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

Radius: Một giá trị khoảng cách được sử dụng để xác định *vùng lân cận radius* của bất kỳ điểm dữ liệu nào.

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 Radius quá cao, các cụm sẽ hợp nhất và phần lớn các điểm sẽ nằm trong cùng một cụm.

Min sample: Là một ngưỡng số điểm dữ liệu tối thiểu được nhóm lại với nhau nhằm xác định một *vùng lân cận radius* có mật độ cao. Số lượng Min sample không bao gồm điểm ở tâm.

Nếu Min sample thấp có nghĩa là nó sẽ xây dựng nhiều cụm hơn từ là tiếng ồn(Min sample ít nhất là 3), trong khi Min sample cao có nghĩa nó có thể bao gồm tất cả các điểm trong tập dữ liệu(nên chọn Min sample = 2* dim, dim là chiều của dữ liệu, còn đối với dữ liệu 2 chiều nên chọn Min sample = 4).

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

Model	Pros	Cons	Use Cases	Example
K-mean	+) Quickest centroid based algorithm +) Very lucid and can scale up for large amount of dataset +) Reduces intra-cluster variance measure	+) Suffers when there is noise in the data +) Outliers can never be identified +) Even though it reduces intra-cluster variance, it faces local mimimum problem +) Not ideal for datasets of non-convex shapes +) Complicated to predict best K value	Even cluster size, flat geometry, not too many clusters and general-purpose	+)Segmentation of customers in business +) Image segmentation +) Genetic analysis in medicine +) Anomaly detection.
DBSCAN	+) Resistant to outliers +) Can handle clusters of different shapes and sizes +) Not required to specify the number of cluster	+) Highly sensitive to the two-parameters: Radius and Min sample +) DBSCAN cannot cluster datasets well with large variances in densities	Uneven cluster sizes and non-flat geometry	+) Face clustering
GMM	+) Robust to outliers +) Provides the BIC score for selecting parameters +) Converges fast given good initialization	The algorithm is highly complex and can be slow	Good for density estimation and flat geometry	+) Clustering of homogeneous bacterial colonies to estimate their size