

팀명: PNU-Saja Boys

팀원: 202155127 김태경, 202155151 신훈교, 202155160 유진영

[학습 데이터 전처리 방식]

녹는점 데이터 중 가장 큰 데이터 셋인 Bradley 오픈 데이터로 학습을 시켰다. 맨처음엔 canonical smile처리를 했다. 그 후 데이터를 살펴보니 같은 물질이라도 출처에 따라(실험에 따라) 녹는점의 차이가 0.5~5정도 났다. 그래서 평균값 처리했다. 근데 여러 화학자들이 수기로 쓰는 녹는점이라서 끓는점을 녹는점으로 잘못 적힌 값들이 있었다. 그 값들은 서칭(구글, pubchem, TCI, Sigma-Aldrich 등)으로 녹는점으로 바꿔줬다. 이성질체가 있는 물질들은 동일한 이성질체라도 cis/trans상태에 따라 녹는점이 다르기에 smile에 있는 괄호로 cis인지 trans인지 판단해서 녹는점을 서칭해서 적었다.

[model]

모델은 extra-regression-tree를 활용하여 학습하였다. 이 ML 모델은 트리를 앙상블하는 기법인데, RF와의 차이는 feature split을 랜덤으로 주어 서로 다른 트리를 학습하게 하여, 일반적인 RF보다 속도도 빠르고 분산도 작다. 또한, 부트스트랩을 활용하지 않기 때문에 bias도 작기 때문에, 이 모델을 사용하여 학습을 진행하였다.

[Feature]

RDkit으로 스마일에서 뽑아내 줄 수 있는 거의 모든 피처를 다 뽑아주었다.

atom Pair FP, Topological Torsion FP, RDKit FP, Avalon FP, MACCS Keys, MolWt, MolLogP, TPSA, NumHDonors, NumHAcceptors, NumRotatableBonds 등등과 3D 관련 피처를 추출하여, 모델이 최대한 많은 피처를 확인하여 잘 예측할 수 있도록 설계하였다.