**빅데이터 마이닝**

:머신러닝



|  |  |
| --- | --- |
| **과목** | 빅데이터 머신러닝 |
| **담당교수** | 손대순 교수님 |
| **학과,이름** | 데이터테크전공 20183226신희진 |
| **제출일** | 2020.10.18 |

**목 차**

**Ⅰ. 서론**

1. 분석의 목적
2. 분석의 의미

-무엇을 하는가

**Ⅱ. 본론**

1. 데이터 설명
2. 변수 설명
3. KNN 알고리즘

**Ⅲ. 결론**

1. 데이터분석 결과 및 해석

-KNN알고리즘 수행 결과 및 해석

1. 새로 알게 된 사실

**Ⅳ. 출처**

**I. 서 론**

**ⅰ. 분석의 목적**

K-nearset neighbor 알고리즘은 기존의 데이터를 활용하여 가장 비슷한 값을 예측하는 알고리즘이다. KNN은 사례기반, lazy learning 알고리즘이다. KNN알고리즘은 주변 값들을 이용하기 때문에 측정하고자 하는 위치와 얼마나 많은 주변 값(k)을 참고할지에 따라서 결과값이 달라진다. 따라서 데이터를 활용해 그 데이터들의 값을 예측하기 하고 주변 값을 측정을 하기 위해 분석을 한다.

**ⅱ. 분석의 의미**

-무엇을 하는가

KNN알고리즘은 최근접 이웃을 사용한 분류이다. 유사한 범주의 아이템들은 수치이고 동질데이터이다. 개념을 정의하기 힘들다면 최근접 이웃이 적합하다. 예를 들면 과일의 속성을 어떻게 정의해야 하는지? 를 분석하여 그 분석의 의미를 알아내는 것이다. 하지만 데이터들의 그룹 사이에 뚜렷한 구별이 없다면 대체로 경계를 식별하는데 적합하지는 않을 것이다.

**Ⅱ. 본 론**

**ⅰ. 데이터 설명**

사용한 데이터는 MASS패키지에 포함된 Pima.tr 학습 데이터이다. 본 데이터는 애리조나 피닉스 지역의 피마 인디언들의 당뇨병을 WHO기준으로 기록한 데이터이다. 학습 데이터와 테스트 데이터를 합쳐서 532명의 데이터가 존재하고, 총 여덟 개의 변수를 포함하고 있다.

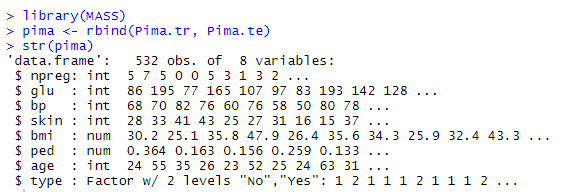
**ⅰⅰ. 변수 설명**

| 변수명 | 설명 | |-----------|---------------------------| | `npreg` | 임신 횟수 | | `glu` | 글루코스 부하 검사 결과[^4] | | `bp` | 확장기 혈압[^5] | | `skin` | 삼두근 피부 주름 두께[^6] | | `bmi` | BMI 수치 (몸무게 / 키^2) | | `ped` | 가족력을 기반으로 당뇨병이 생길 확률[^7] | | `age` | 나이 | | `type` | 당뇨병 여부 |

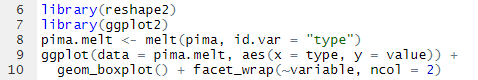
로 구성되어있고, 만약의 위해 기존 데이터들을 합쳐서 다시 샘플링하여 학습 데이터와 테스트 데이터를 새로 생성한다.

1. **KNN 알고리즘**

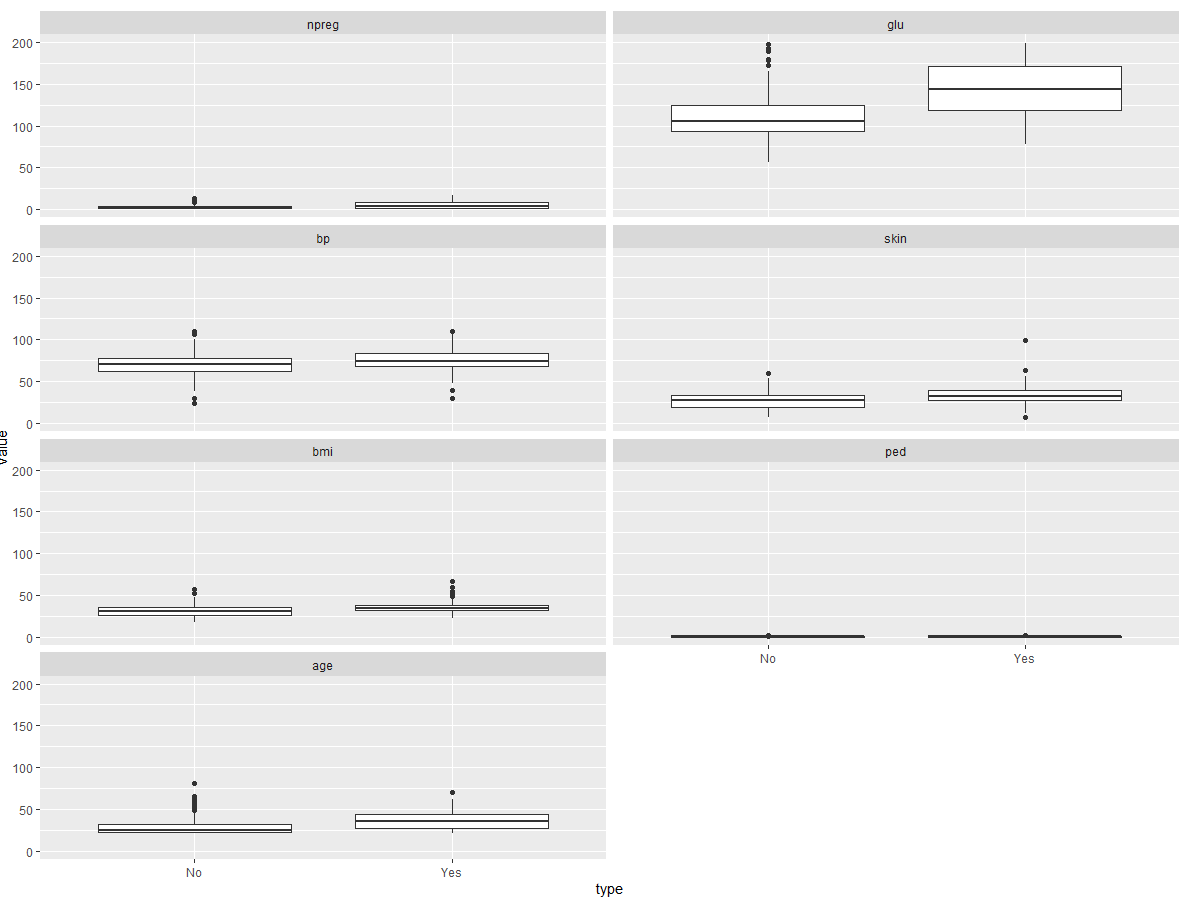
MASS패키지의 Pima.tr, Pima.te데이터를 가져와 데이터를 합치고, 데이터의 분포를 확인해 본 결과 532개의 데이터와 8개의 변수를 가지고 있는 것을 볼 수 있다.



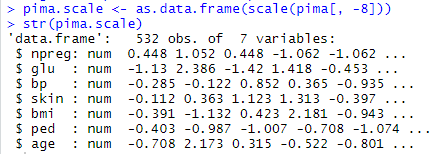
탐색적 데이터 분석을 위해 reshape2패키지의 melt()함수를 이용해 데이터의 형태를 바꿔준다. 바꾼 데이터를 각각의 변수를 기준으로 상자그림을 그려준다.



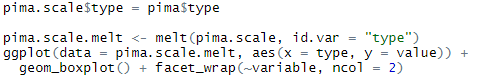
상자그림을 보면 모든 데이터의 측정 범위가 상이한 것을 볼 수 있다. 따라서 표준화 해주는데, z-표준화를 이용하여, 표준화 후, 위의 과정을 다시 진행한다.



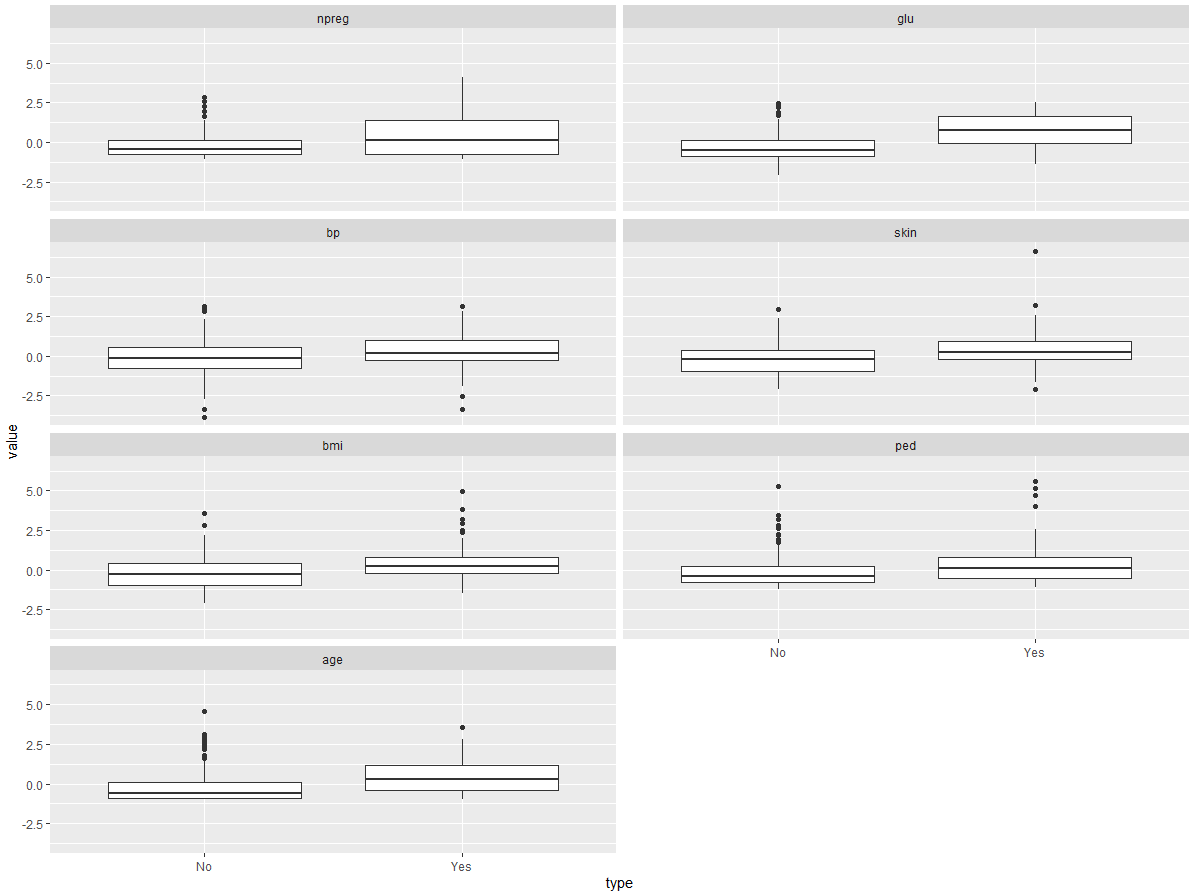
z-표준화를 하여 pima.scale로 넣어주고 다시 데이터의 구조를 확인한다.



표준화 해준 값으로 다시 위의 과정과 같이 상자그림을 그려준 것이다.

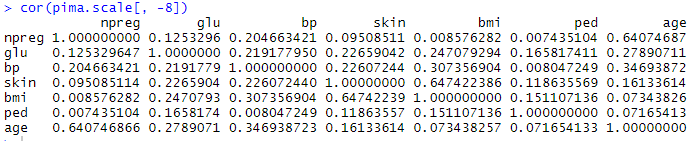


표준화 후 상자그림을 확인한 결과 달라진 것을 볼 수 있다.

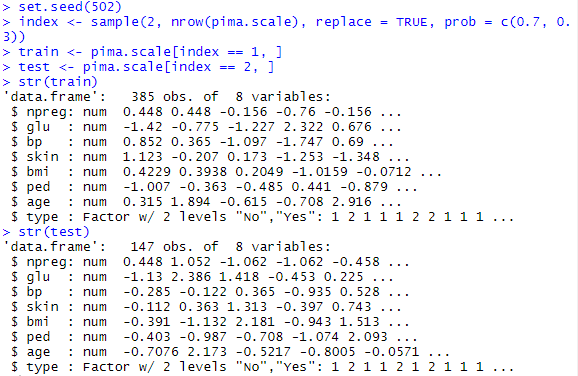


각 변수에 대해 KNN알고리즘의 경우 서로 간의 상관도가 모델링에 있어 문제가 없지만, 선형 회귀 법 등의 방법에서는 문제가 될 수 있으므로 확인을 한다.

상관계수를 확인하면 임신 횟수와 나이, 삼두근 피부 주름 두께와 bmi가 제법 강한 양의 상관관계를 보여주고 있다는 것을 알 수 있다.



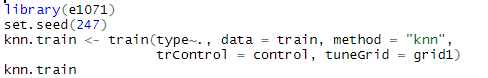
표준화 한 데이터를 학습 데이터와 테스트 데이터로 나눈다. 비율을 7:3으로 하고, 샘플링하여 분리를 한다.



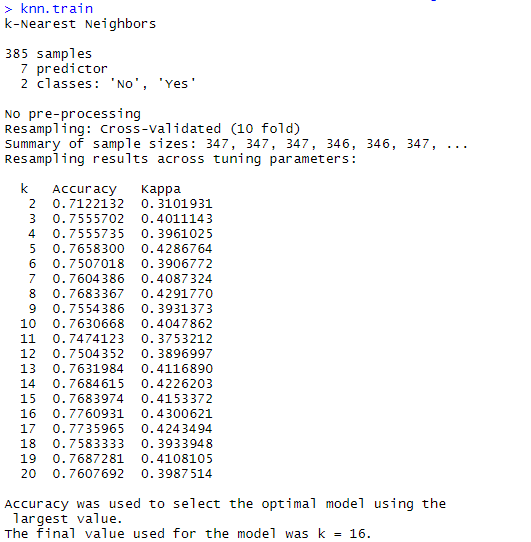
KNN 알고리즘의 적절한 k값을 찾기 위해 expand.grid()함수를 이요해 하나의 데이터 프레임을 만들어 준다. 그 다음에는 caret패키지를 불러와 trainControl()함수를 사용해 학습 데이터를 제어해준다.



설정을 해준 후 k값을 찾는다.



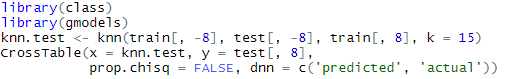
최적의 값은 16으로 나왔다. 최적의 k값은 정확도와 카파 계수를 통해서 얻어진다. 최적의 k값을 이용해 KNN 알고리즘을 실행한다.

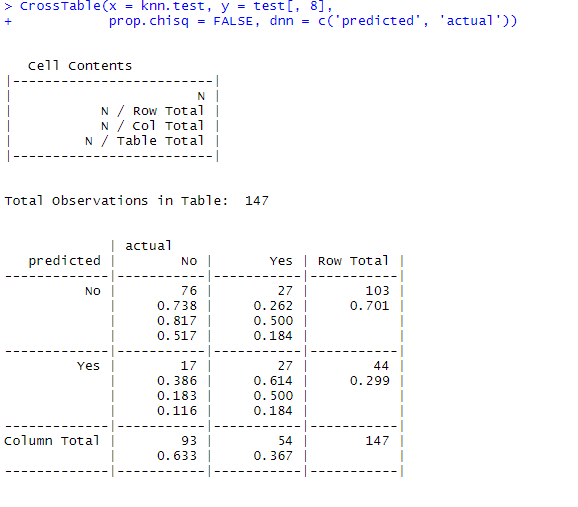


**Ⅲ. 결론**

**ⅰ. 데이터 분석 결과 및 해석**

확인 결과 KNN알고리즘 모델의 정밀도는 73.8%, 재현율은 81.7%, 정확도는 70.0% 인것을 확인 할 수 있다.





**ⅰⅰ. 새로 알게 된 사실**

KNN 알고리즘을 실습해보면서 단순하고 효율적이라는 것을 알 게 되었다. 하지만 사용할 데이터도 잘 정리를 해야 된다는 사실도 알게 되었다. 단순하지만 데이터 사이에 연관성이 없다면 알고리즘을 찾기 힘들 것 같다고 생각이 들었다. 그에 따라 많은 메모리들이 필요하여 매번 사용하기엔 좀 어렵지 않을까라는 생각도 들었다. 모델을 따로 생성하지 않아 분류단계가 느린 사실을 알게 되었다. KNN 알고리즘을 사용하는 것을 처음 배워보았는데, 막상 무슨 데이터를 사용하면 좋을까 고민이 많았다. 여러 데이터들을 사용해보기도 하였고, 중간중간 실패도 하였다. 하지만 어렵게 찾아서 그런지 배우는 하나하나가 나에게는 많은 도움을 주었고 유익한 수업이 되었던 것 같았다. 기회가 된다면 다른 데이터를 사용하여 또 다른 KNN 알고리즘을 찾아보고 싶었다.

**Ⅳ. 출처**

<https://blog.naver.com/kyoungin90/222090575303>

https://bandibell.tistory.com/348

https://otzslayer.github.io/machine-learning/kNN/