

RESEARCH

Open Access



Machine learning based classification of catastrophic health expenditures: a cross-sectional study of Korean low-income households

Seok Min Ji^{1†}, Jeewuan Kim^{2†} and Kyu Min Kim^{2*}

Abstract

Background Despite the National Health Insurance (NHI) system implemented in South Korea, concerns persist regarding access to health coverage for low-income households. To address this issue, this study aims to use machine learning-based data mining techniques to classify whether such households will face catastrophic health expenditures (CHEs).

Methods A total of 4,031 low-income people were extracted using 2019 data from the Korea Health Panel Survey. The classification model was developed using four machine learning algorithms: Random Forest, Gradient boosting, Decision tree, Ridge regression, Neural network, and AdaBoost. Ten-fold cross validation was carried out to ensure the reliability of the analysis results. The model was evaluated based on the Area Under Receiver Operating Characteristics (AUROC) as well as accuracy, precision, recall, and F-1 score.

Results The study's findings revealed that the incidence of CHE was 26.2% in low-income households. The AdaBoost model had the highest classifiable power. It showed AUROC of 89.8%, accuracy of 83.1%, precision of 82.4%, recall of 83.1, and F1 score of 82.1%. The study found that economic activity, chronic disease, and age were significant factors that could lead to CHEs. Therefore, individuals over 65, with chronic conditions, and unemployed had the highest likelihood of developing CHE.

Conclusion It is essential to identify low-income households that are at risk of CHEs in advance before facing the economic burden. This research is expected to provide fundamental data that can aid in developing an integrated support program to prevent and manage CHEs more effectively.

Keywords Catastrophic health expenditure, CHE, Machine learning, Population health, Health policy

[†]Seok Min Ji and Jeewuan Kim contributed equally to this work.

*Correspondence:

Kyu Min Kim
perves@gtec.ac.kr

¹Department of Cancer AI and Digital Health, Graduate School of Cancer Science and Policy, National Cancer Center, Siheung, Gyeonggi-do, Republic of Korea

²Department of Health Administration, Gyeonggi College of Science and Technology, Siheung, Gyeonggi-do, Republic of Korea



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Background

Catastrophic health expenditures (CHEs) have been an everlasting phenomenon worldwide. Although South Korea provides health security with the National Health Insurance, the social insurance covering all citizens and managed by a single insurer, issues regarding health security have constantly been raised, particularly in low-income households [1, 2]. Thus, it is necessary to measure how much economic loss a household suffers from medical expenses, analyzing its CHEs; the CHE is a representative indicator of how much medical expenses have been spent compared to the household's ability to pay, which means that the share of medical expenses in household income is above a certain level [3].

According to previous studies, it is reported that when CHE occurs, the risk of financial hardships and poverty increases [4]. Most importantly, expenses tend to be high in low-income households since medical services are associated with low-income elasticity [5]. As the high medical expenditure is unavoidable, rising medical expenses are becoming fatal in certain low-income households [6]. According to Choi et al., between 2010 and 2015, the rate of CHE in Korea decreased from 0.9 to 0.5% for households earning 150% or more of the standard median income; however, households earning less than 50% of the standard median income increased from 10.0 to 12.8%, verifying the proportion of CHEs in low-income groups was high [7]. The incidence of CHEs is expected to be low when public health expenses are high, and the proportion of out-of-pocket (OOP) medical expenses is low; however, in the case of Korea, it is reported that there are numerous non-coverage items and high co-payments due to rapid population aging, an increase in chronic diseases, and the emergence of new medical technology [8]. Comparing the size of household OOP payments by Organization for Economic Cooperation and Development (OECD) major countries and individuals over the past three years, South Korea ranked second in 2020, third in 2021, and 2022. It was reported to be the country with the 4th highest medical expenses [9].

A combination of various factors causes CHEs, and it is known that age, number of household members, economic activity, chronic diseases, and subjective health awareness are affected [10–13]. According to Zhang et al., it is likely to experience CHEs for those with lower income levels, poorer subjective health awareness; have low economic activity; have several chronic diseases, and aged [10]. The risk of CHEs was higher in Malaysia if women were the heads of the household; families lived in rural areas and had fewer household members [11]. Moreover, those having lower educational levels, experiences of depression, and experiences of being unsatisfied with medical services have been known to have a higher

probability of CHEs [12, 13]. Until now, it appears that most of the studies analyzed factors affecting the occurrence of CHEs targeting general adults or patients with specific diseases.

Therefore, the aim of this study is to classify the occurrence of CHE in low-income households using data mining techniques. Hence, it is essential to detect risk factors in advance before CHEs occur and prevent them from falling into poverty through appropriate management and support [14]. Data mining has higher accuracy in classifying results than other methodologies and is widely used in diagnosing and classifying various diseases in the medical field [15]. More importantly, this study should be able to provide fundamental data for developing an integrated program to prevent and manage CHEs in low-income households in Korea.

Methods

Data source and study population

This research utilized the Korea Health Panel Survey (KHPS), which is a nationally conducted survey that provides in-depth information on household members' medical usage patterns, medical expenses, and factors influencing medical usage and expenses. The study's final analysis involved adults who were 18 years or older, utilizing data from the 2019 KHPS based on specific criteria. Firstly, individuals who had previously received medical services such as hospitalization, outpatient services, and emergency visits were excluded. Secondly, those whose income exceeded 100% of the standard median income in 2019 were also excluded. Standard median income refers to the person standing in the center when all citizens are lined up in order of income [16]. The Korean government's CHE program is a well-known medical expense support service for socially vulnerable groups, primarily for individuals or households with a standard median income below 100%. Therefore, only those with low-income households were selected as participants based on the income standard for the CHE program. In the final analysis, a total of 4,031 participants were included (Fig. 1).

Data availability and ethics statement

The KHPS data are available, and researchers can download the raw data upon request (<https://www.khp.re.kr:444>). The raw KHPS data do not include any personal information. This study was reviewed and approved by the Korea University Institutional Review Board (IRB No. 2023-0043).

Variables

Target variables

The focus of this study is to determine whether or not CHE has been incurred. The most widely known Wagstaff

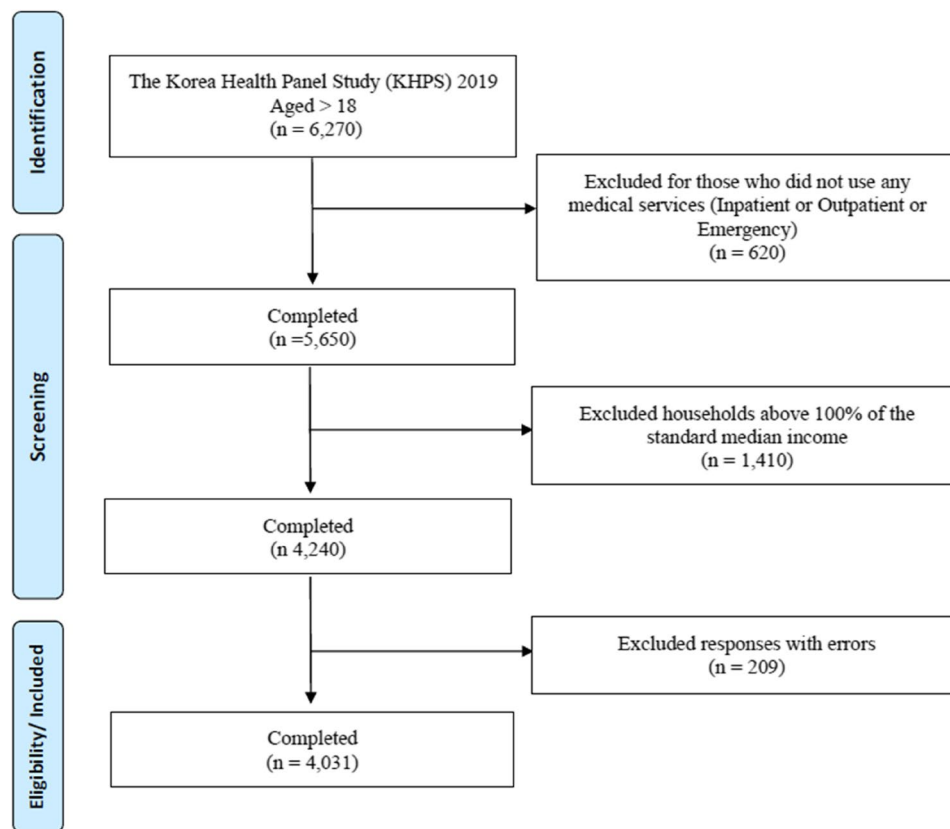


Fig. 1 Flow diagram of the participants selection

& van Doorslaer method, the standard measure to quantify income-related inequalities in health economics, was used [17]. The total household income is used as the denominator, which is generally defined as the ability to pay. The KHPS provides the total household income as one variable, including all income within the household, such as earned income, property income, pension income, financial income, private transfers, etc. The OOP expenses, which correspond to the numerator, mainly include the minimum OOP expenses at medical institutions [18]. The OOP pay comprises emergency medical expenses, inpatient medical expenses, outpatient medical expenses, emergency prescription drug costs, inpatient prescription drug costs, and outpatient prescription drug costs. Overall, CHEs are considered to have occurred when the proportion of OOP expenses compared to the ability to pay exceeded a certain level [17, 19].

$$\text{Catastrophic Health Expenditure} = \frac{\text{Out of Pocket}}{\text{Total household income}} \geq 10\%$$

This study determined that CHE occurs when medical expenses exceed 10% of ability to pay based on previous research [18, 20]. There is no clear consensus on whether CHEs have occurred so far [21]. In general, 10% and 40%

thresholds are used a lot [22, 23]. When comparing CHE rates between countries, 40% is mainly used, and 10% is used when research is conducted on a small individual basis [24]. However, it is noted that this threshold can be interpreted differently depending on the circumstances of a country [25]. The threshold level of 10% is widely used as a minimum baseline in numerous studies [17, 19–21]. Therefore, to increase comparability between studies, this study also selected 10% as the threshold level.

Input variables

The study used certain input factors which are explained below. Gender was divided into male (0) and female (1), and age was 18–64 (0) and over 65 (1), based on 65 years of age. The education level was categorized as elementary school (0), middle school (1), high school (2), and bachelor's degree or higher (3). The household size was classified as one person (0), two people (1), and three or more people (2). Marital status was categorized as married (0) and single (1). The economic activity was categorized as employed (0) and unemployed (1). In addition, subsidies, which are cash services provided by government and private organizations, were classified as either received (0) or not received (1). Private medical insurance was categorized as enrolled (0) or not enrolled (1). Disability, chronic disease, unmet medical experience,

and depression were classified as either present (0) or absent (1). Further, the study categorized participants into those with and without disabilities based on whether a physician assigned them a disability grade. Chronic disease was defined as the use of medication for more than six months [26]. Assessing the unmet medical care, the participants were asked whether required any treatment or examination at a hospital or clinic in the past year but could not receive it, excluding dentistry services. An unmet medical need was recorded if the answer to the question was yes, indicating one or more experiences that did not receive the required medical care. Depression, defined as a subjective feeling that individuals, was assessed as present or absent. Lastly, perceived health awareness was measured on a 5-point Likert scale with the question, 'How do you think your health is in general?', with higher scores indicating higher subjective health awareness.

Statistical analysis

The frequency analysis, chi-square test, fisher test and machine learning techniques were used for this study. First, frequency analysis and chi-square test, fisher test were performed to verify whether statistical differences occurred depending on the demographic characteristics of the study participants and the occurrence of CHE. Second, Random Forest, Gradient boosting, Decision tree, Ridge regression, Neural network and AdaBoost were used to build a classification model for the occurrence of CHE for low-income households. Neural Network is one of the most widely used machine learning methodologies that classify the category of a target factor by combining input factors with a non-linear model, delivering to each hidden unit while providing the combination of hidden units to the output unit [27]. In this study, multilayer perceptron of neural networks was used, and a backpropagation algorithm was used to reduce errors [learning ratio: 9:1, batch size: 10, epochs: 500]. Decision Tree classifies the categories of target factors by charting decision rules as a tree structure. Since it is expressed in a tree structure, it is easy to interpret the classification results and has the advantage of being able to obtain information on significant classifiable factors [28]. In this study, the CART (classification and regression tree) algorithm was used. Specifically, the Gini Index was used as the separation criterion, and the maximum tree depth was designated as 100, and the number of cases of the upper and lower nodes was designated as 20 and 5, respectively. Ridge Regression is a well-established method to tackle the multi collinearity problem [29]. It involves the introduction of some bias into the regression equation to reduce the variance of the estimators of the parameters. Random forest has been reported to have relatively high predictive power and model stability when performed on

data that contains a large number of input variables [30]. Detailed tuning was conducted to find the most suitable model in this study [n_estimators:10, max_features: auto, min_samples_split:5, others are set to defaults]. Gradient boosting has been reported to have a low probability of overfitting and a high classifiable performance when combined with a variety of algorithms, such as decision trees, neural networks based on boosting [31]. In this study, a combination of hyperparameters representing optimal results was derived through random search [number of trees: 100, learning rate: 0.1, limit depth of individual trees: 3, do not split subsets smaller than: 2, others are set to defaults]. AdaBoost also known as Adaptive Boosting, is a boosting algorithm that is widely used to refine imbalanced data and used as an Ensemble Method in Machine Learning. AdaBoost is an advanced ensemble algorithm with a high detection rate and is less prone to overfitting [32]. In this study, a decision tree was used to increase the speed and performance of weak learners [base estimator: tree, number of estimators: 50, learning rate: 1.0, classification algorithm: SAMME.R].

To develop and evaluate the classifiable model, the k-fold cross validation method was applied to ensure the reliability of the results. We adopted the tenfold cross validation in which the original sample is randomly partitioned into 10 subsamples of equal size. A single subsample is retained as the validation data for testing the model, and the remaining nine subsamples are used as training data. And then the cross-validation process is repeated ten times with each of the ten subsamples employed exactly once for the validation. And feature selection to determine the relative importance of classifiable factors that led to the occurrence of CHEs in low-income households, Wrapper's stepwise backward elimination was employed to identify the optimal model. This step involved removing the least relevant factors in a sequential manner. This model was evaluated according to its Area Under the Receiver Operating Characteristic (AUROC) as well as its accuracy, precision, recall, and F-1 score. The analysis was performed using R and SAS Enterprise Miner.

Results

General characteristics

Table 1 presents the demographic characteristics of the study participants and the differences in occurrences of CHE. First of all, regarding gender, that men ($n=2,775$; 68.8%) accounted for a larger proportion than women ($n=1,256$; 31.2%). In terms of age distribution, 2,806 individuals, making up 69.6% of the population, were over the age of 65. In addition, 1,321 people (32.8%) had an elementary school diploma, while 1,178 people (29.2%) had a high school diploma as the highest level of education. Two people were the most common when asked

Table 1 General characteristics of the participants and differences in CHEs

Variables		N=4,031 (100%)	Occurrence of CHE (26.2%)	χ^2
Sex	Men	2775 (68.8)	17.1	7.600**
	Women	1256 (31.2)	9.1	
Age	18–64	1225 (30.4)	3.9	159.511***
	65 and older	2806 (69.6)	22.3	
Educational level	elementary school	1321 (32.8)	10.9	77.530***
	middle school	765 (19.0)	5.5	
	high school	1178 (29.2)	6.6	
	college or higher	767 (19.0)	3.2	
Number of Household members	1	1079 (26.8)	7.7	115.089***
	2	1892 (46.9)	14.8	
	3 or more people	1060 (26.3)	3.7	
Marital status	Married	2471 (61.3)	15.8	0.647
	Single	1560 (38.7)	10.4	
Economic status	Employed	2421 (60.1)	12.1	118.407***
	Unemployed	1610 (39.9)	14.1	
Subsidies	yes	3610 (89.6)	24.7	34.817***
	no	421 (10.4)	1.5	
Private health insurance	yes	2350 (58.3)	12.8	50.876***
	no	1681 (41.7)	13.4	
Disability	yes	457 (11.3)	4.4	43.162***
	no	3574 (88.7)	21.8	
Chronic disease	yes	3100 (76.9)	23.5	129.876***
	no	931 (23.1)	2.7	
Unmet medical needs	yes	526 (13.0)	3.9	4.554*
	no	3505 (87.0)	22.3	
Depression	yes	359 (8.9)	2.9	9.001**
	no	3672 (91.1)	23.3	

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

about the number of household members, at 1,892 people (46.9%). Regarding marital status, 2,471 (61.3%) were married, and 2,421 (60.1%) were revealed as employed. In relation to subsidies, 3,610 people (89.6%) showed that most were receiving subsidies. Regarding private medical insurance enrollment, 2,350 people (58.3%) accounted for more than half; most people ($n = 3,574$; 88.7%) reported having no disability. Further, 3,100 people (76.9%) indicated having chronic diseases, and 3,505 people (87.0%) stated not having unmet medical needs. Additionally, it was demonstrated that approximately 91% of people did not have depression. The study participants had a perceived health awareness level of 3.1, which was higher than the average (1–5 points, SD: 0.87). Lastly, after examining the demographic characteristics of the study participants and the occurrence of CHEs, significant differences were found in all variables except marital status ($\chi^2 = 0.647$, $p < 0.005$).

Feature importance

The relative importance of classifiable factors contributing to classifying CHE in low-income households, which is the purpose of this study, was examined using feature selection. The higher the importance ranking of

Table 2 The significance of the variables that impact the CHEs

Variables	Ranking	Variables	Ranking
Economic activities	1	Education level	8
Chronic disease	2	Subsidies	9
Age	3	Depression	10
Marital status	4	Perceived health awareness	11
Number of household members	5	Disability	12
Private health insurance	6	Unmet medical needs	13
Gender	7		

the classifiable factor, the greater its influence on classifying CHE. Among the variables in this study, the highest top ranking was confirmed to be the economic activities, chronic disease, age. This finding suggests that economic activities, chronic disease, age are the most significant factors in classifying low-income CHE compared to other factors. Subsequently, marital status and the number of household members were found to be ranked high. On the other hand, perceived health awareness, disability, and unmet medical needs were displayed to be distributed in low rankings Table 2.

Classification model performance

This study employed various classification techniques, including random forest, gradient boosting, decision tree, ridge regression, neural network, adaBoost machine analysis, to develop a classification model that accurately classifies CHE in low-income households. Tables 3 and 4 displays the results of the classification analysis. The method used was Wrapper’s stepwise elimination, which was applied sequentially to the factors’ relative importance identified in Table 2. The analysis revealed that the adaBoost algorithm had a higher AUROC than the other three algorithms. Specifically, the AUROC for the Random Forest was 84.9%, for Ridge Regression analysis, 68.7%, for Gradient boosting, 73.5%, for Decision tree, 84.5%, and for Neural network, 78.5%. As the number of input factors increases in decision trees, the AUROC tends to improve. When all 13 factors were used, the AdaBoost demonstrated 89.8% AUROC, 83.1% Accuracy, 82.4% Precision, 83.1% Recall, 82.1% F1 score (Fig. 2). To ensure that the main outcome was reliable and robust, a sensitivity analysis was conducted by dividing the dependent variable, CHE incidence, into three thresholds (20%, 30%, 40%); the analysis revealed that the main outcome did not change in Tables 5, 6, 7, 8, 9 and 10.

Discussion

The purpose of this research was to investigate whether certain factors were linked to the appearance of CHEs in low-income households. Machine learning techniques were used to develop an accurate classification model. The analysis showed that the AdaBoost was the best model for classifying the occurrence of CHEs in low-income households. The study also identified several significant factors that influenced the incidence of CHEs, such as economic activity, chronic disease, age, the number of household members, marital status, disability status, and perceived health awareness. The main research findings can be summarized as follows.

The study’s findings revealed that the incidence of CHE was 26.2% in low-income households. This is a significant difference compared to the incidence of CHE, which was 3.9% (in 2020, with a threshold of 40% or higher) among the general population in Korea [33]. Looking at the characteristics of the population that experienced CHE in this study, it was found that older adults, those with lower education levels, those with chronic illnesses, and those receiving subsidies experienced CHEs. In summary, CHE occurs more frequently in specific groups than in general groups. Importantly, the characteristics of the group experiencing CHE need to be identified and regularly investigated and managed.

After applying Wrapper’s step-by-step elimination method, the AdaBoost was found to have the highest classifiable power (89.8%) for determining the relative

Table 3 Distribution of AUROC, accuracy, precision, recall, F1 score by machine learning algorithms

No	Random forest					Ridge regression					Gradient boosting				
*	AUROC	Accuracy	Precision	Recall	F1	AUROC	Accuracy	Precision	Recall	F1	AUROC	Accuracy	Precision	Recall	F1
1	0.591	0.735	0.541	0.735	0.623	0.591	0.735	0.541	0.735	0.623	0.591	0.735	0.541	0.735	0.623
2	0.633	0.735	0.541	0.735	0.623	0.633	0.735	0.541	0.735	0.623	0.633	0.735	0.541	0.735	0.623
3	0.657	0.735	0.541	0.735	0.623	0.656	0.735	0.541	0.735	0.623	0.657	0.735	0.541	0.735	0.623
4	0.667	0.735	0.541	0.735	0.623	0.663	0.735	0.541	0.735	0.623	0.667	0.735	0.541	0.735	0.623
5	0.678	0.735	0.541	0.735	0.623	0.666	0.735	0.541	0.735	0.623	0.678	0.735	0.541	0.735	0.623
6	0.688	0.735	0.541	0.735	0.623	0.667	0.735	0.541	0.735	0.623	0.685	0.735	0.541	0.735	0.623
7	0.704	0.738	0.704	0.738	0.639	0.670	0.736	0.710	0.736	0.627	0.693	0.737	0.725	0.737	0.628
8	0.736	0.745	0.712	0.745	0.676	0.672	0.736	0.707	0.736	0.626	0.701	0.736	0.718	0.736	0.625
9	0.749	0.749	0.724	0.749	0.678	0.672	0.736	0.707	0.736	0.626	0.704	0.737	0.736	0.737	0.630
10	0.767	0.755	0.732	0.755	0.697	0.675	0.736	0.683	0.736	0.627	0.706	0.738	0.741	0.738	0.632
11	0.823	0.784	0.771	0.784	0.751	0.681	0.737	0.695	0.737	0.636	0.722	0.741	0.702	0.741	0.661
12	0.834	0.793	0.782	0.793	0.765	0.686	0.733	0.660	0.733	0.639	0.729	0.742	0.701	0.742	0.673
13	0.849	0.805	0.796	0.805	0.785	0.687	0.731	0.654	0.731	0.638	0.735	0.744	0.708	0.744	0.677

*Input variables, 1: economic activities, 2: 1+ chronic disease, 3: 2+ age, 4: 3+ marital status, 5: 4+ number of household members, 6: 5+ private health insurance, 7: 6+ gender, 8: 7+ education level, 9: 8+ subsidies, 10: 9+ depression, 11: 10+ perceived health awareness, 12: 11+ disability, 13: 12+ unmet medical needs

Table 4 Distribution of AUROC, accuracy, precision, recall, F1 score by machine learning algorithms (continued)

No	Decision tree					AdaBoost					Neural network				
	AUROC	Accuracy	Precision	Recall	F1	AUROC	Accuracy	Precision	Recall	F1	AUROC	Accuracy	Precision	Recall	F1
*															
1	0.591	0.735	0.541	0.735	0.623	0.591	0.735	0.541	0.735	0.623	0.591	0.735	0.541	0.735	0.623
2	0.633	0.735	0.541	0.735	0.623	0.633	0.735	0.541	0.735	0.623	0.633	0.735	0.541	0.735	0.623
3	0.657	0.735	0.541	0.735	0.623	0.657	0.735	0.541	0.735	0.623	0.656	0.735	0.541	0.735	0.623
4	0.667	0.735	0.541	0.735	0.623	0.667	0.735	0.541	0.735	0.623	0.667	0.735	0.541	0.735	0.623
5	0.679	0.735	0.541	0.735	0.623	0.679	0.735	0.541	0.735	0.623	0.676	0.735	0.541	0.735	0.623
6	0.689	0.736	0.700	0.736	0.625	0.689	0.736	0.700	0.736	0.625	0.683	0.735	0.541	0.735	0.623
7	0.705	0.738	0.720	0.738	0.636	0.707	0.739	0.716	0.739	0.638	0.694	0.737	0.725	0.737	0.628
8	0.737	0.744	0.712	0.744	0.668	0.748	0.749	0.723	0.749	0.682	0.706	0.738	0.703	0.738	0.638
9	0.750	0.748	0.720	0.748	0.680	0.766	0.756	0.733	0.756	0.701	0.720	0.741	0.702	0.741	0.660
10	0.764	0.754	0.734	0.754	0.690	0.788	0.766	0.753	0.766	0.716	0.731	0.741	0.715	0.741	0.649
11	0.815	0.775	0.762	0.775	0.735	0.858	0.802	0.793	0.802	0.779	0.754	0.755	0.727	0.755	0.711
12	0.832	0.786	0.777	0.786	0.751	0.878	0.817	0.810	0.817	0.798	0.769	0.763	0.746	0.763	0.711
13	0.845	0.793	0.786	0.793	0.763	0.898	0.831	0.824	0.831	0.821	0.785	0.772	0.760	0.772	0.726

* Input variables, 1: economic activities, 2: 1 + chronic disease, 3: 2 + age, 4: 3 + marital status, 5: 4 + number of household members, 6: 5 + private health insurance, 7: 6 + gender, 8: 7 + education level, 9: 8 + subsidies, 10: 9 + depression, 11: 10 + perceived health awareness, 12: 11 + disability, 13: 12 + unmet medical needs

importance of influencing factors when classifying CHEs in low-income households. These findings are similar to studies with the highest predictive power of AdaBoost (95.30%) by applying five machine learning techniques (logistic regression, support vector machine, random forest, XGBoost, and AdaBoost) to predict employee promotion [34]. Another study compared various algorithms such as naive bays and decision tree for heart disease prediction and found that the ensemble technique adaBoost showed the highest accumulation when applying adaBoost [35]. In addition, a study that made customer turn predictions in the e-commerce industry reported that adaBoost showed the highest predictive power [32].

After using the Feature Selection to identify major risk factors associated with a high likelihood of developing CHEs, economic activity, chronic disease and age emerged as critical factors. According to another study that examined the trends in the incidence of CHEs, older people with chronic diseases were identified as the group with the highest risk of incurring CHEs [7]. In a study that compared the factors affecting the occurrence of CHEs between urban and rural areas, the main common factor was found to be employment status, whether the person was employed or unemployed [5]. Specifically, the study revealed that the incidence of CHEs tends to increase in low-income households [27, 36].

Consequently, AdaBoost can effectively handle complex predictors and show superior performance compared to other classification techniques. With that said, it may be beneficial to systematically screen and manage risk groups for CHEs. The community can be screened in two stages, with the first stage identifying groups at risk of CHEs, such as those who frequent public health centers and community-care centers. In the second stage, systematic management can be provided to prevent unnecessary medical expenses by identifying the causes of CHEs and fostering improvements through in-depth consultation and appropriate support. That said, it may be essential to establish social networks to maintain continuous management with the local community. In the long term, a health management program will be necessary for factors that can be modified through education, such as lifestyle habits. There is a policy to systematically manage chronic diseases such as high blood pressure and diabetes with a focus on primary health care in South Korea. Therefore, it may be worth considering expanding the target and scope of the policy.

While private health insurance did not seem to be a major factor in this study, its importance should not be overlooked. Previous research has found that individuals with private health insurance are less likely to face financial ruin from medical expenses compared to those without coverage [26, 37].

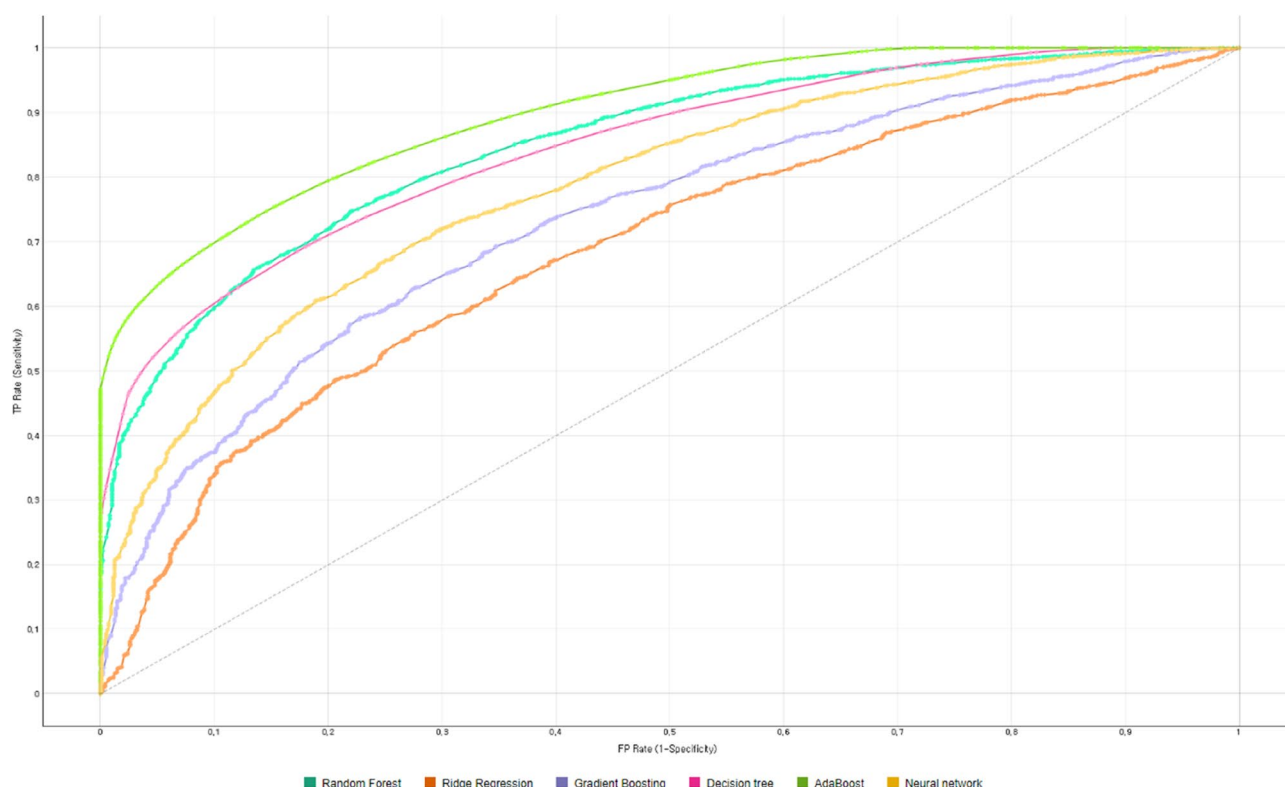


Fig. 2 The AUROC of CHE classification models using various machine learning algorithms

Above all, active support was found crucial for enabling low-income households to engage in economic activity. Thus, it is necessary to implement a support policy that encourages independent economic activity rather than simply providing subsidies for medical expenses in the long run. In South Korea, low-income households have access to support services for medical and living expenses, but the scale of support could be higher. In addition, substantial services cannot be expected as services that enable self-reliance are not provided. Therefore, taking proactive steps beyond providing income support and implementing various measures to facilitate economic activities is necessary. An expansion of a government initiative, the Senior Employment Project, could be a plausible option to consider. This scheme offers employment counseling and job-specific training to help seniors become independent. It also provides job information and resources that cater to the unique needs and abilities of the older population [38]. The results of this study can help expand support by considering the characteristics of the risk group for CHEs.

Finally, the limitations of the study are as follows. First, this study was conducted as a cross-sectional study, and it is difficult to identify the causal relationship over time. Second, much detailed characteristics of low-income

households were not considered in the impact of medical expenses. For instance, within the low-income households, the relative severity varies depending on age, gender, residence, and medical insurance type; therefore, the occurrence of CHEs is likely to show different patterns. In future studies, it will be necessary to classify and analyze these factors. Third, one of the main limitations is the generalizability of the findings. This is because there is a lack of research on creating classification models for the occurrence of CHEs targeting low-income households. Thus, it was difficult to compare whether the results of this study were reasonably understandable. Fourth, the group can change depending on how you configure the Threshold setting (10-40%). Using a threshold of 40% instead of 10% identifies less risky factors and risk groups for CHE. As this study focuses on low-income households that are vulnerable to CHE, we decided that 10% is appropriate to identify more sensitive groups. Fifth, although the AUROC in this study is similar when compared to previous studies, the relatively disproportionate number of CHE experiences may be responsible for the reduced AUROC. Nevertheless, the study holds significance as it examines CHEs in low-income households and identifies risk groups by developing the classification model.

Table 5 Sensitivity analysis results (threshold: 20%)

No	Random forest					Ridge regression					Gradient boosting				
	AUROC	Accuracy	Precision	Recall	F1	AUROC	Accuracy	Precision	Recall	F1	AUROC	Accuracy	Precision	Recall	F1
*	0.600	0.880	0.775	0.880	0.824	0.600	0.880	0.775	0.880	0.824	0.600	0.880	0.775	0.880	0.824
1	0.600	0.880	0.775	0.880	0.824	0.600	0.880	0.775	0.880	0.824	0.600	0.880	0.775	0.880	0.824
2	0.640	0.880	0.775	0.880	0.824	0.640	0.880	0.775	0.880	0.824	0.640	0.880	0.775	0.880	0.824
3	0.656	0.880	0.775	0.880	0.824	0.652	0.880	0.775	0.880	0.824	0.656	0.880	0.775	0.880	0.824
4	0.667	0.880	0.775	0.880	0.824	0.650	0.880	0.775	0.880	0.824	0.666	0.880	0.775	0.880	0.824
5	0.680	0.880	0.775	0.880	0.824	0.650	0.880	0.775	0.880	0.824	0.679	0.880	0.775	0.880	0.824
6	0.690	0.880	0.775	0.880	0.824	0.652	0.880	0.775	0.880	0.824	0.685	0.880	0.775	0.880	0.824
7	0.708	0.881	0.895	0.881	0.826	0.657	0.880	0.775	0.880	0.824	0.691	0.880	0.775	0.880	0.824
8	0.754	0.881	0.845	0.881	0.832	0.657	0.880	0.775	0.880	0.824	0.702	0.881	0.895	0.881	0.826
9	0.771	0.883	0.861	0.883	0.834	0.658	0.880	0.775	0.880	0.824	0.708	0.880	0.775	0.880	0.824
10	0.790	0.884	0.868	0.884	0.836	0.663	0.880	0.775	0.880	0.824	0.718	0.881	0.895	0.881	0.826
11	0.853	0.890	0.881	0.890	0.850	0.668	0.880	0.775	0.880	0.824	0.729	0.881	0.896	0.881	0.827
12	0.866	0.893	0.885	0.893	0.859	0.674	0.880	0.775	0.880	0.824	0.736	0.882	0.881	0.882	0.829
13	0.880	0.897	0.892	0.897	0.865	0.674	0.880	0.775	0.880	0.824	0.743	0.882	0.881	0.882	0.829

*Input variables, 1: economic activities, 2: 1+ chronic disease, 3: 2+ age, 4: 3+ marital status, 5: 4+ number of household members, 6: 5+ private health insurance, 7: 6+ gender, 8: 7+ education level, 9: 8+ subsidies, 10: 9+depression, 11: 10+ perceived health awareness, 12: 11+ disability, 13: 12+ unmet medical needs

Table 6 Sensitivity analysis results (threshold: 20%) (continued)

No	Decision tree					AdaBoost					Neural network				
	AUROC	Accuracy	Precision	Recall	F1	AUROC	Accuracy	Precision	Recall	F1	AUROC	Accuracy	Precision	Recall	F1
*	0.600	0.880	0.775	0.880	0.824	0.600	0.880	0.775	0.880	0.824	0.600	0.880	0.775	0.880	0.824
1	0.600	0.880	0.775	0.880	0.824	0.600	0.880	0.775	0.880	0.824	0.600	0.880	0.775	0.880	0.824
2	0.639	0.880	0.775	0.880	0.824	0.640	0.880	0.775	0.880	0.824	0.640	0.880	0.775	0.880	0.824
3	0.652	0.880	0.775	0.880	0.824	0.656	0.880	0.775	0.880	0.824	0.656	0.880	0.775	0.880	0.824
4	0.660	0.880	0.775	0.880	0.824	0.667	0.880	0.775	0.880	0.824	0.666	0.880	0.775	0.880	0.824
5	0.669	0.880	0.775	0.880	0.824	0.682	0.880	0.775	0.880	0.824	0.673	0.880	0.775	0.880	0.824
6	0.678	0.880	0.775	0.880	0.824	0.694	0.880	0.775	0.880	0.824	0.683	0.880	0.775	0.880	0.824
7	0.689	0.881	0.895	0.881	0.826	0.712	0.881	0.856	0.881	0.827	0.691	0.880	0.775	0.880	0.824
8	0.720	0.881	0.872	0.881	0.827	0.767	0.884	0.857	0.884	0.840	0.717	0.881	0.895	0.881	0.826
9	0.734	0.882	0.870	0.882	0.829	0.790	0.887	0.870	0.887	0.847	0.725	0.881	0.895	0.881	0.826
10	0.748	0.883	0.865	0.883	0.834	0.814	0.892	0.882	0.892	0.857	0.741	0.881	0.895	0.881	0.826
11	0.803	0.887	0.869	0.887	0.846	0.887	0.904	0.895	0.904	0.883	0.772	0.883	0.886	0.883	0.831
12	0.820	0.889	0.876	0.889	0.851	0.907	0.911	0.904	0.911	0.894	0.784	0.883	0.886	0.883	0.831
13	0.836	0.892	0.877	0.892	0.858	0.925	0.918	0.916	0.918	0.902	0.805	0.888	0.887	0.888	0.845

*Input variables, 1: economic activities, 2: 1+ chronic disease, 3: 2+ age, 4: 3+ marital status, 5: 4+ number of household members, 6: 5+ private health insurance, 7: 6+ gender, 8: 7+ education level, 9: 8+ subsidies, 10: 9+depression, 11: 10+ perceived health awareness, 12: 11+ disability, 13: 12+ unmet medical needs

Table 7 Sensitivity analysis results (threshold: 30%)

No	Random forest					Ridge regression					Gradient boosting				
	AUROC	Accuracy	Precision	Recall	F1	AUROC	Accuracy	Precision	Recall	F1	AUROC	Accuracy	Precision	Recall	F1
*															
1	0.611	0.940	0.884	0.940	0.912	0.611	0.940	0.884	0.940	0.912	0.611	0.940	0.884	0.940	0.912
2	0.618	0.940	0.884	0.940	0.912	0.618	0.940	0.884	0.940	0.912	0.618	0.940	0.884	0.940	0.912
3	0.654	0.940	0.884	0.940	0.912	0.651	0.940	0.884	0.940	0.912	0.655	0.940	0.884	0.940	0.912
4	0.667	0.940	0.884	0.940	0.912	0.661	0.940	0.884	0.940	0.912	0.668	0.940	0.884	0.940	0.912
5	0.698	0.940	0.884	0.940	0.912	0.666	0.940	0.884	0.940	0.912	0.697	0.940	0.884	0.940	0.912
6	0.726	0.940	0.884	0.940	0.912	0.666	0.940	0.884	0.940	0.912	0.721	0.940	0.884	0.940	0.912
7	0.752	0.941	0.944	0.941	0.913	0.670	0.940	0.884	0.940	0.912	0.730	0.940	0.884	0.940	0.912
8	0.808	0.941	0.925	0.941	0.914	0.670	0.940	0.884	0.940	0.912	0.755	0.940	0.884	0.940	0.912
9	0.819	0.942	0.945	0.942	0.914	0.671	0.940	0.884	0.940	0.912	0.757	0.941	0.944	0.941	0.912
10	0.838	0.942	0.935	0.942	0.915	0.675	0.940	0.884	0.940	0.912	0.767	0.941	0.944	0.941	0.912
11	0.897	0.943	0.934	0.943	0.918	0.677	0.940	0.884	0.940	0.912	0.783	0.940	0.884	0.940	0.912
12	0.905	0.943	0.939	0.943	0.920	0.679	0.940	0.884	0.940	0.912	0.788	0.941	0.945	0.941	0.914
13	0.914	0.943	0.939	0.943	0.920	0.680	0.940	0.884	0.940	0.912	0.791	0.941	0.945	0.941	0.914

*Input variables, 1: economic activities, 2: 1 + chronic disease, 3: 2 + age, 4: 3 + marital status, 5: 4 + number of household members, 6: 5 + private health insurance, 7: 6 + gender, 8: 7 + education level, 9: 8 + subsidies, 10: 9 + depression, 11: 10 + perceived health awareness, 12: 11 + disability, 13: 12 + unmet medical needs

Table 8 Sensitivity analysis results (threshold: 30%) (continued)

No	Decision tree				AdaBoost				Neural network						
	AUROC	Accuracy	Precision	Recall	F1	AUROC	Accuracy	Precision	Recall	F1	AUROC	Accuracy	Precision	Recall	F1
*															
1	0.611	0.940	0.884	0.940	0.912	0.611	0.940	0.884	0.940	0.912	0.612	0.940	0.884	0.940	0.912
2	0.612	0.940	0.884	0.940	0.912	0.618	0.940	0.884	0.940	0.912	0.618	0.940	0.884	0.940	0.912
3	0.652	0.940	0.884	0.940	0.912	0.655	0.940	0.884	0.940	0.912	0.654	0.940	0.884	0.940	0.912
4	0.663	0.940	0.884	0.940	0.912	0.668	0.940	0.884	0.940	0.912	0.664	0.940	0.884	0.940	0.912
5	0.688	0.940	0.884	0.940	0.912	0.699	0.940	0.884	0.940	0.912	0.685	0.940	0.884	0.940	0.912
6	0.708	0.940	0.884	0.940	0.912	0.730	0.940	0.884	0.940	0.912	0.712	0.940	0.884	0.940	0.912
7	0.722	0.941	0.944	0.941	0.913	0.756	0.941	0.925	0.941	0.914	0.727	0.940	0.884	0.940	0.912
8	0.756	0.940	0.884	0.940	0.912	0.820	0.942	0.927	0.942	0.917	0.744	0.940	0.884	0.940	0.912
9	0.760	0.940	0.884	0.940	0.912	0.838	0.943	0.933	0.943	0.920	0.769	0.940	0.884	0.940	0.912
10	0.773	0.941	0.925	0.941	0.913	0.858	0.946	0.938	0.946	0.927	0.785	0.940	0.884	0.940	0.912
11	0.800	0.942	0.935	0.942	0.917	0.924	0.953	0.948	0.953	0.941	0.823	0.941	0.944	0.941	0.913
12	0.806	0.942	0.935	0.942	0.917	0.939	0.955	0.951	0.955	0.944	0.829	0.942	0.945	0.942	0.914
13	0.816	0.943	0.932	0.943	0.920	0.951	0.959	0.955	0.959	0.952	0.851	0.943	0.946	0.943	0.917

*Input variables, 1: economic activities, 2: 1 + chronic disease, 3: 2 + age, 4: 3 + marital status, 5: 4 + number of household members, 6: 5 + private health insurance, 7: 6 + gender, 8: 7 + education level, 9: 8 + subsidies, 10: 9 + depression, 11: 10 + perceived health awareness, 12: 11 + disability, 13: 12 + unmet medical needs

Table 9 Sensitivity analysis results (threshold: 40%)

No	Random forest					Ridge regression					Gradient boosting				
	AUROC	Accuracy	Precision	Recall	F1	AUROC	Accuracy	Precision	Recall	F1	AUROC	Accuracy	Precision	Recall	F1
*															
1	0.629	0.966	0.934	0.966	0.950	0.629	0.966	0.934	0.966	0.950	0.629	0.966	0.934	0.966	0.950
2	0.657	0.966	0.934	0.966	0.950	0.653	0.966	0.934	0.966	0.950	0.657	0.966	0.934	0.966	0.950
3	0.668	0.966	0.934	0.966	0.950	0.665	0.966	0.934	0.966	0.950	0.668	0.966	0.934	0.966	0.950
4	0.675	0.966	0.934	0.966	0.950	0.670	0.966	0.934	0.966	0.950	0.678	0.966	0.934	0.966	0.950
5	0.718	0.966	0.934	0.966	0.950	0.667	0.966	0.934	0.966	0.950	0.721	0.966	0.934	0.966	0.950
6	0.758	0.966	0.934	0.966	0.950	0.670	0.966	0.934	0.966	0.950	0.747	0.966	0.934	0.966	0.950
7	0.786	0.967	0.968	0.967	0.951	0.672	0.966	0.934	0.966	0.950	0.754	0.966	0.934	0.966	0.950
8	0.854	0.967	0.968	0.967	0.951	0.671	0.966	0.934	0.966	0.950	0.793	0.966	0.934	0.966	0.950
9	0.869	0.967	0.968	0.967	0.951	0.675	0.966	0.934	0.966	0.950	0.799	0.966	0.934	0.966	0.950
10	0.884	0.967	0.969	0.967	0.953	0.682	0.966	0.934	0.966	0.950	0.807	0.966	0.934	0.966	0.950
11	0.926	0.968	0.969	0.968	0.954	0.683	0.966	0.934	0.966	0.950	0.822	0.966	0.934	0.966	0.950
12	0.937	0.968	0.969	0.968	0.954	0.683	0.966	0.934	0.966	0.950	0.830	0.97	0.968	0.967	0.951
13	0.943	0.968	0.964	0.968	0.954	0.681	0.966	0.934	0.966	0.950	0.832	0.967	0.968	0.967	0.951

*Input variables, 1: economic activities, 2: 1 + chronic disease, 3: 2 + age, 4: 3 + marital status, 5: 4 + number of household members, 6: 5 + private health insurance, 7: 6 + gender, 8: 7 + education level, 9: 8 + subsidies, 10: 9 + depression, 11: 10 + perceived health awareness, 12: 11 + disability, 13: 12 + unmet medical needs

Table 10 Sensitivity analysis results (threshold: 40%) (continued)

No	Decision tree					AdaBoost					Neural network				
	AUROC	Accuracy	Precision	Recall	F1	AUROC	Accuracy	Precision	Recall	F1	AUROC	Accuracy	Precision	Recall	F1
*															
1	0.500	0.966	0.934	0.966	0.950	0.629	0.966	0.934	0.966	0.950	0.629	0.966	0.934	0.966	0.950
2	0.500	0.966	0.934	0.966	0.950	0.657	0.966	0.934	0.966	0.950	0.657	0.966	0.934	0.966	0.950
3	0.500	0.966	0.934	0.966	0.950	0.668	0.966	0.934	0.966	0.950	0.667	0.966	0.934	0.966	0.950
4	0.500	0.966	0.934	0.966	0.950	0.678	0.966	0.934	0.966	0.950	0.672	0.966	0.934	0.966	0.950
5	0.500	0.966	0.934	0.966	0.950	0.724	0.966	0.934	0.966	0.950	0.698	0.966	0.934	0.966	0.950
6	0.500	0.966	0.934	0.966	0.950	0.762	0.966	0.934	0.966	0.950	0.729	0.966	0.934	0.966	0.950
7	0.500	0.966	0.934	0.966	0.950	0.791	0.967	0.960	0.967	0.952	0.744	0.966	0.934	0.966	0.950
8	0.500	0.966	0.934	0.966	0.950	0.867	0.967	0.959	0.967	0.953	0.809	0.966	0.934	0.966	0.950
9	0.500	0.966	0.934	0.966	0.950	0.883	0.969	0.970	0.969	0.955	0.811	0.966	0.934	0.966	0.950
10	0.500	0.966	0.934	0.966	0.950	0.904	0.970	0.970	0.970	0.960	0.825	0.967	0.968	0.967	0.951
11	0.500	0.966	0.934	0.966	0.950	0.951	0.974	0.972	0.974	0.967	0.863	0.967	0.968	0.967	0.951
12	0.961	0.975	0.972	0.975	0.971	0.961	0.975	0.972	0.978	0.971	0.865	0.967	0.968	0.967	0.951
13	0.500	0.966	0.934	0.966	0.950	0.970	0.978	0.976	0.978	0.974	0.899	0.967	0.969	0.967	0.953

* Input variables, 1: economic activities, 2: 1 + chronic disease, 3: 2 + age, 4: 3 + marital status, 5: 4 + number of household members, 6: 5 + private health insurance, 7: 6 + gender, 8: 7 + education level, 9: 8 + subsidies, 10: 9 + depression, 11: 10 + perceived health awareness, 12: 11 + disability, 13: 12 + unmet medical needs

Conclusion

The purpose of the study was to find an optimal classification model regarding the occurrence of CHE based on the factors and causes identified in previous studies. To the best of the authors' knowledge, this study may be the first to utilize machine learning techniques for the classification model for the occurrence of CHEs in low-income households in South Korea. As a result, the AdaBoost had the highest classifiable power, and the top factors were found to be that affected the occurrence of CHEs were identified as economic activity, chronic disease, and age. Notably, the influencing factors on the occurrence of CHEs in low-income households may vary according to gender and age group, suggesting the need for further research.

Abbreviations

CHE Catastrophic Health Expenditure
KHPS Korea Health Panel Survey
OECD Organization for Economic Cooperation and Development
OOP Out-of-pocket
WHO World Health Organization

Acknowledgements

Not applicable.

Authors' contributions

KMK had full access to all the data used in this study and takes responsibility for the integrity of the data and accuracy of the data analysis. Study concept and design: KMK, Data acquisition: SMJ, JWK, KMK, Statistical analysis: KMK, JWK, SMJ, Interpretation of the results: KMK, SMJ, Manuscript drafting: KMK, JWK. All the authors have read and approved the final version of the manuscript.

Funding

Not applicable.

Data availability

The data will be made available for special purposes only upon request to the corresponding authors.

Declarations

Ethics approval and consent to participate

This study was approved by the Korea University Institutional Review Board (IRB No. 2023-0043). The IRB of Korea University waived informed consent since this study was retrospective and blinding of the personal information in the data was performed. This data is publicly accessible and written informed consent is obtained from all the participants before participating in the survey. Respondents' information was completely anonymized for use for research purposes and unidentified prior to analysis. The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2000.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 7 February 2025 / Accepted: 1 July 2025

Published online: 07 August 2025

References

- Böhm K, Schmid A, Götze R, Landwehr C, Rothgang H. Five types of OECD healthcare systems: empirical results of a deductive classification. *Health Policy*. 2013;113(3):258–69. <https://doi.org/10.1016/j.healthpol.2013.09.003>.
- Shin SM. Household catastrophic health expenditure related to pain in Korea. *Korean J Pain*. 2023;36(3):347–57. <https://doi.org/10.3344/kjp.23041>.
- Koo JH, Jung HW. Which indicator should be used? A comparison between the incidence and intensity of catastrophic health expenditure: using difference-in-difference analysis. *Health Econ Rev*. 2022;12(1):58. <https://doi.org/10.1186/s13561-022-00403-w>.
- Azzani M, Roslani AC, Su TT. Determinants of household catastrophic health expenditure: A systematic review. *Malaysian J Med Sciences: MJMS*. 2019;26(1):15–43. <https://doi.org/10.21315/mjms2019.26.1.3>.
- Lee HY, Oh J, Kawachi I. Changes in catastrophic health expenditures for major diseases after A 2013 health insurance expansion in South Korea. *Health Aff*. 2022;41(5):722–31. <https://doi.org/10.1377/hlthaff.2021.01320>.
- Aittomäki A, Martikainen P, Laaksonen M, Lahelma E, Rahkonen O. Household economic resources, labour-market advantage and health problems - a study on causal relationships using prospective register data. *Soc Sci Med*. 2012;75(7):1303–10. <https://doi.org/10.1016/j.socscimed.2012.05.015>.
- Choi JW, Shin JY, Cho KH, Nam JY, Kim JY, Lee SG. Medical security and catastrophic health expenditures among households containing persons with disabilities in Korea: a longitudinal population-based study. *Int J Equity Health*. 2016;15(1):119. <https://doi.org/10.1186/s12939-016-0406-9>.
- Sohn M, Che X, Park HJ. Unmet healthcare needs, catastrophic health expenditure, and health in South Korea's universal healthcare system: progression towards improving equity by NHI type and income level. *Healthc (Basel Switzerland)*. 2020;8(4):408. <https://doi.org/10.3390/healthcare8040408>.
- Organization for Economic Cooperation and Development. (2023). Death and death rate by cause. Retrieved from <https://stats.oecd.org/Index.aspx?DataSetCode=SHA>. (Accessed 8 February 2024).
- Zhang F, Jiang J, Yang M, Zou K, Chen D. Catastrophic health expenditure, incidence, trend and socioeconomic risk factors in China: A systematic review and meta-analysis. *Front Public Health*. 2023;10:997694. <https://doi.org/10.3389/fpubh.2022.997694>.
- Sayuti M, Sukeri S. Assessing progress towards sustainable development goal 3.8.2 and determinants of catastrophic health expenditures in Malaysia. *PLoS ONE*. 2022;17(2):e0264422. <https://doi.org/10.1371/journal.pone.0264422>.
- Mekonen AM, Gebregziabher MG, Teferra AS. The effect of community based health insurance on catastrophic health expenditure in Northeast Ethiopia: A cross sectional study. *PLoS ONE*. 2018;13(10):e0205972. <https://doi.org/10.1371/journal.pone.0205972>.
- Bodhisane S, Pongpanich S. The influence of the National health insurance scheme of the Lao people's Democratic Republic on healthcare access and catastrophic health expenditures for patients with chronic renal disease, and the possibility of integrating organ transplantation into the health financing system. *Health Res Policy Syst*. 2022;20(1):71. <https://doi.org/10.1186/s12961-022-00869-4>.
- Mulaga AN, Kamndaya MS, Masangwi SJ. Decomposing socio-economic inequality in catastrophic out-of-pocket health expenditures in Malawi. *PLOS Global Public Health*. 2022;2(2):e0000182. <https://doi.org/10.1371/journal.pgph.0000182>.
- El-Hasnony IM, Elzeki OM, Alshehri A, Salem H. Multi-Label active learning-Based machine learning model for heart disease prediction. *Sensors*. 2022;22(3):1184. <https://doi.org/10.3390/s22031184>.
- Flores G, Krishnakumar J, O'Donnell O, van Doorslaer E. Coping with health-care costs: implications for the measurement of catastrophic expenditures and poverty. *Health Econ*. 2008;17(12):1393–412. <https://doi.org/10.1002/hecl.1338>.
- Wagstaff A, van Doorslaer E. Catastrophe and impoverishment in paying for health care: with applications to Vietnam 1993–1998. *Health Econ*. 2003;12(11):921–34. <https://doi.org/10.1002/hecl.776>.
- Doshmangir L, Hasanpoor E, Abou Jaoude GJ, Eshtiaq B, Haghighat-Bidgoli H. Incidence of catastrophic health expenditure and its determinants in Cancer patients: A systematic review and Meta-analysis. *Appl Health Econ Health Policy*. 2021;19(6):839–55. <https://doi.org/10.1007/s40258-021-00672-2>.
- Mutyambizi C, Pavlova M, Hongoro C, Booyesen F, Groot W. Incidence, socio-economic inequalities and determinants of catastrophic health expenditure and impoverishment for diabetes care in South Africa: a study at two public hospitals in Tshwane. *Int J Equity Health*. 2019;18(1):73. <https://doi.org/10.1186/s12939-019-0977-3>.

20. Dhanekar A, Kumari R, Bahurupi YA. Out-of-Pocket, catastrophic health expenditure and distress financing on Non-Communicable diseases in India: A systematic review with Meta-Analysis. *Asian Pac J Cancer Prevention: APJCP*. 2021;22(3):671–80. <https://doi.org/10.31557/APJCP.2021.22.3.671>.
21. Koch KJ, Pedraza C, Schmid A. Out-of-pocket expenditure and financial protection in the Chilean health care system-A systematic review. *Health Policy*. 2017;121(5):481–94. <https://doi.org/10.1016/j.healthpol.2017.02.013>.
22. Njagi P, Arsenijevic J, Groot W. Understanding variations in catastrophic health expenditure, its underlying determinants and impoverishment in sub-Saharan African countries: a scoping review. *Syst Rev*. 2018;11(1):136. <https://doi.org/10.1186/s13643-018-0799-1>.
23. Goryakin Y, Suhrcke M. The prevalence and determinants of catastrophic health expenditures attributable to non-communicable diseases in low and middle-income countries: a methodological commentary. *Int J Equity Health*. 2014;7(1):107. <https://doi.org/10.1186/s12939-014-0107-1>.
24. Sriram S, Albadrani M. A study of catastrophic health expenditures in India - evidence from nationally representative survey data: 2014-2018. *F1000Res*. 2022;3(11):141.
25. Frohlich KL, Potvin L. Transcending the known in public health practice: the inequality paradox: the population approach and vulnerable populations. *Am J Public Health*. 2008;98(2):216–21. <https://doi.org/10.2105/AJPH.2007.114777>.
26. Boscart V, Crutchlow LE, Taucar S, Johnson L, Heyer K, Davey M, Costa M, A. P., Heckman G. Chronic disease management models in nursing homes: a scoping review. *BMJ Open*. 2020;10(2):e032316. <https://doi.org/10.1136/bmjopen-2019-032316>.
27. Kim KM, Kim JH, Rhee HS, Youn BY. Development of a prediction model for the depression level of the elderly in low-income households: using decision trees, logistic regression, neural networks, and random forest. *Sci Rep*. 2023;13(1):11473. <https://doi.org/10.1038/s41598-023-38742-1>.
28. Diz J, Marreiros G, Freitas A. Applying data mining techniques to improve breast Cancer diagnosis. *J Med Syst*. 2016;40(9):203. <https://doi.org/10.1007/s10916-016-0561-y>.
29. Rajan MP. An efficient ridge regression algorithm with parameter Estimation for data analysis in machine learning. *SN Comput Sci*. 2022;3(2):171.
30. Wallace ML, Mentch L, Wheeler BJ, Tapia AL, Richards M, Zhou S, Yi L, Redline S, Buysse DJ. Use and misuse of random forest variable importance metrics in medicine: demonstrations through incident stroke prediction. *BMC Med Res Methodol*. 2023;23(1):144. <https://doi.org/10.1186/s12874-023-01965-x>.
31. Li W, Huang G, Tang N, Lu P, Jiang L, Lv J, Qin Y, Lin Y, Xu F, Lei D. Effects of heavy metal exposure on hypertension: A machine learning modeling approach. *Chemosphere*. 2023;337:139435. <https://doi.org/10.1016/j.chemosphere.2023.139435>.
32. Xiahou X, Harada Y. Customer churn prediction using adaboost classifier and BP neural network techniques in the E-commerce industry. *Am J Industrial Bus Manage*. 2022;12(3):277–93.
33. National Health and Nutrition Study. (2023). Incidence of catastrophic health expenditure. Retrieved from https://www.khp.re.kr:444/web/research/board/surveyresult_view.do?&bbsid=60&seq=14. (Accessed 18 February 2024).
34. Jafor MA, Wadud MAH, Nur K, Rahman MM. Employee promotion prediction using improved adaboost machine learning approach. *AJSE*. 2023;22(3):2520–4890.
35. Mahesh TR, Kumar D, Vinoth Kumar V, Asghar V, Geman J, Arulkumaran O, Arun G, N. AdaBoost ensemble methods using K-fold cross validation for survivability with the early detection of heart disease. *Comput Intell Neurosci*. 2022;4(18):9005278. <https://doi.org/10.1155/2022/9005278>.
36. Krishnan R, Rajpurkar P, Topol EJ. Self-supervised learning in medicine and healthcare. *Nat Biomedical Eng*. 2022;6(12):1346–52. <https://doi.org/10.1038/s41551-022-00914-1>.
37. Aryankhesal A, Etemadi M, Mohseni M, Azami-Aghdash S, Nakhaei M. Catastrophic health expenditure in Iran: A review Article. *Iran J Public Health*. 2018;47(2):166–77.
38. Kang EN. The multidimensional changes of senior social activities and job program's participants: focused on civic service activities and market-type working group. *Korean J Gerontological Social Welf*. 2018;73(3):29–53.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.