scientific reports

Check for updates

OPEN

Exploring explainable machine learning algorithms to model predictors of tobacco use among men in Sub Sahara Africa between 2018 and 2023

Mequannent Sharew Melaku[1]✉, Nebebe Demis Baykemagn[1], Lamrot Yohannes[2] & Adem Tsegaw Zegeye[1]

Tobacco smoking is a significant public health issue in sub-Saharan Africa, with its prevalence shaped by various demographic factors. This study aimed to model predictors of tobacco use among men in Sub Sahara Africa between 2018 and 2023 using machine learning algorithms. Data from Demographic and Health Surveys covering 147,466 men were analyzed. STATA version 17 was used for data cleaning and descriptive statistics, while Python 3.9 was employed for machine learning predictions. The study utilized several machine learning models, including Decision Tree, Logistic Regression, Random Forest, KNN, eXtreme Gradient Boosting (XGBoost), and AdaBoost, to identify the key predictors of tobacco use among men. Hyperparameter optimization was performed using Randomized Search with tenfold cross-validation, enhancing model performance. The Additive Explanations (SHAP) method was used to assess predictor significance. Model performance was evaluated based on accuracy, precision, recall, F1 score, and area under the curve (AUC). The study found a pooled tobacco use prevalence of 14.73%, with no significant variation between countries. High tobacco use was observed in Mozambique, Zambia, Benin, Mali, Mauritania, Senegal, Guinea, Sierra Leone, and Liberia, with Tanzania, Benin, and Senegal reporting the highest rates. The XGBoost algorithm attained an accuracy of 98% and an AUC score of 97%. SHAP analysis revealed that age, education, wealth index, religion, residence, internet use, occupation, age at first sex, number of sexual partners, and marital status were key predictors. These findings underscore the need for targeted public health interventions and highlight the value of machine learning in identifying at-risk populations and addressing socio-cultural and economic factors influencing tobacco use.

Keywords Determinants, Tobacco smoking, Smokeless tobacco, Prediction, Sub-Saharan Africa

Abbreviations

DHS	Demographic and health survey
WHO	World Health Organization
DALYs- Disability	Adjusted life years
UNFPA	United Nations Population Fund
IMF	International monetary fund
IR	Individual record
AUC-ROC	Area under the receiver operating characteristic curve
SMOTE	Synthetic minority over-sampling technique
XGBoost	Extreme gradient boosting
AUC	Area under the curve
SHAP	SHapley additive exPlanations
RFE	Recursive feature elimination

[1]Department of Health Informatics, Institute of Public Health, University of Gondar, Gondar, Ethiopia. [2]Department of Environmental and Occupational Health and Safety, Institute of Public Health, College of Medicine and Health Science, University of Gondar, Gondar, Ethiopia. ✉email: Mequannent.sharew@uog.edu.et

SSA	Sub Sahara Africa
SDGs	Sustainable development goals

Tobacco use remains a major global health threat, especially in low and middle income countries, where over 80% of the world's tobacco user reside[1–5]. Despite a clear evidence of its risks, more than one billion people continued to use tobacco. Tobacco consumption is responsible for approximately 8 million deaths each year and contributes to 6.9% of total years of life lost and 5.5% of disability-adjusted life years[6,7]. If the present pattern persists, smoking may result in up to 1 billion deaths by the end of this century[8]. In addition, tobacco use causes an annual global economic loss exceeding 1.5 trillion dollars[9,10]. Although Sub-Saharan Africa (SSA) presently records the lowest tobacco consumption rates globally, it is witnessing the most rapid increase in tobacco use[11–14]. Tobacco users in the region increased from 66 million in 2015 to 84 million by 2025, the fastest growth among all world health organization (WHO) regions[15,16]. Tobacco use is increasing across Sub-Saharan Africa, with smoking prevalence varying significantly from as low as 4% in Ghana to as high as 27.2% in Lesotho[13,17]. Tobacco use expected to reach an epidemic levels in SSA by 2040[18].

In SSA, tobacco use varies widely across countries and cultures, with a range of products including cigarettes, pipes, chewing tobacco, snuff (nasal and oral), kreteks, cigars, cheroots, cigarillos, and water pipes[19]. Broadly, tobacco products can be classified into two main categories: smoking tobacco (including cigarettes, cigars, and water pipes) and smokeless tobacco such as chewing tobacco and various forms of snuff[19–21]. The tobacco industry continues to view Sub Saharan Africa as a promising market due to the region's large and growing youth population[16]. Moreover, markets in SSA are often unregulated, cigarette prices are low, and tobacco control laws are either weak or not fully implemented and enforced. Similarly, healthcare resources are limited, health infrastructure is poor[7], and industry interference is prevalent. Additionally, lack of effective national tobacco governance and political will to enforce legislation could worsen health outcomes if the tobacco epidemic spreads in the region[22–24].

Tobacco and nicotine use are associated with numerous health conditions, including asthma, wheezing, and cancer. Tobacco use is a major contributor to the global disease burden, being linked to around 71% of lung cancer cases, 42% of chronic respiratory disorders, and 10% of cardiovascular diseases[25]. Similarly, tobacco use increases the risk of respiratory diseases among pregnant women in their unborn babies[26,27]. Moreover, tobacco use poses high level of health risks to others through secondhand smoke exposure[28,29]. Previous findings have been highlighted the various determinant factors of tobacco use. Socioeconomic status, religion, marital status, wealth, education, and residence are among the factors identified in various studies[8,17,21,30–36].

Moreover, tobacco use is more prevalent among men than women, with 47% of men and 11% of women reporting that they are using tobacco[37]. Gender difference in tobacco use is influenced by societal norms that discourage smoking among women while associating traits like defiance and risk-taking with men, shaping each gender's perception of tobacco's risks and perceived benefits[38]. Despite ongoing efforts to reduce tobacco use, it remains widespread among men in SSA. However, the factors driving this issue remain insufficiently examined. Only a handful of nationally representative studies based on Demographic and Health Surveys (DHS) data have focused on describing and modeling the factors associated with tobacco use in the region. A study conducted among 14 SSA countries (2000–2006)[39] reported higher adjusted odds of tobacco use among men[11]. On the other hand, a time-trend analysis of DHS data revealed a steady decline in socioeconomic disparities in tobacco use across Sub-Saharan African countries from 2003 to 2019[12].

Although previous studies have identified key determinants of tobacco use, they were limited by small study areas and predominantly relied on traditional statistical methods[40–42]. However, these conventional methods have several limitations, such as the assumption of linearity, difficulty in handling complex interactions between variables, and reliance on manual feature engineering. It also performs less well with large, high-dimensional datasets and may struggle with non-linear relationships. Despite this, the complex interplay among these determinants has yet to be rigorously explored using machine learning techniques applied to comprehensive datasets. Machine learning, with its superior predictive capabilities and ability to capture complex, nonlinear relationships, offers considerable potential for uncovering these intricate dynamics. This approach could provide crucial insights necessary for advancing effective tobacco control strategies at the regional level. In analyzing tobacco use, it reveals hidden patterns and interactions often missed by traditional methods. Unlike conventional approaches, it handles large, noisy, and high-dimensional datasets, uncovering trends, subgroup differences, and dynamic influences. Tools like SHAP improve model interpretability, making results more transparent. As shown in practice, machine learning is a powerful and convenient method for analyzing complex data and uncovering insights beyond the reach of traditional statistical techniques[42]. Hence, this study aimed to model predictors of tobacco use among men in SSA between 2018 and 2023 using machine learning algorithms.

Methodology

Study design and study period

Numerous Sub Saharan African countries collaborate with international organizations, including the United Nations Population Fund (UNFPA), to conduct comprehensive Demographic and Health Surveys (DHS). Typically, these surveys are conducted every five years, while Mini-DHS surveys have been conducted at intervals of two to three years since 1990. Data collection in each country followed a cross-sectional study design. The present study utilized secondary data from Demographic and Health Surveys (DHS) conducted between 2018 and 2023. A design science approach was applied for the in-depth analysis of DHS data gathered during this period in SSA.

Study area

SSA consist of 48 to 51 countries found in the south of the Sahara Desert, had an estimated population of 1.26 billion in 2023. The population is projected to approach 2 billion by 2043, driven by high fertility rates and a predominantly young demographic, with many individuals under the age of 15 https://www.statista.com/statistics/805605/total-population-sub-saharan-africa/. Many SSA countries, in collaboration with organizations like the UNFPA, conduct Demographic and Health Surveys (DHS) every five years, with Mini-DHS surveys every two to three years. Accordingly, DHS data conducted among 20 SSA countries (Kenya, Madagascar, Mauritania, Rwanda, Tanzania, Ghana, Côte d'Ivoire, Burkina Faso, Liberia, Senegal, Sierra Leone, Gabon, Gambia, Zambia, Cameroon, Benin, Mali, Mozambique, Guinea, and Nigeria) between 2018 to 2023 included in this study (Fig. 1).

Source and study population

The source population comprises all men aged 15 to 64 years who are interviewed for tobacco use variable in the most recent DHS surveys across Sub Saharan African countries. The study population includes men within the same age group who are interviewed for tobacco use variable in the selected country within the region.

Study variables

Outcome variable

The main outcome variable in this study was tobacco consumption among men. This was determined based on self-reported consumption of various tobacco products such as cigarettes, pipes, chewing tobacco, nasal snuff, oral snuff, kreteks, cigars/cheroots/cigarillos, water pipes, and other country-specific forms of tobacco. Men who reported using at least one of these products were categorized as 'Tobacco Users,' while those who did not use any were classified as 'Non-Tobacco Users.

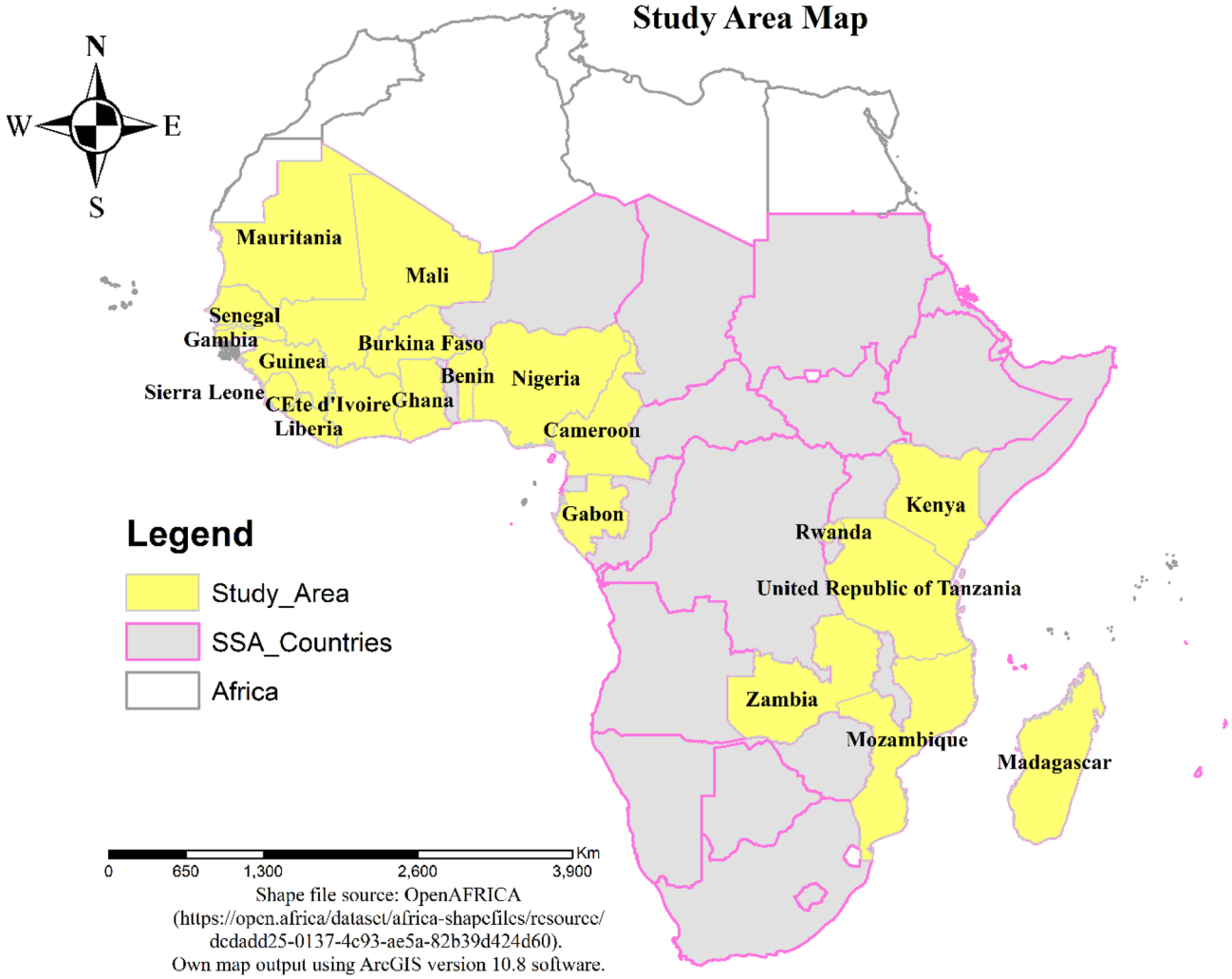

Fig. 1. Study area map of SSA for tobacco use among men (Shape file source: OpenAFRICA. (https://open.africa/dataset/africa-shapefiles/resource/dcdadd25-0137-4c93-ae5a82b39d424d60). Own map output using ArcGIS version 10.8 software.)

Independent variables

The independent variables were divided into two categories: factors at the individual level and factors at the community level. Individual-level factors comprised age, religion, occupation, literacy status, marital status, educational attainment of men, age at first sexual intercourse, number of sexual partners, household wealth status, and exposure to media. Community-level factors included place of residence, overall men's literacy within the community, community-wide media exposure, and the level of poverty at the community level.

Sample size determination and sampling procedure

The DHS program adopts standardized methods involving uniform questionnaires, manuals, and field procedures to collect data that is comparable across countries in the world. A total of 147,466 (weighted) men aged between 15 to 64 years among 20 countries in sub Saharan Africa included in this study. Households were selected through a stratified multi-stage cluster sampling approach, starting with the random selection of enumeration areas within census strata, followed by random household selection. The men's questionnaire collected information on men's health and household characteristics. This study used aggregated data from Demographic and Health Surveys carried out in 20 SSA countries between 2018 and 2023. Out of the 48 SSA countries initially considered for this study, a systematic selection process was employed to ensure the inclusion of countries with recent, standard, and complete Demographic and Health Survey (DHS) data collected between 2018 and 2023. Countries were excluded for various reasons, including the absence of publicly available DHS data (such as Djibouti, Somalia, South Sudan, Mauritius, Reunion, Botswana, Equatorial Guinea, and Seychelles) or restricted access to datasets, as in the case of Eritrea. Additionally, countries like Angola, Burundi, South Africa, Zimbabwe, Chad, the Democratic Republic of Congo, Congo, Namibia, Comoros, Central African Republic, Swaziland, Sao Tome and Principe, Cape Verde, and Sudan were excluded due to outdated survey data. Furthermore, Ethiopia, Malawi, Niger, Togo, and Uganda were omitted because they either conducted Mini DHS surveys during the study period or lacked key variables necessary for this analysis. After applying these criteria, 20 countries with complete and recent standard DHS data were included: Benin, Burkina Faso, Cameroon, Côte d'Ivoire, Gambia, Ghana, Gabon, Guinea, Kenya, Lesotho, Liberia, Madagascar, Mali, Mozambique, Nigeria, Rwanda, Senegal, Sierra Leone, Tanzania, and Zambia. The final pooled dataset consisted of 147,466 men, with 62,639 residing in urban areas and 84,827 in rural areas (Fig. 2; Table 1).

Data collection tool and procedures

Since the early 1990s, the DHS has been a nationally representative household survey in developing nations, conducted via questionnaire. The five questions employed for DHS, were the household questionnaire, the woman's questionnaire, the man's questionnaire, the biomarker questionnaire, and the health facility questionnaire. Survey techniques, sampling plans, questionnaires, and data collection and processing are consistent across countries. Households and survey respondents typically selected using a stratified, two-stage cluster sampling technique. The men's Questionnaire was designed to collect data from all eligible men aged 15 to 64 on various issues, including questions about men's health including tobacco use. Relevant household and demographic variables determining tobacco use was collected. Data for this study gathered from the men record file of the DHS conducted in SSA countries.

Data quality control

The questionnaire was pretested before the survey in regions not included in the actual sample. This pretest served as part of the training for data collectors and supervisors before fieldwork. It aimed to assess the training agenda, data collection instruments, and the Computer-Assisted Personal Interviewing (CAPI) approach. Additionally, it evaluated personnel competence, workload, training procedures, and the feasibility of the data collection timeline. The pretest also reviewed administrative, financial, and logistical aspects, tested data transmission reliability and quality monitoring systems, and assessed the effectiveness of publicity, advocacy, and data processing strategies. To maintain data integrity, clusters used for the pretest were excluded from the main survey. After field practice, a debriefing session was conducted with the pretest field staff to identify challenges and areas for improvement. Based on insights gained, necessary modifications were made to the questionnaires. Feedback was provided to individuals and teams both during the debriefing session and in daily discussions. Using paper questionnaires from the field practice, participants were trained on the electronic data entry system, which was programmed on tablet computers. The conclusions drawn from the debriefing process were used to refine the questionnaires, CAPI program, and field logistics before launching the main training and data collection. This study used STATA version 17 software to extract, clean, and recode data. The aggregated data was weighted to ensure the sample data representativeness.

Data management and analysis

Data extraction, cleaning, and analysis were conducted using STATA version 17, utilizing the Men Record (MR) dataset. Prior to statistical analysis, the data were weighted using the sampling weight, which was calculated by dividing the men sample weight (mv005) by 1,000,000 to ensure the survey's representativeness. Descriptive statistics were presented as frequencies, percentages, and text, through tables and graphs.

Data analysis for machine learning algorithms was conducted using Python version 3.9 software. Missing data were imputed using median values for numerical and mode for categorical variables. Outliers were identified and removed using the Local Outlier Factor with a 5% contamination setting. One-hot encoding was used to convert categorical variables into dummy variables, while numerical features were standardized through Z-score normalization to ensure consistent scaling. To mitigate class imbalance, the Synthetic Minority Over-Sampling Technique (SMOTE) was applied, adjusting the initial 4:6 minority-to-majority class ratio to a balanced 1:1 distribution prior to model training. Feature selection was conducted through a hybrid approach: Recursive

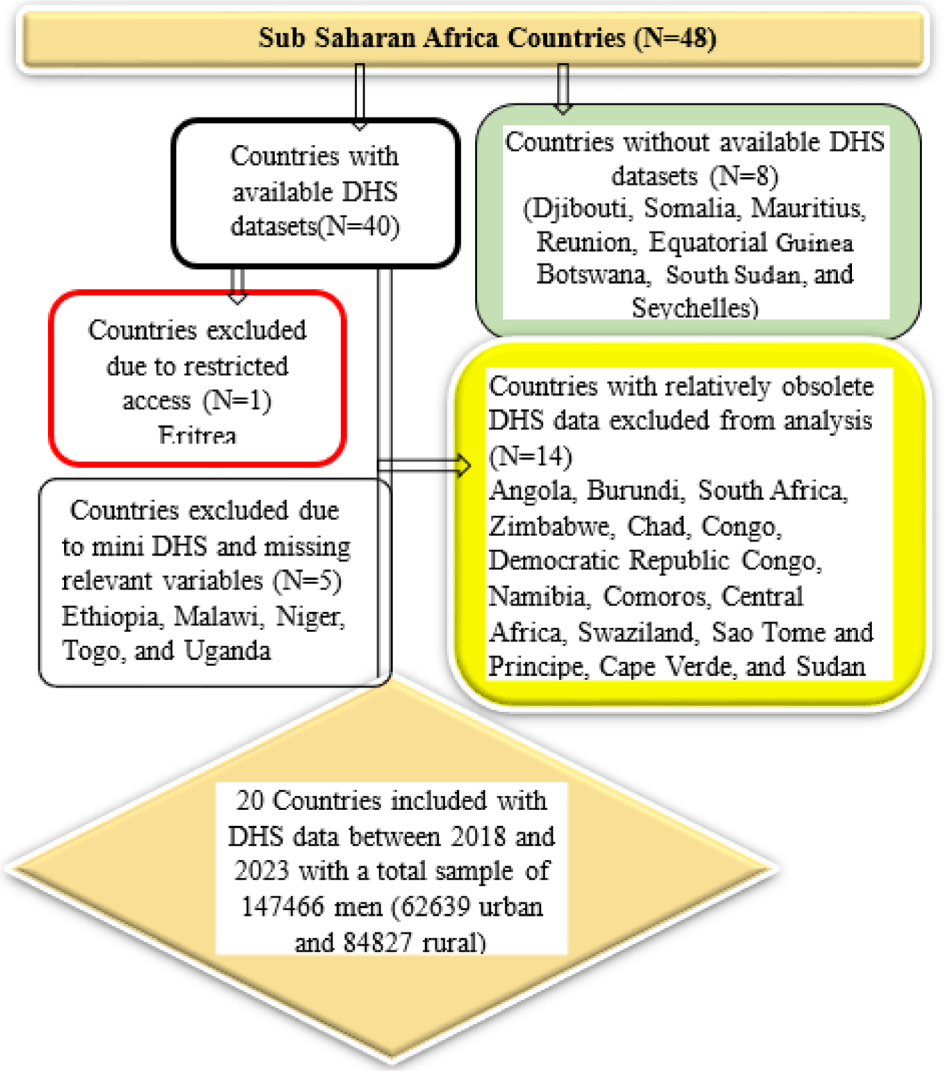

Fig. 2. Shows the detailed process of included countries and the total weighted sample.

Feature Elimination with logistic regression (L2 regularization) as the base estimator, iteratively removing 5 features per step. The best feature set was selected by maximizing AUC with fivefold cross-validation, while RFE removed unnecessary features to improve model efficiency.

Model selection and development

Six classification algorithms namely **AdaBoost Classifier**, XGB Classifier, Random Forest (RF), KNN, Extreme Gradient Boosting (XGBoost), Decision Tree were selected to balance interpretability and predictive power[43–47]. Logistic regression was used as a baseline for comparison, while ensemble techniques like Random Forest and XGBoost were selected for their strength in capturing non-linear patterns, managing high-dimensional data, and reducing overfitting via bagging and boosting. Simpler models such as decision trees were also evaluated to explore the balance between model complexity and performance. To assess the performance of the model, we used an 80:20 train-test split, where 80% of the dataset was allocated for training and the remaining 20% for testing. Additionally, we applied tenfold cross-validation to ensure robust evaluation by dividing the training data into 10 subsets, iteratively training the model on nine subsets while using the remaining one for validation. This approach helps mitigate overfitting and provides a more reliable estimate of the model's generalization performance.

Model optimization

In this study, hyperparameter optimization was carried out using Randomized Search with stratified fivefold cross-validation, a computationally efficient and robust method for tuning model parameters. This approach enables a stochastic exploration of the hyperparameter space, allowing the model to avoid local optima and identify parameter configurations that yield superior generalization performance. By randomly sampling from predefined distributions of hyperparameter values and validating model performance across multiple folds, this method balances computational efficiency with thoroughness in optimization. Key hyperparameters such as

SSA countries	Unweighted sample size	Weighted sample size	Residence	
			Urban	Rural
Burkina Faso	7720(5.2%)	9716(6.6%)	5580	4135
Benin	7595(5.2%)	8101(5.5%)	3678	4422
Cote Divoire	7591(5.2%)	8959(6.1%)	5025	3934
Cameroon	6978(4.7%)	8270(5.6%)	4161	4109
Gabon	6894(4.7%)	6787(4.6%)	3181	3606
Ghana	7044(4.8%)	7512(5.1%0	2022	5491
Gambia	4636(3.1%)	4889(3.3%)	2093	2796
Guinea	4117(2.8%)	3751(2.5%)	1469	2283
Kenya	14,453(9.8%)	13,130(8.9%)	5035	8094
Liberia	4249(2.9%)	3874(2.6%)	1546	2328
Madagascar	9037(6.1%)	7633(5.2%)	3174	4460
Mali	4618(3.1%)	4210(2.9%)	1640	2570
Mauritania	5673(3.9%)	4948(3.4%)	2309	2638
Mozambique	5380(3.7%)	5297(3.6%)	1669	3628
Nigeria	13,311(9.0%0	12,146(8.2%)	4228	7918
Rwanda	6513(4.4%)	6284(4.3%)	1698	4586
Sierra Leone	7197(4.9%)	7064(4.8%)	2459	4605
Senegal	6321(4.3%)	5338(3.6%)	1680	3657
Tanzania	5763(3.9%)	5156(3.5%)	1795	3361
Zambia	12,132(8.2%)	14,402(9.8%)	8196	6206
Total	147,222	147,466	62,639	84,827

Table 1. Shows the detailed presentation of sample size determination with urban rural stratification among the included countries.

the number of estimators, maximum tree depth, minimum samples per leaf, and learning rate (for boosting algorithms) were systematically tuned for each algorithm. The optimal parameter sets were selected based on maximizing the Area Under the Receiver Operating Characteristic Curve (AUC) on the validation sets, ensuring a model that not only fits the training data well but also maintains high discriminative power on unseen data.

Performance metrics

To evaluate the effectiveness of the models, we used a comprehensive set of performance metrics, ensuring a balanced assessment of predictive accuracy and reliability. Accuracy, precision, recall, and AUC-ROC were used to evaluate model performance, offering a balanced view of prediction quality and class separation, especially for imbalanced data.

Shapley additive explanations (SHAP)

The relationship between predictors and the outcome was analyzed using SHAP (Shapley Additive Explanations), which also highlighted the most influential variables in predicting tobacco use[48]. SHAP analysis uses a game theory framework to offer a global or local interpretation and explanation for any machine learning model's prediction[48].

Association rule mining

Association rule mining is among the most important and popular data mining techniques, used for discovering hidden patterns and relationships based on specific confidence intervals and lift, thereby addressing limitations in feature selection[49,50]. The Apriori algorithm is a widely used method in association rule mining, designed to uncover hidden patterns and relationships in large datasets. It operates by identifying frequent item sets that meet a minimum support threshold (1 to 5%) and generating association rules that satisfy confidence ($\geq 20\%$) and lift (≥ 1) criteria. The algorithm leverages the "Apriori property," which states that all subsets of a frequent item set also be frequent, enabling efficient pruning of the search space and reducing computational load[51,52].

Candidate item sets are iteratively generated and evaluated. Rules are constructed from these item sets by dividing them into antecedent and consequent parts, and their strength is measured using support, confidence, and lift. Lift values greater than 1 indicate positive associations. These thresholds help exclude weak or coincidental rules, focusing on significant, actionable insights[53,54].

Meta-analysis

Prevalence and standard error for each country were calculated using the binomial distribution formula. Heterogeneity was assessed using the chi-square test, I^2 statistic, and p-value. Subgroup analysis was conducted based on region (East, West, South, and Central Africa) and world bank economic classification of the country. A forest plot illustrated prevalence estimates with 95% confidence intervals, where box sizes reflected study weights, and lines indicated confidence intervals.

Heterogeneity

Heterogeneity in meta-analysis is commonly assessed using Cochran's Q test and the I^2 statistic. A Q test *p*-value < 0.10 suggests moderate heterogeneity. The I^2 statistic shows the percentage of variation due to differences between studies, ranging from 0% (none) to 100% (high), with higher values indicating more inconsistency[55]. When heterogeneity is significant and high, it may indicate that there is a significant variation between studies and may suggest the choice of the model. Assessing heterogeneity is essential for drawing a valid conclusion from meta-analyses.

Ethical consideration

This study involved secondary analysis of publicly accessible DHS data. All data collection procedures were approved by the IRB, ensuring that no participants, households, or communities could be identified. The datasets contain no personal names or household addresses, maintaining strict confidentiality standards as outlined by the MEASURE DHS Program, which prioritizes respondent privacy. The data is found at MEASURE DHS website (https://dhsprogram.com/data/dataset_admin/login_main.cfm;jsessionid=595E1DC0FC22A8C647F89A2368E7906C.cfusion?CFID=1242368&CFTOKEN=c892a8da9855f981-8D71EDAA-BF68-5950-1D6BA7F0BB37D5E8).

Result

This study contained a total of **147,466** (weighted) study participants. The largest number of the study participants were from Zambia, 14,402(9.8%) while the smaller study participants were from Guinea, 3751(2.5%). The overall tobacco use among men is 14.73% ranges from 1280(8.9%) in Zambia to 874(17.0%) in Tanzania. In addition, the highest cluster of tobacco use was detected in Mozambique, Zambia, Benin, Mali, Mauritania, Senegal, Guinea, Sierra Leone, and Liberia while the lowest is clustered in Zambia (Fig. 3). Among the study participants

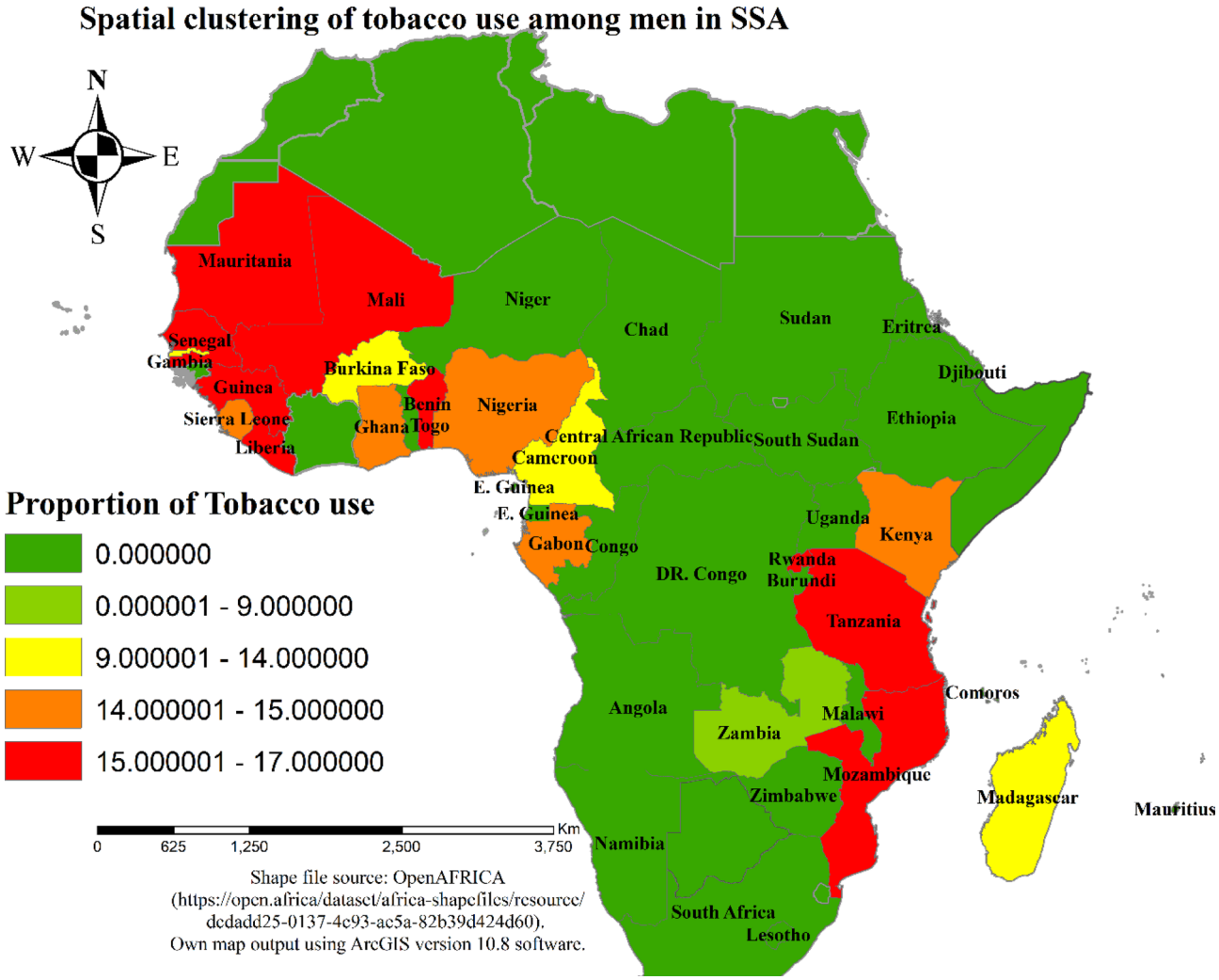

Fig. 3. The spatial distribution of tobacco uses among men in SSA (Shape file source: OpenAFRICA (https://open.africa/dataset/africa-shapefiles/resource/dcdadd25-0137-4c93-ae5a82b39d424d60). Own map output using ArcGIS version 10.8 software.)

Tanzania 874(17.0%), Bennie 1361(16.8%), and Senegal, 864(16.2%) were the largest number of tobacco use in Sub Saharan African countries (Table 2).

The prevalence of tobacco uses among men

This study revealed that the pooled prevalence of tobacco use among men in sub-Saharan Africa is 14.73 with a confidence interval of 11.26 to 18.19 with no heterogeneity between countries (I2 = 0.0, p value = 1.000) (Fig. 4). Subgroup analysis also conducted by region and income classification to detect if there is any source of heterogeneity.

Subgroup analysis

A subgroup analysis has been performed based on geographical regional location and world bank income classification of SSA countries. Regarding the geographical location, the maximum proportion was observed in East Africa with proportion of 15.45 and the least was observed in Central Africa region with the proportion of 11.53 (Fig. 5). Heterogeneity was not detected within groups and between groups. A subgroup analysis by world bank income classification result revealed that low income countries, 15.16 had a relatively highest proportion of tobacco users compared to middle income countries 14.68. Heterogeneity was not detected (Fig. 6).

Sociodemographic characteristics of the study participant

The majority of study participants were between the ages of 15–24 and 25–34 accounting for around 54,428(36.9%), and 38,013(25.8%) of the total sample, respectively. Of the participants in the sample, about 84,827(57.5%) lived in rural and 62,639(42.5%) in urban areas. The majority of the study participants, 66,331(45.0%) were married. Among the study participants, 38.6% (56,960) had completed secondary education, 40.5% (59,734) identified as other Christian denominations, and 71.7% (105,781) reported having their first sexual experience between the ages of 15 and 19, representing the majority of the participants 118,050(80.1%) had media exposures, about 66,939(45.4%) of the subject were rich wealth status, and 77,730(52.7%) of the participants had high media exposure (Table 3).

Machine learning analysis result

Synthetic minority over-sampling technique (SMOTE)

During the first study of the dataset, a significant class imbalance was observed, with the majority class accounting for considerably more samples than the minority class. Predictions based on this skewness may be distorted, with the model performing well for the majority class 124,587(84.63%), and the minority class of smoked cigarettes 22,635 (15.37%). To address class imbalance, the Synthetic Minority Over-Sampling Technique (SMOTE) was applied. This widely used method creates synthetic examples of the minority class, enhancing its representation within the dataset. SMOTE uses class distribution balancing to improve the model's ability to recognize underlying patterns in the data. The first study revealed a considerable imbalance in class distribution, with the majority class having a much higher percentage of samples than the minority class. The minority class samples

SSA countries	Weighted sample size (%)	Tobacco use	
		Yes	No
Burkina Faso	9716(6.6%)	1330(13.7%)	8386(86.3%)
Benin	8101(5.5%)	1361(16.8%)	6739(83.2%)
Cote Divoire	8959(6.1%)	1371(15.3%)	7587(84.7%)
Cameroon	8270(5.6%)	1173(14.2%)	7097(85.8%)
Gabon	6787(4.6%)	996(14.7%)	5791(85.3%)
Ghana	7512(5.1%0	1116(14.9%)	6397(85.1%)
Gambia	4889(3.3%)	676(13.8%)	4213(86.2%)
Guinea	3751(2.5%)	595(15.9%)	3156(84.1%)
Kenya	13,130(8.9%)	1978(15.1%)	11,151(84.9%)
Liberia	3874(2.6%)	601(15.5%)	3273(84.5%)
Madagascar	7633(5.2%)	1090(14.3%)	6544(85.7%)
Mali	4210(2.9%)	659(15.7%)	3551(84.3%)
Mauritania	4948(3.4%)	774(15.6%)	4173(84.4%)
Mozambique	5297(3.6%)	842(15.9%)	4455(84.1%)
Niger	12,146(8.2%)	1808(14.9%)	10,338(85.1%)
Rwanda	6284(4.3%)	987(15.7%)	5297(84.3%)
Sierra Leone	7064(4.8%)	1053(14.9%)	6011(85.1%)
Senegal	5338(3.6%)	864(16.2%)	4474(83.8%)
Tanzania	5156(3.5%)	874(17.0%)	4282(83.0%)
Zambia	14,402(9.8%)	1280(8.9%)	13,122(91.1%)
Total	147,466	21,428	126,037

Table 2. The distribution of weighted sample size and tobacco use among men in SSA.

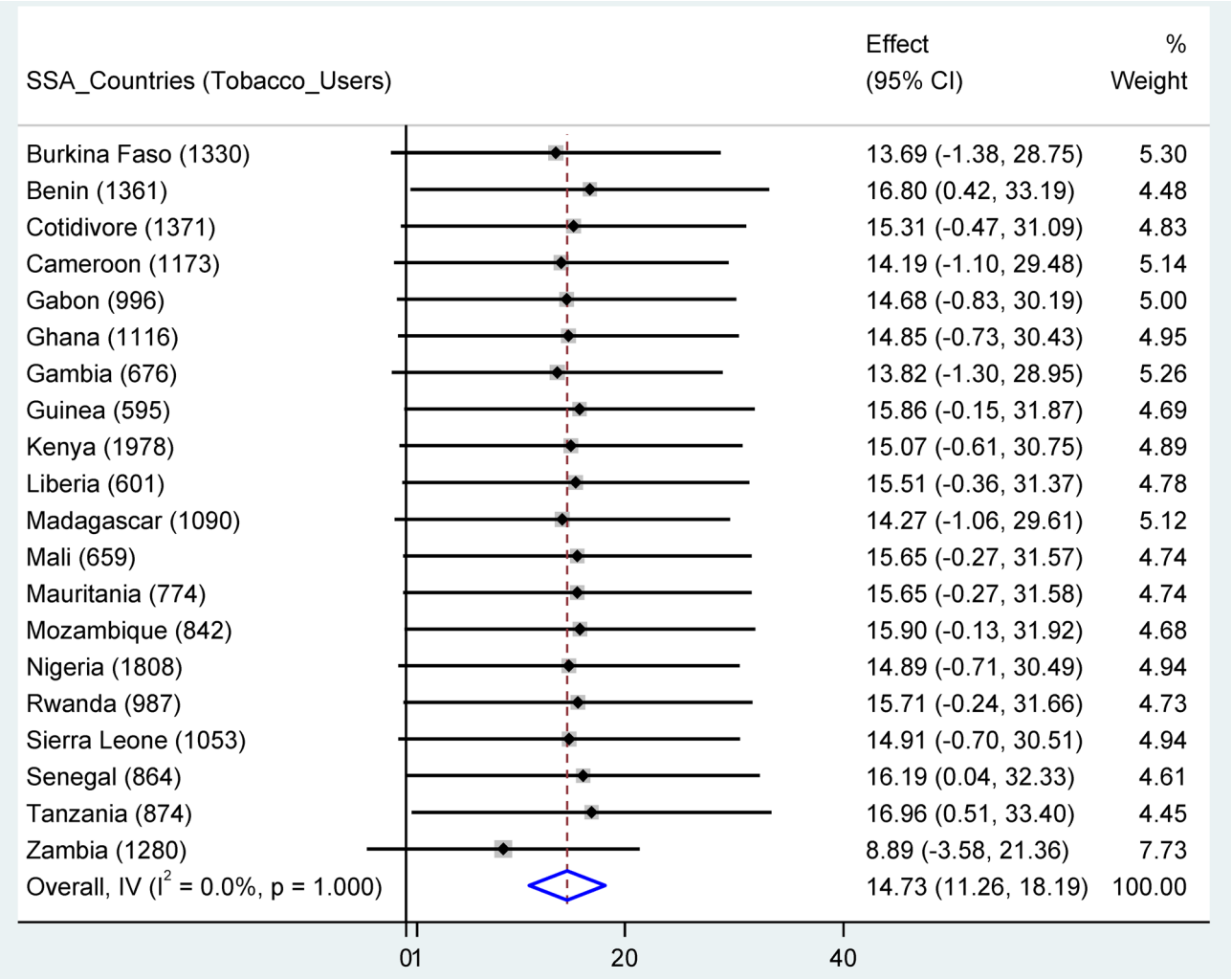

Fig. 4. The pooled prevalence of tobacco use among men in SSA countries.

were extended to match the majority class when the SMOTE oversampling strategy was used, resulting in a more balanced class distribution. To balance the outcome variable's unbalanced distribution, to use the SMOTE oversampling technique produced 101,952 additional synthetic observations from the minority category. To create symmetric distributions for both categories and to enable the development of dependable predictive models, the overall distribution of smoked cigarettes was modified from 22,635 to 124,587 grouped as Yes (for smoked cigarettes) and No (for no smoked cigarettes) to give 124,587 in each class (Fig. 7).

Model performance comparison

After comparing machine learning models, the XGBoost classifier emerged as the best predictive model. The model demonstrated strong performance metrics, achieving an accuracy of 98%, an AUC (Area Under the Curve) of 97, a precision score of 95, a recall of 94, and an F1 score of 90. These results indicate a high level of predictive accuracy and a good balance between precision and recall. To validate the robustness and generalizability of the model, a tenfold cross-validation technique was employed. This method divides the dataset into ten equal parts, using nine parts for training and one for testing in each iteration, thereby offering a more comprehensive and reliable assessment of the model's performance across different data subsets. Hence, the XGBoost classifier is the most accurate prediction model to predict tobacco use among men in Sub Saharan African (Table 4; Fig. 8).

Importance feature selection

Similar to how the majority of conventional statistical methods, such as logistic and linear regression, use t-statistics and p-values to find significant variables, the variable importance feature selection methodology helps in identifying the most significant predictors of an outcome variable. The average measure of a feature's significance about other characteristics that are included in the ensemble model to predict the result is called feature importance. The results of the XGBoost Classifier feature importance analysis revealed that the top ten most important factors influencing the likelihood of smoking cigarettes were: age, age at first sex, education, residence, wealth index, number of sexual partners, religion, internet use, marital status, and occupation (Fig. 9).

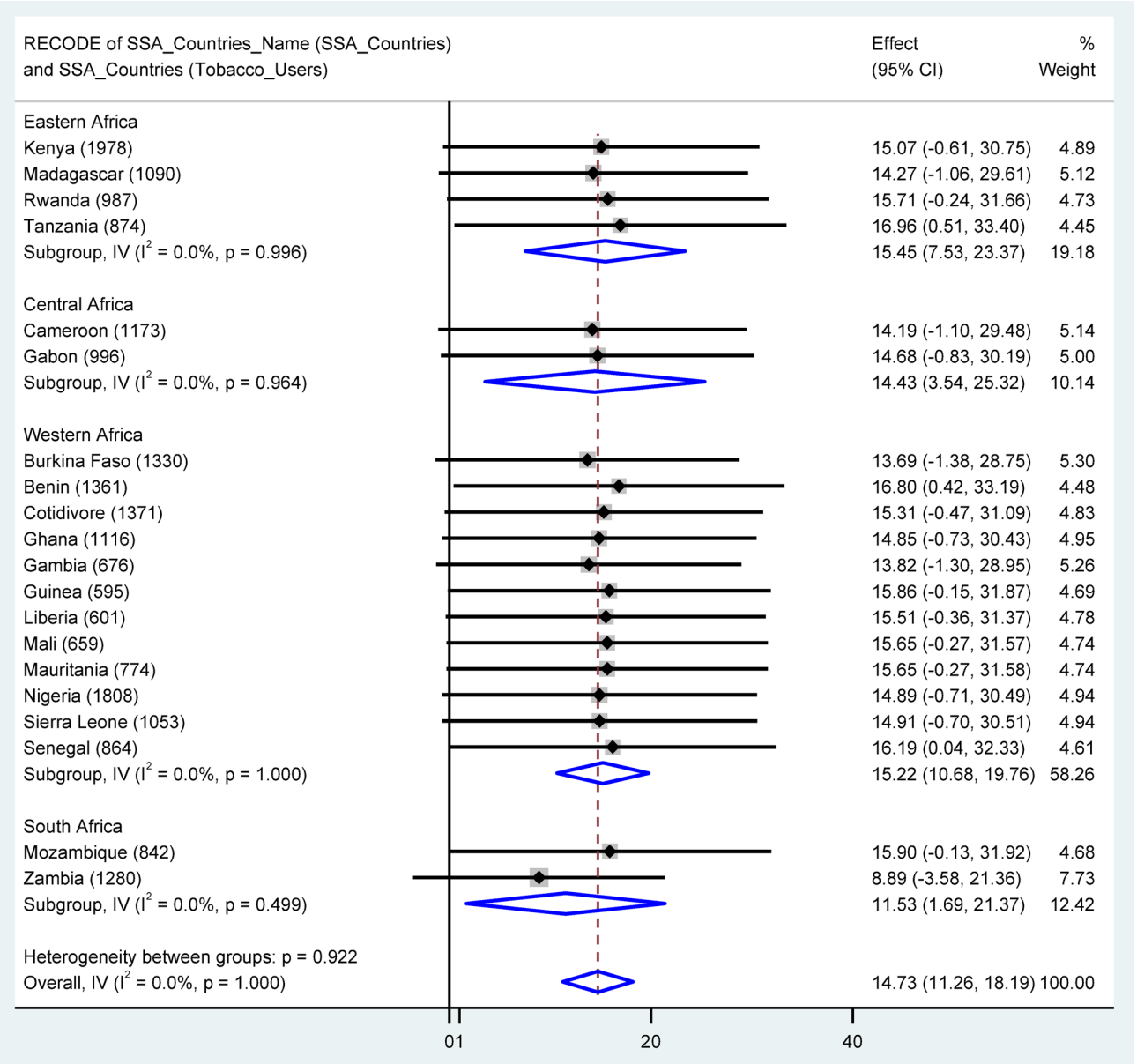

Fig. 5. The sub group analysis of tobacco uses among men by regional location of SSA countries.

Beeswarm plot

The beeswarm plot visualizes feature contributions to the model's predicted probability of tobacco use, where each point represents an observation's feature value and its directional impact on the log-odds output. Features positioned to the right of the vertical baseline (SHAP = 0) increase the predicted probability of the positive class (class 1: "No tobacco use"), with higher feature values (encoded in red, e.g., elevated age, age at first sex, wealth index, or internet use) correlating strongly with reduced tobacco use likelihood. Conversely, lower values of these features (blue) diminish class 1 probability, shifting predictions toward class 0 ("no tobacco use"). Features left of the baseline (SHAP < 0) exhibit an inverse relationship: higher values (red) for attributes such as age, wealth index, internet use, residence amplify the predicted risk of tobacco use (class 0), while lower values (blue) attenuate this effect. The color gradient (red = high/positive, blue = low/negative) clarifies how feature magnitudes modulate directional influence, emphasizing critical interactions for instance, higher education level (red) strongly predicts class 1, whereas its absence (blue) inversely associates with tobacco use (Fig. 10).

Association rule mining

The association rule mining is used for discovering hidden patterns and relationships based on specific confidence intervals and lift, thereby addressing limitations in feature selection. The association rule mining has extracted the following association rules based on the aproiri algorithm.

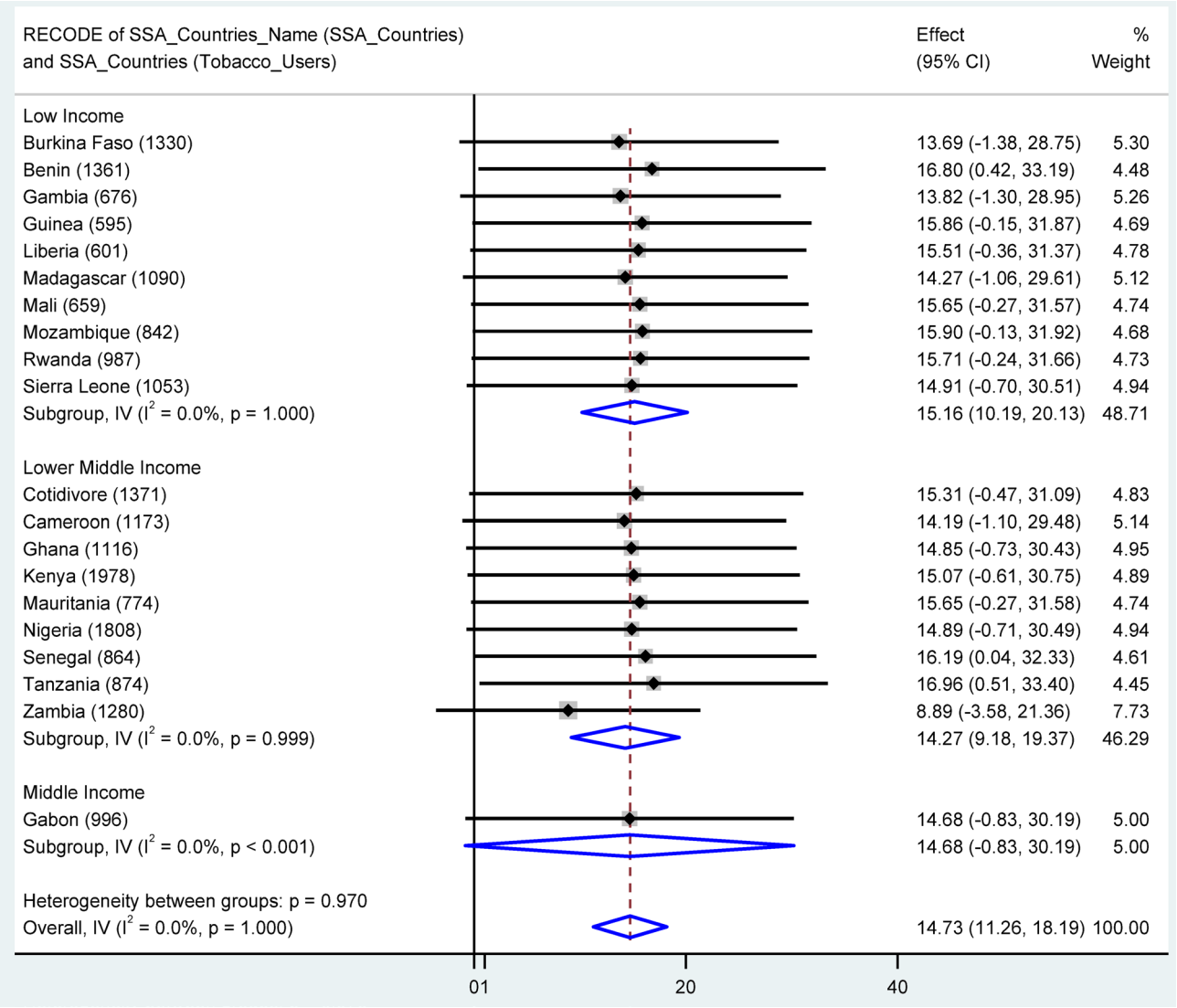

Fig. 6. The sub group analysis of tobacco uses among men by world bank income classification of SSA countries.

Rule 1: This rule states that if a participant is from rural residence, not use internet, and high community poverty the possibility of tobacco use is (24.3% confidence). The lift value of 1.22 indicates that this association is highly stronger than expected by chance.
Rule 2: This rule states that if a participant is from rural residence, and not use internet the possibility of tobacco use is (24.0% confidence). The lift value of 1.21 indicates that this association is highly stronger than expected by chance.
Rule 3: This rule states that if a participant is not use internet, and high community poverty, and high media exposure the possibility of tobacco use is (24.0% confidence). The lift value of 1.21 indicates that this association is highly stronger than expected by chance.
Rule 4: This rule states that if a participant is not use internet, and high community media exposure, the possibility of tobacco use is (24.1% confidence). The lift value of 1.21 indicates that this association is highly stronger than expected by chance.
Rule 5: This rule states that if a participant is not use internet, and high media exposure, the possibility of tobacco use is (23.67% confidence). The lift value of 1.21 indicates that this association is highly stronger than expected by chance.
Rule 6: This rule states that if a participant is not use internet, high community poverty and high media exposure, the possibility of tobacco use is (24.06% confidence). The lift value of 1.21 indicates that this association is highly stronger than expected by chance.
Rule 7: This rule states that if a participant is not use internet and has media exposure, the possibility of tobacco use is (23.67% confidence). The lift value of 1.19 indicates that this association is highly stronger than expected by chance.

Variables	Category	Weighted frequency (%)	Tobacco use	
			Yes	No
Age	15–24	54,428(36.9%)	3656(6.7%)	50,772(93.3%)
	25–34	38,013(25.8%)	6286(16.5%)	31,727(83.5%)
	35–44	30,086(20.4%)	6062(20.1%)	24,024(79.9%)
	45 & above	24,939(16.9%)	5425(21.8%)	19,514(78.2%)
Residence	Urban	62,639(42.5%)	8234(13.1%)	54,405(86.9%)
	Rural	84,827(57.5%)	13,195(15.6%)	71,632(84.4%)
Education	No education	29,916(20.3%)	5476(18.3%)	24,441(81.7%)
	Primary	46,295(31.4%)	8033(17.4%)	38,262(82.6%)
	Secondary	56,960(38.6%)	6922(12.2%)	50,038(87.8%)
	Higher	14,294(9.7%)	998(7.0%)	13,296(93.0%)
Religion	Catholic	52,054(35.3%)	7685(148%)	44,369(85.2%)
	Islam	3616(2.5%)	538(14.9%)	3078(85.1%)
	Protestants	4856(3.3%)	673(13.9%)	4184(86.1%)
	Other Christians	59,734(40.5%)	8376(14.0%)	51,358(86.0%)
	Universal	11,613(7.9%)	1219(10.5%)	10,394(89.5%)
	Other/no religion	15,592(10.6%)	2938(18.8%)	12,654(81.2%)
marital status	Never in union	62,384(42.3%)	5195(8.3%)	57,189(91.7%)
	Married	66,331(45.0%)	11,784(17.8%)	54,547(82.2%)
	Living with partner	13,069(8.9%)	2502(19.1%)	10,567(80.9%)
	Single	5682(3.9%)	1948(34.3%)	3734(65.7%)
Age at first sex	15–19	105,781(71.7%)	15,458(14.6%)	90,323(85.4%)
	20–24	29,442(20.0%)	4350(14.8%)	25,092(85.2%)
	25 & above	12,242(8.3%)	1621(13.2%)	10,622(86.8%)
Number of sexual partners	0	107,103(72.6%)	13,907(13.0%)	93,196(87.0%)
	1	30,070(20.4%)	5169(17.2%)	24,900(82.8%)
	2–4	10,293(7.0%)	2353(22.9%)	7940(77.1%)
Internet Use	Yes	55,199(37.4%)	5907(10.7%)	49,292(89.3%)
	No	92,267(62.6%)	15,521(16.8%)	76,745(83.2%)
Media exposure	Yes	118,050(80.1%)	17,161(14.5%)	100,889(85.5%)
	No	29,415(19.9%)	4268(14.5%)	25,148(85.5%)
Literacy	Unable read/write	39,008(26.5%)	7453(19.1%)	31,555(80.9%)
	Able to read/write	107,961(73.2%)	13,893(12.9%)	94,068(87.1%)
	Blind/visually impaired	497(0.3%)	82(16.5%)	415(83.5%)
wealth index	Poor	51,245(34.8%)	9814(19.2%)	41,431(80.8%)
	Middle	29,281(19.9%)	4232(14.5%)	25,049(85.5%)
	Rich	66,939(45.4%)	7382(11.0%)	59,557(89.0%)
Occupation	Not working	22,261(15.1%)	1606(7.2%)	20,655(92.8%)
	Professional	36,076(24.5%)	4342(12.0%)	31,734(88.0%)
	Agriculture	47,407(32.1%)	8648(18.2%)	38,759(81.8%)
	Manual/skilled/unskilled	32,186(21.8%)	5775(17.9%)	26,411(82.1%)
	Others	9535(6.5%)	1058(111%)	8478(88.9%)
Community Poverty	Low poverty	3747(2.5%)	324(8.6%)	3424(91.4%)
	High poverty	143,718(97.5%)	21,105(14.7%)	122,613(85.3%)
Community Media exposure	Low media exposure	69,735(47.3%)	10,398(14.9%)	59,337(85.1%)
	High media exposure	77,730(52.7%)	11,030(14.2%)	66,700(85.8%)
Community Education level	High education	72,763(49.3%)	10,228(14.1%)	62,535(85.9%)
	Low education	74,702(50.7%)	11,200(15.0%)	63,502(85.0%)

Table 3. Socio-demographic characteristics of men in SSA between 2018 and 2023.

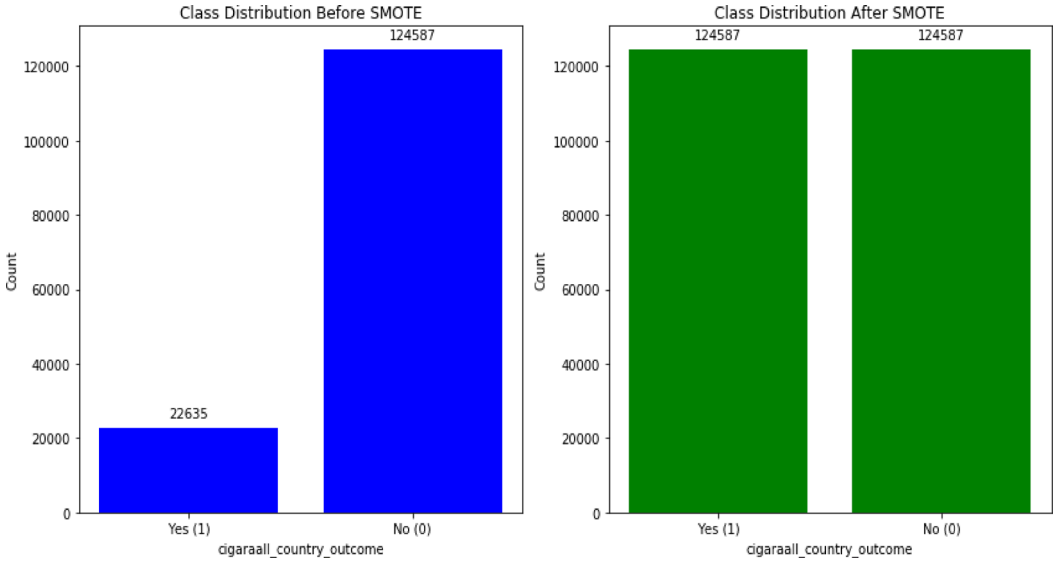

Fig. 7. Class Distribution of tobacco use before and after SMOTE.

Machine learning models	Dataset	Accuracy	AUC	Precision	Recall	F1 score
Logistic regression	Unbalanced (%)	55	52	50	51	56
	Balanced (%)	85	53	52	62	63
K nearest neighbor	Unbalanced (%)	79	50	55	52	54
	Balanced (%)	85	94	87	92	86
Decision Tree Classifier	Unbalanced (%)	85	50	65	68	55
	Balanced (%)	97	92	88	90	94
Random Forest Classifier	Unbalanced (%)	85	45	55	50	54
	Balanced (%)	98	95	94	93	92
XGBoost classifier	Unbalanced (%)	56	58	60	57	59
	Balanced (%)	98	97	95	94	90
AdaBoost Classifier	Unbalanced (%)	84	43	52	65	45
	Balanced (%)	98	94	93	95	96

Table 4. Model performance comparison of machine learning performance metrics to predict tobacco use among men in SSA.

Rule 8: This rule states that if a participant is married and within high community poverty region, the possibility of tobacco use is (23.04% confidence). The lift value of 1.16 indicates that this association is highly stronger than expected by chance.

In general, the association rules consistently point to a higher probability of tobacco use among participants: rural residence, non-use of internet, high community poverty, high community media exposure, and high media exposure (Fig. 11).

Discussion

Although tobacco smoking is the leading cause of preventable diseases and deaths worldwide (5.4 million deaths in 2010), the rate of tobacco consumption has been rising in LMICs over the past few decades[56]. The development of successful tobacco control measures in this area of growing worldwide interest region in tobacco use will be facilitated by an understanding of the prevalence of tobacco use in SSA and their connection to key social health factors. Out of the 47 nations where the most recent DHS data are available, our analysis of the DHS data produced national-level estimates of tobacco use among men in 20 SSA countries. Among men, the pooled prevalence of tobacco use was 14.73(11.26, 18.19) percent. The prevalence ranges from 8.90 in Zambia to 17.00% in Tanzania. There is a wide range of disparities in tobacco use among the countries in the SSA. Among all the SSA countries, the prevalence of tobacco use among men was high in Tanzania, Bennie, and Senegal. Moreover, the spatial analysis found the highest cluster of tobacco use in Mozambique, Zambia, Benin, Mali, Mauritania, Senegal, Guinea, Sierra Leone, and Liberia while the lowest clustering was found in Zambia.

Although countries in SSA are variant and might follow different health system and approach a significant heterogeneity among countries was not detected. Furthermore, subgroup analysis by regional classification and

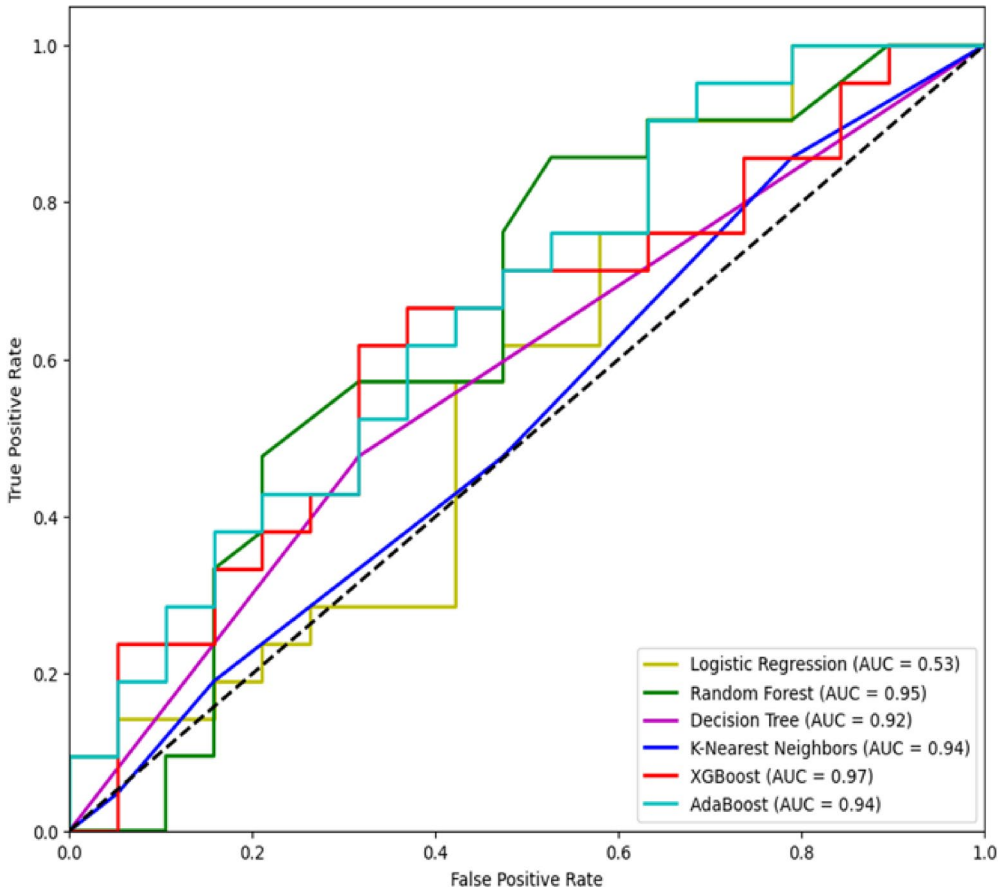

Fig. 8. Receiver operating characteristics and area under the curve for tobacco use among men in Sub Saharan Africa.

world bank income classification was conducted to further investigate if there is any source of heterogeneity yet the finding revealed that significant heterogeneity was not found between regions and among middle and low income countries within the region. Although Sub-Saharan African (SSA) countries differ significantly in terms of geography, economic development, and health system structures, the analysis did not reveal significant heterogeneity in the prevalence of tobacco use among men. This consistent pattern can be explained by several factors. Many SSA countries share similar socio-cultural norms and face common economic and public health challenges that may drive comparable smoking behaviors. The tobacco industry's marketing strategies often target these nations uniformly, exploiting weak regulatory frameworks and limited enforcement capacity. Furthermore, the widespread adoption of the WHO Framework Convention on Tobacco Control (WHO FCTC) has led to the implementation of similar tobacco control policies across countries in region[57,58]. These aligned policy environments may help explain the uniformity in prevalence rates. Subgroup analyses based on regional classifications and World Bank income levels also not found significant differences, suggesting that neither regional proximity nor economic status strongly influences tobacco use patterns. Additionally, the use of standardized data sources, particularly the Demographic and Health Surveys (DHS), which apply consistent sampling techniques and measurement tools across countries, likely contributed to the low methodological variation and the absence of statistical heterogeneity observed in the findings.

The finding of this study is in line with the finding of a study among 10 East African countries, which reported a prevalence of 14.69 among men[59], 30 SSA[11]. The finding of this study is lower than that of a study among four SSA countries, which reported a 34.5 prevalence among men[14,60], local studies conducted in Ethiopia[61], and Democratic Republic Congo[62]. The possible explanation for the difference might be that these studies are local-based and country-specific, which leads to a high prevalence of tobacco use depending on the sociodemographic and norms of each country. In addition, the difference between the four SSA results could be due to the small sample size and the few countries included. The findings of this study is higher than those of local studies conducted in Ethiopia[31], Gabon[63], 21 SSA[64]. These notable variations may stem from differing cultural expectations and societal restrictions surrounding tobacco consumption. This may also be associated with increased production and aggressive marketing in SSA countries[65].

To model tobacco use, six machine learning algorithms including Random Forest, XGBoost, K-Nearest Neighbors (KNN), Decision Tree, Logistic Regression, and AdaBoost were applied to both balanced and imbalanced datasets. Among these, XGBoost demonstrated superior performance, achieving an accuracy of 98%, an AUC of 0.97, a precision of 95%, a recall of 94%, and an F1 score of 90. The evaluation of these models

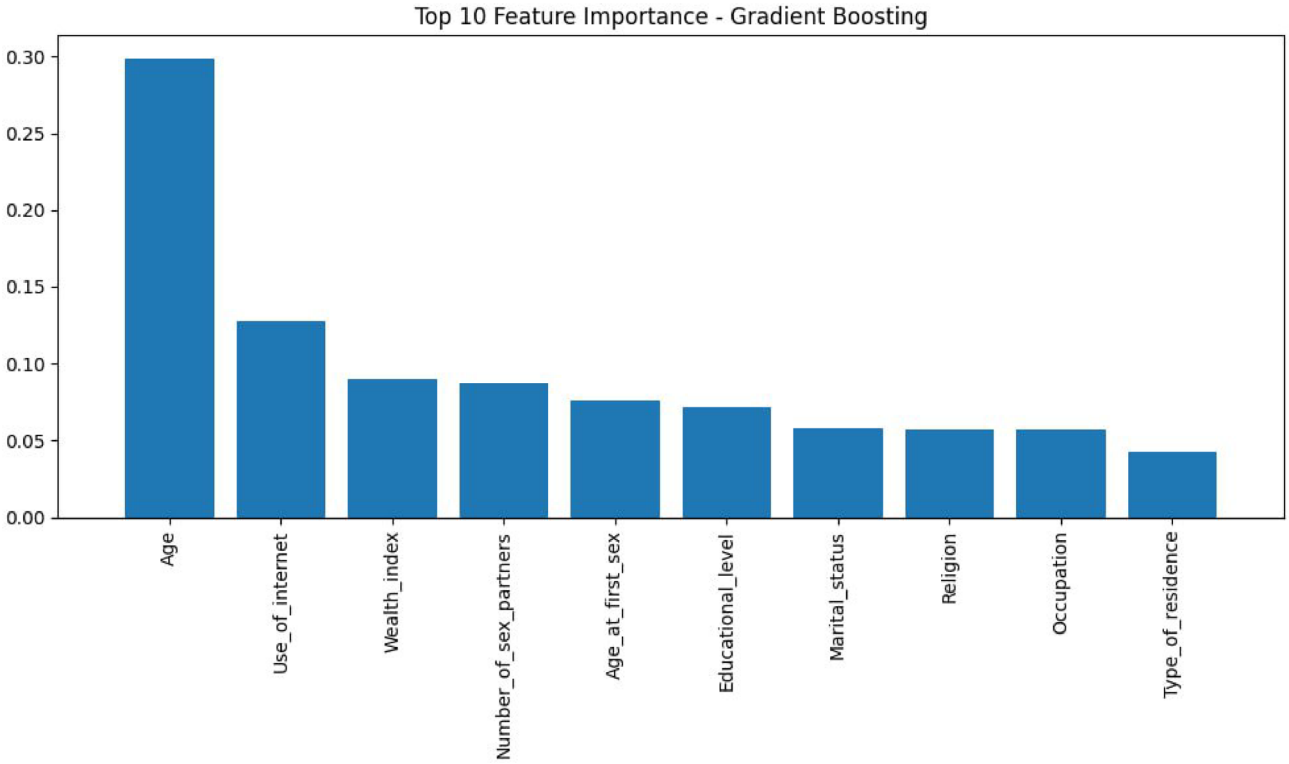

Fig. 9. XGBOOST feature importance selection for tobacco use in Sub Saharan Africa.

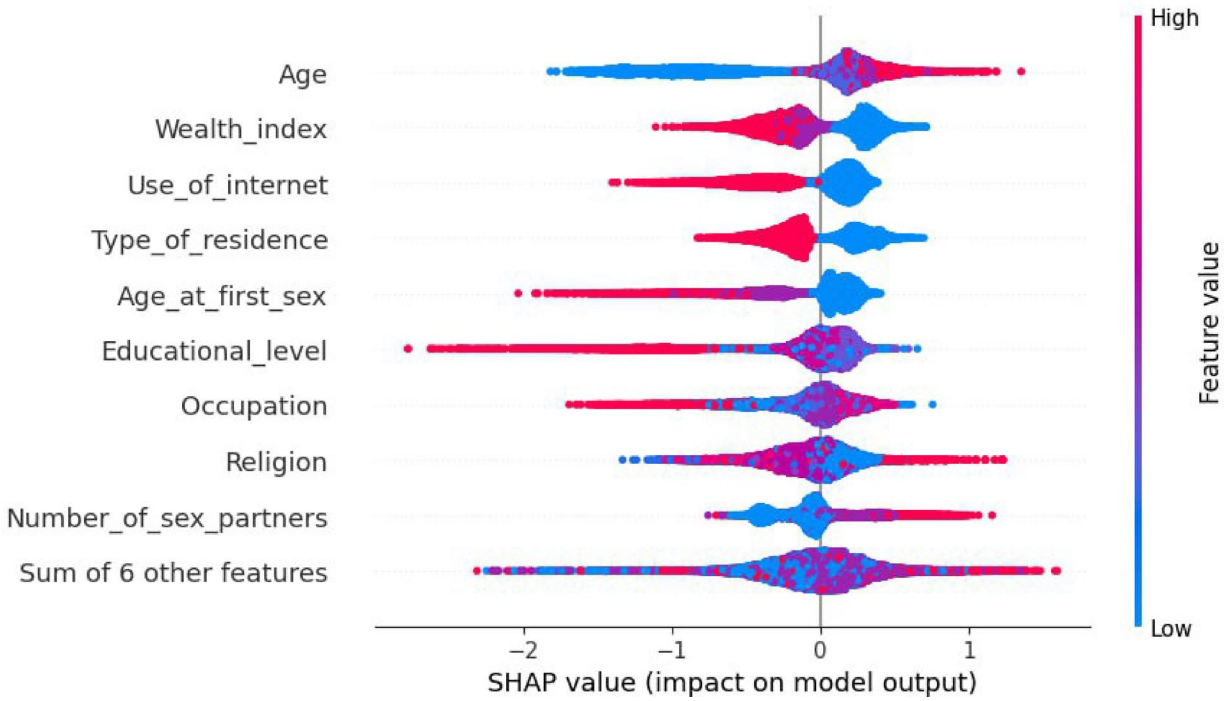

Fig. 10. Important features SHAP value impact on model to predict tobacco use among men across in SSA.

employed tenfold cross-validation to ensure the reliability and robustness of the results by minimizing overfitting through repeated train-test splitting. To address class imbalance, the Synthetic Minority Over-Sampling Technique (SMOTE) was utilized, enabling the models to learn from both majority and minority classes equally and thereby improving prediction fairness and model generalizability. The balanced dataset showed considerable

Smokes cigarette =Yes

```
**Rule** 1: Use of internet_No, community_poverty_high poverty, Type of place of residence_rural -> Smokes_cigarette_yes
Support: 0.1019506248095093, Lift: 1.223606668411035, Confidence: 0.24287529497186422

**Rule** 2: Use of internet_No, Type of place of residence_rural -> community_poverty_high poverty, Smokes_cigarette_yes
Support: 0.1019506248095093, Lift: 1.223606668411035, Confidence: 0.24287529497186422

**Rule** 3: Use of internet_No, Type of place of residence_rural -> Smokes_cigarette_yes
Support: 0.1019506248095093, Lift: 1.223606668411035, Confidence: 0.24287529497186422

**Rule** 4: Use of internet_No, community_media_exposure_high_media_exposure, community_poverty_high poverty -> Smokes_cigarette_yes
Support: 0.11818043279487961, Lift: 1.2122159407283184, Confidence: 0.24061433447098973

**Rule** 5: Use of internet_No, community_media_exposure_high_media_exposure -> Smokes_cigarette_yes
Support: 0.11818043279487961, Lift: 1.2122159407283184, Confidence: 0.24061433447098973

**Rule** 6: Use of internet_No, community_media_exposure_high_media_exposure -> community_poverty_high poverty, Smokes_cigarette_yes
Support: 0.11818043279487961, Lift: 1.2122159407283184, Confidence: 0.24061433447098973

**Rule** 7: Use of internet_No, media_exposure_recoded_yes -> Smokes_cigarette_yes
Support: 0.11764705882352941, Lift: 1.192500065470388, Confidence: 0.23670090449179826

**Rule** 8: Use of internet_No, media_exposure_recoded_yes -> community_poverty_high poverty, Smokes_cigarette_yes
Support: 0.11764705882352941, Lift: 1.192500065470388, Confidence: 0.23670090449179826

**Rule** 9: Use of internet_No, community_poverty_high poverty, media_exposure_recoded_yes -> Smokes_cigarette_yes
Support: 0.11764705882352941, Lift: 1.192500065470388, Confidence: 0.23670090449179826

**Rule** 10: Marital_status_recode_Married -> community_poverty_high poverty, Smokes_cigarette_yes
Support: 0.10553185004571777, Lift: 1.1606179834828105, Confidence: 0.2303725881570193
```

Fig. 11. Association rules for determinants of tobacco use among men in SSA.

improvement in performance metrics, which informed the subsequent feature selection and SHAP value analysis. This approach allowed the discovery of intricate patterns and complex interactions between tobacco use and its determinants relationships that traditional statistical methods might overlook. Unlike conventional models that depend on predetermined assumptions, machine learning algorithms are capable of handling large, high-dimensional datasets and can automatically identify associations across multiple variables. This enhances the depth of analysis, revealing trends, subgroup differences, and dynamic factors influencing tobacco use that linear models may fail to capture.

Moreover, the SHAP analysis improved the interpretability of the model by illustrating how each variable influenced the predictions. The ten most impactful factors associated with tobacco use were found to be age, internet use, household wealth index, education level, occupation, religious affiliation, age at first sexual experience, marital status, type of residence, and the number of sexual partners.

Socio-cultural factors, including religion, have been identified as significant determinants of tobacco use. The role of religion in influencing tobacco consumption can be complex and multifaceted. Religious beliefs often shape social norms, values, and behaviors within a community, and these norms can either promote or discourage certain practices, including the use of tobacco. However, within different religious groups, there may be a variety of confounding factors that can influence tobacco use. For instance, cultural practices, historical traditions, or community attitudes toward tobacco may vary significantly across religious contexts. In some cases, religious communities may have less stringent regulations or social pressures regarding tobacco use, while in others, tobacco use may be socially accepted or even normalized due to cultural traditions. Additionally, the

socio-economic conditions of religious groups could further contribute to tobacco consumption. For example, in communities where poverty, stress, or lack of access to healthcare is prevalent, tobacco use might be seen as a coping mechanism or an outlet for emotional relief. Furthermore, peer influence and social networks within religious groups may also play a significant role, as individuals may be more likely to adopt behaviors, including tobacco use, based on what aligns with the norms and values of their religious or social groups.

Likewise, occupation has also been identified as a critical factor in determining tobacco use among men. This can be explained by the nature of certain occupations that involve physically demanding or energy-exhausting work. Workers in such environments—such as those in manual labor, construction, agriculture, or industrial sectors—may experience significant physical fatigue, stress, and mental strain due to the intensity of their daily tasks. For these individuals, tobacco use may be perceived as a coping mechanism to manage the stress and exhaustion they face. Smoking could offer temporary relief or a sense of escape from the physical and emotional toll of their jobs, providing them with a momentary break from the demands of their work.

Moreover, the feeling of being "sacrificed for survival" could stem from a perceived lack of control over their circumstances, especially if the work is low-paying or comes with few benefits. In such situations, tobacco use may become a way to cope with the harsh realities of their occupation and the pressures of maintaining their livelihood. Social and environmental factors in the workplace such as peer pressure, smoking being normalized among colleagues, or a lack of workplace health initiatives can further contribute to the adoption of tobacco use to managing stress. Over time, these occupational and social factors can create a strong association between work and tobacco use, making it more difficult for individuals to quit. Age and the prevalence of tobacco use in all its forms were positively correlated, indicating that the older the individual, the higher the relative risk of using any tobacco product. In line with other research, the findings showed that men who were 25 years of age or older had a higher risk of smoking[31,60–63,66]. This might be due to the reason older men experience stress, disturbed moods, and mental health problems more than younger men and use smoking to overcome all this problems[67]. In addition, older men are likely to have high exposure to tobacco, which leads to malfunctioning tobacco use.

Likewise, this study revealed that men who resided in poverty households were more likely to use tobacco products than those who resided in rich households, consistent with previously published studies[60,68–70]. This phenomenon is closely linked to individuals from lower socioeconomic backgrounds, who often experience heightened life stresses. Faced with financial instability, housing insecurity, and limited access to education or healthcare, these individuals may turn to tobacco smoking and chewing as a way to cope with the psychological and emotional strain of their circumstances. Research consistently shows that those in lower socioeconomic groups are more vulnerable to such stressors, and this vulnerability often leads to unhealthy coping mechanisms, such as substance use, with tobacco being one of the most prevalent choices. Moreover, individuals in poverty may lack adequate access to health professionals or resources to effectively manage their lifestyles. Financial barriers, insufficient healthcare infrastructure, and societal factors like stigma or cultural norms can discourage them from seeking help or guidance. Consequently, they may not receive the necessary support to address the root causes of their stress, mental health challenges, or unhealthy behaviors, such as tobacco use. This lack of control over their circumstances, combined with limited access to healthcare, perpetuates a cycle of harmful behaviors that can be difficult to break.

In addition, marital status was a significant predictor of tobacco use; men who live together and are single are more likely to use tobacco than married men. The findings of this study are supported by other studies conducted in SSA[11,14]. This might be explained as men who have not been married will encounter different life events and may use tobacco and other substances to cope with the stressful situations. On the other hand, less educated men are more likely to use tobacco products which is consistent with prior studies[59,60,63,65]. This might be due to the fact that education can contribute to greater socioeconomic stability, reducing the chance of using smoking as a coping mechanism in times of stress[71]. In addition, education is the most powerful weapon for combating any form of substance use behavior[72]. Through education, individuals gain the understanding and skills needed to make decisions that support their overall well-being.

Furthermore, men who use internet to were less likely to use tobacco products, consistent with earlier studies[64,73]. The internet plays a crucial role in influencing population behaviors. Attitudes towards tobacco are significantly influenced by internet use, and evidence indicates that exposure to internet messages about tobacco use impacts both tobacco use and prevention[74]. Moreover, this study revealed that age at first sex among men was a significant determinants of tobacco use among men, consistent with previously published studies. Those who have sexual exposure at early age were more likely to use tobacco products among men. This might be interlinked with those who have early sexual exposure might have exposed different stressful situations like sexually transmitted disease which leads to stressful situation and tobacco use. Lastly, men who has more than one sexual partner been more likely to use tobacco products. This is also linked with men with many sexual partners have a fear of sexually transmitted disease and might pressure by their partners to Live together leads to stressful situation and exposed for tobacco use.

This study provides comprehensive insights into the widespread challenge of tobacco use among men across sub-Saharan Africa, revealing substantial disparities in both prevalence and contributing factors. The finding indicates that the pooled prevalence across the region was approximately 14.73 with a confidence interval of 11.26 to 18.19 with no heterogeneity between countries. In addition, the highest cluster of tobacco use was detected in Mozambique, Zambia, Benin, Mali, Mauritania, Senegal, Guinea, Sierra Leone, and Liberia. The study also demonstrates the potential of machine learning approaches in identifying key predictors of tobacco use with high accuracy and reliability. XGBoost showed a strong predictive capability, with an accuracy of 98%, AUC of 0.97, precision of 95, and recall of 94 using the balanced dataset. Factors like age, internet use, wealth index, education level, occupation, religion, age at first sex, marital status, residence, and number of sexual partners emerging as significant predictors of tobacco use. These results emphasize the multifaceted nature of tobacco

use in the region and point to the urgent need for tailored, evidence-based interventions that address both individual-level behaviors and broader systemic barriers. To effectively address the high burden of tobacco use across SSA, a multifaceted and inclusive strategy is needed. Investment in education, empowerment, and access to information through digital platforms, and community health programs is crucial for delivering information about the health impacts of tobacco use. Interventions must also address socio-cultural and economic barriers, such as early sexual experience, and poverty. Furthermore, integrating data-driven approaches, including machine learning and predictive analytics, can support more efficient identification of at-risk populations and better-targeted interventions. A collaborative effort among policymakers, health professionals, researchers, and communities is vital to reduce tobacco use across the region.

Limitations and strengths of the study

This study's primary strength was the use of large sample sizes and nationally representative data. The use of a sophisticated and suitable statistical method (machine learning technique), which revealed previously undiscovered relationships and patterns in the field, was another key point. To determine the relative significance of each predictor and to understand how each component contributed to the model's predictions, the researchers employed a variety of methodologies, including SHAP. This helped them understand how various factors affected the model's predictions.

Despite the wide nature and high sample size of this study, the study has limitations, such as self-reporting being the primary method used to measure tobacco use. As a result, the possibility exists that responses were influenced by social desirability bias, leading to underreporting. In addition, the removal of country-specific local variables may result in the loss of some country-specific attributes that may have contributed to an improved investigation.

Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Received: 7 August 2024; Accepted: 26 June 2025
Published online: 09 July 2025

References

1. Smoking, ITJL, France: IARC. IARC monographs for the evaluation of the Carcinogenic Risk of Chemicals to Humans. 1986:312–4.
2. Cancer IWGotEoCRtHJLIAfRo. Tobacco smoke and involuntary smoking (IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Vol. 83). 2004.
3. Bartal, M. Health effects of tobacco use and exposure. *Monaldi Arch. Chest Dis.* **56**(6), 545–554 (2001).
4. Mitchell, B. E., Sobel, H. L. & Alexander, M. H. The adverse health effects of tobacco and tobacco-related products. *Primary Care: Clin. Office Pract.* **26**(3), 463–498 (1999).
5. Musk, A. W. & De Klerk, N. H. History of tobacco and health. *Respirology* **8**(3), 286–290 (2003).
6. Gilmore, A. B., Fooks, G., Drope, J., Bialous, S. A. & Jackson, R. R. Exposing and addressing tobacco industry conduct in low-income and middle-income countries. *The Lancet* **385**(9972), 1029–1043 (2015).
7. Cairney, P., Studlar, D. & Mamudu, H. *Global tobacco control: Power, policy, governance and transfer* (Springer, 2011).
8. Warren, C. W. *The GTSS atlas* (CDC Foundation, 2009).
9. Organization WH. WHO study group on tobacco product regulation. Report on the scientific basis of tobacco product regulation: ninth report of a WHO study group: World Health Organization; 2023.
10. Organization WH. WHO study group on tobacco product regulation: report on the scientific basis of tobacco product regulation: fifth report of a WHO study group: World Health Organization; 2015.
11. Sreeramareddy, C. T., Pradhan, P. M. & Sin, S. Prevalence, distribution, and social determinants of tobacco use in 30 sub-Saharan African countries. *BMC Med.* **12**, 1–13 (2014).
12. Sreeramareddy, C. T. & Acharya, K. Trends in prevalence of tobacco use by sex and socioeconomic status in 22 sub-Saharan African countries, 2003–2019. *JAMA Netw. Open* **4**(12), e2137820–e2137820 (2021).
13. Brathwaite, R., Addo, J., Smeeth, L. & Lock, K. A systematic review of tobacco smoking prevalence and description of tobacco control strategies in sub-Saharan African countries 2007 to 2014. *PLoS ONE* **10**(7), e0132401 (2015).
14. Boua, P. R. et al. Prevalence and socio-demographic correlates of tobacco and alcohol use in four sub-Saharan African countries: A cross-sectional study of middle-aged adults. *BMC Public Health* **21**(1), 1126 (2021).
15. Reitsma, M. B. et al. Smoking prevalence and attributable disease burden in 195 countries and territories, 1990–2015: A systematic analysis from the Global Burden of Disease Study 2015. *The Lancet* **389**(10082), 1885–1906 (2017).
16. Ou, Z. et al. Global trends in death, years of life lost, and years lived with disability caused by breast cancer attributable to secondhand smoke from 1990 to 2019. *Front. Oncol.* **12**, 853038 (2022).
17. Guliani, H., Gamtessa, S. & Çule, M. J. B. P. H. Factors affecting tobacco smoking in Ethiopia: Evidence from the demographic and health surveys. *BMC Public Health* **19**, 1–17 (2019).
18. Thun, M., Peto, R., Boreham, J. & Lopez, A. D. Stages of the cigarette epidemic on entering its second century. *Tob. Control* **21**(2), 96–101 (2012).
19. Onoh, I. et al. Prevalence, patterns and correlates of smokeless tobacco use in Nigerian adults: An analysis of the Global Adult Tobacco Survey. *PLoS ONE* **16**(1), e0245114 (2021).
20. Othman, M., Farid, N. D. N., Aghamohammadi, N. & Danaee, M. Determinants of smokeless tobacco use and prevalence among Sudanese adolescents. *Arch. Public Health* **79**, 1–10 (2021).
21. Agaku, I. T., Ayo-Yusuf, O. A., Vardavas, C. I. & Connolly, G. Predictors and patterns of cigarette and smokeless tobacco use among adolescents in 32 countries, 2007–2011. *J. Adolesc. Health* **54**(1), 47–53 (2014).
22. Cairney, P. & Mamudu, H. The global tobacco control 'endgame': Change the policy environment to implement the FCTC. *J. Public Health Policy* **35**, 506–517 (2014).
23. Barnett, P., Zhang, W. & Jiang, S. Policy environments for tobacco control. *Smoking Environ. China: Challenges Tobacco Control* **2011**, 211–245 (2021).
24. Vargas, L. S. et al. Determinants of tobacco use by students. *Rev. Saude Publica* **51**, 36 (2017).
25. Ezzati, M. & Lopez, A. D. Estimates of global mortality attributable to smoking in 2000. *The Lancet* **362**(9387), 847–852 (2003).

26. Landau, L. I. Tobacco smoke exposure and tracking of lung function into adult life. *Paediatric Respiratory Rev.* **9**(1), 39–44 (2008).
27. Maritz, G. S. & Mutemwa, M. Tobacco smoking: Patterns, health consequences for adults, and the long-term health of the offspring. *Global J. Health Sci.* **4**(4), 62 (2012).
28. Organization WH. WHO Report on the Global Tobacco Epidemic, 2009: Implementing smoke-free environments: executive summary. (World Health Organization; 2009).
29. The, W. Curbing the epidemic: Governments and the economics of tobacco control. *Tob. Control* **8**(2), 196 (1999).
30. Baleta, A. J. T. L. Africa's struggle to be smoke free. *The Lancet* **375**(9709), 107–108 (2010).
31. Mengesha, S. D. et al. Tobacco use prevalence and its determinate factor in Ethiopia-finding of the 2016 Ethiopian GATS. *BMC Public Health* **22**(1), 555 (2022).
32. Ngaruiya, C. et al. Tobacco use and its determinants in the 2015 Kenya WHO STEPS survey. *BMC Public Health* **18**, 1–13 (2018).
33. Nketiah-Amponsah, E., Afful-Mensah, G. & Ampaw, S. Determinants of cigarette smoking and smoking intensity among adult males in Ghana. *BMC Public Health* **18**, 1–10 (2018).
34. Ogbodo, S. C. & Onyekwum, C. A. Social determinants of health, religiosity, and tobacco use in sub-Saharan Africa: Evidence from the global adult tobacco surveys in seven countries. *J. Public Health* **32**(6), 895–908 (2024).
35. Colwell, B., Mosema, K. B., Bramble, M. S. & Maddock, J. Comparisons of social and demographic determinants of tobacco use in the democratic Republic of the Congo. *Globalization Health* **16**, 1–11 (2020).
36. Sreeramareddy, C. T., Pradhan, P. M. S., Mir, I. A. & Sin, S. Smoking and smokeless tobacco use in nine South and Southeast Asian countries: Prevalence estimates and social determinants from Demographic and Health Surveys. *Popul. Health Metrics* **12**, 1–16 (2014).
37. Jha, P., Ranson, M. K., Nguyen, S. N. & Yach, D. Estimates of global and regional smoking prevalence in 1995, by age and sex. *Am. J. Public Health* **92**(6), 1002–1006 (2002).
38. Waldron, I. Patterns and causes of gender differences in smoking. *Soc. Sci. Med.* **32**(9), 989–1005 (1991).
39. Pampel, F. Tobacco use in sub-Sahara Africa: estimates from the demographic health surveys. *Soc. Sci. Med.* **66**(8), 1772–1783 (2008).
40. Ij, H. Statistics versus machine learning. *Nat. Methods* **15**(4), 233 (2018).
41. Iniesta, R., Stahl, D. & McGuffin, P. Machine learning, statistical learning and the future of biological research in psychiatry. *Psychol. Med.* **46**(12), 2455–2465 (2016).
42. Chen, J. H. & Asch, S. M. Machine learning and prediction in medicine—beyond the peak of inflated expectations. *N. Engl. J. Med.* **376**(26), 2507 (2017).
43. Wyner, A., Olson, M., Bleich, J. & Mease, D. Explaining the success of adaboost and random forests as interpolating classifiers. ArXiv. 2015;abs/1504.07676.
44. Irmalasari, I. & Dwiyanti, L. Algorithm analysis of decision tree, gradient boosting decision tree, and random forest for classification (Case Study: West Java House of Representatives Election 2019). In *2023 international conference on electrical engineering and informatics (ICEEI)*. 1-5 (2023).
45. Dev, V. & Eden, M. Gradient boosted decision trees for lithology classification. In *Computer aided chemical engineering*. (2019).
46. Basha, S., Rajput, D. & Vandhan, V. Impact of gradient ascent and boosting algorithm in classification. *Int. J. Intell. Eng. Syst.* **11**, 41–49 (2018).
47. Azmi, S. S. & Baliga, S. An overview of boosting decision tree algorithms utilizing AdaBoost and XGBoost boosting strategies. (2020).
48. Bifarin, O. O. Interpretable machine learning with tree-based shapley additive explanations: Application to metabolomics datasets for binary classification. *PLoS ONE* **18**(5), e0284315 (2023).
49. Altaf, W., Shahbaz, M. & Guergachi, A. Applications of association rule mining in health informatics: A survey. *Artif. Intell. Rev.* **47**, 313–340 (2017).
50. Hahsler, M., Grün, B. & Hornik, K. arules-A computational environment for mining association rules and frequent item sets. *J. Stat. Softw.* **14**, 1–25 (2005).
51. Savasere, A., Omiecinski, E. & Navathe, S. An Ecient algorithm for mining association rules in large databases. In *Proceedings of the 21st international conference on very large databases (VLDB)*; (1995).
52. Kumar, A. S. & Wahidabanu, R. Data mining association rules for making knowledgeable decisions. In *Data mining applications for empowering knowledge societies*. p. 43–53 (IGI Global; 2009).
53. Han, J., Kamber, M. & Pei, J. *Data mining: Concepts and* (Morgan Kaufmann Publishers, 2012).
54. Tan, P.-N., Steinbach, M. & Kumar, V. Introduction to data mining: Pearson Education India; (2016).
55. Rücker, G., Schwarzer, G., Carpenter, J. R. & Schumacher, M. Undue reliance on I 2 in assessing heterogeneity may mislead. *BMC Med. Res. Methodol.* **8**(1), 79 (2008).
56. Organization WH. WHO report on the global tobacco epidemic 2015: Raising taxes on tobacco: (World Health Organization; 2015).
57. Organization WH. WHO report on the global tobacco epidemic 2021. (2021).
58. Organization WH. WHO report on the global tobacco epidemic, 2021: addressing new and emerging products: executive summary. WHO report on the global tobacco epidemic, 2021: addressing new and emerging products: executive summary2021.
59. Terefe, B., Jembere, M. M., Chekole, B., Assimamaw, N. T. & Gebeyehu, D. A. Frequency of cigarette smoking and its associated factors among men in East Africa: A pooled prevalence analysis of national survey using multinomial regression. *BMC Public Health* **24**(1), 668 (2024).
60. Mamudu, H. M., John, R. M., Veeranki, S. P. & Ouma, A. E. O. The odd man out in Sub-Saharan Africa: Understanding the tobacco use prevalence in Madagascar. *BMC Public Health* **13**, 1–11 (2013).
61. Gutema, B. T. et al. Tobacco use and associated factors among adults reside in Arba Minch health and demographic surveillance site, southern Ethiopia: A cross-sectional study. *BMC Public Health* **21**, 1–10 (2021).
62. Colwell, B., Mosema, K. B., Bramble, M. S. & Maddock, J. Comparisons of social and demographic determinants of tobacco use in the Democratic Republic of the Congo. *Glob. Health* **16**, 1–11 (2020).
63. Islam, M. S. et al. Prevalence of and factors associated with tobacco smoking in the Gambia: A national cross-sectional study. *BMJ Open* **12**(6), e057607 (2022).
64. Darteh, E. K. M., Yaya, S., Dickson, K. S. & Seidu, A.-A. Prevalence and drivers of tobacco use among young men in Sub-Saharan Africa: Evidence from 21 nationally representative surveys. (2020).
65. Sreeramareddy, C. T. & Pradhan, P. M. S. Prevalence and social determinants of smoking in 15 countries from North Africa, Central and Western Asia, Latin America and Caribbean: Secondary data analyses of demographic and health surveys. *PLoS ONE* **10**(7), e0130104 (2015).
66. Sreeramareddy, C. T. & Acharya, K. Trends in prevalence of tobacco use by sex and socioeconomic status in 22 sub-Saharan African countries, 2003–2019. *JAMA Netw. Open* **4**(12), e2137820 (2021).
67. Hall, S. M. et al. Older versus younger treatment-seeking smokers: Differences in smoking behavior, drug and alcohol use, and psychosocial and physical functioning. *Nicotine Tob. Res.* **10**(3), 463–470 (2008).
68. Yaya, S., Bishwajit, G., Shah, V. & Ekholuenetale, M. Socioeconomic disparities in smoking behavior and early smoking initiation among men in Malawi. Tobacco use insights. 2017;10:1179173X17726297.

69. Palipudi, K. et al. Prevalence and sociodemographic determinants of tobacco use in four countries of the World Health Organization: South-East Asia region: findings from the Global Adult Tobacco Survey. *Indian J. Cancer* **51**(Suppl 1), S24–S32 (2014).
70. Ntoimo, L. F. C., Odimegwu, C. O. & Alex-Ojei, C. A. Tobacco use among men in sub-Saharan Africa: Does family structure matter?. *Stud. Soc. Populat. Int. Perspect*. 343–61 (2019).
71. Businelle, M. S. et al. Mechanisms linking socioeconomic status to smoking cessation: A structural equation modeling approach. *Health Psychol.* **29**(3), 262 (2010).
72. Bugbee, B. A., Beck, K. H., Fryer, C. S. & Arria, A. M. Substance use, academic performance, and academic engagement among high school seniors. *J. Sch. Health* **89**(2), 145–156 (2019).
73. Achia, T. N. Tobacco use and mass media utilization in sub-Saharan Africa. *PLoS ONE* **10**(2), e0117219 (2015).
74. Davis, R. M. *The role of the media in promoting and reducing tobacco use: US Department of Health and Human Services*, (National Institutes of Health; 2008).

Acknowledgements
We would like to thank the DHS program for providing the data set

Author contributions
All authors agreed to be responsible for all elements of the work and, including Melaku, MS: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, writing original draft, Writing review & editing. Baykemagn, ND: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, writing original draft, Writing review & editing. AT: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, writing original draft, Writing review & editing. LY: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, writing original draft, Writing review & editing. All authors made a significant contribution to the work reported.

Declarations

Competing interest
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to M.S.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.