심장병이 있는 사람 예측 및 분류

어드벤쳐디자인:빅데이터마이닝:패턴탐색

곽명빈(데이터테크), 정혜정(임상의학통계), 황유은(데이터테크)

지도교수 : 박현숙

요약

미국에서는 매년 647,000명이 심장병으로 사망하고 있고, 한국에서도 암 다음으로 심장질환이 높은 사망률을 보이는 상황임. BRFSS의 약 20만 명 가량의 심장병 데이터를 바탕으로 심장병에 영향을 주는 요인을 분석할 계획임. EDA를 통해 데이터의 분포를 살펴보고, 의사결정 나무를 이용하여 심장병이 있는 사람들을 예측 및 분류를 통해 심장병으로 인한 사망자를 줄이는 것을 목표로 연구를 진행하였음.

연구배경

미국에서는 매년 647,000명이 심장병으로 사망하고 있고, 대부분의 사람들은 가슴통증, 심정지 같은 증상 이후에 병이 있다는 사실이 있음. 또한 한국에서도 암 다음으로 심장질환이 높은 사망원인임을 확인 하였음. 분류와 예측을 통해 심장질환이 있는 사람들의 특성을 파악 후, 심장병으로 인한 사망자를 줄이는 것을 목적으로 프로젝트를 진행하게 되었음.

연구방법

데이터 설명

Heart disease health indicators BRFSS2015

- **종속변수**

HeartDiseaseorAttack

- **독립변수**

HighBP, HIghChol, Diabetes, BMI, Stroke, …

분석방법

1. EDA

- 막대그래프를 이용하여 데이터별 분포 확인
- 상관계수를 이용하여 종속변수와 독립변수 간 선형성 확인

2. Decision Tree

- 의사결정 나무를 이용해 예측 및 분류
- K-fold, Grid Search를 이용해 모델 최적화
- Confusion Matrix를 이용해 모델 성능 검증
- Over Sampling(SMOTE)을 이용해 데이터 수 조정

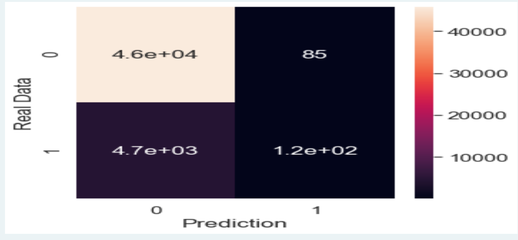
연구결과

Decision Tree

- Sklearn의 train test split을 이용하여 8:2 비율로 분리 후, 모델 설정
- K-fold와 Grid search를 이용하여 최적의 파라미터를 찾아 모델 생성 후 검증하였지만 재현율이 낮은 문제 발생
- Oversampling(SMOTE) 후, 같은 작업을 통하여 재현율 확인

```
Accuracy_score:0.9066 → 0.7219
Recall_score:0.02473 → 0.8167
Precision_score: 0.5813 → 0.6867
```

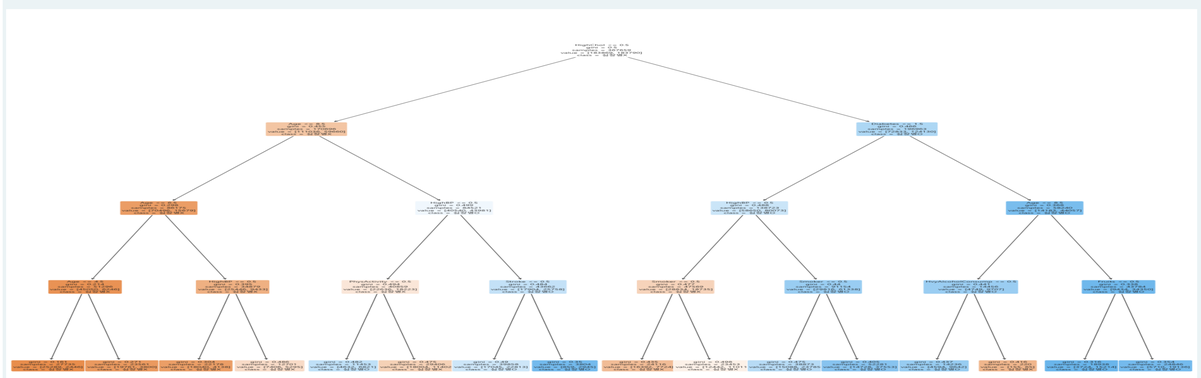
정확도는 떨어졌지만, 재현율과 정밀도가 상승




```
criterion = "gini", max_depth = 4, max_features= 'log2'
```

GridSearch로 찾은 최적의 모델을 바탕으로 의사결정나무 시각화

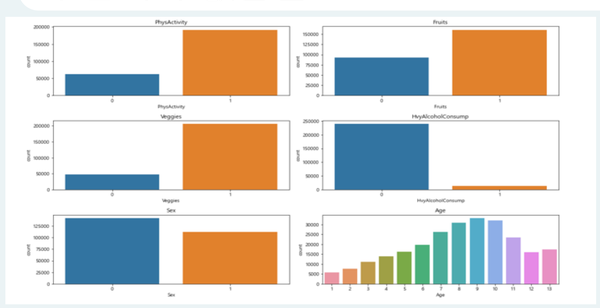
- 콜레스테롤 수치 ↓ 45세 ↑, 고혈압, 뇌졸중 유무에 관계없이 심장병
- 콜레스테롤 수치 ↓ 45세 ↑, 운동 X 면 심장병
- 콜레스테롤 수치↑ 고혈압, 흡연 여부에 관계없이 심장병
- 콜레스테롤 수치↑ 당뇨병 O, 나이와 관계없이 심장병

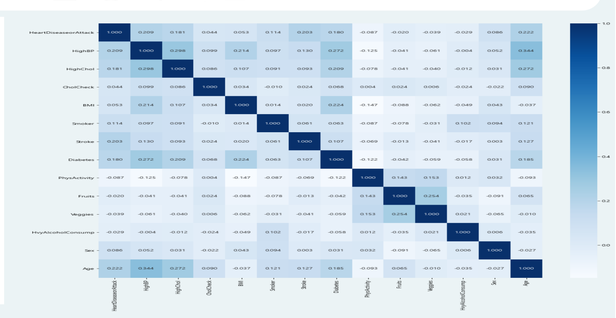


연구결과

EDA

막대그래프로 데이터 불균형 문제 확인, 종속변수와 독립변수 간 선형적 관계가 없음을 correlation matrix를 통해 확인





결론

1. 콜레스테롤 수치가 높을 때는 당뇨병과 고혈압을 예방해야 함.
2. 당뇨병이 있을 때 심장병이 발생할 위험이 커짐.
3. 콜레스테롤 수치가 낮아도 나이가 많으면 고혈압을 예방해야 함.
4. 고혈압이 없어도 운동을 하지 않으면 심장병의 위험이 커짐.
5, 노화와 고혈압, 당뇨병은 모두 심장 질환에 대한 원인 및 위험 요소.
6. Centers for Disease Control and Prevention의 주요 위험요소와 비슷함.