

머신러닝을 활용한 성인 비만 예측 및 중요 변수 분석

Predicting and Analyzing Adult Obesity Using Machine Learning

홍재성¹⁾, 김건우²⁾, 박은지³⁾, 조수연⁴⁾, 정태원⁵⁾

요약

본 연구는 20대 이상의 성인들을 대상으로, 어떠한 변수들이 비만에 영향을 주는지 살펴보고, 해당 변수들을 통해 비만을 줄이기 위한 방향성을 제시하고자 하였다. 이를 위하여 제8기 국민건강영양조사(2019)에서 제공한 결과를 바탕으로, BMI를 측정할 수 있는 성인 6,199명의 설문 조사 데이터를 분석 자료로써 사용하였다. 저체중부터 제3도 비만까지 6개의 집단을 분류하기 위해, 분석은 머신러닝 기반의 모형인 XGBoost, LightGBM, Random Forest를 이용하여 Voting Classifier 모형을 사용하였다. 개별 모형에 대해 Bayesian Optimization을 토대로 하이퍼 파라미터를 튜닝했으며, 최종적으로 테스트 데이터에 대해 약 60% 이상의 정확도를 보였다. Voting Classifier를 구축하는 데 사용했던 3가지 모형에 대해 개별적으로 SHAP 패키지를 적용해 어떠한 변수들이 비만 등급 예측에 있어 중요한 비중을 가졌는지 확인해본 결과, 자기 체형 인식, 혈액 검사, 약력이 있었다. 이는 비만과 당뇨병 사이의 관계로 인한 변수 중요도 등 기타 질병과의 유사성을 머신러닝 분석을 통해 확인할 수 있음을 객관화하였다. 또한 자기 체형 인식과 체중 감량 노력이 주요 변수임을 미루어 보아, 비만은 객관적 지표임에 더불어, 사회적인 영향도 크게 받고 있다는 것을 알 수 있다. 더불어 약력이 중요 변수 중 하나임을 통해 기존 BMI를 기준으로 한 비만이 근육과 지방의 구분을 못 한다는 문제점도 머신러닝을 통해 확인할 수 있었다.

Keywords : 비만, Voting, 머신러닝, XGBoost, LightGBM, Random Forest

-
- 1) 홍재성/ 국민대학교 빅데이터경영통계전공 학부생
 - 2) 김건우/ 국민대학교 빅데이터경영통계전공 학부생
 - 3) 박은지/ 국민대학교 바이오발효융합과 학부생
 - 4) 조수연/ 국민대학교 임산생명공학과 학부생
 - 5) 정태원/ 국민대학교 소프트웨어융합전공 학부생

1. 서론

세계보건기구(WHO : World Health Organization)에서 발표한 2021년 세계영양보고서[1]에 따르면, 세계 인구의 절반은 영양 상태가 불량하며, 48%가 저체중 혹은 과체중, 비만이다. 이를 수치상으로 확인해보면, 성인 22억 명이 현재 과체중 혹은 비만이라는 것을 알 수 있다. 선진국에서는 비만 인구가 수십 년 동안 빠르게 증가하고 있었으며, 국내에서도 현재 비만 인구는 급격한 상승세를 보이고 있다.

위에 대한 자료로 비만 예방과 치료를 주제로 개최된 2021 추계국제학술대회에서, 한국영양학회는 코로나 19로 비만 발병률이 가속화되고 있다고 밝혔다[2]. 4~14세 소아비만은 코로나 19 이전 23.9%에서 31.4%로 증가하였으며, [3]. 보건복지부에서 발표한 제7기 국민건강영양조사 주요 결과[4]를 살펴보면 대한민국 성인(19세 이상)의 비만 유병률은 1998년 26.0%에서 2007년 31.7%, 2016년 34.8%, 2020년에 38.3%[5]로 계속해서 유병률이 증가하는 추세가 나타나고 있음을 밝히고 있다.

이러한 비만은 단순히 외적인 문제로만 볼 것이 아니다. 1997년 세계보건기구는 비만을 질병으로 분류하였으며, [6]. 비만은 당뇨병, 돌연사, 부정맥 등의 심혈관 질환, 폐쇄성 천식, 무호흡증, 암 등에 상관관계를 보인다[7]. 따라서 의학, 교육 및 경제 수준이 발달하였음에도 불구하고, 2009년 옥스퍼드 연구진은 비만으로 인한 인류의 수명은 보통은 3년에서 많게는 10년까지 줄어들 가능성이 크다고 발표하였고, [8]. 2014년 미국 국립보건원에서는 제3도 비만의 기대수명은 14년까지 단축된다고 밝혔다[9].

비만은 이뿐만 아니라 사회·경제적 차원에도 영향을 미치고 있다[10]. 실제 비만으로 인한 대한민국의 사회·경제적 손실은 최근 10년간 약 2배 가까이 증가하였으며, 상대적으로 소득 수준이 낮을수록 고도비만율이 높은 경향을 보이는 등 저소득 계층 및 사회적으로 취약한 계층에서 비만율이 더 높은 것으로 조사되었다[11].

정부는 “국가 비만관리 종합대책(2018~2022)”을 발표하며, 비만에 대한 범정부 차원의 대책을 발표하였다[12]. 대한민국 국민의 비만율을 34.8%로 유지하겠다는 목표였다. 이 외에도 모유 수유 권장, 과음 및 폭식 등을 유발하는 문화 지양, 환경 개선을 위한 가이드라인 구축, 영양표시 의무화, 스포츠강좌이용권 지원 확대, 신체 활동 증진 및 비만 관리에 우수한 기업을 건강 친화 기업 인증제도 등을 통해 정부가 인증하여, 인재 확보 및 건강보험료 감면 등에 대한 여러 방안을 계획하였다.

하지만 2020년 국민건강영양조사에서 발표한 자료와 같이, 해당 정책은 코로나 이후에 비만율이 급증하며 실패하였다. 이에 본 연구는 대한민국 성인의 비만에 영향을 주는 변수들을 머신러닝 기법을 통해 분석하고, 코로나 시기 대한민국 성인의 비만을 예방하기 위한 중재 전략 및 정책을 수립하는 데 방향성을 제공하고자 한다.

2. 이론적 배경

2.1 비만의 정의

비만이란 세계보건기구의 정의에 따라 “지방 조직이 체내에 과잉으로 축적되어 건강에 이상을 초래하는 상태”를 의미한다. 하지만 실제 지방량을 정확히 측정하기는 어려워 비만을 간접적으로 평가한다. 그 방법에는 체질량지수와 허리둘레의 측정, 생체전기저항분석, DXA 등이 있다.

본 연구에서는 대한비만학회에서 세운 지침[13]에서 허리둘레를 기준으로 한 위험성을 배제하고, 카우프 지수로 측정한 BMI 지수만을 사용해 비만을 구분하였다. 즉, 이하 내용에서는 수식 1을 통해 구한 BMI 지수가 18.5 미만일 경우 저체중, 18.5 이상 23 미만일 경우 정상 체중, 23 이상 25 미만일 경우 과체중, 25이상 30미만일 경우 1단계 비만, 30 이상 35 미만일 경우 2단계 비만, 35 이상일 경우 3단계 비만으로 정의한다.

$$\text{체질량지수 (BMI)} = \frac{\text{몸무게 (kg)}}{[\text{신장 (m)}]^2}$$

수식 1 BMI 지수 산식

2.2 머신러닝

머신러닝이란 인공지능의 한 분야로, 통계학적인 방법론을 기반으로 데이터를 이용해 특정 분야에 적합한 기계를 학습시키는 연구 분야를 의미한다.

머신러닝에는 크게 지도학습, 비지도 학습, 강화학습 3가지가 존재한다. 지도학습은 주어진 데이터 X와 목표 레이블 Y를 이용해 값을 예측하는 학습방법이다. 지도학습에는 집값, 주식 종가 등 연속성을 띠는 값을 예측하는 회귀, 성별, 스팸메일, 운동 동작 등을 예측하는 분류, 사용자의 선호도를 예측하는 추천, 순위를 예측하는 랭킹 등이 있다. 대표적인 지도학습 알고리즘은 나이브 베이즈 (naive bayes), 서포트 벡터 머신(Support Vector Machine), K-근접 이웃 모형 (K-nearest neighbors), 의사결정나무 (Decision Tree) 등이 있다.

비지도 학습은 주어진 데이터 X 를 통해 위상 공간상의 데이터 분포를 학습하는 방법이다. 비지도 학습에는 데이터를 분포를 이용해 군집을 만드는 군집화, 위상공간의 분포를 이용해 행렬을 분해하는 저차원 행렬분해 등이 있다. 대표적인 비지도 학습 알고리즘은 K -평균 알고리즘, Topic Modeling, Autoencoder 등이 있다.

강화학습은 주어진 환경 E 속에서, 할 수 있는 행동 A 가 주어졌을 때, 장기적 관점의 보상 R 을 최대화하도록 학습하는 방법이다. 대표적인 예시로는 구글에서 개발한 알파고 가 있다.

본 연구에서는, 6개의 비만 척도를 분류하는 것이 목적이다. 따라서 지도학습의 분류 알고리즘 중, Random Forest Classifier, LighGBM Classifier, XGBoost Classifier 3가지 Tree Model를 적용한 Voting Classifier를 구축하여서 비만을 예측했다.

2.3 선행연구

Galobardes et al.(2000)은 스위스에 거주하는 직장인 약 3,000명에 대해 교육수준과 직업이 BMI에 영향을 끼치는지에 대한 연구를 진행하였다. 연구 결과 남성이 여성보다 과체중이 많으며, 교육수준과 직업이 독립적으로 BMI에 영향을 끼친다는 결론을 얻어냈다[14].

Ball et al.(2003)은 호주에서 1995년에 진행된 호주 건강 및 영양 설문조사에 참여한 8,667명의 성인을 대상으로, 비만과 가족구성원, 고용 상태, 집안 형편, 이민 여부 등 사회적 변수와의 상관관계에 관한 연구를 진행하였다. 연구 결과 건강상의 변수를 통제하였을 때, 사회적 변수가 중요한 변수라는 결론과 사회적 변수와 건강상의 변수는 개별적으로 유의미한 결과를 내지 못한다는 결론을 얻어냈다[15].

Wen et al.(2004)은 대만 북부에 거주하면서 건강검진센터를 방문한 6,318명의 대만인을 대상으로 15가지 질병과 삶의 질에 36가지 설문조사에 대한 자료를 기반으로, 비만을 예측하는 연구를 진행하였다. 연구 결과 비만은 고혈압, 고콜레스테롤혈증, 고중성지방혈증, 제2형 당뇨병, 고혈중, 폐 기능 장애, 지방간 질환, 골관절염의 질병과 일일 운동량, 적당한 활동량의 설문조사 상관관계를 보인다는 결론을 얻어냈다[16].

Lu Qi(2014)는 비만에 대한 다양한 음식과 영양성분의 연관성을 연구를 통해 생유전체학, 대사체학과 같은 “오믹스” 연구에서 개인의 비만 수용성에는 다양한 차이가 존재하며, 맞춤형 영양성분이 필요함을 제언했다[17].

김종임 외(2016)는 대한민국에 거주하는 대학생 289명을 대상으로 BMI, 체질량지

수, 체형 인식, 비만 스트레스, 자아존중감에 따른 체중조절 행동 예측을 진행하였다. 연구 결과 비만 스트레스, 타인평가체형이 체중조절 행동 예측 변수로 나타난다는 결론을 얻어냈다[18].

Trattner et al.(2017)은 미국에 거주하면서 Allrecipes.com을 이용하는 사용자의 9년간의 데이터로, 144,839명에 대해 bookmark와 비만 사이의 연관성을 연구하였다. 연구 결과 지방과 당이 높은 음식을 북마크한 사람들이 비만과 연관이 높다는 결론을 얻어냈다[19].

Batterham et al.(2017)은 93명의 임상 시험 대상자의 인구통계학적인 정보를 이용해 Logistic Regression 모델을 학습시켜 몸무게 감소에 대한 예측을 진행하였다[20].

김은주 외(2020)는 머신러닝을 이용해 한국 비만 클리닉 센터에 방문하는 사용자 25,988명의 1차 방문, 2차 방문, 3차 방문 시의 몸무게 감소 여부를 예측하는 Decision Tree, Random Forest, Logistic Regression, Artificial Neural Network 의 총 4가지 모델을 만들어 5-fold Cross validation으로 성능을 확인하며 예측하였다. 정확도의 측면에서 ANN의 성능이 가장 좋았으며, AUC의 측면에서 Random Forest의 성능이 가장 좋았다는 결론을 얻어냈다[21].

본 연구와 선행연구와의 차별점은 다음과 같다. 첫째, 본 연구는 비만의 6진 분류를 시행한다. 과거의 연구들은 WHO의 비만 척도(BMI 25 이상)를 기준으로 비만인지 비만인지 아닌지, 혹은 몸무게가 증가인지 감소인지에 대한 이진 분류 연구를 시행하였다. 따라서 본 연구는 머신러닝 기반의 모형들을 통해 6진 분류를 한 것에서 차이가 있다.

둘째, 본 연구는 다양한 특징을 이용해 예측한다. 국민건강영양조사는 약 717개의 특징을 제공하고 있다. 사람의 지식이 먼저 반영되어 변수를 선택해서 모형에 사용하기 보다는, 성인에게 주어진 모든 공통적인 질문을 정리하여 모두 사용하였다. 따라서 본 연구는 주관적인 변수 선택보다 객관적인 결과를 목적으로 두어, 기존에 생각하지 못했던 결과를 확인해 볼 수 있다는 점에서 차별점이 있다.

셋째, 본 연구는 제8기 국민건강연구조사를 바탕으로, 키와 몸무게의 정보가 있는 성인 6,199명을 대상으로 연구를 진행한다. 선행연구의 연구 대상자의 수는 적게는 93명부터 많게는 25,988명까지 있으나, 6000명이 넘는 선행연구는 적었다. 따라서 본 연구는 다른 연구들에 비해 연구 대상자의 수가 많다는 점에서 차별점이 있다.

3. 자료의 구성 및 분석모형

3.1 자료의 구성

3.1.1. 데이터의 구성

본 연구는 비만의 등급을 예측하기 위해 보건복지부 산하단체인 질병 관리청에서 매년 실시하는 국민건강영양조사의 원시 자료를 사용하였다. 국민건강영양조사는 1998년 11월부터 12월까지 조사한 제1기 조사부터, 2019년 1월부터 12월까지 조사한 제8기 조사까지의 원시 자료를 제공하고 있다. 본 연구는 이들 중 코로나 19 이후인 2020년과 가장 가까운 사회통계학적 데이터를 담고 있을 것으로 추정되는 2019년의 조사 자료를 사용하였다.

해당 국민건강영양조사 데이터는 1번 기본 조사, 2번 건강설문 조사, 3번 검진 조사, 4번 영양조사의 설문 응답 결과를 사용자 ID 별로 묶어 총 8,110명에 대한 특징 717개를 제공한다.

성인들의 BMI를 기준으로 비만의 등급 분류가 목표이기에, 키 혹은 몸무게를 기재하지 않은 설문과 만 19세 이하의 설문을 제외한 총 6,199명의 데이터를 활용하였다. 이후에 scikit-learn에서 제공하는 train_test_split 모듈에 stratify 인자를 주어 학습용 데이터 4959명, 테스트 데이터 1240명으로 활용하였다.

3.1.2 변수설명

중속 변수는 대한비만협회의 비만 진료지침에 따라 BMI를 기준으로 나눈 6개의 비만 등급이다. BMI와 6개의 등급은 표 1을 따른다.

분류	체질량지수 (kg/m^2)
저체중	< 18.5
정상	18.5 - 22.9
비만 전단계	23 - 24.9
1단계 비만	25 - 29.9
2단계 비만	30 - 34.9
3단계 비만	≥ 35

표 1 체질량지수 기반 비만 분류

본 연구에서 적용한 독립변수는 설문마다 정리하면 표 2와 같다.

표 2의 Data Type에서 Numerical 변수의 NaN 값은 학습용 데이터의 평균으로 채워 주었으며, 이후 분위수를 기반으로한 5개의 그룹을 만들어서 범주형 변수로도 변

환하였다.

표 2의 Data Type에서 Categorical 변수의 NaN 값은 최빈값으로 채웠으며, scikit-learn 패키지에서 지원하는 OneHotEncoder 모듈에 handle_unknown = 'ignore' 인자를 지정하여 변환하였다. handle_unknown 인자를 'ignore' 로 지정할 경우, 학습용 데이터에서는 나타나지 않고 테스트 데이터에서 처음 다루는 변수일 경우, 1이 아닌 0을 채우는 기법이다.

3.1.3 분석 방법

본 연구는 6단계의 비만 등급을 예측하기 위해 (Multi-class classification) 머신러닝 모형인 Random Forest, LightGBM, XGBoost 3가지 모형을 Bayesian Optimization을 이용해 교차 검증 정확도가 높아지는 방향으로 하이퍼 파라미터 튜닝을 진행한다. 이후 이를 바탕으로 튜닝이 완료된 3가지 모형의 분류기를 Scikit-learn에서 제공하는 Voting Classifier 모듈을 사용해 하나로 압축 후 최종 모형으로 사용한다.

전체적인 데이터 분석의 과정은 수집한 데이터의 정제 및 전처리, 탐색적 데이터 분석, 모델링 기법 선택 및 모델링 수행, 검증 및 평가의 과정으로 이루어졌다

3.2 분석모형

3.2.1 Random Forest Classifier

랜덤포레스트는 Breiman(2001)에 의해 처음 소개된 앙상블 기법으로 의사결정 나무의 단점을 개선하기 위한 알고리즘 중 하나이다. 랜덤포레스트는 Bagging과 Bootstrapping을 적용한 모형으로, 모형 내부에서 다양한 의사결정 나무 모형을 (Decision Tree) 새로운 행과 열을 이용해 학습한다. 그리고 이 과정에서 선택되지 않은 데이터를 OOB라고 부르며, OOB 데이터를 분류하는 OOB error를 최소화하는 방향으로 학습한다. 랜덤포레스트는 내부에서 부트 스트래핑 과정을 적용하기 때문에, 변수 중요도를 추정할 수 있다는 장점이 있기에 본 연구의 분석에 활용되었다. 이는 scikit-learn에 내장되어있어 사용할 수 있다.

3.2.2 XGBoost

XGBoost는 기존 Gradient Boosting 기법의 단점을 해소하며 등장한 모형이다. XGBoost는 Gradient Boosting과 동일하게 손실함수를 최소화하는 방향으로 학습한다. 또한, 손실함수를 학습하는 과정에서, 기존 모형이 잘 설명하지 못하는 부분을 보완한다. 이는 xgboost 패키지에

XGBClassifier로 분류기가 내장되어있어 사용할 수 있다.

설문지	변수명	설명	Data Type
기본 변수	region	17개 시도	Categorical
	town_t	동/읍면 구분	Categorical
	apt_t	아파트 구분	Categorical
	sex	성별	Categorical
	age	만나이	Numerical
	incm5	소득 5분위수 (개인)	Categorical
	ho_incm5	소득 5분위수 (가구)	Categorical
	edu	교육수준 재분류 코드	Categorical
가구조사	occip	직업재분류 및 실업/비경제활동 상태 코드	Categorical
	cfam	가구원수	Categorical
	genertn	가구 세대구성코드	Categorical
	allownc	기초생활수급 여부	Categorical
	house	주택소유여부	Categorical
	live_t	주택형태	Categorical
	marri_1	결혼여부	Categorical
	fam_rela	가구주와의 관계	Categorical
건강설문 이환	tins	건강보험종류	Categorical
	D_1_1	주관적 건강상태	Categorical
	D_2_1	최근2주간 몸이 불편했던 경험 유무	Categorical
	DI1_dg	고혈압 의사진단 여부	Categorical
	DI2_dg	이상지질혈증 의사진단 여부	Categorical
	DI3_dg	뇌졸중 의사진단여부	Categorical
	DI4_dg	심근경색증 또는 협심증 의사진단 여부	Categorical
	DI5_dg	심근경색증 의사진단 여부	Categorical
	DI6_dg	협심증 의사진단 여부	Categorical
	DM1_dg	관절염 의사진단 여부	Categorical
	DM2_dg	골관절염 의사진단 여부	Categorical
	DM3_dg	류마티스성 관절염 의사진단 여부	Categorical
	DM4_dg	골다공증 의사진단 여부	Categorical
	DJ2_dg	폐결핵 의사진단 여부	Categorical
	DJ4_dg	천식 의사진단 여부	Categorical
	DE1_dg	당뇨병 의사진단 여부	Categorical
	DE2_dg	갑상선 질환 의사진단 여부	Categorical
	DC1_dg	위암 의사진단 여부	Categorical
	DC2_dg	간암 의사진단 여부	Categorical
	DC3_dg	대장암 의사진단 여부	Categorical
	DC4_dg	유방암 의사진단 여부	Categorical
	DC5_dg	자궁경부암 의사진단 여부	Categorical
	DC6_dg	폐암 의사진단 여부	Categorical
	DC7_dg	갑상선암 의사진단 여부	Categorical
	DC11_dg	기타암1 의사진단 여부	Categorical
	DF2_dg	우울증 의사진단 여부	Categorical
	DL1_dg	아토피피부염 의사진단 여부	Categorical
	DJ8_dg	알레르기비염 의사진단 여부	Categorical
	DJ6_dg	부비동염 의사진단 여부	Categorical
	DH4_dg	중이염 의사진단 여부	Categorical
	DN1_dg	신부전 의사진단 여부	Categorical
	DK8_dg	B형간염 의사진단 여부	Categorical
	DK9_dg	C형간염 의사진단 여부	Categorical
	DK4_dg	간경변증 의사진단 여부	Categorical
의료 이용	M_2_yr	필요 의료서비스 미충족 여부	Categorical

건강검진 및 예방접종	BH9_11	인플루엔자(독감) 예방접종 여부	Categorical
	BH1	(성인) 건강검진 수진여부	Categorical
	BH2_61	(성인) 2년간 암검진 여부	Categorical
	LQ4_00	활동제한 여부	Categorical
	LQ1_sb	최근 1달간 외병여부	Categorical
활동 제한 및 삶의 질	LQ2_ab	최근 1달간 결근여부	Categorical
	LQ_1EQL	EuroQoL : 운동능력	Categorical
	LQ_2EQL	EuroQoL : 자기관리	Categorical
	LQ_3EQL	EuroQoL : 일상활동	Categorical
	LQ_4EQL	EuroQoL : 통증/불편	Categorical
	LQ_5EQL	EuroQoL : 불안/우울	Categorical
	LQ_1HT	HINT-8 : 계단오르기	Categorical
	LQ_2HT	HINT-8 : 통증	Categorical
	LQ_3HT	HINT-8 : 기운	Categorical
	LQ_4HT	HINT-8 : 일하기	Categorical
	LQ_5HT	HINT-8 : 우울	Categorical
	LQ_6HT	HINT-8 : 기억	Categorical
	LQ_7HT	HINT-8 : 잠자기	Categorical
	LQ_8HT	HINT-8 : 행복	Categorical
손상(사고 및 중독)	AC1_yr	1년간 손상 발생여부	Categorical
입원이용	MH1_yr	1년간 입원이용 여부	Categorical
외래이용	MO1_wk	2주간 외래이용 여부	Categorical
교육 및 경제활동	educ	교육수준 - 학력	Categorical
	graduat	교육수준 - 졸업여부	Categorical
	EC1_1	경제활동 상태	Categorical
	EC_lgw_2	최장일자리: 표준직업분류	Categorical
	EC_lgw_4	최장일자리: 종사상지위	Categorical
	EC_pedu_1	유년기환경: 아버지 교육수준	Categorical
	EC_pedu_2	유년기환경: 어머니 교육수준	Categorical
비만 및 체중조절	BO1	주관적 체형인식	Categorical
	BO1_1	1년간 체중 변화 여부	Categorical
	BO2_1	1년간 체중 조절 여부	Categorical
음주	BD1	(만12세이상) 평생음주경험	Categorical
	BD1_11	(만12세이상) 1년간 음주빈도	Categorical
	BD2_1	(만12세이상) 한번에 마시는 음주량	Categorical
	BD7_5	(성인) 1년간 음주문제 상담 여부	Categorical
안전의식	BA2_12	(성인) 운전시 안전벨트 착용여부	Categorical
	BA2_13	자동차 앞좌석 안전벨트 착용 여부	Categorical
	BA2_14	자동차 뒷자리 안전벨트 착용 여부	Categorical
	BA2_2_1	(성인) 1년간 음주운전 여부: 자전거	Categorical
	BA2_2_3	(성인) 1년간 음주운전 여부: 오토바이	Categorical
	BA2_2_5	(성인) 1년간 음주운전 여부: 자동차	Categorical
	BA2_22	(성인) 1년간 음주운전 여부: 동승 횡수	Categorical
정신건강	BP16_1	(만12세이상) 주중(또는 일하는 날) 하루 평균 수면 시간	Categorical
	BP16_2	(만12세이상) 주말(또는 일하지 않는 날, 일하지 않는 전날) 하루 평균 수면 시간	Categorical
	BP1	평소 스트레스 인지 정도	Categorical
	BP5	2주이상 연속 우울감 여부	Categorical
	BP6_10	1년간 자살 생각 여부	Categorical
	BP6_2	1년간 자살 계획 여부	Categorical

	BP6_31	1년간 자살 시도 여부	Categorical
	BP7	1년간 정신문제 상담	Categorical
	BS1_1	(성인) 평생흡연 여부	Categorical
흡연	BS12_37	(성인) 궤련형 전자담배 평생사용여부	Categorical
	BS12_1	(성인) 액상형 전자담배 평생사용여부	Categorical
신체활동	BE3_71	일_고강도 신체활동 여부	Categorical
	BE3_81	일_중강도 신체활동 여부	Categorical
	BE3_91	장소이동 신체활동 여부	Categorical
	BE3_75	여가_고강도 신체활동 여부	Categorical
	BE3_85	여가_중강도 신체활동 여부	Categorical
	BE8_1	평소 하루 앉아서 보내는 시간(시간)	Numerical
	BE3_31	1주일간 걷기 일수	Categorical
	BE5_1	1주일간 근력운동 일수	Categorical
검진기본조사	HE_prg	임신여부	Categorical
	HE_fst	공복시간	Numerical
	HE_fst_5group	HE_fst의 group	Categorical
	HE_HPfh1	고혈압 의사진단 여부(부)	Categorical
	HE_HPfh2	고혈압 의사진단 여부(모)	Categorical
	HE_HPfh3	고혈압 의사진단 여부(형제자매)	Categorical
	HE_HLfh1	고지혈증 의사진단 여부(부)	Categorical
	HE_HLfh2	고지혈증 의사진단 여부(모)	Categorical
	HE_HLfh3	고지혈증 의사진단 여부(형제자매)	Categorical
	HE_IHDfh1	허혈성심장질환 의사진단 여부(부)	Categorical
	HE_IHDfh2	허혈성심장질환 의사진단 여부(모)	Categorical
	HE_IHDfh3	허혈성심장질환 의사진단 여부(형제자매)	Categorical
	HE_STRfh1	뇌졸중 의사진단 여부(부)	Categorical
	HE_STRfh2	뇌졸중 의사진단 여부(모)	Categorical
	HE_STRfh3	뇌졸중 의사진단 여부(형제자매)	Categorical
	HE_DMfh1	당뇨병 의사진단 여부(부)	Categorical
	HE_DMfh2	당뇨병 의사진단 여부(모)	Categorical
	HE_DMfh3	당뇨병 의사진단 여부(형제자매)	Categorical
	HE_THfh1	갑상선질환 의사진단 여부(부)	Categorical
	HE_THfh2	갑상선질환 의사진단 여부(모)	Categorical
	HE_THfh3	갑상선질환 의사진단 여부(형제자매)	Categorical
	HE_HBfh1	B형간염 의사진단 여부(부)	Categorical
	HE_HBfh2	B형간염 의사진단 여부(모)	Categorical
	HE_HBfh3	B형간염 의사진단 여부(형제자매)	Categorical
혈압 측정	HE_rPLS	맥박 규칙성 여부	Categorical
	HE_sbp	최종 수축기 혈압(2,3차 평균)	Numerical
	HE_dbp	최종 이완기 혈압(2,3차 평균)	Numerical
	HE_HP	고혈압 유병여부(19세이상)	Categorical
혈액검사	HE_glu	공복혈당	Numerical
	HE_HbA1c	당화혈색소	Numerical
	HE_insulin	인슐린	Numerical
	HE_HDL_st2	HDL-콜레스테롤(보정값)	Numerical
	HE_TG	중성지방	Numerical
	HE_LDL_drct	LDL-콜레스테롤(직접검사)	Numerical
	HE_HBsAg	B형간염표면항원(수치)	Numerical
	HE_ast	AST(SGOT)	Numerical
	HE_alt	ALT(SGPT)	Numerical
	HE_hcv	C형간염항체(수치)	Numerical
	HE_HB	헤모글로빈	Numerical
	HE_HCT	헤마토크리트	Numerical
	HE_BUN	혈중요소질소	Numerical
	HE_crea	혈중크레아티닌	Numerical
	HE_WBC	백혈구	Numerical
	HE_RBC	적혈구	Numerical
	HE_Bplt	혈소판	Numerical

	HE_Uacid	요산	Numerical
소변검사	HE_Uph	요산도	Numerical
	HE_Unitr	아질산염	Categorical
	HE_Usq	요비중	Numerical
	HE_Upro	요단백	Categorical
	HE_Uglu	요당	Categorical
	HE_Uket	요잠혈요케톤	Categorical
	HE_Ubil	요빌리루빈	Categorical
	HE_Ubld	요잠혈	Categorical
	HE_Uro	유로빌리노겐	Categorical
	HE_Ucrea	요크레아티닌	Numerical
	HE_Una	요나트륨	Numerical
	HE_Ualb	요알부민	Numerical
	HE_Ukal	요칼륨	Numerical
구강검사	OR1	본인인지 구강건강상태	Categorical
	O_pain	최근 1년간 치통 경험유무	Categorical
	O_ortho	교정치료 경험 유무	Categorical
	BM1_0	어제 하루 칫솔질 여부	Categorical
	BM1_1	칫솔질 시기: 아침식사 전	Categorical
	BM1_2	칫솔질 시기: 아침식사 후	Categorical
	BM1_3	칫솔질 시기: 점심식사 전	Categorical
	BM1_4	칫솔질 시기: 점심식사 후	Categorical
	BM1_5	칫솔질 시기: 저녁식사 전	Categorical
	BM1_6	칫솔질 시기: 저녁식사 후	Categorical
	BM1_7	칫솔질 시기: 간식 후	Categorical
	BM1_8	칫솔질 시기: 잠자기 전	Categorical
	BM2_1	사용용품: 치실(Floss)	Categorical
	BM2_3	사용용품: 치간칫솔	Categorical
	BM2_2	사용용품: 양치용액(구강세정액)	Categorical
	BM2_4	사용용품: 전동칫솔	Categorical
	BM2_5	사용용품: 기타(위터픽(물사출기), 허클리너, 점단칫솔(임플란트용칫솔), 틀니관리용품)	Categorical
	BM13	운동 또는 사고로 인한	Categorical
악력검사		치아손상 경험여부	Categorical
	BM7	(성인) 씹기 문제	Categorical
	BM14	치과진료 미치료 여부	Categorical
	GS_use	주요 사용하는 손	Numerical
	GS_mea_r_1	1차 악력측정_오른손	Numerical
	GS_mea_r_2	2차 악력측정_오른손	Numerical
	GS_mea_r_3	3차 악력측정_오른손	Numerical
식생활조사	GS_mea_l_1	1차 악력측정_왼손	Numerical
	GS_mea_l_2	2차 악력측정_왼손	Numerical
	GS_mea_l_3	3차 악력측정_왼손	Numerical
	L_BR	식품섭취조사 1일 전 아침결식 여부	Categorical
	L_LN	식품섭취조사 1일 전 점심결식 여부	Categorical
	L_DN	식품섭취조사 1일 전 저녁결식 여부	Categorical
	L_BR_FQ	최근 1년 동안 1주 동안 아침식사 빈도	Categorical
	L_LN_FQ	최근 1년 동안 1주 동안 점심식사 빈도	Categorical
	L_DN_FQ	최근 1년 동안 1주 동안 저녁식사 빈도	Categorical
	L_BR_TO	최근 1년 동안 아침식사 시 가족 및 가족 외 사람과의 동반 여부	Categorical
	L_LN_TO	최근 1년 동안 점심식사 시 가족 및 가족 외 사람과의 동반 여부	Categorical
	L_DN_TO	최근 1년 동안 저녁식사 시 가족 및 가족 외 사람과의 동반 여부	Categorical
	L_OUT_FQ	외식 횟수	Categorical
	LS_1YR	최근 1년 동안 2주 이상 식이보충제 복용여부	Categorical
	LK_EDU	영양교육 여부	Categorical

	LK_LB_CO	영양표시 인지여부	Categorical
	N_DIET	식사요법여부	Categorical
	N_WAT_C	물섭취량(컵)	Numerical
	N_EN	에너지	Numerical
	N_WATER	수분	Numerical
	N_PROT	단백질	Numerical
	N_FAT	지방	Numerical
	N_SFA	포화지방산	Numerical
	N_MUFA	단일불포화지방산	Numerical
	N_PUFA	다가불포화지방산	Numerical
	N_N3	n-3계 지방산	Numerical
	N_N6	n-6계 지방산	Numerical
	N_CHOL	콜레스테롤	Numerical
	N_CHO	단수화물	Numerical
	N_TDF	식이섬유	Numerical
	N_SUGAR	당	Numerical
	N_CA	칼슘	Numerical
	N_PHOS	인	Numerical
	N_FE	철	Numerical
	N_NA	나트륨	Numerical
	N_K	칼륨	Numerical
	N_VA_RAE	비타민A	Numerical
	N_CAROT	베타카로틴	Numerical
	N_RETIN	레티놀	Numerical
	N_B1	티아민	Numerical
	N_B2	리보플라빈	Numerical
	N_NIAC	나이아신	Numerical

표 2 변수 설명

3.2.3 LightGBM

LightGBM 또한 Decision Tree 모형에 Boosting 알고리즘을 적용한 모형이다. XGBoost와는 다르게, LightGBM은 Leaf-wise 분할을 선택하기에, 학습 시간이 매우 짧은 데 반해 과적합에 강하며 나무의 깊이를 최소화하는 특징이 있다. 이는 lightgbm 패키지에 LGBMClassifier로 분류기가 내장되어있어 사용할 수 있다.

3.2.4 Voting

Voting은 여러 알고리즘을 기반으로 투표를 하는 방법이다. 위에서 정의한 3개의 알고리즘을 이용해 3가지 모형의 개별적인 투표를 바탕으로, 가장 많이 나타난 혹은 평균 예측 확률이 가장 높은 특성치를 선택하는 방법이다. 이는 scikit-learn에 모듈로 내장되어있어 사용할 수 있다.

3.2.5 SHAP

SHAP은 게임이론에 기반해 만들어진 모형 해석 기법의 하나이다. 이는 변수 간의 의존성을 고려하고 클래스 분류에 영향을 끼친 변수에 기여도를 나타낸다. 본 연구에서는 개별 모형에 SHAP를 적용해 변수의 기

여도를 판단하고, 비만에 영향을 준 변수를 확인할 것이다. 이는 shap 패키지로 구현되어 있어 사용할 수 있다.

4. 연구 결과

4.1 비만 예측 모형 설계

본 연구에서는 비만의 6개 등급을 분류하기 위해 사전 설문조사에 약 240개의 변수를 사용하며, 앞에서 설명한 것과 같은 변환을 통해 약 1,400개의 변수를 활용한다. 이 변수 중 각 비만의 등급을 예측하는데 유의미한 변수를 찾기 위해 총 3가지 모형을 활용한다.

본 연구에서는 3가지 모형을 Bayesian Optimization을 활용해 하이퍼 파라미터를 탐색한다. Bayesian Optimization은 모형의 하이퍼 파라미터를 탐색하는 방법이다. 주어진 하이퍼 파라미터의 범주 내에서, N 번의 임의의 파라미터를 이용해서 모형의

모형	하이퍼 파라미터			검증 성능
	명칭	탐색 범주	최종 선택	
Random Forest Classifier	max_depth	10 ~ 50	33	0.5889
	min_samples_split	2 ~ 30	12	
	n_estimators	30 ~ 300	294	
LightGBM Classifier	colsample_bytree	0.4 ~ 1.0	0.5225	0.6158
	learning_rate	0.01 ~ 0.1	0.012	
	max_depth	5 ~ 30	22	
	min_child_samples	20 ~ 40	24	
	n_estimators	30 ~ 300	176	
	subsample	0.4 ~ 1	0.7619	
XGBoost Classifier	colsample_bytree	0.4 ~ 1.0	0.4302	0.6182
	learning_rate	0.01 ~ 0.1	0.0182	
	max_depth	5 ~ 15	6	
	n_estimators	30 ~ 300	128	
	subsample	0.4 ~ 1	0.858	

표 3 모형별 하이퍼파라미터 및 교차검증 성능

성능을 평가하고, 앞서 진행한 N 번의 하이퍼 파라미터 중 모형의 성능이 좋았던 결과들을 바탕으로 그 주변의 하이퍼 파라미터를 Bayesian 통계학을 바탕으로 M 번 탐색하는 방법이다. Bayesian Optimization 과정에서 모형의 평가는, 편향(Underfitting과 과적합(overfitting)의 균형을 맞추기 위해 Stratified K-fold Cross Validation을 활용한 정확도를 바탕으로 사용하였다. K-fold cross validation은 사용할 전체 데이터를 K개의 블록으로 나누고, 한 블록을 테스트 세트(test set)로 나머지는 훈련 세트(train set)가 되도록 하며, 이 과정을 K번 반복하여 모든 데이터가 테스트 데이터로 쓰이는 경우가 존재 하도록 만든 검증 기법이다. 본 연구에서는 각 과정(iteration)을 통해 나온 정확도(accuracy)를 평가지표로 삼았으며 4번의 결과를 평균 내어 나온 값을 모델을 평가하는데 사용하였다.

매개변수 (Hyperparameter)의 선택은 매개 변수를 튜닝할 때 설정해 둔 각 범위에 대하여 탐색을 진행하며, 사전 모델링 정보를 최적 값 탐색에 반영하는 방법인 베이시안 최적화 (Bayesian Optimization)를 이용하였다

하이퍼 파라미터를 탐색 후, 최종 3가지 모형에 대해 Voting Classifier를 구축하여 최종 예측 모형을 만들어 성능을 확인한다. 또한, SHAP를 통해 3가지 각각의 모형에 대해 어떤 변수가 모형 예측에 중요한 변수로 작용하였는지 확인한다.

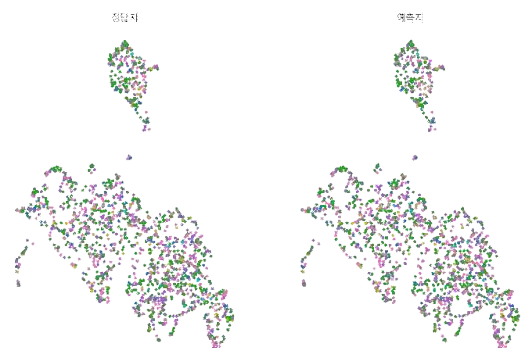
4.2 최종 모형 활용

본 연구에서는 3가지 모형 모두 각각의 하이퍼 파라미터 범주에 대해서, N을 20, M을 20으로 설정하여 발생한 40개의 파라

미터 중, Stratified 4-fold의 평균 정확도가 가장 높은 하이퍼 파라미터를 모형의 최종 하이퍼 파라미터로 사용하였다. 모형별 하이퍼 파라미터의 범주와 최종 하이퍼 파라미터는 표 3과 같다.

해당 최종 모형은 검증 성능을 바탕으로, Random Forest에 가중치 1을, LightGBM에 가중치 2를, XGBoost에 가중치 1.5를 주어 생성했다. Voting Classifier를 Stratified 4-fold를 사용해 테스트 데이터를 예측한 결과, 최종 모형의 성능은 정확도 기준 0.60이다.

모형의 개별 예측결과는 그림 1과 같다. 그림 1은 최종 모형을 토대로 예측한 결과를 Umap을 이용해 2차원에 분포시킨 후 비교한 이미지이다. 그림 1을 살펴보면, 모델이 전반적으로 정상 채종과 제1도 비만을 예측함을 알 수 있다. 또한, 예측결과를 살펴보면 테스트 데이터에 대해 제3도 비만을 예측하지 못한다.



시각화를 통해 모델 성능을 한눈에 파악하기 위한 목적으로 진행한 분석에서는 ROC 커브를 통해 나오는 AUC의 값을 통

해 모델성능을 확인한다. ROC 커브는 AUC의 값에 따라 0.90 - 1.00는 excellent, 0.80 - 0.90는 good, 0.70 - 0.80는 fair, 0.60 - 0.70는 poor, 0.50 - 0.60는 fail로 해석하는데, 본 연구 예측분석에서는 모델별 AUC값이 0.87 (LGBM), 0.86(RF), 0.87(XGB), 0.88(Voting Classifier)로 모두 good 범주에 들었다.

4.3 모형별 중요 변수 확인

4.3.1 Random Forest Classifier

Random Forest Classifier 모형은 테스트 데이터에 대한 정확도가 0.58323이다. Random Forest의 정확도에 영향을 끼친 변수를 SHAP으로 분석한 결과, 그림 1과 같은 결과가 나왔다.

Random Forest Classifier는 주관적 체형을 의미하는 “BO1”, 당뇨병과 관련 있는 “HE_insulin”, 1년간 체중 변화 시도를 의미하는 “BO2_1”이 주요한 변수이다.

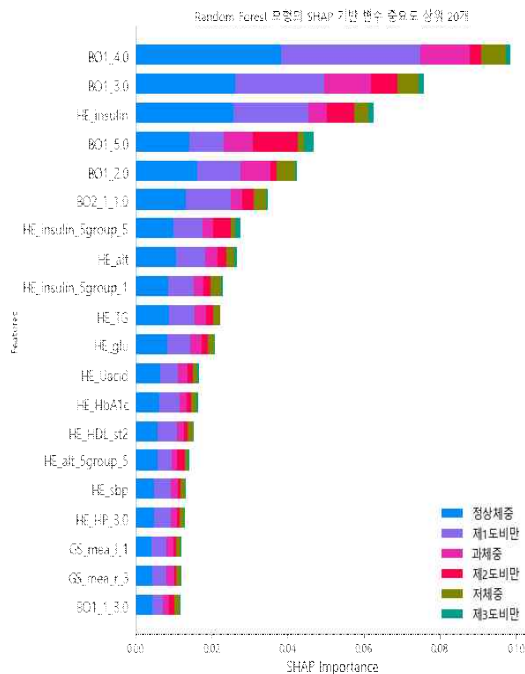


그림 2 Random Forest 모형의 SHAP 기반 변수 중요도 상위 20개

4.3.2 LightGBM Classifier

LightGBM Classifier 모형은 테스트 데이터에 대한 정확도가 0.6064이다. LightGBM Classifier의 정확도에 영향을 끼친 변수를 SHAP으로 분석한 결과, 그림 2와 같은 결과가 나왔다.

LightGBM Classifier는 주관적 체형을 의미하는 “BO1”, 당뇨병과 관련 있는 “HE_insulin”, 1년간 체중 변화 시도를 의미하는 “BO2_1”, 나이를 의미하는 “age”가 주요한 변수이다.

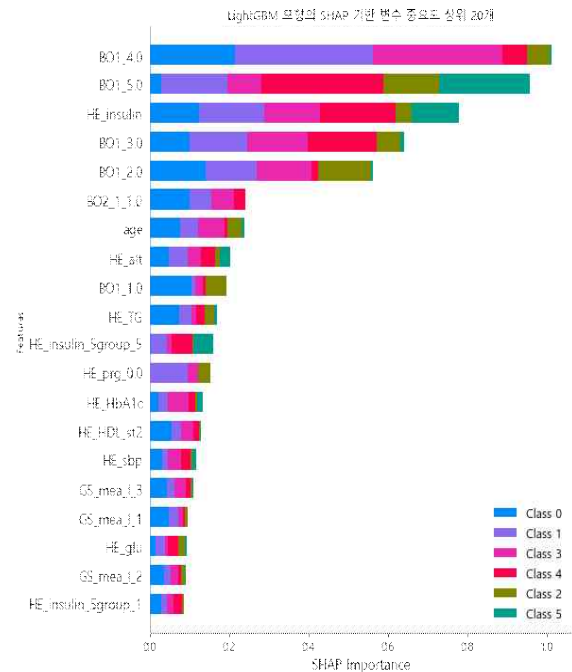


그림 3 LightGBM 모형의 SHAP 기반 변수 중요도 상위 20개

4.3.3 XGBoost Classifier

XGBoost Classifier 모형은 테스트 데이터에 대한 정확도가 0.6013이다. XGBoost Classifier의 정확도에 영향을 끼친 변수를 SHAP으로 분석한 결과, 그림 3과 같은 결과가 나왔다.

XGBoost Classifier는 주관적 체형을 의미하는 “BO1”, 간 검사 수치 중 하나인 ALT(SGPT)를 의미하는 “HE_alt”, 1년간 체중 변화 시도를 의미하는 “BO2_1”, 나이를 의미하는 “age”가 주요한 변수이다.

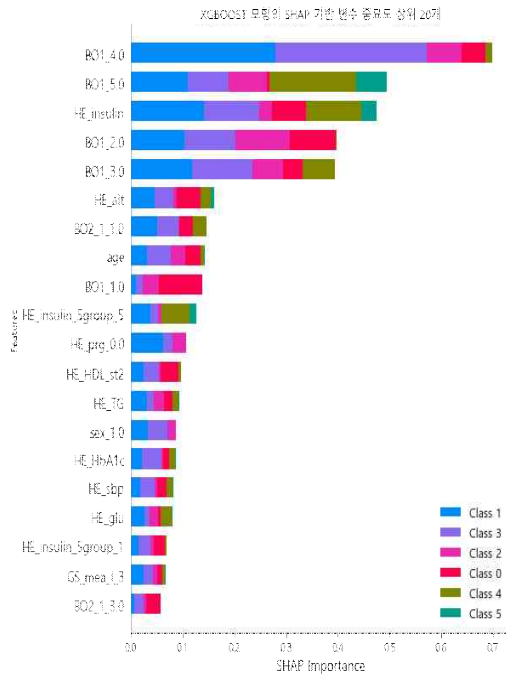


그림 4 XGBoost 모형의 SHAP 기반 변수 중요도 상위 20개

4.3.4 변수 중요도 결과 해석

3가지 모형이 공통적으로 중요히 판단한 변수로는 “BO1”, “BO2_1”, “HE_insulin”이며, 특히 BO시리즈의 경우 정상 체중과 2도 비만을 분류하는데 큰 비중을 가진다. 이는 곧 자신의 체형을 인식하고, 체중 감량을 위해서 노력하는 객관화된 정보와 당뇨병과 관련이 있는 인슐린 분비가 주요 변수로 작용하였음을 알 수 있다. 또한, Boosting 기반의 모형들을 사용한 결과 나이와 간 수치 이상도 주요한 변수임을 알 수 있다. 그 이외로는 혈액 속 중성지방, 공복 혈당, 요산, 당화혈색소 등 당뇨병과 관련이 있는 변수들과, 근력적 수치와 연관되는 악력이 주요 변수임을 알 수 있다.

5. 결론 및 개선 사항

본 연구에서는 제8기 국민영양조사 원시 자료를 바탕으로, 다양한 설문 조사 결과를 통해 BMI 비만을 예측을 (Multi-class classification) 통해, 비만에 영향을 주는 변수에 무엇이 있는지 확인하고자 했다. 이를 위해 전체 데이터 중, 성인의 정보를 사용했으며, 공통으로 응답한 조사의 정보를 Random Forest, LightGBM, XGBoost를 활용하여 만든 Voting Classifier로 예측했

으며, 각각의 모형으로 어떤 변수가 모형 예측에서 중요한 역할을 했는지 알아보았다.

연구 결과는 정확도 60%로, 신뢰도가 크게 높지 않지만, 기존 선행연구들이 진행했던 비만과 기타 질병 사이의 유의미한 상관 관계를 밝히는 모습을 보여준다.

특히 당뇨와 관련이 높은 변수인 인슐린, 혈당 등이 3가지 분류모델 모두에서 중요 변수로 확인되는 점이 위 내용을 뒷받침해주는 근거가 된다.

또한, 개인의 체형 인식과 체중 감량 시도가 3가지 모형에 있어서 중요한 변수임을 토대로 선행연구에서 제안한 바와 같이, 비만이 질병의 개념을 넘어 사회적 의미도 함께 고려해야 함을 확인할 수 있다.

그 외에도 악력이 중요 변수로 확인되는 것은 BMI를 기반으로 한 비만이기때, 지방과 근력 사이의 관계를 정확하게 판단하지 못한다는 문제점을 예측을 통해서도 확인할 수 있다.

이러한 결과를 바탕으로 비만은 질병의 차원에서 다양한 요인들과 관계가 있으며, 다양한 질병의 근원이 되고, 사회적인 문제점을 야기할 수 있음을 알 수 있다. 또한, 비만을 예측하는데 가장 중요한 변수가 자신의 체형을 인식하고 있다는 점에서 출발하며, 체중 감량을 위해 노력한다는 점에서 비만을 예측할 수 있기에 일상생활 속에서 체중 감량을 지원하는 다양한 방안도 고려해볼 필요가 있다.

본 연구에서는 기존 다른 연구들보다는 더 많은 양의 데이터를 활용해 예측을 진행했다. 그럼에도 비만의 등급별 데이터 포인트의 수는 적어, 학습에 어려움이 있었다. 특히 제3도 비만은 하나도 예측하지 못하는 등의 문제점도 보였다. 또한, 조사에 응답 없음 값이 많이 존재했다. 이를 해결하기 위해 과거부터 수집된 모든 설문 조사 결과를 활용할 수 있었으나, 본 연구에서는 코로나 이후의 영향력과 가장 유사한 최신의 데이터만 사용하였기에 해결하기 어려웠다. 이를 해결하기 위해서는 좀 더 세밀화된 조사를 통해 정제된 데이터를 더 많이 수집해 해결할 수 있을 것이다.

6. 참고문헌

- [1] WHO 연간 보고서 (2021), 2021 Global Nutrition Report. Available from: <https://globalnutritionreport.org/reports/2021-global-nutrition-report/>
- [2] 한국영양학회, “코로나19로 가속화된 비만 발병률” 한국영양학회, ‘비만 예방과 치료’ 주제로 2021 추계국제학술대회 개최, NewsWire, 2021년 10월 13일자. Available

- f r o m :
<https://www.newswire.co.kr/newsRead.php?no=932118>
- [3] Kang HM, Jeong DC, Suh BK, Ahn MB. . The Impact of the Coronavirus Disease-2019 Pandemic on Childhood Obesity and Vitamin D Status. J Korean Med Sci. 2021 Jan;36(3):e21. Available from: <https://doi.org/10.3346/jkms.2021.36.e21>
- [4] 질병관리청 국민건강영양조사 주요결과 - 만 성질환 유병률(2021). Available from: https://knhanes.kdca.go.kr/knhanes/sub01/sub01_05.do#none
- [5] 통계청 국가지표체계(2021), 국민삶의질지표 - 비만을. Available from: <https://www.index.go.kr/unify/idx-info.do?idxCd=8021>
- [6] WHO Consultation on Obesity (1997: Geneva, Switzerland), World Health Organization. Division of Noncommunicable Diseases & World Health Organization. Programme of Nutrition, Family and Reproductive Health. (1998). Obesity : preventing and managing the global epidemic : report of a WHO Consultation on Obesity, Geneva, 3-5 June 1997. World Health Organization. Available from: <https://apps.who.int/iris/handle/10665/63854>
- [7] Fruh S. M. (2017). Obesity: Risk factors, complications, and strategies for sustainable long-term weight management. Journal of the American Association of Nurse Practitioners, 29(S1), S3-S14. Available from: <https://doi.org/10.1002/2327-6924.12510>
- [8] University of Oxford (2009), Moderate obesity takes years off life expectancy. Available from: <https://www.ox.ac.uk/news/2009-03-18-moderate-obesity-takes-years-life-expectancy>
- [9] Kitahara CM, Flint AJ, Berrington de Gonzalez A, Bernstein L, Brotzman M, MacInnis RJ, et al. (2014) Association between Class III Obesity (BMI of 40-59 kg/m²) and Mortality: A Pooled Analysis of 20 Prospective Studies. PLoS Med 11(7): e1001673. Available from: <https://doi.org/10.1371/journal.pmed.1001673>
- [10] 변순용. 2020, "비만의 개인윤리적 문제에 대한 연구", 초등도덕교육, pp.79-100. Available from: doi:10.17282/ethics.2020...79
- [11] 국민건강보험공단 (2017), 건강수명 향상을 위한 보험자 비만관리사업 개선방안 연구 (연구 보고서 2017-1-0008). Available from: <http://lib.nhis.or.kr/search/detail/CATXAZ00000037289>
- [12] 대한민국 정책브리핑 (2018), 비만, 더 이상 개인 문제 아냐...국가가 관리한다. Available from: <https://www.korea.kr/news/policyNewsView.do?newsId=148853068>
- [13] 대한비만협회 (2020) 비만진료지침 2020 Available from: http://general.kosso.or.kr/html/user/core/view/reaction/main/kosso/inc/data/Quick_reference_guideline.pdf
- [14] Galobardes, Bruna & Morabia, Alfredo & Bernstein, Martine. (2000). The Differential Effect of Education and Occupation on Body Mass and Overweight in a Sample of Working People of the General Population. Annals of epidemiology. 10. 532-7. 10.1016/S1047-2797(00)00075-2.
- [15] Ball K, Mishra GD, Crawford D. Social factors and obesity: an investigation of the role of health behaviours. Int J Obes Relat Metab Disord. 2003 Mar;27(3):394-403. doi: 10.1038/sj.ijo.0802237. PMID: 12629569.
- [16] Tsai WL, Yang CY, Lin SF, Fang FM. Impact of obesity on medical problems and quality of life in Taiwan. Am J Epidemiol. 2004 Sep 15;160(6):557-65. doi: 10.1093/aje/kwh251. PMID: 15353416.
- [17] Qi L. Personalized nutrition and obesity. Ann Med. 2014;46(5):247-252. doi:10.3109/07853890.2014.891802
- [18] Kim, J.-I. (2016, June 30). Predictors of Weight Control Behavior According to College Students' BMI, Perception of Body Shape, Obesity Stress, and Self-Esteem. Journal of the Korea Academia-Industrial cooperation Society. The Korea Academia-Industrial Cooperation Society. <https://doi.org/10.5762/kais.2016.17.6.438>

- [19] Trattner C, Parra D, Elswiler D. Monitoring obesity prevalence in the United States through bookmarking activities in online food portals. PLoS One. 2017 Jun 21;12(6):e0179144. doi: 10.1371/journal.pone.0179144. PMID: 28636665; PMCID: PMC5479550.

- [20] Batterham, M., Tapsell, L., Charlton, K., O'Shea, J., Thorne, R., Using data mining to predict success in a weight loss trial. JOURNAL OF HUMAN NUTRITION AND DIETETICS; AUG 2017, 30 4, p471-p478, 8p. doi: 10.1111/jhn.12448.

- [21] 김은주, 박영배, 최가혜, 임영우, 옥지명, 노은영, 송태민, 강지훈, 이향숙, 김서영.(2020).한국의 체중 조절 프로그램에 참여한 과체중, 비만 환자에서의 머신러닝 기법을 적용한 체중 감량 예측 연구.대한한의학회지,41(2),58-79.