과 목 명: 데이터 마이닝

담당교수: 진창호 교수

학 과: 산업경영공학과

학 번: 2016100937 김성수

2016100946 김효준

2016101000 홍성희

Advanced DATA MINING

Intermediate Report

<청소년 비만 예측 및 패턴 도출>



발표 영상:

목 차

1. Introdduction 4
2. 과제 수행 배경 및 동기
3. 문제 정의 및 도출
4. 기존 연구 및 기술의 문제점
5. 제안하는 방법의 개요
6. 논문 주요 결과의 개요
7. methods 7
8. 연구 대상 및 데이터셋 생성
9. Data Understanding
10. 데이터 (남녀 통합, 남자/여자 데이터) 소개
11. 연구에 사용한 변수 및 Pre-processing
12. 분석 방법
13. **MODELING 15**
14. results 21
15. 모델링 결과
16. 변수 중요도 확인
17. discussion 74
18. limitation 76

그림 목차

그림 1 비만의 심각성 - 5

그림 2 비만으로 인한 영향 - 5

그림 3 PaientsLikeme - 5

그림 4 PatientsLikeMe 활용 사례 - 6

그림 5 데이터 표준화 및 정규화 - 14

그림 6 Cross Validation 설명 - 15

그림 7 Naïve Bayes 설명 - 16

그림 8 Random Forest - 17

그림 9 그림 9 로지스틱 회귀분석 - 18

그림 10 Ridge Cost Function - 18

그림 11 Lasso Cost Function - 18

그림 12 Elastic Net Cost Function - 18

그림 13 KNN - 19

그림 14 Gradient Descent - 20

그림 15 Gradient Decent 비교 - 20

그림 16 SVM - 21

Summary  
■ **Introduction** ( 대상 기업 : PatientsLikeMe – “환자들의 페이스북” )  
- 세계 비만 인구는 2008년 기준 총 14억 명, 매년 약 280만 명이 비만 관련 질병으로 사망   
- 청소년기 비만은 성인으로 이어짐 🡺 청소년 비만 사전 예측 및 비만 유발 요인 파악이 중요  
■ **Methods**  
- 국민건강영양조사 및 청소년 건강행태 조사에 참여한 청소년 대상으로 분석 수행  
- 기존에 발표된 청소년 비만 관련 논문을 바탕으로 Feature Selection  
- EDA (탐색적 분석) : 다중 공선성, 데이터 시각화를 통한 Feature의 Value 분포, 통계량 등 확인  
- 3개 데이터셋, 4가지 결측치 처리 방법, 2가지 데이터 스케일링, K=5/10인 K-fold 교차 검증 수행  
- 6가지 분류기법(NB Classifier, RF, LR, K-NN, SGD, SVM)사용  
🡺모델 별 그리드 서치 수행하며 청소년 통합/남자/여자 데이터 각각에 대해 144개의 모델 수립.   
■ **Results 및 Discussion**

|  |  |  |  |
| --- | --- | --- | --- |
| **모델** | **결측치 처리 방법**  **(통합/남자/여자)** | **스케일링 효과 ( O / X )** | **K-fold 교차 검증 K값 변화 효과 ( O / △ /X )** |
| **Naïve Bayes Classifier** | 중앙,최빈/완전/완전 | X | X |
| **Random Forest** | 완전/완전/완전 | X | △ |
| **Logistic Regression** | 완전/완전/완전 | X | O |
| **K-NN** | 완전/완전/완전 | O | △ |
| **SVM** | 완전/완전/완전 | O | △ |

|  |  |
| --- | --- |
| **데이터** | **유의미한 인자 ( 상위 5개 요인 중 2번 이상의 빈도로 등장 )** |
| **통합** | 주관적 체형인지, 만 나이, 하루 평균 앉아있는 시간, 주관적 건강상태, 성별,  평소 스트레스 인지 정도 |
| **남자** | 주관적 체형인지, 만 나이, 하루 평균 수면시간, 2주 이상 연속 우울감 여부,  평소 스트레스 인지 정도 |
| **여자** | 주관적 체형인지, 만 나이, 하루 평균 앉아있는 시간 |

- “결측치 처리 방법” 에는 각 데이터별로 가장 높은 성능을 보인 결측치 처리 방법을 기재함  
 ( 평균=평균값 대체, 중앙=중앙값 대체, 최빈=최빈값 대체, 완전=완전 제거법을 의미함)  
- △는 세 데이터 중 하나라도 K값을 증가시켰을 때 성능이 하락한 경우를 의미  
참고 : SGD는 파라미터 변경 시 사용하는 모델 기법이 달라져서 위 표에서 제외함- SGD가 남녀 통합 데이터/남자 데이터에서, NB Classifier가 여자 데이터에서 Recall이 우수  
🡺 청소년 남자 비만과 청소년 여자 비만에 유의미한 인자에 차이가 있음을 확인  
■ **Limitation** - 컴퓨팅 리소스의 한계, 가중치 적용의 한계, 결측치 처리 방법의 한계

**[ Introduction ]**

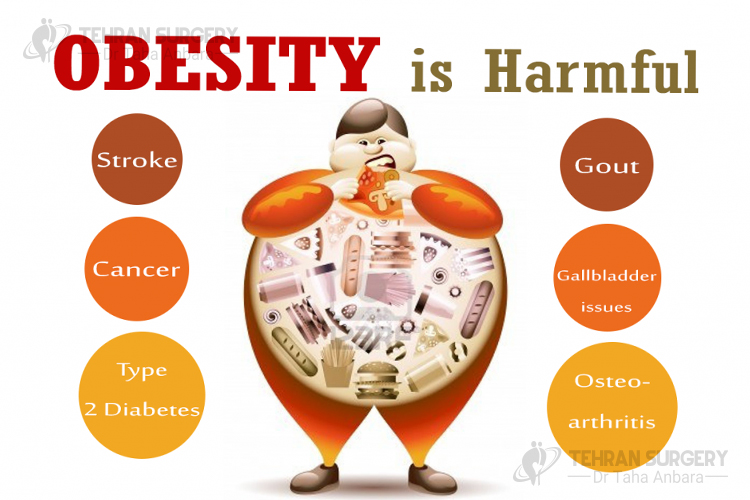
****■ **과제 수행 배경 및 동기**

그림 2 비만으로 인한 영향

출처:

그림 1 비만의 심각성

1. 세계보건기구 (WHO)에서는 비만을 21세기 신종 전염병으로 선포 (윤난희, 2016)
2. 2008년 기준, 세계 비만 인구는 총 14억 명에 달하고 매년 약 280만 명이 비만과 관련된 질병으로 사망 (윤난희, 2016)
3. 청소년기 비만 : 외모로 사회적 낙인(stigma)을 경험할 위험 존재 (Sunhee Park**,** 2011)
4. 사회적 고립, 자살, 왜곡된 신체 이미지 등 정신적 질환에 노출 가능성 (Sunhee Park**,** 2011)
5. 청소년기에 비만이면 성인까지 이어질 확률이 높음 (Sunhee Park**,** 2011)
6. 심장병, 호흡기 질환 등의 만성질환에 더 큰 노출 가능성 (Sunhee Park**,** 2011)
7. 고혈압, 제2형 당뇨병 등의 다른 심혈관 위험요인의 위험 증가 (Sunhee Park**,** 2011)

🡺 청소년 비만을 사전에 예측하는 모델을 수립하고 청소년 비만을 유발하는 요인들을 파악함으로써 추후 본 연구를 공공 기관, 민간 기업 등 다양한 분야에서 활용될 것으로 예상

■ **문제 정의 및 도출**

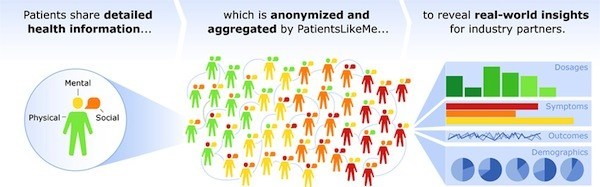
□ 대상 기업 : PatientsLikeMe - “환자들의 페이스북”

그림 3 PatientsLikeMe

: 2,800개 이상의 질병에 대한 전세계의 83만명 이상의 환자들이 모인 거대한 SNS 서비스

- 환자들이 직접 자신의 상태가 어떤지, 증세가 어떻게 진행되고 있는지, 어떤 약을 언제부터   
 먹었고, 효능/부작용은 어땠는지 등을 작성

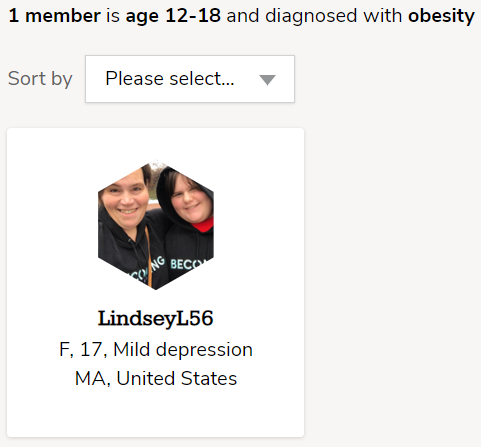
 Ex) PatientsLikeMe에 가입한 환자는 본인의 관심을 끄는 가입자(가령 특정 약을 2년 이상 복용한 환자 혹은 특정 약을 매일 5㎎씩 복용하는 환자)를 찾아 증세의 진행 경과, 효능/부작용, 개인 노하우 등의 정보를 상호 공유

그림 4 : PatientsLikeMe 활용 사례 (출처 : https://www.patientslikeme.com/)

- 83만 명이 넘는 PatientsLikeMe 가입자들 중 비만이 있는 가입자의 수는 3231명에 불과하고 이 중 청소년의 수는 단 한 명에 불과 ( 2020.10.17 기준 )  
- PatientsLikeMe에는 환자들이 자신의 상태, 증세 등을 직접 쓰기 때문에 사실적이라는 장점은 있으나 이는 주관적이기 때문에 글쓴이 스스로가 인지하지 못하는 다른 요인에 의해 비만이 유발되거나 비만 관련 질환들로 이어질 가능성 존재

🡺 PatientsLikeMe는 데이터를 기업에 판매함으로써 수익을 창출한다. 그런데 비만인 청소년의 데이터는 단 한 개에 불과하다. 만약 PatientsLikeMe 사이트에서 청소년 비만 예측 서비스와 청소년 비만 유발 요인들을 제공한다면 자연스레 청소년 사용자뿐만 아니라 청소년을 가진 부모의 PatientsLikeMe 가입이 증가할 것이다. 사용자가 늘어나면 데이터는 늘어나므로 PatientsLikeMe는 이로 인해 수익이 증가할 것으로 기대된다.

■BP : 청소년 비만에 대한 예측 및 패턴 도출

■DSP : SGD, SVM, Random Forest, Logistic Regression, Naïve Bayes Classifier, KNN Classifier등   
 분류 기법을 이용한 청소년 비만 Mechanism 파악

■ 기존 연구 및 기술의 문제점

- 기존 연구들은 단순 상관관계나 평균 차 검정, 회귀분석 등을 바탕으로 한 체중에 영향을 주는 요인연구가 주를 이룸. (정보람, 2020; 유지호, 2014; 박소연, 2019 등)

- 성인 및 청소년 비만 영향 요인을 연구한 논문은 다수 발표된 바 있으나, 청소년 성별에 따른 각 그룹의 비만 영향 요인에 대한 연구는 부족한 실정. (2020.10.17 기준)

- 최근 비만과 관련된 분야에서도 머신 러닝을 적용한 시도들이 증가하고 있으나, 해외 연구들이 주를 이룸.

■ 제안하는 방법의 개요

- 국민건강영양조사 자료와 청소년건강행태조사 자료를 이용하여 국내 청소년을 대상으로 함

- 기 등재된 논문을 이용해 Domain Knowledge를 보완해 청소년 비만 영향 요인들을 추출함.

- Logistic Regression, SVM, K-NN, Random Forest, SGD, Naive Bayes Classifier 즉, 분류 분석에 사용되는 머신 러닝 기법을 이용해 청소년 비만 예측 모델 수립 및 청소년 비만 영향 요인별 중요도 파악

■ 논문 주요 결과의 개요

- 남/녀 통합 청소년, 남자 청소년, 여자 청소년 데이터 모두에서 ‘주관적 체형인지’가 가장 영향이 큰 Feature로 도출되었다. 이는 사회적 낙인을 성인에 비해 쉽게 경험하는 청소년의 특성이 반영되어 있다고 해석할 수 있으며 비만 예방 및 조기교육에 대한 시사점을 제공한다.

- 본 연구에서 추출한 인자들에 대하여 가장 유의미한 척도라고 판단한 Recall을 기준으로 각 데이터 셋 별로 남/녀 통합 데이터셋과 남자 데이터는 SGD Classifier가, 여자 데이터 셋에 대해서는 Naïve Bayes Classifier가 가장 좋은 성능을 드러낸 것을 확인하였다. 또한 데이터별로 비만 영향 요인에 차이가 있음을 확인하였다.

- 본 연구 ‘청소년 비만 예측 및 매커니즘 도출 연구’는 PatientsLikeMe의 부족한 청소년 표본에 대한 객관적 지표의 필요성을 시사한다. 이에 ‘국민건강영양조사’ 및 ‘청소년 건강행태 조사’ 데이터를 활용하여 객관화된 실질적 지표를 제공하고 가입자수 증대에 이바지할 수 있는 예측 모델의 제공으로 수익률 창출에 기여하는데 의의를 지닌다.

**[ Method ]**

■ 연구 대상 및 데이터셋 생성

* 본 연구는 2018년도에 시행된 제7기 1~3차년도(2016~2018년) 국민건강영양조사에 참여한 만 12세~18세의 청소년 및 제14차년도(2018년) 청소년 건강행태 조사에 참여한 청소년을 대상으로 진행함
* 만 12~18세의 청소년에 대하여 전체 청소년, 남자 청소년, 여자 청소년으로 분류하여 각각 분석을 진행함
* 국민건강영양조사는 보건복지부와 질병관리본부가 전국 규모의 건강 및 영양을 조사한 통계자료로 해당 조사는 전 국민을 대상으로 하여 건강설문조사, 검진조사, 영양조사를 1년 단위로 집계됨
* 청소년 건강 행태조사는 교육부, 보건복지부, 질병관리본부가 중1 ~ 고3 학생을 대상으로 실시하는 익명성 자기 기입식 온라인 조사로 흡연, 음주 등 약 15개의 영역에 대해 1년 단위로 집계됨
* 초기 연구 진행 과정에서는 국민건강영양조사 제 7기 전체 자료와 청소년 건강행태 조사 자료 2016년, 2017년, 2018년도 데이터를 다 통합하여 분석을 수행하려 했으나, 본 연구에서는 Grid Search와 다양한 스케일링 기법 등을 활용하여432개의 모델을 수립하여 컴퓨팅 리소스의 한계 및 과다한 시간적 비용이 생겨 분석을 수행하지 못했음.

■ Data Understanding

① Strength  
1) 원시자료 구성이 건강설문조사, 검진조사, 영양조사 별로 분류된 데이터  
: 필요한 데이터에 대한 접근성과 활용성 용이  
2) 연령층(성인, 청소년, 소아)별로 분류된 데이터  
: 청소년에 대한 데이터를 추출해내기 용이  
3) 질병관리본부 건강영양조사과에서 공개한 공공 데이터  
: 해당 데이터 전반에 대한 높은 신뢰도  
4) 대대적으로 공개된 공공 데이터  
: 사용시 별도의 비용이 발생하지 않음, 데이터 수집에 소요되는 시간이 적음.

② Limitation  
1) 국민건강영양조사 응답 시 청소년의 경우 외부적 요인의 작용(부모를 동반해 진행)  
: 설문조사의 특정 변수에 대한 신뢰도가 떨어질 가능성이 존재한다.  
2) 제7기로 지정된 16년, 17년, 18년도의 설문 항목들 중 다른 항목이 존재  
: 일부 데이터 사용 불가  
3) 국민건강영양조사 자료는 단면조사의 특성 : 변수들 간의 선후 관계 확인에 어려움  
4) 자가 보고된 신장과 체중 수치  
: 산출된 체질량지수를 기준으로 비만 여부를 판단 시, 비만율이 과소 추정될 가능성

③ Benefit  
1) 청소년 비만 사전 예측 및 청소년 비만 영향 요인 파악  
: 공공 기관에서는 이를 통해 초중고 교육에 활용하여 공익을 증대시킬 수 있으며, 민간 기업에서는 이를 통해 건강관리 프로그램 등에 활용하여 수익을 창출할 수 있음  
2) Patientslikeme의 부족한 청소년 비만과 관련된 정보 제공  
: ‘Obesity’ 커뮤니티 내 유의미한 결과를 제공하여 고객 유치 및 기업 활성화

④ Cost  
1) 700개가 넘는 attribute(2018년 기준)와 다수의 결측치  
: 유의미한 변수선정 및 결측치 대체에 있어 많은 시간과 비용 소요  
2) 다수의 범주형 데이터와 수치형 데이터 혼재  
: 785개의 변수들 중 범주형 변수 파악 및 One-hot-encoding에 시간과 노력이 소요

■ 데이터 (남녀 통합, 남자/여자 데이터) 소개

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **데이터 명** | **Instance 개수** | **비만 아닌 인원 수 (명)** | **비만 인원 수 (명)** | **비만율 (%)** |
| **df\_all** | 48706 | 41642 | 7064 | 14.503 |
| **df\_all\_b** | 23871 | 19103 | 4768 | 19.974 |
| **df\_all\_g** | 24835 | 22539 | 2296 | 9.245 |
| **df\_mean** | 59654 | 51137 | 8517 | 14.277 |
| **df\_mean\_b** | 29707 | 23922 | 5785 | 19.474 |
| **df\_mean\_g** | 29947 | 27215 | 2732 | 9.123 |
| **df\_median** | 59654 | 51137 | 8517 | 14.277 |
| **df\_median\_b** | 29707 | 23922 | 5785 | 19.474 |
| **df\_median\_g** | 29947 | 27215 | 2732 | 9.123 |
| **df\_mode** | 59654 | 51137 | 8517 | 14.277 |
| **df\_mode\_b** | 29707 | 23922 | 5785 | 19.474 |
| **df\_mode\_g** | 29947 | 27215 | 2732 | 9.123 |

표 1 데이터 소개

비만율 (%) = (비만 인원 수 / Instance 개수 ) \* 100

* 남녀 통합/남자/여자 데이터 비만율을 확인한 결과 결측치 있는 행 제거, 결측치 대체 방법을 했을 경우 둘 다 청소년 남자의 비만율이 여자 청소년의 비만율보다 2배 이상 큼

■ 연구에 사용한 변수 및 Pre-processing

□ Feature selection   
Data의 Dimensionality를 줄여GIGO(Garbage In, Garbage Out) 현상 및 Curse of Dimensionality 방지, Domain Knowledge 보완을 위해 논문을 통하여 ‘청소년 비만’과 관련 있는 8개의 Feature를 선별함.

- 선별한 Feature의 목록

1. 만나이 [age] (윤영숙 외 1명, 2009)  
   ② 주중 하루 평균 수면시간[Total\_slp\_wk],   
   주중 하루 평균 수면시간[Total\_slp\_wd] (Sunhee Park, 2011)  
   ③ 주관적 건강 상태[D\_1\_1] (윤영숙 외 1명, 2009)  
   ④ 평소 스트레스 인지 정도[BP1] (Sunhee Park, 2011)  
   ⑤ 주관적 체형 인지 [BO1] (NA YOUNG SHIN 외 1명, 2008)  
   ⑥ 1주일간 근력운동 일수 [BE5\_1] (Sunhee Park, 2011)  
   ⑦ 평소 하루 앉아서 보내는 시간(시간) [BE8\_1],   
   평소 하루 앉아서 보내는 시간(분)[BE8\_2] (Sunhee Park, 2011)  
   ⑧ 2주이상 연속 우울감 여부 [BP5] (지영주¸ 2011)

① 만나이[age]: 윤영숙 외 1명의 연구(2009)에 따르면 나이가 심혈관질환 위험 요인을 매개로 하여 비만율이 증가한다고 제시

② 주중 하루 평균 수면시간[Total\_slp\_wk]: Sunhee Park의 연구(2011)에 따르면 감소된 수면시간 과체중과 비만이 유의하다고 제시하며 늦게 수면에 취하기에 밤에 음식을 섭취하게 되는 등의 과학적 근거를 제시

③ 주관적 건강 상태[D\_1\_1]: 윤영숙 외 1명의 연구(2009)에 따르면 주관적 건강 상태를 안 좋다고 생각할수록 비만율이 증가한다고 제시

④ 평소 스트레스 인지 정도[BP1]: Sunhee Park의 연구(2011)에 따르면 정신적 스트레스가 글루코코티코이드의 증가로 비만 및 과체중에 영향을 끼친다고 제시

⑤ 주관적 체형 인지 [BO1]: NA YOUNG SHIN 외 1명의 연구(2008)에 따르면 청소년기 비만과 가장 유의한 인자는 주관적 체형 인지이며 주관적 체형인지가 안 좋을수록 유의하다고 제시

⑥ 1주일간 근력운동 일수 [BE5\_1]: 근력운동을 많이 할수록 비만과 과체중의 확률을 낮추는데 도움을 주는데 유의하고 에너지를 사용하기에 비만과 과체중과 유의미한 관계가 있다고 제시

⑦ 평소 하루 앉아서 보내는 시간(시간) [BE8\_1], 평소 하루 앉아서 보내는 시간(분) [BE8\_2]: 학업에 많은 시간을 투자하는 청소년 상 앉아있는 시간이 많을수록 비만과 과체중에 유의하다고 제시

⑧ 2주 이상 연속 우울감 여부 [BP5]: 지영주 외 1명의 연구(2011)에 따르면 우울의 경험이 있는 비만군이 비 비만군보다 높게 나타났고 다변량 접근법에서도 유의한 예측인자로 파악됨

□ Feature 세부 전처리 내용

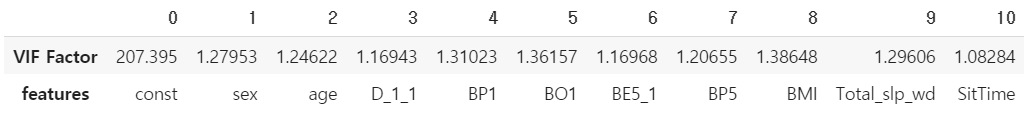
① 만나이[age]: 청소년만을 추리기 위하여 만 12세에서 18세의 청소년만을 필터링  
② 주중 하루 평균 수면시간[Total\_slp\_wd]: 연속형 변수의 특성상 이상치를 제거  
③ 주관적 건강 상태[D\_1\_1]: 9(모름, 무응답)을 NA로 처리하고 순서형 변수이므로 Category화를 시켜주고 순서를 지정   
④ 평소 스트레스 인지 정도[BP1]: 8(비해당), 9(모름, 무응답)을 NA로 처리하고 순서형 변수이므로 Category화를 시켜주고 순서를 지정  
⑤ 주관적 체형 인지 [BO1]: 8(비해당), 9(모름, 무응답)을 NA로 처리하고 순서형 변수이므로 Category화를 시켜주고 순서를 지정  
⑥ 1주일간 근력운동 일수 [BE5\_1]: 8(비해당), 9(모름, 무응답)을 NA로 처리   
⑦ 하루 평균 앉아서 보내는 시간(분) [SitTime]: 88(비해당), 99(모름, 무응답)에 대해서는 NA로 처리하고, [BE8\_1]의 (시) 특성과 [BE8\_2]의 (분) 특성을 활용하여 하루에 앉아서 보내는 시간을 (분)으로 통일하고 이상치를 제거하여 파생변수 생성  
⑧ 2주이상 연속 우울감 여부 [BP5]: 8(비해당), 9(모름, 무응답)을 NA로 처리하고 명목형 변수로써 Category화 실시  
⑨ 비만 여부 [BMI]: BMI를 구하는 식인 (체중/신장(m))^2을 python에 그대로 적용하여 산출하였다. 그 후 적절한 Target Attribute로 mapping시켜주기 위하여 비만이 아닌 사람을 0으로(BMI < 25), 비만인 사람(BMI >= 25)을 1로 바꿔준 후 Category화를 하여 파생변수 생성

□ EDA (탐색적 데이터 분석)

|  |  |
| --- | --- |
| [성별] | [만 나이] |
| : 남성 혹은 여성의 값만 가짐 | : 20대~30대 중반 Instance가 비교적 적음 |
| [주관적 건강 상태] | [평소 스트레스 인지 정도] |
| : 9. “모름”에 대해 NA로 처리 필요 | : 8. 비해당(소아) , 9. 모름, 무응답에 대해 NA로 처리 필요 |
| [주중 하루 평균 수면시간] | [주중 하루 평균 수면시간] |
| : 다수의 이상치 처리 필요 | 🡪 이상치 제거 후, 연속형 데이터의 균일한 분포 확인 가능 |
| [1주일간 근력운동 일수] | [평소 하루 앉아서 보내는 시간(시간)] |
| : 8. 비해당(소아) , 9. 모름, 무응답에 대해 NA로 처리 필요 | : 88. 비해당(소아), 99. 모름, 무응답에 대해 NA로 처리 필요 |
| [평소 하루 앉아서 보내는 시간(분)] | [2주이상 연속 우울감 여부] |
| : 88. 비해당(소아), 99. 모름, 무응답에 대해 NA로 처리 필요 | : 8. 비해당(소아), 9. 모름, 무응답에 대해 NA로 처리 필요 |
| [신장] | [체중] |
| 파생변수 ‘BMI’를 위한 신장의 분포 확인 | 파생변수 ‘BMI’를 위한 체중의 분포 확인 |
| [주관적 체형 인지] | [앉아있는 시간] |
| : 8.비해당, 9.모름/무응답에 대해 NA로 처리 필요 | : [BE8\_1], [BE8\_2] 의 결측치 처리 후 파생변수 [SitTime] 의 분포 확인 |

□ 다중공선성 파악

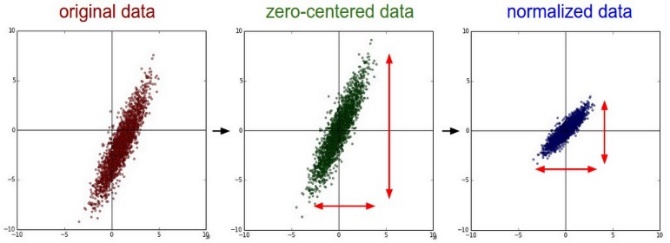
각 변수 간에 상관관계가 높으면 부정적인 영향을 미치는 것을 ‘다중공선성’이라 함  
🡪 다중공선성 문제를 확인하기 위해, 분산팽창요인(VIF, Variance Inflation Factor)를 파악  
🡪 그 결과, 추출된 변수 간의 다중공선성 문제는 없음을 확인



□ 데이터 전처리

1) 결측치 처리

① 완전제거법 : 결측치 포함 Data Object 삭제  
② 결측치 특정 퍼센트(Ex. 10%, 20%) 이상인 Attribute 삭제  
🡪 본 연구의 추출된 Attribute에는 결측치가 10% 이하이므로, attribute 자체 삭제 불가  
③ 단일대체법 : 중앙값 대체, 평균값 대체, 최빈값 대체

2) 데이터 스케일링

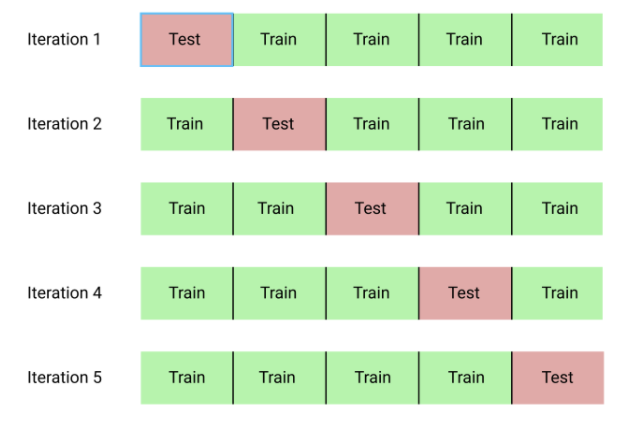
① 표준화 : 모든 데이터 포인트가 동일한 정도의 스케일로 조정 (이상치에 민감하다는 단점 존재)  
🡪 원래 데이터의 분포 손실되지 않음  
② 정규화 : 데이터의 범위를 0과 1로 변환하여 데이터 분포를 조정  
**□ K-fold 교차검증(K-fold Cross Validation)**: 전체 데이터의 일부를 Validation Set으로 사용해 모델 성능을 평가하는 것  
 🡪 데이터 셋의 크기가 작은 경우 테스트셋에 대한 성능 평가 개선에 효과적

그림 5 데이터 표준화 및 정규화  
 출처: cs231n, Stanford University (데이터 스케일링)

그림 6 Cross Validation 설명

* Modeling

1. **Naive Bayes Classifier**

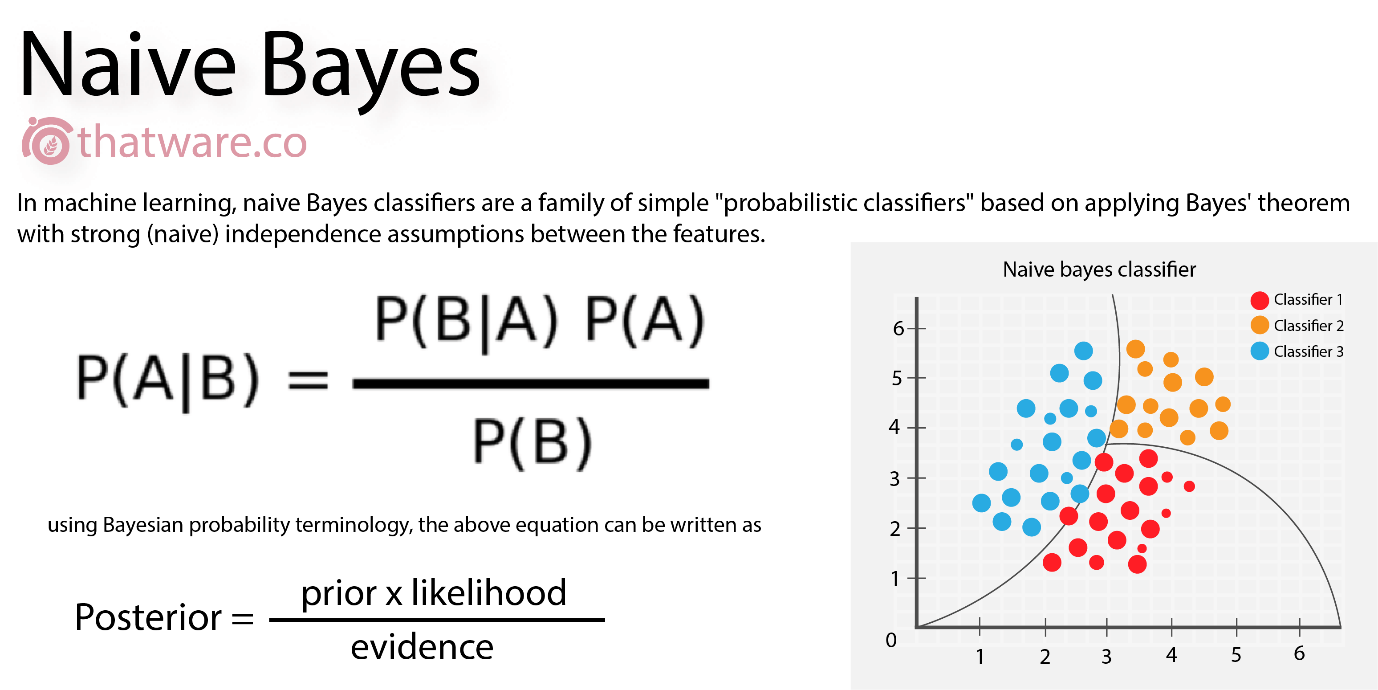


그림 7 Naïve Bayes 설명

□ 정의  
- 베이즈 정리를 적용한 변수들 간 독립을 가정하는 확률 분류기의 일종  
Ex) 특정 과일을 사과로 분류 가능하게 하는 변수들 (둥글다, 빨갛다, 지름 10cm)은 Naïve Bayes Classifier는 특성들 사이에 발생할 수 있는 연관성이 없음을 가정하고 각각의 변수들이 특정 과일이 사과일 확률에 독립적으로 기여하는 것으로 간주  
- 훈련 데이터를 활용해 특징 값이 제공하는 증거를 기반으로 결과가 관측될 확률을 계산   
- Label이 없는 새 데이터에 적용할 때 결과가 관측될 확률을 이용해 가장 유력한 클래스를 예측

□ 파라미터 : 모델의 복잡도를 조절하는 α (alpha)

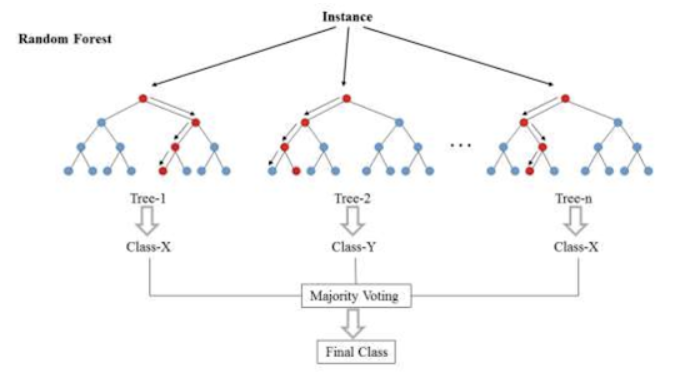
(α가 주어지면 모든 변수에 양의 값을 가진 가상의 Data Point를 alpha 개수만큼 추가)

**□** 장단점

- 장점  
1) Computation Cost가 작음.  
2) Multiple Class 예측을 위해서도 사용 가능  
3) 큰 데이터셋에도 빠른 수행 가능

- 단점  
1) Feature간의 독립성이 있어야 함  
2) 수치 속성으로 구성된 많은 데이터셋에 적합하지 않음

1. **Random Forest**



* 정의  
  : 분류 및 회귀분석 등에 사용되는 앙상블 기법의 일종으로 훈련과정에서 구성한 다수의 결정 트리로부터 부류나 평균 예측치를 통해 예측하는 모형

□ 원리  
1) 부트스트랩(bootstrap)방법으로 훈련 데이터셋을 생성  
2) 부트스트랩한 데이터를 활용하여 기초 분류기(트리)를 훈련  
3) 기초 분류기(트리)를 하나의 분류기로 결합 (이때 평균이나 과반수투표 방식을 활용)

그림 8 Random Forest

□ 중요 파라미터  
1) 트리의 개수: 총 몇 개의 트리로 구성할 것인지 정하는 것으로 트리수가 적으면 시간이 적게 걸리는 대신 일반화 능력이 떨어지고, 트리가 많다면 시간은 오래 걸리지만 일반화 능력이 우수  
2) 최대 허용 깊이: 하나의 트리에서 Root 노드에서 Terminal 노드까지 최대 몇 개의 노드를 거칠 것인지 결정하는 것으로 깊이가 너무 작으면 과소적합이 일어나고 너무 깊으면 과대적합이 일어날 수 있기에 적절한 값의 설정이 중요

□ 장단점  
장점  
① 결정 트리가 하나면 훈련 데이터에 과대적합이 되는 경향이 있으나 여러 결정 트리를  
 활용 시 과대적합을 방지할 수 있음  
② 분류모델에서 상대적으로 중요한 변수를 선정가능  
③ 이상치에 덜 민감

단점  
① 모델의 설명적인 측면이 부족함(Non-descriptive)  
② 텍스트 데이터와 같은 매우 차원이 높고 희소한 데이터에는 잘 작동하지 않음  
③ 많은 트리를 만들수록 많은 메모리와 긴 훈련시간이 필요

**3. Logistic Regression**

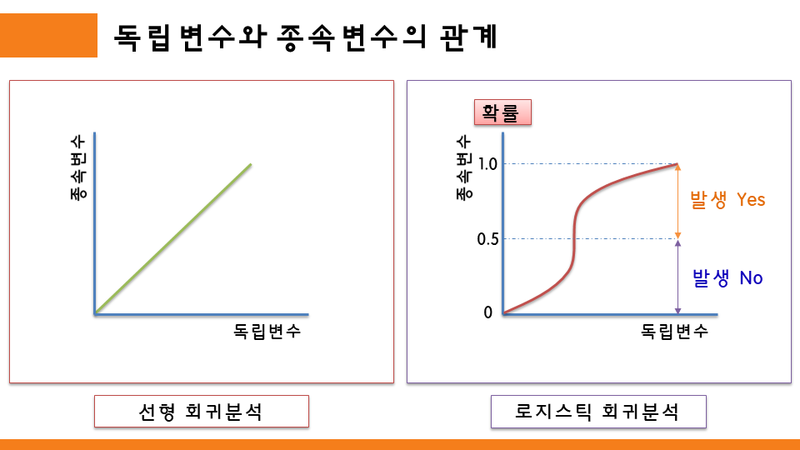
**-** 종속변수의 계급이 0과 1 두 가지 값을 가지고 관심의 대상이 되는 계급이 1인 경우에 종속변수가 1이 될 확률을 예측하는 모형  
Ex) 종속변수 값, 즉 확률이 0.5보다 크면 사건이 일어나며, 0.5보다 작으면 그 사건이 일어나지 않는 것으로 예측  
**-** 독립 변수는 연속형, 범주형 모두 가능  
**-** 종속 변수는 범주형

그림 9 로지스틱 회귀분석

□ 정규화 회귀분석  
- 선형 회귀모델의 Cost Function (비용함수)에 페널티를 적용  
- 정규화 방법을 사용하여 회귀계수 추정값의 변동을 줄임으로써 모형의 예측 오류를 개선 가능  
 (다중공선성 문제 및 Over Fitting 문제 방지)

1. Ridge : L2 정규화

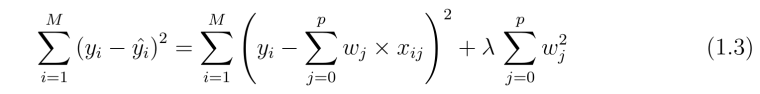


그림 10 Ridge Cost Function

1. Lasso : L1 정규화

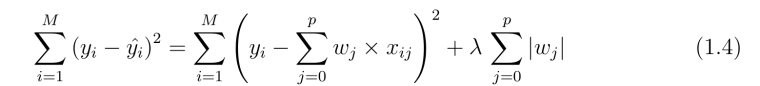


그림 11 Lasso Cost Function

1. Elastic Net : L1 정규화와 L2 정규화 혼합

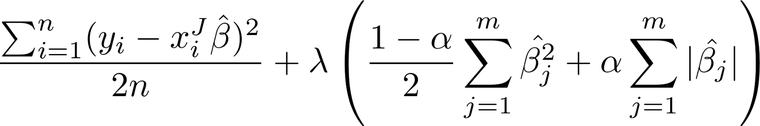
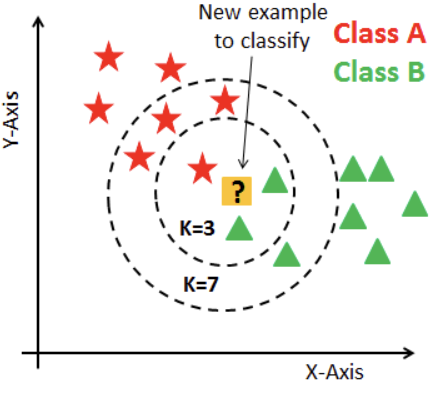


그림 12 Elastic Net Cost Function

**4. KNN(K-Nearest Neighbor)**

□ 정의  
- 새로운 데이터를 입력받았을 때 유클리디안 거리 등을 활용하여 가장 가까이 있는 K개의 데이터를 통해 새로운 데이터를 예측하는 알고리즘  
- K값은 사람이 정하는 하이퍼 파라미터이며 분류 모델의 경우 K개의 최근접 이웃 데이터에 대해 과반수 의결에 의해 분류되며 회귀 모델의 경우 K개의 최근접 이웃이 가진 값의 평균을 출력

그림 13 KNN

□ 중요 매개변수  
1) K: 일반적으로 K값이 커질수록 분류에서 잡음의 영향이 줄어드나 항목간 경계가 불분명해짐. K가 너무 작으면 과소적합이 될 우려가 있으며 너무 크다면 과대적합의 위험성이 존재함. 특히 분류 문제의 경우 동률의 투표를 막기위해 홀수의 K를 선택하는 것이 바람직하며 이러한 최적의 K는 교차검증이나 부트스트랩 방식 등을 활용하여 유추가능

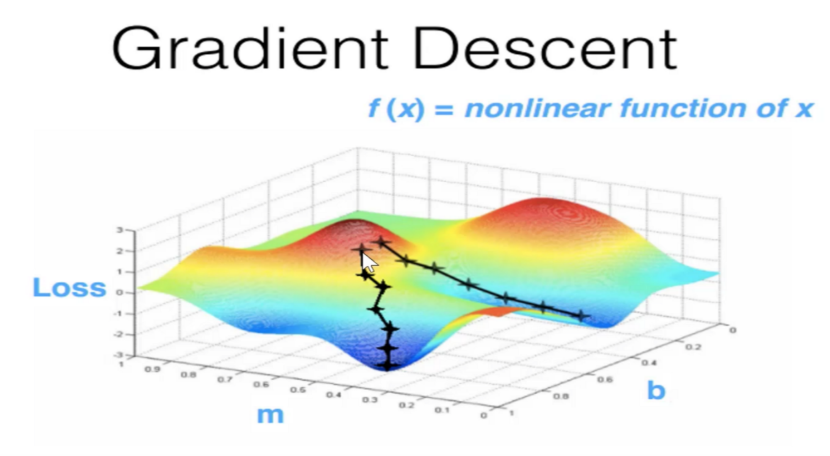
2) 거리척도: 거리를 측정하는 방식에 따라 결과가 크게 달라지는 알고리즘  
① 유클리디안 거리: 두 관측치 사이의 직선 최단거리  
② 맨하탄 거리: 각 좌표축 방향으로만 이동했을 때 거리  
③ 마할라노비스 거리: 변수 내 분산, 공분산, 상관계수를 모두 고려한 거리  
④ 상관거리: Pearson correlation을 거리척도로 활용하여 데이터의 유사도를 반영한 거리

* 장단점

(1) 장점  
 ① Descriptive한 모델로 이해하기 용이하며 많은 조정 없이 좋은 성능을 발휘  
 ② 데이터가 많을수록 매우 일관적인 결과를 도출하며 데이터 분포에 강건

(2) 단점  
① Noise, 무관한 특징이 존재하거나 특징이 중요성과 일치하지 않을 경우 정확도가 감소  
② 데이터가 많아질수록 계산량이 늘어나 예측이 느리고 많은 특성을 처리하는 능력이 떨어짐  
③ 최적의 K선정에 있어 어려움이 있음

**5. SGD (Stochastic Gradient Descent)**

확률적 경사하강법을 이용하여 선형모델을 구현하는 방법이며 convex 형태의 loss 함수를 통한 최적화 문제에 매우 효율적인 접근방법

🡺 구체적으로 loss function을 계산할 때 전체 데이터(batch) 대신 일부 조그마한 데이터의 모음(mini-batch)에 대해서만 loss function을 계산 (본 연구에서는 손실함수를 ‘hinge’, ‘log’, ‘squared hinge’, ‘modified huber’ 로 지정하여 grid search 를 통해 최적의 파라미터 도출함)

그림 14 Gradient Descent

□ 장점  
- 계산 속도가 빠르기 때문에 같은 시간에 더 많은 step을 갈 수 있으며 여러 번 반복할 경우 보통 batch의 결과와 유사한 결과로 수렴  
-  Batch Gradient Descent에서 빠질 local minimum에 빠지지 않고 더 좋은 방향으로 수렴할 가능성  
- 해석하기가 단순하며 회귀 분석에 널리 사용되어 왔지만 선형 매핑만 학습 가능

□ 단점  
- 반복이 늘어날수록 노이즈가 심함  
- 확률적 경사하강법의 여러 변형 함수의 최저점에 가까운 점을 찾을 가능성이 높지만, 항상 보장되는 것은 아님

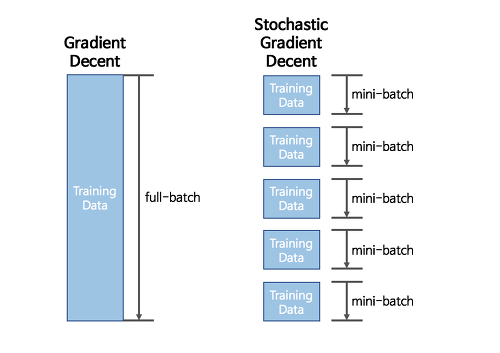
□ 일반적인 Gradient Descent와의 비교   
1) Gradient Decent  
- 모든 데이터를 계산  
- 최적의 한 step을 나아감  
- 매우 느림

그림 15 Gradient Decent 비교

2) Stochastic Gradient Decent  
- 일부 데이터만 계산  
- 빠르게 전진  
- 적절한 learning rate 설정 중요  
🡺 grid search 통해 최적의 파라미터 도출

**6. SVM (Support Vector Machine)**

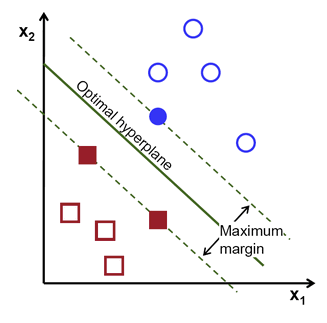
훈련 데이터들을 서로 다른 두 개의 클래스로 분류할 때, 분류의 기준이 되는 분리 경계면을 학습 알고리즘을 이용하여 모색

그림 16 SVM

🡪 입력벡터 x를 고차원의 특정공간으로 매핑한 후 두 클래스 사이의 마진을 최대화하는 분리 경계면을 찾는 것을 목적으로 함

이 때, 최대마진 분리 경계면에 가장 근접한 훈련 데이터를 support vector라 함

□ 장점

- 분류모델을 해결하기 위해 최적의 분리 경계면을 제공  
- 입력공간과 관련된 비선형문제를 고차원의 특정공간의 선형문제로 대응시켜 나타내므로 수학적으로 분석하는 것이 수월  
- 조정해야 할 파라미터의 수가 많지 않아, 비교적 간단하게 학습에 영향 미치는 요소들을 규명 가능  
- 구조적 위험을 최소화 -> 과대적합문제에서 벗어남  
- 볼록함수를 최소화하는 학습을 진행 -> global 최적해를 구할 수 있음

□ 단점

- 데이터 전처리 및 매개변수 설정에 신경 많이 써야 함  
🡺 표준화 및 정규화 데이터별 performance를 grid search를 통해 최적의 파라미터 모색

**[ Result ]**

각 모델들에 대하여 성별, CV, 스케일링, 결측치 제거방안에 대하여 정리한 내용은 아래와 같음.

**■ Naive Bayes Classifier Performance Results  
\*참고 : 각 Performance Measure 기재 시 소수점 넷 째 자리에서 반올림함**

**□ CV = 5 적용**

**[ 청소년 전체 데이터 (남/녀 통합) ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Naïve-Bayes Classifier** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.885 | 0.679 | 0.391 | 0.969 | 0.496 |
| 결측행제거,표준화 | 0.885 | 0.679 | 0.391 | 0.969 | 0.496 |
| 결측행제거,정규화 | 0.885 | 0.679 | 0.391 | 0.969 | 0.496 |
| 평균값대체,원본 | 0.874 | 0.582 | 0.408 | 0.951 | 0.479 |
| 평균값대체,표준화 | 0.874 | 0.582 | 0.408 | 0.951 | 0.479 |
| 평균값대체,정규화 | 0.874 | 0.582 | 0.408 | 0.951 | 0.479 |
| 중앙값대체,원본 | 0.874 | 0.582 | 0.409 | 0.951 | 0.480 |
| 중앙값대체,표준화 | 0.874 | 0.582 | 0.409 | 0.951 | 0.480 |
| 중앙값대체,정규화 | 0.874 | 0.582 | 0.409 | 0.951 | 0.480 |
| 최빈값대체,원본 | 0.873 | 0.579 | 0.409 | 0.950 | 0.479 |
| 최빈값대체,표준화 | 0.873 | 0.579 | 0.409 | 0.950 | 0.479 |
| 최빈값대체,정규화 | 0.873 | 0.579 | 0.409 | 0.950 | 0.479 |

**[ 청소년 남자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Naïve-Bayes Classifier** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.849 | 0.615 | 0.654 | 0.898 | 0.633 |
| 결측행제거,표준화 | 0.849 | 0.614 | 0.654 | 0.898 | 0.633 |
| 결측행제거,정규화 | 0.849 | 0.614 | 0.654 | 0.898 | 0.633 |
| 평균값대체,원본 | 0.835 | 0.571 | 0.607 | 0.890 | 0.588 |
| 평균값대체,표준화 | 0.835 | 0.571 | 0.607 | 0.890 | 0.588 |
| 평균값대체,정규화 | 0.835 | 0.571 | 0.607 | 0.890 | 0.588 |
| 중앙값대체,원본 | 0.835 | 0.571 | 0.606 | 0.890 | 0.588 |
| 중앙값대체,표준화 | 0.835 | 0.571 | 0.606 | 0.890 | 0.588 |
| 중앙값대체,정규화 | 0.835 | 0.571 | 0.606 | 0.890 | 0.588 |
| 최빈값대체,원본 | 0.833 | 0.568 | 0.604 | 0.889 | 0.585 |
| 최빈값대체,표준화 | 0.833 | 0.568 | 0.604 | 0.889 | 0.585 |
| 최빈값대체,정규화 | 0.833 | 0.568 | 0.604 | 0.889 | 0.585 |

**[ 청소년 여자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Naïve-Bayes Classifier** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.917 | 0.626 | 0.256 | 0.984 | 0.363 |
| 결측행제거,표준화 | 0.917 | 0.626 | 0.256 | 0.984 | 0.363 |
| 결측행제거,정규화 | 0.917 | 0.626 | 0.256 | 0.984 | 0.363 |
| 평균값대체,원본 | 0.911 | 0.522 | 0.218 | 0.980 | 0.308 |
| 평균값대체,표준화 | 0.911 | 0.523 | 0.218 | 0.980 | 0.308 |
| 평균값대체,정규화 | 0.911 | 0.523 | 0.218 | 0.980 | 0.308 |
| 중앙값대체,원본 | 0.911 | 0.526 | 0.221 | 0.980 | 0.311 |
| 중앙값대체,표준화 | 0.911 | 0.524 | 0.221 | 0.980 | 0.311 |
| 중앙값대체,정규화 | 0.911 | 0.524 | 0.221 | 0.980 | 0.311 |
| 최빈값대체,원본 | 0.911 | 0.522 | 0.224 | 0.979 | 0.314 |
| 최빈값대체,표준화 | 0.911 | 0.522 | 0.224 | 0.979 | 0.314 |
| 최빈값대체,정규화 | 0.911 | 0.522 | 0.224 | 0.979 | 0.314 |

□ **CV = 10 적용**

**[ 청소년 전체 데이터 (남/녀 통합) ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Naïve-Bayes Classifier** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.885 | 0.679 | 0.391 | 0.969 | 0.496 |
| 결측행제거,표준화 | 0.885 | 0.679 | 0.391 | 0.969 | 0.496 |
| 결측행제거,정규화 | 0.885 | 0.679 | 0.391 | 0.969 | 0.496 |
| 평균값대체,원본 | 0.873 | 0.581 | 0.407 | 0.951 | 0.478 |
| 평균값대체,표준화 | 0.874 | 0.582 | 0.408 | 0.951 | 0.479 |
| 평균값대체,정규화 | 0.874 | 0.582 | 0.408 | 0.951 | 0.479 |
| 중앙값대체,원본 | 0.874 | 0.582 | 0.407 | 0.951 | 0.479 |
| 중앙값대체,표준화 | 0.874 | 0.582 | 0.409 | 0.951 | 0.480 |
| 중앙값대체,정규화 | 0.874 | 0.582 | 0.409 | 0.951 | 0.480 |
| 최빈값대체,원본 | 0.873 | 0.578 | 0.407 | 0.950 | 0.477 |
| 최빈값대체,표준화 | 0.873 | 0.579 | 0.409 | 0.950 | 0.479 |
| 최빈값대체,정규화 | 0.873 | 0.579 | 0.409 | 0.950 | 0.479 |

**[ 청소년 남자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Naïve-Bayes Classifier** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.849 | 0.614 | 0.654 | 0.898 | 0.633 |
| 결측행제거,표준화 | 0.849 | 0.614 | 0.654 | 0.898 | 0.633 |
| 결측행제거,정규화 | 0.849 | 0.614 | 0.654 | 0.898 | 0.633 |
| 평균값대체,원본 | 0.835 | 0.571 | 0.607 | 0.890 | 0.588 |
| 평균값대체,표준화 | 0.835 | 0.571 | 0.607 | 0.890 | 0.588 |
| 평균값대체,정규화 | 0.835 | 0.571 | 0.607 | 0.890 | 0.588 |
| 중앙값대체,원본 | 0.835 | 0.571 | 0.606 | 0.890 | 0.588 |
| 중앙값대체,표준화 | 0.835 | 0.571 | 0.606 | 0.890 | 0.588 |
| 중앙값대체,정규화 | 0.835 | 0.571 | 0.606 | 0.890 | 0.588 |
| 최빈값대체,원본 | 0.833 | 0.568 | 0.603 | 0.889 | 0.585 |
| 최빈값대체,표준화 | 0.833 | 0.568 | 0.604 | 0.889 | 0.585 |
| 최빈값대체,정규화 | 0.833 | 0.568 | 0.604 | 0.889 | 0.585 |

**[ 청소년 여자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Naïve-Bayes Classifier** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.917 | 0.626 | 0.256 | 0.984 | 0.363 |
| 결측행제거,표준화 | 0.917 | 0.626 | 0.256 | 0.984 | 0.363 |
| 결측행제거,정규화 | 0.917 | 0.624 | 0.256 | 0.984 | 0.363 |
| 평균값대체,원본 | 0.911 | 0.523 | 0.218 | 0.980 | 0.308 |
| 평균값대체,표준화 | 0.911 | 0.523 | 0.218 | 0.980 | 0.308 |
| 평균값대체,정규화 | 0.911 | 0.523 | 0.218 | 0.980 | 0.308 |
| 중앙값대체,원본 | 0.911 | 0.526 | 0.221 | 0.980 | 0.311 |
| 중앙값대체,표준화 | 0.911 | 0.524 | 0.221 | 0.980 | 0.311 |
| 중앙값대체,정규화 | 0.911 | 0.524 | 0.221 | 0.980 | 0.311 |
| 최빈값대체,원본 | 0.911 | 0.522 | 0.224 | 0.979 | 0.314 |
| 최빈값대체,표준화 | 0.911 | 0.522 | 0.224 | 0.979 | 0.314 |
| 최빈값대체,정규화 | 0.911 | 0.522 | 0.224 | 0.979 | 0.314 |

□ 데이터 (남녀 통합, 남자/여자 데이터) 기준 해석

* 나이브 베이즈 분류 분석 결과를 데이터 기준으로 해석한 결과는 아래와 같음

1. 남녀 통합 청소년 데이터  
   : CV = 5, CV = 10인 경우를 통틀어 Recall이 가장 높은 경우는 Recall이 0.409  
    CV = 5, CV = 10 통합 평균 Accuarcy : 0.876, 평균 Recall : 0.404, 평균 F1 score : 0.483
2. 청소년 남자 데이터  
   : CV = 5, CV = 10인 경우를 통틀어 Recall이 가장 높은 경우는 Recall이 0.654  
    CV = 5, CV = 10 통합 평균 Accuracy : 0.838, 평균 Recall : 0.618, F1 score : 0.599
3. 청소년 여자 데이터  
   : CV = 5, CV = 10인 경우를 통틀어 Revall이 가장 높은 경우는 Recall이 0.256  
    CV = 5, CV = 10 통합 평균 Accuracy : 0.913, Recall : 0.230, F1 score : 0.324

🡺 청소년 여자 데이터의 평균 Accurcy가 청소년 남자 데이터의 평균 Accuracy보다 약 0.075 높지만 본 연구에서의 주 관심 측도인 Recall은 여자가 남자보다 평균 0.388 낮음

🡺 Accuracy는 평균적으로 여자가 남자보다 높음. 그러나 Recall은 평균적으로 여자가 남자보다 낮음. 여자 청소년의 경우 비만율이 남자에 비해 절반 이하이기 때문에 비만인 경우에 대한 학습에 사용할 데이터 부족으로 인해 Recall이 작은 것으로 판단. 이 과정에서 모델의 성능을 단순히 Accuracy만을 이용해 판단해서는 안 됨을 확인함

□ 결측치 처리 (결측치 있는 행 제거, 평균값/중앙값/최빈값 대체) 기준 해석

: 남녀 통합 청소년 데이터에서는 중앙값 대체, 최빈값 대체를 적용했을 경우가 Test Data에 대한 퍼포먼스가 가장 높게 나옴. CV = 5와 CV = 10의 경우 둘 다 중앙값 대체, 최빈값 대체 적용 시 최고 퍼포먼스는 0.409로 동일함.

청소년 남자 데이터와 청소년 여자 데이터에서는 결측치가 있는 행을 제거하는 방법 즉, 완전 제거법을 사용한 경우가 평균값 대체, 중앙값 대체, 최빈값 대체와 같은 단일 대체법을 이용한 경우보다 Test Data에 대한 모델의 퍼포먼스가 높음.

* 동일한 기법을 사용하더라도 모델의 Training에 사용된 데이터에 어떤 결측치 처리 방법을 적용하는지에 따라 모델의 퍼포먼스가 영향을 받음을 확인함

□ 스케일링 (원본, 표준화, 정규화) 기준 해석

- 나이브 베이즈 분류 분석의 경우 분석 전 입력 변수 값들의 표준화, 정규화 등의 스케일링이 필요한지 여부에 대한 정보가 매우 적었으며 필요 여부 의견이 나뉨

- 본 연구에서는 연속형 변수인 “주중 하루 평균 수면 시간”과 “하루 평균 앉아있는 시간” 변수를 표준화 적용한 경우, 정규화를 적용한 경우, 원본 그대로를 사용한 경우에 대해 모두 모델을 수립하여 결과를 비교함

- 남녀 통합 데이터의 경우 CV = 5인 경우, 원본을 사용한 경우, 표준화 혹은 정규화를 적용한 경우 모델의 퍼포먼스에는 차이가 없었음. CV = 10인 경우, 표준화 및 정규화를 적용한 경우가 원본을 사용한 경우와 Accuracy는 같았지만, Recall은 0.002가 높았음

- 남자 데이터 및 여자 데이터에서는 원본을 사용한 경우, 표준화 혹은 정규화를 적용한 경우 모델의 퍼포먼스에는 차이가 없었음.

🡺 남녀 통합 데이터에 CV = 10인 경우를 제외하고는 원본, 표준화 혹은 정규화 적용에 상관없이 모델의 퍼포먼스는 같음. CV = 10인 경우 남녀 통합 데이터에서 표준화 및 정규화를 적용했을 경우 원본을 사용한 경우 수치 비교를 하면 차이가 있긴 하나 0.002로 매우 미미함.

🡺 나이브 베이즈 분류기에서는 표준화 및 정규화와 같은 스케일링의 효과가 없다고 판단

그러나 본 연구에서는 세 가지 데이터를 기준으로 스케일링의 효과를 확인하였기에 한계가 존재하며 남녀 통합 데이터에 CV = 10인 경우 어떠한 이유로 Recall에 차이가 존재하였는지 추가적인 연구가 필요.

□ CV 기준 해석

: 모델 수립 시 5-fold Cross Validation과 10-fold Cross Validation을 수행한 결과는 아래와 같음  
\*참고 : 아래 표에는 CV = 5와 CV = 10 각각의 모델 퍼포먼스의 평균치를 기재함

**[ 남녀 통합 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CV** | **Accuracy** | **Precision** | **Recall** | **Specificity** | **F1 Score** |
| **5** | 0.877 | 0.605 | 0.404 | 0.955 | 0.484 |
| **10** | 0.876 | 0.605 | 0.404 | 0.955 | 0.483 |

**[ 남자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CV** | **Accuracy** | **Precision** | **Recall** | **Specificity** | **F1 Score** |
| **5** | 0.838 | 0.581 | 0.618 | 0.892 | 0.599 |
| **10** | 0.838 | 0.581 | 0.618 | 0.892 | 0.599 |

**[ 여자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CV** | **Accuracy** | **Precision** | **Recall** | **Specificity** | **F1 Score** |
| **5** | 0.913 | 0.549 | 0.230 | 0.981 | 0.324 |
| **10** | 0.913 | 0.549 | 0.230 | 0.981 | 0.324 |

- 남녀 통합 데이터에서 CV = 5 인 경우와 CV = 10인 경우 Accuracy와 F1 Score에서 소수점 셋째 자리에서 1의 차이가 있으나 이는 소수점 넷째 자리에서 반올림해서 기재를 하는 중 발생했기에 사실상 Accuracy와 F1 Score에서의 차이는 이보다 작음.

🡺 K-fold Cross Validation에서 K 값의 변화는 나이브 베이즈 분류기의 성능에 매우 적거나 없음을 확인

**■ Random Forest Performance Results  
\*참고 : 각 Performance Measure 기재 시 소수점 넷 째 자리에서 반올림함**

**□ CV = 5 적용**

**[ 청소년 전체 데이터 (남/녀 통합) ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Random Forest** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.891 | 0.659 | 0.516 | 0.955 | 0.579 |
| 결측행제거,표준화 | 0.893 | 0.672 | 0.508 | 0.958 | 0.579 |
| 결측행제거,정규화 | 0.891 | 0.66 | 0.517 | 0.955 | 0.58 |
| 평균값대체,원본 | 0.875 | 0.586 | 0.429 | 0.95 | 0.495 |
| 평균값대체,표준화 | 0.875 | 0.586 | 0.429 | 0.949 | 0.495 |
| 평균값대체,정규화 | 0.875 | 0.586 | 0.429 | 0.949 | 0.495 |
| 중앙값대체,원본화 | 0.876 | 0.59 | 0.437 | 0.949 | 0.502 |
| 중앙값대체,표준화 | 0.876 | 0.59 | 0.436 | 0.949 | 0.502 |
| 중앙값대체,정규화 | 0.876 | 0.59 | 0.436 | 0.949 | 0.502 |
| 최빈값대체,원본화 | 0.876 | 0.586 | 0.449 | 0.947 | 0.508 |
| 최빈값대체,표준화 | 0.876 | 0.586 | 0.449 | 0.947 | 0.508 |
| 최빈값대체,정규화 | 0.876 | 0.586 | 0.449 | 0.947 | 0.508 |

**[ 청소년 남자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Random Forest** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.856 | 0.652 | 0.599 | 0.92 | 0.624 |
| 결측행제거,표준화 | 0.856 | 0.652 | 0.599 | 0.92 | 0.624 |
| 결측행제거,정규화 | 0.856 | 0.652 | 0.599 | 0.92 | 0.624 |
| 평균값대체,원본 | 0.852 | 0.633 | 0.567 | 0.92 | 0.598 |
| 평균값대체,표준화 | 0.851 | 0.632 | 0.566 | 0.92 | 0.597 |
| 평균값대체,정규화 | 0.851 | 0.632 | 0.566 | 0.92 | 0.598 |
| 중앙값대체,원본화 | 0.852 | 0.633 | 0.569 | 0.92 | 0.599 |
| 중앙값대체,표준화 | 0.852 | 0.632 | 0.569 | 0.92 | 0.599 |
| 중앙값대체,정규화 | 0.852 | 0.632 | 0.569 | 0.92 | 0.599 |
| 최빈값대체,원본화 | 0.849 | 0.62 | 0.574 | 0.915 | 0.596 |
| 최빈값대체,표준화 | 0.848 | 0.619 | 0.573 | 0.915 | 0.595 |
| 최빈값대체,정규화 | 0.849 | 0.62 | 0.574 | 0.915 | 0.596 |

**[ 청소년 여자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Random Forest** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.926 | 0.701 | 0.355 | 0.985 | 0.472 |
| 결측행제거,표준화 | 0.926 | 0.701 | 0.355 | 0.985 | 0.472 |
| 결측행제거,정규화 | 0.926 | 0.701 | 0.355 | 0.985 | 0.472 |
| 평균값대체,원본 | 0.91 | 0.521 | 0.199 | 0.982 | 0.288 |
| 평균값대체,표준화 | 0.91 | 0.519 | 0.199 | 0.981 | 0.288 |
| 평균값대체,정규화 | 0.91 | 0.521 | 0.199 | 0.982 | 0.288 |
| 중앙값대체,원본화 | 0.913 | 0.628 | 0.111 | 0.993 | 0.189 |
| 중앙값대체,표준화 | 0.913 | 0.628 | 0.111 | 0.993 | 0.189 |
| 중앙값대체,정규화 | 0.913 | 0.628 | 0.111 | 0.993 | 0.189 |
| 최빈값대체,원본화 | 0.911 | 0.577 | 0.11 | 0.992 | 0.185 |
| 최빈값대체,표준화 | 0.912 | 0.58 | 0.111 | 0.992 | 0.187 |
| 최빈값대체,정규화 | 0.912 | 0.58 | 0.111 | 0.992 | 0.187 |

□ **CV = 10 적용**

**[ 청소년 전체 데이터 (남/녀 통합) ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Random Forest** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.89 | 0.662 | 0.498 | 0.957 | 0.568 |
| 결측행제거,표준화 | 0.89 | 0.662 | 0.498 | 0.957 | 0.568 |
| 결측행제거,정규화 | 0.891 | 0.66 | 0.517 | 0.955 | 0.58 |
| 평균값대체,원본 | 0.875 | 0.586 | 0.429 | 0.95 | 0.495 |
| 평균값대체,표준화 | 0.875 | 0.586 | 0.429 | 0.949 | 0.495 |
| 평균값대체,정규화 | 0.875 | 0.586 | 0.429 | 0.949 | 0.495 |
| 중앙값대체,원본화 | 0.876 | 0.59 | 0.437 | 0.949 | 0.502 |
| 중앙값대체,표준화 | 0.876 | 0.59 | 0.436 | 0.949 | 0.502 |
| 중앙값대체,정규화 | 0.876 | 0.59 | 0.436 | 0.949 | 0.502 |
| 최빈값대체,원본화 | 0.877 | 0.592 | 0.436 | 0.95 | 0.503 |
| 최빈값대체,표준화 | 0.877 | 0.592 | 0.437 | 0.95 | 0.503 |
| 최빈값대체,정규화 | 0.877 | 0.592 | 0.436 | 0.95 | 0.503 |

**[ 청소년 남자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Random Forest** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.857 | 0.655 | 0.596 | 0.922 | 0.625 |
| 결측행제거,표준화 | 0.857 | 0.656 | 0.597 | 0.922 | 0.625 |
| 결측행제거,정규화 | 0.857 | 0.655 | 0.596 | 0.921 | 0.624 |
| 평균값대체,원본 | 0.851 | 0.629 | 0.576 | 0.918 | 0.601 |
| 평균값대체,표준화 | 0.851 | 0.629 | 0.576 | 0.918 | 0.601 |
| 평균값대체,정규화 | 0.851 | 0.629 | 0.576 | 0.918 | 0.601 |
| 중앙값대체,원본화 | 0.851 | 0.631 | 0.566 | 0.92 | 0.597 |
| 중앙값대체,표준화 | 0.852 | 0.632 | 0.569 | 0.92 | 0.599 |
| 중앙값대체,정규화 | 0.852 | 0.632 | 0.569 | 0.92 | 0.599 |
| 최빈값대체,원본화 | 0.849 | 0.617 | 0.587 | 0.912 | 0.602 |
| 최빈값대체,표준화 | 0.852 | 0.629 | 0.58 | 0.917 | 0.604 |
| 최빈값대체,정규화 | 0.852 | 0.632 | 0.569 | 0.92 | 0.599 |

**[ 청소년 여자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Random Forest** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.926 | 0.699 | 0.343 | 0.985 | 0.46 |
| 결측행제거,표준화 | 0.926 | 0.7 | 0.345 | 0.985 | 0.462 |
| 결측행제거,정규화 | 0.926 | 0.696 | 0.347 | 0.985 | 0.463 |
| 평균값대체,원본 | 0.909 | 0.513 | 0.145 | 0.986 | 0.226 |
| 평균값대체,표준화 | 0.909 | 0.51 | 0.143 | 0.986 | 0.224 |
| 평균값대체,정규화 | 0.909 | 0.508 | 0.143 | 0.986 | 0.224 |
| 중앙값대체,원본화 | 0.912 | 0.557 | 0.151 | 0.988 | 0.237 |
| 중앙값대체,표준화 | 0.913 | 0.628 | 0.111 | 0.993 | 0.189 |
| 중앙값대체,정규화 | 0.913 | 0.628 | 0.111 | 0.993 | 0.189 |
| 최빈값대체,원본화 | 0.914 | 0.582 | 0.202 | 0.985 | 0.3 |
| 최빈값대체,표준화 | 0.914 | 0.585 | 0.202 | 0.986 | 0.3 |
| 최빈값대체,정규화 | 0.914 | 0.585 | 0.202 | 0.986 | 0.3 |

□ 데이터 (남,녀 통합/남자/여자 데이터) 기준 해석

1. 남녀 통합 청소년 데이터  
   : CV = 5, CV = 10인 경우를 통틀어 Recall이 가장 높은 경우는 Recall이 0.517   
    CV = 5, CV = 10 통합 평균 Accuarcy : 0.880, 평균 Recall : 0.454, 평균 F1 score : 0.520
2. 남자 청소년 데이터  
   : CV = 5, CV = 10인 경우에서 각각 Recall이 가장 높은 경우를 기준으로 Recall이 0.599  
    CV = 5, CV = 10 통합 평균 Accuarcy : 0.852, 평균 Recall : 0.578, 평균 F1 score : 0.605
3. 여자 청소년 데이터  
   : CV = 5, CV = 10을 통틀어 Recall이 가장 높은 경우 0.355  
    CV = 5, CV = 10 통합 평균 Accuarcy : 0.915, 평균 Recall : 0.199, 평균 F1 score : 0.291

:여자 청소년의 평균 Accuracy는 남/녀 통합과 남자 청소년 데이터의 평균 Accuracy보다 높은 것을 볼 수 있고 본 연구에서의 주 관심 측도인 Recall은 여자 청소년이 가장 낮은 것을 볼 수 있음

: Recall을 기준으로 봤을 때 ‘남자 청소년 > 남/녀 통합 > 여자 청소년’ 순으로 높은 성능을 나타냄.

* 논문에서 추출한 인자들의 경우 남자 청소년 > 남/녀 통합 > 여자 청소년 순으로 적합할 것으로 판단

□ 결측치 처리 (결측치 있는 행 제거, 평균값/중앙값/최빈값 대체) 기준 해석

: 결측치 처리 방안에 대해서는 Recall이 가장 높은 경우는 결측치가 있는 행을 제거한 것이 남녀 통합, 남자, 여자 모두를 비교해보았을 때 가장 성능이 좋음.

: 남/녀 통합과 남자 청소년에 대해서는 단일대체법(평균값 대체, 중앙값 대체, 최빈값 대체)에 대해서는 큰 차이를 찾아볼 수 없을 만큼 비슷하였으나 여자 청소년에 대해서는 ‘최빈값 대체 > 중앙값 대체 > 평균값 대체’ 순으로 좋은 성능을 보임.

* 랜덤 포레스트에서는 단일 대체법을 활용한 방법론보다 결측행을 제거하는 방식이 좋은 효과를 보임

□ 스케일링 (원본, 표준화, 정규화) 기준 해석

: 스케일링(표준화, 정규화)을 한 것과 안 한 것의 차이가 가장 드러나지 않는 모델로 Recall 뿐만 아니라 나머지 척도에 대해서도 남/녀 통합, 남자 청소년, 여자 청소년에서 모두 큰 차이를 찾아볼 수 없음

* 랜덤 포레스트의 경우 스케일링이 모델 성능에 큰 영향을 끼치지 않는다는 사실을 확인함

□ CV 기준 해석

: 5-fold Cross Validation과 10-fold Cross Validation을 수행한 결과는 아래와 같음

참고: 아래 표에는 CV = 5와 CV = 10 각각의 모델 퍼포먼스의 평균치를 기재함

**[ 남녀 통합 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CV** | **Accuracy** | **Precision** | **Recall** | **Specificity** | **F1 Score** |
| **5** | 0.880 | 0.606 | 0.457 | 0.950 | 0.521 |
| **10** | 0.880 | 0.607 | 0.452 | 0.951 | 0.518 |

**[ 남자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CV** | **Accuracy** | **Precision** | **Recall** | **Specificity** | **F1 Score** |
| **5** | 0.852 | 0.634 | 0.577 | 0.919 | 0.604 |
| **10** | 0.853 | 0.636 | 0.580 | 0.919 | 0.606 |

**[ 여자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CV** | **Accuracy** | **Precision** | **Recall** | **Specificity** | **F1 Score** |
| **5** | 0.915 | 0.607 | 0.194 | 0.988 | 0.284 |
| **10** | 0.915 | 0.599 | 0.204 | 0.987 | 0.298 |

: CV의 경우 5보다 10으로 정하는 것처럼 더 크게 설정할수록 대체로 결과가 좋아짐을 볼 수 있었으나 ‘여자 청소년: 평균값 대체’와 같이 오히려 성능이 하락한 케이스도 있음

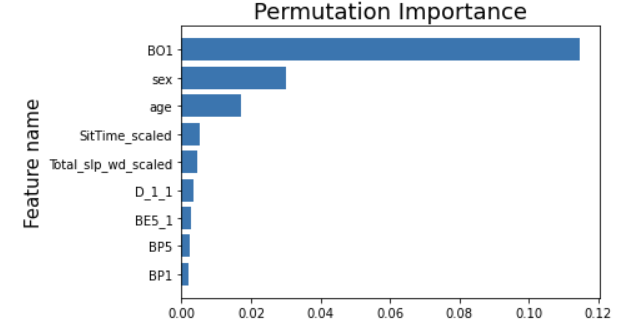
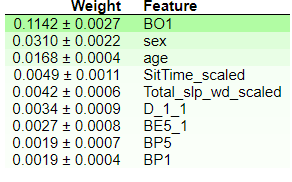
* CV를 높인 결과 성능이 향상된 것이 많아 교차검증의 K를 늘리는 것은 성능을 높이는데 합리적이라고 판단할 수 있으나 모델을 만드는 시간적 비용이 더 많이 들어가 성능을 좋게 하기 위하여 무조건 크게 잡는 것은 좋은 판단이 아님을 확인

**□ Random Forest활용 변수 중요도 확인  
\*참고: sklearn 패키지의 permutation\_importance를 활용하여 변수중요도 추출**

1. 청소년 전체 데이터 (cv = 5, 10) : 결측행 제거, 정규화 (결과 동일함)

1-1) cv=5

* 최적 파라미터 : max\_depth = 10, min\_samples\_leaf = 8, mim\_samples\_split = 20, n\_estimators = 100
* 변수 중요도



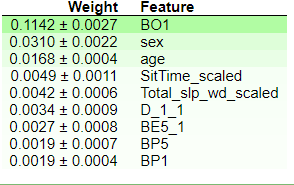
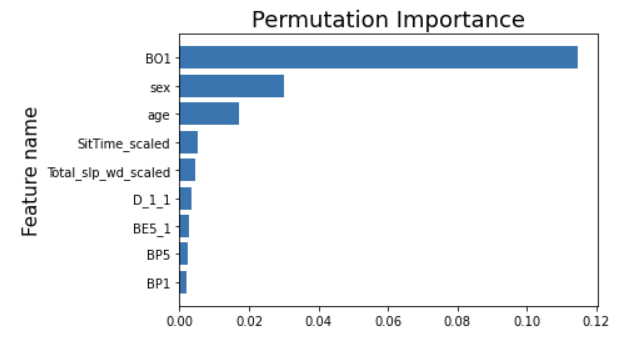
* Permutation Importance 해석 예시

: 위 표에 나타나는 Weight는 value를 shuffling 후 shuffling 전과 비교했을 때 Score가 떨어진 정도를 보여줌

설명변수 ‘BO1’에 대해 0.1142만큼 Score가 가장 많이 떨어졌다면 이는 해당 feature가 예측에 가장 영향력이 있으며 모델이 해당 변수에 가장 많이 의존한다고 해석가능

1-2) cv = 10

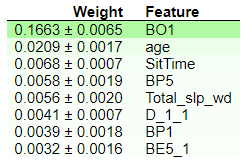
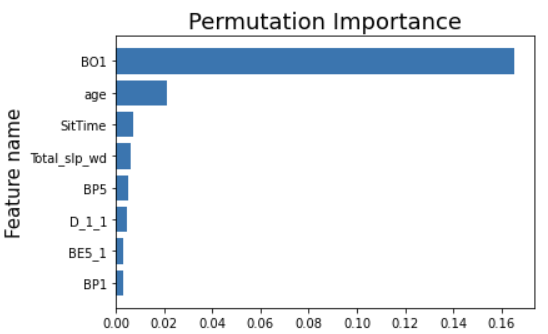
* 최적 파라미터 : max\_depth = 10, min\_samples\_leaf = 8, mim\_samples\_split = 20, n\_estimators = 100
* 변수 중요도

1. 남자 데이터 ( cv = 5) : 결측행 제거, 원본/표준화/정규화 (결과 동일함)

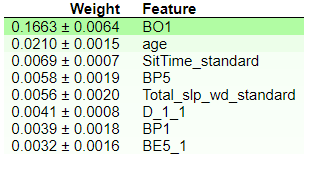
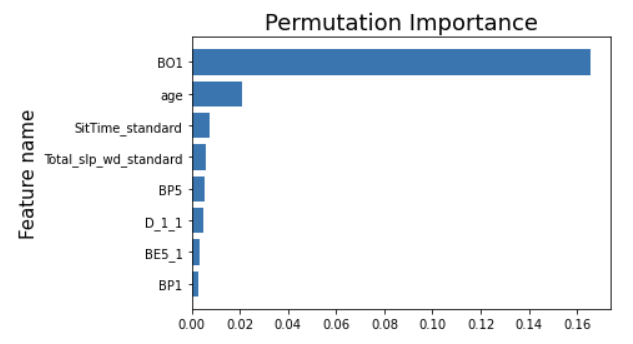
2-1) 원본

* 최적 파라미터 : max\_depth = 8, min\_samples\_leaf = 12, mim\_samples\_split = 8, n\_estimators = 10
* 변수 중요도

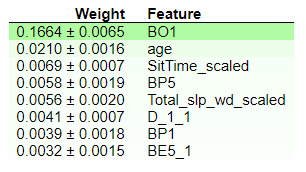
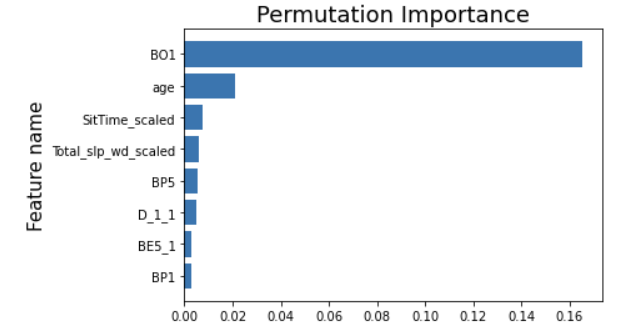
2-2) 표준화

* 최적 파라미터 : max\_depth = 8, min\_samples\_leaf = 12, mim\_samples\_split = 8, n\_estimators = 10
* 변수 중요도

2\_3) 정규화

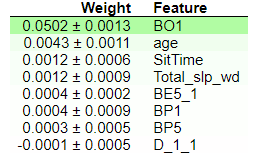
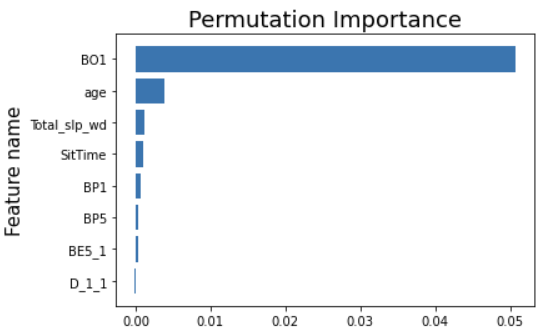
* 최적 파라미터 : max\_depth = 8, min\_samples\_leaf = 12, mim\_samples\_split = 8, n\_estimators = 10
* 변수 중요도

1. 여자 데이터 ( cv = 5) : 결측행 제거, 원본/표준화/정규화 (결과 동일함)

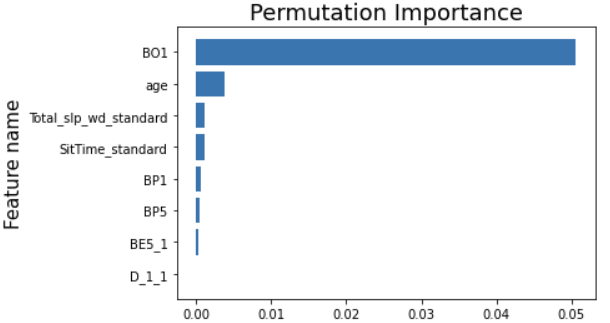
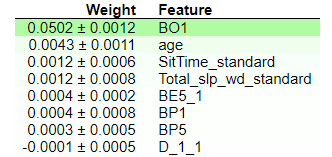
3-1) 원본

* 최적 파라미터 : max\_depth = 12, min\_samples\_leaf = 12, mim\_samples\_split = 8, n\_estimators = 100
* 변수 중요도

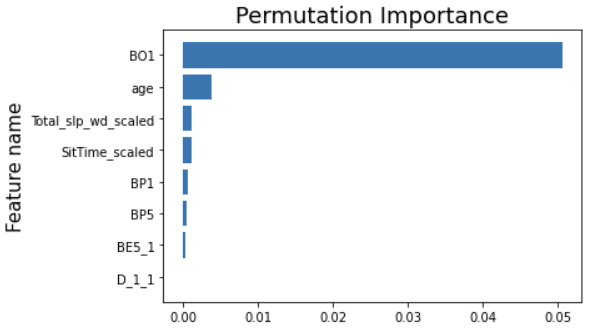
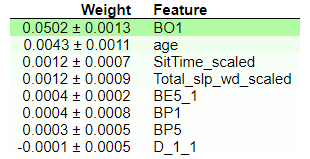
3-2) 표준화

* 최적 파라미터 : max\_depth = 12, min\_samples\_leaf = 12, mim\_samples\_split = 8, n\_estimators = 10
* 변수 중요도



3-3) 정규화

* 최적 파라미터 : max\_depth = 8, min\_samples\_leaf = 12, mim\_samples\_split = 8, n\_estimators = 10
* 변수 중요도



* 변수 중요도 확인 결과 해석

: 모델의 Confusion matrix가 완전히 동일하여 5개의 척도가 비슷한 것들을 열거하고 변수 중요도를 확인한 결과 완전히 동일하진 않다는 것을 확인

* 모델의 예측 결과는 동일하지만 데이터가 다르면 중요도의 정량화된 값은 바뀔 수 있음을 확인

: [데이터 (남/녀 통합/남자/여자 데이터) 기준 해석] 에서 본 것처럼 ‘남자 청소년 > 남/녀 통합 > 여자 청소년’ 순으로 예측 성능이 좋았는데 이에 비례하여 변수 중요도도 동일한 흐름을 따라감

* 세 그룹에서 가장 중요도가 높은 [BO1]의 경우 변수 중요도가 남자 청소년이 여자 청소년 보다 3배 높은 경향을 보임

: 유의미한 인자로 상위 3개 [BO1], [age], [SitTime]이 모두 동일하게 선정되었으며 그 이하 항목들에 대해서는 데이터 별로 다른 성향을 보임

* 각 남/녀 통합, 남자 청소년, 여자 청소년에 대하여 유의미한 비만 인자는 [BO1], [age], [SitTime] 으로 결론지을 수 있고 그 이하 항목들은 그룹의 특성에 따라 다름을 확인

**■ Logistic Regression Performance Results  
\*참고 : 각 Performance Measure 기재 시 소수점 넷 째 자리에서 반올림함**

**□ CV = 5 적용**

**[ 청소년 전체 데이터 (남/녀 통합) ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Logistic Regression** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.878 | 0.622 | 0.399 | 0.959 | 0.486 |
| 결측행제거,표준화 | 0.891 | 0.670 | 0.492 | 0.959 | 0.567 |
| 결측행제거,정규화 | 0.891 | 0.670 | 0.492 | 0.959 | 0.567 |
| 평균값대체,원본 | 0.868 | 0.572 | 0.287 | 0.964 | 0.383 |
| 평균값대체,표준화 | 0.877 | 0.627 | 0.342 | 0.966 | 0.443 |
| 평균값대체,정규화 | 0.877 | 0.627 | 0.342 | 0.966 | 0.443 |
| 중앙값대체,원본 | 0.868 | 0.568 | 0.305 | 0.961 | 0.397 |
| 중앙값대체,표준화 | 0.877 | 0.627 | 0.342 | 0.966 | 0.442 |
| 중앙값대체,정규화 | 0.877 | 0.627 | 0.342 | 0.966 | 0.442 |
| 최빈값대체,원본 | 0.870 | 0.595 | 0.272 | 0.970 | 0.374 |
| 최빈값대체,표준화 | 0.877 | 0.629 | 0.337 | 0.967 | 0.439 |
| 최빈값대체,정규화 | 0.877 | 0.627 | 0.337 | 0.967 | 0.438 |

**[ 청소년 남자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Logistic Regression** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.834 | 0.627 | 0.414 | 0.938 | 0.499 |
| 결측행제거,표준화 | 0.859 | 0.678 | 0.561 | 0.933 | 0.614 |
| 결측행제거,정규화 | 0.861 | 0.683 | 0.563 | 0.935 | 0.617 |
| 평균값대체,원본 | 0.829 | 0.610 | 0.340 | 0.948 | 0.436 |
| 평균값대체,표준화 | 0.842 | 0.672 | 0.367 | 0.957 | 0.474 |
| 평균값대체,정규화 | 0.842 | 0.674 | 0.367 | 0.957 | 0.475 |
| 중앙값대체,원본 | 0.831 | 0.678 | 0.254 | 0.971 | 0.369 |
| 중앙값대체,표준화 | 0.842 | 0.672 | 0.367 | 0.957 | 0.474 |
| 중앙값대체,정규화 | 0.842 | 0.673 | 0.367 | 0.957 | 0.475 |
| 최빈값대체,원본 | 0.830 | 0.676 | 0.247 | 0.971 | 0.362 |
| 최빈값대체,표준화 | 0.842 | 0.682 | 0.351 | 0.960 | 0.464 |
| 최빈값대체,정규화 | 0.842 | 0.685 | 0.349 | 0.961 | 0.462 |

**[ 청소년 여자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Logistic Regression** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.917 | 0.660 | 0.213 | 0.989 | 0.321 |
| 결측행제거,표준화 | 0.925 | 0.680 | 0.366 | 0.982 | 0.476 |
| 결측행제거,정규화 | 0.925 | 0.680 | 0.366 | 0.982 | 0.476 |
| 평균값대체,원본 | 0.913 | 0.540 | 0.324 | 0.972 | 0.405 |
| 평균값대체,표준화 | 0.913 | 0.538 | 0.324 | 0.972 | 0.404 |
| 평균값대체,정규화 | 0.915 | 0.574 | 0.243 | 0.982 | 0.342 |
| 중앙값대체,원본 | 0.913 | 0.542 | 0.338 | 0.971 | 0.417 |
| 중앙값대체,표준화 | 0.913 | 0.538 | 0.324 | 0.972 | 0.404 |
| 중앙값대체,정규화 | 0.915 | 0.574 | 0.243 | 0.982 | 0.342 |
| 최빈값대체,원본 | 0.909 | 0.857 | 0.009 | 0.999 | 0.017 |
| 최빈값대체,표준화 | 0.913 | 0.538 | 0.324 | 0.972 | 0.404 |
| 최빈값대체,정규화 | 0.914 | 0.570 | 0.243 | 0.982 | 0.341 |

□ **CV = 10 적용**

**[ 청소년 전체 데이터 (남/녀 통합) ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Logistic Regression** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.891 | 0.671 | 0.489 | 0.989 | 0.566 |
| 결측행제거,표준화 | 0.891 | 0.670 | 0.492 | 0.959 | 0.567 |
| 결측행제거,정규화 | 0.891 | 0.670 | 0.493 | 0.959 | 0.568 |
| 평균값대체,원본 | 0.876 | 0.635 | 0.313 | 0.970 | 0.419 |
| 평균값대체,표준화 | 0.877 | 0.627 | 0.341 | 0.966 | 0.442 |
| 평균값대체,정규화 | 0.877 | 0.627 | 0.341 | 0.966 | 0.442 |
| 중앙값대체,원본 | 0.877 | 0.635 | 0.322 | 0.969 | 0.428 |
| 중앙값대체,표준화 | 0.877 | 0.626 | 0.341 | 0.966 | 0.442 |
| 중앙값대체,정규화 | 0.877 | 0.626 | 0.341 | 0.966 | 0.442 |
| 최빈값대체,원본 | 0.877 | 0.629 | 0.335 | 0.967 | 0.437 |
| 최빈값대체,표준화 | 0.877 | 0.628 | 0.337 | 0.967 | 0.439 |
| 최빈값대체,정규화 | 0.877 | 0.630 | 0.336 | 0.967 | 0.439 |

**[ 청소년 남자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Logistic Regression** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.859 | 0.680 | 0.550 | 0.936 | 0.609 |
| 결측행제거,표준화 | 0.859 | 0.677 | 0.560 | 0.933 | 0.613 |
| 결측행제거,정규화 | 0.860 | 0.684 | 0.563 | 0.935 | 0.618 |
| 평균값대체,원본 | 0.842 | 0.682 | 0.358 | 0.960 | 0.470 |
| 평균값대체,표준화 | 0.842 | 0.678 | 0.359 | 0.959 | 0.469 |
| 평균값대체,정규화 | 0.842 | 0.673 | 0.364 | 0.957 | 0.473 |
| 중앙값대체,원본 | 0.842 | 0.692 | 0.335 | 0.964 | 0.452 |
| 중앙값대체,표준화 | 0.842 | 0.678 | 0.359 | 0.959 | 0.469 |
| 중앙값대체,정규화 | 0.842 | 0.674 | 0.360 | 0.958 | 0.470 |
| 최빈값대체,원본 | 0.841 | 0.691 | 0.330 | 0.964 | 0.447 |
| 최빈값대체,표준화 | 0.841 | 0.681 | 0.351 | 0.960 | 0.463 |
| 최빈값대체,정규화 | 0.842 | 0.684 | 0.349 | 0.961 | 0.462 |

**[ 청소년 여자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Logistic Regression** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.928 | 0.683 | 0.406 | 0.951 | 0.509 |
| 결측행제거,표준화 | 0.928 | 0.683 | 0.406 | 0.981 | 0.509 |
| 결측행제거,정규화 | 0.928 | 0.683 | 0.406 | 0.981 | 0.509 |
| 평균값대체,원본 | 0.913 | 0.536 | 0.327 | 0.972 | 0.406 |
| 평균값대체,표준화 | 0.913 | 0.537 | 0.340 | 0.971 | 0.416 |
| 평균값대체,정규화 | 0.913 | 0.538 | 0.332 | 0.971 | 0.411 |
| 중앙값대체,원본 | 0.913 | 0.539 | 0.332 | 0.971 | 0.411 |
| 중앙값대체,표준화 | 0.913 | 0.537 | 0.340 | 0.971 | 0.416 |
| 중앙값대체,정규화 | 0.913 | 0.538 | 0.332 | 0.971 | 0.411 |
| 최빈값대체,원본 | 0.913 | 0.537 | 0.338 | 0.970 | 0.415 |
| 최빈값대체,표준화 | 0.913 | 0.556 | 0.341 | 0.970 | 0.417 |
| 최빈값대체,정규화 | 0.913 | 0.538 | 0.335 | 0.971 | 0.413 |

□ 데이터 (남녀 통합, 남자/여자 데이터) 기준 해석

* 로지스틱 회귀분석 결과를 데이터 기준으로 해석한 결과는 아래와 같음

1. 남녀 통합 청소년 데이터  
   : CV = 5, CV = 10인 경우를 통틀어 Recall이 가장 높은 경우는 Recall이 0.493   
    CV = 5, CV = 10 통합 평균 Accuarcy : 0.879, 평균 Recall : 0.365, 평균 F1 score : 0.461
2. 청소년 남자 데이터  
   : CV = 5, CV = 10인 경우에서 각각 Recall이 가장 높은 경우를 기준으로 Recall이 0.563  
    CV = 5, CV = 10 통합 평균 Accuarcy : 0.844, 평균 Recall : 0.391, 평균 F1 score : 0.489
3. 청소년 여자 데이터  
   : CV = 5, CV = 10을 통틀어 Recall이 가장 높은 경우 0.406  
    CV = 5, CV = 10 통합 평균 Accuarcy : 0.916, 평균 Recall : 0.315, 평균 F1 score : 0.400

* 청소년 여자 데이터의 평균 Accuracy가 청소년 남자 데이터의 평균 Accuracy보다 약 0.072 높지만 본 연구에서의 주 관심 측도인 Recall은 여자가 남자보다 평균 0.076 낮음
* Accuracy는 평균적으로 여자가 남자보다 높음. 그러나 Recall은 평균적으로 여자가 남자보다 낮음. 여자 청소년의 경우 비만율이 남자에 비해 절반 이하이기 때문에 비만인 경우에 대한 학습에 사용할 데이터 부족으로 인해 Recall이 작은 것으로 판단. 이 과정에서 모델의 성능을 단순히 Accuracy만을 이용해 판단해서는 안 됨을 확인함

□ 결측치 처리 (결측치 있는 행 제거, 평균값/중앙값/최빈값 대체) 기준 해석

: CV = 5일 때와 CV = 10일 때를 통틀어 남녀 통합/남자/여자 청소년 데이터 각각에서 Recall이 가장 높은 경우가 모두 결측치가 있는 행을 제거했을 때에 해당함

* 결측치가 있는 행을 제거하는 방법 즉, 완전 제거법을 사용한 경우가 평균값 대체, 중앙값 대체, 최빈값 대체와 같은 단일 대체법을 이용한 경우보다 Test Data에 대한 모델의 퍼포먼스가 높음. 이는 본 연구에 사용된 데이터에서는 단일 대체법을 사용했을 경우 정보의 손실이 완전 제거법을 사용했을 경우 손실된 정보량보다 많은 것으로 사료됨.

□ 스케일링 (원본, 표준화, 정규화) 기준 해석

- 로지스틱 회귀분석의 경우 분석 전 입력 변수 값들의 표준화, 정규화 등의 스케일링이 필요한지 여부에 대해 다양한 의견이 존재

- 본 연구에서는 연속형 변수인 “주중 하루 평균 수면 시간”과 “하루 평균 앉아서 보내는 시간(분)” 변수를 표준화 적용한 경우, 정규화를 적용한 경우, 원본 그대로를 사용한 경우에 대해 모두 모델을 수립하여 결과를 비교

- 남녀 통합 데이터의 경우 표준화 및 정규화를 적용시킨 경우가 원본을 사용한 경우보다 Accuracy는 최소0.01, 최대 0.13, Recall은 최소 0.03, 최대 0.28, F1 score는 최소 0.01, 최대 0.8 높음.

- 남자 데이터의 경우 원본을 사용한 경우가 표준화 및 정규화를 적용시킨 경우보다 Measure 값들이 높은 경우도 있고, 반대의 경우도 존재

- 여자 데이터의 경우도 남자 데이터의 경우와 동일

- 표준화를 적용한 경우와 정규화를 적용한 경우의 Measure 비교 결과 차이는 미미함

- 데이터에 따라 표준화 및 정규화를 이용한 모델의 성능이 뛰어날 수도, 원본을 이용한 모델의   
 성능의 뛰어날 수도 있음을 확인

🡺 로지스틱 회귀분석에서는 표준화 및 정규화와 같은 스케일링의 효과가 크지 않다고 판단

그러나 본 연구에서는 세 가지 데이터를 기준으로 스케일링의 효과를 확인하였기에 한계 존재

□ CV 기준 해석

: 모델 수립 시 5-fold Cross Validation과 10-fold Cross Validation을 수행한 결과는 아래와 같음

참고 : 아래 표에는 CV = 5와 CV = 10 각각의 모델 퍼포먼스의 평균치를 기재함

**[ 남녀 통합 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CV** | **Accuracy** | **Precision** | **Recall** | **Specificity** | **F1 Score** |
| **5** | 0.877 | 0.622 | 0.357 | 0.964 | 0.452 |
| **10** | 0.880 | 0.640 | 0.373 | 0.968 | 0.470 |

**[ 남자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CV** | **Accuracy** | **Precision** | **Recall** | **Specificity** | **F1 Score** |
| **5** | 0.841 | 0.668 | 0.379 | 0.954 | 0.477 |
| **10** | 0.846 | 0.681 | 0.403 | 0.953 | 0.501 |

**[ 여자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CV** | **Accuracy** | **Precision** | **Recall** | **Specificity** | **F1 Score** |
| **5** | 0.915 | 0.608 | 0.277 | 0.980 | 0.362 |
| **10** | 0.917 | 0.575 | 0.353 | 0.971 | 0.437 |

- 남녀 통합 데이터, 남자/여자 데이터 모두에서 Accuracy, Precison, Recall, F1 Score가 CV = 10일 때가 CV = 5일 때보다 높음

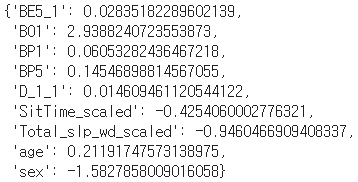
- K-fold Cross Validation에서 K 값을 증가시킴으로 인해 테스트 데이터에 대한 모델의 성능이 뛰어남을 확인

**□ Logistic Regression활용 변수 중요도 확인**

**참고 : sklearn 패키지는 회귀분석 시 설명변수의 p-value를 추출하는 방법이 아직 나오지 않아, 대안으로 statsmodels 패키지의 OLS(선형회귀분석) 함수를 사용하여 p-value 확인**

1. 청소년 전체 데이터 ( cv = 10 ) : 결측행 제거, 정규화

* 최적 파라미터 : C = 100, penalty = ‘l2’
* 회귀 계수

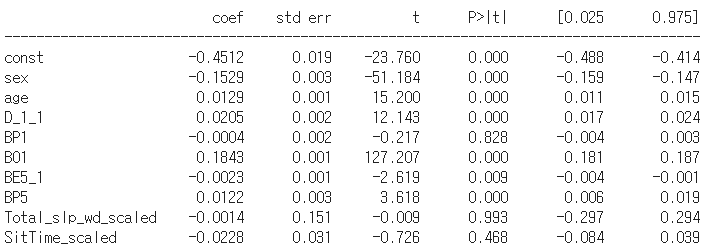


회귀 계수 해석 예시

: 설명변수 ‘age’을 제외한 나머지 변수들이 주어질 때 ‘age’ 값이 한 단위 증가할 때 BMI = 1(비만)일 오즈가 exp(0.212) = 1.236배, 즉 나이가 한 살 증가할 때마다 BMI = 1 (비만)의 오즈가 21.2 % 증가

설명변수 ‘sex’를 제외한 나머지 변수들이 주어질 때 오즈에 대한 ‘sex’의 효과는 exp(-1.583) = 0.205배, 즉 여자인 경우 남자에 비해 BMI = 1(비만)의 오즈가 약 80% 감소

* 설명 변수 p-value 확인

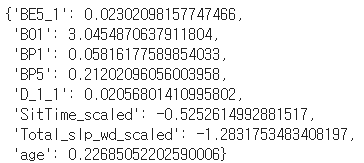


* BP1 (평상시 스트레스 인지 정도), Total\_slp\_wd\_scaled(주중 평균 수면 시간)와 SitTime\_scaled(하루 평균 앉아서 보내는 시간(분))의 p-value가 매우 높아서 본 연구에서 사용한 남녀 통합 청소년 비만 여부를 예측하는데 유의미하지 않은 요인임을 확인
* 스트레스, 수면 시간, 앉아서 보내는 시간이 비만에 영향을 끼치는 요인 중 하나라는 기존 논문과는 다른 결과 (Sunhee Park, 2011)
* P-value가 매우 높은 위 세 가지 변수를 제외하고 회귀 계수 기준 남녀 통합 청소년의 비만 여부에 영향을 끼치는 상위 3개의 요인은 BO1(주관적 체형 인지), sex(성별), age(나이)

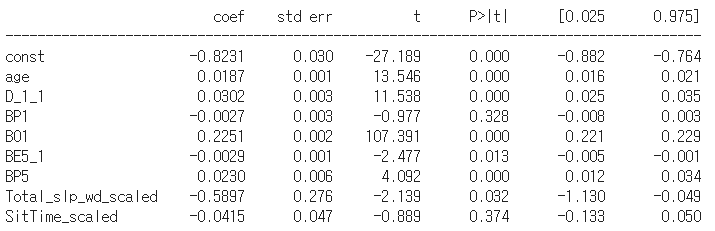
1. 남자 데이터 ( cv = 10 ) : 결측행 제거, 정규화

- 최적 파라미터 : C = 100, penalty = ‘l2’

- 회귀 계수



* 설명 변수 p-value 확인

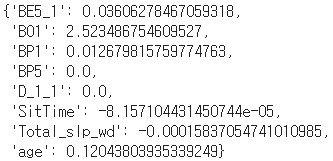


* BP1 (평상시 스트레스 인지 정도), SitTime\_scaled(하루 평균 앉아서 보내는 시간(분))의 p-value가 각각 0.328, 0.374로 높아서 본 연구에서 사용한 청소년 남자 비만 여부를 예측하는데 유의미하지 않은 요인임을 확인
* 스트레스, 앉아서 보내는 시간이 비만에 영향을 끼치는 요인 중 하나라는 기존 논문과는 다른 결과 (Sunhee Park, 2011)
* P-value가 높은 위 두 가지 변수를 제외하고 회귀 계수 기준 청소년 남자의 비만 여부에 영향을 끼치는 상위 3개의 요인은 BO1 (주관적 체형 인지), Total\_slp\_wd\_scaled(주중 평균 수면시간), age(나이)

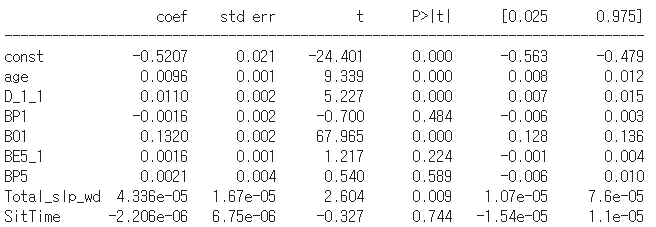
3) 여자 데이터 ( cv = 10 ) : 결측행 제거, 원본/표준화/정규화 (결과 동일함)

3-1) 원본

* 최적 파라미터: C = 0.1, penalty = ‘l1’
* 회귀 계수



* 설명 변수 p-value 확인



참고: 여자 데이터의 경우 결측행 제거, 원본/표준화/정규화 세 가지 경우에 대해 Recall이 동일하게 나와서 세 모델 각각의 회귀 계수, p-value 확인함.  
 세 모델 각각의 회귀 계수, p-value 확인 결과가 동일하여 여자 데이터 결측행 제거, 원본을 이용하여 해석함

* BP1(평상시 스트레스 인지정도), BE5\_1(1주일간 근력운동 일수), BP5(2주 이상 연속 우울감 여부), SitTime(하루 평균 앉아서 보내는 시간(분))의 p-value가 각각 0.484, 0.224, 0.589, 0.774로 높아서 본 연구에서 사용한 청소년 여자 비만 여부를 예측하는데 유의미하지 않은 요인임을 확인
* 스트레스, 근육운동 일수, 앉아서 보내는 시간, 우울경험이 비만에 영향을 끼치는 요인 중 하나라는 기존 논문과는 다른 결과 (Sunhee Park, 2011; 지영주, 2011)
* P-value가 높은 위 네 가지 변수를 제외하고 회귀 계수 기준 청소년 여자의 비만 여부에 영향을 끼치는 상위 2개의 요인은 BO1(주관적 체형 인지), age(나이)  
  ( Total\_slp\_wd (주중 평균 수면 시간)는 회귀 계수가 약 4.336e-05으로써 설명력이 매우 떨어져서 사실상 비만 여부를 예측하는데 영향이 없다고 판단하여 제외함 )
* 회귀 계수 결과 BP5(2주 이상 연속 우울감 여부)와 D\_1\_1(주관적 건강 상태) 두 변수의 회귀 계수는 0임. 즉, 청소년 여자 비만 여부에 2주 이상 연속 우울감 여부와 주관적 건강 상태는 어떠한 영향도 끼치지 않음을 확인. 또한 Total\_slp\_wd의 회귀 계수는 약 -0.00016로써 사실상 청소년 여자 비만 여부 예측에 영향이 거의 없음

1. 변수 중요도 확인 결과 결론

* 남녀 통합, 남자/여자 데이터 모두 “주관적 체형 인지”가 1순위로써 비만 여부를 예측하는데 가장 중요함
* 남녀 통합 데이터에서 “성별”이 2순위로써 성별이 비만 여부를 예측하는데 중요함
* “나이”는 세 개의 데이터에서 모두 3순위 안에 포함됨을 통해 나이는 성별과 상관없이 청소년의 비만 여부를 예측하는데 있어서 중요함을 확인
* 여자의 경우 “2주 이상 연속 우울감 여부”와 “주관적 건강 상태”의 회귀 계수가 0으로써 비만 여부에 대한 설명력이 없음. 반면, 남자의 경우 “2주 이상 연속 우울감 여부”의 회귀 계수는 0.212로써 회귀 계수 기준 변수 중요도가 4위에 해당
* “2주 이상 연속 우울감 여부”는 남자 청소년의 비만 여부를 예측하는데 꽤나 중요하지만 여자 청소년의 비만 여부를 예측하는데 있어서 의미가 없음

**■ KNN Performance Results  
\*참고 : 각 Performance Measure 기재 시 소수점 넷 째 자리에서 반올림함**

**□ CV = 5 적용**

**[ 청소년 전체 데이터 (남/녀 통합) ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **KNN** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.793 | 0.258 | 0.229 | 0.888 | 0.242 |
| 결측행제거,표준화 | 0.846 | 0.469 | 0.458 | 0.912 | 0.463 |
| 결측행제거,정규화 | 0.851 | 0.486 | 0.475 | 0.915 | 0.48 |
| 평균값대체,원본 | 0.791 | 0.249 | 0.232 | 0.884 | 0.24 |
| 평균값대체,표준화 | 0.841 | 0.44 | 0.421 | 0.911 | 0.43 |
| 평균값대체,정규화 | 0.841 | 0.441 | 0.429 | 0.909 | 0.435 |
| 중앙값대체,원본화 | 0.795 | 0.26 | 0.234 | 0.889 | 0.246 |
| 중앙값대체,표준화 | 0.839 | 0.435 | 0.421 | 0.909 | 0.428 |
| 중앙값대체,정규화 | 0.841 | 0.441 | 0.431 | 0.909 | 0.436 |
| 최빈값대체,원본화 | 0.794 | 0.258 | 0.235 | 0.887 | 0.246 |
| 최빈값대체,표준화 | 0.839 | 0.435 | 0.421 | 0.909 | 0.428 |
| 최빈값대체,정규화 | 0.841 | 0.44 | 0.432 | 0.909 | 0.436 |

**[ 청소년 남자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **KNN** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.728 | 0.305 | 0.284 | 0.838 | 0.294 |
| 결측행제거,표준화 | 0.814 | 0.533 | 0.529 | 0.885 | 0.531 |
| 결측행제거,정규화 | 0.837 | 0.596 | 0.568 | 0.904 | 0.582 |
| 평균값대체,원본 | 0.738 | 0.303 | 0.265 | 0.853 | 0.283 |
| 평균값대체,표준화 | 0.828 | 0.564 | 0.503 | 0.906 | 0.532 |
| 평균값대체,정규화 | 0.836 | 0.59 | 0.524 | 0.912 | 0.555 |
| 중앙값대체,원본화 | 0.736 | 0.295 | 0.259 | 0.851 | 0.276 |
| 중앙값대체,표준화 | 0.824 | 0.552 | 0.507 | 0.901 | 0.529 |
| 중앙값대체,정규화 | 0.837 | 0.591 | 0.526 | 0.912 | 0.556 |
| 최빈값대체,원본화 | 0.737 | 0.305 | 0.274 | 0.849 | 0.289 |
| 최빈값대체,표준화 | 0.826 | 0.559 | 0.5 | 0.905 | 0.528 |
| 최빈값대체,정규화 | 0.831 | 0.572 | 0.532 | 0.904 | 0.551 |

**[ 청소년 여자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **KNN** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.853 | 0.166 | 0.146 | 0.925 | 0.156 |
| 결측행제거,표준화 | 0.877 | 0.327 | 0.312 | 0.935 | 0.319 |
| 결측행제거,정규화 | 0.875 | 0.322 | 0.319 | 0.931 | 0.32 |
| 평균값대체,원본 | 0.854 | 0.168 | 0.151 | 0.925 | 0.159 |
| 평균값대체,표준화 | 0.873 | 0.285 | 0.256 | 0.935 | 0.27 |
| 평균값대체,정규화 | 0.875 | 0.302 | 0.286 | 0.934 | 0.293 |
| 중앙값대체,원본화 | 0.856 | 0.165 | 0.142 | 0.928 | 0.153 |
| 중앙값대체,표준화 | 0.874 | 0.289 | 0.258 | 0.936 | 0.272 |
| 중앙값대체,정규화 | 0.875 | 0.302 | 0.284 | 0.934 | 0.293 |
| 최빈값대체,원본화 | 0.855 | 0.159 | 0.138 | 0.927 | 0.148 |
| 최빈값대체,표준화 | 0.873 | 0.286 | 0.265 | 0.934 | 0.275 |
| 최빈값대체,정규화 | 0.874 | 0.299 | 0.283 | 0.934 | 0.291 |

□ **CV = 10 적용**

**[ 청소년 전체 데이터 (남/녀 통합) ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **KNN** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.794 | 0.262 | 0.231 | 0.889 | 0.245 |
| 결측행제거,표준화 | 0.846 | 0.469 | 0.458 | 0.912 | 0.463 |
| 결측행제거,정규화 | 0.851 | 0.486 | 0.475 | 0.915 | 0.48 |
| 평균값대체,원본 | 0.792 | 0.25 | 0.229 | 0.885 | 0.239 |
| 평균값대체,표준화 | 0.841 | 0.44 | 0.418 | 0.911 | 0.429 |
| 평균값대체,정규화 | 0.841 | 0.441 | 0.429 | 0.909 | 0.435 |
| 중앙값대체,원본화 | 0.795 | 0.26 | 0.234 | 0.889 | 0.246 |
| 중앙값대체,표준화 | 0.841 | 0.44 | 0.421 | 0.911 | 0.43 |
| 중앙값대체,정규화 | 0.841 | 0.441 | 0.431 | 0.909 | 0.436 |
| 최빈값대체,원본화 | 0.794 | 0.257 | 0.235 | 0.887 | 0.246 |
| 최빈값대체,표준화 | 0.839 | 0.435 | 0.421 | 0.909 | 0.428 |
| 최빈값대체,정규화 | 0.84 | 0.44 | 0.431 | 0.909 | 0.436 |

**[ 청소년 남자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **KNN** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.726 | 0.303 | 0.284 | 0.837 | 0.293 |
| 결측행제거,표준화 | 0.835 | 0.598 | 0.532 | 0.911 | 0.563 |
| 결측행제거,정규화 | 0.845 | 0.622 | 0.572 | 0.913 | 0.596 |
| 평균값대체,원본 | 0.737 | 0.302 | 0.267 | 0.851 | 0.283 |
| 평균값대체,표준화 | 0.831 | 0.574 | 0.513 | 0.908 | 0.542 |
| 평균값대체,정규화 | 0.841 | 0.6 | 0.548 | 0.912 | 0.572 |
| 중앙값대체,원본화 | 0.736 | 0.295 | 0.257 | 0.852 | 0.275 |
| 중앙값대체,표준화 | 0.83 | 0.573 | 0.509 | 0.908 | 0.539 |
| 중앙값대체,정규화 | 0.837 | 0.593 | 0.524 | 0.913 | 0.566 |
| 최빈값대체,원본화 | 0.737 | 0.305 | 0.275 | 0.849 | 0.289 |
| 최빈값대체,표준화 | 0.826 | 0.559 | 0.499 | 0.905 | 0.527 |
| 최빈값대체,정규화 | 0.84 | 0.598 | 0.538 | 0.913 | 0.566 |

**[ 청소년 여자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **KNN** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.852 | 0.16 | 0.141 | 0.925 | 0.15 |
| 결측행제거,표준화 | 0.877 | 0.327 | 0.312 | 0.935 | 0.319 |
| 결측행제거,정규화 | 0.875 | 0.322 | 0.319 | 0.931 | 0.32 |
| 평균값대체,원본 | 0.853 | 0.161 | 0.145 | 0.924 | 0.153 |
| 평균값대체,표준화 | 0.873 | 0.285 | 0.256 | 0.935 | 0.27 |
| 평균값대체,정규화 | 0.875 | 0.302 | 0.286 | 0.934 | 0.293 |
| 중앙값대체,원본화 | 0.855 | 0.163 | 0.142 | 0.927 | 0.152 |
| 중앙값대체,표준화 | 0.874 | 0.289 | 0.258 | 0.936 | 0.272 |
| 중앙값대체,정규화 | 0.875 | 0.302 | 0.284 | 0.934 | 0.293 |
| 최빈값대체,원본화 | 0.855 | 0.16 | 0.139 | 0.927 | 0.149 |
| 최빈값대체,표준화 | 0.873 | 0.287 | 0.265 | 0.934 | 0.276 |
| 최빈값대체,정규화 | 0.874 | 0.299 | 0.283 | 0.934 | 0.291 |

□ 데이터 (남,녀 통합/남자/여자 데이터) 기준 해석

1. 남녀 통합 청소년 데이터  
   : CV = 5, CV = 10인 경우를 통틀어 Recall이 가장 높은 경우는 Recall이 0.475  
    CV = 5, CV = 10 통합 평균 Accuarcy : 0.826, 평균 Recall : 0.368, 평균 F1 score : 0.376
2. 남자 청소년 데이터  
   : CV = 5, CV = 10인 경우에서 각각 Recall이 가장 높은 경우를 기준으로 Recall이 0.568  
    CV = 5, CV = 10 통합 평균 Accuarcy : 0.800, 평균 Recall : 0.441, 평균 F1 score : 0.463
3. 여자 청소년 데이터  
   : CV = 5, CV = 10을 통틀어 Recall이 가장 높은 경우 0.319  
    CV = 5, CV = 10 통합 평균 Accuarcy : 0.868, 평균 Recall : 0.236, 평균 F1 score : 0.245

: 여자 청소년의 평균 Accuracy는 반면 남/녀 통합과 남자 청소년 데이터보다 높은 반면, 본 연구에서 주 척도로 활용되는 Recall의 경우 여자 청소년이 가장 낮은 수치를 보여주고 있음

: Recall을 기준으로 보았을 때 ‘남자 청소년 > 남/녀 통합 > 여자 청소년’순으로 높은 성능을 나타냄

* 논문에서 추출한 인자들이 남자 청소년 > 남/녀 통합 > 여자 청소년 순으로 적합할 것이라고 판단

□ 결측치 처리 (결측치 있는 행 제거, 평균값/중앙값/최빈값 대체) 기준 해석

: 결측치 처리 방안에 대해서는 Recall의 경우 결측행을 제거한 것이 남/녀 통합, 남자 청소년, 여자 청소년 모두에게 가장 성능이 좋았음

: 그 외 나머지 결측치 처리 방안에 대해서는 평균값 대체, 중앙값 대체, 최빈값 대체에 대해서는 큰 차이를 볼 수 없을 만큼 비슷했음

* KNN에서는 단일대체법을 활용한 결측치 처리 방안보다 결측행을 제거하는 방식이 더 좋은 효과를 보임

□ 스케일링 (원본, 표준화, 정규화) 기준 해석

: 결과적으로 남,녀 통합, 남자 청소년, 여자 청소년에 대하여 ‘정규화 > 표준화 > 원본’ 의 결과를 보여주었으며 특히 원본의 경우 Recall 뿐만 아니라 다른 척도들에 대해서도 뒤떨어지는 결과를 보여줌

* KNN알고리즘 특성상 거리를 활용하여 유사도를 추출하기에 변수들의 Value에 대하여 표준화 및 정규화 등의 전처리가 필요하다는 사실을 확인함

□ CV 기준 해석

: 5-fold Cross Validation과 10-fold Cross Validation을 수행한 결과는 아래와 같음

참고: 아래 표에는 CV = 5와 CV = 10 각각의 모델 퍼포먼스의 평균치를 기재함

**[ 남녀 통합 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CV** | **Accuracy** | **Precision** | **Recall** | **Specificity** | **F1 Score** |
| **5** | 0.826 | 0.384 | 0.368 | 0.903 | 0.376 |
| **10** | 0.826 | 0.385 | 0.368 | 0.903 | 0.376 |

**[ 남자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CV** | **Accuracy** | **Precision** | **Recall** | **Specificity** | **F1 Score** |
| **5** | 0.798 | 0.480 | 0.439 | 0.885 | 0.459 |
| **10** | 0.802 | 0.494 | 0.443 | 0.889 | 0.468 |

**[ 여자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CV** | **Accuracy** | **Precision** | **Recall** | **Specificity** | **F1 Score** |
| **5** | 0.868 | 0.256 | 0.237 | 0.932 | 0.246 |
| **10** | 0.868 | 0.255 | 0.236 | 0.931 | 0.245 |

: 대체적으로는 CV 값을 높게 설정할수록 좋은 성능을 배출한 것을 볼 수 있었으나 성능이 유지되거나 오히려 동결된 Case도 있었음

* CV를 높일수록 이론적으로 성능이 향상되는 것은 합리적인 판단이나 오히려 모델을 만들 때 시간적 비용이 더 많이 들어가므로 무조건적으로 높게 잡기보다는 적절한 수준의 CV의 값을 설정하는 것이 이상적이라는 것을 확인

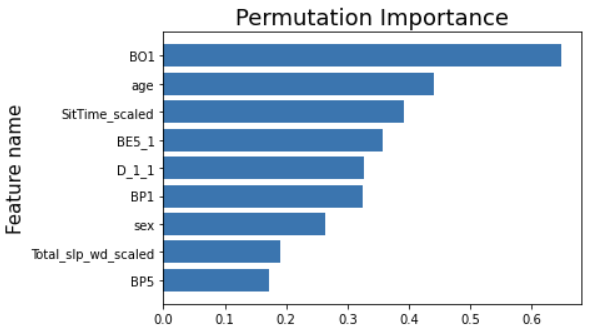
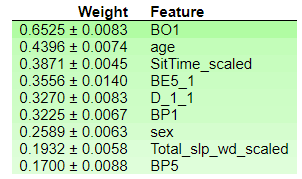
**□ KNN활용 변수 중요도 확인**

**참고: sklearn 패키지의 permutation\_importance를 활용하여 변수중요도 추출**

1. 청소년 전체 데이터 (cv = 5, 10) : 결측행 제거, 정규화 (결과 동일함)

1-1) cv = 5

* 최적 파라미터 : leaf\_size = 30, n\_neighbors = 1, weights = uniform
* 변수 중요도



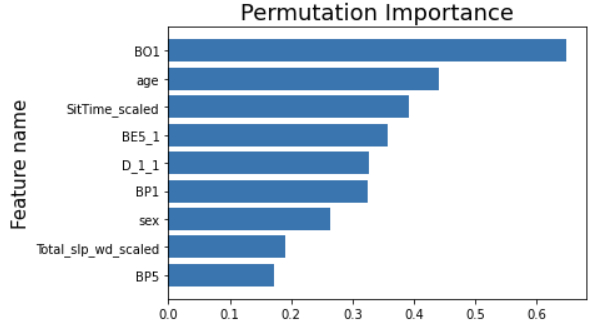
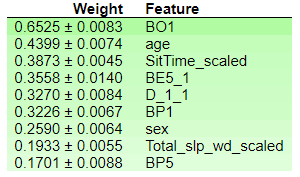
* Permutation Importance 해석 예시

: 위 표에 나타나는 Weight는 value를 shuffling 후 shuffling 전과 비교했을 때 Score가 떨어진 정도를 보여줌

설명변수 ‘BO1’에 대해 0.6525만큼 Score(neg\_mean\_squared\_error)가 가장 많이 떨어졌다면 이는 해당 feature가 예측에 가장 영향력이 있으며 모델이 해당 변수에 가장 많이 의존한다고 해석가능

1-2) cv = 10

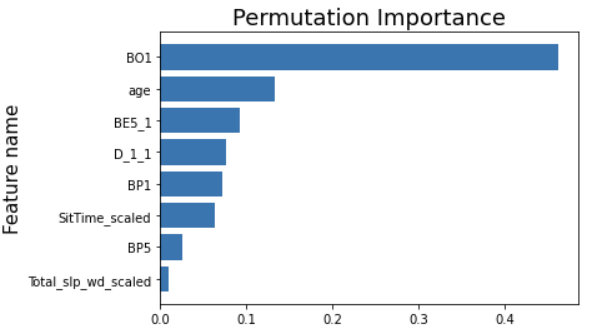
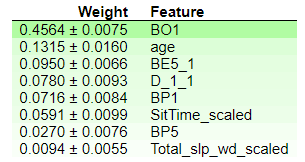
* 최적 파라미터 : leaf\_size = 5, n\_neighbors = 1, weights = uniform
* 변수 중요도



1. 남자 데이터 ( cv = 10) : 결측행 제거, 정규화

2-1) 원본

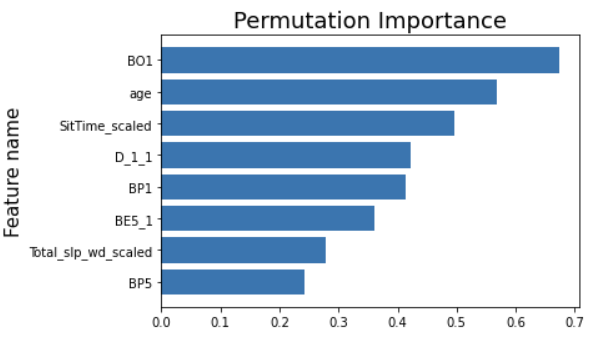
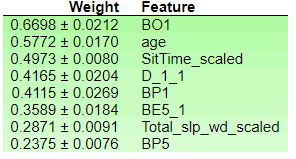
* 최적 파라미터 : leaf\_size = 5, n\_neighbors = 5, weights = uniform
* 변수 중요도



1. 여자 데이터 ( cv = 5, 10) : 결측행 제거, 정규화 (결과 동일함)

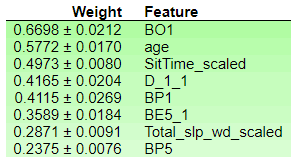
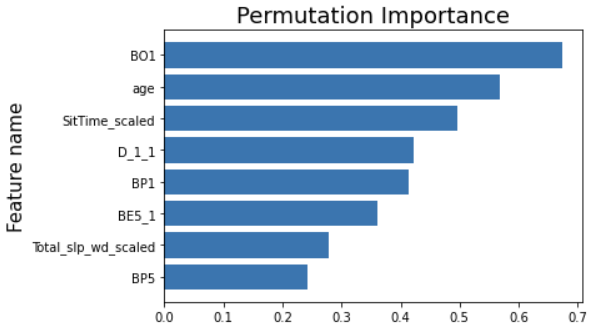
3-1) cv = 5

* 최적 파라미터 : leaf\_size = 5, n\_neighbors = 1, weights = uniform
* 변수 중요도



3-2) cv = 10

* 최적 파라미터 : leaf\_size = 5, n\_neighbors = 1, weights = uniform
* 변수 중요도

* 변수 중요도 확인 결과 해석

: 모델의 Confusion matrix가 완전히 동일하여 5개의 척도가 비슷한 것들을 열거하고 변수 중요도를 확인한 결과 완전히 동일하진 않다는 것을 확인

* 모델의 예측 결과는 동일하지만 데이터가 다르면 중요도의 정량화된 값은 바뀔 수 있음을 확인

: [데이터 (남/녀 통합/남자/여자 데이터) 기준 해석] 에서는 ‘남자 청소년 > 남/녀 통합 > 여자 청소년’ 순으로 예측 성능이 좋았는데 모델 별 변수 중요도에서는 오히려 역순으로 ‘여자 청소년 > 남/녀 통합 > 남자 청소년’의 결과를 보여줌

* 모델의 성능에 따라 변수중요도가 무조건적으로 비례하여 증가하는 것은 아님을 확인

: 유의미한 인자로 상위 3개 [BO1], [age]가 모두 동일하게 선정되었으며 그 이하 항목들에 대해서는 데이터 별로 다른 성향을 보임

* 각 남/녀 통합, 남자 청소년, 여자 청소년에 대하여 유의미한 비만 인자는 [BO1], [age]로 결론지을 수 있고 그 이하 항목들은 그룹의 특성에 따라 다름을 확인

**■ SGD (Stochastic Graident Descent) Performance Results  
\*참고 : 각 Performance Measure 기재 시 소수점 넷째 자리에서 반올림함**

**□ CV = 5 적용**

**[ 청소년 전체 데이터 (남/녀 통합) ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **모델명** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.864 | 0.564 | 0.440 | 0.940 | 0.490 |
| 결측행제거,표준화 | 0.880 | 0.578 | 0.640 | 0.921 | 0.608 |
| 결측행제거,정규화 | 0.89 | 0.685 | 0.472 | 0.963 | 0.559 |
| 평균값대체,원본 | 0.867 | 0.541 | 0.447 | 0.937 | 0.490 |
| 평균값대체,표준화 | 0.879 | 0.618 | 0.404 | 0.958 | 0.488 |
| 평균값대체,정규화 | 0.877 | 0.661 | 0.280 | 0.976 | 0.394 |
| 중앙값대체,원본 | 0.864 | 0.526 | 0.485 | 0.927 | 0.505 |
| 중앙값대체,표준화 | 0.879 | 0.600 | 0.467 | 0.948 | 0.525 |
| 중앙값대체,정규화 | 0.88 | 0.621 | 0.403 | 0.959 | 0.488 |
| 최빈값대체,원본 | 0.847 | 0.464 | 0.442 | 0.915 | 0.453 |
| 최빈값대체,표준화 | 0.877 | 0.634 | 0.341 | 0.967 | 0.444 |
| 최빈값대체,정규화 | 0.877 | 0.584 | 0.472 | 0.944 | 0.522 |

**[ 청소년 남자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **모델명** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.822 | 0.544 | 0.678 | 0.859 | 0.604 |
| 결측행제거,표준화 | 0.858 | 0.668 | 0.578 | 0.928 | 0.620 |
| 결측행제거,정규화 | 0.854 | 0.634 | 0.638 | 0.908 | 0.636 |
| 평균값대체,원본 | 0.822 | 0.550 | 0.439 | 0.910 | 0.491 |
| 평균값대체,표준화 | 0.844 | 0.653 | 0.424 | 0.945 | 0.515 |
| 평균값대체,정규화 | 0.849 | 0.640 | 0.511 | 0.931 | 0.568 |
| 중앙값대체,원본 | 0.830 | 0.598 | 0.381 | 0.938 | 0.466 |
| 중앙값대체,표준화 | 0.849 | 0.623 | 0.571 | 0.917 | 0.596 |
| 중앙값대체,정규화 | 0.849 | 0.652 | 0.487 | 0.937 | 0.557 |
| 최빈값대체,원본 | 0.829 | 0.615 | 0.331 | 0.950 | 0.431 |
| 최빈값대체,표준화 | 0.851 | 0.617 | 0.622 | 0.907 | 0.620 |
| 최빈값대체,정규화 | 0.853 | 0.624 | 0.616 | 0.910 | 0.620 |

**[ 청소년 여자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **모델명** | **accuracy** | **precision** | **Recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.910 | 0.527 | 0.307 | 0.972 | 0.388 |
| 결측행제거,표준화 | 0.928 | 0.690 | 0.395 | 0.981 | 0.503 |
| 결측행제거,정규화 | 0.922 | 0.613 | 0.416 | 0.973 | 0.496 |
| 평균값대체,원본 | 0.909 | Nan | 0.000 | 1.000 | Nan |
| 평균값대체,표준화 | 0.889 | 0.405 | 0.455 | 0.932 | 0.428 |
| 평균값대체,정규화 | 0.911 | 0.514 | 0.372 | 0.965 | 0.432 |
| 중앙값대체,원본 | 0.908 | Nan | 0.000 | 1.000 | nan |
| 중앙값대체,표준화 | 0.912 | 0.544 | 0.199 | 0.983 | 0.292 |
| 중앙값대체,정규화 | 0.913 | 0.533 | 0.343 | 0.970 | 0.417 |
| 최빈값대체,원본 | 0.908 | Nan | 0.000 | 1.000 | Nan |
| 최빈값대체,표준화 | 0.912 | 0.525 | 0.366 | 0.966 | 0.431 |
| 최빈값대체,정규화 | 0.913 | 0.532 | 0.356 | 0.969 | 0.426 |

□ **CV = 10 적용**

**[ 청소년 전체 데이터 (남/녀 통합) ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **모델명** | **accuracy** | **precision** | **Recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.753 | 0.339 | 0.740 | 0.756 | 0.465 |
| 결측행제거,표준화 | 0.889 | 0.643 | 0.524 | 0.951 | 0.577 |
| 결측행제거,정규화 | 0.892 | 0.658 | 0.527 | 0.954 | 0.585 |
| 평균값대체,원본 | 0.611 | 0.267 | 0.988 | 0.549 | 0.420 |
| 평균값대체,표준화 | 0.879 | 0.597 | 0.459 | 0.948 | 0.519 |
| 평균값대체,정규화 | 0.879 | 0.609 | 0.430 | 0.954 | 0.504 |
| 중앙값대체,원본 | 0.867 | 0.571 | 0.273 | 0.966 | 0.369 |
| 중앙값대체,표준화 | 0.878 | 0.630 | 0.355 | 0.965 | 0.454 |
| 중앙값대체,정규화 | 0.879 | 0.623 | 0.388 | 0.960 | 0.478 |
| 최빈값대체,원본 | 0.729 | 0.338 | 0.930 | 0.696 | 0.495 |
| 최빈값대체,표준화 | 0.876 | ..6118 | 0.409 | 0.958 | 0.490 |
| 최빈값대체,정규화 | 0.878 | 0.605 | 0.415 | 0.954 | 0.492 |

**[ 청소년 남자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **모델명** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.829 | 0.681 | 0.272 | 0.968 | 0.388 |
| 결측행제거,표준화 | 0.855 | 0.634 | 0.654 | 0.906 | 0.644 |
| 결측행제거,정규화 | 0.855 | 0.631 | 0.664 | 0.903 | 0.647 |
| 평균값대체,원본 | 0.817 | 0.525 | 0.608 | 0.867 | 0.564 |
| 평균값대체,표준화 | 0.848 | 0.615 | 0.589 | 0.911 | 0.602 |
| 평균값대체,정규화 | 0.851 | 0.630 | 0.574 | 0.919 | 0.601 |
| 중앙값대체,원본 | 0.821 | 0.538 | 0.557 | 0.885 | 0.547 |
| 중앙값대체,표준화 | 0.851 | 0.638 | 0.546 | 0.925 | 0.588 |
| 중앙값대체,정규화 | 0.849 | 0.636 | 0.525 | 0.927 | 0.575 |
| 최빈값대체,원본 | 0.801 | 0.493 | 0.708 | 0.824 | 0.581 |
| 최빈값대체,표준화 | 0.852 | 0.630 | 0.588 | 0.916 | 0.608 |
| 최빈값대체,정규화 | 0.849 | 0.622 | 0.571 | 0.916 | 0.596 |

**[ 청소년 여자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **모델명** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.908 | Nan | 0.000 | 1.000 | Nan |
| 결측행제거,표준화 | 0.918 | 0.570 | 0.439 | 0.966 | 0.496 |
| 결측행제거,정규화 | 0.928 | 0.688 | 0.402 | 0.981 | 0.508 |
| 평균값대체,원본 | 0.909 | Nan | 0.000 | 1.000 | Nan |
| 평균값대체,표준화 | 0.911 | 0.515 | 0.353 | 0.967 | 0.419 |
| 평균값대체,정규화 | 0.913 | 0.539 | 0.335 | 0.971 | 0.413 |
| 중앙값대체,원본 | 0.909 | 0.000 | 0.000 | 0,.99 | Nan |
| 중앙값대체,표준화 | 0.915 | 0.562 | 0.293 | 0.977 | 0.385 |
| 중앙값대체,정규화 | 0.912 | 0.522 | 0.363 | 0.967 | 0.428 |
| 최빈값대체,원본 | 0.908 | Nan | 0.000 | 1.000 | Nan |
| 최빈값대체,표준화 | 0.912 | 0.547 | 0.206 | 0.983 | 0.299 |
| 최빈값대체,정규화 | 0.912 | 0.531 | 0.315 | 0.970 | 0.395 |

□ 데이터 (남,녀 통합/남자/여자 데이터) 기준 해석   
1) 남녀 통합 청소년 데이터  
: CV = 5, CV = 10인 경우를 통틀어 Recall이 가장 높은 경우는 Recall이 0.988   
 CV = 5, CV = 10 통합 평균 Accuarcy : 0.853, 평균 Recall : 0.488, 평균 F1 score : 0.492

2) 남자 청소년 데이터  
: CV = 5, CV = 10인 경우에서 각각 Recall이 가장 높은 경우를 기준으로 Recall이 0.708  
 CV = 5, CV = 10 통합 평균 Accuarcy : 0.841, 평균 Recall : 0.547, 평균 F1 score : 0.569

3) 여자 청소년 데이터  
: CV = 5, CV = 10을 통틀어 Recall이 가장 높은 경우 0.455  
 CV = 5, CV = 10 통합 평균 Accuarcy : 0.911, 평균 Recall : 0.246, 평균 F1 score : 0.420

🡪 Accuracy는 평균적으로 여자가 남자보다 높음. 그러나 Recall은 평균적으로 여자가 남자보다 낮음. 여자 청소년 데이터의 경우, 스케일링에 따른 영향을 많이 받아 원본 데이터에 대해 recall값을 0으로 반환하여 평균 Recall 값이 남자 청소년 데이터의 평균 Recall보다 현저히 낮은 것을 파악함   
🡪 또한, 남녀 통합 청소년 데이터에서 가장 높은 Recall값이 도출됨.   
남자, 여자 청소년 데이터에 비해 월등히 많은 양의 데이터가 높은 Recall 성능에 영향이 있는 것으로 사료됨

□ 결측치 처리 (결측치 있는 행 제거, 평균값/중앙값/최빈값 대체) 기준 해석  
: 결측치 처리 방안에 대해서는 Recall이 가장 높은 경우는 결측치가 있는 행을 제거한 것이 남녀 통합, 남자, 여자 모두를 비교해보았을 때 전반적으로 성능이 좋았으나, 가장 높은 Recall을 도출한 모델은 결측치를 평균값으로 대체한 경우이다.

□ 스케일링 (원본, 표준화, 정규화) 기준 해석  
: 스케일링(표준화, 정규화)을 한 것과 안 한 것의 Recall 측면에서 남/녀 통합, 남자 청소년, 여자 청소년에서 모두 달리 나타난 것으로 확인되었음  
1) 전체 – 표준화(cv=5), 원본(cv=10) 데이터가 Recall 측면에서 좋은 성능  
2) 남자 – 원본 데이터가 Recall 측면에서 좋은 성능  
3) 여자 – 표준화 데이터가 Recall 측면에서 좋은 성능

□ CV 기준 해석   
: 5-fold Cross Validation과 10-fold Cross Validation을 수행한 결과는 아래와 같음  
 \*참고 : 아래 표에는 CV = 5와 CV = 10 각각의 모델 퍼포먼스의 평균치를 기재함

**[ 남녀 통합 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CV** | **Accuracy** | **Precision** | **Recall** | **Specificity** | **F1 Score** |
| **5** | 0.871 | 0.619 | 0.284 | 0.969 | 0.502 |
| **10** | 0.872 | 0.621 | 0.273 | 0.972 | 0.491 |

**[ 남자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CV** | **Accuracy** | **Precision** | **Recall** | **Specificity** | **F1 Score** |
| **5** | 0.833 | 0.621 | 0.383 | 0.943 | 0.596 |
| **10** | 0.832 | 0.618 | 0.367 | 0.945 | 0.582 |

**[ 여자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CV** | **Accuracy** | **Precision** | **Recall** | **Specificity** | **F1 Score** |
| **5** | 0.915 | 0.609 | 0.179 | 0.989 | 0.327 |
| **10** | 0.915 | 0.609 | 0.179 | 0.989 | 0.368 |

: CV가 5보다 10인 경우가 훨씬 향상된 performance를 도출함

SGDClassifier모델의 경우, loss함수 파라미터 (“"hinge", "log", "squared\_hinge", "modified\_huber")를 조정해가며 GridSearch 방식으로 최적의 파라미터를 찾아감

🡺 따라서, 모든 데이터마다 최적의 loss함수가 변화하며 이에 따라 자연스레 최적의 결측치 처리 및 최적의 스케일링 방식이 각기 달리 적용됨을 확인할 수 있음

**■ SVM (Support Vector Machine) Performance Results  
\*참고 : 각 Performance Measure 기재 시 소수점 넷 째 자리에서 반올림함**

**□ CV = 5 적용**

**[ 청소년 전체 데이터 (남/녀 통합) ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **모델명** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.857 | Nan | 0.000 | 1.000 | Nan |
| 결측행제거,표준화 | 0.891 | 0.673 | 0.484 | 0.960 | 0.563 |
| 결측행제거,정규화 | 0.892 | 0.665 | 0.515 | 0.956 | 0.581 |
| 평균값대체,원본 | 0.857 | Nan | 0.000 | 1.000 | Nan |
| 평균값대체,표준화 | 0.875 | 0.606 | 0.364 | 0.961 | 0.454 |
| 평균값대체,정규화 | 0.877 | 0.602 | 0.411 | 0.955 | 0.488 |
| 중앙값대체,원본 | 0.857 | Nan | 0.000 | 1.000 | Nan |
| 중앙값대체,표준화 | 0.875 | 0.606 | 0.363 | 0.961 | 0.454 |
| 중앙값대체,정규화 | 0.877 | 0.602 | 0.411 | 0.955 | 0.488 |
| 최빈값대체,원본 | 0.857 | Nan | 0.000 | 1.000 | Nan |
| 최빈값대체,표준화 | 0.856 | 0.601 | 0.389 | 0.934 | 0.462 |
| 최빈값대체,정규화 | 0.879 | 0.598 | 0.469 | 0.948 | 0.526 |

**[ 청소년 남자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **모델명** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.800 | Nan | 0.000 | 1.000 | Nan |
| 결측행제거,표준화 | 0.855 | 0.638 | 0.633 | 0.911 | 0.635 |
| 결측행제거,정규화 | 0.854 | 0.628 | 0.666 | 0.901 | 0.647 |
| 평균값대체,원본 | 0.800 | Nan | 0.000 | 1.000 | Nan |
| 평균값대체,표준화 | 0.844 | 0.616 | 0.533 | 0.920 | 0.571 |
| 평균값대체,정규화 | 0.848 | 0.620 | 0.567 | 0.916 | 0.592 |
| 중앙값대체,원본 | 0.805 | Nan | 0.000 | 1.000 | Nan |
| 중앙값대체,표준화 | 0.844 | 0.616 | 0.533 | 0.920 | 0.571 |
| 중앙값대체,정규화 | 0.848 | 0.620 | 0.567 | 0.916 | 0.592 |
| 최빈값대체,원본 | 0.805 | Nan | 0.000 | 1.000 | Nan |
| 최빈값대체,표준화 | 0.844 | 0.617 | 0.534 | 0.920 | 0.572 |
| 최빈값대체,정규화 | 0.846 | 0.616 | 0.568 | 0.914 | 0.591 |

**[ 청소년 여자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **모델명** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.908 | Nan | 0 | 1 | nan |
| 결측행제거,표준화 | 0.928 | 0.683 | 0.406 | 0.981 | 0.509 |
| 결측행제거,정규화 | 0.928 | 0.683 | 0.406 | 0.981 | 0.509 |
| 평균값대체,원본 | 0.909 | Nan | 0 | 1 | Nan |
| 평균값대체,표준화 | 0.915 | 0.605 | 0.202 | 0.987 | 0.303 |
| 평균값대체,정규화 | 0.914 | 0.57 | 0.245 | 0.981 | 0.342 |
| 중앙값대체,원본 | 0.909 | Nan | 0 | 1 | Nan |
| 중앙값대체,표준화 | 0.915 | 0.605 | 0.202 | 0.987 | 0.303 |
| 중앙값대체,정규화 | 0.914 | 0.57 | 0.245 | 0.981 | 0.342 |
| 최빈값대체,원본 | 0.909 | Nan | 0 | 1 | 0 |
| 최빈값대체,표준화 | 0.914 | 0.588 | 0.201 | 0.986 | 0.299 |
| 최빈값대체,정규화 | 0.914 | 0.567 | 0.242 | 0.981 | 0.339 |

□ **CV = 10 적용**

**[ 청소년 전체 데이터 (남/녀 통합) ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **모델명** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.855 | Nan | 0.0 | 1.0 | Nan |
| 결측행제거,표준화 | 0.891 | 0.673 | 0.484 | 0.960 | 0.563 |
| 결측행제거,정규화 | 0.892 | 0.665 | 0.515 | 0.956 | 0.581 |
| 평균값대체,원본 | 0.857 | Nan | 0.0 | 1.0 | Nan |
| 평균값대체,표준화 | 0.875 | 0.606 | 0.363 | 0.961 | 0.454 |
| 평균값대체,정규화 | 0.877 | 0.602 | 0.411 | 0.955 | 0.489 |
| 중앙값대체,원본 | 0.857 | Nan | 0.0 | 1.0 | Nan |
| 중앙값대체,표준화 | 0.875 | 0.606 | 0.363 | 0.961 | 0.454 |
| 중앙값대체,정규화 | 0.875 | 0.606 | 0.363 | 0.961 | 0.454 |
| 최빈값대체,원본 | 0.857 | Nan | 0.0 | 1.0 | Nan |
| 최빈값대체,표준화 | 0.875 | 0.606 | 0.394 | 0.951 | 0.469 |
| 최빈값대체,정규화 | 0.875 | 0.604 | 0.383 | 0.954 | 0.464 |

**[ 청소년 남자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **모델명** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.805 | Nan | 0.000 | 1.000 | Nan |
| 결측행제거,표준화 | 0.844 | 0.616 | 0.533 | 0.919 | 0.571 |
| 결측행제거,정규화 | 0.848 | 0.620 | 0.567 | 0.916 | 0.592 |
| 평균값대체,원본 | 0.805 | Nan | 0.000 | 1.000 | Nan |
| 평균값대체,표준화 | 0.845 | 0,.6171 | 0.533 | 0.920 | 0.572 |
| 평균값대체,정규화 | 0.848 | 0.620 | 0.567 | 0.916 | 0.592 |
| 중앙값대체,원본 | 0.805 | Nan | 0.000 | 1.000 | Nan |
| 중앙값대체,표준화 | 0.844 | 0.616 | 0.533 | 0.920 | 0.571 |
| 중앙값대체,정규화 | 0.848 | 0.620 | 0.567 | 0.916 | 0.592 |
| 최빈값대체,원본 | 0.805 | Nan | 0.000 | 1.000 | Nan |
| 최빈값대체,표준화 | 0.845 | 0.617 | 0.533 | 0.920 | 0.572 |
| 최빈값대체,정규화 | 0.847 | 0.616 | 0.568 | 0.914 | 0.591 |

**[ 청소년 여자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **모델명** | **accuracy** | **precision** | **recall** | **specificity** | **F1-score** |
| 결측행제거,원본 | 0.908 | Nan | 0 | 1 | Nan |
| 결측행제거,표준화 | 0.928 | 0.683 | 0.406 | 0.981 | 0.509 |
| 결측행제거,정규화 | 0.928 | 0.683 | 0.406 | 0.981 | 0.509 |
| 평균값대체,원본 | 0.908 | Nan | 0 | 1 | Nan |
| 평균값대체,표준화 | 0.915 | 0.605 | 0.202 | 0.987 | 0.303 |
| 평균값대체,정규화 | 0.914 | 0.57 | 0.245 | 0.981 | 0.342 |
| 중앙값대체,원본 | 0.908 | Nan | 0 | 1 | Nan |
| 중앙값대체,표준화 | 0.915 | 0.605 | 0.202 | 0.987 | 0.303 |
| 중앙값대체,정규화 | 0.914 | 0.57 | 0.245 | 0.981 | 0.342 |
| 최빈값대체,원본 | 0.908 | Nan | 0 | 1 | Nan |
| 최빈값대체,표준화 | 0.914 | 0.588 | 0.201 | 0.986 | 0.299 |
| 최빈값대체,정규화 | 0.914 | 0.567 | 0.242 | 0.981 | 0.339 |

□ 데이터 (남,녀 통합/남자/여자 데이터) 기준 해석

1) 남녀 통합 청소년 데이터  
: CV = 5, CV = 10인 경우를 통틀어 Recall이 가장 높은 경우는 Recall이 0.515  
 CV = 5, CV = 10 통합 평균 Accuarcy : 0.871, 평균 Recall : 0.278, 평균 F1 score : 0.497

2) 남자 청소년 데이터  
: CV = 5, CV = 10인 경우에서 각각 Recall이 가장 높은 경우를 기준으로 Recall이 0.666  
 CV = 5, CV = 10 통합 평균 Accuarcy : 0.833, 평균 Recall : 0.375, 평균 F1 score : 0.589

3) 여자 청소년 데이터  
: CV = 5, CV = 10을 통틀어 Recall이 가장 높은 경우 0.406  
 CV = 5, CV = 10 통합 평균 Accuarcy : 0.915, 평균 Recall : 0.179, 평균 F1 score : 0.347

🡪 Accuracy는 평균적으로 여자가 남자보다 높음. 반면, Recall은 평균적으로 여자가 남자보다 낮음. 여자 청소년 데이터의 경우, 스케일링에 따른 영향을 많이 받아 원본 데이터에 대해 recall값을 0으로 반환하여 평균 Recall 값이 남자 청소년 데이터의 평균 Recall보다 현저히 낮은 것을 파악함  
🡪 통합/남자/여자 데이터 모두 결측행 제거/정규화를 한 경우가 가장 높은 performance를 보임

□ 결측치 처리 (결측치 있는 행 제거, 평균값/중앙값/최빈값 대체) 기준 해석  
: 결측치에 대해서 평균값, 중앙값, 최빈값 대체 방법보다 결측치가 존재하는 행을 모두 제거하는 것이 모델의 performance가 가장 높음을 확인

□ 스케일링 (원본, 표준화, 정규화) 기준 해석  
: SVM의 경우, 연속형 데이터의 스케일링에 따른 성능 차이가 매우 상이함

🡪 본 연구에서는 연속형 변수인 “주중 하루 평균 수면 시간”과 “하루 평균 앉아있는 시간” 변수를 표준화 또는 정규화를 통해 조정된 입력변수가 모델의 performance를 향상시키는데 매우 큰 영향이 있음을 확인

□ CV 기준 해석

: 5-fold Cross Validation과 10-fold Cross Validation을 수행한 결과는 아래와 같음  
 \*참고 : 아래 표에는 CV = 5와 CV = 10 각각의 모델 퍼포먼스의 평균치를 기재함

**[ 남녀 통합 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CV** | **Accuracy** | **Precision** | **Recall** | **Specificity** | **F1 Score** |
| **5** | 0.873 | 0.59 | 0.441 | 0.946 | 0.497 |
| **10** | 0.834 | 0.535 | 0.537 | 0.884 | 0.487 |

**[ 남자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CV** | **Accuracy** | **Precision** | **Recall** | **Specificity** | **F1 Score** |
| **5** | 0.843 | 0.618 | 0.523 | 0.92 | 0.56 |
| **10** | 0.84 | 0.606 | 0.571 | 0.906 | 0.578 |

**[ 여자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CV** | **Accuracy** | **Precision** | **Recall** | **Specificity** | **F1 Score** |
| **5** | 0.911 | 0.543 | 0.267 | 0.976 | 0.424 |
| **10** | 0.913 | 0.497 | 0.226 | 0.98 | 0.418 |

: cv가 5일 때와 10 일 때를 비교하였을 때, 별다른 차이가 보이지 않음

**[ Discussion ]**

□ 척도 설명

1) Accuracy - TP + TN / TP + TN + FP + FN)  
: 실제 데이터에서 예측 데이터가 얼마나 같은지를 판단하는 지표

앞서 본 연구의 target attribute인 BMI 여부의 경우, 비만인 사람이 전체 데이터에서 14%임을 확인하였다. 이 경우 ‘무조건 비만이 아니다’ 라는 결과를 도출하는 모델도 성능이 86%로 파악된다.

🡺 본 연구의 모델 performance를 accuracy로만 측정하는데 적합하지 않음

2) Precision: TP / (TP + FP) VS Recall: TP / (TP + FN)

Precision의 성능 향상 🡪 FP 감소 🡪 실제로 정상인데 비만이라고 예측하는 비율 감소  
Recall의 성능 향상 🡪 FN 감소 🡪 실제로 비만인데 정상이라고 예측하는 비율 감소

본 연구에서는 실제로 비만인 청소년이 비만으로 예측되는 확률이 무엇보다 중요함

🡺 본 연구의 모델 performance를 측정하기 위해서는 Recall이 Precision보다 적합

□ Predictive 측면 해석

**[ 남녀 통합 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Method** | **Accuracy** | **Precision** | **Recall** | **Specificity** | **F1 Score** |
| **Naive Bayes Classifier** | 0.874 | 0.582 | 0.409 | 0.951 | 0.480 |
| **Random Forest** | 0.891 | 0.66 | 0.517 | 0.955 | 0.58 |
| **Logistic Regression** | 0.891 | 0.670 | 0.493 | 0.959 | 0.568 |
| **K-Nearest Neighbor** | 0.851 | 0.486 | 0.475 | 0.915 | 0.48 |
| **SGDClassifier** | 0.611 | 0.267 | 0.988 | 0.549 | 0.420 |
| **SVM** | 0.892 | 0.665 | 0.515 | 0.956 | 0.581 |

**[ 남자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Method** | **Accuracy** | **Precision** | **Recall** | **Specificity** | **F1 Score** |
| **Naive Bayes Classifier** | 0.849 | 0.615 | 0.654 | 0.898 | 0.633 |
| **Random Forest** | 0.856 | 0.652 | 0.599 | 0.92 | 0.624 |
| **Logistic Regression** | 0.860 | 0.684 | 0.563 | 0.935 | 0.618 |
| **K-Nearest Neighbor** | 0.845 | 0.622 | 0.572 | 0.913 | 0.596 |
| **SGDClassifier** | 0.801 | 0.493 | 0.708 | 0.824 | 0.581 |
| **SVM** | 0.854 | 0.628 | 0.666 | 0.901 | 0.647 |

**[ 여자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Method** | **Accuracy** | **Precision** | **Recall** | **Specificity** | **F1 Score** |
| **Naive Bayes Classifier** | 0.849 | 0.614 | 0.654 | 0.898 | 0.633 |
| **Random Forest** | 0.926 | 0.701 | 0.355 | 0.985 | 0.472 |
| **Logistic Regression** | 0.891 | 0.670 | 0.493 | 0.959 | 0.568 |
| **K-Nearest Neighbor** | 0.875 | 0.322 | 0.319 | 0.931 | 0.32 |
| **SGDClassifier** | 0.889 | 0.405 | 0.455 | 0.932 | 0.428 |
| **SVM** | 0.928 | 0.683 | 0.406 | 0.981 | 0.509 |

- 전체적으로 SGDClassifer모델이 남녀 통합 데이터/ 남자 데이터에서 Recall 측면에서   
다른 모델링 기법보다 모델링 우월함을 확인할 수 있음

- 반면, 여자 데이터에서 Naïve Bayes Classifier 모델이 가장 높은 Recall을 도출함

□ Descriptive측면 해석

**[ 남/녀 통합 청소년 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **데이터** | **1위** | **2위** | **3위** | **4위** | **5위** |
| **로지스틱 회귀분석** | 주관적 체형 인지 | 성별 | 만 나이 | 2주이상 연속 우울감 여부 | 주관적  건강상태 |
| **Random**  **Forest** | 주관적 체형 인지 | 성별 | 만 나이 | 하루 평균  앉아있는 시간 | 주중 하루  평균 수면시간 |
| **KNN** | 주관적 체형 인지 | 만 나이 | 하루 평균  앉아있는 시간 | 1주일간  근력운동 일수 | 주관적  건강상태 |

**[ 남자 청소년 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **데이터** | **1위** | **2위** | **3위** | **4위** | **5위** |
| **로지스틱 회귀분석** | 주관적 체형 인지 | 주중 하루  평균 수면시간 | 만 나이 | 2주이상 연속 우울감 여부 | 평소 스트레스 인지 정도 |
| **Random**  **Forest** | 주관적 체형 인지 | 만 나이 | 하루 평균  앉아있는 시간 | 2주이상 연속 우울감 여부 | 주중 하루  평균 수면시간 |
| **KNN** | 주관적 체형 인지 | 만 나이 | 1주일간  근력운동 일수 | 주관적  건강상태 | 평소 스트레스 인지 정도 |

**[ 여자 청소년 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **데이터** | **1위** | **2위** | **3위** | **4위** | **5위** |
| **로지스틱 회귀분석** | 주관적 체형 인지 | 만 나이 | - | - | - |
| **Random**  **Forest** | 주관적 체형 인지 | 만 나이 | 하루 평균  앉아있는 시간 | 주중 하루  평균 수면시간 | 1주일간  근력운동 일수 |
| **KNN** | 주관적 체형 인지 | 만 나이 | 하루 평균  앉아있는 시간 | 주관적  건강상태 | 평소 스트레스 인지 정도 |

* 대다수 항목에서 차이가 났으나, 공통적으로 ‘주관적 체형인지’는 남/녀 통합, 남자 청소년, 여자청소년 모두에서 유의미하게 드러난 것으로 보아 가장 유의미한 인자로 여겨짐
* 남/녀 통합 데이터를 보면 ‘주관적 체형 인지’ 이외에 ‘만 나이’, ‘하루 평균 앉아 있는 시간’, ‘주관적 건강상태’, ‘성별’, ‘평소 스트레스 인지정도’가 상위 5개안에 2번이상의 Frequency빈도로 추출되어 비만 유의미한 인자로 추측됨
* 남자 청소년 데이터를 보면 ‘주관적 체형 인지’ 이외에 ‘만 나이’ 또한 3개의 모델에서 모두 추출됨. 그 외에 ‘하루 평균 수면시간’, ‘2주 이상 연속 우울감 여부’, ‘평소 스트레스 인지 정도’이 상위 5개안에 2번이상의 Frequency빈도로 추출되어 유의미한 인자로 추측됨
* 여자 청소년 데이터를 보면 ‘주관적 체형 인지’ 이외에도 ‘만 나이’ 또한 3개의 모델에서 모두 추출됨. 그 외에 ‘하루 평균 앉아있는 시간’이 상위 5개 안에 2번이상의 Frequency빈도로 추출되어 유의미한 인자로 추측됨.
* 특히 여자 청소년의 경우 로지스틱 회귀분석에서 유의미한 변수가 2개만 추출되었는데 이는 우선적으로 P-value에 의해 통계적으로 유의미하지 않다고 판단했고 더불어 coefficient의 값이 너무 작아 분석가의 판단하에 불필요한 변수라고 판단

□ PatientsLikeMe의 Business Problem 및 청소년 비만 솔루션 제안

1. PatientsLIkeMe의 가입자 증가 및 수익 증대

* 전세계의 83만명 환자들이 모인 거대한 SNS서비스인 PatientsLikeMe의 특성상 환자들이 자신의 상태가 어떠한지, 증세가 어떤지, 어떤 약을 언제부터 복용했고 효능과 부작용이 어떠했는지 자기 중심적으로 그리고 개인적으로 작성.
* 특히 청소년 비만의 경우 PatientsLikeMe에는 청소년 비만에 관심있다고 한 부류 단 1명만 존재하는데 한 명의 의견으로는 객관적 성격을 띌 수 없음.
* 비만은 다양한 요인의 복합적인 상호작용을 통해 나타나는 결과로 이는 사람에 따라 증상이 같아도 사람의 신체적 구조상 차이가 있고 주관적인 성격을 띄고 있기에 일반화하기에는 한계가 존재.
* 이번 ‘청소년 비만 예측 및 패턴 도출’은 주요 만성질환의 발생 위험을 높이는 건강문제의 결정적 요인인 비만을 주관적인 지표들을 모아 비만에 영향을 미치는 개인의 생활관련 특성과 심리/정신적 특성을 알고자 시행하였으며 이는 차후 잠재적인 비만 청소년을 위한 예방교육 및 상담과 조기중재에 필요한 기초자료가 될 것이며 PatientsLikeMe입장에서 통계적으로 추출된 유의미한 인자들을 제시한다면 가입자 수가 늘어나 PatientsLikeMe입장에서도 데이터가 축적되어 수익이 증가할 것으로 기대 (Sunhee Park, 2011)

1. 사회적 낙인에 민감한 청소년의 비만 예방 중요성 재고

* 비만은 게으르거나 의지가 약하다는 등의 부정적인 인식으로 사회적 낙인이 되는 경우가 있어 이를 해결하기 보다는 회피하고 사회적으로 고립되는 경우가 많아 신체적/정신적 건강에 모두 부정적인 영향을 미침 (Sunhee Park, 2011)
* 이는 본 연구에서 남/녀 통합 청소년, 남자 청소년, 여자 청소년에서 모두 유의미하다고 나온 ‘주관적 체형 인지’를 설명할 수 있으며 주관적으로 뚱뚱하다고 생각하는 만큼 비만인 청소년들이 많았으며 이는 오히려 부정적 인식으로 인해 사회적 고립이 되어 문제를 해결하기보단 피하려고만 하는 특성이 잘 반영됨
* 특히 청소년의 경우 정신적으로 성숙하지 않아 더 민감하며 지속적인 비만예방과 관리가 이루어질 수 있도록 학교에서의 제도적 환경의 노력이 개인적인 수준에서 건강생활습관 실천으로 이어질 수 있게 토대를 만들어 주는 것의 중요성 재고

1. 향후 활용방향

* 본 연구에서 제안한 비만 여부 예측을 위한 분석 기법은 비만 예측에 활용될 수 있을 뿐만 아니라 이를 바탕으로 통합/남/녀 데이터별 비만 예방 정책이나. 건강 설문조사 항목 등에 효과적으로 쓰일 수 있을 것으로 기대한다. 본 연구에서 제안한 6가지 기법의 성능 지표 비교를 통해 최적의 솔루션을 제시할 수 있고 향후 Feature를 더 보완함으로써 Recall을 높이는 개선 작업도 기대된다.

**[ Limitation ]**

■ **컴퓨팅 리소스의 한계**

**:** 본 연구에서는 여섯 가지 분석 기법, 네 가지 결측치 처리 방법, 세 가지 데이터, 두 가지 교차 검증, 스케일링 관련 세 가지 방법 (표준화/정규화 및 스케일링을 적용하지 않은 경우)를 고려하여 이 모든 경우에 대해 총 432개의 모델을 수립하였다. 또한 각 모델 수립 시 모델의 성능을 개선하기 위하여 각 모델별로 Grid Search를 진행했다.

앞서 Method 단계에서 간단하게 언급한 바와 같이, 국민건강영양조사 2016년도, 2017년도, 2018년도 자료와 청소년건강행태조사 2016년도, 2017년도, 2018년도 자료를 통합하고 전처리를 한 결과 18만 개 이상의 Instance로 구성된 데이터셋을 만들었다. 팀원 개인이 가진 컴퓨팅 리소스로는 부족하다고 판단하여 구글 코랩을 이용했다. 하지만 너무 많은 Instance가 존재해서 Grid Search 및 K-Fold Cross Validation을 이용하여 모델을 수립하기에는 너무 많은 시간이 소요되어 정해진 기간 내에 프로젝트를 수행하는 것이 불가능하다고 판단했다. 일례로, Support Vector Machine에서 5-fold cross validation을 이용해 모델을 하나 수립하는데 3시간 이상이 소요되었다. 즉, 컴퓨팅 리소스의 부족으로 인해 원래 사용하려고 했던 데이터를 사용하지 못하고 차선책으로 국민건강영양조사 2016년도, 2017년도, 2018년도 자료와 청소년 건강행태조사 2018년도 자료를 통합하여 사용했다.

**■ 가중치 적용의 한계**

**:** 제7기로 지정된 16년, 17년, 18년도의 설문 항목들이 서로 동일하지 않아 부재하는 항목이 존재해 일부 데이터를 사용할 수 없다. 다른 기에 해당하는 데이터를 가져올 때는 가중치를 처리해야하는 어려움이 있다. .

**■ 결측치 처리 방법의 한계**

: 결측치는 분석결과의 편향이나 통계적 검정력의 문제를 일으킨다. 이에 연구결과가 모집단의 특성을 대표하지 못하는 자료에 의존하는 문제점이 발생할 수 있다. 결측치를 보완하는 가장 완벽한 방법은 모름 및 무응답자들에 대한 재설문조사이지만 이는 시간과 비용이 많이 든다는 단점이 있다.

완전제거법의 경우, 실제 설문자료를 분석하여 완전제거법을 적용하여 분석하는 경우에 평균 및 분산과 같은 기술통계치 문제, 변수 간 관계를 분석하는 회귀분석의 결과에도 영향을 미침.

단일 대체법은 평균값, 중앙값, 최빈값 대체 등을 통해 완전한 데이터셋을 구성할 수 있다는 장점이 있으나 대부분 편의추정량을 발생시키는 것으로 알려짐. 추정량의 또한 결측치를 하나의 값으로 대체하기 때문에 표준오차가 과소추정 되는 문제가 있음. 특히 평균대체법은 결측된 변수의 분산을 작게 하는 문제와 함께 다른 변수와의 상관관계를 낮추는 등 분석결과의 편의를 가져옴. (고길곤 외 1인, 2016)

따라서 완전제거법과 단일대체법과 같은 전통적 결측치 처리방법보다 발전된 다중대체법을 사용하거나 시간과 비용의 문제를 해결 가능하다면 재조사법을 이용하여 결측치를 처리한 후 분석을 수행하면 분석 결과의 타당성이 향상될 것으로 예상됨.

**■ 참고문헌**

1. 윤영숙, 박혜순, "비만과 만성 신장 질환", 대한비만학회지 18-4, 2009, p.127.
2. 지영주, 김영혜, "Factors Influencing Obesity among Adolescent: Analysis of 2011 Korean Youth Risk Behavior Survey" , Korean J Obes Vol.22(1), 2011, pp. 45 ~ 46.
3. Md. Saiful Islam, Abdullah Al Mahmud, Md. Rafiqul Islam, "Machine Learning Approaches for Modeling Spammer Behavior", 2010, pp. 11~13.
4. NA YOUNG SHIN, MA AND MIN SUP SHIN, PHD, "Body Dissatisfaction, Self-Esteem, and Depression in Obese Korean Children", The Journal of Pediatr 152:502-6, 2008, pp. 503~506.
5. Sunhee Park, Association Between Short Sleep Duration and Obesity Among South Korean Adolescents, Western Journal of Nursing Research 33(2), 2011, pp. 207–223.
6. 정보람, "소아·청소년 비만에 영향을 미치는 요인." 국내석사학위논문 중앙대학교 건강간호대학원, 2020
7. 유지호, "청소년의 비만 관련요인과 체중 조절 실태에 관한 연구." 국내박사학위논문 조선대학교, 2014
8. 박소연 ( Park So-youn ), "청소년의 비만 관련 요인에 대한 다층모형 분석." 保健社會硏究 39.3,2019, pp. 314-347.
9. 고길곤, 탁현우, “설문자료의 결측치 처리방법에 관한 연구: 다중대체법과 재조사법을 중심으로”, 행정논총2016;54(4). pp. 291-319.
10. Picuki[웹사이트]. (2020.09.25). URL: <https://www.picuki.com/media/2075086259748794609>
11. Findhealthclinics[웹사이트]. (2020.09.25). URL: <https://www.findhealthclinics.com/US/Los-Angeles/518301005193479/Concierge-Detox-La>
12. Harvard Business School[웹사이트]. (2020.09.25). URL: <https://www.hbs.edu/openforum/openforum.hbs.org/goto/challenge/understand-digital-transformation-of-business/patientslikeme-applying-analytics-to-patient-reported-data.html>
13. PatientsLikeMe[웹사이트]. (2020.10.17). URL: <https://www.patientslikeme.com/>
14. cc.gatech.edu[웹사이트]. (2020.10.18). URL: <https://www.cc.gatech.edu/~jarulraj/courses/8803-f19/slides/13-training-2.pdf>
15. towardsdatascience[웹사이트]. (2020.10.18). URL: <https://www.google.co.kr/url?sa=i&url=https%3A%2F%2Ftowardsdatascience.com%2Fa-glance-into-competitive-data-science-the-best-practices-for-computer-vision-2c77c5d98d19&psig=AOvVaw0UMRQYM3dDukdjT73ZoMp_&ust=1603989151300000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCKiS-Lbb1-wCFQAAAAAdAAAAABAD>
16. towardsdatascience[웹사이트]. (2020.10.18). URL: <https://towardsdatascience.com/introduction-to-na%C3%AFve-bayes-classifier-fa59e3e24aaf>
17. Wikipedia[웹사이트]. (2020.10.19). URL: <https://en.wikipedia.org/wiki/Random_forest>
18. 임우택 박사의 SPSS 통계 이야기[웹사이트]. (2020.10.19). URL: <https://blog.naver.com/y4769/221851780608>
19. Entheoscientist[웹사이트]. (2020.10.19). URL: <https://nurilee.com/data-science-model-summary-linear-ridge-lasso-elasticnet/>
20. meme[웹사이트]. (2020.10.20). URL: <https://me.me/i/new-example-to-classify-class-a-class-b-k-ec823b9d710e48009c8ff6568c53d676>
21. medium[웹사이트]. (2020.10.21). URL: <https://medium.com/@KCVC/taking-a-stab-at-data-science-day-1-c429b4c802d5>
22. ugong2san[웹사이트]. (2020.10.21). URL: <https://ugong2san.tistory.com/2062>
23. Researchgate[웹사이트]. (2020.10.21). URL: <https://www.researchgate.net/publication/315797348_VOICE_COMMAND_BASED_OBJECT_RECOGNIZING_ROBOT_USING_SPEECH_AND_IMAGE_FEATURE_EXTRACTION/figures?lo=1>