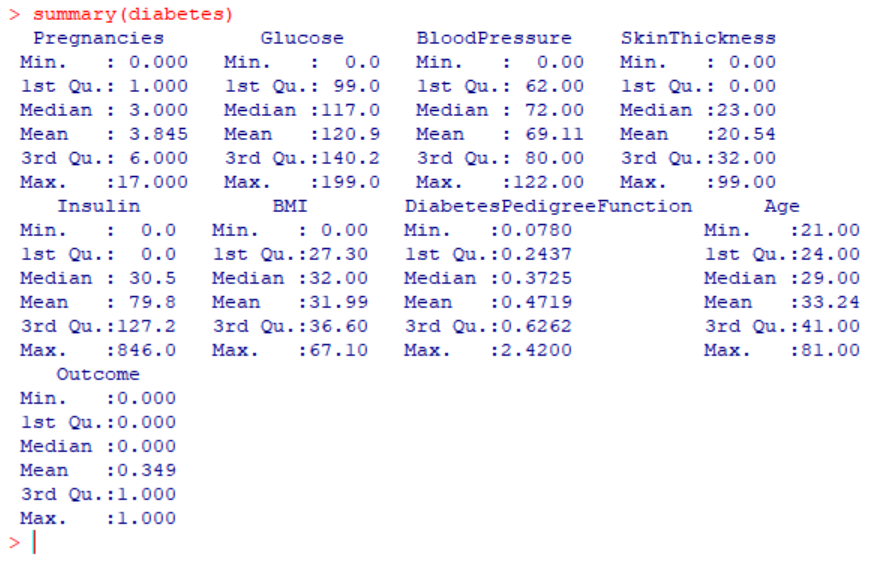
당뇨병 원인 분석

이은주, 나유정

분석 목적 : 당뇨병의 원인을 찾고 발병을 예측하기

데이터 정리 및 변수 설명



pregnancies : 임신 횟수

glucose : 경구 포도당 내성 검사에서 혈장 포도당 농도

blood pressure : 혈액 압력(확장기 혈압)

skin thickness : 삼두근 피부 두께

insulin : 혈청 인슐린

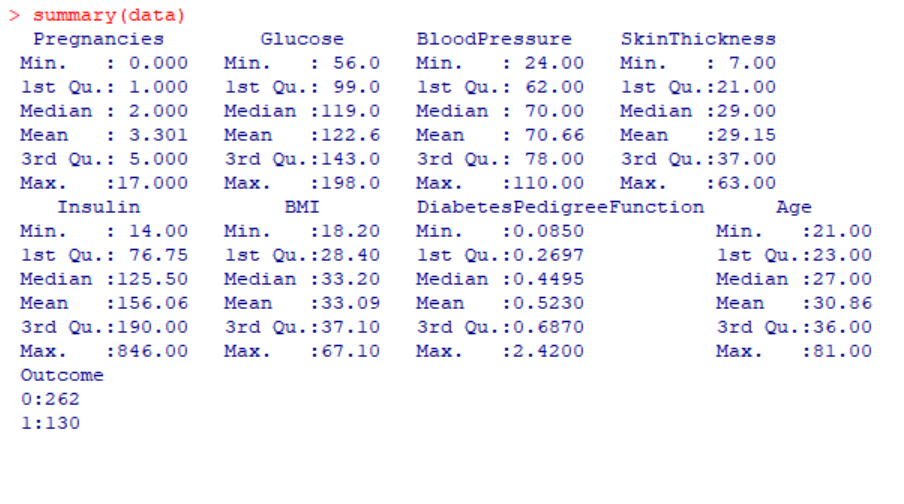
BMI : 체질량 지수(체중/(신장)^2)

diabetes pedigree function : 당뇨병 혈통 기능

age : 나이

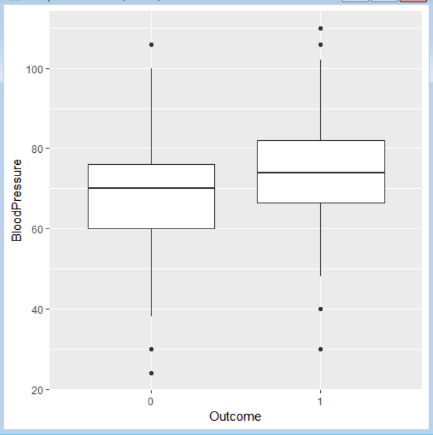
outcome : 결과(0 또는 1)

#불필요한 자료 제거



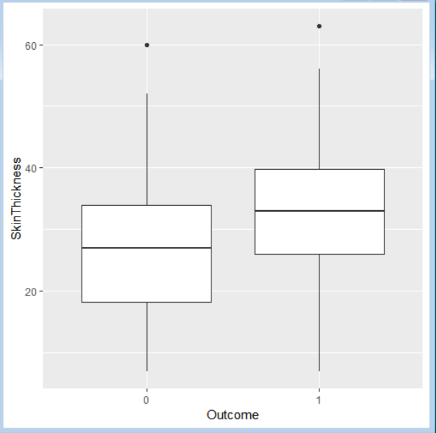
#변수와 당뇨병과의 관련성 비교

#혈압



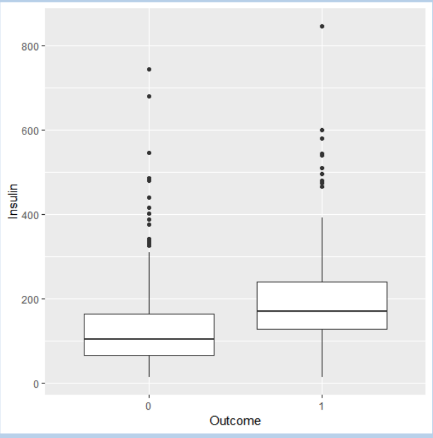
중앙값이 거의 비슷하여 큰 차이가 없어 보입니다.

#피부 두께



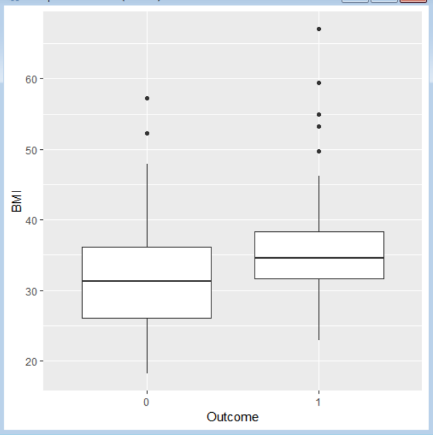
박스플랏 모양이 거의 비슷하기에 큰 차이가 없다고 생각할 수 있습니다.

#인슐린



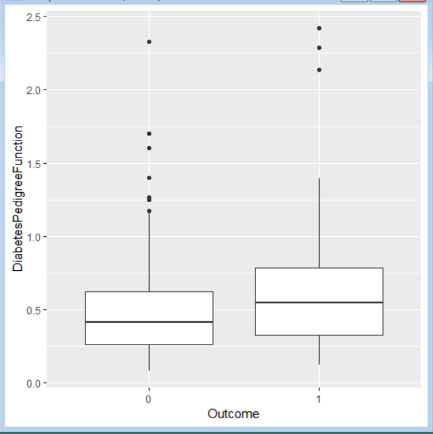
인슐린에 따른 당뇨병 발생은 큰 차이가 없어보입니다.

#BMI



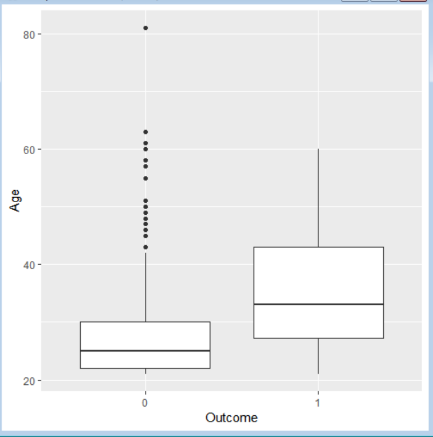
BMI에 따른 당뇨병 발생은 큰 차이가 없어보입니다.

#당뇨병 혈통 기능



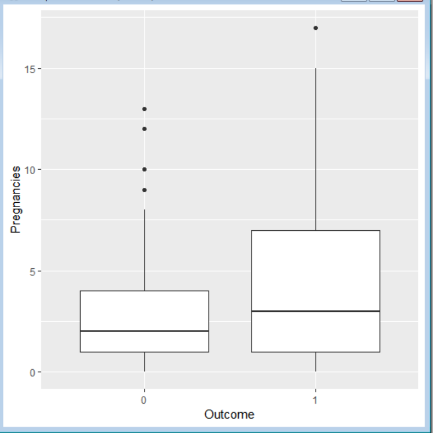
당뇨병은 혈통에 큰 영향을 받지 않습니다.

#나이



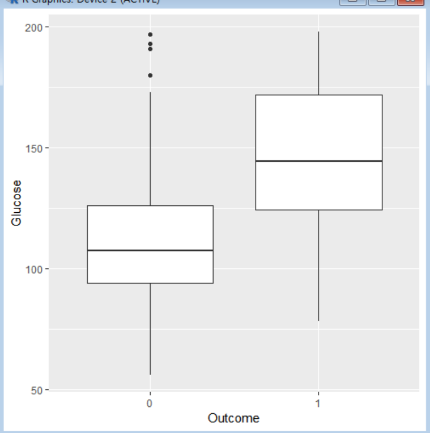
나이가 많을수록 당뇨병일 확률은 높습니다.

#임신횟수



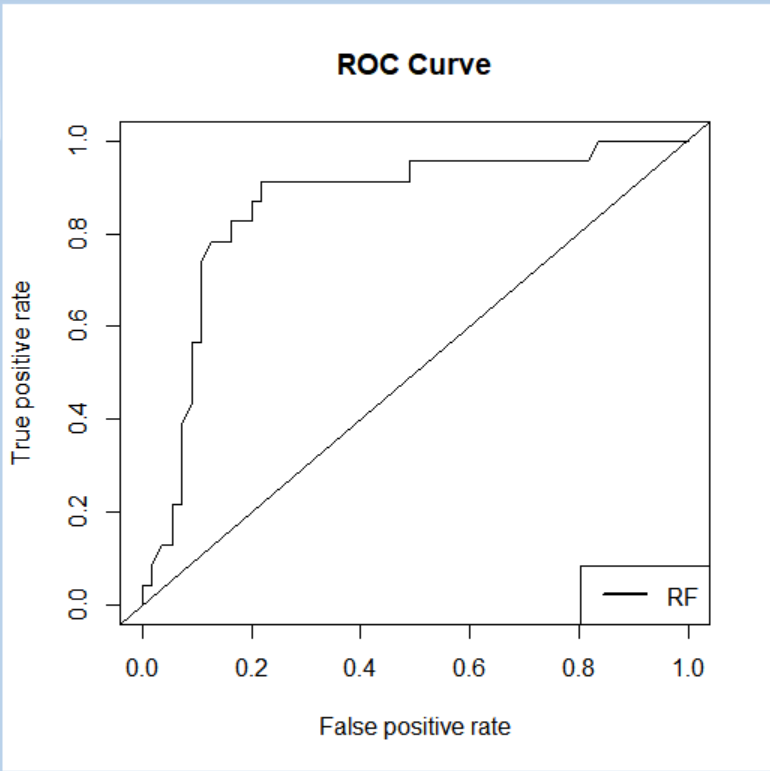
임신 횟수가 많을수록 당뇨병일 확률이 높습니다.

#글루코스



글루코스에 따른 분포가 확실하게 차이가 나 높으면 당뇨병일 확률이 큽니다.

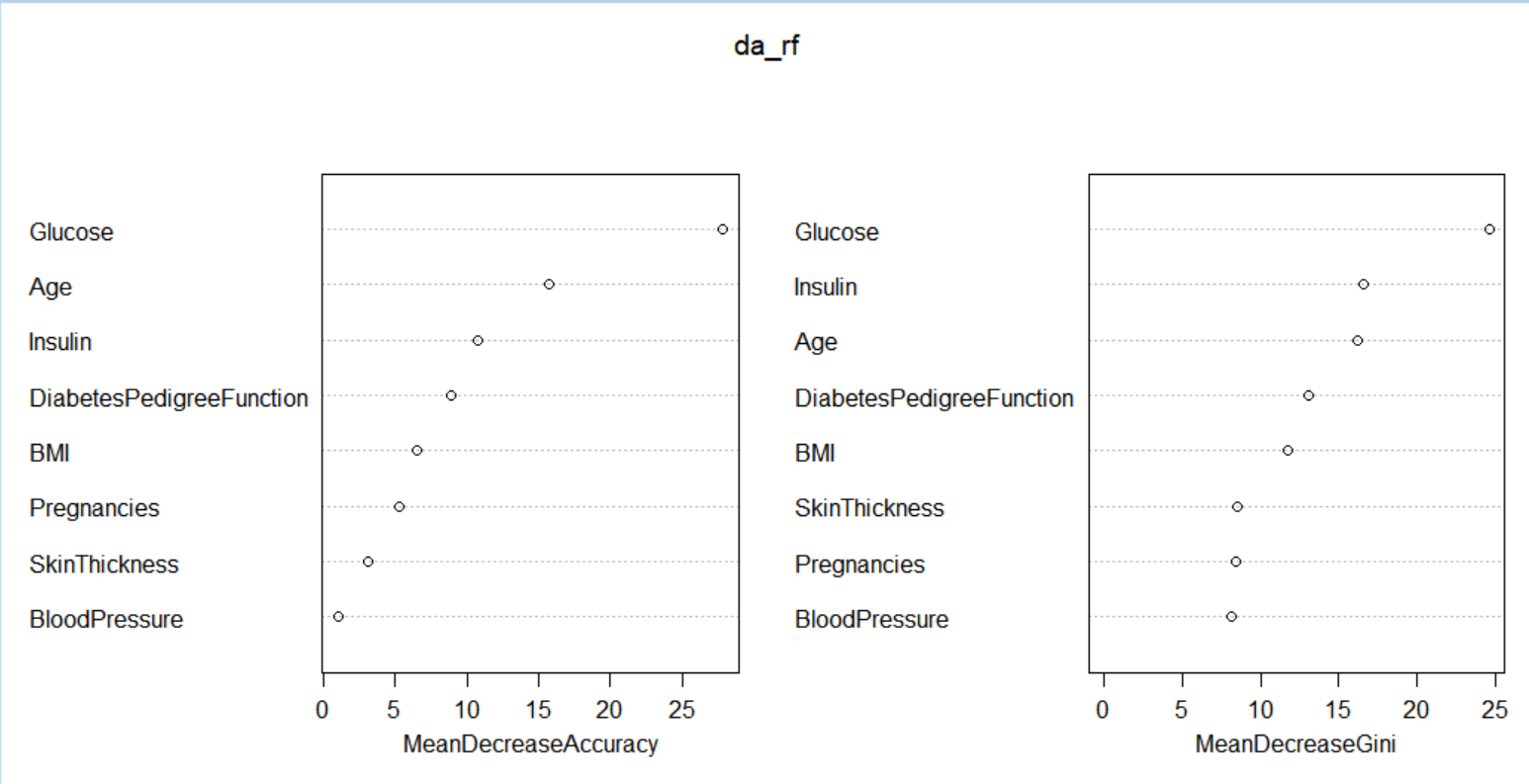
#랜덤포레스트 - ROC커브와 AUC



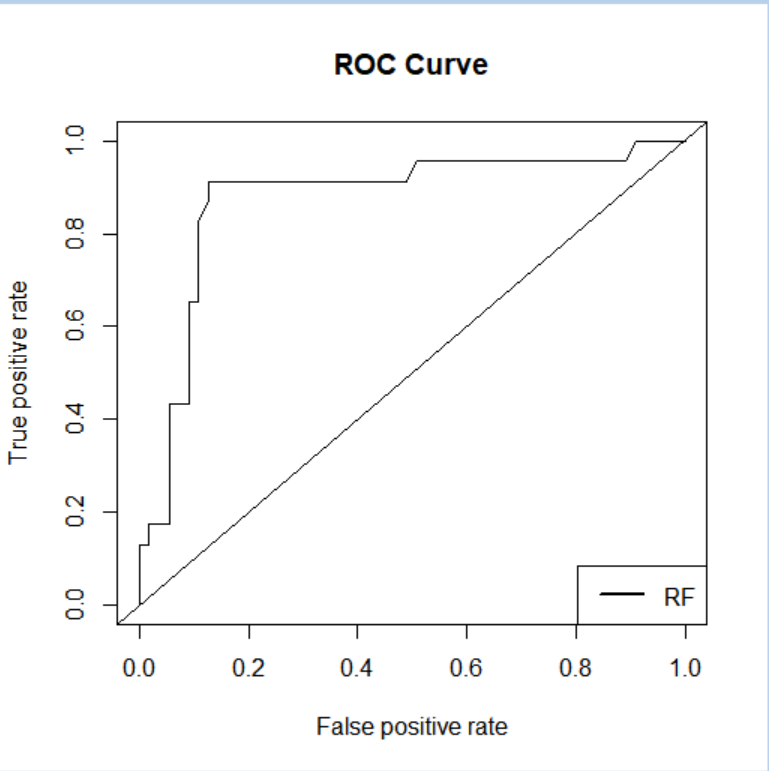


-> 본래 데이터의 ROC커브와 AUC

-> AUC를 올리기 위해 중요도 확인



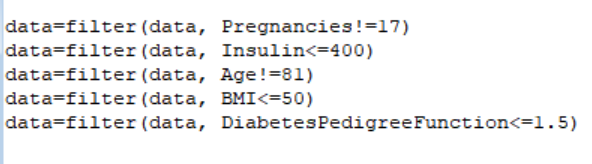
-> 중요도 낮은 BloodPressure과 SkinThickness 변수 제거



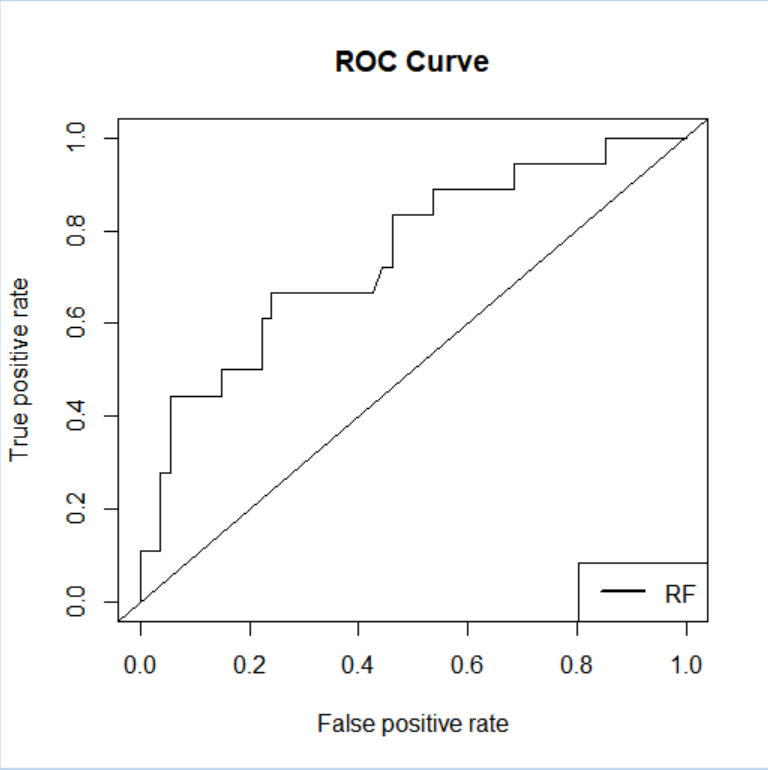


-> AUC값을 더 올리기 위해 이상치 제거

###이상치라 생각했던 모든 데이터 제거

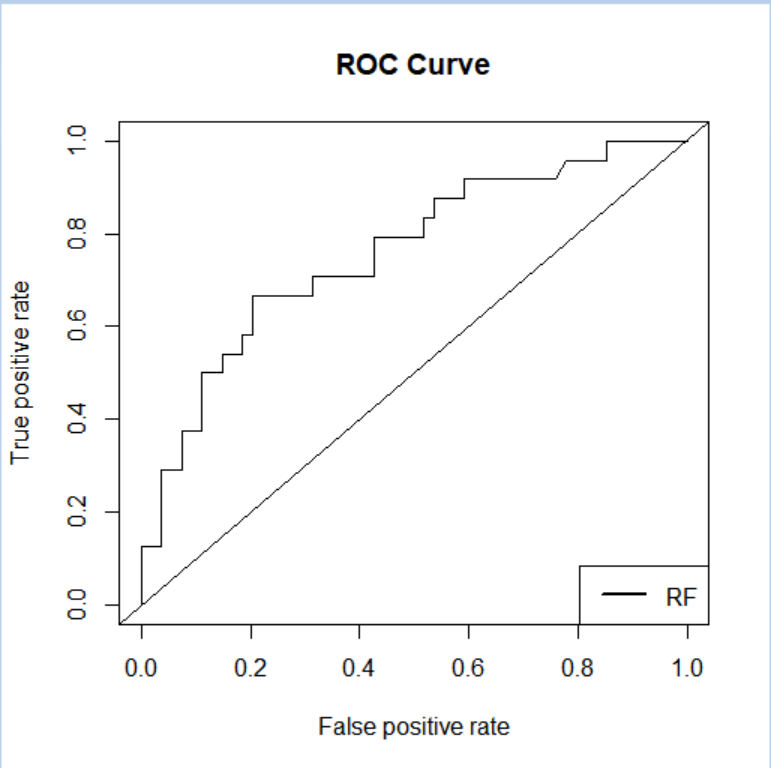






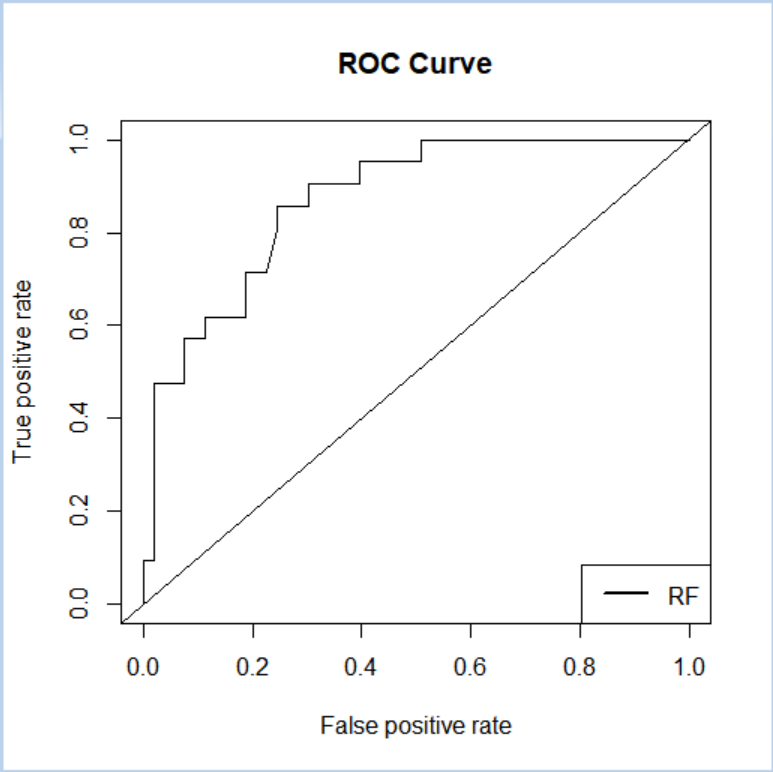
###Pregnancies의 이상치만 제거





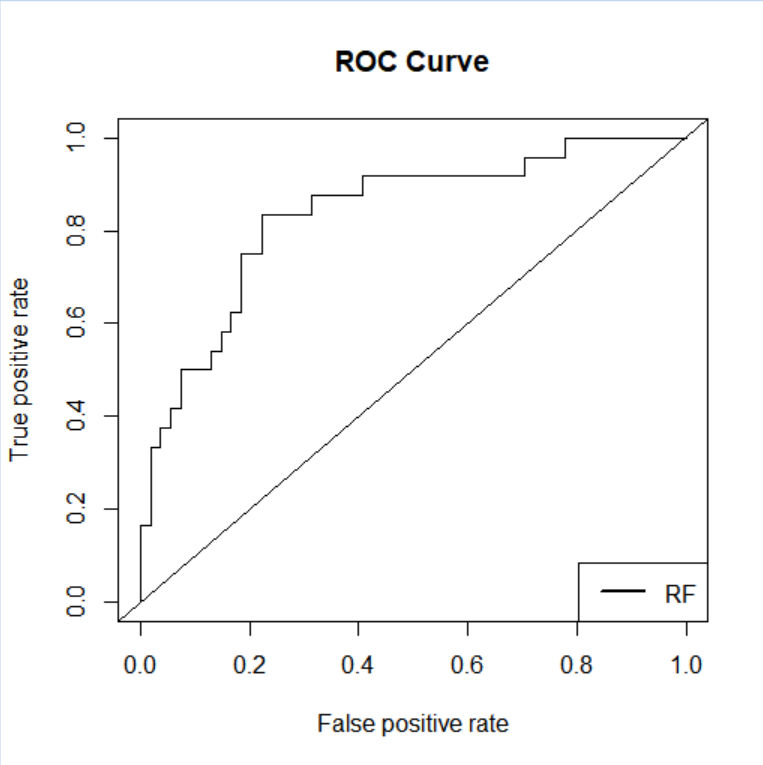
###Insulin의 이상치만 제거





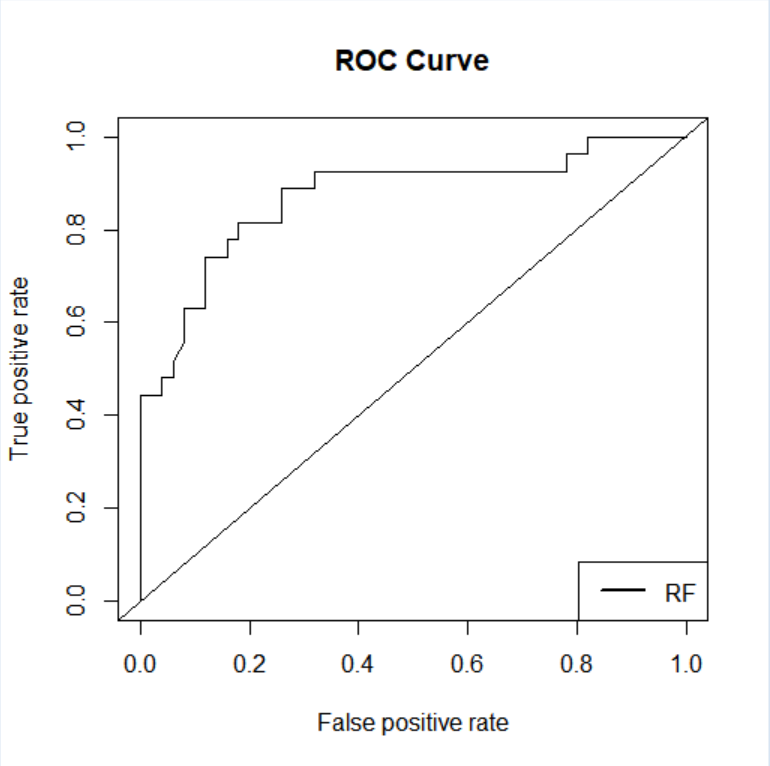
###Age의 이상치만 제거





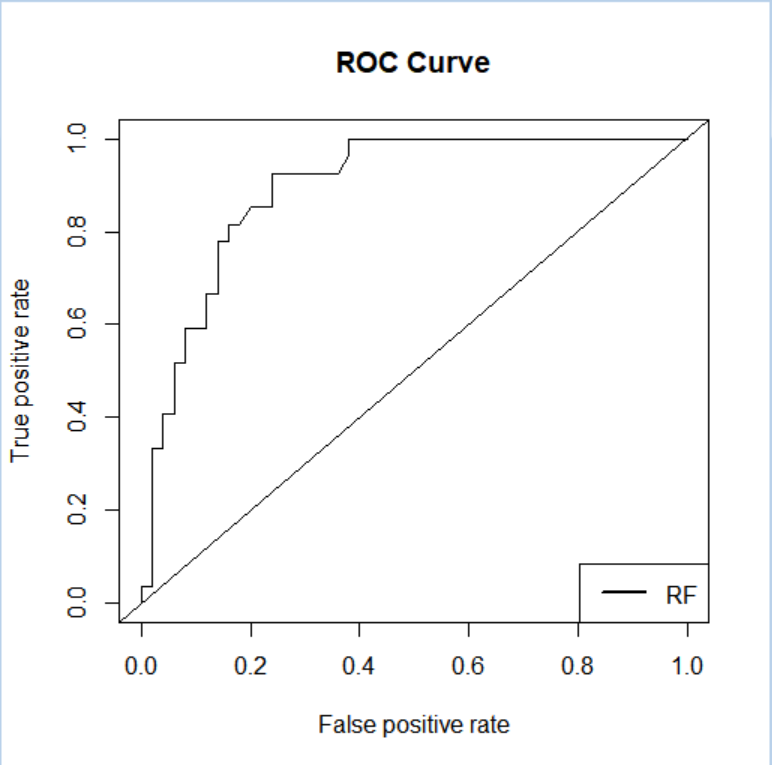
###BMI의 이상치만 제거





###DiabetesPedigreeFuncion의 이상치만 제거







-> 가장 AUC값이 올라가는 모델을 찾음.