데이터 수집 방법 결정 및 데이터셋 확보

Aiden.ji(지윤수)

Luna.jung(정지연)

1. 서론 (Introduction)

- **문서 목적 :** 이 문서는 AI 모델 학습을 위해 필요한 이미지 및 정형 데이터를 수집하고, 데이터 품질을 관리하며, 이를 최적화하는 방법에 대해 설명하기 위해 작성되었다.
- **배경:** 헬스케어 데이터는 민감한 개인의 신체 정보(키, 체중, 칼로리 섭취량 등)를 포함하여 철저한 보호가 필요합니다. 또한, 이러한 데이터는 진단과 예측 모델을 통해 개인 맞춤형 식단 추천 등의 다양한 응용 가능성을 가지고 있습니다.

2. 데이터 수집 방법 (Data Collection Methods)

- **데이터 유형 분류:**
    - 음식 사진 데이터는 AI가 시각적으로 식품을 인식하는 데 필요한 정보를 제공하고, 칼로리 데이터는 식단 계획에 필요한 영양 정보를 제공합니다.
- **데이터 수집 경로:**
    - 회원 조사: 사용자가 직접 입력하는 신체 정보와 식사 기록을 바탕으로 수집
    - 공공 데이터베이스: 공공기관에서 제공하는 공개 데이터를 통해 확보
    - 구글 크롤링: 웹 크롤러를 통해 음식 이미지를 자동으로 수집
- **데이터 수집 도구 및 기술:**
    - Python 기반의 웹 크롤링 도구인 BeautifulSoup 및 Selenium을 사용하여 웹에서 데이터를 수집하며, API 통합(GET 요청)을 통해 효율적으로 데이터를 관리합니다."

3. 데이터 품질 관리 (Data Quality Control)

- 아래와 같은 기준으로 지속해서 관리가 필요.

- **정확성 (Accuracy):** 신뢰성 있는 출처에서 수집된 데이터를 우선으로 사용
- **완전성 (Completeness):** 데이터가 얼마나 누락되지 않고 수집
- **일관성 (Consistency):** 서로 다른 출처에서 수집된 데이터 간의 일관성 여부 확인
- **정규화 (Normalization):** 데이터를 통일된 형식으로 정리
- **데이터 클리닝 (Data Cleaning):**
    - 데이터 정규화는 식품 이름과 칼로리 정보를 통일된 형식으로 변환하여 AI 모델의 학습에 일관성을 부여
    - 클리닝 과정에서는 누락된 데이터나 이상값을 제거하거나 보완

4. 데이터 셋 확보 전략 (Dataset Acquisition Strategy)

- **공공 데이터셋 활용:**
    - 공공기관의 오픈 데이터 플랫폼(AI-hub)에서 이용 가능한 헬스케어 데이터 셋
- **파트너십:** 의료기관, 연구기관과의 협력을 통한 데이터 확보 전략
- **프라이버시 및 보안 고려:** 개인정보보호법 및 관련 규정 준수 방법
    - 데이터 익명화 및 비식별화 방법
- **데이터 활용 동의서 (Consent Form):** 환자나 참여자로부터 데이터 수집 동의를 받는 절차

5. 데이터 저장 및 관리 (Data Storage and Management)

- **클라우드 기반 저장소:** AWS, Azure와 같은 클라우드 플랫폼 활용
- **데이터베이스 선택:** 구조화된 데이터와 비구조화된 데이터 저장

6. 결론 (Conclusion)

- 데이터 수집 및 확보의 중요성 요약
    - 현재 데이터에 대한 제한(사진 5000장)으로 인해 음식 이미지의 학습에 제한사항이 존재.
    - 이후 좀 더 확장된 저장소 및 음식 데이터 학습으로 음식 종류의 다양화 필요
- 향후 데이터 활용 및 분석 방안
    - 향후 음식 데이터에서 Kcal 정보 외에도 여러가지 정보를 얻어 일일 섭취량 추천 및 식단 추천