An Application of Machine Learning for the Identification of Adolescent Smoking Risk Factors

Sophia J. Chung, PhD, MSN, RN[1,] & Young Ji Lee, PhD, MS, RN[2,3]

[1]Department of Nursing, University of Ulsan, Ulsan, South Korea; [2]School of Nursing;
[3]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA

Aims

To identify the predictors of adolescents smoking behavior in South Korea using a machine-learning approach

Introduction

- Smoking is known as a modifiable risk behavior that causes various health problems including cancer
- Globally, adolescent smoking is likely to persist through adulthood
- In South Korea, smoking behaviors among Korean adolescents remains a significant social problem
- For an effective intervention, the factors that underlie and influence the behavior of smoking should be identified.
- Machine learning is an approach that is well suited to reveal patterns of information in large, complex datasets that are useful in predicting outcomes (Chekroud, 2016)

Method

- Data
  - ➢ The 2015 Korean Youth Risk Behaviors Web-based Survey (KYRBS) was used.
  - ➢ Data completed items concerning smoking were used.
  - ➢ Therefore, data from 5,123 among 68,043 students in grades 7 through 12 were used.
- Machine-learning approach
  - ➢ Machine-learning pipeline developed by Fayyad (1996) and Yoon (2015) was applied.
  - ➢ Clinically meaningful features based on the conceptual framework for adolescent risk behaviors (Jessor, 1991) were selected for analysis.
  - ➢ This process was expected to reduce the “curse of dimensionality" (a high number of inter-related variables in large dataset interfere with the accuracy of the machine-learning model)
  - ➢ Three machine learning algorithms embedded in Weka (i.e., J48, Naïve Bayes, and Logistic Regression) were applied.

Framework

Adolescent smoking behavior

Biology /Genetics	Social environment	Perceived environment	Personality	Behaviors
• Family history of smoking	• SES • Exposure to tobacco advertising	• Best friends’ or teachers’ smoking behavior	• Depression • Suicidal ideation	• Alcohol consumption
• School achievement	• Exposure to anti-smoking advertising campaigns	• Smoking cessation education in school	• Perceived health status • Perceived stress level	• Physical activity • Nutrition

Results

- Majority of the adolescents in this study were male and high school students (80% and 78%, respectively)
- The Logistic Regression algorithm demonstrated the highest level of accuracy (84.0%, F-measure = 0.795).
- Grade (-0.06) and alcohol consumption (-0.56) were the top two features with the highest coefficients.
  - ➢ Middle school students and students who had never drank alcohol were highly associated with the behavior of smoking

Conclusions

- This study demonstrates the behavioral predicators associated with smoking using the KYRBY.
- The results were inconsistent with the previous studies.
- Further study with association between smoking behaviors and alcohol consumption among Korean adolescent is needed

References

- Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., ... & Corlett, P. R. (2016). Cross-trial prediction of treatment outcome in depression: a machine learning approach. *The Lancet Psychiatry, 3(3),* 243-250.
- Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996, August). *Knowledge Discovery and Data Mining: Towards a Unifying Framework*. In KDD (Vol. 96, pp. 82-88).
- Jessor, R. (1991). Risk behavior in adolescence: A psychosocial framework for understanding and action. *Journal of adolescent Health, 12(8),* 597-605.
- Yoon, S., Suero-Tejeda, N., & Bakken, S. (2015). A Data Mining Approach for Examining Predictors of Physical Activity Among Urban Older Adults. *Journal of Gerontological Nursing.*

Acknowledgments Funded by the University Of Ulsan.