KIPS Trans. Softw. and Data Eng.
Vol.11, No.3 pp.107~114
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2022.11.3.107

Comparison of Feature Selection Methods Applied on Risk Prediction for Hypertension

Dashdondov Khongorzul† · Mi-Hye Kim††

ABSTRACT

In this paper, we have enhanced the risk prediction of hypertension using the feature selection method in the Korean National Health and Nutrition Examination Survey (KNHANES) database of the Korea Centers for Disease Control and Prevention. The study identified various risk factors correlated with chronic hypertension. The paper is divided into three parts. Initially, the data preprocessing step of removes missing values, and performed z-transformation. The following is the feature selection (FS) step that used a factor analysis (FA) based on the feature selection method in the dataset, and feature importance (FI) and multicollinearity analysis (MC) were compared based on FS. Finally, in the predictive analysis stage, it was applied to detect and predict the risk of hypertension. In this study, we compare the accuracy, f-score, area under the ROC curve (AUC), and mean standard error (MSE) for each model of classification. As a result of the test, the proposed MC-FA-RF model achieved the highest accuracy of 80.12%, MSE of 0.106, f-score of 83.49%, and AUC of 85.96%, respectively. These results demonstrate that the proposed MC-FA-RF method for hypertension risk predictions is outperformed other methods.

Keywords : KNHANES, Hypertension, Feature Selection, Multicollinearity, Factor Analysis

고혈압 위험 예측에 적용된 특징 선택 방법의 비교

Dashdondov Khongorzul† · 김 미 혜††

요 약

본 논문에서는 질병관리청 국민건강영양조사(KNHANES: Korea National Health and Nutrition Examination Survey) 데이터베이스에서 특징선택 방법으로 고혈압을 감지 예측하는 방법을 개선했다. 또한 만성 고혈압과 관련된 다양한 위험 요인을 확인하였다. 본 논문은 3가지로 나누어, 첫째 결측값을 제거하고 Z-변환을 하는 데이터 전처리 단계이다. 다음은 데이터 셋에서 특징선택법을 기반으로 하는 요인분석(FA)을 사용하는 특징선택 단계이며, 특징선택을 기반으로 다중공선형 분석(MC)와 특징중요도(FI)을 비교했다. 마지막으로 예측분석단계에서 고혈압 위험을 감지하고 예측하는데 적용했다. 본 연구에서는 각 분류 모델에 대해 ROC 곡선(AUC) 아래의 평균 표준 오차(MSE), F1 점수 및 면적을 비교한다. 테스트 결과 제안한 MC-FA-RF모델은 80.12% 가장 높은 정확도를 보이고, MSE, f-score, AUC 모델의 경우 각각 0.106, 83.49%의, 85.96% 으로 나타났다. 이러한 결과는 고혈압위험 예측에 대한 제안된 MC-FA-RF 방법이 다른 방법에 비해 우수함을 보이고 있다.

키워드 : 국민건강영양조사, 고혈압, 특징선택, 다중공선형성, 요인분석

1. Introduction

Hypertension is a serious health condition that increases the risk of brain, heart, kidney, and other diseases [1,2]. It is the leading cause of premature death worldwide and affects more than 1 in 4 men and 1 in 5 women [1].

In recent years, the incidence of hypertension among young people has increased dramatically, not only in the elderly. Some researchers have tried to prove the predisposing factors for adult hypertension in adolescents aged 10-19 years. They concluded using MC analysis, that the wrist circumference and total cholesterol have an important role in the risk of hypertension through adolescence to adulthood [2]. In addition, [3] focused on the sodium component of salt in the study of dietary factors associated with hyper-

※ This research was financially supported by the Ministry of Trade, Industry, and Energy (MOTIE) and Korea Institute for Advancement of Technology (KIAT) through the National Innovation Cluster R&D program (Cooperative Regional Industry Development Program with the relocated Public Institutes, P0002072).
※ 이 논문은 2021년 한국정보처리학회 춘계학술발표대회에서 "Feature selection-based Risk Prediction for Hypertension in Korean men"의 제목으로 발표된 논문을 확장한 것임.
† 정 회 원 : 충북대학교 컴퓨터공학과 연구원
†† 정 회 원 : 충북대학교 컴퓨터공학과 교수
Manuscript Received : June 30, 2021
First Revision : July 29, 2021
Accepted : August 19, 2021
* Corresponding Author : Mi-Hye Kim(mhkim@cbnu.ac.kr)

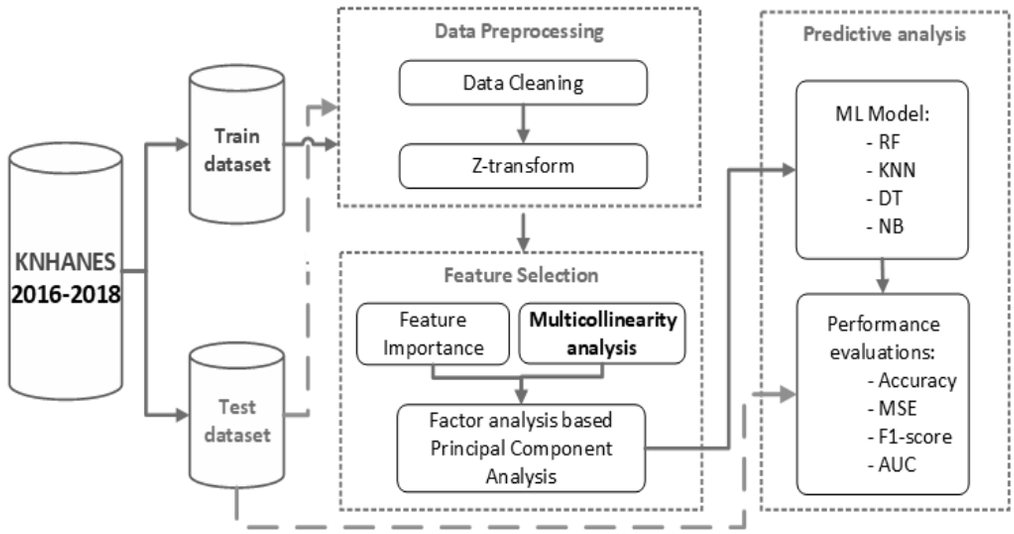

Fig. 1. The Experimental Model of the Proposed Method

tension using MC analysis. In this consider, the use of ML algorithms to diagnose the reason and risks of hypertension has raised in recent years. ML is a learning process that starts with observations or data, such as examples, first-hand experience, guidance, etc., to make well determinations in the future based on given examples. The main goal is to approve computers to learn and adjust their actions automatically without human agency or assistance. [4] explored the hypertension awareness, treatment, prevalence, and control rate between 2008 to 2017 of KNHANES. Our previous work [5,6] implemented a method of removal to multivariate outliers to improve the performance of predictive analyzes. In this study, we proposed the feature selection based on FA, taking into account the social and demographic characteristics associated with hypertension, which were then detected in more detail using the commonly used ML algorithm.

2. System Techniques

In this part, we explain the components of the proposed prediction model. Fig. 1 shows the proposed system based on the FA-based FS method. The proposed structure consists of three main parts: data preprocessing, feature selection, and predictive analysis.

2.1 Data Preprocessing

System missing values are values that are fully absent from the data. Data missing values are values that are imperceptible while analyzing or correcting data. To prevent the model issues, we eliminated all of the missing values and no responses are omitted from the main dataset. Also, we removed unrelated attributes with hypertension from the dataset.

2.2 Feature Selection

The feature selection module is performed into two parts: 1) DT classifier feature importance (FI) with FA based principal component analysis (PCA) feature selection; 2) MC analysis with FA based PCA feature selection. The purpose of this study is to predict hypertension from the KNHANES dataset. This dataset contained a health examination of several numbers of diseases, health questions, and nutrition information of the Korean population, contents with 1193 attributes. Therefore, we selected the important features for proposed a simple accurate model. We compared decision tree (DT) based FI and MC analysis are two methods to select features. For decision tree-based FS, we changed input parameters several times and chose the optimal value that gave the best result. In this study, the optimal value of criterion used for the “gini”. Then, we compared this method with our proposed method, and the proposed method outperformed.

1) Feature Importance

We applied a DT classifier to selected important features in hypertension. In this stage, if the importance score is greater than zero then the features will

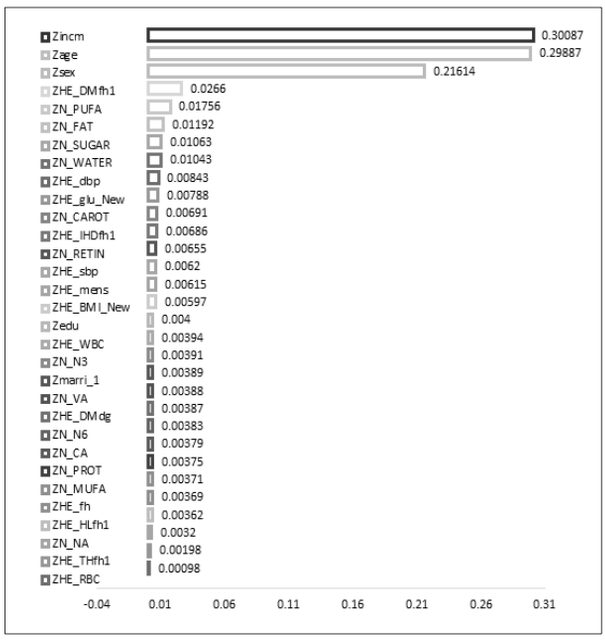

Fig. 2. DT-based Feature Importance in the KNHANES Dataset

be selected. Our suggested DT-based important features explained the prediction algorithm across the clean dataset in Fig. 2. For the KNHANES dataset, "income", "age", "sex", "Diabetes Doctor Diagnosis", "polyunsaturated fatty acids", "Fat", "Sugar", and "Water" were provided as most valuable features with importance scores of 0.30087, 0.29887, 0.21614, 0.0266, 0.01756, 0.01192, 0.01063, and 0.01043 to predict among the Korean nations shown Fig. 2.

2) Multicolinearity Analysis

We verify to check the collinearity between selected health examination, nutrition, and basic information features with hypertension using MC in regression analysis. MC is a statistical notion for when two or more input attributes, the valuations from the features, in this case, are very extremely correlated [6]. If highly correlated variables can relate to each other, the features should be removed. The tolerance results and variance inflation factor (VIF) are verified using MC analysis. If the VIF value is greater than 10, and the tolerance of less than 0.10, then there is a MC problem [7,8]. In total, 33 features remained at the end of this analysis and are is used as inputs to the next factor analysis. Fig. 3 demonstrates the result of the Variance Inflation Factor (VIF) value and Fig. 4 demonstrates the tolerance value for the selected features. The VIF value of the Z Score (NBA) feature is not calculated, because the tolerance value was equal to 0, and Z Score (N_VA_RAE) feature VIF value has very high. In that case, these features are excluded from the collinearity analysis. The full results of the correlation and MC analysis of the KNHANES dataset in Table A1 are shown in Appendix A.

3) Factor Analysis

FA is a method used to moderate the number of attributes to a any factors. This method estimates and scores the most usual variance of all attributes. In this paper, we used the PCA for the extracted factor from the experimental dataset by the maximum variance and assigns them to the first factor [9]. Then the variance defined by the initial factor is estimated and the high-

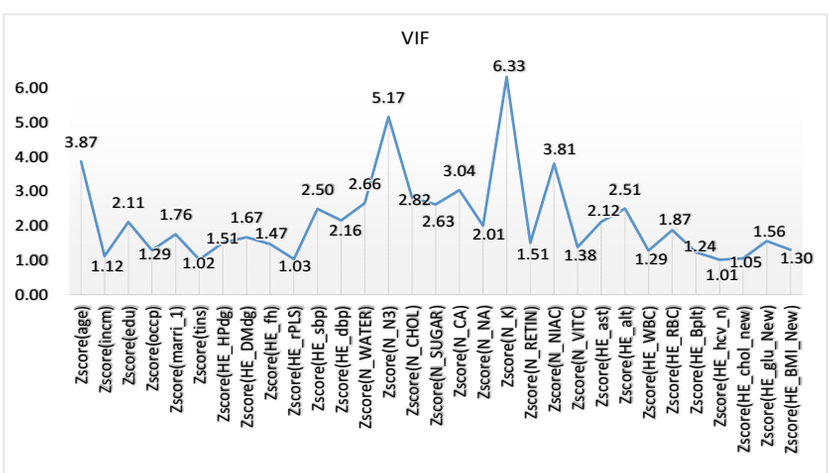

Fig. 3. Results of the Variance Inflation Factor of MC Analysis on the Selected Features

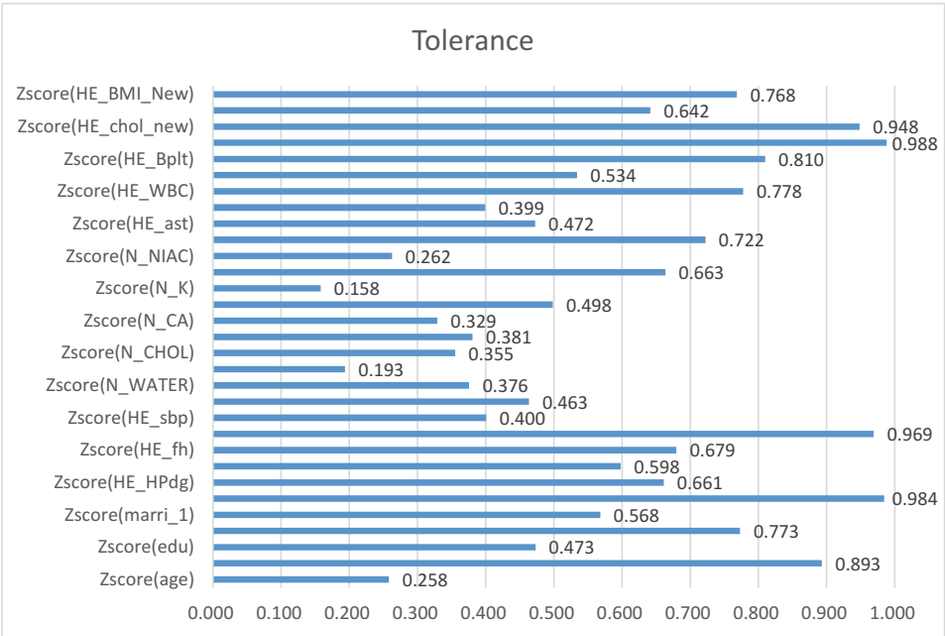

Fig. 4. Results of the Tolerance Value of Multicollinearity Analysis on the Selected Features

est variance of the second factor is computed. This operation simulated to the final factor. The model shows the correlation between the variables because KMO factor = 0,786, determinant = 0.0012 pattern sizes are adaptable, and significance is equal to 0.001 (alpha<0.005), rejected the null hypothesis. Here, we can perform FA.

2.3. Predictive Analysis

To develop the execution of the predictive analysis, we aim to at the training database. That is, the dimension of the features from the training database was reduced by using PCA-based FA instead of direct train classification. The k-Nearest Neighbors (KNN), DT, random memory (RF), and Naïve Bayes (NB) algorithms were then used in the training database prepared [5,6,9,10].

1) Preparing Experimental Dataset

In this research, the Korean National Health and Nutrition Examination Survey dataset were used to make the suggested model for the hypertension risk prediction. KNHANES is handled by the Disease Control and Prevention (KCDC) [1]. It included a diseases health inspection, questions, and nutrition information of Koreans. We evaluated the pattern duration for 2016-2018. Additionally, we created the target value for hypertension patients greater than 19 years old. Furthermore, this target group was created by subjects who had a history of heart disease, heart attack, diabetes, prediabetes, and stroke. We eliminated 7,176 of the 16,489 entries and 58 of the 1,192 attributes for unrelated features to hypertension and missing values of the row. Following, we removed attributes based on FI and MC analysis which there a total of 31 features for FI, and 33 features for the MC method. After that we selected eliminated features based on FA which separated method FI and MC, there 10 factor for MC-FA, and 9 factor for FI-FA. Fig. 5 shows the procedure of creating the target dataset.

2) Details of the Data Structure

In the initial configuration of the training (70%) and test (30%) sets, we identified three categories of labels: normal, pre-hypertension, and hypertension for the target characteristics of the experimental data set. Target variables descriptive statistics are described in Table 1 [12].

This article describes hypertension in patients with systolic blood pressure (SBP) above 140, diastolic blood pressure (DBP) above 90, and the use of antihypertensive drugs (AHD) as prescribed by a doctor. The criteria

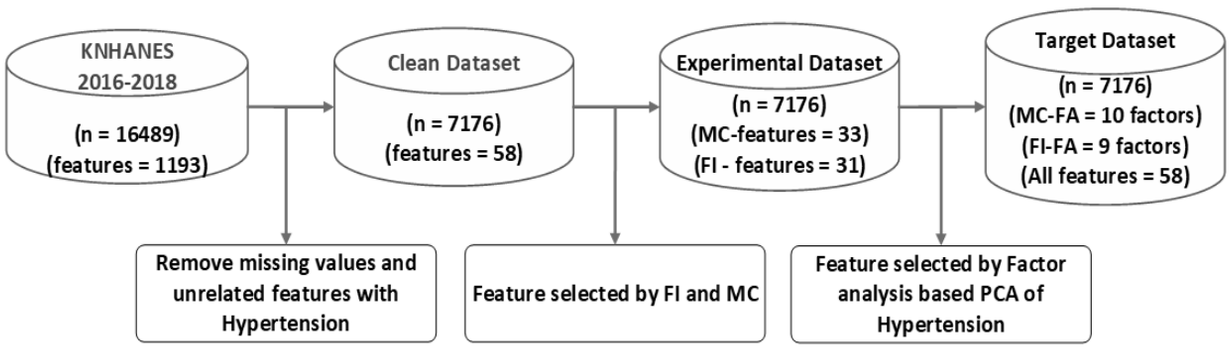

Fig. 5. Experimental Dataset Preparation Procedure of Hypertension

Table 1. Descriptions of Target Variable

Class	Total	Train (70%)	Test (30%)
Normal	3172	2216	956
Pre-hypertension	1685	1222	463
Hypertension	2319	1585	734
Total	7176	5023	2153

for labeling hypertension are as follows. If the DBP ranges from 0 to 80 and the SBP ranges from 0 to 120, it is labeled normal. If the DBP ranges from 80 to 90 and the SBP ranges from 120 to 140, the label indicates pre-hypertension. Conversely, a classification label will hypertension.

3. Experimental Results

The data set was used to suggest a algorithm for predicting the hypertension risk in previous experimental data [6,13]. At first, we deleted a row of missing values and irrelevant attributes to hypertension. Following, we choose DT-based model important features which there a total of 31 attributes from 58; and the MC analysis-based model which there a total of 33 attributes from 58. Later, we selected features according to the feature importance with factor analysis (FI-FA) and multicollinearity with factor analysis (MC-FA) model of hypertension risk detection, in this case, five attributes.

The accuracy, MSE, f-score, and AUC [5] evaluations of the experimental results are presented in Table 2, and the highest values of evaluation scores are marked in bold. If we did not use the MC-FA and FI-FA models to predict risk factor levels, the KNN algorithm shows

Table 2. Comparison Results of the Suggested Methods for Experimental Target Dataset

Algorithms		Acc (%)	MSE	f-score (%)	AUC (%)
MC-FA 10 factor	**RF**	**80.12**	**0.106**	**83.49**	**85.96**
	KNN	**72.13**	**0.169**	**74.12**	**78.41**
	DT	68.88	0.142	78.84	82.78
	NB	65.54	0.165	69.09	74.79
FI-FA 9 factor	RF	70.87	0.146	76.46	80.34
	KNN	65.86	0.205	68.12	73.99
	DT	56.43	0.205	69.56	75.79
	NB	61.82	0.185	65.20	72.16
All features	RF	64.17	0.176	68.44	75.47
	KNN	65.67	0.204	67.47	74.28
	DT	51.27	0.238	65.03	72.25
	NB	58.03	0.210	59.34	70.38

the highest performance than other algorithms. However, FA with FI and MC model-based algorithms outperformed all features algorithms. Our proposed MC-FA-RF model executed the highest accuracy of 80.12%, MSE of 0.106, f-score of 83.49%, and AUC of 85.96%. Afterwards, the second-best accuracy of 72.13%, MSE of 0.169, f-score of 74.12%, and AUC of 78.41% were executed by KNN with the MC-FA.

It can be obviously seen from Table 2 that 33 and 31 dimensions were diminished to 10 and 9 dimensions by FA, the AUC of risk prediction growed in all methods. This described that the suggested FA-based feature reduction method is suitable for the hypertension risk prediction.

The ROC curve is the main evaluation measure of detection performance. We presented the ROC curve for each method on the target data set in Fig. 6 ~ Fig. 8.

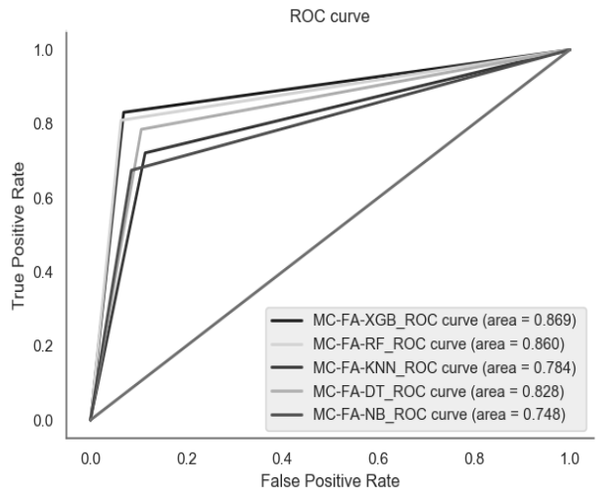

Fig. 6. ROC Area Curve of Proposed Algorithm Based on MC-FA Method

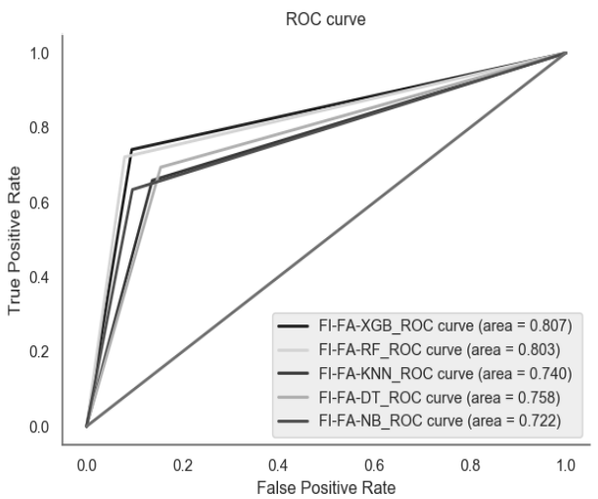

Fig. 7. ROC Area curve of Proposed Algorithm Based on FI-FA Method

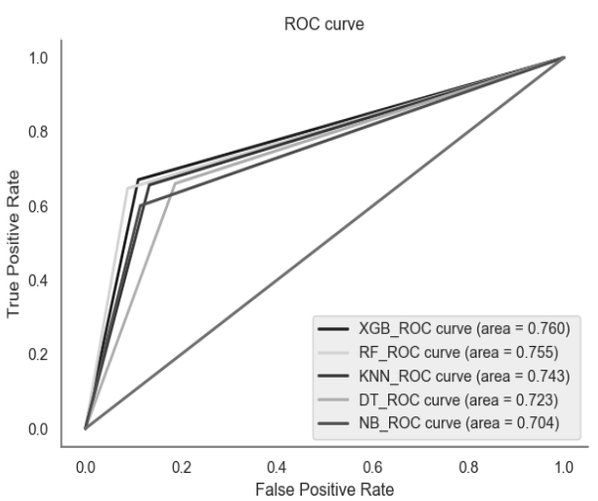

Fig. 8. ROC Area Curve of Proposed Algorithm for All Features

4. Conclusion

In this research, we enhanced the prediction of hypertension risk with the help of an FA-based selection for the Korean National Health Information open data. These analyses have identified a character of risk factors correlated with hypertension. These important characteristics were selected based on the MC analysis. Following, extracted from a PCA-based FA model. Studies have shown that high blood pressure is related to social and demographic factors such as age, household income, education, occupation, BMI, and other health factors. Besides, hypertension is determined by nutrients such as energy, water, sugar, calcium, and others. Experimental results show that the proposed MC-FA-RF method increased accuracy, MSE, f-scores, and AUC results by 15.95%, 0.07, 15.05%, and 10.49%, respectively.

References

[1] Korea Centers for Disease Control & Prevention [Internet], http://knhanes.cdc.go.kr.

[2] S., Kalantari, et al., "Predictors of early adulthood hypertension during adolescence: A population-based cohort study," *BMC Public Health*, Vol.17, No.1, pp.1-8, 2017.

[3] J. van der Leeuw, M. H. de Borst, L. M. Kieneker, S. J. Bakker, R. T. Gansevoort, and M. B. Rookmaaker, "Separating the effects of 24-hour urinary chloride and sodium excretion on blood pressure and risk of hypertension: Results from PREVEND," *PloS one*, Vol.15, No.2, pp.e0228490, 2020.

[4] K. Kim, E. Ji, J. Y. Choi, S. W. Kim, S. Ahn, and C. H. Kim, "Ten-year trends of hypertension treatment and control rate in Korea," *Scientific Reports*, Vol.11, No.1, pp.1-8, 2021.

[5] K. Dashdondov and M. H. Kim, "Multivariate outlier removing for the risk prediction of gas leakage based methane gas," *Journal of the Korea Convergence Society*, Vol.11, No.12, pp.23-30, 2020.

[6] K. Dashdondov and M. H. Kim, "Prediction of hypertension in Korean men using the outlier detection method," *International Conference on the Multimedia and Ubiquitous Engineering (MUE2021)*, Jeju, Korea, Apr. 22-24, 2021.

[7] D. E. Farrar and R. R. Glauber, "Multicollinearity in regression analysis: The problem revisited," *Review of Economics and Statistics*, Vol.49, No.1 pp.92-107, 1967.

[8] R. M. O'brien, "A caution regarding rules of thumb for variance inflation factors," *Quality & Quantity*, Vol.41, No.5, pp.673-690, 2007.

[9] V. N. Vapnik, “The nature of statistical learning theory,” Springer, New York, 1995.
[10] W., Chang, et al., “A machine-learning-based prediction method for hypertension outcomes based on medical data,” *Diagnostics*, Vol.9, No.4, pp.178, 2019.
[11] D. J. Denis, “Applied univariate, bivariate, and multivariate statistics: Understanding statistics for social and natural scientists, With Applications in SPSS and R.,” John Wiley & Sons, 2021.
[12] K. Dashdondov and M. H. Kim, “Mahalanobis distance based multivariate outlier detection to improve performance of hypertension prediction,” *Neural Processing Letters*, pp.1-13, 2021.

Dashdondov Khongorzul
https://orcid.org/0000-0001-5113-8542
e-mail : khongor@chungbuk.ac.kr
She received a Ph.D. degree in Radio and Communication Engineering from Chungbuk National Univ. in 2013. She has been a post-doctoral researcher at Chungbuk National Univ. since 2017. Her research interests are in the area of Probability and Statistics, Queueing theory, Image processing, Machine Learning.

Mi-Hye Kim
https://orcid.org/0000-0001-5859-5471
e-mail : mhkim@cbnu.ac.kr
She received a Ph.D. degree in Science from Chungbuk National Univ. in 2001. She has been a professor at Chungbuk National Univ. since 2004. Her research interests are in the area of Big Data, Functional Games, Ubiquitous Games, Platforms, Fuzzy Measures and Fuzzy Integration, Gesture Recognition.

APPENDIX A

Table A1. Correlation and MC Analysis for KNHANES Dataset

	Features	Tolerance	VIF	p-value
1	Zscore(HE_hcv_n)	0.99	1.01	0.00
2	Zscore(tins)	0.98	1.02	0.04
3	Zscore(HE_HNsAg_NKh	0.98	1.03	**0.65**
4	Zscore(HE_rPLS)	0.97	1.03	0.00
5	Zscore(HE_chol_new)	0.95	1.05	0.03
6	Zscore(incm)	0.89	1.12	0.00
7	Zscore(HE_Bplt)	0.81	1.24	0.00
8	Zscore(HE_WBC)	0.78	1.29	0.00
9	Zscore(occp)	0.77	1.29	0.00
10	Zscore(HE_BMI_New)	0.77	1.30	0.00
11	Zscore(N_CAROT)	0.72	1.38	**0.05**
12	Zscore(N_VITC)	0.72	1.38	0.00
13	Zscore(HE_fh)	0.68	1.47	0.00
14	Zscore(N_RETIN)	0.66	1.51	0.00
15	Zscore(HE_HPdg)	0.66	1.51	0.00
16	Zscore(HE_glu_New)	0.64	1.56	0.00
17	Zscore(HE_DMdg)	0.60	1.67	0.00
18	Zscore(marri_1)	0.57	1.76	0.00
19	Zscore(HE_RBC)	0.53	1.87	0.03
20	Zscore(N_NA)	0.50	2.01	0.00
21	Zscore(edu)	0.47	2.11	0.00
22	Zscore(HE_ast)	0.47	2.12	0.00
23	Zscore(HE_dbp)	0.46	2.16	0.00
24	Zscore(N_FE)	0.43	2.32	**0.06**
25	Zscore(HE_sbp)	0.40	2.50	0.00
26	Zscore(HE_alt)	0.40	2.51	0.00
27	Zscore(N_SUGAR)	0.38	2.63	0.00
28	Zscore(N_WATER)	0.38	2.66	0.00
29	Zscore(N_CHOL)	0.35	2.82	0.00
30	Zscore(N_CA)	0.33	3.04	0.00
31	Zscore(N_NIAC)	0.26	3.81	0.00
32	Zscore(age)	0.26	3.87	0.00
33	Zscore(N_TDF)	0.24	4.20	**0.21**
34	Zscore(N_N3)	0.19	5.17	0.00
35	Zscore(N_K)	0.16	6.33	0.00
36	Zscore(N_CHO)	0.16	6.40	**0.52**
37	Zscore(N_EN)	**0.09**	**11.13**	0.00
38	Zscore(N_PROT)	**0.08**	**11.96**	0.00
39	Zscore(N_PHOS)	**0.06**	**16.09**	0.00
40	Zscore(HE_HPfh1)	**0.06**	**17.96**	0.00
41	Zscore(HE_STRfh1)	**0.05**	**21.54**	0.00
42	Zscore(HE_HLfh1)	**0.03**	**35.80**	0.00
43	Zscore(HE_DMfh1)	**0.02**	**41.73**	0.00
44	Zscore(N_MUFA)	**0.02**	**55.85**	0.00
45	Zscore(N_SFA)	**0.02**	**59.35**	0.00
46	Zscore(HE_IHDfh1)	**0.02**	**63.69**	0.00
47	Zscore(N_N6)	**0.01**	**101.34**	0.00
48	Zscore(HE_HBfh1)	**0.01**	**106.15**	0.00
49	Zscore(HE_THfh1)	**0.01**	**129.52**	0.00
50	Zscore(N_PUFA)	**0.01**	**148.54**	0.00
51	Zscore(N_FAT)	**0.00**	**307.47**	0.00
52	Zscore(HE_mens)	**0.00**	**426.50**	0.00
53	Zscore(HE_prg)	**0.00**	**4575.05**	0.00
54	Zscore(sex)	**0.00**	**5111.12**	0.00
55	Zscore(N_VA_RAE)	-	-	0.00
56	Zscore(N_VA)	-	-	0.00