 TRƯỜNG ĐẠI HỌC

**SƯ PHẠM KỸ THUẬT THÀNH PHỐ HỒ CHÍ MINH**





**KHOA ĐÀO TẠO CHẤT LƯỢNG CAO**

NGÀNH CÔNG NGHỆ THÔNG TIN

**BÁO CÁO ĐỒ ÁN 3**

**TÌM HIỂU VỀ THUẬT TOÁN K - MEANS**

**SVTH: NGUYỄN HUY CƯỜNG 17110107**

**BÙI VĂN HÀ 17110129**

**Khóa: 2017 – 2021**

**` GVHD: TS.Lê Văn Vinh**



TP. Hồ Chí Minh, tháng 01 năm 2021

# **LỜI CẢM ƠN**

Lời đầu tiên, nhóm thực hiện xin được phép gửi lời cảm ơn chân thành đến khoa Đào tạo Chất Lượng Cao – Trường Đại học Sư phạm Kỹ thuật Thành phố Hồ Chí Minh đã tạo mọi điều kiện thuận lợi nhất cho nhóm thực hiện có cơ hội được tự do tiếp cận, tham khảo, mở rộng thêm kiến thức trong lĩnh vực Công nghệ thông tin nói chung và môn Học máy nói riêng.

Lời cảm ơn trân trọng nhất nhóm thực hiện xin chân thành được gửi đến Thầy **Lê Văn Vinh** người đã dùng mọi tâm huyết và tri thức của người Cô, cùng đồng hành và giảng dạy, hướng dẫn và tạo mọi điều kiện thuận lợi giúp đỡ cho nhóm phát huy hết khả năng cũng như nâng cao kiến thức trong suốt quá trình học tập, đặc biệt là trong quá trình chuẩn bị và thực hiện báo cáo môn học Đồ Án 3. Cảm ơn sự nhiệt tình của Thầy, là động lực vô cùng to lớn giúp nhóm thực hiện báo cáo đồ án có thể thực hiện, kiên trì trong suốt quá trình thực hiện báo cáo và khám phá ra những kiến thức mới đầy thú vị và bổ ích liên quan đến báo cáo mà cụ thể là báo cáo đồ án 3 về chủ đề **Tìm hiểu về thuật toán K-Means.**

Cuối cùng nhóm thực hiện kính chúc thầy dồi dào sức khỏe, niềm tin để tiếp tực sự nghiệp cáo quý của mình là truyền đạt kiến thức cho thế hệ mai sau.

Nhóm xin chân thành cám ơn.

*TP.HCM*, ngày tháng 01 năm 2021

Sinh viên thực hiện

# **LỜI MỞ ĐẦU**

* Trong việc khai phá dữ liệu hiện nay, việc dữ liệu ngày càng trở nên dần một lớn hơn đã và đang là một nỗi khó khăn đối với những người làm công nghệ số nói chung, cũng như là ngành dữ liệu nói riêng. Dữ liệu lớn quá mức sẽ rất khó khăn trong việc kiểm soát và phân loại dữ liệu, khiến dữ liệu của chúng ta rất dễ bị trùng lặp và không được minh bạch. Chính vì lẽ đó, cùng với sự phát triển mạnh mẽ của nền tảng công nghệ AI, Machine learning, chúng ta đã cho ra đới những phương pháp có thể giải quyết được bài toán nan giải trên. Đó là thuật toán K-Means, thuộc về nhóm thuật toán gom cụm Clustering.

# **MỤC LỤC**

[**LỜI CẢM ƠN** 1](#_Toc61810305)

[**LỜI MỞ ĐẦU** 2](#_Toc61810306)

[**MỤC LỤC** 3](#_Toc61810307)

[**DANH MỤC TỪ, TỪ VIẾT TẮT** 5](#_Toc61810308)

[**DANH MỤC BẢNG BIỂU** 6](#_Toc61810309)

[**DANH MỤC HÌNH ẢNH** 7](#_Toc61810310)

[**CHƯƠNG 1: TỔNG QUAN VỀ BÁO CÁO ĐỒ ÁN** 9](#_Toc61810311)

[1.1. Giới thiệu chung về đề tài 9](#_Toc61810312)

[1.2. Lý do chọn đề tài 9](#_Toc61810313)

[1.3. Mục tiêu nghiên cứu 9](#_Toc61810314)

[1.4. Công nghệ sử dụng 9](#_Toc61810315)

[**CHƯƠNG 2: CƠ SỞ LÝ THUYẾT** 10](#_Toc61810316)

[2.1. Clustering – Thuật toán gom cụm 10](#_Toc61810317)

[2.1.1. Định nghĩa 10](#_Toc61810318)

[2.1.2. Giới thiệu chung 10](#_Toc61810319)

[2.1.3. Chức năng 11](#_Toc61810320)

[2.1.4. Các yêu cầu cần có khi sử dụng thuật toán gom cụm 11](#_Toc61810321)

[2.1.5. Các thuật toán sử dụng gom cụm Clustering: 11](#_Toc61810322)

[2.1.6. Các ứng dụng sử dụng Clustering 13](#_Toc61810323)

[2.1.7. Ưu nhược điểm 14](#_Toc61810324)

[2.2. Thuật toán K – Means 16](#_Toc61810325)

[2.2.1. Giới thiệu chung 16](#_Toc61810326)

[2.2.2. Định nghĩa 16](#_Toc61810327)

[2.2.3. Chức năng và cách thức hoạt động 16](#_Toc61810328)

[2.2.4. Phân tích toán học 17](#_Toc61810329)

[2.2.5. Ưu nhược điểm của thuật toán K-Means 22](#_Toc61810330)

[**CHƯƠNG 3: BÀI TOÁN MINH HỌA** 24](#_Toc61810331)

[3.1. Bài toán 1: Phân chia phân khúc khách hàng trong siêu thị 24](#_Toc61810332)

[3.1.1. Đặc tả bài toán 24](#_Toc61810333)

[3.1.2. Mô tả chức năng 24](#_Toc61810334)

[3.1.3. Xử lý bài toán 24](#_Toc61810335)

[3.2. Bài toán 2: Nén dung lượng hình ảnh 30](#_Toc61810336)

[3.2.1. Đặc tả bài toán 30](#_Toc61810337)

[3.2.2. Mô tả chức năng 30](#_Toc61810338)

[3.2.3. Xử lý bài toán 30](#_Toc61810339)

[**PHẦN TỔNG KẾT** 43](#_Toc61810340)

[1. Kết quả đạt được 43](#_Toc61810341)

[2. Khó khăn 43](#_Toc61810342)

[3. Bài học kinh nghiệm 43](#_Toc61810343)

[**TÀI LIỆU THAM KHẢO** 44](#_Toc61810344)

# **DANH MỤC TỪ, TỪ VIẾT TẮT**

*Supervising Learning*: Học có giám sát

*Unsupervising Learning*: Học không giám sát

***Segmention analysis*: Phân cụm phân khúc**

***Classify*: Phân loại**

***Estimate*: Ước lượng**

***Predict*: Dự báo**

***Linear Regression*: Hồi quy tuyến tính**

***Logistic Regression*: Hồi quy Logic**

# **DANH MỤC BẢNG BIỂU**

[**Bảng 1:** Bảng so sánh các kỹ thuật gom cụm 13](#_Toc61080612)

# **DANH MỤC HÌNH ẢNH**

[**Hình 1:** Khai báo thư viện cho bài toán 1 24](#_Toc61080613)

[**Hình 2:** Đọc file dataset đầu vào cho bài toán 1 25](#_Toc61080614)

[**Hình 3:** In ra số dòng và số cột của dataset 25](#_Toc61080615)

[**Hình 4**: Kiểm tra dữ liệu bị thiếu 26](#_Toc61080616)

[**Hình 5:** In ra số dòng bị thiếu dữ liệu 26](#_Toc61080617)

[**Hình 6:** Tìm số cụm tối ưu cho bài toán 1 27](#_Toc61080618)

[**Hình 7:** Sơ đồ gấp khúc Elbow 27](#_Toc61080619)

[**Hình 8:** Xây dựng mô hình gom cụm 28](#_Toc61080620)

[**Hình 9:** Xây dựng sơ đồ gom cụm 28](#_Toc61080621)

[**Hình 10:** Biểu đồ gom cụm dữ liệu khách hàng 29](#_Toc61080622)

[**Hình 11:** Diễn giải về biểu đồ gom cụm 29](#_Toc61080623)

[**Hình 12:** Khai báo thư viện cho bài toán 2 30](#_Toc61080624)

[**Hình 13:** Phân tích màu sắc 31](#_Toc61080625)

[**Hình 14:** Đọc hình ảnh đầu vào 31](#_Toc61080626)

[**Hình 15:** Xem thống số của ảnh gốc 32](#_Toc61080627)

[**Hình 16:** Dung lượng ban đầu của ảnh 32](#_Toc61080628)

[**Hình 17:** Số màu gốc của ảnh 32](#_Toc61080629)

[**Hình 18:** Phương sai của ảnh 32](#_Toc61080630)

[**Hình 19:** Biểu diển các RGB pixel trên mô hình 3D 33](#_Toc61080631)

[**Hình 20:** Tìm giá trị RGB cho cụm 2 màu 33](#_Toc61080632)

[**Hình 21:** Tìm màu sắc cho các giá trị RGB thu được 34](#_Toc61080633)

[**Hình 22:** Xuất ra hình ảnh với chỉ 2 màu 35](#_Toc61080634)

[**Hình 23:** Tính WCSS, BCSS, Explained Variance và Image size 36](#_Toc61080635)

[**Hình 24:** Dùng K-Means gom cụm các hình ảnh theo vòng lặp số cụm k cho trước 36](#_Toc61080636)

[**Hình 25:** Các thông tin hình ảnh thu thập được 37](#_Toc61080637)

[**Hình 26:** Xuất ra các hình có số màu tăng dần 37](#_Toc61080638)

[**Hình 27:** Bảng các hình ảnh với số màu tăng dần 38](#_Toc61080639)

[**Hình 28:** Vẽ đồ thị gấp khúc cho các giá trị 39](#_Toc61080640)

[**Hình 29**: Đồ thị gấp khúc 39](#_Toc61080641)

[**Hình 30:** Khai báo phương thức toán học cho thuật toán k tối ưu 40](#_Toc61080642)

[**Hình 31**: Tìm ra k tối ưu 41](#_Toc61080643)

[**Hình 32:** So sánh thông tin giữa ảnh nén tối ưu với ảnh gốc 41](#_Toc61080644)

[**Hình 33:** Kết quả của bài toán 2 42](#_Toc61080645)

# **CHƯƠNG 1: TỔNG QUAN VỀ BÁO CÁO ĐỒ ÁN**

## Giới thiệu chung về đề tài

* Báo cáo đồ án mà nhóm thực hiện xây dựng là tìm hiểu một cách rõ ràng những nội dung cơ bản vê thuật toán K-Means qua những định nghĩa,đặc điểm tính chất của nó. Từ đó xây dựng cho mình một ví dụ đơn giản, dễ hiểu áp dụng thuật toán K-Means.

## Lý do chọn đề tài

* Nhằm giúp cho việc nghiên cứu về một thuật toán gom cụm được dễ dàng và nắm được kiến thức tốt hơn.
* Tăng cường khả năng lập trình, làm việc nhóm.
* Tiếp thu kiếm các kiến thức về Python nói chung và Machine Learning nói riêng.

## Mục tiêu nghiên cứu

* Hiểu rõ được bản chất, đặc điểm, công thức và cách thức vận hành của thuật toán K-Means trong việc gom cụm và xử lý dữ liệu trong thực tế.

## Công nghệ sử dụng

* Công nghệ sử dụng cho Ví dụ Demo: Python trên ứng dụng Jupyter Notebook

# **CHƯƠNG 2: CƠ SỞ LÝ THUYẾT**

## Clustering – Thuật toán gom cụm

### Định nghĩa

* ‘Clustering’ là cách nhóm 1 tập hợp các đối tượng (objects) dựa trên các đặc điểm, thuộc tính của , tổng hợp lại theo từng nhóm dựa trên độ tương đồng (similarities).
* Liên quan đến da mining, clustering phân chia da bằng 1 thuật toán join được chỉ định, thích hợp nhất cho các phân tích thông tin mong muốn.
* Theo cách hiểu đơn giản nhất, Clustering là phương pháp phân tích qua đó tập dữ liệu sẽ được phân thành nhiều cụm hoặc nhóm khác nhau, trong mỗi cụm/ nhóm các điểm dữ liệu hay các quan sát sẽ giống nhau, và giữa các cụm hoặc nhóm có sự khác biệt.

### Giới thiệu chung

* Thuật toán Clustering có tên gọi khác là **segmention analysis,**phân tích phân khúc, vì thuật toán này được ứng dụng khá nhiều trong marketing, sales, và CRM với nhiệm vụ xác định các phân khúc khách hàng để đưa ra các chiến dịch quảng cáo, bán hàng nhắm mục tiêu hiệu quả.
* Clustering còn được gọi là **unsupervised classification**(phân loại không giám sát) là phương pháp trong **unsupervised learning**(học không giám sát) – phương pháp xây dựng các model phân tích – dựa trên tập dữ liệu “không có nhãn”, các điểm dữ liệu chưa được phân loại – mục đích tìm hiểu và trích xuất được những thông tin giá trị về đặc điểm, tính chất của những quan sát bên trong.
* Khác với **supervised learning**(học có giám sát), Clustering không cố gắng phân loại (classify), không cố gắng ước lượng (estimate), hay dự báo (predict) giá trị của biến mục tiêu.
* Cơ sở để phương pháp Clustering có thể được triển khai và giúp phân cụm tối ưu chính là dựa vào các thuộc tính, các biến,… nằm trong tập dữ liệu, đặc điểm và tính chất của dữ liệu.

### Chức năng

* Như đã biết, dữ liệu của  có thể là bất cứ thứ gì, chẳng hạn như dữ liệu về khách hàng: Thuật toán phân cụm sẽ rất hữu ích trong việc đánh giá và chia thành các nhóm người dùng khác nhau, rồi từ đó có thể đưa ra những chiến lược phù hợp trên từng nhóm người dùng đó.

### Các yêu cầu cần có khi sử dụng thuật toán gom cụm

* Có khả năng mở rộng.
* Thích nghi với các kiểu dữ liệu khác nhau
* Khám phá ra các cụm với hình thức bất kỳ
* Ít nhạy cảm với thứ tự của dữ liệu vào
* Khả năng giải quyết dữ liệu nhiễu
* Ít nhạy cảm với tham số đầu vào
* Thích nghi với dữ liệu đa chiều.
* Dễ hiểu, dễ cài đặt và có tính khả thi.

### Các thuật toán sử dụng gom cụm Clustering:

* Có bốn thuật toán tiêu biểu cho việc sử dụng thuật toán gom cụm dữ liệu Clustering, bao gồm:
* Hard Clustering:
* K – Means (Phân cụm phân hoạch)
* Hierarchical Clustering (Phân cụm phân cấp)
* DBSCAN (Phân cụ theo mật độ)
* EM (Phân cụm theo mô hình)
* Phân cụm theo lưới
* Soft Clustering:
* Fuzzy Clustering (Phân cụm mờ)

|  |  |
| --- | --- |
| **Kỹ thuật gom cụm** | **Chức năng** |
| K - Means | Ý tưởng chính của kỹ thuật này là phân hoạch một tập hợp dữ liệu có n phần tử cho trước thành k nhóm dữ liệu sao mỗi phần tử dữ liệu chỉ thuộc về một nhóm dữ liệu có tối thiểu ít nhất một phần tử dữ liệu. Số các cụm được thiết lập là các đặc trưng được lựa chọn trước. |
| Hierarchical Clustering | Phương pháp này xây dựng một phân cấp trên cơ sở các đối tượng dữ liệu đang xem xét. Nghĩa là sắp xếp một tập dữ liệu đã cho thành một cấu trúc có dạng hình cây, cây phân cấp này được xây dựng theo kỹ thuật đệ quy. |
| DBSCAN | Phương pháp này nhóm các đối tượng theo hàm mật độ xác định. Mật độ xác định được định nghĩa như là số các đối tượng lân cận của một đối tượng dữ liệu theo một ngưỡng nào đó. |
| EM | Phương pháp này cố gắng khám phá các phép xấp xỉ tốt của các tham số mô hình sao cho khớp với dữ liệu một cách tốt nhất. Chúng có thể sử dụng chiến lược phân cụm phân hoạch hoặc phân cụm phân cấp, dựa trên cấu trúc hoặc mô hình mà chúng giả định về tập dữ liệu và cách chúng hiệu chỉnh các mô hình này để nhận dạng ra các phân hoạch. |
| Phân cụm theo lưới | Đây là phương pháp dựa trên cấu trúc dữ liệu lưới để phân cụm dữ liệu, phương pháp này chủ yếu tập trung áp dụng cho lớp dữ liệu không gian. |
| Fuzzy Clustering | Phân cụm mờ là một hình thức phân cụm trong đó mỗi điểm dữ liệu có thể thuộc về nhiều hơn một cụm. Phân tích cụm hoặc phân cụm liên quan đến việc gán các điểm dữ liệu cho các cụm sao cho các mục trong cùng một cụm càng giống nhau càng tốt, trong khi các mục thuộc các cụm khác nhau càng giống nhau càng tốt |

**Bảng 1:** Bảng so sánh các kỹ thuật gom cụm

### Các ứng dụng sử dụng Clustering

* Tương tự như các thuật toán phân loại là CART, hay Logistic regression, Clustering được ứng dụng phổ biến rất nhiều ở các lĩnh vực khác nhau từ kinh tế, y tế, đến khoa học xã hội.
* Clustering sẽ giúp các công ty tăng sự hiệu quả của các hoạt động marketing và bán hàng đa mục tiêu. Ứng dụng clustering trong market segmention hay customer segmention thực chất đã được nhắc đến từ rất lâu (từ trước những năm 1970) bởi các chuyên gia phân tích (như Martin Christopher với tài liệu “Cluster analysis and market segmention” năm 1969, John Saunders với tài liệu “Cluster analysis for market segmention) nhưng được ứng dụng phổ biến trong 10 năm trở lại đây. Phần lớn nhờ vào sự phát triển của công nghệ kỹ thuật và khoa học dữ liệu.
* Công việc sẽ dễ dàng hơn rất nhiều nếu có clustering. Clustering cho phép phân nhóm các khách hàng trong tập dữ liệu theo nhân khẩu học, hành vi mua sắm,… kết hợp với những hiểu biết, kinh nghiệm về các yếu tố thu hút khách hàng mua hàng, các đặc điểm thể hiện khả năng khách hàng sẽ mua hàng. Từ đó tinh chỉnh những giải pháp bán hàng, marketing phù hợp cho từng nhóm khách hàng tìm được từ Clustering.
* Ưu điểm của việc phân khúc khách hàng trong lĩnh vực kinh doanh nói chung không những giúp các công ty tăng được doanh số, tăng được lợi nhuận mà còn có thể tăng được trải nghiệm khách hàng, thỏa mãn nhu cầu khách hàng một cách cá nhân hóa. Hơn nữa có cơ hội giữ chân khách hàng cao hơn, mối quan hệ với khách hàng sẽ bền vững hơn, giảm tỷ lệ rời dịch vụ khi họ luôn thấy được thứ mình cần mà công ty mang lại.

### Ưu nhược điểm

#### Ưu điểm

* Phân cụm được sử dụng trong nhiều ứng dụng, mang lại nhiều lợi ích, bao gồm:
* **Đối với phân khúc khách hàng**: có thể phân nhóm khách hàng của mình dựa trên việc mua hàng của họ, hoạt động của họ trên trang web của
* **Đối với phân tích dữ liệu**: khi phân tích một tập dữ liệu mới, trước tiên thường hữu ích là phát hiện ra các cụm của các trường hợp tương tự, vì việc phân tích các cụm riêng biệt thường dễ dàng hơn.
* **Là một kỹ thuật giảm kích thước**: một khi tập dữ liệu đã được phân nhóm, thường có thể đo lường mức độ phù hợp của từng cá thể với mỗi cụm
* **Đối với phát hiện bất thường** (còn gọi là phát hiện ngoại lệ): bất kỳ trường hợp nào có ái lực thấp với tất cả các cụm đều có khả năng là bất thường. có thể phát hiện những người dùng có hành vi bất thường, chẳng hạn như số lượng yêu cầu bất thường mỗi giây
* **Đối với việc học bán giám sát (Semi – Supervised Learning)**: nếu chỉ có một vài nhãn, có thể thực hiện việc phân loại và truyền các nhãn cho tất cả các phiên bản trong cùng một cụm.
* **Đối với công cụ tìm kiếm**: một số công cụ tìm kiếm cho phép tìm kiếm những hình ảnh tương tự như một hình ảnh tham chiếu.
* **Để phân đoạn một hình ảnh**: bằng cách nhóm các pixel theo màu của , sau đó thay thế màu của mỗi pixel bằng màu trung bình của cụm đó, có thể giảm đáng kể số lượng các màu khác nhau trong hình ảnh. Kỹ thuật này được sử dụng trong nhiều hệ thống theo dõi và phát hiện đối tượng, vì nó giúp phát hiện đường viền của từng đối tượng dễ dàng hơn

#### Nhược điểm

* Thuật toán phân cụm do sự sát nhập của người sử dụng và khi ứng dụng, chứa các ràng buộc định hướng
* Sự lựa chọn trong thuật toán phân cụm phụ thuộc hoàn toàn vào loại dữ liệu có sẵn và mục đích cụ thể của ứng dụng
* Một số các thuật toán phân cụm tích hợp các ý tưởng của một số phương pháp phân nhóm để đôi khi khó phân loại thuật toán đã cho dưới dạng duy nhất thuộc một loại phương pháp phân cụm.

## Thuật toán K – Means

### Giới thiệu chung

* Sau khi dạo qua những thuật toán Machine Learning được dựa trên thuật toán gom cụm, thấy có một thuận toán phân nhóm khá phổ biến, đó là K-Means: đó là một thuật toán tuy đơn giản nhưng lại khá hiệu quả và được sử dụng rộng khắp.
* Trước khi bắt y vào làm, cần phải xác định sẵn 2 thứ: đó là hàm khoảng cách được sử dụng (ví dụ như khoảng cách Euclid) và số lượng nhóm mong muốn (kí hiệu là k)
* Ý tưởng của thuật toán:
* Khởi tạo K điểm dữ liệu trong bộ dữ liệu và tạm thời coi nó là tâm của các cụm dữ liệu.
* Với mỗi điểm dữ liệu trong bộ dữ liệu, tâm cụm của nó sẽ được xác định là 1 trong K tâm cụm gần nó nhất.
* Sau khi tất cả các điểm dữ liệu đã có tâm, tính toán lại vị trí của tâm cụm để đảm bảo tâm của cụm nằm ở chính giữa cụm.
* Bước 2 và bước 3 sẽ được lặp đi lặp lại cho tới khi vị trí của tâm cụm không thay đổi hoặc tâm của tất cả các điểm dữ liệu không thay đổi.

### Định nghĩa

* Thuật toán phân cụm K-means là một phương pháp được sử dụng trong phân tích tính chất cụm của dữ liệu. Nó đặc biệt được sử dụng nhiều trong khai phá dữ liệu và thống kê. Nó phân vùng dữ liệu thành k cụm khác nhau. Giải thuật này giúp xác định được dữ liệu của nó thực sử thuộc về nhóm nào.

### Chức năng và cách thức hoạt động

* Trong thuật toán K-means clustering, không biết nhãn (label) của từng điểm dữ liệu. Mục đích và chức năng của thuật toán này 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*.*
* Thuật toán bắt đầu với việc chọn ra tâm của từng cụm. có thể đơn giản chọn k điểm ngẫu nhiên trong bộ, hoặc sử dụng một số hướng tiếp cận nào khác, nhưng nhìn chung ngẫu nhiên vẫn là cách tốt nhất. Rồi kế tiếp, luân phiên lặp lại 2 giai đoạn sau:
* **Giai đoạn gán**: gán từng phần tử trong bộ dữ liệu của vào các cụm. Cách thức tiến hành đó là: với mỗi điểm, hãy tính khoảng cách từ điểm đó tới vị trí các tâm, sau cùng: tâm nào gần nhất thì gán vào cụm ứng với cái tâm đó
* **Giai đoạn cập nhật**: duyệt từng cụm, cập nhật lại tọa độ của tâm: Như đã biết, sau giai đoạn 1, đã thu được k cụm ứng với dãy các điểm được gán cho từng cụm. Tọa độ tâm mới của cụm sẽ bằng trung bình cộng tọa độ các điểm trong cụm
* Sau càng nhiều vòng lặp, các tâm càng di chuyển chậm dần, và tổng khoảng cách từ mỗi điểm trong cụm tới tâm cụm lại càng nhỏ đi. Quá trình sẽ kết thúc cho tới khi hàm tổng khoảng cách hội tụ (tức là không có sự thay đổi nào xảy ra ở giai đoạn gán nữa). Lúc này tọa độ tâm vẫn sẽ bằng trung bình cộng các điểm hiện tại trong cụm, hay nói cách khác tâm sẽ không còn di chuyển tiếp nữa.

### Phân tích toán học

* Mục đích cuối cùng của thuật toán phân nhóm này là: từ dữ liệu đầu vào và số lượng nhóm muốn tìm, hãy chỉ ra center của mỗi nhóm và phân các điểm dữ liệu vào các nhóm tương ứng. Giả sử thêm rằng mỗi điểm dữ liệu chỉ thuộc vào đúng một nhóm.

#### Một số ký hiệu toán học

* Giả sử có N điểm dữ liệu là X= và K < N là số cluster muốn phân chia. Cần tìm các center và label của mỗi điểm dữ liệu.
* Với mỗi điểm dữ liệu đặt =[,,…,] là label vector của nó, trong đó nếu được phân vào cluster k thì và ,. Điều này có nghĩa là có đúng một phần tử của vector là bằng 1 (tương ứng với cluster của ), các phần tử còn lại bằng 0. Ví dụ: nếu một điểm dữ liệu có label vector là [1,0,0,…,0] thì nó thuộc vào cluster 1, là [0,1,0,…,0] thì nó thuộc vào cluster 2, …. Cách mã hóa label của dữ liệu như thế này được goi là biểu diễn [*one-hot*](https://en.wikipedia.org/wiki/One-hot).
* Ràng buộc của có thể viết dưới dạng toán học như sau:
* Gọi ràng buộc trên là **(1)**

#### Hàm mất mát và bài toán tối ưu

* Nếu coi center là center (hoặc representive) của mỗi cluster và *ước lượng* tất cả các điểm được phân vào cluster này bởi , thì một điểm dữ liệu được phân vào cluster k sẽ bị sai số là Với muốn sai số này có trị tuyệt đối nhỏ nhất nên sẽ tìm cách để đại lượng sau đây đạt giá trị nhỏ nhất:
* Hơn nữa, vì được phân vào cluster k nên Khi đó, biểu thức bên trên sẽ được viết lại là:
* Sai số cho toàn bộ dữ liệu sẽ là:
* Trong đó lần lượt là các ma trận được tạo bởi label vector của mỗi điểm dữ liệu và center của mỗi cluster. Hàm số mất mát trong bài toán K-means clustering là hàm với ràng buộc như được nêu trong phương trình **(1)**. Tóm lại, cần tối ưu bài toán sau:

**(2)**

*Thỏa mãn điều kiện sau*:

#### Nhắc lại khái niệm argmin

* Đã biết ký hiệu min là *giá trị nhỏ nhất của hàm số*, argmin chính là *giá trị của biến số để hàm số đó đạt giá trị nhỏ nhất đó*. Nếu thì giá trị nhỏ nhất của hàm số này bằng 0, đạt được khi x=1. Trong ví dụ này và . Thêm ví dụ khác, nếu thì nói vì 1 là chỉ số để đạt giá trị nhỏ nhất (bằng 0). Biến số viết bên dưới min là biến số cúng cần tối ưu. Trong các bài toán tối ưu, thường quan tâm tới argmin hơn là min

#### Thuật toán tối ưu hàm mất mát

* Bài toán (2) là một bài toán khó tìm điểm tối ưu vì nó có thêm các điều kiện ràng buộc. Tuy nhiên, trong một số trường hợp vẫn có thể tìm được phương pháp để tìm được nghiệm gần đúng hoặc điểm cực tiểu.
* Một cách đơn giản để giải bài toán (2) là xen kẽ giải Y và M khi biến còn lại được cố định. Đây là một thuật toán lặp, cũng là kỹ thuật phổ biến khi giải bài toán tối ưu, và sẽ lần lượt giải quyết hai bài toán sau đây:
* Cố định **M**, tìm **Y**
* **Giả sử đã tìm được các centers, hãy tìm các label vector để hàm mất mát đạt giá trị nhỏ nhất.** Điều này tương đương với việc tìm cluster cho mỗi điểm dữ liệu.
* Khi các centers là cố định, bài toán tìm label vector cho toàn bộ dữ liệu có thể được chia nhỏ thành bài toán tìm label vector cho từng điểm dữ liệu như sau:

**(3)**

*Thỏa mãn điều kiện:*

* Vì chỉ có một phần tử của label vector  bằng 11 nên bài toán (3)(3) có thể tiếp tục được viết dưới dạng đơn giản hơn:
* Vì  chính là bình phương khoảng cách tính từ điểm  tới center , có thể kết luận rằng **mỗi điểm  thuộc vào cluster có center gần nó nhất**. Từ đó có thể dễ dàng suy ra label vector của từng điểm dữ liệu.
* Cố định **Y**, tìm **M**
* **Giả sử đã tìm được cluster cho từng điểm, hãy tìm center mới cho mỗi cluster để hàm mất mát đạt giá trị nhỏ nhất.**
* Một khi đã xác định được label vector cho từng điểm dữ liệu, bài toán tìm center cho mỗi cluster được rút gọn thành:

* Tới đây, có thể tìm nghiệm bằng phương pháp giải đạo hàm bằng 0, vì hàm cần tối ưu là một hàm liên tục và có đạo hàm xác định tại mọi điểm. Và quan trọng hơn, hàm này là hàm convex (lồi) theo nên sẽ tìm được giá trị nhỏ nhất và điểm tối ưu tương ứng.
* Đặt là hàm bên trong dấu , được đạo hàm:
* Giải phương trình có đạo hàm bằng 0, được:
* Mẫu số chính là phép đếm số lượng các điểm dữ liệu trong cluster j . Còn tử số chính là tổng các điểm dữ liệu trong cluster j. Hay nói một cách đơn giản hơn nhiều: **là trung bình cộng của các điểm trong cluster j.**

#### Tóm tắt lại thuật toán như sau

* **Đầu vào:** Dữ liệu X và số lượng cluster cần tìm K.
* **Đầu ra:** Các center M và label vector cho từng điểm dữ liệu Y
* Chọn K điểm bất kỳ làm các center ban đầu.
* Phân mỗi điểm dữ liệu vào cluster có center gần nó nhất.
* Nếu việc gán dữ liệu vào từng cluster ở bước 2 không thay đổi so với vòng lặp trước nó thì dừng thuật toán.
* Cập nhật center cho từng cluster bằng cách lấy trung bình cộng của tất các các điểm dữ liệu đã được gán vào cluster đó sau bước 2.
* Quay lại bước 2.
* Có thể đảm bảo rằng thuật toán sẽ dừng lại sau một số hữu hạn vòng lặp.

### Ưu nhược điểm của thuật toán K-Means

#### Ưu điểm

* Với thuật toán K-Means, có thể giúp được trong việc xử lý một số bài toán thực tế
* Tương đối đơn giản để thực hiện.
* Chia tỷ lệ cho các tập dữ liệu lớn.
* Sự hội tụ của các nhóm được đảm bảo.
* Có thể khởi động được vị trí của các trung tâm.
* Dễ dàng thích ứng với các bài toán mới.
* Tổng quát hóa thành các cụm có hình dạng khác nhau và kích thước, chẳng hạn như các cụm hình elip.

#### Nhược điểm

* Có một vài hạn chế của thuật toán phân cụm K-means:
* Cần biết số lượng cluster cần clustering
* Nghiệm cuối cùng phụ thuộc vào các centers được khởi tạo ban đầu
* Các cluster cần có só lượng điểm gần bằng nhau
* Các cluster cần có dạng hình tròn
* Khi một cluster nằm phía trong 1 cluster khác, K-Means sẽ khó cho ra được kết quả có độ chính xác cao

# **CHƯƠNG 3: BÀI TOÁN MINH HỌA**

## Bài toán 1: Phân chia phân khúc khách hàng trong siêu thị

### Đặc tả bài toán

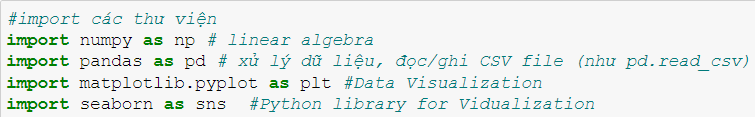
* Một siêu thị muốn phân tích dữ liệu của một nhóm khách hàng theo các tiêu chí khác nhau dựa vào thông tin dữ liệu của họ bao gồm: Giới tính, Tuổi tác, Thu nhập trung bình và Độ chi tiêu

### Mô tả chức năng

* Sử dụng thuật toán K-Means, tạo một lượng cụm nhất định, phân tích dữ liệu từ file đầu vào và tiến hành phân cụm thành các nhóm khách hàng khác nhau, mỗi nhóm sẽ chứa dữ liệu của những khách hàng cùng thỏa mãn một tiêu chí nào đó.

### Xử lý bài toán

#### Khai báo thư viện



**Hình 1:** Khai báo thư viện cho bài toán 1

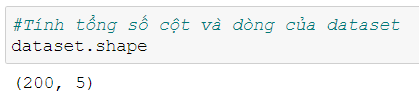
#### Đọc dữ liệu đầu vào

* Đọc file đầu vào và in lên màn hình 10 dòng đầu tiên của Dataset



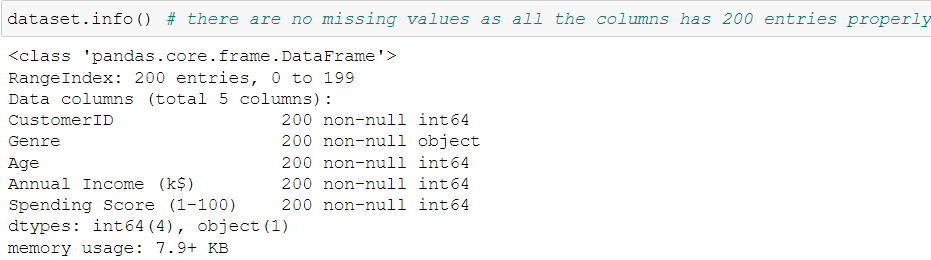
**Hình 2:** Đọc file dataset đầu vào cho bài toán 1

* Tính tổng số dòng và số cột chứa dữ liệu của Dataset



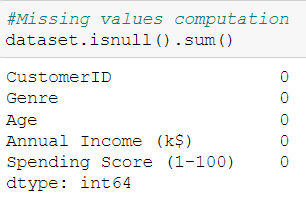
**Hình 3:** In ra số dòng và số cột của dataset

* Kiểm tra xem có dòng nào thiếu dữ liệu hay không, kết quả cho ra là không có dòng nào bị thiếu dữ liệu



**Hình 4**: Kiểm tra dữ liệu bị thiếu

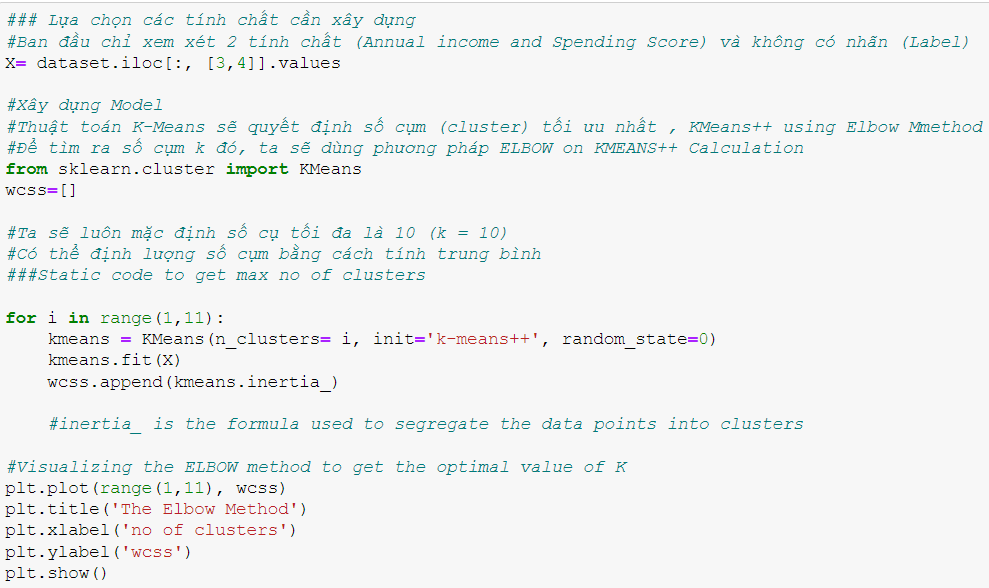
* Tính số dòng bị thiếu dữ liệu: vì không có dòng bị thiếu, nên tất cả sẽ trả về 0



**Hình 5:** In ra số dòng bị thiếu dữ liệu

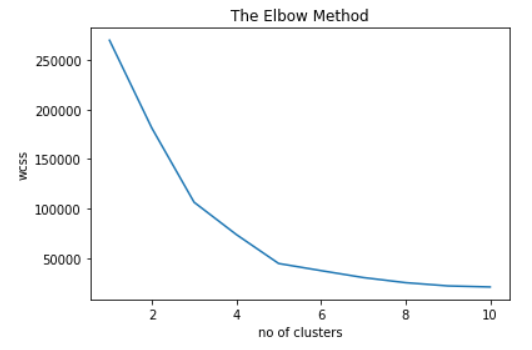
#### Phân tích dữ liệu

* Khai báo các giá trị tham số cho Model bằng cách sử dụng thuật toán K-Means tìm ra số cụm k tối ưu nhất thông qua phương thức biểu đồ gấp khúc (Elbow method)



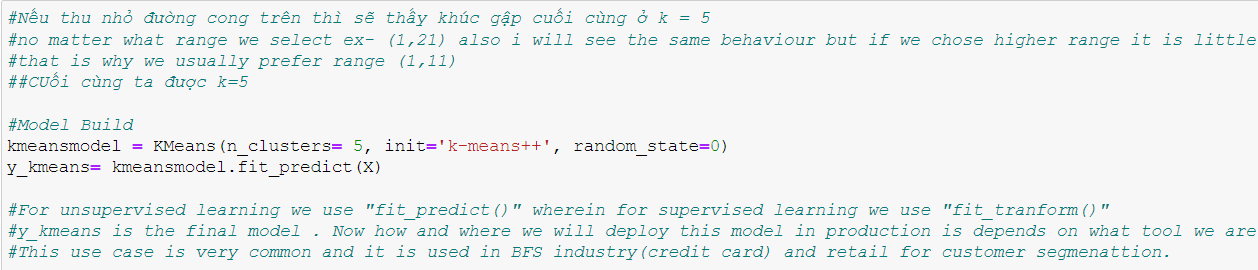
**Hình 6:** Tìm số cụm tối ưu cho bài toán 1

* Kết quả ta được biểu đồ như sau:

****

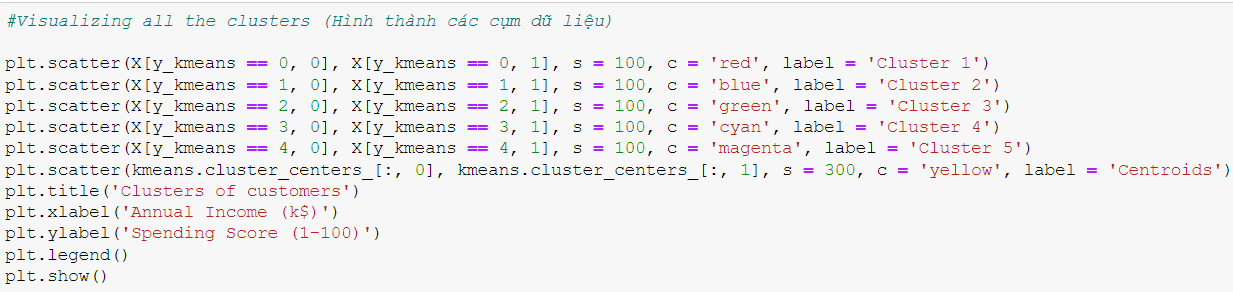
**Hình 7:** Sơ đồ gấp khúc Elbow

* Dựa vào biểu đồ, ta thấy điểm gấp khúc sau cùng nằm ở vị trí k = 5 🡪 Do đó, ta có được số lượng cụm (Cluster) tối ưu nhất là k=5.



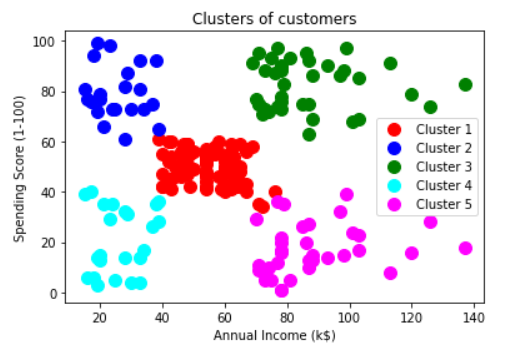
**Hình 8:** Xây dựng mô hình gom cụm

* Xây dựng biểu đồ gom cụm cho 5 cụm đã được phân tích



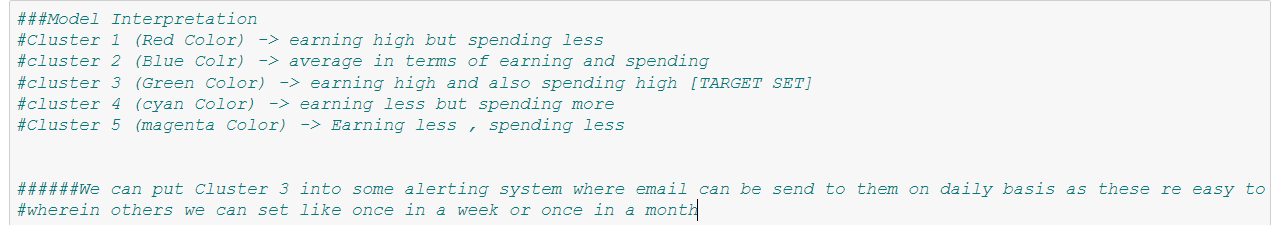
**Hình 9:** Xây dựng sơ đồ gom cụm

* Kết quả cuối cùng ta được một biểu đồ gom cụm với các cụm đã được phân nhóm theo cùng một tính chất



**Hình 10:** Biểu đồ gom cụm dữ liệu khách hàng

* Diễn giải về biểu đồ gom cụm như sau:



**Hình 11:** Diễn giải về biểu đồ gom cụm

* Cụm 1 (Màu đỏ): Có thu nhập cao nhưng mức chi tiêu ít
* Cụm 2 (Màu xanh dương): Có trung bình giữa thu nhập và mức chi tiêu là tương đồng, cân đối với nhau
* Cụm 3 (Màu xanh lục): Có thu nhập và mức chi tiêu đều cao
* Cụm 4 (Màu xanh da trời): Có thu nhập thấp nhưng mức chi tiêu lại cao
* Cụm 5 (Màu hồng): Có thu nhập và mức chi tiêu đều thấp

## Bài toán 2: Nén dung lượng hình ảnh

### Đặc tả bài toán

* Chúng ta có một file ảnh có kích cỡ là 220x220, với khoảng vài chục ngàn màu, kích cỡ khoảng vài trăm kb. Liệu chúng ta có thể giảm bớt dung lượng màu bằng các thay đổi chất lượng cũng như là màu sắc của file ảnh này hay không?

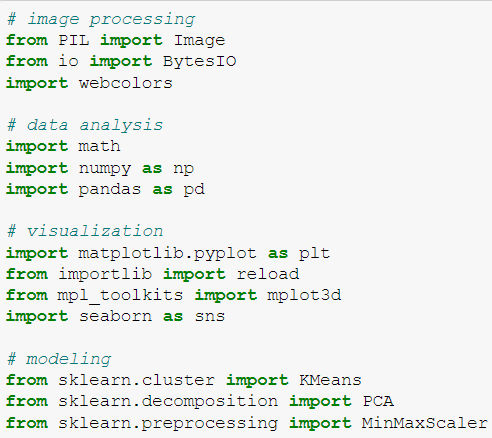
### Mô tả chức năng

* Thật vậy, với thuật toán K-Means, chúng ta hoàn toàn có thể làm được điều đó. Thuật toán K-Means sẽ cho phép chúng ta giảm được một lượng lớn kích cỡ bức ảnh (nén ảnh) để có thể lưu trữ một cách thuận tiện hơn bằng cách gom cụm các tinh thể màu sắc thành các nhóm, sau đó chỉ giử lại một số điểm màu RGB nhất định.

### Xử lý bài toán

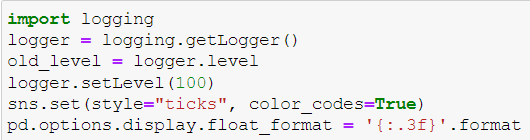
#### Khai báo thư viện

* Import các thư viện cần sử dụng



**Hình 12:** Khai báo thư viện cho bài toán 2

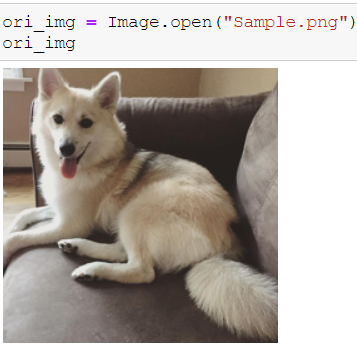
* Ngăn chặn cắt xén ảnh, chuyển đổi định dạng màu sắc của ảnh



**Hình 13:** Phân tích màu sắc

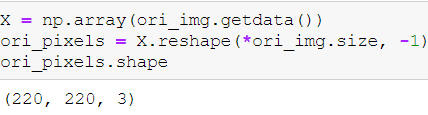
#### Đọc dữ liệu đầu vào

* Đọc file ảnh đầu vào và in ra màn hình



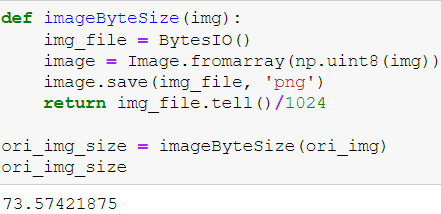
**Hình 14:** Đọc hình ảnh đầu vào

* Xem thông só ảnh (Chiều rộng, chiều cao, số đếm chỉ định RGB)



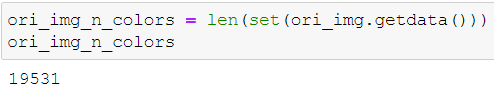
**Hình 15:** Xem thống số của ảnh gốc

* Tính dung lượng ban đầu của ảnh



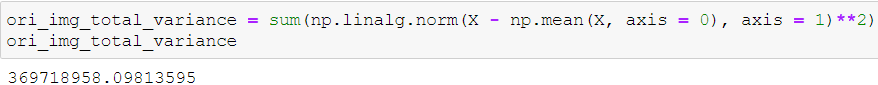
**Hình 16:** Dung lượng ban đầu của ảnh

* Tính số màu ban đầu của ảnh



**Hình 17:** Số màu gốc của ảnh

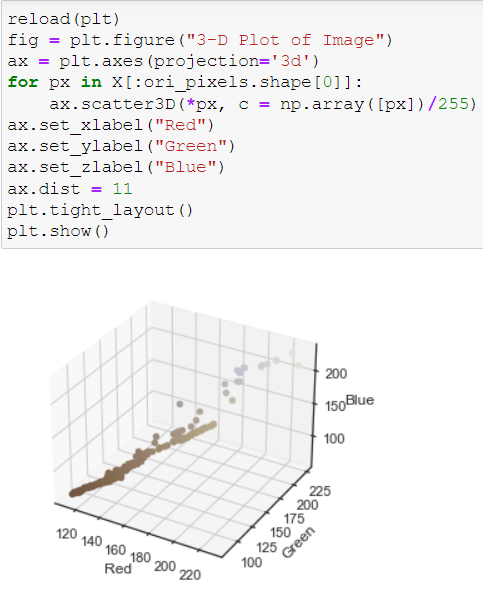
* Tính phương sai của bức ảnh



**Hình 18:** Phương sai của ảnh

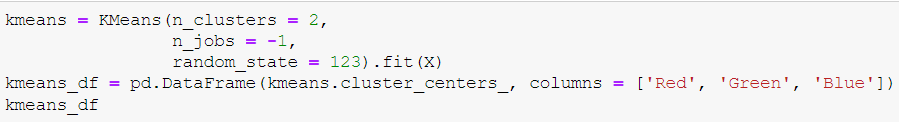
#### Phân tích dữ liệu

* Biểu diễn các điểm RGB của ảnh trên biểu đồ dạng 3D

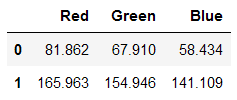


**Hình 19:** Biểu diển các RGB pixel trên mô hình 3D

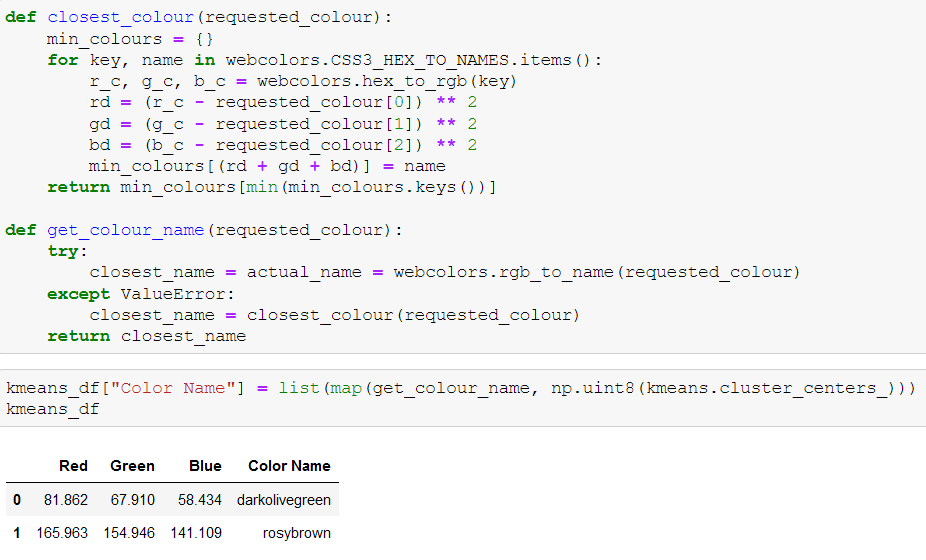
* Đầu tiên, ta có thể xem bức ảnh với 2 màu bằng cách khai báo số cụm k=2 và sau đó sử dụng K-Means để lấy ra các giá trị RGB của 2 màu đó



**Hình 20:** Tìm giá trị RGB cho cụm 2 màu



* Sau đó, ta sẽ tìm xem đây chính xác là 2 màu gì bằng cách chuyển các giá trị RGB rồi so sánh các màu khác nhau rồi chuyển hóa



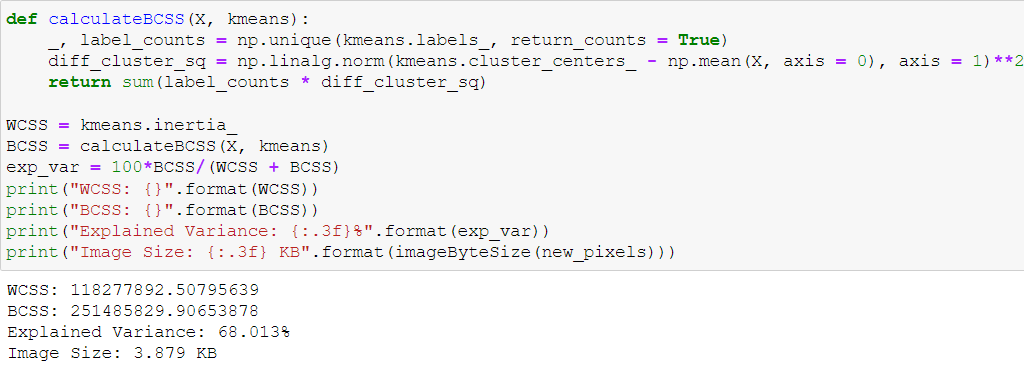
**Hình 21:** Tìm màu sắc cho các giá trị RGB thu được

* Ta có được 2 màu: darkolivegreen và rosybrown. Giờ ta sẽ hiển thị 2 màu này lên trên bức ảnh rồi xuất lên màn hình



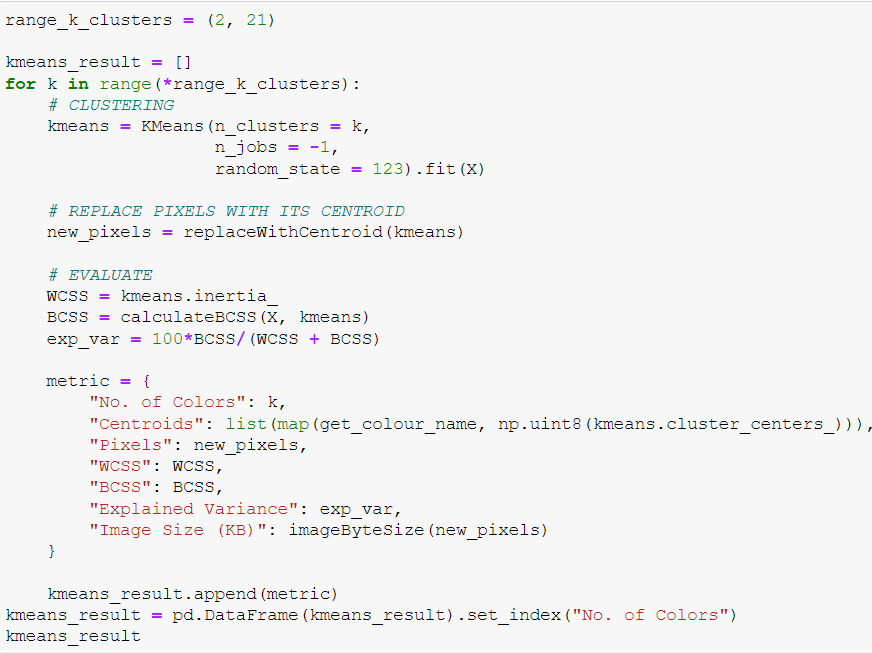
**Hình 22:** Xuất ra hình ảnh với chỉ 2 màu

* Bằng cách dùng phương pháp thay các điểm Pixel bằng các Centroid (Nhân cụm) rồi gom chúng lại với nhau thì chúng ta đã có bức ảnh ban đầu nhưng chỉ với 2 màu duy nhất
* Tính cụm tổng bình phương bao gồm hai giá trị WCSS và BCSS



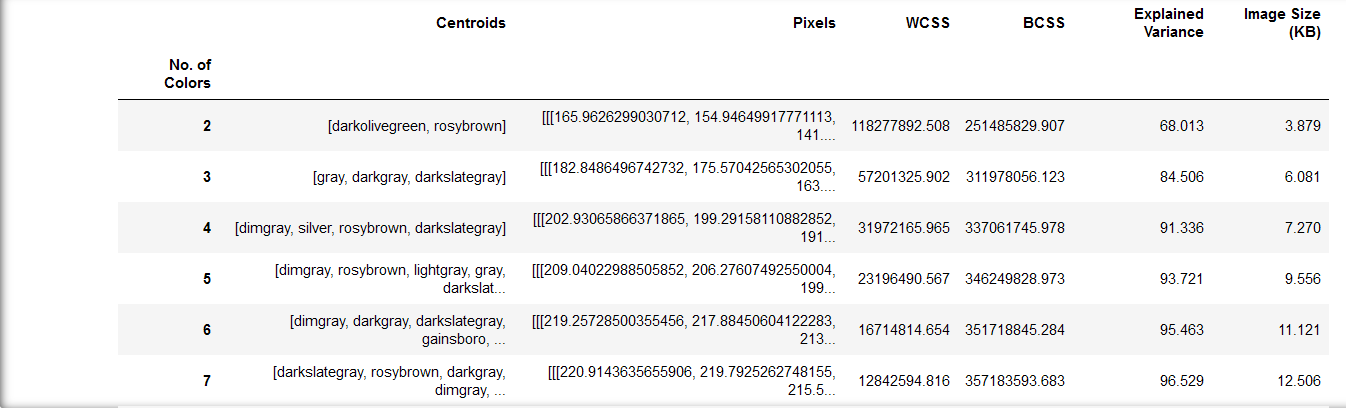
**Hình 23:** Tính WCSS, BCSS, Explained Variance và Image size

* Bằng hai giá trị WCSS và BCSS, ta có thể tìm ra Explained Variance (độ sai màu cao nhất) là 68% và kích cỡ ảnh đã giảm xuống đáng kể, từ 73,574kb xuống còn 3,879kb.
* Tiếp theo, có thể lặp lại các bước này cho một đoạn số cụm k bất kỳ, ở đây sẽ là từ 2 màu đến 20 màu



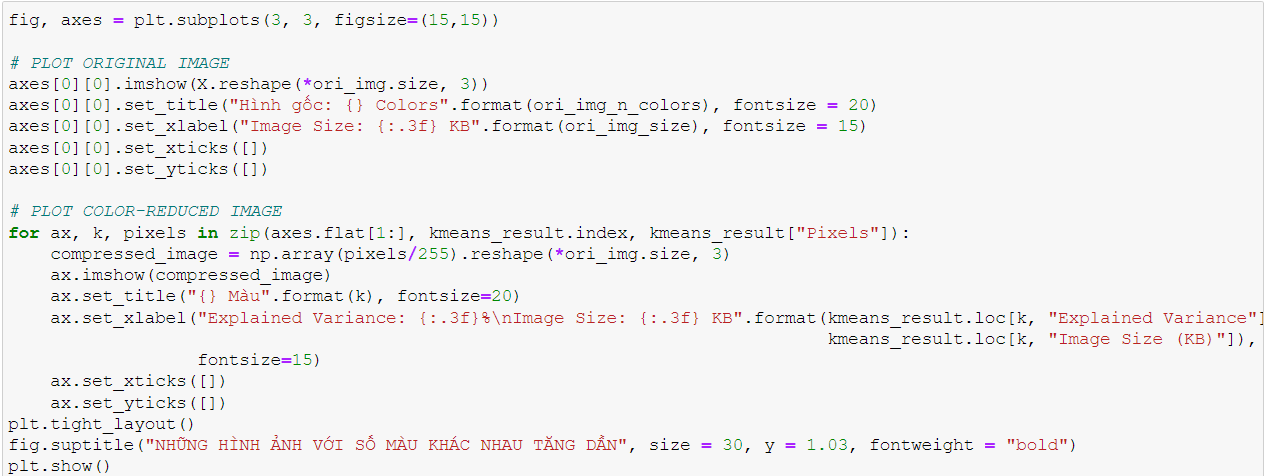
**Hình 24:** Dùng K-Means gom cụm các hình ảnh theo vòng lặp số cụm k cho trước

* Ta thu được kết quả như sau:



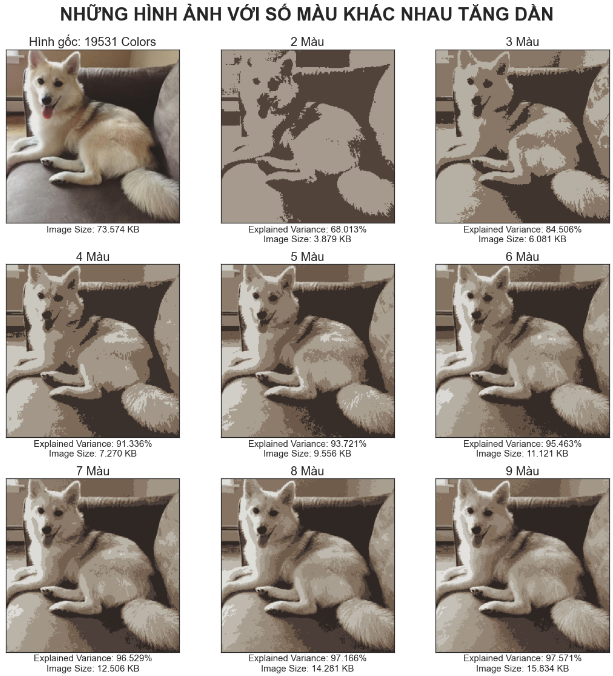
**Hình 25:** Các thông tin hình ảnh thu thập được

* Sau đó, ta sẽ xuất từng dòng dưới dạng hình ảnh với số màu được tăng dần



**Hình 26:** Xuất ra các hình có số màu tăng dần

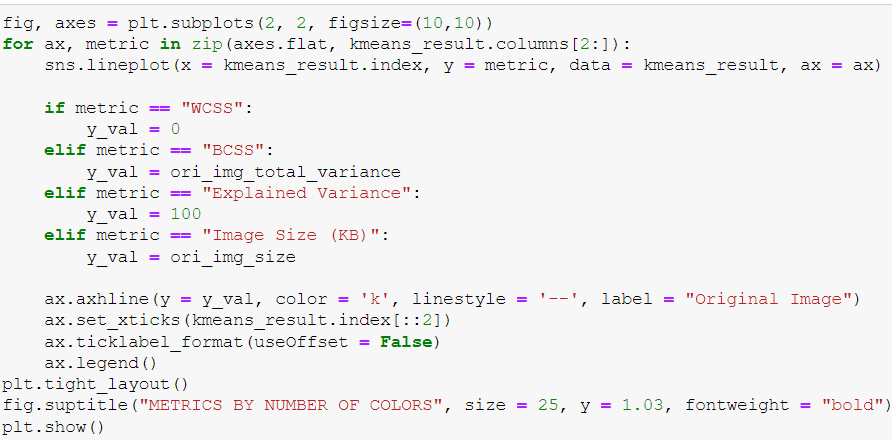
* Ta có bảng các hình ảnh sau



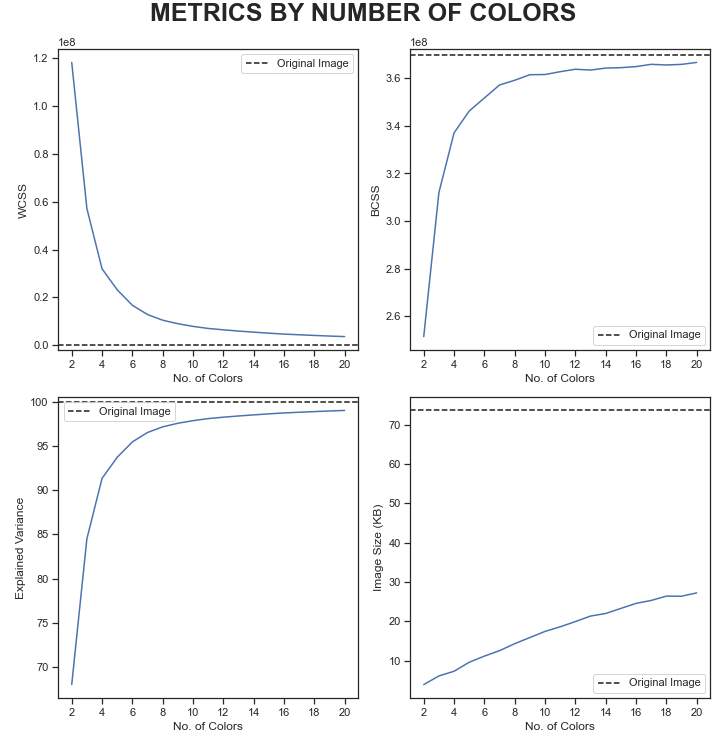
**Hình 27:** Bảng các hình ảnh với số màu tăng dần

#### Tìm hình ảnh nén lại có giá trị tối ưu nhất

* Đầu tiên, ta vẽ ra các sơ đồ tương ứng cho các giá trị WCSS, BCSS, Explained Variance và Image size

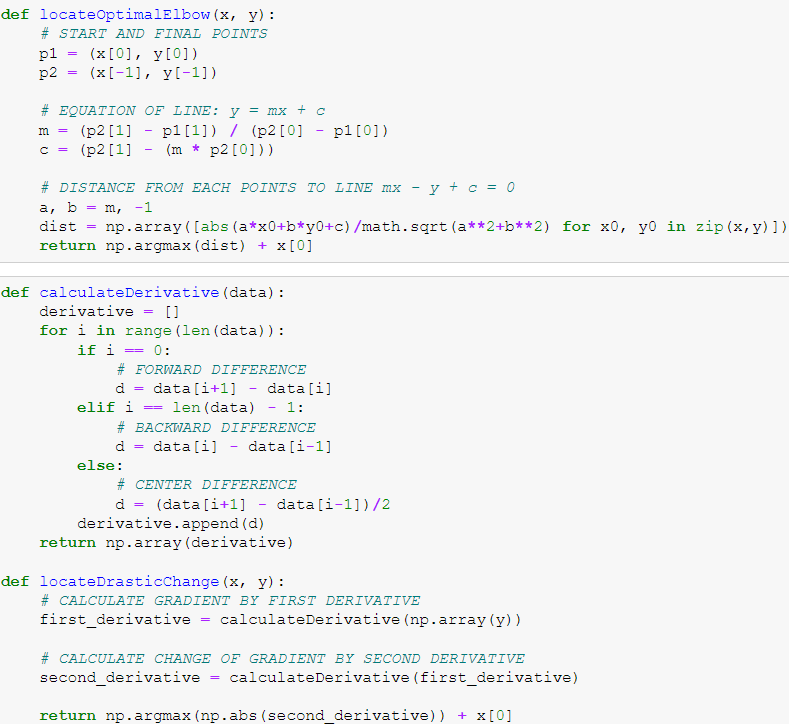


**Hình 28:** Vẽ đồ thị gấp khúc cho các giá trị



**Hình 29**: Đồ thị gấp khúc

* Sau đó, dựa vào các phương pháp toán học để tìm ra số cụm k tối ưu nhất



**Hình 30:** Khai báo phương thức toán học cho thuật toán k tối ưu



**Hình 31**: Tìm ra k tối ưu

* Cuối cùng, so sánh hình ảnh ứng với số cụm k tối ưu nhất với ảnh gốc ban đầu



**Hình 32:** So sánh thông tin giữa ảnh nén tối ưu với ảnh gốc



**Hình 33:** Kết quả của bài toán 2

* Từ đó cho thấy từ một file ảnh 73.574kb và 19531 màu, ta có thể nén tới 84.885% so với ảnh gốc, tương ứng với 11.121kb và 6 màu mà giữ được độ chính xác (Explained Variance) là cao nhất (95.463%).

# **PHẦN TỔNG KẾT**

## Kết quả đạt được

* Sau khi thực hiện báo cáo đồ án, nhóm thực hiện đã có được những kết quả, thành tựu nhất định:
* Hiểu được những kiến thức cơ bản của thuật toán K-Means
* Nắm được cách thức vận hành của thuật toán K-Means thông qua các công thức toán học
* Thấy được tầm quan trọng của thuật toán K-Means trong các bài toán thực tế

## Khó khăn

* Trong quá trình thực hiện báo cáo đồ án nhóm thực hiện nhóm đã gặp phải những khó khăn nhất định:
* Khó tiếp cận ban đầu với ngôn ngữ lập trình Python
* Mất nhiều thời gian trong việc hiểu cách thức hoạt động cúa các công thức toán học
* Khó khăn trong việc sử dụng Jupyter Notebook
* Khó khăn trong việc tiếp cận và thực hiện các ví dụ nâng cao
* Thời gian hạn chế khiến việc hoàn thiện ví dụ vẫn chưa được chỉnh chu nhất có thể

## Bài học kinh nghiệm

* Sau khi thực hiện báo cáo đồ án, nhóm thực hiện báo cáo đồ án đã rút ra được những bài học kinh nghiệm có ích:
* Thuật toán K-Means nói riêng và Machine Learning nói chung đang ngày càng trở phổ biến và chiếm thế thượng phong trong lĩnh vực công nghệ thông tin và trí tuệ nhân tạo AI.
* Biết cách xây dựng một bài toán đơn giản bằng thuật toán K-Means.
* Hiểu hơn về cách thức phối hợp giữa các nền tảng, công nghệ lập trình trong lĩnh vực Machine Learning.

# **TÀI LIỆU THAM KHẢO**

[1] Tham khảo các kỹ thuật gom cụm

https://text.xemtailieu.com/tai-lieu/so-sanh-mot-so-thuat-toan-phan-cum-du-lieu-pdf-82418.html

[2] Tham khảo về thuật toán K-Means

Ebook: Vu 2019 Machine Learning Co Ban

[3] Tham khảo demo Bài toán 1

https://www.kaggle.com/roshansharma/mall-customers-clustering-analysis

[4] Tham khảo demo Bài toán 2

https://ichi.pro/vi/nen-hinh-anh-bang-cach-su-dung-k-means-clustering-va-phan-tich-thanh-phan-chinh-trong-python-150907224231445

[5] Tham khảo chuyển đổi ảnh RGBA sang RGB

https://stackoverflow.com/questions/50331463/convert-rgba-to-rgb-in-python/50332356