

**APPLICATION OF SUPPORT VECTOR MACHINES IN IDENTIFYING FACTORS RELATED TO CD4 CELL COUNT LEVELS AMONG HIV PATIENTS**

**A RESEARCH PROJECT SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF BACHELORE OF SCIENCE IN APPLIED STATISTICS WITH COMPUTING IN THE DEPARTMENT OF MATHEMATICS, PHYSICS AND COMPUTING**

**FEBRUARY, 2023**

# **DECLARATION**

This Proposal is our original work and has not been presented for a Degree in any other University.

NAME REG. NO. SIGN DATE

TONY N. APINDI AST/08/19 \_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_

# **SUPERVISORS DECLARATION**

This proposal has been submitted with my/our approval as University supervisor(s).

1. Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Name of the Supervisor: Dr. Gregory Kerich

2. Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Name of the supervisor: Mr. Dennis Mwan

# **ABBREVIATIONS AND ACRONYMS**

**AIDS:** Acquired Immunodeficiency Syndrome

**ART:** Antiretroviral Therapy

**COVID-19:** Corona Virus Disease – 2019

**HIV:** Human Immunodeficiency Virus

**HIV-ASES:** HIV Treatment Adherence Self-Efficacy Scale

**KENPHIA:** Kenya Population-based HIV Impact Assessment

**PHIA:** Population-based HIV Impact Assessment

**PMTCT:** Preventation of Mother to Child Transmission

**TB:** Tuberculosis

**UN:** United Nations

**UNAIDS:** Joint United Nations Programme on HIV/AIDS

**WHO:** World Health Organization

# **ABSTRACT**

The spread of HIV/AIDS in Kenya has ravished many communities over many decades without a vaccine being found. Many researchers have found out that in Kenya, most of the affected people are adolescent girls and women who form a larger percentage of people who are HIV positive as compared to males. Therefore, this study seeks to find out which factors affect the CD4 cell count levels in females by using support vector machine (SVM). The general objective of the study is to apply machine learning models in the identification of factors associated with CD4 cell count levels. The specific objectives include investigating demographic and socio-economic factors related to CD4 cell count levels to fit a support vector machine learning model and to investigate the factors affecting the CD4 cell levels on HIV+ women in Kenya. The results were arrived at by using SVM to identify factors related to CD4 cell count levels. The data is the latest data set of HIV positive women in Kenya extracted from Kenya’s PHIA. From the analysis, 90% of individuals who did not enroll in school have a low cd4 count while 10% have a high CD4 cell count. Furthermore, 90% of respondents related to the head had a low CD4 count, and 9 % of heads of families had a high CD4 count. 50% of the whole relationship with the head of the family was a relative who had a high CD4 count. The performance evaluation revealed that the SVM model had an accuracy of 0.949, which indicates that the model was able to correctly classify 94.9% of the test data. The precision for the CD4 category 1 and 2 were 0.992 and 0.143 respectively. The recall for the CD4 category 1 and 2 were 0.956 and 0.5 respectively. The F1 score for the CD4 category 1 and 2 were 0.974 and 0.222 respectively. The significant relationships were Relationship With Head, Sick to Work last 3 Months, Ever Attended School, Ever Enrolled in School, Work for Pay, Married/Live Together, Number of Pregnancies, Pregnant Currently, Ever Avoided Pregnancy, Ever Sought TB Treatment, Duration on ART and ARVs Detected. The recommendations for this research include healthcare providers and policymakers prioritize the education of individuals on the significance of enrolling in school and obtaining employment, particularly those living with HIV. Additionally, that healthcare providers offer comprehensive HIV management programs that focus on the social and economic factors that affect CD4 cell count levels. Finally, further studies should be conducted to explore the role of other social and economic factors on CD4 cell count levels.

# **TABLE OF CONTENTS**

[DECLARATION i](#_Toc133502975)

[SUPERVISORS DECLARATION iii](#_Toc133502976)

[ABBREVIATIONS AND ACRONYMS iv](#_Toc133502977)

[ABSTRACT v](#_Toc133502978)

[TABLE OF CONTENTS vi](#_Toc133502979)

[CHAPTER ONE: INTRODUCTION 1](#_Toc133502980)

[1.1 Background of the Study 1](#_Toc133502981)

[1.2 Problem Statement 2](#_Toc133502982)

[1.3 Justification 3](#_Toc133502983)

[1.4 Purpose of the Study 3](#_Toc133502984)

[1.5 Objectives of the Study 3](#_Toc133502985)

[*1.5.1 General Objective* 3](#_Toc133502986)

[*1.5.2 Specific Objectives* 3](#_Toc133502987)

[1.6 Study Hypothesis 3](#_Toc133502988)

[*1.6.1 Null Hypothesis* 3](#_Toc133502989)

[1.7 Significance of the Study 4](#_Toc133502990)

[1.8 Scope of the Study 4](#_Toc133502991)

[1.9 Limitations 4](#_Toc133502992)

[CHAPTER TWO: LITERATURE REVIEW 5](#_Toc133502993)

[2.1 HIV AIDS in Kenya 5](#_Toc133502994)

[2.2 CD4 cells 5](#_Toc133502995)

[2.3 Treatment of HIV 7](#_Toc133502996)

[2.4 Support Vector Machines (SVM) 8](#_Toc133502997)

[CHAPTER THREE: METHODOLOGY 9](#_Toc133502998)

[3.1 Data Description 9](#_Toc133502999)

[3.2 Data Pre-Processing 12](#_Toc133503000)

[3.3 Data Analysis 13](#_Toc133503001)

[*3.3.1. To investigate demographic and socio-economic factors related to CD4 count levels.* 13](#_Toc133503002)

[*3.3.2. To fit an appropriate model to the data.* 13](#_Toc133503003)

[*3.3.3. To find out the factors affecting the CD4 levels on HIV positive women in Kenya* 15](#_Toc133503004)

[CHAPTER FOUR: RESULTS 16](#_Toc133503005)

[4.1. To investigate demographic and socio-economic factors related to CD4 cell count levels. 16](#_Toc133503006)

[*4.1.1. Education* 16](#_Toc133503007)

[*4.1.2 Relationship with head* 24](#_Toc133503008)

[4.2. Machine Learning Model 24](#_Toc133503009)

[*4.2.1. Support Vector Machine* 24](#_Toc133503010)

[4.3. To find out the factors affecting the CD4 cell levels on HIV positive women in Kenya 27](#_Toc133503011)

[*4.3.1. Household Characteristics* 33](#_Toc133503012)

[CHAPTER FIVE: CONCLUSION & RECOMMENDATIONS 34](#_Toc133503013)

[5.1 Conclusion 34](#_Toc133503014)

[CHAPTER SIX: REFERENCES 36](#_Toc133503015)

[APPENDIX 38](#_Toc133503016)

# **CHAPTER ONE: INTRODUCTION**

## **1.1 Background of the Study**

The HIV epidemic has been present for more than thirty years yet there is still no cure or an alternative vaccine for the disease. Therefore, HIV/AIDS has remained a major health crisis, especially in Sub-Saharan Africa, where adolescent girls and young women are at higher risk of infection as they account for about one in four new infections. In addition, in Eastern and Southern African regions, adolescent girls and young women accounted for 26% of new infections (UNAIDS, 2020).

HIV has caused immense human suffering in Sub-Saharan Africa, with the most obvious impact on individuals being death and suffering. The larger effect has been felt in the health and socio-economic sectors. For instance, in Sub-Saharan Africa, people living with HIV/AIDS-related problems occupy half of the hospital beds. The World Health Organization (WHO) had associating an increase in HIV vulnerability to legal and social factors (World Health Organization, 2022). However, great milestones had been achieved in treating HIV virus. For instance, availability and rapid scale-up of antiretroviral therapy (ART) drugs disbursement, voluntary male medical circumcision, antiretroviral medication for the prevention of mother to child transmission, pre-exposure prophylaxis, among others. The UNAIDS has also been keen on setting targets to specific countries while giving grants to research studies to attain zero new infections, zero discrimination and zero AIDS-related deaths.

The trends in the new HIV infections across countries in Africa had declined by more than 33%, from an estimated 2.2 million in 2005 to 1.5 million in 2013 (UNAIDS, 2022). The scale-up and widespread coverage of ART had led to substantial declines in new HIV infections. Despite these declines, HIV incidences rates remained unacceptably high, with the largest number of new infections coming from; South Africa (22%), Nigeria (15%), Uganda (10%), Rwanda (7%) and Kenya (7%). The epidemic in other Sub-Saharan countries had seen a substantial decline due to the impact of modest ART coverage at CD4 cell counts ranging from <200 to 500 per milliliter of blood. This had resulted in significant declines in mortality, with life expectancy increasing by an additional ten years. These studies provided evidence on the benefits of early ART initiation to HIV positive Individuals (UNAIDS, 2014).

Through previous researches, CD4 cell count is one of the parameters used to measure disease progression. HIV attacks CD4 cells reducing their levels in the body, making it difficult to fight diseases. Furthermore, CD4 cell count had been used for immunological classification of HIV infection where the levels had been shown to correlate with clinical stages of HIV related diseases (Barnett et al., 2008). In 2019, 1.5 million people were living with HIV, 4.9% adult prevalence (ages 15 – 49), 77,000 new infections, and 22,000 AID related deaths in Kenya (Kamer, 2022). The HIV prevalence in women was 6.6% while that of men was at 3.1% (Ministry of Health, 2020). Therefore, there was a higher burden on women as compared to men with the disease. The country recorded the following results; 79% of people living with HIV were aware of their status, 78.9% were on antiretroviral therapy, and 85.3% were virally suppressed (79-78.9-85.3) (NSCOP, 2020). This results was better when compared to the UNAIDS target of actualizing the 90-90-90 goal by 2020.Women were unreasonably affected by HIV in Kenya. In 2018, 890,000 women aged above the age of 15 were living with HIV compared to 510,000 men of the same age group. Similarly, the same year, 36,000 women were newly infected with HIV compared to 27,000 (UNAIDS, 2020).

In Kenya, there was a great disparity between the two genders. Women were mainly discriminated against men, with statistics showing that approximately 45% of women aged 15-49 who had never been married or in a long-term relationship were estimated to have experienced physical or sexual violence from an intimate male partner in 2019 (Odhiambo, 2020). Women also tend to be infected earlier because they had older partners and got married off earlier. Therefore, women were a key indicator of the country’s progress towards eliminating HIV/AIDS as a public health threat. They made up most of the total HIV infections and were more likely to experience more challenges accessing antiretroviral therapy treatment.

## **1.2 Problem Statement**

The prevalence rate of HIV among adults aged 15 to 64 years in Kenya was 4.9 per cent in 2020 (KENPHIA, 2020) hence it remains a public health concern, with a significant number of adults living with the virus. The country had made progress towards the UNAIDS 90-90-90 goal but needed to accelerate efforts towards the 2030 vision of ending AIDS as a public health threat (Frescura et al., 2022). To achieve this, it was crucial to understand factors that impact CD4 cell count among the HIV population in Kenya. Machine learning techniques, particularly support vector machines, had shown promise in analysing existing data sets related to HIV and CD4 cell count in the country. However, there is limited public research available on the use of these methods.

## **1.3 Justification**

Gender plays a big role in the community, women living with HIV in Kenya experience inequality compared to their male counterparts. This study aimed to break down barriers that prevent women from accessing quality and affordable testing and treatment services by using CD4 level as an indicator of good health. The analysis utilised the machine learning model Support Vector Machines (SVM). The study aimed to achieve this by utlising the machine learning models to provide a basis of estimation with a dichotomous outcome. The model would either predict a low or high CD4 cell count by using socio-demographic and other factors, with a high being a more likeable outcome. Therefore, the significant factors used would act as a compass in improving HIV+ infected women’s health, thereby improving general public health. Globally, this would make great strides towards attaining the UNAIDS target of eliminating HIV/AIDS as a public health threat while attaining sustainable development goals. In conclusion, quality affordable health is a basic right to everyone, and this study strived to provide statistical inferences toward the same.

## **1.4 Purpose of the Study**

To provide an insight into various factors that affect the CD4 cell count levels among women and give recommendations on the factors to be mitigated.

## **1.5 Objectives of the Study**

### ***1.5.1 General Objective***

The general objective of this study was to apply machine learning models in the identification of factors associated with CD4 cell count levels.

### ***1.5.2 Specific Objectives***

1. To investigate demographic and socio-economic factors related to CD4 cell count levels.
2. To fit a support vector machine learning model.
3. To find out the factors affecting the CD4 cell levels on HIV+ women in Kenya.

## **1.6 Study Hypothesis**

### ***1.6.1 Null Hypothesis***

There exists an association between the selected variable factors and CD4 cell count levels.

## **1.7 Significance of the Study**

This study would enable people and the government to get appraised on factors associated with CD4 cell count levels. This would enable the government to develop appropriate ways of mitigating these factors to alleviate the HIV pandemic. The models identified the main demographic and socio-economic factors associated with CD4 cell levels to help the relevant stakeholders provide solutions. Therefore, providing data insights to scientists, students and the general public doing related research topics. The government could implement relevant policies to bridge inequalities faced by women. This would greatly improve the health sector that would consequentially improve the country’s economy. In addition, due to COVID -19 outbreaks, most health resources had been channeled towards combating the COVID-19 pandemic. This move had resulted in the abandonment of other health threats such as HIV+ infections and a keen interest in creating solution indicators towards eliminating HIV as a public threat through machine models.

## **1.8 Scope of the Study**

The study was centered on the women’s population at risk that were already infected with HIV, aged 15-64 years in Kenya.

## **1.9 Limitations**

Lack of funds and adequate time confined the study to secondary data sources.

# **CHAPTER TWO: LITERATURE REVIEW**

## **2.1 HIV AIDS in Kenya**

According to UNAIDS, in 2019, 1.5 million people were living with HIV translating to a 4.7% adult age prevalence (ages 15-49), an increase from 2018 when it was 4.8%. In the same year, 42,000 new HIV infections were recorded, decreasing from 46,000 the previous year (UNAIDS, 2020). Approximately 50% of new infections were from ages 15-29 (AIDSinfo | UNAIDS, 2021). The HIV load in urban areas is higher in urban areas than in rural areas with 6.3% and 3.6%. HIV prevalence among women towers that of men with 6.6% against 3.1% among men. HIV prevalence is highest among women aged 45-49 at 12%. Among the young people (ages 15-24), the HIV prevalence is 1%, with women having a higher prevalence of 2% compared to 0.6% in men. The prevalence among children under 15 years is 0.3%. The PHIA report established that 79% of people living with HIV know their status, 78.9% of those under ART, with 85.3% of those under treatment have virally suppressed HIV (2021). 80 % of pregnant women living with HIV received ART for PMTCT, while 67% of children living with HIV are on ART (UNAIDS, 2020).

## **2.2 CD4 cells**

CD4 helper cell or T cell, also known as CD4 T lymphocyte, is a type of white blood cells responsible for fighting off bacteria, viruses, and other invading germs. CD4 count is a test that estimates the number of CD4 cells in a cubic millimetre of blood (Okoye and Picker, 2013). The test aids in establishing how much destruction have been done to your immune system and the likely outcome in the event antiretroviral treatment (ART) does not use this test. For a HIV negative person, the CD4 count should be anything between 500 and 1500 (Nall, 2021). People with HIV who have a CD4 count of over 500 are considered in good health. HIV negative people with less than 200 CD4 are considered to be at a greater risk of developing a serious illness.

Upon gaining entry to the body, HIV targets the immune system (white blood cells), particularly the CD4 cells. When too many CD4 cells are lost, the immune system gets weak, and it faces challenges facing infections. HIV lacks mechanisms to replicate on its own; hence, it attaches itself to the surface of the CD4, gets inside and becomes part of the cell (Nall, 2021). Once the CD4 cell is dead, HIV begins releasing more copies of HIV into the bloodstream. The newly released bits of HIV take over more CD4 cells, and the cycle continues reducing the number of HIV-free, working CD4 cells (Nall, 2021). When to destruction advances to a stage, the CD4 count drops below 200, the host is then said to have AIDS.

CD4 count, also known as CD4 lymphocyte count, CD4+ count, T4 count, enables the health care provider to check if an individual is at risk of any complications from HIV. The CD4 count can also be used to analyse and observe how HIV affects an individual’s immune system and if the individual is developing any complications from HIV. The test shows the advancement of your immune system regarding HIV, indicating if a change in medication will suffice to manage the situation. Also, when the CD4 cell count is too low, the patient can be diagnosed with AIDS.AIDS is a serious form of HIV virus, and it opens the body to opportunistic infections due to the damages it does to the immune system. Over more than two decades, CD4 cell counts have been critical in understanding the progression of HIV virus (Ford et al., 2015). This measurement determines when a patient should begin their antiretroviral therapy (ARV), and it shows the progression of the virus during the administration of this therapy. According to Ford et al. (2017), advancement in technology pushes out CD4 cell count in marking the beginning of ARV administration. The introduction of using viral load testing to monitor the virus’s progression in patients makes using CD4 cell counts void (Ford et al., 2017).

Studies have been carried out to identify the various factors that may affect CD4 cell counts in HIV positive individuals. The factors can be medically related or even related t the socio-economic situation of the individuals. These credible studies show that the CD4 cell counts are very flexible due to various environmental factors. Jones et al. (1993) and Ickovics et al. (2001) show the reaction of CD4 cell counts to medical issues and conditions. According to Jones et al. (1993), tuberculosis has a relationship with the CD4 cell counts in HIV positive patients. The study showed that the patients with low CD4 cell counts had more chances of contracting severe tuberculosis and those with a higher CD4 cell count had fewer chances of contracting the disease. This study showed how critical CD4 cell count is to an HIV positive individual. Ickovics et al. (2001) carried out a study to determine the association of depressive symptoms with HIV related mortality and the decline in CD4 cell count among women with HIV. Using a progressive and longitudinal cohort study and a multivariate analysis, the study showed that depressive symptoms are associated with the progression of HIV virus (Ickovics et al., 2001). The symptoms directly affect the CD4 cell counts, which progress the disease. These studies show the effect of mental and general diseases on CD4 cell counts, which enable the progression of the HIV virus.

Montarroyos et al. (2014) carried out a study meant to identify the factors related to the variations of CD4 counts in HIV positive patients. The study implemented a multilevel model using three levels of aggregation to analyse the association of the predictor variables and the fluctuations in CD4 level count over time (Montarroyos et al., 2014). The study found that CD4 counts level is related to factors like treatment adherence, patients’ habits, change in treatment or doctor and use of ART. The lives the patients live greatly determines the levels of CD4 count and the progression of the virus. The patients should be responsible for monitoring the lives they lead concerning the progression of the HIV virus in their bodies (Montarroyos et al., 2014).

## **2.3 Treatment of HIV**

Today there is no cure for HIV. The only existing remedies are medications that benchmark HIV and avert complications (HIV/AIDS - Diagnosis and treatment - Mayo Clinic, 2021). These medications are known as antiretroviral therapy (ART). ART prevents HIV from replicating and from destroying the immune system of an infected person. ART is usually a combination of three or more medications from several different drug classes (HIV/AIDS - Diagnosis and treatment - Mayo Clinic, 2021). The treatment uses drugs from different classes to cater to individual drug resistance, avoid generating new drug-resistance strains of HIV and optimise blood suppression. This combination is defined as Highly Active Antiretroviral Therapy (HAART). These medications help lower the amount of viral load in the body. Anyone diagnosed with HIV should immediately be enrolled on these medications.

Although there is no treatment for this virus, guidelines have been set to ensure the patients are well cared for and their health improves. Studying the factors that may affect individuals with HIV is critical when providing them with care and guidance. Studies have been conducted to try and come up with strategies that will reduce the occurrence of HIV. According to (Hayes et al., 2019), a combination prevention intervention with ART provided per the local guidelines resulted in a 30% lower incidence of HIV infection (Hayes et al., 2019). The study compares the combined method of intervention to standard care to point out what can be improved to provide quality treatment for HIV positive individuals. The presence of ARTs for the population led to a significant decline in the incidence of HIV virus.

Johnson et al. (2007) stress that adherence to treatment for HIV positive patients are critical in managing the virus. The study focused on the adherence of self-efficacy for treatment for HIV virus. Also, the paper validates the use of the HIV Treatment Adherence Self-Efficacy Scale (HIV-ASES) using two samples of HIV+ adults on ART (Johnson et al., 2007). The successful development of Highly Active Antiretroviral Therapy (HAART) was a great achievement in the management of HIV (Floridia et al., 2008). This study conducted by Floridia et al. (2008) investigated the gender differences in HIV therapeutics. The data on the drug response showed a similar outcome in men and women in the study. However, female candidates appear to be more vulnerable to adverse events related to the treatment (Floridia et al., 2008). This disparity between the genders poses an unprecedented challenge, and the treatment needs to be optimised to cover this disparity.

## **2.4 Support Vector Machines (SVM)**

Support Vector Machines (SVM) is a popular machine learning technique used for classification and regression analysis. SVM is particularly useful when the data is non-linearly separable, meaning that a linear decision boundary cannot accurately separate the data points. SVM works by finding the hyperplane that maximizes the margin between the support vectors, which are the data points closest to the decision boundary.

SVM is based on the idea of finding the optimal trade-off between minimizing the classification error and maximizing the margin, which makes it a powerful algorithm for complex datasets. SVM can be applied to a wide range of applications, including image classification, natural language processing, and financial prediction. However, one limitation of SVM is that it can be computationally expensive for large datasets, and the choice of kernel function can also have a significant impact on the accuracy of the model.

There is some evidence documented online about the use of machine learning techniques to analyze existing data sets related to HIV and CD4 count in Kenya. However, the use of machine learning in this context is still a relatively new area of research, and there may not be as much literature available on this topic compared to more established areas of HIV research. In regards to the use of machine learning, a study by Daniel Niguse Mamo et al. (2023) used random forests classifier outperformed in predicting and identifying the relevant predictors of virological failure in Ethiopia. This outcome suggested that these techniques may have utility in improving HIV care for people. Another study published in the journal PLOS ONE in 2021 used machine learning techniques to identify factors associated with virologic failure among HIV-positive individuals receiving antiretroviral therapy in Kenya. The study found that several clinical and demographic factors were strongly associated with virologic failure, including age, sex, baseline CD4 count, and viral load (Masaba et al., 2023).

# **CHAPTER THREE: METHODOLOGY**

## **3.1 Data Description**

The data is the latest data set of HIV positive women in Kenya extracted from Kenya’s PHIA. The sample size is 1242 women, and the variables under survey were 28 variables. The variables include categorical and numerical variables. Different levels of the categorical variables are coded in numerical form. They are described as follows:

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Description of Variable** | **Coded Values and Labels** |
| **Dependent variable** |  |  |
| * CD4.category | CD4 level category | 1. High   0- Low |
| **Independent variables** |  |  |
| * age | Age in years | Between 15 and 65 |
| * agelim | Age groups for population pyramid | 0-4 years  5-9 years  10-14 years  15-19 years  20-24 years  25-29 years  30-34 years  35-39 years  40-44 years  45-49 years  50-54 years  55-59 years  60-64 years  65-69 years |
| * shipwhead | Relationship of the individual with the head of the family | 1. Head 2. Wife/husband/partner 3. Son/ daughter 4. Son/daughter in law 5. Grandchild 6. Parent 7. Parent in law 8. Brother/sister 9. Co wife 10. Other relatives 11. Adopted/foster/stepchild 12. Not related |
| * liveinhold | Does the individual live in the house | 1. No 2. Yes |
| * sickwork | Has the individual been very sick for at least three months | 1. No 2. Yes |
| * gosch | Ever attended school | 1. No 2. Yes |
| * enrollsch | Ever enrolled in school | 1. No 2. Yes |
| * leveledu | Highest level of school you attended | 1. primary 2. post-primary training 3. secondary (O- level) |
| * gradelevel | Highest grade at the school level | Between 0 and 14 |
| * workpay | Work for payment past 12 months | 0 - No  1 - Yes |
| * livetogether | Ever married or lived together | 0 - No  1 - Yes |
| * pregnancies | Number of pregnancies | Between 0 and 13 |
| * liveborn | Ever had a pregnancy that resulted in a live birth | 1. No 2. Yes |
| * numchild2012 | Number of children given birth since 2012 | 0,1,…,10 |
| * pregnow | Current pregnancy status | 1. Not currently pregnant 2. Currently pregnant |
| * avoidpreg | Avoiding pregnancy | 1. No 2. Yes |
| * age1stsex | Age at first sex | Between 8 and 35 |
| * age1stsexlim | Age groups for population pyramid for age at first sex | 0-4 years  5-9 years  10-14 years  15-19 years  20-24 years  25-29 years  30-34 years  35-39 years |
| * tbtreat | Ever sought TB treatment | 1. No 2. Yes |
| * alcofreq | How often does the individual have a drink containing alcohol | 1. NEVER 2. MONTHLY OR LESS 3. 2-4 TIMES A MONTH 4. 2-3 TIMES A WEEK 5. 4 OR MORE TIMES A WEEK |
| * urban | Urban area indicator | 1. Rural 2. Urban |
| * knownstat | Known HIV status | 1. STATED HIV NEGATIVE 2. STATED HIV POSITIVE |
| * wealthq | Wealth quantile | 1. Lowest 2. Second 3. Middle 4. Fourth 5. Highest |
| * sexlast12 | Respondent had sexual intercourse in the past 12 months | 1. No 2. Yes |
| * everhadsex | Respondent ever had sexual intercourse | 1. No 2. Yes |
| * buysellsex12 | Bought/sold sex past 12 months | 1. No 2. Yes |
| * onart | Indicator whether the respondent is on ART | 1. On ART 2. Not on ART |
| * timeonart | Duration of time on ART | 1. On ART 24 months or more 2. On ART 12-23 months 3. On ART <12 months 4. Not on ART |
| * arvsdetected | Indicator whether ARVs detected | 1. ARVs not detected 2. ARVs detected |

## **3.2 Data Pre-Processing**

The pre-processing of data was carried out by doing the data wrangling process, also known as data cleaning. Ridzuan (2022) explains data cleaning as the process of modifying data to ensure that it is free of irrelevances and incorrect information expounds on data cleaning steps and weighs the advantages and disadvantages of data cleaning. The data was cleaned in preparation for its analysis. R Programming was used to clean the data. The removal of irrelevant observations followed in the cleaning process. The data structure was acknowledged in the software; categorical and numerical variables were recognised. Also, outliers were identified and filtered out. The variables in the data representing the ages, age and age1st sex variables were recognised into age groups to deeper understand the characteristics of the age-based variables. In addition, the data variables were renamed to aid in the imputation and avoid overlapping in the visualisations.

Data was checked for the missing data in terms of percentage. If it is below thirty per cent (30%), then imputation should be done. In addition, the method of handling the missing data depending on whether the missing values are missing at random or not was determined. In the case of missing data, the basic assumption for our data was that the missing values are missing at random for the missing values to be imputed. The missing data is addressed using visualisations to understand the distribution of missing data. The missing values will be imputed using the tidyverse package in R programming software. The tidyverse package in R-Software is a suitable line of data pre-processing (Schober & Vetter, 2020).

## **3.3 Data Analysis**

### ***3.3.1. To investigate demographic and socio-economic factors related to CD4 count levels.***

Socio-demographics are nothing more than characteristics of a population. Generally, characteristics such as age, gender, ethnicity, education level, income, years of experience, location, etc., are considered socio-demographic factors. Cross tabulations (also referred to as cross-tabs) are a quantitative research method appropriate for analysing the relationship between two or more variables. Cross tabulations provide a way of analysing and comparing the results for one or more variables with other(s) results.

Dwumoh et al. (2014), state that Determinant of factors associated with child health outcomes and service utilisation in Ghana: Multiple indicator cluster survey conducted in 2011. Using cross-table(s) to investigate Cross-tabulation of socio-demographic characteristics and National Health Insurance Scheme Membership of children under-five based on Chi-square test statistic with the corresponding p-value in Ghana, 2011 (Dwumoh et al., 2014). The socio-demographic factors would be crossed with our dependent variable (CD4 level) to determine if there is an association between the socio-demographics and the outcome of CD4 level in the data.

R-Software will be used in determining the associations.

### ***3.3.2. To fit an appropriate model to the data.***

#### ***3.3.2.1. Support Vector Machines (SVM)***

SVM is a supervised machine learning algorithms used in classification as well as regression problems. The goal of this algorithm is to create best line or decision boundary, which we can use to partition n-dimensional space into classes to help fit new data points in right categories. This best line is known as hyperplane. SVM chooses extreme points or vectors best known as support vectors to create a hyperplane hence algorithm is known as support vector machine. The algorithm is popularly used in face detection and image classification.

**Types of SVM**

1. Linear SVM: used in linearly separable data
2. Non-linear SVM: used in non-linearly separable data.

The linear SVM is used as a classifier that segregates classes or categories into n-dimensional space. The closest data points to the decision boundaries are known as support vectors and determine the position of this decision boundary. The distance between the vectors or data points and decision boundary is known as margin. The goal of SVM is to maximise this margin and find the hyperplane/decision boundary with the maximum distance from vectors known as optimal hyperplane. Linear SVM: The mathematical model for a linear SVM is:

f(x) = wT x + b……….(i)

Where: w is the weight vector

x is the input vector

b is the bias term

In case of non-linear SVM we cannot have a straight-line separating data point. The classifier will segregate classes into more than 2 dimensional spaces. For our case we will be using non-linear SVM since our dataset cannot be classified by using a straight line. The mathematical equation for a non-linear SVM involves the use of a kernel function to transform the input data into a higher-dimensional space. The most commonly used kernel functions are the Gaussian kernel and the polynomial kernel. In our project, the kernel function would be used to identify the factors related to CD4 cell count among HIV positive patients that may not be linearly separable in the original feature space. The mathematical model for a non-linear SVM is:

f(x) = ∑(αi \* yi \* K(xi, x)) + b……….(ii)

Where: α is a vector of coefficients,

y is the output vector,

K is the kernel function,

xi is the input vector,

b is the bias term.

xi represent the input features related to CD4 cell count in HIV positive patients, and y represent the low or high CD4 cell count class. The goal of the non-linear SVM is to find the hyperplane that separates the data with the maximum margin in the transformed feature space.

R-Software will be used to fit in a support vector machine model to the data.

### ***3.3.3. To find out the factors affecting the CD4 levels on HIV positive women in Kenya***

Researchers research to develop new theories, ideas and products that shape our society and our everyday lives. The purpose of research is to understand the further world and learn how this knowledge can be applied to better everyday life. It is an integral part of problem-solving. Using the data, we hope to conduct a detailed analysis to investigate the factors that affect HIV positive women in Kenya. Information on variables affecting the levels of CD4 level in HIV-Positive women in Kenya was presented to promote intervention studies and surveys.

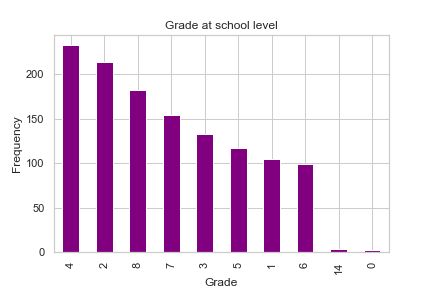
The variables that significantly affect the result of the CD4 level of an individual shall be obtained. The association and relationship of the factors to the CD4 levels was extracted using cross-tables and results interpreted. These procedures were accomplished using R-Software. Understanding the functions of the programs was critical to understand and internalise the entire project.

# **CHAPTER FOUR: RESULTS**

## **4.1. To investigate demographic and socio-economic factors related to CD4 cell count levels.**

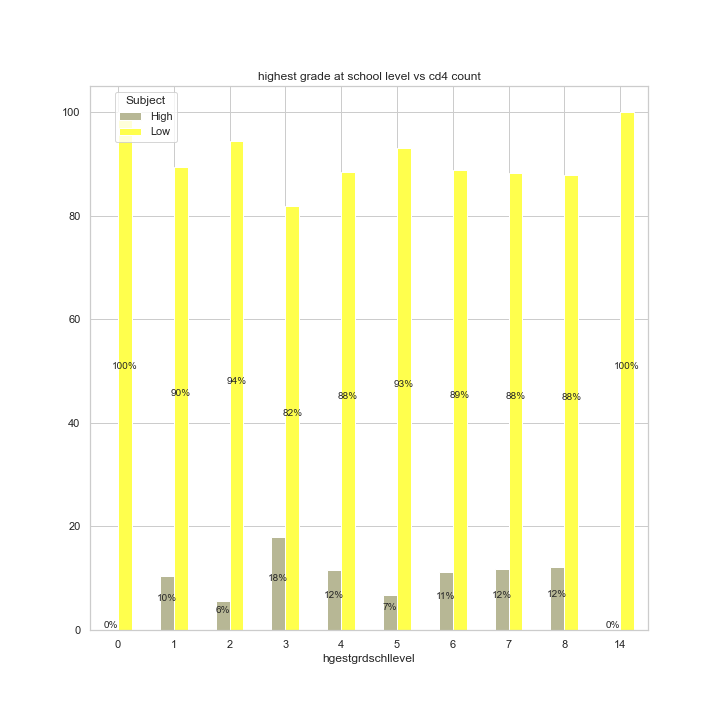
### ***4.1.1. Education***

#### ***4.1.1.1 Highest grade at school***



*Figure 1: Highest grade in school*

The majority of the respondents had grade 4 as their highest-grade level (233), followed by grade 2 with (214) respondents. Those with the highest-grade level as 14 and 0 recorded relatively few respondents.



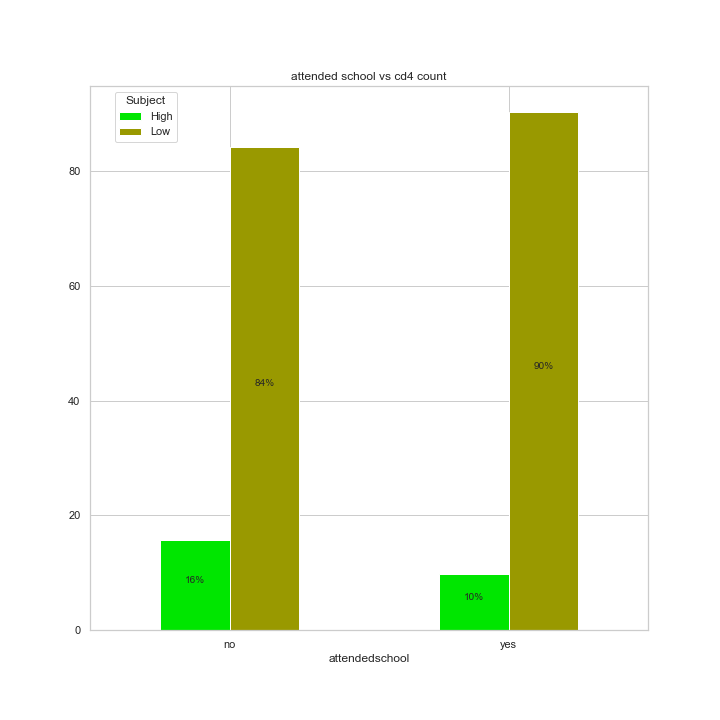
*Figure 2: Highest grade in school against CD4 cell count*

From *figure 2*, the respondents in grade 2 recorded the highest low CD4 count at 94%, while those in grade 4 recorded the least low cd4 count at 82%. On the other hand, the highest low cd4 count was 18% recorded for grade 3, while the least percentage of high cd4 count was 6% for respondents with grade 2 as the highest-grade level. Respondents in grade 0 and 14 are considered as outliers and have no significant value.

#### ***4.1.1.2 Attended school***

*Figure 3: Attended school*

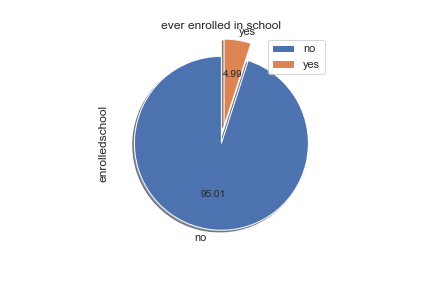
The figure below indicates the association between those who attended school and their cd4 levels. With 84% of the respondents having attended school [*figure 3*], 16% of those who attended school had a high CD4 count compared to 10 % of those who did not attend school and had a high CD4 count [*figure 4*]. On the other hand, 90% of those who attended school have a low CD4 count compared to 84% of those who did not attend school and had a low CD4 count.



*Figure 4: Attended school against CD4 Cell count*

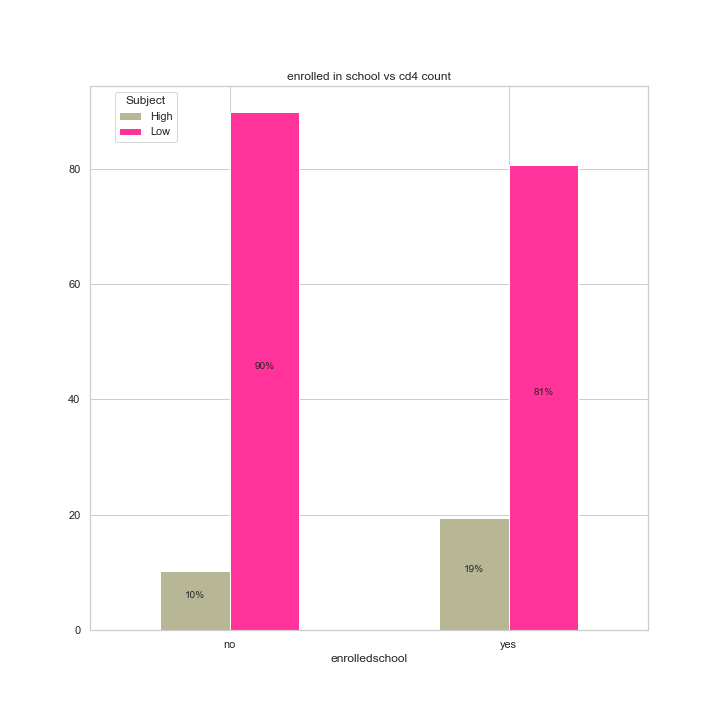
#### ***4.1.1.3 Enrolled in school***

According to *figure 5,* 95% of the respondents have never enrolled in school.



*Figure 5: Enrolled in school*

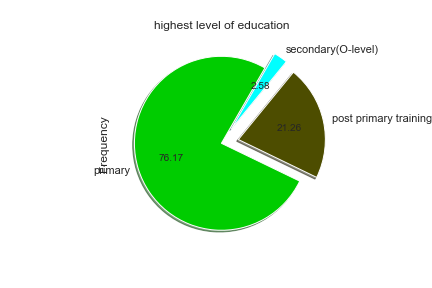
In figure 6, 90% of individuals who did not enroll in school have a low cd4 count while 10% have a high cd4 count. Furthermore, 81% of those who enrolled in school have a low cd4 count, while 19% have a high cd4 count.



*Figure 6: Enrolled in school against CD4 cell count*

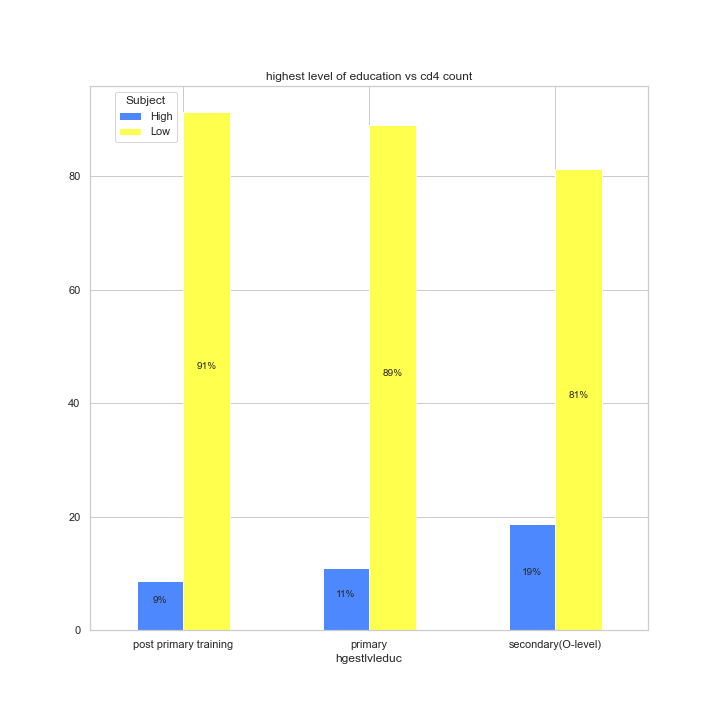
#### ***4.1.1.4 Highest level of education***

Figure 7 shows that 76% of the respondent had primary school education, 21% had post-primary training. Only 3 % of the respondents had secondary education.



*Figure 7: Highest level of education*

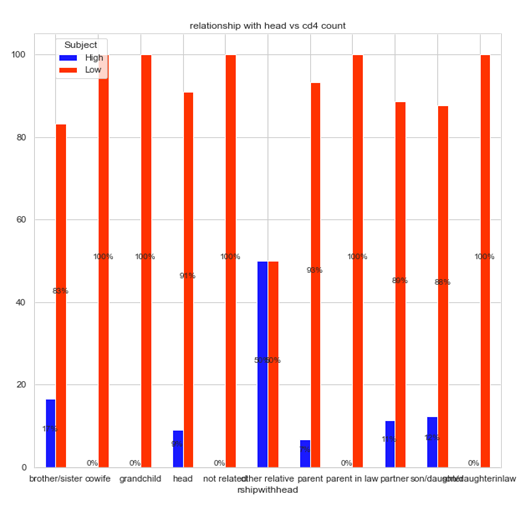
From figure 8*,* 91% of respondents with post-primary training had a low CD4 count, and 9% with the same training had a high CD4 count. 89% of the respondents with primary education had a low CD4 count compared to 11% with a high CD4 count with the same education. 81% of those with secondary education had a low CD4 count compared with their 19% counterparts with a high CD4 count.



*Figure 8: Highest level of education against CD4 cell count*

### ***4.1.2 Relationship with head***

From *figure 9,* the respondent with a relationship with co-wife, grandchild, not related, a parent in law, daughter in law all had no record of high CD4 count resulting to a 100% record of low CD4 count. Therefore, 90% of respondents related to the head had a low CD4 count, and 9 % of heads of families had a high CD4 count. 50% of the whole relationship with the head of the family was a relative who had a high CD4 count. Sons/daughters of the head had 12% with high CD4 counts. 11% of family head partners had a high CD4 count.



*Figure 9: Relationship with head against CD4 cell count*

## **4.2. Machine Learning Model**

### ***4.2.1. Support Vector Machine***

To fit a support vector machine (SVM) model, we first installed and loaded the e1071 package. We then split the dataset into training and testing sets with 70% of the data used for training and 30% for testing. Next, we built the SVM model using the ‘svm’ function from the e1071 package. We specified CD4.category as the target variable and used all other variables as predictors with a linear kernel. We then made predictions on the test data using the ‘predict’ function and evaluated the performance of the SVM model using various metrics such as accuracy, precision, recall, and F1 score.

The performance evaluation revealed that the SVM model had an accuracy of 0.949, which indicates that the model was able to correctly classify 94.9% of the test data. The precision for the CD4 category 1 and 2 were 0.992 and 0.143 respectively. The recall for the CD4 category 1 and 2 were 0.956 and 0.5 respectively. The F1 score for the CD4 category 1 and 2 were 0.974 and 0.222 respectively.

These results indicate that the SVM model was able to perform well in predicting the CD4 category of individuals based on the selected predictors. We therefore agreed on using SVM rather than the Logistic regression which had an accuracy of 88 percent

**SVM model**

Call:

svm(formula = CD4.category ~ ., data = train\_data, kernel = "linear")

Parameters:

SVM-Type: C-classification

SVM-Kernel: linear

cost: 1

Number of Support Vectors: 53

* SVM-Type: This shows that the SVM is a C-classification type, which means that the SVM is trained to perform classification tasks.
* SVM-Kernel: This shows that a linear kernel was used in this model, which means that the decision boundary is a straight line.
* cost: This shows that the cost parameter was set to 1. The cost parameter controls the trade-off between maximizing the margin and minimizing the classification error.
* Number of Support Vectors: This shows that 53 support vectors were used in the model. Support vectors are the data points that are closest to the decision boundary and have the most influence on the classification.

**1. Accuracy**: Accuracy is a metric used to measure the overall performance of a classification model. It represents the proportion of correctly classified samples out of the total number of samples.

Formula: Accuracy = (True Positive + True Negative) / (True Positive + False Positive + True Negative + False Negative)

Explanation: True Positive (TP): The number of samples that are actually positive and are correctly classified as positive by the model. False Positive (FP): The number of samples that are actually negative but are incorrectly classified as positive by the model. True Negative (TN): The number of samples that are actually negative and are correctly classified as negative by the model. False Negative (FN): The number of samples that are actually positive but are incorrectly classified as negative by the model.

**2.Precision**: Precision is a metric that measures the proportion of correctly classified positive samples out of all the samples classified as positive.

Formula: Precision = True Positive / (True Positive + False Positive)

Explanation: True Positive (TP): The number of samples that are actually positive and are correctly classified as positive by the model. False Positive (FP): The number of samples that are actually negative but are incorrectly classified as positive by the model.

**Recall**: Recall is a metric that measures the proportion of correctly classified positive samples out of all the actual positive samples.

Formula: Recall = True Positive / (True Positive + False Negative)

Explanation: True Positive (TP): The number of samples that are actually positive and are correctly classified as positive by the model. False Negative (FN): The number of samples that are actually positive but are incorrectly classified as negative by the model.

**F1 Score**: F1 score is the harmonic mean of precision and recall. It provides a balanced measure between precision and recall, which is useful when the classes are imbalanced.

Formula: F1 Score = 2 \* (Precision \* Recall) / (Precision + Recall)

Explanation: Precision: The proportion of correctly classified positive samples out of all the samples classified as positive. Recall: The proportion of correctly classified positive samples out of all the actual positive samples.

**# Calculate performance metrics**

table <- table(svm\_pred, test\_data$CD4.category)

accuracy <- sum(diag(table)) / sum(table)

precision <- diag(table) / colSums(table)

recall <- diag(table) / rowSums(table)

f1\_score <- 2 \* precision \* recall / (precision + recall)

**Output**

|  |  |  |
| --- | --- | --- |
| **Metric** | **Low CD4** | **High CD4** |
| Accuracy | 0.9496403 | **-** |
| Precision | 0.9924242 | 0.1428571 |
| Recall | 0.9562044 | 0.5 |
| F1 Score | 0.9739777 | 0.2222222 |

## **4.3. To find out the factors affecting the CD4 cell levels on HIV positive women in Kenya**

The significant characteristics of the study participants are summarised in Table 1 below;

|  | | | |
| --- | --- | --- | --- |
|  |  | **CD4 Level** | |
| **Characteristics** | **Total (N=1242)** | **High (N=133)** | **Low (N=1109)** |
| **Relationship With Head** |  |  |  |
| brother/sister | 24 (1.9%) | 4 (3.0%) | 20 (1.8%) |
| Co-wife | 1 (0.1%) | 0 (0%) | 1 (0.1%) |
| grandchild | 8 (0.6%) | 0 (0%) | 8 (0.7%) |
| head | 681 (54.8%) | 62 (46.6%) | 619 (55.8%) |
| not related | 6 (0.5%) | 0 (0%) | 6 (0.5%) |
| other relative | 20 (1.6%) | 10 (7.5%) | 10 (0.9%) |
| parent | 15 (1.2%) | 1 (0.8%) | 14 (1.3%) |
| parent in law | 1 (0.1%) | 0 (0%) | 1 (0.1%) |
| partner | 412 (33.2%) | 47 (35.3%) | 365 (32.9%) |
| son/daughter | 73 (5.9%) | 9 (6.8%) | 64 (5.8%) |
| Son-in-law/daughter-in-law | 1 (0.1%) | 0 (0%) | 1 (0.1%) |
| **Sick to Work last 3 Months** |  |  |  |
| no | 1096 (88.2%) | 103 (77.4%) | 993 (89.5%) |
| yes | 146 (11.8%) | 30 (22.6%) | 116 (10.5%) |
| **Ever Attended School** |  |  |  |
| no | 197 (15.9%) | 31 (23.3%) | 166 (15.0%) |
| yes | 1045 (84.1%) | 102 (76.7%) | 943 (85.0%) |
| **Ever Enrolled in School** |  |  |  |
| no | 1180 (95.0%) | 121 (91.0%) | 1059 (95.5%) |
| yes | 62 (5.0%) | 12 (9.0%) | 50 (4.5%) |
| **Work for Pay** |  |  |  |
| no | 919 (74.0%) | 104 (78.2%) | 815 (73.5%) |
| yes | 323 (26.0%) | 29 (21.8%) | 294 (26.5%) |
| **Married/Live Together** |  |  |  |
| no | 90 (7.2%) | 8 (6.0%) | 82 (7.4%) |
| yes | 1152 (92.8%) | 125 (94.0%) | 1027 (92.6%) |
| **Number of Pregnancies** |  |  |  |
| Mean (SD) | 4.15 (2.39) | 3.62 (2.11) | 4.22 (2.41) |
| Median [Min, Max] | 4.00 [0, 13.0] | 4.00 [0, 9.00] | 4.00 [0, 13.0] |
| **Pregnant Currently** |  |  |  |
| Currently not pregnant | 1190 (95.8%) | 121 (91.0%) | 1069 (96.4%) |
| Currently pregnant | 52 (4.2%) | 12 (9.0%) | 40 (3.6%) |
| **Ever Avoided Pregnancy** |  |  |  |
| no | 669 (53.9%) | 96 (72.2%) | 573 (51.7%) |
| yes | 573 (46.1%) | 37 (27.8%) | 536 (48.3%) |
| **Ever Sought TB Treatment** |  |  |  |
| no | 1009 (81.2%) | 99 (74.4%) | 910 (82.1%) |
| yes | 233 (18.8%) | 34 (25.6%) | 199 (17.9%) |
| **Duration on ART** |  |  |  |
| < 12 months | 118 (9.5%) | 8 (6.0%) | 110 (9.9%) |
| 12-23 months | 120 (9.7%) | 18 (13.5%) | 102 (9.2%) |
| 24 months or more | 346 (27.9%) | 70 (52.6%) | 276 (24.9%) |
| Not on ART | 658 (53.0%) | 37 (27.8%) | 621 (56.0%) |
| **ARVs Detected** |  |  |  |
| ARVs detected | 916 (73.8%) | 61 (45.9%) | 855 (77.1%) |
| ARVs not detected | 326 (26.2%) | 72 (54.1%) | 254 (22.9%) |
|  |  |  |  |

**Crosstab of HIV status and sex**

|  | **HIV Positive** | **HIV Negative** | **Total** |
| --- | --- | --- | --- |
| Male | 63 | 17 | 80 |
| Female | 172 | 62 | 234 |
| Transgender | 1 | 0 | 1 |
| Total | 236 | 79 | 315 |

**Crosstab of HIV status and marital status**

|  | **HIV Positive** | **HIV Negative** | **Total** |
| --- | --- | --- | --- |
| Married/Cohabiting | 122 | 37 | 159 |
| Single | 87 | 39 | 126 |
| Separated/Divorced | 25 | 1 | 26 |
| Widowed | 2 | 2 | 4 |
| Total | 236 | 79 | 315 |

**Crosstab of HIV status and age group**

|  | **HIV Positive** | **HIV Negative** | **Total** |
| --- | --- | --- | --- |
| 15-24 | 23 | 16 | 39 |
| 25-34 | 72 | 23 | 95 |
| 35-44 | 65 | 17 | 82 |
| 45-54 | 49 | 13 | 62 |
| 55+ | 27 | 10 | 37 |
| Total | 236 | 79 | 315 |

**Crosstab of HIV status and education level**

|  | **HIV Positive** | **HIV Negative** | **Total** |
| --- | --- | --- | --- |
| No education | 24 | 6 | 30 |
| Primary incomplete | 60 | 15 | 75 |
| Primary complete | 57 | 18 | 75 |
| Secondary incomplete | 54 | 23 | 77 |
| Secondary complete | 30 | 10 | 40 |
| Tertiary | 10 | 7 | 17 |
| Vocational/technical | 1 | 0 | 1 |
| Other | 0 | 0 | 0 |
| Total | 236 | 79 | 315 |

**CD4 Cell Count by Age Group**

| **CD4 Cell Count** | **18-24 years** | **25-34 years** | **35-44 years** | **45-54 years** | **55+ years** |
| --- | --- | --- | --- | --- | --- |
| < 200 | 18 (40.9%) | 36 (37.5%) | 52 (38.2%) | 39 (34.2%) | 26 (29.5%) |
| 200-499 | 17 (38.6%) | 36 (37.5%) | 50 (36.8%) | 43 (37.7%) | 34 (38.6%) |
| >= 500 | 9 (20.5%) | 21 (21.9%) | 35 (25.7%) | 33 (28.9%) | 31 (35.2%) |

**Table CD4 Cell Count by Gender**

| **CD4 Cell Count** | **Female** | **Male** |
| --- | --- | --- |
| < 200 | 97 (36.7%) | 74 (32.7%) |
| 200-499 | 102 (38.6%) | 85 (37.6%) |
| >= 500 | 68 (25.7%) | 1. 27.4%) |

### 

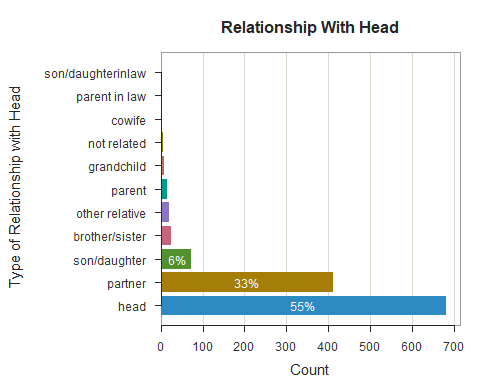
Based on the tables above, we can see that several characteristics of the study participants are associated with their CD4 cell levels. For instance, being sick to work in the last three months and having ever sought TB treatment are associated with low CD4 levels. On the other hand, being on ART for 24 months or more, currently pregnant, and having ARVs detected are associated with high CD4 levels. Moreover, the table suggests that some characteristics do not have a significant relationship with CD4 levels, such as being a co-wife or a parent-in-law. Additionally, some characteristics are not evenly distributed among the high and low CD4 categories, such as the duration on ART and ever avoided pregnancy. Women who have been on ART for 24 months or more are more likely to have high CD4 levels than those who have been on ART for less than 24 months, and women who have ever avoided pregnancy are more likely to have low CD4 levels than those who have never avoided pregnancy.

Overall, these findings suggest that various factors, such as ART adherence, TB treatment, and pregnancy status, are associated with CD4 cell levels in HIV-positive women in Kenya. Healthcare providers can use this information to target interventions and provide appropriate care to improve the CD4 cell levels of HIV-positive women in Kenya.

### ***4.3.1. Household Characteristics***

#### ***4.3.1.1 Relationship with the household head***

From Figure 1 below, most (55%) of the respondents were the heads and therefore produced 44.6% of the respondents with high CD4 level, as shown in Table 1. Only 5% were not related to the head of the household, and all of them recorded low CD4 levels.



# **CHAPTER FIVE: CONCLUSION & RECOMMENDATIONS**

## **5.1 Conclusion**

The general objective of this study was to investigate if the level of CD4 cells can be affected by social and economic factors. Using data from Kenya, we achieved our objective by performing descriptive statistics on the selected variables. Based on the analysis and findings of this study, it can be concluded that social and economic factors have a significant impact on the level of CD4 cells in individuals with HIV. Specifically, factors such as education, employment, and the ability to work can affect CD4 cell count, which plays a critical role in managing HIV.

The results of our SVM model with a linear kernel were able to predict the CD4 category with high accuracy of 94.96% using a set of 27 predictors. The precision for low CD4 category was 99.24% and for high CD4 category was 14.29%, indicating that the model was highly accurate in predicting individuals who have a low CD4 count, but less accurate in identifying those who do not have a low CD4 count. The recall for low CD4 category was 95.62% and for high CD4 category was 50%, indicating that the model was highly effective in identifying individuals who have a low CD4 count, but less effective in identifying those who have a high CD4 count. The F1 score for low CD4 category was 97.40% and for high CD4 category was 22.22%.

The SVM model performed well in predicting the CD4 category using the set of 27 predictors, with high accuracy, precision, and recall for low CD4 category and lower accuracy, precision, and recall for high CD4 category. This suggests that the model can effectively identify individuals who are at higher risk of having a low CD4 count, but may require further refinement to identify those who are at lower risk. Furthermore, this study emphasizes the crucial role of enrolling in school and having gainful employment in the management of HIV. Education and employment can provide individuals with HIV access to resources, which can help them better manage their condition and improve their overall quality of life. These findings can be valuable for informing clinical decisions and interventions related to HIV treatment and manage

In conclusion, the findings of this study can be used to inform policymakers and healthcare providers on the importance of addressing social and economic factors in the management of HIV. This can help improve the effectiveness of HIV management programs and ultimately lead to better health outcomes for individuals living with HIV.

**Top of Form**

**5.2. Recommendations**

This study covered numerous issues that affect the public and, most importantly, a group of vulnerable people. Understanding what HIV positive individuals need is the first step to successfully managing this virus. Our research should shed light on the various challenges these people face in our communities and the effort to address these challenges. Although the data we used is for Kenyan women living with HIV, the information gathered in this study projects challenges faced by other people living with HIV worldwide. The government and relevant stakeholders should consider these recommendations in Kenya to improve the management of this virus. Our research pointed out some issues that would improve the management of HIV in Kenya. As pointed out by Anand et al. (2009), implementing comprehensive positive prevention measures will come a long way in reducing the impact of HIV in Sub-Saharan Africa.

Based on the findings of this study, we recommend that healthcare providers and policymakers prioritize the education of individuals on the significance of enrolling in school and obtaining employment, particularly those living with HIV. This recommendation is based on our results, which showed that enrolling in school and being employed are positively correlated with CD4 cell count. Therefore, promoting education and employment opportunities for individuals living with HIV may have a positive impact on their CD4 cell count levels and, ultimately, their overall health.

Moreover, we recommend that healthcare providers offer comprehensive HIV management programs that focus on the social and economic factors that affect CD4 cell count levels. This study has shown that these factors have a significant impact on the management of HIV, and therefore, healthcare providers must consider these factors when designing treatment plans for individuals living with HIV. Lastly, we recommend that further studies be conducted to explore the role of other social and economic factors on CD4 cell count levels. This study only focused on a limited number of variables, and there is a need for more research to be conducted to gain a deeper understanding of the social and economic factors affecting CD4 cell count levels. By doing so, we can enhance the current knowledge on HIV management and develop more effective strategies to manage this pandemic.

# **CHAPTER SIX: REFERENCES**

Anand, P., Hunter, G., Carter, I., Dowding, K., Guala, F., & Van Hees, M. (2009). The Development of Capability Indicators. *Journal of Human Development and Capabilities*, *10*(1), 125–152. https://doi.org/10.1080/14649880802675366

Barnett, D., Walker, B., Landay, A., & Denny, T. N. (2008). CD4 immunophenotyping in HIV infection. *Nature Reviews Microbiology*, *6*(S11), S7–S15. https://doi.org/[10.1038/nrmicro1998](file:///C:\Users\NORWICH%20SERVICES\AppData\Local\Microsoft\Windows\INetCache\IE\Q0TZ9KFM\10.1038\nrmicro1998)

Daniel Niguse Mamo, Tesfahun Melese Yilma, Makida Fekadie, Sebastian, Y., Tilahun Bizuayehu, Mequannent Sharew Melaku, & Agmasie Damtew Walle. (2023). Machine learning to predict virological failure among HIV patients on antiretroviral therapy in the University of Gondar Comprehensive and Specialized Hospital, in Amhara Region, Ethiopia, 2022. *BMC Medical Informatics and Decision Making*, *23*(1). https://doi.org/10.1186/s12911-023-02167-7

Elul, B., Basinga, P., Nuwagaba-Biribonwoha, H., Saito, S., Horowitz, D., Nash, D., Mugabo, J., Mugisha, V., Rugigana, E., Nkunda, R., & Asiimwe, A. (2013). High Levels of Adherence and Viral Suppression in a Nationally Representative Sample of HIV-Infected Adults on Antiretroviral Therapy for 6, 12 and 18 Months in Rwanda. *PLoS ONE*, *8*(1), e53586. https://doi.org/[10.1371/journal.pone.0053586](file:///C:\Users\NORWICH%20SERVICES\AppData\Local\Microsoft\Windows\INetCache\IE\Q0TZ9KFM\10.1371\journal.pone.0053586)

Frescura, L., Godfrey-Faussett, P., Feizzadeh A., A., El-Sadr, W., Syarif, O., & Ghys, P. D. (2022). Achieving the 95 95 95 targets for all: A pathway to ending AIDS. *PLOS ONE*, *17*(8), e0272405. https://doi.org/[10.1371/journal.pone.0272405](file:///C:\Users\NORWICH%20SERVICES\AppData\Local\Microsoft\Windows\INetCache\IE\Q0TZ9KFM\10.1371\journal.pone.0272405)

Kamer, L. (2022). *AIDS-related deaths leading countries worldwide 2021*. Statista. <https://www.statista.com/statistics/281396/countries-with-highest-number-of-aids-deaths/>

KENPHIA. (2020). *KENPHIA Preliminary Report*. Www.health.go.ke. <https://www.health.go.ke/wp-content/uploads/2020/02/KENPHIA-2018-PREL-REP-2020-HR3-final.pdf>

Masaba, R., Woelk, G., Siamba, S., Ndimbii, J., Ouma, M., Khaoya, J., Kipchirchir, A., Boniface Ochanda, & Okomo, G. (2023). Antiretroviral treatment failure and associated factors among people living with HIV on therapy in Homa Bay, Kenya: A retrospective study. *PLOS Global Public Health*, *3*(3), e0001007–e0001007. https://doi.org/10.1371/journal.pgph.0001007

Ministry of Health. (2020). *Kenya’s National HIV Survey Shows Progress Towards Control of the Epidemic. Nairobi, 20th February 2020 – MINISTRY OF HEALTH*. Health.go.ke. <https://www.health.go.ke/kenyas-national-hiv-survey-shows-progress-towards-control-of-the-epidemic-nairobi-20th-february-2020/#:~:text=The%20Government%20today%20released%20preliminary>

Nsanzimana, S., Rwibasira, G. N., Malamba, S. S., Musengimana, G., Kayirangwa, E., Jonnalagadda, S., Fazito Rezende, E., Eaton, J. W., Mugisha, V., Remera, E., Muhamed, S., Mulindabigwi, A., Omolo, J., Weisner, L., Moore, C., Patel, H., & Justman, J. E. (2022). HIV incidence and prevalence among adults aged 15-64 years in Rwanda: Results from the Rwanda Population-based HIV Impact Assessment (RPHIA) and District-level Modeling, 2019. *International Journal of Infectious Diseases*, *116*, 245–254. https://doi.org/[10.1016/j.ijid.2022.01.032](file:///C:\Users\NORWICH%20SERVICES\AppData\Local\Microsoft\Windows\INetCache\IE\Q0TZ9KFM\10.1016\j.ijid.2022.01.032)

NSCOP. (2020). *Division of National AIDS & STI Control Program | Fight Against HIV and AIDS*. Www.nascop.or.ke. <https://www.nascop.or.ke/#:~:text=In%202019%2C%20a%20total%20of>

Odhiambo, A. (2020, April 8). *Tackling Kenya’s Domestic Violence Amid COVID-19 Crisis*. Human Rights Watch. <https://www.hrw.org/news/2020/04/08/tackling-kenyas-domestic-violence-amid-covid-19-crisis>

Ridzuan, F. (2022). *A Review on Data Cleansing Methods for Big Data*. Sciencedirect.com. <https://www.sciencedirect.com/science/article/pii/S1877050919318885/pdf?md5=c8d975a00d9baaf0fdbcf1c527ccc96a&pid=1-s2.0-S1877050919318885-main.pdf>

Schober, P., & Vetter, T. R. (2020). Missing Data and Imputation Methods. *Anesthesia & Analgesia*, *131*(5), 1419–1420. <https://doi.org/10.1213/ane.0000000000005068>

UNAIDS. (2014). *UNAIDS report shows that 19 million of the 35 million people living with HIV today do not know that they have the virus*. Www.unaids.org. <https://www.unaids.org/en/resources/presscentre/pressreleaseandstatementarchive/2014/july/20140716prgapreport>

UNAIDS. (2020). *UNAIDS data 2020*. Www.unaids.org. <https://www.unaids.org/sites/default/files/media_asset/2020_aids-data-book_en.pdf>

UNAIDS. (2022). *2022 GLOBAL HIV STATISTICS*. <https://www.unaids.org/sites/default/files/media_asset/UNAIDS_FactSheet_en.pdf>

UNICEF. (2017). *Recent study finds that over 50% of children in Rwanda are victims of sexual, physical or emotional violence*. Www.unicef.org. <https://www.unicef.org/rwanda/press-releases/recent-study-finds-over-50-children-rwanda-are-victims-sexual-physical-or-emotional>

World Health Organization. (2022). *Vulnerable groups and key populations at increased risk of HIV*. World Health Organization - Regional Office for the Eastern Mediterranean. <https://www.emro.who.int/asd/health-topics/vulnerable-groups-and-key-populations-at-increased-risk-of-hiv.html>

# **APPENDIX**

#Reading Data and Extracting Only KENYA Data

library(dplyr)

library(readr)

hivdt <- read\_csv("C:/Users/HP/Downloads/phiacd4.csv")

hivtz <- hivdt[hivdt$Kenya==1,]#subsetting my dataframe to only include rows where the Kenya column is equal to 1

head(hivtz)

colnames(hivtz)

hivtz2 <- select(hivtz,-c(31,32,33)) #remove Country labelled variables

colnames(hivtz2)

hivtz3<-hivtz2 #duplicate datasets

#Number to Factor for hivtz2 dataset

str(hivtz2) #check for type of datatype in columns

numcol<-c(2:10,12,14:15,17:30) #number columns to be factors

hivtz2[numcol]<-lapply(hivtz2[numcol],factor) #convert to categorical variables

str(hivtz2) #check the structure of the dataset

```

```{r}

#Check for NAs

sapply(hivtz2, function(x) sum(is.na(x)))#check for no. of NAs in columns

mean(is.na(hivtz2))#overall missing data proportion

apply(hivtz2,2,function(col)sum(is.na(col))/length(col))#per column missing data proportion

nalist <- colnames(hivtz2)[apply(hivtz2,2,anyNA)]

nalist #list of columns with NAs

```

```{r}

#install.packages("simputation")

#Addressing the NAs

library(visdat)#visualize data

library(naniar)#visualizing and work with missing data

library(simputation)#simple imputation

library(tidyverse)

colnames(hivtz2)

hivtz2tbl <- as\_tibble(hivtz2)

#########MISSING DATA VISUALIZATIONS #########

hivtz2tbl %>% vis\_dat()#types and na distr

```

```{r}

hivtz2tbl %>% vis\_miss()#distr of missing na

```

```{r}

hivtz2tbl %>% gg\_miss\_upset()#upset plot- nas in columns and interaction

```

```{r}

#Fill missing data using mice - 5% max for imputation for per column

library(tidyverse)

library(mice)

#install.packages("mice")

set.seed(5)

# Check for missing values pattern

md.pattern(hivtz2)

hivtz2$`Pregnacy status now`<- as.numeric(hivtz2$`Pregnacy status now`)

# Impute missing values with mean

hivtz2$`Pregnacy status now`[is.na(hivtz2$`Pregnacy status now`)] <- mean(hivtz2$`Pregnacy status now`, na.rm = TRUE)

# Check for missing values after imputation

sum(is.na(hivtz2$`Pregnacy status now`))

###Exporting Dataset

write.csv(hivtz2,"C:/Users/HP/Downloads/filtered phiacd4 (1).xlsx"

```

\*\*DATA ANALYSIS\*\*

```{r message=FALSE, warning=FALSE}

library(readxl)

df<-read\_excel("C:/Users/HP/Downloads/filtered phiacd4 (1) (1).xlsx")

head(df)

numcol2<-c(2:11,13,15:16,18:30) #number columns to be factors

df[numcol2]<-lapply(df[numcol2],factor) #convert to categorical variables

str(df)

colnames(df)

library(gmodels)

library(MASS)

#AGE

CrossTable(df$Age.at.first.sex,df$CD4.category, chisq = TRUE)

#HH ARRANGEMENTS

CrossTable(df$Relationship.with.family.head,df$CD4.category, chisq = TRUE)

CrossTable(df$Respondent.live.in.household,df$CD4.category, chisq = TRUE)

CrossTable(df$Ever.married.lived.together,df$CD4.category, chisq = TRUE)

#EDUCATION STATUS

CrossTable(df$Ever.attended.school,df$CD4.category, chisq = TRUE)

CrossTable(df$Ever.enrolled.in.school,df$CD4.category, chisq = TRUE)

CrossTable(df$Highest.level.of.education,df$CD4.category, chisq = TRUE)

CrossTable(df$Highest.grade.at.that.school.level,df$CD4.category, chisq = TRUE)

#ALCOHOL

CrossTable(df$Alcohol.drink.frequency,df$CD4.category, chisq = TRUE)

#URBAN

CrossTable(df$Urban.area.indicator,df$CD4.category, chisq = TRUE)

#WEALTHQ

CrossTable(df$Wealth.quintile,df$CD4.category, chisq = TRUE)

```

\*\*OBJECTIVE 2: To fit a binary logistic regression model\*\*

```{r}

#Dataset with numeric variables

library(readr)

mine <- df

```

```{r}

head(df)

```

```{r}

#Check and Remove Highly Correlated Columns

library(dplyr)

library(corrr)

library(tidyverse)

minenocd4 <- subset(mine, select = -CD4.category)

#inenocd4<- select(minenocd4, -c(12,14,30))

minenocd4 <- mine[, -c(12, 14, 30)]

```

```{r}

#check unique values for dataset

sapply(minenocd4, function(x) length(unique(x)))

```

```{r}

res.cor<-correlate(minenocd4, method = "pearson", use = "pairwise.complete.obs")

res.cor

res.cor %>% gather(-term, key = "colname", value = "cor") %>% filter(abs(cor)>0.85)

```

```{r}

#check unique values for dataset

sapply(minenocd4, function(x) length(unique(x)))

```

```{r}

library(gtsummary)

df <- select(minenocd4, -Bought.sold.sex.in.the.past.12.months)

head(minenocd4)

```

```{r}

#Model Formulation

modl<-glm(minenocd4$CD4.category ~ Age + Bought.sold.sex.in.the.past.12.months + Whether.ARVs.detected+ Duration.of.time.on.ART + On.ART + LAg..recent.long.term.infection+ Wealth.quintile,data = minenocd4,family = "binomial")

print(summary(modl), signif.stars = TRUE)

modl %>%

tbl\_regression(exponentiate = FALSE)

```

```{r}

# Split data into training and testing sets

train\_index <- sample(nrow(minenocd4), 0.8 \* nrow(minenocd4))

trainer <- minenocd4[train\_index, ]

tester <- minenocd4[-train\_index, ]

```

```{r}

#Final Model

library(gtsummary)

final\_modl <- glm(trainer$CD4.category ~ Relationship.with.family.head + worksicklast3mon + attendedschool + enrolledschool + wrkpaymtlst12mon + marriedorlivedtogether + nopregnancies + pregstatusnw + avoidpregnancy + soughtTBtrtment + timeonART + ARVsdetected, data = trainer,family = "binomial")

print(summary(final\_modl), signif.stars = TRUE)

final\_modl %>%

tbl\_regression(exponentiate = TRUE)

#Odds Ratio

exp(coef(final\_modl))

#Testing

res<-predict(final\_modl,tester, type="response")

res

#Confusion Matrix

table(ActualValue = tester$CD4cat, PredictedValue = res > 0.5)

```

```{r}

#Finding the correct threshold for the model to reduce the false positive rate

#Change res to training dataset

res<-predict(final\_modl,trainer, type="response")

library(ROCR)

ROCRPred <- prediction(res, trainer$CD4cat)#check prediction

ROCRPerf <- performance(ROCRPred,"tpr","fpr")

plot(ROCRPerf,colorize = TRUE,print.cutoffs.at = seq(0.1, by= 0.1))#check performance

#tpr = true positive rate

#fpr = false positive rate

```

```{r}

#Check the threshold

res<-predict(final\_modl,tester, type="response")

table(ActualValue = tester$CD4cat, PredictedValue = res > 0.2)

(180+9)/(180+9+18+42)#using 0.2 = 75.9%

table(ActualValue = tester$CD4cat, PredictedValue = res > 0.3)

(200+7)/(200+7+20+22)#using 0.3 = 83.13%

table(ActualValue = tester$CD4cat, PredictedValue = res > 0.4)

(215+4)/(215+4+23+7)#using 0.4 = 87.95%

#Using 0.4, which does not excessively reduce the efficiency but reduces false positive by one

```

\*\*OBJECTIVE 2: To fit a support vector machine model\*\*

```{r}

# Install and load e1071 package

library(e1071)

```

Before building the SVM model, we need to split the dataset into training and testing sets. We will use 70% of the data for training and 30% for testing. The following code splits the data into training and testing sets:

```{r}

# Set seed for reproducibility

set.seed(123)

library(dplyr)

#check unique values for dataset

sapply(df, function(x) length(unique(x)))

#remove the last variable because it has one level and and two which are not scalable not relevant

df <- select(df, c(12,14,30))

# Split data into training and testing sets

train\_index <- sample(nrow(df), 0.7 \* nrow(df))

train\_data <- df[train\_index, ]

test\_data <- df[-train\_index, ]

```

Now, we can build the SVM model using the 'svm' function from the 'e1071' package. We will use a linear kernel and the default values for other parameters. The code for building the SVM model is as follows:

```{r}

# Build SVM model

svm\_model <- svm(CD4.category ~ ., data = train\_data, kernel = "linear")

```

In the above code, we specified CD4.category as the target variable and used all other variables as predictors. We also specified the kernel as 'linear'.

We can now use the model to make predictions on the test data using the 'predict' function. The code for making predictions is as follows:

```{r}

# Make predictions on test data

svm\_pred <- predict(svm\_model, newdata = test\_data)

```

Finally, we can evaluate the performance of the SVM model using various metrics such as accuracy, precision, recall, and F1 score. Here's the code for calculating these metrics:

```{r}

#subset

test\_data<-test\_data[0:122,]

# Calculate performance metrics

table <- table(svm\_pred, test\_data$CD4.category)

accuracy <- sum(diag(table)) / sum(table)

precision <- diag(table) / colSums(table)

recall <- diag(table) / rowSums(table)

f1\_score <- 2 \* precision \* recall / (precision + recall)

# Print performance metrics

cat("Accuracy:", accuracy, "\n")

cat("Precision:", precision, "\n")

cat("Recall:", recall, "\n")

cat("F1 Score:", f1\_score, "\n")

```

In the above code, we first calculated the confusion matrix using the 'table' function. Then, we calculated accuracy, precision, recall, and F1 score using the confusion matrix. Finally, we printed the performance metrics to the console.

\*\*OBJECTIVE 3: To investigate the factors affecting the CD4 levels on HIV+ women in kenya\*\*

```{r}

library(ggplot2)

library(psych)#describe()

library(reshape2)#from wide format to long format

library(scales)

library(moments)

library(lessR)#barcharts for categorical and normality detailed visual test

library(DT) #generating datatables

library(dplyr)#selecting and sorting

library(tidyverse)

library(ggpubr)#normality graphs

library(gmodels)#Crosstables

```

```{r}

#Chart For Categorical Variables

#Relationship with Head

table(df$Relationship.with.family.head)

prop.table(table(df$Relationship.with.family.head))

BarChart(Relationship.with.family.head, data = df, horiz = TRUE, sort = "-", stat = "count", main = "Relationship With Head",ylab = "Count", xlab = "Type of Relationship with Head")

CrossTable(df$Relationship.with.family.head,df$CD4.category)

```

```{r}

#Time On ART

table(df$Duration.of.time.on.ART)

prop.table(table(df$Duration.of.time.on.ART))

BarChart(Duration.of.time.on.ART, data = df, sort = "-", stat = "proportion", main = "Time on ART for the Repondents",ylab = "Proportion", xlab = "Time Range")

CrossTable(df$Duration.of.time.on.ART, df$CD4.category)

```

```{r}

#ARVs Detected

table(df$Whether.ARVs.detected)#Put it as a statement in paper

prop.table(table(df$Whether.ARVs.detected))

CrossTable(df$Whether.ARVs.detected, df$CD4.category)

```

```{r}

#Categorical Summaries

#YES/NO CATEGORIES##BOOLEAN

dfsickwork <- table(df$Sick.to.work.last.three.months)

dfgosch <- table(df$Ever.attended.school)

dfenrollsch <- table(df$Ever.enrolled.in.school)

dfworkpay <- table(df$Work.for.payment.in.last.12.months)

dflivetogether <- table(df$Ever.married.lived.together)

dfavoidpreg <- table(df$Avoiding.pregnancy)

dftbtreat <- table(df$Ever.sought.TB.treatment)

```

```{r}

df.cat2 <- rbind(dfsickwork,dfgosch,dfenrollsch,dfworkpay,dflivetogether,dfavoidpreg,dftbtreat)

df.cat2

rownames(df.cat2)<-c("Sick to Work Last 3 Months","Ever Attended School","Ever Enrolled in School","Worked for Pay in Last 12 Months","Ever Married/Lived Together","Avoiding Pregnancy","Ever Sought TB Treatment")

colnames(df.cat2)<-c("No","Yes")

#Transform data to long format

long <- melt(df.cat2, id.vars = c("No", "Yes"))

long

colnames(long)<-c("Variable","Condition","Value")

colnames(long)

# Grouped barplot using ggplot2

Variable <- long$Variable

Value <- long$Value

Condition <- long$Condition

x<-Value/1242

ggplot(long,

aes(x = Variable,

y = x,

fill = Condition,

label = scales::percent(x))) +

geom\_bar(stat = "identity",

position = "dodge") +

scale\_y\_continuous(labels = function(x) paste0(x\*100, "%")) +

labs(x = "Variable", y = "Frequency (%)", title = "Boolean Categorical Variables") +

theme\_classic() +

geom\_text(position = position\_dodge(width = .9), # move to center of bars

vjust = -0.5, # nudge above top of bar

size = 1.9) +

coord\_flip()

```

```{r}

#NUMERICAL DATA#################

dfpreg<-describe(as.numeric(df$Number.of.pregnancies))

dfpreg

datatable(dfpreg)

#value of the Shapiro-Wilk Test is greater than 0.05, the data is normal. If it is below 0.05, the data significantly deviate from a normal distribution

shapiro.test(as.numeric(df$Number.of.pregnancies))

#Density Plot

ggdensity(as.numeric(df$Number.of.pregnancies),

main = "Density plot of Number of Pregnancies",

xlab = "Number of Pregnancies")

#CROSSTABLE WITH CD4

library(dplyr)

library(table1)

colnames(df)

df3 <- select(df, c(2,4,5,6,9,10,11,13,15,26,21,23))

colnames(df3)

```

```{r}

head(df3)

```

```{r}

labels <- list(variables=list(Relationship.with.family.head = "Relationship With Head",

Sick.to.work.last.three.months = "Sick to Work last 3 Months",

Ever.attended.school= "Ever Attended School",

Ever.enrolled.in.school = "Ever Enrolled in School",

Work.for.payment.in.last.12.months= "Work for Pay",

Ever.married.lived.together = "Married/Live Together",

Number.of.pregnancies = "Number of Pregnancies",

Avoiding.pregnancy = "Ever Avoided Pregnancy",

Ever.sought.TB.treatment = "Ever Sought TB Treatment",

Ever.had.sexual.intercourse = "Duration on ART",

Whether.ARVs.detected ="ARVs Detected"),

groups=list("", "CD4 Level"))

levels(df$CD4.category) <- c("High", "Low")

strata <- c(list(Total = df), split(df, df$CD4.category))

dftbl3 <- table1(strata,

labels,

groupspan=c(1,2),

rowlabelhead = "Characteristics",

overall = "Total",

caption = "CD4 Levels against the Characteristics",

footnote = "CD4 Levels against the Significant Variables",

data = df)

print(dftbl3)

```