과 목 명: 데이터 마이닝

담당교수: 진창호 교수

학 과: 산업경영공학과

학 번: 2016100937 김성수

2016100946 김효준

2016101000 홍성희

Advanced DATA MINING

Final Report

<청소년 비만 예측 및 패턴 도출>



발표 영상: https://drive.google.com/file/d/18Hn-yED0aTbL2a8DEBTzzCxySNP29z6M/view?usp=sharing

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 8
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 18**
14. results 25
15. 모델링 결과
16. 변수 중요도 확인
17. discussion 55
18. limitation 60

**Summary**  
■ **Introduction**  
- 세계 비만 인구는 2008년 기준 총 14억 명, 매년 약 280만 명이 비만 관련 질병으로 사망   
- 청소년기 비만은 성인으로 이어질 확률이 높으며 따라서 Random Forest, Logistic Regression 등 분류 기법을 이용한 청소년 비만 사전 예측 및 청소년 비만 유발 요인 파악이 중요  
- 대상 기업 : PatientsLikeMe – “환자들의 페이스북”  
■ **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  
- Intermediate Report 연구 결과를 바탕으로 각 데이터별로 모델의 최고 성능을 보여준 결측치 처리 방법, 스케일링 여부, K = 5/10인 K-fold 교차 검증 수행  
- 4가지 분류기법(Random Forest, Logistic Regression, SVM, MLP)사용  
- 추가적으로 오버샘플링, PCA, K-Means 기법을 이용하여 모델의 Performance 변화를 확인  
🡺모델 별 그리드 서치 수행하며 청소년 통합/남자/여자 데이터 각각에 대해 모델 수립.  
■ **Results 및 Discussion**

|  |  |  |  |
| --- | --- | --- | --- |
| **모델** | **오버 샘플링 효과**  **(O/X)** | **기법 별 효과 (O / △ / X)** | **교차 검증 CV값 변화 효과 (O / △ / X)** |
| **Random Forest** | O | △ | O |
| **Logistic Regression** | O | X | △ |
| **SVM** | O | O | △ |
| **MLP** | O | - | - |

- 성능척도에 AUC와 F2 Score를 추가하였고 성능 비교 간 F2 Score을 활용하여 기존 F1 Score에 본 연구의 주요 척도인 recall에 대한 가중치를 부여하여 비교  
- “기법 별 효과”는 Random Forest는 PCA의 효과를, Logistic Regression은 K-mean전처리의 효과를, SVM은 비선형 Kernel SVM의 성능향상에 대한 효과를 기재  
- △는 세 데이터 중 하나라도 성능이 하락한 경우 기재, -는 해당사항이 없는 항목에 대해 기재  
- 비선형 커널 트릭으로 데이터의 비선형성까지 포착하여 model의 예측력 견인

|  |  |
| --- | --- |
| **데이터** | **유의미한 인자(상위 5개 요인 중 2번 이상의 빈도로 등장)** |
| **통합** | 주관적 체형인지, 성별, 만나이, 1주일간 근력운동 일수, 주관적 건강상태 |
| **남자** | 주관적 체형인지, 만나이, 2주이상 연속 우울감 여부, 주관적 건강상태 |
| **여자** | 주관적 체형인지, 만나이, 하루 평균 앉아있는 시간, 주관적 건강상태 |

**🡺 청소년 남자, 여자 비만에 영향을 주는 인자가 거의 비슷하지만 미세하게 차이를 보임**■ **Limitation** – 컴퓨팅 리소스 및 가중치 적용의 한계, 결측치 처리법의 한계, PCA 해석의 한계

**[ Introduction ]**

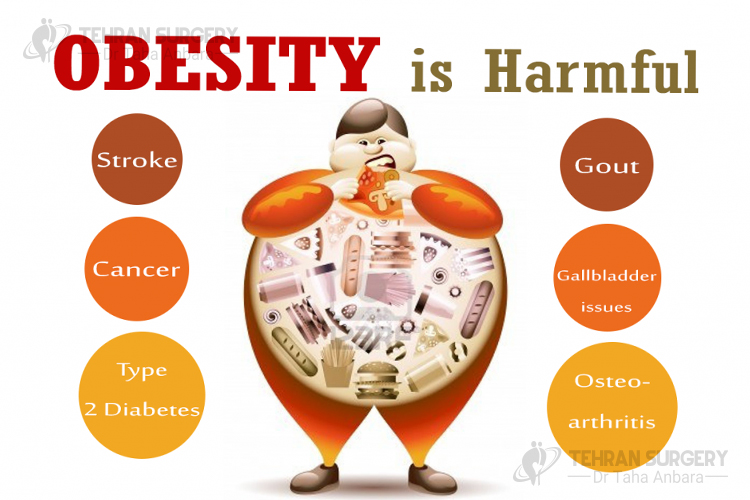
****■ **과제 수행 배경 및 동기**

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**

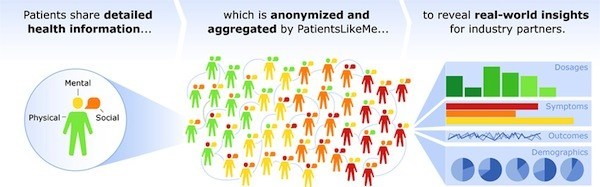
□ 대상 기업 : PatientsLikeMe - “환자들의 페이스북”

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

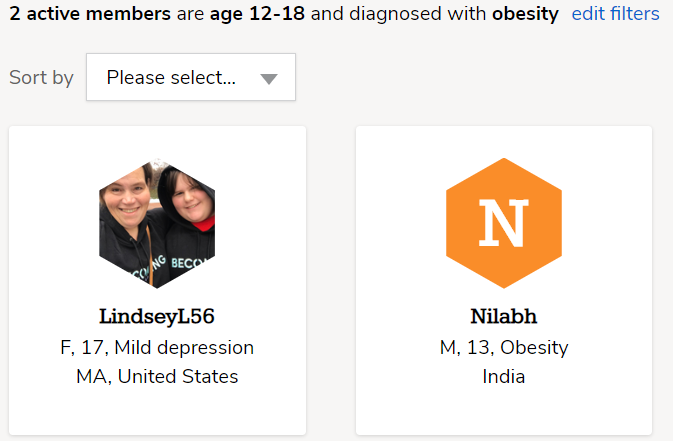


그림 4 PatientsLikeMe 활용 사례

- 83만 명이 넘는 PatientsLikeMe 가입자들 중 비만이 있는 가입자의 수는 3248명에 불과하고 이 중 청소년의 수는 단 두 명에 불과 ( 2020.11.25 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 : Kernel 및 PCA, 오버 샘플링을 활용하여 SVM, Random Forest, Logistic Regression 등   
 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. (2020.11.30 기준)

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

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

표 1 방법론별 성능 정리

- 위 표는 Intermediate Report에서 연구한 각 모델 별 결측치 처리 방법, 스케일링 효과, K-fold 교차 검증 K값 변화에 따른 성능 변화 결과를 요약, 정리한 표다. 본 연구에서는 청소년 비만 Mechanism파악을 위해 변수 중요도를 파악하기 위하여 Random Forest와 Logistic Regression을 사용했다. 또한 SVM을 이용하여 청소년 비만 여부를 예측하고자 한다.

- SVM은 변수 중요도를 파악하기 위해서 Linear SVM을 사용해야 하므로 본 연구에서는 Intermediate Report와 달리 Linear SVM을 사용하여 변수 중요도를 도출하고자 한다. 이를 커널을 사용한 경우와 비교를 통해 SVM의 효과를 강조한다. 대략적으로 말하자면 커널 트릭을 사용하여 계산 비용을 크게 줄이며 마치 특성을 추가하는 듯한 효과를 지닌다.

- 또한 본 연구의 데이터셋은 Class Imbalance문제가 심각하기에 오버 샘플링 기법을 통하여 Class Imbalance문제를 해결하고자 한다. 선행 연구에 의하면, SVM으로 모델링 시 기존 데이터셋보다 Minority Class에 대해 Class Imbalance를 해결하여 비율을 맞춰준 데이터 셋에 대하여 더 우수한 성능을 보여주었다고 밝혀졌다. (Alexander Liu, 2007) 따라서 본 연구에서 SVM모델을 사용하여 높은 성능을 보여줄 것으로 사료된다.

- 변수가 많아져 고차원이 될수록 모델은 학습속도가 느려 지며 과대적합에 빠지기 쉽다. 이러한 원인은 차원의 저주에 비롯되는데 이를 통해 일반화성능이 저하될 수 있다. 이를 PCA기법을 활용해 기존 데이터의 설명력은 유지하면서 차원의 저주를 해결하여 일반화 성능 및 학습속도를 개선하는 효과가 기대된다.

- 즉, 본 연구에서는 Logistic Regression, SVM, Random Forest와 같은 분류 기법, 고차원 커널, 차원 축소 PCA기법을 이용해 청소년 비만 예측 모델 수립 및 청소년 비만 영향 요인 별 중요도를 파악하고자 한다.

■ 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, Random Forest 즉, 분류 분석에 사용되는 머신 러닝 기법을 이용해 청소년 비만 예측 모델 수립 및 청소년 비만 영향 요인별 중요도 파악

■ 논문 주요 결과의 개요

- 커널을 통한 모델링이 기존 선형적인 모델에 비해 성능이 좋았으며, PCA를 통해 차원축소를 한 결과 모델의 Complexity가 25%가 감소하면서 거의 비슷한 성능을 얻을 수 있었다. 하지만 K-mean전처리 방법은 예측에 있어 좋은 성능을 이끌어내지 못했고 오히려 성능 하락을 가져왔다.

- 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 인자의 적합성은 ‘남자 청소년 > 여자 청소년 > 남/녀 통합 청소년’ 순으로 적합하다는 것을 F2 Score를 통해 파악했으며, 모든 모델에서 위와 같은 순서를 보여주어 신뢰성을 확보할 수 있었다.

- 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 기 전체 자료와 청소년 건강행태 조사 자료 2016년, 2017년, 2018년도 데이터를 다 통합하여 분석을 수행하려 했으나, 본 연구에서는 Grid Search와 오버 샘플링을 활용하여 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\_all\_over** | 83282 | 41642 | 41642 | 50 |
| **df\_all\_b\_over** | 38206 | 19103 | 19103 | 50 |
| **df\_all\_g\_over** | 45078 | 22539 | 22539 | 50 |

표 2 데이터 소개

본 연구에서는 앞서 Introduction에서 언급한 바와 같이 Logistic Regression, Random Forest, SVM, MLP를 사용했다. 따라서 [표 2]를 참고하여 각 기법 별로 가장 높은 성능을 보인 결측치 처리 방법, 스케일링 방법에 알맞은 데이터셋을 생성하였다. 추가적으로 클래스 불균형 문제를 보완하기 위해 오버샘플링 기법을 적용한 데이터셋을 생성하였다.

* 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  
   \*비만율 (%) = (비만 인원 수 / Instance 개수 ) \* 100
* 오버 샘플링을 이용하여 남녀 통합/남자/여자 데이터의 class minority인 비만인 인원을 비만이 아닌 인원의 수만큼 증가시켜 주어 비만율을 50%로 맞춤

■ 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)

: “남녀 통합 청소년” 데이터 셋에 대하여 오버 샘플링 전과 후 데이터 셋에 대해 실시하였다.

(1) 오버 샘플링 전 기존 데이터

|  |  |
| --- | --- |
| [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 |

표 3 기존 데이터 EDA

(2) 오버 샘플링 후 데이터 (전처리 후)  
 \*앉아있는 시간의 시간과 분을 통합하였고 무응답 및 NA값을 제거하였다.

|  |  |
| --- | --- |
| [성별] | [만나이] |
| : 남성 혹은 여성의 값만 가짐 | : 12세 청소년은 다소 적고, 18세 청소년이 다소 많음 |
| [주관적 건강 상태] | [평소 스트레스 인지 정도] |
| : Positive skew되어있으며 ‘5’로 응답한 인원의 수는 매우 적음 | : 비교적 정규분포의 형태를 지님 |
| [주중 하루 평균 수면시간] | [평소 하루 앉아서 보내는 시간(시간)] |
| : 이상치 제거 후 정규분포의 형태를 지님 | : 이상치 제거 후 정규분포의 형태를 지님 |
| [1주일간 근력운동 일수] | [2주이상 연속 우울감 여부] |
| : 완벽하게 Positive skew되어있고 값들이 1에 많이 몰려 있는 것을 볼 수 있음. | : 이상치가 제거된 후 설문조사에 필요한 1과 2 값만 가짐 |
| [BMI] | [BO1] |
| : 기존 신장과 체중에 대한 결측치를 제거한 후 체중/(신장)^2을 실시한 후 25를 기준으로 25이상을 1로, 25미만은 0으로 할당 | : 정규분포의 형태를 띄고있음 |

표 4 오버 샘플링 데이터 EDA

□ 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

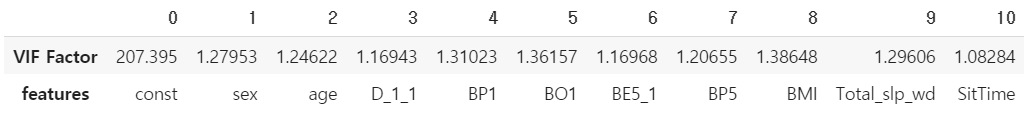
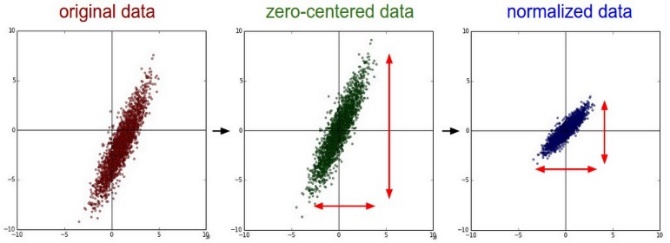


표 5 다중공선성 VIF

□ 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

2) 데이터 스케일링   
①정규화 : 데이터의 범위를 0과 1로 변환하여 데이터 분포를 조정

그림 5 데이터 표준화 및 정규화

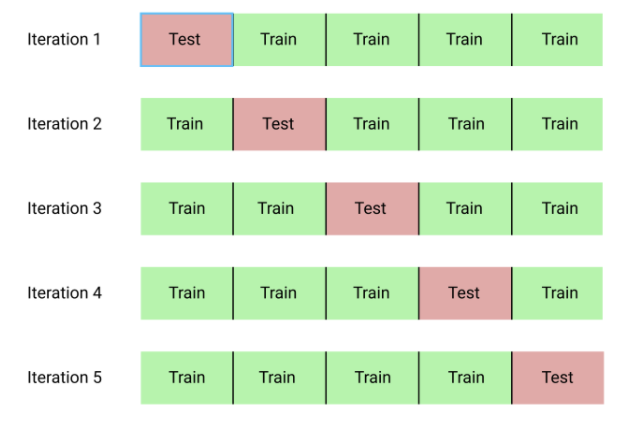
**□ K-fold 교차검증(K-fold Cross Validation)  
**: 전체 데이터의 일부를 Validation Set으로 사용해 모델 성능을 평가하는 것  
 🡪 데이터 셋의 크기가 작은 경우 테스트셋에 대한 성능 평가 개선에 효과적

그림 6 Cross Validation 설명

**□ 언더샘플링 & 오버샘플링**

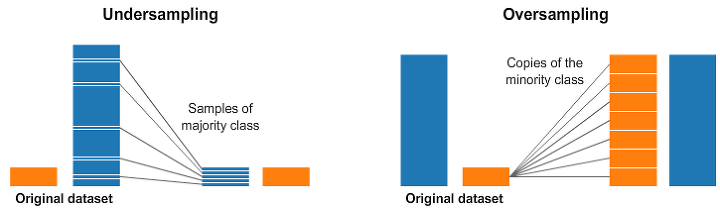


그림 7 언더샘플링 & 오버샘플링

□ 언더샘플링  
: 많은 데이터 셋을 적은 데이터 셋 수준으로 감소시키는 방식  
Ex) 비만 = 0 레이블을 가진 인스턴스가 10,000개, 비만 = 1 레이블을 가진 인스턴스가 100개  
🡪 비만 = 0 레이블인 인스턴스를 100개로 줄임

□ 오버샘플링  
: 적은 데이터 셋을 많은 데이터 셋 수준으로 증가시키는 방식  
Ex) 비만 = 0 레이블을 가진 인스턴스가 10,000개, 비만 = 1 레이블을 가진 인스턴스가 100개  
🡪 비만 = 1 레이블인 인스턴스를 10,000개로 늘림

🡺 지도 학습 시, Class Imbalance 문제를 해결하기 위해 사용

**[ Modeling ]**

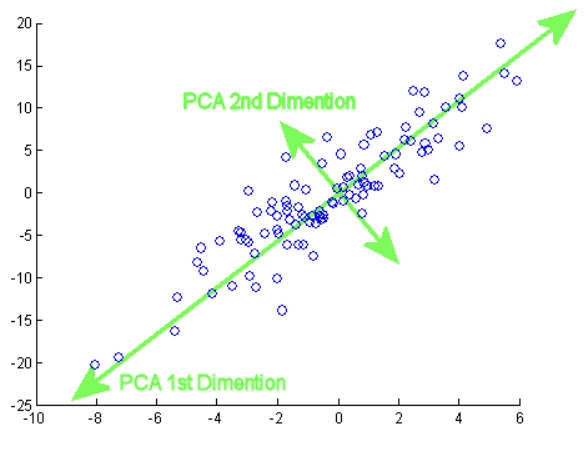
**1. PCA(Principle Component Analysis)**

그림 8 PCA

□ 정의

- MNIST 등 데이터 셋의 경우 샘플에 대하여 수 백 개 이상 심지어 수백만 개 이상의 특징을 가진 데이터들이 있다. 이렇게 많은 데이터는 오히려 학습을 느리게 하고, 좋은 솔루션을 찾기 어렵게 하는 차원의 저주에 빠지기 쉬워 과대적합에 빠지기 쉽다.

- PCA에서 차원 축소 시 정보를 최대한 보존하는 방향으로 투영(Projection)을 시키는데 이러한 축을 주성분(Principle Component)라고 한다. 주성분은 내림차순으로 배열되며 첫 주성분의 설명력이 가장 높다.

□ 원리

- 핵심은 차원축소 된 더 적은 데이터로 기존의 데이터만큼 충분한 설명력을 유지하는 것이며 분산을 최대화하는 방향으로 원본 데이터와 원본 데이터의 차원을 낮은 차원으로 투영한 초 평면 사이의 평균 제곱거리를 최소화하는 축으로 투영하는 것이다.

- 특잇값 분해를 활용하여 훈련 행렬인 X를 로 분해한다. 여기서 주성분의 단위벡터는 V에 담겨있다. 🡪 축소된 데이터 셋 위와 같은 고유벡터에 기존 훈련 행렬 X를 곱함으로써 얻을 수 있다.

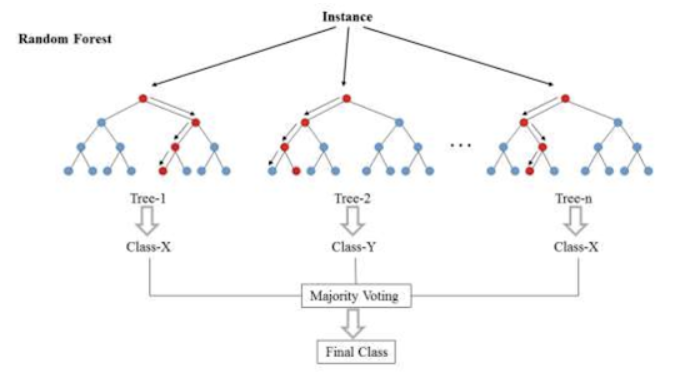
□ 장단점

- 압축된 데이터를 획득하여 기존 데이터셋 보다 feature의 수가 줄어 value수가 줄어들게 되면서 데이터는 소량의 정보손실을 얻지만 계산속도는 상승하고 과대적합을 피할 수 있다.

- 우리가 구한 방향으로 Projection한 것이 우리가 찾고자 하는 방향인지 알 수 없다. 분산이 가장 큰 방향으로 Projection한다고해도 이상적인 것인가에 대한 고찰이 필요하다.

- 생성된 Principle component에 대하여 해석이 어렵다. PCA의 치명적인 문제로 성능적으로는 손실 대비 우수한 측면을 제공하나, descriptive측면에서는 기존 데이터셋에 비해 확연히 떨어진다.

**2. 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 9 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**

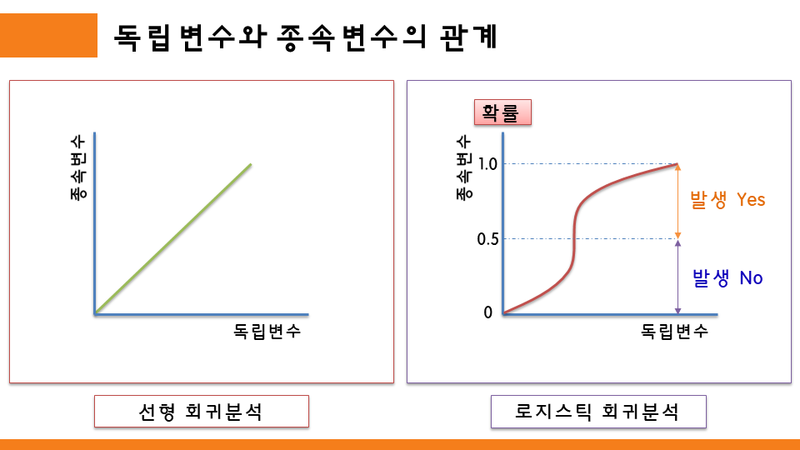
**-** 종속변수의 계급이 0과 1 두 가지 값을 가지고 관심의 대상이 되는 계급이 1인 경우에 종속변수가 1이 될 확률을 예측하는 모형  
Ex) 종속변수 값, 즉 확률이 0.5보다 크면 사건이 일어나며, 0.5보다 작으면 그 사건이 일어나지 않는 것으로 예측  
**-** 독립 변수는 연속형, 범주형 모두 가능  
**-** 종속 변수는 범주형

그림 10 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

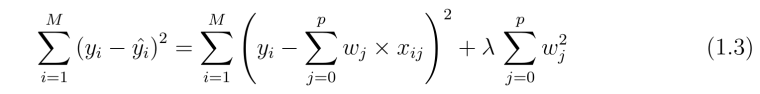


그림 11 Ridge Cost Function

1. Lasso : L1 정규화

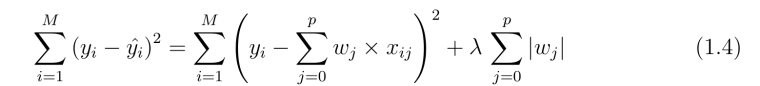


그림 12 Lasso Cost Function

1. Elastic Net : L1 정규화와 L2 정규화 혼합

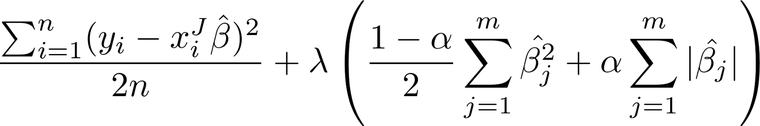


그림 13 Elastic Net Cost Function

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

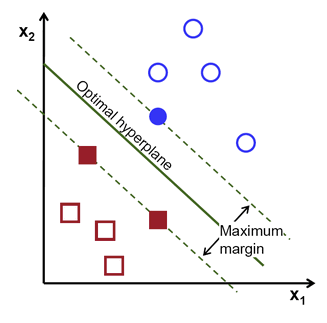
훈련 데이터들을 서로 다른 두 개의 클래스로 분류할 때, 분류의 기준이 되는 분리 경계면을 학습 알고리즘을 이용하여 모색

그림 16 SVM

그림 14 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

**5. MLP (Multi-Layer Perceptron)**

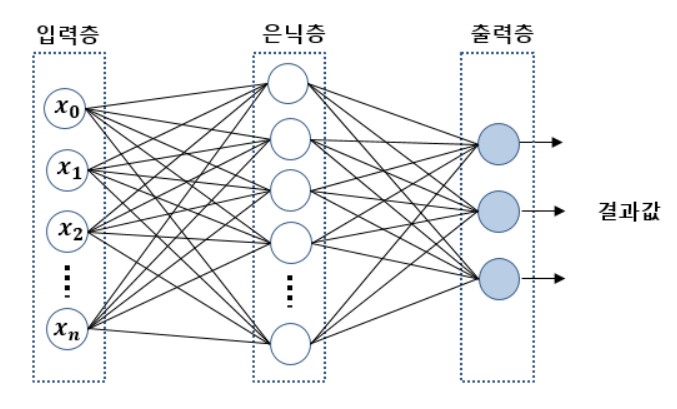
□ **정의**  
- 단층 퍼셉트론의 한계: 비선형적으로 분리되는 데이터에 대해 학습 불가능  
Ex) AND 연산에 대해서는 학습이 가능하지만, XOR에 대해서는 학습이 불가능  
🡺 입력층 (Input Layer)과 출력층 (Output Layer) 사이에 하나 이상의 중간층(은닉층, Hidden Layer)을 두어 비선형적으로 분리되는 데이터에 대해서도 학습이 가능하도록 고안된 인공 신경망

그림 15 MLP

□ **MLP의 기본 구조**  
1) 한 개의 입력층  
2) TLU들로 구성된 하나 이상의 층 = 은닉층  
3) TLU들로 구성된 하나의 마지막 층 = 출력층  
4) 층들 간 Fully Connected ( Ex. 입력층의 모든 노드들과 은닉층의 모든 노드들 사이가 연결됨 )  
5) 출력층을 제외한 모든 층에 하나의 Bias 뉴런   
6) 활성 함수: None, ReLU, Softmax 등  
- 은닉층이 2개 이상인 MLP는 DNN (Deep Neural Network, 심층 신경망)이라고 불림

□ **Backpropagation (역전파)**  
- 다층 퍼셉트론의 은닉층에 있는 노드에는 Target Value가 존재하지 않아 기존의 학습 방법으로는 학습이 불가  
🡺 오류 역전파를 이용해 MLP 학습 가능

□ **MLP의 문제**  
- Overfitting: 은닉층의 수 또는 노드 개수를 과도하게 늘림으로 인해 과대적합 발생 가능  
- Vanishing Gradient: 층이 깊어짐에 따라 역전파를 하는 과정에서 에러 값이 현저히 작아져 학습이 제대로 이루어지지 않을 수 있음

**[ Result ]**

**■** 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

3) F2 Score

|  |  |  |  |
| --- | --- | --- | --- |
| Confusion Matrix | | 실제 정답 | |
| True | False |
| 분류 결과 | Positive | TP | FP |
| False | FN | TN |

표 6 Confusion Matrix

- True Positive (TP): 실제 True인 정답을 True라고(Positive) 예측  
- False Positive (FP): 실제 False인 정답을 True라고(Positive) 예측  
- False Negative (FN): 실제 True인 정답을 False라고(Negative) 예측  
- True Negative (TN): 실제 False인 정답을 False라고(Negative) 예측  
 ①Precision (정밀도): 모델이 True라고 분류한 것 중에서 실제 True인 것의 비율  
 = TP / ( TP + FP )  
 ②Recall (재현율): 실제 True인 것 중에서 모델이 True라고 예측한 것의 비율  
 = TP / ( TP + FN )  
 ③F1 score: 정밀도와 재현율을 동일한 중요도로 반영한 지표  
 =   
 ④F2 score: 정밀도를 2배 더 중요하게 생각하여 정밀도와 재현율을 반영한 지표  
 =

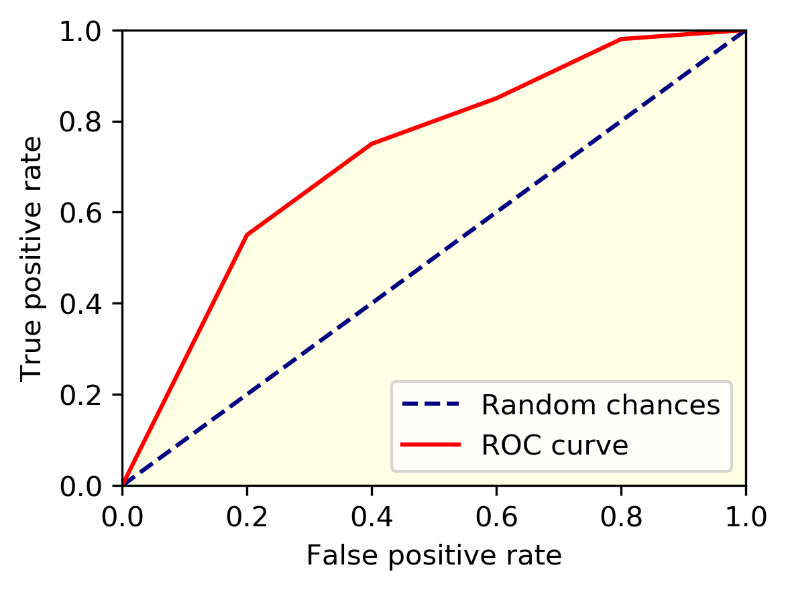
4) AUC ( Area Under Curve )  
- AUC란? ROC 그래프 아래의 면적값  
- Curve가 왼쪽 위 모서리에   
 가까울수록 모델의 성능이 좋다고 평가  
 : Recall이 크고 FPR(False Positive Rate)   
 이 작은 모형이 좋은 모형

그림 16 AUC

- AUC의 최대값 = 1  
 : FPR에 비해 Recall이 클수록 AUC가   
 1에 가까운 값을 가짐

- AUC의 최소값 = 0.5

**[ Result ]**

**■ Random Forest Performance Results  
\*참고 : 각 Performance Measure 기재 시 소수점 넷 째 자리에서 반올림함  
\*RandomForest: n\_estimators, max\_depth, min\_samples\_leaf, min\_samples\_split 그리드 서치  
\*PCA로 age, Total\_slp\_wk, SitTime을 하나의 변수로 차원 축소**

**□ CV = 5 적용**

**[ 청소년 전체 데이터 (남/녀 통합) ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Random Forest** | **Accuracy** | **Precision** | **Recall** | **F2** | **AUC** |
| 원본 | 0.891 | 0.66 | 0.517 | 0.540 | 0.922 |
| 원본(PCA) | 0.886 | 0.646 | 0.467 | 0.494 | 0.914 |
| 오버 샘플링 | 0.777 | 0.388 | 0.936 | 0.730 | 0.924 |
| 오버 샘플링(PCA) | 0.757 | 0.369 | 0.955 | 0.725 | 0.916 |

표 7 RandomForest 청소년 전체 데이터 (CV=5)

**[ 청소년 남자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Random Forest** | **Accuracy** | **Precision** | **Recall** | **F2** | **AUC** |
| 원본 | 0.856 | 0.652 | 0.599 | 0.609 | 0.923 |
| 원본(PCA) | 0.845 | 0.627 | 0.55 | 0.564 | 0.908 |
| 오버 샘플링 | 0.831 | 0.543 | 0.948 | 0.825 | 0.933 |
| 오버 샘플링(PCA) | 0.829 | 0.541 | 0.949 | 0.825 | 0.922 |

표 8 RandomForest 청소년 남자 데이터 (CV=5)

**[ 청소년 여자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Random Forest** | **Accuracy** | **Precision** | **Recall** | **F2** | **AUC** |
| 원본 | 0.926 | 0.701 | 0.355 | 0.394 | 0.885 |
| 원본(PCA) | 0.923 | 0.664 | 0.347 | 0.384 | 0.868 |
| 오버 샘플링 | 0.689 | 0.227 | 0.975 | 0.587 | 0.900 |
| 오버 샘플링(PCA) | 0.689 | 0.226 | 0.975 | 0.587 | 0.884 |

표 9 RandomForest 청소년 여자 데이터 (CV=5)

□ **CV = 10 적용**

**[ 청소년 전체 데이터 (남/녀 통합) ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Random Forest** | **Accuracy** | **Precision** | **Recall** | **F2** | **AUC** |
| 원본 | 0.891 | 0.66 | 0.517 | 0.540 | 0.922 |
| 원본(PCA) | 0.887 | 0.666 | 0.449 | 0.481 | 0.917 |
| 오버 샘플링 | 0.782 | 0.394 | 0.932 | 0.732 | 0.925 |
| 오버 샘플링(PCA) | 0.757 | 0.369 | 0.955 | 0.725 | 0.916 |

표 10 RandomForest 청소년 전체 데이터 (CV=10)

**[ 청소년 남자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Random Forest** | **Accuracy** | **Precision** | **Recall** | **F2** | **AUC** |
| 원본 | 0.857 | 0.656 | 0.597 | 0.608 | 0.932 |
| 원본(PCA) | 0.846 | 0.64 | 0.523 | 0.543 | 0.91 |
| 오버 샘플링 | 0.831 | 0.544 | 0.951 | 0.827 | 0.945 |
| 오버 샘플링(PCA) | 0.83 | 0.542 | 0.95 | 0.825 | 0.927 |

표 11 RandomForest 청소년 남자 데이터 (CV=10)

**[ 청소년 여자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Random Forest** | **Accuracy** | **Precision** | **Recall** | **F2** | **AUC** |
| 원본 | 0.926 | 0.696 | 0.347 | 0.386 | 0.924 |
| 원본(PCA) | 0.923 | 0.664 | 0.347 | 0.384 | 0.868 |
| 오버 샘플링 | 0.690 | 0.227 | 0.975 | 0.588 | 0.899 |
| 오버 샘플링(PCA) | 0.689 | 0.226 | 0.974 | 0.586 | 0.88 |

표 12 RandomForest 청소년 여자 데이터 (CV=10)

□ 데이터 (남,녀 통합/남자/여자 데이터) 기준 해석

1. 남녀 통합 청소년 데이터  
   : CV = 5, CV = 10인 경우를 통틀어 Recall이 가장 높은 경우는 F2 Score: 0.732, AUC: 0.925   
    CV = 5, CV = 10 통합 평균 Accuarcy : 0.829, 평균 Recall : 0.716  
    CV = 5, CV = 10 통합 평균 F2 score : 0.621, 평균 AUC: 0.920
2. 남자 청소년 데이터  
   : CV = 5, CV = 10인 경우를 통틀어 Recall이 가장 높은 경우는 F2 Score: 0.827, AUC: 0.945  
    CV = 5, CV = 10 통합 평균 Accuarcy : 0.841, 평균 Recall : 0.758  
    CV = 5, CV = 10 통합 평균 F2 score : 0.703, 평균 AUC: 0.925
3. 여자 청소년 데이터  
   : CV = 5, CV = 10인 경우를 통틀어 Recall이 가장 높은 경우는 F2 Score: 0.588, AUC: 0.900  
    CV = 5, CV = 10 통합 평균 Accuarcy : 0.807, 평균 Recall : 0.662  
    CV = 5, CV = 10 통합 평균 F2 score : 0.487, 평균 AUC: 0.889

:본 연구에서 관심있는 지표들에 대해 평균을 산출한 결과, 모든 지표에 대하여 ‘남자 청소년 > 남/녀 통합 > 여자청소년’순으로 높은 성능을 보였다. 이는 중간 보고서 때와 같은 흐름을 보였다.

: 가장 높은Recall을 기준으로 봤을 때 ‘여자 청소년 > 남/녀 통합 > 남자 청소년’ 순으로 높은 성능을 나타냈는데 오히려 평균수치와는 반대되는 흐름을 보였다.

🡺 지표 전체적으로 보았을 때 논문에서 추출한 인자들의 경우 남자 청소년 > 남/녀 통합 > 여자 청소년 순으로 적합할 것으로 판단하였다.

□ 오버 샘플링 (학습 데이터에 대해서만 동일이 되도록 오버 샘플링) 기준 해석

: 오버 샘플링 결과, class imbalance문제를 갖고 있던 기존 데이터 셋과 달리 class imbalance해소 결과 눈에 띄는 성능 향상을 보여주었다. 특히 본 연구에서 관심대상인 recall에 대하여 남자, 여자, 남/녀 통합 청소년 모두에 대하여 0.9를 넘는 수치를 보여주었으며 AUC에 대해서도 좋은 성능을 보여주었다.

: 기존 중간 보고서에서는 비만인 청소년을 잘 맞추지 못하여 recall이 낮은 성능을 보였으나, 기존에 잘 맞추던 비만이 아닌 청소년에 대해서는 잘 맞추지 못하여 accuracy는 다소 하락한 수치를 보였음.

🡺 accuracy에서는 다소 약간의 손해를 얻었으나 recall 및 AUC에서 큰 수확을 얻어 오버 샘플링으로 predictive측면을 향상시켰다고 판단하였다.

□ PCA 기준 해석

: 거의 모든 상황에 대하여 PCA를 적용 후 비만인 청소년을 예측한 결과, 중간 보고서 때의 성능과 유사하나 다소 조금씩 하락한 성능을 보여주었다.

: 하지만 이는 모델의 predictive측면을 희생했지만, 계산량 부분에 있어 기존 8개의 변수를 6개로 줄여 데이터의 25%가 소멸되었음에도 불구하고 거의 유사한 성능을 보이는 것으로 설명할 수 있다.

🡺 PCA를 통해 계산량을 획기적으로 줄임과 동시에 기존 데이터의 설명력을 100% 보존할 수 없지만 그에 버금가는 설명력을 지닌 것을 볼 수 있었다.

□ 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** | **F2** | **AUC** |
| **5** | 0.828 | 0.516 | 0.719 | 0.622 | 0.919 |
| **10** | 0.829 | 0.522 | 0.713 | 0.620 | 0.920 |

표 13 RandomForest 남녀 통합 데이터 CV기준 해석

**[ 남자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CV** | **Accuracy** | **Precision** | **Recall** | **F2** | **AUC** |
| **5** | 0.840 | 0.591 | 0.762 | 0.706 | 0.922 |
| **10** | 0.841 | 0.596 | 0.755 | 0.701 | 0.929 |

표 14 RandomForest 남자 데이터 CV기준 해석

**[ 여자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CV** | **Accuracy** | **Precision** | **Recall** | **F2** | **AUC** |
| **5** | 0.807 | 0.455 | 0.663 | 0.488 | 0.884 |
| **10** | 0.807 | 0.453 | 0.661 | 0.486 | 0.893 |

표 15 RandomForest 여자 데이터 CV기준 해석

: CV의 경우 5와 10을 비교해보았을 때 10을 설정함으로써 성능이 확연히 상승하는 것은 볼 수 없었으며 다소 비슷한 수치를 보여주었다.

🡺 CV를 높인 결과 성능이 향상된 척도도 있어 연구의 목적에 따라 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 아닌 것을 알 수 있다.

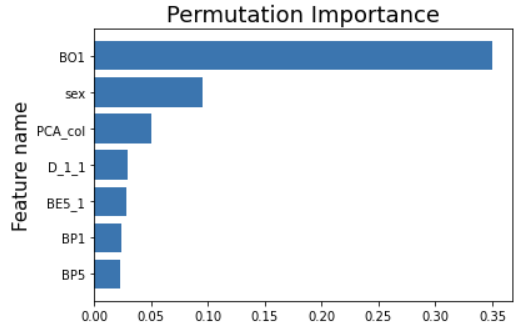
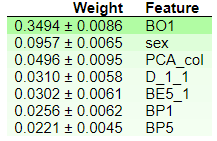
**□ Random Forest활용 변수 중요도 확인  
\*참고: sklearn 패키지의 permutation\_importance를 활용하여 변수중요도 추출  
\*중간보고서 때 cv에 따른 중요도 차이는 유의미하지 않다고 입증하여 분석에서 제외하였음**

1. 청소년 전체 데이터

1-1) 기존 데이터 PCA

* 최적 파라미터 : max\_depth = 12, min\_samples\_leaf = 8, mim\_samples\_split = 20, n\_estimators = 10
* 변수 중요도

그림 17 RandomForest 청소년 전체 데이터 변수 중요도(기존 데이터 PCA)



* Permutation Importance 해석 예시

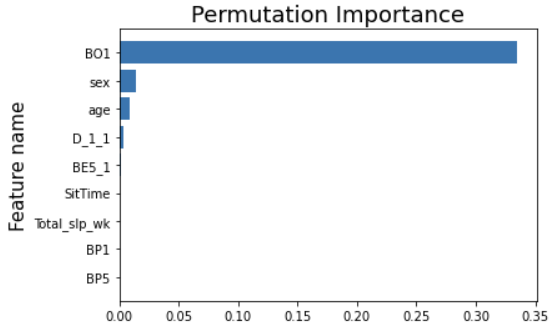
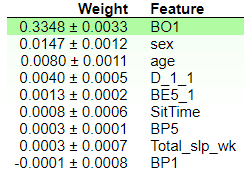
: 위 표에 나타나는 Weight는 value를 shuffling 후 shuffling 전과 비교했을 때 Score가 떨어진 정도를 보여줌

설명변수 ‘BO1’에 대해 0.3494만큼 Score가 가장 많이 떨어졌다면 이는 해당 feature가 예측에 가장 영향력이 있으며 모델이 해당 변수에 가장 많이 의존한다고 해석가능

1-2) 오버 샘플링

* 최적 파라미터 : max\_depth = 6, min\_samples\_leaf = 12, mim\_samples\_split = 8, n\_estimators = 10
* 변수 중요도

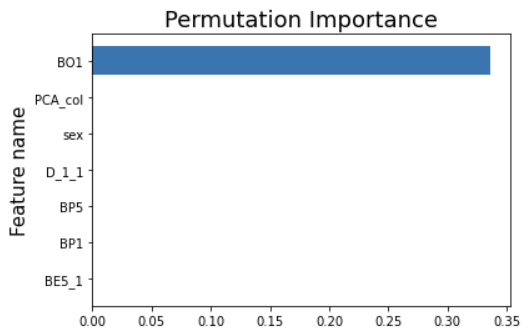
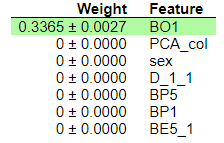
그림 18 RandomForest 청소년 전체 데이터 변수 중요도(오버 샘플링 데이터)



1-3) 오버 샘플링 PCA

* 최적 파라미터 : max\_depth = 6, min\_samples\_leaf = 8, min\_samples\_split = 8, n\_estimators = 100
* 변수 중요도

그림 19 RandomForest 청소년 전체 데이터 변수 중요도(오버 샘플링 데이터 PCA)

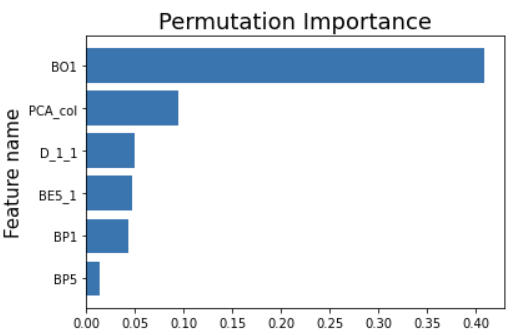
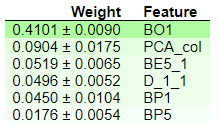


1. 남자 청소년 데이터

2-1) 기존 데이터 PCA

* 최적 파라미터 : max\_depth = 12, min\_samples\_leaf = 8, mim\_samples\_split = 20, n\_estimators = 10
* 변수 중요도

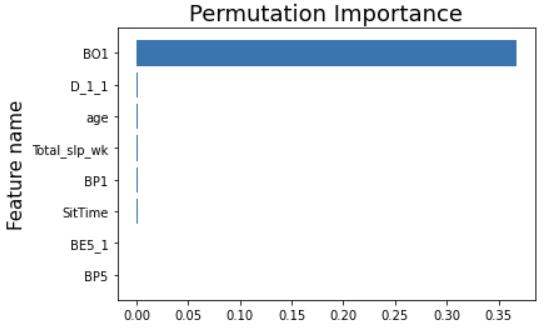
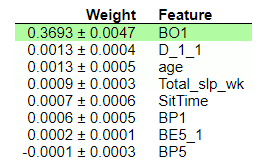
그림 20 RandomForest 청소년 남자 데이터 변수 중요도(기존 데이터 PCA)



2-2) 오버 샘플링

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

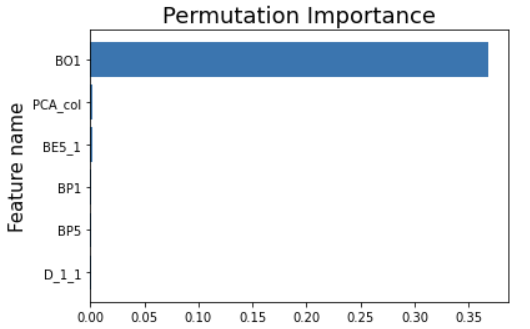
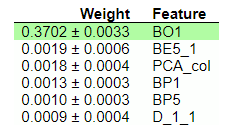
그림 21 RandomForest 청소년 남자 데이터 변수 중요도(오버 샘플링 데이터)



2-3) 오버 샘플링 PCA

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

그림 22 RandomForest 청소년 남자 데이터 변수 중요도(오버 샘플링 데이터 PCA)

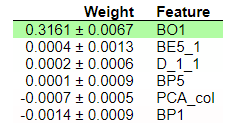
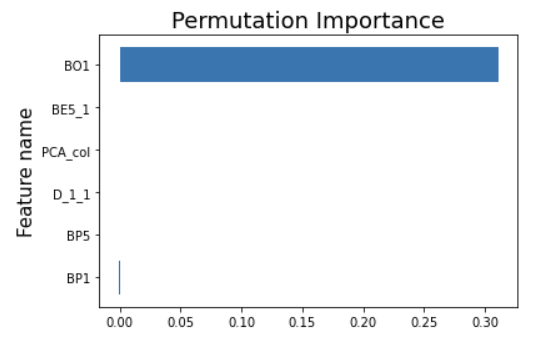


1. 여자 청소년 데이터

3-1) 기존 데이터 PCA

* 최적 파라미터 : max\_depth = 6, min\_samples\_leaf = 18, mim\_samples\_split = 8, n\_estimators = 100
* 변수 중요도

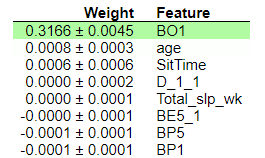
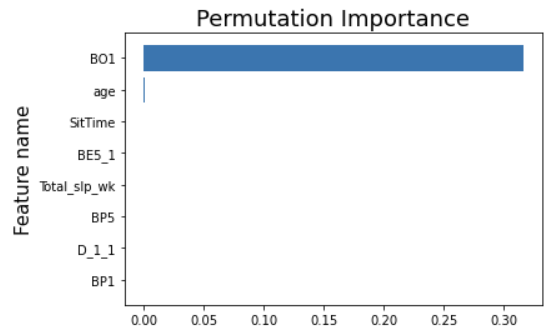
그림 23 RandomForest 청소년 여자 데이터 변수 중요도(기존 데이터 PCA)



3-2) 오버 샘플링

* 최적 파라미터 : max\_depth = 6, min\_samples\_leaf = 12, mim\_samples\_split = 8, n\_estimators = 100
* 변수 중요도

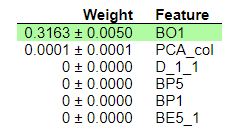
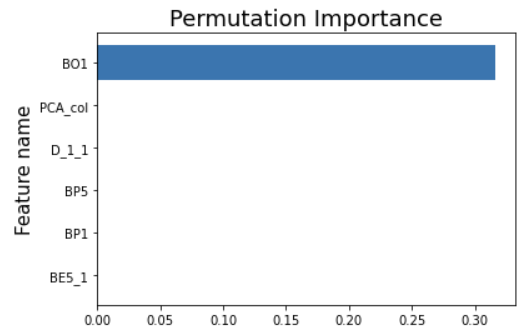
그림 24 RandomForest 청소년 여자 데이터 변수 중요도(오버 샘플링 데이터)



3-3) 오버 샘플링 PCA

* 최적 파라미터 : max\_depth = 6, min\_samples\_leaf = 8, min\_samples\_split = 8, n\_estimators = 100
* 변수 중요도

그림 25 RandomForest 청소년 여자 데이터 변수 중요도(오버 샘플링 데이터 PCA)



* 변수 중요도 확인 결과 해석

: 중간 보고서 당시 사용한 기존 데이터와 해당 데이터에 PCA를 적용한 후 변수 중요도를 확인한 결과 순서의 변동이 크진 않으나 하위항목에 대하여 발견되는 경향이 있었고, 기존 3개의 변수가 1개의 변수로 줄어들면서 기존 변수는 사라지고 빈자리를 하위 변수가 채웠다. PCA로 만들어진 변수는 여자 청소년의 기존 데이터 PCA를 제외하고는 2~3위 정도로 예측에 있어 기여도가 있음을 알 수 있었다.

🡺 모델의 예측 결과는 PCA를 통해 기존에서 약간 하락하였으나 변수 중요도의 순은 크게 바뀌지 않았다. 그러나 PCA변수가 추가되면서 해당 변수의 순위가 높으나 명확한 해석이 불가능한 치명적인 단점을 보여준다.

: [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

🡺 가장 유의미한 인자로 [BO1]가 모두 동일하게 선정되었으며 그 이하 항목들에 대해서는 데이터 별로 다른 성향을 보임:

: 오버 샘플링 결과 변수 중요도가 한 변수에 엄청나게 집중됨을 볼 수 있었다. 기존 데이터는 하위 변수들의 어느정도 기여도가 존재했으나, bar graph에서 보이는 것처럼 사실상 [BOI]의 영향력에 의해 변수의 성능이 결정되었다고 볼 수 있다.

🡺 오버 샘플링을 통해 성능이 경이롭게 상승한 것은 [BOI]의 영향력이 컸음을 알 수 있으며 비만인지 아닌지 예측에 있어 가장 큰 공헌을 하는 변수[BOI]가 오버 샘플링을 하면서 더 강력해 졌음을 알 수 있다.

**■ Logistic Regression Performance Results**

**\*참고: 각 Performance Measure 기재 시 소수점 넷 째 자리에서 반올림함  
1) L1 norm과 L2 norm penalty 적용 시 C값과 penalty 그리드 서치  
2) ElasticNet penalty 적용 시 C값과 l1\_ratio 그리드 서치  
3) Logistic Regression 기법 사용 전 K-Means 기법을 이용한 전처리 적용 여부에 따라 구분  
4) 훈련 데이터셋에 오버샘플링 (업샘플링) 적용 여부에 따라 구분**

**□ CV = 5 적용**

**[ 청소년 전체 데이터 (남/녀 통합) ]**

**- L1 Norm, L2 Norm Penalty 적용**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Accuracy** | **Precision** | **Recall** | **F2** | **AUC** |
| **K-평균 전처리 X**  **오버샘플링 X** | 0.895 | 0.694 | 0.491 | 0.521 | 0.727 |
| **K-평균 전처리 O**  **오버샘플링 X** | 0.855 | Nan | 0.0 | Nan | 0.5 |
| **K-평균 전처리 X**  **오버샘플링 O** | 0.819 | 0.438 | 0.875 | 0.729 | 0.842 |
| **K-평균 전처리 X**  **오버샘플링 O** | 0.5 | Nan | 0.0 | Nan | 0.5 |

표 16 Logistic Regression 청소년 전체 데이터 (CV=5) (1)

\*특이사항: K-평균 전처리 후 모든 인스턴스를 0으로 예측함

**- ElasticNet Penalty 적용**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Accuracy** | **Precision** | **Recall** | **F2** | **AUC** |
| **K-평균 전처리 X**  **오버샘플링 X** | 0.881 | 0.689 | 0.331 | 0.370 | 0.653 |
| **K-평균 전처리 O**  **오버샘플링 X** | 0.855 | Nan | 0.0 | Nan | 0.5 |
| **K-평균 전처리 X**  **오버샘플링 O** | 0.812 | 0.425 | 0.842 | 0.704 | 0.825 |
| **K-평균 전처리 X**  **오버샘플링 O** | 0.5 | Nan | 0.0 | Nan | 0.5 |

표 17 Logistic Regression 청소년 전체 데이터 (CV=5) (2)

\*특이사항: K-평균 전처리 후 모든 인스턴스를 0으로 예측함

**[ 청소년 남자 데이터 ]**

**- L1 Norm, L2 Norm Penalty 적용**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Accuracy** | **Precision** | **Recall** | **F2** | **AUC** |
| **K-평균 전처리 X**  **오버샘플링 X** | 0.860 | 0.688 | 0.547 | 0.570 | 0.742 |
| **K-평균 전처리 O**  **오버샘플링 X** | 0.800 | Nan | 0.0 | Nan | 0.5 |
| **K-평균 전처리 X**  **오버샘플링 O** | 0.832 | 0.546 | 0.942 | 0.822 | 0.873 |
| **K-평균 전처리 X**  **오버샘플링 O** | 0.5 | Nan | 0.0 | Nan | 0.5 |

표 18 Logistic Regression 청소년 남자 데이터 (CV=5) (1)

\*특이사항: K-평균 전처리 후 모든 인스턴스를 0으로 예측함

**- ElasticNet Penalty 적용**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Accuracy** | **Precision** | **Recall** | **F2** | **AUC** |
| **K-평균 전처리 X**  **오버샘플링 X** | 0.840 | 0.660 | 0.409 | 0.443 | 0.678 |
| **K-평균 전처리 O**  **오버샘플링 X** | 0.800 | Nan | 0.0 | Nan | 0.5 |
| **K-평균 전처리 X**  **오버샘플링 O** | 0.824 | 0.534 | 0.920 | 0.804 | 0.860 |
| **K-평균 전처리 X**  **오버샘플링 O** | 0.5 | Nan | 0.0 | Nan | 0.5 |

표 19 Logistic Regression 청소년 남자 데이터 (CV=5) (2)

\*특이사항: K-평균 전처리 후 모든 인스턴스를 0으로 예측함

**□ 여자 데이터**

**- L1 Norm, L2 Norm Penalty 적용**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Accuracy** | **Precision** | **Recall** | **F2** | **AUC** |
| **K-평균 전처리 X**  **오버샘플링 X** | 0.926 | 0.682 | 0.383 | 0.420 | 0.682 |
| **K-평균 전처리 O**  **오버샘플링 X** | 0.907 | Nan | 0.0 | Nan | 0.5 |
| **K-평균 전처리 X**  **오버샘플링 O** | 0.695 | 0.228 | 0.964 | 0.586 | 0.816 |
| **K-평균 전처리 X**  **오버샘플링 O** | 0.5 | Nan | 0.0 | Nan | 0.5 |

표 20 Logistic Regression 청소년 여자 데이터 (CV=5) (1)

\*특이사항: K-평균 전처리 후 모든 인스턴스를 0으로 예측함

**- ElasticNet Penalty 적용**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Accuracy** | **Precision** | **Recall** | **F2** | **AUC** |
| **K-평균 전처리 X**  **오버샘플링 X** | 0.910 | 0.551 | 0.126 | 0.149 | 0.558 |
| **K-평균 전처리 O**  **오버샘플링 X** | 0.907 | Nan | 0.0 | Nan | 0.5 |
| **K-평균 전처리 X**  **오버샘플링 O** | 0.717 | 0.225 | 0.845 | 0.545 | 0.774 |
| **K-평균 전처리 X**  **오버샘플링 O** | 0.5 | Nan | 0.0 | Nan | 0.5 |

표 21 Logistic Regression 청소년 여자 데이터 (CV=5) (2)

\*특이사항: K-평균 전처리 후 모든 인스턴스를 0으로 예측함

\*\*\* K-평균 전처리 적용 시 특이사항  
: CV = 5를 적용한 상황에서, K-평균 전처리 후 로지스틱 회귀분석을 수행한 결과 청소년 전체 데이터, 남자 데이터, 여자 데이터 각각에 대해 오버샘플링 전,후 모든 경우에 각 Instance를 모두 0 (비만이 아님)이라고 예측함

🡪 K-평균 전처리한 모델의 Performance를 보면 Accuracy는 각 데이터별로 전체 Instance 중 비만이 아닌 Instance의 비율과 동일, AUC = 0.5로써 가장 낮은 AUC가 관측됨.

🡪 K-평균 전처리 후 Logistic Regression 적용시 Performance가 올라간 예시와 달리 본 연구에서는 오히려 Logistic Regression Model의 Performance가 현저히 낮아짐.

🡺 CV = 10 적용 시 K-평균 전처리를 하지 않고 Logistic Regression Modeling 수행

**□ CV = 10 적용 (참고: K-평균 전처리를 하지 않고 Logistic Regression Modeling 수행)**

**[ 청소년 전체 데이터 (남/녀 통합) ]**

**- L1 Norm, L2 Norm Penalty 적용**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Accuracy** | **Precision** | **Recall** | **F2** | **AUC** |
| **오버샘플링 X** | 0.895 | 0.694 | 0.491 | 0.521 | 0.727 |
| **오버샘플링 O** | 0.819 | 0.438 | 0.875 | 0.729 | 0.842 |

표 22 Logistic Regression 청소년 전체 데이터 (CV=10) (1)

**- ElasticNet Penalty 적용**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Accuracy** | **Precision** | **Recall** | **F2** | **AUC** |
| **오버샘플링 X** | 0.881 | 0.689 | 0.328 | 0.366 | 0.651 |
| **오버샘플링 O** | 0.814 | 0.428 | 0.843 | 0.706 | 0.826 |

표 23 Logistic Regression 청소년 전체 데이터 (CV=10) (2)

**[ 청소년 남자 데이터 ]**

**- L1 Norm, L2 Norm Penalty 적용**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Accuracy** | **Precision** | **Recall** | **F2** | **AUC** |
| **오버샘플링 X** | 0.860 | 0.688 | 0.547 | 0.570 | 0.742 |
| **오버샘플링 O** | 0.832 | 0.546 | 0.942 | 0.822 | 0.873 |

표 24 Logistic Regression 청소년 남자 데이터 (CV=10) (1)

**- ElasticNet Penalty 적용**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Accuracy** | **Precision** | **Recall** | **F2** | **AUC** |
| **오버샘플링 X** | 0.839 | 0.658 | 0.406 | 0.439 | 0.676 |
| **오버샘플링 O** | 0.810 | 0.534 | 0.921 | 0.805 | 0.860 |

표 25 Logistic Regression 청소년 남자 데이터 (CV=10) (2)

**[ 청소년 여자 데이터 ]**

**- L1 Norm, L2 Norm Penalty 적용**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Accuracy** | **Precision** | **Recall** | **F2** | **AUC** |
| **오버샘플링 X** | 0.926 | 0.682 | 0.383 | 0.420 | 0.682 |
| **오버샘플링 O** | 0.688 | 0.225 | 0.972 | 0.585 | 0.816 |

표 26 Logistic Regression 청소년 여자 데이터 (CV=10) (1)

**- ElasticNet Penalty 적용**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Accuracy** | **Precision** | **Recall** | **F2** | **AUC** |
| **오버샘플링 X** | 0.910 | 0.551 | 0.126 | 0.149 | 0.558 |
| **오버샘플링 O** | 0.717 | 0.226 | 0.845 | 0.545 | 0.775 |

표 27 Logistic Regression 청소년 여자 데이터 (CV=10) (2)

**\* 참고: 아래 데이터 기준 해석, 오버 샘플링 기준 해석, CV 기준 해석에서 기재한 퍼포먼스의 평균치를 계산할 때는 K-평균 전처리를 적용한 경우의 결과값은 제외함. 그 이유는 아래와 같다.**1) K-평균 전처리 적용 후 결과값에는 NaN이 존재하여 평균값 계산이 불가능함.  
2) K-평균 전처리를 적용한 경우의 결과를 데이터 기준, 오버 샘플링 기준, CV 기준 해석에 포함하면 K-평균 전처리 후 퍼포먼스 값이 매우 낮기 때문에 데이터, 오버 샘플링, CV의 효과 해석에 왜곡이 발생함.

**□ 데이터 (남,녀 통합/남자/여자 데이터) 기준 해석**

1) 남녀 통합 청소년 데이터  
: CV = 5, CV = 10를 통틀어 F2 Score가 가장 높은 경우는 F2 Score: 0.729, AUC: 0.842  
 CV = 5, CV = 10 통합 평균 Accuarcy : 0.852, 평균 Recall : 0.635  
 CV = 5, CV = 10 통합 평균 F2 score : 0.581, 평균 AUC: 0.762

2) 남자 청소년 데이터  
: CV = 5, CV = 10를 통틀어 F2 Score가 가장 높은 경우는 F2 Score: 0.822, AUC: 0.873  
 CV = 5, CV = 10 통합 평균 Accuarcy : 0.837, 평균 Recall : 0.704  
 CV = 5, CV = 10 통합 평균 F2 score : 0.659, 평균 AUC: 0.788

3) 여자 청소년 데이터  
: CV = 5, CV = 10를 통틀어 F2 Score가 가장 높은 경우는 F2 Score: 0.586, AUC: 0.816  
 CV = 5, CV = 10 통합 평균 Accuarcy : 0.811, 평균 Recall : 0.581  
 CV = 5, CV = 10 통합 평균 F2 score : 0.425, 평균 AUC: 0.708

: 본 연구에서 가장 관심있는 지표인 F2 Score에 대해 평균을 산출한 결과, ‘남자 청소년’ > ‘남/녀 통합’ > ‘여자 청소년’ 순으로 F2 Score를 보였다. 이는 중간 보고서에서 Recall을 중심으로 산출한 결과와 동일하다.

**□ 오버샘플링 (학습 데이터에 대해서만 동일이 되도록 오버샘플링) 기준 해석**

: 오버 샘플링 결과, Class Imbalance 문제를 갖고 있던 기존 데이터 셋과 달리 오버 샘플링을 통해 Class Imbalance 문제를 해소한 데이터셋에서 눈에 띄는 성능 향상을 보였다. 특히 본 연구의 주 관심 대상인 F2 Score에서 최소 0.16에서 최대 0.36의 성능 향상이 있었다. 그러나 오버 샘플링을 한 결과, 비만인 청소년에 대한 패턴을 잘 학습하여 비만인 청소년에 대해서는 잘 예측하였으나, 비만이 아닌 청소년에 대해 잘 예측하지 못하여 Accuracy와 Precision은 감소함

🡺 Accuracy 및 Precision 측면에서는 손해가 있었으나, 주 관심대상인 F2 Score 및 Recall에서 큰 성능 향상이 있어서 오버 샘플링을 통해 본 연구의 목적을 달성하는데 큰 성능 개선이 있었다고 판단됨

**□ 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** | **F2** | **AUC** |
| **5** | 0.852 | 0.562 | 0.635 | 0.581 | 0.762 |
| **10** | 0.852 | 0.562 | 0.635 | 0.581 | 0.762 |

표 28 Logistic Regression 남녀 통합 데이터 CV기준 해석

**[ 남자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CV** | **Accuracy** | **Precision** | **Recall** | **F2** | **AUC** |
| **5** | 0.839 | 0.607 | 0.705 | 0.660 | 0.788 |
| **10** | 0.835 | 0.607 | 0.704 | 0.659 | 0.788 |

표 29 Logistic Regression 남자 데이터 CV기준 해석

**[ 여자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CV** | **Accuracy** | **Precision** | **Recall** | **F2** | **AUC** |
| **5** | 0.812 | 0.422 | 0.580 | 0.425 | 0.708 |
| **10** | 0.810 | 0.421 | 0.582 | 0.425 | 0.708 |

표 30 Logistic Regression 여자 데이터 CV기준 해석

: CV = 5와 CV = 10을 적용한 결과를 비교해보았을 때 보다 많은 횟수의 교차 검증을 훈련에 사용함으로써 성능이 확연히 상승하는 것은 볼 수 없었으며 다소 비슷한 수치를 보여주었다.

🡺 K 값을 높인 결과 데이터에 따라 성능이 향상된 척도도 있으나, 오히려 감소한 척도도 있다. 또한 가장 많은 Instance를 포함하고 있는 남녀 통합 데이터에서는 CV = 5와 CV = 10의 차이가 없다. 따라서 특정 횟수 이상으로 K값을 높이는 것은 모델의 퍼포먼스 향상에는 큰 영향이 없고, K값을 늘림에 따라 모델링 시간에 더 많은 시간이 소요되므로 무조건 K를 증가시키는 것은 좋지 않다는 것을 확인하였다.

**□ Logistic Regression Penalty 기준 해석**

: Logistic Regression에서는 Cost Function에 L1 norm (Lasso), L2 norm (Ridge), ElasticNet ( L1 norm + L2 norm) 페널티를 적용하여 회귀계수 추정값을 변동을 줄임으로써 모형의 예측 오류를 개선 가능하다. 이를 통해 다중공선성 문제 또는 오버 피팅 문제를 방지할 수 있는 것으로 알려져있다.

본 연구에서는 Python의 sklearn 패키지를 이용해 로지스틱 회귀분석 및 그리드 서치를 수행했다. 이 때, 사용되는 함수 자체의 차이로 인해 1) L1 norm과 L2 norm을 그리드 서치에 포함한 경우와 2) ElasticNet을 사용하고 C, l1 ratio를 그리드 서치한 경우로 나누어 모델을 수립했다.

🡺 그 결과 본 연구에서 수립한 모든 Logistic Regression 모델에서 l1 norm과 l2 norm으로 그리드 서치를 통해 수립된 모델의 퍼포먼스가 ElasticNet을 사용한 경우의 퍼포먼스보다 뛰어났음.

**□ 변수 중요도 확인**

**\* 참고 :**

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

**2) 중간보고서 때 cv에 따른 중요도 차이는 유의미하지 않다고 입증하여 분석에서 제외하였음**

**3) 청소년 전체 데이터, 남자 데이터, 여자 데이터 각각에 대해 가장 높은 F2 Score를 보인 모델을 이용하여 변수 중요도를 확인함**

**[ 청소년 전체 데이터 ]**

- 최적 파라미터: C = 10, Penalty = L1 norm (Lasso)

- 회귀 계수

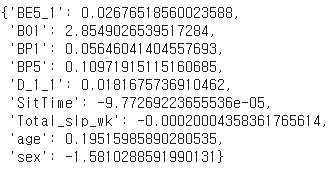


그림 26 청소년 전체 데이터 회귀 계수

회귀 계수 해석 예시

1) 설명변수 ‘age’을 제외한 나머지 변수들이 주어질 때 ‘age’ 값이 한 단위 증가할 때 BMI = 1(비만)일 오즈가 exp(0.195) = 1.215배, 즉 나이가 한 살 증가할 때마다 BMI = 1 (비만)의 오즈가 21.5% 증가한다.

2) 설명변수 ‘sex’를 제외한 나머지 변수들이 주어질 때 오즈에 대한 ‘sex’의 효과는 exp(-1.581) = 0.206배, 즉 여자(sex = 2)인 경우 남자(sex = 1)에 비해 BMI = 1(비만)의 오즈가 약 80% 감소

- 설명 변수 p-value 확인

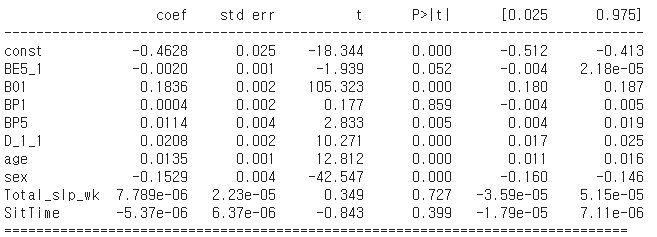


그림 27 청소년 전체 데이터 p-value

- BE5\_1 (1주일간 근력운동 일수), BP1 (평상시 스트레스 인지 정도), Total\_slp\_wk (주중 평균 수면 시간), SitTime (하루 평균 앉아서 보내는 시간 (분))의 p-value가 0.05보다 높아서 남녀 통합 청소년 비만 여부를 예측하는데 유의미하지 않은 요인임을 확인

- P-value가 0.05보다 높은 위 네 가지 변수를 제외하고 회귀 계수 기준 남녀 통합 청소년의 비만 여부에 큰 영향을 끼치는 상위 3개의 요인은 BO1 (주관적 체형 인지), sex (성별), age (나이)

**[ 남자 데이터 ]**

- 최적 파라미터: C = 10, Penalty = L1 norm (Lasso)

- 회귀 계수

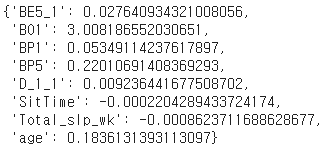


그림 28 청소년 남자 데이터 회귀 계수

- 설명 변수 p-value 확인

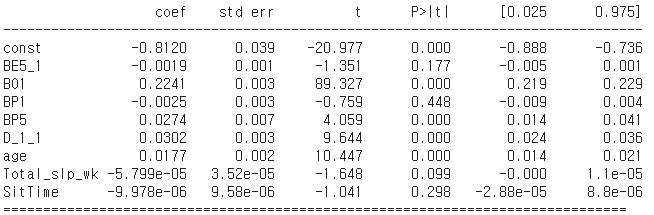
****

그림 29 청소년 남자 데이터 p-value

**-** BE5\_1 (1주일간 근력운동 일수), BP1 (평상시 스트레스 인지 정도), Total\_slp\_wk(주중 평균 수면 시간), SitTime (하루 평균 앉아서 보내는 시간 (분))의 P-value가 0.05보다 높아서 남자 청소년 비만 여부를 예측하는데 유의미하지 않은 요인임을 확인

- P-value가 0.05보다 높은 위 네 가지 변수를 제외하고 회귀 계수 기준 청소년 남자의 비만 여부에 큰 영향을 끼치는 상위 3개의 요인은 BO1 (주관적 체형 인지), BP5 (2주 이상 연속 우울감 여부), age (나이)

**[ 여자 데이터 ]**

- 최적 파라미터: C = 1, Penalty = L1 norm (Lasso)

- 회귀 계수

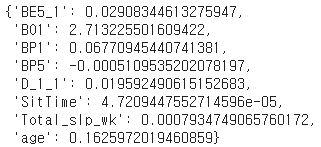


그림 30 청소년 여자 데이터 회귀 계수

- 설명 변수 p-value 확인

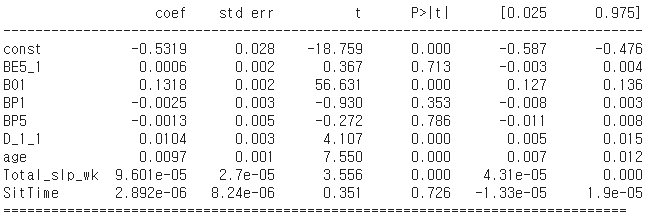


그림 31 청소년 여자 데이터 p-value

: BE5\_1 (1주일간 근력운동 일수), BP1 (평상시 스트레스 인지 정도), BP5 (2주 이상 연속 우울감 여부), SitTime(하루 평균 앉아서 보내는 시간 (분))의 P-value가 0.05보다 높아서 청소년 여자 비만 여부를 예측하는데 유의미하지 않은 요인임을 확인

: P-value가 0.05보다 높은 네 가지 변수를 제외하고 회귀 계수 기준 청소년 여자의 비만 여부에 큰 영향을 끼치는 상위 3개의 요인은 BO1 (주관적 체형 인지), age (나이) D\_1\_1 (주관적 건강상태)

: 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

: 남자의 경우 “2주 이상 연속 우울감 여부”의 회귀 계수는 0.22로써 회귀 계수 기준 변수 중요도가 2위에 해당하지만, 여자의 경우 “2주 이상 연속 우울감 여부”의 회귀 계수는 -0.001로써 비만 여부에 대한 설명력이 매우 낮음.

🡺 “2주 이상 연속 우울감 여부”는 남자 청소년의 비만 여부를 예측하는데 꽤나 중요하지만 여자 청소년의 비만 여부를 예측하는데 있어서 유의한 영향을 끼치지 못함.

이는 중간 보고서의 결과와 동일.

**■ Support vector machine Results**

**\*참고: 각 Performance Measure 기재 시 소수점 넷째 자리에서 반올림함  
1) 선형SVM : C값 그리드 서치  
2) 비선형 SVM : 커널(‘rbf’,’poly’) 사용 시 C값과 gamma 그리드 서치  
3) 훈련 데이터셋에 오버샘플링 (오버샘플링) 적용 여부에 따라 구분**

**(1) 선형SVM ( C 그리드 서치)**

**□ CV = 5 적용**

**[ 청소년 전체 데이터 (남/녀 통합) ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F2 | AUC |
| 원본 | 0.923 | 0.664 | 0.350 | 0.387 | 0.666 |
| 오버샘플링 | 0.691 | 0.227 | 0.968 | 0.586 | 0.816 |

표 31 선형 SVM 청소년 전체 데이터 (CV=5)

**[ 청소년 남자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F2 | AUC |
| 원본 | 0.866 | 0.674 | 0.638 | 0.645 | 0.781 |
| 오버샘플링 | 0.829 | 0.541 | 0.949 | 0.825 | 0.874 |

표 32 선형 SVM 청소년 남자 데이터 (CV=5)

**[ 청소년 여자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F2 | AUC |
| 원본 | 0.923 | 0.664 | 0.350 | 0.387 | 0.666 |
| 오버샘플링 | 0.692 | 0.227 | 0.968 | 0.586 | 0.816 |

표 33 선형 SVM 청소년 여자 데이터 (CV=5)

□ **CV = 10 적용**

**[ 청소년 전체 데이터 (남/녀 통합) ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F2 | AUC |
| 원본 | 0.892 | 0.697 | 0.453 | 0.487 | 0.710 |
| 오버샘플링 | 0.811 | 0.427 | 0.888 | 0.730 | 0.843 |

표 34 선형 SVM 청소년 전체 데이터 (CV=10)

**[ 청소년 남자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F2 | AUC |
| 원본 | 0.867 | 0.693 | 0.599 | 0.616 | 0.766 |
| 오버샘플링 | 0.829 | 0.541 | 0.949 | 0.825 | 0.874 |

표 35 선형 SVM 청소년 남자 데이터 (CV=10)

**[ 청소년 여자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F2 | AUC |
| 원본 | 0.923 | 0.664 | 0.350 | 0.387 | 0.666 |
| 오버샘플링 | 0.692 | 0.227 | 0.968 | 0.586 | 0.816 |

표 36 선형 SVM 청소년 여자 데이터 (CV=10)

**(2) 비선형 SVM ( kernel(‘poly’,’rbf’), C, gamma 그리드 서치 )**

\*참고 : 각 Performance Measure 기재 시 소수점 넷째 자리에서 반올림함

**□ CV = 5 적용**

**[ 청소년 전체 데이터 (남/녀 통합) ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F2 | AUC |
| 원본 | 0.896 | 0.680 | 0.531 | 0.556 | 0.744 |
| 오버샘플링 | 0.841 | 0.476 | 0.938 | 0.785 | 0.881 |

표 37 비선형 SVM 청소년 전체 데이터 (CV=5)

[ 청소년 남자 데이터 ]

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F2 | AUC |
| 원본 | 0.866 | 0.668 | 0.650 | 0.653 | 0.785 |
| 오버샘플링 | 0.855 | 0.583 | 0.956 | 0.848 | 0.893 |

표 38 비선형 SVM 청소년 남자 데이터 (CV=5)

**[ 청소년 여자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F2 | AUC |
| 원본 | 0.923 | 0.670 | 0.340 | 0.377 | 0.662 |
| 오버샘플링 | 0.803 | 0.315 | 0.961 | 0.682 | 0.874 |

표 39 비선형 SVM 청소년 여자 데이터 (CV=5)

□ **CV = 10 적용**

**[ 청소년 전체 데이터 (남/녀 통합) ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F2 | AUC |
| 원본 | 0.896 | 0.680 | 0.531 | 0.55 | 0.744 |
| 오버샘플링 | 0.841 | 0.476 | 0.938 | 0.785 | 0.881 |

표 40 비선형 SVM 청소년 전체 데이터 (CV=10)

**[ 청소년 남자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F2 | AUC |
| 원본 | 0.866 | 0.668 | 0.650 | 0.653 | 0.785 |
| 오버샘플링 | 0.855 | 0.583 | 0.956 | 0.848 | 0.893 |

표 41 비선형 SVM 청소년 남자 데이터 (CV=10)

**[ 청소년 여자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F2 | AUC |
| 원본 | 0.923 | 0.670 | 0.340 | 0.377 | 0.662 |
| 오버샘플링 | 0.803 | 0.315 | 0.961 | 0.682 | 0.874 |

표 42 비선형 SVM 청소년 여자 데이터 (CV=10)

**□ 데이터 (남,녀 통합/남자/여자 데이터) 기준 해석**

**- 선형 SVM 모델**

1) 남녀 통합 청소년 데이터  
: CV = 5, CV = 10를 통틀어 F2 Score가 가장 높은 경우는 F2 Score: 0.73, AUC: 0.843  
 CV = 5, CV = 10 통합 평균 Accuracy : 0.829, 평균 Recall : 0.665  
 CV = 5, CV = 10 통합 평균 F2 score : 0.548, 평균 AUC: 0.759

2) 남자 청소년 데이터  
: CV = 5, CV = 10를 통틀어 F2 Score가 가장 높은 경우는 F2 Score: 0.825, AUC: 0.874  
 CV = 5, CV = 10 통합 평균 Accuracy : 0.848, 평균 Recall : 0.784  
 CV = 5, CV = 10 통합 평균 F2 score : 0.728, 평균 AUC: 0.824

3) 여자 청소년 데이터  
: CV = 5, CV = 10를 통틀어 F2 Score가 가장 높은 경우는 F2 Score: 0.586, AUC: 0.816  
 CV = 5, CV = 10 통합 평균 Accuracy : 0.808, 평균 Recall : 0.659  
 CV = 5, CV = 10 통합 평균 F2 score : 0.487, 평균 AUC: 0.741

: 본 연구에서 가장 관심있는 지표인 F2 Score에 대해 평균을 산출한 결과, ‘남자 청소년’ > ‘남/녀 통합’ > ‘여자 청소년’ 순으로 F2 Score를 보였다. 이는 중간 보고서에서 Recall을 중심으로 산출한 결과와 동일하다.

**- 비선형 SVM 모델**

1) 남녀 통합 청소년 데이터  
: CV = 5, CV = 10를 통틀어 F2 Score가 가장 높은 경우는 F2 Score: 0.785, AUC: 0.881  
 CV = 5, CV = 10 통합 평균 Accuracy : 0.869, 평균 Recall : 0.735  
 CV = 5, CV = 10 통합 평균 F2 score : 0.6705, 평균 AUC: 0.813

2) 남자 청소년 데이터  
: CV = 5, CV = 10를 통틀어 F2 Score가 가장 높은 경우는 F2 Score: 0.848, AUC: 0.893  
 CV = 5, CV = 10 통합 평균 Accuracy : 0.8605, 평균 Recall : 0.803  
 CV = 5, CV = 10 통합 평균 F2 score : 0.751, 평균 AUC: 0.839

3) 여자 청소년 데이터  
: CV = 5, CV = 10를 통틀어 F2 Score가 가장 높은 경우는 F2 Score: 0.682, AUC: 0.874  
 CV = 5, CV = 10 통합 평균 Accuracy : 0.863, 평균 Recall : 0.651  
 CV = 5, CV = 10 통합 평균 F2 score : 0.530, 평균 AUC: 0.768

: 본 연구에서 가장 관심있는 지표인 F2 Score에 대해 평균을 산출한 결과, ‘남자 청소년’ > ‘남/녀 통합’ > ‘여자 청소년’ 순으로 F2 Score를 보였다. 이는 중간 보고서에서 Recall을 중심으로 산출한 결과와 동일하다.

🡺 선형 SVM 모델과 비선형 SVM 모델을 비교한 결과, 비선형 SVM 모델이 3가지 데이터 모두 높은 성능을 보였다. 이는 커널 트릭의 수학적 기교를 적용하여 계산량을 줄여줌과 동시에 좋은 성능을 효과를 얻을 수 있었다.

**□ 오버샘플링 (학습 데이터에 대해서만 동일이 되도록 오버샘플링) 기준 해석**

: 오버 샘플링 결과, Class Imbalance 문제를 갖고 있던 기존 데이터 셋과 달리 오버 샘플링을 통해 Class Imbalance 문제를 해소한 데이터셋에서 눈에 띄는 성능 향상을 보였다. 특히 본 연구의 주 관심 대상인 F2 Score에서 최소 0.18에서 최대 0.305의 성능 향상이 있었다. 그러나 오버 샘플링을 한 결과, 비만인 청소년에 대한 패턴을 잘 학습하여 비만인 청소년에 대해서는 잘 예측하였으나, 비만이 아닌 청소년에 대해 잘 예측하지 못하여 Accuracy와 Precision은 감소함

🡺 Accuracy 및 Precision 측면에서는 손해가 있었으나, 주 관심대상인 F2 Score 및 Recall에서 큰 성능 향상이 있어서 오버 샘플링을 통해 본 연구의 목적을 달성하는데 큰 성능 개선이 있었다고 판단됨

**□ 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** | **F2** | **AUC** |
| **5** | 0.838 | 0.512 | 0.697 | 0.579 | 0.777 |
| **10** | 0.86 | 0.57 | 0.703 | 0.640 | 0.795 |

표 43 SVM 남녀 통합 데이터 CV기준 해석

**[ 남자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CV** | **Accuracy** | **Precision** | **Recall** | **F2** | **AUC** |
| **5** | 0.854 | 0.617 | 0.798 | 0.743 | 0.833 |
| **10** | 0.854 | 0.621 | 0.789 | 0.736 | 0.830 |

표 44 SVM 남자 데이터 CV기준 해석

**[ 여자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **CV** | **Accuracy** | **Precision** | **Recall** | **F2** | **AUC** |
| **5** | 0.835 | 0.469 | 0.655 | 0.508 | 0.755 |
| **10** | 0.835 | 0.469 | 0.655 | 0.508 | 0.755 |

표 45 SVM 여자 데이터 CV기준 해석

: CV = 5와 CV = 10을 적용한 결과를 비교해보았을 때 보다 많은 횟수의 교차 검증을 훈련에 사용함으로써 성능이 확연히 상승하는 것은 볼 수 없었으며 다소 비슷한 수치를 보여주었다.

🡺 K 값을 높인 결과 데이터에 따라 가장 많은 Instance를 포함하고 있는 남녀 통합 데이터에서는 CV = 5와 CV = 10의 차이가 존재하였지만 나머지 데이터는 차이가 없다. 따라서 특정 횟수 이상으로 K값을 높이는 것은 모델의 퍼포먼스 향상에는 큰 영향이 없고, K값을 늘림에 따라 모델링 시간에 더 많은 시간이 소요되므로 무조건 K를 증가시키는 것은 좋지 않다는 것을 확인하였다.

**□ Support vector machine Penalty 기준 해석**

: Support vector machine에서는 파라미터 C를 통해 값이 클수록 모델이 훈련데이터에 과대적합 되는 경향이 생긴다. 또한, 비선형 모델 시, 파라미터 gamma는 값이 클수록 결정 경계가 조금 더 불규칙해지고 각 샘플에 따라 구불구불하게 휘어진다. 반대로 작은 gamma 값은 넓은 범위에 걸쳐 영향을 주므로 경계가 더 부드러워진다.

본 연구에서는 Python의 sklearn 패키지를 이용해 SVM모델링 및 그리드 서치를 수행했다. 이 때, 선형 모델인 경우와 비선형에서의 rbf, poly 커널 트릭을 사용한 모델로 나누어 모델을 수립했다.

🡺 그 결과 본 연구에서 수립한 모든 커널 트릭을 사용한 비선형 모델에서 그리드 서치를 통해 수립된 모델의 퍼포먼스가 선형 SVM 모델을 사용한 경우의 퍼포먼스보다 뛰어났음.

□ 변수 중요도 확인

: 변수 중요도는 선형 SVM에서 계수(coefficient)를 통해 확인 가능하여 앞서 도출한 남녀/남/여 데이터별 최적 파라미터를 적용하여 산출한 변수 시각화는 다음과 같다.

[남녀 통합 데이터]

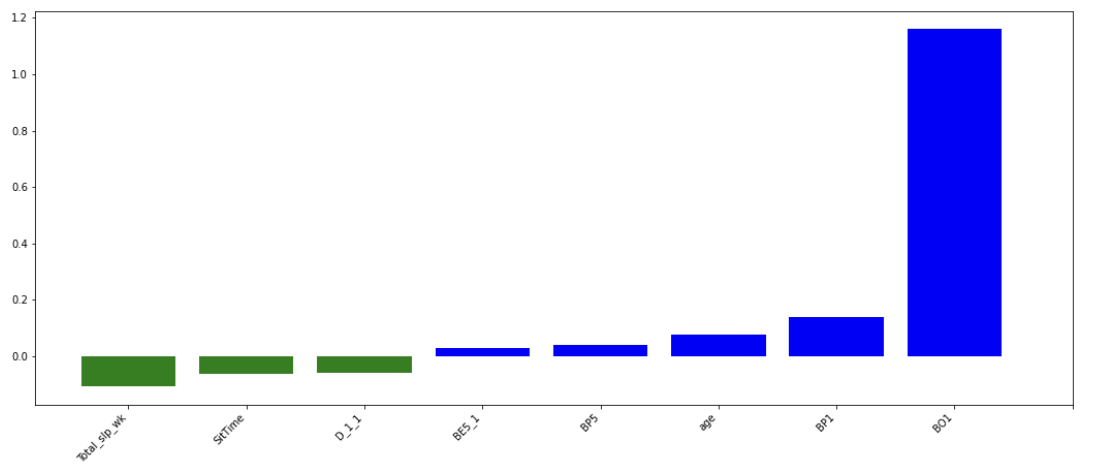


그림 32 SVM 청소년 전체 데이터 변수중요도 시각화

[남자 데이터]

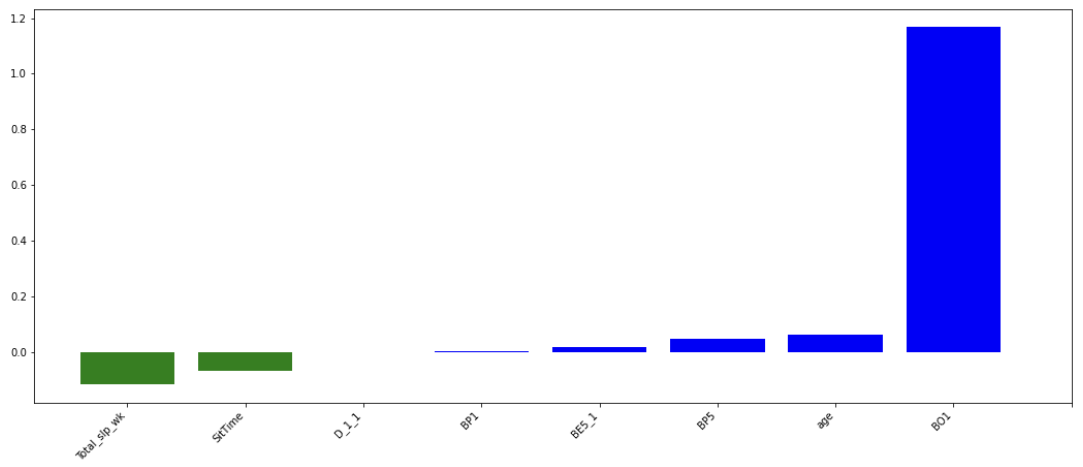


그림 33 SVM 청소년 남자 데이터 변수중요도 시각화

[여자 데이터]

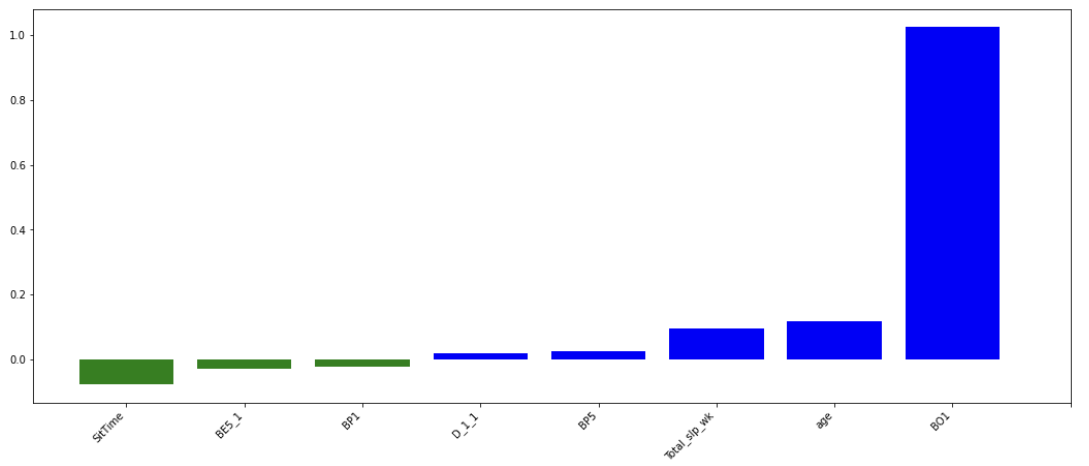


그림 34 SVM 청소년 여자 데이터 변수중요도 시각화

: 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

: “주중 평균 수명 시간”이 여자 청소년의 비만 여부를 예측하는데 양의 계수를 띄지만, 다른 두 데이터에는 음의 계수를 띄며 유의하다고 판단됨

■ MLP Results   
**\*참고: 각 Performance Measure 기재 시 소수점 넷째 자리에서 반올림함  
1) 은닉층 2개를 쌓음 (‘RELU’함수)  
2) 출력층은 ‘Sigmoid’ 함수를 통하여 자살 여부를 분류해줌  
3) 훈련 데이터셋에 오버샘플링 (오버샘플링) 적용 여부에 따라 구분**

**[ 청소년 전체 데이터 (남/녀 통합) ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F2 | AUC |
| 원본 | 0.870 | 0.626 | 0.251 | 0.285 | 0.613 |
| 오버샘플링 | 0.844 | 0.478 | 0.802 | 0.706 | 0.827 |

표 46 MLP 청소년 전체 데이터

**[ 청소년 남자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F2 | AUC |
| 원본 | 0.828 | 0.627 | 0.348 | 0.382 | 0.648 |
| 오버샘플링 | 0.815 | 0.520 | 0.923 | 0.799 | 0.855 |

표 47 MLP 청소년 남자 데이터

**[ 청소년 여자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F2 | AUC |
| 원본 | 0.907 | 0.465 | 0.048 | 0.058 | 0.521 |
| 오버샘플링 | 0.736 | 0.225 | 0.756 | 0.513 | 0.745 |

표 48 MLP 청소년 여자 데이터

: 본 연구에서 가장 관심있는 지표인 F2 Score에 대해 평균을 산출한 결과, ‘남자 청소년’ > ‘남/녀 통합’ > ‘여자 청소년’ 순으로 F2 Score를 보였다. 이는 앞서 머신러닝 기법에서의 결과와 동일하다.

**□ 오버샘플링 (학습 데이터에 대해서만 동일이 되도록 오버샘플링) 기준 해석**: 오버 샘플링 결과, Class Imbalance 문제를 갖고 있던 기존 데이터 셋과 달리 오버 샘플링을 통해 Class Imbalance 문제를 해소한 데이터셋에서 눈에 띄는 성능 향상을 보였다. 특히 본 연구의 주 관심 대상인 F2 Score에서 세가지 데이터셋 모두 0.4 이상의 성능 향상이 있었다. 그러나 오버 샘플링을 한 결과, 비만인 청소년에 대한 패턴을 잘 학습하여 비만인 청소년에 대해서는 잘 예측하였으나, 비만이 아닌 청소년에 대해 잘 예측하지 못하여 Accuracy와 Precision은 감소하였지만 앞에서의 머신러닝 모델보다는 덜하다. 또한, 머신러닝 모델보다 오버샘플링이 큰 효과를 불러옴을 확인할 수 있다.🡺 주 관심대상인 F2 Score 및 Recall에서 큰 성능 향상이 있어서 오버 샘플링을 통해 본 연구의 목적을 달성하는데 큰 성능 개선이 있었다고 판단됨

**[ Discussion ]**

□ Predictive 측면 해석  
: 각 모델에 대하여 F2 Score기준 가장 좋은 성능을 보인 것을 기입하였다.

**[ 남녀 통합 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Method** | **Accuracy** | **Precision** | **Recall** | **F2 Score** | **AUC** |
| **Random Forest** | 0.782 | 0.394 | 0.932 | 0.732 | 0.925 |
| **Logistic Regression** | 0.831 | 0.544 | 0.951 | 0.827 | 0.945 |
| **SVM** | 0.690 | 0.227 | 0.975 | 0.588 | 0.899 |
| **MLP** | 0.844 | 0.478 | 0.802 | 0.706 | 0.827 |

표 49 청소년 통합 데이터 모델 별 최우수 성능

**[ 남자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Method** | **Accuracy** | **Precision** | **Recall** | **F2 Score** | **AUC** |
| **Random Forest** | 0.819 | 0.438 | 0.875 | 0.729 | 0.842 |
| **Logistic Regression** | 0.832 | 0.546 | 0.942 | 0.822 | 0.873 |
| **SVM** | 0.695 | 0.228 | 0.964 | 0.586 | 0.816 |
| **MLP** | 0.815 | 0.520 | 0.923 | 0.799 | 0.855 |

표 50 청소년 남자 데이터 모델 별 최우수 성능

**[ 여자 데이터 ]**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Method** | **Accuracy** | **Precision** | **Recall** | **F2 Score** | **AUC** |
| **Random Forest** | 0.841 | 0.476 | 0.938 | 0.785 | 0.881 |
| **Logistic Regression** | 0.855 | 0.583 | 0.956 | 0.848 | 0.893 |
| **SVM** | 0.803 | 0.315 | 0.961 | 0.682 | 0.874 |
| **MLP** | 0.736 | 0.225 | 0.756 | 0.513 | 0.745 |

표 51 청소년 여자 데이터 모델 별 최우수 성능

- 전체적으로 Logistic Regression모델이 남녀 통합 데이터/남자 데이터/여자 데이터 모두에서 F2 Score을 비교해보면 다른 모델링 기법보다 모델링 우월하며 MLP기법이 가장 좋지 않은 성능을 보임

- 모든 예측자료를 보면 기존 중간 보고서 때와 다르게 여성에서 recall만 가장 높고 그 외 척도에 대해서는 변별력이 없는 수치를 보임. 이는 F2 Score(Precsion과 recall의 가중 합)이 모델 성능 비교에 있어 적절한 척도로 제시될 수 있음을 시사

- 오버 샘플링 후 모든 모델에서 성능이 눈에 띄게 증가함. 이는 데이터가 많아질수록  
일반화 성능이 높아짐을 시사하며 class imbalance문제를 해결하여 기존에 뒤떨어지던 minority class예측에 있어서도 모델의 학습이 수월하게 이뤄짐을 볼 수 있음.

- CV를 보았을 때 랜덤 포레스트 알고리즘을 제외하고는 모두 CV의 효과를 보지 못함. 랜덤 포레스트의 경우에도 CV=10일 때가 효과가 좋긴 했으나, 모델이 학습하는 시간을 고려해 본다면 미미함. 이는 모델의 성능을 높이고자 무작정 CV를 높이는 것은 바람직하지 않다는 것을 제시

- 기존 모델 외에도 PCA, K-mean 전 처리, 커널 SVM을 활용하였으나 PCA와 K-mean전처리에서는 오히려 성능이 하락했음. 커널 SVM 같은 경우는 선형 SVM에 비선형의 특징을 추가하여 분류를 더 수월하게 해내는 것을 성능을 통해 알 수 있었음

□ Descriptive측면 해석  
: 각 모델에 대하여 F2 Score기준 가장 좋은 성능을 보인 것을 기입하였다.

**[ 남/녀 통합 청소년 ]**

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

표 52 청소년 전체 데이터 상위 5개 유의미한 인자

**[ 남자 청소년 ]**

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

표 53 청소년 남자 데이터 상위 5개 유의미한 인자

**[ 여자 청소년 ]**

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

표 54 청소년 여자 데이터 상위 5개 유의미한 인자

* 대다수 항목에서 차이가 났으나, 공통적으로 ‘주관적 체형인지’는 남/녀 통합, 남자 청소년, 여자청소년 모두에서 유의미하게 드러난 것으로 보아 가장 유의미한 인자로 여겨짐
* MLP의 경우 blackbox모델로 변수의 중요도에 대한 해석이 불가하여 descriptive측면에서는 배제하고 해석하였음
* 남/녀 통합 데이터를 보면 ‘주관적 체형 인지’ 이외에 ‘만 나이’가 3개의 모델에서 모두 추출됨. 이들은 세 모델에서 공통적으로 유의미한 인자로 도출됨으로써 청소년 전체로 보았을 때 가장 유의미한 인자라고 해석 가능
* 남자 청소년 데이터를 보면 ‘주관적 체형 인지’ 이외에 ‘만 나이’, ‘2주 이상 연속 우울감 여부’ 또한 3개의 모델에서 모두 추출됨. 이들은 세 모델에서 공통적으로 유의미한 인자로 도출됨으로써 남자 청소년 비만에 유의한 영향을 준다고 해석할 수 있음
* 여자 청소년 데이터를 보면 ‘주관적 체형 인지’ 이외에도 ‘만 나이’, ‘주관적 건강상태’ 또한 3개의 모델에서 모두 추출됨. 이들은 세 모델에서 공통적으로 유의미한 인자로 도출됨으로써 남자 여자 청소년 비만에 유의미한 영향을 준다고 해석 할 수 있음
* 특히 여자 청소년의 경우 로지스틱 회귀분석에서 유의미한 변수가 4개만 추출되었는데 이는 우선적으로 P-value에 의해 통계적으로 유의미하지 않다고 판단했고 더불어 coefficient의 값이 너무 작아 분석가의 판단하에 불필요한 변수라고 판단
* 오버 샘플링 후 변수 중요도를 보았을 때 시각화가 가능했던 랜덤 포레스트와 SVM의 경우 BO1[주관적 체형 인지]를 제외하고 나머지는 모두 극단적으로 0에 수렴하는 것을 볼 수 있었음. 이는 오버 샘플링 후 한 변수의 영향이 오버 샘플링을 함으로써 모델내 중요성이 증가했고 ‘가장’ 유의미한 인자 도출에 있어 추후 사회과학연구에서 활용 가능성 제시

□ 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 3가지 모델의 성능 지표 비교를 통해 최적의 솔루션을 제시할 수 있고 향후 Feature를 더 보완함으로써 F2 Score를 높이는 개선 작업도 기대된다.

**[ Limitation ]**

■ **컴퓨팅 리소스의 한계**

**-**  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.

- svm 모델의 커널 트릭에는 다항식 커널, 가우시안 RBF 커널 이외에 많은 커널이 존재한다. 예를 들어 라플라스 RBF, 시그모이드 커널, 베셀 커널 등이 있다. 그러나 컴퓨터 리소스 및 시간의 제약으로 두 가지 커널(다항식 커널, 가우시안 RBF 커널)에 국한하여 모델링하였다.

**■ 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. 또한 국민건강영양조사 자료와 청소년 건강 행태조사 자료를 함께 분석에 사용했기에 이에 대한 가중치도 적용하는 것이 타당하나, 이 때 적용하는 가중치에 대한 정보가 존재하지 않아 가중치를 적용하지 못하였다.

**■ 결측치 처리 방법의 한계**

: 본 연구에서는 Intermediate Report 연구 결과를 바탕으로 완전 제거법을 적용한 데이터를 사용했다. 분석 대상 데이터에 만약 결측치가 존재한다면, 이는 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 미친다.

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

**■ PCA 결과 해석의 한계**

: 본 연구에서는 PCA를 활용한 차원축소로 feature engineering을 통해 모델의 complexity를 줄임과 동시에 데이터의 설명력을 유지하고자 하였다. 결과적으로 모델의 predictive 측면은 기존 데이터셋과 맞먹을 정도로 충분했으나 1개로 축소된 기존 3개의 변수에 대한 설명력을 잃어 descriptive한 측면은 오히려 하락했다.

: 기존에 있던 변수(age, Total\_slp\_wk, SitTime)의 변수들은 변수 중요도 측면에서 중상위권을 차지하는 변수들로, PCA로 차원축소를 한 결과 기존의 중상위권의 영향력을 유지하였다. 하지만 blackbox형태의 차원축소로 ‘청소년 비만 매커니즘’에 대한 설명에 있어 기존 데이터에 비해 현저히 떨어진다는 한계를 지닌다.

**■ MLP 결과 한계**

: 앞서 머신러닝 모델에서는 변수 중요도를 통해비만 유발 인자를 도출한 바 있다. 그러나, 블랙박스의 특성을 가지는 딥러닝의 한계가 존재하여 변수 중요도를 확인할 수 없었다. 따라서 본 연구는 딥러닝 모델을 통하여 예측하는데 의의를 두었다. 후속 연구에서 설명 가능한 AI 기법이 발전되고 적용된다면 더 큰 의의를 지닐 수 있다는 것으로 사료된다.

**■ 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. Leo Breiman, “RandomForest”, Machine learning, 45, 3~32, 2001
11. Alexander Liu, “Generative Oversampling for Mining Imbalanced Datasets”, IEEE, 42~49, 2001
12. Farrar, Glauber, “Multicollinearity in regression analysis”, The Problem Revisited 105-64, 1964
13. Picuki[웹사이트]. (2020.09.25). URL: <https://www.picuki.com/media/2075086259748794609>
14. Findhealthclinics[웹사이트]. (2020.09.25). URL: <https://www.findhealthclinics.com/US/Los-Angeles/518301005193479/Concierge-Detox-La>
15. 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>
16. PatientsLikeMe[웹사이트]. (2020.10.17). URL: <https://www.patientslikeme.com/>
17. cc.gatech.edu[웹사이트]. (2020.10.18). URL: <https://www.cc.gatech.edu/~jarulraj/courses/8803-f19/slides/13-training-2.pdf>
18. 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>
19. towardsdatascience[웹사이트]. (2020.10.18). URL: <https://towardsdatascience.com/introduction-to-na%C3%AFve-bayes-classifier-fa59e3e24aaf>
20. Wikipedia[웹사이트]. (2020.10.19). URL: <https://en.wikipedia.org/wiki/Random_forest>
21. 임우택 박사의 SPSS 통계 이야기[웹사이트]. (2020.10.19). URL: <https://blog.naver.com/y4769/221851780608>
22. Entheoscientist[웹사이트]. (2020.10.19). URL: <https://nurilee.com/data-science-model-summary-linear-ridge-lasso-elasticnet/>
23. meme[웹사이트]. (2020.10.20). URL: <https://me.me/i/new-example-to-classify-class-a-class-b-k-ec823b9d710e48009c8ff6568c53d676>
24. medium[웹사이트]. (2020.10.21). URL: <https://medium.com/@KCVC/taking-a-stab-at-data-science-day-1-c429b4c802d5>