Section2 Project

뇌졸중 걸릴 가능성이 있는
지에 대한 예측

박윤아

목차

1. 문제제시

1)뇌졸중

-뇌에 혈액을 공급하는 혈관이 막히거나 터지면서 뇌가 손상되는 질환.

-언제 어디서 찾아올지 모르고 생존해도 치명적인 후유증이 남을 수 있음

-국내 3대 사망원인 중 하나

1. 문제제시

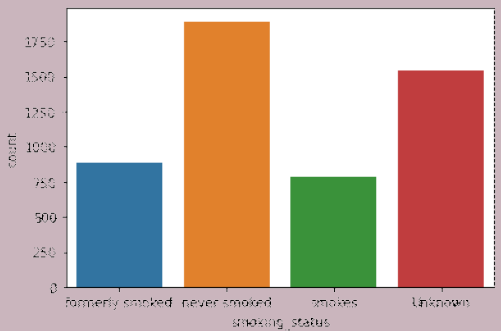
2) 데이터 설명

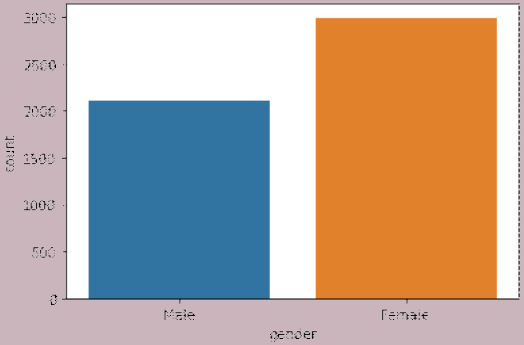
- 5110명의 환자의 성별, 나이, 고혈압 유무, 결혼 여부 등등이 나타난 데이터 셋

- BMI 201개가 값이 나타나있지 않아 임의로 BMI평균값으로 처리

->뇌졸중 걸릴 가능성이 있는지에 대한 예측

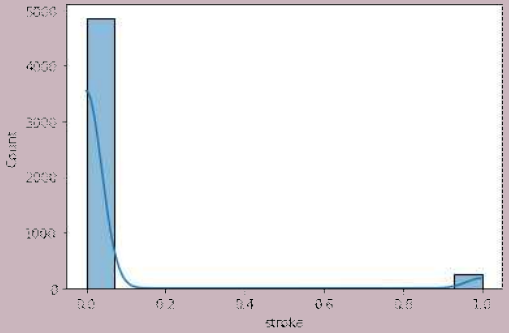
데이터 분포

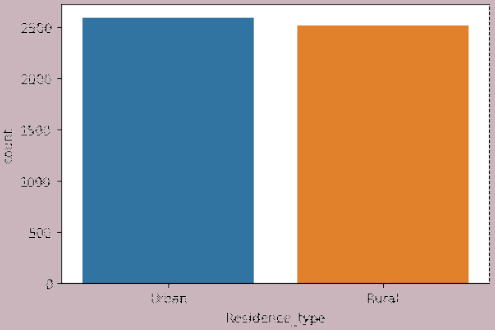




흡연을 하는지 안하는지 알수 없는 자와 비흡연자의 수가 높고
여성의 수가 더 많다는 것을 알 수 있었다.

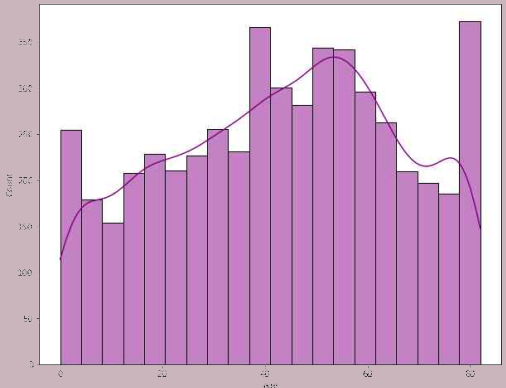
데이터 분포

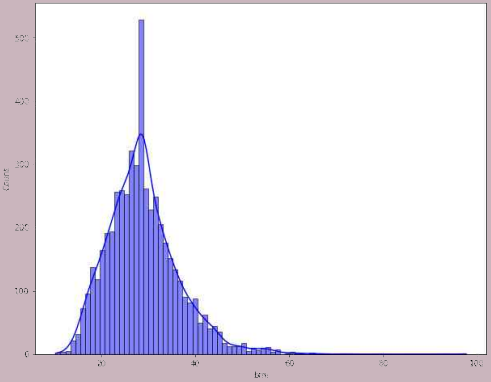




뇌졸중에 걸린 사람과 걸리지 않은 사람의 데이터 불균형이 크다는 것과
도시에 사는 사람과 시골에 사는 사람의 수가 비슷하다는 것을 알게 되었습니다.

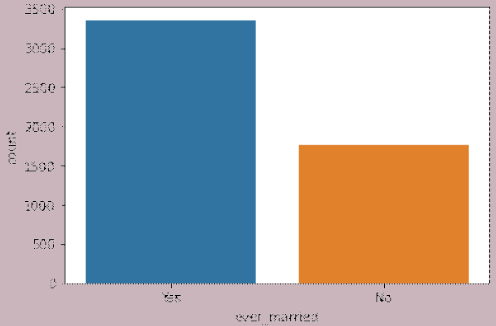
데이터 분포

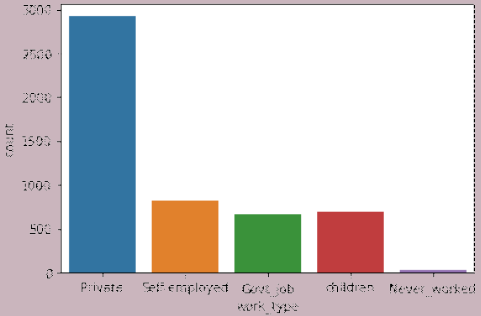




나이는 78세가 많았음을 알 수 있었습니다.
Bmi지수도 28.893237에서 높은 수치를 보인다는 것을 알 수 있었습니다.

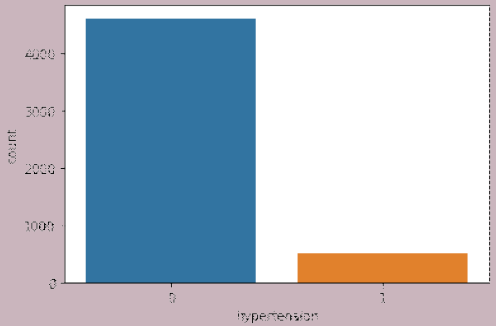
데이터 분포

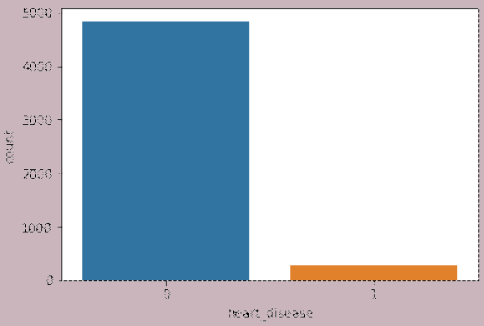




결혼비율에서는 결혼을 한 사람의 수가 많았고,
직업의 분포에서는 공적업무를 하는 사람이 많음을 알 수 있었습니다.

데이터 분포





고혈압 유무는 고혈압이 없는 사람,
심장 질환 유무에서는 심장 질환이 없는 사람이 많았습니다.

2. 가설설정

1

나이 많을수록

뇌졸중에 걸릴 가능성이 높다

2

평균 혈당 지수가 높을수록

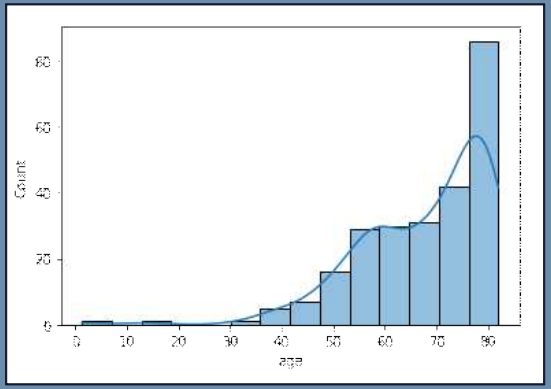
뇌졸중에 걸릴 가능성이 높다

3

체질량 지수가 높을수록

뇌졸증에 걸릴 가능성이 높다

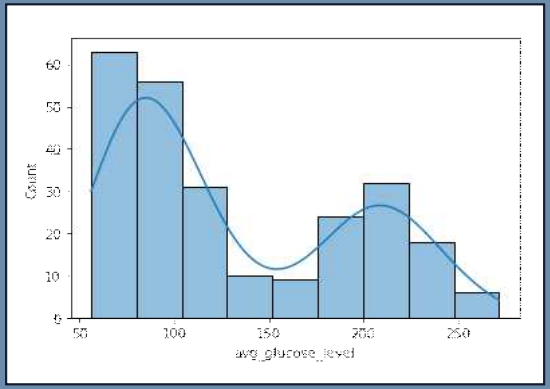
2. 가설설정



가설 1. 나이가 많을수록 뇌졸중에 걸릴 가능성이 높다?

분포도를 통해 age가 높을수록 뇌졸중에 걸릴 가능성이 높다는 가설은 채택되었다.

2. 가설설정

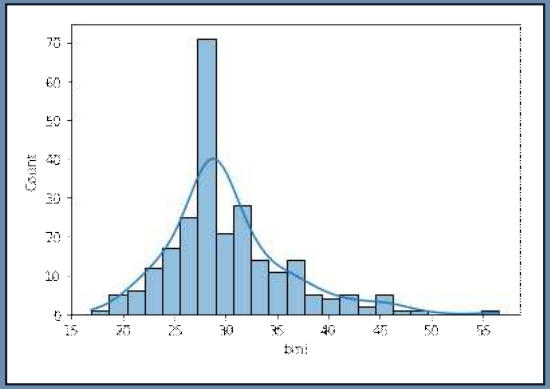


가설 2. 평균 혈당지수가 높을수록 뇌졸중에 걸릴 가능성이 높다

-분포도를 통해 혈액내 평균 혈당지수가 높을수록 뇌졸중에 걸릴 가능성이 높다는 가설은 기각되었습니다.

-50~100사이와 200부근에 많이 분포한다는 것을 알 수 있습니다.

2. 가설설정



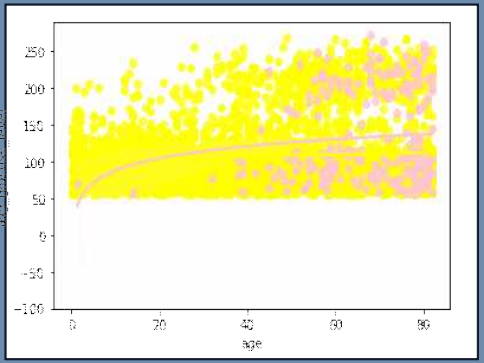
가설 3. 체질량 지수가 높을수록 뇌졸증에 걸릴 가능성이 높다

-체질량 지수가 높을수록 뇌졸중에 걸릴 가능성이 높다는 가설은 기각되었습니다.

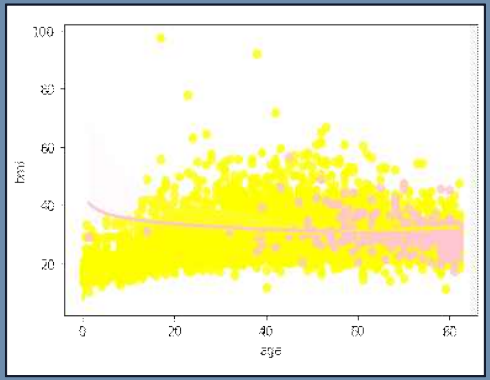
-bmi지수가 25~30사이에 많이 분포되어 있음을 알 수 있다.

2. 가설설정

나이 vs 평균혈당수치



나이 vs BMI



핑크색의 분포를 통하여 나이와 혈당수치, BMI는 연관이 있음을 알 수 있었습니다.

3. 모델학습

데이터 불균형?
- SMOTE

데이터의 개수가 적은 클래스의 표본을 가져온 뒤 임의의 값을 추가하여 새로운 샘플을 만들어 데이터에 추가하는 방식이다.

Base line:

기준모델
-정확도: 95 %
-재현율: 4%

-> 뇌졸중 예방을 목적,
재현율 점수가 중요

	Random Forest	Logistic Regression	
cv_	94%	82%	12
F1 score _	14%	23%	9

3. 모델학습(RandomForest)

	precision	recall	f1-score	support
0	0.94	0.91	0.93	960
1	0.11	0.18	0.14	62
accuracy			0.87	1022
macro avg	0.53	0.54	0.53	1022
weighted avg	0.89	0.87	0.88	1022

1. 재현율 점수: 0.8659491193737769
2. 재현율 점수 : 0.87279843444227

GridSearchCV -> 교차 검증 기반,
알고리즘의 예측 성능을 개선해보았습니다.

	precision	recall	f1-score	support
0	0.94	0.92	0.93	960
1	0.10	0.13	0.11	62
accuracy			0.87	1022
macro avg	0.52	0.52	0.52	1022
weighted avg	0.89	0.87	0.88	1022

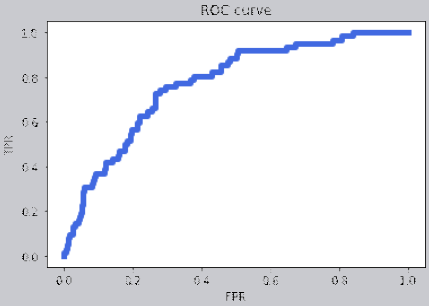
3. 모델학습(Logistic Regression)

	precision	recall	f1-score	support
0	0.97	0.79	0.87	960
1	0.15	0.56		62
accuracy			0.77	1022
macro avg	0.56	0.68	0.55	1022
weighted avg	0.92	0.77	0.83	1022

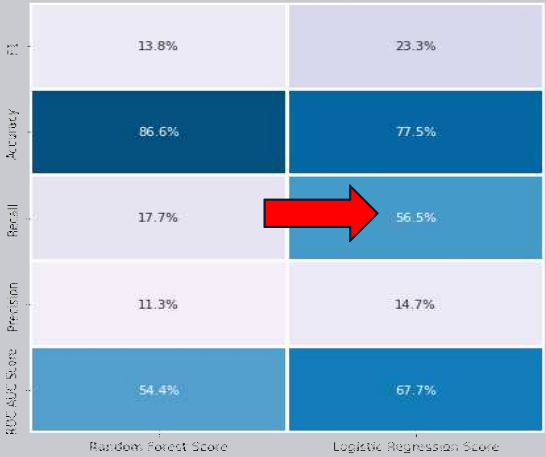
ROC curve에 대한 설명:

x축:뇌졸중 환자에 대해 뇌졸중이라고 진단,
y축: 정상에 대해 뇌졸중이라고 진단

그래프가 좌상단으로 가장 많이 치우친 그래프를 갖는 모델이 높은 성능을 보인다



3. 모델학습(Random Forest vs Logistic Regression)



Random Forest는
Logistic Regression에 비해
정확도가 높았지만
재현율이 낮음

Logistic Regression은
RandomForest에 비해
정확도는 낮지만
재현율이 높음

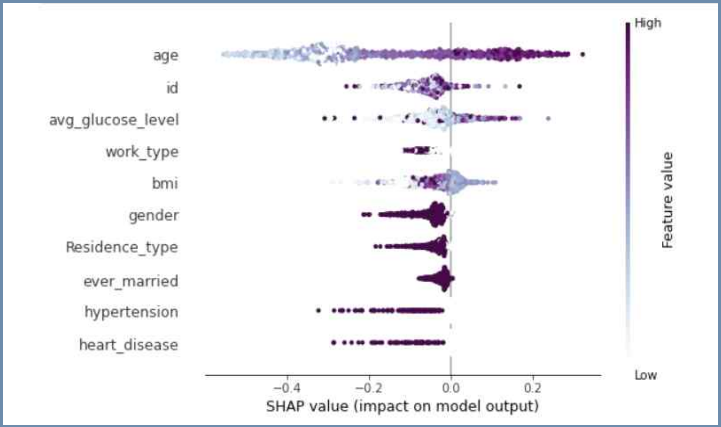
=> 뇌졸중 예방을 목적,
재현율 점수가 중요함

3. 해석 및 결론 (RandomForest)

	Feature	Importance
	age	0.405449
0	id	0.143186
	avg_glucose_level	0.139878
	bmi	0.125112
	work_type	0.077036
1	gender	0.034006
7	Residence_type	0.024185
5	ever_married	0.022593
3	hypertension	0.016663
4	heart_disease	0.011892

나이, 혈당수치, bmi, worktype이
중요한 수치임을 알게 되었습니다

3. 해석 및 결론 (RandomForest)

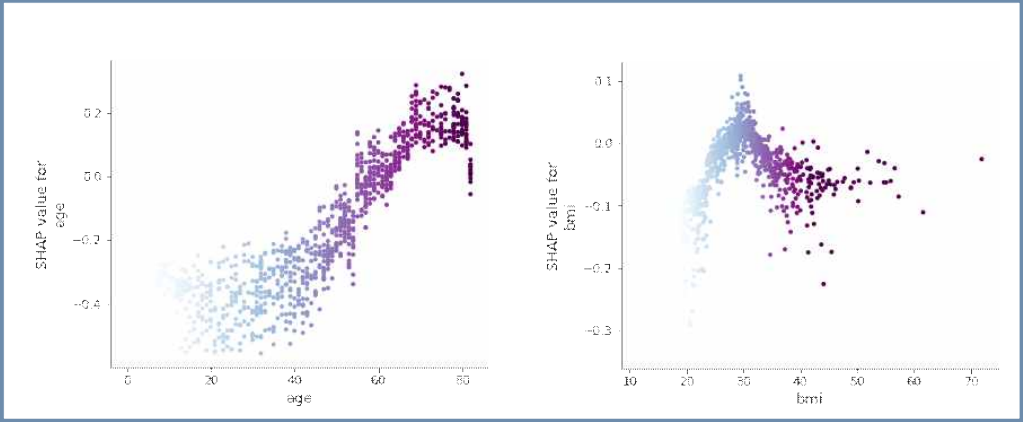


SHAP?

여러가지의 항목을 토대로 얼마만큼의 영향을 주는지 나타내는 지표

뇌졸중을 앓지 않은 경우를 예측하는 쪽으로 크게 치우침을 확인할 수 있었습니다.

3. 해석 및 결론 (RandomForest)



나이가 많을 수록 , bmi가 30이상일 수록 뇌졸중 확률이 높아짐을 알 수 있다.

3. 해석 및 결론 (Logistic Regression)

Weight[?]	Feature
+2.314	bmi
+0.344	hypertension
+0.258	ever_married
+0.128	age
+0.056	heart_disease
-0.164	<BIAS>
-0.329	gender
-0.356	work_type
-0.500	Residence_type
-0.533	avg_glucose_level
-0.551	id

bmi, 고혈압, 결혼여부, 나이, 심장질환여부가 중요한 변수로 확인

3. 해석 및 결론 (RandomForest)

랜덤포레스트와 로지스틱 회귀모델을 통하여 모델들의 하이퍼파라미터 조정을 시도, 결과를 개선해보았습니다.

-> 랜덤포레스트는 높은 정확도를 보였고
로지스틱 회귀 모델은 높은 재현율과 f1 score가 나왔습니다.

랜덤포래스트에서는 나이, 혈당수치, bmi, worktype이 중요한 수치임을 알게 되었습니다
eli5를 사용하여 로지스틱 회귀 모델에서는 bmi, 고혈압, 결혼여부, 나이, 심장질환여부가 중요한 변수로 확인되었습니다.

SHAP을 통해 feature들의 영향을 살펴보면서 나이가 많을 수록 , bmi가 30이상일 수록 뇌졸중 확률이 높아짐을 알 수 있었습니다.

출처

-

https://m.khan.co.kr/life/health/article/202110291033002#c2bhtt
ps://m.khan.co.kr/life/health/article/202110291033002#c2b

감사합니다