

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH**

**KHOA ĐÀO TẠO CHẤT LƯỢNG CAO**

**NGÀNH CÔNG NGHỆ THÔNG TIN**

**🙞🕮🙜**



**BÁO CÁO CUỐI KÌ**

**MÔN: ĐỒ ÁN 3**

**TÌM HIỂU VÀ DEMO GOM CỤM DỮ LIỆU SỬ DỤNG THUẬT TOÁN DBSCAN**

**GVHD: Ts. Lê Văn Vinh**

**Sinh viên thực hiện MSSV**

**Luyện Ngọc Thanh 17110221**

**Nguyễn Thanh Lập 17110169**

**TP. Hồ Chí Minh, tháng 12 năm 2020**

**LỜI CẢM ƠN**

Để hoàn thành báo cáo môn học này, lời đầu tiên, nhóm chúng tôi xin bày tỏ lòng biết ơn chân thành và sâu sắc nhất đến giảng viên hướng dẫn: Tiến sĩ Lê Văn Vinh đã tận hình hướng dẫn chúng tôi trong suốt quá trình nghiên cứu và thực hiện đề tài. Chúng tôi xin gửi lời cảm ơn đến quý thầy cô trong khoa Đào Tạo Chất Lượng Cao, quý thầy cô trong khoa Công Nghệ Thông Tin đã trang bị cho chúng tôi những kiến thức và kinh nghiệm quý giá trong quá trình học tập và nhiệt tình giúp đỡ chúng tôi thực hiện đề tài này.

Chúng tôi cũng xin chân thành cảm ơn phòng thư viện đã cung cấp tài liệu và tạo mọi điều kiện thuận lợi để chúng tôi hoàn thành báo cáo này. Mặc dù đã có nhiều cố gắng, nhưng do thời gian có hạn, trình độ, kỹ năng của bản thân còn nhiều hạn chế nên chắc chắn không tránh khỏi những hạn chế, thiếu sót. Rất mong được sự đóng góp, chỉ bảo, bổ sung thêm của thầy cô và các bạn.

Chúng tôi xin chân thành cảm ơn.

**MỤC LỤC**

[**MỤC LỤC HÌNH 2**](#_Toc61032758)

[**CHƯƠNG 1: XÁC ĐỊNH YÊU CẦU 3**](#_Toc61032759)

[**1.1. Lý do chọn đề tài 3**](#_Toc61032760)

[**1.2. Mục tiêu nghiên cứu 3**](#_Toc61032761)

[**CHƯƠNG 2: GIỚI THIỆU VỀ MACHINE LEARNING 4**](#_Toc61032762)

[**2.1. Machine learning là gì? 4**](#_Toc61032763)

[**2.2. Các loại hệ thống Machine learning 4**](#_Toc61032764)

[**2.3. Giải pháp 4**](#_Toc61032765)

[**2.3.1. Supervised learning 4**](#_Toc61032766)

[**2.3.2. Unsupervised learning 5**](#_Toc61032767)

[**CHƯƠNG 3: TÌM HIỂU THUẬT TOÁN DBSCAN 7**](#_Toc61032768)

[**3.1. Sơ lược về DBSCAN 7**](#_Toc61032769)

[**3.2. Các khái niệm 8**](#_Toc61032770)

[**3.3. Giải thuật 10**](#_Toc61032771)

[**CHƯƠNG 4. DEMO 12**](#_Toc61032772)

[**4.1. Mô tả tập dữ liệu 12**](#_Toc61032773)

[**4.2. Demo gom cụm với thuật toán dbscan dựa trên ngôn ngữ python 12**](#_Toc61032774)

[**4.3. Đánh giá kết quả phân cụm 16**](#_Toc61032775)

[**CHƯƠNG 5: KẾT LUẬN 16**](#_Toc61032776)

[**5.1. Tổng kết 16**](#_Toc61032777)

[**5.2. Ưu điểm của thuật toán DBSCAN 16**](#_Toc61032778)

[**5.3. Nhược điểm của thuật toán DBSCAN 16**](#_Toc61032779)

[**5.4. Độ phức tạp của thuật toán DBSCAN 16**](#_Toc61032780)

[**CHƯƠNG 6: TÀI LIỆU THAM KHẢO 17**](#_Toc61032781)

# **MỤC LỤC HÌNH**

[Hình 1: Mô phỏng quy trình gom cụm 8](#_Toc61021135)

[Hình 2: Directly density-reachable 8](#_Toc61021136)

[Hình 3: Density-reachable 9](#_Toc61021137)

[Hình 4: Density-connected 10](#_Toc61021138)

[Hình 5: Giải thuật dbscan 11](#_Toc61021139)

[Hình 6: Tập dữ liệu mall\_customers 12](#_Toc61021140)

[Hình 7: Đọc dữ liệu từ file mall\_customers.csv 12](#_Toc61021141)

[Hình 8: Đọc và xử lý dữ liệu từ dataset 13](#_Toc61021142)

[Hình 9: Xác định eps và min\_samples 14](#_Toc61021143)

[Hình 10: Heatplot xác định số cluster 14](#_Toc61021144)

[Hình 11: Heatplot xác định eps và min\_samples 15](#_Toc61021145)

[Hình 12: Hàm phân cụm với 2 tham số eps và min\_samples 15](#_Toc61021146)

[Hình 13: Tiến hành phân cụm và visualize kết quả 16](#_Toc61021147)

[Hình 14: Visualize kết quả phân cụm với thuật toán DBSCAN 16](#_Toc61021148)

[Hình 15. Thông tin các cluster 17](#_Toc61021149)

# **CHƯƠNG 1: XÁC ĐỊNH YÊU CẦU**

## **1.1. Lý do chọn đề tài**

Trong những năm gần đây, trí tuệ nhân tạo (artificial interlligence – AI) nổi lên như một bằng chứng cho cuộc cách mạng lần thứ tư (1-động cơ hơi nước, 2-năng lượng điện, 3-công nghệ thông tin). Trí tuệ nhân tạo đã và đang trở thành thành phần cốt lõi trong các hệ thống công nghệ cao. Nó có mặt hầu hết ở các lĩnh vực trong đời sống như xe tự hành của Google và Tesla, hệ thống tự tag khuôn mặt trong ảnh của Facebook, trợ lý ảo Siri của Apple,…Đó chỉ là vài ứng dụng nổi bật trong vô vàng những ứng dụng của trí tuệ nhân tạo.

Học máy (machine learning – ML) là một tập con của trí tuệ nhân tạo. Nó là một lĩnh vực nhỏ trong khoa học máy tính có khả năng tự học hỏi dựa trên dữ liệu được đưa vào mà không cần phải được lập trình cụ thể và nó có vai trò then chốt cho việc phát triển trí tuệ nhân tạo

Sau khi được giáo viên Ts. Lê Văn Vinh phổ biến nhóm đã thấy được tầm quan trọng của machine learning trong lĩnh vực công nghệ ngày nay nên nhóm đã quyết định chọn đề tài môn đồ án 3 là “Tìm hiểu và demo gom cụm dữ liệu sử dụng thuật toán DBSCAN”

## **1.2. Mục tiêu nghiên cứu**

* Tìm hiểu về Machine Learning.
* Tìm hiểu và hiểu được thuật toán DBSCAN.
* Xây dựng một chương trình demo gom cụm với thuật toán DBSCAN dựa trên ngôn ngữ python

# **CHƯƠNG 2: GIỚI THIỆU VỀ MACHINE LEARNING**

**2.1. Machine learning là gì?**

Machine learning là khoa học (và nghệ thuật) lập trình máy tính để chúng có thể học từ dữ liệu.

**2.2. Các loại hệ thống Machine learning**

Có rất nhiều loại hệ thống Machine learning khác nhau có thể được phân loại dựa trên:

* Có được đào tạo với sự giám sát của con người hay không:
* Supervised learning (Học có giám sát)
* Unsupervised learning (Học không giám sát)
* Semi-supervised learning (Học bán giám sát)
* Reinforcement learning (Học tăng cường)
* Có thể học tăng dần một cách nhanh chóng hay không (online versus batch learning).
* Có hoạt động đơn giản bằng cách so sánh các điểm dữ liệu mới với các điểm dữ liệu đã biết hay thay vào đó là phát hiện các mẫu trong dữ liệu đào tạo và xây dựng mô hình dự đoán, giống như các nhà khoa học làm (instance-based versus model-based learning).

**2.3. Giải pháp**

* Sau khi tìm hiểu các dạng Machine learning khác nhau, nhóm nghiên cứu quyết định tìm hiểu sâu về hai dạng của Machine learning là Supervised learning và Unsupervised learning để tìm ra giải pháp cho bài toán. Từ đó sẽ tìm hiểu một số thuật toán phù hợp để giải quyết vấn đề đặt ra ứng với mỗi dạng.
* Cụ thể với Supervised learning nhóm sẽ tìm hiểu về thuật toán DBSCAN.

### **2.3.1. Supervised learning**

* Trong học tập có giám sát, dữ liệu training bạn cung cấp cho thuật toán bao gồm các giải pháp mong muốn, được gọi là nhãn.
* Một ví dụ về học tập có giám sát điển hình là phân loại. Bộ lọc thư rác là một ví dụ điển hình về điều này: nó được đào tạo với nhiều email mẫu cùng với lớp của chúng (spam hoặc ham) và nó phải học cách để phân loại các email mới.
* Dưới đây là một số thuật toán học có giám sát quan trọng:
* k-Nearest Neighbors
* DBSCAN
* Linear Regression
* Logistic Regression
* Support Vector Machines (SVMs)
* Decision Trees and Random Forests
* Neural networks

### **2.3.2. Unsupervised learning**

* Theo như tên gọi của thuật toán, việc học máy máy sẽ không có sự hướng dẫn rõ ràng và dữ liệu được đưa vào sẽ không được tiền xử lý bởi con người.
* Một số thuật toán không giám sát quan trọng:
* Clustering
* k-Means
* Hierarchical Cluster Analysis (HCA)
* Expectation Maximization
* Visualization and dimensionality reduction
* Principal Component Analysis (PCA)
* Kernel PCA
* Locally - Linear Embedding (LLE)
* t-distributed Stochastic Neighbor Embedding (t-SNE)
* Association rule learning
* Apriori
* Eclat
* Một ví dụ điển hình học tập không giám sát là giả sử bạn có nhiều dữ liệu về khách truy cập trang web đọc sách của mình. Bạn có thể muốn chạy một thuật toán phân cụm để cố gắng phát hiện các nhóm đọc giả truy cập tương tự (như nhóm yêu thích thể loại khoa học viễn tưởng, hay nhóm thích sách về kinh tế, …). Bạn không cho thuật toán biết trước là khách truy cập thuộc nhóm nào: nó sẽ phải tìm thấy các kết nối đó mà không cần sự trợ giúp của bạn.

# **CHƯƠNG 3: TÌM HIỂU THUẬT TOÁN DBSCAN**

**3.1. Sơ lược về DBSCAN**

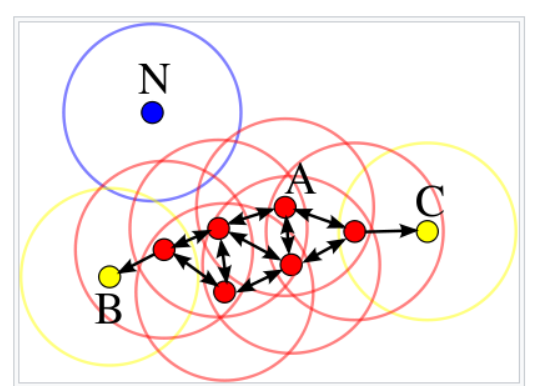
DBSCAN hay Density Base Spatial Clustering of Application with Noise, ý tưởng chính của thuật toán là một điểm sẽ thuộc về 1 cluster nếu nó gần các điểm thuộc cluster đó.

Có 2 tham số chính của DBSCAN:

* eps: Là khoảng cách của các “hàng xóm”. Hai điểm được coi là “hàng xóm” nếu khoảng cách giữa chúng nhỏ hơn hoặc bằng eps.
* minPts: Số điểm tối thiểu để tạo thành 1 cluster.

Dựa vào 2 tham số trên, các điểm được phân loại thành điểm lõi, điểm biên hoặc điểm ngoài.

* Điểm lõi: Một điểm sẽ là điểm lõi nếu có tối thiểu minPts số điểm tính cả chính nó được bao quanh bởi 1 khu vực với bán kính là eps.
* Điểm biên: Một điểm sẽ là điểm biên nếu nó có thể được “chạm tới” từ điểm lõi và có ít hơn minPts số điểm ở trong khu vực của nó.
* Điểm ngoài: Một điểm sẽ là điểm ngoài nếu nó không phải là điểm lõi cũng như là điểm biên.

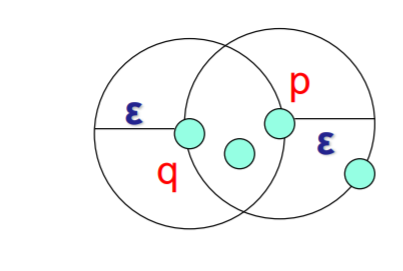


Hình : Mô phỏng quy trình gom cụm

Ở trường hợp này minPts sẽ là 4, điểm A sẽ là điểm lõi vì quanh nó có tối thiểu 4 điểm với bán kính là eps, B và C sẽ là điểm biên vì nó có thể “chạm tới” từ điểm A và có ít hơn 4 điểm bao quanh nó, cuối cùng điểm N sẽ là điểm ngoài bởi vì N không thể được “chạm tới” bởi bất cứ điểm nào nằm trong cluster.

**3.2. Các khái niệm**

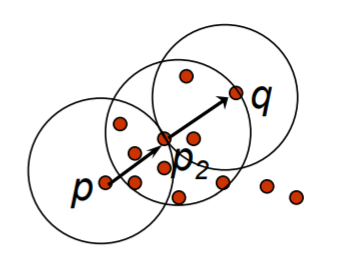
*Directly density-reachable* (khả năng đạt được trực tiếp): q có thể đạt được trực tiếp từ p nếu q trong vùng láng giềng ε-neighborhood của p và p phải là core object.



Hình : Directly density-reachable

*Density-reachable* (khả năng đạt được):

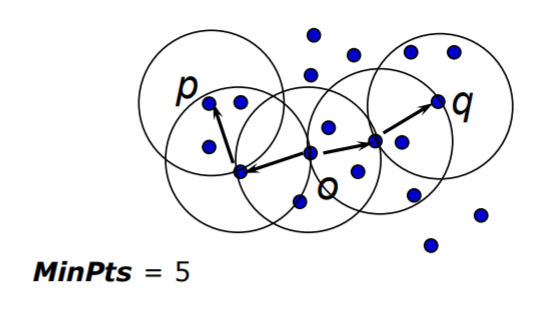
* Cho trước tập đối tượng D, ε và MinPts
* q density-reachable từ p nếu ∃ chuỗi các đối tượng p1, ..., pn ∈ D với p1 = p và pn = q sao cho pi+1 directly density-reachable từ pi theo các thông số ε và MinPts, 1 ≤ i ≤ n.
* Bao đóng truyền (transitive closure) của directly densityreachable ⁄ Quan hệ bất đối xứng (asymmetric relation)



Hình : Density-reachable

*Density-connected* (nối kết dựa trên mật độ):

* Cho trước tập các đối tượng D, ε và MinPts
* p, q ∈ D
* q density-connected với p nếu ∃ o ∈ D sao cho cả q và p đều density-reachable từ o theo các thông số ε và MinPts.
* Quan hệ đối xứng

****

Hình : Density-connected

**3.3. Giải thuật**

* Xác định minPts và eps.
* Điểm bắt đầu được chọn ngẫu nhiên ở một khu vực được xác định bởi bán kính eps. Nếu ở đó có tối thiểu minPts số điểm trong vùng phụ cận thì điểm đó được đánh dấu là điểm lõi nếu không nó sẽ được coi là điểm gây nhiễu (noise point). Một khi thông tin về 1 cluser được hình thành (tạm gọi là cluster A), tất cả các điểm ở trong khu vực của điểm khởi tạo sẽ trở thành một phần của cluster A. Nếu có một điểm mới mà điểm đó cũng là điểm lõi thì các điểm trong khu vực của điểm đó cũng được thêm vào cluster A.

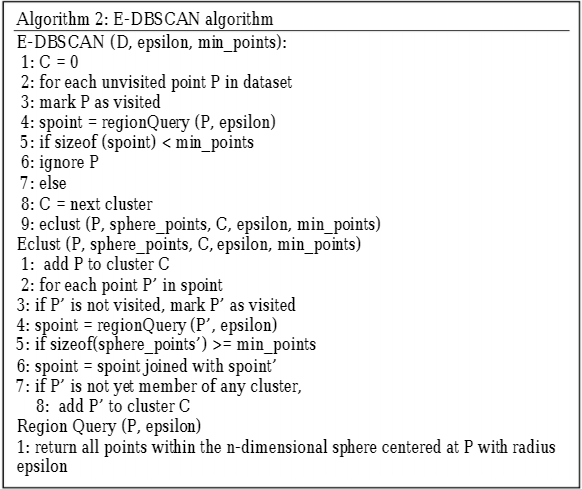
*Chú ý****:*** Điểm được coi là điểm gây nhiễu cũng có thể được kiểm tra lại và trở thành một phần của cluster nếu thỏa mãn điều kiện.

* Bước tiếp theo là chọn ngẫu nhiên một điểm ở giữa các điểm mà điểm đó chưa được “ghé thăm” ở bước trước đó. Và vẫn là quy trình kiểm tra được áp dụng.
* Tiến trình này sẽ kết thúc khi tất cả các điểm đã được “ghé thăm”.

*Chú ý****:*** Khoảng cách giữa các điểm được xác định bằng cách sử dụng phương pháp tính khoảng cách được đề cập ở giải thuật k-means.

Bằng cách áp dụng các bước trên, giải thuật DBSCAN phù hợp để tìm các khu vực có tỉ trọng cao và tách nó ra khỏi khu vực có tỉ trọng thấp.

Điều kiện cần để hình thành một cluster đó là phải có tối thiểu một điểm lõi. Trường hợp xấu nhất cluster sẽ có chỉ một điểm lõi và các điểm biên của nó.



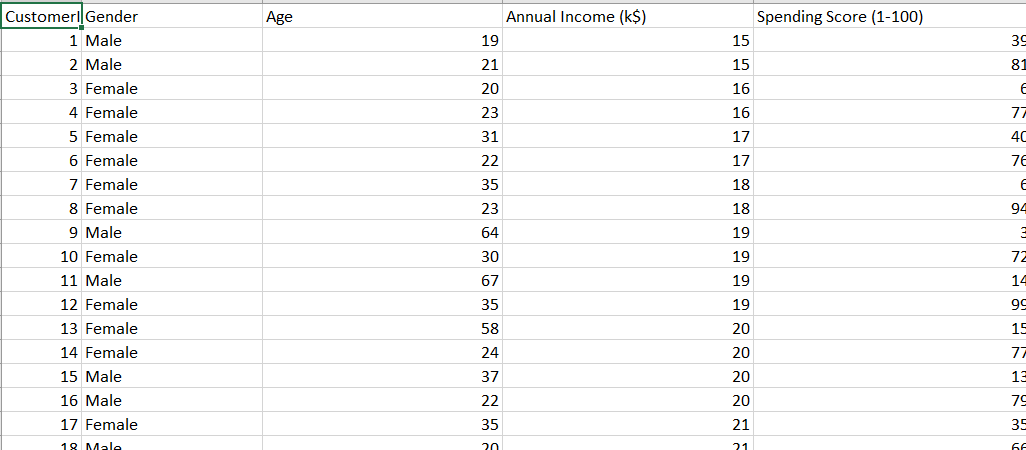
Hình : Giải thuật dbscan

**CHƯƠNG 4. DEMO**

**4.1. Mô tả tập dữ liệu**

Tập dữ liệu nhóm sử dụng là tập dữ liệu được thu thập về phần trăm sử dụng thu nhập cá nhân của khách hàng tại 1 mall, với những thông tin về giới tính, độ tuổi, thu nhập và phần trăm sử dụng thu nhập.

Thu nhập cá nhân có đơn vị là k$(ngàn $) và phần trăm sử dụng thu nhập được tính từ 1 – 100.

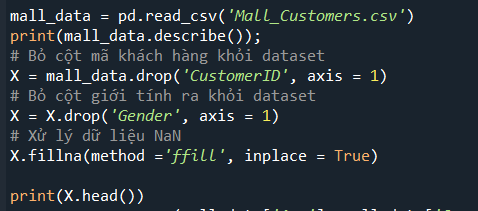


Hình : Tập dữ liệu mall\_customers

Nhóm dự định sử dụng tập dữ liệu này để tiến hành phân cụm dựa trên độ tuổi, thu nhập và phần trăm sử dụng thu nhập.

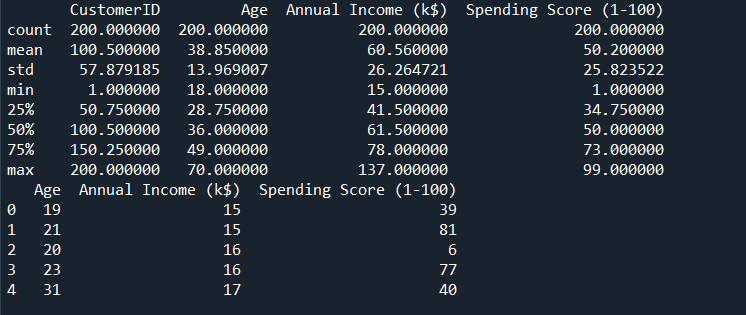
**4.2. Demo gom cụm với thuật toán dbscan dựa trên ngôn ngữ python**

Đầu tiên chúng ta tiến hành đọc dữ liệu từ dataset



Hình : Đọc dữ liệu từ file mall\_customers.csv

Kết quả thu được



Hình : Đọc và xử lý dữ liệu từ dataset

Để xác định được tham số eps và min\_samples (min Pts) trong thuật toán DBSCAN, ta phải dựa vào tham số silhouette score để đưa ra lựa chọn tốt nhất.

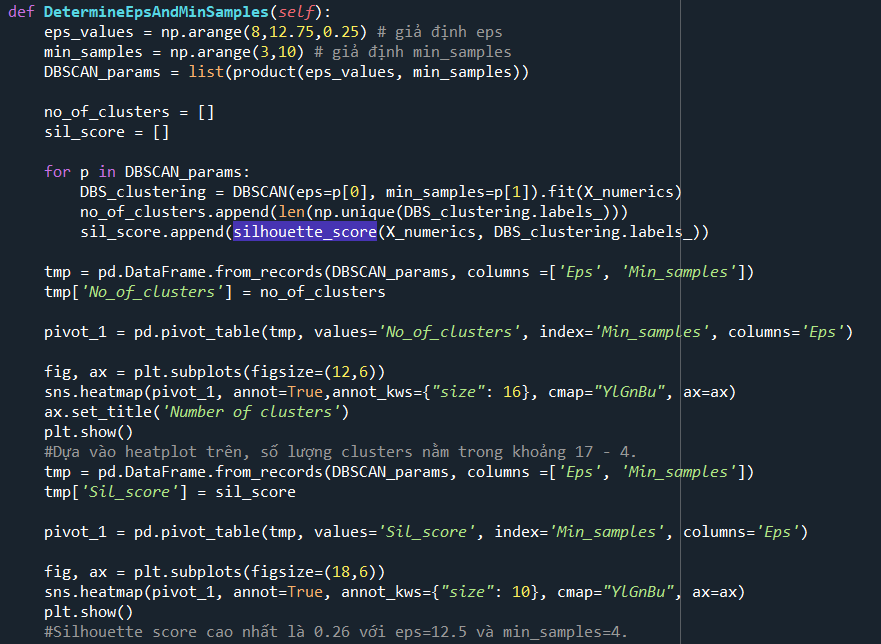
*Silhouette score* làmột thông số đánh giá hiệu năng của việc phân cụm. Cụ thể chúng ta sử dụng silhouette score. Với mỗi ví dụ, giá trị silhouette score của nó được tính như sau:

Silhouette score =

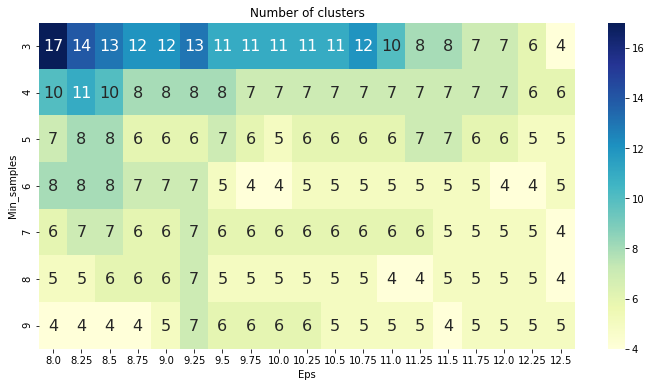
Trong đó:

* a là khoảng cách trung bình tới các ví dụ khác ở trong cùng một cụm
* b là khoảng cách trung bình tới các ví dụ trong cụm gần nhất  
  Giá trị silhouette score nằm trong khoảng từ -1 đến 1
* silhouette score nằm gần 1 nghĩa là ví dụ đang được phân cụm chính xác, xa các cụm khác
* silhouette score gần 0 nghĩa là ví dụ đang nằm gần đường bao của cụm
* silhouette score gần -1 nghĩa là ví dụ đang bị phân sai cụm  
  Chúng ta có thể lấy giá trị silhouette score thông qua hàm silhouette\_score của thư viện scikit-learn.

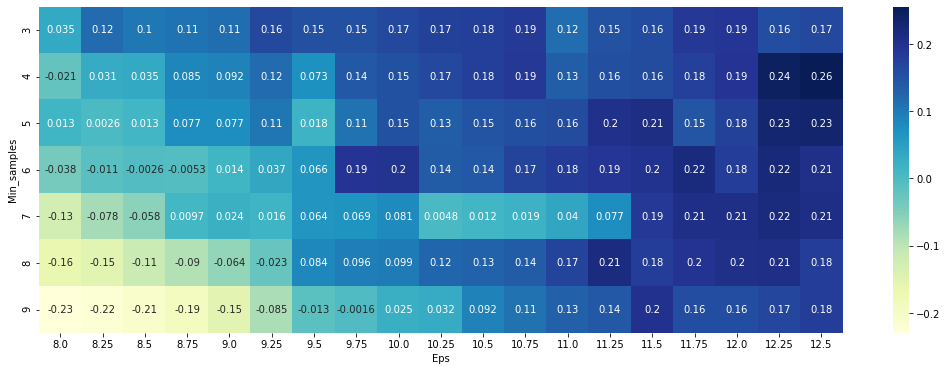
Tiến hành giả định eps và min\_samples sẽ nằm trong một khoảng giá trị từ đó thực hiện tính toán tham số silhouette score. Dựa vào kết quả tính toán, xây dựng heatplot để đưa ra kết luận về eps và min\_samples.



Hình : Xác định eps và min\_samples

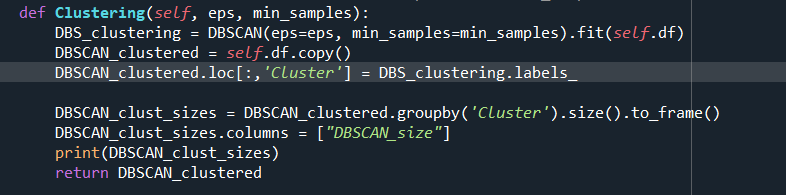


Hình : Heatplot xác định số cluster



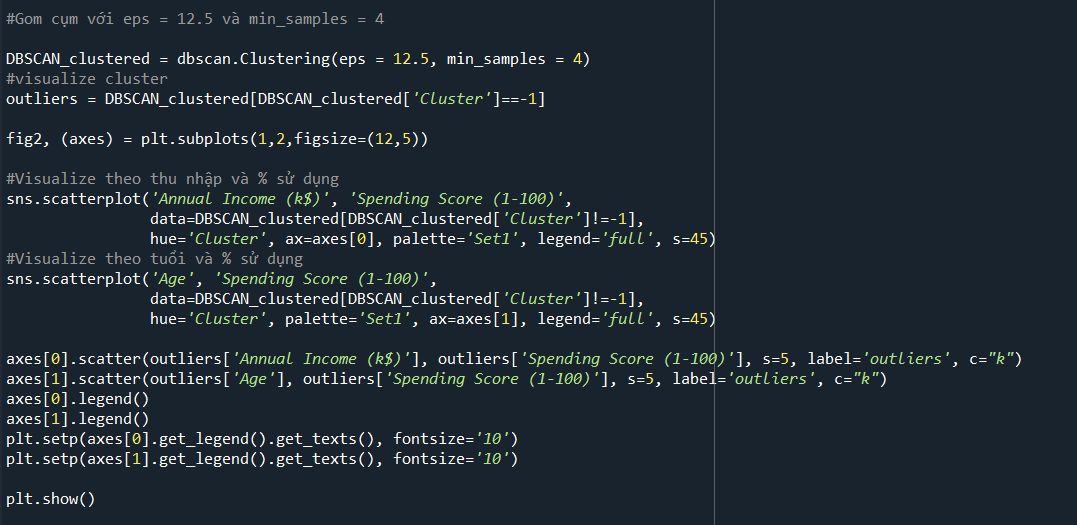
Hình : Heatplot xác định eps và min\_samples

Dựa vào 2 heatplot trên ta thấy với eps = 12.5 và min\_samples = 4 thì tham số silhouette score đạt được giá trị lớn nhất là 0.26 chính vì vậy ta chọn eps = 12.5 và min\_samples = 4 để tiến hành phân cụm.



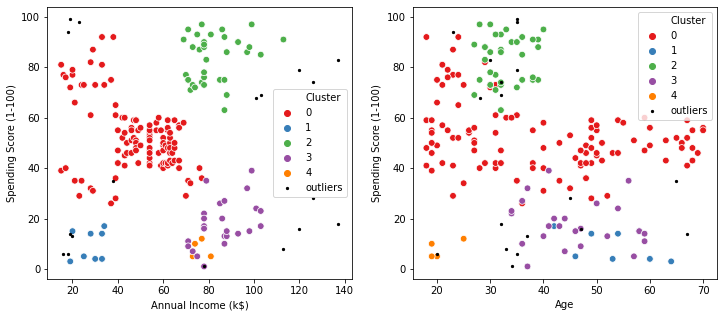
Hình : Hàm phân cụm với 2 tham số eps và min\_samples

Metric được sử dụng để tính toán khoảng cách giữa các điểm là Euclidean có công thức như sau:



Hình : Tiến hành phân cụm và visualize kết quả

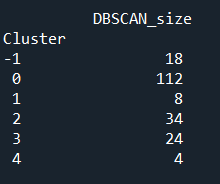
Sau khi thực hiện phân cụm, bởi vì việc gắn label cho từng điểm ngay trên dataframe gốc nên ta dễ dàng visualize data theo từng cặp dữ liệu. Dưới đây là data khi được visualize theo thu nhập – phần trăm chi tiêu và tuổi – phần trăm chi tiêu.



Hình : Visualize kết quả phân cụm với thuật toán DBSCAN

**4.3. Đánh giá kết quả phân cụm**

Bằng việc sử dụng thuật toán DBSCAN để phân cụm cho dữ liệu chi tiêu của người dùng, chúng ta đã phân ra được 5 cụm tương ứng là cluster 0, 1, 2, 3, 4 và các điểm noise hay outliner.



Hình . Thông tin các cluster

Ta thấy cụm được đánh số 4 có vừa đủ số minPts để trở thành 1 cluster, qua đó ta thấy được việc lựa chọn 4 là minPts là hợp lý.

Có 18 điểm không thuộc bất cứ cluster nào trên tổng số 200 điểm data được cung cấp qua đó cũng cho thấy rằng kết quả phân cụm lần này được đánh giá là tốt.

**CHƯƠNG 5: KẾT LUẬN**

**5.1. Tổng kết**

Trong quá trình thực hiện đề tài nhóm đã tiến hành tìm hiểu sâu và rõ hơn về ngành học máy. Nhìn chung nhóm đã đạt được những mục tiêu đề ra ban đầu. Qua đó giúp nhóm em cải thiện được nhiều kỹ năng hơn và có thêm nhiều kinh nghiệm hơn.

Sau thời gian học tập và thực hiện nhóm đã hiểu về:

* Có kiến thức cơ bản về Machine Learning
* Lập trình với ngôn ngữ python.
* Các thuật toán cơ bản trong ngành học máy.

**5.2. Ưu điểm của thuật toán DBSCAN**

* Có thể phát hiện cluster có hình dạng bất kì
* Chỉ yêu cầu một hàm đo khoảng cách và hai tham số đầu vào: Eps và MinPts
* Cho ra kết quả tốt và thực thi hiệu quả trên nhiều tập dữ liệu

**5.3. Nhược điểm của thuật toán DBSCAN**

* Không thích hợp cho việc tìm các cluster trong CSDL cực lớn
* Nếu tập dữ liệu có mật độ thay đổi lớn, thuật toán quản lý kém hiệu quả

**5.4. Độ phức tạp của thuật toán DBSCAN**

* Độ phức tạp trung bình của mỗi truy vấn là O(logn)
* Độ phức tạp của thuật toán là O(nlogn), n là kích thước tập dữ liệu

# **CHƯƠNG 6: TÀI LIỆU THAM KHẢO**

[1]. <http://www.jaist.ac.jp/~bao/VNAlectures/Clustering_VNA_Hien(M6M7).pdf>

[2].<https://www.kaggle.com/datark1/customers-clustering-k-means-dbscan-and-ap/notebook>

[3]. <http://www.cit.ctu.edu.vn/~dtnghi/rech/dir_0/Duy-et-al-New.pdf>

[4]. <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html>