Turk J Public Health
Available online at: http://dergipark.org.tr/tjph/
Doi: 10.20518/tjph.1565381

ORIGINAL ARTICLE

Predicting smoking cessation success: a machine learning approach using the global adult tobacco survey

Mouhib Alnoukari[1], Enas Albaghdadi[2], Lujain Shebli[3]

[1]Syrian Private University Faculty of Engineering, Department of Informatics Engineering, Daraa, Syria

[2]Syrian Private University Faculty of Engineering, Department of Informatics Engineering, Daraa, Syria

[3]Syrian Private University Faculty of Engineering, Department of Informatics Engineering, Daraa, Syria

ABSTRACT

Objective: This paper aimed to evaluate the effectiveness of machine learning (ML) algorithms in predicting smoking cessation outcomes using Global adult tobacco surveys (GATS) data. Specifically, we investigated the influence of sociodemographic, behavioral, and environmental factors on smoking cessation success.

Method: GATS data from multiple countries were analyzed using various ML models, including: K-nearest neighbors, decision trees, random forests, neural networks, Naive Bayes and logistic regression. These models were utilized to classify smoking behaviors and predict smoking cessation outcomes.

Results: Logistic regression exhibited the highest accuracy (69.8%) in predicting smoking cessation, surpassing other ML models and emphasizing the impact of sociodemographic, behavioral, and environmental factors on cessation. Additionally, the study highlights the role of education, employment, and the family environment in shaping smoking behaviors and cessation outcomes.

Conclusion: This research underscores the potential of ML to inform effective smoking cessation strategies. By identifying influential factors and predicting cessation outcomes, policymakers can tailor interventions to specific populations and enhance their effectiveness.

Keywords: Smoking cessation, Machine learning, Predictive models, Data analysis, Global adult tobacco survey

Correspondence: Mouhib Alnoukari, Syrian Private University Faculty of Engineering, Department of Informatics Engineering, Daraa, Syria **ORCID/ID:** 0000-0002-3982-2074 **E-mail:** mouhib.alnoukari@gmail.com
Received: 12.10.2024 / **Accepted:** 02.07.2025

Cite This Article: Alnoukari M, Albaghdadi E, Shebli L. Predicting smoking cessation success: a machine learning approach using the global adult tobacco survey. Turk J Public Health 2025;23(2): 144-162.

Turk J Public Health 2025 Open Access http://dergipark.org.tr/tjph/.
This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

INTRODUCTION

Tobacco consumption has long been acknowledged as a persistent global public health issue, significantly contributing to the burden of chronic diseases and premature mortality. Despite numerous initiatives aimed at supporting smoking cessation, long-term success rates have remained suboptimal. Therefore, it is crucial to thoroughly understand the multifactorial determinants of cessation behavior to inform the development of more effective, data-driven interventions that cater to diverse population needs.

Related works

Over the past two decades, machine learning (ML) methods have been increasingly employed to predict smoking-related behaviors. This evolution has seen a shift from using basic classifiers on localized datasets to adopting advanced ensemble techniques and neural network-based models applied to large-scale, multi-country survey data.

In the literature, the capability of ML algorithms utilizing the Preferred Reporting Items for Systematic Reviews is essential for identifying factors associated with smoking cessation and predicting cessation outcomes. Historically, numerous studies have validated the necessity of employing ML algorithms in the field of smoking control[1].

In the study conducted by Lai et al.[2] seven classification algorithms, including Artificial Neural Networks (ANN), Support Vector Machines (SVM), Random Forest (RF), Logistic Regression (LR), Classification and Regression Trees, and Naïve Bayes (NB), were implemented using data collected from a Taiwanese smoking cessation program. Among these algorithms, the ANN achieved the highest predictive accuracy of 64%.[2] Similarly, in the work presented by Rijhwani et al.[3], NB, Sequential Minimal Optimization (SMO), RF, J48, and Decision Stump algorithms were employed on datasets obtained from dental clinic patients in India. In that study, the Decision Stump algorithm reached the highest accuracy of 55.87%.[3]

During the mid-2010s, tree-based models and boosting methods gained prominence. Thangaraju et al.[4] applied the NB, Decision Tree, and J48 algorithms on a dataset of lung cancer patients from the University of California to identify risk factors associated with lung cancer, classify smokers and non-smokers with the disease, and predict lung cancer, with the NB algorithm achieving an accuracy of 83.4%. Kharabsheh et al.[5] utilized various algorithms including SVM, RF, C4.5, NB, SMO, LR, Instance-Based K (IBK), K-star, k-Nearest Neighbors (k-NN), and Decision Tree on data from student smokers in Jordan to predict depression among young student smokers. A prediction accuracy of 97% was achieved using the C4.5 algorithm to identify the characteristic most closely associated with the type of depression, which was age.[5] In Bangladesh, Roy et al.[6] applied K-Nearest Neighbors (k-NN), RF, and SVM algorithms to predict the impact of smoking on the lives of undergraduate students, with k-NN and RF attaining the highest accuracy of 87.93%.

Subsequent research focused on diverse populations and employed more complex techniques. Ismail evaluated the NB, K-Star, Partial Decision Tree (PART), LogitBoost, and Reduce Error Pruning Tree (REPTree) algorithms on Iraqi NYTS data to predict students' smoking behaviors and cessation, as well as the impact of the school environment on these factors, with the K-Star algorithm achieving 91.02% accuracy in behavioral classification, while the LogitBoost algorithm excelled in predicting smoking cessation and analyzing the influence of the school environment.[7] Choi et al.[8] utilized the Least Absolute Shrinkage and Selection Operator (LASSO) and RF (with Relief) algorithms using data from the NYTS, which studies e-cigarette and hookah use in the United States, to predict students' nicotine addiction levels and identify predictive variables contributing to nicotine addiction, with RF outperforming LASSO (73.42% vs. 63.7%). A further study by Kharabsheh

et al.[9] explored the level of nicotine dependence and identified the most influential predictors in Jordanian women using SMO, NB, K-Star, IBK, and C4.5 algorithms, where SMO achieved only 43% accuracy. Davagdorj et al.[10] utilized LR, SVM, RF, and NB algorithms on data from women smokers at Chungbuk University in Korea to predict the success or failure of smoking cessation attempts, with the NB algorithm reaching the highest accuracy of 90.2%.

From 2020 onward, ML applications expanded to include sensor-based and cross-national survey data. Thakur et al.[11] utilized wearable sensor data from a wrist-worn device and employed SVM, k-NN, Decision Tree, RF, and LR algorithms to classify smoking activity amidst daily life confounders in real time, with the SVM algorithm achieving an accuracy of 98.78%. Durmuşoğlu and Çiftçi[12] classified participants' smoking status (smoker, non-smoker) based on their socio-demographic factors using three classification algorithms: k-NN, Multilayer Perceptron (MLP), and C4.5 on GATS data from Türkiye. The results indicated that the C4.5 algorithm attained the highest accuracy (76.977%), and its performance was evaluated in six other European countries.12 Davagdorj et al.[13] employed two types of SMOTE oversampling techniques to balance the classes in Korean health data (KNHANES 2009-2017) for predicting smoking cessation success using Gradient Boosted Trees, RF, Multi-Layer Perceptron Neural Network (MLPNN), LR, and NB algorithms, where the NB algorithm reached an accuracy of 86-87%.[13]

Ensemble methods were also refined. Issabakhsh et al.[14] tested RF, Gradient Boosting Machine, Extreme Gradient Boosting (XGBoost), and Generalized Linear Models algorithms on data from adult smokers to identify factors influencing and predicting smoking cessation in the general population, with the RF algorithm achieving an accuracy of around 76%. Siddiqui et al.[15] employed LR, J48, and ANN algorithms to determine the factors that classify an individual's smoking status as a smoker, ex-smoker, or non-smoker, with the LR algorithm attaining 95.77%. Durmuşoğlu and Kocabey Çiftçi[16] proposed the MLPNN algorithm to classify smoking status based on five traits related to passive smoking among adults in Türkiye. The results indicated that the MLPNN algorithm could accurately classify smoking status with an accuracy of up to 68%. Poynton and McDaniel[17] compared backpropagation neural networks (BPNN) and LR algorithms on NHIS 2000 data to classify smoking cessation status; the BPNN algorithm reached 75.32-77.9%, while the LR algorithm slightly outperformed it.

Recent works have integrated app-based and digital platform data. Wanga et al.[18] applied NB, LR, J48, SVM (with polynomial kernel), and AdaBoost with two weak learners (Decision Stump and J48) algorithms on data from the BecomeAnEX online community to determine individuals' smoking status by evaluating user-generated content, where the combination of AdaBoost and the J48 algorithm achieved 75.9%. Martinović utilized the J48 algorithm to analyze smoking habits and identify the factors influencing smokers' choices between cigarettes and rolling tobacco in Croatia.[19] Perski et al.[20] trained RF, SVM, Punished LR, and Severe Gradient Enhancement algorithms to classify lapses of smokers attempting to quit among UK cessation app users (Smoke Free), reporting accuracies of up to 96%. Finally, in India, Al-ssabbagh et al.[21] analyzed predictors of smoked and smokeless tobacco use using ML models, reporting that RF and XGBoost exceeded 80% in most scenarios.

Research gap and significance

The previous research can be categorized into three main axes: First, many studies have been confined to a specific geographical scale or a particular population category, which adversely affects the ability to generalize the results. Second, the variety of data collection frameworks and processing methods employed in each study complicates objective comparisons between their findings. Third, there is

a lack of in-depth analysis of Global Adult Tobacco Survey (GATS) data at the level of the Arab region, despite its availability in a uniform and comparative format across multiple countries.

This study aims to bridge the research gap through:

- The utilization of GATS data, developed by the World Health Organization (WHO) and Centers for Disease Control (CDC) to gather information on tobacco use among individuals over 15 years old, encompassing 32 countries with significant social and economic diversity.

- The application of six ML algorithms (Logistic Regression, SVM, k-NN, Random Forest, Decision Tree, Neural Networks) to evaluate their predictive performance in classifying results related to smoking.

- The incorporation of socio-demographic, behavioral, and environmental factors in the success of cessation attempts, through a precise analysis of the algorithms' results and their interpretation at a multicultural level.

METHODS

This section outlines the methodology of this study. It is divided into the following sub-sections: data acquisition, data pre-processing, building business intelligence solutions, and implementing ML algorithms. Figure 1 illustrates the entire approach.

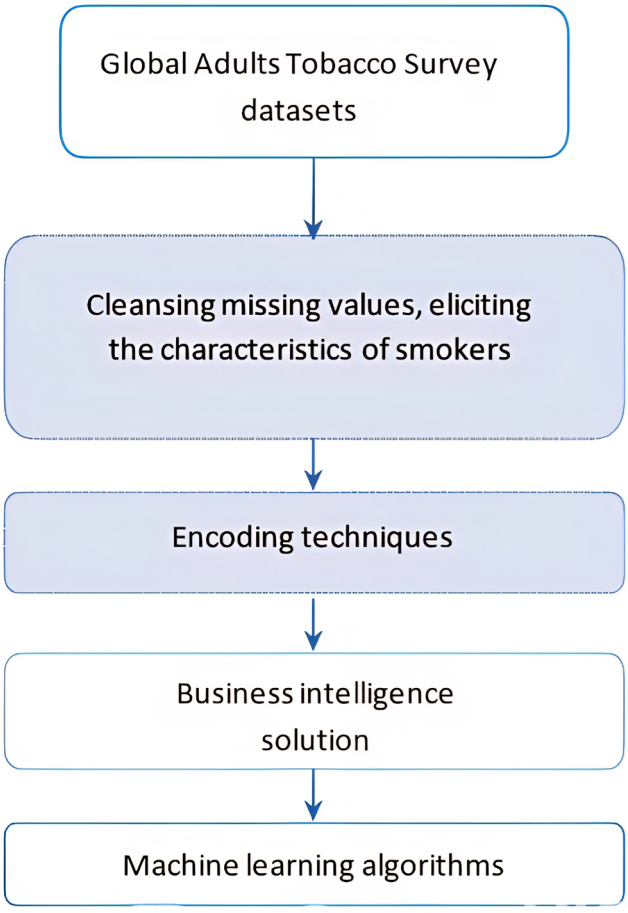


Figure 1. Method workflow

Datasets

This study primarily relies on data from the GATS, a standardized international survey developed in collaboration between the WHO and the CDC and Prevention. GATS is designed to systematically collect nationally representative data on tobacco use and related behaviors among individuals aged 15 years and older.

The original datasets were provided In SPSS and STATA formats and were subsequently converted into CSV files for preprocessing and model development. These datasets are structured based on the available questionnaires.

It is noteworthy that the core individual questionnaire typically consists of around 200 questions, with countries having the option to include additional questions or sections according to their national priorities. As a result, the average number of questions in a GATS questionnaire ranges between 200 and 450 closed-ended questions, depending on country-specific customizations. These questions are distributed across eight thematic sections (labeled A through H) as shown in Table 1, each covering a specific domain related to tobacco-related behaviors or conditions such as smoking history, cessation attempts, exposure to secondhand smoke, media influence, and awareness of tobacco control policies.

Table 1. Surveys sections

Section code	Main content	Description
A	Demographic characteristics	Age, education, current economic status
B	Smoking	Type, frequency, initiation age, quantity (daily/weekly)
C	Smokeless tobacco use	Use of products like gutkha, khaini, snuff, etc.
D	Quit attempts	Tried to quit? Duration? Received help?
E	Secondhand smoke exposure	Exposure to secondhand smoke at home or workplace
F	Expenditure on tobacco	How much is spent on tobacco
G	Media exposure	Viewing ads, warning messages
H	Knowledge, attitudes, and health awareness	Awareness of health risks, attitudes toward smoking

Each question in the dataset is associated with a unique alphanumeric identifier (e.g., A04, B07), a full question text, and a list of predefined, numerically encoded responses (e.g., 1= primary school, 2= secondary school, 77= do not know, 99= refused). For example, question (A04) asks: ”What is the highest level of education you have completed?” with responses ranging from 1 (no formal education) to 8 (postgraduate degree).

To select relevant variables for analysis, a thorough review was conducted of the full questionnaire, including all codes, question texts, and answer categories. While minor variations were found across countries, the core structure and content of the questionnaire remained highly consistent, which enabled harmonized comparative analysis across the 32 countries included in this study.

Table 2. WHO surveys

#	Country	Survey year	Data volume (rows)
1	Türkiye	2008	9,030
		2012	9,851
		2016	8,760
2	Vietnam	2010	9,925
		2015	8,996
3	India	2009	69,296
		2016	74,037
4	Uruguay	2009	5,581
		2017	4,966
5	Ukraine	2010	8,158
		2017	8,298
6	Uganda	2013	8,508
7	Thailand	2009	20,566
		2011	20,606
8	Tanzania	2018	4,797
9	Senegal	2015	4,347
10	Russia	2009	11,406
		2016	2,87
11	Romania	2011	4,517
		2018	4,571
12	Qatar	2013	8,398
13	Poland	2009	7,840
14	Pakistan	2014	7,831
15	Philippines	2009	9,701
		2015	11,644
16	Panama	2013	16,962
17	Nigeria	2012	9,765
18	Mexico	2009	13,617
19	Argentina	2015	14,664
		2012	6,645
20	Bangladesh	2009	9,629
		2017	12,783
21	Botswana	2017	4,643
22	Costa Rica	2015	8,607
23	Greece	2013	4,359
24	Ethiopia	2016	10,150
25	Indonesia	2011	8,305
26	Kazakhstan	2014	4,425
		2019	10,677
27	Egypt	2009	20,924
28	Malaysia	2011	4,250
29	Cameroon	2013	5,271
30	Kenya	2014	4,408
31	Brazil	2008	39,425
32	China	2010	13,354
		2018	19,376

This meticulous review and harmonization of the dataset served as a foundational step to ensure high-quality preprocessing, reliable feature selection, and robust ML model development.

The dataset used comprises data for 561,276 individuals from 46 surveys conducted in various years across all 32 participating countries. Table 2 shows the year and volume of surveys for each country.

Data pre-processing

Following the data collection phase, the datasets underwent a comprehensive cleaning process to ensure consistency and reliability for analysis. For instance, a considerable number of missing values were identified in the column related to the question "How old were you when you first started using smoked tobacco?". These missing values were imputed by utilizing two other available variables:

the respondent's current age and the response to "How many years ago did you first start using smoked tobacco daily?".

In order to enhance the analytical value of the data, questions and responses that were not directly useful for modeling were restructured into meaningful variables. This involved deriving smoker-related characteristics-such as environmental, behavioral, and social attributes-by aggregating and interpreting multiple questionnaire items. All variables were categorized into dependent and independent groups (Table 3). The definition, classification, and structure of these variables were based on the methodological framework proposed by Al-ssabbagh et al.[21]

Table 3. Description of the dependent and independent variables

Dependent variables		
Name of the variable	**Categories**	**Definitions**
Quit attempt among current smokers	Yes No	Current smokers tobacco users responded "daily" or "less than daily" to question: "Do you currently smoke tobacco?" Making a quit attempt: Current smokers users who tried to quit during the past 12 months for more than one day. Those attempted to quit responded "Yes" and those did not attempt to quit responded "No" to the question: "During the past 12 months, have you tried to stop smoking tobacco", and the following question was used to calculate if they quit for more than one day: "Thinking about the last time you tried to quit, how long did you stop smoking tobacco?" (if a current smoker responded "Yes" to the question: "During the past 12 months, have you tried to stop smoking tobacco" and "less than one day" to the question "Thinking about the last time you tried to quit, how long did you stop smoking tobacco?", it was categorized as current smokers/smokeless user who did not attempt to quit)
Successful quitting	Yes No	Former smokers tobacco users responded: a) "not at all" to question "Do you currently smoke tobacco? b) "daily" or "less than daily" to question "In the past, have you smoked tobacco on a daily basis, less than daily, or not at all?" Being successful in quitting: Successful quitters responded "more than 6 months and up to 24 months" to the question "How long has it been since you stopped smoking tobacco?"
Independent variables		
Name of the variable	**Categories**	**Remarks if any**
Socio-demographic characteristics		

Table 3 Continued.		
Age	15-24 25-44 45-64 >65	-
Sex	Male Female	-
Education	No formal education < Primary to primary < Secondary to higher secondary college	
Employment	Employed Self & ad-hoc employed Student Not employed	Individuals with missing information were excluded from the analyses
Environmental characteristics		
Smokers at home	No Yes	If there are one or more smokers at home then it is "Yes", and any other case is "No".
Permissibility of smoking at home	Not allowed allowed Do not know/refused/ missing/not applicable	"Allowed" include: Allowed, not allowed but exceptions and no rules.
Magnitude of second-hand exposure at home	Never Daily Occasionally Do not know/refused/ missing/not applicable	"Occasionally" include: Weekly, monthly and less than monthly.
Permissibility of smoking at work	Not allowed Allowed No/do not work Do not know/refused/ missing/not applicable	Not allowed Allowed No/do not work Do not know/refused/ missing/not applicable "Allowed" include: Allowed anywhere, allowed only in some indoor areas and there is no policy. "No/do not work" is one possible answer to the question: Do you currently work outside of your home?
Advise to quit by health care provider	No Yes Does not apply Refused/missing/not applicable	"Does not apply" are those who either did not visit health care provider or those who were not asked by health care provider if they used tobacco

Table 3 Continued.		
Behavioral characteristics		
Age at initiation	1-24 25-44 >45 Do not know/refused/ missing/ not applicable	Data on age at initiation of first smoke/ smokeless use was available for the same participant in two different variables. In this case, the variable with less missing values namely the question, "How old were you when you first started smoking tobacco daily?" was used as the primary variable; for those missing, data was then imputed from the second variable, "How many years ago did you first start smoking tobacco daily?", the data was computed by subtracting this value from the age.
Time of first smoke (minutes)/ time of first smokeless use (minutes)	Up to 30 >30 Not asked Refused/missing/not applicable	"Not asked" category includes less than daily current smokers tobacco users from whom this data was not collected. Data was not collected for the successful quitters.
Typical daily consumption	<1 1 to10 >10 Do not know/refused/ missing/ not applicable	This variable was created by summing up the number of times a participant used any of the smoked products per day. For less than daily users, the numbers smoked used per week was divided by 7. Data was not available for the successful quitters.
Purchase type	For smoked tobacco: Loose Pack Both Others	For smoked tobacco: "Loose" if both cigarettes and bidis were purchased as loose "Pack" if both cigarettes and bidis were purchased as pack or carton "Both" if one product was purchased as loose and the other as pack Everything else was considered as "Others". Data was not collected for successful quitters.
Periodicity	Daily Less than daily	
Knowledge on the ill effects of smoking tobacco	No Yes Do not know/refused/ missing/ not applicable	Number of questions was different in GATS surveys. One common question on the serious ill effects of smoking tobacco has been used as a surrogate to assess the knowledge.

Table 3 Continued.		
Noticing anti-smoking advertisement		This variable was created by counting the number of "Yes" as an answer to questions related to noticing anti-smoking advertisement in the last 30 days in the two surveys and dividing it by number of questions. This was done for each survey separately to account for the different number of questions between the two. For smoking, questions used from GATS 1: G01a1, G01a2, G01b1, G01b2, G01c1, G01c2, G01d1, G01d2, G01e1, G01e2, g02, and from GATS 2: G01A, G01B, G01C, G01D, G01E, G01F, G01G, G01H, G01I, G02, G02B.
Noticing smoking tobacco promotions		This variable was created by counting the number of "Yes" as an answer to questions related to noticing smoking tobacco promotions in the last 30 days in the two surveys and dividing it by number of questions. This was done for each survey separately to count for the different number of questions between the two. For smoking, questions used from GATS: G04a1, G04a2, G04b1, G04b2, G04c1, G04c2, G04d1, G04d2, G04e1, G04e2, G04f1, G04f2, G04g1, G04g2, G04h1, G04h2, G04i1, G04i2, G04j1, G04j2, G04k1, G04k2, G06a1, G06a2, G06b1, G06b2, G06c1, G06c2, G06d1, G06d2, G06e1, G06e2, G06f1, G06f2, G06g1, G06g2, and G04A, G04B, G04C, G04D, G04E, G04F, G04G, G04H, G04I, G04J, G04K, G06A, G06B, G06C, G06D, G06E, G06F, G06G, G206A, G206B, G206C, G206D, G206E, G206F, G206G.
Users of the other tobacco form	Mono Dual Former	Mono: Have not used the other form. Dual: currently using the two forms of tobacco. Former: used the other form of tobacco in the past.
GATS: Global Adult Tobacco Survey		

Given the importance of employment and education as proxies for economic status, cases with missing data in these two variables were excluded from the analysis. Moreover, smokers were classified into three distinct outcome groups under the dependent variable:

1. Current smoker who attempted to quit but failed (32.35%).

2. Current smoker who did not attempt to quit (61.49%).

3. Smoker who successfully quit smoking (6.16%).

Note: To address class imbalance, SMOTE was explicitly applied during the training phase using Orange's dedicated SMOTE widget. This allowed for oversampling of minority classes prior to model evaluation.

Records with missing outcome classification were also removed to preserve the integrity of the target variable used for predictive modeling.

Appropriate preprocessing techniques were applied to prepare the dataset for modeling. Categorical variables from the GATS questionnaire were encoded using Label Encoding for binary variables (e.g., yes =1, no =0) and One-Hot Encoding for multi-class variables (e.g., education level, employment level).

The continuous variables, like age, were transformed into categorical age groups (e.g., 15-24, 25-34) using binning techniques in R to enhance interpretability and align with the dataset's categorical structure. A similar transformation was applied to daily cigarette consumption and monthly tobacco expenditure, grouping them into defined ranges to reduce skewness and improve model performance.

These preprocessing steps ensured the dataset was in a consistent and analyzable format suitable for ML algorithms. The final cleaned dataset processed using R programming language version 3.35 and Power Bi version 2.124 consisted of 24 variables and included 107,329 individuals who reported using smoked tobacco.

Business intelligence system development

A business intelligence (BI) system was developed to facilitate the analysis and visualization of patterns related to social, environmental, and behavioral factors associated with tobacco use. The BI reports provided insights into why certain individuals do not attempt to quit smoking, while others relapse despite multiple attempts. Furthermore, the system enabled the identification of enablers and barriers smoking and smokeless tobacco cessation. The structured visualizations supported the interpretation of complex relationships between variables and contributed to formulating data-driven hypotheses for subsequent predictive modeling.

Application of machine learning algorithms

Two primary ML methodologies were applied to analyze the prepared dataset. The first approach involved a classification task using six supervised learning algorithms: k-NN, SVM, Neural Networks, Random Forest, Decision Tree, and Logistic Regression. These models were employed to classify smokers into predefined categories and to determine the most influential factors contributing to successful smoking cessation. The classification processes were carried out using the Orange data mining platform.

The second approach focused on predictive modeling, in which the same classification algorithms were utilized to forecast the likelihood of a smoker belonging to a specific category. This predictive analysis aimed to enhance the understanding of behavioral trends and to support the design of targeted cessation strategies.

RESULTS

The considerable gap between the number of individuals trying to quit smoking and those who succeed underscores the significant challenges encountered by smokers. This struggle to quit, despite repeated attempts, indicates that a range of environmental, social, and behavioral factors are pivotal.

Figure 2 highlights the significant variation in smoking cessation success rates across different countries. Nations such as Brazil, Costa Rica, and Mexico exhibit the highest percentages of successful quitters. These results likely reflect the presence of robust public health strategies. Considering that the dataset includes data from various years for different

countries, this could introduce temporal biases and misrepresent the findings.

Descriptive statistics

Data indicates a notable shift in workplace smoking policies over the years, with a trend toward decreasing smoking allowances and increasing prohibitions. This environmental change has led to more quiet attempts, demonstrating the impact of policy on smoking behaviors. Furthermore, unemployed individuals show a higher success rate in quitting smoking, suggesting that the lack of workplace stressors and smoking culture supports cessation efforts (see Figure 3).

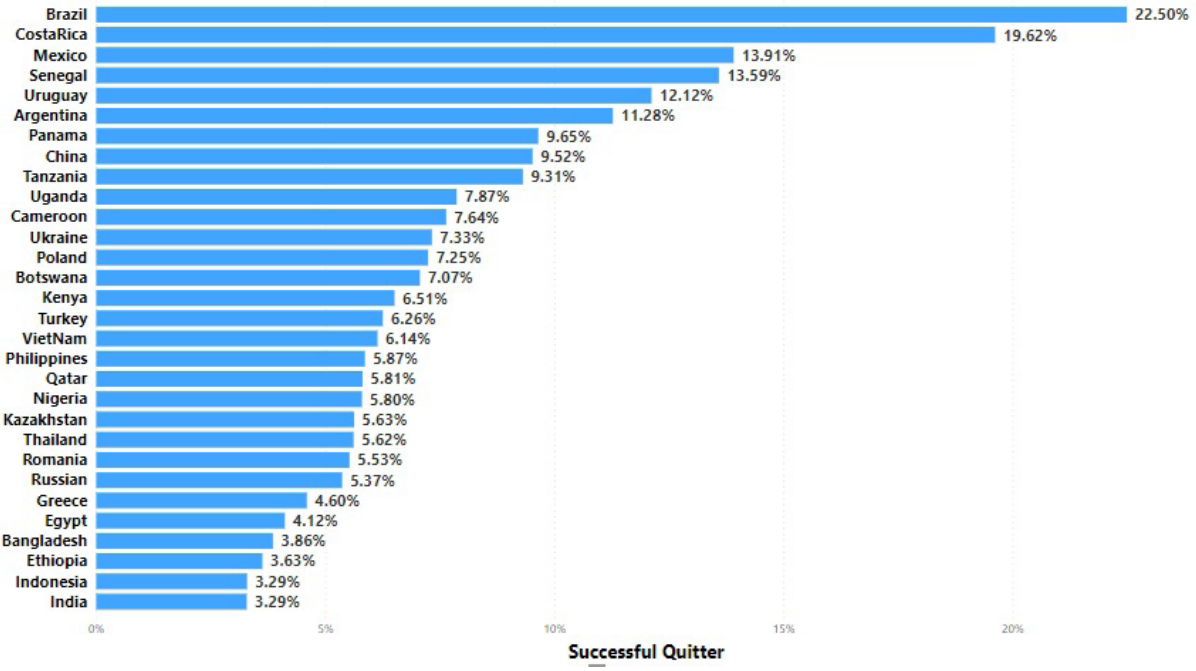


Figure 2. Comparative analysis of smoking cessation success across countries

Age and education significantly impact smoking cessation. Younger individuals (ages 15-24) tend to smoke more heavily, influenced by social pressures and a reduced awareness of health risks. In contrast, smoking rates decline markedly with age (see Figure 4). dependence on smoking cessation.

Individuals with higher education levels show greater rates of quit attempts and success, indicating that social factors such as education and awareness are crucial in smoking cessation. Additionally, types of employment influence smoking behaviors and cessation efforts. Individuals in self-employed or temporary positions demonstrate lower quit attempt rates compared to employees and students, highlighting the effect of job security and work environment on smoking habits (see Figure 5).

Gender differences are evident, with women showing a higher success rate in quitting smoking than men. While men have higher smoking rates and make more attempts to quit, their success rates are lower, suggesting potential behavioral differences in smoking cessation strategies between genders.

The family environment is another vital factor, as the presence of other smokers in the household can greatly hinder an individual's attempts to quit.

Furthermore, individuals who smoke their first cigarette soon after waking are more nicotine-dependent, making it harder for them to quit. This underscores the impact of daily routines and nicotine dependence on smoking cessation.

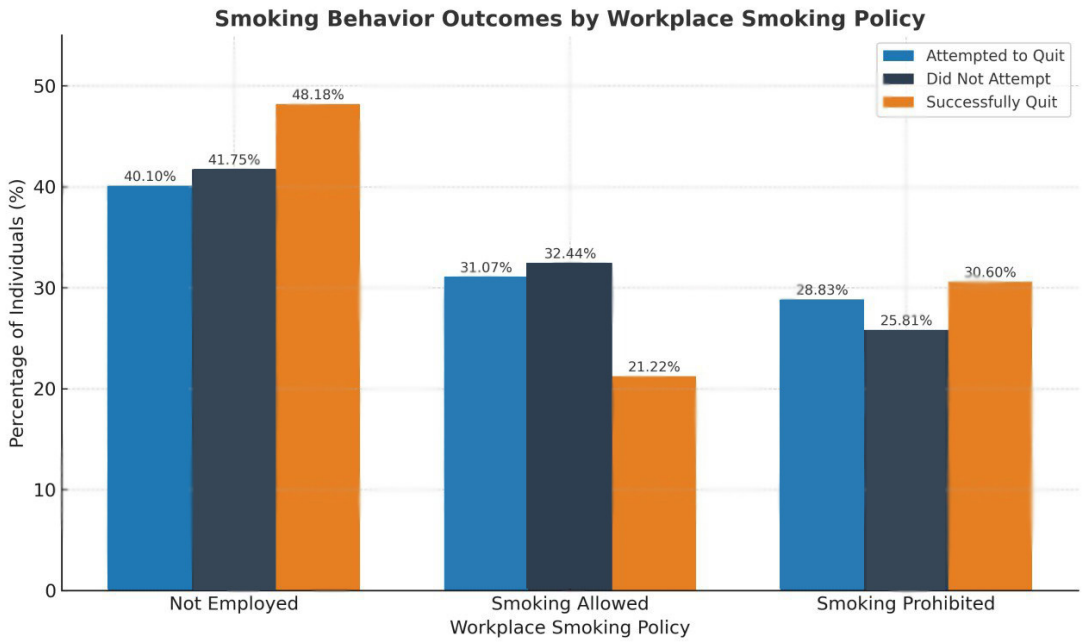


Figure 3. Smoke categories by workplace smoking policy

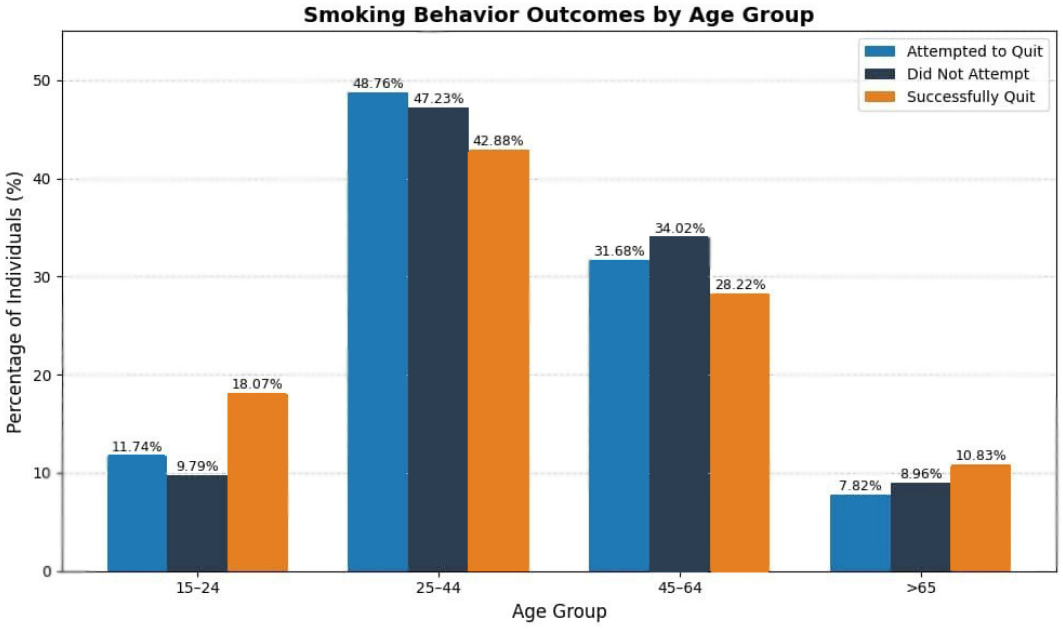


Figure 4. Smoke categories by age group

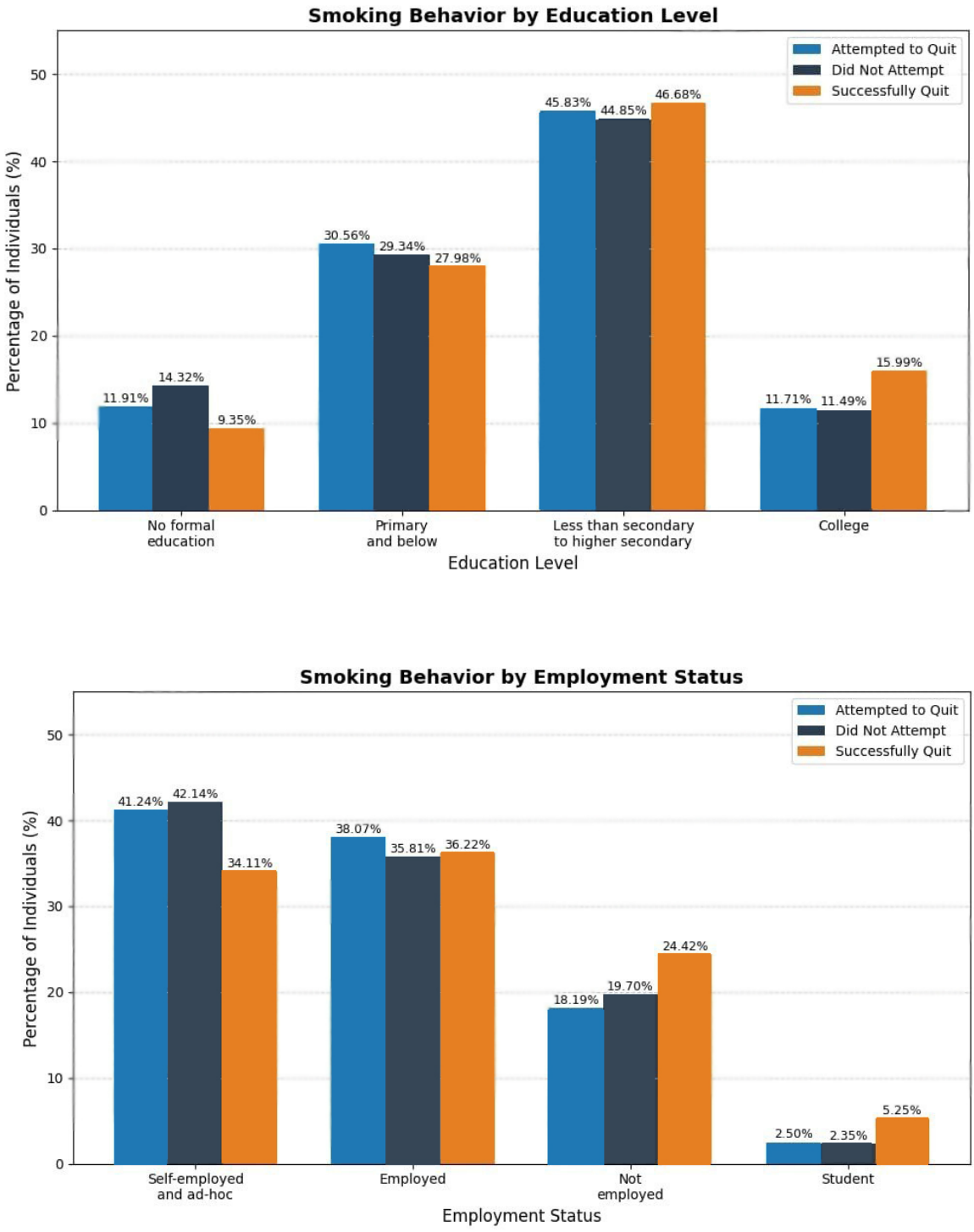


Figure 5. Smoke categories by education level & employment status

Knowledge of the harmful effects of smoking, receiving guidance from healthcare providers, and exposure to anti-smoking advertisements are positively associated with increased quit attempts and success rates. These educational and social interventions are crucial in promoting smoking cessation. In contrast, individuals who lack this knowledge, guidance, and exposure show significantly lower rates of attempts and success.

Although the overall quit attempt rates are lower, individuals who are not exposed to smoking at home have a comparatively higher success rate in quitting. This underscores the significance of a supportive and smoke-free home environment in promoting smoking cessation.

A comparison between rural and urban areas indicates that quit attempts and success rates are somewhat lower in rural regions, highlighting the need for targeted support and interventions to address environmental and social disparities.

In summary, while social and behavioral factors are closely linked to smoking habits and cessation success, social factors such as age, education, and family environment, along with behavioral routines, seem to be most strongly associated with smoking behaviors and cessation outcomes.
In summary, although social and behavioral factors are closely connected to smoking habits and cessation success, social elements like age, education, and family environment, along with behavioral routines, appear to be most strongly associated with smoking behaviors and cessation outcomes.

The performance of various classification algorithms was assessed for a multi-class prediction task, aimed at differentiating among three distinct smoking behavior categories: individuals who attempted to quit smoking, those who did not attempt to quit, and those who successfully quit. The algorithms evaluated included Logistic Regression, Naïve Bayes, Neural Networks, k-NN, Random Forest, and Decision Tree. To ensure reliable model evaluation and prevent overfitting, the dataset was initially split using an 80/20 train-test ratio, where 80% of the samples were allocated for model training and 20% for testing. This division was used consistently across all algorithms during the initial assessment phase.

In addition to this split, a 10-fold cross-validation procedure was employed during model training using Orange's "Test & Score" widget. This approach ensures that each model is evaluated on multiple subsets of the data, thereby providing more robust and generalizable performance metrics.

While the Orange data mining platform does not support automated hyperparameter optimization (e.g., grid search or randomized search) natively, we conducted manual hyperparameter tunning to improve model performance and fairness in comparison. Key adjustments included:

k-Nearest Neighbors (k-NN): Values of k=1, 3, and 5 were tested. Based on cross-validation performance, k=5 was selected as it offered a favorable balance between sensitivity and generalization.

Random forest: The number of trees was varied (n=10, 50, 100), and the setting that yielded the best macro-average accuracy was chosen for final evaluation (n=10).

Neural networks: the settings were meticulously chosen based on best practices outlined in the scientific literature and initial parameter tuning experiments. Specifically, the hidden layer comprised 100 neurons to strike a balance between predictive accuracy and computational efficiency. The ReLU activation function was utilized due to its established effectiveness in deep learning architectures. The Adam optimizer was adopted for its stability and efficiency, accompanied by a small regularization parameter (0.0001) to mitigate overfitting. Training was conducted over 200 iterations, allowing adequate convergence without overtraining. Furthermore, reproducible training was implemented to guarantee the replicability of results.

These manual adjustments, while limited compared to automated tuning frameworks, contributed to enhanced model stability and performance. A summary of the chosen configurations is now included to improve transparency and reproducibility of the methodology.

Considering the multi-class nature of the target variable, performance was mainly assessed using macro-average accuracy.

This averaging method was selected to ensure a balanced evaluation across the three outcome categories. The comprehensive performance results for all classification models are summarized in Table 4.

Among the algorithms tested, LR demonstrated the highest accuracy in classifying smoking behavior categories, achieving an accuracy score (area under curve) of 0.711, indicating a strong predictive capability based on sociodemographic, behavioral, and environmental factors. Naive Bayes and RF followed closely with accuracy scores of 0.704 and 0.666, respectively, showcasing reasonable predictive power. The Neural Networks and k-NN algorithms also yielded meaningful predictions, attaining accuracy scores of 0.656 and 0.645, respectively. In contrast, the decision tree algorithm exhibited the lowest accuracy, with a score of 0.599, suggesting its relatively lower effectiveness in this context.

Further analysis revealed that LR not only achieved higher accuracy but also effectively identified the most influential factors affecting classification outcomes. As illustrated in Figure 6, behavioral factors emerged as the most significant predictors, highlighting their critical role in determining both the likelihood of attempting to quit smoking and the success of these attempts. These findings emphasize the importance of model selection in accurately classifying smoking behaviors, particularly when a range of sociodemographic and environmental factors is considered. The superior performance of LR suggests its reliability as a tool for predictive tasks in this field, providing valuable insights that can inform public health strategies aimed at smoking cessation interventions.

Table 4. Performance measures of different classification models

Model	AUC	CA (%)	F1	Precision	Recall	MCC
Logistic regression	0.711	69.8	0.641	0.673	0.698	0.327
Decision tree	0.599	60.4	0.606	0.609	0.604	0.194
Random forest	0.666	64.9	0.642	0.637	0.649	0.258
k-NN	0.645	65.4	0.630	0.624	0.654	0.239
Neural network	0.656	66.0	0.634	0.629	0.660	0.249
Naive Bayes	0.704	69.4	0.655	0.666	0.694	0.318

AUC: Area under curve, CA: Classification accuracy, MCC: Matthews correlation coefficient, k-NN: k-Nearest Neighbors

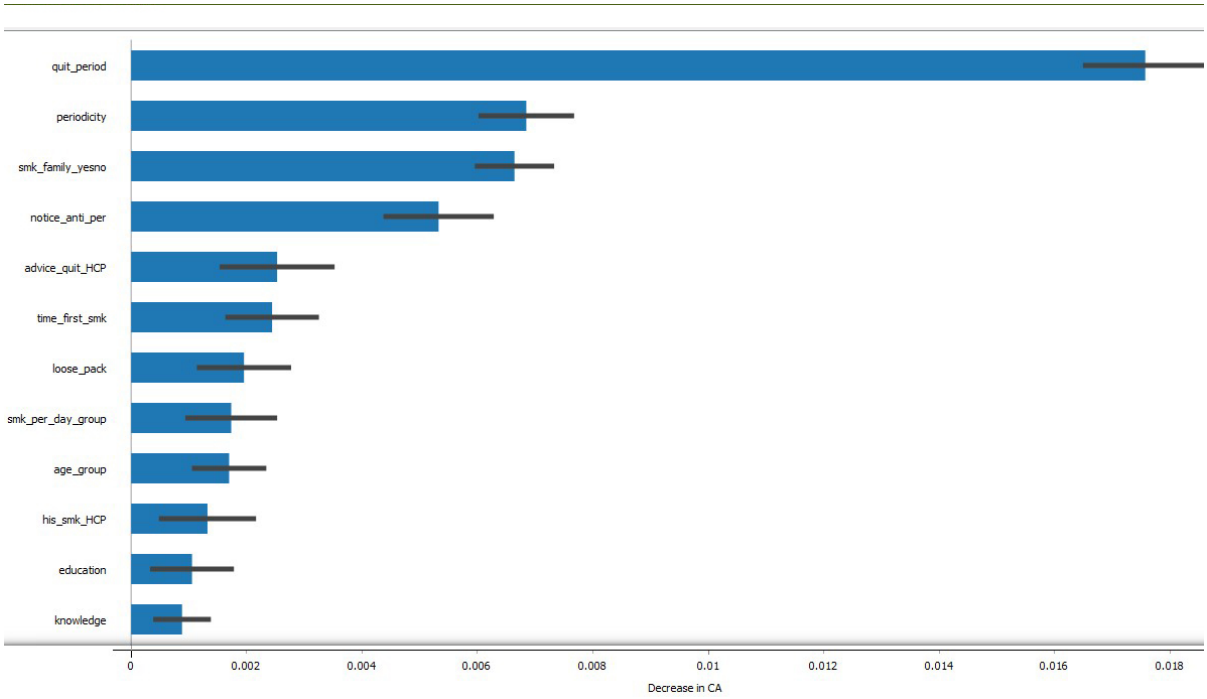


Figure 6. Permutation feature importance technique

DISCUSSION

This study illustrates the potential of ML algorithms in predicting smoking cessation outcomes and identifying influential factors using data from the GATS. The application of various ML algorithms, including logistic regression, naive bayes, neural networks, k-NN, random forests, and decision trees, offered insights into the complex interplay of factors affecting smoking behaviors and cessation success.

Our findings indicate that LR surpassed other algorithms, achieving the highest accuracy (69.8%) in classifying smoking behavior categories. This suggests that linear relationships between predictors and outcomes may significantly influence smoking cessation dynamics. The strong performance of LR aligns with previous studies in this field, such as the work by Davagdorj et al.[10] which demonstrated that LR was effective in predicting smoking cessation outcomes.

The study emphasizes the significance of social, demographic, behavioral, and environmental factors in influencing smoking cessation. Notably, education level, employment status, and family environment emerged as crucial determinants of cessation success. These findings reinforce and highlight the necessity for tailored interventions that address these factors.

The variation in cessation success rates across countries, as depicted in Figure 2, highlights the influence of national policies and cultural contexts on smoking behaviors. Countries with higher success rates, such as Brazil, Costa Rica, and Mexico, may provide valuable insights for policy development in other regions.

Our analysis also revealed gender differences in cessation outcomes, with women exhibiting higher success rates despite lower attempt rates. This observation necessitates further investigation into

gender-specific cessation strategies and support mechanisms. The role of workplace smoking policies in influencing quitting attempts illustrates the potential of environmental interventions in promoting healthy behaviors. Likewise, the impact of knowledge regarding smoking risks and exposure to anti-smoking messages on quit attempts underscores the importance of education and awareness campaigns.

CONCLUSION

This study contributes to the application of ML in public health, specifically in the area of smoking cessation. The findings can guide the creation of more targeted and effective smoking cessation interventions, considering the intricate interplay of individual, social, and environmental factors. Further research is necessary to examine the generalizability of these algorithms across diverse populations and to explore the potential of integrating ML algorithms with traditional public health strategies to bolster global smoking cessation efforts.

It is important to note that this study encountered limitations concerning smokeless tobacco data. Due to the scarcity of available data on smokeless tobacco use in many countries included in GATS, this study could not perform a thorough analysis of this form of tobacco consumption. This limitation may lead to an underestimation of the overall picture of tobacco use, particularly in areas where smokeless tobacco is prevalent. This highlights the necessity for more extensive data collection on all types of tobacco consumption in future studies.

ACKNOWLEDGEMENT

Conflict of Interest: The authors report no declarations of interest.

Financial Support: No support has been received for this study.

Ethical Declaration: Since secondary data were used in the study, there is no need for ethics committee approval.

Author Contrubition: Concept: MA, EA, LS, Design: MA, EA, LS, Supervising: MA, EA, LS, Data collection and entry: MA, EA, LS, Literature search: MA, EA, LS, Writing: MA, EA, LS

REFERENCES

1. Warren KB, Devin CT, William HC, al. Predictors of smoking cessation outcomes identified by machine learning:A systematic review. Addict Neurosci 2023;6:100068.
2. Lai CC, Huang WH, Chang BC, Hwang LC. Development of Machine Learning Models for Prediction of Smoking Cessation Outcome. Int J Environ Res Public Health 2021;18(5):2584.
3. Rijhwani K, Mohanty VR, Aswini YB, Singh V, Hashmi S. Applicability of Data Mining and Predictive Analysis for Tobacco Cessation: An Exploratory Study. Front Dent 2020;17:24.
4. Thangaraju P, Barkavi G, Karthikeyan T. Mining Lung Cancer Data for Smokers and Non-Smokers by Using Data Mining Techniques. IJARCCE. 2014;3(7):7622-7626.
5. Kharabsheh M, Qawasmeh A, Megdadi O, Jawabreh N, Mudallal R, Alzyoud S. A Critical Analysis of the Relationship between Depression and Smoking using Machine Learning. IJSTR 2019;8(12):22-26.
6. Roy P, Hossain MF, Jahan N. Machine Learning Approach to Predict Influence of Smoking on Student Life. ICCCNT Proc. 2021;1-6.
7. Ismail RM. Using machine learning algorithms to study the smoking behavior of Iraqi students. Eurasian Research Bulletin 2023;16:91-101.
8. Choi J, Jung H.T, Ferrell A, Woo S, Haddad L. Machine Learning-Based Nicotine Addiction Prediction Models for Youth E-Cigarette and Waterpipe (Hookah) Users. J Clin Med 2021;10(5):972.

9. Kharabsheh M, Meqdadi O, Alabed M, Veeranki S, Abbadi A, Alzyoud S. A Machine Learning Approach for Predicting Nicotine Dependence. IJACSA 2019;10(3):179-184.
10. Davagdorj K, Yu SH, Kim SY, et al. Prediction of 6 Months Smoking Cessation Program among Women in Korea. Int J of Machine Learning and Computing 2019;9(1):83-90.
11. Thakur SS, Poddar P, Roy RB. Real-time prediction of smoking activity using machine learning based multi-class classification model. Multimed Tools Appl 2022;81(10):14529-14551.
12. Durmuşoğlu UZD, Çiftçi KP. Socio-demographic determinants of smoking: A data mining analysis of the Global Adult Tobacco Surveys. Turk J Public Health 2021;19(3):251-262.
13. Davagdorj K, Park KH, Lee JS, Ryu KH. A machine-learning approach for predicting success in smoking cessation intervention. IEEE 2019.
14. Issabakhsh M, Sanchez-Romero LM, Le TTT, et al. Machine learning application for predicting smoking cessation among US adults: An analysis of waves 1-3 of the PATH study. PLoS ONE 2023;18(6):e0286883.
15. Siddiqui MA, Khan AS, Witjaksono G. Classification of the factors for smoking cessation using logistic regression, decision tree & neural networks. AIP Conf Proc 2020;2203:020036020036-72.
16. Durmuşoğlu ZDU, Kocabey Çiftçi P. Classification of Smoking Status: The Case of Turkey. Symposium Series on Computational Intelligence 2016;1-5.
17. Poynton MR, McDaniel AM. Classification of smoking cessation status with a backpropagation neural network. J Biomed Inform 2006;39(6):680-686.
18. Wanga X, Zhao K, Cha S, et al. Mining user-generated content in an online smoking cessation community to identify smoking status: A machine learning approach. Decis Support Syst. 2019;116:26-34.
19. Martinović T. Investigating Tobacco Usage Habits using Data Mining Approach. Entrenova 2015;1-10.
20. Perski O, Li K, Pontikos N, Simons D, et al. Classification of Lapses in Smokers Attempting to Stop: A Supervised Machine Learning Approach using Data from a Popular Smoking Cessation Smartphone App. Nicotine Tob Res 2023;25(7):1330-1339.
21. Al-ssabbagh M, Elango V, Winkler V. What makes people quit tobacco and succeed at it? An exploratory analysis of smoked and smokeless tobacco from India. Prev Med 2022;158:107033.