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

□□□□□

XÂY DỰNG MÔ HÌNH DỰ ĐOÁN CHỈ SỐ PHÁT TRIỂN CON NGƯỜI VÀ PHÂN CỤM TỈNH THÀNH ĐÁNG SỐNG TẠI VIỆT NAM

Sinh viên thực hiện:		
STT	Họ tên	MSSV
1	Võ Hoàng An	21520555
2	Trần Ngọc Yến Nhi	21521231
3	Trần Thái Hoà	21522082

TP. HỒ CHÍ MINH – 12/2022

1. GIỚI THIỆU

Mỗi năm, nước ta đều thực hiện các thống kê, khảo sát số liệu về nhiều khía cạnh khác nhau: kinh tế, dân số, môi trường, y tế, chất lượng cuộc sống. Đây là một nguồn dữ liệu giá trị có nhiều tiềm năng khai thác bằng cách sử dụng các phương pháp được học trong môn DS105. Với những dữ liệu thu được nhóm đặt mục tiêu xếp hạng được mức độ đáng sống với những chỉ số do nhóm tự thiết kế và sử dụng các mô hình hồi quy để dự đoán chỉ số phát triển con người của 63 tỉnh thành ở Việt Nam.

Bộ dữ liệu được nhóm thu thập tại trang web của tổng cục thống kê Việt Nam [1]. Tổng cục thống kê cho phép sử dụng các dữ liệu thống kê được đăng tải trên trang web với điều kiện trích dẫn nguồn đầy đủ. Dữ liệu nhóm thu thập là các thuộc tính đặc trưng trên nhiều khía cạnh khác nhau như: y tế, an ninh xã hội, kinh tế, lao động, việc làm, thu nhập bình quân đầu người của các tỉnh thành từ năm 2002 đến năm 2022. Bộ dữ liệu được chuẩn hoá và kết hợp bởi nhiều bộ dữ liệu nhỏ với các định dạng khác nhau, sau đó sử dụng hai thuật toán liên quan đến hồi quy và phân cụm cho phần mô hình và đánh giá. Bên cạnh đó, nhóm sử dụng các thư viện hỗ trợ xử lý dữ liệu chính là Pandas, Numpy, Tableau và Seaborn để trực quan hoá dữ liệu. Nhóm đã đề xuất được bốn độ đo dành cho việc đánh giá mức độ đáng sống và xây dựng mô hình dự đoán của 63 tỉnh thành ở Việt Nam dựa trên bộ dữ liệu của nhóm. Bộ dữ liệu và đề tài được nhóm tự phân tích và thiết kế, không dựa trên bất cứ đề tài có sẵn nào. Bộ dữ liệu chỉ phục vụ riêng cho môn học DS105.

2. MÔ TẢ BỘ DỮ LIỆU

Sau khi phân tích và thu thập dữ liệu từ tổng cục thống kê [1], nhóm thu lại được bộ dữ liệu thô. Toàn bộ dữ liệu được thu thập được tổng cục thống kê cho phép sử dụng với mục đích học tập và nghiên cứu. Bộ dữ liệu được nhóm tự lên ý tưởng, thu thập, chắt lọc và thiết kế hình thái dữ liệu. Do đặc thù của tổng cục thống kê[1] nên dữ liệu khuyết có trong bộ dữ liệu rất nhiều, vì có những năm không thực hiện việc thống kê, khảo sát. Chi tiết nội dung các thuộc tính được trình bày chi tiết ở Bảng 1.

STT	Tên thuộc tính	Kiểu dữ liệu	Nội dung thuộc tính
1	Main_Feature	Text	Các chỉ số chỉ đặc điểm của các tỉnh ở Việt Nam. Chi tiết được trình bày ở bảng 4 phụ lục.
2	Sub_Feature	Text	Là các chỉ số phụ của chỉ số chính. Mô tả chi tiết hơn, bổ sung chi tiết cho các đặc điểm chính. Chi tiết được trình bày ở bảng 4 phụ lục.
3	Province	Text	Các tỉnh thành/vùng ở Việt Nam.
4	2002 → 2022	Numeric	Chỉ số của Main_Feature hoặc Sub_Feature tương ứng với các năm.

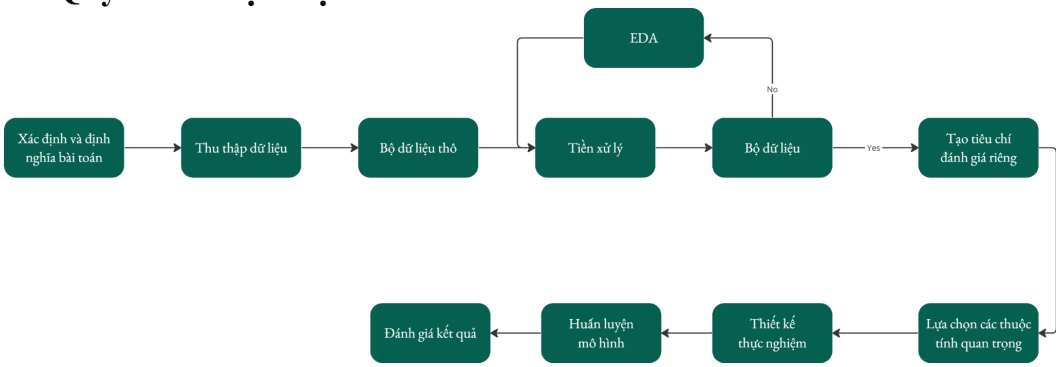
***Bảng 1**: Mô tả chi tiết các thuộc tính trong bộ dữ liệu.*

Bộ dữ liệu có 26 thuộc tính (3 thuộc tính văn bản và 20 thuộc tính số) và 7955 dòng dữ liệu. Dữ liệu của nhóm được thu thập tại website của tổng cục thống kê [1], đây là một cơ quan của nhà nước chuyên thống kê các số liệu tại Việt Nam vì thê số liệu tại trang web là uy tín và đáng tin cậy. Ngoài ra dữ liệu của trang được công khai và được phép sử dụng. Nhóm đã sử dụng công cụ tải xuống tải xuống file excel được cung cấp sẵn trên website. Sau quá trình chắt lọc, nhóm quyết được thu thập 55 file tương ứng với 55 thuộc tính. Tiếp theo nhóm phân tích và đưa ra hình thái chung cho dữ liệu và tiến hành gộp **55 thuộc tính** lại cho ra bộ dữ liệu hoàn chỉnh ở dạng csv. Hình 1 mô tả toàn bộ quá trình thu thập và xử lý dữ liệu. Các điểm dữ liệu mẫu được trình bày chi tiết và đầy đủ tại bảng 5 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

❖ ***Tiền xử lý cho phân tích, thăm dò:***

Chuẩn hóa unicode cho tên các tỉnh và vùng miền. Chuẩn hóa kiểu số cho các giá trị trong các thuộc tính. Các thuộc tính con có thể được tách thành một thuộc tính riêng phục vụ cho việc phân tích. Nhóm tiến hành thăm dò thấy được rằng trong dữ liệu khi bị khuyết sẽ có ký tự "..", nhóm tiến hành chuẩn hoá về NaN. Với các thuộc tính có dữ liệu không phù hợp với định hướng ban đầu của nhóm sẽ được loại bỏ do dữ liệu không đầy đủ ở các năm, phạm vi dữ liệu từ năm 2018 → 2022. Các thuộc tính bị khuyết dữ liệu các năm quá nhiều cũng sẽ bị loại bỏ vì không mang lại nhiều phát hiện. Cuối cùng nhóm thu lại được bộ dữ liệu phụ gồm 44 thuộc tính.

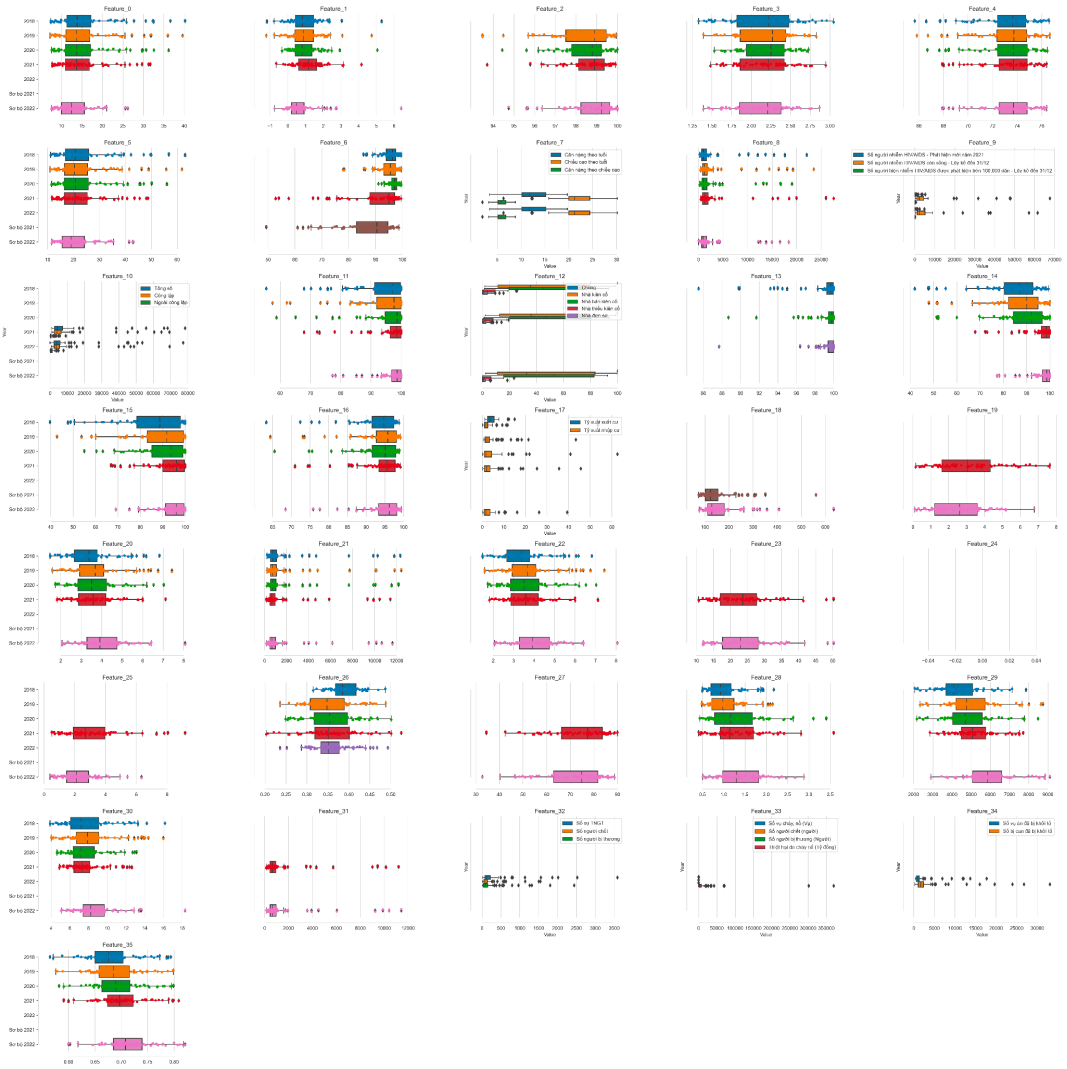
❖ ***Tiền xử lý cho dữ liệu huấn luyện:***

Nhóm tăng cường dữ liệu bằng cách lấy dữ liệu tương ứng với giá trị của biến mục tiêu (HDI) theo từng năm (2018, 2019, 2020, 2021). Các thuộc tính con được tách

thành thuộc tính riêng. Loại bỏ các thuộc tính có lượng dữ liệu khuyết lớn hơn 2 năm. Đồng thời lựa chọn giá trị điền khuyết là 0, mean và mode do các biến định lượng đa phần có phân phối chuẩn. Cuối cùng nhóm thu được bộ dữ liệu cho huấn luyện mô hình gồm 36 thuộc tính và 1 thuộc tính biến mục tiêu (HDI). Bộ dữ liệu này được chia thành 2 tập train và test phục vụ cho việc huấn luyện và đánh giá mô hình. Khi huấn luyện mô hình nhóm lựa chọn các phương pháp điền khuyết là ***0, mean và mode*** do các biến định lượng đa phần có phân phối chuẩn.

3.3. Tìm đặc trưng dữ liệu

❖ *Thống kê mô tả từng thuộc tính*

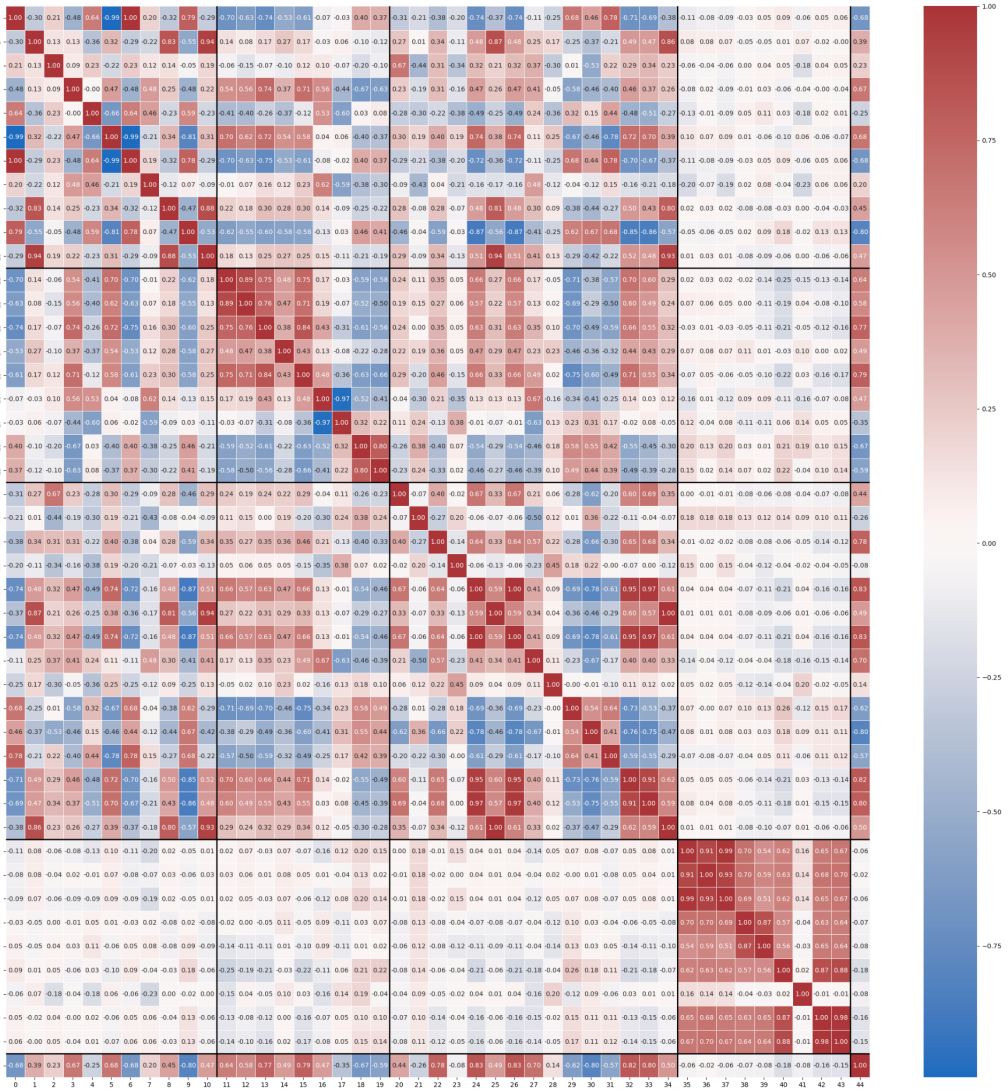


Hình 3: *Thống kê mô tả của từng thuộc tính*

Hầu hết các thuộc tính chính đều có giá trị outlier. Tuy nhiên phần lớn số lượng outliers nhỏ hơn 10 (10/63 tỉnh). Điều này cho thấy dữ liệu các main_features đều có phân phối đồng đều và ít biến động. Các main_features có giá trị outlier nhiều là feature_9/10/17/32/33/34. Lý do là các main_features này đều có chứa sub_features và đơn vị đo của các sub_features có thể khác nhau nên dẫn đến xuất hiện các outlier. Có

tổng cộng 13/37 features chứa dữ liệu ít hơn 3 năm. Các features này sẽ được cân nhắc loại bỏ hoặc điền khuyết ở bước phân tích và đào tạo mô hình.

❖ ***Độ tương quan giữa các thuộc tính***



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

Nhóm quy ước thang đo của độ tương quan: thấp: 0 → 0.4, trung bình: 0.4 → 0.8, cao: 0.8 → 1. Sự tương quan giữa các thuộc tính trong cùng một chỉ số (được định nghĩa ở mục 3.4) thường ở mức trung bình đến cao. Trong khi sự tương quan giữa các thuộc tính ở các chỉ số khác nhau ở mức thấp đến trung bình. Tất cả thuộc tính trong chỉ số *antt* đều có mức tương quan thấp đối với các thuộc tính còn lại. Thuộc tính mục tiêu (HDI) có mức tương quan trung bình-cao với các thuộc tính còn lại.

3.4. Thiết kế tiêu chí đánh giá

Với những dữ liệu thu thập được cùng các phân tích mà nhóm thu được, nhóm đã tạo ra 4 chỉ số phù hợp để đánh giá và đưa ra các kết luận chứa nhiều các thông tin có giá trị. Bốn chỉ số là: *An ninh trật tự*, *Lao động - việc làm - thu nhập*, *Y tế*, *Chất lượng cuộc sống*. Mỗi chỉ số được tạo ra bằng cách lấy trung bình cộng các giá trị đã chuẩn hóa của các thuộc tính có trong chỉ số đó. Sau đó giá trị này được chuẩn hóa về thang

đó [0;1] với 3 mức: Thấp: 0 → 0.34, trung bình: 0.34 → 0.67, cao: 0.67 → 1. Chi tiết được thể hiện ở Bảng 2 tại phụ lục.

4. PHÂN TÍCH THĂM DÒ CHUYÊN SÂU

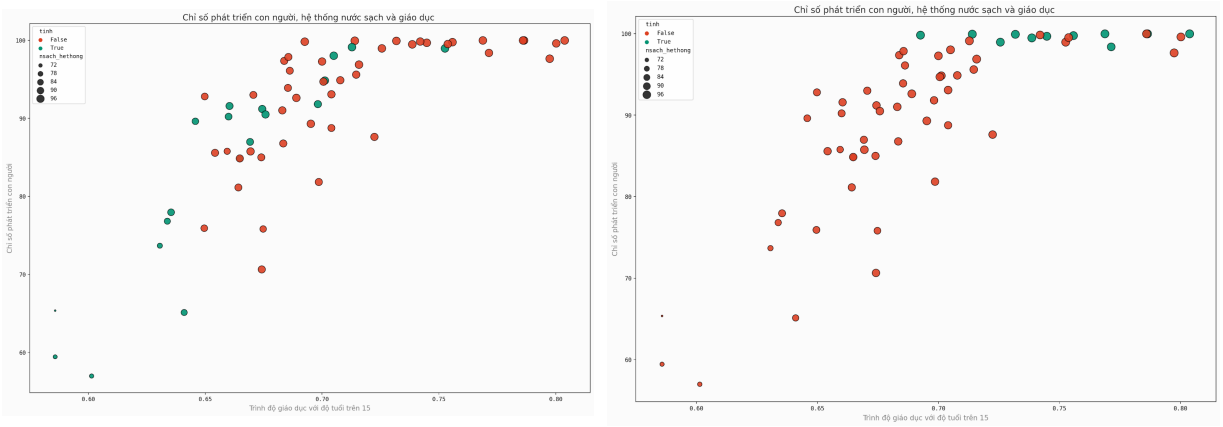
4.1. Phân tích các thuộc tính tiêu biểu trong metrics

❖ *Y tế và sức khỏe*

Tỷ suất chết của trẻ em dưới 1 tuổi và dưới 5 tuổi cao nhất ở Lai Châu và thấp nhất tại TP. HCM, cho thấy có sự chênh lệch lớn trong việc tiếp cận các dịch vụ y tế giữa các vùng. Có sự giảm tỷ suất chết qua các năm nhưng lại tăng nhẹ vào năm 2021, có thể do ảnh hưởng của đại dịch COVID-19 gây ra sự quá tải cho hệ thống y tế và làm gián đoạn các dịch vụ chăm sóc sức khỏe cơ bản. Tỷ lệ tăng dân số giảm đáng kể tại Bình Dương và Đồng Bằng Sông Hồng vào năm 2021 có thể phản ánh sự chuyển dịch dân cư, biến động kinh tế hoặc các chính sách kế hoạch hóa gia đình. Tỷ lệ tiêm chủng đầy đủ giảm mạnh ở Đồng Nai và các tỉnh ở Đồng Bằng Sông Hồng vào năm 2021 (giảm 52%), có thể do đại dịch đã ảnh hưởng đến việc triển khai các chiến dịch tiêm chủng thông thường. Tỷ lệ đăng ký khai sinh cao (trên 93% cho tất cả) cho thấy sự nỗ lực của chính phủ trong việc đảm bảo quyền công dân từ sớm và cung cấp các dịch vụ công cơ bản cho trẻ em.

Có sự tăng tuổi thọ trung bình qua các năm, nhưng lại thấp nhất ở Lai Châu và cao nhất tại TP. HCM, phản ánh sự phân hóa về điều kiện sống và tiếp cận dịch vụ y tế giữa các vùng. Tỷ lệ suy dinh dưỡng cao nhất ở Kon Tum (khoảng 18.87%) và thấp nhất ở TP. HCM (khoảng 4.13%), cho thấy sự khác biệt trong tình trạng an ninh lương thực và tiếp cận dinh dưỡng giữa các vùng. Số lượng người nhiễm HIV/AIDS cao ở TP. HCM và Hà Nội có thể liên quan đến dân số lớn và di chuyển dân cư nhiều hơn. Có thể có một mối tương quan giữa tỷ suất chết và suy dinh dưỡng của trẻ em. Điều này cho thấy cần phải tập trung vào việc cải thiện dinh dưỡng và chăm sóc sức khỏe cho trẻ. Mối liên hệ giữa số bác sĩ, số giường bệnh và số người nhiễm HIV có thể phản ánh năng lực phát hiện và chăm sóc bệnh nhân của hệ thống y tế.

❖ *Chất lượng cuộc sống*



Hình 4: *Biểu đồ mô tả tương quan nhiều thuộc tính - lần lượt là a và b.*

a. Màu xanh thể hiện khu vực TDMNBB và TN. b. Màu xanh thể hiện khu vực ĐBSH

Hình 4a. cho thấy rằng các cụm liên quan đến khu vực Tây Nguyên và Tây Bắc

Việt Nam, được biểu thị bằng chấm xanh, có kích thước nhỏ hơn các cụm khác. Điều này cho thấy hệ thống nước sạch ở khu vực này kém phát triển hơn. Đồng thời cũng cho thấy rằng trình độ người trên 15 tuổi có khả năng biết chữ ở khu vực Tây Nguyên và Tây Bắc Việt Nam chưa cao. Tỷ lệ này tỉ lệ thuận với chỉ số HDI tại khu vực.

Tỷ lệ dân số đô thị được cung cấp nước sạch qua hệ thống cấp nước tập trung ở khu vực đồng bằng Sông Hồng đạt 96,4% năm 2022, cao hơn đáng kể so với tỷ lệ trung bình của cả nước là 92,1%. Tỷ lệ người trên 15 tuổi có khả năng biết chữ ở khu vực đồng bằng Sông Hồng cao hơn đáng kể so với tỷ lệ trung bình của cả nước. Cụ thể, tỷ lệ này đạt 94,1% năm 2022, cao hơn so với tỷ lệ trung bình của cả nước là 91,4%. Qua hình 4 và số liệu thống kê, tỷ lệ dân số từ 15 tuổi trở lên biết chữ ở các vùng kinh tế cũng có sự chênh lệch khá lớn. Vùng đồng bằng sông Hồng có tỷ lệ biết chữ cao nhất, đạt 94,1% năm 2022. Vùng trung du và miền núi phía Bắc có tỷ lệ biết chữ thấp nhất, đạt 84,3% năm 2022. Đồng thời cho thấy rằng cụm liên quan đến khu vực đồng bằng Sông Hồng, được biểu thị bằng chấm xanh, có kích thước lớn hơn so với các cụm khác. Tỷ lệ dân số đô thị được cung cấp nước sạch qua hệ thống cấp nước tập trung và tỷ lệ dân số từ 15 tuổi trở lên biết chữ ở các địa phương cũng có mối quan hệ tương quan thuận. Các tỉnh, thành phố có tỷ lệ cung cấp nước sạch cao nhất là các tỉnh, thành phố thuộc vùng đồng bằng sông Hồng và đô thị lớn, như Thành phố Hồ Chí Minh, Hà Nội, Đà Nẵng, Hải Phòng,... Các tỉnh, thành phố có tỷ lệ cung cấp nước sạch thấp nhất là các tỉnh thuộc vùng trung du và miền núi phía Bắc, như Lai Châu, Điện Biên, Lào Cai,...

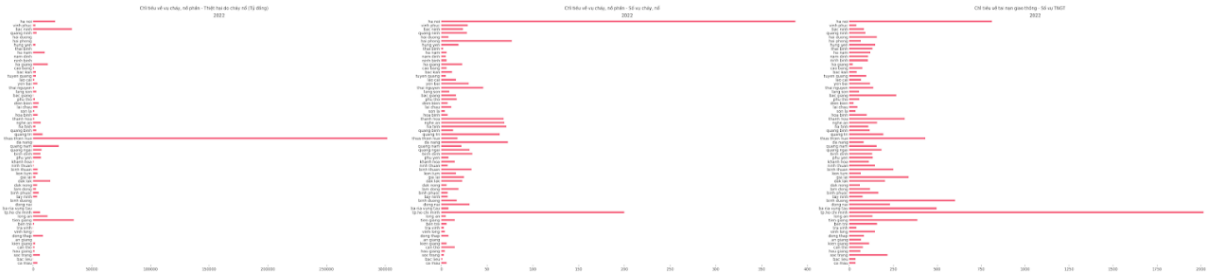
- ***Lao động - việc làm - thu thập***

Bình Dương có tỷ lệ nhập cư cao nhất và cũng là tỉnh có thu nhập bình quân đầu người cao nhất, điều này cho thấy rằng sự thu hút lao động từ nơi khác có thể đang góp phần vào sự tăng trưởng kinh tế của tỉnh. Tuy nhiên, việc tỷ lệ nhập cư giảm trong năm 2021 có thể là kết quả của đại dịch COVID-19, khi mà các biện pháp hạn chế di chuyển được áp dụng. Bắc Ninh đã vươn lên vị trí cao nhất về tỷ suất nhập cư trong năm 2021, có thể do sự phát triển của ngành công nghiệp tại đây, đặc biệt là sự mở rộng của các khu công nghiệp và sự phát triển của ngành điện tử.

Đồng bằng sông Cửu Long và Đồng bằng sông Hồng đều có tỷ lệ nhập cư và năng suất lao động tăng qua các năm, nhưng Đồng Bằng Sông Cửu Long lại có tỷ lệ thiếu việc làm cao nhất. Điều này cho thấy rằng sự phát triển kinh tế không đồng đều và có thể có một lượng lớn lao động nhập cư không tìm được việc làm phù hợp hoặc không có đủ việc làm cho tất cả mọi người. Điện Biên và Tây Nguyên thường xuyên nằm trong số các vùng có thu nhập và năng suất lao động thấp nhất, điều này cho thấy sự cần thiết của việc đầu tư hơn nữa vào cơ sở hạ tầng và giáo dục để tăng cơ hội việc làm và thu nhập cho người dân. Mặc dù Đông Nam Bộ là một trong những vùng phát triển mạnh nhất, nhưng cũng có tỷ lệ thất nghiệp trong độ tuổi lao động cao nhất, có thể là do sự cạnh tranh cao và sự chênh lệch giữa lực lượng lao động có kỹ năng và nhu cầu thực tế của thị trường việc làm. Tỷ lệ cao của lao động có việc làm phi chính

thức cho thấy rằng nhiều người lao động vẫn chưa được hưởng lợi từ sự bảo vệ và lợi ích của việc làm chính thức. Điều này cũng chỉ ra một sự chuyển động lớn về lao động từ nông thôn sang đô thị, nơi mà việc làm phi chính thức thường phổ biến hơn.

❖ ***An ninh trật tự***



Hình 5: *Biểu đồ so sánh lần lượt về thiệt hại cháy nổ - số vụ cháy nổ - số vụ TNGT*

Các tỉnh/khu vực có số vụ TNGT và số người chết nhiều nhất đều thuộc khu vực đô thị, đặc biệt là các thành phố lớn như Hà Nội, TP. Hồ Chí Minh, Bình Dương,... Các tỉnh/khu vực miền núi và miền núi phía Bắc có số vụ TNGT và số người chết ít hơn so với các khu vực khác. Dựa trên dữ liệu thống kê, có thể thấy các "điểm nóng" an toàn giao thông trên cả nước tập trung ở các thành phố lớn, đặc biệt là Hà Nội và TP. Hồ Chí Minh. Số người chết và bị thương trong các vụ cháy nổ thường tỷ lệ thuận với số vụ cháy nổ. Có thể thấy các tỉnh/thành phố có tỷ lệ cháy nổ cao tập trung ở các khu vực thành phố lớn. Các tỉnh/thành phố có tỷ lệ cháy nổ thấp tập trung ở các khu vực miền núi phía Bắc, khu vực có mật độ dân cư thấp. Các điểm nóng cháy nổ trên cả nước là các thành phố lớn, đặc biệt là Hà Nội và TP. Hồ Chí Minh, các khu vực kinh tế phát triển, có mật độ dân cư cao, như Đồng Nai, Bình Dương, Bà Rịa - Vũng Tàu,... và các khu vực có nhiều công trình xây dựng như Hải Phòng, Đà Nẵng, Hà Tĩnh,...

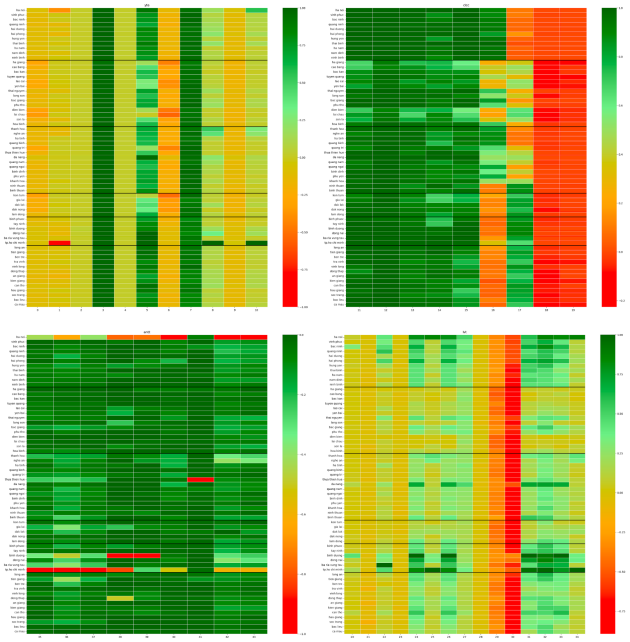
4.2. Phân tích các metrics

Vùng	Y tế		CLCS		LVT		ANTT	
	Mean	**Std**	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**
TDMNPB	0.33	0.18	0.58	0.29	0.14	0.13	**0.95**	**0.04**
DBSH	0.60	**0.06**	**0.97**	**0.03**	0.50	0.17	0.84	0.28
BTBDHMN	0.5	0.14	0.86	0.07	0.27	0.11	0.88	0.08
TN	0.28	0.14	0.72	0.12	0.17	**0.08**	0.92	0.07
DNB	**0.65**	0.20	0.90	0.12	**0.66**	0.27	0.64	0.35
DBSCL	0.56	0.06	0.74	0.1	0.28	**0.08**	0.94	0.05

Bảng 2: *Thống kê mô tả của từng chỉ số phân theo vùng miền.*

Trên thang đo của các heatmap bên dưới, các thuộc tính mang giá trị nhỏ hơn 0 cho thấy mức độ ảnh hưởng tiêu cực và các thuộc tính lớn hơn 0 cho thấy mức độ ảnh hưởng tích cực.

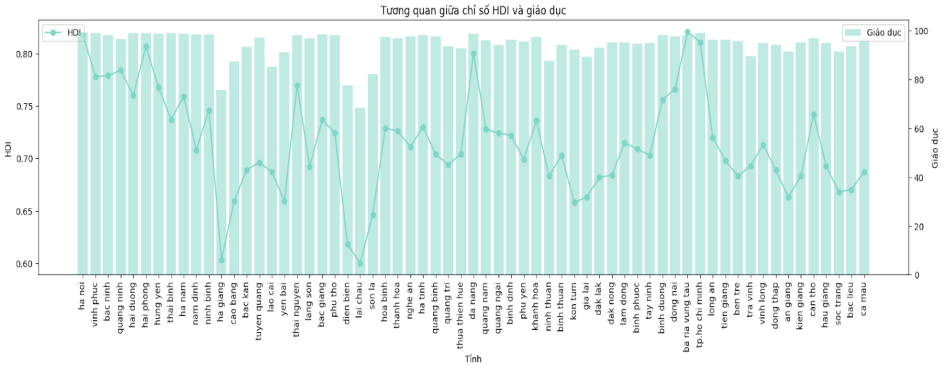
Nhận xét biểu đồ heatmap



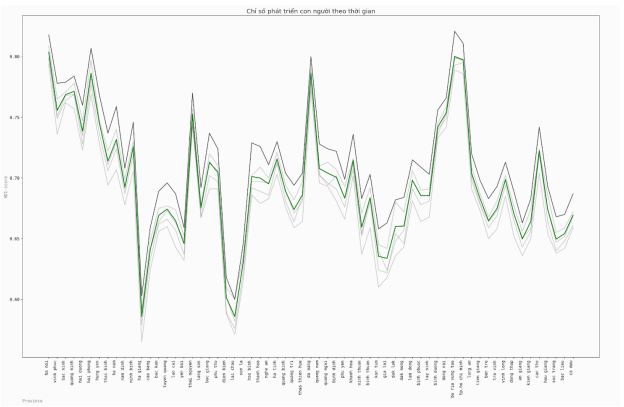
DBSH, DNB là 2 vùng có điểm số cao ở hầu hết các chỉ số, trong khi điểm số của các chỉ số ở 2 vùng TN và TDMNPB là thấp nhất. Tuy nhiên chỉ số antt ở DBSH, DNB lại có biến động cao hơn ở các vùng khác. Việc quản lý và duy trì an ninh trật tự ở các vùng phát triển vẫn còn là một thách thức.

Dựa vào các biểu đồ heatmap tương ứng với các chỉ số. Ta có thể kết luận rằng ở TN và TDMNPB, trong mỗi chỉ số còn có nhiều thuộc tính có giá trị thấp. Ví dụ, ở chỉ số y tế, thuộc tính tỷ suất chết ở trẻ em(dưới 1 tuổi, dưới 5 tuổi) và tỷ lệ suy dinh dưỡng ở trẻ em dưới 5 tuổi còn cao (~ 21.50%).

4.3. Phân tích biến mục tiêu cho mô hình



Chỉ số Phát triển Con người (HDI) của Việt Nam đã có xu hướng tăng qua các năm, từ 0,794 vào năm 2018 lên 0,821 vào năm 2022. Các tỉnh có HDI cao nhất ở Việt Nam bao gồm Bà Rịa - Vũng Tàu, Hà Nội, Thành phố Hồ Chí Minh, Hải Phòng, và Đà Nẵng. Những tỉnh này đều là các trung tâm kinh tế lớn, với cơ sở hạ tầng phát triển, chất lượng giáo dục và y tế cao. Nếu xem xét HDI qua các năm, năm 2022, tỉnh Hà Giang có mức tăng cao nhất với 2,1%, từ



0,565 lên 0,603. Đây là mức tăng cao hơn rất nhiều so với mức tăng trung bình của cả nước (0,7%). Vùng Đồng Bằng Sông Hồng là khu vực có nhiều tỉnh có HDI cao nhất, với 6 tỉnh trong top 10 vào năm 2022. Vùng này có ưu thế về kinh tế, văn hóa, giáo dục, và y tế, đồng thời là trung tâm quan trọng về mặt kinh tế, chính trị, và văn hóa của cả nước. Tuy nhiên, năm 2022, có 6 trong 10 tỉnh nằm trong khu vực Trung du và Miền núi phía Bắc có HDI thấp. Các tỉnh miền núi thường có nền kinh tế kém phát triển, cơ sở hạ tầng chưa hoàn chỉnh, và chất lượng giáo dục, y tế còn hạn chế. Khi xem xét theo vùng, vùng Đồng Bằng Sông Hồng có mức tăng HDI cao nhất, với mức tăng bình quân đạt 0,9%/năm. Vùng Bắc Trung Bộ và Duyên hải miền Trung có mức tăng HDI thấp hơn so với mức tăng bình quân của cả nước, với mức tăng bình quân đạt 0,7%/năm. Vùng Tây Nguyên có mức tăng HDI thấp nhất trong cả nước, với mức tăng bình quân đạt 0,6%/năm. Các tỉnh trong vùng gặp khó khăn về kinh tế - xã hội, cơ sở hạ tầng còn thiếu thốn, và dân trí còn thấp.

5. KẾT QUẢ THÍ NGHIỆM

❖ ***Sử dụng thuộc tính HDI làm biến mục tiêu.***

Mô hình	Điền giá trị 0			Điền giá trị trung bình			Điền giá trị xuất hiện nhiều nhất		
	R2	MSE	MAE	R2	MSE	MAE	R2	MSE	MAE
Linear	0.7891	0.0004	0.0161	0.7979	0.0004	0.0158	0.7979	0.0004	0.0158
SGD	0.3841	0.0012	0.0272	0.6865	0.0006	0.0212	0.6865	0.0006	0.0212
RandomForest	0.8881	0.0002	0.0118	0.8843	0.0002	0.0123	0.8843	0.0002	0.0123
GradientBoosting	**0.9132**	**0.0002**	**0.0102**	**0.9143**	**0.0002**	**0.0103**	**0.9143**	**0.0002**	**0.0103**
LGBM	**0.9155**	**0.0002**	**0.0102**	**0.9164**	**0.0002**	**0.0101**	**0.9164**	**0.0002**	**0.0101**
CatBoost	**0.9359**	**0.0001**	**0.0093**	**0.9337**	**0.0001**	**0.0096**	**0.9337**	**0.0001**	**0.0096**

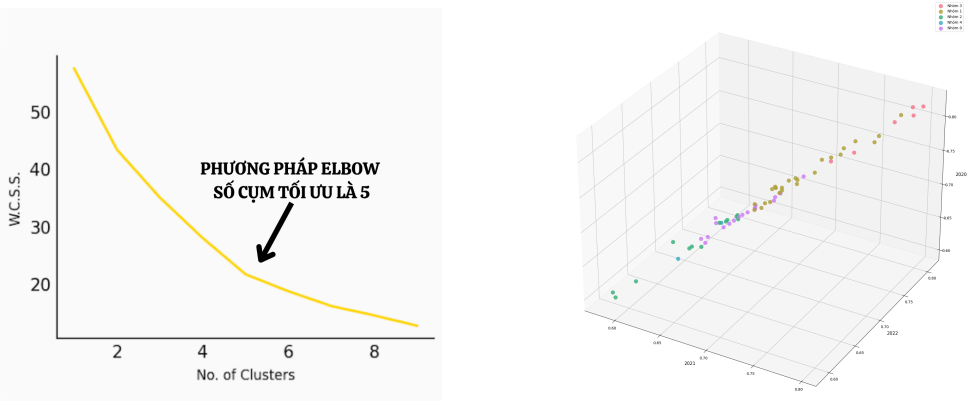
Bảng 3: *Kết quả dự đoán của các mô hình trên nhiều thí nghiệm.*

Nhóm đã sử dụng 3 độ đo phổ biến cho bài toán hồi quy là: R2, Mean squared error, Mean absolute error để đánh hiệu suất của các mô hình. Mô hình CatBoost cho kết quả tốt nhất với R2 = 0.9359 ở thực nghiệm điền giá trị 0. Mô hình SGD cho kết quả thấp nhất trong tất cả mô hình ở tất cả thực nghiệm ở cả ba độ đo. Nhóm nhận thấy MSE của tất cả các mô hình rất bé, điều này có thể lý giải bởi việc biến mục tiêu từ ban đầu chỉ có miền dữ liệu từ (0,1).

❖ ***Phân cụm để đánh giá dựa trên các độ đo đã tạo ra***

Kết quả mô hình cho thấy nhóm 0 và nhóm 1 có quy mô dân số và diện tích tự nhiên lớn, cùng với trình độ phát triển kinh tế - xã hội cao. Nhóm 2, ngược lại, có quy mô nhỏ và đối diện với nhiều khó khăn phát triển. ***Nhóm 3, bao gồm các thành phố***

thể hiện quy mô lớn và trình độ phát triển rất cao, được đánh giá là nhóm tiềm năng với nhiều tỉnh thành đáng sống tại Việt Nam - thể hiện qua hình 6 nhóm 3 đa phần chiếm vị trí ở phía trên cùng của biểu đồ, cho thấy rằng các khu vực trong nhóm



Hình 6: *Biểu đồ thể hiện phương pháp Elbow và phân cụm của mô hình Kmeans.*

này có chỉ số HDI cao và ổn định qua các năm. . Cuối cùng, nhóm 4 tập trung vào các tỉnh đặc biệt trong bối cảnh đa dạng đặc điểm địa lý và xã hội của chúng.

6. KẾT LUẬN

Nhóm đã thực hiện một quá trình thu thập và tiền xử lý dữ liệu kỹ lưỡng, bắt đầu bằng việc chọn lọc và gộp các thuộc tính từ 55 file dữ liệu để tạo ra một bộ dữ liệu hoàn chỉnh. Trước khi bắt đầu phân tích, nhóm đã 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, nhóm 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 HDI, nhóm đã chọn các mô hình như GradientBoosting, LGBM, CatBoost, SGD, Linear, RandomForest 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 0.9359. Mô hình SGD, mặc dù có kết quả thấp nhất, nhưng cũng đóng góp vào việc đánh giá toàn diện về hiệu suất của các mô hình. Nhóm cũng đã thực hiện phân cụm để đánh giá các đặc trưng của các tỉnh thành. Kết quả cho thấy sự chia rõ giữa các 5 cụm với quy mô dân số và diện tích tự nhiên phù hợp với từng khu vực, đồng thời nhận diện các nhóm các tỉnh thành đáng sống tiềm năng. Tổng cộng, quá trình nghiên cứu đã đem lại cái nhìn toàn diện về tình hình phát triển của các tỉnh thành tại Việt Nam và cung cấp những thông tin hữu ích cho việc đưa ra quyết định và chiến lược phát triển trong tương lai.

Các mô hình dự đoán của nhóm cũng trả về trọng số cho các thuộc tính. Trong tương lai nhóm 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 HDI, đồng thời cũng tiếp tục khám phá và phân tích các thuộc tính tiềm năng có trong bộ dữ liệu.

PHỤ LỤC

STT	Ghi tắt	Nội dung
1	TDMNPB	Trung Du Miền Núi Phía Bắc
2	DBSH hoặc ĐBSH	Đồng Bằng Sông Hồng
3	BTBDHMN	Bắc Trung Bộ Duyên Hải Miền Nam
4	TN	Tây Nguyên
5	DNB hoặc ĐNB	Đông Nam Bộ
6	DBSCL hoặc ĐBSCL	Đồng Bằng Sông Cửu Long
7	CLCS	Chất lượng cuộc sống
8	ANTT	An ninh trật tự
9	LVT	Lao động - Việc làm - Thu nhập
10	TP. HCM	Thành phố Hồ Chí Minh

Bảng 1: *Chi tiết các ghi tắt trong bài.*

STT	Chỉ số	Các thuộc tính
1	An ninh trật tự	- Một số chỉ tiêu về tai nạn giao thông. - Một số chỉ tiêu về vụ cháy, nổ. - Số vụ án và bị cáo đã xét xử sở thẩm.
2	Lao động - việc làm - Thu nhập	- Thu nhập bình quân đầu người một tháng theo giá hiện hành. - Thu nhập bình quân đầu người một tháng theo giá hiện hành phân theo trung bình 5 nhóm thu nhập. - Thu nhập bình quân đầu người một tháng (Nghìn đồng) - Nhóm thu nhập thấp nhất. - Thu nhập bình quân đầu người một tháng (Nghìn đồng) - Nhóm thu nhập cao nhất. - Thu nhập bình quân một lao động đang làm việc. - Hệ số bất bình đẳng trong phân phối thu nhập (hệ số

		GINI). - Tỷ lệ hộ nghèo. - Tỷ suất nhập cư, xuất cư và di cư thuần. - Lực lượng lao động từ 15 tuổi trở lên. - Tỷ lệ thiếu việc làm trong độ tuổi lao động. - Số lao động có việc làm trong nền kinh tế. - Tỷ lệ lao động có việc làm phi chính thức. - Tỷ lệ thất nghiệp trong độ tuổi lao động. - Năng suất lao động. - Tỷ lệ lao động từ 15 tuổi trở lên đã qua đào tạo.
3	Y tế	- Tỷ suất chết của trẻ em dưới 1 tuổi. - Tỷ lệ tăng dân số. - Tỷ lệ trẻ em dưới 05 tuổi được đăng ký khai sinh. - Tổng tỷ suất sinh phân theo địa phương. - Tuổi thọ trung bình tính từ lúc sinh. - Tỷ suất chết của trẻ em dưới 5 tuổi. - Tỷ lệ trẻ em dưới một tuổi được tiêm chủng đầy đủ các loại vắc xin. - Tỷ lệ trẻ em dưới 5 tuổi bị suy dinh dưỡng. - Số bác sĩ. - Số người nhiễm HIVAIDS. - Số giường bệnh phân theo địa phương.
4	Chất lượng cuộc sống	- Chỉ số phát triển con người. - Tỷ lệ dân số được sử dụng nguồn nước hợp vệ sinh. - Tỷ lệ hộ có nhà ở. - Tỷ lệ hộ dùng điện sinh hoạt. - Tỷ lệ dân số đô thị được cung cấp nước sạch qua hệ thống cấp nước. - Tỷ lệ dân số dùng hố xí hợp vệ sinh. - Tỷ lệ dân số từ 15 tuổi trở lên biết chữ.

Bảng 2: *Chi tiết về các độ đo.*

Chỉ mục	Thuộc tính
0	Tỷ suất chết của trẻ em dưới 1 tuổi phân theo địa phương chia theo Địa phương và Năm.

1	Tỷ lệ tăng dân số phân theo địa phương chia theo Tỉnh-Thành phố và Năm.
2	Tỷ lệ trẻ em dưới 05 tuổi được đăng ký khai sinh phân theo địa phương chia theo Tỉnh-Thành phố và Năm.
3	Tổng tỷ suất sinh phân theo địa phương chia theo Địa phương và Năm.
4	Tuổi thọ trung bình tính từ lúc sinh phân theo địa phương chia theo Tỉnh-Thành phố và Năm.
5	Tỷ suất chết của trẻ em dưới 5 tuổi phân theo địa phương chia theo Tỉnh-Thành phố và Năm.
6	Tỷ lệ trẻ em dưới một tuổi được tiêm chủng đầy đủ các loại vắc xin phân theo địa phương chia theo Địa phương và Năm.
7	Tỷ lệ trẻ em dưới 5 tuổi bị suy dinh dưỡng phân theo địa phương chia theo Địa phương, Năm và Phân tổ.
8	Số bác sĩ phân theo địa phương.
9	Số người nhiễm HIVAIDS phân theo địa phương chia theo Địa phương, Năm và Phân tổ.
10	Số giường bệnh phân theo địa phương.
11	Tỷ lệ dân số được sử dụng nguồn nước hợp vệ sinh phân theo địa phương
12	Tỷ lệ hộ có nhà ở phân theo loại nhà và phân theo địa phương chia theo Địa phương, Năm và Loại nhà
13	Tỷ lệ hộ dùng điện sinh hoạt phân theo địa phương chia theo Địa phương và Năm
14	Tỷ lệ dân số đô thị được cung cấp nước sạch qua hệ thống cấp nước tập trung phân theo địa phương() chia theo Địa phương và Năm(khó)
15	Tỷ lệ dân số dùng hố xí hợp vệ sinhphân theo địa phương chia theo Địa phương và Năm
16	Tỷ lệ dân số từ 15 tuổi trở lên biết chữ phân theo địa phương chia theo Địa phương và Năm

17	Tỷ suất nhập cư, xuất cư và di cư thuần phân theo địa phương chia theo Tỉnh-Thành phố, Tỷ suất và Năm
18	Năng suất lao động phân theo địa phương. Đơn vị tính Triệu đồng người
19	Tỷ lệ thiếu việc làm trong độ tuổi lao động phân theo địa phương. Đơn vị tính
20	Thu nhập bình quân đầu người một tháng theo giá hiện hành phân theo địa phương chia theo Địa phương
21	Lực lượng lao động từ 15 tuổi trở lên phân theo địa phương chia theo Địa phương và Năm
22	Thu nhập bình quân đầu người một tháng theo giá hiện hành phân theo trungn bình 5 nhóm thu nhập và phân theo địa phương chia theo Địa phương, Nhóm thu nhập và Năm
23	Tỷ lệ lao động từ 15 tuổi trở lên đã qua đào tạo phân theo địa phương. Đơn vị tính
24	Tỷ lệ hộ nghèo phân theo địa phương chia theo Địa phương và Năm
25	Tỷ lệ thất nghiệp trong độ tuổi lao động phân theo địa phương. Đơn vị tính
26	Hệ số bất bình đẳng trong phân phối thu nhập (hệ số GINI) phân theo địa phương chia theo Địa phương và Năm
27	Tỷ lệ lao động có việc làm phi chính thức phân theo địa phương. Đơn vị tính.
28	Thu nhập bình quân đầu người một tháng (Nghìn đồng) - Nhóm thu nhập thấp nhất
29	Thu nhập bình quân một lao động đang làm việc
30	Thu nhập bình quân đầu người một tháng (Nghìn đồng) - Nhóm thu nhập cao nhất
31	Số lao động có việc làm trong nền kinh tế phân theo địa phương. Đơn vị tính Nghìn người
32	Một số chỉ tiêu về tai nạn giao thông phân theo địa phương chia theo Địa phương, Năm và Chỉ tiêu.

33	Một số chỉ tiêu về vụ cháy, nổ phân theo địa phương chia theo Địa phương, Năm và Chỉ tiêu
34	Số vụ án và bị cáo đã xét xử sở thẩm phân theo địa phương chia theo Địa phương, Năm và Chỉ tiêu
35	Chỉ số phát triển con người phân theo địa phương

Bảng 3: *Ánh xạ chỉ mục biểu đồ tương quan*

STT	Các thuộc tính Chính (Main_Features)	Các thuộc tính phụ (Sub_Features)
1	Tỷ suất chết của trẻ em dưới 1 tuổi phân theo địa phương chia theo Địa phương và Năm	
2	Số nhân lực ngành Y trực thuộc sở Y tế phân theo địa phương	Bác sĩ
		Y sĩ
		Y tá
		Hộ sinh
3	Tỷ suất sinh thô, tỷ suất chết thô và tỷ lệ tăng tự nhiên của dân số phân theo địa phương chia theo Địa phương, Phân tổ và Năm	Tỷ suất chết thô
		Tỷ suất sinh thô
4	Số người nhiễm HIVAIDS phân theo địa phương chia theo Địa phương, Năm và Phân tổ	Số người nhiễm HIV/AIDS - Phát hiện mới năm 2021.
		Số người nhiễm HIV/AIDS còn sống - Lũy kế đến 31/12
		Số người hiện nhiễm HIV/AIDS được phát hiện trên 100.000 dân - Lũy kế đến 31/12

5	Tỷ lệ tăng dân số phân theo địa phương chia theo Tỉnh-Thành phố và Năm	None
6	Tỷ lệ trẻ em dưới 05 tuổi được đăng ký khai sinh phân theo địa phương chia theo Tỉnh-Thành phố và Năm	None
7	Tổng tỷ suất sinh phân theo địa phương chia theo Địa phương và Năm	None
8	Số người nhiễm HIVAIDS và số người chết do AIDS phân theo địa phương	Số người nhiễm HIV - Phát hiện mới
		Số bệnh nhân AIDS - Phát hiện mới
		Số người chết do AIDS
		Số người nhiễm HIV còn sống - Lũy kế tính đến 31/12
		Số bệnh nhân AIDS còn sống - Lũy kế tính đến 31/12
9	Tuổi thọ trung bình tính từ lúc sinh phân theo địa phương chia theo Tỉnh-Thành phố và Năm	None
10	Số giường bệnh trực thuộc sở Y tế phân theo địa phương	Tổng số
		Bệnh viện
		Phòng khám khu vực
		B.V điều dưỡng và phục hồi chức năng
		Trạm y tế xã, phường, cơ quan, XN
11	Tỷ suất chết của trẻ em dưới 5 tuổi phân theo địa phương chia theo Tỉnh-Thành phố và Năm	None
12	Số giường bệnh phân theo địa phương	Công lập

		Công lập
		Ngoài công lập
13	Tỷ lệ trẻ em dưới 5 tuổi bị suy dinh dưỡng phân theo địa phương chia theo Địa phương, Năm và Phân tổ	Cân nặng theo tuổi
		Chiều cao theo tuổi
		Cân nặng theo chiều cao
14	Số bác sĩ phân theo địa phương	None
15	Số nhân lực ngành dược trực thuộc sở Y tế phân theo địa phương	Dược sĩ cao cấp
		Dược sĩ trung cấp
		Dược tá
16	Số cơ sở khám, chữa bệnh trực thuộc sở Y tế phân theo địa phương	Tổng Số
		Bệnh viện
		Phòng khám khu vực
		Bệnh viện điều dưỡng và phục hồi chức năng
		Trạm y tế xã, phường, cơ quan, xí nghiệp
17	Chỉ số phát triển con người phân theo địa phương	None
18	Tỷ lệ dân số được sử dụng nguồn nước hợp vệ sinh phân theo địa phương	None
19	Tỷ lệ hộ có nhà ở phân theo loại nhà và phân theo địa phương chia theo Địa phương, Năm và Loại nhà	Nhà kiên cố
		Chung
		Nhà bán kiên cố
		Nhà thiếu kiên cố
		Nhà đơn sơ
20	Tỷ lệ hộ dùng điện sinh hoạt phân theo địa	None

	phương chia theo Địa phương và Năm	
21	Tỷ lệ dân số đô thị được cung cấp nước sạch qua hệ thống cấp nước tập trung phân theo địa phương() chia theo Địa phương và Năm(khó)	None
22	Tỷ lệ dân số dùng hố xí hợp vệ sinhphân theo địa phương chia theo Địa phương và Năm	None
23	Tỷ lệ dân số từ 15 tuổi trở lên biết chữ phân theo địa phương chia theo Địa phương và Năm	None
24	Một số chỉ tiêu về tai nạn giao thông phân theo địa phương chia theo Địa phương, Năm và Chỉ tiêu	Số người bị thương
		Số người chết
25	Một số chỉ tiêu về vụ cháy, nổ phân theo địa phương chia theo Địa phương, Năm và Chỉ tiêu	Số người chết (người)
		Thiệt hại do cháy nổ (Tỷ đồng)
		Số người bị thương (Người)
26	Số vụ án và số bị can đã bị khởi tố phân theo địa phương chia theo Địa phương, Năm và Chỉ tiêu	Số bị can đã bị khởi tố
		Tổng số
27	Số vụ án và bị cáo đã xét xử sở thẩm phân theo địa phương chia theo Địa phương, Năm và Chỉ tiêu	Số bị can đã bị khởi tố
		Tổng số
28	Tỷ suất nhập cư, xuất cư và di cư thuần phân theo địa phương chia theo Tỉnh-Thành phố, Tỷ suất và Năm	Tỷ suất xuất cư
		Tỷ suất nhập cư
29	Năng suất lao động phân theo địa phương. Đơn vị tính Triệu đồngngười	None
30	Chỉ số phát triển con người phân theo địa phương	None
31	Tỷ lệ thiếu việc làm trong độ tuổi lao động phân theo địa phương. Đơn vị tính	None
32	Thu nhập bình quân đầu người một tháng theo giá hiện hành phân theo địa phương chia	None

	theo Địa phương	
33	Lực lượng lao động từ 15 tuổi trở lên phân theo địa phương chia theo Địa phương và Năm	None
34	Thu nhập bình quân đầu người một tháng theo giá hiện hành phân theo trungn bình 5 nhóm thu nhập và phân theo địa phương chia theo Địa phương, Nhóm thu nhập và Năm	None
35	Tỷ lệ lao động từ 15 tuổi trở lên đã qua đào tạo phân theo địa phương. Đơn vị tính	None
36	Tỷ lệ hộ nghèo phân theo địa phương chia theo Địa phương và Năm	None
37	Tỷ lệ thất nghiệp trong độ tuổi lao động phân theo địa phương. Đơn vị tính	None
38	Hệ số bất bình đẳng trong phân phối thu nhập (hệ số GINI) phân theo địa phương chia theo Địa phương và Năm	None
39	Tỷ lệ lao động có việc làm phi chính thức phân theo địa phương. Đơn vị tính	None
40	Thu nhập bình quân đầu người một tháng (Nghìn đồng) - Nhóm thu nhập thấp nhất	None
41	Thu nhập bình quân một lao động đang làm việc	None
42	Số lao động có việc làm trong nền kinh tế phân theo địa phương. Đơn vị tính Nghìn người	None

Bảng 4: *Ánh xạ các thuộc tính có trong dữ liệu.*

STT	Main_Features	Sub_Features	Province	2002	...	Sơ bộ 2021	2022	Sợ bộ 2022
1	Tỷ lệ hộ có nhà ở phân theo loại nhà và phân theo địa phương chia	Nhà kiên cố	Hà Nội	NaN	…	NaN	NaN	93.63

	theo Địa phương, Năm và Loại nhà							
2	Tỷ lệ dân số được sử dụng nguồn nước hợp vệ sinh phân theo địa phương	NaN	Hà Nội	NaN	…	NaN	NaN	100.0

***Bảng 5**: Một vài mẫu dữ liệu trong bộ dữ liệu*

TÀI LIỆU THAM KHẢO

[1] TỔNG CỤC THỐNG KÊ. Link: https://www.gso.gov.vn (15/11/2023).

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Võ Hoàng An	- Thu thập dữ liệu - Đề xuất phương pháp - Xử lý dữ liệu lần 2 - Khám phá dữ liệu - Góp ý báo cáo - Thuyết trình
2	Nguyễn Ngọc Yến Nhi	- Thu thập dữ liệu - Đề xuất phương pháp - Khám phá dữ liệu - Xây dựng mô hình phân cụm - Góp ý báo cáo - Thuyết trình
3	Trần Thái Hoà	- Thu thập dữ liệu - Đề xuất phương pháp - Xử lý dữ liệu lần 1 - Xây dựng mô hình tuyến tính - Viết báo cáo - Thuyết trình