

---

## 통계모델링 및 컨설팅2

건강검진 자료를 이용한 흡연 여부 예측

---



---

과 목 명

통계모델링 및 컨설팅2

담 당 교 수

최성은 교수님

2 조

2018110476 배정민

2019113415 임혜원

2018111365 정유정

# INDEX

## I. 서론 : 연구 목적

## II. 본론

1. 분석자료 설명
2. 데이터 전처리
  - 2.1. 데이터 확인
  - 2.2. 범주형변수
  - 2.3. 수치형변수
  - 2.4. 결측치 처리
  - 2.5. 파생변수
3. 분석 방법 및 결과
  - 3.1. Bayesian Optimization을 활용한 RandomForest
  - 3.2. Bayesian Optimization을 활용한 XGBoost
  - 3.3 Logistic Regression

## III. 결론

1. 주요 변수 분석과 인사이트 도출
2. 모형 활용 방안

## IV. 부록

## I. 서론 : 연구목적

흡연이 인체에 미치는 위험성은 꾸준히 제기되고 있으며, 지속적인 흡연은 암, 당뇨병, 고혈압, 고지혈증 등 이외에도 수많은 질환을 유발하는 요인으로 알려져 있다. 이에 따라, 건강검진 시행 시에 작성하는 문진표, 니코틴 직접 측정, 니코틴 간접 측정 등의 방법으로 흡연 여부를 판단하고 있다. 하지만, 문진표 작성자가 흡연관련 항목에 응답을 하지 않거나 거짓응답을 했을 경우에는 문진표 작성을 통해 흡연여부를 확인하는 데 어려움이 생길 수 있고, 니코틴 혈중 농도를 직접 측정하거나 소변이나 혈액에서 코티닌(cotinine)을 측정하는 방법은 비용이 많이 드는 단점이 있다. 본 프로젝트에서는 국민건강보험공단에서 제공하는 건강검진데이터를 이용하여 흡연여부를 예측하는 적합한 모델을 찾아 이러한 어려움과 단점을 해결하는 데 도움이 되고자 한다.

## II. 본문

### 1. 분석자료 설명

본 분석에 사용된 자료는 국가 공공개방 포털에 공개되어 있는 대한민국 건강보험공단의 건강검진 자료이다. 해당 자료는 건강검진을 받은 건강보험 가입자 100만명을 무작위로 선별하여 이들의 일반 건강검진 결과와, 일반 건강검진 대상자 중 만 40세 혹은 만 66세에 도달한 자들이 받는 생애전환기건강진단의 결과로 이루어져 있다.

총 100만개의 행과 31개의 변수로 구성되어 있으며 변수는 크게 건강검진 수진자의 기본정보, 건강검진 결과, 그리고 문진정보로 이루어져있다. 31개 변수의 설명은 다음과 같다.

변수	설명	종류
기준년도	해당 정보의 기준년도	int64
가입자 일련번호	건강보험 가입자에 부여한 일련번호	int64
성별 코드	대상자의 성별 - 1(남자), 2(여자)	category
연령대 코드	대상자의 나이를 5세 단위로 범주화한 코드	category
시도 코드	대상자 거주지의 시도코드	int64
신장	대상자의 키 (5cm 단위) 예) 160~164cm -> 160cm	int64
체중	대상자의 몸무게 (5kg 단위) 예) 25~29kg -> 25kg	int64
허리둘레	대상자의 허리둘레	float64
시력(좌)	대상자의 좌측 시력 - 0.1이하의 시력은 0.1, 실명은 9.9로 표기	float64
시력(우)	대상자의 우측 시력 - 0.1이하의 시력은 0.1, 실명은 9.9로 표기	float64
청력(좌)	대상자의 좌측 귀 청력 - 1(정상), 2(비정상)	category
청력(우)	대상자의 우측 귀 청력 - 1(정상), 2(비정상)	category
수축기 혈압	대상자의 최고 혈압	float64

이완기 혈압	대상자의 최저 혈압	float64
식전혈당 (공복혈당)	대상자의 식사 전 혈당 수치	float64
총 콜레스테롤	혈청 중 에스텔형, 비에스텔형 콜레스테롤의 합	float64
트리글리세 라이드	중성지방 수치	float64
HDL 콜레스테롤	고밀도 리포단백질에 포함되는 콜레스테롤	float64
LDL 콜레스테롤	저밀도 리포단백질에 함유된 콜레스테롤.	float64
혈색소	혈액이나 혈구 속에 존재하는 색소단백	float64
요단백	소변에 단백질이 섞여 나오는 것	float64
혈청 크레아티닌	신기능장애에 의해 증량하는 수치	float64
(혈청지오티) AST	간 기능을 나타내는 혈액검사 상의 수치. 심장, 신장, 뇌, 근육의 세포들이 손상되는 경우 농도 증가함	float64
(혈청지오티) ALT	간 기능을 나타내는 혈액검사 상의 수치. 간세포가 손상되는 경우 농도 증가함	float64
감마 지티피	간 기능을 나타내는 혈액검사 상의 수치. 쓸개즙 배설 장애, 간세포 장애 발생 시 혈중에 증가함	float64
흡연	대상자의 흡연 상태 여부 1(피우지 않는다), 2(이전에 피웠으나 끊었다), 3(현재도 피우고 있다)	category
음주 여부	대상자의 음주 상태 여부 0(마시지 않는다), 1(마신다)	category
구강검진 수검여부	대상자가 구강검진을 선택하여 검했는지 여부 0(미수검), 1(수검)	category
치아 우식증 유무	대상자의 충치 유무 0(없음), 1(있음)	category
치석	대상자의 치석 여부 0(없음), 1(있음)	category
데이터 공개일자	데이터 작성 기준일자	object

## 2. 데이터 전처리

### 2.1. 데이터 확인

#### 2.1.1. 검진 정보 없는 행

성별, 연령대, 신장, 체중, 허리둘레, 시력, 청력정보, 흡연, 음주여부와 건강검진수검여부 외에 15개의 변수에 대한 건강검진 정보가 전혀 없는 행 33개를 삭제하였다.

### 2.1.2. 변수와 결측치 제거

“기준년도”, “가입자 일련번호”, “시도코드”, “데이터 공개일자”의 4가지 변수는 흡연과 관련성이 전혀 없다고 판단하여 제거하였다.

구강검진 수검을 받지 않은 인원 668,583명에 대해 충치와 치석의 존재 여부가 없었으며, 이는 충치와 치석의 약 60%로 분석에 이용하기 어렵고 충치와 치석의 존재여부가 흡연에 큰 영향을 미치지 않을 변수라고 판단하여 “충치”, “치석”, 그리고 “수검여부” 변수를 제거하였다.

데이터의 범주형 변수들은 문진결과를 통해 만들어진 것으로 결측값이 많았으며, 수치형 변수들 중 총 콜레스테롤(Cholesterol), 트리글리세라이드(Triglyceride), HDL 콜레스테롤, LDL 콜레스테롤 변수들은 2008년 이후부터 건강검진의 문진항목으로 추가되면서 2008년까지는 값이 결측이다. 흡연이 콜레스테롤에 미치는 영향과 관련한 학술자료를 통해 흡연여부를 예측하는 데 콜레스테롤 관련 변수들은 중요한 변수라고 판단하였기에 콜레스테롤과 관련된 변수인 “총 콜레스테롤”, “HDL”, “트리글리세라이드”를 기준으로 결측값이 있는 402,139행을 모두 제거하여 2009년부터의 데이터만 사용하였다.

또한, “흡연”은 타겟 변수로 사용될 것이므로 결측값이 있는 343행도 제거했다.

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		

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

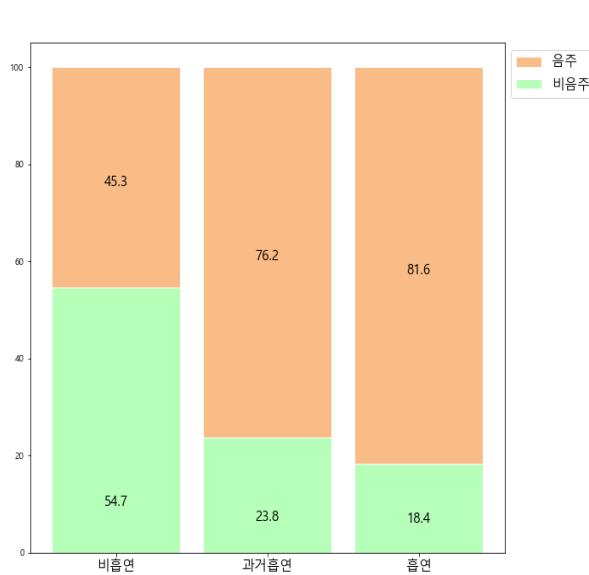
< 표1. 결측치 처리 전 >

< 표2. 결측치 처리 후 >

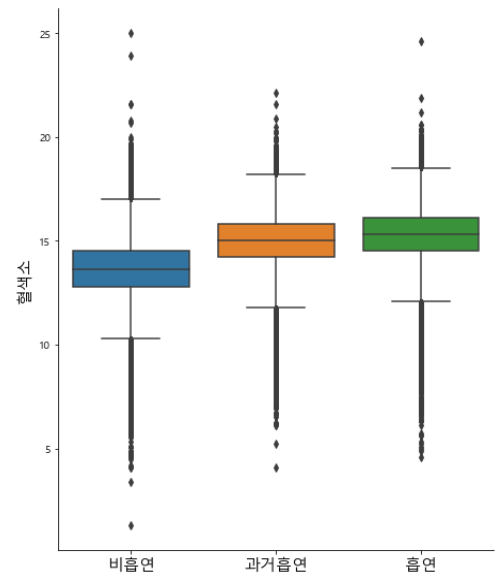
## 2.2. 범주형변수

### 2.2.1. 흡연

흡연 여부를 구분하는 이진분류를 하기 위해서는 타겟 변수인 “흡연”의 재범주화가 필요하였다. 비흡연, 과거흡연, 흡연에 따른 다른 변수들의 비율 및 수치를 확인해 본 결과, 특히 “음주”와 “혈색소”에서 과거흡연 그룹이 비흡연 그룹보다 흡연그룹의 비율 및 수치와 유사함을 보였다. 그리고 서울대학교 건강검진센터에 따르면, 과거흡연자가 비흡연자와 같은 건강상태가 되기 위해서는 21년 이상이 걸린다. 그렇기에 과거흡연 그룹을 흡연그룹에 포함시켜 재범주화하였으며, 비흡연은 0, 흡연은 1로 표현하였다.



< 그림1. 흡연 그룹에 따른 음주여부의 비율 >



< 그림2. 흡연 그룹에 따른 혈색소의 수치 >

## 2.2.2. 연령대

“연령대” 변수는 40세에서 90세까지 5단로 각각 9에서 18까지 범주화되어 있다. 따라서 알아보기 쉽도록 40에서 85로 값을 변환하였다.

## 2.2.3. 요단백

“요단백”은 시험지 검사로 측정된 단백질 수치를 1(정상, ~15mg/dL), 2(약산성, 15~30mg/dL), 3(30~100mg/dL), 4(100~300mg/dL), 5(300~1000mg/dL), 6(1000mg/dL~)으로 범주화되어 있다. 단백질이 하루에 300mg/dL 이상 배출되면 단백뇨로 정의하므로, 5와 6 범주를 하나의 범주 5로 재범주화하였다.

“요단백”의 범주에 따른 흡연과 비흡연의 비율을 비교했을 때, 요단백이 정상일 때의 범주가 비흡연 비율이 가장 많이 나왔으며, 표본 수가 매우 많았다. 이외 다른 범주는 모두 비슷한 수준의 비흡연 비율로 비슷한 경향을 보였다. 이에 따라 “요단백” 변수를 2개의 범주, 15mg/dL 을 기준으로 정상과 비정상으로 재범주해본 결과, 두 집단에서 흡연과 비흡연의 차이가 있지만 확연한 차이는 없어 보였기에 카이스퀘어 독립성 검정을 진행하였다.

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

유의수준 0.05 하에서 기각, 요단백과 흡연은 독립적이지 않고 종속이다. 따라서 2개의 재범주화한 것이 타당하고 흡연과 관계도 있다고 판단하였다.

## 2.3. 수치형변수

### 2.3.1. 신장, 체중, 허리둘레, 시력(좌), 시력(우)

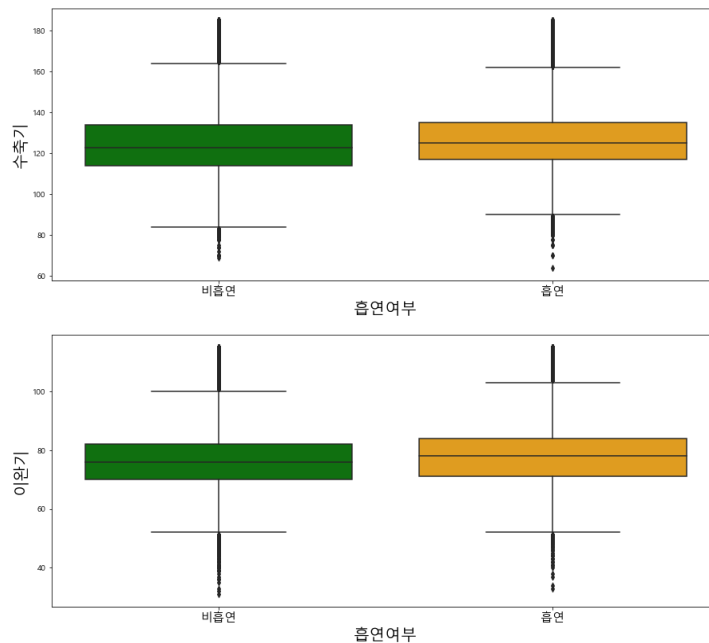
다음은 수치형 변수에 대해 요약된 통계량이다. 허리둘레가 999cm인 최댓값이 있었으며, 시각화를 통해 666cm의 매우 큰 값을 발견하였다. 이 값들은 이상치로 판단하여 결측값으로 처리하였다. 또한, 시력 변수의 값들 중 9.9로 표기되어 있는 값은 실명으로 시력을 측정할 수 없기 때문에 결측값으로 변환하였다.

	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

< 표3. 신장, 체중, 허리둘레, 시력(좌), 시력(우)의 요약 통계량 >

### 2.3.3. 혈압 : 수축기, 이완기

수축기 혈압과 이완기 혈압의 수치는 최댓값이 평균에서 다소 벗어나는 듯 보였으나, 수축기 혈압과 이완기 혈압에 따른 다음의 <표 4>에 따라 . 고혈압 수치보다도 한참 높은 수치에 해당되는 99.9 percentile의 수축기 혈압은 186, 이완기 혈압은 116이상에 해당하는 부분은 이상치로 간주하여 행을 삭제하였다.



< 그림3. 흡연 집단에 따른 수축기와 이완기 혈압 >

이상치 제거 후 “흡연” 변수의 집단에 따른 수축기 혈압 차이를 **boxplot**으로 확인하였을 때 눈에 띄는 차이를 확인하지 못하였기에 이에 대한 차이를 확인하기 위하여 등분산성 검정, 독립표본 T 검정을 진행하였다.

등분산성 검정(Bartlett test)	
H0: 흡연, 비흡연 두 집단의 수축기 분산은 같다. vs. H1: 흡연, 비흡연 두 집단의 수축기 분산은 다르다.	
statistic=1304.7554	p-value=1.046650544969171e-285

유의수준 0.05하에서 귀무가설 H0을 기각할 근거가 충분하며, 흡연과 비흡연 두 집단의 분산은 다르다. 따라서 이분산 독립 표본 T검정을 진행했다.

이분산 독립 표본 T검정(T test)	
H0: 흡연, 비흡연 두 집단의 수축기 평균은 같다. vs. H1 : 흡연, 비흡연 두 집단의 수축기 평균은 다르다.	
statistic=-32.0771	p-value=2.1037716345925608e-225

유의수준 0.05하에서 귀무가설 H0을 기각할 근거가 충분하다. 즉, 흡연과 비흡연 두 집단의 수축기 평균은 다르다.

이분산 독립 표본 T검정(T test)	
H0: 흡연, 비흡연 두 집단의 수축기 평균은 같다. vs. H1: 비흡연 집단의 수축기 평균이 흡연 집단보다 작다.	
statistic=-32.0771	p-value=1.0518858172962804e-225

유의수준 0.05하에서 귀무가설 H0을 기각할 근거가 충분하다. 즉, 비흡연 집단의 수축기 평균이 흡연 집단보다 작다.

다음으로, “흡연” 변수의 집단에 따라 이완기 혈압의 차이를 확인하기 위하여 등분산 검정과 T 검정을 진행하였다.

등분산성 검정(Bartlett test)	
H0: 흡연, 비흡연 두 집단의 이완기 분산은 같다. vs. H1: 흡연, 비흡연 두 집단의 이완기 분산은 다르다.	
statistic=1.6315	p-value=0.2015

유의수준 0.05하에서 귀무가설 H0을 기각할 근거가 충분하지않다. 그러므로 두 집단의 분산은 같다.



이분산 독립 표본 T검정(T test)	
H0: 흡연, 비흡연 두 집단의 이완기 평균은 같다. vs. H1: 흡연, 비흡연 두 집단의 이완기 평균은 다르다.	
statistic=-74.4009	p-value=0.0

유의수준 0.05하에서 귀무가설 H0을 기각할 근거가 충분하다. 즉, 두 집단의 이완기 평균은 다르다.

이분산 독립 표본 T검정(T test)	
H0: 흡연, 비흡연 두 집단의 이완기 평균은 같다. vs. H1: 비흡연 집단의 이완기 평균이 흡연 집단보다 작다.	
statistic=-74.4633	p-value=0.0

유의수준 0.05하에서 귀무가설 H0을 기각할 근거가 충분하다. 즉, 비흡연 집단의 이완기 평균이 흡연집단보다 작다.

따라서 수축기 혈압과 이완기 혈압의 수치는 모두 비흡연자 집단보다 흡연자 집단에서 조금 더 크게 나타났다.

혈압분류	수축기 혈압 (mmHg)		이완기혈압 (mmHg)
정상혈압*	< 120		그리고 < 80
고혈압 전단계	1기	120 - 129	또는 80 - 84
	2기	130 - 139	또는 85 - 89
고혈압	1기	140 - 159	또는 90 - 99
	2기	≥ 160	또는 ≥ 100
수축기 단독고혈압	≥ 140		그리고 < 90

\* 심혈관질환의 발병위험이 가장 낮은 최적 혈압

< 표 4. 수축기혈압, 이완기혈압 수치에 따른 혈압 상태 분류>

#### 2.3.4. 콜레스테롤 : 총콜레스테롤, 트리글리세라이드, hdl, ldl

건강검진데이터에서 LDL콜레스테롤은 2012년부터 계산 값과 측정치 값이 통합되었다. 총콜레스테롤, HDL콜레스테롤, 트리글리세라이드 수치로 계산하되, 트리글리세라이드 측정값이 400mg/dL 이상인 경우 실측정한 값으로 기록되어있다.<sup>1</sup> 이에 따라 “ldl”의 결측 행 7842개 중 트리글리세라이드 측정값이 400mg/dL 미만인 경우 Friedwald equation<sup>2</sup>을 이용하여

<sup>1</sup> 국민건강보험공단, “국가중점 개방데이터(건강검진정보) 사용자 매뉴얼(ver 4.0)”

<sup>2</sup> Friedewald equation : LDL - C = TC - HDL - C - TG/5

307개 행을 대체하였으며, 트리글리세라이드가 400mg/dL 이상이면서 800미만인 경우는 LDL 추정공식 Sampson equation<sup>3</sup>을 이용하여 6601개 행을 대체하였다.

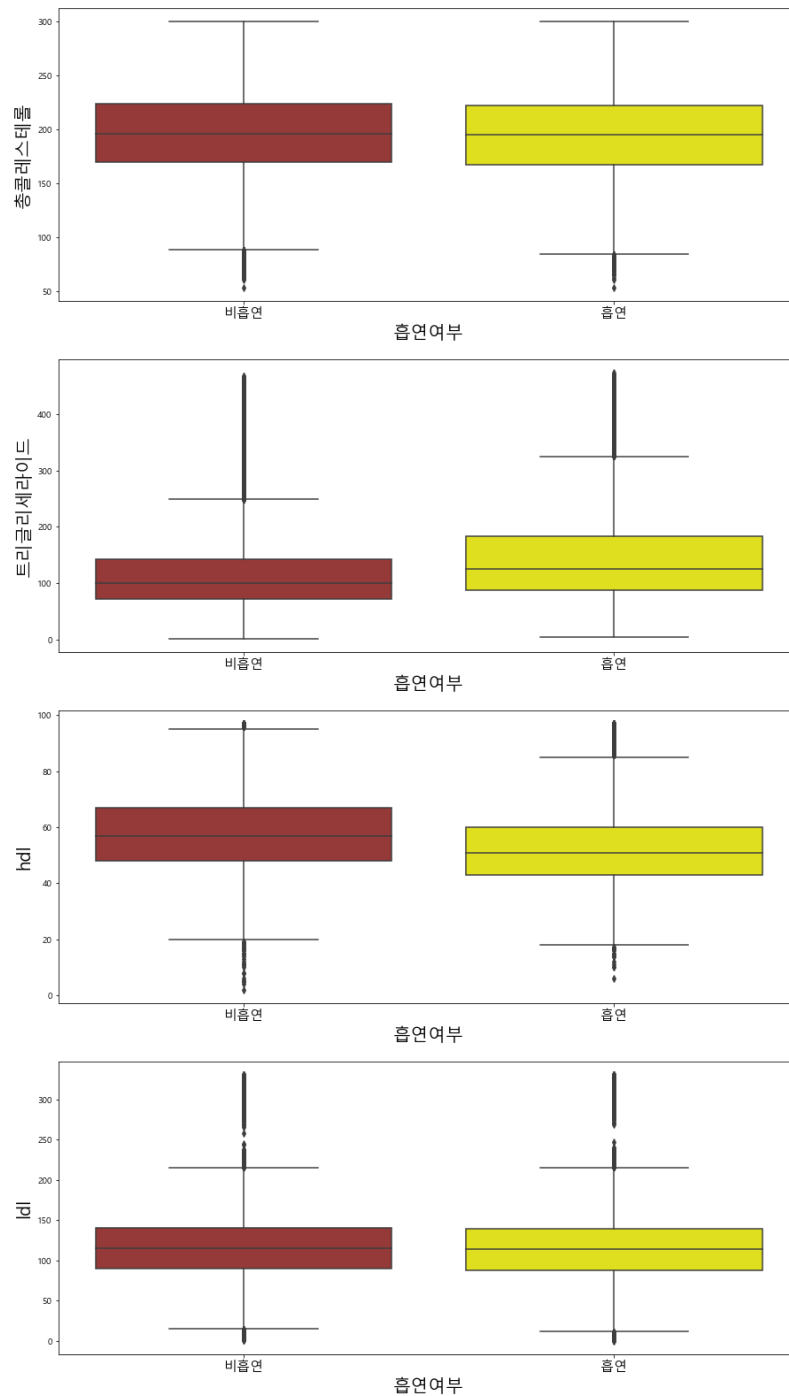
하지만, 트리글리세라이드가 400mg/dL 미만으로 건강검진데이터 기록 시에 사용하는 Friedwald equation을 이용하여 값을 추정했음에도 불구하고 계산된 LDL콜레스테롤 값이 음수값이었기에 이는 계산에 이용된 총콜레스테롤, 트리글리세라이드, HDL콜레스테롤 세 변수의 값 또한 신뢰할 수 없다고 판단하여 대체한 값에서 음수로 나온 18개 행 중 16개 행은 제거하였다. 나머지 2개 행도 음수값이기에 신뢰할 수 값이라 판단하여 제거하였다. 2013년 건강검진통계연보에 따르면, 국가건강검진에서 중성지방을 측정한 11,380,246명 중 중성지방 농도가 400mg/dL 이상인 사람은 236,436명으로 0.02%<sup>4</sup>에 불과하였다. 이에 따라 800mg/dL은 매우 특이한 경우라고 여기고 트리글리세라이드가 800mg/dL 이상인 행에 대한 LDL콜레스테롤 수치를 추정할 수 없으므로 제외하였다.

다음 콜레스테롤 관련 변수들의 boxplot과 요약된 통계량을 통해 총콜레스테롤, HDL콜레스테롤, LDL콜레스테롤 모두 이상치가 존재함을 확인하였으며, 이 수치들은 극히 일부의 수치라고 판단했기 때문에 99 percentile에 해당하는 수치를 기준으로 제거함으로써 이상치를 처리하였다. 총콜레스테롤은 301이상, hdl의 경우에는 98이상, ldl의 경우에는 331이상에 해당하는 수치로, 대한진단검사의학회 정하고 있는 총콜레스테롤, HDL콜레스테롤, LDL콜레스테롤의 “높음/아주높음” 수준의 기준보다도 훨씬 높은 수치에 해당된다.

---

<sup>3</sup> Sampson equation :  $LDL - C = TC/0.948 - HDL - C/0.971 - TG/8.56 + [TG * non - HDL - C]/2140 - TG/16100 - 9.44$

<sup>4</sup> 국민건강보험, 2014.



< 그림4. 흡연 집단에 따른 총콜레스테롤, 트리글리세라이드, hdl, ldl >

이상치 처리 후 흡연 집단에 따른 총콜레스테롤, 트리글리세라이드, HDL콜레스테롤, LDL콜레스테롤의 수치를 Box plot으로 봤을 때, 트리글리세라이드는 흡연집단에서 더 높고 HDL콜레스테롤은 비흡연집단에서 더 높았다. 하지만, 총콜레스테롤과 LDL콜레스테롤은 흡연자와 비흡연자 두 집단에서의 차이를 확인하기 어려웠으므로, 흡연 집단에 따른 총콜레스테롤의 등분산성 검정과 독립표본 T 검정을 진행하였다.

등분산성 검정(Bartlett test)	
H0: 흡연, 비흡연 두 집단의 총콜레스테롤 분산은 같다. vs. H1: 흡연, 비흡연 두 집단의 총콜레스테롤 분산은 다르다.	
statistic=115.7866	p-value=5.292850149405619e-27

유의수준 0.05하에서 귀무가설 H0을 기각할 근거가 충분하다. 즉, 두 집단에서 총콜레스테롤의 분산은 다르다.

이분산 독립 표본 T검정(T test)	
H0: 흡연, 비흡연 두 집단의 총콜레스테롤 평균은 같다. vs. H1: 흡연, 비흡연 두 집단의 총콜레스테롤 평균은 다르다.	
statistic=18.4605	p-value=4.7647064564166156e-76

유의수준 0.05하에서 귀무가설 H0을 기각할 근거가 충분하다. 그러므로 흡연, 비흡연 두 집단의 총콜레스테롤 평균은 다르다.

이분산 독립 표본 T검정(T test)	
H0: 흡연, 비흡연 두 집단의 총콜레스테롤 평균은 같다. vs. H1: 비흡연 집단의 총콜레스테롤 평균이 흡연 집단보다 크다.	
statistic=18.4605	p-value=2.3823532282083078e-76

유의수준 0.05하에서 귀무가설 H0을 기각할 근거가 충분하다. 즉, 비흡연 집단의 총콜레스테롤 평균이 흡연 집단의 총콜레스테롤보다 크다.

### 2.3.5. 공복혈당, 혈색소, 혈청크레아티닌

공복혈당이란 8시간 이상 공복 상태일 때 측정된 혈당 수치를 의미하며 126 mg/dL 이상일 경우 당뇨병으로 진단한다. 공복혈당의 요약통계량을 통해, 75 percentile 값은 109, 최대값은 901인 것을 확인했고 이상치 제거가 필요하다고 판단하여 95, 99 percentile 값을 각각 살펴본 결과 206의 값을 가지는 99 percentile 기준으로 행 제거를 했다. 공복혈당 수치가 70 mg/DL미만이면 저혈당, 70~99 mg/dL인 경우 정상, 100~125 mg/DL는 공복혈당장애, 125 mg/DL이상인 경우 당뇨병이라고 진단하므로 위 기준으로 범주화 시켜주었다.

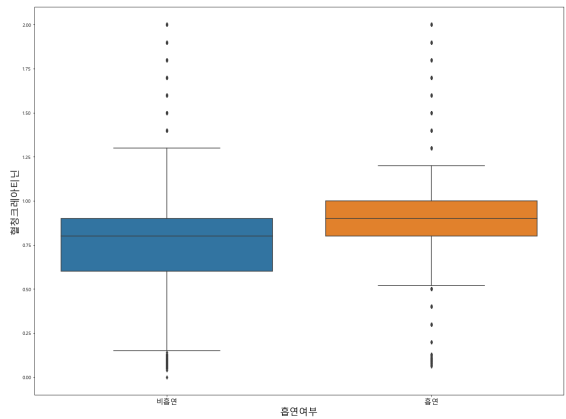
각 범주별 흡연자 비율을 살펴본 결과 <그림3>에서 볼 수 있듯이, 정상 범위의 혈당수치를 가진 집단의 흡연자 비율이 가장 낮았고 당뇨 집단의 흡연자 비율이 가장 높은 것을 확인했다.

혈색소는 성별에 따라 정상 범위가 다르다. 남성의 경우 13.5-17.5g/dL, 여성은 12.5-15.5g/dL일 때 정상이라고 판단한다. 해당 변수의 성별에 따른 분포를 Box plot으로 그려본 결과 정상 범위 내에 대부분이 분포한다는 것을 확인했다.

혈청크레아티닌은 신장 기능장애에 의해 증가하는 신장 관련 건강 수치로. 0.8에서 1.7 mg/dL일 때 정상이라고 판단한다. 통계청의 연령별 혈청크레아티닌 분포 현황을 참고했을 때 2.1이상은 극히 드물기 때문에 이보다 큰 값은 이상치라고 판단했다. 본

자료에서 혈청크레아티닌이 2.1이상인 행은 전체의 0.3%인 것을 확인하였고 해당 행은 삭제하였다. 흡연 여부에 따른 혈청 크레아티닌 수치를 비교해보았을 때 <그림5>와 같이 흡연자의 혈청 크레아티닌 수치가 비교적 더 높은 것을 볼 수 있다.

공복혈당, 혈색소, 혈청크레아티닌 모두 비흡연자에 비해 흡연자의 경우 더 높은 값을 가지는 경향이 있고, 이분산 t 검정 결과 흡연 집단이 비흡연 집단보다 더 큰 수치를 가진다는 것을 확인했다.



<그림5. 흡연 여부에 따른 혈청 크레아티닌 수치의 분포>

2.3.6. 간 : ast, alt, gtp

ast, alt, gtp 는 모두 간 기능을 나타내는 혈액검사 수치이다.

gtp의 경우 999의 값을 가지는 행이 66개 존재하였는데, 정상범위 밖에 있는 단일 값이 많이 존재한다는 것은 측정치가 없는 것으로 간주하고 결측치 처리를 해주었다.

ast와 alt이 정상 범위를 벗어났을 때의 위험수준은 아래의 표와 같다.

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 간손상
중증	400 IU/L 이상 정상 상한치의 10배 이상 증가	급성 바이러스성 간염, 허혈성 및 독성 간손상, 자가면역성 간염, 알코올에 의한 간 손상

따라서 본 자료에서 ast와 alt가 중증수치인 400을 넘는 행을 살펴보았을 때, 전체 자료의 1%미만 이었고 이와 같은 수치는 매우 드물게 나타나는 경우라고 판단하여 해당 행은 삭제하였다.

간수치 변수의 흡연 여부에 따른 분포를 그려보았을 때, 흡연 집단이 비흡연 집단보다 간수치가 높은 경향을 보였다. 이 세 변수들은 분포가 한쪽에 몰려있어 정규성을 따른다고 보기 어려우므로 비모수 검정인 맨휘트니 U Test를 통해 흡연여부에 따른 차이 여부를 확인해보고자

했다. 그 결과 세 변수 모두 유의수준 0.05 하에서 “해당 변수는 흡연 여부 집단에 따른 차이가 없다.”는 귀무가설을 기각할 근거가 충분하다는 것을 확인했다.

## 2.4. 결측치 처리

본 자료의 결측치는 크게 두가지 방법으로 대체하였다. 첫번째로 범주형 변수는 해당 변수의 최빈값으로 대체하였고, 두번째로 수치형 변수는 MICE(Multivariate Imputation by Chained Equation) 알고리즘을 활용하여 대체하였다.

범주형 변수 중에서 시력, 청력에서 많은 결측치가 있는 것을 확인했으나 이는 흡연 여부에 따른 차이가 없으므로 본 분석에 있어 유의미하지 않을 것이라 판단하여 해당 변수는 제거하였다. 요단백과 음주여부 변수에서 결측치가 존재하므로 해당 변수는 사이킷런의 simple imputer를 활용하여 최빈값으로 대체했다.

수치형 변수 중 허리둘레, 혈색소, 수축기, 이완기, alt, gtp에 결측치가 있는 것을 확인하였다. knn은 이상치에 민감하고, 본 자료의 크기에 비해 결측치의 수가 매우 적은 편이라고 판단하여 knn 대신 MICE 알고리즘을 사용하여 대체하였다.

## 2.5. 파생변수

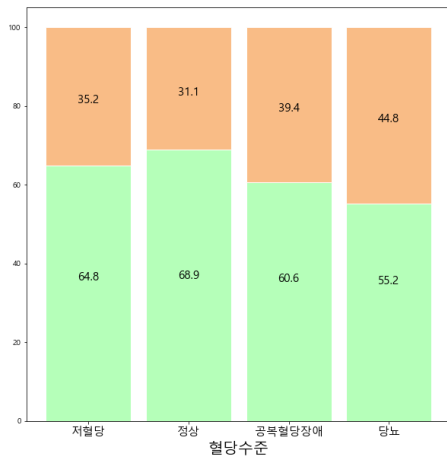
본 자료는 신장, 체중, 허리둘레의 세가지 신체정보를 가지고 있는데, 변수를 줄여 다중공선성 등의 문제를 해결하기 위해 신체정보가 내포된 bmi와 복부비만도 변수를 생성하였다.

체중을 신장의 제곱으로 나누어서 체질량 지수인 bmi 변수를 생성하였고, 허리둘레와 선형 관계를 보여준다는 것을 산점도 그림을 통해 확인했다.

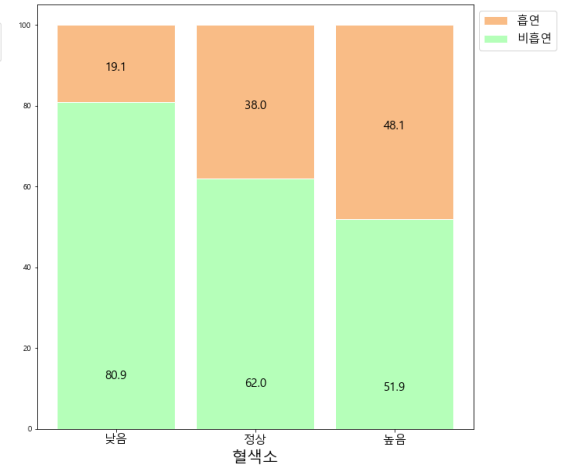
허리둘레를 기준으로 여성은 85이상, 남성은 90이상을 복부비만이라고 판단하므로 이 기준에 맞춰 복부비만 여부를 나타내는 변수를 생성했다. 복부비만 여부에 따른 bmi와 흡연여부 분포를 살펴본 결과 복부비만인 집단의 bmi가 더 높고, 흡연자 집단의 복부비만 비율이 더 높은 것을 확인했다.

혈색소는 정상수치보다 작을 경우 ‘낮음’, 클 경우를 ‘높음’, 그리고 정상 범위 내에 있으면 ‘정상’으로 범주화하여 파생변수를 생성했다. 성별에 따라 기준 값이 다르기 때문에 이에 맞게 구분해주었고, 성별에 따른 혈색소 수준의 차이는 없었다. 범주화 한 혈색소 변수와 흡연 변수 간의 관계를 살펴본 결과를 <그림7>와 같이 시각화 하였다. 혈색소가 정상 수치보다 높은 집단의 흡연자 비율이 정상 수치보다 낮은 집단의 흡연자 비율의 2.5배인 것으로 보아, 혈색소 수치가 높을 수록 흡연자의 비율이 더 높아지는 것을 확인했다.

공복혈당 수치가 70 mg/dL미만이면 저혈당, 70~99 mg/dL인 경우 정상, 100~125 mg/dL는 공복혈당장애, 125 mg/dL이상인 경우 당뇨병이라고 진단하므로 위 기준으로 범주화 시켜주었다. 각 범주별 흡연자 비율을 살펴본 결과 <그림6>에서 볼 수 있듯이, 정상 범위의 혈당수치를 가진 집단의 흡연자 비율이 가장 낮았고 당뇨 집단의 흡연자 비율이 가장 높은 것을 확인했다.



<그림6. 혈당수준에 따른 흡연 여부 비율>



<그림7. 혈색소 수준에 따른 흡연 여부>

### 3. 분석방법 및 결과

본 프로젝트에서는 흡연 여부를 예측하기 위하여 분류방법인 1) Random Forest, 2) XGBoost, 3) Logistic Regression 모델을 이용하였다. 또한, 이 모형의 예측률을 비교해보고 예측 모형에 포함된 중요 변수에 대해 살펴보았다. Random Forest와 XGBoost 모형은 베이지안 최적화(Bayesian Optimization)를 수행하여 모형을 개선하였으며, Logistic Regression 모형은 MinMaxScaler를 이용하여 정규화한 값을 이용하여 인사이트도출과 예측 분석에 활용하였다.

분석에 사용되는 자료는 376874행으로 분석하는 데 충분히 많은 양으로 판단하여 이를 train data : validation data : test data = 6 : 2 : 2로 나누어 사용하였다. train data와 validation data를 사용하여 모형을 학습시켰으며, test data로 데이터를 예측하였다.

사용된 변수 집단 1 : 신장, 체중, 허리둘레, 수축기, 이완기, 공복혈당, 총콜레스테롤, 트리글리세라이드, hdl, ldl, 혈색소, 혈청크레아티닌, ast, alt, gtp, bmi, 복부비만, 성별\_1.0, 성별\_2.0, 연령대\_40.0, 연령대\_45.0, 연령대\_50.0, 연령대\_55.0, 연령대\_60.0, 연령대\_65.0, 연령대\_70.0, 연령대\_75.0, 연령대\_80.0, 연령대\_85.0, 요단백\_1.0, 요단백\_2.0, 음주여부\_0.0, 음주여부\_1.0, 혈당수준\_1, 혈당수준\_2, 혈당수준\_3, 혈당수준\_4, 혈색소\_cat\_낮음, 혈색소\_cat\_정상, 혈색소\_cat\_높음, 흡연

→ 청력(좌), 청력(우), 시력(좌), 시력(우) 변수는 흡연 집단에 따른 차이가 없었으므로 흡연자와 비흡연자를 예측하는 데 중요한 변수가 아니므로 제외하였다.

사용된 변수 집단 2 : 수축기, 이완기, 공복혈당, 총콜레스테롤, 트리글리세라이드, hdl, 혈청크레아티닌, alt, gtp, bmi, 복부비만, 성별\_1.0, 성별\_1.0, 연령대\_40.0, 연령대\_45.0, 연령대\_50.0, 연령대\_55.0, 연령대\_60.0, 연령대\_65.0, 연령대\_70.0, 연령대\_75.0, 연령대\_80.0, 연령대\_85.0, 요단백\_1.0, 요단백\_2.0, 음주여부\_0.0, 음주여부\_1.0, 혈당수준\_1, 혈당수준\_2, 혈당수준\_3, 혈당수준\_4, 혈색소\_cat\_낮음, 혈색소\_cat\_정상, 혈색소\_cat\_높음, 흡연

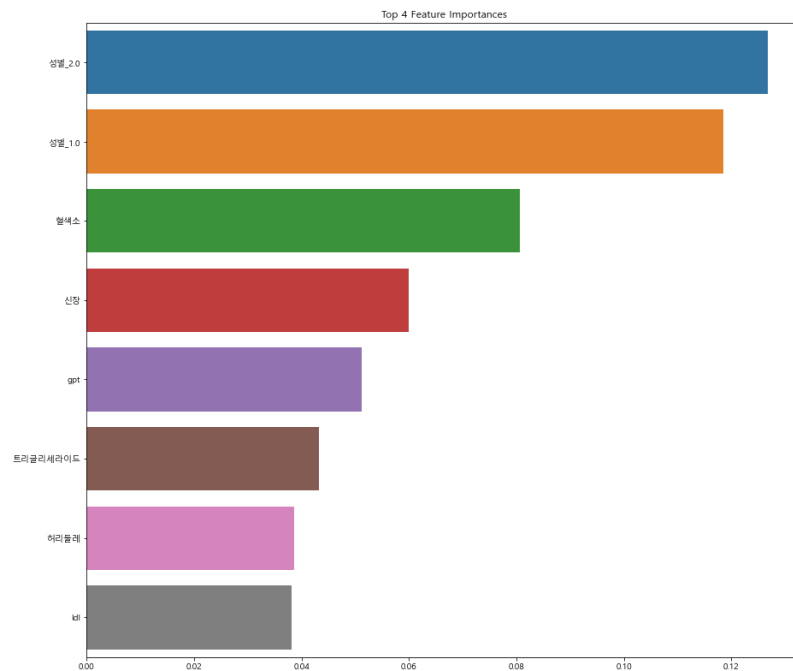
→ “사용된 변수들 1”에서 높은 상관관계를 갖고 있었던 변수들 중 일부를 제외하였다. (신장, 체중, 허리둘레)를 제외하고 파생변수들인 (bmi, 복부비만)을 남겨두었으며, ast와 alt는 상관관계가 높고 둘 다 간염의 정도를 보여주는 지표이므로 ast는 제외하였다. 또한, 총콜레스테롤과 ldl의 상관관계수가 매우 높고, ldl의 결측값을 총콜레스테롤, 트리글리세라이드, hdl로 추정하였기 때문에 오차나 변동성이 클 수 있기에 ldl을 제외하였다.

사용된 변수 집단 3 : 총콜레스테롤, 트리글리세라이드, hdl, 혈청크레아티닌, ast, alt, gtp, bmi, 복부비만, 성별\_1.0, 성별\_2.0, 연령대\_40.0, 연령대\_45.0, 연령대\_50.0, 연령대\_55.0, 연령대\_60.0, 연령대\_65.0, 연령대\_70.0, 연령대\_75.0, 연령대\_80.0, 연령대\_85.0, 요단백\_1.0, 요단백\_2.0, 음주여부\_0.0, 음주여부\_1.0, 혈당수준\_1, 혈당수준\_2, 혈당수준\_3, 혈당수준\_4, 혈색소\_cat\_낮음, 혈색소\_cat\_정상, 혈색소\_cat\_높음, 흡연

### 3.1. Bayesian Optimization을 활용한 RandomForest

의사결정나무(Decision Tree) 모형은 하나의 데이터에 의해 한 가지 의사결정나무 모형을 적합시키므로 데이터에 대한 설명력은 높지만 예측력이 높지 않은 문제가 있다. 데이터에 과적합되어 예측모형을 구축하는 데 사용된 데이터에 대해서는 높은 설명력을 보이지만, 구축에 사용되지 않은 새로운 데이터를 예측할 시 정확도가 확보되지 않을 수 있다. 이러한 이유로 의사결정나무 모형의 문제점을 개선한 랜덤포레스트 모형을 선택하였다.

튜닝 여부	사용된 변수	accuracy	recall	precision	f1
튜닝 전	1	0.818	0.842	0.699	0.764
튜닝 후	2	0.82	0.861	0.697	0.77



< 그림8. 베이지안 최적화 후 Random Forest 모델 변수 중요도 >

튜닝 전 모델의 변수 중요도를 살펴보았을 때, 상위 8개의 변수 중요도에서 성별, 혈색소, 신장, gpt, 트리글리세라이드, 허리둘레, hdl순으로 중요도가 높았다. 베이지안 최적화 알고리즘을 통해 Accuracy 기준으로 최적의 파라미터 조합을 찾아본 결과 최종적으로 아래의 파라미터를 사용하여 모델 튜닝을 진행했다.

파라미터	값
max_depth	18
min_samples_leaf	7



min_samples_split	11
n_estimators	258
n_jobs	-1

앞서 설명한 바와 같이 비슷한 의미를 갖거나 상관관계가 높은 변수를 제거한 집단2를 사용하여 해당 모델에 적합해 본 결과 성능이 향상 되었다. 튜닝 후의 모델에서는 성별, **gtp**, 트리글리세라이드, 총콜레스테롤, **hdl**, 혈청크레아티닌, **alt** 순으로 변수 중요도가 높았다. 혈당, 요단백, 혈당수준, 혈색소\_cat은 중요도가 다른 변수들에 비해 매우 낮았다.

### 3.2. Bayesian Optimization을 활용한 XGBoost

**XGBoost**는 병렬처리로 **GBM**보다 학습과 분류 속도가 빠르며, 과적합을 규제해주는 장점을 가지고 있어 내구성이 강하다. 특정 반복 횟수만큼 더 이상 비용함수가 감소하지 않으면 지정된 반복횟수를 다 완료하지 않아도 수행을 종료하는 조기중단과 같은 기능이 있고, 분류의 영역에서 뛰어난 예측 성능을 발휘한다는 특징을 가지고 있기에 **XGBoost** 모델을 선택하였다.

파라미터	값
n_estimators	100
max_depth	5
learning_rate	0.01
n_jobs	-1

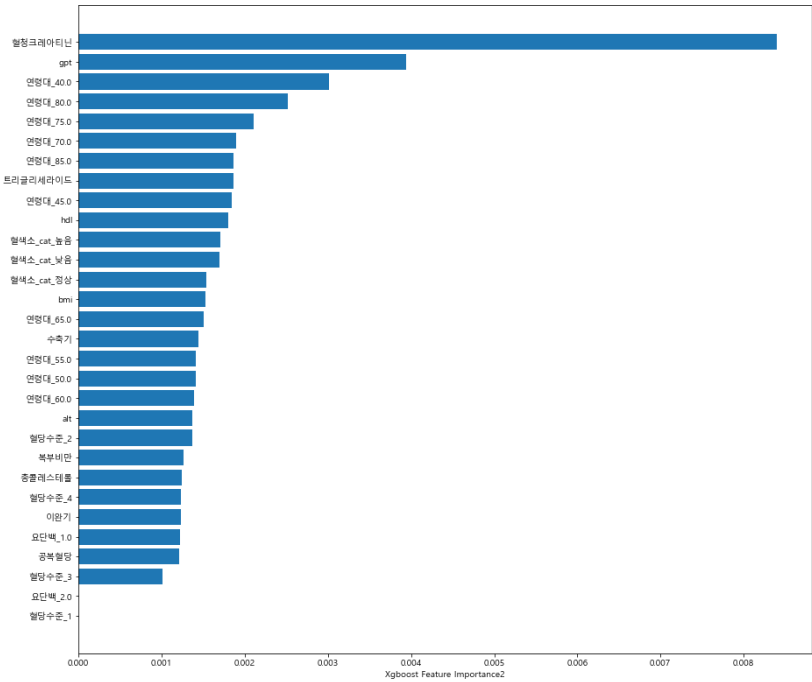
베이지안 최적화를 통한 파라미터 조정을 하기 이전에, 위의 파라미터와 모든 변수를 사용하여 해당 모델에 적합하였을 때, '성별\_1.0', 즉 남성 집단을 나타내는 변수의 중요도가 월등히 높았고, 이어서 음주여부, **gtp**, **hdl** 그리고 연령대 변수가 높은 중요도를 가진다는 결과를 얻었다.

파라미터	값
colsample_bytree	0.9
eta	0.1
gamma	9
max_depth	9
min_child_weight	12
subsample	0.68

이어서 베이지안 최적화를 통해 **Accuracy** 기준 최적의 파라미터 조합을 찾은 결과 위의 파라미터를 사용하였다. 앞서 언급한 변수 집단2를 사용하여 파라미터 조정을 거친 모델에 학습시키고 **validation** 셋으로 성능을 확인해본 결과는 아래와 같다. 성능에 있어서 큰 변화가

없었으므로 변수를 줄이고 파라미터를 조정하는 과정을 통해 모델의 복잡도를 줄이면서 훈련에 성공했다고 볼 수 있다.

튜닝 여부	사용된 변수	accuracy	recall	precision	f1
튜닝 전	1	0.818	0.875	0.69	0.772
튜닝 후	2	0.82	0.856	0.699	0.77



< 그림9. 베이지안 최적화 후 XGBoost 변수 중요도>

튜닝과 변수 선택의 과정을 거친 후의 변수 중요도는 성별, 음주여부가 월등히 높고 그 다음으로 높은 변수의 중요도는 <그림9>와 같다.

### 3.3. Logistic Regression

로지스틱 회귀모델은 여러 독립변수의 선형 결합을 통해 특정 범주에 속할 확률을 예측하는 확률모델로서 확률이 더 높은 범주로 분류하는 지도 학습 기법이다. 종속변수가 연속형 변수인 회귀모델과는 다르게 로지스틱 회귀모델은 범주형 변수가 종속변수가 될 수 있고, 이항 로지스틱 회귀의 경우 종속변수는 0 또는 1의 binary한 값을 가진다. 본 자료는 흡연 여부를 예측하는 이진분류가 목적이므로 로지스틱 회귀 모델이 적합하다고 판단하여 해당 모델을 선택했다.

Validation Dataset의 결과

사용된 변수	accuracy	recall	precision	f1	roc_auc
3	0.826	0.853	0.701	0.769	0.885

## Test Dataset의 결과

사용된 변수	accuracy	recall	precision	f1
3	0.819	0.859	0.696	0.769

Confusion Matrix	
39481	9671
3919	22680

## III. 결론

### 1. 주요 변수 분석과 인사이트 도출

모델 사용에 앞서, 여러 변수들과 흡연여부 간의 관계를 살펴보는 과정을 통해 크게 세가지의 결론을 얻을 수 있었다.

첫째, 성별은 흡연을 예측하는데 큰 영향을 준다. 본 자료에서 남성의 흡연자 비율은 **67.8%**인 것에 반해 여성의 흡연자 비율은 **5.3%**로 여성의 대부분이 비흡연자라고 볼 수 있다. 이와 같은 극단적인 비율의 차이로 인해 다른 건강수치 변수보다 성별을 통해 흡연 여부를 예측하기 더 쉬워진다. 하지만 여성의 흡연을 부정적인 시선으로 보는 사회적 분위기로 인해 여성의 흡연률은 실제보다 낮게 보고된다고 한다<sup>5</sup>. 따라서 특히 여성의 경우에는 다른 건강검진 결과 값을 이용하여 흡연여부 예측에 함께 활용할 필요가 있다.

둘째, 음주하는 집단의 흡연 비율이 높다. 음주를 하지 않는 집단의 흡연자 비율이 **17.8%**인 반면, 음주를 하는 집단의 흡연자 비율은 **49.4%**이다. 이를 통해 흡연과 음주를 모두 하는 사람이 많고, 해당 집단은 건강 관리에 소홀 하다고 볼 수 있다. 따라서 해당 집단은 건강에 문제가 생길 확률이 높으므로 본 분석의 결과를 통해 건강 관리의 필요성을 전달할 수 있다.

셋째, 흡연자의 경우 혈색소 수치가 높은 경우가 많다. 혈색소 수치를 정상 범위 기준으로 “낮음”, “정상”, “높음” 세가지로 범주화 했을 때, 차례대로 흡연자의 비율이 **19.1%**, **38%**, **48.1%**로 비율이 점차 커지는 경향이 있다. 실제로 흡연을 하면 흡연 중에 생기는 담배 연기의 일산화탄소가 헤모글로빈과 결합하게 되면서 적혈구가 많이 만들어 진다. 따라서 혈색소가 높은 적혈구 과다증의 주된 원인이 흡연<sup>6</sup>이라는 데에 타당한 근거가 된다. 본 분석을 통해 흡연자의 경우 혈전에 따라 발생할 수 있는 합병증의 위험성을 알려줌으로써 질병을 예방할 수 있다.

다음으로 모델링 과정을 통해 크게 두가지의 결론을 얻을 수 있었다.

먼저, **Random Forest** 모델을 통해 흡연 여부를 예측하는 데에 콜레스테롤, 간수치, 성별이 큰 영향을 준다고 해석할 수 있다. 콜레스테롤과 간수치는 음주, 식습관 등의 생활 습관에 영향을

<sup>5</sup> 국가지표체계 -국가발전 지표 - 건강 - 건강행태 - 현재흡연율 - 해설

<sup>6</sup> Jane Liesveld, MSD Manual, 적혈구 증가증

받는 건강 수치이므로 본 분석을 통해 흡연자로 예측 되는 집단은 생활 습관 개선을 통해 건강 관리에 유의해야 할 필요가 있다.

마지막으로, **XGBoost** 모델을 통해 성별, 음주여부, 혈청 크레아티닌, **gtp** 그리고 연령대 변수가 흡연 여부를 예측하는 데에 있어서 유의미하다고 볼 수 있다. 이 결과를 통해 중요도가 높게 나온 특정 연령대를 대상으로 금연의 필요성, 흡연의 부정적인 영향을 담은 캠페인을 더 적극적으로 홍보해 볼 수 있다. 추가적으로 본 분석을 통해 흡연자로 예측된 대상자에게는 추가적인 신장 기능 검사를 권유함으로써 더 세부적인 건강관리를 유도하고 질병을 예방할 수 있다.

## 2. 모형 활용 방안

본 분석에서는 **Random Forest**와 **XGBoost**의 두가지 모델을 사용하여 흡연 예측 성능을 비교해보았다. 그 결과 **XGBoost**를 최종 모델로 선정하였고, 학습시킨 모델을 통해 문진 결과 없이 건강검진 데이터를 통해 흡연 여부를 예측해 볼 수 있다. 모형을 확장시켜서 적용한다면 건강검진 결과에 따라 흡연자일 확률이 높은 집단에게는 추가적인 검진 또는 문진을 유도하여 해당 집단이 흡연자라서 발생할 수 있는 합병증이나 질병을 예방할 것을 기대한다.

## IV. 부록

### 1. 참고문헌

- 1) 대한의학회, 질병관리본부, “고혈압 가이드라인 일차 의료용 요약 정보”, 2018.01.
- 2) Youhyun Song et al. (2021) Comparison of the effectiveness of Martin's equation, Friedewald's equation, and a Novel equation in low-density lipoprotein cholesterol estimation
- 3) 국민건강보험공단 빅데이터 운영실, “국가중점 개방데이터(건강검진정보) 사용자 매뉴얼 (ver 4.0)”, 2017.
- 4) 국가법령정보센터, 일반건강검진 및 의료급여생애전환기검진 결과 판정기준, 2022.
- 5) 장성옥, “LDL-콜레스테롤의 추정:Friedwald 공식과 Martin 방법의 비교”, < 통계연구 제20권 제2호>, 2015.
- 6) 김양현, “소변검사 이상의 해석과 임상적 적용”, 고려대학교 의과대학 가정의학교실, 2012.
- 7) 원진희, “국민건강영양조사를 이용한 간기능검사수치에 영향을 주는 요인들에 대한 연구”, 2020.
- 8) 이현웅, “간기능 검사의 올바른 해석”, 대한내과학회 추계학술대회, 2017.

### 2. 참고자료 및 시각화자료

#### 2.1. 수치형 변수 관련 범주 기준

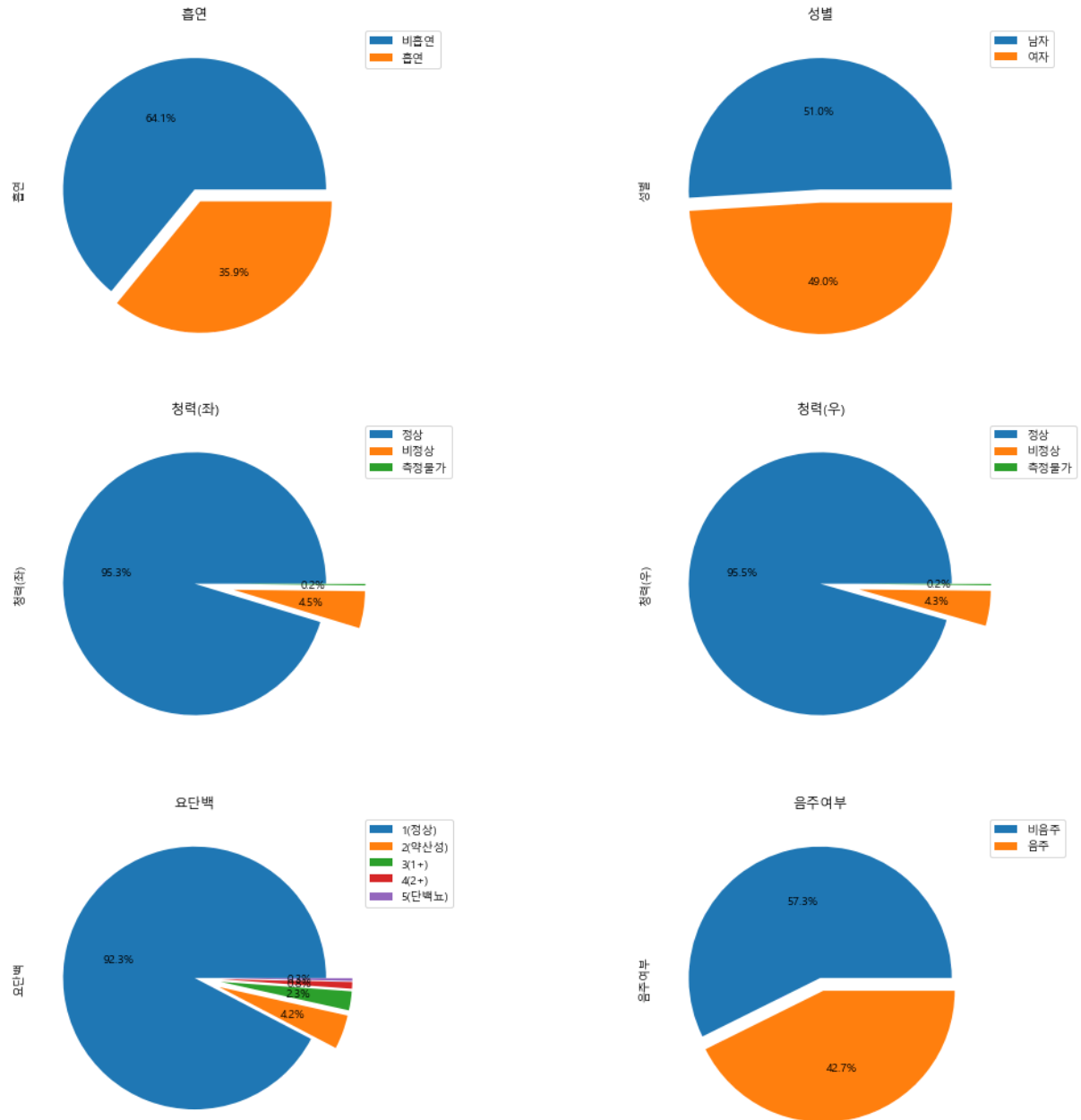
LDL-콜레스테롤		
	< 100	정상 (optimal)
	100 ~ 129	정상 이상 (near optimal/above optimal)
	130 ~ 159	약간 높음 (borderline high)
	160 ~ 189	높음 (high)
	190 이상	아주 높음 (very high)
총콜레스테롤		
	< 200	적절 (desirable)
	200 ~ 239	약간 높음 (borderline high)
	240 이상	높음 (high)
HDL-콜레스테롤		
	< 40	낮음 (low)
	60 이상	높음 (high)

< LDL콜레스테롤, 총콜레스테롤, HDL콜레스테롤 수치별 범주 기준 >

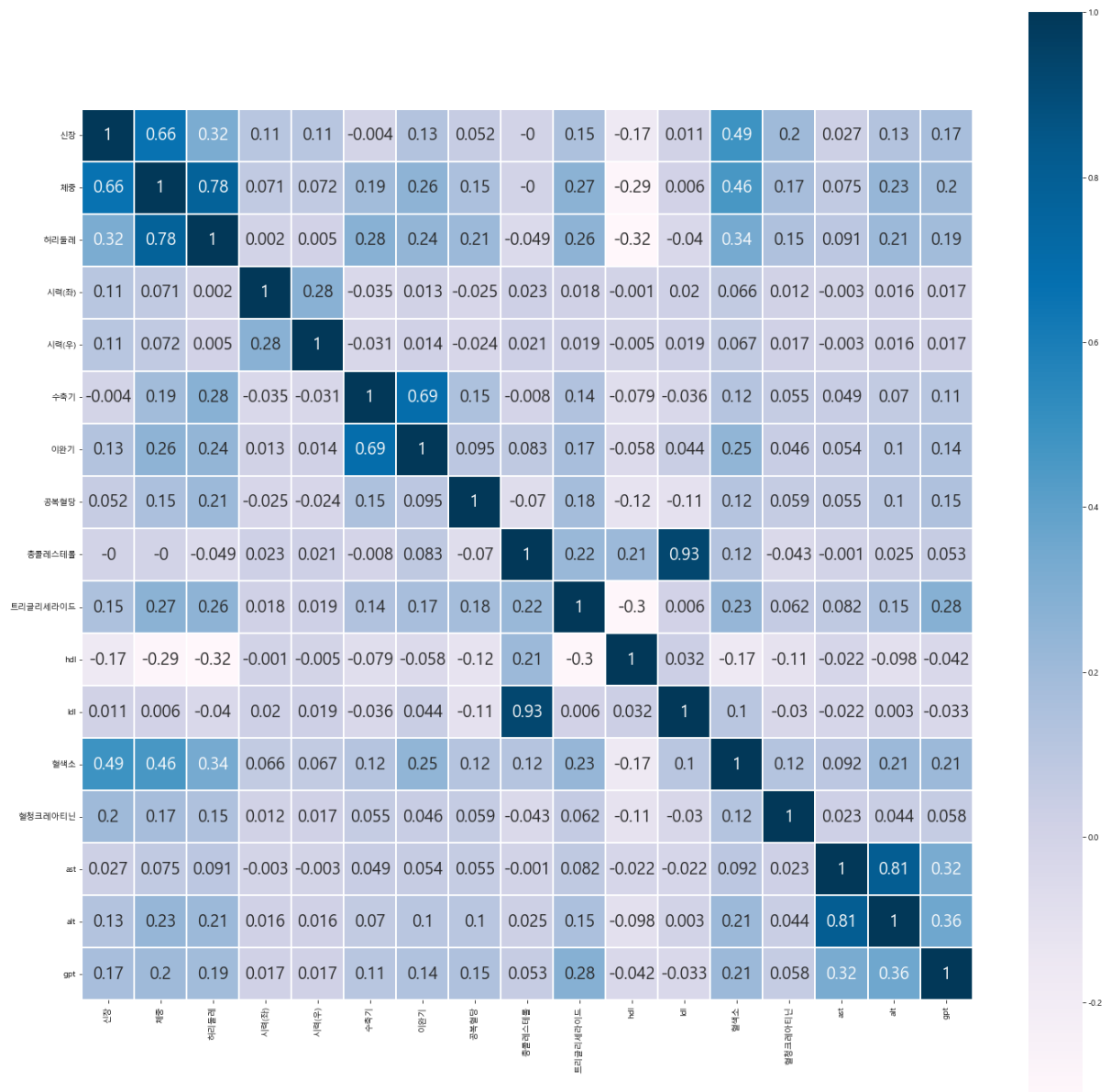
중성지방 수치	결과	치료
< 150mg/dL	정상	-
150 ~ 199mg/dL	약간 높음	체중감량과 운동요법
200 ~ 499mg/dL	높음	생활요법과 함께 약물치료 고려
> 500mg/dL	아주높음	즉시 약물치료

< 중성지방 수치별 범주 기준 >

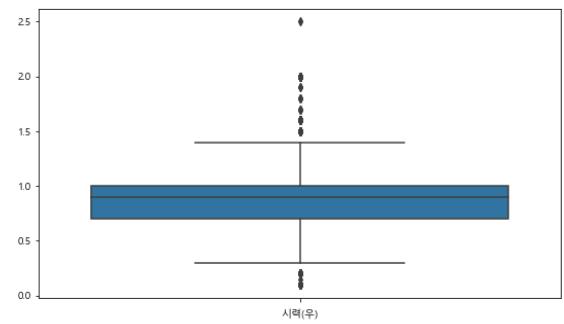
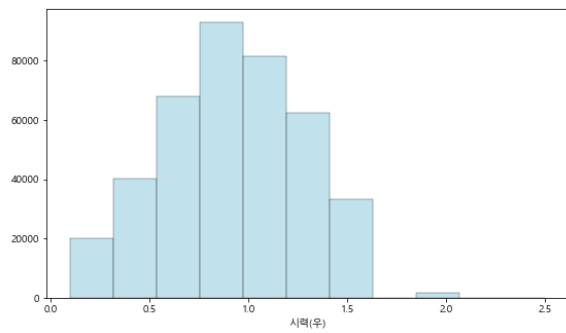
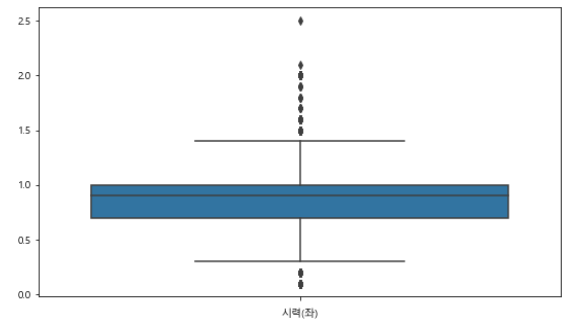
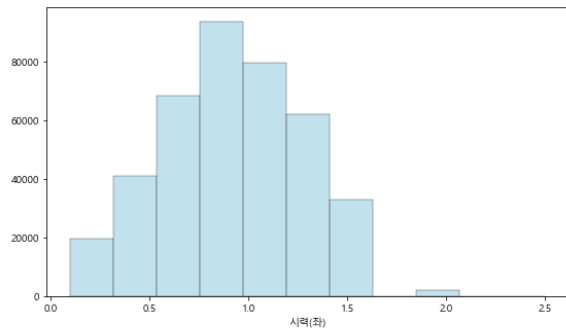
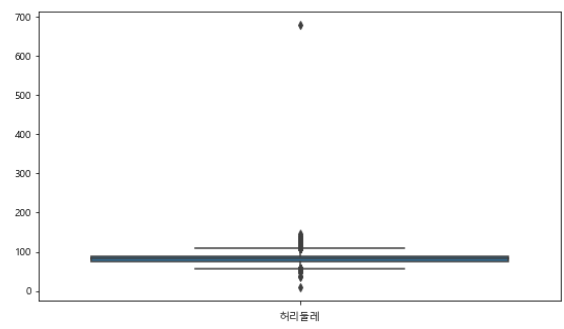
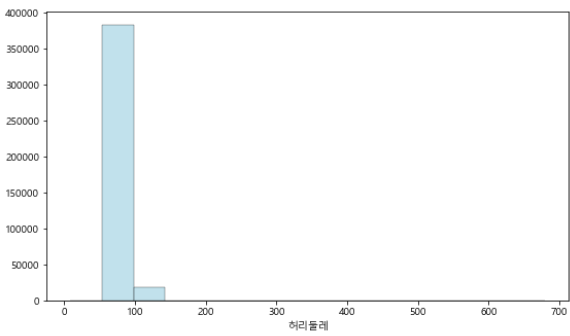
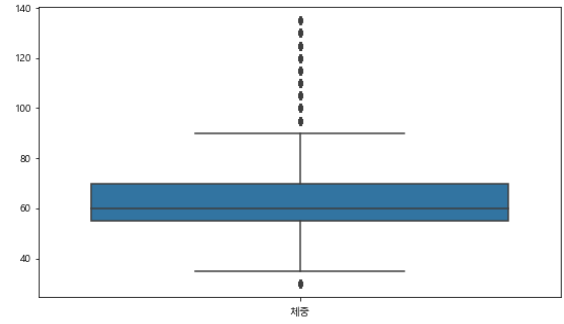
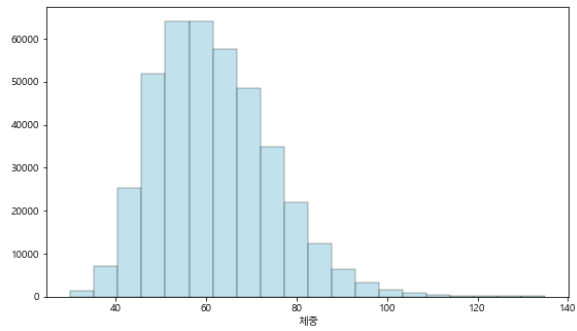
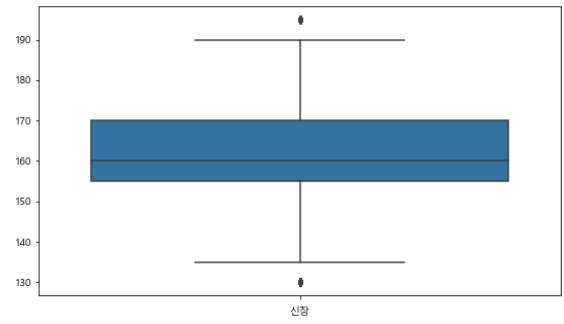
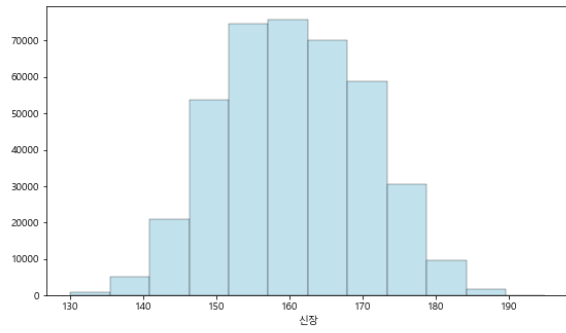
## 2.2. 시각화



< 범주형 변수 pie chart >

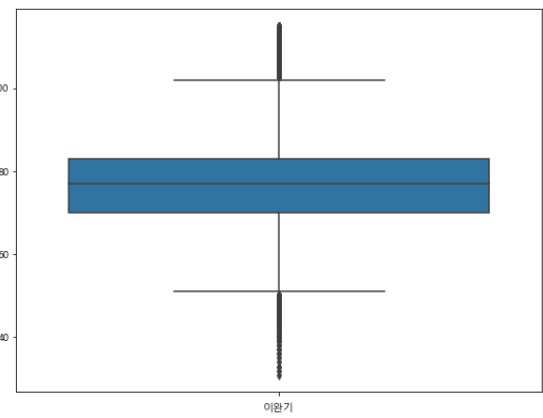
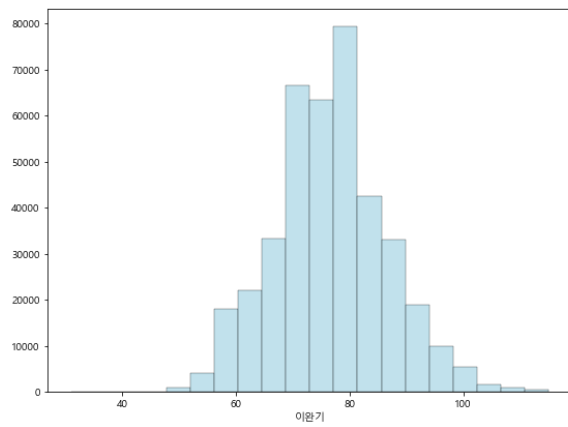
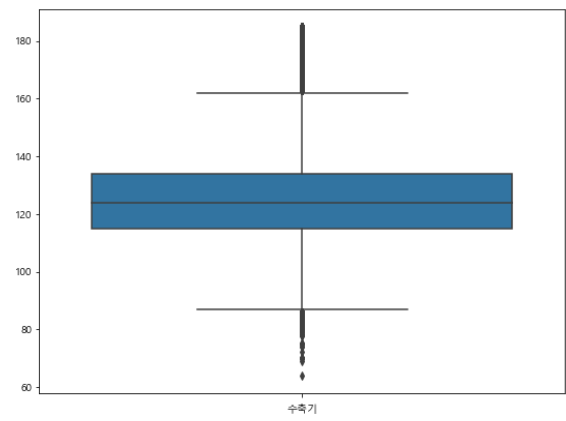
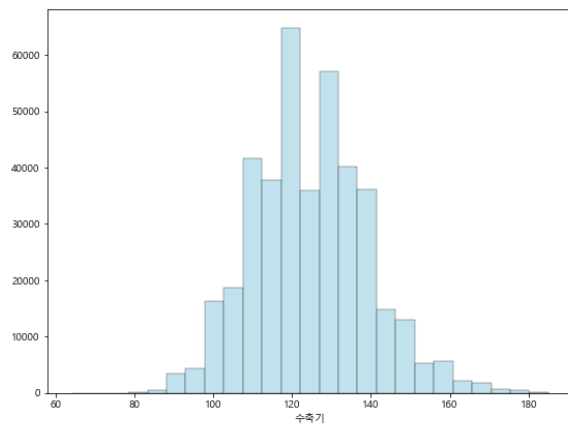


< 수치형 변수 상관관계 heatmap >

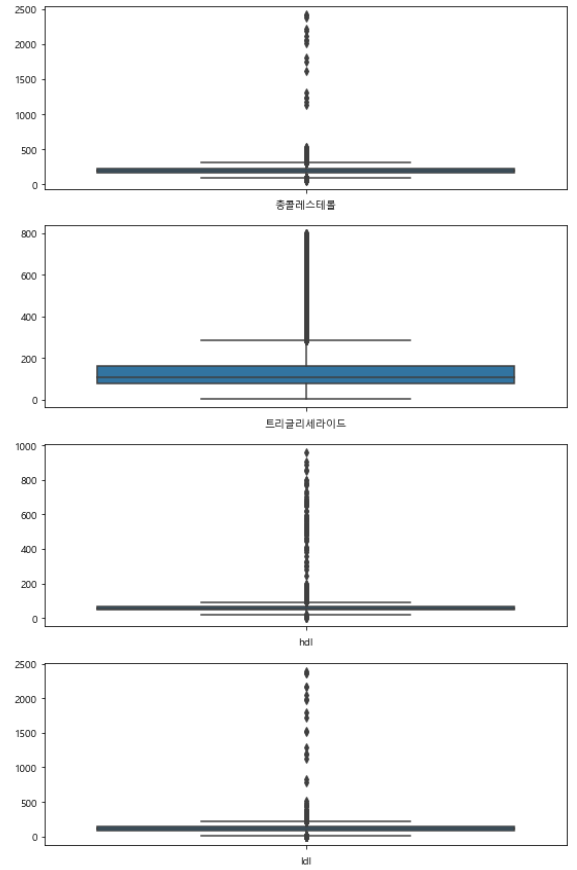
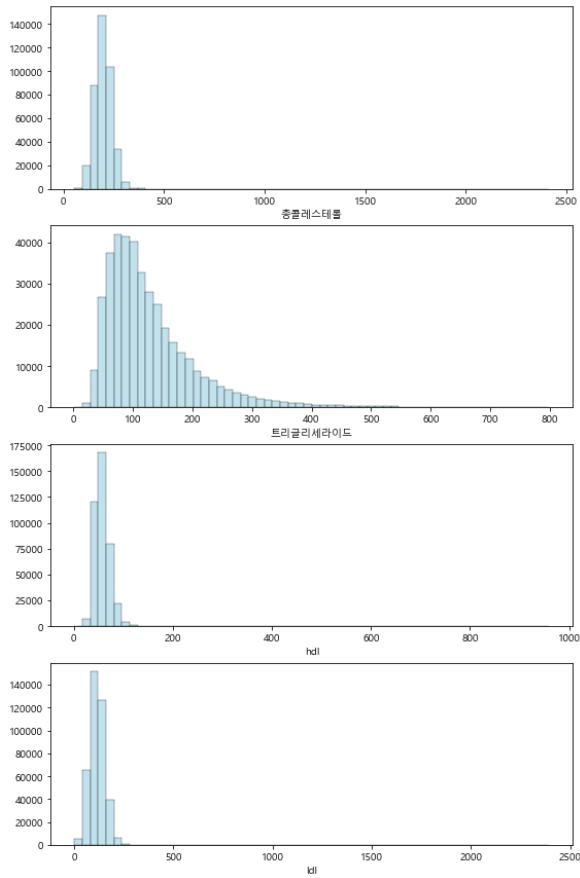


< 신장, 체중, 허리둘레, 시력(좌), 시력(우)의 막대그래프와 boxplot >

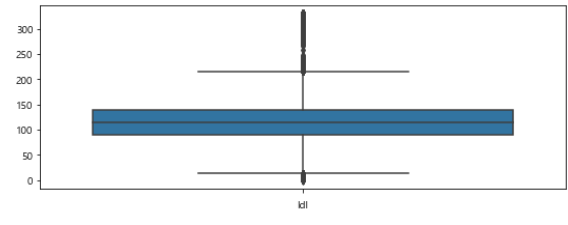
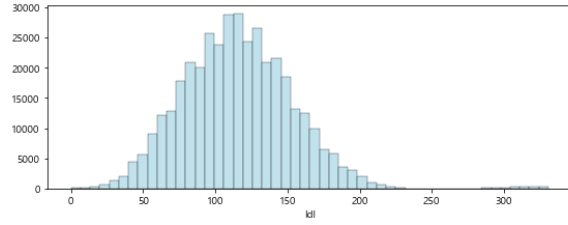
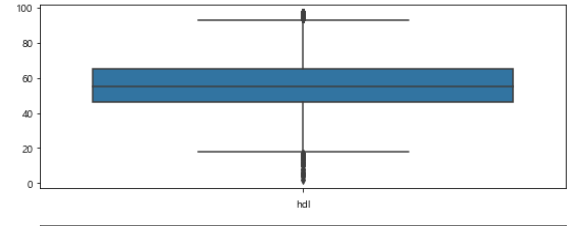
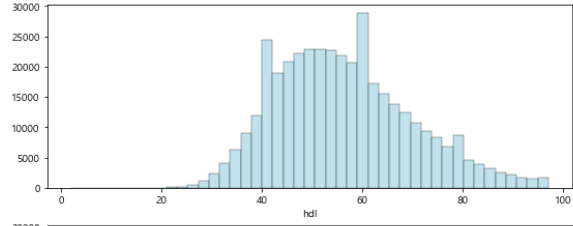
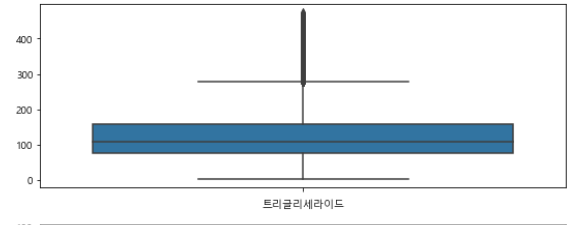
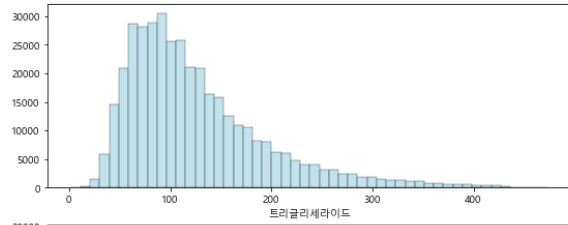
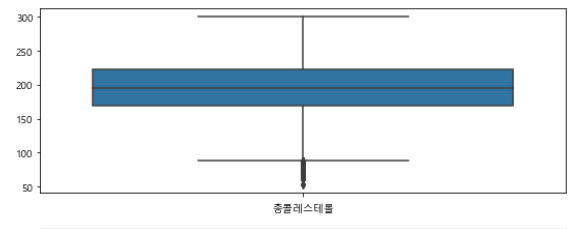
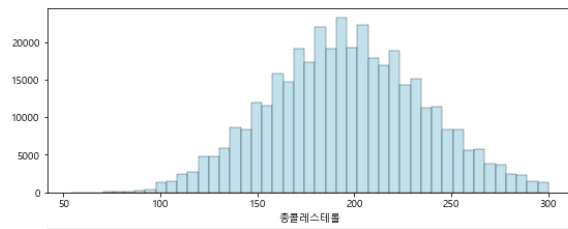




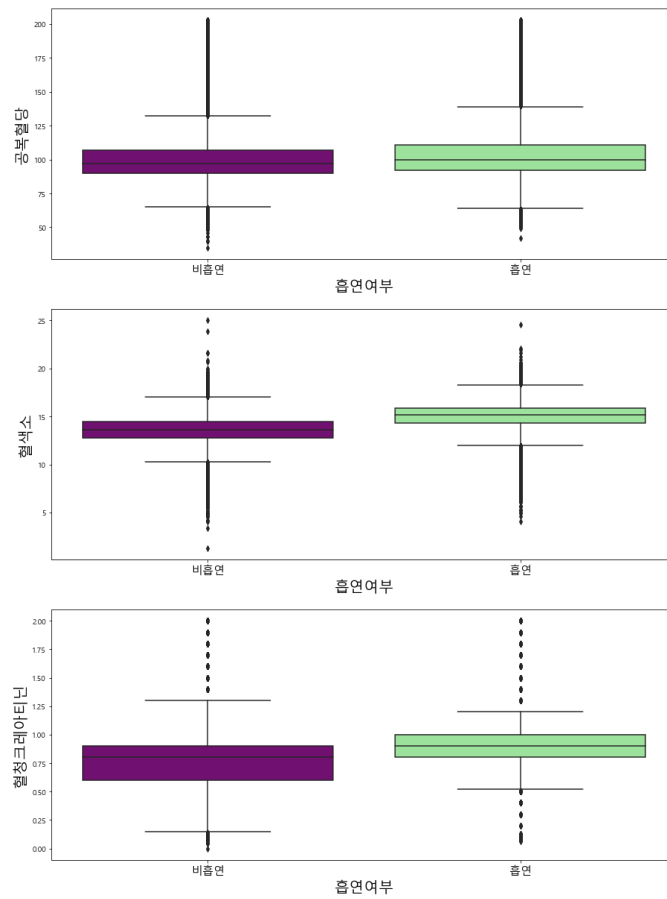
< 수축기혈압, 이완기혈압의 막대그래프와 boxplot >



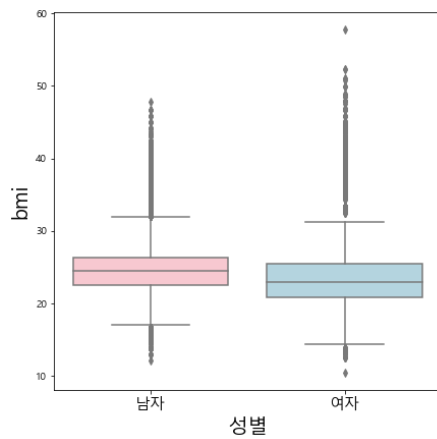
< 이상치 처리 전 총콜레스테롤, 트리글리세라이드, hdl, ldl 분포 >



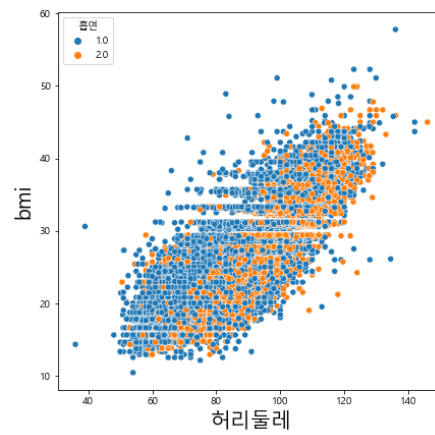
< 이상치 처리 후 총콜레스테롤, 트리글리세라이드, hdl, ldl 분포 >



< 흡연 집단에 따른 공복혈당, 혈색소, 혈청크레아티닌 boxplot >



< 성별에 따른 bmi의 boxplot >



< 허리둘레와 bmi의 scatter plot >