

데이터 마이닝 기법을 통한 남아프리카 CHD(관상동맥질환) 예측



수업명	데이터 마이닝			
교수명	김희경 교수님		조 번호	8조
학 과	수학과	통계학과	통계학과	통계학과
학 번	2016110418	2016110484	2018110514	2018110486
이 름	김상현	오병찬	윤수연	최솔
제출일	2021-06-04			

Contents

I. 서론

- 1. 연구 배경 및 목적 ----- 1p
- 2. 연구 방법 ----- 1p

II. 본론

- 1. 데이터 설명 -----2p
- 2. EDA ----- 5p
- 3. 데이터 전처리 ----- 10p
 - 3.1 변수 제거 및 결측값 처리 -----10p
 - 3.2 이상치 제거 -----10p
 - 3.3 파생변수 생성 ----- 13p
 - 3.4 다중공선성 ----- 14p
- 4. 모델링 ----- 15p
 - 4.1 데이터 준비 -----15p
 - 4.2 모델 후보 생성 -----16p
 - 4.3 모델 선정 기준 -----17p
 - 4.4 최종 모델 선정 -----17p
 - 4.5 변수 중요도 -----18p

III. 결론

- 1. 한계점 ----- 18p
- 2. 활용방안 -----19p

IV. 부록

I. 서론

1. 연구 배경 및 목적

1.1 데이터 생성 배경

1970년대 후반 남아프리카 공화국의 아프리칸스어 사용 백인 계층에서 이례적으로 높은 빈혈성 심장질환 발생률이 관찰되었다(Wyndham, 1982). 특히 이 지역 사회에서 남성 사망률은 여성의 약 2.5 배였다(Rossouw et al., 1983).

1979년, Rossouw et al. (1983)은 집중적인 우편 캠페인을 통해 케이프 주 남서부에 있는 3개의 아프리카너 지역 주민 중 백인 남성 3,357명, 백인 여성 3,831명 중 목표 인구 82명을 모집하였다. Hastie와 Tibshirani (1987)는 3,357명의 백인 남성 중에서 465명의 피험자를 선별하였다. 465명의 피험자는 관상 동맥 심장 질환을 앓고 있는 162명의 환자와 나머지 조사 대상자 집합에서 표본으로 추출된 303명의 대조군으로 구성되었다. 여기서 본 조의 데이터셋은 피험자 465명 중 462명에 대한 데이터로 이루어져 있다.

1.2 프로젝트 목적

고혈압 및 흡연 여부처럼 가역적이라고 간주할 수 있는 주요 위험 요인과 그 강도를 조사한다. 높은 심장질환 발생률을 보이는 지역사회에서 위험 요인에 주의를 기울이고, 그 발생률을 낮출 수 있는 해결방안을 제시하고자 한다.

2. 연구 방법

우선 반응변수인 'chd' 변수와 설명 변수들 간의 분포, 관계를 전체적으로 살펴본다. 그 후 정확한 분석 결과를 위해 이상치 제거와 파생 변수 생성을 비롯한 전처리 과정을 거친다. 다양한 데이터 마이닝 기법을 이용하여 어떤 요인이 영향을 주는지 파악한 후, 결과적으로 CHD 발병의 주요 위험 요인과 강도를 집중적으로 조사한다. 끝으로 발병률을 낮추는 여러 해결방안을 제시한다.

II. 본론

1. 데이터 설명

본 데이터는 160 명의 관상 동맥 심장질환 환자와 302 명의 대조군으로 총 462 개의 행을 가지며, 종속 변수 chd 와 9 개의 목적변수로 구성되어있다.

1) sbp

수축기 혈압 (systolic blood pressure)으로 심장이 수축할 때 혈관에 가해지는 압력을 말하는 수치형 변수이다. 프레이밍 코호트 연구로 SBP 가 CHD 중요한 위험 인자라는 사실이 알려졌으며, 일반적으로 혈압과 CHD 사이에 강한 연관성을 나타낸다고 한다. 최소 115 mm Hg 수축기 수준까지 연관성이 나타나며, 전반적으로 60 ~ 69 세의 경우 수축기 혈압이 10mmHg 낮아지면 CHD 발병 위험이 약 5분의 1로 낮아진다. ¹

2) tobacco

누적된 담배 양(kg)을 나타내는 수치형 변수이다. 미국 존스 홉킨스 대학에서 흡연율과 CHD 발생률 관계를 분석한 결과, 40 년 이상 담배를 피운 사람은 담배를 전혀 피우지 않은 사람에 비해 CHD 발병률 2 배로 나타났다. ²

3) ldl

저밀도 지단백 콜레스테롤의 양을 나타내는 수치형 변수이다. LDLC(low density lipoprotein cholesterol)는 지단백의 한 종류로 혈중 콜레스테롤을 운반한다. LDLC 은 혈관벽에 과도한 콜레스테롤 침착을 유발하여 동맥이 굳어지게 되고, 심질환을 발생시키기 때문에 나쁜 콜레스테롤로 분류된다. 동맥경화증 및 관상동맥질환과 가장 밀접한 관련이 있으며, LDLC 가 10% 증가하면 CHD risk 가 20% 증가한다고 알려져 있다. ³

¹ "Blood pressure and coronary heart disease: a review of the evidence," National Library of Medicine, n.d. 수정, 2021년 5월 14일 접속, <https://pubmed.ncbi.nlm.nih.gov/16222626/>.

² "연합뉴스," "흡연, 말초동맥질환 위험 ↑", 2019년 8월 6일 수정, 2021년 5월 14일 접속, <https://www.yna.co.kr/view/AKR20190806046900009>.

³ 병원약사학회(2011), 제28 권 제3 호 J. Kor. Soc. Health-Syst. Pharm., Vol. 28, No. 3, 296 ~ 304

4) adiposity

Body Adiposity Index(BAI) 수치를 뜻하며 BMI⁴의 한계를 보완하기 위해 제안된 지표이다. BAI는 키와 엉덩이 둘레를 사용하여 체지방 지수를 측정한다. 특히, BAI는 체중을 고려하지 않고 비만을 평가할 수 있으며 모든 인종의 성인에게 적용될 수 있다. 일부 연구에서는 BAI가 심혈관 및 대사 증후군 위험 인자와 관련이 있다고 발표했다.⁵

5) famhist

심장 질환의 가족력 (family history of heart disease)으로 Present와 Absent 두 개의 class를 갖는 범주형 변수이다. 가족력은 환자의 가족이나 가까운 친척의 의학적 내력을 나타내는 말로 CHD의 위험요인으로 알려져있다. 데이터셋에서 피험자 192명이 가족력이 존재하며 270명은 존재하지 않는다.

6) typea

A형 행동⁶을 측정하기 위해 고안된 시험 점수이다. A유형 성격은 성공에 대한 욕구와 경쟁심이 강하며, 강박관념, 난폭함, 착실하고 꼼꼼하면서도 성급한 특징을 가진다. 정신사회학적인 변수 중 A형 성격이 CHD의 중요 위험인자로 보고되었다. A형 성격을 가진 사람은 쉽게 불안증이 오고 이로 인해 심근경색을 유발하는 것이 주원인이다. 실제로 A유형 성격을 가진 사람은 정반대 행동 양상을 보이는 B유형 행동 양상의 소유자에 비해 CHD에 걸릴 위험률이 2배 이상 높다는 조사 결과가 나왔다.⁷

7) obesity

BMI 수치를 나타낸다. BMI는 체질량 지수이며 체중을 키의 제곱으로 나눈 값이다. 비만, 혹은 비대는 특징적인 면에서 질병으로 정의된다. CHD 환자의 80% 이상이 과체중이거나 비만으로

⁴ 체중에 정비례하는 BMI는 나이와 성별에 따라 차이를 구분하지 않고 근육량이 더 많은 운동 선수나 어린이에게 유용하지 않더라도 가장 널리 사용되는 지방 지수(Gallaher et al., 1996 ; Jackson et al., 2002). Bergman et al.

⁵ Schulze MB, Thorand B, Fritsche A, Häring HU, Schick F, Zierer A, Rathmann W, Kroger J, Peters A, Boeing H 및 Stefan N. 체지방량 지수, 체지방 함량 및 제 2형 당뇨병 발병률. 당뇨병 2012

⁶ 1970년대 미국의 심장내과 전문의인 프리드만과 로슨만이 정립한 개념.

⁷ 김수봉, 염근상, "A형 행동유형과 고지혈증의 연관성에 관한 환자-대조군 연구," Korean Journal of Health Promotion 9 no.2 (2009): 142-147. <https://www.dbpia.co.kr/Journal/articleDetail?nodeId=NODE08860763>.

나타났으며 특히 중앙 비만과 고혈압, II 형 당뇨병, 이상지질혈증 등 CHD 위험요인과의 높은 연관성을 갖는다.

8) alcohol

피실험자의 최근 알코올 소비량을 측정한 자료이다. 알코올 섭취와 CHD 발병률은 J/U 자 관계를 보여준다. 매일 적정 수준의 알코올 섭취는 심장질환의 위험성을 낮추는 효과가 있지만 일정 수준을 넘어가면 혈압을 증가시켜 위험성이 높아진다. 적정량은 성인남성 기준 소주 2 잔, 포도주 2 잔, 맥주 2 캔에 해당한다.

9) age

연령이 증가할수록 혈관이 딱딱해지고 탄성을 잃어가며 확장 능력이 감소하는 등 혈관의 변화가 발생하게 되고 이로 인한 혈관질환의 위험도가 높아진다. 또한 연령이 증가할수록 고혈압, 당뇨병, 고지혈증 등의 유병율이 증가하게 되어 복합적인 작용으로 CHD 의 발생 위험을 증가시킨다.

10) chd

관상동맥이란 심장의 표면에 위치하는 혈관을 말하는데 관상동맥은 심장이 움직일 때 심장근육에 혈액을 공급한다. 관상동맥질환은 관상동맥이 좁아져 심장근육 일부에 혈액 공급이 부족해질 때 발생한다. 심장에 혈액을 공급하는 관상동맥의 동맥경화증이 주된 원인이 된다. 원인 인자로는 연령, 흡연, 당뇨, 고지혈증, 고혈압, 가족력 등이 있으며 관상동맥질환이 발생하면 흉통 증상이 전형적으로 발생하며 심장 기능의 저하로 호흡 곤란, 심근경색, 심정지가 일어날 수 있다.

변수에 대한 도메인 지식을 알아본 결과, 모두 목적변수인 'chd'와 연관이 있음을 확인하였고 모든 변수를 분석에 사용하고자 한다.

2. EDA

종속 변수인 chd 의 0(음성)과 1(양성)은 데이터 상에서 음성 관측치는 302 행(65.4%), 양성 관측치는 160 행(34.6%)으로 약 2 배의 차이를 보인다. 첫 번째 독립변수인 sbp 변수를 살펴보면, CHD 음성과 양성 분포에서 모두 정규분포와 유사한 형태를 보인다. 양성인 경우 높은 수축기 혈압을 가지는 빈도가 음성보다 약간 많은 것을 확인할 수 있다. [부록 1] 2017 년 미국심장학회·심장협회(ACC·AHA) 고혈압 가이드라인에서는 정상혈압을 120/80mmHg 미만으로 설정하였고, 수축기 혈압을 기준으로 120~129mmHg 인 경우는 고혈압전단계, 130~139mmHg 를 1 기 고혈압, 140mmHg 이상은 2 기 고혈압으로 정의하였다. 가이드라인을 바탕으로 sbp 변수를 고혈압 범주별로 시각화 하였다.

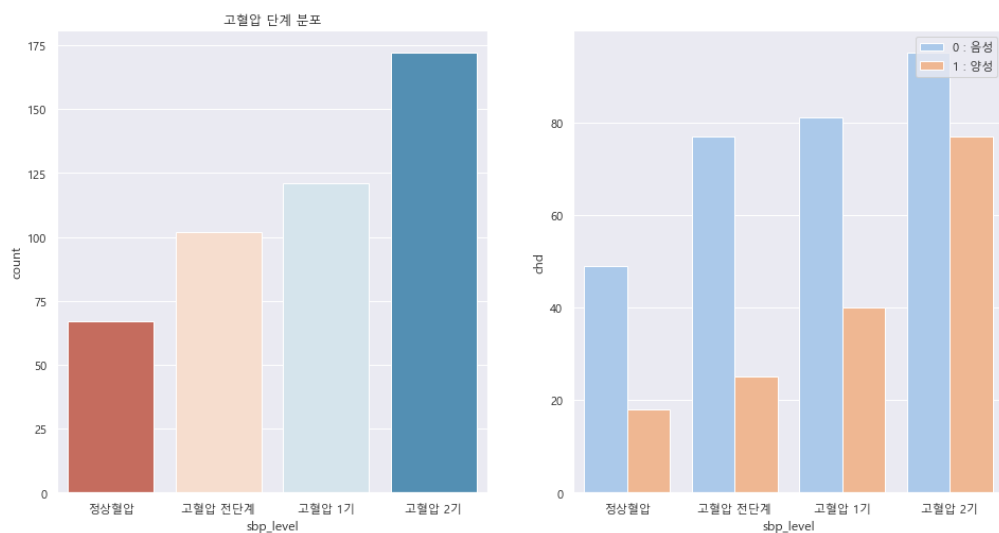


그림 1) 고혈압 단계별 분포

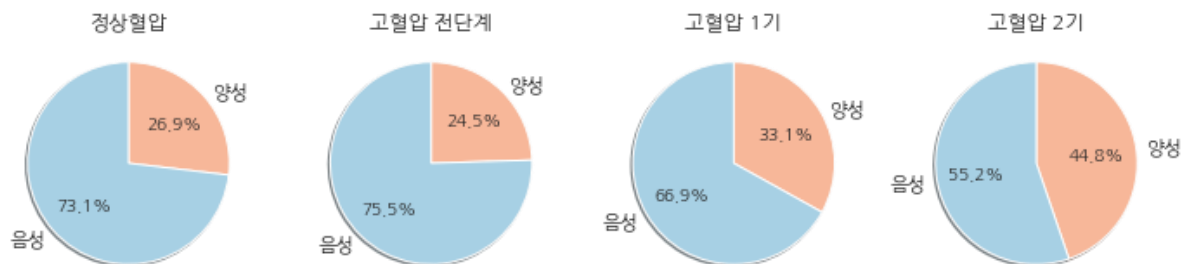


그림 2) 고혈압 단계에 따른 chd 발병률

정상혈압과, 고혈압 전단계, 고혈압 1기에서 CHD 양성이 전체에서 각각 26.9%, 24.5%, 33.1%의 값으로 존재하였고, 고혈압 2기에서는 44.8%를 차지하며 다소 높은 비율을 보인다.

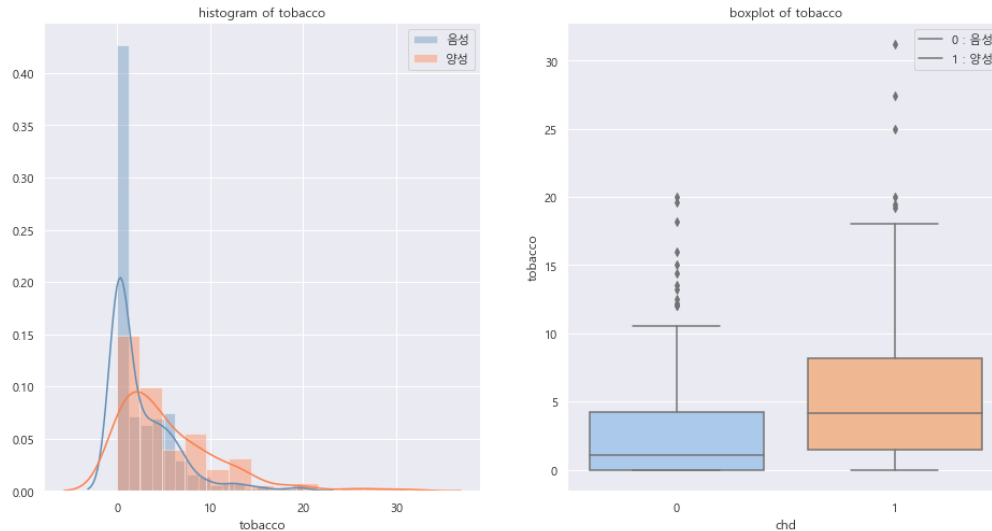


그림 3) chd에 따른 tobacco 분포

다음으로 tobacco 변수는 오른쪽 꼬리를 가진 비대칭 분포를 가진다. CHD 양성인 음성보다 누적 담배양이 큰 경향을 가진다. 누적 담배양이 0kg 일 때 비흡연자, 이를 초과하면 흡연자라고 했을 때 흡연자는 비흡연자보다 CHD 양성 비율이 약 3 배로 나타났다.

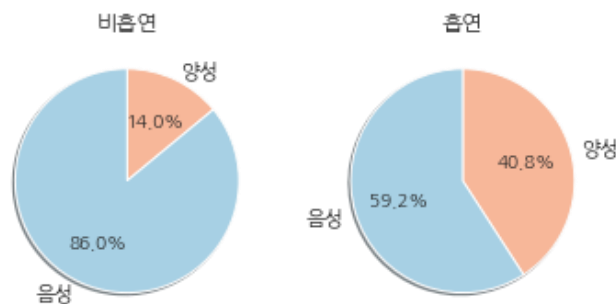


그림 4) 흡연에 여부에 따른 chd 발병률

미국 존스 홉킨스 대학에서 흡연율과 CHD 발생률 관계를 분석한 결과, 40 년 이상 담배를 피운 사람은 담배를 전혀 피우지 않은 사람에 비해 CHD 발병률 2 배로 나타났다고 한다. 이는 본 조의 데이터에서 도출된 결과와 유사하다고 볼 수 있다.

ldl 변수는 양성, 음성 모두 정규분포와 유사한 분포를 가지며, 양성인 음성보다 저밀도 지단백 콜레스테롤의 양이 많다. [부록 2] Mayo Clinic 과 US National Library of Medicine 에 따르면 미국 성인 기준 최적 LDL 수치는 100mg/dL 미만이다. 100~129mg/dL 에서는 양호, 130~159mg/dL 는 경계, 160~189mg/dL 높음, 190mg/dL 이상이면 매우 높음으로 범위를 지정하였다.

본 조의 LDL 콜레스테롤의 단위는 mmol/L 이기 때문에 위의 기준을 단위를 변환하여 범주화를 진행하였다. LDL 콜레스테롤 수치가 증가할수록 CHD 양성 비율이 따라서 증가함을 볼 수 있다.

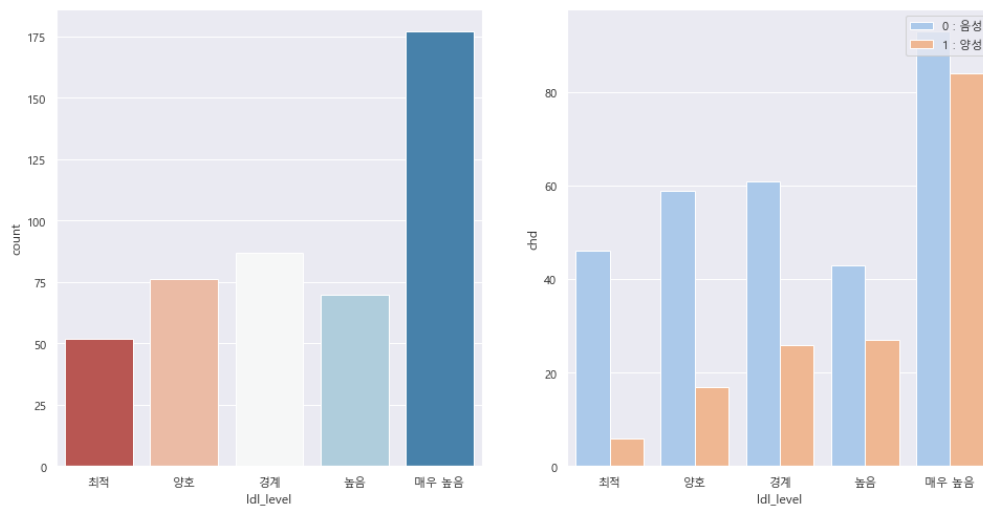


그림 5) LDL 콜레스테롤 단계별 분포

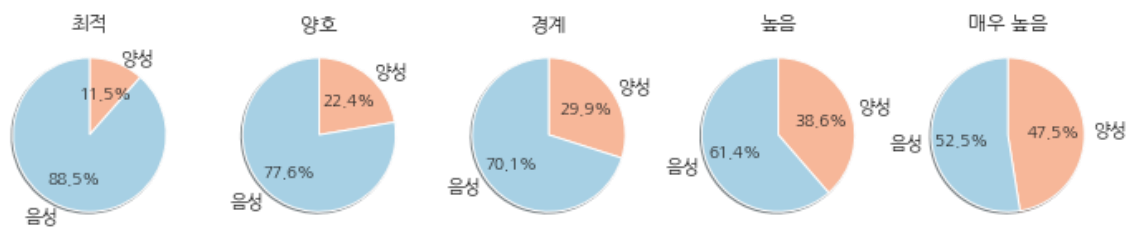


그림 6) LDL 콜레스테롤 단계에 따른 chd 발병률

adiposity 변수의 histogram에서는 양성 그룹이 우측에 위치한 것을 확인할 수 있다. boxplot 또한 CHD 양성 그룹의 adiposity가 더 높은 수치를 갖는다. [부록 3] 따라서 adiposity 변수를 저체중, 정상, 과체중, 비만으로 범주화해 음성과 양성 그룹을 살펴보았다. 비만 그룹에서 가장 많은 CHD 양성이 발생했고 체중이 올라갈수록 양성 그룹의 비율이 커지는 것을 확인할 수 있었다. [부록 4]

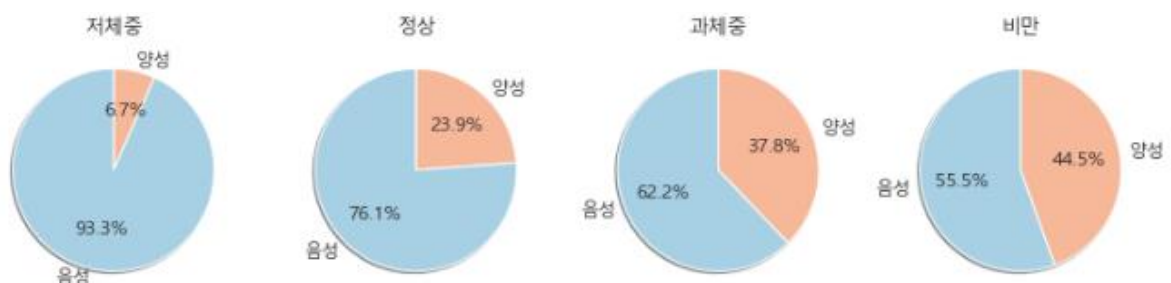


그림 7) adiposity 그룹별 chd 발병률

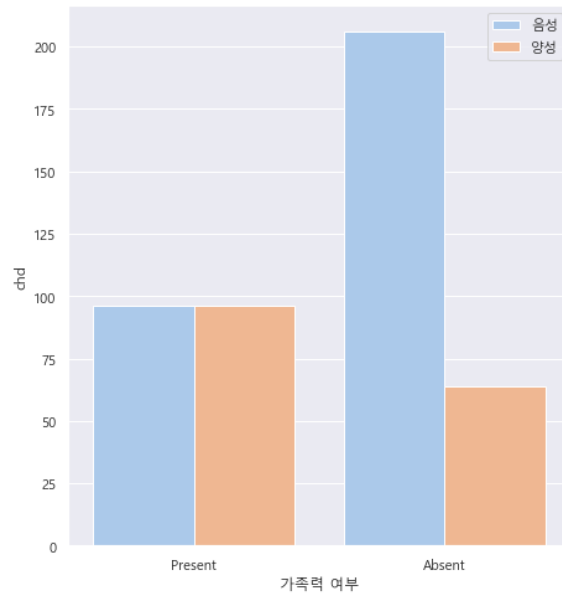


그림 8) 가족력에 따른 chd 발병률

심장 질환의 가족력이 있는 경우 CHD 양성 비율이 50%로 가족력이 없을 때의 비율인 23.7%의 2 배로 나타났다. 이 차이는 famhist 변수가 유의함을 보여준다.

obesity 변수의 분포를 살펴보면 환자와 정상 그룹은 유사한 형태의 보여주고 있다. histogram 과 boxplot 에서는 환자 그룹의 분포가 정상 그룹의 분포보다 오른쪽에 위치하고 평균이 더 높지만 유의한 차이로 고려할 수 없었다. [부록 6]



그림 9) obesity에 따른 chd 발병률

따라서 세계보건기구의 비만기준⁸에 따라 obesity 변수를 저체중, 정상, 과체중, 비만으로 범주화해 음성과 양성 그룹을 살펴보았다. 비만 그룹에서 가장 많은 CHD 환자가 발생했고 체중이 올라갈수록 환자 그룹의 비율이 커지는 것을 확인할 수 있었다. [부록 6] [부록 7]

⁸ 김경애, "비만 기준 'BMI 25' 올리자는 국회에 전문가 난색 "근거 있어야" ", 메디포뉴스, 2018-10-15

alcohol 변수에서는 대부분 알코올 소비량이 대부분 0에 가까운 right-skewed의 모습을 보여주고 있다. [부록 8] 알코올 소비량이 0인 사람을 알코올 미섭취자 그룹으로, 0보다 큰 사람을 섭취자 그룹으로 나누어 발병률을 비교해보았을 때 큰 비율의 차이는 없었다. 알코올 소비량의 3분위수를 기준으로 알코올 과섭취자 그룹과 적정섭취자 그룹으로 나누어 비율을 살펴봐도 유의미한 차이를 보이지 않는다. [부록 9]

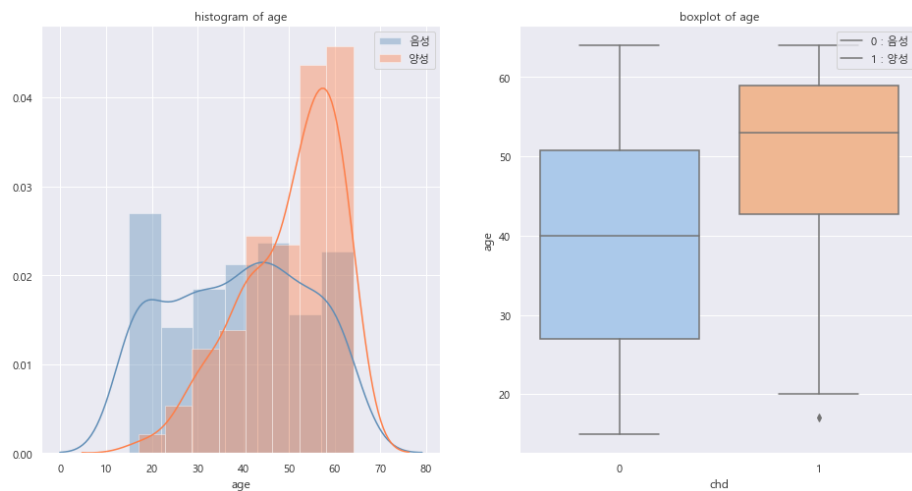


그림 10) chd에 따른 age 분포

age 변수에서 음성인 경우 나이대가 고르게 분포되어 있지만 양성인 경우 전반적으로 높은 나이대에 분포되어있다. 나이대별 변수를 생성한 후 음성과 양성 빈도를 살펴보면 연령대가 높아질수록 빈도수가 커지며 심지어 50대와 60대에서는 CHD 환자가 정상인 사람보다 더 많아지거나 같아지는 것을 볼 수 있다. 환자와 정상인의 비율 또한 고연령대로 갈수록 환자의 비율이 급격히 커진다.

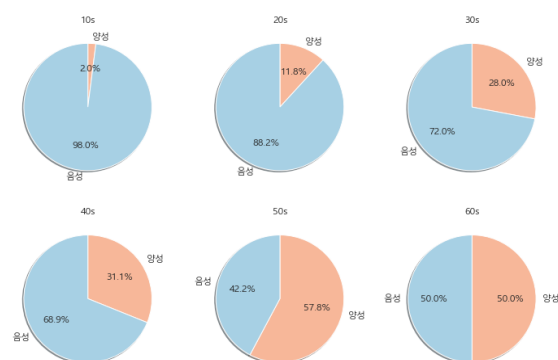


그림 11) 나이대별 chd 발병률

3. 데이터 전처리

3.1 변수 제거 및 결측값 처리

행의 순서를 나타내는 row.name 변수를 제거한다. 데이터 상에서 어떠한 결측치도 발견되지 않았다.

3.2 이상치 제거

1. tobacco 변수와 alcohol 변수

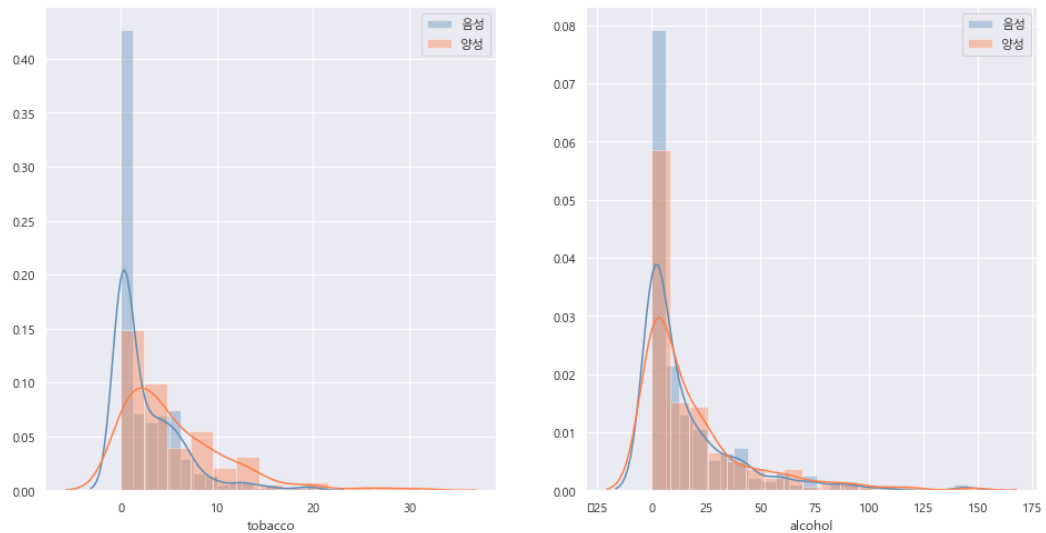


그림 12) tobacco, alcohol 변수 분포

흡연량과 음주량을 나타내는 tobacco와 alcohol 변수 다른 변수에 비해 왜도값이 양의 방향(Right-skewed)으로 큰 것을 알 수 있다. 각 변수의 값이 0인 행이 많으며 이는 흡연과 음주를 하지 않는 피실험자가 많음을 의미한다. alcohol 변수의 중앙값은 7.51이지만 평균값은 17.04로 이상치의 영향을 많이 받는 대푯값인 평균값이 중앙값과 많이 차이가 난다. 이는 tobacco 변수에서도 중앙값과 3사분위수가 각각 2.0과 5.5이지만 최대값이 31.2로 이상치가 존재함을 뜻한다.

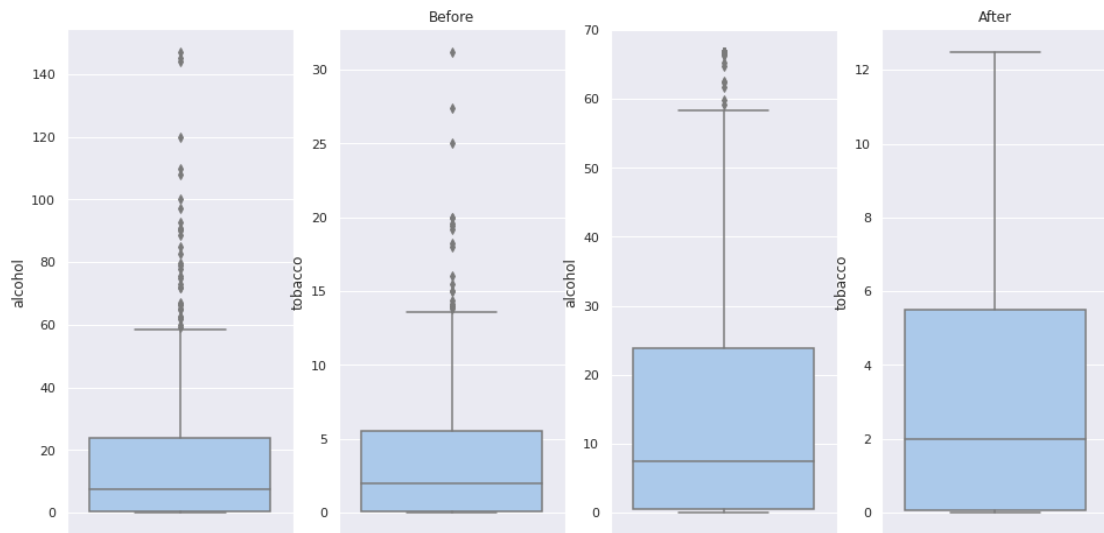


그림 13) 95 percentile 대체 전 / 후

이에 따라 각 변수의 값이 각 변수의 95 percentile 이상의 값을 가진다면 이는 95 percentile 값으로 대체하는 처리를 하였다. 이를 통해 극단치들의 값을 줄여 영향을 줄이는 효과를 낼 수 있다.

2. adiposity 변수와 obesity 변수

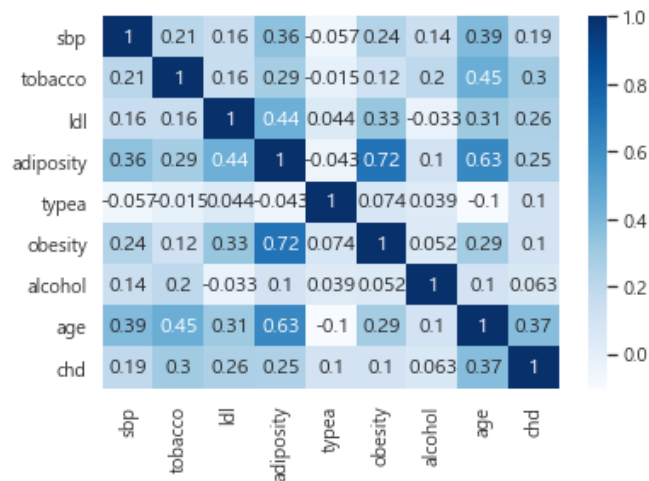


그림 14) 변수 상관계수 hitmap

변수 간 상관관계를 확인했을 때 adiposity와 obesity, adiposity와 age가 각각 0.73, 0.63으로 0.5가 넘는 상관관계를 보였다.

그 중 adiposity와 obesity의 joint plot에서 두 변수 중 한 변수의 값만 크고 나머지 변수의 값은 작은 관측치를 발견했다. 따라서 두 변수의 비율을 나타내는 변수를 생성하였다. $\text{weight_rate1} = \text{adiposity} / \text{obesity}$ 이며, $\text{weight_rate2} = \text{obesity} / \text{adiposity}$ 이다. 두 변수의 분포를 boxplot에서 확인했을 때, 각각 2와 4보다 큰 값을 이상치로 판단하고 제거하였다.

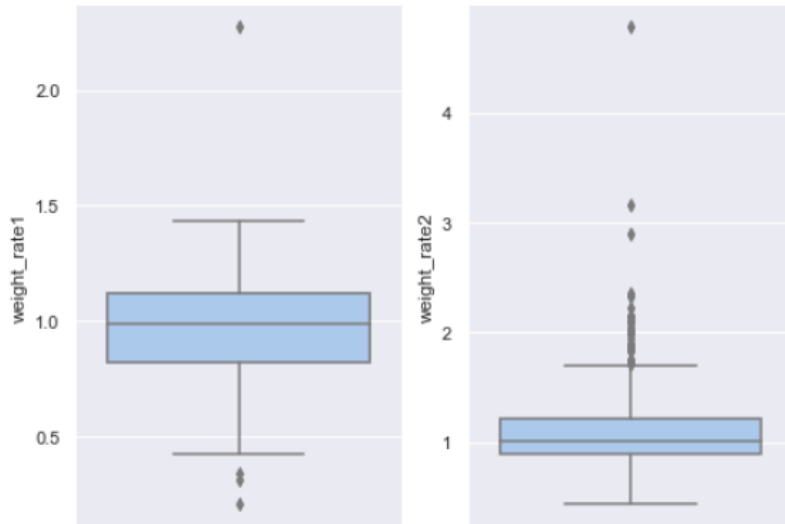


그림 15) weight_rate1, weight_rate2 변수 boxplot

이상치 제거 후 obesity와 adiposity 변수의 산점도는 다음과 같다.

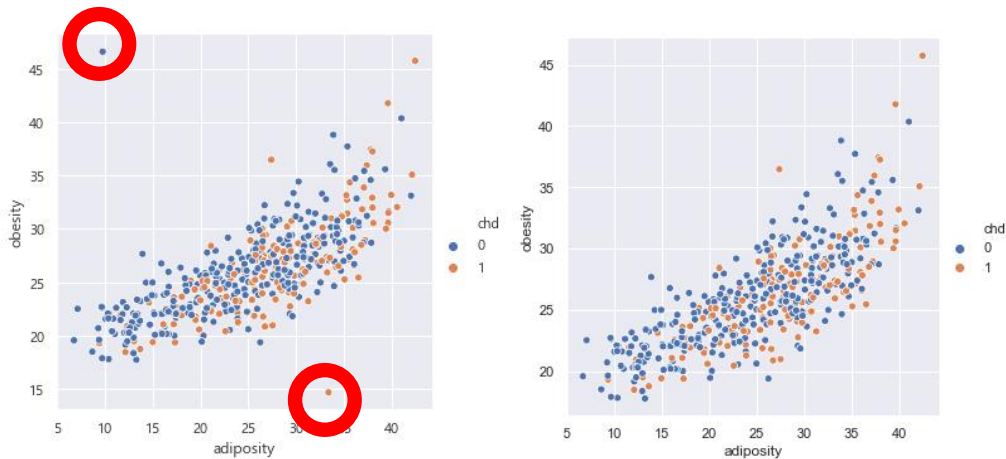


그림 16) 이상치 제거 전 / 후 obesity, adiposity 변수 scatter plot

3.3 파생변수 생성

1. 연령대 (ages) 변수 생성

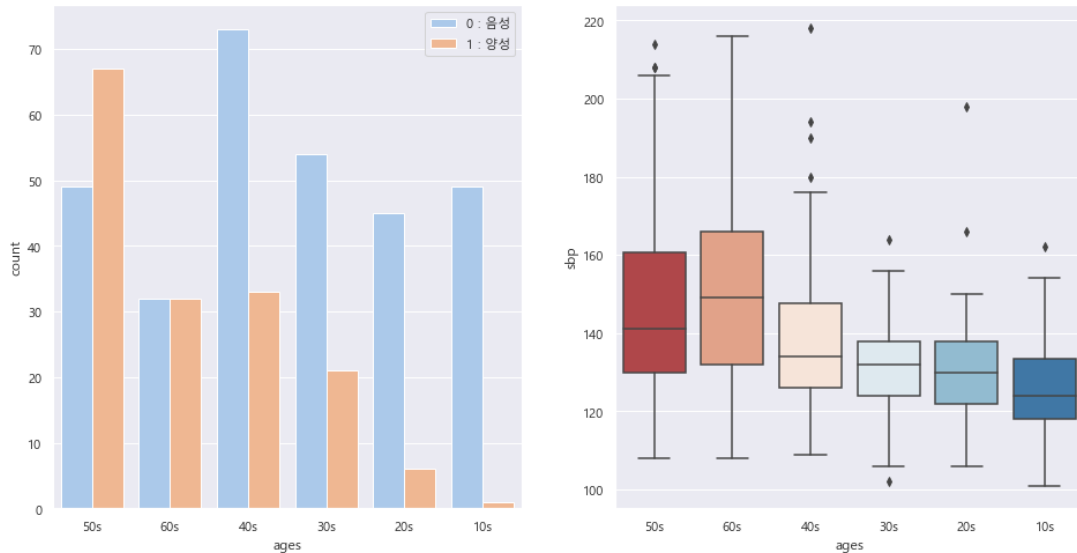


그림 17) 나이 별 chd 분포

피실험자들의 나이를 연령대별로 범주화하여 파생변수를 생성하였다. 높은 연령대일수록 CHD 양성 비율이 높음을 확인할 수 있었다.

2. 상대적 혈압(sbp_rel) 변수 생성

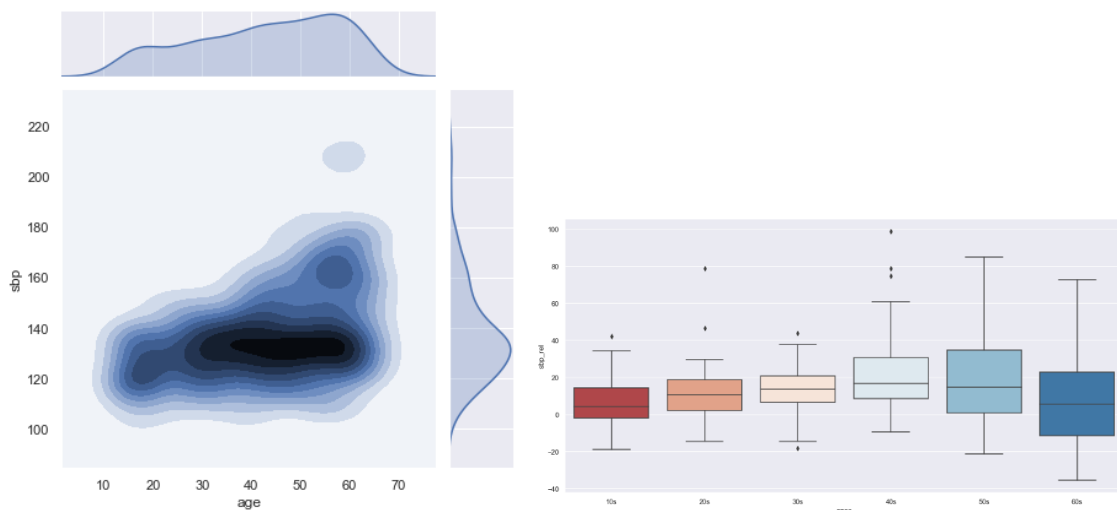


그림 18) sbp, age 변수 joint plot

그림 19) 나이에 따른 sbp_rel

나이가 들수록 혈관의 탄력성이 떨어지고 굳어가기 때문에 혈압이 높아지는 것은 자연스러운 현상이며 본 조의 데이터에서도 나이가 감소할수록 이를 확인할 수 있다. 따라서 평균 혈압의 수치는 연령대에 따라 달라진다. 연령에 따라 평균 혈압 수치가 다를 것을 고려하여 나이(age) 변수에 따라 수축기 혈압(sbp) 변수가 평균 수치로부터 얼마큼 떨어져 있는지 나타내는 파생변수 sbp_rel 을 생성하였다. 평균 수치 값은 미국 남성의 연령 별 평균 혈압 범위를 수축기 혈압과 이완기 혈압을⁹ 활용하였다.

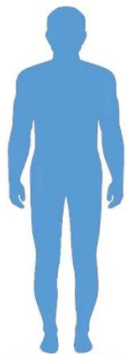
Male	Age	SBP	DBP
	21-25	120.5	78.5
	26-30	119.5	76.5
	31-35	114.5	75.5
	36-40	120.5	75.5
	41-45	115.5	78.5
	46-50	119.5	80.5
	51-55	125.5	80.5
	56-60	129.5	79.5
	61-65	143.5	76.5

그림 20) average BP

3.4 다중공선성

adiposity 변수와 obesity 변수는 모두 비만지수 지표로 본 데이터에서도 높은 상관관계를 가지며 유사도가 높음을 확인할 수 있다. 따라서 VIF(분산팽창요인)을 확인한 후 독립변수들 간 상관관계를 보이는 경우 변수를 제거하고자 한다.

VIF Factor features					
5	107.295706	obesity			
0	101.065930	sbp			
3	50.507558	adiposity			
4	29.271284	typea			
7	27.757084	age			
2	7.929405	ldl			
8	4.129590	sbp_rel			
1	2.529237	tobacco			
6	1.821648	alcohol			

VIF Factor features		
2	21.762682	adiposity
5	17.461977	age
3	9.696577	typea
1	7.888195	ldl
0	2.460869	tobacco
4	1.819809	alcohol
6	1.667994	sbp_rel

그림 21) 변수 제거 전 / 후 VIF

⁹ 『What Are Normal Blood Pressure Ranges by Age For Men and Women?』 (OnHealth, 2021)

최종적으로 가장 높은 VIF 를 보인 obesity 변수와 sbp 변수를 제거하기로 하였다. 앞서 말했듯 obesity 변수는 체지방을 나타내는 BMI 에 대한 변수이다. 해당 변수를 삭제하더라도 높은 상관관계를 가지며 체지방을 나타내는 또 다른 지표 BAI 를 나타내는 adiposity 변수로 설명이 가능할 것이다. 또한 sbp 변수는 기존의 수축기혈압 변수에서 나이를 고려한 sbp_rel 변수로 설명이 가능하다고 판단했다. 이러한 근거를 바탕으로 다중공선성이 발생한 두 변수를 제거한다.

4. 모델링

4.1 데이터 준비

최종 데이터셋은 다음과 같이 총 9 개의 독립변수와 461 행의 데이터로 이루어져있다.

	tobacco	ldl	adiposity	famhist	typea	alcohol	age	chd	sbp_rel
0	12.00	5.73	23.11	1	49	66.8495	52	1	34.5
1	0.01	4.41	28.61	0	55	2.0600	63	1	0.5
2	0.08	3.48	32.28	1	52	3.8100	46	0	-1.5
3	7.50	6.41	38.03	1	51	24.2600	58	1	40.5
4	12.49	3.50	27.78	1	60	57.3400	49	1	14.5
...
456	0.40	4.11	42.06	1	56	2.0600	57	0	40.5
457	0.40	5.98	31.72	0	64	0.0000	58	0	84.5
458	4.20	4.41	32.10	0	52	18.7200	52	1	56.5
459	3.00	1.59	15.23	0	40	26.6400	55	0	-21.5
460	5.40	11.61	30.79	0	64	23.9700	40	0	2.5

그림 22) 최종 데이터셋

설명변수와 반응변수를 각각 3:7 비율로 train set, test set 으로 분리한다. 이는 train 데이터로 학습한 모델이 학습에 사용되지 않은 test 데이터에서 얼마나 정확하게 예측할 수 있는지 확인하기 위함이다.

또한 계층적 k-겹 교차검증 (Stratified k-fold cross validation)을 적용하였다. 이는 현재 데이터의 양이 500 개 미만으로 매우 적은 양에 속한다. 따라서 별도의 검증 데이터를 학습에 재활용하여 최대한 많은 데이터셋을 활용하고자 했다. 하나의 학습 / 검증 데이터로 이루어진 모델은 해당 학습 데이터에만 과적합(overfitting) 될 가능성이 높아 교차 검증을 통해 일반화 된 모델을 생성하고자

한다. 추가로 본 데이터는 양성이 35% 채 되지 않는 불균형한 데이터이다. 따라서 데이터를 분할할 때, 양성과 음성 비율을 유지하여 외적 타당도를 높여 최대한 일반화가 가능하도록 하였다.

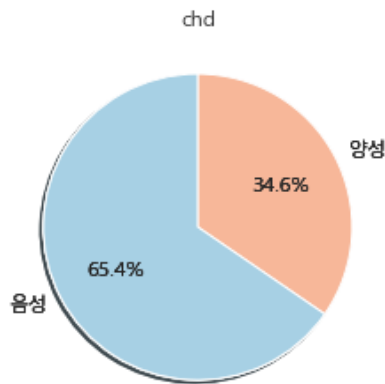


그림 23) Imbalanced Data

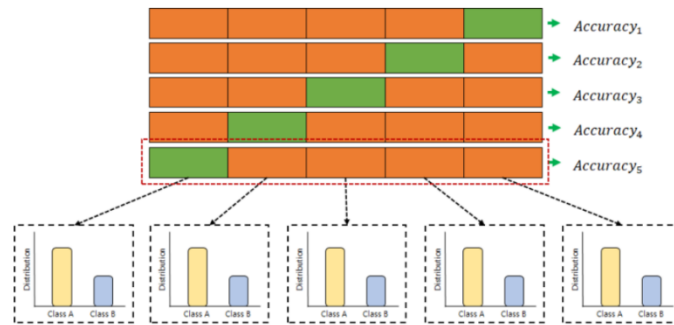


그림 24) 계층적 k-겹 교차검증 이해도

4.2 모델 후보 생성

머신러닝 모델 중 일반적으로 이진 분류 (binary classifier)에 많이 사용되는 아래에 3 가지를 후보로 결정했다.

SVM 은 패턴인식, 자료 분석을 위한 지도학습 모델이며 분류에 많이 사용한다. 주어진 데이터 점들이 두 개의 그룹 안에 각각 속해 있다고 가정했을 때, 새로운 데이터 점이 두 그룹 중 어느 곳에 속하는지 판단하는 이진 선형 분류 모델을 만들게 된다. 만들어진 분류 모델은 데이터가 사상된 공간에서 경계로 표현되는데 SVM 알고리즘은 그 중 가장 큰 폭을 가진 경계를 찾는 알고리즘이다.

로지스틱 회귀분석(Logistic Regression)은 분석하고자 하는 대상들이 두 집단 혹은 그 이상의 집단(다변수 데이터)으로 나누어진 경우에 개별 관측치들이 어느 집단으로 분류될 수 있는가를 분석하고 이를 예측하는 모형을 개발하는데 사용되는 대표적인 통계 알고리즘이다.

랜덤 포레스트(Random Forest)은 다수의 의사결정나무모델에 의한 예측을 종합하는 앙상블 방식이다. 의사 결정 나무와 bagging 을 혼합한 형태라고 볼 수 있다. 부트스트랩을 이용해 학습 집하에서 다양한 샘플을 추출하고 입력 변수 중 일부의 입력 변수만 사용한다. 데이터 샘플링 및 변수 선택을 통해 의사 결정 나무의 다양성을 확보할 수 있다.

각각의 모델을 학습할 때 최적 파라미터를 탐색하는 그리드서치(grid-search)를 함께 진행하였다. 랜덤포레스트의 경우 트리의 깊이, 생성 트리의 개수가 그 예가 될 것이다. 로지스틱 회귀분석에서는 패널티의 유형, 패널티 강도가 있고 SVM에선 가우시안 커널 폭을 제어하는 gamma 값이 있다. 이러한 파라미터의 후보군을 설정하여 최적의 하이퍼 파라미터 (hyper-parameter)를 찾아낸다.

Model	Hyper-parameter Grid Search
LogisticRegression	'penalty' : ['l1', 'l2', 'elasticnet'] 'C' : [0.001, 0.01, 0.1, 1, 10, 100]
RandomForest	'max_depth' : [4, 6, 8] 'n_estimators' : [10, 50, 100] 'min_samples_split' : [10, 100]
SVM	'gamma' : [0.01, 0.1, 1, 10] 'C' : [0.01, 0.1, 1, 10]

그림 25) 3가지 모델 그리드서치

4.3 모델 선정 기준

머신러닝 모델의 성능을 나타내는데 많은 지표들이 있다. 정확도(accuracy), 재현율(recall), F1 스코어(f1 score)등이 그 예인데 일반적으로 많이 사용하는 정확도 지표는 본 데이터에서 적합하지 않다고 판단했다. 그 이유는 불균형한 데이터에서 보다 적은 가능성으로 발생할 상황에 대해 제대로 된 분류를 해주는지 평가가 불가능하기 때문이다. 이를 정확도 역설(accuracy paradox)이라고도 한다. 따라서 정밀도와 재현율의 조화평균인 F1 score 를 평가지표로 사용하고자 한다. 조화평균을 사용한 이유는 차지하는 비중이 큰 경우의 bias를 줄여주어 불균형이 일어났을 때도 모델의 성능을 정확히 평가하기 위함이다. [부록 9]

4.4 최종 모델 선정

최종적으로 f1 score 가 0.59로 C는 0.1, 패널티는 'l2'를 사용한 로지스틱 회귀모형이 선정되었다. 회귀식은 다음과 같다.

$$\text{logit}(chd) = -6.442 + 0.088 * tobacco + 0.114 * ldl - 0.017 * adiposity + 0.631 * famhist + 0.038 * typea - 0.003 * alcohol + 0.065 * age + 0.004 * sbprel$$

그림 26) 최종 모델 로지스틱 회귀식

이 식은 tobacco 가 1 단위 증가할 때마다, chd 발병 오즈(odds)가 $\exp(0.0876)$ 배 즉 1.09 배 증가함을 의미한다. 또한 famhist 와 같은 더미변수의 경우 가족력이 있을 때의 오즈가 없을 때보다 $\exp(0.631)$ 배 즉, 1.8 배 더 높다고 해석할 수 있다.

4.5 변수 중요도

로지스틱 회귀모형의 경우 표준화(scale)를 통해 변수 중요도를 비교할 수 있다. 각 변수의 로지스틱 회귀계수 절대값이 0.666 으로 가장 큰 age 변수가 chd 에 가장 영향을 많이 주는 것을 알 수 있다. 이는 나이가 많을수록 몸의 전체적인 면역력이 낮아져 쉽게 병에 노출됨으로 해석된다.

	columns_scale	coef
0	intercept	-0.845355
0	tobacco	0.328977
1	ldl	0.207478
2	adiposity	-0.026989
3	famhist	0.452744
4	typea	0.281083
5	alcohol	-0.053249
6	age	0.665992
7	sbp_rel	0.054462

그림 27) 표준화 회귀계수

5. 결론

5.1 한계점

해당 데이터는 피실험자들이 CHD 양성 판정 이후 혈압 감소 치료와 위험요인을 줄이기 위한 조치를 받고 측정한 데이터이다. sbp 변수의 경우 혈압 감소 치료때문에 실제 혈압을 파악하기 힘들고 데이터상의 수치와 차이가 있을 수 있다. 그에 따라 관련 변수인 sbp_rel 변수의 중요도가 낮게 나왔음을 고려해야한다. 이와 같이 사전 조치로 인해 특정 변수에서는 일반적으로 CHD 가 양성인 경우의 특징을 보여주지 못할 가능성이 있고 정확한 측정치라고 판단하기 어렵다. 또한 데이터는 462 개의 행으로 관측치가 매우 적어 미세한 차이에도 분석 결과가 상이해질 수 있다. 제공된 변수

설명에서는 단위와 측정기간의 언급이 없어 ldl 변수와 같이 수치로 유추해야 했다. alcohol 변수는 ‘최근의 알코올 소비량’이라는 설명이 있지만 ‘최근’의 기준을 알 수 없어 단위의 유추가 불가능하고 분석 기준을 결정하기 어려웠다. 마지막 한계점으로, 본 데이터는 앞선 설명에서도 언급했듯이 남아프리카 공화국에서 백인남성을 대상으로 실시된 조사이다. CHD 발병은 인종과 성별에 따라 위험요인의 영향과 그 정도가 다르다. 특정 인종과 성별에 편향된 데이터이기 때문에 분석결과를 전 지구적으로 활용하기 어렵다.

5.2 활용방안

1. 효율적인 보험 관리

2018 년 기준 대한민국의 가구당 평균 보험 상품의 개수는 12 개로 조사되었다.¹⁰ 금융소비자연맹은 GDP 를 고려했을 때, 국민 1 인당 보험료 지출 순위가 세계 6 위로 경제력 대비 보험료 지출이 많은 편이라 평가한다. 이러한 배경에는 자신의 건강이 언제 나빠질지 모르는 불확실성이 있다. 만약을 대비하고자 하는 생각은 좋으나 과하지 않은 적정선을 유지하는 것이 중요하다. 현재 보험료 계산을 할 때, 기준으로 삼는 것은 나이이다. 나이에 따라 보험료가 크게 달라지며 그 다음으로 성별, 과거 입원 이력 등으로 세부 조정된다. 이는 본 프로젝트에서 로지스틱 회귀분석의 표준화 회귀계수 중 age 변수 크기가 가장 높았다는 분석과 동일하게 흘러간다. 하지만 피보험자 입장에서 자신이 특정 질병에 대한 가능성을 어느 정도 유추해볼 수 있다면 먼저 필요한 보험을 판단할 수 있을 것이며, 이는 단순히 근거 없이 막연한 걱정만으로 들었던 보험의 수를 줄여주는 순기능을 낳는다. 따라서 본 프로젝트에서 진행한 방식을 차용하여 다양한 질병 데이터를 대상으로 모델링하고 예측한 후 자신의 현재 건강 상태에 필요한 보험을 판단할 수 있게 한다.

내 보험료는 얼마?!

이름

생년월일

연락처

보험료 계산 ▼

무료상담신청 ▼

님의 예상 월보험료는

월

원 입니다.

그림 28) 라이나 생명 나이 기준 보험료 계산 시스템

¹⁰ "가구당 보험상품 12개 가입... 월소득 18% 보험료로 지출," *한국경제*, 2018년 1월 9일 수정, 2021년 6월 2일 접속, <https://www.hankyung.com/economy/article/201801090678i>.

2. 웨어러블 기기 건강 정보 앱

현재 다양한 웨어러블 기기(wearable device)가 시중에 나와 있다. 이러한 시장의 열풍은 앞으로도 계속 그치지 않을 것이라 많은 전문가들이 말한다.¹¹ [부록 11]

이와 더불어 쉽게 자신의 심박수를 체크하고 운동량을 체크하는 등 건강에 대한 관심도 커졌다. 본 프로젝트에서 분석한 결과를 웨어러블 기기에 적용해보고자 한다. 간단한 개인 정보를 입력하고 웨어러블 기기를 통해 직접 측정 가능한 건강 상태 정보들은 바로 앱으로 전달하여 현재 상태에서 특정 질병의 위험도를 파악할 수 있게 해준다.

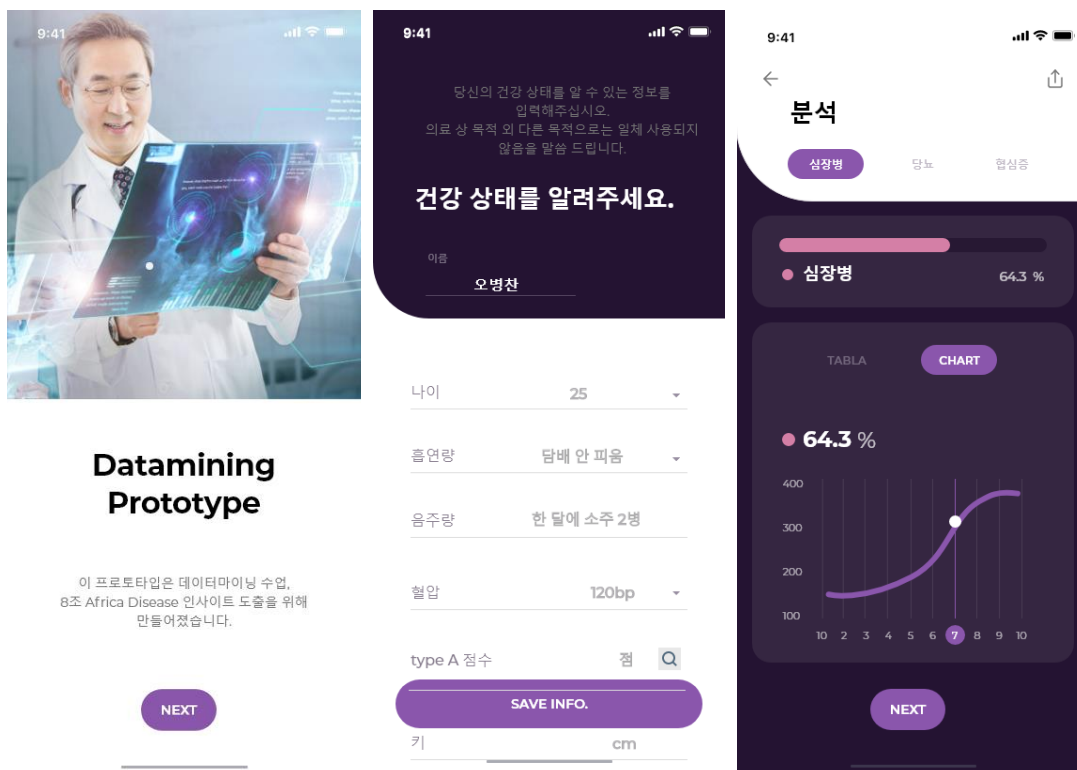
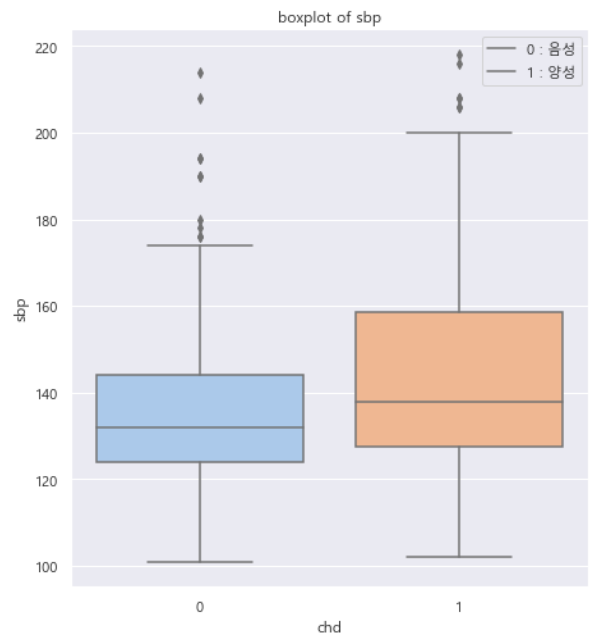
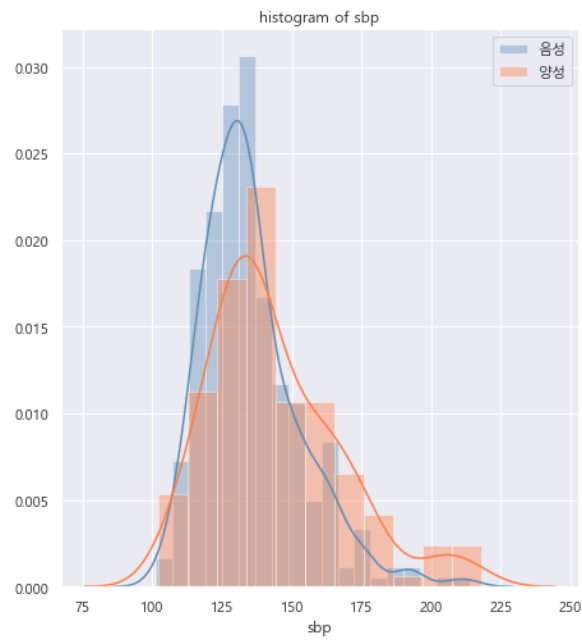


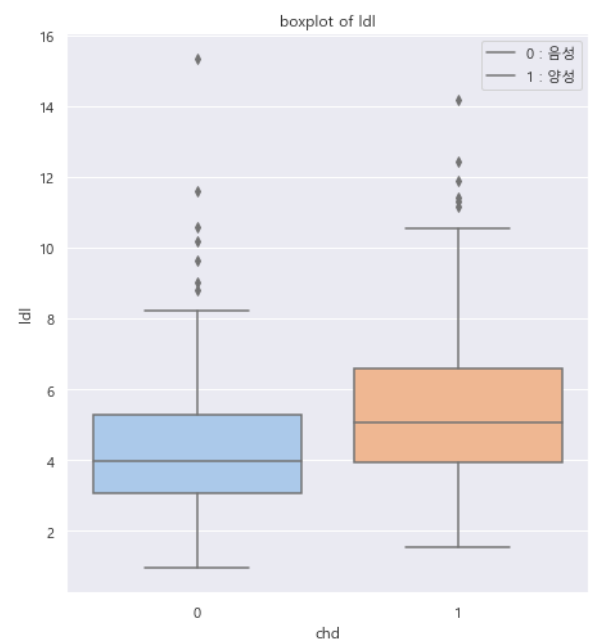
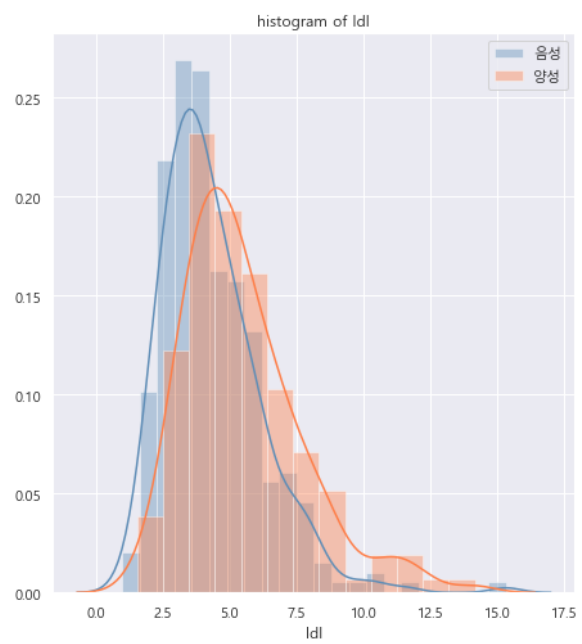
그림 29) 자가 건강 진단 어플리케이션 프로토타입

¹¹ "IDC Forecasts Steady Double-Digit Growth for Wearables," *Printed Electronics*, 2019년 수정, 2021년 6월 2일 접속, https://www.printedelectronicsnow.com//contents/view_breaking-news/2019-03-20/idc-forecasts-steady-double-digit-growth-for-wearables/.

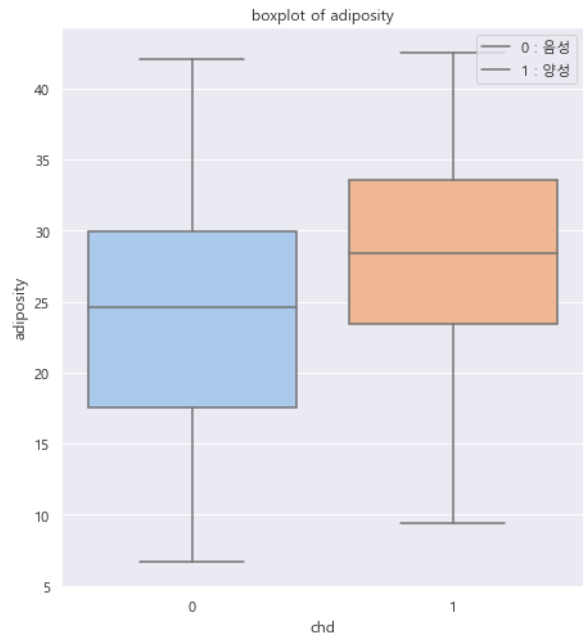
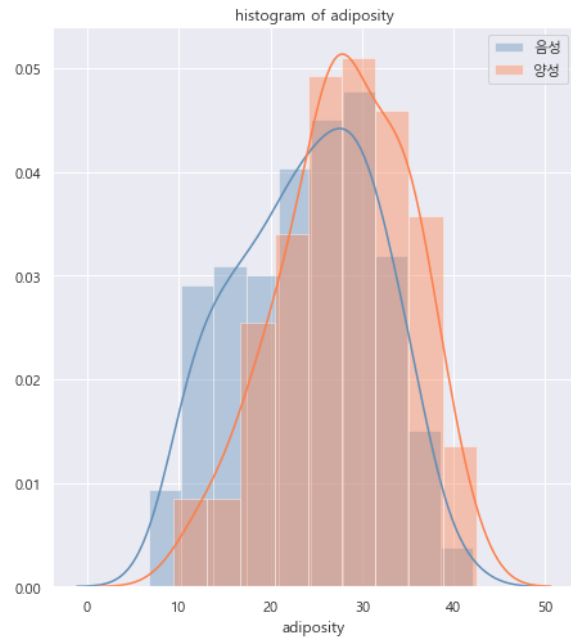
IV. 부록



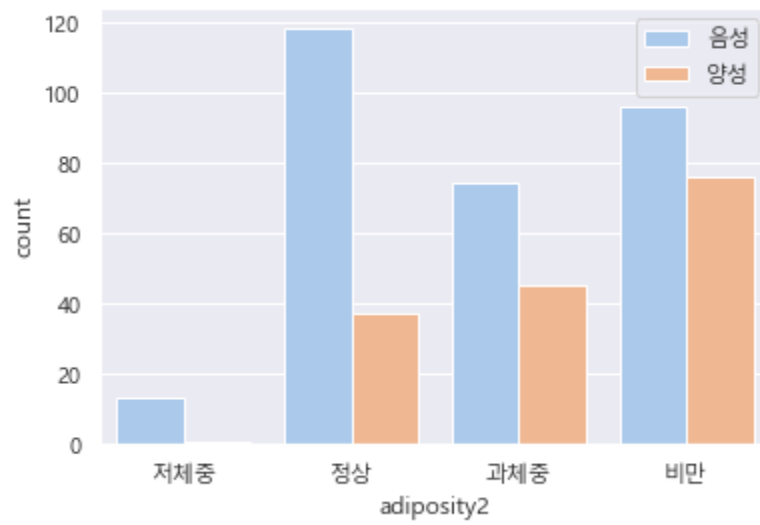
부록 1) sbp 분포



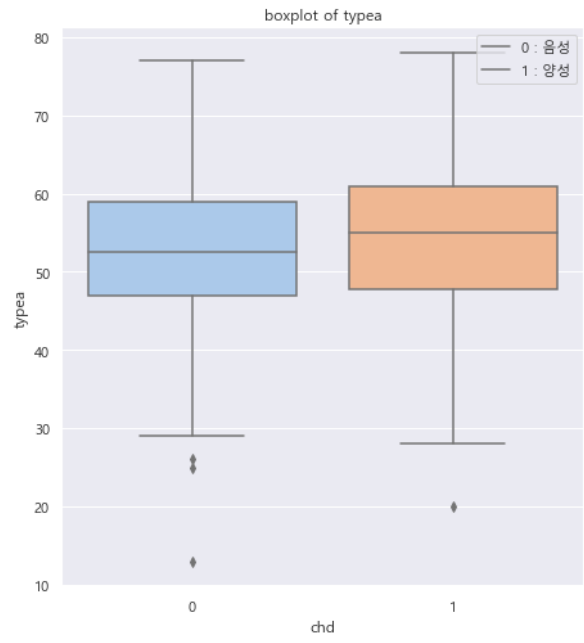
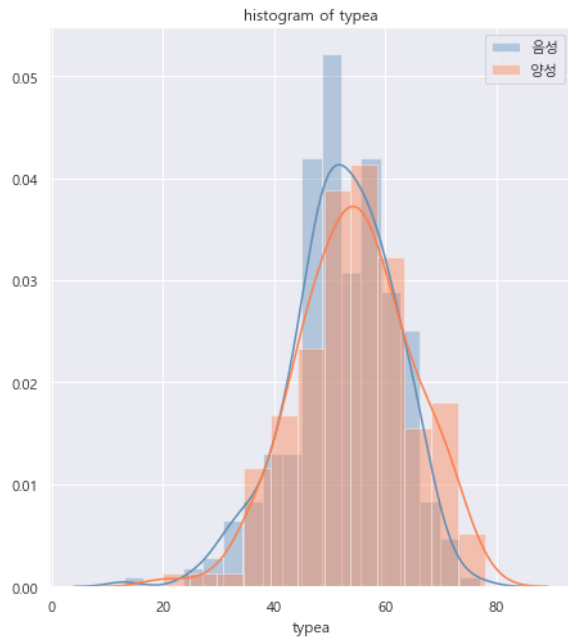
부록 2) ldl 분포



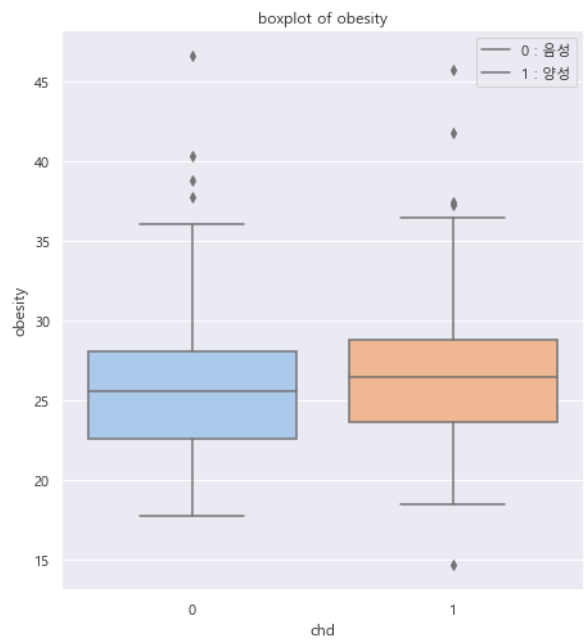
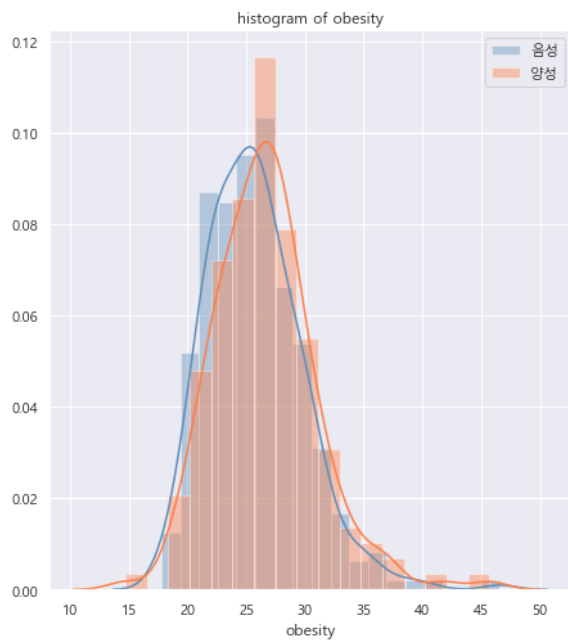
부록 3) adiposity 분포



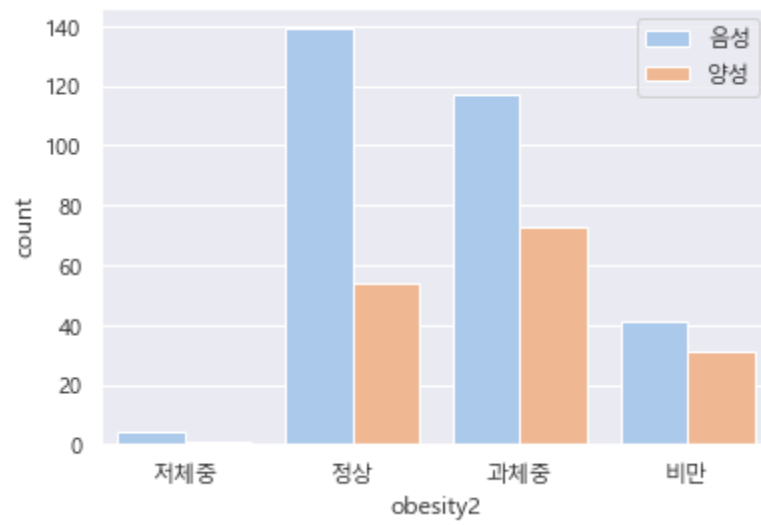
부록 4) adiposity 그룹별 chd 발병빈도



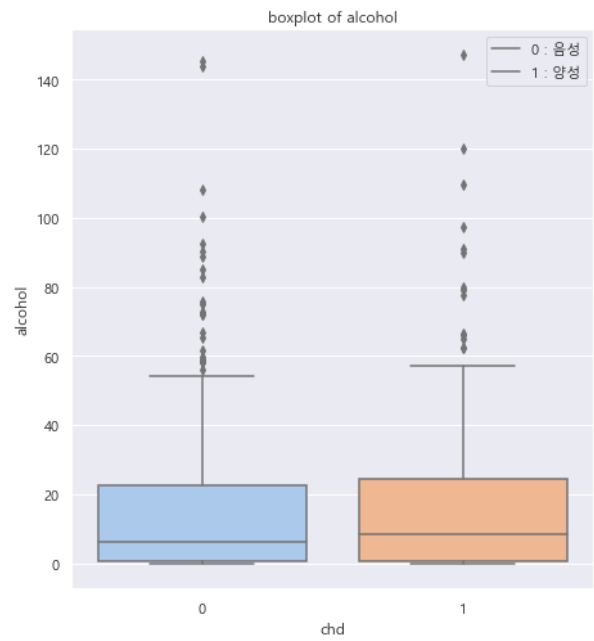
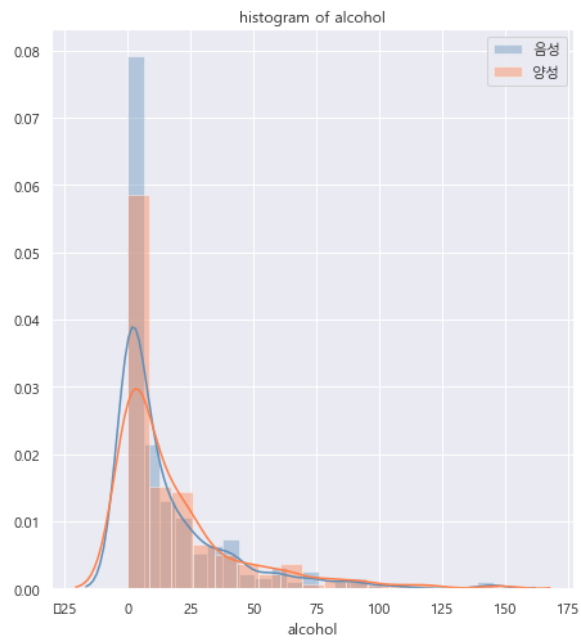
부록 5) typea 분포



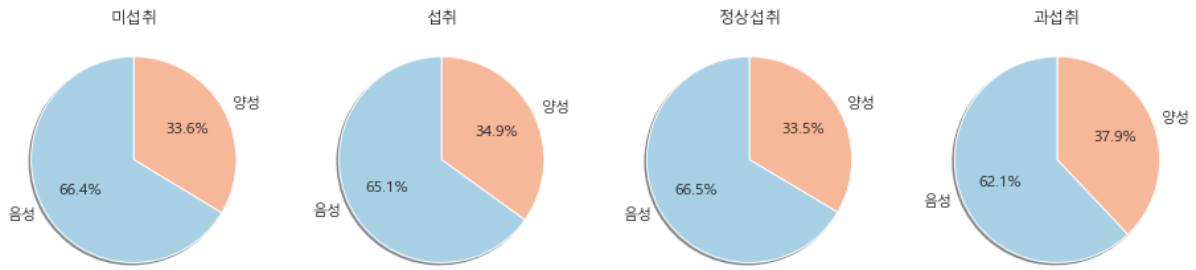
부록 6) obesity 분포



부록 7) obesity 그룹별 chd 발병빈도



부록 8) alcohol 분포



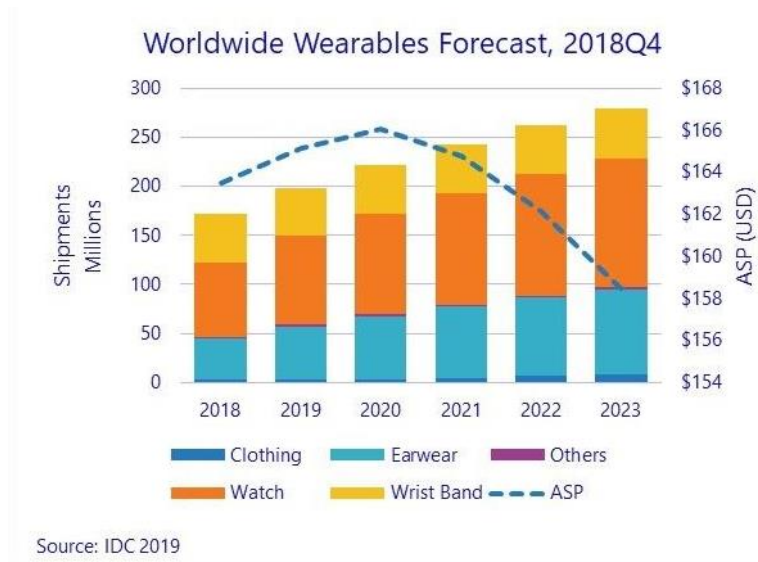
부록 9) 알코올 섭취량에 따른 chd 발병률

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 - score = 2 \times \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}}$$

부록 10) 모델 성능 지표



부록 11) 웨어러블 기기 시장 동향

참고문헌

<https://great-northern-diver.github.io/loon.data/reference/SAheart.html>

<https://www.sciencedirect.com/science/article/abs/pii/S0146280620301328>

건양대학교 심장혈관센터 https://www.kyuh.ac.kr/department/cardiology/sub2_6.html

인천권역심뇌혈관질환센터

<https://m.blog.naver.com/PostView.nhn?blogId=iccv&logNo=221766677319&proxyReferer=https:%2F%2Fwww.google.com%2F>

(https://www.heart.org/idc/groups/heart-public/@wcm/@sop/@smd/documents/downloadable/ucm_449846.pdf)-그래프 참고

<https://www.kaim.or.kr/bbs/index.html?code=health&category=&gubun=&page=1&number=9987&mode=view&keyfield=&key=>