**Advanced Data Mining Report**

**경희대학교 산업경영공학과**

청소년 비만 예측 및 패턴 도출



고급데이터마이닝

진창호

산업경영공학과 2016100937 김성수  
산업경영공학과 2016100946 김효준  
산업경영공학과 2016101000 홍성희

과목명

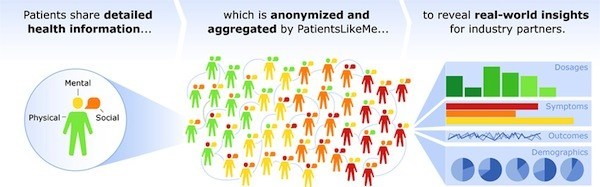
담당교수

팀원

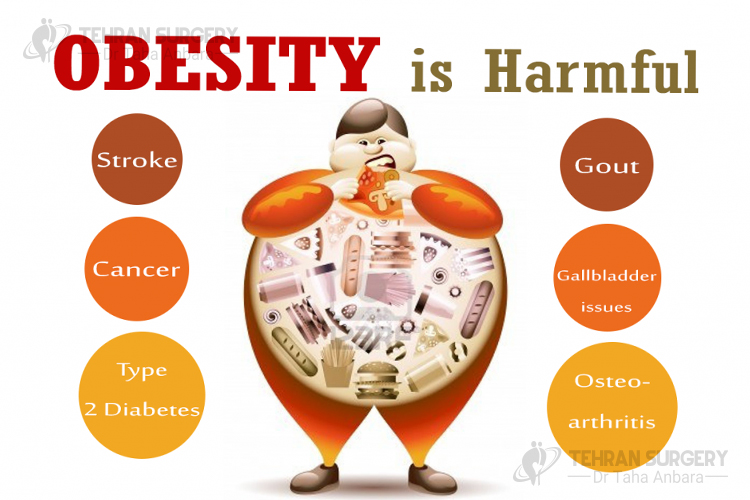
동영상URL

**■ Business Understanding**

**□ Business Problem → Data Science Problem  
-** BP : 청소년 비만에 대한 예측 및 패턴 도출  
**-** DSP : SGD, SVM, Random Forest, Logistic Regression, Naïve Bayes Classifier, KNN Classifier 등 분류기법을 통한 청소년 비만 패턴 파악

**□ 대상 기업 파악**

PatientsLikeMe : “환자들의 페이스북”   
- 1,800개 이상의 질병에 대한 전세계의 200,000명 이상의 환자들이 모인 거대한 SNS 서비스  
- 환자들이 직접 자신의 상태가 어떤지, 증세가 어떻게 진행되고 있는지, 어떤 약을 언제부터   
 먹었고, 효능/부작용은 어땠는지 등을 작성

**□ 연구의 필요성 및 중요성**

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 Park1**,**2000)
4. 청소년기 비만 : 성인까지 이어질 확률이 높음 (Sunhee Park1**,**2000)
5. 심장병, 호흡기 질환 등의 만성질환에 더 큰 노출 가능성 (Sunhee Park1**,**2000)
6. 고혈압, 제2형 당뇨병 등의 다른 심혈관 위험요인의 위험 증가 (Sunhee Park1**,**2000)
7. 사회적 고립, 자살, 식사 장애 등 정신적 질환에 노출 가능성 (Sunhee Park1**,**2000)

**□ 연구 목표 및 기여**  
 국민건강영양조사 자료의 환경요인과 정신 건강 상태 등을 활용하여 청소년 비만 패턴 파악을 통한 본 연구의 기여 사항은 아래와 같음  
1. 설문조사 지표, 비만 예방 교육, 학교 식단 개선 등 청소년의 비만관리를 위한 프로그램 개선  
2. 성인 및 청소년 비만 영향 요인을 연구한 논문은 다수 발표된 바 있으나, 청소년 성별에 따른  
 각 그룹의 비만 영향 요인에 대한 연구는 부족한 실정 (2020.09.23 기준)   
    🡪 로지스틱 회귀분석 및 랜덤 포레스트 기법을 통해 그룹별 비만 영향 요인 파악  
3. PatientsLikeMe에 수립된 모델 제공  
 🡪 국가별 청소년 비만 예방 등 후속 연구에 PatientsLikeMe에 축적된 다량의 데이터를 활용

**■ Data Understanding**

**□ 국민건강영양조사 원시자료에 대한 이해**1) 1998년부터 3년 주기로 기를 나누고 있으며 2016년부터 2018년 자료는 7기에 해당  
2) 크게 건강설문조사, 검진조사, 영양조사로 3가지로 분류  
3) 우리가 사용할 건강설문조사는 조사방법에 따라 가구조사, 건강면접조사, 건강행태조사로 구분  
4) 제 7기 데이터 구조 :   
 \*2016년: (8150, 799) / 2017년: (8127, 857) / 2018년: (8150, 785) ( 년도 (행,열) )

**□ 국민건강영양조사 원시자료의 Strength, Limitation, Benefit, Cost  
- 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  
**- 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  
**- Benefit**  
1) 대대적으로 공개된 공공 데이터  
    : 사용시 별도의 비용이 발생하지 않음, 데이터 수집에 소요되는 시간이 적음.  
**- 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

**■ (Expected) Data Preparation**

**□ 분석 데이터셋 생성**기준: 만 12~18세, 성별 🡪 청소년, 남자 청소년, 여자 청소년으로 3개의 데이터셋 생성

**□ Model 수립에 사용될 Attribute 선정**- 비만 관련 논문 바탕 변수 추출① 만나이 [age] (윤영숙 외 1명, 2009)② 주중 하루 평균 수면시간[Total\_slp\_wk], 주말 하루 평균 수면시간[Total\_slp\_wd] (Sunhee Park1, 2011)③ 주관적 건강 상태[D\_1\_1] (윤영숙 외 1명, 2009)④ 평소 스트레스 인지 정도[BP1] (Sunhee Park1, 2011)⑤ 주관적 체형 인지 [BO1] (NA YOUNG SHIN 외 1명, 2008)⑥ 1주일간 근력운동 일수 [BE5\_1] (Sunhee Park1, 2011)⑦ 평소 하루 앉아서 보내는 시간(시간) [BE8\_1], 평소 하루 앉아서 보내는 시간(분) [BE8\_2] (Sunhee Park1, 2011)⑧ 2주이상 연속 우울감 여부 [BP5] (지영주¸ 2011)

**□ 새로운 변수 생성 및 할당**1) 체질량 지수 [BMI]  
: ‘신장[HE\_ht]’과 ‘체중[HE\_wt]’을 활용하여 파생변수 생성  
: 값이 25 이상이면 1(비만), 25 미만이면 0(비만 아님)으로 범주화 후 Target Attribute로 사용

2) 하루 평균 수면시간 [Oneday\_slp]  
: ‘주중 하루 평균 수면시간[Total\_slp\_wk]’ 과 ‘주말 하루 평균 수면시간[Total\_slp\_wd]’의   
 가중평균을 통해 파생변수 생성

3) 하루 평균 앉아서 보내는 시간 [Oneday\_sit]  
: ‘하루 앉아서 보내는 시간(시간)[BE8\_1]’ 과 ‘평소 하루 앉아서 보내는 시간(분) [BE8\_2]’을 분단위로 환산하여 파생변수 생성

4) Categorical Attribute → Dummy Attribute  
: Random Forest, Logistic Regression 등의 모델링에 적합한 입력 변수인 더미변수로 변환

**□ 결측치 처리**1) 결측치 포함 Data Object 삭제  
2) 결측치가 특정 퍼센트(Ex. 10%, 20%) 이상인 Attribute 삭제  
3) 결측치 대체 (중앙값, 평균값, 최빈값 등)  
🡪 세 가지 방법 모두 수행 후 정보 손실이 더 적은 방법 선택

**□ 탐색적 분석**: 시각화를 이용한 각 변수 기초 통계량 (Ex. 최소값, 4분위수, 표준편차 등) 확인, 이상치 확인 등

**□ 데이터 스케일링**

1. 표준화 : 🡪 모든 데이터 포인트가 동일한 정도의 스케일로 조정  
    (이상치에 민감하다는 단점 존재)
2. 정규화 : 🡪 데이터의 범위를 0과 1로 변환하여 데이터 분포를 조정

**□ 데이터 샘플링: 층화 추출법**: 모집단을 구분하는 2개 이상의 class에 비례하게 ‘단순 무작위 추출방법’으로 표본을 추출

**■ (Expected) Modeling**

**□ Model 수립 시 사용할 방법론**1) SGD Classifier: 확률적 경사 하강법(SGD) 학습을 사용하여 정규화된 선형 모델(Ex. SVM, logistic regression)

2) Support Vector Machine: n차원의 데이터 공간에서 샘플 그룹들을 구분해 내는 최적의 분할선(optimal decision boundary)을 찾아내는 것

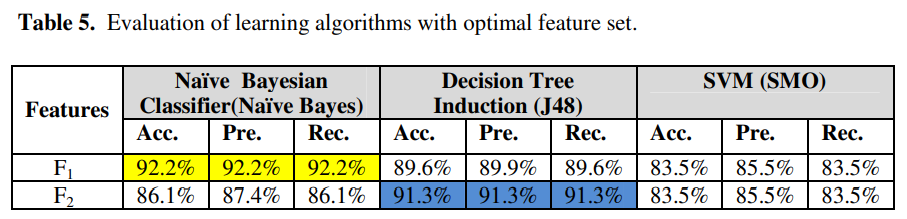
3) Random Forest: 다수의 표본으로 수립된 다수의 의사결정 모형 결과를 종합하는 앙상블 방법 ①종속 변수 값 예측(분류) ②변수의 중요도 확인

4) Logistic regression: 종속변수(범주형)와 독립변수 간의 관계를 구체적인 함수로 나타내어 분류모델에 사용하는 것

5) Naive bayes classifier: 모든 특성들 사이의 독립을 가정하는 베이즈 정리를 적용한 분류방법론

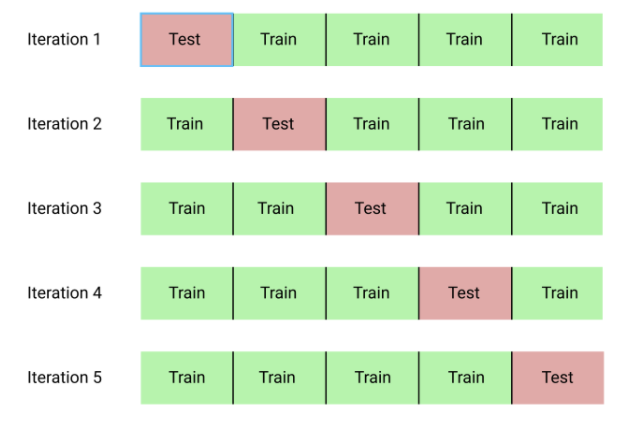
6) KNN: 새로운 데이터가 주어졌을 때 기존 데이터 가운데 가장 가까운 K개 이웃의 정보로 새로운 데이터를 예측하는 방법론

**□ [Expected] Model 성능 비교**



1) 데이터의 특성에 따라Naïve Bayes Classifier와 Decision Tree가 Support Vector Machine에 비해 성능이 좋은 것으로 드러남. (Saiful Islam 외 2명, 2010)  
2) Random Forest의 경우 Accuracy와 AUC측면에서 Logistic regression보다 우수하다고 나타남(Raphael Couronné 외 2명, 2018)  
3) 위의 참고문헌들을 종합해보면 Naïve Bayesian classifier, Decision Tree, Random Forest 기법이 본 연구의 성능을 높일 것으로 예상됨

**□ [Expected] 클래스 불균형 해소 방법론**1) Weight balancing: 학습 데이터에서 loss를 계산 시 특정 클래스에 대해서는 더 큰 loss를 계산  
 Ex) ‘비만’ 클래스에는 더 큰 정확도가 필요하므로 ‘비만’이라는 클래스에 더 큰 loss를 취함  
2) Upsampling & Downsampling: 데이터 상 청소년(남)은 비만율이 20.4%, 청소년(여)는 비만율이   
 11.7%로 불균형  
 ①Upsampling: 데이터가 적은 쪽을 표본으로 더 많이 추출  
 ②Downsampling: 데이터가 많은 쪽을 표본으로 더 적게 추출

**■ (Expected) Evaluation**

**□ 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

**□ Confusion matrix와 AUC 척도를 이용한 평가**① Accuracy(정확도): 전체 예측 중 맞춘 예측의 비율② Precision(정밀도): 모델이 True라고 분류한 것 중 실제 True인 것의 비율③ Recall(재현율): 실제 True인 것 중에서 모델이 True라고 예측한 것의 비율④ : Precision과 Recall의 가중치를 부여한 조화평균⑤ AUC: ROC커브 밑의 면적(1에 가까울수록 좋은 성능)

**□ 분석의 타당성**1) 분석과정의 타당성: [국민건강영양조사] 데이터의 일부를 직접 추출하여 분석 & 분석에 필요한 Attribute를 선정하는데 있어서 논문을 바탕으로 선별  
2) 분석결과의 타당성: 다양한 분류 기법마다 적절히 변형된 Input Data를 통해 도출된 Performance Measure로 높은 신뢰도

**□ 분석한계**1) [국민건강영양조사] 데이터를 기반한 분석은 우리나라 청소년에 국한되어 있으므로 전세계 청소년 대상 적용에 한계 존재  
2) 데이터의 부족으로 모델의 성능 저하

**■ 참고문헌**

1. Eric Stice, Katherine Presnell, and Heather Shaw, "Psychological and Behavioral Risk Factors for Obesity Onset in Adolescent Girls: A Prospective Study", Journal of Consulting and Clinical Psychology Vol.73(2), 2005, pp. 195~202.
2. 윤영숙, 박혜순, "비만과 만성 신장 질환", 대한비만학회지 18-4, 2009, p.127.
3. 윤난희, “비만과 비만관리에서 개인 건강행동과 지역 환경요인의 영향 분석”, 서울대학교 대학원, 2016, pp. 1~4.
4. 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.
5. Md. Saiful Islam, Abdullah Al Mahmud, Md. Rafiqul Islam, "Machine Learning Approaches for Modeling Spammer Behavior", 2010, pp. 11~13.
6. 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.
7. Raphael Couronné, Philipp Probst and Anne-Laure Boulesteix, "Random forest versus logistic regression: a large-scale benchmark experiment", Couronné et al. BMC Bioinformatics , 2018, pp. 9~10.
8. Sunhee Park1, Association Between Short Sleep Duration and Obesity Among South Korean Adolescents, Western Journal of Nursing Research 33(2), 2011, pp. 207–223.
9. Seulggie Choi, MD 외 9명, "Association of Obesity or Weight Change With Coronary Heart Disease Among Young Adults in South Korea", JAMA Internal Med 178(8), 2018, pp. 1065 ~ 1067.