2023-09-13

SEMINAR: Classification for COPD

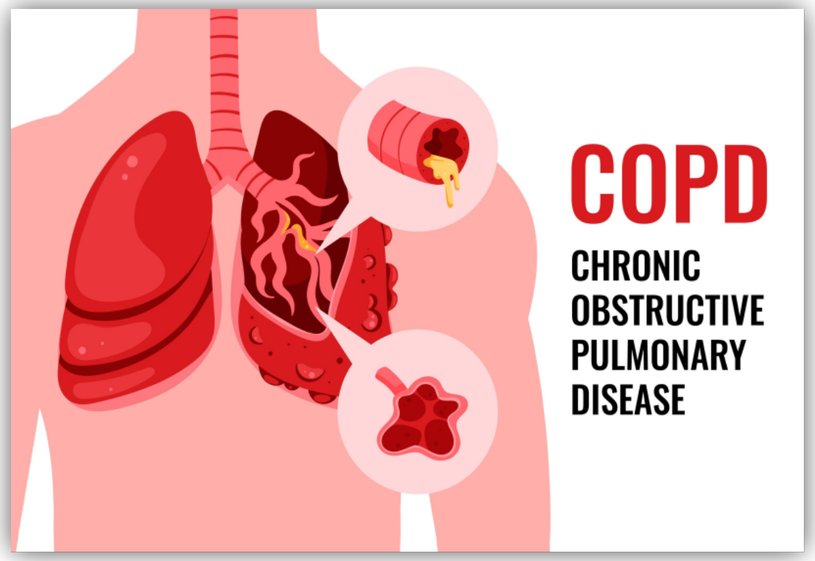
Ungsik Kim (Computer Science Undergraduate)

Research Objective

연구 목적

- COPD 소개
- 연구 목적 및 중요성

COPD 소개



COPD는 폐 건강을 해칠 수 있는 만성 폐질환으로,
폐 기관의 공기 통로가 좁아져 호흡 곤란과 폐 기능 저하를 일으키는 질환

연구 목적 및 중요성

"머신러닝을 통해 COPD 발병을 예측 및
COPD에 영향을 미치는 요인들을 분석"

Experimental Setup

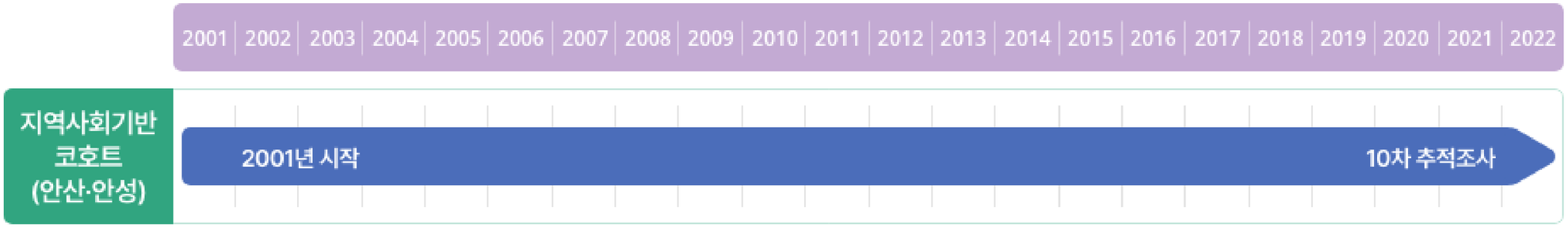
연구 방법 계획

- 데이터 수집 방법 (영양 및 유전데이터)
- Classification 도구 (SVM, KNN, ...)

데이터 수집 방법

코호트 현황

- KoGES는 40세 이상의 일반인구 집단을 대상으로 구축한 '일반인 기반(population-based) 코호트'와 만성질환의 유전-환경 상호작용 위험요인 규명을 위한 '유전-환경(gene-environment) 모델 코호트'로 구성됩니다. 2001년부터 약 23만 5천 여명 규모의 기반조사 참여자를 모집하였으며 2~4년 주기로 코호트 참여자를 재접촉하여 조사 및 검진을 수행하는 반복 추적조사를 실시하였습니다. 현재는 통계청 사망자료, 건강보험공단 수진자료 및 암센터 암등록자료 등의 연계를 통한 수동 추적조사도 병행하여 실시하고 있습니다.



Classification 도구

전통적 머신러닝 기법

- KNN
- SVM

앙상블 기법

- RandomForest
- XGboost

Data Processing

데이터 처리

- 결측치 처리
- 데이터 전처리
- 데이터 시각화 예시

결측치 처리

연속형 변수

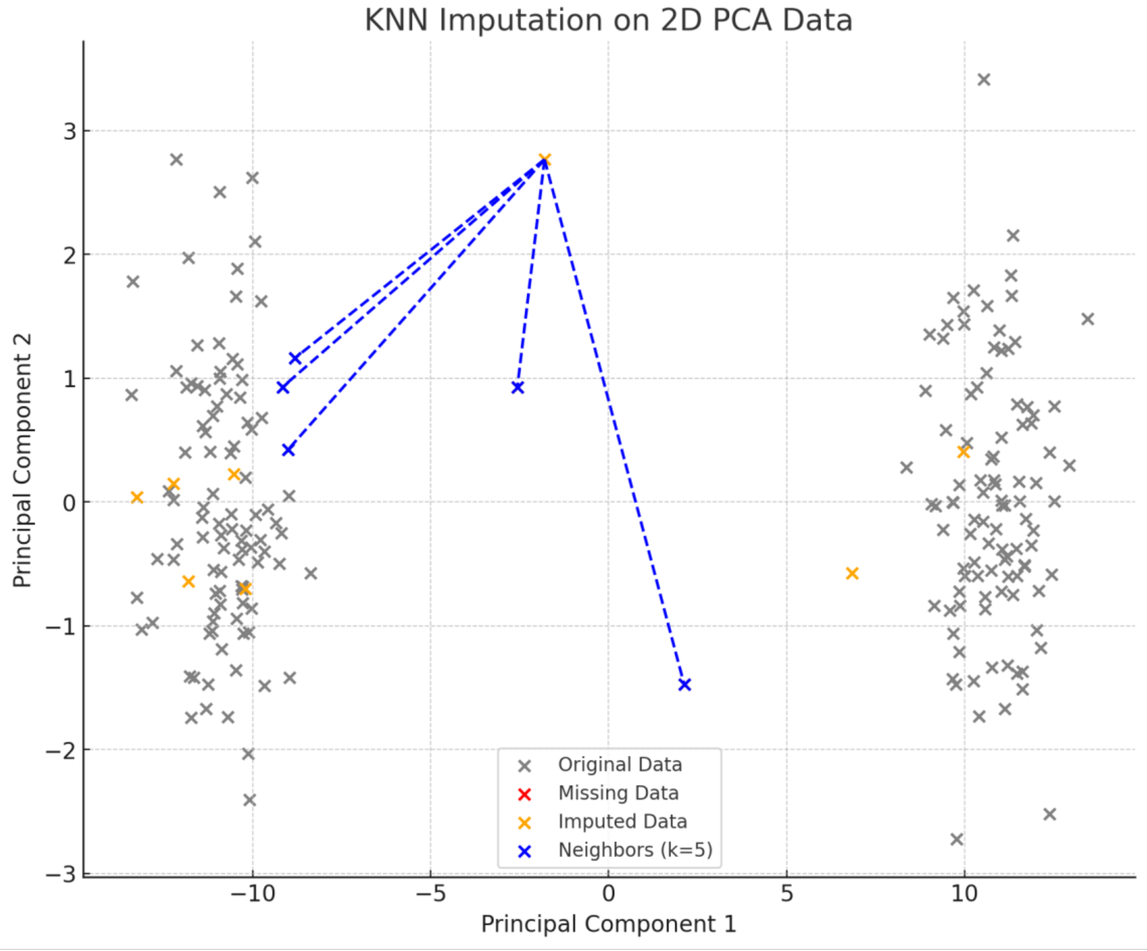
KNN Imputer로 보간하여 처리

범주형 변수

최빈값을 이용하여 처리

유전체 데이터

결측값을 하나의 의미로 보고
"0 0"으로 처리



데이터 전처리

Data Split

- 6 : 2 : 2 (Train : Validation : Test)

Scaling

- Robust Scaler

CatBoost Encoder

문자열로 구성된 범주형 데이터를
y값 참고해서 연속형 변수로 변환

데이터 전처리

Data Split

- 6 : 2 : 2 (Train : Validation : Test)

Scaling

- Robust Scaler

CatBoost Encoder

문자열로 구성된 범주형 데이터를
y값 참고해서 연속형 변수로 변환

```
X_all_train, y_all_train, X_all_val, y_all_val, X_all_test, y_all_test = split(df)
temp_df
```

✓ 0.0s

```
Shape of train : (5775, 58)
Shape of validation : (1925, 58)
Shape of test : (1925, 58)
==========================================================
```

데이터 전처리

Data Split

- 6 : 2 : 2 (Train : Validation : Test)

Scaling

- Robust Scaler

CatBoost Encoder

문자열로 구성된 범주형 데이터를
y값 참고해서 연속형 변수로 변환

```python
def scale(name):
    scaler = RobustScaler()
    temp = name.drop(["AS1_COPD"], axis=1)
    scale = pd.DataFrame(scaler.fit_transform(temp))
    scale.index = temp.index
    scale.columns = temp.columns
    return pd.concat([scale, df["AS1_COPD"]], axis=1)
```

```python
df = scale(df)
df
```

✓ 0.0s

DIST_ID	AS1_AGE	AS1_PACKYR	AS1_CHEMJOBDU	AS1_DUSTJOBDU	AS1_SLPAMTM
NIH2308038847	0.1875	0.222222	-0.407407	-0.593750	-1.0
NIH2308676988	-0.3750	0.000000	-0.222222	0.437500	0.5
NIH2308004412	-0.4375	0.000000	-0.629630	0.468750	0.0
NIH2308744669	0.6875	1.904762	0.851852	-0.531250	-0.5
NIH2308018034	0.8750	0.679365	-1.037037	0.250000	0.0
...	...	...	...	...	...
NIH2308853044	0.6250	0.000000	5.925926	4.750000	-1.0
NIH2308650703	0.7500	0.000000	-0.622222	0.718750	-1.0
NIH2308993350	0.9375	0.000000	-0.037037	-0.406250	-1.0
NIH2308295674	1.1875	0.000000	0.074074	-0.281250	0.0
NIH2308896274	0.5625	0.000000	0.777778	-0.359375	-0.5

9625 rows × 55 columns

데이터 전처리 (SMOTE)

Data Split

- 6 : 2 : 2 (Train : Validation : Test)

Scaling

- Robust Scaler

CatBoost Encoder

문자열로 구성된 범주형 데이터를
y값 참고해서 연속형 변수로 변환

DIST_ID	SNP_A_4291320	SNP_A_2284008	SNP_A_2298582	SNP_A_2184029
NIH2308913474	C T	A G	G G	T T
NIH2308248701	C T	A G	G G	C T
NIH2308155196	T T	G G	T G	T T
NIH2308965960	0 0	0 0	0 0	0 0
NIH2308961596	C T	A G	G G	C T

DIST_ID	SNP_A_4291320	SNP_A_2284008	SNP_A_2298582	SNP_A_2184029
NIH2308913474	1.051119	1.050281	1.051126	1.046231
NIH2308248701	1.051119	1.050281	1.051126	1.059083
NIH2308155196	1.056186	1.056824	1.059293	1.046231
NIH2308965960	1.062127	1.062217	1.063129	1.058946
NIH2308961596	1.051119	1.050281	1.051126	1.059083

데이터 전처리

Over Sampling

- SMOTE

Feature Selection

- Pearson Correlation
- Forward Selection

데이터 전처리

Over Sampling

- SMOTE

Feature Selection

- Pearson Correlation
- Forward Selection

```
X_all_smote, y_all_smote = smote(X_all_train, y_all_train)
```

```
length of original data is  5775
Proportion of True data in original data is 5.45%
Proportion of False data in original data is 94.55%
length of oversampled data is  10920
Proportion of True data in oversampled data is 50.00%
Proportion of False data in oversampled data is 50.00%
315
5460
```

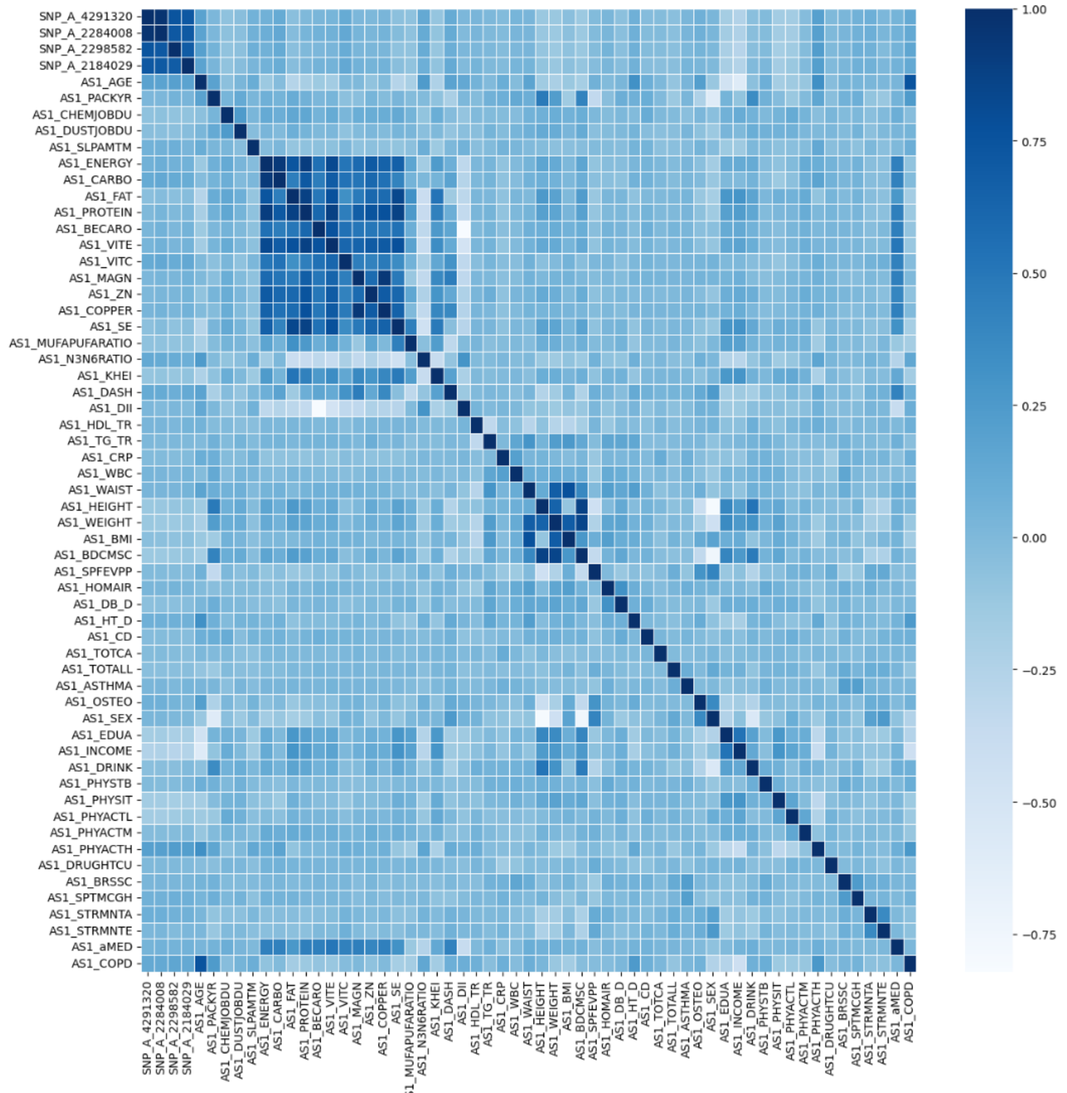
데이터 전처리 (Pearson)

Over Sampling

- SMOTE

Feature Selection

- Pearson Correlation
- Forward Selection



데이터 전처리 (Forward)

Over Sampling

- SMOTE

Feature Selection

- Pearson Correlation
- Forward Selection

```
    dt = DecisionTreeClassifier()
    sfs = SequentialFeatureSelector(dt, n_features_to_select=23)
    sfs.fit(X, y)
    sfs.get_feature_names_out(X.columns)
```

```
array(['SNP_A_2284008', 'SNP_A_2298582', 'AS1_AGE', 'AS1_SLPAMTM',
       'AS1_FAT', 'AS1_TG_TR', 'AS1_HEIGHT', 'AS1_BDCMSC', 'AS1_HOMAIR',
       'AS1_DB_D', 'AS1_HT_D', 'AS1_CD', 'AS1_TOTALL', 'AS1_ASTHMA',
       'AS1_OSTEO', 'AS1_SEX', 'AS1_INCOME', 'AS1_DRUGHTCU', 'AS1_BRSSC',
       'AS1_SPTMCGH', 'AS1_STRMNTA', 'AS1_STRMNTE', 'AS1_aMED'],
      dtype=object)
```

데이터 전처리 (Feature Selection)

- Pearson Corr 기반 예측

- Forward Selection 기반 예측

- Pearson and Forward 기반 예측 (Union)

- Pearson and Forward 기반 예측 (InterSection)

Classification Evaluation

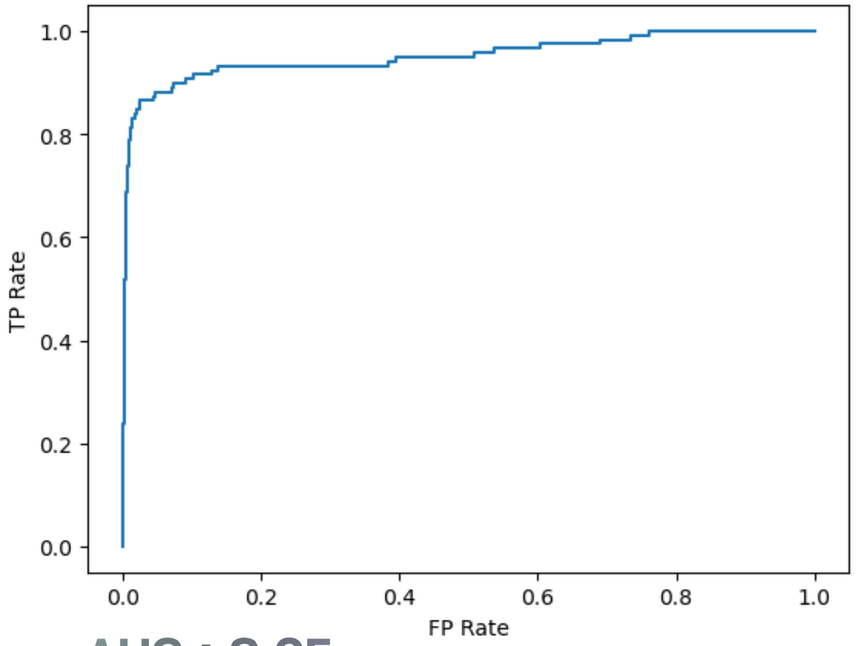
모델 성능 평가

- Confusion Matrix 및 F1-Score
- ROC 곡선

모델 성능 평가

	Accuracy	Precision	Recall	F1-Score
KNN	0.88	0.32	0.80	0.46
SVM	0.90	0.38	0.86	0.53
XGBoost	0.98	0.81	0.82	0.82
Random Forest	0.98	0.79	0.81	0.80

XGBoost	False Predict	True Predict
실제 False	1786	21
실제 True	22	96



AUC : 0.95

Data Analysis

데이터 분석

- 데이터 시각화 및 변수 중요도

데이터 시각화 및 변수 중요도