The Study of Chronic Kidney Disease Classification using KHANES data

Hong-Ki Lee, Sungmin Myoung

Dept. of Management, Jungwon University,
Dept. of Health Administration, Jungwon University

I. Introduction

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, Classification, Chronic Kidney Disease, Naive Bayes, CART, Artificial Neural Network

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 nationwide 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), prehypertension, 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 naïve 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.

In this analysis, the techniques of classification considered naïve 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.

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

<table>
<thead>
<tr>
<th>Measure</th>
<th>NB</th>
<th>Logistic</th>
<th>CART</th>
<th>ANN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>98.18%</td>
<td>97.16%</td>
<td>97.22%</td>
<td>93.95%</td>
</tr>
<tr>
<td>TP</td>
<td>16.24</td>
<td>15.79</td>
<td>14.11</td>
<td>15.32</td>
</tr>
<tr>
<td>TN</td>
<td>81.94</td>
<td>81.37</td>
<td>83.11</td>
<td>78.63</td>
</tr>
<tr>
<td>FP</td>
<td>0.84</td>
<td>0.87</td>
<td>0.87</td>
<td>2.86</td>
</tr>
<tr>
<td>FN</td>
<td>0.98</td>
<td>1.97</td>
<td>0.95</td>
<td>3.19</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>98.44%</td>
<td>88.14%</td>
<td>92.51%</td>
<td>89.43%</td>
</tr>
<tr>
<td>Specificity</td>
<td>100%</td>
<td>100%</td>
<td>98.19%</td>
<td>98.39%</td>
</tr>
</tbody>
</table>

In table 2., algorithm classification present about test data set in KHANES. In case of test dataset, NB(naive Bayes) algorithm gives that the highest classification accuracy 98.15% as compared with all other algorithm results.

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, naïve 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 nationwide 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