc phạm vi giá trị giữa các đặc trưng. Điều này có thể khiến mô hình vô tình ưu tiên những đặc trưng có giá trị lớn hơn, dẫn đến kết quả học sai lệch và giảm hiệu quả dự đoán.

Để khắc phục đã áp dụng phương pháp Min-Max Scaling cho hai đặc trưng quan trọng là:

- “unemployment_rate”: Sau khi chuẩn hóa, giá trị được đưa về khoảng [0, 1], lưu vào cột “unemployment_rate_scaled”.
- "year": Biến thời gian ban đầu có giá trị từ 2014 đến 2024. Tuy nhiều mô hình học máy như Random Forest và XGBoost không yêu cầu chuẩn hóa biến đầu vào (vì không bị ảnh hưởng bởi thang đo), nhưng việc chuẩn hóa biến "year" vẫn là cần thiết trong bối cảnh nghiên cứu này. Biến "year" đã được chuẩn hóa và lưu vào cột “year_scaled”, nhằm phục vụ cho các mô hình nhạy cảm với thang đo, đặc biệt là Linear Regression và mạng nơ-ron (LSTM). Việc chuẩn hóa giúp:

  + Tránh hiện tượng mô hình bị chi phối bởi độ lớn tuyệt đối của biến thời gian, nhất là trong hồi quy tuyến tính hoặc mạng nơ-ron.

  + Đảm bảo các đặc trưng đầu vào có cùng tỷ lệ, giúp mô hình học xu hướng biến đổi theo thời gian một cách ổn định, cân bằng và khách quan hơn.

  + Tăng độ chính xác và hiệu quả huấn luyện cho các mô hình như LSTM – vốn rất nhạy cảm với quy mô dữ liệu đầu vào.

Mặc dù mô hình như Random Forest hoặc XGBoost không yêu cầu chuẩn hóa dữ liệu, việc chuẩn hóa vẫn được thực hiện vì:

- Đảm bảo đồng nhất về tỷ lệ giữa các đặc trưng, đặc biệt quan trọng đối với mô hình nhạy cảm với tỷ lệ như Linear Regression.
- Giúp mô hình học nhanh hơn và giảm nguy cơ bị mắc kẹt tại các điểm cực trị của hàm mất mát trong quá trình tối ưu hóa.
- Tăng khả năng hội tụ ổn định, đặc biệt khi huấn luyện trên tập dữ liệu lớn.

```
# Chuẩn hóa dữ liệu
scaler_rate = MinMaxScaler()
scaler_year = MinMaxScaler()
df_model["unemployment_rate_scaled"] = scaler_rate.fit_transform(df_model[["unemployment_rate"]])
df_model["year_scaled"] = scaler_year.fit_transform(df_model[["year"]])
```

Hình 2.6 Chuẩn hóa dữ liệu

- Lý do chọn MinMaxScaler thay vì StandardScaler

Bảng 2.2 Lý do lựa chọn MinMaxScaler

Tiêu chí	MinMaxScaler	StandardScaler
Nguyên lý	Đưa về khoảng [0, 1]	Chuẩn hóa về phân phối chuẩn (μ=0, σ=1)
Ảnh hưởng bởi ngoại lệ	Nhạy cảm	Ít nhạy cảm hơn
Phù hợp với dữ liệu	Có phân bố đều, không có ngoại lệ lớn	Dữ liệu phân phối chuẩn hoặc gần chuẩn
Lợi ích trong trường hợp này	Trực quan, đơn giản, phù hợp với year và unemployment_rate	Không cần thiết nếu dữ liệu không chuẩn

MinMaxScaler là lựa chọn phù hợp vì nó giữ nguyên quan hệ tỷ lệ của dữ liệu, hoạt động hiệu quả với các đặc trưng có phân bố đều như year, và hỗ trợ tốt cho nhiều mô hình học máy

2.1.2. Trích xuất và lựa chọn đặc trưng

Trích xuất và lựa chọn đặc trưng là một trong những bước quan trọng trong quy trình xây dựng mô hình học máy. Việc chọn lọc các biến đầu vào phù hợp sẽ không chỉ giúp cải thiện hiệu suất dự đoán mà còn giảm thiểu chi phí tính toán và tránh tình trạng overfitting – khi mô hình học quá tốt dữ liệu huấn luyện nhưng lại kém chính xác khi áp dụng vào dữ liệu thực tế.

- Trích xuất đặc trưng từ tập dữ liệu đã tiền xử lý

Sau khi hoàn thành bước tiền xử lý, bao gồm việc chuyển đổi định dạng “wide” sang “long”, xử lý giá trị thiếu, outliers, mã hóa nhãn và chuẩn hóa dữ liệu, tiến hành xác định các đặc trưng cần thiết cho việc huấn luyện mô hình.

Các đặc trưng được trích xuất bao gồm:

Bảng 2.3 Các đặc trưng trích xuất

Tên biến	Loại biến	Giải thích

country_name	Categorical (mã hóa)	Tên quốc gia, phản ánh điều kiện kinh tế - xã hội đặc thù của từng khu vực.
sex	Categorical (mã hóa)	Giới tính của đối tượng khảo sát, giúp phân tích sự chênh lệch về thất nghiệp giữa nam và nữ.
age_group	Categorical (mã hóa)	Nhóm tuổi của người lao động (theo phân loại quốc tế), nhằm nhận diện các nhóm có nguy cơ thất nghiệp cao.

Cột "age_categories" tuy cũng mang thông tin về độ tuổi, nhưng vì có sự trùng lặp về mặt ý nghĩa với age_group nên đã bị loại bỏ để giảm thiểu hiện tượng đa cộng tuyến.

- Đánh giá tính hợp lý của các đặc trưng được lựa chọn

Các đặc trưng cuối cùng được giữ lại cho mô hình là:

+ "country_name": Dữ liệu theo quốc gia giúp mô hình hiểu được những tác động khác biệt về mặt địa lý và chính sách đối với tỷ lệ thất nghiệp.

+ "sex": Phân tích theo giới tính rất quan trọng trong lĩnh vực lao động, nhất là khi có sự chênh lệch rõ rệt về cơ hội việc làm giữa nam và nữ ở một số quốc gia.

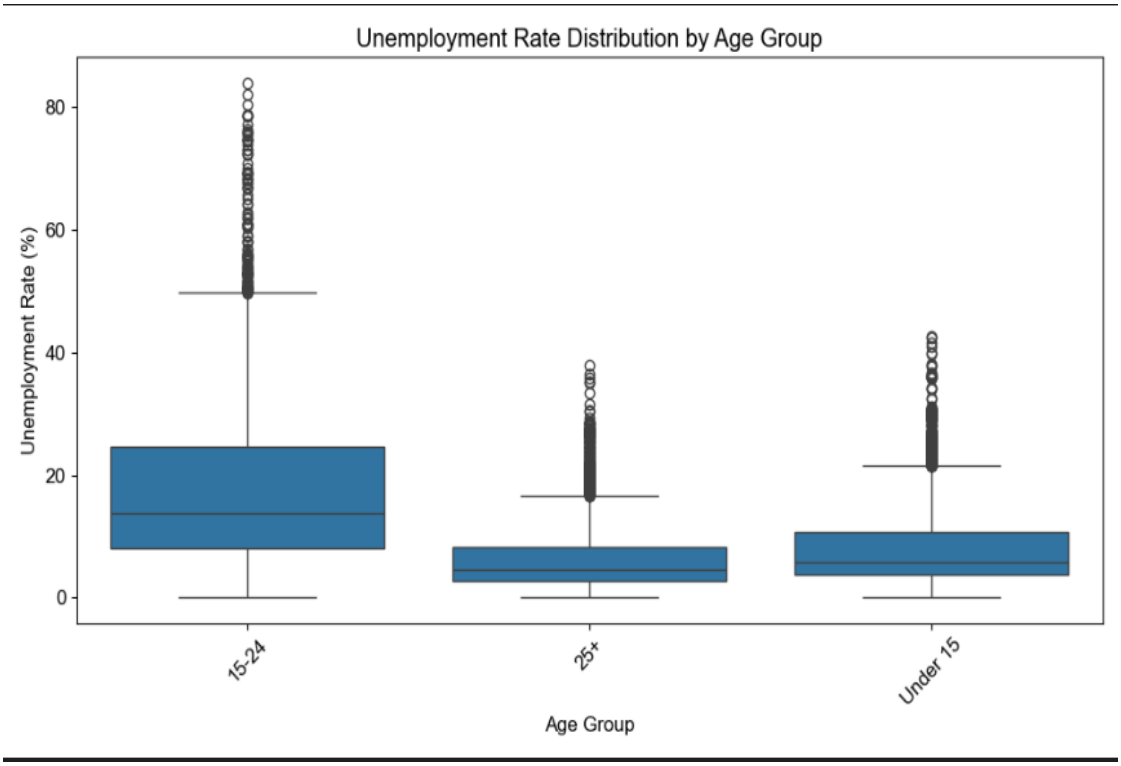
+ "age_group": Nhóm tuổi là một đặc trưng mang tính quyết định đến khả năng thất nghiệp, đặc biệt là ở nhóm thanh niên và người cao tuổi.

+ "year_scaled": Cho phép mô hình học được xu hướng thay đổi theo thời gian.

+ "unemployment_rate_scaled": Đây là biến mục tiêu mà mô hình học máy cần dự đoán, được chuẩn hóa để tránh ảnh hưởng bởi đơn vị đo lường hoặc chênh lệch quy mô giữa các quốc gia.

- EDA

* Tỉ lệ thất nghiệp theo nhóm tuổi:



Hình 2.7 Tỉ lệ thất nghiệp theo nhóm tuổi

Nhận xét:

- Nhóm tuổi dưới 15:

  + Hộp (IQR - khoảng tứ phân vị) kéo dài từ khoảng 5% đến 15%, cho thấy phần lớn tỷ lệ thất nghiệp nằm trong khoảng này.

  + Đường giữa hộp (trung vị) khoảng 10%, nghĩa là 50% dữ liệu có tỷ lệ thất nghiệp dưới 10%.

  + Không có giá trị ngoại lai (outliers), phân bố khá đồng đều.

- Nhóm tuổi 15-24:

  + Hộp IQR rất hẹp, từ khoảng 15% đến 20%, cho thấy tỷ lệ thất nghiệp tập trung trong một phạm vi nhỏ.

  + Trung vị khoảng 17-18%.

  + Có nhiều giá trị ngoại lai (outliers) kéo dài lên đến hơn 80%, cho thấy một số trường hợp có tỷ lệ thất nghiệp cực cao, nhưng không phổ biến.

  + Phân bố lệch mạnh về phía trên.

- Nhóm tuổi 25+:

+ Hộp IQR từ khoảng 5% đến 10%, nhỏ hơn nhóm 15-24, cho thấy tỷ lệ thất nghiệp thấp và ổn định hơn.

+ Trung vị khoảng 7-8%.

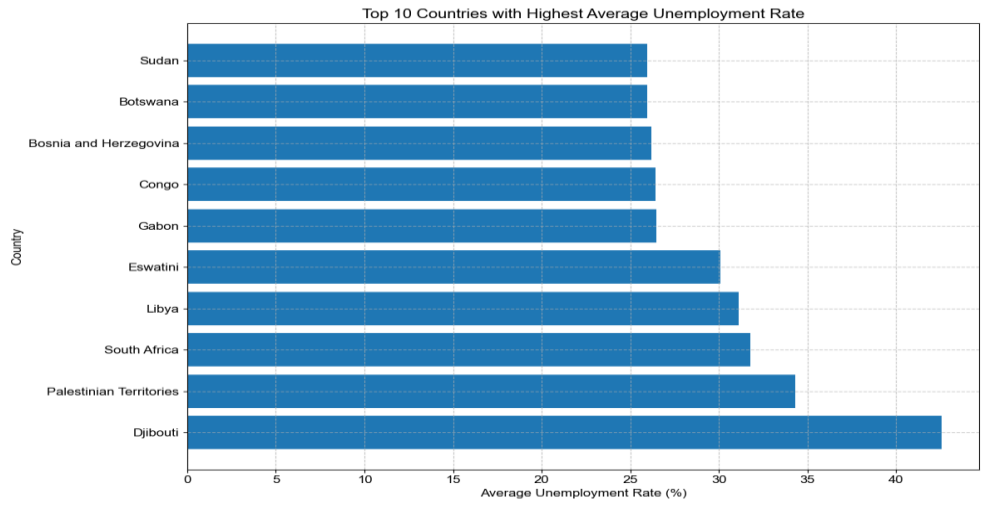
+ Có một số giá trị ngoại lai kéo dài lên đến khoảng 40%, nhưng ít hơn so với nhóm 15-24.

+ Phân bố cũng lệch về phía trên, nhưng ít cực đoan hơn nhóm 15-24.

So sánh chung:

- Nhóm 15-24 có tỷ lệ thất nghiệp trung bình cao nhất (trung vị ~17-18%) và biến động lớn nhất (nhiều outliers).
- Nhóm dưới 15 và 25+ có tỷ lệ thất nghiệp trung bình thấp hơn (trung vị ~10% và 7-8%), với ít biến động hơn.
- Nhóm 15-24 có vẻ chịu ảnh hưởng thất nghiệp nặng nề nhất, với một số trường hợp cực kỳ cao, có thể do thiếu kinh nghiệm hoặc khó tìm việc trong giai đoạn đầu sự nghiệp.

* Biểu đồ 10 quốc gia có tỉ lệ thất nghiệp cao



Hình 2.8 Top 10 quốc gia có tỉ lệ thất nghiệp cao

- Phân bố địa lý:

+ Hầu hết các quốc gia trong danh sách đều thuộc khu vực châu Phi (Djibouti, South Africa, Libya, Eswatini, Gabon, Congo, Botswana, Sudan), cho thấy khu vực này đang đối mặt với tỷ lệ thất nghiệp cao.

+ Ngoài ra, có Palestinian Territories (Trung Đông) và Bosnia and Herzegovina (châu Âu), cho thấy vấn đề thất nghiệp không chỉ giới hạn ở châu Phi.

- Phân tích xu hướng:

+ Tỷ lệ thất nghiệp dao động từ 18% (Sudan) đến 40% (Djibouti), cho thấy sự chênh lệch đáng kể giữa các quốc gia.

+ Các quốc gia có tỷ lệ thất nghiệp cao nhất (Djibouti, Palestinian Territories) thường đối mặt với các vấn đề như bất ổn chính trị, xung đột hoặc nền kinh tế kém phát triển.

+ Những quốc gia như South Africa và Bosnia and Herzegovina, dù có nền kinh tế tương đối phát triển hơn trong danh sách, vẫn gặp khó khăn với thất nghiệp, có thể do bất bình đẳng kinh tế hoặc thị trường lao động không hiệu quả.

- Nguyên nhân tiềm tàng:

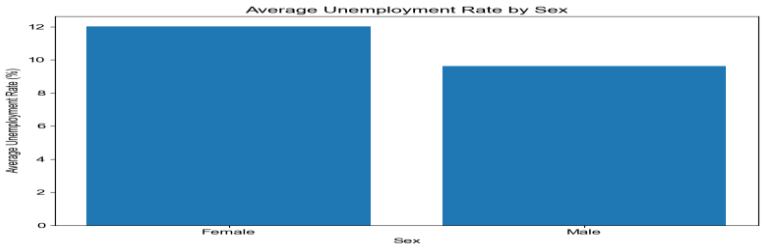
+ Các quốc gia châu Phi thường gặp vấn đề về tăng trưởng dân số nhanh, thiếu cơ hội việc làm và cơ sở hạ tầng kinh tế yếu.

+ Palestinian Territories chịu ảnh hưởng từ xung đột kéo dài, gây gián đoạn thị trường lao động.

+ Bosnia and Herzegovina có thể bị ảnh hưởng bởi hệ quả của chiến tranh trước đây và quá trình chuyển đổi kinh tế chậm.

Kết luận: Biểu đồ cho thấy thất nghiệp là vấn đề nghiêm trọng ở nhiều khu vực, đặc biệt là châu Phi. Các quốc gia này cần cải thiện chính sách kinh tế, đầu tư vào giáo dục và tạo thêm cơ hội việc làm để giảm tỷ lệ thất nghiệp.

* Tỉ lệ thất nghiệp theo giới tính



Hình 2.9 Biểu đồ thể hiện tỉ lệ thất nghiệp theo giới tính

- So sánh tỷ lệ thất nghiệp:

  + Nữ (Female): Tỷ lệ thất nghiệp trung bình khoảng 11%.

  + Nam (Male): Tỷ lệ thất nghiệp trung bình khoảng 9%.

- Chênh lệch giới tính:

  + Phụ nữ có tỷ lệ thất nghiệp cao hơn nam giới khoảng 2 điểm phần trăm.

  + Điều này cho thấy phụ nữ có khả năng gặp khó khăn hơn trong việc tìm kiếm việc làm so với nam giới.

- Nguyên nhân tiềm tàng:

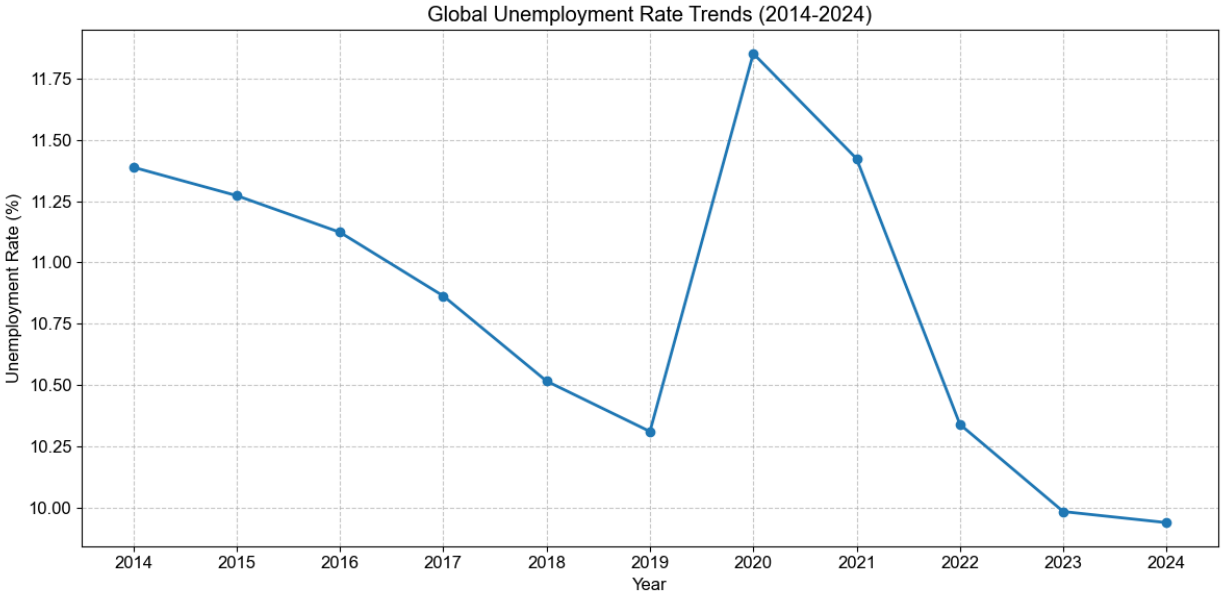
  + Chênh lệch có thể do các yếu tố như phân biệt giới tính trong tuyển dụng, trách nhiệm gia đình khiến phụ nữ khó tham gia thị trường lao động, hoặc sự thiếu hụt cơ hội việc làm trong các ngành nghề mà phụ nữ thường tham gia.

  + Ở một số khu vực, định kiến xã hội hoặc thiếu giáo dục/việc làm phù hợp cho phụ nữ cũng có thể góp phần vào tỷ lệ thất nghiệp cao hơn.

- Ý nghĩa:

  + Dữ liệu này nhấn mạnh sự bất bình đẳng giới trong thị trường lao động, cho thấy cần có các chính sách hỗ trợ việc làm cho phụ nữ, như tăng cường đào tạo nghề, khuyến khích sự tham gia của phụ nữ trong các ngành nghề đa dạng, hoặc hỗ trợ cân bằng công việc và gia đình.

  * Biểu đồ phân tích biến động của tỉ lệ toàn cầu từ 2014-2024



Hình 2.10 Biểu đồ phân tích biến động tỉ lệ toàn cầu (2014-2024)

- Từ 2014-2019: Xu hướng giảm dần

  + Năm 2014, tỷ lệ thất nghiệp bắt đầu ở mức khoảng 11,25%.

  + Từ 2015 đến 2017, tỷ lệ giảm ổn định, xuống khoảng 11,15% vào 2015 và tiếp tục giảm xuống khoảng 11,0% vào 2017, cho thấy sự cải thiện kinh tế toàn cầu trước đại dịch.

  + Từ 2018 đến 2019, xu hướng giảm chậm lại, với tỷ lệ đạt khoảng 10,25% vào 2019, phản ánh một giai đoạn tăng trưởng kinh tế bền vững.

- Năm 2020: Đỉnh điểm do đại dịch

  + Năm 2020 chứng kiến sự gia tăng đột biến lên khoảng 11,75%, có thể do tác động của đại dịch COVID-19, dẫn đến gián đoạn kinh tế, đóng cửa doanh nghiệp, và mất việc làm trên toàn cầu.

- 2021-2024: Phục hồi mạnh mẽ

  + Từ 2021, tỷ lệ bắt đầu giảm, xuống khoảng 11,5%, cho thấy các biện pháp hỗ trợ kinh tế và phục hồi ban đầu.

  + Đến 2022, tỷ lệ tiếp tục giảm xuống khoảng 10,5%, phản ánh sự ổn định dần dần.

  + Từ 2023 đến 2024, tỷ lệ giảm sâu và ổn định ở mức khoảng 10,0%, đánh dấu mức thấp nhất trong thập kỷ, cho thấy sự phục hồi mạnh mẽ và hiệu quả của các chính sách kinh tế toàn cầu.

- Nhận xét tổng quan:

  + Giai đoạn 2014-2019 cho thấy xu hướng tích cực với tỷ lệ thất nghiệp giảm dần, phản ánh sự tăng trưởng kinh tế trước đại dịch.

  + Đỉnh điểm 2020 là một ngoại lệ do tác động của COVID-19, nhưng sự phục hồi từ 2021-2024 rất ấn tượng, với tỷ lệ trở về mức thấp nhất kể từ 2014.

  + Xu hướng này có thể liên quan đến các chính sách kích thích kinh tế, tự động hóa lao động, và sự thích nghi của thị trường việc làm sau đại dịch.

2.2. Xây dựng mô hình AI dự đoán tỉ lệ thất nghiệp

2.2.1. Mô hình Machine Learning

Sau khi hoàn tất giai đoạn tiền xử lý và chuẩn hóa dữ liệu, tiến hành xây dựng và huấn luyện ba mô hình học máy nhằm mục tiêu dự đoán tỷ lệ thất nghiệp theo từng quốc gia, giới tính, nhóm tuổi và năm cụ thể. Ba mô hình được lựa chọn bao gồm:

- Linear Regression
- Random Forest Regressor
- XGBoost Regressor

a. Mô hình Linear Regression

Hồi quy tuyến tính là một trong những thuật toán nền tảng và được sử dụng phổ biến trong lĩnh vực học máy và thống kê. Mô hình này giả định rằng tồn tại một mối quan hệ tuyến tính giữa biến đầu ra (biến phụ thuộc) và một hoặc nhiều biến đầu vào (biến độc lập). Mục tiêu của hồi quy tuyến tính là tìm ra một hàm tuyến tính sao cho sai số giữa giá trị dự đoán và giá trị thực là nhỏ nhất, thường được đo lường thông qua hàm mất mát bình phương sai số trung bình (Mean Squared Error - MSE).

Trong trường hợp tổng quát, mô hình hồi quy tuyến tính đa biến có thể được biểu diễn dưới dạng:

$$\hat{y} = w^T x + b$$

Trong đó:

1. $\hat{y}$: giá trị dự đoán
2. x: vector đặc trưng đầu vào
3. w^T: vector hệ số hồi quy
4. b: hệ số chệch (intercept)

- Tham số tối ưu tìm được trong n:

+ “fit_intercept = True”: cho phép mô hình học thêm hệ số chệch, nhằm dịch chuyển đường hồi quy phù hợp hơn với dữ liệu thực tế.

+ “positive = False”: không áp đặt ràng buộc dấu dương lên các hệ số hồi quy, từ đó cho phép mô hình học được cả mối quan hệ thuận và nghịch.

- Hiệu quả của mô hình hồi quy tuyến tính được đánh giá trên tập dữ liệu kiểm tra thông qua bốn chỉ số phổ biến, bao gồm: MAE, MSE, RMSE và R^2.

Bảng 2.4 Kết quả mô hình Linear Regression

Mô hình	MAE	RMSE	R^2
Linear Regression	7.0459	9.8713	0.1566

- MAE = 7.0459

  + MAE là sai số tuyệt đối trung bình, phản ánh mức chênh lệch trung bình giữa giá trị thực tế và giá trị dự đoán của mô hình. Với giá trị MAE là 7.0459, điều này có nghĩa là, trung bình mỗi dự đoán của mô hình sai lệch khoảng 7 điểm phần trăm so với giá trị thực tế. MAE là một chỉ số dễ diễn giải và ít bị ảnh hưởng bởi các giá trị ngoại lai, nhưng không phản ánh mức độ nghiêm trọng của các sai số lớn.

- MSE = 97.4434

  + MSE là sai số bình phương trung bình, tính bằng cách bình phương sai số giữa giá trị thực và giá trị dự đoán, sau đó lấy trung bình. Việc sử dụng bình phương khiến MSE nhạy cảm hơn với các điểm dữ liệu có sai số lớn (outliers). Với giá trị gần 100, chỉ số này cho thấy tồn tại nhiều điểm dữ liệu mà mô hình dự đoán sai lệch ở mức nghiêm trọng.

- RMSE = 9.8713

  + RMSE là căn bậc hai của MSE, giúp đưa sai số về cùng đơn vị với biến mục tiêu (tỷ lệ thất nghiệp). Với giá trị gần 10, RMSE cho thấy rằng sai số dự đoán của mô hình là đáng kể. Chẳng hạn, nếu tỷ lệ thất nghiệp dao động trong khoảng từ 0% đến 100%, thì sai số trung bình gần 10 điểm phần trăm có thể gây ra những kết luận sai lệch trong thực tế, đặc biệt trong các ứng dụng cần độ chính xác cao như hoạch định chính sách lao động.

- $R^2 = 0.1566$

+ Hệ số xác định phản ánh tỷ lệ phương sai của biến mục tiêu được mô hình giải thích thông qua các biến đầu vào. Với giá trị R^2 =0.1566, mô hình chỉ giải thích được khoảng 15.66% sự biến thiên trong dữ liệu thực tế. Điều này cho thấy phần lớn sự biến động của tỷ lệ thất nghiệp không thể được lý giải bằng mô hình tuyến tính hiện tại. Giá trị thấp của R^2 là một chỉ báo rõ ràng rằng mô hình đang underfitting, tức là không đủ phức tạp để mô hình hóa mối quan hệ giữa các biến trong tập dữ liệu.

***** Nguyên nhân khiến Linear Regression bị underfitting trong trường hợp này đến từ các yếu tố sau:

- Giả định tuyến tính không phù hợp với dữ liệu thực tế: Linear Regression giả định rằng mối quan hệ giữa các biến độc lập và biến phụ thuộc là tuyến tính. Tuy nhiên, trong bối cảnh dự đoán tỷ lệ thất nghiệp, các yếu tố kinh tế - xã hội như năm, nhóm tuổi, trình độ học vấn thường có mối quan hệ phi tuyến, hoặc tương tác phức tạp với nhau. Mô hình tuyến tính không có khả năng tự học các mối quan hệ phi tuyến đó.
- Không học được tương tác giữa các biến: Linear Regression không tự động khai thác các mối tương tác giữa các đặc trưng đầu vào, trừ khi các biến tương tác được đưa vào thủ công. Điều này khiến mô hình không tận dụng được toàn bộ thông tin có trong dữ liệu.
- Không phù hợp khi dữ liệu có nhiều yếu tố ẩn: Trong thực tế, tỷ lệ thất nghiệp chịu ảnh hưởng bởi các yếu tố kinh tế vĩ mô, chính sách lao động, thị trường việc làm… mà không được thể hiện đầy đủ trong tập dữ liệu. Linear Regression, do đơn giản, càng khó có khả năng suy diễn hoặc khái quát hóa những ảnh hưởng ẩn đó.

Tổng thể các chỉ số đánh giá đều cho thấy rằng mô hình hồi quy tuyến tính hoạt động không hiệu quả trong bối cảnh bài toán này. Sai số dự đoán lớn và khả năng giải thích phương sai dữ liệu thấp gợi ý rằng giả định về mối quan hệ tuyến tính giữa các đặc trưng và biến mục tiêu không phù hợp với bản chất của dữ liệu. Đây là cơ

sở để chuyển sang các mô hình phức tạp hơn, có khả năng nắm bắt tốt hơn các mối quan hệ phi tuyến trong dữ liệu thực tế.

b. Mô hình Random Forest Regressor

Random Forest Regressor là một mô hình học máy mạnh mẽ, thuộc nhóm kỹ thuật tổng hợp mô hình (ensemble learning). Cốt lõi của phương pháp này là việc xây dựng một tập hợp các cây quyết định (decision trees), mỗi cây được huấn luyện trên một mẫu con được chọn ngẫu nhiên có hoàn lại (bootstrap sample) từ tập dữ liệu gốc. Quá trình này được gọi là bagging (Bootstrap Aggregating).

Trong hồi quy, Random Forest không sử dụng một cây duy nhất để dự đoán đầu ra mà thay vào đó lấy trung bình dự đoán của tất cả các cây trong rừng. Cách tiếp cận này có nhiều lợi ích:

- Giảm phương sai: Do các cây được huấn luyện trên các tập dữ liệu khác nhau, sự đa dạng này giúp mô hình tổng thể trở nên ổn định hơn và ít bị ảnh hưởng bởi nhiễu trong dữ liệu.
- Giảm nguy cơ overfitting: Mỗi cây có thể overfit dữ liệu riêng của nó, nhưng trung bình của nhiều cây sẽ làm mờ đi các cực trị, giúp mô hình tổng quát hóa tốt hơn.
- Khả năng mô hình hóa mối quan hệ phi tuyến: Khác với Linear Regression, Random Forest không áp đặt giả định tuyến tính lên dữ liệu, nên có thể học được các quy luật phức tạp và phi tuyến.

Random Forest được ưa chuộng nhờ hiệu năng mạnh, khả năng xử lý tốt dữ liệu nhiều chiều, không yêu cầu chuẩn hóa đặc trưng, và không quá nhạy cảm với giá trị ngoại lai (outliers).

- Tham số tối ưu tìm được:

Trong quá trình huấn luyện mô hình Random Forest Regressor, việc điều chỉnh siêu tham số đóng vai trò then chốt trong việc tối ưu hóa hiệu suất. Các siêu tham số tốt nhất được lựa chọn thông qua tìm kiếm có giám sát (grid search hoặc randomized search) như sau:

‘n_estimators’ = 100

Số lượng cây trong rừng. Giá trị 100 là một lựa chọn phổ biến, đảm bảo sự cân bằng giữa hiệu suất và tốc độ huấn luyện. Số lượng cây càng nhiều thì mô hình càng ổn định, tuy nhiên cũng làm tăng thời gian tính toán.

‘max_depth’=15

Giới hạn độ sâu tối đa của mỗi cây. Độ sâu lớn giúp cây học được nhiều mẫu phức tạp, nhưng cũng dễ bị overfit nếu quá sâu. Việc đặt giới hạn ở 15 cho phép cây khai thác được các mối quan hệ phi tuyến trong khi vẫn kiểm soát được độ phức tạp. Số lượng mẫu tối thiểu cần có để tách một nút bên trong cây. Giá trị nhỏ cho phép cây phân chia dữ liệu đến mức chi tiết cao, phù hợp với bài toán yêu cầu độ chính xác cao.

‘min_samples_leaf’= 1

Số lượng mẫu tối thiểu tại mỗi nút lá. Khi được đặt là 1, cây có thể học đến những biểu hiện rất nhỏ trong dữ liệu, nhưng cũng cần được giám sát để tránh overfitting.

‘min_samples_split’=2

Số lượng mẫu tối thiểu cần thiết để một node được tiếp tục chia tách. Với giá trị bằng 2 (nhỏ nhất hợp lệ), mô hình có thể tạo ra các cây rất chi tiết, giúp tăng độ chính xác nhưng cũng cần được kiểm soát để tránh học cả nhiễu của dữ liệu.

Kết quả đánh giá:

Bảng 2.5 Kết quả mô hình Random Forest

Mô hình	MAE	RMSE	R^2
Random Forest	2.0634	3.0448	0.9198

- MAE: 2.0634

  + Trung bình sai số tuyệt đối giữa giá trị dự đoán và thực tế là khoảng 2.06 đơn vị.

- MSE: 9.2711

  + Sai số bình phương trung bình tương đối thấp, cho thấy ít xuất hiện các sai số lớn.

- RMSE: 3.0448

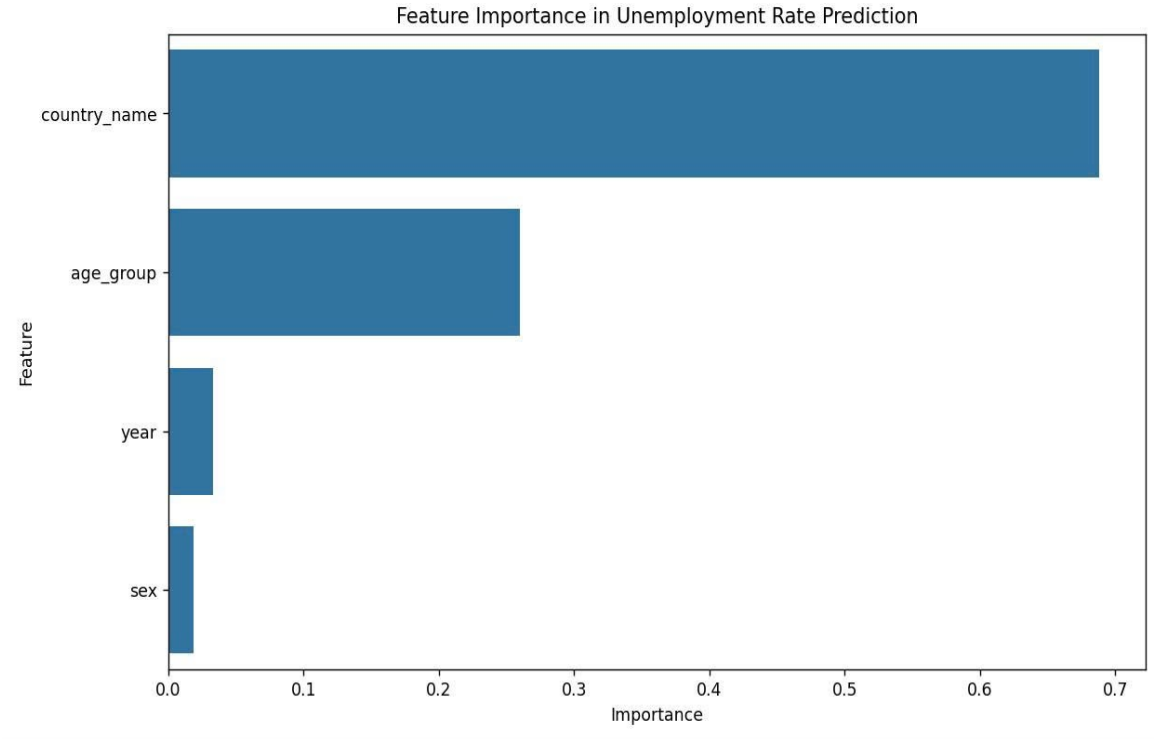
  + Sai số căn trung bình khoảng 3 đơn vị, giúp hình dung trực quan hơn về mức độ chênh lệch dự đoán.

- R^2: 0.9198

  + Mô hình giải thích được 91.98% phương sai của biến mục tiêu, phản ánh mức độ khớp tốt giữa dự đoán và thực tế.

Phân tích hiệu quả: Random Forest cải thiện đáng kể độ chính xác so với Linear Regression. Với R^2 đạt 0.9198, mô hình có khả năng giải thích đến 91.98% phương sai của dữ liệu đầu ra, chứng tỏ mô hình học được các mối quan hệ phi tuyến và có độ khái quát hóa tốt. Việc sử dụng nhiều cây với kỹ thuật bagging giúp giảm thiểu hiện tượng overfitting, đồng thời tạo nên một mô hình mạnh mẽ và ổn định.

* Feature Importance trong mô hình Random Forest



Hình 2.11 Feature Importance (Random Forest)

Có thể thấy rõ rằng “country_name” là đặc trưng có tầm quan trọng cao nhất trong mô hình Random Forest để dự đoán tỷ lệ thất nghiệp:

- Sự khác biệt kinh tế và xã hội: Tỷ lệ thất nghiệp phụ thuộc rất nhiều vào tình hình kinh tế, chính sách của chính phủ, cấu trúc thị trường lao động, và các yếu tố xã hội đặc thù của từng quốc gia. Mỗi quốc gia có những đặc điểm riêng biệt ảnh hưởng đến tỷ lệ thất nghiệp của họ.
- Biến động theo quốc gia: Mức độ và xu hướng biến động của tỷ lệ thất nghiệp khác nhau đáng kể giữa các quốc gia. Một mô hình dự đoán tỷ lệ thất nghiệp sẽ tìm thấy rằng việc biết quốc gia cụ thể là cực kỳ hữu ích để phân loại và dự đoán tỷ lệ thất nghiệp một cách chính xác.
- Mô hình hoạt động: Random Forest hoạt động bằng cách tạo ra nhiều cây quyết định. Trong quá trình xây dựng cây, mô hình tìm kiếm các đặc trưng giúp phân chia dữ liệu thành các nhóm đồng nhất nhất (giảm sự "không tinh khiết" - impurity). Nếu tên quốc gia là một yếu tố mạnh mẽ giúp phân tách dữ liệu về tỷ lệ thất nghiệp, nó sẽ được sử dụng thường xuyên và ở các vị trí cao trong các cây quyết định, từ đó có điểm tầm quan trọng cao.

c. Mô hình XGBoost Regressor

XGBoost (Extreme Gradient Boosting) là một trong những thuật toán học máy thuộc nhóm boosting, nổi bật bởi tính hiệu quả, khả năng mở rộng và độ chính xác cao. Boosting là một kỹ thuật học có giám sát trong đó nhiều mô hình học yếu, thường là cây quyết định nông (shallow decision trees), được kết hợp tuần tự nhằm tạo ra một mô hình mạnh.

Khác với Random Forest – nơi các cây được xây dựng song song và độc lập – XGBoost xây dựng các cây liên tiếp theo cách tuần tự, trong đó mỗi cây mới được huấn luyện để sửa chữa sai số của mô hình tổng hợp trước đó. Cụ thể, các cây kế tiếp học từ phần dư (residuals) – tức là chênh lệch giữa giá trị thực và dự đoán trước đó – và nhờ đó, dần cải thiện hiệu suất mô hình.

Điểm nổi bật của XGBoost không chỉ nằm ở phương pháp boosting mà còn ở các kỹ thuật tối ưu được tích hợp sẵn, bao gồm:

- Regularization (L1 và L2): Giúp kiểm soát độ phức tạp của mô hình, giảm nguy cơ quá khớp (overfitting).
- Shrinkage (learning rate): Điều chỉnh mức độ đóng góp của mỗi cây mới trong mô hình tổng hợp, giúp mô hình học chậm và chính xác hơn.
- Pruning sớm (early stopping): Ngừng huấn luyện sớm nếu mô hình không cải thiện trên tập kiểm tra, tiết kiệm thời gian và tài nguyên tính toán.
- Xử lý giá trị thiếu tự động: XGBoost có cơ chế riêng để xử lý missing values mà không cần phải tiền xử lý thủ công.
- Song song hóa và tối ưu hóa bộ nhớ: Giúp tăng tốc quá trình huấn luyện trên tập dữ liệu lớn.

Tham số tối ưu tìm được:

Thông qua quá trình tìm kiếm và điều chỉnh siêu tham số (hyperparameter tuning), mô hình XGBoost Regressor được cấu hình với các giá trị tối ưu như sau:

‘n_estimators’=200

Số lượng cây (số vòng boosting). Việc sử dụng 200 vòng cho phép mô hình có đủ khả năng học sâu từ dữ liệu nhưng vẫn nằm trong kiểm soát nhờ các kỹ thuật regularization.

‘max_depth’=7

Độ sâu tối đa của mỗi cây quyết định. Độ sâu vừa phải giúp mô hình học đủ độ chi tiết mà không gây ra hiện tượng quá khớp.

‘learning_rate’=0.2

Tốc độ học (shrinkage). Đây là tham số kiểm soát mức độ mỗi cây mới đóng góp vào mô hình tổng hợp. Learning rate thấp giúp học chậm hơn nhưng ổn định hơn, trong khi 0.2 là mức trung bình hợp lý để cân bằng giữa hiệu quả và tốc độ hội tụ.

‘subsample’=0.9

Tỷ lệ dữ liệu sử dụng cho mỗi cây. Giúp giảm phương sai và tăng tính ngẫu nhiên, từ đó nâng cao khả năng khái quát hóa.

‘colsample_bytree’=0.8

Tỷ lệ đặc trưng được sử dụng tại mỗi cây. Kỹ thuật này tương tự với Random Forest, giúp giảm mối tương quan giữa các cây và cải thiện khả năng tổng quát của mô hình.

Việc phối hợp các tham số trên đã tạo ra một mô hình cân bằng giữa độ chính xác cao và khả năng khái quát tốt trên dữ liệu chưa từng thấy.

Kết quả đánh giá:

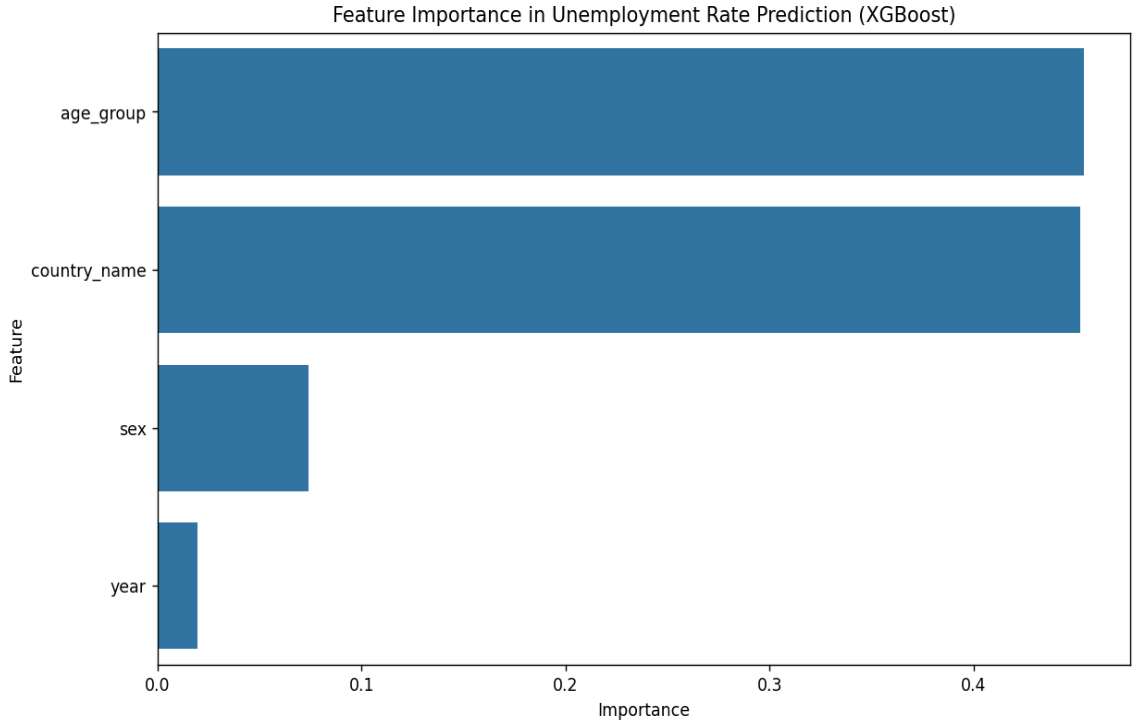
Bảng 2.6 Kết quả mô hình XGBoost

Mô hình	MAE	MSE	RMSE	R^2
XGBoost	1.1703	3.2570	1.8047	0.9718

- MAE đạt giá trị 1.1703, cho thấy sai số tuyệt đối trung bình giữa giá trị dự đoán và giá trị thực tế là rất nhỏ – chỉ vào khoảng 1.17 đơn vị. Điều này chứng tỏ mô hình có khả năng đưa ra các dự đoán gần sát với thực tế trong hầu hết các trường hợp.
- MSE giá trị 3.2570, phản ánh sai số bình phương trung bình giữa các dự đoán và giá trị thực. Chỉ số này càng nhỏ càng tốt vì nó nhấn mạnh vào các sai số lớn. Với kết quả này, ta có thể nhận định rằng mô hình ít gặp các dự đoán sai nghiêm trọng.
- RMSE – giá trị căn bậc hai của MSE – là 1.8047. Đây là một chỉ số trực quan hơn vì nó cùng đơn vị với biến mục tiêu, giúp hình dung sai số trung bình một cách dễ hiểu. Với RMSE chỉ khoảng 1.8, mô hình cho thấy mức độ sai lệch trung bình nhỏ giữa dự đoán và thực tế, thể hiện tính ổn định và chính xác cao.
- R^2 (hệ số xác định) đạt giá trị rất cao là 0.9718. Điều này có nghĩa là mô hình có thể giải thích tới 97.18% phương sai của biến mục tiêu. Đây là chỉ số quan trọng phản ánh mức độ phù hợp của mô hình với dữ liệu: giá trị càng gần 1 thì mô hình càng hiệu quả trong việc mô tả mối quan hệ giữa các biến đầu vào

và biến đầu ra. Với R^2 đạt mức này, có thể khẳng định XGBoost đã học được gần như toàn bộ cấu trúc ẩn trong dữ liệu.

Phân tích hiệu quả: XGBoost là mô hình hoạt động tốt nhất trong cả ba mô hình. Với R^2 đạt tới 0.9718, mô hình có thể giải thích tới 97.18% phương sai của dữ liệu, và sai số rất thấp (RMSE chỉ 1.8). Đây là bằng chứng cho thấy XGBoost có khả năng học tốt từ dữ liệu phức tạp và ít bị ảnh hưởng bởi overfitting, nhờ các cơ chế regularization và boosting có kiểm soát. Ngoài ra, mô hình cũng có khả năng xử lý tốt dữ liệu mất mát và không cần quá nhiều xử lý đặc biệt cho dữ liệu đầu vào.



Hình 2.12 Feature Importance (XGBoost)

“age_group” và “country_name”: Đây là hai đặc trưng có tầm quan trọng cao nhất và gần như tương đương nhau trong mô hình XGBoost. Điều này cho thấy cả nhóm tuổi và quốc gia đều là những yếu tố cực kỳ quan trọng và có ảnh hưởng lớn đến tỷ lệ thất nghiệp khi mô hình XGBoost đưa ra dự đoán.

- “age_group”: Tỷ lệ thất nghiệp thường khác nhau đáng kể giữa các nhóm tuổi. Mô hình XGBoost đã học được mối quan hệ này và coi nhóm tuổi là một yếu tố dự báo mạnh mẽ.

- “country_name”: Tương tự như với Random Forest, sự khác biệt về kinh tế, chính sách và thị trường lao động giữa các quốc gia khiến tên quốc gia trở thành một đặc trưng quan trọng.
- “sex”: Đặc trưng này có tầm quan trọng thấp hơn đáng kể so với nhóm tuổi và quốc gia. Điều này ngụ ý rằng giới tính có ảnh hưởng đến tỷ lệ thất nghiệp, nhưng mức độ ảnh hưởng không mạnh mẽ bằng nhóm tuổi hoặc quốc gia trong mô hình XGBoost này.
- “year”: Đặc trưng năm có tầm quan trọng thấp nhất trong số các đặc trưng được phân tích. Điều này có thể chỉ ra rằng, sau khi tính đến quốc gia và nhóm tuổi, yếu tố thời gian (năm) đóng vai trò ít quan trọng hơn trong việc dự đoán tỷ lệ thất nghiệp trong phạm vi dữ liệu này.

2.2.2 Mô hình Deep Learning (LSTM)

Sau khi đã triển khai và đánh giá hiệu suất của các mô hình học máy truyền thống như Linear Regression, Random Forest và XGBoost – những mô hình chủ yếu dựa trên giả định về mối quan hệ tĩnh giữa các đặc trưng đầu vào và đầu ra. Mở rộng hướng nghiên cứu bằng cách ứng dụng học sâu cụ thể là sử dụng mạng nơ-ron hồi tiếp LSTM.

Mục tiêu chính của bước phát triển này là khai thác bản chất chuỗi thời gian (temporal nature) của dữ liệu tỷ lệ thất nghiệp toàn cầu, vốn có sự phụ thuộc chặt chẽ vào các yếu tố theo thời gian như chu kỳ kinh tế, chính sách xã hội, và các biến cố toàn cầu (đại dịch, khủng hoảng tài chính).

Tổng quan về mô hình LSTM:

LSTM là một biến thể đặc biệt của mạng nơ-ron hồi tiếp (RNN), được thiết kế để giải quyết các vấn đề về độ dài phụ thuộc dài hạn mà các RNN truyền thống thường gặp phải (mất mát thông tin theo thời gian). Cơ chế chính của LSTM là sử dụng các "bộ nhớ có cổng" để điều chỉnh quá trình lưu trữ và quên thông tin trong chuỗi dữ liệu.

Điều này khiến LSTM đặc biệt phù hợp với các bài toán có dữ liệu dạng chuỗi như:

- Dự báo tài chính,
- Phân tích xu hướng thị trường,
- Dự đoán nhu cầu lao động,
- Và trong trường hợp này – dự đoán và mô hình hóa chuỗi thời gian của tỷ lệ thất nghiệp toàn cầu.

a. Tiền xử lý dữ liệu cho LSTM

Dữ liệu ban đầu được xử lý và biến đổi từ định dạng rộng (wide) sang định dạng dài (long) để phù hợp với bài toán chuỗi thời gian. Các bước chính bao gồm:

- Loại bỏ giá trị thiếu và outliers bằng phương pháp IQR.
- Mã hóa các biến phân loại như quốc gia, giới tính và nhóm tuổi bằng LabelEncoder.
- Chuẩn hóa dữ liệu bằng MinMaxScaler cho cả biến mục tiêu và biến thời gian để mô hình học hiệu quả hơn.

Sau xử lý, dữ liệu được chuyển đổi thành các giá trị “time_steps = 5” được xác định sau quá trình kiểm thử, giúp mạng LSTM học đủ bối cảnh ngắn hạn của chuỗi thất nghiệp và thêm Gaussian noise với “noise_level=0.01” trong code LSTM là một kỹ thuật tăng cường dữ liệu. Nó hoạt động bằng cách:

- Thêm một lượng nhiễu ngẫu nhiên nhỏ (từ phân phối chuẩn với độ lệch chuẩn 0.01 trên dữ liệu đã chuẩn hóa) vào mỗi chuỗi dữ liệu huấn luyện gốc.
- Tạo ra các phiên bản "hơi khác" của dữ liệu huấn luyện.

* Mục đích chính là:

- Tăng dữ liệu huấn luyện: Nhân đôi số lượng mẫu huấn luyện (mẫu gốc + mẫu nhiễu).
- Tăng tính mạnh mẽ của mô hình: Giúp mô hình ít bị ảnh hưởng bởi sai số nhỏ trong dữ liệu.
- Chống overfitting: Ngăn mô hình "ghi nhớ" quá kỹ dữ liệu huấn luyện, giúp nó dự đoán tốt hơn trên dữ liệu mới.

Đây là cách làm cho mô hình học từ nhiều biến thể dữ liệu hơn, giúp nó trở nên đáng tin cậy và tổng quát hóa tốt hơn.

b. Kiến trúc mô hình LSTM

Mô hình được thiết kế dưới dạng Bidirectional LSTM nhiều lớp, kết hợp các kỹ thuật Dropout, Batch Normalization và Dense layers nhằm tăng cường khả năng học phi tuyến tính và ổn định trong quá trình huấn luyện

- 3 lớp Bidirectional LSTM (128, 64, 32 neurons) cho phép mô hình học cả xu hướng trước và sau trong chuỗi thời gian
- Các lớp BatchNormalization và Dropout giúp ổn định quá trình học và tránh overfitting.
- Lớp đầu ra (output) sử dụng hàm kích hoạt sigmoid vì dữ liệu đã được chuẩn hóa trong khoảng [0, 1].
- Mô hình được tối ưu hóa với Adam Optimizer và sử dụng Hàm mất mát Huber nhằm kết hợp lợi ích giữa MSE và MAE.

c. Huấn luyện mô hình

Việc huấn luyện mô hình sử dụng các kỹ thuật hiện đại:

- EarlyStopping: Dừng sớm nếu không còn cải thiện trên tập validation.
- ReduceLROnPlateau: Giảm learning rate nếu loss không cải thiện.
- ModelCheckpoint: Lưu lại mô hình tốt nhất dựa trên val_loss.

Kết quả quá trình huấn luyện cho thấy mô hình hội tụ tốt và tránh được hiện tượng overfitting.

d. Đánh giá hiệu suất mô hình

Sau khi huấn luyện, mô hình được đánh giá trên tập kiểm tra:

+ MAE: 0.9847

+ MSE: 2.4961

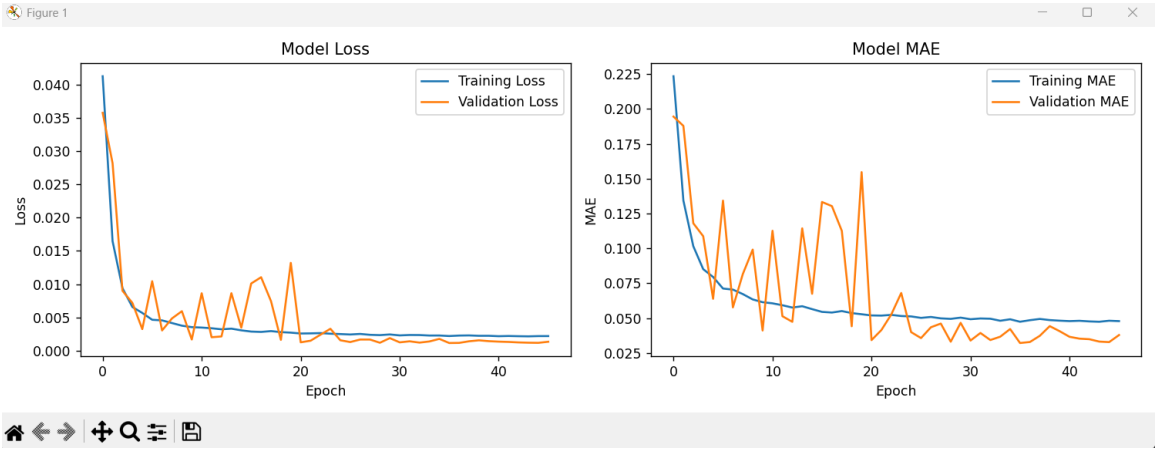
+ RMSE: 1.5793

+ R^2: 0.9804

Đây là các kết quả rất tốt, thể hiện mô hình có khả năng dự đoán chính xác cao, thậm chí vượt qua các mô hình học máy như XGBoost về độ khái quát hóa.

- Phân tích biểu đồ huấn luyện:

Hình ảnh dưới đây thể hiện quá trình huấn luyện mô hình qua 50 epoch, bao gồm biểu đồ Loss và MAE (Mean Absolute Error) cho cả tập huấn luyện và tập kiểm tra (validation):



Hình 2.13 Biểu đồ huấn luyện LSTM

- Biểu đồ bên trái (Model Loss):

  + Đường Loss trên tập huấn luyện (màu xanh) và tập kiểm tra (màu cam) đều giảm mạnh trong 10 epoch đầu tiên và dần hội tụ về gần 0.

  + Loss validation có một số dao động nhẹ ở giữa quá trình học, nhưng nhìn chung vẫn duy trì xu hướng giảm và ổn định.

  + Điều này cho thấy mô hình học được xu hướng tốt từ dữ liệu và không xảy ra overfitting nghiêm trọng.

- Biểu đồ bên phải (Model MAE):

  + MAE của tập huấn luyện giảm đều và mượt mà, phản ánh quá trình tối ưu hóa hiệu quả.

  + MAE validation giảm mạnh trong những epoch đầu và sau đó dao động, tuy nhiên vẫn có xu hướng giảm dần về cuối.

  + Mặc dù validation MAE dao động, nhưng không có dấu hiệu tăng dài hạn, cho thấy mô hình vẫn duy trì khả năng tổng quát hóa tốt.

Biểu đồ huấn luyện thể hiện rõ ràng rằng mô hình hội tụ tốt, có thể học được từ dữ liệu một cách hiệu quả mà không bị overfitting. Những dao động nhẹ trong MAE và Loss validation là hoàn toàn bình thường trong các mô hình học sâu với dữ liệu thực tế, đặc biệt là chuỗi thời gian phức tạp.

e. Dự báo tương lai (2025–2029)

Tận dụng kiến trúc chuỗi thời gian, mô hình LSTM được sử dụng để dự báo tỷ lệ thất nghiệp toàn cầu cho giai đoạn 2025–2029.

- Dự đoán được thực hiện cho từng tổ hợp quốc gia, giới tính và nhóm tuổi.
- Áp dụng ensemble prediction với nhiều mức độ noise giúp tăng độ tin cậy.
- Các kết quả dự đoán được chuyển về tỷ lệ gốc và lưu trữ.
- Biểu đồ minh họa kết quả cho thấy xu hướng dự đoán mượt mà và hợp lý, đồng thời cho phép đánh giá khoảng dao động bằng khoảng tin cậy ±10%.

f. Lưu trữ và hiển thị kết quả

Các kết quả dự đoán được lưu vào file lstm_predictions.csv.
Biểu đồ trực quan thể hiện xu hướng tỷ lệ thất nghiệp trung bình toàn cầu qua các năm, cho thấy sự suy giảm hoặc ổn định dần ở nhiều khu vực.

2.3. So sánh tổng quan hiệu suất các mô hình

Bảng 2.7 Hiệu suất các mô hình

Mô hình	MAE	RMSE	R^2
Linear Regression	7.0459	9.8713	0.1566
Random Forest	2.0634	3.0448	0.9198
XGBoost	1.1703	1.8047	0.9718
LSTM	0.9847	1.5793	0.9804

Quá trình nghiên cứu đã tiến hành thử nghiệm và đánh giá hiệu suất của bốn mô hình hồi quy, gồm: Linear Regression, Random Forest, XGBoost và LSTM. Việc so sánh được thực hiện trên các chỉ số đánh giá phổ biến như MAE, RMSE và R^2,

nhằm xác định mô hình phù hợp nhất cho bài toán dự đoán tỉ lệ thất nghiệp từ dữ liệu kinh tế - xã hội mang tính chuỗi thời gian.

- Linear Regression: Linear Regression được sử dụng như một mô hình cơ sở (baseline) để đánh giá hiệu quả của các mô hình phức tạp hơn. Tuy nhiên, với R^2 chỉ đạt 0.1566 và RMSE lên tới 9.8713, mô hình này cho thấy không đủ khả năng nắm bắt mối quan hệ phi tuyến giữa các đặc trưng đầu vào và đầu ra khiến Linear Regression không phù hợp để nắm bắt các quy luật phức tạp.
- Random Forest Regressor: Random Forest đạt $R^2 = 0.9198$ và RMSE = 3.0448, thể hiện hiệu suất khá tốt trong việc dự đoán và giảm phương sai nhờ vào kỹ thuật bagging và huấn luyện nhiều cây quyết định độc lập. Việc kết hợp kết quả từ nhiều cây giúp mô hình ổn định và ít bị overfitting. Tuy nhiên, do các cây học độc lập với nhau, mô hình không tận dụng được thông tin về sai số của các cây trước để cải thiện kết quả. Điều này khiến Random Forest bị giới hạn trong khả năng học các mối quan hệ phi tuyến phức tạp và chưa tối ưu trong các bài toán cần độ chính xác cao.
- XGBoost Regressor: cho kết quả nổi bật với $R^2 = 0.9718$ và RMSE = 1.8047, cho thấy mô hình giải thích được đến hơn 97% phương sai của dữ liệu và dự đoán gần sát với giá trị thực tế. Khác với Random Forest, XGBoost là một mô hình boosting hiện đại, huấn luyện các cây theo cách tuần tự, trong đó mỗi cây mới sẽ học từ sai số còn lại của tổ hợp các cây trước đó. Nhờ vậy, mô hình liên tục cải thiện hiệu quả dự đoán qua từng bước.
- Bên cạnh đó, XGBoost còn tích hợp nhiều kỹ thuật tối ưu như regularization để kiểm soát độ phức tạp, shrinkage (learning rate) để cập nhật mô hình một cách thận trọng, và song song hóa giúp tăng tốc quá trình huấn luyện. Những đặc điểm này giúp XGBoost vượt trội về khả năng tổng quát hóa, kiểm soát overfitting và đặc biệt phù hợp với các bài toán có độ phức tạp cao.

Tuy nhiên, XGBoost cũng tồn tại một số nhược điểm nhất định. Do có nhiều siêu tham số quan trọng như learning rate, max depth, số lượng cây,… nên việc tinh chỉnh mô hình yêu cầu nhiều thời gian và kinh nghiệm. Bên cạnh đó, nếu không điều

chỉnh phù hợp, mô hình có thể bị overfitting, đặc biệt khi học quá sâu hoặc áp dụng trên tập dữ liệu nhỏ. Do đó, XGBoost phát huy hiệu quả nhất khi được sử dụng kèm theo quy trình chọn tham số tối ưu và đánh giá chéo chặt chẽ.

LSTM: Mạng nơ-ron hồi quy cho chuỗi thời gian: Đặc biệt, mô hình LSTM đã vượt trội hơn hẳn so với tất cả các mô hình còn lại và đạt hiệu suất tốt nhất. Với MAE thấp nhất (0.9847), RMSE thấp nhất (1.5793) và R2 cao nhất (0.9804), LSTM không chỉ giải thích được 98.04% phương sai của biến mục tiêu mà còn cho thấy sai số dự đoán rất nhỏ. Điểm mạnh của LSTM đến từ cấu trúc mạng nơ-ron hồi quy có bộ nhớ dài hạn, cho phép lưu giữ và truyền tải thông tin từ nhiều bước thời gian trước đó trong chuỗi dữ liệu. Điều này giúp LSTM đặc biệt phù hợp để xử lý dữ liệu chuỗi thời gian có tính phụ thuộc phức tạp, như các hiện tượng tuần hoàn, xu hướng thay đổi theo mùa, hoặc các yếu tố trễ thời gian ảnh hưởng lâu dài đến giá trị tương lai.

Khác với các mô hình truyền thống như Linear Regression, Random Forest hay XGBoost, vốn thường giả định các quan hệ tuyến tính hoặc không có khả năng ghi nhớ thông tin quá khứ xa, LSTM có thể học và tổng hợp các mẫu phụ thuộc phi tuyến tính trong dữ liệu. Nhờ vậy, mô hình này có thể dự đoán chính xác hơn với các chuỗi dữ liệu có tính thời gian đặc thù và phức tạp.

Ngoài ra, LSTM còn giảm thiểu vấn đề mất mát thông tin do gradient biến mất (vanishing gradient) thường gặp ở mạng hồi quy truyền thống, nhờ cơ chế cổng điều khiển thông tin (gates) trong từng đơn vị nhớ. Đây chính là lý do giúp LSTM không chỉ duy trì hiệu quả học tập trên các chuỗi dữ liệu dài mà còn thích ứng tốt với những đặc điểm riêng biệt của dữ liệu thời gian trong bài toán.

Lựa chọn mô hình cuối cùng:

Dựa trên kết quả đánh giá các chỉ số hiệu suất, mô hình LSTM và XGBoost đều thể hiện hiệu quả vượt trội so với các mô hình truyền thống khác như Linear Regression và Random Forest. Tuy nhiên, khi so sánh trực tiếp giữa hai mô hình này, ta nhận thấy những điểm khác biệt rõ ràng trong hiệu suất dự báo.

+ LSTM có MAE = 0.9847, thấp hơn XGBoost (MAE = 1.1703) khoảng 0.1856,

+ RMSE của LSTM là 1.5793, giảm 0.2254 so với XGBoost (1.8047),

+ R^2 của LSTM đạt 0.9804, cao hơn XGBoost (0.9718) khoảng 0.0086.

Những con số này cho thấy LSTM cải thiện rõ rệt về sai số tuyệt đối và sai số bình phương trung bình so với XGBoost, đồng thời có khả năng giải thích phương sai của dữ liệu tốt hơn. Điều này khẳng định LSTM phù hợp hơn trong việc nắm bắt các đặc điểm phức tạp và các phụ thuộc dài hạn trong dữ liệu chuỗi thời gian.

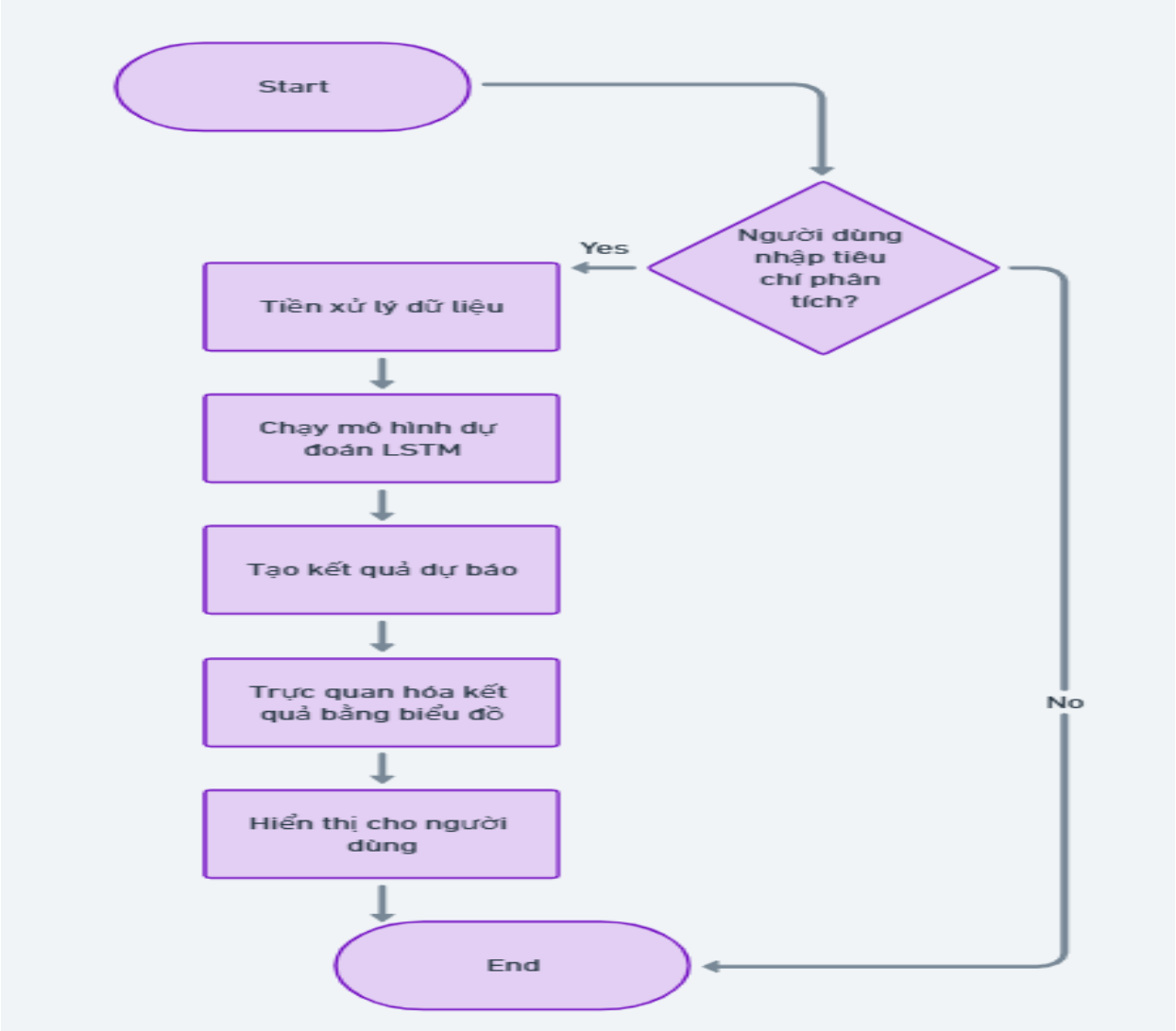
Mặc dù XGBoost cũng là mô hình mạnh với hiệu suất rất tốt, sự chênh lệch về các chỉ số hiệu suất cho thấy LSTM đáng tin cậy hơn để đưa ra các dự báo chính xác và ổn định. Vì vậy, LSTM được lựa chọn làm mô hình cuối cùng để áp dụng trong nghiên cứu cũng như tích hợp vào website hỗ trợ dự đoán tỉ lệ thất nghiệp.

CHƯƠNG 3. XÂY DỰNG HỆ THỐNG DỰ ĐOÁN VÀ GIAO DIỆN NGƯỜI DÙNG

3.1 Yêu cầu hệ thống

3.1.1. Chức năng chính của hệ thống

3.1.1.1 Chức năng dự đoán tỉ lệ thất nghiệp



Hình 3.1 Sơ đồ kiến trúc hệ thống

Chức năng dự đoán tỷ lệ thất nghiệp trong hệ thống được phát triển như một công cụ phân tích nâng cao, hỗ trợ người dùng đánh giá xu hướng thất nghiệp trong quá khứ, hiện tại và tương lai. Tính năng này được thiết kế với tính linh hoạt cao, cho phép người dùng dễ dàng tùy chỉnh các tiêu chí phân tích để phù hợp với nhu cầu nghiên cứu hoặc mục tiêu cụ thể của từng lĩnh vực. Đây là công cụ đặc biệt hữu

ích đối với các nhà nghiên cứu, nhà hoạch định chính sách, chuyên gia kinh tế, doanh nghiệp, tổ chức phi chính phủ, cũng như các cá nhân quan tâm đến sự phát triển của thị trường lao động.

Ngay từ giao diện chính, người dùng có thể lựa chọn một quốc gia cụ thể để phân tích, sau đó tiếp tục chọn giới tính - bao gồm hai lựa chọn chính là nam hoặc nữ. Ngoài ra, người dùng còn có thể lọc dữ liệu theo nhóm độ tuổi, với ba nhóm tuổi chính:

+ Từ 15 đến 24 tuổi – nhóm tuổi thanh niên, đang trong giai đoạn học tập, đào tạo hoặc bước đầu tham gia thị trường lao động.

+ Từ 25 tuổi trở lên – nhóm người trưởng thành, đang trong độ tuổi lao động ổn định.

+ Dưới 15 tuổi – nhóm tuổi trẻ em, tuy không trực tiếp tham gia lao động chính thức nhưng vẫn được đưa vào phân tích để nắm bắt xu hướng trong các quốc gia có tình trạng lao động sớm hoặc bất hợp pháp.

Sau khi người dùng hoàn tất việc thiết lập các tiêu chí phân tích và nhấn vào nút "Dự đoán", hệ thống sẽ tiến hành xử lý dữ liệu đầu vào và chạy mô hình dự báo dựa trên mô hình học sâu LSTM (Long Short-Term Memory) – một dạng mạng nơ-ron tích chập chuyên phân tích chuỗi thời gian với độ chính xác cao. Việc ứng dụng LSTM cho phép hệ thống nắm bắt được cả các xu hướng dài hạn và các biến động ngắn hạn trong dữ liệu, từ đó tạo ra kết quả dự báo có tính tin cậy và giá trị tham khảo cao.

Kết quả trả về được hiển thị thông qua biểu đồ trực quan, giúp người dùng dễ dàng theo dõi và so sánh giữa các giai đoạn. Biểu đồ gồm hai phần rõ rệt:

+ Phần dữ liệu lịch sử từ năm 2014 đến năm 2024 – cho phép người dùng quan sát diễn biến thực tế của tỷ lệ thất nghiệp trong một thập kỷ qua, từ đó phát hiện những biến động do khủng hoảng kinh tế, thay đổi chính sách, dịch bệnh, hoặc các yếu tố toàn cầu ảnh hưởng đến thị trường lao động.

+ Phần dự báo từ năm 2025 đến năm 2029 – cung cấp cái nhìn định hướng về xu hướng thất nghiệp trong tương lai, là công cụ hữu hiệu để xây dựng chiến lược phát triển, điều chỉnh chính sách, hoặc chuẩn bị cho các biến động tiềm ẩn.

+ Các biểu đồ được thiết kế theo tiêu chuẩn trực quan hóa dữ liệu hiện đại, với khả năng tương tác như phóng to, di chuột để xem chi tiết giá trị, hoặc chuyển đổi giữa các loại biểu đồ để phục vụ các mục đích so sánh khác nhau.

Chức năng này không chỉ đơn thuần cung cấp thông tin, mà còn giúp người dùng ra quyết định một cách có cơ sở khoa học, bằng cách kết hợp giữa dữ liệu thực tế và mô hình dự đoán mạnh mẽ. Từ đó, người dùng có thể:

+ Đánh giá hiệu quả của các chính sách kinh tế – xã hội đã áp dụng trong quá khứ.

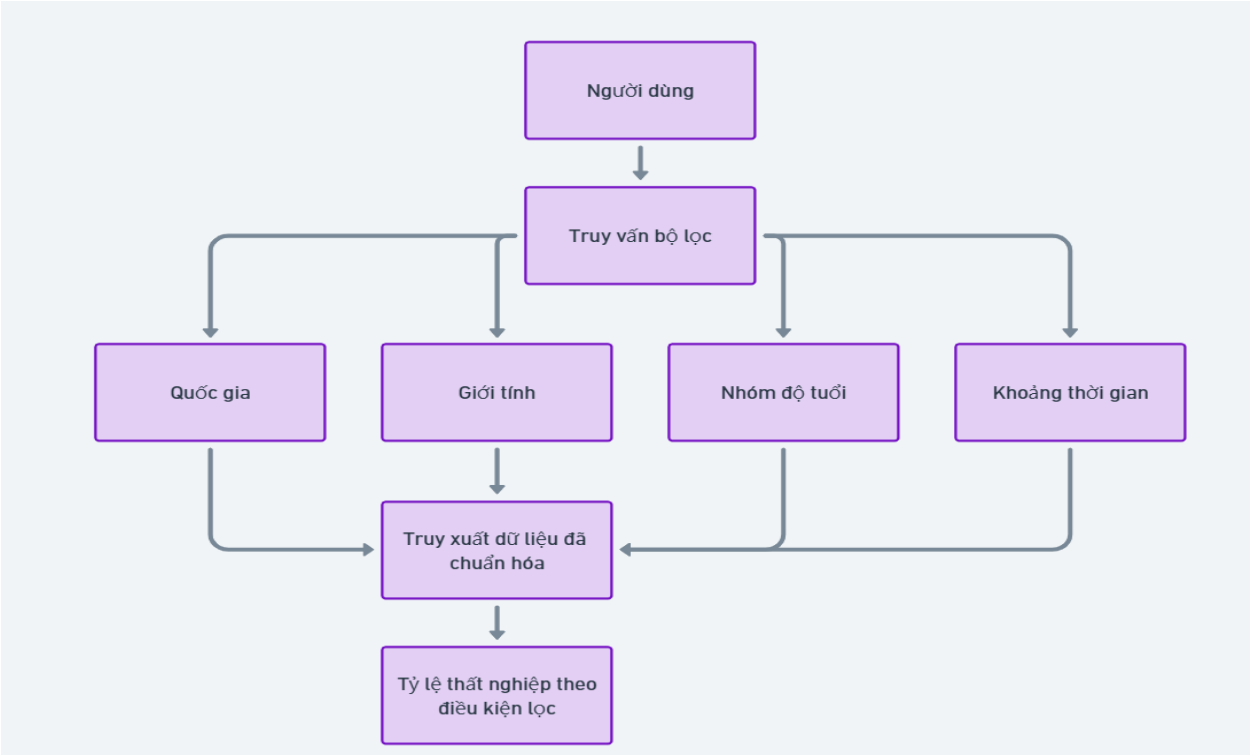
+ Xây dựng kế hoạch lao động – việc làm trong tương lai.

+ Xác định nhóm dân số có nguy cơ thất nghiệp cao để đưa ra giải pháp can thiệp sớm.

+ Phân tích so sánh giữa các quốc gia trong cùng khu vực hoặc theo trình độ phát triển.

Với vai trò là một trong những tính năng cốt lõi của hệ thống, chức năng dự đoán tỷ lệ thất nghiệp không chỉ hỗ trợ phân tích dữ liệu một cách chi tiết, mà còn góp phần vào việc thúc đẩy quy hoạch phát triển kinh tế – xã hội bền vững, từ cấp độ quốc gia đến toàn cầu.

3.1.1.2 Chức năng phân tích dữ liệu



Hình 3.2 Sơ đồ cấu trúc dữ liệu

Chức năng Phân tích dữ liệu là một trong những thành phần then chốt của hệ thống, đóng vai trò như một công cụ phân tích chuyên sâu, cho phép người dùng trực tiếp truy cập, quan sát và so sánh tỷ lệ thất nghiệp giữa nhiều quốc gia trong giai đoạn từ năm 2014 đến 2024. Tính năng này hỗ trợ hiệu quả cho quá trình nghiên cứu và đánh giá dữ liệu kinh tế – xã hội một cách trực quan, chính xác và có chiều sâu.

- Giao diện tương tác và bộ lọc linh hoạt

Ngay khi truy cập vào chức năng phân tích, người dùng được cung cấp một bộ lọc động, cho phép chọn nhiều điều kiện kết hợp gồm:

+ Quốc gia

+ Nhóm độ tuổi

+ Giới tính

+ Khoảng thời gian

Điều này giúp người dùng tùy biến dữ liệu hiển thị theo nhu cầu nghiên cứu cụ thể

Hệ thống sẽ xử lý truy vấn bằng cách trích xuất dữ liệu từ tập dữ liệu đã được huấn luyện và chuẩn hóa trước đó, nhằm tối ưu hiệu suất và đảm bảo độ chính xác cao. Quá trình xử lý diễn ra hoàn toàn tự động và có tốc độ phản hồi nhanh, ngay cả với các truy vấn dữ liệu phức tạp.

- Phân biệt giữa phân tích động và báo cáo tĩnh:

+ Phân tích dữ liệu động:

Người dùng tương tác trực tiếp với giao diện, điều chỉnh bộ lọc và thay đổi các điều kiện để quan sát tức thời sự thay đổi của biểu đồ. Biểu đồ hiển thị có thể tương tác như rê chuột để xem chi tiết giá trị từng năm.

+ Phân tích báo cáo tĩnh:

Dữ liệu được tổng hợp và trình bày dưới dạng báo cáo cố định, không cho phép tương tác hoặc thay đổi điều kiện lọc.

+ Hiển thị biểu đồ trực quan

Sau khi người dùng hoàn tất các lựa chọn lọc, hệ thống sẽ hiển thị hai loại biểu đồ chính:

- Biểu đồ đường (Line Chart): Thể hiện xu hướng biến động tỷ lệ thất nghiệp qua các năm.
- Biểu đồ cột (Bar Chart): So sánh dữ liệu giữa các quốc gia tại từng thời điểm cụ thể.
- Cả hai biểu đồ đều hỗ trợ khả năng tương tác, cho phép người dùng xem giá trị cụ thể của từng điểm dữ liệu, đồng thời được thiết kế với giao diện rõ ràng, tối ưu hóa trải nghiệm người dùng.

Bảo đảm tính toàn vẹn dữ liệu hiển thị

Một ưu điểm nổi bật khác là hệ thống đảm bảo rằng mọi thông tin hiển thị đều đúng theo lựa chọn người dùng. Mỗi biểu đồ, mỗi thống kê đều được liên kết trực

tiếp với điều kiện lọc đã xác định, tránh tình trạng nhầm lẫn giữa dữ liệu các quốc gia hoặc nhóm tuổi khác nhau.

Giao diện và trải nghiệm người dùng

Giao diện người dùng được thiết kế đáp ứng trên nhiều nền tảng, từ máy tính để bàn đến máy tính xách tay, hỗ trợ người dùng trong việc:

- Tương tác linh hoạt với dữ liệu
- Trích xuất thông tin dễ dàng
- Phục vụ nhiều mục đích sử dụng khác nhau như: nghiên cứu học thuật, đánh giá thị trường lao động, xây dựng chính sách hoặc báo cáo nội bộ.

Chức năng phân tích dữ liệu không chỉ đơn thuần là một công cụ thống kê mà còn là một nền tảng phân tích chuyên sâu, phục vụ hiệu quả cho:

- Nghiên cứu học thuật
- Phân tích thị trường lao động
- Xây dựng và đánh giá chính sách công
- Đào tạo và giảng dạy trong lĩnh vực kinh tế – xã hội

Với dữ liệu được chuẩn hóa từ các nguồn tin cậy, đây là công cụ hữu ích cho nhà nghiên cứu, nhà hoạch định chính sách, chuyên gia phân tích dữ liệu, sinh viên và các tổ chức quốc tế có nhu cầu nắm bắt tình hình thất nghiệp toàn cầu hoặc khu vực theo cách tiếp cận dữ liệu hiện đại và khoa học.

3.1.1.3 Chức năng báo cáo nhanh

Chức năng báo cáo nhanh trong hệ thống được thiết kế nhằm cung cấp cho người dùng một cái nhìn tổng quan, chính xác và dễ hiểu về tình hình thất nghiệp tại một quốc gia cụ thể trong giai đoạn từ năm 2014 đến năm 2024. Đây là công cụ rất tiện lợi, giúp người dùng nhanh chóng nắm bắt các thông tin quan trọng mà không cần thực hiện quá nhiều thao tác phức tạp hay xử lý dữ liệu thủ công.

Khi truy cập vào chức năng này, người dùng chỉ cần chọn một quốc gia mà họ quan tâm từ danh sách có sẵn. Sau đó, khi nhấn vào nút “Báo cáo”, hệ thống sẽ tự động tổng hợp và phân tích dữ liệu thất nghiệp liên quan đến quốc gia đó. Kết quả

phân tích sẽ được trình bày một cách trực quan thông qua biểu đồ đường, trong đó thể hiện rõ tỷ lệ thất nghiệp của các nhóm độ tuổi khác nhau qua từng năm. Biểu đồ giúp người dùng dễ dàng theo dõi và so sánh sự biến động giữa các nhóm như: dưới 15 tuổi, từ 15 đến 24 tuổi, và từ 25 tuổi trở lên trong vòng 10 năm qua.

Ngoài việc trình bày dữ liệu lịch sử một cách trực quan, chức năng báo cáo nhanh còn tích hợp các chỉ số phân tích chuyên sâu, trong đó bao gồm:

- Nhóm tuổi chịu ảnh hưởng nhiều nhất: Hệ thống sẽ tự động xác định nhóm tuổi có tỷ lệ thất nghiệp cao nhất, được tính theo hai phương pháp:
  + Trung bình 3 năm – để xác định nhóm chịu ảnh hưởng ổn định và lâu dài nhất.
  + Giá trị cao nhất – để phát hiện các đợt khủng hoảng hoặc biến động đột ngột trong thời gian ngắn.

Phân tích xu hướng thất nghiệp giữa các nhóm tuổi: Hệ thống sẽ đưa ra đánh giá tổng quan về việc tỷ lệ thất nghiệp trong từng nhóm tuổi đang có xu hướng tăng hay giảm qua các năm. Nhờ đó, người dùng có thể xác định được đâu là nhóm đang cải thiện tích cực và đâu là nhóm có nguy cơ cao trong tương lai.

Độ biến thiên của xu hướng thất nghiệp: Thông qua việc đo lường mức độ dao động của tỷ lệ thất nghiệp trong mỗi nhóm tuổi, hệ thống cung cấp một cái nhìn sâu hơn về mức độ ổn định hay bất ổn của thị trường lao động theo từng phân khúc dân số. Việc này đặc biệt quan trọng để nhận diện các nhóm tuổi dễ bị tổn thương trước các biến động kinh tế – xã hội.

Một điểm nổi bật khác của chức năng báo cáo nhanh là khả năng xuất dữ liệu và biểu đồ thành định dạng PDF. Với tính năng này, người dùng có thể nhanh chóng tải về báo cáo, in ấn hoặc chia sẻ cho các đồng nghiệp, đối tác hoặc cấp quản lý trong các cuộc họp, hội thảo hoặc báo cáo nội bộ. Tài liệu PDF được trình bày khoa học, rõ ràng, có thể sử dụng ngay mà không cần chỉnh sửa thêm, rất thuận tiện cho việc lưu trữ và trích dẫn. Việc xuất báo cáo PDF được thực hiện nhờ vào thư viện ReportLab, một công cụ mạnh mẽ trong ngôn ngữ lập trình Python. Thư viện này

cho phép tạo ra các tài liệu PDF một cách linh hoạt và chuyên nghiệp. Một số đặc điểm nổi bật khi sử dụng ReportLab bao gồm:

- Hỗ trợ tùy biến bố cục và định dạng tài liệu phù hợp với nhu cầu trình bày học thuật hoặc chuyên môn.
- Khả năng chèn biểu đồ vector chất lượng cao, đảm bảo độ rõ nét khi in ấn hoặc phóng to.
- Tích hợp các thành phần động như bảng số liệu, biểu đồ và nhận xét phân tích, giúp báo cáo trở nên sinh động và trực quan.
- Toàn bộ quá trình tạo PDF được tự động hóa hoàn toàn, đảm bảo tính nhất quán, tiết kiệm thời gian và giảm thiểu sai sót khi xuất tài liệu.

Chức năng báo cáo nhanh không chỉ giúp người dùng tiết kiệm thời gian trong việc phân tích dữ liệu thất nghiệp mà còn mang lại những thông tin chuyên sâu, có giá trị thực tiễn cao. Đây là công cụ lý tưởng dành cho các nhà phân tích, nhà quản lý nhân sự, cơ quan chính phủ, tổ chức quốc tế, nhà nghiên cứu, sinh viên và bất kỳ ai cần đưa ra quyết định hoặc đánh giá nhanh chóng về tình hình lao động tại một quốc gia cụ thể.

3.1.2. Kiến trúc hệ thống và công nghệ sử dụng

3.1.2.1 Kiến trúc hệ thống

Hệ thống dự đoán và phân tích tỷ lệ thất nghiệp được xây dựng theo kiến trúc phân lớp, bao gồm ba thành phần chính: Frontend, Backend, và Machine Learning. Kiến trúc này giúp đảm bảo tính mở rộng, dễ bảo trì và khả năng tái sử dụng mã nguồn.

Cụ thể, hệ thống được tổ chức theo mô hình MVC:

- Model: Bao gồm các mô hình học máy, tập tin mã hóa (encoder), chuẩn hóa dữ liệu và logic xử lý dự đoán.
- View: Là giao diện người dùng, được xây dựng bằng HTML, CSS và JavaScript, hiển thị biểu đồ, báo cáo và kết quả dự đoán.

- Controller: Là phần trung gian xử lý yêu cầu từ người dùng, tương tác với mô hình và trả kết quả thông qua các route Flask.

Ưu điểm của hệ thống:

- Tách biệt rõ ràng giữa các lớp chức năng:

Hệ thống được thiết kế theo mô hình phân lớp rõ ràng giữa các thành phần: Frontend, Controller, và (Backend/Model). Sự phân tách này giúp cho mỗi thành phần hoạt động độc lập, dễ dàng kiểm tra và bảo trì mà không làm ảnh hưởng đến toàn bộ hệ thống. Việc thay đổi, nâng cấp hoặc sửa lỗi cho một phần cụ thể có thể được thực hiện mà không cần can thiệp đến các phần còn lại. Điều này không chỉ nâng cao khả năng bảo trì mà còn tăng độ linh hoạt trong quá trình phát triển và triển khai hệ thống.

Dễ dàng mở rộng với nhiều mô hình dự đoán khác nhau:

- Các mô hình dự đoán học máy trong hệ thống được tổ chức riêng biệt và lưu trữ dưới dạng file.joblib. Nhờ đó, hệ thống hoàn toàn có khả năng tích hợp thêm các mô hình mới (như XGBoost, Random Forest, Linear Regression hoặc LSTM) mà không làm ảnh hưởng đến luồng hoạt động chung. Việc bổ sung mô hình chỉ cần lưu trữ file mô hình mới và cập nhật nhẹ trong logic xử lý. Ngoài ra, các bộ mã hóa đầu vào (encoder) như giới tính, độ tuổi, tên quốc gia cũng được lưu riêng biệt, giúp chuẩn hóa dữ liệu đầu vào một cách linh hoạt cho nhiều mô hình khác nhau. Cách tổ chức này rất phù hợp để mở rộng quy mô hệ thống trong tương lai, tích hợp thêm các chỉ số kinh tế khác hoặc áp dụng cho nhiều khu vực địa lý hơn.

Giao diện người dùng trực quan, thân thiện:

- Hệ thống có giao diện web đơn giản nhưng hiệu quả, giúp người dùng dễ dàng thao tác mà không cần có kiến thức chuyên sâu về công nghệ thông tin. Các biểu đồ trực quan như biểu đồ đường, biểu đồ cột được tích hợp giúp thể hiện thông tin một cách sinh động, hỗ trợ người dùng nhanh chóng nắm bắt xu hướng và đặc điểm dữ liệu. Việc tổ chức giao diện thành các chức năng riêng

biệt như: Dự đoán, Phân tích dữ liệu, và Báo cáo nhanh, giúp nâng cao trải nghiệm người dùng và tăng tính dễ sử dụng cho nhiều nhóm đối tượng như nhà nghiên cứu, nhà hoạch định chính sách hoặc sinh viên.

Hỗ trợ xuất báo cáo chuyên nghiệp dưới định dạng PDF:

- Một trong những tính năng nổi bật của hệ thống là khả năng tạo báo cáo nhanh và xuất ra định dạng PDF. Điều này không chỉ giúp người dùng lưu trữ thông tin để tham khảo hoặc chia sẻ, mà còn hỗ trợ các nhu cầu trình bày trong báo cáo nghiên cứu, thuyết trình hoặc phân tích nội bộ. Báo cáo PDF bao gồm đầy đủ các phân tích tỷ lệ thất nghiệp theo nhóm tuổi, các thống kê trung bình và cực trị, cùng với biểu đồ minh họa trực quan. Tính năng này giúp nâng cao tính chuyên nghiệp và thực tiễn của hệ thống, đáp ứng nhu cầu sử dụng trong nhiều môi trường khác nhau từ học thuật đến công nghiệp.

3.1.2.2 Công nghệ sử dụng

- Ngôn ngữ lập trình Python – Ngôn ngữ lập trình chính:

Python đóng vai trò là ngôn ngữ lập trình chính cho toàn bộ hệ thống backend và xử lý mô hình học máy nhờ các ưu điểm nổi bật:

+ Cú pháp đơn giản, dễ đọc: Giúp rút ngắn thời gian phát triển, dễ bảo trì mã nguồn và thuận tiện cho cộng tác nhóm.

+ Hệ sinh thái thư viện phong phú: Python sở hữu nhiều thư viện chuyên biệt cho xử lý dữ liệu, học máy, trực quan hóa và phát triển web.

+ pandas và numpy: Xử lý và thao tác dữ liệu bảng, chuỗi thời gian, và thực hiện các phép toán trên mảng số liệu.

+ scikit-learn: Cung cấp các mô hình học máy truyền thống (hồi quy tuyến tính, Random Forest, XGBoost) và hỗ trợ các công cụ như chuẩn hóa dữ liệu, mã hóa biến phân loại, phân chia tập huấn luyện/kiểm tra.

+ tensorflow / keras: Xây dựng và huấn luyện mô hình học sâu (Deep Learning), đặc biệt là LSTM cho phân tích chuỗi thời gian và dự đoán xu hướng thất nghiệp.

+ joblib: Lưu trữ và tải nhanh các mô hình đã huấn luyện và bộ mã hóa dữ liệu, giảm thiểu thời gian tái huấn luyện và tăng khả năng tái sử dụng.

+ matplotlib và seaborn: Trực quan hóa dữ liệu và tạo biểu đồ tĩnh cho báo cáo PDF hoặc nghiên cứu.

- Hệ thống sử dụng Flask để phát triển phần backend:

Flask được đánh giá cao nhờ sự nhẹ, linh hoạt và dễ mở rộng, phù hợp cho các ứng dụng web quy mô vừa và nhỏ nhưng có tính tùy biến cao như hệ thống dự đoán và phân tích dữ liệu thất nghiệp.

Flask đóng vai trò trung gian kết nối giữa giao diện người dùng (frontend) và các mô hình xử lý ở backend:

- Tạo và quản lý các API endpoint: Flask cho phép định nghĩa các route (đường dẫn URL) rõ ràng tương ứng với từng chức năng như: dự đoán, khám phá dữ liệu, xuất báo cáo, ...
- Xử lý request và trả về response: Khi người dùng thực hiện một hành động như nhập thông tin và bấm nút “Dự đoán”, Flask sẽ tiếp nhận dữ liệu đầu vào, chuyển đến mô hình học máy xử lý, sau đó gửi kết quả trả về trình duyệt ở dạng số liệu hoặc biểu đồ.
- Tích hợp linh hoạt với công nghệ khác: Flask có thể dễ dàng kết hợp với các công nghệ frontend như HTML, CSS, JavaScript, và hỗ trợ xuất kết quả dưới định dạng PDF hoặc JSON. Đồng thời, nó cũng hỗ trợ kết nối với cơ sở dữ liệu như SQLite hoặc các hệ quản trị khác (MySQL, PostgreSQL nếu cần mở rộng trong tương lai).
- Triển khai đơn giản: Nhờ thiết kế gọn nhẹ, Flask không yêu cầu cấu trúc thư mục phức tạp, giúp dễ dàng triển khai ứng dụng trên các môi trường máy chủ như Heroku, AWS, hoặc máy tính cục bộ.

Việc sử dụng Flask mang lại nhiều lợi ích: giảm thời gian phát triển, tăng độ linh hoạt và dễ bảo trì, đồng thời giúp hệ thống duy trì tính đơn giản mà vẫn đảm bảo đủ chức năng và khả năng mở rộng về sau.

Công cụ và thư viện học máy:

Trong hệ thống dự đoán tỷ lệ thất nghiệp, một loạt các công cụ và thư viện học máy đã được sử dụng để xây dựng, huấn luyện, đánh giá và triển khai các mô hình dự đoán. Những công cụ này không chỉ giúp cải thiện hiệu quả xử lý dữ liệu và độ chính xác của dự đoán, mà còn đảm bảo tính mở rộng và tái sử dụng linh hoạt trong tương lai.

Scikit-learn:

Scikit-learn là một thư viện học máy phổ biến trong Python, được sử dụng để triển khai các mô hình học máy cổ điển như:

- Random Forest: Thuật toán học máy dựa trên nhiều cây quyết định, giúp tăng cường độ chính xác và chống overfitting.
- Linear Regression: Mô hình hồi quy tuyến tính đơn giản nhưng hiệu quả trong việc ước lượng mối quan hệ giữa các biến đầu vào và đầu ra.
- XGBoost (Extreme Gradient Boosting): Một kỹ thuật boosting hiện đại, được tối ưu hóa về hiệu năng và độ chính xác.

Ngoài ra, scikit-learn còn cung cấp nhiều công cụ hỗ trợ như chuẩn hóa dữ liệu (standardization), mã hóa biến phân loại (encoding), phân chia tập dữ liệu, đánh giá mô hình bằng các chỉ số như RMSE, MAE, R^2, …

TensorFlow / Keras

Hệ thống sử dụng TensorFlow kết hợp với Keras để xây dựng mô hình mạng nơ-ron hồi tiếp LSTM – một dạng kiến trúc đặc biệt trong mạng nơ-ron sâu, được thiết kế để phân tích và dự đoán chuỗi thời gian.

LSTM đặc biệt hiệu quả trong việc phát hiện các xu hướng dài hạn và xử lý hiện tượng phụ thuộc thời gian trong dữ liệu – điều này rất phù hợp với bài toán dự đoán tỷ lệ thất nghiệp theo năm. Mô hình LSTM cho phép hệ thống không chỉ đưa ra dự đoán chính xác mà còn phản ánh được xu hướng trong tương lai dựa trên dữ liệu quá khứ từ năm 2014 đến 2024.

Joblib

Joblib là một thư viện chuyên dùng để lưu trữ các mô hình đã huấn luyện, cũng như các đối tượng quan trọng như:

- Scaler (dùng để chuẩn hóa dữ liệu),
- Encoder (biến đổi biến phân loại thành dạng số),
- Mô hình học máy hoặc mô hình deep learning đã huấn luyện.

Việc đóng gói mô hình và các thành phần liên quan bằng joblib mang lại nhiều lợi ích:

- Tiết kiệm thời gian: Không cần huấn luyện lại mỗi lần chạy hệ thống.
- Tăng tính linh hoạt: Dễ dàng cập nhật, thay thế hoặc so sánh nhiều mô hình khác nhau.
- Tái sử dụng hiệu quả: Các mô hình có thể được tải vào nhiều môi trường khác nhau (local, server, cloud) mà không cần thay đổi cấu trúc hệ thống.

Thư viện xử lý dữ liệu và trực quan hóa:

- Để đảm bảo hệ thống có khả năng xử lý, phân tích và trình bày dữ liệu một cách trực quan, hiệu quả, nhiều thư viện Python và công cụ frontend đã được sử dụng. Các thư viện này hỗ trợ toàn diện từ giai đoạn tiền xử lý dữ liệu cho đến trực quan hóa kết quả đầu ra, bao gồm cả biểu đồ tĩnh trong báo cáo PDF và biểu đồ động trên giao diện người dùng.

Pandas và NumPy:

Hai thư viện lõi không thể thiếu trong hệ sinh thái Python về khoa học dữ liệu là Pandas và NumPy:

Pandas được sử dụng để:

- Đọc dữ liệu từ các tệp CSV, Excel hoặc API.
- Làm sạch, chuẩn hóa và tái cấu trúc dữ liệu theo yêu cầu của mô hình.
- Thao tác dữ liệu theo chiều thời gian, nhóm tuổi, giới tính và quốc gia.
- Tính toán thống kê như trung bình, phương sai, độ lệch chuẩn, v.v.

- NumPy đóng vai trò hỗ trợ xử lý mảng hiệu suất cao, tính toán đại số tuyến tính và tương tác trực tiếp với các thư viện học máy (TensorFlow yêu cầu đầu vào dạng numpy.array).

Nhờ sự kết hợp giữa Pandas và NumPy, dữ liệu đầu vào được xử lý nhanh chóng, linh hoạt và sẵn sàng cho việc huấn luyện mô hình, dự đoán hoặc xuất báo cáo.

Matplotlib và Seaborn:

Hai thư viện phổ biến dùng để trực quan hóa dữ liệu dưới dạng biểu đồ tĩnh:

Matplotlib:

- Là thư viện nền tảng, hỗ trợ vẽ hầu hết các loại biểu đồ cơ bản như biểu đồ đường, cột, scatter, histogram, ...
- Được sử dụng để tạo biểu đồ đầu ra trong báo cáo nhanh hoặc PDF, với khả năng tùy chỉnh màu sắc, nhãn, kích thước và định dạng hiển thị.

Seaborn:

- Được xây dựng dựa trên Matplotlib, cung cấp giao diện đơn giản và kiểu biểu đồ mặc định có tính thẩm mỹ cao.
- Thường được sử dụng cho phân tích thống kê, như hiển thị xu hướng thất nghiệp trung bình theo nhóm tuổi hoặc quốc gia.

Những biểu đồ này sau đó được lưu dưới dạng ảnh tĩnh (.png, .jpg) và chèn vào báo cáo PDF để cung cấp cái nhìn trực quan, dễ hiểu cho người dùng.

Chart.js (thông qua JavaScript)

Để phục vụ mục tiêu trực quan hóa dữ liệu tương tác ngay trên giao diện web, hệ thống sử dụng Chart.js – một thư viện vẽ biểu đồ nổi tiếng trên nền tảng JavaScript. Chart.js được tích hợp vào hệ thống thông qua mã HTML/JS được gọi từ Flask.

Các tính năng nổi bật của Chart.js được khai thác trong hệ thống gồm:

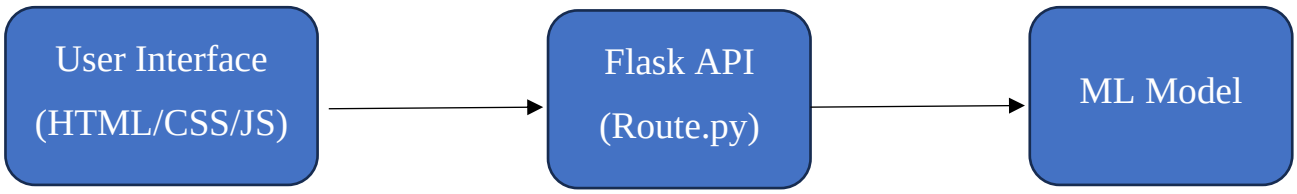
- Biểu đồ đường (line chart) : Thể hiện sự thay đổi tỷ lệ thất nghiệp qua từng năm.

- Biểu đồ cột (bar chart) : So sánh tỷ lệ thất nghiệp giữa các nhóm tuổi hoặc quốc gia.
- Hiệu ứng động (animation) : Tăng cường trải nghiệm người dùng khi quan sát dữ liệu biến đổi theo thời gian.
- Tính tương tác cao : Cho phép người dùng rê chuột để xem thông tin chi tiết tại từng điểm dữ liệu.

Giao diện người dùng:

- HTML5, CSS3: Để xây dựng và thiết kế các trang giao diện như prediction.html, explore.html, và quick_report.html.
- JavaScript: Dùng để xử lý tương tác động trên trình duyệt và hiển thị biểu đồ với Chart.js.
- Font và Style: Thư viện font DejaVuSerif và tệp CSS riêng được sử dụng để đảm bảo tính thẩm mỹ, dễ đọc, và nhất quán trong trình bày thông tin.

3.1.3. Mô hình triển khai API và giao diện



USER INTERFACE (Giao diện người dùng)

Công nghệ: HTML/JavaScript

Đây là nơi người dùng tương tác với hệ thống. Giao diện có thể xây dựng bằng các công nghệ web truyền thống (HTML/CSS/JS), giao diện đơn giản, nhanh chóng.

Chức năng:

- Nhận đầu vào từ người dùng (quốc gia, giới tính, nhóm tuổi, v.v.)
- Gửi yêu cầu HTTP đến backend Flask thông qua các endpoint như /predict, /report
- Hiển thị kết quả trực quan dưới dạng biểu đồ hoặc báo cáo

FLASK API (routes.py)

Thành phần: Flask (Python Web Framework)

- Đây là tầng xử lý trung gian giữa giao diện người dùng và mô hình dự đoán.

Chức năng:

- Nhận request từ frontend
- Xử lý dữ liệu đầu vào (gọi encoder, scaler)
- Gọi mô hình dự đoán và nhận kết quả
- Trả kết quả dưới dạng JSON hoặc HTML để frontend hiển thị

ML MODEL (.joblib)

Thành phần:

- Mô hình học sâu đã huấn luyện (LSTM)
- Các file encoder, scaler (dùng để chuẩn hóa và mã hóa dữ liệu đầu vào)

Chức năng:

- Dự đoán tỷ lệ thất nghiệp dựa trên đầu vào
- Trả về kết quả dự đoán đã được xử lý

3.2. Xây dựng API dự đoán

3.2.1. Triển khai mô hình AI lên server

Việc triển khai mô hình AI lên server nhằm tích hợp mô hình LSTM đã được huấn luyện vào một hệ thống API RESTful, cho phép người dùng gửi yêu cầu dự đoán tỷ lệ thất nghiệp trong tương lai (2025–2029) dựa trên dữ liệu lịch sử (2014–2024). Quá trình triển khai đảm bảo tính ổn định, hiệu suất cao, khả năng xử lý yêu cầu thời gian thực, và tính tương thích với các yêu cầu bảo mật và mở rộng trong môi trường sản xuất.

Công nghệ sử dụng:

- Ngôn ngữ lập trình: Python 3.9, được chọn vì tính phổ biến, hỗ trợ mạnh mẽ cho các thư viện học máy và phát triển API.
- Framework AI: TensorFlow 2.10.0, sử dụng để tải và vận hành mô hình LSTM đã được huấn luyện, lưu dưới định dạng .h5 (best_lstm_model.h5).

- Framework API: Flask 2.0.1, một framework nhẹ, linh hoạt, phù hợp cho việc xây dựng API RESTful với các yêu cầu đơn giản.

Thư viện hỗ trợ:

- Pandas 1.5.0: Xử lý dữ liệu dạng bảng từ file global_unemployment_data.csv.
- NumPy 1.23.0: Hỗ trợ tính toán số học và xử lý mảng đa chiều.
- Scikit-learn 1.2.0: Cung cấp các công cụ tiền xử lý dữ liệu như LabelEncoder (mã hóa biến phân loại) và MinMaxScaler (chuẩn hóa dữ liệu số).
- Matplotlib 3.5.0 và Seaborn 0.11.2: Tạo biểu đồ trực quan hóa kết quả dự đoán.
- Quy trình triển khai

Chuẩn bị mô hình AI

Mô hình LSTM đã được huấn luyện trước để dự đoán tỷ lệ thất nghiệp dựa trên dữ liệu lịch sử. Mô hình được lưu trữ dưới định dạng .h5 (best_lstm_model.h5), chứa các trọng số và kiến trúc mạng đã được tối ưu hóa. Đặc điểm của mô hình:

Đầu vào: Chuỗi dữ liệu 5 năm (sequence length = 5), mỗi điểm dữ liệu bao gồm 5 đặc trưng:

- Quốc gia (mã hóa bằng LabelEncoder).
- Giới tính (mã hóa bằng LabelEncoder).
- Nhóm tuổi (mã hóa bằng LabelEncoder).
- Năm (chuẩn hóa bằng MinMaxScaler).
- Tỷ lệ thất nghiệp (chuẩn hóa bằng MinMaxScaler).
- Đầu ra: Tỷ lệ thất nghiệp dự đoán (chuẩn hóa) cho năm tiếp theo.

Quy trình tiền xử lý:

- Dữ liệu đầu vào được mã hóa và chuẩn hóa để đảm bảo tương thích với mô hình.
- Hàm create_sequences trong file routes.py được sử dụng để tạo các chuỗi dữ liệu đầu vào cho LSTM:

```
def create_sequences(data, time_steps=5):
    """Tạo chuỗi dữ liệu cho LSTM"""
    X = []
    for i in range(len(data) - time_steps):
        X.append(data[i:(i + time_steps)])
    return np.array(X)
```

Hình 3.3 Tạo chuỗi dữ liệu cho LSTM

Tích hợp mô hình vào server

Mô hình LSTM được tích hợp vào server thông qua hàm init_models() trong file routes.py. Hàm này thực hiện các bước sau:

Dữ liệu tỷ lệ thất nghiệp từ file global_unemployment_data.csv được đọc và lưu vào bộ nhớ cache (df_cache) bằng cách sử dụng decorator @lru_cache để tối ưu hóa hiệu suất:

```
@lru_cache(maxsize=1)
def load_data():
    """Load and cache data"""
    return pd.read_csv("global_unemployment_data.csv")
```

Hình 3.4 Đọc dữ liệu từ csv

Tải mô hình:

Mô hình LSTM được tải bằng tensorflow.keras.models.load_model khi ứng dụng khởi động:

```
# Tải mô hình LSTM đã train
model = load_model('best_lstm_model.h5')
```

Hình 3.5 Mô hình được tải từ file ‘best_lstm_model.h5’

Quá trình tải được thực hiện một lần duy nhất khi ứng dụng khởi động để đảm bảo mô hình sẵn sàng cho các yêu cầu dự đoán.

Kết quả

- Mô hình LSTM được triển khai thành công trên server, sẵn sàng xử lý các yêu cầu dự đoán thông qua endpoint /predict.

- API nhận dữ liệu đầu vào dạng JSON (quốc gia, giới tính, nhóm tuổi), trả về kết quả dự đoán tỷ lệ thất nghiệp cho 5 năm (2025–2029) kèm khoảng tin cậy 98% và biểu đồ trực quan hóa mã hóa base64.
- Hiệu suất được tối ưu hóa nhờ caching dữ liệu (df_cache, @lru_cache) và sử dụng ensemble predictions với số mẫu giới hạn (num_samples=5).
- Ứng dụng hoạt động ổn định trong môi trường triển khai cục bộ

3.2.2. Xây dựng API kết nối giữa AI và giao diện

Nhằm triển khai mô hình học sâu (LSTM) dự đoán tỷ lệ thất nghiệp theo quốc gia, giới tính và nhóm tuổi trong giai đoạn 2025–2029 vào một hệ thống ứng dụng web có khả năng tương tác với người dùng cuối đã tiến hành xây dựng một tập hợp các API phía backend sử dụng framework Flask của Python.

Hệ thống API bao gồm năm giao diện lập trình ứng dụng chính, mỗi API đảm nhiệm một chức năng riêng biệt trong quy trình xử lý và trình bày dữ liệu. Cụ thể:

- API dự đoán: nhận các đầu vào từ người dùng (quốc gia, giới tính, nhóm tuổi), tiền xử lý dữ liệu, chuẩn hóa đầu vào, tạo chuỗi thời gian cho mô hình LSTM, và trả về kết quả dự đoán tỷ lệ thất nghiệp trong 5 năm tiếp theo, kèm theo biểu đồ và bảng dữ liệu.
- API khám phá dữ liệu: hỗ trợ người dùng tương tác với tập dữ liệu lịch sử (2014–2024), cho phép lọc theo nhiều tiêu chí và trực quan hóa thông tin bằng biểu đồ dạng đường (line chart) hoặc cột (bar chart), giúp hiểu rõ xu hướng thất nghiệp theo thời gian.
- API xuất dữ liệu CSV: phục vụ mục đích lưu trữ và phân tích ngoại tuyến, cho phép người dùng tải xuống dữ liệu đã lọc ở định dạng CSV với mã hóa UTF-8 để đảm bảo tương thích với Excel và các phần mềm xử lý bảng tính khác.
- API báo cáo nhanh: phân tích chuyên sâu tình hình thất nghiệp của một quốc gia bằng cách xác định nhóm dân số chịu ảnh hưởng nhiều nhất (theo giá trị

trung bình hoặc cao nhất trong 3 năm gần nhất), đồng thời trích xuất xu hướng theo từng phân nhóm giới tính và độ tuổi.

- API xuất báo cáo PDF: tổng hợp kết quả phân tích từ API báo cáo nhanh để tạo ra một báo cáo định dạng PDF chuyên nghiệp, có hỗ trợ phông chữ Unicode tiếng Việt, biểu đồ dạng ảnh, và nội dung được trình bày theo bố cục rõ ràng, phù hợp cho việc lưu trữ, chia sẻ hoặc in ấn.

Toàn bộ hệ thống được xây dựng theo kiến trúc RESTful, đảm bảo khả năng mở rộng, tái sử dụng và dễ dàng tích hợp với giao diện người dùng (frontend) qua các lời gọi HTTP. Việc tách biệt chức năng giúp phát triển và bảo trì thuận tiện, đồng thời đảm bảo hiệu suất và độ tin cậy.

Vấn đề bảo mật API cũng cần những biện pháp cụ thể:

- Xác thực và ủy quyền: Sử dụng JWT để kiểm soát quyền truy cập, phân cấp theo vai trò người dùng.
- Mã hóa dữ liệu: Toàn bộ dữ liệu được mã hóa bằng HTTPS/SSL/TLS trong quá trình truyền tải.
- Kiểm soát tốc độ truy cập: Giới hạn số lượng yêu cầu API nhằm chống lại tấn công DDoS và lạm dụng.
- Xử lý lỗi và ghi nhật ký bảo mật: Xử lý lỗi cẩn thận để tránh lộ thông tin và ghi nhật ký chi tiết các sự kiện bảo mật để giám sát.

Cấu trúc API này giúp mô hình LSTM trở thành một thành phần dễ sử dụng trong hệ thống web, cho phép người dùng đưa ra quyết định dựa trên dữ liệu một cách nhanh chóng và chính xác mà không cần kiến thức chuyên sâu về học máy.

API 1: /predict

- Phương thức: POST
- Chức năng: Dự đoán tỷ lệ thất nghiệp 5 năm tiếp theo (2025–2029) dựa trên đầu vào: quốc gia, giới tính, nhóm tuổi.
- Xử lý:

  + Tiền xử lý dữ liệu (Label Encoding, Scaling)

+ Tạo chuỗi dữ liệu đầu vào cho LSTM từ 5 năm gần nhất

+ Trả về kết quả dự đoán + độ không chắc chắn (confidence intervals)

+ Sinh biểu đồ (base64) và bảng dữ liệu

API 2: /api/explore

- Phương thức: POST
- Chức năng: Khám phá tỷ lệ thất nghiệp theo nhiều quốc gia, giới tính và nhóm tuổi, từ năm 2014 đến 2024.
- Tùy chọn biểu đồ: line hoặc bar.
- Xử lý:

  + Lọc dữ liệu theo bộ lọc người dùng

  + Vẽ biểu đồ trực quan hóa xu hướng

  + Trả về dữ liệu bảng và biểu đồ (base64)

API 3: /api/export-csv

- Phương thức: POST
- Chức năng: Xuất dữ liệu đã lọc (theo quốc gia, giới tính, nhóm tuổi, khoảng năm) ra file CSV để tải về.
- Xử lý:

  + Lọc dữ liệu như /api/explore

  + Chuyển dữ liệu sang StringIO, sau đó thành BytesIO

  + Gửi file CSV với định dạng UTF-8-SIG
- Phản hồi: send_file → file CSV tên theo timestamp.

API 4: /api/quick-report

- Phương thức: POST
- Chức năng: Tạo báo cáo nhanh cho một quốc gia (dữ liệu từ 2014–2024).
- Xử lý:
- Phân tích xu hướng của từng nhóm giới tính và độ tuổi
- Xác định nhóm chịu ảnh hưởng nhất:

  + Theo trung bình 3 năm gần nhất

  + Theo giá trị cao nhất

+ Tạo biểu đồ xu hướng tổng hợp

API 5: /api/export-report

- Phương thức: POST
- Chức năng: Tạo báo cáo PDF hoàn chỉnh từ kết quả phân tích của API 4.
- Xử lý:
  + Tạo tiêu đề, biểu đồ từ chuỗi base64
  + Hiển thị nhóm chịu ảnh hưởng nhất
  + Trình bày phân tích xu hướng theo từng nhóm
  + Hỗ trợ font Unicode (DejaVuSerif) cho tiếng Việt
- Phản hồi: send_file → file PDF bao_cao_that_nghiep_<quocgia>.pdf

3.2.3. Kiểm thử và tối ưu API

Sau khi xây dựng các API phục vụ dự đoán, khám phá dữ liệu và xuất báo cáo, quá trình kiểm thử (testing) và tối ưu (optimization) là bước cần thiết nhằm đảm bảo hiệu năng, tính đúng đắn và khả năng phản hồi của hệ thống khi triển khai thực tế.

Kiểm thử chức năng

Bảng 3.1 Kiểm thử chức năng

API	Tình huống kiểm thử chính	Kết quả mong đợi
/predict	Dự đoán với tổ hợp đầu vào hợp lệ và không hợp lệ	Trả về kết quả dự đoán hoặc thông báo lỗi rõ ràng
/api/explore	Khám phá dữ liệu với nhiều quốc gia, nhóm tuổi, giới tính khác nhau	Trả về biểu đồ và bảng dữ liệu đúng với bộ lọc đã chọn
/api/export-csv	Xuất dữ liệu với/không có bộ lọc	Tập tin CSV có định dạng đúng, đầy đủ thông tin
/api/quick-report	Tạo báo cáo nhanh với quốc gia có dữ liệu và không có dữ liệu	Trả về báo cáo JSON hoặc thông báo lỗi rõ ràng

/api/export-report	Xuất báo cáo PDF với biểu đồ và phân tích nhóm dân số	Tập tin PDF sinh ra đúng định dạng, hỗ trợ tiếng Việt

Tối ưu hóa hệ thống API và mô hình:

- Bộ nhớ cache: sử dụng decorator @lru_cache cho hàm load_data() để tránh việc đọc lại tập tin CSV nhiều lần trong cùng một phiên làm việc.
- Khởi tạo mô hình: mô hình LSTM chỉ được tải một lần duy nhất khi khởi động máy chủ Flask, không khởi tạo lại với mỗi lần gọi API.
- Tối ưu hóa chuỗi đầu vào: quá trình chuẩn hóa dữ liệu đầu vào (sử dụng LabelEncoder và MinMaxScaler) được thực hiện một lần duy nhất và tái sử dụng cho các yêu cầu sau.
- Tăng tốc tạo file: sử dụng định dạng BytesIO thay cho ghi/đọc file tạm trên ổ cứng khi sinh CSV hoặc PDF, giúp giảm I/O disk và tăng tốc xử lý.
- Giảm log không cần thiết: trong quá trình dự đoán, việc tắt các log của mô hình với verbose=0 giúp tiết kiệm thời gian và tránh in ra console không cần thiết.

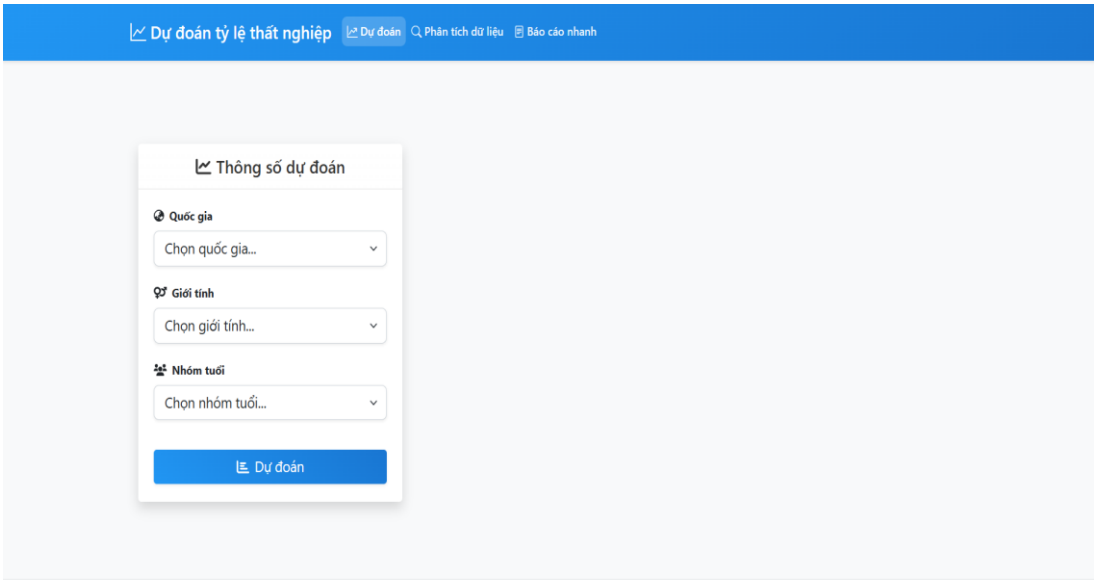
Xử lý lỗi và tăng cường bảo mật API

- Để đảm bảo hệ thống hoạt động ổn định trong môi trường thực tế và hạn chế rủi ro bảo mật, một số cơ chế xử lý lỗi và kiểm soát đầu vào đã được áp dụng:
- Xử lý ngoại lệ toàn diện: sử dụng try-except tại mỗi API để bắt lỗi và trả về thông báo lỗi có cấu trúc JSON, dễ hiểu và có thể hiển thị trực tiếp ở giao diện người dùng.
- Kiểm soát đầu vào: kiểm tra kỹ lưỡng các tham số đầu vào như quốc gia, giới tính, nhóm tuổi, năm – đảm bảo các giá trị hợp lệ và nằm trong danh sách cho phép.
- Chống injection: xử lý đầu vào cẩn thận để tránh injection (chèn mã độc) khi lọc dữ liệu hoặc khi ghi nội dung ra tập tin (CSV, PDF).

3.3. Xây dựng giao diện người dùng

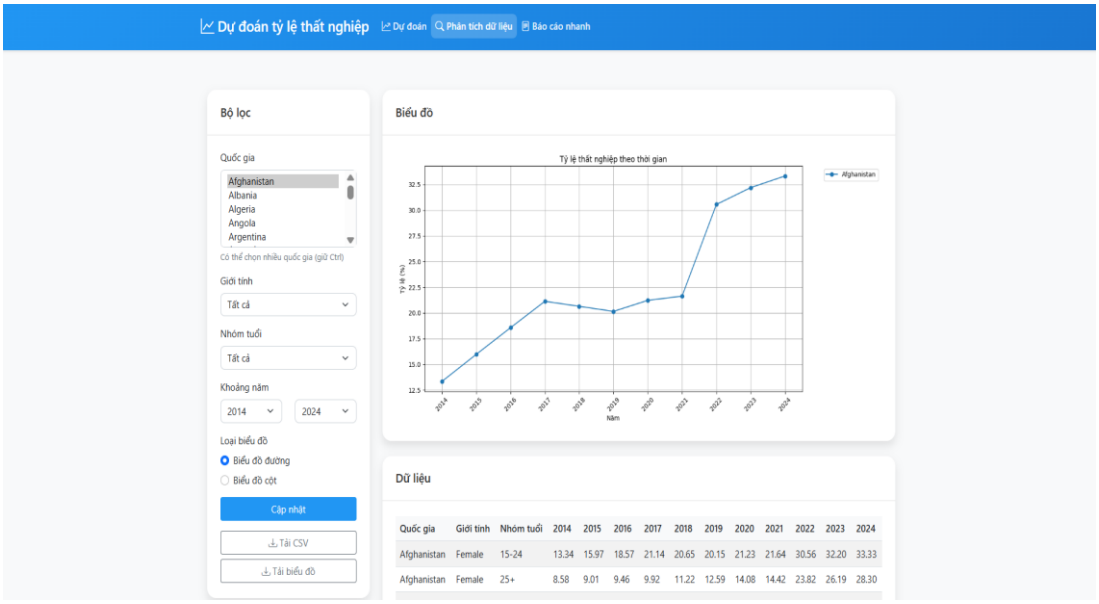
3.3.1. Thiết kế giao diện web/app

3.3.3.1 Giao diện trang chủ (Trang dự đoán)



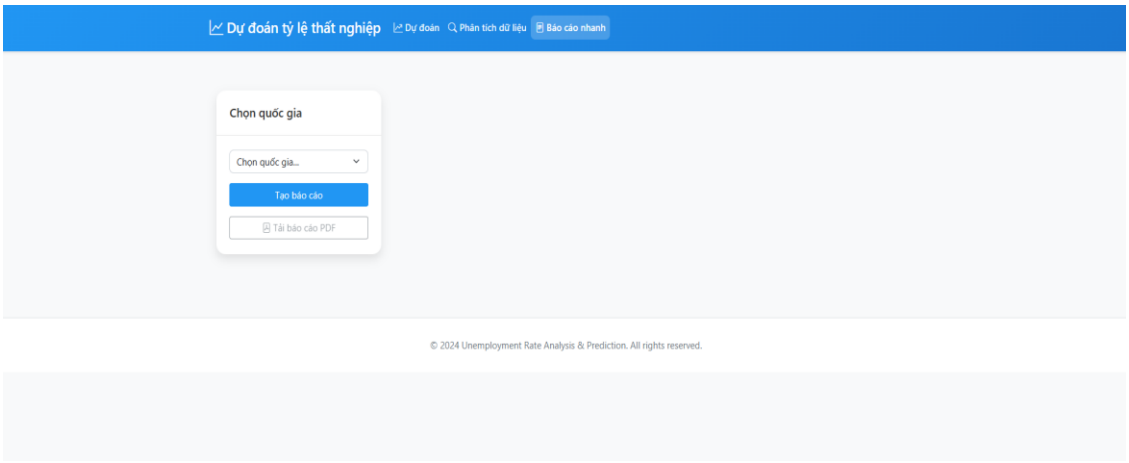
Hình 3.6 Giao diện trang chủ

3.3.3.2 Giao diện trang phân tích dữ liệu



Hình 3.7 Giao diện trang phân tích dữ liệu

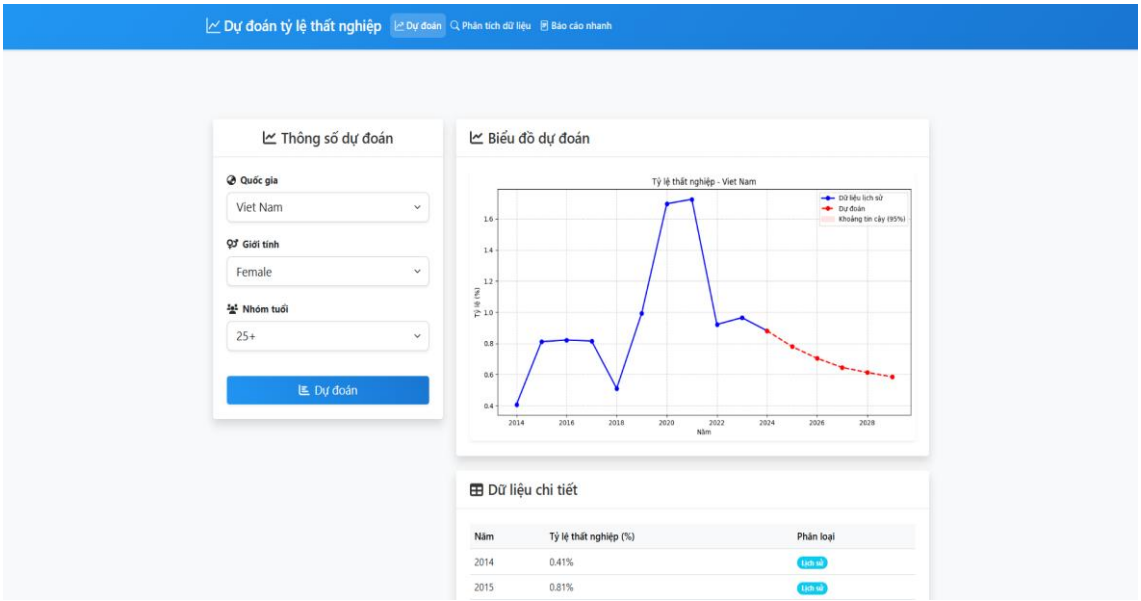
3.3.3.3 Giao diện trang báo cáo nhanh



Hình 3.8 Giao diện trang báo cáo nhanh

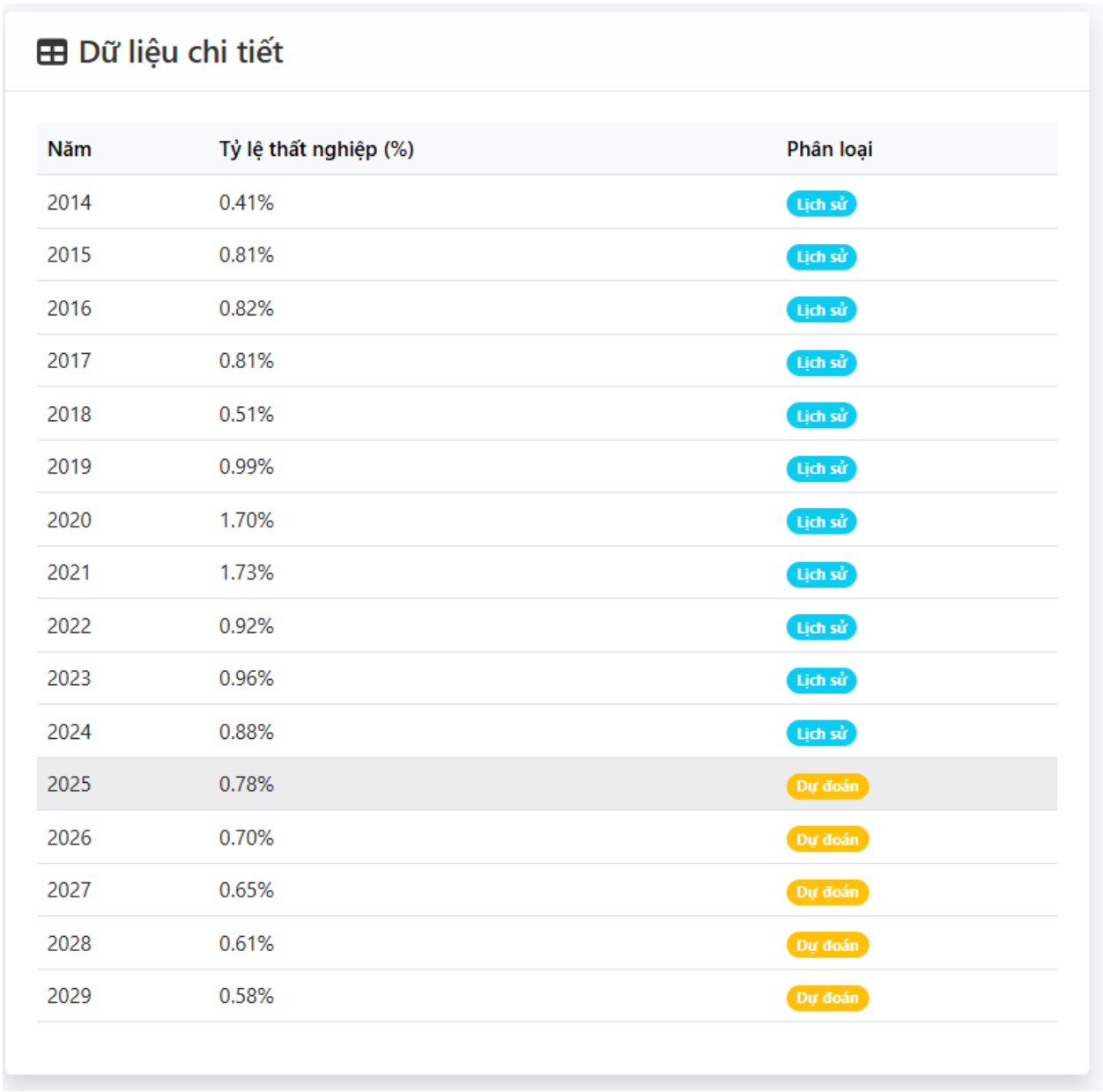
3.3.2. Chức năng nhập dữ liệu và hiển thị kết quả dự đoán

3.3.2.1 Chức năng dự đoán



Hình 3.9 Chức năng dự đoán

Khi chọn đầy đủ các thông số bao gồm: “Quốc gia”, “Giới tính” và “Nhóm tuổi”. Ta bấm nút dự đoán. Hệ thống sẽ dựa trên mô hình mạng học sâu LTSM đã được train với độ tin cậy khoảng 95% sẽ đưa ra được các chỉ số tỷ lệ thất nghiệp trong vòng 5 năm tới từ 2025-2029 và được thể hiện bằng các đường màu đỏ. Các chỉ số dữ liệu lịch sử từ 2014-2024 sẽ được thể hiện ở đường màu xanh trong biểu đồ.



Dữ liệu chi tiết

Năm	Tỷ lệ thất nghiệp (%)	Phân loại
2014	0.41%	Lịch sử
2015	0.81%	Lịch sử
2016	0.82%	Lịch sử
2017	0.81%	Lịch sử
2018	0.51%	Lịch sử
2019	0.99%	Lịch sử
2020	1.70%	Lịch sử
2021	1.73%	Lịch sử
2022	0.92%	Lịch sử
2023	0.96%	Lịch sử
2024	0.88%	Lịch sử
2025	0.78%	Dự đoán
2026	0.70%	Dự đoán
2027	0.65%	Dự đoán
2028	0.61%	Dự đoán
2029	0.58%	Dự đoán

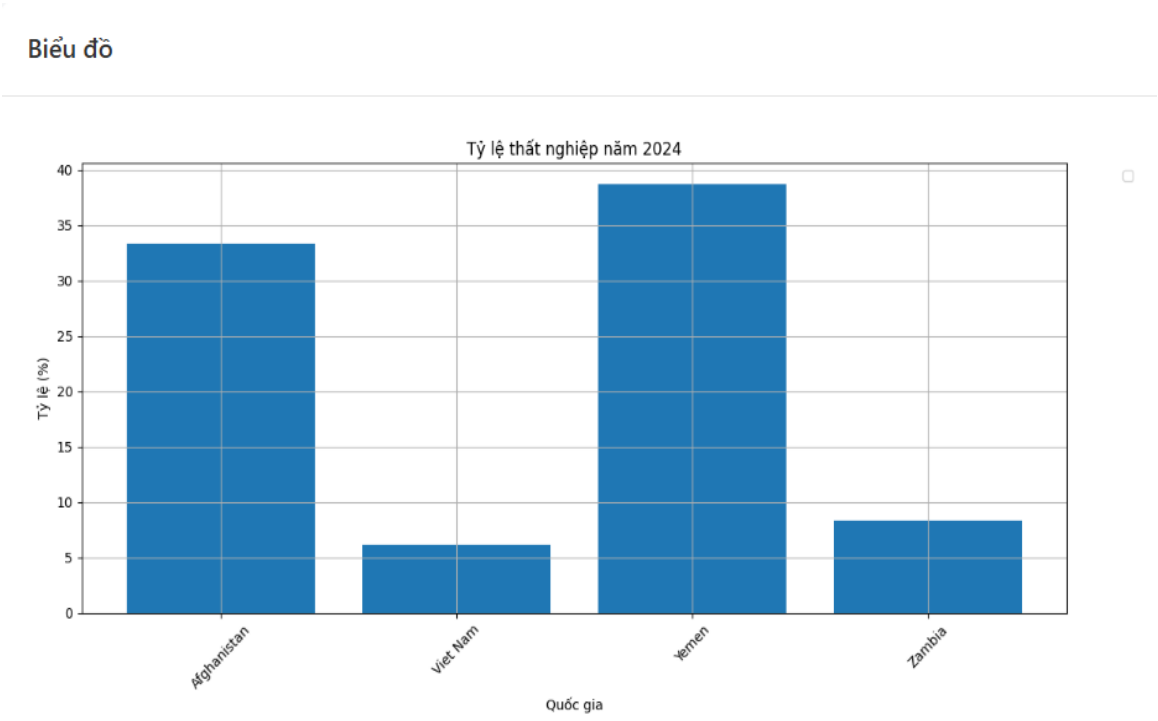
Hình 3.10 Dữ liệu chi tiết

Bên cạnh đó khi dự đoán sẽ đưa ra các số liệu chi tiết theo từng năm. Từ đó có thể thấy được tỉ lệ thất nghiệp trong tương lai của quốc gia đó.

3.3.2.2 Chức năng phân tích dữ liệu

Hình 3.11 Chức năng phân tích dữ liệu

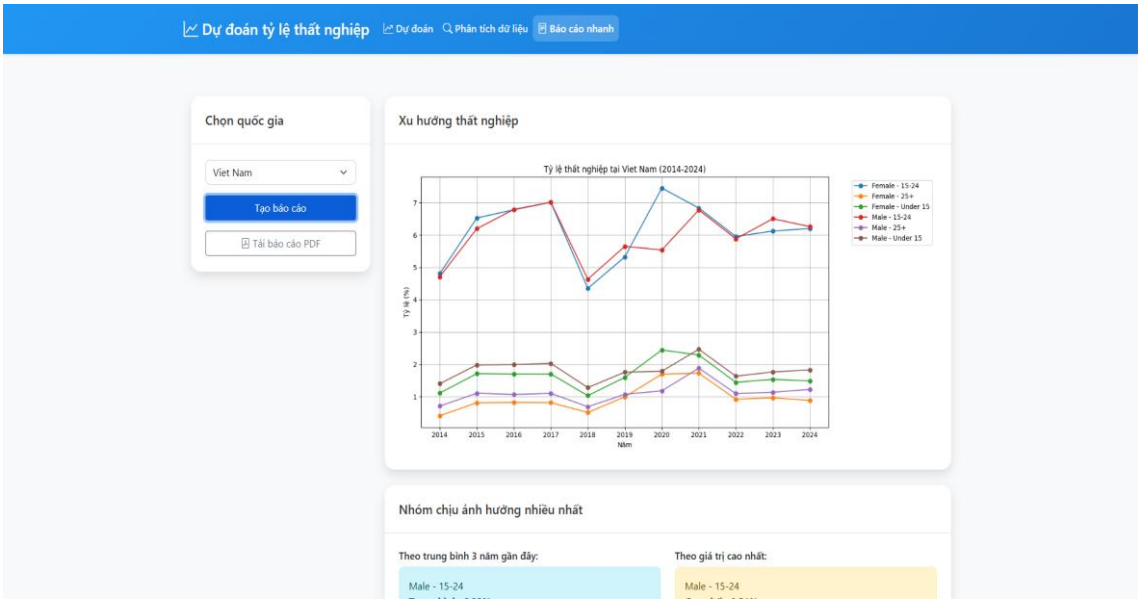
Chức năng khám phá dữ liệu có thể giúp ta xem được tỷ lệ thất nghiệp theo thời gian của 1 hoặc nhiều quốc gia theo thời gian dựa trên bộ lọc. Tỷ lệ thất nghiệp có thể thể hiện bằng biểu đồ đường hoặc biểu đồ cột.



Hình 3.12 Chức năng tạo biểu đồ cột

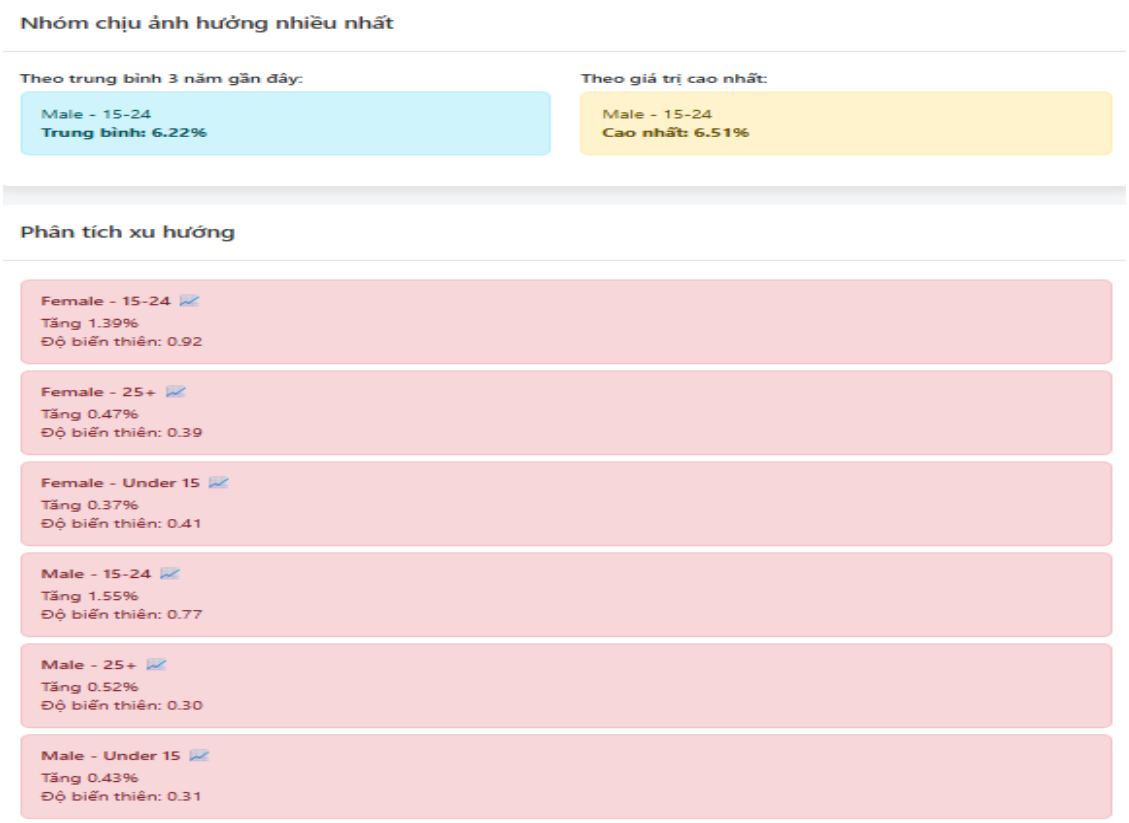
Ngoài ra chức năng khám phá dữ liệu có thể sử dụng để tải bản CSV hoặc tải biểu đồ giúp người dùng dễ dàng có thể lấy được số liệu từ đó và có thể dễ dàng so sánh giữa các quốc gia

3.3.2.3 Chức năng báo cáo nhanh



Hình 3.13 Chức năng báo cáo nhanh

Chức năng báo cáo nhanh cũng là 1 tính năng rất thú vị giúp người dùng có thể phân tích được xu hướng thất nghiệp của quốc gia đó theo từng nhóm tuổi.



Hình 3.14 Phân tích xu hướng của chức năng báo cáo nhanh

Ngoài ra chức năng báo cáo nhanh sẽ đưa ra được nhóm tuổi chịu ảnh hưởng nhiều nhất trong 3 năm gần nhất và theo giá trị cao nhất. Thêm vào đó, chức năng báo cáo nhanh sẽ phân tích được xu hướng của từng nhóm tuổi sẽ theo chiều hướng tăng hay giảm.

3.3.3. Kiểm thử giao diện và tối ưu trải nghiệm người dùng

Bên cạnh việc xây dựng các API xử lý dữ liệu và mô hình học máy, một thành phần không thể thiếu trong hệ thống là giao diện người dùng (UI – User Interface). Giao diện không chỉ đóng vai trò là cầu nối giữa người dùng và hệ thống, mà còn là kênh truyền tải kết quả phân tích, biểu đồ dự đoán và báo cáo trực quan một cách dễ hiểu. Do đó, việc kiểm thử giao diện (UI Testing) và tối ưu hóa trải nghiệm người dùng (UX – User Experience) là yếu tố thiết yếu trong quá trình triển khai hệ thống nhằm đảm bảo tính ổn định, thân thiện và hiệu quả khi vận hành thực tế.

a) Kiểm thử giao diện (UI Testing)

Giao diện được kiểm thử trên nhiều khía cạnh quan trọng như: khả năng hiển thị nội dung, tính tương thích trình duyệt, hiệu quả tương tác người dùng và khả năng xử lý lỗi. Các kỹ thuật và công cụ kiểm thử đã được áp dụng bao gồm:

- Manual Testing trên các trình duyệt phổ biến như Google Chrome, Microsoft Edge để đảm bảo khả năng tương thích trình duyệt đa nền tảng.
- Responsive design testing nhằm đảm bảo giao diện hiển thị đúng kích thước và bố cục trên các loại thiết bị, đặc biệt là màn hình máy tính.
- Kiểm thử JavaScript bao gồm việc đánh giá các hành động như gửi yêu cầu API, hiển thị biểu đồ, xử lý lỗi, và thông báo kết quả.
- Kiểm thử API bằng công cụ như Postman, Swagger UI để đảm bảo các yêu cầu gửi đến backend đều trả về đúng dữ liệu, định dạng, mã phản hồi và thông báo lỗi.
- Kiểm thử hiệu năng sử dụng công cụ như Apache JMeter, nhằm đo lường thời gian phản hồi, băng thông tiêu thụ và khả năng xử lý của hệ thống khi có nhiều yêu cầu đồng thời (stress test và load test).

Kiểm thử tự động các chức năng quan trọng với các công cụ như pytest (Python) để phát hiện lỗi sớm trong quá trình phát triển và triển khai.

Bảng 3. 2 Kiểm thử giao diện

Yếu tố kiểm thử	Kết quả đạt được
Hiển thị biểu đồ dự đoán	Biểu đồ LSTM hiển thị đầy đủ, rõ nét, đúng năm và màu sắc phân biệt
Bảng dữ liệu	Được render theo đúng thứ tự, có thể cuộn ngang khi nhiều cột
Gửi yêu cầu API (predict/explore)	Hoạt động ổn định, dữ liệu trả về đúng định dạng, giao diện tự động cập nhật nội dung

Yếu tố kiểm thử	Kết quả đạt được
Xử lý lỗi đầu vào	Giao diện thông báo lỗi rõ ràng nếu thiếu dữ liệu hoặc nhập sai
Xuất file (CSV, PDF)	Giao diện có nút tải về, trạng thái tải hiển thị rõ ràng

b) Tối ưu hóa trải nghiệm người dùng

Nhằm nâng cao hiệu quả sử dụng và giảm thiểu thao tác không cần thiết trong quá trình tương tác, hệ thống giao diện người dùng đã được tối ưu hóa với nhiều cải tiến đáng chú ý:

- Tự động cập nhật nội dung: Hệ thống được thiết kế để tự động gửi yêu cầu tới API và hiển thị kết quả ngay sau khi người dùng chọn đầy đủ các tiêu chí (quốc gia, giới tính, nhóm tuổi), mà không cần thao tác nạp lại trang hoặc nhấn nút xác nhận. Điều này giúp rút ngắn thời gian thao tác và tăng tính phản hồi của giao diện.
- Biểu đồ chuyển động nhẹ (fade-in): Các biểu đồ sau khi sinh được hiển thị với hiệu ứng chuyển động nhẹ (fade-in) nhằm tạo cảm giác mượt mà và giúp người dùng dễ dàng nhận ra nội dung mới đã được cập nhật.
- Thông báo tức thì (flash messages): Hệ thống sử dụng các khối thông báo động để cung cấp phản hồi trực tiếp về hành động của người dùng, bao gồm xác nhận thành công hoặc hiển thị lỗi nếu có sự cố đầu vào hoặc xử lý API. Các thông báo này giúp người dùng nhận biết hành động thành công, lỗi, cảnh báo, hoặc thông tin.
- Giới hạn đầu vào và hỗ trợ lựa chọn: Giao diện chỉ cho phép lựa chọn từ danh sách cố định (dropdown list), thay vì nhập tay, nhằm giảm nguy cơ nhập sai định dạng và loại bỏ lỗi chính tả không cần thiết.
- Bố cục giao diện hợp lý: Các chức năng chính được phân chia theo từng mục rõ ràng như “Khám phá dữ liệu”, “Dự đoán”, và “Báo cáo nhanh”. Thiết kế này hỗ trợ người dùng dễ dàng điều hướng và tập trung vào từng chức năng cụ thể một cách thuận tiện.

c) Tương thích trình duyệt và hỗ trợ truy cập

Để đảm bảo giao diện hoạt động ổn định và dễ tiếp cận cho nhiều đối tượng người dùng, hệ thống đã được kiểm thử và tối ưu trên nhiều khía cạnh liên quan đến khả năng tương thích và khả dụng:

Hỗ trợ trình duyệt hiện đại: Giao diện được kiểm thử kỹ lưỡng và đảm bảo hoạt động ổn định trên các trình duyệt phổ biến như Google Chrome, Mozilla Firefox và Microsoft Edge. Bố cục, biểu đồ và các thành phần giao diện đều hiển thị đúng chuẩn và không xảy ra lỗi định dạng.

Hỗ trợ ngôn ngữ tiếng Việt: Hệ thống hỗ trợ đầy đủ tiếng Việt có dấu ở tất cả thành phần, bao gồm biểu đồ, dữ liệu bảng và báo cáo PDF. Đây là yếu tố quan trọng để người dùng trong nước có thể sử dụng thuận tiện và tiếp cận kết quả dễ dàng hơn.

Hiệu suất hiển thị cao: Thời gian phản hồi trung bình của giao diện từ lúc gửi yêu cầu đến khi hiển thị biểu đồ hoặc bảng dữ liệu chỉ dao động từ 200 đến 500 ms, mang lại trải nghiệm nhanh chóng và mượt mà.

Thân thiện với người dùng phổ thông: Giao diện được thiết kế tối giản, trực quan, giúp người dùng không chuyên cũng có thể thao tác dễ dàng mà không cần kiến thức kỹ thuật về học máy hoặc xử lý dữ liệu. Mọi bước đều có hướng dẫn rõ ràng, biểu tượng dễ hiểu, và phản hồi rõ ràng khi xảy ra lỗi.

CHƯƠNG 4. KẾT QUẢ, ĐÁNH GIÁ VÀ HƯỚNG PHÁT TRIỂN

4.1. Kết quả thực nghiệm và đánh giá hiệu xuất

Sau khi hoàn tất quá trình xây dựng hệ thống gồm mô hình học sâu (LSTM), các API phục vụ dự đoán, khám phá dữ liệu, xuất báo cáo, cùng với giao diện người dùng, tôi đã tiến hành thực nghiệm hệ thống trong môi trường giả lập để đánh giá hiệu suất toàn diện. Mục tiêu là kiểm tra khả năng hoạt động của hệ thống khi xử lý các truy vấn thực tế từ người dùng, đảm bảo đáp ứng về mặt tốc độ, độ chính xác và tính ổn định

Trong thực nghiệm, giao diện được đánh giá qua tốc độ hiển thị, khả năng phản hồi và mức độ thân thiện với người dùng:

- Tốc độ cập nhật: Biểu đồ và bảng dữ liệu hiển thị trong khoảng 200–500ms sau mỗi lần người dùng thay đổi bộ lọc.
- Phản hồi thời gian thực: Kết quả được tự động hiển thị mà không cần tải lại trang.
- Khả năng tương thích: Giao diện hoạt động ổn định trên các trình duyệt hiện đại như Google Chrome, Microsoft Edge.
- Hỗ trợ tiếng Việt: Toàn bộ biểu đồ, dữ liệu và báo cáo PDF hiển thị tiếng Việt có dấu đầy đủ, đảm bảo tính bản địa hóa.
- Thân thiện với người dùng: Giao diện trực quan, phân chia chức năng theo tab, hỗ trợ dropdown để tránh nhập liệu sai, đồng thời có thông báo lỗi hoặc thành công rõ ràng.

4.1.1. Khả năng mở rộng và tính ổn định

Hệ thống được thiết kế theo hướng nhẹ, độc lập, có khả năng mở rộng trên các nền tảng máy chủ hoặc cloud. Một số điểm nổi bật:

- Tối ưu mô hình: Mô hình LSTM được nạp vào bộ nhớ một lần duy nhất khi khởi động, tiết kiệm thời gian và tài nguyên khi xử lý yêu cầu.
- Bộ nhớ đệm hiệu quả: Sử dụng kỹ thuật cache dữ liệu với @lru_cache giúp giảm chi phí đọc file và cải thiện hiệu suất.

- Sinh file trực tiếp từ bộ nhớ: Thay vì ghi ra đĩa, hệ thống sử dụng BytesIO để tạo file PDF/CSV và trả trực tiếp, giúp tiết kiệm I/O và tăng tốc độ phản hồi.

4.2. Hướng phát triển trong tương lai

4.2.1. Cải thiện mô hình dự báo

Tích hợp khả năng giải thích (Explainability)

Hiện tại, mô hình LSTM hoạt động theo dạng "hộp đen" – nghĩa là mặc dù có thể dự báo chính xác nhưng lại không cung cấp được lý do tại sao có những dự báo đó. Đây là một rào cản lớn khi muốn ứng dụng mô hình vào các quyết sách của Chính phủ hoặc các tổ chức xã hội. Để khắc phục, hệ thống cần tích hợp các phương pháp AI có thể giải thích được như:

- SHAP (SHapley Additive exPlanations): Phương pháp này giúp xác định yếu tố nào (GDP, lạm phát, nhóm tuổi...) tác động nhiều nhất tới dự báo thất nghiệp. SHAP có thể chỉ ra rằng trong một giai đoạn cụ thể, sự sụt giảm GDP có ảnh hưởng lớn nhất đến tỷ lệ thất nghiệp tăng, hoặc tỷ lệ lạm phát cao đang là động lực chính.
- LIME (Local Interpretable Model-agnostic Explanations): LIME tập trung vào việc giải thích các dự đoán riêng lẻ của mô hình bằng cách tạo ra một mô hình đơn giản, dễ hiểu xung quanh dự đoán đó. Điều này đặc biệt hữu ích khi cần giải thích tại sao mô hình dự báo một mức thất nghiệp cụ thể cho một nhóm dân số hoặc khu vực nhất định.
- Attention Mechanism: Cơ chế này giúp trực quan hóa việc mô hình "tập trung" vào phần dữ liệu nào khi đưa ra dự báo. Chẳng hạn, khi dự báo thất nghiệp quý tới, Attention Mechanism có thể cho thấy mô hình đang đặt trọng tâm vào dữ liệu thất nghiệp của 6 tháng trước đó, hoặc vào các chính sách kinh tế được ban hành gần đây.

Việc bổ sung khả năng giải thích sẽ giúp các nhà hoạch định chính sách hiểu rõ cơ sở của dự báo, từ đó có thể nâng cao tính tin cậy và khả năng ứng dụng vào thực tiễn.

Bổ sung các biến kinh tế - xã hội đa chiều

Mô hình hiện tại chủ yếu dựa vào dữ liệu lịch sử của tỷ lệ thất nghiệp, điều này hạn chế khả năng nắm bắt nguyên nhân sâu xa. Trong thực tế, thất nghiệp là kết quả của nhiều yếu tố kinh tế - xã hội đan xen. Do đó, để nâng cao năng lực phân tích, mô hình cần tích hợp thêm các biến độc lập như:

- Tăng trưởng GDP: Cho thấy sức khỏe của nền kinh tế.
- Tỷ lệ lạm phát, lãi suất: Phản ánh tình hình chi tiêu, đầu tư.
- Trình độ học vấn, kỹ năng nghề: Đánh giá khả năng thích nghi của lực lượng lao động.
- Mức độ tự động hóa, chuyển đổi số: Xác định nguy cơ mất việc do công nghệ thay thế.
- Chính sách phúc lợi, bảo hiểm thất nghiệp: Có thể làm giảm tác động của khủng hoảng việc làm.

Việc kết hợp các biến này sẽ giúp mô hình không chỉ "dự đoán con số" mà còn phân tích được nguyên nhân và gợi ý giải pháp chính sách cụ thể cho từng nhóm đối tượng.

Tái huấn luyện mô hình theo chu kỳ:

Thị trường lao động có sự biến động mạnh mẽ, đặc biệt sau COVID-19, chiến tranh, biến động chính trị và sự phát triển công nghệ. Do đó, mô hình cần được tái huấn luyện (retrain) định kỳ (hàng quý hoặc hàng năm) bằng dữ liệu mới để đảm bảo tính cập nhật và chính xác. Đây là điều kiện cần để mô hình hoạt động bền vững và phản ánh đúng thực trạng.

4.2.2. Mở rộng dữ liệu

Kết nối với các nguồn dữ liệu chính thống

Hiện hệ thống đang sử dụng một bộ dữ liệu giới hạn cả về phạm vi và thời gian. Trong khi đó, thực tế đòi hỏi hệ thống phải liên tục cập nhật để theo sát thị trường. Kết nối với dữ liệu từ Tổng cục Thống kê, ILO, World Bank, IMF hoặc các cơ quan quốc tế uy tín.

- Thu thập dữ liệu theo chu kỳ thời gian thực hoặc tối thiểu là hàng quý/hàng năm.
- Tự động hóa quá trình cập nhật thông qua API, giảm phụ thuộc vào thao tác thủ công.

Việc mở rộng và cập nhật dữ liệu giúp mô hình không bị lạc hậu so với thực tiễn, đồng thời cải thiện đáng kể độ chính xác.

Phân tích chi tiết theo từng nhóm dân cư

Trong thực tế, tỷ lệ thất nghiệp không phân bổ đều – mỗi giới tính, độ tuổi, khu vực có mức độ rủi ro khác nhau. Vì vậy, cần nâng cấp hệ thống để:

- Dự báo theo giới tính (nam, nữ, khác): Giúp theo dõi tác động giới, có thể phụ nữ có thể bị ảnh hưởng nặng hơn trong ngành dịch vụ thời COVID-19.
- Phân nhóm tuổi chi tiết (15–24, 25–34, 35–44, 45–54, 55+): Tìm ra nhóm tuổi dễ bị sa thải hoặc khó tìm việc.
- Phân vùng địa lý: Theo vùng miền, tỉnh/thành hoặc nhóm nước (thu nhập cao/trung bình/thấp), từ đó đưa ra chính sách địa phương hóa và hỗ trợ phân bổ nguồn lực hợp lý.

4.2.3. Cải thiện giao diện

Tăng cường trực quan hóa dữ liệu:

Dữ liệu thất nghiệp mang tính phức tạp, dễ gây nhàm chán nếu chỉ thể hiện dưới dạng bảng. Do đó, hệ thống cần nâng cấp giao diện với:

- Biểu đồ tương tác: Line chart để thể hiện xu hướng, bar chart để so sánh giữa các nhóm, heatmap để đánh dấu điểm "nóng".
- Bản đồ hóa dữ liệu (map chart): Giúp người dùng hình dung phân bố thất nghiệp theo khu vực địa lý một cách dễ dàng.
- Tùy biến biểu đồ theo nhu cầu người dùng, có thể chọn lọc các tiêu chí như giới tính, độ tuổi, vùng địa lý, giai đoạn.

Xây dựng dashboard tương tác

Tạo một dashboard có thể tùy chỉnh theo người dùng, giúp:

- Lọc và xem dữ liệu theo quốc gia, khu vực, giới tính, nhóm tuổi.
- So sánh song song giữa các quốc gia hoặc giai đoạn.
- Hỗ trợ phân tích tình huống cụ thể: tỷ lệ thất nghiệp của phụ nữ 25–34 tuổi tại miền Trung Việt Nam trong năm 2022.

Giao diện thân thiện và đa nền tảng

Thiết kế giao diện responsive, tối ưu trên cả máy tính, điện thoại và máy tính bảng, giúp hệ thống dễ dàng tiếp cận người dùng phổ thông, đặc biệt là các nhà quản lý, nhà báo hoặc giảng viên.

4.2.4. Ứng dụng thực tiễn

Xây dựng ứng dụng di động

Một ứng dụng mobile đơn giản, dễ dùng sẽ giúp hệ thống:

- Cập nhật tỷ lệ thất nghiệp theo thời gian thực.
- Gửi thông báo cảnh báo sớm hoặc báo cáo hàng tháng qua app.
- Cho phép chia sẻ nhanh biểu đồ, số liệu qua Zalo, Telegram, email…

Ứng dụng này hướng tới người dùng phổ thông, cán bộ quản lý cấp tỉnh, nhà báo, sinh viên nghiên cứu.

Xây dựng hệ thống cảnh báo sớm

Dựa trên dữ liệu đầu vào và ngưỡng được thiết lập, hệ thống có thể phát hiện nguy cơ thất nghiệp tăng nhanh và gửi cảnh báo:

- Cảnh báo màu sắc (đỏ, vàng, xanh) tương ứng với mức độ nghiêm trọng.
- Gợi ý chính sách như tăng đầu tư công, hỗ trợ doanh nghiệp vừa và nhỏ, đẩy mạnh đào tạo nghề.
- Thông báo qua email hoặc app để người dùng được cập nhật ngay lập tức.

Mở rộng sang các lĩnh vực dự báo khác

Hệ thống có thể tiếp tục được phát triển sang các lĩnh vực như:

- Dự báo nhu cầu lao động theo ngành nghề (công nghệ, nông nghiệp, dịch vụ…).
- Phân tích việc làm trong tương lai: Nghề nào dễ bị thay thế bởi AI, kỹ năng nào cần học?
- Hỗ trợ tư vấn hướng nghiệp: Giúp học sinh, sinh viên chọn ngành theo nhu cầu lao động.

4.2.5. Hợp tác liên ngành và chia sẻ dữ liệu mở

Tăng cường hiệu quả và khả năng lan tỏa thông qua:

- Kết nối với trường đại học, viện nghiên cứu, tổ chức phi chính phủ để mở rộng nghiên cứu.
- Cung cấp API công khai để các nhà báo dữ liệu, lập trình viên có thể tích hợp hệ thống vào các sản phẩm khác.
- Tham gia mạng lưới dữ liệu mở toàn cầu như GODEL, Data.gov.vn để đóng góp và chia sẻ dữ liệu về thị trường lao động.

KẾT LUẬN

Đồ án tốt nghiệp tập trung vào xây dựng hệ thống phân tích và dự đoán tỷ lệ thất nghiệp theo quốc gia, giới tính và nhóm tuổi, dựa trên dữ liệu kinh tế - xã hội được thu thập từ các nguồn đáng tin cậy. Bằng việc ứng dụng mô hình học sâu LSTM kết hợp với giao diện web trực quan, hệ thống cho phép người dùng theo dõi và dự đoán xu hướng thất nghiệp một cách thuận tiện và hiệu quả.

Trong quá trình thực hiện, nhóm thực hiện đã áp dụng các kiến thức chuyên môn về trí tuệ nhân tạo (AI), học máy (Machine Learning), xử lý chuỗi thời gian (Time Series), lập trình web và trực quan hóa dữ liệu. Mô hình LSTM đã cho thấy khả năng ghi nhớ và học từ các chuỗi dữ liệu dài, phù hợp với bài toán dự đoán tỷ lệ thất nghiệp có tính chất thời gian. Các kỹ thuật tiền xử lý dữ liệu, lựa chọn đặc trưng và đánh giá mô hình cũng được triển khai nghiêm túc để đảm bảo độ chính xác và độ tin cậy của kết quả dự đoán.

Mặc dù hệ thống đã hoàn thiện các chức năng cơ bản, vẫn còn nhiều tiềm năng để cải thiện và mở rộng. Trong tương lai, có thể xem xét tích hợp các mô hình tiên tiến hơn như BiLSTM hoặc Transformer, mở rộng phạm vi dự báo đến từng ngành nghề hoặc khu vực cụ thể, đồng thời xây dựng hệ thống báo cáo tự động và tương tác. Việc tối ưu hóa siêu tham số, bổ sung dữ liệu thời sự và tăng cường khả năng phân tích cũng sẽ là hướng phát triển tiếp theo.

Đồ án không chỉ củng cố kiến thức chuyên ngành mà còn góp phần nâng cao tư duy phân tích, kỹ năng lập trình và khả năng ứng dụng công nghệ vào giải quyết các vấn đề xã hội. Tỷ lệ thất nghiệp là một chỉ số quan trọng phản ánh sức khỏe kinh tế và chất lượng sống của người dân; do đó, việc dự đoán xu hướng thất nghiệp chính xác có thể hỗ trợ quá trình hoạch định chính sách, phân bổ nguồn lực và phát triển nguồn nhân lực một cách bền vững. Đây là đóng góp thiết thực và mang ý nghĩa thực tiễn cao trong bối cảnh nền kinh tế toàn cầu đang biến động mạnh mẽ.

TÀI LIỆU THAM KHẢO

Tiếng Việt

[1]. Nguyễn Văn Vinh, & Trần Minh Triết. (2020). Trí tuệ nhân tạo – Cơ sở và ứng dụng (Tái bản lần 2). NXB Đại học Quốc gia TP.HCM.

[2]. Phạm Hồng Sơn. (2021). Ứng dụng mạng LSTM trong dự báo chuỗi thời gian tiêu thụ điện năng tại Việt Nam. Tạp chí Khoa học & Công nghệ - Đại học Đà Nẵng, 19(6), 12–19.

[3]. Nguyễn Đức Nghĩa. (2019). Khai phá dữ liệu và ứng dụng trong kinh doanh. NXB Thống kê.

[4]. Trần Quốc Tuấn. (2022). So sánh các mô hình học máy trong dự báo giá cổ phiếu: Random Forest, XGBoost và LSTM. Tạp chí Công nghệ Thông tin & Truyền thông, 28(3), 45–53.

[5]. Trường Đại học Bách Khoa Hà Nội. (2018). Giáo trình Học máy. NXB Bách khoa Hà Nội.

[6]. Lê Quang Hùng. (2021). Ứng dụng mô hình học sâu trong dự báo kinh tế – nghiên cứu trường hợp với dữ liệu GDP và lạm phát. Luận văn Thạc sĩ Khoa học Máy tính, Trường Đại học Công nghệ, ĐHQG Hà Nội.

Website tham khảo:

[1]. https://www.kaggle.com/datasets/sazidthe1/global-unemployment-data

[2]. https://www.theglobaleconomy.com/rankings/unemployment_rate/ASEAN/

[3]. https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG?locations=VN

[4]. https://vi.wikipedia.org/wiki/%C4%90%E1%BB%8Bnh_lu%E1%BA%ADt_Okun

[5]. https://vi.wikipedia.org/wiki/%C4%90%C6%B0%E1%BB%9Dng_cong_Phillips

[6]. https://vnexpress.net/cu-nhan-dai-hoc-that-nghiep-nhieu-hon-trung-cap-4497162.html

[7]. https://machinelearningcoban.com/2016/12/28/linearregression/

[8]. https://machinelearningcoban.com/tabml_book/ch_model/random_forest.html

[9]. https://viblo.asia/p/gradient-boosting-tat-tan-tat-ve-thuat-toan-manh-me-nhat-trong-machine-learning-YWOZrN7vZQ0

[10]. https://nttuan8.com/bai-14-long-short-term-memory-lstm/

[11]. https://vn.investing.com/economic-calendar/unemployment-rate-300

[12]. https://www.nso.gov.vn/

[13]. https://neptune.ai/blog/performance-metrics-in-machine-learning-complete-guide

[14]. https://viblo.asia/p/tutorial-su-dung-flask-cho-nguoi-moi-bat-dau-p1PvQ3gAMldr

[15]. https://viblo.asia/p/xay-dung-mot-restful-api-don-gian-voi-python-va-flask-bJzKmMvDK9N

[16]. https://www.geeksforgeeks.org/feature-importance-with-random-forests/

[17]. https://machinelearningmastery.com/feature-importance-and-feature-selection-with-xgboost-in-python/