데이터 전처리 결과서

# **AI 푸드닥터(푸닥) 데이터 수집 보고서**

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

| **데이터 수집 목적** | 기업 임직원을 위한 맞춤형 식단 추천 서비스 개발을 위해 음식별 영양성분, 식품 분류, 알레르기 정보, 선호도 정보 등을 수집하여 분석 기반을 마련한다. 이를 통해 건강관리, 시간 절약, 메뉴 스트레스 해소를 동시에 해결하는 AI 서비스 구현을 목표로 한다. |
| --- | --- |
| **데이터 수집 방법** | - 식품의약품안전처 공공 데이터포털에서 CSV 파일 형태로 음식 데이터 14,751건 수집  - 네이버 백과사전에서 음식 이미지 크롤링  - 알러지 및 건강 정보는 문서 형태로 수집 후 벡터화하여 검색용 데이터로 활용 |
| **수집 데이터(요약)** | - 총 14,751건 → 전처리 후 552건으로 정제  - 주요 컬럼: 식품명, 분류(대/중/소), 중량, 에너지, 탄수화물, 단백질, 지방, 당류, 식이섬유, 나트륨 등  - 불필요한 식품군 제거 (음료, 유제품 등)  - 각 식품을 '밥/반찬/국/기타'로 식단 항목 매핑  - 전처리 도구: Python / Pandas 활용 |