Advanced Science and Technology Letters
Vol.65 (Database 2014), pp.54-56
http://dx.doi.org/10.14257/astl.2014.65.13

Analysis of Smoking in Korean Adolescents Based on Data Mining Techniques

Haewon Byeon[1,2*] , Raeho Lee[3*]

[1] Dept. of Speech Language Pathology & Audiology, Nambu University, Gwangju, Korea
[2] Speech-Language Pathology Center, Nambu University, Gwangju, South Korea
[3] Dept. of Korean Language Education, Nambu University, Gwangju, Korea
byeon@nambu.ac.kr

Abstract. The purpose of this study was to analyze the risk factors of adolescent smoking using data mining techniques. Data were from the 2012 Korea Youth Risk Behavior Web-based Survey. Subjects were 36,889 high school students (18,938 men, 17,951 women). A prediction model was developed by the use of a Classification and Regression Tree Algorithm. In the decision tree analysis, academic performance, suicidal ideation, type of school they attend and indirect smoking experience were significantly associated with adolescent smoking.

Keywords: data mining, adolescents, smoking, risk factors, decision tree

1 Introduction

While smoking rate of the population over the age of 40 has been on the decrease for the last 10 years due to the national no-smoking policy, smoking population in adolescents are still on the increase [1].

Especially, surveys show that smoking rate of High School male students stays around 30%, the highest rate in Asia, which is 3 times more than that of Japan (8%) and whopping 10 times more than Singapore (3%) [1]. This difference in smoking trend between adolescents and adults implies the possibility that there might be difference in factors related to smoking between them.

As numerous studies so far conducted on adolescent smoking have only remained in fact-finding research or exploration on individual risk factors [6-8], the application of their prediction models to population group with complex and various risk factors cannot help but have limitations.

The purpose of this study was to analyze the risk factors of adolescent smoking using data mining techniques.

[*] Co-corresponding author

ISSN: 2287-1233 ASTL

2 Methods

2.1 Study Population

The source of data from 2012 Korea Youth Risk Behavior Web-based Survey (KYRBS) conducted by Ministry for Health, Welfare and Family Affairs, Korea Center for Disease Control and Prevention and Ministry of Education, Science and Technology on middle and high school students nationwide [2]. This study selected 36,889 high school students (18,938 males and 17,951 females) as its subjects of analysis.

2.2 Measurement

Explanatory variables included gender, home economics, current drinking, passive smoking, Body Mass Index (BMI, kg/m^2), weight perception, experience of depression, suicidal thinking experience, subjective level of happiness, type of school (academic, vocational), academic achievement, and scale of residing city.

2.3 Statistical analysis

When the related factors of smoking were identified in the chi-square test, the related factors of smoking were statistically classified and a prediction model was established, using decision tree.

The Classification and Regression Tree (CART) algorithm was used as a method to predict the related factors in the decision tree model [3]. Measuring impurity with Gini Index [4], CART is an algorithm that performs the binary split, where only 2 child nodes are formed from the parent node.

$$G = \sum_{j=1}^{c} P(i)(1-P(j)) = 1 - \sum_{j=1}^{c} P(j)^2 = 1 - \sum_{j=1}^{c} (n_j/n)$$

In the CART, the Alpha value for the criteria of splitting and merging was set at 0.05. The number of parent nodes was 1,000 and that of child nodes was 500, and the number of branches was limited to 4. The validity of the model was first tested using a misclassification table, and then the risks of the model were compared using the 10-fold cross-validation.

3 Results

The result shows that the classification variables that affect significantly include current drinking, gender, academic achievement, suicidal thinking experience, type of school, and passive smoking. The most primary prediction factor was current drinking (Fig. 1). Second, gender was the related classification variable. The third related classification variable were academic achievement and suicidal thinking experience. Lastly, type of school and passive smoking were the related classification variable.

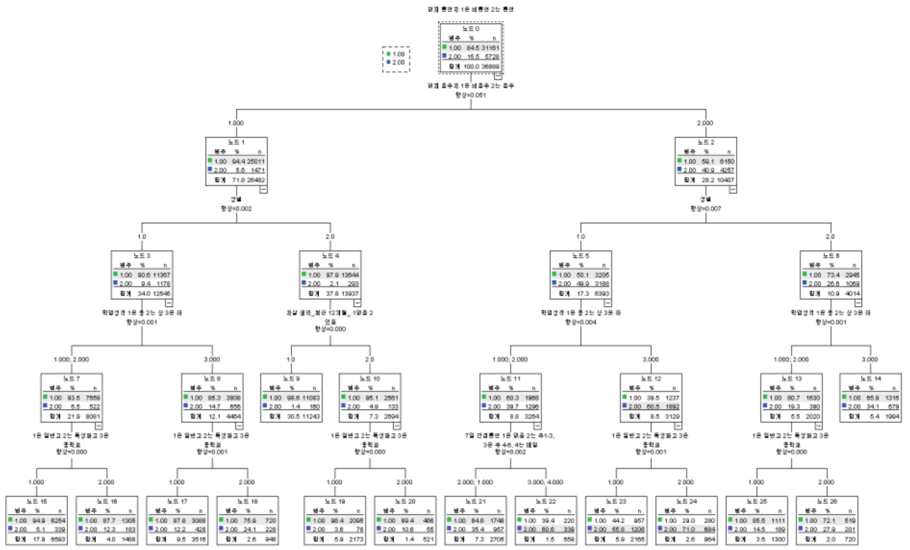

Fig. 1. Predictor model by CART algorithm for Smoking

4 Conclusion

In the decision tree analysis, academic performance, suicidal ideation, type of school they attend and indirect smoking experience were significantly associated with adolescent smoking.

References

1. The Korean Association of Smoking and Health.: 2008 Smoking Prevalence of the Korean Middle School and High School Students. The Korean Association of Smoking and Health. Seoul. (2008)
2. Ministry of Health and Welfare.: Korea Youth Risk Behavior Web-based Survey 2012. Ministry of Health and Welfare. Seoul. (2013)
3. Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J.: Classification and Regression Trees. CA: Wadsworth. Belmont. (1984)
4. Tan, P., Steinbach, M., Kumar, V.: Introduction to data mining. Addison Wesley. Boston. (2006)