

# 018 데이터 전처리

Data Preprocessing

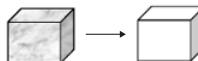
## AI 학습에 사용될 데이터를 정리하고 변환해 품질을 높이는 과정

- 다양한 출처에서 수집된 데이터를 분석 가능한 형태로 정제·가공해, AI 모델이 안정적으로 학습할 수 있도록 만드는 절차
- 데이터를 바르게 이해·예측할 수 있도록 하는 단계로, AI 개발의 출발점이자 성능을 좌우하는 단계

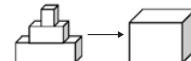
### 데이터 전처리의 개념

데이터 전처리는 AI가 학습하기 전에 원시 데이터를 분석 가능한 형태로 다듬는 과정입니다. 실제 수집된 데이터에는 누락값, 오류, 중복, 잡음 등 다양한 결함이 포함될 수 있으며, 이러한 데이터는 그대로 학습에 사용하면 왜곡된 결과를 초래할 수 있습니다. 전처리는 이런 문제를 사전에 수정하고 구조를 표준화해, 모델이 의미 있는 패턴을 정확히 학습할 수 있도록 돕습니다. 즉, 데이터 전처리는 AI 학습 전반의 품질과 신뢰성을 확보하기 위한 준비 과정 이자, 모델의 성능을 결정하는 핵심 절차입니다.

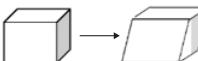
데이터 클리닝



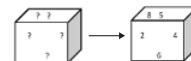
데이터 정규화



데이터 변환



결측치 처리



데이터 통합



노이즈 식별



출처 : Big Data Analytics

### 데이터 전처리의 개념

전처리는 일반적으로 정제, 통합, 변환, 축소의 네 단계로 이루어집니다. 정제 단계에서는 누락된 데이터를 보완하거나 잘못된 값을 교정하고, 이상치나 불필요한 정보를 제거해 데이터의 오류를 바로잡습니다. 통합 단계에서는 여러 출처의 데이터를 결합해 형식과 단위를 일관되게 정리하며, 변환 단계에서는 값의 범위를 조정하고 비정형 데이터를 수치 형태로 변환해 모델이 처리할 수 있도록 만듭니다. 마지막으로 축소 단계에서는 분석에 꼭 필요한 변수만 남기거나 일부 데이터를 표본으로 추출해 연산 효율을 높입니다. 최근에는 이러한 과정을 자동화한 전처리 도구와 파이프라인이 도입되어, 대규모 데이터셋의 품질을 안정적으로 유지하고 처리 속도와 정확도를 동시에 향상시키고 있습니다.

### 데이터 전처리의 의의

데이터 전처리는 AI 성능을 좌우하는 핵심 품질 관리 기술입니다. 부정확하거나 편향된 데이터는 모델의 판단 오류, 과적합, 공정성 문제로 이어질 수 있습니다. 따라서 전처리를 통해 데이터의 정확성과 다양성을 확보하는 일은 AI의 신뢰성과 책임성을 높이는 데 필수적입니다. 또한 전처리는 데이터 랭글링, 라벨링, 피처 엔지니어링 등 후속 과정의 효율을 향상시켜 AI 개발 전반의 자동화와 품질 향상을 가능하게 합니다.