|  |  |
| --- | --- |
| Research title | 주요 독성 (hERG 외 3종) 데이터를 활용한 모델 구축 및 평가 |
| Name | 고강빈 |

1. 논문 데이터 수집

1) hERG / DILI 데이터 수집

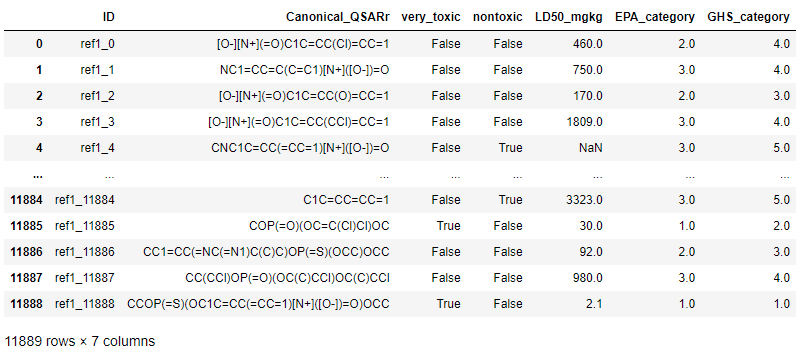
2) Acute Oral Toxcity 데이터 수집

Gadaleta et al, 2019에서는 rat acute oral systemic toxicity (LD50) 관련 데이터셋을 공개하고 있다. 해당 데이터는 8,994개의 Training set과 2,895개의 Evaluation set으로 이루어져 있으며, NICEATM과 NCCT의 Cheminformatics community에서 수집할 수 있다.

데이터 내 Feature의 총 개수는 13개로, 그 중 SMILES와 활성정보를 선택적으로 수집하였다. 처리 후의 Feature의 개수는 ID를 포함하여 총 7개로, 각각에 대한 설명은 다음과 같다.

|  |  |  |
| --- | --- | --- |
| LD50\_mgkg | mg/kg 단위의 연속적 값 | |
| very\_toxic | LD50 < 50mg/kg (TRUE) | |
| nontoxic | LD50 > 2000 mg/kg (TRUE) | |
| EPA\_category | 1 | LD50 ≤ 50 mg/kg |
| 2 | 50 < LD50 ≤ 500 mg/kg |
| 3 | 500 < LD50 ≤ 5000 mg/kg |
| 4 | LD50 > 5000 mg/kg |
| GHS\_category | 1 | LD50 ≤ 5 mg/kg |
| 2 | 5 < LD50 ≤ 50 mg/kg |
| 3 | 50 < LD50 ≤ 300 mg/kg |
| 4 | 300 < LD50 ≤ 2000 mg/kg |
| 5 | LD50 > 2000 mg/kg |

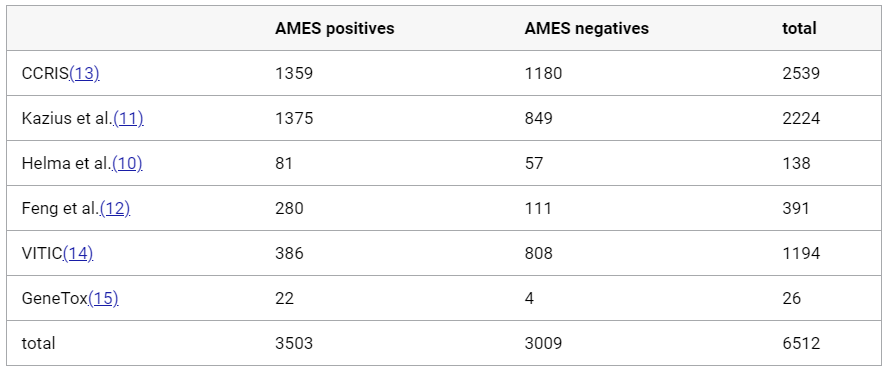
TS와 ES는 해당 논문 내에서 모델링에 활용하기 위해 나눈 것이므로 본 연구에서는 병합하여 하나의 포맷의 데이터로 만들었으며, ID는 ref1\_(index)의 형태로 추가하였다.



3) AMES Mutagenecity 데이터 수집

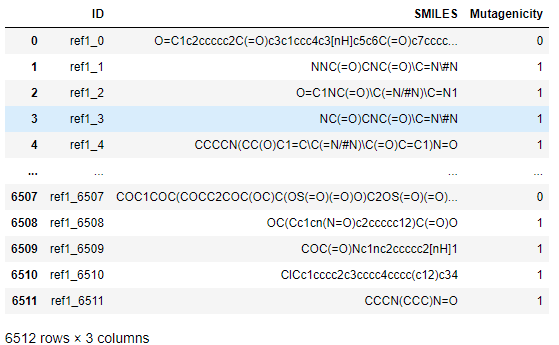
3-1) Hansen et al, 2009

해당 논문에서는 Ames Mutagenicity Benchmark Data Set을 공개하고 있다. 이 Dataset은 CCRIS, VITIC 등 총 6개의 source로부터 수집되었으며, 각각에 대한 자세한 내용은 다음 표와 같다.



각 source로부터 수집된 데이터를 하나의 테이블로 병합 후 서로 다른 source간 중복 데이터를 제거하였다. 또한 extraordinary / inorganic molecules 20개가 추가로 제외되었으며, DEREK 또는 MultiCASE 내부 데이터와 모순된 결과를 보이는 25개의 데이터를 제거하여 최종적으로 6,512개의 데이터를 얻었다.

이렇게 얻어진 데이터 중 SMILES와 Mutagenicity column을 분리하였으며, ID column을 추가하였다. ID는 ref1\_(index)의 형태로 설정되었다.

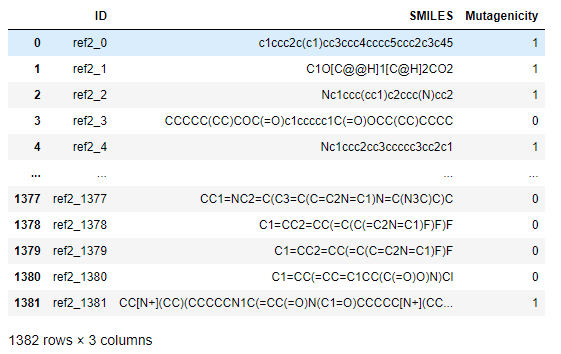


3-2) MutagenPred-GCNN. 2021

해당 논문에서는 총 2개의 데이터셋, Training set과 External validation set을 공개하고 있다. 이 중 TS의 경우 위 Hansen에서 얻어진 데이터와 같은 데이터이므로 ES만을 추가로 수집하였다.

이는 CCRIS, NTP, ISSSTY로부터 수집되었으며, 703개의 mutagens와 680개의 non-mutagens 데이터를 합하여 총 1,383개의 데이터를 가진다.

마찬가지로, 위 데이터 중 SMILES와 Mutagenicity column을 분리하여 ID를 추가하였다. 해당 ID는 ref2\_(index)의 형태를 가진다.

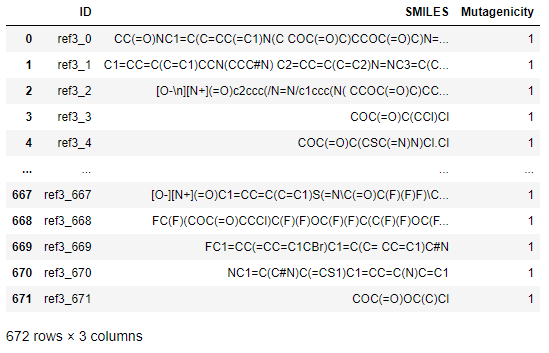


3-3) AMES/QSAR International Collaborative Study

총 12,140개의 화합물에 대하여 Mutagenicity 확인이 이루어졌으며, 이 중 Ames mutagenicity와 강한 association을 가지는 class A 화합물을 공개하고 있다.

각 phase별로 약 4,000개의 화합물이 검사되었으며, 이 중 각각 183개, 253개, 236개로 총 672개의 class A 화합물이 선별되었다.

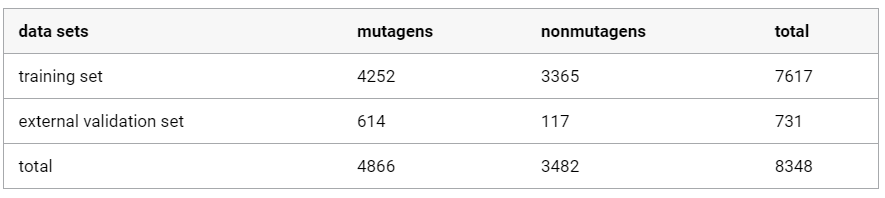
마찬가지로 이로부터 SMILES와 Mutagenicity column을 분리하여 ref3\_(index) 형태의 ID를 추가하였다.



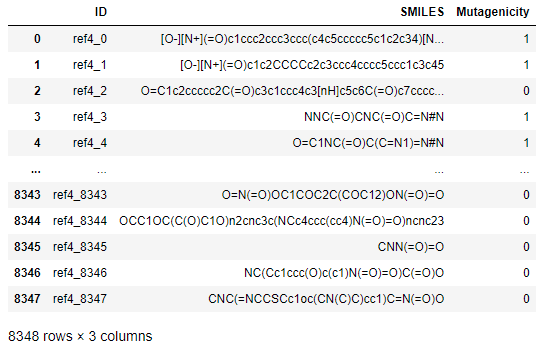
데이터의 통일성을 위해 class A를 Mutagenicity 1로 다시 labeling하였다.

3-4) Xu et al, 2012

해당 논문에서는 Training set, External validation set을 공개하고 있다. TS는 총 4개의 논문(Cheng et al 2011, Ewing et al 2006, Wang et al 2012, Chen et al 2011)으로부터 수집되었으며, ES는 Lazar toxicity predictions 웹사이트로부터 수집되었다. 각각에 대한 내용은 다음과 같다.



전체 데이터에 대하여 잘못된 SMILES와 중복된 데이터가 제거되었으며, E 또는 Z configuration이 부정확한 데이터와 inorganic compounds가 추가로 제거되었다. 마지막으로 무게가 40 미만 혹은 800 이상인 molecules와 tautomers를 제거하여 최종 데이터를 얻었다. 실험의 정확도를 위해 최종 데이터 중 여러 source에 대해 experimental mutagenicity data가 여러 종류인 화합물은 제외되었으며, 비슷한 수치의 experimental mutagenicity data를 가지는 경우 하나만을 남기고 나머지는 제거하였다.

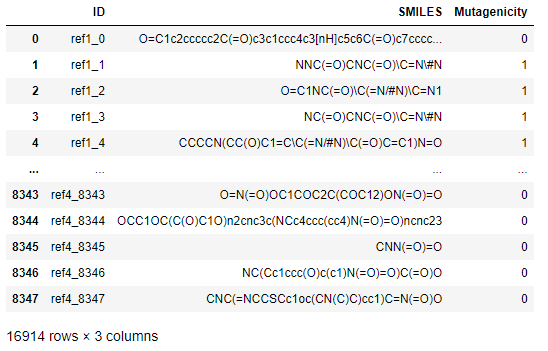


데이터의 통일성을 위하여 mutagens로 label된 데이터는 1로, non-mutagens로 label된 데이터는 0으로 다시 labeling하였다. 최종적으로 분리된 데이터에 ref4\_(index)형태의 ID를 추가하였다.

3-5) AMES 최종

위에서 얻은 데이터들을 하나의 csv로 병합하여 AMES\_total을 만들었다.

총 데이터의 개수는 16,914개로, 9,734개의 mutagens와 7,171개의 non-mutagens 데이터를 가진다.



2. Canonicalization

3. 데이터 테이블 병합 및 중복처리

Canonicalization된 SMILES를 기준으로 중복된 값의 레이블을 처리한다. 동일 SMILES에 대한 Label의 값이 다를 경우 다음과 같이 처리한다.