Check for updates

RESEARCH NOTE

REVISED Application of machine learning techniques to profile smoking behavior of adolescent girls in Ghana

[version 2; peer review: 1 approved, 2 approved with reservations]

Sara V. Flanagan, Ariadna Vargas, Jana Smith

ideas42, New York, New York, 10004, USA

v2 **First published:** 03 Jan 2024, **8**:2
https://doi.org/10.12688/gatesopenres.14991.1
Latest published: 20 Nov 2024, **8**:2
https://doi.org/10.12688/gatesopenres.14991.2

Abstract

Background

Tobacco use trends among adolescents in low- and middle-income countries, and in particular narrowing gender gaps, highlight the need for interventions to prevent and/or reduce tobacco use among adolescent girls. We evaluated a social marketing program in Ghana discouraging tobacco use among adolescent girls and additionally investigated the pathways influencing smoking behaviors to identify programmatic opportunities for impact. Leveraging the data collected through the stepped wedge cluster randomized trial and panel survey of 9000 girls aged 13–19 , we sought to apply machine learning (ML) techniques to identify the most important variables for predicting initiation of smoking.

Methods

To identify predictors of smoking initiation we sought to develop a model which could accurately differentiate smokers from non-smokers and evaluated various ML approaches for training classifier algorithms to achieve this. We selected a Synthetic Minority Over-sampling Technique (SMOTE) because it optimized the recall and precision of the model. We then utilized the technique of feature importance for greater insight into how the model arrived at its decisions and to rank the most important variables for predicting smokers. To explore different dimensions of smoking behavior, including initiation and continuation, we trained our model by using several combinations of target outcomes and input variables from the panel survey.

Open Peer Review

Approval Status ✓ ? ?

	1	2	3
version 2 (revision) 20 Nov 2024	✓ view		? view
version 1 03 Jan 2024	? view	? view	

1. **Ruoyan Sun**, The University of Alabama at Birmingham, Birmingham, USA
2. **Jennifer M. Murray**, Queens University Belfast, Belfast, UK
3. **Runqiu Wang**, University of Nebraska Medical Center, Omaha, USA

Any reports and responses or comments on the article can be found at the end of the article.

Results

The resulting features of smokers highlight the importance of girls' independence and connectivity, social environment, and peer influence on likelihood of smoking, and in particular subsequent initiation. These results were largely consistent with our formative research findings based on qualitative interviews informed by behavioral science.

Conclusions

This novel application of ML techniques demonstrates how data science approaches can generate new programmatic insights from rigorous evaluation data, especially when data collection is informed by behavioral theory. Such insights about the relative importance of different features can be valuable input for program planning and outreach.

Keywords
machine learning, synthetic data, smoking, tobacco, adolescent girl, algorithm, behavioral science

Corresponding author: Sara V. Flanagan (sara@ideas42.org)

Author roles: Flanagan SV: Data Curation, Investigation, Project Administration, Writing – Original Draft Preparation; **Vargas A**: Formal Analysis, Methodology, Software, Writing – Original Draft Preparation; **Smith J**: Conceptualization, Funding Acquisition, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the Bill and Melinda Gates Foundation [INV005809].

Copyright: © 2024 Flanagan SV *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Flanagan SV, Vargas A and Smith J. **Application of machine learning techniques to profile smoking behavior of adolescent girls in Ghana [version 2; peer review: 1 approved, 2 approved with reservations]** Gates Open Research 2024, **8**:2 https://doi.org/10.12688/gatesopenres.14991.2

First published: 03 Jan 2024, **8**:2 https://doi.org/10.12688/gatesopenres.14991.1

REVISED **Amendments from Version 1**

This version is a response to viewer comments. Additional details and references have been added to the Introduction section. Additionally, we have clarified the secondary analyses performed, added the notable results, and expanded on a few points in the Discussion section.

Any further responses from the reviewers can be found at the end of the article

Introduction

Machine learning (ML) is a discipline at the intersection of data science and artificial intelligence with a focus on building algorithms to make predictions without requiring explicit programming to do so[1]. The approach involves building a model using sample training data. These methods are increasingly being applied to support evidence-based decision making across a wide range of fields, including global health. Here we report our experience applying ML to generate programmatic insight from data collected for a rigorous impact evaluation.

Rising tobacco use among adolescents in low and middle-income countries (LMICs)[i] and increasing tobacco use among girls relative to boys[ii] highlight the need for interventions to prevent and reduce tobacco use among adolescent girls. Non-cigarette tobacco products are contributing to the narrowing gender gap,[iii] and surveys suggest shisha use is now more prevalent among girls than boys in Ghana, although with significant regional variation.[iv] As external evaluators of a social marketing program in Ghana focused on discouraging tobacco use among adolescent girls, we conducted a stepped wedge cluster randomized trial and panel survey of 9000 girls aged 13–19 in select neighborhoods of Accra and Kumasi over twelve months of 2021–2022. A secondary objective was to provide additional behavioral insights on the pathways influencing smoking behavior among teenage girls in Ghana and to identify programmatic opportunities for impact. Although a formative research phase with qualitative interviews informed by the behavioral science literature was conducted to support this objective[2], we were interested to explore how recent advances in ML techniques might be applied to our panel survey data to generate additional insights related to the predictors of smoking initiation and to compare and contrast findings from these two approaches.

Although our study had a relatively large sample, smoking behavior was quite rare. Only 1.6% of girls reported having ever tried smoking (defined as cigarette or shisha use) at baseline and 0.1% having smoked in the past 30 days. We wanted to think more creatively about ways to profile potential smokers in this population beyond basic demographic slices of the sample. Our objective was to first build an algorithm that would predict which girls are more likely to be smokers and then identify which factors are most important to make those predictions. In this research note we describe how we explored several ML methods to build a classifier model and then applied a ML explainability method to profile the most important predictors of smokers using different smoking definitions, as well as different data subsamples.

Methods

Data source

Data collection for the step wedge evaluation included a survey of all 9000 girls over four rounds—Baseline (before any implementation), Midline 1 (after the first period of implementation), Midline 2 (after the second), and Endline (after all areas had been activated). Participants aged 13–19 were recruited through multi-stage sampling in which neighborhood clusters in Accra and Kumasi were first randomly selected for the study, and then households within each cluster were systematically approached through a community mapping process until the target number of adolescent girls willing to participate in the study were identified. To be considered eligible for the study, girls had to have access to a phone and had to intend to remain at their residence over the subsequent year, to enable enumerators to reach them at future rounds of the panel survey. The questionnaire was informed by the formative research and a smoking-oriented theory of change and comprised about 100 questions covering: background and demographic information; social context, confidence, and self-efficacy; sources of influence on girls' decisions and actions; tobacco perceptions and norms; tobacco use, opportunity, and refusal; and program exposure and perceptions. Tobacco use questions were adapted from the Global Youth Tobacco Survey.[v] The questionnaire was reviewed for face validity with the program team and portions were piloted with 396 adolescent girls during the formative research phase[2]. Most responses were binary or Likert scales of 1 to 4 that were recoded to binary (agree vs. disagree, likely vs. not likely) for this exploratory exercise. Several indices of multiple survey items that were confirmed to have high internal consistency served as indicators

[i] Stone E, Peters M. Young low and middle-income country (LMIC) smokers—implications for global tobacco control. Transl Lung Cancer Res. 2017 Dec;6(Suppl 1):S44–6.

[ii] WHO | Regional Office for Africa [Internet]. [cited 2022 Oct 11]. Tobacco Control. Available from: https://www.afro.who.int/health-topics/tobacco-control

[iii] Agaku IT, Sulentic R, Dragicevic A, Njie G, Jones CK, Odani S, *et al.* Gender differences in use of cigarette and non-cigarette tobacco products among adolescents aged 13–15 years in 20 African countries. Tob Induc Dis. 2024 Jan 22;22:10.18332/tid/169753.

[iv] Logo DD, Kyei-Faried S, Oppong FB, Ae-Ngibise KA, Ansong J, Amenyaglo S, *et al.* Waterpipe use among the youth in Ghana: Lessons from the Global Youth Tobacco Survey (GYTS) 2017. Tob Induc Dis. 2020 May 29; 18:47.

[v] Global Youth Tobacco Survey Collaborative Group. Global Youth Tobacco Survey (GYTS): Core Questionnaire with Optional Questions, Version 1.2. Atlanta, GA: Centers for Disease Control and Prevention, 2014.

of intermediate outcomes in the impact evaluation; however, for this application of ML techniques survey items were maintained as individual variables. Ethical review and approval for the evaluation study were provided by Innovations for Poverty Action IRB (protocol #15798) and the Ghana Health Service Ethics Review Committee (GHSERC: 003/11/20). Written informed consent to participate in the study was obtained from participants aged 18 years or older. For younger participants, written parental consent was first obtained followed by written assent of the adolescent.

Building the classification model

Initially, we trained a basic Random Forest classifier using the baseline data to identify girls who have ever smoked, which performed well in terms of accuracy (Table 1). Accuracy is defined as the proportion of all girls who are correctly classified as either smokers or non-smokers (Figure 1).

To optimize our baseline model, we chose to focus on increasing the recall metric, rather than just accuracy. Recall is a measure of the model's ability to correctly identify actual smokers; a high recall value indicates that the model is able to identify most girls who smoke, while a low recall value suggests that the model is missing a significant number of smokers. Accuracy, which is affected by the low prevalence of smoking in our data, can bias the model towards classifying most individuals as non-smokers. Therefore, we were more concerned with correctly identifying those who did smoke to ensure effective programmatic targeting towards likely smokers, even if it meant that some non-smokers were mistakenly identified as smokers.

Additionally, we focused on improving the model's precision. If the precision of the model is low, it means that the model is classifying many individuals as smokers even though they are not, which could result in potentially unnecessary or less well-targeted interventions. Therefore, improving precision helps to reduce targeting of non-smokers and ensures that the resources allocated to smoking interventions are still used effectively and efficiently.

With the objective of optimizing the recall and precision, we evaluated the following ML approaches using the "imbalance-library" Python package[3] (which is built upon the "scikit-learn" ML library) to determine which method produced the optimal outcome:

- *RandomOverSampler*: This technique works by randomly oversampling the minority class (smokers) in the dataset, i.e., it generates additional random samples of the minority class to balance the class distribution.
- *ADASYN (Adaptive Synthetic Sampling)*[4]: ADASYN works by generating synthetic samples of the minority class based on the density of the minority class in the feature space and the distance between minority samples. The goal is to increase the number of samples of the minority class in the training data while maintaining a balanced representation of the classes in the feature space.
- *SMOTE (Synthetic Minority Over-sampling Technique)*[5]: Like ADASYN, SMOTE works by generating synthetic samples of the minority class, but instead of generating samples based on the density and distance of the minority class, it generates samples based on the

Table 1. Performance of classification algorithm using machine learning techniques.

Model	Accuracy	Precision	Recall
Without oversampling	0.9594	0.2500	0.0448
RandomOverSampler	0.9586	0.6494	0.3117
ADASYN	0.9669	0.5445	0.2675
SMOTE	0.9901	0.7900	0.7299
Class weighting	0.9174	0.3576	0.2595
Hellinger Distance	0.9131	0.7413	0.6603

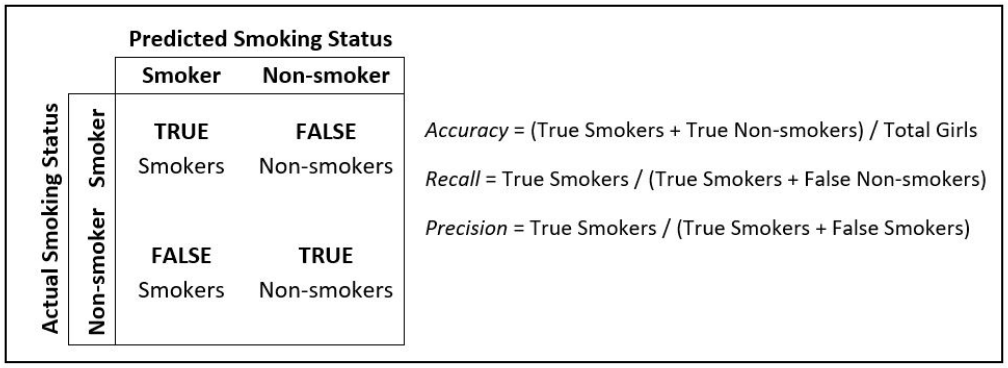


Figure 1. Actual vs. predicted smoking status.

k-nearest neighbors of the minority samples. In this study 5 k-neighbors were used.

- *Class weighting*: This method assigns higher weights to the minority class, encouraging the algorithm to pay more attention to the minority class and make better predictions.

- *Hellinger Distance as a Tree Split Criterion*[6]: Hellinger distance is a measure of the difference between two probability distributions. In decision tree algorithms, Hellinger distance can be used as a criterion for splitting nodes in the tree, to determine the most informative split that maximizes the difference between the classes in the feature space. Using Hellinger distance as a split criterion can improve the accuracy of decision tree algorithms in imbalanced classification tasks by giving more attention to the minority class.

We ultimately selected the SMOTE technique of generating synthetic data for training the classification algorithm because it optimized its performance in terms of recall and precision (Table 1).

Generating smoker profiles

After successfully building a high-performing classification model, our goal was to build a profile of girls who smoke. However, ML models are often referred to as "black boxes" because they can be difficult to interpret and comprehend. This means that the model might make accurate predictions, but it's not clear how it arrived at those predictions or why a particular prediction was made.

To gain more transparency, we utilized the technique of feature importance. This method explains predictions made by the model by identifying which features or input data had the most significant impact on those predictions. Essentially, it helps us understand which characteristics have the most influence in predicting a girl to become a smoker. By calculating feature importance we aimed to generate greater insight into how the model arrived at its decisions and to rank the most important variables. We determined whether an important feature is a risk for or protective against smoking based on the direction of that variable's association with the smoking outcome in our data.

To determine the profiles of girls most likely to smoke, we trained our model by using the following combinations of target outcomes and input variables from our survey data:

Model #1: Predicted outcome: reported ever tried smoking at baseline.

Training inputs: baseline survey responses

Model #2: Predicted outcome: reported having smoked during the previous month at endline.

Training inputs: endline survey responses

Model #3: Predicted outcome: among non-smokers at baseline, reported ever tried smoking at endline.

Training inputs: baseline survey responses

As a secondary analysis we looked at these same combinations within sub-groups of respondents, including younger (13–15) vs. older girls (16–19) and Accra vs. Kumasi, to support interpretation of findings and identify heterogeneity in predictors. We also analyzed additional combinations of data inputs on these predicted outcomes at different time points for robustness.

Results

Model #1 (Table 2) differentiates between girls based on smoking history prior to the study. The results highlight that girls who had ever tried smoking report different social settings and experiences at present than non-smokers. They are more likely to have regular phone access, recently attended a party, and/or been recently offered alcohol, which suggest more independence and social activity. They are also more likely to think their close friends have tried shisha, that their friends would approve of them smoking shisha, and to say that they would not likely refuse an offer of shisha from friends. Several indicators included in the evaluated program's theory of change, particularly those related to girls' confidence in expressing preferences to friends, were not important to classifying likely smokers.

Given that smoking in this population is rare and infrequent, Model #2 (Table 3) identifies which factors are most important to differentiate recent smokers from recent non-smokers. Endline responses were used given very few recent smokers at baseline. Recent smokers are more likely to be older and live in the capital Accra and report lower parental influence than non-smokers, another sign of greater independence. They are more likely to report that smoking is common among their peers and that they would accept an offer to smoke from friends, and less likely to perceive shisha as harmful. The belief that people who say no are admired by others is one of the top factors predictive of being a recent smoker; this may suggest positive views of smoking as demonstrating independence or rebelliousness against norms.

From Model #3 (Table 4) we profiled the small number of girls who were non-smokers at baseline but reported they had tried smoking by endline, in order to identify which factors are most important to set them apart from girls who remained non-smokers during the study period. We see that age and phone access were most important to subsequent smoking initiation, which suggests that these girls may already have had more independence and connectedness than other non-smoking girls. The girls that started smoking within the study period were also more likely to perceive that most social activities for girls their age involve shisha and that most girls their age have tried shisha, which suggests that they are also experiencing or perceiving a different social environment at baseline

Table 2. Model 1, reported ever smoking at baseline.

Variable Input	Importance
Went to a party last 30 days (+)	0.128
Thinks most friends smoke shisha (+)	0.084
Has regular access to phone (+)	0.069
Would refuse shisha from friends (-)	0.053
Was offered alcohol last 30 days (+)	0.043
Friends would approve smoking shisha (+)	0.040
Thinks most other girls smoke shisha (+)	0.040
Lives in Accra (+)	0.034
Age (+)	0.033
Friends would shun if smoked shisha (-)	0.032
Friends would shun if smoked cigarette (-)	0.032
Went to a bar last 30 days (+)	0.028
Would refuse shisha from boyfriend (-)	0.024
Lives with both parents (-)	0.022
Believes smoking shisha is harmful (-)	0.022
Friends would approve smoking (+)	0.021
In school (-)	0.019
Parents influence their decisions (-)	0.017
Has close circle of friends (-)	0.017
Believes not smoking shisha is important (-)	0.016
Thinks people who say no are admired (+)	0.014
Friends would shun if said no (+)	0.013
Boyfriend would shun if said no (+)	0.013
Friends influence their decisions (+)	0.012
Others think smoking cigarettes is guy-guy (+)	0.012
Most activities involve smoking shisha (+)	0.012
Others think smoking shisha is guy-guy (+)	0.011
Girls like other confident girls (-)	0.011
Boys like confident girls (-)	0.010
Most activities involve smoking cigarettes (+)	0.010
Believes smoking cigarettes is harmful (-)	0.009
Boys like girls who smoke shisha (+)	0.009
Would refuse cigarette from friends (-)	0.007
Feels comfortable expressing likes/dislikes (-)	0.007
Stands firm with friends (-)	0.006
Would refuse cigarette from boyfriend (-)	0.005
Thinks most other girls smoke cigarettes (+)	0.005
Boys like girls who smoke cigarettes (+)	0.005
Can tell friends if uncomfortable (+)	0.005
Believes not smoking cig is important (-)	0.003
Think most friends smoke cigarettes (+)	0.003
Other girls influence their decisions (+)	0.001

Note: +/- indicate whether that variable is associated with increased or decreased risk of smoking

Table 3. Model 2, recent smoker at endline (abbreviated).

Variable Input	Importance
Age (+)	0.120
Thinks most other girls smoke cigarettes (+)	0.108
Lives in Accra (+)	0.106
Would refuse cigarette from friends (-)	0.090
People who say no are admired (+)	0.079
Parents influence decisions (-)	0.060
Thinks smoking shisha is harmful (-)	0.045
Went to a party last 30 days (+)	0.035
Friends would shun if smoked shisha (-)	0.031
Has close circle of friends (-)	0.030
Think most friends smoke shisha (+)	0.029
Lives with both parents (-)	0.027
Most activities involve smoking shisha (+)	0.022
Most activities involve smoking cigarettes (+)	0.018

Note: +/- indicate whether that variable is associated with increased or decreased risk of smoking

than other non-smokers. And these girls are more likely to report their friends have influence on their decisions, and less likely to report parent influence than other non-smokers, which suggests they might be more susceptible to social opportunity and persuasion.

In secondary analyses of Model 3, living in Accra and thinking other girls smoke shisha were the most important factors for smoking initiation among younger girls (13–15) who were non-smokers at baseline. Having friends that influence decisions was less important, and school status was one of the least important, likely because very few younger girls are

Table 4. Model 3, non-smoker at baseline, smoked by endline (abbreviated).

Variable Input	Importance
Age (+)	0.129
Has regular access to phone (+)	0.079
Friends influence decisions (+)	0.053
In school (-)	0.050
Lives with both parents (-)	0.041
Most activities involve smoking shisha (+)	0.037
Parents influence decisions (-)	0.035
Think other girls smoke shisha (+)	0.033
People who say no are admired (+)	0.031
Has close circle of friends (-)	0.030
From Accra (+)	0.028
Most activities involve smoking cigarettes (+)	0.028
Friends would shun if said no (+)	0.027

Note: +/- indicate whether that variable is associated with increased or decreased risk of smoking

not in school. Among older girls, having friends or parents who influence decisions were the most important factors. This suggests that the spaces and behaviors girls are exposed to may be more relevant for smoking initiation among younger girls, whereas interpersonal influence matters more among older girls.

Conclusion/discussion

We report here a novel application of ML techniques to generate insights around a behavior that is still relatively rare among urban teen girls in Ghana. Creating synthetic data through the SMOTE technique created a more balanced training data set for the classifier, and by opening up the black box of the algorithm we learned which features are most important towards predicting girls likely to be smokers.

While very few girls in Ghana have tried smoking, even fewer report having smoked in the previous 30 days. Models 1 and 2 suggest both groups are more likely to be older and have the freedom to attend social events where they may be exposed to smoking opportunities or behaviors that shape their perceptions of norms among peers. Living in Accra, the more cosmopolitan and diverse of the two study cities, was also common among both groups. However, a perceived higher prevalence of cigarette smoking and willingness to accept cigarettes from friends are distinctively strong predictors of recent smokers. This may reflect that shisha is more commonly experimented with among girls this age and is more likely to be smoked on an infrequent basis (less than monthly). For that reason, most strong predictors of ever having smoked reflect beliefs about shisha, whereas recent smokers who smoke cigarettes, while still the minority, may have particularly strong and distinctive views on cigarette smoking. Other distinctively strong predictors of recent smoking include reporting low influence of parents and expressing admiration for people who are willing to go against trends. Parents in Ghana are generally highly protective of their teenage girls and there is a strong social norm against smoking[2]; these findings may reflect the absence of such parental figures in smokers' lives, or their willful rebellion against them.

Model 3 supports that changes in independence and the social environments shaping girls' opportunities to smoke do in fact precede initiation and experimentation. Cross-sectional surveys can support associations between smoking behavior and social environment; the Model 3 classifier leverages survey data from two timepoints to highlight the pre-existing differences among non-smoking girls that are most predictive of subsequent smoking initiation. Among younger girls, this may reflect different home environments, where girls have more mobility and less parental oversight, putting some girls at more risk of smoking opportunities. These differences may also explain why older age is such a strong predictor of smoking, as adolescent girls are often given more freedom to attend social events and activities as they age. Furthermore, many girls in urban Ghana enroll in a boarding school for their Senior High School (SHS) education where they may gain exposure to girls from less sheltered backgrounds who may influence subsequent opportunities for smoking[2].

Such insights about the relative importance of different features to a target behavior can be valuable input for program planning and outreach that is responsive to a specific population and context. This is especially true in the case of programs aimed at smoking prevention, given that smoking in teens is addictive and the literature suggests that people who avoid smoking in adolescence are unlikely to ever start[vi]. Better understanding of the risk factors for recent and future smoking behavior, and in particular the role of the social environment, can suggest programmatic opportunities for targeting adolescents most at risk or exploring more promising programmatic directions, such as reshaping the environments they are exposed to.

Although the smoker profiles generated through this approach are limited by the inputs we provide the machine, i.e., our survey data, they were largely consistent with our formative qualitative research findings around influences on girls' behavior in urban Ghana[2]. The large sample size of this study obtained through cluster randomized sampling is one of its strengths, although there are potential limitations related to collecting data on sensitive topics from adolescent girls, as discussed in that study[2]. Data collection tools in both cases were informed by insights from the tobacco literature and behavioral science; we believe grounding survey data collection in strong behavioral science theory strengthened the utility of the ML outputs and resulted in better alignment of findings with more in-depth qualitative research techniques. This novel application of ML techniques demonstrates the potential synergism between

[vi] Tyas SL, Pederson LL. Psychosocial factors related to adolescent smoking: a critical review of the literature. Tob Control. 1998 Dec 1;7(4):409–20.

data science and behavioral science to generate insights about predictors of behavior and highlights the importance of basing quantitative data collection in behavioral theory, especially if opportunities for rich qualitative investigation are limited in other settings.

Data availability

Source data

The data that support the findings of this study are not publicly available due to privacy and ethical restrictions but are available from the corresponding author on reasonable request. A reasonable request would be from a legitimate party with a specific research objective, would not violate the privacy or ethical protections of participants, and would not require additional reformatting or repackaging of the data by the authors.

Extended data

The baseline survey questions are available here: https://doi.org/10.6084/m9.figshare.24581616[7]

This project contains the following extended data:

- Baseline Survey_Final.pdf

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Software availability

The RandomForestClassifier that was used for classification tasks belongs to the open-source Python scikit-learn library[8] available at https://github.com/scikit-learn/scikit-learn, with documentation here: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. The feature importances are provided by the "feature_importances" attribute of the scikit-learn's RandomForestClassifier, which offers a way to assess the importance of each feature in making accurate predictions with the random forest model.

All the oversampling methods were performed using open-source Python imbalanced-learn library[3] hosted at https://github.com/scikit-learn-contrib/imbalanced-learn with documentation here: https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.RandomOverSampler.html.

Acknowledgements

We wish to thank the survey participants and JMK Consulting Limited for assisting with data collection. We are also grateful to Good Business and Now Available Africa for their partnership and input into the survey instruments. We also thank Lois Aryee and Jeremy Barofsky of ideas42 for their contributions to the research activities of this project, and Jean Paullin of the Bill & Melinda Gates Foundation for her continued support and guidance on these efforts.

References

1. Jordan MI, Mitchell TM: **Machine learning: trends, perspectives, and prospects.** *Science.* 2015; **349**(6245): 255–260.
PubMed Abstract | Publisher Full Text

2. Aryee LNA, Flanagan SV, Trupe L, *et al.*: **Social norms and social opportunities: a qualitative study of influences on tobacco use among urban adolescent girls in Ghana.** *BMC Public Health.* 2024; **24**(1): 2978.
PubMed Abstract | Publisher Full Text | Free Full Text

3. Lemaitre G, Nogueira F, Aridas CK: **Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning.** *J Mach Learn Res.* 2017; **18**(1): 559–563.
Reference Source

4. He H, Bai Y, Garcia EA, *et al.*: **ADASYN: adaptive synthetic sampling approach for imbalanced learning.** In: *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence).* IEEE; 2008; 1322–1328.
Publisher Full Text

5. Chawla NV, Bowyer KW, Hall LO, *et al.*: **SMOTE: Synthetic Minority Over-sampling Technique.** *J Artif Intell Res.* 2002; **16**(1): 321–357.
Publisher Full Text

6. Cieslak DA, Hoens TR, Chawla NV, *et al.*: **Hellinger distance decision trees are robust and skew-insensitive.** *Data Min Knowl Disc.* 2012; **24**(1): 136–158.
Publisher Full Text

7. Flanagan S, Vargas A, Smith J: **Application of machine learning techniques to profile smoking behavior of adolescent girls in Ghana: baseline questionnaire.** *figshare.* [Data], 2023.
http://www.doi.org/10.6084/m9.figshare.24581616.v1

8. Pedregosa F, Varoquaux G, Gramfort A, *et al.*: **Scikit-learn: *Machine Learning in Python*.** *J Mach Learn Res.* **12**: 2825–2830.
Reference Source

Open Peer Review

Current Peer Review Status: ✔ ? ?

Version 2

Reviewer Report 26 August 2025

https://doi.org/10.21956/gatesopenres.17698.r39728

© **2025 Wang R.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Runqiu Wang
University of Nebraska Medical Center, Omaha, Nebraska, USA

The paper "Application of machine learning techniques to profile smoking behavior of adolescent girls in Ghana" applies machine learning (ML) methods, mostly random forest classifier, to identify predictors of smoking initiation among adolescent girls using data from a stepped wedge cluster randomized trial and panel survey involving 9,000 girls aged 13-19 years. The study leverages the Synthetic Minority Over-sampling Technique (SMOTE) to optimize model precision and recall and employs feature importance techniques to enhance interpretability.

Below are my comments:
1.Introduction: Strengthen the introduction by briefly discussing current tobacco use prevalence among adolescents in Ghana and the gender gap trend. Cite recent literature to contextualize the relevance and urgency of this issue, for example:
a. (Ref 1)
b. (Ref 2)

2. Methods:
(1).Suggest including results from a conventional logistic regression model as a baseline comparison. This would provide a useful benchmark and highlight the specific advantages or limitations of the Random Forest classifier in this context.
(2).While the paper presents itself as applying machine learning broadly, only the Random Forest classifier is used. The authors may consider exploring additional ML models: such as gradient boosting machines (e.g., XGBoost), support vector machines to assess model robustness and validate the consistency of key predictors. This would strengthen the claim that machine learning approaches (not just Random Forests) offer valuable insights in this context.
(3) The manuscript does not describe the validation strategy used for model training and evaluation. It is unclear whether cross-validation, train/test split, or another method was used to compute the reported metrics (accuracy, precision, recall). Clarifying the validation framework and specifying how these metrics were calculated would strengthen the credibility and reproducibility of the results.
(4) SMOTE should only be applied to the training set to avoid data leakage. The authors should

clarify whether SMOTE was applied correctly after splitting the data and not on the entire dataset. Improper use could compromise the validity of performance metrics and model generalizability.

3. Results:
The reported feature directions (positive or negative associations) are not derived from the Random Forest model itself, as standard feature importance metrics do not provide directionality. The authors should clarify how these directions were determined, e.g., through separate bivariate analyses, and discuss the limitations of inferring directionality outside the trained model context. Alternatively, the authors may consider using SHAP (SHapley Additive exPlanations). SHAP provides both the direction and magnitude of each feature's contribution to a prediction, offering a unified and theoretically grounded method for interpreting machine learning models.
4. Conclusion/Discussion
I have no comments here, authors did a good job for this part.

References
1. Aryee L, Flanagan S, Trupe L, Yucel M, et al.: Social norms and social opportunities: a qualitative study of influences on tobacco use among urban adolescent girls in Ghana. *BMC Public Health*. 2024; **24** (1). Publisher Full Text
2. D LOGO, D. I. V. I. N. E., Oppong, F. B., Singh, A., Amenyaglo, S., Wiru, K., T ANKRAH, S. A. M. P. S. O. N. Owusu-Dabo,: Profile and predictors of adolescent tobacco use in Ghana: evidence from the 2017 Global Youth Tobacco Survey (GYTS). *J Prev Med Hyg.* Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Biostatistics, Machine Learning and Deep Learning, multiple testing in high-dimensional data, infectious disease

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have

significant reservations, as outlined above.

Reviewer Report 11 December 2024

https://doi.org/10.21956/gatesopenres.17698.r38488

© **2024 Sun R.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Ruoyan Sun
The University of Alabama at Birmingham, Birmingham, Alabama, USA

The authors have addressed most of my previous comments. I have two minor questions.

1. In the Introduction, new references are added as footnotes (Footnote I to VI) but not cited. Is this because of limited references allowed (up to 8)? If there is space for more than 8 references, then the authors should cite these new references properly.

2. Tysa & Pederson (Footnote vi) was published in 1998 and smoking behaviors have changed substantially since then in many countries. Is there any more recent references that can be used?

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Tobacco control; policy evaluation; health economics; modeling and simulation.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 17 July 2024

https://doi.org/10.21956/gatesopenres.16324.r36854

© **2024 Murray J.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Jennifer M. Murray
Queens University Belfast, Belfast, England, UK

This Research Note applies machine learning (ML) techniques to identify the most important

variables for predicting initiation of smoking amongst 9000 adolescent girls (aged 13-19 years) participating in a social marketing program in Ghana. The data was collected through a stepped wedge cluster randomized trial with panel survey over four time-points during 12 months of 2021-2022. The results show the importance of adolescent girls' independence and connectivity, social environment, and peer influence for predicting previous smoking behavior, and subsequent smoking initiation. The study design and methods are appropriate for answering the research questions, and the work is technically sound. The statistical analysis and its interpretation are appropriate, and sufficient details of the methods and analysis have been provided to allow replication. The conclusions drawn are adequately supported by the results. The source data underlying the results are not publicly available due to privacy and ethical restrictions. However, the authors have said that they will provide the data on reasonable request. The authors have also provided a link to the survey questions in the "Extended data" section. Overall, the work is clearly and accurately presented. However, there is little reference to the current literature on adolescent smoking initiation or the behavioral theory underlying the study and outcome measures. I believe that the work is OK to be indexed in its current form, but I have chosen the "approved with reservations" approval status because I have several minor comments that I think could improve the published product.

Key strengths: Large sample size, methods are appropriate for answering the research questions, comprehensive dataset with outcome variables and predictors informed by behavioral theory and qualitative research findings.

Key weaknesses: Little discussion of the current literature on adolescent smoking or the relevant research on behavioral theory.

Comments:

1. There is little discussion on the current literature on adolescent smoking in the introduction section of the main text. For example, although the background section of the abstract mentions tobacco trends in low- and middle-income countries, and narrowing gender gaps, this is not expanded upon in the main text and there is no mention of the adolescent smoking prevalence (for girls in Ghana). I appreciate that the article is a Research Note and therefore has a limited word count, but I would still expect to see some reference to the issue of adolescent smoking rates in the introduction section.
2. Similarly, throughout the paper the authors highlight that their survey items (and the ML inputs) were informed by the formative/qualitative research findings, and the behavior change theory underlying the evaluated program. However, the qualitative findings or the behavioral theory have not been summarized anywhere. I think it would be interesting to see a brief overview of the social marketing program, and how it relates to the study's outcome measures, in the methods section.
3. At the end of the methods section, you state that you conducted some secondary analyses using the same data inputs and predicted outcomes within subgroups of the respondents (including younger versus older girls, and different regions), and also used different combinations of data inputs and predicted outcomes. However, you were not very specific about these analyses, and the results are not reported. I suggest you should add specific details about these analyses (e.g., what age groups and regions were compared, what alternative combinations of data inputs and predicted outcomes were used) and summarize the findings in the results section. You could also upload the results tables for the

secondary analyses to a repository and provide the link in the "Extended data" section. I do not think this should be left as it is because it is vague, and the results have not been discussed anywhere in the article. It would be interesting to see the results if you had used the "intentions for smoking cigarettes or shisha over the next 30 days" outcome at the end of the survey.

4. Similarly, in the methods section you state that you recoded some of your data's Likert scales to binary. This would cause some loss of information. Have you considered presenting alternative (sensitivity) analyses to determine the impact of using the original uncategorized outcome variables?
5. Although you have provided a link to the full survey in the "Extended data" section, I think you could provide some more information on the outcome measures in the methods section (e.g., insert the relevant references if they have been adapted from previous research studies, and describe whether they have been validated for use with your research population).
6. The study's strengths and limitations have not been discussed in the discussion section.
7. On page 4, where you provide an overview of the ML techniques, you state that the "RandomOverSampler" technique works by "randomly oversampling the minority class (non-smokers)". Are smokers not the minority class?
8. Where you describe the SMOTE technique (the optimal technique whose results you ultimately report on), you say that it works by generating samples based on the k-nearest neighbors of the minority samples. What was "k" in your models?
9. In your models, "lives in Accra" is frequently an important predictor. Can you comment on this result (e.g., the differences between Accra and the other communities in your sample, and why adolescents living in Accra should have increased risk of smoking)?

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Public health, health behavior change, physical activity, adolescent smoking, peer influence, mediation analyses, SIENA modelling

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 19 Nov 2024

Sara Flanagan

1. There is little discussion on the current literature on adolescent smoking in the introduction section of the main text. For example, although the background section of the abstract mentions tobacco trends in low- and middle-income countries, and narrowing gender gaps, this is not expanded upon in the main text and there is no mention of the adolescent smoking prevalence (for girls in Ghana). I appreciate that the article is a Research Note and therefore has a limited word count, but I would still expect to see some reference to the issue of adolescent smoking rates in the introduction section.
    - **Response**: This background info has been added to the Introduction section of the main text along with citations.
2. Similarly, throughout the paper the authors highlight that their survey items (and the ML inputs) were informed by the formative/qualitative research findings, and the behavior change theory underlying the evaluated program. However, the qualitative findings or the behavioral theory have not been summarized anywhere. I think it would be interesting to see a brief overview of the social marketing program, and how it relates to the study's outcome measures, in the methods section.
    - **Response**: The qualitative study has just recently been published and the citation has been updated. That paper includes more explanation of the context for the study and an extensive discussion of the findings relative to the literature.
3. At the end of the methods section, you state that you conducted some secondary analyses using the same data inputs and predicted outcomes within subgroups of the respondents (including younger versus older girls, and different regions), and also used different combinations of data inputs and predicted outcomes. However, you were not very specific about these analyses, and the results are not reported. I suggest you should add specific details about these analyses (e.g., what age groups and regions were compared, what alternative combinations of data inputs and predicted outcomes were used) and summarize the findings in the results section. You could also upload the results tables for the secondary analyses to a repository and provide the link in the "Extended data" section. I do not think this should be left as it is because it is vague, and the results have not been discussed anywhere in the article. It would be interesting to see the results if you had used the "intentions for smoking cigarettes or shisha over the next 30 days" outcome at the end of the survey.
    - **Response:** We have updated the methods to specify the age groups (13-15 vs. 16-19) and regions (Accra vs. Kumasi), and summarized the age subgroup results for model 3 at the end of the results section. We also edited the Methods so that it does not seem like we explored additional outcomes (we did not explore intention to smoke), rather we looked at the same outcomes (ever

smoking, recent smoking) at different time points.

4. Similarly, in the methods section you state that you recoded some of your data's Likert scales to binary. This would cause some loss of information. Have you considered presenting alternative (sensitivity) analyses to determine the impact of using the original uncategorized outcome variables?
    - **Response:** SMOTE uses Euclidean distance to figure out how close minority samples are to each other, but can struggle with high-dimensional or ordinal data, like raw Likert scales, because of a problem called “distance distortion.” In high-dimensional spaces, data points are more spread out, which makes their distances more similar and makes it hard to find true nearest neighbors. This distortion could make a sensitivity analysis confusing, since it will be difficult to know if any differences are from real predictive value or just from high-dimensional effects. Moreover, transforming the Likert scale variables was also to make interpretation of the feature importance technique easier, so using the original scales again may not add much value, given potential distortion, and could reintroduce complexity to the interpretation.
5. Although you have provided a link to the full survey in the "Extended data" section, I think you could provide some more information on the outcome measures in the methods section (e.g., insert the relevant references if they have been adapted from previous research studies, and describe whether they have been validated for use with your research population).
    - **Response:** We added a comment that the tobacco use questions were adapted from the Global Youth Tobacco Survey, and cited the formative research, which is the context in which the questionnaire was tested in this population.
6. The study's strengths and limitations have not been discussed in the discussion section.
    - **Response:** The final paragraph of the discussion mentions the limitations of data inputs and the strengths of grounding data collection in behavioral science. We have expanded there to mention the large sample size as a strength and referenced a longer discussion of limitations related to collecting data from this study population in our formative research paper, to save on space here.
7. On page 4, where you provide an overview of the ML techniques, you state that the "RandomOverSampler" technique works by "randomly oversampling the minority class (non-smokers)". Are smokers not the minority class?
    - **Response:** This was a typo and has been corrected, we thank the reviewer for catching it.
8. Where you describe the SMOTE technique (the optimal technique whose results you ultimately report on), you say that it works by generating samples based on the k-nearest neighbors of the minority samples. What was "k" in your models?
    - **Response:** We used 5 k-neighbors and added this detail to the description of the technique.
9. In your models, "lives in Accra" is frequently an important predictor. Can you comment on this result (e.g., the differences between Accra and the other communities in your sample, and why adolescents living in Accra should have increased risk of smoking)?
    - **Response:** We have added a comment to the results discussion that Accra is

the more cosmopolitan and diverse of the two study cities.

Competing Interests: No competing interests were disclosed.

Reviewer Report 16 May 2024

https://doi.org/10.21956/gatesopenres.16324.r36278

© 2024 Sun R. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Ruoyan Sun
The University of Alabama at Birmingham, Birmingham, Alabama, USA

Applying novel machine learning techniques to a panel survey of 9000 girls aged 13-19 in Ghana, this study identified important variables that predict smoking initiation among adolescent girls. Strengths of the study include clearly written descriptions of various machine learning approaches and model selection based on performance measures (accuracy vs prevision vs recall). However, there are a few problems.

1. There is a lack of information or justification of why smoking among adolescent girls is an important issue in Ghana.
   - In the background part of the abstract, the authors mentioned "..., and in particular narrowing gender gaps, highlight the need for interventions to prevent and/or reduce tobacco use among adolescent girls". It is not clear what narrowing gender gaps mean here. Are we saying boys have higher smoking prevalence and girls are catching up? If this is the case, then in the Introduction, the authors need to expand on this part with proper citations.
   - In the last paragraph of the Introduction, the authors mentioned "only 1.6% of girls reported ever tried smoking at baseline and 0.1% having smoked in the past 30 days." These rates are extremely low, especially compared to those in developing countries. For example, past 30-day cigarette smoking among middles school students in the US was 1.1% in 2023.[1] Why is smoking among adolescent girls in Ghana an important topic to study? Has the smoking prevalence increased in the recent years? The authors need to add more background information and justification.

2. While the machine learning techniques are cool and novel, it is not clear how the findings advance our existing knowledge. Many of the factors identified, such as peer influence (is smoking common among friends) and exposure to smoking (most activities involve smoking cigarettes), are well-known factors that are associated with smoking initiation. The authors need to compare their results with existing literature and highlight their contribution. This can be done by adding a paragraph or two in the Discussion.

3. To demonstrate the advantage or benefits of machine learning techniques, the authors could

conduct the same analysis using convention regression approaches and compare the results. For example, logistic regressions can also identify risk factors that are significantly associated with ever smoking or past 30-day smoking.

References

1. Birdsey J, Cornelius M, Jamal A, Park-Lee E, et al.: Tobacco Product Use Among U.S. Middle and High School Students - National Youth Tobacco Survey, 2023.*MMWR Morb Mortal Wkly Rep*. 2023; **72** (44): 1173-1182 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Tobacco control; policy evaluation; health economics; modeling and simulation.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 19 Nov 2024

Sara Flanagan

1. There is a lack of information or justification of why smoking among adolescent girls is an important issue in Ghana.
   - In the background part of the abstract, the authors mentioned "..., and in particular narrowing gender gaps, highlight the need for interventions to prevent and/or reduce tobacco use among adolescent girls". It is not clear what narrowing gender gaps mean here. Are we saying boys have higher smoking prevalence and girls are catching up? If this is the case, then in the Introduction, the authors need to expand

on this part with proper citations.
    - **Response**: We have expanded on this background information in the Introduction section and added citations.
- In the last paragraph of the Introduction, the authors mentioned "only 1.6% of girls reported ever tried smoking at baseline and 0.1% having smoked in the past 30 days." These rates are extremely low, especially compared to those in developing countries. For example, past 30-day cigarette smoking among middles school students in the US was 1.1% in 2023.[1] Why is smoking among adolescent girls in Ghana an important topic to study? Has the smoking prevalence increased in the recent years? The authors need to add more background information and justification.
    - **Response**: Other studies had suggested increasing tobacco use, and in particular shisha, among adolescent girls in Ghana. We have expanded on this background information in the Introduction section and added citations.

2. While the machine learning techniques are cool and novel, it is not clear how the findings advance our existing knowledge. Many of the factors identified, such as peer influence (is smoking common among friends) and exposure to smoking (most activities involve smoking cigarettes), are well-known factors that are associated with smoking initiation. The authors need to compare their results with existing literature and highlight their contribution. This can be done by adding a paragraph or two in the Discussion.
- **Response**: We recognize the word limit for research notes is a constraint to significantly expanding the Discussion, but we have edited the text to specify that these insights have programmatic value for planning interventions that are responsive to this specific population and context. We also encourage readers to refer to our recently published qualitative research study (cited) for a more extensive discussion of these findings relative to the existing tobacco literature.

3. To demonstrate the advantage or benefits of machine learning techniques, the authors could conduct the same analysis using convention regression approaches and compare the results. For example, logistic regressions can also identify risk factors that are significantly associated with ever smoking or past 30-day smoking.
- **Response**: Comparing our model against a logistic regression model may not be the best comparison, since Random Forests can better leverage the additional samples created through SMOTE as it builds multiple trees on different subsets of the data, which may help to generalize better. Random Forests also provide additional benefits over logistic regression. Random Forests provide insights into the importance of different predictors, allowing you to identify which features contribute most to the prediction, whereas logistic regression provides coefficients that can be harder to interpret, especially with interactions. Random Forests are also less prone to overfitting, especially in complex datasets, because they aggregate predictions from multiple trees, which helps to smooth out noise.

Competing Interests: No competing interests were disclosed.