

016 데이터 랭글링

Data Wrangling

데이터를 분석 목적에 맞게 정리하고 구조화하는 과정

- 정제된 데이터를 다시 가공해 분석이나 AI 학습에 활용하기 적합한 형태로 만드는 절차로, 데이터 전처리 이후 실제 활용 단계와 분석 효율을 높이는 핵심과정

데이터 랭글링의 개념

데이터 랭글링은 전처리를 거친 데이터를 분석이나 AI 학습에 직접 활용할 수 있도록 재구성하는 과정입니다. 전처리가 데이터의 오류를 수정하고 품질을 높이는 단계라면, 랭글링은 목적에 맞게 데이터를 선별·결합·변환해 활용 가능한 형태로 만드는 단계입니다. 예를 들어 여러 데이터셋을 합치거나 특정 항목을 추출해 새로운 구조를 구성하는 것이 이에 해당합니다. 이 과정을 통해 방대한 데이터를 분석 목적에 맞게 간결화하고, 모델이 이해할 수 있는 입력 형태로 준비할 수 있습니다.

데이터 랭글링의 과정

데이터 랭글링은 탐색 → 구조화 → 정제 → 확충 → 검증 → 배포의 순환 과정을 거칩니다. 먼저 탐색 단계에서 데이터의 형태와 품질을 살펴보고, 구조화 단계에서는 서로 다른 형식의 데이터를 공통 구조로 맞추습니다. 정제 단계에서는 잘못된 값이나 불필요한 정보를 제거하며, 확충 단계에서는 부족한 정보를 보완하거나 필요한 속성을 추가합니다. 검증 단계에서는 오류나 불일치를 점검해 신뢰성을 확보하고, 마지막 배포 단계에서는 완성된 데이터를 분석 환경이나 AI 학습 시스템에 전달해 활용합니다. 이를 통해 데이터 랭글링은 품질과 활용성을 동시에 높입니다.



데이터 랭글링의 의의

데이터 랭글링은 AI 개발과 데이터 분석의 실질적 생산성을 높이는 핵심 과정입니다. 이 단계를 통해 방대한 데이터를 빠르게 다루고, 필요한 정보만 추출해 인사이트를 도출할 수 있습니다. 특히 AI 모델의 입력 데이터가 구조화되어 있을수록 학습 속도와 예측 정확도가 향상되기 때문에, 랭글링은 단순한 정리 작업을 넘어 AI 학습의 효율과 해석 가능성을 높이는 기술로 평가됩니다. 또한 자동화된 랭글링 시스템은 반복적 수작업을 줄여 데이터 사이언스 전반의 접근성과 속도를 개선하고, 데이터 활용의 범위를 한층 확장시키고 있습니다.