1 데이터 특성 분석

1) 독성 데이터 분포 분석
   가) 수집한 독성데이터가 numerical data인 경우
      (1) 독성 데이터의 히스토그램 등 기초통계량 분석
   나) 수집한 독성데이터가 categorical data인 경우
      (1) 독성 데이터의 비율 분석 (얼마나 불균형 한지 등)

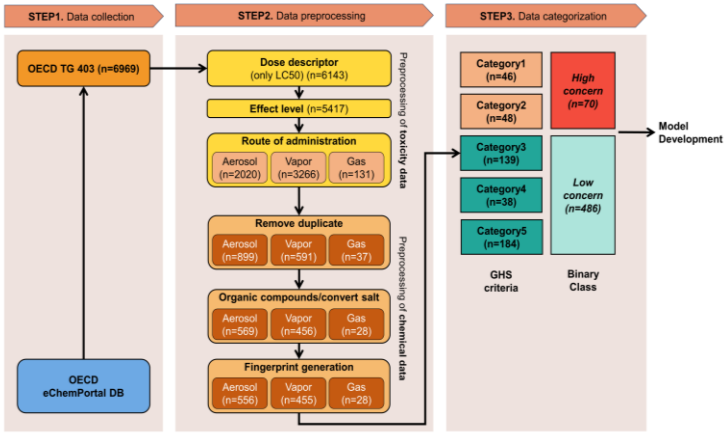
2) 화학물질 구조 유사성 분석
   가) Tanimoto coefficient 등 화학물질 간 유사도 계산

2 데이터 전처리

1) 독성 데이터 처리
   가) Dose descriptor 선정
   나) Effect level 데이터 없는 것 제외
   다) 노출 경로에 따른 데이터 분할

2) 화학물질 데이터 처리
   가) 중복값 제거
   나) 염, 금속, 이온 물질제거

3) 화학물질 독성 카테고리 분류



3 분자지문 생성

1) SMILES를 화학물질 분자지문으로 변환 RDKit
2) 분자지문 생성되지 않는 물질 제거