Original Article

Machine Learning Helps in Prediction of Tobacco Smoking in Adolescents

Abstract

Background: Considering the increasing prevalence of adolescent smoking in recent years, this study proposes a machine learning (ML) approach for distinguishing adolescents who are prone to start smoking and those who do not directly confess to smoking. **Methods:** We used two repeated measures cross-sectional studies, including data from 7940 individuals as distinct training and test datasets. Utilizing the randomized least absolute shrinkage and selector operator (LASSO), the most influential factors were selected. We then investigated the performance of different ML approaches for the automatic classification of students into smoker/nonsmoker and low-risk/high-risk categories. **Results:** Randomized LASSO feature selection prioritized 15 factors, including peer influence, risky behaviors, attitude and school policy toward smoking, family factors, depression, and sex as the most influential factors in smoking. Applying different ML approaches to the three study plans yielded an AUC of up to 0.92, sensitivity of up to 0.88, PPV of up to 0.72, specificity of up to 0.98, and NPV of up to 0.99. **Conclusions:** The results showed the capability of our ML approach to distinguish between classes of smokers and nonsmokers. This model can be used as a brief screening tool for automated prediction of individuals susceptible to smoking for more precise preventive intervention plans focusing on adolescents.

Keywords: *Adolescent, classification, machine learning, prediction, tobacco*

Hamidreza Roohafza*, Elahe Mousavi[1]*, Razieh Omidi[2], Masoumeh Sadeghi[3], Mohammadreza Sehhati[4]*, Ahmad Vaez[4,5]*

Isfahan Cardiovascular Research Centre, Cardiovascular Research Institute, Isfahan University of Medical Sciences, Isfahan, Iran, [1]Department of Bioelectrics and Biomedical Engineering, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan, Iran, [2]Isfahan Province Health Centre, Isfahan University of Medical Sciences, Isfahan, Iran, [3]Cardiac Rehabilitation Research Centre, Cardiovascular Research Institute, Isfahan University of Medical Sciences, Isfahan, Iran, [4]Department of Bioinformatics, Isfahan University of Medical Sciences, Isfahan, Iran, [5]Department of Epidemiology, University of Groningen, University Medical Centre Groningen, Groningen, The Netherlands

**Contributed equally*

Address for correspondence:
Dr. Ahmad Vaez,
Department of Bioinformatics and System Biology, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Po Box 8174673461, Hezar Jerib Street, Isfahan, Iran.
E-mail: a.vaez@umcg.nl

Introduction

Tobacco use is among the public health threats worldwide. Based on the World Health Organization (WHO) report, 22% of the world's population aged +15 are tobacco consumers. This preventable problem causes six million deaths each year and despite global efforts towards building awareness of smoking risks, eight million annual deaths are predicted by 2030.[1-3]

Most tobacco consumers start smoking during adolescence and before the age of 18.[4] The highest rates of adolescent smoking occur during the transition from middle to high school when particular physical, cognitive, and emotional changes happen[5] and the development of detrimental behavior in this period could contribute to serious health issues and unhealthy lifetime patterns.[6] Two major groups of influencing factors on smoking trends in adolescents are individual and environmental factors. The first group includes depressive symptoms, lack of self-efficacy, attitude toward smoking, and risky behavior. The second group includes peer influence, familial, and school factors.[7,8]

Two steps of providing effective preventive intervention plans aimed at adolescents are (i) detecting the main discriminative factors, and (ii) screening of the tobacco users. Screening is particularly important as the identified individuals may receive well-timed interventions. Brief screening tools with the ability to efficiently identify tobacco or other substance use disorders have been developed.[9-11] However, these tools are mostly based on direct questions about the status of usage and are not based on individual characteristics and risk levels. Hence, the current brief screening tools cannot appropriately guide clinical interventions.

To identify individuals at risk of smoking initiation, the "susceptibility to smoking index" has been introduced. Susceptibility to smoking is based on three questions: "Do you think that in the future you might experiment with cigarettes?"; "At any time during the next year do you think you will smoke a cigarette?" and, "If one of your best friends were to offer you a cigarette, would you smoke it?".[12] It was shown that this index only identifies one-third of future smokers. Furthermore, adding curiosity to the original susceptibility to the

How to cite this article: Roohafza H, Mousavi E, Omidi R, Sadeghi M, Sehhati M, Vaez A. Machine learning helps in prediction of tobacco smoking in adolescents. Int J Prev Med 2025;16:27.

Access this article online

Website:
www.ijpvmjournal.net/www.ijpm.ir

DOI:
10.4103/ijpvm.ijpvm_306_23

Quick Response Code:

Dr. Mohammadreza Sehhati,
Department of Bioinformatics and System Biology, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences,
Po Box 8174673461, Hezar Jerib Street, Isfahan, Iran.
E-mail: mr.sehhati@amt.mui.ac.ir

smoking index could increase the sensitivity of the index. However, this index lacks well-known factors of smoking initiation[12-14] and is based on leading questions that are not appropriate for assessing adolescents who do not want to confess to smoking or identify as future smokers. Another study utilized a bootstrap-enhanced least absolute shrinkage operator to select the most influential factors in initiation usage and ended up with a brief prognostic tool to identify adolescents at risk of transitioning from never to ever smoking.[15] However, the current smokers individuals were completely excluded from the study. This selective strategy could insert bias into the results and hence yield wrong predictions.[16]

The main aim of our study is to provide a brief automatic screening tool with the ability to predict the smoking status (SS) of adolescents based on indirect questions about personal factors. To gain sensitivity and specificity, we took advantage of machine learning (ML) methods not only for automatic feature selection but also for the classification of adolescents into smokers and nonsmokers.

Methods

Datasets of smoker and nonsmoker adolescents

We used the data from the two phases of a repeated measure cross-sectional study entitled "Isfahan Tobacco Use Prevention Program (ITUPP)," which was originally designed to investigate SS in adolescent students. The two separate phases of this study were conducted in Isfahan, Iran in 2010 and 2015. More details about the design and conduct of ITUPP can be found elsewhere.[8] In brief, a multistage stratified cluster random sampling procedure was designed to select the adolescent students. This procedure was conducted using repeated measures for the selection of 5408 and 2682 students in 2010 and 2015, respectively. Data collection was based on a self-administered questionnaire which is comprised of 119 questions. The questionnaire tries to investigate different influential factors on smoking as follows. (I) Social factors include peer smoking and perceived social norms of smoking. (II) Family factors include family SS, parental advice, and the ability of the parents to prevent family conflict. (III) School policy toward smoking control includes teacher SS, rules of school about smoking, and the student's attitude toward school smoking policy. (IV) Psychological factors include refusal skills, self-efficacy, risk-taking, smoking intention, SCL-90 depression subscale, and general self-efficacy scale. (V) Attitudinal and belief factors toward smoking. Furthermore, SS for cigarette and water pipe separately was asked directly and classified into five subgroups as[1] never smoker,[2] tried at least one puff,[3] tried at least once a month but less than once a week,[4] tried at least once a week but less than once a day, and[5] at least once a day.[8,17] Students with no responses to these two questions, were eliminated from the study, and the study population was reduced to 5336 and 2604 individuals for the phases of 2010 and 2015, respectively.[8] The study of ITUPP was approved by ethic committee of the Isfahan University of Medical Science (87139).

Experimental setting

To avoid leakage in the evaluation of any ML method, the training and test datasets should be completely independent. Thus, we considered the independent datasets of 2010 and 2015 of ITUPP as 'training' and 'test' datasets, respectively.

We assessed 119 questions as well as two scores (family conflict and risky behavior scores) for each individual as input features. The questionnaire has also two other questions asking the individuals' current SS including cigarette and water pipe separately. Considering the main aim of this study which is focused on the prediction of SS of adolescents, the five subgroups according to individuals' answers to SS questions (see above) were merged in different ways to create three different binary outcome definitions (OD) and were used as the ground truth for model evaluation [Figure 1]. Hence, our classification task forms a two-class problem. In this way, the intermediate individuals (i.e., occasional users and trainers) which are always difficult to identify, were merged into either smokers or nonsmokers based on three different ODs, as described below, to eventually end up with two classes of SS per OD. We then evaluated different ML methods (classifiers) in the prediction of SS in the three ODs.

In the first OD (OD1), models were constructed based on two classes of 'never smokers' and 'smokers'. Class 1, i.e., 'never-smokers', contains the students using neither cigarettes nor water-pipe (choosing item 1 to question about their current SS). Class 2, i.e., 'smokers', contains individuals either using cigarettes or water pipes (choosing items 4 or 5). In this situation, individuals with other responses (occasional users and trainers) were dropped from the investigated population.

The second OD (OD2) is focused on discriminating the 'low-risk' individuals from 'high-risk' individuals. In this plan individuals responding neither use of cigarettes nor water-pipe (choosing item 1) were considered as class 1, i.e., 'low-risk', whereas others including occasional users, trainers, and smokers were considered as class 2, i.e., 'high-risk' individuals.

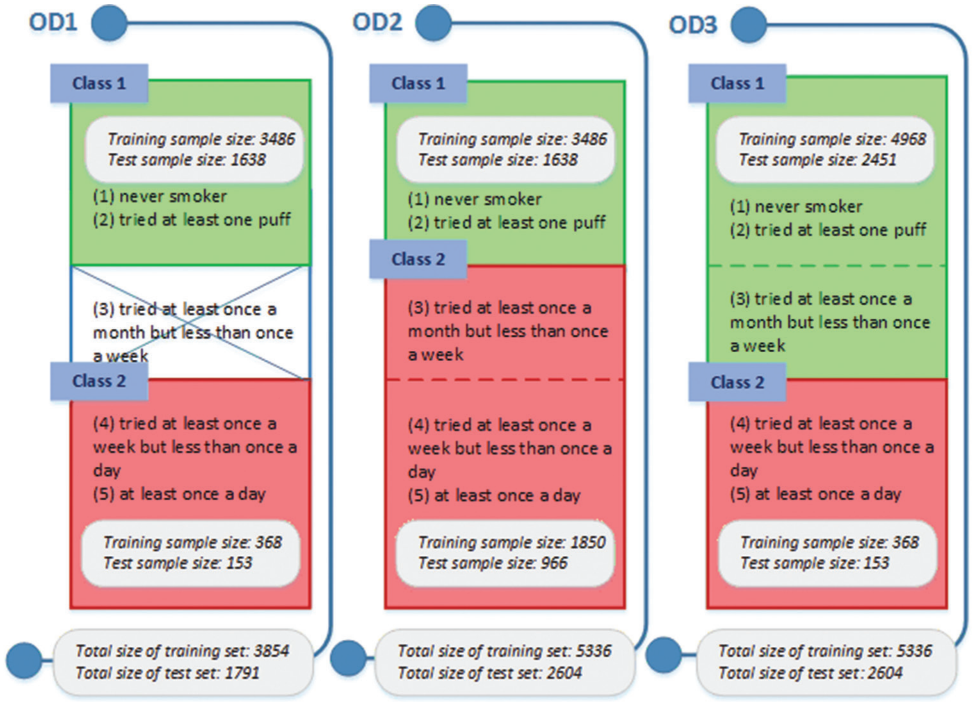

OD1
Class 1
Training sample size: 3486
Test sample size: 1638
(1) never smoker
(2) tried at least one puff
(3) tried at least once a month but less than once a week
Class 2
(4) tried at least once a week but less than once a day
(5) at least once a day
Training sample size: 368
Test sample size: 153
Total size of training set: 3854
Total size of test set: 1791
OD2
Class 1
Training sample size: 3486
Test sample size: 1638
(1) never smoker
(2) tried at least one puff
Class 2
(3) tried at least once a month but less than once a week
(4) tried at least once a week but less than once a day
(5) at least once a day
Training sample size: 1850
Test sample size: 966
Total size of training set: 5336
Total size of test set: 2604
OD3
Class 1
Training sample size: 4968
Test sample size: 2451
(1) never smoker
(2) tried at least one puff
(3) tried at least once a month but less than once a week
Class 2
(4) tried at least once a week but less than once a day
(5) at least once a day
Training sample size: 368
Test sample size: 153
Total size of training set: 5336
Total size of test set: 2604

Figure 1: Description of the three output definition (OD)

The third OD (OD3) is focused on smokers so that the individuals who tried at least once a week but less than once a day, or at least once a day (choosing items 4 or 5) were considered as 'smokers' (class 2) while others including occasional users, trainers, and never-smokers were considered as 'nonsmokers' (class 1).

Selection of influential features

The process of selecting the most significant subset of the existing variables without transforming them leads to an improved feature space that makes the model easier to interpret, reduces overfitting, and enables the algorithms to work faster. Traditional feature selection methods that are based on an evaluation of specific criteria consider each variable independently.[18] In contrast, there are other groups of feature selection methods such as weighted naive Bayes[19] and Least Absolute Shrinkage and Selection Operator (LASSO)[20] which search for an optimal subset of features, within the classifier construction. These selection methods have become popular due to their joint feature selection property.

Here we used a well-known extension of LASSO, named randomized LASSO,[21] as the feature selection step to find the most influential factors. The important characteristic of randomized LASSO is the stability of selection. The main idea of the LASSO is forcing the L1 constraint on the model parameters that can deal with the multi-co-linearity in the data matrix while penalizing the coefficients of the regression variables shrinking some of them to zero. The variables with non-zero coefficients would be selected. The high-level idea of randomized LASSO is to apply the LASSO on various subsets of data and different subsets of features. After repeating this process a number of times, the most frequently selected features would be recognized as the most important features.[21,22] Stability selection results in much less sensitive features to the choice of the regularization.[18]

Classification problem

To achieve the best results, five classifiers including Logistic Regression (LR),[23] Support Vector Machine (SVM),[24] Random Forest Classifier (RFC), Adaptive Boosting Classifier (ABC),[25] and Gradient Boosting Classifier (GBC)[26] were investigated. A brief introduction to the utilized classifiers is provided in the Appendix.

After finding the most influential factors in the prediction of smoking usage using the randomized LASSO feature selection method, the classification of individuals was performed based on the selected features in conjunction with five classifiers. We repeated the same procedure for each of the three above-mentioned ODs. Grid search[27] and cross-validation on the training dataset were then used to set the parameters of each classifier to attain its best performance. The results of the classifiers were evaluated based on five criteria including Area Under the receiver operating characteristic curve (AUC), sensitivity, specificity, Positive Predictive Value (PPV), and Negative Predictive Value (NPV). The flowchart of the proposed method is summarized in Figure 2.

Results

We included 7940 individuals of ITUPP comprising

5336 and 2604 individuals for the phases of 2010 and 2015, respectively.[8] Of all participants, 4002 (50.4%) were girls and 3938 (49.6%) were boys, and 4610 (58%) were high school students and 3330 (42%) were more junior students (so-called guidance school). Details of the study population and distribution of the individuals in the three described plans (ODs) are summarized in Table 1 and Figure 1, respectively.

Feature selection

To prioritize the list of 119 features and find the most influential factors on tobacco usage and end up with a stable set of features, we applied randomized LASSO to create 1000 models based on 75% of randomly selected samples and features. In this step, we used the target groups of OD1 consisting of two completely separate groups, 'never smokers' and 'smokers' to help the algorithm find the more discriminative features. The algorithm could eventually find 15 questions belonging to most of the assessed fields but especially has pointed towards 1) Peer Influence (PI), 2) School Policies toward smoking (SP), 3) Family factors (F), 4) Risky Behaviours (RB), 5) Attitude toward smoking (A), 6) Depression (D) and, 7) Sex (S). The set of 15 selected questions are listed in Table 2. Figure 3 shows the average answers of the participants to the 15 selected questions per two completely distinct groups, 'never smokers' and 'smokers'.

Classification of individuals

As shown in Table 3, in OD1, where the intermediate level of the spectrum of SS is ignored, all of the reported classifiers had a high ability to discriminate 'never smokers' from 'smokers'. All of the classifiers resulted in an AUC of more than 0.90, specificity of more than 0.81, and NPV of more than 0.96.[23] Based on a one-way Analysis of Variance (ANOVA), there is no significant difference between the results of the different investigated classifiers (*P* Value = 0.99). Considering the sensitivity, SVM, LR and, RFC achieved higher values. As shown in Figure 1, the prevalence of class 2 ('smokers') in OD1 is about 10% both in the training and test data. Here we see a trade-off between sensitivity and PPV which can be due to the low prevalence of smokers. Depending on the goal of screening, if PPV is more important than sensitivity, ABC and GBC are the selected methods with PPV of 0.64 and 0.62, respectively. The experiments including training and testing the model, were repeated 10 times for all of the classifiers and ended up with a variance of <0.01 confirming the repeatability of the models. Besides, a Chi-Squared test was applied to all the selected features comparing class 1 and class 2. For all the features there are significant differences between the mean features of the two groups of 'smokers' and 'never smokers' (*P* value <0.001).

Since the inclusion of occasional users and trainers in the test population could change the problem, we also investigated OD2 and OD3 (see Methods section) by inversely including the occasional users and trainers in class 2 (OD2) or class 1 (OD3). In OD2, the AUC of all classifiers is more than 0.81 which shows the high performances of the models. However, the AUC of the models in this OD is about 0.10 less than the AUCs of OD1 which can be due to the inclusion of borderline users, i.e., occasional users and trainers. Based on the second section of Table 3, RFC, LR, and SVM obtained the highest sensitivity of 0.74, 0.73, and 0.72, respectively, and GBC and ABC achieved a specificity of more than

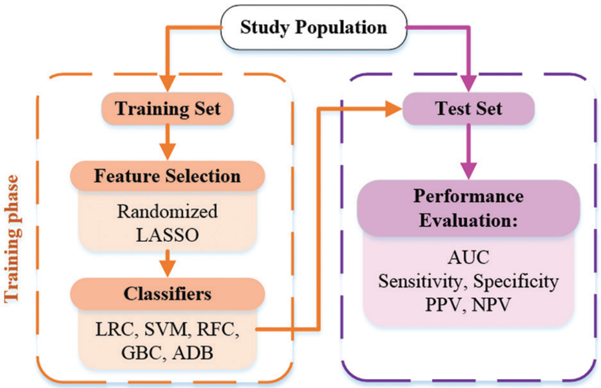
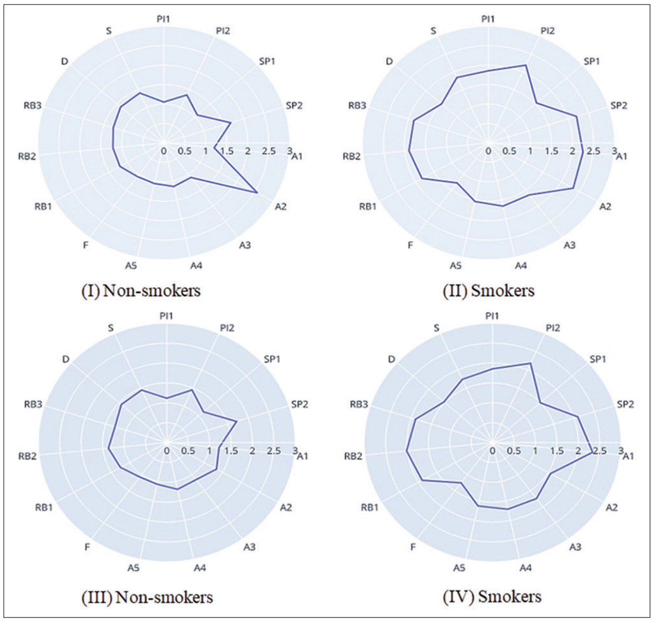


Figure 2: Flowchart of the proposed method. The utilized classifier are Logistic Regression Classifier (LRC), Support Vector Machine (SVM), Random Forest Classifier (RFC), Gradient Boosting classifier (GBC), and AdaBoost Classifier (ADB). The performance evaluation metrics are: Area Under Curve (AUC), Sensitivity, Specificity, Positive Predictive Value (PPV) and Negative Predictive Value (NPV)



(I) Non-smokers

(II) Smokers

(III) Non-smokers

(IV) Smokers

Figure 3: Average of the answers to the 15 selected variables, in training and test data sets. PI, Peer Influences; SP, School Policies toward smoking; A, Attitude toward smoking; F, Family factor; RB, Risky Behaviors; D, Depression; and S, Sex. Figures (I) and (II) are related to the study population of 2010 (Training samples) and, Figures (III) and (IV) are related to the study population of 2015 (Test samples)

Table 1: Distribution of the two study populations

Feature	Subcategory	Study Population of 2010		Study Population of 2015	
		Girl (*n*) (%)	Boy (*n*) (%)	Girl (*n*) (%)	Boy (*n*) (%)
Age (Mean±SD)		15.3+1.78	15.44+1.6	15.32+1.6	15.3+1.65
Educational Level	Guidance school	1204 (22.6)	1198 (22.5)	463 (17.8)	465 (17.8)
	High school	1462 (27.4)	1472 (27.5)	873 (33.5)	803 (30.8)
Residency Area	Urban	2375 (44.5)	2368 (44.4)	1149 (44.3)	1086 (41.4)
	Rural	291 (5.4)	302 (5.7)	187 (7.1)	182 (7.2)

Table 2: The set of 15 selected questions as the most influential factors on tobacco usage

Feature category	Name	Questions	Response categories
Peer Influence	PI1	How many of your friends smoke?	1) None 2) About half 3) Almost all
	PI2	The percentage of your peers you think they have experienced smoking (even a puff).	1) Less than 30% 2) Between 30% -60% 3) More than 60%
School Policies Toward Smoking	SP1	Do any of your teachers smoke?	1) Less than half 2) About half 3) More than half
	SP2	At our school the students themselves work hard to prevent smoking at school.	1) Agree 2) No idea 3) Disagree
Attitude Toward Smoking	A1	Sometimes I feel like I need a cigarette or a puff of water pipe.	1) Disagree 2) No idea 3) Agree
	A2	Smoking also affects the health of non-smokers.	1) Agree 2) No idea 3) Disagree
	A3	I like water pipe because of its smoke.	1) No 2) Yes
	A4	I like water pipe because of its smell.	1) No 2) Yes
	A5	Water pipe is useful for relieving boredom and reducing stress.	1) No 2) Yes
Family factors	F	Does your sibling smoke cigarette or water pipe?	1) No 2) Yes
Risky Behaviors	RB1	It's worth to be in trouble for fun and entertainment.	1) Rarely 2) Sometimes 3) Usually
	RB2	I like to take risks.	1) Rarely 2) Sometimes 3) Usually
	RB3	I enjoy doing things that others believe should not be done	1) Rarely 2) Sometimes 3) Usually
Depression	D	You've felt less energized or less active in the past month.	1) No 2) Yes
Sex	S	What is your sex?	1) Female 2) Male

0.86. The inclusion of occasional users and trainers into this OD increased the prevalence of class 2 to about 36% which caused a sizable increase in PPV to more than 0.62. The Chi-square test of all the features in this OD resulted in a *P* value <0.001 for all of the features except for one of the attitude questions, which asks individuals' opinions about the effects of smoking on the health of non-smokers.

In OD3, the same as OD1 and OD2, high AUC (more than 0.85) achieved for all of the classifiers. As shown in Table 3, SVM, LRC, and RFC reached the

Table 3: The results of classification of adolescents in 3 described plans

OD	Explanation	Classifier	AUC	Sensitivity	Specificity	PPV	NPV
OD1	Class 1: Never smokers Class 2 : Smokers	LR	0.92	0.88	0.81	0.30	0.99
		SVM	0.92	0.88	0.84	0.34	0.99
		RFC	0.93	0.87 (±0.01)	0.84 (±0.01)	0.34 (±0.01)	0.99
		ABC	0.92	0.61	0.97	0.64	0.96
		GBC	0.92	0.59 (±0.01)	0.96 (±0.01)	0.62 (±0.01)	0.96
OD2	Class 1: Never smokers Class 2: Occasional users, trainers, smokers	LR	0.81	0.73	0.73	0.62	0.82
		SVM	0.82	0.72	0.76	0.64	0.82
		RFC	0.81	0.74 (±0.01)	0.73 (±0.01)	0.62 (±0.01)	0.83
		ABC	0.81	0.58	0.87	0.72	0.78
		GBC	0.81	0.58 (±0.01)	0.86 (±0.01)	0.70 (±0.01)	0.78
OD3	Class 1: Never smokers, occasional users, trainers Class 2: Smokers	LR	0.88	0.85	0.74	0.17	0.99
		SVM	0.88	0.86	0.77	0.19	0.99
		RFC	0.88	0.83 (±0.01)	0.79 (±0.01)	0.19 (±0.01)	0.99
		ABC	0.85	0.33	0.98	0.53	0.96
		GBC	0.87	0.41 (±0.02)	0.97 (±0.01)	0.49 (±0.02)	0.96

The results are the mean of 10 repetitions of experiments. Not-mentioned standard deviation are less than 0.001. OD indicates outcome definition; AUC, area under the curve; PPV, positive predictive value; NPV, negative predictive value; LR, logistic regression; SVM, support vector machine; RFC, random forest classifier; ABC, adaptive boosting classifier; and GBC, gradient boosting classifier

highest sensitivity of 0.86, 0.85 and 0.83, respectively. Furthermore, the specificity and NPV of more than 0.74 and 0.98 achieved for the mentioned classifiers. However, the low prevalence of the class 2 (<6%) resulted in low PPV. The Chi-squared test of features was significant (*P* value <0.001) confirming the significant difference between the mean of smoker class and nonsmoker.

Discussion

In this study we took the advantage of ML to automatically predict adolescent's SS using a set of indirect questions. This prediction model can be then used for appropriate intervention in the adolescent populations.

To avoid tiredness of the respondent and hence, imprecise answers, we tried to introduce a brief and comprehensive questionnaire that can be used for the prediction of individuals' SS and identification of high-risk individuals. To find the most influential factors in tobacco use of adolescents, we used the randomized LASSO method and ended up with a 15-item screening questionnaire. These 15 selected features interestingly cover different affective fields on smoking including 2 peer influence items, 2 school policies toward smoking items, 3 risky behaviour items, 5 attitudes toward smoking items, 1 family factor item, and 1 depression item and sex [Table 2, Figure 3]. Several other studies have identified peer influence[14,28-32] and sex[14,15,31-33] as the effective factors for smoking experimentation. Depression,[15,29,30] risky behaviours,[15,34] family factors,[30,32] attitudes[14,29,32], and school policies toward smoking[33,35] have also been shown to be associated with smoking in several studies.

Our model is similar to two other studies that also introduced a risk model for predicting 1-year risk of smoking initiation. Sylvestre *et al.*[15] investigated the effect of alcohol use in smoking initiation and found it as one of the main items of their model. However, as alcohol use is illegal in our study population, this factor was not considered in our study. The second study, Talluri *et al.,*[32] also considered cognitive susceptibility in the model. Although our randomized LASSO found cognitive susceptibility as an influential factor in the identification of susceptible individuals to smoking, but due to the directive nature of this factor, we eliminated it from our final model based on the expert comment. However, if inserting this question into the model, the sensitivity of our model would increase.

Based on the different patterns of classes shown in Figures 3, questions related to peer influence, school policies and risky behaviours toward smoking can be underscored as the more discriminative factors.

To automatically predict adolescent's SS, we examined three distinct ODs, so that (i) in the first OD we overlooked the borderline group of trainers and occasional users, (ii) in the second OD we merged the trainers and occasional users with the high-risk group (class 2), and (iii) in the third OD we considered the borderline group as the nonsmoker group (class 1). Based on our results, the introduced models could reach AUC of more than 0.80 which shows the high ability of introduced classifiers in discriminating low-risk and high-risk individuals or nonsmoker and smoker individuals for further intervention.

Our proposed models could provide high sensitivity, the low prevalence of smoking in the population resulted in lower PPV of the classifiers. However, we think an appropriate population-based intervention program aims towards successful identification of true negative individuals, i.e., nonsmokers, who do not need further intervention. The rest, including true positive and false positive individuals can be then followed-up for further interventions. All these mean that high AUC, high specificity and very high NPV of our models promise successful identification of true negative individuals and hence, further preventive interventions can be appropriately aimed towards susceptible individuals While we used a large validated data set to test and train the models in this work, this data set imposed two constraints on the study, which if removed in future studies could result in more precise and upgraded models. The models' prediction capacity is limited by cross-sectional data sets, and using longitudinal data sets could better confirm the potential of ML applications. Furthermore, the study's nationwide scope could be a limitation for social science research. However, the main aim of this study is focused on the concept prove of ML utilization in the field of prediction of individuals' SS, which might also be applied to the study of other social behaviours.

Conclusions

In this study we investigated two repeated measures cross-sectional studies including 7940 individuals and applied the randomized LASSO method to extract the most influential factors on smoking and eventually introduced a 15-item tobacco screening questionnaire. Then, the application of different ML classifiers in the automatic prediction of adolescents' SS was investigated. All different plans of our study yielded high AUC, high specificity, and very high NPV. Our proposed ML based method provides an automatic screening tool geared toward distinguishing high-risk individuals who need further interventions.

Acknowledgments

We would like to especially thank the staff of the Isfahan Provincial Health Center for their enthusiasm and commitment to conduct the study and all students who participated in this study.

Financial support and sponsorship

The study was funded by Isfahan University of Medical Sciences (Number 87139 and 3981000).

Conflicts of interest

There are no conflicts of interest.

Received: 15 Nov 23 **Accepted:** 11 Sep 24
Published: 24 Apr 25

References

1. WHO Report on the Global Tobacco Epidemic 2019: Offer Help to quit tobacco use. World Health Organization; 2019.
2. Barnes J, McRobbie H, Dong CY, Walker N, Hartmann-Boyce J. Hypnotherapy for smoking cessation. Cochrane Database Syst Revi 2019;2019:CD001008.
3. Perez-Warnisher MT, de Miguel MdPC, Seijo LM. Tobacco use worldwide: Legislative efforts to curb consumption. Ann Glob Health 2018;84:571:9.
4. Warren CW, Riley L, Asma S, Eriksen MP, Green L, Blanton C, *et al*. Tobacco use by youth: A surveillance report from the Global Youth Tobacco Survey project. Bull World Health Organ 2000;78:868-76.
5. Duncan LR, Pearson ES, Maddison R. Smoking prevention in children and adolescents: A systematic review of individualized interventions. Patient Educ Couns 2018;101:375-88.
6. Roohafza H, Heidari K, Alinia T, Omidi R, Sadeghi M, Andalib E, *et al*. Smoking motivators are different among cigarette and waterpipe smokers: The results of ITUPP. J Epidemiol Glob Health 2015;5:249-58.
7. Gritz ER, Prokhorov AV, Hudmon KS, Jones MM, Rosenblum C, Chang C-C, *et al*. Predictors of susceptibility to smoking and ever smoking: A longitudinal study in a triethnic sample of adolescents. Nicotine Tob Res 2003;5:493-506.
8. Roohafza H, Heidari K, Omidi R, Alinia T, Rajabi F, Bagheri S, *et al*. Methodology of Isfahan tobacco use prevention program: First phase. Adv Prev Med 2013;2013:182170.
9. Kelly SM, Gryczynski J, Mitchell SG, Kirk A, O'Grady KE, Schwartz RP. Validity of brief screening instrument for adolescent tobacco, alcohol, and drug use. Pediatrics 2014;133:819-26.
10. McNeely J, Strauss SM, Saitz R, Cleland CM, Palamar JJ, Rotrosen J, *et al*. A brief patient self-administered substance use screening tool for primary care: Two-site validation study of the Substance Use Brief Screen (SUBS). Am J Med 2015;12:784.e9-19.
11. McNeely J, Wu L-T, Subramaniam G, Sharma G, Cathers LA, Svikis D, *et al*. Performance of the tobacco, alcohol, prescription medication, and other substance use (TAPS) tool for substance use screening in primary care patients. Ann Intern Med 2016;165:690-9.
12. Strong DR, Hartman SJ, Nodora J, Messer K, James L, White M, *et al*. Predictive validity of the expanded susceptibility to smoke index. Nicotine Tob Res 2015;17:862-9.
13. Carey FR, Wilkinson AV, Harrell MB, Cohn EA, Perry CL. Measurement and predictive value of susceptibility to cigarettes, e-cigarettes, cigars, and hookah among Texas adolescents. Addict Behav Rep 2018;8:95-101.
14. Pierce JP, Choi WS, Gilpin EA, Farkas AJ, Merritt RK. Validation of susceptibility as a predictor of which adolescents take up smoking in the United States. Health Psychol 1996;15:355.
15. Sylvestre M-P, Hanusaik N, Berger D, Dugas E, Pbert L, Winickoff J, *et al*. A tool to identify adolescents at risk of cigarette smoking initiation. Pediatrics 2018;142:e20173701.
16. Klein JD. Screening tools for who will start smoking and the future of clinical prediction. Pediatrics 2018;142:e20182298.
17. Roohafza H, Omidi R, Alinia T, Heidari K, Mohammad-Shafiee G, Jaberifar M, *et al*. Factors associated with smoking contemplation and maintenance among Iranian adolescents. East Mediterr Health J 2018;24:714-21.
18. Kamkar I, Gupta SK, Phung D, Venkatesh S. Stable feature selection for clinical prediction: Exploiting ICD tree structure using Tree-Lasso. J Biomed Inform 2015;53:277-90.
19. Duda RO, Hart PE, Stork DG. Pattern Classification. John Wiley and Sons; 2012.
20. Tibshirani R. Regression shrinkage and selection via the lasso. J R Statist Soc B 1996;58:267-88.
21. Meinshausen N, Bühlmann P. Stability selection. J R Statist Soc B 2010;72:417-73.
22. Fonti V, Belitser E. Feature selection using lasso. VU Amsterdam Research Paper in Business Analytics, 2017. p. 1-25.

23. Hosmer DW Jr, Lemeshow S, Sturdivant R×. Applied Logistic Regression. John Wiley & Sons; 2013.
24. Noble WS. What is a support vector machine? Nat Biotechnol 2006;24:1565-7.
25. Schapire RE, Freund Y. Boosting: Foundations and Algorithms. Kybernetes. 2013.
26. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot 2013;7:21.
27. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al*. Scikit-learn: Machine learning in Python. J Mach Learn Res 2011;12:2825-30.
28. Bonilha AG, Ruffino-Netto A, Sicchieri MP, Achcar JA, Rodrigues-Júnior AL, Baddini-Martinez J. Correlates of experimentation with smoking and current cigarette consumption among adolescents. J Bras Pneumol 2014;40:634-42.
29. Dahlui M, Jahan NK, Majid H, Jalaludin M, Murray L, Cantwell M, *et al*. Risk and protective factors for cigarette use in young adolescents in a school setting: What could be done better? PloS One 2015;10:e0129628.
30. Tercyak KP. Psychosocial risk factors for tobacco use among adolescents with asthma. J Pediatr Psychol 2003;28:495-504.
31. Abroms L, Simons-Morton B, Haynie DL, Chen R. Psychosocial predictors of smoking trajectories during middle and high school. Addiction 2005;100:852-61.
32. Talluri R, Wilkinson AV, Spitz MR, Shete S. A risk prediction model for smoking experimentation in Mexican American youth. Cancer Epidemiol Prev Biomarkers 2014;23:2165-74.
33. June KJ, Sohn SY, So AY, Yi GM, Park SH. A study of factors that influence Internet addiction, smoking, and drinking in high school students. J Korean Acad Nurs 2007;37:872-82.
34. Veselska Z, Geckova AM, Orosova O, Gajdosova B, van Dijk JP, Reijneveld SA. Self-esteem and resilience: The connection with risky behavior among adolescents. Addict Behav 2009;34:287-91.
35. Breslau N, Fenn N, Peterson EL. Early smoking initiation and nicotine dependence in a cohort of young adults. Drug Alcohol Depend 1993;33:129-37.

Appendix

A brief introduction to the utilized classifiers

Logistic Regression (LR)

Logistic regression (LR) is one of the first supervised learning type algorithms that due to its efficient, straightforward nature and easy implementation is commonly used for the predicting dichotomous targets. Utilizing a logit function, linear regression could transform to logistic regression.[23]

Support Vector Machine (SVM)

One of the best-known ML methods is SVM which can be used in both regression and classification problems. The main idea of SVM as a classifier is seeking the optimal separating hyper-plane between observations of different classes with the largest possible amount of margin. Utilization of soft margin concept lets the SVM to accept some amount of misclassification to get the separating hyper-plane in case of mixing the class of marginal data. To provide a solution for non-linear separable data, SVM uses the kernel trick that projects the data to a higher-dimensional space. This projection could help to find the separating hyper-plane. Here we used Radial Based Function (RBF) as the kernel of SVM.[24]

Random Forest Classifier (RFC)

The major reasons of difference between the prediction of different classifiers return to bias and variance of the models and one of the best methods to reduce these factors is ensemble methods. An ensemble uses combinations of classifiers to give the final results instead of relying on one single classifier. Two class of ensemble techniques are bagging and boosting. In the bagging technique, aggregation of different classifiers could be performed using methods such as model averaging or voting, while in boosting, the classifiers are not made independently, rather in a sequential manner the misclassified observation by any classifier is learned by the subsequent classifier. RFC is a bagging technique that is flexible and easy to use. For different sets of randomly selected data samples (created by replacement), RFC creates decision trees, gets the prediction from each tree and, find the best solution by means of voting.

Adaptive Boosting Classifier (ABC)

ABC was the first realization of boosting technique with great successes. The main idea in ABC is defining weights for observations and classifiers. At each iteration, it puts more weight on harder instances defined by the previous iteration so that the current classifier be more focused on those instances to get better classification results. Further to weighting the observation, the weight assigned to each classifier is updating iteratively according to its accuracy. Finally, for the new observations, the prediction is done using the voting strategy among the base classifiers.[25]

Gradient Boosting Classifier (GBC)

Gradient boosting classifier is another classifier of boosting family that consecutively fits base learners to reach a more accurate result. In GBC the new base-learners are constructed so that the negative gradient of the loss function respect to the prediction be minimized. By training later models on the gradient of the error with respect to the loss predictions of the previous model, the model has learned to correct the mistakes of the previous model.[26]