피마 인디언 당뇨병 예측 모델

20181110493 정정룡

목차

Part 1

EDA

변수

변수 이름	변수 설명
Pregnancies	임신 횟수
Glucose [NA]	2시간 동안의 경구포도당 내성 검사에서 혈장 포도당 농도
BloodPressure [NA]	이완기 혈압(mm Hg)
SkinThickness [NA]	팔 삼두근 뒤쪽의 피하지방 측정값(mm)
Insulin [NA]	2시간 혈청 인슐린 수치(mm U/ml)
BMI [NA]	체질량 지수
DiabetesPedigreeFunction	당뇨 내력 가중치 값
Age	나이
Outcome	당뇨병 여부(0, 1)

Levene's test

- 정규성을 만족할 수 없어 Levene's test로 등분산 검정 진행

Shapiro_Wilks test

- 표본수가 2000미만이므로 Shapiro_Wilks test 로 정규성 검정 진행

- 독립성과 등분산성을 가정하고 진행

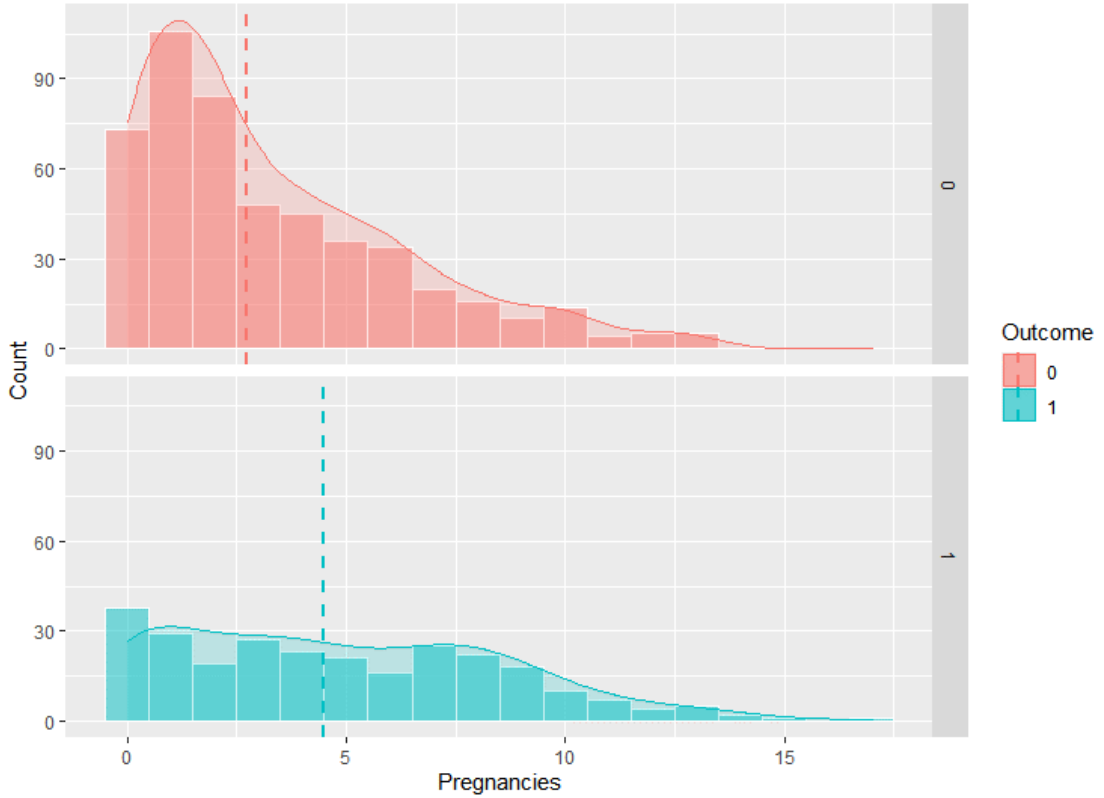
Student t-test

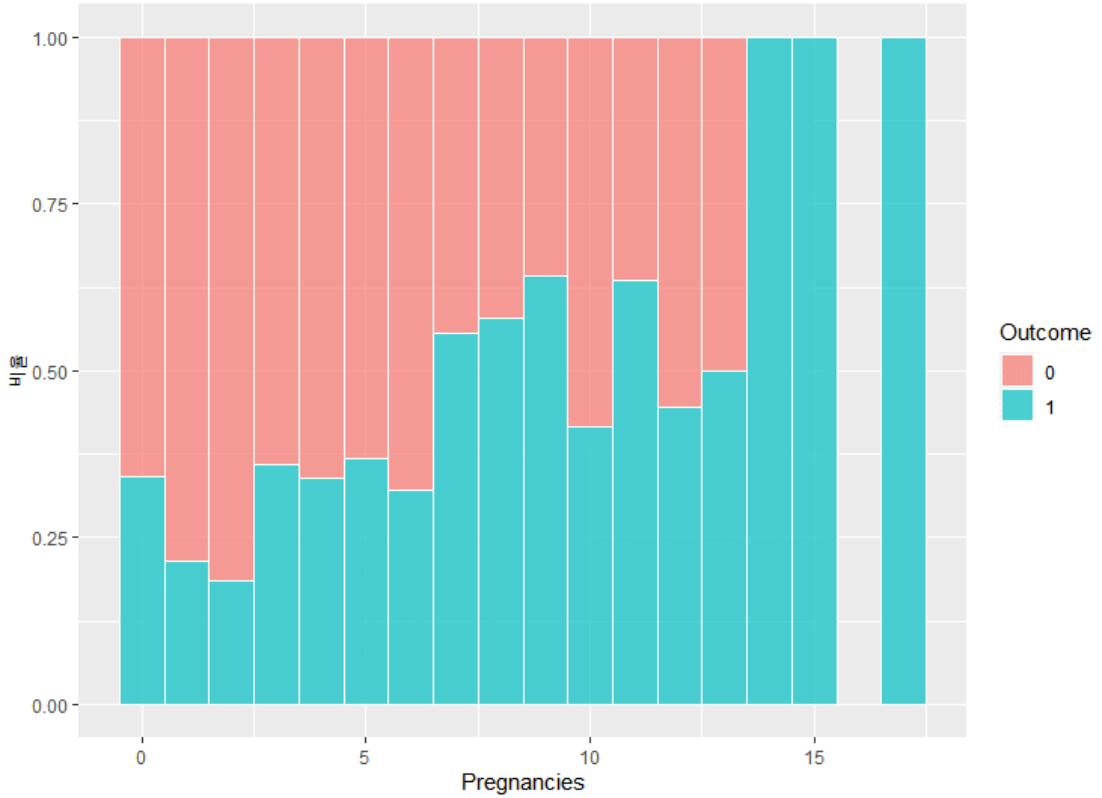
- 정규성
- 등분산성

Wilcoxon rank sum test

- 정규성 X

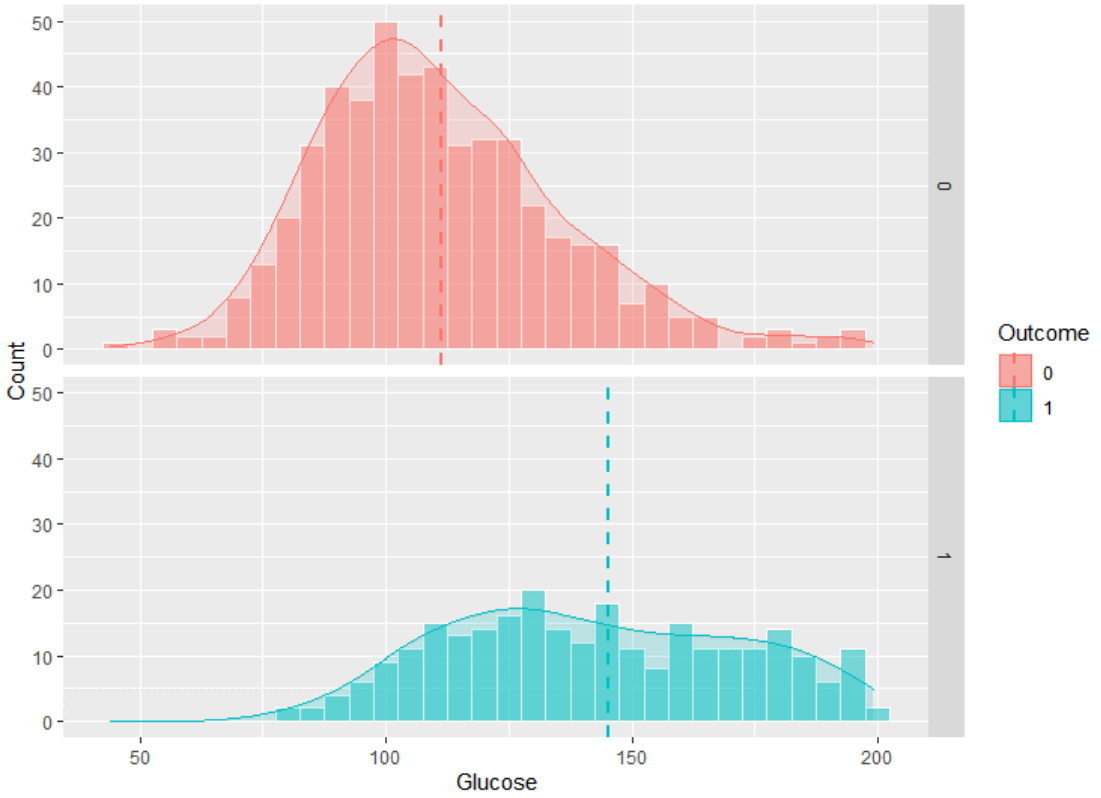
Pregnancies

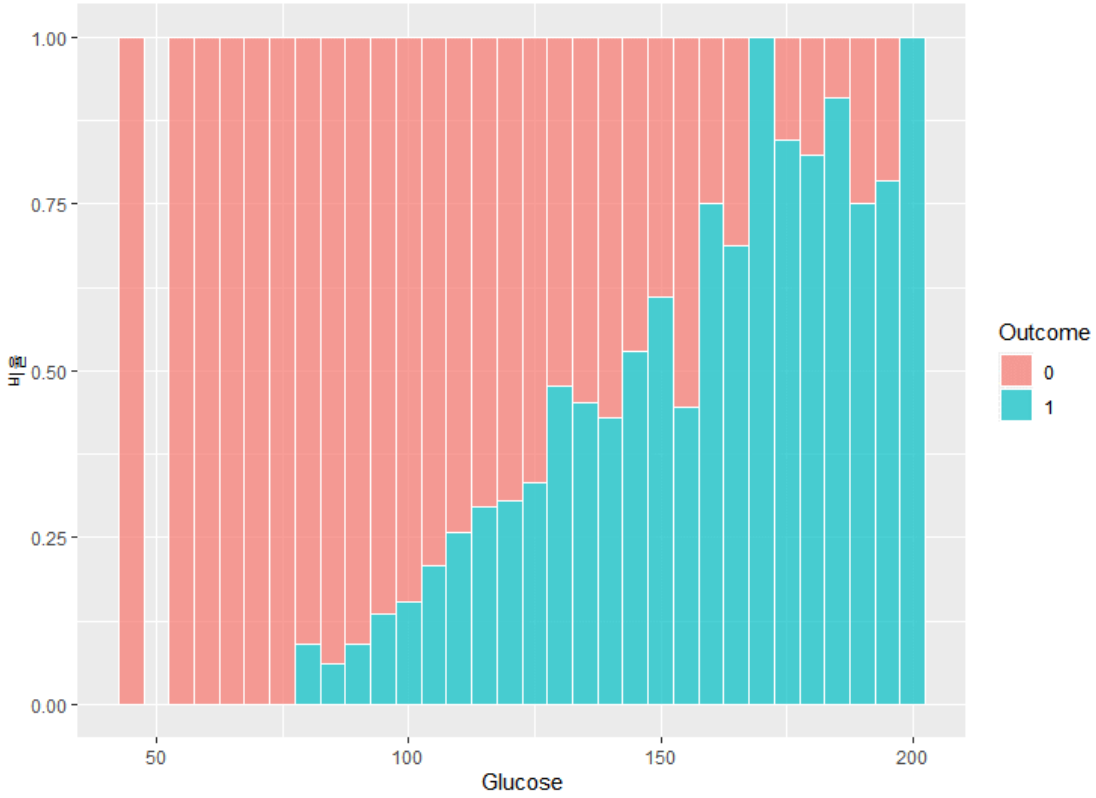




등분산 따르지 않으나, 두 집단 차이는 유의

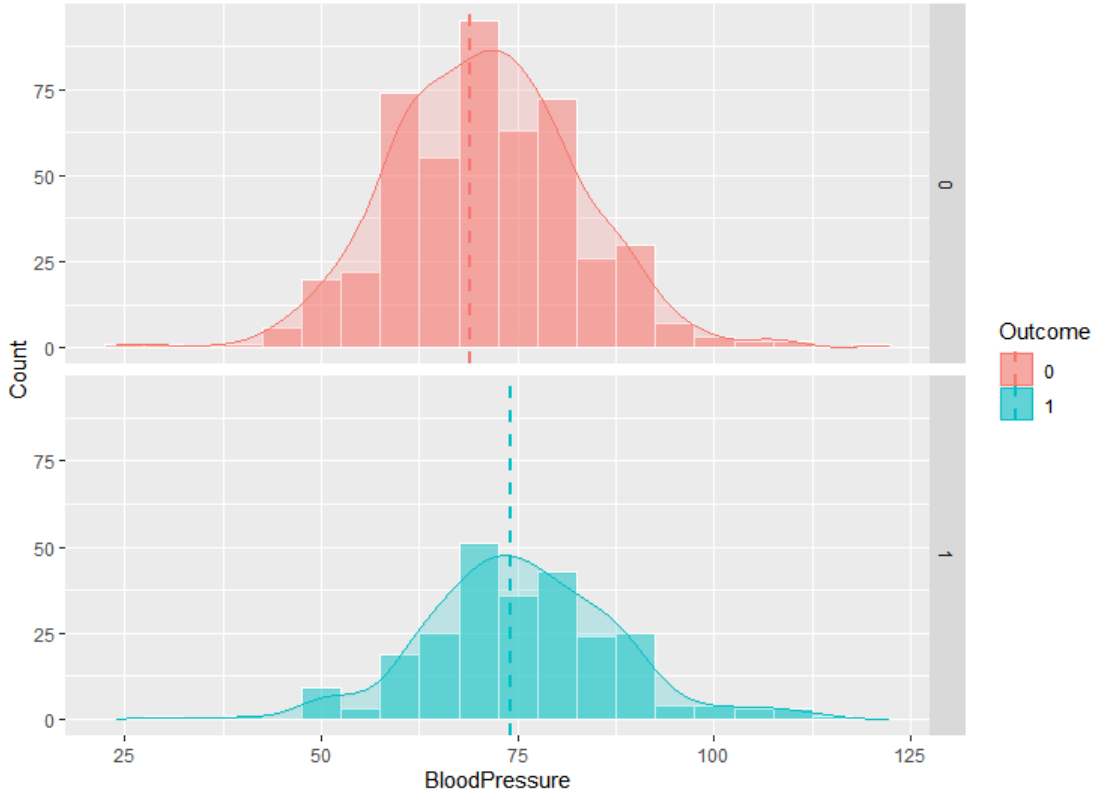
Glucose

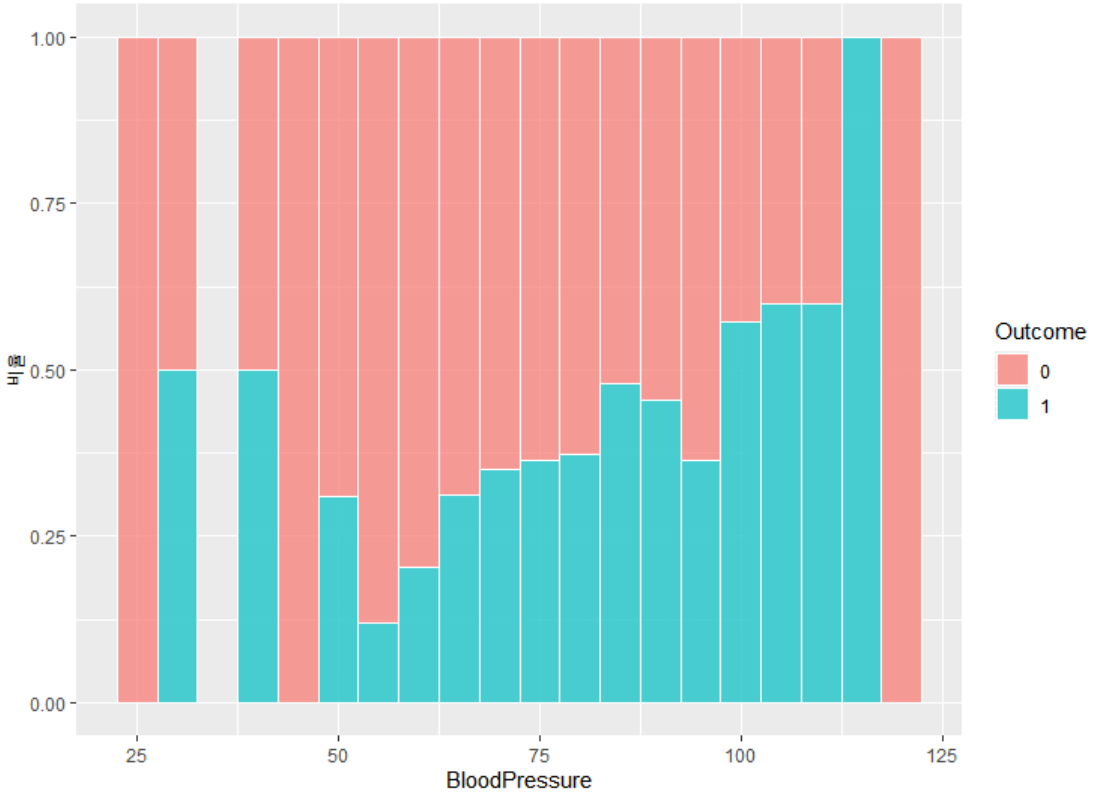




등분산 따르지 않으나, 두 집단 차이는 유의

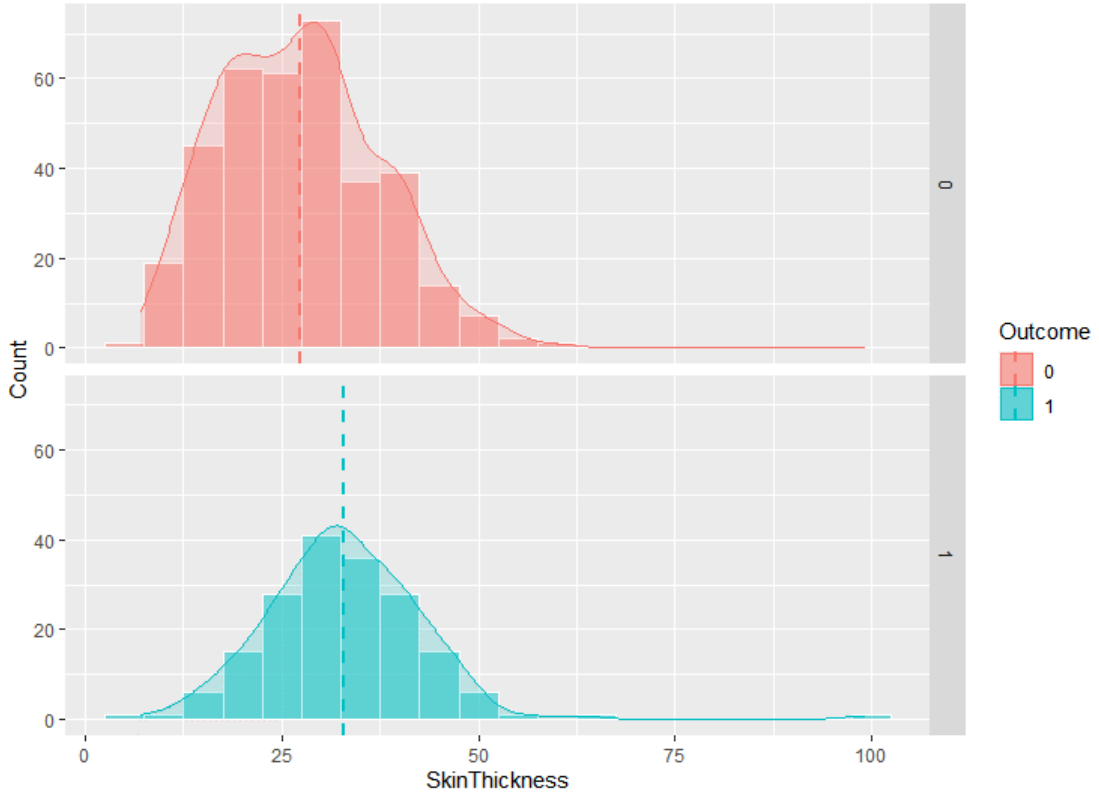
BloodPressure

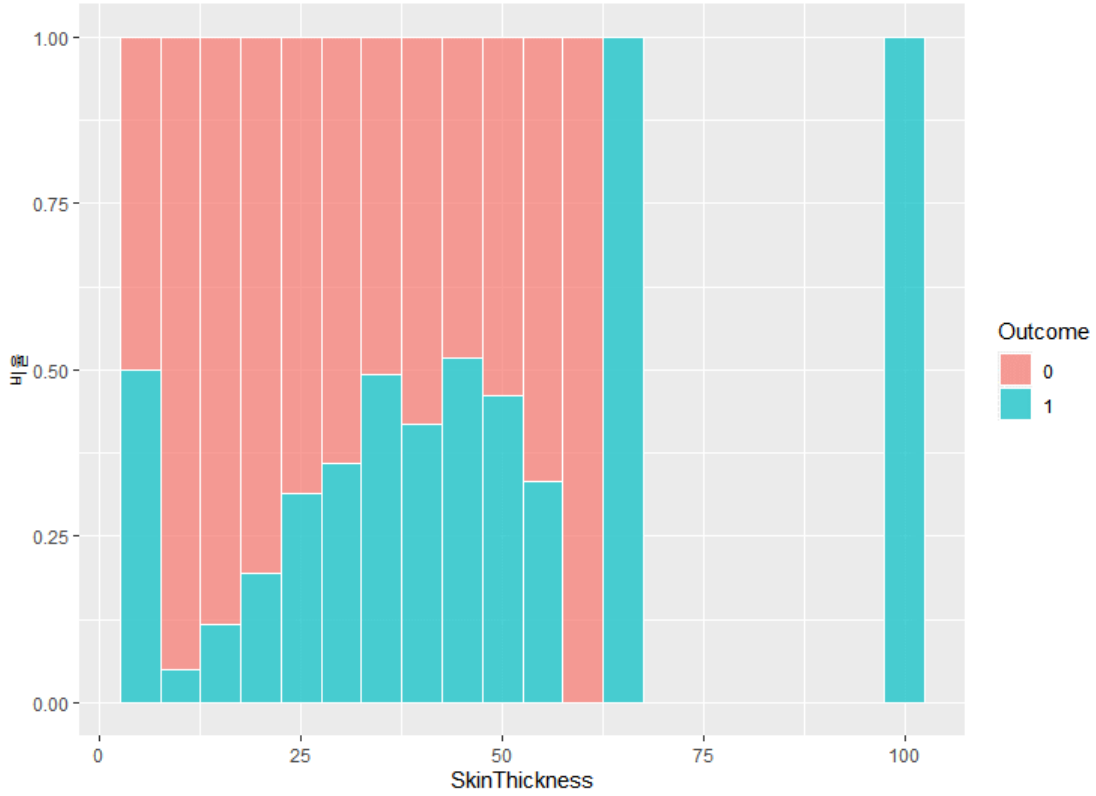




등분산, 정규성 따르며, 두 그룹 차이도 유의

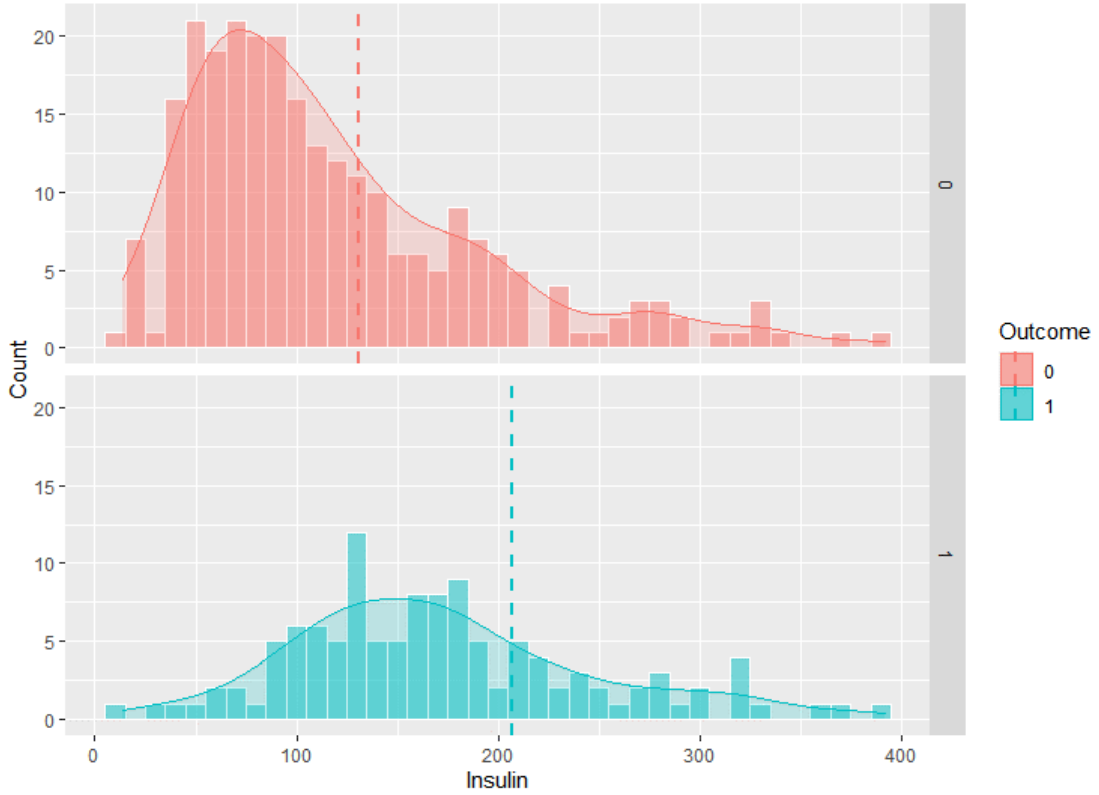
SkinThickness

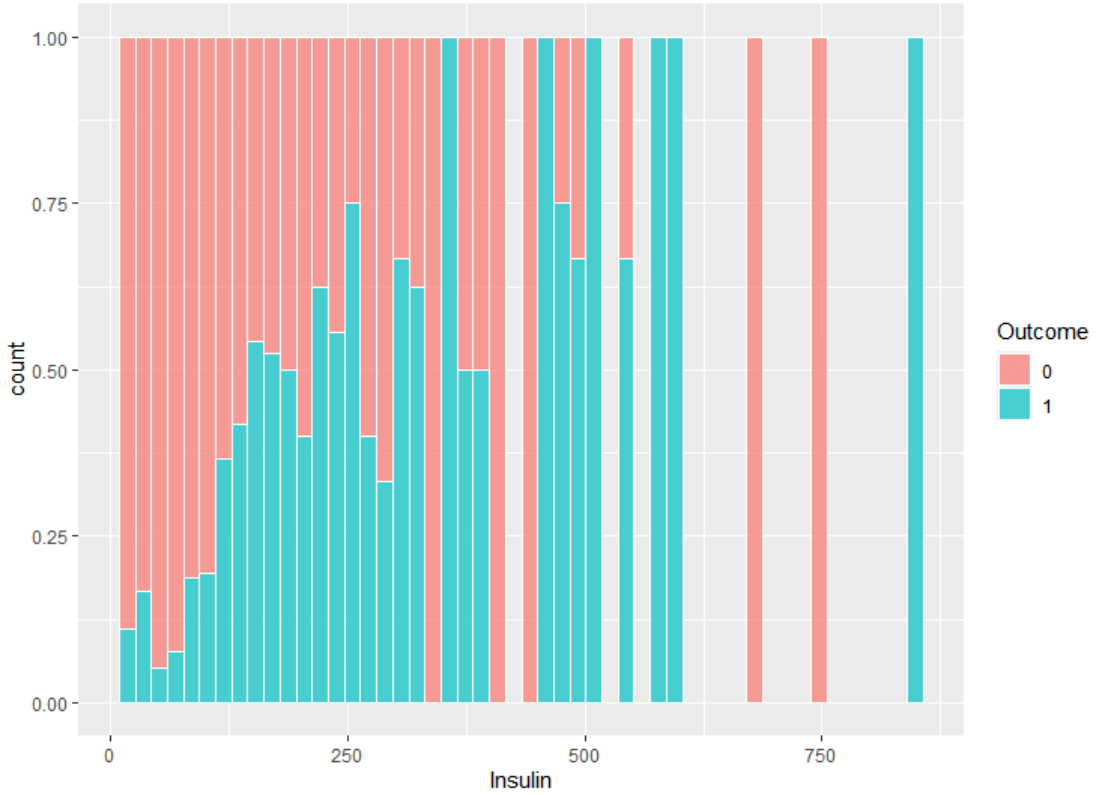




등분산 따르나, 정규성 따르지 않고, 두 그룹 차이는 유의

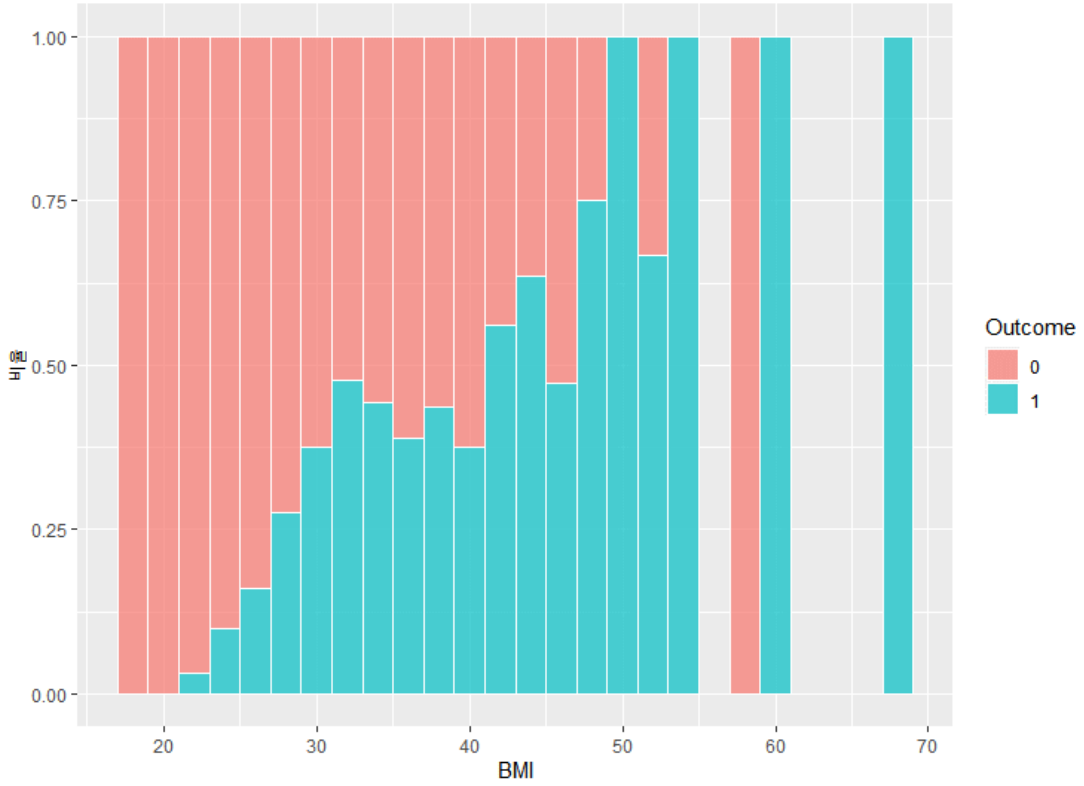
Insulin





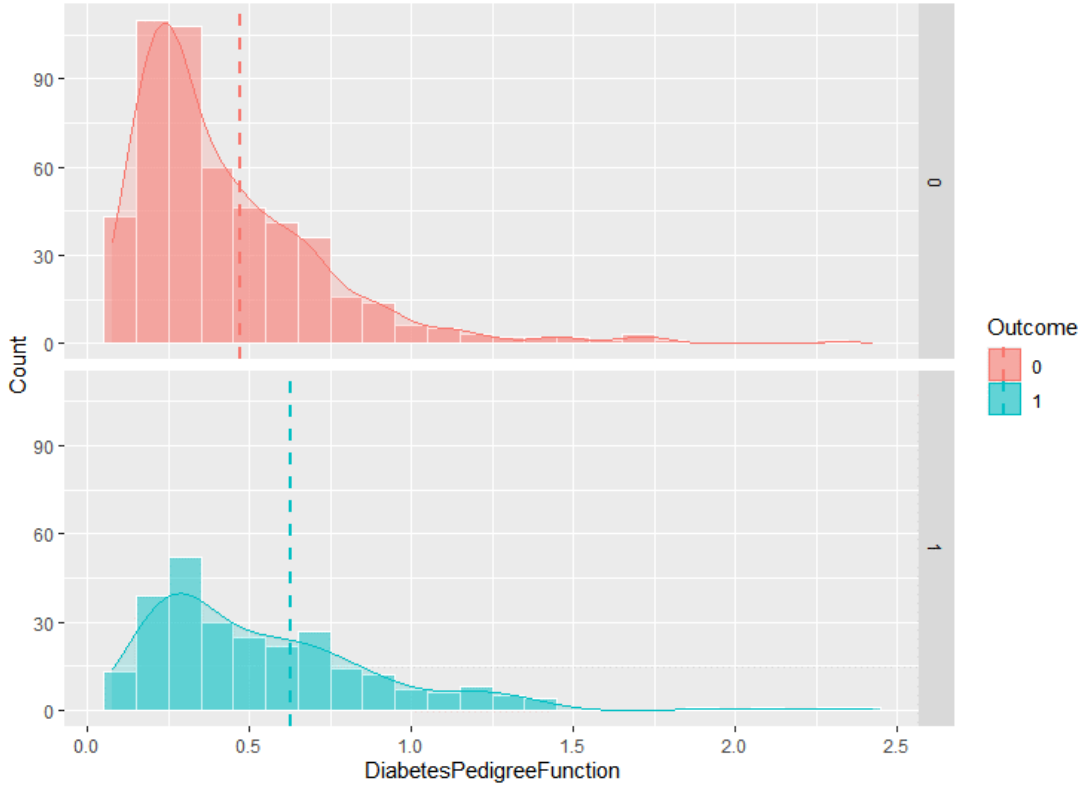
등분산 따르지 않으나, 두 그룹 차이는 유의

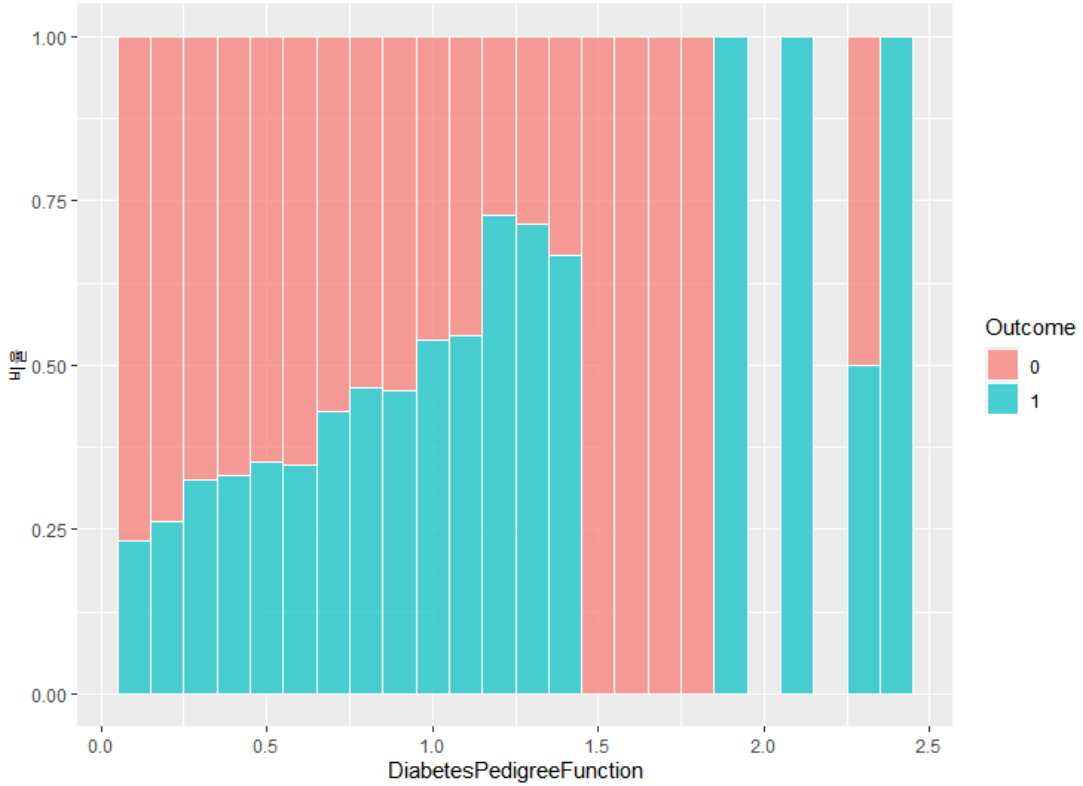
BMI



등분산 따르나, 정규성 따르지 않고, 두 그룹 차이는 유의

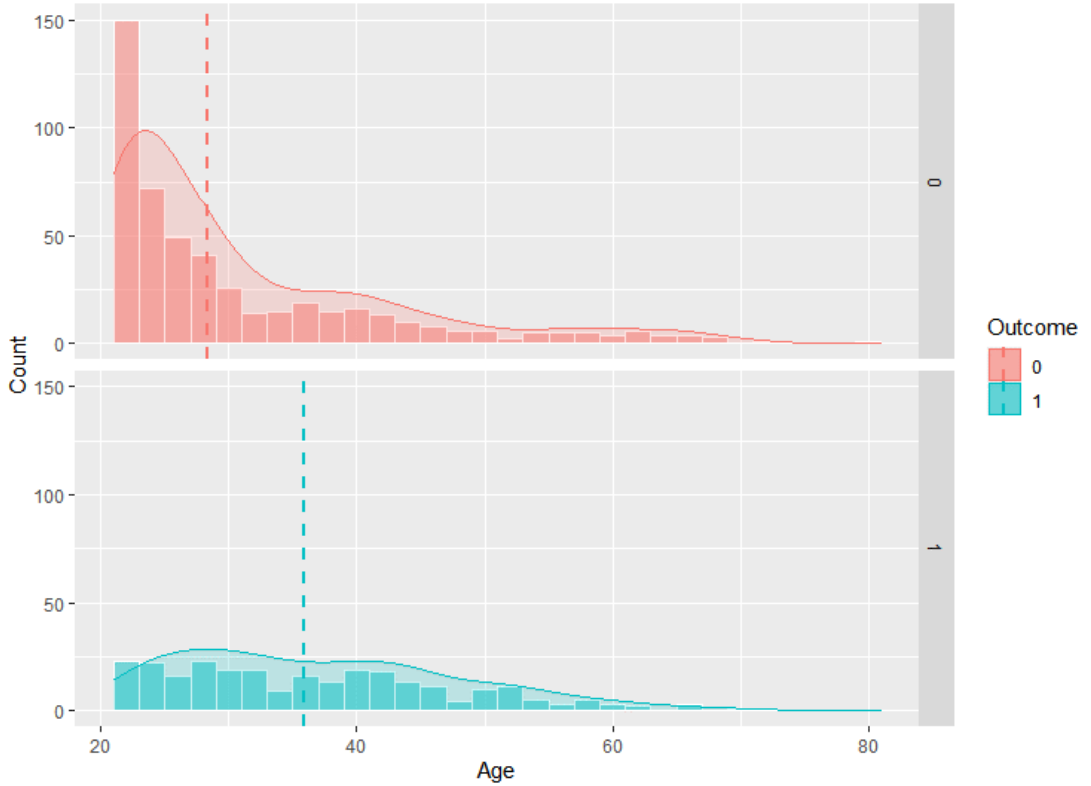
DiabetesPedigreeFunction





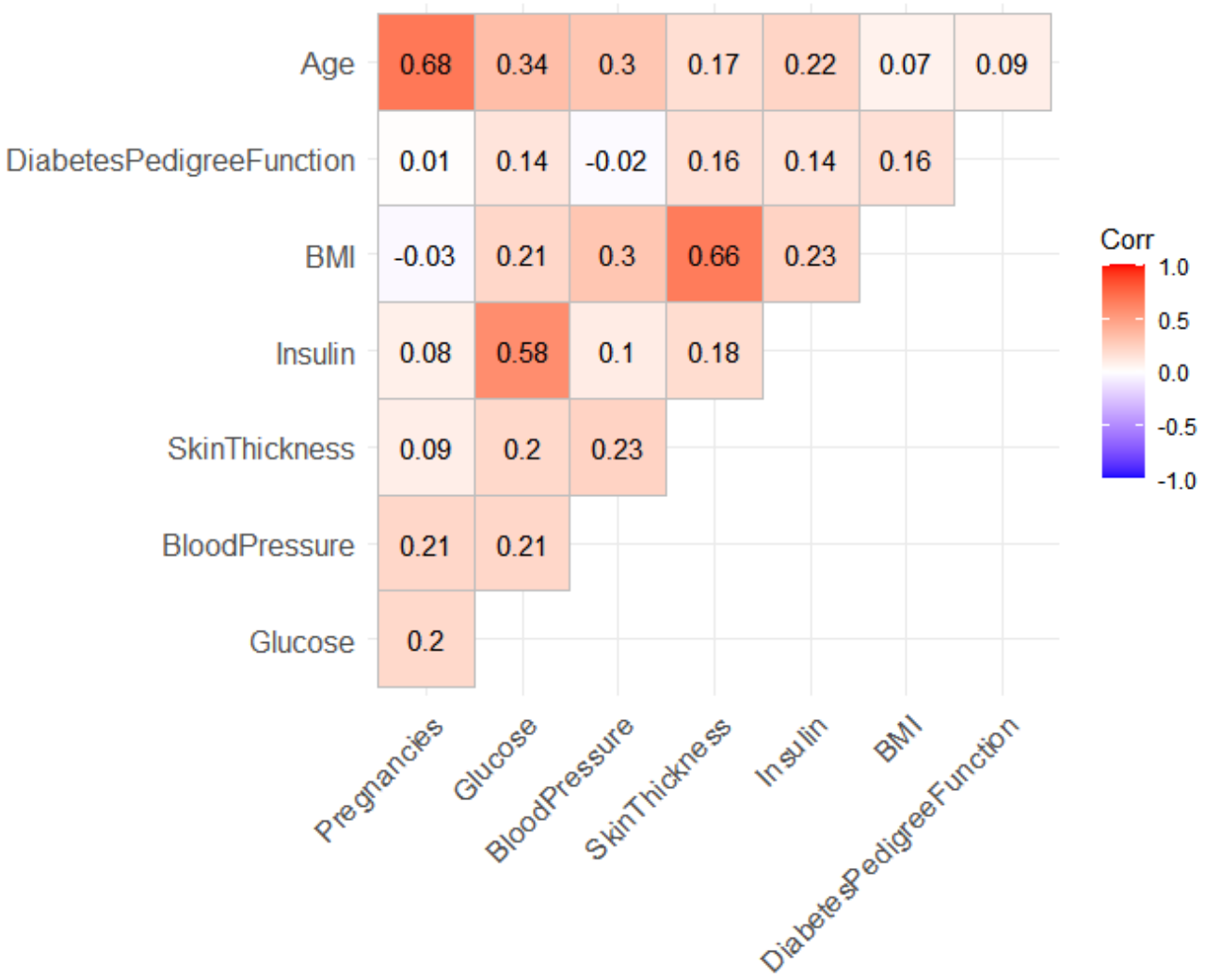
등분산 따르나, 정규성 따르지 않고, 두 그룹 차이는 유의

Age



등분산 따르지 않으나, 두 그룹 차이는 유의

Correlation

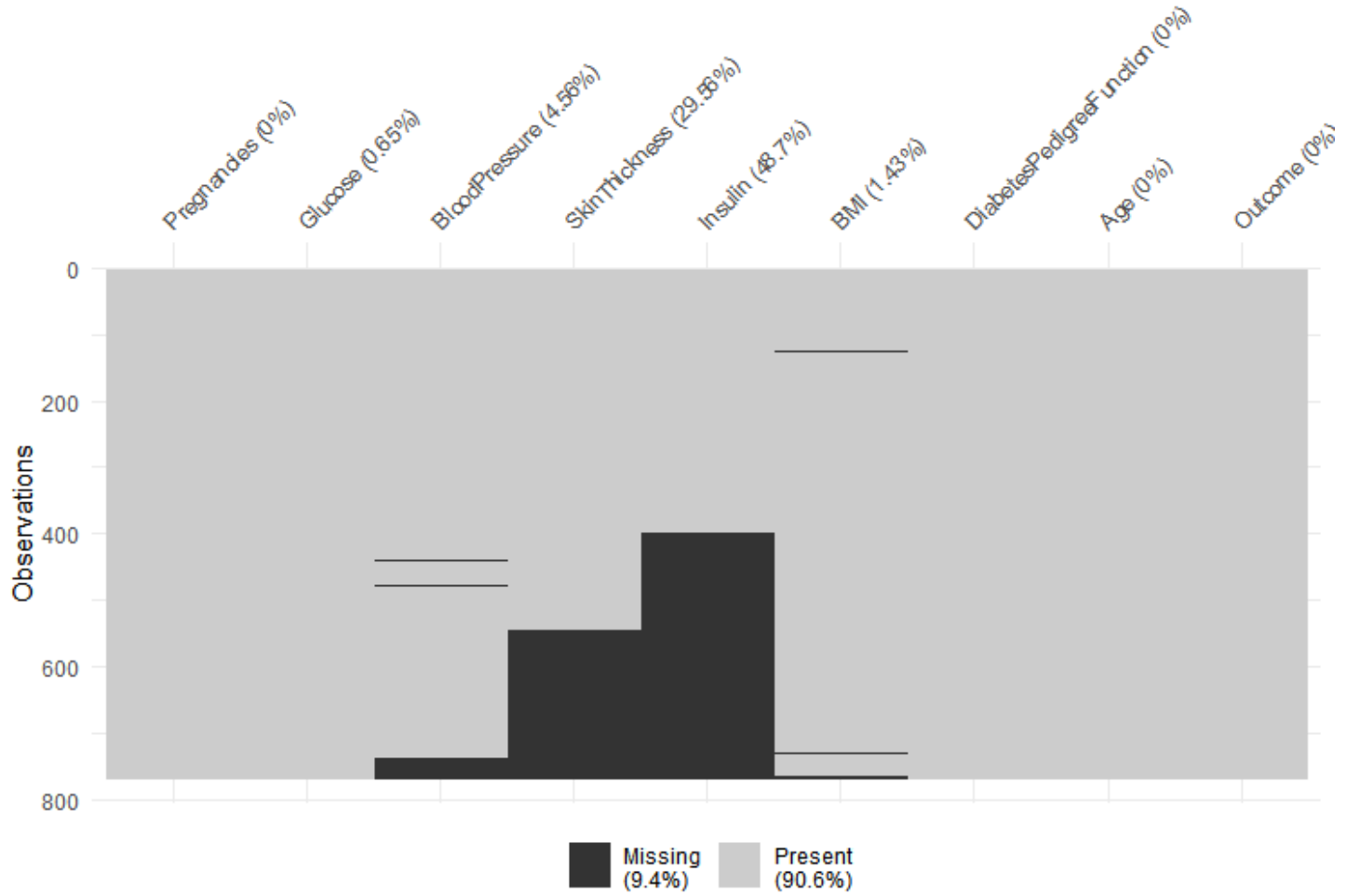


Age와 Pregnancies : 0.68

BMI와 SkinThickness : 0.66

Insulin과 Glucose : 0.58

결측치



Insulin : 374개

SkinThickness : 227개

BloodPressure : 35개

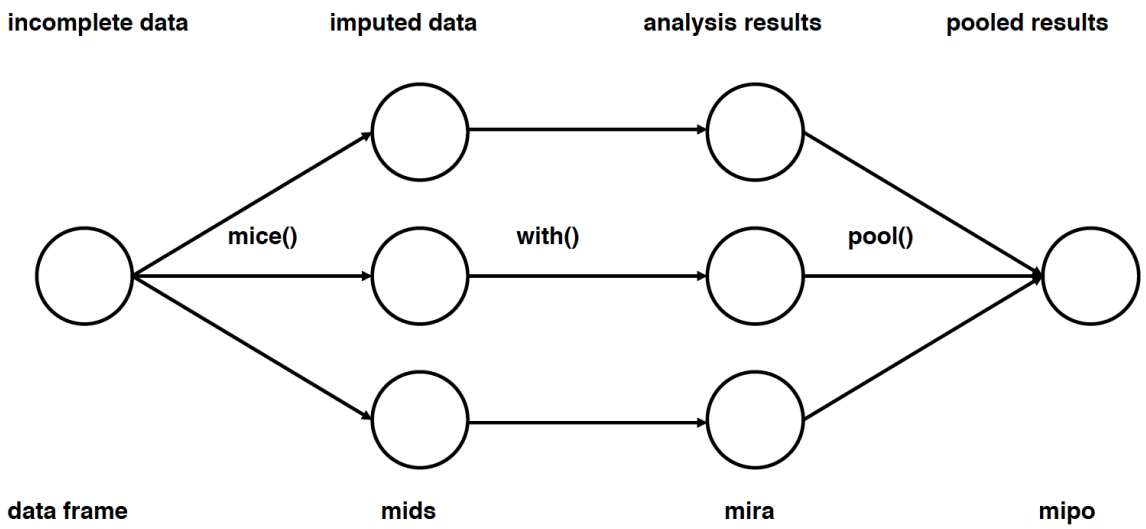
BMI : 11개

Glucose : 5개

결측치 처리 방법

결측 비율	처리 방법
10% 미만	제거 **or** 방법 상관없이 **Imputation**
10% 이상 **20%**미만	**Model based method, Regression**
20% 이상	**Model based method, Regression**

Mice



모든 변수들이 정규분포에 따른 다는 가정 없음

1. Fill-in

모든 변수의 결측치를 변수의 순서대로 채우며,

앞서 채워진 변수는 다음 채워지는 변수의 독립변수로 활용

2. Imputation

앞서 채워진 값들을 변수의 순서대로 대체하는 과정,

이 과정을 충분히 길게 하여 대체된 데이터셋에서 결측치가 독립적인 추출이 될 때까지 시행

Part 2

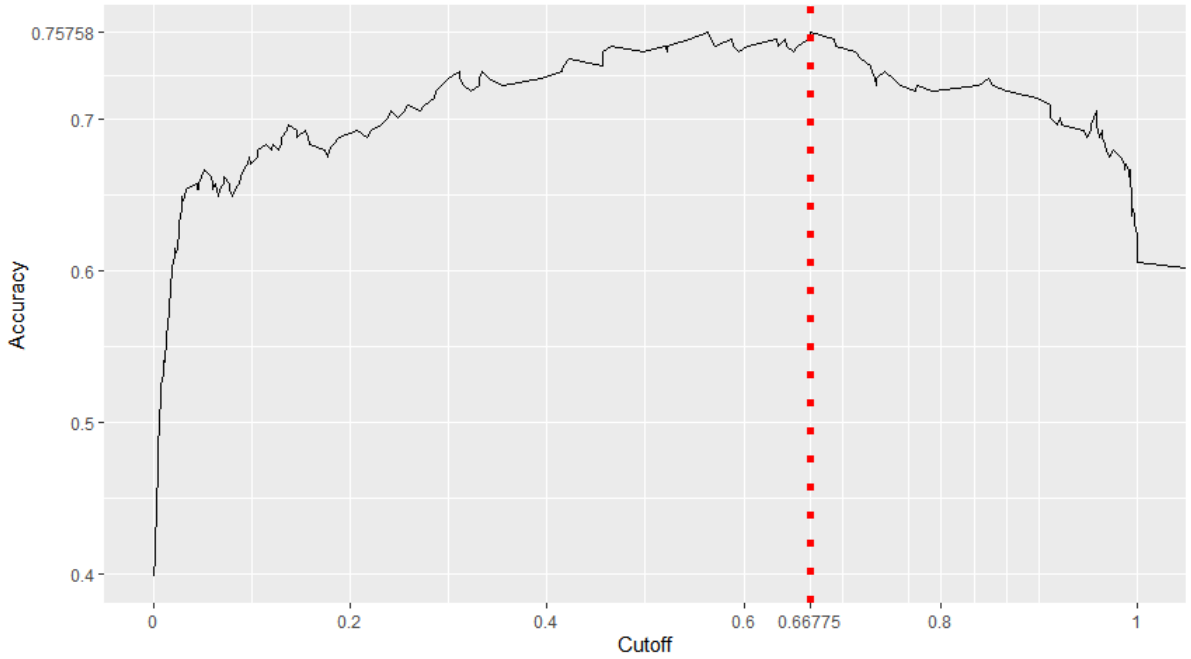
Modeling

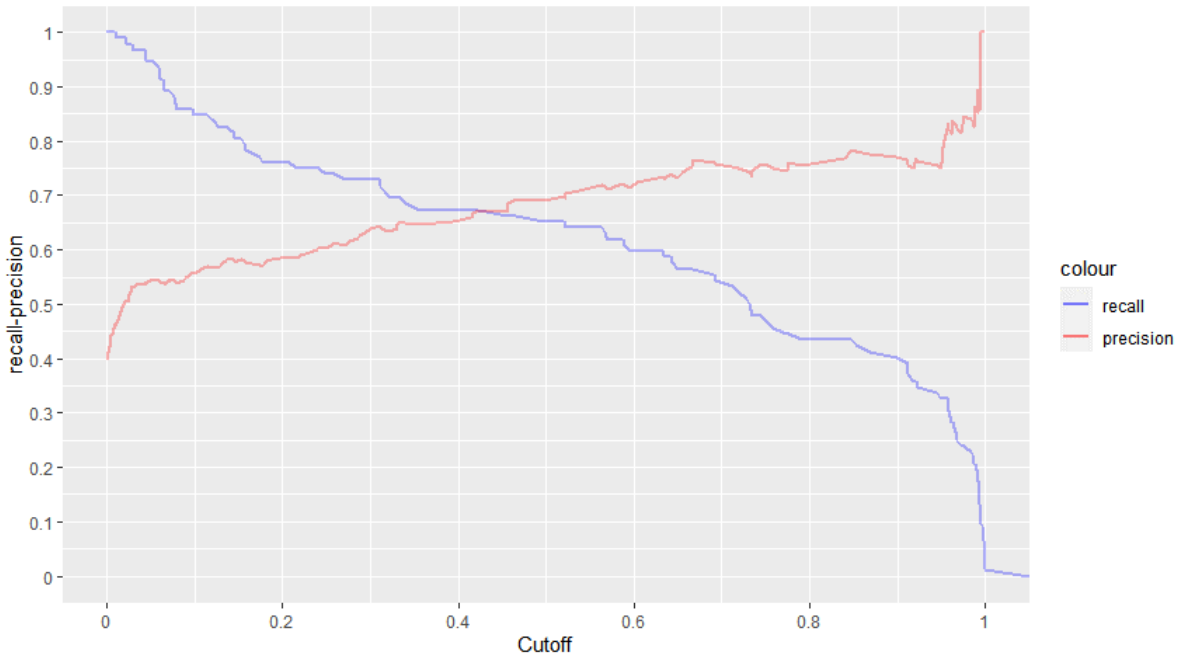
- Naïve Bayes, Logistic, Decision Tree
  KNN, SVM, RandomForest,
  XGB, Ensemble 이용

- 표준화 진행

- train: test=7:3 으로 진행

Naive Bayes

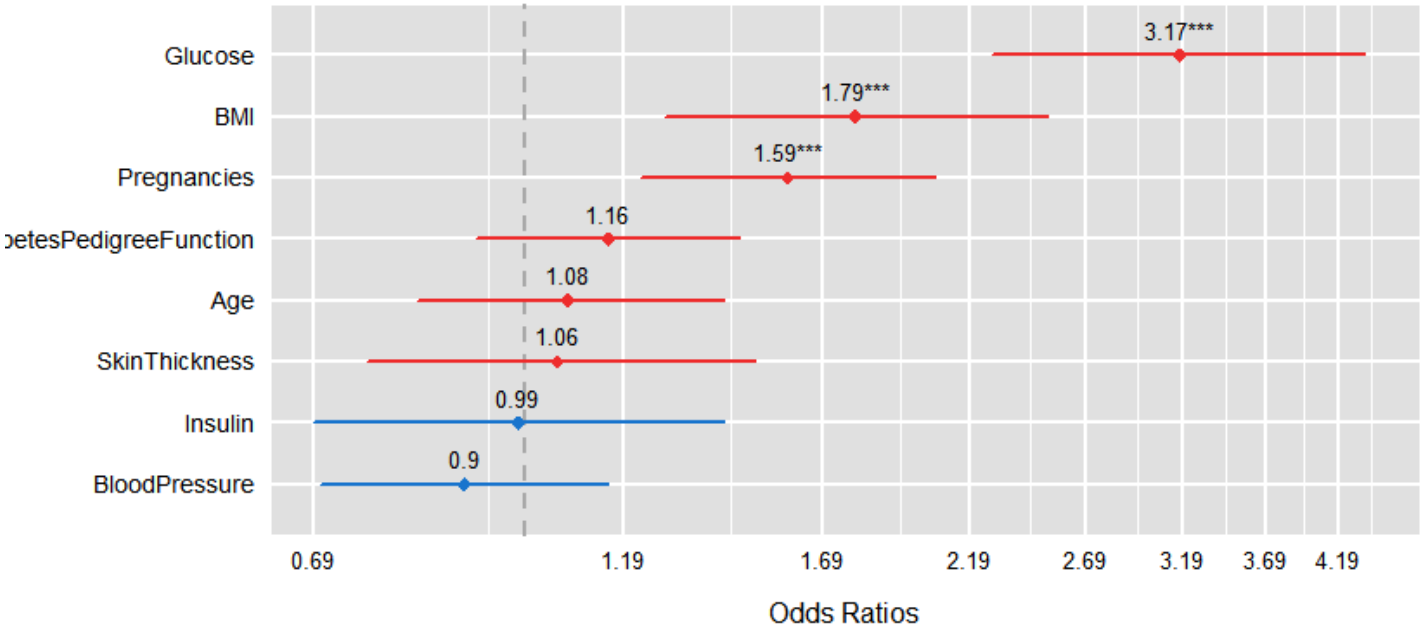




Cutoff 가 0.66775에서 Accuracy 0.75758

Logistic Regression

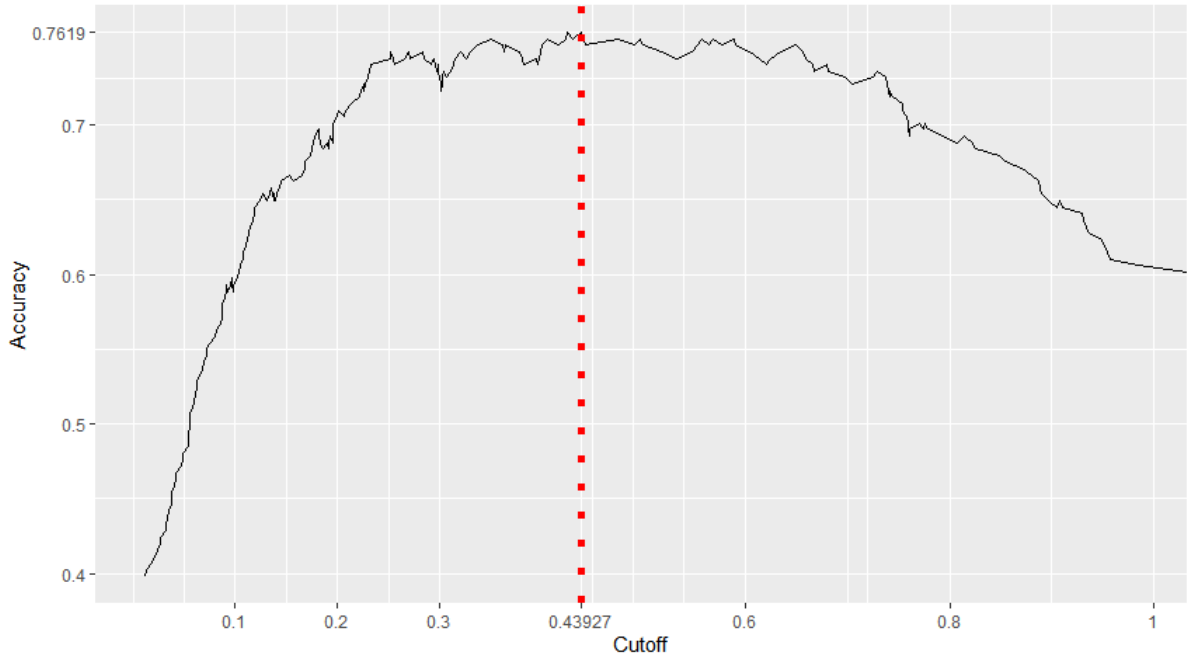
변수 이름	VIF
Pregnancies	1.441564
Glucose	1.716390
BloodPressure	1.232609
SkinThickness	1.817627
Insulin	1.745497
BMI	2.034133
DiabetesPedigreeFunction	1.018205
Age	1.575832

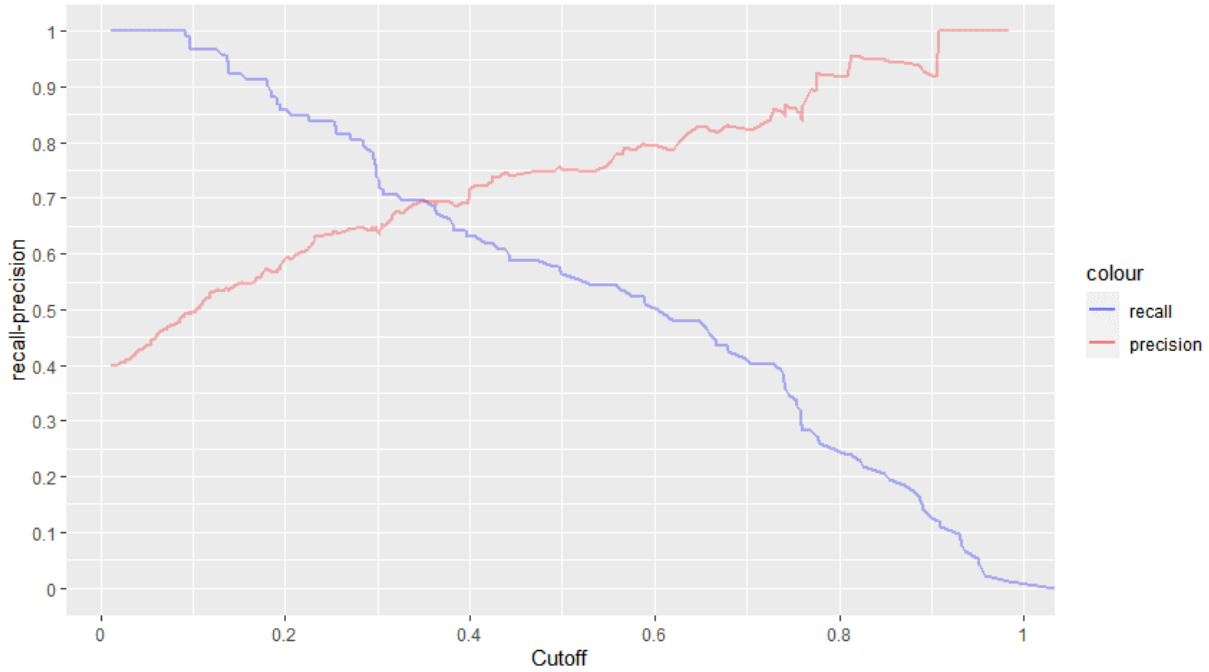


$$VIF = \frac{1}{1-R_i^2}$$

$$Odds_i = \exp(\boldsymbol{\beta}_i)$$

Logistic Regression



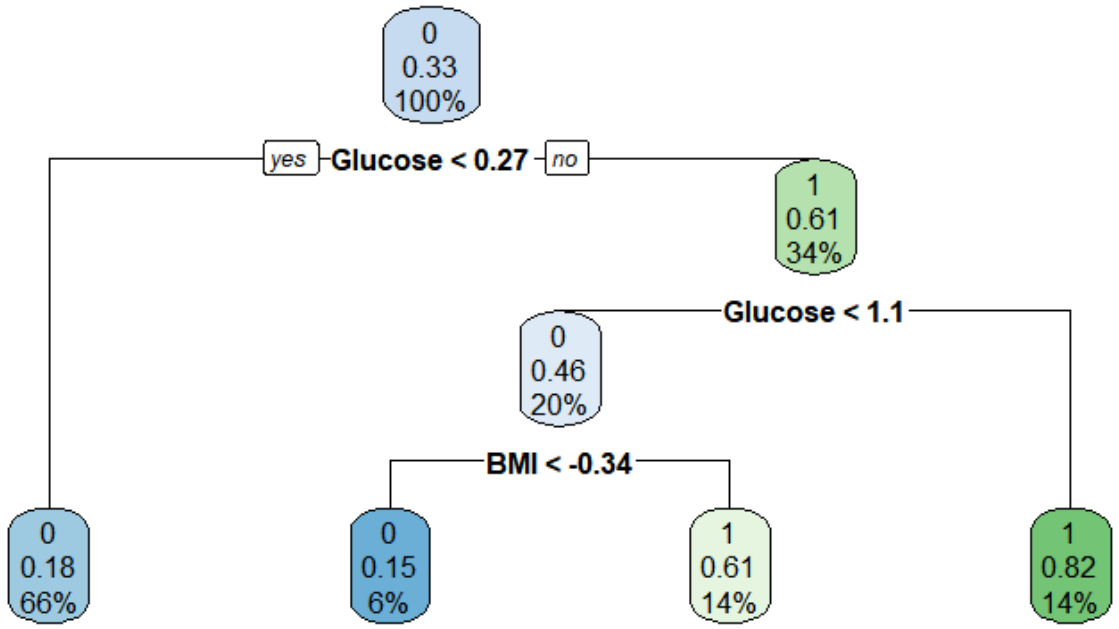


Cutoff 가 0.43927에서 Accuracy 0.7619

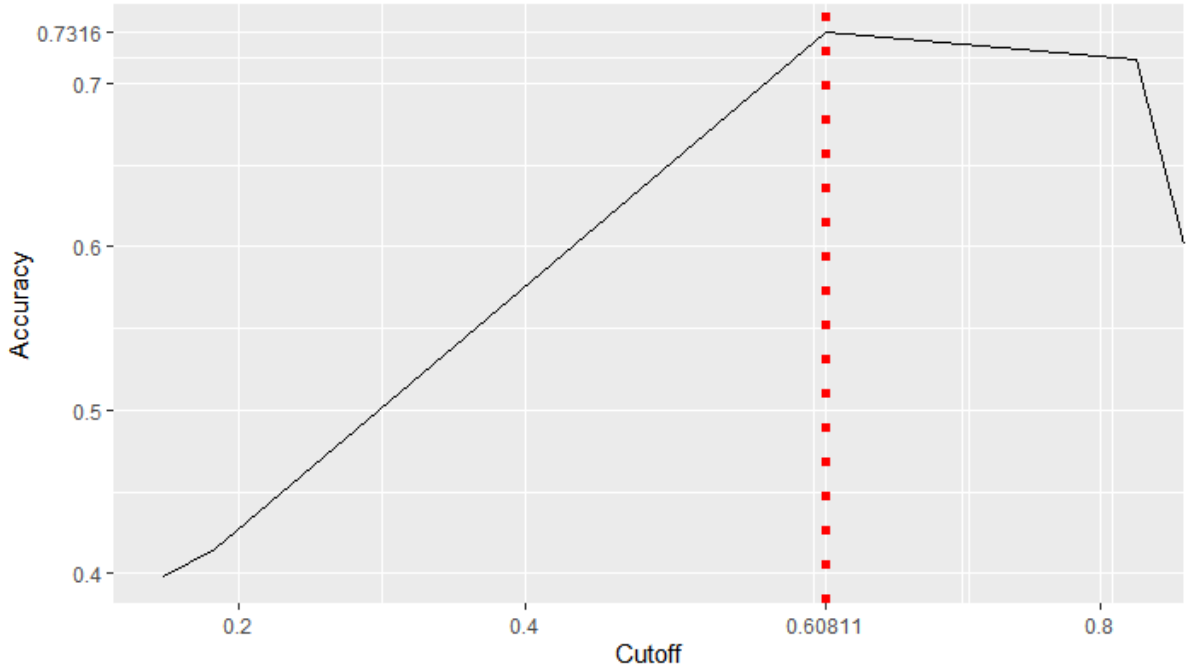
Decision Tree

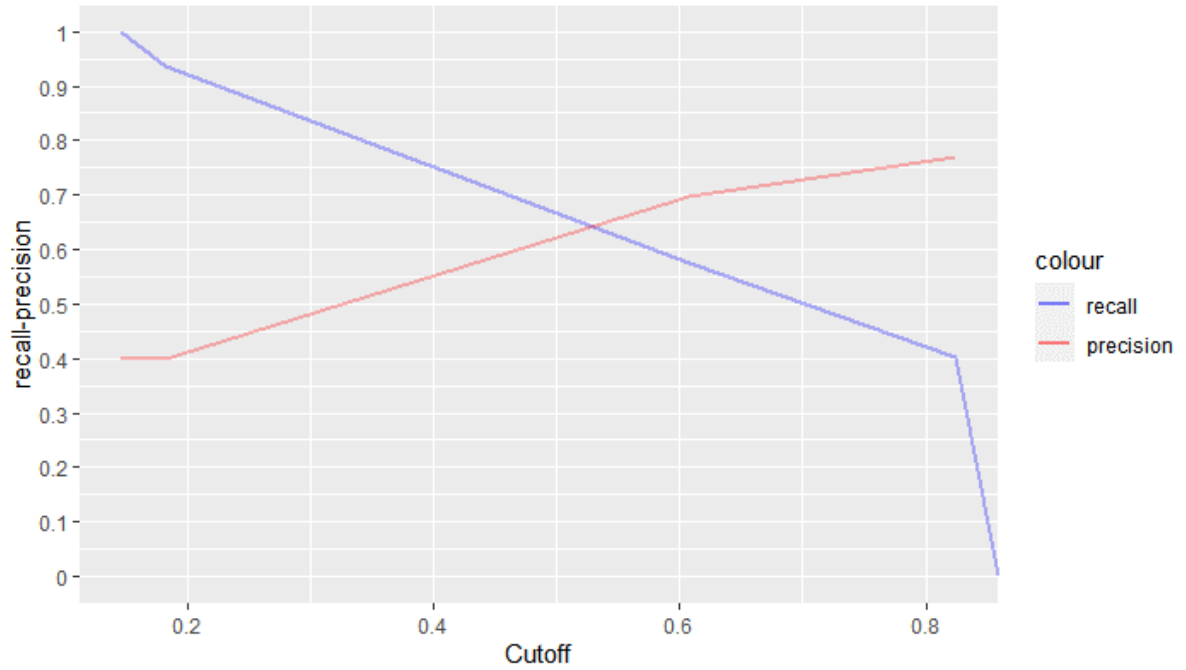
- mlr 패키지 사용하여 k-fold 교차검증 및 grid search 시행

- maxdepth : 9
- cp : 0.0289
- minsplit : 9



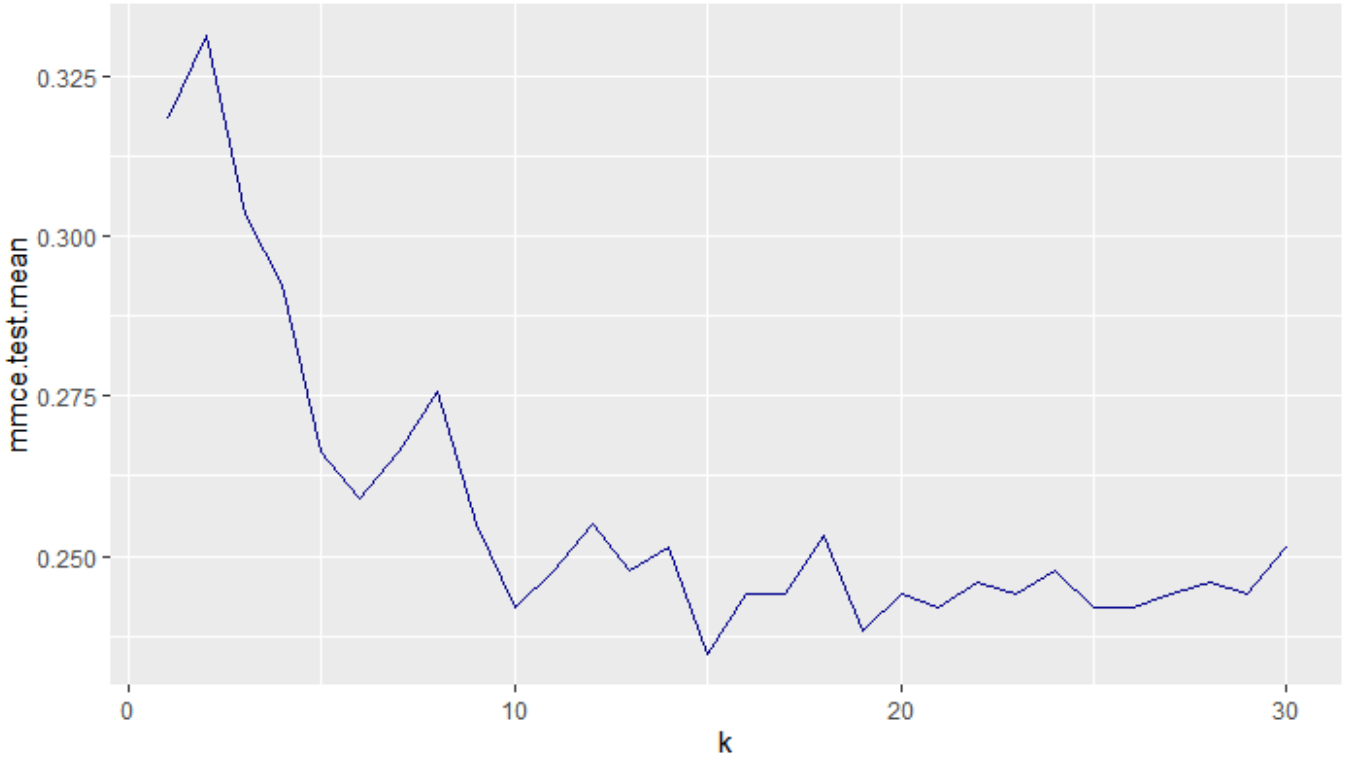
Decision Tree





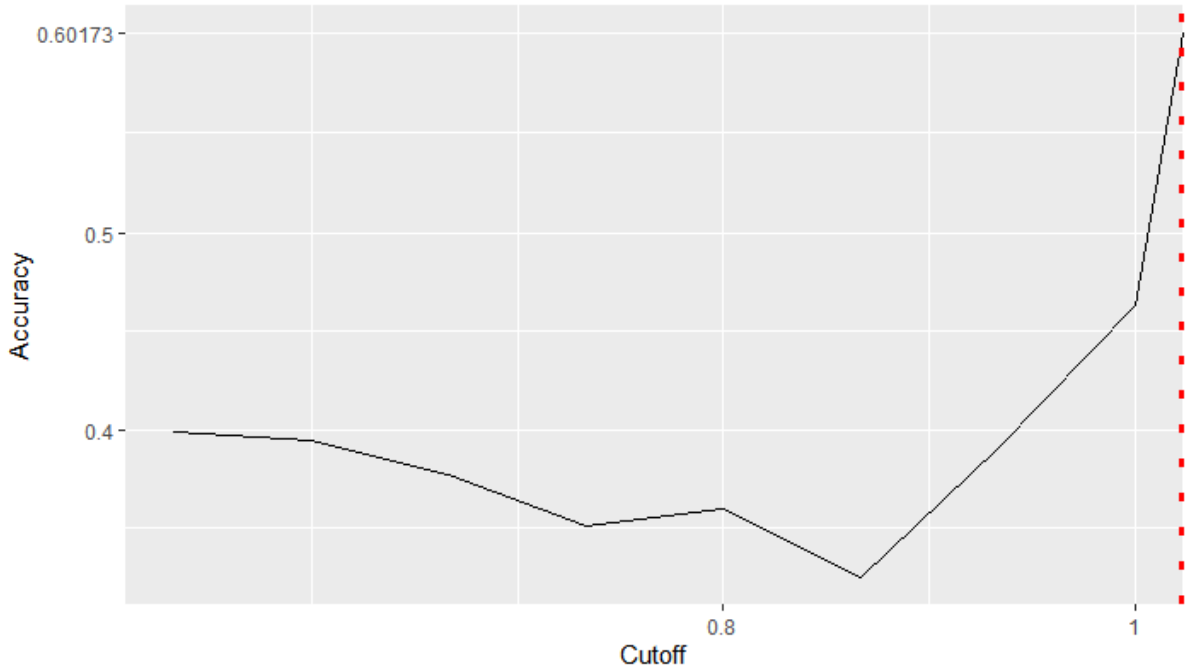
Cutoff 가 0.60811에서 Accuracy 0.7316

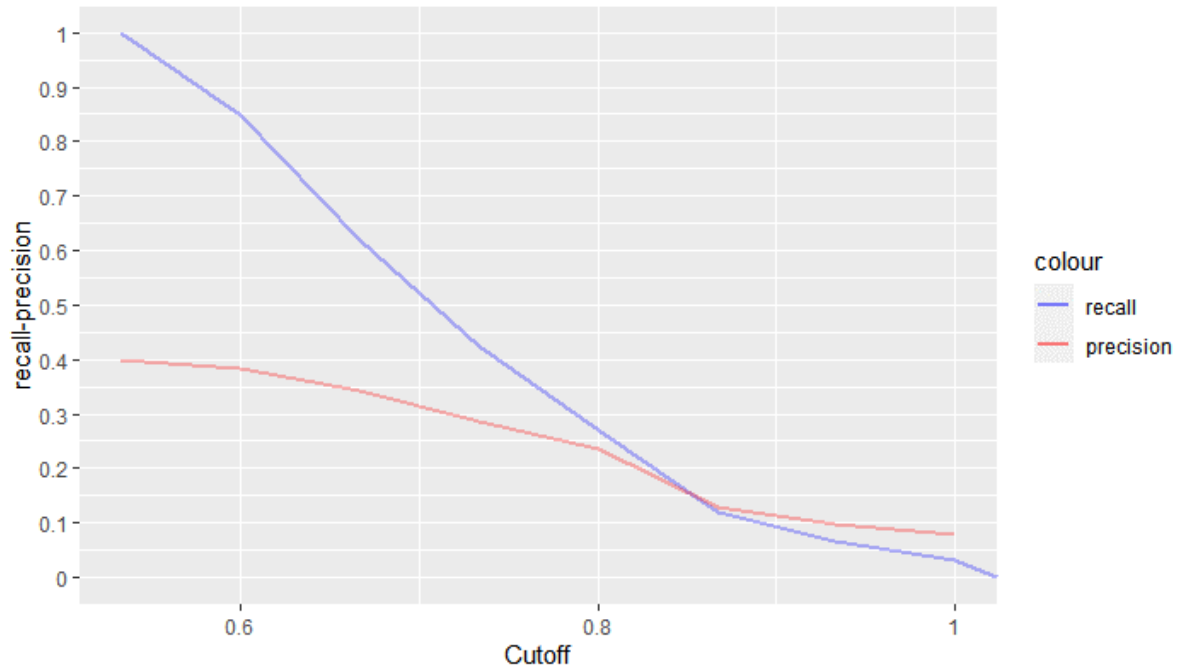
KNN



K는 15일 때 최적

KNN





KNN 모형은 이 데이터에 적합하지 않은 것으로 판단

SVM

- Gaussian Kernel 사용

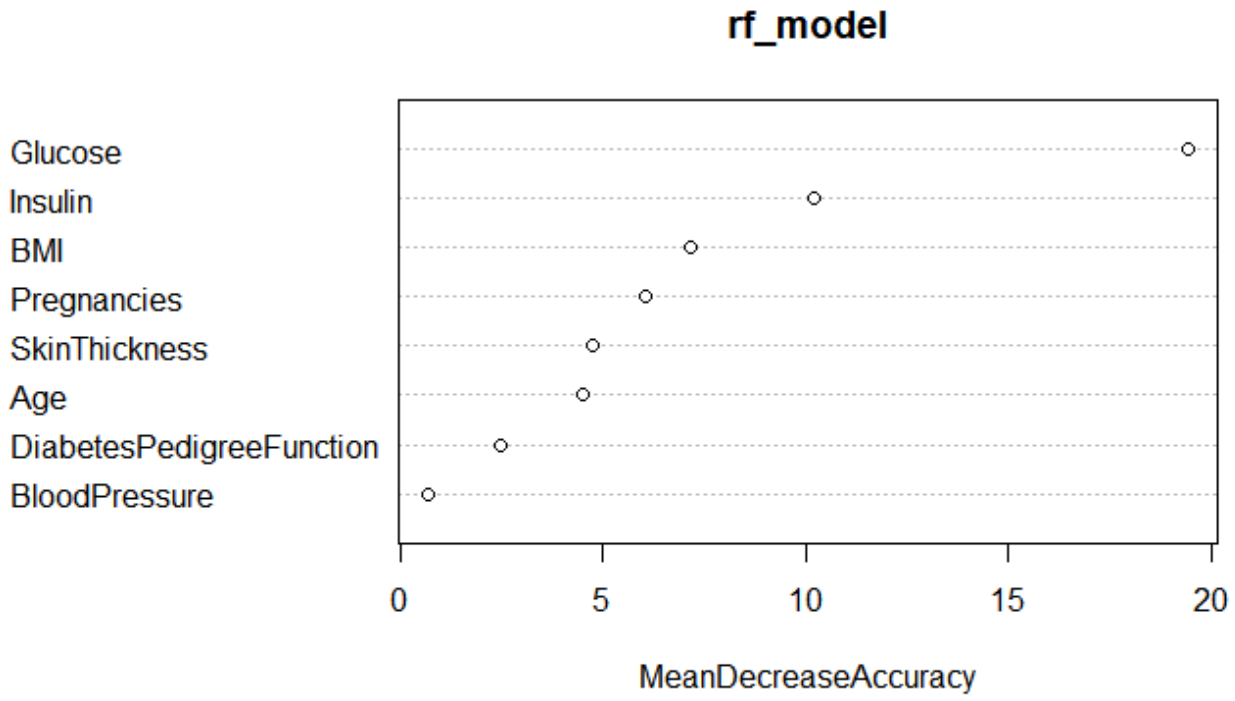
- C : 12.9
- Sigma : 0.000464

		True	
		0	1
Predict	0	124	46
	1	15	46

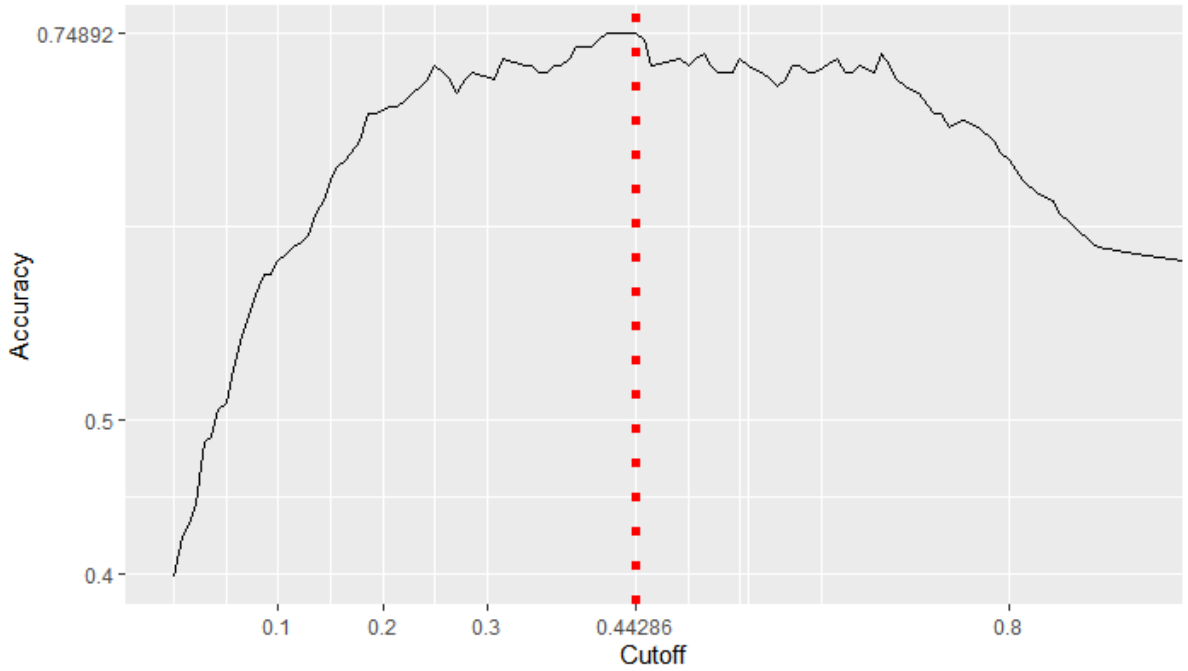
Accuracy : 0.7359

Random Forest

- mtry : 2
- ntree : 140
- nodesize : 7



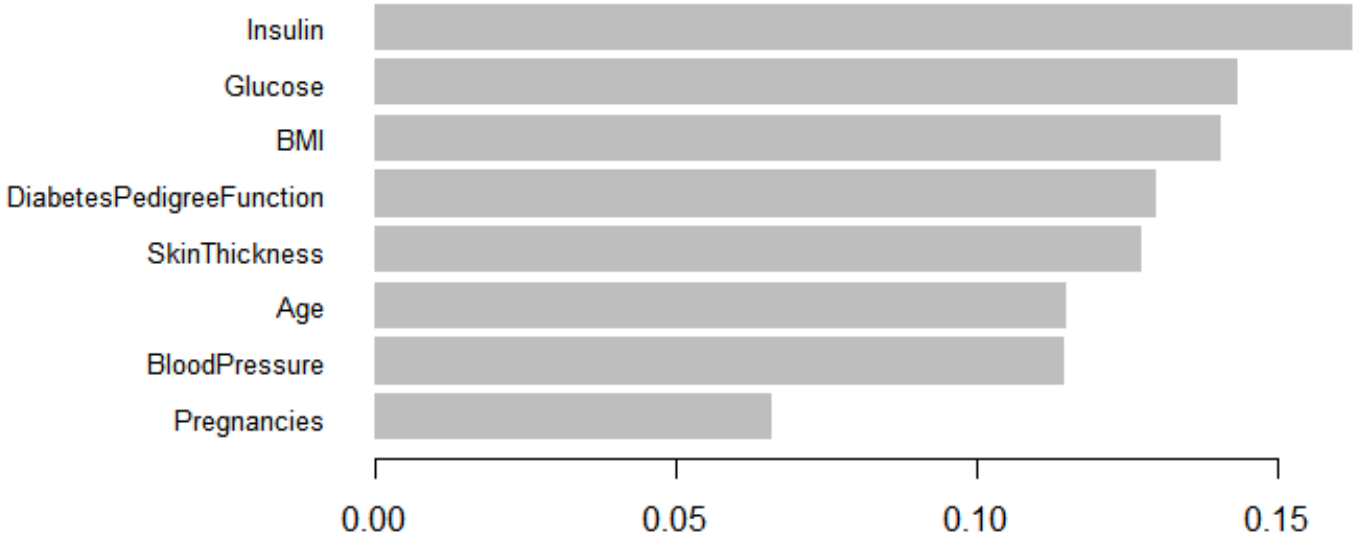
Random Forest



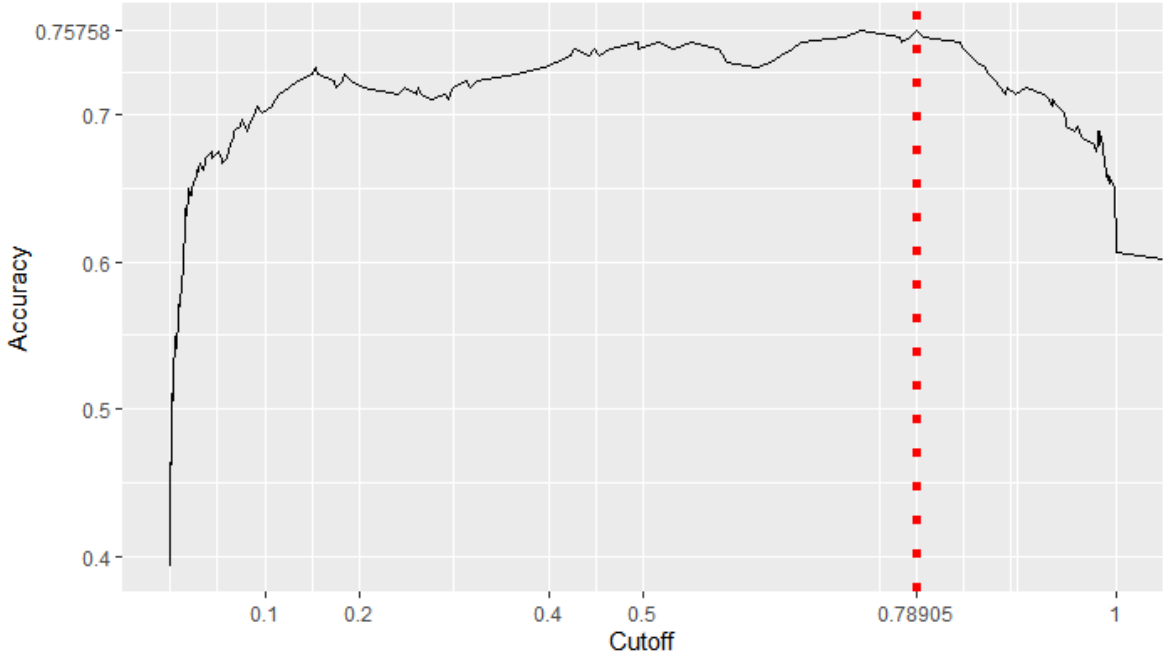
Cutoff 가 0.44286에서 Accuracy 0.74892

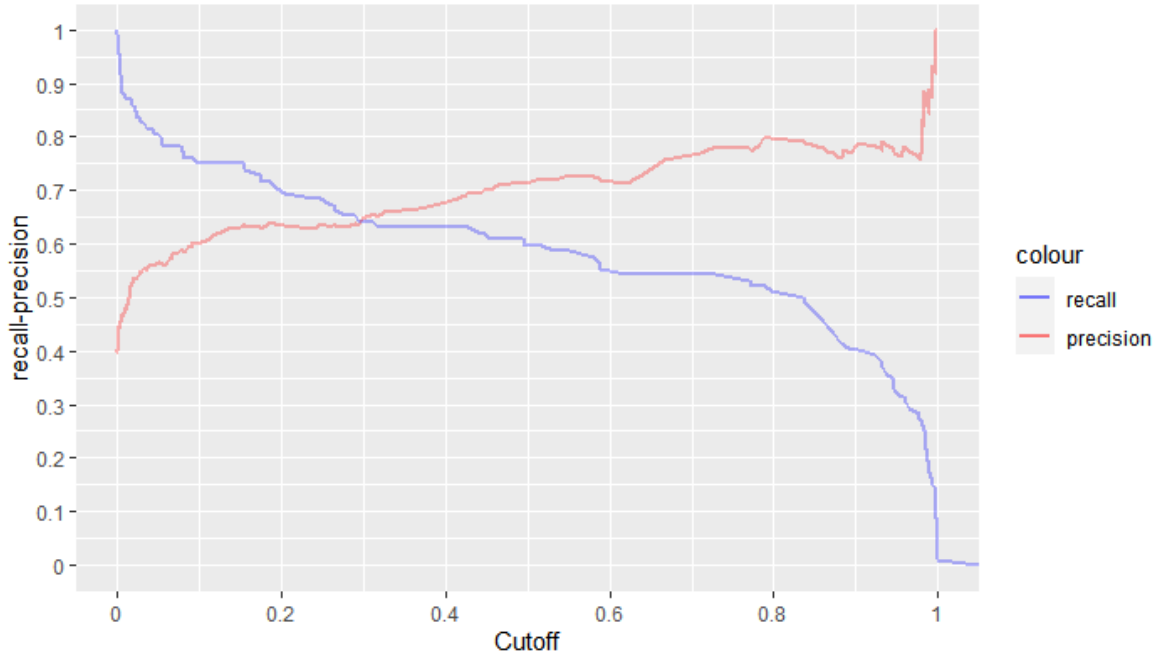
XGBoost

- nrounds : 189
- lambda : 0.167
- subsample : 0.311



XGBoost





Cutoff 가 0.78905에서 Accuracy 0.75758

Ensemble-soft voting

활용한 Model

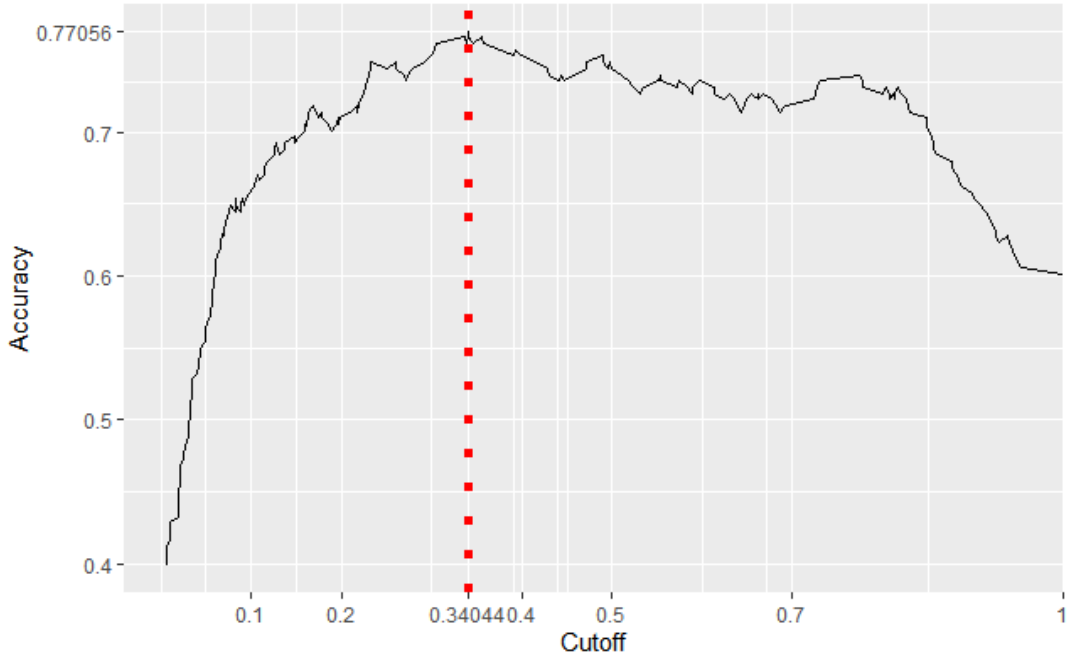
- Naive Bayes
- Logistic Regression
- Random Forest
- XGBoost

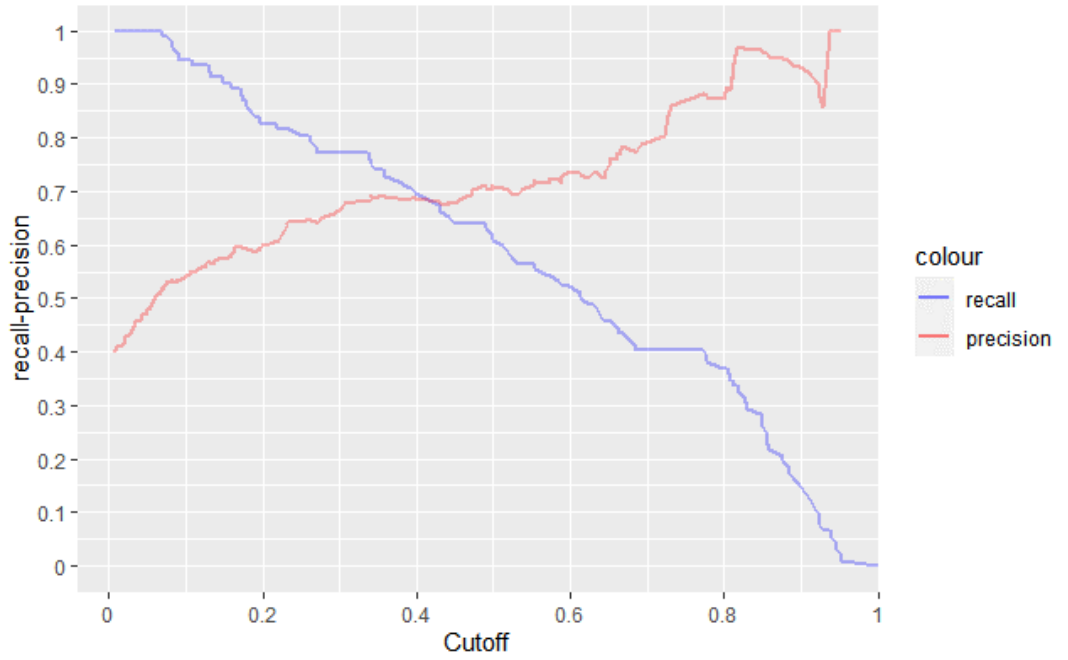
제외한 Model

- Decision Tree
- KNN
- SVM

Hard voting으로는 KNN만 제외해서 해본 결과 Accuracy 0.7489

Ensemble–soft voting





Cutoff 가 0.34044에서 Accuracy 0.77056

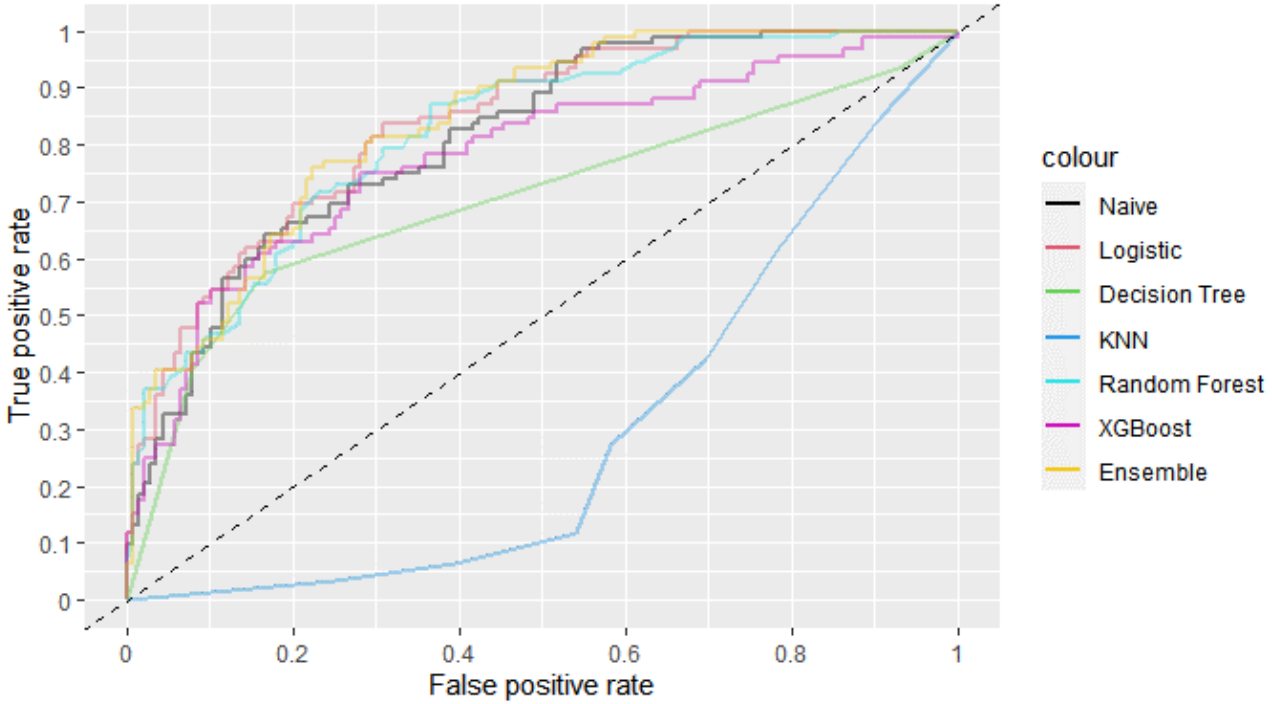
Part 3

Select Model

Part 3

Select Model

ROC Plot



AUC(Area under the ROC curve)

- Naïve : 0.8125586
- Logistic : 0.8368001
- Decision Tree : 0.7038239
- KNN : 0.2941038
- Random Forest : 0.821356
- XGBoost : 0.7790116
- Ensemble : 0.8389115

감사합니다