



**Đề tài**

# **PHÂN CỤM QUỐC GIA THEO TÌNH TRẠNG KINH TẾ - XÃ HỘI BẰNG THUẬT TOÁN OPTICS**

**Môn học:** Máy học

**Giảng viên hướng dẫn:** Nguyễn An Tế

**Nhóm sinh viên thực hiện:** Nhóm 1

- Đỗ Ngọc Phương Anh – 31221020325
- Nguyễn Ngọc Thúy Anh – 31221020005
- Nguyễn Trần Thế Anh – 31221026655
- Phan Huỳnh Ngọc Anh – 33221025199
- Phạm Bằng – 31221024364

## DANH MỤC HÌNH ẢNH

Hình 1. Phân cụm dữ liệu	3
Hình 2. Điểm lõi (Core points) và điểm biên (Border points)	6
Hình 3. Density-reachable và Density-connected	7
Hình 4. Reachability Plot	9
Hình 5. Minh họa sự khác biệt giữa OPTICS và DBSCAN	10
Hình 6. So sánh khoảng cách giữa các điểm	11
Hình 7. Mô tả cách thuật toán hoạt động	13
Hình 8. Sự khác biệt giữa các thuật toán phân cụm	16
Hình 9. Một số dòng dữ liệu trong bộ dữ liệu	20
Hình 10. Thống kê mô tả dữ liệu	21
Hình 11. Kết quả kiểm tra giá trị null	21
Hình 12. Trực quan hóa các giá trị ngoại lai	22
Hình 13. Các bước thực hiện PCA	24
Hình 14. Kết quả xác định thành phần chính thứ k	25
Hình 15. Biểu đồ Reachability	30
Hình 16. Kết quả phân cụm OPTICS trong không gian PCA	30
Hình 17. Kết quả đếm số lượng đối tượng từng nhãn	31
Hình 18. Bảng dữ liệu thêm cột gán nhãn 'Cluster'	31
Hình 19. Dữ liệu của các quốc gia thuộc nhãn -1	32
Hình 20. Mô tả tuyến tính dữ liệu của các quốc gia thuộc nhãn -1	32
Hình 21. Dữ liệu của các quốc gia thuộc nhãn 0	33
Hình 22. Mô tả tuyến tính dữ liệu của các quốc gia thuộc nhãn 0	33
Hình 23. Silhouette Score của thuật toán OPTICS	35
Hình 24. Kết quả phân cụm K-Means trong không gian 3D	36
Hình 25. Silhouette Score của thuật toán K-Means	37
Hình 26. Kết quả phân cụm của thuật toán DBSCAN trong không gian 3D	38
Hình 27. Silhouette Score của thuật toán DBSCAN	38

## DANH MỤC BẢNG BIỂU

Bảng 1. Các loại thuật toán phân cụm	5
Bảng 2. So sánh OPTICS với các thuật toán phân cụm khác	16
Bảng 3. Mô tả thuộc tính và kiểu dữ liệu trong dữ liệu	20

# MỤC LỤC

<b>DANH MỤC HÌNH ẢNH.....</b>	<b>.....</b>
<b>DANH MỤC BẢNG BIỂU.....</b>	<b>.....</b>
<b>LỜI CẢM ƠN.....</b>	<b>.....</b>
<b>LỜI MỞ ĐẦU.....</b>	<b>.....</b>
<b>CHƯƠNG 1. TỔNG QUAN ĐỀ TÀI.....</b>	<b>1</b>
1.1. Giới thiệu đề tài.....	1
1.2. Mục tiêu nghiên cứu.....	1
1.3. Phương pháp nghiên cứu.....	1
1.4. Tài nguyên sử dụng.....	2
<b>CHƯƠNG 2. CƠ SỞ LÝ THUYẾT.....</b>	<b>3</b>
2.1. Giới thiệu về phân cụm dữ liệu.....	3
2.1.1. Giới thiệu tổng quan về phân cụm dữ liệu.....	3
2.1.2. Giới thiệu tổng quan về phân cụm theo mật độ.....	5
2.2. Thuật toán OPTICS.....	7
2.2.1. Giới thiệu tổng quan về thuật toán OPTICS.....	7
2.2.2. Mô tả thuật toán.....	10
2.2.3. Thuật toán hoạt động.....	13
2.2.4. So sánh thuật toán OPTICS với một số thuật toán phân cụm khác.....	15
2.3. Đánh giá thuật toán phân cụm.....	16
2.3.1. Cơ sở lý thuyết.....	16
2.3.2. Chỉ số Silhouette.....	17
<b>CHƯƠNG 3. ÁP DỤNG THUẬT TOÁN TRÊN BỘ DỮ LIỆU.....</b>	<b>19</b>
3.1. Tổng quan bộ dữ liệu.....	19
3.1.1. Nguồn gốc bộ dữ liệu.....	19
3.1.2. Vấn đề được đặt ra trong bộ dữ liệu.....	19
3.1.3. Mô tả bộ dữ liệu.....	19
3.2. Tiền xử lý dữ liệu.....	20
3.2.1. Mô tả dữ liệu.....	20
3.2.2. Làm sạch dữ liệu.....	21
3.2.3. Giảm chiều dữ liệu bằng phương pháp PCA.....	23
3.2.4. Phương pháp ELBOW.....	24
3.3. Áp dụng thuật toán OPTICS:.....	26
3.3.1. Mô hình khởi tạo.....	26
3.3.2. Tìm kiếm các hyperparameters.....	27
3.3.3. Tạo mô hình từ các siêu tham số vừa tìm được và tiến hành phân cụm.....	29
<b>CHƯƠNG 4. ĐÁNH GIÁ THUẬT TOÁN.....</b>	<b>34</b>
4.1. Độ đo Silhouette Score.....	34

4.2. So sánh các phương pháp phân cụm khác.....	35
<b>CHƯƠNG 5. THẢO LUẬN VÀ KẾT LUẬN.....</b>	<b>39</b>
5.1. Thảo luận.....	39
5.2. Kết luận.....	39
5.3. Ứng dụng thực tế và triển vọng.....	40
<b>TÀI LIỆU THAM KHẢO.....</b>	

## LỜI CẢM ƠN

Nhóm chúng em xin chân thành gửi lời cảm ơn đến thầy Nguyễn An Tế đã tận tình giảng dạy cũng như hướng dẫn và góp ý chân thành cho chúng em trong suốt quá trình thực hiện bài tiểu luận này. Những kiến thức quý báu, những chỉ dẫn tận tình của thầy không chỉ giúp chúng em hoàn thành bài tiểu luận mà còn tạo tiền đề cho những kinh nghiệm và bài học quý giá.

Tuy nhiên, trong quá trình thực hiện đồ án môn học, nhóm nghiên cứu vẫn sẽ không thể tránh khỏi một số sai sót và hạn chế do kiến thức còn chưa đầy đủ. Tuy nhiên, chúng em tin rằng những khó khăn và thử thách này sẽ là cơ hội quý giá giúp nhóm củng cố nền tảng vững chắc và từ đó giúp nhóm hoàn thiện hơn trong những nghiên cứu và dự án tương lai.

Chúng em thực sự biết ơn sự tận tâm và nhiệt huyết của thầy trong quá trình giảng dạy, điều này là nguồn động lực lớn để chúng em nỗ lực và phát triển hơn trong quá trình học tập và nghiên cứu.

Trân trọng cảm ơn thầy!

Nhóm 1

## LỜI MỞ ĐẦU

Trong kỷ nguyên số hiện đại, sự phát triển nhanh chóng của công nghệ cũng như nhu cầu ứng dụng công nghệ trong nhiều lĩnh vực đã tạo ra một lượng dữ liệu khổng lồ và vẫn tiếp tục gia tăng mỗi ngày. Từ tài chính, giáo dục, y tế đến mọi lĩnh vực khác, dữ liệu đã trở thành “mỏ vàng” quý giá giúp thúc đẩy các giải pháp sáng tạo, đồng thời làm thay đổi cách chúng ta vận hành và ra quyết định. Tuy nhiên sự gia tăng không ngừng của dữ liệu lại đặt ra một thách thức to lớn trong việc phân tích, quản lý và khai thác để biến chúng thành những tri thức hữu ích.

Đứng ở trung tâm của làn sóng này, Machine Learning - một nhánh quan trọng trong lĩnh vực trí tuệ nhân tạo (AI), đã nổi lên như một “người dẫn đường”, giúp giải quyết thách thức to lớn mà chúng ta đang phải đối mặt. ML tập trung vào việc áp dụng thuật toán vào dữ liệu và liên tục “học” để từ đó cải thiện độ chính xác và đưa ra những phân tích hữu ích giúp chúng ta hiểu rõ hơn về thế giới.

Đó là lý do nhóm nghiên cứu lựa chọn thuật toán OPTICS để áp dụng vào bài toán phân cụm các quốc gia theo các yếu tố kinh tế - xã hội, dữ liệu sử dụng được lấy từ bộ dữ liệu Country Socioeconomic Data. Mục tiêu của nhóm là xây dựng một mô hình phân cụm các quốc gia dựa trên các yếu tố quan trọng như GDP bình quân đầu người, thu nhập ròng, chi tiêu y tế, tuổi thọ trung bình và các chỉ số khác. Việc phân cụm này không chỉ là để nhận diện và gom nhóm các quốc gia có đặc điểm tương đồng, mà còn hỗ trợ trong việc đưa ra quyết định phân bổ ngân sách viện trợ như thế nào cho hiệu quả.

# CHƯƠNG 1. TỔNG QUAN ĐỀ TÀI

## 1.1. Giới thiệu đề tài

Trong bối cảnh các cuộc khủng hoảng kinh tế, xã hội và môi trường không ngừng gia tăng, việc tìm kiếm các chiến lược hiệu quả để giảm nghèo và thúc đẩy phát triển bền vững trở nên cấp thiết hơn bao giờ hết. Các tổ chức phi lợi nhuận và các chương trình viện trợ quốc tế vẫn luôn đóng vai trò quan trọng trong việc cải thiện tình trạng nghèo đói, song, một câu hỏi lớn được đặt ra là: *Làm thế nào để xác định quốc gia nào cần được ưu tiên viện trợ?*

Đề tài được nhóm nghiên cứu xây dựng với mong muốn tìm hiểu và áp dụng thuật toán OPTICS trên bộ dữ liệu *Country Socioeconomic Data*, nhằm đánh giá hiệu quả của thuật toán trong việc phân cụm và phát hiện các cấu trúc tiềm ẩn trong dữ liệu. Thông qua việc phân tích các yếu tố kinh tế - xã hội như GDP, thu nhập ròng, chi tiêu y tế, tỷ lệ tử vong và các chỉ số phát triển khác, nhóm nghiên cứu mong muốn phân nhóm các quốc gia theo mức độ phát triển tương đồng. Từ đó, đề xuất các chiến lược viện trợ hiệu quả cho các tổ chức và chương trình viện trợ quốc tế, giúp các quốc gia khó khăn nâng cao chất lượng cuộc sống và tạo cơ hội phát triển nền kinh tế bền vững trong tương lai.

## 1.2. Mục tiêu nghiên cứu

Mục tiêu của đề án là ứng dụng thuật toán OPTICS để phân cụm các quốc gia có các đặc điểm kinh tế - xã hội tương đồng nhau. Nghiên cứu sẽ không chỉ tìm hiểu cơ chế hoạt động và ưu nhược điểm của thuật toán OPTICS trong việc phân cụm và xử lý dữ liệu mà còn giúp nhận diện các nhóm quốc gia có đặc điểm phát triển tương đồng. Dựa trên kết quả phân cụm, các tổ chức và các chương trình viện trợ quốc tế có thể phân tích và sử dụng các phân tích để:

- Nắm được tình hình kinh tế - xã hội của các quốc gia.
- Phát triển các chính sách viện trợ phù hợp, phân bổ tài nguyên hiệu quả để nâng cao chất lượng cuộc sống và tạo điều kiện để phát triển kinh tế của các nhóm quốc gia khác nhau.
- Dự đoán xu hướng kinh tế - xã hội thế giới trong tương lai

## 1.3. Phương pháp nghiên cứu

Để có thể đảm bảo được cái nhìn sâu sắc và toàn diện về khả năng ứng dụng của thuật toán OPTICS vào việc phân nhóm các quốc gia dựa trên các yếu tố kinh tế - xã hội, nhóm nghiên cứu sẽ áp dụng phương pháp nghiên cứu kết hợp giữa nghiên cứu lý thuyết và thực nghiệm trên bộ dữ liệu thực tế.

- *Nghiên cứu tài liệu:* Nghiên cứu các tài liệu khoa học, bài báo, các nghiên cứu liên quan đến thuật toán OPTICS, phân cụm, các thuật toán phân cụm khác và các nội dung liên quan đến các phân tích dữ liệu kinh tế - xã hội. Mục đích là để nắm vững nền tảng, hiểu rõ ưu điểm và hạn chế của thuật toán OPTICS, cũng như có cái nhìn tổng quan về các vấn đề cần giải quyết trong nghiên cứu này.
- *Thực nghiệm trên bộ dữ liệu thực:* Nhóm sử dụng bộ dữ liệu Country Socioeconomic Data với các chỉ số kinh tế - xã hội quan trọng như GDP, tỷ lệ tử vong, chi tiêu y tế và các chỉ số khác. Thuật toán OPTICS sẽ được áp dụng để phân cụm các quốc gia có đặc điểm tương đồng về các yếu tố, nhằm nhận diện các quốc gia có mức độ phát triển tương đương nhau.
- *Phân tích, đánh giá và thảo luận:* Kết quả phân cụm sẽ được đánh giá qua chỉ số như Silhouette Score để đo lường mức độ đồng nhất giữa các cụm và sự khác biệt giữa các cụm. Đồng thời, nhóm cũng sẽ so sánh thuật toán OPTICS với các thuật toán phân cụm khác như K-Means và DBSCAN để đánh giá hiệu quả của thuật toán. Các kết quả thu được sẽ được nhóm thảo luận chi tiết để làm rõ các ưu nhược điểm của thuật toán trong việc phân tích dữ liệu kinh tế - xã hội và hỗ trợ ra chiến lược hiệu quả.

#### 1.4. Tài nguyên sử dụng

Bộ dữ liệu *Country Socioeconomic Data* được lấy từ Kaggle

Công cụ lập trình và phần mềm:

- Ngôn ngữ lập trình: Python
- Thư viện hỗ trợ:
  - Scikit-learn: Thực thi thuật toán OPTICS và các thuật toán phân cụm khác.
  - Pandas và Numpy: Xử lý và phân tích dữ liệu.
  - Matplotlib và Seaborn: Trực quan hóa dữ liệu.
- Môi trường làm việc: Google Colab



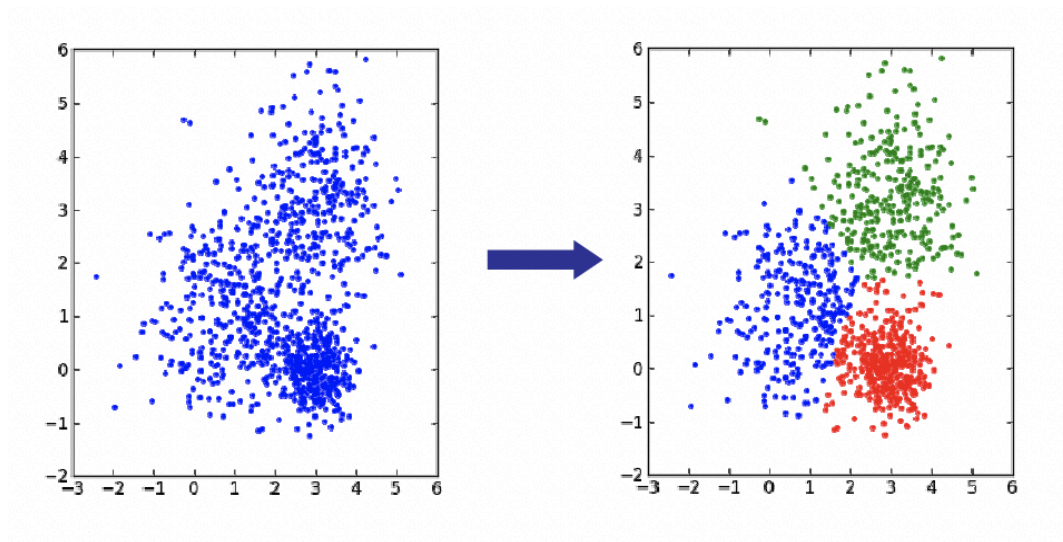
## CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

### 2.1. Giới thiệu về phân cụm dữ liệu

#### 2.1.1. Giới thiệu tổng quan về phân cụm dữ liệu

Phân cụm dữ liệu là một kỹ thuật học máy không giám sát (Unsupervised Learning) , nhằm gom nhóm các đối tượng dữ liệu thành các cụm nhằm tìm ra các cụm có tính tương đồng dựa trên một số tiêu chí đặc trưng nào đó và các đối tượng thuộc các cụm khác nhau thì có tính khác biệt. Phân cụm dữ liệu được sử dụng để khám phá và nhận diện các cụm hoặc mẫu dữ liệu tự nhiên tiềm ẩn trong tập dữ liệu lớn. Mục tiêu của quá trình này là trích xuất thông tin và tri thức hữu ích nhằm hỗ trợ việc ra quyết định hiệu quả hơn.

Độ tương tự giữa các đối tượng được đo lường dựa trên một tiêu chí cụ thể, được xác định từ trước và phù hợp với mục tiêu phân tích. Khác với phân loại (classification), trong đó các đặc điểm của lớp (hay cụm) đã được biết trước và dùng để gán nhãn cho các đối tượng mới. Quá trình phân cụm dữ liệu thì không biết trước về tính chất của các cụm. Thay vào đó, nó dựa vào mối quan hệ giữa các đối tượng để phát hiện các nhóm có tính tương tự nhau, từ đó hình thành các cụm có đặc trưng riêng biệt dựa trên một thước đo nhất định.



Hình 1. Phân cụm dữ liệu

*Nguồn: Nguyễn, A. T. (2023). Học không giám sát. UEH LMS*

Các ứng dụng của phân cụm dữ liệu rất đa dạng, có thể kể đến như phân tích phân khúc khách hàng, phân tích mạng xã hội để phân tích hành vi, xây dựng hệ thống

đề xuất trong lĩnh vực thương mại điện tử, phân loại bệnh nhân theo các đặc điểm sức khỏe trong y tế và nhiều lĩnh vực khác.

Loại thuật toán	Đặc điểm	Các thuật toán phổ biến
<b>Partitioning Clustering</b> (Phân cụm phân hoạch)	Phân cụm phân hoạch chia dữ liệu có $n$ phần tử thành $k$ cụm với giá trị $k$ được định, trong đó mỗi cụm được đại diện bởi một trung tâm hoặc điểm centroid. Mỗi điểm dữ liệu sẽ được gán vào cụm mà nó gần nhất với centroid của cụm đó. Khoảng cách thường được sử dụng để xác định độ tương đồng là khoảng cách Euclidean, có thể sử dụng các loại khoảng cách khác tùy thuộc vào đặc điểm của dữ liệu. Sau khi gán đối tượng vào các cụm, centroid của mỗi cụm sẽ được cập nhật dựa trên các đối tượng mà nó đang chứa.	K-Means, K-medoids (PAM, CLARA, CLARANS), Fuzzy C-Means
<b>Hierarchical Clustering</b> (Phân cụm phân cấp)	Phân cụm phân cấp là phương pháp mà trong đó các đối tượng được phân chia thành các cụm có cấu trúc phân cấp, thường được biểu diễn bằng cây phân cấp (dendrogram). Nó tạo ra một cấu trúc phân cấp giữa các cụm, cho phép người dùng dễ dàng hiểu mối quan hệ giữa các cụm và đối tượng. Mỗi cụm có thể được chia thành các cụm con, tạo thành nhiều cấp độ phân chia.	Diana, ANGNES, BIRCH, CAMELEON
<b>Density-Based</b> (Phân cụm theo mật độ)	Phân cụm dựa trên mật độ là một phương pháp phân tích dữ liệu dùng để nhóm các điểm dữ liệu thành các cụm dựa trên mật độ của chúng trong không gian dữ liệu. Ý tưởng chính của phương pháp này là các điểm dữ liệu nằm gần nhau trong các khu vực dày đặc (có mật độ điểm cao) sẽ được gộp thành cụm, trong khi các điểm nằm ở những khu vực thưa thớt sẽ bị coi là nhiễu (hoặc không thuộc cụm nào).	DBSCAN, OPTICS, DenClue
<b>Model-Based</b> (Phân cụm theo mô hình)	Phân cụm dựa trên mô hình là một phương pháp phân cụm sử dụng các mô hình thống kê để biểu diễn cấu trúc của dữ liệu. Thay vì chỉ dựa vào khoảng cách giữa các điểm dữ liệu như trong K-Means hoặc phân cụm phân cấp, phân cụm dựa trên mô hình giả định rằng dữ liệu được sinh ra từ một tập hợp các phân phối xác suất. Các điểm dữ liệu trong cùng một cụm được cho	EM, SOM, COBWEB

	là tuân theo cùng một phân phối xác suất, chẳng hạn như phân phối Gaussian	
--	--	--

Bảng 1. Các loại thuật toán phân cụm

### 2.1.2. Giới thiệu tổng quan về phân cụm theo mật độ

Các phương pháp phân cụm hầu hết đều hoạt động theo cách nhóm các điểm dữ liệu dựa trên khoảng cách của các điểm dữ liệu. Tuy nhiên, các phương pháp này chỉ có thể xác định các cụm hình cầu và gặp trở ngại khá lớn trong việc phát hiện các cụm có hình dạng bất kỳ. Để giải quyết vấn đề này, một số phương pháp phân cụm đã được phát triển dựa trên khái niệm về mật độ, gọi chung là phương pháp phân cụm dựa trên mật độ ((Density-Based Clustering). Đây là một phương pháp mạnh mẽ, phù hợp cho việc phát hiện các cụm có hình dạng bất kỳ và xử lý các điểm nhiễu.

Phân cụm theo mật độ (Density-Based Clustering) là một phương pháp học máy không giám sát, đầu tiên ta sẽ tìm hiểu các khái niệm cơ bản trong việc phân cụm dựa trên mật độ (Ester et al., 1996):

- **Định nghĩa 1:** Vùng lân cận Eps (bán kính  $\epsilon$ ) của một điểm  $p$ , ký hiệu là  $N_{Eps}(p)$ , được định nghĩa như sau:

$$N_{Eps}(p) = \{q \in D \mid dist(p, q) \leq Eps\}$$

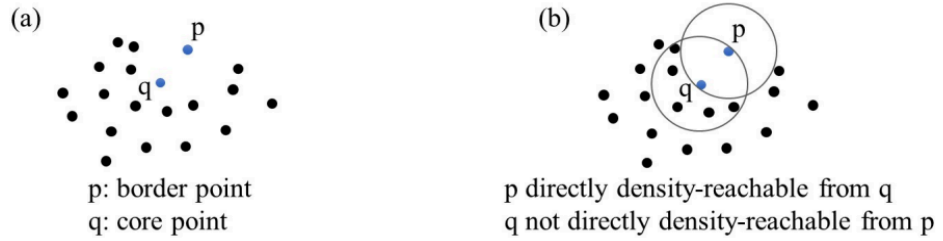
Một cách tiếp cận đơn giản được yêu cầu cho mỗi điểm trong một cụm là phải có ít nhất số điểm tối thiểu (MinPts) trong vùng lân cận Eps của điểm đó. Tuy nhiên, cách tiếp cận này không hoàn toàn hiệu quả vì có 2 loại điểm trong một cụm: điểm lõi (core points) và điểm biên (border points). Nói một cách đơn giản, điểm lõi và điểm biên có thể giải thích như sau:

- **Điểm lõi (Core points):** Là những điểm mà trong phạm vi vùng lân cận Eps xung quanh chúng, số lượng điểm lớn hơn hoặc bằng ngưỡng MinPts. Các điểm này có thể xem là “trung tâm” của một cụm vì chúng tạo ra được một nhóm các điểm bao quanh chúng.
- **Điểm biên (Border points):** Là những điểm nằm trong phạm vi vùng lân cận Eps của các điểm lõi, nhưng xung quanh chúng không có đủ số lượng điểm MinPts để chúng có thể trở thành điểm lõi, tuy chúng không phải là “trung tâm” của cụm, nhưng chúng vẫn là một phần của cụm.

- **Định nghĩa 2:** *Tiếp cận trực tiếp dựa trên mật độ (Directly Density-reachable)* Một điểm  $p$  được gọi là tiếp cận trực tiếp dựa trên mật độ từ một  $q$  theo bán kính  $\epsilon$  và  $MinPts$  nếu:

1.  $p \in N_{Eps}(q)$
2.  $|N_{Eps}(p)| \geq MinPts$  (điều kiện để trở thành điểm lõi)

Điều này có nghĩa là, nếu  $q$  là một điểm lõi và  $p$  nằm trong vùng lân cận  $Eps$  (bán kính  $\epsilon$ ) của  $q$ , và trong vùng lân cận  $Eps$  có số lượng điểm lớn hơn hoặc bằng ngưỡng  $MinPts$  thì  $p$  sẽ được gọi là tiếp cận dựa trên mật độ từ  $q$ .



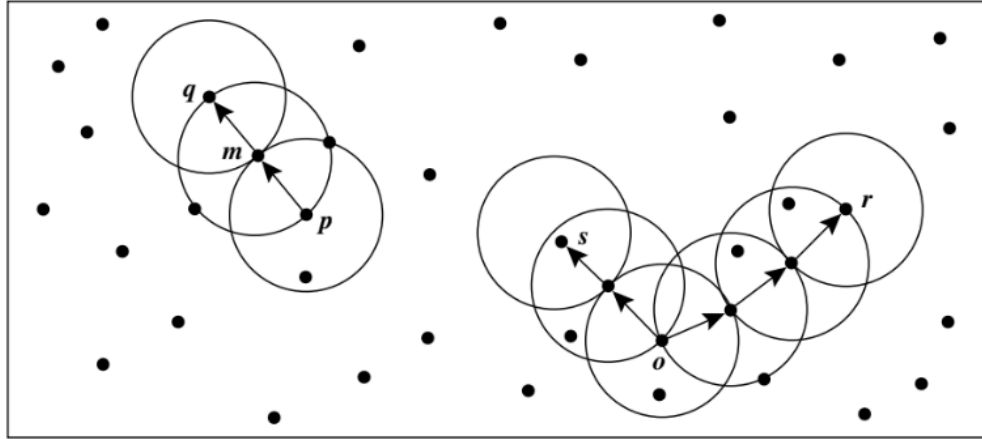
Hình 2. Điểm lõi (Core points) và điểm biên (Border points)

Nguồn: Gao et al. (2023)

- **Định nghĩa 3:** *Tiếp cận dựa trên mật độ (Density-reachable)* Một điểm  $p$  được gọi là tiếp cận dựa trên mật độ từ một  $q$  theo bán kính  $\epsilon$  và  $MinPts$  nếu tồn tại một chuỗi các điểm  $p_1, \dots, p_n, p_1 = q, p_n = p$  sao cho mỗi điểm  $p_{i+1}$  có thể tiếp cận trực tiếp dựa trên mật độ từ điểm  $p_i$ .

Tiếp cận dựa trên mật độ có nghĩa là có thể di chuyển từ điểm này đến điểm khác thông qua một chuỗi các điểm lõi (core points) mà mỗi điểm trong chuỗi đều có ít nhất  $MinPts$  điểm nằm trong phạm vi vùng lân cận  $Eps$  của nó. Hình 3 mô tả cách tiếp cận dựa trên mật độ.

- **Định nghĩa 4:** *Kết nối dựa trên mật độ (Density-connected)* Một điểm  $s$  được gọi là kết nối dựa trên mật độ với điểm  $r$  theo bán kính  $\epsilon$  và  $MinPts$  nếu tồn tại một điểm  $o$  sao cho cả hai điểm  $s$  và  $r$  đều có thể tiếp cận dựa trên mật độ từ điểm  $o$ .



Hình 3. Density-reachable và Density-connected

Nguồn: Han (2011)

Nói cách khác, s và r được kết nối thông qua một chuỗi các điểm mà các điểm ấy đều có thể kết nối dựa trên mật độ từ điểm trước đó trong chuỗi. Điều này giúp hình thành sự kết nối giữa các điểm không thể trực tiếp tiếp cận theo mật độ từ nhau, nhưng lại có thể kết nối thông qua một điểm chung.

- **Định nghĩa 5:** *Cụm (Cluster)* Với  $D$  là một cơ sở dữ liệu chứa các điểm. Một cụm  $C$  theo bán kính  $\epsilon$  và MinPts là một tập không rỗng của  $D$  thỏa mãn các điều kiện sau:
  1. *Tính cực đại (Maximality):* Với mọi điểm dữ liệu  $p$  và  $q$  thuộc tập  $D$ . Nếu điểm  $p$  thuộc phân cụm  $C$  và  $q$  có thể tiếp cận dựa trên mật độ từ  $p$ , thì  $q$  cũng là một điểm thuộc phân cụm  $C$ .
  2. *Tính liên thông (Connectivity):* Với mọi điểm dữ liệu  $p$  và  $q$  thuộc phân cụm  $C$ ,  $p$  có thể kết nối dựa trên mật độ với  $q$ .
- **Định nghĩa 6:** *Nhiều (Noise)* Giả sử  $C_1, \dots, C_k$  là các cụm của  $D$  với các tham số  $Eps_i$  và  $MinPts_i$  với  $i = 1, 2, \dots, k$ .

$$noise = \{p \in D \mid \forall i : p \notin C_i\}$$

Nhiều được định nghĩa là tập hợp các điểm thuộc  $D$  không thuộc bất kỳ cụm nào, tức là không đạt điều kiện để là điểm lõi hoặc điểm biên.

## 2.2. Thuật toán OPTICS

### 2.2.1. Giới thiệu tổng quan về thuật toán OPTICS

- Lịch sử hình thành và phát triển

Thuật toán OPTICS (Ordering Points To Identify the Clustering Structure) là thuật toán phân cụm dữ liệu kinh điển, được giới thiệu bởi Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel và Jörg Sander vào năm 1999. Thuật toán này được phát triển để khắc phục một số hạn chế của các phương pháp phân cụm truyền thống như DBSCAN, đặc biệt trong việc xử lý dữ liệu có mật độ không đồng đều và cấu trúc phức tạp.

Tiền thân của OPTICS là DBSCAN, một trong những thuật toán phân cụm dựa trên mật độ đầu tiên và phổ biến nhất. DBSCAN tiến hành phân cụm dữ liệu dựa trên khoảng cách Euclid, yêu cầu xác định trước hai tham số quan trọng là  $\epsilon$  (bán kính lân cận) và MinPts (số lượng điểm tối thiểu trong một cụm). Tuy nhiên, hiệu quả của DBSCAN phụ thuộc lớn vào việc lựa chọn các tham số này, dẫn đến hạn chế khi xử lý các tập dữ liệu có mật độ không đồng đều.

Kế thừa các ưu điểm từ DBSCAN, thuật toán OPTICS mở rộng thêm các khả năng vượt trội khác. Thay vì chỉ phân cụm dữ liệu dựa trên một bán kính cố định, OPTICS tiến hành sắp xếp các cụm dữ liệu theo thứ tự tăng dần và xác định các điểm dữ liệu láng giềng phù hợp khi xem xét một bán kính tối thiểu. Thuật toán này cho phép phát hiện các cụm có hình dạng tùy ý và xử lý hiệu quả dữ liệu có mật độ thay đổi phức tạp.

Một điểm nổi bật khác của OPTICS là khá ổn định với dữ liệu nhiễu (Noise) và các điểm ngoại lai (Outliers). Ngoài ra, thuật toán còn cung cấp một biểu đồ khả năng tiếp cận (reachability plot), cho phép người dùng trực quan hóa cấu trúc cụm và dễ dàng chọn ngưỡng để xác định các cụm.

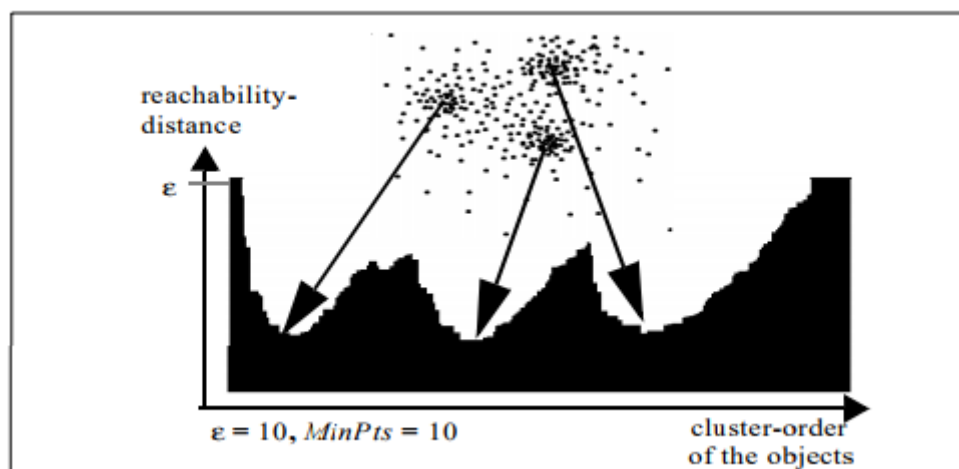
- Giới thiệu khái niệm

Thuật toán OPTICS là thuật toán phân cụm dữ liệu dựa trên mật độ, được mở rộng từ thuật toán DBSCAN, giảm bớt tham số đầu vào. OPTICS hoạt động dựa trên ý tưởng sắp xếp các điểm dữ liệu dựa trên mật độ, từ đó phân cụm tự động các cụm dữ liệu dựa trên thứ tự này.

OPTICS không yêu cầu bán kính  $\epsilon$  cố định như là DBSCAN, nhưng vẫn sử dụng tham số  $\text{max\_eps}$  để giới hạn khoảng cách tối đa trong việc tìm kiếm các điểm lân cận. Điều này giúp OPTICS có sự linh hoạt trong việc phát hiện các cụm có mật độ không đồng đều. Ngoài ra, OPTICS giới thiệu một số khái niệm quan trọng giúp hiểu rõ hơn về cách thức hoạt động của thuật toán:

- *Khoảng cách lõi (Core Distance)*: Khoảng cách tối thiểu cần thiết để một điểm dữ liệu  $p$  được coi là điểm lõi. Nếu  $p$  không phải là điểm lõi, thì khoảng cách lõi của nó không được xác định.

- *Khoảng cách khả năng tiếp cận (Reachability Distance)*: Khoảng cách có thể tiếp cận của điểm  $q$  so với một điểm  $p$  khác là khoảng cách nhỏ nhất sao cho  $q$  có thể tiếp cận trực tiếp từ  $p$  nếu  $p$  là điểm lõi.
- *Biểu đồ khả năng tiếp cận (Reachability - plot)*: Đồ thị biểu diễn thứ tự các điểm và khoảng cách khả năng tiếp cận của chúng, giúp trực quan hóa cấu trúc cụm. Các điểm dữ liệu trong cùng một cụm có khoảng cách khả năng tiếp cận đến các lân cận gần nhất của chúng thấp, tạo thành các “thung lũng” trong đồ thị. Độ sâu của “thung lũng” tỷ lệ thuận với mật độ của cụm.



Hình 4. Reachability Plot

*Nguồn: Redinger & Hunner (2017)*

- Mục tiêu và ứng dụng chính

Mục tiêu chính của OPTICS là phân cụm dữ liệu bằng cách tìm các vùng có mật độ đối tượng cao hơn các vùng lân cận, đồng thời khắc phục hạn chế của thuật toán DBSCAN khi xử lý dữ liệu có mật độ không cố định. Bên cạnh đó, OPTICS giúp hiểu sâu hơn về cấu trúc tổng thể của dữ liệu nhờ việc tạo ra dãy thứ tự các điểm và cũng dễ dàng phát hiện ra Outliers.

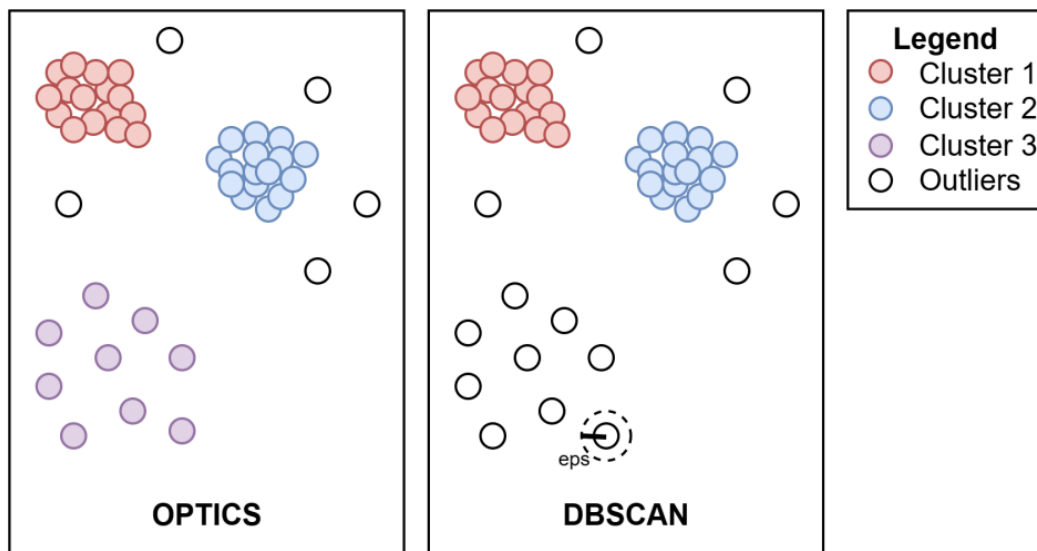
- Một số ứng dụng chính của OPTICS:

- Phân cụm dữ liệu liên quan đến địa lý, thời gian như xác định mật độ dân cư, phân tích hành vi giao dịch tài chính...
- Phân cụm dữ liệu khoa học như dữ liệu y sinh học...

- Phân tích dữ liệu trong lĩnh vực E-commerce như phân tích đánh giá của khách hàng...
- Xử lý ảnh

### 2.2.2. Mô tả thuật toán

Để có thể dễ hiểu hơn cách thuật toán OPTICS hoạt động, thay vì chỉ mô tả bằng các khái niệm, nhóm sẽ sử dụng minh họa bằng hình ảnh để giải thích các thuật ngữ một cách đơn giản nhất.



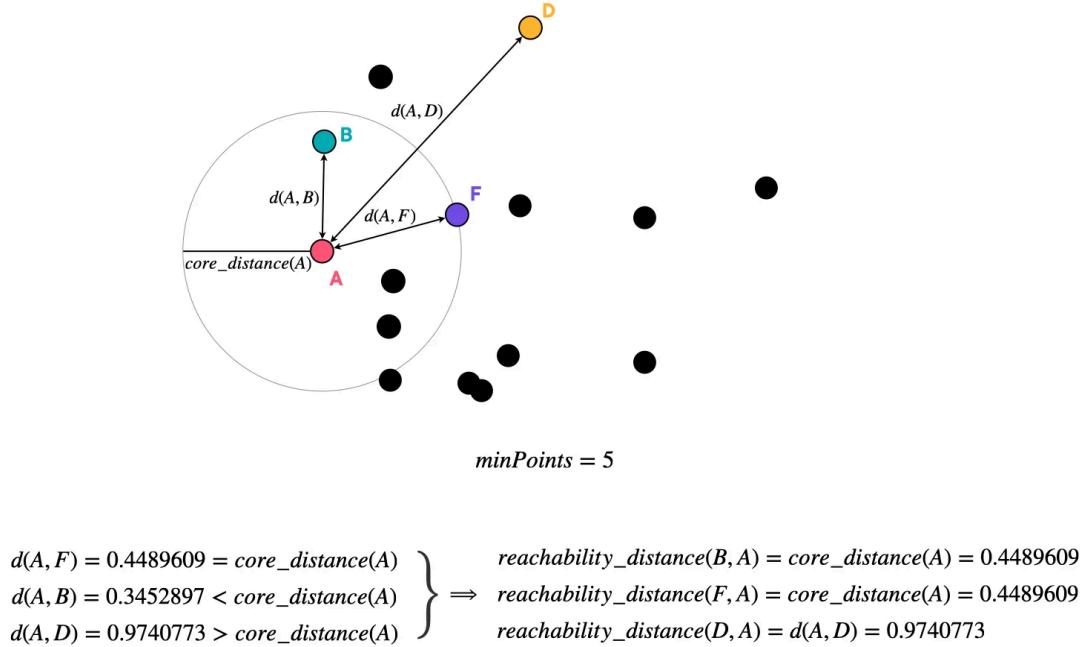
Hình 5. Minh họa sự khác biệt giữa OPTICS và DBSCAN

Nguồn: Tran et al. (2023)

Đầu tiên, hãy cùng giải thích rõ hơn tại sao OPTICS lại được coi là phần mở rộng của thuật toán DBSCAN. Như chúng ta có thể thấy qua hình minh họa 5, DBSCAN hoạt động dựa trên hai tham số chính là bán kính  $\epsilon$  (Eps) và số lượng điểm tối thiểu MinPts. Khi xét một điểm dữ liệu, nếu trong phạm vi Eps của nó không có đủ số lượng MinPts, điểm đó sẽ không được coi là điểm lõi và sẽ bị bỏ qua trong quá trình phân cụm. Nếu tăng Eps lên, cụm Cluster 3 có thể được xác định, tuy nhiên, nó có khả năng cũng sẽ khiến cho hai cụm ở phía trên (Cluster 1 và Cluster 2) cũng như các outlier hợp nhất lại với nhau. Điều này cho thấy hạn chế lớn của thuật toán DBSCAN khi hoạt động với dữ liệu có mật độ hoặc kích thước khác nhau, rất khó để có thể xác định một tham số Eps hợp lý.



Một vấn đề cũng từ đó được đặt ra là: Làm thế nào để có thể xác định những cụm khác nhau với khoảng cách giữa các điểm trong mỗi cụm là khác nhau. Đây chính là điểm vượt trội của OPTICS so với DBSCAN.



Hình 6. So sánh khoảng cách giữa các điểm

Nguồn: Plakalovic, A. (2023)

Để xác định tham số khoảng cách hợp lý, OPTICS sẽ so sánh khoảng cách giữa các điểm với các khoảng cách liên quan đến mật độ lân cận. Ví dụ trong hình 6, với A là điểm lõi được chọn để xem xét, tính toán khoảng cách lõi (*core distance*) và khoảng cách khả năng tiếp cận (*reachability distance*) và các điểm B, D, F là các điểm lân cận của A. Vòng tròn bao quanh điểm A là bán kính có đủ  $\text{minPoints} = 5$  điểm lân cận, bao gồm cả điểm A.

Khoảng cách lõi của điểm A ( $\text{core\_distance}(A)$ ) là khoảng cách từ điểm A đến điểm xa nhất trong tập hợp  $\text{minPoints} - 1$  điểm gần nhất xung quanh nó. Trong ví dụ này,  $\text{minPoints} = 5$ , nên  $\text{core\_distance}(A)$  sẽ được tính dựa trên điểm xa nhất trong 4 điểm lân cận gần nhất của A. Có thể thấy  $d(A, F) = \text{core\_distance}(A) = 0.4489609$  vì trong 4 điểm lân cận gần A nhất, điểm F có khoảng cách xa nhất.

Sau khi có được  $\text{core\_distance}(A)$ , ta sẽ tiếp tục xem xét để tính khoảng cách khả năng tiếp cận (*reachability distance*). Trước khi tính khoảng cách khả năng tiếp

cận, ta sẽ tính khoảng cách giữa A và các điểm khác, được ký hiệu là  $d(A, X)$ . Trong ví dụ minh họa, ta có khoảng cách từ A đến các điểm B, D, F lần lượt là:

- $d(A, B) = 0.3452897$
- $d(A, D) = 0.9740773$
- $d(A, F) = 0.4489609$

Để có thể tính được khoảng cách khả năng tiếp cận của từng điểm so với A ( $reachability\_distance(X, A)$ ), ta sẽ so sánh  $core\_distance(A)$  với khoảng cách  $d(A, X)$  của từng điểm dữ liệu, cụ thể:

- Nếu khoảng cách  $d(A, X) \leq core\_distance(A)$ , nghĩa là điểm X nằm trong vùng mật độ lõi của điểm A. Khi đó, khoảng cách khả năng tiếp cận của X sẽ được xác định là:

$$reachability\_distance(X, A) = core\_distance(A)$$

- Nếu khoảng cách  $d(A, X) > core\_distance(A)$ , nghĩa là điểm X nằm ngoài vùng mật độ lõi của điểm A. Khi đó, khoảng cách khả năng tiếp cận của X sẽ được xác định là:

$$reachability\_distance(X, A) = d(A, X)$$

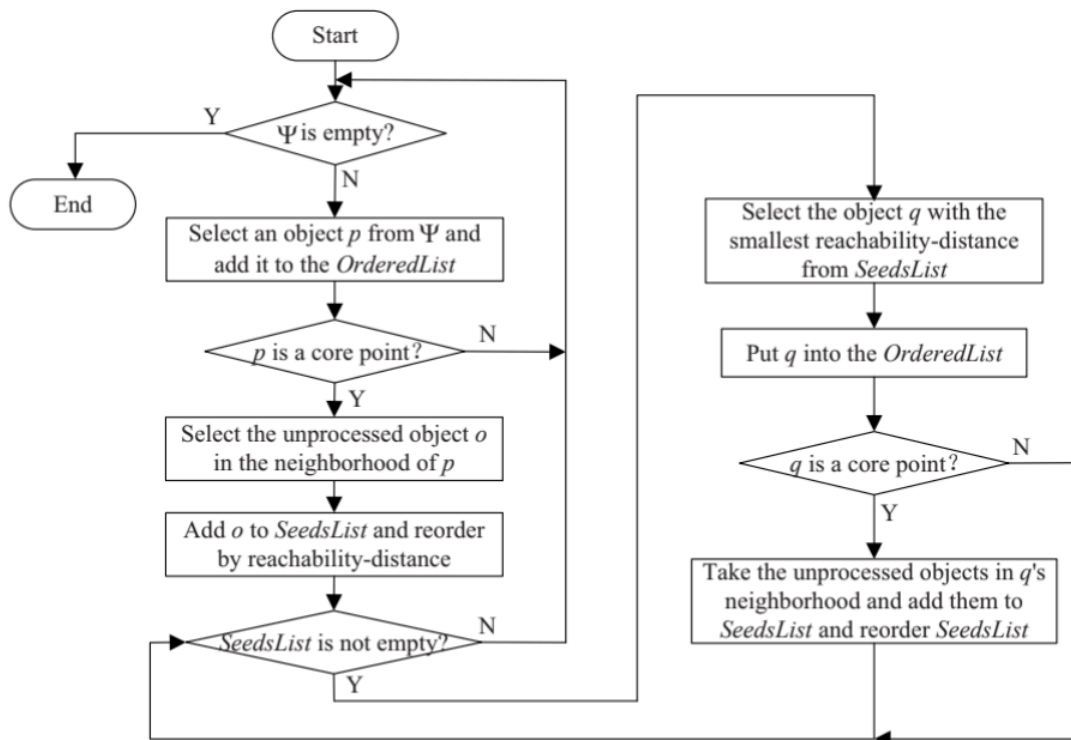
Ở hình 6, ta có thể lần lượt nhận xét khả năng tiếp cận của từng điểm B, D, F so với A như sau:

- $d(A, B) = 0.3452897 < core\_distance(A) = 0.4489609$   
 $\Rightarrow reachability\_distance(B, A) = core\_distance(A) = 0.4489609$
- $d(A, D) = 0.9740773 > core\_distance(A) = 0.4489609$   
 $\Rightarrow reachability\_distance(D, A) = d(A, D) = 0.9740773$
- $d(A, F) = core\_distance(A) = 0.4489609$   
 $\Rightarrow reachability\_distance(F, A) = core\_distance(A) = 0.4489609$

Việc xác định khoảng cách khả năng tiếp cận (*reachability distance*) của từng điểm dữ liệu đóng vai trò quan trọng trong việc sắp xếp thứ tự các điểm dữ liệu trong thuật toán OPTICS, không chỉ giúp nâng cao hiệu quả xử lý mà còn mang lại nhiều lợi ích đáng kể:

- *Phân cụm linh hoạt*: Thứ tự sắp xếp các điểm phản ánh mật độ các điểm dữ liệu trong cục bộ một vùng, cho phép phân tích cụm ở nhiều cấp độ khác nhau, đặc biệt hữu ích với dữ liệu có mật độ thay đổi.
- *Tạo đồ thị khoảng cách tiếp cận*: Trực quan hóa, giúp phát hiện các cụm và phân bố mật độ điểm, từ đó nhận diện cụm và nhiễu dễ dàng.
- *Cho phép điều chỉnh ngưỡng tùy ý*: OPTICS cho phép điều chỉnh ngưỡng sau khi đã sắp xếp điểm, giúp tối ưu phân cụm và linh hoạt thử nghiệm nhiều mật độ khác nhau.

### 2.2.3. Thuật toán hoạt động



Hình 7. Mô tả cách thuật toán hoạt động

Nguồn: Wang et al. (2021)

#### 1. Khởi tạo:

- Tạo một danh sách đã sắp xếp (Ordered file) để ghi kết quả sắp xếp, danh sách ban đầu rỗng.
- Đánh dấu tất cả đối tượng là chưa được xử lý.

## 2. Lặp qua tất cả đối tượng:

- Duyệt qua từng đối tượng trong cơ sở dữ liệu, nếu đối tượng chưa được xử lý, tiến hành mở rộng cụm.

## 3. Mở rộng thứ tự cụm:

- Tạo một Seed list rỗng. Seed list là danh sách hàng đợi ưu tiên dùng để theo dõi các điểm lân cận trong quá trình mở rộng cụm. Các đối tượng trong Seed list được sắp xếp theo thứ tự reachability distance tăng dần.
- Xác định vùng lân cận (neighbors) của đối tượng hiện tại.
- Đánh dấu đối tượng hiện tại là đã được xử lý.
- Tính toán CoreDist của đối tượng hiện tại đến neighbor gần nhất sao cho số lượng lân cận đủ lớn (tối thiểu là bằng Minpts).
- Thêm đối tượng hiện tại vào danh sách đã sắp xếp.
- Nếu đối tượng hiện tại được xác định là lõi (core point), tiến hành mở rộng cụm:
  - Cập nhật Seed list bao gồm đối tượng hiện tại và các lân cận của đối tượng đó.
  - Chọn đối tượng có reachability distance nhỏ nhất trong Seed list, xác định vùng lân cận và CoreDist của đối tượng được chọn này.
  - Thêm đối tượng được chọn vào danh sách Ordered file.
  - Nếu đối tượng được chọn được xác định là lõi, thêm các lân cận mà chưa được xử lý của đối tượng được chọn đó vào Seed list.
  - Sắp xếp lại thứ tự trong Seed list.

- Lặp lại quá trình cho đến khi Seed list rỗng, hoàn tất xác định cụm hiện tại.

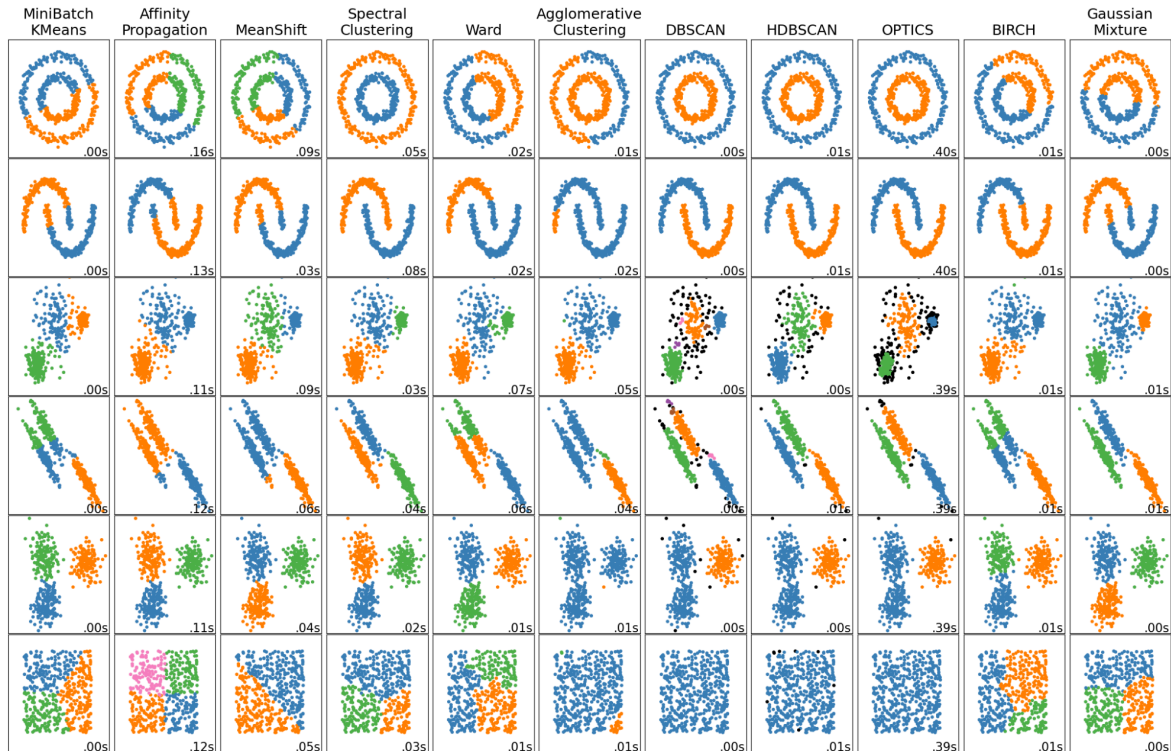
4. Lặp lại bước 3. Thuật toán kết thúc và đóng Ordered file khi duyệt qua tất cả đối tượng trong cơ sở dữ liệu.

#### 2.2.4. So sánh thuật toán OPTICS với một số thuật toán phân cụm khác

Tiêu chí so sánh	DBSCAN	OPTICS	HDBSCAN	DENCLUE
<b>Nguyên lý hoạt động</b>	Phân cụm dựa trên mật độ của các điểm gần nhau	Sắp xếp các điểm theo khoảng cách tiếp cận và phân cụm theo mật độ	Phân cấp dựa trên mật độ thay đổi	Phân cụm quanh các mode mật độ cao
<b>Khả năng phát hiện cụm ở mật độ khác nhau</b>	Không thể, chỉ phù hợp với cụm mật độ đồng nhất	Có thể, thực hiện hiệu quả cho cụm mật độ thay đổi	Có thể, tự động tối ưu mật độ	Có thể, sử dụng ước lượng mật độ
<b>Khả năng phát hiện nhiễu</b>	Tốt	Tốt	Tốt, độ chính xác cao	Khá tốt, độ hiệu quả phụ thuộc vào thiết lập hàm mật độ hạt nhân (Kernel Density Estimation)
<b>Tham số đầu vào</b>	Epsilon ( $\epsilon$ ), số điểm tối thiểu (MinPts)	Số điểm tối thiểu (MinPts), không cần epsilon cố định	MinPts, không cần epsilon cố định	Ngưỡng mật độ và hàm mật độ hạt nhân
<b>Độ phù hợp với từng dạng dữ liệu</b>	Cụm có mật độ cố định, hình dạng bất kỳ	Cụm có mật độ và hình dạng bất kỳ	Dữ liệu phân cấp, mật độ bất kỳ	Dữ liệu đa chiều, cụm phức tạp
<b>Độ phức tạp</b>	$O(n \log n)$ , xử lý chậm với dữ liệu cực lớn	$O(n \log n)$ , hiệu quả trên dữ liệu lớn	$O(n \log n)$ , phụ thuộc vào cấu trúc dữ liệu	$O(n \log n)$ , tùy thuộc vào số lượng mode
<b>Khả năng ứng dụng</b>	Phân tích dữ liệu địa lý, nhận diện mẫu trong ảnh, xử lý dữ liệu sinh học	Phân tích địa lý phức tạp, dữ liệu sinh học đa mật độ, bài toán cần khám phá cụm mà không	Phân cụm dữ liệu đa mật độ, phân tích sinh học đa cấp, dữ liệu tài chính xã hội	Phân tích hình ảnh đa chiều, nghiên cứu sinh học, dữ liệu phức tạp

	biết trước tham số cố định		
--	----------------------------	--	--

Bảng 2. So sánh OPTICS với các thuật toán phân cụm khác



Hình 8. Sự khác biệt giữa các thuật toán phân cụm

Nguồn: Scikit-learn. (n.d.)

## 2.3. Đánh giá thuật toán phân cụm

### 2.3.1 Cơ sở lý thuyết

Để đánh giá chất lượng của phân cụm, có thể dựa trên 2 tiêu chí chung là sự kết dính (cohesion) để đo mức độ tương đồng trong cụm, và sự khác biệt (distinction) để đo sự mức độ tương đồng giữa các cụm.

Mục tiêu của các phương pháp phân cụm đó chính là có thể đạt được mức độ tương đồng cao trong một cụm và giữa các cụm khác nhau thì có mức độ tương đồng thấp, tức là các đối tượng trong cùng một cụm sẽ giống nhau và các đối tượng khác cụm thì sẽ khác nhau.

Để đo sự tương đồng giữa hai đối tượng có n thuộc tính *định lượng*, ví dụ  $x=(x_1, x_2, \dots, x_n)$  và  $y=(y_1, y_2, \dots, y_n)$  thì có thể sử dụng 2 chỉ số là cosine hoặc tích vô hướng giữa 2 đối tượng.

Cosine định lượng cụ thể cosin của góc giữa các vector, cung cấp phép đo hướng thay vì độ lớn của chúng. Cách này hữu ích trong các trường hợp mà độ lớn của các vector ít được quan tâm và tập trung vào hướng của chúng. Công thức:

$$sim(x, y) = \frac{\sum_{i=1}^n x_i \cdot y_i}{\sqrt{\sum_{i=1}^n x_i^2} \cdot \sqrt{\sum_{i=1}^n y_i^2}}$$

Tích vô hướng (scalar product) là một phép đo độ tương đồng theo hướng và độ lớn của các vector. Trong trường hợp mà cả hướng và kích thước các vector đều mang thông tin có ý nghĩa thì tích vô hướng là một phép đo hữu ích. Giá trị của tích vô hướng giao động từ âm vô cùng đến dương vô cùng, giá trị âm biểu thị chúng có hướng ngược nhau và giá trị dương thì biểu thị chúng cùng hướng, giá trị 0 khi các vector vuông góc với nhau. Công thức:

$$sim(x, y) = \sum_{i=1}^n x_i \cdot y_i$$

### 2.3.2 Chỉ số Silhouette

Chỉ số Silhouette dùng để đánh giá chất lượng kết quả phân cụm dữ liệu, nó đo lường mức độ tương đồng giữa các đối tượng trong cùng một cụm so với các đối tượng trong cùng một cụm các cụm khác.

Chỉ số Silhouette được tính theo công thức sau:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Trong đó:

- $a(i)$  là khoảng cách trung bình từ đối tượng  $i$  đến mọi đối tượng khác trong cùng một cụm.
- $b(i)$  là khoảng cách trung bình từ đối tượng  $i$  đến tất cả các đối tượng khác trong cụm gần nhất.

Chỉ số Silhouette có giá trị nằm trong khoảng từ -1 đến 1, giá trị càng gần 1 cho biết đối tượng được phân cụm tốt, nghĩa là nó gần với các đối tượng trong cụm của nó hơn là các đối tượng trong các cụm khác. Ngược lại, giá trị càng gần -1 cho biết đối tượng có thể đã được phân vào cụm không phù hợp.

Ngoài ra, biểu đồ silhouette càng nhiều dao (càng bầu) càng tốt, tức nó đều nhau, còn nhiều kiếm thì có nghĩa là nó cho kết quả phân cụm xấu.



## CHƯƠNG 3. ÁP DỤNG THUẬT TOÁN TRÊN BỘ DỮ LIỆU

### 3.1. Tổng quan bộ dữ liệu

#### 3.1.1 Nguồn gốc bộ dữ liệu

Bộ dữ liệu “Country Socioeconomic Data” được lấy từ một cuộc thi cùng tên do tác giả AYOUB AZIAMIMER tổ chức trên Kaggle. Bộ dữ liệu cung cấp thông tin chi tiết về tình hình kinh tế - xã hội của 167 quốc gia. Dữ liệu bao gồm 10 thuộc tính và 167 quan sát, tương ứng với từng quốc gia.

#### 3.1.2 Vấn đề được đặt ra trong bộ dữ liệu

HELP International là một tổ chức phi chính phủ nhân đạo quốc tế cam kết chống lại đói nghèo và cung cấp các tiện nghi cơ bản cũng như cứu trợ cho người dân ở các quốc gia kém phát triển trong thời kỳ thảm họa và thiên tai.

HELP International đã huy động được khoảng 10 triệu đô la. Hiện tại, CEO của tổ chức phi chính phủ này cần quyết định cách sử dụng số tiền này một cách chiến lược và hiệu quả. Vì vậy, CEO phải đưa ra quyết định lựa chọn các quốc gia đang cần viện trợ cấp thiết nhất. Nhiệm vụ mà nhóm phải đối mặt là phân loại các quốc gia dựa trên một số yếu tố kinh tế - xã hội và sức khỏe quyết định đến sự phát triển tổng thể của quốc gia. Sau đó đưa ra những đề xuất những quốc gia mà CEO nên tập trung hỗ trợ nhiều nhất.

#### 3.1.3 Mô tả bộ dữ liệu

Bộ dữ liệu bao gồm các thuộc tính được liệt kê sau đây:

Thuộc tính	Mô tả	Kiểu dữ liệu
country	Tên quốc gia	object
child_mort	Tỷ lệ tử vong của trẻ em dưới 5 tuổi trên 1000 trẻ em	float64
exports	Xuất khẩu hàng hóa và dịch vụ, tính theo % trên GDP bình quân đầu người	float64
health	Tổng chi tiêu y tế theo % tuổi trong GDP bình quân đầu người	float64
imports	Nhập khẩu hàng hóa và dịch vụ. Tính theo % tuổi trong Tổng GDP	float64
income	Thu nhập ròng mỗi người	int64

inflation	Đo lường tốc độ tăng trưởng hàng năm của Tổng GDP	float64
life_expec	Số năm trung bình mà một đứa trẻ mới sinh có thể sống được nếu mô hình tử vong hiện tại được giữ nguyên	float64
total_fer	Số con mà mỗi phụ nữ sẽ sinh ra nếu tỷ suất sinh theo độ tuổi hiện tại không đổi	float64
gdpp	GDP bình quân đầu người. Được tính bằng Tổng GDP chia cho tổng dân số.	int64

Bảng 3. Mô tả thuộc tính và kiểu dữ liệu trong dữ liệu

## 3.2. Tiền xử lý dữ liệu

### 3.2.1 Mô tả dữ liệu

Một số dòng dữ liệu trong bộ dữ liệu country data:

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553
1	Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090
2	Algeria	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460
3	Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530
4	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200

Hình 9. Một số dòng dữ liệu trong bộ dữ liệu

*Nguồn: Tác giả*

Mô tả dữ liệu:

	count	mean	std	min	25%	50%	75%	max
child_mort	167.0	38.270060	40.328931	2.6000	8.250	19.30	62.10	208.00
exports	167.0	41.108976	27.412010	0.1090	23.800	35.00	51.35	200.00
health	167.0	6.815689	2.746837	1.8100	4.920	6.32	8.60	17.90
imports	167.0	46.890215	24.209589	0.0659	30.200	43.30	58.75	174.00
income	167.0	17144.688623	19278.067698	609.0000	3355.000	9960.00	22800.00	125000.00
inflation	167.0	7.781832	10.570704	-4.2100	1.810	5.39	10.75	104.00
life_expec	167.0	70.555689	8.893172	32.1000	65.300	73.10	76.80	82.80
total_fer	167.0	2.947964	1.513848	1.1500	1.795	2.41	3.88	7.49
gdpp	167.0	12964.155689	18328.704809	231.0000	1330.000	4660.00	14050.00	105000.00

Hình 10. Thống kê mô tả dữ liệu

*Nguồn: Tác giả*

Qua hình ta thấy được tất cả các cột đều có số lượng nhất quán là 167. Một điều lưu ý nhỏ ở đây là cột lạm phát (inflation) xuất hiện giá trị tối thiểu min = -4.2 điều này có nghĩa là đất nước này đang có sự hiện diện của tình trạng giảm phát (do lạm phát có thể nhận giá trị âm nên đây không phải dữ liệu nhập sai).

Trực quan hóa dữ liệu để đưa ra nhận xét

### 3.2.2 Làm sạch dữ liệu

- Kiểm tra các giá trị null:

```
country      0
child_mort   0
exports      0
health       0
imports      0
income       0
inflation    0
life_expec   0
total_fer    0
gdpp         0
dtype: int64
```

Hình 11. Kết quả kiểm tra giá trị null

*Nguồn: Tác giả*

Bộ dữ liệu không chứa bất kỳ giá trị null nào. Vì thế nhóm không cần thực hiện xử lý các giá trị null.

- Chuyển đổi đơn vị cho các thuộc tính nhập khẩu (imports), xuất khẩu (exports), Tổng chi tiêu y tế (health) từ giá trị phần trăm (%) sang giá trị thực (USD):

Quy trình tính toán như sau:

- Nhập khẩu: Giá trị thực tế của nhập khẩu = (Tỷ lệ phần trăm nhập khẩu / 100) × GDP bình quân đầu người
- Xuất khẩu: Giá trị thực tế của xuất khẩu = (Tỷ lệ phần trăm xuất khẩu / 100) × GDP bình quân đầu người

- Chi tiêu y tế: Chi tiêu y tế thực tế = (Tỷ lệ phần trăm chi tiêu y tế / 100) × GDP bình quân đầu người

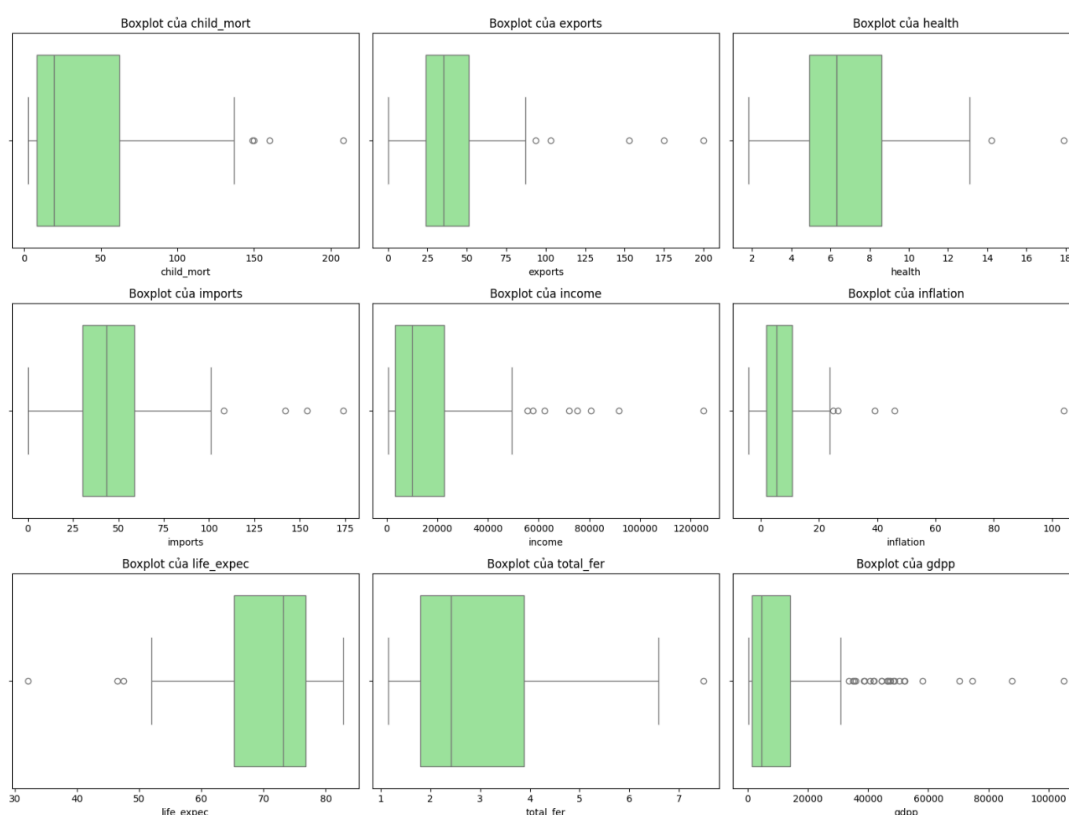
```
#Chuyển đổi dữ liệu % (health, imports, exports) thành dữ liệu thực
data['exports'] = (data['exports'] / 100) * data['gdp']
data['imports'] = (data['imports'] / 100) * data['gdp']
data['health'] = (data['health'] / 100) * data['gdp']
```

*Nguồn: Tác giả*

Cần phải thay đổi đổi từ giá trị phần trăm (%) sang giá trị thực (USD) để có thể dễ dàng đánh giá tác động thực tế, so sánh giữa các quốc gia và cung cấp đầy đủ thông tin về số liệu từ đó hỗ trợ việc phân tích và ra quyết định chính xác hơn. Ví dụ: 10% của GDP bình quân đầu người ở một quốc gia giàu (GDP cao) sẽ khác rất nhiều so với 10% của GDP ở một quốc gia nghèo (GDP thấp). Việc chỉ sử dụng tỷ lệ phần trăm có thể dẫn đến sự hiểu nhầm về mức độ ảnh hưởng kinh tế hoặc xã hội.

- Outlier:

Nhóm sử dụng quy tắc IQR để xác định các giá trị ngoại lai (Outlier). Sử dụng biểu đồ hộp để có một cái nhìn rõ hơn.



Hình 12. Trực quan hóa các giá trị ngoại lai

*Nguồn: Tác giả*

Khi kiểm tra các điểm ngoại lai trong các thuộc tính khác nhau bằng biểu đồ hộp (boxplot), nhóm quyết định không loại bỏ các điểm ngoại lai bởi vì một phần mục tiêu của bài hướng đến việc tìm ra những nước cần sự giúp đỡ của tổ chức HELP do đó có thể các điểm ngoại lai là các nước còn chậm phát triển và phần còn lại là thuật toán Optics vẫn hoạt động tốt khi có sự xuất hiện của các điểm ngoại lai.

- Chuẩn hóa dữ liệu:

Chuẩn hóa dữ liệu để đảm bảo các đặc trưng có phạm vi giá trị đồng nhất, có tác động như nhau tới việc phân tích, giúp làm giảm tác động của các biến có phạm vi giá trị vượt trội hơn các biến khác. Nhóm đã tiến hành chuẩn hóa dữ liệu bằng phương pháp Standardization, tức là đưa dữ liệu về một phân bố có giá trị trung bình ( $\bar{x}$ ) bằng 0 và độ lệch chuẩn ( $\sigma$ ) bằng 1. Công thức:

$$z = \frac{x - \bar{x}}{\sigma}$$

```
scaled_data = StandardScaler().fit_transform(data)
```

### 3.2.3 Giảm chiều dữ liệu bằng phương pháp PCA

PCA (principal component analysis) là phương pháp tìm một hệ cơ sở mới sao cho phần lớn thông tin của dữ liệu sẽ tập trung vào một vài tọa độ, để đơn giản trong tính toán thì PCA sẽ tìm một hệ trục chuẩn mới để làm cơ sở. PCA giúp giảm số lượng biến (features) mà vẫn giữ được phần lớn các thông tin ban đầu.

Các bước thực hiện PCA:

- Tính vector kỳ vọng của bộ dữ liệu:

$$\bar{x} = \frac{1}{N} \sum_{n=1}^N x_n$$

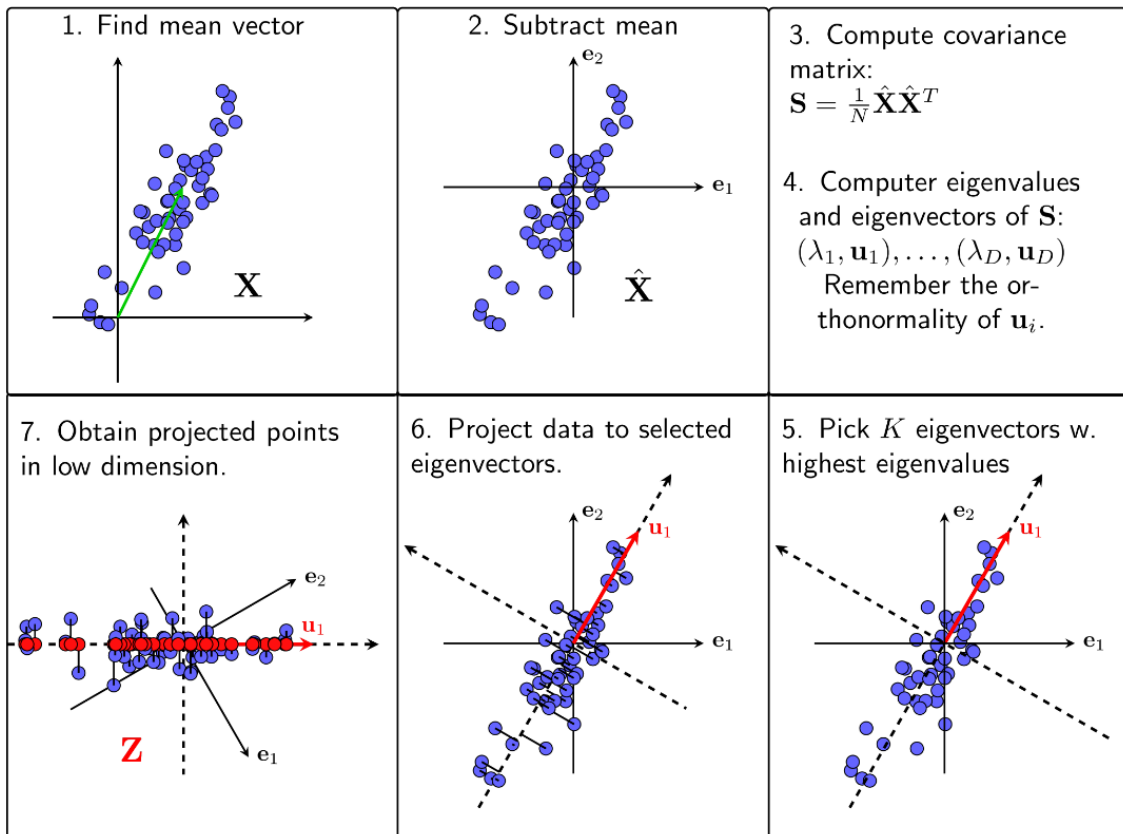
- Tính  $\hat{x}_n = x_n - \bar{x}$
- Tính ma trận hiệp phương sai (Covariance Matrix):

$$S = \frac{1}{N} \hat{X} \hat{X}^T$$

- Tính các trị riêng, vector riêng có chuẩn bằng 1 và sắp xếp theo thứ tự giảm dần của trị riêng
- Chọn K vector riêng ứng với K trị riêng lớn nhất

- Thực hiện phép chiếu dữ liệu ban đầu được chuẩn hóa  $\hat{X}$  xuống không gian con vừa tìm được.
- Dữ liệu mới là tọa độ các điểm trên không gian mới:

$$Z = U_K^T \hat{X}$$



Hình 13. Các bước thực hiện PCA

Nguồn: *Machine Learning Cơ Bản. (2017, June 15)*

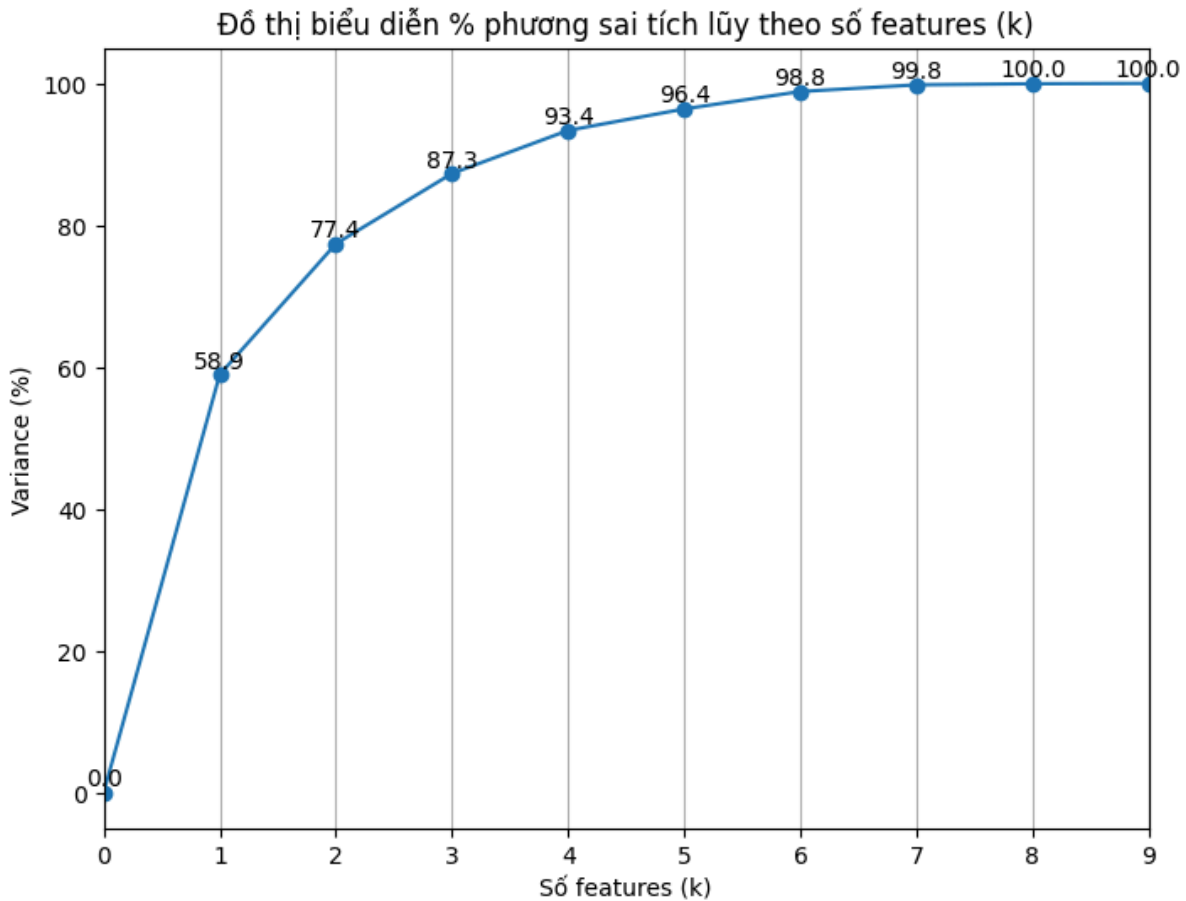
### 3.2.4 Phương pháp ELBOW

Phương pháp ELBOW là phương pháp phổ biến để xác định số lượng k cụm trong tập dữ liệu đối với các thuật toán phân cụm với các thuật toán phân cụm xác định trước tham số này. Ngoài ra, phương pháp này còn được dùng để xác định số k thành phần chính cần thiết để giải thích một tỉ lệ phương sai nhất định trong dữ liệu của bài toán PCA.

Sau khi tiến hành phân tích thành phần chính (PCA) để thu được các thành phần chính (principal components), các giá trị phương sai được tính toán từ các giá trị

riêng (eigenvalues) của ma trận hiệp phương sai. Phương sai tích lũy của thành phần chính thứ k được tính bằng công thức sau:

$$Cumulative\ Variance = \sum_{i=1}^k \lambda_i$$



Hình 14. Kết quả xác định thành phần chính thứ k

*Nguồn: Tác giả*

Để lựa chọn số k, ta vẽ đồ thị phương sai tích lũy và xác định điểm ELBOW, tức điểm mà đồ thị có hình dạng như một khuỷu tay tức là từ điểm này trở đi thì sự đóng góp của các thành phần chính trở nên không đáng kể nữa.

Dựa vào đồ thị trên thì có thể chọn  $k = 3$ , vì tại điểm này, giá trị phương sai tích lũy nằm trong khoảng tin cậy từ 85% đến 95%. Cho thấy rằng đây là một lựa chọn hợp lý để cân bằng giữa việc giảm số lượng thành phần và đảm bảo mức độ giải thích dữ liệu đủ cao mà không làm tăng thêm sự phức tạp không cần thiết.

### 3.3. Áp dụng thuật toán OPTICS:

#### 3.3.1 Mô hình khởi tạo

##### Parameters:

- *min\_samples* : int > 1 hoặc float nằm giữa 0 và 1, mặc định là 5

Định nghĩa số điểm lân cận của một core point. Nếu nó là số nguyên thì nó sẽ xác định số lượng điểm lân cận tối thiểu (bao gồm chính nó) để trở thành core point. Nếu là số thực thì nó tính theo tỷ lệ phần trăm của tổng số điểm trong dữ liệu, sau đó làm tròn lên ít nhất 2.

- *max\_eps* : float, mặc định là *np.inf*

Là khoảng cách tối đa giữa hai điểm để chúng được xem là lân cận của nhau. Giảm *max\_eps* sẽ làm thời gian chạy ngắn hơn.

- *metric*: str hoặc callable, mặc định là '*minkowski*'

Đây là phép đo khoảng cách được dùng để tính khoảng cách giữa các điểm.

Một số các phép đo như '*euclidean*', '*manhattan*', '*cosine*', '*chebyshev*',.....

- *cluster\_method* : str, mặc định là '*xi*'

Phương pháp dùng để trích xuất các cụm từ thứ tự và khoảng cách tiếp cận (reachability).

- '*xi*' : sử dụng phương pháp xi để xác định ranh giới cụm dựa trên độ dốc trong reachability plot
- '*dbscan*' : dùng phương pháp DBSCAN để trích xuất cụm từ khoảng cách tiếp cận.

- *xi* : float trong khoảng (0,1), mặc định là 0.05

Tham số này xác định độ dốc tối thiểu trong reachability plot để xác định ranh giới các cụm. Một điểm được xem là điểm phân tách nếu độ dốc giữa điểm đó và điểm tiếp theo nhỏ hơn 1-xi.

- *min\_cluster\_size* : int hoặc float, mặc định là None

Đây là số lượng điểm tối thiểu trong một cụm để xem nó là một cụm hợp lệ. Nếu là None thì nó bằng với *min\_samples*.

##### Attributes (các thuộc tính):



- *labels\_* : nhãn cụm cho mỗi điểm trong bộ dữ liệu, những điểm không thuộc cụm nào sẽ có nhãn là -1.
- *reachability\_* : Khoảng cách tiếp cận cho mỗi điểm trong bộ dữ liệu.
- *ordering\_* : Danh sách chỉ mục của các điểm theo thứ tự trong quá trình xử lý.
- *core\_distances\_* : Khoảng cách mà mỗi điểm trở thành core point, được lập chỉ mục theo thứ tự các đối tượng, các điểm không bao giờ có khoảng cách là inf.

### 3.3.2 Tìm kiếm các hyperparameters

Sau khi tiến hành PCA để giảm chiều dữ liệu, nhóm tiến hành tìm kiếm các siêu tham số cho mô hình để đạt được kết quả phân cụm được đánh giá bằng chỉ số Silhouette tốt nhất. Việc thử các giá trị là để tìm ra sự kết hợp tốt nhất cho cấu trúc dữ liệu, từ đó tối ưu kết quả phân nhóm.

Với bộ dữ liệu gồm 200 dòng, nhóm thử *min\_samples* với các giá trị {3,5,7}, giúp chỉnh độ "chặt chẽ" của các cụm. Giá trị nhỏ cho phép phát hiện nhiều cụm nhỏ, trong khi giá trị lớn giúp loại bỏ nhiễu và chỉ nhận diện các cụm lớn hơn.

Tiếp theo là *xi* với các giá trị {0.05, 0.1, 0.2} để kiểm soát sự phân tách giữa các cụm. Giá trị nhỏ giúp phân biệt rõ ràng các cụm, giá trị lớn có thể hợp nhất các cụm gần nhau.

Cuối cùng là *min\_cluster\_size* với các giá trị {0.3, 0.5, 0.7} đảm bảo mỗi cụm có kích thước tối thiểu. Giá trị nhỏ cho phép phát hiện các cụm nhỏ, trong khi giá trị lớn chỉ nhận diện các cụm có đủ điểm.

```
# Thử nghiệm với các giá trị min_samples khác nhau
for min_samples in [3, 5, 10]:
    # Thử nghiệm với các giá trị xi khác nhau
    for xi in [0.05, 0.1, 0.2]:
        # Thử nghiệm với các giá trị min_cluster_size khác nhau
        for min_cluster_size in [0.3, 0.5, 0.7]:
            # Thực hiện thuật toán OPTICS
            OP = OPTICS(min_samples=min_samples, xi=xi,
min_cluster_size=min_cluster_size, metric='euclidean')
            OP_df = OP.fit_predict(scores)

            # Check for the number of unique labels
            n_clusters = len(np.unique(OP_df))
```

```

# Calculate silhouette score only if more than 1 cluster
is found

if n_clusters > 1:
    silhouette_avg = silhouette_score(scores, OP_df)
    if silhouette_avg > 0.5:
        print(f"min_samples: {min_samples}, xi: {xi},
min_cluster_size: {min_cluster_size}, Silhouette Score:
{silhouette_avg}")

```

Và kết quả đạt được như sau:

```

min_samples: 3, xi: 0.05, min_cluster_size: 0.3, Silhouette Score:
0.09136405494708613
min_samples: 3, xi: 0.05, min_cluster_size: 0.5, Silhouette Score:
0.2323771597126316
min_samples: 3, xi: 0.05, min_cluster_size: 0.7, Silhouette Score:
0.44598822308197283
min_samples: 3, xi: 0.1, min_cluster_size: 0.3, Silhouette Score:
0.44598822308197283
min_samples: 3, xi: 0.1, min_cluster_size: 0.5, Silhouette Score:
0.44598822308197283
min_samples: 3, xi: 0.1, min_cluster_size: 0.7, Silhouette Score:
0.44598822308197283
min_samples: 3, xi: 0.2, min_cluster_size: 0.3, Silhouette Score:
0.5722623738943571
min_samples: 3, xi: 0.2, min_cluster_size: 0.5, Silhouette Score:
0.5722623738943571
min_samples: 3, xi: 0.2, min_cluster_size: 0.7, Silhouette Score:
0.5722623738943571
min_samples: 5, xi: 0.05, min_cluster_size: 0.3, Silhouette Score:
0.13541228468298433
min_samples: 5, xi: 0.05, min_cluster_size: 0.5, Silhouette Score:
0.32790041723013524
min_samples: 5, xi: 0.05, min_cluster_size: 0.7, Silhouette Score:
0.32790041723013524
min_samples: 5, xi: 0.1, min_cluster_size: 0.3, Silhouette Score:
0.5722623738943571
min_samples: 5, xi: 0.1, min_cluster_size: 0.5, Silhouette Score:
0.5722623738943571
min_samples: 5, xi: 0.1, min_cluster_size: 0.7, Silhouette Score:
0.5722623738943571
min_samples: 5, xi: 0.2, min_cluster_size: 0.3, Silhouette Score:
0.5722623738943571
min_samples: 5, xi: 0.2, min_cluster_size: 0.5, Silhouette Score:
0.5722623738943571
min_samples: 5, xi: 0.2, min_cluster_size: 0.7, Silhouette Score:
0.5722623738943571
min_samples: 10, xi: 0.05, min_cluster_size: 0.3, Silhouette Score:
0.14275926371275827
min_samples: 10, xi: 0.05, min_cluster_size: 0.5, Silhouette Score:
0.38206770658694444
min_samples: 10, xi: 0.05, min_cluster_size: 0.7, Silhouette Score:
0.38206770658694444

```

```
min_samples: 10, xi: 0.1, min_cluster_size: 0.3, Silhouette Score: 0.38206770658694444
min_samples: 10, xi: 0.1, min_cluster_size: 0.5, Silhouette Score: 0.38206770658694444
min_samples: 10, xi: 0.1, min_cluster_size: 0.7, Silhouette Score: 0.38206770658694444
```

Từ kết quả trên, ta có thể chọn *min\_samples: 3, xi: 0.2, min\_cluster\_size: 0.3* để triển khai mô hình OPTICS.

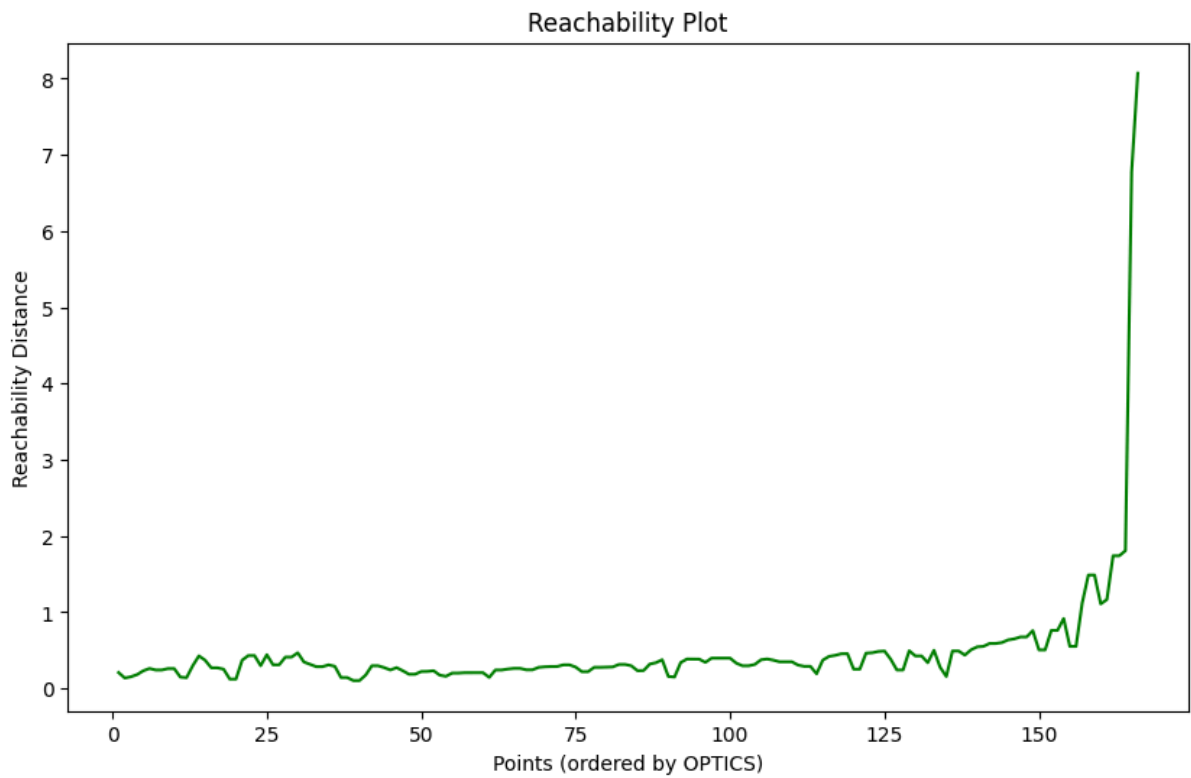
### 3.3.3 Tạo mô hình từ các siêu tham số vừa tìm được và tiến hành phân cụm

Nhóm tiến hành phân cụm dữ liệu dựa trên các tham số vừa tìm được *min\_samples: 3, xi: 0.2, min\_cluster\_size: 0.3, metric = 'euclidean'*.

```
#Áp dụng thuật toán OPTICS
optics = OPTICS(min_samples=3, xi=0.2, min_cluster_size=0.3, metric = 'euclidean')
OP_df = optics.fit_predict(scores)
# Thêm nhãn cluster vào DataFrame
data['Cluster'] = OP_df
```

Sau khi phân cụm, trực quan reachability plot để có thể quan sát kết quả phân cụm cũng như là khả năng tiếp cận của các điểm dữ liệu.

```
# Vẽ Reachability Plot
reachability = optics.reachability_[optics.ordering_]
plt.figure(figsize=(10, 6))
plt.plot(reachability, 'g-', linewidth=1.5)
plt.xlabel('Points (ordered by OPTICS)')
plt.ylabel('Reachability Distance')
plt.title('Reachability Plot')
plt.show();
```

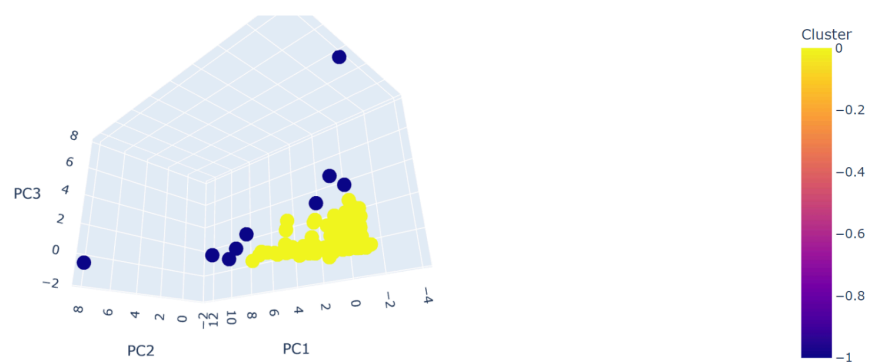


Hình 15. Biểu đồ Reachability

*Nguồn: Tác giả*

Ngoài ra, nhóm cũng thể hiện kết quả phân cụm trong không gian 3D để có thể dễ dàng quan sát kết quả phân cụm.

Phân cụm OPTICS trong không gian PCA



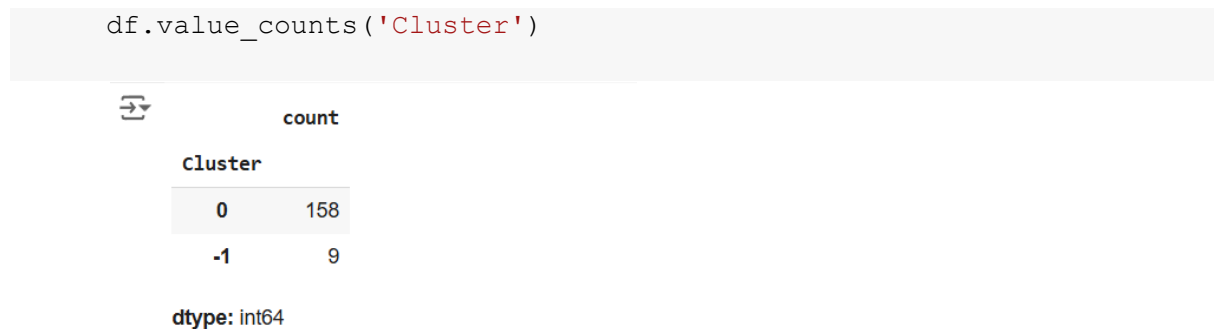
Hình 16. Kết quả phân cụm OPTICS trong không gian PCA

*Nguồn: Tác giả*

Có thể thấy thuật toán phân cụm khá tốt, mật độ các điểm trong 1 cụm dày và chỉ số Silhouette đo được với các điểm được phân cụm là 0.57 cho thấy các điểm trong một cụm có sự tương đồng với nhau cao.

Ngoài ra, thuật toán cũng xác định được 9 điểm không nằm trong một cụm nào cả, nằm xa các vùng có mật độ các điểm dày.

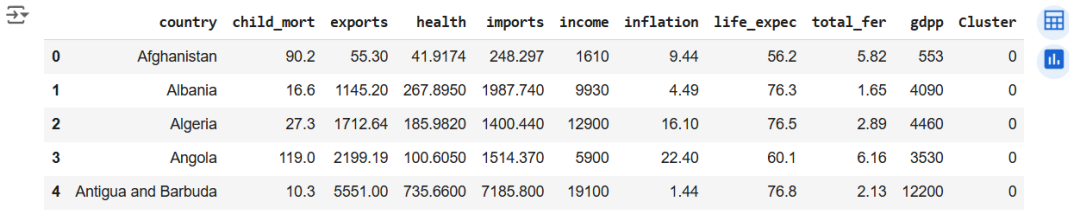
Tiến hành đếm số lượng đối tượng của từng nhãn, thu được kết quả sau:



Hình 17. Kết quả đếm số lượng đối tượng từng nhãn

Có tất cả 158 quốc gia được gán nhãn 0 và 9 quốc gia được gán nhãn -1 ( tức là 9 quốc gia này không thuộc cụm cả).

```
[193] df.head()
```



	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	Cluster
0	Afghanistan	90.2	55.30	41.9174	248.297	1610	9.44	56.2	5.82	553	0
1	Albania	16.6	1145.20	267.8950	1987.740	9930	4.49	76.3	1.65	4090	0
2	Algeria	27.3	1712.64	185.9820	1400.440	12900	16.10	76.5	2.89	4460	0
3	Angola	119.0	2199.19	100.6050	1514.370	5900	22.40	60.1	6.16	3530	0
4	Antigua and Barbuda	10.3	5551.00	735.6600	7185.800	19100	1.44	76.8	2.13	12200	0

Hình 18. Bảng dữ liệu thêm cột gán nhãn 'Cluster'

In thông tin các quốc gia thuộc từng nhãn:

- Nhãn -1:

```
df_1 = df[df['Cluster'] == -1]
df_1
```

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdp	Cluster
49	Equatorial Guinea	111.0	14671.80	766.080	10071.90	33700	24.900	60.9	5.21	17100	-1
91	Luxembourg	2.8	183750.00	8158.500	149100.00	91700	3.620	81.3	1.63	105000	-1
103	Mongolia	26.1	1237.55	144.160	1502.55	7710	39.200	66.2	2.64	2650	-1
113	Nigeria	130.0	589.49	118.131	405.42	5150	104.000	60.5	5.84	2330	-1
114	Norway	3.2	34856.60	8323.440	25023.00	62300	5.950	81.0	1.95	87800	-1
123	Qatar	9.0	43796.90	1272.430	16731.40	125000	6.980	79.5	2.07	70300	-1
133	Singapore	2.8	93200.00	1845.360	81084.00	72100	-0.046	82.7	1.15	46600	-1
145	Switzerland	4.5	47744.00	8579.000	39761.80	55500	0.317	82.2	1.52	74600	-1
163	Venezuela	17.1	3847.50	662.850	2376.00	16500	45.900	75.4	2.47	13500	-1

Hình 19. Dữ liệu của các quốc gia thuộc nhãn -1

	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdp	Cluster
count	9.000000	9.000000	9.000000	9.000000	9.000000	9.000000	9.000000	9.000000	9.000000	9.0
mean	34.055556	47077.093333	3318.883444	36228.452222	52184.444444	25.646778	74.411111	2.720000	46653.333333	-1.0
std	49.854491	59317.590771	3814.291353	49484.239769	40587.560321	34.016494	9.291185	1.662521	39228.433247	0.0
min	2.800000	589.490000	118.131000	405.420000	5150.000000	-0.046000	60.500000	1.150000	2330.000000	-1.0
25%	3.200000	3847.500000	662.850000	2376.000000	16500.000000	3.620000	66.200000	1.630000	13500.000000	-1.0
50%	9.000000	34856.600000	1272.430000	16731.400000	55500.000000	6.980000	79.500000	2.070000	46600.000000	-1.0
75%	26.100000	47744.000000	8158.500000	39761.800000	72100.000000	39.200000	81.300000	2.640000	74600.000000	-1.0
max	130.000000	183750.000000	8579.000000	149100.000000	125000.000000	104.000000	82.700000	5.840000	105000.000000	-1.0

Hình 20. Mô tả tuyến tính dữ liệu của các quốc gia thuộc nhãn -1

- Nhãn 0:



```
df_0 = df[df['Cluster'] == 0]
```

```
df_0
```

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdp	Cluster
0	Afghanistan	90.2	55.30	41.9174	248.297	1610	9.44	56.2	5.82	553	0
1	Albania	16.6	1145.20	267.8950	1987.740	9930	4.49	76.3	1.65	4090	0
2	Algeria	27.3	1712.64	185.9820	1400.440	12900	16.10	76.5	2.89	4460	0
3	Angola	119.0	2199.19	100.6050	1514.370	5900	22.40	60.1	6.16	3530	0
4	Antigua and Barbuda	10.3	5551.00	735.6600	7185.800	19100	1.44	76.8	2.13	12200	0
...	...	...	...	...	...	...	...	...	...	...	...
161	Uzbekistan	36.3	437.46	80.1780	393.300	4240	16.50	68.8	2.34	1380	0
162	Vanuatu	29.2	1384.02	155.9250	1565.190	2950	2.62	63.0	3.50	2970	0
164	Vietnam	23.3	943.20	89.6040	1050.620	4490	12.10	73.1	1.95	1310	0
165	Yemen	56.3	393.00	67.8580	450.640	4480	23.60	67.5	4.67	1310	0
166	Zambia	83.1	540.20	85.9940	451.140	3280	14.00	52.0	5.40	1460	0

158 rows × 11 columns

Hình 21. Dữ liệu của các quốc gia thuộc nhãn 0

	df_0.describe()									
		child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp Cluster
count	158.000000	158.000000	158.000000	158.000000	158.000000	158.000000	158.000000	158.000000	158.000000	158.0
mean	38.510127	5161.705744	927.876545	4899.991975	15148.753165	6.764209	70.336076	2.960949	11045.151899	0.0
std	39.899032	8185.544828	1542.626579	7128.033110	15317.154904	6.312360	8.850019	1.509675	14422.828412	0.0
min	2.600000	1.076920	12.821200	0.651092	609.000000	-4.210000	32.100000	1.230000	231.000000	0.0
25%	8.625000	428.284500	67.989000	571.837500	3290.000000	1.790000	65.300000	1.810000	1310.000000	0.0
50%	19.750000	1733.350000	297.416000	1981.770000	9925.000000	5.025000	72.950000	2.410000	4550.000000	0.0
75%	62.150000	6051.475000	890.445000	6356.390000	21100.000000	10.075000	76.675000	3.895000	12500.000000	0.0
max	208.000000	50161.000000	8663.600000	42125.500000	80600.000000	26.500000	82.800000	7.490000	58000.000000	0.0

Hình 22. Mô tả tuyến tính dữ liệu của các quốc gia thuộc nhãn 0

## CHƯƠNG 4. ĐÁNH GIÁ THUẬT TOÁN

### 4.1. Độ đo Silhouette Score

Để đánh giá xem kết quả chất lượng kết quả phân cụm, nhóm dùng chỉ số Silhouette để đo sự tương đồng giữa các điểm trong cụm vừa tìm được.

```
# Calculate and print the silhouette score
silhouette_avg = silhouette_score(scores, OP_df)
print(f"Silhouette Score: {silhouette_avg}")

# Compute the silhouette scores for each sample
sample_silhouette_values = silhouette_samples(scores, OP_df)

# Create a silhouette plot
import matplotlib.cm as cm
fig, ax = plt.subplots(figsize=(10, 6))

y_lower = 10

for i in range(len(np.unique(OP_df))):
    ith_cluster_silhouette_values = sample_silhouette_values[OP_df == i]
    ith_cluster_silhouette_values.sort()

    size_cluster_i = ith_cluster_silhouette_values.shape[0]
    y_upper = y_lower + size_cluster_i

    color = cm.nipy_spectral(float(i) / len(np.unique(OP_df)))
    ax.fill_betweenx(np.arange(y_lower, y_upper), 0,
ith_cluster_silhouette_values,
                    facecolor=color, edgecolor=color, alpha=0.7)

    ax.text(-0.05, y_lower + 0.5 * size_cluster_i, str(i))
    y_lower = y_upper + 10

ax.set_title("Silhouette Plot")
ax.set_xlabel("Silhouette Coefficient Values")
ax.set_ylabel("Cluster Label")

ax.axvline(x=silhouette_avg, color="red", linestyle="--")

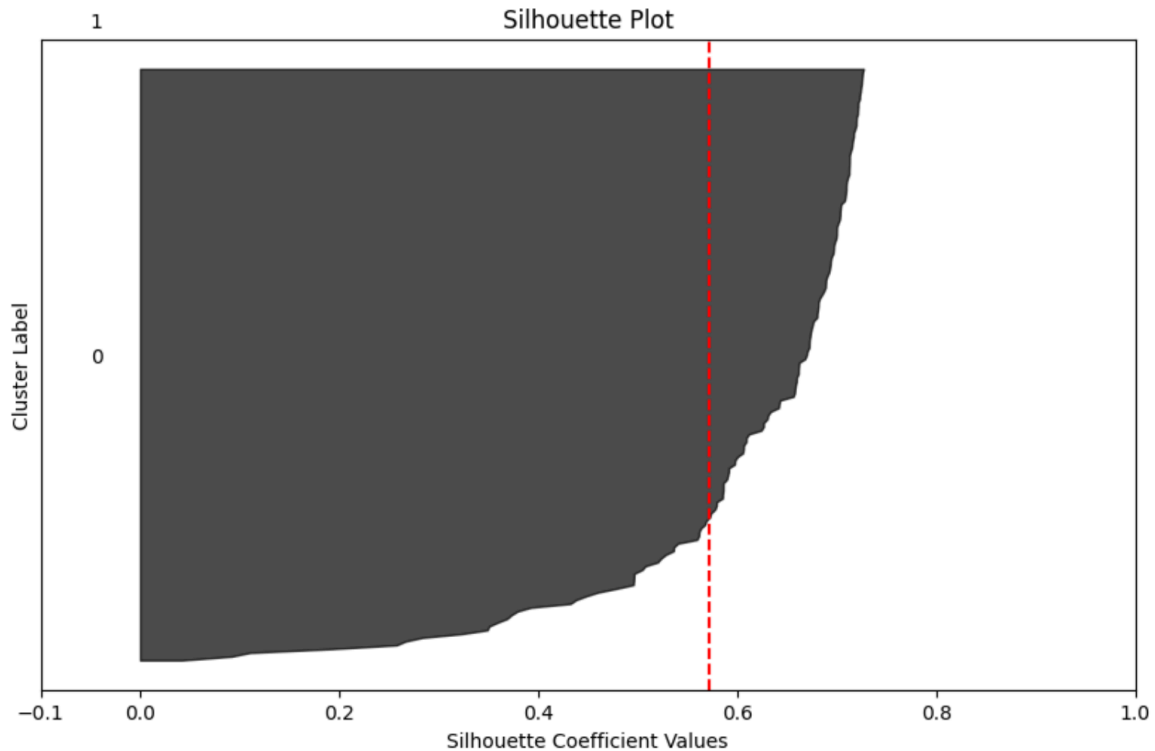
ax.set_yticks([]) # Clear the yaxis labels / ticks
```



```
ax.set_xticks([-0.1, 0, 0.2, 0.4, 0.6, 0.8, 1])

plt.show()
```

Silhouette Score: 0.5722623738943571



Hình 23. Silhouette Score của thuật toán OPTICS

*Nguồn: Tác giả*

Chỉ số Silhouette đạt được là 0.57, cho thấy chất lượng phân cụm khá tốt, các điểm trong cùng một cụm có sự tương đồng với nhau khá tốt.

## 4.2. So sánh các phương pháp phân cụm khác

Nhóm đánh giá thuật toán bằng phương pháp đánh giá tương đối qua việc so sánh với thuật toán là K-Means và DBSCAN.

- **K-Means**

Nhóm tiến hành phân cụm bộ dữ liệu với thuật toán K-Means từ thư viện sklearn với  $k = 3$ .

```
from sklearn.cluster import KMeans
X_numerics = scores_df[['PC1', 'PC2', 'PC3']]
KM_3_clusters = KMeans(n_clusters=3,
```

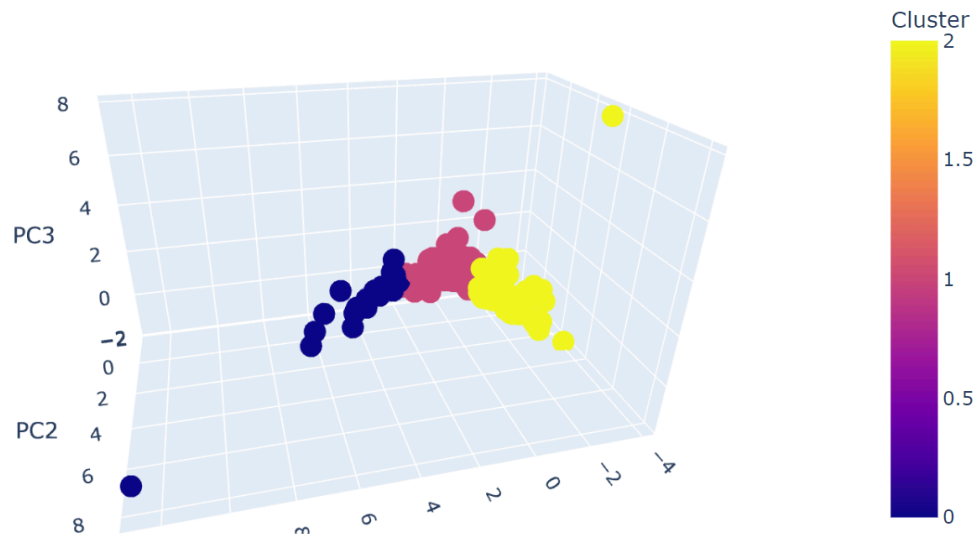
```

init='k-means++').fit(X_numerics) # khởi tạo và điều chỉnh mô hình
K-Means
KM3_clustered = X_numerics.copy()
KM3_clustered.loc[:, 'Cluster'] = KM_3_clusters.labels_ # gắn nhãn vào
điểm

```

Kết quả phân cụm được trực quan hoá bằng biểu đồ 3D. Ta có thể thấy với hai thuật toán thuộc hai phương pháp phân cụm khác nhau có thể cho ra kết quả số cụm khác nhau như trong bộ dữ liệu mà nhóm sử dụng thì K-Means phân thành 3 cụm trong khi đó OPTICS và DBSCAN chỉ cho kết quả 1 cụm.

Phân cụm với K-means trong không gian 3D

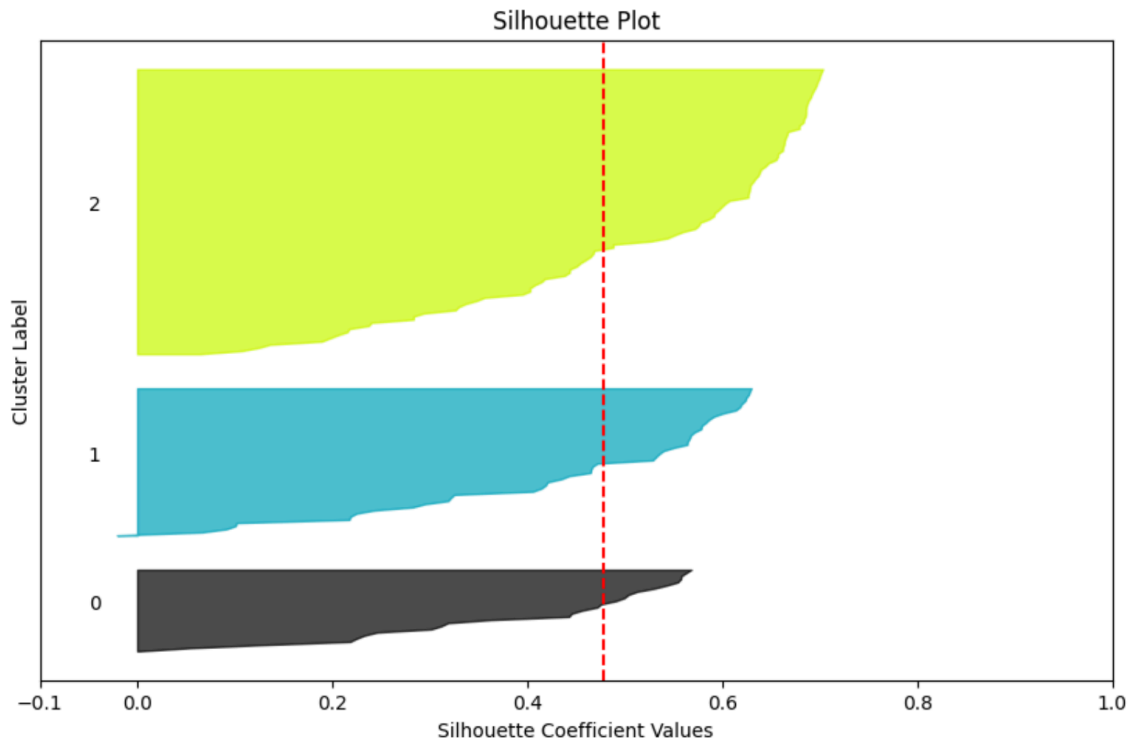


Hình 24. Kết quả phân cụm K-Means trong không gian 3D

*Nguồn: Tác giả*

Đánh giá kết quả phân cụm bằng chỉ số Silhouette

Silhouette Score: 0.4786795843501643



Hình 25. Silhouette Score của thuật toán K-Means

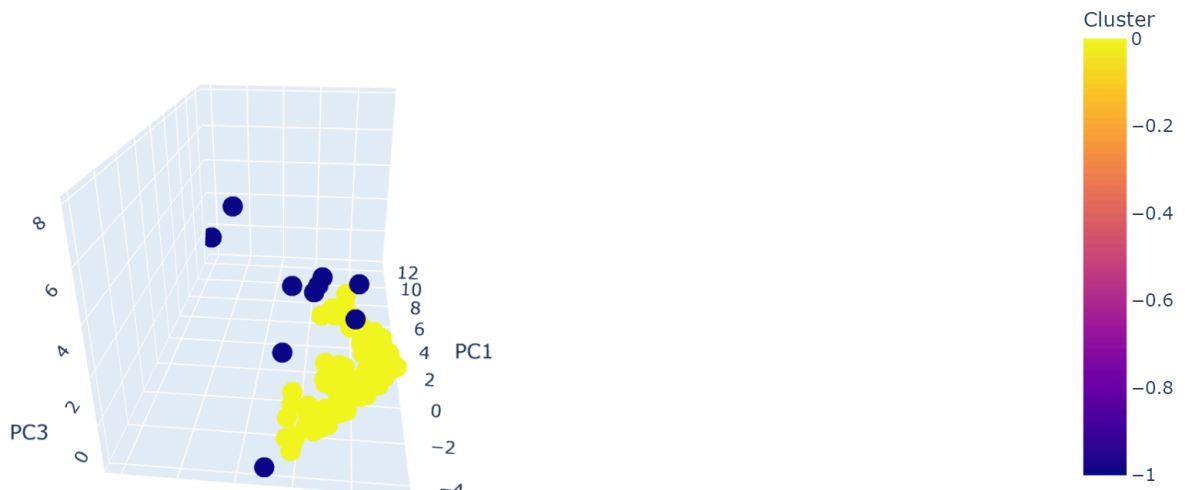
*Nguồn: Tác giả*

- **DBSCAN**

Tương tự, nhóm cũng chạy bộ dữ liệu với thuật toán DBSCAN từ thư viện sklearn với tham số  $\text{eps} = 1$  và  $\text{min\_samples} = 4$ .

```
DBS_clustering = DBSCAN(eps=1, min_samples=4).fit(X_numerics)
DBSCAN_clustered = X_numerics.copy()
DBSCAN_clustered.loc[:, 'Cluster'] = DBS_clustering.labels_ # gán
(label) nhãn vào điểm
```

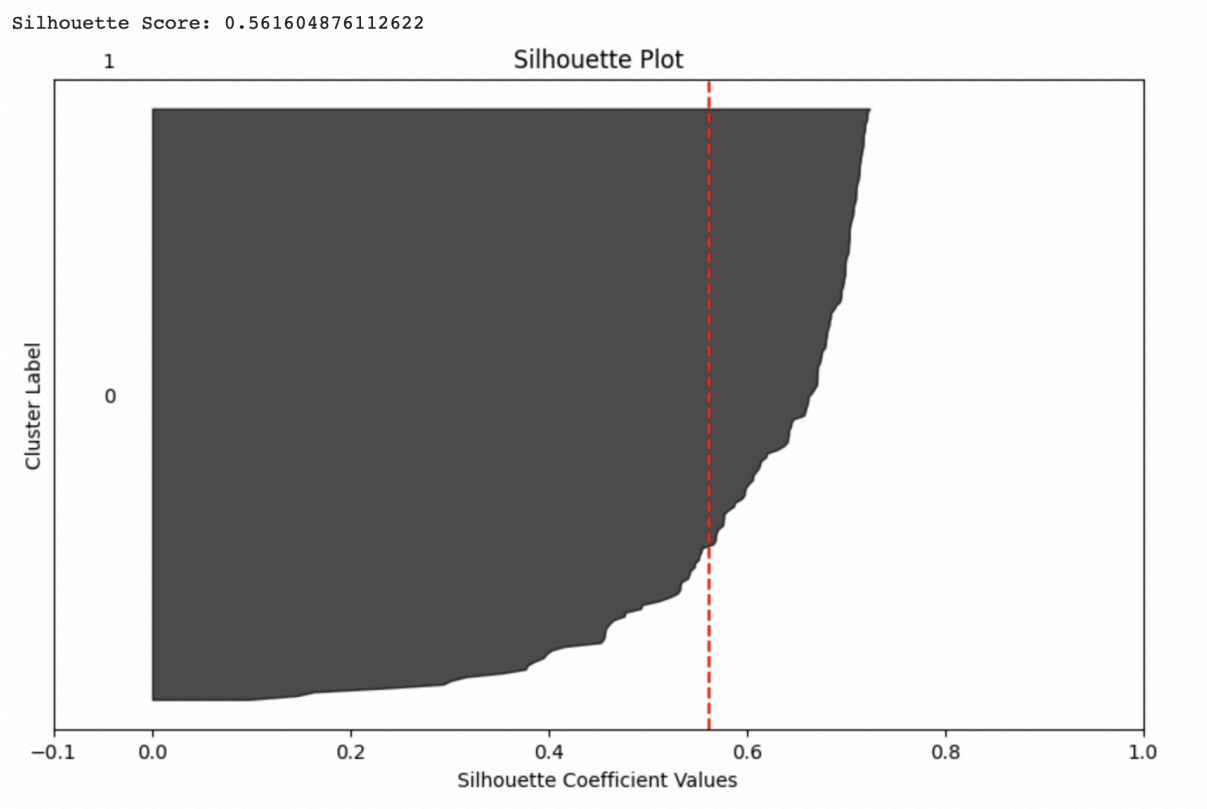
Kết quả phân cụm được biểu diễn trong không gian 3D. Từ kết quả cho thấy với bộ dữ liệu mà nhóm sử dụng thì thuật toán OPTICS và DBSCAN cho kết quả phân cụm có sự tương tự nhau.



Hình 26. Kết quả phân cụm của thuật toán DBSCAN trong không gian 3D

*Nguồn: Tác giả*

Đánh giá kết quả phân cụm bằng chỉ số Silhouette



Hình 27. Silhouette Score của thuật toán DBSCAN

*Nguồn: Tác giả*

## CHƯƠNG 5. THẢO LUẬN VÀ KẾT LUẬN

### 5.1. Thảo luận

Dựa trên kết quả phân cụm của thuật toán OPTICS, nhóm nghiên cứu có thể đưa ra một số ý kiến sau:

- Kết quả phân cụm cho thấy cụm lớn có 158 quốc gia và 9 quốc gia thuộc giá trị ngoại lệ. Điều này phản ánh sự phân hóa đáng kể giữa các quốc gia dựa trên các chỉ số kinh tế - xã hội. Tuy nhiên, từ kết quả phân tích về cụm lớn, có thể chỉ ra rằng mặc dù tất cả các quốc gia đều nằm trong cùng một cụm, nhưng chúng vẫn có những vấn đề riêng biệt cần được giải quyết.
- Nhóm quốc gia thuộc phân cụm lớn có xu hướng tương đồng về độ phát triển kinh tế - xã hội với GDP và thu nhập ở mức trung bình thấp, y tế còn khá hạn chế, tuổi thọ ngắn, tỷ lệ trẻ em tử vong cao. Trong khi đó 9 quốc gia ngoại lệ có xu hướng kinh tế phát triển cao vượt trội, với các chỉ số GDP, thu nhập, chi tiêu y tế và tuổi thọ cao, trong khi tỷ lệ tử vong trẻ em rất thấp.
- Dựa trên kết quả phân tích này, tổ chức HELP International nên xem xét hỗ trợ các quốc gia có tỷ lệ tử vong trẻ em cao và chỉ tiêu cho sức khỏe thấp. Đồng thời, nên nghiên cứu thêm về các quốc gia không được phân cụm để hiểu rõ hơn về tình hình của những quốc gia này và quyết định xem có nên đầu tư vào đó hay không.

### 5.2. Kết luận

Việc sử dụng thuật toán OPTICS đã giúp xác định một cách rõ ràng cấu trúc của dữ liệu xã hội và kinh tế trong các quốc gia. Chỉ số Silhouette đạt 0.57 cho thấy kết quả phân cụm có chất lượng khá, phản ánh sự tương đồng cao trong nội bộ cụm.

Đồng thời, so với K-Means và DBSCAN, OPTICS xử lý phân cụm tốt hơn trong việc nhận diện cấu trúc cụm, đặc biệt đối với dữ liệu có mật độ thay đổi và tồn tại điểm nhiễu mà không cần phải tiền xử lý outlier(s) trước đó.

Tuy nhiên, để hiểu rõ hơn về đặc điểm của cụm cũng như lý giải được nguyên nhân các điểm ngoại lai không thể phân cụm được, cần tiến hành nghiên cứu sâu hơn, có thể sử dụng kết hợp thêm một số phương pháp nghiên cứu khác hoặc bổ sung thêm các chỉ số kinh tế - xã hội khác để bộ dữ liệu đầy đủ và phản ánh được tính toàn cảnh hơn.

### 5.3. Ứng dụng thực tế và triển vọng

- Ứng dụng thực tế:
  - Kết quả phân tích có thể được dùng để hỗ trợ các tổ chức quốc tế như HELP International đề ra các quyết định ưu tiên viện trợ hiệu quả hơn, tập trung vào các quốc gia có mức độ phát triển thấp hoặc đặc biệt cần hỗ trợ.
  - Đồng thời kết quả cũng hỗ trợ việc đề ra các chính sách định hướng phát triển và phân bổ các nguồn lực kinh tế - xã hội phù hợp với đặc điểm từng cụm.
- Triển vọng phát triển:
  - Trong tương lai, có thể kết hợp thuật toán OPTICS cùng các kỹ thuật phân tích khác để cải thiện độ chính xác và hiệu quả phân cụm.
  - Ngoài ra có thể xem xét thêm các yếu tố thời gian hoặc xu hướng kinh tế thay đổi vào phân tích để cung cấp thêm giá trị trong việc dự đoán và lập kế hoạch phát triển dài hạn.

## TÀI LIỆU THAM KHẢO

1. Nguyễn, A. T. (2023). *Học không giám sát (Unsupervised Learning)*. Tải từ: UEH LMS.
2. Gao, S., Li, M., Rao, J., Mai, G., Prestby, T., Marks, J., & Hu, Y. (2023). *Automatic urban road network extraction from massive GPS trajectories of taxis*.
3. Han, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.
4. Redinger, G., & Hunner, M. (2017). *Visualization of Clustering Algorithms: VIS2017 - Project Proposal*. Tải từ: [https://gregorredinger.github.io/vis\\_clustering\\_algorithms/M2/index.html](https://gregorredinger.github.io/vis_clustering_algorithms/M2/index.html).
5. Fournier-Viger, P. (n.d.). *Using OPTICS to Extract a Cluster-Ordering of Points and DBSCAN-Style Clusters*. Tải từ: <https://www.philippe-fournier-viger.com/spmf/Optics.php>.
6. Guowei Wang, Yu Chen, Jian Li, Yunpeng Hao. The Application of the OPTICS Algorithm in the Maize Precise Fertilization Decision-Making. 9th International Conference on Computer and Computing Technologies in Agriculture (CCTA), Sep 2015, Beijing, China. pp.317-324, ff10.1007/978-3-319-48357-3\_31ff. fhal-01557832f
7. Yu, Z., Jin, Y., Parmar, M., & Wang, L. (n.d.). *Application of Modified OPTICS Algorithm in E-Commerce Sites Classification and Evaluation*.
8. Tran, H.; Vu-Van, T.; Bang, T.; Le, T.-V.; Pham, H.-A.; Huynh-Tuong, N. *Data Mining of Formative and Summative Assessments for Improving Teaching Materials towards Adaptive Learning: A Case Study of Programming Courses at the University Level*. Electronics 2023, 12, 3135. <https://doi.org/10.3390/electronics12143135>
9. Plakalovic, A. (2023). *Clustering Algorithms: DBSCAN vs. OPTICS*. Atlantbh. Tải từ: <https://www.atlantbh.com/clustering-algorithms-dbscan-vs-optics/>
10. Wang, Q.; Zhang, Y.; Yin, S.; Wang, Y. *A Novel Underdetermined Blind Source Separation Method Based on OPTICS and Subspace Projection*. Symmetry 2021, 13, 1677. <https://doi.org/10.3390/sym13091677>
11. Market Brew. (n.d.). *The Role of Cosine Similarity in Vector Space and its Relevance in SEO*. Tải từ: <https://marketbrew.ai/a/cosine-similarity>

12. Scikit-learn. (n.d.). *Comparing different clustering algorithms on toy datasets*. Tải từ:  
[https://scikit-learn.org/dev/auto\\_examples/cluster/plot\\_cluster\\_comparison.html#sphx-glr-auto-examples-cluster-plot-cluster-comparison-py](https://scikit-learn.org/dev/auto_examples/cluster/plot_cluster_comparison.html#sphx-glr-auto-examples-cluster-plot-cluster-comparison-py)
13. Medium. (2021). *Understanding Vector Similarity*. Tải từ:  
<https://medium.com/advanced-deep-learning/understanding-vector-similarity-b9c10f7506de>
14. Big Data Uni. (n.d.). *Các phương pháp đánh giá trong thuật toán clustering*. Tải từ:  
<https://bigdatauni.com/tin-tuc/cac-phuong-phap-danh-gia-trong-thuat-toan-clustering.html>
15. Machine Learning Cơ Bản. (2017, June 15). *PCA - Phân tích thành phần chính*. Tải từ: <https://machinelearningcoban.com/2017/06/15/pca/>
16. Quy, N. D. (n.d.). *Các phương pháp scaling trong học máy*. Tải từ  
<https://ndquy.github.io/posts/cac-phuong-phap-scaling/>
17. GeeksforGeeks. (n.d.). *What is Standardization in Machine Learning?* Tải từ:  
<https://www.geeksforgeeks.org/what-is-standardization-in-machine-learning/>
18. BigDataUni. (n.d.). *Các phương pháp đánh giá trong thuật toán clustering*. Tải từ:  
<https://bigdatauni.com/tin-tuc/cac-phuong-phap-danh-gia-trong-thuat-toan-clustering.html>
19. Nguyen Huu Dien. (2023) *Phân cụm dữ liệu K-Means, DBSCAN và AP*. Tải từ:  
[https://www.kaggle.com/code/nguyenhuudien/ph-n-c-m-d-li-u-k-means-dbscan-v-ap/code?fbclid=IwZXh0bgNhZW0CMTAAR2FTdXa2tWPpUzaW3irhFIEH\\_vGfg8KCIH14gzHYrqj5ZpFNFH4vpE-VuI\\_aem\\_p4NYHnZoxNVSuEb7y505Bg](https://www.kaggle.com/code/nguyenhuudien/ph-n-c-m-d-li-u-k-means-dbscan-v-ap/code?fbclid=IwZXh0bgNhZW0CMTAAR2FTdXa2tWPpUzaW3irhFIEH_vGfg8KCIH14gzHYrqj5ZpFNFH4vpE-VuI_aem_p4NYHnZoxNVSuEb7y505Bg)
20. Test. (2021, January 11). CH 10 OPTICS [Video]. YouTube. Tải từ:  
<https://www.youtube.com/watch?v=CV0mWaHOTA8>
21. Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J. (1999). OPTICS: Ordering points to identify the clustering structure. In *Proceedings of the ACM SIGMOD International Conference on Management of Data* (pp. 49–60). ACM Press. Tải từ:  
<https://www.dbs.uni.lmu.de/Publicationen/Papers/OPTICS.pdf>