데이터 전처리 결과서

# **AI 푸드닥터(푸닥) 데이터 전처리 보고서**

* **팀 정보**
  + 팀원: 서주혁, 대성원, 신동익, 윤정연
  + 깃허브: <https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN07-FINAL-3Team>
* **개요**
  + 산출물 단계 : 모델링 및 평가
  + 평가 산출물 : 수집된 데이터 및 전처리 문서
  + 제출 일자 : 2025.04.04
  + 깃허브 경로 : <https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN07-FINAL-3Team>

| **개요** | * **데이터 설명:** 식품의약품안전처에서 수집한 음식 영양 정보 및 분류 데이터 * **데이터 수집목적:** 개인 맞춤형 식단 추천 시스템을 개발하기 위한 기초 데이터 확보 및 전처리 |
| --- | --- |
| **데이터 자동화 및 검증** | * **데이터 수집 자동화 프로세스:** Python(Pandas)을 활용하여 대량의 음식 데이터를 자동 수집 및 전처리 (결측치 제거, 분류 매핑 등) * **검증:** 중복 및 이상치 제거, 불필요 식품군(음료, 유제품, 장류 등) 필터링 |
| **데이터 저장 및 관리** | * **데이터 저장 방식:** MySQL 관계형 데이터베이스에 정제된 데이터를 저장 * **관리 방식:** Django ORM을 통해 CRUD 처리 및 FastAPI와 연동하여 백엔드 서비스와 통신 |
| **데이터 전처리 과정** | * **전처리 단계:**   ① 주요 컬럼(식품명, 중량, 영양소) 선택  ② 결측치 및 중복 제거  ③ 대/중/소분류 기반 식단 카테고리화   * **사용 도구:** Python, Pandas, Jupyter |
| **데이터 전처리 결과** | * **결과:** 원본 14,751건 → 전처리 후 552건 정제 완료 * **향후 데이터 사용 계획:**   ① LLM 기반 식단 추천 시스템에 활용  ② 식단 구성 AI 알고리즘 학습용 데이터로 사용  ③ 식단 만족도 기반 개선 모델 개발 예정 |