# BÀI THỰC HÀNH SỐ 10

## Phân cụm

**Hàm sử dụng:**

***model = KMeans(init = 'k-means++', n_clusters = clusters,  n_init = 12)***

***model.fit(X)***

Hàm này gọi *KMeans ()* từ thư viện sklearn và chỉ định một số tham số chính:

- **init** - Phương thức khởi tạo của centroid. Giá trị sẽ là: 'k-mean ++'. **k-means ++** - Chọn các trung tâm cụm ban đầu cho phân cụm k-means theo cách thông minh để tăng tốc độ hội tụ.
- **n_clusters** - Số lượng cụm sẽ hình thành cũng như số lượng trọng tâm cần tạo. Giá trị sẽ là 3
- **n_init** - Số lần thuật toán k- **mean** sẽ được chạy với các hạt centroid khác nhau. Kết quả cuối cùng sẽ là đầu ra tốt nhất của n_init chạy liên tiếp về quán tính. Giá trị sẽ là 12
- ***random_state:*** sử dụng một số nguyên để xác định việc tạo ngẫu nhiên các centroid ban đầu

***phương thức lấy kết quả của phân cụm:***

- ***cluster_centres_:*** trả về một mảng các vị trí centroid
- ***label_:*** nhãn cụm được gán cho mỗi điểm dữ liệu

**Bài 1: Thực hiện phân cụm**

Làm theo bài mẫu **Taodlngau nhien**
**làm theo bài mẫu tạo ra 5 đặc trương (x0, x1, x2, x3, x4) và tiến hành phân cụm và cho nhận xét vớt số cụm khác nhau.**
**Bài 2:** Bài toán phân cụm khách hàng:

Hãy tưởng tượng rằng bạn có tập dữ liệu khách hàng và bạn cần áp dụng phân đoạn khách hàng trên dữ liệu lịch sử này. Phân khúc khách hàng là hoạt động phân chia cơ sở khách hàng thành các nhóm cá nhân có các đặc điểm giống nhau. Đó là một chiến lược quan trọng vì một doanh nghiệp có thể nhắm mục tiêu đến những nhóm khách hàng cụ thể này và phân bổ hiệu quả các nguồn lực tiếp thị. Ví dụ: một nhóm có thể chứa những khách hàng có lợi nhuận cao và ít rủi ro, tức là, có nhiều khả năng mua sản phẩm hoặc đăng ký dịch vụ hơn. Nhiệm vụ của doanh nghiệp là giữ chân những khách hàng đó. Một nhóm khác có thể bao gồm khách hàng từ các tổ chức phi lợi nhuận. Bây giờ hãy sử dụng thuật toán K-Means để phân khúc khách hàng dựa trên các đặc điểm được cung cấp trong dữ liệu với python.
(làm theo bài mẫu **khachhang) trên tệp dữ liệu khách hàng.**
**Bài 3**: Sử dụng tệp dữ liệu: **Mall_Customers.csv**
Thực hiện tiền xử lý dữ liệu, phân tích dữ liệu theo các trường: Genre, Age, Annual Income (k$), Spending Score (1-100) theo quan điểm của mình.

Phân cụm dữ liệu cho tệp dữ liệu
**Bài 4**: Thực hiên phân cụm trên tệp dữ liệu ThiTHPT2018 dựa vào điểm thi theo từng khối
Xử lý dữ liệu trước khi phân cụm.