데이터 전처리 및 정제 과정 설계 및 구현

Aiden.ji(지윤수)

Luna.jung(정지연)

1. 개요

- **개인정보 보호:** 입력받는 개인 데이터가 무엇인지에 따라 개인정보보호법이나 GDPR 등을 준수해야한다. 이름, 나이, 성별 등의 정보가 포함된다면 익명화나 암호화 등 보안 조치가 필요

- **음식 사진 처리:** 음식 사진은 머신러닝이나 딥러닝 알고리즘을 통해 자동으로 음식 종류를 식별하고 칼로리를 추정하는 과정에 사용해야 한다. 사진의 품질을 일정하게 유지하고, 사진을 분석할 수 있는 알고리즘의 정확성을 높이는 방법을 고려해야 함

- **Kcal 데이터의 정확성:** 저장 되어있는 데이터에서 1인분, 2인분 등의 양의 차이에 따라 변화할 수 있기 때문에 Kcal 데이터를 각각 변환할 수 있도록 한다

- **데이터셋의 확장 가능성:** 현재는 간단한 데이터와 음식 사진만 다루고 있지만, 향후 추가적으로 더 복잡한 데이터를 다루게 될 수 있으므로 음식에 대한 정보를 어디까지 제공할 지 기준을 정해둘 것

2. 데이터 수집 목표

- **데이터 품질 향상:** 체계적인 수집 방법을 통해 데이터의 정확성, 일관성, 신뢰성을 확보.
- **분석 효율성 향상:** 필요한 데이터를 빠르고 정확하게 수집하여 이후 분석 과정에서 효율적으로 활용.

3. 데이터 수집 원칙

- **적합성:** 분석 목적에 맞는 데이터를 수집.
- **정확성:** 신뢰할 수 있는 소스에서 데이터를 수집하고, 중복 및 오류를 최소화.
- **완전성:** 필요한 모든 변수를 수집하여 누락된 데이터 없이 완벽한 데이터셋 구축.

- **일관성:** 동일한 규칙과 형식에 맞추어 데이터를 정규화하여 일관성 있게 통합.

4. 데이터 수집 방법

- **1차 데이터 수집:** 직접적인 입력, 센서 데이터 등을 통해 수집된 원시 데이터.
    - **예시:** 환자의 생체 신호, 일일 운동량, 음식 섭취 기록 등.
- **2차 데이터 수집**: 추후 서비스를 확장할 때 결정

5. 데이터셋 확보 과정

- **데이터 요청 및 접근 권한:** 공공 기관 또는 파트너 기관으로부터 데이터 접근 권한 확보 절차 설명.
- **데이터 수집 도구:** 데이터 크롤링, API, 센서 장치 등의 활용 방안.
- **데이터 수집 및 저장:** 수집된 데이터를 안전하게 저장하고, 접근성을 고려한 데이터 저장소 구성.

6. 데이터 정제 및 전처리 과정 (추후 추가 작성 필요)

- **중복 데이터 제거:** 중복된 이미지 식별 및 제거.
- **데이터 표준화:** 단위 통일, 날짜 형식 변환 등 데이터 형식 통일 작업.

7. 데이터 품질 관리

- **데이터 검증:** 수집된 데이터의 품질을 검토하고, 오류가 없는지 확인.
- **일관성 검사:** 다른 소스에서 수집된 데이터와의 일관성 확인.
- **정확성 검사:** 수집된 데이터가 실제 값을 잘 반영하는지 검토.