PIMA 여성 인디언의 당뇨병 진단 예측 모델 선정

김연희 박소담 오현진 이가영
이수연 조현우 황수진

목차

CONTENTS

01

연구배경

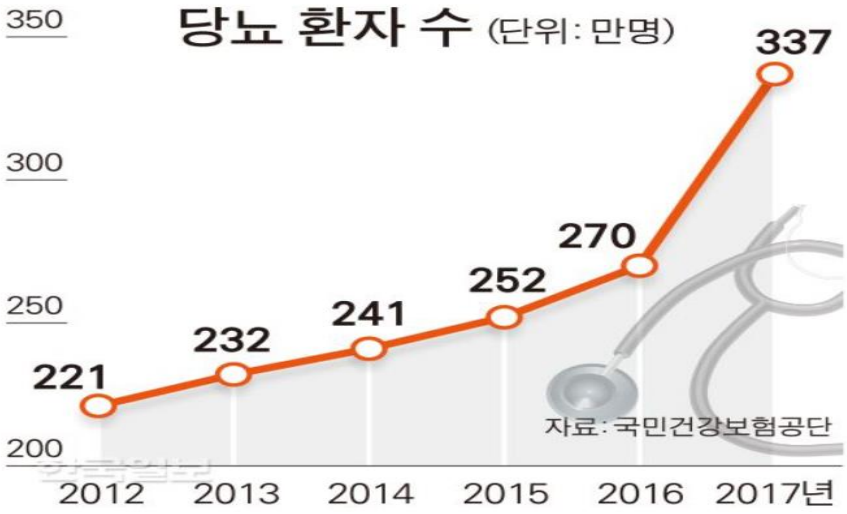
1.1 연구동기

1.2 분석 tool

1.3 데이터 출처

연구배경

연구동기



최근 급격하게 당뇨병 진단 환자의 수가 증가 ,
빅데이터의 발전으로 개인 의료 시스템 부문에 관심 증가

→ 그 중 임산부인 경우 혈당이 과도하게 증가 하면
임신성 당뇨병으로 이어짐 ,
계류유산과 심장질환을 가진 아이를 출산할 가능성 높아진다는
사실을 발견

연구목표

1. 당뇨병과 임산부 사이의 어떤관련이 있는지 임신횟수를 포함하여 여러 변수에 대해 알아보고 ,
당뇨병과 임산부와 관련된 변수들 간의 어떤관련이 있는지 알아보고자 함
2. 대표적인 PIMA인디언 여성의 당뇨병 데이터를 이용
→ **당뇨병 발생의 예측모델로서 적합한 모델을 찾고자 함**

연구동기

WHY ? PIMA 여성인디언 데이터를 선택한 이유

PIMA 인디언들의 주 음식은 '백색위험' 으로
불리는 백설탕, 정제밀가루이다.

'당뇨병의 원인' 이 되는 이들의 음식이
현시대 우리가 쉽게 노출되고 , 주식으로 삼는
비율이 높아졌다.

변수를 포함하여 여러 변수에 대해 알아

→ 이들을 표본으로 두고 연구를 진행함

2. 대표적인 PIMA인디언 여성의 당뇨병 데이터를 이용
   → **당뇨병의 발생의 예측모델로서 적합한 모델을 찾고자 함**

연구배경

분석 tool

분석도구로 R 을 이용
함

데이터 출

미국 국립과학재단 (USF) 의 지원을 받아 UCI 에서 운영하는
UCI머신러닝 저장소로 부터 PIMA Indian diabets 데이터를 얻음

02

변수설명

2.1 예측변수

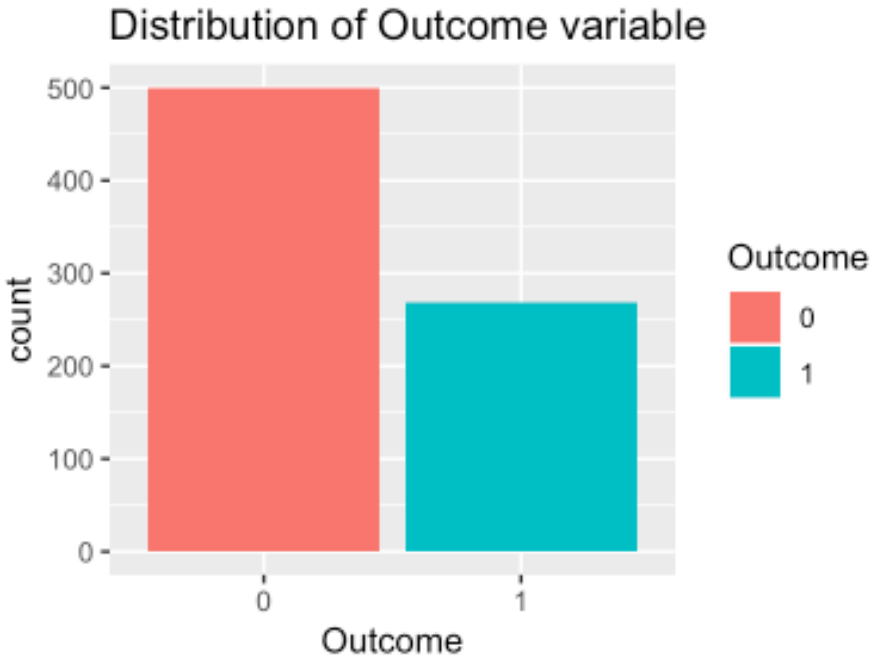
2.2 반응변수

예측변수

Pregnancies : 임신횟수
Glucose : 포도당
Blood Pressure : 혈압
Skin Thickness : 피부 두께
Insulin : 인슐린
BMI : 체질량지수
Diabetes Pedigree Function: 당뇨병 가족력
Age : 나이

반응변수

결과(outcome)
이 자료에는 당뇨병 진단을 받은 268명의 여성과 당뇨 진단을 받지 않은 500명의 여성이 있다.
[당뇨=1 , 당뇨 아님 =0]



03

데이터 분석

3.1 변수들과 당뇨병의 관계

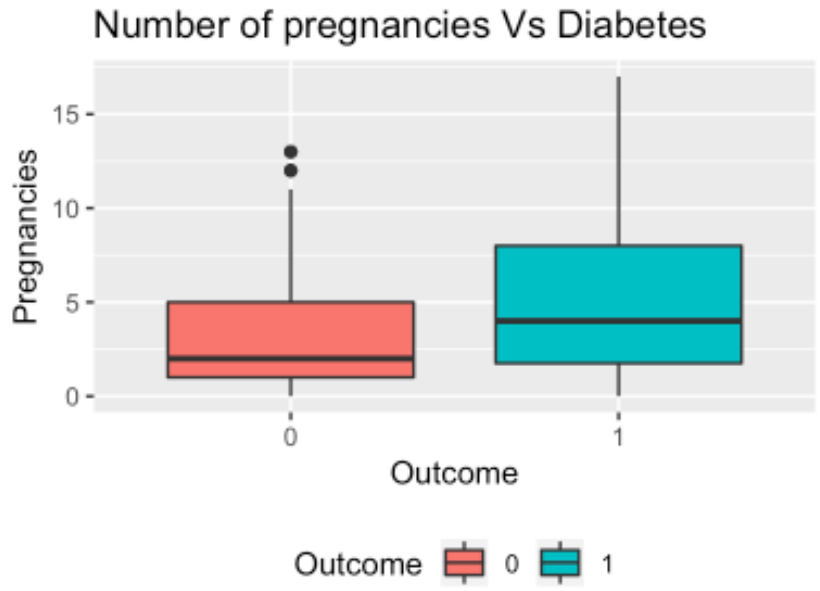
3.2 독립변수들 간의 상관관계

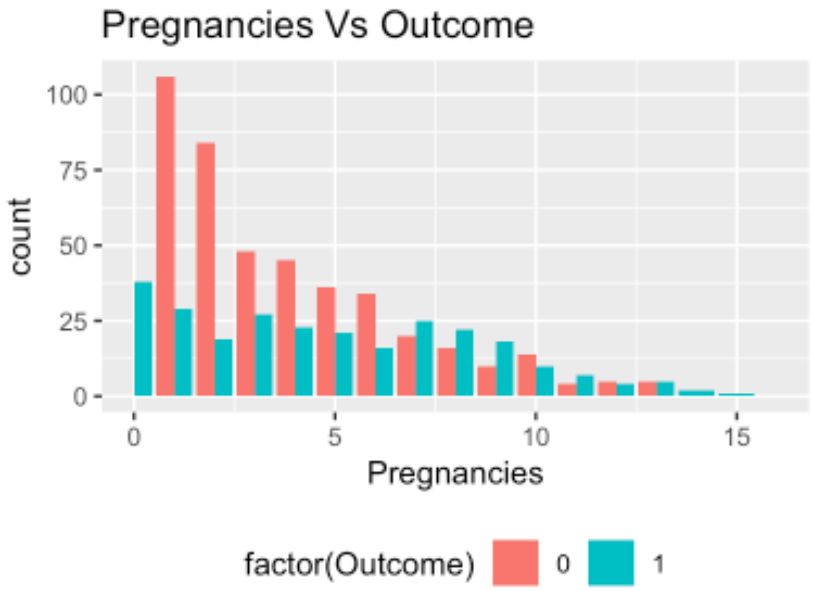
데이터 분석
변수들과 당뇨병의 상관관계

임신 횟수

박스 플롯에서 당뇨병 진단을 받은 여성들이 그렇지 않은 여성들보다 더 많은 임신을 한 것을 알 수 있음
히스토그램을 보면 임신 횟수와 당뇨병 발생 사이에 명확한 관계가 없는 것을 알 수 있음

그래프



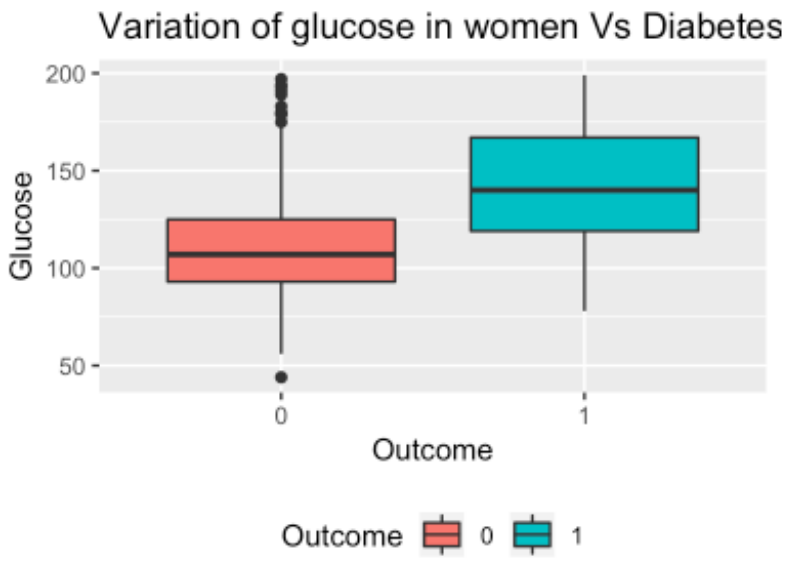


데이터 분석
변수들과 당뇨병의 상관관계

포도당

밀도도에서는 여성의 두 범주 포도당 수준이 중복되는 것을 나타내는 반면에 박스 플롯에서 당뇨 진단을 받은 여성들과 그렇지 않은 여성들에게 존재하는 포도당 양의 명확한 차이를 보여줌

그래프

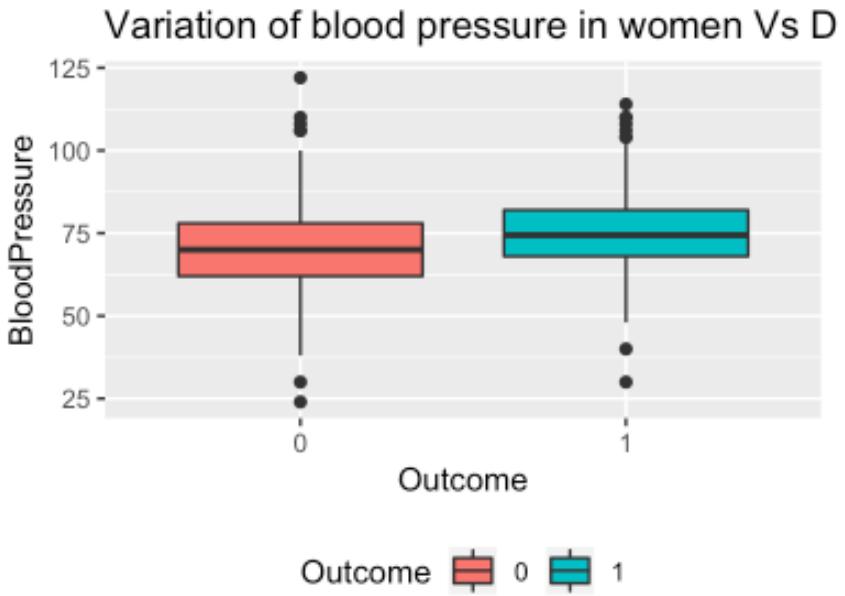


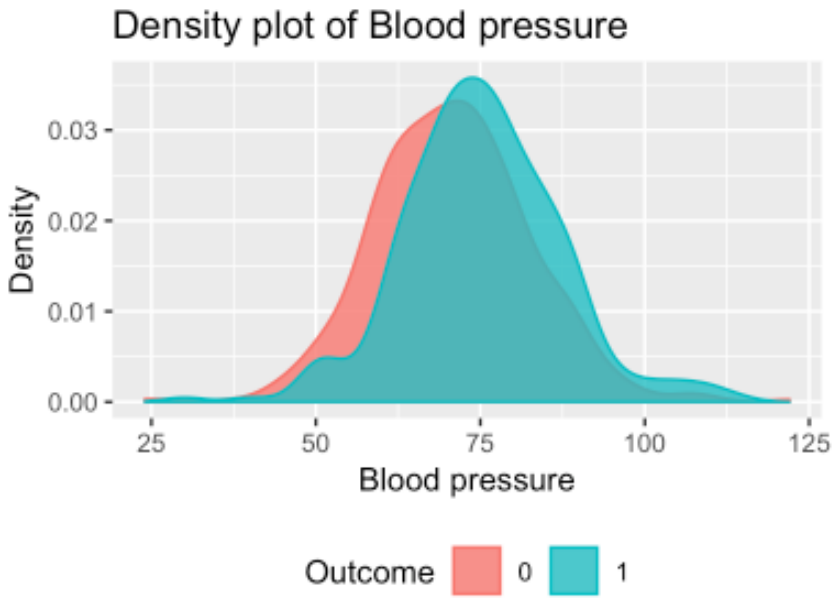
데이터 분석
변수들과 당뇨병의 상관관계

혈압

당뇨병을 가지고 있는 여성과 그렇지 않은 여성의 두 범주에는 뚜렷한 차이가 보이지 않음
혈압이 반응 변수의 좋은 예측 변수가 아닐 수도 있다는 것을 보여줌

그래프



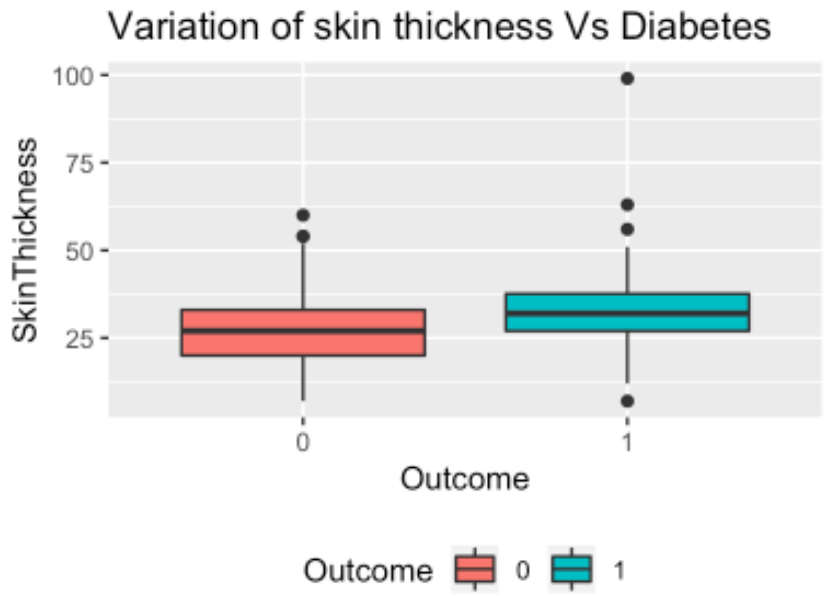


데이터 분석
변수들과 당뇨병의 상관관계

피부두께

당뇨병을 가지고 있는 여성과 그렇지 않은 여성의 두 범주에 뚜렷한 차이가 없음
마찬가지로 피부 두께가 반응 변수의 좋은 예측 변수가 아닐 수 있다는 것을 보여줌

그래프

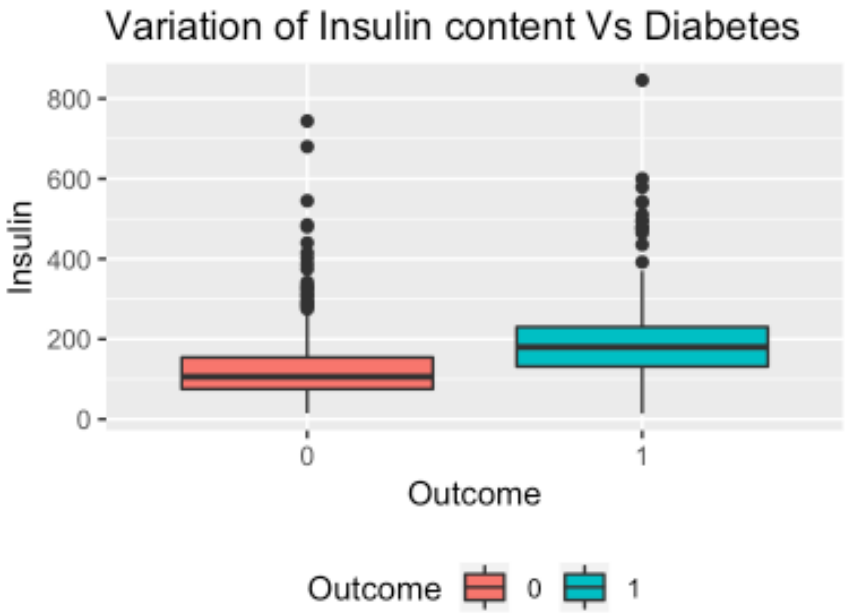


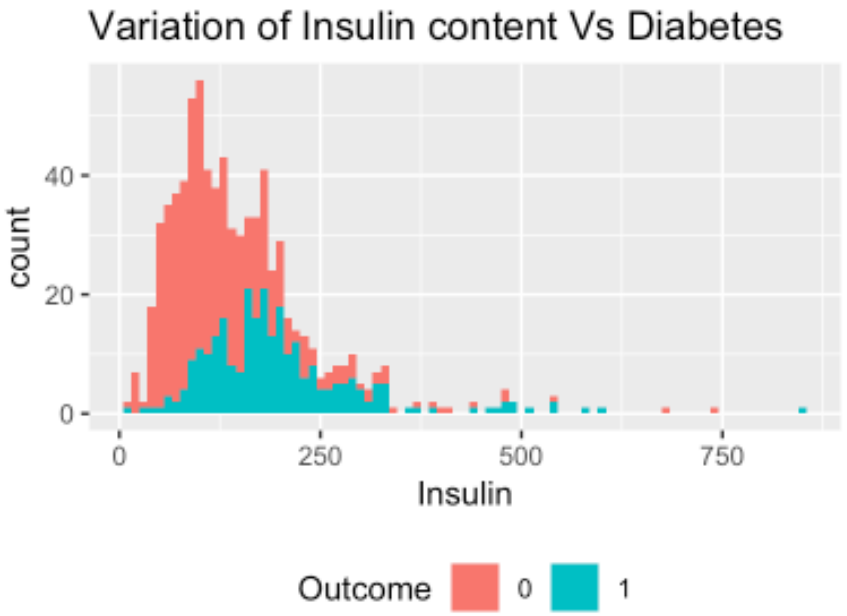
데이터 분석
변수들과 당뇨병의 상관관계

인슐린

당뇨병을 가지고 있는 여성과 그렇지 않은 여성의 두 범주에 뚜렷한 차이가 없음
인슐린이 반응 변수의 좋은 예측 변수가 아닐 수 있다는 것을 보여줌

그래프



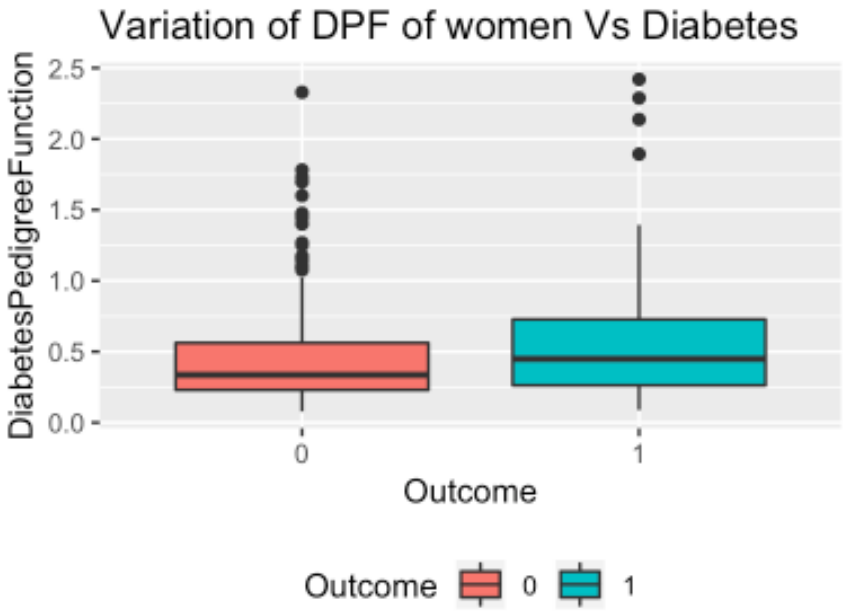


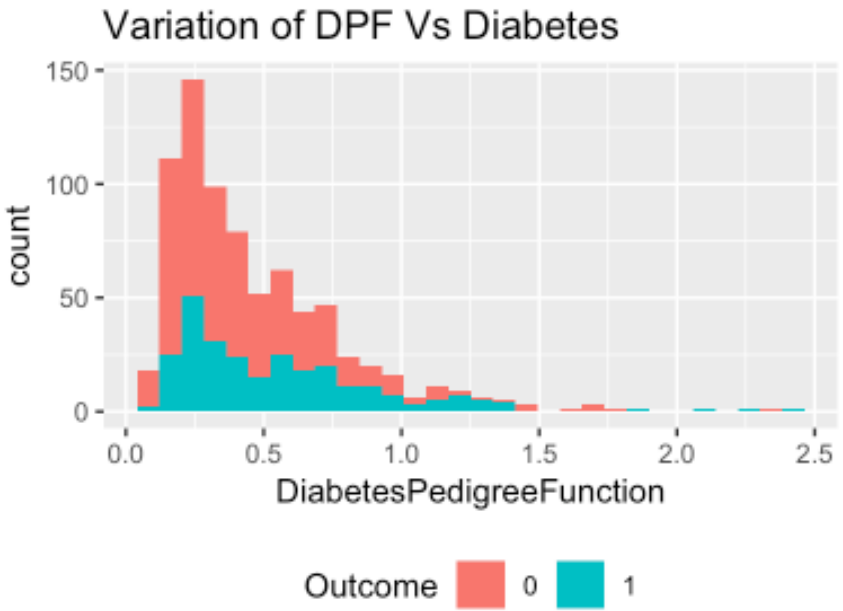
데이터 분석
변수들과 당뇨병의 상관관계

당뇨병 가족력

당뇨병을 가지고 있는 여성과 그렇지 않은 여성의 두 범주에 뚜렷한 차이가 없음
DPF가 반응 변수의 좋은 예측 변수가 아닐 수 있다는 것을 보여줌

그래프



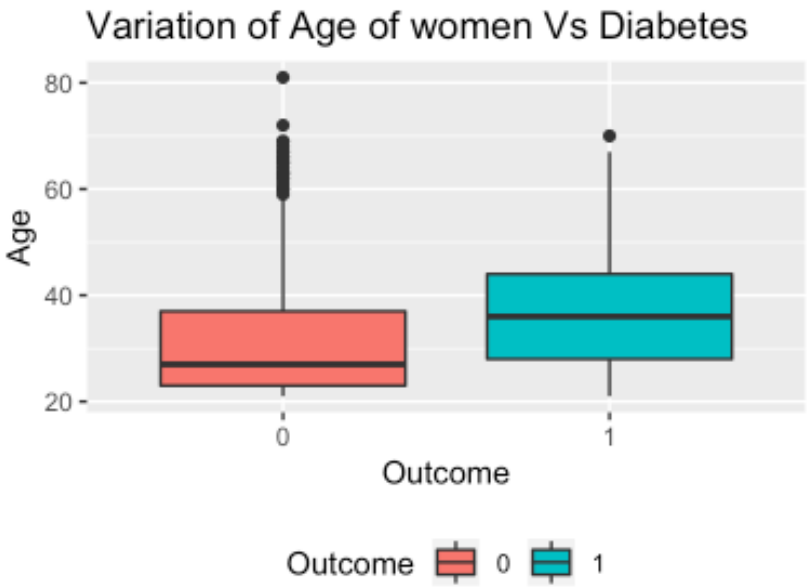


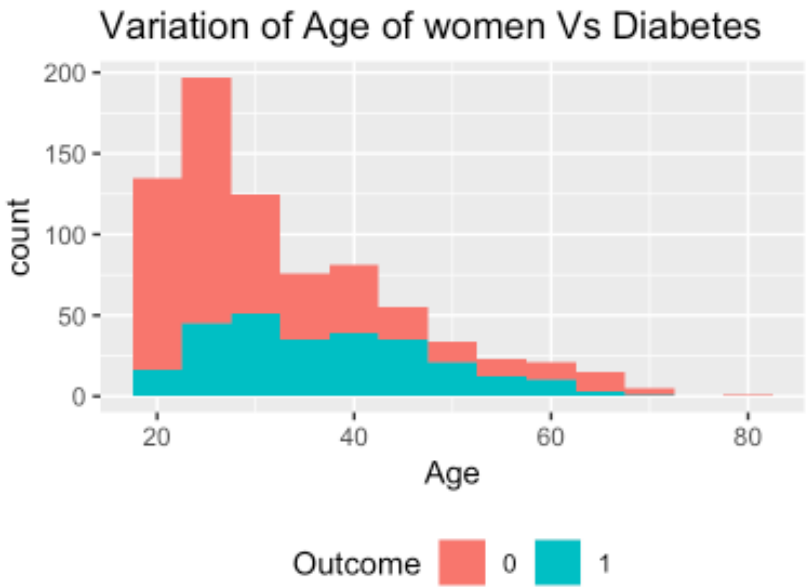
데이터 분석
변수들과 당뇨병의 상관관계

나이

당뇨병을 가지고 있는 여성과 그렇지 않은 여성의 두 범주에 뚜렷한 차이가 없음
나이가 반응 변수의 좋은 예측 변수가 아닐 수 있다는 것을 보여줌

그래프

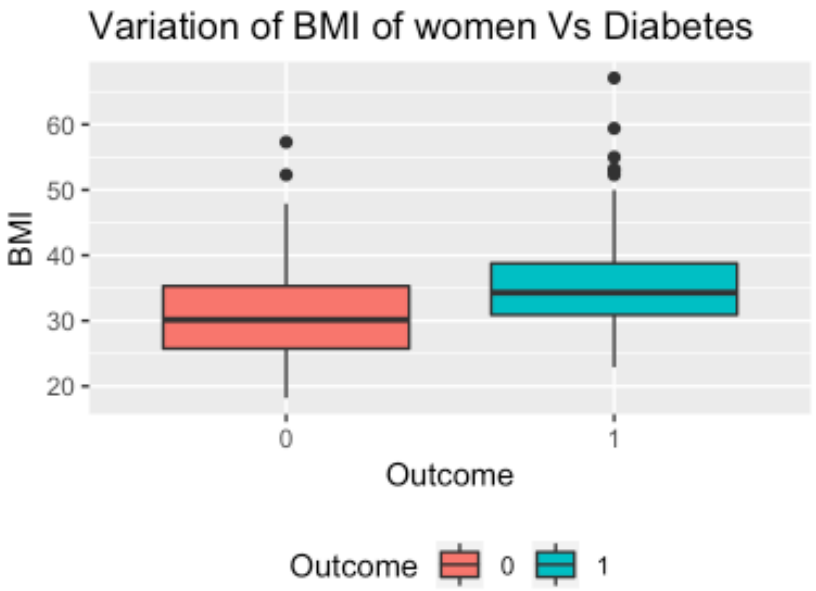


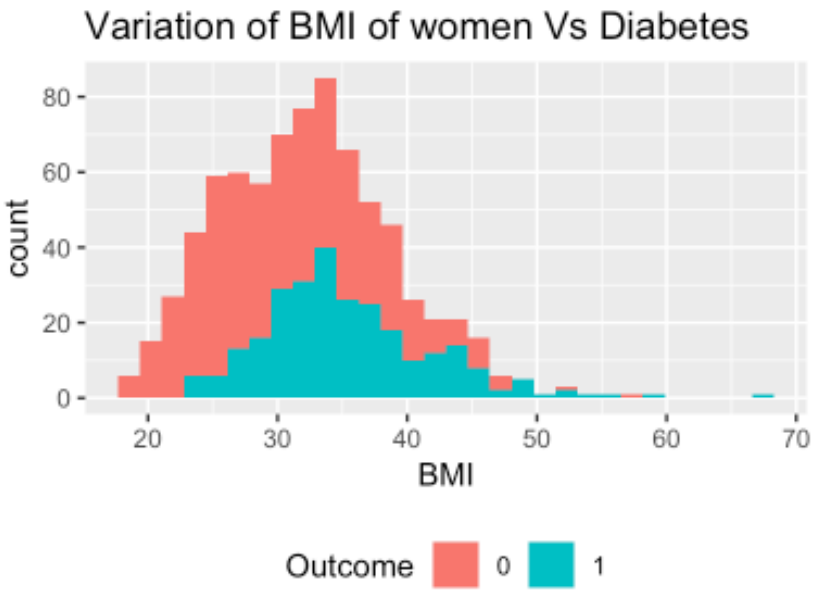


BMI

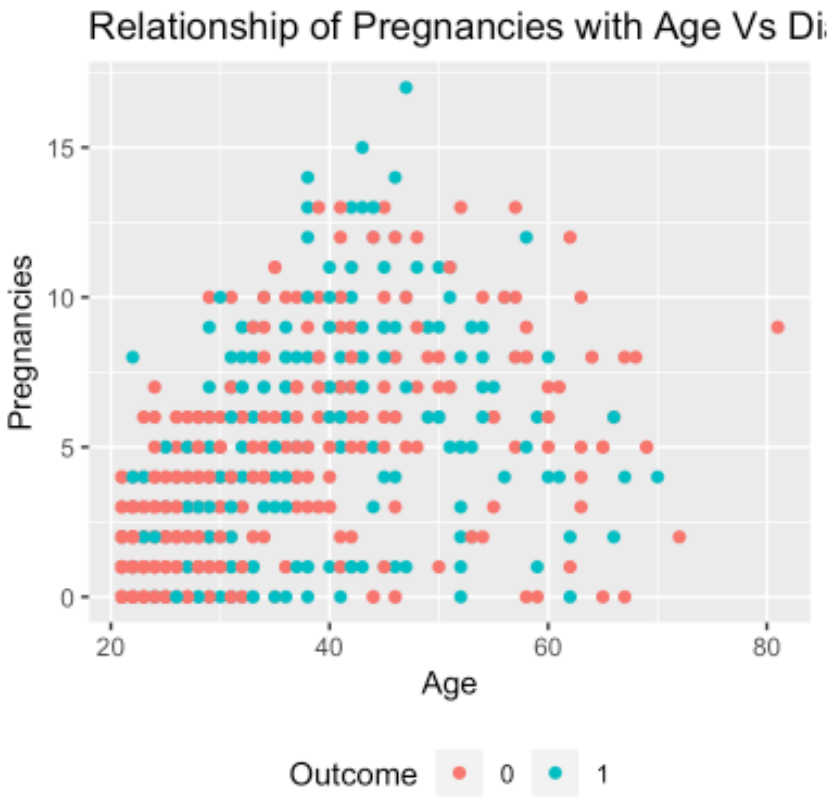
당뇨병을 가진 여성들이 그렇지 않은 여성보다 25이상의 높은 BMI 수치를 가지고 있다는 것을 보여줌

그래프



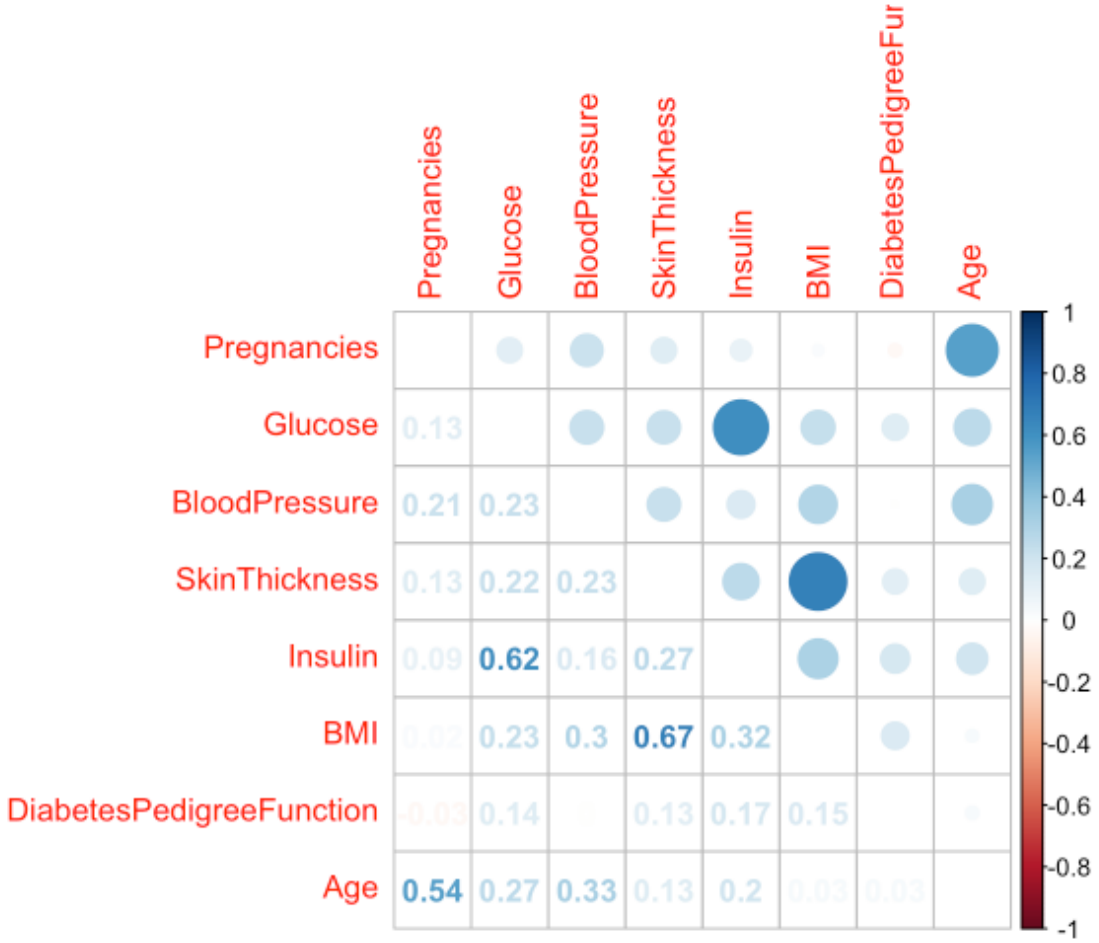


데이터 분석
이중 변량 연관



- 당뇨가 있는 여성과 당뇨가 없는 여성을 연령대별로 구분하는 명확한 경계선은 없음
- 당뇨병에 걸리지 않은 여성들이 당뇨병에 걸린 여성들에 비해 포도당 수치가 낮은 것으로 보임

데이터 분석
예측변수들 간의 상관관계



상호 간의 선형 관계를 설정하기 위해 모든 변수 사이에 상관 관계도를 그렸을 때 인슐린과 포도당, BMI와 피부 두께는 높은 선형상관 관계를 보임

04

Modeling

4.1 로지스틱 회귀

4.2 의사결정 나무

4.3 랜덤 포레스트

4.4 SVM

Modeling

로지스틱 회귀(Logistic Regression)

정의

- 분석하고자 하는 대상들이 두 집단 혹은 그 이상의 집단으로 나누어진 경우 개별 관측치 들이 어느 집단으로 분류 될 수 있는가를 분석 , 예측하는 모형을 개발하는 대표적인 통계적인 알고리즘

- 로지스틱 회귀분석은 목적, 절차에 있어서 일반 회귀분석과 유사 그러나 , 종속 변수가 명목척도로 측정된 범주형 질적 변수인 경우에 사용한다는 것이 차이점

장점

1 . 간단한 방법 , 예측의 신뢰도를 평가하는 데 사용할 수 있는 가능성을 계산 할 수 있음

2. 결과변수가 연속변수가 아니거나 정규분포 하지 않는 경우 예측변수에 범주형 변수를 투입할 수 있음

단

1 . 선형관계에 있다고 가정해야 함

1. AIC 값

```
Step:  AIC=593.85
Outcome ~ Pregnancies + Glucose + BMI + DiabetesPedigreeFunction

                          Df Deviance    AIC
<none>                         583.85 593.85
- DiabetesPedigreeFunction  1   587.72 595.72
- Pregnancies               1   605.27 613.27
- BMI                       1   626.00 634.00
- Glucose                   1   691.67 699.67
```

(1) Outcome을 반응 변수로 사용하여 나머지 8개의 예측 변수와 함께 전체 모델을 만듬

(2) 가장 중요한 변수를 식별하기 위해 단계별 변수 선택 방법을 사용
→ AIC 를 선택 기준으로 선택한 최종 모델은 AIC 값이 593.85 의 가장 낮은 값을 갖는
로지스틱 회귀모형을 생성함

2. 로지스틱 회귀분석 결과 생성된 모델

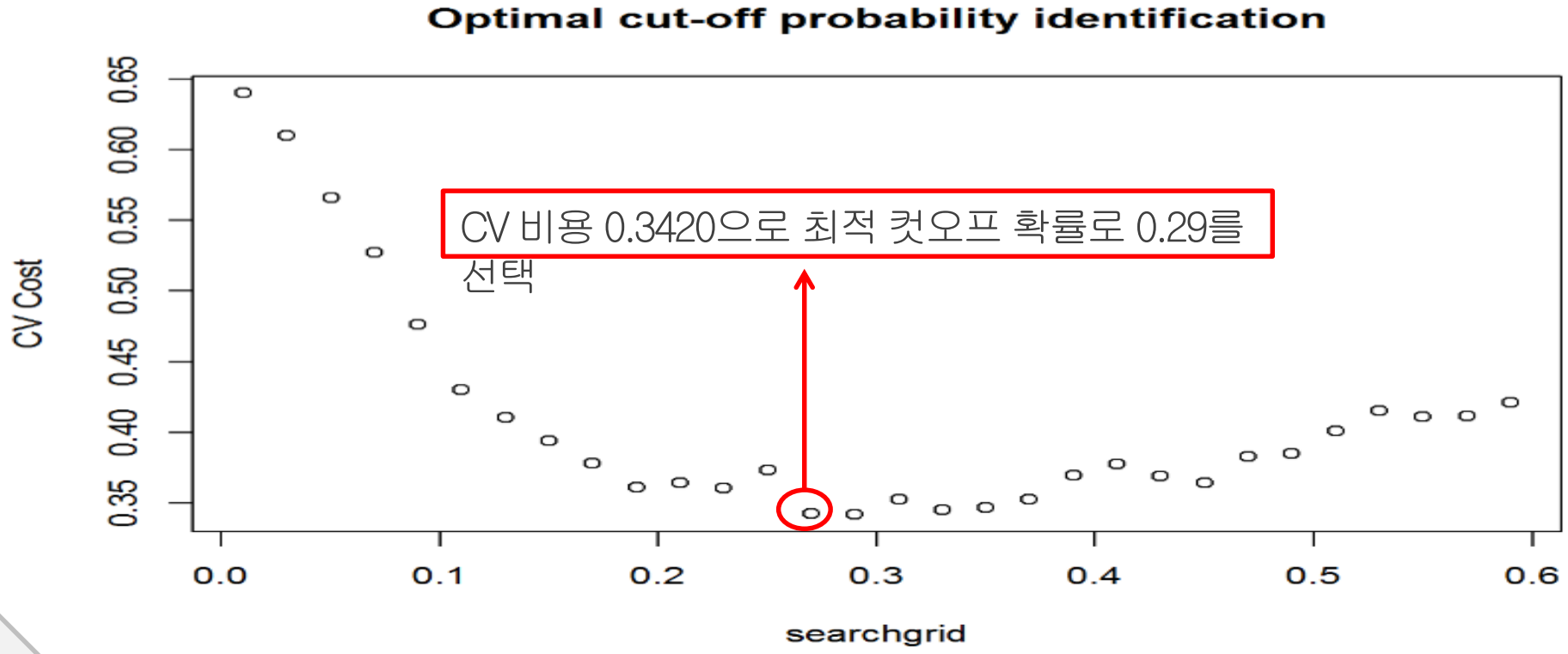
Outcome = −9.161189 + 0.035024 Glucose + 0.13832 Pregnancies + 0.100968BMI + 0.642931 DPF

```
Call:
glm(formula = Outcome ~ Pregnancies + Glucose + BMI + DiabetesPedigreeFunction,
    family = binomial, data = train.data)

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -9.161189   0.783300 -11.696  < 2e-16 ***
Pregnancies              0.138323   0.030588   4.522 6.12e-06 ***
Glucose                  0.035024   0.003797   9.223  < 2e-16 ***
BMI                      0.100968   0.016547   6.102 1.05e-09 ***
DiabetesPedigreeFunction 0.642931   0.328343   1.958   0.0502 .
```

3. 로지스틱 회귀 – 모수 조정

- 전체 모델을 사용하여 교차 검증을 실시 , 최적의 컷오프 확률을 식별함
- 불균형 비용 함수는 당뇨병 환자를 잘못 식별한 모델을 처벌하는 것으로 정의
- 정의된 비대칭 비용 함수 사용(거짓 양성의 경우 2:1): Flase Negatives)

→ 최적의 컷오프 확률이 계산됨 , 최적의 컷오프 확률은 모델이 가장 낮은 판별율을 가질 확률



Modeling

의사결정 나무 (Decision Tree)

정의

- 의사결정 규칙을 나무구조로 도표화하여 분류와 예측을 수행하는 방법

장점

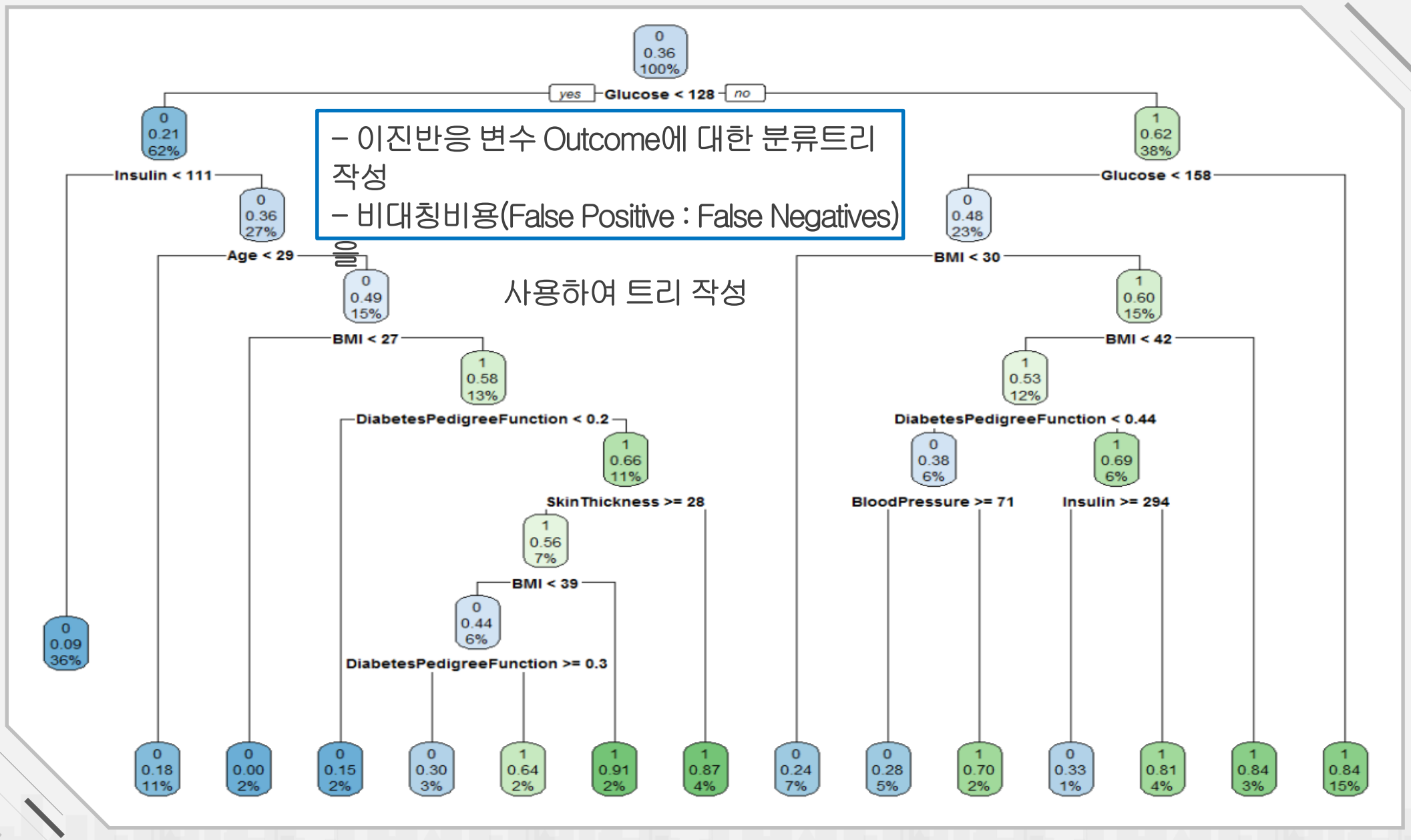
1. 나무구조로 표현되어 모형을 사용자가 쉽게 이해할 수 있음
2. 두 개 이상의 변수가 결합하여 목표변수에 어떻게 영향을 주는지 쉽게 알 수 있음 (교호작용 효과)
3. 선형성 , 정규성 , 등분산성 등의 가정이 필요하지 않음

단

1. 분석용 자료에만 의존하므로 새로운 자료의 예측에서는 불안정할 가능성이 높음
2. 연속형 변수를 비연속적 값으로 취급 → 분리의 경계점 부근에서 예측오류가 클 가능성이 높음
3. 선형 모형에서 주 효과는 다른 예측변화와 관련시키지 않아도 각 변수의 영향력을 해설 할 수 있지만 , 의사나무는 그렇지 않다.

Modeling

의사결정 나무 (Decision Tree) _ R 결과값

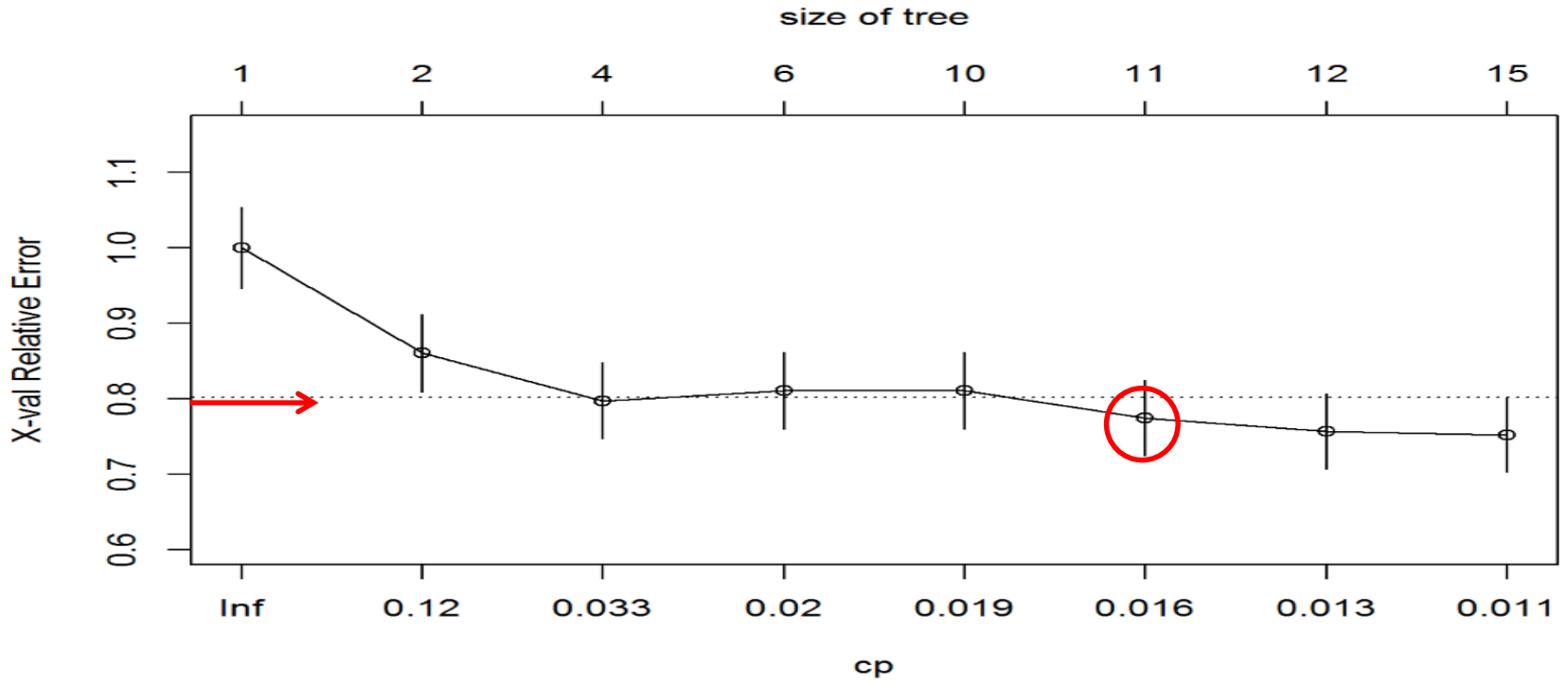


- 이진반응 변수 Outcome에 대한 분류트리 작성
- 비대칭비용(False Positive : False Negatives)을 사용하여 트리 작성

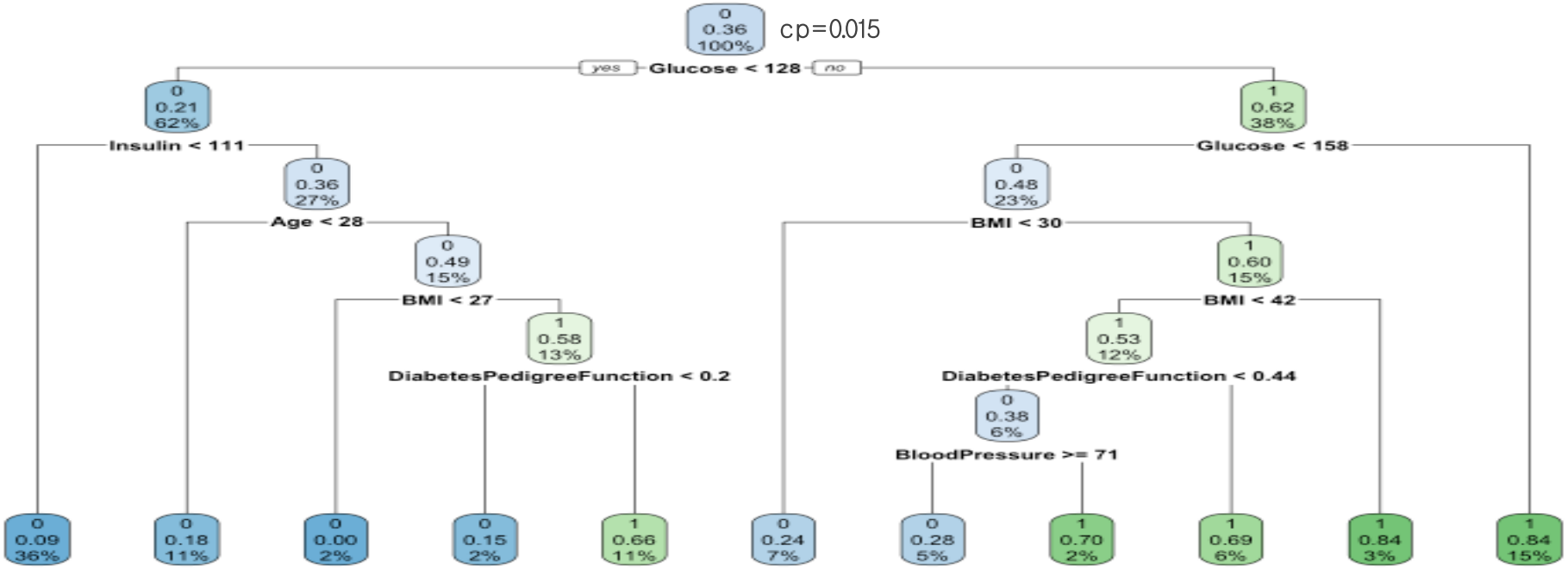
CP 조정

상대적 오차가 크게 감소하지 않는 복잡도 매개 변수 값은 최종 나무의 Cp(트리의 복잡성 매개변수)로 선택

Cp값 0.015를 사용하여 의사 결정 나무를 폐기함



가지치기 시행 후 분류 트리



포도당이 당뇨병 발생에 가장 큰 영향을 미쳤던 것을 의미
인슐린, 나이, 체질량지수, 당뇨병 가족력, 혈압에 따라서도 당뇨병 발생 유무에 조금은 관련이 있다는 것을 확인

Modeling

랜덤 포레스트 (Random Forest)

정의

- 하나의 데이터를 랜덤 샘플링하여 다수의 의사결정 나무를 만듬
  → 만들어진 의사결정나무들의 결과들을 모아 다수결로 최종 결과를 도출하는 알고리즘
- 집단학습을 기반으로 분류 , 회귀 , 클러스터링 등을 구현하는 앙상블 학습방법의 일종

장점

1 . 성능이 좋으며 정확도가 높음
2. 간편하며 빠름
3. 큰 데이터 셋 에서도 잘 사용되며 많은 입력 변수들을 다룰 수 있음

단

1 . 속도와 메모리의 비용이 큼
2. 과적합이 발생할 수 있음

랜덤 포레스트 모델 출력

- 랜덤포레스트는 차례로 개별나무의 결정을 결합하여 최종 결정을 내림
- OOB(Out-of-Beg)의 오차율은 25.04 %

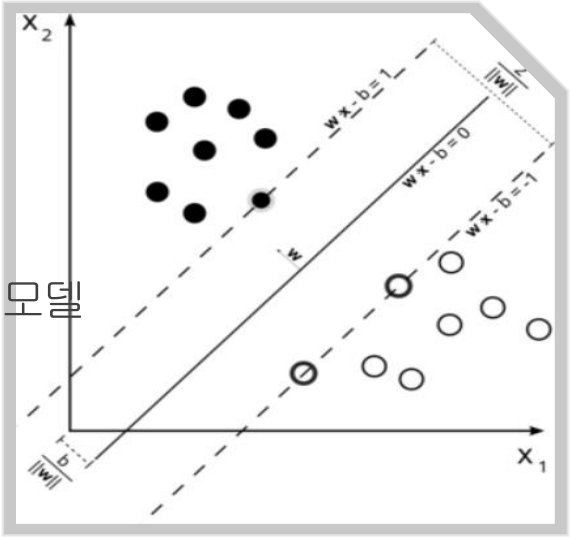
```
Call:
 randomForest(formula = Outcome ~ ., data = train.data, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 25.04%
Confusion matrix:
    0   1 class.error
0 331  62   0.1577608
1  92 130   0.4144144
```

Modeling

SVM (Support Vector Machine)

정의

- 지도학습의 기법 중 하나의 패턴인식 , 자료분석을 위한 지도학습 모델
  즉 , 2개의 범주를 분류하는 이진분류기



장점

1. 고차원에서 모두 효과적이다.
2. Decision fucntion에서 메모리 효율성
3. 차원수 > 데이터 수 일 때도 효율적이다
4. 커널함수 커스터마이징 가능

단

1.데이터가 너무 많으면 속도가 느리고
메모리적으로 힘듬
2.확률 추정치를 제공하지 않고 ,
5분할 교차검증을 사용하여
소비 리소스가 큼

SVM 모델 실행

```
svm.model <- svm(Outcome~., data = train.data,kernel = "radial", cost = 1, gamma = 0.1,probability = TRUE)
```

재구성된 SVM 모델 – 매개변수

```
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  0.2
      gamma:  0.125
```

비용 매개 변수는 전체적인 잘못 해석되는 오류 비율을
최소화하는 최상의 모델을 식별하도록 조정

→ cost : 2 , gamma : 0.125 로 잘못 해석된 값이 가장 낮아
최적의 매개변수가 됨

수정된 SVM 모델 실행

```
> summary(svm.model)

Call:
svm(formula = Outcome ~ ., data = train.data, kernel = "radial", cost = 0.2,
    gamma = 0.125, probability = TRUE)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  0.2
      gamma:  0.125

Number of Support Vectors:  398

 ( 202 196 )


Number of Classes:  2

Levels:
 0 1
```

05

평가 및 해석

5.1 평가

5.2 검증

5.3 결론

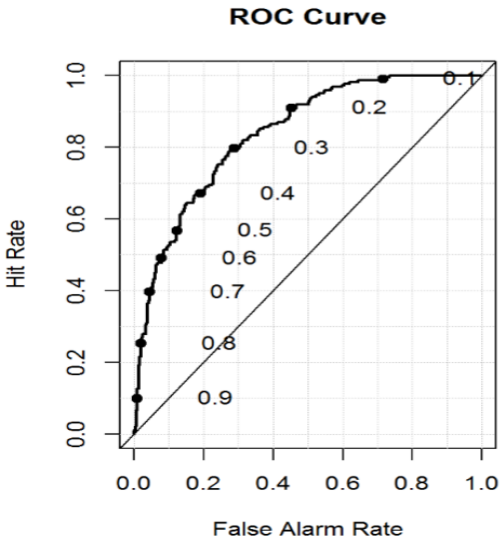
평가 및 해석

모델성능평가 – 로지스틱 회귀, 의사결정 나무

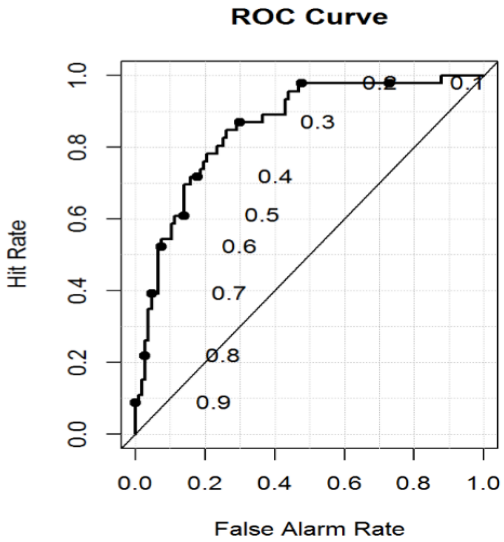
모델 성능 평가

1. 모델의 성능은 학습 데이터(데이터의 80%)와 평가 데이터(데이터의 20%)간에 비교
2. 최상의 모델은 비표본 데이터의 성능에 기초하여 선택되었으며 민감도가 의사 결정 기준으로 선택

로지스틱 회귀분

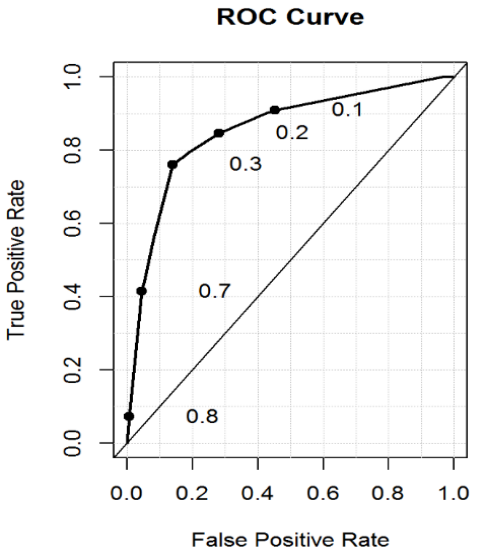


학습 데이터
AUC=0.839

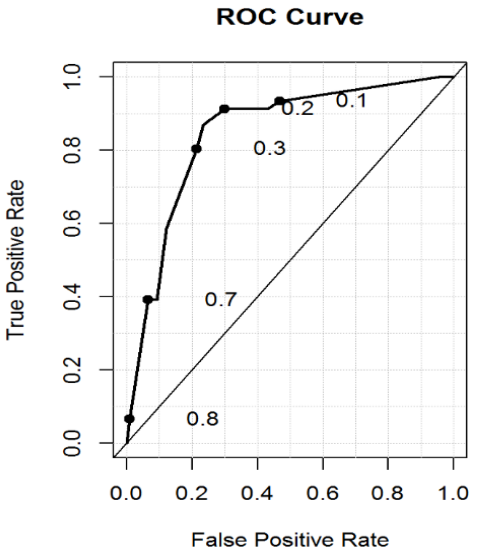


평가 데이터
AUC=0.857

의사결정 나무



학습 데이터
AUC=0.854

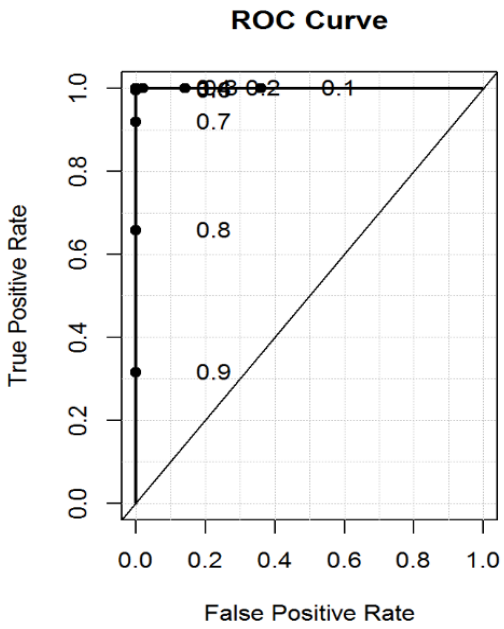


평가 데이터
AUC=0.848

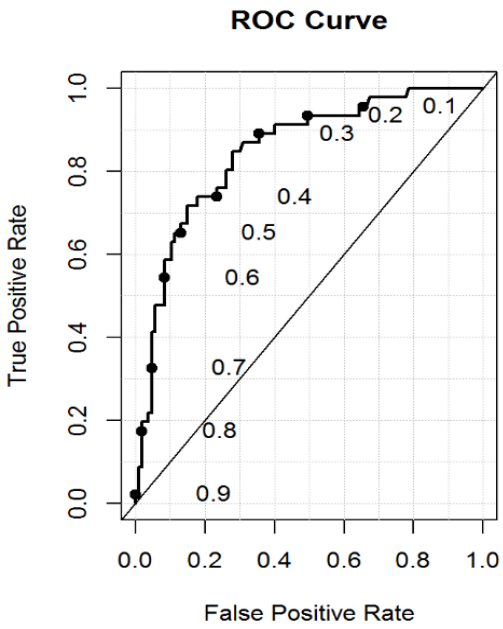
평가 및 해석

모델성능평가 – 랜덤 포레스트 , SVM

랜덤 포레스트

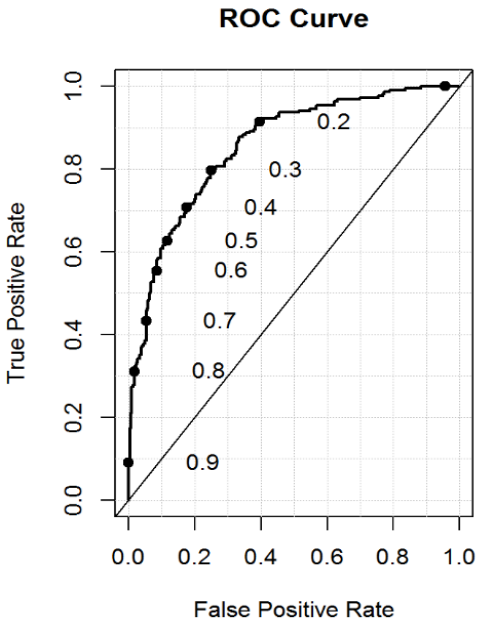


학습 데이터
AUC=1

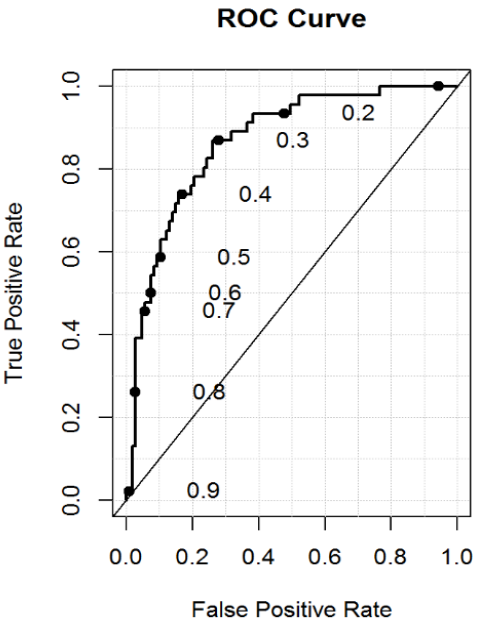


평가 데이터
AUC=0.847

SVM



학습 데이터
AUC=0.855



평가 데이터
AUC=0.866

평가 및 해석

최적의 성능 모델 결정 – 교차 검증

- 데이터 분할로 인한 무작위성의 영향을 제거하기 위해 최적의 모델을 결정하기 위하여
  5배의 교차 검증을 사용

교차 검증된 민감도	
[1] 0.7984117	로지스틱 회귀분석
[1] 0.7715438	분류나무
[1] 0.59497	랜덤 포레스트
[1] 0.5059086	SVM

- 로지스틱 회귀모델은 가장 높은 민감도를 보임

→ PIMA 여성 인디언의 당뇨병 발생을 예측하는 데
가장 적합한 모델로 선택

결론

연구 결론

PIMA인디언 여성의 당뇨병 발생을 예측하기 위한
최상의 모델을 식별하기 위해
로지스틱 회귀 분석, 분류 트리, 랜덤 포리스트 및 SVM과 같은
분류 모델을 구축 , 평가

→ 민감도의 교차 검증된 성과 측정에서, **로지스틱 회귀 분석 모델**은
가장 우수한 성과/예측 모델로 결론지음

Q & A

감사합니다

Thank you