DrivenData 캐글처럼 데이터분석과제 올라오는 사이트!

[Flu Shot Learning: Predict H1N1 and Seasonal Flu Vaccines](https://www.drivendata.org/competitions/66/flu-shot-learning/page/210/)

**문제상황**

코로나 백신개발과 보급이 가장 전세계 주요 이슈인 요즘, 코로나 이전 가장 유행했던 감염질환인 신종플루 백신과 계절독감 백신을 맞은 사람들의 survey 데이터를 통해 두 가지 백신을 맞았는가 / 하나만 맞았는가 / 둘다 안 맞았는가를 예측하는 multilabel 분류 문제

**특이점**

* **주어진 데이터의 feature들은 전부 범주형 데이터 (이항변수 또는 다항변수)**
* **신종플루와 계절독감 백신 접종여부를 모두 예측해야 하므로 Y값이 두개인 multi-label classification문제로 multi category와는 다른 개념**
* **이항 분류예측 모델을 두개 만들 수도 있지만 두 Y변수간에 의존성이 존재하는 것을 고려하는 다른 multi label 분류 방법이 따로 존재하는 것 같아서 좀 더 새로운 머신러닝 모델링 적용 가능할 듯!**
* **최종 예측 결과는 분류 라벨이 아니라 확률이어야 함**

**걱정되는 점**

* **X변수 Y변수 모두 범주형 변수인 설문조사 데이터라는 점에서 작년 범주와 겹치지 않을까 하는 점이 걱정됨. 작년 범주팀은 결측치 처리에 힘을 줬다면, 이번에는 다중 분류 모델이라는 점에 집중해보려고 함**
* **결국은 단순 예측 위주로 갈까 걱정됨 – 단순히 성능 올리기에 집중하기보다 다양한 EDA, 시각화 통해서 데이터셋 이해하기 + 사후 모델링 해석 방법들도 적용해보려고 함!**

**분석흐름 (아직은 아이디어)**

2주차:

* 데이터 전처리(NA처리, 범주형 변수 인코딩 등)
* 시각화
* 클러스터링 통해서 유형 나눠볼 수 있지 않을까
* 클래스 불균형 해결 위해 샘플링 방법 적용?

4주차:

* 모델링!
* 다중라벨분류 방법들 적용
  + - Problem Transformation : 일반 이항분류 모델 2가지 적용
    - Adapted Algorithm
    - Ensemble approaches
* 모델링 해석 방법들 적용? (가능하다면)