

차별적 영향 분석 (Disparate Impact Analysis)

Disparate Impact: 중립적으로 보이는 정책이나 관행이 특정 집단에 불균형적인 부정적 영향을 미치는 현상

4/5ths Rule

Disparate Impact Ratio

$$\frac{P(\hat{Y}=1|\text{Group}=\text{B})}{P(\hat{Y}=1|\text{Group}=\text{A})}$$

✓ ≥ 0.8 : 공정함

✗ < 0.8 : 차별적 영향 의심

탐지 방법

- 통계적 유의성 검정 (Chi-square test)
- 그룹별 성능 메트릭 비교
- 교차분석 (Intersectionality)
- 시간에 따른 변화 추적
- 인과관계 분석

의료 AI에서의 사례

- 특정 인종의 낮은 진단율
- 노인 환자의 치료 접근성 제한
- 여성 환자의 과소 치료
- 저소득층 환자의 낮은 권장 등급