ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

PHÂN TÍCH CÁC YẾU TỐ MÔI TRƯỜNG ẢNH HƯỞNG ĐẾN THIÊN TAI VÀ XÂY DỰNG MÔ HÌNH DỰ ĐOÁN THIÊN TAI Ở CHÂU Á

Nhóm 10			
Sinh viên thực hiện:			
STT	Họ tên	MSSV	Ngành
1	Hồ Ngọc Mai	22520839	KHDL
2	Lê Ngọc Thiên Phúc	22521117	KHDL
3	Nguyễn Thành Đạt	22520224	KHDL

TP. HỒ CHÍ MINH – 12/2024

1. GIỚI THIỆU

Đề tài "Phân tích các yếu tố môi trường ảnh hưởng đến thiên tai và xây dựng mô hình dự đoán thiên tai ở Châu Á" được chúng tôi xây dựng nhằm mục tiêu phân tích các yếu tố môi trường như nhiệt độ bề mặt, trữ lượng carbon trong rừng,… tác động như thế nào đến sự xuất hiện cũng như tần suất của các thiên tai tại các quốc gia Châu Á. Đề tài của chúng tôi sử dụng các thư viện hỗ trợ xử lý dữ liệu chính là Pandas, Numpy, Seaborn và Matplotlib để trực quan hóa dữ liệu và mô hình dự đoán Linear Regression, LGBM, Catboost để xây dựng một hệ thống có khả năng dự đoán các thiên tai xảy ra. Các mô hình như Linear Regression, Catboost được áp dụng để tối ưu hóa độ chính xác của dự đoán. Kết quả của đề tài cho thấy các mô hình này có khả năng dự đoán thiên tai với độ đo R2 là 0.45.

Bộ dữ liệu phân tích trong đề tài này do chúng tôi tự thu thập từ trang web Climate Change Indicators Dashboard [1]. Đây là tập hợp dữ liệu bao gồm các thông tin về khí hậu, và các yếu tố môi trường liên quan. Bộ dữ liệu và đề tài do nhóm tự phân tích thiết kế và không dựa trên đề tài có sẵn nào khác. Bộ dữ liệu chỉ phục vụ riêng cho môn học DS105 – Phân tích và trực quan hóa dữ liệu. Mọi dữ liệu và kết quả phân tích đều được công khai minh bạch và có thể được xác thực từ các nguồn tham khảo được ghi rõ trong tài liệu.

2. MÔ TẢ BỘ DỮ LỆU

Bộ dữ liệu mà nhóm sử dụng được thu thập từ trang web Climate Change Indicators Dashboard [1]. Đây là tập hợp dữ liệu đã được thu thập và chuẩn hóa, bao gồm nhiều thuộc tính đặc trưng cho các khía cạnh khác nhau của biến đổi khí hậu, môi trường cung cấp một cái nhìn toàn diện về tác động của các yếu tố này đến tần suất và cường độ của các thiên tai ở các quốc gia từ năm 1980 đến năm 2022.

Bộ dữ liệu này được chuẩn hóa và tổng hợp từ nhiều nguồn nhỏ với các định dạng khác nhau, phục vụ cho việc xây dựng mô hình và đánh giá trong quá trình nghiên cứu. Toàn bộ đề tài và dữ liệu được nhóm tự phân tích và thiết kế dựa trên các bộ dữ liệu gốc. Bộ dữ liệu chỉ dành riêng cho môn học Phân tích và Trực quan Dữ liệu (DS105).

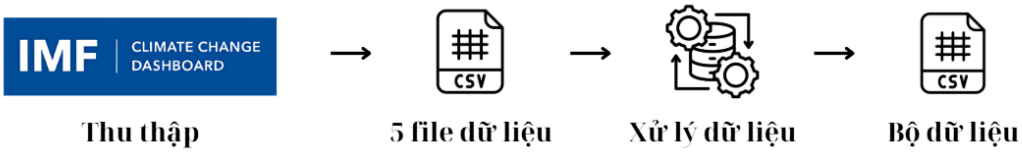
Dữ liệu thô sau khi phân tích và thu thập từ Quỹ Tiền Tệ Quốc Tế (IMF) đã được nhóm sử dụng. IMF cho phép sử dụng dữ liệu này cho mục đích học tập và nghiên cứu, với yêu cầu trích dẫn nguồn. Dữ liệu được nhóm tự phân tích, chắt lọc và phát triển ý tưởng. Vì IMF thu thập dữ liệu từ nhiều đơn vị, cơ quan khác nhau và từ chính IMF, nên bộ dữ liệu có một số giá trị bị thiếu ở các năm và các quốc gia khác nhau. Chi tiết các thuộc tính của dữ liệu được trình bày trong **Bảng 1**.

***Bảng 1**: Mô tả chi tiết các thuộc tính trong bộ dữ liệu.*

SST	Tên thuộc tính	Kiểu dữ liệu	Nội dung thuộc tính
1	Country	Text	Các quốc gia trên thế giới.
2	Measure	Text	Các biển trên thế giới

3	ISO3	Text	Viết tắt tên của các quốc gia trên thế giới
4	Indicator	Text	Các chỉ số về các yếu tố môi trường. Chi tiết được trình bày ở **Bảng 2** phụ lục.
5	Unit	Text	Đơn vị đo lường.
6	2018 → 2022	Numeric	Giá trị của Indicator tương ứng với các năm.

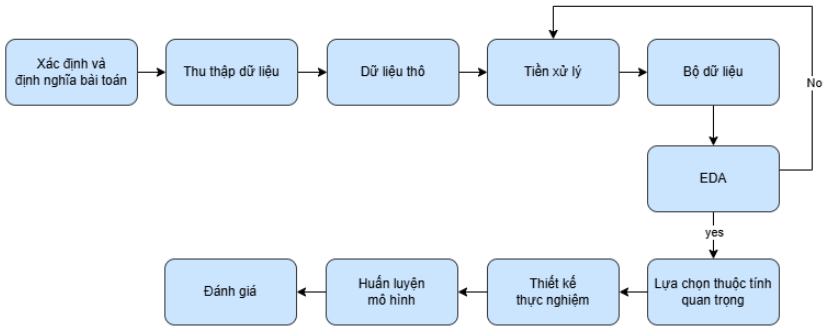
Bộ dữ liệu có 10 thuộc tính (5 thuộc tính văn bản và 5 thuộc tính số) và có 567 dòng dữ liệu. Dữ liệu của nhóm thu thập tại website Climate Change Indicators Dashboard – một bảng điều khiển các chỉ số biến đổi khí hậu của IMF, đây là một cơ quan, tổ chức quốc tế uy tín, dữ liệu hợp tác phát triển bởi IMF và các cơ quan uy tín khác trên thế giới (OECD, WBG, UN, Eurostat, …) vì thế số liệu tại website này là uy tín và đáng tin cậy. Ngoài ra dữ liệu của trang được công khai và được phép sử dụng nhằm mục đích học tập và nghiên cứu với điều khoản về nguồn trích dẫn. Nhóm đã chọn lọc và tải xuống các file .csv được cung cấp trên website. Sau quá trình phân tích và chắt lọc, nhóm đã thu thập được 5 bộ dữ liệu tương ứng với 5 chỉ số chính. Cuối cùng nhóm phân tích và đưa ra cấu trúc chung cho dữ liệu sau đó tiến hành xử lý và gộp các bộ dữ liệu lại cho ra bộ dữ liệu hoàn chỉnh ở dạng csv. **Hình 1** mô tả toàn bộ quá trình thu thập, xử lý dữ liệu để có được bộ dữ liệu thô cuối cùng. Các điểm dữ liệu mẫu được trình bày chi tiết và đầy đủ tại **Bảng 3** phụ lục.



Hình 1: *Quy trình thu thập và xây dựng bộ dữ liệu.*

3. PHƯƠNG PHÁP PHÂN TÍCH

3.1 Quy trình thực hiện



Hình 2: *Toàn bộ quy trình thực hiện.*

3.2 Tiền xử lý dữ liệu

❖ *Tiền xử lý cho phân tích thăm dò*

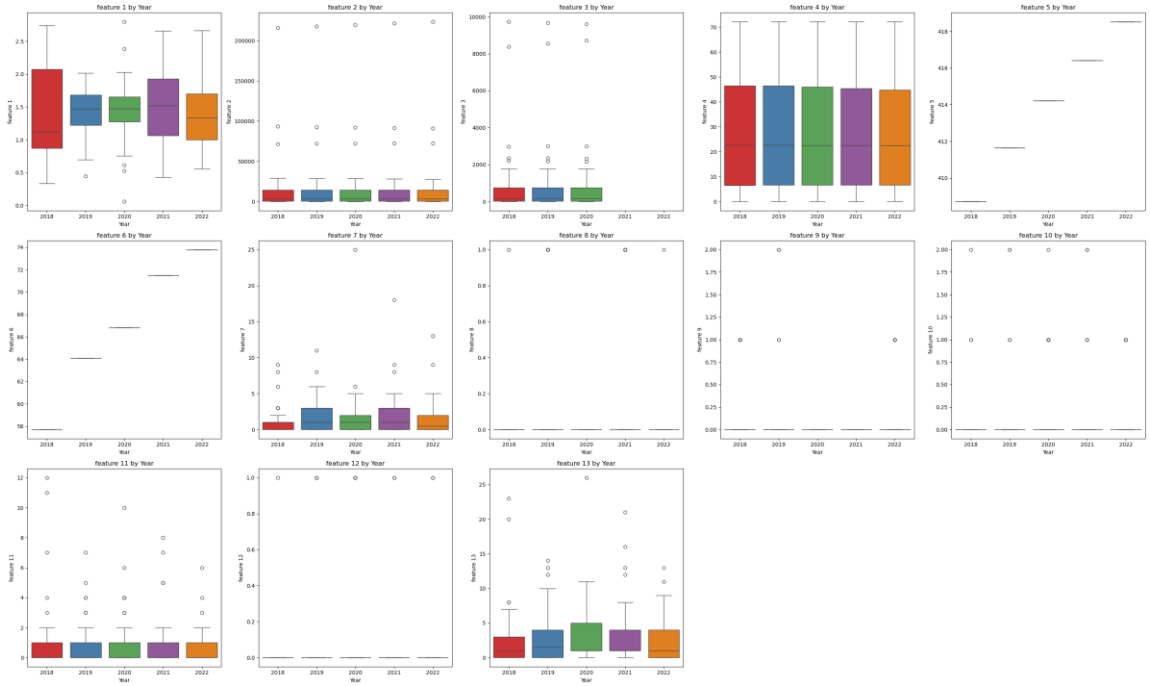
Trong quá trình phân tích dữ liệu ban đầu, chúng tôi nhận thấy bộ dữ liệu chứa hơn 200 quốc gia và nhiều năm có dữ liệu khuyết, đặc biệt là giai đoạn trước năm 2018. Để đảm bảo tính đầy đủ và chất lượng của dữ liệu, chúng tôi quyết định thực hiện các bước tiền xử lý để giữ lại dữ liệu của các quốc gia thuộc khu vực Châu Á, giới hạn thời gian phân tích từ năm 2018 đến năm 2022. Để đảm bảo tính nhất quán giữa các bộ dữ liệu chúng tôi lọc và thống nhất tên quốc gia thông qua chỉ số ISO3. Trong quá trình tìm hiểu dựa trên các chỉ số cần thiết để phân tích chuyên sâu, chúng tôi loại đi các quốc gia không có đủ các chỉ số yêu cầu. Sau khi tiền xử lý, chúng tôi thu được bộ dữ liệu gồm 42 nước, 13 chỉ số và dữ liệu từ năm 2018 đến 2022.

❖ *Tiền xử lý cho dữ liệu huấn luyện*

Trong quá trình chuẩn bị dữ liệu huấn luyện, chúng tôi đã thực hiện các bước tiền xử lý nhằm đảm bảo tính đầy đủ và cấu trúc phù hợp. Đầu tiên, chúng tôi loại bỏ các dữ liệu không cần thiết, cụ thể là các dòng dữ liệu tổng toàn cầu để tập trung vào dữ liệu của từng quốc gia. Dữ liệu sau đó được chuyển đổi từ dạng rộng sang dạng dài, tiếp theo, dữ liệu được pivot lại sao cho mỗi chỉ số (Indicator) trở thành một cột riêng biệt, với chỉ số được sắp xếp theo từng quốc gia và từng năm. Kết quả là một bộ dữ liệu có cấu trúc rõ ràng, trong đó các hàng tương ứng với từng quốc gia và năm, còn các cột thể hiện các chỉ số cụ thể. Cuối cùng, dữ liệu đã được làm sạch và chuẩn hóa này được xuất ra tệp CSV với tên **data_for_model.csv** để phục vụ cho quá trình huấn luyện.

3.3 Tìm đặc trưng dữ liệu

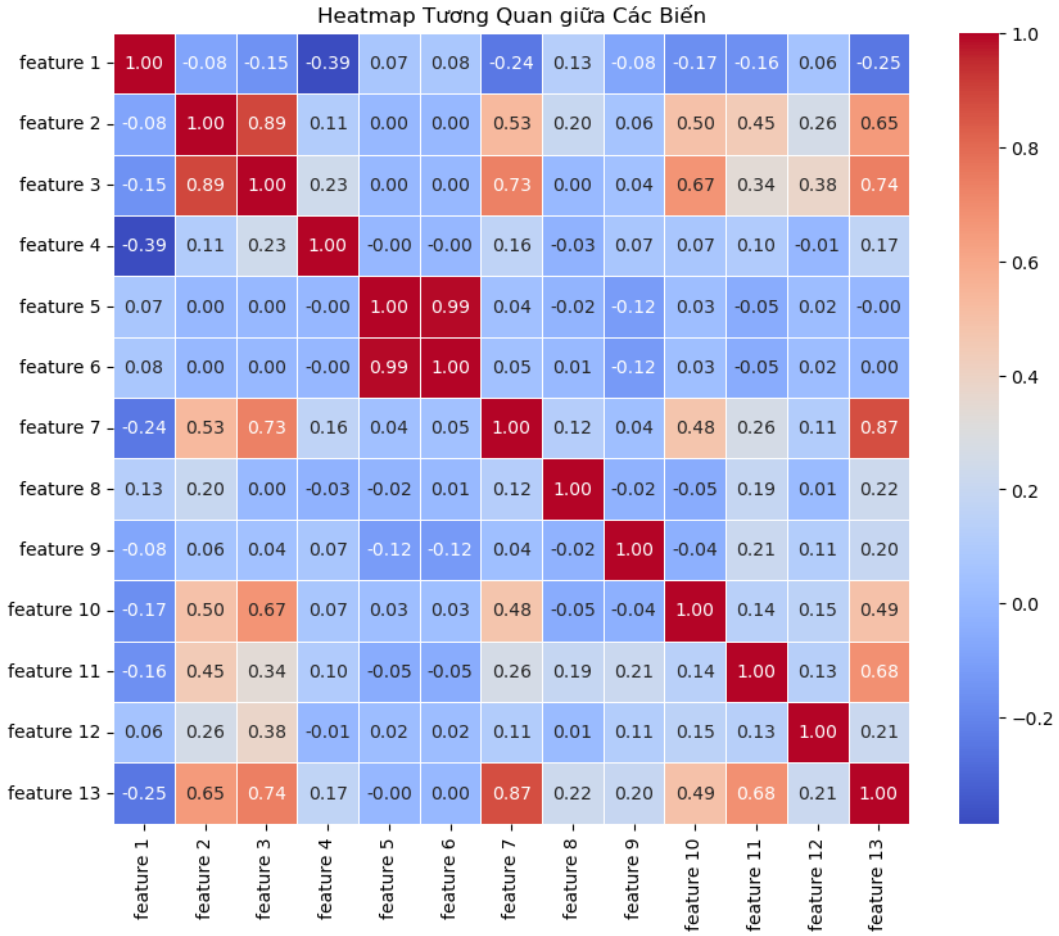
❖ *Thống kê mô tả từng thuộc tính*



Hình 3: *Thống kê mô tả của từng thuộc tính*

Phân tích các boxplot cho thấy sự đa dạng trong phân bố của các thuộc tính theo thời gian. Đáng chú ý, các thuộc tính 1, 2, 3, 4, 5, 6 (chi tiết các thuộc tính được trình bài ở **Bảng 1** phụ lục) thường có phân bố không đối xứng với sự xuất hiện của các giá trị ngoại lệ, cho thấy có thể có một số yếu tố tác động mạnh đến một số quan sát. Ngược lại, các thuộc tính 7, 8, 9, 10, 11, 12, 13 có xu hướng phân bố đối xứng hơn. Sự biến động của giá trị trung bình và phân tán qua các năm cho thấy các yếu tố bên ngoài có thể đã ảnh hưởng đến các thuộc tính này. Nhìn chung, các boxplot cho thấy rõ ràng những tác động của biến đổi khí hậu đến hệ sinh thái, môi trường và cuộc sống của con người. Sự gia tăng nhiệt độ, tăng mực nước biển và gia tăng tần suất các hiện tượng cực đoan là những bằng chứng rõ ràng cho thấy tình hình khí hậu đang thay đổi nhanh chóng.

❖ *Độ tương quan giữa các thuộc tính*



Hình 4: *Biểu đồ tương quan giữa các thuộc tính*

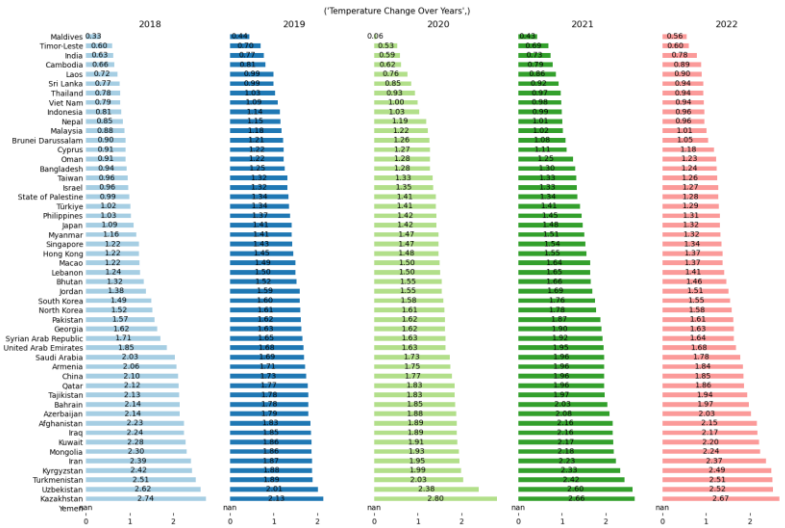
Chúng tôi quy ước thang đo của độ tương quan: không tương quan: 0 → 0.3, yếu: 0.3 → 0.5, vừa: 0.5 → 0.7, mạnh: 0.7 → 1. Nhìn chung, biểu đồ tương quan cho thấy một cấu trúc tương quan phức tạp giữa các biến. Sự tồn tại của các nhóm biến có mối quan hệ chặt chẽ và sự đa dạng trong mối quan hệ giữa các chỉ số cung cấp những thông tin quan trọng cho việc xây dựng mô hình và phân tích dữ liệu. Mối quan hệ giữa các nhóm chỉ số khác nhau nhìn chung ở mức thấp hoặc trung bình. Điều này cho thấy các

chỉ số này đo lường những khía cạnh khác nhau của dữ liệu và có thể cung cấp thông tin bổ sung cho nhau. Các chỉ số riêng lẻ cũng thể hiện sự đa dạng trong mối liên hệ với các nhóm chỉ số còn lại, cho thấy sự phức tạp của cấu trúc dữ liệu. Tuy nhiên, mức độ tương quan cao giữa các biến trong cùng một nhóm chỉ số cũng đặt ra vấn đề về đa cộng tuyến, có thể ảnh hưởng đến độ tin cậy của các mô hình thống kê.

4. PHÂN TÍCH THĂM DÒ CHUYÊN SÂU

4.1. Phân tích các chỉ số của từng quốc gia

4.1.1. Sự thay đổi nhiệt độ bề mặt



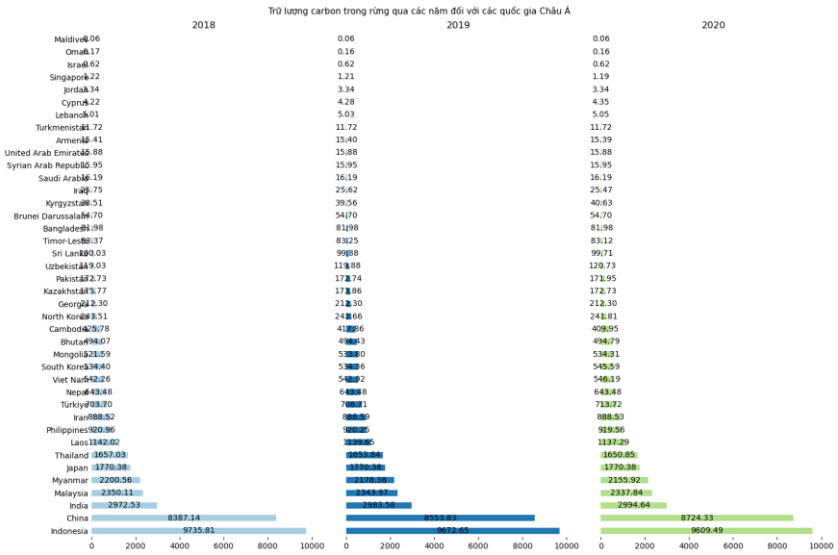
Hình 5: *Biểu đồ so sánh sự thay đổi nhiệt độ bề mặt giữa các quốc gia Châu Á*

Biểu đồ thể hiện tỷ suất thay đổi nhiệt độ qua các năm từ 2018 đến 2022 tại các quốc gia châu Á. Trong đó, sự chênh lệch rõ rệt giữa quốc gia có mức thay đổi nhiệt độ cao nhất và thấp nhất phản ánh tình trạng bất bình đẳng trong điều kiện môi trường và khả năng thích ứng. Cụ thể, Kazakhstan liên tục ghi nhận mức nhiệt độ thay đổi cao nhất qua các năm (2.74 vào 2018 và tăng đến 2.67 vào 2022), trong khi Maldives có mức thay đổi nhiệt độ thấp nhất (chỉ 0.33 vào năm 2018 và 0.56 vào năm 2022). Sự gia tăng này có thể liên quan đến các yếu tố như hoạt động công nghiệp, nạn phá rừng và ảnh hưởng của biến đổi khí hậu lên từng khu vực. Năm 2021 chứng kiến sự gia tăng nhiệt độ rõ rệt tại các quốc gia như Kazakhstan, Turkmenistan và Uzbekistan với mức thay đổi xấp xỉ từ 2.4 trở lên. Điều này có thể liên quan đến các hiện tượng thời tiết cực đoan, và sự nóng lên toàn cầu. Sự khác biệt lớn về tỷ suất nhiệt độ thay đổi giữa các quốc gia như Maldives (thấp nhất) và các quốc gia Trung Á như Kazakhstan (cao nhất) có thể kéo theo sự phân hóa về nguy cơ thiên tai như hạn hán, lũ lụt có thể ảnh hưởng nghiêm trọng đến cộng đồng. Mặt khác, các quốc gia có nhiệt độ thay đổi thấp hơn như Maldives hoặc Timor-Leste có thể đang hưởng lợi từ các chính sách bảo vệ môi trường hoặc khả năng thích ứng tốt hơn với biến đổi khí hậu. Việc nghiên cứu sự gia tăng nhiệt độ qua từng năm là cơ sở quan trọng để xây dựng các mô hình dự đoán thiên tai, đồng thời giúp các chính phủ có chiến lược cải thiện điều kiện sống và đảm bảo hệ thống y tế ứng phó

hiệu quả với biến đổi khí hậu. Không chỉ vậy, điều này cho thấy cần phải tập trung vào việc triển khai các biện pháp cải thiện đồng bộ, tập trung vào các lĩnh vực chính như bảo vệ môi trường, năng lượng, và phát triển bền vững.

Tổng quan ta thấy rằng hầu hết các quốc gia trên thế giới đều có sự gia tăng nhiệt độ bề mặt qua từng năm. Điều này có thể giải thích vì diện tích rừng giảm dẫn đến khả năng hấp thụ CO2 của rừng giảm, do đó nồng độ CO2 trong không khí tăng dẫn đến sự nóng lên toàn cầu. Tuy nhiên, bên cạnh việc nhiệt độ bề mặt có xu hướng tăng qua các năm, một số quốc gia lại cho thấy nhiệt độ có xu hướng giảm nhẹ trong một số giai đoạn. Hiện tượng này có thể được giải thích bởi nhiều nguyên do. Một là sự thay đổi trong mức độ che phủ rừng và cây xanh ở các quốc gia, các quốc gia đẩy mạnh chính sách bảo vệ và phục hồi rừng, trồng cây xanh và cải thiện độ che phủ tự nhiên sẽ giúp hạ nhiệt độ bề mặt. Hai là việc giảm hoạt động công nghiệp và ô nhiễm cục bộ, một số quốc gia trong giai đoạn này có thể giảm thiểu hoạt động công nghiệp hoặc khai thác tài nguyên (do suy thoái kinh tế hoặc tác động của đại dịch COVID-19). Điều này làm giảm lượng khí thải nhà kính, bụi mịn và bức xạ nhiệt từ các khu công nghiệp. Các quốc gia có diện tích nhỏ như Maldives, Timor-Leste hoặc đảo quốc nhỏ khác dễ chịu tác động bởi các yếu tố khí hậu đại dương (làm mát từ biển). Điều này khiến nhiệt độ thay đổi ít hoặc giảm nhẹ theo thời gian. Tuy nhiên, hiện tượng nhiệt độ bề mặt giảm không có nghĩa là biến đổi khí hậu ngừng lại. Thay vào đó, nó phản ánh sự dao động cục bộ tạm thời do tác động của tự nhiên hoặc các chính sách can thiệp của con người. Tuy nhiên, xu hướng tổng thể vẫn cho thấy nhiệt độ toàn cầu tiếp tục tăng, và các hiện tượng giảm chỉ là dấu hiệu mang tính ngắn hạn.

4.1.2. Trữ lượng carbon trong rừng



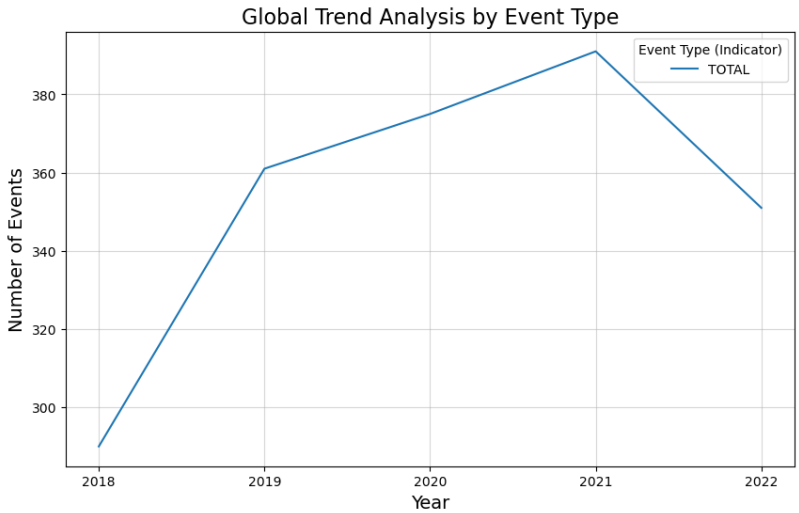
Hình 6: *Biểu đồ so sánh trữ lượng carbon trong rừng giữa các quốc gia Châu Á*

Dữ liệu biểu đồ thể hiện trữ lượng carbon trong rừng của các quốc gia ở Châu Á từ năm 2018 đến 2020. Trữ lượng carbon có sự gia tăng qua các năm ở một số quốc gia,

đặc biệt là các quốc gia có diện tích rừng lớn. Một số quốc gia duy trì trữ lượng carbon ổn định, trong khi một số khác có xu hướng giảm nhẹ. Indonesia, Trung Quốc và Ấn Độ luôn đứng đầu với mức trữ lượng carbon lớn nhất. Indonesia: Từ 9735.81 (2018) lên 9609.49 (2020), giảm nhẹ trong khoảng thời gian này. Trung Quốc: Tăng từ 8387.14 (2018) lên 8724.33 (2020), cho thấy sự cải thiện trong bảo tồn rừng. Ấn Độ: Có sự tăng nhẹ từ 2972.53 (2018) lên 2994.64 (2020). Đây là những quốc gia có rừng nhiệt đới và diện tích rừng lớn. Bên cạnh đó, một số quốc gia như Maldives, Oman, Israel, Singapore có trữ lượng carbon rất thấp, chỉ ở mức từ 0.06 đến 1.2. Điều này có thể do diện tích rừng hạn chế hoặc địa hình đặc biệt như đảo quốc hoặc sa mạc. Bên cạnh đó, qua biểu đồ cũng có thể thấy được sự thay đổi đáng chú ý đến từ các quốc gia. Campuchia: Tăng từ 25.78 (2018) lên 409.95 (2020), cho thấy nỗ lực bảo tồn rừng đã cải thiện đáng kể. Bhutan và Nepal cũng có sự tăng nhẹ trong giai đoạn này. Các quốc gia như Iran và Philippines có trữ lượng tương đối ổn định nhưng không tăng trưởng nhiều.

Đây là kết quả của chính sách bảo tồn và phục hồi rừng của các quốc gia như Trung Quốc và Campuchia cũng như trồng rừng, tái tạo rừng và các chương trình giảm phát thải carbon. Bên cạnh những ảnh hưởng tích cực, nạn phá rừng để phát triển nông nghiệp và đô thị hóa ở một số quốc gia như Indonesia và Malaysia cũng như vấn đề khai thác rừng và các hoạt động công nghiệp làm giảm trữ lượng carbon diễn ra. Đây là nguyên do khiến các ảnh hưởng tiêu cực đến trữ lượng carbon trong rừng xảy ra. Chính vì lý do đó, các quốc gia cần tăng cường chính sách bảo tồn rừng, đặc biệt là ngăn chặn nạn phá rừng, phát triển các chương trình tái trồng rừng và bảo vệ các khu rừng nguyên sinh, cũng như áp dụng công nghệ và giám sát để quản lý rừng hiệu quả hơn.

4.1.3. Tần suất thiên tai



Hình 7: *Biểu đồ xu hướng tần suất thiên tai ở Châu Á*

Cụ thể, biểu đồ thể hiện xu hướng thiên tai toàn cầu từ năm 2018 đến 2022, cụ thể là tổng số lượng thiên tai được ghi nhận trong từng năm. Trong giai đoạn 2018 – 2019,

số lượng thiên tai tăng từ khoảng 512 lên 665, cho thấy sự gia tăng đáng kể. Đây là kết quả của việc hiện tượng toàn cầu xảy ra với tần suất nhiều hơn, và dữ liệu thu thập đầy đủ hơn trong giai đoạn này. Đó là kết quả của các tác động của biến đổi khí hậu và việc quản lý tài nguyên thiên nhiên không bền vững làm gia tăng các hiện tượng cực đoan như bão, lũ lụt, hạn hán, cháy rừng và nhiệt độ cao kỷ lục. Chính những điều trên có thể khiến số lượng thiên tai bất thường tăng lên.

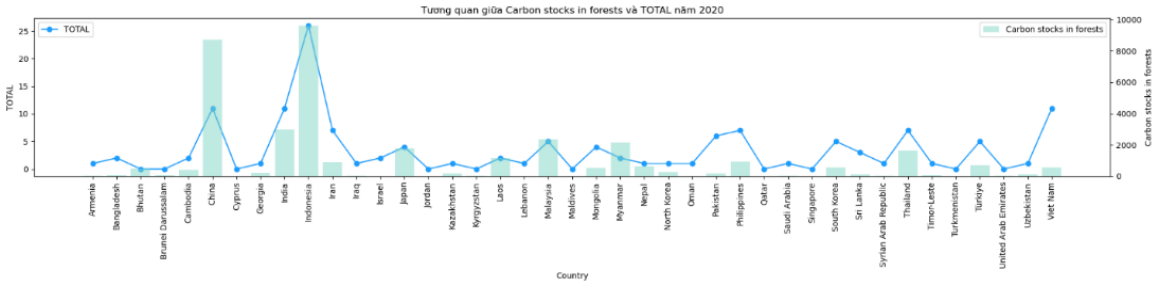
Trong giai đoạn năm 2021-2022, số lượng thiên tai giảm nhẹ xuống còn 555, đánh dấu sự sụt giảm đáng kể so với năm 2021. Có thể đây là kết quả của việc sau năm 2021, tình hình đại dịch được kiểm soát tốt hơn. Sự giảm nhẹ số lượng thiên tai cũng có thể do các yếu tố như giới hạn trong việc thu thập, phân loại hoặc báo cáo dữ liệu, dẫn đến thiếu sót trong thống kê.

Biểu đồ thể hiện xu hướng gia tăng số lượng sự kiện trong giai đoạn 2018 - 2021, chủ yếu do biến đổi khí hậu và tiến bộ trong giám sát dữ liệu. Sau đó giảm nhẹ vào năm 2022, có thể liên quan đến giảm thiểu thiên tai cực đoan và sự ổn định trở lại của kinh tế - xã hội toàn cầu. Điều này cho thấy sự biến động của các sự kiện toàn cầu, trong đó một số yếu tố như thiên tai, hoặc biến đổi khí hậu có thể là nguyên nhân chính ảnh hưởng đến xu hướng này.

4.2. Phân tích các chỉ số theo thế giới

Từ năm 2018 đến 2022, nồng độ CO2 trong khí quyển tăng đều, nguyên nhân chính dẫn đến việc này là do diện tích rừng trên toàn cầu ngày càng giảm dẫn đến trữ lượng carbon của rừng cũng giảm theo, giảm sự hấp thụ CO2 trong khí quyển, dẫn đến việc nóng lên toàn cầu, cũng là nguyên nhân chính là cho băng tan dẫn đến mực nước biển trung bình trên toàn cầu ngày càng tăng. Những sự thay đổi về khí hậu này cũng dẫn đến các thiên tai xuất hiện nhiều hơn.

4.3. Phân tích biến mục tiêu cho mô hình



Hình 8: *Biểu đồ tương quan giữa trữ lượng carbon trong rừng và tần suất thiên tai của từng quốc gia Châu Á năm 2020*

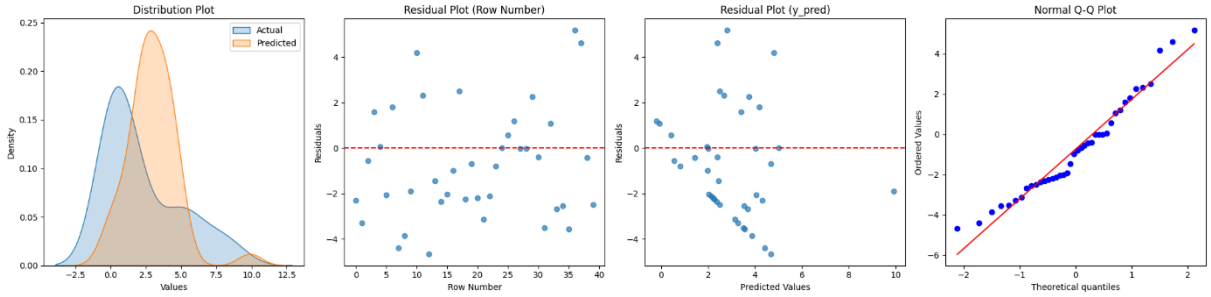
Các quốc gia có tổng số lượng thiên tai cao nhất ở Châu Á bao gồm Trung Quốc, Ấn Độ, Nhật Bản, Indonesia, Pakistan và đặc biệt là ở Việt Nam. Các quốc gia này đều có mật độ dân số cao và phát triển đô thị nhanh chóng ở các khu vực dễ bị thiên tai (gần sông, biển), làm gia tăng thiệt hại khi thiên tai xảy ra. Các yếu tố tự nhiên như vị trí địa lý, hoạt động địa chấn, điều kiện thời tiết cùng với tác động của con người (đô thị hóa, biến đổi khí hậu) là nguyên nhân khiến những quốc gia này thường xuyên đối mặt với

thiên tai nghiêm trọng. Chẳng hạn như, Trung Quốc, Ấn Độ, Pakistan, Việt Nam đều nằm gần hoặc trong khu vực vành đai nhiệt đới nơi chịu ảnh hưởng của các hiện tượng khí hậu, đồng thời có đồng bằng ven biển và hệ thống sông lớn dẫn đến dễ bị ngập lụt trong mùa mưa.

Qua biểu đồ hình trên, biểu đồ cho thấy mối tương quan giữa số lượng disasters và lượng carbon stocks in forests. Chúng tôi nhận định rằng các quốc gia có carbon stocks in forests cao thường có số lượng disasters cũng cao hơn so với các quốc gia còn lại. Điều này có thể do thiên tai liên quan đến rừng, các quốc gia có diện tích rừng lớn hơn thường dễ gặp các thảm họa liên quan đến rừng, như cháy rừng, lũ quét hoặc sạt lở đất. Bên cạnh đó là do rủi ro từ khí hậu nhiệt đới, nhiều quốc gia có diện tích rừng lớn nằm trong khu vực nhiệt đới, dễ bị ảnh hưởng bởi mưa bão, lũ lụt, và các hiện tượng thời tiết cực đoan. Mối tương quan giữa thiên tai và carbon stocks in forests cho thấy sự phức tạp trong mối quan hệ giữa thiên nhiên và thiên tai. Mặc dù rừng đóng vai trò quan trọng trong việc điều hòa khí hậu và giảm thiểu thiên tai, nhưng chúng cũng dễ bị ảnh hưởng bởi các loại thảm họa liên quan đến rừng như cháy rừng và lũ lụt.

5. KẾT QUẢ THÍ NGHIỆM

Có thể thấy Hình 8 các thiên tai xuất hiện ở các nước Châu Á chủ yếu là bão và lũ lụt. Do đó, chúng tôi sử dụng các thuộc tính Storm, Flood và Total làm biến mục tiêu để dự đoán tần suất xuất hiện thiên tai ở các nước Châu Á.



***Hình 9**: Đánh giá hiệu suất của mô hình Linear Regression với biến mục tiêu Total*

Mô hình	Biến mục tiêu Storm			Biến mục tiêu Flood			Biến mục tiêu Total		
	R2	MSE	MAE	R2	MSE	MAE	R2	MSE	MAE
Linear	-0.0598	3.1264	1.2393	-0.2090	2.2851	1.3100	0.1133	6.3728	2.1017
LGBM	**0.3647**	**1.8739**	**0.8904**	**-5.6659**	**12.598**	**2.0115**	**-0.0005**	**7.2277**	**1.9010**
CatBoost	**0.4035**	**1.7597**	**0.8330**	**0.3642**	**1.2016**	**0.7823**	**0.4567**	**3.9048**	**1.4195**

***Bảng 2:** Kết quả dự đoán của mô hình của 3 biến mục tiêu*

Chúng tôi sử dụng mô hình Linear Regression và 3 độ đo phổ biến cho bài toán hồi quy là: R2, Mean squared error, Mean absolute error cùng với Residual Plot để đánh hiệu

suất của mô hình. Ở cả 3 độ đo Linear Regression cho kết quả rất tệ và Residual Plot cho thấy dữ liệu không phù hợp với mô hình này. Chúng tôi sử dụng thêm các mô hình, Catboost cho kết quả tốt nhất ở cả 3 độ đo với 3 biến mục tiêu. Từ những kết quả trên, bộ dữ liệu còn thiếu những đặc trưng tốt để các mô hình có thể dự đoán tốt hơn các thiên tai.

6. KẾT LUẬN

Chúng tôi đã thực hiện một quá trình thu thập và tiền xử lý kỹ lưỡng, bắt đầu bằng việc chọn lọc và gộp các thuộc tính từ 5 file dữ liệu để tạo ra một bộ dữ liệu hoàn chỉnh. Trước khi bắt đầu phân tích chuyên sâu, chúng tôi đã phân tích cơ bản để thu hẹp phạm vi phân tích và thực hiện tiền xử lý để chuẩn hóa dữ liệu, xử lý giá trị bị khuyết và loại bỏ các thuộc tính không phù hợp. Trong quá trình huấn luyện mô hình, chúng tôi đã áp dụng các chiến lược như lựa chọn giá trị điền khuyết, loại bỏ biến có lượng dữ liệu khuyết lớn, và chọn các biến có tương quan cao. Bộ dữ liệu đã được chia thành hai tập train và test để đảm bảo tính khách quan của việc đánh giá mô hình. Với mục tiêu dự đoán thiên tai, chúng tôi đã chọn các mô hình như LGBM, CatBoost, Linear và tiến hành đánh giá sự hiệu quả của chúng bằng các độ đo phổ biến. Mô hình CatBoost đã đạt được kết quả tốt nhất với R2 đạt 0.4567. Mô hình Linear, mặc dù có kết quả thấp nhất, nhưng cũng đóng góp vào việc đánh giá toàn diện về hiệu suất của các mô hình. Tổng quan, quá trình nghiên cứu đã đem lại cái nhìn toàn diện về các yếu tố môi trường ảnh hưởng đến thiên tai ở Châu Á cũng như cung cấp những thông tin hữu ích cho việc đưa ra quyết định và chiến lược dự đoán và khắc phục thiên tai trong tương lai.

Các mô hình dự đoán của chúng tôi cũng trả về trọng số cho các thuộc tính. Trong tương lai chúng tôi sẽ tiến hành phân tích các trọng số này để tìm ra các thuộc tính quan trọng, đóng góp nhiều vào quá trình dự đoán thiên tai, đồng thời cũng tiếp tục khám phá và phân tích các thuộc tính tiềm năng có trong bộ dữ liệu.

PHỤ LỤC

STT	Các thuộc tính	Các thuộc tính tương ứng của bộ dữ liệu
1	feature 1	Temperature change with respect to a baseline climatology, corresponding to the period 1951-1980
2	feature 2	Forest area
3	feature 3	Carbon stocks in forests
4	feature 4	Share of forest area
5	feature 5	Yearly Atmospheric Carbon Dioxide Concentrations
6	feature 6	Change in mean sea level: Sea level
7	feature 7	Flood
8	feature 8	Drought
9	feature 9	Extreme temperature
10	feature 10	Landslide
11	feature 11	Storm
12	feature 12	Wildfire
13	feature 13	TOTAL

Bảng 2. Ánh xạ các thuộc tính có trong dữ liệu.

STT	Các thuộc tính Chính	Chỉ số yêu tố môi trường

1	Annual Surface Temperature Change	Temperature change with respect to baseline climatology, corresponding to the period 1951-1980
2	Atmospheric CO_2 Concentrations	Atmospheric Carbon Dioxide Concentrations
3	Change in Mean Sea Levels	Change in mean sea level: Sea level
4	Climate-related Disasters Frequency	Climate related disasters frequency, Number of Disasters: Drought
		Climate related disasters frequency, Number of Disasters: Extreme temperature
		Climate related disasters frequency, Number of Disasters: Flood
		Climate related disasters frequency, Number of Disasters: Landslide
		Climate related disasters frequency, Number of Disasters: Storm
		Climate related disasters frequency, Number of Disasters: TOTAL
		Climate related disasters frequency, Number of Disasters: Wildfire
5	Forest and Carbon	Carbon stocks in forests
		Forest area
		Index of carbon stocks in forests
		Index of forest extent

		Land area
		Share of forest area

Bảng 3. Phân tích tổng quan các bộ dữ liệu.

STT	Bộ dữ liệu	Số cột ban đầu	Số dòng ban đầu	Số biến phân loại	Số biến số	Số lượng giá trị khuyết
1	Annual Surface Temperature Change	72	225	9	63	Không nhiều
2	Atmospheric CO_2 Concentrations	12	1570	9	3	Không khuyết
3	Change in Mean Sea Levels	13	39617	10	3	Khá nhiều
4	Climate-related Disasters Frequency	51	975	6	45	Khá nhiều
5	Forest and Carbon	41	1333	7	34	Khá nhiều

TÀI LIỆU THAM KHẢO

[1] Climate Change Indicators Dashboard. [Trực tuyến]. Link: https://climatedata.imf.org/. [Truy cập lần cuối 17/12/2024]

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Hồ Ngọc Mai	- Thu thập dữ liệu - Đề xuất phương pháp - Xử lý dữ liệu lần 2 - Khám phá dữ liệu - Viết báo cáo - Thuyết trình
2	Lê Ngọc Thiên Phúc	- Thu thập dữ liệu - Đề xuất phương pháp - Xử lý dữ liệu lần 2 - Khám phá dữ liệu - Viết báo cáo - Thuyết trình
3	Nguyễn Thành Đạt	- Thu thập dữ liệu - Đề xuất phương pháp - Xử lý dữ liệu lần 1 - Xây dựng các mô hình dự đoán - Viết báo cáo - Thuyết trình