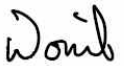
2021학년도 기초과학융합연구소 학부생 동계 인턴십 결과보고서

○ **연구주제**	부스팅 기법을 사용하여 암 질병률의 연관성을 분석 및 예측한다.		
○ **성명**	서영석	○ **학번**	20171297
○ **지도교수**	정원일 교수님  (서명/인)		

1. 연구목표

- 편향과 지도학습의 차이를 줄이기 위한 부스팅 기법을 사용하여 COVID-19 및 각종 질병률을 예측
- 설명변수에 따라 달라지는 질병률과 연관성을 분석하고 연구

2. 연구내용

각종 질병 중 'LUNG'(폐암)을 선택하여 연구를 진행했다.

1) 데이터 전처리

- 처음 phenotype에서 많은 데이터를 사용하기 위해 전처리 작업에서 1) 결측값이 많은 MDM_B, MHTN_B, MLPD_B, PHTN_B, PDM_B, PLPD_B 를 제거했다. (단순 결측값이 많기 때문이 아닌 LUNG과의 상관관계, 연관성이 적다고 생각하여 제거했다.)
- 가족력과 과거력 데이터인 PCAN_00, FCAN_00 데이터에서 변수가 1인 개수 중 LUNG인 경우의 확률이 높은 PCAN80 (0.33), FCAN80 (0.053)을 넣었고 그 외 데이터는 확률이 0 혹은 0.02 보다 낮게 나와 제거하였다. PCAN80, FCAN80에서 결측값은 모두 0으로 대체하였다.
- ALCO_B(음주량), SMOKA_B(흡연량)이 있기 때문에 ALCO_AMOUNT_B, SMOKA_MOD_B는 겹치는 데이터이기에 제거하였다.
- HT_B(신장), WT_B(몸무게)를 이용하여 BMI (WT/(HT * HT)*10000)의 피쳐를 만들어 추가하였고 HT_B와 WT_B를 제거한 뒤 진행하였다.
- FVC와 FEV1은 폐와 관련된 피처지만 약 40~50%의 결측값을 가지므로 FEV1은 제거, FVC의 경우 넣은 결과값과 뺀 결과값으로 나누어 진행하였다.

Phenotype에서는 총 23개(24개)를 사용하였고 내용은 아래와 같다.

AGE_B SMOK_B ALCO_B EXER_B SBP_B DBP_B CHO_B LDL_B TG_B HDL_B FBS_B GOT_B GPT_B GGT_B URIC_B BIL WBC CREAT LUNG SEX1 CRC PCAN80 FCAN80 BMI (FVC)

2) 데이터 피처 수

Train과 Test 데이터는 8:2로 진행하였고 phenotype과 SNPs를 합친 뒤 Lasso를 이용하여 300-400개의 피처 수를 뽑았다.

- 1) FVC 존재

데이터의 수는 총 8763개고 그 중 LUNG의 개수는 232개이다. 약 0.026%이다. 그 중 train 데이터는 7010개, test 데이터는 1753개이다.

phenotype을 고정시킨 뒤, Lasso로 사용한 피처의 수는 330개(SNPs)+24개(phenotype)으로 총 354개이다.

2) FVC 존재 X

데이터의 수는 총 13505개고 그 중 LUNG의 개수는 350개이다. 약 0.026%이다. 그 중 train 데이터는 10804개, test 데이터는 2701개이다.

phenotype을 고정시킨 뒤 Lasso로 사용한 피처의 수는 401개(SNPs)+23개(phenotype)로 총 424개이다.

3) 모델링 작업

모델링은 RandomForestClassifier, DecisionTreeClassifier, KNeighborsClassifier, AdaboostClassifier, XGBClassifier, LGBMClassifier로 총 6개의 모델을 사용했다.

Optuna(Automl 기법)을 통해 분류 모델의 파라미터를 자동적으로 지정하여 AUC를 높였다.

1) FVC 존재

Algorithm	Best trial	Best AUC score
Random forest	'max_depth': 6, 'max_leaf_nodes': 157, 'n_estimators': 162	0.703
Decision tree	'max_depth': 3, 'max_leaf_nodes': 970	0.714
KNeighbors	'n_neighbors': 182, 'leaf_size': 184	0.657
Adaboost	'n_estimators': 375	0.668
XGBoost	'n_estimators': 157, 'min_child_weight': 156	0.761
LGBM	'n_estimators': 59, 'max_depth': 866	0.700

2) FVC 존재 X

Algorithm	Best trial	Best AUC score
Random forest	'max_depth': 2, 'max_leaf_nodes': 305, 'n_estimators': 310	0.745
Decision tree	'max_depth': 6, 'max_leaf_nodes': 867	0.764
KNeighbors	'n_neighbors': 157, 'leaf_size': 156	0.679
Adaboost	'n_estimators': 59	0.702
XGBoost	'n_estimators': 833, 'min_child_weight': 212	0.767
LGBM	'n_estimators': 182, 'max_depth': 182	0.754

3) 최종 AUC 결과

AUC	Algorithm	Clinical + SNPs (FVC 존재 354개)	Clinical + SNPs (FVC x. 424개)
SNPs after LD pruning	Random forest	0.703	0.745
	Decision tree	0.714	0.764
	KNeighbors	0.657	0.679
	Adaboost	0.668	0.702
	XGBoost	0.761	0.767
	LGBM	0.700	0.754

결론은 XGBoost가 0.761,0.767로 가장 높은 예측치를 보이는 모델임을 볼 수 있었고, FVC가 존재하지 않는 피처에서 가장 높은 AUC 값(0.767)을 얻음을 확인할 수 있었다.

3. 후속연구 계획

Deep learning(Classification)이나 AutoEncoder를 이용한 방법으로 후속연구를 진행하고, 피쳐

들의 정교함을 추가하게 된다면 AUC를 높일 수 있지 않을까하는 바람이 있다.