Journal of the Korea Academia-Industrial cooperation Society
Vol. 25, No. 4 pp. 324-332, 2024

https://doi.org/10.5762/KAIS.2024.25.4.324
ISSN 1975-4701 / eISSN 2288-4688

캣부스트 모델을 이용한 대사증후군 예측

유보현[1,4], 최아라[2,4*], 김태훈[3]
[1]빛고을 전남대학교병원, [2]전남대학교병원 의생명연구지원센터, [3]전라남도 공공보건의료지원단, [4]전남대학교 일반대학원 간호학과

Prediction of Metabolic Syndrome Using The Catboost Model

Bo Hyun Yu[1,4], Ah-ra Choi[2,4*], Tae Hun Kim[3]
[1]Chonnam National University Bitgoeul Hospital
[2]Chonnam National University Hospital Biochemical Research Institute
[3]Jeollanamdo Public Health Policy Institute
[4]Department of Nursing, Graduate School, Chonnam National University

요 약 본 연구는 머신러닝 중 하나인 캣부스트(catboost) 알고리즘을 통해 비침습적 방법으로 대사증후군을 예측하고자 시도되었다. 예측모델의 학습을 위한 자료는 국민건강영양조사 제 8기(2019-2021) 자료 중 대사증후군이 없는 그룹 11,542명, 대사증후군이 있는 4,008명으로 총 11,545명이며, 투입변수는 비침습적 요인들로만 구성된 14개였다. 본 연구의 모델구축 및 성능평가를 위한 모든 코드는 Python 3.9.7로 작성되었으며, 통계 및 모델 구축을 위해서 SciPy 1.614, SHAP(Shapley Additive exPlanations), Scikit-learn 1.2.2 패키지가 사용되었다. 연구결과 대사증후군 예측에 가장 중요한 요인은 수면시간(1.354), 체질량 지수(BMI: body mass index, 이하 BMI)(1.153), 성별(1.118)로 확인되었다. 또한 연령(0.658), 현재 흡연(0.229), 1년간 체중변화 (0.199), 만성질환 가족력(0.176), 나쁜 주관적 건강인지(0.158), 1개월에 1잔 이상의 음주(0.132), 만성질환(0.092), 높은 스트레스 인지(0.073), 유산소 신체활동 실천 없음(0.063), 활동제한(0.060), 암(0.016) 순으로 나타났다. 예측모델의 전반적 성능(AUC)은 0.874 (95% CI, 0.874-0.874)로 확인되었다. 본 연구에서 구축한 모델을 활용하여 간편한 질문을 통해 대사증후군을 빠르고 정확하게 식별할 수 있으며, 이는 개인 및 집단 수준에서 예방프로그램을 개발하는 데 중요한 기초자료로 활용될 수 있을 것으로 기대한다.

Abstract This study attempted to predict the metabolic syndrome in a non-invasive manner using the CatBoost algorithm, a type of machine learning. The dataset for training the prediction model comprised 11,545 individuals, including 11,542 without metabolic syndrome and 4,008 with metabolic syndrome, sourced from the 8th Korea National Health and Nutrition Examination Survey (KNHANES, 2019-2021). The fourteen input variables consisted solely of non-invasive factors. All the code for the model construction and performance evaluation was written in Python 3.9.7, utilizing SciPy 1.614, SHAP (Shapley Additive exPlanations), and Scikit-learn 1.2.2 packages for statistics and model building. The study found that the most important factors in predicting metabolic syndrome were sleep duration (1.354), body mass index (BMI, 1.153), and gender (1.118). Additionally, age, smoking, weight change over the past year, family history of chronic disease, poor perceived health status, alcohol consumption of more than one drink per month, and chronic diseases followed in order of importance. The overall performance of the prediction model (area under the curve [AUC]) was confirmed to be 0.874 (95% CI, 0.874-0.874). The model constructed in this study could provide foundational data for the early identification of metabolic syndrome in individuals and populations through simple questions, as well as assist in the development of prevention programs for diseases related to the metabolic syndrome.

Keywords : Machine Learning, Catboost, Metabolic Syndrome, Prediction Model, KNHANES

*Corresponding Author : Ah-Ra Choi(Chonnam National University Hospital Biochemical Research Institute)
email: 7violet8@hanmail.net
Received March 6, 2024 Revised March 25, 2024
Accepted April 5, 2024 Published April 30, 2024

1. 서론

대사증후군(metabolic syndrome)은 National Cholesterol Education Program's Adult Treatment Panel III (NCEP ATP III)의 진단기준에 따라 공복 시 혈당, 혈압, 중성지방, 고밀도 지단백 콜레스테롤, 복부 둘레(허리둘레) 중 3개 이상에서 비정상 소견을 가지는 경우 대사증후군으로 정의한다 [1].

최근 식생활과 생활양식의 서구화로 국내에서도 대사증후군의 유병률은 지속적으로 증가하는 추세이다. 심장대사증후군학회(KSCMS: Korean Society of CardioMetabolic Syndrome)에서 분석한 Metabolic syndrome Fact sheet in Korea 2021 자료에 따르면, 우리나라의 성인 대사증후군의 유병률은 2007년부터 2018년, 12년간 대사증후군의 유병률은 21.6%에서 24.6%의 수준을 보이며 꾸준히 증가하고 있다. 또한 연령이 증가함에 따라 유병률이 증가되는데, 60세에서 69세 미만에서 41.0%, 70세 이상에서 47.4%로 높은 수치를 보였다. 이렇게 높은 유병률을 보이는 대사증후군은 심혈관질환 발병 위험을 2배로 증가시키고 전체 사망률도 1.5배 높이는 것으로 나타났다 [2]. 대사증후군은 심뇌혈관질환의 주요 위험 요인으로 밀접한 관련성을 보이므로 대사증후군의 조기진단과 치료는 임상적으로 매우 중요하다 [3,4].

또한, 대사증후군 예방에 대한 관심이 높아짐에 따라, 대사 증후군 예측에 관한 연구도 적극적으로 시도되고 있으며, 결과는 인종, 연령, 성별 등에 따라 다양하다 [5]. 그러나 대부분의 연구는 대사증후군과 특정 위험인자와의 연관성에 관한 연구[6-9]였으며, 예측요인의 탐색에 그친 연구 [5,10-12]가 많았고, 침습적 절차와 비침습적 절차를 통해 확인될 수 있는 요인들을 모두 포함하고 있다. 이는 비용이 많이 들고 시간이 많이 걸릴 수 있으며 [13], 침습적 검사는 고통스럽고 불편한 절차일 수 있다. 따라서 대사증후군을 선별하기 위한 빠르고 간단한 비침습적이며 비용 효율적인 예측모델을 개발하는 것이 필요하다.

최근 보건의료 연구에서는 다양한 요인들을 분류 및 예측할 수 있는 머신러닝 방법을 사용한 연구가 활발히 진행되고 있으며 [14-16], 선별 및 진단과정을 보다 효율적이고 객관적이며 신뢰할 수 있게 만들 수 있는 도구를 개발하기 위해 다양한 기계 학습 방법이 사용되고 있다 [15,17,18].

전통적인 회귀모형은 요인이 많을 때 수렴하지 않거나, 성립되지 않는다 [19]. 또한 표본이 많을수록 유의성 검증으로 인한 문제가 생길 수 있다 [19]. 하지만, 머신러닝 기반모델은 이러한 전통적인 통계기반 회귀모형의 제약을 극복할 수 있다 [20]. 더불어 머신러닝은 예방 헬스케어 시대의 가능성을 높여주고, 특정 환경에 맞는 데이터를 활용하여 정확성을 높일 수 있다 [21].

따라서, 대사증후군의 영향요인과 예측을 위해 보다 포괄적이고 다양한 비침습적 요인들을 기계학습을 사용한 예측모델을 통해 확인할 필요가 있다. 이를 통해 대사증후군에 대한 침습적 검사가 없이도 대사증후군과 관련된 질병을 조기 발견하여, 개인 및 집단 수준에서의 건강한 삶을 증진시킬 수 있을 것이다. 이에 본 연구에서는 국민건강영양조사 제 8기(2019~2021)의 데이터를 기반으로 하여 대사증후군을 비침습적으로 예측하는 모델을 개발하고자 한다. 이를 통해 대사증후군 환자를 조기에 식별하고, 침습적 검사에 따른 부담을 줄일 수 있는 방법으로 제시하고자 한다.

2. 연구 방법

2.1 연구 설계

본 연구는 국민건강영양조사 제 8기(2019~2021년) 데이터를 기반으로 하여, 캣부스트(catboost) 알고리즘을 사용하여 대사증후군을 선별하기 위한 횡단연구(cross-sectional study)이다. 본 연구에서는 캣부스트(catboost) 알고리즘을 활용하였다.

CatBoost 알고리즘은 범주형 변수를 별도로 처리하지 않고 모델의 입력으로 직접 사용할 수 있어 범주형 변수의 전처리와 과적합(overfitting)문제를 해결할 수 있다는 장점이 있어 [22,23] 범주형 자료가 대부분인 본 연구에 적합하다고 판단하여 선택하였다.

2.2 자료원 및 연구대상

본 연구에 사용된 자료는 국민건강영양조사 제 8기(2019~2021년) 데이터를 활용하였다. 원시자료는 총 22,559명이었고, 19세 미만 대상자와 관련변수 자료가 누락된 대상자를 제외한 총 11,545명을 대상으로 분석하였다(Table 1).

Table 1. Variables excluded because did not quality for the study

Variables	process	count	n(%)
Under age 19	Drop	3,868	
Height	Drop	1,020	
Weight	Drop	865	
Smoking status	Drop	1,013	
Drinking status	Drop	999	
Sleeping time	Drop	803	
Perceived health status	Drop	792	
Perceived stress	Drop	1,019	
Weight change over 1 year	Drop	800	
Aerobic physical activity	Drop	2,059	
Physical activity restrictions	Drop	792	
Family history of chronic disease	Drop	1,251	
Stroke	Drop	792	
Coronary disease	Drop	1,986	
Osteoarthritis	Drop	792	
Rheumatoid arthritis	Drop	792	3,146(13.9)*
Osteoporosis	Drop	792	
Gout	Drop	792	
Thyroid disease	Drop	792	
Kidney disease	Drop	792	
Liver Cirrhosis	Drop	792	
Hepatitis B	Drop	792	
Hepatitis C	Drop	792	
Stomach cancer	Drop	792	
Liver cancer	Drop	792	
Colon cancer	Drop	792	
Breast cancer	Drop	792	
Cervical cancer	Drop	792	
Lung cancer	Drop	792	
Thyroid cancer	Drop	792	
Extra cancer	Drop	792	
Final raw datas			15,545(68.9)

*Combine duplicates for each variable

2.3 모델의 학습에 사용된 데이터셋의 지표설정

본 연구에 사용된 국민건강영양조사 제 8기 데이터의 총 변수는 838개이며, 그 중 데이터베이스관리의 수정일, 개인 및 가구 식별을 위한 변수를 제외한 835개의 변수가 모델학습을 위한 데이터셋으로 검토되었다. 또한 복합표본설계인 국민건강영양조사 자료의 특성을 고려하여, 모델학습을 위한 데이터의 가중치인 wt_itvex가 모델 구축에 사용되었다.

본 연구는 의료전문가와 비전문가를 막론하고 빠르고 간편하게 사용할 수 있는 방법을 제안하기 위해, 대사증후군과 관련된 요인 중 비침습적 변수만을 선택하였다. 데이터셋으로 포함된 변수들은 임상현장에서 심혈관질환자를 진료중인 순환기내과 교수 1인, 순환기내과 간호사 2명, 성인간호학 교수 1명의 자문을 받아 선택하였으며, 각 전문가는 선행논문을 검토한 후 자문이 이루어졌다.

2.3.1 투입 변수

자문을 통해 선택된 데이터셋의 변수는 성별, 나이, BMI, 흡연, 1개월에 1잔 이상의 음주, 수면시간, 주관적 건강인지, 스트레스 인지, 최근 1년간 체중 변화, 유산소 신체활동 실천, 활동제한 여부, 만성질환 가족력으로 결정되었다. 또한 동반질환이 영향을 미치는지 확인을 위해 만성질환 (뇌졸중, 관상동맥질환, 골관절염, 류마티스 관절염, 골다공증, 통풍, 갑상선질환, 신장질환, 간경화, B형간염, C형간염) 진단 여부와 암(위암, 간암, 대장암, 유방암, 자궁암, 갑상샘암, 기타암) 진단 여부를 선택하였다.

2.3.2 대사증후군 진단 변수

대사증후군은 NCEP-ATP III 개정안 [24]과 대한비만학회에서 제시한 복부 비만의 기준 [25]에 근거하여 정의하였다. 포함되는 기준은 다음과 같다.

1) 허리둘레: 남자 ≥ 90cm, 여자 ≥ 80cm, 2) 중성지방 ≥ 150mg/dL, 3) 고밀도지단백콜레스테롤 (HDL-C: high density lipoprotein-Choesterol) : 남자 〈 40mg/dL,여자 〈 50mg/dL, 4) 혈압 ≥ 130/85 mmHg 또는 혈압강하제를 복용중, 5) 공복혈당 ≥ 100mg/dL 또는 혈당강하제(또는 인슐린)를 사용 중인 경우로 이 5가지 중 3가지 이상을 가지고 있는 경우 대사증후군으로 분류하였다.

2.4 모델 학습을 위한 데이터 전처리

본 연구에서 범주형 변수는 캣부스트(catboost) 알고리즘을 사용하므로 원-핫 인코딩(one-hot encoding)을 이용한 변환이 필수는 아니지만, 예측에 기여하는 요인들을 확인하기 위해 변환 처리하였다 [26]. 그리고 만성질환과 암은 해당 변수들을 하나로 통합하여 진단유무를 기준으로 새로운 범주형 변수인 만성질환과 암 변수로 생성하였다. 또한 각 자료에서 결측치와 해당 없음에 응답하거나 응답을 거부한 문항의 경우 NaN(not a number)으로 변환하여, 이산형 변수들을 최빈값 및 미스포레스트(missforest) 알고리즘을 통해 대치하였으나,

결과 해석상 문제가 있을 것으로 판단되어 최종적으로 제외하였다(Table 1).

본 연구에 사용된 대사증후군이 없는 대상자와 있는 대상자의 비율은 2.9:1 정도로 불균형하다. 이렇게 불균형한 데이터로 학습된 머신러닝 모델은 다수 데이터에 대한 과적합(overfitting)으로 인해 소수 데이터 분류 예측 시, 성능이 좋지 않은 경향을 보일 수 있다. 이 문제를 해결하기 위해 데이터가 가진 표본 수의 불균형을 해결하기 위한 방법인 샘플링 기법 [27]을 사용하였다. 본 연구에서 사용된 샘플링 기법은 낮은 비율인 데이터를 k-최근접 이웃(k-nearest neighbor, k-NN) 알고리즘을 활용하여 새로 생성하는 방법인 소수 클래스 오버 샘플링(Synthetic Minority Over-sampling Technique, SMOTE) 기법을 통해 대사증후군이 없는 군과 있는 군의 표본수가 일치하도록 처리하였다 [28].

2.5 캣부스트(catboost) 모델

캣부스트(catboost)는 Category Boosting의 약자로 얀덱스사가 개발한 의사결정 트리 라이브러리를 강화한 그래디언트 부스팅(gradient boosting) 방법이다[29]. 캣부스트(catboost) 모델은 gradient boosting machine (GBM)을 기반으로 한 머신러닝 알고리즘으로 ordered boosting을 사용하여 기존 GBM 알고리즘에서 각 단계별 모형의 구축 시 해당 시점의 예측 대상이 되는 target 변수를 포함하여 모형의 과적합(overfitting) 가능성이 높아지는 문제를 해결하였으며, 범주형 변수처리에 유용한 모형이다[22]. 본 연구에서 설정된 캣부스트(catboost)의 최적의 매개변수는 다음과 같다(Table 2).

Table 2. Variables excluded because did not quality for the study

Best parameters	Value
learning rate	0.2
12_leaf_reg	3
iterations	300
depth	3

2.6 캣부스트(catboost) 모델 구축

캣부스트(catboost) 모델구축을 위해 데이터셋을 학습용과 테스트용으로 분할하였다. 데이터의 분할을 위해, 무작위분할을 지원하는 Python의 Scikit-learn 패키지를 이용하였으며, 전체 데이터셋의 80%는 학습용, 20%는 테스트용으로 사용되었다.

캣부스트(catboost) 모델의 학습은 지도학습방식으로, 학습용 데이터셋의 투입변수와 출력변수인 대사증후군 진단 변수를 동시에 입력받아 구축하였다. 구축된 모델의 성능평가는 테스트용 데이터셋을 대상으로 하여 대사증후군 여부를 선별하고 실제 데이터셋과 결과를 비교하여 구축된 모델의 성능을 확인하였다. 모델의 성능 평가는 총 100회 반복 수행된 결과를 평균으로 측정하였다.

2.7 통계 분석

본 연구의 통계 분석 및 모델의 구축을 위해 Python 3.9.7(Python software foundation, Wilmington, DE, USA)언어와 SciPy 1.614, SHAP(Shapley Additive exPlanations), Scikit-learn 1.2.2 패키지를 이용하였으며, 구체적 통계 방법은 다음과 같다.

1) 주입변수 특성에 대한 평균, 표준편차, 교차분석은 SciPy패키지를 이용하여 분석하였다.
2) SHAP(SHapley Additive exPlanations) 패키지를 사용하여 모델 구축에 기여하는 변수들간 상대적인 중요도를 확인하였다. SHAP는 게임이론의 Shapley value를 이용한 알고리즘으로, 결과값에 대한 각 feature의 기여도를 의미한다 [30]. SHAP를 사용하면 모델의 예측값에 대한 변수의 영향력과 입력변수에 따른 예측값의 변화정도를 알 수 있다 [31].
3) 구축된 모델의 성능은 scikit-learn 패키지를 이용하여 확인하였다. 테스트용 데이터셋에서 구축된 모델의 전반적 성능인 AUC(area under the receiver operating characteristics curve), 정확도(accuracy), 정밀도(precision)와 재현율(recall)의 조화평균인 F1 score, 특이도(specificity), 민감도(sensitivity) 그리고 민감도와 특이도의 기하평균인 G-mean(Geometric mean)을 사용하여 균형적인 성능을 평가하였다 [32]. 모든 성능 지표는 bootstrapping을 사용한 95% 신뢰구간(confidence interval, CI)과 평균으로 나타냈다 [33].

3. 연구 결과

3.1 대사증후군 예측모델 투입변수의 특성

연구 대상의 대사증후군 예측모델에 투입된 변수의 특성을 확인하기 위해 독립표본 T검정과 교차분석을 실시하였다.

Table 3. Input data Characteristics of metabolic syndrome prediction model

Variables	Categories	Metabolic syndrome				p value
		Absent(n=11,542)		Present(4,003)		
		n(%)	Mean±SD	n(%)	Mean±SD	
Age			49.8±17.1		55.35±15.0	〈.0.001
Sex	Male	3,819(33.1)		3,131(78.2)		〈.0.001
	Female	7,723(66.9)		872(21.8)		
Body mass index	〈25kg/m^2	8,479(73.5)		1,433(35.8)		〈.0.001
	≥25kg/m^2	3,063(26.5)		2,570(64.2)		
Smoking status	Non-smoker	10,035(86.9)		2,915(72.8)		〈.0.001
	Current smoker	1,507(13.1)		1,089(27.2)		
Drinking status	No drinking	5,831(50.5)		1,529(38.2)		〈.0.001
	≥1 cup/month	5,711(49.5)		2,474(61.8)		
sleeping time			6.8±1.4		6.7±1.4	〈.0.001
Perceived health status	Poor	7,750(67.1)		2,897(72.4)		〈.0.001
	Good	3,792(32.9)		1,106(27.6)		
Perceives stress	Low	8,436(73.1)		2.987(74.6)		0.059
	High	3,106(26.9)		1,016(25.4)		
Weight change over 1 year	No	7,105(61.6)		2,381(59.5)		0.020
	Yes	4,437(38.4)		1,622(40.5)		
Aerobic physical activity	No	5,067(43.9)		1.622(40.5)		〈.0.001
	Yes	6,475(56.1)		2.381(59.5)		
Physical activity restrictions	No	10,791(93.5)		3,667(91.6)		〈.0.001
	Yes	751(6.5)		336(8.4)		
Family history of chronic disease	No	4,296(37.2)		1,415(35.3)		0.034
	Yes	7,246(62.8)		2,588(64.7)		
Chronic disease	No	8,562(74.2)		2,969(74.2)		0.988
	Yes	2,980(25.8)		1,034(25.8)		
Cancer	No	10,920(94.6)		3,796(94.8)		0.597
	Yes	622(5.4)		207(5.2)		

원자료에서 추출한 연구대상자는 총 15,545명으로, 대사증후군이 없는 그룹은 11,542명, 있는 그룹은 4,003명이었다. 평균연령은 대사증후군 그룹(55.35± 15.0)이 대사증후군이 없는 그룹(49.8±17.1)에 비해 더 높았고(p〈0.001), 성별은 대사증후군이 있는 그룹은 남성(78.2%)의 비율이 더 많았다(p〈0.001). BMI는 대사증후군 그룹에서 25kg/m^{2}이상에 해당되는 비율이 64.2%로 더 높았고(p〈0.001), 1년간 체중 변화(38.4% vs. 40.5%, p=0.020) 또한 대사증후군 그룹에서 더 많은 것으로 확인되었다. 흡연(13.1% vs. 27.2%, p〈0.001)과 1개월에 1잔 이상의 음주(49.5% vs. 61.8%, p〈0.001)는 모두 대사증후군 그룹에서 더 많았다.

수면시간은 대사증후군이 없는 그룹이 6.8±1.4시간, 대사증후군이 있는 그룹이 6.7±1.4시간으로 나타났다(p〈0.001). 주관적 건강인지는 두 군 모두 '나쁘다'에 해당하는 비율이 67.1%와 72.4%로 '좋다'에 해당하는 비율보다 더 많았고(p〈0.001), 스트레스인지 정도는 두 군 모두 스트레스를 적게 느끼는 비율이 73.1%와 74.6%로 많이 느끼는 비율보다 더 높은 것으로 확인되었다(p=0.020). 유산소 운동을 하는 비율은 대사증후군이 없는 그룹은 56.1%, 대사증후군이 있는 그룹은 59.5%였으며(p〈0.001), 활동제한 여부는 두 군 모두 없는 비율이 94.6%와 94.8%였다(p〈0.001). 동반질환을 확인하기 위한 만성질환과 암은 두 군 모두 '만성질환이 없음(74.2% vs. 74.2%, p=0.988)'과 '암 진단 없음(95.6% vs. 94.8%, p=0.597)이 다수지만, 두 군간 차이는 없었다. 하지만 만성질환 가족력 여부는 두 군 모두 있다(62.8% vs. 64.7%, p=0.034)고 응답한 비율이 더 높았다(Table 3).

3.2 대사증후군 예측모델의 순열 기능 중요도 (PFI: permutation feature Importance, 이하 PFI)

PFI는 구축된 모델에서 변수들의 상대적 중요도를 의미하는 것으로, 대사증후군 예측에 중요한 변수는 다음과 같다(Table 4). 모델 구축 결과 대사증후군 선별에 가장 중요한 세 가지 변수는 수면시간(1.354), 25kg/m^2

이상 BMI(1.153), 성별(1.118)로 확인되었다.

또한 연령(0.658), 현재 흡연(0.229), 1년간 체중변화(0.199), 만성질환 가족력(0.176), 나쁜 주관적 건강인지(0.158), 1개월에 1잔 이상의 음주(0.132), 만성질환(0.092), 높은 스트레스 인지(0.073), 유산소 신체활동 실천 없음(0.063), 활동제한(0.060), 암(0.016) 순으로 나타났다.

Table 4. Permutation feature Importance of the metabolic syndrome prediction model

feature	Variable importance score
Sleeping time	1.354
Body mass index≥25kg/m^2	1.153
Male	1.118
Age	0.658
Current smoker	0.229
Weight change over 1 year	0.199
Family history of chronic disease	0.176
Poor perceived health status	0.158
Drinking≥1 cup/month	0.132
Chronic disease	0.092
High perceived stress	0.073
No aerobic physical activity	0.063
Physical activity restrictions	0.060
Cancer	0.016

3.3 대사증후군 예측모델의 성능평가

본 연구에서 구축된 대사증후군을 예측모델의 성능평가 결과는 다음과 같다(Table 5). 모델의 AUC는 0.874 (95% CI, 0.874-0.874)로 확인되었고, 실제 데이터와 예측 데이터가 얼마나 일치하는지를 나타내는 정확도(accuracy)는 0.875(95% CI, 0.874-0.876)였다. F1 Score은 0.881(95% CI, 0.880-0.882)이었으며, 특이도(specificity)는 0.829(95% CI, 0.827-0.830), 민감도(sensitivity)는 0.921(95% CI, 0.920-0.922)로 확인되었다. 균형적인 성능 지표인 G-mean은 0.873(95% CI, 0.873-0.874)로 나타났다.

Table 5. Performance of the metabolic syndrome prediction model

Metrics	Performance score	95% CI
AUC	0.874	0.874-0.874
Accuracy	0.875	0.874-0.876
F1 score	0.881	0.880-0.882
Specificity	0.829	0.827-0.830
Sensitivity	0.921	0.920-0.922
G-mean	0.873	0.873-0.874

4. 논의

본 연구는 국민건강영양조사 8기 데이터에서 추출한 대상자의 비침습적 요인들을 활용하여 대사증후군을 선별하는 캣부스트(catboost) 알고리즘 기반 예측모델을 구축한 후, 구축된 모델의 성능을 평가하고, 대사증후군 선별에 영향을 주는 주요 요인들에 대해 확인하였다.

본 연구에서 구축된 캣부스트(catboost)알고리즘 기반 모델은 빠르고 쉽게 활용할 수 있는 비침습적 변수만을 사용하였음에도 0.87 이상의 AUC 성능이 확인되었고, 정확도(accuracy)는 0.875, 조화평균(F1 score) 0.881, 특이도(specificity) 0.829, 민감도(sensitivity) 0.921, G-mean 0.873의 우수한 성능지표를 보였다. 이러한 결과는 본 연구에서 구축한 모델이 대사증후군을 가진 개인을 효과적으로 식별하기 위해 활용될 수 있는 가능성을 나타낸다.

모델에 기여한 변수 중요도 결과 수면시간, 25kg/m^2 이상의 BMI, 성별, 연령, 현재 흡연, 1년간 체중변화, 만성질환 가족력, 나쁜 주관적 건강인지, 1개월에 1잔 이상의 음주, 만성질환, 높은 스트레스 인지, 유산소 신체활동 실천 없음, 활동제한, 암이 확인되었다. 본 연구에서 가장 큰 기여요인으로 확인된 수면시간은 대사증후군의 위험요인으로, 수면시간이 대사증후군과 관련성이 있다는 선행연구들과 일치하였다 [34-36]. 또한, 비만 진단지표로 사용되는 BMI는 대사증후군의 진단기준에 포함되는 허리둘레와 연관이 있는 요인으로 영향력이 높게 나타났다. 이는 비만이 대사증후군의 대표적인 위험인자일 뿐 아니라, BMI가 대사증후군의 유병에 큰 영향을 줄 수 있음을 나타낸다는 선행연구 결과와 일치한다 [5]. 다음은 성별이 중요한 변수 중 상위 수준으로 확인되었다. 이는 성별에 따른 대사증후군의 위험요인 탐색 연구에서 남성의 대사증후군 유병율이 여성보다 높았던 선행연구 결과와 일치하였다 [5,37]. 본 연구에서 구축한 모델의 성능에 순열 기능 중요도의 결과(Table 4)에서 현재 흡연, 1년간 체중변화, 만성질환 가족력, 나쁜 주관적 건강인지, 1개월에 1잔 이상의 음주, 만성질환, 높은 스트레스 인지, 유산소 신체활동 실천 없음, 활동제한, 암의 기여중요도의 합보다 수면시간, 25kg/m^2이상의 BMI, 성별의 중요도가 상대적으로 높게 나타났음을 확인할 수 있다. 이는 향후 대사증후군 예방을 위한 중재 프로그램 개발 시 적절한 수면시간관리, 비만관리 프로그램 등을 포함시킬 필요가 있음을 시사한다. 더불어 이 요인들을 포함하여 더 적은 변수만으로도 더 높은 성능의 예측모

델이 개발될 수 있는 가능성을 시사한다.

본 연구에서 구축한 모델의 종합성능은 기계학습을 활용한 대사증후군을 예측한 모델들에 비해 기계학습을 적용한 다른 모델들 [16]과 비교할 때 비슷하거나 더 높은 성능을 보여준다. 하지만 다른 모델들은 침습적 요인들을 포함하고 있어 비침습적 요인들만으로 구축한 본 연구의 모델은 침습적 검사에 대한 부담없이 전문가 및 개인이 대사증후군을 조기에 식별할 수 있을 것이다.

본 연구는 다음과 같은 제한점이 있다. 첫째, 모델 구축에 활용한 국민건강영양조사는 추적조사 없이 데이터를 수집한 횡단 자료를 활용하여 분석한 연구로 대사증후군의 영향요인 및 인과관계의 방향을 설명하는 데 한계가 있다. 둘째, 국민건강영양조사는 여러 목적을 위해 수행된 복합표본설계이지만, 본 연구는 비침습적 요인만을 활용할 목적으로 수행되었기에 대사증후군에 영향을 주는 모든 변수가 포함되지 못하였으며, 전문가의 자문으로 결정된 변수만을 사용했기 때문에 대사증후군과의 연관성에 대해 의미있게 분석하지 못하였다.

본 연구의 결과는 이러한 제한점에도 불구하고, 우리나라 인구를 대표하는 대규모 인구집단을 대상으로 한 국민건강영양조사 패널 데이터를 활용하여 대사증후군에 영향을 미칠 수 있는 비침습적 요인을 분석하였다는 점에서 의의가 있다. 본 연구에서 구축한 모델을 활용하여 간편한 질문을 통해 대사증후군을 빠르고 정확하게 식별할 수 있으며, 또한 구축한 모델을 통해 확인된 예측인자들은 임상현장이나 지역사회의 개인 및 집단 수준에서 대사증후군을 관리하기 위한 예방프로그램 개발에 활용되기를 기대한다.

5. 결론

본 연구는 국민건강영양조사 제 8기(2019~2021)의 데이터를 기반으로 하여 대사증후군을 비침습적으로 예측하는 모델을 개발하고자 시도되었다. 본 연구에서는 빠르고 쉽게 활용할 수 있는 비침습적 변수만을 사용하여 캣부스트(catboost) 알고리즘을 통해 예측모델을 구축하였고, 0.87 이상의 AUC 성능을 보였다. 이는 의료전문가와 비전문가를 막론하고 빠르고 간편하게 사용할 수 있어 대사증후군 환자를 조기에 식별하고, 침습적 검사에 따른 부담을 줄일 수 있는 방법으로 생각된다.

References

[1] S. N. Grundy, D. Becker, L. T. Clark, R. S. Cooper, M. A. Denke, W. J. Howard et al, "Third report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) Final Report" *Circulation,* pp.3143-3421, Vol.106, No.25, Dec. 2002.
DOI: https://doi.org/10.1161/circ.106.25.3143

[2] S. Mottillo, K. B. Filion, J. Genest, L. Joseph, L. Pilote et al., "The Metabolic Syndrome and Cardiovascular Risk: A Systematic Review and Meta-Analysis," *Journal of the American College of Cardiology*, Vol.56, No.14, pp.1113-1132, Sep. 2010.
DOI: https://doi.org/10.1016/j.jacc.2010.05.034

[3] H. W. Lee, "Current Clinical Practice : Diagnosis and treatment of metabolic syndrome," *The Korean Journal of Medicine*, Vol.71, No.4, pp.463-467, Oct. 2006.

[4] Y. J. Kim, "Metabolic Syndrome and Stroke," *Journal of the Korean Neurological Association,* Vol.23, No.5, pp.585-594, Oct. 2005.

[5] S. E. Lee, H. S. Rhee, "Convergence study to detect metabolic syndrome risk factors by gender difference," *Journal of Digital Convergence*, Vol.19, No.12, pp.477-486, Dec. 2021.
DOI: https://doi.org/10.14400/jdc.2021.19.12.477

[6] Y. Park, "The Association of Metabolic Syndrome and Serum Vitamin E in Korean Adults", *Journal of the Korean Applied Science and Technology*, Vol.40, No.3, pp.385-391, Jun. 2023.
DOI: https://doi.org/10.12925/jkocs.2023.40.3.385

[7] S. Y. Bang, "The Association between Physical Activity and Metabolic Syndrome Index in Middle-aged Adults", *Journal of Information Technology Applications and Management*, Vol.30, No.1, pp.71-80, Feb. 2023.
DOI: https://doi.org/10.21219/jitam.2023.30.1.071

[8] J. Yoo, J. I. Jeong, C. G. Park, S. W. Kang, J. Ahn, "Impact of Life Style Characteristics on Prevalence Risk of Metabolic Syndrome", *Journal of Korean Academy of Nursing*, Vol.39, No.4, pp.594-601, Aug. 2009.
DOI: https://doi.org/10.4040/jkan.2009.39.4.594

[9] H. J. Chae, M. J. Kim, "Prevalence and Related Factors of Metabolic Syndrome among Postmenopausal Adult Women", *Journal of muscle and joint health*, Vol.30, No.3, pp.179-188, Dec. 2023.
DOI: https://doi.org/10.5953/jmjh.2023.30.3.179

[10] K. Yun, "A Study on the Predictive Factors of the Risk of Metabolic Syndrome in Households Living Alone", *The Korean Journal of Health Service Management*, Vol.16, No.1, pp.41-51, Mar. 2022.
DOI: https://doi.org/10.12811/kshsm.2022.16.1.041

[11] J. I. Ramírez-Manent, A. M. Jover, C. S. Martinez, P. Tomás-Gil, P. Martí-Lliteras, et al, "Waist circumference is an essential factor in predicting insulin resistance and

early detection of metabolic syndrome in adults," *Nutrients,* Vol.15, No.2, pp.257-268, Jan. 2023.
DOI: https://doi.org/10.3390/nu15020257

[12] Y. Li, J. Gui, H. Liu, T. Yuan, D. Zhang, et al, "Predicting metabolic syndrome by obesity-and lipid-related indices in mid-aged and elderly Chinese: a population-based cross-sectional study," *Frontiers in Endocrinology*, Vol.14, pp.1-18, Jul. 2023.
DOI: https://doi.org/10.3389/fendo.2023.1201132

[13] W. Xu, Z. Zhang, K. Hu, P. Fang, R. Li et al., "Identifying Metabolic Syndrome Easily and Cost Effectively Using Non-Invasive Methods with Machine Learning Models," *Diabetes, Metabolic Syndrome and Obesity,* Vol.16, pp.2141-2151, Jul. 2023.
DOI: https://doi.org/10.2147/DMSO.S413829

[14] B. Jeong, J. H. Kim, T. Y. Heo, "A Study on the Application and Comparison of Statistical Models and Machine Learnning-based Techniques for Predicting the Onset of Dementia," *Journal of the Korean Data Analysis Society*, Vol.22, No.5. pp.1819-1834, Oct. 2020.
DOI: https://doi.org/10.37727/jkdas.2020.22.5.1819

[15] S. Jeong, M. Lee, S. Yoo, "Machine Learning-based Stroke Risk Prediction using Public Big Data," *Journal of Advanced Navigation Technology*, Vol.25, No.1, pp.96-101, Feb. 2021.
DOI: https://doi.org/10.12673/jant.2021.25.1.96

[16] D. Seong, K. Jeong, S. Lee, Y. Baek, "Metabolic Syndrome Prediction Model for Koreans in Recent 20 Years: A Systematic review," *The Journal of the Korea Contents Association,* Vol.21, No.8, pp.662-674, Aug. 2021.
DOI: https://dx.doi.org/10.5392/JKCA.2021.21.08.662

[17] S. Mohseni-Takalloo, H. Mozaffari-Khosravi, H. Mohseni, M. Mirzaei, M. Hosseinzadeh, "Metabolic syndrome prediction using non-invasive and dietary parameters based on a support vector machine," *Nutrition, Metabolism and Cardiovascular Diseases*, Vol.34, No.1, pp.126-135, Jan. 2024.
DOI: https://doi.org/10.1016/j.numecd.2023.08.018

[18] Y. Park, H. Kang, "Performance Comparison of Various Classification Algorithms of Machine Learning Applications for Predicting Diabetic Nephropathy," *Journal of the Korea Academia-Industrial cooperation Society*, Vol.23, No.7, pp.184-191, Jul. 2022.
DOI: https://dx.doi.org/10.5762/KAIS.2022.23.7.184

[19] T. H. Kim, H. J. Jeong, J. Y. Song, N. Kim, E. M. Lee, "Analysis of Influencing Factors of Suicide Ideation Using Random Forest Model : Focusing on the National Health and Nutrition Examination Survey," *The Korean Data Analysis Society*, Vol.25, No.3, pp.1121-1132, Jun. 2023.
DOI: https://doi.org/10.37727/jkdas.2023.25.3.1121

[20] A. Geron, Hands-on Machine Learning with Scikit-Learn and Tensor Flow: Concepts, Tools, and Techniques to Build Intelligent Systems, p.572, O'Reily Media, 2017.

[21] Y. C. Woo, S. Y. Lee, W. Choi, C. W. Ahn, O. K. Baek, "Trend of Utilization of Machine Learning Technology for Digital Healthcare Aata Analysis," *Electronics and Telecommunications Trends*, Vol.34, No.1, pp.98-110, Feb. 2019.
DOI: https://dx.doi.org/10.22648/ETRI.2019.J.340109

[22] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, A. Gulin, "CatBoost: unbiased boosting with categorical features," *32nd Conference on Neural Information Processing Systems*, Montréal, Canada. *31,* pp.1-11, Dec. 2018.

[23] J. Kim, J. Park, "Evaluation of Multi-classification Model Performance for Algal Bloom Prediction Using CatBoost," *Journal of Korean Society on Water Environment*, Vol.39, No.1, pp.1-8, Jan. 2023.
DOI: https://doi.org/10.15681/KSWE.2023.39.1.1

[24] E. Expert Panel on Detection, "Executive Summary of the Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III)," *Jama*, Vol.285, Vo.19, pp.2486-2497, May 2001.
DOI: https://doi.org/10.1001/jama.285.19.2486

[25] Y. S. Yoon, S. W. Oh, "Optimal Waist Circumference Cutoff Values for the Diagnosis of Abdominal Obesity in Korean Adults," *Endocrinology and Metabolism*, Vol.29, No.4, pp.418-426, Dec. 2014.
DOI: https://doi.org/10.3803/EnM.2014.29.4.418

[26] D. M. Harris, S. L. Harris, Digital Design and Computer Architecture, p.720, Morgan Kaufmann, 2012.

[27] K. Lee, J. Lim, K. Bok, J. Yoo, "Handling Method of Imbalance Data for Machine Learning: Focused on Sampling," *The Journal of the Korea Contents Association*, Vol.19, No.11, pp.567-577, Nov. 2019.
DOI: https://doi.org/10.5392/JKCA.2019.19.11.567

[28] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of artificial intelligence research*, Vol.16, pp.321-357, Jun. 2002.
DOI: https://doi.org/10.1613/jair.953

[29] A. V. Dorogush, V. Ershov, A. Gulin, "CatBoost: gradient boosting with categorical features support," arXiv preprint arXiv:1810.11363, pp.1-7, Oct. 2018.
DOI: https://doi.org/10.48550/arXiv.1810.11363

[30] J. Seo, N. Kang, "Exploration of Factors on Pre-service Science Teachers' Major Satisfaction and Academic Satisfaction Using Machine Learning and Explainable AI SHAP," *Journal of Science Education*, Vol.47, No.1, pp.37-51, Apr. 2023.
DOI: https://dx.doi.org/10.21796/jse.2023.47.1.37

[31] S. M. Lundberg, S. I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in neural information processing systems*, 30, Nov, 2017.
DOI: https://doi.org/10.48550/arXiv.1705.07874

[32] M. Kubat, S. Matwin, "Addressing the Curse of

Imbalanced Training Sets: One-Sided Selection," *In Icml,* Vol.97, No.1, pp.179-186, Jul. 1997.

[33] T. J. DiCiccio, B. Efron, "Bootstrap confidence intervals," *Statistical science*, Vol.11, No.3, pp.189-228, Aug. 1996. DOI: https://dx.doi.org/10.1214/ss/1032280214

[34] Y. Park, "Effects of Sleep Duration and Quality on Prevalence of Metabolic Syndrome and Metabolic Syndrome Components in Korean Blue-collar Workers," *Korean Journal of Occupational Health Nursing*, Vol.29, No.1, pp.69-77, Feb. 2020. DOI: https://doi.org/10.5807/kjohn.2020.29.1.69

[35] H. Park, "The Effects of Shift Work and Hours of Sleep on Metabolic Syndrome in Korean Workers," *Korean Journal of Occupational Health Nursing*, Vol.25, No.2, pp.96-107, May 2016. DOI: http://dx.doi.org/10.5807/kjohn.2016.25.2.96

[36] B. G. Lee, J. Y. Lee, S. A. Kim, D. M. Son, O. K. Ham, "Factors associated with Self-Rated Health in Metabolic Syndrome and Relationship between Sleep Duration and Metabolic Syndrome Risk Factors," *Journal of Korean Academy of Nursing*, Vol.45, No.3. pp.420-428, Jun. 2015. DOI: http://dx.doi.org/10.4040/jkan.2015.45.3.420

[37] E. O. Park, K. J. Kang, "The influences of Gender and Obesity on the Metabolic Syndrome among Korean Adults: Based on Korea National Health and Nutrition Examination Survey," *Journal of the Korea Academia-Industrial cooperation Society*, Vol.22, No.9, pp.692-699, Sep. 2021. DOI: https://doi.org/10.5762/KAIS.2021.22.9.692

유 보 현(Bo Hyun Yu) [정회원]

- 2021년 2월 : 전남대학교 일반대학원 간호학과 (간호학 석사)
- 2024년 2월 : 전남대학교 일반대학원 간호학과 (간호학 박사수료)
- 2001년 3월 ~ 2024년 2월 : 전남대학교병원 간호사
- 2024년 2월 ~ 현재 : 빛고을전남대학교병원 진료비심사팀장

〈관심분야〉
간호, 의/생명공학

최 아 라(Ah-Ra Choi) [정회원]

- 2018년 2월 : 전남대학교 일반대학원 보건학 협동과정 (보건학 석사)
- 2022년 3월 ~ 현재 : 전남대학교 일반대학원 간호학과 박사과정
- 2012년 6월 ~ 현재 : 전남대학교병원 연구 간호사

〈관심분야〉
간호, 의/생명공학

김 태 훈(Tae-Hun Kim) [정회원]

- 2024년 2월 : 전남대학교 일반대학원 간호학과 (간호학 박사)
- 2023년 6월 ~ 현재 : 전라남도 공공보건의료지원단 연구원

〈관심분야〉
정신간호, 빅데이터, 인공지능