Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Predicting Chronic Kidney Disease Risk Factors Using Machine Learning: Using the 2021 Korea National Health and Nutrition Examination Survey

Junga Kim
Kyungpook National University Hospital
Sung Hee Lee
leesh@knu.ac.kr

Kyungpook National University

Research Article

Keywords: Machine learning, Age, Kidney diseases, Prediction algorithm, Risk factors, Korea

Posted Date: March 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-6364577/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background: The prevalence of chronic kidney disease (CKD) in Korea increases annually. With the rapidly aging population in Korea, the number of patients with CKD is expected to increase further. CKD imposes a significant burden on both individuals and the country. However, due to the lack of awareness of CKD, most patients are diagnosed in the end stage of CKD. Therefore, this study aims to develop a machine learning model for CKD to identify at-risk patients, slow disease progression, and prevent complications.

Methods: Based on the Rainbow model, 61 variables were considered explanatory variables. Among the adult and elderly, 197 (5.1%) of 3,868 participants and 135 (11.1%) of 1,216 participants were classified as having CKD, respectively. Six machine learning methods were used to explore risk factors for CKD and identify the model with the highest performance power. Logistic regression analysis was used to confirm the importance of key variables in each selected machine-learning model.

Results: In adults, the boosting method demonstrated the highest predictive power for CKD (accuracy, 0.974; precision, 0.975; recall, 0.974; F1 score, 0.968; AUC, 0.886). Analysis of the elderly population revealed the Naïve Bayes model as the most effective for predicting CKD (accuracy, 0.905; precision, 0.922; recall, 0.905; F1 score, 0.912; AUC, 0.744). Logistic regression analysis reaffirmed the risk factors identified. In adults, these included urine protein, age, private insurance, hypertension, diabetes, residential area, and anemia. In the elderly, they included urine protein, anemia, age, diabetes, private insurance, moderate to high physical activity levels, household composition, sex, number of elderly leisure welfare facilities per 1,000 people, and subjective health perception.

Conclusion: The machine learning risk model for CKD developed in this study serves as a foundation for the formulation of nursing plans, the establishment of early warning systems through prediction, and the development of nursing guidelines in nursing practice.

Background

Chronic kidney disease (CKD) is a progressive condition characterized by an irreversible decline in kidney function, impairing its ability to function normally [1]. The Korea Disease Control and Prevention Agency reports a 6.3% prevalence of CKD in Korea in 2021 [2]. According to the International Society of Nephrology, Korea ranks 6th among 59 countries, including the United States, and 4th among Asian countries for CKD prevalence [3]. With an average annual growth rate of 18.8% per million people, Korea has the second-highest CKD growth rate globally after Thailand [4]. The number of patients receiving treatment for CKD in Korea more than doubled from 137,003 in 2012 to 296,397 in 2022 over the past 10 years [5]. In 2024, the elderly population (aged 65 years or older) in Korea accounted for approximately 19.2%, marking its transition into a super-aged society [7]. Kidney function declines with age, and CKD prevalence is expected to increase further as the population ages. In Korea, 17.5% of patients with CKD are aged 65 years or older, which is approximately three times greater than the overall population [6].

Early-stage CKD can be easily detected through screening tests such as blood and urine tests and is manageable with regular checkups [7]. However, early-stage CKD often goes undetected due to the absence of symptoms [8]. Advanced-stage CKD progresses rapidly, causing severe renal dysfunction and requiring need renal replacement therapy [1]. Patients with end-stage renal disease undergoing renal replacement therapy often experience persistent depression and poor quality of life [9]. Additionally, complications increase the prevalence and mortality of cardiovascular disease [9]. CKD incurred the highest per capita medical expenses in Korea, imposing a significant economic burden on both individuals and the country [10]. Therefore, preventing progression to end-stage CKD, which is difficult to treat, is important [11]. Predicting CKD enables nurses to identify risk factors for patients at risk of CKD, strengthen personalized care, provide individual education, and guide timely hospital visits.

The health of an individual is affected by the environment. Dahlgren and Whitehead's rainbow model is a theory that explains the determinants of health based on social-ecological theory (Figure 1) [12]. The multicausal model states that the health of an individual is a complex interaction of biological characteristics as well as internal and external factors across four layers [12]. This study provides a theoretical framework for understanding how factors across various layers affect health outcomes, especially health problems. Therefore, this model guides the development of targeted interventions or policies [13]. In a previous study, when the local government applied the rainbow model to health policies and environmental improvements, the health level of residents improved [14]. In addition, research highlights the effects of social health determinants based on the rainbow model on hypertension [15]. Thus, the rainbow model is a theory that effectively explains diseases with various risk factors, such as CKD. In this study, the risk factors for CKD based on the rainbow model were examined.

The Korea National Health and Nutrition Examination Survey (KNHANES) is a national statistic of Korea conducted annually nationwide that systematically collects data on health status, health behaviors, and environmental factors influencing health [16]. It includes screening surveys, nutrition surveys, and health questionnaires related to CKD, making it a valuable data source for CKD research that considers the characteristics of the Korean population. The KNHANES data includes health determinants outlined in the rainbow model [16]. In addition, integrating the community health-related database with the KNHANES expands research by identifying factors influencing community health and disparities.

Machine learning is an effective method for analyzing and predicting diseases with complex and diverse causes [17]. Existing statistical methods struggle to analyze diseases with complex and diverse causes, determine causal relationships and address sample imbalances [18]. Trends have been identified and predicted using machine learning based on the National Health and Nutrition Survey, health insurance records, and medical panel surveys conducted in Korea, producing highly accurate prediction models [19]. Machine learning is a recognized and essential research method for analyzing such surveys [19,20]. However, machine learning studies on CKD have focused primarily on physiological indicators, with limited research examining and predicting individuals and their surrounding health determinants [21-23]. When analyzing CKD using the rainbow model, machine learning classification algorithms provide an appropriate approach [24]. The rainbow theory incorporates diverse variables by comprehensively

addressing numerous factors influencing disease causation [25]. Machine learning effectively explores and models the relationships between variables in such multidimensional data [12].

CKD is difficult to detect or treat due to lack of awareness of the disease. Developing a CKD classification model is a meaningful study to address this lack of awareness and information. Evidence-based nursing practice relies on applying the latest research methods to verify theories and generate reliable data for patient care [26]. This study used machine learning to enable evidence-based nursing practices by scientifically classifying and analyzing CKD data. The secured machine learning model serves as an objective and valid indicator for effectively conveying and explaining the condition of a patient. It functions as an early warning system, enabling the early identification of at-risk patients to slow the progression of CKD or prevent complications. Nurses must scientifically understand the conditions of patients and provide effective interventions. This enables nurses to take a leading role in patient management, deliver targeted interventions, and collaborate with multidisciplinary experts to develop CKD management plans. It also serves as foundational data for developing national and local CKD-related health policies.

Therefore, this study aims to develop a CKD classification model using a machine learning classification algorithm based on a rainbow model using KNHANES data, focusing on adult and elderly populations to identify risk factors. It would be possible to detect asymptomatic CKD patients and expect appropriate intervention before CKD symptoms appear.

Methods and Materials

Methodology

This study is a secondary data analysis study that integrated data from the 2021 KNHANES with that of the 2021 Community Health Database to develop and evaluate a CKD classification model. This study aims to compare the performance of various CKD classification models developed using machine learning algorithms, identify the most effective classification model, and determine risk factors associated with CKD. For machine learning classification algorithm, Jeffreys's Amazing Statistics Program (JASP) 0.19.0.0 and for logistic regression analysis, Statistical Package for the Social Sciences (SPSS) 29.0.2.0 was used.

Research population

The sampling frame of KNHANES is based on the most recent population and housing census data available at the time of sample design to ensure a representative sample of the target population—the Korean people aged ≥ 1 year. A two-stage stratified cluster sampling method was employed, selecting survey areas (1st) and households (2nd) as the sampling units. This study focused on CKD in Korean adults and elderly individuals, including participants aged ≥ 19 years and elderly individuals aged ≥ 65 years.

Data collection

The KNHANES is a government-led statistical survey conducted under relevant laws. The Korea Disease Control and Prevention Agency discloses KNHANES data annually, making it publicly available for academic research. The data were anonymized in accordance with the Personal Information Protection Act and the Statistics Act to prevent individual identification. In this study, data were downloaded from the Korea Disease Control and Prevention Agency (KNHANES) website after completing the statistical data compliance pledge and used per the National Health and Nutrition Survey Usage Guidelines. The community health-related database was publicly available on the community health survey website, allowing linkage of health determinants to community health levels and disparities. The community health-related factor database was downloaded from the website and used in this study.

This study was approved by the Kyungpook National University Bioethics Committee for exemption from the IRB (2024-0278) after reviewing the overall research, including the purpose and methods.

Research variables and definitions.

Chronic kidney disease

CKD was classified as “yes” if the glomerular filtration rate was < 60 ml/min/1.73 m² or the urine albumin-to-creatinine ratio was ≥ 30 mg/g, based on the CKD prevalence criteria of the Korea Disease Control and Prevention Agency. Otherwise, it was classified as “no.”

The dependent variable, CKD, was calculated using serum creatinine (Scr), urine albumin, urine creatinine, sex, weight, and age data from the KNHANES. The glomerular filtration rate was determined via the CKD-EPI eGFR formula (2021) provided by the Korean Society of Nephrology. The CKD-EPI eGFR (mL/min/1.73m^2) was calculated using the formula = 141*min(Scr/κ,1)$^{\alpha}$*max(Scr/κ,1)$^{-1.209}$*0.993^{Age}*1.018[if female]*1.159[if black], where K = 0.7 (females) or 0.9 (males), α = -0.241 (females) or -0.302 (males), min = indicates the minimum of Scr/K or 1, max = indicates the maximum of Scr/K or 1. The urine albumin-to-creatinine ratio was estimated via the expected daily creatinine excretion (g/day). For men, it was calculated as (28-(age/6))*Bwt/1000(g/day), and for women, it was calculated as (22-(age/9))*Bwt/1000(g/day).

Explanatory variables

Based on Dahlgren and Whitehead’s Rainbow Model (1991) and a literature review of factors influencing CKD identified in previous studies, 61 explanatory variables were selected. These variables were synthesized from previous studies (Table 1).

Age, sex, and constitution factors from the center of the Rainbow Model were selected, including age, sex, hypertension, diabetes, obesity, anemia, urine protein, and urine occult blood. Hypertension, diabetes, obesity, and anemia were derived variables defined based on prevalence data from physical measurements and examination results in the National Health Survey.

Individual lifestyle factors from the first layer of the model were assessed using health survey data. Smoking and drinking were evaluated based on current data, while anxiety, stress, activity limitations, subjective health status, aerobic physical activity, occupation, and nutritional surveys (food and water intake) were evaluated through self-reports.

Social network variables from the second layer of the model were derived from family-related data in the KNHANES. These included marital status, household size, and generational composition.

Variables corresponding to the third layer, living and working conditions, included dietary composition, nutritional labeling awareness, economic activity status, recipient of basic living assistance, household and individual income, health insurance, personal insurance, education, medical examinations, annual unmet medical needs for clinics or hospitals, unmet need for necessary medical services, house ownership, and housing type, as outlined in the model.

In the fourth layer, general socioeconomic, cultural, and environmental conditions, the corresponding KNHANES variables included residence area, city/county, and town/village. Since residential areas alone did not fully capture general socioeconomic, cultural, and environmental conditions, a community health-related database was integrated for a more comprehensive analysis. A previous study on the determinants of health-related quality of life showed machine learning, medical personnel, medical facilities, and specific regional indicators as key factors influencing health status at the regional level [124]. In addition, regional economic and development factors—such as the population, crime rate, and number of medical cases—should be thoroughly examined when formulating and implementing health policies for a specific region [124]. Consequently, variables were selected from the community health survey database.

Table 1. Variables input into machine learning.

Category	Variable	DB
Age, sex & Constitutional factors (8)	Age, Sex, Hypertension (create), Diabetes Mellitus (create), Obesity (create), Anemia (create), Urine Protein (checkup), Urine blood (checkup)	KNHANES
Individual lifestyle factors (13)	Smoking, Drinking, Stress, Anxiety, Subjective health evaluation (Quality of Life), Activity limitation, Aerobic physical activity, Occupation, Water intake (cup), Food intake (water, sodium, potassium, protein)	
Social Network (3)	Marriage, Number of household members, Family household	
Living and working conditions (14)	Dietary composition, Nutritional labeling awareness, Economic activity state, Recipient of basic living, Income (house, individual), Health insurance, Personal insurance, Education, Medical examination, Annual unmet need for medical clinics or hospitals, Unmet need for necessary medical services, House ownership, Housing type	
General socioeconomic cultural and environmental conditions (2+21)	Residence area, District-town & village	
	Moderate or higher level of physical activity practice rate, Healthy lifestyle practice rate, Number of sports facilities, Number of parks, Traffic culture index, Walking practice rate, Elderly population ratio, Financial Independence, Financial autonomy, Number of cultural infrastructure facilities per 100,000 people, Number of elderly leisure and welfare facilities per 1000 elderly people, Number of residents per rescue worker, Number of doctors and nurses working in medical institutions per 1,000 people, Number of medical institution beds per 1,000 people, number of people in out-of-town clinics (total), Treatment rate for those diagnosed with hypertension, treatment rate for those diagnosed with diabetes, Annual Diabetic Kidney Disease Complication Screening Rate, Annual diabetic kidney disease complication screening rate, Annual unmet medical needs rate	Community Health Database

KNHANES, Korea National Health and Nutritional Survey

Data preprocessing

The 2021 KNHANES included 7,090 participants, of whom 5,953 were aged ≥ 19 years. Exclusions comprised 976 individuals who did not participate in the nutritional survey, 269 who did not participate in the physical examination survey, and 33 people who did not participate in the health survey. The final sample comprised 4,674 adults aged ≥ 19 years and 1,838 elderly people aged ≥ 65 years.

The KNHANES comprised three components: a physical examination, health (health behavior and health interview surveys), and nutrition surveys [16]. In this study, missing values (weight, age, urine creatinine, urine protein, and creatinine) used to derive the dependent variable, CKD, could not be replaced with specific values. In addition, assigning specific values to examination items related to generated

variables, such as diabetes, hypertension, anemia, and obesity (risk factors for CKD), was challenging. Therefore, a simple elimination method was applied.

Data preprocessing was performed on 4,674 adults aged ≥ 19 years, excluding non-participators, by sequentially removing missing values from the nutrition, physical examination, and health survey. In this study, three participants with missing nutrition survey data, 531 with missing physical examination data, and 272 with missing health survey data were excluded. The appendix presents the homogeneity analysis results for the population and analysis group, along with the characteristics of participants excluded due to missing data. Finally, the analysis included 3,868 participants aged ≥ 19 years, of whom 1,216 were aged ≥ 65 years. The community health-related database was combined using the residential area information of the participants (Figure 2).

Data analysis

In this study, data analysis was conducted on two groups: adults aged ≥ 19 years and elderly individuals aged ≥ 65 years. A CKD classification model was developed, the final model was selected, and key risk variables were analyzed using logistic regression to determine the risk levels (Figure 3).

The community health-related database was integrated using the 2021 KNHANES. After data collection, SPSS 29.0 was used for data cleansing to confirm the participants before merging the community health-related database. Sixty-one explanatory variables were selected based on Dahlgren and Whitehead's Rainbow Model to identify CKD risk factors.

Subsequently, a machine learning classification algorithm was performed using the JASP 19.0 program. Among the machine learning classification algorithms provided by the program, six classification algorithms were applied, including Naïve Bayes, K-nearest neighbors, support vector machines, decision trees, random forests, and boosting. The linear discriminant and neural network models were excluded, as they require continuous explanatory variables.

The models were then evaluated to identify the best-performing predictive model using five indicators: accuracy, precision, recall, F1 score, and area under the curve (AUC). An AUC of 0.6 or higher was generally considered significant [27]. Additionally, previous research recommended examining various performance indicators alongside the receiver operating characteristic curve [28]. The final model was selected by comparing performance indicators among models with AUC values of ≥ 0.6 or higher.

Risk factors were identified based on the variable importance of the selected model. JASP provided two metrics for CKD risk factors: variable importance (Relative Influence) and average loss value (Mean dropout loss). Ensemble-based algorithms additionally reported both metrics, while the other algorithms provided only the mean dropout loss value. The mean dropout loss value of variables indicated how much each variable contributes to the prediction performance of the model.

To quantitatively interpret the risk factors, variables with high importance were selected for logistic regression analysis [29,30]. Given the 1-year duration of the study, cross-sectional weighting was applied

[31]. Logistic regression analysis was used to identify the magnitude of risk variables and their associations with CKD.

Results

Adult population (≥ 19 years)

Performance comparison of chronic kidney disease classification models developed in the adult population (≥ 19 years)

The performance of machine learning models was evaluated using five key metrics: accuracy, precision, recall, F1-score, and AUC. The results showed that the boosting, decision tree, Naïve Bayes, random forest, and support vector machine models achieved significant performance, while K-nearest neighbors had an AUC < 0.6 and were excluded. Table 2 shows the performance of each model based on the evaluation metrics. Since recall is crucial for minimizing Type II errors in disease prediction, the boosting model, which achieved the highest recall, was selected as the final model. Additionally, it outperformed the other models across the remaining.

Table 2. Performance comparison of CKD classification models (≥ 19 years)

Performance indicator	Model	Total	Rank
Accuracy	Naïve Bayes	0.821	5
	KNN	0.944	-
	SVM	0.972	2
	Decision Tree	0.962	3
	Random Forest	0.948	4
	Boosting	0.975	1
Precision	Naïve Bayes	0.919	5
	KNN	0.911	-
	SVM	0.968	2
	Decision Tree	0.954	3
	Random Forest	0.951	4
	Boosting	0.975	1
Recall	Naïve Bayes	0.821	5
	KNN	0.944	-
	SVM	0.972	2
	Decision Tree	0.962	3
	Random Forest	0.948	4
	Boosting	0.974	1
F1 Score	Naïve Bayes	0.862	5
	KNN	0.925	-
	SVM	0.967	2
	Decision Tree	0.955	3
	Random Forest	0.927	4
	Boosting	0.968	1
AUC	Naïve Bayes	0.654	4
	KNN	0.558	-
	SVM	0.681	3
	Decision Tree	0.622	5

	Random Forest	0.889	1
	Boosting	0.886	2

AUC, area under the curve; CKD, chronic kidney disease

Variable importance of the final selected boosting model in the adult population (≥ 19 years)

To identify CKD risk factors, the variable importance of the boosting model was analyzed (Table 3). The relative influence values ranked as follows: urinary protein (68.088), age (15.145), private insurance (8.807), hypertension (3.232), diabetes (2.055), residential area (1.489), and anemia (1.233).

In the boosting model, mean dropout loss did not directly influence the variable performance but it was used to evaluate model performance. Although these variables did not have a direct influence, they enhanced the predictive accuracy of the model. The mean dropout loss value for the remaining variables was 0.095.

Table 3. Variable Importance of Boosting (≥ 19 years)

Rank	Variable	R.I	M.D.L
1	Urine protein (Age, sex and constitutional)	68.088	0.186
2	Age (Age, sex and constitutional)	15.145	0.128
3	Personal insurance (Living and working conditions)	8.807	0.111
4	Hypertension (Age, sex and constitutional)	3.232	0.117
5	Diabetes Mellitus (Age, sex and constitutional)	2.005	0.1
6	Area of residence (General socioeconomic cultural and environmental conditions)	1.489	0.105
7	Anemia (Age, sex and constitutional)	1.233	0.097
8	(Age, sex and constitutional) Sex, Obesity, Urine blood	0	0.095
	(Individual lifestyle factors) Smoking, Drinking, Stress, Anxiety, Subjective health evaluation, Activity limitation, Aerobic physical activity, Occupation, Water intake (cup), water/Sodium/Potassium/Protein intake		
	(Social and community networks) Marriage, Number of household members, Family household		
	(Living and working conditions) Dietary composition, Nutritional labeling awareness, Economic activity state, Recipient of basic living, Health insurance, Personal insurance, Home income, Individual income, Education, Medical examination, Annual unmet need for necessary medical services, House ownership, Housing type		
	(General socioeconomic cultural and environmental conditions) District, town & village Moderate or higher level of physical activity practice rate, Healthy lifestyle practice rate, Number of sports facilities, Number of parks, Traffic culture index, Walking practice rate, Elderly population ratio, Financial Independence, Financial autonomy, Number of cultural infrastructure facilities per 100000 people, Number of elderly leisure and welfare facilities per 1000 elderly people, Number of residents per rescue worker, Number of doctors and nurses working in medical institutions per 1000 people, Number of medical institution beds per 1000 people, number of people in out-of-town clinics (total), Treatment rate for those diagnosed with hypertension, treatment rate for those diagnosed with diabetes, Annual Diabetic Kidney Disease Complication Screening Rate, Annual diabetic kidney disease complication screening rate, Annual unmet medical needs rate		

RI, relative influence; M.D. L, mean dropout loss

Logistic regression analysis to determine variable importance in the adult population (≥ 19 years)

Logistic regression analysis was conducted to determine the risk levels of the top seven variables that were significantly identified in the boosting model (Table 4).

Age positively influenced CKD prevalence, with each year of increase raising the risk by 1.07 times (OR = 1.07, CI: 1.07–1.07, p < .001). Hypertension, diabetes, anemia and urinary protein also exhibited significant positive associations with CKD. The presence of hypertension increased CKD risk by 1.83 times (OR = 1.83, CI: 1.82–1.83, p < .001), while diabetes increased the risk by 1.41 times (OR = 1.41, CI: 1.40–1.41, p < .001). Additionally, anemia was linked to a 3.33-fold increase in CKD risk (OR = 3.33, CI: 3.31–3.35, p < .001). The strongest association was observed with urinary protein, which raised the risk of CKD by 9.76 times (OR = 9.76, CI: 9.72–9.80, p < .001).

Compared to individuals the private insurance, those uncertain about their coverage exhibited a negative association. Compared to individuals with private insurance, those uncertain about their coverage had a 0.19-fold lower risk of CKD (OR = 0.81, CI: 0.78–0.83, p < .001). Compared to individuals with private insurance, those without coverage had a positive effect on CKD, a 2.37-fold greater risk of CKD (OR = 2.37, CI: 2.35–2.38, p < .001).

Risk factors for CKD were analyzed according to residential area. In Seoul, the capital of South Korea, Gyeonggi, Gyeongbuk, Gwangju, Daegu, Daejeon, Busan, Sejong, Ulsan, Jeju, and Chungbuk were associated with a lower risk of CKD. Compared to Seoul, the risk of CKD was 0.44 times greater in Gyeonggi (OR = 0.56, CI: 0.56–0.57, p < .001), 0.37 times greater in Gyeongbuk (OR = 0.63, CI: 0.62–0.64, p < .001), 0.31 times greater in Gwangju (OR = 0.69, CI: 0.68–0.70, p < .001), 0.13 times greater in Daegu (OR = 0.87, CI: 0.85–0.88, p < .001), 0.57 times greater in Daejeon (OR = 0.43, CI: 0.43–0.44, p < .001), 0.27 times greater in Busan (OR = 0.73, CI: 0.72–0.74, p < .001), and 0.09 times greater in Sejong (OR = .91, CI: 0.87–0.95, p < .001), Ulsan by 0.04 times (OR = 0.96, CI: 0.95–0.98, p < .001), Jeju, by 0.09 times (OR = 0.09, CI: 0.08–0.09, p < .001), and Chungbuk by 0.72 times (OR = 0.28, CI: 0.28–0.29, p < .001). In contrast, Seoul, Gangwon, Gyeongnam, Incheon, Jeonnam, Jeonbuk, and Chungnam were associated with an increased risk of CKD. Compared to Seoul, the risk of CKD was higher in Gangwon (OR = 1.47, CI: 1.45–1.49, p < .001), Gyeongnam (OR = 1.01, CI: 1.00–1.02, p < .001), Incheon (OR = 1.78, CI: 1.76–1.80, p < .001), Jeonnam (OR = 1.56, CI: 1.54–1.58, p < .001), 1 Jeonbuk (OR = 1.21, CI: 1.20–1.23, p < .001), and 1 Chungnam (OR = 1.44, CI: 1.43–1.46, p < .001).

Table 4. Risk factors for CKD (≥ 19 years, *N = 3,868)*

Rainbow model (category)	Variable	B	S.E.	Sig.	OR	95% CI
	Intercept	-10.31	.009	<. 001	.000	
Age, sex and constitutional factors	Age	.06	.000	<. 001	1.07	1.07–1.07
	Hypertension	.60	.002	<. 001	1.83	1.82–1.83
	Diabetes Mellitus	.34	.002	<. 001	1.41	1.40–1.41
	Anemia (yes)	1.20	.003	<. 001	3.33	3.31–3.35
	Urine protein	2.28	.002	<. 001	9.76	9.72–9.80
Living and working condition	Personal insurance (yes*)			<. 001		
	Personal insurance (unknown)	-.22	.015	<. 001	0.81	0.78–0.83
	Personal insurance (No)	.86	.003	<. 001	2.37	2.35–2.38
General socioeconomic cultural environmental condition	Aria of residence (Seoul; Capital*)			<. 001		
	Gangwon	.39	.006	<. 001	1.47	1.45–1.49
	Gyeonggi	-.58	.004	<. 001	0.56	0.56–0.57
	Kyungnam	.01	.006	.178	1.01	1.00–1.02
	Kyungbuk	-.46	.006	<. 001	0.63	0.62–0.64
	Gwangju	-.37	.008	<. 001	0.69	0.68–0.70
	Daegu	-.14	.008	<. 001	0.87	0.85–0.88
	Daejeon	-.83	.010	<. 001	0.43	0.43–0.44
	Pusan	-.32	.006	<. 001	0.73	0.72–0.74

Sejong	-.10	.020	<.001	0.91	0.87–0.95
Ulsan	-.04	.010	<.001	0.96	0.95–0.98
Incheon	.58	.005	<.001	1.78	1.76–1.80
Jeonnam	.45	.006	<.001	1.56	1.54–1.58
Jeonbuk	.19	.006	<.001	1.21	1.20–1.23
Jeju	-2.47	.015	<.001	.09	0.08–0.09
Chungnam	.37	.006	<.001	1.44	1.43–1.46
Chungbuk	-1.27	.009	<.001	.28	0.28–0.29

Note. Groups marked with an asterisk are the reference groups. CKD, chronic kidney disease

Elderly population (≥ 65 years)

Performance comparison of chronic kidney disease classification models developed for the elderly population (≥ 65 years)

The performance of machine learning models was evaluated using five metrics: accuracy, precision, recall, F1 score, and AUC. In individuals aged ≥ 65, the boosting, Naïve Bayes, and random forest models achieved AUC values of 0.6 or higher, indicating significant predictive performance. Table 5 shows the detailed performance metrics of each model based on these indicators.

Since recall is important for minimizing Type II errors in disease prediction, the Naïve Bayes model, which had the highest recall, was selected as the final model. It also outperformed the other models across the remaining evaluation metrics.

Table 5. Performance comparison of CKD classification models (≥ 65 years)

Performance indicator	Model	Total	Rank
Accuracy	Naïve Bayes	0.905	1
	KNN	0.909	-
	SVM	0.885	-
	Decision Tree	0.909	-
	Random Forest	0.901	2
	Boosting	0.893	3
Precision	Naïve Bayes	0.922	1
	KNN	0.918	-
	SVM	0.898	-
	Decision Tree	0.885	-
	Random Forest	0.911	2
	Boosting	0.863	3
Recall	Naïve Bayes	0.905	1
	KNN	0.909	-
	SVM	0.885	-
	Decision Tree	0.909	-
	Random Forest	0.901	2
	Boosting	0.893	3
F1 Score	Naïve Bayes	0.912	1
	KNN	0.870	-
	SVM	0.912	-
	Decision Tree	0.890	-
	Random Forest	0.858	3
	Boosting	0.867	2
AUC	Naïve Bayes	0.744	2
	KNN	0.527	-
	SVM	0.517	-
	Decision Tree	0.582	-

	Random Forest	0.788	1
	Boosting	0.708	3

AUC, area under the curve; CKD, chronic kidney disease

Variable importance of the final selected Naïve Bayes model in the elderly population (≥ 65 years)

To identify risk factors for CKD in the elderly population ≥ 65 years), the variable importance of the Naïve Bayes model was analyzed (Table 6).

The Age, sex, & Constitutional factors variables identified included urine protein (0.328), anemia (0.285), age (0.284), diabetes (0.282), sex (0.278), hypertension (0.273), obesity (0.271), and urine occult blood (0.268).

The individual lifestyle factors identified included subjective health perception (0.278), anxiety (0.277), protein intake (0.277), potassium intake (0.276), stress (0.270), water intake (0.269), alcohol consumption (0.264), smoking (0.264), sodium intake (0.263), and activity restrictions (0.26).

The social network factors identified included generational composition (0.278), household size (0.273), and marital status (0.269).

The living and working conditions identified included private insurance (0.280), home ownership (0.278), health checkups (0.272), health insurance coverage (0.272), dietary habits (0.268), basic livelihood security (0.267), employment status (0.267), housing type (0.266), education level (0.266), household income (0.262), personal income (0.262), and awareness of nutrition labels (0.260).

The general socioeconomic, cultural, and environmental conditions identified includes: the rate of moderate-to-high physical activity (0.279), number of elderly leisure and welfare facilities per 1,000 elderly individuals (0.278), transportation culture index (0.277), rate of healthy lifestyle practices (0.276), rate of walking practices (0.275), annual rate of unmet medical needs (0.275), treatment rate for individuals diagnosed with hypertension (0.274), city/province classification (0.274), annual screening rate for diabetes and kidney disease complications (0.273), treatment rate for individuals diagnosed with diabetes (0.271), occupation (0.271), rate of aerobic physical activity (0.271), number of hospital beds per 1,000 individuals (0.271), financial independence (0.270), proportion of the elderly population (0.270), number of emergency medical personnel per 1,000 individuals (0.268), number of doctors per 1,000 individuals in medical institutions (0.267), and township/village (0.267). Additionally, the following factors were confirmed: financial independence (0.267), unmet medical needs (0,265), number of cultural infrastructure facilities per 100,000 individuals (0.265), the annual rate of unmet hospital needs (0.265), number of nurses (0.264), number of beds in dialysis units (0.263), number of sports facilities (0.262), number of individuals in out-of-town clinics (0.260), and number of parks (0.260).

Table 6. Variable Importance of Naïve Bayes (≥ 65 years)

Rank	Variable	M.D.L
1	Urine protein (Age, sex & Constitutional factors)	0.328
2	Anemia (Age, sex & Constitutional factors)	0.285
3	Age (Age, sex & Constitutional factors)	0.284
4	Diabetes Mellitus (Age, sex & Constitutional factors)	0.282
5	Personal insurance (Living and working conditions)	0.28
6	Moderate or higher level of physical activity practice rate (Community)	0.279
7	Family household (Social Network)	0.278
	Sex (Age, sex & Constitutional factors)	
	Number of elderly leisure and welfare facilities per 1,000 elderly people (Community)	
	Subjective health evaluation (Individual lifestyle factors)	
8	Anxiety (Individual lifestyle factors)	0.277
	Traffic culture index (Community)	
	Protein intake (Individual lifestyle factors)	
9	House ownership (Living and working conditions)	0.276
	Potassium intake (Individual lifestyle factors)	
	Healthy lifestyle practice rate(Community)	
10	Walking practice rate (Community)	0.275
	Annual unmet medical needs rate (Community)	
11	Treatment rate for those diagnosed with hypertension (Community)	0.274
	Area of residence (General socioeconomic cultural environmental conditions)	
12	Annual Diabetic Kidney Disease Complication Screening Rate (Community)	0.273
	Number of households (Social Network)	
	Hypertension (Age, sex & Constitutional factors)	
13	Medical examination (Living and working conditions)	0.272
	Health insurance (Living and working conditions)	
14	treatment rate for those diagnosed with diabetes (Community)	0.271
	Occupation (Individual lifestyle factors)	

	Aerobic physical activity (Individual lifestyle factors)	
	Number of medical institution beds per 1000 people (Community)	
	Obesity (Age, sex & Constitutional factors)	
15	Financial autonomy (Community)	0.27
	Stress (Individual lifestyle factors)	
	Elderly population ratio (Community)	
16	Marriage (Social Network)	0.269
	Water intake (Individual lifestyle factors)	
17	Number of residents per rescue worker (Community)	0.268
	Dietary composition (Living and working conditions)	
	Urine blood (Age, sex & Constitutional factors)	
18	Recipient of basic living (Living and working conditions)	0.267
	Number of doctors working in medical institutions per 1000 people (Community)	
	District, town & village (General socioeconomic cultural environmental conditions)	
	Financial Independence (Community)	
	Economic activity (Living and working conditions)	
19	Housing type (Living and working conditions)	0.266
	Education (Living and working conditions)	
20	Unmet need for necessary medical services (Living and working conditions)	0.265
	Number of cultural infrastructure facilities per 100000 people (Community)	
	Annual unmet need for medical clinics or hospitals (Living and working conditions)	
21	Number of nurses working in medical institutions per 1000 people (Community)	0.264
	Drinking (Individual lifestyle factors)	
	Smoking (Individual lifestyle factors)	
22	Sodium intake (Individual lifestyle factors)	0.263
	Number of beds in dialysis unit (Community)	0.263
	Water intake (cup) (Individual lifestyle factors)	
23	Home income (Living and working conditions)	0.262
	Personal income (Living and working conditions)	

	Number of sports facilities (Community)	
24	number of people in out-of-town clinics (Community)	0.26
	Nutritional labeling awareness (Living and working conditions)	
	Activity limitation (Individual lifestyle factors)	

* Mean dropout loss: M.D. L

Logistic regression analysis to determine variable importance in the elderly population (≥ 65 years)

In addition, logistic regression analysis was conducted to identify risk factors for the top 10 variables that were most significant in the Naïve Bayes model (Table 7).

Compared to males, females had a lower risk of CKD, with a 0.66-fold reduction (OR = 0.34, 95% CI: 0.34–0.35, p < .001). Age positively influenced CKD risk, with each increase in age raising the risk by 1.11 times (OR = 1.11, CI: 1.11–1.11, p < .001). Similarly, diabetes was a significant risk factor, increasing CKD risk by 1.36 times (OR = 1.36, CI: 1.35–1.36, p < .001). Urine protein and anemia were both significant risk factors for CKD. The presence of anemia increased CKD risk by 2.49 times (OR = 2.49, CI: 2.45–2.51, p < .001), while the presence of urine protein increased the risk by 6.06 times (OR = 6.06, CI: 6.02–6.10, p < .001).

Worsening subjective health perception was associated with an increased risk of CKD. Each decline in perceived health status increased the risk by 1.45 times (OR = 1.45, CI: 1.45–1.46, p < .001).

Lack of private insurance was associated with an increased risk of CKD. Individuals without private insurance had a 1.43-fold higher risk of CKD (OR = 1.43, CI: 1.42–1.44, p < .001).

Generational composition negatively influenced CKD risk compared to the first-generation (others) in the standard military. The risk of CKD was 0.65 times lower in the first-generation (single-person household) group (OR = 0.35, CI: 0.34–0.36, p < .001) than in the first-generation (others). The risk of CKD in second-generation (others) was 0.56 times lower (OR = 0.44, CI: 0.43–0.45, p < .001) than that in the first-generation (others). In the third-generation or older, the risk of CKD was 0.46 times lower (OR = 0.54, CI: 0.53–0.56, p < .001) than that in the first-generation (others). In the first-generation (couples), the risk of CKD was 0.77 times lower (OR = 0.23, CI: 0.23–0.24, p < .001) than that in the first-generation (others). In the second-generation (couples + unmarried children), the risk of CKD was 0.75 times lower (OR = 0.25, CI: 0.24–0.25, p < .001) than that in the first-generation (others). In the second-generation (single parents + unmarried children), the risk was 0.47 times lower (OR = 0.53, CI: 0.52–0.55, p < .001) than that in the first-generation (others).

A higher regional rate of moderate to high physical activity was associated with a lower risk of CKD. In cities with higher rates of moderate to high physical activity, the risk of CKD decreased by 0.09 times (OR = 0.91, CI: 0.91–0.91, p < .001). Conversely, a higher number of leisure and welfare facilities per 1,000

elderly individuals was associated with an increased risk of CKD. The risk increased by 1.04 times (OR = 1.04, CI: 1.04–1.04, p < .001) in regions with many such facilities.

Table 7. Risk factors for CKD (≥ 65 years, *N = 1216)*

Rainbow model	Variable	B	S.E.	Sig.	OR	95% C.I
	Intercept	-8.18	.038	<. 001	.00	
Age, sex and constitutional	Sex (Men*)					
	Sex (Women)	-1.07	.004	<. 001	0.34	0.34–0.35
	Age	.10	.000	<. 001	1.11	1.11–1.11
	Diabetes Mellitus	.30	.003	<. 001	1.36	1.35–1.36
	Anemia (Yes)	.91	.004	<. 001	2.49	2.45–2.51
	Urine protein	1.80	.003	<. 001	6.06	6.02–6.10
Individual lifestyle	Subjective health evaluation (Yes)	.38	.002	<. 001	1.45	1.45–1.46
Social and community networks	General composition 1st generation (and so on*)			<. 001		
	1st generation (one person)	-1.05	.013	<. 001	0.35	0.34–0.36
	2nd generation (and so on)	-.82	.013	<. 001	0.44	0.43–0.45
	3rd generation or more	-.61	.013	<. 001	0.54	0.53–0.56
	1st generation (Couple)	-1.46	.012	<. 001	0.23	0.23–0.24
	2nd generation (Couple and unmarried child)	-1.40	.013	<. 001	0.25	0.24–0.25
	2nd generation (Lone parent and unmarried child)	-.63	.014	<. 001	0.53	0.52–0.55
Living and working conditions	Personal insurance (Registered*)			<. 001		
	Personal insurance (Unknown)	-19.6	199.843	.922	.00	
	Personal insurance (Not registered)	.36	.004	<. 001	1.43	1.42–1.44

General socioeconomic cultural environmental conditions	Moderate or higher level of physical activity practice rate	-.096	.001	<.001	0.91	0.91–0.91
	Number of elderly leisure and welfare facilities per 1000 elderly people	.038	.000	<.001	1.04	1.04–1.04

Note. Groups marked with an asterisk are the reference groups. CKD, chronic kidney disease

Discussion

Adult population (≥ 19 years)

In this study, we developed a CKD classification model using a machine learning algorithm for the adult population. The boosting model demonstrated the highest performance in the adult population. This evaluation aligns with previous machine learning studies on CKD, where boosting algorithms demonstrated excellent performance [32,33]. Boosting sequentially learns from the given data while supplementing the performance of the model [34-36]. Although boosting requires more analysis time than that of other models due to its sequential learning process, this drawback is offset by its performance improvement. This analysis technique enhances reliability and reproducibility, which are essential in medical applications.

The risk factors for CKD were identified in the final boosting model for the adult population. The identified biological factors include urinary protein, age, hypertension, diabetes, and anemia, with a urinary protein having the greatest influence. Urinary protein, a key indicator of kidney damage, accumulates when kidney function declines due to impaired protein filtration. Proteinuria is divided into simple and disease-related proteinuria depending on its underlying cause, and its occurrence may not always correlate with albuminuria due to differing mechanisms [37]. Since proteinuria does not always indicate CKD, urinary protein was evaluated as a risk factor in this study, distinguishing it from microalbuminuria, an indicator for CKD diagnosis. Studies report biological factors as key variables [3,38,39]. Regarding work and living standards, CKD risk is higher in the nonprivate insurance group than in the enrolled group. This finding aligns with that of previous studies showing a higher CKD risk in the nonprivate insurance group [38]. Reducing medical expenses through private insurance may improve CKD prevention, early detection, and treatment accessibility by facilitating regular checkups and treatment [40]. For individuals with low living standards, policy support is needed to encourage them to undergo the national health checkup, which includes a renal function test for adults aged 20 years or older every 2 years, improving access to healthcare. Residential area (city/province) was identified as a factor indicating the cultural, general, and social environment. In the capital city of Seoul, CKD risk varies significantly by region. This finding aligns with that of a previous study showing that regional inequality is associated with CKD [41] and highlighting the influence of sociodemographic indices [42]. Community nursing staff should receive education on CKD risk factors and be categorized into risk groups to

facilitate timely treatment when necessary. Policy efforts should focus on monitoring the CKD-related therapeutic environment in each region in collaboration with regional medical institutions to address disparities in healthcare access.

We also analyzed the mean dropout loss values for 54 variables, excluding the seven variables identified as influential in the boosting model. The 54 variables were indirectly significant to the performance of the model. As the rainbow model indicates, health results from multiple factors affecting it through various methods and pathways [43]. Therefore, CKD-related policies for adults should prioritize changes in biological factors and implement interventions that enhance individual capacity to reduce the effect. Individualized health education and guidance should focus on lifestyle improvements, regular checkups and management, as well as medication guidance to help individuals manage their disease, slow its progression or prevent complications. In addition, support and guidance for checkups should align with living and working conditions, while policy should address unmodifiable environmental factors, such as general social, cultural, and environmental factors. Interventions should then gradually expand to lifestyle factors and social networks, prioritized based on mean dropout loss value.

Elderly population (≥ 65 elderly individuals)

A CKD classification model for the elderly population aged ≥ 65 years was developed using a machine learning classification algorithm, with the Naïve Bayes model demonstrating the highest performance. This finding aligns with that of a previous study [44] reporting that the Naïve Bayes model exhibits sensitivity equal to or greater than that of other models in predicting CKD. Naïve Bayes is a data mining algorithm and one of the oldest machine learning classification techniques. It excels in binary classification and it is commonly used for document or spam mail classification, demonstrating excellent performance with large, high-dimensional data [45]. Naïve Bayes is also an effective algorithm in multivariate cases [24]. In the elderly population, major CKD risk factors such as hypertension and diabetes are more clearly defined in adults, making the relatively simple Naïve Bayes model likely effective.

Variable importance was assessed using the mean dropout loss, the average loss function value in the Naïve Bayes model built for the elderly population. All 61 variables fall values within the range of 0.26–0.33, suggesting that the explanatory variables selected based on the rainbow model affect CKD in elderly individuals. In addition, the interactions between variables were examined through logistic regression analysis among the top 10 important variables.

In the Naïve Bayes model developed for the elderly population, biological factors were identified as the most influential variables. Women had a lower risk of CKD than men, while the risk increased with age, diabetes, anemia, and proteinuria. Subjective health perception was identified as a lifestyle-related variable in individual health behavior. The risk of CKD increases with poorer subjective health perception. A previous study reports that positive health perceptions in elderly individuals were associated with health behavior and survival rate [46]. Therefore, for individuals with low subjective health perceptions, health services should be provided in vulnerable areas essential to enhance perception, promote positive

health views through health education, nursing counseling, and health behavior improvement programs, and, if necessary, increase the participation of patients in treatment and self-management.

Generational composition represents the social network variable, indicating whether a household includes 1st, 2nd, or 3rd generations. This study showed that first-generation households have the highest risk of CKD compared to that of other generational compositions. “First-generation other” refers to all multi-person first-generation households, excluding single-person or couple-only households. This category includes cohabiting individuals, unmarried adult siblings, and grandparent-grandchild households. Grandparent-grandchild households consist solely of grandparents and grandchildren. In the absence of the parent generation, grandparents assume parental roles and caregiving responsibilities. In these first-generation households, emotional stability or security may be weak. This challenge is particularly difficult for the elderly to overcome alone, posing a significant health risk. Considering individual circumstances and unique family structures, enhancing medical accessibility through public health center visits, linking social support, conducting regular health checkups and status evaluations, and coordinating management plans with family members are necessary.

Private insurance status, a key factor in living and working standards, was also identified as a significant variable, similar to the adult population. The greater risk of CKD among those without private insurance highlights the need for policies that enhance medical accessibility for elderly individuals with lower living standards. Elderly people without private insurance may neglect health checkups or management after retirement, underscoring the need for appropriate medical support.

Social, cultural, and environmental factors include the rate of moderate to high physical activity and the number of elderly leisure welfare facilities per 1,000 elderly people. A higher rate of moderate to high physical activity correlates with a lower CKD risk. Exercise benefits both physical and renal function in elderly people with CKD [47]. Therefore, encouraging moderate to high levels of physical activity at the community level or creating supportive environments can aid in CKD prevention and management in the elderly. These findings suggest the effect of community environments on health behaviors and CKD, emphasizing the need for local government health policies to promote physical activity among the elderly.

Contrary to expectations, CKD risk increases with the number of elderly leisure welfare facilities per 1,000 elderly individuals. This may be due to the natural increase in CKD prevalence with a growing elderly population. As of 2021, Statistics Korea reports an average of five elderly leisure welfare facilities per 1,000 elderly individuals nationwide. Regional figures vary, with 15 in Jeonnam, 12 in Jeonbuk, 10 in Chungnam, Chungbuk, and Gyungbuk, nine in Gyungnam and Sejong, seven in Gangwon, four in Gwangju and Ulsan, three in Busan, Daegu, Daejeon, and Jeju, and two in Seoul and Incheon [48]. A higher number of elderly leisure welfare facilities also indicates a larger elderly population in the area. Jeonnam had the highest proportion of elderly individuals nationwide at 24%, followed by Gyungbuk at 23%, Jeonbuk and Gangwon at 22%, Busan and Chungnam at 20%, Chungbuk at 19%, Daegu and Gyungnamat 18%, Seoul at 17%, Jeju at 16%, Incheon, Gwangju and Daejeon at 15%, and Ulsan and Gyeonggi at 14% [49]. Beyond

the number of leisure welfare facilities, their quality may also contribute to CKD risk. Inactive programs due to insufficient facility operation or management or structural limitations that hinder the transition from leisure activities to health management may contribute to CKD risk. Therefore, expanding leisure welfare facilities should be accompanied by program improvements that directly support health management. Programs should consider the health status of elderly individuals, and an environment that encourages active participation should be established to enhance health outcomes.

In the elderly population model for those aged ≥ 65 years, community health-related database and social network variables have a relatively greater effect than those in the adult population model. Elderly individuals are more influenced by community health-related factors than those of younger generations. Elderly individuals may be excluded from policies due to retirement or limited access to information, making them vulnerable to inadequate health checkups. Therefore, securing a budget and implementing policies to ensure CKD screening for elderly individuals is essential. To address the growing elderly population, nursing plans and support measures should incorporate environmental factors and social support systems.

General Discussion

This study showed that major risk factors for CKD identified in previous studies were significant variables in the CKD classification algorithm for both adults and elderly individuals. The boosting model for adults and the Naïve Bayes model for the elderly effectively capture the key variables in CKD prediction, producing reliable results by accounting for differences in prevalence and health behaviors between the two populations. In machine learning, selecting an appropriate model that aligns with the characteristics of the problem is essential. Implementing these developed machine learning classification models in clinical nursing settings can extend the role of nursing beyond traditional human-centered care [50].

This study is significant because, unlike previous studies that primarily focused on individual physiological indicators, it selects variables based on theoretical foundations to develop a CKD classification model and examines the interactions among various influencing factors. This approach differs from those of previous studies by identifying multidimensional CKD risk factors and enabling personalized nursing and preventive interventions. In addition, this study is valuable as it integrates all four major areas of nursing, i.e., research, education, policy, and practice, enabling a multidimensional approach.

The study has some limitations. First, the use of cross-sectional survey data makes it difficult to identify the temporal relationship between risk factors and CKD prevalence. Second, the analysis was limited to factors collected from the National Health Survey and community database, excluding some potential risk factors. Third, the data preprocessing process did not ensure population and analysis group homogeneity (see additional file). Machine learning classification algorithms can be continuously

improved and supplemented through additional learning. Further training, analysis, and comparisons are necessary to enhance various aspects of the developed model.

Conclusion

This study was conducted to develop a CKD classification model and identify key risk factors affecting CKD by integrating a community database with the 2021 KNHANES. Using the rainbow model, key factors affecting CKD served as explanatory variables, and six machine learning classification analyses were applied to develop a CKD classification model.

This study showed boosting as the most predictive CKD classification model for individuals aged ≥ 19 years and Naïve Bayes for those aged ≥ 65 years. The selected model effectively classified CKD, confirming the influence of variables selected based on the social decision model.

The CKD classification machine learning model developed in this study serves as a foundation for nursing practice. It can identify CKD risk factors, predict at-risk groups, establish nursing plans for patients, and provide personalized care and education.

Abbreviations

CKD, Chronic Kidney Disease; KNHANES, Korea National Health and Nutrition Examination Survey; Jeffreys's Amazing Statistics Program, JASP; Statistical Package for the Social Sciences, SPSS; RI, Relative Influence; M.L.D, mean dropout loss; AUC, area under the curve.

Declarations

Ethics approval and consent to participate

All methods were carried out in accordance with the Declaration of Helsinki. This study was approved as a secondary data analysis study and received exemption from IRB review by the Kyungpook National University Bioethics Committee (2024-0278).

Consent for publication

Not applicable.

Availability of data and materials

Data sharing is not applicable to this article as no datasets were generated during the current study. In this study, we downloaded Korea National Health and Nutrition Examination Survey data publicly available on the Korea Centers for Disease Control and Prevention (KNHNES) website (https://knhanes.kdca.go.kr). We used the community health-related factor database downloaded from a

publicly available website (https://chs.kdca.go.kr). Guided research data are provided upon reasonable request and with the permission of the relevant institution.

Competing interests

The authors declare that they have no competing interests.

Funding

Not applicable.

Author contributions

JK contributed to the study design, performed the statistical analyses, interpreted the data and drafted the manuscript. SHL designed and supervised the study, interpreted the data, and critically revised the manuscript. All the authors read and approved the final manuscript. This paper is based on JK's Doctoral dissertation research completed under the direction of Dr. SHL.

Acknowledgments

Not applicable.

Additional materials

Additional files.docx. provide information on general characteristics of missing values removed during data preprocessing and on tests for homogeneity of population and analysis groups.

References

1. Beto JA, Bansal VK. Nutrition interventions to address cardiovascular outcomes in chronic kidney disease. Adv Chronic Kidney Dis. 2004;11:391-7.
2. Trends in the Prevalence of Chronic Kidney Disease, 2011-2021. Public Health Wkly Rep. 2023;16:238-9. https://doi.org/10.56786/PHWR.2023.16.8.3
3. Saran R, Robinson B, Abbott KC, Agodoa LY, Bhave N, Bragg-Gresham J, et al. US renal data system 2017 annual data report: epidemiology of kidney disease in the United States. Am J Kidney Dis. 2018;71:A7. https://doi.org/10.1053/j.ajkd.2018.01.002
4. Korean Society of Nephrology. End-stage kidney disease fact sheet. 2024.
5. Centers for Disease Control and Prevention. Trends in the Number of Chronic Kidney Disease Patients and Medical Expenses, During 2012-2022. Public Health Wkly Rep. 2024;17:381-2. https://doi.org/10.56786/PHWR.2024.17.9.3
6. Kim DH, Hyun YY, Cha JJ, Lee S, Lee HK, Choi JW, et al. Kidney Health Plan 2033 in Korea: bridging the gap between the present and the future. Kidney Res Clin Pract. 2024;43:8. https://doi.org/10.23876/j.krcp.23.232.

7. Kim YH. Recent therapeutic strategies based on the stage of chronic kidney disease. Korean J Med. 2009;76:521-7.
8. Hong YA, Ban TH, Kang C, Hwang SD, Choi SR, Lee H, et al. Trends in epidemiologic characteristics of end-stage renal disease from 2019 Korean Renal Data System (KORDS). Kidney Res Clin Pract. 2021;40:52. https://doi.org/10.23876/j.krcp.20.202
9. Chen C, Tsai Y, Hsu H, Wu I, Sun C, Chou C, et al. Depression and suicide risk in hemodialysis patients with chronic renal failure. Psychosomatics. 2010;51:528-6e. https://doi.org/10.1016/S0033-3182(10)70747-7
10. Kang EJ, Oh KH. Prognosis of chronic kidney disease patients in Korea through long-term follow-up study of survival and preservation of renal function in patients with chronic kidney disease. Public Health Wkly Rep. 2023;16:271-9.
11. Lee SH. Opening Ceremony Seminar: Step-by-step management of patients with chronic kidney disease. Korean Soc Intern Med Spring Conf Papers. 2014:27-30.
12. Göran D, Whitehead M. Policies and strategies to promote social equity in health. 1991.
13. Kim MJ, Park NH. Analysis of spatial distribution of hypertension prevalence and its related factors based on the model of social determinants of health. J Korean Acad Community Health Nurs. 2018;29:414-28. https://doi.org/10.12799/jkachn.2018.29.4.414
14. Lee J. Regional health inequalities and individual and regional-level health determinants. Health Soc Res. 2016;36:345-84.
15. Da Silva MVB, Ferreira ET, de Lima Filho CA, Basílio VKV, Lobo MJS, Gava PHR, et al. Efeitos dos determinantes sociais da saúde na hipertensão: uma revisão sob a luz do modelo de Dahlgren e Whitehead. J Educ Sci Health. 2023;3:1-13. https://doi.org/10.52832/jesh.v3i1.172
16. Korea Disease Control and Prevention Agency (KDCA). Korea National Health and Nutrition Examination Survey raw data usage guidelines, 8th edition (2019-2021). 2023.
17. Agarwal A, Verma A, Khari M. Comparative Assessment of Machine Learning Methods for Early Prediction of Diseases Using Health Indicators. In: Approaches to Human-Centered AI in Healthcare. IGI Global; 2024. p. 160-86. https://doi.org/10.4018/979-8-3693-2238-3.ch007
18. Ma Q. Recent applications and perspectives of logistic regression modelling in healthcare. Theor Nat Sci. 2024; 36:185-90. https://doi.org/10.54254/2753-8818/36/20240614
19. Jeong S, Lee M, Yoo S. Machine learning-based stroke risk prediction using public big data. J Adv Navig Technol. 2021;25:96-101. https://doi.org/10.12673/jant.2021.25.1.96
20. Lim HK. Prediction of myocardial infarction/angina and selection of major risk factors using machine learning. J Korean Data Anal Soc. 2018; 20:647-56. https://doi.org/10.37727/jkdas.2018.20.2.647
21. Aoki J, Kaya C, Khalid O, Kothari T, Silberman MA, Skordis C, et al. CKD progression prediction in a diverse US population: A machine-learning model. Kidney Med. 2023;5:100692. https://doi.org/10.1016/j.xkme.2023.100692

22. Luo W, Gong L, Chen X, Gao R, Peng B, Wang Y, et al. Lifestyle and chronic kidney disease: a machine learning modeling study. Front Nutr. 2022; 9:918576. https://doi.org/10.3389/fnut.2022.918576
23. Segal Z, Kalifa D, Radinsky K, Ehrenberg B, Elad G, Maor G, et al. Machine learning algorithm for early detection of end-stage renal disease. BMC Nephrol. 2020; 21:1-10. https://doi.org/10.1186/s12882-020-02093-0
24. Bucur IG, Claassen T, Heskes T. MASSIVE: Tractable and Robust Bayesian Learning of Many-Dimensional Instrumental Variable Models. In: Proceedings of the Conference on Uncertainty in Artificial Intelligence. 2020. p. 1049-58.
25. Dahlgren G, Whitehead M, World Health Organization. Levelling up (part 2): A discussion paper on European strategies for tackling social inequities in health. Copenhagen: WHO Regional Office for Europe; 2006.
26. Mahalakshmi AL. Evidence based practice. In: Futuristic trends in pharmacy & nursing. Vol. 3, Book 5, p. 509-19. IIP Series; 2024. https://doi.org/10.58532/V3BAPN5CH45
27. Carter JV, Pan J, Rai SN, Galandiuk S. ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves. Surgery. 2016;159:1638-1645. https://doi.org/10.1016/j.surg.2015.12.029
28. Wald NJ, Bestwick JP. The area under the ROC curve: is it a valid measure of screening performance? J Med Screen. 2014;21:220. https://doi.org/10.1177/0969141314548908
29. Park YJ, Kang HK. A study on the performance comparison of classification algorithms for applying machine learning to predict diabetic kidney disease. J Korea Acad Ind Coop Soc. 2022;23:184-191.
30. Jeong SO. Development of a machine learning-based prediction model for influenza vaccination and exploration of influencing factors [dissertation]. Ewha Womans University Graduate School; 2023. Available from: https://www.riss.kr/link?id=T16806105.
31. Kim YP. Cautionary points for statistical analysis of complex sample design data. Korean J Fam Pract. 2016;6:61-62.
32. Ogunleye A, Wang Q. XGBoost model for chronic kidney disease diagnosis. IEEE/ACM Trans Comput Biol Bioinform. 2019;17:2131-40. https://doi.org/10.1109/TCBB.2019.2911071
33. Park YJ, Kang HK. A study on the performance comparison of classification algorithms for applying machine learning to predict diabetic kidney disease. J Korea Acad Ind Coop Soc. 2022;23:184-91.
34. Kim G. JASP data analysis: From basic statistics to machine learning. Hanarae Academy; 2023.
35. Kwon CM. Python machine learning complete guide. Wikibooks; 2022.
36. Yoon YJ. Boosting algorithm for large-scale and sequential data. J Appl Stat. 2010;23:197-206.
37. Hong SC. Correlation between albuminuria and proteinuria in patients with chronic kidney disease [dissertation]. Seoul: Hanyang University; 2015.
38. Kim E. Factors affecting the occurrence of chronic kidney disease in adult: secondary analysis using data from the 7th Korea National Health and Nutrition Examination Survey. J Korean Data Anal Soc. 2018;20:1037-50.

39. Seo HJ. Effects for comorbidities of chronic kidney disease on the progression to end-stage renal disease. J Health Inform Stat. 2020;45:356-64. https://doi.org/10.21032/jhis.2020.45.4.356
40. Lee Y, Im B, Park Y. The determinants and comparison of health behavior and health service by private medical insurance on National Health-Nutrition Survey. J Korea Contents Assoc. 2010;10:190-204. https://doi.org/10.5392/JKCA.2010.10.12.190
41. Akrawi DS, Li X, Sundquist J, Sundquist K, Zöller B. End stage renal disease risk and neighbourhood deprivation: a nationwide cohort study in Sweden. Eur J Intern Med. 2014;25:853-9. https://doi.org/10.1016/j.ejim.2014.09.016
42. Bikbov B, Purcell CA, Levey AS, Smith M, Abdoli A, Abebe M, et al. Global, regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet. 2020;395:709-33. https://doi.org/10.1016/S0140-6736(20)30045-3
43. Lee SE. The moderating effect of economic activity participation on the relationship between subjective health status and life satisfaction in older adults with chronic diseases. Ment Health Soc Work. 2012;40:234-62.
44. Bai Q, Su C, Tang W, Li Y. Machine learning to predict end stage kidney disease in chronic kidney disease. Sci Rep. 2022;12:8377. https://doi.org/10.1038/s41598-022-12316-z
45. Kim M, Choi H, Park C. Categorical variable selection in Naive Bayes classification. Korean J Appl Stat. 2015;28:407-15. http://dx.doi.org/10.5351/KJAS.2015.28.3.407
46. Deeg DJ, Bath PA. Self-rated health, gender, and mortality in older persons: introduction to a special section. Gerontologist. 2003;43:369-71. https://doi.org/10.1093/geront/43.3.369
47. Hyeon-Ju L, Youn-Jung S, Eun JS. Effectiveness of exercise for improving physical and renal function in older adults with pre dialysis chronic kidney disease: A systematic review and meta-analysis. J Korean Crit Care Nurs. 2023;16(3). https://doi.org/10.34250/jkccn.2023.16.3.34
48. Ministry of Health and Welfare, Long-term Care Insurance Division. Number of elderly leisure and welfare facilities per 1,000 elderly population (by province/city/county). Available from: https://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1YL20961&conn_path=I2. 2021.
49. Statistics Korea. 2021 statistics on elderly individuals. 2021.
50. Lee HB, Moon WJ, Kim SA, Lee JH, Jang OJ. Exploring the applicability of artificial intelligence for improving nursing practice in Korea. J Korean Academy Nurse Adm. 2023;29:564-76. https://doi.org/10.11111/jkana.2023.29.5.564

Figures

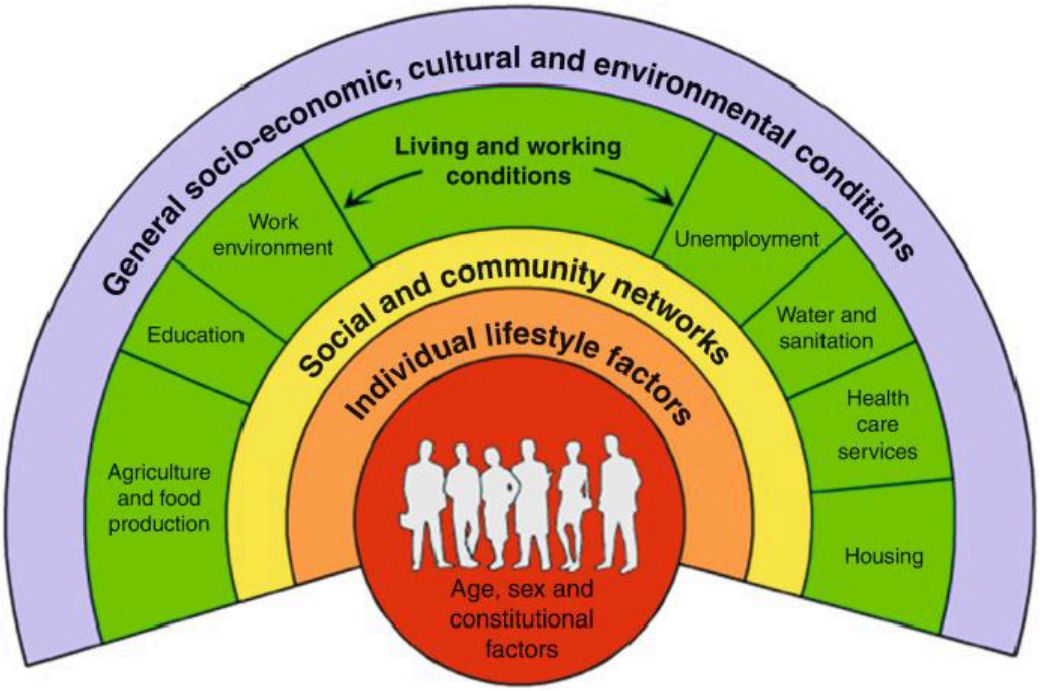


Figure 1

Dahlgren and Whitehead’s rainbow model (2007) [40]

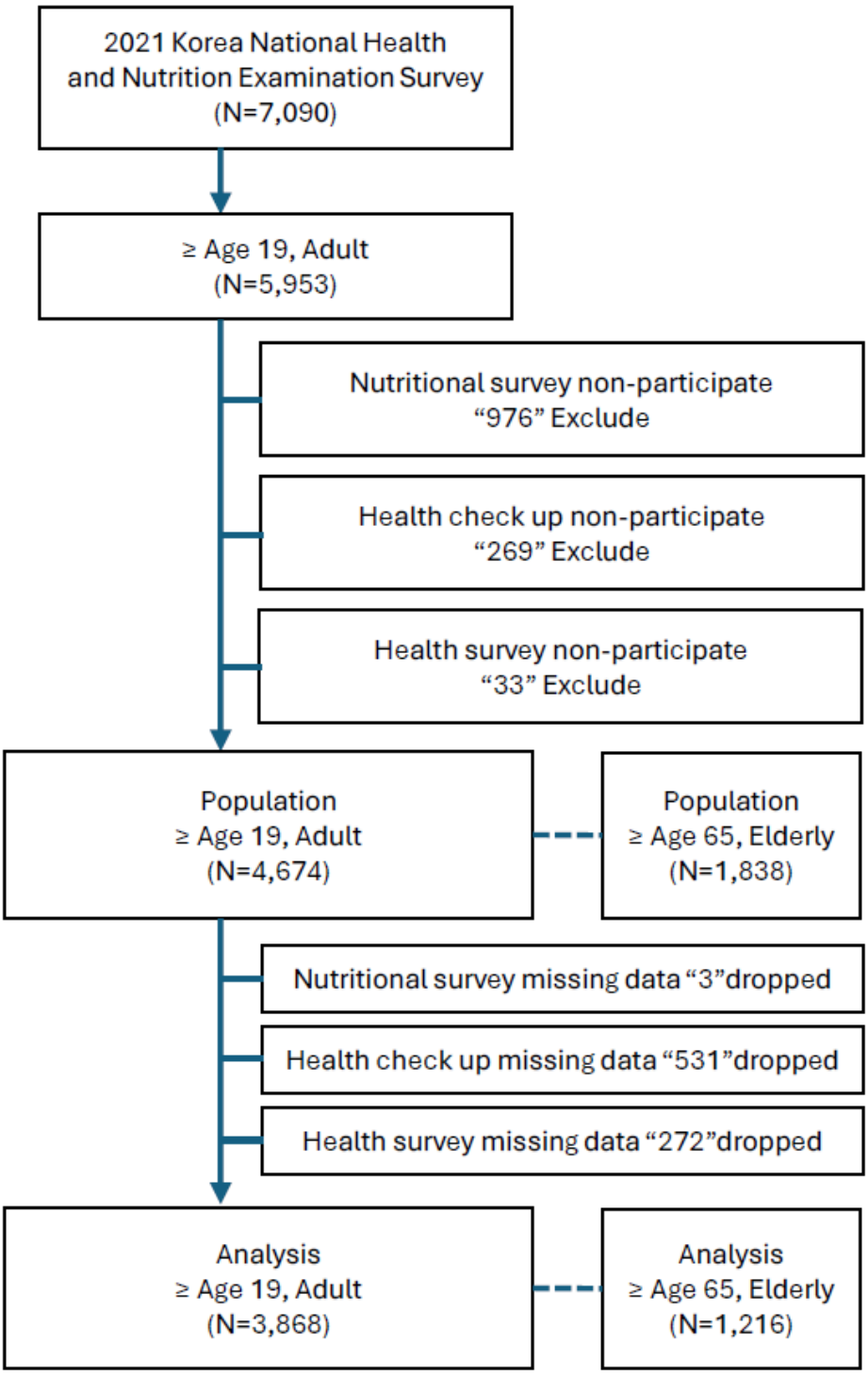


Figure 2

Flow Chart of the Participants

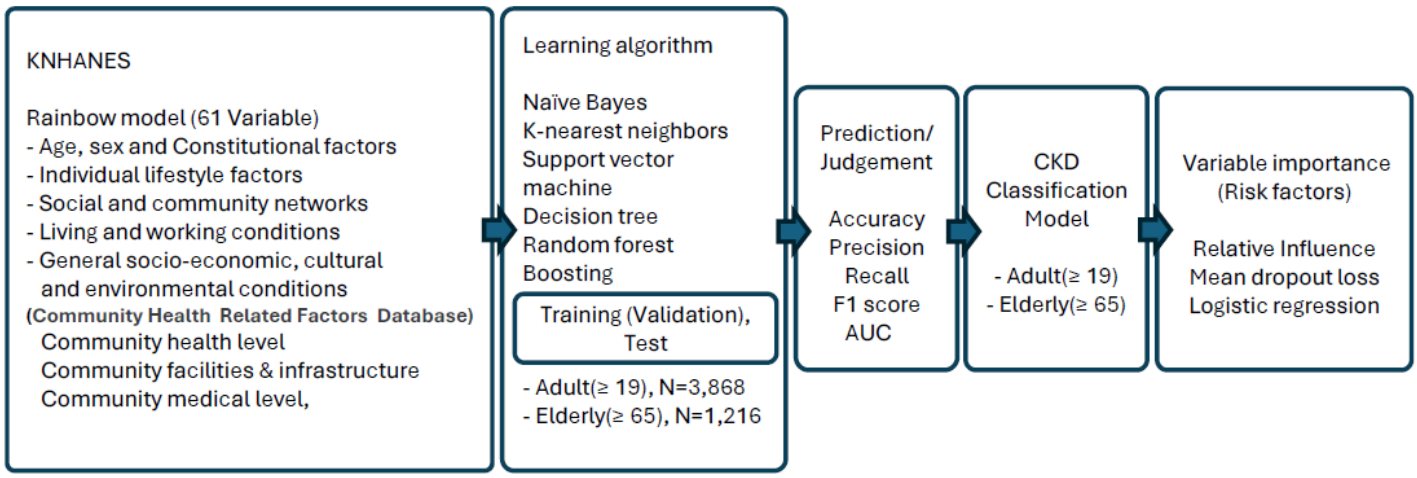


Figure 3

Process of Data Analysis

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- additionalfilePredictingChronicKidneyDiseaseRiskFactorsUsingMachineLearning.docx