Use of Machine Learning Techniques to Identify HIV Predictors for Screening

Patrick E McSharry
Oxford Man Institute of Quantitative Finance, Oxford University, Oxford OX2 6ED

Charles Mutai (charlimtai@gmail.com)
African Center of Excellence in Data Science, University of Rwanda, Kigali BP 4285

Innocent Ngaruye
College of Science and Technology, University of Rwanda, Kigali

Edouard Musabanganji
College of Business and Economics, University of Rwanda, Kigali

Research Article

Keywords: Socio-behavioral, Screening, High-risk, predictors, XGBoost

Posted Date: December 2nd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-118786/v1

Version of Record: A version of this preprint was published at BMC Medical Research Methodology on July 31st, 2021. See the published version at https://doi.org/10.1186/s12874-021-01346-2.

Abstract

Aim: HIV prevention measures at sub-Saharan Africa are still short of attaining the UNAIDS 90-90-90 fast track targets set in 2014. Identifying predictors for HIV status may facilitate targeted screening interventions that improve health care. We aimed at identifying HIV predictors as well as predicting persons at high risk of the infection.

Method: We applied six machine learning approaches for building models using population-based HIV Impact Assessment (PHIA) data for 41,939 male and 45,105 female respondents with 24 and 29 variables respectively from four countries in sub-Saharan countries. We trained and validated the six algorithms on 80% of data and tested on the remaining 20% where we rotated around the left-out country. An algorithm with the best mean f1 score was retained and trained on the most predictive variables. We used the model to identify people living with HIV and individuals with a higher likelihood of contracting the disease.

Results: Application of XGBoost algorithm appeared to significantly improve identification of HIV positivity over the other six algorithms by f1 scoring mean of 78.9% and 92.8% for males and females respectively. Amongst the eight most predictor features in both sexes were: age, relationship with family head, the highest level of education, highest grade at that school level, work for payment, avoiding pregnancy, age at the first experience of sex, and wealth quintile. Model performance using these variables increased significantly compared to having all the variables included. We identified five males and seven females individuals that would require testing to find one HIV positive individual. We also predicted that 4·14% of males and 10.81% of females are at high risk of the infection.

Conclusion: Our findings provide a potential use of XGBoost algorithm with socio-behavioural-driven data at substantially identifying HIV predictors and predicting individuals at high risk of infection for targeted screening.

Background Literature

HIV continues to be significantly the most threatening infectious disease and a burden to public health globally. In the year 2019, global estimates show that 38 million people are living with HIV while 1.7 million and 0.69 million are reported new cases and deaths respectively, despite the remarkable progress in diagnosis and access to antiretroviral therapy (ART) [1]. HIV is disproportionately burdening people living in East and Southern Africa, more than half (54.5%) of people are living with HIV, 42.9% are newly infected with HIV and 43.5% of deaths are due to AIDS respectively [1]. In 2018, 1.6 million, 1 million, 0.21 million and 1.2 million people were living with HIV, 72,000, 38,000, 7,800 and 48,000 were newly infected people and 24,000, 13,000, 2,400 and 17,000 deaths were from AIDS-related illness in Tanzania, Malawi, Eswatini and Zambia respectively [2]. The Joint United Nations Programme (UNAIDS) had set goals towards stopping AIDS as a public health threat by 2030 [3],[4]. However, COVID-19 pandemic is already thwarting the progress made, and it can adversely lead to additional AIDS-related deaths in sub-Saharan Africa by the end of 2021 [5].

Despite universal HIV intervention efforts in East and Southern Africa, the geographical distribution of the HIV epidemic is still widely varied [6],[7]. The region being a resource-constraint can not have every intervention to everyone and everywhere. Granular information concerning the HIV epidemic needs tailor-made solutions to address and help protect specific individuals [8]. To identify the most vulnerable individuals for the infection globally, strategies are geared towards optimal allocation of resources and thus higher impact and efficiency contrary to a homogeneous distribution of resources [9],[10]. Behavioural and social-demographic factors are among significant contributions of HIV transmission and require investigation on the nature of the impact on the HIV epidemic in a particular population [11]. Despite HIV screening being an effective method of identifying individual status, it has challenges and constraints [12]. Community-based HIV screening has successfully improved the identification of people living with HIV [13]. However, gaps persist that require innovative screening strategies to address them [14],[15],[16].

Machine learning entails the utilisation of computational and statistical algorithms to determine hidden associations of data that might increase predictions through relaxation of the modelling postulates advanced by standard approaches [17]. Among the recent advances in prediction tools and identification techniques in HIV statistical data, machine learning offers greater capability in processing huge amounts of data. Its recent application in the identification of potential candidates for preexposure prophylaxis (PrEP) in the USA and Denmark and a population-based research setting in Eastern Africa highlights some of its capabilities [18]. Classification as a non-parametric machine learning method has also been predominantly utilised in health-related outcomes task showcasing high-performance [17]. [19], used Laplacian-modified naïve Bayesian to identify active inhibitor compounds from a target database. Another example is the use of electronic health record data in developing HIV prediction models for identifying PrEP candidates in an extensive health-care system [20]. Various significant data analytical avenues existing in the healthcare system for patients from the perspective of multiple stakeholders have been reviewed [21]. A machine learning algorithm has been developed that can automatically select important HIV risk-related variables using patients' demographic and clinical data [22].

A review of the use of machine learning approaches in studying HIV/AIDS infection was previously published [23]. The paper [24], used machine learning approaches in classifying patients with and without the toxicity of biomarkers of mitochondrial in HIV. Recently, [25] used machine learning techniques on the Demographic Health Survey of 10 countries to identify HIV Positive individuals. [26], applied SuperLearner in the classification of HIV virology failure in selection strategy using adherence data. SuperLearner improved the accuracy of classifying accelerometry data.

This paper aims at using the machine learning XGBoost algorithm to identify the HIV predictors of persons using socio-behavioural features and predict those at increased risk of infection in the East and Southern African countries.

One of the ways of diagnosing for people living with HIV is through universal screening of individuals attending health care facilities, but this can be costly for the low-risk population compared to selective testing of those at high risk [27]. Including social-demographic factors in the analysis may extensively improve the potential of predicting those at higher risk of the infection, enhancing optimal choices in the screening process, and helping to facilitate testing and counselling for HIV [28]. This may also disclose individuals who may need PrEP, among other necessary early interventions[29],[30].

Methods

Data

We used data from the Population-based HIV Impact Assessment (PHIA) project whose goal is to evaluate the HIV disease trial at different levels of the population [31].

ICAP, based in Columbia University in collaboration with the US Centers for Disease Control and Prevention (CDC) and the ministries of Health, manages and implements the PHIA project. The PHIA project is assessing programs of HIV in countries supported by the President's Emergency Plan for AIDS Relief (PEPFAR) by national household surveys.

It was established in 2015 and geared towards documenting the achievements of HIV programs in participating countries as well as ensuring a better understanding of the regional burden trends of the disease in developing countries. PHIA conducts surveys in 14 countries: Côte d'Ivoire, Cameroon, Ethiopia, Eswatini, Haiti, Kenya, Lesotho, Zimbabwe, Malawi, Namibia, Rwanda, Tanzania, Uganda and Zambia. More details on the PHIA survey have been reported elsewhere [32].

We only included individuals tested for HIV in our analysis from the recently released PHIA survey data for Tanzania (2016-2017), Zambia (2016), Malawi (2015-2016) and Eswatini (2016-2017). We merged adult datasets and HIV test results from the four countries to obtain two sets of data, comprising 41,939 male and 45,105 female respondents with 238 variables each, Table 1. We considered two HIV test outcomes for respondents, positive and negative, thereby requiring the construction of a binary classifier using machine learning.

Data pre-processing

We pooled datasets from the four countries and merged HIV test results with adult interview datasets and resampled utilising sample weights of HIV test outcomes per country. We removed variables with more than 30% (137 females, 136 males) missing values and the ones with no variance. We also encoded both the nominal and ordinal variables using the label-code and one-hot encode methods appropriately based on the information from the survey [33]. Also, 38 and 42 variables for males and females respectively were not included in the final dataset because they were non-informative. This resulted in 41,939 males and 45,105 females in the final dataset corresponding to 24 and 29 variables respectively as shown in, Table 2. From this final dataset, 23 variables of the total variables were similar for both sexes.

Table 1. Background characteristics of PHIA dataset

Characteristics	Levels	Male	Female
Total number of individual		41,939	45,105
Country (n, percentage of total)	Malawi	9,587 (22.9%)	10,242 (22.7%)
	Eswatini	5,482 (13.1%)	6,393 (14.2%)
	Tanzania	16,584 (39.5%)	17,476 (38.7%)
	Zambia	10,286 (24.5%)	10,994 (24.4%)
Current age(n, percentage of total)	15-19	8,670 (20.7%)	8,996 (21.5%)
	20-24	7,262 (17.3%)	7,620 (18.2%)
	25-29	5,920 (14.1%)	6,426 (15.3%)
	30-34	4,917 (9.6%)	5,598 (13.3%)
	35-39	4,040 (9.6%)	4,403 (10.5%)
	40-44	3,315 (7.9%)	3,486 (8.3%)
	45-49	2,559 (6.1%)	2,615 (6.2%)
	50-54	1,802 (4.3%)	2,043 (4.9%)
	55-59	1,476 (3.5%)	1,598 (3.8%)
	60-64	834 (2.0%)	959 (2.3%)
	65-69	402 (1.0%)	464 (1.1%)
	70-74	270 (0.6%)	375 (0.9%)
	75-79	220 (0.5%)	255 (0.6%)
	80+ years	252 (0.6%)	267 (0.6%)
Type of residence,n (% of total)	Urban	13,855 (33.0%)	15,388 (36.7%)
	Rural	28,084 (67.0%)	29,717 (70.9%)

Model validation

We left out one country for later testing and this was rotated around for testing the generalisation of the models separately for males and females. 80% of the datasets were selected for training while 20% were used as test samples, and multiple imputations with chained equations (MICE) [34] was utilized in imputing the missing values in each of these categories. A similar imputation approach was applied to the samples that were left out. We imputed on the samples fives times and averaged the results. Finally, we were further harmonised and scaled the data to ensure a fair penalisation of the scheme used for all the regressors.

We randomly picked from a grid 50 sets of hyperparameters, and these were in training and validation of data using each of Elastic Net (EN) [35], k-Nearest Neighbors (kNN) [36], RandomForest (RF) [37], Support Vector Machine (SVM)[38], XGBoost [39] and Light Gradient Boosting (LGBT) [40] algorithms.

In this study, our machine learning task was structured to solve a binary classification problem. Our dataset comprises healthy individuals labelled negative in one class while the infected individuals are labelled positive in the other class. We determined the average scores of f1 for each of these 50 sets over the validated samples and the most powerful set of hyperparameters were picked. f1 score is a metric that is the most-used member of the parametric family of the f-measures, named after the parameter value $\beta = 1$ [41]. It is defined as the harmonic mean of precision and recall.

Table 2. PHIA summary of the features used

Features	Female	Male	Category
Age	√	√	15 to 80 years
Relationship with the family head	√	√	Head, Wife/Husband/Partner, Son or Daughter, Son-in-law/Daughter-in-law, Grandchild, Parent, Parent-in-law, Brother/Sister, Co-wife, Other Relative, Adopted/Foster/Stepchild, Not related
Live here	√	√	Yes, No
Sick to work in the last three months	√	√	Yes, No
Ever attended school	√	√	Yes, No
Ever enrolled in school	√	√	Yes, No
Highest level of education	√	√	Pre-primary, Primary, Post-primary training, Secondary (O-Level), Post-secondary (O-Level) Training, Secondary(A-Level), Post-secondary-Level) Training, University
Highest grade at that school level	√	√	
Work for payment in the last 12 months	√	√	Yes, No
Ever married/lived together	√	√	Yes, No
Avoiding pregnancy	√	√	Yes, No
Age at first sex	√	√	8 to 56 years
Ever tested HIV	√	√	Ever tested, Never tested
Ever sought TB treatment	√	√	Yes, No
Alcohol drink frequency	√	√	Never, once a month or less, 2 to 4 times a month, 2 to 3 times a week, 4 or more times a week
Urban area indicator	√	√	Urban, Rural
Known HIV status	√	√	Stated Positive, Stated Negative, Never Tested, Don't
Wealth quintile	√	√	Lowest, Second, Middle, Fourth, Highest
HIV status	√	√	Positive, Negative
Had sexual intercourse past 12 months	√	√	Yes, No
Ever had sexual intercourse	√	√	Yes, No
Country	√	√	Tanzania, Zambia, Malawi, Eswatini
Marital status	√	√	Married, Living together, Widowed, Divorced, Separated
No. of times pregnant	√		0 to 23 times
No. of births since 2012	√		0 to 10 children
Ever had a successful birth	√		Yes, No
Whether pregnant now	√		Yes, No
Circumcision status		√	Yes, No

f1 score = 2*(Recall * Precision) / (Recall + Precision).

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations and a recall (sensitivity) is the ratio of correctly predicted positive observations to all observations in the actual class. It accounts for both false positives and false negatives and it can not be influenced by the uneven class distribution and therefore preferred over the accuracy metric [42]. Importantly, the metric used was very sensitive and high yielding in predicting the number of HIV positive persons. For one to be classified as HIV positive, the probability threshold set was 50%.

Model and variable selection with the direction of the association

The algorithm with the best f1 score was used in the subsequent analysis where all countries were included. We used the algorithm along with 250 sets of parameters random search in training and validation with a five-fold cross-validation plan. First, we estimated and compared sensitivity, f1 score and positive predictive value results on all the variables by use of MICE.

Then, a sequential forward floating selection (SFFS) procedure [43] with their f1 scores were conducted on the 80% training samples. We evaluated the association of the features with the probability of being HIV positive and picked the portion of features where f1 scores stabilized using SHapley Additive exPlanations (SHAP) [44]. SHAP is an important technique which is used in explaining the contribution of each feature in prediction.

Results

Algorithms results in test samples

Figure 1, shows that the XGBoost algorithm achieved the highest f1 score of 78% and 92% for males and females respectively, among the five algorithms that were used on all 8 (4 per sex) test samples. This was followed closely by the RF algorithm with a score of 76% for males and 93% for females. EN algorithm performance was 72% and 90% for males and females respectively. SVM performance was 71% for males and 89% for females. KNN f1 score was 54% for males and 88% for females while LGBM performed dismally with an f1 score of 48% for males and 88% for females, Table 3.

Algorithms results in left-out samples

Similarly, the six algorithms were trained on all the four left-out samples. The f1 scores between males and females substantially varied all the algorithms, Figure 1. However, XGBoost algorithm got the highest f1 score of 46% and 85% for males and females respectively, among the six algorithms. This was followed closely by the KNN algorithm with a score of 45% for males and 85% for females. SVM algorithm performance was 86% and 35% for males and females respectively. LGBM secured the highest score of 87% for males and a low score of 33% for females. RF scored 86% for males and the lowest 30% for females while EN was the worst-performing algorithm with an f1 score of 76% for males and 33% for females, Table 3. These out-of-sample results are generally worse than the in-sample results previously presented due to over-fitting the data.

Table 3. F1 score for Algorithms on the test, left-out and train samples

samples	XGBoost	KNN	SVM	RF	EN	LGBM
males test	0.78	0.54	0.71	0.76	0.72	0.48
female test	0.92	0.88	0.89	0.93	0.90	0.88
males left-out	0.46	0.45	0.86	0.86	0.76	0.87
females left-out	0.85	0.85	0.35	0.30	0.33	0.33
males train	0.74	0.52	0.67	0.73	0.68	0.45
females train	0.91	0.87	0.89	0.92	0.88	0.88

Imputation results for the entire datasets

We used the XGBoost algorithm in the subsequent analysis where we included all countries in the data. Table 4 shows the results of the two different imputation methods on all variables. XGBoost imputation methods performed slightly higher than the MICE imputation technique on

all features. The f1 scores on the validation set are 79.3% versus 73.3% for males and 93.1% versus 91.2% for females. As a result of the given performance, we used the XGBoost imputation method in further analysis.

Table 4. Performance of imputation methods on variables with test samples

TN: True Negative, TP: True Positive, FN: False Negative, FP: False Positive, PPV: Positive Predictive Value, S: Sensitivity,

	TP	FP	FN	TN	F1 (%)	S (%)	PPV (%)
Complete with MICE imputation (males)	30,176	636	788	1,952	73.3	71.2	75.4
Complete with MICE imputation (females)	31,172	220	576	4,116	91.2	87.7	94.9
Complete with XGBoost imputation (males)	7,593	110	163	522	79.3	76.2	82.6
Complete with XGBoost imputation females	7,815	33	122	1,051	93.1	89.6	97.0
Complete with XGBoost imputation males (15 variables)	7,586	117	163	522	78.9	76.2	81.7
Complete with XGBoost imputation females (12 variables)	7,518	33	128	1,045	92.8	89.1	96.9

Variable selection and direction of associations

We conducted an SFFS procedure in determining the saturation limit selecting variables based on f1 scoring. As a result, 15 and 12 most influential features of males and females were selected respectively, Figure 2. To understand how a feature contributes to the output of the model, we plot SHAP values, Figure 3 and 4 for males and females respectively. These variables are displayed after ranking in descending order, bearing the highest values of Shapley at the bottom. Here, all the values on the left represent the observations that shift the predicted value in the negative direction while the points on the right contribute to shifting the prediction in a positive direction. The graph summarises the impact of explanatory features on the model output. Features that increase or decrease the risk of HIV infection are coded in red and blue respectively. Being older, never attending school, at the highest level of education, at highest grade a school level, in avoidance of pregnancy, in TB treatment, in use alcohol drink, an urban dweller, aware of HIV status, wealthy, nonmarital and circumcised is predictive of HIV positivity.

Situations

A. 95% of individuals living with HIV know their state

Table 5, shows confusion matrices on test samples. A sensitivity of 95% for males would need 4,154 individuals out of 8,388 (49.52%) tested to identify 651 HIV positives from 685 persons living with HIV. Correspondingly, therefore, five individuals would need testing to find one person who is HIV positive with a PPV of 15.67%. 3,301 individuals out of 9,021 (36.59%) of females would require testing to detect 1,115 HIV positives out of the 1,173 persons living with HIV. Similarly, the PPV is 33.77% and needs 19 individuals tested.

B. 95% or higher probability of being HIV positive

We identified 348 (4·14%) males out of the 8,388 and 975 (10.81%) females out of 9,021 as a high-risk population. We find that 350 males would have been identified HIV positives out of 685 people living with HIV while 969 females would have been correctly identified HIV positives from 1,173 individuals, Table 5.

Table 5. PLHIV know their status and 95% or more probability of being HIV positive

95% of those with HIV	TP	FP	FN	TN	PPV %
Know their status (males)	4200	3503	34	651	15.67
Know their status (females)	5662	2186	58	1115	33.77
95% or > probability of being HIV positive (males)	7690	13	350	335	99.26
95% or more probability of being HIV positive (females)	7842	6	204	969	96.26

Discussion

We used a large dataset of over 80,000 respondents in four countries from the East and Southern Africa region to predict the HIV status of persons by use of socio-demographic factors. We used the XGBoost method in identification of the most predictive factors of HIV positivity, which delivered better results than the other six algorithms with f1 score on the sample test of 79.3 and 93.1% when all variables are included in males and females respectively.

The method enabled us to establish the most predictive features for HIV status in both sexes: age, relationship with the head of the family, ever enrolled in school, the highest level of education, highest grade at that school level, work for payment in the last 12 months, wether avoiding pregnancy, age at first sex experienced, ever sought TB treatment, frequency of alcohol drinking, urban area indicator, known HIV status, wealth quintile, number of pregnancies, number of births since 2012, marital status and circumcision status. Although prior investigations have addressed risk factors of HIV through logistic regression models, such approaches, however, do not consider the complex nature of various predictors of HIV [45],[46]. The XGBoost algorithm can discover the drivers of the epidemic for planning effective interventions and eventually attain the treatment targets set globally[47].

The course of the association between predictor features and HIV status of individuals was determined through the use of XGBoost along with SHAP plots and illustrated specific feature importance to give an intuitive understanding of the key features. The age of an individual has the highest overall impact on HIV status than other features, and any change of age can have a more remarkable influence than others. More aged individuals have a higher probability of the infection in both sexes. Several of those avoiding pregnancy by various methods stand a higher chance of contracting the disease in both sexes. Potential reasons for this may be an increased exposure to sex making them more vulnerable. The majority of those living in urban regions seemed to be more exposed to the disease than their counterparts living in rural areas. Those with a little level of education have low knowledge of mitigation measures of HIV risk and pose a greater risk of HIV in both sexes. A higher number of those seeking TB treatment are associated with the positivity of the disease (or already infected with HIV, opportunistic diseases). Similarly being uncircumcised exposes males to the disease which is consistent with studies by [48] while in females alone, being exposed to sex at an older age, attaining higher grade at school, an increase in the number of children born may lead to a reduction of HIV positivity. Our results are consistent with those of [49] that indicated literacy and urbanity as strong predictors of HIV acquisition, [50], that found that urban dwellers may increase HIV positivity through more contact with high-risk sexual individuals than rural residents. Age, little education and gender being predictors of the disease assert the findings in [51],[52].

A 95% sensitivity was required in ensuring that 95% of PLWH knew their status. With the XGBoost algorithm, we utilized 15 and 12 most predictive variables of males and females accordingly to establish 5 and 19, the number required to screen to know one individual with HIV in males and females respectively. These are within the range of 3 to 86 and 4 to 154 for community-based and facility-based screening respectively given in previous studies [53]. We identified 4·14% males and 10.81% females as a high-risk population in the second situation, which is consistent with previous studies that indicated that about seven women get new infections with HIV for every four men infected [54],[55]. In general, females performance in all our algorithms were very high compared to males counterparts in this study.

Our conclusion might reveal the social-behavioural identification of HIV and can enhance screening approaches in limited resources situations. There is a need to adapt HIV screening strategies that better target the adult population, those using contraceptives, urban dwellers, the little educated population, TB patients and uncircumcised men. There is an increased number of available surveys with individual-level data that is rich in demographic characteristics, social history, laboratory tests and results of various diseases. More

advanced approaches to utilize them can effectively assist in preventing, diagnosing and testing HIV and other diseases. This approach can be implemented by clinicians and community health care workers in identifying individuals who might benefit from the testing and counselling for HIV processes. This may also disclose individuals who may need PrEP, among other risk reduction strategies.

List Of Abbreviations

UNAIDS: The Joint United Nations Programme; HIV: Human Immunodeficiency Virus; AIDS: Acquired Immune Deficiency Syndrome; PHIA: Population-based HIV Impact Assessment; DHS: Demographic and Health Surveys; ART: antiretroviral therapy; UNAIDS: The Joint United Nations Programme; PrEP: preexposure prophylaxis; CDC: Centers for Disease Control and Prevention; PEPFAR: President's Emergency Plan for AIDS Relief; EN: Elastic Net; kNN : k-Nearest Neighbors; RF: RandomForest; SVM: Support Vector Machine; LGBT: Light Gradient Boosting; MICE: multiple imputations with chained equations; SFFS: sequential forward floating selection; SHAP: SHapley Additive exPlanations.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

Not applicable

Competing interests

The authors declare that they have no competing interests. All authors approved the final

manuscript.

Funding

Not applicable

Authors' contributions

CM conceptualised the idea, processed data into the software and delivered the results for this manuscript. PM contributed to the interpretation of the results. CM drafted the initial manuscript and all authors made substantial revisions to the work. PM commented on the final draft of the manuscript and CM finalised the text. All authors read and approved the final manuscript.

Authors' information

[1]African Center of Excellence in Data Science, University of Rwanda, Kigali BP 4285, Rwanda. [2]College of Engineering, Carnegie Mellon University Africa, Kigali BP 6150, Rwanda. [3]Oxford Man Institute of Quantitative Finance, Oxford University, Oxford OX2 6ED, UK. [4]College of Science and Technology, University of Rwanda, Kigali, Rwanda. [5]College of Business and Economics, University of Rwanda, Kigali, Rwanda

Acknowledgements

Not applicable.

References

[1] ‘UNAIDS data 2020’. https://www.unaids.org/en/resources/documents/2020/unaids-data (accessed Oct. 29, 2020).

[2] ‘Zambia | UNAIDS’. https://www.unaids.org/en/regionscountries/countries/zambia (accessed Nov. 10, 2020).

[3] 'Fast-track commitments to end AIDS by 2030 | UNAIDS'. https://www.unaids.org/en/resources/documents/2016/fast-track-commitments (accessed Nov. 30, 2020).

[4] '2016 United Nations Political Declaration on Ending AIDS sets world on the Fast-Track to end the epidemic by 2030'. https://www.unaids.org/en/resources/presscentre/pressreleaseandstatementarchive/2016/june/20160608_PS_HLM_PoliticalDeclaration (accessed Oct. 29, 2020).

[5] L. E. Bain, C. Nkoke, and J. J. N. Noubiap, 'UNAIDS 90–90–90 targets to end the AIDS epidemic by 2020 are not realistic: comment on "Can the UNAIDS 90–90–90 target be achieved? A systematic analysis of national HIV treatment cascades"', *BMJ Glob. Health*, vol. 2, no. 2, Mar. 2017, doi: 10.1136/bmjgh-2016-000227.

[6] D. F. Cuadros *et al.*, 'Mapping the spatial variability of HIV infection in Sub-Saharan Africa: Effective information for localized HIV prevention and control', *Sci. Rep.*, vol. 7, no. 1, Art. no. 1, Aug. 2017, doi: 10.1038/s41598-017-09464-y.

[7] L. C. Zulu, E. Kalipeni, and E. Johannes, 'Analyzing spatial clustering and the spatiotemporal nature and trends of HIV/AIDS prevalence using GIS: the case of Malawi, 1994-2010', *BMC Infect. Dis.*, vol. 14, no. 1, p. 285, May 2014, doi: 10.1186/1471-2334-14-285.

[8] H. Huerga *et al.*, 'Who Needs to Be Targeted for HIV Testing and Treatment in KwaZulu-Natal? Results From a Population-Based Survey', *J. Acquir. Immune Defic. Syndr. 1999*, vol. 73, no. 4, pp. 411–418, Dec. 2016, doi: 10.1097/QAI.0000000000001081.

[9] S. Blower and B. J. Coburn, 'Maximising the effect of combination HIV prevention in Kenya', *Lancet Lond. Engl.*, vol. 384, no. 9952, p. 1426, Oct. 2014, doi: 10.1016/S0140-6736(14)61859-6.

[10] S. O. Aral, E. Torrone, and K. Bernstein, 'Geographical targeting to improve progression through the sexually transmitted infection/HIV treatment continua in different populations', *Curr. Opin. HIV AIDS*, vol. 10, no. 6, pp. 477–482, Nov. 2015, doi: 10.1097/COH.0000000000000195.

[11] 'Social and Behavioural Aspects of the HIV Epidemic–A Review on JSTOR'. https://www.jstor.org/stable/2982186 (accessed Nov. 06, 2020).

[12] S. A. Rizza, R. J. MacGowan, D. W. Purcell, B. M. Branson, and Z. Temesgen, 'HIV Screening in the Health Care Setting: Status, Barriers, and Potential Solutions', *Mayo Clin. Proc.*, vol. 87, no. 9, pp. 915–924, Sep. 2012, doi: 10.1016/j.mayocp.2012.06.021.

[13] C. Celum and R. Barnabas, 'Reaching the 90-90-90 target: lessons from HIV self-testing', *Lancet HIV*, vol. 6, no. 2, pp. e68–e69, Feb. 2019, doi: 10.1016/S2352-3018(18)30289-3.

[14] D. Kwarisiima *et al.*, 'High rates of viral suppression in adults and children with high CD4+ counts using a streamlined ART delivery model in the SEARCH trial in rural Uganda and Kenya', *J. Int. AIDS Soc.*, vol. 20, no. Suppl 4, 2017, doi: 10.7448/IAS.20.5.21673.

[15] R. V. Barnabas *et al.*, 'Uptake of antiretroviral therapy and male circumcision after community-based HIV testing and strategies for linkage to care versus standard clinic referral: a multisite, open-label, randomised controlled trial in South Africa and Uganda', *Lancet HIV*, vol. 3, no. 5, pp. e212–e220, May 2016, doi: 10.1016/S2352-3018(16)00020-5.

[16] C. C. Iwuji *et al.*, 'Universal test and treat and the HIV epidemic in rural South Africa: a phase 4, open-label, community cluster randomised trial', *Lancet HIV*, vol. 5, no. 3, pp. e116–e125, Mar. 2018, doi: 10.1016/S2352-3018(17)30205-9.

[17] J. Sidey-Gibbons and C. Sidey-Gibbons, 'Machine learning in medicine: a practical introduction', *BMC Med. Res. Methodol.*, vol. 19, Mar. 2019, doi: 10.1186/s12874-019-0681-4.

[18] J. L. Marcus, W. C. Sewell, L. B. Balzer, and D. S. Krakower, 'Artificial Intelligence and Machine Learning for HIV Prevention: Emerging Approaches to Ending the Epidemic', *Curr. HIV/AIDS Rep.*, vol. 17, no. 3, pp. 171–179, 2020, doi: 10.1007/s11904-020-00490-6.

[19] A. E. Klon, M. Glick, and J. W. Davies, 'Application of machine learning to improve the results of high-throughput docking against the HIV-1 protease', *J. Chem. Inf. Comput. Sci.*, vol. 44, no. 6, pp. 2216–2224, Dec. 2004, doi: 10.1021/ci0497861.

[20] J. L. Marcus, L. B. Hurley, D. S. Krakower, S. Alexeeff, M. J. Silverberg, and J. E. Volk, 'Use of electronic health record data and machine learning to identify candidates for HIV pre-exposure prophylaxis: a modelling study', *Lancet HIV*, vol. 6, no. 10, pp. e688–e695, Oct. 2019, doi: 10.1016/S2352-3018(19)30137-7.

[21] V. Palanisamy and R. Thirunavukarasu, 'Implications of big data analytics in developing healthcare frameworks – A review', *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 31, no. 4, pp. 415–425, Oct. 2019, doi: 10.1016/j.jksuci.2017.12.007.

[22] A. I. L. LearningNeuroscienceNeurotech·July 8 and 2019, 'Two new algorithms can identify patients at risk of HIV', *Neuroscience News*, Jul. 08, 2019. https://neurosciencenews.com/hiv-algorithms-14441/ (accessed Nov. 02, 2020).

[23] S. Kumari, U. Chouhan, and S. K. Suryawanshi, 'Machine learning approaches to study HIV / AIDS infection : A Review', 2012, doi: 10.21786/bbrc/10.1/6.

[24] J. S. Lee, E. Paintsil, V. Gopalakrishnan, and M. Ghebremichael, 'A comparison of machine learning techniques for classification of HIV patients with antiretroviral therapy-induced mitochondrial toxicity from those without mitochondrial toxicity', *BMC Med. Res. Methodol.*, vol. 19, no. 1, p. 216, Nov. 2019, doi: 10.1186/s12874-019-0848-z.

[25] E. Orel, R. Esra, J. Estill, S. Marchand-Maillet, A. Merzouki, and O. Keiser, 'Machine learning to identify socio-behavioural predictors of HIV positivity in East and Southern Africa', *medRxiv*, p. 2020.01.27.20018242, Jan. 2020, doi: 10.1101/2020.01.27.20018242.

[26] M. L. Petersen *et al.*, 'Super learner analysis of electronic adherence data improves viral prediction and may provide strategies for selective HIV RNA monitoring', *J. Acquir. Immune Defic. Syndr. 1999*, vol. 69, no. 1, pp. 109–118, May 2015, doi: 10.1097/QAI.0000000000000548.

[27] W. Zheng, L. Balzer, M. van der Laan, and M. Petersen, 'Constrained binary classification using ensemble learning: an application to cost-efficient targeted PrEP strategies', *Stat. Med.*, vol. 37, no. 2, pp. 261–279, 2018, doi: 10.1002/sim.7296.

[28] C. M. Obermeyer and M. Osborn, 'The utilization of testing and counseling for HIV: a review of the social and behavioral evidence', *Am. J. Public Health*, vol. 97, no. 10, pp. 1762–1774, Oct. 2007, doi: 10.2105/AJPH.2006.096263.

[29] K. M. De Cock, J. L. Barker, R. Baggaley, and W. M. El Sadr, 'Where are the positives? HIV testing in sub-Saharan Africa in the era of test and treat', *AIDS Lond. Engl.*, vol. 33, no. 2, pp. 349–352, 01 2019, doi: 10.1097/QAD.0000000000002096.

[30] S. Ahmed *et al.*, 'Lost opportunities to identify and treat HIV-positive patients: results from a baseline assessment of provider-initiated HIV testing and counselling (PITC) in Malawi', *Trop. Med. Int. Health*, vol. 21, no. 4, pp. 479–485, 2016, doi: 10.1111/tmi.12671.

[31] 'PHIA Project Document Manager - Datasets'. https://phia-data.icap.columbia.edu/files (accessed Oct. 29, 2020).

[32] 'Population-based HIV Impact Assessment (PHIA) Data Use Manual - Google Search'. https://www.google.com/search?q=Population-based+HIV+Impact+Assessment+(PHIA)+Data+Use+Manual+%E2%80%8B&oq=Population-based+HIV+Impact+Assessment+(PHIA)+Data+Use+Manual+%E2%80%8B&aqs=chrome..69i57.585j0j7&client=ubuntu&sourceid=chrome&ie=UTF-8 (accessed Oct. 29, 2020).

[33] M. Kuhn and K. Johnson, *Feature Engineering and Selection: A Practical Approach for Predictive Models*, 1st edition. Chapman and Hall/CRC, 2019.

[34] S. Buuren and C. Groothuis-Oudshoorn, 'MICE: Multivariate Imputation by Chained Equations in R', *J. Stat. Softw.*, vol. 45, Dec. 2011, doi: 10.18637/jss.v045.i03.

[35] H. Zou and T. Hastie, 'Regularization and variable selection via the elastic net', *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, 2005, doi: 10.1111/j.1467-9868.2005.00503.x.

[36] Z. Zhang, 'Introduction to machine learning: k-nearest neighbors', *Ann. Transl. Med.*, vol. 4, no. 11, Jun. 2016, doi: 10.21037/atm.2016.03.37.

[37] L. Breiman, 'Random Forests', *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[38] C. Cortes and V. Vapnik, 'Support-vector networks', *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[39] T. Chen and C. Guestrin, 'Xgboost: A scalable tree boosting system', in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[40] J. Fan, X. Ma, L. Wu, F. Zhang, X. Yu, and W. Zeng, 'Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data', *Agric. Water Manag.*, vol. 225, p. 105758, Nov. 2019, doi: 10.1016/j.agwat.2019.105758.

[41] B. C. Vickery, 'Reviews : van Rijsbergen, C. J. Information retrieval. 2nd edn. London, Butterworths, I978. 208pp', *J. Librariansh.*, vol. 11, no. 3, pp. 237–237, Jul. 1979, doi: 10.1177/096100067901100306.

[42] C. Goutte and E. Gaussier, 'A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation', Apr. 2005, vol. 3408, pp. 345–359, doi: 10.1007/978-3-540-31865-1_25.

[43] G. Heinze, C. Wallisch, and D. Dunkler, 'Variable selection – A review and recommendations for the practicing statistician', *Biom. J.*, vol. 60, no. 3, pp. 431–449, 2018, doi: https://doi.org/10.1002/bimj.201700067.

[44] 'A Unified Approach to Interpreting Model Predictions'. https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions (accessed Nov. 02, 2020).

[45] K. Johnson and A. Way, 'Risk Factors for HIV Infection in a National Adult Population: Evidence From the 2003 Kenya Demographic and Health Survey', *JAIDS J. Acquir. Immune Defic. Syndr.*, vol. 42, no. 5, pp. 627–636, Aug. 2006, doi: 10.1097/01.qai.0000225870.87456.ae.

[46] K. L. Dunkle, R. K. Jewkes, H. C. Brown, G. E. Gray, J. A. McIntryre, and S. D. Harlow, 'Gender-based violence, relationship power, and risk of HIV infection in women attending antenatal clinics in South Africa', *Lancet Lond. Engl.*, vol. 363, no. 9419, pp. 1415–1421, May 2004, doi: 10.1016/S0140-6736(04)16098-4.

[47] 'Miles to go - Global AIDS update 2018 | UNAIDS'. https://www.unaids.org/en/20180718_GR2018 (accessed Nov. 30, 2020).

[48] K. E. Agot, J. O. Ndinya-Achola, J. K. Kreiss, and N. S. Weiss, 'Risk of HIV-1 in Rural Kenya: A Comparison of Circumcised and Uncircumcised Men', *Epidemiology*, vol. 15, no. 2, pp. 157–163, 2004.

[49] Z. Baranczuk *et al.*, 'Socio-behavioural characteristics and HIV: findings from a graphical modelling analysis of 29 sub-Saharan African countries', *J. Int. AIDS Soc.*, vol. 22, no. 12, p. e25437, 2019, doi: 10.1002/jia2.25437.

[50] R. K. Sing and S. Patra, 'What Factors are Responsible for Higher Prevalence of HIV Infection among Urban Women than Rural Women in Tanzania?', *Ethiop. J. Health Sci.*, vol. 25, no. 4, Art. no. 4, Oct. 2015, doi: 10.4314/ejhs.v25i4.5.

[51] A. B. M. Kharsany and Q. A. Karim, 'HIV Infection and AIDS in Sub-Saharan Africa: Current Status, Challenges and Opportunities', *Open AIDS J.*, vol. 10, pp. 34–48, Apr. 2016, doi: 10.2174/1874613601610010034.

[52] M. N. I. Mondal and M. Shitan, 'Factors affecting the HIV/AIDS epidemic: An ecological analysis of global data', *Afr. Health Sci.*, vol. 13, no. 2, Art. no. 2, Sep. 2013, doi: 10.4314/ahs.v13i2.15.

[53] A. B. Suthar *et al.*, 'Towards Universal Voluntary HIV Testing and Counselling: A Systematic Review and Meta-Analysis of Community-Based Approaches', *PLOS Med.*, vol. 10, no. 8, p. e1001496, Aug. 2013, doi: 10.1371/journal.pmed.1001496.

[54] R. C. Dellar, S. Dlamini, and Q. A. Karim, 'Adolescent girls and young women: key populations for HIV epidemic control', *J. Int. AIDS Soc.*, vol. 18, no. 2 Suppl 1, p. 19408, 2015, doi: 10.7448/IAS.18.2.19408.

[55] 'Women and girls, HIV and AIDS', *Avert*, Jul. 20, 2015. https://www.avert.org/professionals/hiv-social-issues/key-affected-populations/women (accessed Nov. 13, 2020).

Figures

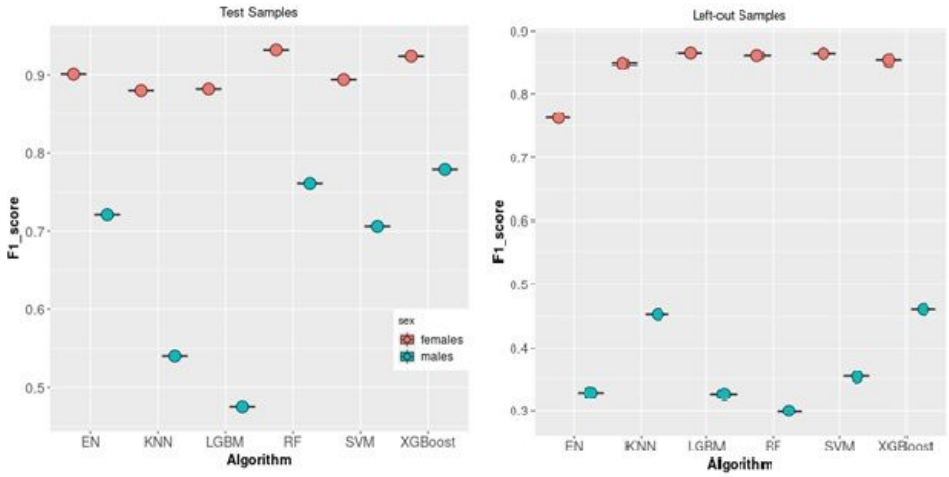


Figure 1

f1 scores Boxplot on methods used on test and left-out samples per sex

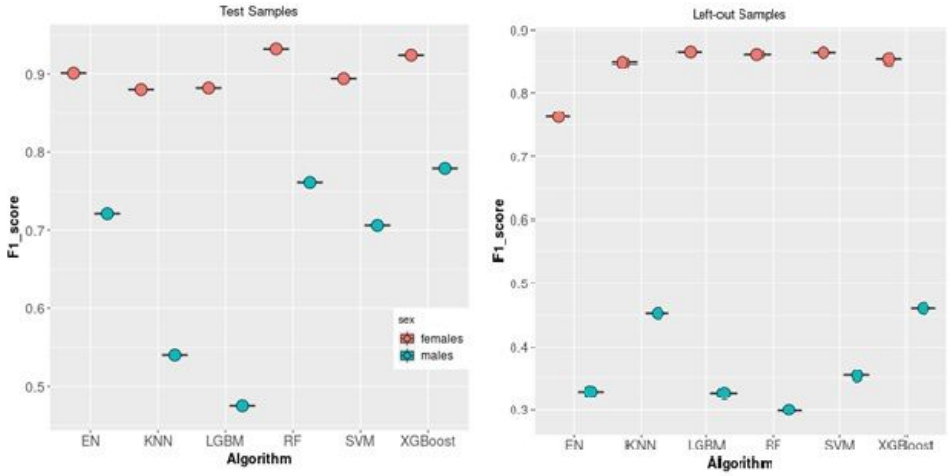


Figure 1

f1 scores Boxplot on methods used on test and left-out samples per sex

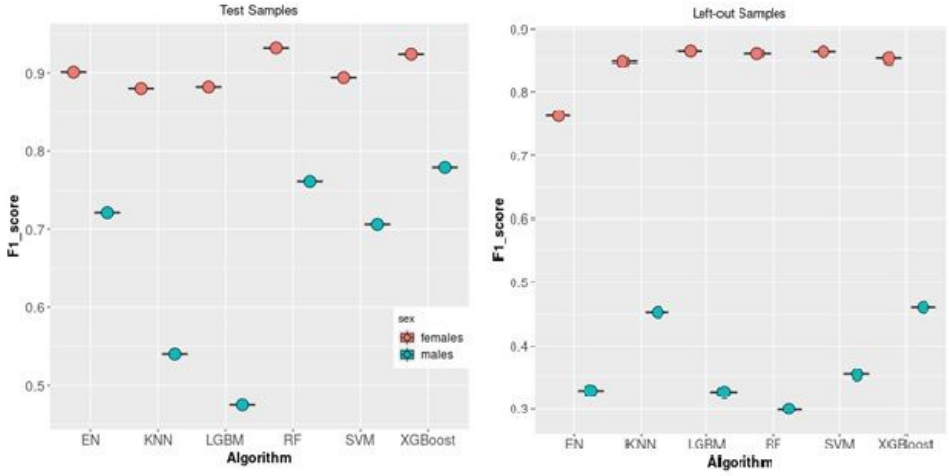
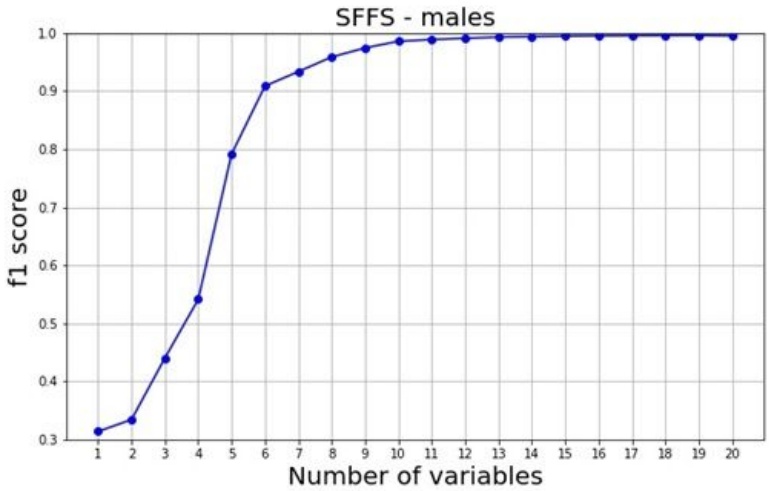


Figure 1

f1 scores Boxplot on methods used on test and left-out samples per sex



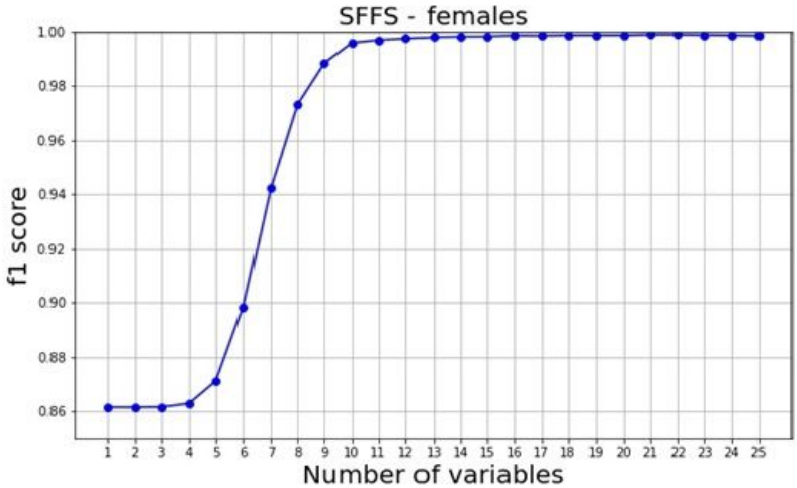
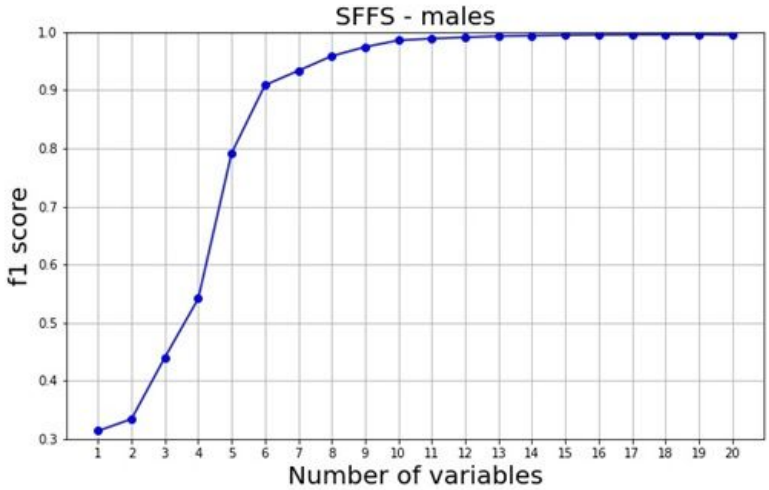


Figure 2

Sequential floating forward selection (SFFS)



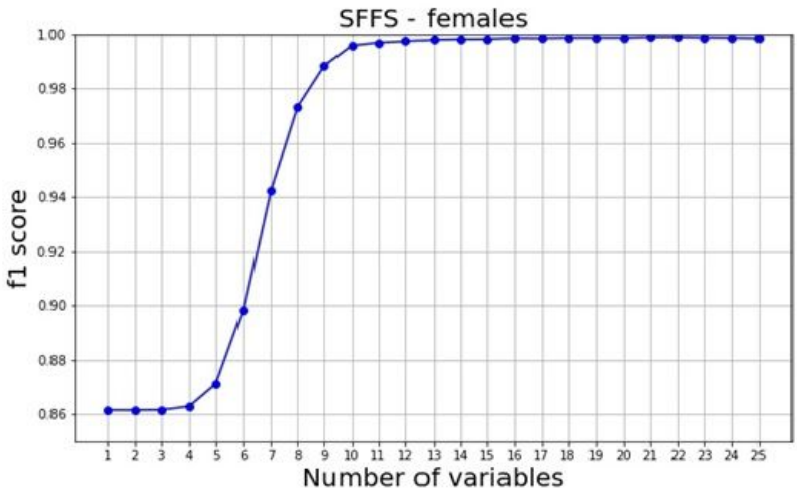


Figure 2

Sequential floating forward selection (SFFS)

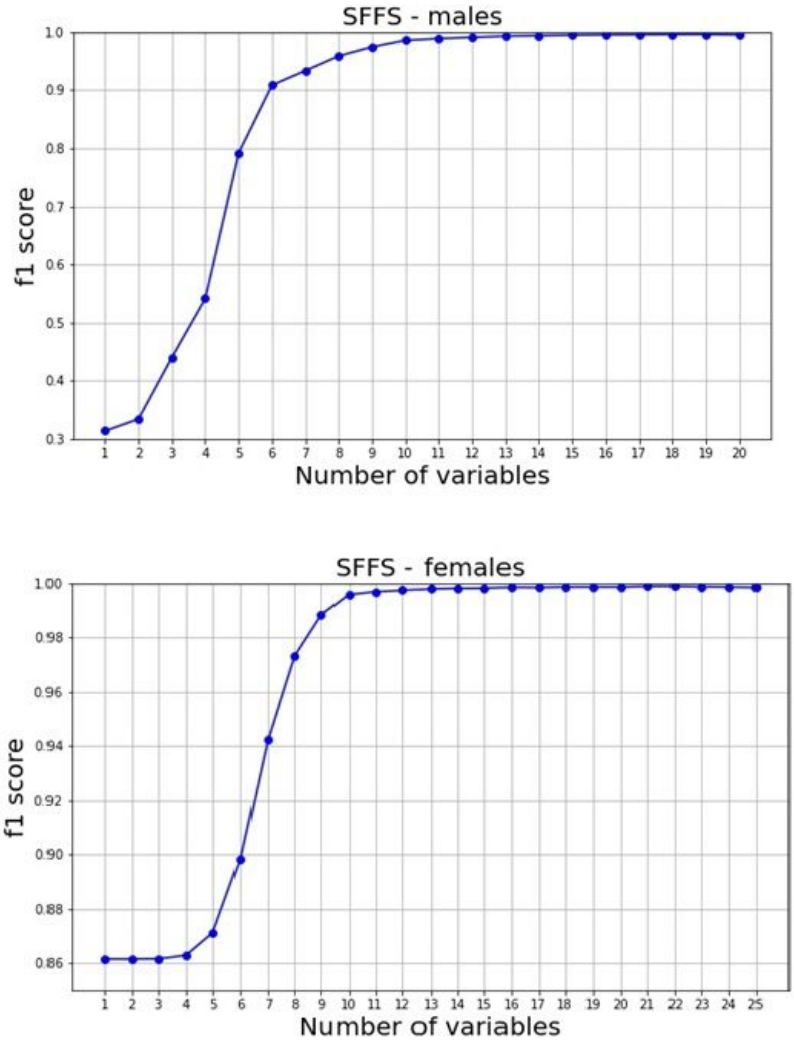


Figure 2

Sequential floating forward selection (SFFS)

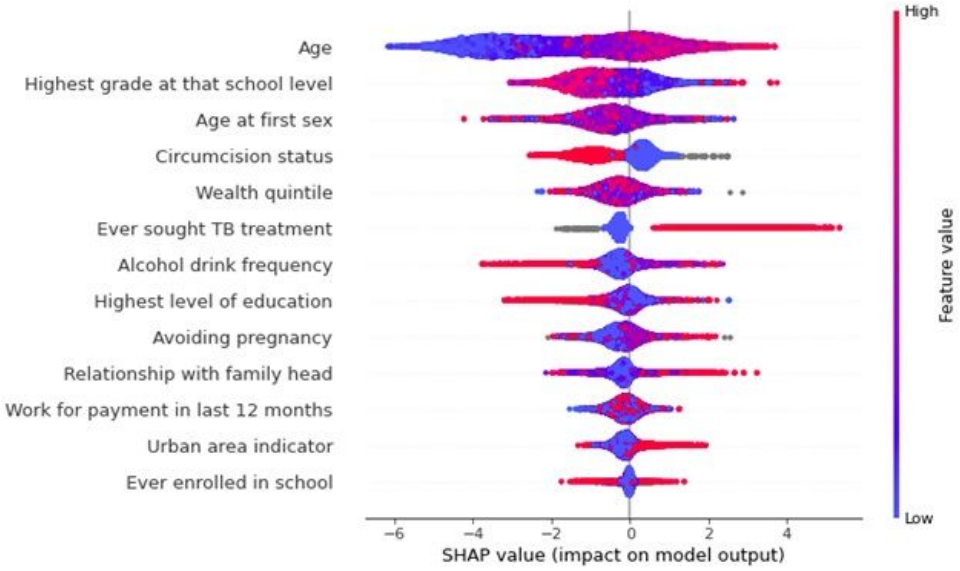


Figure 3

SHAP summary plots for HIV status predictors in male individuals

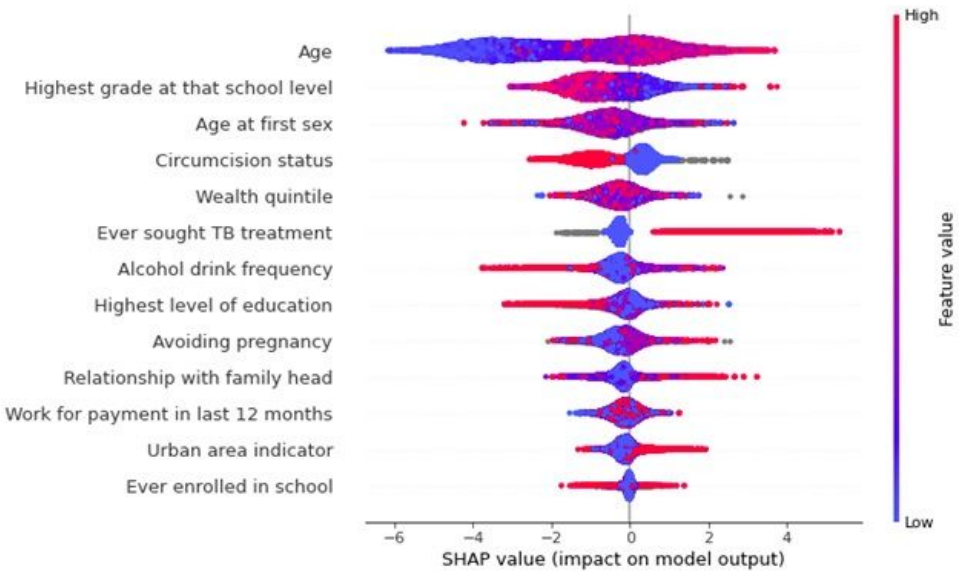


Figure 3

SHAP summary plots for HIV status predictors in male individuals

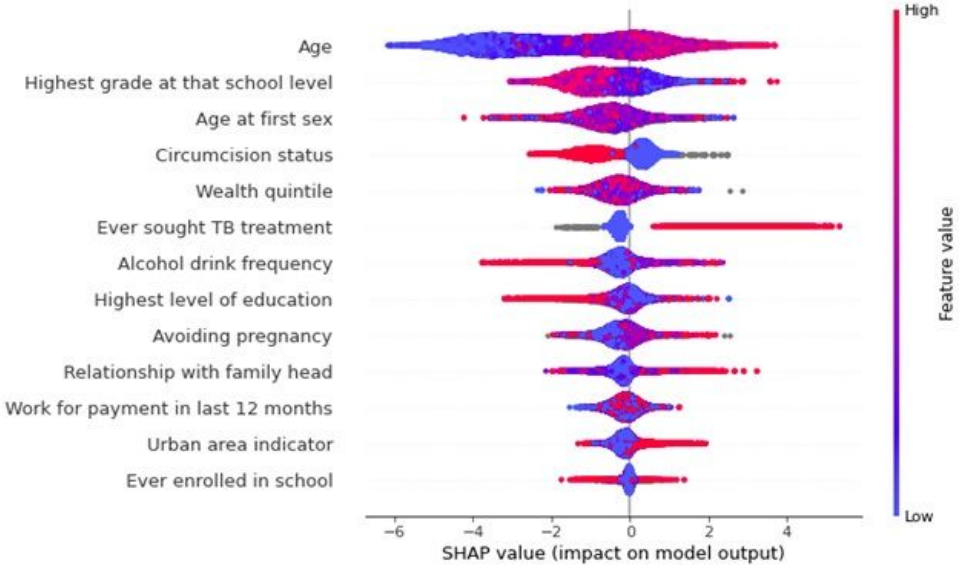


Figure 3

SHAP summary plots for HIV status predictors in male individuals

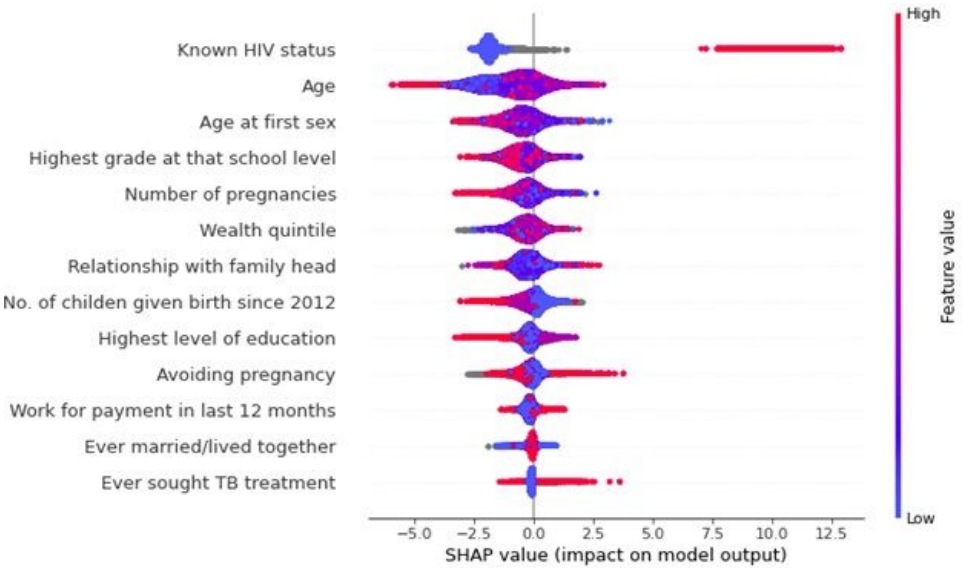


Figure 4

SHAP summary plots for HIV status predictors in female individuals

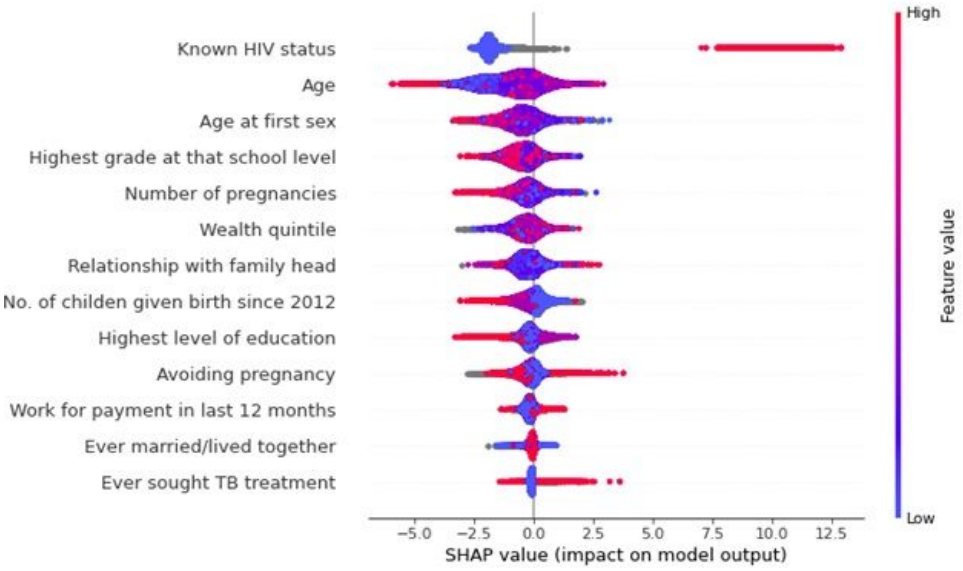


Figure 4

SHAP summary plots for HIV status predictors in female individuals

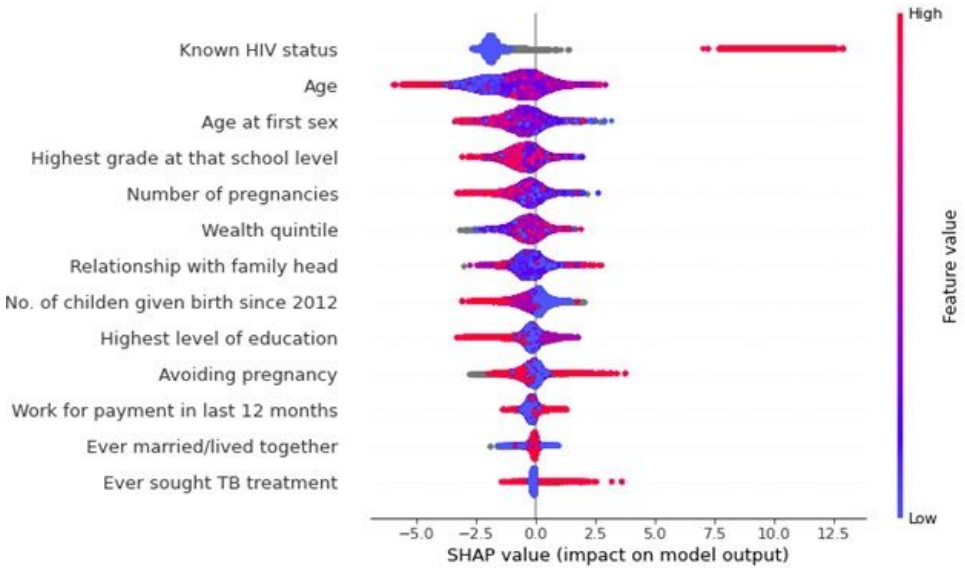


Figure 4

SHAP summary plots for HIV status predictors in female individuals