

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

KHOA ĐÀO TẠO SAU ĐẠI HỌC

A red circle with a yellow star in the middle

AI-generated content may be incorrect.

BÁO CÁO

HỌC PHẦN KHAI PHÁ DỮ LIỆU

ĐỀ TÀI: ỨNG DỤNG THUẬT TOÁN DBSCAN VÀ HIERARCHICAL CLUSTERING CHO BÀI TOÁN PHÂN BỔ NGUỒN VỐN HỖ TRỢ CHO CÁC QUỐC GIA

|  |  |  |
| --- | --- | --- |
| GIẢNG VIÊN: | TS. Đặng Hoàng Long | |
| NHÓM HỌC VIÊN: | Trần Quang Hiệp  Đặng Quang Dũng  Cao Minh Quyền  Lê Mạnh Hùng | B24CHKH  B24CHKH004  B24CHKH  B24CHKH |

Hà Nội – 03/2025

MỤC LỤC

[1. Giới thiệu đề tài 7](#_Toc192331404)

[1.1. Lý do lựa chọn đề tài 7](#_Toc192331405)

[1.2. Mục tiêu nghiên cứu 7](#_Toc192331406)

[1.3. Phạm vi và giới hạn của đồ án 7](#_Toc192331407)

[2. Tổng quan về bài toán 8](#_Toc192331408)

[2.1. Bối cảnh và ý nghĩa thực tế 8](#_Toc192331409)

[2.2. Nguồn dữ liệu và đặc trưng của dữ liệu 8](#_Toc192331410)

[2.3. Hướng tiếp cận: Unsupervised Learning 9](#_Toc192331411)

[3. Khai phá dữ liệu 10](#_Toc192331412)

[3.1. Thông tin chung về dữ liệu 10](#_Toc192331413)

[3.2. Mô tả thống kê cơ bản 13](#_Toc192331414)

[3.3. Tóm tắt các phát hiện chính từ EDA 17](#_Toc192331415)

[4. Tiền xử lý dữ liệu & Feature Engineering - Hiệp 18](#_Toc192331416)

[5. Các thuật toán phân cụm 21](#_Toc192331417)

[5.1. Giới thiệu chung về học không giám sát (Unsupervised Learning) và bài toán phân cụm 21](#_Toc192331418)

[5.2. DBSCAN 22](#_Toc192331419)

[5.2.1. Giới Thiệu 22](#_Toc192331420)

[5.2.2. Định Nghĩa Thuật Toán DBSCAN 22](#_Toc192331421)

[5.2.3. Các Tham Số Chính Của DBSCAN 22](#_Toc192331422)

[5.2.4. Các Đặc Điểm Của DBSCAN 23](#_Toc192331423)

[5.2.5. Hoạt Động Của DBSCAN 23](#_Toc192331424)

[5.2.6. Ví Dụ Minh Họa DBSCAN 24](#_Toc192331425)

[5.2.8. Ứng Dụng Thực Tế Của DBSCAN Trong Phân Cụm Quốc Gia 29](#_Toc192331426)

[5.2.9. Tổng kết lại về thuật DBSCAN: 32](#_Toc192331427)

[5.3. Hierarchical Clustering 33](#_Toc192331428)

[6. Xây dựng mô hình và so sánh kết quả 35](#_Toc192331429)

[6.1. Triển khai từng mô hình phân cụm trên bộ dữ liệu 35](#_Toc192331430)

[6.2. Đánh giá chất lượng mô hình 36](#_Toc192331431)

[6.2.1. DBSCAN Clustering 36](#_Toc192331432)

[6.2.2. Hierarchical Clustering 37](#_Toc192331433)

[6.3. So sánh kết quả trực quan 38](#_Toc192331434)

[6.3.1 DBSCAN Clustering 38](#_Toc192331435)

[6.3.2. Hierarchical Clustering 40](#_Toc192331436)

[6.3.3. So sánh cụm của DBSCAN và Hierarchical về mặt phân bố. 45](#_Toc192331437)

[6.4. Nhận xét, phân tích lý do mô hình cho kết quả tốt/không tốt 48](#_Toc192331438)

[7. Kết luận 48](#_Toc192331439)

[7.1. Tổng kết những điểm chính đạt được 48](#_Toc192331440)

[7.2. Hạn chế của nghiên cứu và hướng phát triển tương lai 48](#_Toc192331441)

[7.3. Bài học kinh nghiệm và đề xuất ứng dụng thực tiễn 49](#_Toc192331442)

DANH MỤC TỪ VIẾT TẮT

|  |  |  |
| --- | --- | --- |
| STT | Ký hiệu chữ viết tắt | Chữ viết đầy đủ |
| 1 | DBSCAN | Density-Based Spatial Clustering of Applications with Noise |
| 2 | KNN | K-nearest neighbors |
| 3 | PCA | Principal component analysis |

DANH MỤC HÌNH ẢNH

[Hình 1: Một số mẫu dữ liệu 10](#_Toc192331456)

[Hình 2: Kiểu dữ liệu của các biến 11](#_Toc192331457)

[Hình 3: Thống kê sơ bộ dữ liệu 11](#_Toc192331458)

[Hình 4: Đồ thị phân phối giá trị của các biến 13](#_Toc192331459)

[Hình 5: Biểu đồ tỷ lệ tử vong trẻ em (child\_mort) 14](#_Toc192331460)

[Hình 6: Biểu đồ chỉ số xuất khẩu 15](#_Toc192331461)

[Hình 7: Biểu đồ tỷ lệ chi tiêu cho y tế 16](#_Toc192331462)

[Hình 8: Ma trận tương quan 18](#_Toc192331463)

[Hình 9: Bảng chuẩn hoá giá trị 19](#_Toc192331464)

[Hình 10: Bảng chuẩn hoá giá trị 20](#_Toc192331465)

[Hình 11: Trực quan hóa PCA 20](#_Toc192331466)

[Hình 12: Ví dụ điểm dữ liệu cho thuật toán DBSCAN 24](#_Toc192331467)

[Hình 13: Tính khoảng cách tới các điểm 25](#_Toc192331468)

[Hình 14: Phân loại các điểm 26](#_Toc192331469)

[Hình 15: Giảm minPts và phân loại điểm 26](#_Toc192331470)

[Hình 16: Trực quan phân cụm DBSCAN 27](#_Toc192331471)

[Hình 17: Tập dữ liệu mẫu 30](#_Toc192331472)

[Hình 18: Kết quả phân cụm DBSCAN 31](#_Toc192331473)

[Hình 19: Biểu đồ kết quả phân cụm DBSCAN 32](#_Toc192331474)

[Hình 20: Biểu đồ k-distance graph DBSCAN với dữ liệu kết hợp 36](#_Toc192331475)

[Hình 21: Biểu đồ k-distance graph DBSCAN với dữ liệu PCA 37](#_Toc192331476)

[Hình 22: Biểu đồ trực quan hóa kết quả phân cụm với DBSCAN với dữ liệu kết hợp 38](#_Toc192331477)

[Hình 23: Biểu đồ Boxplot trực quan hóa kết quả phân cụm với DBSCAN với dữ liệu kết hợp 38](#_Toc192331478)

[Hình 24: Biểu đồ trực phân bổ nguồn vốn theo quốc gia với DBSCAN với dữ liệu kết hợp 39](#_Toc192331479)

[Hình 25: Biểu đồ trực quan hóa kết quả phân cụm với DBSCAN với dữ liệu PCA 39](#_Toc192331480)

[Hình 26: Biểu đồ boxplot trực quan hóa kết quả phân cụm với DBSCAN với dữ liệu PCA 40](#_Toc192331481)

[Hình 27: Biểu đồ trực phân bổ nguồn vốn theo quốc gia với DBSCAN với dữ liệu PCA 40](#_Toc192331482)

[Hình 28: Biểu đồ dendrogram kết quả phân cụm với Hierachical Clustering với dữ liệu kết hợp 41](#_Toc192331483)

[Hình 29: Biểu đồ trực quan hóa kết quả phân cụm với Hierachical Clustering với dữ liệu kết hợp 42](#_Toc192331484)

[Hình 30: Biểu đồ Boxplot kết quả phân cụm với Hierachical Clustering với dữ liệu kết hợp 42](#_Toc192331485)

[Hình 31: Biểu đồ trực phân bổ nguồn vốn theo quốc gia với Hierarchical Clustering với dữ liệu kết hợp 43](#_Toc192331486)

[Hình 32: Biểu đồ dendrogram kết quả phân cụm với Hierachical Clustering với dữ liệu PCA 43](#_Toc192331487)

[Hình 33: Biểu đồ trực quan hóa kết quả phân cụm với Hierachical Clustering với dữ liệu PCA 44](#_Toc192331488)

[Hình 34: Biểu đồ Boxplot kết quả phân cụm với Hierachical Clustering với dữ liệu PCA 44](#_Toc192331489)

[Hình 35: Biểu đồ trực phân bổ nguồn vốn theo quốc gia với Hierarchical Clustering với dữ liệu PCA 45](#_Toc192331490)

DANH MỤC BẢNG BIỂU

[Bảng 1: Bảng kết quả số lượng cụm phân loại của các thuật toán 46](#_Toc192331492)

[Bảng 2: Bảng kết quả số lượng của từng cụm phân loại của các thuật toán 47](#_Toc192331493)

[Bảng 3: Bảng kết quả Điểm Silhouse của từng cụm phân loại của các thuật toán 48](#_Toc192331494)

# 1. Giới thiệu đề tài

## 1.1. Lý do lựa chọn đề tài

Trong bối cảnh toàn cầu hoá, sự chênh lệch về điều kiện kinh tế - xã hội giữa các quốc gia ngày càng được thể hiện rõ nét. Hàng triệu người tại nhiều khu vực trên thế giới vẫn phải đối mặt với nghèo đói, dịch vụ y tế lạc hậu và thiên tai triền miên. Trước hiện trạng này, Tổ chức HELP International – một tổ chức phi chính phủ (NGO) – đã gây quỹ thành công khoảng 10 triệu đô la Mỹ nhằm hỗ trợ các quốc gia khó khăn. Tuy nhiên, câu hỏi đặt ra là: Làm thế nào để phân bổ khoản kinh phí này một cách công bằng, hiệu quả và dựa trên những bằng chứng dữ liệu xác thực nhất?

Sự phát triển vượt bậc của khoa học dữ liệu gợi mở một hướng đi mới: Sử dụng kỹ thuật học máy không giám sát (Unsupervised Learning) để phân tích dữ liệu kinh tế – xã hội và y tế của các quốc gia, từ đó nhận diện những nhóm nước có đặc điểm tương đồng. Giải pháp này không chỉ giúp tối ưu hoá việc sử dụng ngân sách mà còn nâng cao tính minh bạch và khách quan trong quá trình ra quyết định.

## 1.2. Mục tiêu nghiên cứu

Đề tài hướng đến việc xây dựng một quy trình phân cụm (clustering) các quốc gia dựa trên đặc điểm kinh tế, xã hội và y tế, từ đó hỗ trợ Tổ chức HELP International xác định thứ tự ưu tiên trong phân bổ nguồn quỹ. Cụ thể:

* **Phân nhóm các quốc gia**: Dựa trên các chỉ số tổng quan (GDP, mật độ dân số, tỷ lệ tử vong, mức độ tiếp cận dịch vụ y tế, v.v.) nhằm tìm ra những nhóm nước có cấu trúc tương đồng.
* **So sánh và đánh giá** các thuật toán phân cụm phổ biến như K-Means, DBSCAN và Hierarchical Clustering thông qua những chỉ số như: Silhouette Score, Calinski-Harabasz).
* **Đề xuất ứng dụng thực tiễn**: Tận dụng kết quả phân cụm để hỗ trợ quá trình quyết định phân bổ nguồn lực khẩn cấp hoặc dài hạn đến các khu vực cần giúp đỡ nhất.

Thông qua việc triển khai các bước từ thu thập, phân tích dữ liệu đến đánh giá mô hình, đồ án hy vọng mang lại góc nhìn cận cảnh về cách thức ứng dụng khoa học dữ liệu vào giải quyết những thách thức nhân đạo quy mô lớn.

## 1.3. Phạm vi và giới hạn của đồ án

Phạm vi của đồ án xoay quanh bộ dữ liệu được cung cấp, bao gồm một số quốc gia cùng các chỉ số điển hình (về kinh tế, xã hội, y tế). Do những ràng buộc về thời gian và nguồn lực, nghiên cứu không bao quát toàn bộ quốc gia trên thế giới hay tất cả các chỉ số mở rộng (như giáo dục, môi trường).

Hơn nữa, đồ án tập trung vào bài toán phân cụm – một phân nhánh của học máy không giám sát. Các chủ đề nâng cao như dự báo xu hướng hay xây dựng mô hình phân loại không nằm trong phạm vi nghiên cứu, song kết quả thu được từ quá trình phân cụm có thể làm nền tảng cho các nghiên cứu tiếp theo về phân bổ nguồn lực hoặc phân tích chính sách vĩ mô.

# 2. Tổng quan về bài toán

## 2.1. Bối cảnh và ý nghĩa thực tế

Với sứ mệnh hỗ trợ các quốc gia đang gặp khó khăn về kinh tế và y tế, Tổ chức HELP International đã và đang nghiên cứu những giải pháp hữu hiệu để sử dụng nguồn quỹ 10 triệu đô la Mỹ một cách hiệu quả nhất. Bối cảnh toàn cầu hiện nay chứng kiến sự bất bình đẳng rõ rệt về mức sống, khả năng tiếp cận y tế và cơ sở hạ tầng, đặc biệt tại những khu vực thường xuyên hứng chịu thiên tai hoặc thiếu thốn điều kiện phát triển. Trong khi đó, các quyết định phân bổ nguồn viện trợ truyền thống vẫn thường dựa vào nhận định chủ quan hay các báo cáo tổng quan thiếu tính hệ thống.

Chính vì thế, việc ứng dụng khoa học dữ liệu đã mở ra hướng tiếp cận khách quan hơn: thông qua các đặc trưng kinh tế, xã hội và y tế, HELP International có thể phân tích và xác định những quốc gia thực sự cần trợ giúp nhiều nhất. Qua đó, quá trình quyết định phân bổ quỹ sẽ trở nên minh bạch, chính xác, và tối ưu hơn. Không chỉ gói gọn trong việc cung cấp số liệu thống kê đơn thuần, khoa học dữ liệu còn giúp tổ chức nhìn sâu hơn vào cấu trúc, mối liên hệ giữa các yếu tố, từ đó tạo ra giá trị thiết thực trong việc hỗ trợ con người và phát triển bền vững.

## 2.2. Nguồn dữ liệu và đặc trưng của dữ liệu

Dữ liệu được sử dụng trong đề tài được tổng hợp từ nhiều nguồn thống kê quốc tế và các báo cáo kinh tế – xã hội đáng tin cậy. Mục tiêu chính là phản ánh bức tranh tổng quan về tình hình phát triển của các quốc gia thông qua những chỉ số liên quan đến y tế, kinh tế và nhân khẩu học. Dưới đây là danh sách các biến quan trọng cùng ý nghĩa của chúng:

* **country**: Tên quốc gia được phân tích.
* **child\_mort**: Tỷ lệ trẻ em dưới 5 tuổi tử vong trên mỗi 1000 trẻ sinh ra còn sống. Chỉ số này phản ánh chất lượng chăm sóc sức khỏe bà mẹ – trẻ em cũng như điều kiện y tế ở quốc gia đó.
* **exports**: Tỷ lệ xuất khẩu hàng hóa và dịch vụ tính trên đầu người, được biểu thị dưới dạng phần trăm GDP bình quân đầu người. Giá trị này thể hiện mức độ hội nhập kinh tế quốc tế và khả năng cạnh tranh của nền kinh tế.
* **health**: Mức chi tiêu cho y tế trên đầu người, tính theo phần trăm GDP bình quân đầu người. Thông qua chỉ số này, ta có thể đánh giá mức độ đầu tư cho hệ thống y tế và tiếp cận dịch vụ khám chữa bệnh.
* **imports**: Tỷ lệ nhập khẩu hàng hóa và dịch vụ trên đầu người, cũng được biểu thị dưới dạng phần trăm GDP bình quân đầu người. Chỉ số này giúp quan sát khả năng phụ thuộc vào nguồn cung quốc tế hoặc nhu cầu nhập khẩu của mỗi quốc gia.
* **Income**: Thu nhập ròng tính trên đầu người. Thu nhập cao thường gắn với mức sống cao hơn và khả năng tiếp cận các dịch vụ xã hội tốt hơn.
* **Inflation**: Tốc độ tăng trưởng hằng năm của Tổng sản phẩm quốc nội (GDP). Lạm phát có thể ảnh hưởng đến ổn định kinh tế vĩ mô và sức mua của người dân.
* **life\_expec**: Tuổi thọ trung bình của trẻ sơ sinh, giả thiết các điều kiện tử vong hiện tại không thay đổi. Đây là một trong những chỉ số chủ chốt để đánh giá chất lượng cuộc sống và điều kiện chăm sóc y tế.
* **total\_fer**: Tổng tỷ suất sinh, tức số con trung bình mà một phụ nữ sinh ra trong suốt độ tuổi sinh sản, nếu các mức sinh hiện tại giữ nguyên. Thông số này cho thấy xu hướng gia tăng dân số tương lai.
* **gdpp**: GDP bình quân đầu người, tính bằng tổng GDP chia cho tổng dân số. Đây là chỉ số kinh tế quan trọng giúp so sánh mức thu nhập và năng lực sản xuất giữa các quốc gia.

Trước khi áp dụng các thuật toán phân cụm, nhóm nghiên cứu tiến hành kiểm tra và làm sạch bộ dữ liệu, loại bỏ những quan sát thiếu hoặc không hợp lệ, đồng thời thực hiện các bước chuẩn hóa cần thiết (ví dụ: xử lý ngoại lai, mã hóa biến) để đảm bảo tính nhất quán. Bộ dữ liệu sau khi được chuẩn hóa sẽ phản ánh khá đầy đủ thực trạng kinh tế – xã hội – y tế của mỗi quốc gia, qua đó tạo nền tảng tin cậy cho các bước phân tích và mô hình hóa tiếp theo.

## 2.3. Hướng tiếp cận: Unsupervised Learning

Trong bối cảnh này, học máy không giám sát (Unsupervised Learning) là lựa chọn phù hợp, bởi mục tiêu của dự án là phát hiện những nhóm quốc gia có chung đặc điểm mà không hề có sẵn nhãn hay phân loại định trước. Cụ thể, kỹ thuật phân cụm (clustering) cho phép hệ thống tìm kiếm các cấu trúc ẩn trong dữ liệu, giúp phân loại các quốc gia thành những cụm khác nhau dựa trên độ tương đồng về kinh tế, xã hội hoặc y tế.

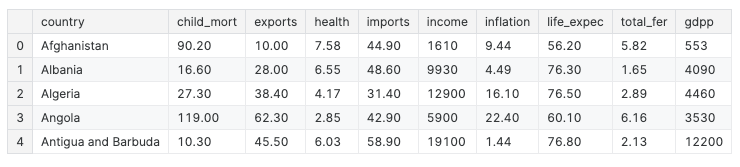
Trong đề tài này, nhóm nghiên cứu đặc biệt tập trung vào hai thuật toán phổ biến khác là DBSCAN và Hierarchical Clustering.

* DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [1] dựa trên khái niệm mật độ điểm, có khả năng phát hiện “nhiễu” và tự động xác định số cụm, phù hợp với bộ dữ liệu có tính phân tán hoặc không gian không đồng nhất.
* Hierarchical Clustering (Phân cụm phân cấp) [2]cung cấp cách tiếp cận theo lối kết hợp dần (agglomerative) hoặc tách dần (divisive) các đối tượng, và thường được biểu diễn trực quan qua dendrogram. Ưu điểm của phương pháp này là dễ quan sát cách các quốc gia được gom nhóm, cũng như linh hoạt thay đổi ngưỡng để quyết định số lượng cụm cuối cùng.

Việc ứng dụng những phương pháp phân cụm đa dạng nhằm mục đích so sánh, đánh giá ưu – nhược điểm của mỗi mô hình, từ đó chọn ra giải pháp phù hợp nhất với cấu trúc dữ liệu và yêu cầu của Tổ chức HELP International.

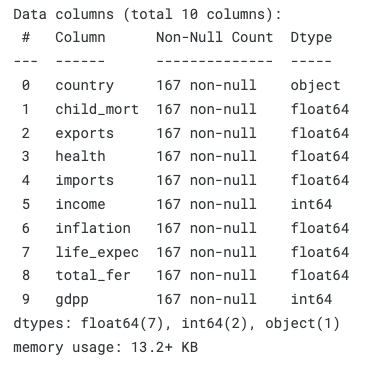
# 3. Khai phá dữ liệu

## 3.1. Thông tin chung về dữ liệu

**

Hình 1: Một số mẫu dữ liệu

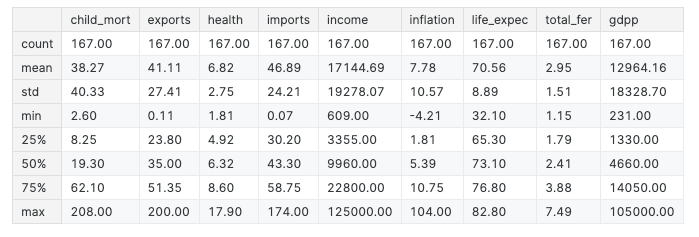
Bộ dữ liệu bao gồm 167 hàng và 10 biến (cột), trong đó có 1 biến dạng string (cột *country*) và 9 biến dạng số (các cột *child\_mort, exports, health, imports, income, inflation, life\_expec, total\_fer, gdpp*). Thông tin này cho thấy mỗi quan sát tương ứng với một quốc gia, kèm theo các chỉ số kinh tế-xã hội như tỉ lệ tử vong trẻ em, xuất khẩu, chi tiêu y tế, thu nhập, lạm phát, tuổi thọ bình quân, v.v.

**

Hình 2: Kiểu dữ liệu của các biến

Đặc điểm kiểu dữ liệu và cấu trúc:

* Cột *country* có kiểu object (string), biểu thị tên quốc gia.
* Các cột còn lại đều là kiểu số (phần lớn là float64, riêng *income* và *gdpp* được ghi nhận dưới dạng int64 hoặc có thể chuyển đổi sang float nếu cần).
* Qua lệnh data.info(), ta thấy không có giá trị khuyết (Non-Null Count = 167 cho mọi biến), đồng nghĩa với việc toàn bộ tập dữ liệu không có ô trống hay lỗi thiếu giá trị.

**

Hình 3: Thống kê sơ bộ dữ liệu

Khi quan sát qua hàm data.describe(), có thể nhận thấy:

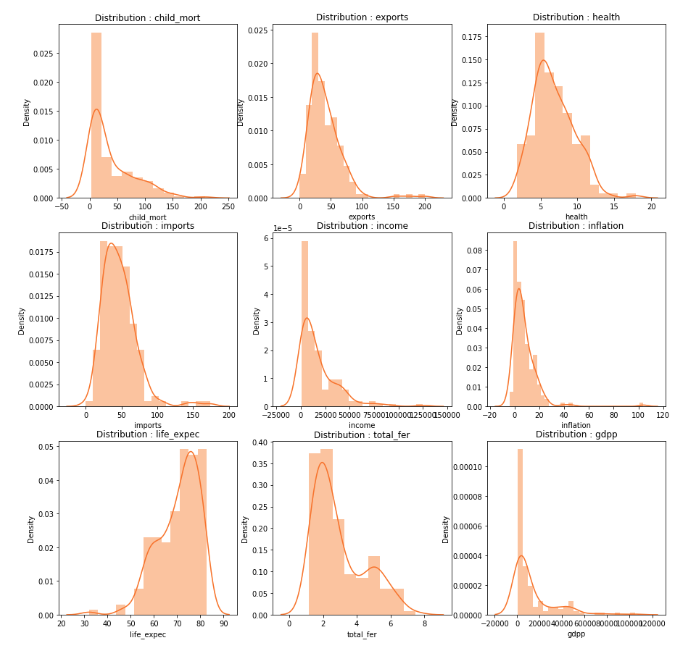
* **child\_mort (tử vong trẻ em)**:
  + Trung bình khoảng 38.27
  + Dao động rất lớn, với giá trị nhỏ nhất là 2.60 và lớn nhất là 208.00
* **gdpp (tổng sản phẩm quốc nội bình quân đầu người)**:
  + Trung bình khoảng 12,964.16
  + Phân bố rất rộng với độ lệch chuẩn hơn 18,000, cho thấy sự chênh lệch đáng kể giữa các quốc gia.
  + Khoảng giá trị từ 231 đến 105,000
* **income (thu nhập)** và một số chỉ số kinh tế khác (*exports, imports, inflation*...) cũng cho thấy sự khác biệt rõ rệt giữa các quốc gia.
* **life\_expec (tuổi thọ bình quân)** trung bình là 70.56, với giá trị thấp nhất là 32.10 và cao nhất là 82.80.

Những thống kê này gợi ý một mức độ phân tán dữ liệu khá lớn giữa các quốc gia, đặc biệt ở khía cạnh kinh tế (GDP/người, thu nhập) và y tế (tử vong trẻ em, tuổi thọ).

Đánh giá bước đầu về cấu trúc và chất lượng dữ liệu:

* Dữ liệu có kích thước nhỏ gọn (167 dòng) và không có giá trị khuyết, do đó thuận tiện cho bước tiền xử lý cũng như cho các bài toán phân tích, mô hình hoá.
* Hầu hết các biến số đều có phạm vi rộng, thể hiện rõ tính đa dạng của các quốc gia về kinh tế, xã hội và y tế. Điều này đòi hỏi việc chuẩn hoá dữ liệu (scaling, normalization) có thể sẽ hữu ích nếu áp dụng các phương pháp phân cụm, hồi quy hoặc phân tích thống kê sâu hơn.

## 3.2. Mô tả thống kê cơ bản



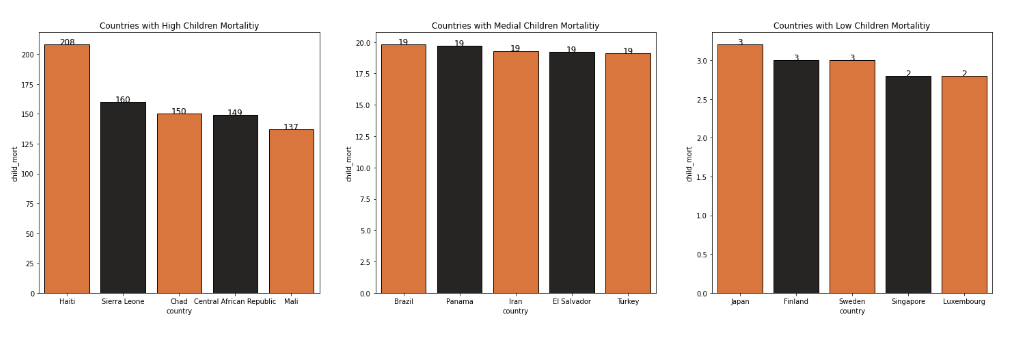
Hình 4: Đồ thị phân phối giá trị của các biến

Khi quan sát đồ thị phân phối (histogram kết hợp đường mật độ) của các biến số, ta nhận thấy phần lớn các biến (như *child\_mort, exports, imports, income, inflation, total\_fer, gdpp*) đều có phân phối lệch phải (*right-skewed*). Điều này cho thấy phần lớn giá trị tập trung ở khoảng tương đối “thấp hoặc trung bình”, trong khi một số giá trị cao hơn hẳn kéo đuôi phân phối về phía phải.

Riêng biến *life\_expec* (tuổi thọ trung bình) có xu hướng lệch trái hoặc phân bố hơi lệch âm (*left/negatively skewed*), nghĩa là phần lớn quốc gia có tuổi thọ ở mức trung bình đến cao (50–80 tuổi), chỉ một số ít nước có tuổi thọ rất thấp. Trong khi đó, biến *health* (chi tiêu y tế, nếu tính theo tỉ lệ % GDP) có phân bố gần chuẩn hơn so với những biến còn lại. Điều này ám chỉ sự tương đối ổn định của tỉ lệ chi tiêu y tế giữa các quốc gia, dù vẫn có một số nước cao hơn hoặc thấp hơn mức trung bình đáng kể.

Nhận xét về các biến và tính đa dạng:

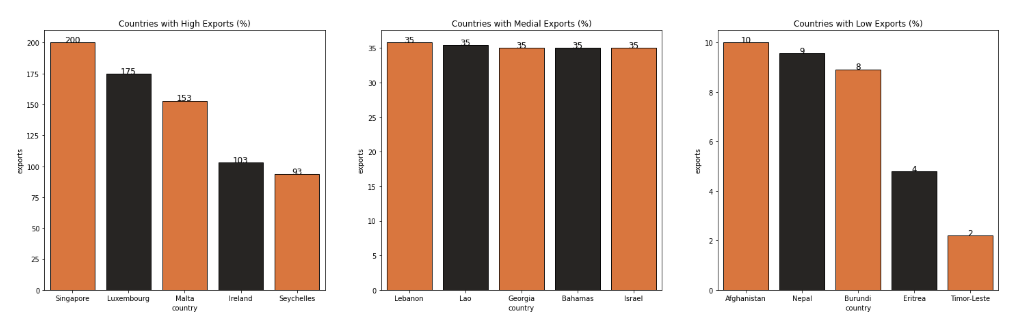
* **child\_mort (tỷ lệ tử vong trẻ em)**: Phân bố lệch phải thể hiện nhiều quốc gia có mức tử vong thấp đến trung bình, nhưng cũng có một nhóm nhỏ chịu cảnh tử vong trẻ em rất cao.
* **income (thu nhập)** và **gdpp (GDP đầu người)**: Biến động lớn và lệch phải nặng, phản ánh rõ rệt khoảng cách phát triển kinh tế giữa các quốc gia giàu (có thu nhập rất cao) và quốc gia nghèo (thu nhập rất thấp).
* **inflation (lạm phát)**: Chủ yếu phân bố quanh mức dưới 10%, nhưng tồn tại một số giá trị âm (giảm phát) hoặc rất cao, cho thấy tình trạng bất ổn ở vài quốc gia.
* **total\_fer (tổng tỉ lệ sinh)**: Tương tự, phần lớn giá trị nằm ở vùng 1–3, nhưng một số quốc gia có tỉ lệ sinh trên 6 gây nên độ lệch phân phối lớn.



Hình 5: Biểu đồ tỷ lệ tử vong trẻ em (child\_mort)

Dựa trên ba biểu đồ thể hiện tỷ lệ tử vong trẻ em (child\_mort) phân thành ba nhóm (cao, trung bình, thấp), ta có thể rút ra một số nhận xét sau:

* 1. **Nhóm có tỷ lệ tử vong trẻ em cao**
* Haiti là quốc gia có tỷ lệ tử vong trẻ em cao nhất trong nhóm, với con số lên tới hơn 200.
* Nhiều nước châu Phi cũng nằm ở mức cao, ví dụ như Sierra Leone, Chad, Cộng hòa Trung Phi và Mali. Điều này cho thấy điều kiện y tế, dinh dưỡng và hạ tầng ở những khu vực này còn nhiều hạn chế, dẫn tới tử vong trẻ em ở mức báo động.
  1. **Nhóm có tỷ lệ tử vong trẻ em trung bình**
     + Những quốc gia như Brazil, Panama, Iran, El Salvador, Thổ Nhĩ Kỳ có tỷ lệ tử vong trẻ em quanh mức từ 18 đến 20.
     + Đây có thể được coi là mức trung gian, nơi điều kiện y tế và phát triển kinh tế - xã hội đã ở mức khá hơn, nhưng vẫn cần cải thiện hơn nữa để tiến tới tỷ lệ tử vong trẻ em thấp hơn.
  2. **Nhóm có tỷ lệ tử vong trẻ em thấp**
     + Các quốc gia như Nhật Bản, Phần Lan, Thụy Điển, Singapore và Luxembourg ghi nhận tỷ lệ tử vong trẻ em từ 2 đến 3, nằm trong nhóm thấp nhất.
     + Đây đều là những nước có nền kinh tế và y tế phát triển, hệ thống chăm sóc sức khỏe toàn diện, điều kiện sống cao, dẫn đến tỷ lệ tử vong trẻ em rất thấp.



Hình 6: Biểu đồ chỉ số xuất khẩu

Dựa trên ba biểu đồ thể hiện chỉ số xuất khẩu, ta có một số nhận xét sau:

1. **Nhóm quốc gia có tỷ lệ xuất khẩu cao**

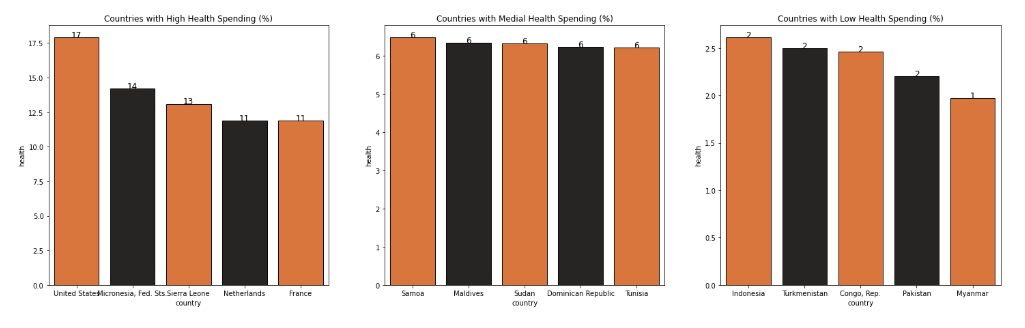
* Singapore, Luxembourg, Malta, Ireland và Seychelles là những nước dẫn đầu về tỷ lệ xuất khẩu.
* Điểm chung của các nước này là sở hữu lợi thế về vị trí địa lý (gần các tuyến giao thương quốc tế), mô hình kinh tế linh hoạt, hoặc nguồn tài nguyên/quy mô thị trường nhỏ nhưng chuyên môn hoá cao.
* Singapore, dù dân số không lớn và nằm trong một khu vực hạn chế về tài nguyên thiên nhiên, lại xây dựng được nền kinh tế dựa trên thương mại và dịch vụ rất phát triển. Tương tự, Luxembourg và Malta cũng tập trung vào dịch vụ tài chính, du lịch và logistics.

1. **Nhóm quốc gia có tỷ lệ xuất khẩu trung bình**

* Biểu đồ cho thấy Lebanon, Lào, Georgia, Bahamas, Israel... (hoặc một vài quốc gia tuỳ theo bảng xếp hạng cụ thể) nằm trong khoảng giá trị trung bình.
* Các nước này thường có tiềm năng xuất khẩu nhất định, nhưng chưa vượt trội do hạn chế về quy mô kinh tế, cơ sở hạ tầng, hoặc phụ thuộc vào tình hình chính trị - xã hội.

1. **Nhóm quốc gia có tỷ lệ xuất khẩu thấp**

* Afghanistan, Nepal, Burundi, Eritrea và Timor-Leste nằm ở nhóm cuối.
* Nhiều quốc gia thuộc nhóm này đối mặt với khó khăn về vị trí địa lý, bất ổn chính trị, hoặc quy mô kinh tế nhỏ. Điều này khiến khả năng sản xuất và tiếp cận thị trường quốc tế bị hạn chế, dẫn đến tỷ lệ xuất khẩu khiêm tốn.



Hình 7: Biểu đồ tỷ lệ chi tiêu cho y tế

Dựa trên ba biểu đồ thể hiện tỷ lệ chi tiêu cho y tế, ta có một số nhận xét sau:

1. **Nhóm có tỷ lệ chi tiêu y tế cao**

* Hoa Kỳ (US) đứng đầu với mức chi tiêu khoảng 17% GDP cho y tế. Điều này phản ánh hệ thống y tế rất phát triển và quy mô chi tiêu lớn của người dân cho chăm sóc sức khỏe.
* Một số quốc gia khác trong nhóm cao như Sierra Leone, Hà Lan (Netherlands), Pháp (France)... cũng duy trì tỷ lệ đầu tư y tế đáng kể, dù không cao bằng Mỹ. Nguyên nhân có thể xuất phát từ chính sách an sinh xã hội, thu nhập bình quân cao, hoặc nhu cầu chăm sóc sức khỏe rộng rãi của người dân.

1. **Nhóm có tỷ lệ chi tiêu y tế trung bình**

* Các nước như Samoa, Maldives, Sudan, Cộng hòa Dominica, Tunisia... có mức chi tiêu y tế nằm xung quanh 6% GDP.
* Đây thường là nhóm quốc gia đang phát triển hoặc có cơ cấu kinh tế–xã hội ở mức trung bình, nên chi tiêu y tế đạt mức vừa phải, chưa thật sự cao nhưng cũng không quá thấp.

1. **Nhóm có tỷ lệ chi tiêu y tế thấp**

* Indonesia, Turkmenistan, Congo, Pakistan, Myanmar thuộc nhóm chi dưới 3% GDP cho y tế.
* Lý do có thể do nguồn lực tài chính hạn chế, quy mô nền kinh tế còn nhỏ, hoặc chính sách ưu tiên chưa tập trung cao vào y tế. Hệ quả là hệ thống y tế có thể gặp nhiều thách thức, thiếu trang thiết bị, nhân lực hoặc dịch vụ chăm sóc sức khỏe đầy đủ cho toàn dân.

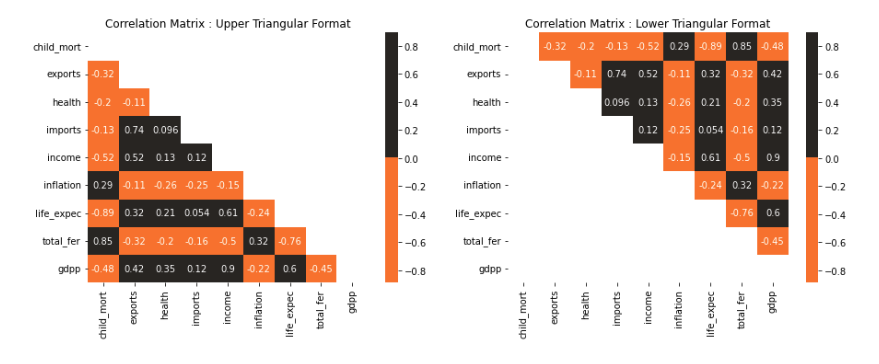
## 3.3. Tóm tắt các phát hiện chính từ EDA

Những phát hiện chính rút ra của nhóm sau quá trình EDA [3]:

1. **Tình trạng y tế và dân số**
   * Các nước châu Phi thường có chỉ số *child\_mort* (tỷ lệ tử vong trẻ em) cao, *life\_expec* (tuổi thọ trung bình) thấp và *total\_fer* (tỷ lệ sinh) cao. Đây là những dấu hiệu nghiêm trọng về điều kiện y tế và mức sống.
   * Haiti nổi bật với tỷ lệ tử vong trẻ em cao nhất, nhấn mạnh nhu cầu cấp thiết về hỗ trợ y tế và cơ sở hạ tầng.
2. **Chi tiêu y tế**
   * Mỹ (US) là quốc gia chi tiêu nhiều nhất cho y tế, nhưng không nằm trong nhóm dẫn đầu về tuổi thọ.
   * Điều đáng chú ý là không có nước nào nằm cả trong top 5 quốc gia có tuổi thọ cao nhất lẫn top 5 quốc gia chi tiêu y tế nhiều nhất.
   * Các nước châu Á có xu hướng chi tiêu y tế ở mức thấp hơn, trong khi châu Phi thường thiếu hụt nguồn lực y tế và phải đối mặt với nhiều gánh nặng bệnh tật hơn.
3. **Xuất nhập khẩu**
   * Singapore, Malta, Luxembourg, Seychelles (và đôi khi cả Qatar) góp mặt trong nhóm top 5 về cả xuất khẩu lẫn nhập khẩu. Vị trí địa lý và chiến lược kinh tế linh hoạt đóng vai trò quan trọng trong việc thúc đẩy thương mại.
   * Sudan là quốc gia châu Phi duy nhất có mức nhập khẩu thấp, trong khi Brazil là nước thấp nhất về nhập khẩu trong toàn bộ dữ liệu.
4. **Lạm phát (inflation)**
   * Các quốc gia châu Phi nhìn chung có xu hướng lạm phát cao hơn so với các khu vực khác. Lạm phát cao kéo theo nhiều hệ luỵ kinh tế – xã hội, làm giảm sức mua và gây bất ổn.
5. **Thu nhập và GDP**
   * Qatar sở hữu mức thu nhập cao nhất thế giới, theo sau là Singapore và Luxembourg.
   * Luxembourg cùng Qatar cũng nằm trong nhóm quốc gia dẫn đầu về GDP bình quân đầu người (gdpp).
   * Ngược lại, nhiều nước châu Phi nằm trong nhóm cuối cả về *income* lẫn *gdpp*, chịu ảnh hưởng nặng nề từ lịch sử thuộc địa và điều kiện phát triển còn hạn chế.

Tựu trung, EDA cho thấy sự chênh lệch rõ rệt về phát triển kinh tế – xã hội giữa các khu vực trên thế giới. Các chỉ số về tử vong trẻ em, tuổi thọ, thu nhập và GDP đều phản ánh khoảng cách phát triển lớn, đặc biệt giữa nhóm nước giàu (châu Âu, Bắc Mỹ, một số nước châu Á) và những quốc gia nghèo, bất ổn (nhiều nước châu Phi).

# 4. Tiền xử lý dữ liệu & Feature Engineering - Hiệp

****

Hình 8: Ma trận tương quan

Dựa vào ma trận tương quan (Correlation Matrix), ta nhận thấy:

* child\_mort (tử vong trẻ em) có tương quan âm rõ rệt với các biến kinh tế như *income, gdpp, exports*. Ngược lại, nó có tương quan dương với *inflation* và *total\_fer*.
* life\_expec (tuổi thọ bình quân) cho thấy tương quan âm với *child\_mort* và *total\_fer*, nhưng tương quan dương với *income* và *gdpp*.
* income và gdpp có tương quan rất cao (lên đến ~0.9), phản ánh rằng quốc gia có GDP đầu người cao thường đồng thời có thu nhập (per capita income) cao.

Từ việc quan sát này, nhóm phân tích đã đi đến đề xuất chia các biến thành 3 nhóm (categories) theo chủ đề:

1. Nhóm Y tế - Dân số (health): *child\_mort, health, life\_expec, total\_fer*
2. Nhóm Thương mại (trade): *exports, imports*
3. Nhóm Tài chính (finance): *income, inflation, gdpp*

Việc phân nhóm dựa trên ý nghĩa kinh tế - xã hội của từng biến (thay vì chỉ dựa trên tương quan) nhằm mục đích:

* Quản lý và trực quan hóa: Dễ phân tích và so sánh các biến có cùng chủ đề.
* Xử lý đặc thù: Áp dụng các phương pháp chuẩn hóa hay biến đổi riêng theo từng nhóm biến.

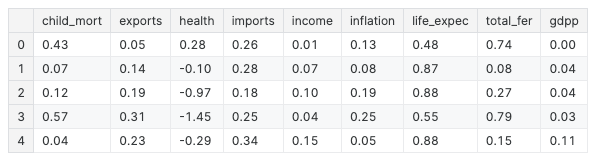
Ở bước đầu tiên, nhóm tạo ra ba biến mới đại diện cho ba khía cạnh kinh tế – xã hội: Health, Trade và Finance. Cách tiếp cận là lấy các cột liên quan (như *child\_mort, health, life\_expec, total\_fer* cho nhóm “Health”), rồi chia giá trị của từng biến cho giá trị trung bình của chính biến đó (một dạng chuẩn hóa cơ bản), sau đó cộng tổng để tạo thành chỉ số gộp.

Tương tự, nhóm “Trade” bao gồm *imports, exports* và nhóm “Finance” gộp *income, inflation, gdpp*. Kết quả là DataFrame với ba cột: Health, Trade, Finance, cho phép thu gọn thông tin từ nhiều cột ban đầu thành ba đặc trưng chính.

****

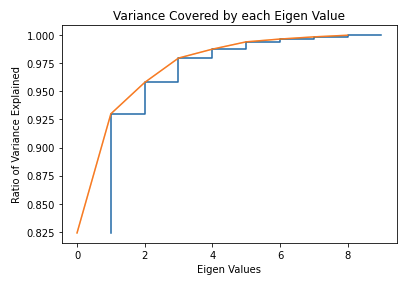
Hình 9: Bảng chuẩn hoá giá trị

Tiếp theo, nhóm sử dụng MinMaxScaler() để chuẩn hóa ba biến Health, Trade và Finance về thang giá trị [0,1]. Kết quả được minh họa bằng bảng với các cột: *Country*, *Health*, *Trade*, *Finance*. Nhờ quá trình này, mỗi chỉ số (Health, Trade, Finance) được đưa về cùng một khoảng, giúp mô hình học máy không bị ảnh hưởng bởi sự chênh lệch đơn vị đo lường hay độ lớn khác nhau của dữ liệu gốc.

****

Hình 10: Bảng chuẩn hoá giá trị

Sau khi chuẩn hóa dữ liệu, ta áp dụng PCA để giảm số chiều.

****

Hình 11: Trực quan hóa PCA

Biểu đồ đường và cột (step plot) cho thấy tỷ lệ phương sai tích lũy theo số lượng thành phần chính. Đường màu cam thể hiện giá trị cộng dồn của tỷ lệ phương sai, còn cột xanh thể hiện giá trị riêng lẻ của mỗi thành phần. Quan sát biểu đồ, ta thấy chỉ cần khoảng 2–3 thành phần chính đã có thể nắm giữ trên 95% phương sai của tập dữ liệu. Từ đó, ta xác định số chiều tối ưu để giữ lại (thường dựa trên ngưỡng ≥95% phương sai tích lũy).

Cuối cùng, dựa trên ngưỡng phương sai mong muốn, ta giữ lại những cột tương ứng với các thành phần chính có phương sai cao (chẳng hạn cột 0, 1, 2) và loại bỏ các thành phần có mức đóng góp phương sai không đáng kể (các cột từ 3 trở đi). Kết quả là bảng mới chỉ còn 3 cột, mỗi cột tượng trưng cho một thành phần chính (PC1, PC2, PC3). Việc giảm chiều dữ liệu như vậy vừa giúp đơn giản hóa mô hình, vừa có thể cải thiện hiệu suất và giảm hiện tượng overfitting khi triển khai các thuật toán học máy hoặc phân cụm.

# 5. Các thuật toán phân cụm

## 5.1. Giới thiệu chung về học không giám sát (Unsupervised Learning) và bài toán phân cụm

Máy học không có giám sát là khi bạn cung cấp dữ liệu đầu vào thuật toán mà không có bất kỳ dữ liệu đầu ra được gắn nhãn nào. Sau đó, tự nó, thuật toán xác định các mẫu và mối quan hệ trong và giữa dữ liệu. Tiếp theo là một số loại kỹ thuật học không có giám sát.

1. ***Phân cụm***

Kỹ thuật học không có giám sát phân cụm sẽ nhóm các đầu vào dữ liệu nhất định lại với nhau, vì vậy chúng có thể được phân loại thành một tổng thể. Có nhiều loại thuật toán phân cụm khác nhau phụ thuộc vào dữ liệu đầu vào. Một ví dụ về phân cụm là xác định các loại lưu lượng truy cập mạng khác nhau để dự đoán các sự cố bảo mật tiềm ẩn.

1. ***Học quy tắc liên kết***

Các kỹ thuật học quy tắc liên kết sẽ khám phá các mối quan hệ dựa trên quy tắc giữa các đầu vào trong tập dữ liệu. Ví dụ, thuật toán Apriori tiến hành phân tích giỏ thị trường để xác định các quy tắc như cà phê và sữa thường được mua cùng nhau.

1. ***Mật độ xác suất***

Các kỹ thuật mật độ xác suất trong học không có giám sát sẽ dự đoán khả năng hoặc tiềm năng của giá trị đầu ra nằm trong phạm vi của những điều được coi là bình thường đối với đầu vào. Ví dụ, một nhiệt kế trong phòng máy chủ thường ghi dữ liệu giữa một phạm vi mức độ nhất định. Tuy nhiên, nếu nhiệt kế đột ngột đo một số thấp dựa trên phân phối xác suất, nó có thể chỉ ra sự cố thiết bị.

1. ***Giảm kích thước***

Giảm kích thước là một kỹ thuật học không có giám sát giúp giảm số lượng các tính năng trong tập dữ liệu. Nó thường được sử dụng để xử lý trước dữ liệu cho các chức năng máy học khác và giảm độ phức tạp và tổng chi phí. Ví dụ: nó có thể làm mờ hoặc cắt các tính năng nền trong ứng dụng nhận dạng hình ảnh.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) là một thuật toán học không có giám sát thuộc nhóm phân cụm (**Clustering**).

DBSCAN hoạt động dựa trên mật độ của các điểm dữ liệu, giúp phát hiện cụm dữ liệu có hình dạng bất kỳ và có thể xử lý dữ liệu nhiễu hiệu quả. Thuật toán xác định cụm bằng cách tìm kiếm các vùng có mật độ điểm cao và mở rộng cụm từ những điểm lõi (core points).

## 5.2. DBSCAN

### 5.2.1. Giới Thiệu

Trong lĩnh vực Khai phá Dữ liệu (Data Mining) và Học máy (Machine Learning), bài toán phân cựm (clustering) giữ vai trò quan trọng trong việc phát hiện cấu trúc ẩn trong dữ liệu không có nhãn (unsupervised learning). Tuy nhiên, các phương pháp phân cụm truyền thống như K-Means hoặc Hierarchical Clustering thường giả định rằng các cụm có cấu trúc hình cầu và yêu cầu xác định trước số cụm. Tuy nhiên, những giả định này không luôn phù hợp với các bối cảnh dữ liệu phức tạp hoặc chứa nhiều nhiễu (outliers).

Nhằm giải quyết những hạn chế trên, thuật toán DBSCAN (Density-Based Spatial Clustering of Applications with Noise) đã được đề xuất bởi Martin Ester và Hans-Peter Kriegel vào năm 1996. DBSCAN cung cấp một cách tiếp cận dựa trên mật độ để phát hiện các cụm mà không cần biết trước số lượng cụm.

### 5.2.2. Định Nghĩa Thuật Toán DBSCAN

DBSCAN là một thuật toán phân cụm dựa trên mật độ, trong đó các điểm dữ liệu được nhóm lại với nhau nếu chúng thuộc cùng một vùng có mật độ cao, trong khi các điểm dữ liệu không đạt ngưỡng mật độ sẽ bị xác định là nhiễu. Cách định nghĩa này giúp DBSCAN không phụ thuộc vào các giả định hình dạng cụm như K-Means.

### 5.2.3. Các Tham Số Chính Của DBSCAN

Thuật toán DBSCAN hoạt động dựa trên hai tham số chính:

* **Eps (ε) - Bán kính lân cận**: Xác định vùng lân cận của mỗi điểm dữ liệu.
  + Nếu ε quá lớn, các cụm dễ bị nhập nhầm vào nhau.
  + Nếu ε quá nhỏ, cụm có thể bị phân mảnh.
* **MinPts - Số điểm tối thiểu**: Xác định số lượng điểm tối thiểu trong vùng lân cận ε để một điểm được coi là điểm lõi (core point).
  + MinPts quá lớn → Số cụm ít đi, nhiều điểm bị xem là nhiễu.
  + MinPts quá nhỏ → Tạo ra nhiều cụm nhỏ không đầy ý nghĩa.

### 5.2.4. Các Đặc Điểm Của DBSCAN

DBSCAN có một số ưu điểm sau:

* **Tự động xác định số cụm**: DBSCAN không yêu cầu xác định trước số cụm, mà tìm các cụm dựa trên mật độ.
* **Xử lý được cụm có hình dạng bất kỳ**: Không giới hạn với các cụm hình tròn hoặc lồi.
* **Phát hiện và xử lý nhiễu**: Tách các điểm nhiễu ra khỏi các cụm.
* **Phân loại dữ liệu thành 3 nhóm**: Điểm lõi (Core Points), Điểm biên (Border Points), Điểm nhiễu (Noise Points).

Tuy nhiên, DBSCAN không hoạt động tốt khi dữ liệu có mật độ thay đổi lớn hoặc trong không gian nhiều chiều.

### 5.2.5. Hoạt Động Của DBSCAN

Thuật toán DBSCAN phân loại các điểm dữ liệu thành ba nhóm chính:

* **Core Point (Điểm lõi)**: Điểm có ít nhất **MinPts** điểm trong vùng lân cận **ε**, bao gồm chính nó.
* **Border Point (Điểm biên)**: Điểm nằm trong vùng lân cận của một **core point**, nhưng không đủ **MinPts** để trở thành điểm lõi.
* **Noise Point (Điểm nhiễu)**: Điểm không thuộc bất kỳ cụm nào.

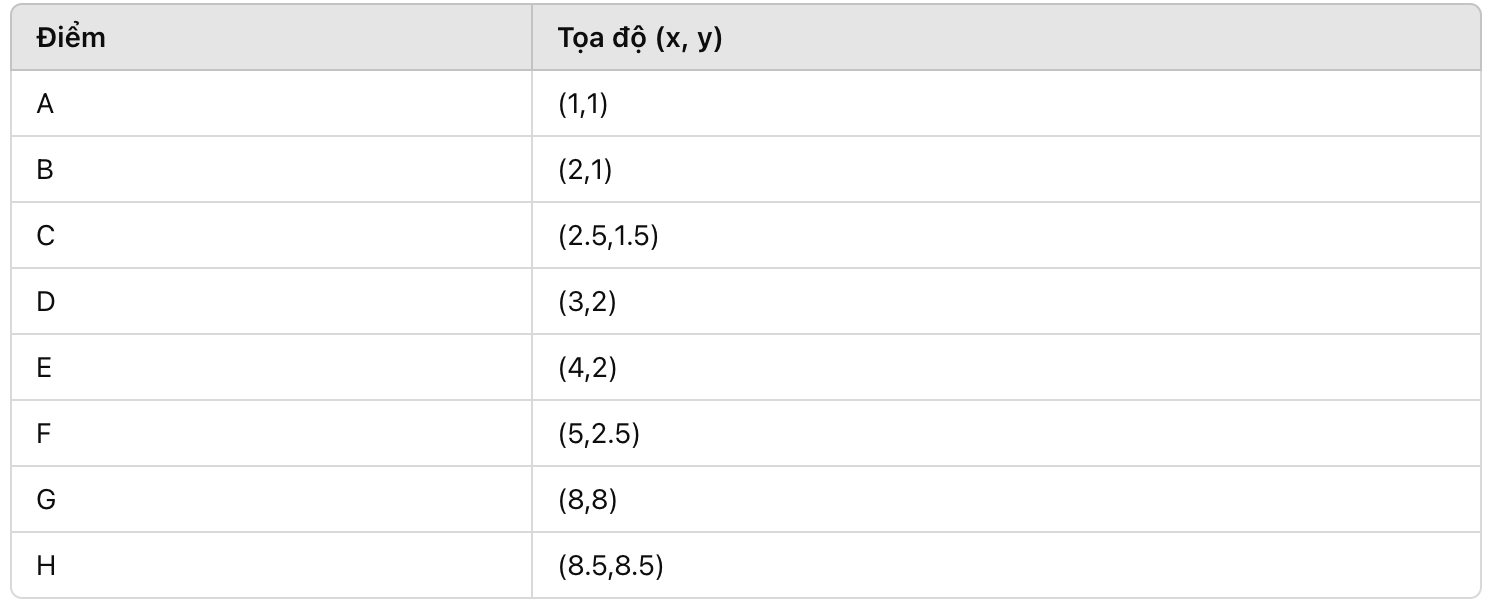
DBSCAN hoạt động theo các bước sau:

|  |
| --- |
| **Bước 1: Chọn một điểm chưa được thăm**.  **Bước 2: Tìm tất cả các điểm trong vùng lân cận** của điểm đó bằng bán kính **ε**.  **Bước 3: Xác định loại điểm**:   * Nếu số lượng điểm lân cận **≥ MinPts**:   + Đánh dấu điểm là **core point**.   + Mở rộng cụm bằng cách thêm các điểm lân cận vào cụm hiện tại.   + Tiếp tục mở rộng cho đến khi không còn điểm nào có thể được thêm vào cụm. * Nếu số lượng điểm lân cận **< MinPts**:   + Đánh dấu điểm là **noise point**.   + Nếu sau này điểm này nằm trong vùng lân cận của một **core point**, nó có thể trở thành **border point**.   **Bước 4: Lặp lại quá trình** cho đến khi tất cả các điểm dữ liệu được xét. |

### 5.2.6. Ví Dụ Minh Họa DBSCAN

1. ***Tập dữ liệu giả định***

Giả sử ta có một tập dữ liệu gồm **8 điểm** trong không gian 2D:



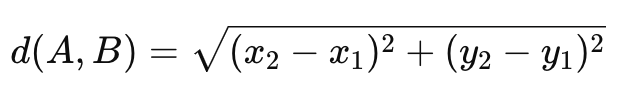
Hình 12: Ví dụ điểm dữ liệu cho thuật toán DBSCAN

Chọn tham số:

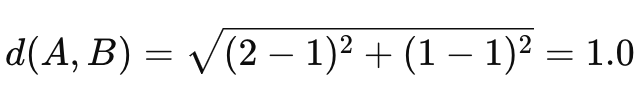
* **Eps (ε) ≈ 2.69** (dựa trên đồ thị k-distance với "elbow point").
* **MinPts ≥ D + 1**, với D là số chiều (**D = 2**), chọn **MinPts = 3**.

1. ***Tính toán khoảng cách Euclidean***

Khoảng cách giữa các điểm được tính theo công thức:



Ví dụ, khoảng cách từ **A(1,1)** đến **B(2, 1)**:



Bảng khoảng cách giữa các điểm với bán kính **Eps = 1.5**:



Hình 13: Tính khoảng cách tới các điểm

Xác định điểm lõi, điểm biên và điểm nhiễu

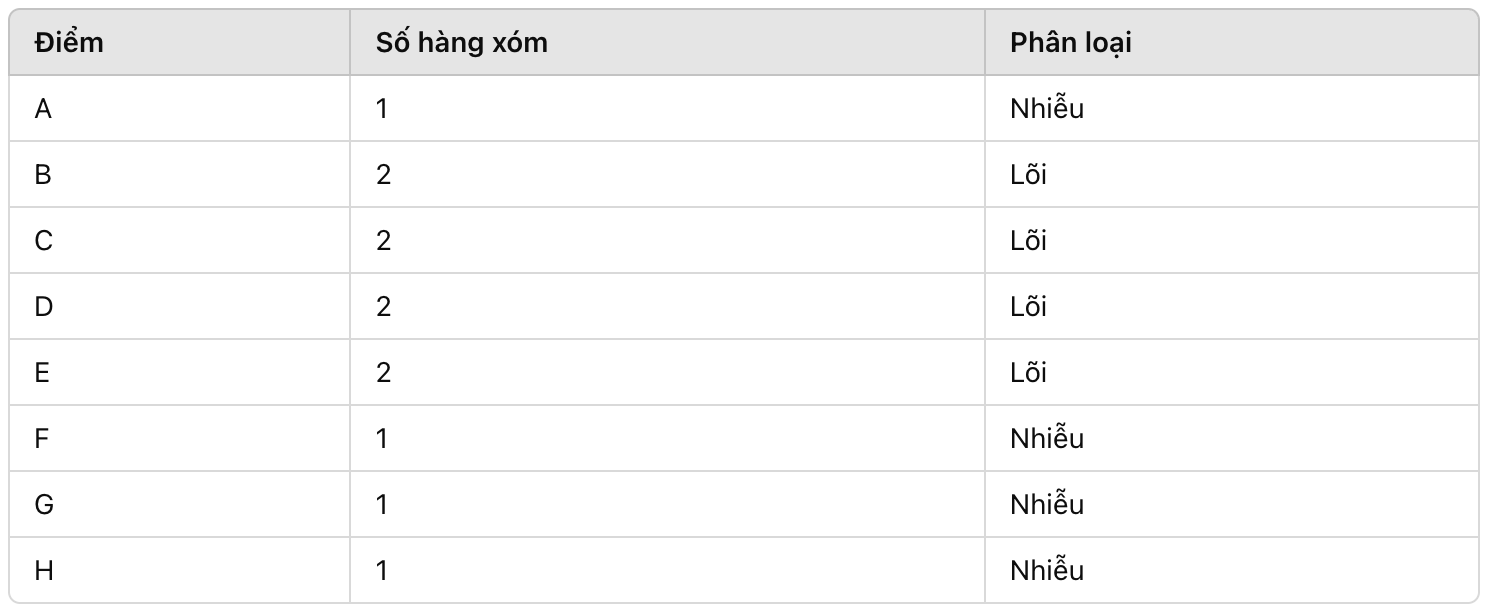


Hình 14: Phân loại các điểm

**Kết quả**: Không có điểm nào có **≥ MinPts = 3**, nên tất cả các điểm đều bị coi là **nhiễu**. DBSCAN thất bại với bộ tham số này!

1. ***Điều chỉnh tham số***

Nếu ta **giảm MinPts = 2**, kết quả sẽ thay đổi:



Hình 15: Giảm minPts và phân loại điểm

**Kết qủa:**

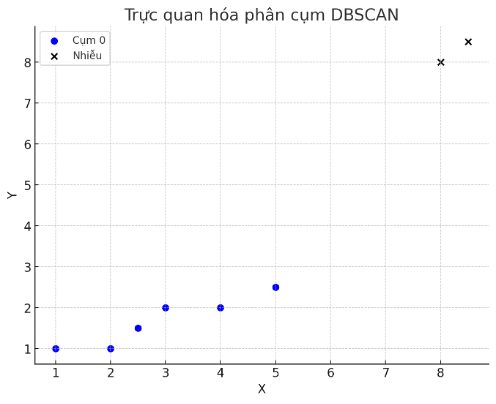
* **Cụm 1**: {B, C, D, E}
* **Nhiễu**: {A, F, G, H}

Như vậy, việc **điều chỉnh tham số phù hợp** sẽ giúp DBSCAN phát hiện được cụm dữ liệu chính xác hơn.

Ví dụ trên cho thấy:

* Việc chọn tham số ε và MinPts có ảnh hưởng lớn đến kết quả phân cụm.
* DBSCAN có thể thất bại nếu tham số không phù hợp.
* Việc điều chỉnh MinPts và Eps có thể cải thiện kết quả phân cụm.
* DBSCAN hiệu quả trong việc xá**c** định các cụm dữ liệu có hình dạng tự do mà không yêu cầu số cụm đầu vào.

**Kết quả thể hiện bằng biểu đồ:**



Hình 16: Trực quan phân cụm DBSCAN

**5.2.7. Đánh Giá Hiệu Suất Của Thuật Toán DBSCAN**

Thuật toán DBSCAN sở hữu nhiều ưu điểm nổi bật so với các phương pháp phân cụm truyền thống, đặc biệt trong môi trường dữ liệu phức tạp và không có cấu trúc rõ ràng.

* Không cần xác định số cụm trước: Không giống như thuật toán **K-Means**, DBSCAN không yêu cầu chỉ định số cụm **K** trước khi chạy thuật toán. Thay vào đó, DBSCAN tự động xác định cụm dựa trên **mật độ của dữ liệu**, giúp tránh tình trạng lựa chọn sai số lượng cụm, một vấn đề phổ biến trong K-Means.
* Khả năng phát hiện cụm có hình dạng bất kỳ: Trong khi K-Means giả định rằng các cụm có dạng hình cầu, DBSCAN có thể xác định **các cụm có hình dạng bất kỳ**, chẳng hạn như **cụm dài, cụm cong, hoặc cụm nhánh**. Điều này làm cho DBSCAN trở thành lựa chọn phù hợp trong các bài toán phân cụm dữ liệu thực tế, nơi dữ liệu thường không tuân theo cấu trúc hình cầu.
* Xử lý tốt dữ liệu có nhiễu và outliers: DBSCAN có **cơ chế phát hiện và xử lý nhiễu (outliers)** bằng cách không gán các điểm nhiễu vào bất kỳ cụm nào. Điều này khác với K-Means, vốn buộc phải gán mọi điểm dữ liệu vào một cụm, kể cả khi điểm đó thực chất là nhiễu.
* Hoạt động tốt với dữ liệu có mật độ không đồng đều: Do DBSCAN phân cụm dựa trên **mật độ điểm**, nó có thể nhận diện các cụm có **mật độ khác nhau**, giúp nó hoạt động tốt hơn so với K-Means, vốn dễ bị ảnh hưởng khi các cụm có kích thước và mật độ khác nhau.
* Không bị ảnh hưởng bởi việc khởi tạo ban đầu: Trong khi **K-Means phụ thuộc vào việc chọn tâm cụm ban đầu**, DBSCAN không yêu cầu khởi tạo trước, giúp tránh rơi vào **cực tiểu cục bộ**. Điều này giúp thuật toán hoạt động ổn định hơn trong nhiều trường hợp.

Mặc dù DBSCAN mang lại nhiều lợi ích, nhưng nó cũng tồn tại một số hạn chế cần được cân nhắc khi áp dụng vào thực tế.

* Nhạy cảm với tham số Eps và MinPts
  + **Eps (ε) và MinPts là hai tham số quan trọng** quyết định cách thuật toán xác định cụm.
  + Nếu chọn **ε quá nhỏ**, thuật toán có thể phân loại nhiều điểm là nhiễu, khiến số lượng cụm bị giảm hoặc không có cụm nào được hình thành.
  + Nếu chọn **ε quá lớn**, nhiều cụm có thể bị gộp chung lại, làm mất đi tính chính xác trong phân cụm.
  + Do đó, cần sử dụng **phương pháp phân tích đồ thị k-distance** để xác định **ε tối ưu**.
* Gặp khó khăn với dữ liệu có mật độ thay đổi lớn. DBSCAN giả định rằng các cụm có **mật độ tương đối đồng nhất**. Khi dữ liệu chứa **cụm dày đặc xen kẽ với cụm thưa thớt**, thuật toán có thể không phân tách được các cụm chính xác.
* Không hoạt động tốt với dữ liệu nhiều chiều (high-dimensional data)
  + Trong không gian nhiều chiều (**high-dimensional space**), **khoảng cách Euclidean trở nên kém ý nghĩa**, dẫn đến việc xác định cụm trở nên không hiệu quả.
  + Khi số chiều cao, các điểm dữ liệu có xu hướng trở nên đồng đều về khoảng cách, gây khó khăn cho việc xác định vùng mật độ cao..
* Độ phức tạp thuật toán và hiệu suất trên tập dữ liệu lớn
  + DBSCAN có **độ phức tạp trung bình là** O(n**log**n), nhưng trong **trường hợp xấu nhất**, độ phức tạp có thể lên đến **O(n2)**.
  + Điều này khiến thuật toán **không hiệu quả trên tập dữ liệu lớn** (vài triệu điểm trở lên).

Tóm lại, DBSCAN là một thuật toán mạnh mẽ khi **phân cụm dữ liệu phức tạp, không có số cụm xác định trước, hoặc chứa nhiều nhiễu**. Tuy nhiên, nó **kém hiệu quả với dữ liệu nhiều chiều** và **không phù hợp với tập dữ liệu lớn** nếu không có cải tiến. Việc lựa chọn tham số **Eps** và **MinPts** đúng cách là **yếu tố quan trọng nhất** quyết định chất lượng phân cụm.

### 5.2.8. Ứng Dụng Thực Tế Của DBSCAN Trong Phân Cụm Quốc Gia

Phân nhóm các quốc gia dựa trên chỉ số kinh tế - xã hội là một nhiệm vụ quan trọng trong hoạch định chính sách và phân bổ nguồn lực. Việc sử dụng thuật toán DBSCAN trong phân cụm giúp phát hiện các nhóm quốc gia có đặc điểm tương đồng, đồng thời nhận diện các quốc gia có điều kiện đặc biệt (outliers) để có chính sách hỗ trợ phù hợp. Với mục tiêu:

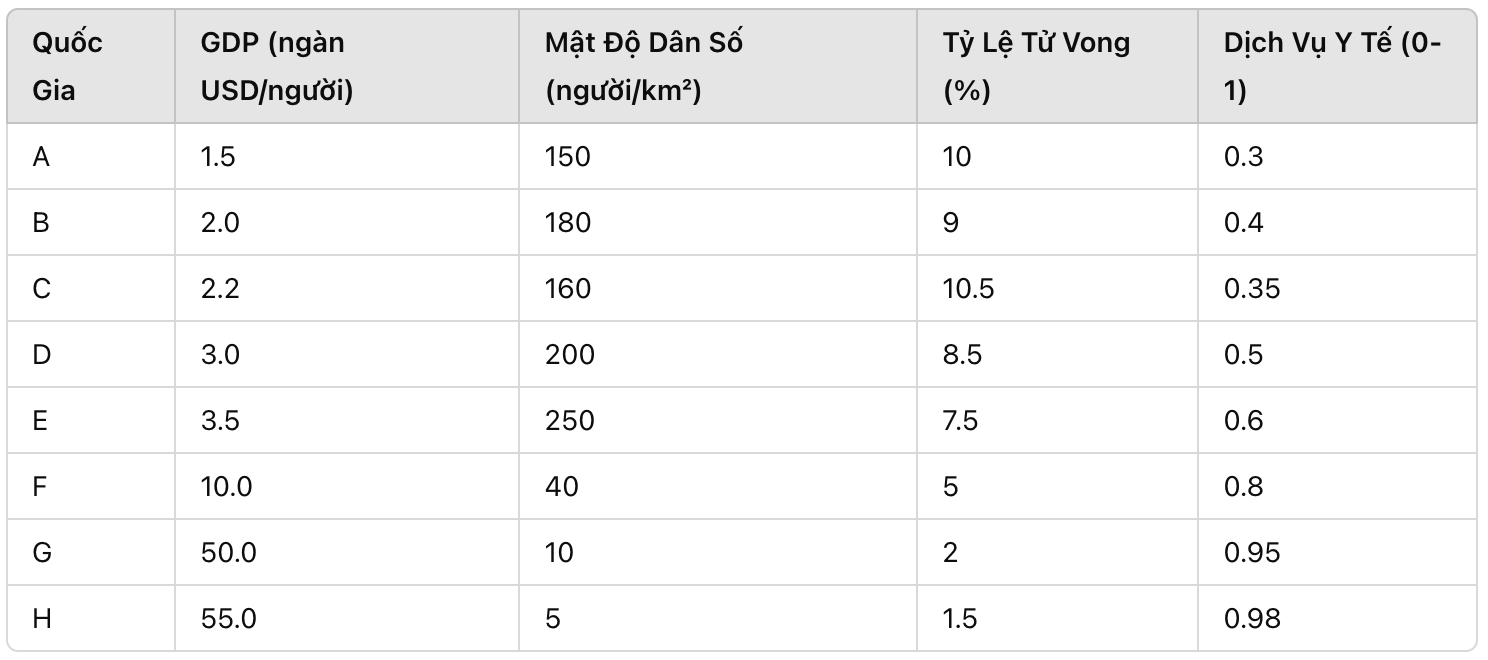
* Nhóm các quốc gia có tình trạng kinh tế - xã hội tương đồng để ưu tiên hỗ trợ phát triển.
* Phát hiện các quốc gia ngoại lệ có điều kiện đặc biệt để đưa ra các chính sách hỗ trợ khác biệt.
* Ứng dụng thuật toán DBSCAN để tự động phân cụm, thay vì phải xác định số lượng cụm trước.

Quy Trình Ứng Dụng DBSCAN:

* Bước 1: Phân Nhóm Các Quốc Gia Dựa Trên Chỉ Số Xã Hội - Kinh Tế. Các chỉ số được lựa chọn để phân nhóm quốc gia:
  + GDP bình quân đầu người (USD/người/năm) - phản ánh mức sống kinh tế.
  + Mật độ dân số (người/km²) - cho thấy áp lực dân số lên tài nguyên.
  + Tỷ lệ tử vong (%) - đánh giá rủi ro về y tế.
  + Mức độ tiếp cận dịch vụ y tế (chỉ số từ 0 đến 1) - phản ánh khả năng chăm sóc sức khỏe.
  + Tỷ lệ nghèo đói, chỉ số phát triển con người (HDI) (nếu có thể thu thập).

Mục tiêu: Xây dựng mô hình phân cụm các quốc gia dựa trên các chỉ số trên để hỗ trợ ra quyết định chính sách phát triển bền vững.

Áp Dụng Thuật Toán DBSCAN Để Phân Cụm tập dữ liệu mẫu. Giả sử tập dữ liệu bao gồm 8 quốc gia với các chỉ số kinh tế - xã hội như sau:

****

Hình 17: Tập dữ liệu mẫu

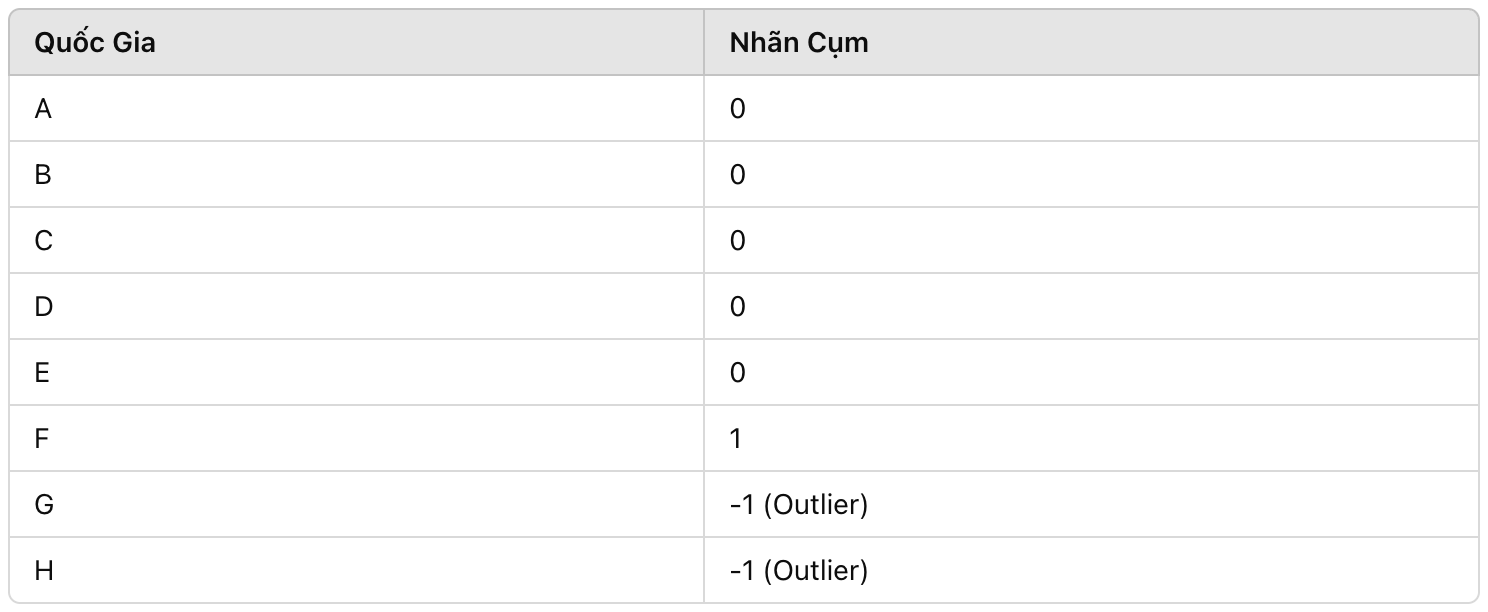
Nhận xét sơ bộ:

* + Nhóm A, B, C, D, E có GDP thấp, mật độ dân số cao và tỷ lệ tử vong cao → Có thể thuộc cùng một cụm.
  + F có điều kiện kinh tế và y tế tốt hơn nhưng chưa phải là nhóm phát triển cao nhất.
  + G và H có GDP rất cao, mật độ dân số thấp và dịch vụ y tế phát triển → Khả năng cao là outliers.

Lựa chọn tham số DBSCAN

* Eps (ε - bán kính lân cận): Dựa trên k-distance plot, chọn ε = 2.5.
* MinPts (số điểm tối thiểu để xác định cụm): Dữ liệu có 4 chỉ số, do đó chọn MinPts = 4.

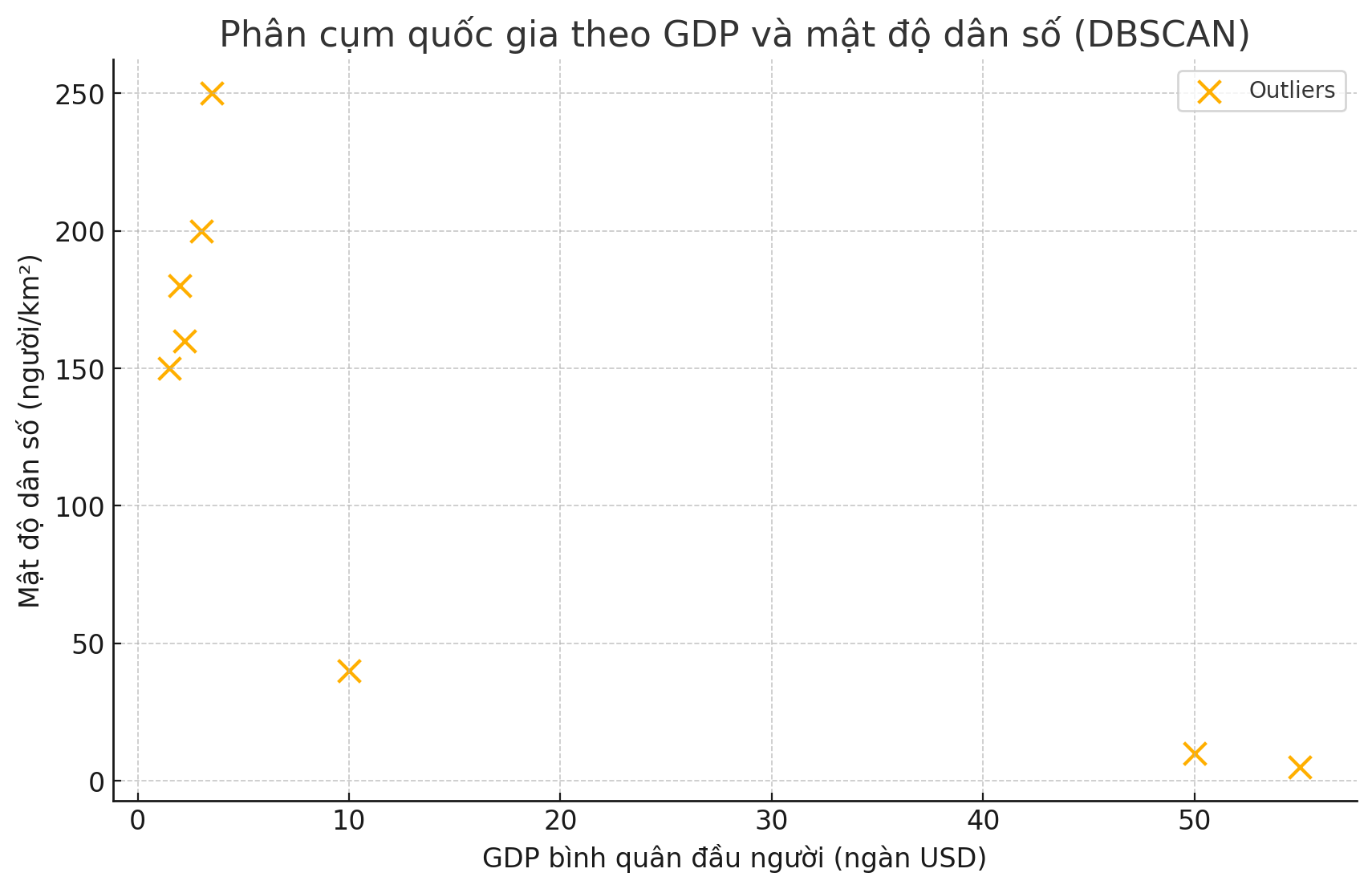
Kết Quả Phân Cụm DBSCAN. Sau khi áp dụng DBSCAN, ta có kết quả sau:



Hình 18: Kết quả phân cụm DBSCAN

Nhận xét kết quả:

* Cụm 0: Gồm các quốc gia A, B, C, D, E → Nhóm cần hỗ trợ mạnh, do có mức GDP thấp, dân số cao và tỷ lệ tử vong cao.
* Cụm 1: Quốc gia F có điều kiện tốt hơn, nhưng chưa đạt mức phát triển cao nhất → Hướng tới phát triển bền vững.
* G & H là outliers: Có GDP cao, dân số thấp và dịch vụ y tế phát triển vượt trội → Không cần trợ cấp, có thể là nhóm quốc gia phát triển.



Hình 19: Biểu đồ kết quả phân cụm DBSCAN

### 5.2.9. Tổng kết lại về thuật DBSCAN:

DBSCAN là một thuật toán phân cụm mạnh mẽ, đặc biệt hữu ích khi xử lý dữ liệu có nhiễu, có mật độ thay đổi và không cần biết trước số cụm.Tuy nhiên, nó cần được sử dụng một cách thận trọng, với việc chọn **Eps và MinPts hợp lý** để đạt hiệu quả cao nhất.

**Khi nào nên dùng DBSCAN:**

* Khi **không biết trước số cụm**.
* Khi **các cụm có hình dạng không chuẩn**.
* Khi **dữ liệu có nhiễu hoặc outliers**.
* Khi **cụm có mật độ đồng đều hoặc khác biệt nhưng có thể điều chỉnh tham số hợp lý**.

**Khi nào không nên dùng DBSCAN:**

* Khi **dữ liệu có nhiều chiều (high-dimensional data)**.
* Khi **dữ liệu có cụm với mật độ quá khác biệt**.
* Khi **dữ liệu quá lớn mà không có tối ưu hóa**.

## 5.3. Hierarchical Clustering

Tư tưởng phân cụm phân cấp: nhóm dần các đối tượng (agglomerative) hoặc tách dần cụm (divisive).

Phân cụm phân cấp là một phương pháp phân cụm không giám sát trong học máy, nhằm tổ chức dữ liệu thành các cụm dựa trên mức độ tương đồng giữa các đối tượng. Có hai cách tiếp cận chính trong phân cụm phân cấp:

* **Agglomerative (Phương pháp kết tụ - từ dưới lên):**
  + - Bắt đầu với mỗi đối tượng (hoặc điểm dữ liệu) được xem là một cụm riêng lẻ.
    - Sau đó, các cụm gần nhau nhất (dựa trên một thước đo khoảng cách như Euclidean) được ghép dần thành các cụm lớn hơn.
    - Quá trình này tiếp tục cho đến khi tất cả các đối tượng thuộc về một cụm duy nhất hoặc đạt được tiêu chí dừng.
* **Divisive (Phương pháp phân chia - từ trên xuống):**
  + - Bắt đầu với toàn bộ tập dữ liệu được xem như một cụm lớn duy nhất.
    - Sau đó, cụm này được chia nhỏ dần thành các cụm nhỏ hơn dựa trên sự khác biệt giữa các đối tượng.
    - Quá trình tiếp tục cho đến khi mỗi đối tượng trở thành một cụm riêng lẻ hoặc đạt tiêu chí dừng.

Phương pháp agglomerative phổ biến hơn do tính đơn giản trong triển khai và được sử dụng rộng rãi trong thực tế.

Các phương pháp liên kết: single, complete, average, Ward.

Trong phân cụm phân cấp, cách tính khoảng cách giữa các cụm (linkage) là yếu tố quan trọng quyết định cấu trúc của các cụm. Dưới đây là bốn phương pháp liên kết chính:

* + **Single Linkage (Liên kết đơn):**
    - Khoảng cách giữa hai cụm được định nghĩa là khoảng cách nhỏ nhất giữa bất kỳ cặp điểm nào thuộc hai cụm đó.
    - Công thức: d(, **) =  d(x,y)
    - Ưu điểm: Tốt cho việc phát hiện các cụm có hình dạng không đều (như chuỗi).
    - Nhược điểm: Dễ bị nhạy cảm với nhiễu (noise) và tạo hiệu ứng "chuỗi" (chaining).
  + **Complete Linkage (Liên kết đầy đủ):**
    - Khoảng cách giữa hai cụm là khoảng cách lớn nhất giữa bất kỳ cặp điểm nào thuộc hai cụm đó.
    - Công thức: d(, **) =  d(x,y)
    - Ưu điểm: Tạo ra các cụm nhỏ gọn, đồng nhất hơn.
    - Nhược điểm: Có thể bỏ qua các cụm lớn hoặc không đều.
  + **Average Linkage (Liên kết trung bình):**
    - Khoảng cách giữa hai cụm là trung bình của tất cả các khoảng cách giữa các cặp điểm thuộc hai cụm.
    - Công thức: d(C\_1, C\_2) = 1/(|C\_1 ||C\_2 |) ∑\_(x∊C1 ) ∑\_(y∊C2) d(x,y)
    - Ưu điểm: Cân bằng giữa single và complete linkage, ít nhạy cảm với nhiễu hơn single linkage.
    - Nhược điểm: Tốn kém về tính toán với tập dữ liệu lớn.
  + **Ward’s Method (Phương pháp Ward):**
    - Dựa trên việc giảm thiểu tổng bình phương sai số (variance) trong cụm khi ghép hai cụm lại với nhau.
    - *Công thức: Tối thiểu hóa ∆ =  – ( + )*, trong đó SSE là tổng bình phương khoảng cách từ các điểm đến tâm cụm.
    - Ưu điểm: Tạo ra các cụm cân bằng và nhỏ gọn, thường cho kết quả tốt trong thực tế.
    - Nhược điểm: Yêu cầu tính toán phức tạp hơn và giả định dữ liệu có phân phối gần giống hình cầu.

Xác định số cụm dựa trên dendrogram. Dendrogram là một biểu đồ cây thể hiện quá trình ghép hoặc chia cụm trong phân cụm phân cấp. Nó giúp xác định số lượng cụm tối ưu bằng cách:

* + **Trục ngang:** Đại diện cho các đối tượng hoặc cụm được ghép lại với nhau.
  + **Trục dọc:** Đại diện cho khoảng cách (hoặc độ đo tương tự) giữa các cụm khi chúng được ghép.

**Cách sử dụng dendrogram để chọn số cụm:**

1. Quan sát các nhánh của dendrogram: Mỗi lần ghép cụm tạo ra một nhánh mới.
2. Tìm "khoảng cách nhảy lớn" (largest gap): Cắt dendrogram tại mức độ cao (height) mà khoảng cách giữa hai lần ghép liên tiếp là lớn nhất. Điều này cho thấy sự phân tách tự nhiên giữa các cụm.
3. Quyết định số cụm: Số lượng nhánh bị cắt tại mức đó chính là số cụm.

Ví dụ: Nếu bạn cắt dendrogram ở một mức mà có 3 nhánh, thì số cụm là 3. Phương pháp này mang tính trực quan và phụ thuộc vào mục tiêu phân tích của người dùng.

# 6. Xây dựng mô hình và so sánh kết quả

## 6.1. Triển khai từng mô hình phân cụm trên bộ dữ liệu

Toàn bộ bài tập được nhóm thực hiện viết với mã Python và các thư viện hỗ trợ. Đối với các thuật toán phân cụm, nhóm sử dụng sklearn và scipy, để trực quan hóa dữ liệu cũng như kết quả phân cụm, nhóm sử dụng các thư viện như matplotlib, seaborn hay plotpy. Cụ thể, nhóm sử dụng sklearn để chạy thuật toán DBSCAN, K-Nearest Neighbors giúp chọn tham số epsilon và minpts cho DBSCAN, thư viện scipy để chạy thuật toán Hierachical Clustering.

Dưới đây là cách cài đặt các mô hình và các tham số ban đầu cho từng thuật toán:

* K-Nearest Neighbors:
  + Tham số k để xác định số cụm = 7
  + Sử dụng Elbow để tìm điểm thay đổi đột ngột để xác định tham số epsilon cho DBSCAN
* DBSCAN:
  + **minPts:** Là số lượng tối thiểu các điểm dữ liệu cần có trong vùng lân cận của một điểm để được coi là **Core Point**.
  + **Epsilon:** Là bán kính của vùng lân cận xung quanh một điểm trung tâm.
  + Cách chọn minpts và epsilon:
    - **minPts:** Nếu D là số lượng chiều hoặc số lượng đặc trưng của tập dữ liệu, thì **minPts** >= D + 1. Thông thường, với các tập dữ liệu nhỏ hoặc có nhiều nhiễu, **minPts** >= 2 \* D.
    - **Epsilon:** Giá trị này thường được xác định bằng cách sử dụng đồ thị **k-distance graph** được tạo ra từ mô hình **KNN**. Điểm mà đồ thị thay đổi đột ngột chính là giá trị được chọn.
* Hierachical Clustering:
  + Tham số chính: n\_clusters, linkage, metric
  + Thước đo khoảng cách giữa các điểm ('euclidean', 'manhattan', v.v.).
  + Phương pháp liên kết các cụm ('single', 'complete', 'average', 'ward').
  + Quá trình tạo cụm của Hierarchical Clustering này được thể hiện bằng **dendrogram**. Từ biểu đồ này, chọn ra số lượng cụm bằng cách xác định số lần hợp nhất ở giai đoạn áp chót. Độ dài của các đường thẳng đứng càng lớn thể hiện khoảng cách giữa các cụm càng xa.
  + Để chọn số lượng cụm, có thể đặt một ngưỡng giá trị và đếm số lượng đường thẳng đứng nằm phía trên ngưỡng này. Số đường thẳng này chính là số lượng cụm. Ngoài ra, các phương pháp như **Silhouette Score Method** và **Elbow Method** cũng có thể được sử dụng.

Nhóm tiến hành chạy 2 mô hình với 2 loại dữ liệu:

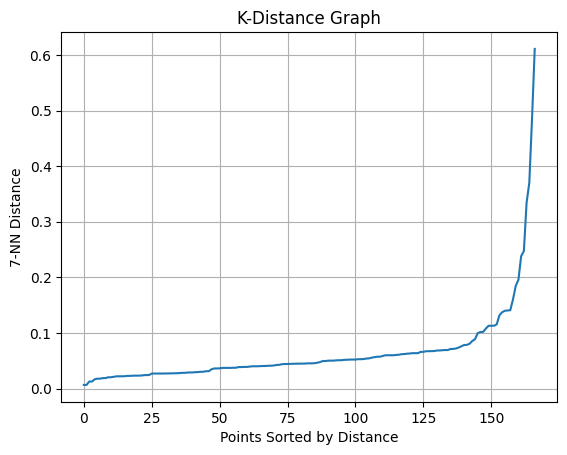
* Dữ liệu 3 nhóm sau tiền xử lý dữ liệu gồm: **Health - Trade - Finance gọi là dữ liệu kết hợp**
* Dữ liệu sử dụng PCA để giảm chiều

## 6.2. Đánh giá chất lượng mô hình

### 6.2.1. DBSCAN Clustering

**a. DBSCAN Clustering với dữ liệu kết hợp**

Sử dụng KNN [4] chạy với **dữ liệu kết hợp** để chọn tham số cho DBSCAN với n\_neighbord=7, thu được đồ thị **k-distance graph** sau:



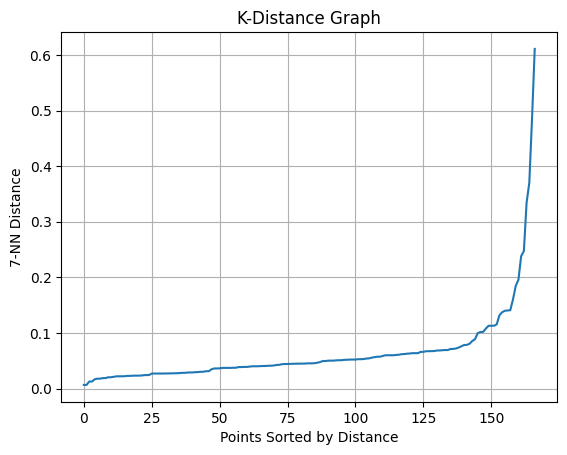
Hình 20: Biểu đồ k-distance graph DBSCAN với dữ liệu kết hợp

Từ biểu đồ kết quả thử nghiệm chạy KNN, có thể chọn **eps** = 0.08 và chọn **minPts** = 8 >= 2 \* 3 số features (Sau khi áp dụng PCA cho dữ liệu xuống 3 chiều).

Giá trị gán cho **n\_neighbors** = **minPts** - 1 = 7.

**b. DBSCAN Clustering với dữ liệu PCA**

Sử dụng KNN chạy với dữ liệu đã giảm chiều dữ liệu để chọn tham số cho DBSCAN với n\_neighbord=7, thu được đồ thị **k-distance graph** sau:



Hình 21: Biểu đồ k-distance graph DBSCAN với dữ liệu PCA

Từ biểu đồ kết quả thử nghiệm chạy KNN, có thể chọn **eps** = 0.08 và chọn

**minPts** = 8 >= 2 \* 3 số features (Sau khi áp dụng PCA cho dữ liệu xuống 3 chiều).

Giá trị gán cho **n\_neighbors** = **minPts** - 1 = 7.

### 6.2.2. Hierarchical Clustering

Với thuật toán Hierarchical Clustering, nhóm tiến hành chạy trước với linkage để thu được dendrogram. Sau đó tiến hành chọn cụm theo phân cấp của cây.

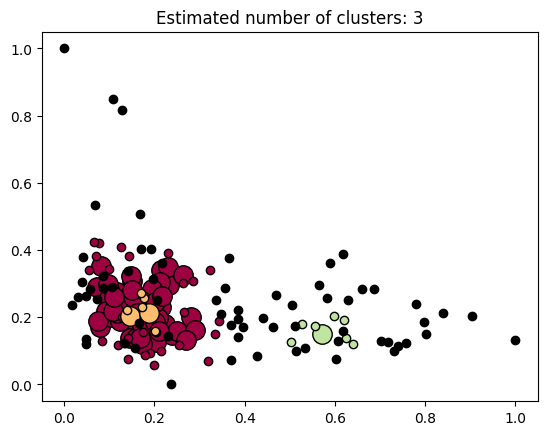
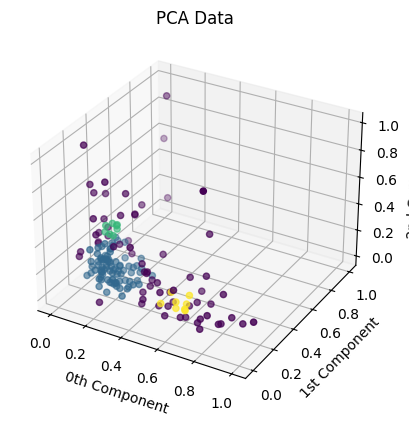
## 6.3. So sánh kết quả trực quan

### 6.3.1 DBSCAN Clustering

**a. DBSCAN với dữ liệu kết hợp**

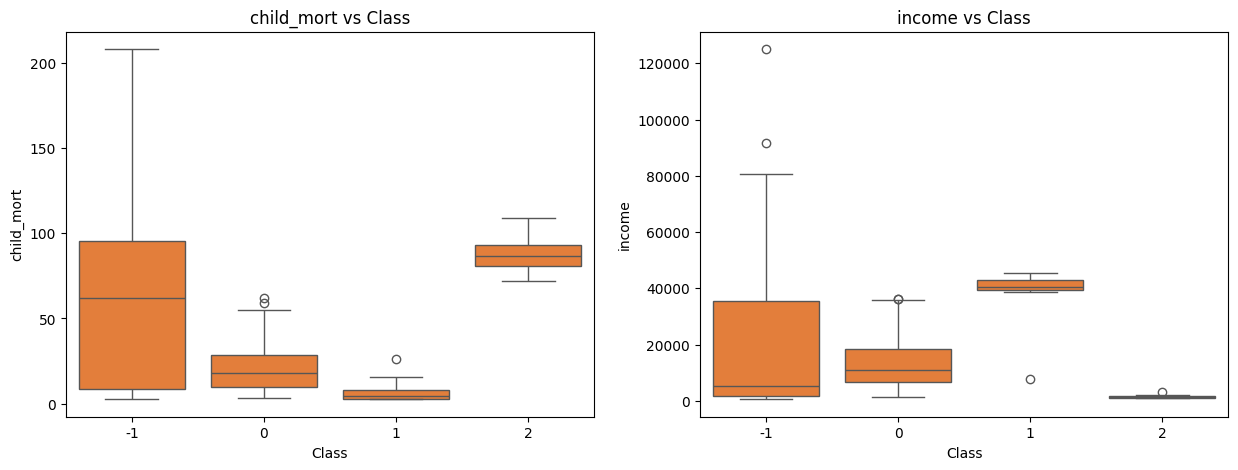
Kết quả chạy DBSCAN với **dữ liệu kết hợp** và các tham số tham số eps=0.08, minPts=8 thu được 3 cụm và 94 outliers.

Dưới đây là biểu đồ trực quan hóa kết quả phân cụm bằng DBSCAN với dữ liệu đã qua PCA:



Hình 22: Biểu đồ trực quan hóa kết quả phân cụm với DBSCAN với dữ liệu kết hợp

Nhìn lại biểu đồ giữa income, child\_mort với các cụm để nhận xét:

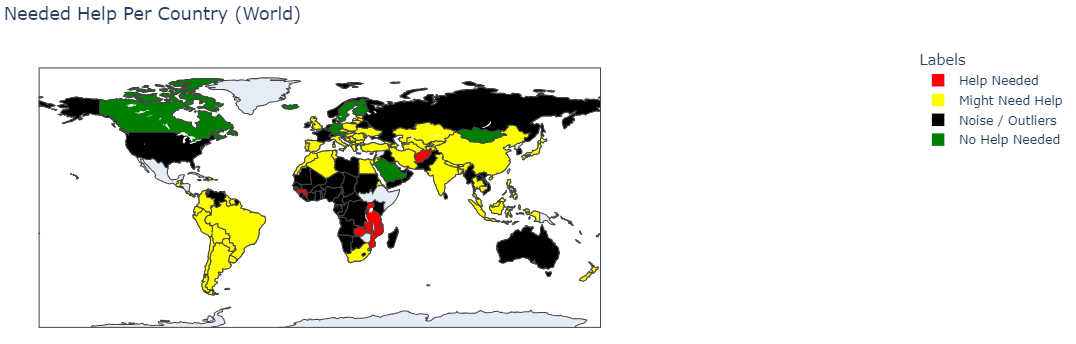


Hình 23: Biểu đồ Boxplot trực quan hóa kết quả phân cụm với DBSCAN với dữ liệu kết hợp

Từ biểu đồ trên và dựa trên **thu nhập thấp và tỷ lệ tử vong trẻ em cao** là dấu hiệu của một **quốc gia kém phát triển về kinh tế** có thể kết luận:

* + - **-1: Nhiễu / Điểm ngoại lai**
    - **2: Cần hỗ trợ**
    - **0: Không cần hỗ trợ**
    - **1: Có thể cần hỗ trợ**

Dưới đây là biểu đồ phân bổ nguồn hỗ trợ dựa theo tên quốc gia:

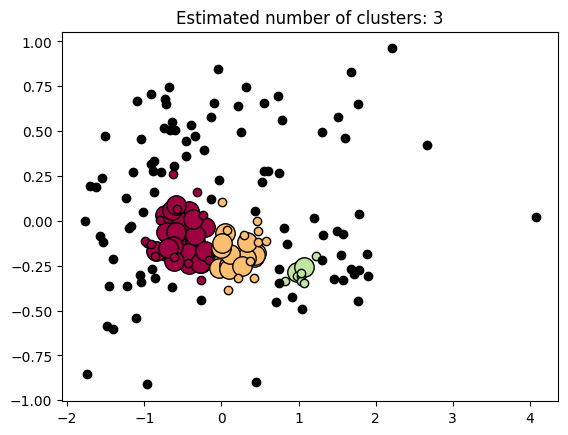
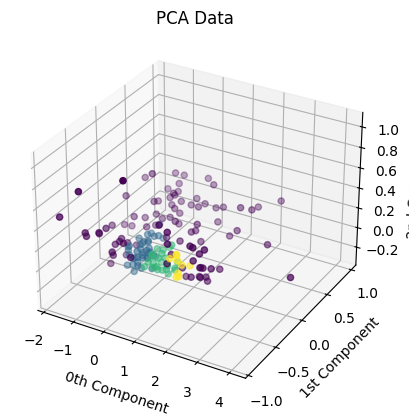


Hình 24: Biểu đồ trực phân bổ nguồn vốn theo quốc gia với DBSCAN với dữ liệu kết hợp

**b. DBSCAN với dữ liệu PCA**

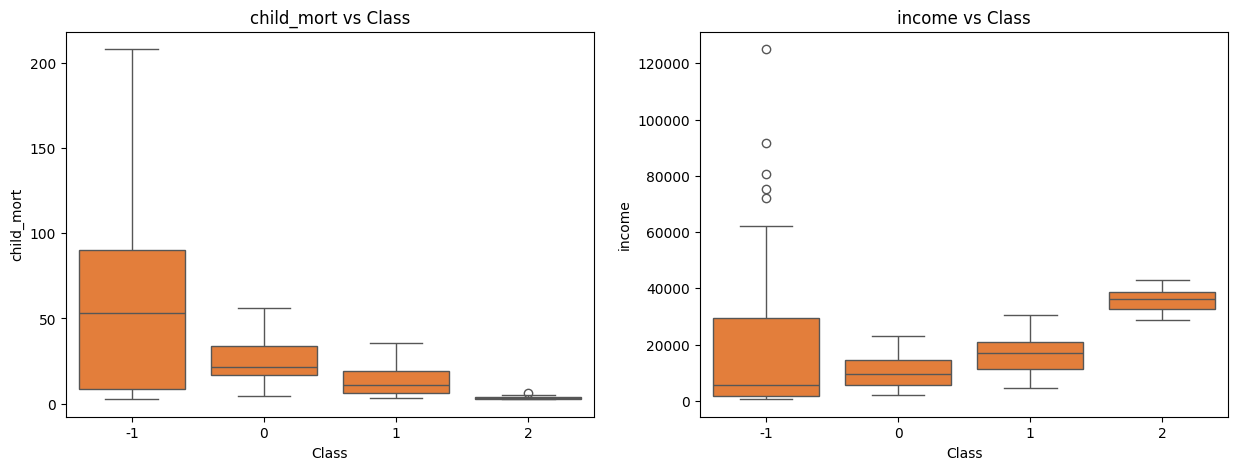
Kết quả chạy DBSCAN với dữ liệu PCA và các tham số tham số eps=0.08, minPts=8 thu được 3 cụm và 94 outliers.

Dưới đây là biểu đồ trực quan hóa kết quả phân cụm bằng DBSCAN với dữ liệu đã qua PCA:



Hình 25: Biểu đồ trực quan hóa kết quả phân cụm với DBSCAN với dữ liệu PCA

Nhìn lại biểu đồ giữa income, child\_mort với các cụm để nhận xét:

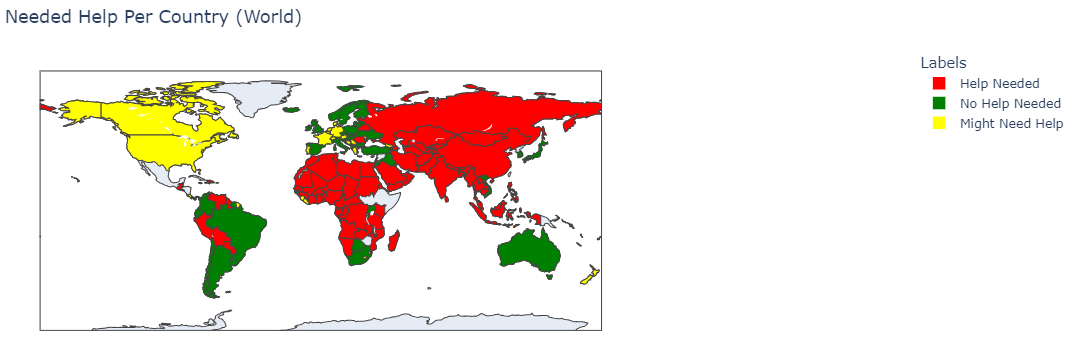


Hình 26: Biểu đồ boxplot trực quan hóa kết quả phân cụm với DBSCAN với dữ liệu PCA

Từ biểu đồ trên và dựa trên **thu nhập thấp và tỷ lệ tử vong trẻ em cao** là dấu hiệu của một **quốc gia kém phát triển về kinh tế** có thể kết luận:

* + **-1: Nhiễu / Điểm ngoại lai**
  + **0: Cần hỗ trợ**
  + **1: Có thể cần hỗ trợ**
  + **2: Không cần hỗ trợ**

Dưới đây là biểu đồ phân bổ nguồn hỗ trợ dựa theo tên quốc gia:

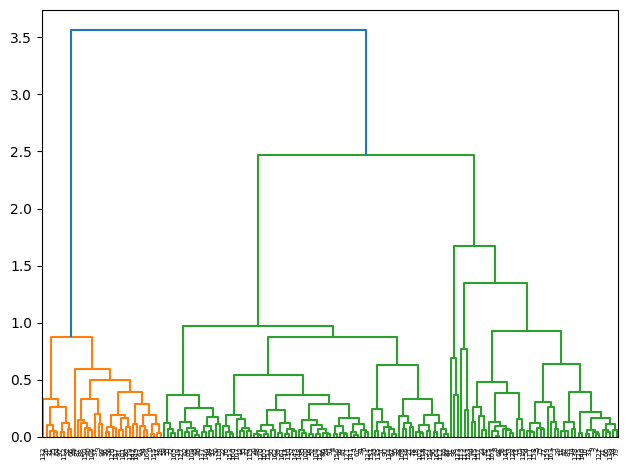


Hình 27: Biểu đồ trực phân bổ nguồn vốn theo quốc gia với DBSCAN với dữ liệu PCA

### 6.3.2. Hierarchical Clustering

**a. Hierarchical Clustering với dữ liệu hợp nhất**

Nhóm tiến hành chạy thuật toán với linkage là ward và metric là euclidean để chọn số cụm. Dưới đây là kết quả sau khi chạy thu được dendrogram như sau:

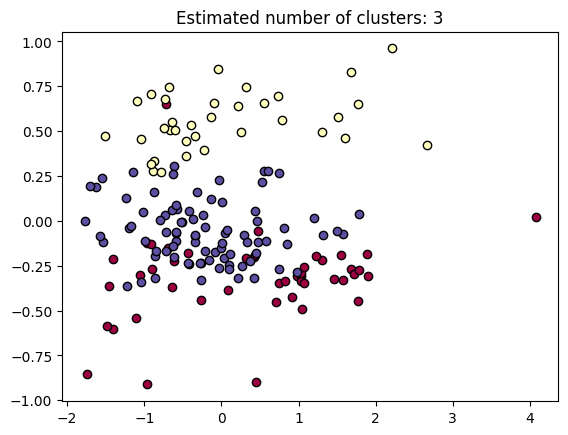
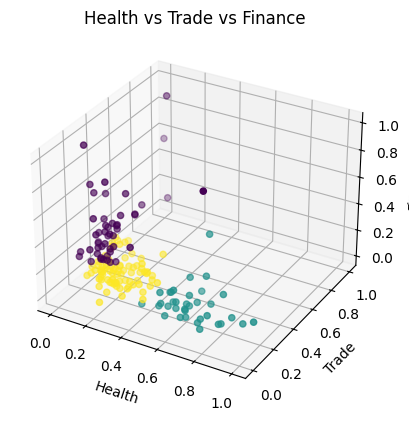


Hình 28: Biểu đồ dendrogram kết quả phân cụm với Hierachical Clustering với dữ liệu kết hợp

Nhận xét biểu đồ dendrogram:

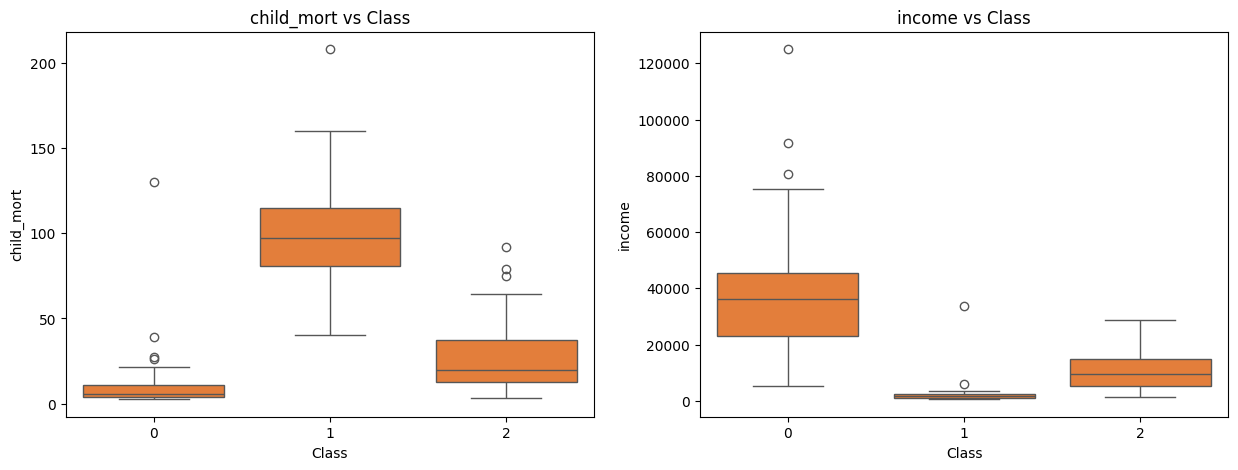
* Từ biểu đồ có thể thấy, cần chia các quốc gia thành 3 nhóm.
* Phân tích dendrogram cho tập dữ liệu này khá dư thừa vì từ biểu đồ thấy 1 **đường màu cam** cùng với 2 **đường màu xanh** là các cụm cuối cùng được tạo ra trước khi kết nối với nhau.
* Dendrogram có 3 nhánh, cho thấy **3 cụm** được tạo ra trước khi hợp nhất thành 1

Kết quả phân cụm phân cấp của Hierachical Clustering với n\_clusters=3 và dữ liệu đã chạy PCA để giảm chiều như sau:



Hình 29: Biểu đồ trực quan hóa kết quả phân cụm với Hierachical Clustering với dữ liệu kết hợp

Nhìn lại biểu đồ giữa income, child\_mort với các cụm để nhận xét:

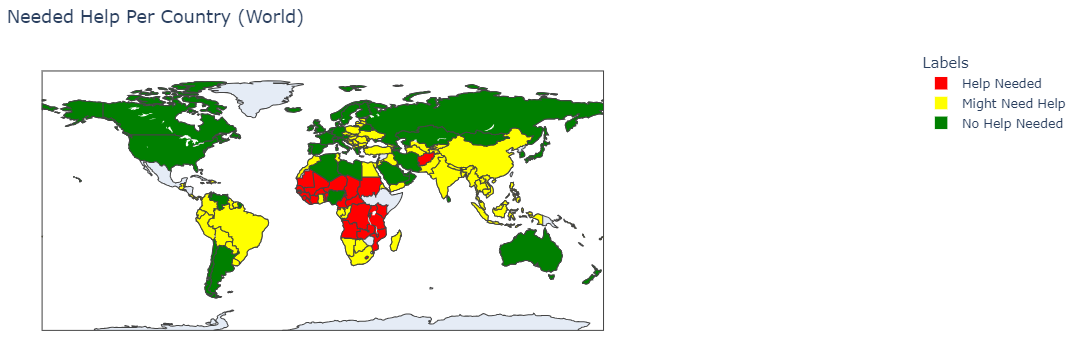


Hình 30: Biểu đồ Boxplot kết quả phân cụm với Hierachical Clustering với dữ liệu kết hợp

Từ biểu đồ trên và dựa trên **thu nhập thấp và tỷ lệ tử vong trẻ em cao** là dấu hiệu của một **quốc gia kém phát triển về kinh tế** có thể kết luận:

* + **0: Không cần hỗ trợ**
  + **1: Cần hỗ trợ**
  + **2: Có thể cần hỗ trợ**

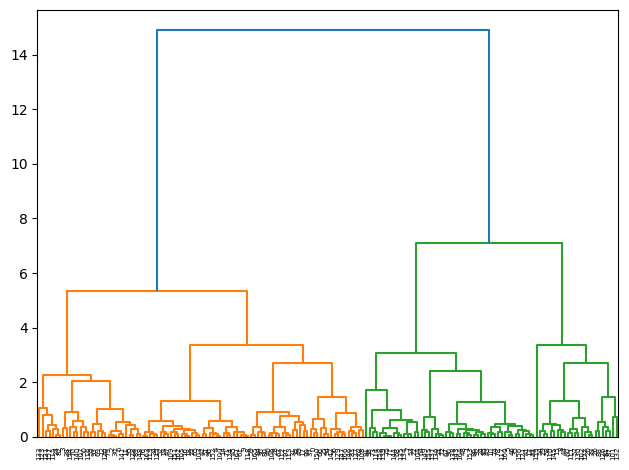
Dưới đây là biểu đồ phân bổ nguồn hỗ trợ dựa theo tên quốc gia:

****

Hình 31: Biểu đồ trực phân bổ nguồn vốn theo quốc gia với Hierarchical Clustering với dữ liệu kết hợp

**b.** **Hierarchical Clustering với dữ liệu PCA**

Nhóm tiến hành chạy thuật toán với linkage là ward và metric là euclidean để chọn số cụm. Dưới đây là kết quả sau khi chạy thu được dendrogram như sau:

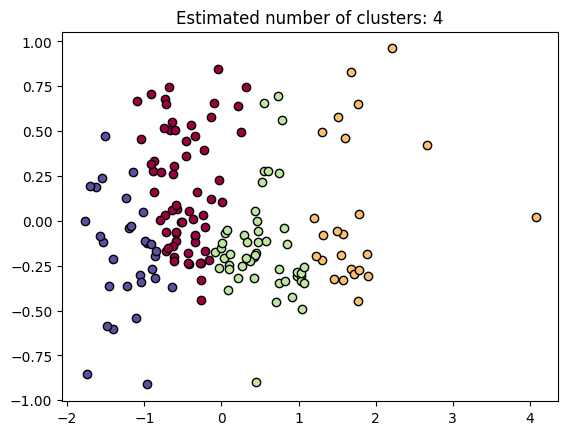
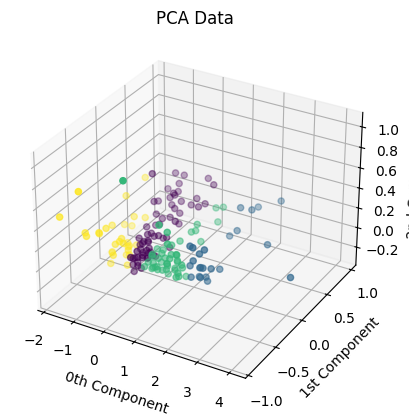


Hình 32: Biểu đồ dendrogram kết quả phân cụm với Hierachical Clustering với dữ liệu PCA

Nhận xét biểu đồ dendrogram:

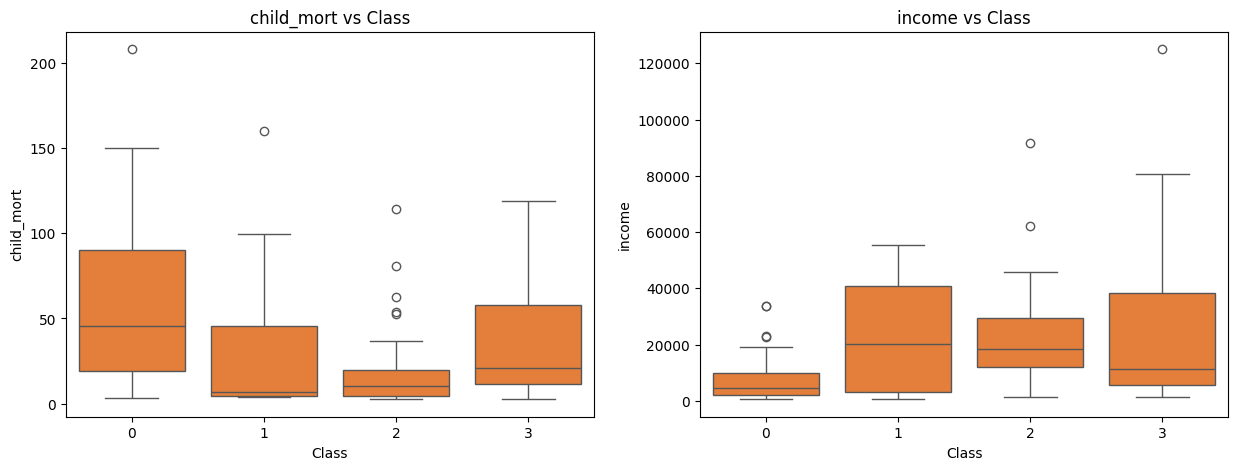
* Từ biểu đồ có thể thấy, cần chia các quốc gia thành 3 nhóm.
* Phân tích dendrogram cho tập dữ liệu này khá dư thừa vì từ biểu đồ thấy 2 **đường màu xanh** cùng với 2 **đường màu cam** là các cụm cuối cùng được tạo ra trước khi kết nối với nhau.
* Dendrogram có 4 nhánh, cho thấy **4 cụm** được tạo ra trước khi hợp nhất thành 1

Kết quả phân cụm phân cấp của Hierachical Clustering với n\_clusters=3 và dữ liệu đã chạy PCA để giảm chiều như sau:



Hình 33: Biểu đồ trực quan hóa kết quả phân cụm với Hierachical Clustering với dữ liệu PCA

Nhìn lại biểu đồ giữa income, child\_mort với các cụm để nhận xét:

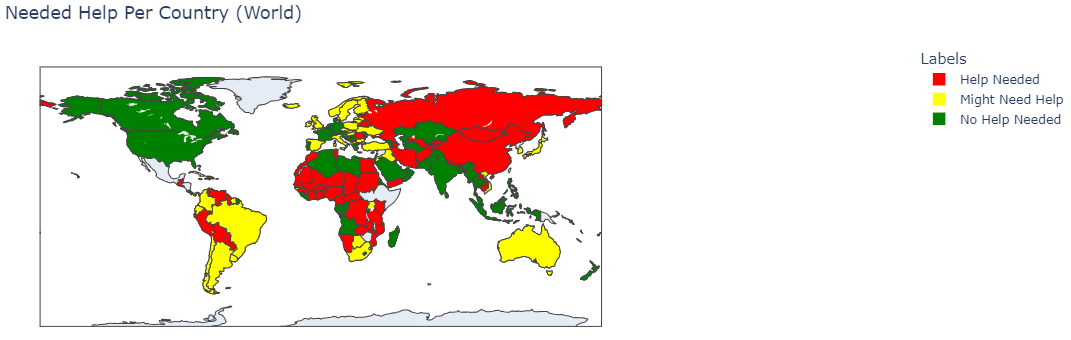


Hình 34: Biểu đồ Boxplot kết quả phân cụm với Hierachical Clustering với dữ liệu PCA

Từ biểu đồ trên và dựa trên **thu nhập thấp và tỷ lệ tử vong trẻ em cao** là dấu hiệu của một **quốc gia kém phát triển về kinh tế** có thể kết luận:

* + **0: Cần hỗ trợ**
  + **2: Có thể cần hỗ trợ**
  + **1, 3: Không cần hỗ trợ**

Dưới đây là biểu đồ phân bổ nguồn hỗ trợ dựa theo tên quốc gia:

****

Hình 35: Biểu đồ trực phân bổ nguồn vốn theo quốc gia với Hierarchical Clustering với dữ liệu PCA

### 6.3.3. So sánh cụm của DBSCAN và Hierarchical về mặt phân bố.

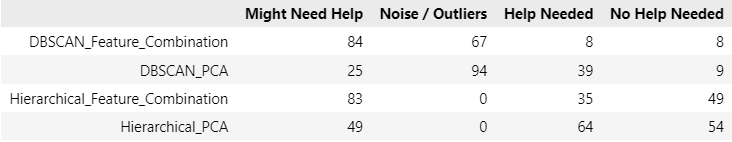
Để so sánh cụm của DBSCAN và Hierarchical về mặt phân bố, nhóm thực hiện so cánh theo các tiêu chí sau:

* **So sánh số lượng cụm:** Đếm số lượng cụm mà mỗi thuật toán tạo ra, bao gồm cả số lượng điểm nhiễu (outliers) của DBSCAN.

|  |  |
| --- | --- |
| Trường hợp thử nghiệm | Số cụm dự đoán |
| DBSCAN\_Feature\_Combination | 3 cụm và 1 cụm nhiễu |
| DBSCAN\_PCA | 3 cụm và 1 cụm nhiễu |
| Hierarchical\_Feature\_Combination | 3 cụm |
| Hierarchical\_PCA | 4 cụm |

Bảng 1: Bảng kết quả số lượng cụm phân loại của các thuật toán

* **Phân tích kích thước cụm:** So sánh số lượng điểm trong từng cụm để xem thuật toán nào tạo ra cụm đồng đều hay chênh lệch.



Bảng 2: Bảng kết quả số lượng của từng cụm phân loại của các thuật toán

Dựa vào bảng so sánh kết quả cụm của DBSCAN và Hierarchical với hai phương pháp xử lý đặc trưng (Feature Combination và PCA), Nhóm có một số nhận xét như sau:

* **DBSCAN\_Feature\_Combination:** DBSCAN với Feature Combination tạo ra rất nhiều điểm nhiễu (67), cho thấy thuật toán này khó tìm ra các cụm rõ ràng với bộ đặc trưng này. Cụm “Help Needed” và “No Help Needed” đều rất nhỏ, có thể cụm không cân bằng.
* **DBSCAN\_PCA:** Sau khi giảm chiều bằng PCA, DBSCAN tạo ra nhiều điểm nhiễu hơn (94). Số lượng cụm "Help Needed" tăng, nhưng các cụm vẫn chưa cân bằng và tỷ lệ noise vẫn quá lớn, cho thấy DBSCAN không hoạt động tốt với PCA trong trường hợp này.
* **Hierarchical\_Feature\_Combination:** Hierarchical với Feature Combination phân cụm tốt hơn, không có điểm nhiễu, và số lượng cụm phân bổ đều hơn. Đây là dấu hiệu của một mô hình ổn định.
* **Hierarchical\_PCA:** Hierarchical với PCA thể hiện rất tốt: không có điểm nhiễu, các cụm được phân bố đều hơn, đặc biệt cụm "Help Needed" và "No Help Needed" chiếm tỷ lệ cao. Đây là mô hình có kết quả phân cụm cân bằng và chính xác nhất.

Xếp hạng các mô hình**:**

* + - **Hierarchical\_PCA** là mô hình tốt nhất: không có noise, cụm phân bố đều.
    - **DBSCAN** tạo ra quá nhiều điểm nhiễu, đặc biệt khi áp dụng PCA.
    - **Hierarchical\_Feature\_Combination** cũng tốt, nhưng chưa cân bằng cụm bằng Hierarchical\_PCA.
    - **DBSCAN\_Feature\_Combination** hoạt động kém nhất với nhiều điểm nhiễu và cụm không cân bằng.
* **Đánh giá chất lượng cụm:** Tính chỉ số như Silhouette Score.

|  |  |
| --- | --- |
| Trường hợp thử nghiệm | Điểm |
| DBSCAN\_Feature\_Combination | 0.0459 |
| DBSCAN\_PCA | -0.0464 |
| Hierarchical\_Feature\_Combination | 0.0338 |
| Hierarchical\_PCA | 0.3401 |

Bảng 3: Bảng kết quả Điểm Silhouse của từng cụm phân loại của các thuật toán

Dựa vào bảng Silhouette Score, Nhóm có một số nhận xét như sau:

* **DBSCAN\_Feature\_Combination:** Silhouette Score: **0.0459.** Điểm rất thấp, cho thấy các cụm không tách biệt rõ ràng, và có sự chồng chéo lớn giữa các cụm. DBSCAN với bộ đặc trưng gốc hoạt động không hiệu quả.
* **DBSCAN\_PCA:** Silhouette Score: **-0.0464** (Âm). Điểm âm là dấu hiệu xấu, nghĩa là các điểm thường gần cụm khác hơn cụm của chính nó. DBSCAN sau PCA thậm chí còn hoạt động tệ hơn, tạo ra các cụm không hợp lý.
* **Hierarchical\_Feature\_Combination:** Silhouette Score: **0.0338.** Điểm thấp, nhưng vẫn dương. Điều này cho thấy Hierarchical với đặc trưng gốc vẫn có thể tìm ra cụm, nhưng chất lượng cụm không cao và các cụm khá chồng lấn.
* **Hierarchical\_PCA:** Silhouette Score: **0.3401.** Điểm cao nhất trong tất cả các mô hình, thể hiện khả năng phân cụm tốt, các cụm tách biệt rõ ràng. Hierarchical kết hợp PCA đã cải thiện đáng kể hiệu quả phân cụm.

Xếp hạng các mô hình**:**

* **Hierarchical\_PCA** vượt trội nhất với Silhouette Score cao nhất (**0.3401**), cho thấy khả năng tạo cụm chặt chẽ và tách biệt tốt.
* **DBSCAN\_PCA** hoạt động kém nhất với điểm âm, chứng tỏ các cụm tạo ra rất kém chất lượng.
* PCA cải thiện đáng kể kết quả cho Hierarchical nhưng làm giảm chất lượng với DBSCAN.
* Hierarchical ổn định hơn DBSCAN với cả hai bộ đặc trưng.

## 6.4. Nhận xét, phân tích lý do mô hình cho kết quả tốt/không tốt

Tổng kết lại, nhóm có nhận xét như sau:

* **Hierarchical\_PCA** – Tốt nhất: Silhouette Score cao nhất (0.3401), cụm chặt chẽ, phân bố đều, không có noise.
* **Hierarchical\_Feature\_Combination** – Tốt nhưng chưa cân bằng cụm bằng Hierarchical\_PCA.
* **DBSCAN\_Feature\_Combination** – Nhiều điểm nhiễu, cụm không cân bằng.
* **DBSCAN\_PCA** – Kém nhất: Silhouette Score âm, cụm kém chất lượng.

# 7. Kết luận

## 7.1. Tổng kết những điểm chính đạt được

Qua quá trình phân tích và áp dụng các phương pháp khoa học dữ liệu lên bộ dữ liệu về tình hình phát triển kinh tế – xã hội – y tế của các quốc gia, có thể khẳng định rằng chất lượng dữ liệu đầu vào là yếu tố tiên quyết ảnh hưởng đến hiệu suất của mô hình. Việc thực hiện EDA đã cung cấp nhiều góc nhìn quan trọng, làm sáng tỏ đặc điểm của từng chỉ số và mối liên hệ giữa chúng. Bên cạnh đó, thí nghiệm cho thấy quá trình chuẩn hóa, tiêu chuẩn hóa (normalization và standardization) có tác động đáng kể đến kết quả phân cụm. Tương tự, bước feature engineering, bao gồm các kỹ thuật kết hợp biến và PCA (Phân tích thành phần chính), đòi hỏi sự cân nhắc thận trọng do ảnh hưởng đến độ phức tạp của mô hình và khả năng giải thích kết quả.

Trong phạm vi đề tài, hai thuật toán được triển khai – Hierarchical Clustering và DBSCAN – đều thể hiện những khác biệt rõ rệt về nguyên lý hoạt động cũng như hiệu suất thực tế. Dù còn những hạn chế về mặt đánh giá và tính đồng nhất, việc so sánh kết quả giữa các thuật toán đã giúp nhóm nghiên cứu có cái nhìn tổng quan về cách tiếp cận học máy không giám sát, nhận biết ưu – nhược điểm của từng phương pháp.

## 7.2. Hạn chế của nghiên cứu và hướng phát triển tương lai

Về hạn chế, do dữ liệu được tổng hợp từ nhiều nguồn với đặc thù thống kê và thời điểm thu thập khác nhau, chất lượng dữ liệu có thể chưa tối ưu và chưa thật sự đồng nhất. Ngoài ra, số lượng features vẫn còn giới hạn, chưa bao quát hết được những khía cạnh quan trọng khác như giáo dục, môi trường, hay mức độ bất bình đẳng xã hội. Thêm vào đó, kết quả đánh giá mô hình cho thấy hiệu quả phân cụm chưa thực sự tốt như mong đợi, một phần đến từ tính chất phân tán của dữ liệu và các biến liên quan.

Trong tương lai, việc mở rộng bộ dữ liệu (bổ sung thêm các chỉ số định tính và định lượng), kết hợp với các phương pháp tiền xử lý nâng cao (chẳng hạn xử lý dữ liệu bị thiếu, dữ liệu lệch, ngoại lai) sẽ giúp cải thiện hiệu suất của mô hình. Đồng thời, việc thử nghiệm các thuật toán phân cụm khác như Spectral Clustering, Gaussian Mixture Model hoặc thậm chí phân cụm lai (hybrid methods) cũng là hướng tiếp cận đáng quan tâm để tìm được mô hình phù hợp nhất với cấu trúc dữ liệu thực tế.

## 7.3. Bài học kinh nghiệm và đề xuất ứng dụng thực tiễn

Từ những kết quả thu được, một bài học quan trọng là quá trình chọn và chuẩn bị dữ liệu không chỉ dừng lại ở việc dọn dẹp, mà còn bao gồm cả sự đánh giá nghiêm túc về các biến đặc trưng. Những kỹ thuật như feature engineering và PCA cần được thực hiện có chủ đích và hiểu rõ bối cảnh, thay vì áp dụng một cách máy móc.

Về ứng dụng thực tiễn, phân cụm các quốc gia dựa trên các yếu tố kinh tế – xã hội – y tế là bước khởi đầu cho nhiều giải pháp nhân đạo và phát triển bền vững. Dữ liệu đầu ra có thể hỗ trợ các tổ chức quốc tế như HELP International trong việc xác định ưu tiên phân bổ nguồn lực, đẩy nhanh tốc độ tiếp cận những khu vực dễ tổn thương nhất. Thêm vào đó, mô hình phân cụm có thể được tích hợp vào các hệ thống dự báo, giám sát và cảnh báo sớm về nguy cơ dịch bệnh, đói nghèo hay thiên tai, giúp nâng cao hiệu quả và tính chủ động trong công tác hỗ trợ.

TÀI LIỆU THAM KHẢO

[1] “15. DBSCAN — Deep AI KhanhBlog.” Accessed: Mar. 08, 2025. [Online]. Available: https://phamdinhkhanh.github.io/deepai-book/ch\_ml/index\_DBSCAN.html

[2] “14. Hierarchical Clustering (phân cụm phân cấp) — Deep AI KhanhBlog.” Accessed: Mar. 08, 2025. [Online]. Available: https://phamdinhkhanh.github.io/deepai-book/ch\_ml/index\_HierarchicalClustering.html

[3] “Mục đích của EDA — Machine Learning cho dữ liệu dạng bảng.” Accessed: Mar. 08, 2025. [Online]. Available: https://machinelearningcoban.com/tabml\_book/ch\_data\_processing/eda\_purpose.html

[4] “Giải thuật k hàng xóm gần nhất,” *Wikipedia tiếng Việt*. Aug. 16, 2021. Accessed: Mar. 08, 2025. [Online]. Available: https://vi.wikipedia.org/w/index.php?title=Gi%E1%BA%A3i\_thu%E1%BA%ADt\_k\_h%C3%A0ng\_x%C3%B3m\_g%E1%BA%A7n\_nh%E1%BA%A5t&oldid=65495740

PHÂN CÔNG CÔNG VIỆC

Dũng – Xây dựg mô hình và so sánh kết quả