국가건강영양조사를 활용한
한국형 고혈압 예측 모델링

강의명: 비즈니스 통계 응용

학과: AI 빅데이터융합경영학과

학번: 20195262

이름: 장예진

목차

제 1 장 서론

1. 연구배경

고혈압은 현대 사회에서 급증하는 만성 질환 중 하나이자 세계적으로 사망 위험 요인 1 위로 꼽힌다. 과다한 칼로리 섭취, 스트레스 등의 건강행태에 따라 각 개인의 건강 상태는 악화되고 있는 상황이다. 최근 2030 세대의 고혈압 유병율이 증가함에 따라, 고혈압의 조기 예측은 개별 건강 상태를 지속적으로 관리하고 예방적인 의료지표를 제공하는 데에 기여할 수 있다.

2. 고혈압의 정의와 종류

먼저, 고혈압은 심장이 온몸으로 피를 뿜어낼 때 혈관이 받는 압력인 혈압이 정상 수치보다 올라가 있는 상태를 의미한다. 혈압 상승을 일으키는 특정한 원인이 확인되었는지에 따라 일차성(본태성)과 이차성 고혈압으로 구분할 수 있다.

1) 일차성(본태성)고혈압은 혈압 상승을 가져올만한 특정한 원인이 발견되지 않은 고혈압으로서 환자의 90~95%가 해당된다. 직접적인 원인은 밝혀지지 않았지만 운동부족, 복부비만, 염분 과다 섭취 등의 인자들이 고혈압 발병과 관련이 있는 것으로 알려진다.

2) 이차성 고혈압이란 혈압 상승을 가져올 만한 특정한 원인 질환이 확인된 고혈압으로, 기저질환을 치료하면 대부분 정상혈압으로 돌아오게 된다. 이차성 고혈압을 유발하는 대표적인 질환에는 심혈관질환, 쿠싱 증후군, 갑상선 기능 항진증, 임신 중독증 등이 있다.

고혈압은 초기에 이렇다 할 뚜렷한 증상이 없기 때문에 간과할 수 있는 위험한 질병이다. 이에 따라 "고혈압 위험요인을 가지고 있는 정상인이 향후 고혈압이 발병할 확률은 어느정도 인지"에 대한 지표를 제공하고자 본 연구 에서는 데이터를 활용하여 한국형 고혈압 예측 모델을 개발하고자 한다. 추가적으로 결과를 적용한 의료비 목록을 시각화 하여 다양한 관점의 결과를 산출해보고자 한다.

3. 이전 고혈압 진단 기준과 연구

이 전까지 고혈압 진단 기준은 수축기 140mmH/이완기 90mmHg 으로 진단되어왔다. 2017 년 이후부터 미국 심장 협회에서 수축기 130mmHg/이완기 80mmHg 으로 기준을 강화하였다. 유럽과 대한민국의 경우에는 기존의 진단 기준인 140/90 의 기준을 유지하고 있다. 이렇게 서로 다른 국가 및 지역에서의 고혈압 진단 기준의 차이는 다양한 요인의 영향을 받지만 이로 인해 특정 환자가 고혈압 진단을 받을 수 있는 기준이 상이하게 적용될 수 있다. 따라서 한국형 고혈압 예측 모델을 개발할 때 이러한 다양성을 고려하여 모델을 구성하고 결과를 해석하는 것이 중요하다.

4. 연구 목적

본 연구의 주요 질문은 "현대 사회에서 고혈압의 예측 요인은 무엇인가?"이다. 이를 통해 고혈압 발생에 영향을 미치는 주요 요인들을 식별하고, 개인의 건강상태를 조기에 파악하여 적절한 예방조치를 취할 수 있는 모델을 개발하는 것이 목표이다. 그에 따라 미국(US)와 유럽(EU)의 진단 기준에 따라 고혈압 예측 모델링을 진행한다. 모델링을 진행한 이후에, 미국과 유럽의 진단 기준에 따른 예측 모델 성능 비교 및 일반화 가능성을 비교하여 결과를 도출한다. 연구의 최종 목적은 예측 모델(interpretation) 해석을 통한 위험요인의 중요도 및 경향성 파악이다.

제 2 장 문헌 검토와 이론적 틀

1. 선행 연구 및 구별점

기존의 머신러닝 기반 고혈압 예측 모델에서는 사회인구학적 특징 정도로 구성된 변수를 구성된 일반적이고 심플한 변수들을 사용하여 일반인이 이해하기 쉽고 활용하기 쉽게 진행한 모델이 대부분이었다. 본 연구에서는 조금은 복잡한 모델과 구조를 사용하지만 다양한 시각화를 통해 쉽게 이해할 수 있도록 접근해보았다. 또한, 다양한 기계학습 모델 종류를 활용하여 모델링을 최적의 모델을 찾는 하이퍼파라미터 튜닝까지 진행하여 강화된 모델을 제시한다.

2. 데이터 소개 및 work flow

본 연구는 질병관리청에서 제공하는 국가건강영양조사(2019~2021) 3 개년의 데이터를 기반으로서 데이터 전 처리 및 특성 선택을 거쳐 다양한 머신 러닝 알고리즘을 활용하며 4 개의 모델을 학습하고 모델을 비교한다. 그 중 평가지표 점수 상위 모델에게 하이퍼파라미터 튜닝을 진행하여 최적의 모델을 학습하고 성능을 평가한다. 또한 모델의 결과를 해석하고 의료비와의 관련성을 분석하여 다각적인 결과 도출을 진행하였다.

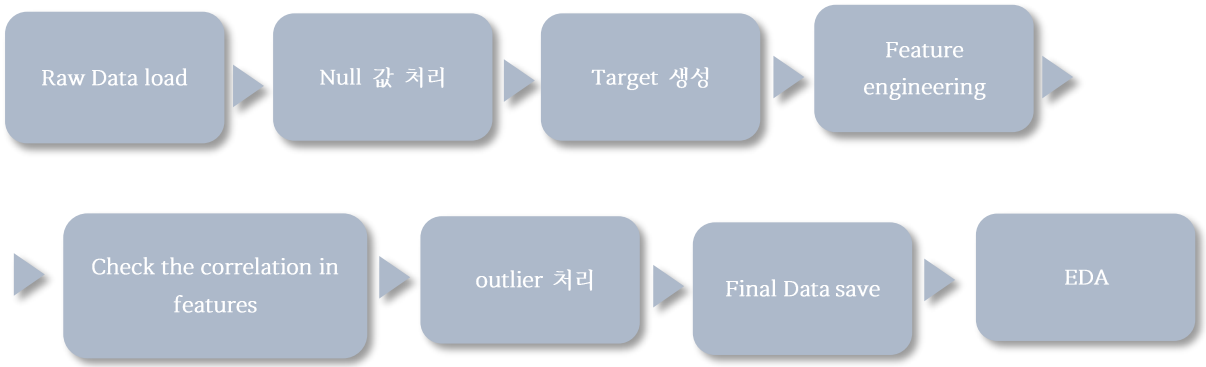
연구에서 식별된 고혈압 발생에 영향을 미치는 주요 요인들은 연구의 특정 그룹 사이에 격차를 초래할 수 있다. 이러한 격차들을 해소하기 위해 본 연구는 다양한 특성 그룹핑 및 건강 행태에 관련된 특성들을 추가적으로 고려하고 분석할 예정이다. 다양한 요인을 고려하여 모델의 일반화 성능을 향상시키고 다양한 인구 집단간의 예측 정확도의 균형은 맞추는 방안을 논의하고자 한다.

제 3 장 본론

1. 데이터 수집

대한민국 질병관리청 국민건강영양조사 3 개년(2019~2021) 병합 데이터셋을 준비한다. 데이터셋을 들여다보니 검진조사는 고혈압, 당뇨, 이상지질혈증, 비만 등, 건강설문조사는 음주, 흡연, 의료이용 등 그리고, 식품 및 영양소 섭취현황 식생활 형태를 나타낸 영양 조사로 구성 되어있다. 다운받은 원시자료는 python 을 활용하여 .SAS(확장자) > .CSV 로 변환(encoding='cp949')

> df 로 정의까지 데이터 준비과정을 진행하였다.

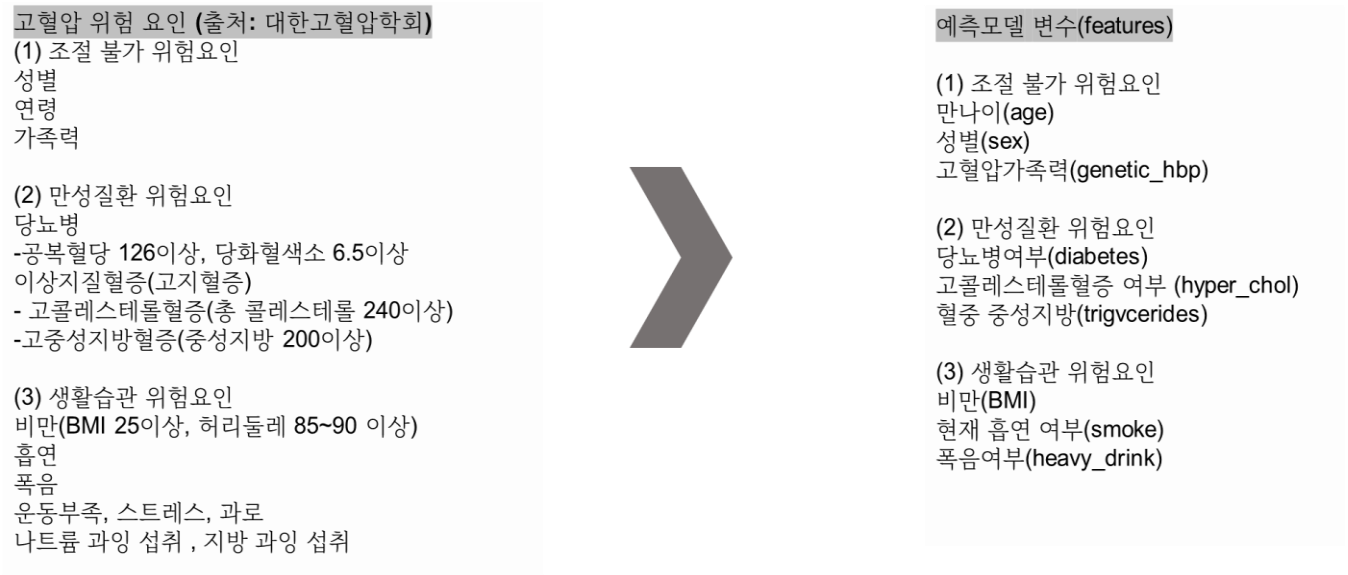


[그림 1] 데이터 전 처리 work flow

2. 데이터 특성 확인 및 변수 처리

고혈압 진단 예측 모델이기에 고혈압 여부를 예:1, 아니오:0 로서 모델의 종속 변수를 정하여 두 가지로 나누는 이진분류문제로 정의하였다. 위의 언급한 것과 같이 미국 진단 기준 130/80 이상이거나 혈압조절제를 복용하는 자, 유럽진단 기준 140/90 이상이거나 혈압 조절제를 복용하는 자를 기준으로 정하였다. 유의한 변수 선택을 위해서 위험요인과 특성은 대한 고혈압 학회에서 정의된 위험요인과 특성 중에서 객관화가 가능하 수치를 선정하여 예측 모델의 특성으로 사용하였다.

[그림 2]아래의 선정된 변수 들의 상세한 기준을 선정하여 데이터 preparation 을 진행하였다.



먼저, 필요한 변수만을 list 에 넣어 정의하고 추출해준다. 예측모델 각 변수 명을 확인해보았다.

(1) 조절불가변수: ID(식별번호), year(연도), sex(성별), age(나이)

HE_HPfh1(가족력: 고혈압의사진단여부(부), HE_HPfh2(가족력: 고혈압의사진단여부(모))

(2) 만성질환변수: DI1_2(혈압조절제복용), HE_dbp(1 차 이완기혈압)

DE1_31(당뇨병혈당치료:인슐린주사), DE1_32(당뇨병혈당치료:당뇨병약), HE_glu(공복혈당),

HE_HbA1c(당화혈색소), DE1_dg(당뇨병 의사 진단 여부),HE_sbp(당뇨병 유병여부)

DI2_2(이상지질혈증 약 복용), HE_chol(총콜레스테롤) , HE_TG(중성지방)

(3) 생활습관변수: BD1_11(음주 율), BD2_1(고 위험 음주 율), sm_presnt(현재흡연율),

HE_wt(체중), HE_wc(허리둘레), HE_BMI(체질량지수)

변수들을 추출하고 결측치를 np.nan 값으로 변환해 주고 결측치가 포함된 row 도 제거해 준 이 후, Info()함수를 통해 확인한 대부분이 범주형 변수로서 ID 를 제외한 모든 column 은 수치형으로 변환해주었다.

모델링을 하기 전, 목적에 맞게 종속변수를 생성해주고자 한다.

	DI1_2	HE_sbp	HE_dbp	bp_drug	HBP_US	HBP_EU
0	1.0	122.0	84.0	1	1	1
1	8.0	111.0	73.0	0	0	0
2	8.0	125.0	85.0	0	1	0
3	8.0	109.0	77.0	0	0	0
4	1.0	131.0	77.0	1	1	1

[그림 3]

고혈압 조절제 복용여부 응답 1,2,3,4 에 따른 column 생성(bp_drug)

130/90(미국진단기준), 혈압조절제 복용에 따른 고혈압 여부 column 생성 (HBP_US)

140/90(유럽진단기준), 혈압조절제 복용에 따른 고혈압 여부 column 생성 (HBP_EU)

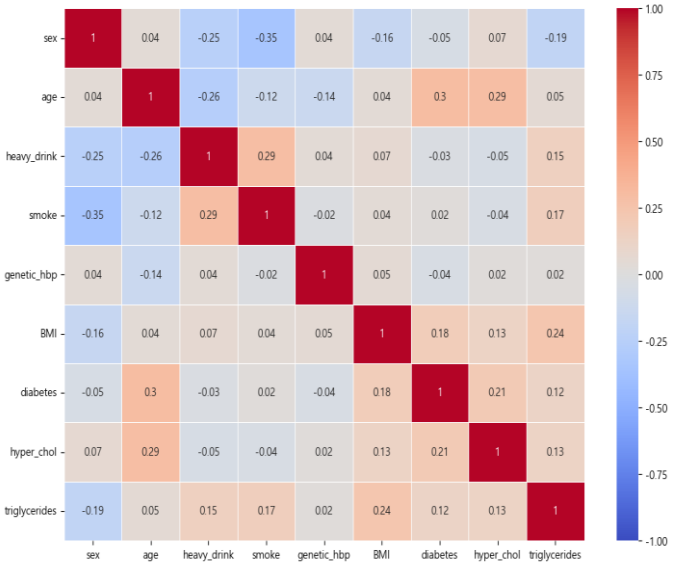
정의된 조절불가, 만성질환, 생활습관 변수에 대한 Feature Engineering(특성 가공) 진행 과정이다.

- ID, year, sex, age column 추가
- BD1_11 ==3,4,5,6, BD2_1==4,5 을 활용하여 폭음 여부 "heavy_drink column" 추가
- sm_presnt 에 따른 현재 흡연 여부"smoke column"추가
- DE1_31, DE1_32 에 해당되거나, DE1_dg>126(공복혈당), HE_glu(공복혈당)<6.5 이상인 당뇨병 여부에 관한 "diabetes column" 추가
- HE_TG 에 따른 중성지방 "triglycerides column"추가
- D1_2 >4 이거나 HE_chol >240 에 해당하는 이상지질혈증(고콜레스테롤혈증)여부에 따른 "hyper chol column"추가
- HE_HPfh2, HE_HPfh2 에 따른 가족력 "genetic_hbp column"추가
- HE_wt, HE_wx HE_BMI 에 따른 체중, 허리둘레, "BMI column" 추가

[그림 4]

	ID	year	sex	age	heavy_drink	smoke	genetic_hbp	weight	waist	BMI	diabetes	hyper_chol	triglycerides
5379	N760288401	2019.0	2.0	53.0	0	0	1.5	68.7	89.5	25.636538	0	0	55.0
3912	H790227201	2019.0	1.0	42.0	1	0	1.5	63.3	88.1	23.708256	0	1	78.0
7093	A835210202	2020.0	2.0	52.0	0	0	2.0	57.6	83.2	22.956841	0	0	83.0
6998	A831291801	2020.0	1.0	66.0	1	0	1.0	67.4	84.3	21.957945	0	0	132.0
5769	O763317302	2019.0	2.0	67.0	0	0	1.0	50.1	83.6	20.586785	1	0	62.0

(1) 상관관계 확인



[그림 5] 데이터 전처리와 준비를 끝 마친 이후에, 위의 변수들이 상관관계를 살펴보기 위한 heatmap 을 출력해보았다.

분석 결과, BMI, 체중, 허리둘레 변수 간의 강한상관관계를 확인하였고, 강한상관관계로 인한 다중공선성 문제가 예상되어 그에 따른 특성 제거를 통해 문제를 해결해 보고자 하였다.

강한 상관관계를 가지는 체중과 허리둘레 변수를 제거한 이후의 상관관계를 확인해보니 다중공선성 문제가 해결된 상관관계 heatmap 을 확인할 수 있다

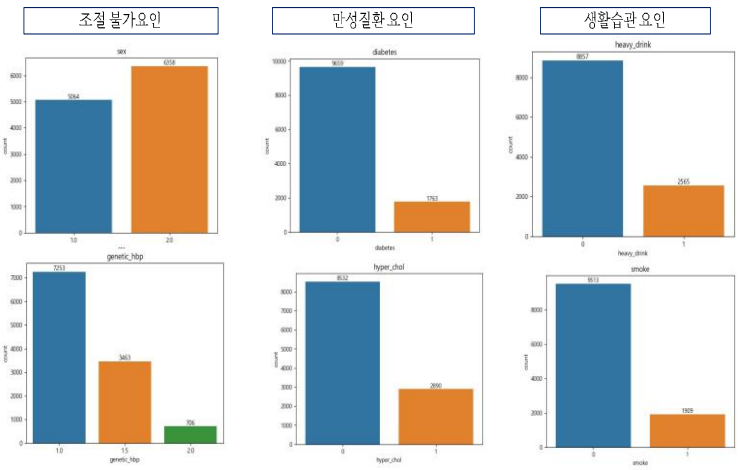
(2) 이상치 제거

변수 간의 전체 분포를 (pairplot) 분포를 시각화 하여 보니, 그중 중성지방과 BMI 분포에서의 극단적으로 큰 값들이 존재하여 이상치들을 제거하였다. 이상치의 기준으로서 중성지방은 은 500mg/dL 이상이거나 BMI 는 40 이상으로 설정하였다.

3. EDA

전처리와 변수 추가를 완료한 최종 데이터 EDA 진행하였다. 독립변수를 기준으로 한 상관관계 시각화를 진행해보니 데이터에서의 불균형은 없다고 판단하였다. 그리하여 Raw data 로 정의한 진단 기준에 따른 분포와 비교하여 확인해보았다.

02 EDA – 범주형 변수 확인



[그림 6]

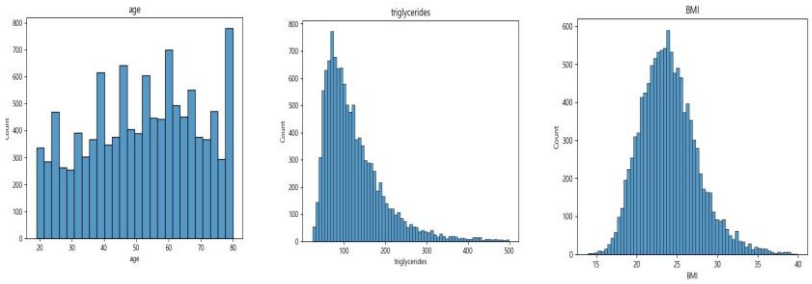
조절불가요인인 성별은 여성보다 남성이 많았고, 가족력 변수는 고혈압 유병 가족력이 없는 비율이 월등하다.

만성질환요인인 당뇨병과 고콜레스테롤혈증 둘 모두 유병하지 않은 사람의 비율이 많다.

생활습관 요인인 과음과 흡연 부분에서도 비음주인과 비흡연자의 비율이 더 많다.

02 EDA – 수치형(연속형) 변수 확인

조절불가요인 | 만성질환요인 | 생활습관요인



[그림 7]

수치형 변수 EDA

나이분포: 대부분의 참가자들이 40~60 대 사이에 분포하며 특히 50 대의 분포가 두드러진다.

중성지방 분포: 대부분의 참가자들은 낮은 중성지방 분포를 보이고 있으며 수치가 150 이하인 참가자가 가장 많다.

체질량지수분포: BMI20~30 대 사이에 참가자가 가장 많다.

4. 모델링

Modeling basic preparation

(1) Train-test split 을 8:2 로 진행하였다.

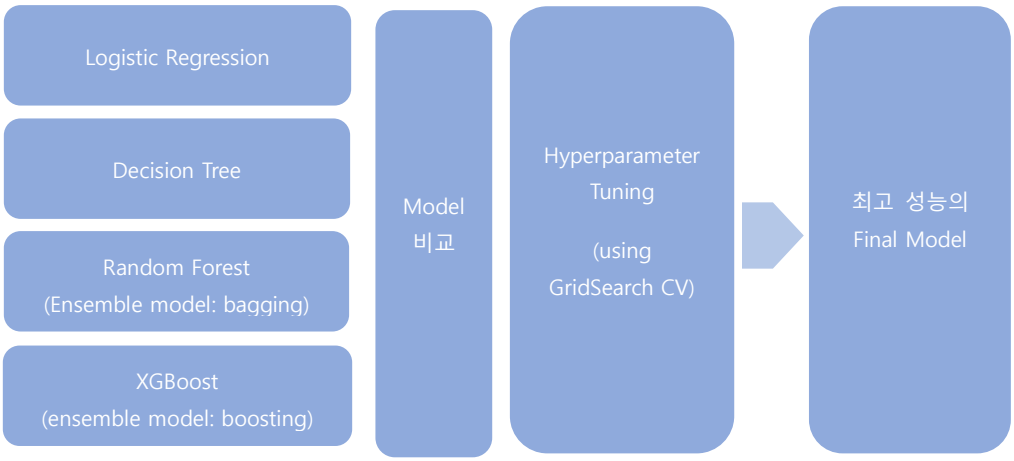
(2) 모델의 예측 reliability 를 높이기 위해 Cross-Validation 진행하며 K 값은 5 에 통일하여 진행하였다.

(3) 일단 먼저 모델평가를 진행하기 전에 US 와 EU 기준에 따른 모델의 baseline 을 확인해보았다.

- US 의 baseline 은 Baseline accuracy(US): 0.52 %, Baseline AUC_score(US): 0.5%
  EU 의 baseline 은 Baseline accuracy(EU): 0.67% , Baseline AUC_score(EU): 0.5%

  ➢ Baseline 0.5 설정

[Modeling 파이프라인]



1) Logistic Regression

선형 모델로서 연속형 변수 간의 분포가 치우치지 않는 것이 중요한 요인이다. 확인된 중성 지방의 분포가 치우쳐져 있어 log transform 을 통해 분포를 적절하게 조절하는 데이터 전처리를 진행한다.

[LR 모델 파이프 라인]

1. Log transformation 처리 된 데이터 input	2. MinMax Scaling (Scaler)	3. Model Fitting (Logistic Regression)	4. Evaluation (Cross Validation, CV=5)

2) Decision Tree

[DT 모델 파이프라인]

1. 위의 preprocessing 진행한 final_data input	2.Hyperparameter 설정: Random state = 42	3. Model Fitting (Decision Tree)	4. Evaluation (Cross Validation, CV=5)

3) Random Forest

[RF 모델 파이프라인]

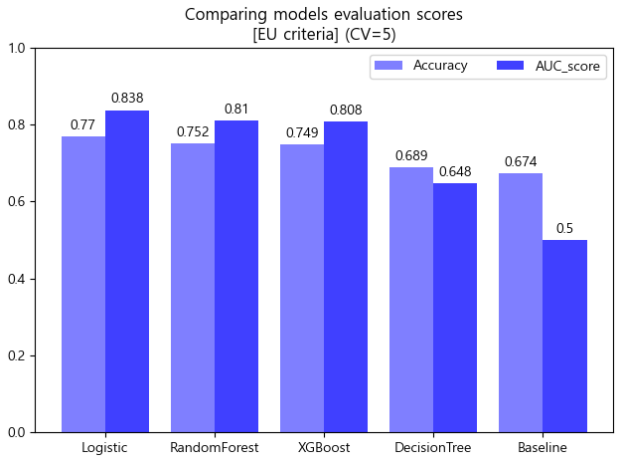
1. 위의 preprocessing 진행한 final_data input	2.Random state = 42 N_jobs = -1	3. Model Fitting (Random Forest)	4. Evaluation (Cross Validation, CV=5)

4) XG Boost

[XGB 모델 파이프라인]

1. 위의 preprocessing 진행한 final_data input	2.Random state = 42 N_jobs = -1	3. Model Fitting (XG Boost)	4. Evaluation (Cross Validation, CV=5)

5. Model Comparison



[그림 8]

최적의 모델을 찾기 위해 4 가지의 모델링을 진행하고 비교하여 보니, Decision Tree model 의 Evaluation score 가 가장 낮다.

그리하여 Decision Tree model 을 제외한 나머지 3 개의 모델로

Grid Search CV 를 활용한 Model Hyper parameter tuning 을 진행하였다.

	Before_accuarcy	After_accuarcy	Before_AUC	After_AUC
XGBoost	0.748713	0.771917	0.808364	0.838031
Logistic	0.769618	0.770384	0.837712	0.837813
RandomForest	0.752436	0.772026	0.809834	0.834052

[그림 9]

Logistic Regression, Random Forest, XGBoost

3 가지의 모델을 GridSearchCV 를 활용한

Hyperparameter Tuning 을 각 특성에 맞게 진행해준 후,

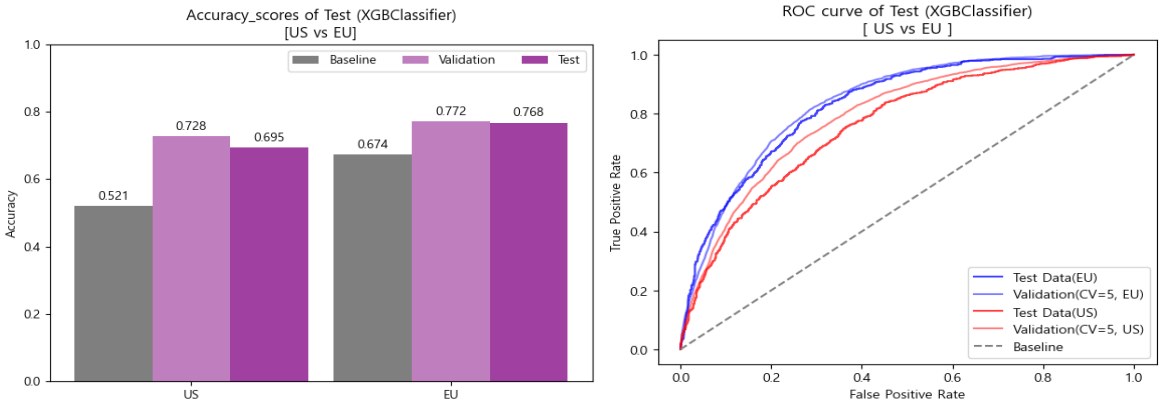
Accuracy 와 AUC score 를 확인하였다.

그 결과, XGboost 모델의 성능이

가장 높게 도출되어, 그에 따른 최종 모델로 선정하였다.

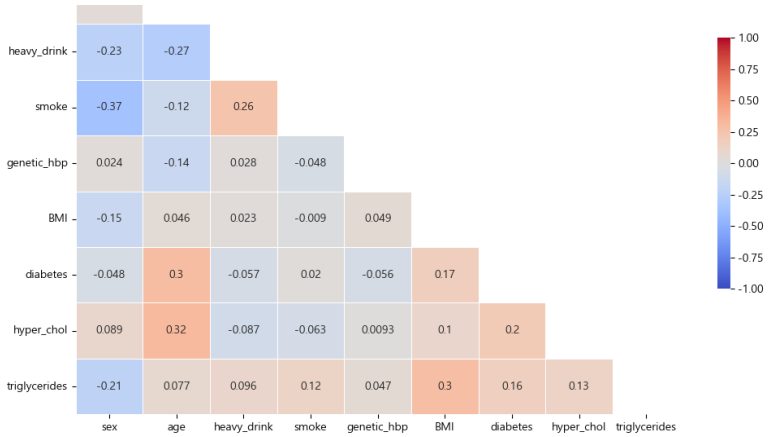
제 4 장 결론

1. 모델링 결과 분석 및 특성 분석



[그림 10,11] 최종 모델에 따른 test 검증, 일반화 가능성을 진행한 시각화 자료

- 최종 모델에서 또한 EU 진단 기준이 US 진단 기준 보다 높은 성능을 나타낸다.
- 종합해서 보았을 때에 성능 score 와 일반화(generalization) 가능성 전반적으로 유럽의 진단 기준이 한국인 특성 고혈압 진단 기준에 더 적합하다고 판단된다.



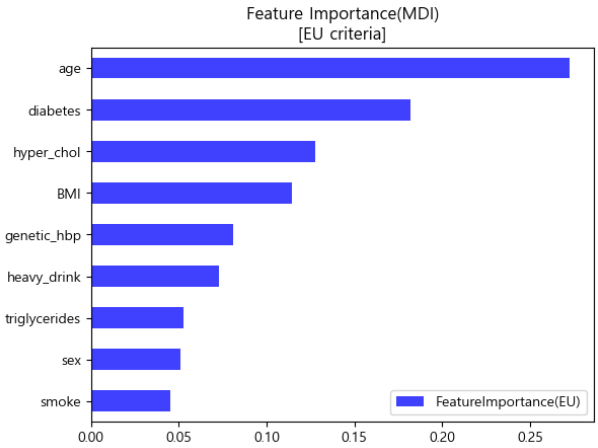
[그림 12]

특성 중요도 분석을 하기 전, 상관관계 확인을 통해 상관관계가 큰 관계 가 존재하면 해석에 유의해야 하므로 시각화를 진행하여 확인하였다.

확인 결과, 변수 간의 상관관계가 크지 않아 해석에 유의하지 않아도 된다는 판단을 내렸다.

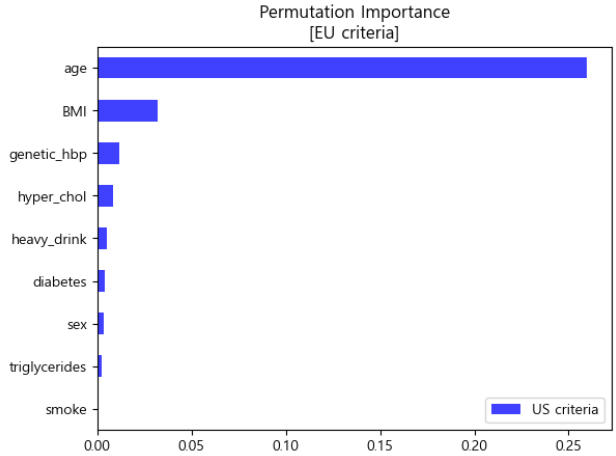
(= 변수 들 간의 독립성이 있다)

(1) MDI (특성중요도)



[그림 13] 특성 중요도는 트리 기반 모델에서 사용되는 중요도 용어로서 평균적인 불순도 감소량에 따른 특성 중요도를 나타낸다. 그렇지만 해당 측정방법은 High cardinality 특성에만 높은 값을 부여하는 문제가 있기에 나이(age)가 가장 큰 영향을 미친다는 것을 고려하고, 나머지 측정치는 제외한다.

(2) Permutation Importance(순열 중요도)



[그림 14] 순열 중요도는 선정된 모델이 어떤 특성(변수)에서 가장 의존하고 있는지를 이해하기 위해 선정하였다. 하나의 특성에 노이즈를 주어 얼마나 성능이 변하는 지 살펴보는 방법으로서 해당 그래프를 확인해보니 나이, 비만도(BMI), 가족력을 가장 중요한 특성으로 나타낸다. 0 에 수렴하는 흡연 여부는 이후에 분석에서 제외하였다.

(3) PDP(Partial Dependencee Plot)

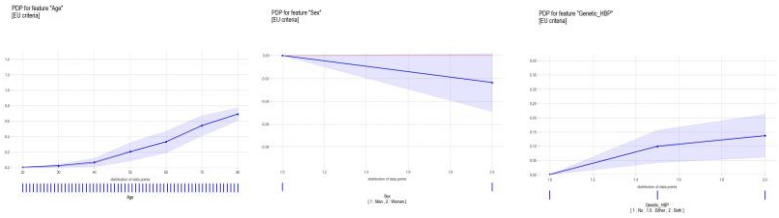
PDP 특성 분석이란 각 변수에 경향성을 파악하기 위해 진행한 방법이다. 평균적으로 특정 변수의 변화에 따라 어떻게 예측하는지를 보여주는 그래프이다. 해당 분석에서 사용되는 그래프는 변수 간의 독립성을 전제로 하므로 PDP 특성 분석이 적절하다고 판단하였다.

04 PDP 특성 분석

조절불가요인
• 나이(연속형)

조절불가요인
• 성별(범주형)

조절불가요인
• 가족력 (범주형)



[그림 15] 조절 불가 요인

나이: 나이가 증가함에 따라 고혈압 발생 위험이 증가함을 시사

성별: 모델이 예측에 미치는 영향이 상대적으로 적지만, 남성이 여성보다 고혈압에 더 잘 걸림을 시사

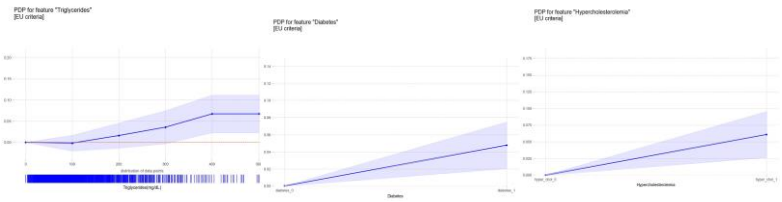
가족력: 유전적 고혈압 유무에 따른 영향력 기준에서 부모 한쪽, 양쪽이 고혈압 유병인 경우에 고혈압 발생 위험이 증가하는 경향성 확인

04 PDP 특성 분석

만성질환요인
• 중성지방(연속형)

만성질환요인
• 당뇨병(범주형)

만성질환요인
• 고콜레스테롤혈증(범주형)



[그림 16] 만성 질환 요인

중성지방: 중성지방 수치가 높을수록, 200 이상으로 증가할 때 고혈압 발생 위험이 증가함을 시사

당뇨병: 당뇨병이 있는 경우 더욱 높은 고혈압 발생 위험이 있다는 것을 시사

고콜레스테롤혈증: 고콜레스테롤 혈증이 있는 경우 고혈압 발생위험이 있다는 것을 시사

04 PDP 특성 분석

생활습관 요인
• 비만

만성질환요인
• 폭음

[그림 17] 생활 습관 요인

비만: BMI(체지방수치)가 증가함에 따라 고혈압 발생 위험도가 증가하는 경향 확인

폭음: 폭음을 하는 경우인

heavy drink == 1 변수를 확인했을 때에, 고혈압 발생위험이 증가하는 경향을 확인

2. 중요도 결과 요약

- 특성중요도(순열중요도): (왼쪽으로 갈수록 높은 중요도)
  - 미국 진단기준 (130mmHg/80mmHg)을 따를 때의 중요도 순위는 다음과 같다.

  - 유럽 진단기준 (140mmHg/90mmHg) 따를 때의 변수 중요도 순위

결과 분석

본 연구에서는 미국과 유럽의 진단 기준을 바탕으로 고혈압 예측모델을 개발하였다. 변수 중요도에 따른 주요 변수로서 나이, BMI, 가족력을 고려하였고, 이 중 나이가 가장 주요한 요인으로 밝혀졌다. 최종 모델을 튜닝 된 XGboost 모델로 설정하였으며, 유럽 진단 기준에 따른 모델의 성능이 미국 진단 기준보다 높음을 확인했다.

모델의 해석을 위해 MDI, PI, PDP 경향분석을 사용하여 특성 중요도와 경향성을 분석하였다. 이 분석 결과들에 따르면 연령이 높을수록, 남성일 경우, 가족력이 강할수록, 고혈압의 예측 확률이 증가한다. 또한, 중성지방 수치가 200 이상, 고콜레스테롤혈증 진단, 당뇨병 유병여부 등이 고혈압 예측 확률을 증가시킨다. 생활습관 요인으로 BMI 지수가 중요한 역할을 하며, 특히 BMI 지수가 25 이상일 경우 고혈압 예측 확률이 급격히 증가한다. 폭음 또한 고혈압 예측 확률을 높이는 요인으로 확인된다.

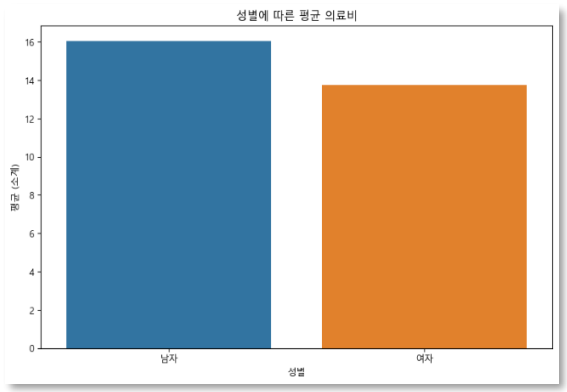
결론적으로, 이 고혈압 예측 모델은 고혈압 진단에 중요한 지표를 제공한다. 이러한 결과 값을 바탕으로 개인이나 의료 전문가들은 고혈압 위험을 줄이기 위해 다음과 같은 변화를 고려할 수 있다. 체중관리를 통해 BMI 를 적정 범위내로 유지하고 식단 조절과 규칙적인 운동이 필요하다. 위와 비슷한 이야기로서 건강한 식습관 관리를 통해 중성지방 수치와 콜레스테롤 관리를 위해 지방이 적은 식단을 유지하고 과일과 채소 등 식이섬유의 섭취를 늘리는 것이 좋다. 또한 과도한 음주의 빈도를 줄이는 것 또한 필연적이다. 정기적인 건강검진을 통해 가족력이나 기타 위험요인이 있는 경우, 정기적인 건강검진을 통해서 혈압 수치를 모니터링하는 것이 중요하다.

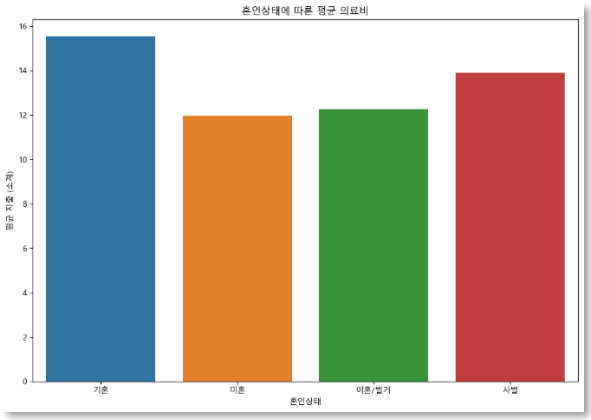
해당 모델을 통해 개인의 생활습관과 건강관리에 대한인식을 높이고, 고혈압 예방 및 관리에 도움이 되는 구체적인 조치를 취할 수 있을 거라 예상할 수 있다.

제 5 장 결과에 따른 의료비 분석(시각화)

고령화 사회에 접어들면서, 나이가 고혈압 발생의 중요한 예측 요인으로 드러난 점은 주목할 만하다. 이에 따라, 고혈압 치료와 관련된 의료비 부담이 어느 정도인지, 특히 고령인구에 미치는 경제적 영향을 이해하는 것이 중요 해졌다고 해석 가능하다. 이러한 의료비 부담은 개인의 건강 관리 뿐 아니라 공중보건 정책 및 의료시스템 설계에도 영향을 미칠 것이라 생각하여 통계자료를 찾게 되었다.

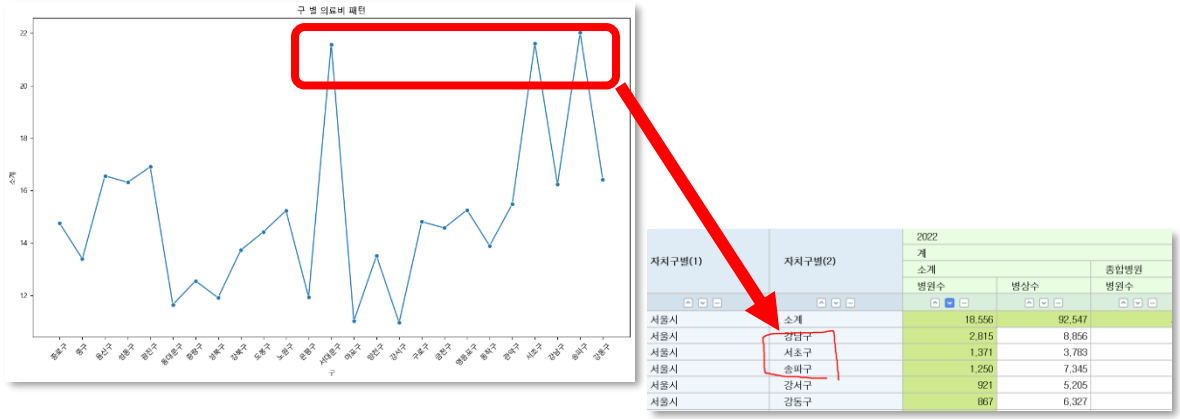
열린 데이터 광장의 "서울특별시 의료비 현황" 데이터를 나이: 50 대 이상, 질병: 고혈압으로 필터링을 진행하여 추가 분석을 진행했다.





[그림 18] 성별에 따른 평균 의료비에서 남성의 고혈압 유병률이 높음을 다시 확인가능 하다.

[그림 19] 혼인 상태에 따른 의료비 분석에서는 기혼자들이 더 많은 의료비를 지출하는 경향이 나타났으며, 이는 가족의 건강관리와 관련이 있을 것으로 추정된다.



[그림 20]지역 별 분석에서는 송파구, 강남구, 서초구가 높은 의료비 지출을 보였는데, 다른 열린 데이터 광장의 "서울특별시 의료기관 현황"데이터를 확인해보니, 이는 해당 지역이 다른 지역에 비해 상대적으로 대형병원의 수가 많은 것과 연관이 있을 것이라 판단하였다.

제 6 장 한계점

해당 프로젝트의 한계점으로는

운동부족, 스트레스, 나트륨과 당류 과잉 섭취 등 고혈압에 영향을 미칠 수 있는 중요한 변수들을 사용하지 못한 점이 있다. 이러한 변수들의 객관적 측정 및 수치화가 자세하게 필요하다고 판단된다. 또한 미국 진단 기준에 따른 가족력 정보의 부재도 한계점으로 지적할 수 있습니다. 더하여, 관련한 상세 의료비 관련 데이터가 존재했다면 고혈압 진단 및 의료비 예측 모델링을 진행하여 실용적이고 폭넓은 예측 모델링을 진행할 수 있지 않았을 까에 대한 아쉬움을 남긴다.

참고문헌

(국가건강영양조사 데이터활용 참고)

https://knhanes.kdca.go.kr/knhanes/sub04/sub04_04_01.do

https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artiId=ART002779811

(전체 flow): https://www.mdpi.com/2075-4418/9/4/178

(model – decision tree) https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full

(model-random forest) https://www.sciencedirect.com/science/article/pii/S0169136815000037?casa_token=ay2iO8vDy88AAAAA:uGKUqNHqsVCVK9Zbpd__yfFoWp3FoPNuA4nfup8GXtslqeIfsN1QLn2CVEzTPcIxfBarvXA

(model-xgboost) https://link.springer.com/article/10.1007/s41666-020-00077-1

(고혈압영향요인 참고 1) https://europepmc.org/article/med/2468976

(고혈압영향요인 참고 2) https://link.springer.com/article/10.1007/s11906-014-0483-3