PREDICTING HIV STATUS AMONG WOMEN IN SOUTH AFRICA USING MACHINE LEARNING: COMPARING DECISION TREE MODEL AND LOGISTIC REGRESSION

OLUWABUKOLA OLUWAPELUMI OLADOKUN

1220736

SUPERVISOR: Prof Rod Alence

CO-SUPERVISOR: Dr Sasha Frade

A project report submitted to the Faculty of Humanities, University of the Witwatersrand, Johannesburg, in partial fulfilment of the requirements for the degree of Master of Arts in E-science (Data Science).

August 2019

Declaration

I declare that this project report is my own, unaided work, except where otherwise acknowledged. It is being submitted for the degree of Master of Art in E-science in the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination in any other university.

Signed this ___ day of 20____

____

Oluwabukola Oluwapelumi Oladokun

Abstract

The HIV epidemic has grown immensely to become a serious public health problem globally. 940,000 people died from HIV in 2017, and approximately 1.8 million new infections were reported worldwide in the same year. Almost half of all new HIV infections are in women aged 15-24 years old in sub-Saharan Africa. In addition, South Africa has the highest HIV rate worldwide with an estimated 7.2 million people living with the virus in the country. To effectively manage this epidemic, better understanding of the sociodemographic factors that influence the risk of seroconversion is needed. This can be obtained by creating a model of the HIV epidemic especially among at-risk populations. More specifically, the aim of this study is to predict the HIV status of an individual, given readily available demographic data using decision tree and comparing the results with traditional logistic regression.

Individual recode data was gotten from DHS 2016 for women in South Africa. The study sample was 7808 women aged 15-49 years living in South Africa. Data was split into training (75%) and testing (25%) datasets. The logistic regression model had the highest accuracy for both training (62.90%) and testing dataset (68.039%). Accuracy for the decision tree model was 63.93%. The AUCs from the ROC curve reported 0.652 and 0.682 for the DT and LG respectively. This means that on average, a woman will be predicted as HIV negative 65.2% of the time as compared to being HIV positive using the DT model and 68.2% using the LG model.

The accuracy of both models was not high enough with the logistic regression unexpectedly having a higher accuracy, the accuracy of the decision tree model could have been impacted due to overfitting. In addition, demographic data might not be enough to accurately predict HIV status especially at the medical classification level, or more variables are needed to build the model. It is also recommended that different input features be tested, as well as automatic relevance detection to assess which inputs contribute to the output of the model.

For my parents, family, friends, colleagues and supervisors, thank you for your understanding, patience and support.

Acknowledgements

I would like to thank my supervisors Prof. Rod Alence and Dr Sasha Frade for their valuable guidance, encouragement, support throughout the research process. Thanks to the National E-science Postgraduate Teaching and Training Platform (NEPTTP) for providing the much-needed financial support, which is much appreciated. The database used in this study was from the Demographic and Health Surveys (DHS) Program (DHS, 2016) and Statistics South Africa (Stats SA), in partnership with the South African Medical Research Council (SAMRC), conducted the South Africa Demographic and Health Survey 2016 (SADHS 2016).

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AIDS	Acquired Immune Deficiency Syndrome
ART	Antiretroviral Treatment
AUC	Area Under the Curve
CV	Cross Validation
DHS	Demographic Health Survey
DT	Decision Tree
HAART	Highly Active Antiretroviral Therapy
HIV	Human Immune Virus
IPV	Intimate Partner Violence
LR	Logistic Regression
MSEM	Modified Social Ecological Model
SSA	Sub-Saharan Africa
STI	Sexually Transmitted Infection
ROC	Receiver Operating Characteristics

CHAPTER ONE: INTRODUCTION

This chapter gives an overview of the HIV burden of disease in the global, regional and country level context. The problem statement and the justification of the research are provided. And finally, the research question and the objectives of the study are also outlined.

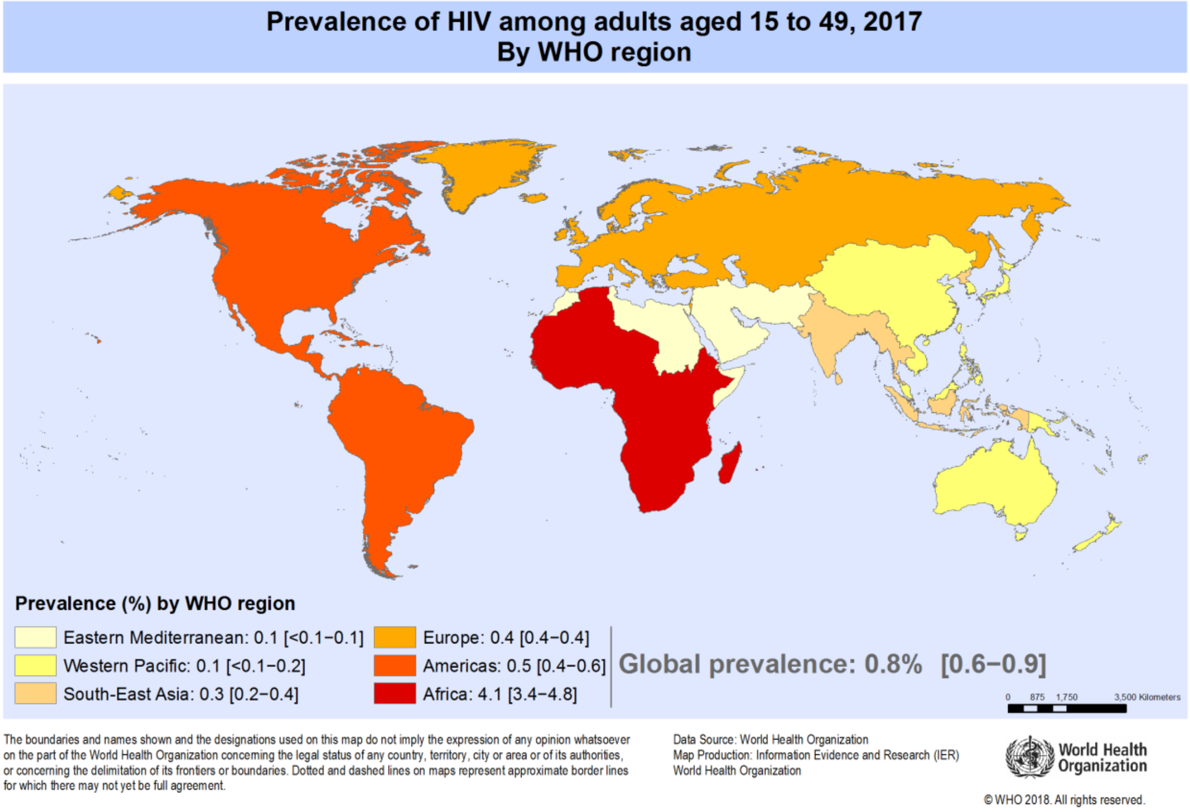
1.1 Background

The human immune virus (HIV) is a retrovirus that infects immune cells such as the CD4 cells which interferes with the body’s ability to fight organisms that cause diseases (Seitz, 2016). Acquired immune deficiency syndrome (AIDS) is as a result of the gradual weakening of the immune system that has been infected with the virus (Seitz, 2016). Although HIV is primarily a sexually transmitted disease, it is also spread and transmitted by contact with infected blood from mother to the child during pregnancy, as well as during childbirth and breast-feeding (Malani, 2016). Without the appropriate treatment, the progress from HIV to AIDS could take years, at which point the immune system is compromised. Currently, there is no treatment for eradicating the virus, but the existing treatments can control and slow down the progress of the virus by preventing replication of the virus, leading to a small amount of the viral load in the blood (Seitz, 2016).

The HIV and AIDS epidemic has grown immensely and is an important public health problem. Globally, it has been the leading cause of death since its discovery in the early 1980s, more than three decades ago (Kharsany & Karim, 2016). Since the onset of the epidemic, a lot of research has been conducted to understanding the virus and how it affects the human system and leads to AIDS - the final stage of the disease. Unfortunately, despite several decades of research, there is still no known cure for HIV and those living with HIV, as well as those at risk of HIV do not have access to prevention, care or treatment (Deeks, Lewin, & Havlir, 2013).

Since the start of the epidemic, more than 74.9 million people have been infected with HIV and more than 32.0 million have died from AIDS-related illnesses (UNAIDS, 2019a). In 2017, 940,000 people died from HIV, and approximately 1.8 million people became newly infected globally (WHO, 2018a). According to UNAIDS (2018) statistics, approximately 37.9 million people were living with HIV globally in 2018, of which 36.2 million are adults and 1.7 million are children less than 15 years old (UNAIDS, 2019a). People newly infected with HIV were approximately 1.7 million, and those that died from AIDS-related illnesses were approximately 770,000 in 2018 (UNAIDS, 2019a). It is reported that 79% of people living with HIV know their status while about 8.1 million do not know they are living with HIV (UNAIDS, 2019a). Although this is a problem that needs to be addressed, new infections have been reduced by 40% since 1997; this decrease is seen 1.7 million in 2018 as compared to 2.9 million in 1997 (UNAIDS, 2019a).

The burden of the HIV epidemic varies significantly between regions and countries. Sub-Saharan Africa (SSA) remains the most severely affected by the HIV epidemic, with nearly 1 in every 25 adults living with HIV. The region also accounts for nearly two-thirds of the people living with HIV globally in 2017 (WHO, 2018b). However, the overall trend shows fewer new infections and decreased AIDS-related mortality in the region since 2000 (Fettig, Swaminathan, Murrill, & Kaplan, 2014). Evidence shows that this is attributed to the increase in the uptake of Antiretroviral medications (ART) which reduces mortality rates, reducing infection rates and as a result, slowly increasing the HIV prevalence in most countries (Ambrosioni, Calmy, & Hirschel, 2011). Global coverage of ART in 2015 reached 46% and in South Africa alone, 3.4 million people living with HIV were on treatment. Due to the increase of prevention and care services, there has been a 45% decrease in the disease rate since its highest peak in 2005. However, HIV/AIDS still remains the leading cause of death and disability worldwide, especially in SSA (del Rio, 2017).


Prevalence of HIV among adults aged 15 to 49, 2017
By WHO region
Prevalence (%) by WHO region
Eastern Mediterranean: 0.1 [<0.1–0.1]
Western Pacific: 0.1 [<0.1–0.2]
South-East Asia: 0.3 [0.2–0.4]
Europe: 0.4 [0.4–0.4]
Americas: 0.5 [0.4–0.6]
Africa: 4.1 [3.4–4.8]
Global prevalence: 0.8% [0.6–0.9]
0 875 1,750 3,500 Kilometers
World Health Organization

The boundaries and names shown and the designations used on this map do not imply the expression of any opinion whatsoever on the part of the World Health Organization concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate border lines for which there may not yet be full agreement.

Data Source: World Health Organization
Map Production: Information Evidence and Research (IER)
World Health Organization

©WHO 2018. All rights reserved.

Figure 1: Global Prevalence of HIV 2017

Over 50% of people living with HIV globally are women, this percentage increase to 59% in SSA which makes them an at-risk population (Fettig et al., 2014). A major risk factor for HIV infection in sub-Saharan Africa is mother-to-child transmission of HIV. However, the introduction of Prevention of Mother to Child Transmission (PMTCT) services has reduced the mother to child transmission rate and approximately 80% of pregnant women with HIV are receiving treatment in sub-Saharan Africa (WHO, 2018c). Regardless of noteworthy breakthroughs in addressing the HIV epidemic in SSA, 80% of all new infections among adolescents are accounted for by adolescent girls. With approximately 7,000 girls and young women aged 15 – 24 years being infected every week and having up to eight-fold higher risk of HIV infection compared to their male counterparts (Williams et al., 2015).

The economic impact of HIV/AIDS is severe as it removes people in their prime from working thus adversely affecting development and growth. Socially, health resources are overly stretched, diverting attention from other health issues. To effectively address this epidemic; 1) accurate and reliable information on the existing number of cases (prevalence), 2) better understanding of the sociodemographic factors that influence the risk of seroconversion, and 3) an understanding of interventions needed to address the socio-demographic factors, to facilitate behavioural changes, as well as treatment interventions. This can be achieved by developing a model of the HIV epidemic especially among at-risk populations.

1.2 Problem Statement

The epidemiological impact of HIV reveals the impact of HIV and AIDS on the society, households as well as individuals. HIV impacts development by lowering the life expectancy of a population (Waziri, Mohamed Nor, Raja Abdullah, & Adamu, 2015). Consequently, reduction in life expectancy affects the economic development due to increased financial strain on the government to treat and prevent the epidemic, as well as make up for the loss of a productive contributor to the economy (Waziri et al., 2015). At the household and individual level, the immediate impact of HIV is experienced when there are members of the household infected with the virus (Hosegood, 2009). These households would have to spend more money to provide ART to infected family members. This extra financial strain for households who already have a lower earning capacity, fosters poverty (Hosegood, 2009). The epidemic generates an increase in funeral, medical, and legal costs for families, and impacts on the capacity of households to stay together (Hosegood, 2009). This is evident in the amount of children that have been orphaned or lost one parent due to AIDS - 63% in South Africa as of 2015 (WHO, 2015).

The World Health Organisation (WHO) also reported that 820 000 women and men aged 15-24 years were newly infected with HIV in developing countries (WHO, 2018a). In 2010, 2.7 million new HIV infections were recorded worldwide, and 70% of these were in SSA (UNAIDS, 2013). SSA is the region most affected by HIV, with 25.7 million people living with HIV in 2017 and accounts for more than half of the global total new HIV infections (Kharsany & Karim, 2016). Women in SSA also continue to carry a disproportionately high burden of HIV in the region. Young women in the region are reported to be infected approximately ten years before their male counterparts (Kharsany & Karim, 2016). It was reported that in SSA, 71% of people living with

HIV aged 15-24 years are women and almost half of all new HIV infections are in women aged 15-24 (Ramjee & Daniels, 2013).

In South Africa, an estimated 7.2 million people were living with HIV in 2017, making South Africa the country with the biggest HIV and AIDS burden globally. In SSA, South Africa accounts for a third of all the new infections in the region. Individuals aged 15-49 accounted for 18.8% of the adult HIV prevalence in the country. It was reported that in 2017, there were 270,000 new HIV infections, and an estimated 110,000 South Africans died from AIDS-related illnesses (UNAIDS, 2018). HIV prevalence is still on the high, 19% amongst the general population, it however varies between regions. In KwaZulu-Natal, HIV prevalence is almost 12.2% as compared with 6% and 7% in the Western Cape and Northern Cape, respectively (SANAC, 2017).

South Africa has made progress towards the 90-90-90 goal whose aim was to diagnose 90% of all HIV-positive persons, provide ART for 90% of those diagnosed, and achieve viral suppression for 90% of those treated by 2020 (UNAIDS, 2019). South Africa runs the largest ART programme in the world and in 2015, was spending $1.34 billion annually to run its HIV programmes (UNAIDS, 2017). However, according to the report from the 9th South African AIDS Conference, only the first goal of diagnosing 90% of all HIV positive persons has been reached, but it has not reached its target for treatment coverage and prevention (UNAIDS, 2019b).

The aim of this study was to predict the HIV serostatus of an individual (HIV positive or HIV negative), using sociodemographic and behavioral data that are readily available in the health sector. At the end of the day, the knowledge gained from this study will be useful in constructing health and social policies for the prevention and management of HIV/AIDS. Machine learning methods have been successfully used in medicine and health informatics for decision making and will be utilized in this study for data mining. Logistic regression and decision trees have very similar purposes, however, decision tree as a machine learning method is a major strength of this study as it is said to be better for prediction than inference (Makridakis, Spiliotis, & Assimakopoulos, 2018). Therefore, it is of great importance to understand both techniques, their strengths and weakness and the difference between the results they produce. Social science variables and concepts have different classes and exhibit complex nonlinear relationship that cannot be explained using coefficients from logistic regression (Patty & Penn, 2015). Machine learning methods can be useful for developing better theories for future predictions and for testing external validity (Grimmer, 2015). In addition, it can be used to understand the future impact of proposed policies in real time. The current availability of cloud computing which is cost effective makes it possible for policy makers to submit potential policy outcomes and then have it analyzed to determine a set of potential policy changes that could achieve those outcomes (Patty & Penn, 2015).

1.3 Justification

To facilitate the goal of improving women's reproductive health outcomes in underdeveloped regions, effective HIV risk reduction strategies are needed. One of the strategies is to further reduce the spread of HIV and AIDS; to assess the level at which an individual is at risk of HIV based on factors such as biological, social, and other behavioral factors (Gerbi, Habtemariam, Robnett, Nganwa, & Tameru, 2012). Although it is imperative that everyone gets tested regardless of risk level, early detection and treatment of the virus plays a critical role in slowing the progression to AIDS and enabling people to live normal lives. One of the key areas researchers and policy makers have is to efficiently identify the maximum number of HIV infected persons, encourage voluntary testing for those at risk, and implementing ART, which the national 90-90-90 target focuses on achieving.

Although the goal of testing 90% people to know their HIV status has been achieved, the targets for prevention and treatment has not been achieved (UNAIDS, 2019b). Due to this, new HIV infections might occur, and continual testing is required (Wong, Murray, Phelps, Vermund, & McCarraher, 2017). Studies in South Africa has shown that factors such as socio-economic, education, employment status, residence type, and perception of risk hinder people from testing (MacPhail, Pettifor, Moyo, & Rees, 2009; Musheke et al., 2013; Peltzer, Matseke, Mzolo, & Majaja, 2009). The new National Strategic Plan has acknowledged that closing these gaps in HIV screening will be a key priority in the future (UNAIDS, 2019b). They plan on decentralising HIV testing, to ensure that places of work and local community settings are able to encourage HIV testing and to provide HIV tests (UNAIDS, 2019b). This study addresses this gap with the aim of predicting HIV status. If health centres can predict HIV status easily through machine learning, based on readily available socio-demographic factors, there might be a greater chance at increasing HIV testing by educating the patient on their risk, because it has been shown that perceived risk increases the chances of HIV testing.

1.4 Research Question

Does machine learning models (decision trees) perform better in predicting HIV status among women in South Africa than the traditional logistic regression model?

1.5 Research Hypothesis

1) Null Hypothesis: There is no significant difference in the predictive accuracy of decision tree and logistic regression in predicting HIV status among women in South Africa.
2) Alternative Hypothesis: Decision trees predicts HIV status among women in South Africa more accurately than logistic regression.

1.6 Research Aim and Objectives

The aim of this study was to develop a decision tree and logistic regression model that predicts HIV status of women in South Africa based on sociodemographic factors.

1) To determine the prevalence of HIV among women in South Africa during 2016 by socio-demographic and behavioural variables
2) To predict HIV status of women in South Africa using decision trees and logistic regression
3) To compare results of traditional logistic regression model and machine learning methods

CHAPTER TWO: LITERATURE REVIEW AND THEORETICAL FRAMEWORK

This chapter provides a pertinent review of the literature. First, it provided a review on the risk factors associated with HIV according to the literature. Thereafter, relevant HIV models and machine learning methods are discussed considering HIV research. Finally, a theoretical and conceptual framework to understand the variables considering the current study.

2.1 Brief Review of the Literature

2.1.1 HIV Socio-demographic and Behavioural Risk factors

The prevalence of HIV varies among countries and within a country's different population groups. This is associated with differences in biological vulnerabilities, socio-behavioral, economic, and cultural differences, as well as socio-demographic factors. Several studies have been done in sub-Saharan Africa to determine the HIV risk factors of women in the region. In SSA, the general risk factors include: cultural practices such as circumcision and ritual cleansing (Zuma et al., 2016), other sexually transmitted infections (Kalichman, Pellowski, & Turner, 2011), contraceptive use, having sex with multiple partners, unprotected sex (Gregson et al., 2005), sexual violence (Ramjee & Daniels, 2013), high migration level, educational attainment and wealth inequality and disparities (Wand & Ramjee, 2012a). These risk factors can be grouped as sexual behavioral factors and influential risk factors. From the literature review, no study has been done to determine how at-risk an individual is of acquiring HIV based on the of the identified risk factors, especially in SSA.

Biosocial Factors

Age and sex are huge contributors to HIV risk. Women in SSA continue to carry a disproportionately high burden of HIV in the region. It was reported that among young women aged 15-24 years, 71% of them are living with HIV and almost half of all new HIV infections are in women in the same age group (Ramjee & Daniels, 2013). A South African study found that approximately 6% of young boys are HIV positive, which is lower than the prevalence of HIV among young girls and women in the country (Fallis, 2013). Gender inequalities, sexual violence and abuse, and inequity in the access to services increases the vulnerability and susceptibility of women to HIV infection. In addition to being biologically vulnerable to HIV and sexually transmitted infections (STI's), adolescent girls are more likely to have older sexual partners who are Injection Drug Users (IDUs) and have other sexual partners, consequently increasing their potential exposure to HIV.

Education

In addition to age, the level of educational attainment has been identified as a risk factor for HIV in SSA. Education increases with age, and with an increase in education, the less likely it is for a girl to marry early (Palamuleni, 2011). When females are formally educated, they will be able to make an informed choice of marriage, negotiate safe sex, earn income for the family, reduce her risk of violence in a relationship, and have health knowledge that would protect her (Ackermann & Klerk, 2002). Poor education and low literacy level have been identified as a major determinant of poor and negative health outcome in SSA, including HIV and it has been found that individuals with low levels of education are at greater risk of being HIV positive (Bärnighausen, Bloom, & Humair, 2007; Gabrysch, Edwards, Glynn, & Study Group on Heterogeneity of HIV Epidemics in African Cities, 2008). Education can play an important role in reducing the burden of HIV and AIDS by increasing health education and creating opportunities for prevention and better disease management (Mondal & Shitan, 2013).

Social Disparities and Inequalities

Due to the colonial history of Africa, race also plays a role in education and educational opportunities due to various disparities and distribution of opportunities between the different racial groups. The Black population has consistently been found to have less access to education than other groups in the South Africa due to financial constraints (Weissman et al., 2015). In South Africa, the prevalence of HIV positive individuals is 13.3% times more among Black individuals, than in other racial groups in the country (Shisana et al., 2014). Besides racial differences, there are urban-rural differences in the spread of HIV in the region, which is also associated with low levels of education and poverty at both the individual and societal levels (Weissman et al., 2015).

A cross national study among African countries found that high HIV prevalence was significantly and positively associated that countries with low national net income, low Gross Domestic Product (GDP), as well as political instability. This, however, wasn’t the case in South Africa and Nigeria, which have a high GDP and high HIV prevalence rates compared to other SSA countries (Andoh, Umezaki, Nakamura, Kizuki, & Takano, 2006).

Low income neighbourhoods have limited access to social and health structures that prevent the spread of HIV, such as a clinic for testing and consultation and an adequate number of trained health specialists. Living in a low-income neighbourhoods also increases the risk of engaging in transactional sex as a means of income (Maganja, Maman, Groves, & Mbwambo, 2007a). A Kenyan study found that HIV prevalence is lower among higher socio-economic classes in urban areas than among low income urban and rural dwellers (Magadi, 2016). In Africa, transactional sex does not only refer to sex work, there are different cultural practices that encourage that men be the material and financial provider in a relationship and this in turn creates an unbalanced power scale leaving the woman vulnerable in the relationship (Maganja, Maman, Groves, & Mbwambo, 2007; Stoebenau, Heise, Wamoyi, & Bobrova, 2016). This imbalance of power increases the chance of engaging in HIV risk behaviours such as not negotiating for safe sex, having multiple sexual partners, and sexual violence (Stoebenau et al., 2011). The results in the studies by (Chirinda & Peltzer, 2014; Davidoff-Gore, Luke, & Wawire, 2011; Rodrigo & Rajapakse, 2010) indicates that poor and less educated people were less likely to use condoms, thus increasing their risk and exposure to HIV. Other identified factors for this urban-rural difference is level of knowledge, stigma, urban poverty, social cohesion and cultural practices (Ayodele & Ayodele, 2016; Naidoo et al., 2009; Yehadji, 2010).

Population Migration

Public Health literatures have identified population mobility and migration as a driver for increase in HIV transmission. In SSA, the focus in literature has been on South African miners and East

African truck drivers (Coffee et al., 2005; Crush, Grant, & Frayne, 2007; Weine & Kashuba, 2012). Although there are contradictory findings on the effects of migration and how it actually plays a role in the epidemic, mostly due to differences in the definition of migration and mobility, there has been a consensus that migration increases the rate of human interaction, and expose people to sexual partners that are coming from high prevalence areas (Deane, Parkhurst, & Johnston, 2010; Weine & Kashuba, 2012). A study among migrants in rural Tanzania found that the prevalence of HIV among recent migrants was 3.7% more than non-migrants to the area, and that migrants contributed significantly to the increase in HIV prevalence in the area (Mmbaga et al., 2008). This is especially true among low level workers such as truck drivers, miners and hawkers (Corno & de Walque, 2012), especially in rural areas (Camlin et al., 2010). A study by McGrath, Hosegood, Newell, and Eaton (2015) suggests that migration has significant consequences and HIV risk which are particularly disadvantageous to women such as; engagement in risk behaviour especially by migrant males such as drug and alcohol use, and commercial sex, in the context of migration may place women at higher risk of acquiring HIV than men. According to McGrath et al., (2015) the high number of sexual partners increases the likelihood of seroconverting to being HIV positive for migrating women to a greater extent than for migrating men.

Violence against Women

Violence against women particularly sexual violence and intimate partner violence (IPV) are major public health problems and a violation of women's human rights (WHO, 2017). In 2013, WHO reported that globally, 35% of women have experienced sexual and intimate partner violence in their lives (WHO, 2014). Violence can negatively affect women's mental, physical, sexual, and reproductive health, and increases the chances of acquiring HIV (WHO, 2017). A study found that HIV positive women in SSA reported higher rates of intimate partner violence (IPV) than HIV positive women in the USA (Campbell et al., 2008). Dunkle, Nduna, and Shai (2010) found that HIV positive women reported more lifetime partner violence compared to HIV negative women. A study in Uganda found that an increase in IPV was related to gender inequality in the African context, multiple partners, alcohol and drug use, and poverty (Kouyoumdjian et al., 2013). In South

Africa, the act of ‘Rape’ is a significant driver of HIV transmission among young women and it is reported that there are over 2000 rape cases in the country every year (R. Jewkes & Morrell, 2010).

Condom Use

In South Africa, contraceptive use has been understood as an important determinant of HIV transmission (Haddad et al., 2014). Although consistent use of contraceptive methods such as condoms are important for women who wish to not get pregnant, it is highly important to prevent and reduce the risk of HIV sero-conversion between the sexual partners (Prat, Planes, Gras, & Sullman, 2016). Despite this, the prevalence of contraceptive use has been found to be low among populations with the highest risk and prevalence of HIV. Among the sexually active in South Africa, the reported rate of condom use ranges between 32.8% and 78.4% (van Loggerenberg et al., 2012). The level of educational attainment was found to influence the probability of condom use. Studies found that females with more years in high school and higher education were more likely to be using different methods of contraceptives especially condoms (Alemayehu, Belachew, & Tilahun, 2012). In addition, van Loggerenberg et al. (2012) found that HIV/AIDS knowledge, gender inequality and access to formal education prevent safer sex practices among women in South Africa. Condom use is also affected by alcohol and drug use which is evident in Africa in the form of binging. It is also associated with HIV transmission rates as it influences sexual decision-making, safe sex negotiation and condom use (Seth C. Kalichman, Simbayi, Vermaak, Jooste, & Cain, 2008).

Community Influences

It was found that especially for females, being an orphan, being unmarried and a partner not being circumcised was associated with increased risk of HIV seroconversion (Gregson et al., 2005; Thurman, Brown, Richter, Maharaj, & Magnani, 2006). Social connectedness has been found to possibly have a positive effect on sexual initiation as an outcome of sexual and reproductive health (Barber & Schluterman, 2008; Regnerus & Luchies, 2006). Social connectedness which is having a supportive relationship and a sense of belonging to social structures such as parents, family, peer groups and the community (Barber & Schluterman, 2008; Regnerus & Luchies, 2006). It is recommended that the integration of these social groups into sexual and reproductive health intervention programs would be critical for enhancing the protective effects (Markham et al., 2010). Although parental monitoring and parental overcontrol might also pose as a risk factor for early sexual debut, condom and contraceptive use (Manlove, Ryan, & Franzetta, 2007). Finally, adolescents who dropped out of school are more likely to have had sex earlier than adolescents who were currently in school (Sambisa et al., 2010).

2.1.2 Existing HIV Models and Machine Learning Related Studies

A handful of population-based surveys conducted nationally have been performed to determine social and behavioural factors that influence the rapid spread of HIV in South Africa. The surveys include; the South African Demographic and Health Survey (SADHS) (Department of Health 1998-2016), the Human Sciences Research Councils (HSRC) surveys (1997-2016), and the oldest, the South African Health Inequalities Survey (SAHIS 1994). However, these studies are reports and do not investigate the association between the sociodemographic factors, health risk behaviours and HIV serostatus.

Past studies have shown that it is not feasible to predict HIV status using non-clinical information. A study by Lallemant et al. (1992) used logistic regression to predict HIV positive status among pregnant women in the Congo. Results from the logistic regression analysis identified a few significant factors found to be independently associated with HIV positivity. These include age, district of residence, duration of the romantic/sexual relationship, number of decreased children, number of living children, and history of blood transfusion and/or hospitalization. However, the predictive accuracy of the model was poor; 80% of the women who were HIV positive were correctly predicted as such by the model, 50% of the truly HIV negative women were misclassified as HIV positive. Due to the poor predictive accuracy of the model, they concluded that it is challenging and close to impossible to identify a subgroup at risk of HIV, to ensure that specific actions could be targeted to them (Lallemant et al., 1992). Ayisi et al. (2000) used Poisson regression. The model showed five factors associated with HIV seropositivity; anaemia, malarial parasitaemia, and a history of previous vaginal discharge treatment, alcohol consumption and fever. Amongst the pregnant women, the researchers could not identify people at risk of HIV infection using non-clinical information, thus indicating that general access to voluntary HIV counselling and testing would be better than targeted screening.

A study in Zaire used logistic regression using various factors with the aim of providing an alternative to HIV serological screening of 15 to 45 year old women. They concluded that to predict HIV serostatus without biomarkers, the model needs to include markers of present illness that are considered AIDS/HIV-related symptoms such as diarrhoea, profound weight loss, and/or chronic fever (Hassig et al., 1990). Despite the numerous applications of artificial intelligence and machine learning for prediction in clinical medicine, little has been done to apply them to HIV/AIDS prevention and planning. A study by Lee and Park (2001) used artificial neural networks to classify and predict the symptomatic status of HIV patients. For training the model, a total of 1,026 cases was used; 667 HIV cases in total for testing the model. The variables used were: sex, IV user, race, sexual identity (heterosexual or homosexual), total number of hospital clinic visits, total number of patient admission, total number of private physician visits, total number of in-patient nights, total number of ambulatory visits, and total number of emergency room visits. After testing the models, an 88% predictive accuracy was reached.

Another study by De Queiroz Mello et al. (2006) tested classification trees and logistic regression in predicting smear negative pulmonary tuberculosis (SNPT). From the logistic regression, they generated a clinical and radiological prediction score. The area under the receiver operator characteristic curve, sensitivity, and specificity were used to evaluate the model's performance in both generation and validation samples. They found that it was possible to generate predictive models for SNPT with sensitivity ranging from 64% to 71% and specificity ranging from 58% to 76%. They concluded that the models might be useful as screening tools for estimating the risk of SNPT, optimizing the utilization of more expensive tests, and avoiding costs of unnecessary anti-tuberculosis treatment.

Kurt, Ture, and Kurum (2008), compared performances of classification techniques to predict the presence of coronary artery disease (CAD). Performances of classification techniques were compared using ROC curve, Hierarchical Cluster Analysis (HCA), and Multidimensional Scaling (MDS). They found that decision tree and logistic regression were better than other techniques in predicting CAD in according to HCA and MDS.

Moving away from the medical sector, logistic regression and decision tree have also been compared in predicting environmental hazards. Hong, Pradhan, Xu, and Tien Bui (2015) compared both methods in predicting landslide hazard in China. They found that the decision tree model yielded better overall performance and accurate results than the logistic regression model. However, concluded that they are both promising data mining techniques which might be considered to use in landslide susceptibility mapping. In the banking sector, Nie, Rowe, Zhang, Tian, and Shi (2011), used decision tree and logistic regression in credit card churn forecasting. After testing the model on the test data, the test result showed that regression performs a little better than decision tree.

Coupled with the fact that these studies are outdated, machine learning has been proposed to address HIV status prediction. Machine learning has been utilized in the medical sector for diagnosing clinical functions, for confirming diagnosis, analysing death and survival rate, and for aiding decision making. A common use of artificial intelligence is classification using predictive models. Using linear regression, linear models are very easy to develop. However, training the dataset is time consuming, and does not necessarily offers a major gain over the calculation by Fishers linear discriminant, or other statistical methods. Regardless of this, regression models are better than most machine learning methods such as decision tree in terms of interpretability. Machine learning methods are often criticised for difficulty in interpretation because none of the coefficients can be interpreted in the way regression models can be interpreted. Machine learning methods are however efficient and strong in modelling multidimensional spaces and are highly reliable in their ability because they can adapt to changes in the data.

2.2 Theoretical Framework

The HIV epidemic is very complex, with multiple interacting individual, social and structural risk factors (Kaufman, Cornish, Zimmerman, & Johnson, 2014). This study makes use of part of the modified social ecological model (MSEM) which was recently extended by Baral, Logie, Grosso, Wirtz, and Beyrer (2013) to build on past ecological frameworks - such as the model of behaviour change (which primarily focuses on individual motivations for behaviour), the theory of planned action, and the health belief model. The individual, social network, community, policy and stage of the HIV epidemic are the five layers of the HIV epidemic MSEM model. The fifth level 'stage of the HIV epidemic' was added based on the theory that the risk of an individual is important for the spread of a disease, but they are inadequate and unsatisfactory in explaining population level epidemic dynamics.

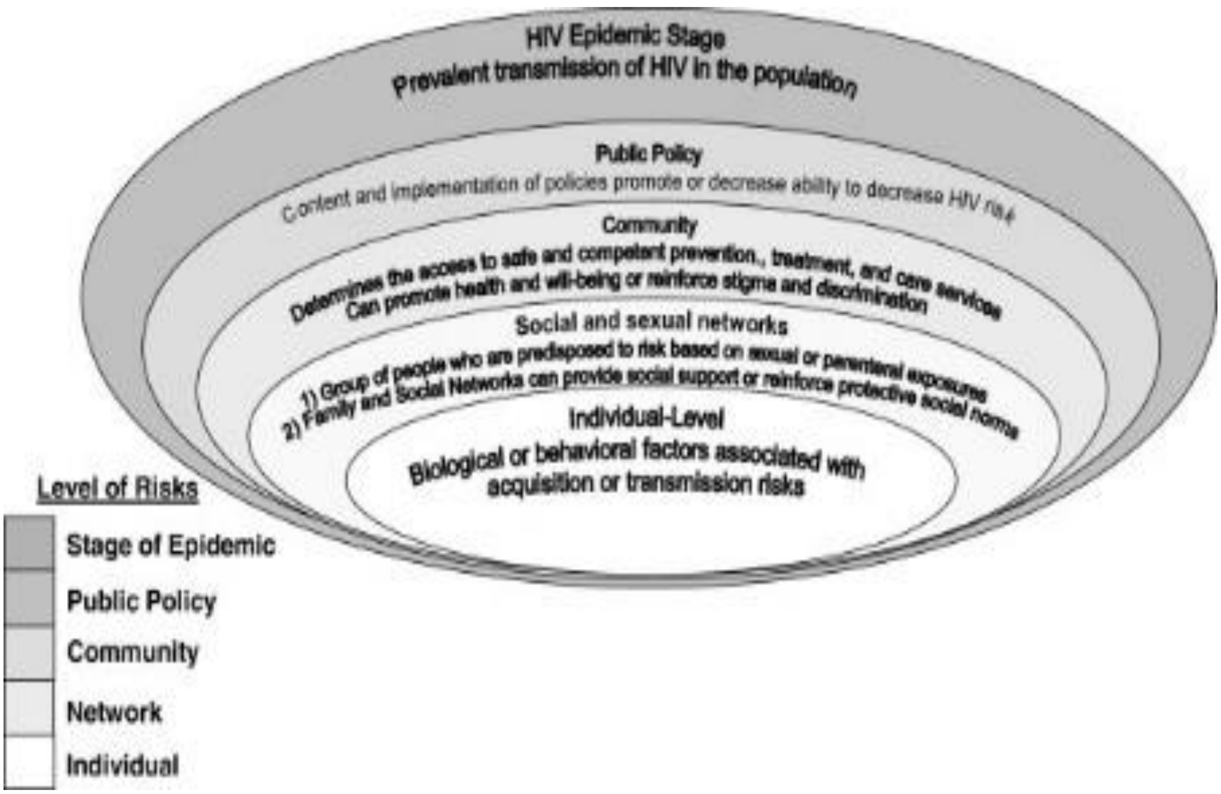

HIV Epidemic Stage
Prevalent transmission of HIV in the population
Public Policy
Content and implementation of policies promote or decrease ability to decrease HIV risk
Community
Determines the access to safe and competent prevention, treatment, and care services
Can promote health and will-being or reinforce stigma and discrimination
Social and sexual networks
1) Group of people who are predisposed to risk based on sexual or parenteral exposures
2) Family and Social Networks can provide social support or reinforce protective social norms
Individual-Level
Biological or behavioral factors associated with acquisition or transmission risks
Level of Risks
Stage of Epidemic
Public Policy
Community
Network
Individual

Figure 2 Modified socio ecological model for HIV risk (Baral et al., 2013)

Individual factors are the biological or behavioural factors that make a person vulnerable to acquire or transmit the disease. Individual factors for HIV risk are age, sex, condom and contraceptive use, and substance use. Social and sexual networks consist of interpersonal relationships such as family, friends, neighbours and others that can influence health and health behaviours directly (Poundstone, Strathdee, & Celentano, 2004). HIV risks on this level are associated with, social engagement, social networks, social influence, access to information, intimate contact, and disease prevalence (Wang, Brown, Shen, & Tucker, 2011).

The community can either be a place of wellbeing and health, or a source of stigma and ill health. Community usually refers to network ties and relationship between organisations, religions, geographic or political region, etc. Sociocultural norms and values always plays out on the community level and they influence interpersonal processes and individual behaviour (Baral et al., 2013). Stigma at the community level prevents populations at risk utilizing the provisions available, as well as uptake of HIV prevention, treatment, and care services.

Laws and policies provide the general framework through policies and financing for shaping the risk of marginalized populations as well as the general population. Worldwide laws such as criminalization of homosexuality and sex work, criminalization of prevention practices such as needle exchange implemented based on morals, culture, and political wills rather than evidence from public health research have driven the increase in HIV infection risk (Baral et al., 2012; Degenhardt et al., 2010). Policies determine budget and resource allocation to healthcare, education, and HIV prevention services, it also affects health by driving conflict and economic disruptions that affects the provision of services (Hecht et al., 2009).

The premise for ‘stage of the epidemic’ is that individual characteristic behaviour, policy, social norms or network influence can create infectious disease. However, they can only create conditions which can either foster or supress the vulnerability of an individual to a disease that is already prevalent (Baral et al., 2013). For example, the risk of a person that engages in unprotected sex should be interpreted within the context of a high burden of HIV infection and viral load in the broader population the individual operates in (Wellings et al., 2006).

2.3 Conceptual Framework

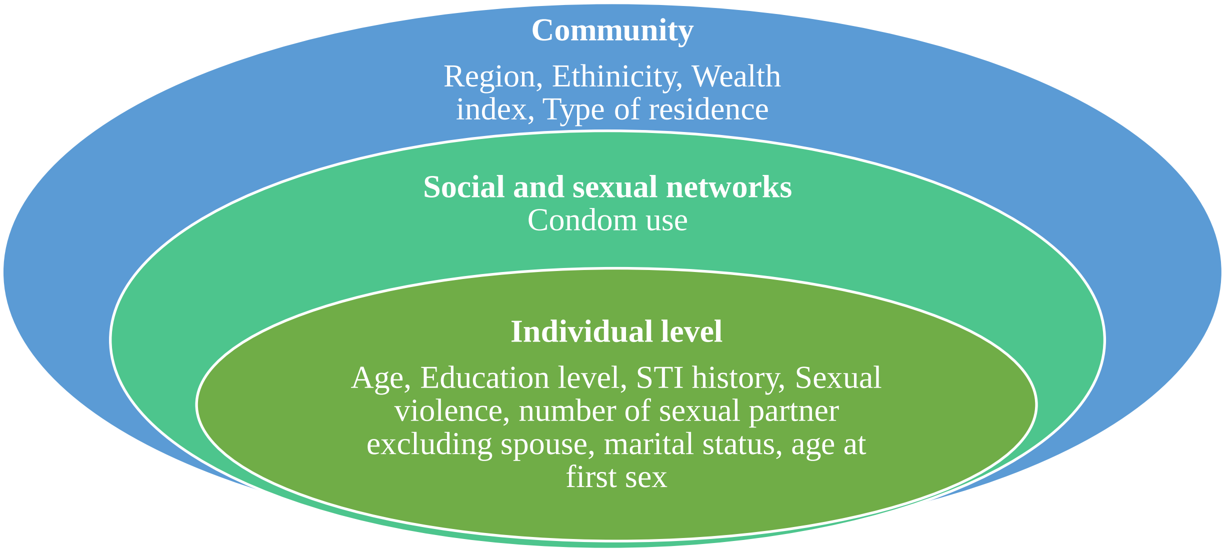

Community
Region, Ethinicity, Wealth index, Type of residence
Social and sexual networks
Condom use
Individual level
Age, Education level, STI history, Sexual violence, number of sexual partner excluding spouse, marital status, age at first sex

Figure 3: Modified ecological model for HIV risk among women in South Africa

Many women in South Africa are at great risk of being infected with HIV. The high prevalence in the country is because of many factors. As identified by the literature, these include poverty, sexual violence against women, cultural practices such as child marriages and sexual cleansing, condom use, high prevalence of STIs, and political and social factors such as stigma and service availability (Zuma et al., 2016).

Individual HIV risk factors have gained great attention in studies, and rightly so. They are the most proximal factors to an individual and they have the highest probability of exposure. HIV infection has been associated with: history of STIs, sexual risk factors such as unprotected sex, and frequency and number of sexual partners, sexual violence, and other socio-demographic factors such as age (15-24 have the highest HIV prevalence rates), and education level (Ackermann & Klerk, 2002).

At the social and sexual risk level, social norms contribute to high HIV risks among women especially among ethnic groups. The variables under examination in this level are ‘condom use’ and ‘multiple wives’. These variables increase risk of HIV among women because of unbalanced power relations with male partners. Men in the African context can have more than one wife/sexual partner which increases a woman’s exposure to the disease. In addition, a woman’s ability to negotiate safer sex is limited (Muula, 2008).

Community risk factors that drive the spread of HIV through the interaction with individual risk factors are least considered in studies. Norms and values in the community that stigmatize sexual practices and sexually diverse populations present significant barriers to access and uptake of HIV prevention and treatment services (Muula, 2008). Region, type of residence, wealth status, and ethnicity are all related and contribute to increasing the risk of an individual to HIV exposure and infection. For example, an individual with low educational level who lives in rural regions and forms part of a low-income group would have limited funds to access HIV prevention services such as HIV education and counselling. In addition, the Black population have been found to bear the greatest burden of HIV in the country (Kharsany & Karim, 2016), therefore, it is important to understand the intersecting role of social and structural discrimination or practices in shaping health risk and outcomes.

CHAPTER THREE: METHODOLOGY

This chapter gives an overview of the methodology used for this study. It outlines and discusses the data source, study population, and variable definition. This chapter also explained how the data was analysed.

3.1 Data Source

This study is makes us of a cross-sectional study design to analyse secondary data gotten from the Demographic and Health Surveys (DHS) Program (DHS, 2016). Statistics South Africa (Stats SA), in partnership with the South African Medical Research Council (SAMRC), carried out the South Africa Demographic and Health Survey 2016 (SADHS 2016) at the request of the National Department of Health (NDoH). Technical assistance was provided through The DHS Program. DHS are nationally representative household surveys that provide comprehensive data for a wide range of monitoring and impact evaluation indicators in the areas of population, health, and nutrition. To date, DHS has collected, analysed, and disseminated accurate and representative data on population, health, HIV, nutrition and more through more than 400 surveys in over 90 countries. This allows for comparative and analytical cross-national analysis of important heath topics (DHS, 2016).

3.2 Study Population

There were 8514 women aged 15 to 49 years in the individual recode of the DHS and 6912 women in HIV dataset who consented to the HIV test. The survey used a stratified two-stage sample design. Probability proportional to size sampling of primary sampling units (PSUs) at the first stage and systematic sampling of dwelling units (DUs) at the second stage was taken into consideration. The PSU size was extrapolated from the Census 2011 DU count. A total of 750 PSUs was chosen from the 26 sampling strata, resulting in 468 PSUs selected in urban areas. In traditional areas, 224 PSUs was selected, and 58 PSUs in farm areas. To get a representative sample of women in South Africa, the distribution of the women in the sample was weighted. Although the numbers may seem low, the weighted number of women in the survey accurately represents the proportion of women who live in Gauteng and the proportion of women who live in the Northern Cape; as well as represent a larger number of women interviewed (South African Department of Health, 2016). The study population is women aged 15-49 years living in South Africa. The total number of observations that will be used is 7808 because of missing HIV test results.

3.3 Variables

The Woman's Questionnaire (Individual recode) was used to collect information from all eligible women aged 15 and older. In all households, eligible women age 15-49 were asked questions on topics such as: age, education, sexual activity, knowledge of HIV/AIDS and methods of HIV transmission. A separate dataset contained the HIV test results of women aged 15 and older who consented to having their samples taken for the test (DHS, 2016). Only 58% of the women interviewed consented and submitted blood samples for testing.

Table 1: Variable definition

Variable Code/Name	Variables	Variable Description	Variable Recode
V012/age	Respondent's current age	Integer	1) 15-19 2) 20-24 3) 25-29 4) 30-34 5) 35-39 6) 40-44 7) 45-49
V024/region	Region	1)Western Cape 2) Eastern Cape 3) Northern Cape 4) Free State 5) Kwazulu-Natal 6) North West 7) Gauteng 8) Mpumalanga 9) Limpopo	1)Western Cape 2) Eastern Cape 3) Northern Cape 4) Free State 5) Kwazulu-Natal 6) North West 7) Gauteng 8) Mpumalanga 9) Limpopo
V025/res	Type of residence	1)Urban 2) Rural	1) Rural 2) Urban
V106/edu	Highest education level	0) No education 1) Primary 2) Secondary 3) Higher	1) <tertiary 1) >tertiary
V190/wealth	Wealth index	1)Poorest 2) Poorer 3) Middle 4) Richer 5) Richest	1) Low 2) Middle 3) High
V131/ethinicity	Ethnicity	1) Black/African 2) White 3) Coloured 4) Asian	1) Black/African 2) White 3) Coloured 4) Asian
V313/ModCon	Current use by contraceptive method type	0) No method 1) Folkloric method 2) Traditional method 3) Modern method	1) Non-modern 2) Modern method
V501/married	Current marital status	0) Never in union 1) Married 2) Living with partner 3) Widowed" 4 "Divorced" 5 "No longer living together/separated"	1) Not married 2) Married

V766a/multpart	Number of sexual partners excluding spouse	Integer	1) <1 partner 2) >1partner
V525/debut	Age at first sex	Integer	1) >18years 2) <18years
V763a/sti	Had any STI in last 12 months	0) No 1) Yes 2) Don't know	1) No 2) Yes
V714/job	Respondent currently working	0) No 1) Yes	1) No 2) Yes
V781/hivknow	Ever been tested for HIV	0) No 1) Yes	1) No 2) Yes
V833a/usecondom	Used condom every time had sex with most recent partners in last 12 months	1) No 2) Yes	1) No 2) Yes
Hiv06 (Dependent variable)/hiv_stat	HIV status		1) HIV negative 2) HIV positive

3.4 Handling the dataset

The data was randomly split into 75% for training dataset and 25% for test datasets based on common practice. Splitting the dataset is important for cross validation, reducing overfitting, and for evaluating the performance of the models. However, the dataset was unbalanced with 70% of the cases being HIV positive. This was a problem because machine learning algorithms are sensitive to highly unequal classes and do not work very well with imbalanced datasets. This is because an algorithm does not get the necessary information about the minority class to make an accurate information, leading to misleading predictions and accuracies. The prediction of a model that has been trained with unbalanced data will be biased towards the more common class but still have very high accuracy.

Furthermore, machine learning algorithms assume that the dataset is balanced classes, and the errors received from the classes have the same cost. Although in real life situations, the chances of obtaining an imbalanced data from the population is high because there are more healthy control samples than disease cases, it is important that this problem be addressed for prediction or classification. In addition, researchers have also shown that results from a balanced dataset improves the overall performance of the classification as compared to an imbalanced dataset.

Over sampling was done on the dataset using ROSE (Randomly Over Sampling Examples) package. The ROSE package provides functions to deal with binary classification problems in the presence of imbalanced classes. An artificial dataset was created based on cross-validation sampling methods.

Table 2 below shows the distribution of HIV negative and HIV positive women in the sample population. A total of 7808 South African women were included in this study, the majority were HIV negative (n=6129, 78.5%) and 1,679 (21.5%) were HIV positive.

Table 2: Frequency of HIV status in the datasets

	Original dataset	**75% Training dataset**	**Balanced Training dataset**	**25% Test dataset**
HIV Negative	6129	4599	2978	1580
HIV Positive	1679	1279	2900	400

3.5 Justification of Splitting the dataset

After the model training is done on the models, it cannot be assumed that the model will work well on a new data set it has not seen before. It is uncertain that model will have the desired accuracy and variance in a needed environment (Kraska et al., 2013). The assurance needed for the accuracy of the predictions of the model is gotten by a process of validation. This process of deciding whether the numerical results quantifying hypothesised relationships between variables, are acceptable as descriptions of the data (Domingos, 2012). To evaluate the performance of any machine learning model, it must be tested on some unseen data. Based on the model's performance on unseen data, it can be determined if the model is Under-fitting/Over-fitting/Well generalised (Domingos, 2012). Cross validation (CV) is one of the techniques used to test the effectiveness of a machine learning models, it is also a re-sampling procedure used to evaluate a model if we have a limited data (Arlot & Celisse, 2010). To perform CV a sample/portion of the data needs to be set aside, on which model training is not done on, and later this sample will be used for testing/validating.

A common technique for CV is "Train-Test Split" approach. In this, the complete dataset is randomly split into training and test sets. Then perform the model training on the training set only and use the test set for validation purpose, ideally split the data into 70:30 or 80:20. With this approach there is a possibility of high bias if we have limited data, because we would miss some information about the data which we have not used for training. If our data is huge and our test sample and train sample has the same distribution, then this approach is acceptable (Reitermanov, 2010).

The current data was randomly split into 75% for training dataset and 25% for test datasets based on common practice. (Reitermanov, 2010)Emphatically, the test data is a hold-out sample that is used to assess the final selected model and estimate its prediction error. Test data are not used until after the model building and selection process is complete. Test data shows how well the model will generalize, i.e., how well the model performs on new data. By new data, what is meant is data that have not been involved in the model building nor the model selection process in any way. In this case, the machine-learning train/test approach treats the testing data "as if" the response (HIV+ or not) is unknown.

3.6 Statistical Analysis Plan

Objective 1: To determine the prevalence of HIV among women in South Africa during 2016

Descriptive statistics, namely frequency and percentage distributions, was used to achieve this objective as it describes the features of the data and give important insight to individual variables. Frequency is the number of participants that fit into the category of either HIV positive or HIV negative group. Percentages were calculated to ascertain the percent of the sample that corresponds with the frequency in the category. These were presented in graphs and table format. To determine which risk factor was used as a feature in predicting, the independent variables must have a strong relationship with the dependent variable (HIV status). This was assessed using chi-square as a measure of association. Chi-square is appropriate for this kind of data because it is a nominal data.

A chi-square is called significant ($p < 0.05$) if there is an association between two variables, and non-significant ($p > 0.05$) if there is not an association (Diener-West, 2008).

Objective 2: To predict HIV risk factor level of women in South Africa in 2016

Decision Trees

Decision tree was used to predict the HIV status of a woman in South Africa. Decision tree builds regression models in the form of a flow-chart. Decision trees are used for both categorical and numerical data types and is considered to have high predictive power with high accuracy and easy to interpret complex relationships. It splits the sample into two or more homogeneous sets based on most significant differentiator variables. It breaks down the dataset into smaller subsets each time it splits and as a result, a decision tree is formed (Kotsiantis, 2013).

Decision tree is a widely used machine learning technique for predictive models as well as exploratory analysis. It is a very effective method for rapid prototyping of models. Decision trees are significant for use in situations when there is a high non-linearity & complex relationship between the dependent variable and the independent variables, in such a case, a tree model is more likely to outperform a classical regression method.

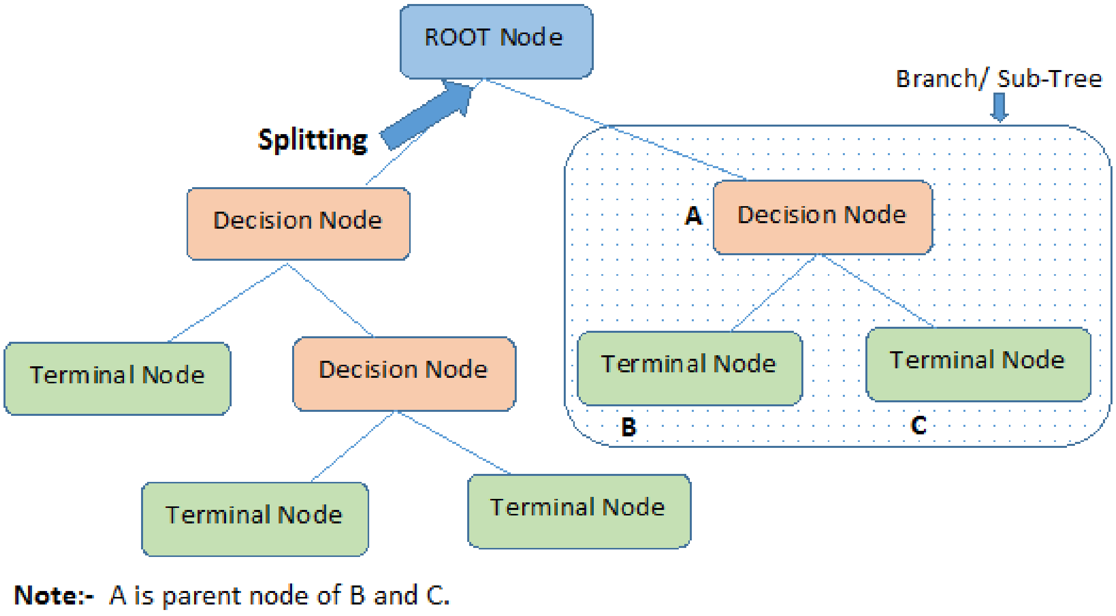


Figure 4: Decision tree structure

Calculating the entropy is how decision trees split into nodes and it is simply an indicator of how messy the data is. The aim in decision trees is to tidy the data and separate the data into the classes they belong to decrease entropy (Kotsiantis, 2013).

$$E = -\sum_i p_{i\log_2 pi}$$

where p_i are the ratios of elements of each label in the set.

Logistic Regression

Logistic regression is a nonlinear regression technique used for classifying dichotomous dependent variables. The logistic regression uses the logistic function to split the output of a linear equation between 0 and 1 instead of fitting a straight line. The logistic function is defined as:

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Where $p(x)$ is the probability of the outcome and $\beta_0 + \beta_1$ are regression coefficients

And it looks like this

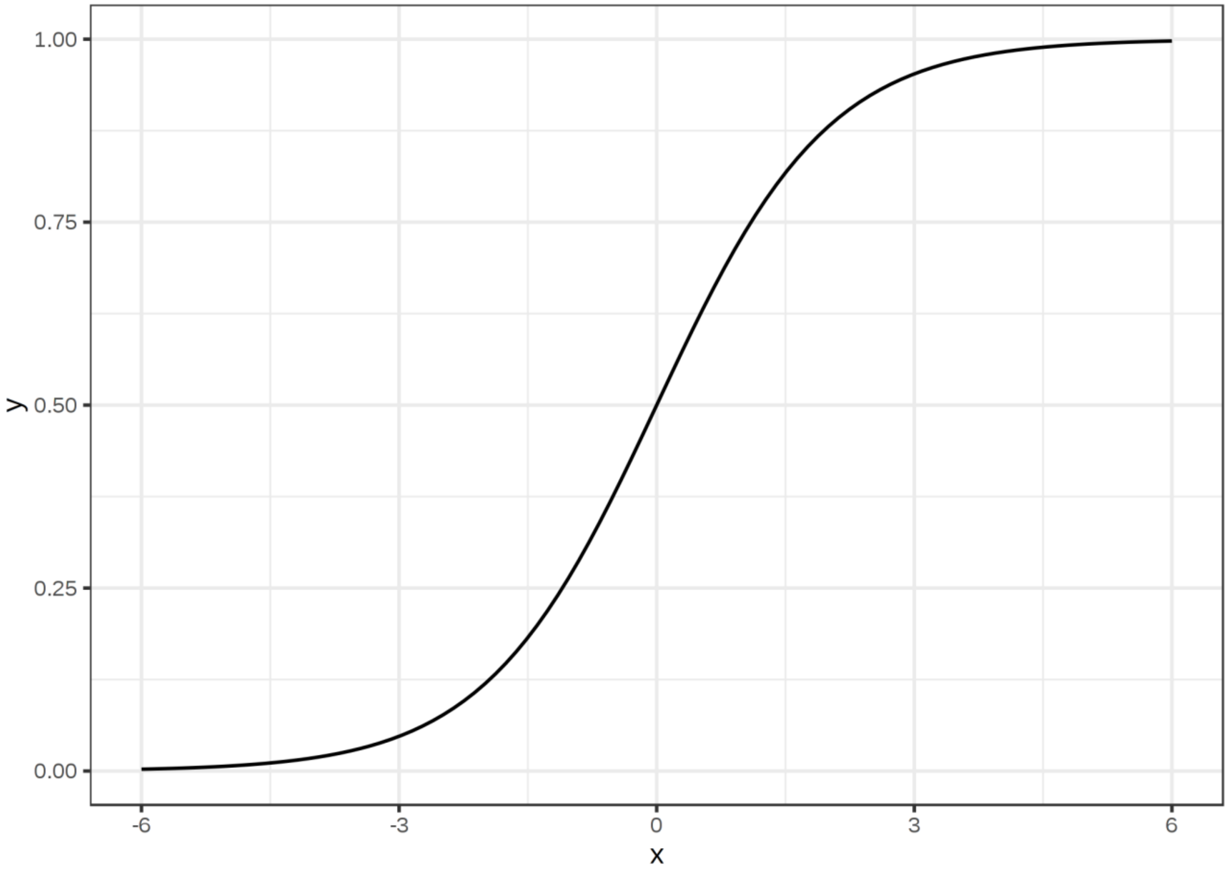


Figure 5: Logistic regression function

Logistic regression is an exponential family model therefore, it would be more informative to present exponentiated coefficients rather than the estimates (Dreiseitl & Ohno-Machado, 2002). Multicollinearities among the predictor factors was ascertained using variance inflation factors (VIF). However, and no collinearity was found among the factors. *D*ifferent models built with the training data with different variable combinations were built to examine the effect of the different factors. The first model was built with all of the variables. Fit was assessed by chi-square statistics proposed by Hosmer–Lemeshow Goodness of Fit Test.

Objective 3: To compare results from traditional logistic regression with results from machine learning methods

The performance of all models was evaluated and analysed both on the training and testing datasets. The performance of both models on the test dataset was used to evaluate the prediction capabilities and prediction accuracy. In this study, the ROC curve and five evaluation metrics was used to assess the performance of both models. The ROC curve is created by plotting the true positive rate (sensitivity) compared to the false positive rate (1 — specificity) along with the various cut-off thresholds. The area under the ROC curve (AUC) is used for the quantitative comparison of the models. The metrics derived from the confusion matrix of the models are

Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$

Sensitivity = $\frac{TP}{TP+FN}$

Specificity = $\frac{TN}{FP+TN}$

Positive predictive value = $\frac{TP}{FP+TP}$

Negative predictive value = $\frac{TN}{FN+TN}$

Where TP (True positive) is the number of cases that are classified correctly as HIV positive. TN (True negative) is the number of cases that are correctly classified as HIV negative. FP (False positive) is the number of cases incorrectly classified as HIV positive. FN (False negative) is the number of cases that are incorrectly classified as HIV negative.

CHAPTER FOUR: RESULTS

This chapter provides results of the analysis; descriptive analysis, decision tree model, and logistic regression model. Finally, a comparison of the performance of both models using statistical evaluation measures and the ROC curve.

4.1 Prevalence of HIV among women in South Africa during 2016

Table 3 below shows the socio-demographic distribution of the sample population. Of the 7808 participants, 21.2% of the participants were between the ages of 15-19 years. Those aged between 20-24 years, 25-19 years, 30-34 years, 35-39 years, 40-44 years, and 45-49 account for; 18.7%, 17.2%, 13.1%, 10.5%, 9.9%, and 9.4% of the total sample population respectively. There was no major difference in the distribution of HIV positive status by age, however women aged 25-29 years (17.6%) accounted for highest rate of HIV positive status. Followed by women aged 30-34 years (16.7%), 15-19 years (16.1%), 20-24 years (16.0%), 35-39 years (13.8%), 40-44 years (11.6%) and finally, 45-49 (8.3%). On the other hand, a decreasing pattern was seen for the distribution of HIV negative status by age starting from ages 15-19 (22.7%), down to ages 45-49 (9.7%).

A greater proportion of women in the sample were from KwaZulu-Natal (17.9%) and Mpumalanga (13.8%). Therefore, it is expected that both provinces will have the highest prevalence of HIV positive women in the sample population with 418 HIV positive women living in KwaZulu-Natal (24.9%) and 254 HIV positive women living in Mpumalanga (15.1%). While Western Cape (3.0%) and Northern Cape (4.5%) had the lowest prevalence of HIV positive women.

There is a marginal difference in the distribution of participants based on the type of place of residence; 3,815 (48.9%) lived in rural type residence and 3,995 (51.1%) lived in Urban residence. As a result, there is little difference in the HIV positive status between the women living in urban areas and those living in rural areas. Although slightly more women from rural areas were HIV positive (50.4%) as compared to women from urban areas (49.6%). Conversely, there were slightly more HIV negative women in urban areas (51.6%) than in rural areas (8.4%).

Most women fell below the tertiary level of education (n=7148, %=91.5), while only 660 women (8.5%) were above tertiary education. Consequently, there is an uneven proportion of the distribution of HIV status by education level among the women. Out of the women that were HIV positive, 1,596 women (95.1%) had an education level below tertiary education and 83 women (4.9%) had an education above tertiary education.

Additionally, the distribution of HIV status by racial group was disproportionate as there were more black women in the sample population. Therefore, black women had higher HIV positive levels, (97.5%) than the rest of the women from the other racial groups. Only 2.4% of Coloured women, 0.1% of Asian women and less than 1% of White women were HIV positive.

In the sample population, 74.7% and 74.3% of non-married women were HIV positive and HIV negative respectively. While, 25.3% and 25.7% of married women were HIV positive and HIV negative respectively.

The distribution of modern contraceptive use was roughly equal; non-modern contraceptive (n=4271, %=54.7) and modern contraceptive (n=3537, %=45.3). There is little difference in HIV positive status between women using non-modern contraceptive methods (n=882, %=52.5) and women who are using modern contraceptive (n=797, %=47.5).

Only 13.2% of the women in the sample are in relationships where the husband has more than one wife. However, 55.7% of women in polygamous relationships are HIV positive as compared to 44.3% who are not in a polygamous relationship. In addition, women who are not in polygamous relationships account for more HIV negative women in this category (n=5233, %=85.4).

Furthermore, 51.8% of the women have more than one sexual partner, and 48.2% have only one sexual partner. Consequently, 55.7% of the women who have multiple sexual partners (n=936) and 44.3% who have only one sexual partner (n=743) are HIV positive.

The majority of the women in the dataset report age of debut less than 18 years of age (n=4744, %=60.8), while 39.2% report age of debut as greater than 18 years old. Women whose age of debut is less than 18 years old report higher frequencies of being HIV positive (n=1307, %=61.8) than women whose age of debut is older than 18 years (n=642, %=38.2).

Among those who are HIV positive, 92.5% do not have a history of STI and 7.5% have a STI history. Women who are HIV positive and not employed (n=1229, %=73.2) are more than those who are employed (n=450, %=26.8) and are HIV positive. Finally, 75.5% report to use condoms and 24.5% report to not using condoms. Only 20% of women who have not used condoms in 12 months are HIV positive. While, 80% of those who report to use condoms are HIV positive.

Table 3: Frequency and Percentage Distribution of Socio-Demographic Factors by HIV Status

Variable description	**Summary n (%)**	**HIV Negative n (%)**	**HIV Positive n (%)**	**P value**
Age				
15-19	1,659 (21.2)	1,389 (22.7)	270 (16.1)	<0.001*
20-24	1,457 (18.7)	1,189 (19.4)	268 (16.0)	
25-29	1,345 (17.2)	1,049 (17.1)	296 (17.6)	
30-34	1,019 (13.1)	739 (12.1)	280 (16.7)	
35-39	821 (10.5)	590 (9.6)	231 (13.8)	
40-44	771 (9.9)	576 (9.4)	195 (11.6)	
45-49	736 (9.4)	597 (9.7)	139 (8.3)	
Province				
Western Cape	347 (4.4)	297 (4.8)	50 (3.0)	<0.001*
Eastern Cape	1,089 (13.9)	844 (13.8)	245 (14.6)	
Northern Cape	583 (7.5)	508 (8.3)	75 (4.5)	
Free State	914 (11.7)	702 (11.5)	212 (12.6)	
KwaZulu-Natal	1,395 (17.9)	977 (15.9)	418 (24.9)	
North West	943 (12.1)	735 (12.0)	208 (12.4)	
Gauteng	518 (6.6)	410 (6.7)	108 (6.4)	
Mpumalanga	1,079 (13.8)	825 (13.5)	254 (15.1)	
Limpopo	940 (12.0)	831 (13.6)	109 (6.5)	
Place of Residence				
Rural	3,815 (48.9)	2,968 (48.4)	847 (50.4)	0.142
Urban	3,993 (51.1)	3,161 (51.6)	832 (49.6)	
Educational Level				
Below Tertiary	7,148 (91.5)	5,552 (90.6)	1,596 (95.1)	<0.001*
Above Tertiary	660 (8.5)	577 (9.4)	83 (4.9)	

Wealth				
Low	1,836 (23.5)	1,405 (22.9)	431 (25.7)	<0.001*
Middle	1,604 (20.5)	1,174 (19.2)	430 (25.6)	
High	4,368 (55.9)	3,550 (57.9)	818 (48.7)	
Race				
Black	7,096 (90.9)	5,459 (89.1)	1,637 (97.5)	<0.001*
White	128 (1.6)	128 (2.1)	0 (0.0)	
Coloured	539 (6.9)	498 (8.1)	41 (2.4)	
Asian	45 (0.6)	44 (0.7)	1 (0.1)	
Marital Status				
Not Married	5,807 (74.4)	4,552 (74.3)	1,255 (74.7)	0.692
Married	2,001 (25.6)	1,577 (25.7)	424 (25.3)	
Modern Contraceptive Use				
Non-Modern	4,271 (54.7)	3,389 (55.3)	882 (52.5)	0.043*
Modern	3,537 (45.3)	2,740 (44.7)	797 (47.5)	
Polygamy				
<1	6780 (86.8)	5,233 (85.4)	1547 (92.1)	<0.001*
>1	1028 (13.2)	896 (14.6)	132 (7.9)	
Multiple Sex Partners				
<1	3,760 (48.2)	3,017 (49.2)	743 (44.3)	<0.001*
>1	4,048 (51.8)	3,112 (50.8)	936 (55.7)	
Age of Debut				
<18	4,744 (60.8)	3,707 (60.5)	1,037 (61.8)	0.341
>18	3,064 (39.2)	2,422 (39.5)	642 (38.2)	
STI history				
No	7,435 (95.2)	5,882 (96.0)	1,553 (92.5)	<.0.001*
Yes	373 (4.8)	247 (4.0)	126 (7.5)	
Employment				
No	5,701 (73.0)	4,472 (73.0)	1,229 (73.2)	0.848
Yes	2,107 (27.0)	1,657 (27.0)	450 (26.8)	
Condom Use				
No	1914 (24.5)	1579 (25.8)	335(20.0)	<0.001*
Yes	5894 (75.5)	4550 (74.2)	1344 (80.0)	

**significant at the p<0.05% level*

4.2 Predicting HIV risks of women in South Africa using decision trees

Decision trees are recursive portioning algorithms which means that when given a subset of training data, it finds the variable that best predicts the outcome using if-then-else decision rules. It further finds a split on that variable that best separates the labels, and splits into two new subsets. A conditional tree approach was used to build a decision tree. In this approach, it selects split on a variable with the lowest p-value and the tree stops building when splits are not statistically significant. The dependent variable of this decision tree (Figure 5) is HIV status which has two levels, Yes and No. The tree shows that the important variables that were useful to predict HIV status are; province, STI history, race, residence type, employment status, age, age of debut, condom use, and wealth status. The root node of the tree or the variable selected for splitting was 'Province'. This means that STI history was the most important attribute and it is closely related to HIV status. The tree further splits into 'STI history' and 'Race'.

Women who meet the criterion of node 4 (n=27) had 80% probability of being HIV positive. These women live in either the Western Cape, North West, or Limpopo Province, have a history of STI and live in the urban part of the provinces. On the other hand, those who live in the rural part of the provinces and has a history of STI (node 5, n=12) were less than 1% likely to be HIV positive. Decision node 9 (n=115) includes black women who live in either the Western Cape, North West, or Limpopo Province, do not have a history of STI, are employed, with age of debut greater than 18 years old. These women over 20% likelihood of being HIV positive. While black women with age of debut less than 18 years (node 10, n=14) have an additional 20% increase in HIV risk (40%) than those with age of debut greater than 18. On the contrary, white and coloured women who are employed, with no history of STI and live in either the Western Cape, North West, or Limpopo Provinces had less than 10% likely to be HIV positive (node 11, n=105).

Women in node 9 (n=652) who do not have a history of STI, are employed, live in either the Northern Cape or Limpopo are 35% more at risk of being HIV positive. Those who live in the Western Cape are further split by race. White and coloured women who are between the ages of 15 to 29 years have a less than 1% probability of being HIV positive (node 16, n=49), while those aged 30-49 years have an increased 60% chance of being HIV positive (node 17, n=17). On the other hand, black women in this node are 75% more likely to be HIV positive (node 18, n=82).

The second terminal node of the province root node splits into Eastern Cape, Free State, KwaZulu Natal, North West, Gauteng and Mpumalanga. This is further split by race; ‘White, Asian or Coloured’, and ‘Black’. The ‘White or Coloured’ terminal node further splits by race to divide these two categories. Coloured women from either the Eastern Cape, Free State, KwaZulu Natal, North West, Gauteng or Mpumalanga have a 25% likelihood to be HIV positive (node 21, n=59). However, White and Asian women from these provinces are less than 1% likely to be HIV positive (node 22, n=86). This means that coloured women are at higher risk of being HIV positive than White and Asian women from the same province.

Furthermore, the black women split further splits by age. Black women between the age of 15-29 and between the age of 45-49 years, from either Eastern Cape, Free State, North West, or Gauteng, who are also in the low or high wealth category are 40% more likely to be HIV positive (node 26, n=131). Those in the middle wealth category who do not use condoms are also 40% likely to be HIV positive (node 28, n=71), while those who use condoms have an unexpectedly higher risk of 60% of being HIV positive (node 29, n=228).

Black women between the age of 15-29 and between the age of 45-49 years, who live in either KwaZulu Natal or Mpumalanga, live in urban areas of the provinces, and belong to the high wealth category have a 40% increased risk of being HIV positive (node 32, n=240). Those in node 33 (n=640) that belong to the low and middle wealth category have a higher 55% probability of being HIV positive.

Women in node 35 (n=90) are black women between the ages of 15-29 and 45-49 years, live in either KwaZulu Natal or Mpumalanga, live in rural areas of the provinces and do not use condoms have a 50% probability of being HIV positive. While those who use condoms unexpectedly have a higher 80% probability of being HIV positive (node 36, n=333).

Black women between the ages of 30 and 44 years old, who live in either the Free State, North West, Gauteng, or Mpumalanga province and belong to the low and middle wealth category are

70% more likely to be HIV positive (node 39, n=327). Those who belong to the high wealth level are only 50% more likely to be HIV positive (node 40, n=584).

Black women who live in the urban areas of the Eastern Cape and are between the ages of 30 and 44 years old are 55% likely to be HIV positive (node 43, n=144). Meanwhile, those in the same age group but who live in the urban area of KwaZulu Natal are 70% more likely to be HIV positive (node 44, n=254). The last, node 45 consists of black women who live in the rural parts of the Eastern Cape and KwaZulu Natal and are between the ages of 30 and 44 years old. These women are 80% more likely to be HIV positive (n=291).

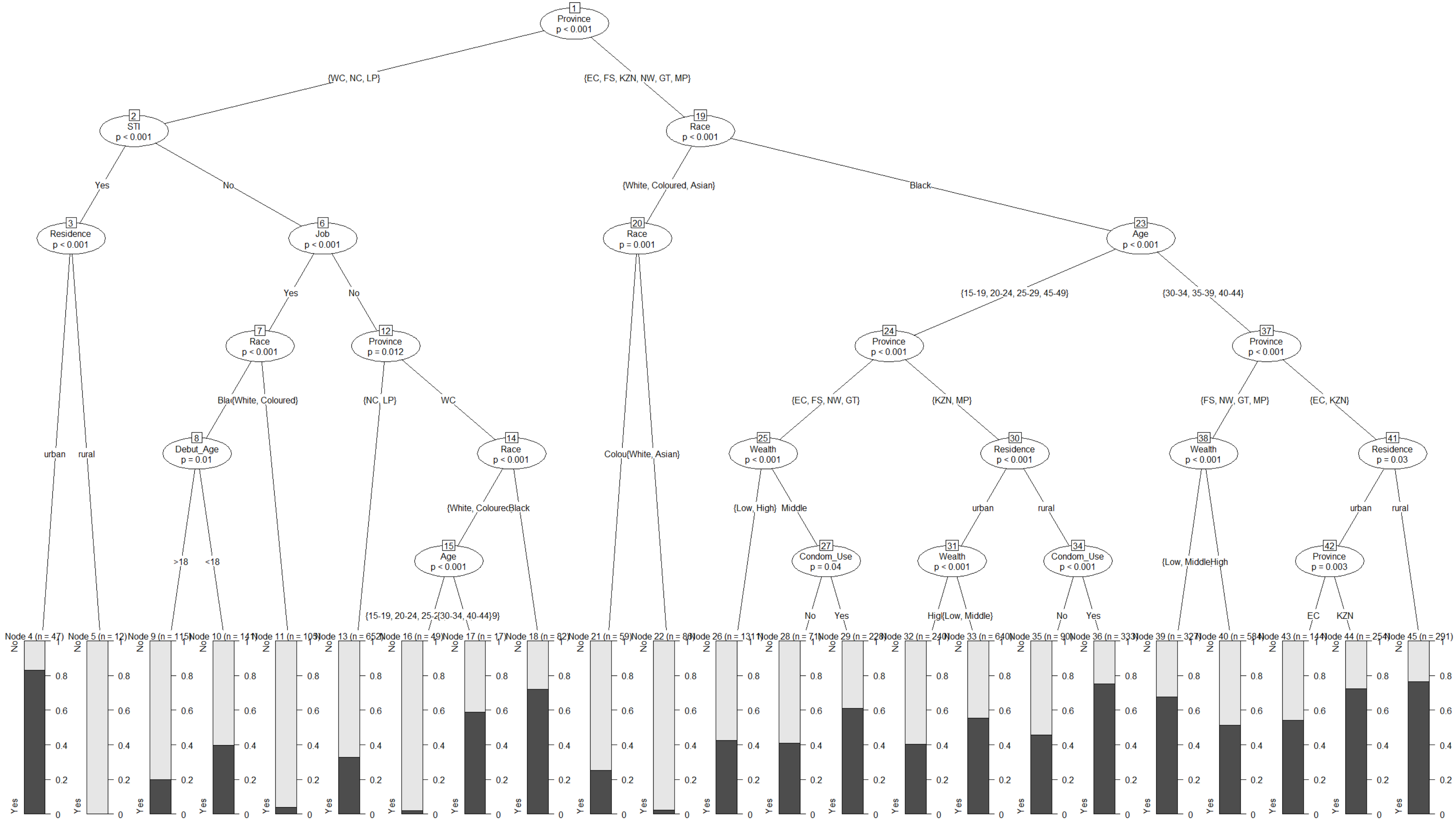


Figure 6: Decision Tree

4.3 Predicting with logistic regression

Table 3 shows the results from the logistic regression models. The model includes all the independent variables.

Ages 20-24 was positively associated with being HIV positive with a 29% increased likelihood of being HIV positive (B=0.115, CI= -0.049, 0.279, p=>0.05). Being between the ages of 25-29 years significantly increased the chances of being HIV positive by 52% (B=0.384, CI= 0.216, 0.552, p=<0.01). It is expected that HIV risk of HIV would be higher among the economically active ages of 30-49 years old. Those aged 30-34 years old are reported as having the highest risk of being HIV positive; 89% significant increased likelihood of being HIV positive (B=0.787, CI= 0.606, 0.968, p=<0.01). Those aged 35-39 years are twice as likely to be HIV positive (OR=2.34, CI= 0.630, 1.003, p=<0.01). Women between the ages of 40-44 years are also over twice as likely to be HIV positive (OR=2.12, CI= 0.519, 0.912, p=<0.01). Finally, there is 57% significant increased likelihood of women aged 45-49 to be reported as being HIV positive (B=0.604, CI= 0.405, 0.803), p=<0.01).

Living in the Eastern Cape significantly increases a woman's chances of being HIV positive by 88% (OR=1.12, CI= (-0.146, 0.366), p=>0.05). No such significance is associated with women living in the Northern Cape, as they have a 42% decreased risk of HIV (B=-0.229, CI= -0.531, 0.072), p=>0.05). The Free State also has an insignificant positive association with 66% increased chances of being HIV positive (OR=1.14, CI= (-0.132, 0.399), p=>0.05). Women living in KwaZulu-Natal had a significantly higher risk of being HIV positive (OR=2.34, CI= 0.580, 1.135), p=<0.01). In the North West, women have only 26% increased likelihood of being HIV positive, however this is not significant (B=0.209, CI= -0.075, 0.494, p=>0.05). Similarly, women living in Gauteng have 10% decreased odds of being HIV positive (B=-0.001, CI= -0.307, 0.304, p=>0.05). In Mpumalanga, the chances of being HIV positive was significantly high (OR=2.03, CI= 0.290, 0.872, p=<0.01). In addition, living in Limpopo significantly lowers the risk of being HIV positive by 32% (B=-0.389, CI= -0.692, -0.087), p=<0.05).

Living in an urban area was significantly and negatively associated with being HIV positive. Living in an urban area lowers the chances of being HIV positive by 19% (B=-0.497, CI=-0.615, -0.378), p=<0.01). Having an education above tertiary level showed a negative and significant relationship to being HIV positive, with a decreased risk of 32% (B=-0.529, CI=-

0.726, -0.331), p=<0.01). In addition, having a high wealth status is significantly associated with a 34% reduced risk of being HIV positive (B=-0.307, CI=-0.439, -0.175, p=<0.01). While, being middle-class is a 22% significant increased risk for HIV (B=0.174, CI= 0.036, 0.311, p=<0.5).

As compared to being Black, being White shows a high but insignificant negative association with HIV positivity (OR=0.00, CI= -282.902, 251.701, p=>0.05). In addition, being Coloured (OR=0.79, CI= -1.258, -0.732, p=<0.01) or Asian (OR=0.33, CI=-4.280, -0.1.814, p=<0.01). showed a significant but negative relationship to having a HIV positive status.

Being married shows an insignificant decreased risk of being HIV positive by 57% (B=-0.098, CI=-0.240, 0.043, p=>0.5). Although not statistically significant, using modern contraceptive lower the risk of being HIV positive by 13% (B=-0.061, CI=-0.160, 0.037, p=>0.05). Having more than one sexual partner significantly increases the risk of HIV by 5% (B=0.155, CI=0.030, 0.279, p= <0.05). Being in a polygamous relationship showed a significant negative association (OR=1.57, CI=-0.484, -0.111, p=<0.01).

Age of debut greater than 18 years old significantly increases the risk of HIV by 1% (B=-0.195, CI=-0.295, -0.095, p=<0.01). Having a history of STI significantly increases the chances of being HIV positive (OR=4.08, CI=0.166, 0.545, p=<0.01). There was a negative and significant association between being employed and being HIV positive, indicating a 43% lower risk of HIV (B=-0.221, CI=-0.334, -0.109, p=<0.01). In addition, using condoms increases the risk of being HIV positive by 68% (B=0.342, CI=0.228, 0.457, p=<0.01).

Table 4: Logistic Regression Models

Variables	B (Coefficients)	Confidence Interval (95%)	OR
Age (ref: age 15-19 years)			
20-24 years	0.115	-0.049,	1.29
25-29 years	0.384***	0.279	1.52
30-34 years	0.787***	0.216, 0.552	1.89
35-39 years	0.817***	0.606, 0.968	2.34
40-44 years	0.716***	0.630, 1.003	2.12
45-49 years	0.604***	0.519, 0.912	1.57
		0.405, 0.803	
Province (ref: Western Cape)			
Eastern Cape	0.299*	0.017, 0.581	1.88
Northern Cape	-0.229	-0.531,	1.42
Free State	0.114	0.072	1.66
KwaZulu-Natal	0.857***	-0.165,	2.34
North West	0.209	0.394	1.26
Gauteng	-0.001	0.580, 1.135	1.10
Mpumalanga	0.581***	-0.075,	2.03
Limpopo	-0.389**	0.494	0.68
		-0.307,	
		0.304	
		0.290, 0.872	
		-0.692, -	
		0.087	
Place of Residence (ref: Rural)			
Urban	-0.497***	-0.615, - 0.378	0.81
Educational Level (ref: Below Tertiary)			
Above Tertiary	-0.529***	-0.726, - 0.331	1.32
Wealth (ref: Low)			
Middle	0.174**	0.036, 0.311	1.22
High	-0.307***	-0.439, -	0.66
		0.175	
Race (ref: Black)			
White	-15.600	-282.902,	0.00
Coloured	-0.995***	251.701	0.79
Asian	-3.047***	-1.258, -	0.33
		0.732	
		-4.280, -	
		1.814	
Marital Status (ref: Not Married)			

Married	-0.098	-0.240, 0.043	1.57
Modern Contraceptive Use (ref: Non-modern)			
Modern	-0.061	-0.160, 0.037	1.13
Multiple Sex Partners (ref:<1)			
>1	0.155**	0.030, 0.279	0.95
Polygamy (ref: <1)			
>1	-0.297***	-0.484, -0.111	1.57
Age of Debut (ref:<18)			
>18	-0.195***	-0.295, -0.095	1.01
STI History (ref: No)			
Yes	1.389***	0.166, 0.545	4.08
Employment (ref: No)			
Yes	-0.221***	-0.334, -0.109	1.43
Condom Use (ref: No)			
Yes	0.342***	0.228, 0.457	0.68
Observations	5,878		
Log Likelihood	-3,737.808		
Akaike Inf. Crit.	7,535.616		
Note: *p<0.1; **p<0.05; ***p<0.01			

4.4 Comparing the performance of logistic regression and decision tree

Table 4 shows the comparison of the performances of the models. In the training dataset, the highest positive predictive value is for the logistic regression model (64%) which indicates the probability of the model to correctly classify HIV status in the training dataset to HIV negative class is 64%. It is followed by the decision tree which correctly classifies HIV negative women 63% of the time. In the test dataset, the positive predictive value for the decision tree and logistic regression is 78% and 80% respectively. For negative predictive value, the decision tree model shows 62% indicating that the probability to correctly classify HIV status in the training dataset to the HIV positive class, 61% for the logistic regression model. The decision tree model was better at classifying the HIV negative class while the logistic regression was better at classifying the HIV positive class correctly. The negative predictive value for the test dataset was 52% and 56% for the decision tree and logistic regression model respectively.

The decision tree model has the highest sensitivity in the training dataset, explaining that 64% of HIV status are correctly classified as HIV negative, whereas the logistic regression was 61%. Specificity is highest for the decision tree model, explaining 65% of the HIV positive which are correctly classified. Specificity score for logistic regression was 61%. The logistic regression model had the highest accuracy for both training (63%) and testing dataset (68.03%). Accuracy for the decision tree model was 62% for the training dataset and 64% for the test dataset.

Table 5: Comparison of model performance using statistical evaluation measures

Parameter	**Training data**		**Testing data**	
	Logistic regression	***Decision tree***	***Logistic regression***	***Decision tree***
Sensitivity*	61.18	64.00	64.00	57.33
Specificity*	64.66	60.93	74.47	74.47
Positive predictive value*	64.00	62.72	80.00	78.18
Negative predictive value*	61.86	62.24	56.45	52.24
Accuracy*	62.90	62.49	68.03	63.93
****expressed in percentage (%)***				

Table 6: Confusion Matrix

		Reference			
		Training		Test Data	
		HIV Negative	HIV Positive	HIV Negative	HIV Positive
Predictions					
Logistic Regression	HIV Negative	1822	1025	48	12
	HIV Positive	1156	1875	27	35
Decision Tree	HIV Negative	1906	1133	43	12
	HIV Positive	1072	1767	32	35

The ROC curve below was created using the test dataset, the AUCs are 0.697 and 0.667 for the decision tree and logistic regression respectively. This means that on average, a woman will be predicted as HIV negative 69.7% of the time as compared to being HIV positive using the decision tree model and 66.7% using the logistic regression model.

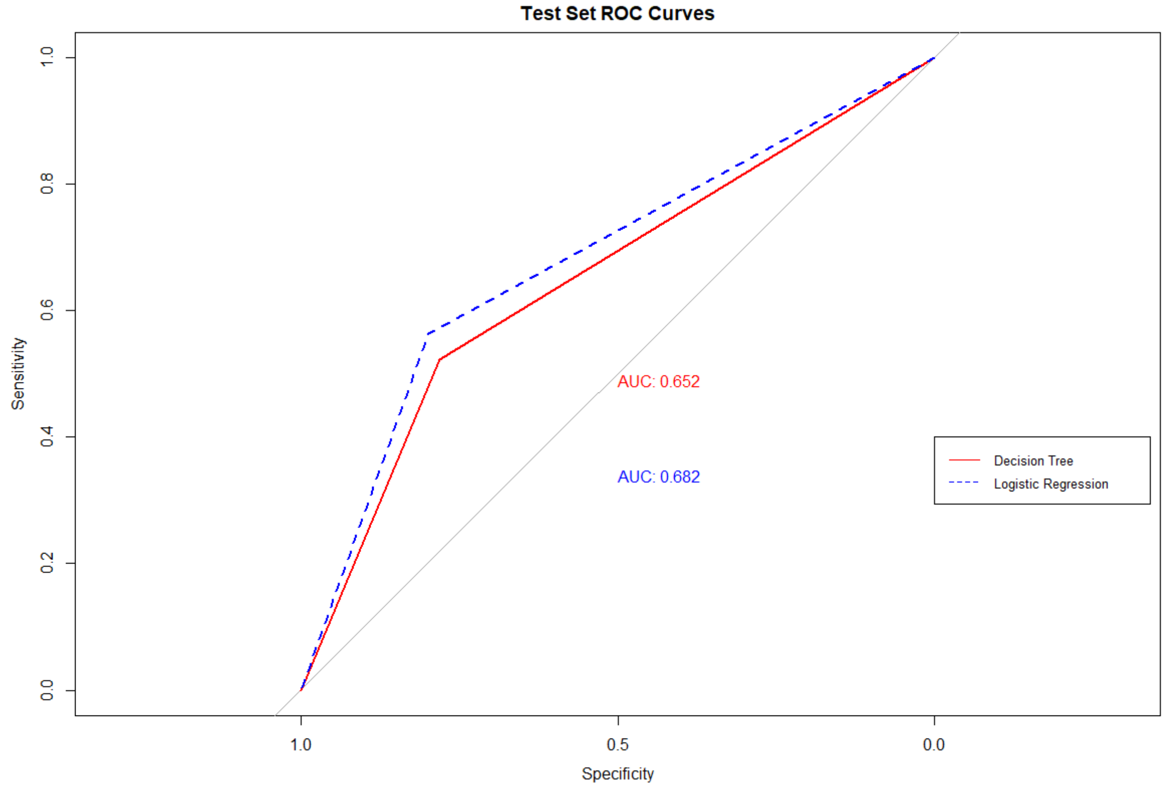


Figure 7: The ROC curve for Logistic regression and Decision tree using test dataset

CHAPTER FIVE: DISCUSSION

This section will discuss the results of the analysis of this research in predicting HIV status of women in South Africa. This section also includes the strengths and limitations of this study, the implication of the result to public health and also provides recommendations in providing solutions to addressing the problem. It reviews the prediction method that has the highest accuracy and the main factors that increase the risk of HIV. Logistic regression and decision tree have been compared in other fields with the latter having the highest accuracy (De Queiroz Mello et al., 2006; Hong, Pradhan, Xu, & Tien Bui, 2015; Kurt et al., 2008a; Nie, Rowe, Zhang, Tian, & Shi, 2011a; Shahiri et al., 2015). However, the exploration of these methods for predicting HIV status has seldom been carried out. This issue was addressed in this paper by comparing decision tree and logistic regression models.

5.1 Sociodemographic and Behavioural Factors- Decision Tree

It is well known that the overall performance of models is influenced by the variables which were used to produce them. Although the same set of variables was used to build both models, De Queiroz Mello et al. (2006) found that the accuracy of decision tree and logistic regression was dependent on the variables used; a certain combination of variables produced better accuracy for logistic regression than decision tree. De Queiroz Mello et al. (2006) also made a similar conclusion from their study that the performance of the models is related to the variables identified in their development. However, the literature reviewed shows that there is no standard guideline available. In this study, fourteen HIV risk factors were selected as independent variables. The assessment of the risk factors to HIV are important to the model.

The result show that province was the most important variable, therefore it can be concluded that it is a major factor in categorizing HIV status and increasing the risk of HIV. According to the national survey, the Northern Cape and Western Cape have the lowest percent of HIV positive women and this is attributed to the significant white population as compared to other provinces such as KwaZulu-Natal which is the country's largest and poorest province (Shisana et al., 2012; Stats SA, 2018). STI history was also selected as the next variable for a split. This is in agreement with other authors such as Bernstein, Marcus, Nieri, Philip, and Klausner (2010) McClelland et al. (2007), and Ward & Rönn (2010). Although these do not investigate the interaction between and the combination of other variables. The probability of being HIV positive was highest for those who had a history of STI and are from rural areas (Magadi, 2016;

Maganja et al., 2007a; Weissman et al., 2015) and those who lived in the Eastern Cape, Free State, KwaZulu-Natal, North West, Gauteng, or Mpumalanga. A correlation could also be made between the prevalence of STI in the provinces and HIV prevalence. According to the decision tree, the split for those who do not have a history of STI was in Western Cape, Northern Cape and Limpopo who are reported to have low HIV prevalence as compared to the rest of the provinces (Stats SA, 2018). Therefore, it could be concluded that the higher the rate of STI's in a region, the higher the increased risk of HIV.

The decision tree also showed a relationship between age and being in a polygamous relationship. The result showed a high probability of being HIV positive for women aged 15-24 and 45-49 and are not in a polygamous relationship i.e. married to a man with more than one wife. This result could be because age 15-24 is a young age category and many of them are not likely to be married, therefore the chances of them being in a polygamous relationship is low. On the other hand, the result showed that women in the age group who do not have a job and are not married have a 60% chance of being HIV positive. In addition, those who are married and those who have multiple sexual partners are 65% and 70% respectively more likely to be HIV positive than the other age groups. Going by reports in South Africa, this age group is considered to be highly sexually active, engaged in sexual risk behavior and are not properly educated on safe sex and health education (Mah, 2010; Pettifor et al., 2013). They are also prone to have multiple sexual partners as they are considered too young to be in a committed relationship (Mah & Halperin, 2010a). In addition, adolescent girls and young women in South Africa are more vulnerable to HIV infection than their male peers due to structural, social and biological factors. For example, engaging in transactional and age disparate relationships, high school drop-out, food insecurity, gender-based violence and sexual abuse, and amplification of biological factors which facilitate transmission (Dellar, Dlamini, & Karim, 2015). From these results, it is recommended that this age group be the focus for health and sexual education programs (Harrison, Newell, Imrie, & Hoddinott, 2010).

According to the decision tree, condom use shows an increased probability of being HIV positive even among those that are in high HIV prevalence provinces (Eastern Cape, Free State, KwaZulu-Natal, North West, Gauteng and Mpumalanga). This inconsistent result for an important variable could be because of missing values for condom use in the dataset, although values were imputed in the training dataset. Other studies further buttresses the fact that condoms remain a frontline defense against the HIV/AIDS pandemic (Crosby, 2013). Several studies have shown that condom use is responsible for HIV prevention between 44% and 66% (Crosby, 2013; Kurth, Celum, Baeten, Vermund, & Wasserheit, 2011).

In addition, multiple sex partner is a risk factor with a 90% increased likelihood of being HIV positive. Understanding sexual network patterns is an important dimension to understanding the spread of HIV/AIDS. Studies have shown that consistent condom use is reduced among people with multiple sexual partners (Johnston et al., 2010; Seth C. Kalichman et al., 2007). Steady sex partners of participants with multiple partnerships were significantly less likely to be protected by condoms than steady partners of individuals with only one sex partner. Individuals with multiple sex partners were also significantly less likely to have disclosed their HIV status or go for voluntary testing (Johnston et al., 2010).

Being Black was also split by age of debut less than 18 years old, with a 40% increase in the likelihood of being HIV positive as compared to black women with age of debut greater than 18 years old. In South Africa, rape is a major reason a young girl would have sex before the legal age, and rape is a significant driver of HIV transmission among young women, especially black women (Jewkes & Morrell, 2010). Therefore, developing and implementing interventions to addressing sexual violence against women will undeniably be advantageous to addressing the HIV epidemic in the country.

5.2 Sociodemographic and Behavioural Factors- Logistic Regression

Although the logistic regression does not show a string of variables that could be related to increased chances of HIV, the results of individual effects of the variables on the outcome was similar to the decision tree result with a few exceptions. Like the decision tree, history of an STI has a strong positive association with HIV by 95.99%, however, the decision tree breaks it down to show the influence of a combination of variables on HIV serostatus.

The logistic regression model reports that women aged 35-39 years have the highest risk of being HIV positive (Table 4) which is not consistent with other studies that report younger ages as being more at risk of HIV (Dellar et al., 2015). However the skewed data between the proportion of HIV by the age group could have influenced this result which the decision tree does well in accounting for (Song & Lu, 2015). In addition, living in KwaZulu-Natal, the

Eastern Cape, and Mpumalanga significantly increased the chances of being HIV positive, while the rest were insignificant. Living in Limpopo significantly lowers the risk of being HIV positive. Which is similar to the decision tree model. In agreement to studies by Munthali and Zulu (2007), Stöckl, Kalra, Jacobi, andWatts (2013), and Wand and Ramjee (2012) who found that girls who first had sex before they turned 18 years were significantly more likely to be HIV positive, the current study showed that age of debut above the age of 18 decreased the likelihood of being HIV positive as opposed to age of debut below the age of 18 which increases HIV risk. And this result was statistically significant.

Having an education above tertiary level was significant in reducing the probability of being HIV positive although this variable did not appear on the decision tree. This variable would have been further understood using the decision tree as it might state what other factors such as age and residence type work in conjunction with education level to influence the risk of HIV. A study found that in rural areas, people with lowers levels of formal education had greater probability of being HIV positive (Gómez-Olivé et al., 2013). Furthermore, the current study found that living in urban areas had a negative and significant association with HIV as opposed to living in rural areas. This means that the probability and risk of being HIV positive in an urban area is less. It is postulated that these variables (education and residence type) might be colinear although the analysis does not show multicollinearity. It is explained that there are more educated people in urban areas, who have been found to also use condoms more consistently and who might have more knowledge of HIV/AIDS (Gómez-Olivé et al., 2013) and are less likely to engage in commercial sex work (Boyer et al., 2017; Mah & Halperin, 2010b). However, although these are hypotheses that need to be further investigated.

5.3 Performance Comparison of the models in predicting HIV status

This research was aimed at predicting the HIV status of women in South Africa using a decision tree and comparing the result with logistic regression. The respective prediction models were developed to predict HIV status. Furthermore, SEN, SPE, PPR, NPR, and Accuracy was used to evaluate the performance of the models. The results from the test dataset indicates that the accuracy for the logistic regression model was 68% and 64% for the decision tree model. Contrary to the expected results, the logistic regression was better at predicting HIV status than decision tree model. This means that out of 100 women, 68% of them would be correctly assigned to their HIV status category using logistic regression, and 64% of them using decision tree.

The results of this study are similar to findings from a few studies who compared decision tree model against logistic regression. Study by Long, Griffith, Selker, and D'agostino (1993) who compared the performance of decision tree and logistic regression in classifying patients with acute cardiac ischemia found that the logistic regression performed better than the decision tree. They stated that this was due to overfitting which they accounted for by pruning, but this also led to the data set not capturing the true statistical nature of the data appropriately and overspecification.

The decision tree method is a powerful statistical tool for classification, prediction and data mining. It simplifies complex relationships making it easy to understand and interpret, it also handles missing and skewed data well and is robust to outliers (Song & Lu, 2015). Logistic regression always makes a parametric assumption of the log-odds transformation, regardless of the predictors in the model. Given that logistic regression is often used in the social sciences, one would expect that it performs very well in its naive/ simple implementations, rather than in its most sophisticated implementations (Westreich, Lessler, & Funk, 2010). Machine learning techniques on the other hand make fewer assumptions than logistic regression and deals with interactions and nonlinearities, in their naive implementations. Therefore, decision trees and other machine learning techniques such as boosting might find recommendation to replace logistic regression as the presumptive mechanism for prediction or classification.

In predicting the presence of coronary artery disease, Kurt, Ture, and Kurum (2008) found that the logistic regression performed slightly better than the decision tree. However, they only compared the statistical significance using the ROC curve. They however stated that the decision tree provides a more comprehensive analytical framework that would be optimal for clinical guidelines and health policy for the prevention and management of cardiovascular diseases. Another study which showed that the regression model performed better than the decision tree is a study by Nie, Rowe, Zhang, Tian, and Shi (2011). The aim of their study was to develop a churn prediction model using credit card data. They concluded that a decision tree model is better because it provides simple and easy to understand rules for making decisions.

Contrary to the findings of this study and other studies highlighted above, a study by Hong et al. (2015) that predicted landslide hazard in China found that the decision tree performed better in terms of prediction capabilities. Similarly, De Queiroz Mello et al. (2006) found that the decision tree was better in predicting smear negative pulmonary tuberculosis.

5.4 Limitations

As with all analytic methods, there are also limitations of the decision tree method that users must be aware of. The main disadvantage is that it can be subject to overfitting and underfitting, particularly when using a small data set. This problem can limit the generalizability and robustness of the resultant models (Kotsiantis, 2013). However, there was no overfitting because the results of the training dataset and the test dataset were close. Logistic regression requires that each variable is independent otherwise the model will tend to overweigh the significance of those observations (Ranganathan, Pramesh, & Aggarwal, 2017). Another potential problem is that strong correlation between different potential input variables may result in the selection of variables that improve the model statistics but are not causally related to the outcome of interest (Ranganathan et al., 2017). Thus, one must be cautious when interpreting the models and when using the results of these models to develop causal hypotheses.

In general, both methods are supervised learning and require that independent variables be selected prior to building the model. Due to the secondary data analysis from a survey, which limits the variety of variables available, the models do not have a high accuracy percentage (Kotsiantis, 2013). A final limitation to be considered is that only approximately 20% of the participants were HIV positive. Therefore, the data was highly skewed, and this influenced the logistic regression results.

CHAPTER SIX: CONCLUSION

The use of data to make decisions has been utilized by different research areas using different approaches. Logistic regression is a major method used in the statistics community for predicting outcomes. In the machine learning community, decision tree is widely recognised. Logistic regression and decision tree are used for similar tasks, however, the former is better for determining probability of an outcome being present in an individual, while the latter is used for deciding for example who gets treatment based on features used to determine an outcome. In this study, supervised learning was used to train HIV data to predict HIV Status among women in South Africa with the use of decision trees.

The HIV and AIDS epidemic has grown immensely and is an important public health problem. Sub-Saharan region of Arica remains the most severely affected with South Africa having the most prevalence of HIV. However, women carry a disproportioned burden of the disease and are regarded as high at-risk population. To effectively manage this epidemic, reliable and accurate information on the prevalence, better understanding of the sociodemographic factors that influence the risk of seroconversion, and an understanding of interventions required to address the socio-demographic factors and improve behaviour and treatment interventions are needed. The current study endeavoured to achieve this by predicting the HIV status of an individual, given readily available demographic data especially in low resource health centres in order to effectively manage and deliver HIV services.

The results were also compared to the logistic regression model and an extreme gradient boosting model. Decision tree and the extreme gradient boosting models are machine learning techniques and according to current literature have not been used for HIV serostatus prediction. The results show that logistic regression results and that of the decision tree produce similar results in terms of understanding variable association, but the logistic regression has significantly better accuracy and a larger area under the ROC curve.

The accuracy for the validation dataset was 68% and 64% for the logistic regression and decision tree models respectively, which is not a high enough accuracy level. This proves that demographic data might not be enough to accurately predict HIV status especially at the medical classification level, or more variables are needed to build the model. It is also recommended that different input features be tested, as well as automatic relevance detection to assess which inputs contribute to the output. By observing these features, it would be easier to find additional relevant input features.

Furthermore, it was expected that the decision tree would do better in terms of predictive accuracy than the logistic accuracy. Although the data was balanced using the SMOTE method, decision trees do not handle imbalanced data classes well and that could be a reason why it did not outperform the logistic regression. In addition, there could have been undetected noise in the dataset that might have caused overfitting of the decision tree model. The model was pruned and cross validated, and it was confirmed that there was no overfitting because the accuracy between the testing and training data were similar, with no huge difference. Otherwise, the decision tree is a very powerful tool that works well with small and big datasets and does not rely on prior knowledge of the variables or any pre-assumed parameter. The results of the decision tree can be improved considerably using techniques such as random forests and boosting. In conclusion, no algorithm is generally better than the other, they have their different purposes, advantages and disadvantages. The logistic regression provides better understanding of the effect of the predictor variable on the response variable, which the decision tree does not offer. However, the decision tree is easy to implement and understand even for non-statisticians.

CHAPTER SEVEN: STUDY IMPLICATIONS AND RECOMMENDATIONS

7.1 Policy and Programme Implications

The result of this study points to areas that health promotion interventions needs to address in tackling the HIV epidemic especially in reaching the 90-90-90 goal by 2020. Studies have shown that utilizing current technologies and methods will enable the achievement of the goal without the need for a cure or vaccine (Harries et al., 2016). HIV testing has come a long way and at this point, it is rapid and easy to use allowing for the decentralisation of task and wide coverage. Diagnosing the millions of people who do not know that they have HIV and starting and retaining those people on ART are daunting prospects and will require increased focus and efficiency including the scaling up of human resources, health systems, innovations in service delivery including community- and patient-driven initiatives, and additional financial resources. This includes utilizing machine learning methods to assist in HIV diagnosis.

The results of this study suggest that STI history is a major factor in determining HIV status. Hence, controlling and preventing STI's should not be neglected or treated in isolation when addressing HIV issues. It is recommended that South Africa strengthens STI programmes in conjunction with interventions aimed at achieving the 90-90-90 goal. This is because STI is strongly linked to HIV seroconversion therefore, it would be futile in isolating. Secondly, a primary target population for health interventions should be adolescent and young adult females. These are highly at-risk and susceptible population groups as highlighted above. Furthermore, preventing HIV transmission in the young generation will reduce future prevalence of the disease.

7.2 Implications for Further Research

These results should also be considered when considering which statistical analysis method should be used in epidemiological/social science research. Future studies should explore a wider range of social science and epidemiological application of machine learning. The decision tree is also considered to be another promising technique. In order to verify the overall performance of both methods, more studies should be conducted with more important variable selection to achieve higher predictive accuracy. Boosting should be the focus as they are available for us on different statistical computing platforms and are resistant to overfitting. However, for those unfamiliar with machine learning methods such as decision tree, it is advised that they start out with logistic regression because it is a simple model and will help in better understanding their specific datasets. But with better categorized data, the decision tree model would do better than the logistic regression.

REFERENCES

(SANAC), S. A. N. A. C. (2017). *LET OUR ACTIONS COUNT SOUTH AFRICA'S NATIONAL STRATEGIC PLAN ON Summary*. Retrieved from www.sanac.org.za

Ackermann, L., & Klerk, G. W. de. (2002). SOCIAL FACTORS THAT MAKE SOUTH AFRICAN WOMEN VULNERABLE TO HIV INFECTION. *Health Care for Women International*, *23*(2), 163–172. https://doi.org/10.1080/073993302753429031

Alemayehu, M., Belachew, T., & Tilahun, T. (2012). Factors associated with utilization of long acting and permanent contraceptive methods among married women of reproductive age in Mekelle town, Tigray region, north Ethiopia. *BMC Pregnancy and Childbirth*, *12*(1), 6. https://doi.org/10.1186/1471-2393-12-6

Ambrosioni, J., Calmy, A., & Hirschel, B. (2011). HIV treatment for prevention. *Journal of the International AIDS Society*, *14*, 28. https://doi.org/10.1186/1758-2652-14-28

Andoh, S. Y., Umezaki, M., Nakamura, K., Kizuki, M., & Takano, T. (2006). Correlation between national income, HIV/AIDS and political status and mortalities in African countries. *Public Health*, *120*(7), 624–633. https://doi.org/10.1016/j.puhe.2006.04.008

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79. https://doi.org/10.1214/09-SS054

Ayisi, J. G., van Eijk, A. M., ter Kuile, F. O., Kolczak, M. S., Otieno, J. A., Misore, A. O., … Nahlen, B. L. (2000). Risk factors for HIV infection among asymptomatic pregnant women attending an antenatal clinic in western Kenya. *International Journal of STD & AIDS*, *11*(6), 393–401. https://doi.org/10.1258/0956462001916119

Ayodele, O., & Ayodele, O. M. (2016). Urban-Rural Differentials in HIV/AIDS Knowledge of Nigerian Senior Secondary School Students. *International Journal of Health Sciences*, *4*(3), 2372–5079. https://doi.org/10.15640/ijhs.v4n3a6

Baral, S., Beyrer, C., Muessig, K., Poteat, T., Wirtz, A. L., Decker, M. R., … Kerrigan, D. (2012). Burden of HIV among female sex workers in low-income and middle-income countries: a systematic review and meta-analysis. *The Lancet Infectious Diseases*, *12*(7), 538–549. https://doi.org/10.1016/S1473-3099(12)70066-X

Baral, S., Logie, C. H., Grosso, A., Wirtz, A. L., & Beyrer, C. (2013). Modified social ecological model: a tool to guide the assessment of the risks and risk contexts of HIV epidemics. *BMC Public Health*, *13*, 482. https://doi.org/10.1186/1471-2458-13-482

Barber, B. K., & Schluterman, J. M. (2008). Connectedness in the Lives of Children and Adolescents: A Call for Greater Conceptual Clarity. *Journal of Adolescent Health*, *43*(3), 209–216. https://doi.org/10.1016/J.JADOHEALTH.2008.01.012

Bärnighausen, T., Bloom, D. E., & Humair, S. (2007). Human Resources for Treating HIV/AIDS: Needs, Capacities, and Gaps. *AIDS Patient Care and STDs*, *21*(11), 799–812. https://doi.org/10.1089/apc.2007.0193

Bernstein, K. T., Marcus, J. L., Nieri, G., Philip, S. S., & Klausner, J. D. (2010). Rectal gonorrhea and chlamydia reinfection is associated with increased risk of HIV seroconversion. *Journal of Acquired Immune Deficiency Syndromes*. https://doi.org/10.1097/QAI.0b013e3181c3ef29

Boyer, C. B., Greenberg, L., Chutuape, K., Walker, B., Monte, D., Kirk, J., … Munoz, M. (2017). Exchange of Sex for Drugs or Money in Adolescents and Young Adults: An Examination of Sociodemographic Factors, HIV-Related Risk, and Community Context. *Journal of Community Health*. https://doi.org/10.1007/s10900-016-0234-2

Camlin, C. S., Hosegood, V., Newell, M. L., Mcgrath, N., Bärnighausen, T., & Snow, R. C. (2010). Gender, migration and HIV in rural Kwazulu-Natal, South Africa. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0011539

Campbell, J. C., Baty, M. L., Ghandour, R. M., Stockman, J. K., Francisco, L., & Wagman, J. (2008). The intersection of intimate partner violence against women and HIV/AIDS: a review. *International Journal of Injury Control and Safety Promotion*, *15*(4), 221–231. https://doi.org/10.1080/17457300802423224

Chirinda, W., & Peltzer, K. (2014). Correlates of inconsistent condom use among youth aged 18-24 years in South Africa. *Journal of Child and Adolescent Mental Health*. https://doi.org/10.2989/17280583.2013.877912

Coffee, M. P., Garnett, G. P., Mlilo, M., Voeten, H. A. C. M., Chandiwana, S., & Gregson, S. (2005). Patterns of Movement and Risk of HIV Infection in Rural Zimbabwe. *The Journal of Infectious Diseases*, *191*(s1), S159–S167. https://doi.org/10.1086/425270

Corno, L., & de Walque, D. (2012). Mines, migration and HIV/AIDS in Southern Africa. *Journal of African Economies*. https://doi.org/10.1093/jae/ejs005

Crosby, R. A. (2013). State of condom use in HIV prevention science and practice. *Current HIV/AIDS Reports*. https://doi.org/10.1007/s11904-012-0143-7

Crush, J., Grant, M., & Frayne, B. (2007). No. 3: Linking Migration, HIV/AIDS and Urban Food Security in Southern and Eastern Africa. *Southern African Migration Programme*. Retrieved from https://scholars.wlu.ca/samp/15

Davidoff-Gore, A., Luke, N., & Wawire, S. (2011). Dimensions of poverty and inconsistent condom use among youth in urban Kenya. *AIDS Care - Psychological and Socio-Medical Aspects of AIDS/HIV*. https://doi.org/10.1080/09540121.2011.555744

De Queiroz Mello, F. C., Do Valle Bastos, L. G., Machado Soares, S. L., Rezende, V. M. C., Conde, M. B., Chaisson, R. E., … Werneck, G. L. (2006). Predicting smear negative pulmonary tuberculosis with classification trees and logistic regression: A cross-sectional study. *BMC Public Health*. https://doi.org/10.1186/1471-2458-6-43

Deane, K. D., Parkhurst, J. O., & Johnston, D. (2010). Linking migration, mobility and HIV. *Tropical Medicine & International Health*, *15*(12), 1458–1463. https://doi.org/10.1111/j.1365-3156.2010.02647.x

Deeks, S. G., Lewin, S. R., & Havlir, D. V. (2013, November 2). The end of AIDS: HIV infection as a chronic disease. *The Lancet*, Vol. 382, pp. 1525–1533. https://doi.org/10.1016/S0140-6736(13)61809-7

Degenhardt, L., Mathers, B., Vickerman, P., Rhodes, T., Latkin, C., & Hickman, M. (2010). Prevention of HIV infection for people who inject drugs: why individual, structural, and combination approaches are needed. *The Lancet*, *376*(9737), 285–301. https://doi.org/10.1016/S0140-6736(10)60742-8

del Rio, C. (2017). The global HIV epidemic: What the pathologist needs to know. *Seminars in Diagnostic Pathology*. https://doi.org/10.1053/j.semdp.2017.05.001

Dellar, R. C., Dlamini, S., & Karim, Q. A. (2015). Adolescent girls and young women: Key populations for HIV epidemic control. *Journal of the International AIDS Society*. https://doi.org/10.7448/IAS.18.2.19408

Demographic and Health Survey. (2016). The DHS Program - Data. Retrieved February 26, 2019, from https://www.dhsprogram.com/Data/

Diener-West, M. (n.d.). *Use of the Chi-Square Statistic*. Retrieved from http://ocw.jhsph.edu/courses/fundepiii/PDFs/Lecture17.pdf

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87. https://doi.org/10.1145/2347736.2347755

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics*. https://doi.org/10.1016/S1532-0464(03)00034-0

Fallis, A. . (2013). SOUTH AFRICAN NATIONAL HEALTH AND NUTRITION SURVEY 2012. In *Journal of Chemical Information and Modeling* (Vol. 53). https://doi.org/10.1017/CBO9781107415324.004

Fettig, J., Swaminathan, M., Murrill, C. S., & Kaplan, J. E. (2014). Global epidemiology of HIV. *Infectious Disease Clinics of North America*. https://doi.org/10.1016/j.idc.2014.05.001

Gabrysch, S., Edwards, T., Glynn, J. R., & Study Group on Heterogeneity of HIV Epidemics in African Cities. (2008). The role of context: neighbourhood characteristics strongly influence HIV risk in young women in Ndola, Zambia. *Tropical Medicine & International Health*, *13*(2), 162–170. https://doi.org/10.1111/j.1365-3156.2007.01986.x

Gerbi, G. B., Habtemariam, T., Robnett, V., Nganwa, D., & Tameru, B. (2012). Psychosocial factors as predictors of HIV/AIDS risky behaviors among people living with HIV/AIDS. *Journal of AIDS and HIV Research (Online)*, *4*(1), 8–16. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22374351

Gómez-Olivé, F. X., Angotti, N., Houle, B., Klipstein-Grobusch, K., Kabudula, C., Menken, J., … Clark, S. J. (2013). Prevalence of HIV among those 15 and older in rural South Africa. *AIDS Care - Psychological and Socio-Medical Aspects of AIDS/HIV*. https://doi.org/10.1080/09540121.2012.750710

Gregson, S., Nyamukapa, C. A., Garnett, G. P., Wambe, M., Lewis, J. J. C., Mason, P. R., … Anderson, R. M. (2005). HIV infection and reproductive health in teenage women orphaned and made vulnerable by AIDS in Zimbabwe. *AIDS Care*, *17*(7), 785–794. https://doi.org/10.1080/09540120500258029

Grimmer, J. (2015). We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together. *PS: Political Science & Politics*, *48*(01), 80–83. https://doi.org/10.1017/S1049096514001784

Haddad, L. B., Polis, C. B., Sheth, A. N., Brown, J., Kourtis, A. P., King, C., … Ofotokun, I. (2014). Contraceptive Methods and Risk of HIV Acquisition or Female-to-Male Transmission. *Current HIV/AIDS Reports*. https://doi.org/10.1007/s11904-014-0236-6

Harries, A. D., Suthar, A. B., Takarinda, K. C., Tweya, H., Kyaw, N. T. T., Tayler-Smith, K., & Zachariah, R. (2016). Ending the HIV/AIDS epidemic in low- and middle-income countries by 2030: is it possible? *F1000Research*. https://doi.org/10.12688/f1000research.9247.1

Harrison, A., Newell, M. L., Imrie, J., & Hoddinott, G. (2010). HIV prevention for South African youth: Which interventions work? A systematic review of current evidence. *BMC Public Health*. https://doi.org/10.1186/1471-2458-10-102

Hassig, S. E., Kinkela, N., Nsa, W., Kamenga, M., Ndilu, M., Francis, H., & Ryder, R. W. (1990). Prevention of perinatal HIV transmission: are there alternatives to pre-pregnancy serological screening in Kinshasa, Zaire? *AIDS (London, England)*, *4*(9), 913–916. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2252564

Hecht, R., Bollinger, L., Stover, J., McGreevey, W., Muhib, F., Madavo, C. E., & de Ferranti, D. (2009). Critical Choices In Financing The Response To The Global HIV/AIDS Pandemic. *Health Affairs*, *28*(6), 1591–1605. https://doi.org/10.1377/hlthaff.28.6.1591

Hong, H., Pradhan, B., Xu, C., & Tien Bui, D. (2015). Spatial prediction of landslide hazard at the Yihuang area (China) using two-class kernel logistic regression, alternating decision tree and support vector machines. *Catena*. https://doi.org/10.1016/j.catena.2015.05.019

Hosegood, V. (2009). The demographic impact of HIV and AIDS across the family and household life-cycle: implications for efforts to strengthen families in sub-Saharan Africa. *AIDS Care*, *21 Suppl 1*(Suppl 1), 13–21. https://doi.org/10.1080/09540120902923063

Jewkes, R. K., Dunkle, K., Nduna, M., & Shai, N. (2010). Intimate partner violence, relationship power inequity, and incidence of HIV infection in young women in South Africa: A cohort study. *The Lancet*. https://doi.org/10.1016/S0140-6736(10)60548-X

Jewkes, R., & Morrell, R. (2010). Gender and sexuality: emerging perspectives from the heterosexual epidemic in South Africa and implications for HIV risk and prevention. *Journal of the International AIDS Society*, *13*(1), 6–6. https://doi.org/10.1186/1758-2652-13-6

Johnston, L., O'Bra, H., Chopra, M., Mathews, C., Townsend, L., Sabin, K., … Kendall, C. (2010). The associations of voluntary counseling and testing acceptance and the perceived likelihood of being HIV-infected among men with multiple sex partners in a

South African Township. *AIDS and Behavior*. https://doi.org/10.1007/s10461-008-9362-8

Kalichman, S. C., Pellowski, J., & Turner, C. (2011). Prevalence of sexually transmitted co-infections in people living with HIV/AIDS: systematic review with implications for using HIV treatments for prevention. *Sexually Transmitted Infections*, *87*(3), 183–190. https://doi.org/10.1136/sti.2010.047514

Kalichman, Seth C., Ntseane, D., Nthomang, K., Segwabe, M., Phorano, O., & Simbayi, L. C. (2007). Recent multiple sexual partners and HIV transmission risks among people living with HIV/AIDS in Botswana. *Sexually Transmitted Infections*. https://doi.org/10.1136/sti.2006.023630

Kalichman, Seth C., Simbayi, L. C., Vermaak, R., Jooste, S., & Cain, D. (2008). HIV/AIDS Risks among Men and Women Who Drink at Informal Alcohol Serving Establishments (Shebeens) in Cape Town, South Africa. *Prevention Science*, *9*(1), 55–62. https://doi.org/10.1007/s11121-008-0085-x

Kaufman, M. R., Cornish, F., Zimmerman, R. S., & Johnson, B. T. (2014). Health behavior change models for HIV prevention and AIDS care: practical recommendations for a multi-level approach. *Journal of Acquired Immune Deficiency Syndromes (1999)*, *66 Suppl 3*(Suppl 3), S250-8. https://doi.org/10.1097/QAI.0000000000000236

Kharsany, A. B. M., & Karim, Q. A. (2016). HIV Infection and AIDS in Sub-Saharan Africa: Current Status, Challenges and Opportunities. *The Open AIDS Journal*, *10*, 34–48. https://doi.org/10.2174/1874613601610010034

Kotsiantis, S. B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review*, *39*(4), 261–283. https://doi.org/10.1007/s10462-011-9272-4

Kouyoumdjian, F. G., Calzavara, L. M., Bondy, S. J., O'Campo, P., Serwadda, D., Nalugoda, F., … Gray, R. (2013). Intimate partner violence is associated with incident HIV infection in women in Uganda. *AIDS*. https://doi.org/10.1097/QAD.0b013e32835fd851

Kraska, T., Berkeley, U. C., Berkeley, U. C., Griffith, R., Franklin, M. J., Berkeley, U. C., & Berkeley, U. C. (2013). *MLbase : A Distributed Machine-learning System*.

Kurt, I., Ture, M., & Kurum, A. T. (2008a). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2006.09.004

Kurt, I., Ture, M., & Kurum, A. T. (2008b). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Systems with Applications*, *34*(1), 366–374. https://doi.org/10.1016/j.eswa.2006.09.004

Kurth, A. E., Celum, C., Baeten, J. M., Vermund, S. H., & Wasserheit, J. N. (2011). Combination HIV prevention: Significance, challenges, and opportunities. *Current HIV/AIDS Reports*. https://doi.org/10.1007/s11904-010-0063-3

Lallemant, M., Lallemant-Le Coeur, S., Cheynier, D., Nzingoula, S., Jourdain, G., Sinet, M., … Larouzé, B. (1992). Characteristics associated with HIV-1 infection in pregnant women in Brazzaville, Congo. *Journal of Acquired Immune Deficiency Syndromes*, *5*(3), 279–285. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1740754

Lee, C. W., & Park, J. A. (2001). Assessment of HIV/AIDS-related health performance using an artificial neural network. *Information and Management*. https://doi.org/10.1016/S0378-7206(00)00068-9

Long, W. J., Griffith, J. L., Selker, H. P., & D'agostino, R. B. (1993). A comparison of logistic regression to decision-tree induction in a medical domain. *Computers and Biomedical Research*, *26*(1), 74–97. https://doi.org/10.1006/cbmr.1993.1005

MacPhail, C., Pettifor, A., Moyo, W., & Rees, H. (2009). Factors associated with HIV testing among sexually active South African youth aged 15-24 years. *AIDS Care - Psychological and Socio-Medical Aspects of AIDS/HIV*. https://doi.org/10.1080/09540120802282586

Magadi, M. A. (2016). Understanding the urban–rural disparity in HIV and poverty nexus: the case of Kenya. *Journal of Public Health*, *39*(3), e63–e72. https://doi.org/10.1093/pubmed/fdw065

Maganja, R. K., Maman, S., Groves, A., & Mbwambo, J. K. (2007a). Skinning the goat and pulling the load: transactional sex among youth in Dar es Salaam, Tanzania. *AIDS Care*, *19*(8), 974–981. https://doi.org/10.1080/09540120701294286

Maganja, R. K., Maman, S., Groves, A., & Mbwambo, J. K. (2007b). Skinning the goat and pulling the load: transactional sex among youth in Dar es Salaam, Tanzania. *AIDS Care*, *19*(8), 974–981. https://doi.org/10.1080/09540120701294286

Mah, T. L. (2010). Prevalence and correlates of concurrent sexual partnerships among young people in South Africa. *Sexually Transmitted Diseases*. https://doi.org/10.1097/OLQ.0b013e3181bcdf75

Mah, T. L., & Halperin, D. T. (2010a). Concurrent sexual partnerships and the HIV epidemics in Africa: evidence to move forward. *AIDS and Behavior*. https://doi.org/10.1007/s10461-008-9433-x

Mah, T. L., & Halperin, D. T. (2010b). Concurrent sexual partnerships and the HIV epidemics in Africa: Evidence to move forward. *AIDS and Behavior*. https://doi.org/10.1007/s10461-008-9433-x

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0194889

Malani, P. N. (2016, July 12). Human immunodeficiency virus. *JAMA - Journal of the American Medical Association*, Vol. 316, p. 238. https://doi.org/10.1001/jama.2016.7995

Manlove, J. S., Ryan, S., & Franzetta, K. (2007). Risk and Protective Factors Associated with the Transition to a First Sexual Relationship with an Older Partner. *Journal of Adolescent Health*, *40*(2), 135–143. https://doi.org/10.1016/J.JADOHEALTH.2006.09.003

Markham, C. M., Lormand, D., Gloppen, K. M., Peskin, M. F., Flores, B., Low, B., & House, L. D. (2010). Connectedness as a Predictor of Sexual and Reproductive Health Outcomes for Youth. *Journal of Adolescent Health*, *46*(3 SUPPL.), S23–S41. https://doi.org/10.1016/j.jadohealth.2009.11.214

McClelland, R. S., Sangaré, L., Hassan, W. M., Lavreys, L., Mandaliya, K., Kiarie, J., … Baeten, J. M. (2007). Infection with Trichomonas vaginalis Increases the Risk of HIV-1 Acquisition . *The Journal of Infectious Diseases*. https://doi.org/10.1086/511278

McGrath, N., Hosegood, V., Newell, M. L., & Eaton, J. W. (2015). Migration, sexual behaviour, and HIV risk: A general population cohort in rural South Africa. *The Lancet HIV*. https://doi.org/10.1016/S2352-3018(15)00045-4

Mmbaga, E. J., Leyna, G. H., Hussain, A., Mnyika, K. S., Sam, N. E., & Klepp, K.-I. (2008). The role of in-migrants in the increasing rural HIV-1 epidemic: results from a village population survey in the Kilimanjaro region of Tanzania. *International Journal of*

Infectious Diseases, *12*(5), 519–525. https://doi.org/10.1016/j.ijid.2008.02.007

Mondal, M. N. I., & Shitan, M. (2013). Factors affecting the HIV/AIDS epidemic: An ecological analysis of global data. *African Health Sciences*. https://doi.org/10.4314/ahs.v13i2.15

Munthali, A. C., & Zulu, E. M. (2007). The timing and role of initiation rites in preparing young people for adolescence and responsible sexual and reproductive behaviour in Malawi. *African Journal of Reproductive Health*.

Musheke, M., Ntalasha, H., Gari, S., McKenzie, O., Bond, V., Martin-Hilber, A., & Merten, S. (2013). A systematic review of qualitative findings on factors enabling and deterring uptake of HIV testing in Sub-Saharan Africa. *BMC Public Health*. https://doi.org/10.1186/1471-2458-13-220

Muula, A. S. (2008). HIV infection and AIDS among young women in South Africa. *Croatian Medical Journal*, *49*(3), 423–435. https://doi.org/10.3325/CMJ.2008.3.423

Naidoo, J. R., Uys, L. R., Greeff, M., Holzemer, W. L., Makoae, L., Dlamini, P., … Kohi, T. (2009). African countries. *African Journal of AIDS Research*, *6*(1), 17–23. https://doi.org/10.2989/16085900709490395

Nie, G., Rowe, W., Zhang, L., Tian, Y., & Shi, Y. (2011a). Credit card churn forecasting by logistic regression and decision tree. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2011.06.028

Nie, G., Rowe, W., Zhang, L., Tian, Y., & Shi, Y. (2011b). Credit card churn forecasting by logistic regression and decision tree. *Expert Systems with Applications*, *38*(12), 15273–15285. https://doi.org/10.1016/j.eswa.2011.06.028

Palamuleni, E. M. (2011). Socioeconomic determinants of age at marriage in Malawi. *International Journal of Sociology and Anthropology*, *3*(7), 224–235.

Patty, J. W., & Penn, E. M. (2015). Analyzing Big Data: Social Choice and Measurement. *PS: Political Science & Politics*, *48*(01), 95–101. https://doi.org/10.1017/S1049096514001814

Peltzer, K., Matseke, G., Mzolo, T., & Majaja, M. (2009). Determinants of knowledge of HIV status in South Africa: Results from a population-based HIV survey. *BMC Public Health*. https://doi.org/10.1186/1471-2458-9-174

Pettifor, A., Bekker, L. G., Hosek, S., DiClemente, R., Rosenberg, M., Bull, S. S., … Cowan,

F. (2013). Preventing HIV among young people: Research priorities for the future. *Journal of Acquired Immune Deficiency Syndromes*. https://doi.org/10.1097/QAI.0b013e31829871fb

Poundstone, K. E., Strathdee, S. A., & Celentano, D. D. (2004). The Social Epidemiology of Human Immunodeficiency Virus/Acquired Immunodeficiency Syndrome. *Epidemiologic Reviews*, *26*(1), 22–35. https://doi.org/10.1093/epirev/mxh005

Prat, F., Planes, M., Gras, M. E., & Sullman, M. J. M. (2016). Perceived Pros and Cons of Condom Use as Predictors of its Consistent Use with a Heterosexual Romantic Partner Among Young Adults. *Current Psychology*, *35*(1), 13–21. https://doi.org/10.1007/s12144-015-9357-3

Ramjee, G., & Daniels, B. (2013). Women and HIV in Sub-Saharan Africa. *AIDS Research and Therapy*, *10*(1), 30. https://doi.org/10.1186/1742-6405-10-30

Ranganathan, P., Pramesh, C. S., & Aggarwal, R. (2017). Common pitfalls in statistical analysis: Logistic regression. *Perspectives in Clinical Research*. https://doi.org/10.4103/picr.PICR_87_17

Regnerus, M. D., & Luchies, L. B. (2006). The Parent-Child Relationship and Opportunities for Adolescents' First Sex. *Journal of Family Issues*, *27*(2), 159–183. https://doi.org/10.1177/0192513X05281858

Reitermanov, Z. (2010). *Data Splitting*. 31–36.

Rodrigo, C., & Rajapakse, S. (2010). HIV, poverty and women. *International Health*. https://doi.org/10.1016/j.inhe.2009.12.003

SAMBISA, W., CURTIS, S. L., & STOKES, C. S. (2010). ETHNIC DIFFERENCES IN SEXUAL BEHAVIOUR AMONG UNMARRIED ADOLESCENTS AND YOUNG ADULTS IN ZIMBABWE. *Journal of Biosocial Science*, *42*(01), 1. https://doi.org/10.1017/S0021932009990277

Seitz, R. (2016). Human Immunodeficiency Virus (HIV). *Transfusion Medicine and Hemotherapy*, *43*(3), 203–222. https://doi.org/10.1159/000445852

Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. *Procedia Computer Science*. https://doi.org/10.1016/j.procs.2015.12.157

Shisana, O., Labadarios, D., Rehle, T., Simbayi, L., Zuma, K., Dhansay, A., … the

SANHANES-1team. (2014). *the south african national health and nutrition survey*. Cape Town: HSRC Press.

Song, Y. Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*. https://doi.org/10.11919/j.issn.1002-0829.215044

South African Department of Health. (2016). *South African Demographic Health Survey*.

Stöckl, H., Kalra, N., Jacobi, J., & Watts, C. (2013). Is Early Sexual Debut a Risk Factor for HIV Infection Among Women in Sub-Saharan Africa? A Systematic Review. *American Journal of Reproductive Immunology*. https://doi.org/10.1111/aji.12043

Stoebenau, K., Heise, L., Wamoyi, J., & Bobrova, N. (2016). Revisiting the understanding of "transactional sex" in sub-Saharan Africa: A review and synthesis of the literature. *Social Science & Medicine*, *168*, 186–197. https://doi.org/10.1016/J.SOCSCIMED.2016.09.023

Stoebenau, K., Nixon, S. A., Rubincam, C., Willan, S., Zembe, Y. Z., Tsikoane, T., … Razafintsalama, V. (2011). More than just talk: the framing of transactional sex and its implications for vulnerability to HIV in Lesotho, Madagascar and South Africa. *Globalization and Health*, *7*(1), 34. https://doi.org/10.1186/1744-8603-7-34

Thurman, T. R., Brown, L., Richter, L., Maharaj, P., & Magnani, R. (2006). Sexual Risk Behavior among South African Adolescents: Is Orphan Status a Factor? *AIDS and Behavior*, *10*(6), 627–635. https://doi.org/10.1007/s10461-006-9104-8

UNAIDS. (n.d.-a). 90-90-90: treatment for all | UNAIDS. Retrieved April 8, 2019, from http://www.unaids.org/en/resources/909090

UNAIDS. (n.d.-b). Ending AIDS: progress towards the 90–90–90 targets | UNAIDS. Retrieved February 26, 2019, from http://www.unaids.org/en/resources/documents/2017/20170720_Global_AIDS_update_2017

UNAIDS. (2012). *Report on the Global AIDS Epidemic.* Retrieved from http://www.unaids.org/en/resources/documents/2012/20121120_UNAIDS_Global_Report_2012

UNAIDS. (2018). AIDSinfo | UNAIDS. Retrieved February 26, 2019, from http://aidsinfo.unaids.org/

UNAIDS. (2019a). *2018 GLOBAL HIV STATISTICS*.

UNAIDS. (2019b). New modelling research shows partial progress in South Africa's response to HIV. Retrieved July 22, 2019, from https://www.unaids.org/en/resources/presscentre/featurestories/2019/june/20190628_south-africa-modelling

van Loggerenberg, F., Dieter, A. A., Sobieszczyk, M. E., Werner, L., Grobler, A., & Mlisana, K. (2012). HIV prevention in high-risk women in South Africa: Condom use and the need for change. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0030669

Wand, H., & Ramjee, G. (2012a). Assessing and evaluating the combined impact of behavioural and biological risk factors for HIV seroconversion in a cohort of South African women. *AIDS Care*, *24*(9), 1155–1162. https://doi.org/10.1080/09540121.2012.687820

Wand, H., & Ramjee, G. (2012b). The relationship between age of coital debut and HIV seroprevalence among women in Durban, South Africa: A cohort study. *BMJ Open*. https://doi.org/10.1136/bmjopen-2011-000285

Wang, K., Brown, K., Shen, S.-Y., & Tucker, J. (2011). Social Network-Based Interventions to Promote Condom Use: A Systematic Review. *AIDS and Behavior*, *15*(7), 1298–1308. https://doi.org/10.1007/s10461-011-0020-1

Ward, H., & Rönn, M. (2010). Contribution of sexually transmitted infections to the sexual transmission of HIV. *Current Opinion in HIV and AIDS*. https://doi.org/10.1097/COH.0b013e32833a8844

Waziri, S. I., Mohamed Nor, N., Raja Abdullah, N. M., & Adamu, P. (2015). Effect of the Prevalence of HIV/AIDS and the Life Expectancy Rate on Economic Growth in SSA Countries: Difference GMM Approach. *Global Journal of Health Science*, *8*(4), 212–220. https://doi.org/10.5539/gjhs.v8n4p212

Weine, S. M., & Kashuba, A. B. (2012). Labor Migration and HIV Risk: A Systematic Review of the Literature. *AIDS and Behavior*, *16*(6), 1605–1621. https://doi.org/10.1007/s10461-012-0183-4

Weissman, S., Duffus, W. A., Iyer, M., Chakraborty, H., Samantapudi, A. V., & Albrecht, H. (2015). Rural–Urban Differences in HIV Viral Loads and Progression to AIDS among New HIV Cases. *Southern Medical Journal*, *108*(3), 180–188.

https://doi.org/10.14423/SMJ.0000000000000255

Wellings, K., Collumbien, M., Slaymaker, E., Singh, S., Hodges, Z., Patel, D., & Bajos, N. (2006). Sexual behaviour in context: a global perspective. *The Lancet*, *368*(9548), 1706–1728. https://doi.org/10.1016/S0140-6736(06)69479-8

Westreich, D., Lessler, J., & Funk, M. J. (2010). Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology*. https://doi.org/10.1016/j.jclinepi.2009.11.020

WHO. (2015). WHO | World AIDS Day: Business Unusual: Time to end the AIDS epidemic. *WHO*. Retrieved from https://www.who.int/woman_child_accountability/ierg/news/ierg_statement_AIDS_1_december_2014/en/

WHO. (2018a). HIV/AIDS Key Facts. Retrieved February 26, 2019, from https://www.who.int/news-room/fact-sheets/detail/hiv-aids

WHO. (2018b). *HIV Report 2017*.

WHO. (2018c). WHO | Prevention of mother-to-child transmission (PMTCT). *WHO*.

Williams, B. G., Gouws, E., Somse, P., Mmelesi, M., Lwamba, C., Chikoko, T., … Gboun, M. (2015). Epidemiological Trends for HIV in Southern Africa: Implications for Reaching the Elimination Targets. *Current HIV/AIDS Reports*, *12*(2), 196–206. https://doi.org/10.1007/s11904-015-0264-x

Wong, V. J., Murray, K. R., Phelps, B. R., Vermund, S. H., & McCarraher, D. R. (2017). Adolescents, young people, and the 90–90–90 goals. *AIDS*. https://doi.org/10.1097/qad.0000000000001539

world health organisation. (2014). *Global and regional estimates of violence against women*. World Health Organization.

world health organisation. (2017). Violence against women. Retrieved from https://www.who.int/news-room/fact-sheets/detail/violence-against-women

Yehadji, D. (2010). *Urban-rural disparities in HIV related knowledge, behavior and attitude in Burkina Faso: Evidence from Burkina Faso Demographic and Health Survey 2010*. Retrieved from https://scholarworks.gsu.edu/iph_theses/390

Zuma, K., Shisana, O., Rehle, T. M., Simbayi, L. C., Jooste, S., Zungu, N., … Abdullah, F.

(2016). New insights into HIV epidemic in South Africa: key findings from the National HIV Prevalence, Incidence and Behaviour Survey, 2012. *African Journal of AIDS Research*, *15*(1), 67–75. https://doi.org/10.2989/16085906.2016.1153491