UNIVERSITY OF THE WITWATERSRAND

FACULTY OF HEALTH SCIENCES

SCHOOL OF PUBLIC HEALTH

Title: Predictive modelling of risk factors associated with HIV infection amongst adults in South Africa in 2016 using supervised machine learning.

Name: Munyaradzi Nyahuma

Student no: 2727881

Supervisor: Prof Eustasius Musenge

A Research Report Submitted to the Faculty of Health Sciences, the University of the Witwatersrand in partial fulfilment of the requirements for the Degree of Masters in Epidemiology in the field of Epidemiology & Biostatistics

June 2025

Dedication

I would like to dedicate this achievement to my wonderful wife Tinashe Tombe Nyahuma, whose unwavering encouragement and belief in me made it possible for me to pursue this course in Epidemiology and Biostatistics. Your support, both emotional and practical, kept me focused and motivated even when the challenge seemed insurmountable. You have been my rock, and this accomplishment is as much yours as it is mine.

To my children Makanakaishe and Elisha Nyahuma, thank you for your patience and understanding during the times I had to be away, immersed in my studies. Your love and smiles were the light at the end of long study days, and I am grateful for the sacrifices you made during this long journey.

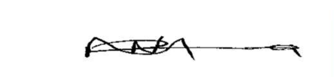
To my brother Lovemore Nyahuma thank you for your constant presence and unwavering support throughout my academic journey and reminding me of the importance of perseverance. Your words of wisdom and belief in me kept me focused and determined.

Above all, I give thanks to God for his guidance, strength and grace throughout this journey.

Declaration

I, Munyaradzi Nyahuma, student number 2727881, declare that this research report is my original work. It is submitted in partial fulfilment of the requirements for the degree of Masters in Epidemiology in the field of Epidemiology and Biostatistics, at the University of Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination to this or any other university.

I have read the sections on referencing and plagiarism in the WITS Plagiarism Policy. I am aware that plagiarism is wrong and that the University of Witwatersrand may take disciplinary action should plagiarism be found in this work. I have followed the required conventions in referencing the thoughts and ideas of others. I confirm that all the work submitted in the thesis titled “Predictive modelling of risk factors associated with HIV infection among adults in South Africa in 2016 using supervised machine learning**”** is my own unaided work except where I have explicitly indicated otherwise.



Signed

Date 12/06/2025

Acknowledgement

I would like to express my sincere gratitude to my supervisor, Prof. E. Musenge, from the Division of Epidemiology and Biostatistics at the School of Public Health, Faculty of Health Sciences, University of Witwatersrand, for his invaluable support and guidance throughout my research journey.

Abstract

Background

HIV incidence has been declining globally but it remains high in Sub Saharan Africa where 60% of cases in the world originate, with South Africa alone contributing 19% of global HIV population. In this era of artificial intelligence, predicting individuals at risk of HIV is vital for targeted prevention, early intervention and ultimately reducing HIV mortality. We identified the risk factors associated with HIV infection among adults in South Africa in 2016.

Method

A secondary data analysis was conducted using cross sectional data from the DHS 2016 survey, which included 2726 women and 2136 men in South Africa, to identify risk factors associated with HIV infection. One variable was removed whenever the correlation was >80%. Lasso regression and generalized variance selection were used to identify the best predictors of HIV infection. Systematic descriptive analysis was first conducted. This was followed by sex stratified multilevel logistic regression to control for confounding. Missing data was handled using multiple imputations. Risk factors were then ranked based on model coefficients, and a risk scoring tool was developed. Finally, the stratum specific predictive performance of the models was compared using F1 scores on the test data.

Results

HIV prevalence was 28.7% in females and 14.9% in males. Common risk factors associated with higher odds of infection in both sexes included older age (females AOR: 2.49; 95% CI: 1.42–4.38; males AOR: 6.72; 95% CI: 3.39–13.33) and condom use during the last sexual encounter (females AOR: 2.11; 95% CI: 1.54–2.89; males AOR: 2.76; 95% CI: 1.26–6.05). Protective factors included male circumcision (AOR: 0.48; 95% CI: 0.32–0.71). For females, XGB model performed best (F1: 0.82 weighted, 0.77 balanced, 0.72 unweighted data). For males, XGB led on weighted (0.92) and balanced (0.83) data; logistic regression was best on unweighted (0.80).

Conclusion

HIV prevalence was twice as much in females compared to males. Of note is highest prevalence among black women and men, residing in the KwaZulu-Natal (KZN) region. Teen girls also have a higher rate of HIV infection than teen boys. Additionally, women and men who used a condom during their last sexual encounter were more likely to be infected. Circumcision was protective amongst men.

XGB was the model with the best predictive performance. Public health efforts should prioritize targeted HIV prevention and treatment strategies for high-burden groups, including Black women and men in KwaZulu-Natal and adolescent girls. Expanding access to youth-centered sexual health education, promoting consistent condom use, and scaling up male circumcision programs could significantly reduce HIV transmission. XGB highlights the potential of machine learning tools to support targeted HIV interventions by early identification of high-risk people.

Key words: HIV risk factors, predictive modelling, South Africa, Machine learning.

Contents

List of figures and tables

ABBREVIATIONS

AIC	Akaike information criterion
AIDS	Acquired immune-deficiency syndrome
ART	Antiretroviral therapy
AUC	Area under the curve
BIC	Bayesian information criterion
ROC	Receiver operating curve
DHS	Demographic and Health Survey
GOF	Goodness of fit
HIV	Human immuno-deficiency virus
KNN	K nearest neighbor
MSP	Multiple sexual partners
PrEP	Pre-exposure prophylaxis
SES	Socioeconomic status
SSA	Sub-Saharan Africa
STI	Sexually transmitted infections
UNAIDS	Joint United Nations Programme on HIV/AIDS
USAID	United States Agency for International Development
XGB	Extreme Gradient boosting

1. INTRODUCTION

1.1 Background

Despite global reduction in human immunodeficiency virus (HIV) incidence, it remains high in areas such as Sub Saharan Africa (SSA) where 60% of cases in the world originate (1). South Africa is one of the countries in SSA. In 2015 in South Africa total population estimated to be infected with HIV was between 6 400 000 to 6 700 000 with an estimated prevalence of 12.7% (2). This accounted for approximately 19% of the global population living with HIV at the time, highlighting a substantial share of the global burden. As of 2022, the prevalence of the disease in South Africa was 17.8% for adults aged 15-49 (3). The prevalence was 23.5% for women and 12.1% for men aged 15-49 respectively. Population size living with HIV including children stood at 7 600 000 (3). If unabated HIV infection among adults will continue to present public health concerns and economic impact.

There have been advances in developing methods to combat HIV infection and knowledge on risk factors associated with HIV is vast but interventions have not proved to be effective (1). Ineffective intervention will lead to delays in attaining the Joint United Nations Programme on HIV/AIDS (UNAIDS) target of 2030 which states that 95% of infected humans are aware of their status, 95% of HIV diagnosed humans are on treatment and 95% have viral suppression (4). In SSA HIV is mainly transmitted through heterosexual intercourse, which significantly contribute to the epidemic in children through vertical transmission (5). Women are mainly affected , representing 58% of all individuals living with HIV in the region (5). Moreover there is still a significant gap in HIV prevention technologies initiated by women particularly for those who find it challenging to negotiate existing options such as abstinence, condom use, behavior change, medical male circumcision, or early initiation of treatment within their relationships (5).

HIV transmission is influenced by the size of high risk groups such as intravenous drug users, sex workers and men who have sex with men and their broader interactions with the general population (6). To curtail the spread of infection interventions targeted at high risk groups should be implemented (6). Use of machine learning in health care to identify risk factors can help with combating HIV transmission.

There are various techniques used to select risk factors used for predictive modelling with Least absolute shrinkage selection operator (LASSO) being the commonly used (4) (7) (8). LASSO has LASSO 1 (L1) which shrinks coefficients of less important features to zero and LASSO 2 (L2) which penalizes coefficients but does not set them to zero (9). Other techniques used include sequential forward floatation selection and SHAP values which can demonstrate the contribution of each feature in prediction (4) (7) (10).

Machine learning is artificial intelligence techniques that can be used to make predictions or decisions in public health and other fields (4) (7) . It uses algorithms that are data driven to identify non-complex linear patterns and detect complex interactive patterns among risk factors (7). When developing predictive models, machine learning is not based on any assumptions or requirements for statistical inference (4). It is a variety of techniques that describes how computers identify patterns among data (10).

Machine learning has two classes namely, supervised and unsupervised learning based on data type used on the algorithm (10). Training of algorithm on labelled data is supervised learning. For unsupervised learning the algorithm determines the data structure on its own. Supervised learning can be classified further depending on nature of outcome variable being predicted. Categorical outcomes utilise classification algorithms whereas continuous outcomes use regression algorithms (10). Machine learning can be used to identify risk factors and rank them based on their impact on the South African population.

Comparison of performance of the machine learning tools can be based on using a confusion matrix, receiver operating curve (ROC) or F1 scores. If the predictive precision, sensitivity, specificity, F1 score and ROC is high the better the model is at predicting infection (4) (11).

This study utilised Demographic and Health Survey data (DHS 16) to identify and rank risks in order of importance so that interventions can be directed at the major risk factors to minimise impact in South Africa, as well as identifying individuals at high risk of infection using machine learning.

1.2 Literature review

Literature indicates that some of the risk factors associated with HIV infection include age at first sexual encounter, highest education level, transactional sex, multiple sexual partners, unprotected intercourse, barriers to use of prophylaxis and presentation to care for high risk groups and recently infected individuals, submission through force, substance abuse before sexual activity and gender based violence (1) (12) (8) (13).

According to Fieggen et al. and Mutai et al. the risk factors associated with HIV infection can be classified into 3 broad categories namely behavioral, socioeconomic and demographic. The risk factors for HIV transmission also differ significantly from one population to the other (7) (10).

1.2.1 Behavioral factors associated with HIV infection

On behavioral factors, sexual behavior and substance abuse are the major driving forces in HIV transmission (14). This is supported by a cohort study by Ayton on adolescents girls and young woman in South Africa showing that most adolescents who were HIV seronegative did not report any use of contraceptive, alcohol, drugs and any pregnancy in the previous years (14). Use of intoxicating substances is also implicated on a Chinese study on men having sex with men (MSM) as a contributing factor to HIV incidence because they impair judgement and allow men to engage in sexual behaviors that they might avoid when sober such as anal sex, engaging in sex with numerous partners and low condom use (8). Identification of high risk groups based on sexual behavior, low risk perception and through screening of individuals attending healthcare or selective testing of those at high risk, can facilitate the use of pre exposure prophylaxis (PrEP) and counselling as an HIV transmission prevention measure (7) (12) (15).

1.2.2 Use of PrEP as a preventative measure

Use of antiretroviral PrEP in United States of America was approved in 2012 and was shown to reduce new HIV infections however uptake of PrEP is still lagging behind in areas with high HIV prevalence (12). Mayer et al. suggests that addressing the barriers in uptake of Prep will improve in reducing new infections. He suggested that knowledge and awareness of PrEP should be improved, perception of low risk for HIV and stigma from family members and health care providers on uptake of PrEP should be addressed. He also emphasized on easy access to PrEP for those who are high risk groups, and its provision at a low cost and perceived side effects of the medication (12). The uptake of PrEP is lower in high risk groups such as MSM, juveniles and adult females and has been attributed to stigma, lack of awareness and support (12).

1.2.3 Demographic factors associated with HIV infection

On demographic factors a cross sectional study conducted in Zimbabwe by Birri and Musenge shows that HIV prevalence is higher in females as compared to males due to their biological susceptibility, social cultural factors where women experience gender inequality and sexual violence (4). Kansiime et al. in a cohort study conducted in four African countries on HIV incidence also confirms the same findings that females among high risk groups had high chances of getting infection compared to males (1) .The disparity in incidence between young women and men of similar age is attributed to the women dating older men who have high probability of HIV infection (16) (17) . Birri also found out that having a younger female household head was associated with HIV positivity (4).

1.2.4 Socioeconomic factors associated with HIV infection

On socioeconomic factors Kansiime et al. found out that poverty was one of the main reasons why transactional sex is a driving force on HIV transmission (1). People living in poverty often lack financial resources to cover essential needs such as food, healthcare, shelter and education thereby engaging in transactional sex to secure the necessities. Adolescents girls and young women who are vulnerable are easily forced into transactional sex (1) (14). A cross sectional study by Sodeinde et al. in Nigeria showed that 41.5% of adolescent females who engage in transactional sex for survival and luxury purposes come from houses without food security (18) .

According to Fieggen et al. some risks are modifiable at population level and individual level such as low condom use and having multiple sexual partners using behavior change communication (10). Some risks are not modifiable such as time travelling to work especially for truck drivers, age, sex, such that PrEP and counselling is recommended in such situations (10). Public health interventions should be targeted and applicable to specific individuals and communities rather than using a blanket approach. Targeting increases efficiency and effectiveness of intervention programmes (10) (19). Machine learning can be incorporated to help combat HIV infection.

1.2.5 Machine learning use in risk prediction

Machine learning can effectively adjust and quantify risk by understanding complex feature patterns, though it risks overfitting (10).

Primary purpose of classification machine learning is accurate prediction compared to inference for traditional statistical methods (10). A confusion matrix is used to evaluate model performance. Model predicted outcome on test dataset is contrasted to the outcome seen on the dataset at set probability threshold. Performance metrics like sensitivity, specificity, accuracy, and precision are calculated. Challenges of the confusion matrix is predictions are made at a particular probability threshold and it’s difficult to assess what would happen at a different probability cut off (10).

Due to challenges presented by the confusion matrix the area under the receiver operating curve (AUROC) provides a way of visualizing a tradeoff between sensitivity and specificity across all levels of probability and is the gold standard of measuring performance of a model (10). Unlike the confusion matrix which determines sensitivity and specificity at set probability threshold, for receiver operating curve (ROC) it determines at all probability levels and the results are not dependent on the relative prevalence of the outcome (10). F1 scores, the harmonic mean of precision and recall can be used if data is unbalanced. The F1 scores can be preferred over ROC which is sensitive to unbalanced data (4).

There are a variety of models that can be used in predicting infections in high risk groups examples include extreme gradient boost network (XGB), k nearest neigh bour (KNN), logistic regression and lasso regression models (11). In their studies on algorithm comparison Birri and Musenge found out that of the six models that they compared XGB was the best performing model followed by the KNN and the Elastic net was the worst performing model (4). Xiao also found out that the XGB was the best performing model as well as logistic regression, however Elastic net did well which is in contradiction with Birri and Musenge’ findings (11). When Mutai et al. compared five predictive algorithms he found out that the XGB was the best performing model while the KNN was the worst performing (7).

1.2.6 Challenges with predictive modelling

Fieggen et al. also highlighted the challenge with predicting HIV infection is that HIV incidence varies greatly by population therefore impacting on the baseline probability of infection (10). The rate at which new HIV infections occur is not uniform across different groups of people (10). Different populations by geographical location, age, gender, sexual orientation and behavior have varying levels of risk for HIV infection. Variation impacts baseline probability of infection which is the initial likelihood of an individual contracting HIV before any intervention or additional factors are considered. Models need to accurately represent the differences across diverse populations to provide reliable predictions (10). Failure to account for differences can lead to incorrect estimates of infection risk and undermine effectiveness of prevention and treatment strategies (10).

Variation on incidence by population can be corrected by segmentation and stratification which is dividing population into subgroups based on age, gender, sexual orientation and geographic location then have specific incidence rate and baseline probability of infection for each group (10). Can also be corrected by including geography as a feature in the model (10).

1.3 Problem statement

In South African context, HIV prevalence has been increasing despite significant advances in treatment and management of HIV (2) (3) (20). If left uncontrolled, HIV can place immense strain on the healthcare systems and significantly reduce economic productivity leading to severe threat of survival to human populations. There is a lack of studies that are comparing predictive modelling techniques for women and men separately to identify the most effective approach for predicting the risk factors (4).

1.4 Justification

Exposures linked with high probability of HIV infection have been studied widely but HIV prevalence continues to increase (1). The research will use machine learning to assist with early detection of individuals at risk and their presentation to care for PrEP or ART (8) (20) .Understanding of risk factors associated with HIV is important for determining strategies to prevent further transmission (21). The research seeks to contribute to the body of knowledge by informing public health interventions and strategies to reduce HIV infection rates through identification of high-risk individuals using a scoring system, recommend PrEP and the use of a model that can best predict infection. Compared to other studies research findings will be applicable to the rest of the country.

1.5 Research question, aim and objectives

1.5.1 Research question

How can supervised machine learning models be effectively utilized to identify risk factors associated with HIV infection, and to predict likelihood of infection among adults 15 years and above in South Africa in 2016?

1.5.2 Aim

To identify and rank key risk factors associated with HIV infection and to recommend a model which can improve prediction accuracy when identifying at risk individuals.

1.5.3 Objectives

1. To describe the characteristics and distribution of HIV infected adults who are 15 years and above in South Africa in 2016.

2. To identify key risk factors associated with HIV infection among adults 15 years and above in South Africa in 2016.

3. To select a predictive model that can identify individuals at risk of HIV infection among adults 15 years and above in South Africa in 2016.

The study sought to determine the risk factors linked with HIV infection among adults 15 years and above in South Africa in year 2016 and to identify and recommend a model that can best predict infection using supervised machine learning. The study will recommend intervention strategies to curb HIV infection at individual and population levels based on key risk factors.

1.6 Conceptual framework and research outline

Shown below is a diagrammatic representation of how the outcome variable is influenced by exposure variables. The conceptual framework presented in Figure 1 below has been done based on priori.

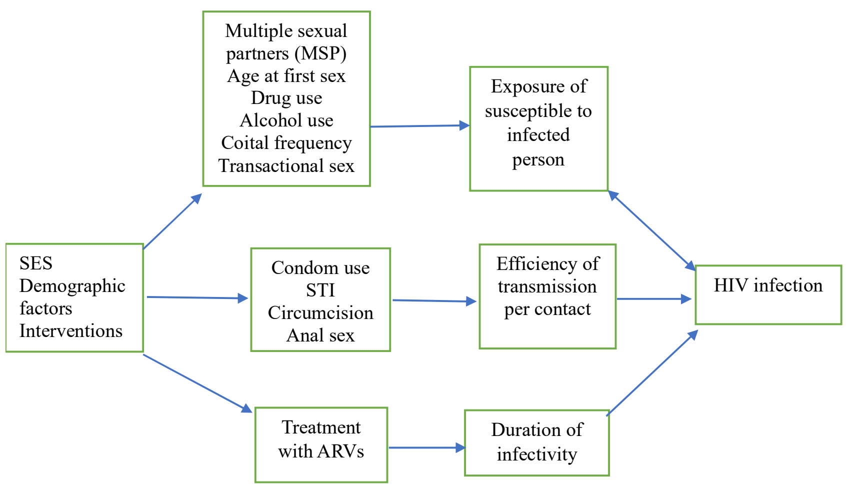


Figure 1.1 Conceptual framework for HIV

The outline of the research is as follows chapter 2 methods, chapter 3 results, chapter 4 discussion and conclusion and chapter 5 strengths and limitations.

2 METHODS

This section provides detailed description of the study setting, data collection, data management and the step-by-step statistical methods used in the research report.

2.1 Primary study setting

The Demographic and Health Survey (DHS) is a United States Agency for International Development (USAID) led initiative that involves nationally representative household surveys collecting data from individuals in selected households (22). The DHS program aims to help improve ways in which data on population, health and nutrition is collected, analyzed and disseminated. The data collected is used to facilitate planning, policy-making, and program management (22).

2.2 Primary study data collection

Data were collected between July 2016 and November 2016 using a two-stage sampling method. A total of 750 primary sampling units (clusters) were randomly selected from all clusters across the nine provinces. Within each of these selected clusters, households were then randomly chosen for participation (22). Provinces were classified into urban, farm and traditional areas resulting in 26 sampling strata from which 750 primary clusters were identified. Of the 750 sampling units 468 were in urban areas, 224 in traditional areas and 58 in farm areas (22). Qualifying adults 15 years and above from selected households in all the 9 provinces of South Africa were interviewed face to face using questionnaires by trained interviewers. Data was collected from a representative sample of eligible women and men aged 15-49 and 15-59 respectively.

2.2.1 Outcome variable (HIV) data collection

Voluntary HIV testing was done on participants who consented. Blood for HIV testing was provided by participants using a finger prick and was transported to the Global clinical laboratory in Durban as dried blood spots (DBS) on a filter paper (22) (23). The test procedure was anonymous as personal identifiers were not retained, and respondents did not receive results. Respondents were offered free HIV counselling and testing coupons at their nearest collaborating centers (23).

Blood samples were tested using enzyme linked immunosorbent assay (ELIZA) test kits and retesting of all positive samples was done with 5-10% of negative samples being retested with a second ELIZA (24). HIV test results sample weight was divided by one million to provide sample weights representative of the entire South African population (22). Sample weights were also applied to the enumeration areas from which individual households were selected (22). Figure below displays the provinces in South Africa from which participants originated (25)

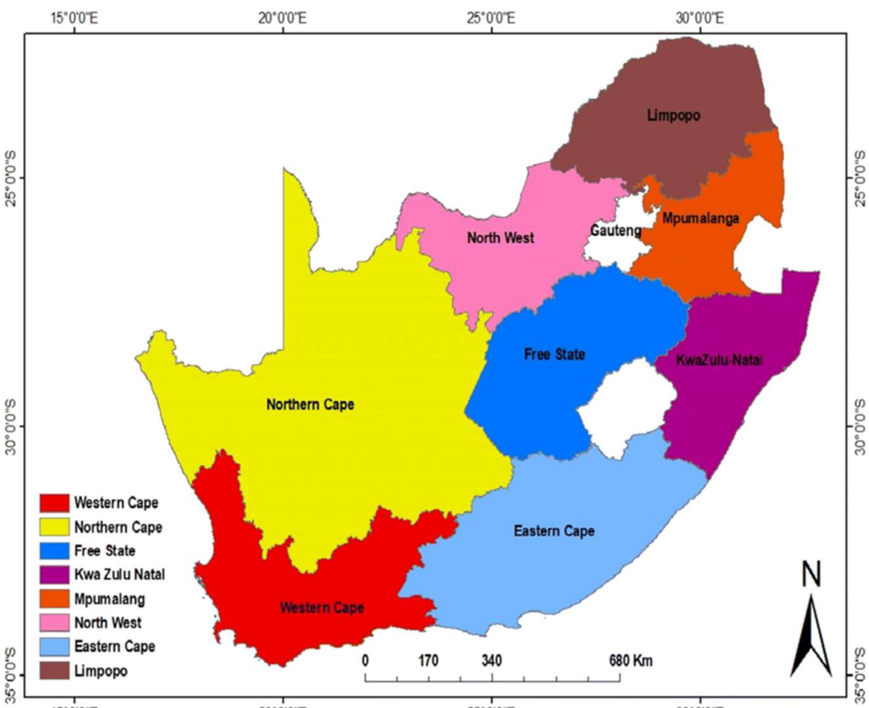


Figure 2.1 Map of South Africa showing different provinces of South Africa

2.2.2 Potential predictors during primary data collection

Potential predictors of HIV status were derived from self-reported data obtained through face-to-face interviews addressing HIV-related behaviors and experiences.

2.3 Current study design description

This was a cross-sectional study design on DHS data collected in 2016 using secondary data analysis. It identified and ranked risk factors linked with HIV infection in adults 15 years and above in South Africa using supervised machine learning.

2.4 Data source for current study

Permission to use the data was obtained from DHS following a written request citing study title and objectives. Authorization was granted to access and download the data. The datasets that were retrieved in dta format were HIV data, Individual recode for woman, men recode for men and PR code for households. The data sets were imported into Stata version18 and then merged to form combined data. Data used in this research report can be requested using the following link (https://dhsprogram.com/data/dataset_admin/login_main.cfm) (26)

2.5 Study population

Study population are adults 15 years old and above residing in selected households from all 9 provinces of South Africa in 2016 and underwent HIV testing.

2.6 Power analysis

For the secondary data analysis, no sampling was done. Power of study was determined in Stata using one proportion with a 19.5% prevalence, size effect of 2%, average cluster size of 9.216, intraclass correlation of 0.03, design effect of 1.25 and significance level of 0.05. A sample size of 6912 people was sufficient to provide a power of 95.82%. Sample size was reduced to 5530 when a design effect of 1.25 was implemented.

2.7 Inclusion criteria

- Records of Males and females 15 years old and above residing in households selected through survey sampling method in all 9 South African provinces.

- Informed consent.

2.8 Exclusion criteria

- Records with more than 30% missing information.

2.9 Data management

Using Stata version 18 the following processes were followed to ensure that the data is clean, accurate and suitable for analysis.

The data was cleaned by retaining relevant variables to the study based on priori, removing one variable when the correlation coefficient was > 80% and checking for duplicates. Variables with greater than 30% missing values and uninformative ones were also removed. Renaming variables in the dataset was done for easy management and referencing. Variables that were similar in both male and female datasets were renamed to a similar variable name for easy appending and merging. After cleaning, data was transformed by grouping or recoding variables of interest as well as creating new variables in Stata. Records without HIV results were excluded from the analysis and data was saved and protected using password.

The diagram below shows the steps involved in creating DHS 16 data

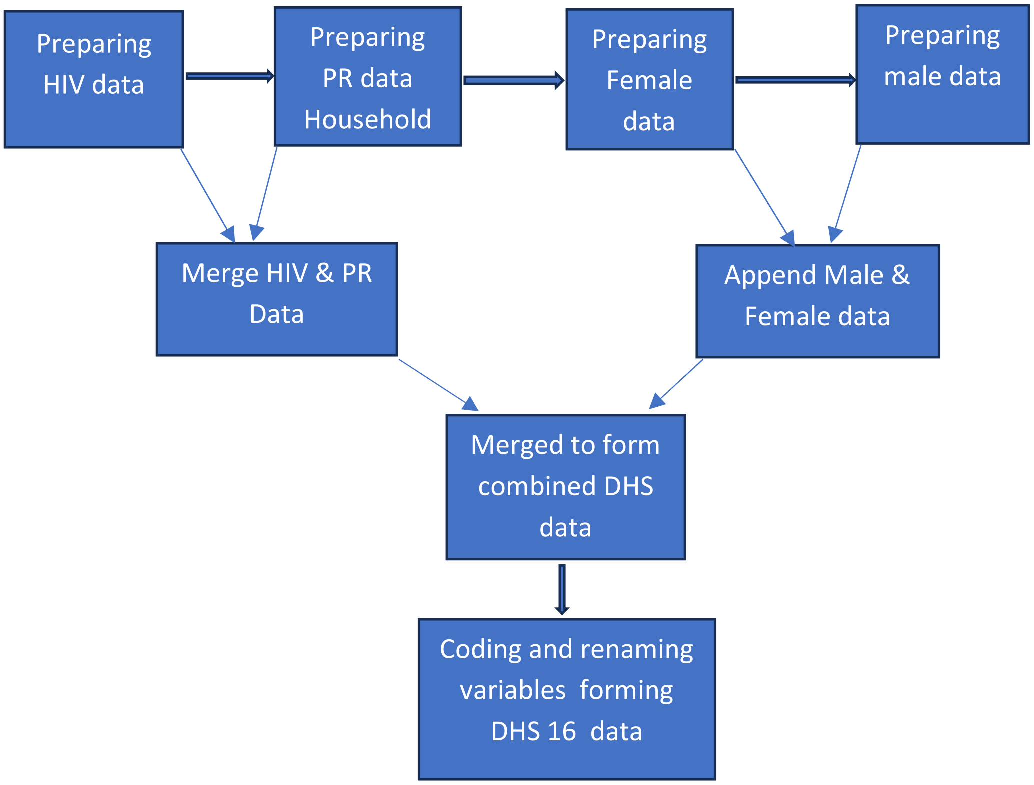


Figure 2.2 Data management process

2.10 Creation of new variables

After merging individual datasets to create a combined DHS dataset, variable names were updated for clarity. A categorical variable for contraceptive use was established, distinguishing between participants not using contraceptives and those using various methods—such as injections, condoms, abstinence, and sterilization. Additionally, the continuous variable for age was categorized into two groups namely young adults (15–24 years) and older adults (25 years and above). These steps resulted in the creation of the DHS 16 dataset.

2.11 Descriptive analysis of combined DHS 16 data and sex stratified data

Data was analyzed as unweighted and as weighted data to determine if there was an effect when using clustered data. When the data was survey set for clustered analysis the primary sampling unit used was cluster, weight variable was wt and strata was region. Descriptive analysis was done using tables with statistics of outcome HIV which is binary using frequencies and percentages and determining number of participants in the study. Continuous variables were described using means, standard deviations, median and IQR whilst categorical variables were described using percentages and frequencies. Following analysis of combined data, the data was stratified by sex for further analysis. Graphical representation of relationships was done using bar plots, box plots and maps on both combined and sex stratified data.

2.12 Feature selection process on stratified data

Data was first stratified by sex and LASSO regression was used for potential feature selection, LASSO fits linear models and utilizes L1 regularization to enhance model performance by penalizing the absolute size of the coefficients (9).

Following LASSO, generalized variance selection was used to select features to be used for prediction. Generalized variance selection is a statistical technique that identifies the most informative variables in a dataset by evaluating their contribution to the overall variability of the data (27) . The method identifies features that contribute significantly to explaining the variance in the target variable and avoids redundant features that overlap in explaining variance thereby preventing multicollinearity issues. Multivariate dispersion or variability for a set of variables is used as the selection criterion (27) (28) . For variables with a correlation greater than 80% one variable was removed. For the female dataset a total of 53 potential predictors were selected and 20 variables were removed because they had more than 30% missing observations. LASSO regression was done on the 33 variables, and 23 potential predictors were left. Of the remaining 23 potential predictors generalised variance selection was conducted and 16 predictors were left. On the male dataset there were 53 predictors as well and 17 were removed because they had more than 30% missing observations. Of the remaining 36 Lasso regression was used and 20 variables remained. On the 20 potential predictors generalized variance selection (GV selection) selected 11. For an overview of missing data patterns see appendices (tables 2.1 to 2.4)

2.13 Goodness of fit of using selected variables

The Hosmerleme show goodness of fit was used to assess the goodness of fit of the variables selected by generalized variance selection. It assesses how well a logistic regression model fits the data to explain the outcome variable. A high p_value (>0.05) suggests a good fit (29).

2.14 Checking for confounding

Confounding was assessed using a logistic regression followed by the chest command to determine which variables were confounders. A 10% change between the crude odds ratio and adjusted odds ratio was regarded as a confounder (30) (31).

2.15 Multivariable analysis on sex stratified data

Multivariable analysis was done using predictors that were selected by generalized variance after potential predictors were selected by lasso. Potential features selected by LASSO regression were 23 for females and 20 for males. Following generalized variance selection, the predictors that were left were used for multivariable analysis on the female dataset and male dataset. The remaining predictors were 16 for females and 11 for males. Interactions terms were also assessed during multivariable analysis. Preprocessed significant predictors were then ranked based on model coefficients along with the development of a risk scoring tool. Statistically, significant preprocessed features from multivariable analysis were later used for predictive modelling.

2.16 Data splitting and handling of missing data for predictive modelling

For machine learning, risk factors with more than 30% missing information were dropped to ensure quality and reliability of the dataset and reduce bias in imputation (32) (33). Multiple imputations with chained equations were done on risk factors with missing data that was less than 30%. Missing data on risk factors with less than 30% missing observations, data was assumed to be missing at random implying that the probability of information missing in a variable depends only on available information (32) (33). Data was stratified by gender and multivariate analysis was done on separate male and female datasets.

Both female and male data was split by clusters into 80 % training and 20% testing and not as individuals in clusters. According to Napoli and White clustering-based splits can lead to better model performance than random splits. This is due to the more representative nature of the training and test sets (34). The clusters maintain integrity by ensuring that similar data points are grouped together (34) . Clustering enhances generalizability of models by ensuring that training and testing datasets reflect real world situations more accurately. Individuals in the same cluster are more likely to be similar to each other than those in different clusters. If splitting is not done by cluster it can lead to data leakage and overfitting thereby affecting generalizability of the model (34).

Random splitting can lead to overly optimistic performance estimates since it does not consider the relationship between data points. Random splits are associated with bias in the results if data has clusters (35). Random splitting is commonly used and straightforward but does not account for the underlying distribution of the data (35) (36).

2.17 Data preprocessing

In preparation for predictive modelling, data preprocessing was conducted and included the conversion of binary variables to numeric, label encoding ordinal variables, transforming and normalising continuous variables, standardising and normalising discrete variables, and applying one-hot encoding to nominal variables.

2.18 Predictive modelling

For predictive modelling comparison of the performance metrics of three models namely XGB, Logistic regression and KNN was done using original weighted data and balanced weighted data in R Studio while performance on original and unweighted data was done in Orange. Data balancing was performed using synthetic minority oversampling technique (SMOTE). In R studio model performance was assessed using only statistically significant predictors from multivariable analysis of preprocessed predictors for both males and females. In Orange, available predictors were used to assess model performance due to how some variables appeared when the data file was exported as a csv file some appeared as strings and could not be used. Performance metrics were compared using F1 scores, Precision recall and AUROC, precision and recall.

2.19 Hyperparameter tuning

Hyperparameter tuning was conducted in two stages for XGB and KNN models. For XGB the first stage learning rate was optimized using a random grid search with 10 points while other parameters were kept at their default values (37). The stage involved 5-fold cross validation to ensure robust evaluation. After identification of optimal learning rate, it was fixed and a second round of tuning was conducted to find the best values for the remaining hyperparameters using a range of values for the hyperparameters (37). For KNN the optimum number of neighbors was determined using a random grid search with 10 points, exploring the number of neighbor values from 3 to 50. On the second stage with fixed neighbors the search for the best weight and distance metrics was performed. For logistic regression a generalized linear model engine was used alongside 5-fold cross validation. Preprocessing was further implemented on categorical variables using one hot encoding.

2.20 Visual workflow of methodology in R Studio

The figure below illustrates the visual workflow of the methodology used in R Studio on original weighted data and balanced weighted data for females and males.

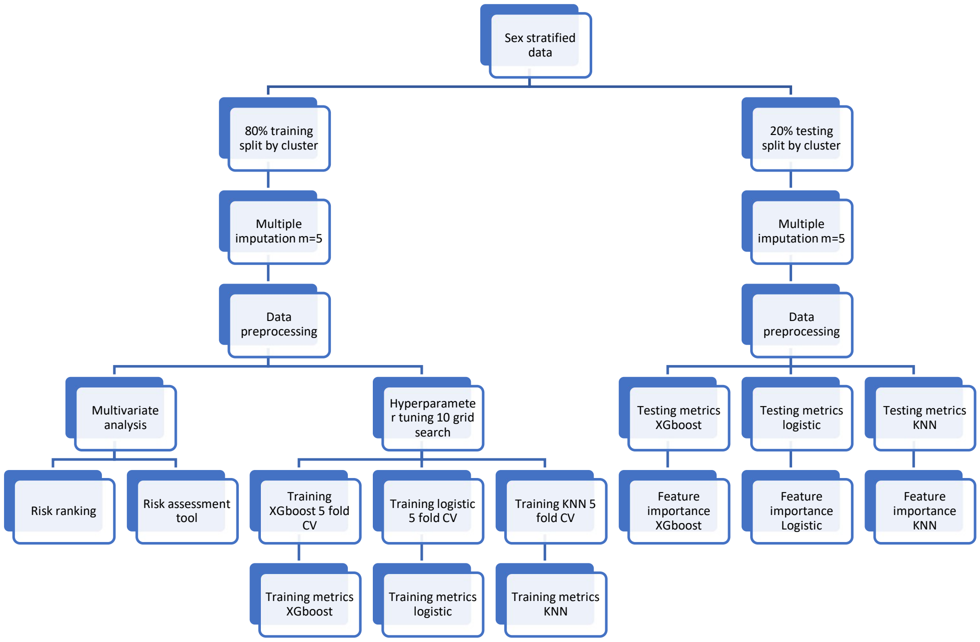


Figure 2.3 Visual workflow of methodology used in R Studio

2.21 Visual workflow of methodology used in Orange

The figure below shows visual workflow methodology in Orange on original unweighted data for males and females.

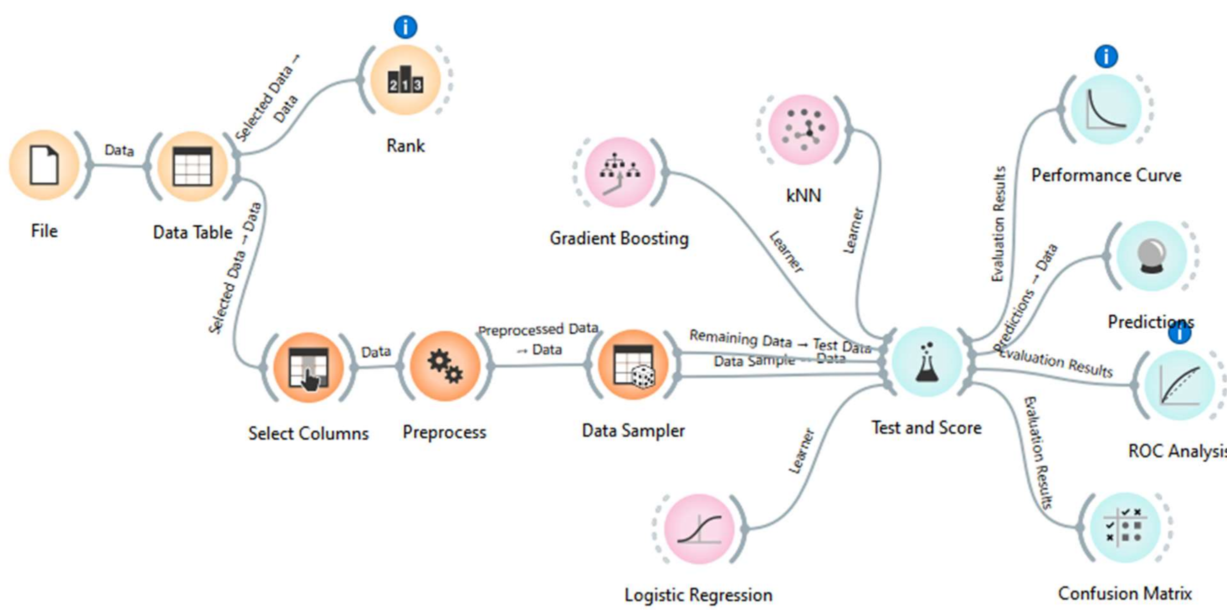


Figure 2.4 Visual workflow of methodology used in Orange

2.22 Ethical considerations

The current study of secondary data analysis was conducted using data collected from humans at household levels. Ethical clearance was obtained from the University of Witwatersrand Human Research Ethics Committee in Johannesburg, South Africa (Clearance Certificate Number M2409110) see appendix 2. Data was received as deidentified data and it is impossible to identify or contact any participants who were part of the survey. Dataset was only accessible to me as the principal investigator and my supervisor and was stored on a computer for conducting the analysis. Data was saved and protected using password. No duplicates were made.

3. RESULTS

3.1 Introduction

The following section outlines the results obtained from the analysis of the objectives of the study in section 1.53. Results are presented in tabular and graphical form as well as summary narratives.

3.2 Distribution of participants by sex and cluster

The figure below illustrates the graphical representation of participants at each level of analysis. DHS 16 data (combined) was utilized to describe the characteristic and distribution of participants, while sex stratified data was used for descriptive and multivariable analysis focusing separately on male and female participants. Some records from DHS 16 were not used on sex stratified data because their gender was not recorded. Figure below illustrates how samples were first stratified by sex and then split by cluster into training and testing sets.

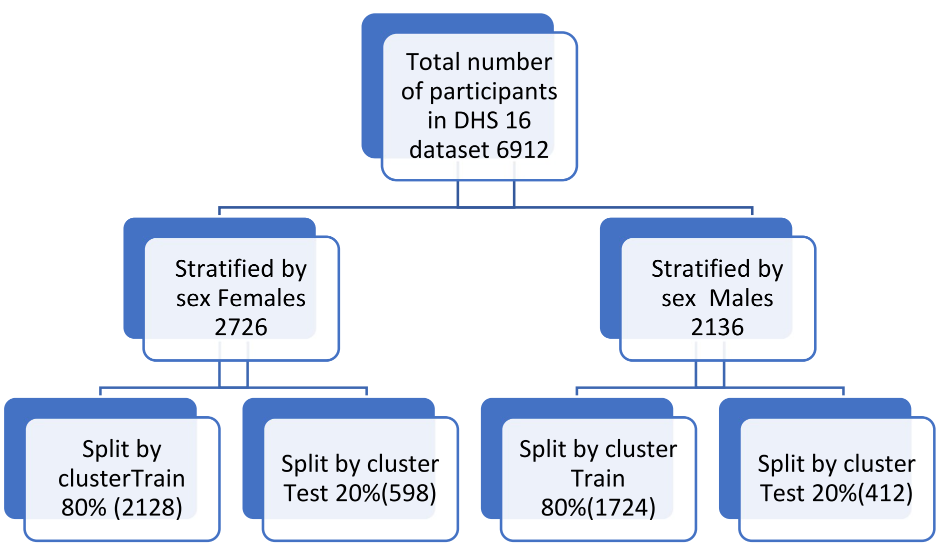


Figure 3.1 How samples for sex stratified analysis were obtained

3.3 Characteristics and distribution of participants in DHS 16 dataset

The tables below show weighted and unweighted characteristics and distribution of participants in the combined datasets. This was done to show the effect of clustering and weighting of data during data analysis.

Table 3.1 Unweighted characteristics and distribution of participants

	HIV negative (N = 5559)	HIV positive (N = 1353)	Total (N = 6912)	p-value
Respondent's current age				<0.001
Mean (SD)	29 (12)	35 (9)	31 (11)	
Median (Q1, Q3)	26 (20, 37)	34 (28, 42)	29 (21, 39)	
Age category				<0.001
Young Adults (15-24)	1634 (29%)	153 (11%)	1787 (26%)	
Old Adults (25+)	3925 (71%)	1200 (89%)	5125 (74%)	
Gender				<0.001
Male	1820 (48%)	316 (29%)	2136 (44%)	
Female	1953 (52%)	773 (71%)	2726 (56%)	
Educational level				<0.001
no education	110 (3%)	40 (4%)	150 (3%)	
primary	505 (13%)	204 (19%)	709 (15%)	
secondary	2809 (74%)	797 (73%)	3606 (74%)	
higher	349 (9%)	48 (4%)	397 (8%)	
Ethnicity				<0.001
black/african	3336 (88%)	1061 (97%)	4397 (90%)	
white	110 (3%)	1 (0%)	111 (2%)	
coloured	295 (8%)	26 (2%)	321 (7%)	
indian/asian	31 (1%)	1 (0%)	32 (1%)	
other	1 (0%)	0 (0%)	1 (0%)	
Marital status				<0.001
never in union	2476 (66%)	580 (53%)	3056 (63%)	
married	811 (21%)	195 (18%)	1006 (21%)	
living with partner	326 (9%)	196 (18%)	522 (11%)	
widowed	44 (1%)	38 (3%)	82 (2%)	
divorced	28 (1%)	11 (1%)	39 (1%)	
no longer living together/separated	88 (2%)	69 (6%)	157 (3%)	
Region				<0.001
western cape	222 (6%)	41 (4%)	263 (5%)	
eastern cape	573 (15%)	151 (14%)	724 (14%)	
northern cape	278 (7%)	40 (4%)	318 (6%)	
free state	450 (11%)	148 (13%)	598 (12%)	

	HIV negative (N = 5559)	HIV positive (N = 1353)	Total (N = 6912)	p-value
kwazulu-natal	514 (13%)	240 (22%)	754 (15%)	
north west	520 (13%)	162 (15%)	682 (14%)	
gauteng	329 (8%)	87 (8%)	416 (8%)	
mpumalanga	526 (13%)	183 (16%)	709 (14%)	
limpopo	511 (13%)	64 (6%)	575 (11%)	
Residence type				0.866
urban	2029 (52%)	574 (51%)	2603 (52%)	
rural	1894 (48%)	542 (49%)	2436 (48%)	
Employement status				0.009
No	2497 (66%)	674 (62%)	3171 (65%)	
Yes	1276 (34%)	415 (38%)	1691 (35%)	

In the unweighted data the overall HIV prevalence was 19.6%. The prevalence among females was nearly double that of males with rates of 28.4% for females and 14.8% for males. HIV prevalence rates by region were as follows Western Cape (15.6%), Eastern Cape (20.9%), Northern Cape (12.6%), Free State (24.7%), Kwazulu-Natal (KZN) (31.8%), North West (23.8%), Gauteng (20.9%), Mpumalanga (25.8%) and Limpopo (11.1%). For older adults, the HIV prevalence was 23.4% while it was 8.6% for young adults.

Among the participants, 56% were female and 44% were male, 74% were old adults and 26% were young adults. Of note is the proportion of HIV positivity by region. Among those who tested positive for HIV, the highest proportion came from KZN at 22% whilst Gauteng had 8%. HIV positivity rate was the same for married individuals and those living with a partner both at 18%. Among the participants who were positive 97% were black and 2% were coloured.

Table 3.2 Weighted characteristics and distribution of participants

	assay 1 result			
	HIV negative	HIV positive	Total	Test
N	3,933 (78.3%)	1,091 (21.7%)	5,023 (100.0%)	
Respondent's current age	30.143 (11.705)	34.609 (9.188)	31.113 (11.356)	<0.001

Age category				
Young Adults (15-24)	1,608 (40.9%)	147 (13.5%)	1,755 (34.9%)	<0.001
Old Adults (25+)	2,325 (59.1%)	944 (86.5%)	3,268 (65.1%)	
Gender				
Male	2,160 (54.9%)	378 (34.7%)	2,538 (50.5%)	<0.001
Female	1,773 (45.1%)	713 (65.3%)	2,485 (49.5%)	
Educational level				
no education	102 (2.6%)	33 (3.0%)	135 (2.7%)	<0.001
primary	489 (12.4%)	170 (15.6%)	659 (13.1%)	
secondary	2,882 (73.3%)	827 (75.8%)	3,709 (73.8%)	
Higher	459 (11.7%)	61 (5.6%)	520 (10.4%)	
Ethnicity				
black/african	3,354 (85.3%)	1,071 (98.2%)	4,424 (88.1%)	<0.001
White	185 (4.7%)	3 (0.2%)	187 (3.7%)	
coloured	336 (8.6%)	16 (1.4%)	352 (7.0%)	
indian/asian	58 (1.5%)	2 (0.2%)	59 (1.2%)	
Other	0 (0.0%)	0 (0.0%)	0 (0.0%)	
Marital status				
never in union	2,448 (62.2%)	568 (52.1%)	3,016 (60.0%)	<0.001
married	935 (23.8%)	202 (18.5%)	1,137 (22.6%)	
living with partner	358 (9.1%)	219 (20.1%)	578 (11.5%)	
widowed	45 (1.1%)	30 (2.8%)	75 (1.5%)	
divorced	45 (1.2%)	11 (1.0%)	56 (1.1%)	
no longer living together/separated	101 (2.6%)	60 (5.5%)	161 (3.2%)	
Region				
western cape	427 (10.8%)	86 (7.9%)	513 (10.2%)	<0.001
eastern cape	449 (11.4%)	105 (9.6%)	554 (11.0%)	
northern cape	87 (2.2%)	12 (1.1%)	100 (2.0%)	
free state	187 (4.8%)	60 (5.5%)	247 (4.9%)	
kwazulu-natal	620 (15.8%)	273 (25.1%)	894 (17.8%)	
north west	264 (6.7%)	79 (7.3%)	343 (6.8%)	
gauteng	1,157 (29.4%)	313 (28.7%)	1,470 (29.3%)	
mpumalanga	315 (8.0%)	109 (10.0%)	424 (8.4%)	
limpopo	427 (10.8%)	53 (4.8%)	479 (9.5%)	
Residence type				
Urban	2,663 (67.7%)	755 (69.2%)	3,418 (68.0%)	0.515
Rural	1,269 (32.3%)	336 (30.8%)	1,605 (32.0%)	
Employement status				
No	2,473 (62.9%)	668 (61.2%)	3,141 (62.5%)	0.555
Yes	1,460 (37.1%)	423 (38.8%)	1,883 (37.5%)	

Total sample: N = 5,023

In the weighted data, the overall HIV prevalence was 21.7%. The prevalence among females was almost double that of males with rates of 28.7% and 14.9% for females and males respectively. HIV prevalence among old adults was 3.5 times that of young adults, with rates of 28.9% for old adults and 8.4% for young adults.

Among the participants 49.5% were females and 50.5% were males, 65.1% were old adults and 34.9% were young adults.

Ethnical background shows that most participants were black making up 88% of the population, followed by coloured at 7%, white at 3.7% with Indian and Asians at 1.2%. The HIV prevalence amongst the groups were 24.2% for blacks, 1.6% whites, 4.5% coloured and 3.4% for Asians and Indians.

Two percent of the participants had no formal education with HIV prevalence of 24.4%. Those with primary education made 13.1% of the participants with a prevalence of 25.8%. Majority of the participants had secondary education 73.8% with a prevalence of 22.3% and those with higher education formed 10.4% of participants with a prevalence of 11.7%.

Results from marital status showed that those widowed had the highest HIV prevalence at 40%, followed by participants living with a partner at 37.9%, no longer living together 37.3%, divorced 19.6%, never in union 18.8% and lastly married participants at 17.8%

Prevalence by province showed KZN as the province with the highest HIV prevalence at 30.6%. Mpumalanga province was second with 25.7% followed by Free State 24.3%, Northwest 23.1%, Gauteng 21.3%, Eastern Cape 19%, Western cape 16.8%, Northern Cape 12.3% and Limpopo 11%. Applying survey weights led to slight changes in prevalence rates compared to unweighted data. For example, KZN, the weighted prevalence was 30.6% versus 31.8% unweighted, in the Western Cape 16.8% weighted versus 15.6% unweighted, Gauteng 21.3% weighted versus 20.9% unweighted and in Limpopo 11% weighted compared to 11.1% unweighted.

The table below shows weighted HIV prevalence by province

Table 3.3 Combined HIV prevalence for each province

Province	HIV Prevalence
KZN	30.6%
Mpumalanga	25.7%
Free State	24.3%
North West	23.1%
Gauteng	21.3%
Eastern Cape	19%
Western Cape	16.8%
Northern Cape	12.3%
Limpopo	11%

The map below shows combined weighted age prevalence for females aged 15-49 and males aged 15-59 in each province.

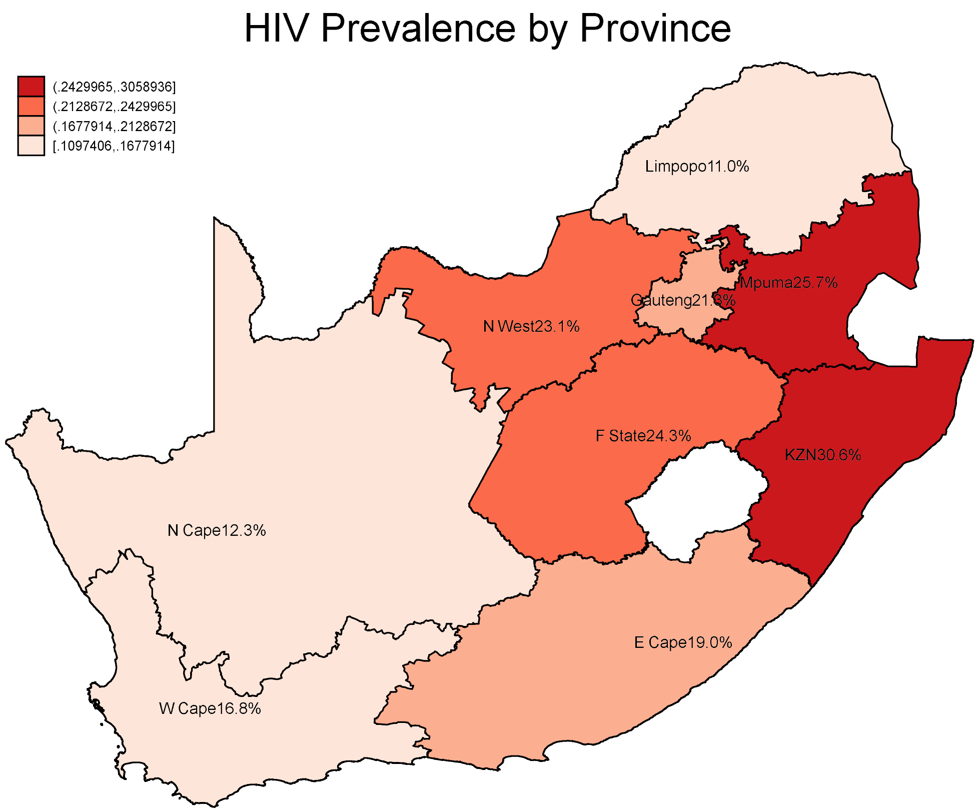


Figure 3.2 Combined HIV prevalence per province

Figure 3.3 below presents weighted HIV prevalence among females aged 15-49 years. Five of the nine provinces had prevalence rates of 30% or higher. KZN had the highest rate at 39.1% followed by Mpumalanga (34.2%), Eastern Cape (30.5%) and both North West and Free State at 30%. Gauteng reported a slightly lower prevalence of 27.6%. The provinces with the lowest rates were Western Cape (19.2%), Limpopo at 15.4% and Northern Cape (14%).

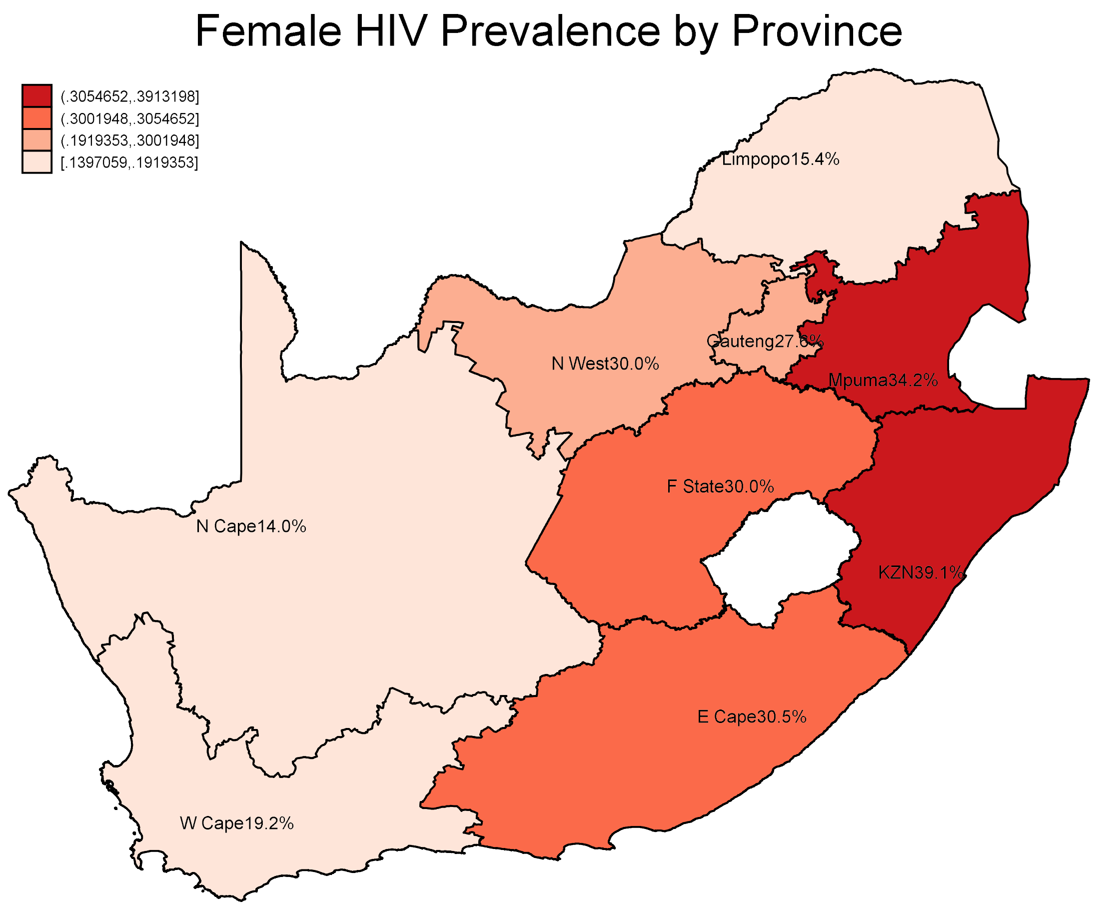


Figure 3.3 HIV prevalence for females in each province

Figure below presents weighted HIV prevalence amongst men aged 15-59. KZN had highest prevalence (20.4%), followed by Free State (18.6%), Mpumalanga (17.3%), North West (16.9%), Gauteng (15.7%) and Western Cape (14.5%). Lower prevalence rates were reported in Northern Cape (10.6%), Eastern Cape (8.4%) and Limpopo (6.1%).

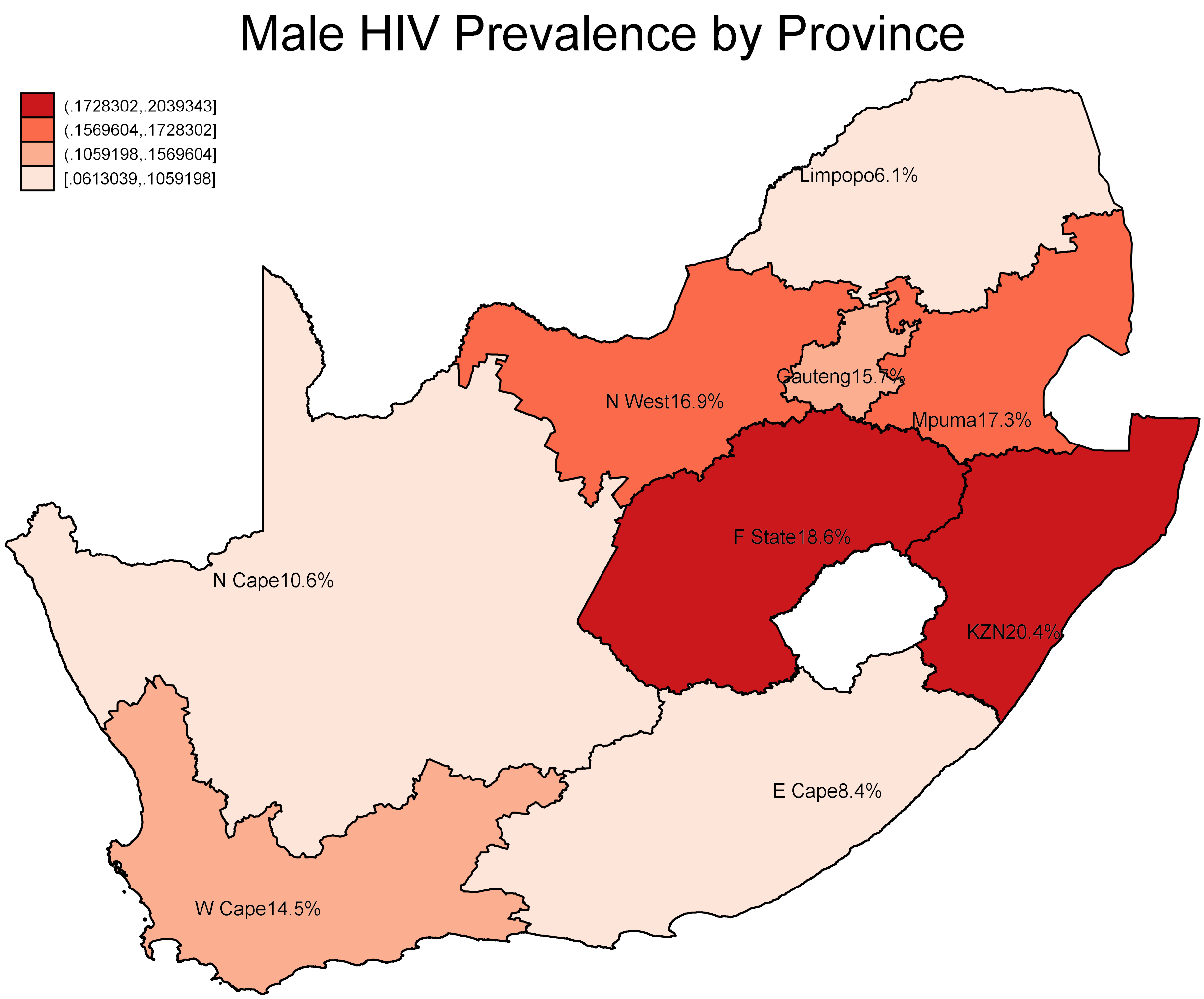


Figure 3.4 HIV prevalence for males in each province

The figure below presents the proportion of individuals with HIV across different ages separately for males (age 15-59) and females (age 15-49) separately. For males' prevalence peaks around 30 and 44 years with lower prevalence at younger and older ages. For females' prevalence starts increasing earlier (late teens), peaks around 30-40 and then declines reaching rates between 40% and 50%. The figure suggest that younger females are disproportionately affected by HIV.

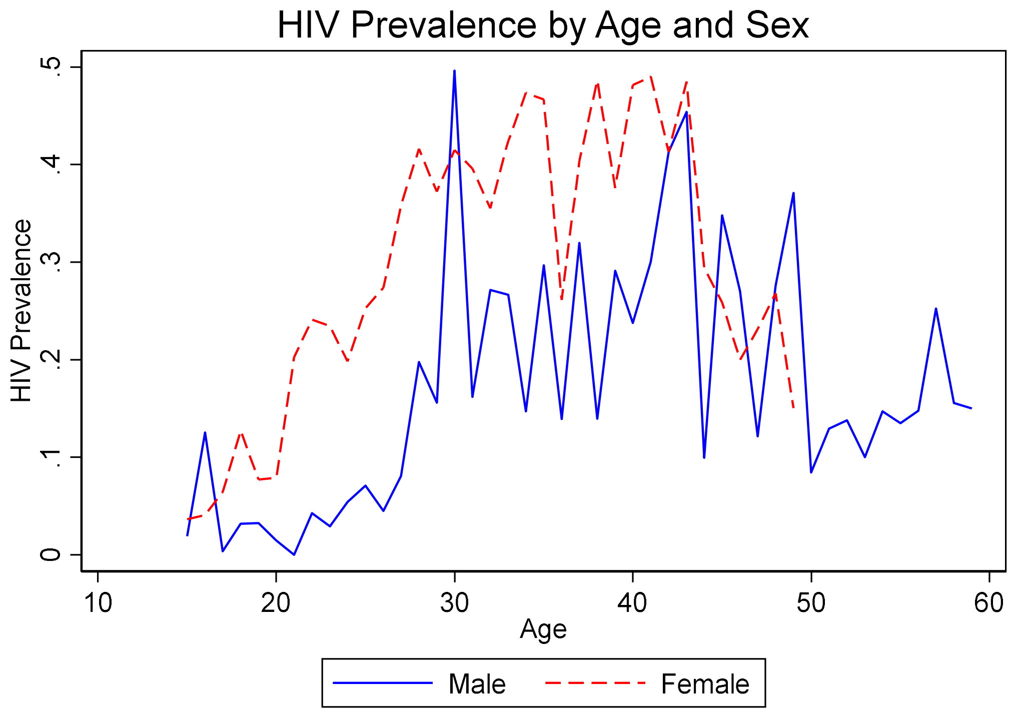


Figure 3.5 HIV prevalence by age and sex

Among those who tested HIV positive, 86.5% were older adults and 13.5% were young adults. Additionally, the proportion of HIV-positive women was nearly double that of HIV-positive men, with women making up 65.3% and men 34.7% of the total HIV-positive participants.

Participants with no formal education accounted for 3% of the cases, while those with primary education represented 15.6%. The highest prevalence was observed among individuals with secondary education, who accounted for 75.8% of all cases. In contrast, participants with higher education contributed 5.6% of the total cases.

For the variable current marital status, the largest proportion of all the positive cases was amongst those who were never in union with a rate of 52.1%. These were followed by participants living with a partner at 20.1% and the rate amongst married couples was 18.5%.

The rate of HIV positivity was 5.5% among participants who were separated or no longer living together, and 2.8% among those who were widowed. The lowest rate was observed among divorced individuals, at just 1%.

3.4 Characteristics and distribution of participants in male and female datasets

Figure below shows that HIV prevalence was high in the old age category for both males and females. Females had higher prevalence in category 2 when compared to males.

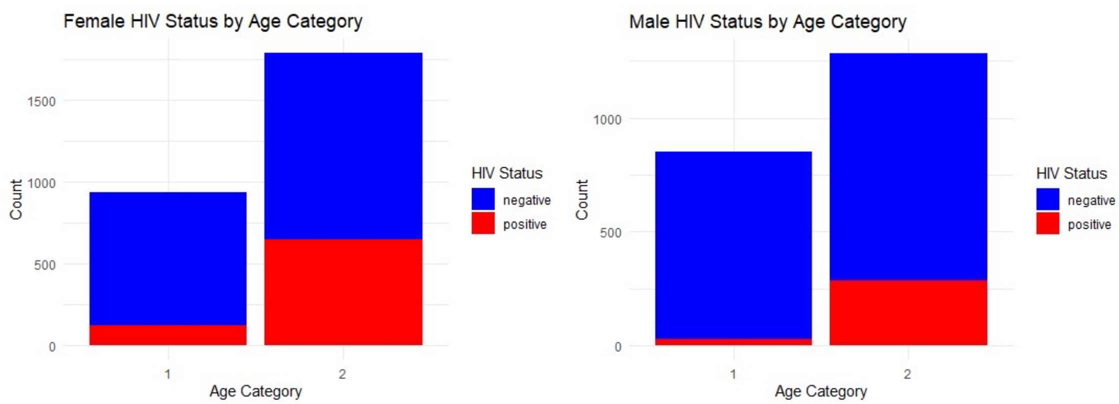


Figure 3.6 HIV status by age category Left panel: females; right panel: males.

The figure below presents box plots illustrating the distribution of HIV status by age. For females, the most affected are age group is between 25 and 40 years, while for males the most affected age range is from 32 to 47 years.

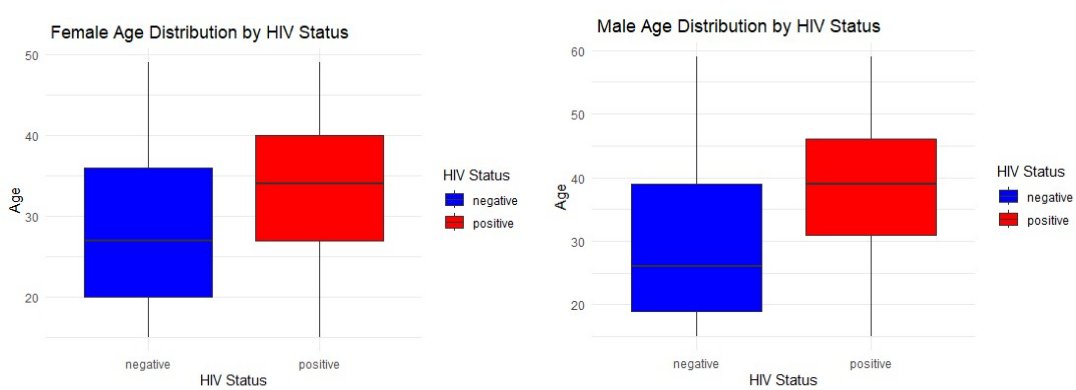


Figure 3.7 Age distribution by HIV status. Left panel: females; right panel: males.

3.5 Features selected by LASSO and GV selection

Potential features selected using LASSO regression for female and male datasets are shown in table 3.4 and those selected by GV select are shown in table 3.5 see the appendices.

3.6 Goodness of fit on variables selected by GV selection

The Hosmerleme show goodness of fit using the variables selected by generalized variance selection was 0.9604 for female dataset which is a very good fit. The accuracy of the model was 73.93%. On the male dataset the goodness of fit was 0.2469 which is also a good fit. The accuracy of the model was 84.18%. The selected predictors explained the outcome variable well.

3.7 Checking for confounding on variables selected using GV selection

Figures below show the confounding effects of variables selected using GV selection on female and male datasets. The graph below showed that age category was a confounder with a 12% confounding effect when condom use during last sex was used as the exposure on female data

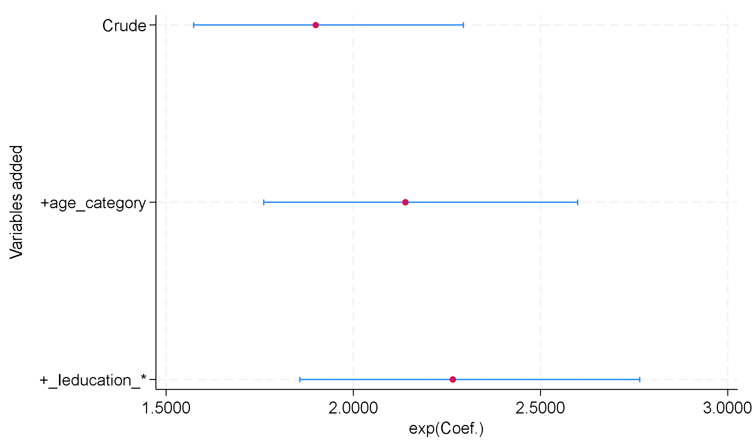


Figure 3.8 Confounding effect of age category on female dataset

The graph below showed that condom use last sex was a confounder with a 17.7% confounding effect when age category was used as the exposure on the male dataset.

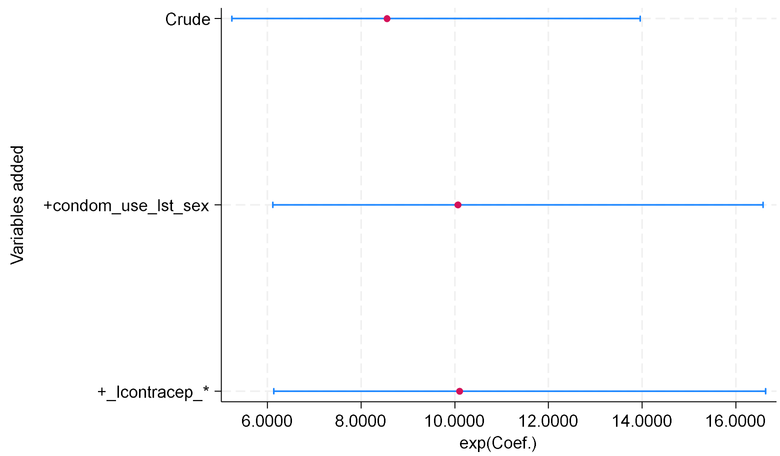


Figure 3.9 Confounding effect of condom use last sex on male data

3.8 Multivariable analysis of sex stratified data

Multivariable analysis was done on both male and female datasets to understand variable relationships. Common risk factors for both male and female datasets were age category, condom use last sex, household socio economic status and cohabitation duration.

The tables below illustrate results on multivariable analysis of female and male datasets.

Table 3.6 Multivariable analysis of factors associated with HIV infection among adult females

HIV-status	AOR	SE	z	p	95% CI
Household members	0.95	(0.03)	-1.85	0.065	0.90,1.00
Region					
eastern cape	1.05	(0.39)	0.12	0.903	0.50,2.19
northern cape	0.59	(0.23)	-1.35	0.178	0.27,1.27
free state	0.81	(0.30)	-0.57	0.566	0.39,1.68
kwazulu-natal	1.62	(0.61)	1.27	0.204	0.77,3.41
north west	0.69	(0.25)	-1.02	0.306	0.33,1.41
gauteng	0.88	(0.32)	-0.34	0.731	0.43,1.80
mpumalanga	1.37	(0.60)	0.72	0.473	0.58,3.25
limpopo	0.39	(0.15)	-2.46	0.014	0.19,0.83

Residence type					
rural	0.84	(0.17)	-0.87	0.386	0.56,1.25
Household socio economic status					
poorer	0.95	(0.22)	-0.21	0.835	0.61,1.49
middle	0.85	(0.22)	-0.61	0.539	0.51,1.42
richer	0.83	(0.25)	-0.62	0.535	0.46,1.50
richest	0.62	(0.22)	-1.34	0.181	0.31,1.25
Respondent's current age	1.04	(0.02)	2.52	0.012	1.01,1.08
Education level					
primary	1.19	(0.59)	0.36	0.721	0.45,3.15
secondary	1.09	(0.53)	0.18	0.861	0.42,2.83
higher	0.48	(0.28)	-1.24	0.215	0.15,1.53
Ethnicity					
white	1.00	(.)	.	.	1.00,1.00
coloured	0.14	(0.07)	-4.10	0.000	0.05,0.35
indian/asian	0.32	(0.35)	-1.03	0.304	0.04,2.84
relationship to household head	0.96	(0.03)	-1.45	0.147	0.91,1.01
Marital status					
married	0.10	(0.07)	-3.44	0.001	0.03,0.37
living with partner	0.16	(0.11)	-2.60	0.010	0.04,0.64
widowed	0.09	(0.07)	-3.08	0.002	0.02,0.42
divorced	0.10	(0.10)	-2.34	0.020	0.01,0.69
no longer living together/separated	0.25	(0.18)	-1.94	0.053	0.06,1.02
Cohabitation duration					
0-4	12.62	(8.79)	3.64	0.000	3.21,49.53
5-9	8.25	(5.64)	3.09	0.002	2.16,31.56
10-14	5.79	(3.97)	2.56	0.011	1.50,22.29
15-19	4.99	(3.44)	2.33	0.020	1.28,19.36
20-24	4.47	(3.14)	2.14	0.033	1.13,17.73
25-29	1.93	(1.38)	0.91	0.361	0.47,7.90
30+	1.00	(.)	.	.	1.00,1.00
Age at first sex	0.93	(0.03)	-2.12	0.034	0.88,1.00
Beating justified					
yes	1.13	(0.67)	0.21	0.838	0.35,3.60
don't know	1.73	(0.78)	1.21	0.227	0.71,4.19
Condom use last sex					
yes	2.11	(0.34)	4.62	0.000	1.54,2.89
Sti					
yes	2.21	(0.60)	2.91	0.004	1.29,3.76
Relationship with	1.01	(0.01)	1.10	0.270	0.99,1.03

most recent sex partner					
Age category					
Old Adults (25+)	2.49	(0.72)	3.19	0.002	1.42,4.38

Table 3.6 above shows results from female multivariable analysis. The results of the multivariable logistic regression analysis are presented as adjusted odds ratios (AORs), with all other variables in the model held constant. Ethnicity was significantly associated with HIV status, with coloured females showing an 84% reduction in the odds of HIV infection compared to Black African females (AOR: 0.16; 95% CI: 0.05–0.35). Individuals reporting a past STI have more than twice the odds of being HIV positive relative to those without such a history (AOR: 2.21; 95% CI: 1.52–2.89). Adults aged 25 years and older had 149% higher odds of HIV infection compared to younger adults (AOR: 2.49; 95% CI: 1.42–4.38). Interestingly for condom use during the last sexual encounter those who used a condom during last sex encounter had a 111% increase in the odds of HIV infection compared to those who did not (AOR: 2.11; 95% CI: 1.54–2.89). Additionally, each one-year increase in respondent age was associated with a 4% increase in the odds of HIV infection (AOR: 1.04; 95% CI: 1.01–1.08).

Cohabitation duration was also significantly associated with HIV status, using 30+ as the reference category. Compared to this group, women who had cohabited for 0–4 years had over 12 times higher odds of being HIV positive (AOR: 12.62; 95% CI: 3.21–49.53), while those with 5–9 years of cohabitation had over 8 times the odds (AOR: 8.25; 95% CI: 2.16–31.56). Participants with 10–14 years of cohabitation had nearly 6 times the odds (AOR: 5.79; 95% CI: 1.50–22.29), and those with 15–19 years and 20–24 years had nearly 5- and 4.5-times higher odds, respectively (AOR: 4.99; 95% CI: 1.28–19.36 and AOR: 4.47; 95% CI: 1.13–17.73). The association for those who had cohabited for 25–29 years was not statistically significant (AOR: 1.93; 95% CI: 0.47–7.90). The 30+ years cohabitation group was used as the internal reference category within the cohabitation duration variable and did not differ significantly from women who had never been in a union.

Table 3.7 Multivariable analysis of factors associated with HIV infection among adult males

HIV status assay 1 result	AOR	SE	z	p	95% CI
Age category					
Old Adults (25+)	6.72	(2.34)	5.46	0.000	3.39,13.33
Condom use last sex					
yes	2.76	(1.10)	2.54	0.011	1.26,6.05
Marital status					
married	0.40	(0.22)	-1.64	0.102	0.13,1.20
living with partner	0.85	(0.55)	-0.25	0.805	0.24,3.00
widowed	1.46	(1.13)	0.49	0.627	0.32,6.66
divorced	0.13	(0.11)	-2.44	0.015	0.02,0.67
no longer living together/separated	0.87	(0.56)	-0.21	0.831	0.25,3.08
Household socio economic status					
poorer	0.84	(0.24)	-0.61	0.542	0.47,1.48
middle	0.81	(0.23)	-0.73	0.465	0.46,1.43
richer	0.48	(0.16)	-2.15	0.032	0.25,0.94
richest	0.49	(0.21)	-1.66	0.098	0.21,1.14
Circumcised					
yes	0.48	(0.10)	-3.62	0.000	0.32,0.71
don't know	2.23	(1.90)	0.94	0.347	0.42,11.88
HIV testing					
yes	2.11	(0.62)	2.55	0.011	1.19,3.76
Cohabitation duration					
0-4	3.51	(2.17)	2.04	0.042	1.05,11.80
5-9	2.29	(1.26)	1.50	0.133	0.78,6.75
10-14	4.05	(2.57)	2.20	0.028	1.16,14.08
15-19	1.34	(0.87)	0.45	0.651	0.38,4.80
20-24	1.03	(0.70)	0.05	0.960	0.27,3.92
25-29	2.10	(1.42)	1.10	0.270	0.56,7.91
30+	1.00	(.)	.	.	1.00,1.00
Contraceptive use					
using contraceptive	0.62	(0.25)	-1.17	0.244	0.28,1.39
Respondent's current age	1.03	(0.01)	1.83	0.068	1.00,1.05
Number of trips in last 12 months	0.98	(0.01)	-1.49	0.136	0.96,1.01
Age at first sex	0.99	(0.03)	-0.33	0.742	0.93,1.05

Table 3.7 above shows results from male multivariable analysis. The results of the multivariable logistic regression analysis are presented as adjusted odds ratios (AORs), with all other variables in the model held constant. Older male adults (25+) have 6.72 higher odds of being HIV-positive than young male adults (15–24) (AOR: 6.72; 95% CI: 3.39–13.33). Interestingly for condom use during the last sexual encounter those who used a condom during last sex encounter had 2.76 times odds of HIV infection compared to those who did not (AOR: 2.76; 95% CI: 1.26–6.05). Circumcision is strongly protective it reduces odds of HIV by about 52% (AOR: 0.48; 95% CI: 0.32–0.71). Richer individuals have about 52% lower odds of being HIV-positive than the poorest group (AOR: 0.48; 95% CI: 0.25–0.94).

Individuals cohabiting for 0–4 years have 3.5 times the odds of being HIV-positive compared to those cohabiting for 30+ years (AOR: 3.51; 95% CI: 1.05–11.80), while Individuals cohabiting for 10–14 years have 4.1 times odds of being HIV-positive compared to those cohabiting for 30+ years (AOR: 4.05; 95% CI: 1.16–14.08).

The findings underscore the complexity of factors influencing HIV risk and suggest areas for further research and intervention strategies to mitigate infection rates effectively.

3.9 Interaction terms from multivariable analysis

Interaction terms that were significant on the female dataset were between age category and ethnicity as well as age category and region. The results showed that the odds of contracting HIV for adult-coloured females are lowered by 91% compared to black young adult females (AOR: 0.09 95%CI: 0.019-0.462). The effect of age on HIV risk varies substantially across provinces. Compared to young adult females in the Western Cape the odds of HIV in adults in the following provinces were high Free State (AOR: 4.4 95%CI 1.02-18.6), North West (AOR 5.06 95%CI: 1.11-23.17), Gauteng (AOR: 6.12 95%CI:1.24-30.12) and Limpopo (AOR:5.56 95%CI:1.13-27.30). Refer to Table 3.8 in appendices.

On the male dataset one interaction term between condom use last sex and age category was significant. The results showed that older adults who used condoms during last sex were 4.4

times likely to get infected compared to young adults not using condoms (AOR:4.32 95%CI:1.16-16.06). Refer to table 3.9 in appendices.

3.10 Imputation of missing data results

The table below shows the imputation process for variables with missing observations of less than 30% in preparation for model training. Splitting of data by clusters has been found to be very effective as there were more significant features on multivariate analysis compared to when the data was split according to individuals in the dataset. This ensured better comparability between clusters as individuals in a single cluster are regarded as dependent rather than independent.

Table 3.10 Complete, missing and imputed variables for female and male datasets

	Female data				Male data			
Variables Training	Complete	Missing	Imputed	Total	Complete	Missing	Imputed	Total
condom use last sex	1679	449	449	2128	1362	362	362	1724
recent partner relationship	1673	455	455	2128	-	-	-	-
Testing	Complete	Missing	Imputed	Total	Complete	Missing	Imputed	Total
condom use last sex	446	152	152	598	360	80	80	412
Recent partner relationship	446	152	152	598	-	-	-	-

3.11 Preprocessing and ranking of risk factors on training data

The risk factors were ranked using coefficients from the multivariate logistic regression of significant predictors on preprocessed training data. Risk score was calculated by dividing the coefficients by the lowest coefficient and rounding it off to the nearest integer. Tables below illustrate the ranking of risk factors associated with HIV infection.

Table 3.12 Ranking risk factors using absolute coefficients, risk scores using significant features from multivariable analysis for females

Variable	Coefficient (β)	Risk Score	Risk rank number
age	1.84304169	25	1
hhmembers	-1.35108592	19	2
ethnicity	-1.27043609	18	3
recent_partner_rltnshp	-0.82479361	11	4
age_category	0.79314431	11	5
condom_use_lst_sex	0.72536171	10	6
sti	0.60214591	8	7
cohabitation_duration	-0.28414858	4	8
hhses	-0.20679550	3	9
marital_status	0.19438812	3	10
region	-0.07222505	1	11

*Smallest predictor: coefficient 0.07222505

*Risk score calculated by dividing absolute coefficient by 0.07222505 and round off to nearest integer.

Table 3.13 Ranking of risk factors using absolute coefficients, risk scores using significant features from multivariable analysis for males

Variable	Coefficient (β)	Risk Score	Risk rank number
age_category	1.94961045	25	1
circumcised	-0.81505817	11	2
HIV_testing	0.77595410	10	3
condom_use_lst_sex	0.55400952	7	4
marital_status	0.22644224	3	5
hhses	-0.22085829	3	6

cohabitation_duration	-0.07673679	1	7

*Smallest predictor: coefficient 0.07673679

*Risk score calculated by dividing absolute coefficient by 0.07673679 and round off to nearest integer.

The predictor with the highest score for females was age, while for males it was age category with both achieving a risk score of 25. For information on the risk score tools, please refer to tables 3.14 and 3.15 in the appendices. Females with a risk score of 34.5 or higher are referred for counselling and testing, while males with a risk score of 30 or higher are also referred for counselling and testing because they exhibit high risk behavior.

3.12 Predictive modelling

Results for comparing predictive performance of models were done using significant features from multi variate analysis in R on original weighted data and on balanced and weighted data. Analysis in orange was done using available predictors on the dataset due to the nature of some variables in the dataset when the dta file was exported as a csv file. For features used in R for predictive modelling refer to table 3.11 and for features used in orange refer to table 3.16 in the appendices.

Table below shows performance metrics female data

Table 3.17 Comparison of performance metrics on female original weighted data, SMOTE (balanced) and weighted data and on original unweighted data

	Original data Females Training Weighted				**Original data Females Testing Weighted**			
Metric	**XGB**	**Logistic regression**	**KNN**	**Ratio N:P**	**XGB**	**Logistic regression**	**KNN**	**Ratio N:P**
PR AUC	0.5762	0.5831	0.5518	2.66:1	0.5407	0.5471	0.5486	2.13:1
F1 score	0.8500	0.8413	0.8682		0.8203	0.8112	0.8128	
AUROC	0.7808	0.7641	0.8433		0.7340	0.7300	0.7251	
Recall	0.6273	0.5775	0.7174		0.6444	0.6004	0.6078	
Precision	0.7842	0.7800	0.8056		0.7424	0.7430	0.7399	

Accuracy	0.7622	0.7497	0.7923		0.7268	0.7171	0.7174	
	SMOTE (balanced) data Females Training weighted				**SMOTE (balanced) data Females Testing weighted**			
Metric	**XGB**	**Logistic regression**	**KNN**	**Ratio**	**XGB**	**Logistic regression**	**KNN**	**Ratio**
PR AUC	0.3217	0.3546	0.3081	1:1	0.3304	0.3682	0.3676	1:1
F1 score	0.8033	0.7116	0.9165		0.7725	0.6813	0.6791	
AUROC	0.8899	0.7768	0.9805		0.8516	0.7416	0.7313	
Recall	0.8108	0.7093	0.7403		0.7868	0.6798	0.6972	
Precision	0.7898	0.7246	0.8055		0.7431	0.7080	0.6771	
Accuracy	0.7999	0.7167	0.7961		0.7631	0.6929	0.6781	
	Original unweighted data Females Training				**Original unweighted data Females Testing**			
Metric	**XGB**	**Logistic**	**KNN**	**Ratio**	**XGB**	**Logistic**	**KNN**	**Ratio**
F1 score	0.7900	0.7360	0.7840	2.53:1	0.7210	0.7090	0.6550	2.53:1
AUROC	0.8520	0.7850	0.8450		0.7490	0.7390	0.6720	
Recall	0.8060	0.7570	0.7960		0.7450	0.7360	0.6770	
Precision	0.7980	0.7380	0.7850		0.7250	0.7130	0.6450	
Accuracy	0.8060	0.7570	0.7960		0.7450	0.7360	0.6770	

For original and weighted female data Table 3.17 showed that during training, all models performed well with the XGB showing the highest F1 score (0.8500) and AUROC (0.7808). KNN has the highest recall (0.7174) and precision (0.8056) suggesting good detection of positive cases. During testing performance drops slightly on the test set with the XGB maintaining the highest F1 score (0.8203) and AUROC (0.7340) whilst the KNN suffers a drop in accuracy (0.7174). Recall remains high for XGB and KNN but decreases slightly for logistic regression.

For balanced (smote) and weighted female data KNN achieved highest F1 score (0.9165) and AUROC (0.9805) during training which is an improvement in detecting positives although precision and recall are a bit balanced as shown in table 3.17 on the female dataset. The logistic regression has the highest PR score (0.3546) despite it being low. During testing XGB showed the highest F1 score (0.7725) and AUROC (0.8516) but performance dropped when compared to training set. KNN has the largest drop in performance with recall and accuracy decreasing significantly.

On original and unweighted female data Table 3.17 showed that all models performed well during training with the XGB having the highest F1 score (0.7900) and AUROC (0.8520). Recall and precision are balanced across all models, but the logistic regression consistently showed lower performance. During testing there is a drop in performance metrics across all models particularly for KNN. XGB maintains the highest F1 score (0.7210) and AUROC (0.7490)

Table below shows performance metrics male data

Table 3.18 Comparison of performance metrics on male original weighted data, SMOTE (balanced) and weighted data and on original unweighted data

	Original data Males Training Weighted				**Original data Males Testing Weighted**			
Metric	**XGB**	**Logistic regression**	**KNN**	**Ratio N:P**	**XGB**	**Logistic regression**	**KNN**	**Ratio N:P**
PR AUC	0.7388	0.7413	0.6947	5.84:1	0.7501	0.7349	0.7344	5.44:1
F1 score	0.9258	0.9230	0.9279		0.9180	0.9141	0.9131	
AUROC	0.7723	0.7760	0.9054		0.7452	0.7672	0.7765	
Recall	0.6551	0.6727	0.8201		0.7500	0.3333	0.4000	
Precision	0.8732	0.8606	0.8689		0.8505	0.8460	0.8477	
Accuracy	0.8652	0.8582	0.8679		0.8495	0.8422	0.8422	
	SMOTE (balanced) data Males Training Weighted				**SMOTE (balanced) data Males Testing Weighted**			
Metric	**XGB**	**Logistic regression**	**KNN**	**Ratio**	**XGB**	**Logistic regression**	**KNN**	**Ratio**
PR AUC	0.3143	0.3511	0.3072	1:1	0.3183	0.3539	0.3466	1:1
F1 score	0.8641	0.6970	0.9328		0.8283	0.7015	0.7114	
AUROC	0.9376	0.7857	0.9932		0.9138	0.7825	0.7891	
Recall	0.8877	0.6907	0.8333		0.8704	0.6966	0.7204	
Precision	0.8343	0.7631	0.8685		0.7748	0.7414	0.6732	
Accuracy	0.8590	0.7211	0.8677		0.8155	0.7167	0.6940	
	Original unweighted data Males Training				**Original unweighted data Males Testing**			
Metric	**XGB**	**Logistic**	**KNN**	**Ratio**	**XGB**	**Logistic**	**KNN**	**Ratio**
F1 score	0.8470	0.8160	0.8570	5.75:1	0.7810	0.8030	0.7760	5.78:1
ROC AUC	0.8820	0.8210	0.8810		0.7660	0.7480	0.6780	
Recall	0.8810	0.8590	0.8770		0.8360	0.8430	0.8150	
Precision	0.8820	0.8200	0.8580		0.8010	0.8160	0.7620	
Accuracy	0.8810	0.8590	0.8770		0.8360	0.8430	0.8150	

For original and weighted male data Table 3.18 showed that during training, all models performed well with highest Precision recall area under the curve (PR AUC) on the logistic model (0.7413) and the KNN showed the highest F1 score (0.9279) and (AUROC) of (0.9054). KNN has the highest metrics apart from PR AUC and precision. During testing performance drops slightly on the test set with the XGB having the highest F1 score (0.9180) and PR AUC (0.7501) whilst the logistic regression and KNN suffers a drop in recall (0.3333 and 0.4000) respectively.

On balanced data Table 3.18 showed that all models had low PR AUC during training, highest being 0.3511 for logistic regression. KNN had the highest F1 score (0.9328) and AUROC (0.9932). Logistic regression struggled with balanced data on most metrics apart from PR AUC. During testing the XGB had the highest F1 score (0.8283) and AUROC (0.9138)

The optimal tuning hyperparameters for the weighted and balanced datasets, for both males and females, are presented in Appendix Table 3.19

Table 3.18 showed that all models performed well during training on original and unweighted data with the KNN having the highest F1 score (0.8570) and the XGB having the highest AUROC (0.8820). During testing there is a drop in performance metrics across all models and surprisingly the logistic regression has the highest F1 score (0.8030) whilst the XGB had the highest AUROC (0.7660). KNN had the lowest performance metrics.

Figure below top left shows female training ROC performance and top right female test ROC performance on original weighted data, bottom left male training ROC and bottom right male test ROC for the 3 models XGB, Logistic and KNN. KNN has highest AUROC (0.8433) during training and XGB has highest AUROC during testing for female dataset (0.7340). For the male dataset KNN has the highest AUROC during training (0.9054) and (0.7765) during testing.

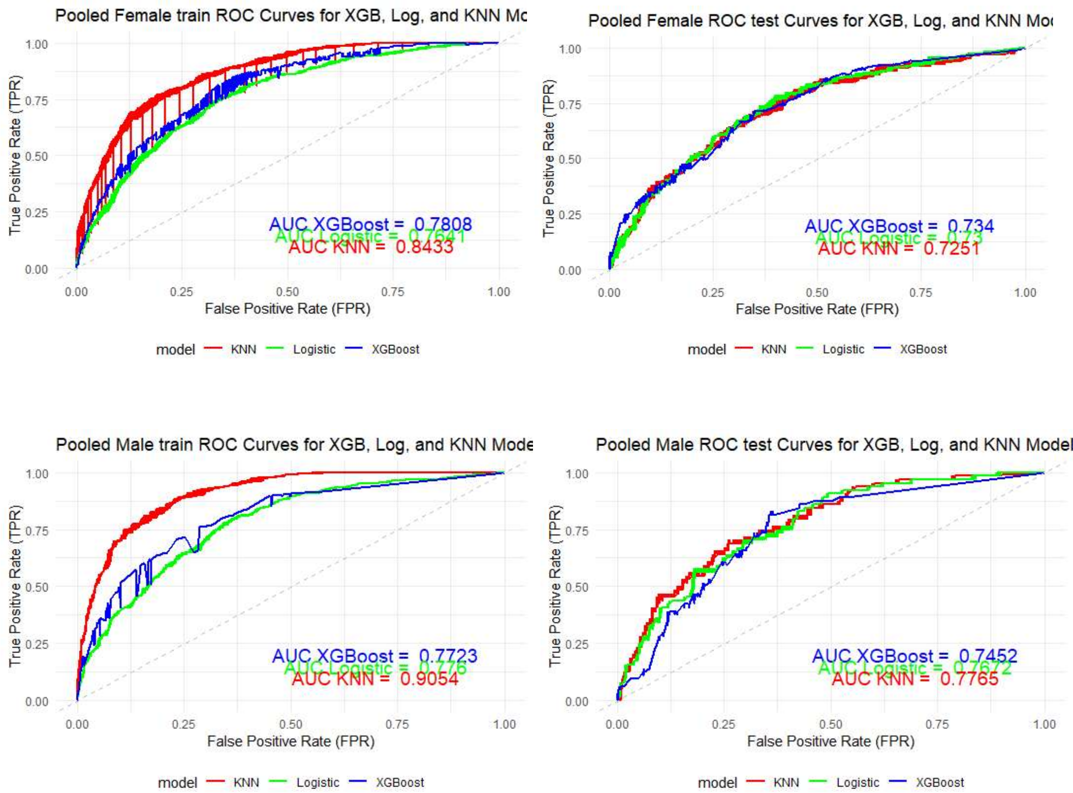


Figure 3.10 ROC curves for females and males training and testing on original weighted data

Figure below top left shows female training PR performance and top right female test PR performance on original weighted data, bottom left male training PR and bottom right male test PR for the 3 models XGB, Logistic and KNN. Logistic regression has highest PR AUC (0.5831) during training and KNN has highest PR AUC during testing for female dataset (0.5486). For the male dataset Logistic regression has the highest PR AUC during training (0.7413) and the XGB (0.7501) during testing.

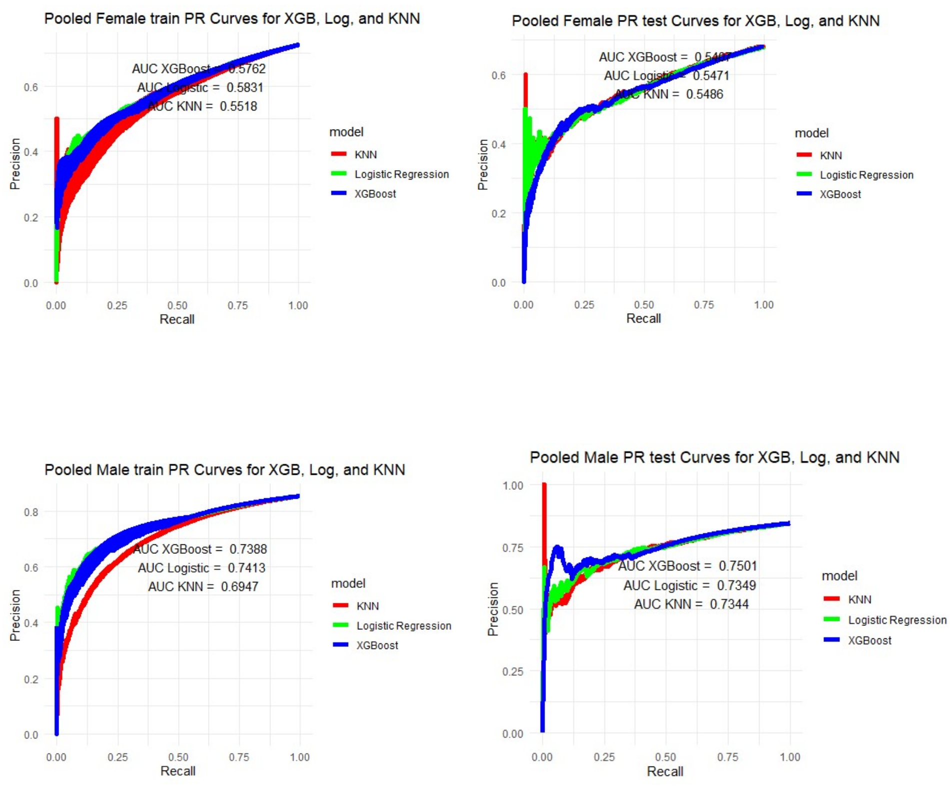
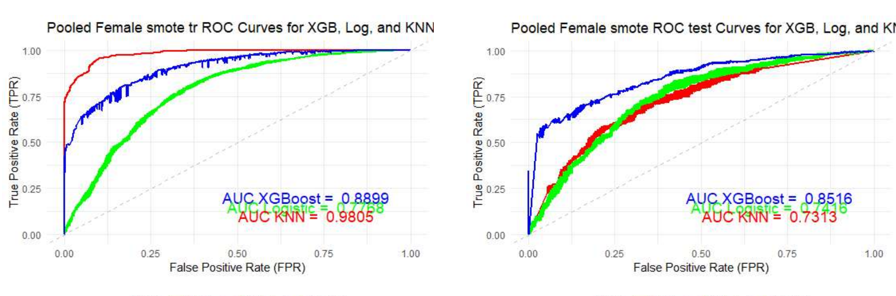


Figure 3.11 PR curves for females and males training and testing on original weighted data

Figure below top left shows female training ROC performance and top right female test ROC performance on balanced (smote) and weighted data, bottom left male training ROC and bottom right male test ROC. KNN has highest AUROC (0.9805) during training and XGB has highest AUROC during testing for female dataset (0.8516). For the male dataset KNN has the highest AUROC during training (0.9932) and XGB during testing (0.9138).



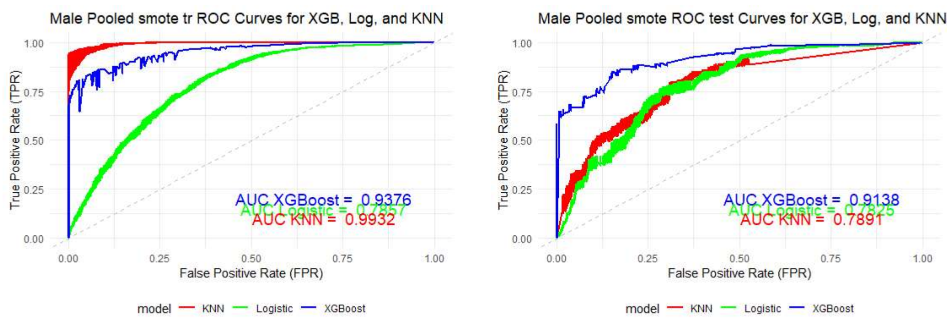


Figure 3.12 ROC curves for females and males on balanced and weighted data

Figure below top left shows female training PR performance and top right female test PR performance on balanced and weighted data, bottom left male training PR and bottom right male test PR. Logistic regression has highest PR AUC (0.3546) during training and testing for female dataset (0.3682). For the male dataset the logistic regression has the highest PR AUC during training (0.3511) and testing (0.3539) as well.

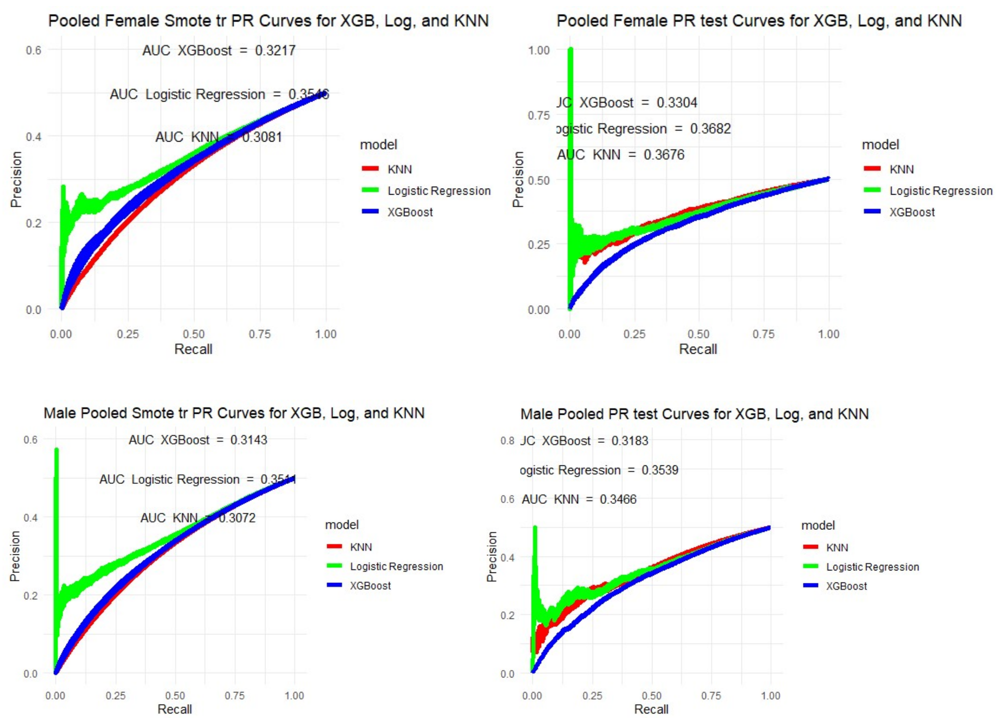


Figure 3.13 PR curves for females and males on balanced and weighted data

Performance metrics for AUROC in orange, KNN is grey, XGB is brown and logistic regression is green. Figure below shows top left female ROC curves during training and top right during testing in orange for original and unweighted datasets and bottom left male ROC curves training curve and right ROC test curve. For female dataset XGB had the highest AUROC during training (0.8520) and testing (0.7490). For the male dataset XGB had the highest AUROC during training (0.8820) and testing (0.7660)

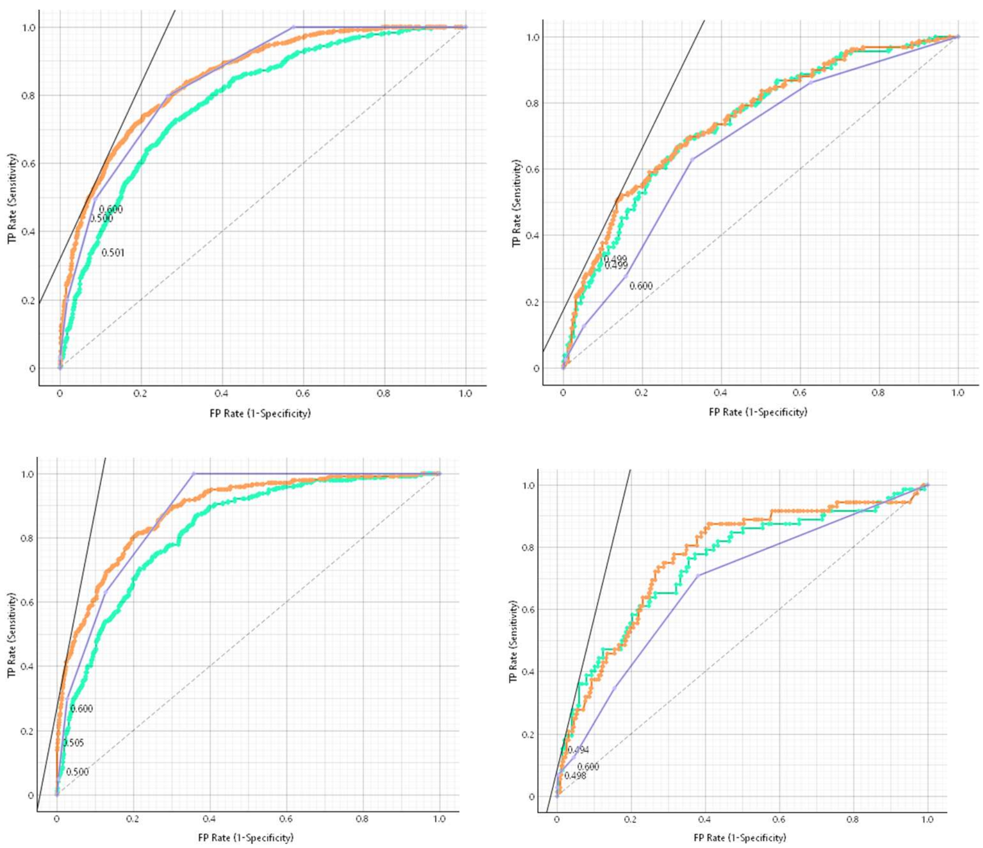


Figure 3.14 ROC on original and unweighted data for females and males

3.13 Feature importance for the models

The figures below demonstrate that various algorithms prioritize different features in terms of their predictive accuracy for both female and male datasets.

3.13.1 Feature importance for XGB model

The graphs below illustrate important features for the XGB left for females and right for males. Age is the most important factor for females and age category for males.

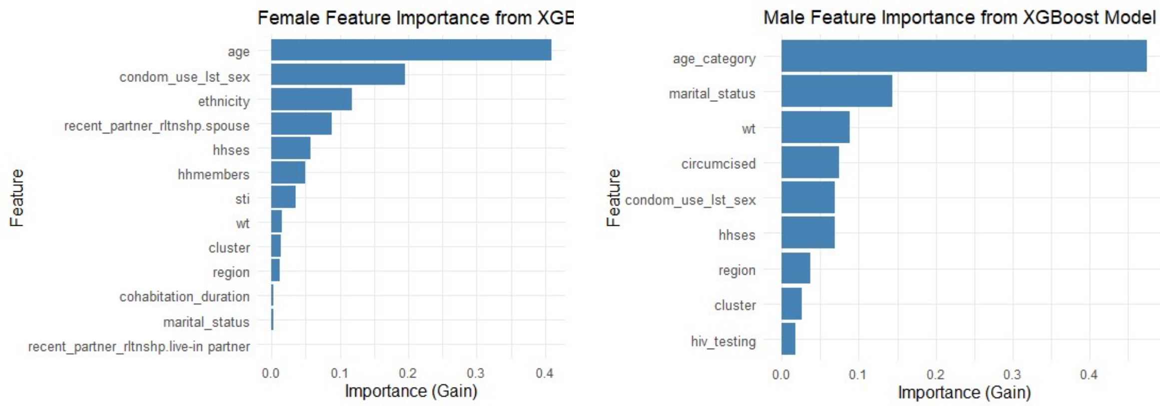


Figure 3.15 Feature importance for XGB model

3.13.2 Feature importance for logistic regression model

Age is the most important feature for females left graph and age category for males' right graph.

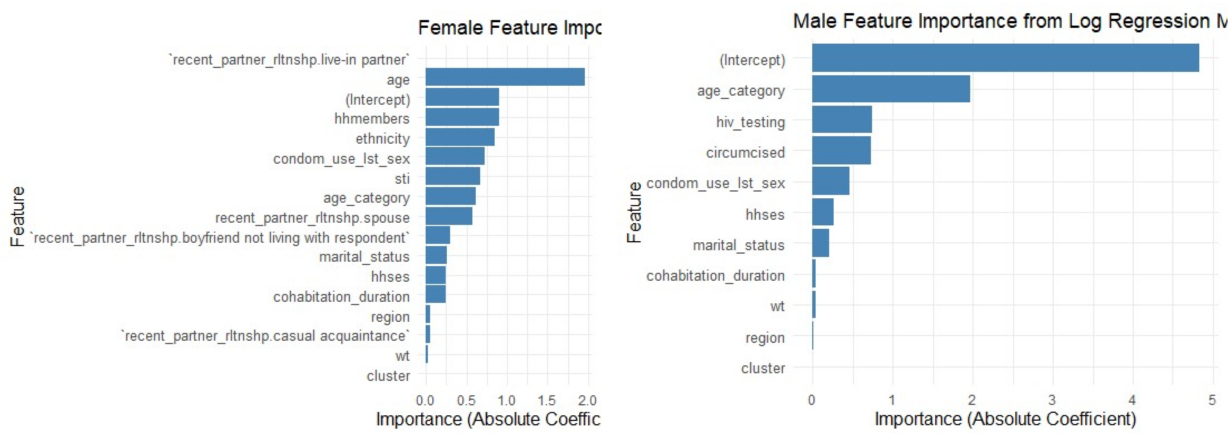


Figure 3.16 Feature importance for the logistic regression model

3.13.3 Feature importance for KNN model

The graphs below illustrate important features for the KNN model left for females and right for males during prediction. Condom use last sex is the most important predictor for females and circumcised for males.

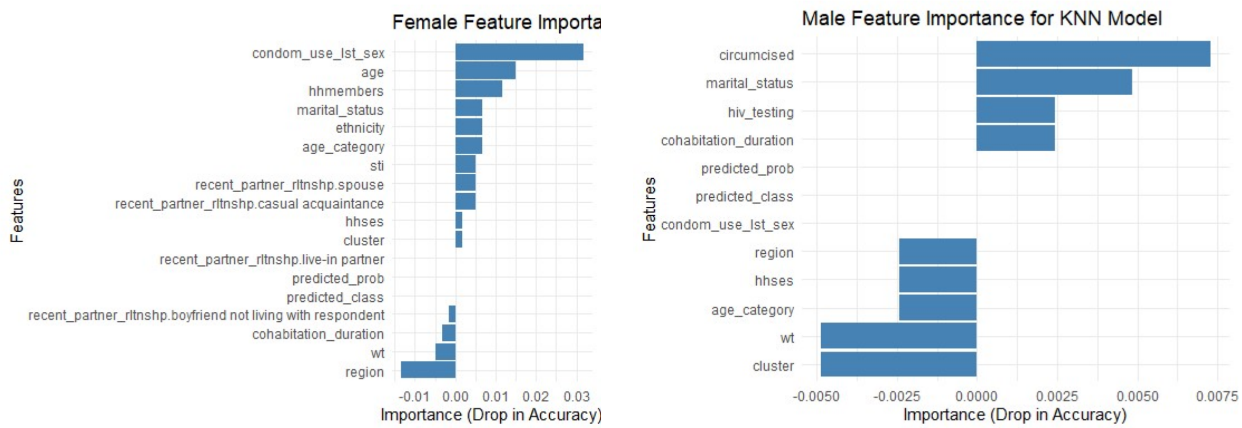


Figure 3.17 Feature importance for the KNN model

The region with highest HIV prevalence for both sexes was KZN. Adolescent girls exhibited a higher prevalence of HIV infection than their male counterparts. Common risk factors for both male and female datasets were age category, condom use last sex, household socio economic status and cohabitation duration. XGB was the model with the best predictive performance followed by logistic regression.

4. DISCUSSION AND CONCLUSION

In this study, we examined the characteristics and distribution of HIV infected adults who are 15 years old and above in South Africa in 2016, identified key risk factors associated with HIV infection and selected a predictive model that can identify individuals at risk of infection. Findings from this study are discussed according to the specific objectives.

4.1 Characteristics and distribution of participants

The result from this study shows that most participants were black, followed by coloured, white and lastly Indian and Asians. HIV prevalence was highest amongst the blacks, followed by coloured, Asians and Indians and white.

The weighted gender of participants was almost similar with 50.5% males and 49.5% females. From the study HIV prevalence was twice as high amongst females than males. This is the same findings by (1) (4). Women have greater biological susceptibility especially younger women during heterosexual intercourse with the vaginal mucosa providing a larger surface area for HIV to enter the blood stream when no protection is used.

The number of participants who were classified as old adults 25+ years were almost twice the number of participants classified as young adults 15-24 years old. HIV prevalence amongst old adults was close to 3.5 times that of young adults. This is also in agreement with (4) (1) (38). This is attributed to older individuals having more years of potential exposure to HIV through risky behaviors such as unprotected sex or sharing needles thereby increasing likelihood of infection over time.

Marital status and relationship dynamics also played a role. More than half of the participants were never in union, followed by those who were married. Participants who were living with a partner were half of those who were married. Participants who were widowed had the highest HIV prevalence and were closely followed by those who were living with a partner. Married participants had the lowest HIV prevalence. The findings may be reflective of past partner loss due to HIV, or the increased risk associated with new or multiple partnerships following marital dissolution.

Study showed KZN as the province with the highest combined HIV prevalence. Mpumalanga province was second followed by Free State, Northwest, Gauteng, Eastern Cape, Western cape, Northern Cape and Limpopo in that order. High prevalence in KZN could be attributed to frequent movement of people for work in trucking, port and industrial sectors which increases the risk of HIV transmission and consistent condom use being low in some communities.

HIV prevalence was lowest amongst participants with higher education followed by secondary education and was almost similar among participants with primary and no education at all. Findings are in agreement with (39) (40). Individuals with higher education are more likely to have better access to comprehensive and accurate information about HIV transmission and prevention. More educated individuals are more likely to seek out and use preventive healthcare services. Higher education is associated with delayed sexual debut and lower risk behaviors.

Teen girls were having higher infection rates than teen boys. This is supported by studies from (14) (21). This could be attributed to early sexual debuts for teen girls dating older men with high chances of having contracted HIV when compared to teen boys who have late sexual debuts.

On the female dataset black females from KZN were having the highest HIV prevalence and on the male dataset males from KZN had highest prevalence which makes KZN a region that needs great attention.

4.2 Common risk factors associated with HIV infection

Common risk factors for both male and female datasets were age category, condom use last sex, household socio economic status and cohabitation duration.

4.2.1 Key risk factors on female dataset

From the study the risk factors that were protective in females were cohabitation duration and age at first sex. Household members were protective though marginally significant. Factors with higher odds of HIV infection were age category, condom use last sex, sti, and age. Significant interaction terms were between ethnicity and age category as well as region and age category.

Our study found that longer cohabitation duration is associated with lower odds of infection, which contrast with the findings of (4). The discrepancy may stem from the fact that in South Africa, unions are more often cohabitation based rather than marriage based. The longer individuals cohabit, the more stable their relationship tends to be, fostering mutual trust. Stability is likely to reduce the risk of exposure to HIV through external sexual partners.

The study also found that late sexual debut is associated with lower odds of infection. Fewer years of sexual activity means fewer opportunities for potential exposure. Starting sex later in life results in fewer lifetime sexual partners. Finding is supported by (14).

Our study also found that having more household members reduces the risks of contracting HIV though it was marginally significant. This could be attributed to increased social control or oversight whereby larger households may offer great supervision especially for young woman or adolescents leading to limited risky sexual behavior. Individuals in larger households may have less privacy which can also reduce the likelihood of engaging in sexual activities at home. Household responsibilities may also limit time for social interaction that leads to sexual partnerships.

We also found out that HIV prevalence is higher in the older age category than in the young age category. This is also in agreement with (4) (1) (38). This is attributed to older individuals having more years of potential exposure to HIV through risky behaviors such as unprotected sex or sharing needles thereby increasing likelihood of infection over time.

Surprisingly the study found out that the odds of HIV infection are higher for females who used condoms during last sex compared to those who did not use condoms. These findings are supported by studies from (41) (42) (43). The variable represents condom use during last sex, but it does not capture consistent or correct use over time. Condoms are highly protective when used properly and consistently. However, individuals who already know they are HIV-positive or who perceive themselves to be at high risk may adopt condom use to protect their partners which is an example of behavioral adaptation. In cross-sectional data, HIV infection precedes the reported condom use, making condom use a consequence of HIV status rather than a cause. The finding may therefore reflect behavior change following diagnosis, counseling, or perceived risk (e.g., multiple partners or prior STIs), rather than a protective effect of condom use itself (41) (42) (43).

Sexually transmitted infections (sti) have been found to be associated with an increase in the chances of contracting HIV. The finding is in agreement with (44). Many sti have been found to cause inflammation or ulcers which makes it easier for HIV to enter the body during sexual contact. HIV and sti are transmitted through unprotected sex such that people engaging in high-risk sexual behaviors are at higher risk of both (44).

We also found out that black woman had higher chances of contracting HIV when compared to woman of other races especially black women residing in KZN. Interaction showed that the odds of HIV infection black old adults are more than twice that of black young adults. Coloured old adults have lower odds of infection compared to young black adults. Interaction between region and age category showed four regions (Free State, North West, Gauteng and Limpopo) with older adults having higher odds of infection compared to young adults from the Western Cape. The effect of age on HIV risk varies substantially across provinces.

4.2.2 Key risk factors on male dataset

Protective risk factors in males were circumcision and household socioeconomic status. Risk factors with higher odds of infection were age category, condom use last sex and HIV testing.

Our study findings on circumcision is in agreement with other studies on the reduction of infection amongst men who are circumcised compared to those who are not circumcised (38). Removal of the foreskin has been shown to reduce the surface area which contains Langerhans cells where HIV virus can attach to (38) (45) (46).

The study also found lower odds of HIV infection for the rich when compared to the poorest. The findings are related to several aspects such as better access to education including health and sexual education. This leads to better understanding of HIV prevention methods. Wealthier households have better access to health care services including HIV testing, counselling and treatment. Lower SES is often associated with circumstances that increase vulnerability to HIV such as engaging in transactional or survival sex due to economic necessity. The findings are in agreement with (18) and different from (4).

On marital status the odds of being infected were higher in males who were never in union compared to other categories. Males not in stable union tend to have more casual or multiple sexual partners thereby increasing the chances of contracting HIV. Casual encounters are associated with lower condom use. Lack of partner communication amongst the man outside a stable relationship might reduce discussions about sexual history, HIV testing or condom use negotiation.

Males who said yes to HIV testing had higher odds of infection compared to those who said no. This could be those who are already at higher risk are more likely to get tested, which can be a form of self-selection bias.

Interaction term that was significant on the male was between age category and condom use last sex. Older adults who used condoms at last sex were having higher odds of infection. The same explanation in females applies to males.

The risk assessment tools created during analysis adds to the knowledge on risk assessment literature in predicting HIV acquisition in the South African population.

4.3 Comparative analysis of predictive models for HIV detection

The three models namely XGB, logistic regression and K Nearest Neighbors were compared on their predictive performance on male and female datasets (weighted, balanced and unweighted).

4.3.1 Predictions on female dataset

Our study also found out that the model with best predictive power on the female dataset was the XGB. It had the highest F1 score on original weighted data, balanced weighted data and unweighted data. This is in agreement with studies on weighted data by (7) (4). This could be attributed to algorithmic efficiency and regularization. By building an ensemble of decision trees sequentially with each trying to correct the errors of previous trees, it allows the algorithm to minimize bias and variance leading to great performance. Regularization prevents the model from overfitting.

4.3.2 Predictions on male dataset

On the male dataset our study found out that XGB had the highest F1 score on original weighted data, balanced weighted data during testing which is in agreement with studies on weighted data by (7) (4). Surprisingly the logistic regression had the highest F1 score during testing in unbalanced and unweighted male data in Orange. The observation could be attributed to the fact that the class distribution and sampling biases may not be accounted for which can lead to overfitting or underfitting based on complexity of class imbalance. Logistic regression could have performed better because it is a simpler model and less prone to overfitting on unbalanced and unweighted data. Further studies on this could provide a better picture.

The concept of predictive modelling was adopted from (4) whilst the risk factor ranking and development of a risk assessment tool was developed from (1)

4.4 Conclusion

HIV prevalence was twice as much in females compared to males. Of note is highest prevalence among black women and men, residing in the KwaZulu-Natal (KZN) region. Teen girls also have a higher rate of HIV infection than teen boys. Additionally, women and men who used a condom during their last sexual encounter were more likely to be infected. Circumcision was protective amongst men.

XGB was the model with the best predictive performance. Public health efforts should prioritize targeted HIV prevention and treatment strategies for high-burden groups, including Black women and men in KwaZulu-Natal and adolescent girls. Expanding access to youth-centered sexual health education and scaling up male circumcision programs could significantly reduce HIV transmission. Condoms remain a highly effective method to prevent HIV when used consistently and correctly. Public health programs must emphasize proper and regular condom use—not just during the last sexual act. XGB highlights the potential of machine learning tools to support targeted HIV interventions by early identification of high-risk people.

5. STRENGTHS AND LIMITATIONS

The following section will discuss the strengths first followed by the limitations of the study.

5.1 High outcome validity and reliability

Outcome was not self-reported. Laboratory testing provides standardised high precision measurement of outcomes thereby reducing misclassification bias. Retesting of inconclusive samples resulted in improved diagnostic accuracy, enhancing internal validity. All samples were tested at a single laboratory thereby minimising interlaboratory differences.

5.2 Stratification by sex

Stratification improves interpretation of the model by ensuring that findings are not confounded be sex differences. Enables targeted interventions by showing sex specific health disparities and social determinants.

5.3 Generalisability of results

The results from the study are generalizable to the rest of the country. Survey weights and design corrections reduce sampling bias and standard errors are adjusted for clustering and stratification. Weighted sampling corrects for non -response and enhance external validity

The limitations of the studies are as follows

5.4 Trends over time not determined

Cross sectional studies cannot determine trends as populations change over time so findings may not reflect longer term trends or variation. The study design is unable to assess high risk sexual behaviours across a lifetime, offering only short-term snapshots.

5.5 Limitations in causal inference

Cross sectional studies cannot establish causal relationship between variables. In a cross-sectional study, both exposures and outcome are measured at the same point in time. This design makes it impossible to determine temporality whether the exposure preceded the outcome, which is a key criterion for establishing causality. As a result, while you can identify associations between variables, you cannot infer whether one factor caused the other.

5.6 Information bias

Measurement issues also arise as self-reported data may be subject to reporting errors, social desirability bias or misclassification. Self-reported behavioral data are vulnerable to several biases. In addition to missing responses, participants may provide socially desirable answers, especially for sensitive topics such as sexual behavior or condom use. This social desirability bias can lead to under-reporting of stigmatized behaviors such as multiple sexual partners, inconsistent condom use and over-reporting of protective behaviors such as HIV testing, safer sex practices. Such misclassification typically biases associations toward the null, though in some cases it can generate spurious associations. Recall bias occurs when participants may not accurately remember past behaviors or events, especially those that are frequent or routine. Based on cultural/contextual influences, responses may be shaped by community norms or perceptions of what is acceptable to disclose. Measurement error can arise due to differences in understanding of survey questions and this can affect consistency and reliability.

5.7 Lack of critical information on other variables

Compounding the above-mentioned challenges is the lack of access to biological factors and clinical records meaning critical variables such as viral load and ART data are often underutilised in research. Viral load and ART treatment are important clinical characteristics, but they are not readily available in population-based survey data. While these variables provide valuable information about disease progression, treatment response, and survival among individuals already living with HIV, they cannot be incorporated into predictive modeling of HIV risk in the general population. This is because both viral load and ART status are post-diagnosis variable which are only relevant once an individual is already infected. Including them in risk prediction models for HIV acquisition would therefore introduce bias and limit the model’s applicability for prevention, as the goal is to identify predictors that are observable prior to infection.

5.8 Scope for further studies

Areas that could be explored for further studies include conducting longitudinal studies to confirm causality and temporal relationships between risk factors and HIV acquisition. Cross sectional studies identify associations and not causation.

Predictive model validation and calibration to assess model performance in external datasets since models built on one dataset might not generalise well.

Spatial Geographic analysis at district level to identify geographical clusters or hotspots of risk. This will help with implementing target interventions effectively.

6. REFERENCES

1. Kansiime S, Hansen CH, Hayes R, Ruzagira E. Developing HIV risk prediction tools in four African settings. Trop Med Int Health TM IH. 2023 Sep;28(9):720–30.

2. UNAIDS. AIDS data 2016 [Internet]. 2016 [cited 2024 Jul 13]. Available from: https://www.unaids.org/sites/default/files/media_asset/2016-AIDS-data_en.pdf

3. UNAIDS. South Africa UNAIDS [Internet]. 2022 [cited 2024 Jul 13]. Available from: https://www.unaids.org/en/regionscountries/countries/southafrica

4. Birri Makota RB, Musenge E. Predicting HIV infection in the decade (2005-2015) pre-COVID-19 in Zimbabwe: A supervised classification-based machine learning approach. PLOS Digit Health. 2023 Jun;2(6):e0000260.

5. Kharsany ABM, Karim QA. HIV Infection and AIDS in Sub-Saharan Africa: Current Status, Challenges and Opportunities. Open AIDS J. 2016;10:34–48.

6. Obermeyer CM. HIV in the Middle East. BMJ. 2006 Oct 21;333(7573):851–4.

7. Mutai CK, McSharry PE, Ngaruye I, Musabanganji E. Use of machine learning techniques to identify HIV predictors for screening in sub-Saharan Africa. BMC Med Res Methodol. 2021 Jul 31;21(1):159.

8. Dong Y, Liu S, Xia D, Xu C, Yu X, Chen H, et al. Prediction Model for the Risk of HIV Infection among MSM in China: Validation and Stability. Int J Environ Res Public Health. 2022 Jan 17;19(2).

9. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. J R Stat Soc Ser B Methodol [Internet]. 2018 Dec;58(1):267–88. Available from: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

10. Fieggen J, Smith E, Arora L, Segal B. The role of machine learning in HIV risk prediction. Front Reprod Health. 2022;4:1062387.

11. Xiao J, Ding R, Xu X, Guan H, Feng X, Sun T, et al. Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. J Transl Med. 2019 Apr 11;17(1):119.

12. Mayer KH, Agwu A, Malebranche D. Barriers to the Wider Use of Pre-exposure Prophylaxis in the United States: A Narrative Review. Adv Ther. 2020 May;37(5):1778–811.

13. Dunkle KL, Jewkes RK, Brown HC, Gray GE, McIntryre JA, Harlow SD. Gender-based violence, relationship power, and risk of HIV infection in women attending antenatal clinics in South Africa. Lancet Lond Engl. 2004 May 1;363(9419):1415–21.

14. Ayton SG, Pavlicova M, Abdool Karim Q. Identification of adolescent girls and young women for targeted HIV prevention: a new risk scoring tool in KwaZulu Natal, South Africa. Sci Rep. 2020 Aug 3;10(1):13017.

15. Krakower DS, Naja-Riese GM, Edelstein ZR, Gandhi AD, Wahnich A, Fischer MA. Academic Detailing to Increase Prescribing of HIV Pre-exposure Prophylaxis. Am J Prev Med. 2021 Nov;61(5 Suppl 1):S87–97.

16. Harrison A, Colvin CJ, Kuo C, Swartz A, Lurie M. Sustained High HIV Incidence in Young Women in Southern Africa: Social, Behavioral, and Structural Factors and Emerging Intervention Approaches. Curr HIV/AIDS Rep. 2015 Jun;12(2):207–15.

17. Harrison A, Cleland J, Frohlich J. Young people's sexual partnerships in KwaZulu-Natal, South Africa: patterns, contextual influences, and HIV risk. Stud Fam Plann. 2008 Dec;39(4):295–308.

18. Sodeinde K, Sholeye O, Adebisi O, Omotosho A, Sodeinde A, Abiodun O. Sex for food: a transactional sex reaction to the COVID-19-induced rise in the prevalence of hunger among in-school female adolescents in Ogun State, Nigeria. Pan Afr Med J. 2023;44:177.

19. Aral SO, Torrone E, Bernstein K. Geographical targeting to improve progression through the sexually transmitted infection/HIV treatment continua in different populations. Curr Opin HIV AIDS. 2015 Nov;10(6):477–82.

20. Smith MK, Rutstein SE, Powers KA, Fidler S, Miller WC, Eron JJJ, et al. The detection and management of early HIV infection: a clinical and public health emergency. J Acquir Immune Defic Syndr 1999. 2013 Jul;63 Suppl 2(0 2):S187-199.

21. Moyo RC, Govindasamy D, Manda SO, Nyasulu PS. A prediction risk score for HIV among adolescent girls and young women in South Africa: identifying those in need of HIV pre-exposure prophylaxis. HIV Res Clin Pract. 2023 Jun 3;24(1):2221377.

22. Croft, Trevor N., Allen, Courtney K., Zachary, Blake W. Guide to DHS Statistics [Internet]. USAID; 2023 [cited 2024 Jul 13]. Available from: https://dhsprogram.com/data/Guide-to-DHS-Statistics/index.cfm

23. Macro International. AIDS Indicator Survey : HIV Testing Manual. Macro International; 2007. [cited 2024 Dec 20]. Available from: https://dhsprogram.com/pubs/pdf/AISM8/HIV_Testing_Manual.pdf

24. Fishel, J.D., and Dean G. Performance of Enzyme Immunoassays for HIV Serology in Surveys Conducted by The Demographic and Health Surveys Program. ICF International; 2016.

25. Omotayo AO, Olagunju KO, Omotoso AB, Ogunniyi AI, Otekunrin OA, Daud AS. Clean water, sanitation and under-five children diarrhea incidence: Empirical evidence from the South Africa's General Household Survey. Environ Sci Pollut Res [Internet]. 2021 Nov 1;28(44):63150–62. Available from: https://doi.org/10.1007/s11356-021-15182-w

26. DHS. The DHS Program. Demographic and Health Surveys [Internet]. 2024 [cited 2024 Dec 22]. Available from: https://dhsprogram.com/data/dataset_admin/login_main.cfm

27. Lakshmi Padmaja D, Vishnuvardhan B. Variance-Based Feature Selection for Enhanced Classification Performance. In: Satapathy SC, Bhateja V, Somanah R, Yang XS, Senkerik R, editors. Information Systems Design and Intelligent Applications. Singapore: Springer Singapore; 2019. p. 543–50.

28. Karami S, Saberi-Movahed F, Tiwari P, Marttinen P, Vahdati S. Unsupervised feature selection based on variance-covariance subspace distance. Neural Netw Off J Int Neural Netw Soc. 2023 Sep;166:188–203.

29. Nattino G, Pennell ML, Lemeshow S. Assessing the goodness of fit of logistic regression models in large samples: A modification of the Hosmer-Lemeshow test. Biometrics. 2020 Jun;76(2):549–60.

30. Lee PH. Is a cutoff of 10% appropriate for the change-in-estimate criterion of confounder identification? J Epidemiol. 2014;24(2):161–7.

31. Burstyn I. The mockery that confounds better treatment of confounding in epidemiology: The change in estimate fallacy. Glob Epidemiol. 2024 Dec;8:100166.

32. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009 Jun 29;338:b2393.

33. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res. 2011 Mar;20(1):40–9.

34. Napoli A, White P. Clustering-Based Validation Splits for Model Selection under Domain Shift. In 2024. Available from: https://api.semanticscholar.org/CorpusID:270123051

35. Dobbin KK, Simon RM. Optimally splitting cases for training and testing high dimensional classifiers. BMC Med Genomics. 2011 Apr 8;4:31.

36. Simm J, Humbeck L, Zalewski A, Sturm N, Heyndrickx W, Moreau Y, et al. Splitting chemical structure data sets for federated privacy-preserving machine learning. J Cheminformatics. 2021 Dec 7;13(1):96.

37. Dancho, M. XGBoost Hyperparameter Tuning (how to do my 2 step process in R) [Internet]. XGBoost Hyperparameter Tuning (how to do my 2 step process in R). 2024 [cited 2024 Dec 20]. Available from: https://www.business-science.io/code-tools/2024/01/12/xgboost-hyperparameter-tuning.html

38. Balkus JE, Brown E, Palanee T, Nair G, Gafoor Z, Zhang J, et al. An Empiric HIV Risk Scoring Tool to Predict HIV-1 Acquisition in African Women. JAIDS J Acquir Immune Defic Syndr [Internet]. 2016 Jul 1 [cited 2024 Jun 25];72(3):333–43. Available from: https://journals.lww.com/00126334-201607010-00016

39. Stoner MCD, Pettifor A, Edwards JK, Aiello AE, Halpern CT, Julien A, et al. The effect of school attendance and school dropout on incident HIV and HSV-2 among young women in rural South Africa enrolled in HPTN 068. AIDS Lond Engl. 2017 Sep 24;31(15):2127–34.

40. David JK, Pant R, Allam RR, Priya VMP, Aridoss S, Arumugam E. The relationship between educational attainment and hiv prevalence among pregnant women attending antenatal clinics in six states of India: Sentinel surveillance from 2010 to 2017. Indian J Public Health. 2020 Apr;64(Supplement):S15–21.

41. Etowa J, Ghose B, Loemba H, Etowa EB, Husbands W, Omorodion F, et al. Factors Associated with Condom Knowledge, Attitude, and Use among Black Heterosexual Men in Ontario, Canada. ScientificWorldJournal. 2021;2021:8862534.

42. McCarthy M, Kauer S, Fisher C. Descriptive norms about condom use predict odds of using a condom during last sexual experience in a large, national survey of adolescents from Australia. Sex Health. 2022 Jun;19(3):157–63.

43. Hendriksen ES, Pettifor A, Lee SJ, Coates TJ, Rees HV. Predictors of condom use among young adults in South Africa: the Reproductive Health and HIV Research Unit National Youth Survey. Am J Public Health. 2007 Jul;97(7):1241–8.

44. Ward H, Rönn M. Contribution of sexually transmitted infections to the sexual transmission of HIV. Curr Opin HIV AIDS. 2010 Jul;5(4):305–10.

45. Gray R, Kigozi G, Kong X, Ssempiija V, Makumbi F, Wattya S, et al. The effectiveness of male circumcision for HIV prevention and effects on risk behaviors in a posttrial follow-up study. AIDS Lond Engl. 2012 Mar 13;26(5):609–15.

46. Weiss HA, Dickson KE, Agot K, Hankins CA. Male circumcision for HIV prevention: current research and programmatic issues. AIDS Lond Engl. 2010 Oct;24 Suppl 4(0 4):S61-69.

7. APPENDICES

7.1 Tables in appendices

Table 2.1 Percentage of missing data for variables in the female dataset

Variable	Missing	Total	Percent Missing
hiv_status	0	2,726	0.00
cluster	0	2,726	0.00
wt	0	2,726	0.00
hhmembers	0	2,726	0.00
region	0	2,726	0.00
residence_~e	0	2,726	0.00
hhhead_sex	0	2,726	0.00
hhhead_age	0	2,726	0.00
hhses	0	2,726	0.00
caseid	0	2,726	0.00
age	0	2,726	0.00
education_~l	0	2,726	0.00
ethnicity	0	2,726	0.00
relationsh~d	0	2,726	0.00
number_of_~s	0	2,726	0.00
time_away_~e	1,477	2,726	54.18
pregnancy_~n	0	2,726	0.00
marital_st~s	0	2,726	0.00
union_times	1,702	2,726	62.44
age_at_coh~n	1,702	2,726	62.44
cohabitati~n	0	2,726	0.00
ageat_firs~x	0	2,726	0.00
rcnt_sex	0	2,726	0.00
employemen~s	0	2,726	0.00
beating_ju~d	0	2,726	0.00
heardof_aids	0	2,726	0.00
condom~t_sex	601	2,726	22.05
sti	0	2,726	0.00
genital_di~e	0	2,726	0.00

sex_partners	0	2,726	0.00
recent_par~p	601	2,726	22.05
advice_aft~i	2,389	2,726	87.64
hiv_testing	0	2,726	0.00
sex_for_gi~s	2,211	2,726	81.11
condom_wit~r	1,762	2,726	64.64
refuse_sex	1,867	2,726	68.49
ask_use_co~m	1,867	2,726	68.49
sex_times_~t	601	2,726	22.05
concurrent~x	2,594	2,726	95.16
counsellin~g	2,183	2,726	80.08
partner_vi~e	1,449	2,726	53.15
sexual_vio~e	1,449	2,726	53.15
drug_use	1,449	2,726	53.15
gender	0	2,726	0.00
mcaseid	2,726	2,726	100.00
circumcised	2,726	2,726	100.00
paid_anyon~x	2,726	2,726	100.00
paying_for~x	2,726	2,726	100.00
condom~r_sex	2,726	2,726	100.00
condom_use~d	2,726	2,726	100.00
age_category	0	2,726	0.00
contracept~e	0	2,726	0.00

Table 2.2 Features with no missing observations and those with missing observations less than 30% for female dataset

Variable	Missing	Total	Percent Missing
hiv_status	0	2,726	0.00
cluster	0	2,726	0.00
wt	0	2,726	0.00
hhmembers	0	2,726	0.00
region	0	2,726	0.00
residence_~e	0	2,726	0.00
hhhead_sex	0	2,726	0.00
hhhead_age	0	2,726	0.00
hhses	0	2,726	0.00
caseid	0	2,726	0.00
age	0	2,726	0.00
education_~l	0	2,726	0.00
ethnicity	0	2,726	0.00
relationsh~d	0	2,726	0.00
number_of_~s	0	2,726	0.00
pregnancy_~n	0	2,726	0.00
marital_st~s	0	2,726	0.00
cohabitati~n	0	2,726	0.00
ageat_firs~x	0	2,726	0.00
rcnt_sex	0	2,726	0.00
employemen~s	0	2,726	0.00
beating_ju~d	0	2,726	0.00
heardof_aids	0	2,726	0.00
condom_use~x	601	2,726	22.05
sti	0	2,726	0.00
genital_di~e	0	2,726	0.00
sex_partners	0	2,726	0.00
recent_par~p	601	2,726	22.05
hiv_testing	0	2,726	0.00
sex_times_~t	601	2,726	22.05

```
      gender |          0          2,726          0.00
age_category |          0          2,726          0.00
contracept~e |          0          2,726          0.00
------------------------------------------------------------------
```

Table 2.3 Percentage of missing data for variables in the male dataset

Variable	Missing	Total	Percent Missing
hiv_status	0	2,136	0.00
cluster	0	2,136	0.00
wt	0	2,136	0.00
hhmembers	0	2,136	0.00
region	0	2,136	0.00
residence_~e	0	2,136	0.00
hhhead_sex	0	2,136	0.00
hhhead_age	0	2,136	0.00
hhses	0	2,136	0.00
caseid	2,136	2,136	100.00
age	0	2,136	0.00
education_~l	0	2,136	0.00
ethnicity	0	2,136	0.00
relationsh~d	0	2,136	0.00
number_of_~s	0	2,136	0.00
time_away_~e	1,214	2,136	56.84
pregnancy_~n	2,136	2,136	100.00
marital_st~s	0	2,136	0.00
union_times	1,367	2,136	64.00
age_at_coh~n	1,354	2,136	63.39
cohabitati~n	0	2,136	0.00
ageat_firs~x	0	2,136	0.00
rcnt_sex	0	2,136	0.00
employemen~s	0	2,136	0.00
beating_ju~d	0	2,136	0.00
heardof_aids	0	2,136	0.00
condom~t_sex	442	2,136	20.69
sti	0	2,136	0.00
genital_di~e	0	2,136	0.00
sex_partners	0	2,136	0.00
recent_par~p	442	2,136	20.69

advice_aft~i	1,982	2,136	92.79
hiv_testing	0	2,136	0.00
sex_for_gi~s	265	2,136	12.41
condom_wit~r	1,232	2,136	57.68
refuse_sex	2,136	2,136	100.00
ask_use_co~m	2,136	2,136	100.00
sex_times_~t	442	2,136	20.69
concurrent~x	1,765	2,136	82.63
counsellin~g	2,136	2,136	100.00
partner_vi~e	2,136	2,136	100.00
sexual_vio~e	2,136	2,136	100.00
drug_use	2,136	2,136	100.00
gender	0	2,136	0.00
mcaseid	0	2,136	0.00
circumcised	0	2,136	0.00
paid_anyon~x	265	2,136	12.41
paying_for~x	265	2,136	12.41
condom~r_sex	2,083	2,136	97.52
condom_use~d	2,091	2,136	97.89
age_category	0	2,136	0.00
contracept~e	0	2,136	0.00

Table 2.4 Features with no missing observations and those with missing observations less than 30% for male dataset

Variable	Missing	Total	Percent Missing
hiv_status	0	2,136	0.00
cluster	0	2,136	0.00
wt	0	2,136	0.00
hhmembers	0	2,136	0.00
region	0	2,136	0.00
residence_~e	0	2,136	0.00
hhhead_sex	0	2,136	0.00
hhhead_age	0	2,136	0.00
hhses	0	2,136	0.00
age	0	2,136	0.00
education_~l	0	2,136	0.00
ethnicity	0	2,136	0.00
relationsh~d	0	2,136	0.00
number_of_~s	0	2,136	0.00
marital_st~s	0	2,136	0.00
cohabitati~n	0	2,136	0.00
ageat_firs~x	0	2,136	0.00
rcnt_sex	0	2,136	0.00
employemen~s	0	2,136	0.00
beating_ju~d	0	2,136	0.00
heardof_aids	0	2,136	0.00
condom_use~x	442	2,136	20.69
sti	0	2,136	0.00

```
genital_di~e |          0          2,136          0.00

sex_partners |          0          2,136          0.00

recent_par~p |        442          2,136         20.69

 hiv_testing |          0          2,136          0.00

sex_for_gi~s |        265          2,136         12.41

sex_times_~t |        442          2,136         20.69

      gender |          0          2,136          0.00

     mcaseid |          0          2,136          0.00

 circumcised |          0          2,136          0.00

paid_anyon~x |        265          2,136         12.41

paying_for~x |        265          2,136         12.41

age_category |          0          2,136          0.00

contracept~e |          0          2,136          0.00

-------------+----------------------------------------------
```

Table 3.4 Male and female potential features selected for prediction using LASSO

	Male Features	Female Features
1.	Household members	Household members
2.	Household socioeconomic status	Region
3.	Age	Residence type
4.	Education level	Household head age
5.	Relationship to household head	Household socioeconomic status
6.	Number of trips (trips in last 12 months)	Age
7.	Marital status	Education level
8.	Cohabitation duration	Ethnicity
9.	Age at first sex	Relationship to household head
10.	Employement status	Marital status
11.	Beating justified if wife refuses sex	Cohabitation duration
12.	Condom use during last sex	Age at first sex
13.	Genital discharge (had any genital discharge in last 12	Recent sexual activity
14.	Recent partner relationship	Employement status
15.	HIV testing (ever been tested for HIV)	Heard of aids
16.	Sex for gifts (sex in return for gifts in last 12 months)	Condom use during last sex
17.	Circumcised	Sti (had any sti in the last 12 months)
18.	Paying for sex (paid for sex in last 12 months)	Genital discharge (had any genital discharge in last 12 months)
19.	Age category	Sex partners (Number of sex partners in last 12 months plus s
20.	Contraceptive use	Recent partner relationship
21.		Sex times with recent partner (number of times having sex)
22.		Age category
23.		Beating justified if wife refuses sex

Table 3.5 Male and female features selected by GV selection method used for multivariable analysis

	Male Features	Female Features
1.	Age category	Household members
2.	Condom use during last sex	Region
3.	Marital status	Residence type
4.	Household socioeconomic status	Household socioeconomic status
5.	Circumcised	Age
6.	HIV testing (ever been tested for HIV)	Education level
7.	Cohabitation duration	Ethnicity
8.	Contraceptive use	Relationship to household head
9.	Age	Marital status
10.	Number of trips (trips in last 12 months)	Cohabitation duration
11.	Age at first sex	Age at first sex
12.		Condom use during last sex
13.		Sti (had any sti in the last 12 months)
14.		Recent partner relationship
15.		Age category
16.		Beating justified if wife refuses sex

Table 3.8 showing interaction terms on multivariable analysis for females

HIV status	AOR	SE	z	p	95% CI
Ethnicity##age category					
Young adults (15-24) black	1.00	.	.	.	1.00,1.00
Old adults 25+	2.65	0.76	3.39	0.001	1.51,4.65
Coloured#Young adults (15-24	1.00	.	.	.	1.00,1.00
Coloured#Old Adults (25+)	0.09	0.07	-3.04	0.002	0.02,0.43
Region##age category					
Young adults (15-24)	1.00	.	.	.	1.00,1.00
Old adults (25+) W Cape	0.95	0.60	-0.09	0.929	0.27,3.26
Free State#Young Adults (15-24)	1.00	.	.	.	1.00,1.00
Free State#Old Adults (25+)	4.36	3.22	1.99	0.047	1.02,18.63
N West#Young Adults (15-24)	1.00	.	.	.	1.00,1.00
N West#Old Adults (25+)	5.06	3.92	2.09	0.037	1.11,23.17
Gauteng Young Adults (15-24)	1.00	.	.	.	1.00,1.00
Gauteng#Old Adults (25+)	6.12	4.96	2.23	0.026	1.24,30.12
Limpopo#Young Adults (15-24)	1.00	.	.	.	1.00,1.00
Limpopo#Old Adults (25+)	5.565	4.51	2.12	0.034	1.13,27.30

Table 3.9 showing interaction terms on multivariable analysis on male data

HIV status	AOR	SE	z	p	95% CI
Age category## condom use last sex					
Young adults (15-24) #no	1.00	.	.	.	1.00,1.00
Old adults25+ # no	1.00	.	.	.	1.00,1.00
Old adults25+ # yes	4.32	2.89	2.19	0.029	1.16,16.06

Table 3.11 Male and female features used for predictive modelling in R

	Male Features	Female Features
1.	Age category	Household members
2.	Condom use during last sex	Region
3.	Marital status	Household socioeconomic status
4.	Household socioeconomic status	Age
5.	Circumcised	Ethnicity
6.	HIV testing (ever been tested for HIV)	Cohabitation duration
7.	Cohabitation duration	Condom use during last sex
8.		Sti (had any sti in the last 12 months)
9.		Recent partner relationship (including dummies)
10.		Age category
11.		Marital status
12.		
13.		
14.		
15.		
16.		

Table 3.14 Female risk score assessment tool

Variable		Risk Score
ethnicity	black	18
	white	0.037
	coloured	0.257
	indian/asian	0.037
recent_partner_rltnshp	casual acquaintance	11
	spouse	9
	boyfriend not living with respo	8
	live in partner	4.64
age_category	old adults	11
	young adults	1.717
condom_use_lst_sex	no	10
sti	yes	8
cohabitation_duration	never married	4
	0-4	1.175
	5-9	0.860
	10-14	0.545
	15-19	0.553
	20-24	0.207
	25-29	0.215
	30+	0.138
hhses	poorest	3
	poorer	2.527
	middle	2.838
	richer	2.031
	richest	1.142
marital_status	never in union	3
	living with partner	1.157
	married	1.065
	separated	0.317
	widowed	0.161
	divorced	0.058
region	Gauteng	1
	KwaZulu Natal	0.875
	Mpumalanga	0.348
	Eastern Cape	0.334
	Western Cape	0.275
	North West	0.254
	Free State	0.192
	Limpopo	0.167
	Northern Cape	0.038

*Overall risk score obtained from risk ranking

*Subcategory scores obtained from using survey set prevalences

HIV prevalence in old adults is 86.5.2% and in young adults is 13.5%

Young adult score (11/86.5) * 13.5= 1.717

Score above 34.5 high risk refer for counselling and testing

Table 3.15 Male risk score assessment tool

Variable		Risk Score
age_category	old adults young adults	25 3.9017
Circumcised	no Yes	11 8.1632
HIV testing	yes no	10 1.2108
condom_use_lst_sex	no	7
marital_status	never in union living with partner married separated widowed divorced	3 1.1574 1.0653 0.3167 0.1612 0.0576
hhses	poorest middle poorer richer richest	3 2.8385 2.5269 2.0308 1.1423
cohabitation_duration	never married 0-4 5-9 10-14 15-19 20-24 25-29 30+	1 0.2937 0.2150 0.1363 0.1382 0.0518 0.0537 0.0345

*Overall risk score obtained from risk ranking

*Subcategory scores obtained from using survey set prevalences

HIV prevalence in circumcised is 57% and in circumcised 42.3%

Circumcised score (11/57) * 42.3= 8.1632

Score above 30. high risk refer for counselling and testing

Table 3.16 Male and female features used in Orange

	Male Features	Female Features
1.	Age category	Household members
2.	Condom use during last sex	Region
3.	Marital status	Household socioeconomic status
4.	Household socioeconomic status	Age
5.	Circumcised	Education level
6.	HIV testing (ever been tested for HIV)	Ethnicity
7.	Cohabitation duration	Relationship to household head
8.	Contraceptive use	Marital status
9.	Age	Cohabitation duration
10.	Number of trips (trips in last 12 months)	Condom use during last sex
11.	Age at first sex	Sti (had any sti in the last 12 months)
12.		Recent partner relationship
13.		Age category
14.		
15.		
16.		

Table 3.19 Optimum tuning hyperparameters for XGB and KNN on weighted and balanced data

Model	Tuning Stage	Stage / Parameters	Females (Weighted)	Males (Weighted)	Females (Balanced)	Males (Balanced)
XGB	1	Learning rate	0.039	0.089	0.071	0.0415
XGB	2	Tree depth	4	4	5	5
XGB	2	Loss reduction	3.16	2.76	2.15	1.58
XGB	2	Stop iterations	27	48	30	46
XGB	2	ROC AUC (mean)	0.741	0.762	0.879	0.941
XGB	2	Std. error	0.015	0.015	0.002	0.004
XGB	2	n	5	5	5	5
KNN	1	Neighbors (k)	47	46	5	7
KNN	2	ROC AUC (mean)	0.752	0.750	0.846	0.925
KNN	2	Std. error	0.008	0.078	0.007	0.003
KNN	2	n	5	5	5	5

7.2 Appendix 1 University of Witwatersrand Plagiarism form

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

FACULTY OF HEALTH SCIENCES

PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

SENATE PLAGIARISM POLICY: APPENDIX ONE

I Munyaradzi Nyahuma (Student number: 2727881) am a student registered for the degree of

Msc Epidemiology (Epi & Bio) Epidemiology in the academic year 2025.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.
- I have included as an appendix a report from "Turnitin" (or other approved plagiarism detection) software indicating the level of plagiarism in my research document.

Signature: ______________________ Date: 05/06/2025

1

7.3 Appendix 2 Ethics clearance certificate

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

R49 Dr M Nyahuma

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M2409110

NAME: (Principal Investigator) Dr M Nyahuma **DEGREE:** MSc

DEPARTMENT: School of Public Health
Division of Epidemiology and Biostatistics
Faculty of Health Sciences

PROJECT TITLE: *Predictive modelling of risk factors associated with HIV infection amongst adults in South Africa in 2016 using supervised machine learning*

DATE CONSIDERED: Ad hoc

DECISION: Approved unconditionally

CONDITIONS: Retrospective record review

NOTE: If contact information regarding student study participants is required, please contact the Registrar's office - <Nicoleen.Potgieter@wits.ac.za>

SUPERVISOR: Professor E Musenge

APPROVED BY: Professor P Ruff, Chairperson, HREC (Medical)

DATE OF APPROVAL: 2024/10/25 **EXPIRY DATE:** 2029/10/24

This Clearance Certificate is valid for 5 years from the date of approval. An extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and ONE COPY returned to the Research Office secretariat on the 3rd floor, Phillip Tobias Building, Parktown, University of the Witwatersrand, Johannesburg.

I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated from the research protocol as approved, I/we undertake to submit details to the Committee. I agree to submit a yearly progress report in the format available at https://wits.ac.za/research/researcher-support/research-ethics/ethics-committees/. This report is due annually, for the duration of the Clearance Certificate, beginning on the first anniversary of Date of Approval above. Unreported changes to the study may invalidate the clearance given by the HREC (Medical).

Signature of Principal Investigator Date

7.4 Appendix 3 Turn it in report

Final Research Report for Munyaradzi Nyahuma 2727881 Epi and Bio edited.docx

ORIGINALITY REPORT

12%	8%	10%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	wiredspace.wits.ac.za Internet Source	2%
2	www.frontiersin.org Internet Source	1%
3	open.uct.ac.za Internet Source	<1%
4	Thankian, Kusathan, Sidney O. C. Mwaba, and Anitha J. Menon. "Domestic Violence and Sexual Health among Young Women in Zambia", African Research Review, 2015. Publication	<1%
5	hdl.handle.net Internet Source	<1%
6	Tougher, Sarah. "Economic analysis of the resources required to support children orphaned or made vulnerable by HIV/AIDS in South Africa with cash transfers", Proquest, 20111003 Publication	<1%

7.5 Codes used for female variable selection

Stata codes used for female variable selection

** Feature selection using LASSO

lasso logit hiv_status hhmembers region residence_type hhhead_sex hhhead_age hhses age education_level ethnicity relationship_hhhead number_of_trips pregnancy_termination marital_status cohabitation_duration ageat_first_sex rcnt_sex employement_status beating_justified heardof_aids condom_use_lst_sex sti genital_discharge sex_partners recent_partner_rltnshp hiv_testing   sex_times_with_recent gender age_category contraceptive_use

***To see selected features

```
 lassocoef

---------------------------------

                      |  active

----------------------+----------

            hhmembers |     x

               region |     x

       residence_type |     x

           hhhead_age |     x

                hhses |     x

                  age |     x

      education_level |     x

            ethnicity |     x

  relationship_hhhead |     x

       marital_status |     x

cohabitation_duration |     x

      ageat_first_sex |     x

             rcnt_sex |     x

   employement_status |     x
```

```
     beating_justified |      x
           heardof_aids |      x
       condom_use_lst_sex |      x
                     sti |      x
       genital_discharge |      x
            sex_partners |      x
  recent_partner_rltnshp |      x
   sex_times_with_recent |      x
            age_category |      x
                   _cons |      x
----------------------------------
```

Generalized variance feature selection codes for female dataset

gvselect <term> hhmembers hhhead_age age cohabitation_duration ageat_first_sex sex_partners sex_times_with_recent region residence_type hhses education_level ethnicity relationship_hhhead marital_status employement_status beating_justified heardof_aids condom_use_lst_sex sti genital_discharge recent_partner_rltnshp rcnt_sex age_category, nmodels(2): logit hiv_status <term>

```
Optimal models:

   # Preds        LL        AIC        BIC

         1 -1260.139  2524.277    2535.6
         1 -1272.495   2548.99  2560.313
         2 -1227.437  2460.873  2477.858
         2 -1232.185  2470.369  2487.354
         3 -1204.898  2417.797  2440.443
         3 -1205.967  2419.934  2442.581
         4 -1191.283  2392.566  2420.874
         4 -1192.644  2395.289  2423.596
         5 -1181.788  2375.575  2409.544
```

5	-1182.406	2376.813	2410.782	
6	-1169.009	2352.019	2391.65	
6	-1171.947	2357.893	2397.524	
7	-1159.698	2335.396	2380.688	
7	-1160.809	2337.618	2382.91	
8	-1151.809	2321.617	2372.571	
8	-1153.487	2324.974	2375.927	
9	-1146.739	2313.478	2370.093	
9	-1147.614	2315.227	2371.842	
10	-1142.93	2307.859	2370.136	
10	-1143.548	2309.096	2371.373	
11	-1140.471	2304.941	2372.88	
11	-1140.644	2305.287	2373.226	
12	-1138.291	2302.582	2376.182	
12	-1138.414	2302.828	2376.428	
13	-1136.481	2300.962	2380.224	
13	-1136.575	2301.15	2380.412	
14	-1134.953	2299.907	2384.829	
14	-1134.958	2299.916	2384.839	
15	-1133.748	2299.496	2390.08	
15	-1133.838	2299.677	2390.261	
16	-1132.535	2299.071	2395.317	* selected model lowest AIC
16	-1132.792	2299.585	2395.831	
17	-1131.64	2299.28	2401.187	
17	-1131.931	2299.861	2401.769	
18	-1131.046	2300.093	2407.662	
18	-1131.062	2300.125	2407.694	

```
19 -1130.424  2300.848  2414.079
19 -1130.485   2300.97    2414.2
20 -1129.854  2301.709  2420.601
20 -1130.035   2302.07  2420.962
21 -1129.467  2302.935  2427.488
21 -1129.709  2303.419  2427.972
22 -1129.361  2304.721  2434.936
22 -1129.384  2304.768  2434.983
23 -1129.279  2306.559  2442.435
```

```
age_category
1 : age
2 : condom_use_lst_sex age_category
2 : ethnicity age_category
3 : condom_use_lst_sex age_category hhses
3 : condom_use_lst_sex ethnicity age_category
4 : condom_use_lst_sex ethnicity age_category hhses
4 : condom_use_lst_sex ethnicity age_category education_level
5 : condom_use_lst_sex cohabitation_duration marital_status ethnicity age
5 : condom_use_lst_sex ethnicity age_category hhses sti
6 : condom_use_lst_sex cohabitation_duration marital_status ethnicity age
hhses
6 : condom_use_lst_sex cohabitation_duration marital_status ethnicity age sti
7 : condom_use_lst_sex cohabitation_duration marital_status ethnicity age
hhses sti
7 : condom_use_lst_sex cohabitation_duration marital_status ethnicity age
age_category hhses
```

8 : condom_use_lst_sex cohabitation_duration marital_status ethnicity age age_category hhses sti

8 : condom_use_lst_sex cohabitation_duration marital_status ethnicity age age_category sti education_level

9 : condom_use_lst_sex cohabitation_duration marital_status ethnicity age age_category hhses sti education_level

9 : condom_use_lst_sex cohabitation_duration marital_status ethnicity age age_category hhses sti residence_type

10 : condom_use_lst_sex cohabitation_duration marital_status ethnicity age age_category hhses sti education_level

residence_type

10 : condom_use_lst_sex cohabitation_duration marital_status ethnicity age age_category hhses sti education_level

region

11 : condom_use_lst_sex cohabitation_duration marital_status ethnicity age age_category hhses sti education_level

region relationship_hhhead

11 : condom_use_lst_sex cohabitation_duration marital_status ethnicity age age_category hhses sti education_level

region hhmembers

12 : condom_use_lst_sex cohabitation_duration marital_status ethnicity age age_category hhses sti education_level

region ageat_first_sex relationship_hhhead

12 : condom_use_lst_sex cohabitation_duration marital_status ethnicity age age_category hhses sti education_level

region residence_type relationship_hhhead

13 : condom_use_lst_sex cohabitation_duration marital_status ethnicity age age_category hhses sti education_level

region ageat_first_sex residence_type relationship_hhhead

13 : condom_use_lst_sex cohabitation_duration marital_status ethnicity age age_category hhses sti education_level

region ageat_first_sex recent_partner_rltnshp hhmembers

14 : condom_use_lst_sex cohabitation_duration marital_status ethnicity age age_category hhses sti education_level

```
    region ageat_first_sex recent_partner_rltnshp relationship_hhhead
hhmembers

14 : condom_use_lst_sex cohabitation_duration marital_status ethnicity age
age_category hhses sti education_level

    region ageat_first_sex recent_partner_rltnshp residence_type
relationship_hhhead

15 : condom_use_lst_sex cohabitation_duration marital_status ethnicity age
age_category hhses sti education_level

    region ageat_first_sex recent_partner_rltnshp residence_type
relationship_hhhead hhmembers

15 : condom_use_lst_sex cohabitation_duration marital_status ethnicity age
age_category hhses sti education_level

    region ageat_first_sex recent_partner_rltnshp residence_type
relationship_hhhead beating_justified

16 : condom_use_lst_sex cohabitation_duration marital_status ethnicity age
age_category hhses sti education_level

    region ageat_first_sex recent_partner_rltnshp residence_type
relationship_hhhead beating_justified hhmembers
```