R을 이용한 당뇨병 여부 예측하기

김동률

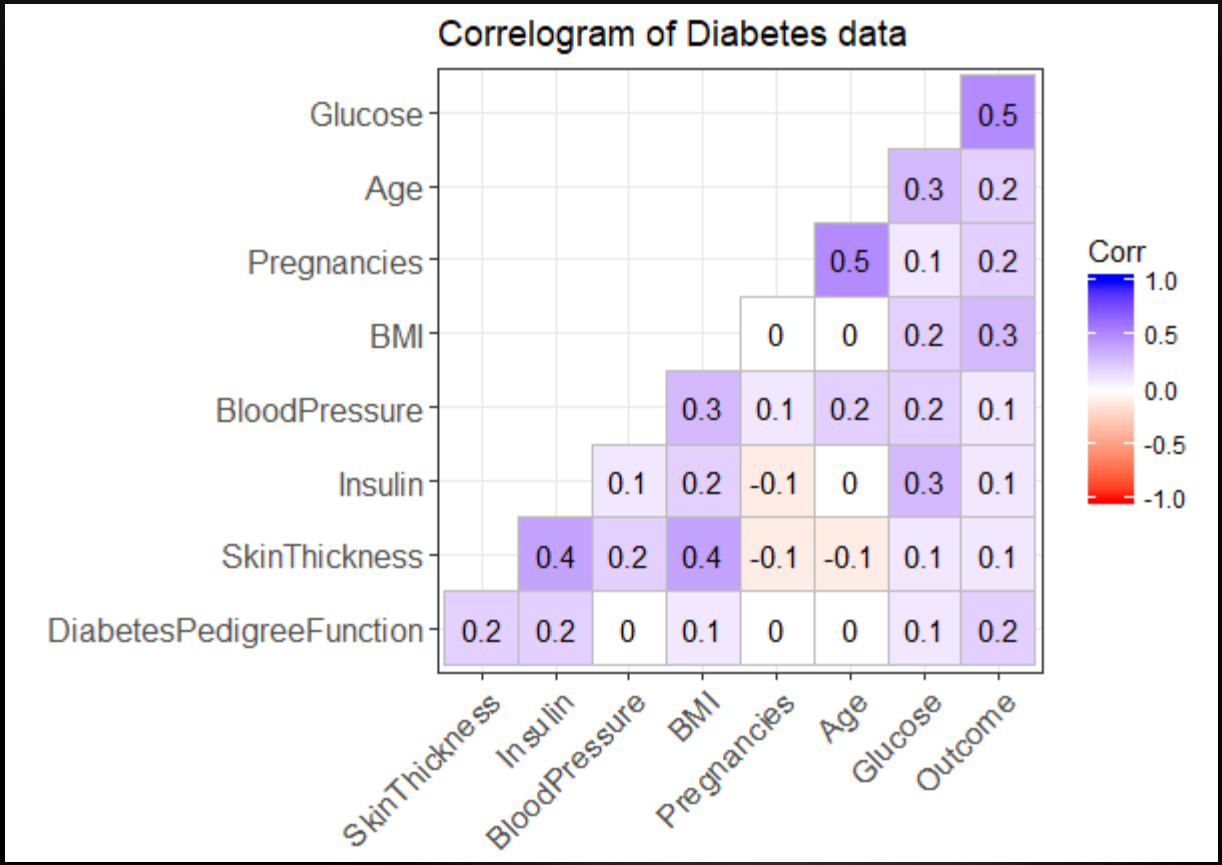
- 출처: 국립 당뇨병 학회 (캐글에서 데이터를 가져옴)
  (https://www.kaggle.com/uciml/pima-indians-diabetes-database/data)
- 분석 목적: 기존 환자의 당뇨병 진단 정보를 기반으로 새로운 환자의 당뇨병 여부를 예측
- 모든 환자는 피마 인디언 유산의 21 세 이상인 여성

DATA SET

-DATA SET은 여러 가지 의료 예측 변수와 하나의 목표 변수(Outcome)로 구성

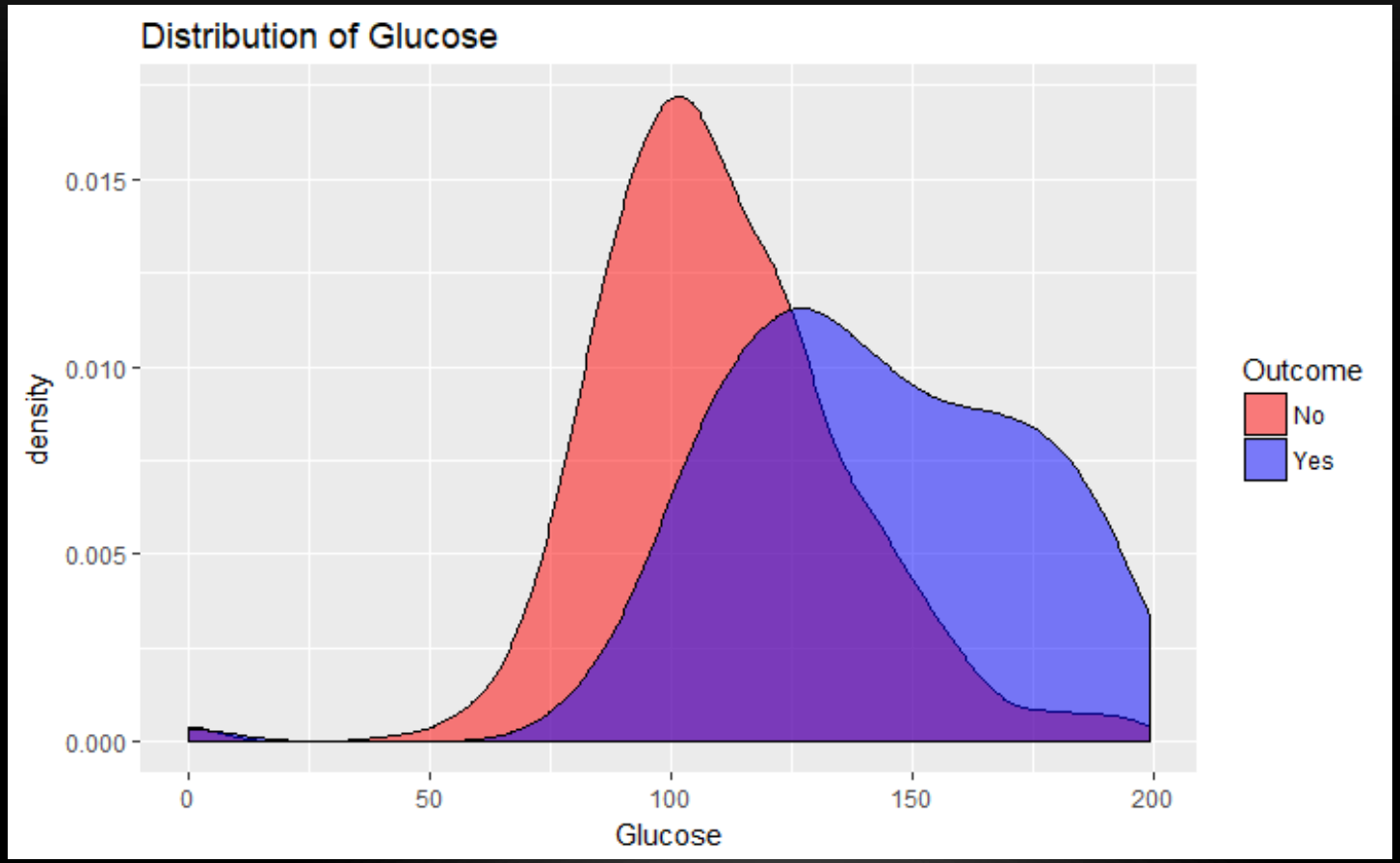
- Pregnancies : 임신 횟수
- Glucose : 포도당 내성 검사에서 2시간 동안의 포도당 농도
- BloodPressure : 혈압 (mm Hg)
- SkinThickness : 상완 삼두근 피부 두께(mm)
- Insulin : 2 시간 동안의 혈청 인슐린 (mu U/ml)
- BMI : 체질량 지수
- DiabetesPedigreeFunction : 당뇨 유전력
- Age : 나이
- Outcome : 당뇨병 진단 (YES: 당뇨병 O , NO : 당뇨병 X)

상관관계



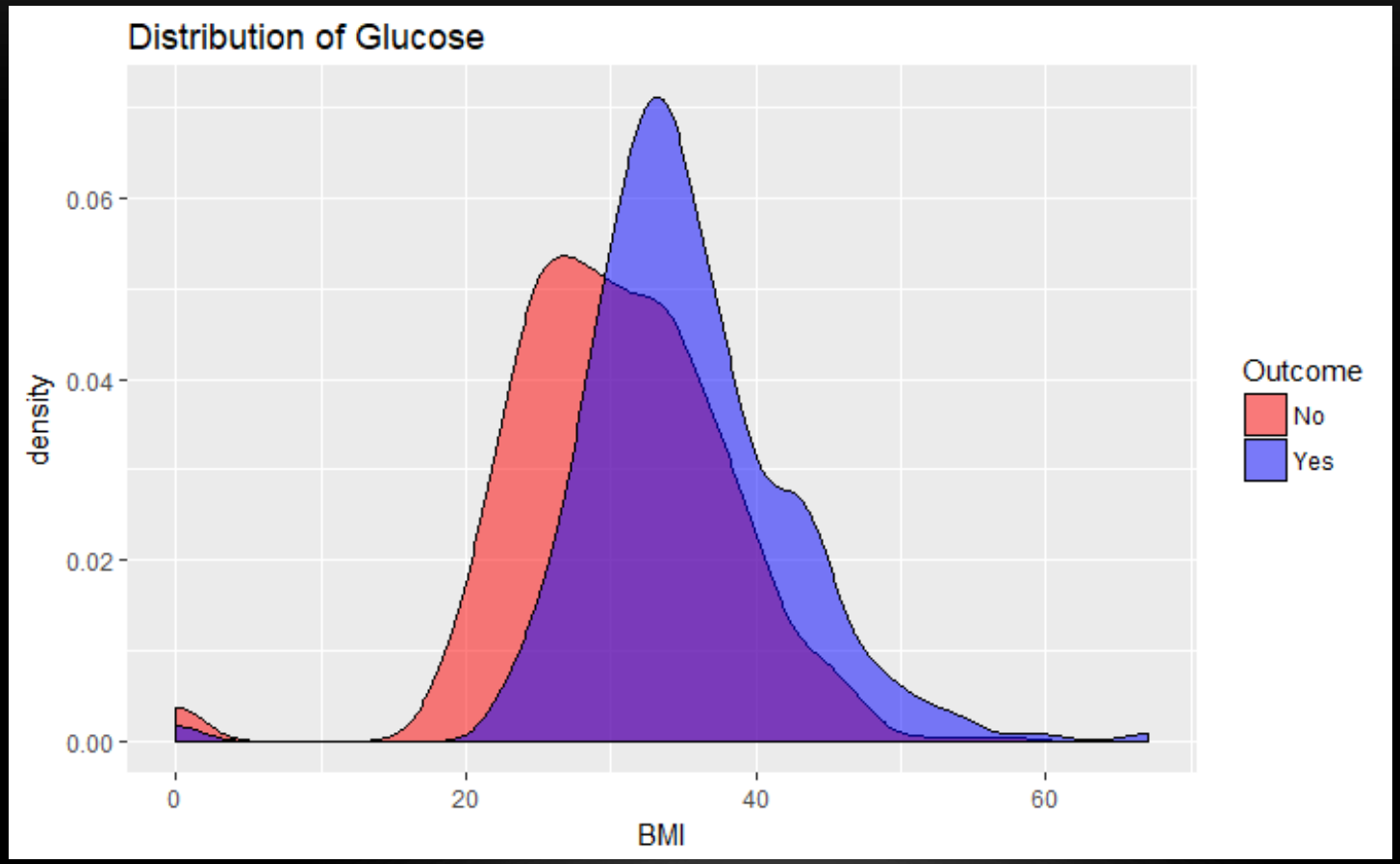
- 분석하기 전에 앞서 각 변수들간의 상관관계 파악
- 목표 변수 Outcome과 상관관계가 높은 변수는 Glucose(0.5), BMI(0.3) 임을 알 수 있음

‘GLUCOSE’와 ‘BMI’ 변수의 당뇨병 여부에 따른 그래프



- ggplot 함수로 시각화
- 당뇨병 여부에 따라 포도당 수치의 분포가 차이가 있다는 걸 알 수 있다.

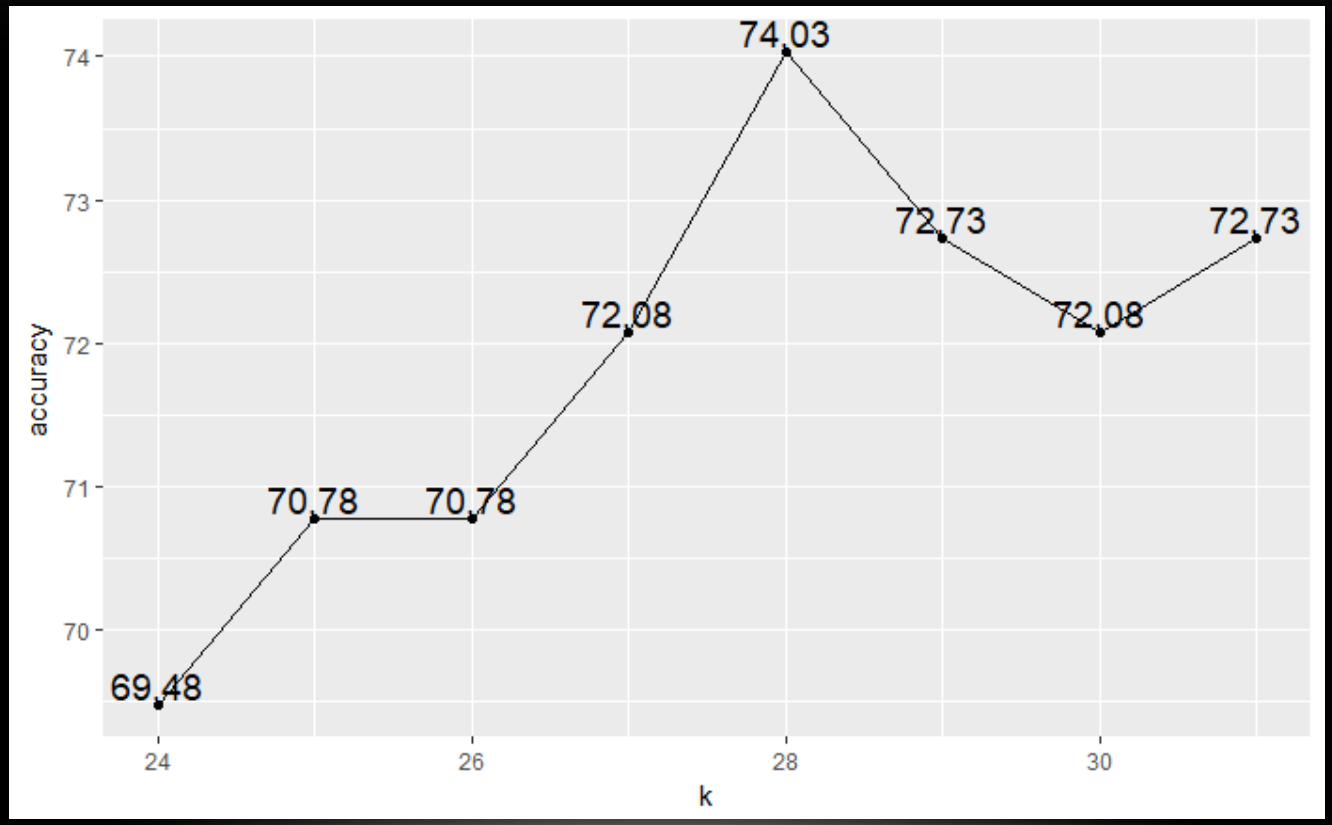
‘GLUCOSE’와 ‘BMI’ 변수의 당뇨병 여부에 따른 그래프



- ggplot 함수로 시각화
- 당뇨병 여부에 따라 BMI 수치의 분포가 차이가 있다는 걸 알 수 있다.

KNN 모델

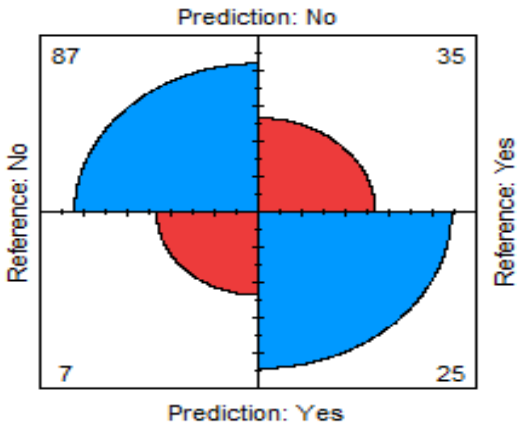
- k 값 지정 : $\sqrt{\text{데이터 수}} = \sqrt{768} = 27.71281$ ∴ k 는 27 근처 값으로 정의
- k 가 28 일 때, 정확도는 최대



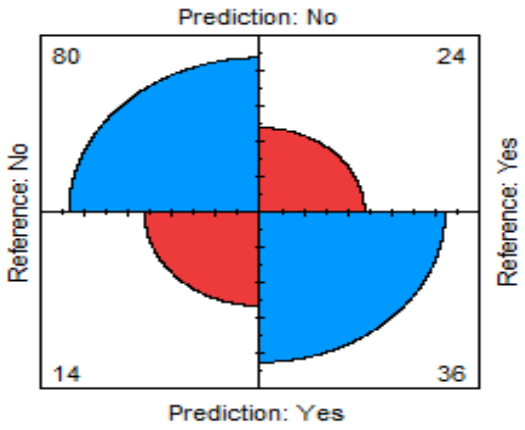
KNN 모델 – KNN, KNN 정규화, KNN 표준화 정확도 비교

- 트레인 데이터셋 (80%) 과 테스트 데이터 셋 (20%) 으로 분류한 뒤 knn 모델 적용
- 데이터의 정규화와 표준화에 따른 모델 정확도가 향상되는 점을 볼 수 있지만 정확도가 크게 차이 나지는 않음

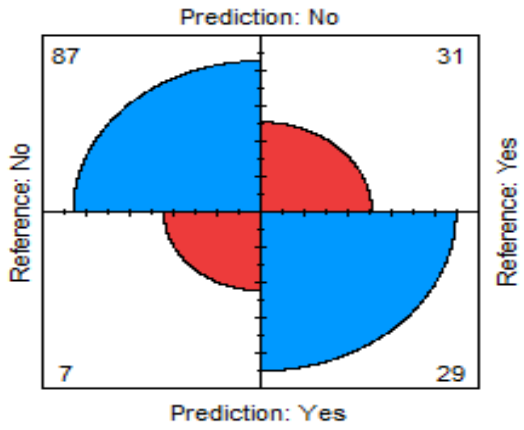
knn (73%)



knn_정규화 (75%)



knn_표준화 (75%)



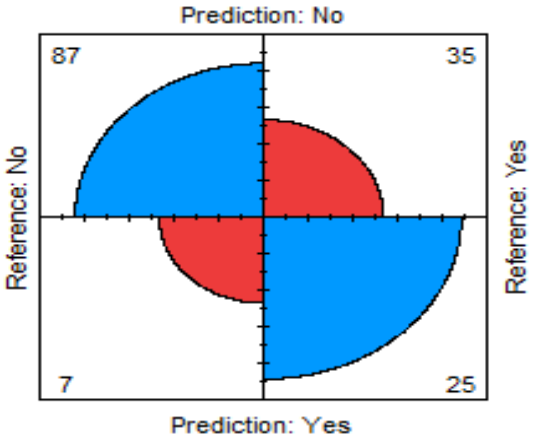
분석 모델

- KNN(K Neares Neighbors) - 범주를 알지 못하는 데이터가 있을 때, 근접한 k개의 데이터를 이용해 범주를 지정
  - 최근접 이웃의 거리를 계산하는 방식은 유클리드거리계산 방식을 이용
  - 훈련데이터로 모델을 훈련한 뒤, 테스트 데이터 범주분류
- SVM(Support Vector Machine) : - 지도학습의 분류모델 중 하나로, KNN과 마찬가지로 훈련데이터/테스트 데이터 필요
  - 분류 방식은 크게 선형 분류와 비선형 분류
  - 퍼셉트론의 개념을 토대로 분류하는 방식
- Naïve Bayes : - 분류를 쉽고 빠르게 하기 위해 분류기에 사용하는 특징들이 서로 확률적 독립이라는 가정
  - 확률적으로 독립이라는 가정에 위반되는 경우 에러가 발생할 수 있음
  - 특징들이 많을 경우, 특징들의 관계를 모두 고려하면 너무 복잡해지기 때문에 단순화 시켜 쉽고 빠르게 판단을 내릴 때 주로 사용

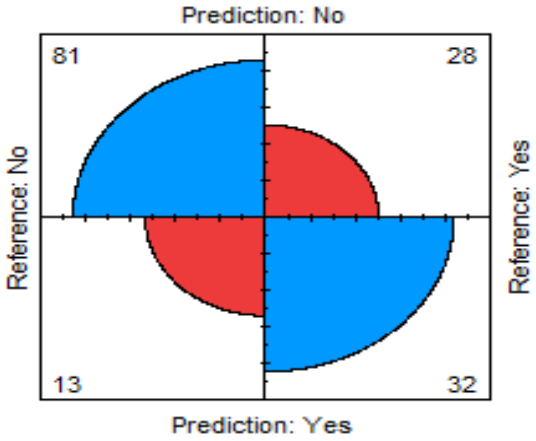
KNN, SVM, NAÏVE 모델 정확도 비교

- 트레인 데이터셋 (80%) 과 테스트 데이터 셋 (20%) 으로 분류한 뒤 각각의 모델을 적용
- Naïve 모델이 가장 정확도가 높은걸 확인 할 수 있음

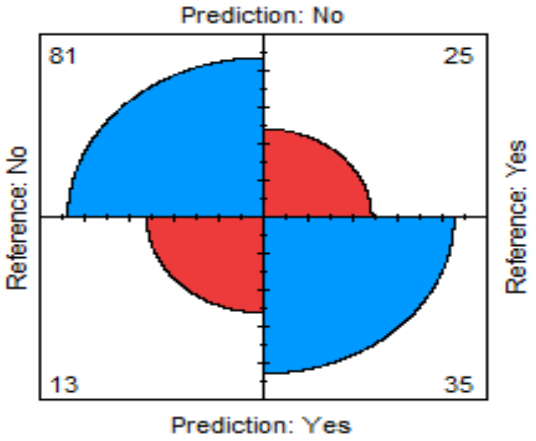
knn (73%)



svm (73%)



naive (75%)



결론

- 당뇨병 여부(Outcome)와 상관관계가 높은 속성들이 존재 (Glucose, BMI 등)
- Knn 모델 중에서 k= 28일 때, 정확도가 최고
- 데이터를 정규화, 표준화 하면 기존 knn 보다 정확도가 높아짐. 하지만 큰 차이를 보이진 못함
- Knn, SVM, Naïve bayes 모델 중에서 정확도가 가장 높은 모델은 Naïve bayes
- 하지만 각각의 속성들을 독립적으로 바라보는 Naïve bayes 모델 특성 상, 앞서 상관관계 분석을 살펴봤을 때 이 모델이 실효성이 있는지는 의문
- 또한 모델들의 정확도가 80%도 채 되지 않는 점을 미루어 보았을 때, 데이터의 갯수(768)가 많이 부족하다는 점을 알 수 있음
- 따라서 데이터를 좀 더 쌓아서 모델을 훈련시키면 최적의 모델이 무엇인지 좀 더 명확하게 나올 것이라고 생각함