The 4th Iranian Conference on Systems Biology

Perdition of Smoking in Young Adults Based on Machine Learning Methods: A System Medicine Approach

Elahe Mousavi[1], Hamidreza Roohafza[2], Mohammadreza Sehhati[3], Ahmad Vaez*[4]

[1]Student Research Committee, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan, Iran.;
[2]Cardiac Rehabilitation Research Centre, Cardiovascular Research Institute, Isfahan University of Medical Sciences, Isfahan, Iran;
[3]Department of Bioelectric and Biomedical Engineering, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan, Iran;
[4]Department of Bioinformatics, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan, Iran
*a.vaez@umcg.nl

Abstract

Tobacco use is one of the main leading cause of preventable death. Numerous studies have shown that intervention to postpone or prevent tobacco use can be an effective strategy to prevent smoking. Considering the reduced onset age of smoking, this study focused on predicting the usage status of teenage students for further prevention. In this study, we propose a machine learning framework for automatic classification of students to smoker and non-smoker based on questionnaire data. The main set of variables are including psychological (depression and self-efficacy), family, social, attitudinal and belief factors and school policy toward smoking. The results of specificity and negative predictive value of 93% and 98% respectively, show the high performance of AdaBoost classifier in predicting and classifying students as smoker or non-smoker. At the next step, using randomized lasso feature selection, the most effective variables for classification were introduced.

Materials and Methods

This study followed two cross-sectional study which were investigated in 2010 and 2015 among high school students in Isfahan province. The self-administered questionnaire used for gathering all data includes 108 questions about psychological (depression and self-efficacy), family, social, attitudinal and belief factors as well as school policy toward smoking (cigarettes and hookah). This study consist of two main stages. At the first stage of the experiments, all the questions were inserted in the study and used as the input features of five classifiers including Logistic Regression Classifier (LR), XGBoost, Support Vector Machine (SVM), Adaboost (AB) and Linear Discriminant Analysis (LDA). At the second stage, Randomized Lasso Feature Selection (RLFS) method has used to introduce the best predictive features in identifying adolescents' smoking status.

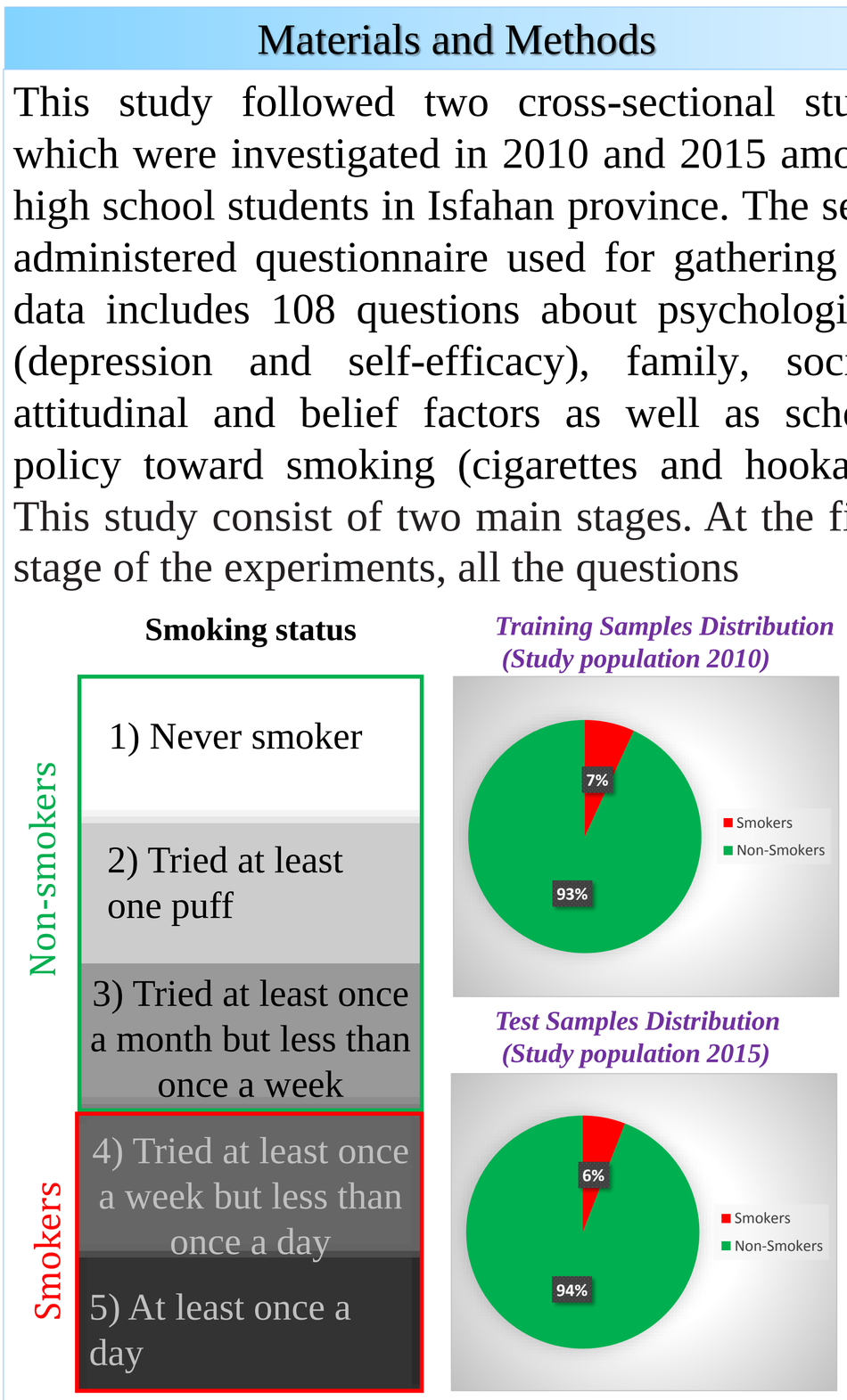

Figure1. Distribution of study population.

Results and Discussion

As a comparison between the results of different classifiers, Area Under Curve (AUC), specificity, sensitivity, Positive Predictive Value (PPV), Negative predictive Value (NPV) of different classifiers ***based on all questions*** of questioner are reported in the Figure 2. The results of different classifiers ***based on the selected features*** are also summarized in Figure 3. A comparison between results shows that it could be possible to obtain high classification rates with only a few selected features (13 questions) from the entire dataset. The high rated features selected by RLFS are: 1) peer influence: “How many of your friends smoke?” 2) Their expectation about smoking in the future. 3) School smoking: “Do teachers smoke at your school?” 4) Household smoking behavior: “Does your brother/sister smoke?” and how is the reaction of your parents toward your smoking?” 5) Their attitude towards smoking.

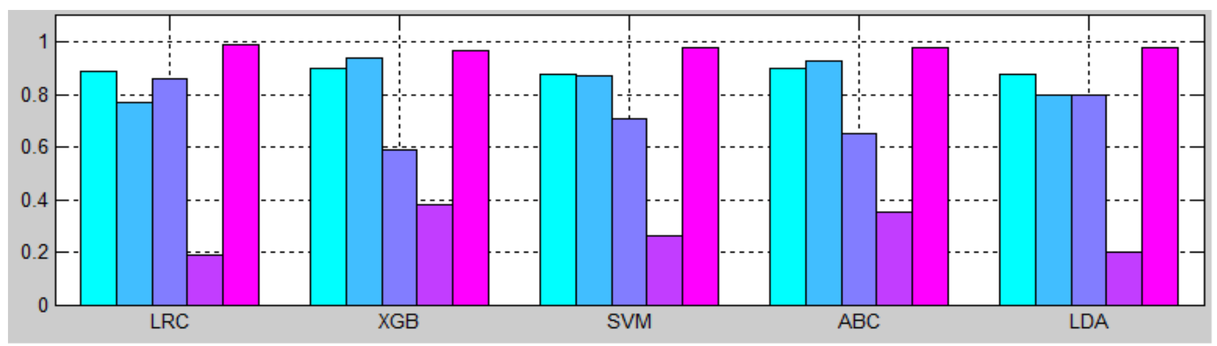

Figure2. Classification results of smoker and non-smoker adolescents based on total questions (108 questions) by different machine learning classifiers.

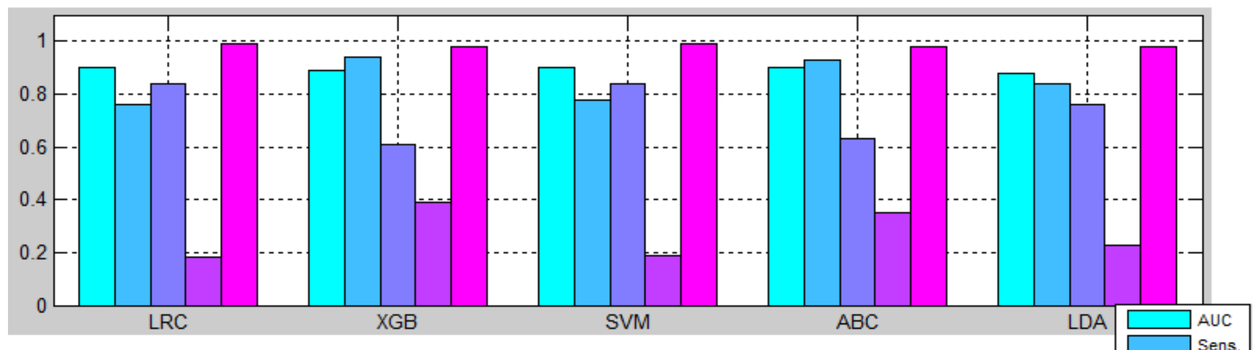

Figure3. Classification results of smoker and non-smoker adolescents based on selected questions (13 questions) by RLFS and different machine learning classifiers.

Conclusion

Systems thinking, as a paradigm for systems medicine, has three major components: Identifying all the players in that system, detecting how the players relate to each other, and understanding the impact of those relationships on each other. We used machine learning as a tools for systems thinking to address how to prevent smoking considering different players from different psycho-socio-cultural layers of human system.

This study shows that machine learning classifiers can help to identify the smoking status and finding the best discriminating features.