당뇨병 예측 모델

1. 프로젝트 제목

- 바이오 마커(Bio-Marker)를 이용한 당뇨병 예측 모델

2. 프로젝트 시작 계기

- 당뇨병은 세계 10대 사망 질환, 연간 400만명이 사망하고 연 800조원의 치료비가 발생한다.
- 총 10783의 feature 를 가지고 feature reduction 과정을 통해 , 최소한의 정보로 정확도가 높은 예측을 하고 싶습니다.

3. 프로젝트 개요

- 출처 (공공데이터-HMPdacc(NIH)에서 다운)
- "RF", "AdaBoost", "GradientBoost", "XGBoost" 모두 써보고 auc 가장 높은 것으로 선택할 예정
- 각 모델의 철학은 이러함
- RandomForest: 훈련과정에서 다수의 결정트리로부터 분류 또는 평균 예측치를 출력함으로써 동작한다.
- AdaBoost(adaptive boosting): 약한 학습기(weak learner, ex: 결정트리 학습법)의 결과물들을 가중치를 두어 더하는 방법으로 가속화 분류기의 최종 결과물을 표현할 수 있다. 이전의 분류기에서 잘못 분류된 것을 약한 학습기를 이용해 수정할 수 있음을 이용해 다양한 상황에 적용(adaptive)할 수 있다. 이상점이 많은 데이터에 취약하나 과적합 문제에는 덜 취약하다.
- GradientBoost: stump나 tree 가 아닌 single leaf에서 학습을 시작한다. 타겟값에 대한 초기 추정 leaf이 tree를 타고가면서 error를 반영한 새로운 tree 를 만든다. 회귀 분석 혹은 분류 분석을 수행하는데 유용하며 머신 러닝 알고리즘 중에서 가장 예측 성능이 높다고 알려져 있다.
- XGBoost: Gradient Boosting 알고리즘을 분산 환경에서 실행할 수 있도록 구현해놓은 라이브러리이다. Regression, Classification 문제를 모두 지원하며, 성능과 자원효율이 좋아서 인기있는 알고리즘이다.

4. 기대효과

- 당뇨병 조기 발견, 치료가 가능해진다.
- 당뇨병은 예측하기 쉬운 질병이지만 방치하는 경우가 많다. 발견만 일찍한다면 사망률과 치료비 감소의 효과를 얻어낼 수 있을 것이다.