통계 모델링 및
컨설팅 2

건강검진 데이터를 이용한 흡연 여부 예측	▼	2조	팀원

배정민	통계학과 2018110476
임혜원	통계학과 2019113415
정유정	경영정보학과 2018111365

INDEX ▼ page

배경

- 흡연이 인체에 미치는 위험성
  : 암, 당뇨병, 고혈압, 고지혈증 등 이외에도 수많은 질환을 유발
- 흡연 여부를 판단 방법
  : 건강검진 시행 시에 작성하는 문진표, 니코틴 직접 측정, 니코틴 간접 측정
  등의 방법으로 흡연 여부를 판단 방법

문제점

- 흡연 항목 미응답 또는 거짓 응답
  → 흡연 여부를 확인 어려움
- 니코틴 직접 측정 또는 코티닌 측정
  → 비용 ↑

목적

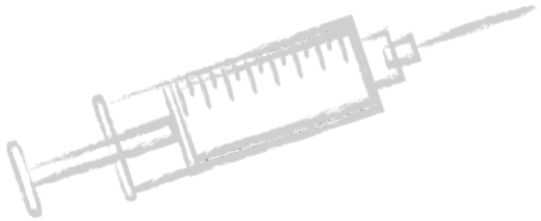
건강검진 데이터를 이용한 흡연 여부 예측 모델

자료 설명

건강보험공단의 건강검진 자료

100만명의 건강검진 결과

100만 행 X 31개 변수



기준년도
가입자 일련번호
시도코드
데이터 공개일자

성별코드
연령대 코드
신장
체중
허리둘레
시력(좌)
시력(우)
청력(좌)
청력(우)

흡연
음주 여부

총 콜레스테롤
트리글리세라이드
HDL 콜레스테롤
LDL 콜레스테롤

AST
ALT
감마 GTP

구강검진 수검여부
충치
치석

공복혈당
혈색소
혈청 크레아티닌
요단백

수축기
이완기

01 변수 제거

- 기준년도
- 가입자 일련번호
- 시도코드
- 데이터 공개일자
- 충치
- 치석
- 구강검진 수검여부

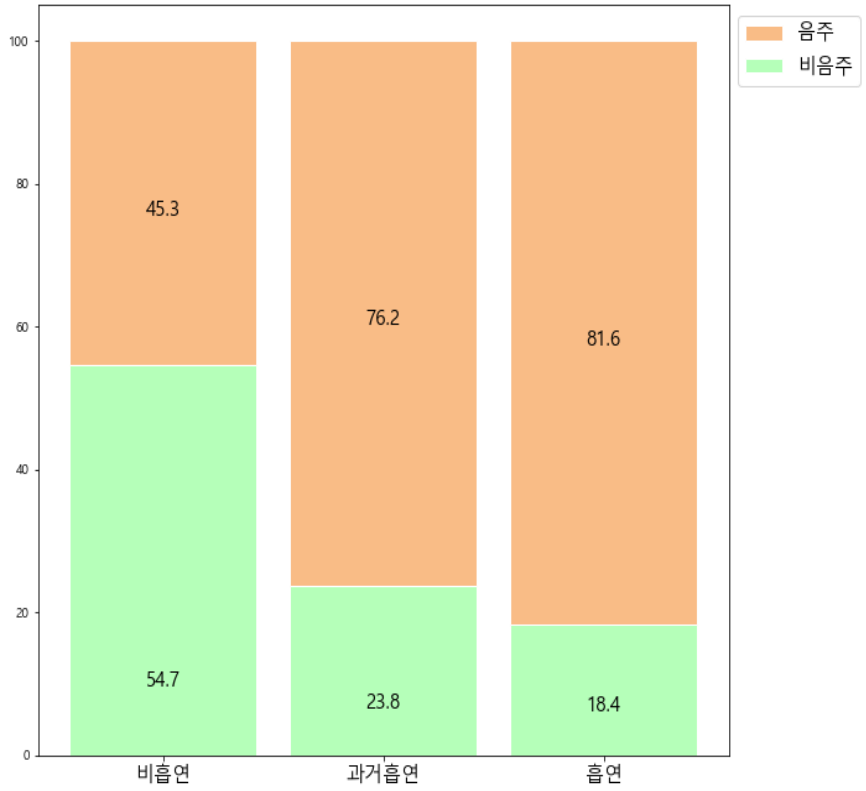
변수	결측치 개수	변수	결측치 개수
~~기준년도~~	~~0~~	식전혈당(공복혈당)	7602
~~가입자 일련번호~~	~~0~~	총 콜레스테롤	597694
~~시도코드~~	~~0~~	트리글리세라이드	597678
성별코드	0	HDL 콜레스테롤	597685
연령대 코드(5세단위)	0	LDL 콜레스테롤	605529
신장(5Cm단위)	0	혈색소	7611
체중(5Kg 단위)	0	요단백	12141
허리둘레	108	혈청크레아티닌	7602
시력(좌)	257	(혈청지오티)AST	7601
시력(우)	252	(혈청지오티)ALT	7602
청력(좌)	222	감마 지티피	7603
청력(우)	230	흡연상태	343
수축기 혈압	7532	음주여부	196
이완기 혈압	7534	~~구강검진 수검여부~~	~~0~~
~~치아우식증유무~~	~~668617~~	~~치석~~	~~668618~~
~~데이터 공개일자~~	~~0~~		

02 결측치 제거

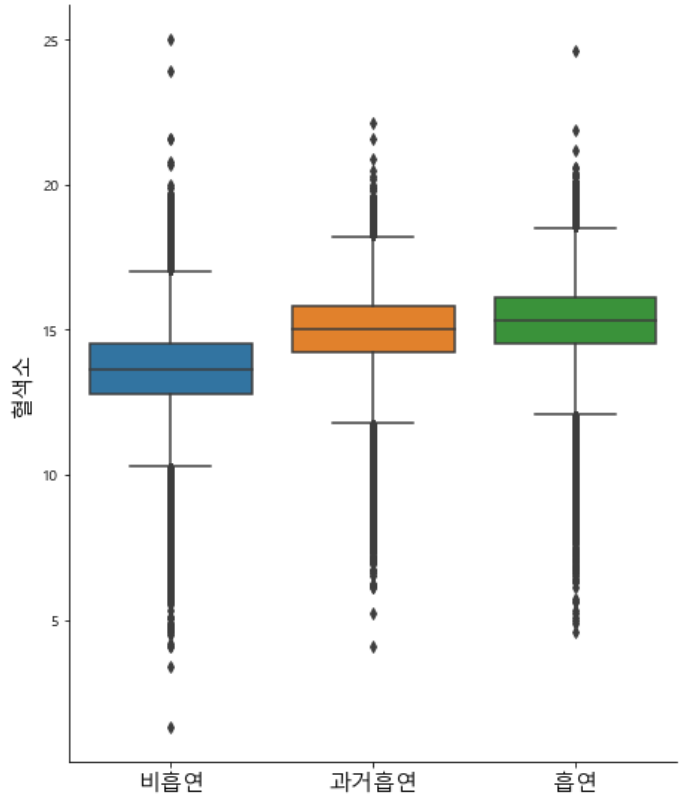
- NA 33개 행
- 총콜레스테롤,
  트리글리세라이드.
  hdl기준 402,139개 행
- 흡연 343개 행

변수	결측치 개수	변수	결측치 개수
성별	0	총콜레스테롤	0
연령대	0	트리글리세라이드	0
신장	0	hdl	0
체중	0	ldl	7842
허리둘레	60	혈색소	7
시력(좌)	120	요단백	2032
시력(우)	116	혈청크레아티닌	0
청력(좌)	98	ast	0
청력(우)	100	alt	1
수축기	10	gpt	1
이완기	10	흡연	0
공복혈당	0	음주여부	62

흡연 그룹 재범주화



흡연 그룹에 따른 혈색소

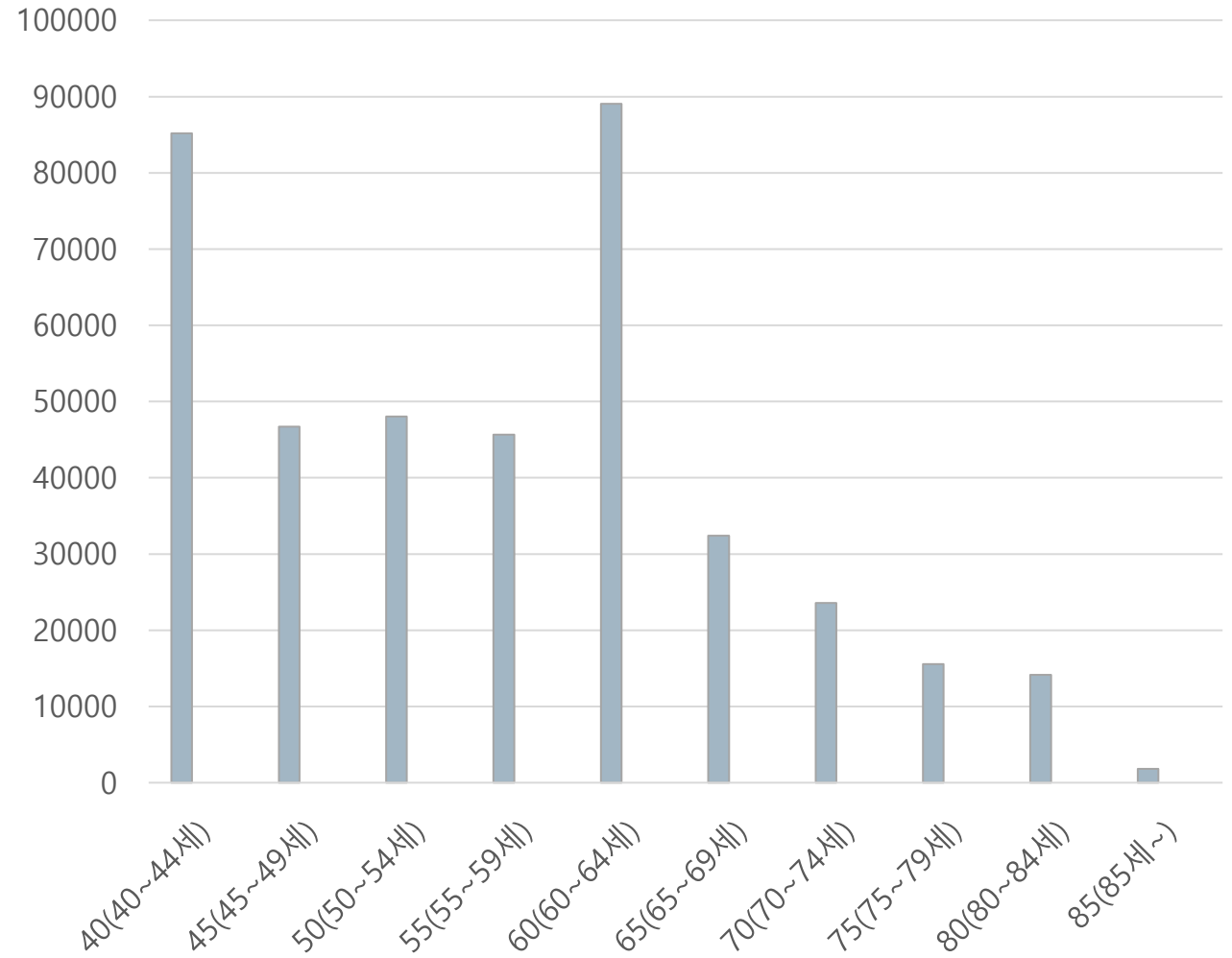


- 기존 범주 :

  비흡연, 과거흡연, 흡연
- 과거흡연 → 흡연
- 재범주화 : 비흡연, 흡연

연령대 재범주화

- 9 → 40
- 10 → 45
- 11 → 50
- 12 → 55
- 13 → 60
- 14 → 65
- 15 → 70
- 16 → 75
- 17 → 80
- 18 → 85



요단백 재범주화

- 1 : 15mg/dL 미만 (정상)
- 2 : 15mg/dL 이상, 30mg/dL 미만 (약산성)
- 3 : 30mg/dL 이상, 100mg/dL 미만
- 4 : 100mg/dL 이상, 300mg/dL 미만
- 5 : 300mg/dL 이상, 1000mg/dL 미만
- 6 : 1000mg/dL 이상

→

- 1 : 15mg/dL 미만 (정상)
- 2 : 15mg/dL 이상 (이상)

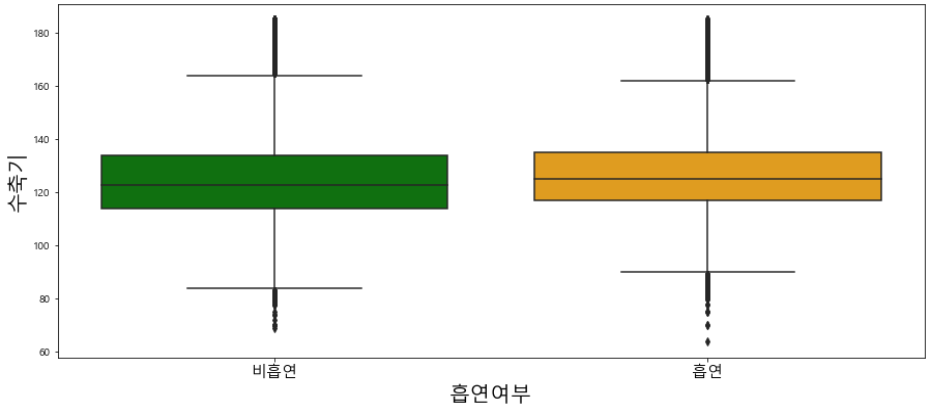
- H0: 요단백과 흡연은 독립이다. vs. H1: 요단백과 흡연은 독립이 아니다.

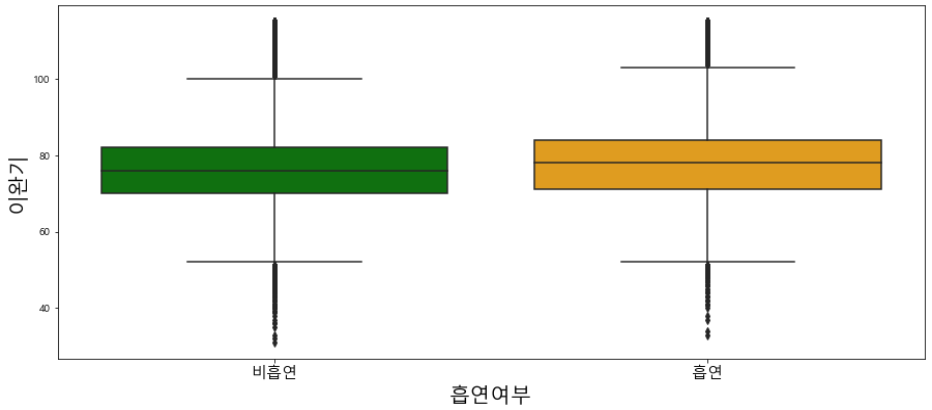
	non_smoker	smoker
정상	238535(236585.27)	130846(132795.73)
이상	17730(19679.73)	12996(11046.27)

Chi-square: 581.7
P-value: 0.0

수치형 변수 – 허리둘레, 시력(좌), 시력(우), 수축기, 이완기

	신장	체중	허리둘레	시력(좌)	시력(우)
count	402139.000000	402139.000000	402079.000000	402019.000000	402023.000000
mean	160.704010	62.751859	82.247045	0.926994	0.929863
std	9.209452	12.300228	9.786217	0.650889	0.664638
min	130.000000	30.000000	8.700000	0.100000	0.100000
25%	155.000000	55.000000	75.900000	0.700000	0.700000
50%	160.000000	60.000000	82.000000	0.900000	0.900000
75%	170.000000	70.000000	88.900000	1.000000	1.000000
max	195.000000	135.000000	999.000000	9.900000	9.900000





01

허리둘레

- 666 cm 이상치 처리 및 제거
- 999cm 이상치 처리 및 제거

02

시력

- 9.9 : 실명
- NA 처리

03

혈압

- 99.9 percentile 기준 이상치 처리
- 수축기 혈압 : 186 이상
- 이완기 혈압 : 116 이상

수치형 변수 – 총콜레스테롤, 트리글리세라이드, hdl, ldl

01

TG < 400mg/dL

Friedewald equation

7842 – 307 = 7535

02

400 <= TG < 800

Sampson equation

7535 – 6601 = 934

03

TG >= 800mg/dL

추정 불가

934 – 934 = 0

04

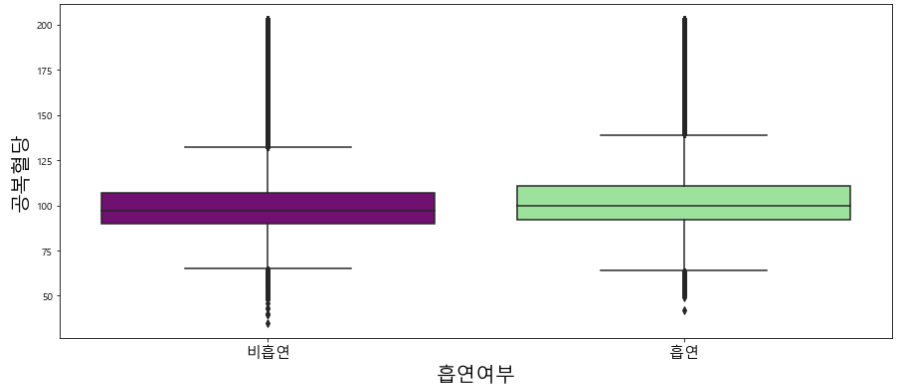
Estimated ldl < 0
&
99 percentile

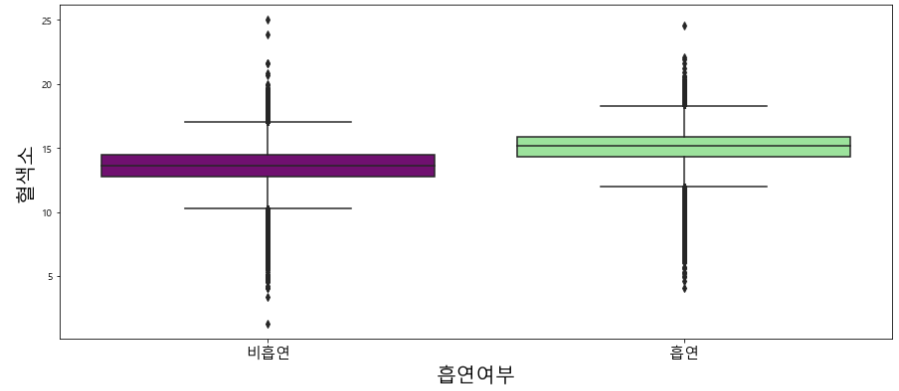
총콜레스테롤 : 301 이상

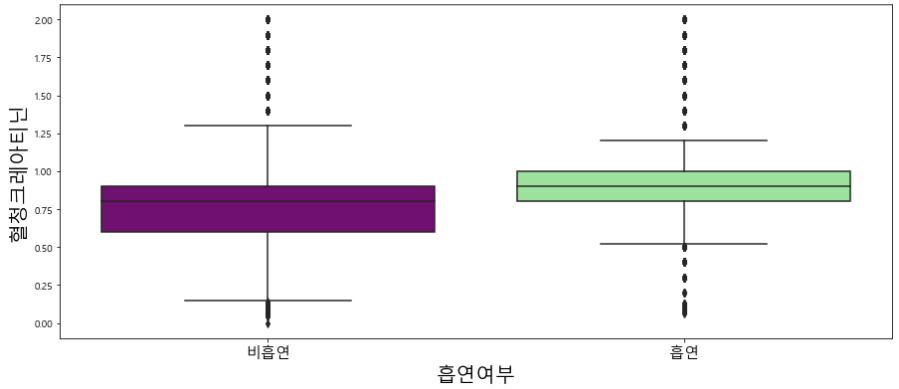
hdl : 98이상

ldl : 331이상

NA 수 = 0







혈중 농도 변수

등분산 검정

H_0 : 흡연, 비흡연 두 집단의 해당 변수의 분산은 같다

세 변수 모두 두 집단에서 분산이 같지 않다.

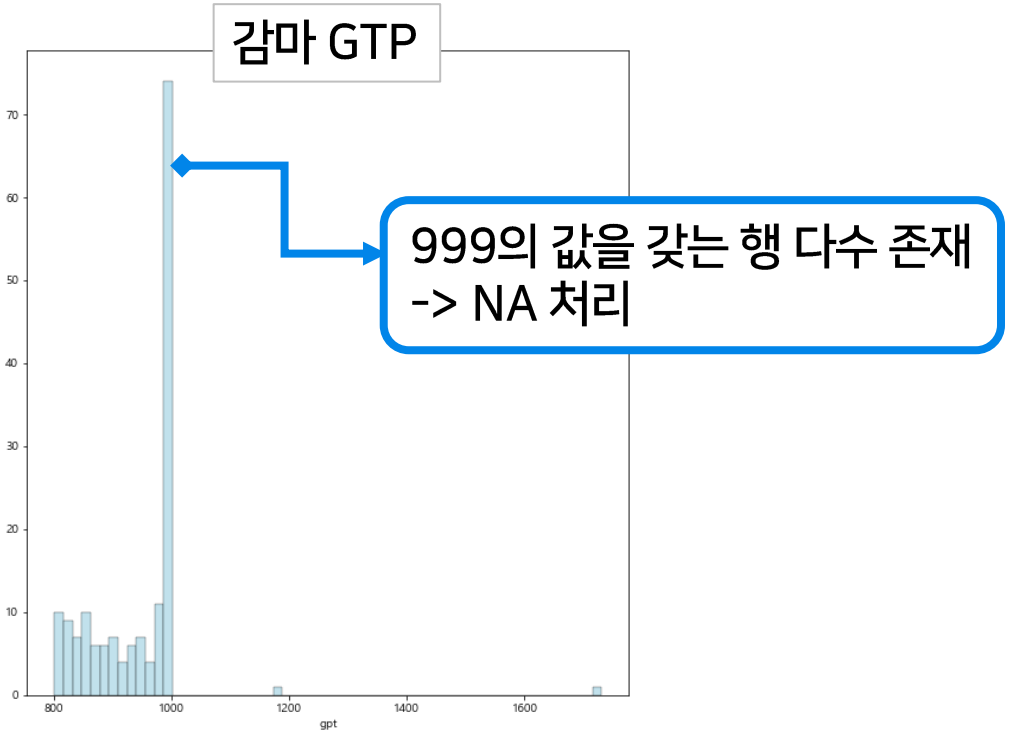
이분산 T test

H_0 : 흡연, 비흡연 두 집단의 해당 변수의 평균은 같다.

세 변수 모두 두 집단에서의 평균은 같지 않다.

따라서 흡연집단이 비흡연 집단보다

큰 수치를 가진다는 것을 확인



간 수치 변수

위험도	수치 범위	원인
경도	40-200 IU/L 정상 상한치의 5배 이내 증가	비알코올성 지방간 질환 만성 B형 간염
중등도	200-400 IU/L 정상 상한치의 5-10배 증가	바이러스 간염 약물에 의한 간 손상
중증	400 IU/L 이상 정상 상한치의 10배 이상 증가	급성 바이러스성 간염 허혈성 및 독성 간 손상 자가면역성 간염 알코올에 의한 간 손상

AST와 ALT의 위험 수준
전체 자료의 1%미만이 중증 범위 -> 행 삭제

비모수 검정 Mann-Whitney U Test

H_0 : 해당 간수치 변수는 흡연여부 집단에 따라 평균의 차이가 없다.

결과 : 세 변수 모두 흡연여부 집단에 따른 평균 값에 차이가 있다.

결측치 처리

01 범주형 변수

Simple Imputer
최빈값 대체

요단백
음주여부

02 수치형 변수

Multivariate Imputation by Chained Equation

허리둘레
수축기
이완기
AST
혈색소
요단백
GTP

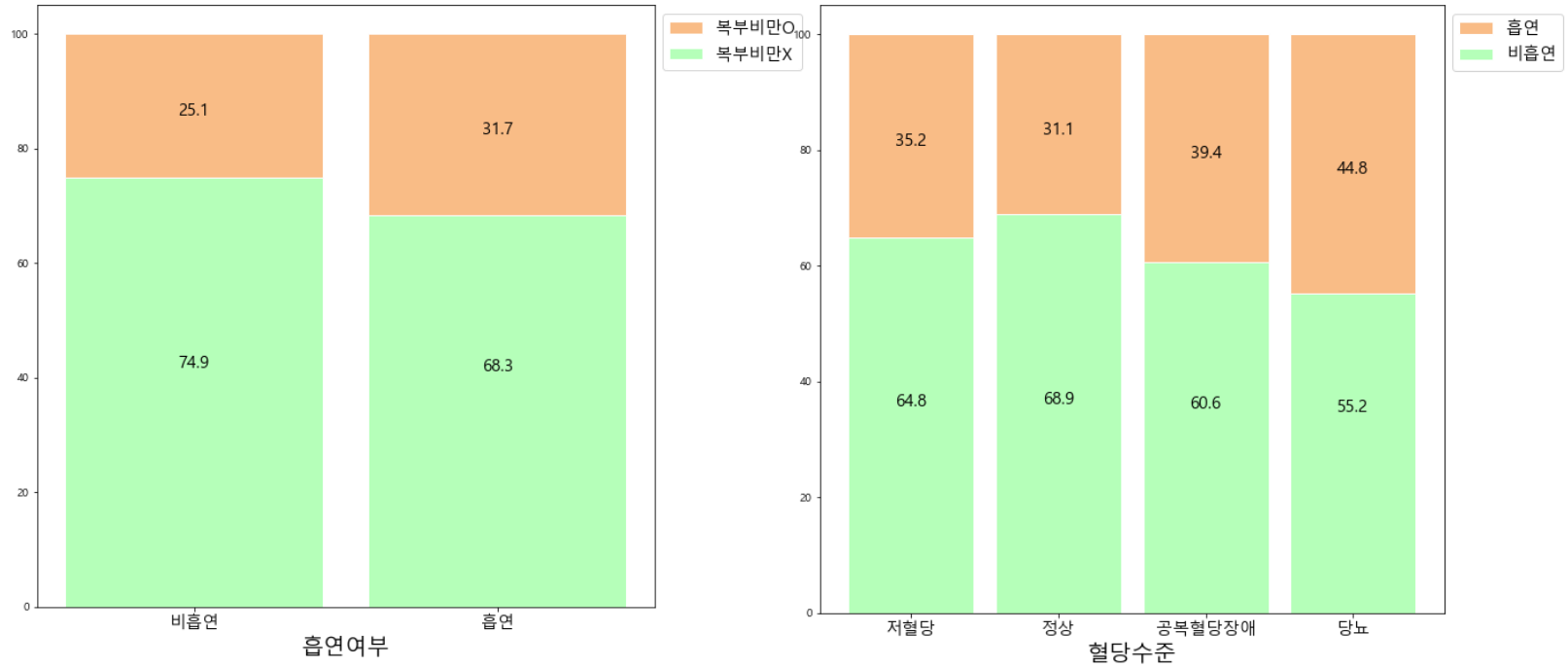
파생변수

01 체질량 지수(BMI)

02 복부비만

03 혈색소_cat

04 혈당수준



Random Forest

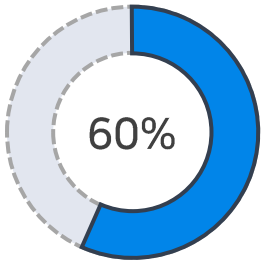
성능 비교 및 중요 변수 파악

XGBoost

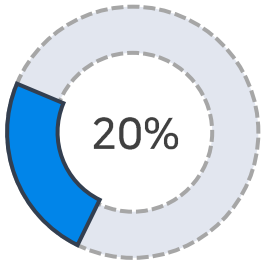
성능 비교 및 중요 변수 파악

총 376874행

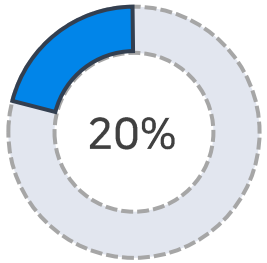
Train / Validation -> 학습과 성능 검증
Test -> 예측



Train



Validation



Test

모델 조정

변수 선택

수축기	BMI
이완기	복부비만
공복혈당	성별
총콜레스테롤	연령대
트리글리세라이드	요단백
HDL 콜레스테롤	음주여부
혈청크레아티닌	혈당수준
ALT	혈색소_cat
감마 GTP	흡연

베이지안 최적화

모델의 하이퍼파라미터
최적화 알고리즘

베이즈 정리를 기반으로
목적함수를 최대화 하는
최적해를 찾아줌

목적함수 정확도로 설정

Random Forest

max_depth = 18

min_samples_leaf = 7

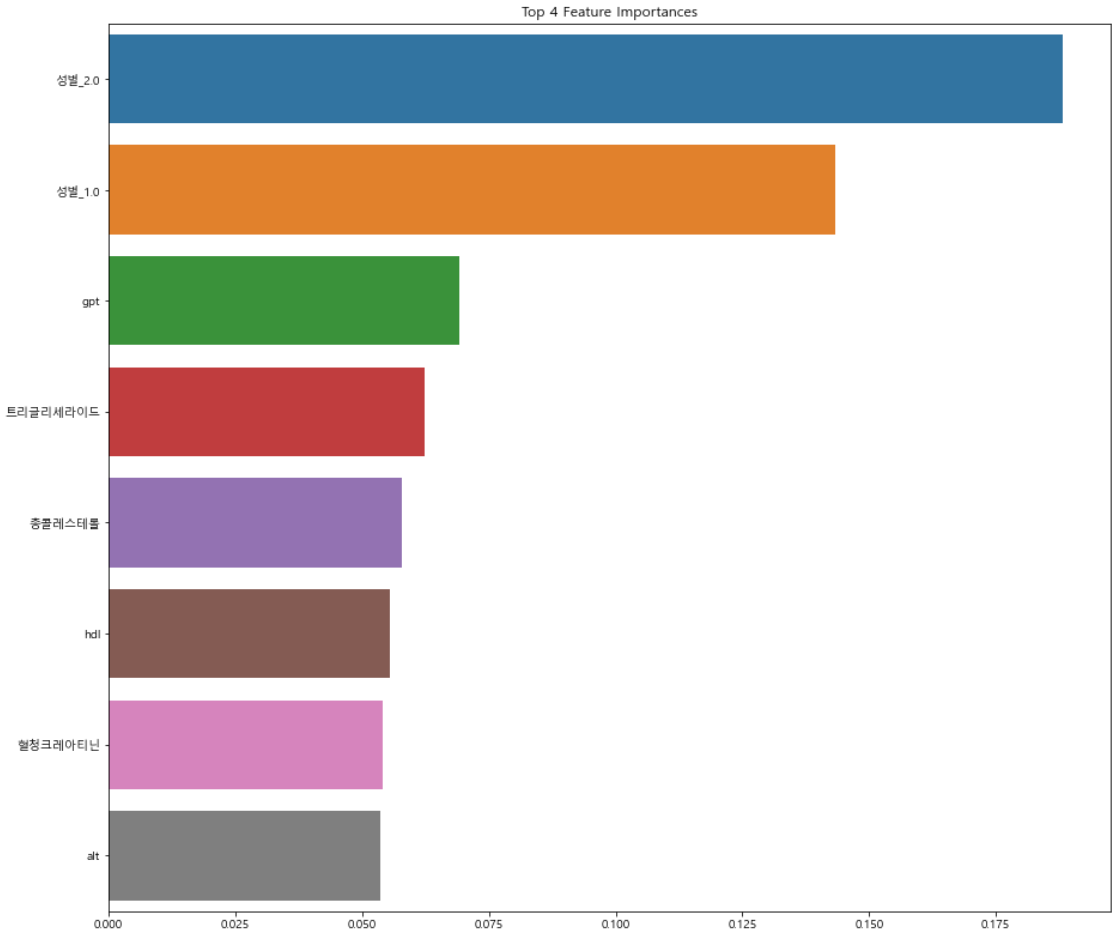
min_samples_split = 11

n_estimators = 258

n_jobs = -1

	Accuracy	Recall	Precision	F1
전	0.818	0.842	0.699	0.764
후	0.82	0.861	0.697	0.77

콜레스테롤, 간수치, 성별의 영향 큼



모델 조정 후의 변수 중요도

XGBoost

colsample_bytree = 0.9

eta = 0.1

gamma = 9

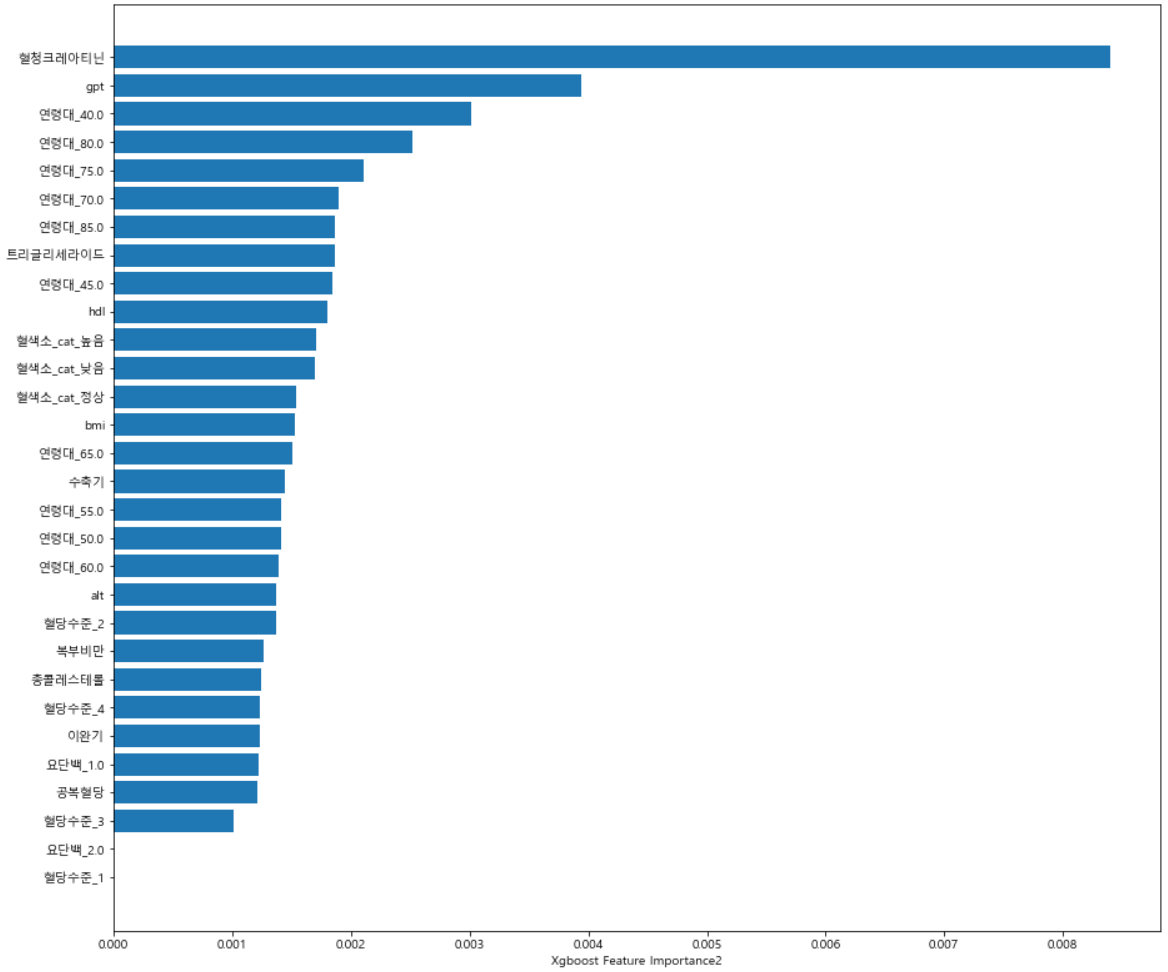
max_depth = 9

min_child_weight = 12

subsample = 0.68

	Accuracy	Recall	Precision	F1
전	0.818	0.875	0.69	0.772
후	0.82	0.856	0.699	0.77

성별, 음주여부, 혈청 크레아티닌, gtp의 영향 큼
특정 연령대의 중요도 높은 편



모델 조정 후의 변수 중요도

최종 모델 선정 – Random Forest

파라미터 조정한 최종 모델에 Test data 적용 한 결과

Accuracy	0.819
Recall	0.859
Precision	0.696
F1	0.769

결론

1) 성별은 흡연을 예측하는데 큰 영향을 준다

2) 음주를 하는 사람 중 흡연자의 비율이 높다

3) 흡연자의 경우 혈색소 수치가 높은 경향이 있다

최종 모델로 Random Forest 선정

모델을 통해 얻은 주요 변수 정보를 이용해

문진이나 별도의 검사 없이 흡연 여부를 예측

흡연자일 확률 높은 집단에게는 추가적인 검진 유도

-> 질병 초기 진단 가능

감 사 합 니 다