국민건강영양조사 자료를 이용한 만성신장질환 분류기법 연구

이홍기[O], 명성민[*]
[O]중원대학교 경영학과,
[*]중원대학교 보건행정학과
e-mail: it2020@jwu.ac.kr[O], smmyoung@jwu.ac.kr[*]

The Study of Chronic Kidney Disease Classification using KHANES data

Hong-Ki Lee[O], Sungmin Myoung[*]
[O]Dept. of Management, Jungwon University,
[*]Dept. of Health Administration, Jungwon University

● 요 약 ●

Data mining is known useful in medical area when no availability of evidence favoring a particular treatment option is found. Huge volume of structured/unstructured data is collected by the healthcare field in order to find unknown information or knowledge for effective diagnosis and clinical decision making. The data of 5,179 records considered for analysis has been collected from Korean National Health and Nutrition Examination Survey(KHANES) during 2-years. Data splitting, referred as the training and test sets, was applied to predict to fit the model. We analyzed to predict chronic kidney disease (CKD) using data mining method such as naive Bayes, logistic regression, CART and artificial neural network(ANN). This result present to select significant features and data mining techniques for the lifestyle factors related CKD.

키워드: Data Mining, Classfication, Chronic Kidney Disease, Naive Bayes, CART, Artificial Neural Network

I. Introduction

Data Mining is one of the most encourging areas of research with the purpose of finding useful information from large volume data sets. In particular, this method is known useful in medical field when no availability of evidence favoring a specific treatment is searched.

Chronic kidney disease(CKD) is associated with end stage renal disease, as well as cardiovascular morbidity and motality. The prevalence of CKD has rapidly increased with obesity epidemmic. Cross sectional and cohort studies showed that an increased body mass index was significantly associated with the development of CKD and deterioration of renal function.

The Korean National Health and Nutrition Examination Survey(KHANES) is a national surveilance system that has been assessing the health and nutritional status of Koreans since 1998 by the Korea Centers for Disease Control and Prevention(KCDC). This survey is cross-sectional study which is covered approximately 10,000 individuals each year and includes socioeconomic status, health-related behaviors, previous and current diseases, nutrition survey, among others.

The purpose of this study is to identify the lifestyle factors related to CKD applying data mining techniques using KHANES data during 2 years(2016~2017). Through this study, we expect that special attention will be paid to CKD in lifestyle strategies based evidence of nationalwide survey data.

II. Methods

The analysis data set has been taken from KHANES during 2 years(2016~2017). The data obtained 6,179 adults after cleaning and removing missing value. Fig 1. shows flowchart of this

research. Chronic kidney disease was defined as dipstick proteinuria or a reduced GFR<60 ml/min/1.73m2). Risk factors known as affect CKD selected 12 attributes which includs age(>60 years), prehypertansion, IFG, BMI(>25), abdominal obesity, smoking(current/heavy/no), alcohol consumption(no/moderate /heavy), inactivity(no, <3times, >3times).

In this analysis, the techniques of classification considered naive Bayes, logistic regression, CART and artificial neural network(ANN). Data analysis were used R-package 3.6.0. Statistical significance was considered for p-values under 0.05.

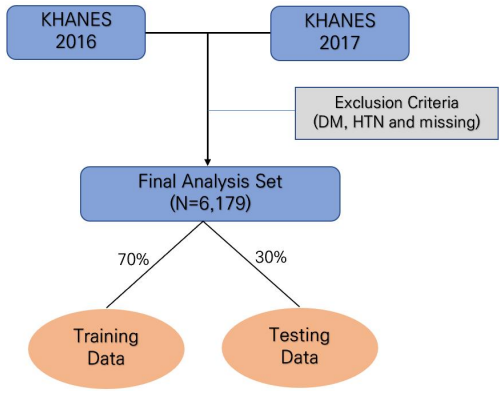

Fig. 1. Research Flowchart

III. Results

Algorithm classification results are shown in Table 1. for training dataset in KHANES, and show that NB(naive Bayes) algorithm gives that the highest classification accuracy 99.77% as compared with all other algorithm results.

Table 1. Summary of Algorithms classification outputs for classifying the CKD – Training Dataset

Measure	NB	Logistic	CART	ANN
Accuaray	99.77%	99.56%	99.01%	98.78%
TP	21.76%	19.55%	17.82%	19.34%
TN	78.01%	80.01%	81.19%	79.44%
FP	0.23%	0.14%	0.44%	0.72%
FN	0.00%	0.30%	0.55%	0.50%
Sensitivity	98.73%	98.49%	98.78%	100%
Specificity	96.55%	98.16%	99.22%	100%

In table 2., algorithm classification present about test data set in KHANES. In case of test dataset, NB method also provides a highest classfication accuracy 98.15% as compared with all other algorithm results.

Table 2. Summary of Algorithms classification outputs for classifying the CKD – Training Dataset

Measure	NB	Logistic	CART	ANN
Accuaray	98.18%	97.16%	97.22%	93.95%
TP	16.24	15.79	14.11	15.32
TN	81.94	81.37	83.11	78.63
FP	0.84	0.87	0.87	2.86
FN	0.98	1.97	0.95	3.19
Sensitivity	98.44%	88.14%	92.51%	89.43%
Specificity	100%	100%	98.19%	98.39%

IV. Conclusions

Our research applied data mining techniques for the lifestyle factors related CKD in KHANES dataset. The result of four data mining algorithms have been compared to define the most accurate algorithm results in predict the CKD. In this results, naive Bayes algorithm recommends the best algorithms. Through this study, we expect that special attention will be paid to CKD in lifestyle strategies based evidence of nationalwide survey data. Also, it can be expected to provide the basis of appropriate diagnosis and treatment to physicians.

ACKNOWLEDGEMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (NRF-2019R1F1A1060865).

REFERENCES

[1] E. A. Rady and A. S. Anwar, "Prediction of kidney disease stages using data mining algorithms," Informatics in Medicine Unlocked, Vol. 15, 100178, 2019.

[2] J. Lee and H. Kim, "Identification of major risk factors associated with respiractory diseases by data mining," Journal of The Korean Data and Information Science Society, Vol. 2, pp. 373-384, 2014.

[3] Y. Kim, D. Chang, S. Kim, I. Park, and S. Kang, "Study on factors of management for diabetes mellitus using data mining," Journal of The Korean Academia-Industrial cooperation Society, Vol. 10, pp. 1100-1108, 2009.