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

**KHOA ĐÀO TẠO CHẤT LƯỢNG CAO**

**NGÀNH CÔNG NGHỆ THÔNG TIN**

**---o0o---**



**BÁO CÁO ĐỀ TÀI**

**TÌM HIỂU GIẢI THUẬT GOM CỤM DỮ LIỆU (CLUSTERING)**

**MÔN: ĐỒ ÁN 3**

**SVTH: LƯ MẠNH QUÂN 17110209**

**TRƯƠNG MINH KHOA 17110163**

**GVHD: TỪ TUYẾT HỒNG**

Hồ Chí Minh, tháng 12 năm 202

**TRANG BÌA**

**LỜI CẢM ƠN**

“Để hoàn thành đồ án này, em xin gửi lời cảm ơn chân thành đến:

Ban giám hiệu trường Đại Học Sư phạm kỹ thuật Thành phố Hồ Chí Minh vì đã tạo điều kiện về cơ sở vật chất với hệ thống thư viện hiện đại, đa dạng các loại sách, tài liệu thuận lợi cho việc tìm kiếm, nghiên cứu thông tin.

Xin cảm ơn giảng viên bộ môn – Cô Từ Tuyết Hồng đã giảng dạy tận tình, chi tiết để em có đủ kiến thức và vận dụng chúng vào đồ án này.

Do chưa có nhiều kinh nghiệm làm để tài cũng như những hạn chế về kiến thức, trong đồ án chắc chắn sẽ không tránh khỏi những thiếu sót. Rất mong nhận được sự nhận xét, ý kiến đóng góp, phê bình từ phía cô để đồ án được hoàn thiện hơn.

Lời cuối cùng, em xin kính chúc cô nhiều sức khỏe, thành công và hạnh phúc.”

TP.HCM, ngày 15 tháng 12 năm 2020

Sinh viên thực hiện

**NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................ ................................................................................................................................................................................................................................................................................................................................................................................................................................................

TP.HCM, ngày 15 tháng 12 năm 2020

Giáo viên hướng dẫn

**LỜI MỞ ĐẦU**

Cùng với sự phát triển của khoa học kĩ thuật, công nghệ thông tin nói chung và bộ môn học máy nói riêng ngày càng được ứng dụng rộng rãi trong nhiều lĩnh vực. Với nhu cầu tiền xử lý dữ liệu đầu vào để phân chia dữ liệu một cách chính xác và hiệu quả, thông thường trước đây con người sẽ phải thực hiện việc đó một cách thủ công. Nhưng với sự ra đời của giải thuật gom cụm dữ liệu (clustering) đã giúp ta giải quyết những phần nặng nhọc, phức tạp trong công việc đó. Từ đó mở ra giải pháp cho các bài toán có độ phức tạp cao như: viết chương trình đề xuất, tìm kiếm thông tin, v.v ..

Mục tiêu của đề tài: Tìm hiểu về giải thuật gom cụm dữ liệu (clustering) và viết chương trình ứng dụng đáp ứng nhu cầu, giải quyết các vấn đề trong thực tế

Nội dung của báo cáo gồm 3 phần:

[1. Giới thiệu thuật toán 8](#_Toc59047750)

[2. Phân tích chi tiết 9](#_Toc59047751)

[3. Chương trình ứng dụng 16](#_Toc59047752)

**DANH MỤC HÌNH ẢNH**

[Figure 1: Minh họa phân cụm 10](#_Toc59050318)

[Figure 2: Bước 1 11](#_Toc59050319)

[Figure 3: Bước 2 12](#_Toc59050320)

[Figure 4: Bước 3 13](#_Toc59050321)

[Figure 5: Bước 4 14](#_Toc59050322)

[Figure 6: Bước 5 16](#_Toc59050323)

[Figure 7: Bước 6 17](#_Toc59050324)

[Figure 8: Bước 7 18](#_Toc59050325)

[Figure 9: Khai báo thư viện 20](#_Toc59050326)

[Figure 10: Đọc dữ liệu và xuất 10 dòng đầu 20](#_Toc59050327)

[Figure 11: The Elbow Method 21](#_Toc59050328)

[Figure 12: Kết quả phân cụm 22](#_Toc59050329)

[Figure 13: Kết quả với 4 cụm 23](#_Toc59050330)

[Figure 14: Kết quả với 6 cụm 24](#_Toc59050331)

**PHẦN VIỆC CỦA TỪNG THÀNH VIÊN**

|  |  |
| --- | --- |
| **Nội dung công việc** | **Chịu trách nhiệm** |
| **1. Tìm hiểu giải thuật gom cụm dữ liệu** | Lư Mạnh Quân, Trương Minh Khoa |
| **2. Giới thiệu thuật toán** | Trương Minh Khoa |
| **3. Phân tích chi tiết** | Trương Minh Khoa |
| **4. Chương trình ứng dụng** | Lư Mạnh Quân |
| **5. Trình bày báo cáo** | Lư Manh Quân |

# 1. Giới thiệu thuật toán

Thuật toán K-means clustering (phân cụm K-means) là một trong những thuật toán cơ bản nhất trong Unsupervised learning. Trong thuật toán K-means clustering, chúng ta không biết nhãn (label) của từng điểm dữ liệu. Mục đích là làm thể nào để phân dữ liệu thành các cụm (cluster) khác nhau sao cho dữ liệu trong cùng một cụm có tính chất giống nhau.

Ví dụ: Một công ty muốn tạo ra những chính sách ưu đãi cho những nhóm khách hàng khác nhau dựa trên sự tương tác giữa mỗi khách hàng với công ty đó (số tiền khách hàng đã chi trả cho công ty; độ tuổi; giới tính; thu nhập; …). Giả sử công ty đó có rất nhiều dữ liệu của rất nhiều khách hàng nhưng chưa có cách nào chia toàn bộ khách hàng đó thành một số nhóm/cụm khác nhau. Nếu một người biết Machine Learning được đặt câu hỏi này, phương pháp đầu tiên ta nghĩ đến sẽ là K-means Clustering. Sau khi đã phân ra được từng nhóm, nhân viên công ty đó có thể lựa chọn ra một vài khách hàng trong mỗi nhóm để quyết định xem mỗi nhóm tương ứng với nhóm khách hàng nào. Phần việc cuối cùng này cần sự can thiệp của con người, nhưng lượng công việc đã được rút gọn đi rất nhiều.

Ý tưởng đơn giản nhất về cluster (cụm) là tập hợp các điểm ở gần nhau trong một không gian nào đó (không gian này có thể có rất nhiều chiều trong trường hợp thông tin về một điểm dữ liệu là rất lớn). Hình bên dưới là một ví dụ về 5 cụm dữ liệu:

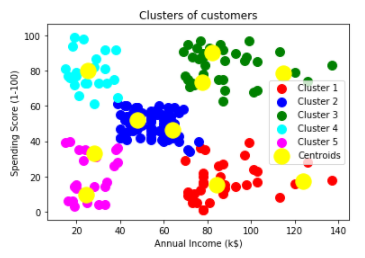


Figure 1: Minh họa phân cụm

Giả sử mỗi cluster có một điểm đại diện (center) màu vàng. Và những điểm xung quanh mỗi center thuộc vào cùng nhóm với center đó. Một cách đơn giản nhất, xét một điểm bất kỳ, ta xét xem điểm đó gần với center nào nhất thì nó thuộc về cùng nhóm với center đó. Tới đây, chúng ta có một bài toán thú vị: Trên một vùng biển hình vuông lớn có năm đảo màu xanh dương, xanh lá, đỏ, hồng, xanh da trời như hình trên.

# 2. Phân tích chi tiết

Đầu tiên là chuẩn bị dữ liệu cần phân cụm. Tiếp theo quyết định số lượng cụm (cluster) cần phân chia. Ở ví dụ này thử chọn số cluster là 3. Ở đây data được thể hiện dưới dạng các điểm cho dễ quan sát. Cự ly của các dữ liệu được hiểu là độ dài đoạn thẳng nối giữa 2 điểm với nhau.

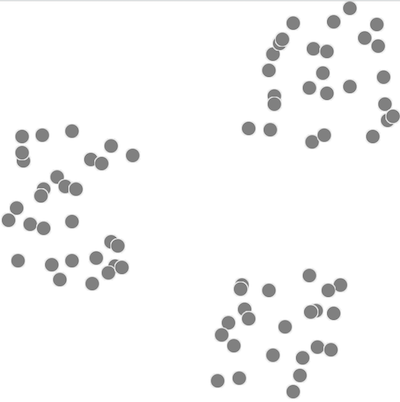


Figure 2: Bước 1

Bước 2：

Chọn ngẫu nhiên 3 điểm làm điểm trung tâm của cluster.

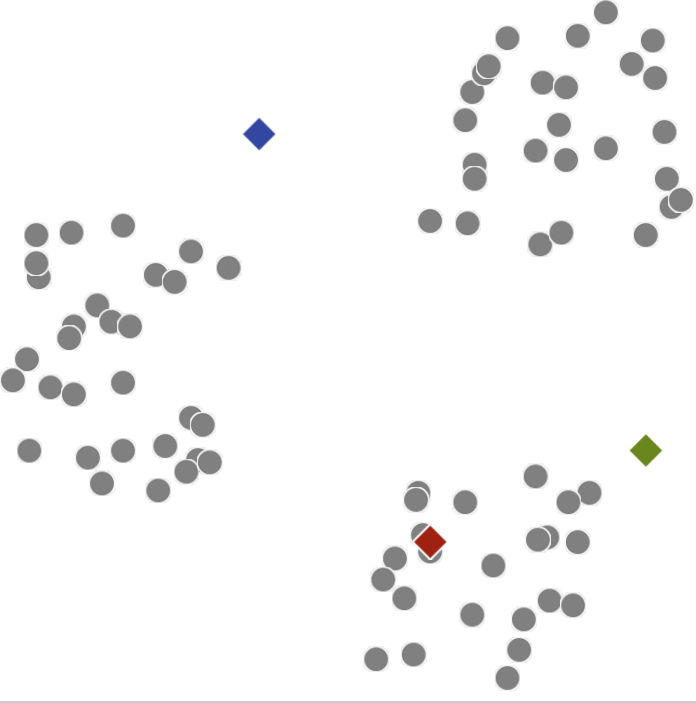


Figure 3: Bước 2

Bước 3:

Với các điểm dữ liệu không được chọn là điểm trung tâm thì tính toán khoảng cách từ chính điểm đó đến các cluster và quyết định cluster nào gần với mình nhất.

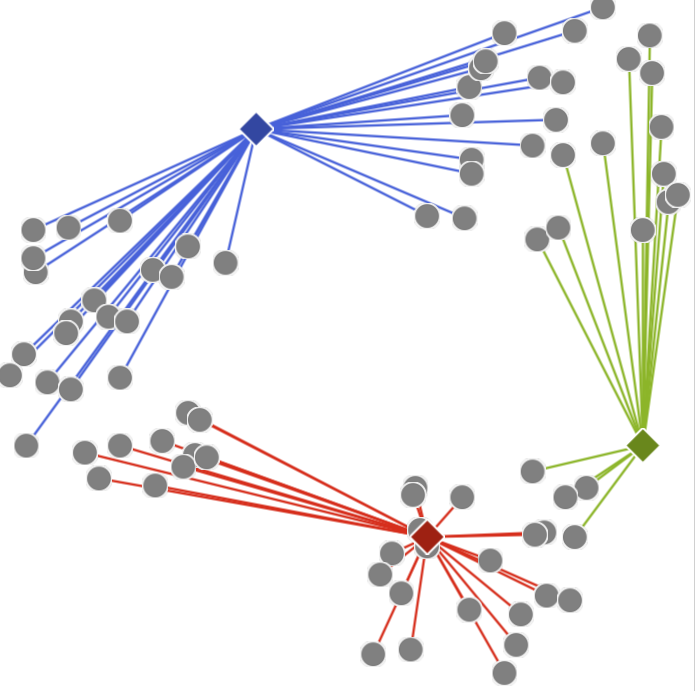


Figure 4: Bước 3

Bước 4:

Từ bước tính toán trên, tiến hành phân loại các điểm về các cluster đã quyết định (cluster gần nó nhất). Vậy là đã phân ra được 3 cụm.

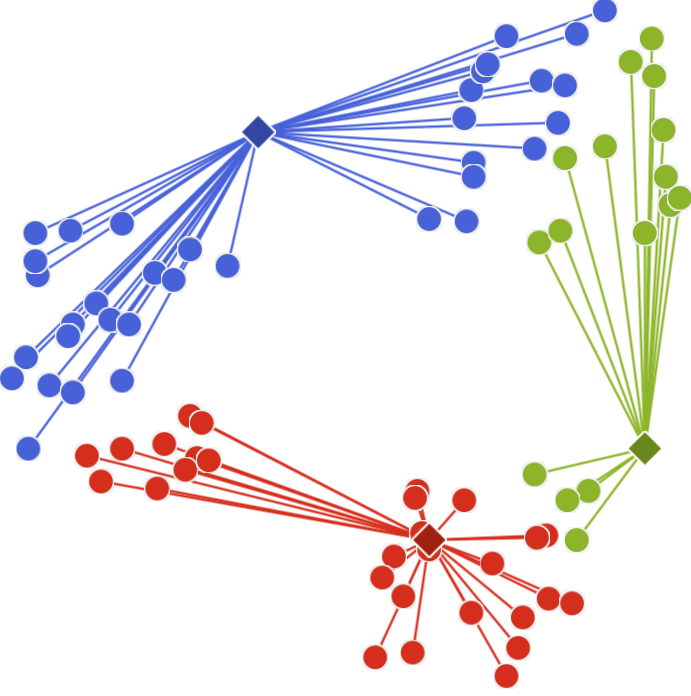


Figure 5: Bước 4

Tuy nhiên nhìn vào hình trên có thể nhận ra ngay các cụm dữ liệu này chưa phải là chuẩn nhất. Chính xác là do điểm trung tâm cụm được chọn chưa chính xác. Nên bước 5 sẽ được tiến hành.

Bước 5:

Bước trên chúng ta đã thu được 3 cụm, bây giờ tiến hành tính trọng tâm của các điểm dữ liệu của từng cụm. Sau đó di chuyển điểm trung tâm của cụm sang vị trí vừa tính được.

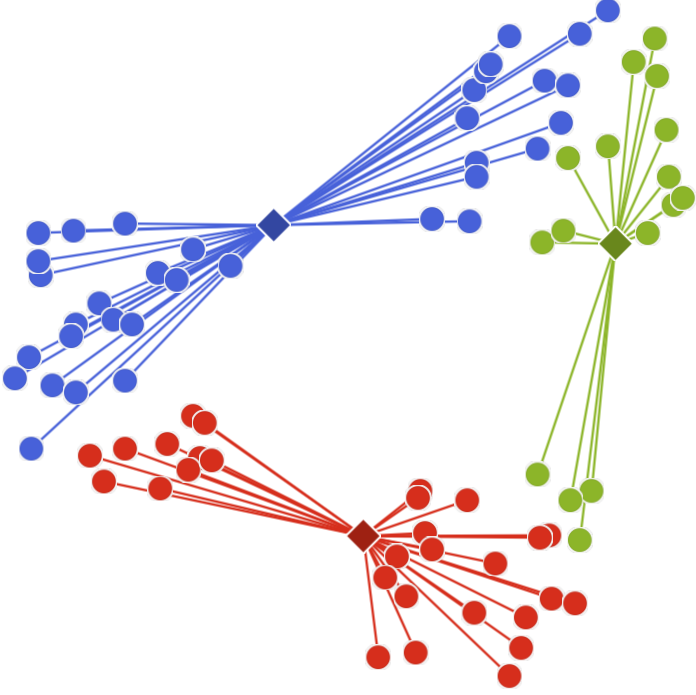


Figure 6: Bước 5

Vị trí mà 3 điểm trung tâm của cluster vừa di chuyển đến được hiểu ngắn gọn chính là điểm trung tâm đang di chuyển đến vị trí chính xác hơn.

Bước 6:

Một lần nữa tiến hành bước 3, tính toán lại khoảng các các điểm đến các điểm trung tâm. Sau đó phân loại lại các điểm dữ liệu về các cụm.

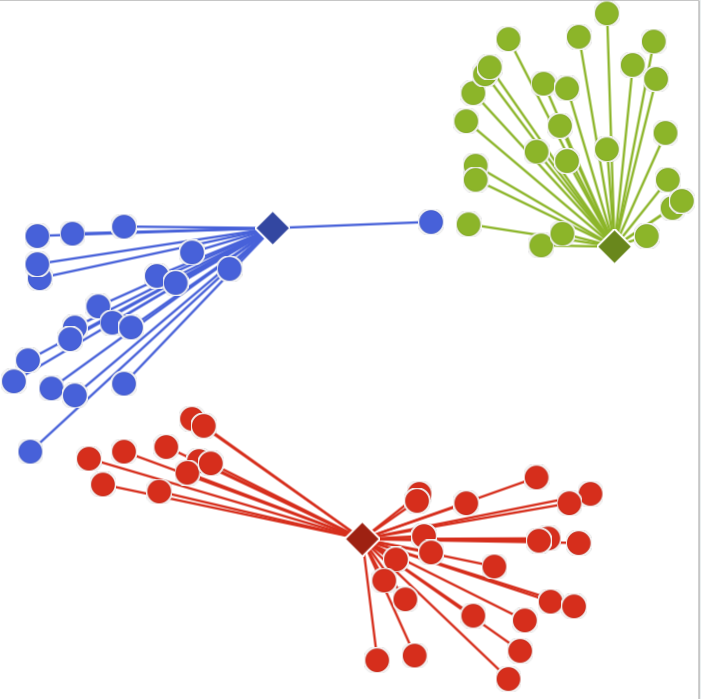


Figure 7: Bước 6

Bước 7:

Sau đó lặp lại quá trình di chuyển cluster trung tâm và phân loại lại các điểm về các cụm gần nhất.

Quá trình này sẽ dừng khi sau khi dữ liệu sau khi phân cụm lại không thay đổi gì so với lần trước.

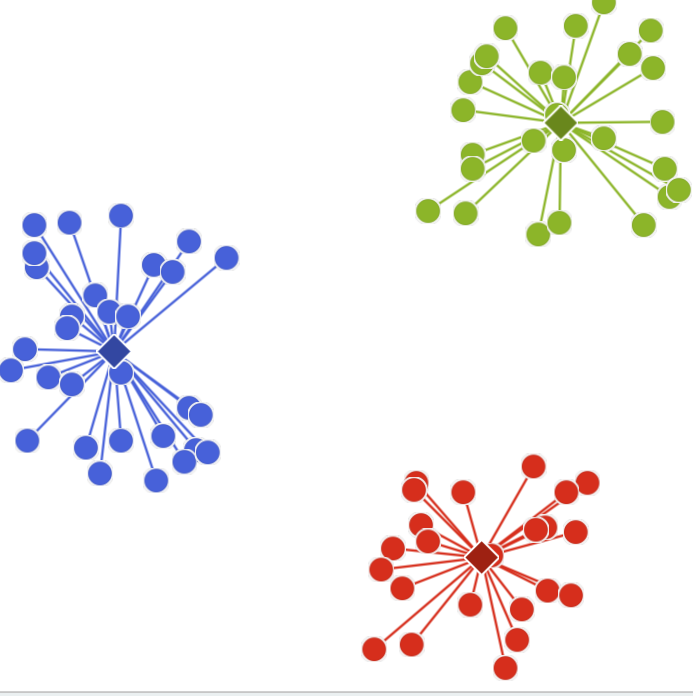


Figure 8: Bước 7

Với ví dụ này thì với với lần lặp thứ 4, thuật toán đã phân cụm thành công.

# 3. Chương trình ứng dụng

\_ Khai báo thư viện sử dụng:

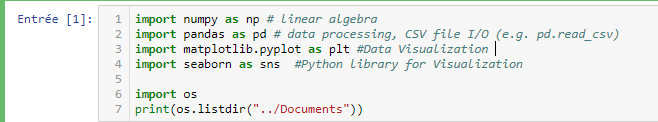


Figure 9: Khai báo thư viện

\_ Đọc file dữ liệu và xuất ra 10 dòng đầu:

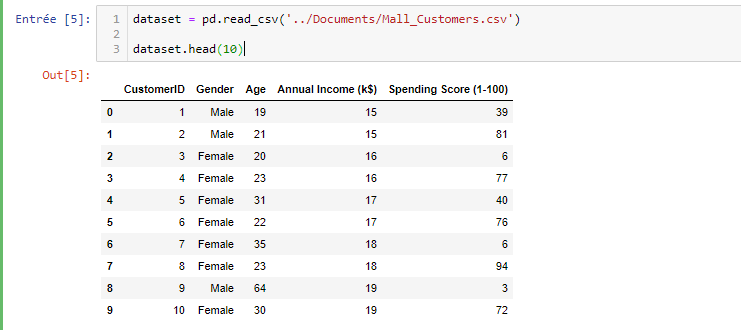


Figure 10: Đọc dữ liệu và xuất 10 dòng đầu

\_ Lấy 2 cột giá trị cần dùng, sử dụng Elbow Method để so sánh tổng khoảng cách giữa các điểm với trung tâm cụm gần nhất ứng với số cụm

\_ Ta thấy ở 5 thì xuất hiện gấp khúc cuối cùng và tổng khoảng cách giảm không đáng kể nên sẽ chia các điểm thành 5 cụm

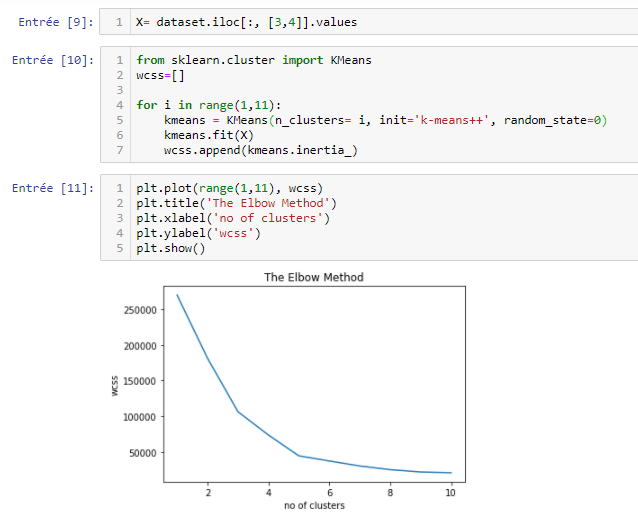


Figure 11: The Elbow Method

\_ Và đây là kết quả phân cụm, ta có thể quan sát được:

* Cụm xanh da trời là những khách hàng có thu nhập thấp nhưng chi tiêu cao
* Cụm hồng là những khách hàng có thu nhập thấp và chi tiêu thấp
* Cụm xanh lá là những khách hàng có thu nhập cao và chi tiêu cao
* Cụm xanh dương là những khách hàng có thu nhập trung bình và chi tiêu trung bình
* Cụm đỏ là những khách hàng có thu nhập cao nhưng chi tiêu thấp

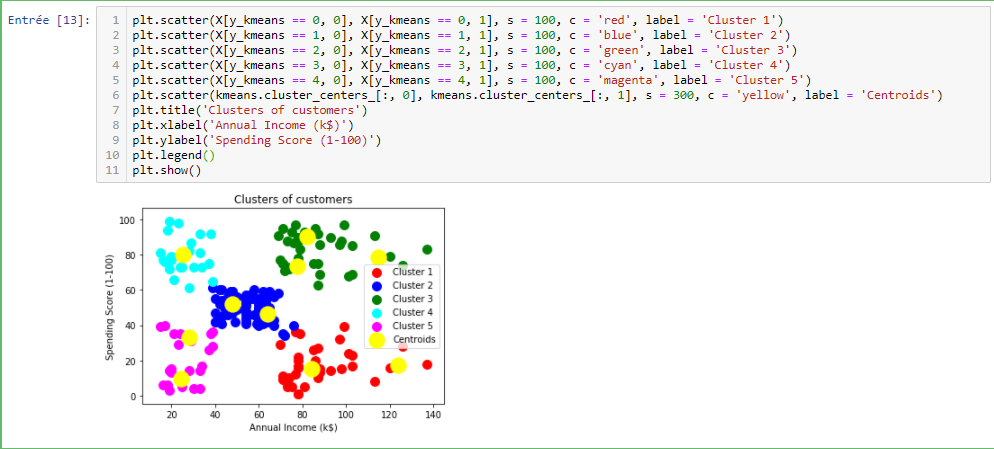


Figure 12: Kết quả phân cụm

* Và từ đây ta có thể đưa ra các chính sách phù hợp với từng nhóm khách hàng, giảm bớt nặng nhọc công việc và đáp ứng nhu cầu thực tế

\_ Nhưng ta cũng cần lưu ý để tránh chọn sai số lượng cụm dẫn đến kết quả không chính xác, hiệu quả:

+ Đây là kết quả nếu ta chọn số lượng cụm là 4, có thể thấy sẽ có 1 phần các điểm trong cụm xanh có giá trị thu nhập, giá trị chi tiêu ở mức trung bình, khác so với phần còn lại trong cụm xanh

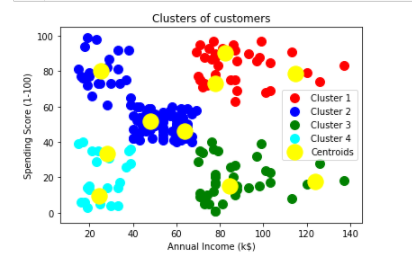


Figure 13: Kết quả với 4 cụm

+ Đây là kết quả nếu ta chọn số lượng cụm là 6, có thể thấy sự tương đồng giữa cụm cam và cụm xanh lá nhưng lại được chia thành 2 cụm riêng biệt dẫn đến sự thiếu chính xác về kết quả

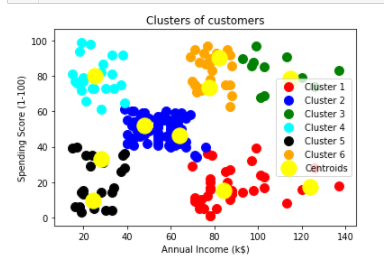


Figure 14: Kết quả với 6 cụm

**KẾT LUẬN**

Sau khi thực hiện đồ án, nhóm em đã nắm được cách giải thuật gom cụm dữ liệu dữ liệu hoạt động, từ đó tiến hành viết chương trình ứng dụng cho giải thuật để giải quyết bài toán thực tế.

Việc lựa chọn số lượng cụm là rất quan trọng, nó ảnh hưởng rất lớn đến kết quả đầu ra. Từ kết quả đó, ta mới có thể có được sự phân chia dữ liệu rõ ràng, dễ dàng áp dụng các phương thức, cách tiếp cận khác nhau cho từng cụm dữ liệu.

Mặc dù chưa thật sự hoàn thành một cách xuất sắc những mục tiêu đã đề ra, nhóm vẫn cảm thấy hài lòng với những công việc và những kinh nghiệm nhóm đạt được trong suốt quá trình thực hiện đồ án. Khoảng thời gian tuy ngắn nhưng đã cho nhóm những trải nghiệm thật sự quý báu, giúp nhóm vững vàng hơn trước con đường học vấn dài ở phía trước

**TÀI LIỆU THAM KHẢO**

<https://www.kaggle.com/vjchoudhary7/kmeans-clustering-in-customer-segmentation?fbclid=IwAR1RWlJrsYs53ktJTyvIQKdYBYNYbX6XQsPTfZpCnID2xQFVaTt_PTqVdnE>

<https://machinelearningcoban.com/2017/01/01/kmeans/#gioi-thieu-bai-toan>

https://codetudau.com/de-dang-hieu-phuong-phap-k-means-qua-hinh-ve/index.html