피마 인디언 당뇨병 자료를 통한 데이터 분석

이지은[1)]

요약

제안된 방법은 가장 좋은 분류기는 로지스틱 회귀이며 ROC곡선과 오분류율을 통해서 이를 확인하였다.

주요용어 : 로지스틱 회귀, 의사결정나무, 단순 베이즈 분류, k-근방 분류, k-평균 군집법, 계층적 군집분석, ROC곡선, 오분류율.

1. 서론

당뇨병은 인슐린의 분비량이 부족하거나 정상적인 기능이 이루어지지 않는 등의 대사질환이다. 혈중 포도당 농도가 높은 것이 특징이다. 당뇨병은 제1형과 제2형으로 구분되는데, 제1형 당뇨병은 이전에 '소아 당뇨병'이라고 불렸었으며, 인슐린을 전혀 생산하지 못하는 것이 원인이 되어 발생하는 질환이다. 당뇨의 대부분은 인슐린이 상대적으로 부족한 제 2형 당뇨병이다. 제2형 당뇨병은 인슐린 저항성(혈당을 낮추는 인슐린 기능이 떨어져 세포가 포도당을 효과적으로 연소하지 못하는 것)을 특징으로 한다.

당뇨병 환자는 전 세계적으로 약 1억 명으로 추산되며, 미국은 전 인구의 약 6%가 당뇨병을 가지고 있는 것으로 추정된다. 선진국과 여러 국가에서 당뇨병이 발생하는 빈도가 늘어가는 반면 당뇨병의 주요 원인은 명확하게 밝혀지지 않았다. 특히 미국 애리조나의 몽고족 선조의 피마 인디언 부족이 제2형 당뇨 발병률이 가장 높은 지역이며 피마 인디언 당뇨 데이터는 확실한 제 2형 당뇨의 사례가 될 것이다.
제2형 당뇨병은 식생활의 서구화에 따른 고열량, 고지방의 식단, 운동 부족, 스트레스 등 환경적인 요인이 크게 작용하는 것으로 보이지만, 이 외에도 특정 유전자의 결함에 의해서도 당뇨병이 생길 수 있으며 수술, 감염, 약제에 의해서도 생길 수 있다.

당뇨병의 정확한 판정과 예측을 위해 당뇨병의 요인 중 주요한 요인들을 로지스틱 회귀, 의사결정나무, 단순 베이즈 분류, k-근방 분류, k-평균 군집분석, 계층적 군집분석등을 통하여 알아볼 것이다.

[1)]30019 세종시 세종로 2511, 고려대학교 공공정책대학 경제통계학부 빅데이터전공 학사과정, E-mail : kamiu199967@korea.ac.kr

2. 분석 데이터

피마 인디언 당뇨병 자료는 1965년부터 당뇨병과 소화기 및 신장 질환 국립 연구소(National Institute of Diabetes and Digestive and Kidney Diseases)에서 조사해왔다. 데이터의 목적은 데이터의 특정 변수 값을 기반으로 환자의 당뇨 진단 예측력을 높이기 위함이다. 피마 인디언 자료의 모든 환자는 21 세 이상의 피마 인디언 부족 여성을 대상으로 진행하였다. 자료는 9개의 변수와 768개의 환자 자료로 이루어져있으며 이 중 결과 변수인 Outcome은 당뇨이면 1, 당뇨가 아니면 0을 갖는 범주형 변수이다. 나머지 8개의 원인 변수는 Pregnancies(이전 임신의 횟수), Glucose(식후 2시간 뒤 포도당 부하 수치 검사의 혈당 농도), BloodPressure(확장기 혈압), SkinThickness(삼두근 피부 주름 두께), Insulin(혈청 인슐린), BMI(체질량 지수), DiabetesPedigreefunction(당뇨 내력 가중치 값), Age(나이)로 구성되어 있다. 피마 인디언 당뇨 자료를 5대5로 훈련자료, 시험자료로 무작위 분리한 후 분석할 것이다.

3. 분석 방법

3.1 로지스틱회귀

결과 변수인 Outcome은 '1'은 '당뇨', '0'이면 '당뇨가 아님'으로 나누어지는 이항 변수인데 이때, p(x)는 x일 때 당뇨 확률을 의미하며 입력 변수 $x_1, x_2, ..., x_p$에 대하여 다중 로지스틱 회귀모형은 아래와 같다.

$$P(Y=1|x) = p(B_0 + B_1 x + ... + B_p x_p) \quad (1)$$

이때, $p(x) = \exp(x)/1 + \exp(x)$이다.

회귀계수인 $B_0, B_1, .., B_p$는 최대우도법으로 추정되며 우도 비를 이용한 우도비 검증으로 유의성을 검증한다. 변수 선택기준은 AIC를 기준으로 후진소거법을 이용한다.

도출된 로지스틱 회귀모형은 결과 변수인 Y가 '당뇨'일 확률을 추정하는데 '당뇨'는 '1', '당뇨가 아님'은 '0의 값을 가질 때, $P(Y=1|X=x)$를 추정 한다. 이때 0.5를 절단값으로 정하여 $P(Y=1|X=x)$가 0.5보다 크면 Y=1, 0.5보다 작으면 Y=0으로 분류한다.

이후 오분류율과 ROC곡선을 구하여 분류기의 성능을 다른 분류기와 비교한다.

3.2 의사결정나무

의사결정나무는 데이터의 특징에 대한 질문을 하면서 응답에 따라 데이터를 분류해가는 알고리즘 이다. 즉 불순 도를 기준으로 사용하여 노드를 재귀적으로 분할하면서 나무 모델을 만든다. 주어진 입력 값에 대하여 출력 값을 예측하는 모형이며 if-then 방식이므로 해석과 구현이 쉬운 장점이 있다. 하지만 연속형 출력변수를 갖는 회귀에서는 예측력이 감소한다.

분류나무에서 카이제곱 통계량, 지니지수, 엔트로피 지수등을 사용하여 불순 도를 측정하고 나무를 성장시킨다. 이때 유의하지 않은 마디를 제거하는 방식으로 가지치기를 진행할 것이다.

$\hat{p}_{mk}$가 최종모형의 영역 R_m에 속하는 자료중 출력변수의 범주가 k인 자료의 비율일 때,

$x \in R_m$이면 예측 값은 아래와 같이 주어진다.

$$\hat{y} = \arg\max_k \hat{p}_{mk} \quad (2)$$

분류나무 모형을 만든 후 노드의 유의성을 기준으로 가지치기 하여 분석 할 것이다. 또한 분류기의 성능은 오분류율과 ROC곡선으로 비교하여 진행한다.

3.3 단순베이즈 분류, k-근방분류

단순 베이즈 분류는 입력변수의 조건부분포가 서로 독립이라는 단순 베이즈 정리를 이용한 확률 분류기의 일종이다. 데이터를 범주로 분리할 때 많이 사용된다. 또한 고차원 확률 문제를 반복적인 일차원 확률 추정 문제로 바꿀 수 있는 장점이 있다.

입력변수인 $x = (x_1, x_2, .., x_p)$일 때, 사후 확률은 아래와 같다.

$$P(Y = k | X_1 = x_1, .., X_p = x_p) \propto P(Y = k) \Pi_{j=1}^{p} P(X_j = x_j | Y = k) \quad (3)$$

단순 베이즈 분류는 훈련자료를 이용하여 모든 j와 k에 대하여 추정 값

$\hat{P}(Y = k), \hat{P}(X_j = x_j | Y = k)$를 얻은 후 Y의 추정 값을 구한다. Y의 추정 값은 아래와 같다.

$$\hat{Y} = \arg max(\hat{P}(Y = k) \Pi_{j=1}^{p} \hat{P}(X_j = z_j | Y = k)) \quad (4)$$

k-근방 분류는 통계적 모형을 적합하지 않고 메모리 기반으로 분류하는 방법이며 x가 주어진 후, x와 가장 가까운 k개의 훈련 자료 점의 y값을 비교하여 가장 많은 클래스로 출력변수 값을 예측하는 원리이다. N(x)을 x와 가장 가까운 k개의 훈련자료 집합인 k-근방으로 두면

$$\hat{y}(x) = \arg\max_{l \in -1,+1} \sum_{x_i \in N(x)} I(y_i = l) \quad (5)$$

이때, k가 증가하면 편의는 커지나 분산은 감소하고 모형의 복잡 도는 감소한다.

이 두 분류기 또한 ROC곡선과 오분류율을 통해 분류기의 성능을 알아볼 것이다.

3.4 k-평균군집법, 계층적군집법

k-평균 군집법은 군집수 k에 대한 군집내 거리 제곱 합을 최소화 하는 것이 목적인 군집분석법이다. 유클리드 거리를 이용하여 데이터를 유사한 특성의 k개 군집으로 분리한다. 이때 가장 최적의 k를 도출하여 사용할 것이다. 계층적 군집법보다 계산 량이 적고 빠르게 처리가 가능하다. 하지만 과정에서 평균을 이용해야 하므로 사용되는 모든 변수가 연속형 변수여야 한다. 따라서 피마

인디언 데이터에서 이진 범주형 변수인 Outcome을 제외한 변수를 분석에 이용할 것이다.

계층적 군집법은 병합방법과 분할 방법이 있는데 병합 방법 중에서도 두 군집이 병합되었을 때 오차제곱합의 증가분 기반으로 군집간 유사성을 이용하는 ward방식을 이용할 것이다. 계층적 군집법은 군집 결과를 나무구조의 덴드로그램을 통하여 간단하고 해석이 쉽게 표현할 수 있으며 전체 군집간 구조적 관계를 쉽게 살펴볼 수 있다.

4. 분석 결과

4.1 로지스틱회귀

전체 피만 인디언 당뇨 데이터에서 유의한 변수를 알기 위해 유의확률을 확인하였다.

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -8.4046964  0.7166359 -11.728  < 2e-16 ***
Pregnancies              0.1231823  0.0320776   3.840 0.000123 ***
Glucose                  0.0351637  0.0037087   9.481  < 2e-16 ***
BloodPressure           -0.0132955  0.0052336  -2.540 0.011072 *
SkinThickness            0.0006190  0.0068994   0.090 0.928515
Insulin                 -0.0011917  0.0009012  -1.322 0.186065
BMI                      0.0897010  0.0150876   5.945 2.76e-09 ***
DiabetesPedigreeFunction 0.9451797  0.2991475   3.160 0.001580 **
Age                      0.0148690  0.0093348   1.593 0.111192
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 993.48  on 767  degrees of freedom
Residual deviance: 723.45  on 759  degrees of freedom
AIC: 741.45

Number of Fisher Scoring iterations: 5
```

첫 번째 변수= x_1인 ‘이전 임신 횟수’는 독립 표본 t검정을 실시한 결과 유의확률이 0.001정도로 유의수준인 0.05보다 작으므로 회귀 계수가 0이라는 귀무가설을 기각할 수 있다. 따라서 임신 횟수는 당뇨병의 결과에 영향을 주는 변수라고 할 수 있다. x_2인 ‘식후 2시간 뒤 포도당 부하 수치 검사의 혈당 농도’의 유의확률은 0.001보다 작고 유의수준인 0.05보다 작으므로 회귀 계수가 0이라는 귀무가설을 기각한다. 따라서 ‘식후 2시간뒤 포도당 부하 수치 혈당 농도’는 당뇨병의 결과에 영향을 주는 변수라고 할 수 있다. x_3인 ‘확장기 혈압’은 독립 표본 t검정을 실시한 결과

유의확률이 0.01이며 유의수준인 0.05보다 작으므로 회귀 계수가 0이라는 귀무가설을 기각할 수 있다. 따라서 확장기 혈압은 당뇨병의 결과에 영향을 주는 변수라고 할 수 있다. x_4인 '삼두근 피부 주름 두께'는 유의확률이 0.92로 유의수준인 0.05보다 훨씬 크므로 귀무가설을 기각할 수 없으며 유의한 변수라고 할 수 없다. x_5인 '혈청 인슐린 수치'는 유의확률이 0.18정도로 유의수준인 0.05보다 크므로 귀무가설을 기각할 수 없다. 또한 당뇨병의 결과에 유의하지 않은 변수이다. x_6인 체질량 지수는 유의확률이 0.05보다 훨씬 작으므로 회귀계수가 0이라는 귀무가설을 기각할 수 있다. 즉 체질량 지수는 당뇨병 결과에 유의한 변수이다. x_7인 '당뇨 내력 가중치 값'은 유의확률이 0.0015정도로 유의수준인 0.05보다 작으므로 귀무가설을 기각할 수 있다. '당뇨 내력 가중치 값'은 당뇨병 결과에 대하여 유의한 변수이다. x_8인 나이는 0.111정도로 유의수준인 0.05보다 크므로 귀무가설을 기각할 수 없고 유의한 변수가 아니다.

이후에 피마 인디언 당뇨병 자료를 384개씩 훈련자료와 시험자료로 나눈 후 로지스틱 회귀 분석을 실시하였다.

```
Call:
glm(formula = Outcome ~ ., family = "binomial", data = dia.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4941  -0.8016  -0.4646   0.8369   2.6797

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -7.4641634  0.9550497  -7.815 5.48e-15 ***
Pregnancies              0.1239232  0.0436103   2.842  0.00449 **
Glucose                  0.0296152  0.0050544   5.859 4.65e-09 ***
BloodPressure           -0.0111196  0.0067691  -1.643  0.10044
SkinThickness            0.0037905  0.0093443   0.406  0.68500
Insulin                 -0.0003177  0.0012164  -0.261  0.79397
BMI                      0.0854355  0.0209576   4.077 4.57e-05 ***
DiabetesPedigreeFunction 0.7216072  0.4040498   1.786  0.07411 .
Age                      0.0080541  0.0123328   0.653  0.51371
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 502.69  on 383  degrees of freedom
Residual deviance: 387.47  on 375  degrees of freedom
AIC: 405.47

Number of Fisher Scoring iterations: 5
```

회귀계수의 부호가 양수이면 변수의 값이 높아질수록 당뇨일 확률이 커진다.

예시로 임신 횟수의 회귀계수는 0.1239이고 오즈 비는 1.13로 임신횟수가 한 단위 증가할 때마다 오즈가 1.13배 증가하는 것을 알 수 있다. '식후 2시간 뒤 포도당 부하 수치 검사의 혈

당 농도'의 회귀계수는 0.0296으로 오즈 비는 1.03이다. 확장기 혈압의 회귀 계수는 -0.01111로 오즈 비는 0.988이다. 즉, 확장기 혈압이 한 단위 증가할수록 당뇨의 가능성은 더 하락하게 된다. 삼두근 피부 주름 두께의 회귀 계수는 0.003으로 오즈 비는 1.003이다. 혈중 인슐린 수치의 회귀 계수는 -0.0003이고 오즈 비는 0.9997이므로 혈중 인슐린 수치가 한 단위 증가하면 아주 미세하게 당뇨의 가능성은 하락하는 것을 의미한다. 체질량 지수의 회귀 계수는 0.085로 오즈 비는 1.088이다. 내력 가중치 값의 회귀 계수는 0.721로 오즈 비는 2.0564이다. 이는 당뇨 내력이 한 단위 증가하면 당뇨의 가능성은 약 2배 정도 증가하는 것을 의미한다. 마지막으로, 나이의 회귀계수는 0.008이며 오즈 비는 1.008이다.

최적의 모형을 선택하기 AIC를 기준으로 후진소거법을 이용한다.

```
Step:  AIC=400.02
Outcome ~ Pregnancies + Glucose + BloodPressure + BMI + DiabetesPedigreeFunction

                         Df Deviance    AIC
<none>                        388.02 400.02
- BloodPressure           1   390.41 400.41
- DiabetesPedigreeFunction 1   391.65 401.65
- Pregnancies             1   401.43 411.43
- BMI                     1   410.12 420.12
- Glucose                 1   439.77 449.77
```

AIC값이 가장 작은 최적의 모형은 임신 횟수, 식후 2시간 뒤 포도당 부하 수치 검사의 혈당 농도, 혈압, 체질량 지수, 당뇨 내력 가중치 값을 사용한 모형이다. 이 이외의 변수는 제거하였다.

추정된 로지스틱 회귀 모형은 아래와 같다.

$$\log(\frac{p}{1-p}) = -7.46 + 0.123x_1 + 0.029x_2 - 0.011x_3 + 0.085x_4 + 0.72x_5 \qquad (6)$$

이때, x_i는 i번째 입력 변수이다.

<표1>변수 선택전 오분류표

actual \ predicted	0	1
0	225	30
1	51	78

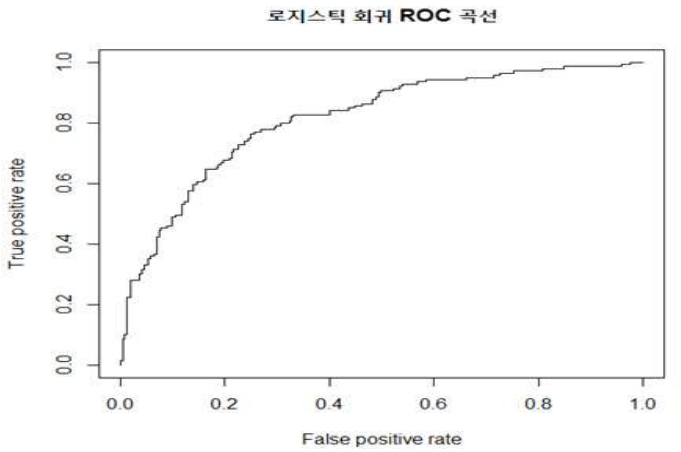
<표2>변수 선택후 오분류표

actual \ predicted	0	1
0	226	29
1	51	78

변수 선택 전 오분류율은 0.2109375이며 변수 선택 이후 오분류율은 0.208333으로 조금 감소한 것을 알 수 있다.

Roc 곡선으로 분류기를 평가할 때는 곡선 아래 면적인 Auc가 클수록 분류기의 예측력과 성능이 좋다. 로지스틱 회귀의 분류기 성능이 좋은 것을 알 수 있다.

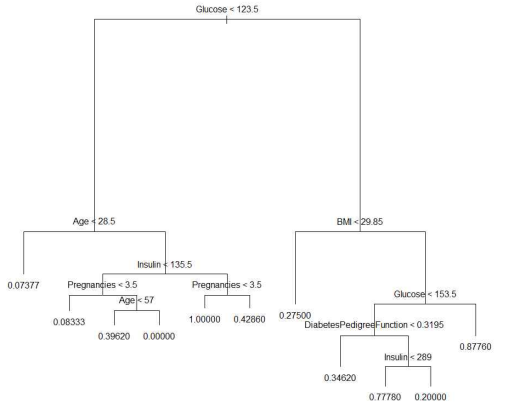
<그림1>로지스틱 회귀 모형 ROC곡선



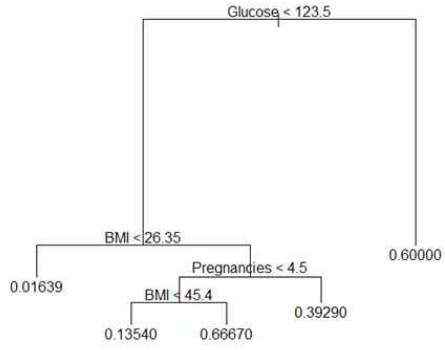
4.2 의사결정나무

전 처리된 훈련 데이터로 의사 결정 나무 모형을 만들고 성장 시켰다. 가지의 수는 모형의 복잡도를 의미하는데 너무 많으면 시험자료에 대입하였을 때 성능이 떨어지는 과대적합문제가 생길 수 있고 가지의 수가 너무 적으면 과소적합 문제가 발생한다.

유의한 노드만 남겨두고 유의하지 않은 노드는 snip.tree()로 제거하였다.



<그림2>가지치기 이전 의사결정나무



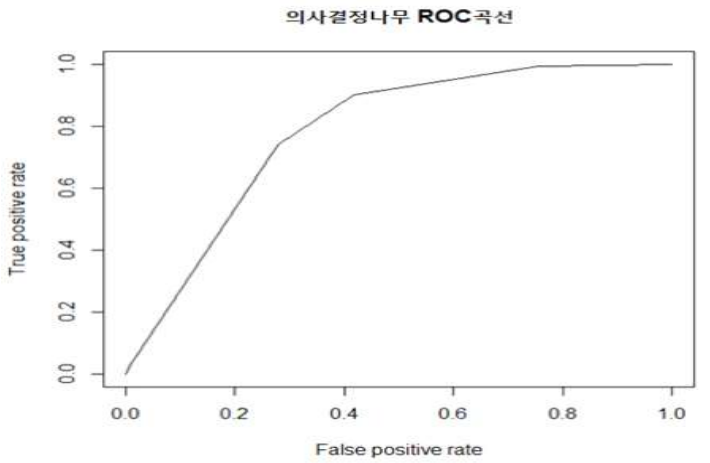
<그림3>가지치기 이후 의사결정나무

가지치기 이후 끝마디는 6개가 된 것을 알 수 있다. 적합된 트리 모델로 Outcome 변수를 예측한 후 실제 값과 비교하여 오분류표와 오분류율을 구하였다. 의사결정나무의 오분류율은 0.28125이다. 로지스틱 회귀보다는 오분류율이 크지만 ROC곡선으로 살펴본 분류기의 성능은 양호하다.

	predicted	
actual	0	1
0	186	69
1	39	90

<표3>의사결정나무 오분류표

<그림4>의사결정나무 ROC곡선



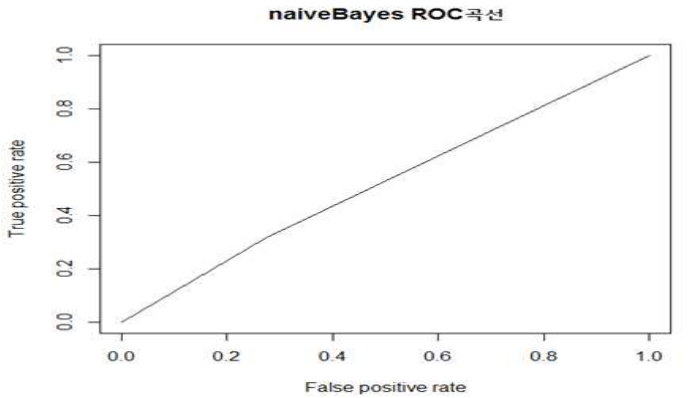
4.3 단순베이즈 분류, k-근방분류

단순 베이즈 가정아래서 훈련자료로 적합 시킨 단순 베이즈 분류 모델을 이용하여 시험자료의 Outcome 변수를 예측하였다. 예측 값과 실제 값을 비교할 때 오분류율은 0.1979167이다.

pred	0	1
0	226	47
1	29	82

<표4>단순 베이즈 오분류표

<그림5 >단순 베이즈 Roc곡선



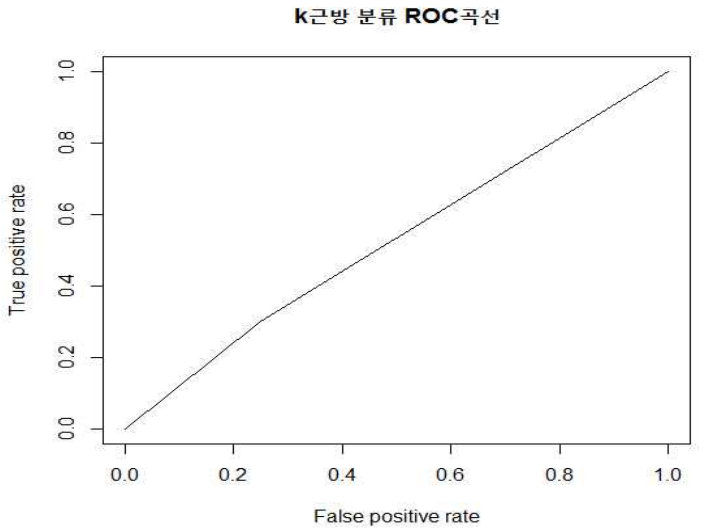
Roc곡선이 기울기가 1인 직선에 가까울수록 분류기의 성능과 예측력이 낮다. 그러나 단순 베이즈 분류의 ROC 곡선 기울기가 1인 직선에 상당히 가까운데 이는 데이터에 적합한 분류기가 아니거나 단순 베이즈 가정을 만족시키지 않기 때문이다

k-근방 분류의 경우, k가 너무 커지면 편의가 커지고 분산이 감소하기 때문에 가장 적합한 군집의 수 k를 구해야한다 1부터 10까지의 숫자중 tune.knn()을 이용한 결과, k=6일 때 가장 적합하다고 도출되었다. 6-근방일 때 오분류율은 0.2291667이다. 만들어진 6개의 근방 중에서 Outcome을 제외한 다른 변수들로 거리가 가까운 클래스를 구하여 출력변수인 Outcome값을 예측하는 방식이다.

	NA	
pred	0	1
0	224	57
1	31	72

<표5>k-근방 분류 오분류표

<그림6>k-근방 분류 ROC곡선



Roc곡선이 거의 기울기 1인 직선은 아니지만 직선에 가까운 정도인데 이는 k-근방 분류기의 예측력과 성능이 떨어지기 때문이다. 피만 인디언 당뇨 데이터 분석에는 매우 적합하지 않은 분류기임을 알 수 있다. 곡선 아래 면적인 Auc또한 0.5에 가까운 값이다.

단순 베이즈 분류와 k-근방 분류 모두 피마 인디언 당뇨병 데이터에는 적합하지 않은 분류기이다.

4.4 k-평균군집법, 계층적군집법

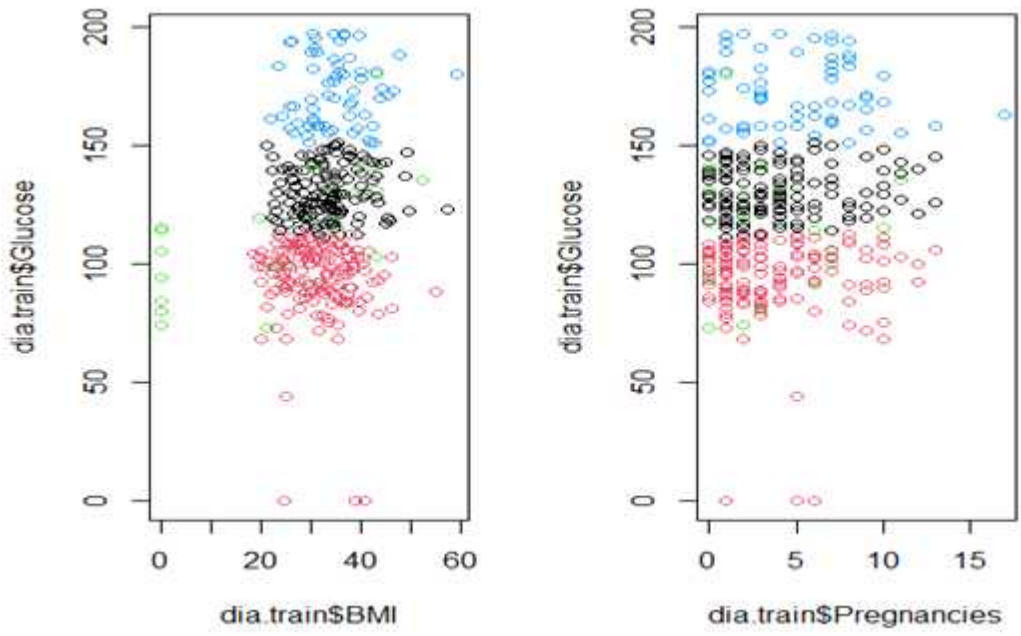
k-평균 군집법은 유클리드 거리를 기준으로 유사한 k개의 군집으로 분리한 후 관측치의 평균을 군집의 중심으로 잡는 방법이다. 평균을 구하기 위해서는 Outcome과 같은 이진 범주형 변수를 사용하지 못하므로 Outcome 변수를 제외한 데이터를 입력하였다. nbClust를 이용하여 최적의 군집 개수를 구한 결과 k=4가 도출되었다. 4개의 군집을 으로 분리하게 된다. nbClust결과를 살펴보면 11개의 수치가 k=4일 때 최적임을 내포하고 있다.

```
* Among all indices:
* 7 proposed 2 as the best number of clusters
* 1 proposed 3 as the best number of clusters
* 11 proposed 4 as the best number of clusters
* 1 proposed 9 as the best number of clusters
* 3 proposed 15 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  4
```

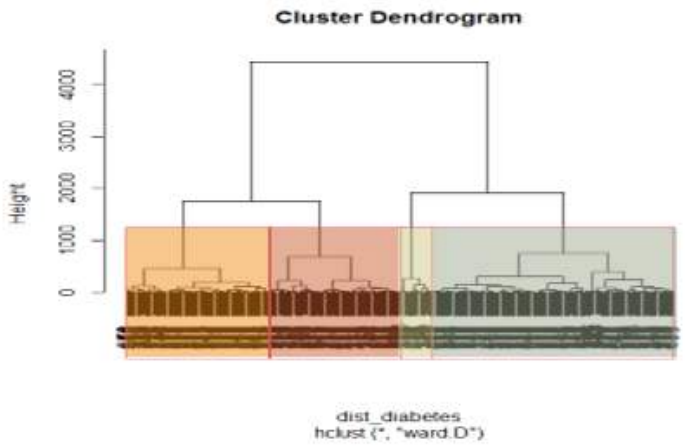
<그림7>k=4일 때 군집



가장 유의한 변수 2쌍인 '식후 2시간 뒤 혈당 부하 수치'와 '체질량 지수', '임신 횟수'와 '식후 2시간 뒤 혈당 부하 수치'를 뽑아서 군집으로 나타내었다. k=4이므로 유사한 군집별로 4개씩 나누어졌다. 군집에서 이탈한 수치도 보이나 전체적으로 군집별로 잘 분리되었음이 보인다.

계층적 군집법에서도 k-평균 군집법에서 구한 군집의 개수인 k=4를 사용하겠다. 계층적 군집법에는 병합방법과 분할방법이 있는데 이중에서도 병합방법인 ward방식을 사용하였다. 일단 유클리드 거리를 구한 후, 두 군집이 병합되었을 때 오차제곱합의 증가분 기반으로 군집 간 유사성을 이용하여 4개의 군집으로 분리한다. 아래 덴드로그램에서 군집의 분리를 확인할 수 있다.

<그림8>계층적 군집 덴드로그램

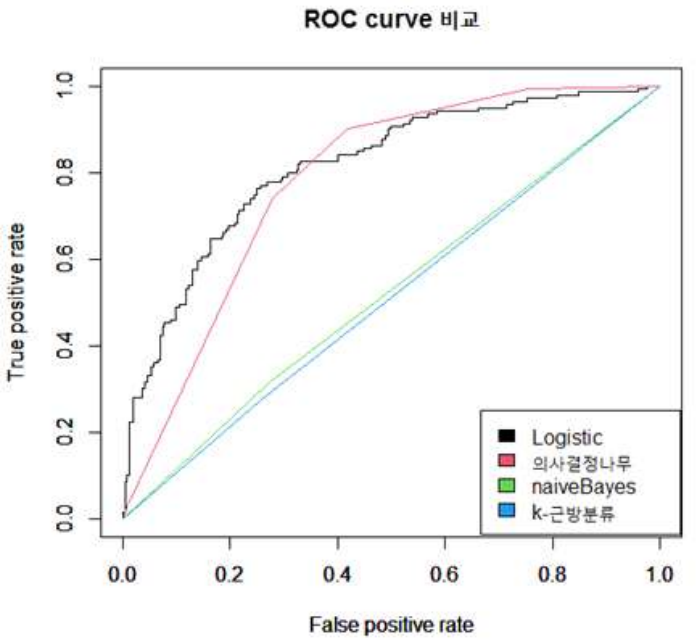


5. 결론

분류기 / 오분류율	로지스틱 회귀	의사결정나무	단순 베이즈 분류	k – 근 방 분류
	0.208333	0.28125	0.197916	0.229166

<표6>분류기별 오분류율

<그림9>분류기별 Roc곡선



Roc곡선과 오분류율로 살펴본 가장 성능이 좋은 분류기는 로지스틱 회귀 모델입니다. 오분류율은 단순 베이즈 분류기가 0.1979로 가장 작지만, Roc곡선 상에서 곡선아래의 영역인 Auc가 로지스틱 회귀에서 가장 크다. 따라서 로지스틱 회귀 모형이 최적의 분류기이다.

로지스틱 회귀 분석결과로 살펴본 피만 인디언 당뇨병에서 유의한 변수들은 '식후 2시간 뒤 포도당 부하 수치 검사의 혈당 농도', '체질량 지수', '임신 횟수', '당뇨 내력 가중치 값'이다. 이때 '식후 2시간 뒤 포도당 부하 수치 검사의 혈당 농도'인 Glucose변수는 당뇨병 판정에 사용되는 변수이므로 당연히 가장 유의하고 중요한 변수이다. 그 다음 '체질량 지수'가 높으면 비만의 가능성이 크므로 비만의 합병증인 당뇨의 가능성 또한 증가한다. 그 뒤로는 '임신 횟수'와 '당뇨 내력 가중치 값'순이다. 즉, 당뇨 내력은 영향은 있지만 크지 않으며 운동이나 올바른 식습관을 통해서 바꿀 수 있는 체질량 지수의 영향이 크다. 당뇨 내력이 있더라도 본인이 관리한다면 당뇨를 예방할 수 있다. '임신 횟수'의 영향은 여성이 임신을 많이 할수록 당뇨의 위험 또한 커지는 것을 의미한다. 나이는 유의하지 않은 변수이므로 당뇨는 나이에 상관없이 발병한다는 것을 알 수 있다.

References

[1] 이왕선 (2014). 피마 인디언 당뇨병 자료를 이용한 다중 로지스틱 회귀모형 비교, 6-16.
[2] 박창이, 김용대, 김진석, 송종우, 최호식 (2015). R을 이용한 데이터마이닝, 교우사.
[3] 서민구 (2014). R을 이용한 데이터 처리&분석 실무, 길벗사.
[4] Paul teetor (2012). R Cookbook, 프로그래밍인사이트.