Journal of the Korea Industrial Information Systems Research Vol.27 No.5, Oct. 2022 :73-82
http://dx.doi.org/10.9723/jksiis.2022.27.5.073 ISSN:1229-3741

머신러닝 기반 생애주기별 고혈압 위험 요인 분석+

(Analysis of Hypertension Risk Factors by Life Cycle Based on Machine Learning)

강 성 안[1], 김 소 희[2], 류 민 호[3]*
(SeongAn Kang, SoHui Kim, and Min Ho Ryu)

요 약 고혈압과 같은 만성질환은 발병의 원인은 다양한 요인들이 복합적으로 작용하기 때문에 생애주기에 따라 차별화된 관리가 필요하다. 본 연구는 머신러닝을 이용해 고혈압 발병에 영향을 미치는 요인들의 생애주기별로 차이를 분석한다. 이를 위해, 질병관리청의 국민건강영양조사 데이터에 대한 전처리 및 변수 선택 과정을 거쳐 총 35개의 변수를 활용했다. 분석결과, 트리기반 머신러닝 모델 중 XGBoost가 중년과 노년 모두 예측 성능이 높은 모델로 나타났다. 변수중요도를 통해 도출된 생애주기별 고혈압 위험요인을 살펴보면 중년의 경우 개인특성 요인, 유전적 요인, 영양섭취 요인이 고혈압 위험요인으로 나타났고, 노년의 경우 영양섭취 요인, 식생활 요인, 생활습관 요인이 고혈압 위험요인으로 도출되었다. 본 연구 결과는 생애주기별 고혈압 관리에 유용한 기초자료로 사용될 수 있을 것으로 기대된다.

핵심주제어: 고혈압, 트리기반 머신러닝, 생애주기, 변수중요도

Abstract Chronic diseases such as hypertension require a differentiated approach according to age and life cycle. Chronic diseases such as hypertension require differentiated management according to the life cycle. It is also known that the cause of hypertension is a combination of various factors. This study uses machine learning prediction techniques to analyze various factors affecting hypertension by life cycle. To this end, a total of 35 variables were used through preprocessing and variable selection processes for the National Health and Nutrition Survey data of the Korea Centers for Disease Control and Prevention. As a result of the study, among the tree-based machine learning models, XGBoost was found to have high predictive performance in both middle and old age. Looking at the risk factors for hypertension by life cycle, individual characteristic factors, genetic factors, and nutritional intake factors were found to be risk factors for hypertension in the middle age, and nutritional intake factors, dietary factors, and lifestyle factors were derived as risk factors for hypertension. The results of this study are expected to be used as basic data useful for hypertension management by life cycle.

Keywords: Hypertension, Tree-based Machine Learning, Lifecycle, Feature Importance

* Corresponding Author: ryumh12@dau.ac.kr
+ 이 논문은 동아대학교 교내 연구과제 지원을 받아 수행됨
Manuscript received August 05, 2022 / revised September 29, 2022 / accepted October 13, 2022

1) 동아대학교 경영정보학과, 제1저자
2) 동아대학교 경영정보학과, 제2저자
3) 동아대학교 경영정보학과, 교신저자

1. 서 론

고혈압이란 수축기 혈압이 140mmHg 이상이거나 이완기 혈압이 90mmHg 이상일 때를 말하며, 국내 사망 원인 2, 3위인 심근경색, 협심증과 같은 심혈관 질환의 주요 요인이다(Lee, 2013). 고혈압은 주요 만성질환으로 유병률이 높고 합병증으로 인한 사망률 또한 매우 높은 질환으로 알려져 있다(Samadian et al., 2016). 국내 고혈압 환자 수는 지속적으로 증가하는 추세로, 2014년 약 1,045만 명에서 2021년 약 1,375만 명으로 약 31.58% 증가하였다(KSH, 2022).

고혈압, 당뇨와 같은 만성질환의 경우 연령 또는 생애주기에 따라 유병률이 다르며, 신체적 특성, 삶의 질, 건강행태 등의 특성적인 차이가 있을 수 있다. 따라서 고혈압 관리를 위해서는 차별적인 접근이 필요하다(Lee and Cho, 2016; Kim and Min, 2020). 또한, 고혈압의 경우 단일 요인이 원인이 되어 발생하는 병이 아니라, 유전적 요인, 식습관, 영양섭취, 운동 부족, 스트레스 등과 같은 다양한 요인들이 복합적으로 작용하여 발생하기 때문에 고혈압 유병에 영향을 미치는 요인을 분석함에 있어 이러한 다양한 측면의 변수들을 종합적으로 고려할 필요가 있다(Lee, 2018; KSH, 2018).

최근 머신러닝, 딥러닝 등 기계학습 및 알고리즘의 발전으로 이를 활용하여 다양한 질병들을 예측하고 진단하기 위한 연구들이 활발하게 진행되고 있다. 머신러닝을 이용한 고혈압 진단 및 예측 모델 연구들에서는 소득·가구원 수·성별·결혼 여부 등의 사회경제적 요인, 음주·흡연·근력운동·유산소 운동 등과 같은 건강행태, 가족력, 건강지표 및 식생활 요인 등 다양한 요인들을 사용하고 있다. 하지만 기존의 연구에서는 고혈압이 복합적인 요인으로 발병하는 병임에도 불구하고 사회인구학적 요인, 식생활 요인 등 개별적인 요인만을 독립변수로 활용하여 분석을 진행한 경우가 대부분이다.

따라서 본 연구는 기존 선행연구에서 밝혀진 고혈압에 영향을 미치는 요인들을 유형에 따라 분류하고, 트리기반 머신러닝 알고리즘인 Random Forest(RF), XGBoost(XGB), LightGBM(LGB) 등을 사용하여 고혈압을 예측하는 분류기를 생성하여, 변수중요도 기반으로 생애주기별 고혈압의 위험인자를 파악하여 제시하는 것을 목적으로 한다.

본 연구는 다음과 같이 구성된다. 2장에서 관련 문헌 연구를 설명하고, 3장에서는 분석에 사용된 데이터와 처리 방법 및 분석 방법에 대해 설명하고, 4장에서는 분석결과를 제시한다. 5장에서는 연구의 최종 결론을 제시한다.

2. 이론적 배경

2.1 고혈압에 영향을 미치는 요인

영양소 섭취량, 식생활 및 생활습관, 인구 사회학적 특성, 가족력, 임상학적 특성 등 다양한 유형의 변수에 대해서 고혈압에 미치는 영향이 유효한 변수를 찾아내는 연구가 활발하게 진행되고 있다.

Gu et al.(2012)은 국민건강영양조사 자료를 이용하여 한국인 성인을 대상으로 영양섭취기준 대비 영양소 섭취 수준에 따른 고혈압 유병 위험도를 분석하였다. 이를 통해 고혈압 유병 위험에 대한 식생활 및 생활습관 요인의 영향력을 규명하였으며, 총 12가지 영양소 중 단백질, 인, 칼륨, 철, 티아민, 리보플라빈, 나이아신 등의 영양소 섭취량이 상대적으로 높은 사람들은 고혈압 유병 위험이 낮았다. 또한, 나트륨 섭취량이 2000mg~4900mg 이하의 섭취 수준을 보인 대상자는 고혈압 유병 위험이 상대적으로 낮다는 것을 밝혔다. Kim(2019)은 제6차 고령화 연구패널 조사를 활용하여 고혈압 유병률에 영향을 미치는 위험요인에 대한 연구를 진행하였다. 65세 미만 성인과 65세 이상 노인으로 연구군을 구분하였으며, 주관적 건강상태, 당뇨병 유무, 심장질환 유무, BMI 등의 요인이 위험요인으로 나타났다. 또한, 연구군 별로 고혈압에 영향을 미치는 위험요인이 다른 것을 밝혔다. Byeon et al.(2015)은 생활습관을 중심으로 고혈압의 발생 위험 확률을 예측할 수 있는 모형을 개발하기 위해 사회인구학적 특성, 임상 건강 행위 특성, 입원과 관련된 내역 등을 활용하여 중년 고혈압을 예측하는 연구를 진행하였다.

고혈압의 경우 다양한 요인이 복합적으로 작용하는 질환임에도 불구하고 선행연구들은 단일 혹은 몇 가지의 요인을 설명변수로 활용하여 분석을 진행하고 있어, 고혈압에 영향을 미치는 다양한 요인들을 모두 포함하지 못하는 한계를 가진다. 본 연구에서는 이러한 한계를 보완하여 고혈압에 영향을 미치는 다양한 요인들을 유형화해 종합적인 분석을 진행한다.

2.2 머신러닝을 활용한 연구

머신러닝 및 딥러닝 등 기계학습 및 알고리즘의 발전으로 이를 활용하여 고혈압을 포함한 다양한 질병들을 예측하고 진단하기 위한 연구들이 활발하게 진행되고 있다.

Lim(2018)은 국민건강영양조사 자료에 포함되어 있는 인구사회학적 특성, 개인과거병력, 건강설문, 건강검진 자료를 활용하여 심근경색증 및 협심증 발생을 예측하고, 주요요인을 찾기 위해 B-LASSO 모형을 제안했다. Lee and Lee (2020)는 폐경 이후 여성의 골다공증 유병 여부를 예측하는 연구를 위해 트리기반 머신러닝 모델을 이용하였다. 연구 결과 모든 모델에서 나이는 골다공증 유병 여부를 예측하는데 가장 큰 영향력을 보인 변수이며, 가장 좋은 성능을 보인 XGBoost를 이용하여 변수선택 후 모델의 성능을 높였다. Hong el al.(2022)은 농촌진흥청의 소비자 패널데이터와 건강보험공단의 진료 데이터를 연계하여 식품 소비 특성을 통한 대사성 질환자군과 대조군을 나누어 식품소비에 따른 대사성 질환 분류모델을 비교하는 연구를 진행했다. 로지스틱 회귀, 의사결정나무, XGBoost를 활용하여 분류모델을 생성하고 모델의 예측 정확도를 비교하였으며, 가장 높은 성능을 보인 모델은 XGBoost였다.

머신러닝을 사용하여 고혈압을 진단 및 예측하는 연구들을 살펴보면, Lee(2021)는 사회인구학적 변수들만을 사용하여 한국인 성인 고혈압 예측 모델을 제안했다. 해당 연구에서는 wrapper-based feature subset selection method, Naive Bayes model 등을 사용하여 활용성이 높은 고혈압 예측 모델을 제시했다. 또한 남녀 각각에 대한 고혈압 예측 모델을 제시하였으며, 전체적으로 남성보다는 여성에서의 고혈압 예측 성능이 높은 것으로 나타났다. LaFreniere et al.(2016)은 캐나다 1차 의료 감시 네트워크(CPCSSN) 데이터를 활용하여 고혈압을 예측하는 연구를 진행했다. 환자의 현재 건강상태, 의료 기록 및 인구통계학적 요인들을 중 고혈압에 중요한 위험요인들을 추출하였다. 추출된 요인들을 활용하여 인공신경망 모델인 ANN을 사용하여 개인의 고혈압 여부를 예측하는 모델을 제시하였다.

AIKaabi et al.(2020)은 로지스틱 회귀, 의사결정 나무, 랜덤 포레스트 등의 머신러닝 기법을 활용하여 고혈압 발병을 예측하기 위한 모델을 구축하고, 모델 간 비교를 목적으로 연구를 진행하였다. 연구를 위해 카타르 바이오 뱅크의 데이터를 사용하였으며, 사회인구학적 변수, 생활 방식 변수를 활용하여 고혈압을 예측하는 모델을 구축하였다. 연구 결과 가장 높은 정확도를 보인 모델은 랜덤 포레스트였으며, 성별, 콜레스테롤 유병 여부, 당뇨병 유병 여부, 허리둘레 등의 요인이 고혈압 예측에 중요한 요인으로 나타났다.

고혈압 예측 및 위험요인과 관련한 선행연구를 살펴보았을 때, 대부분의 연구에서는 성별을 기준으로 연구대상의 차이를 비교하였다. 또한 고혈압의 경우 생애주기에 따라 유병률이 다르지만(Lee and Cho, 2016; Kim and Min, 2020) 이러한 차이를 고려하지 않고 고혈압 유병 여부를 예측하거나 고혈압 위험요인을 분석한 연구가 대부분이었다. 따라서 본 연구에서는 생애주기 별로 고혈압에 영향을 미치는 요인을 찾고, 트리기반 머신러닝 알고리즘을 활용하여 고혈압을 예측하는 분류기를 생성하고, 변수중요도를 기반으로 생애주기별 고혈압의 위험 인자를 파악하는 것을 목적으로 한다.

3. 데이터 및 연구 방법

3.1 분석자료 및 연구대상

본 연구에서는 질병관리청에서 제공하는 국민건강영양조사 원시 데이터를 활용하였다. 수집

된 데이터의 범위는 2016년부터 2020년까지이며, 5년간의 데이터를 통합하여 분석에 사용하였다. 국민건강영양조사 데이터는 횡단적 데이터이며, 건강설문조사, 검진조사, 영양조사 등 3가지 부문으로 구성되어 있다.

고혈압을 예측하기 위해 사용된 설명변수는 고혈압에 영향을 미치는 요인과 관련된 선행연구를 참고하였으며, 관련 분야의 전문가들과의 협의를 통해 '유전적 요인', '개인특성', '생활습관 요인', '식생활 요인', '영양섭취 요인'으로 구분되는 47개의 특성을 추출하였다.

분석을 위해 추출된 데이터의 수는 22,914명이며, 만 19세 이하의 참가자 224명은 분석에서 제외하였다. 또한 생애주기별 고혈압 위험요인을 확인하기 위해 데이터를 생애주기 기준에 따라 나누었다. 생애주기의 경우 연구 분야 및 연구자에 따라 분류하는 기준이 다르다. 본 연구에서는 건강행태학적 특성을 반영한 연구를 참고하여 20~39세는 청년기, 40~64세는 중년기, 65세 이상은 노년기로 구분하였다(An, 2010; Kim and Jung, 2019). 본 연구에서는 사용되는 청년 데이터의 경우 정상군과 고혈압군 간의 높은 편향을 보였다. 이러한 목표변수의 편향은 머신러닝 모델이 제대로 학습되지 않고, 새로 입력되는 데이터의 예측 성능이 저하되는 문제가 발생할 수 있어 청년 데이터는 본 연구에서 제외하였다(Mun et al., 2016; Yoon and Bang 2021).

Table 1 Number of data used for analysis

Life Cycle	Total	Normal	Hypertensive
Youth (20~39)	6,296	5,683	613
Middle age (40~64)	10,175	5,855	4,320
Old age (more than 65 age)	6,443	1,308	5,135

3.2 데이터 전처리

본 연구에서 사용되는 국민건강영양조사의 경우 변수들에 존재하는 결측치의 비중이 높았다. 일반적으로 결측치를 제거하고 분석을 진행하지만, 결측치 제거 시 데이터의 수가 적어져 학습할 수 있는 데이터가 부족하게 되면 올바른 예측을 할 수 없다. 따라서 본 연구에서는 결측치를 범주형 변수의 경우 최빈값, 수치형 변수는 평균값으로 대치하는 보간법을 사용하여 처리했다. 또한, 수치형 변수들의 상관관계를 확인하여 높은 상관관계를 보이는 변수들은 연구에서 제외하고, 수치형 변수들의 범위와 크기 차이의 문제를 해결하기 위해 표준화 작업을 진행하였다. 전처리를 완료한 후 최종적으로 선택한 변수는 총 35개이며, 이를 유형화한 결과는 Table 2와 같다.

3.3 분류 모형 구축

전처리를 수행한 후 훈련 데이터(Train Data)와 검증 데이터(Test Data)를 7:3의 비율로 나누었고, 훈련 데이터를 활용하여 분류 모형을 학습하였다. 머신러닝 분류는 훈련 데이터의 입력변수와 목표변수인 고혈압 여부를 동시에 입력받아 학습하는 지도학습(Supervised Learning) 방법으로 모형을 구축하였다. 구축된 모형으로 결과를 추론할 때, 훈련 데이터셋을 대상으로 고혈압 유무를 예측하고 실제값과 비교하여 분류 모형을 평가하였다.

본 연구에서는 분류 알고리즘으로 다수의 의사결정 트리를 학습하여 예측하는 앙상블 모델인 Random Forest, XGBoost, LightGBM을 사용하였다. Random Forest란 전체 특성 중 무작위로 일부만 선택해 하나의 결정 트리를 만들고 선택된 설명변수의 집합 중 가장 최적의 결과를 내는 알고리즘으로 모든 변수를 사용하여 최적의 결과를 내는 의사결정 나무의 단점인 과적합 문제를 보완한 방법이다(Kim, 2019). XGBoost는 랜덤 포레스트와 다르게 다수의 트리 결합보다는 트리의 오차를 보완하는 방식으로 학습하는 방식인 그레디언트 부스팅(Gradient Boosting) 모델에서 발전한 모델이다(Chen et al, 2016; Son, 2021). LightGBM은 XGBoost와 같은 그래디언트 부스팅을 기반으로 하는 알고리즘으로, 다른 GBM 알고리즘 보다 분석을 수행하는 속도가 빠르다는 장점이 있는 알고리즘이다

Table 2 Final Variable

Factor	Explanatory Variable
Genetic Factors (4)	Chronic disease doctor diagnosis family history, Hypertension doctor diagnosis(father), Hypertension doctor diagnosis(mother), Hypertension doctor diagnosis(siblings)
Personal Characteristics (7)	Sex, Personal Income, Household income, Household Member, Marital Status, Body Mass Index(BMI), Diabetes
Lifestyle (5)	Smoking, Drinking, strength training day(1 week), Cardiovascular exercise, Fasting time
Dietary Factors (3)	Breakfast frequency per week, Lunch frequency per week, Dinner frequency per week
Nutritional Factors (16)	Water, Protein, Fat, Cholesterol, Carbohydrate, Dietary fiber, Sugar, Calcium, Ferrum, Sodium, Vitamin A, Thiamine, Riboflavin, Niacin, Folate, Vitamin C

(Guolin Ke et al, 2016). 이러한 트리 기반 모델은 비선형적이고 비모수적인 모형으로, 데이터에 대한 가정이 없이 자유롭게 사용할 수 있으며, 종속변수에 영향을 미치는 요인들을 직관적으로 탐색할 수 있다는 장점이 있다(Lee and Lee, 2020; Oh et al., 2021).

3.4 평가 지표

본 연구에서 사용된 데이터는 정상군의 데이터가 고혈압 군의 데이터보다 많은 편향이 존재하기 때문에 학습된 모형의 성능을 평가하기 위해 혼동행렬(Confusion Matrix)을 통한 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1 score를 성능평가지표로 사용하였다. Fig. 1은 평가 지표에 대한 수식을 나타낸 표이다.

		Actual Values	
		True	False
Predict Values	True	True Positive	False Positive
	False	False Negative	True Negative

Indicator	Formula
Accuracy	(TP+TN)/(TP+TN+FP+FN)
Precision	TP/(TP+FP)
Recall	TP/(TP+FN)
F1 Score	2*Precision*Recall/ (Precision+Recall)

Fig. 1 Confusion Matrix / Performance Indicators

4. 분석결과

4.1 분류 모형 평가

Table 3은 35개의 변수를 활용하여 각 모델별로 중년 및 노년의 고혈압 유병 여부를 예측한 결과이다. 중년 데이터의 경우 정확도 및 고혈

Table 3 Detect Model Performance Indicators

Life Cycle	Middle age			Old age		
Model	Random Forest	XGBoost	LightGBM	Random Forest	XGBoost	LightGBM
Accuracy	70.03	70.95	70.46	78.89	79.41	78.22
Precision	55.63	58.88	60.03	98.05	98.18	95.91
Recall	68.00	68.30	66.96	80.00	80.35	80.50
F1-Score	61.18	63.24	63.30	88.10	88.38	87.53

Table 4 The Importance of Variables by Life Cycle

Middle age						
No	Random Forest		XGBoost		LightGBM	
	Feature	Weight	Feature	Weight	Feature	Weight
1	BMI	0.1430 ± 0.0112	BMI	0.0952 ± 0.0104	BMI	0.0878 ± 0.0079
2	Hypertension doctor diagnosis (siblings)	0.0541 ± 0.0080	Sex	0.0393 ± 0.0079	Sex	0.0363 ± 0.0098
3	Sex	0.0510 ± 0.0095	Diabetes	0.0292 ± 0.0100	Diabetes	0.0346 ± 0.0048
4	Diabetes	0.0305 ± 0.0104	Hypertension doctor diagnosis (siblings)	0.0252 ± 0.0073	Hypertension doctor diagnosis (siblings)	0.0253 ± 0.0104
5	Household members	0.0271 ± 0.0059	Water	0.0129 ± 0.0068	Water	0.0173 ± 0.0018
6	Hypertension doctor diagnosis (mother)	0.0221 ± 0.0032	Sugar	0.0093 ± 0.0043	Riboflavin	0.0106 ± 0.0066
7	Sugar	0.0199 ± 0.0061	Household members	0.0072 ± 0.0050	Vitamin A	0.0085 ± 0.0045
8	Chronic disease doctor diagnosis family history	0.0184 ± 0.0040	Household Income	0.0068 ± 0.0052	Household members	0.0079 ± 0.0020
9	Smoking	0.0177 ± 0.0054	Riboflavin	0.0060 ± 0.0019	Dietary fiber	0.0062 ± 0.0071
10	Household Income	0.0174 ± 0.0020	Vitamin A	0.0057 ± 0.0055	Fat	0.0061 ± 0.0070

Table 5 The Importance of Variables by Life Cycle in Old age

Old age						
No	Random Forest		XGBoost		LightGBM	
	Feature	Weight	Feature	Weight	Feature	Weight
1	BMI	0.0011 ± 0.0010	Riboflavin	0.0038 ± 0.0023	Riboflavin	0.0099 ± 0.0033
2	Riboflavin	0.0007 ± 0.0003	BMI	0.0037 ± 0.0053	BMI	0.0048 ± 0.0038
3	Sugar	0.0006 ± 0.0004	Personal Income	0.0023 ± 0.0017	Thiamine	0.0034 ± 0.0021
4	Thiamine	0.0005 ± 0.0005	Cardiovascular Exercise	0.0016 ± 0.0022	Cardiovascular Exercise	0.0023 ± 0.0018
5	Water	0.0005 ± 0.0003	Breakfast frequency per week	0.0013 ± 0.0012	Lunch frequency per week	0.0019 ± 0.0011
6	Sodium	0.0004 ± 0.0005	Thiamine	0.0013 ± 0.0018	Vitamin C	0.0019 ± 0.0017
7	Potassium	0.0004 ± 0.0003	Household Income	0.0010 ± 0.0008	Smoking	0.0017 ± 0.0007
8	Household Income	0.0004 ± 0.0004	Hypertension doctor diagnosis (father)	0.0009 ± 0.0004	Sodium	0.0017 ± 0.0019
9	Diabetes	0.0004 ± 0.0003	Lunch frequency per week	0.0008 ± 0.0008	Personal Income	0.0013 ± 0.0018
10	Smoking	0.0003 ± 0.0004	Folate	0.0008 ± 0.0023	Sex	0.0013 ± 0.0010

압을 고혈압으로 잘 예측한 지표인 Recall 값 모두 XGBoost가 높게 나왔기 때문에 중년 데이터의 적합한 모델은 XGBoost으로 볼 수 있다. 노년 데이터의 경우 정상군보다 고혈압군으로 데이터가 편향되어 모델의 정밀도(Precision)와 재현율(Recall)이 정확도에 비해 높은 값을 보였다. 노년의 경우 정확도 및 F1 Score는 XGBoost가 다른 모델에 비해 높았으며, Recall 값의 경우 세 가지 모델 모두 큰 차이를 보이지 않았지만, 정밀도와 재현율의 조화평균 값인 F1 Score의 값이 높은 XGBoost가 적합한 모델이라고 할 수 있다.

4.2 변수중요도

학습이 완료된 모델을 활용하여 고혈압의 위험인자를 확인하기 위해 변수중요도를 측정하였다. 변수중요도를 측정하기 위해 Permutation Feature Importance(PFI)를 활용하였다. PFI는 각 변수가 모델의 의사결정에 공헌한 정도를 상대적으로 정량화하는 알고리즘으로, 모델의 성능이 감소하는 수치를 계산하여 변수의 중요도를 측정하는 방법으로 계산이 빠르고, 사용범위가 넓으며, 일관된 변수의 중요도를 측정할 수 있다는 장점이 있다(Jeon el al., 2021; Oh el al., 2022). Table 4와 Table 5는 각각 중년과 노년을 대상으로 PFI를 사용하여 고혈압 분류에 공헌도(Weight)가 가장 높은 상위 10개의 변수를 모델별로 나타낸 표이다. 중년의 경우 모든 모델에서 공통적으로 BMI가 가장 중요한 변수로 나타났다. 또한 고혈압 의사 진단여부(형제자매), 성별, 당뇨병 유병 여부, 수분 섭취가 상위에 배치되었다. 또한 상위 변수들 간 순서의 차이만 존재할 뿐 비슷한 변수들이 상위에 나타났다. 모델 간 차이점으로는 Random Forest의 경우 만성질환 가족력 여부, 흡연 여부가 고혈압 위험요인으로 나타났으며, XGBoost와 LightGBM의 경우 영양섭취 요인이 상위에 나타났지만, 리보플라빈, 비타민 A, 식이섬유, 지방 등으로 차이를 보였다.

노년의 경우 XGBoost, LightGBM 모델에서는 리보플라빈이, Random Forest에서는 BMI가 중요한 변수로 나타났다. 노년의 경우 모델 간 요인의 차이가 다양하게 나타났으며, XGBoost와 LightGBM의 경우 유산소 신체활동 변수 및 식생활 요인의 변수가 상위에 나타났고 Random Forest의 경우 영양섭취 요인이 상위에 나타났다.

5. 결 론

본 연구는 생애주기별 고혈압 위험요인을 도출하여 제시하는 것을 목적으로 한다. 이를 위해 트리기반 머신러닝 알고리즘을 사용하여 고혈압을 예측하는 머신러닝 모델을 구축하였다. 구축된 모델을 바탕으로 도출된 변수중요도를 기반으로 생애주기별 고혈압의 위험인자를 파악하였다.

중년과 노년 모두 XGBoost 기반의 모델이 고혈압 환자를 분류하는 성능이 우수하게 나타났다. 기계학습 모델을 이용하여 생애주기별 고혈압 위험요인을 도출한 결과, 중년의 고혈압 위험요인의 경우 사용된 XGBoost, Random Forest, Light GBM 모두 상위에 분포한 요인들이 일치하였다. 또한 성별·당뇨병 유병 여부·가구원 수·가구 소득분위·개인 소득분위 등과 같은 개인특성 요인, 당·비타민 A·지방·리보플라빈·수분·식이섬유 등과 같은 영양섭취 요인 그리고 만성질환 가족력 여부·고혈압 의사진단 여부(형제자매) 등과 같은 유전적 요인이 중년 고혈압의 위험요인으로 도출되었다. 노년의 경우 중년과 마찬가지로 BMI가 고혈압 위험요인으로 도출되었다. 하지만 노년의 경우 리보플라빈 또한 고혈압 위험요인으로 상위에 도출되었다. 뿐만 아니라 당·염분·티아민·엽산·칼륨 등 영양섭취 요인이 다양하게 도출되었다. 또한 1주일 아침식사 빈도·1주일 점심식사 빈도 등 식생활 요인이 주요 위험요인으로 도출되었으며, 유산소 신체활동 일수·흡연 여부 등과 같은 생활습관 요인이 고혈압 위험요인으로 도출되었다.

분석 결과, 중년과 노년 모두 공통적으로 BMI가 높은 주요 위험인자로 도출되었으며, 영양섭취 요인 또한 공통적인 위험인자인 것을 확

인할 수 있었다. 하지만 중년의 경우 유전적 요인 및 개인특성 요인이 위험인자로 도출된 반면, 노년의 경우에는 식생활 요인 및 생활습관 요인이 주요 인자로 도출된 것을 확인하였다. 이처럼 생애주기에 따라 고혈압의 위험요인의 요소가 다르므로 생애주기별 고혈압 관리를 위한 차별화된 접근이 필요할 것이다.

본 연구에서는 중년, 노년 간 고혈압 위험요인의 차이점을 도출하였으며, 해당 결과를 바탕으로 생애주기별 고혈압 관리에 대한 유용한 기초자료로 사용될 수 있다는 점에서 의의를 가진다.

그럼에도 불구하고 본 연구는 다음과 같은 한계점을 가진다. 첫째, 본 연구에서 사용된 국민건강영양조사의 경우 단면적 조사이기 때문에 해당 데이터만을 가지고는 고혈압의 위험인자와의 인과관계를 설명하는데 한계가 있다. 둘째, 선행연구를 통해 고혈압에 영향을 미치는 요인 및 위험요인을 선정하였지만, 제공하는 모든 변수가 포함되지 못해 고혈압과의 연관성에 대해 의미 있는 분석을 하지 못한 한계를 가진다. 셋째, 청년 데이터의 경우 데이터의 불균형으로 인해 연구에 포함할 수 없었다. 향후 이러한 불균형 문제를 해결하기 위해 다양한 샘플링 방법을 시도해 보거나, 관련 데이터 확보가 이루어진다면 이러한 문제점을 해결할 수 있을 것이다.

References

AlKaabi, L. A., Ahmed, L. S., Al Attiyah, M. F., & Abdel-Rahman, M. E. (2020). Predicting hypertension using machine learning: Findings from Qatar Biobank Study. *Plos one*, 15(10), e0240370.

An. H. M. (2010). Factors of health related quality of life of Korea male and female adults according to life cycle : by using 4th national health and nutrition examination survey, Master's Thesis, Graduate School of YonSei University, Seoul, Korea.

Byeon, H. W. and Cho, S. H. (2015). The Predictive Modeling of Middle-aged Hypertension using Integrated Method of Decision Tree and Neural Network, *Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology*, 23(4). 13-28.

Chen. T. and Guestrin. C. (2016). XGBoost: A Scalable Tree Boosting System, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

Gu, S., Kim, Y. O., Kim, M. K., Yoon, J. S. and Park. K. (2012). Nutrient Intake, Lifestyle Factors and Prevalent Hypertension in Korean Adults: Results from 2007-2008 Korean National Health and Nutrition Examination Survey, *Korean Journal of community Nutrition*, 17(3), 329-340.

Hong, J. H., Lee, K. H., Lee, H. R., Cheong, H. S. and Cho, W. S. (2022). Metabolic Diseases Classification Models according to Food Consumption using Machine Learning, *The Journal of the Korea Contents Association*, 22(3), 354-360.

Jeon. W. J., Lee.1 Y. B. and Geum. Y. J. (2021). Airline Service Quality Evaluation Based on Customer Review Using Machine Learning Approach and Sentiment Analysis, *The Journal of Society for e-Business Studies*, 26(4), 15-36.

Ke. G., Meng. Q., Finley. T., Wang. T., Chen. W., Ma. W., Ye. Q. and Liu. T. Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree, *Advances in Neural Information Processing Systems*, 30.

Kim, K. Y. (2019). Risk factors for hypertension in elderly people aged 65 and over, and adults under age 65, *Journal of Korea Academia-Industrial cooperation Society*, 20(1), 162-169.

Kim. H. J. and Min. E. S. (2020). Health behaviors and quality of life by life cycle of hypertensive patients, *Journal of Convergence for Information Technology (JCIT)*, 10(7), 58-66.

Kim. P. J. (2019). An Analytical Study on Automatic Classification of Domestic Journal articles Using Random Forest, *Journal of the Korean Society for Information Management (JKOSIM)*, 36(2), 57-77.

Kim. S. A. and Jung. H. S. (2019). The Determinants of Life Satisfaction in Different Age Groups and Their Policy Implications, *Health and welfare policy forum*, 4(270), 95-104.

LaFreniere, D., Zulkernine, F., Barber, D., Martin, K. (2016). Using machine learning to predict hypertension from a clinical dataset. *In 2016 IEEE symposium series on computational intelligence (SSCI),* IEEE. 1-7.

Lee, B. J. (2021). Prediction Model of Hypertension Using Sociodemographic Characteristics Based on Machine Learning, *KIPS Transactions on Software and Data Engineering*, 10(11), 541-546.

Lee, E. K. (2013). Factors associated with Hypertension Control in Korean Adults : The Fifth Korea National Health and Nutrition Examination Survey (KNHANES V-2), *Journal of The Korean Data Analysis Society*, 15(6), 3203-3217.

Lee, I. J. and Lee, J. H. (2020). Predictive of Osteoporosis by Tree-based Machine Learning Model in Post-menopause Woman, *Journal of Radiological Science and Technology*, 43(6), 495-502.

Lee. H. Y. (2018). Evaluation and Management of Hypertensive Patients According to New Hypertension Guideline, *The Korean Journal of Medicine*, 93(5), 447-451.

Lee. K. E and Cho. E. H. (2016). Factors Influencing Health related Quality of Life in Patients with Hypertension : Based on the 5th Korean National Health and Nutrition Examination Survey, *The Journal of the Korea Contents Association*, 16(5), 399-409.

Lim. H. K. (2018). Prediction of Myocardial Infarction/Angina and Selection of Major Risk Factors Using Machine Learning. *Journal of The Korean Data Analysis Society*, 20(2), 647-656.

Mun. S. E., Jang. S. B., Lee. J. H. and Lee. J. S. (2016). Technology Trends in Machine Learning and Deep Learning, *Information and Communications Magazine*, 33(10), 49-56.

Oh. T. S., Kim. D. K., Won. C. W., Kim. S. Y., Jeong. E. J., Yang. J. S., Yu. J. H., Kim. B. S. and Lee. J. H. (2022). A Machine-Learning-Based Risk Factor Analysis for Hypertension: Korea National Health and Nutrition Examination Survey 2016 - 2019, *Korean Journal of Family Practice*, 12(3), 173-178.

Samadian F., Dalili N., Jamalian A. (2016). Lifestyle modifications to prevent and control hypertension, *Iran J Kidney Dis,* 10(5), 237-263.

Son. J. W. (2021). *A study on the forecasting model for contract group of apartment by using machine learning methods, Master's Thesis*, M. Graduate School of HanYang University, Seoul, Korea.

The Korean Society of Hypertension(KSH) (2018). *2018 Hypertension Treatment Guidelines*.

The Korean Society of Hypertension (2022). *Press release as a result of analysis of the national prevalence rate using big data from the National Health Insurance Corporation.*

Yoon. S. and Bang. H. T. (2021). Evaluation

of a Thermal Conductivity Prediction Model for Compacted Clay Based on a Machine Learning Method, *KSCE Journal of Civil and Environmental Engineering Research*, 41(2), 123-131.

강 성 안 (SeongAn Kang)

- 학생회원
- 동아대학교 경영정보학과 경영학사
- (현재) 동아대학교 경영정보학과 석사과정
- 관심분야: 데이터 사이언스, 경영정보, 텍스트 마이닝, etc.

김 소 희 (SoHui Kim)

- 학생회원
- 동아대학교 경제학과 경제학사
- (현재) 동아대학교 경영정보학과 석사과정
- 관심분야: 빅데이터, 텍스트마이닝, 금융데이터분석, etc.

류 민 호 (Min Ho Ryu)

- 정회원
- 성균관대학교 산업공학과 공학사
- KAIST 기술경영학부 석사
- KAIST 기술경영학부 박사
- (현재) 동아대학교 경영정보학과 부교수
- 관심분야: 빅데이터분석, 텍스트마이닝, 데이터시각화, IT Management, etc.