📝 밀도 기반 군집화(DBSCAN)

🎯 목차

- DBSCAN vs K-means vs 계층적 군집화
- 언제 DBSCAN을 선택해야 하는가?

Part 7: 결론 및 추가 자료 (Conclusion & Resources)

- 핵심 요약
- 참고 논문 및 튜토리얼

세부사항

Part 1: 도입 (Introduction)

🔍 왜 군집화가 필요한가?

데이터 속에서 "비슷한 것끼리 묶기"는 인공지능과 데이터 과학의 핵심 과제입니다.
예를 들어:

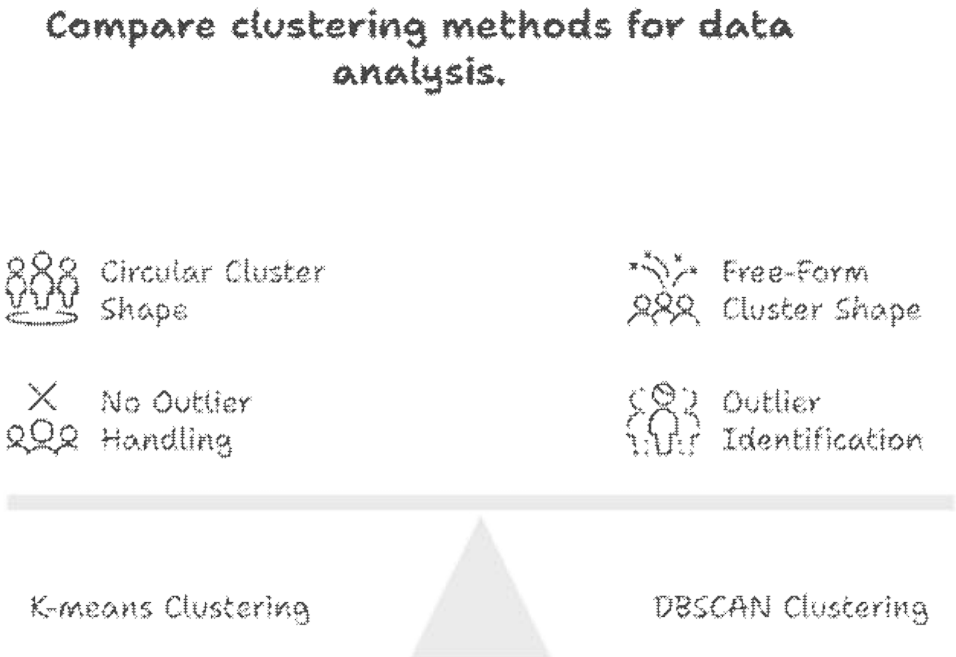
- 고객을 구매 패턴별로 그룹화
- 지도에서 비슷한 위치의 이벤트를 묶어 핫스팟 탐지
- 의료 데이터에서 질병 발생 지역 군집 파악

⚠ K-means의 한계

전통적인 K-means 군집화는 다음과 같은 문제가 있습니다:

1. **클러스터 개수(k)를 미리 지정**해야 함
2. **구형(spherical) 클러스터만** 잘 감지
3. **이상치(outlier)도 강제로 군집에 포함**
4. 비선형·복잡한 형태의 군집은 감지 못함

이러한 한계를 극복하기 위해 밀도 기반 군집화(Density-Based Clustering)가 등장했습니다.



Part 2: 핵심 개념 (Core Concepts)

📌 DBSCAN이란?

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)는 데이터 포인트의 밀도(density)를 기준으로 군집을 형성하는 알고리즘입니다.

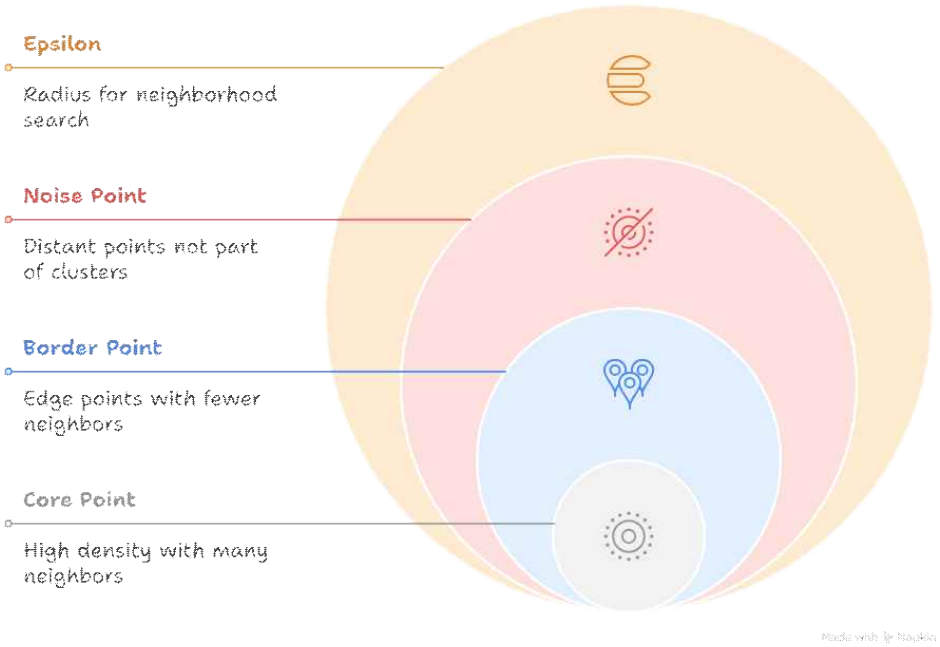
🔑 핵심 용어

용어	설명
Epsilon (ε)	각 포인트 주변의 반경(거리). 이웃을 찾는 기준
MinPts	Core Point로 정의되기 위한 최소 이웃 개수
Core Point	ε 반경 내에 MinPts 이상의 이웃을 가진 포인트 (군집의 중심)
Border Point	Core Point의 이웃이지만, 자신은 Core가 아닌 포인트
Noise Point	어느 군집에도 속하지 않는 이상치

🎬 알고리즘 작동 원리

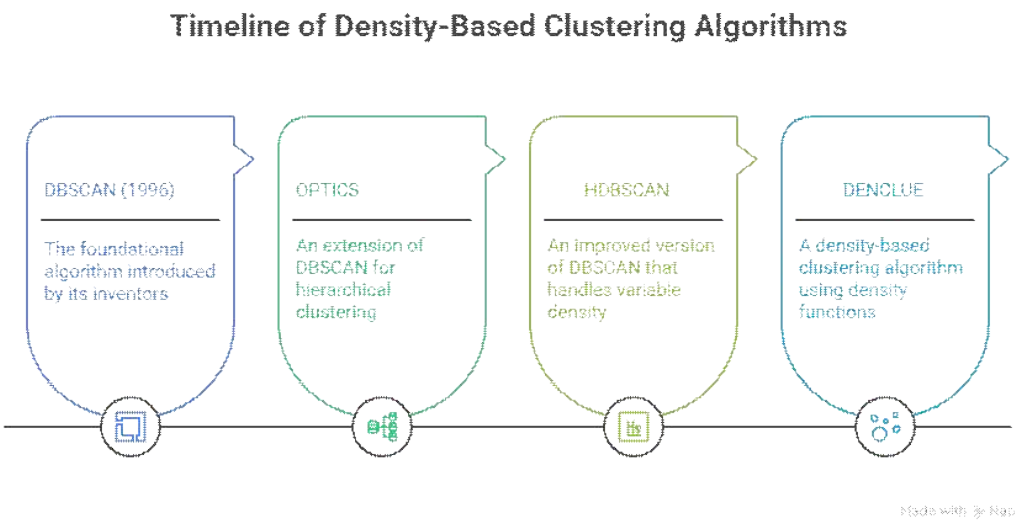
```
1. 임의의 포인트 P 선택
2. P의 ε 반경 내 이웃 개수 확인
3. 이웃 ≥ MinPts → P는 Core Point
   - P로부터 연결된 모든 Core/Border Point를 같은 군집으로 확장
4. 이웃 < MinPts이지만 Core Point의 이웃 → Border Point
5. 그 외 → Noise Point
6. 모든 포인트가 분류될 때까지 반복
```

DBSCAN Clustering Concepts

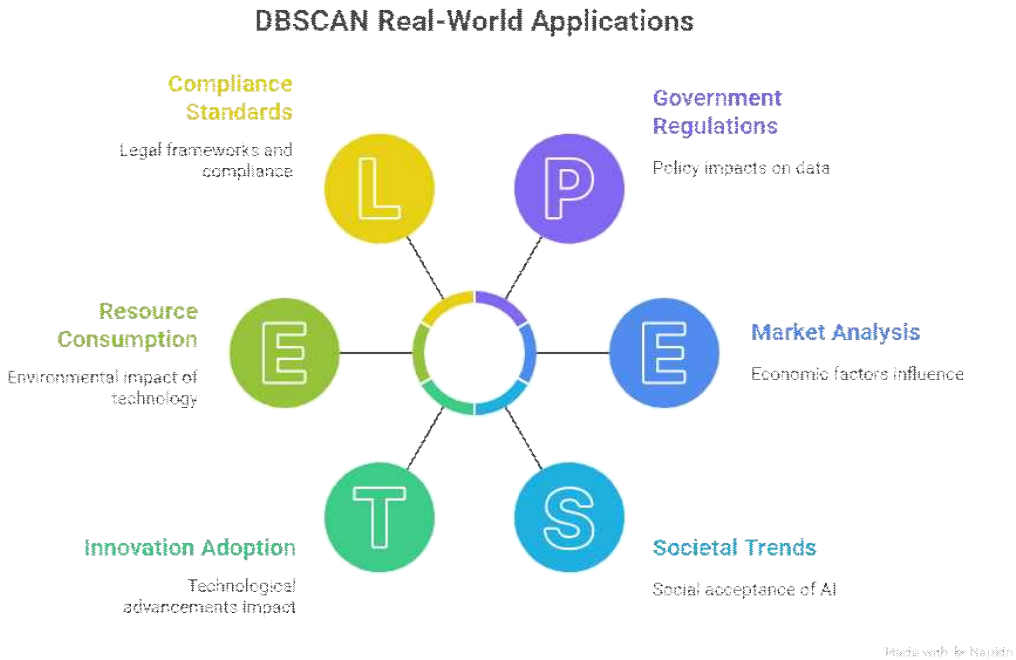


Part 3: 역사와 발전 (History & Evolution)

- DBSCAN은 1996년 Ester, Kriegel, Sander, Xu가 논문에서 공개한 이래('A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise') 군집 알고리즘의 표준이 되었습니다.
- 고정밀 공간 데이터와 이상치 분리 필요성이 늘어나면서 활용도가 폭증했습니다.
- 이후 OPTICS, HDBSCAN, DENCLUE 같은 진화된 밀도기반 알고리즘이 개발되어, 다양한 밀도·복잡구조 데이터에 더 강력하게 대응하도록 변형되었습니다.



Part 4: 실제 응용 사례 (Real-World Applications)



1. **도시 교통 빅데이터**
   - 택시·버스 GPS 좌표를 DBSCAN으로 군집화 → 승하차 핫스팟 분석
   - 결과: 도시 내 교통 혼잡구역, 신규 노선 후보 자동 발굴
2. **이상거래/금융사기 탐지**
   - 거래 금액·장소·패턴의 밀도기반 군집화 → 노이즈가 의미있는 이상거래로 분리
   - 결과: 정상 패턴과 다른 거래감지(사기방지 시스템 백본)
3. **반도체 공정결함 분석**
   - 웨이퍼 표면 결함점 좌표로 DBSCAN → 결함핫스팟 자동탐지
   - 결과: 장비 이상, 공정 편차 빠르게 발견하여 품질 개선
4. **감염병 핫스팟 식별**
   - 환자 위치 데이터로 질병확산 군집 분석
   - 결과: 고위험 지역 우선 방역전략 수립
5. **부동산 시장 분석**
   - 실거래가, 위치 기반 DBSCAN → 유사 가격권/이상거래군 자동분류

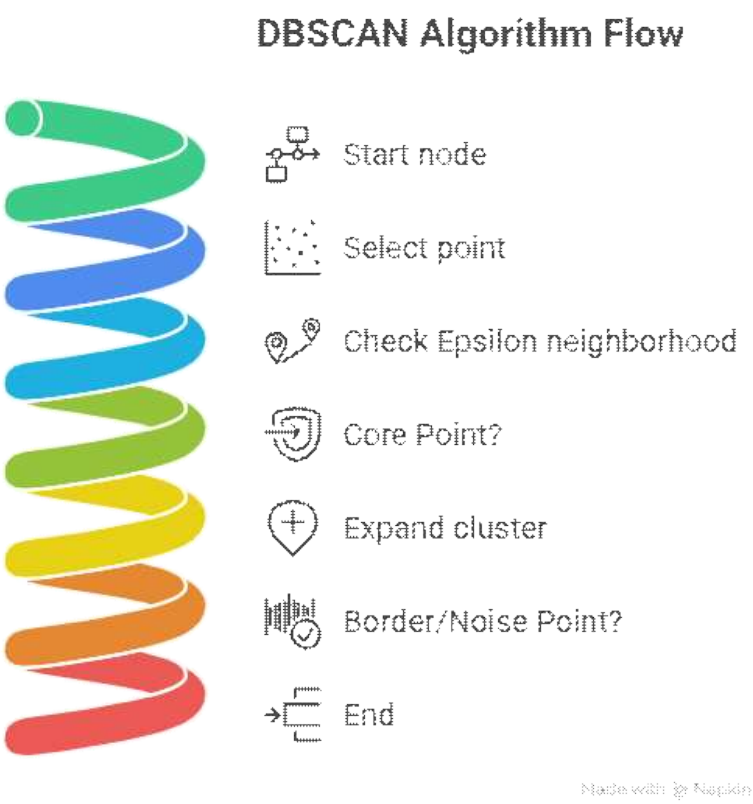
Part 5: 실습 가이드 (Hands-on Tutorial)

- **Python scikit-learn 라이브러리로 DBSCAN 사용 예**

```
from sklearn.cluster import DBSCAN
import numpy as np

X = np.array([[1, 2], [2, 2], [2, 3], [8, 7], [8, 8], [25, 80]])
dbscan = DBSCAN(eps=3, min_samples=2)
labels = dbscan.fit_predict(X)
print(labels)  # 각각의 점이 어느 군집(-1=노이즈)에 속하는지 출력
```

- **최적의 eps, MinPts 선정법:**
    - k-distance plot(팔꿈치 지점)
    - 여러 값 실험하여 군집, 노이즈 분포 직관적으로 점검



Part 6: 장단점 및 선택 가이드 (Pros, Cons & Selection Guide)

장점

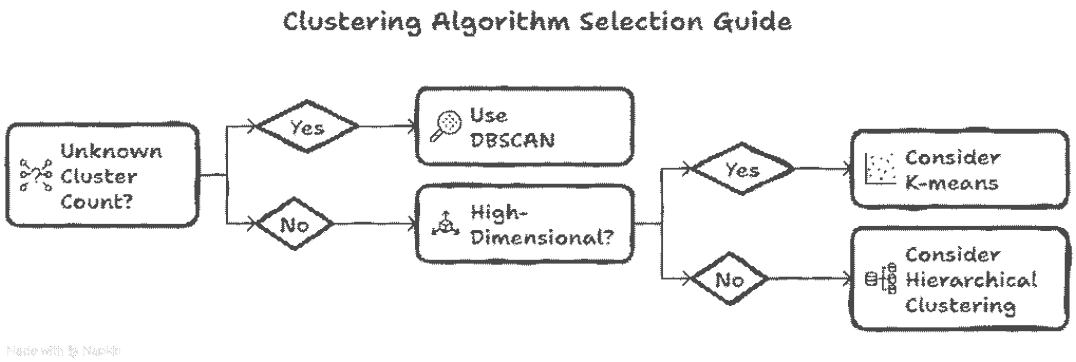
- 클러스터 개수 사전지정이 필요 없음
- 임의의 형태(원형·불규칙·곁가지 등) 군집 감지
- 이상치 자동 분리
- 비선형 데이터도 우수

단점

- 파라미터(eps, MinPts) 튜닝 중요, 데이터 분포 강하게 영향
- 다양한 밀도/잡음 데이터에서는 오동작 가능
- 고차원 데이터(10차원↑)에서는 K-means에 비해 성능 저하

적합상황

- 이상치 감지 우선, 자연현상·지도·GPS 등 위치 기반 데이터
- 비정형, 분포가 복잡한 데이터
- 미리 군집 수를 모를 때



Part 7: 결론 및 추가 자료 (Conclusion & Resources)

핵심 요약

- DBSCAN은 K-means로는 잡히지 않는 복잡·비구형 군집이나 노이즈를 잘 구분하는 강력한 도구
- 단, eps·MinPts 등 파라미터 조정 필요성과 데이터의 밀도 특성 고려가 핵심