Câu 2:

- Minpoint (minpts): minPts tối thiểu có thể được tính theo số chiều D trong tập dữ liệu là minPts ≥ D+1, là một ngưỡng số điểm dữ liệu tối thiểu được nhóm lại với nhau nhằm xác định một vùng lân cận epsilon có mật độ cao. Số lượng minpoint không bao gồm điểm ở tâm.
- Radius (Epsilon): một giá trị khaongr cách được sử dụng để xác định cùng lân cận radius của bất kỳ điểm nào.
- Nếu chỉ số radius hoặc minPts quá lớn, dữ liệu được phân thành cụm lớn (số lượng cụm ít) => có thể dẫn tới phân tách không tốt; Thậm chí các điểm sẽ nằm trong một cụm.
- Nếu chỉ số radius hoặc MinPits càng nhỏ, dữ liệu được chia thành các cụm nhỏ (độ chính xác phân cụm không cao).

Câu 3:

1. Kmeans: bài toán giảm chiều dữ liệu. Chia các datapoints thành các cụm dựa vào đặc tính của nó nhưng dữ liệu không có nhãn
   Xét bài toán: Dataset $\{x_1, x_2, \dots, x_n\}$, mỗi dữ liệu có D features => muốn chia thành K cụm.
   Ý tưởng: Dữ liệu nào có đặc tính gần giống nhau thì cho thành 1 cụm, xa nhau cho vào cụm khác.

   Có n samples, k clusters:
   + $r_{nk} = 1\ nếu\ x_n\ \in\ C_K: ||x_n - \mu_k||^2\ \leq\ ||x_t - \mu_t||^2$
   + $r_{nk} = 0\ nếu\ ngược\ lại.$
   Bài toán: Tìm centroid sao cho khoảng cách từ centroid đến các điểm dữ liệu trong từng cụm là nhỏ nhất.
   $L = \sum_{i=1}^{n}\sum_{j=1}^{K} r_{ij}||x_i - \mu_j||^2$ => minimize L tìm r_{ij}, μ_j.

   3 steps:
   B1: chọn centroid bất kỳ => minize L.
   B2: fix $\mu_k, tìm\ r_{nk}$.
   B3: fix $r_{nk}, tìm\ \mu_k$.
   Lặp lại bước 2 và 3 đến khi các cluster không thay đổi.

2. Gaussian Mixture Model.
   Bài toán: Có các điểm dữ liệu, tìm $\theta = \{\pi_k, \mu_k, \sum_k\}$ tham số sinh ra dữ liệu.

3. DBSCAN: DSCAN không dựa vào khoảng cách, sử dụng density để phân tách các cụm (dựa vào mật độ)
   Thuật toán: 2 bước.
   - Bước 1: Lựa chọn một điểm dữ liệu bất kỳ. Sau đó xác định các corepoint và border point thông qua epsilon.

- Bước 2: Cụm hoàn toàn được xác định không thể mở rộng thêm. Khi đó lặp lại đệ quy toàn bộ quá trình với điểm khởi tạo trong số các điểm dữ liệu còn lại để xác định cụm mới.

*** So sánh:

K-means	GMM	DBSCAN
Các cụm được hình thành có dạng hình cầu hoặc lồi và phải cùng kích thước, đặc điểm.	Xử lý nhiều hình dạng hơn, chủ yếu là các cụm tạo thành hình elip. Các điểm dữ liệu được tạo ra từ sự kết hợp tuyến tính của các phân phối Gaussian đa biến với tham số chưa biết.	Các cụm hình thành có dạng tùy ý và có thể không cùng kích thước đối tượng.
Phân cụm K-mean sẽ tuân theo số cụm được chỉ định. Hard-assigment (mỗi điểm dữ liệu chỉ ở một cụm).	Soft-assigment. Phân cụm dựa trên xác suất hoặc khả năng điểm dữ liệu tồn tại cụm đó.	Số lượng cụm không cần khai báo trước.
K-mean hiệu quả hơn cho tập dữ liệu lớn.	Phù hợp với dữ liệu mảng.	DBSCAN clustering không xử lý hiệu quả với các tập dữ liệu có nhiều chiều.
K-mean không tốt đối với bộ dữ liệu nhiều outliers và noises (Do xác định cụm liên quan đến khoảng cách của centroid và datapoint của cụm đó).	Xử lý tốt với cụm nhiễu, chồng chéo, kéo dài.	DBSCAN xử lý hiệu quả outlier và noise.
1 tham số: k cụm	2 tham số: r_{nk} và μ_k	2 tham số: Radius và minpoint
Mật độ thay đổi của các điểm không ảnh hưởng đến thuật toán phân cụm.	Hoạt động tốt với các phân bố hình học phi tuyến tính.	Không hoạt động tốt với bộ dữ liệu thưa thớt hoặc điểm dữ liệu có mật độ thay đổi.