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 (Eps):Khoảng cách chỉ định các vùng lân cận. Hai điểm được coi là hàng xóm nếu khoảng cách giữa chúng nhỏ hơn hoặc bằng eps.
- Min sample (minPts): Số điểm dữ liệu tối thiểu để xác định một cụm. Số lượng min sample không bao gồm điểm ở tâm và min sample ít nhất phải là 3.
- Nếu radius được chọn quá nhỏ, một phần lớn dữ liệu sẽ không được phân cụm và được xem là nhiễu, trong khi đối với giá trị quá cao, các cụm sẽ hợp nhất và phần lớn các điểm sẽ nằm trong cùng một cụm.
- Min sample thấp giúp thuật toán xây dựng nhiều cụm hơn với nhiều nhiễu hoặc ngoại lệ hơn. Min sample cao hơn sẽ đảm bảo các cụm mạnh mẽ hơn nhưng nếu nó quá lớn, các cụm nhỏ hơn sẽ được kết hợp thành các cụm lớn hơn.

Ex3: Biến đổi lại và so sánh ba thuật toán: kmean, GMM, dbscan. Khi nào nên sử dụng thuật toán nào? Cho ví dụ?

1. **K-mean**

Pros:

- Quickest centroid based algorithm
- Very lucid and can scale up for large amount of dataset
- Reduces intra-cluster variance measure

Cons:

- Suffers when there is noise in the data
- Outliers can never be identified
- Even though it reduces intra-cluster variance, it faces local minimum problem
- Not ideal for datasets of non-convex shapes
- Complicated to predict best K value

Use Cases: Even cluster size, flat geometry, not too many clusters and general-purpose
Example: Image segmentation, Genetic analysis in medicine…

2. **DBSCAN**

Pros:

- Resistant to outliers
- Can handle clusters of different shapes and sizes
- Not required to specify the number of cluster

Cons:

- Highly sensitive to the two-parameters: Radius and Min sample
- DBSCAN cannot cluster datasets well with large variances in densities

Use cases: Uneven cluster sizes and non- flat geometry
Example: Face clustering

3. **GMM**

Pros:

- Robust to outliers
- Provides the BIC score for selecting parameters
- Converges fast given good initialization

Cons: The algorithm is highly complex and can be slow
Use cases: Good for density estimation and flat geometry
Example: