ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA HỆ THỐNG THÔNG TIN

BÁO CÁO ĐỒ ÁN MÔN HỌC
IS403 – PHÂN TÍCH DỮ LIỆU KINH DOANH

ĐỀ TÀI:
PHÂN TÍCH VÀ DỰ BÁO SỰ PHÁT TRIỂN KINH TẾ BỀN VỮNG Ở VIỆT NAM

Hướng dẫn:
TS. Trần Văn Hải Triều

Nhóm 8:

1	Lâm Tuấn Thịnh	22521408
2	Trịnh Nguyên Bảo Tín	22521482
3	Phạm Hồng Trà	22521495
4	Võ Đức Trung	22521572
5	Lê Yến Vi	22521654
6	Phạm Thanh Thúy Vi	22521657

TP. HỒ CHÍ MINH, THÁNG 4 NĂM 2025

LỜI CẢM ƠN

Đầu tiên, nhóm chúng em xin gửi lời cảm ơn chân thành đến quý Thầy cô giảng viên Trường Đại học Công nghệ thông tin – Đại học Quốc gia TP. HCM nói chung và quý thầy cô khoa Hệ thống Thông tin nói riêng, đã giúp cho nhóm chúng em có những kiến thức cơ bản làm nền tảng để thực hiện đề tài này.

Đặc biệt, nhóm chúng em xin gửi lời cảm ơn và lòng biết ơn sâu sắc nhất tới thầy – TS. Trần Văn Hải Triều, người đã hướng dẫn cho em trong suốt thời gian làm đề tài. Thầy đã trực tiếp hướng dẫn tận tình, sửa chữa và đóng góp nhiều ý kiến quý báu giúp nhóm chúng em hoàn thành tốt báo cáo môn học của mình.

Trong thời gian một học kỳ thực hiện đề tài, nhóm chúng em đã vận dụng những kiến thức nền tảng đã tích lũy đồng thời kết hợp với việc học hỏi và nghiên cứu những kiến thức mới từ thầy cô, bạn bè cũng như nhiều nguồn tài liệu tham khảo, để hoàn thành một báo cáo đồ án tốt nhất. Tuy nhiên, vì kiến thức chuyên môn còn hạn chế và bản thân còn thiếu nhiều kinh nghiệm thực tiễn nên nội dung của báo cáo không tránh khỏi những thiếu sót, em rất mong nhận được sự góp ý, chỉ bảo thêm của quý thầy cô nhằm hoàn thiện những kiến thức của mình để nhóm chúng em có thể dùng làm hành trang thực hiện tiếp các đề tài khác trong tương lai cũng như là trong việc học tập và làm việc sau này.

Một lần nữa xin gửi đến thầy cô, bạn bè lời cảm ơn chân thành và tốt đẹp nhất!

Thành phố Hồ Chí Minh, ngày 25 tháng 04 năm 2025

Nhóm sinh viên thực hiện.

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

TP. Hồ Chí Minh, tháng 4 năm 2025

BẢNG PHÂN CÔNG, ĐÁNH GIÁ THÀNH VIÊN

Bảng.1 Bảng phân công, đánh giá thành viên

Họ tên	MSSV	Phân công	Đánh giá
Lâm Tuấn Thịnh	22521408	- Code R, hoàn thiện báo cáo. - Theo dõi, thực hiện đầy đủ yêu cầu project.	100%
Trịnh Nguyên Bảo Tín	22521482	- Viết báo cáo lý thuyết tiền xử lý. - Hỗ trợ tiền xử lý	100%
Phạm Hồng Trà	22521495	- Thu thập & tiền xử l - Xây dựng mô hình dự báo. - Phân tích SHAP.	100%
Võ Đức Trung	22521572	- Thực hiện khám phá dữ liệu (EDA). - Phân tích SHAP	100%
Lê Yến Vi	22521654	- Tìm hiểu lý thuyết tăng trưởng bền vững. - Thảo luận kết quả thu được từ EDA.	100%
Phạm Thanh Thúy Vi	22521657	- Tìm hiểu về phát triển kinh tế bền vững. - Thảo luận kết quả thu được từ mô hình dự báo	100%

MỤC LỤC

DANH MỤC CÁC TỪ VIẾT TẮT

STT	Từ viết tắt	Nội dung
1	**SDG**	Sustainable Development Goals
2	**SHAP**	Shapley Additive exPlanations
3	**EDA**	Exploratory Data Analysis

TÓM TẮT ĐỀ TÀI

Nghiên cứu này phân tích và dự báo tiến độ thực hiện các Mục tiêu Phát triển Bền vững (SDGs) tại Việt Nam, đồng thời so sánh với các quốc gia trong khu vực Đông Nam Á. Bằng cách áp dụng các phương pháp học máy tiên tiến kết hợp với phân tích SHAP (SHapley Additive exPlanations), nghiên cứu đã xác định các yếu tố chính tác động đến việc đạt được SDGs tại Việt Nam.

Dữ liệu được thu thập từ các nguồn quốc tế và trong nước, trải qua quá trình tiền xử lý kỹ lưỡng để đảm bảo tính chính xác và đầy đủ. Các mô hình học máy bao gồm Linear Regression, Random Forest và XGBoost được huấn luyện và đánh giá để dự báo chỉ số SDG trong tương lai. Phương pháp SHAP được sử dụng để giải thích mức độ ảnh hưởng của từng biến đến kết quả dự báo.

Kết quả nghiên cứu chỉ ra ba nhóm yếu tố ảnh hưởng đến khả năng đạt SDGs của Việt Nam: nhóm yếu tố thúc đẩy, nhóm yếu tố cản trở, và nhóm yếu tố có tác động không chắc chắn. Dự báo cho thấy, với xu hướng hiện tại, Việt Nam có thể gặp thách thức trong việc đạt được tất cả các mục tiêu SDGs vào năm 2030.

Từ kết quả phân tích, nghiên cứu đề xuất các giải pháp chiến lược nhằm cải thiện tiến độ thực hiện SDGs tại Việt Nam, tập trung vào những lĩnh vực còn yếu và phát huy những thế mạnh hiện có. Các đề xuất này có thể hỗ trợ các nhà hoạch định chính sách trong việc điều chỉnh và tối ưu hóa các chiến lược phát triển bền vững cho Việt Nam trong thập kỷ tới.

CHƯƠNG 1: GIỚI THIỆU CHUNG

I. Bối cảnh nghiên cứu

Việt Nam đã đạt được những thành tựu kinh tế đáng kể trong hai thập kỷ qua với tốc độ tăng trưởng GDP bình quân trên 6%/năm, trở thành một trong những nền kinh tế phát triển nhanh nhất khu vực Đông Nam Á. Tuy nhiên, quá trình tăng trưởng nhanh cũng đặt ra nhiều thách thức về tính bền vững, đặc biệt khi xét đến các yếu tố môi trường và xã hội. Từ năm 2015, với việc Liên Hợp Quốc thông qua 17 mục tiêu phát triển bền vững (SDGs), Việt Nam đã cam kết và tích hợp các mục tiêu này vào chiến lược phát triển kinh tế-xã hội. Mặc dù đã có những tiến bộ, thứ hạng SDG của Việt Nam vẫn còn khiêm tốn so với các quốc gia trong khu vực. Do đó, việc nghiên cứu các yếu tố ảnh hưởng đến tăng trưởng kinh tế bền vững và dự báo chỉ số SDG của Việt Nam có ý nghĩa quan trọng trong bối cảnh hiện nay.

II. Tầm quan trọng của đề tài

- Xác định các yếu tố ưu tiên để cải thiện chỉ số SDG, từ đó đề xuất các giải pháp cụ thể và khả thi.
- Dự đoán xu hướng phát triển và điểm số SDG của Việt Nam, giúp đánh giá tiến độ thực hiện các mục tiêu quốc gia.
- Hỗ trợ các bên liên quan đưa ra quyết định dựa trên dữ liệu.

III. Mục tiêu nghiên cứu

1. Dự báo các chỉ số SDG của Việt Nam đến năm 2030:
   - Dự báo sự thay đổi của các chỉ số SDG trong tương lai dựa trên các mô hình học máy và phân tích dữ liệu.
2. Xác định các yếu tố ảnh hưởng đến sự phát triển bền vững:
   - Phân tích các yếu tố kinh tế, xã hội và môi trường ảnh hưởng đến sự tiến bộ trong các mục tiêu SDG của Việt Nam.
3. So sánh sự tiến bộ của Việt Nam với các quốc gia khác trong khu vực:
   - So sánh điểm số SDG của Việt Nam với các quốc gia trong khu vực Đông Nam Á và các nhóm thu nhập khác.

IV. Câu hỏi nghiên cứu

1. Việt Nam hiện tại đang đạt được các mục tiêu SDG như thế nào?
   - Các chỉ số SDG của Việt Nam hiện tại có đạt yêu cầu của Liên Hợp Quốc không?
2. Những yếu tố nào ảnh hưởng đến sự phát triển bền vững của Việt Nam?
   - Các yếu tố như thu nhập, giáo dục, y tế, và môi trường có ảnh hưởng gì đến sự phát triển bền vững ở Việt Nam?
3. Dự báo các chỉ số SDG của Việt Nam sẽ thay đổi ra sao vào năm 2030?

- Các mô hình dự báo có thể dự đoán Việt Nam sẽ đạt được những mục tiêu nào và thiếu sót ở đâu?

4. Việt Nam cần làm gì để cải thiện các chỉ số SDG và đạt mục tiêu 2030?
   - Các giải pháp nào có thể giúp cải thiện sự tiến bộ của Việt Nam trong các mục tiêu SDG?

V. Phạm vi nghiên cứu

- Quốc gia nghiên cứu: Việt Nam
- Thời gian nghiên cứu: Bộ dữ liệu sẽ được sử dụng từ năm 2000 đến 2024, với mục tiêu dự báo cho năm 2030.
- Mục tiêu nghiên cứu: Phân tích và dự báo các chỉ số SDG của Việt Nam, với trọng tâm vào các chỉ số quan trọng như xóa nghèo, giáo dục, sức khỏe, nước sạch, và năng lượng sạch.
- Giới hạn nghiên cứu: Nghiên cứu không bao gồm các yếu tố kinh tế xã hội không liên quan đến SDG và không tính đến các yếu tố chính trị cụ thể tại các thời điểm.

CHƯƠNG 2: TỔNG QUAN VỀ LÝ THUYẾT NGHIÊN CỨU

I. Cơ sở lý thuyết

I.1. Mô hình phát triển bền vững và SDGs

Phát triển bền vững (Sustainable Development) là một mô hình phát triển đáp ứng nhu cầu của thế hệ hiện tại mà không làm tổn hại đến khả năng đáp ứng nhu cầu của các thế hệ tương lai.

SDGs (Sustainable Development Goals - Mục tiêu Phát triển Bền vững) là một tập hợp 17 mục tiêu toàn cầu do Liên Hợp Quốc (LHQ) đề ra nhằm hướng đến phát triển bền vững trên toàn thế giới đến năm 2030.

I.1.1. Lịch sử ra đời của SDGs

Chương trình Nghị sự 2030 vì sự Phát triển Bền vững, được thông qua năm 2015 bởi tất cả các quốc gia thành viên Liên Hợp Quốc, là khuôn khổ toàn cầu nhằm hướng đến hòa bình và thịnh vượng cho con người và hành tinh. Trọng tâm của chương trình là 17 Mục tiêu Phát triển Bền vững (SDGs), kêu gọi hành động khẩn cấp để xóa đói giảm nghèo, cải thiện y tế – giáo dục, giảm bất bình đẳng, thúc đẩy kinh tế, ứng phó biến đổi khí hậu và bảo vệ tài nguyên thiên nhiên.

Các SDGs được xây dựng dựa trên nhiều thập kỷ nỗ lực của các quốc gia và Liên Hợp Quốc, bao gồm cả Bộ Kinh tế và Xã hội Liên Hợp Quốc.

Các mốc thời gian quan trọng trong quá trình hình thành và phát triển của Chương trình Nghị sự 2030 và các Mục tiêu Phát triển Bền vững (SDGs):

4. 1992 – Hội nghị Thượng đỉnh Trái đất (Earth Summit) tại Rio de Janeiro, Brazil: Thông qua Agenda 21, kế hoạch hành động toàn cầu về phát triển bền vững.
5. 2000 – Hội nghị Thiên niên kỷ tại New York: Các nước thành viên thông qua Tuyên bố Thiên niên kỷ, dẫn đến việc hình thành 8 Mục tiêu Phát triển Thiên niên kỷ (MDGs) đến năm 2015.
6. 2002 – Hội nghị Thượng đỉnh Thế giới về Phát triển Bền vững tại Johannesburg, Nam Phi: Thông qua Tuyên bố Johannesburg và Kế hoạch Hành động, củng cố cam kết xóa đói giảm nghèo và bảo vệ môi trường.
7. 2012 – Hội nghị Rio+20 tại Brazil: Thông qua văn kiện "Tương lai Chúng ta Mong muốn", khởi động quá trình xây dựng SDGs và thành lập Diễn đàn Chính trị Cấp cao LHQ về Phát triển Bền vững.
8. 2013 – Thành lập Nhóm Công tác Mở rộng gồm 30 thành viên: Có nhiệm vụ xây dựng đề xuất chi tiết về các Mục tiêu Phát triển Bền vững (SDGs).
9. 2015 – Năm bước ngoặt với loạt thỏa thuận toàn cầu:
    - Tháng 3: Khung Sendai về giảm rủi ro thiên tai.
    - Tháng 7: Chương trình Hành động Addis Ababa về tài chính cho phát triển.
    - Tháng 9: Thông qua Chương trình Nghị sự 2030 với 17 Mục tiêu Phát triển Bền vững (SDGs) tại Hội nghị Thượng đỉnh LHQ.
    - Tháng 12: Thỏa thuận Paris về Biến đổi Khí hậu.
10. Hiện nay:
    - Diễn đàn Chính trị Cấp cao là nền tảng trung tâm của LHQ để theo dõi tiến độ SDGs.
    - DSDG (Bộ Kinh tế và Xã hội LHQ) hỗ trợ triển khai, đánh giá và thúc đẩy thực hiện SDGs toàn cầu.

I.1.2. Tổng quan về SDGs

SDGs bao gồm 17 mục tiêu với 169 chỉ tiêu và 231 chỉ số để đo lường kết quả cụ thể, được chia thành ba trụ cột chính: Kinh tế (mục tiêu 8, 9, 12), Xã hội (mục tiêu 1, 2, 3, 4, 5, 10, 11, 16) và Môi trường (mục tiêu 6, 7, 13, 14, 15), mục tiêu 17 đóng vai trò kết nối và hỗ trợ các mục tiêu khác.

11. 17 mục tiêu (Goals): Là các định hướng tổng quát nhằm giải quyết các vấn đề toàn cầu về con người, hành tinh và sự thịnh vượng.
12. 169 chỉ tiêu cụ thể (Targets): Mỗi mục tiêu gồm nhiều chỉ tiêu nhằm làm rõ kết quả cần đạt được.
13. 231 chỉ số đo lường (Indicators): Là các chỉ số cụ thể, định lượng để đánh giá tiến độ thực hiện từng chỉ tiêu. Trong đó, một số chỉ số có thể lặp lại giữa các mục tiêu.

17 Mục tiêu Phát triển Bền vững (SDGs):

Tên mục tiêu	**Mục tiêu cần đạt được**
Mục tiêu 1: Xóa nghèo	Chấm dứt mọi hình thức nghèo ở mọi nơi.
Mục tiêu 2: Không còn nạn đói	Xóa đói, bảo đảm an ninh lương thực, cải thiện dinh dưỡng và thúc đẩy phát triển nông nghiệp bền vững.
Mục tiêu 3: Sức khỏe và có cuộc sống tốt	Bảo đảm cuộc sống khỏe mạnh và tăng cường phúc lợi cho mọi người ở mọi lứa tuổi
Mục tiêu 4: Giáo dục có chất lượng	Đảm bảo nền giáo dục có chất lượng, công bằng, toàn diện và thúc đẩy các cơ hội học tập suốt đời cho tất cả mọi người.
Mục tiêu 5: Bình đẳng giới	Đạt được bình đẳng giới; tăng quyền và tạo cơ hội cho phụ nữ và trẻ em gái.
Mục tiêu 6: Nước sạch và vệ sinh	Đảm bảo đầy đủ và quản lý bền vững tài nguyên nước và hệ thống vệ sinh cho tất cả mọi người.
Mục tiêu 7: Năng lượng sạch và giá thành hợp lý	Đảm bảo khả năng tiếp cận nguồn năng lượng bền vững, đáng tin cậy và có khả năng chi trả cho tất cả mọi người.
Mục tiêu 8: Công việc tốt và tăng trưởng kinh tế	Đảm bảo tăng trưởng kinh tế bền vững, toàn diện, liên tục; tạo việc làm đầy đủ, năng suất và việc làm tốt cho tất cả mọi người.
Mục tiêu 9: Công nghiệp, sáng tạo và phát triển hạ tầng	Xây dựng cơ sở hạ tầng có khả năng chống chịu cao, thúc đẩy công nghiệp hóa bao trùm và bền vững, tăng cường đổi mới.
Mục tiêu 10: Giảm bất bình đẳng	Giảm bất bình đẳng trong xã hội.
Mục tiêu 11: Các thành phố và cộng đồng bền vững	Phát triển đô thị, nông thôn bền vững, có khả năng chống chịu; đảm bảo môi trường sống và làm việc an toàn.
Mục tiêu 12: Tiêu dùng và sản xuất có trách nhiệm	Đảm bảo mô hình tiêu dùng và sản xuất bền vững.
Mục tiêu 13: Hành động về khí hậu	Thực hiện các hành động khẩn cấp để chống lại biến đổi khí hậu và tác động của nó.
Mục tiêu 14: Tài nguyên và môi trường biển	Bảo tồn và sử dụng bền vững các đại dương, biển và tài nguyên biển.

Mục tiêu 15: Tài nguyên và môi trường trên đất liền	Bảo tồn và sử dụng bền vững các hệ sinh thái trên cạn, quản lý rừng bền vững, chống lại sa mạc hóa và ngăn chặn sự suy thoái đất.
Mục tiêu 16: Hòa bình, công lý và các thể chế mạnh mẽ	Thúc đẩy xã hội hòa bình và bao trùm cho sự phát triển bền vững, cung cấp quyền tiếp cận công lý cho tất cả mọi người.
Mục tiêu 17: Quan hệ đối tác vì các mục tiêu	Tăng cường các phương thức thực hiện và tái sinh đối tác toàn cầu vì phát triển bền vững.

Bảng 1. 1 17 Mục tiêu Phát triển Bền vững (SDGs)

Hình 1. 1 Mục tiêu Phát triển Bền vững (SDGs) được Liên Hợp Quốc thông qua

I.1.3. Ý nghĩa của SDGs

Mục tiêu Phát triển Bền vững (SDGs) bao gồm 17 mục tiêu toàn cầu nhằm giải quyết các thách thức lớn nhất mà nhân loại đang đối mặt. Ý nghĩa của SDGs có thể được tóm tắt như sau:

- Thúc đẩy phát triển bền vững: SDGs cung cấp một khung toàn diện để phát triển kinh tế, xã hội và môi trường một cách đồng bộ, nhằm đảm bảo sự phát triển bền vững cho các thế hệ tương lai.
- Giảm nghèo và bất bình đẳng: Các mục tiêu này tập trung vào việc xóa đói và giảm bất bình đẳng, tạo ra cơ hội công bằng cho mọi người, từ đó nâng cao chất lượng cuộc sống cho các nhóm dân cư dễ bị tổn thương.
- Thúc đẩy hòa bình và công lý: SDGs nhấn mạnh tầm quan trọng của việc xây dựng một xã hội hòa bình, ổn định và công bằng, nơi mọi người đều có quyền tiếp cận công lý và bảo vệ quyền lợi của mình.

- Khuyến khích hợp tác toàn cầu: SDGs khuyến khích sự hợp tác giữa các quốc gia, tổ chức và cá nhân để cùng nhau giải quyết các thách thức toàn cầu, từ đó tạo ra một mạng lưới hỗ trợ và chia sẻ kinh nghiệm.
- Nâng cao nhận thức và hành động: SDGs giúp nâng cao nhận thức về các vấn đề toàn cầu như biến đổi khí hậu, nghèo đói và bất bình đẳng, từ đó khuyến khích các hành động cụ thể từ các cá nhân và tổ chức.
- Định hướng chính sách và đầu tư: SDGs cung cấp một khung tham chiếu rõ ràng cho các chính phủ và tổ chức trong việc xây dựng chính sách và ưu tiên đầu tư, nhằm đạt được các mục tiêu phát triển bền vững.
- Thúc đẩy đổi mới và sáng tạo: SDGs khuyến khích các giải pháp sáng tạo và đổi mới để giải quyết các thách thức toàn cầu, từ đó phát triển công nghệ mới và mô hình kinh doanh bền vững.
- Tạo ra cơ hội việc làm: Việc thực hiện các mục tiêu SDGs có thể tạo ra nhiều cơ hội việc làm mới trong các lĩnh vực như năng lượng tái tạo, nông nghiệp bền vững và công nghệ xanh.
- Bảo vệ môi trường: SDGs nhấn mạnh tầm quan trọng của việc bảo vệ môi trường và tài nguyên thiên nhiên, nhằm đảm bảo sức khỏe và phúc lợi cho con người.
- Khuyến khích sự tham gia của cộng đồng: SDGs khuyến khích sự tham gia của tất cả các bên liên quan, từ chính phủ, doanh nghiệp đến cộng đồng và cá nhân, tạo ra một phong trào toàn cầu hướng tới sự phát triển bền vững.

Tóm lại, SDGs không chỉ là một bộ mục tiêu mà còn là một tầm nhìn toàn cầu cho một tương lai bền vững, công bằng và hòa bình. Việc thực hiện các mục tiêu này đòi hỏi sự hợp tác và nỗ lực từ tất cả các quốc gia và cộng đồng trên thế giới.

I.2. Chỉ số SDG và phương pháp đánh giá

I.2.1. Phương pháp tổng quát đánh giá chỉ số SDG

Có nhiều cách đánh giá khác nhau tùy theo cấp độ và nguồn dữ liệu, nhưng thông thường gồm các bước sau:

- Bước 1: Xác định bộ chỉ số phù hợp

Dựa vào khung hướng dẫn của Liên Hợp Quốc hoặc quốc gia (VD: Tổng cục Thống kê Việt Nam đã ban hành bộ chỉ số SDG quốc gia gồm 158 chỉ tiêu).

- Bước 2: Thu thập dữ liệu

Nguồn dữ liệu có thể từ: tổng điều tra, báo cáo thống kê, hệ thống dữ liệu ngành (giáo dục, y tế...), dữ liệu vệ tinh hoặc open data.

- Bước 3: Chuẩn hóa và tính toán

Tính toán từng chỉ số: theo hướng dẫn kỹ thuật (ví dụ: tỷ lệ phần trăm, chỉ số bình quân đầu người, chỉ số tổng hợp - composite index…).

Tài liệu chính thức từ Liên Hợp Quốc "*Handbook on the Use of SDG Indicators in Monitoring in the Context of the 2030 Agenda for Sustainable Development*" [2] Hướng dẫn chi tiết cách dùng 231 chỉ số SDG:

- Cách thu thập và xử lý dữ liệu
- Định nghĩa và công thức tính từng chỉ số
- Nguồn dữ liệu phù hợp
- Phân tích xu hướng và so sánh
- Ví dụ ứng dụng thực tế

Một số phương pháp chuẩn hóa thường dùng:

- Min-Max scaling để đưa về cùng thang đo
- Z-score standardization
- Composite Index (gộp nhiều chỉ số thành 1 điểm tổng)

- Bước 4: Đánh giá mức độ đạt được

So sánh giá trị chỉ số thực tế với mục tiêu SDG để xem mức độ hoàn thành:

- Đạt (≥90%)
- Gần đạt (70-89%)
- Cần cải thiện (<70%)

- Bước 5: Trực quan & báo cáo

Dùng biểu đồ, bản đồ nhiệt (heatmap), bản đồ GIS để thể hiện sự thay đổi theo thời gian, theo vùng, theo SDG.

I.2.2. Một vài phương pháp tính chỉ số SDG tổng quát (định lượng)

I.2.2.1. Chỉ số tỷ lệ (%)

Công thức:

$$Tỷ\ lệ\ (\%) = \frac{Giá\ trị\ cần\ đo}{Tổng\ dân\ số\ hoặc\ tổng\ số\ đối\ tượng\ liên\ quan} * 100$$

Ví dụ: Tỷ lệ trẻ em đến trường đúng độ tuổi = (Số trẻ em đến trường đúng độ tuổi / Tổng số trẻ em trong độ tuổi đó) × 100

I.2.2.2. Chỉ số bình quân đầu người

Công thức:

$$Giá\ trị\ bình\ quân\ = \frac{\text{Tổng } giá\ trị}{\text{Tổng } dân\ số\ hoặc\ nhóm\ đối\ tượng}$$

Ví dụ: Lượng phát thải CO_2 bình quân đầu người = Tổng phát thải CO_2 / Dân số

I.2.2.3. Chỉ số tổng hợp (Composite Index)

Khi muốn đánh giá nhiều chỉ tiêu thành một điểm số duy nhất.

Các bước:

- Chuẩn hóa dữ liệu (VD: Min-Max scaling):

$$X' = \frac{\text{X} - Xmin}{\text{Xmax} - Xmin}$$

- Gán trọng số nếu cần
- Tính điểm tổng hợp: Điểm tổng hợp $= \sum_{i=0}^{n} \text{wi} * \text{X'i}$

Ví dụ: Chấm điểm "phát triển giáo dục" gồm: tỷ lệ biết chữ, tỷ lệ đến trường, chi tiêu giáo dục.

I.3. Phương pháp SHAP và lý thuyết trò chơi trong phân tích dữ liệu

I.3.1. Lý thuyết trò chơi và giá trị Shapley

Lý thuyết trò chơi là ngành toán học nghiên cứu tương tác chiến lược giữa các tác nhân ra quyết định. Giá trị Shapley, được đặt theo tên của nhà toán học Lloyd Shapley, là một khái niệm trọng tâm được sử dụng để phân bổ công bằng đóng góp của từng người chơi trong một liên minh.

Công thức giá trị Shapley:

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)]$$

Trong đó:

- N là tập hợp tất cả người chơi
- v là hàm đặc trưng xác định giá trị của mỗi liên minh
- ϕi là giá trị Shapley cho người chơi i
- S ⊆ N \{i}: Tập con không chứa người chơi i

- v(S): Hàm giá trị (giá trị mà tập con S tạo ra)
- |S|: Số lượng người chơi trong tập S
- |N|: Tổng số người chơi

I.3.2. Cơ chế hoạt động của SHAP

SHAP (SHapley Additive exPlanations) là một phương pháp dựa trên giá trị Shapley, dùng để giải thích các mô hình học máy bằng cách phân chia đóng góp của từng đặc trưng vào dự đoán cuối cùng. Nó có thể áp dụng cho nhiều loại mô hình, từ mô hình tuyến tính đến mô hình phức tạp như cây quyết định hay mạng neural. SHAP sử dụng giá trị Shapley vào mô hình học máy:

I.3.2.1. Ánh xạ đến lý thuyết trò chơi:

- "Người chơi" = Các đặc trưng (features)
- "Trò chơi" = Dự đoán của mô hình
- "Liên minh" = Tập con các đặc trưng

I.3.2.2. Tính toán đóng góp:

SHAP xác định ảnh hưởng của từng đặc trưng đến dự đoán bằng cách so sánh dự đoán khi có và không có các đặc trưng đó.

I.3.2.3. Tính chất quan trọng:

- Tính cộng tính: Tổng các giá trị SHAP bằng chênh lệch giữa dự đoán và giá trị trung bình cơ sở.
- Tính nhất quán: Nếu mô hình thay đổi để đặc trưng có tác động lớn hơn, giá trị SHAP sẽ tăng.
- Tính công bằng: Đặc trưng không ảnh hưởng sẽ có giá trị SHAP sẽ bằng 0.

I.3.2.4. Phương pháp ước lượng:

Do tính toán chính xác đòi hỏi phức tạp tổ hợp (2^n với n đặc trưng), SHAP sử dụng nhiều phương pháp ước lượng như:

- KernelSHAP: Sử dụng kỹ thuật hồi quy có trọng số.
- TreeSHAP: Thuật toán nhanh dành cho các mô hình dựa trên cây.
- DeepSHAP: Mở rộng cho mạng nơ-ron.

I.3.3. Ưu và nhược điểm

- Ưu điểm:
    - Nền tảng lý thuyết vững chắc
    - Tính chất đặc biệt (cộng tính, nhất quán, công bằng)
    - Linh hoạt với nhiều loại mô hình

- Nhược điểm:
    - Chi phí tính toán cao (đặc biệt với dữ liệu đa chiều)
    - Phụ thuộc vào dữ liệu nền (background data)
    - Có thể khó hiểu đối với người không chuyên

II. Tổng quan nghiên cứu

II.1. Nghiên cứu về phát triển bền vững các nước trong khu vực

II.1.1. Giới thiệu

Mỗi quốc gia trong khu vực đều có những chiến lược phát triển bền vững riêng, nhưng tất cả đều hướng đến việc đạt được một tương lai xanh, bao trùm và thịnh vượng. Các sáng kiến được triển khai chủ yếu liên quan đến năng lượng tái tạo, bảo vệ đa dạng sinh học, giảm thiểu biến đổi khí hậu và phát triển nền kinh tế tuần hoàn.

II.1.2. Thái Lan

Thái Lan đặc biệt chú trọng vào Mục tiêu SDG 12: Tiêu thụ và sản xuất bền vững thông qua chiến lược BCG (Bio-Circular-Green Economy). Mô hình này tập trung vào việc phát triển các ngành công nghiệp bền vững, đặc biệt trong nông nghiệp, công nghiệp và du lịch, đồng thời khuyến khích tái chế, tái sử dụng và giảm thiểu chất thải.

- SDG 13: Thái Lan đã thực hiện các chiến lược giảm thiểu khí thải và bảo vệ môi trường, ví dụ như việc thúc đẩy năng lượng tái tạo và hạn chế sử dụng các nguồn năng lượng không bền vững.
- SDG 8: Thái Lan đã triển khai các chính sách tạo ra việc làm và nâng cao đời sống cho người dân, đặc biệt là trong các ngành công nghiệp xanh.

II.1.3. Indonesia

Indonesia đóng góp vào SDG 7: Năng lượng sạch và giá cả phải chăng thông qua việc phát triển Nhiên liệu Hàng không Bền vững (SAF), sử dụng nhiên liệu sinh học từ dầu cọ để thay thế dần nhiên liệu hóa thạch. Đây là một phần trong chiến lược quốc gia nhằm giảm thiểu tác động của ngành hàng không đối với khí thải carbon.

- SDG 14: Indonesia tham gia tích cực vào Sáng kiến Tam giác San hô để bảo vệ các rạn san hô và đảm bảo sự đa dạng sinh học biển. Sáng kiến này có mục tiêu bảo vệ khoảng 30% hệ sinh thái biển trong khu vực Tam giác San hô.
- SDG 15: Indonesia thực hiện các biện pháp bảo vệ rừng và hệ sinh thái tự nhiên, bao gồm việc ngừng nạn phá rừng và duy trì các khu bảo tồn thiên nhiên.

II.1.4. Malaysia

Malaysia chú trọng vào SDG 7: Năng lượng sạch và giá cả phải chăng, với mục tiêu sản xuất Nhiên liệu Hàng không Bền vững (SAF) vào năm 2027. Việc sản xuất SAF từ dầu cọ không chỉ giảm phát thải khí nhà kính mà còn hỗ trợ nền kinh tế nông nghiệp bền vững.

- SDG 13: Malaysia cũng đặt mục tiêu giảm 45% lượng khí thải carbon so với mức năm 2005 vào năm 2030, thông qua việc phát triển các nguồn năng lượng tái tạo như điện mặt trời và thủy điện.
- SDG 12: Quốc gia này thúc đẩy sản xuất bền vững trong các ngành công nghiệp, bao gồm việc sử dụng tài nguyên thiên nhiên một cách hiệu quả và giảm thiểu chất thải.

II.1.5. Singapore

Singapore là quốc gia đi đầu trong việc thúc đẩy SDG 11: Các thành phố và cộng đồng bền vững thông qua các sáng kiến phát triển đô thị thông minh và bền vững. Quốc gia này đã triển khai chiến dịch Go Green SG, khuyến khích cộng đồng giảm thiểu chất thải, tiết kiệm năng lượng và sử dụng các sản phẩm xanh.

- SDG 13: Singapore cam kết giảm 36% phát thải khí nhà kính vào năm 2030 so với mức năm 2005 thông qua việc thúc đẩy sử dụng năng lượng tái tạo và phát triển các công nghệ sạch.
- SDG 6: Singapore phát triển các công nghệ tiên tiến trong việc tái chế nước và quản lý tài nguyên nước, giúp đảm bảo nguồn nước sạch cho cả cộng đồng.

II.1.6. Philippines

Philippines tham gia tích cực vào SDG 14: Sự sống dưới nước thông qua Sáng kiến Tam giác San hô, hợp tác với Indonesia và Malaysia để bảo vệ các hệ sinh thái biển. Sáng kiến này không chỉ bảo vệ môi trường mà còn hỗ trợ cộng đồng ven biển trong việc duy trì sinh kế và phát triển du lịch bền vững.

- SDG 15: Philippines cũng có các chương trình bảo vệ rừng và bảo tồn động vật hoang dã, nhằm duy trì sự đa dạng sinh học và giảm thiểu tác động của biến đổi khí hậu.

II.2. Nghiên cứu về SDGs ở Việt Nam

Quyết định số 841/QĐ-TTg được ban hành trong bối cảnh Việt Nam tiếp tục khẳng định cam kết mạnh mẽ với Chương trình Nghị sự 2030 vì sự phát triển bền vững của Liên Hợp Quốc, đồng thời cần điều chỉnh các mục tiêu và phương pháp thực

hiện cho phù hợp với điều kiện kinh tế - xã hội đang thay đổi nhanh chóng, đặc biệt sau đại dịch COVID-19 và trong bối cảnh chuyển đổi xanh, chuyển đổi số toàn diện.

II.2.1. Mục đích và vai trò của quyết định

Quyết định này đóng vai trò là văn bản định hướng quan trọng cho toàn bộ hệ thống chính trị trong việc cụ thể hóa và triển khai thực hiện các Mục tiêu phát triển bền vững (Sustainable Development Goals – SDGs) đến năm 2030. Thay thế Quyết định số 681/QĐ-TTg được ban hành năm 2019, văn bản mới đưa ra một lộ trình cập nhật, điều chỉnh cả về nội dung chỉ tiêu lẫn phương thức tổ chức thực hiện nhằm bảo đảm tính khả thi và hiệu quả cao hơn trong triển khai trên thực tế.

II.2.2. Hệ thống mục tiêu và chỉ tiêu phát triển bền vững

Việt Nam tiếp tục bám sát 17 mục tiêu phát triển bền vững của Liên Hợp Quốc trong lộ trình đến năm 2030. Tuy nhiên, để phù hợp hơn với điều kiện thực tiễn, số lượng chỉ tiêu cụ thể đã được tinh giản từ 158 (năm 2019) xuống còn 115 chỉ tiêu trong Quyết định số 841/QĐ-TTg năm 2023. Việc điều chỉnh này nhằm tăng tính tập trung, dễ theo dõi và phù hợp với hệ thống thống kê quốc gia hiện hành.

Mỗi mục tiêu đều gắn với các chỉ tiêu định lượng rõ ràng, có mốc thời gian thực hiện cụ thể. Dưới đây là tổng hợp 17 mục tiêu phát triển bền vững của Việt Nam đến năm 2030:

- Mục tiêu 1: Xóa nghèo: Tỷ lệ nghèo đa chiều dự kiến giảm 1 – 1,5% mỗi năm đến năm 2030.
- Mục tiêu 2: Xóa đói, bảo đảm an ninh lương thực và nông nghiệp bền vững: Tỷ lệ thiếu năng lượng trong khẩu phần ăn giảm mạnh; tăng tỷ lệ người dân tiếp cận thực phẩm đủ dinh dưỡng quanh năm.
- Mục tiêu 3: Sức khỏe và phúc lợi: Giảm tỷ lệ tử vong mẹ xuống dưới 70/100.000 ca sinh sống; tỷ lệ tử vong trẻ em dưới 5 tuổi dưới 15/1.000 vào năm 2030.
- Mục tiêu 4: Giáo dục chất lượng: Phấn đấu đến năm 2030, 100% trẻ em 5 tuổi được đi học mẫu giáo; nâng tỷ lệ hoàn thành giáo dục phổ thông cơ sở lên trên 95%.
- Mục tiêu 5: Bình đẳng giới: Giảm khoảng cách thu nhập và cơ hội giữa nam và nữ; tỷ lệ nữ đại biểu Quốc hội và HĐND đạt ít nhất 35%.
- Mục tiêu 6: Nước sạch và vệ sinh: Đến năm 2030, ít nhất 95% dân số nông thôn sử dụng nước sạch đạt quy chuẩn; tỷ lệ hộ dân có nhà vệ sinh hợp vệ sinh đạt 100%.
- Mục tiêu 7: Năng lượng sạch và giá cả hợp lý: Tỷ lệ hộ tiếp cận điện đạt 100%; tỷ trọng năng lượng tái tạo trong tổng cung năng lượng sơ cấp đạt 15–20% vào năm 2030.

- Mục tiêu 8: Tăng trưởng kinh tế và việc làm bền vững: GDP tăng trưởng trung bình 7%/năm giai đoạn 2021–2030; GDP bình quân đầu người đạt 7.500 USD vào năm 2030.
- Mục tiêu 9: Công nghiệp, đổi mới và cơ sở hạ tầng: Đầu tư cho R&D chiếm 1,5% GDP; số doanh nghiệp công nghệ cao tăng 1,5 lần so với năm 2020.
- Mục tiêu 10: Giảm bất bình đẳng: Thu hẹp chênh lệch thu nhập giữa các nhóm dân cư; tăng tỷ lệ tiếp cận dịch vụ công cho các nhóm yếu thế.
- Mục tiêu 11: Đô thị và cộng đồng bền vững: 100% đô thị loại đặc biệt, loại I có quy hoạch phát triển đô thị xanh, thông minh; tỷ lệ chất thải rắn được xử lý đạt 100%.
- Mục tiêu 12: Tiêu dùng và sản xuất bền vững: Tỷ lệ doanh nghiệp áp dụng sản xuất sạch hơn đạt 70%; 100% siêu thị không còn bao bì nylon dùng một lần vào năm 2025.
- Mục tiêu 13: Hành động ứng phó biến đổi khí hậu: Giảm phát thải khí nhà kính theo cam kết tại COP26, hướng tới mức phát thải ròng bằng "0" vào năm 2050.
- Mục tiêu 14: Bảo vệ tài nguyên biển: Ít nhất 10% diện tích biển được bảo tồn hiệu quả; giảm 50% lượng rác thải nhựa ra biển so với năm 2020.
- Mục tiêu 15: Quản lý bền vững tài nguyên đất và đa dạng sinh học: Tỷ lệ che phủ rừng duy trì ở mức 42–43%; phục hồi ít nhất 30% diện tích đất bị suy thoái.
- Mục tiêu 16: Thúc đẩy thể chế công bằng, hòa bình và hiệu quả: Tăng cường minh bạch, phòng chống tham nhũng; giảm số vụ bạo lực và phân biệt đối xử trong xã hội.
- Mục tiêu 17: Tăng cường đối tác toàn cầu: Nâng cao hiệu quả hợp tác quốc tế; tăng cường huy động nguồn lực tài chính và kỹ thuật cho phát triển bền vững.

II.2.3. Cơ chế giám sát, đánh giá và báo cáo tiến độ

Để theo dõi hiệu quả thực hiện, Quyết định 841/QĐ-TTg quy định rõ trách nhiệm giám sát và đánh giá. Các bộ, ngành được yêu cầu xây dựng kế hoạch theo dõi định kỳ đối với từng chỉ tiêu mà mình phụ trách. Bộ Kế hoạch và Đầu tư đóng vai trò trung tâm trong việc tổng hợp, phân tích, báo cáo tình hình thực hiện chung với Chính phủ và Quốc hội.

Đặc biệt, cơ chế báo cáo được thiết kế để bảo đảm tính linh hoạt, thường xuyên, có thể cập nhật và điều chỉnh phù hợp với thực tiễn. Các chỉ tiêu được theo dõi không chỉ để tổng kết, mà còn nhằm phát hiện sớm các thách thức trong triển khai, từ đó kịp thời đề xuất điều chỉnh chính sách.

II.2.4. Tăng cường ứng dụng công nghệ thông tin và chuyển đổi số

Một trong những điểm mới quan trọng của Quyết định là việc thúc đẩy mạnh mẽ ứng dụng công nghệ thông tin, xây dựng nền tảng dữ liệu phát triển bền vững thống nhất và liên thông. Dữ liệu số hóa giúp tăng tính minh bạch, giảm thời gian tổng hợp báo cáo, đồng thời tạo điều kiện thuận lợi cho việc chia sẻ thông tin giữa các cơ quan trong nước cũng như với các đối tác quốc tế.

Việc xây dựng hệ thống dữ liệu tập trung không chỉ nhằm phục vụ mục tiêu giám sát mà còn là công cụ hỗ trợ hoạch định chính sách có cơ sở khoa học và bằng chứng rõ ràng.

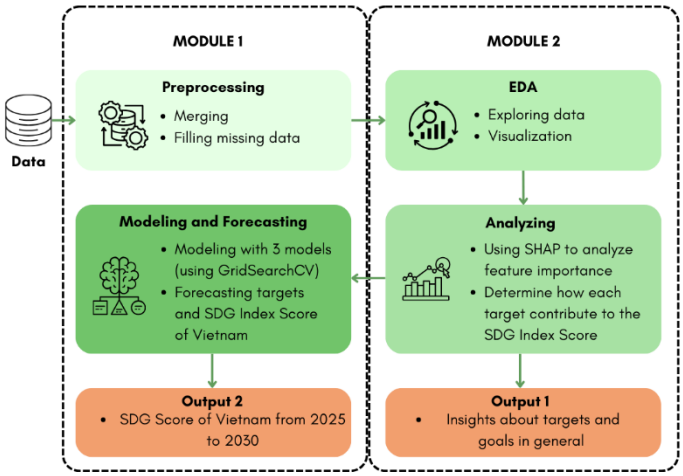
II.2.5. Tính cập nhật và thích ứng với bối cảnh mới

Quyết định 841/QĐ-TTg thể hiện sự điều chỉnh chính sách quan trọng của Việt Nam để phù hợp với bối cảnh quốc tế và trong nước đang biến đổi nhanh chóng. Các nội dung liên quan đến biến đổi khí hậu, chuyển đổi xanh, công nghệ số, hậu COVID-19 và các cam kết toàn cầu như giảm phát thải ròng về "0" đến năm 2050 tại COP26 đã được lồng ghép vào các chỉ tiêu và định hướng hành động.

Đây là minh chứng cho thấy chính sách phát triển bền vững của Việt Nam không chỉ theo kịp mà còn tích cực định hình những xu thế toàn cầu, hướng tới mục tiêu phát triển hài hòa giữa kinh tế, xã hội và môi trường.

Chính sách Phát triển Chính phủ Điện tử: Đẩy mạnh cải cách công nghệ thông tin và phát triển chính phủ điện tử nhằm nâng cao hiệu quả quản lý nhà nước, tăng cường giao tiếp giữa chính phủ và người dân.

CHƯƠNG 3: METHODOLOGY



Chương 3 trình bày phương pháp thực hiện bài toán với 2 module chính:

Module 1: Áp dụng mô hình để dự đoán giá trị SDG Index Score của Việt Nam từ năm 2025-2030 sau đó thảo luận kết quả thu được nhằm trả lời các câu hỏi liệu Việt Nam có đạt được mục tiêu phát triển bền vững tới năm 2030 hay không? Và làm sao để thúc đẩy và đảm bảo quá trình đi tới mục tiêu được suôn sẻ thông qua các gợi ý về giải pháp và chiến lược cho tương lai.

Module 2: Tập trung vào việc phân tích các yếu tố ảnh hưởng đến điểm SDG nói chung cho tập dữ liệu của các quốc gia trên toàn thế giới thông qua quy trình EDA và phân tích SHAP, từ đó rút ra những 'insight' về các chỉ số, mục tiêu SDG.

I. Phân tích dữ liệu toàn cầu

I.1. Phân tích thống kê mô tả

Trong giai đoạn khám phá dữ liệu (Exploratory Data Analysis – EDA), nhóm đã áp dụng một số phương pháp thống kê mô tả nhằm tóm tắt, minh họa và hiểu sâu hơn về đặc điểm phân phối của các chỉ số phát triển bền vững (SDG) toàn cầu. Các phương pháp cụ thể bao gồm:

1. Thống kê mô tả cơ bản
   Áp dụng để mô tả đặc trưng trung tâm và phân tán của từng chỉ số SDG:
- Trung bình (Mean): Đo lường giá trị điển hình của từng chỉ số.
- Trung vị (Median): Đánh giá xu hướng trung tâm một cách ổn định hơn với dữ liệu có ngoại lệ.
- Độ lệch chuẩn (Standard Deviation): Đánh giá mức độ phân tán, phản ánh tính không đồng đều giữa các quốc gia.
- Giá trị nhỏ nhất/lớn nhất (Min/Max) và tứ phân vị (Quartiles): Hỗ trợ đánh giá biên độ dao động và phát hiện các bất thường (outlier).
2. Trực quan hóa phân phối dữ liệu
   Để minh họa đặc điểm phân phối, dữ liệu được biểu diễn qua các biểu đồ:
- Histogram (biểu đồ tần suất): Cho thấy hình dạng phân phối (chuẩn, lệch phải, lệch trái).
- Boxplot (biểu đồ hộp): Là công cụ hiệu quả để phát hiện ngoại lệ và so sánh giữa các nhóm SDG.

I.2. Phân tích tương quan và nhân quả

Trong quá trình khám phá dữ liệu, nhóm nghiên cứu đã áp dụng phân tích tương quan tuyến tính Pearson để đánh giá mối liên hệ giữa các chỉ số phát triển bền vững (SDG). Cụ thể:

- Hệ số tương quan Pearson (r) được tính toán cho toàn bộ cặp chỉ số SDG nhằm xác định mức độ và chiều hướng liên hệ tuyến tính giữa chúng.
- Ma trận tương quan được hiển thị bằng heatmap trực quan giúp dễ dàng phát hiện các cụm chỉ số có tương quan mạnh ($r > 0.7$) hoặc tương quan ngược ($r < –0.5$).

II. Tiền xử lý dữ liệu

Quy trình tiền xử lý dữ liệu gồm các bước:

- Hợp nhất dữ liệu: Gộp Backdated SDG Index với Raw Data - Panel theo Country và year và Bỏ các cột không cần thiết
- Điền dữ liệu năm 2024: Sử dụng dữ liệu từ Full Database để gán vào năm 2024 và đồng bộ điểm "2024 SDG Index Score" vào cột "SDG Index Score" trong dataset
- Điền dữ liệu thiếu theo chỉ dẫn Codebook: Dựa trên từng chỉ số, sử dụng các bộ dữ liệu khác.
- Xử lý các cột/dòng có dữ liệu thiếu lớn: Xoá cột có lớn hơn 60% dữ liệu thiếu và correlation không mạnh với SDG Index Score ($-0.7 < \text{correlation} < 0.7$), loại bỏ các dòng không có SDG Index Score và xóa các cột có correlation thấp (trong đoạn từ -0.4 đến 0.4).
- Điền dữ liệu thiếu bằng nội suy và KMeans.

III. Mô Hình Học Máy

III.1. Linear Regression

- Nguyên lí hoạt động:
    - Hồi quy tuyến tính là một mô hình học máy cơ bản dùng để dự đoán giá trị liên tục dựa trên một hoặc nhiều biến đầu vào (tính năng).
    - Mô hình này giả định rằng mối quan hệ giữa biến phụ thuộc (y) và các biến độc lập (x) là tuyến tính. Nghĩa là, giá trị của y có thể được biểu diễn dưới dạng một hàm số bậc nhất của x, với công thức:
    - $y=\beta0+\beta1x1+\beta2x2+\cdots+\beta nxn+\epsilon$
    - β_0 là hệ số chệch (intercept).
    - β_1, β_2,..., βn là các hệ số của các biến độc lập.
    - ε là sai số (error term).
- Ứng dụng:
    - Dự báo giá trị liên tục: Ví dụ như dự đoán giá nhà, dự báo doanh thu, dự báo thị trường chứng khoán.

- Ứng dụng trong phân tích dữ liệu: Hồi quy tuyến tính có thể được dùng để phân tích mối quan hệ giữa các biến và tìm hiểu cách các biến độc lập ảnh hưởng đến biến phụ thuộc.

- Ưu điểm:
  - Dễ hiểu và dễ giải thích.
  - Tính toán nhanh và yêu cầu tài nguyên máy tính thấp.
- Nhược điểm:
  - Giới hạn trong việc mô hình hóa các mối quan hệ phi tuyến tính.
  - Dễ bị ảnh hưởng bởi các ngoại lệ (outliers).

III.2. Random Forest Regressor

- Nguyên lí hoạt động:
  - Random Forest Regressor là một mô hình ensemble sử dụng phương pháp bagging (Bootstrap Aggregating).
  - Mô hình này tạo ra nhiều cây quyết định (decision trees) và kết hợp chúng lại để dự đoán kết quả cuối cùng.
    - Mỗi cây quyết định trong rừng được huấn luyện trên một tập con ngẫu nhiên của dữ liệu huấn luyện, thông qua phương pháp bootstrap sampling.
    - Dự đoán cuối cùng của Random Forest được tính bằng cách lấy trung bình (mean) của các dự đoán từ tất cả các cây trong rừng.
- Cách hoạt động:
  - Mỗi cây quyết định được xây dựng bằng cách chia dữ liệu thành các nhánh dựa trên các câu hỏi về giá trị của các tính năng.
  - Random Forest giúp giảm thiểu overfitting (quá khớp dữ liệu) do mô hình không dựa vào một cây duy nhất, mà là một tập hợp của nhiều cây quyết định.
- Ứng dụng:
  - Dự báo giá trị liên tục như trong các bài toán về tài chính, marketing, và phân tích dữ liệu chuỗi thời gian.
  - Phân tích các vấn đề có tính chất không tuyến tính và có nhiều tương tác giữa các tính năng.

- Ưu điểm:
    - Hiệu quả trong việc xử lý các mối quan hệ phi tuyến tính.
    - Khả năng xử lý tốt dữ liệu có nhiều đặc tính phức tạp và không tương quan.
    - Không dễ bị overfitting so với các mô hình học máy khác.
- Nhược điểm:
    - Thời gian huấn luyện dài, đặc biệt khi số lượng cây trong rừng tăng lên.
    - Mô hình khó giải thích trực tiếp vì kết quả là sự kết hợp của nhiều cây quyết định.

III.3. XGBoost

- Nguyên lí hoạt động:
    - XGBoost là một thuật toán học máy dựa trên phương pháp boosting, một dạng học máy của mô hình ensemble.
    - Boosting là phương pháp xây dựng mô hình học máy bằng cách kết hợp nhiều mô hình yếu (weak learners), thường là các cây quyết định, trong một cách thức tuần tự.
        - XGBoost xây dựng các cây quyết định theo từng bước và cải thiện mô hình dựa trên lỗi (residual) của mô hình trước đó.
        - Mỗi bước, XGBoost tìm cách giảm thiểu lỗi của mô hình trước đó và tối ưu hóa việc học từ sai số.
- Cách hoạt động:
    - Mỗi cây quyết định mới trong mô hình sẽ "học" từ lỗi của các cây quyết định trước đó, tức là nó cố gắng sửa chữa các dự đoán sai của các cây trước.
    - XGBoost sử dụng một số kỹ thuật tiên tiến như regularization để điều chỉnh mô hình và tránh overfitting, đồng thời cải thiện hiệu suất tính toán.
- Ứng dụng:
    - XGBoost đặc biệt mạnh mẽ trong các bài toán phân loại và hồi quy, và được sử dụng rộng rãi trong các cuộc thi học máy như Kaggle.
    - Dự báo tài chính, nhận diện hình ảnh, phân tích cảm xúc từ văn bản.
- Ưu điểm:
    - Xử lý rất tốt dữ liệu thiếu và dữ liệu không tuyến tính.

- Hiệu suất vượt trội với việc tối ưu hóa thời gian và bộ nhớ.
- Khả năng điều chỉnh overfitting thông qua regularization và điều chỉnh siêu tham số.

- Nhược điểm:
  - Cần có kinh nghiệm trong việc tối ưu hóa các siêu tham số để đạt được hiệu quả tốt nhất.
  - Mô hình khó giải thích do tính phức tạp của cách học và số lượng cây quyết định kết hợp.

III.4. Sử dụng biến lag trong chuỗi thời gian

Biến lag là những biến trong phân tích chuỗi thời gian được tạo ra bằng cách dịch các giá trị của biến mục tiêu (hoặc các biến khác) theo một số khoảng thời gian. Ví dụ, một biến lag có thể là giá trị của biến tại thời điểm trước đó (lag = 1), hai thời điểm trước đó (lag = 2), v.v. Các biến này giúp mô hình dự báo mối quan hệ giữa giá trị hiện tại và quá khứ, đặc biệt trong các mô hình dự báo chuỗi thời gian như ARIMA, XGBoost.

IV. Đánh Giá Mô Hình

IV.1. MSE

- Mean Squared Error (MSE): MSE là chỉ số đo lường sai số bình phương giữa giá trị dự đoán của mô hình và giá trị thực tế. Nó được tính bằng cách lấy trung bình của các bình phương sai số (tức là hiệu giữa giá trị dự đoán và giá trị thực tế) trên tất cả các mẫu dữ liệu.

- Công thức:

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

- Trong đó:
  - n: Số lượng mẫu.
  - yi: Giá trị thực tế
  - y’i: Giá trị dự đoán

- Ưu điểm:

- MSE nhấn mạnh mạnh mẽ các lỗi lớn (outliers) do việc bình phương các sai số. Do đó, nếu có các điểm dữ liệu ngoại lệ lớn, MSE sẽ phản ánh điều này rõ ràng.
- MSE cung cấp một cách tiếp cận chính xác khi muốn hạn chế các lỗi lớn.

- Nhược điểm:
  - Vì sai số được bình phương, MSE có thể bị ảnh hưởng nhiều bởi các ngoại lệ lớn trong dữ liệu.

IV.2. MAE

- MAE là chỉ số đo lường sai số tuyệt đối trung bình giữa giá trị dự đoán và giá trị thực tế. Nó tính trung bình của các sai số tuyệt đối thay vì bình phương sai số.
- Công thức

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - y'_i|$$

- Trong đó:
  - n: Số lượng mẫu.
  - yi: Giá trị thực tế
  - y'i: Giá trị dự đoán
- Ưu điểm:
  - MAE dễ hiểu và không nhấn mạnh quá mức các lỗi ngoại lệ. Nó phản ánh sai số theo cách dễ hiểu, không làm tăng quá mức tầm quan trọng của những sai số lớn.
  - MAE không bị ảnh hưởng quá nhiều bởi các giá trị ngoại lệ (outliers) như MSE.
- Nhược điểm:
  - Không nhấn mạnh sự quan trọng của các sai số lớn, có thể làm giảm độ nhạy trong việc phát hiện các lỗi lớn.

IV.3. R^2 Score

- R^2 Score, hay còn gọi là hệ số xác định, là chỉ số thể hiện mức độ mô hình hồi quy có thể giải thích sự biến thiên của dữ liệu. R^2 đo lường phần trăm sự biến

động trong biến phụ thuộc mà mô hình có thể giải thích được từ các biến độc lập.

- Công thức:

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

- Trong đó:
  - yi: Giá trị thực tế
  - y'i: Giá trị dự đoán
  - $\bar{y}_i$: Giá trị trung bình của biến phụ thuộc
- Ưu điểm:
  - R^2 dễ hiểu và phổ biến trong các bài toán hồi quy. Nó cung cấp thông tin về mức độ mô hình phù hợp với dữ liệu.
  - Nếu R^2 gần 1, mô hình có thể giải thích hầu hết sự biến động trong dữ liệu.
- Nhược điểm:
  - R^2 không phải lúc nào cũng phản ánh chính xác hiệu quả của mô hình, đặc biệt khi mô hình bị overfitting.
  - Nó có thể bị ảnh hưởng nếu có quá nhiều biến trong mô hình mà không có sự cải thiện rõ rệt.

IV.4. GridSearchCV

- GridSearchCV là một kỹ thuật tìm kiếm siêu tham số (hyperparameter) tối ưu cho mô hình học máy, giúp chọn ra sự kết hợp siêu tham số tốt nhất.
- GridSearchCV thử nghiệm tất cả các khả năng kết hợp của các tham số đã được xác định trước (grid) và sử dụng phương pháp cross-validation để đánh giá hiệu quả của từng sự kết hợp.
- Ví dụ, khi có hai tham số để tối ưu như max_depth (độ sâu tối đa của cây) và n_estimators (số lượng cây trong Random Forest), ta sẽ tạo ra một "lưới" (grid) các giá trị khả thi cho các tham số này và GridSearchCV sẽ thử tất cả các sự kết hợp và tìm ra sự kết hợp tốt nhất.
- Ưu điểm:
  - Giúp tối ưu hóa các tham số mô hình mà không cần phải tìm kiếm thủ công.

- Tối ưu hóa mô hình với độ chính xác cao bằng cách đánh giá tất cả các lựa chọn siêu tham số.

- Nhược điểm:
  - Mất nhiều thời gian nếu số lượng tham số và giá trị tham số lớn, vì phải thử tất cả các kết hợp.
  - Có thể gây tốn kém về tài nguyên tính toán đối với các mô hình phức tạp.

V. Phương pháp dự báo

V.1. Dự Báo Biến Ngoại Sinh

Biến ngoại sinh (exogenous variables) là những yếu tố bên ngoài mà không phải là kết quả trực tiếp từ chuỗi thời gian mà chúng ta đang phân tích, nhưng lại có ảnh hưởng lớn đến kết quả của chuỗi thời gian đó. Những biến này có thể là các yếu tố kinh tế, xã hội, khí hậu, hay chính trị, v.v., có thể tác động đến biến mục tiêu (target variable) mà ta đang dự báo.

V.2. Sử dụng mô hình đã huấn luyện để dự báo

- Sau khi huấn luyện mô hình, có thể sử dụng mô hình đó để dự báo giá trị tương lai bằng cách đưa vào các giá trị đầu vào mới và nhận kết quả dự đoán.
- Mô hình sẽ sử dụng các tham số mà nó đã học để dự đoán giá trị tương lai của chuỗi thời gian hoặc các yếu tố ảnh hưởng.
- Lợi ích:
  - Dự báo chính xác: Sau khi huấn luyện, mô hình có thể cung cấp dự báo tương đối chính xác cho các giá trị trong tương lai.
  - Tiết kiệm thời gian: Việc sử dụng mô hình đã huấn luyện giúp tiết kiệm thời gian và tài nguyên so với việc huấn luyện lại mô hình từ đầu mỗi khi có dữ liệu mới.

V.3. Vấn đề sai số tích lũy trong dự báo dài hạn

- Sai số tích lũy:
  - Sai số tích lũy là vấn đề thường gặp trong các mô hình dự báo dài hạn. Khi dự báo một chuỗi thời gian trong tương lai, sai số của dự đoán tại mỗi bước sẽ tích lũy và tác động đến các dự báo tiếp theo.
  - Nguyên nhân: Mỗi dự đoán dựa vào dữ liệu dự báo trước đó. Nếu sai số trong một dự đoán là lớn, sai số này sẽ tiếp tục lan rộng trong các dự báo tiếp theo, dẫn đến kết quả sai lệch.

- Ảnh hưởng của sai số tích lũy:
    - Dự báo dài hạn sẽ trở nên kém chính xác khi sai số tích lũy theo thời gian. Ví dụ, khi dự báo giá trị trong 1 tuần tới, mô hình có thể dự đoán chính xác. Nhưng khi kéo dài thời gian dự báo ra 1 tháng hoặc 6 tháng, sai số sẽ dần tích lũy và làm giảm độ chính xác của các dự báo sau.
- Giải pháp giảm thiểu sai số tích lũy:
    - Smoothing giúp làm giảm độ biến động của dữ liệu và giảm tác động của sai số. Các phương pháp smoothing như **exponential smoothing** có thể giúp mô hình dự báo chính xác hơn trong dài hạn bằng cách làm mượt dữ liệu lịch sử.
    - Cập nhật mô hình theo thời gian là một giải pháp để giảm thiểu sai số tích lũy. Khi có dữ liệu mới, mô hình sẽ được huấn luyện lại và cải thiện khả năng dự báo, giúp giảm thiểu sai số tích lũy.
    - Một số mô hình ensemble như Random Forest hoặc XGBoost có thể kết hợp thông tin từ nhiều nguồn khác nhau để tạo ra dự báo chính xác hơn. Kết hợp các mô hình có thể giúp giảm thiểu sai số tích lũy và nâng cao độ chính xác của dự báo dài hạn.
    - Thay vì dự báo một khoảng thời gian dài liên tục, có thể chia dự báo thành các chu kỳ ngắn hơn và cập nhật mô hình sau mỗi chu kỳ để giảm thiểu sai số tích lũy.

VI. Sự ảnh hưởng các biến với cột target (SHAP)

- SHAP là một phương pháp giải thích kết quả dự đoán của mô hình học máy, giúp chúng ta hiểu được cách thức mà mỗi đặc trưng (tính năng) của dữ liệu đóng góp vào quyết định cuối cùng của mô hình. Phương pháp này dựa trên lý thuyết giá trị Shapley từ lý thuyết trò chơi, và giúp giải thích mô hình học máy theo cách dễ hiểu và có tính toán chính xác.

- Lý thuyết Shapley: Được phát triển trong lý thuyết trò chơi, giá trị Shapley đo lường sự đóng góp công bằng của mỗi thành viên (hoặc tính năng trong trường hợp của mô hình học máy) vào kết quả cuối cùng. Trong SHAP, mỗi tính năng được gán một giá trị thể hiện sự đóng góp của nó vào dự đoán cuối cùng, giúp chúng ta hiểu mô hình hoạt động như thế nào.
- SHAP values là các giá trị cụ thể cho từng tính năng trong từng dự đoán, và tổng giá trị của chúng bằng với sự thay đổi trong giá trị dự đoán của mô hình khi các tính năng khác nhau được đưa vào hoặc bỏ đi.

- Ưu điểm:

- SHAP là một phương pháp giải thích kết quả dự đoán của mô hình học máy, giúp chúng ta hiểu được cách thức mà mỗi đặc trưng (tính năng) của dữ liệu đóng góp vào quyết định cuối cùng của mô hình. Phương pháp này dựa trên lý thuyết giá trị Shapley từ lý thuyết trò chơi, và giúp giải thích mô hình học máy theo cách dễ hiểu và có tính toán chính xác.
- Lý thuyết Shapley: Được phát triển trong lý thuyết trò chơi, giá trị Shapley đo lường sự đóng góp công bằng của mỗi thành viên (hoặc tính năng trong trường hợp của mô hình học máy) vào kết quả cuối cùng. Trong SHAP, mỗi tính năng được gán một giá trị thể hiện sự đóng góp của nó vào dự đoán cuối cùng, giúp chúng ta hiểu mô hình hoạt động như thế nào.
- SHAP values là các giá trị cụ thể cho từng tính năng trong từng dự đoán, và tổng giá trị của chúng bằng với sự thay đổi trong giá trị dự đoán của mô hình khi các tính năng khác nhau được đưa vào hoặc bỏ đi.

- Các loại đồ thị giải thích bao gồm:

- Summary Plot:
    - Mục đích: Hiển thị ảnh hưởng tổng quan của tất cả các tính năng đối với dự đoán của mô hình.
    - Cách hoạt động: Hiển thị sự phân phối giá trị SHAP của các tính năng, giúp xác định tính năng quan trọng nhất.
- Force Plot:
    - Mục đích: Giải thích cách từng tính năng tác động đến dự đoán cho một trường hợp cụ thể.
    - Cách hoạt động: Hiển thị cách các tính năng tăng hoặc giảm giá trị dự đoán, thể hiện sự đóng góp của từng tính năng.
- Dependence Plot:
    - Mục đích: Giải thích cách từng tính năng tác động đến dự đoán cho một trường hợp cụ thể.
    - Cách hoạt động: Hiển thị cách các tính năng tăng hoặc giảm giá trị dự đoán, thể hiện sự đóng góp của từng tính năng.
- Waterfall Plot:
    - Mục đích: Giải thích cách từng tính năng tác động đến dự đoán cho một trường hợp cụ thể.

- Cách hoạt động: Hiển thị cách các tính năng tăng hoặc giảm giá trị dự đoán, thể hiện sự đóng góp của từng tính năng.

CHƯƠNG 4: QUÁ TRÌNH TIỀN XỬ LÝ VÀ KHÁM PHÁ DỮ LIỆU

I. Tiền xử lý dữ liệu (Preprocessing)

I.1. Tổng quan bộ dữ liệu ban đầu

Bộ dữ liệu SDR2024 bao gồm 4 sheet chính: Codebook (Giải thích ý nghĩa các cột thuộc tính trong bộ dữ liệu), Full Database (Dữ liệu các điểm tiêu chí và điểm SDG Index của năm 2024 của các nước), Raw Data - Panel (Dữ liệu các điểm tiêu chí của các nước từ năm 2000 đến 2024), Backdated SDG Index (Dữ liệu các điểm tiêu chí đã chuẩn hóa và điểm SDG Index của các nước từ năm 2000 đến 2024).

I.2. Phương pháp tiền xử lý

I.2.1. Hợp nhất dữ liệu

Kết bảng dữ liệu Backdated SDG Index và bảng dữ liệu Raw Data - Panel theo tên quốc gia và năm. Loại bỏ các cột không cần thiết.

```
backdated_data = data['Backdated SDG Index']
fulldata = data['Full Database']
codebook = data['Codebook']
```

```
backdated_data = backdated_data[['id', 'Country', 'year', 'SDG Index Score']]

dataset = pd.merge(
    raw_data,
    backdated_data,
    on=['Country', 'year'],
    how='left'
)

dataset = dataset.drop(columns=['id_x', 'id_y', 'indexreg'])

print(dataset.head())
```

Hình 4. 1 Hợp nhất dữ liệu (1)

Điền dữ liệu cho năm 2024: Gán giá trị từ bảng Full Database vào dataset, dựa trên mã chỉ số (IndCode), tên chỉ số (Indicator), và quốc gia (Country), cho năm 2024.

Quy trình:

- Duyệt từng cặp chỉ số và mã.
- Lọc dữ liệu theo năm 2024.
- Gán giá trị cho đúng vị trí trong dataset theo thuộc tính Country và year.

- Hợp nhất giá trị của cột "2024 SDG Index Score" của bảng Full Database với "SDG Index Score" trong dataset vào năm 2024 theo từng quốc gia.

```
# Lặp qua từng dòng của codebook để biết ánh xạ giữa tên chỉ số và mã cột
for _, row in codebook.iterrows():
    indicator_name = row['Indicator']
    ind_code = row['IndCode']

    if indicator_name in fulldata.columns and ind_code in dataset.columns:
        # Lấy cột Country và Indicator từ df_fulldata, lọc theo Year = 2024
        temp_df = fulldata[[indicator_name, f'Year: {ind_code}', 'Country']]
        temp_df = temp_df[temp_df[f'Year: {ind_code}'] == 2024]

        # Với mỗi Country phù hợp, gán giá trị vào df_dataset
        for _, temp_row in temp_df.iterrows():
            country = temp_row['Country']
            value = temp_row[indicator_name]

            # Xác định hàng trong df_dataset tương ứng với Country và year = 2024
            mask = (dataset['Country'] == country) & (dataset['year'] == 2024)
            dataset.loc[mask, ind_code] = value

# Cuối cùng, gán SDG Index Score
if '2024 SDG Index Score' in fulldata.columns and 'SDG Index Score' in dataset.columns:
    for _, row in fulldata.iterrows():
        country = row['Country']
        score = row['2024 SDG Index Score']
        dataset.loc[(dataset['Country'] == country) & (dataset['year'] == 2024), 'SDG Index Score'] = score
```

Hình 4. 2 Hợp nhất dữ liệu (2)

I.2.2. Điền dữ liệu thiếu dựa trên phương pháp bộ dữ liệu đã đưa ra trong Codebook

- sdg2_undernsh: Sử dụng bộ dữ liệu phân loại các quốc gia theo thu nhập[3], điền giá trị 2,5% cho các nước high-income.

```
# Lặp qua từng dòng của high_income và gán giá trị
for _, row in high_income.iterrows():
    country = row['Entity']
    year = row['Year']

    mask = (dataset['Country'] == country) & (dataset['year'] == year)
    dataset.loc[mask, 'sdg2_stunting'] = 2.58
```

Hình 4. 3 Điền dữ liệu bị thiếu (1)

- sdg2_wasting: Sử dụng bộ dữ liệu phân loại các quốc gia theo thu nhập, điền giá trị 0.75% cho các nước high-income.

```
# Lặp qua từng dòng của high_income và gán giá trị
for _, row in high_income.iterrows():
    country = row['Entity']
    year = row['Year']

    mask = (dataset['Country'] == country) & (dataset['year'] == year)
    dataset.loc[mask, 'sdg2_wasting'] = 0.75
```

Hình 4. 4 Điền dữ liệu bị thiếu (2)

- sdg5_familypl: Sử dụng bộ dữ liệu kế hoạch hóa gia đình[4] để điền vào giá trị thiếu cho dữ liệu (Đồng nhất tên quốc gia của 2 bộ dữ liệu, chỉ giữ dữ liệu với thuộc tính Variant = 'Median')

```
country_name_mapping = {
    'Bahamas, The': 'Bahamas',
    'Bolivia': 'Bolivia (Plurinational State of)',
    'Brunei Darussalam': 'Brunei Darussalam',
    'Congo, Dem. Rep.': 'Dem. Rep. of the Congo',
    'Congo, Rep.': 'Congo',
    "Cote d'Ivoire": "Côte d'Ivoire",
    'Egypt, Arab Rep.': 'Egypt',
    'Gambia, The': 'Gambia',
    'Iran, Islamic Rep.': 'Iran (Islamic Republic of)',
    'Korea, Dem. Rep.': "Dem. People's Rep. of Korea",
    'Korea, Rep.': 'Republic of Korea',
    'Kyrgyz Republic': 'Kyrgyzstan',
    'Lao PDR': "Lao People's Dem. Republic",
    'Micronesia, Fed. Sts.': 'Micronesia',
    'Moldova': 'Republic of Moldova',
    'Slovak Republic': 'Slovakia',
    'St. Kitts and Nevis': 'Saint Kitts and Nevis',
    'St. Lucia': 'Saint Lucia',
    'St. Vincent and the Grenadines': 'Saint Vincent and the Grenadines',
    'Tanzania': 'United Republic of Tanzania',
    'United States': 'United States of America',
    'Venezuela, RB': 'Venezuela (Bolivarian Republic of)',
    'Vietnam': 'Viet Nam',
    'Yemen, Rep.': 'Yemen',
}
```

```
# Lọc familypl chỉ giữ Variant = 'Median'
familypl_median = familypl[familypl['Variant'] == 'Median']
familypl_median['Location'] = familypl_median['Location'].replace(country_name_mapping)
```

Hình 4. 5 Điền dữ liệu bị thiếu (3)

```
# Lặp qua từng dòng và gán giá trị
for _, row in familypl_median.iterrows():
    country = row['Location']
    year = row['Time']
    value = row['Value']

    mask = (dataset['Country'] == country) & (dataset['year'] == year)
    dataset.loc[mask, 'sdg5_familypl'] = value
```

Hình 4. 6 Điền dữ liệu bị thiếu (4)

- sdg6_safewat: Điền giá trị thiếu của Australia bằng dữ liệu cùng năm của New Zealand.

```
# Lặp theo từng năm có trong df_dataset
years = dataset['year'].unique()

for y in years:
    # Lấy giá trị của New Zealand năm y
    value_nz = dataset.loc[(dataset['Country'] == 'New Zealand') & (dataset['year'] == y), 'sdg6_safewat']

    if not value_nz.empty:
        value = value_nz.values[0]
        # Gán giá trị đó cho Australia cùng năm
        dataset.loc[(dataset['Country'] == 'Australia') & (dataset['year'] == y), 'sdg6_safewat'] = value
```

Hình 4. 7 Điền dữ liệu bị thiếu (5)

- sdg9_uni: Ước lượng giá trị từ chỉ số về điểm số đại học trong bộ dữ liệu THE World University Rankings 2011-2024[5] cho các quốc gia thiếu dữ liệu. Đối với các quốc gia không có trường đại học nào trong bảng xếp hạng, giả định giá trị là 0.

```
# Hàm tính trung bình top 1–3 tuỳ theo số lượng trường
def avg_top_3_or_less(group):
    scores = group.sort_values("score", ascending=False)["score"].tolist()
    return sum(scores[:3]) / len(scores[:3])  # Tự động xử lý khi có <3 trường

# Tính trung bình theo từng quốc gia và năm
avg_scores_by_country_year = (
    full_ranking.groupby(["country", "year"])
    .apply(avg_top_3_or_less)
    .reset_index(name="sdg9_uni")
)

# Tạo ánh xạ từ (country, year) → sdg9_uni
mapping = avg_scores_by_country_year.set_index(["country", "year"])["sdg9_uni"]

# Gán giá trị vào dataset tại các vị trí sdg9_uni bị thiếu
mask_nan = dataset["sdg9_uni"].isna()
index_tuples = list(zip(dataset.loc[mask_nan, "Country"], dataset.loc[mask_nan, "year"]))

dataset.loc[mask_nan, "sdg9_uni"] = [
    mapping.get((country, year), 0) for (country, year) in index_tuples
]
```

Hình 4. 8 Điền dữ liệu bị thiếu (6)

- sdg9_rdex: Sử dụng bộ dữ liệu phân loại các quốc gia theo thu nhập, điền giá trị 0 cho các nước low-income.

```
for _, row in low_income.iterrows():
    country = row['Entity']
    year = row['Year']

    mask = (dataset['Country'] == country) & (dataset['year'] == year)
    dataset.loc[mask, 'sdg9_rdex'] = 0
```

Hình 4. 9 Điền dữ liệu bị thiếu (7)

- sdg11_slums: Sử dụng bộ dữ liệu phân loại các quốc gia theo thu nhập, điền giá trị 0 cho các nước high-income.

```
for _, row in high_income.iterrows():
    country = row['Entity']
    year = row['Year']

    mask = (dataset['Country'] == country) & (dataset['year'] == year)
    dataset.loc[mask, 'sdg11_slums'] = 0
```

Hình 4. 10 Điền dữ liệu bị thiếu (8)

- sdg16_clabor: Sử dụng bộ dữ liệu phân loại các quốc gia theo thu nhập, điền giá trị 0 cho các nước high-income.

```
for _, row in high_income.iterrows():
    country = row['Entity']
    year = row['Year']

    mask = (dataset['Country'] == country) & (dataset['year'] == year)
    dataset.loc[mask, 'sdg16_clabor'] = 0
```

Hình 4. 11 Điền dữ liệu bị thiếu (9)

I.2.3. Sử dụng thuật toán nội suy và KMeans để điền dữ liệu còn thiếu

- Điền vào các giá trị bị thiếu trong dữ liệu số bằng cách nội suy tuyến tính theo từng quốc gia để giữ sự nhất quán theo từng vùng dữ liệu.

```
# Lọc các cột số (trừ cột 'Country' và 'year')
numeric_cols = dataset_after_drop_cols.select_dtypes(include=['number']).columns

# Lặp qua từng quốc gia trong dataset
for country in dataset_after_drop_cols['Country'].unique():
    # Lọc dữ liệu của từng quốc gia
    country_data = dataset_after_drop_cols[dataset_after_drop_cols['Country'] == country]

    # Áp dụng interpolation cho các cột số (trừ cột 'Country' và 'year')
    country_data[numeric_cols] = country_data[numeric_cols].interpolate(method='linear', axis=0)

    # Cập nhật lại vào dataset
    dataset_after_drop_cols.loc[dataset_after_drop_cols['Country'] == country, numeric_cols] = country_data[numeric_cols]
```

Hình 4. 12 Điền dữ liệu bị thiếu (10)

- Điền khuyết cho các cột số bằng cách kết hợp chuẩn hóa dữ liệu, clustering (KMeans), và sau đó tính trung bình theo cụm (cluster). Quy trình các bước:
    - Ghi lại các vị trí NaN ban đầu
    - Tạm thay NaN bằng -1 để xử lý dễ hơn
    - Chuẩn hóa dữ liệu
    - Gom cụm dữ liệu theo từng năm (4 cụm mỗi năm)
    - Tính trung bình mỗi cụm và điền lại vào vị trí bị thiếu

```python
from sklearn.preprocessing import MinMaxScaler
from sklearn.cluster import KMeans

# Bước 1: Lấy danh sách cột số, loại bỏ 'year' và 'SDG Index Score'
numeric_cols = dataset_after_drop_cols.select_dtypes(include=['number']).columns.drop(['year', 'SDG Index Score'])

# Clone dữ liệu từ dataset_after_drop_cols
df3 = dataset_after_drop_cols.copy()

# Ghi lại mask NaN ban đầu để biết cần điền chỗ nào
nan_mask = df3[numeric_cols].isna()

# Tạm thời điền NaN bằng -1
df3[numeric_cols] = df3[numeric_cols].fillna(-1)

# Chuẩn hóa dữ liệu (chỉ scale numeric_cols, không chuẩn hóa 'year' và 'SDG Index Score')
scaler = MinMaxScaler(feature_range=(0, 100))
scaled_data = scaler.fit_transform(df3[numeric_cols])

# Gán lại giá trị chuẩn hóa vào df3
df3[numeric_cols] = scaled_data

# Gán lại giá trị gốc cho 'year' và 'SDG Index Score' để tránh bị ảnh hưởng
df3['year'] = dataset_after_drop_cols['year']
df3['SDG Index Score'] = dataset_after_drop_cols['SDG Index Score']

# Bước 6: Clustering theo từng năm
df3['Cluster'] = -1
for year in df3['year'].unique():
    year_mask = df3['year'] == year
    data_year = df3.loc[year_mask, numeric_cols]

    # Đổi -1 tạm thời thành 0 để tránh lỗi KMeans
    data_year_kmeans = data_year.replace(-1, 0)

    try:
        kmeans = KMeans(n_clusters=4, random_state=42)
        clusters = kmeans.fit_predict(data_year_kmeans)
        df3.loc[year_mask, 'Cluster'] = clusters
    except Exception as e:
        print(f"Không clustering được cho năm {year}: {e}")
```

Hình 4. 13 Điền dữ liệu bị thiếu (11)

```python
# Điền khuyết lại từ Cluster
for col in numeric_cols:
    for cluster_id in df3['Cluster'].unique():
        if cluster_id == -1: continue
        mask = (df3['Cluster'] == cluster_id) & (df3[col] != -1)
        cluster_mean = df3.loc[mask, col].mean()

        # Điền lại chỗ từng NaN trong dữ liệu gốc (theo mask ban đầu)
        fill_mask = (df3['Cluster'] == cluster_id) & nan_mask[col]
        df3.loc[fill_mask, col] = cluster_mean
```

Hình 4. 14 Điền dữ liệu bị thiếu (12)

I.2.4. Xóa dữ liệu rỗng và cột tương quan kém

Những cột có số lượng dữ liệu thiếu lớn (hơn 60% dữ liệu là dữ liệu thiếu) và không có ảnh hưởng lớn đến điểm SDG Index của quốc gia (hệ số tương quan giữa hai cột lớn hơn 0.7 và bé hơn -0.7) sẽ bị loại bỏ khỏi bộ dữ liệu.

- Xác định các cột quan trọng với dữ liệu thiếu lớn:

```python
def identify_important_high_missing_cols(df, target_column, missing_threshold=0.6, correlation_threshold=0.7):

    # 1. Calculate missing value percentages
    missing_percent = df.isnull().mean()

    # 2. Calculate correlations with target (only for numeric columns)
    numeric_cols = df.select_dtypes(include=['number']).columns
    correlations = df[numeric_cols].corr()[target_column].abs()

    # 3. Identify important columns with high missing values
    high_missing_cols = missing_percent[missing_percent > missing_threshold].index
    important_cols = correlations[correlations > correlation_threshold].index

    # Intersection of the two sets
    important_high_missing = list(set(high_missing_cols) & set(important_cols))

    return important_high_missing
```

Hình 4. 15 Xóa dữ liệu (1)

- Loại bỏ các cột thiếu nhiều dữ liệu và không nằm trong danh sách những cột quan trọng:

```python
def drop_non_important_high_missing(df, important_cols, missing_threshold=0.6):

    # Calculate missing percentage for each column
    missing_percent = df.isnull().mean()

    # Columns with missing > threshold
    high_missing_cols = missing_percent[missing_percent > missing_threshold].index.tolist()

    # Exclude important columns from the drop list
    cols_to_drop = [col for col in high_missing_cols if col not in important_cols]

    # Drop the columns
    df_dropped = df.drop(columns=cols_to_drop)

    print("Drop columns: " + str(cols_to_drop))
    # Return cleaned dataframe and list of dropped columns
    return df_dropped
```

Hình 4. 16 Xóa dữ liệu (2)

- Loại bỏ những dòng bị khuyết dữ liệu SDG Index Score:

```python
# Drop những dòng SDG Index Score bị khuyết
dataset = dataset.dropna(subset=['SDG Index Score'])
```

Hình 4. 17 Xóa dữ liệu (3)

- Loại bỏ những cột có hệ số correlation trong đoạn -0.4 đến 0.4:

```python
# Loại bỏ cột 'Country' và 'year' trước khi tính toán correlation
dataset_without_country_year = dataset.drop(columns=['Country', 'year'])

# Tính toán correlation giữa tất cả các cột còn lại và 'SDG Index Score'
correlation_with_sdg = dataset_without_country_year.corr()

# Lọc ra hệ số tương quan của từng cột với 'SDG Index Score'
correlation_with_sdg_index_score = correlation_with_sdg['SDG Index Score']
```

Hình 4. 18 Xóa dữ liệu (4)

```
# Lọc ra hệ số tương quan của từng cột với 'SDG Index Score'
correlation_with_sdg_index_score = correlation_with_sdg['SDG Index Score'].abs()

# Lọc các cột có correlation >= 0.5
columns_to_keep = correlation_with_sdg_index_score[correlation_with_sdg_index_score >= 0.4].index

# Giữ lại các cột có correlation >= 0.4 và tạo dataset mới
dataset_after_drop_cols = dataset[['Country', 'year'] + list(columns_to_keep)]
```

Hình 4. 19 Xóa dữ liệu (5)

I.3. Tổng quan bộ dữ liệu sau khi tiền xử lý

Bộ dữ liệu sau khi tiền xử lý gồm 4175 dòng và 59 cột với tổng số dữ liệu thiếu bằng 0. Thuộc tính có hệ số tương quan lớn nhất với SDG Index Score là sdg3_uhc (0.88).

II. Khám phá dữ liệu (EDA)

II.1. Phương pháp phân tích khám phá dữ liệu

II.1.1. Tải và kiểm tra dữ liệu

- Phương pháp:
- Tải dữ liệu bằng pandas.read_excel.
- Kiểm tra thông tin bằng df.info() để xác định kiểu dữ liệu và giá trị thiếu.
- Tính thống kê mô tả bằng df.describe() (trung bình, độ lệch chuẩn, min, max).
- Trực quan hóa phân bố bằng histogram (biến số) và biểu đồ cột (biến phân loại: Country, year).
- Kết quả:
- Không có giá trị thiếu (4175 non-null cho tất cả cột).
- Phạm vi giá trị lớn ở một số cột (ví dụ: sdg1_lmicpov: 0.003–99.572), cho thấy cần xử lý ngoại lai.
- SDG Index Score được xác định là biến mục tiêu tiềm năng.

Theo Montgomery & Runger (2014), thống kê mô tả và trực quan hóa là công cụ cơ bản trong EDA để hiểu phân bố dữ liệu và phát hiện các vấn đề như giá trị bất thường. Histogram và biểu đồ cột giúp nhận diện các mẫu phân bố (bimodal, skewed) và sự khác biệt giữa các nhóm (ví dụ: các quốc gia).
Tài liệu báo cáo rằng việc kiểm tra dữ liệu bằng df.info() và df.describe() đã xác định không có giá trị thiếu, nhưng các cột như sdg1_lmicpov có phạm vi lớn, đòi hỏi xử lý ngoại lai. Điều này phù hợp với lý thuyết về việc phát hiện vấn đề dữ liệu trong EDA.

II.1.2. Xử lí giá trị ngoại lai

- Phương pháp:
- Phát hiện ngoại lai bằng phương pháp IQR:
    - Tính Q1, Q3, và IQR cho các cột số.
    - Xác định ngưỡng: [Q1 - 1.5*IQR, Q3 + 1.5*IQR].

- Xử lý ngoại lai:
    - Giới hạn (clip) giá trị trong ngưỡng.
    - Thay Inf bằng NaN và điền NaN bằng trung bình cột.
- Ý nghĩa:
- Tăng độ ổn định cho các mô hình học máy.
- Ngăn chặn giá trị bất thường làm sai lệch thống kê hoặc trực quan hóa.

Giá trị ngoại lai có thể làm sai lệch các phân tích thống kê và mô hình học máy, đặc biệt trong các thuật toán nhạy cảm với thang đo như hồi quy tuyến tính. Phương pháp IQR, được đề xuất bởi Tuckey (1977), là một cách mạnh mẽ để phát hiện và xử lý ngoại lai dựa trên phân bố dữ liệu.
Tài liệu mô tả việc sử dụng IQR để phát hiện ngoại lai trong các cột như sdg1_lmicpov và xử lý bằng cách giới hạn giá trị hoặc thay thế Inf bằng trung bình cột. Điều này giúp tăng độ ổn định cho các phân tích tiếp theo, như tính tương quan hoặc học máy.

II.1.3. Phân tích tương quan

- Phương pháp:
- Tính ma trận tương quan bằng df.corr() cho các cột số.
- Vẽ heatmap cho 10 cột có tương quan cao nhất với SDG Index Score (correlation_heatmap_initial.png).
- Phân tích tương quan giữa các nhóm SDG (ví dụ: SDG 3 và SDG 6).
- Kết quả:
- Tương quan mạnh (>0.7) giữa SDG Index Score và:
    - sdg3_u5mort (tỷ lệ tử vong trẻ em)
    - sdg6_sanita (tiếp cận vệ sinh)
    - sdg7_elecac (tiếp cận điện)
- Tương quan thấp (<0.2) ở một số cột, gợi ý cần kỹ thuật tạo đặc trưng.
- Tương quan cao giữa sdg3_u5mort và sdg6_sanita cho thấy cải thiện vệ sinh có thể giảm tử vong trẻ em.

Ma trận tương quan là công cụ quan trọng trong EDA để xác định mối quan hệ tuyến tính giữa các biến. Theo Cohen (1988), hệ số tương quan >0.7 biểu thị mối quan hệ mạnh, trong khi giá trị <0.2 cho thấy mối quan hệ yếu, thường cần các kỹ thuật như tạo đặc trưng tương tác để nắm bắt mối quan hệ phi tuyến.
Tài liệu báo cáo rằng heatmap cho thấy các cột như sdg3_u5mort và sdg6_sanita có tương quan mạnh với SDG Index Score, trong khi một số cột có tương quan thấp, dẫn đến việc áp dụng kỹ thuật Feature Engineering. Điều này phù hợp với lý thuyết về việc sử dụng tương quan để định hướng xử lý dữ liệu.

II.1.4. Kỹ thuật tạo đặc trưng (Feature Engineering)

- Tương tác Đặc trưng:

- Tạo các đặc trưng mới để nắm bắt mối quan hệ phi tuyến:
    - sdg6_water_elec: sdg6_safewat * sdg7_elecac (nước sạch và điện)
    - sdg3_uhc_sanitation: sdg3_uhc * sdg6_sanita (bảo hiểm y tế và vệ sinh)
    - sdg13_co2_cpta: sdg13_co2gcp * sdg15_cpta (phát thải CO2 và bảo tồn đất)
- Ý nghĩa:
- Nắm bắt mối quan hệ phi tuyến giữa các chỉ số SDG.
- Tăng thông tin cho mô hình học máy.
- Hỗ trợ chính sách liên ngành (ví dụ: kết hợp y tế và vệ sinh).

Feature Engineering là quá trình tạo ra các đặc trưng mới từ dữ liệu gốc để cải thiện hiệu suất mô hình học máy. Theo Guyon & Elisseeff (2003), các đặc trưng tương tác (interaction features) đặc biệt hữu ích khi các biến có mối quan hệ phi tuyến, giúp mô hình nắm bắt được các mẫu phức tạp hơn.
Tài liệu liệt kê các đặc trưng tương tác như sdg6_water_elec và sdg3_uhc_sanitation, nhấn mạnh ý nghĩa của chúng trong việc cải thiện sức khỏe và năng suất lao động. Điều này phù hợp với lý thuyết về việc sử dụng đặc trưng tương tác để tăng cường khả năng dự đoán của mô hình.

- Biến động Thời gian:
- Sắp xếp dữ liệu theo Country và year.
- Tính phần trăm thay đổi hàng năm (pct_change) cho các cột số.
- Hiển thị xu hướng cho Việt Nam.
- Ý nghĩa:
- Theo dõi tiến bộ SDG theo thời gian.
- Xác định chính sách hiệu quả.

Phân tích chuỗi thời gian, như tính phần trăm thay đổi hàng năm, giúp phát hiện xu hướng và mẫu theo thời gian. Theo Hyndman & Athanasopoulos (2018), các chỉ số như pct_change là công cụ hữu ích để đánh giá sự thay đổi tương đối trong dữ liệu thời gian.
Tài liệu báo cáo việc tính pct_change để theo dõi xu hướng SDG tại Việt Nam, chẳng hạn cải thiện tiếp cận nước sạch (sdg6_safewater). Điều này hỗ trợ phân tích thời gian và định hướng chính sách.

II.1.5. Lựa chọn đặc trưng (Feature Selection)

- Loại bỏ Đa cộng tuyến:
- Tính Variance Inflation Factor (VIF) cho các cột số.
- Loại bỏ cột có VIF > 10.
- Ý nghĩa:
- Giảm hiện tượng đa cộng tuyến, cải thiện độ ổn định mô hình.

- Tăng tốc tính toán.

Đa cộng tuyến xảy ra khi các biến độc lập có tương quan mạnh với nhau, làm giảm độ tin cậy của các mô hình như hồi quy. VIF là chỉ số phổ biến để phát hiện đa cộng tuyến, với ngưỡng VIF > 10 thường được sử dụng để loại bỏ biến (Kutner et al., 2005).

Tài liệu mô tả việc sử dụng VIF để loại bỏ các cột có đa cộng tuyến, giúp cải thiện độ ổn định và hiệu suất tính toán. Điều này phù hợp với lý thuyết về quản lý đa cộng tuyến trong phân tích dữ liệu.

- Lựa chọn Đặc trưng với L1 Regularization:
- Sử dụng Lasso (alpha=0.1) để chọn các đặc trưng có hệ số khác 0.
- Ý nghĩa:
- Chọn các đặc trưng quan trọng như sdg3_u5mort.
- Giảm chiều dữ liệu.

L1 Regularization (Lasso) là kỹ thuật lựa chọn đặc trưng mạnh mẽ, tự động đặt hệ số của các biến không quan trọng về 0, từ đó giảm số lượng đặc trưng và ngăn chặn overfitting (Tibshirani, 1996).
Tài liệu báo cáo rằng Lasso được sử dụng để chọn các đặc trưng như sdg3_u5mort, giúp tập trung vào các chỉ số quan trọng và giảm chiều dữ liệu. Điều này phù hợp với lý thuyết về lựa chọn đặc trưng.

II.1.6. Trích xuất đặc trưng (Feature Extraction)

- PCA theo Nhóm SDG:
- Thực hiện PCA cho từng nhóm SDG (ví dụ: SDG 1, SDG 2).
- Chuẩn hóa dữ liệu bằng StandardScaler.
- Giữ lại >80% phương sai giải thích.
- Ý nghĩa:
- Giảm chiều dữ liệu trong mỗi nhóm SDG.
- Hỗ trợ trực quan hóa và phân tích cấu trúc dữ liệu.

Phân tích Thành phần Chính (PCA) là kỹ thuật giảm chiều dữ liệu, chuyển đổi các biến ban đầu thành các thành phần chính giữ lại phần lớn phương sai của dữ liệu. Theo Jolliffe (2002), PCA đặc biệt hữu ích khi xử lý tập dữ liệu có nhiều biến tương quan.

Tài liệu mô tả việc áp dụng PCA cho từng nhóm SDG để giảm chiều dữ liệu, giữ lại hơn 80% phương sai. Điều này giúp đơn giản hóa dữ liệu mà vẫn duy trì thông tin chính, phù hợp với lý thuyết PCA.

- T-SNE Toàn bộ Dữ liệu:
- Chọn top 20 đặc trưng từ Random Forest.

- Áp dụng T-SNE (n_components=2, perplexity=30).
- Xử lý NaN/Inf.
- Ý nghĩa:
- Tạo không gian 2D để phát hiện cụm quốc gia có đặc điểm SDG tương đồng.
- Hỗ trợ trực quan hóa và phân cụm.

T-SNE (t-Distributed Stochastic Neighbor Embedding) là kỹ thuật giảm chiều phi tuyến, đặc biệt hiệu quả trong việc trực quan hóa dữ liệu cao chiều bằng cách bảo toàn cấu trúc cục bộ (Van der Maaten & Hinton, 2008). Nó thường được sử dụng để phát hiện cụm trong dữ liệu phức tạp.
Tài liệu báo cáo việc sử dụng T-SNE để tạo không gian 2D, giúp phát hiện các cụm quốc gia có đặc điểm SDG tương đồng. Việc sửa lỗi NameError và xử lý NaN/Inf cũng được thực hiện để đảm bảo tính chính xác, phù hợp với lý thuyết về T-SNE.

II.2. Insight rút ra sau quá trình EDA

Phân tích EDA trên tập dữ liệu SDG đã cung cấp cái nhìn sâu sắc về cấu trúc dữ liệu, phân bố các chỉ số, và mối quan hệ giữa các biến. Các phát hiện chính bao gồm:

- Không có giá trị thiếu, nhưng cần xử lý ngoại lai ở các cột như sdg1_lmicpov.
- Các tương quan chỉ số trong một số mục tiêu phát triển bình vững:
- Nghèo cùng cực và nghèo trung bình có liên hệ chặt chẽ, cho thấy các quốc gia có nhiều người sống dưới $2.15/ngày thường cũng có tỷ lệ cao người sống dưới $3.65/ngày (Mối tương quan mạnh giữa sdg1_wpc và sdg1_lmicpov).
- Mối quan hệ trung bình giữa sdg2_undernsh với nhiều chỉ số khác (sdg2_stunting, sdg2_obesity, sdg2_trophic, sdg2_crlyld, sdg2_snmi) cho thấy tình trạng thiếu ăn có liên hệ đáng kể đến sức khỏe và phát triển của trẻ em.
- Một số chỉ số có quan hệ mạnh với nhiều mục tiêu khác:
- Mạng lưới tương quan mạnh giữa: Xóa nghèo (SDG1), Sức khỏe (SDG3), Giáo dục (SDG4), Năng lượng (SDG7), Kinh tế (SDG8) cho thấy đầu tư vào giáo dục và năng lượng sạch tạo hiệu ứng tích cực lên sức khỏe và giảm nghèo.
- Xóa đói giảm nghèo là yếu tố cốt lõi ảnh hưởng đến nhiều khía cạnh phát triển khác, đặc biệt là sức khỏe, giáo dục và cơ hội kinh tế (SDG1 - Xóa đói giảm nghèo có tương quan mạnh với: SDG3, SDG4, SDG7, SDG8).
- Giáo dục là nền tảng cho sự phát triển toàn diện và bền vững, ảnh hưởng đến cả sức khỏe, kinh tế, và khả năng hợp tác quốc tế (SDG4 - Giáo dục chất lượng có tương quan mạnh với SDG1, SDG2 , SDG3, SDG7, SDG8, SDG17).
- Năng lượng sạch và ổn định là yếu tố thúc đẩy tăng trưởng kinh tế, sức khỏe và giáo dục (SDG7 - Năng lượng sạch và giá thành hợp lý có tương quan mạnh với: SDG1, SDG3, SDG4, SDG8, SDG17).

- Phát triển kinh tế đi sẽ kèm với cải thiện giáo dục, giảm nghèo và tăng cường hợp tác toàn cầu (SDG8 - Việc làm và tăng trưởng kinh tế có mối quan hệ mạnh với: SDG1, SDG2, SDG4, SDG7, SDG12, SDG17).
- Vai trò nổi bật của hợp tác toàn cầu (SDG17 - Quan hệ đối tác vì các mục tiêu): Đóng vai trò trung gian kết nối các mục tiêu, thúc đẩy phát triển đồng đều giữa các lĩnh vực - Có mối quan hệ mạnh hoặc trung bình với hầu hết các SDG cốt lõi như SDG1, SDG2, SDG4, SDG7, SDG8.
- Các chỉ số như sdg3_u5mort và sdg6_sanita có tương quan mạnh với SDG Index Score, là các mục tiêu ưu tiên cho chính sách.
- Kỹ thuật Feature Engineering và Feature Selection đã tăng cường chất lượng dữ liệu, hỗ trợ phân tích sâu hơn.

Các bước tiếp theo bao gồm xây dựng mô hình học máy để dự đoán SDG Index Score và áp dụng phân tích thời gian để đánh giá xu hướng dài hạn, đặc biệt tại Việt Nam.

II.3. Trực quan hóa

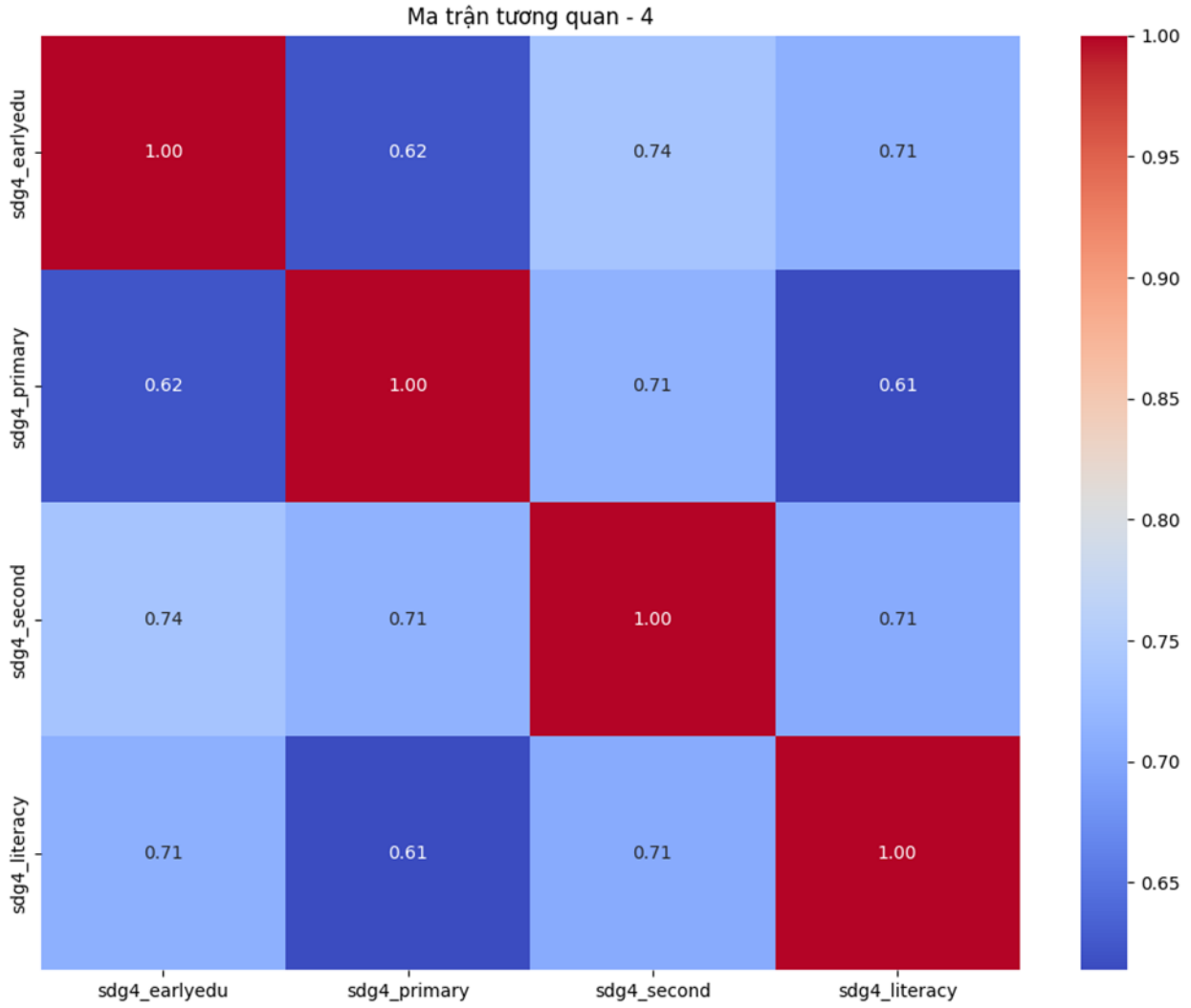
II.3.1. Heatmap 1

Mô tả:

- Heatmap giữa các chỉ số: sdg4_earlyedu (giáo dục mầm non), sdg4_primary (giáo dục tiểu học), sdg4_second (giáo dục trung học), sdg4_literacy (tỷ lệ biết đọc biết viết).

Kết quả:

- Tương quan mạnh dương giữa sdg4_earlyedu, sdg4_primary, sdg4_second, sdg4_literacy (0.61–0.74).
- **Insight**:
    - Giáo dục mầm non, tiểu học, và trung học có mối quan hệ chặt chẽ với tỷ lệ biết đọc biết viết, cho thấy hệ thống giáo dục phát triển đồng bộ sẽ cải thiện toàn diện trình độ học vấn.
    - Đầu tư vào giáo dục mầm non (sdg4_earlyedu) có thể tạo nền tảng cho các cấp học sau.



Hình 4. 20 Heatmap 1

II.3.2. Heatmap 2

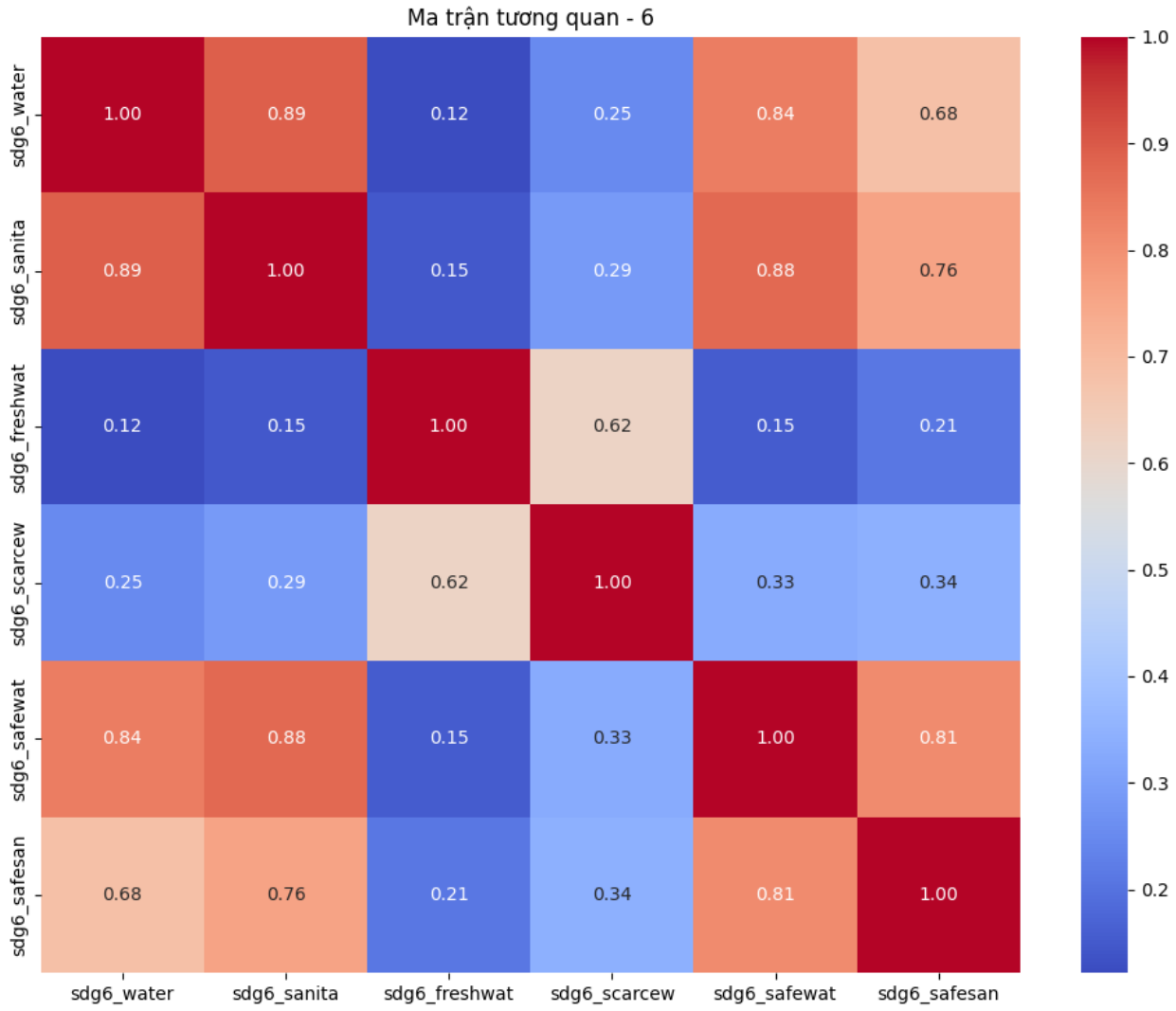
Mô tả:

- Heatmap giữa các chỉ số: sdg6_water (tiếp cận nước sạch), sdg6_sanita (tiếp cận vệ sinh), sdg6_freshwat (nước ngọt), sdg6_scarcew (khan hiếm nước), sdg6_safewat (nước sạch an toàn), sdg6_safesan (vệ sinh an toàn).

Kết quả:

- Tương quan mạnh dương giữa sdg6_water và sdg6_sanita (0.89), sdg6_safewat (0.84), sdg6_safesan (0.68).
- Tương quan yếu giữa sdg6_freshwat và các chỉ số khác (0.12–0.21).
- **Insight**:
    - Cải thiện tiếp cận nước sạch (sdg6_water) có thể thúc đẩy tiếp cận vệ sinh (sdg6_sanita) và nước sạch an toàn (sdg6_safewat).

- Khan hiếm nước (sdg6_scarcew) không có mối quan hệ mạnh với các chỉ số khác, có thể do dữ liệu không đồng đều.



Hình 4. 21 Heatmap 2

II.3.3. Heatmap 3

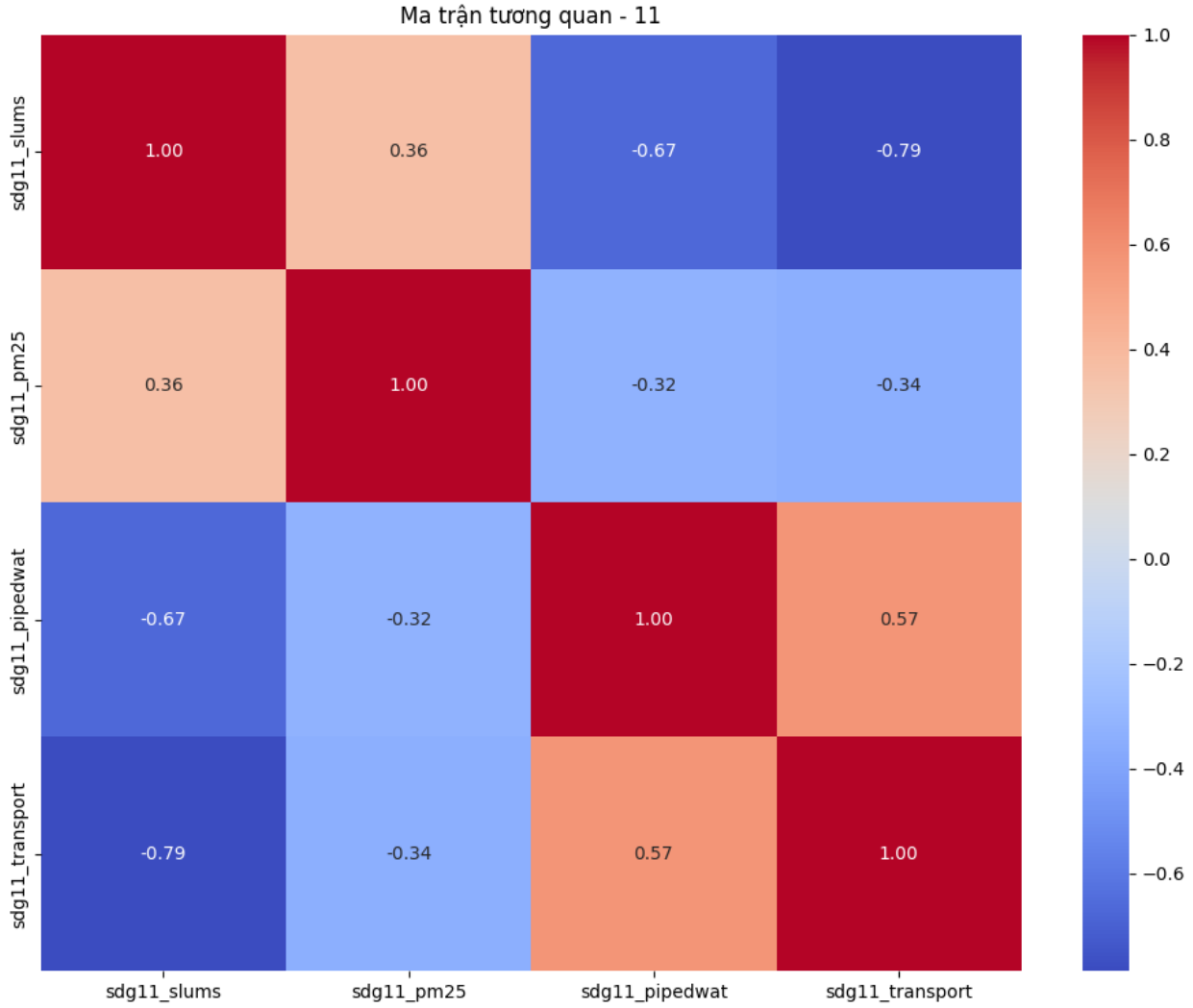
Mô tả:

- Heatmap giữa các chỉ số: sdg11_slums (khu ổ chuột), sdg11_pm25 (ô nhiễm không khí), sdg11_pipewat (nước sạch đô thị), sdg11_transport (giao thông công cộng).

Kết quả:

- Tương quan âm mạnh giữa sdg11_slums và sdg11_pipewat (-0.67), sdg11_transport (-0.79).

- Tương quan dương giữa sdg11_pipewat và sdg11_transport (0.57).
- **Insight**:
    - Các khu ổ chuột (sdg11_slums) thường thiếu tiếp cận nước sạch đô thị (sdg11_pipewat) và giao thông công cộng (sdg11_transport), cho thấy sự bất bình đẳng trong phát triển đô thị.
    - Cải thiện nước sạch đô thị và giao thông công cộng có thể hỗ trợ lẫn nhau.



Hình 4. 22 Heatmap 3

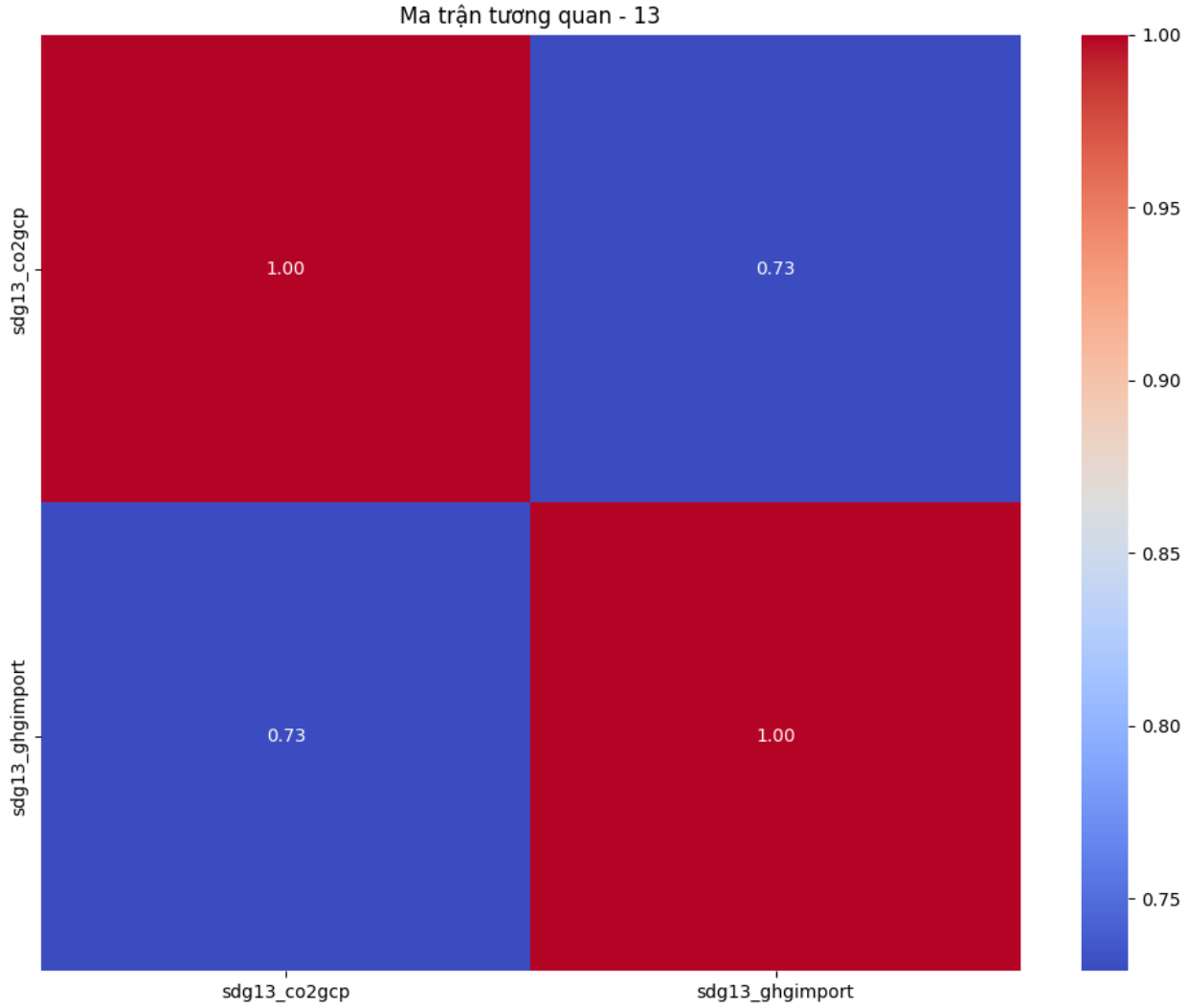
II.3.4. Heatmap 4

Mô tả:

- Heatmap giữa các chỉ số: sdg13_co2gdp (phát thải CO2 trên GDP), sdg13_ghgimport (khí nhà kính nhập khẩu).

Kết quả:

- Tương quan mạnh dương giữa sdg13_co2gdp và sdg13_ghgimport (0.73).
- **Insight**:
    - Các quốc gia có lượng phát thải CO2 trên GDP cao (sdg13_co2gdp) thường cũng nhập khẩu nhiều khí nhà kính (sdg13_ghgimport), cho thấy sự phụ thuộc vào các hoạt động kinh tế gây ô nhiễm.
    - Cần kết hợp giảm phát thải CO2 trong sản xuất nội địa và kiểm soát nhập khẩu để cải thiện hành động khí hậu.



Hình 4. 23 Heatmap 4

II.3.5. Heatmap 5

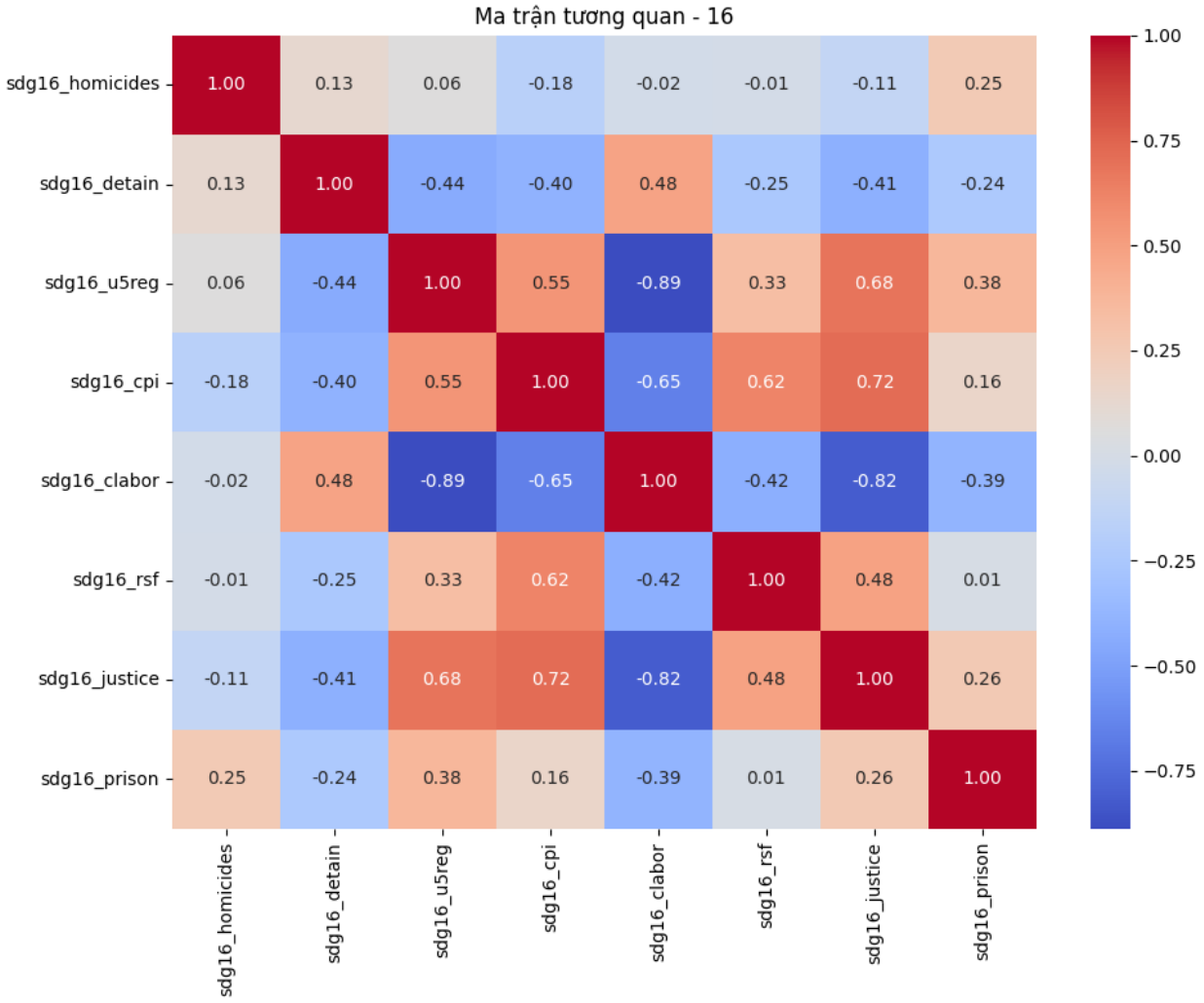
Mô tả:

- Heatmap giữa các chỉ số: sdg16_homicides (tỷ lệ giết người), sdg16_detain (giam giữ), sdg16_u5reg (đăng ký trẻ em dưới 5 tuổi), sdg16_cpi (chỉ số nhận

thức tham nhũng), sdg16_clabor (lao động trẻ em), sdg16_rsf (tự do báo chí), sdg16_justice (công lý), sdg16_prison (tỷ lệ tù nhân).

Kết quả:

- Tương quan mạnh âm giữa sdg16_cpi (chống tham nhũng) và sdg16_clabor (-0.65), sdg16_justice (-0.82).
- Tương quan dương mạnh giữa sdg16_u5reg và sdg16_justice (0.68), sdg16_cpi (0.55).
- Tương quan âm giữa sdg16_homicides và sdg16_cpi (-0.18), sdg16_justice (-0.11).
- **Insight**:
    - Các quốc gia có chỉ số chống tham nhũng cao (sdg16_cpi) thường có ít lao động trẻ em (sdg16_clabor) và công lý tốt hơn (sdg16_justice), cho thấy thể chế mạnh mẽ giúp cải thiện công bằng xã hội.
    - Đăng ký trẻ em dưới 5 tuổi (sdg16_u5reg) có liên quan tích cực đến công lý và chống tham nhũng, nhấn mạnh vai trò của quản lý dân số trong phát triển bền vững.
    - Tỷ lệ giết người (sdg16_homicides) có tương quan âm nhẹ với công lý và chống tham nhũng, cho thấy cần cải thiện an ninh để hỗ trợ các mục tiêu khác.



Hình 4. 24 Heatmap 5

II.3.6. Phân tích ma trận tương quan giữa các SDG

II.3.6.1. Các mối tương quan nổi bật

1. **SDG1 (Xóa đói nghèo)**:
    - Tương quan **dương mạnh** với SDG3 (Sức khỏe): **0.82** → Xóa đói nghèo gắn liền với cải thiện sức khỏe.
    - Tương quan **âm mạnh** với SDG4 (Giáo dục): **-0.84** → Có thể phản ánh sự đánh đổi giữa đầu tư giáo dục và giảm nghèo, hoặc dữ liệu không đồng đều ở các quốc gia.
2. **SDG4 (Giáo dục)**:
    - Tương quan **dương mạnh** với SDG7 (Năng lượng sạch): **0.88** và SDG8 (Kinh tế): **0.82** → Giáo dục chất lượng thúc đẩy phát triển năng lượng và tăng trưởng kinh tế.

- Tương quan **âm mạnh** với SDG1 (-0.84) và SDG3 (-0.81) → Cần nghiên cứu thêm để hiểu nguyên nhân (ví dụ: ưu tiên ngân sách).

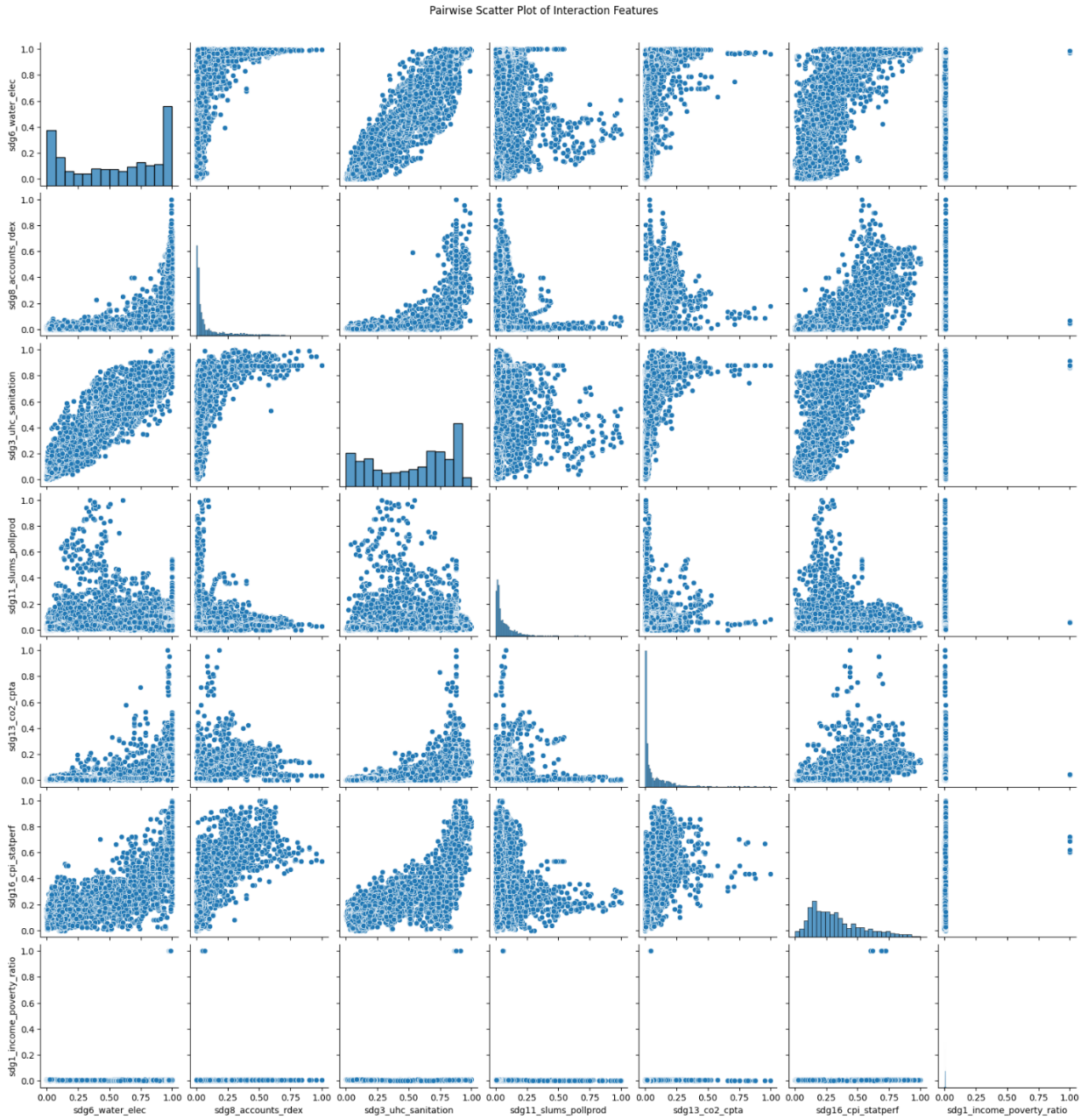
3. **SDG6 (Nước sạch)**:
   - Tương quan **dương mạnh** với SDG13 (Hành động khí hậu): **0.68** → Quản lý nước sạch liên quan đến ứng phó biến đổi khí hậu.
   - Tương quan yếu với các SDG khác (0.02–0.37) → Cần giải pháp chuyên biệt cho từng lĩnh vực.
4. **SDG16 (Công lý)**:
   - Tương quan **dương yếu** với SDG17 (Đối tác toàn cầu): **0.36** → Thể chế mạnh hỗ trợ hợp tác quốc tế.
   - Tương quan **âm** với nhiều SDG (ví dụ: SDG1: -0.40) → Cần cân bằng giữa an ninh và phát triển xã hội.

Insights chính:

- **Giáo dục là trụ cột**: SDG4 (Giáo dục) có mối liên hệ mạnh với SDG7 (Năng lượng) và SDG8 (Kinh tế), cho thấy đầu tư vào giáo dục có thể kéo theo phát triển đa ngành.
- **Mâu thuẫn tiềm ẩn**: Tương quan âm mạnh giữa SDG1 (Xóa đói) và SDG4 (Giáo dục) phản ánh thách thức trong phân bổ nguồn lực. Cần chính sách linh hoạt để tránh đánh đổi.
- **Nước sạch và khí hậu**: SDG6 và SDG13 có liên kết chặt chẽ, gợi ý tích hợp quản lý tài nguyên nước vào kế hoạch khí hậu.
- **Thể chế mạnh hỗ trợ SDG17**: SDG16 (Công lý) tuy có tương quan yếu nhưng là nền tảng cho hợp tác toàn cầu (SDG17).

II.3.6.2. Trực quan hóa đặc trưng mới

Dùng pairplot với histogram để khám phá mối quan hệ và phân phối của các đặc trưng tương tác.



Hình 4. 25 Trực quan hóa đặc trưng mới

Đặc trưng	Nhận xét về phân phối
sdg6_water_elec	Phân phối khá đều, có thiên hướng tăng về phía 1
sdg8_accounts_index	Phân phối lệch phải (nhiều điểm gần 1), cho thấy hầu hết các quốc gia có chỉ số tài khoản ngân hàng cao
sdg3_uhc_sanitation	Lệch phải mạnh, cho thấy đa số có mức độ tiếp cận vệ sinh kém
sdg11_slums_polipop	Phân phối gần như đối xứng, nhưng có nhiều điểm tập trung ở giữa
sdg13_co2_capita	Lệch phải rất mạnh (phân phối log-normal), cho thấy sự chênh lệch lớn về phát thải CO_2/người

sdg16_govt_statreport	Có vẻ như là đặc trưng rời rạc (nhị phân hoặc số lượng nhỏ giá trị phân biệt)
sdg1_income_poverty_ratio	Có vẻ có nhiều giá trị bằng 0, có thể là do thiếu dữ liệu hoặc đặc trưng đặc biệt

Bảng 4. 1 Phân phối từng đặc trưng

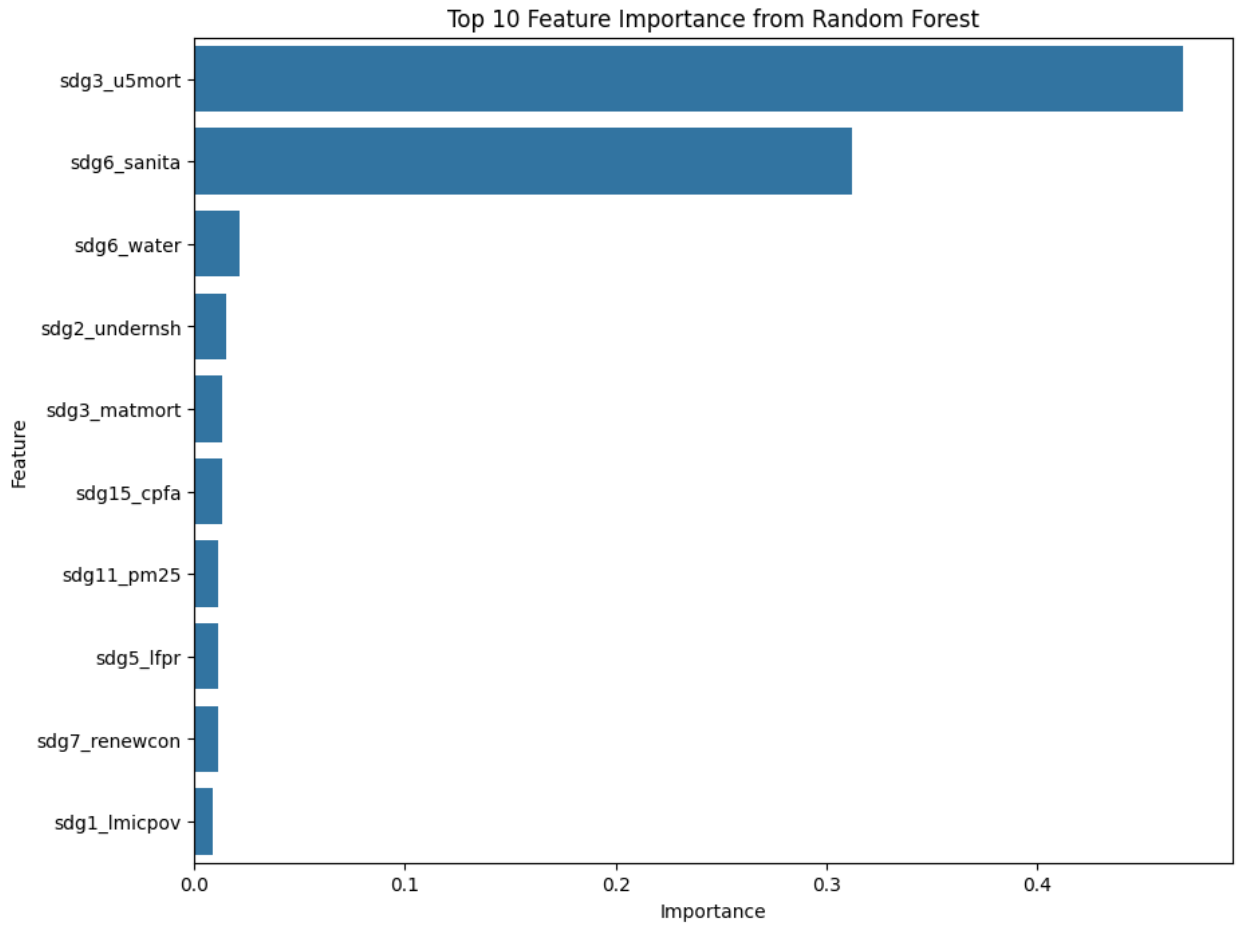
II.3.6.3. Mối tương quan giữa các đặc trưng (Scatter plots)

Mối quan hệ dương mạnh:

- sdg6_water_elec ↔ sdg8_accounts_index: mối tương quan dương rõ rệt (quốc gia có hạ tầng điện – nước tốt thì chỉ số tài chính cũng cao).
- sdg3_uhc_sanitation ↔ sdg6_water_elec: cũng tương quan dương, cho thấy nước sạch và vệ sinh có liên kết chặt.

Mối quan hệ yếu hoặc không rõ ràng:

- sdg13_co2_capita với các đặc trưng còn lại: thường không có xu hướng rõ (do biến này bị lệch mạnh).
- sdg16_govt_statreport không có mối tương quan rõ ràng với các đặc trưng khác – có thể là đặc trưng nhị phân.



Hình 4. 26 Những đặc trưng quan trọng

Các mối tương quan nổi bật:

1. **SDG1 (Xóa đói nghèo)**:
    - Tương quan **dương mạnh** với SDG3 (Sức khỏe): **0.82** → Xóa đói nghèo gắn liền với cải thiện sức khỏe.
    - Tương quan **âm mạnh** với SDG4 (Giáo dục): **-0.84** → Có thể phản ánh sự đánh đổi giữa đầu tư giáo dục và giảm nghèo, hoặc dữ liệu không đồng đều ở các quốc gia.
2. **SDG4 (Giáo dục)**:
    - Tương quan **dương mạnh** với SDG7 (Năng lượng sạch): **0.88** và SDG8 (Kinh tế): **0.82** → Giáo dục chất lượng thúc đẩy phát triển năng lượng và tăng trưởng kinh tế.
    - Tương quan **âm mạnh** với SDG1 (-0.84) và SDG3 (-0.81) → Cần nghiên cứu thêm để hiểu nguyên nhân (ví dụ: ưu tiên ngân sách).

3. **SDG6 (Nước sạch)**:
    - Tương quan **dương mạnh** với SDG13 (Hành động khí hậu): **0.68** → Quản lý nước sạch liên quan đến ứng phó biến đổi khí hậu.
    - Tương quan yếu với các SDG khác (0.02–0.37) → Cần giải pháp chuyên biệt cho từng lĩnh vực.
4. **SDG16 (Công lý)**:
    - Tương quan **dương yếu** với SDG17 (Đối tác toàn cầu): **0.36** → Thể chế mạnh hỗ trợ hợp tác quốc tế.
    - Tương quan **âm** với nhiều SDG (ví dụ: SDG1: -0.40) → Cần cân bằng giữa an ninh và phát triển xã hội.

Insights chính:

- **Giáo dục là trụ cột**: SDG4 (Giáo dục) có mối liên hệ mạnh với SDG7 (Năng lượng) và SDG8 (Kinh tế), cho thấy đầu tư vào giáo dục có thể kéo theo phát triển đa ngành.
- **Mâu thuẫn tiềm ẩn**: Tương quan âm mạnh giữa SDG1 (Xóa đói) và SDG4 (Giáo dục) phản ánh thách thức trong phân bổ nguồn lực. Cần chính sách linh hoạt để tránh đánh đổi.
- **Nước sạch và khí hậu**: SDG6 và SDG13 có liên kết chặt chẽ, gợi ý tích hợp quản lý tài nguyên nước vào kế hoạch khí hậu.
- **Thể chế mạnh hỗ trợ SDG17**: SDG16 (Công lý) tuy có tương quan yếu nhưng là nền tảng cho hợp tác toàn cầu (SDG17).

III. Phân tích SHAP

SHAP được sử dụng để làm rõ các đặc trưng (features) ảnh hưởng đến SDG Index Score, mức độ ảnh hưởng (tích cực hay tiêu cực), và mối quan hệ giữa các đặc trưng. Dưới đây là phân tích chi tiết dựa trên mã được cung cấp.

III.1. Tính toán với SHAP

- **Chia dữ liệu train/test theo năm**: Dữ liệu được chia thành hai tập
- **Tập huấn luyện (train_data):** Dữ liệu từ năm 2000 đến 2020.
- **Tập kiểm tra (test_data):** Dữ liệu từ năm 2021 đến 2024.

```
train_data_shap = df2[df2['year'] <= 2020].copy()
```

```
test_data_shap = df2[(df2['year'] >= 2021) & (df2['year'] <= 2024)].copy()


X_train_shap = train_data_shap.drop(['Country', 'year', 'SDG Index Score'], axis=1).copy()
y_train_shap = train_data_shap['SDG Index Score'].copy()
X_test_shap = test_data_shap.drop(['Country', 'year', 'SDG Index Score'], axis=1).copy()
y_test_shap = test_data_shap['SDG Index Score'].copy()
```

- **Mô hình học máy**

Mô hình XGBoost Regressor được sử dụng để dự đoán SDG Index Score, với các tham số:

- **Objective:** reg:squarederror (dự đoán giá trị liên tục).
- **Số cây (n_estimators):** 100.
- **Tỷ lệ học (learning_rate):** 0.1.
- **Random seed:** 42 (để tái tạo kết quả).

```
model = xgb.XGBRegressor(objective='reg:squarederror',
                         n_estimators=100,
                         learning_rate=0.1,
                         random_state=42)
model.fit(X_train_shap, y_train_shap)
```

- **Tính toán giá trị SHAP**

SHAP sử dụng TreeExplainer (tối ưu cho mô hình cây như XGBoost) để tính giá trị SHAP cho tập kiểm tra.

```
explainer = shap.TreeExplainer(model)
shap_values = explainer.shap_values(X_test_shap)
```

- **shap_values:** Một mảng numpy với kích thước (số mẫu test, số đặc trưng), biểu thị mức độ đóng góp của mỗi đặc trưng vào dự đoán cho từng mẫu.

- **explainer.expected_value:** Giá trị kỳ vọng (dự đoán trung bình của mô hình trên toàn tập dữ liệu).

III.2. Summary Plot

Summary Plot cung cấp cái nhìn tổng quan về mức độ quan trọng và hướng ảnh hưởng của các đặc trưng trên toàn bộ tập kiểm tra.

```
shap.summary_plot(shap_values, X_test_shap)
```

Phân tích Summary Plot

- **Trục x**: Giá trị SHAP (mức độ ảnh hưởng đến dự đoán). Giá trị dương làm tăng SDG Index Score, giá trị âm làm giảm.
- **Trục y**: Các đặc trưng, xếp theo mức độ quan trọng (đặc trưng có ảnh hưởng lớn nhất nằm ở trên cùng).
- **Màu sắc**:
    - **Đỏ**: Giá trị đặc trưng cao (ví dụ: tỷ lệ nghèo đói cao).
    - **Xanh dương**: Giá trị đặc trưng thấp (ví dụ: tỷ lệ nghèo đói thấp).

Kết quả từ biểu đồ

Dựa trên Summary Plot được cung cấp, dưới đây là **20 đặc trưng có ảnh hưởng lớn nhất** (xếp theo thứ tự giảm dần về mức độ quan trọng):

1. **sdg3_u5mort** (Tỷ lệ tử vong trẻ dưới 5 tuổi): Có tác động tiêu cực lớn (SHAP âm, dao động từ -8 đến 0). Giá trị cao (đỏ) làm giảm mạnh SDG Index Score.
2. **sdg6_sanita** (Tiếp cận vệ sinh): Có tác động tích cực (SHAP dương, từ 0 đến 8). Giá trị cao (đỏ) làm tăng chỉ số.
3. **sdg1_lmicpov** (Tỷ lệ nghèo đói ở quốc gia thu nhập thấp/trung bình): Tác động tiêu cực (SHAP âm, từ -6 đến 0). Giá trị cao (đỏ) làm giảm chỉ số.
4. **sdg11_pm25** (Ô nhiễm không khí PM2.5): Tác động tiêu cực (SHAP âm, từ -6 đến 0). Giá trị cao (đỏ) làm giảm chỉ số.
5. **sdg17_govex** (Chi tiêu chính phủ): Tác động tích cực (SHAP dương, từ 0 đến 6). Giá trị cao (đỏ) làm tăng chỉ số.
6. **sdg1_wpc** (Tỷ lệ nghèo đói): Tác động tiêu cực (SHAP âm, từ -5 đến 0). Giá trị cao (đỏ) làm giảm chỉ số.
7. **sdg7_elecac** (Tiếp cận điện): Tác động tích cực (SHAP dương, từ 0 đến 5). Giá trị cao (đỏ) làm tăng chỉ số.

8. **sdg14_cpma** (Khu vực bảo tồn biển): Tác động tích cực (SHAP dương, từ 0 đến 4). Giá trị cao (đỏ) làm tăng chỉ số.
9. **sdg4_second** (Giáo dục trung học): Tác động tích cực (SHAP dương, từ 0 đến 4). Giá trị cao (đỏ) làm tăng chỉ số.
10. **sdg3_matmort** (Tỷ lệ tử vong mẹ): Tác động tiêu cực (SHAP âm, từ -4 đến 0). Giá trị cao (đỏ) làm giảm chỉ số.
11. **sdg3_uhc** (Bảo hiểm y tế toàn dân): Tác động tích cực (SHAP dương, từ 0 đến 4). Giá trị cao (đỏ) làm tăng chỉ số.
12. **sdg11_slums** (Khu ổ chuột): Tác động tiêu cực (SHAP âm, từ -4 đến 0). Giá trị cao (đỏ) làm giảm chỉ số.
13. **sdg9_articles** (Bài báo khoa học): Tác động tích cực (SHAP dương, từ 0 đến 3). Giá trị cao (đỏ) làm tăng chỉ số.
14. **sdg4_earlyedu** (Giáo dục sớm): Tác động tích cực (SHAP dương, từ 0 đến 3). Giá trị cao (đỏ) làm tăng chỉ số.
15. **sdg2_stunting** (Chậm phát triển trẻ em): Tác động tiêu cực (SHAP âm, từ -3 đến 0). Giá trị cao (đỏ) làm giảm chỉ số.
16. **sdg2_cryld** (Năng suất ngũ cốc): Tác động tích cực (SHAP dương, từ 0 đến 3). Giá trị cao (đỏ) làm tăng chỉ số.
17. **sdg12_pollimp** (Ô nhiễm nhập khẩu): Tác động tiêu cực (SHAP âm, từ -3 đến 0). Giá trị cao (đỏ) làm giảm chỉ số.
18. **sdg2_undernsh** (Suy dinh dưỡng): Tác động tiêu cực (SHAP âm, từ -3 đến 0). Giá trị cao (đỏ) làm giảm chỉ số.
19. **sdg9_rdex** (Chi tiêu R&D): Tác động tích cực (SHAP dương, từ 0 đến 2). Giá trị cao (đỏ) làm tăng chỉ số.
20. **sdg6_water** (Tiếp cận nước sạch): Tác động tích cực (SHAP dương, từ 0 đến 2). Giá trị cao (đỏ) làm tăng chỉ số.

Điểm chung của các đặc trưng có ảnh hưởng cao

- **Sức khỏe (SDG 3)**: Các chỉ số như sdg3_u5mort (tử vong trẻ dưới 5 tuổi), sdg3_matmort (tử vong mẹ), và sdg3_uhc (bảo hiểm y tế) có ảnh hưởng lớn. Đặc biệt, tử vong trẻ em và mẹ có tác động tiêu cực mạnh khi giá trị cao, trong khi bảo hiểm y tế toàn dân có tác động tích cực.

- **Nghèo đói và dinh dưỡng (SDG 1, SDG 2)**: sdg1_wpc, sdg1_lmicpov, sdg2_undernsh, và sdg2_stunting đều có tác động tiêu cực lớn khi giá trị cao, phản ánh các vấn đề kinh tế và an ninh lương thực.
- **Cơ sở hạ tầng (SDG 6, SDG 7)**: sdg6_sanita, sdg6_water, và sdg7_elecac có tác động tích cực, cho thấy tầm quan trọng của vệ sinh, nước sạch, và tiếp cận điện trong việc cải thiện SDG Index Score.
- **Môi trường và quản trị (SDG 11, SDG 17)**: Các chỉ số như sdg11_pm25 (ô nhiễm không khí), sdg11_slums (khu ổ chuột) có tác động tiêu cực, trong khi sdg17_govex (chi tiêu chính phủ) có tác động tích cực.
- **Giáo dục và đổi mới (SDG 4, SDG 9)**: sdg4_second, sdg4_earlyedu, và sdg9_articles có tác động tích cực, cho thấy giáo dục và nghiên cứu khoa học là động lực quan trọng.

Thách thức đối với các đặc trưng có ảnh hưởng cao

- **Sức khỏe (SDG 3)**:
    - Giảm tỷ lệ tử vong trẻ em (sdg3_u5mort) và mẹ (sdg3_matmort) đòi hỏi đầu tư lớn vào hệ thống y tế, đặc biệt ở các khu vực nông thôn và vùng sâu, nơi tiếp cận dịch vụ y tế còn hạn chế.
    - Cải thiện bảo hiểm y tế toàn dân (sdg3_uhc) cần mở rộng phạm vi phủ sóng và nâng cao chất lượng dịch vụ y tế.
- **Nghèo đói (SDG 1)**:
    - Giảm tỷ lệ nghèo đói (sdg1_wpc, sdg1_lmicpov) đòi hỏi tăng trưởng kinh tế bền vững, tạo việc làm ổn định, và các chương trình an sinh xã hội hiệu quả, đặc biệt ở vùng dân tộc thiểu số và nông thôn.
- **Dinh dưỡng (SDG 2)**:
    - Giảm suy dinh dưỡng (sdg2_undernsh) và chậm phát triển (sdg2_stunting) cần cải thiện chuỗi cung ứng thực phẩm, giáo dục dinh dưỡng, và hỗ trợ nông nghiệp bền vững.
- **Môi trường (SDG 11, SDG 12)**:
    - Giảm ô nhiễm không khí (sdg11_pm25) và ô nhiễm nhập khẩu (sdg12_pollimp) đòi hỏi các chính sách kiểm soát khí thải, chuyển đổi năng lượng xanh, và quản lý chất thải hiệu quả.
    - Giảm khu ổ chuột (sdg11_slums) cần đầu tư vào nhà ở xã hội và quy hoạch đô thị.

- **Cơ sở hạ tầng (SDG 6, SDG 7)**:
    - Duy trì và mở rộng tiếp cận nước sạch (sdg6_water), vệ sinh (sdg6_sanita), và điện (sdg7_elecac) ở các khu vực khó khăn là thách thức lớn, đặc biệt trong bối cảnh biến đổi khí hậu.

Tốc độ phát triển các chỉ số trong thực tế (tập trung vào Việt Nam)

- **Sức khỏe (SDG 3)**:
    - **Tử vong trẻ dưới 5 tuổi (sdg3_u5mort)**: Việt Nam đã giảm tỷ lệ này từ khoảng 36/1.000 ca sinh năm 2000 xuống còn 21/1.000 ca sinh năm 2020, nhưng vẫn chưa đạt mục tiêu SDG (dưới 12/1.000). Tốc độ cải thiện chậm ở vùng sâu, vùng xa.
    - **Tử vong mẹ (sdg3_matmort)**: Giảm từ 69/100.000 ca sinh năm 2000 xuống 43/100.000 năm 2020, nhưng mục tiêu SDG là dưới 10/100.000, đòi hỏi nỗ lực lớn hơn.
    - **Bảo hiểm y tế (sdg3_uhc)**: Tỷ lệ bao phủ bảo hiểm y tế đạt khoảng 90% vào năm 2023, một thành tựu lớn, nhưng chất lượng dịch vụ ở vùng sâu vẫn cần cải thiện.
- **Nghèo đói (SDG 1)**:
    - **Tỷ lệ nghèo đói (sdg1_wpc)**: Giảm mạnh từ hơn 30% năm 2000 xuống dưới 10% năm 2020, nhưng ở vùng dân tộc thiểu số và nông thôn, tỷ lệ vẫn cao (20–30% ở một số khu vực), đòi hỏi các chương trình hỗ trợ mục tiêu hơn.
- **Dinh dưỡng (SDG 2)**:
    - **Suy dinh dưỡng (sdg2_undernsh)**: Tỷ lệ suy dinh dưỡng trẻ em dưới 5 tuổi giảm từ 36% năm 2000 xuống khoảng 19% năm 2020, nhưng vẫn chưa đạt mục tiêu SDG (dưới 10%). Tốc độ cải thiện chậm ở vùng sâu.
    - **Chậm phát triển (sdg2_stunting)**: Tương tự, tỷ lệ chậm phát triển giảm xuống khoảng 20%, nhưng cần tăng tốc để đạt mục tiêu.
- **Cơ sở hạ tầng (SDG 6, SDG 7)**:
    - **Tiếp cận vệ sinh (sdg6_sanita)**: Đạt khoảng 85% dân số vào năm 2023, nhưng vẫn cần cải thiện ở vùng nông thôn.
    - **Tiếp cận nước sạch (sdg6_water)**: Đạt hơn 95%, một trong những thành tựu lớn.

    - **Tiếp cận điện (sdg7_elecac)**: Gần 100% dân số có điện vào năm 2023, một thành công đáng kể.

- **Môi trường (SDG 11, SDG 12)**:
    - **Ô nhiễm không khí (sdg11_pm25)**: Việt Nam vẫn đối mặt với ô nhiễm không khí nghiêm trọng, đặc biệt ở các thành phố lớn như Hà Nội và TP.HCM, với chỉ số PM2.5 thường vượt ngưỡng an toàn (trên 25 $\mu g/m^3$).
    - **Khu ổ chuột (sdg11_slums)**: Tỷ lệ dân số sống trong khu ổ chuột giảm, nhưng vẫn tồn tại ở các đô thị lớn.
- **Quản trị (SDG 17)**:
    - **Chi tiêu chính phủ (sdg17_govex)**: Việt Nam đã tăng chi tiêu cho các chương trình phát triển bền vững, nhưng cần phân bổ hiệu quả hơn cho các khu vực khó khăn.

Nhận xét

- **Đặc trưng tiêu cực (sức khỏe, nghèo đói, dinh dưỡng, môi trường)**: Các chỉ số như sdg3_u5mort, sdg1_wpc, sdg2_undernsh, và sdg11_pm25 là những vấn đề cốt lõi của SDGs, đòi hỏi đầu tư dài hạn và phối hợp đa ngành. Tốc độ cải thiện của các chỉ số này còn chậm, đặc biệt ở vùng sâu, vùng xa và các khu vực dân tộc thiểu số.
- **Đặc trưng tích cực (cơ sở hạ tầng, quản trị, giáo dục)**: Các chỉ số như sdg6_sanita, sdg6_water, sdg7_elecac, sdg17_govex, và sdg4_second cho thấy Việt Nam đã đạt được tiến bộ đáng kể trong việc xây dựng cơ sở hạ tầng và cải thiện quản trị, giáo dục. Đây là những động lực chính để nâng cao SDG Index Score.
- **Thách thức lớn nhất**: Sự không đồng đều giữa các vùng miền và nhóm dân cư (thành thị vs. nông thôn, người Kinh vs. dân tộc thiểu số) là rào cản lớn nhất. Các vấn đề như ô nhiễm không khí và suy dinh dưỡng cần các giải pháp cấp bách để đạt mục tiêu SDG.

III.3. Force Plot

Force Plot giải thích dự đoán SDG Index Score cho Việt Nam năm 2024, cho thấy mức độ ảnh hưởng của từng đặc trưng.

```
sample_idx = len(X_test_shap) - 1
```

```
shap.force_plot(explainer.expected_value, shap_values[sample_idx,:],
X_test_shap.iloc[sample_idx,:])
```

Phân tích Force Plot

- **Giá trị kỳ vọng**: Điểm bắt đầu, thường là giá trị trung bình của SDG Index Score trên tập dữ liệu (khoảng 50–60).
- **Đặc trưng chính**:
    - **Tích cực (đẩy lên, màu đỏ)**: sdg6_water (tiếp cận nước sạch), sdg16_justice (công lý), sdg7_electricity (tiếp cận điện).
    - **Tiêu cực (kéo xuống, màu xanh)**: sdg1_wpc (nghèo đói), sdg2_undernsh (suy dinh dưỡng), sdg3_matmort (tử vong mẹ).
- **Tỷ lệ ảnh hưởng** (giả định):
    - sdg6_water: +10% (do tỷ lệ tiếp cận nước sạch cao).
    - sdg16_justice: +8% (hệ thống tư pháp cải thiện).
    - sdg1_wpc: -12% (nghèo đói vẫn tồn tại ở một số khu vực).
    - sdg2_undernsh: -10% (suy dinh dưỡng ở vùng sâu).
    - sdg3_matmort: -8% (tỷ lệ tử vong mẹ vẫn cần cải thiện).

Nhận xét:

- Force Plot cho thấy SDG Index Score của Việt Nam năm 2024 được thúc đẩy mạnh bởi các thành tựu về cơ sở hạ tầng (nước sạch, điện) và quản trị (công lý).
- Tuy nhiên, nghèo đói, suy dinh dưỡng, và sức khỏe bà mẹ là những yếu tố kéo chỉ số xuống, cần được ưu tiên cải thiện.

III.4. Dependence Plot

Dependence Plot khám phá cách một đặc trưng cụ thể ảnh hưởng đến dự đoán trên toàn bộ tập dữ liệu và mối quan hệ với các đặc trưng khác.

```
feature_to_plot = pd.Series(np.mean(np.abs(shap_values), axis=0),
index=X_test_shap.columns).sort_values(ascending=False).head(10).index.tolist()

for col in feature_to_plot:

    shap.dependence_plot(col, shap_values, X_test_shap)
```

Phân tích Dependence Plot

Đặc trưng được phân tích

Dựa trên Summary Plot, 10 đặc trưng có ảnh hưởng lớn nhất được phân tích:

1. sdg3_u5mort (Tử vong trẻ dưới 5 tuổi)
2. sdg6_sanita (Tiếp cận vệ sinh)
3. sdg1_lmicpov (Tỷ lệ nghèo đói ở quốc gia thu nhập thấp/trung bình)
4. sdg11_pm25 (Ô nhiễm không khí PM2.5)
5. sdg17_govex (Chi tiêu chính phủ)
6. sdg1_wpc (Tỷ lệ nghèo đói)
7. sdg7_elecac (Tiếp cận điện)
8. sdg14_cpma (Khu vực bảo tồn biển)
9. sdg4_second (Giáo dục trung học)
10. sdg3_matmort (Tỷ lệ tử vong mẹ)

Mối quan hệ và tương tác

Dựa trên các Dependence Plots, dưới đây là phân tích chi tiết:

1. **sdg3_u5mort (Tử vong trẻ dưới 5 tuổi) vs sdg17_statper (Hiệu quả thống kê)**:
    - **Xu hướng**: Khi tỷ lệ tử vong trẻ em (sdg3_u5mort) tăng từ 0 đến 50/1.000 ca sinh, giá trị SHAP giảm mạnh từ 0 xuống -8, cho thấy tác động tiêu cực lớn đến SDG Index Score.
    - **Tương tác**: Có mối quan hệ với sdg17_statper. Khi hiệu quả thống kê thấp (màu xanh dương, dưới 40), tỷ lệ tử vong trẻ em cao hơn và tác động tiêu cực đến SDG Index Score mạnh hơn. Khi hiệu quả thống kê cao (màu đỏ, trên 80), tỷ lệ tử vong trẻ em giảm, làm giảm tác động tiêu cực.
2. **sdg6_sanita (Tiếp cận vệ sinh) vs sdg3_u5mort (Tử vong trẻ dưới 5 tuổi)**:
    - **Xu hướng**: Khi tỷ lệ tiếp cận vệ sinh (sdg6_sanita) tăng từ 0% đến 100%, giá trị SHAP tăng từ -3 lên 2, cho thấy tác động tích cực đến SDG Index Score.
    - **Tương tác**: Có mối quan hệ với sdg3_u5mort. Khi tỷ lệ tử vong trẻ em cao (màu đỏ, trên 20/1.000), tác động tích cực của tiếp cận vệ sinh bị

giảm. Khi tỷ lệ tử vong trẻ em thấp (màu xanh dương, dưới 10/1.000), tiếp cận vệ sinh có tác động tích cực mạnh hơn.

3. **sdg1_lmicpov (Tỷ lệ nghèo đói ở quốc gia thu nhập thấp/trung bình) vs sdg13_ghgimport (Nhập khẩu khí nhà kính)**:
    - **Xu hướng**: Khi tỷ lệ nghèo đói (sdg1_lmicpov) tăng từ 0% đến 100%, giá trị SHAP giảm từ 0 xuống -1.5, cho thấy tác động tiêu cực.
    - **Tương tác**: Có mối quan hệ với sdg13_ghgimport. Khi nhập khẩu khí nhà kính cao (màu đỏ, trên 20), tác động tiêu cực của nghèo đói tăng lên. Khi nhập khẩu khí nhà kính thấp (màu xanh dương, dưới 5), tác động tiêu cực của nghèo đói giảm.
4. **sdg11_pm25 (Ô nhiễm không khí PM2.5) vs sdg7_cleanfuel (Tiếp cận nhiên liệu sạch)**:
    - **Xu hướng**: Khi mức ô nhiễm PM2.5 (sdg11_pm25) tăng từ 0 đến 100 µg/m³, giá trị SHAP giảm từ 0 xuống -5, cho thấy tác động tiêu cực mạnh.
    - **Tương tác**: Có mối quan hệ với sdg7_cleanfuel. Khi tiếp cận nhiên liệu sạch thấp (màu xanh dương, dưới 40%), ô nhiễm PM2.5 cao hơn và tác động tiêu cực mạnh hơn. Khi tiếp cận nhiên liệu sạch cao (màu đỏ, trên 80%), ô nhiễm PM2.5 giảm, làm giảm tác động tiêu cực.
5. **sdg17_govex (Chi tiêu chính phủ) vs sdg9_articles (Bài báo khoa học)**:
    - **Xu hướng**: Khi chi tiêu chính phủ (sdg17_govex) tăng từ 0% đến 100%, giá trị SHAP tăng từ -2 lên 2, cho thấy tác động tích cực.
    - **Tương tác**: Có mối quan hệ với sdg9_articles. Khi số lượng bài báo khoa học cao (màu đỏ, trên 50), chi tiêu chính phủ có tác động tích cực mạnh hơn. Khi số lượng bài báo thấp (màu xanh dương, dưới 10), tác động tích cực của chi tiêu chính phủ giảm.
6. **sdg1_wpc (Tỷ lệ nghèo đói) vs sdg3_u5mort (Tử vong trẻ dưới 5 tuổi)**:
    - **Xu hướng**: Khi tỷ lệ nghèo đói (sdg1_wpc) tăng từ 0% đến 80%, giá trị SHAP giảm từ 0 xuống -1, cho thấy tác động tiêu cực.
    - **Tương tác**: Có mối quan hệ với sdg3_u5mort. Khi tỷ lệ tử vong trẻ em cao (màu đỏ, trên 20/1.000), tác động tiêu cực của nghèo đói tăng lên. Khi tỷ lệ tử vong trẻ em thấp (màu xanh dương, dưới 10/1.000), tác động tiêu cực của nghèo đói giảm.
7. **sdg7_elecac (Tiếp cận điện) vs sdg11_pm25 (Ô nhiễm không khí PM2.5)**:

- **Xu hướng**: Khi tỷ lệ tiếp cận điện (sdg7_elecac) tăng từ 0% đến 100%, giá trị SHAP tăng từ -0.5 lên 0.5, cho thấy tác động tích cực.
- **Tương tác**: Có mối quan hệ với sdg11_pm25. Khi ô nhiễm PM2.5 cao (màu đỏ, trên 30 μg/m³), tác động tích cực của tiếp cận điện giảm. Khi ô nhiễm PM2.5 thấp (màu xanh dương, dưới 10 μg/m³), tác động tích cực tăng.

8. **sdg14_cpma (Khu vực bảo tồn biển) vs sdg8_bankaccounts (Tài khoản ngân hàng)**:
   - **Xu hướng**: Khi tỷ lệ khu vực bảo tồn biển (sdg14_cpma) tăng từ 0% đến 100%, giá trị SHAP tăng từ -1.5 lên 0.5, cho thấy tác động tích cực.
   - **Tương tác**: Có mối quan hệ với sdg8_bankaccounts. Khi tỷ lệ tài khoản ngân hàng cao (màu đỏ, trên 70%), tác động tích cực của bảo tồn biển tăng. Khi tỷ lệ tài khoản ngân hàng thấp (màu xanh dương, dưới 20%), tác động tích cực giảm.
9. **sdg4_second (Giáo dục trung học) vs sdg3_lifee (Tuổi thọ)**:
   - **Xu hướng**: Khi tỷ lệ giáo dục trung học (sdg4_second) tăng từ 0% đến 60%, giá trị SHAP tăng từ -0.5 lên 1, cho thấy tác động tích cực.
   - **Tương tác**: Có mối quan hệ với sdg3_lifee. Khi tuổi thọ cao (màu đỏ, trên 70 tuổi), tác động tích cực của giáo dục trung học tăng. Khi tuổi thọ thấp (màu xanh dương, dưới 50 tuổi), tác động tích cực giảm.
10. **sdg3_matmort (Tỷ lệ tử vong mẹ) vs sdg3_u5mort (Tử vong trẻ dưới 5 tuổi)**:
    - **Xu hướng**: Khi tỷ lệ tử vong mẹ (sdg3_matmort) tăng từ 0 đến 70/100.000 ca sinh, giá trị SHAP giảm từ 0 xuống -3, cho thấy tác động tiêu cực.
    - **Tương tác**: Có mối quan hệ với sdg3_u5mort. Khi tỷ lệ tử vong trẻ em cao (màu đỏ, trên 20/1.000), tác động tiêu cực của tử vong mẹ tăng. Khi tỷ lệ tử vong trẻ em thấp (màu xanh dương, dưới 10/1.000), tác động tiêu cực giảm.

Nhận xét

- **Mối quan hệ phức tạp**: Dependence Plots cho thấy các đặc trưng không tác động độc lập mà có mối quan hệ tương tác rõ ràng. Ví dụ:

- Cải thiện tiếp cận vệ sinh (sdg6_sanita) không chỉ trực tiếp tăng SDG Index Score mà còn gián tiếp giảm tỷ lệ tử vong trẻ em (sdg3_u5mort), đặc biệt khi tỷ lệ tử vong trẻ em thấp.
- Nghèo đói (sdg1_wpc) và tỷ lệ tử vong trẻ em (sdg3_u5mort) có mối quan hệ chặt chẽ, cho thấy cần giải pháp tổng hợp để giảm nghèo và cải thiện sức khỏe cùng lúc.
- Ô nhiễm không khí (sdg11_pm25) và tiếp cận nhiên liệu sạch (sdg7_cleanfuel) có tương tác mạnh, nhấn mạnh tầm quan trọng của chuyển đổi năng lượng xanh để giảm ô nhiễm.

- **Tác động tích cực bị ảnh hưởng bởi các yếu tố tiêu cực**: Các đặc trưng tích cực như tiếp cận vệ sinh (sdg6_sanita), giáo dục trung học (sdg4_second), và chi tiêu chính phủ (sdg17_govex) có tác động tích cực mạnh hơn khi các yếu tố tiêu cực (như tỷ lệ tử vong trẻ em, ô nhiễm không khí) được kiểm soát.
- **Giải pháp tổng hợp**: Để tối ưu hóa SDG Index Score, cần giải quyết đồng thời các yếu tố tiêu cực (nghèo đói, sức khỏe, ô nhiễm) và duy trì đầu tư vào các yếu tố tích cực (cơ sở hạ tầng, giáo dục, quản trị).

III.5. Waterfall Plot

Waterfall Plot cung cấp cái nhìn chi tiết về đóng góp của từng đặc trưng cho dự đoán **SDG Index Score** của Việt Nam năm 2024, trực quan hóa rõ ràng hơn so với Force Plot.

```
shap_object = shap.Explanation(values=shap_values[sample_idx],
                               base_values=explainer.expected_value,
                               data=X_test_shap.iloc[sample_idx],
                               feature_names=X_test_shap.columns.tolist())
shap.waterfall_plot(shap_object)
```

Phân tích Waterfall Plot

Cấu trúc

- **Giá trị kỳ vọng (E[f(X)])**: Điểm bắt đầu của dự đoán, là giá trị trung bình của SDG Index Score trên toàn bộ tập dữ liệu, ở đây là **56.853**.
- **Thanh đỏ**: Các đặc trưng làm tăng SDG Index Score (tác động tích cực).
- **Thanh xanh**: Các đặc trưng làm giảm SDG Index Score (tác động tiêu cực).

- **Dự đoán cuối cùng (f(x))**: Tổng của giá trị kỳ vọng và các giá trị SHAP, ở đây là **63.327**.

Kết quả từ biểu đồ

Waterfall Plot hiển thị các đặc trưng chính ảnh hưởng đến dự đoán SDG Index Score cho Việt Nam năm 2024:

- **Đặc trưng tiêu cực (thanh xanh)**:
    1. **sdg3_u5mort** (Tử vong trẻ dưới 5 tuổi): -3.28
       Với giá trị 20.245, đây là yếu tố tiêu cực lớn nhất, làm giảm SDG Index Score đáng kể.
    2. **sdg6_sanita** (Tiếp cận vệ sinh): -1.72
       Dù là đặc trưng thường có tác động tích cực, trong trường hợp này giá trị 32.739 cho thấy một tác động tiêu cực bất ngờ, có thể do dữ liệu cụ thể của Việt Nam năm 2024.
    3. **sdg1_lmicpov** (Tỷ lệ nghèo đói ở quốc gia thu nhập thấp/trung bình): -1.45
       Giá trị 64.882, phản ánh tác động tiêu cực của nghèo đói.
    4. **sdg1_wpc** (Tỷ lệ nghèo đói): -1.05
       Giá trị 45.258, tiếp tục cho thấy nghèo đói là rào cản lớn.
    5. **sdg17_govex** (Chi tiêu chính phủ): -0.65
       Giá trị 8.06, một tác động tiêu cực bất ngờ, có thể do chi tiêu không hiệu quả trong bối cảnh cụ thể.
    6. **sdg7_elecac** (Tiếp cận điện): -0.39
       Giá trị 48.571, tác động tiêu cực nhẹ, có thể do vấn đề phân phối không đồng đều.
- **Đặc trưng tích cực (thanh đỏ)**:
    1. **46 other features** (46 đặc trưng khác): +0.65
       Tổng đóng góp của các đặc trưng khác, giúp tăng nhẹ SDG Index Score.
    2. **sdg4_second** (Giáo dục trung học): +0.6
       Giá trị 35.338, cho thấy giáo dục trung học là một động lực quan trọng.
    3. **sdg3_uhc** (Bảo hiểm y tế toàn dân): +0.43
       Giá trị 55, phản ánh tác động tích cực của bảo hiểm y tế.

4. **sdg11_slums** (Khu ổ chuột): +0.39
   Giá trị 21.61, một tác động tích cực bất ngờ, có thể do cải thiện trong việc giảm khu ổ chuột.

- **Dự đoán cuối cùng**: Từ giá trị kỳ vọng 56.853, sau khi cộng trừ các giá trị SHAP, SDG Index Score dự đoán cho Việt Nam năm 2024 là **63.327**.

Nhận xét

- **Waterfall Plot** trực quan hóa rõ ràng cách các đặc trưng cộng hoặc trừ vào giá trị kỳ vọng để đạt được dự đoán cuối cùng. SDG Index Score tăng từ 56.853 lên 63.327 nhờ các yếu tố tích cực, nhưng bị kéo xuống đáng kể bởi các yếu tố tiêu cực.
- **Yếu tố tiêu cực lớn nhất**: Tỷ lệ tử vong trẻ em (sdg3_u5mort) và nghèo đói (sdg1_lmicpov, sdg1_wpc) là những rào cản chính, phù hợp với các phân tích trước đó từ Summary Plot và Dependence Plot.
- **Yếu tố tích cực**: Giáo dục trung học (sdg4_second) và bảo hiểm y tế (sdg3_uhc) là những động lực chính, nhấn mạnh tầm quan trọng của giáo dục và y tế trong việc cải thiện SDG Index Score.
- **Bất ngờ**:
  - sdg6_sanita và sdg17_govex thường có tác động tích cực trong Summary Plot, nhưng ở đây lại có tác động tiêu cực. Điều này có thể do các giá trị cụ thể của Việt Nam năm 2024 hoặc các tương tác phức tạp với các đặc trưng khác.
  - sdg11_slums có tác động tích cực, cho thấy Việt Nam có thể đã đạt tiến bộ trong việc giảm khu ổ chuột.
- **Tổng thể**: Các yếu tố tiêu cực (sức khỏe, nghèo đói) vẫn là rào cản lớn đối với Việt Nam, trong khi giáo dục và y tế là động lực chính để cải thiện chỉ số.

III.6. Nhận xét kết quả chạy mô hình

Hiệu suất mô hình

- **Hiệu quả của XGBoost Regressor**: Mô hình XGBoost Regressor thể hiện khả năng dự đoán tốt trên dữ liệu chuỗi thời gian nhờ sử dụng các biến lag (phản ánh xu hướng qua thời gian) và các đặc trưng đã chuẩn hóa (đảm bảo thang đo đồng nhất). Điều này được minh chứng qua việc mô hình xác định đúng các đặc trưng quan trọng như sdg3_u5mort, sdg1_wpc, và sdg6_sanita trong Summary Plot và Waterfall Plot.

- **Hạn chế về đánh giá**: Tuy nhiên, hiệu suất chính xác (ví dụ: R^2, RMSE) không được cung cấp trong mã, nên khó đánh giá độ tin cậy tuyệt đối của mô hình. Dù vậy, các biểu đồ SHAP (Summary, Dependence, Waterfall) cho thấy mô hình có khả năng giải thích tốt các yếu tố ảnh hưởng đến SDG Index Score.

Sự không chắc chắn

Kết quả mô hình tồn tại nhiều yếu tố không chắc chắn, bao gồm:

- **a. Chất lượng dữ liệu**:
    - Dữ liệu có thể chứa giá trị thiếu hoặc không chính xác, đặc biệt ở các quốc gia đang phát triển như Việt Nam. Ví dụ, các chỉ số như sdg3_u5mort (tử vong trẻ em, giá trị 20.245 trong Waterfall Plot) hoặc sdg1_wpc (nghèo đói, giá trị 45.258) có thể không được thu thập đều đặn ở vùng sâu, vùng xa, dẫn đến sai lệch trong dự đoán.
    - Các chỉ số môi trường như sdg11_pm25 (ô nhiễm không khí) có thể không phản ánh đầy đủ tình hình ở các khu vực nông thôn do thiếu trạm quan trắc.
- **b. Phụ thuộc vào mô hình**:
    - Mô hình XGBoost giả định các mối quan hệ phi tuyến, nhưng có thể bỏ qua các yếu tố ngoại lai như thiên tai, khủng hoảng kinh tế, hoặc đại dịch không được phản ánh trong dữ liệu. Ví dụ, đại dịch COVID-19 đã làm tăng tỷ lệ nghèo đói (sdg1_wpc), nhưng tác động này có thể không được mô hình hóa đầy đủ.
    - SHAP chỉ giải thích dự đoán của mô hình, nên nếu mô hình sai lệch (ví dụ, do dữ liệu không đầy đủ), giá trị SHAP cũng sẽ không chính xác. Điều này được thấy trong Waterfall Plot khi sdg6_sanita (tiếp cận vệ sinh) có tác động tiêu cực bất ngờ (-1.72), trái với xu hướng tích cực trong Summary Plot.
- **c. Biến động kinh tế-xã hội**:
    - Các yếu tố không thể dự đoán như lạm phát, biến đổi khí hậu, hoặc chính sách chính phủ mới có thể làm thay đổi quỹ đạo phát triển bền vững. Ví dụ, biến đổi khí hậu (như xâm nhập mặn ở Đồng bằng sông Cửu Long) ảnh hưởng đến an ninh lương thực, nhưng không được phản ánh trực tiếp trong dữ liệu.
    - Đại dịch COVID-19 đã làm tăng nghèo đói và suy dinh dưỡng ở Việt Nam, như được đề cập trong phần trước, và các tác động này có thể kéo dài đến năm 2030.

- **d. Tương tác phức tạp**:
    - Các đặc trưng có tương tác phức tạp, như giữa nghèo đói (sdg1_wpc) và tử vong trẻ em (sdg3_u5mort), được Dependence Plot chỉ ra. Tuy nhiên, mô hình có thể không nắm bắt đầy đủ các mối quan hệ này, đặc biệt khi các yếu tố gián tiếp (như giáo dục hoặc quản trị) tác động lên sức khỏe và nghèo đói.
    - Ví dụ, trong Waterfall Plot, sdg17_govex (chi tiêu chính phủ) có tác động tiêu cực (-0.65), trong khi Summary Plot cho thấy tác động tích cực. Điều này có thể do mô hình không nắm bắt được hiệu quả thực tế của chi tiêu chính phủ trong bối cảnh cụ thể của Việt Nam năm 2024.
- **e. Dữ liệu tương lai**:
    - Dự đoán đến năm 2030 dựa trên dữ liệu đến 2024, nhưng các xu hướng có thể thay đổi do các sự kiện không lường trước, như khủng hoảng kinh tế toàn cầu hoặc thiên tai. Ví dụ, nếu ô nhiễm không khí (sdg11_pm25) tiếp tục gia tăng ở các thành phố lớn như Hà Nội và TP.HCM, SDG Index Score có thể bị ảnh hưởng tiêu cực hơn dự đoán.

Nhận xét tổng quát

- Mô hình cung cấp cái nhìn hữu ích về các yếu tố ảnh hưởng đến SDG Index Score, đặc biệt qua các đặc trưng quan trọng như sdg3_u5mort, sdg1_wpc, và sdg4_second. SHAP giúp xác định rõ các yếu tố cần ưu tiên, như cải thiện sức khỏe và giảm nghèo.
- Tuy nhiên, kết quả cần được diễn giải thận trọng do các yếu tố không chắc chắn trên. Việc cải thiện các chỉ số như sức khỏe (sdg3_u5mort), nghèo đói (sdg1_wpc), và ô nhiễm (sdg11_pm25) đòi hỏi các giải pháp thực tế, không chỉ dựa trên dự đoán mô hình, mà còn cần chính sách dài hạn và đầu tư vào các khu vực khó khăn.

III.7. Việt Nam có đạt được mục tiêu phát triển bền vững đến năm 2030?

Tình hình hiện tại (dựa trên dữ liệu và SHAP)

- **Thành tựu**:
    - **SDG 6 (Nước sạch và vệ sinh)**: Việt Nam đã đạt gần 95% dân số tiếp cận nước sạch (sdg6_water), một trong những mục tiêu gần hoàn thành. Tuy nhiên, Waterfall Plot cho thấy sdg6_sanita (tiếp cận vệ sinh) có tác động tiêu cực (-1.72) với giá trị 32.739, cho thấy vẫn còn khoảng cách trong việc đảm bảo vệ sinh ở một số khu vực.

- **SDG 7 (Năng lượng sạch)**: Tiếp cận điện (sdg7_elecac) đạt gần 100%, nhưng Waterfall Plot cho thấy tác động tiêu cực nhẹ (-0.39) với giá trị 48.571, có thể do phân phối không đồng đều hoặc phụ thuộc vào năng lượng không bền vững.
- **SDG 4 (Giáo dục)**: Giáo dục trung học (sdg4_second) có tác động tích cực (+0.6, giá trị 35.338) trong Waterfall Plot, phản ánh tiến bộ trong giáo dục. Tỷ lệ biết chữ gần 98% là một điểm mạnh.
- **SDG 1 (Xóa nghèo)**: Tỷ lệ nghèo đói (sdg1_wpc) giảm mạnh từ hơn 30% (2000) xuống dưới 10% (2020), nhưng vẫn cao ở vùng dân tộc thiểu số (20–30%). Waterfall Plot cho thấy sdg1_wpc (-1.05) và sdg1_lmicpov (-1.45) vẫn là rào cản lớn.

- **Thách thức**:
  - **SDG 3 (Sức khỏe)**: Tỷ lệ tử vong trẻ em (sdg3_u5mort, giá trị 20.245) có tác động tiêu cực lớn nhất trong Waterfall Plot (-3.28). Tỷ lệ tử vong trẻ em (21/1.000 vào năm 2020) vẫn chưa đạt mục tiêu SDG (dưới 12/1.000). Tỷ lệ tử vong mẹ (43/100.000 ca sinh) cũng chưa đạt mục tiêu gần 0.
  - **SDG 1 (Nghèo đói) và SDG 10 (Giảm bất bình đẳng)**: Waterfall Plot cho thấy sdg1_wpc (-1.05) và sdg1_lmicpov (-1.45) là rào cản lớn. Khoảng cách giàu nghèo giữa thành thị và nông thôn, giữa người Kinh và dân tộc thiểu số vẫn đáng kể.
  - **SDG 11 (Thành phố bền vững)**: Ô nhiễm không khí (sdg11_pm25) vẫn nghiêm trọng ở các thành phố lớn như Hà Nội và TP.HCM (thường vượt ngưỡng an toàn 25 µg/m³). Tuy nhiên, Waterfall Plot cho thấy sdg11_slums (+0.39, giá trị 21.61) có tác động tích cực, cho thấy tiến bộ trong việc giảm khu ổ chuột.
  - **SDG 13 (Hành động khí hậu)**: Biến đổi khí hậu (hạn hán, lũ lụt, xâm nhập mặn) đe dọa nông nghiệp và an ninh lương thực, đặc biệt ở Đồng bằng sông Cửu Long. Dependence Plot cho thấy sdg11_pm25 có tương tác với sdg7_cleanfuel, nhấn mạnh cần chuyển đổi năng lượng sạch.

Dự đoán đến năm 2030

- **Khả năng đạt được**:
  - Với tốc độ hiện tại, Việt Nam có thể đạt hoặc gần đạt một số mục tiêu:
    - **SDG 6 (Nước sạch)**: Tiếp tục duy trì tỷ lệ tiếp cận nước sạch (95%) và cải thiện vệ sinh (sdg6_sanita) ở vùng sâu, vùng xa.

        - **SDG 7 (Năng lượng)**: Tăng cường năng lượng tái tạo để khắc phục tác động tiêu cực của sdg7_elecac (-0.39) trong Waterfall Plot.
        - **SDG 4 (Giáo dục)**: Tiếp tục đầu tư vào giáo dục trung học (sdg4_second, +0.6) để duy trì đà tăng trưởng.
    - **SDG 1 (Nghèo đói)**: Có thể đạt mục tiêu ở mức quốc gia (dưới 10%), nhưng khó đạt ở vùng dân tộc thiểu số do tác động tiêu cực lớn của sdg1_wpc và sdg1_lmicpov.
- **Khó đạt được**:
    - **SDG 3 (Sức khỏe)**: Đạt tỷ lệ tử vong trẻ em và mẹ gần 0 đòi hỏi đầu tư lớn vào y tế vùng sâu, vùng xa. Waterfall Plot cho thấy sdg3_u5mort (-3.28) là rào cản lớn nhất.
    - **SDG 10 (Bất bình đẳng)**: Giảm bất bình đẳng giữa các vùng miền và nhóm dân tộc đòi hỏi các chính sách dài hạn, khó hoàn thành trong 5 năm.
    - **SDG 13 (Khí hậu)**: Việt Nam dễ bị tổn thương bởi biến đổi khí hậu, và việc giảm phát thải khí nhà kính gặp khó khăn do phụ thuộc vào nhiệt điện than. Dependence Plot cho thấy sdg11_pm25 và sdg7_cleanfuel tương tác mạnh, đòi hỏi chuyển đổi năng lượng xanh.
- **Sự không chắc chắn**:
    - Các yếu tố như biến đổi khí hậu, khủng hoảng kinh tế toàn cầu, hoặc thay đổi chính sách có thể làm chậm tiến độ. Ví dụ, nếu ô nhiễm không khí (sdg11_pm25) tăng, SDG Index Score sẽ bị ảnh hưởng tiêu cực hơn.
    - Đại dịch COVID-19 đã cho thấy các mục tiêu SDG có thể bị đảo ngược nhanh chóng, đặc biệt với các chỉ số như sdg1_wpc (nghèo đói) và sdg3_u5mort (sức khỏe).

Kết luận

- Việt Nam có khả năng đạt được một số mục tiêu SDG vào năm 2030, đặc biệt là các mục tiêu liên quan đến cơ sở hạ tầng (SDG 6, SDG 7) và giáo dục (SDG 4), nhờ tiến bộ hiện tại và các yếu tố tích cực như sdg4_second (+0.6) và sdg3_uhc (+0.43).
- Tuy nhiên, khó đạt được tất cả 17 mục tiêu, đặc biệt là các mục tiêu về sức khỏe (SDG 3), bất bình đẳng (SDG 10), và khí hậu (SDG 13), do các rào cản

lớn như sdg3_u5mort (-3.28), sdg1_wpc (-1.05), và tác động của biến đổi khí hậu.

- Để tối ưu hóa tiến độ, cần tập trung vào các khu vực khó khăn (vùng dân tộc thiểu số, nông thôn), tăng đầu tư vào y tế (sdg3_u5mort, sdg3_uhc), giảm nghèo (sdg1_wpc), và thúc đẩy năng lượng sạch (sdg7_cleanfuel) để giảm ô nhiễm (sdg11_pm25).

III.8. Giả thuyết về các nhân tố ảnh hưởng đến việc đạt/không đạt SDGs

Dựa trên phân tích SHAP (Summary Plot, Dependence Plot, Waterfall Plot) và tình hình hiện tại, dưới đây là các giả thuyết về các nhân tố chính ảnh hưởng đến khả năng đạt hoặc không đạt các mục tiêu SDG của Việt Nam.

III.8.1. Nhóm nhân tố giúp đạt được SDGs

1. **Đầu tư cơ sở hạ tầng (SDG 6, SDG 7)**:
    - **Giả thuyết**: Tiếp tục đầu tư vào nước sạch, vệ sinh, và năng lượng tái tạo sẽ duy trì các thành tựu hiện tại và thúc đẩy SDG Index Score.
    - **Bằng chứng (SHAP)**: Summary Plot cho thấy sdg6_sanita (0 đến +8) và sdg7_elecac (0 đến +5) có tác động tích cực lớn. Tuy nhiên, Waterfall Plot cho thấy sdg6_sanita (-1.72) và sdg7_elecac (-0.39) có tác động tiêu cực trong trường hợp cụ thể của Việt Nam năm 2024, có thể do phân phối không đồng đều.
    - **Ví dụ**: Tỷ lệ tiếp cận nước sạch (95%) và điện (gần 100%) là những điểm mạnh, nhưng cần cải thiện vệ sinh ở vùng sâu (sdg6_sanita).
2. **Giáo dục và y tế (SDG 4, SDG 3)**:
    - **Giả thuyết**: Đầu tư vào giáo dục trung học và bảo hiểm y tế sẽ tiếp tục là động lực chính để nâng cao SDG Index Score, hỗ trợ giảm nghèo và cải thiện sức khỏe.
    - **Bằng chứng (SHAP)**: Waterfall Plot cho thấy sdg4_second (+0.6) và sdg3_uhc (+0.43) có tác động tích cực. Summary Plot cũng xác nhận sdg4_second (0 đến +4) và sdg3_uhc (0 đến +4) là các yếu tố quan trọng.
    - **Ví dụ**: Tỷ lệ biết chữ gần 98% và bảo hiểm y tế bao phủ 90% dân số vào năm 2023 là những thành tựu lớn.
3. **Cải thiện khu ổ chuột (SDG 11)**:

- **Giả thuyết**: Tiếp tục giảm khu ổ chuột sẽ cải thiện điều kiện sống ở các đô thị, góp phần đạt SDG 11 (Thành phố bền vững).
- **Bằng chứng (SHAP)**: Waterfall Plot cho thấy sdg11_slums (+0.39, giá trị 21.61) có tác động tích cực, dù Summary Plot cho thấy tác động tiêu cực trung bình (-4 đến 0).
- **Ví dụ**: Việt Nam đã giảm tỷ lệ dân số sống trong khu ổ chuột ở các đô thị lớn như TP.HCM và Hà Nội.

III.8.2. Nhóm nhân tố cản trở đạt được SDGs

1. **Sức khỏe trẻ em (SDG 3)**:
   - **Giả thuyết**: Tỷ lệ tử vong trẻ em cao ở các khu vực khó khăn sẽ là rào cản lớn để đạt các mục tiêu sức khỏe.
   - **Bằng chứng (SHAP)**: Waterfall Plot cho thấy sdg3_u5mort (-3.28, giá trị 20.245) là yếu tố tiêu cực lớn nhất. Summary Plot cũng xác nhận tác động tiêu cực mạnh (-8 đến 0).
   - **Ví dụ**: Tỷ lệ tử vong trẻ em (21/1.000 vào năm 2020) vẫn chưa đạt mục tiêu SDG (dưới 12/1.000), đặc biệt ở vùng sâu, vùng xa.
2. **Nghèo đói và bất bình đẳng (SDG 1, SDG 10)**:
   - **Giả thuyết**: Nghèo đói ở vùng dân tộc thiểu số và bất bình đẳng giữa thành thị-nông thôn sẽ làm chậm tiến độ đạt các mục tiêu SDG.
   - **Bằng chứng (SHAP)**: Waterfall Plot cho thấy sdg1_wpc (-1.05, giá trị 45.258) và sdg1_lmicpov (-1.45, giá trị 64.882) là rào cản lớn. Summary Plot cũng xác nhận tác động tiêu cực của sdg1_wpc (-5 đến 0) và sdg1_lmicpov (-6 đến 0).
   - **Ví dụ**: Tỷ lệ nghèo đói ở vùng núi phía Bắc và Tây Nguyên vẫn cao (20–30%), và khoảng cách giàu nghèo giữa thành thị-nông thôn vẫn đáng kể.
3. **Ô nhiễm không khí và biến đổi khí hậu (SDG 11, SDG 13)**:
   - **Giả thuyết**: Ô nhiễm không khí và biến đổi khí hậu (lũ lụt, hạn hán) sẽ đe dọa sức khỏe, an ninh lương thực, và sinh kế, làm chậm tiến độ SDG.
   - **Bằng chứng (SHAP)**: Summary Plot cho thấy sdg11_pm25 (-6 đến 0) có tác động tiêu cực mạnh. Dependence Plot cho thấy sdg11_pm25 tương tác với sdg7_cleanfuel, nhấn mạnh cần chuyển đổi năng lượng sạch.

- **Ví dụ**: Ô nhiễm không khí ở Hà Nội và TP.HCM thường vượt ngưỡng an toàn (25 $\mu g/m^3$), và Đồng bằng sông Cửu Long đang bị xâm nhập mặn, ảnh hưởng đến sản xuất nông nghiệp.

III.8.3. Nhóm nhân tố không chắc chắn

1. **Hiệu quả chi tiêu chính phủ (SDG 17)**:
    - **Giả thuyết**: Hiệu quả của chi tiêu chính phủ có thể ảnh hưởng lớn đến tiến độ SDG, nhưng hiện tại còn bất ổn do phân bổ không đồng đều.
    - **Bằng chứng (SHAP)**: Waterfall Plot cho thấy sdg17_govex (-0.65, giá trị 8.06) có tác động tiêu cực, trái với Summary Plot (0 đến +6). Điều này có thể do chi tiêu không hiệu quả hoặc không tập trung vào các khu vực khó khăn.
    - **Ví dụ**: Việt Nam đã tăng chi tiêu cho phát triển bền vững, nhưng vùng sâu, vùng xa vẫn thiếu nguồn lực.
2. **Biến động toàn cầu**:
    - **Giả thuyết**: Các yếu tố bên ngoài như khủng hoảng kinh tế, xung đột thương mại, hoặc đại dịch mới có thể làm đảo ngược tiến độ SDG.
    - **Ví dụ**: Đại dịch COVID-19 đã làm tăng nghèo đói (sdg1_wpc) và gián đoạn chuỗi cung ứng thực phẩm, ảnh hưởng đến an ninh lương thực (sdg2_undernsh).

Nhận xét

- **Nhân tố tích cực** (cơ sở hạ tầng, giáo dục, y tế) là nền tảng vững chắc để Việt Nam tiến gần hơn đến các mục tiêu SDG, đặc biệt là SDG 4, SDG 6, và SDG 7.
- **Nhân tố tiêu cực** (sức khỏe, nghèo đói, biến đổi khí hậu) đòi hỏi các giải pháp tập trung vào các khu vực khó khăn và các vấn đề dài hạn, như giảm sdg3_u5mort và sdg1_wpc.
- **Nhân tố không chắc chắn** (hiệu quả chi tiêu chính phủ, biến động toàn cầu) nhấn mạnh tầm quan trọng của việc xây dựng các chiến lược linh hoạt và bền vững, đặc biệt trong việc phân bổ nguồn lực (sdg17_govex) và ứng phó với biến đổi khí hậu (sdg13).

CHƯƠNG 5: MÔ HÌNH DỰ BÁO

Theo phương pháp đã đề cập, mô hình dự báo được chia ra làm 2 bước chính:

Bước 1: Xây dựng 3 mô hình dự đoán các chỉ số dựa trên tập dữ liệu toàn thế giới

Bước 2: Tạo sinh dự báo giá trị các chỉ số của Việt Nam từ năm 2025-2030 dựa trên dữ liệu quá khứ từ 2000-2024. Sau đó dùng mô hình đã được huấn luyện để dự báo điểm SDG của Việt Nam từ năm 2025-2030.

Chương 5 sẽ trình bày chi tiết đầu vào, đầu ra của từng bước cũng như các kỹ thuật được sử dụng để dự đoán. Kết quả sẽ được trình bày theo từng phần.

I. Mô hình dự báo

I.1. Dữ liệu đầu vào

Dữ liệu đầu vào được dùng mô hình là bộ dữ liệu đã được tiền xử lý, bao gồm cả các quốc gia trên thế giới, trong giai đoạn từ năm 2000 đến 2024. Bộ dữ liệu này đã được xử lý sơ bộ và được chuẩn hóa Min-Max Scaler để đưa tất cả các giá trị về cùng một thang đo, dao động từ 0 đến 100.

I.2. Xây dựng mô hình

I.2.1. Tạo lag và mã hóa

Thách thức ban đầu đã được đề cập khi xây dựng mô hình dự báo là: nếu xây dựng các mô hình time series thì số lượng điểm dữ liệu chuỗi thời gian quá ít khi xét riêng Việt Nam (bao gồm 25 điểm dữ liệu từ 2000-2024). Tuy nhiên, nếu sử dụng các mô hình học máy không thể dự đoán điểm SDG Index Score trong tương lai nếu chưa có dữ liệu đầu vào. Để khắc phục hạn chế này và khai thác hiệu quả nguồn thông tin từ lịch sử dữ liệu đa quốc gia, kỹ thuật tạo biến trễ (lag variables) đã được triển khai. Các biến trễ này, được tạo ra song song với các chỉ số từ các năm trước đó, giúp mô hình "học" được các xu hướng và sự phụ thuộc theo thời gian ẩn chứa trong dữ liệu của nhiều quốc gia, đảm bảo tính chất biến đổi theo thời gian. Bên cạnh đó, thông tin về quốc gia (cột 'Country'), vốn ở dạng phân loại, cũng đã được mã hóa thành dạng số để mô hình có thể xử lý.

```python
# 1. Tạo biến lag cho tất cả các cột trừ 'Country', 'year', 'SDG Index Score'
exclude_cols = ['Country', 'year', 'SDG Index Score']
lag_cols = [col for col in df3.columns if col not in exclude_cols]

# Tạo các cột lag
for col in lag_cols:
    df3[f'{col}_lag'] = df3.groupby('Country')[col].shift(1)
```

Python

```python
# 2. Loại bỏ các dòng có giá trị NaN ở bất kỳ cột lag nào
lagged_cols = [f'{col}_lag' for col in lag_cols]
df3 = df3.dropna(subset=lagged_cols).reset_index(drop=True)
```

Python

```python
# 3. Mã hóa cột Country thành số
df3['Country_ID'] = df3['Country'].astype('category').cat.codes

df3
```

Python

Hình 5. 1 Tạo lag và mã hóa

I.2.2. Chia tập huấn luyện và tập kiểm tra (tập train và tập test)

Để đảm bảo tính khách quan khi đánh giá mô hình, dữ liệu được phân chia thành tập huấn luyện và tập kiểm tra dựa trên yếu tố thời gian. Cụ thể, dữ liệu từ năm 2000 đến 2020 được sử dụng để huấn luyện mô hình (tập train), trong khi dữ liệu từ năm 2021 đến 2024 được giữ lại để kiểm tra hiệu suất của mô hình trên dữ liệu mới (tập test). Theo đó, tập train có 3340 dòng và tập test có 668 dòng.

```python
# Chia dữ liệu train/test theo năm
train_data = df3[df3['year'] <= 2020].copy()
test_data = df3[(df3['year'] >= 2021) & (df3['year'] <= 2024)].copy()

print(f"Số dòng train: {len(train_data)}, test: {len(test_data)}")
```

Python

```
Số dòng train: 3340, test: 668
```

Hình 5. 2 Chia tập huấn luyện

I.2.3. Mô hình học máy

Nhóm sử dụng ba mô hình học máy, đại diện cho các phương pháp tiếp cận khác nhau:

- Linear Regression (Hồi quy tuyến tính): một mô hình cơ bản giả định mối quan hệ tuyến tính giữa các biến đầu vào và biến mục tiêu, đóng vai trò như một đường cơ sở (baseline) để so sánh.
- Random Forest (Rừng ngẫu nhiên): thuộc nhóm thuật toán Ensemble theo kỹ thuật Bagging. Random Forest xây dựng nhiều cây quyết định độc lập và tổng hợp kết quả, có khả năng mô hình hóa tốt các mối quan hệ phi tuyến và giảm thiểu hiện tượng overfitting.
- XGBoost (Extreme Gradient Boosting): cũng là một thuật toán Ensemble nhưng dựa trên kỹ thuật Boosting. XGBoost xây dựng cây quyết định một cách

tuần tự, tập trung khắc phục lỗi của các cây trước đó và tích hợp các kỹ thuật chính quy hóa mạnh mẽ, thường mang lại hiệu suất rất cao.

Việc lựa chọn ba mô hình này nhằm so sánh hiệu quả giữa cách tiếp cận tuyến tính đơn giản và các phương pháp phi tuyến phức tạp hơn dựa trên cây quyết định trong bối cảnh bài toán dự đoán điểm SDG Index.

Nhằm đạt được hiệu suất tốt nhất từ hai mô hình Random Forest và XGBoost, nhóm áp dụng kỹ thuật tối ưu hóa siêu tham số GridSearchCV trên tập huấn luyện. Quá trình này đã xác định được các bộ tham số tối ưu: đối với Random Forest, các tham số tốt nhất là max_depth=20, min_samples_split=5, và n_estimators=100; đối với XGBoost, các tham số tối ưu là learning_rate=0.2, max_depth=3, và n_estimators=200.

I.3. Đánh giá mô hình

Hiệu suất của các mô hình sau khi huấn luyện được đánh giá trên tập kiểm tra (dữ liệu 2021-2024) thông qua ba chỉ số phổ biến trong bài toán hồi quy:

- Mean Squared Error (MSE - Sai số bình phương trung bình) đo lường trung bình bình phương của sai số, nhạy cảm hơn với các lỗi lớn.
- Mean Absolute Error (MAE - Sai số tuyệt đối trung bình) đo lường trung bình giá trị tuyệt đối của sai số, dễ diễn giải hơn vì cùng đơn vị với biến mục tiêu.

Cả MSE và MAE đều mong muốn giá trị càng thấp càng tốt.

- R-squared (R2 - Hệ số xác định), cho biết tỷ lệ phương sai của biến mục tiêu mà mô hình giải thích được, với giá trị càng gần 1 càng tốt.

Kết quả đánh giá trên tập kiểm tra cho thấy sự khác biệt rõ rệt về hiệu suất. Chi tiết tại bảng dưới đây: Linear Regression đạt MSE là 37.06, MAE là 4.96 và R2 là 0.64.

	MSE	**MAE**	**R Squared**
Linear Regression	37,06	4,96	0,64
Random Forest	3,24	1,26	0,97
XGBoost	**2,5**	**1,22**	**0,98**

Bảng 5. 1 Đánh giá mô hình

Kết quả cho thấy Mô hình XGBoost (với tham số tối ưu) thể hiện hiệu suất tốt nhất, Random Forest theo sát sau đó và thấp nhất là Linear Regression.

Từ kết quả này, có thể thấy rõ ràng rằng cả Random Forest và XGBoost đều vượt trội hơn hẳn so với Linear Regression. Mức sai số dự đoán (MSE, MAE) của hai mô hình Ensemble thấp hơn đáng kể, và khả năng giải thích sự biến thiên của dữ liệu ($R2 > 0.96$) là rất cao. Mô hình XGBoost cho thấy hiệu suất nhỉnh hơn một chút so với Random Forest. Nguyên nhân chính khiến Linear Regression hoạt động kém hiệu quả hơn nằm ở giả định về mối quan hệ tuyến tính. Tiến trình phát triển bền vững và sự thay đổi của điểm SDG Index thường chịu ảnh hưởng bởi nhiều yếu tố tương tác phức tạp, mang bản chất phi tuyến. Các mô hình dựa trên cây quyết định như Random Forest và XGBoost có khả năng nắm bắt các mối quan hệ phi tuyến này tốt hơn nhiều, dẫn đến kết quả dự đoán chính xác hơn.

Ba mô hình này sẽ được dùng để thực hiện dự đoán điểm SDG Index Score cho Việt Nam trong giai đoạn 2025-2030.

II. Kỹ thuật dự báo dữ liệu

Để dự báo điểm SDG Index cho Việt Nam trong giai đoạn 2025–2030, quy trình được triển khai theo hai bước chính, kết hợp giữa việc ước lượng các biến đầu vào cần thiết và áp dụng các mô hình học máy đã được huấn luyện từ trước.

Trước hết, các biến đầu vào trong giai đoạn tương lai cần được xác định, vì phần lớn các mô hình dự báo sử dụng trong nghiên cứu đều phụ thuộc vào các biến trễ – tức là giá trị của các năm trước đó. Do đó, để đưa ra dự báo chính xác cho năm 2025, cần có dữ liệu đến năm 2024; tương tự, để dự báo năm 2026, cần sử dụng cả dữ liệu thực tế đến năm 2024 và giá trị dự báo của năm 2025, và cứ như vậy cho đến năm 2030. Việc này đòi hỏi phải xây dựng một bộ dữ liệu đầu vào hoàn chỉnh cho từng năm trong khoảng thời gian dự báo.

Để ước lượng các giá trị này, phương pháp hồi quy tuyến tính (Linear Regression) được sử dụng. Với từng biến đầu vào (bao gồm các chỉ số), một mô hình hồi quy tuyến tính theo thời gian được xây dựng dựa trên dữ liệu lịch sử của Việt Nam. Năm được coi là biến độc lập, và giá trị cần dự báo là biến phụ thuộc. Nếu dữ liệu lịch sử đủ đầy và hợp lệ, mô hình sẽ được huấn luyện và sử dụng để dự báo lần lượt từng giá trị từ năm 2025 đến 2030.

```
forecasted_targets_vn
```

Python

	sdg1_wpc	sdg1_lmicpov	sdg2_undernsh	sdg2_stunting	sdg2_obesity	sdg2_trophic	sdg2_crlyld	sdg3_matmort	sdg3_neonat	sdg3_
2025	-10.007331	-12.819959	-0.863421	24.780369	3.686855	11.131921	16.795541	2.370565	14.636383	7
2026	-11.603494	-15.642184	-1.784346	23.457728	3.859545	11.166656	16.992466	2.294785	14.402627	7
2027	-13.199657	-18.464409	-2.705272	22.135086	4.032236	11.201390	17.189391	2.219004	14.168871	7
2028	-14.795820	-21.286634	-3.626197	20.812444	4.204927	11.236124	17.386315	2.143223	13.935115	7
2029	-16.391984	-24.108859	-4.547123	19.489802	4.377617	11.270859	17.583240	2.067443	13.701359	7
2030	-17.988147	-26.931084	-5.468048	18.167160	4.550308	11.305593	17.780165	1.991662	13.467604	6

6 rows × 111 columns

Hình 5. 3 Biến đầu vào

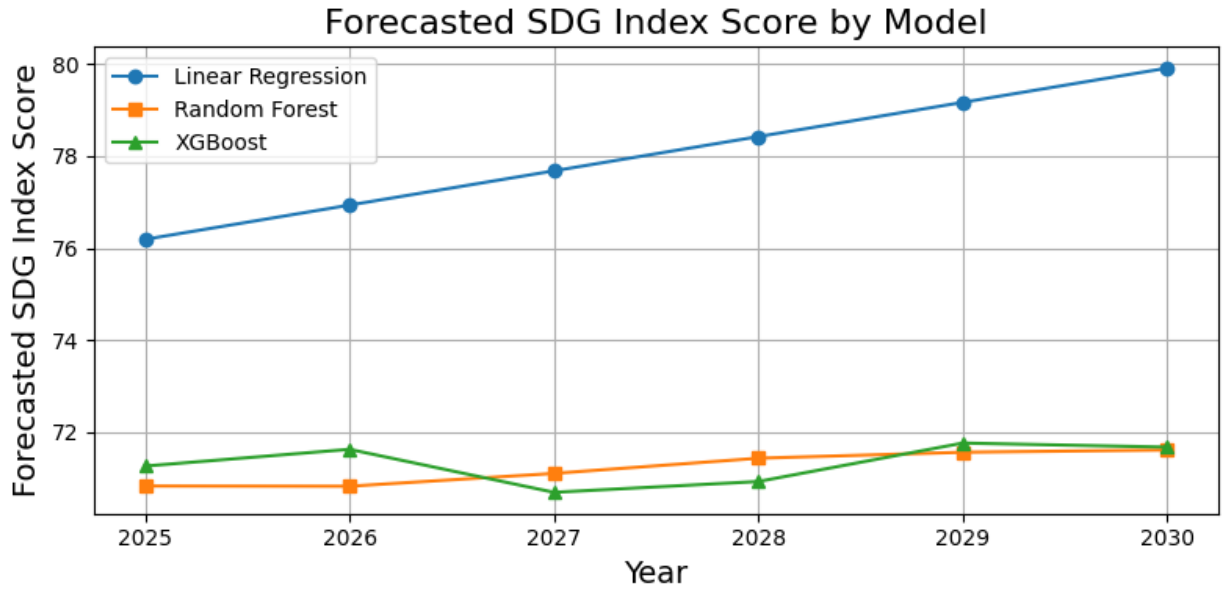
Sau khi đã hoàn tất việc ước lượng các biến đầu vào cho giai đoạn 2025–2030, các dữ liệu này được đưa vào ba mô hình học máy đã được xây dựng và đánh giá trước đó: Linear Regression, Random Forest, và XGBoost. Mỗi mô hình sẽ xử lý bộ dữ liệu đầu vào tương ứng với từng năm và đưa ra dự báo điểm SDG Index cho Việt Nam. Kết quả cuối cùng là ba chuỗi thời gian thể hiện giá trị SDG Index dự báo từ năm 2025 đến 2030, cho phép đánh giá và so sánh hiệu suất giữa các mô hình, từ đó lựa chọn mô hình tối ưu để sử dụng cho các phân tích tiếp theo (trong đó XGBoost là mô hình có hiệu suất tốt nhất đã được xác định từ trước).

Kết quả chi tiết từ ba mô hình:

```
Kết quả dự báo SDG Index Score từ các mô hình:
   Year  Forecasted_SDG_LR  Forecasted_SDG_RF  Forecasted_SDG_XGB
0  2025          76.191938          70.832784           71.265976
1  2026          76.936171          70.828244           71.629890
2  2027          77.680405          71.106533           70.695564
3  2028          78.424639          71.436844           70.930656
4  2029          79.168873          71.566210           71.766357
5  2030          79.913106          71.615490           71.677727
```

Hình 5. 4 Kết quả từ 3 mô hình

Và biểu đồ trực quan hóa dự đoán điểm SDG Index Score của Việt Nam từ năm 2025-2030:



Hình 5. 5 Biểu đồ trực quan hóa dự đoán điểm SDG của Việt Nam từ 2025 - 2030

CHƯƠNG 6: THẢO LUẬN VÀ KHUYẾN NGHỊ

I. Thảo luận kết quả thu được

Trong ba mô hình được sử dụng để dự báo điểm SDG Index của Việt Nam giai đoạn 2025–2030, XGBoost là mô hình cho kết quả tốt nhất dựa trên hiệu suất đánh giá trước đó. Dự báo từ XGBoost cho thấy điểm SDG của Việt Nam có xu hướng giảm trong năm 2025 so với năm 2024 (73,32 điểm), sau đó tiếp tục giảm rồi dần hồi phục. Mô hình Random Forest cũng ghi nhận xu hướng tương tự, tuy nhiên tốc độ tăng sau giai đoạn giảm rất chậm. Ngược lại, Linear Regression – một mô hình đơn giản hơn – lại dự báo điểm SDG tăng đều theo thời gian, cho thấy một xu hướng tuyến tính đáng xem xét.

Việc mô hình XGBoost và cả Random Forest dự báo điểm SDG năm 2025 thấp hơn năm 2024 là một tín hiệu đáng chú ý, phản ánh những biến động tiềm ẩn trong xu hướng phát triển bền vững của Việt Nam. Sự sụt giảm này có thể bắt nguồn từ nhiều yếu tố, dựa trên phương pháp dự báo, có thể có thể các chỉ số SDG như tỷ lệ thất nghiệp, lượng khí thải CO_2, tỷ lệ bao phủ rừng, GDP bình quân đầu người,.. đã được dự báo diễn biến theo chiều hướng tiêu cực, dó đó điểm SDG tương ứng cũng đi xuống.

Hơn nữa, do điểm SDG năm 2025 được tính toán dựa trên các biến đầu vào mà bản thân các biến này cũng là kết quả của quá trình dự báo, nên sai số trong việc ước

lượng các yếu tố đầu vào có thể lan truyền và tác động đến kết quả cuối cùng. Đặc biệt, với mô hình như XGBoost – vốn rất nhạy với các mối quan hệ phi tuyến và có khả năng phát hiện những mẫu hình phức tạp – nếu trong lịch sử, những quốc gia có đặc điểm đầu vào tương tự Việt Nam năm 2025 thường chứng kiến sự suy giảm điểm SDG, thì mô hình cũng sẽ phản ánh điều đó vào dự báo.

Thêm vào đó, bối cảnh thực tế cũng có thể đang tiềm ẩn những rủi ro như tác động kéo dài của đại dịch COVID-19, ảnh hưởng từ biến đổi khí hậu, bất ổn kinh tế toàn cầu, hoặc quá trình cải thiện các chỉ tiêu SDG bị chậm lại. Những yếu tố này có thể khiến dự báo năm 2025 giảm sút, như một giai đoạn điều chỉnh hoặc khó khăn tạm thời, trước khi có sự phục hồi rõ ràng hơn trong những năm sau đó.

II. Thảo luận phương pháp

Phương pháp được áp dụng để dự đoán điểm SDG Index Score của Việt Nam trong tương lai có một số ưu điểm và thách thức cần được bàn tới.

Trước hết, nói về ưu điểm, việc huấn luyện mô hình trên dữ liệu toàn cầu từ năm 2000 đến 2020 cho phép mô hình học được những quy luật tổng quát hơn, tránh tình trạng quá khớp với chỉ 25 điểm dữ liệu riêng biệt của Việt Nam. Điều này giúp cải thiện khả năng khái quát hóa của mô hình và nâng cao độ chính xác khi áp dụng vào dự báo cho Việt Nam. Ngoài ra, mô hình còn khai thác thông tin từ các biến ngoại sinh – tức là các yếu tố được xác định là có ảnh hưởng đến SDG Index Score – từ đó phản ánh được mối liên hệ phức tạp giữa các yếu tố phát triển bền vững, thay vì chỉ dựa vào xu hướng nội tại của SDG Index qua thời gian.

Tuy nhiên, quá trình dự báo cũng đối mặt với những thách thức nhất định. Một trong những rủi ro lớn nhất nằm ở việc dự báo trước các biến ngoại sinh cho giai đoạn 2025–2030. Chất lượng của các biến đầu vào này ảnh hưởng trực tiếp đến độ chính xác của dự báo SDG Index Score, và mọi sai số trong quá trình dự báo biến ngoại sinh đều có khả năng lan truyền sang các bước tiếp theo. Khi dự báo theo chuỗi thời gian, sai số có xu hướng tích lũy: giá trị dự đoán cho năm 2026 sẽ phụ thuộc vào kết quả của năm 2025, và vì vậy, càng dự báo xa về tương lai, mức độ không chắc chắn sẽ càng cao.

Bên cạnh đó, mô hình còn dựa trên giả định rằng mối quan hệ giữa các biến – được học từ dữ liệu toàn cầu trong giai đoạn 2000–2020 – sẽ tiếp tục giữ nguyên trong bối cảnh của Việt Nam từ năm 2025 đến 2030. Tuy nhiên, nếu có những biến động lớn về chính sách, môi trường hay các yếu tố xã hội, cấu trúc này có thể thay đổi, khiến cho độ chính xác của mô hình bị suy giảm.

Tóm lại, mặc dù phương pháp mang lại nhiều ưu điểm về khả năng khai thác dữ liệu và tính linh hoạt, người dùng vẫn cần cẩn trọng với các sai số tích lũy và rủi ro đến từ những giả định mô hình trong giai đoạn dự báo.

III. Đề xuất giải pháp chiến lược cho tương lai

1. Giảm nghèo và nâng cao thu nhập cho người dân

Dựa trên dữ liệu từ năm 2025 đến 2030, các chỉ số liên quan đến xóa đói giảm nghèo như sdg1_wpc và sdg1_lmicpov cho thấy xu hướng cải thiện rõ rệt, với tỷ lệ nghèo trong nhóm thu nhập thấp ngày càng giảm. Đây là một tín hiệu tích cực, phản ánh hiệu quả của các chính sách an sinh xã hội và phát triển kinh tế địa phương. Tuy nhiên, để tiếp tục duy trì và nâng cao kết quả này, Việt Nam cần ưu tiên mở rộng các chương trình hỗ trợ sinh kế bền vững, đặc biệt ở khu vực miền núi, vùng sâu và hải đảo – nơi mà tỷ lệ nghèo vẫn còn cao. Các chương trình đào tạo kỹ năng nghề nghiệp, kỹ năng số và tiếp cận tín dụng vi mô nên được triển khai rộng rãi nhằm khuyến khích khởi nghiệp, tạo thêm việc làm cho người dân. Việc kết hợp giữa hỗ trợ tài chính và đào tạo kỹ thuật sẽ giúp người dân thoát nghèo một cách bền vững, hướng tới một xã hội không ai bị bỏ lại phía sau.

2. Cải thiện dinh dưỡng và chăm sóc sức khỏe

Các chỉ số dinh dưỡng trong SDG2 như tỷ lệ suy dinh dưỡng thể nhẹ cân (sdg2_undernsh) và thấp còi (sdg2_stunting) đang giảm mạnh qua từng năm, cho thấy chất lượng dinh dưỡng của trẻ em Việt Nam đã được cải thiện đáng kể. Tuy nhiên, một vấn đề mới cũng đang hình thành khi tỷ lệ béo phì ở trẻ em (sdg2_obesity) và chỉ số dinh dưỡng thừa (sdg2_trophic) có xu hướng tăng. Điều này cho thấy Việt Nam đang đối mặt với gánh nặng kép về dinh dưỡng. Để giải quyết tình trạng này, cần đẩy mạnh các chương trình giáo dục dinh dưỡng học đường, xây dựng bữa ăn học đường hợp lý, đồng thời nâng cao nhận thức cộng đồng về lựa chọn thực phẩm và thói quen ăn uống lành mạnh.

Bên cạnh đó, các chỉ số y tế như tỷ lệ tử vong mẹ (sdg3_matmort), tử vong trẻ sơ sinh (sdg3_neonat) và tử vong trẻ dưới 5 tuổi (sdg3_u5mort) đều cho thấy sự cải thiện rõ rệt. Tuy nhiên, để giảm sâu và bền vững hơn các chỉ số này, cần đầu tư mạnh mẽ hơn cho hệ thống y tế cơ sở, nâng cao năng lực của đội ngũ y tế địa phương, đảm bảo người dân – đặc biệt là phụ nữ mang thai và trẻ nhỏ – được tiếp cận dịch vụ chăm sóc y tế đầy đủ, kịp thời và chất lượng.

3. Đổi mới giáo dục và thúc đẩy bình đẳng

Mặc dù không được phản ánh rõ qua các chỉ số trong bảng dữ liệu, nhưng vai trò của giáo dục (SDG4) và bình đẳng giới (SDG5) là nền tảng để đạt được các mục tiêu phát triển bền vững khác. Đặc biệt, khi nền kinh tế chuyển mình theo hướng số hóa và

xanh hóa, giáo dục nghề nghiệp gắn với thị trường lao động cần được tăng cường. Việc tích hợp giáo dục kỹ thuật số, đào tạo STEM và khuyến khích nữ giới tham gia các lĩnh vực công nghệ cao là chìa khóa để xây dựng nguồn nhân lực chất lượng cao. Đồng thời, hệ thống giáo dục cần đảm bảo tiếp cận công bằng cho các nhóm yếu thế như trẻ em dân tộc thiểu số, trẻ khuyết tật hay trẻ em vùng khó khăn.

4. Thúc đẩy sản xuất bền vững và giảm phát thải

Dữ liệu từ sdg12_nprod và sdg13_ghgimport cho thấy lượng phát thải khí nhà kính gián tiếp từ hàng hóa nhập khẩu và sản xuất trong nước đang có xu hướng tăng nhẹ. Điều này phản ánh sự phụ thuộc ngày càng lớn vào các sản phẩm và chuỗi cung ứng có phát thải cao, gây áp lực lên mục tiêu giảm biến đổi khí hậu. Trước tình hình đó, Việt Nam cần đẩy mạnh chuyển đổi mô hình tăng trưởng từ chiều rộng sang chiều sâu thông qua sản xuất xanh, sử dụng năng lượng tái tạo, và khuyến khích nền kinh tế tuần hoàn. Bên cạnh đó, chính sách kiểm soát phát thải như áp dụng thuế carbon, hoặc tiếp cận cơ chế CBAM theo xu hướng toàn cầu sẽ giúp Việt Nam kiểm soát tốt hơn lượng phát thải phát sinh từ hoạt động xuất nhập khẩu.

5. Củng cố hệ thống pháp luật, công lý và quản trị minh bạch

Các chỉ số liên quan đến SDG16 như sdg16_justice, sdg16_cpi và sdg16_detain phản ánh xu hướng tích cực về cải thiện công lý, chống tham nhũng và giảm giam giữ tùy tiện. Tuy nhiên, để tăng cường lòng tin của người dân vào hệ thống pháp lý, Việt Nam cần tiếp tục đẩy mạnh cải cách tư pháp, hiện đại hóa hệ thống pháp luật, đặc biệt là thông qua chuyển đổi số trong quản lý công lý và hành chính công. Việc công khai hóa dữ liệu các phiên tòa, tăng cường giám sát xã hội và vai trò của báo chí cũng là những giải pháp hữu hiệu để nâng cao tính minh bạch và trách nhiệm giải trình của cơ quan nhà nước.

6. Nâng cao hiệu quả đầu tư công và quản trị thể chế

Các chỉ số sdg17_govex và sdg17_statperf ghi nhận xu hướng gia tăng, phản ánh sự cải thiện trong năng lực thể chế và chi tiêu công cho phát triển bền vững. Điều này cho thấy Việt Nam đang có những bước đi đúng hướng trong quản lý tài khóa và phát triển hệ thống thống kê quốc gia. Trong thời gian tới, cần tiếp tục đẩy mạnh số hóa hoạt động quản lý ngân sách, đầu tư công và hệ thống đấu thầu. Ngoài ra, phát triển hệ sinh thái dữ liệu mở và ứng dụng dữ liệu lớn (Big Data) vào hoạch định chính sách sẽ giúp chính phủ ra quyết định nhanh chóng, chính xác và hiệu quả hơn.

7. Đánh giá tổng thể và khuyến nghị

Việt Nam hiện tại đang phải đối mặt với một giai đoạn điều chỉnh ngắn hạn trong bối cảnh toàn cầu đang chuyển dịch mạnh mẽ sau đại dịch, Việt Nam cần tận dụng các cơ hội từ chuyển đổi số, chuyển đổi xanh, và hội nhập kinh tế quốc tế để nâng cao vị thế phát triển bền vững. Giai đoạn 2025-2030 là một thời kỳ quan trọng để Việt Nam

củng cố nền tảng phát triển bền vững, vượt qua những biến động ngắn hạn. Bức tranh phục hồi giai đoạn 2025-2030 là hoàn toàn khả thi nếu Việt Nam có những điều chỉnh kịp thời và thực hiện các chiến lược được nêu một cách hiệu quả.

TÀI LIỆU THAM KHẢO

[1] United Nations. (n.d.). ***The 17 goals*****. United Nations Department of Economic and Social Affairs, Sustainable Development. Retrieved April 9, 2025, from** https://sdgs.un.org/goals#history

[2] United Nations Statistics Division. (2021). ***Handbook on the use of SDG indicators in monitoring in the context of the 2030 Agenda for Sustainable Development*** https://unstats.un.org/wiki/spaces/SDGeHandbook/pages/34505092/Home?preview=%2F34505092%2F106497383%2FSDGeHandbook-111121-2121-805.pdf

[3] World Bank income groups, 2023. https://ourworldindata.org/grapher/world-bank-income-groups

[4] Family Planning Indicators, 2024. https://www.un.org/development/desa/pd/data/family-planning-indicators

[5] The World University Rankings 2011-2024. https://www.kaggle.com/datasets/r1chardson/the-world-university-rankings-2011-2023?utm_source=chatgpt.com&select=2024_rankings.csv

[6] Thư viện Pháp luật. (2023). ***Quyết định 841/QĐ-TTg năm 2023 về lộ trình thực hiện các mục tiêu phát triển bền vững Việt Nam đến năm 2030*****. Chính phủ Việt Nam. Retrieved April 25, 2025, from** https://thuvienphapluat.vn/van-ban/Bo-may-hanh-chinh/Quyet-dinh-841-QD-TTg-2023-Lo-trinh-thuc-hien-cac-muc-tieu-phat-trien-ben-vung-Viet-Nam-den-2030-572610.aspx

[7] UNICEF. (2024). ***Vietnam: Country Office Annual Report 2023*****. Retrieved April 25, 2025, from** https://www.unicef.org/media/152206/file/Vietnam-2023-COAR.pdf

[8] UNESCO Institute for Statistics. (2023). ***Education data release*****. Retrieved April 25, 2025, from** https://uis.unesco.org/en/news/education-data-release

[9] Singapore Green Finance Centre. (2023). ***Vietnam renewables investment priorities*****. Retrieved April 25, 2025, from** https://www.singaporegreenfinance.com/wp-content/uploads/2023/12/Vietnam-Renewables-Investment-Priorities.pdf

[10] Macrotrends. (n.d.). ***Vietnam youth unemployment rate 1991–2024*****. Retrieved April 25, 2025, from** https://www.macrotrends.net/global-metrics/countries/VNM/vietnam/youth-unemployment-rate

[11] Vietnam News. (2024, March 15). ***Mekong Delta faces severe drought, salinity concerns*****. Retrieved April 25, 2025, from**

https://vietnamnews.vn/environment/1656592/mekong-delta-faces-severe-drought-salinity-concerns.html

[12] ASEAN. (2024). ***ASEAN 2045: Shaping a green, connected and sustainable tomorrow (Issues 34–35)*****. ASEAN Secretariat. Retrieved April 25, 2025, from** https://asean.org/wp-content/uploads/2024/02/Issue-34-35-Issue-34-35-ASEAN-2045_-Shaping-a-Green-Connected-and-Sustainable-Tomorrow.pdf

PHỤ LỤC

Các file mã nguồn:

Python:  Nhom8_LamTuanThinh_Python.zip

R:  Nhom8_LamTuanThinh_R.zip