Ensemble project

심장 질환 예측 모델

pronounce12@gmail.com
조현상
github notion

01 심장 건강 관련 데이터셋 소개 및 분석 목적

1. 목적 :

심장 질환 예측 모델은 사람들의 건강 관련 데이터와 생활 패턴 데이터를 활용하여 개인의 심장 질환 여부를 예측, 분류하는 모델입니다.

2. 프로젝트 전제 조건:

- 데이터셋은 2020년에 수집된 약 32만명의 의료 및 생활양식 데이터로 구성됩니다.
- 데이터는 환자의 의료 정보와 라이프스타일에 관한 정보를 각각의 데이터셋에 포함하고 있습니다.
- 프로젝트의 주요 관심사 및 중점 분석 대상은 "HeartDisease" 컬럼으로, 이 컬럼은 심장 질환 발생 여부를 나타내는 중요한 변수입니다. 이를 통해 심장 질환과 관련된 패턴 및 예측 모델을 개발하고자 합니다.

3. 데이터 병합:

medical_dataset과 lifestyle_dataset의 환자번호(Patient_Id)로 병합하였습니다.

4. 컬럼 소개 :

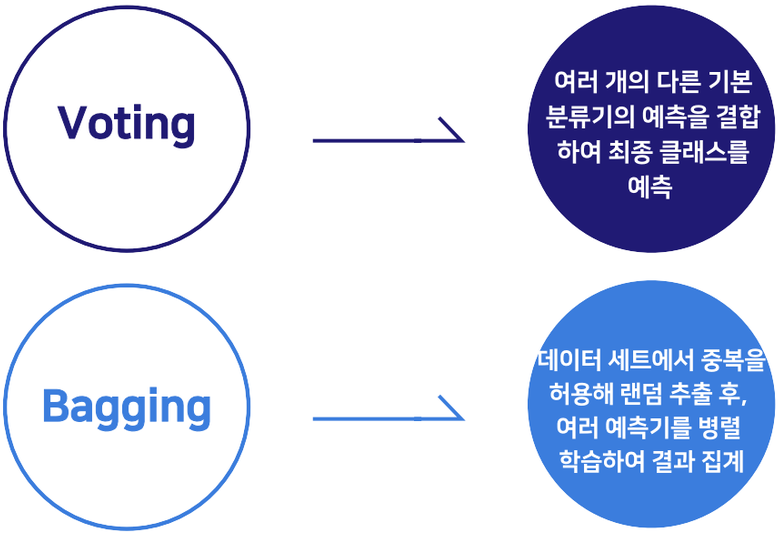
medical_dataset

- BMI (체질량 지수)
- Stroke (뇌졸중 여부)
- PhysicalHealth (신체 건강 상태)
- AgeCategory (연령 범주)
- MentalHealth (정신 건강 상태)
- DiffWalking (걷기 어려움 여부)
- Diabetic (당뇨병 여부)
- GenHealth (일반 건강 상태)
- Asthma (천식 여부)
- KidneyDisease (신장 질환 여부)
- SkinCancer (피부암 여부)
- Sex (성별)
- HeartDisease (심장 질환 여부)
- PatientID (환자 ID)

lifestyle_dataset

- Smoking (흡연 여부)
- AlcoholDrinking (음주 여부)
- PhysicalActivity (신체 활동 빈도)
- SleepTime (수면 시간)
- Race (인종)
- PatientID (환자 ID)

02 앙상블 기법의 설명 및 비교



Voting

"하나의 데이터 세트"에 대해 서로 다른 알고리즘을 가진 분류기를 결합하는 방식입니다.
예측값을 다수결로 투표해서 가장 많은 표를 얻은 예측값을 최종 예측값으로 결정하는 **하드보팅**과,
각 분류기가 예측한 타겟별 확률을 평균내어 가장 높은 확률의 타겟을 최종 예측값으로 결정하는 **소프트 보팅**이 있습니다.

Bagging

다양한 데이터 샘플과 특성을 사용하여 과적합의 위험을 줄일 수 있습니다.
특성 중요도를 제공하여 모델의 해석이 용이합니다.
병렬 처리가 가능하므로 대규모 데이터셋에서도 **빠르게** 학습과 예측이 가능합니다.
단일 의사결정트리보다 **높은 정확도와 안정성**을 제공합니다.

03 데이터 전처리

결측치와 중복행 확인

결측치 확인: isna().sum()
중복행 확인: duplicated().sum()

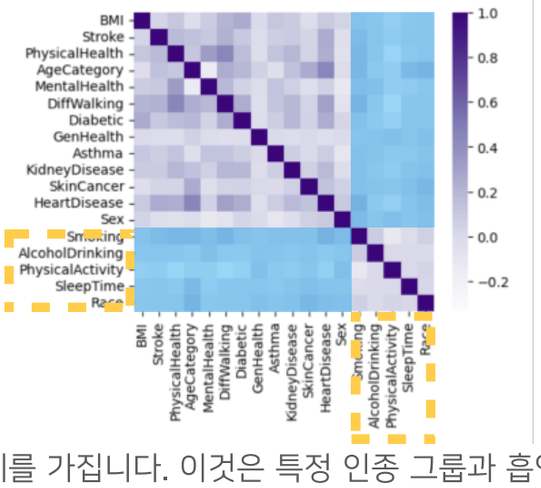
결과: 결측치, 중복행 모두 0

```
In [4]: import pandas as pd

        # 결측치 확인
        print(heart_df.isna().sum())

        # 중복행 확인
        print(heart_df.duplicated().sum())
```

```
BMI               0
Stroke            0
PhysicalHealth    0
AgeCategory       0
MentalHealth      0
DiffWalking       0
Diabetic          0
GenHealth         0
Asthma            0
KidneyDisease     0
SkinCancer        0
HeartDisease      0
Sex               0
PatientID         0
Smoking           0
AlcoholDrinking   0
PhysicalActivity  0
SleepTime         0
Race              0
PatientID         0
dtype: int64
0
```

04 Heatmap

데이터 세트 병합 후, HeatMap을 통해 확인된 연관성



HeatMap 연관성 지표

Smoking 0.175500
Race 0.066313
SleepTime 0.007141
AlcoholDrinking -0.083140
PhysicalActivity -0.174041

해석

HeatMap에 따르면 Smoking과 Race는 심장질환과 양의 상관 관계를 가집니다. 이것은 특정 인종 그룹과 흡연이 심장질환 발생과 연관될 수 있음을 뜻합니다.

한편, AlcoholDrinking과 PhysicalActivity는 음의 상관 관계를 가지며, 즉, 두 변수 중 하나가 증가할 때 다른 변수가 감소하는 경향을 나타냅니다.

이를 통해 라이프스타일에 따라 심장건강에 영향을 미칠 수 있다고 볼 수 있습니다.

05 언더샘플링과 Histogram 그래프

```
HeartDisease
No      292422
Yes      27373
Name: count, dtype: int64
```

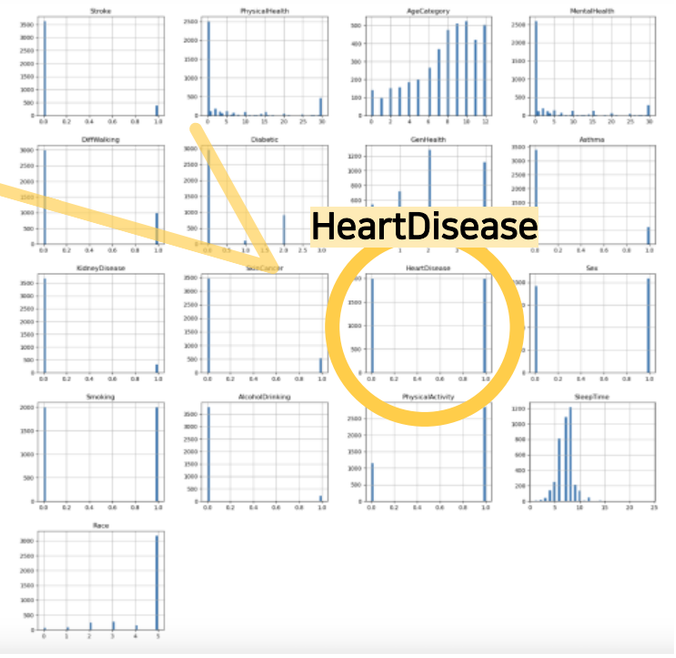
언더 샘플링을 하기 전에는 HeartDisease 변수의 데이터가 크게 불균형했습니다.

우측의 HeartDisease 변수에 대한 히스토그램 그래프가 2000씩만 나오는 이유는, 언더샘플링을 활용하여

데이터셋에서 HeartDisease가 발생하지 않은 경우(0)와 발생한 경우(1)의 샘플 수를 균형있게 맞추었기 때문입니다.

이로 인해 두 클래스 간의 샘플 수 차이가 줄어들어 2000개의 샘플만 표시됩니다.

언더샘플링을 통해 데이터 불균형 문제를 완화하고 데이터 분석을 더 정확하게 수행할 수 있었습니다.



06 Voting Ensemble 평가

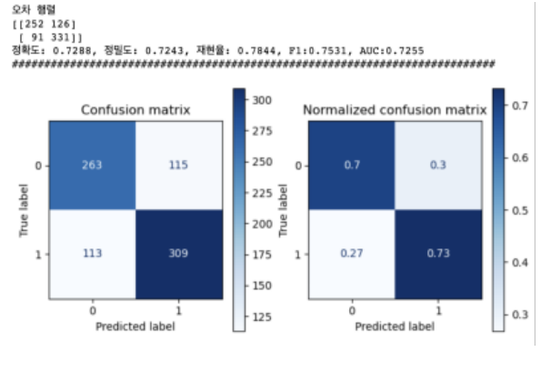
Voting Ensemble

✔ 의학 분야에서는 건강 상태를 판단하거나 질병을 감지하는 모델을 다루므로, 건강 상태를 놓치지 않고 실제 양성 사례를 잘 감지하는 것이 중요하다 판단하여 임계값을 조절해 재현율을 높히는 것에 초점을 뒀습니다.

DTC
```
prediction = Binarizer(threshold=0.47619047619047616).fit_transform(prediction_prob_class1)
get_evaluation(y_test, prediction, grid_dt_classifier, X_test)
```

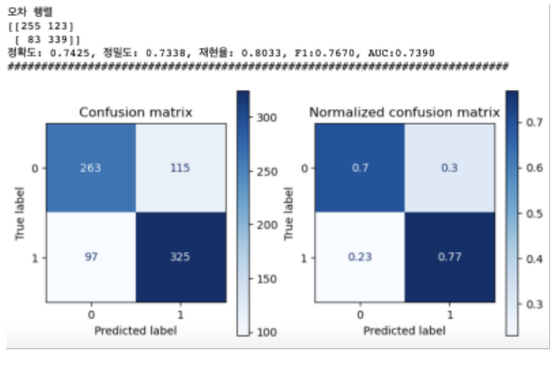
SVC
```
prediction = Binarizer(threshold=0.44365198336980655).fit_transform(prediction_prob_class1)
get_evaluation(y_test, prediction, grid_svc_classifier, X_test)
```

KNN
```
prediction = Binarizer(threshold=0.42857142857142855).fit_transform(prediction_prob_class1)
get_evaluation(y_test, prediction, grid_knn_classifier, X_test)
```



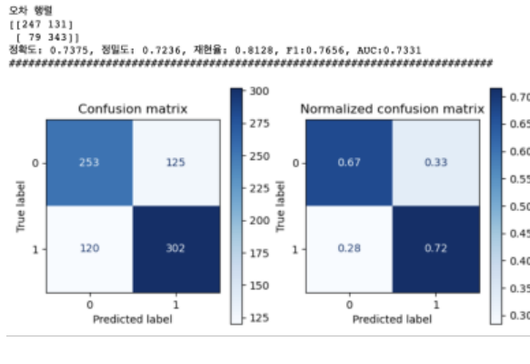
1. DecisionTree

정확도 (Accuracy): 0.7288
정밀도 (Precision): 0.7243
재현율 (Recall): 0.7844
F1 스코어 (F1 Score): 0.7531
AUC (Area Under the ROC Curve): 0.7255



2. Support Vector Machine

정확도 (Accuracy): 0.7425
정밀도 (Precision): 0.7338
재현율 (Recall): 0.8033
F1 스코어 (F1 Score): 0.7670
AUC (Area Under the ROC Curve): 0.7390



3. KNN

정확도 (Accuracy): 0.7375
정밀도 (Precision): 0.7236
재현율 (Recall): 0.8128
F1 스코어 (F1 Score): 0.7656
AUC (Area Under the ROC Curve): 0.7331

07 Bagging Ensaemble 평가

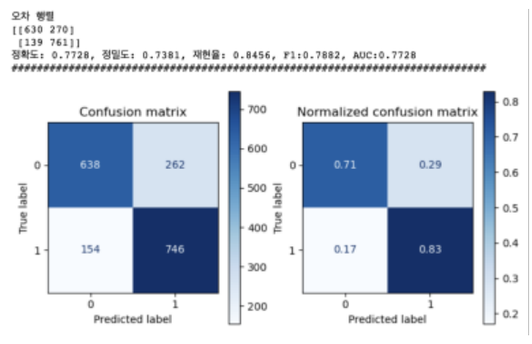
Bagging Ensaemble

✔ 의학 분야에서는 건강 상태를 판단하거나 질병을 감지하는 모델을 다루므로, 건강 상태를 놓치지 않고 실제 양성 사례를 잘 감지하는 것이 중요하다 판단하여 임계값을 조절해 재현율을 높히는 것에 초점을 뒀습니다.

↓

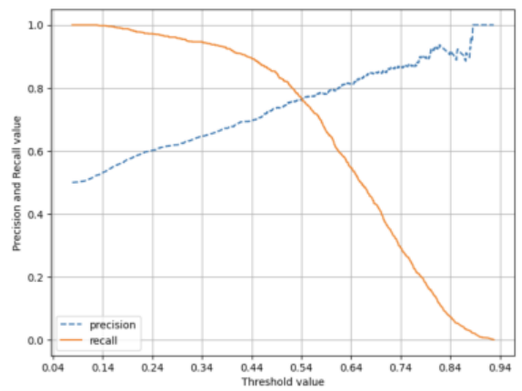
Random Forest

```
prediction = Binarizer(threshold=0.49053045757631125).fit_transform(prediction_prob_class1)
get_evaluation(y_test, prediction, grid_random_forest, X_test)
```



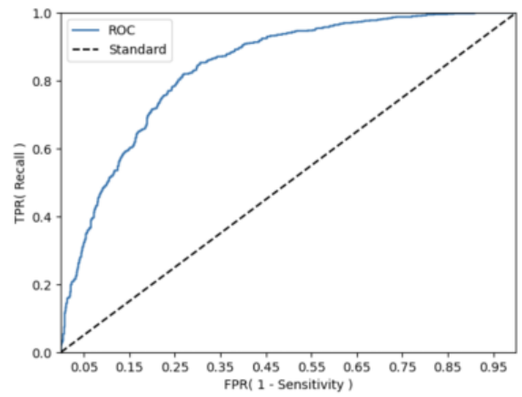
1. Random Forest

정확도 (Accuracy): 0.7728
정밀도 (Precision): 0.7381
재현율 (Recall): 0.8456
F1 스코어 (F1 Score): 0.7882
AUC (Area Under the ROC Curve): 0.7728



2. Precision-Recall Curve

그래프를 확인해보면, 정밀도(Precision)와 재현도 (Recall)이 0.54의 임계값에서 교차하는것을 확인할 수 있습니다.모델의 전제조건을 수행하기 위해 재현도를 높히기 위해 0.49의 임계치를 선택하였습니다.



3. Roc Curve

ROC 커브는 이 그래프는 임계값 변화에 따른 진짜 양성 비율(TPR)과 거짓 양성 비율(FPR) 사이의 관계를 나타냅니다. ROC 커브는 곡선 아래 영역(AUC)이 1에 가까울수록 모델의 성능이 좋습니다.

08 모델 전체 평가

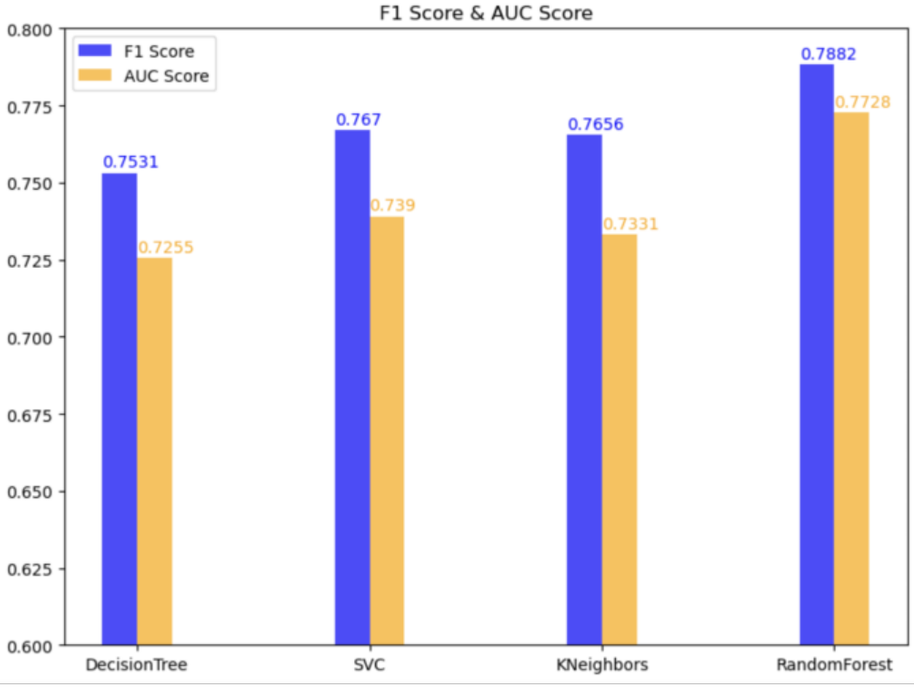
모델 선택과 평가

RandomForest의 F1 스코어와 AUC 스코어가 다른 모델에 비해 상대적으로 높습니다.
따라서 RandomForest가 다른 모델보다 예측 성능이 우수하다고 판단할 수 있습니다.

Bagging은 Bootstrap 샘플링을 사용하여 각 분류기에게 독립적인 데이터 세트를 제공합니다.
또, 중복을 허용하는 샘플링 방식으로, 각 분류기가 다양한 데이터에 대해 학습하게 됩니다.

이에 반해 Voting은 동일한 데이터 세트를 모든 분류기에게 제공합니다. 모든 분류기가 동일한 데이터에 대해 학습하게 되므로 다양성이 부족할 수 있습니다.

따라서 Bagging은 다양한 데이터 세트와 다양한 분류기를 사용하여 모델의 안정성을 향상시키고, 과적합을 줄이는 데 도움을 줘서 더 좋은 스코어를 받을 수 있었다고 생각됩니다.



09 느낀점

언더샘플링

언더샘플링 기법을 적용했을 때, 일부 데이터를 제거함으로써 원래 데이터셋에서 포함된 중요한 정보나 패턴이 손실될 수 있다는 점이 크게 아쉬웠습니다. 특히, 32만개의 데이터를 수집하여 최대한 많은 정보를 활용하기 위해, 또 좋은 성능을 얻기위해 학습을 진행하는데, 언더샘플링을 통해 데이터의 일부를 제거함으로써 그 가치를 줄이는 것에 대해 아쉬웠습니다.

학습 속도

DTC, KNN, SVC, 그리고 랜덤포레스트를 사용하여 작업을 진행하였는데 예상보다 학습 속도가 상당히 느려서 프로젝트 진행에 많은 시간이 소요되었습니다. 특히 SVC와 랜덤포레스트의 경우 데이터의 크기나 복잡도에 따라 학습 시간이 급격히 증가하고, 컴퓨터에 굉장히 부담이 되는 경향이 있어, 이를 해결하려는 여러 시도가 필요했습니다.

감사합니다