3.4 결과 해석

3.4.1 k-means 군집 평가 및 시각화

오차제곱합(Sum of Squared Errors, SSE)

- 각 클러스터 내 샘플이 해당 클러스터 중심으로부터 얼마나 흩어져 있는지 수치화한 지표.
- 수식

$$SSE = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$$

구현

- `sklearn.cluster.KMeans`의 `inertia_` 속성 사용
- 여러 K에 대해 SSE를 계산한 뒤, K값에 따른 SSE 변화를 엘보(elbow) 그래프로 표시

실루엣 계수(Silhouette Coefficient)

- 각 샘플의 응집도(a)와 분리도(b)를 기반으로, "군집 내 응집력 대비 다른 군집과의 분리도"를 –1에서 1 사이로 정규화한 점수
- 한 샘플 i에 대한 계산

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

- $a(i)$: 같은 클러스터 내 다른 샘플들과의 평균 거리
- $b(i)$: 가장 가까운 다른 군집의 모든 샘플들과의 평균 거리
- 전체 군집화 품질: 모든 $s(i)$의 평균
- 구현 및 시각화
  - `sklearn.metrics.silhouette_score`로 전체 점수 계산
  - `sklearn.metrics.silhouette_samples`와 matplotlib으로 각 클러스터별 실루엣 플롯 작성

2차원 산점도 시각화

- 원본 특성 공간이 고차원이면, PCA나 t-SNE를 적용해 2차원으로 축소
- 축소된 좌표에 클러스터 레이블별 색상 및 중심점 표시

3.4.2 DBSCAN 군집 평가 및 시각화

핵심 지표

- 추정된 군집 개수 (excluding noise)
- 노이즈 비율: 라벨이 –1로 판정된 샘플의 비중(군집에 포함되지 않음)
- 실루엣 계수: 노이즈 샘플 제외 후 계산

파라미터 민감도 분석

- ε(eps)와 최소 샘플(min_samples) 조합별로 군집 수 및 노이즈 비율 변화
- 파라미터 그리드(grid) 탐색 결과를 Heatmap으로 시각화(eps, min_samples를 변경해 가면서 시도함)

군집 분포 시각화

- 2차원 축소(PCA/t-SNE) 후,
  - Core 샘플: ●
  - Border 샘플: ○
  - Noise 샘플: ×
- 클러스터별 색상을 달리해 표시

3.4.3 PCA 결과 해석 및 시각화

설명 가능한 분산 비율(Explained Variance Ratio)

- 각 주성분이 전체 데이터 분산을 얼마나 설명하는지 측정
- j번째 성분

$$\mathrm{EVR}_j = \frac{\lambda_j}{\sum_{l=1}^{L} \lambda_l}$$

- 스크리(scree) 플롯: 주성분 번호에 따른 EVR 및 누적 EVR 표시

주성분 산점도

- 1차원/2차원 주성분 축에 원본 데이터 투영
- 군집 라벨별 색상으로 분포 확인

피처 로딩스(Loading Scores)

- 각 원본 피처가 주성분에 기여하는 정도
- 주요 피처 상위 5~10개를 막대그래프로 시각화

3.4.4 추가 해석 및 시각화 기법

- t-SNE 등의 비선형 차원 축소 기법으로 임베딩 후 군집 결과 색상 표시
- 병렬 좌표(parallel coordinates) 플롯: 클러스터별 특성 비교
- 클러스터×피처 히트맵: 클러스터 중심의 특성값 시각화