Ex2: Ý nghĩa tham số radius, min sample trong thuật toán dbscan? Nếu chỉ số lớn, nhỏ ảnh hưởng thế nào tới thuật toán?

Radius: khoảng cách để xác định vùng lân cận (neighbor) của bất kỳ điểm dữ liệu nào.

- Nếu radius quá nhỏ, 1 phần lớn dữ liệu sẽ không được phân cụm và được xem là outliers
- Nếu radius quá cao, các cụm sẽ bị đồng nhất và phần lớn các điểm sẽ nằm trong cùng 1 cụm.

*Min sample***:** ngưỡng số điểm dữ liệu tối thiểu được nhóm lại với nhau nhằm xác định vùng neighbor có mật độ cao (xác định bởi đường tròn bán kính = radius). Số lượng min sample không gồm điểm ở tâm.

- Min sample quá nhỏ => nhiều core point hơn, cluster sẽ nhiễu hơn. Min sample = 1: giá trị không có ý nghĩa, vì khi đó mọi điểm bản thân nó đều là 1 cụm.
- Min sample quá lớn: nhiều điểm sẽ bị coi là noise hơn, kích cỡ cluster bé đi và có thể không bao quát được cluster thật.
- Min sample phải được chọn ít nhất là 3.

Ex3: So sánh 3 thuật toán: kmeans, GMM, dbscan. Khi nào nên sử dụng thuật toán nào? cho ví dụ?

	Ưu điểm	***Nhược điểm***	***Dùng trong trường hợp nào?***
Kmeans	Đơn giản, dễ sử dụng, dễ implement	Phải xác định trước số cụm cho thuật toán Vị trí tâm của cụm phụ thuộc vào điểm khởi tạo ban đầu của chúng: Những vị trí khởi tạo khác nhau có thể dẫn tới cách phân cụm khác nhau. Đối với những bộ dữ liệu có hình dạng phức tạp hoặc mất cân bằng thì thuật toán không hội tụ về qui luật phân chia tổng quát. Nhạy cảm với outliers: Khi xuất hiện outliers thì thường khiến cho tâm cụm bị chệch, ảnh hưởng đến performance.	

		Nhạy cảm với độ lớn đơn vị của biến nên cần chuẩn hoá biến để loại bỏ sự khác biệt đơn vị trước khi đưa vào train. Không phù hợp đối với dữ liệu kích cỡ lớn do k-Means yêu cầu phải tính khoảng cách từ một điểm tới toàn bộ các tâm cụm để tìm ra tâm cụm gần nhất	
GMM	Xử lý dữ liệu có hình thù cụm đa dạng hơn, chủ yếu là các cụm tạo thành hình elip (Kmeans chỉ thực sự tốt ở các cụm có dạng gần giống hình cầu) Soft assignment: trong k-means 1 điểm chỉ thuộc 1 cluster do k-means là hard assigment. Tuy nhiên, ở trong GMM, 1 điểm có thể thuộc vào nhiều cluster với mức độ khác nhau. Điều này hữu ích trong một số task như một bài báo có thể thuộc nhiều chủ đề,..	Không xác định chính xác được mật độ dữ liệu của một số cụm có hình dạng đặc thù	Ước tính mật độ và hình học phẳng
DBSCAN	Tự động loại bỏ outliers Hoạt động tốt đối với những dữ liệu có hình dạng phân phối đặc thù Tốc độ tính toán nhanh	Không hiệu quả đối với những dữ liệu có phân phối đều khắp nơi. Nhạy cảm với hyperparameters radius và min sample	Kích thước cụm không đồng đều và hình học không phẳng