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

**KHOA ĐÀO TẠO CHẤT LƯỢNG CAO**

**NGÀNH CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN 3**

**Báo cáo thuật toán K-mean**

**GVHD: TS Lê Văn Vinh**

**SVTH: Trần Ngọc Hùng**

**MSSV: 17110157**

**SVTH: Châu Huỳnh Phước Toàn**

**MSSV: 17110240**

**TP. Hồ Chí Minh, tháng 11 năm 2020**

**NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN**

|  |  |
| --- | --- |
| **ĐIỂM** |  |

Giảng viên hướng dẫn

TS. Lê Văn Vinh

# **Mục Lục**

[**LỜI CẢM ƠN**](#_30j0zll) **4**

[**DANH MỤC HÌNH VẼ**](#_adlxgtcf6vya) **5**

[**CHƯƠNG 1: MỞ ĐẦU**](#_wrp1xniffina) **6**

[1.1 Lý do chọn đề tài](#_2s8eyo1) 6

[1.2 Mục tiêu đề tài](#_740g5rqun6cq) 6

[Hoàn thành tập tài liệu nghiên cứu sử dụng thuật toán K-means](#_rr7j7w46f0rk) 6

[**CHƯƠNG 2: TỔNG QUAN VỀ THUẬT TOÁN GOM CỤM (CLUSTERING)**](#_uih86bm9kyi4) **7**

[2.1 Thuật toán gom cụm là gì ?](#_slu433tmyb6p) 7

[2.2 Ví dụ về gom cụm](#_iomup7appyun) 7

[**CHƯƠNG 3: TỔNG QUAN VỀ THUẬT TOÁN GOM CỤM K-MEANS (K-MEANS CLUSTERING)**](#_q2groldvtur2) **8**

[3.1 Thuật toán gom cụm k-means là gì ?](#_1plxnsyysvx2) 8

[3.2 Các bước thực hiện thuật toán K-means](#_g5oxat5auovj) 8

[**CHƯƠNG 4: CHƯƠNG TRÌNH MINH HỌA**](#_bfkhhslevy1d) **10**

[4.1 Cấu hình máy và phần mềm sử dụng](#_laajdtqt66x2) 10

[4.2 Đầu vào và kết quả chương trình](#_9d9j8khsop9z) 10

[**CHƯƠNG 5: KẾT LUẬN**](#_va146id5rlgo) **14**

[5.1 Kiến thức tìm hiểu được](#_x64t8qyqamos) 14

[5.2 Hướng phát triển](#_nlf927t4dgde) 14

# **LỜI CẢM ƠN**

Qua đề tài này chúng em đã có thêm những kiến thức về học máy. Chúng em xin chân thành gửi lời cảm ơn đến những người đã hỗ trợ, giúp đỡ chúng em rất nhiều trong thời gian qua thực hiện đề tài.

Đầu tiên chúng em xin gửi lời cảm ơn tới các Thầy Cô trong trường Đại Học Sư Phạm Kỹ Thuật TP.Hồ Chí Minh, đặc biệt là các Thầy Cô trong Khoa Công Nghệ Thông Tin và Khoa Đào Tạo Chất Lượng Cao, đã nhiệt tình giảng dạy, trao dồi, vun đắp kiến thức cho chúng em trong nhiều năm ngồi trên ghế nhà trường.

Đặc biệt chúng em chân thành cảm ơn Thầy Lê Văn Vinh trong thời gian qua đã cung cấp cho chúng em những kiến thức về cũng như tài liệu rất hữu ích và hướng dẫn chi tiết trong quá trình chúng em làm để tài. Thầy luôn giải đáp tận tình những vướng mắc khi chúng em gặp phải. Và chúng em cũng xin gửi lời cảm ơn đến bạn bè trong và ngoài lớp - những người bạn đã cùng chia sẻ, trao đổi kiến thức với chúng em.

Trong phạm vi khả năng cho phép, chúng em đã rất cố gắng để hoàn thành đề tài một cách tốt nhất. Tuy nhiên không thể tránh khỏi nhiều thiếu sót, chúng em kính mong nhận được sự cảm thông và những ý kiến đóng góp của quý thầy cô và các bạn.

Sinh viên thực hiện

Trần Ngọc Hùng

Châu Huỳnh Phước Toàn

# **DANH MỤC HÌNH VẼ**

Hình 1 : Các bước thực hiện thuật toán K-means

Hình 2: Implement các thư viện môi trường.

Hình 3: Khởi tạo tập dữ liệu.

Hình 4: Kết quả dataset đã khởi tạo dữ liệu

Hình 5: Import thư viện KMeans và khởi tạo các tâm cụm.

Hình 6: Kết quả sau khi khởi tạo KMeans

Hình 7: Sử dụng thư viện matplotlib để in các giá trị trong tập dữ liệu (dataset).

Hình 8: Thể hiện tập điểm dữ liệu của tập dữ liệu lên biểu đồ.

Hình 9: Sử dụng thư viện KMean để tìm tọa độ tâm cụm.

Hình 10: In ra các giá trị tâm cụm.

Hình 11: Sử dụng thư viện matplotlib để thể hiện các cụm và tâm cụm.

Hình 12: Hiển thị dữ liệu trên biểu đồ.

# **CHƯƠNG 1: MỞ ĐẦU**

## **1.1 Lý do chọn đề tài**

Với mong muốn tiếp cận với bộ môn học máy, và thực hiện đồ án 3 nhóm chúng em đã chọn đề tài K-means thuộc nhóm gom cụm (clustering) để thực hiện. Với tính chất dễ tiếp cận, nhóm em mong muốn thông qua đề tài có thể hiểu hơn về học máy để có thể áp dụng vào thực tế và các đồ án sau này..

## **1.2 Mục tiêu đề tài**

## Hoàn thành tập tài liệu nghiên cứu sử dụng thuật toán K-means

# **CHƯƠNG 2: TỔNG QUAN VỀ THUẬT TOÁN GOM CỤM (CLUSTERING)**

## **2.1 Thuật toán gom cụm là gì ?**

Phân cụm là :

* Kỹ thuật rất quan trọng trong khai phá dữ liệu
* Có rất nhiều định nghĩa khác nhau về kỹ thuật này
* Về bản chất có thể hiểu phân cụm là các quy trình tìm cách nhóm các đối tượng đã cho vào các cụm (clusters), sao cho các đối tượng trong cùng một cụm tương tự nhau và các đối tượng khác cụm thì không tương tự nhau
* Mục đích của phân cụm là tìm ra bản chất bên trong các nhóm của dữ liệu
* Tuy nhiên, không có tiêu chí nào được xem là tốt nhất để đánh giá hiệu quả của phân tích phân cụm, điều này phụ thuộc vào mục đích của phân cụm như: giảm dữ liệu (data reduction), “cụm tự nhiên” (“natural clusters”), cụm “có ích” (“useful” clusters), phát hiện phần tử ngoại lai (outlier detection).

## **2.2 Ví dụ về gom cụm**

Giả sử có một công ty định ra mắt một mẫu sản phẩm mới T-shirt vào thị trường. Tất nhiên họ sẽ phải sản xuất rất nhiều size để phù hợp với sự đa dạng của thị trường người dùng. Với định hướng đo, công ty đã tiến hành khảo sát dữ liệu chiều cao và cân nặng của người dùng,

Công ty này ko thể đủ nguồn lực để có thể sản xuất áo với tất cả mọi size. Thực tế này trong kinh doanh bạn cũng dễ dàng hiểu được. THay vì đó, họ sẽ chia số lượng người dùng thành các size như là Small, Medium, Large và sản xuất chỉ 3 mẫu như thế. 3 mẫu này là đủ khớp với tất cả mọi người và thị trường. Ở đây việc phân chia các người dùng vào 3 nhóm trên sẽ được xử lý bằng kỹ thuật gom cụm Thuật toán này sẽ cho ta 3 size áo tối ưu nhất - thỏa mãn tất cả mọi người. Tất nhiên, nếu như ko thể tìm được 3 size áo vừa vặn thoả mãn mọi người trong nhóm, công ty sẽ chia nhỏ nhóm thêm thành nhiều nhóm khác, có thể là 5, có thể là nhiều hơn nữa

# **CHƯƠNG 3: TỔNG QUAN VỀ THUẬT TOÁN GOM CỤM K-MEANS (K-MEANS CLUSTERING)**

## **3.1 Thuật toán gom cụm K-means là gì ?**

K-means là thuật toán rất quan trọng và được sử dụng phổ biến trong kỹ thuật phân cụm dữ liệu.Ý tưởng chính của thuật toán K-means là tìm cách phân nhóm các đối tượng (objects) đã cho vào k cụm (k là số các cụm được xác định trước, k là số nguyên dương) sao cho tổng bình phương khoảng cách giữa các đối tượng đến tâm nhóm (centroid) là nhỏ nhất.

Thuật toán K-means áp dụng cho các đối tượng được biểu diễn bởi các điểm trong không gian vectơ d chiều U = {xi | i = 1, … , N}, với xi ∈ ℜ biểu thị đối tượng (hay điểm dữ liệu) thứ i

Thuật toán k-means gom cụm toàn bộ các điểm dữ liệu trong U thành k cụm C = {C1, C2,…, Ck} sao cho mỗi điểm dữ liệu xi nằm trong một cụm duy nhất. Để biết điểm dữ liệu thuộc cụm nào người ta gán cho nó một mã cụm. Các điểm có cùng mã cụm ở cùng cụm, trong khi các điểm khác mã cụm ở trong các cụm khác nhau. Một cụm có thể biểu thị bằng vectơ liên thuộc cụm v có độ dài N, với vi là mã cụm của xi. Giá trị k là đầu vào của thuật toán. Giá trị k dựa trên tiêu chuẩn tri thức trước đó. Trong các thuật toán gom cụm, các điểm được nhóm theo khái niệm “độ gần” hay “độ tương tự”. Với K-means, phép đo mặc định cho “độ tương tự” là khoảng cách Euclide

## **3.2 Các bước thực hiện thuật toán K-means**

Thuật toán k-means thực hiện qua các bước chính sau:

1. Chọn ngẫu nhiên k tâm cụm (centroid) cho k cụm (cluster). Mỗi cụm được đại diện bằng các tâm của cụm.

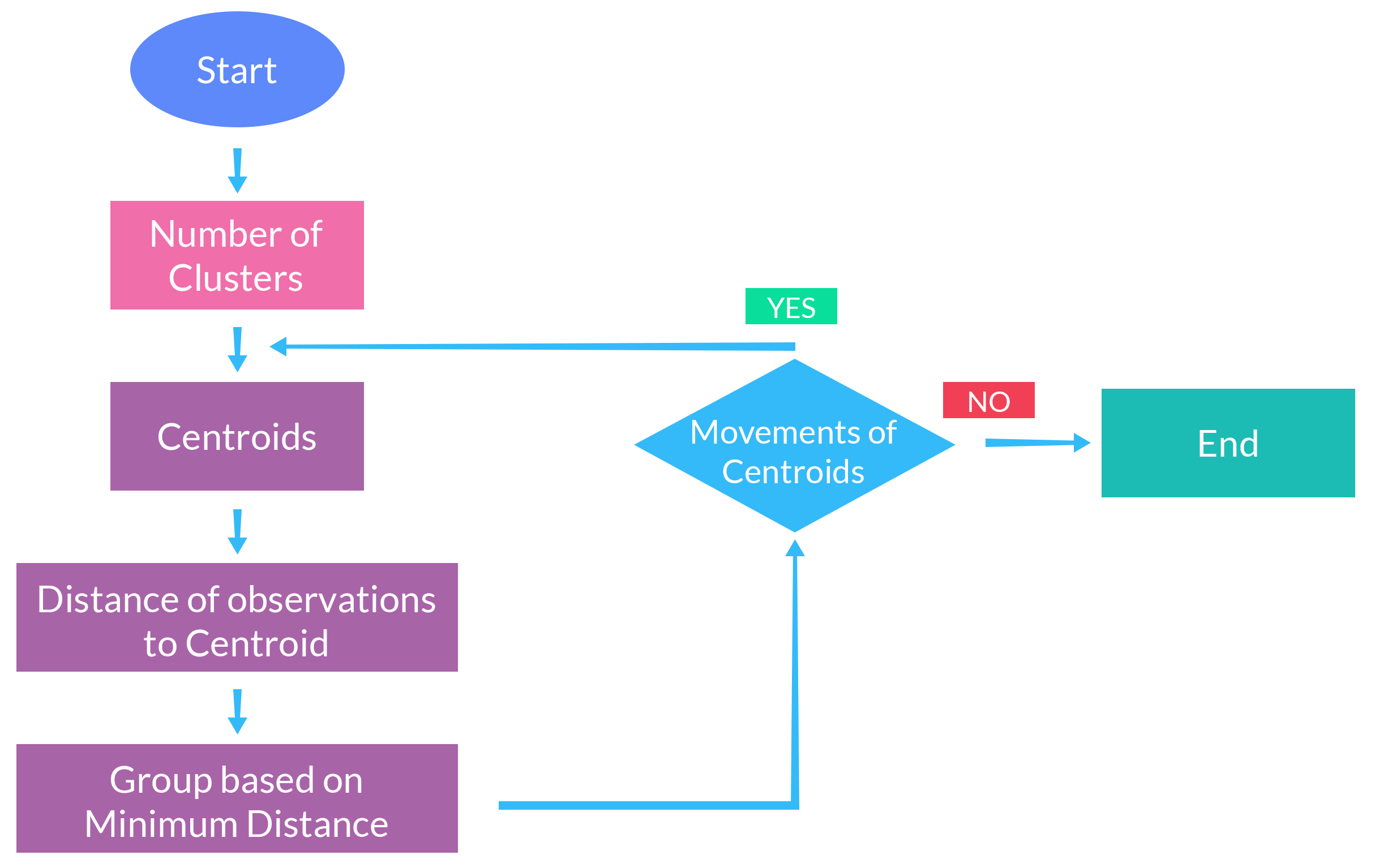
2. Tính khoảng cách giữa các đối tượng (objects) đến k tâm (thường dùng khoảng cách Euclide).

3. Nhóm các đối tượng vào nhóm gần nhất.

4. Xác định lại tâm mới cho các nhóm.

5. Thực hiện lại bước 2 cho đến khi không có sự thay đổi nhóm nào của các đối tượng.

Thuật toán K-means được phát triển bởi nhiều nhà nghiên cứu khác nhau, điển hình là Lloyd (1957), Forgey (1965), Friedman (1967), Rubin (1967), McQueen (1967).



Hình 1 : Các bước thực hiện thuật toán K-means

# **CHƯƠNG 4: CHƯƠNG TRÌNH MINH HỌA**

## **4.1 Cấu hình máy và phần mềm sử dụng**

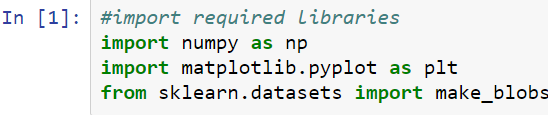
Cấu hình máy:

* Intel core i7 7700
* 16GB RAM
* Win 10 - 32bit

Phần mềm sử dụng:

## **4.2 Đầu vào và kết quả chương trình**

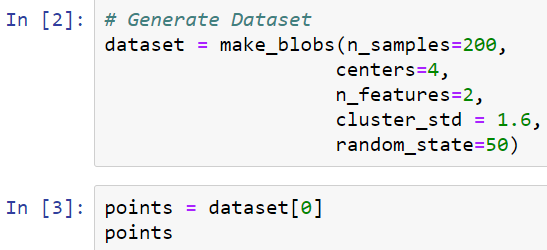
Mã lệnh:



Hình 2: Implement các thư viện môi trường.

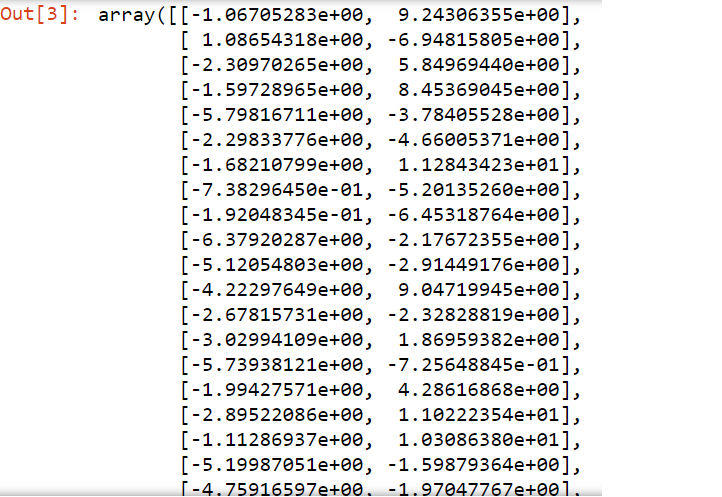
Chức năng dòng lệnh: import các thư viện xử lý cần thiết như numpy để xử lý mảng, matplotlib.pyplot thực hiện các thao tác liên quan đến biểu đồ, sklearn.datasets để load và trả về tập dữ liệu.

Mã lệnh:



Hình 3: Khởi tạo tập dữ liệu.

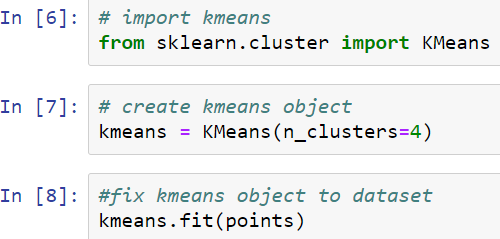
Kết quả trả về:



Hình 4: Kết quả dataset đã khởi tạo dữ liệu

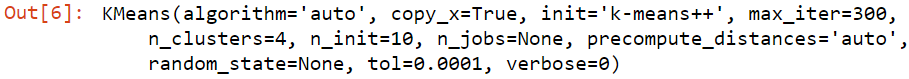
Chức năng dòng lệnh: khởi tạo tập dataset ngẫu nhiên làm dữ liệu đầu vào ( ở đây tập giá trị sẽ gồm 200 điểm với 2 thuộc tính x và y, 4 tâm cụm và giá trị sẽ được random).

Mã lệnh:



Hình 5: Import thư viện K-Means và khởi tạo các tâm cụm.

Kết quả trả về:



Hình 6: Kết quả sau khi khởi tạo K-Means

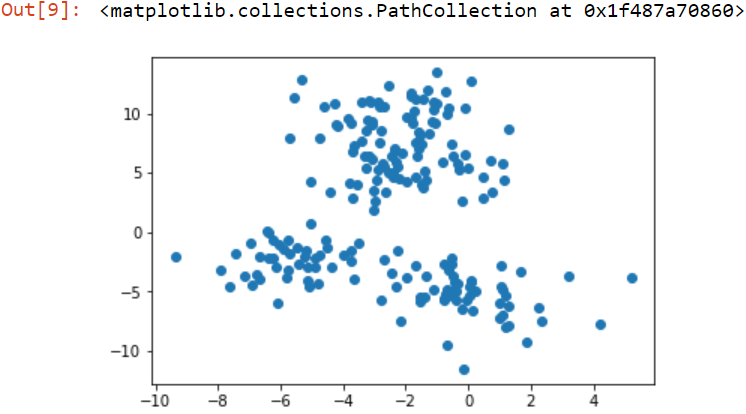
Chức năng dòng lệnh: import thư viện xử lý K-means sẵn có của Python, khởi tạo giá trị số cụm cần gom nhóm (chương trình demo là 4 cụm) và lấy dataset đã khởi tạo ở trên làm dữ liệu thực thi.

Mã lệnh:



Hình 7: Sử dụng thư viện matplotlib để in các giá trị trong tập dữ liệu (dataset).

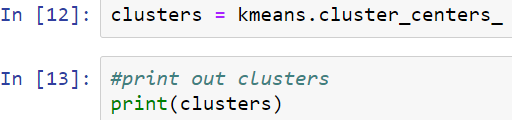
Kết quả trả về:



Hình 8: Thể hiện tập điểm dữ liệu của tập dữ liệu lên biểu đồ.

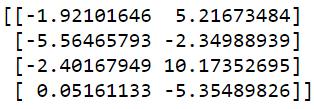
Chức năng dòng lệnh: sử dụng thư viện matplotlib để thể hiện các điểm trong tập dữ liệu dataset lên biểu đồ.

Mã lệnh:



Hình 9: Sử dụng thư viện KMean để tìm tọa độ tâm cụm.

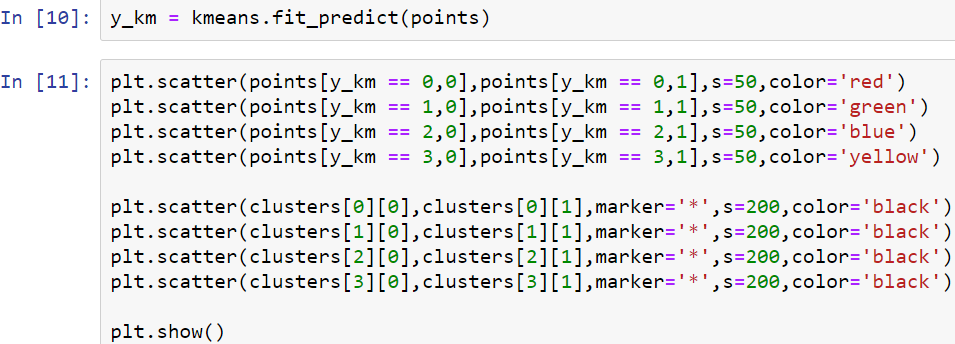
Kết quả trả về:



Hình 10: In ra các giá trị tâm cụm.

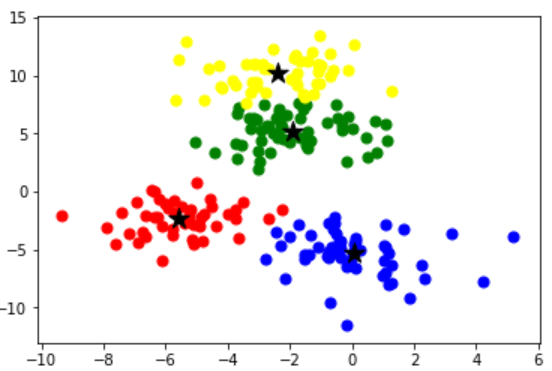
Chức năng dòng lệnh: sử dụng thư viện K-means để tính toán các giá trị tâm cụm và in ra màn hình.

Mã lệnh:



Hình 11: Sử dụng thư viện matplotlib để thể hiện các cụm và tâm cụm.

Kết quả trả về:



Hình 12: Hiển thị dữ liệu trên biểu đồ.

Chức năng dòng lệnh: sử dụng thư viện matplotlib thể hiện các giá trị sau khi gom cụm và các tâm cụm lệnh biểu đồ.

# **CHƯƠNG 5: KẾT LUẬN**

## **5.1 Kiến thức tìm hiểu được**

Nắm bắt được thuật toán gom cụm K-means, có thể sử dụng được tuy nhiên do thời gian và khả năng của các thành viên trong nhóm có hạn nên 1 số phần còn chưa hiểu rõ.

## **5.2 Hướng phát triển**

Nắm bắt chắc chắn hơn thuật toán gom cụm K-means, có thể sử dụng trong khóa luận tốt nghiệp.

.