한국리듬운동학회지, 2023, 제16권 제3호, 11-25
Journal of Korean Society for Rhythmic Exercises
2023, Vol. 16, No. 3, 11-25
https://doi.org/10.34270/jksre.2023.16.3.11
ISSN 2734-066X (Print)
ISSN 2734-0678 (Online)

LASSO와 LightGBM을 활용한 지체장애인 고혈압 유병 분류 알고리즘

박지영[1] · 이정민[2] · 박영빈[3] · 강동헌[4*]
보건복지부 국립재활원 연구원[1,4] · 경찰대학 석사과정[2] · 아워랩 연구원[3]

Prevalence and Classification of Hypertension with LASSO and LightGBM Algorithms for People with External Impairment

Park, Jiyoung[1] · Lee, JeongMin[2] · Park, Youngbin[3] · Kang, Dongheon[4*]
National Rehabilitation Center, Ministry of health and welfare[1,4] ·
Korean National Police University[2] · OurLab[3]

Abstract

We suggest a classification model that predicts the prevalence of hypertension, the No. 1 chronic disease suffered by people with external impairment. We use the data from the “National survey of the Disabled Persons” conducted in 2017 by the Ministry of Health and Welfare and the Korea Institute for Health and Social Affairs. We performed penalized regression analysis(Ridge and LASSO) to select a group of variable candidates and chose α=0.01 for the LASSO model with optimal performance. Based on the selected variables, we constructed a model that can predict the prevalence of hypertension in three classification models (SVM, Logistic Regression, and LightGBM) and evaluated their performance with confusion matrices. We confirmed that lightGBM(learning_rate: 0.05, max_depth : 3) showed the best performance with accuracy of 0.67 and F1 Score of 0.76. This study provide information on hypertension management and prevention by predicting high-risk groups of high blood pressure, a chronic disease of people with external impairment, using the actual condition survey data. It is also expected to contribute to improving the quality of life of people with disabilities.

Key words : People with physical disabilities, hypertension, penalized regression analysis, lasso, lightGBM

* jakekang@korea.kr

Ⅰ. 서론

질병관리청의 “2021 만성질환 현황과 이슈”에 따르면 우리나라 사망원인 10위 내 8개가 만성질환으로 인한 사망이며, 전체 사망의 약 80%를 차지한다(정은경, 2021). 특히 장애인의 경우 비장애인에 비해 장애로 인해 부차적으로 발생하는 건강 문제에 직면하고 있으며, 다소 취약한 건강 상태로 인하여 만성질환이 조기에 발병할 가능성이 높다(호승희, 2017). 또한 2018년 장애인 건강보건통계에 따르면, 장애인의 47.6%가 고혈압을 앓고 있는 것으로 나타났으며, 앞의 선행 연구에 따르면 장애인 2명 중 1명꼴로 고혈압에 시달리고 있는 것으로 파악된다(국립재활원, 2020). 고혈압은 “수축기 혈압이 140mmHg 이상이거나 이완기 혈압이 90mmHg 이상인 경우”를 말하며, “뚜렷한 증상이 없어 대부분 심각한 합병증이 생겨 고혈압이 있음을 알게 되는 경우”가 많은 편이다(질병관리청 국가건강정보포털). 고혈압은 “뇌, 눈, 심장, 신장의 혈관을 손상시키고 뇌졸중, 협심증, 심근경색증, 실명, 심부전, 신부전 등의 위험을 증가”시키기에(인하대병원 인천권역심뇌혈관질환센터, 2020), 만성질환은 지속적인 관리가 필요할 뿐만 아니라 의료비 부담을 가중시키고 심한 경우 장애로 이어질 가능성이 있다. 2차 장애나 복합 장애로 이어지거나 조기 사망의 위험이 있기에 장애인에게서 만성질환 관리는 더욱 중요하므로, 장애인의 만성질환 관리 및 예방과 더불어 생활습관 관리를 포함한 조기 중재(intervention)의 필요성을 시사하였다(김지영, 강민욱, 서욱영, & 이지원, 2020).

2020년 말 기준 장애 유형에 따른 등록장애인 비율은 지체장애인이 45.8%로 가장 많고, 신규 등록장애인은 지체장애인이 16.6%로 2순위를 차지하였다(보건복지부, 2021). 지체장애는 “신경계, 근골격계에서 발생한 다양한 원인으로 인해 몸의 기능이 영구적으로 제한된 질환을 의미”하며, 지체장애인은 운동기능 장애가 있는 경우가 많기 때문에 과체중 및 비만에 노출될 위험이 있다(서울아산병원, 지체장애; 김동일, 김지영, & 송기호, 2020). 비만일 경우 동반되는 합병증으로는 고혈압, 이상지질혈증, 당뇨, 심혈관계 질환, 조기 사망 등이 있다(국립재활원, 2022). 또한 지체장애인에게 자주 발생하는 질환은 등 통증, 고혈압, 무릎관절증, 당뇨병 순으로 나타났다(보건복지부, 2015). 의료분야는 정보가 가지고 있는 부정확성, 개인정보보호 등의 특징 등으로 적용에 제한이 있으나 데이터마이닝 기법을 적용하여 임상에서의 의사결정 예측력을 높이며, 고위험 인자를 발견하여 발생률을 낮출 수 있다(이슬기, & 신택수, 2018). 고혈압과 당뇨병에 대하여 데이터마이닝을 활용하여 사회경제적 요인과 건강행위요인을 선별하여 유병여부를 예측하는 모형이 있었으나 비장애인 대상으로 국한되었다(김한결, 최근호, 임성원, & 이현실, 2016). Garcia(2019)와 Kan(2019) 등의 연구에서 노인 미래 의료비, 고협압 환자의 심형관 질환 예측 등을 위해 LASSO를 활용한 변수 선정, 의사결정알고리즘인 LightGBM을 활용하였다. 2017년 장애인 실태조사의 지체장애인의 만성질환에 대한 조사 결과 고혈압이 20.14%로 가장 높은 비율을 차지하였다. 따라서 본 연구는 2017 장애인 실태조사 데이터를 활용하여 지체장애인의 고혈압에 영향을 미치는 요인을 파악하고, 고혈압 유병 분류 모델 알고리즘을 제안하고자 한다.

Ⅱ. 이론적 배경

1. 벌점화 회귀분석

벌점화 회귀분석에는 Ridge(능형 회귀), Least Absolute Shrinkage and Selection Operator(LASSO) 방법론이 존재한다. Ridge(능형 회귀)는 종속변수에 영향을 미치지 못하는 변수에 벌점을 부과해 회귀계수를 0에 가

깝게 줄이지만, 변수 선택의 기능은 없어 모든 변수가 모형에 포함된다(송상윤, 2015). 반면에 LASSO는 Ridge와 비슷하지만 종속변수에 영향을 미치지 못하는 변수의 회귀계수를 0으로 만들어 모델에서 제외시키기 때문에 변수 선택의 기능을 한다. 랜덤 포레스트 모형과 LASSO 회귀 모형을 사용하여 비만 환자의 CPAP 적정압력을 예측하는 모형을 만들고, 최소제곱법을 이용한 선형 회귀 모형들과 비교하여 더 우수한 성능을 확인한 연구도 있었다(김승수, & 양광익, 2018). 노인의 미래 의료비 예측에 있어 Oridinary Least Squares (OLS)와 벌점 선형 회귀 모형의 예측 성과를 심층적으로 비교하여, LASSO 회귀 모형이 우수한 예측 비율을 보여준다고 입증한 연구도 있다(Kan, Kharrazi, Chang, Bodycombe, Lemke & Weiner, 2019). 또 다른 연구의 경우, 고혈압 환자의 심혈관 질환을 예측하기 위해 콕스 비례위험 회귀 모형(Cox proportional hazards regression)과 벌점 회귀 모형(LASSO, Elastic Net)을 비교하였다(Garcia, Barquero, Mora, Soguero, Goya & Ramos, 2019). 성능은 세 가지 모델 모두에서 유사하지만, 두 벌점 회귀 모형 모두 콕스 비례위험 회귀 모형보다 적합성이 좋고 특성(feature)이 적어 더 우수하다고 판단된다.

본 연구에서는 데이터 부족으로 인한 과소적합을 방지하고 주어진 데이터를 최대한 활용하기 위해, 학습 데이터와 검증 데이터로 나누어 검증 데이터에 대한 최적의 모델을 선정하는 교차 검증(Cross Validation) 기법을 활용하였다(조남훈, 2006).

2. 분류 모델 알고리즘

본 연구에서는 고혈압 유병 예측을 위한 분류 모델을 구현하기 위해 SVM(Support Vector Machine), Logistic Regression, LightGBM 알고리즘을 활용하였다.

1) Support Vector Machine (SVM)

SVM은 퍼셉트론 기반 판별함수 모형으로 데이터를 선형으로 분리하여 이진 혹은 다중 분류하는 모델이다(김도형, 2020). 비선형일 경우 차원을 늘려 선형으로 분리하며, 2차원일 경우 선으로 3차원일 경우 면으로 분리한다. 이러한 선과 면을 판별 경계선 또는 초평면이라고 한다(박주완, 배진성, & 윤혁준, 2019). SVM은 서로 다른 클래스를 최대의 마진(margin)으로 분리하는 초평면을 찾아내는 과정이다(정우진, 2020). 보통 대출 승인 여부나 질병 발병 여부 등을 분류하는 데 사용된다. SVM은 조정해야 할 파라미터의 수가 적어, 다른 모델에 비해 간단하게 영향을 미치는 요인을 파악할 수 있다는 장점을 지닌다(안현철, 김경재, & 한인구, 2005).

SVM의 하이퍼 파라미터는 ‘kernel’, ‘C’, ‘gamma’이다. ‘kernel’에는 linear, poly, rbf 등이 존재한다. ‘C’는 오차 허용 정도를 결정하고, ‘gamma’는 결정 경계의 곡률을 조정하며 두 하이퍼 파라미터 모두 값이 클수록 과적합의 위험이 있다(Scikit Learn developers, 2022a).

2) 로지스틱 회귀분석(Logistic Regression)

로지스틱 회귀분석은 2개 이상의 집단에 대하여 개별 관측치들이 어느 집단에 분류될 수 있을지 예측하는 통계적 기법이다(엄영호, 황정윤, & 정헌주, 2014). 로지스틱 회귀분석은 일반적인 회귀모형과 달리 종속변수를 이항변수로 하여 발병할 것인가, 아닌 것인가 하는 분류 문제를 해결하는 데 사용된다. 로지스틱 회귀분석의 하이퍼 파라미터는 ‘penalty’와 ‘C’이다. ‘penalty’는 규제 방법을 선택하는 파라미터이며, ‘C’는 ‘penalty’에 대한 계수를 결정하고 값이 클수록 복잡한 모델에 대한 규제를 강화한다(Scikit Learn developers, 2022b).

3) LightGBM

LightGBM 알고리즘은 의사 결정 트리 알고리즘에 기반하여 순위 또는 분류를 위한 기계학습 작업에 사용된다(이현미, 장정아, & 전교석, 2020). 기존에 Gradiant Boosting 알고리즘에서 주로 사용되는 트리의 깊이(depth wise)나 균형 트리(level wise)가 아닌, 가장 잘 맞는 트리의 리프 중심(leaf wise)으로 분할한다는 차별점을 지닌다. 최대 손실값을 갖는 리프 노드를 지속적으로 분할하기 때문에, 균형 트리 분할 방식에 비해 예측 오류 손실을 최소화하며 보다 빠른 수행이 가능하게 한다. LightGBM의 하이퍼 파라미터는 'learning_rate'와 'max_depth'이다. 'learning rate'는 부스팅 스텝을 반복적으로 수행할 때 업데이트되는 학습률이며, 'max_depth'는 tree의 최대 깊이를 조절하며 모델 과적합을 다룰 때 사용한다(최태정, 박지인, 손상원, & 윤주범, 2022).

또한 모델별로 최적의 성능을 낼 수 있는 하이퍼 파라미터 선정을 위해 GridSearchCV 기법을 활용하였다. GridSearchCV는 주어진 범위 내에서 일정 간격으로 설정된 하이퍼 파라미터 값을 조절하는 iteration을 거쳐 모델의 성능을 평가하고, 최적의 성능을 보이는 모델을 도출한다(변중무, 윤대웅, 최용욱, & 최준환, 2020).

3. 분류 모델 성능 평가

혼동행렬(confusion matrix)이란 실제값과 예측값이 일치하는지, 일치하지 않는지 보여주는 표로, TP(True Positive, 실제값과 예측값이 양성인 경우), TN(True Negative, 실제값과 예측값이 음성인 경우), FP(False Positive, 실제값이 음성이고 예측값이 양성인 경우), FN(False Negative, 실제값이 양성이고 예측값이 음성인 경우)로 구성된다(Chris Albon, 2019).

표 1. 혼동행렬

		실제	
		Positive	Negative
예측	Positive	True Positive(TP)	True Negative(FN)
	Negative	False Positive(FP)	False Negative(TN)

기계학습 분류 기반으로 한 예측 모형의 평가 및 성능 비교에는 정확도, 정밀도, 재현율, F1 score 네 가지 지표가 있으며, 그중 정확도(Accuracy)는 실제값과 예측값의 일치율을 나타내는 평가 지표로 식(1)과 같다(신백균, 2022). 정밀도(Precision)는 양성이라고 예측한 값 중 실제 양성인 비율을 의미하며, 식(2)와 같다(신백균, 2022). 재현율(Recall)은 실제 양성인 값 중 양성으로 잘 예측한 값의 비율을 의미하며, 식(3)과 같다(신백균, 2022). 식(4)의 F1 score는 정밀도와 재현율을 조합한 조화평균으로, 서로 상충하는 두 지표가 모두 높아야 높아지는 특성을 지녀 모델의 정확성을 종합하여 나타낼 수 있는 지표이다(Heo, Kwon, Kim, Han, & An, 2018). 본 연구에서는 혼동행렬로부터 계산된 4가지 지표로 성능을 비교해 상대적으로 높은 성능을 보이는 분류 모델을 선정하였다.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad \text{식(1)}$$

$$\mathrm{Precision} = \frac{TP}{TP + FP} \quad \text{식(2)}$$

$$Recall = \frac{TP}{TP + FN} \quad \text{식(3)}$$

$$F1\,score = \frac{2}{\frac{1}{\mathrm{Precision}} + \frac{1}{Recall}} = 2 \times \frac{\mathrm{Precision} \times Recall}{\mathrm{Precision} + Recall} \quad \text{식(4)}$$

전통적인 회귀 모형에서는 많은 변수를 한 모델에 투입할 시, 다중공선성과 과적합 문제가 발생하여 해석에 어려움이 있다(김유경, 이정원, 김동호, 2022,). 이러한 방법론적 한계를 극복하고 지체장애인의 고혈압 예측 요인들을 통합적으로 분석하고자 벌점화 회귀분석을 실시하였다. 벌점화 회귀분석을 통해 선정된 변수들을 토대로, 고혈압 유병을 분류하는 알고리즘을 구현하였다.

Ⅲ. 연구 방법

1. 연구 대상

본 연구에서는 고혈압 여부에 응답한 지체장애인을 연구 대상으로 선정하였다. 보건의료 및 운동과 보조기기와 관련한 설문 문항을 독립변수로 한정하여, 해당 요인들이 지체장애인의 고혈압 유병에 미치는 영향을 파악하기 위한 분류 모델을 개발하고자 하였다. 연구에 활용한 데이터는 장애인 실태조사 원시자료이다. 장애인 실태조사는 1980년부터 3-4년 주기로 조사되고 있으며, “우리나라 장애인구 및 장애 출현율 등의 인구통계학적 특성과 장애인의 생활실태 및 복지욕구를 파악하여 장단기 장애인복지정책 수립 및 시행을 위한 기초자료를 생산하는 데” 목적을 둔다(보건복지부, 2018). 본 연구에서는 2017년 조사 자료를 활용하여 분석하였다. 2017년 조사 모집단은 “전국 17개 시도의 일반 주거시설에 거주하는 일반 가구 및 가구원”으로 약 45,000가구가 참여하여, 36,200가구가 조사 완료한 조사이다(보건복지부, 2018). 조사 완료된 가구의 가구원 수는 91,405명으로, 그중 장애인은 6,549명이었다(보건복지부, 2018). 장애인 실태조사의 공통 문항으로는 “보건 의료・건강, 일상생활 지원, 장애인 보조 기기, 교육, 취업 및 직업생활” 등이 있고, 총 15개 장애유형에 따라 설문 문항이 세분화되어 있다.

장애특성 중에 지체장애 대상의 설문 문항은 다음과 같다. 불편한 부위가 상지/하지/척추, 총 3가지로 제시되었고, 그 부위가 좌, 우, 양측인지를 선택하는 문항이 있었다. 가장 불편한 부위는 절단/마비/관절장애/변형 크게 4가지 부위가 제시되었으며, 4가지 부위별로 세부 부위로도 나뉘어 있어서 한 개만 선택하도록 이루어져 있었다. 또한 장애의 주된 원인(선천적 원인, 출생시 원인, 후천적 원인, 원인불명)과 주된 진단명을 선택하는 문항이 존재하였다.

보건의료 및 건강 관련 설문 문항은 다음과 같다. 건강보험 가입여부 및 형태, 현재 치료, 재활, 건강관리 등 지속적인 진료를 받고 있는지 여부, 주관적인 몸 상태 판단, 3개월 이상 계속되는 만성질환 여부, 음주 및 흡연 여부 등이 존재하였다.

일상생활 지원 관련 문항이 있었으며, 다음과 같은 행동을 스스로 할 수 있는가에 대한 여부를 묻는 문항이다. 옷 갈아입기, 목욕하기, 음식물 넘기기, 약 챙겨먹기, 빨래하기 등이며, 남의 도움을 얼마나 필요로 하는지, 어느 정도로 도움이 되는지에 대한 문항도 존재하였다.

장애인 보조기기에 대한 필요, 소지, 사용 여부 등에 대한 문항과 사용 빈도 및 시간, 만족도, 구매경로 등을 묻는 문항이 존재하였다. 보조기기 종류는 장애유형별로 나뉘어 있었며, 지체장애 대상 보조기기 종류는 상하지 의지, 상하지 보조기, 맞춤형 교정용 신발, 자세보조용구 등이 있었으며, 공통 기기는 지팡이, 목발, 전동/수동 휠체어, 전동 스쿠터가 존재하였다.

해당 자료는 보건복지데이터 포털에 원시자료를 요청하여 sav 형식으로 제공받아 활용하였다. 아래 <그림 1> 은 2017 장애인 실태조사 원시자료를 이용하여 지체장애인의 3개월 이상 지속되는 만성질환 시각화한 것으로 고혈압이 가장 많은 응답 비율을 차지하고 있음을 알 수 있다.

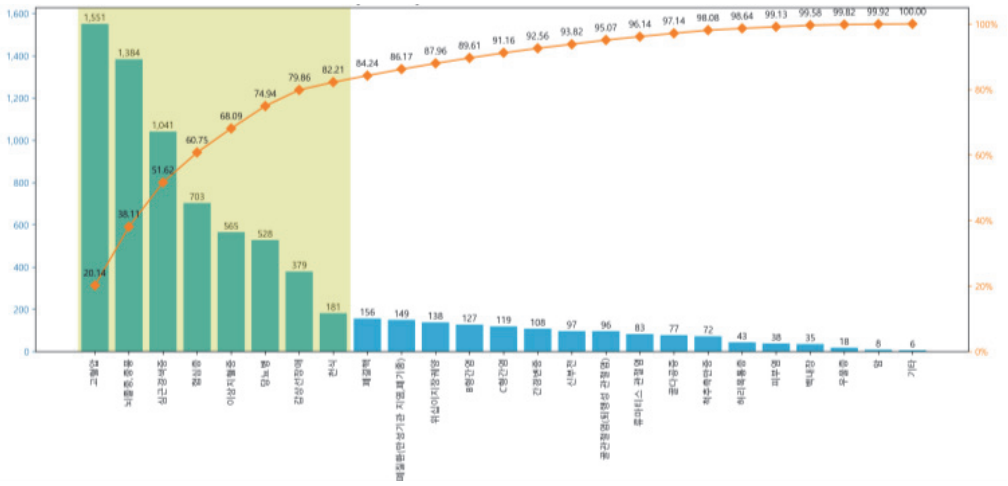


그림 1. 지체장애인의 만성질환 분포

2. 연구방법

1) 활용 도구 소개 및 연구 수행 절차

본 연구에서는 분석을 위하여 R, Python을 활용하였다. 고혈압에 영향을 미치는 변수 선정을 위한 벌점화 회귀분석을 시행하기 전, R의 dplyr, readxl 패키지 등을 활용하여 데이터 정제 작업을 실시하였다. Python에서는 pandas, sklearn 패키지 등을 활용하여 R에서 정제한 데이터를 벌점화 회귀분석하고 다양한 예측 모형을 구현하였다. 연구 수행 절차는 아래 <그림 2>와 같다.

그림 2. 연구 순서도

2) 데이터 정제

지체장애인과 등록장애인 대상자만 1차 추출(3,253명)하고, 추출된 장애인 중 고혈압 여부에 응답자를 최종 추출(2,686명)하였다. 또한 본 연구는 지체장애인을 대상으로 하므로 지체장애인 외의 다른 장애유형에 대한 설문 문항 변수는 모두 제외하였다. 재활치료 서비스 관련 문항에서는 '이용 여부' 변수만 남기고 고혈압 발병과 상대적으로 연관성이 떨어지는 변수(이용 시간, 비용, 바우처 여부 등)는 제거하였다. 고혈압 유병과 직접적인 연관이 있는 월 혈압 약 일수와 혈압약 복용 여부 변수는 제거하여 최종 데이터 셋을 구축하였다.

3) 연구방법

본 연구에서는 벌점화 회귀분석(penalized regression)을 통해 모델링에 사용할 변수를 선정하였다. 예측성과 지표로는 RMSE(평균 제곱근 오차)를 활용하였으며, 각 α값에 따른 회귀 모형의 변수 후보군들을 추출하여 분류 알고리즘 모델에 적용하였다.

모델 생성과 이를 위한 전처리 과정은 Python에서 수행하였다. 먼저 모델 정확도를 높이기 위해 주어진 데이터를 예측 모델의 문제를 잘 표현할 수 있는 Feature로 변형시키는 Feature Engineering을 수행하였다. 이 과정에서는 One-Hot encoding을 통해 범주형 변수를 수치형 자료로 바꾸어 정보 손실 없이 범주형 자료를 그룹화 하였으며, 연속형 변수는 데이터의 범

위가 다르기에 Standardization(Z-score normalization, 표준화) 스케일링을 사용하였다. 그리고 모델 생성에 앞서 학습 데이터(70%)와 테스트 데이터(30%)를 분할하였다. 그리고 변수 전처리와 모델을 생성하는 일련의 과정을 순차적으로 처리하기 위해 파이프라인을 생성하였다.

본 연구에서는 SVM, Logistic Regression, LightGBM 알고리즘을 활용하여 하이퍼 파라미터를 다양하게 조절해 보며 최적의 성능을 내는 대표 모델을 알고리즘당 하나씩 제시하였다. 벌점화 회귀분석을 통해 추출된 변수 후보군들을 이용하여 분류 모델을 생성한 후, GridSearchCV를 이용해 모형별로 최적의 하이퍼 파라미터를 도출하고, 교차 검증을 통해 각 알고리즘별 최적의 모형을 구축하였다(이현미, 전교석, & 장정아, 2020). 최종적으로 시행할 분석 파이프라인은 <그림 3>~<그림 4>와 같다.

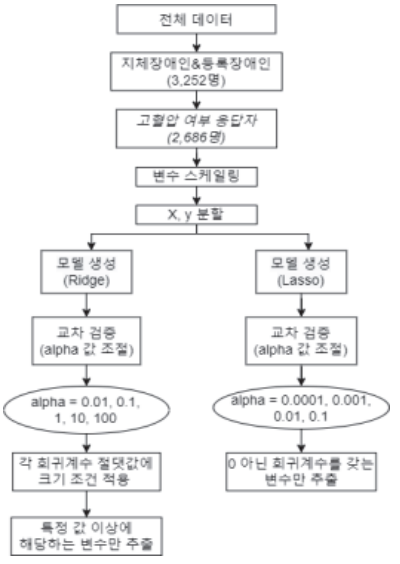


그림 3. 데이터 정제 및 벌점화 회귀분석 과정

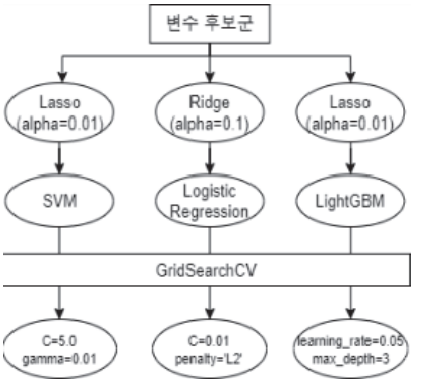


그림 4. 추출된 변수 후보군들을 이용한 분류 모델 생성 및 평가

Ⅳ. 연구 결과

1. 인구사회학적 특성

연구 대상의 인구사회학적 특성은 아래 <표 2>와 같다.

표 2. 연구 대상자 인구사회학적 특성

변수	범주	N(%) or mean±SD
성별	남성	1337 (49.8)
	여성	1349 (50.2)
나이		67.22±11.956
장애 등급	1등급	59 (2.2)
	2등급	129 (4.8)
	3등급	318 (11.8)
	4등급	535 (19.9)
	5등급	898 (33.4)
	6등급	747 (27.8)
지난 2주간 아팠던 날수 (단위 : 일)		12.893±3.539

2. 벌점화 회귀분석 결과 및 도출된 변수

Ridge의 α는 0.01, 0.1, 1, 10, 100으로, LASSO는 0.0001, 0.001, 0.01, 0.1로 값을 조절하며 적합한 α값을 선택하였다. Ridge는 변수 선택의 기능이 없기 때문에 α가 0.1이고 회귀계수의 절댓값이 0.3 이상인 경우, LASSO는 α = 0.01인 경우를 고려하였다. Ridge의

경우 조건을 만족하는 특성 수가 28개, LASSO의 경우 22개일 때 실질적으로 설명 가능하고 분류 모델 개발에 활용할 적절한 개수라 판단하였기 때문이다. 5-fold CV(Cross Validation)를 실시해 모형의 예측성과를 확인한 결과는 다음 <표 3>와 같다.

표 3. Ridge와 LASSO의 5-fold CV 결과

벌점화 회귀분석	α(alpha)	5-fold의 평균 RMSE	사용한 특성 수
Ridge*	0.01	0.529	62
	0.1	0.521	28
	1	0.502	12
	10	0.479	3
	100	0.463	0
LASSO	0.0001	0.497	517
	0.001	0.457	190
	0.01	0.461	22
	0.1	0.496	0

* Ridge는 변수 선택의 기능이 없기 때문에 α가 0.1이고 회귀계수의 절댓값이 0.3 이상인 경우의 특성 수를 뜻함.

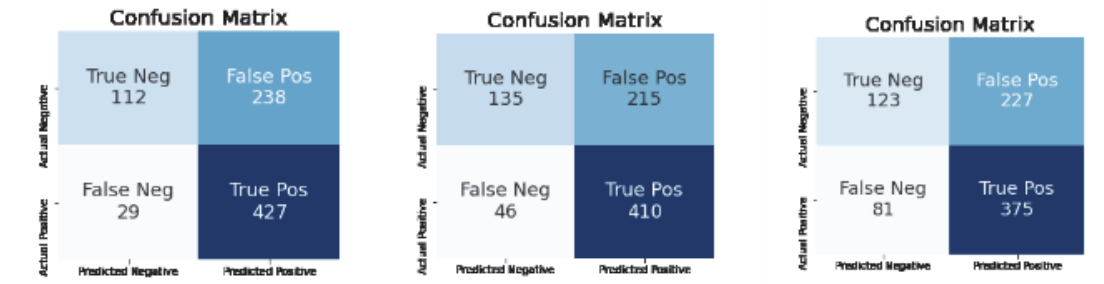
Ridge 모델은 α가 0.1일 때 회귀계수의 절댓값이 0.3 이상인 조건을 만족하는 변수는 28개로, '생년'(-), '월 평균 총가구소득'(+), '키(센티)'(-), '몸무게(kg)'(+), '지난 2주간 아팠던 날수'(+), '가구 유형', '가장 불편한 부위-[절단]어깨 이상(-), [마비]상지 양쪽(-), [마비]전신(전신마비)(+)', '장애주된 원인'(-), '질병명-부정맥(+), 기타 심혈관질환(-), 골다공증(-), 알콜 및 약물의존(+)', '주된 진단명'(-), '건강보험가입여부 및 형태'(-), '사고발생내용2-운수사고(교통사고)(-), 열상/자상/절단/관통상(베임)(+)', '희귀난치성질환 등록 여부'(+), '암 종류1'(+), '암 종류2-대장암(-), 전립선암(+)', '언어치료-이용여부'(-), '놀이치료-이용여부'(+), '심리행동치료-이용여부'(-), 'ADL-음식물 넘기기'(-), 'ADL-옮겨앉기'(+), 'ADL-배변'(-), 'ADL-빨래하기'(-), 'ADL-약 챙겨먹기-지원 불필요(+), 전적인 지원 필요(-)', '현재 일상생활 도와주는 이 유무'(-), '주로 도와주는 사람-손자녀(+), 친척(-)', '팔받침대-사용'(-), '의사소통 보조기기-필요'(+), '이동식 전동리프트-필요'(-)이다. Ridge 회귀분석의 결과는 변수 후보군 추출을 위한 최종 모델로 선정되지 않았기에 회귀계수 부호만 언급하였다. LASSO 회귀분석을 통해 최종 모델링에 활용한 변수는 <표 4>와 같다.

표 4. α = 0.01일 때 LASSO에서 도출된 변수와 회귀계수

구분	변수	회귀계수
거동 불편	가장 불편한 부위(무릎)	0.006
	주된 진단명(관절질환)	0.003
	지팡이 필요(Y)	0.030
	지팡이 소지(Y)	0.007
	장애등급(4급)	0.015
질병	지난 2주간 아팠던 날수	0.388
	혈당관리 치료 여부(비해당)	-0.032
정보통신 기기 사용 여부	휴대폰 사용 여부(Y)	0.009
	스마트폰 사용 여부(Y)	-0.037
	컴퓨터 사용 여부(Y)	-0.016
건강관리 - 합병증	현재 지속적 진료여부(Y)	0.139
	만성질환 관리여부(Y)	0.029
	장애관리 및 재활서비스 여부(N)	0.019
	인플루엔자 예방접종여부(Y)	0.017
체형	본인 체형 평가(비만)	0.062
생활 양식	가구 유형(부부+미혼자녀)	-0.047

3. 분류 모델 결과 및 모델 성능 평가

예측 데이터와 실제 데이터를 행렬로 나타낸 혼동행렬을 이용해 분류 모델 성능을 파악하였다. 본 연구에서는 혼동행렬로부터 계산된 4가지 지표로 성능을 비교해 상대적으로 높은 성능을 보이는 분류 모델을 선정하였다. 각 모델별 혼동행렬은 다음 <그림 5>와 같다. 각 최적모형의 성능 평가 지표 결과는 <표 5>



Ver.1(SVM) 모형 Ver.2(Logistic Regression) 모형 Ver.3(LightGBM) 모형

그림 5. 혼동행렬을 활용한 성능 평가 결과

와 같다. 모형 평가 결과로 모형의 정확도 지표는 Ver.3(LightGBM, 67.99%), Ver.2(Logistic Regression, 67.72%), Ver.1(SVM, 66.87%) 순으로 높았다. 재현율 지표는 Ver.1(SVM, 93.64%), Ver.3(LightGBM, 93.2%), Ver.2(Logistic, 89.91%) 순으로, Ver.2(Logistic Regression)은 재현율이 상당히 떨어져 최종 모델 선택 과정에서 제외되었다.

Python API의 feature importance(F1-Score 지표)를 통해 모델에서의 각 변수의 중요도를 알 수 있으며 Ver.3 (Light GBM, 76.71%), Ver.1(SVM, 76.18%), Ver.2(Logistic Regression, 75.86%), 순으로 높게 나타났다. F1-Score는 서로 상충하는 정밀도와 재현율 두 지표가 모두 높아야 높아지는 특성을 지녀 모델의 정확성을 종합하여 나타낼 수 있는 지표이므로, F1-Score가 높은 Ver.3 (LightGBM)을 최종 모형으로 채택하였다. 최종 선택된 모델에서 각 변수의 중요도 순위는 다음 <그림 6>과 같다.

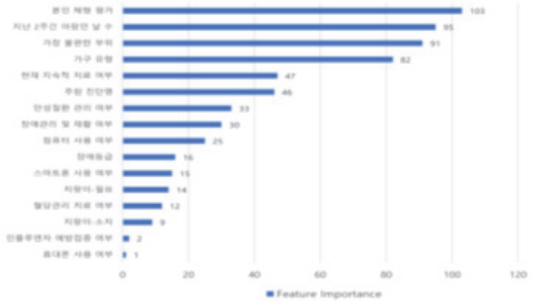


그림 6. Ver.3분류 모델(LightGBM)의 변수 중요도 (Feature Importance)

Ⅴ. 논의

본 연구는 LASSO 회귀분석을 통해 모델링에 활용될 변수를 추출하여 예측모형에 주로 활용되고 있는 3가지 모형(SVM, Logistic Regression, LightGBM)에 적용하여 고혈압 유병을 분류한 모델이다. 3가지 모형 중 최종 선정된 LightGBM 모델은 거동 불편과 관련된 변수로는 가장 불편한 부위(무릎), 주된 진단명(관절질환), 지팡이 필요(Y), 지팡이 소지(Y), 장애등급(4급)이 도출되었다. 가장 불편한 부위를 묻는 문항에 관절장애 중에서도 하지 관절에 해당하는 무릎이 가장 불편한 부위라고 답할수록 고혈압일 확률이 높은 변수로 채택되었다. 우리나라에서 가장 높은 이환율을 보이는 질환이 고혈압과 골관절염인 것을 염두하면 (김은숙, 2021), 타당한 결과라고 판단된다. 동 연구는 골관절염을 앓는 고혈압 노인을 대상으로 건강 상태를 긍정적으로 수용하고 활동 제한을 최소화할 중재 프로그램을 제안하였고, 체질량지수 관리를 위한 다학제적 프로그램의 개발과 중재의 필요성을 언급하였다 (김은숙, 2021). 주된 진단명을 '관절 질환'이라고 응

답할수록 고혈압을 앓고 있을 확률이 높았으며 이는 가장 불편한 부위와 동일 사유로 판단된다. 지팡이가 필요하다고 응답할수록, 지팡이를 소지하고 있다고 응답할수록, 고혈압을 앓고 있을 확률이 높았다. 지팡이는 주로 걸을 때 도움을 얻기 위해 짚는 보조 기기이며, 신체 능력만으로는 제대로 걷기 어려운 노인이나 장애인 등이 주로 이용한다. 이러한 결과는 앞서 언급했던 하지 관절 장애가 고혈압에 걸릴 확률을 높였던 것과 결부시켜 해석할 수 있다. 장애등급도 마찬가지로 살펴보면, 4급의 경우 최대 두 다리 각각의 3대 관절 중 2개의 운동 범위가 각각 50% 이상 75% 미만 감소된 사람, 최대 두 다리의 모든 3대 관절의 운동 범위가 각각 25% 이상 50% 미만 감소된 사람이 포함된다(보건복지부고시 제2018-151호, 2018).

거동 불편과 관련된 변수들을 종합적으로 보면 하지 관절과 관련한 장애가 고혈압과 가장 관련성이 큰 것으로 판단된다. 거동 불편이 비만 및 합병증을 유발하여, 고혈압에 이르게 되는 것으로 추측되므로, 향후 연구에서는 하지 관절 장애와 고혈압의 관계를 파악하는 연구가 진행될 필요가 있다. 또한 지난 2019년 7월, 6등급으로 나뉘었던 장애등급은 이를 경증과 중증으로만 구분하도록 하였다(보건복지부, 2019). 연구에 사용한 데이터는 2017년 자료이지만 공개된 시기는 2021년이므로 추후 없어진 장애등급을 대체하여 조사된 경증, 중증의 데이터를 활용하여 다시 연구가 필요할 것으로 판단된다.

질병과 관련된 변수로는 지난 2주간 아팠던 날수, 혈당관리 치료 여부(비해당)가 있다. 지난 2주간 질병이나 사고 등으로 아팠던 적이 있었는지 묻는 문항은 고혈압, 당뇨 등 만성질환으로 약을 복용하는 경우에도 아팠다고 응답하도록 조사하였기 때문에, 고혈압을 비롯한 만성질환자는 모두 ‘예’라고 답하였을 것이다. 그 문항에 ‘예’라고 답하는 경우 세부 문항으로 지난 2주간 아팠던 날수를 최대 14일까지 주관식으로 기재하는 문항이 있다. 이 문항에서 지난 2주간 아팠던 날수가 증가할수록 고혈압에 걸릴 확률이 높았다. 그러나 지난 2주간 아팠던 날수가 고혈압으로 인한 것일지라도 고혈압 유병 확률이 높아질 수 있어 인과관계를 판단하기 어렵다. 혈당관리 치료 여부에 비해당(당뇨병 앓지 않음)에 응답한 사람일수록, 고혈압일 확률이 낮았다. 해당 문항은 당뇨병을 앓고 있는지 여부에서 파생된 문항으로 당뇨병을 앓고 있지 않은 응답자는 혈당관리 ‘비해당’에 응답을 하였다. 김수영, 나은희, 조선(2018)에 의하면, 당뇨병에 동반되는 질환 중 공동 1위가 고혈압과 비만(50.3%)으로 나타났다. 당뇨병과 고혈압은 흔히 동반되는 질환으로, 적절한 혈압 조절은 당뇨병 관리의 가장 중요한 부분 중의 하나이다. 또한 당뇨병 환자에게 발병하는 미세혈관(망막병증, 알부민뇨증)과 대혈관(심근경색, 뇌경색) 합병증은 고혈압이 없는 환자보다 고혈압을 동반한 환자에서 통계적으로 유의미하게 높았다(김상용, 2020). 이러한 사실들을 종합하여 판단해보면, 당뇨병 환자가 심혈관계 질환으로 사망할 확률을 낮추기 위해서는 고혈압에 대한 적극적인 대처가 필요하다.

정보통신기기 사용 여부와 관련된 변수로 휴대폰, 스마트폰, 컴퓨터 사용 여부(Y)가 도출되었다. 회귀계수의 절댓값이 가장 낮은 변수이지만, 정보통신기기의 사용 여부에 관한 변수들도 고혈압 유병률을 예측하는 데 영향을 미치는 변수로 나타났다. 휴대폰을 사용하는 사람일수록 고혈압일 가능성이 높지만 스마트폰과 컴퓨터를 사용하는 사람일수록 고혈압이 아닐 가능성이 높다. 일반적으로 연령이 증가할수록 고혈압 유병률이 높아지는데, 본 연구에서 사용한 자료를 이용해 연속형 자료인 연령과 이분형 자료인 정보통신기기 사용 여부 간의 상관관계를 파악할 수 있는 Point biserial correlation 계수를 살펴본 결과는 다음과 같다. 휴대폰 사용 여부는 연령과 양의 상관관계(r=0.44)를 보였지만, 스마트폰, 컴퓨터 사용 여부와 연령은 각각 음의 상관관계(r=-0.58, r=-0.53)를 보였다. 연령이 높을수록 스마트폰과 컴퓨터보다는 휴대폰을

사용할 가능성이 높으므로, 연령에 의한 매개효과일 것으로 추측된다. 정보통신기기의 사용 여부에 따라 고혈압 가능성에 차이가 있다는 결과는 흥미로웠지만, 추후 구조방정식을 통해 정보통신기기 사용 여부의 직접 효과가 있는지 검증한다면 정확한 인과관계를 파악할 수 있을 것이다.

다음 건강관리와 관련된 변수로는 현재 지속적 진료여부(Y), 만성질환 관리여부(Y), 장애관리 및 재활서비스 여부(N), 인플루엔자 예방접종여부(Y)가 도출되었다. 현재 지속적 진료여부에 대해 지속적 진료를 받고 있을수록 고혈압을 앓고 있을 확률이 높았다. 또한 최근 1년 동안 만성질환 관리 서비스와 장애관리 및 재활 서비스를 받고 있다고 응답한 사람일수록, 고혈압일 확률이 높았다. 만성질환 중 고혈압을 앓고 있는 응답자들이 보건 의료 서비스를 받기 위해 해당 서비스를 이용하기에 만성질환 관리 서비스를 받을수록 고혈압의 확률이 높은 것으로 사료된다. 따라서 이 두 변수는 고혈압과의 인과관계가 명료하게 파악되지 않았다. 인플루엔자 예방접종 여부에 관해 인플루엔자 예방접종을 한 사람은 그렇지 않은 사람보다 고혈압의 위험이 높았다. 서초구 보건소에 따르면 인플루엔자 예방접종 대상자는 “만성 호흡기 질환(천식, 기관지확장증, 만성 기관지염, 폐기종 등), 심혈관 질환(류마티스성・허혈성 심질환, 고혈압 등)을 앓는 사람”이라고 명시되었다. 따라서 만성 호흡기 질환이나 심혈관 질환을 기저질환으로 앓고 있다면 그만큼 예방접종을 권고하기 때문에, 이와 같은 결과가 나왔을 것이라 사료된다. 여기에 고혈압이 포함되기 때문에 이 또한 인과관계가 분명하지 않을 수 있다.

본인 체형 평가와 가구 유형도 유의미한 변수로 드러났다. 본인 체형 평가 문항에서 본인의 체형이 비만이라고 평가할수록 고혈압일 확률이 높은 변수로, 고혈압을 예측하는 가장 중요한 변수로 채택되었다. 비만은 대표적으로 고혈압 등의 합병증을 유발하는 원인으로 알려져 있다. 최근 측정한 키와 몸무게에 대한 변수가 있었음에도 중요한 변수로 채택되지 못한 데에는 체형이 키와 몸무게에 영향을 받는데 모델이 두 변수를 함께 반영하지 못하였기 때문이라 추측한다. 본 연구는 장애인 실태조사의 원자료를 그대로 사용하였지만, 키와 몸무게를 함께 고려하는 BMI 지수가 있었다면 중요한 변수로 채택될 수 있었을 것이라 판단된다. 김수정, 박세환, 서영성, 배철영, 신동학(1994)은 비만도를 고려한 상대 체중과 고혈압의 상관성을 밝혔다.

장애인 실태조사는 다양한 가구 유형 범주를 구성하여 정확한 가구 유형을 파악하고자 하였다. 가구유형이 ‘부부+미혼 자녀’인 사람일수록 고혈압이 아닐 확률이 높았다. 선행연구를 토대로 고찰해본다면, 두 가지 가능성을 제시할 수 있다. 첫 번째로, 응답자의 연령과 결부지어 해석할 수 있다. 일반적으로 미혼자녀의 연령대는 기혼자녀보다 낮기 때문에, 미혼자녀를 둔 부부는 그렇지 않은 부부보다 상대적으로 젊을 가능성이 높다. 고혈압은 연령과 상관이 크기 때문에, 가구유형 중 상대적으로 젊은 연령대에 속하는 부부+미혼 자녀 유형은 고혈압일 가능성이 낮을 것이라 판단된다. 그러나 응답자가 가구 유형에서 어느 구성원인지를 파악할 수 없고, 응답자가 더 젊을 가능성이 있는 가구 유형(ex. 부부, 1인 가구 등)이 존재하기 때문에 해석의 한계를 가진다. 후속 연구를 통해 가구유형과 고혈압의 인과관계를 파악하고자 한다면, 연령과의 매개효과를 고려해 볼 수 있다.

두 번째로, 가구 유형별 삶의 질을 원인으로 해석할 수 있다. Gee(2000)은 가구 유형은 노인의 일상 생활과 삶의 질에 영향을 미치는 중요한 환경적 특성이 될 수 있다고 보았다. 가구 유형별로 건강 상태, 의료경험, 건강에 대한 관심도와 생활습관이 다르다는 선행연구(이유현, 김윤진, & 조덕영, 2014; 이은숙, 2021; 이혜재, & 허순임, 2017; 윤미순, 최은희, 김유진, & 최시은, 2021)와 함께 결부 지을 수 있다. 윤미순 등(2021)은 가구 유형이 노인의 건강과 삶의 질에 미치

는 영향요인을 파악하고자 하였는데, 1인 가구(노인인 경우)는 삶의 질이 노인 부부보다 떨어지고 교육을 포함한 전반적인 생활수준이 낮았으며(안경숙, 2005), 조손 가구(조모의 경우)는 경제, 우울, 건강과 관련된 삶의 질이 낮았고, 질병이 있는 경우 삶의 질은 악화되었다(양경순 & 한재희, 2013). 이를 통해 가구 유형별로 건강 문제에 대한 정책적 지원과 사회적 지지 체계를 달리할 필요성을 제기하였다.

마지막으로 본 연구의 한계점은 다음과 같다. 첫째, 충분한 데이터를 사용한 것이 아니므로 분석 결과를 일반화하기 부족하다. 장애인 실태조사는 6,549명의 응답자 정보가 있지만 실제 분석에 사용한 응답자는 2,686명이다. 이는 2021년 기준 전국 지체장애인 1,191,462명 중 0.23%에 해당하는 비율로 충분한 표본이라고 보기 어렵다. 더 많은 표본을 확보하여 분석할 경우 대표성 문제를 개선할 수 있을 것이다. 또한 장애인 실태조사는 조사 대상자의 기억에 의존하여 진행되는 설문조사인 만큼 오류나 누락이 있어서 정확도가 떨어질 가능성이 있다.

둘째, 모델에 활용한 독립변수를 의료보건 분야로만 한정지었다. 이외에 다양한 요인들도 환자들의 질병 및 건강에 영향을 미칠 수 있으므로 향후 연구에서는 인구학적 및 사회경제적 요인 등 다양한 변수를 고려한 모델 개발이 필요할 것으로 사료된다.

셋째, 인과관계에서 원인과 결과 순서를 파악할 수 없는 변수가 다수 존재하였다. 현재 지속적 진료여부(Y), 만성질환 관리여부(Y), 장애관리 및 재활서비스 여부(N), 인플루엔자 예방접종여부(Y)는 고혈압과의 선후 관계를 파악할 수 없다는 횡단 조사와 벌점화 회귀분석의 단점을 드러낸다. 장애인 실태조사는 패널조사와 같은 종단 조사가 아니기 때문에, 고혈압 발병과 변수들 간의 선후 관계를 명확히 파악하기 어렵다. 또한 벌점화 회귀분석의 경우, 인과관계를 도출하려는 목적보다는 수많은 변수들 사이에서 종속변수에 상대적으로 큰 영향을 미치는 변수를 추출해내는 데 의의가 있기 때문에 인과 관계를 규명할 수 없다는 한계를 지닌다. '지난 2주간 아팠던 날수' 역시 조사표를 살펴보면 "고혈압, 당뇨 등 만성질환으로 계속 약을 복용하는 경우도 '아팠다'로 조사합니다."라고 기재되어 있다(보건복지부, 2018). 이미 고혈압을 앓고 있어 아팠던 날수를 기재한 것인지, 다른 질병의 유병으로 인한 것인지 파악하지 못하였다. 하지만 이 변수는 최대 14일까지의 연속형으로 측정되었고 아팠던 날이 길수록 고혈압에 걸릴 가능성이 높다고 해석할 수 있다. 향후 조사의 타당성을 위하여 사유를 묻는 문항을 추가하는 방안을 고려할 수 있다.

Ⅵ. 결론

본 연구는 국가승인통계인 2017년 장애인 실태조사의 마이크로 데이터를 이용하여 장애인의 고혈압 유병을 예측하였다. 벌점화 회귀분석(Ridge)와 Least Absolute Shrinkage and Selection Operator(LASSO)을 통해 선정된 변수를 3가지 모형(SVM, Logistic Regression, LightGBM)에 적용하여 고혈압 유병을 분류한 모델이다. 3가지 모형 중 최종 선정된 LightGBM 알고리즘에 적용하여 약 68% 정확도를 나타내는 모델을 제안하였으며, 대규모로 진행되는 국가승인통계 조사의 장점을 활용하여 새로운 변수를 발견하는 연구는 질병에 영향을 미친다고 알려진 기존의 변수들 외에 잠재적인 위험요인을 찾아 가시화할 수 있다는 점에서 가치가 있을 것으로 판단된다.

관련된 선행연구는 벌점화 회귀분석으로 변수의 영향력을 파악하거나 분류 모델을 구현하는 데에 그쳤다면, 본 연구는 벌점화 회귀분석을 이용하여 종속변수에 유의한 변수를 선정하고, 이를 기반으로 분류 모델을 제안하였다는 점에서 의의를 가진다.

또한 본 연구에서 제시된 가장 불편한 부위(무릎), 주된 진단명(관절질환), 지팡이 필요(Y), 지팡이 소지

(Y) 등 고혈압에 영향을 미치는 주요 변수는 지체장애인의 고혈압 예방을 위한 복합적인 관리와 선제적 예방 조치 등에 활용될 수 있을 것으로 사료된다. 또한 다른 장애유형이나 만성질환을 예측하는 후속 연구를 진행할 수 있으며, 만성질환과 관련된 변수를 확인하여 관리함으로써 장애인의 건강한 삶과 삶의 질 개선에 기여할 수 있을 것으로 기대한다.

참고문헌

국립재활원(2022, 10. 8.). **장애인 건강생활 건강체중 유지**. http://www.nrc.go.kr/portal/html/content.do?depth=hl&menu_cd=02_04

권수영, 김예순, 문종훈, 박재민, 백유진, 송이슬(2019). **2018 장애인 건강관리 사업**. 서울: 국립재활원.

권수영, 김예순, 이경현, 이범석, 호승희(2020). **2018년도 장애인건강보건통계, 그림으로 보는 주요 통계**. 서울: 국립재활원.

김도형 (2020). **서포트 벡터 머신.** https://datascienceschool.net/03%20machine%20learning/13.02%20%EC%84%9C%ED%8F%AC%ED%8A%B8%20%EB%B2%A1%ED%84%B0%20%EB%A8%B8%EC%8B%A0

김상용(2016). 당뇨병 환자의 고혈압 관리. **Journal of Korean Diabetes, 17**(2), 88-95.

김수영, 나은희, 조선(2018). 당뇨병에서의 동반질환 유병률 및 조합유병률. **보건정보통계학회지, 43**(3), 237-244.

김수정, 박세환, 서영성, 배철영, 신동학(1994). 비만의 유병율과 질환과의 관계. **가정의학회지, 15.**

김승수, 양광익(2018). 비만 폐쇄수면무호흡 환자에서 기계학습을 통한 적정양압 예측모형. **대한수면연구학회, 2384**(2423), 2384-2431.

김아름, 최민혁(2018). 장애여부가 고혈압 및 당뇨병 이환에 미치는 영향: 성향점수매칭법을 활용하여. **한국자료분석학회, 20**(3), 1503-1517.

김유경, 이정원, 김동호(2022). Elastic net 회귀분석을 활용한 고등학생들의 자기주도학습능력 예측요인 탐색. **한국교육문제연구, 40**(2), 83-109.

김은숙(2021). 골관절염이 동반된 고혈압 노인의 건강관련 삶의 질 영향 요인. **한국산학기술학회논문지, 22**(3), 169-180.

김지영, 강민욱, 서욱영, 이지원(2020). 장애인의 만성질환, 건강행태 및 사망위험: 국민건강보험공단 건강검진자료 분석. **보건사회연구, 40**(2), 121-150.

김지영, 송기호, 김동일(2020). 지체장애인을 위한 개별 운동프로그램 개발. **한국생활환경학회지, 27**(5), 586-597.

김한결, 최근호, 임성원, 이현실(2016). 국민건강영양조사를 활용한 대사증후군 유병 예측모형 개발을 위한 융복합 연구: 데이터마이닝을 활용하여. **한국디지털정책학회, 14**(2), 325-332.

박주완, 배진성, 윤혁준(2019). **빅데이터 분석 기법을 이용한 소상공인 신용평가 모형 구축 연구**. 대전: KOREG(신용보증재단중앙회).

보건복지부(2015. 11. 4.). **장애인 다빈도 질환 살펴보니 근골격계 질환, 고혈압, 당뇨 등 발생율 높아.** http://www.mohw.go.kr/react/al/sal0301vw.jsp?PAR_MENU_ID=04&MENU_ID=0403&CONT_SEQ=327194

보건복지부(2018. 9. 5.). **2017년 장애인실태조사.** http://www.mohw.go.kr/react/jb/sjb030301vw.jsp?PAR_MENU_ID=03&MENU_ID=032901&CONT_SEQ=345972

보건복지부(2019. 4. 28.). **장애등급? 이젠 장애정도로 바꾸세요 -행안부, 장애등급제 개편에 따른 자치법규 일제정비 추진-.** https://www.mois.go.kr/frt/bbs/type010/commonSelectBoardArticle.do?bbsId=BBSMSTR_000000000008&nttId=70372

보건복지부(2021. 4. 19.). **2020년 한 해 동안 새롭게 등록한 장애인 8만 3000명.** http://www.mohw.go.kr/react/al/sal0301vw.jsp?PAR_MENU_ID=04

&MENU_ID=0403&page=1&CONT_SEQ=365339

서울아산병원, **질환백과.** https://www.amc.seoul.kr/asan/healthinfo/disease/diseaseDetail.do?contentId=33883&tabIndex=0

서초구 보건소. **감기와 인플루엔자.** https://www.seocho.go.kr/site/sh/07/10702040100002015072310.jsp

송상윤(2015). 예대금리차 결정요인 모형의 예측력 비교 연구 - Ridge, LASSO 및 Elastic Net 방법론을 중심으로-. **금융지식연구, 13**(3), 41-65.

송재철(2013). **예방의학과 공중보건학(제2판).** 서울: 계축문화사.

신백균(2022). **Must Have 머신러닝 · 딥러닝 문제해결 전략 캐글 수상작 리팩터링으로 배우는 문제해결 프로세스와 전략**. 서울: 골든래빗(주).

안경숙(2005). 노인부부가구 · 노인독신가구의 사회적 지지가 삶의 질에 미치는 요인에 관한 연구. **한국노년학, 25**(1), 1-20.

양경순, 한재희(2013). 조손가정 조모의 심리적 경험-서울지역 기초생활수급대상 조손가정 조모를 중심으로. **상담학연구, 14**(2), 911-930.

엄영호, 황정윤, 정헌주(2014). 한국 ODA 의 구속성 여부에 관한 경험적 분석. **행정논총, 52**(1), 123-144.

윤미순, 최은희, 김유진, 최시은(2021). 가구 유형에 따른 노인의 건강 관련 삶의 질에 미치는 영향요인. **근관절건강학회지, 28**(2), 174-182.

이슬기, 신택수(2018). SVM 과 meta-learning algorithm 을 이용한 고지혈증 유병 예측모형 개발과 활용. **지능정보연구, 24**(2), 111-124.

이유현, 김윤진, 조덕영(2014). 노인에서 가구유형과 건강행태: 제 5 기 국민건강영양조사 (2010-2012). **보건의료산업학회지, 8**(4), 199-207.

이은숙(2021). 세대별 가구 유형이 건강생활습관, 의료 서비스 이용 및 건강결과에 미치는 효과. **보건교육건강증진학회지, 38**(3), 1-12.

이현미, 전교석, 장정아(2020). LightGBM 알고리즘을 활용한 고속도로 교통사고심각도 예측모델 구축. **한국전자통신학회 논문지, 15**(6), 1123-1130.

이혜재, 허순임(2017). 노인의 미충족 의료 경험의 결정요인-가구 유형을 중심으로. **보건경제와 정책연구, 23**(2), 85-108.

인하대병원 인천권역심뇌혈관질환센터(2020). **고혈압(안내책자).** 인하대병원, 13-19.

장애등급판정기준, 보건복지부고시 제2018-151호 (2018).

정우진(2020). 데이터 사이언스를 활용한 사회안전망 강화: 의료보장제도 가입자의 위험 예측 모형 구축. **정부학연구, 26**(2), 29-61.

정은경(2021). **2021 만성질환 현황과 이슈**. 청주:질병관리청.

조남훈(2006). 교차검증을 이용한 SVM 전력수요예측. **전기학회논문지A, 55**(11), 485-491

질병관리청 국가건강정보포털 (2021.01.13). **노인 고혈압.** https://health.kdca.go.kr/healthinfo/biz/health/gnrlzHealthInfo/gnrlzHealthInfo/gnrlzHealthInfoView.do?cntnts_sn=4687

최용욱, 윤대웅, 최준환, 변중무(2020). 베이지안 최적화를 이용한 암상 분류 모델의 하이퍼 파라미터 탐색. **지구물리와 물리탐사, 23**(3), 157-167.

최태정, 박지인, 손상원, 윤주범(2022). 네트워크 이상탐지를 위한 트리 기반과 인공 신경망 모델 성능 비교 연구. **한국통신학회 학술대회논문집,** 1321-1322.

호승희(2017). **2017 장애인백서.** 한국장애인개발원.

Albon, C. (2019). **파이썬을 활용한 머신러닝 쿡북**. 서울: 한빛미디어.

Garcia-Carretero, R., Barquero-Perez, O., Mora-Jimenez, I., Soguero-Ruiz, C., Goya-Esteban, R.,& Ramos-Lopez, J. (2019). Identification of clinically relevant features in hypertensive patients using penalized regression: a case study of cardiovascular events. *Medical & Biological Engineering & Computing, 57*, 2011-2026.

Gee, E. M. (2000). Living arrangements and quality of life among Chinese Canadian elders. *Social Indicators Research, 51,* 309-329.

Heo, J. S., Kwon, D. H., Kim, J. B., Han, Y. H., &

An, C. H. (2018). Prediction of cryptocurrency price trend using gradient boosting. *KIPS Transactions on Software and Data Engineering, 7*(10), 387-396.

Kan, H. J., Kharrazi, H., Chang, H. Y., Bodycombe, D., Lemke, K., & Weiner, J. P. (2019). Exploring the use of machine learning for risk adjustment: A comparison of standard and penalized linear regression models in predicting health care costs in older adults. *PloS one, 14*(3), e0213258.

Python API, lightgbm.plot_importance(2022). https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.plot_importance.html

Scikit Learn developers(2022a). 1.4. Support Vector Machines. https://scikit-learn.org/stable/modules/svm.html

Scikit Learn developers(2022b). sklearn.linear_model.

LogisticRegression. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

논문투고일 : 2023. 11. 14.
심 사 일 : 2023. 11. 26.
게재확정일 : 2023. 12. 08.