건강 검진 결과를 토대로 흡연 여부 예측하기

팀 : 와글바글

팀원 : 박영원(대표), 서문세완, 홍수호

KU KONKUK UNIVERSITY

데이터 소개

국민건강보험공단이 실시한
2009~2020 기초건강검진 결과

총 12만 행
30개의 열

22 / 건강검진 결과

키, 몸무게, 혈압, 콜레스테롤 등

8 / 인적사항 및 그 외

검사연도, 나이, 성별, 지역 등

프로젝트의 목표

건강검진 결과만으로 흡연 여부를 판단하는 모델 만들기

여러 가지 예측 모델과 성능을 향상시킬 수 있는 기법을 동원하여
최선의 모델 찾기

흡연 여부를 판단하는데 큰 영향을 끼친 피처 찾기

모델에서 제공하는 메서드를 활용하여 특성 중요도를 확인하고
시각화해보기

데이터 전처리

변수 삭제

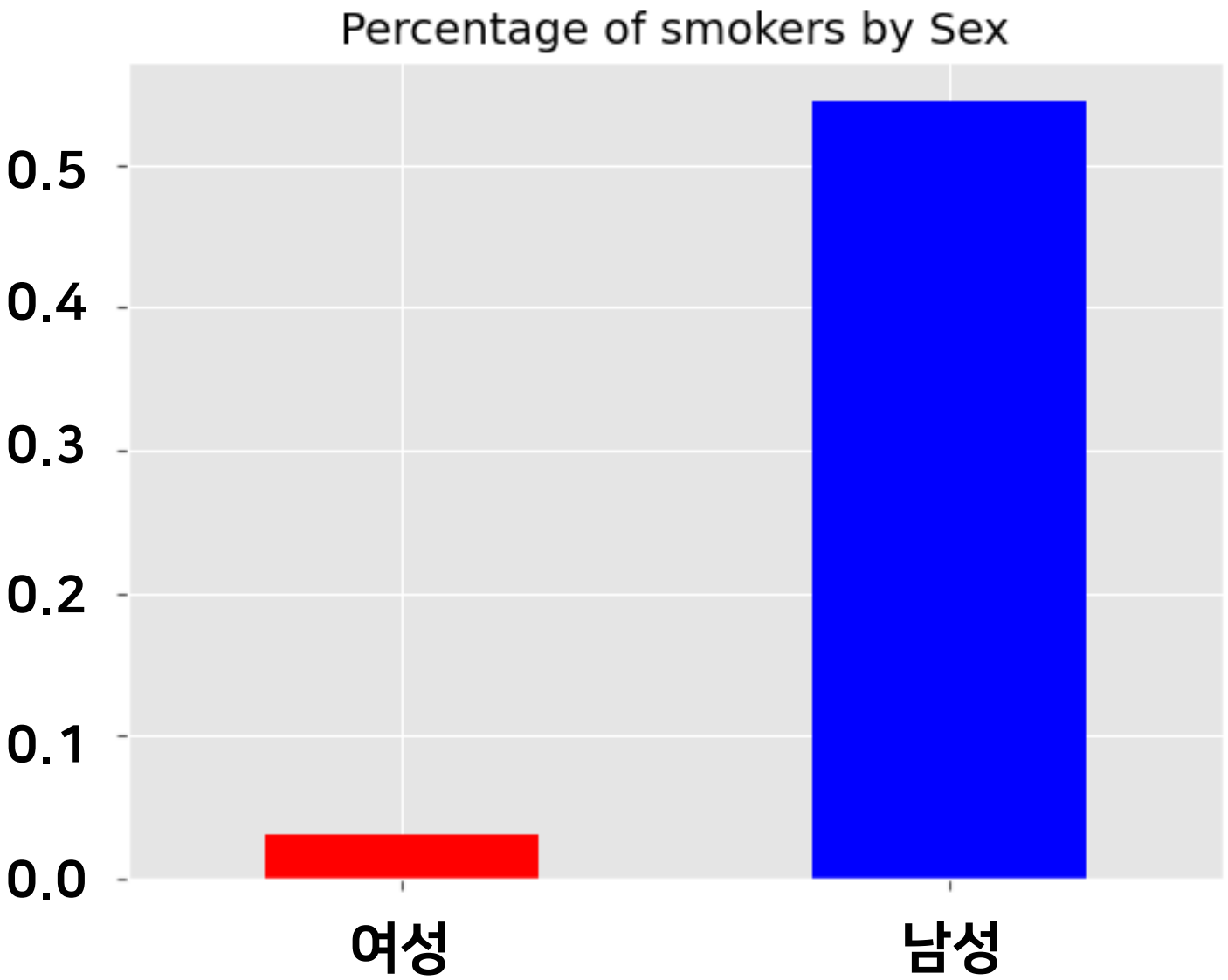
- 건강검진의 결과가 아닌 변수 삭제
- 결측값 비율이 높은 변수 삭제
- 다른 변수로부터 영향을 받는 변수 삭제

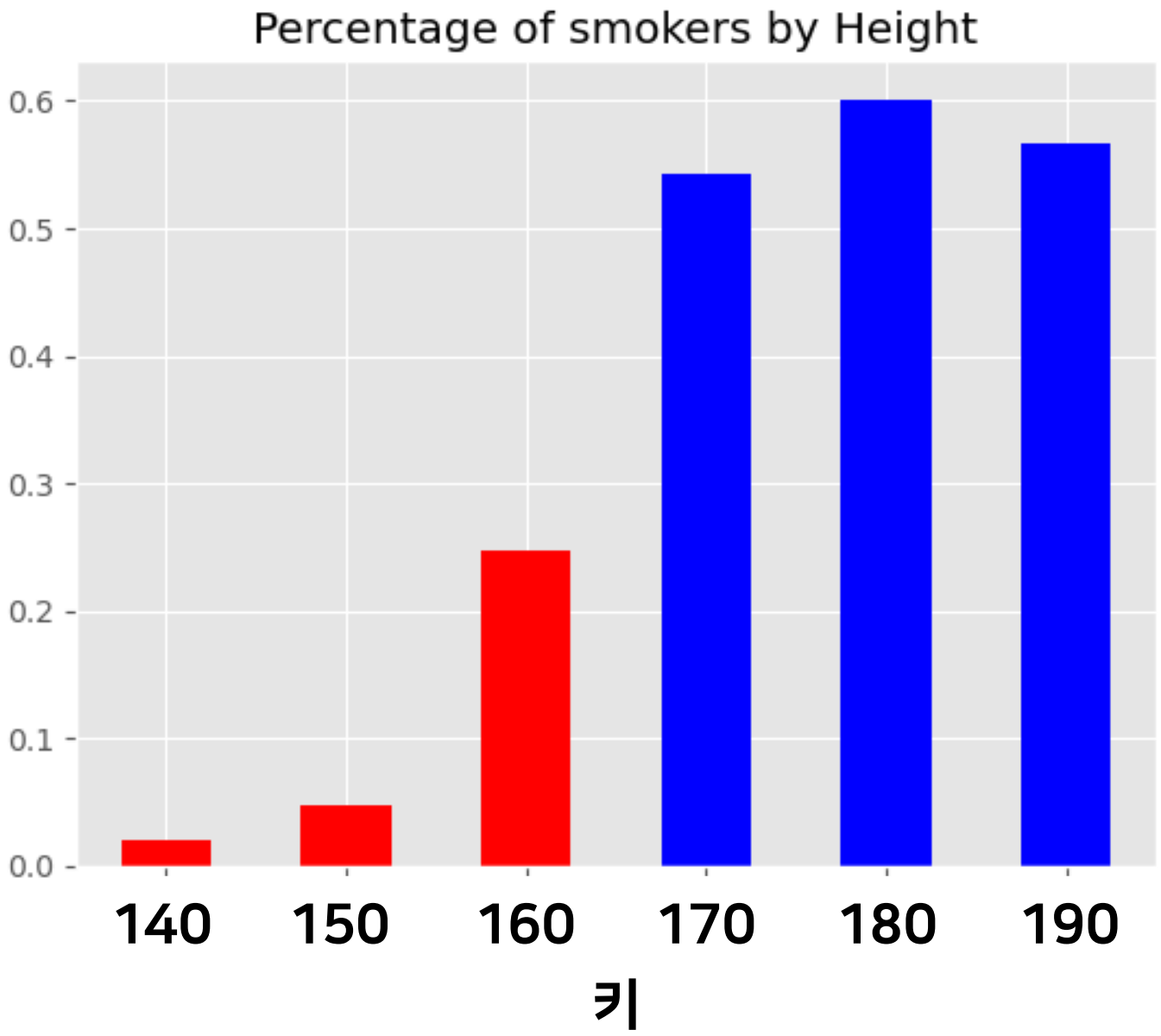
행 삭제

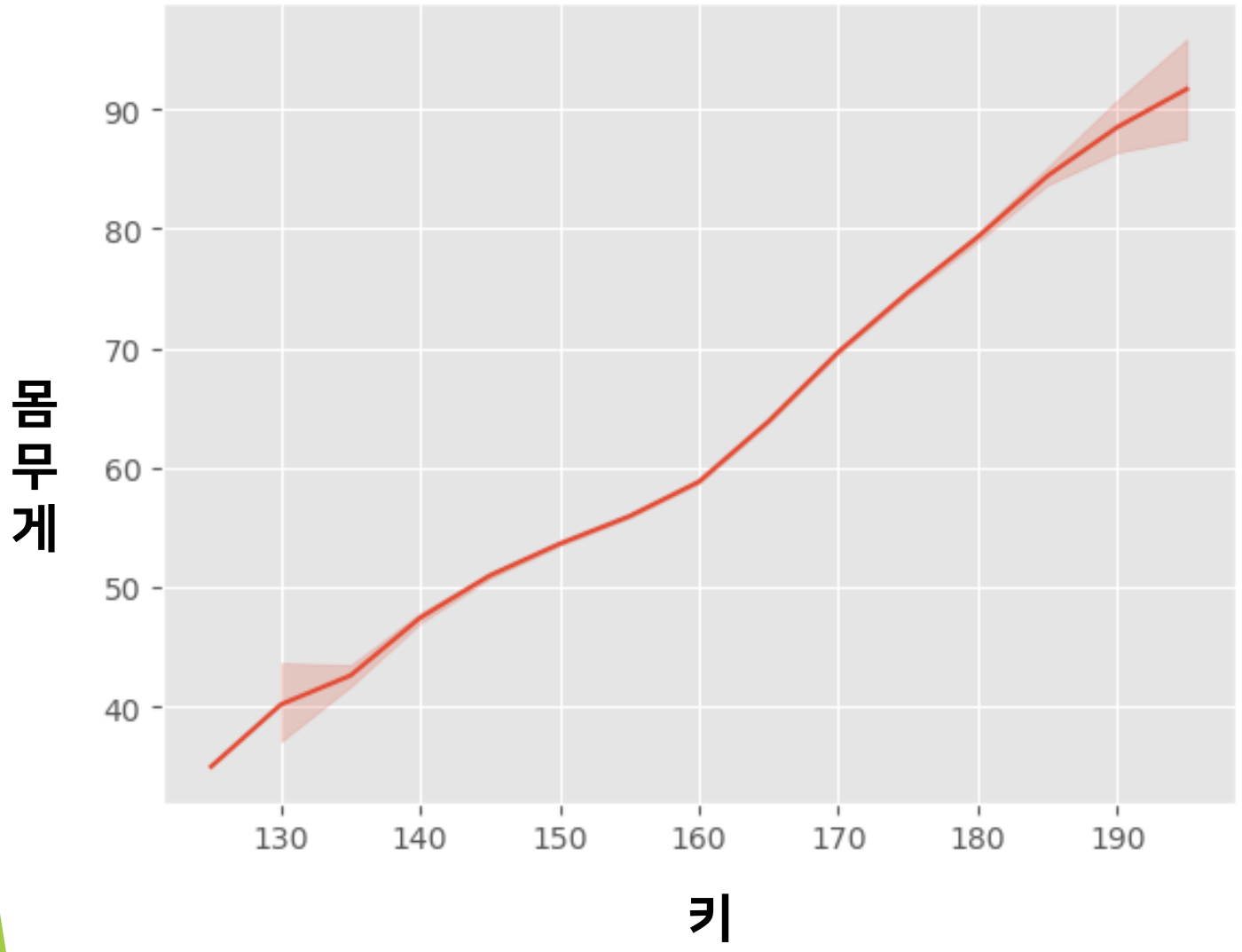
- 타깃변수가 결측된 행 삭제
- 금연자 데이터 삭제
- 혈액 검사 관련 변수가 결측된 행 삭제

데이터 전처리

변수 변환	몸무게 -> BMI 지수로 변환
결측치/이상치 대체	결측값은 중앙값으로 대체
	청력과 시력 변수의 이상치 대체
데이터 분리	전체 데이터의 10%는 테스트 데이터로 분리
	남은 데이터의 10%는 검증 데이터로 분리


Percentage of smokers by Sex
0.5
0.4
0.3
0.2
0.1
0.0
여성
남성


Percentage of smokers by Height
0.6
0.5
0.4
0.3
0.2
0.1
0.0
140
150
160
170
180
190
키



몸무게와의 상관계수
키 : 0.66
성별 : 0.56

BMI와의 상관계수
키 : 0.06
성별 : 0.16

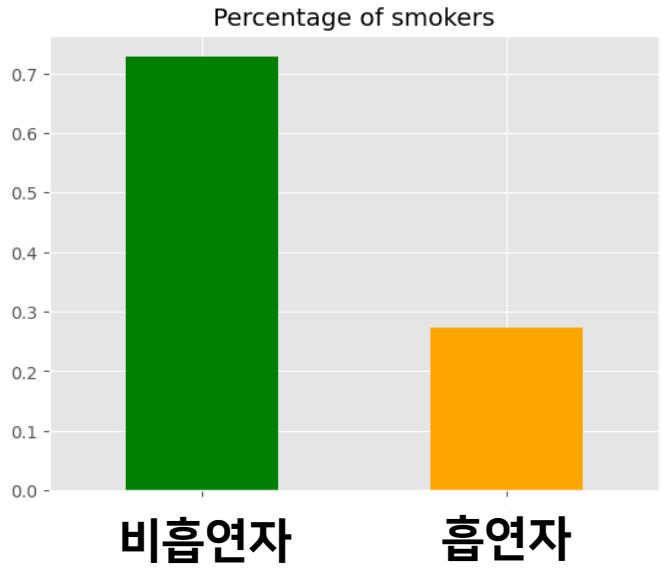
데이터 전처리

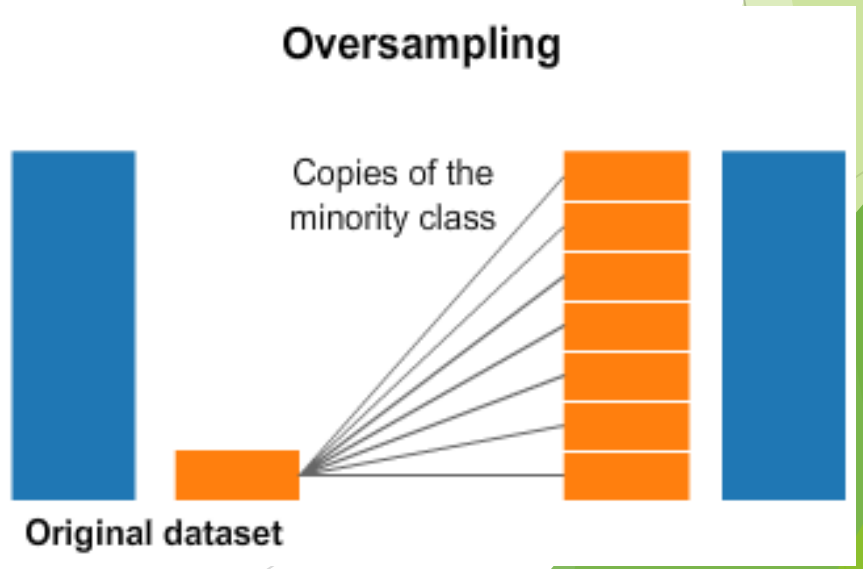
데이터 표준화

StandardScaler를 활용하여 일괄적으로 표준화

데이터 불균형 해소

훈련 데이터의 흡연자 데이터 오버샘플링





모형 적합

- lightGBM
- 로지스틱 회귀
- 랜덤포레스트
- XGBoost
- K-최근접 이웃
- 나이브 베이즈

- GridSearchCV
- RandomizedSearchCV

튜닝 파라미터

로지스틱 회귀	규제 강도	solver	
나이브 베이즈	평탄화 변수		
K-최근접 이웃	이웃 K		
랜덤포레스트	결정트리 수	불순도 측정기준	최대 깊이
lightGBM	학습률	규제 강도	
XGBoost	결정트리 수	최대 리프 수	최대 깊이

모델 검증 정확도

lightGBM	로지스틱 회귀	랜덤포레스트
80.494%	78.789%	78.628%
XGBoost	K-최근접 이웃	나이브 베이즈
78.212%	72.882%	72.453%

모델 개선 방안

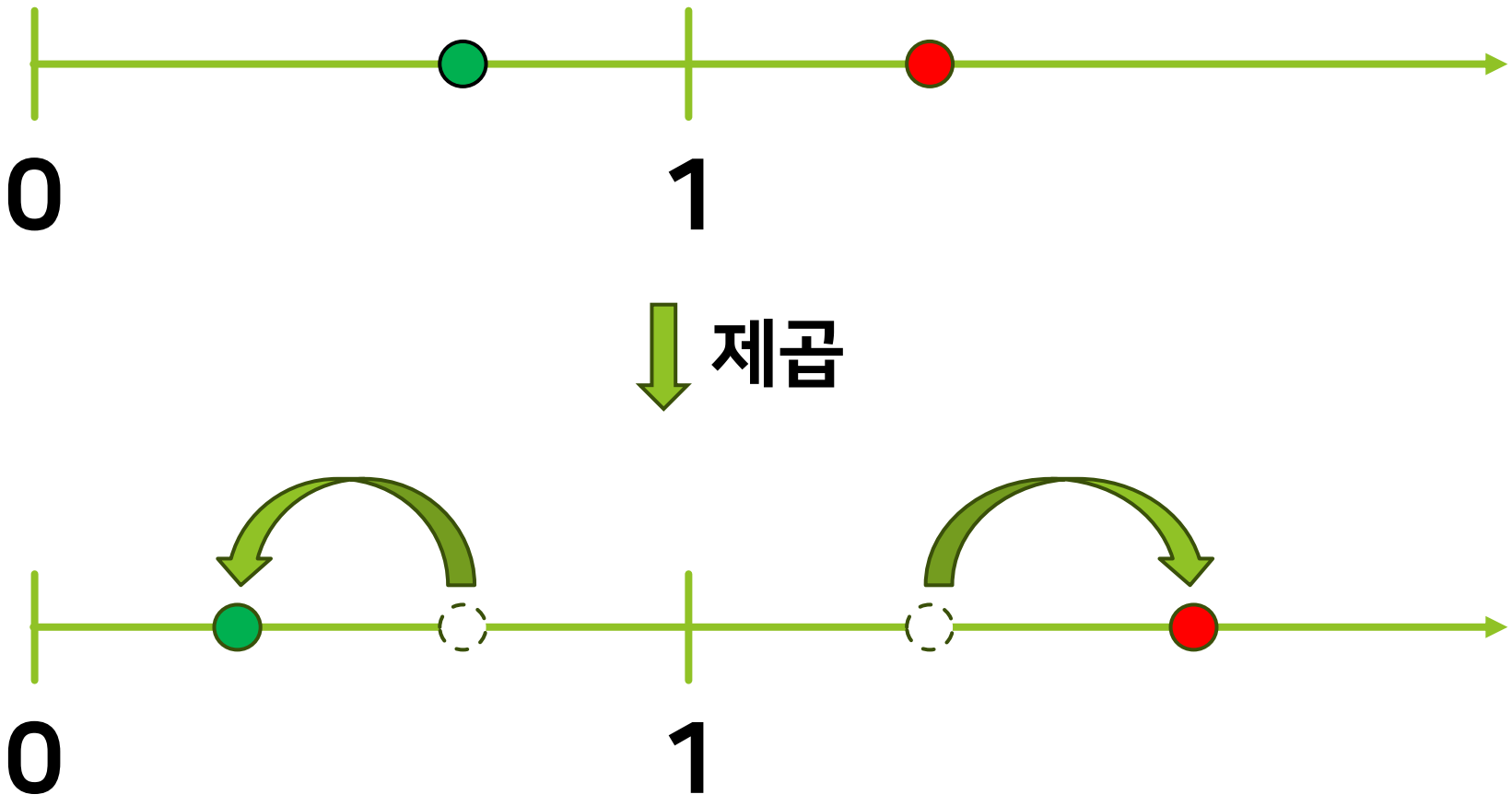
$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

▲ 최소-최대 정규화

$$x_{scaled} = \frac{\text{x – 정상 범주 하한}}{\text{정상 범주 상한 – 정상 범주 하한}}$$

▲ 정상 범주를 고려한 스케일링

모델 개선 방안



모델 개선 방안

기존 모델

80.494%

모든 피처에 적용한 경우

80.709%

SGPT_ALT, 크레아티닌, 혈색소

81.407%

최종 모델 평가

최종 모델 : lightGBM

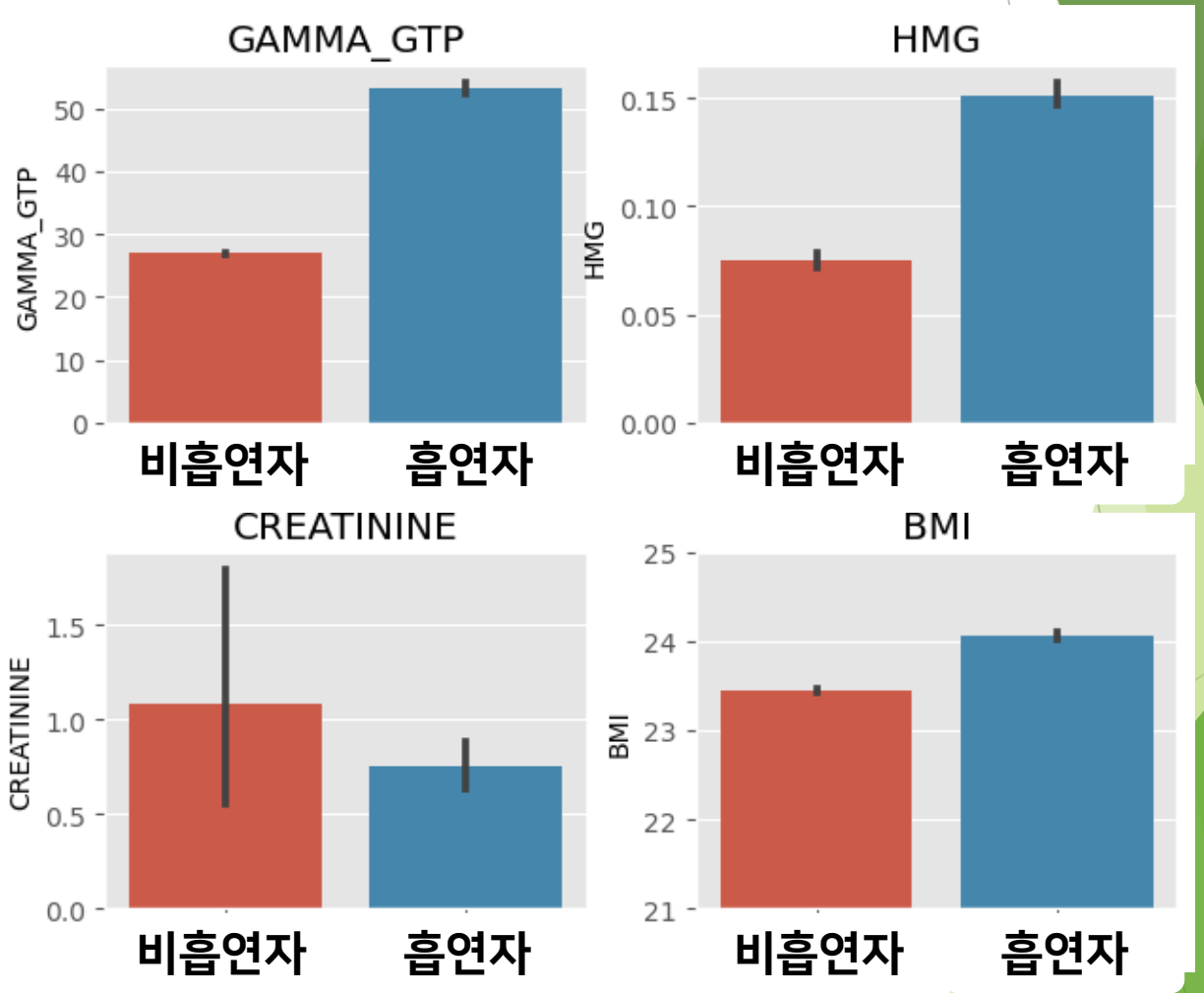
테스트 정확도 : 79.159%

테스트 정밀도 : 57.945%

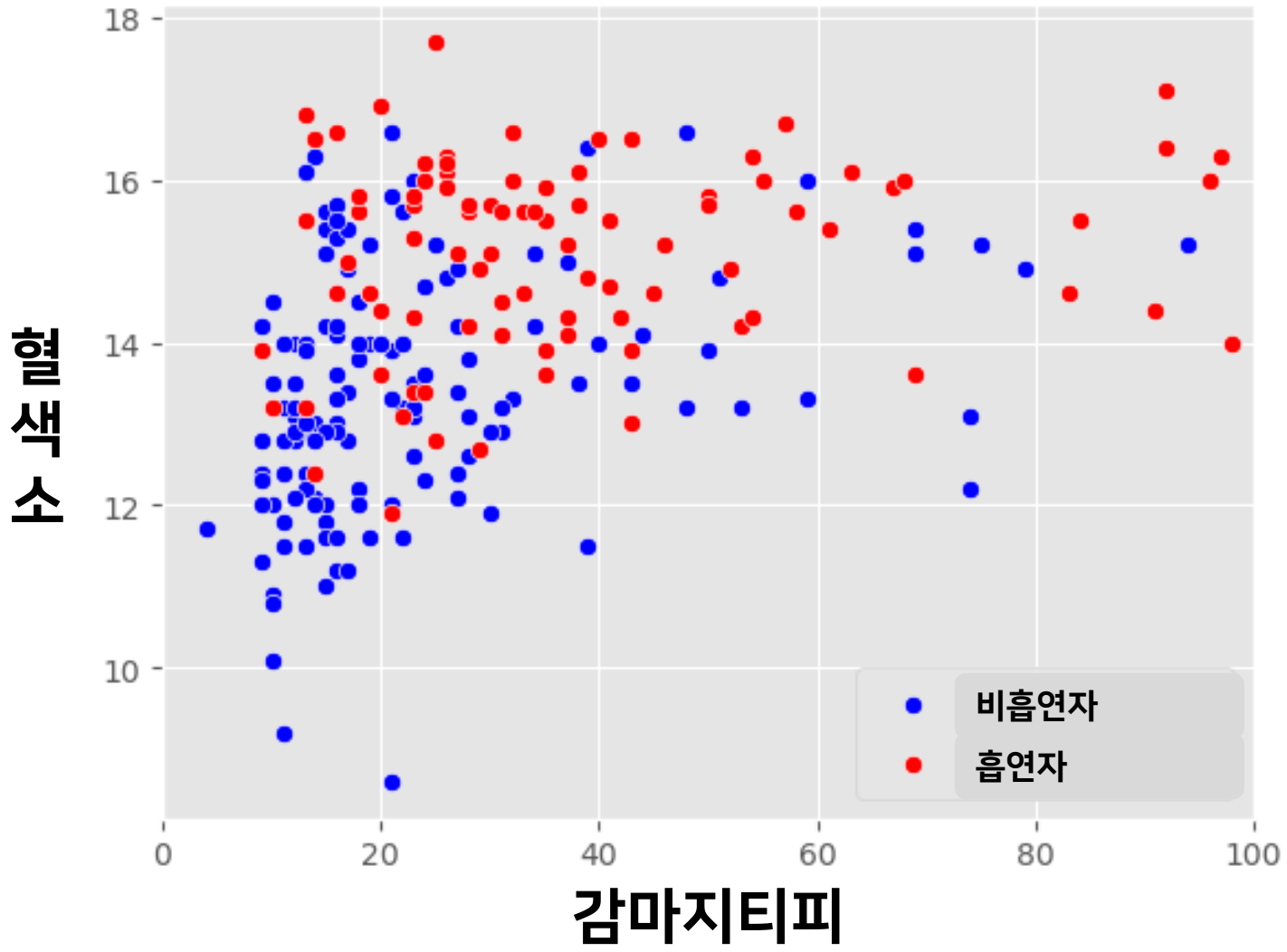
테스트 재현율 : 85.955%

특성 중요도 분석

피처	특성 중요도 비율
감마지티피	0.303
혈색소	0.229
크레아티닌	0.145
BMI	0.087



특성 중요도 분석



아쉬운 점

감사합니다.