Journal of the Korean Institute of Industrial Engineers
Vol. 47, No. 6, pp. 504-513, December 2021.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2021.47.6.504

머신러닝 알고리즘을 이용한 중년 남성 건강과 밀접한 라이프로그 분석

김지용[1] · 이지수[2] · 박민서[3†]

[1]광운대학교 수학과 / [2]고려대학교 보건정책관리학부 / [3]서울여자대학교 데이터사이언스학과

Analysis of Lifelong for Health of Middle-Aged Men by Using Machine Learning Algorithm

Jiyong Kim[1] · Jisoo Lee[2] · Minseo Park[3]

[1]Department of Mathematics, Kwangwoon University
[2]Department of Division of Health Policy and Management, Korea University
[3]Department of DataScience, Seoul Women's University

This study aims to find out important factors related to BMI (Body Mass Index) by using machine learning algorithms. BMI is highly related to the health of middle-aged men, such as various chronic diseases. 71 middle-aged men's sleep data, step data, and body weight data were collected from a smartwatch device. Then the data divided into 3 groups by person's height and analyzed by using regression and tree-based machine learning. Moreover, the results were visualized by using explainable AI, SHAP (SHapley Additive exPlanations) to show positive and negative effect of each variable to BMI. In results, the factors have a close relationship with BMI were different in each height group and it shows that considering a method of clustering people into physical characteristics such as height is important to predict an individual's BMI. Further, through results of this study, it is expected to contribute to a personalized health management for each individual.

Keywords: Lifelog, Smartwatch, BMI, Machine Learning, SHAP(Shapley Additive exPlanations)

1. 서 론

라이프로그란 다양한 디지털 센서로부터 수집되는 개인의 경험으로 구성된 통합 디지털 기록을 말하며(Dodge *et al.*, 2007), 여기에는 활동량, 수면 정보, 체중 변화, 체질량, 근육량, 지방량 등이 포함될 수 있다. 라이프로그는 기존에 주로 스마트폰을 통해 측정되었으나, 스마트폰 휴대 시에만 기록이 이루어진다는 점과 스마트폰 내장 센서의 한계로 인해 기록이 불연속적이며 수집 가능한 라이프로그의 종류도 제한적인 경우가 많았다. 그러나, 웨어러블 디바이스가 출시되면서 더욱 정확하고 정밀한 측정이 가능하게 되었으며, 웨어러블 디바이스로 측정 가능한 걸음, 수면, 체중 등의 라이프로그 특성들은 만성질환 발생 및 건강관리에 밀접하게 사용되고 있다(Kim, 2009; Luyster *et al.*, 2012; Zheng *et al.*, 2017). 하지만 현재 라이프로그를 활용한 헬스케어 서비스들은 단순 기록이나 단편적인 통계치를 제시하는 수준에 그치고 있으며, 라이프로그에 따른 운동 및 생활 습관 피드백 기능이 있다 하더라도 이용자별 특성을 고려한 피드백이 아닌 모든 사용자에게 같은 기준의 목표를 제시하고 있다. 따라서, 본 연구에서는 만성질환의 위험이 크고, 건강에 관심도가 높은 중년 남성의 라이프로그 분석을 통해 개개인의 건강관리에 효과적인 인자를 도출하고자 한다. 이를 위해 건강을 대표하는 변수로서 BMI

† 연락저자 : 박민서 교수, 52828 서울특별시 노원구 공릉2동 화랑로 621 서울여자대학교 데이터사이언스학과, Tel : 02-970-7504, E-mail : mpark@swu.ac.kr
2021년 9월 6일 접수; 2021년 10월 14일 수정본 접수; 2021년 11월 6일 게재 확정.

를 종속변수로 설정하고, 수면 및 걸음 데이터로부터 생성된 다양한 라이프로그 변수를 설명변수로 활용하여 BMI를 예측하였다. 또한, SHAP 기법을 활용해 BMI 예측에 가장 크게 영향을 미친 변수를 시각화하였고 이를 개인 건강관리 방안에 활용하고자 한다.

본 연구의 구성은 다음과 같다. 제2장에서는 관련 선행 연구에 대해 기술하고, 제3장에서는 라이프로그 분석에 사용한 머신러닝 알고리즘을 소개한다. 제4장에서는 제안하는 연구 방법을 설명하며, 제5장과 제6장에서는 연구의 결과 및 요약을 기술할 것이다.

2. 선행 연구

2.1 체중과 질병과의 관계

체중은 과체중과 비만을 측정할 수 있는 지표이며, 체중 증가는 2형 당뇨, 관상동맥질환, 고혈압(Hu, 2008), 담석증(Maclure *et al.*, 1989), 몇 가지 암(Song *et al.*, 2015)의 위험을 높이는 것과 관련 있는 것으로 알려져 있다. 미국 여성 92,837명과 남성 25,303명에 대한 코호트조사 결과를 보면, 청년기부터 중년기 사이(여성은 18세부터 55세, 남성은 21세부터 55세)의 체중 변화가 55세 이후 각종 만성질환의 발병과 밀접한 관련이 있음을 알 수 있다(Zheng *et al.*, 2017). 체중이 증가할수록 제2형 당뇨, 고혈압, 심혈관질환, 비만 관련 암, 담석증, 심각한 골관절염, 백내장 등 각종 만성질환의 발생 위험이 커졌다. 특히 제2형 당뇨의 경우 기준집단(체중 2.5kg 감소에서 2.5kg 증가 사이)에 비해 20kg 이상 체중이 증가한 집단에서 여성의 경우 10.51배, 남성의 경우 7.75배 위험이 증가하는 것으로 나타나며 다른 만성질환에 비해 체중 증가에 따른 발병 위험도가 크게 증가하였다.

이에 본 연구에서는 개인의 건강을 나타내는 종속변수를 체중으로 설정하였다. 그러나, 같은 체중이라도 신장 차이에 따라 적정 체중인지가 달라지기 때문에 체중 대신 체질량지수(Body Mass Index, BMI)를 종속변수로써 사용하였다. BMI는 신장과 체중의 비율(BMI = Kg/㎡)을 사용하여 쉽게 계산될 수 있어 체지방량과 비만의 측도로 많이 사용되고 있으며, 선행 연구에서도 단순히 체중이 아니라 BMI를 활용하여 BMI와 만성질환 및 사망률과의 연관성을 확인하였고(Yiengprugsawan *et al.*, 2014; Halim *et al.*, 2019), 중년 남성의 건강 상태에 대해 예측 및 관리할 수 있는 지표로서 활용하였다(Wang *et al.*, 2016; Stenholm *et al.*, 2017).

2.2 수면, 걷기와 BMI와의 관계

통계적 방법론을 활용하여 수면 또는 걷기와 같은 라이프로그 데이터와 BMI의 상관관계 분석한 연구로서, 소아 및 청소년의 BMI Z점수와 수면 시간, 수면 효율성, 그리고 렘수면 비율 간의 유의미한 관련성을 도출한 연구(Liu *et al.*, 2008), 대한민국 성인 수면 시간의 증가와 BMI의 감소와의 연관성 및 수면 시간과 비만과의 관련성을 도출한 연구(Park *et al.*, 2007), 중년 여성 대상 8주간 걷기 운동 프로그램을 추적해 비만도별 걷기 운동이 체중, 근육량, BMI에 긍정적인 변화를 주는 것을 확인한 연구(Shin *et al.*, 2016), 그리고 BMI에 대해 6분 걷기 테스트(6MWT)가 유의미한 효과를 보이는 것을 확인한 연구(Donini *et al.*, 2013; Correia *et al.*, 2015) 등이 있다.

2.3 머신러닝을 활용한 BMI 예측

머신러닝을 활용해 BMI에 영향을 주는 변수 추출 및 모델 평가 관련 연구로서는, 소아 비만을 조기 발견할 수 있게 위험요인을 추출하는 영국의 코호트 연구(Singh *et al.*, 2020), 한방 비만 프로그램에 참여한 과체중 및 비만 성인 환자들의 체중 감량 예측 및 주요 변수 추출 관련 연구(Kim *et al.*, 2020), 인도네시아 성인 비만의 위험 요소들을 탐색하고 이를 기반으로 비만 상태를 예측하는 연구(Thamrin *et al.*, 2021), 심리적 변수를 설명변수로 사용하여 BMI 값과 BMI 상태를 예측하는 연구(Delnevo *et al.*, 2021) 등이 있다.

3. 라이프로그 분석에 사용한 알고리즘

3.1 Ridge Regression

회귀분석에서 모형에 입력할 독립변수들 사이에 다중공선성(multicollinearity)이 존재하는 경우 회귀계수들의 분산이 매우 커지게 되어 추정량으로서의 정도가 나쁘게 된다(Kim, 2010). 이러한 경우 작은 편의(bias)를 허용하여 회귀계수의 크기를 축소함으로써 회귀계수들의 분산을 작게 할 수 있는데, 릿지 회귀 모형(Ridge Regression)은 회귀계수의 크기에 패널티(penalty)를 부여함으로써 회귀계수를 축소하는 방법이다(Hoerl *et al.*, 1970).

3.2 Support Vector Regression(SVR)

SVR의 손실 함수는 회귀계수 크기를 작게 하여 실제값과 예측값의 차이가 작은 회귀선을 찾는 것으로, 데이터에 노이즈가 있다고 가정하여 노이즈가 있는 실제 값을 완벽히 추정하는 것을 목표로 하지 않아, 적정 범위(2ε) 내에서의 실제값과 예측값의 차이를 허용하는 방식으로 계산한다(Drucker *et al.*, 1997; Basak *et al.*, 2007).

3.3 Extreme Gradient Boosting: XGBoost

그래디언트 부스팅 머신(GBM) 알고리즘은 성능이 약한 모델(weak simple model) 여러 개를 손실 함수 상에서 경사 하강

법 최적화 방식을 통해 기울기가 가장 큰 방향으로 점진적(greedy stagewise) 추가하는 앙상블 방법론 중 부스팅 계열에 속하는 알고리즘이며 회귀분석 또는 분류 분석을 수행할 수 있다(Natekin and Knoll, 2013). XGBoost는 그래디언트 부스팅 머신 중 하나로 지도 학습에 활용될 수 있다(Chen and Guestrin, 2016). 또한 GPU 병렬 처리가 가능하여 GBM 대비 빠르게 수행할 수 있는 장점이 있어 많이 활용되고 있다(Mitchell et al, 2018).

3.4 Categorical Boosting: CatBoost

CatBoost 알고리즘은 모든 종류의 그래디언트 부스팅 머신에 존재하는 목표 누수(target leakage)로 인한 예측 변화(prediction shift) 문제와 높은 카디널리티(high cardinality)를 가진 범주형 변수에 대한 전처리 문제를 해결하기 위해 순서형 부스팅(Ordered Boosting) 알고리즘 활용하였다(Prokhorenkova *et al.*, 2017). 순서형 부스팅은 ordering principle이라는 핵심 개념으로 이뤄졌는데, 이는 서로 비교할 수 없는 범주를 가진 범주형 변수들에 대해 높은 카디널리티를 가졌을 때 발생하는 이슈를 각 범주의 예상 목표값을 추정하는 목표 통계량(target statistics)별로 범주를 그룹화하는 방식에 그치지 않고, 학습 데이터에 무작위 순열(Random Permutation)을 적용해 인위적인 시간이라는 개념을 도입하여 효과적으로 과적합을 방지하는 방식이다.

3.5 XAI(eXplainable Artifical AI): SHAP(SHapley Additional exPlanations)

설명 가능한 인공지능(XAI)은 사람들에게 이해 가능한 쉬운 언어로 예측된 결과에 대해 직관적인 설명을 제공하고, 알고리즘 동작을 정확하게 묘사하여 동작에 대한 예측을 할 수 있게 한다 (Gilpin *et al.*, 2018). SHAP 기법은 게임 이론을 따르는 새플리 값(Shapley Value)을 기반으로 예측에 대한 각 변수의 기여도를 계산하여 예측값을 설명하는 기법이다 (Lundberg and Lee, 2017). 새플리 값은 여러 변수의 조합을 구성한 뒤 중요도를 확인하려는 변수의 유무에 따른 평균 변화를 통해 해당 변수의 변수 중요도(feature importance)를 계산하여 얻을 수 있다(Fryer *et al.*, 2021). SHAP 기법은 종속변수에 대한 해당 독립변수의 양(+)/음(-) 영향력을 확인할 수 있다는 장점이 있다.

4. 연구 방법

4.1 데이터 수집

본 연구에서는 2021년 2월부터 2021년 8월까지의 기간동안 ㈜지아이비타 앱(Vitameans)과 삼성 갤럭시 워치(Samsung Galaxy watch active2)를 통해 수집한 35세부터 59세까지의 중년 남성 사용자들의 수면, 걸음 및 체중 데이터를 사용하였다. 데이터 셋은 위의 기간에 수집된 사용자의 취침 시각, 기상 시각, 분 및 일별 걸음 수, 분 및 일별 걸은 거리, 분 및 일별 순간 속도 그리고 일별 체중 정보를 담고 있으며, 이는 일별 수면 데이터, 분별 수면 데이터, 일별 걸음 데이터, 분별 걸음 데이터, 그리고 일별 체중 데이터 각각 6,223개, 241,068개, 6,380개, 1,797,590개, 그리고 6,729개 행으로 구성되어 있다.

4.2 데이터 전처리

개인별 BMI의 변동성을 설명하는 최적의 변수를 찾기 위해 삼성 갤럭시 워치를 통해 수집된 일별 수면 데이터와 분별 수면 데이터를 이용해 일별 수면 총시간, 수면 효율성 등 여러 가지 파생변수를 생성하였다. 마찬가지로 일별 및 분별 걸음 데이터를 이용해 일별 총 걸음 수, 아침 시간대 평균 걸음 속도 등의 파생변수를 생성하였다. 또한, 사용자의 신장 등의 신체 데이터와 일별 체중 데이터를 걸음 및 수면 데이터와 측정날짜를 기준으로 통합하였다. 생성된 파생변수 및 사용한 변수는 <Table 1>과 같다.

4.3 변수 선택

본 연구에서 생성한 파생변수 55개를 모두 모델에 입력할 경우, 모델이 복잡해져 과적합의 우려가 있으므로(Ying, 2019) 중복변수를 제거하는 관련 Feature Selection을 수행 후 모델에 변수 입력을 진행하였다. Feature Selection을 위해 Wrapper 방식을 사용하였으며 그중에서도 VIF Stepwise Selection(Van Breugel *et al.*, 2016; Akinwande *et al.*, 2015; Głąb *et al.*, 2015)을 사용했다. 본 논문에서 사용된 예측 모델은 변수 간 다중공선성에 큰 영향을 받지 않으나 <Table 1>에서 사용된 55개의 입력변수 모두 사용은 과적합 우려 문제 외에도, 비슷한 해석 또는 역할을 하는 변수들이 중복되게 상위 주요 변수를 차지하는 이슈가 발생하였다. 따라서 BMI에 대한 해석에 있어, 변수 간의 강한 상관관계로 비슷한 역할을 하는 변수를 일부 제거하기 위해 VIF(Variance Inflation Factor) 값이 큰 변수를 찾아 제거하기로 하였다. VIF > 15일 경우, 해당 변수는 설명변수에서 빠져도 나머지 변수들이 종속변수를 90% 이상 설명할 수 있기에, 각 변수에 대한 VIF를 확인하여 15 이상의 높은 다중공선성을 가진 변수들을 차례로 제외하였다. 이때, 일반적으로 VIF 기준값을 10으로 선택하나(Vittinghoff *et al.*, 2012), 본 연구에서 VIF 값의 기준을 15로 택한 이유는 선행 연구에서(Ekelund *et al.*, 2015) '연속해서 20분 이상 걸은 시간'과 '횟수'가 BMI 및 건강에 중요한 변수로 작용한 것을 참고하여 이 두 변수를 제거하지 않고 유지하기 위함이다. <Table 2>는 15 이상의 VIF 값을 가진 변수를 제외하고 남은 변수들의 종류이다.

Table 1. Derived Variables and Description

Derived Variables	Description	Derived Variables	Description
DATE	Date of data	MUSCLE	User's muscle mass measured by smart scale
USER_CODE	User's unique code	WEIGHT	User's weight measured by smart scale
STEP_COUNT	Total steps per day	BMI	User's BMI measured by smart scale
DISTANCE	Total distance walked per day	FAT	User's amount of fat measured by smart scale
CALORIE	Calories burned per day	BMI_INDEX	Categories according to BMI values
AVG_SPD	Average daily walking speed	BMI_STATUS	Description of status for BMI_INDEX
REAL_SUM_WALK_TIME	Total time walked per day based on distance walked	BED_TIME	User's bed time
SUM_WALK_TIME	Total steps per day according to Samsung Galaxy collection method	TOTAL_SLEEP_TIME_HOUR	Total sleep time of users per day
STEP_STD	Deviation of daily step	TOTAL_SLEEP_TIME_VARIABILITY	Variance in total sleep time of users per day
DIST_STD	Deviation of distance walked per day	SLEEP_EFFICIENCY	The ratio of sleep efficiency to the user's total sleep
SPD_STD	Deviation of daily walking speed	DEEP_SLEEP_RATE	The ratio of deep sleep to the user's total sleep
MOR_AVG_SPD	Average walking speed in the morning time period(7 am to 10 am) per day	REM_SLEEP_RATE	The ratio of REM sleep to user's total sleep
MOR_REAL_WALK_TIME	Total walking time based on distance walked for each morning time zone	NAP_COUNT	Number of naps per day by user
MOR_WALK_TIME	Total walking time calculated by Galaxy based on each morning time zone	BED_TIME_VARIANCE	User's bedtime variation from the previous day
MOR_WALK_DIST	Distance walked by users in the morning hours per day	BED_TIME_VARIANCE_FLAG	A binary flag indicating whether the change in bedtime from the previous day is within 2 hours or not
TOTAL_TIME_CONTINUOUS_WALK_20MINUTES	Total time of continous walked for more than 20 minutes	BED_TIME_AT_10PM_TO_12PM_FLAG	A binary flag indicating whether bedtime is between 10pm and 12 midnight or not
TOTAL_COUNT_CONTINUOUS_WALK_20MINUTES	Number of continuous walks for more than 20 minutes	GENDER	User's gender
LNC_AVG_SPD	Average walking speed during lunch time period(10 am to 1 pm) per day	AGE	User's age
LNC_REAL_WALK_TIME	Distance-based total steps walked by users during lunch hour per day	AGE_CATEGORY_10	Age category in 10 units
LNC_WALK_TIME	Total walking time calculated by Galaxy based on each lunch time zone	AGE_CATEGORY_5	Age category in 5 units
LNC_WALK_DIST	Distance walked by users in the lunch hours per day	HEIGHT	User's height
AFT_AVG_SPD	Average walking speed of users during the evening hours(5pm to 8pm) per day	HEIGHT_CATEGORIZE_10	Height category(in units of 10 cm)
AFT_REAL_WALK_TIME	Distance-based total steps walked by users during evening hour per day	HEIGHT_CATEGORIZE_5	Height category(in units of 5 cm)
AFT_WALK_TIME	Total number of steps calculated based on the Galaxy Watch in each evening time zone	WEEKDAY	Day of the week
AFT_WALK_DIST	Distance walked by users in the evening per day	WEEKEND	A binary flag indicating whether it is a weekend or not
NT_AVG_SPD	Average walking speed of users during the night time(9pm to 12 midnight) per day	HOLIDAY	A binary flag indicating whether it is a holiday or not
NT_REAL_WALK_TIME	Distance-based total steps walked by users during night hour per day	NT_WALK_TIME	Total number of steps calculated based on the Galaxy Watch in each night time zone
NT_WALK_DIST	Distance walked by users in the night time per day		

Table 2. Variables without High Multicollinearity

Variable	VIF	Variable	VIF	Variable	VIF	Variable	VIF
Intercept	0.000000	NT_AVG_SPD	1.234312	MOR_WALK_DIST	2.996044	TOTAL_SLEEP_TIME_VARIABILITY	5.547898
GENDER	0.000000	BED_TIME_AT_10PM_TO_12PM_FLAG	1.253070	MOR_WALK_TIME	3.832779	SUM_WALK_TIME	5.601399
HOLIDAY	1.037140	STEP_COUNT	1.284651	LNC_WALK_DIST	3.849704	AGE_CATEGORY_10	5.686721
WEEKEND	1.072393	REM_SLEEP_RATE	1.333314	LNC_WALK_TIME	4.091017	AGE	5.778830
DEEP_SLEEP_RATE	1.077002	LNC_AVG_SPD	1.335287	NT_WALK_DIST	4.124434	TOTAL_SLEEP_TIME_HOUR	5.950027
AVG_SPD	1.100169	SLEEP_EFFICIENCY	1.343735	REAL_SUM_WALK_TIME	4.125097	TOTAL_COUNT_CONTINUOUS_WALK_20MINUTES	13.824525
HEIGHT	1.147705	MOR_AVG_SPD	1.609265	NT_WALK_TIME	4.564117	TOTAL_TIME_CONTINUOUS_WALK_20MINUTES	14.084656
NAP_COUNT	1.175760	BED_TIME_VARIANCE_FLAG	1.960718	AFT_WALK_DIST	4.602295		
AFT_AVG_SPD	1.219505	BED_TIME_VARIANCE	1.997267	AFT_WALK_TIME	4.733533		

4.4 데이터 모델링

본 연구는 전처리 및 변수 선택(Feature Selection)을 통해 얻은 최종적인 변수를 네 가지 모델(Ridge 회귀모델, SVR 모델, XGBoost 모델, CatBoost 모델)로 학습하였다. 모델 생성 및 학습 단계에서 train set과 test set를 8:2로 나누어서 학습하였으며, 보다 정확한 검증을 위해 5-fold cross validation을 사용하였다. 더불어, 과적합이 생기지 않게 early_stopping과 하이퍼파라미터 값을 유의하면서 조정하였다. <Figure 1>은 데이터 모델링의 전체적인 흐름을 나타낸다(Oh *et al.*, 2021). 본 연구에서는 하이퍼파라미터 값의 최적값을 찾는 방법으로 GridSearch와 성능은 비슷하지만 속도가 빠른 베이지안 최적화의 접근 방식인 Hyperopt(Bergstra *et al.*, 2013)를 활용하였으며, Hyperopt 검색 공간의 옵션으로는 hp.uniform과 hp.quniform(Putatunda and Kiran, 2018)를 사용하였다. 또한, 하이퍼파라미터의 범위를 설정하는 데에는 선행 연구를 참고하였다(SVR: Crone *et al.*, 2006; Parveen *et al.*, 2016;, XGBoost: Ogunleye and Wang, 2020; Ryu *et al.*, 2020;, CatBoost: Bassi *et al.*, 2021; Zhou *et al.*, 2021).

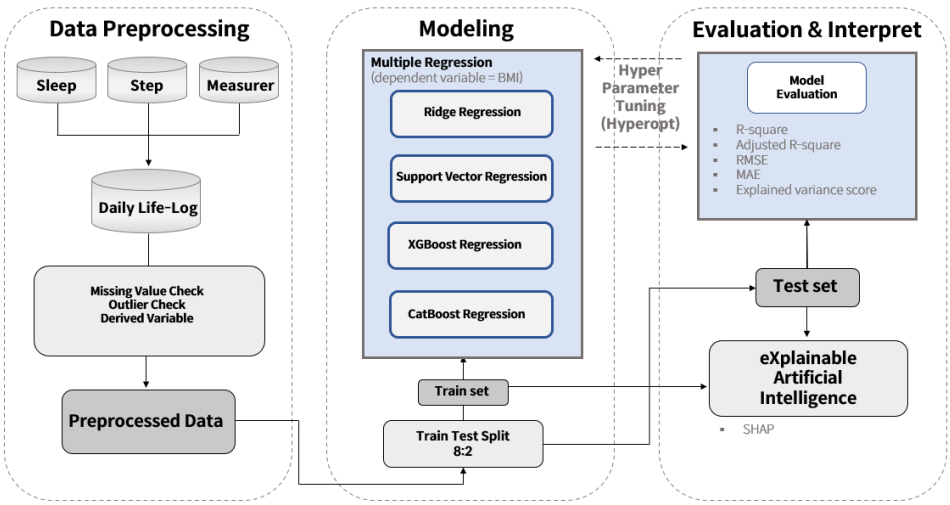

Figure 1. Flow Chart of the BMI Prediction Modeling

Table 3. Hyperparameters for XGBoost, CatBoost, Ridge Regression, SVR model

XGBoost		CatBoost	
colsample_bylevel	0.614374098996575	depth	5
colsample_bytree	0.856746338644902	iterations	800
gamma	0.497626873710138	learning_rate	0.0100096020927548
reg_alpha	0.759634599717601	objective	MAE
learning_rate	0.038069566520957	Ridge Regression	
max_depth	3	alpha	6.682015737206296
min_child_weight	3	SVR	
n_estimators	800	C	49.84009007578617
objective	reg:linear	epsilon	0.3639222481694366
subsample	0.729165370718636	gamma	auto
		kernel	linear

Table 4. Comparison of 4 Machine Learning Algorithms for All Data: Ridge Regression, SVR model, XGBoost, CatBoost

모델	Explained_Variance_Score	R-square	Adj. R-Square	MAE	RMSE
Ridge	0.090	0.089	0.077	2.222	2.910
SVR	0.245	0.245	0.234	1.941	2.650
XGBoost	**0.442**	**0.441**	**0.434**	**1.697**	**2.279**
CatBoost	0.365	0.365	0.356	1.841	2.431

데이터에 5-fold cross validation 적용한 후, Hyperopt로 구한 각 모델의 최적 하이퍼파라미터 값을 <Table 3>에 정리하였고, 최적의 하이퍼파라미터를 적용한 각 모델의 성능지표는 <Table 4>에 정리하였다.

4.5 신장별 군집생성

모델링을 완료한 머신러닝 모델을 사용하여 데이터를 학습한 결과, 네 가지 모델 모두 공통으로 신장 변수의 중요도가 매우 높은 것으로 나타났다(<Figure 2> 참고). 따라서, 신장을 기준으로 군집을 나누어 분석하였다. 통계청 및 기존 연구에 따르면, 우리나라 남성의 30대, 40대, 50대의 평균 키는 각각 174.05cm, 172.15cm, 169.39cm(National Health Insurance Service, 2021)로 평균 신장을 기준으로 각각 군집 1인 165cm~170cm, 군집 2인 170cm~175cm, 그리고 평균 신장과 조금 거리가 있는 175cm~180cm을 군집 3으로 나눠 군집 간 특징에 대해 분석을 하였다.

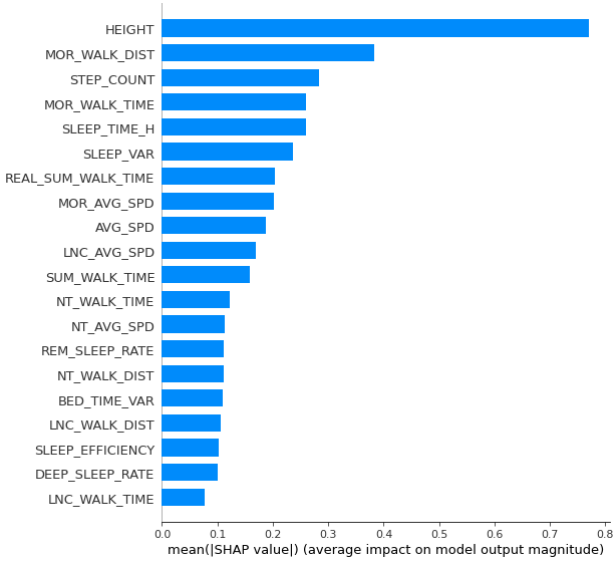

Figure 2. Feature Importance Calculated by the SHAP Method

5. 결 과

군집 1은 사용자 13명으로 구성되었으며, <Table 5>는 4가지 모델의 성능지표를 정리한 것이다. 가장 성능이 좋았던 모델은 Catboost 알고리즘이었으며, Catboost을 통해 군집 1의 데이터 중 임의의 데이터를 SHAP 기법을 활용해 BMI에 영향력이 큰 변수를 산출한 결과는 <Figure 3>와 같다. 붉은색은 target value인 BMI에 대해 높은 값을 예측하도록 영향을 주며, 파란색은 낮은 값을 예측하도록 영향을 주는 변수를 의미한다. 해당 데이터에 대해 예측 모델은 BMI를 23.95로 예측하였고 실제 값은 22.8이었다. 이때 BMI에 음(-)의 영향을 준 요소로써 아침 시간대 평균 걸음 속도가 가장 큰 영향을 주는 것으로 해석된다. 즉, 아침 시간대 평균 걸음 속도가 느려지면 BMI 감소가 발생할 확률이 낮아진다고 볼 수 있다.

군집 2는 사용자 30명으로 구성되었으며, <Table 6>은 4가지 모델의 성능지표를 정리한 것이다. 가장 높은 성능을 보인 모델은 XGBoost 알고리즘이었으며, XGBoost을 통해 군집 2의 데이터 중 임의의 데이터를 SHAP 기법을 활용해 BMI에 영향력이 큰 변수를 산출한 결과는 <Figure 4>와 같다. 군집 2의 해당 데이터에 대해 예측 모델은 BMI를 24.17로 예측하였고 실제 값은 24이었다. BMI에 음(-)의 영향을 주는 요소로서 아침 시간대 평균 걸음 속도와 수면 효율성이 큰 영향을 주는 요소로 해석된다.

군집 3은 사용자 26명으로 구성되었으며, 가장 높은 성능을 보인 모델은 SVR이었다 (<Table 7> 참고). SVR 모델과 SHAP 기법을 이용하여 BMI에 영향력이 큰 변수를 산출한 결과는 <Figure 5>와 같다. <Figure 5>는 추출한 군집 3의 데이터에서 변수들이 BMI에 어떤 영향을 주는지를 보여준다. 예측 모델은 BMI를 26.31로 예측하였고 실제 값은 26.5이었다. BMI에 음(-)의 영향을 준 요소로써 아침 시간대 평균 걸음 속도가 가장 큰 영향을 주는 것으로 해석된다. 즉, 아침 시간대 평균 걸음 속도가 느려지면 BMI 감소가 발생할 확률이 낮아진다고 볼 수 있다.

군집 내 임의의 데이터가 아닌 군집 내 전체 데이터에 대해 변수 영향력을 분석하였을 때, 군집 1에서는 아침 평균 걸음 속도, 밤 시간 대 평균 걸음 속도, 수면 총시간, 평균 걸음 속도, 수면 효율성이 BMI 예측에 대한 주요 변수이며, 군집 2에서는 아침 시간대 걸은 거리, 아침 평균 걸음 속도, 점심 평균 걸음

Table 5. Comparison of 4 Machine Learning Algorithms for Group 1: Ridge Regression, SVR model, XGBoost, CatBoost

모델	Explained_Variance_Score	R-square	Adj. R-Square	MAE	RMSE
Ridge	-0.040	-20.487	-22.171	11.807	12.104
SVR	0.159	0.149	0.082	1.875	2.409
XGBoost	0.597	0.591	0.559	1.248	1.670
CatBoost	**0.611**	**0.605**	**0.574**	**1.240**	**1.641**

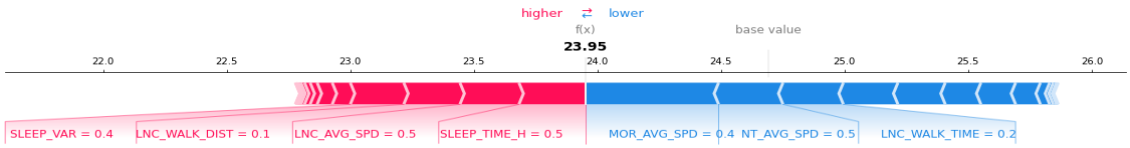

Figure 3. Variables that Affect BMI for Group 1

Table 6. Comparison of 4 Machine Learning Algorithms for Group 2: Ridge Regression, SVR model, XGBoost, CatBoost

모델	Explained_Variance_Score	R-square	Adj. R-Square	MAE	RMSE
Ridge	-0.013	-1.434	-1.510	2.871	3.490
SVR	0.069	0.054	0.024	1.622	2.176
XGBoost	**0.453**	**0.453**	**0.436**	**1.209**	**1.654**
CatBoost	0.319	0.316	0.295	1.305	1.850

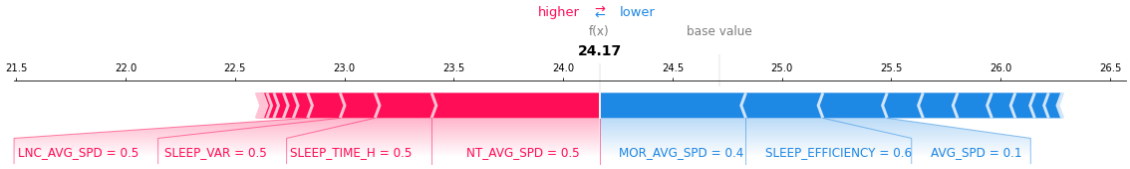

Figure 4. Variables that Affect BMI for Group 2

Table 7. Comparison of 4 Machine Learning Algorithms for Group 3: Ridge Regression, SVR model, XGBoost, CatBoost

모델	Explained_Variance_Score	R-square	Adj. R-Square	MAE	RMSE
Ridge	0.316	-107.124	-134.390	38.207	38.328
SVR	**0.837**	**0.830**	**0.787**	**1.245**	**1.522**
XGBoost	0.720	0.691	0.613	1.431	2.048
CatBoost	.732	0.698	0.621	1.343	2.027

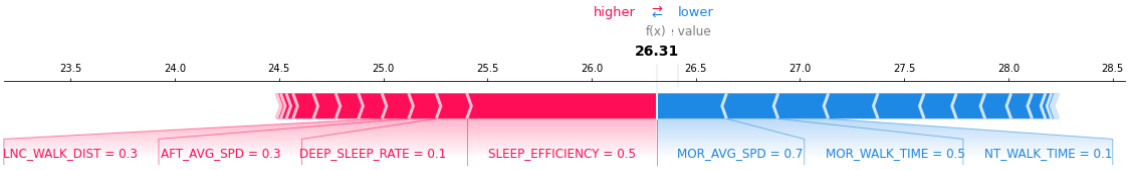

Figure 5. Variables that Affect BMI for Group 3

속도, 수면 효율성, 점심 시간대 걸은 거리가 BMI 예측에 대한 중요변수이며, 군집 3에서는 일별 총수면 시간 편차, 일별 걸음 수, 아침 시간대 걸은 시간, 점심 시간대 걸은 거리, 거리 기반으로 산출한 일별 걸은 총시간이 BMI 예측에 대한 주요 변수임을 확인할 수 있었다. 또한, 신장에 따라 군집을 나눠 실험한 결과(<Table 5>, <Table 6>, <Table 7>)가 군집을 나누지 않은 전체 데이터(<Table 4>)로 실험한 결과보다, 모든 성능지표에서 나은 결과를 보여주었다.

6. 결 론

본 연구에서는 웨어러블 디바이스를 통해 측정된 수면 및 걸음 데이터로 대한민국 중년남성의 건강관리 솔루션을 제시하기 위해 신장 그룹별 라이프로그 변수가 BMI에 미치는 영향력을 연구하였다. 스마트워치로부터 수집된 중년 남성 71명의 걸음, 수면 데이터를 신장에 따라 세 가지 군집(165cm~170cm, 170cm~175cm, 175cm~180cm)으로 나누었다. 걸음, 수면 데이터로부터 55개의 파생 변수를 생성하였고, 과적합과 결과 해석의 용이성을 위해 파생변수 중 다중공선성이 낮은 변수들만 선택하여 설명변수로 사용하였다. 머신러닝 기법으로는 회귀분석의 대표적인 네 가지(Ridge Regression, Support Vector Regression, XGBoost, CatBoost) 모델을 사용하고, 이에 대한 성능을 비교하였다. 또한, SHAP 기법을 활용하여 군집 중 임의의 데이터에 대하여 라이프로그 변수들이 BMI를 예측하는데 미치는 영향력을 양(+)과 음(-)으로 구분하여 시각화하였다. 실험 결과, 신장에 따라 군집을 나눠 실험했을 때 군집을 나누지 않은 경우보다 BMI를 더 잘 예측하는 것을 알 수 있었다. 또한, 신장 군집에 따라 변수의 영향력 및 중요도가 다르다는 것을 알 수 있었다. 즉, 개인의 BMI 예측을 위해서는 신장과 같은 신체적 특성으로 군집을 나눠 분석하는 방식이 고려되어야 함을 보여준다. 따라서, 본 연구 결과를 기반으로 라이프로그 데이터의 군집 별 분석을 통해, BMI 예측을 비롯해 건강관리에 영향력 있는 변수를 도출함으로써, 향후, 개인별 맞춤 건강관리에 기여할 수 있을 것으로 기대한다.

향후 연구에 있어서는 데이터가 매우 큰 경우, 본 연구에서 사용한 VIF stepwise 방식의 변수 선택 방식은 계산량이 매우 커지는 문제가 발생할 수 있으므로 변수 중요도에 대한 시각화뿐만 아니라 Feature Selection에서도 효과적으로 적용할 수 있는 Boruta-Shap 기법(Marcílio and Danilo, 2020; Chieregato *et al.*, 2021)을 활용하는 것을 제안한다.

참고문헌

Akinwande, M. O., Dikko, H. G., and Samson, A. (2015), Variance Inflation Factor: As a Condition for the Inclusion of Suppressor Variable (s) in Regression Analysis, *Open Journal of Statistics*, **5**(07), 754.

Basak, D., Pal, S., and Patranabis, D. C. (2007), Support vector regression, *Neural Information Processing−Letters and Reviews*, **10**(10), 203-224.

Bassi, A., Shenoy, A., Sharma, A., Sigurdson, H., Glossop, C., and Chan, J. H. (2021), Building Energy Consumption Forecasting: A Comparison of Gradient Boosting Models, *The 12th International Conference on Advances in Information Technology*, 1-9.

Bergstra, J., Yamins, D., and Cox, D. D. (2013), Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms, *Proceedings of the 12th Python in Science Conference*, 13, 20

Chen, T. and Guestrin, C. (2016), *Xgboost: A scalable tree boosting system*, 785-794.

Chieregato, M., Frangiamore, F., Morassi, M., Baresi, C., Nici, S., Bassetti, C., Bnà, C., and Galelli, M. (2021), A hybrid machine learning/deep learning COVID-19 severity predictive model from CT images and clinical data, *ArXiv Preprint ArXiv*:2105.06141.

Correia de Faria Santarém, G., de Cleva, R., Santo, M. A., Bernhard, A. B., Gadducci, A. V., Greve, J. M. D., and Silva, P. R. S. (2015), Correlation between body composition and walking capacity in severe obesity, *PloS One*, **10**(6), e0130268.

Crone, S. F., Guajardo, J., and Weber, R. (2006), A study on the ability of support vector regression and neural networks to forecast basic time series patterns, *Springer US*, 149-158.

Delnevo, G., Mancini, G., Roccetti, M., Salomoni, P., Trombini, E., and Andrei, F. (2021), The Prediction of Body Mass Index from Negative Affectivity through Machine Learning: A Confirmatory Study, *Sensors*, **21**(7), 2361.

Dodge, M. and Kitchin, R. (2007), 'Outlines of a world coming into existence': Pervasive computing and the ethics of forgetting, *Environment and Planning B: Planning and Design*, **34**(3), 431-445.

Donini, L. M., Poggiogalle, E., Mosca, V., Pinto, A., Brunani, A., and Capodaglio, P. (2013), Disability affects the 6-minute walking distance in obese subjects (BMI> 40 kg/㎡), *PloS One*, **8**(10), e75491.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., and Vapnik, V. (1997), Support vector regression machines, *Advances in Neural Information Processing Systems*, **9**, 155-161.

Ekelund, U., Ward, H. A., Norat, T., Luan, J., May, A. M., Weiderpass, E., Sharp, S. J., Overvad, K., Østergaard, J. N., and Tjønneland, A. (2015), Physical activity and all-cause mortality across levels of overall and abdominal adiposity in European men and women: The European Prospective Investigation into Cancer and Nutrition Study (EPIC), *The American Journal of Clinical Nutrition*, **101**(3),613-21.

Fryer, D., Strumke, I., and Nguyen, H. (2021), Shapley values for feature selection: The good, the bad, and the axioms, *arXiv*:2102.10936

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018), *Explaining explanations: An overview of interpretability of machine learning*, 80-89.

Głąb, T., Sadowska, U., and Żabiński, A. (2015), Application of image analysis for grass tillering determination, *Environmental Monitoring and Assessment*, **187**(11), 1-9.

Halim, A. A., Basu, A., and Kirk, R. (2019), The Prevalence of Body Mass Index-Associated Chronic Diseases in Diverse Ethnic Groups in New Zealand, *Asia Pacific Journal of Public Health*, **31**(1), 84-91.

Hoerl, A. E., and Kennard, R. W. (1970), Ridge Regression: Biased Estimation for Nonorthogonal Problems, *Technometrics*, **12**(1), 55-67.

Hu, F. B. (2008), *Obesity Epidemiology*, Physical activity, sedentary behaviors, and obesity, 301-319, Oxford University Press, Inc, New York, US.

Kim, B-Y. and Shin, M-H. (2010), Procedure for the Selection of Principal Components in Principal Components Regression, *The Korean Statistical Society*, **23**(5), 967-975.

Kim D-H. (2009), Effect of Walking Exercise, *Korean J Fam Med.*, **30**(3), S329-S331.

Kim, E. J., Park Y-B., Choi, K. H., Lim, Y.-W., Ok, J.-M., Noh, E.-Y., Song, T. M., Kang, J. H., Lee, H. S., and Kim, S.-Y. (2020), Application of Machine Learning to Predict Weight Loss in Overweight, and Obese Patients on Korean Medicine Weight Management Program, *The Journal of Korean Medicine*, **41**(2), 58-79.

Liu, X., Forbes, E., Ryan, N. D., and Rofey, D. (2008), Rapid Eye Movement Sleep in Relation to Overweight in Childrenand Adolescents, *Arch Gen Psychiatry*, **65**(8), 924-932.

Luyster, F. S., Strollo, P. J., Zee, P. C., and Walsh, J. K. (2012), Sleep: A health imperative, *Sleep*, **35**(6), 727-734.

Lundberg, S. M. and Lee, S. I. (2017), A Unified Approach to Interpreting Model Predictions, *Neural Information Processing 17*.

Maclure, K. M., Hayes, C. K., Colditz, A. G., Stampfer, J. M., Speizer, E. F, and Walter C. W. (1989), Weight, diet, and the risk of symptomatic gallstones in middle-aged women, *N Engl J Med.*, **321**(9), 563-569.

Marcílio, W. E. and Danilo M. E. (2020), From explanations to feature selection: assessing SHAP values as feature selection mechanism, *IEEE 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 340-347.

Mitchell, R., Adinets, A., Rao, T., and Frank, E. (2018), Scalable GPU Accelerated Learning, *arXiv*:1806.11248, 1-4

Natekin, A. and Knoll, A. (2013), Gradient Boosting Machines, A Tutorial, *Front. Neurorobotics*, **7**, 1-21.

National Health Insurance Service (2021), Average height distribution by province, age, and gender: General, https://kosis.kr/statHtml/statHtml.do?orgId=350&tblId=DT_35007_N130.

Ogunleye, A. and Wang, Q. G. (2020), XGBoost Model for Chronic Kidney Disease Diagnosis, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **17**(6), 2131-2140.

Oh, H.-R., Son, A.-L., and Lee, Z.-K. (2021), Occupational accident prediction modeling and analysis using SHAP, *Digital Contents Society*, **22**(7), 1115-1123.

Park, Y.-J., Lee, W.-C., Yim, H.-W, and Park, Y.-M. (2007), The Association between Sleep and Obesity in Korean Adults, *J Prev Med Public Health*, **40**(6), 454-460.

Parveen, N., Zaidi, S., and Danish, M. (2016), Support vector regression model for predicting the sorption capacity of Lead(II), *Perspectives in Science*, **8**, 629-631.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, V. A., and Gulin, A. (2017), CatBoost: unbiased boosting with categorical features, *arXiv*:1706.09516, 3-17.

Putatunda, S. and Kiran, R. (2018), A Comparative Analysis of Hyperopt as Against Other Approaches for Hyper-Parameter Optimization of XGBoost, *SPML '18: Proceedings of the 2018 International Conference on Signal Processing and Machine Learning*, 6-10.

Ryu, S. E., Shin, D. H., and Chung, K. (2020), Prediction Model of Dementia Risk Based on XGBoost Using Derived Variable Extraction and Hyper Parameter Optimization, *IEEE Access*, **8**, 177708-177720.

Shin, J.-H. and Kim, M.-S. (2016), The Effects on Body Composition, Health Related Physical Fitness and Metabolic Syndrome Factors of Working Exercise Program in Obese Middle-Aged Women, *Korea Coaching Development Center*, **18**(1), 39-46.

Singh, B. and Tawfik, H. (2020), Machine Learning Approach for the Early Prediction of the Risk of Overweight and Obesity in Young People, *Computational Science*, 523-535.

Song, M., Hu, F. B., Spiegelman, D., Chan, A. T, and Wu, K. (2015), Adulthood weight change and risk of colorectal cancer in the Nurses' Health Study and Health Professionals Follow-up Study, *Cancer Prev Res (Phila)*, **8**(7), 620-627.

Stenholm, S., Head, J., and Vahtera, J. (2017), Body mass index as a predictor of healthy and disease-free life, expectancy between ages 50 and 75: A multicohort study, *International journal of Obesity*, **41**(5), 769-775.

Thamrin, S. A., Arsyad, D., and Nasir, S. (2021), Predicting Obesity in Adults Using Machine Learning Techniques: An Analysis of Indonesian Basic Health Research 2018, *Frontiers in Nutrition*, **8**, 252.

Van Breugel, P., Friis, I., Demissew, S., Lillesø, J.-P. B., and Kindt, R. (2016), Current and future fire regimes and their influence on

natural vegetation in Ethiopia, *Ecosystems*, **19**(2), 369-386.
Vittinghoff, E., Glidden, D. V., Shiboski, S. C., and McCulloch, C. E. (2012), *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*, 2nd edition, Springer, Boston, MA, US.
Wang, M., Yi, Y., Roebothan, B., and Colbourne, C. (2016), Body Mass Index Trajectories among Middle-Aged and Elderly Canadians and Associated Health Outcomes, *Journal of Environmental and Public Health*, **2016**, 7014857.
Yiengprugsawan, V., Banwell, C., Zhao, J., Seubsman, S-A., and Sleigh, A. (2014), Relationship between Body Mass Index Reference and All-Cause Mortality: Evidence from a Large Cohort of Thai Adults, *Journal of Obesity*, **2014**, 6.
Ying, X. (2019), An Overview of Overfitting and its Solutions, *Journal of Physics: Conference Series*, **1168**, 022022.
Zheng, Y., Manson, J. E., and Yuan, C. (2017), Associations of Weight Gain From Early to Middle Adulthood With Major Health Outcomes Later in Life, *JAMA*, **318**(3), 255-269.
Zhou, F., Pan, H., Gao, Z., and Huang, X. (2021), Fire Prediction Based on CatBoost Algorithm, *Mathematical Problems in Engineering*, **2021**, 1929137.

저자소개

김지용: 김지용은 광운대학교 수학과에서 2021년 학사를 취득하였다. 연구분야는 라이프로그 및 헬스케어 데이터 분석, 머신러닝이다.

이지수: 이지수는 고려대학교 보건정책관리학부에서 2021년 학사학위를 취득하였다. 연구분야는 라이프로그 및 헬스케어 데이터 분석, 머신러닝이다.

박민서: 박민서 교수는 2009년 메사추세츠 대학교 컴퓨터사인언스(머신러닝) 전공으로 박사학위를 취득하였다. 삼성 SDS Bioinformatics Lab 및 성균관대학교 삼성융합의과학원 수석연구원, SK 텔레콤 팀리더, 한화시스템 상무(AI Lab 장)을 거쳐 현재 서울여자대학교 데이터사이언스학과 교수로 재직하고 있다. 연구분야는 라이프로그 및 헬스케어 데이터 분석, 바이오인포메틱스, 머신러닝이다.