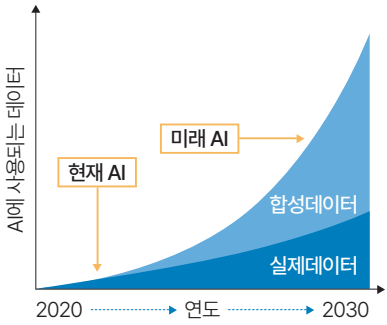
합성데이터의 활용



출처 : Gartner

합성데이터는 모델 학습, 성능 검증, 위험 분석 등 다양한 과정에서 활용됩니다. 의료 분야에서는 실제 환자 정보를 공유하기 어려운 상황에서 합성된 진료 기록이나 의료 영상을 활용해 연구와 알고리즘 개발을 진행할 수 있습니다. 금융 분야에서는 거래 기록이나 신용 패턴을 합성해 위험 평가 모델을 안전하게 검증할 수 있으며, 공공 행정 영역에서는 민감 정보를 포함한 데이터를 합성 버전으로 제공해 데이터 개방성을 높이는 데 기여합니다. 또한 실제 환경에서 수집하기 어려운 희귀 상황이나 극단적 사건을 인위적으로 생성할 수 있어, 드문 패턴을 학습해야 하는 보안·사기 탐지 분야에서도 효과적입니다. 합성데이터는 데이터 부족을 해소하고 민감 정보를 보호하며, 특정 조건의 데이터를 자유롭게 구성할 수 있다는 점에서 AI 개발의 효율성과 접근성을 크게 높이는 기술적 기반이 됩니다.

합성데이터의 과제

합성데이터는 활용 가치가 높지만 몇 가지 한계를 가지고 있습니다. 먼저 원본 데이터의 품질이 낮거나 편향이 심한 경우, 합성데이터도 동일한 한계를 그대로 복제할 수 있습니다. 생성형 모델을 사용할 때는 현실성과 일관성을 확보하는 것이 중요하며, 품질이 떨어진 합성데이터는 모델 성능을 저하시킬 위험이 있습니다. 또한 합성데이터가 원본 데이터를 완전히 대체할 수 있는지에 대한 기준이 명확하지 않아, 실제 모델 평가나 규제 준수 측면에서 신뢰성 검증 절차가 필요합니다. 지나치게 원본 데이터와 유사하게 생성될 경우 재식별 위험이 다시 발생할 수 있다는 점도 주의해야 합니다. 이러한 과제를 해결하기 위해서는 데이터 품질 평가 기준, 안전성 검증 방법, 생성 절차의 투명성 확보가 함께, 합성데이터 활용에 대한 정책적·기술적 가이드라인 마련이 요구됩니다.

관련 용어

업샘플링 (Up-sampling) & 다운샘플링 (Down-sampling)

업샘플링은 데이터 분포가 한쪽으로 치우쳐 있을 때, 소수 클래스의 데이터를 인위적으로 늘려 학습 균형을 맞추는 기법입니다. 기존 데이터를 단순 복제하거나 변형해 늘리거나, 생성형 모델을 활용해 새로운 합성 데이터를 만들어 보완하는 방식이 사용됩니다. 이를 통해 모델이 특정 클래스에만 편향되는 현상을 줄이고, 소수 클래스의 패턴을 안정적으로 학습하도록 돕습니다.

반대로 다운샘플링은 다수 클래스의 데이터 양을 줄여 전체 분포를 균형 있게 만드는 방법입니다. 불필요한 데이터를 제거해 학습 속도를 높이거나, 한 클래스가 전체 모델 판단을 지배하는 상황을 방지하는 데 효과적입니다.

두 기법은 모두 데이터 불균형 문제를 해결하기 위한 대표적 방법으로, 합성데이터와 함께 사용할 때 부족한 영역을 보완하고 편향을 줄이는 데 유용합니다. 특히 소수 클래스가 중요한 의미를 갖는 의료·보안·사기 탐지 분야에서는 업샘플링과 다운샘플링을 적절히 조합해 모델의 일반화 능력과 예측 신뢰도를 높일 수 있습니다.