KUBIG 2022-2R ML 분반
의료데이터팀

Heart Attack 분류모델

김진서 신인섭 이영노 천원준

목차

Target 변수

심장마비를 경험할 확률이 적음 = 0
심장마비를 경험할 확률이 더 높음 = 1
이항형(Binary) 변수

→

원데이터 변수들

- Age 나이
- Sex 성별(2범주)
- Exang :운동 유발성 협심증(이항)
- cp: 가슴 통증 종류 (4범주)
- trtbps: 휴식기 혈압
- chol : 콜레스테롤 (in mg/dl)
- fbs : 단식 후 혈당 (이항형)
- rest_ecg : 심전도 결과 (3범주)
- thalach : 최대심박수
- oldpeak : 이전 peak
- slp : 기울기 (3범주)
- thall : 협심증 검사 결과 (3범주)
- caa : 중요 정맥숫자(4범주)

1) 아웃라이어 처리

Z*3 이상치 제거

2) 다중공선성(multicollinearity) 문제

(correlation plot 다음 슬라이드에 있음)

Lidge와 Lasso의 경우 종속 변수가 이항변수여서 X

변수제거의 경우, 정량적이지 않고 논리가 부족할 것으로 예상

따라서, PCA를 통해 문제를 해결하기로 하였음

Scree plot 기준 완만지점인 12를 기준으로 주성분 개수 선정

추후, PCA 사용 유무에 따른 성능 판단을 위해

100번, 1000번의 test 결과 평균끼리 비교하였음

(Train&Test split의 무작위성으로 인함)

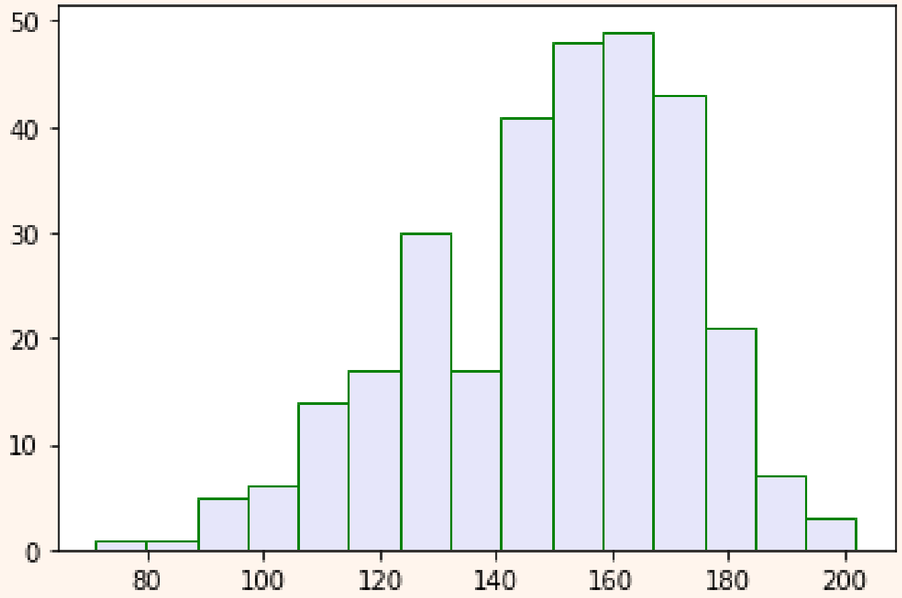
최종 Features

범주형 데이터

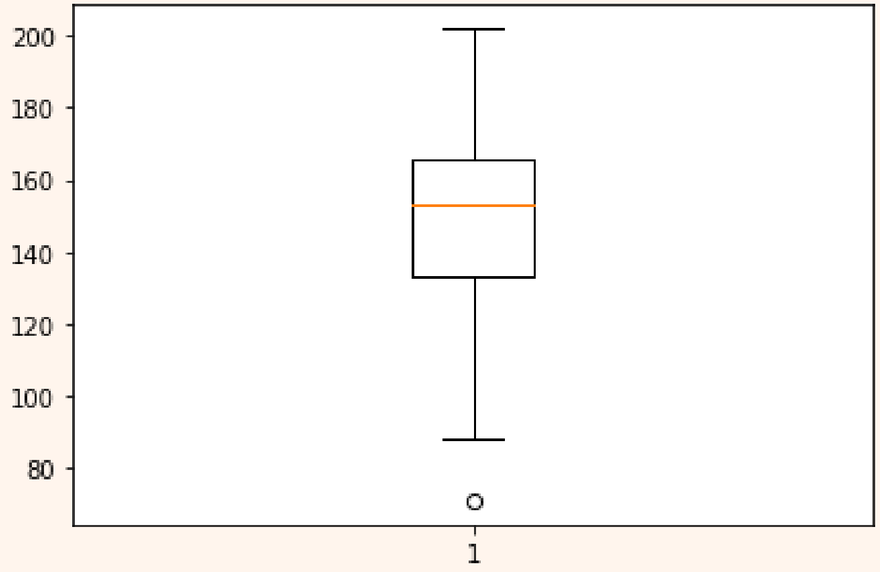
sex, cp, fbs, restecg, exang, slp, caa, thall

연속형 데이터

age, trtbps, chol, thalachh, oldpeak

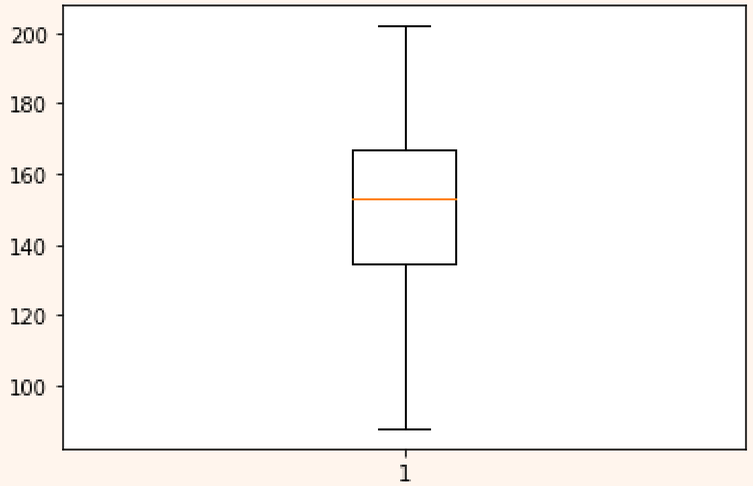


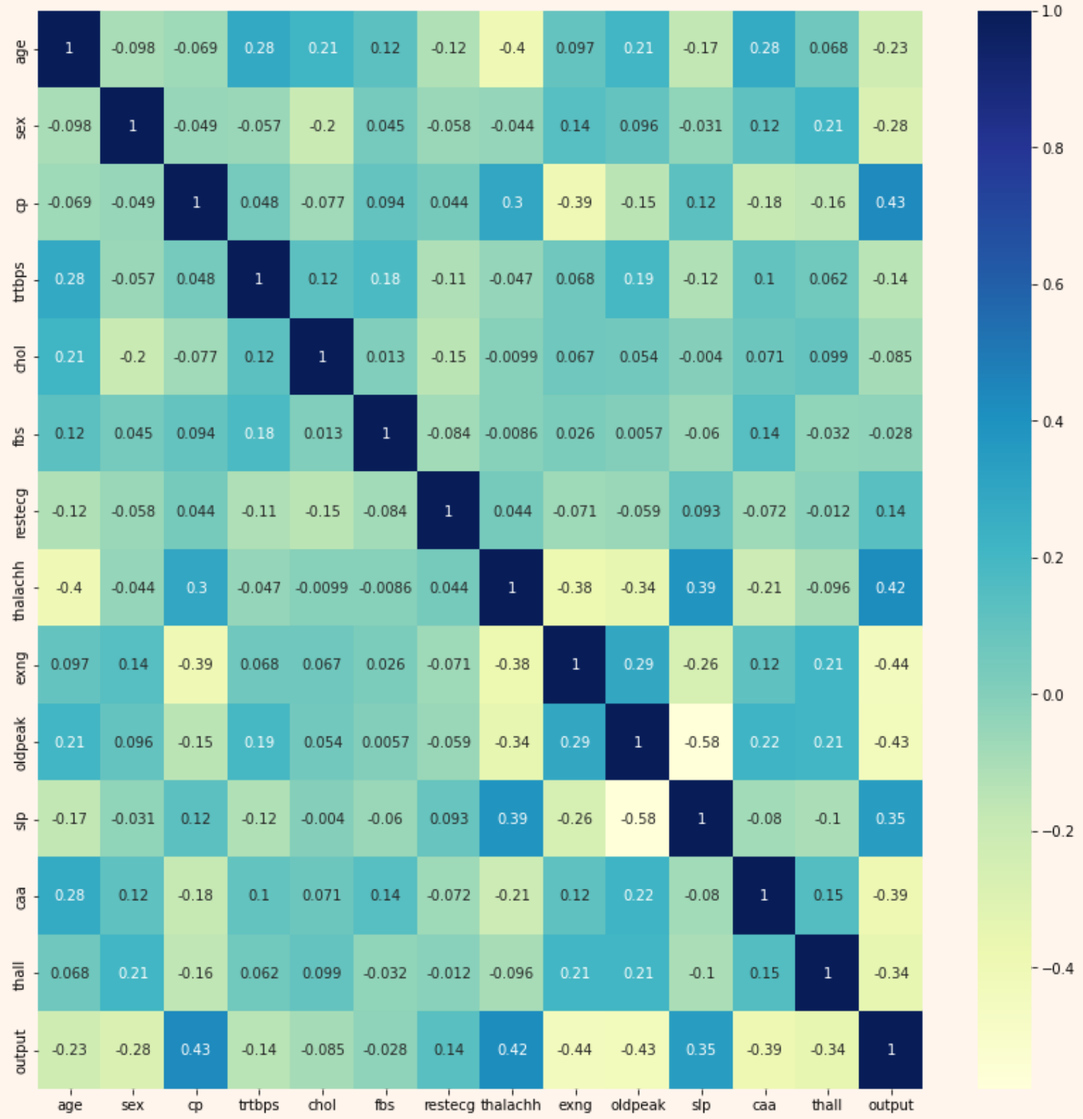
thalachh 변수의 histogram



thalachh 변수의 boxplot

3*Z 이상/이하의 결측치 제거 후, 동변수 boxplot





Slope과 peak의 correlation이 높고,
심근 경색과 관련있는 동시에 feature 끼리도
상관있는 변수들이 있음을 추론 가능

cp, exang, oldpeak, thalachh가
각각 output 과 상관관계 0.43, 0.44, 0.43,
0.42씩 존재함을 볼 수 있음.

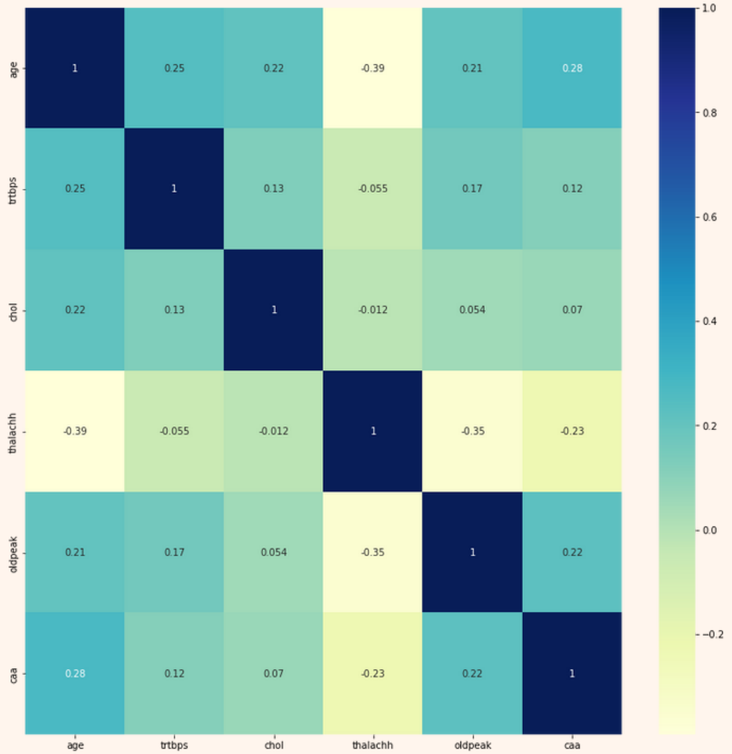


Figure 1 : 연속형 변수간 상관관계 plot
=> 상관계수 크게 유의미 X

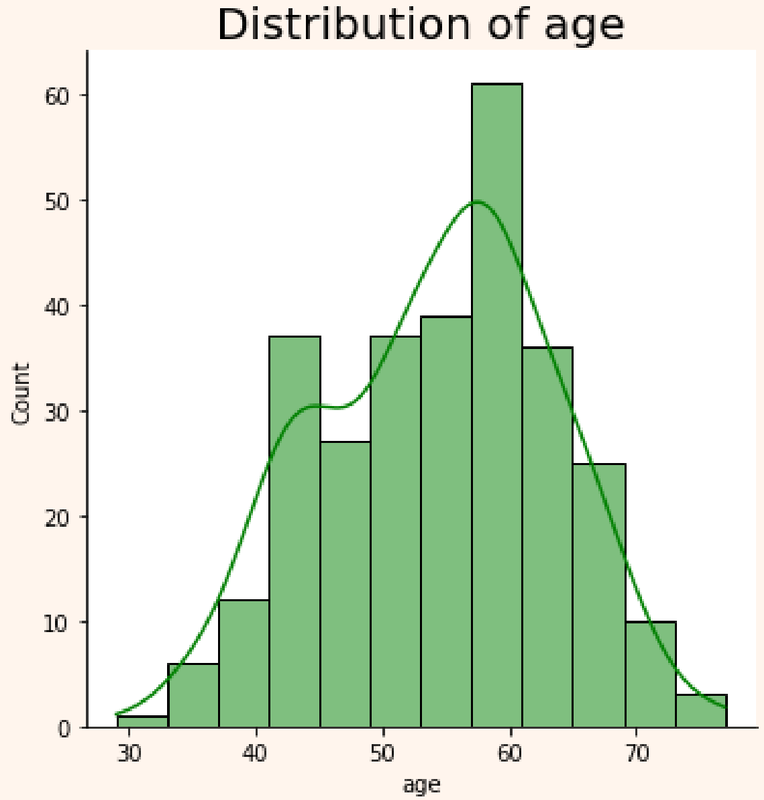


Figure 2 : 나이 변수 분포

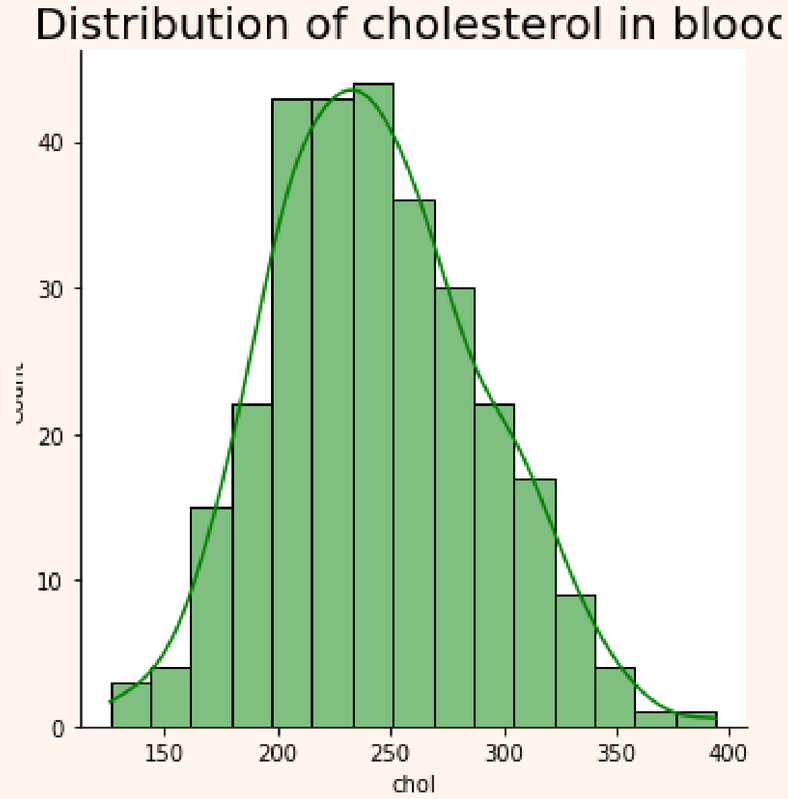


Figure 3 : 혈중 콜레스테롤 농도 분포

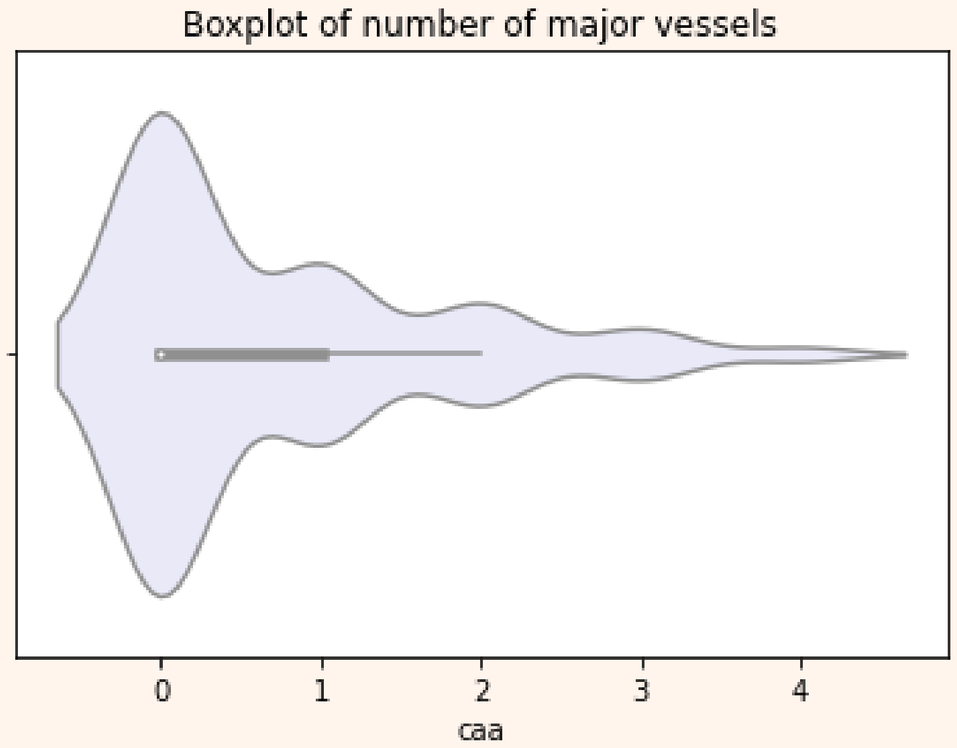


Figure 4 : Major Vessels Violin Plot

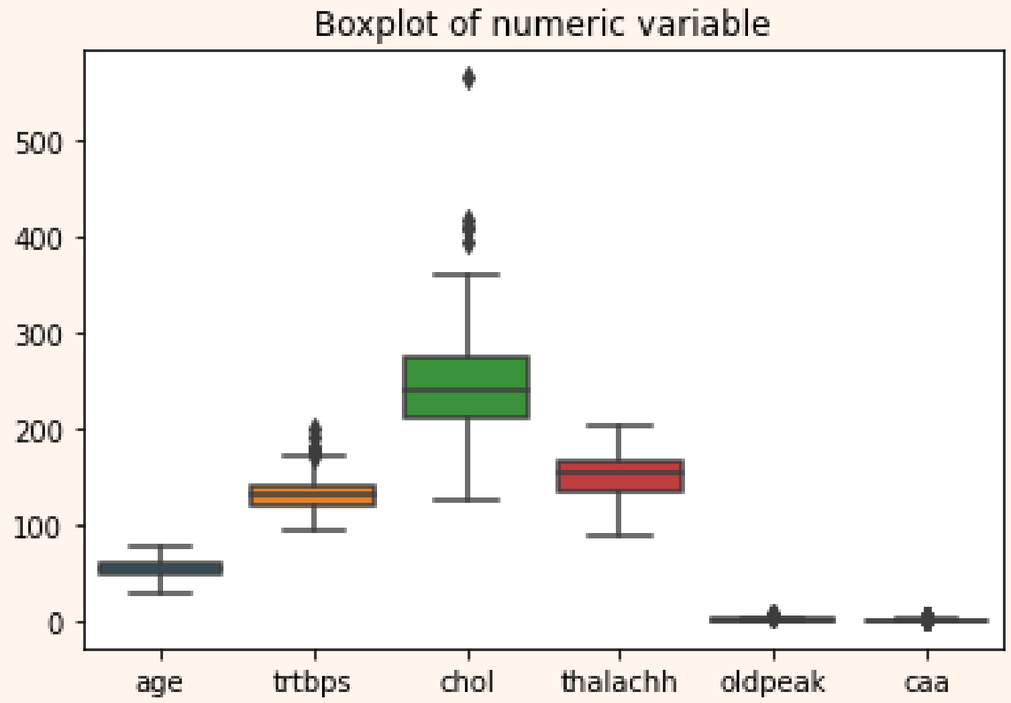


Figure 5: 수치형 변수들의 boxplots

outliers 확인 가능, 앞선 기준으로 제거 완료

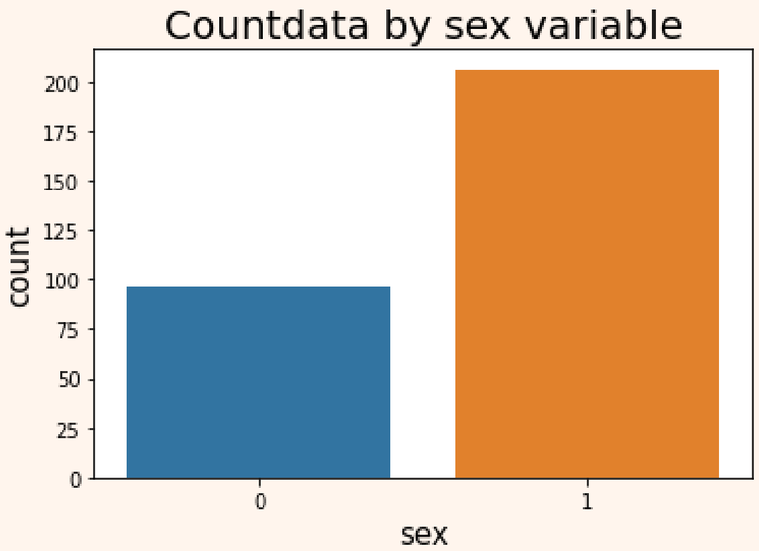


Figure 6: 성별 countplot

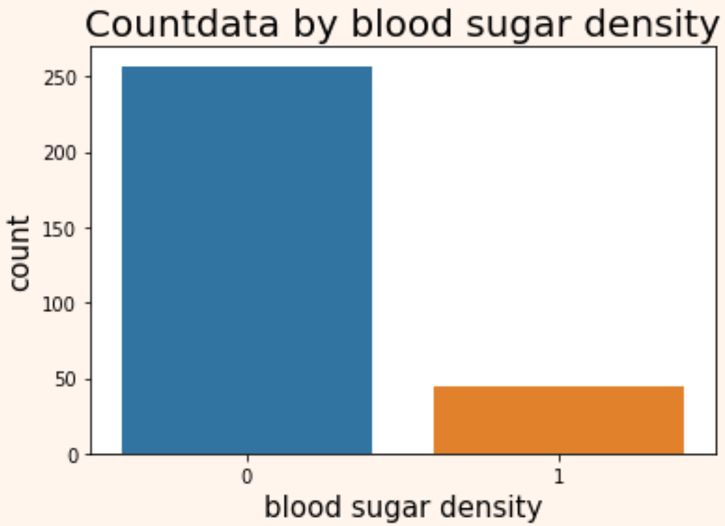


Figure 7 : 혈당 countplot

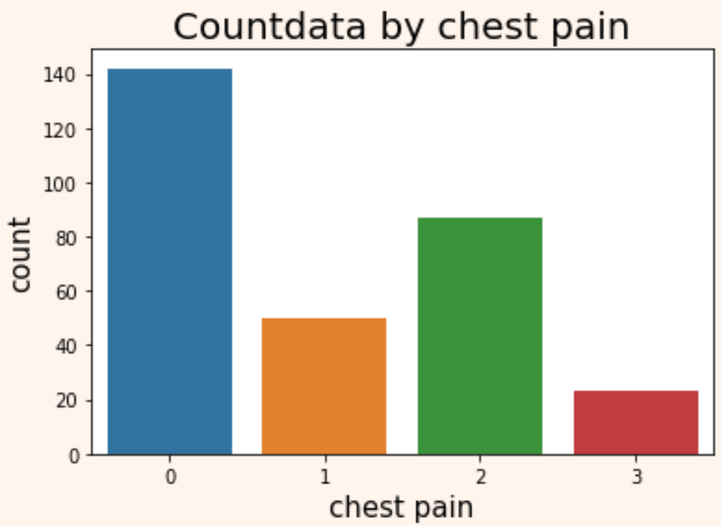


Figure 8: chest pain countplot

다소 불균형

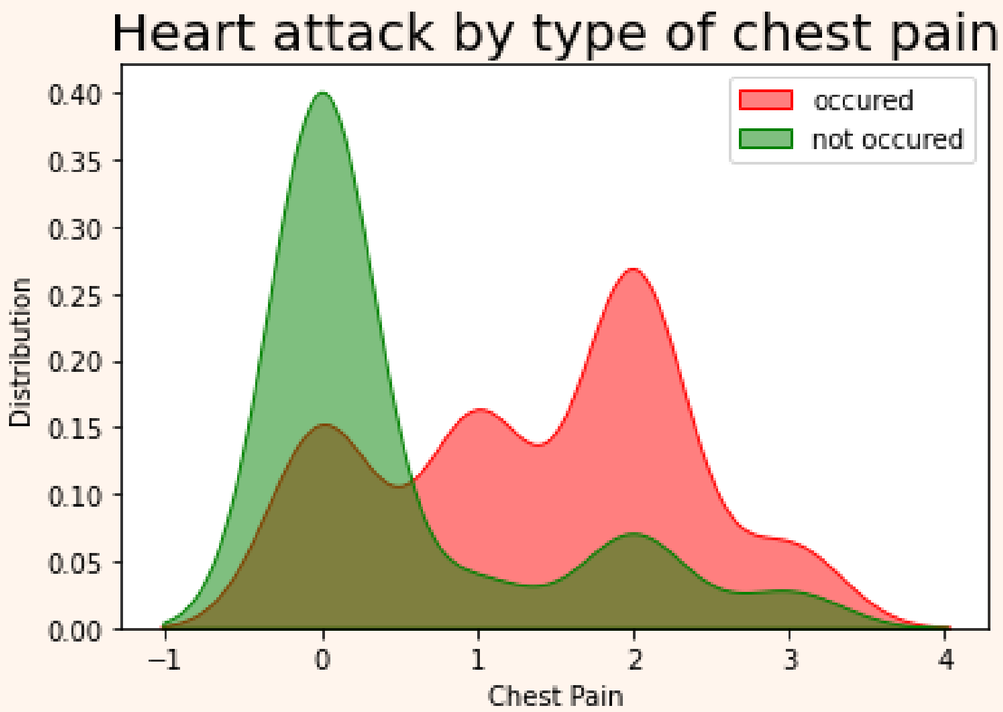


Figure 9: Heart attack by type of chest paint

cp=2 인 사람들이 심장마비를
경험할 확률이 더 높음

Figure 10: Heart attack by ages

나이와 심장마비 경험 확률이 더 높은 것은
양의 상관관계가 없어보인다.

분류 모델

훈련, 실험데이터 나누기

Training & Testing split는 8:2
(test size = 20%)

분류모델

5가지 분류 모델 사용 (PCA O,X)

1. Logistic Regression
2. Decision Tree
3. Naive Bayes
4. Random Forest Classifier
5. SVM Classifier

iteration 사용 이유

- 각 알고리즘별 PCA이전과 PCA이후의 성능비교
  -> 동일한 train/test set 필요
- 그러나 train_test_split 메소드는 train set과 test set을 랜덤하게 스플릿
- train_test_split 메소드 이전에 PCA를 해야하는
  PCA의 전처리적인 성격 때문에 original data와
  pca이후의 train/test set이 다르게 됨
- 따라서, 성능 비교를 위해 train/test set을 여러개 만들어서,
  각각 set마다 알고리즘별로 나온 성능을 평균값을 내어
  수렴하는 근사치를 비교하는 방법을 사용했음.

5가지 분류 모델 정확도

	Non pca(100)	Non pca(1000)	Pca(100)	Pca(1000)
Logistic Regression	84.6%	84.9%	83.9%	84.9%
Decision Tree	74.9%	75.4%	74.4%	74%
Naive Bayes	78.5%	81.3%	77.5%	81.7%
Random Forest Classifier	83.3%	80.2%	82.3%	79.9%
SVM Classifier	83.7%	83.3%	82.9%	83.7%

Logistic Regression

모든 경우의 수에서, 가장 성능이 좋은 모델

결과 요약 1

PCA 적용 이후,
Logistic Regression > SVM Classifier(Linear Kernel) > Naive Bayes > Random Forest Classifier > Decision Tress
순으로 성능이 좋음

결과 요약 2

PCA 사용은 정확도 향상에 전반적으로 긍정적인 결과 유도
(특히 naive bayes의 경우 성능 6.2% 개선
그러나, Random Forest Classifier의 경우 성능이 3.3% 감소)
=> PCA가 언제나 더 좋은 결과만 주는 것이 아님

결과 요약 3

PCA 적용 이후, Logistic Regression이
PCA 이전보다 1% 성능 개선 되었음

결과 요약 4

결론적으로, PCA 사용은 overfitting 방지와
성능 개선 둘 다 가능하게 하였음

결론

심장마비 예측을 위해서는 PCA전처리를 한
로지스틱 회귀분석 알고리즘이 정확도 85%로
제일 정확도가 높게 나왔음.

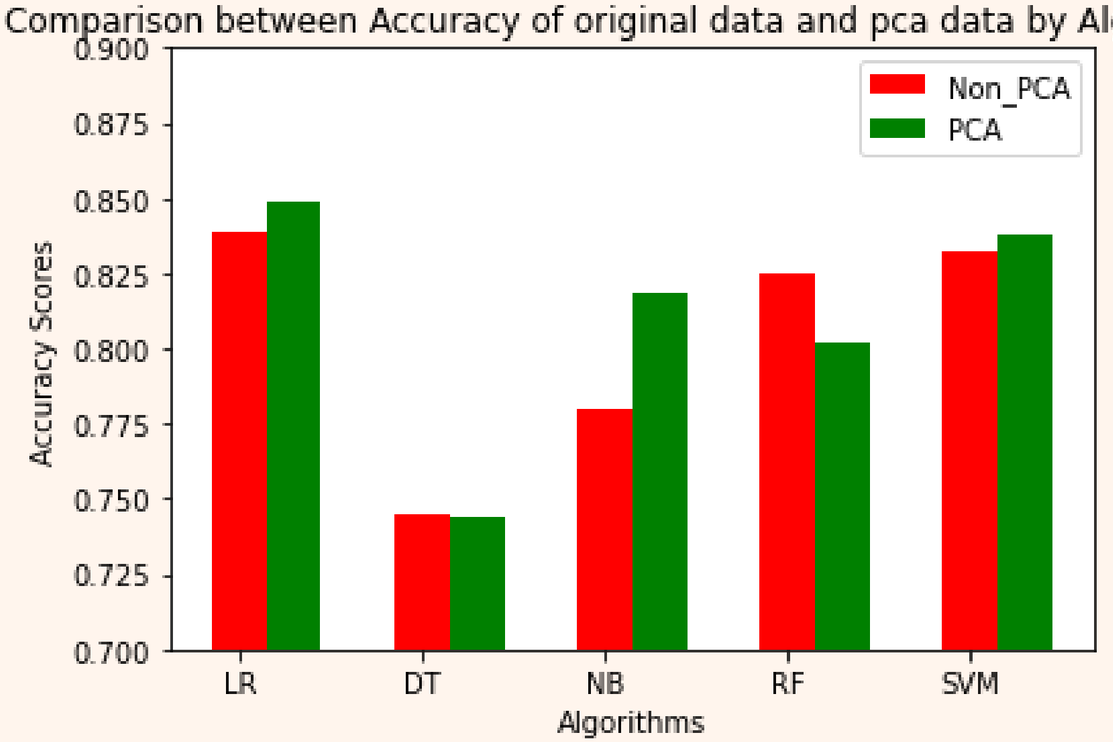


Figure 11: Comparison between Accuracy of original data and pca data by Algorithms

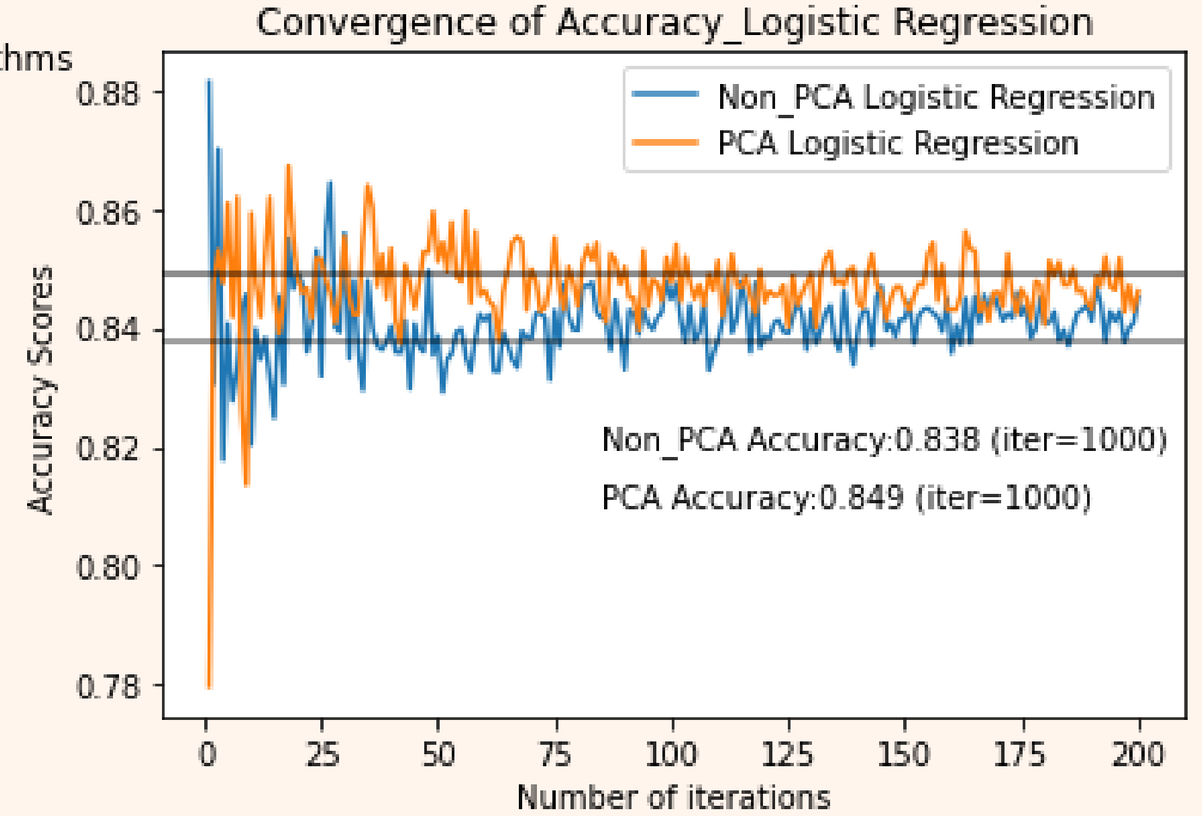


Figure 12: Convergence of Accuracy - Logistic Regression

- Accuracy Scores : 각 iteration number까지 누적된 성능들의 평균값
- Iteration수가 많아질수록 분산이 줄어들고 특정한 값에 수렴
- 로지스틱 회귀분석의 경우 PCA를 사용할 때, 정확도 및 성능이 더 좋음

감사합니다!