1. K-평균 알고리즘 이해

- K-평균이란?
    - 군집화에서 가장 일반적으로 사용되는 알고리즘
    - 군집 중심점이라는 특정한 임의의 지점을 선택해 해당 중심에 가장 가까운 포인트들을 선택해서 반복적으로 프로세스를 수행한다.
- 장점
    - 일반적인 군집화에서 가장 많이 활용
    - 알고리즘이 쉽고 간결
- 단정
    - 거리 기반 알고리즘으로 속성의 개수가 매우 많을 경우 군집화 정확도가 떨어진다
    - 반복 수행하므로 횟수가 많아지면 수행시간 느려진다
    - 몇 개의 군집을 선택해야 할지 가이드하기 어렵다
- 사이킷런 패키지는 K-평균을 구현하기 위해 KMeans 클래스 제공
- 데이터 생성기
    - Make_blobs(): 개별 군집의 중심점과 표준 편차 제어 기능이 추가
    - Make_classification(): 노이즈를 포함한 데이터를 만드는 데 유용

2. 군집 평가(Cluster Evaluation)

- 군집화는 분류와 유사해 보일 수 있으나
    - (1)데이터 내에 숨어있는 별도의 그룹을 찾아서 의미를 부여하거나

    - (2)동일한 분류 값에 속하더라도 그 안에서 더 세분화된 군집화를 추구하거나
    - (3)서로 다른 분류 값의 데이터도 더 넓은 군집화 레벨화 등의 영역을 가지고 있다.
- 실루엣 분석: 군집화 평가 방법으로 얼마나 효율적으로 분리돼 있는지 평가
    - 효율적으로 잘 분리됐다 = 다른 군집과의 거리는 떨어져 있고 동일 군집끼리의 데이터는 서로 가깝게 잘 뭉쳐 있다
    - 실루엣 계수: 실루엣 분석이 기반하고 있는 것, 개별 데이터가 가지는 군집화 지표. 해당 데이터가 같은 군집 내의 데이터와 얼마나 가깝게 군집화돼 있고 다른 군집에 있는 데이터와는 얼마나 멀리 분리돼 있는지 나타낸다.
- 좋은 군집화 조건
    - 전체 실루엣 계수 평균값이 1에 가깝다
    - 전체 실루엣 계수의 평균값과 개별 군집 평균값의 편차가 크지 않아야 한다. 즉, 개별 군집의 실루엣 계수 평균값이 전체 실루엣 계수의 평균값에서 크게 벗어나지 않아야 한다.

3. 평균이동

- 평균이동: K-평균과 유사하게 중심을 군집의 중심으로 지속적으로 움직이면서 군집화 수행. 중심을 데이터가 모여 있는 밀도가 가장 높은 곳으로 이동한다.
    - 데이터의 분포도를 이용해 군집 중심점을 찾는다.
    - 특정 데이터를 반경 내의 데이터 분포 확률 밀도가 가장 높은 곳으로 이동하기 위해 주변 데이터와의 거리 값을 KDE 함수 값으로 입력한 뒤 그 반환 값을 현재 위치에서 업데이트하면서 이동한다.

4. GMM(Gaussian Mixture Model)

- GMM: 군집화를 적용하고자 하는 데이터가 여러 개의 가우시안 분포를 가진 데이터 집합들이 섞여서 생성된 것이라는 가정 하에 군집화를 수행하는 방식.

5. DBSCAN

- DBSCAN: 밀도 기반 군집화의 대표적인 알고리즘으로, 특정 공간 내에 데이터 밀도 차이를 기반 알고리즘으로 하고 있어서 복잡한 기하학적 분포도를 가진 데이터 세트에 대해서도 군집화를 잘 수행할 수 있다.