폐암 데이터 분석

2021년도 2학기 인공지능 변영철 교수님

Woongstar

Positive Vibes

20141108196 컴퓨터공학전공 양지웅

폐암데이터 분석

1. 데이터 이해하기

1. EDA
2. 데이터 표본 확인하기
3. 데이터 특성 상관관계 확인하기

02. 모델 학습하기

1.데이터 전처리
2.모델 만들기
3.더 좋은 모델로 튜닝하기

03. 결론

1. 결론
2. 느낀점

01. 데이터 이해하기

1. EDA
2. 데이터 표본 확인하기
3. 데이터 특성 상관관계 확인하기

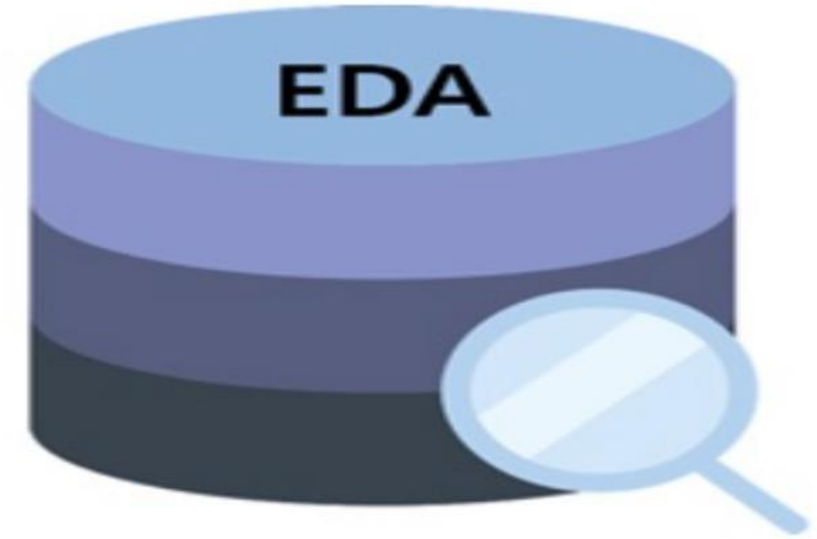
EDA(Exploratory Data Analysis)

탐색적 데이터 분석

1) 정의
수집한 데이터가 들어왔을 때, 이를 다양한 각도에서 관찰하고 이해하는 과정입니다. 한마디로 데이터를 분석하기 전에 그래프나 통계적인 방법으로 자료를 직관적으로 바라보는 과정입니다.

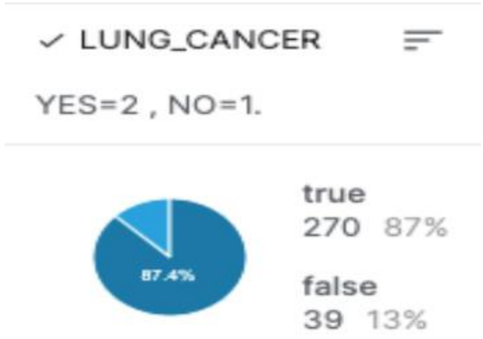
2) 필요한 이유
데이터의 분포 및 값을 검토함으로써 데이터가 표현하는 현상을 더 잘 이해하고, 데이터에 대한 잠재적인 문제를 발견할 수 있습니다. 이를 통해, 본격적인 분석에 들어가기에 앞서 데이터의 수집을 결정할 수 있습니다.

다양한 각도에서 살펴보는 과정을 통해 문제 정의 단계에서 미쳐 발생하지 못했을 다양한 패턴을 발견하고, 이를 바탕으로 기존의 가설을 수정하거나 새로운 가설을 세울 수 있습니다.

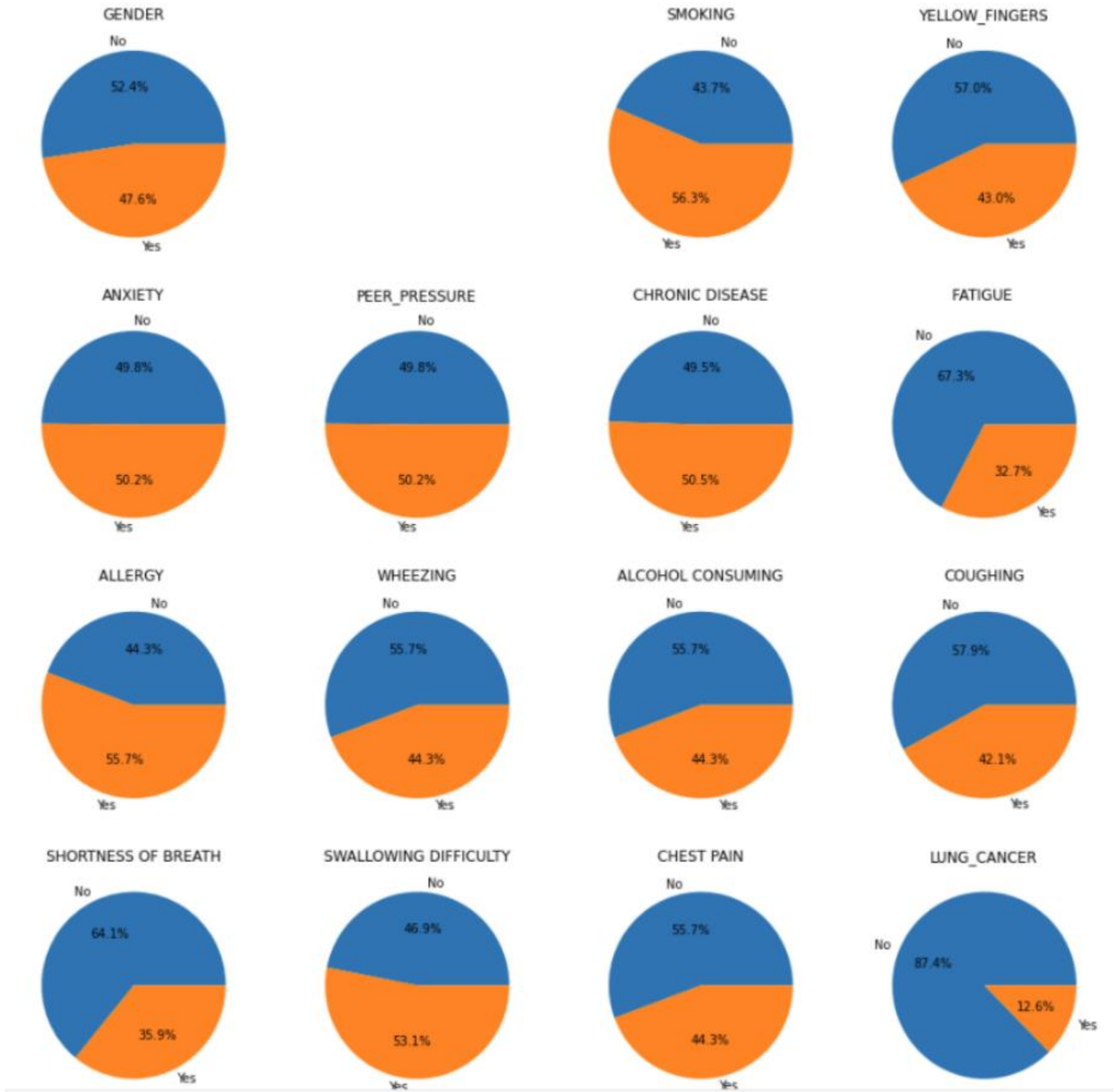


309 data, 16 column

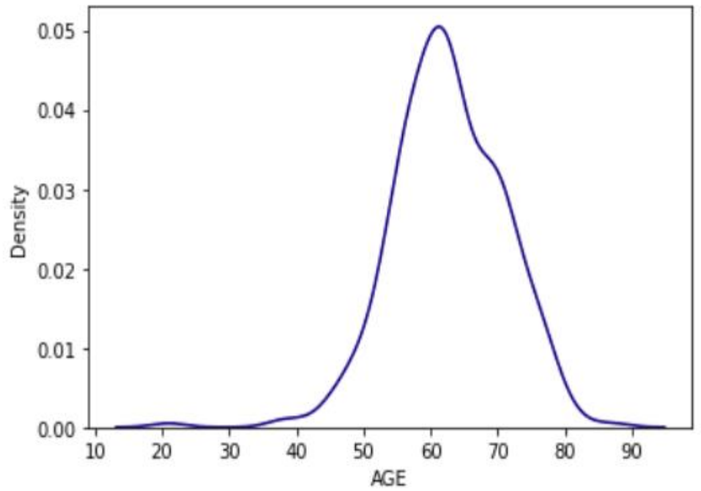
Gender: M(male), F(female)	성별
Age: Age of the patient	나이
Smoking: YES=2 , NO=1.	흡연여부
Yellow fingers: YES=2 , NO=1.	노란색 손가락
Anxiety: YES=2 , NO=1.	걱정
Peer_pressure: YES=2 , NO=1.	동조 압력
Chronic Disease: YES=2 , NO=1.	만성질환
Fatigue: YES=2 , NO=1.	피곤
Allergy: YES=2 , NO=1.	알래르기
Wheezing: YES=2 , NO=1.	쌕색거림
Alcohol: YES=2 , NO=1.	술
Coughing: YES=2 , NO=1.	기침
Shortness of Breath: YES=2 , NO=1.	숨가쁨
Swallowing Difficulty: YES=2 , NO=1.	삼키기 어려움
Chest pain: YES=2 , NO=1.	가슴통증
Lung Cancer: YES , NO.	폐암 여부



데이터 표본 확인하기




```
plt.figure(figsize=(16,16))
for i in range(0,len(data.columns)):
    if i==1:
        continue
    else:
        plt.subplot(4,4,i+1)
        plt.title("{0}".format(data.columns[i]))
        plt.pie(data.iloc[:,i].value_counts(sort=False),labels=['No','Yes'],autopct='%.1f%%')
```



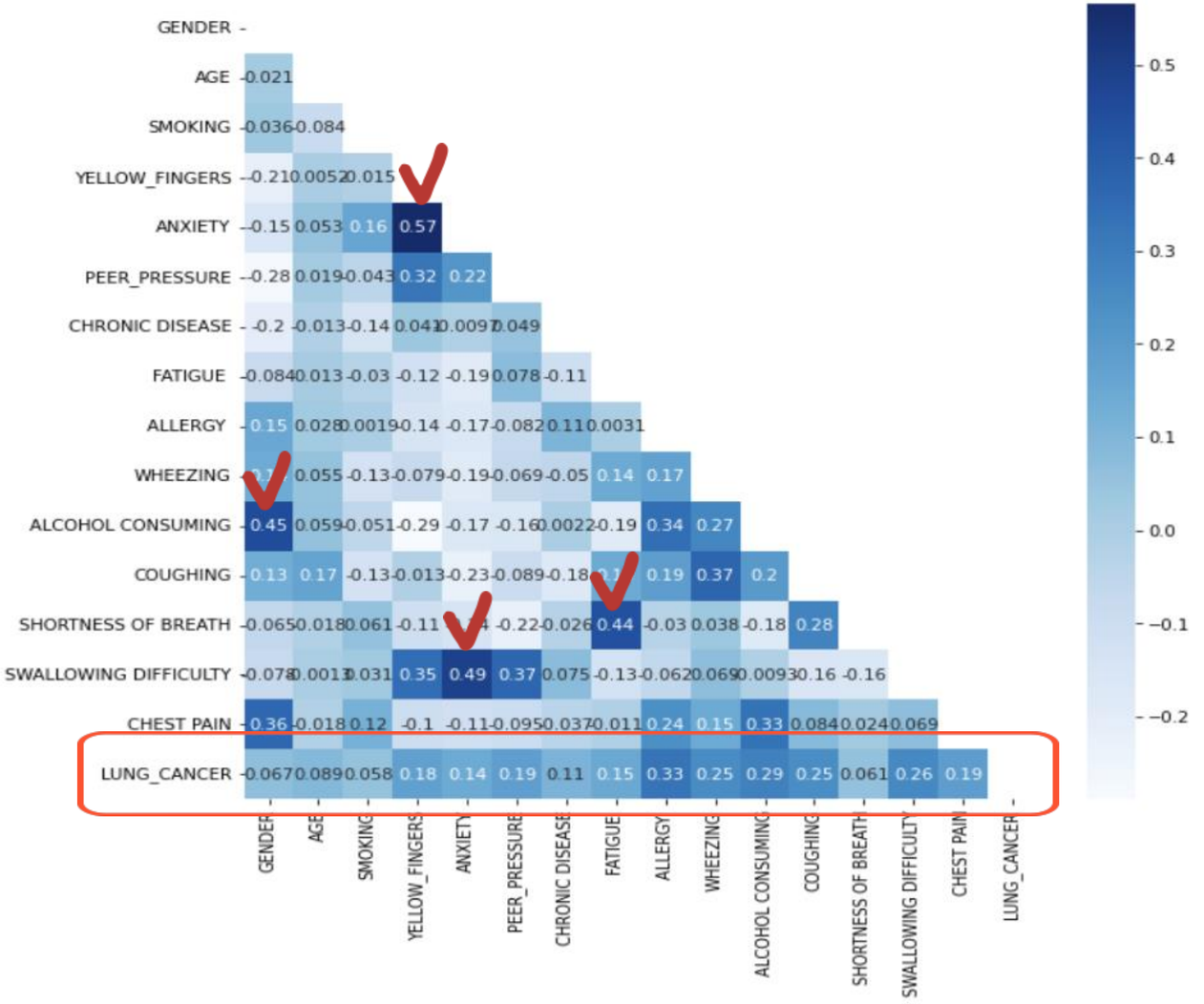

```
sns.kdeplot(data['AGE'],color='navy')
```

제가 분석할 캐글 Lung Cancer 데이터셋의 표본(조사대상)은
대부분 50대 부터 70세 사이의 사람이 많으며
왼쪽 그림과 같은 비율로 특성별 표본이 이루어져 있습니다.

데이터 상관관계 확인하기

```
#heatmap
mask=np.zeros_like(data.corr(),dtype=bool)
mask[np.triu_indices_from(mask)]=True

plt.figure(figsize=(10,10))
sns.heatmap(data.corr(),annot=True,mask=mask,cmap='Blues')
```



상관관계?

두 대상이 서로 관련성이 있다고 추측되는 관계를 말합니다.

상관관계를 왜 먼저 파악하는데?

우리는 여러 변수들이 폐암과 어떠한 관계를 가지는치 찾는 것이 목표입니다.

Pandas 의 corr()함수는 각 변수들 별로 상관관계를 -1에서 1의 값으로 나타내줍니다.

(pandas 라이브러리는 피어슨 상관계수로 계산합니다.)

상관관계를 그려서 얻는게 뭔데?

1. 폐암과 어떠한 특성(독립변수)이 가장 상관관계가 있는지 파악할 수 있습니다. 옆의 그림에서 알수 있듯이 ALLERGY라는 변수가 가장 LUNG_CANCER에 영향을 미치는 변수인 것을 알 수 있었고 두번째로는 ALCOHOL CONSUMING(알코올 섭취) 변수가 영향을 크게 미쳤습니다.

2. 두번째로 다중공선성 문제를 해결하기 위함입니다.
다중 공선성문제는 내가 폐암 유무(종속변수)를 설명할 때 , 높은 상관관계인 특성(독립변수)들로 해당 종속변수를 설명할 경우 오히려 예측확률이 낮아지는 현상입니다.

->우리는 특정 변수만을 가지고 폐암 예측을 하려고 할때에는
Fatigue 변수와 SHORTNESS OF BREATH변수, 혹은
SWALLOWING DIFFICULTY 변수와 ANXIETY 조합의 변수들로
폐암예측을 하면 예측확률이 떨어질 수 있겠구나 등의 정보를 알 수 있습니다.

02.모델 학습

1. 데이터 전처리
2. 모델 만들기
3. 더 좋은 모델로 튜닝하기

데이터 전처리

1. 결측치 확인하기

```
lung_cancer_data.isnull().sum()
```

```
lung_cancer_data = lung_cancer_data = lung_cancer_data.dropna(axis=0)
```

```
GENDER                    0
AGE                       0
SMOKING                   0
YELLOW_FINGERS            0
ANXIETY                   0
PEER_PRESSURE             0
CHRONIC DISEASE           0
FATIGUE                   0
ALLERGY                   0
WHEEZING                  0
ALCOHOL CONSUMING         0
COUGHING                  0
SHORTNESS OF BREATH       0
SWALLOWING DIFFICULTY     0
CHEST PAIN                0
LUNG_CANCER               0
dtype: int64
```

2. 데이터 일치시키기

```
# label encoding
lung_cancer_data.replace({"LUNG_CANCER":{'YES':0,'NO':1}},inplace=True)
# printing the first 5 rows of the dataframe
lung_cancer_data.head(5)
```

```
# label encoding
lung_cancer_data.replace({"GENDER":{'M':0,'F':1}},inplace=True)
# printing the first 5 rows of the dataframe
lung_cancer_data.head(5)
```

Stirng(YES,NO),String(F,M) -> INT

ALCOHOL CONSUMING	COUGHING	SHORTNESS OF BREATH	SWALLOWING DIFFICULTY	CHEST PAIN	LUNG_CANCER
2	2	2	2	2	0
1	1	2	2	2	0
1	2	2	1	2	1
2	1	1	2	2	1
1	2	2	1	1	1

데이터 스플릿, 스케일링

왜 데이터를 스플릿하죠?

내가 만든 모델의 predict 를 계산하기 위해서
모델을 만들기 전에 학습용 데이터와 검증용 데이터를 분리해야 합니다.

```
x_train, x_test, y_train, y_test = train_test_split (X, y , test_size = 0.
2, random_state = 35)
```

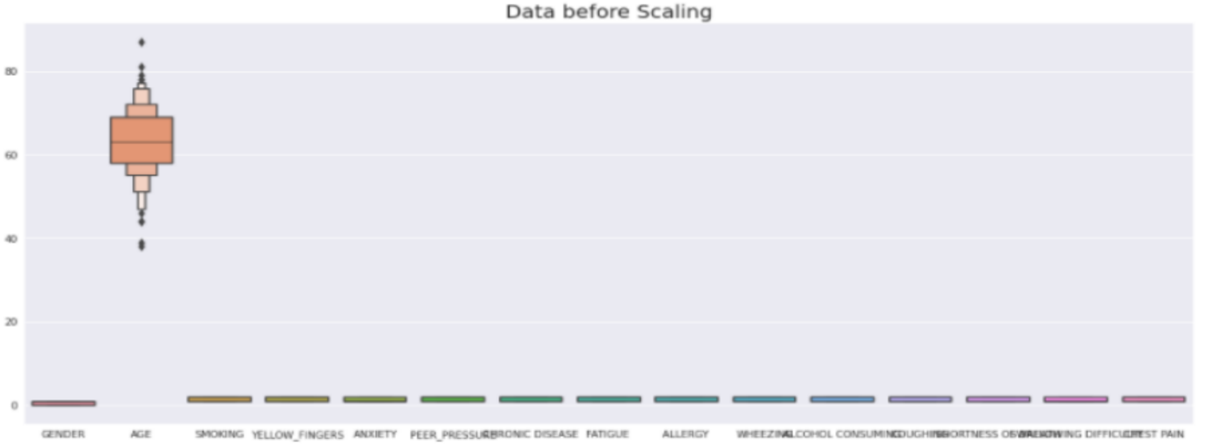
```
len(x_test), len(x_train)
```

```
(62, 247)
```

```
# Scaling data:
scaler = StandardScaler()
x_train = scaler.fit_transform(x_train)
x_test = scaler.transform(x_test)
```

```
plt.figure(figsize=(20,8))
plt.title("Data after Scaling", fontsize = 20, y=1.0)
sns.boxenplot(data=x_train)
plt.show()
```

왜 데이터 스케일링이 필요하죠?



데이터를 스케일링 하기전의 시각화 자료를 보면 AGE이외의 값은 모두 0과 1로 변주형 변수를 표현하고 있지만 AGE의 값은 최대 90까지의 값을 표현하고 있습니다. 이 상태로 모델을 학습시킬 경우 상대적으로 AGE란 변수에 많은 가중치를 두어서 학습되버릴 수 있습니다.
저는 StandardScaler함수를 이용해서 평균 0 표준편차 1로 모든 변수를 스케일링 하였습니다.

모델 학습시키기

```
models = [("LR", LogisticRegression(max_iter=1000)),
          ("SVC", SVC()),
          ("KNC", KNeighborsClassifier(n_neighbors=10)),
          ("DTC", DecisionTreeClassifier()),
          ("GNB", GaussianNB()),
         ("SGDC", SGDClassifier()),
         ("Perc", Perceptron()),
          ("NC", NearestCentroid()),
          ("Ridge", RidgeClassifier()),
          ("BNB", BernoulliNB()),
          ("RF", RandomForestClassifier()),
          ("ADA", AdaBoostClassifier()),
          ("XGB", GradientBoostingClassifier()),
          ("PAC", PassiveAggressiveClassifier())

]
    for name, model in models:
        model.fit(x_train, y_train)
        model_results = model.predict(x_test)
        score= precision_score(y_test, model_results, average='macro')
        results.append(score)
        names.append(name)
        finalresults.append((name,score))
```

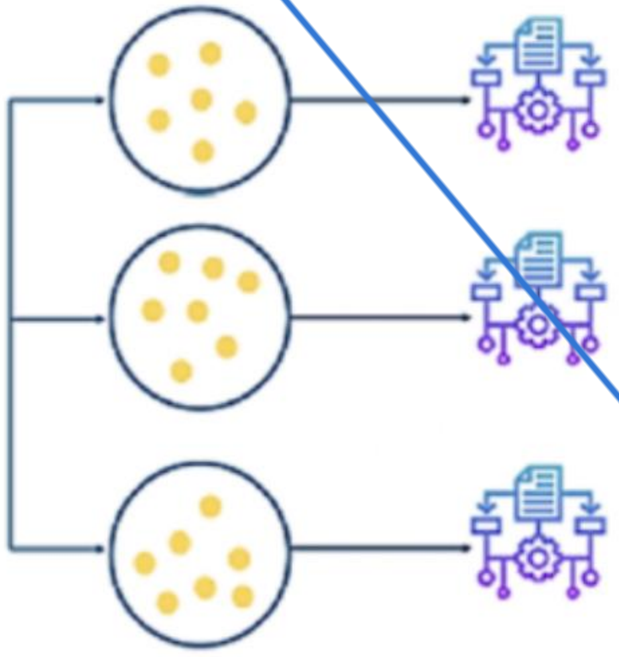
```
[('RF', 0.9583333333333333),
 ('LR', 0.8405172413793103),
 ('SVC', 0.8405172413793103),
 ('GNB', 0.8405172413793103),
 ('Ridge', 0.8405172413793103),
 ('XGB', 0.8405172413793103),
 ('BNB', 0.8065476190476191),
 ('KNC', 0.7649122807017543),
 ('SGDC', 0.7649122807017543),
 ('ADA', 0.7649122807017543),
 ('DTC', 0.7584415584415585),
 ('Perc', 0.7584415584415585),
 ('NC', 0.6939203354297694),
 ('PAC', 0.650462962962963)]
```

Random Forest 알고리즘이 95.8%의 예측확률로 폐암데이터 예측을 하는데는 가장 좋은 알고리즘인 것을 알 수 있습니다.

RandomForest 알고리즘은 무엇일까요?

Random Forest

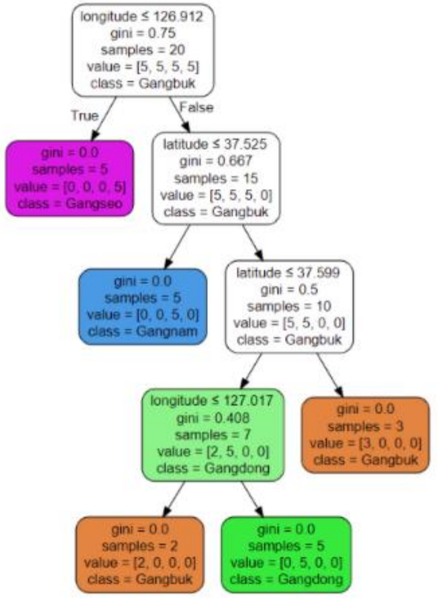
랜덤포레스트는 여러 Decision Tree를 배깅해서 예측을 실행하는 모델입니다.



DecisionTree

BAGGING

RANDOM FOREST



Decision Tree

결정 트리는 스무고개 하듯이 예/아니오 질문을 이어가며 학습합니다. 의사결정 트리는 훈련데이터에 대해 과대적합이 쉽게 될 수 있는 모델입니다.

BAGGING

한 가지 분류 모델을 여러개 만들어 서로 다른 학습 데이터로 학습시킨 후 투표를 통해 가장 높은 예측값으로 최종 결론을 내리는 앙상블 기법입니다.

Random Forest

랜덤 포레스트의 포레스트는 숲(Forest)입니다. 결정 트리는 트리는 나무(Tree)입니다. 나무가 모여 숲을 이룹니다. 즉, 결정 트리(Decision Tree)가 모여 랜덤 포레스트(Random Forest)를 구성합니다. 여러 개의 결정 트리를 통해 랜덤 포레스트를 만들면 과대적합 되는 단점을 해결할 수 있습니다.

더 좋은 모델 만들기 : Hyperparameter Tuning

하이퍼 파라미터?

최적의 인공지능 모델 구현을 위해 학습률이나 배치크기, 훈련 반복 횟수, 가중치 초기화 방법 등 인간의 선험적 지식을 기반으로 머신러닝이나, 딥러닝 모델에 설정하는 변수를 뜻합니다.
즉! 더 좋은 랜덤포레스트 모델을 만들기 위해 모델 개발자가 직접적으로 설정하는 파라미터를 뜻합니다.

```python
# Grid search and space:
models_params= {
    "RF":{'model':RandomForestClassifier(),
          'params':{
              'max_features': list(range(1,10)),
              'n_estimators':[10,100,1000]
          }}}

# Evaluation:

cv = RepeatedStratifiedKFold(n_splits=5,n_repeats=20)

# Search:
scores=[]

for model_name, params in models_params.items():
    rs = RandomizedSearchCV(params['model'], params['params'], cv=cv , n_iter=10)
    rs.fit(x_train,y_train)
    scores.append([model_name,dict(rs.best_params_),rs.best_score_])
data=pd.DataFrame(scores,columns=['Model','Parameters','Score'])
data
```

n_estimators	- 결정트리의 갯수를 지정 - Default = 10 - 무작정 트리 갯수를 늘리면 성능 좋아지는 것 대비 시간이 걸릴 수 있음
max_features	- 최적의 분할을 위해 고려할 최대 feature 개수 - Default = 'auto' (결정트리에서는 default가 none이었음) - int형으로 지정 →피처 갯수 / float형으로 지정 →비중 - sqrt 또는 auto : 전체 피처 중 √(피처개수) 만큼 선정 - log : 전체 피처 중 log2(전체 피처 개수) 만큼 선정

	Model	Parameters	Score
0	RF	{'n_estimators': 1000, 'max_features': 3}	0.918457

RandomizedSearchCV 알고리즘을 이용하여서 가장 좋은 랜덤 포레스트 모델은 최대 3개의 특성(독립변수)을 이용해서 1000개의 Decision Tree를 이용한 Random Forest 모델이 가장 좋은 모델임을 알 수 있었습니다.

03. 결론

1. Conclusion

2. 느낀점

Conclusion

우리는 이번 분석을 통해서 어떠한 가치를 알 수 있었을까요?

가장 영향을 많이 미치는 특성

폐암에 가장 영향을 미치는 특성은?

ALLERGY라는 변수가 상관관계 0.33로 가장 폐암에 영향을 많이 끼쳤습니다.
두번째로는 Alchohol 변수가 0.29로 두번째로 폐암 여부에 가장 많은 영향을 끼쳤습니다.

-> 폐암에 걸리지 않기 위해서는 알러지 여부를 미리 파악하고, 술을 적당히 드시는게 좋겠습니다.

생각보다?

SMOKING 변수

제 일반적인 상식으로는 Smoking 변수 즉 흡연 여부가 폐암에 아주 큰 영향을 미칠 꺼라 생각하였으나 0.05정도로 아주 작은 영향을 끼쳤습니다.

앞으로 무슨 모델로 예측해야해?

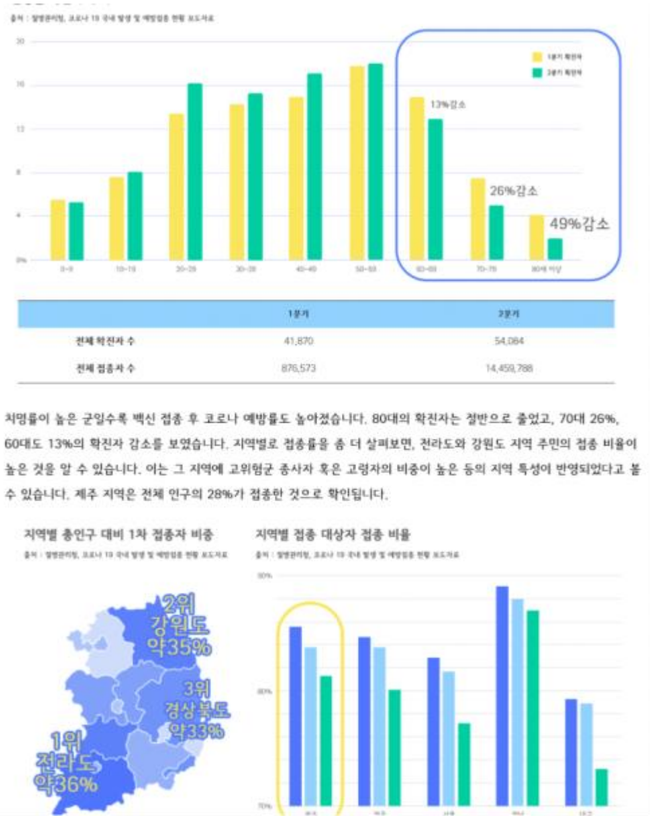
예측확률이 95%이상인 RandomForest모델을 활용하자

우리는 최대 3개의 특성을 가진 1000개의 Decision Tree를 이용한 Random Forest 모델이 가장 좋은 모델을 만들 수 있습니다.

결론

이 폐암 데이터 분석을 통해서 만약 환자의 위 16개의 특성을 알 수 있다면 우리가 만든 모델을 이용하여 95%의 정확도로 해당 환자의 폐암 여부를 알 수 있을 것입니다.

느낀점



https://www.jejudatahub.net/report/serial/view/40

빅데이터 융합전공을 하면서 항상 데이터 분석에 관심이 많았고, 빅데이터와 함께 따라오는 키워드는 인공지능 이었습니다.
이번 인공지능 수업을 들으면서 마냥 어렵게만 느껴졌던 AI란 단어가 조금은 생소하지 않게 느껴졌습니다.
Know-How 에서 Know - Where 로 생각하는 방법이 인공지능을 어렵게만 생각하던 저에게 커다란 도움이 되었던 것 같습니다.
모델이 어떠한 코드나 계산 방식으로 이루어진 것 보다, 해당 모델의 기능에 대해서 잘 알고 , 내가 분석하고자 하는 데이터 셋이나 분석방법론에 맞춰 좋은 모델을 찾는 것이 좋은 경험이 되었던 것 같습니다.
이번 학기에는 좋은 기회로 10월 17일 글로 제가 기획하고 분석한 제주 데이터허브에 2분기 리포트도 쓰게 되어
많은 성장을 일궈낸 한 학기였던것 같습니다.
더 공부하고 성장하여 더 나은 분석을 고민하는 좋은 개발자가 되고 싶습니다.
감사합니다.