2022-07

Machine learning model for prediction and visualization of HIV index testing in northern Tanzania

Chikusi, Happyness

NM-AIST

https://doi.org/10.58694/20.500.12479/1591

MACHINE LEARNING MODEL FOR PREDICTION AND VISUALIZATION OF HIV INDEX TESTING IN NORTHERN TANZANIA

Happyness S. Chikusi

A Project Report Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Embedded and Mobile Systems of the Nelson Mandela African Institution of Science and Technology

Arusha, Tanzania

July 2022

ABSTRACT

Infection with the human immunodeficiency virus and acquired immunodeficiency syndrome (HIV/AIDS) continue to pose a threat to Tanzanian society. Various tactics have been used to improve the number of persons who are aware of their HIV status. Index testing stands out among these methods as the most effective way to count the number of HIV contacts who may be at risk of catching HIV from HIV-positive individuals. The current HIV index testing, however, is manual, which presents a number of difficulties, including inaccuracies, is time-consuming, and is expensive to operate. In order to forecast and depict HIV index testing, this study presents the findings of the machine-learning model. The software development procedure was in accordance with agile software development principles. The regions of Kilimanjaro, Arusha, and Manyara in Tanzania are where the data was gathered which consisted of 11 features and 6346 samples. The dataset was then separated into training sets with 5075 samples each and testing sets with 1270 samples (80/20). The datasets were subjected to the methods Random Forest (RF), XGBoost, and Artificial Neural Networks (ANN). Random forest MAE (1.1261), XGBoost MAE (1.2340), and ANN MAE (1.1268) were the three results obtained. Random forest algorithms had the lowest mean absolute errors (MAE). Therefore, RF appearing to have the highest performance when compared to the other two algorithms. In comparison to men (17.4%), data visualization reveals that females are more likely to test for HIV and to name their partners (82.6%). Additionally, there were higher instances of persons listing and mentioning their partners in the Kilimanjaro region. This work helped us realize the importance of machine learning in predicting and visualizing HIV index tests in general. The created model can help decision-makers build a viable intervention to stop the spread of HIV and AIDS in our communities. The report suggests that health centers in other areas employ this concept to make their work more straightforward.

DECLARATION

I, Happyness S. Chikusi, do hereby declare to the Senate of Nelson Mandela African Institution of Science and Technology that this project report is my original work and that it has neither been submitted nor being concurrently submitted for degree award in any other institution.

Happyness S. Chikusi 21.07.2022

Candidate Name **Signature** **Date**

The above declaration is confirmed by:

Dr. Judith Leo JLeo.

Name of Supervisor 1 **Signature** **Date**

Prof. Shubi Kaijage 5-8-2022

Name of Supervisor 2 **Signature** **Date**

COPYRIGHT

CERTIFICATION

The undersigned certify that they have read and hereby recommend for acceptance by The Nelson Mandela African Institution of Science and Technology, a project report titled**, "Machine Learning Model for Prediction and Visualization of HIV Index Testing in Northern Tanzania",** in partial fulfillment of the requirements for the degree of Master of science in Embedded and Mobile Systems of the Nelson Mandela African Institution of Science and Technology.

Dr. Judith Leo Leo.

Name of Supervisor 1 **Signature** **Date**

Prof. Shubi Kaijage 5-8-2022

Name of Supervisor 2 **Signature** **Date**

ACKNOWLEDGEMENTS

First, I thank Jesus Christ, my Lord, and savior of my life, for His mercy and love. It is by his grace the accomplishment of this study was possible.

Secondly, the Nelson Mandela African Institution of Science and Technology is to be commended for giving me the opportunity to pursue my master's degree, and I want to express my gratitude for that. Thanks to Soft Med Company for allowing me to perform an internship there and for giving me a project to complete in order to earn my project report, as well as the Center of Excellence in ICT in East Africa (CENIT@EA) for funding my studies.

Thirdly, I want to convey my sincere gratitude to Dr. Judith Leo and Prof. Shubi Kaijage who served as my mentors and supervisors during this study. Their support and guidance assisted me in shaping myself into a research scientist.

Finally, I am grateful to my parents, Mr. and Mrs. Samwel Chikusi, my lovely husband, Mr. Pastory Msanya, my son, Gabriel Pastory Msanya, my sisters, and brother, as well as my closest friends for their moral, material support and prayers throughout my entire academic career.

I pray to Almighty God to grant you a precious life to all of you.

DEDICATION

I dedicate this project report to my lovely husband, Mr. Pastor Msanya, and my dearer son, Gabriel Msanya. I always feel blessed to be with you in my life. Therefore, I dedicate this to you to symbolize unity and infinite love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

AIDS	Acquired Immune Deficiency Syndrome
ANN	Artificial Neural Network
CD_4	Cluster of differentiation 4
CENIT@EA	Center of Excellence of Information Communication and Technology in East Africa
COP	Country Operational Plan
CSS	Cascading Style Sheets
HIV	Human Immunodeficiency Virus
HTML	Hypertext and Markup Language
ICT	Information Communication and Technology
MAE	Mean Absolute Error
ML	Machine Learning
MSE	Mean Squared Error
NM-AIST	Nelson Mandela African Institution of Science and Technology
PEPFAR	President's Emergency, Plan for AIDS Relief
PLHIV	People Living with Human Immunodeficiency Virus
RF	Random Forest
RMSE	Root Mean Squared Error
SAP	System Application and Product
UNAID	Joint United Nations Programme on HIV/AIDS
USA	United State of America
VCT	Voluntary HIV counseling and Testing
WHO	World Health Organization

CHAPTER ONE

INTRODUCTION

1.1 Background of the Problem

Human Immunodeficiency Virus (HIV) infection is the main cause of the illness known as acquired immunodeficiency syndrome (AIDS). Acquired immunodeficiency syndrome is caused by significant immune system impairment, which leaves the body exposed to various diseases (KHAN & KHATTAK, 2006).

Global public health is at risk from the viral disease HIV. World Health Organization estimates that 38% of persons worldwide are HIV positive. But only 19% of people are aware of their situation (UNAIDS, 2020) With an estimated 68% in sub-Saharan Africa, many people with HIV reside in middle- and low-income nations.

Human rights and equity are addressed in the WHO Strategy for 2016–2021 along with a dramatic drop in new HIV infections and a reduction in mortality. The goal is to eliminate HIV as a public health issue by 2030 and reduce new infections to less than five hundred thousand by 2020. The current goal is 90 90 90, which means that 90% of people are aware of their situation, 90% of them receive appropriate medical attention, and the remaining 90% continue to improve their health (Bain *et al.*, 2017; United Nations Joint Programme on HIV/AIDS (UNAIDS), 2014).

In the southern region of Africa, a number of nations have made significant progress in the HIV/AIDS Program, ensuring that 90% of persons with HIV are aware of their status. The main step towards reaching the Joint United Nations Program on HIV/AIDS aim of 90 90 90 is HIV testing and counseling. However, for 2025, the goal is 95%, 95%, 95 %

Index testing is a case-finding technique used to identify exposed contacts of Human Immunodeficiency Virus (HIV) positive people who should be offered HIV testing. Alternatively called partner notification (Jeren and Abebe, 2017). The new person who has been diagnosed as HIV positive and is engaged in HIV treatment is the fundamental component of the index testing strategy. The term "index client" refers to this person. Counselors and healthcare professionals ask index clients to list all of their relationships, including those who use drugs intravenously or sexually, as well as their children. Figure 1 depicts well the information. The procedure is optional and private.

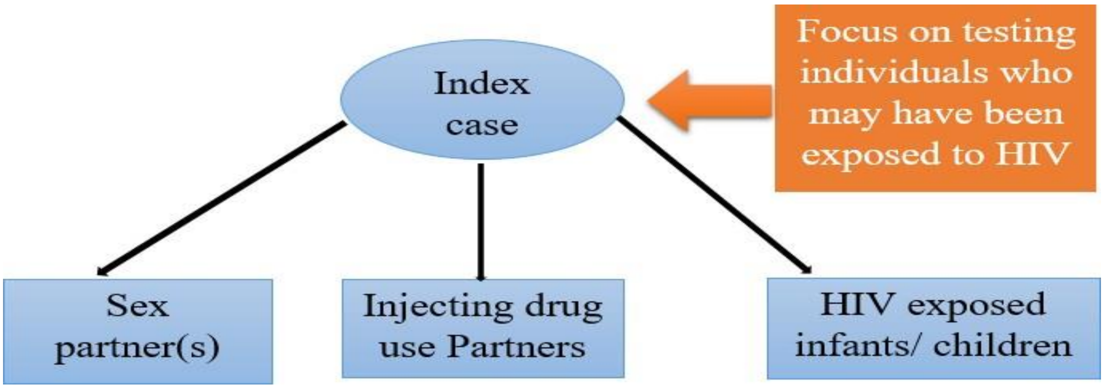

Focus on testing individuals who may have been exposed to HIV
Index case
Sex partner(s)
Injecting drug use Partners
HIV exposed infants/ children

Figure 1: **Index case testing** (Pepfar, 2015)

Each partner and child are contacted, told of the HIV exposure, and given the option to test voluntarily. The goal of index testing is to stop the HIV transmission cycle (PEPFAR, 2015b). Additionally, if the index client's result status is positive, they will be connected to treatment, and if it is negative, they will be connected to preventative services.

Various HIV testing methods exist, including home-based community VCT, mobile testing, outreach testing, and voluntary HIV counseling and testing (VCT). Mobile and home-based outreach, however, is expensive. In order to enhance the number of people who are aware of their status, index case testing was established, and it is the best method available.

In order to increase the ability of policymakers to plan, prioritize, and implement effective interventions, this study uses machine learning approaches to forecast the number of HIV index tests and uses the data for visualization items of age, sex, location, and kind of relationship. Additionally, the index client can support them in changing their behavior, improving their capacity for self-care, and becoming more dedicated to treatment with hope and empowerment.

1.2 Statement of the Problem

The data are manually calm for the present HIV index client testing system, which lacks an automated method. As a result, it is difficult to forecast HIV index testing from data analysis. In addition, skilled data entry workers and data analysts are needed for the job. Therefore, getting the desired result will cost more and take more time. Additionally, it cannot be avoided that human error will occur.

There is no machine-learning model for making predictions of HIV index testing or application for making data visualization in order to know what factors may prevent or influence an HIV-positive client to list his or her partners, despite the fact that HIV index is the leading modality in finding and increasing the number of people to know their status.

The suggested machine-learning model and visualization application will assist experts in making predictions and generating current, legible, and intelligible data visualization. In order to effectively plan an intervention strategy to eradicate HIV/AIDS as a public health issue, the machine-learning model will aid in predicting the number of HIV index testing.

1.3 Rationale of the Project

This study focused on identifying and predicting number of HIV index testing and data visualization for given HIV client. Thus, the study did not consider ranking according to their probability of HIV status or forecasting HIV trend in the future.

1.4 Project Objectives

1.4.1 Main Objective

To develop a machine-learning model for predicting HIV index testing in northern Tanzania.

1.4.2 Specific Objectives

(i) To identify the requirement for developing HIV index testing model.

(ii) To develop the model that will accurately predict the number of HIV index testing.

(iii) To evaluate the performance of the HIV index testing model.

(iv) To deploy machine learning model for HIV index testing to web app to be used in real world.

1.5 Research Questions

(i) What information is needed to create a machine-learning model that can predict the frequency of HIV index testing?

(ii) What algorithms will be used to create a machine-learning model that predicts the frequency of HIV index testing?

(iii) How well does the created ML model for HIV index testing perform?

(iv) Which technology will be applied in the actual world to implement the HIV index-testing model?

1.6 Significance of the Study

With the developed machine learning model and visualization app timely and, easy production of report can be achieved. The proper usage of the developed application and the model can significantly save the financial and time cost to health workers and HIV stakeholders. This is due to the fact that, health workers must not have to find data scientist expertise for data analysis and making predictions.

Moreover, the developed machine-learning model and application platform will help all other HIV/AIDS stakeholders such as policy makers, the government, as well as HIV/AIDS related Non-Government Organizations (NGOs) in making decision for ending up HIV as a health threat by 2030.

1.7 Delineation of the Study

In order for Tanzania and other Sub-Saharan nations to achieve sustainable development through sustaining the population's good health, the HIV/AIDS crisis must be effectively addressed. In order to identify and predict people who are at risk of contracting HIV/AIDS early on and to aid health professionals and policymakers in making informed decisions and initiating early intervention, the use of machine learning models that are trained using local datasets is desired by taking into account end user interaction. These methods will make it easier to determine the extent to which variables influence HIV index testing, allow health stakeholders without any prior machine learning experience to readily interact with, and make use of large amounts of data. Therefore, the study attempted to develop a machine learning model approach using the dataset from northern Tanzania in order to facilitate an intervention program and create HIV/AIDS awareness in the society.

CHAPTER TWO

LITERATURE REVIEW

2.1 Overview of Literature Survey

The purpose of this chapter is to go through different literature reviews related to this study. This literature review was collected from various sources such as journals, conference articles, books and official reports. Several related works were reviewed to find out different information on this study. For instance, the status of HIV/AIDS in Tanzania, the application of machine learning in health care, the machine learning application on making a prediction of HIV index testing as well as other related research work already done.

2.2 Status of HIV in Tanzania

According to Tanzania's HIV statistics, there are 1.7 million HIV-positive individuals living there, 77 000 new HIV infections, and 27 000 deaths from AIDS (COP, 2019). Therefore, for the HIV program to be successful and effectively has to yield 95 95 95 by 2025. This means that 95% know their HIV status, 95% are in treatment, and 95% on ART are virally suppressed. Figure 2 illustrates the responsiveness of the 90 90 90 Target in Tanzania.

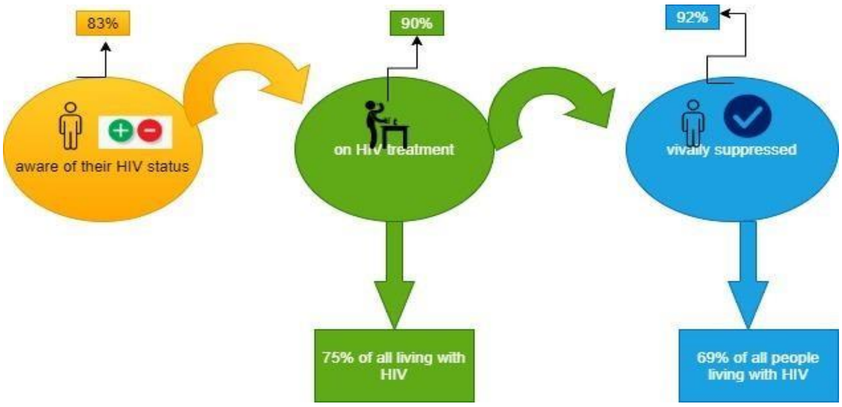

83%
aware of their HIV status
90%
on HIV treatment
92%
vivally suppressed
75% of all living with HIV
69% of all people living with HIV

Figure 2: Tanzania's progress towards 90% 90% 90% (Avert, 2020)

In addition, the HIV status in Tanzania shows new infection occurs in the context of stable heterosexual relationships 38.8%, causal heterosexual 28.9%, sex workers 1.3% sexual worker’s clients 8.7%, partners of sex workers clients 3.3%, partner of people engaged in casual sex 7.6%, people who inject drugs 2.1% and men having sex with men 6.8% (National Bureau of Statistics, 2018).

In order to identify People Living with HIV (PLHIV), the Ministry of Health is aiming to include partner notification services and index testing as national strategies from 2018 to 2022 as reported by Gwajima and Wright. (2021). A promising method for locating new HIV cases in Tanzania is index case testing. This plan will help Tanzania reach its first goal of 90 HIV cases detected (from 2017 to 2022) and its second goal of 95 cases detected (for both males, adolescents, and children) by 2025.

2.3 Machine Learning in Health Care

The usage and development of computer systems that can learn and adapt without explicitly following instructions are known as machine learning (ML). These systems utilize algorithms to analyze and derive conclusions from patterns in data (SAP Insights, 2019). The creation of features for machine learning algorithms relies on the domain expertise of the data. Computer vision, automatic speech recognition, business analysis, natural language processing, and health care are a few of the disciplines where machine learning has been applied. To achieve the greatest results, features must be able to extract pertinent data from enormous amounts of varied data, which requires a lot of time and work.

When used with high-quality data, machine learning approaches reliably anticipate outcomes in a variety of applications, including drug discovery and disease diagnosis. Diabetes, autism, subtyping, and cancer detection are all areas of interest for machine learning in the medical field. ML was also used to forecast the cholera outbreak (Leo *et al.,* 2019).

In terms of HIV/AIDS, machine learning was used as follows: HIV medication resistance prediction (Raposo *et al.,* 2020). In order to monitor the evolution of human immunodeficiency, machine learning is used to anticipate each patient's individual present CD4 cell count (Singh *et al.,* 2013). Internet search forecasting of new HIV infections in China (Zhang *et al.,* 2018), Machine learning for identification of persons who are at high risk of contracting AIDS (Balzer *et al.,* 2019). Also ML in improving HIV case detection (Smyrnov *et al.,* 2016). Despite of applying machine learning in HIV /AIDS. The authors only developed a model for the prediction of HIV status, where a person is positive or negative and did not predict how many people will be at risk of contracting HIV for that specific person.

Other related works use community testing to predict HIV index testing in order to improve HIV case discoveries among men. Chi-Square test estimated index cascade, and descriptive statistics were employed in the procedure (Mwango *et al.,* 2020). High-volume index testing for HIV utilizing a services registry and Microsoft Excel (Mahachi *et al.*, 2019). Another study focused on using machine learning for planning HIV/AIDS diagnosis and treatment (Prabhakaran, 2014). However, in these studies, authors used the methods like Chi-Square that are used for the testing relationship between categorical variables. In addition, Mahach (2019) used Microsoft excel which has limitations such as lack of control, excel run slowly when data file is large; Excell is susceptible to human error, and difficulties in troubleshooting.

Therefore, the goal of this work is to create a flexible, user-friendly, and error-free model to predict HIV index testing and visualization items of age, sex, location, and kind of relationship. The visualization tool will also improve the capacity to organize, prioritize, and carry out a successful intervention. Table 1 shows the summary of literature review on various Machine-learning algorithms, The Author of paper, area where they had been applied and their application.

Table 1: Literature review summary

Author	Application	Ml Model	Area
Leo *et al*. (2020)	A reference Machine learning model for prediction of cholera based on weather changes	XGBoost, Decision tree, Support Vector Machine, K-Nearest Neighbors', Linear regression	Tanzania
Raposo *et al.* (2020)	HIV medication resistance prediction	Random forest	Rwanda
Singh *et al*. (2013)	Machine learning is used to anticipate each patient's individual present CD4 cell count.	Linear regression, XGBoost	China
Zhang *et al.* (2018)	Internet search forecasting of new HIV infections in China	Logistic regression random forestry	China
Smyrnov *et al.* (2016)	ML in improving HIV case detection	Sampling technique	Tanzania
Mwango *et al.* (2020)	The community testing to predict HIV index testing in order to improve HIV case discoveries among men		Zambia
Mahachi *et al*. (2019)	High-volume index testing for HIV utilizing a services registry	Microsoft Excel	Zimbabwe

CHAPTER THREE

MATERIALS AND METHODS

3.1 Introduction

This chapter covers the different tools, technologies and methods, which have been employed to undertake this study. It describes the area of study, data required for model development, machine learning life cycle, data visualization, algorithms for machine learning model development procedures, model evaluation, model deployment, and testing. The approach and techniques described here were applied to ensure not only that study was carried out at the expected level but also the developed application meet the user's expectations.

3.2 Materials

3.2.1 Area of Study

The area of study selected was the northern part of Tanzania as shown in Fig. 3. Northern regions are Tanga, Kilimanjaro, Arusha, and Manyara. Arusha, Manyara, and Kilimanjaro had been selected to represent the group. The selection was because regions of Arusha, Kilimanjaro and Manyara have low rate of HIV prevalence of 2.2%, 3.1% and 2.7% respectively (National Bureau of Statistics, 2018). In addition, due to the sensitivity of HIV data, we only managed to get the data from mention area with the help of the hosting company.

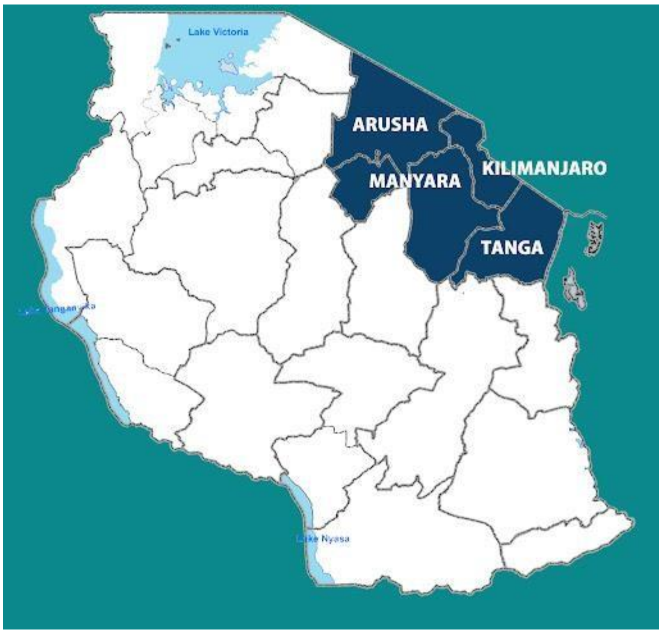
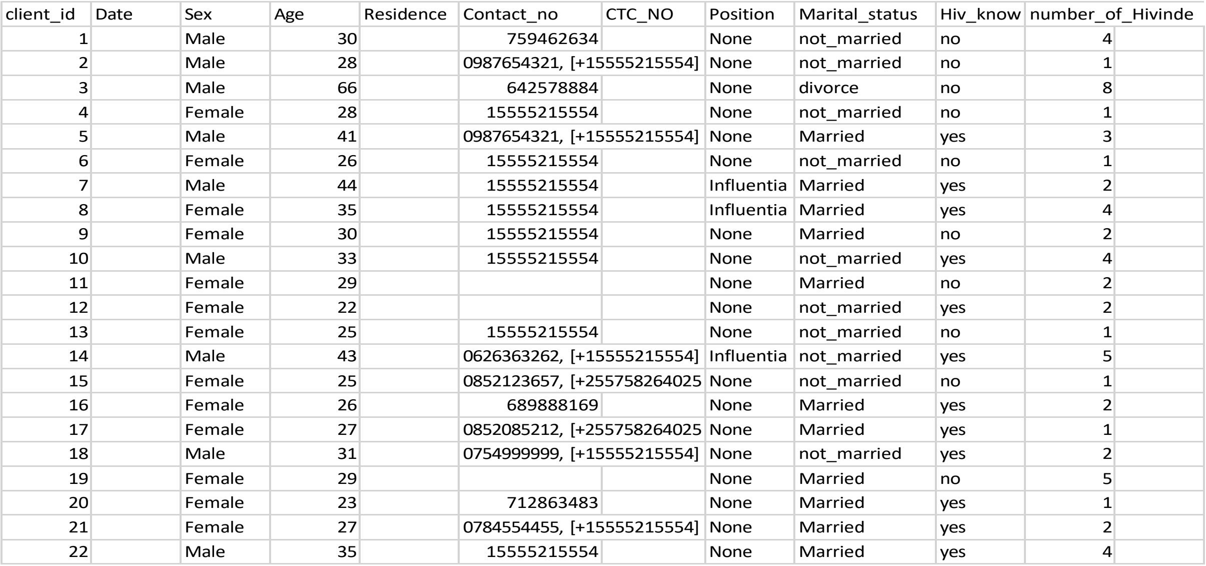


Figure 3: Map of Tanzania showing northern regions (Tawasanet, 2019)

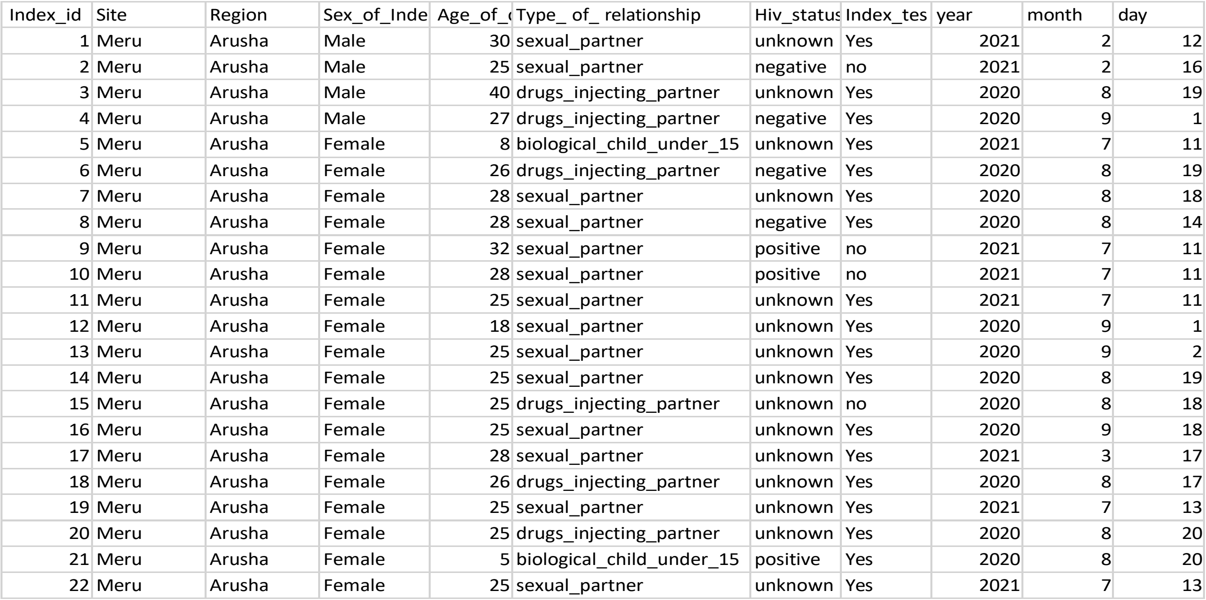
3.2.2 Data used for Model Development

The data for this study came from several health facilities and community locations in the cities of Arusha, Kilimanjaro, and Manyara. As indicated, the client information is made up of 6346 samples and 11 features, while the index-client data is made up of 7228 samples and 11 features as shown in Fig. 4 and Fig. 5.



client_id	Date	Sex	Age	Residence	Contact_no	CTC_NO	Position	Marital_status	Hiv_know	number_of_Hivinde	
1		Male	30		759462634		None	not_married	no	4	
2		Male	28		0987654321, [+15555215554]		None	not_married	no	1	
3		Male	66		642578884		None	divorce	no	8	
4		Female	28		15555215554		None	not_married	no	1	
5		Male	41		0987654321, [+15555215554]		None	Married	yes	3	
6		Female	26		15555215554		None	not_married	no	1	
7		Male	44		15555215554		Influentia	Married	yes	2	
8		Female	35		15555215554		Influentia	Married	yes	4	
9		Female	30		15555215554		None	Married	no	2	
10		Male	33		15555215554		None	not_married	yes	4	
11		Female	29				None	Married	no	2	
12		Female	22				None	not_married	yes	2	
13		Female	25		15555215554		None	not_married	no	1	
14		Male	43		0626363262, [+15555215554]		Influentia	not_married	yes	5	
15		Female	25		0852123657, [+255758264025		None	not_married	no	1	
16		Female	26		689888169		None	Married	yes	2	
17		Female	27		0852085212, [+255758264025		None	Married	yes	1	
18		Male	31		0754999999, [+15555215554]		None	not_married	yes	2	
19		Female	29				None	Married	no	5	
20		Female	23		712863483		None	Married	yes	1	
21		Female	27		0784554455, [+15555215554]		None	Married	yes	2	
22		Male	35		15555215554		None	Married	yes	4	

Figure 4: Client information original dataset before preprocessing



Index_id	Site	Region	Sex_of_Inde	Age_of_	Type_ of_ relationship	Hiv_status	Index_tes	year	month	day
1	Meru	Arusha	Male	30	sexual_partner	unknown	Yes	2021	2	12
2	Meru	Arusha	Male	25	sexual_partner	negative	no	2021	2	16
3	Meru	Arusha	Male	40	drugs_injecting_partner	unknown	Yes	2020	8	19
4	Meru	Arusha	Male	27	drugs_injecting_partner	negative	Yes	2020	9	1
5	Meru	Arusha	Female	8	biological_child_under_15	unknown	Yes	2021	7	11
6	Meru	Arusha	Female	26	drugs_injecting_partner	negative	Yes	2020	8	19
7	Meru	Arusha	Female	28	sexual_partner	unknown	Yes	2020	8	18
8	Meru	Arusha	Female	28	sexual_partner	negative	Yes	2020	8	14
9	Meru	Arusha	Female	32	sexual_partner	positive	no	2021	7	11
10	Meru	Arusha	Female	28	sexual_partner	positive	no	2021	7	11
11	Meru	Arusha	Female	25	sexual_partner	unknown	Yes	2021	7	11
12	Meru	Arusha	Female	18	sexual_partner	unknown	Yes	2020	9	1
13	Meru	Arusha	Female	25	sexual_partner	unknown	Yes	2020	9	2
14	Meru	Arusha	Female	25	sexual_partner	unknown	Yes	2020	8	19
15	Meru	Arusha	Female	25	drugs_injecting_partner	unknown	no	2020	8	18
16	Meru	Arusha	Female	25	sexual_partner	unknown	Yes	2020	9	18
17	Meru	Arusha	Female	28	sexual_partner	unknown	Yes	2021	3	17
18	Meru	Arusha	Female	26	drugs_injecting_partner	unknown	Yes	2020	8	17
19	Meru	Arusha	Female	25	sexual_partner	unknown	Yes	2021	7	13
20	Meru	Arusha	Female	25	drugs_injecting_partner	unknown	Yes	2020	8	20
21	Meru	Arusha	Female	5	biological_child_under_15	positive	Yes	2020	8	20
22	Meru	Arusha	Female	25	sexual_partner	unknown	Yes	2021	7	13

Figure 5: Index information before preprocessing

3.2.3 Tools and Technology

The study used python programing language. The reason for selecting this language is that it is open-source and offers various libraries supporting machine learning. Others are HTML5, CSS, source code editor, Jupiter notebook, flask framework, streamlit, and Heroku server. Heroku is a platform as a service that enables programmers to create, launch, and manage cloud-based apps. In addition, it is scalable across all programming languages.

3.3 Methods

3.3.1 Agile Specifically Evolutionary Prototyping

The selection of this methodology is due to the following reasons: The model is suitable for interactive refine at each step; it saves time and effort and is used in Artificial Intelligence (AI) and machine learning applications. Then followed Machine-learning life cycle.

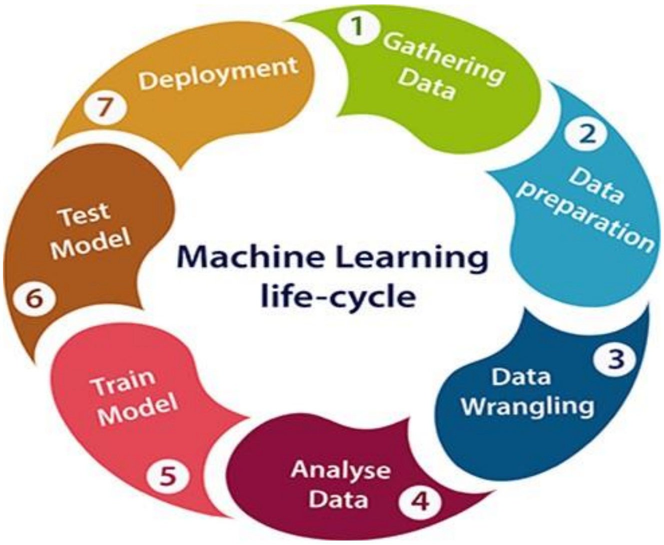


Figure 6: Machine-learning life cycle

Building machine-learning projects involve a cyclical process called the machine-learning life cycle. The life cycle's primary goal is to find a solution for the project. Figure 6 shows the stages of the life cycle, and Table 2 explains the stages in detail.

Table 2: Summary of main stages of machine-learning life cycle

Name	Explanation
Gathering data	All stages are built upon it. Understanding the subject requires study to get the necessary data and have domain knowledge.
Data preparation	In this stage, we identify and select the data set. It helps to understand the data that we have to work with, Understanding the characteristics, quality and data format
Data wrangling	During this phase, raw data were cleaned and transformed into a usable. It has data reduction, cleaning, and sampling. Outliers and skewed data are removed in order to improve the quality of the data.
Analyze data	In this step, the data is analyzed, and determine the type of the data so as to choose the right techniques,
Train the model	In this step, we train the model with a variety of machine-learning techniques using datasets. A model must be trained in order for it to comprehend the numerous patterns, laws, and features.
Test the model	Focus on checking or evaluating the model.
Deployment	It is the action of applying the model.

3.3.2 Data Gathering

We first focused on literature review and contacting focus group discussions with the key stakeholders to understand the problem. The aim of contacting literature review was to identify the gap from other researchers' work to understand the topic better and come up with the best solution to answer the research questions. Then, we carried out the focus group discussions to extract knowledge regarding HIV index testing from the stakeholder. Peoples participated in focus group discussion were 10, 2 health workers at the facility from each region and 4 technical persons at hosting company. The guiding questions are attached (Appendix 3).

3.3.3 Stage of Data Preprocessing

Data selection, cleansing, integration, reduction, and transformation are some of the techniques used in the preparation of data. Two different types of datasets were included in the data obtained from Kilimanjaro, Arusha, and Manyara. The first dataset included client information and a client index. According to the data's nature and the results of the literature review, the following preprocessing methods were used.

The first stage of data preprocessing was data selection. The dataset had other data concerning tuberculosis (TB). The dataset was selected only for HIV information and left the others.

The second stage-involved data cleaning, which involves fixing errors and removing incomplete, inaccurate, and inappropriate portions of the dataset. Not only that, but also the cleaning process involved ignoring features with no values. There were no misspelling errors in the cleaning process and duplicated were identified and cleared. Data cleaning steps were described in Table 3, and steps were represented well.

Table 3: Data cleaning steps

Steps	Explanation
Data Analysis	Detecting noise and dirty data by looking at the dataset and data quality.
Workflow definition	Provide an explanation of the cleaning rules, which use maximizing to handle missing values.
Execute	Set out the results and provide the orders for the dataset processing.
Verification	To ensure the accuracy and efficiency of the cleaning rules to users, literature, and the responsible offices.

Data integration was the third stage of data preprocessing, entails combining data from several sources into a coherent dataset. The integration process adhered to conventional measures consistency. Utilizing sophisticated Microsoft Excel procedures, all files were integrated.

Data reduction was the fourth stage of the data pretreatment procedure. The dataset's integrity was not compromised during data reduction. For example, the following features were diminished: Client_id, contact number, CTC number, date, and residence. The reduction process was done due to the following reasons ClientID was removed because we have already worked with it to get a number of HIV indexes for client. A residence was reduced because it had no value. The fourth stage of data preprocessing was data transformation. All the data were transformed into a format that was acceptable for analysis and model creation. It involved converting binary data into 1 and 0 correspondingly, and categorical data into numerical values.

3.3.4 Visualizing Data

In order for a user to gain an understanding from the studied data and make the best decision possible, it must be presented graphically. This is known as data visualization (Chawla et al., 2018; Khalid, 2021). The web app for data visualization was done by using streamlit, which is an open-source app framework in python language (Streamlit, 2022).

3.3.5 Machine Learning Algorithms

Figure 7 illustrates approaches to solving ML problems. Supervised learning, unsupervised learning, and reinforcement learning are the three main categories of machine learning. Depending on the problem under which category will follow (Rashidi *et al.*, 2019).

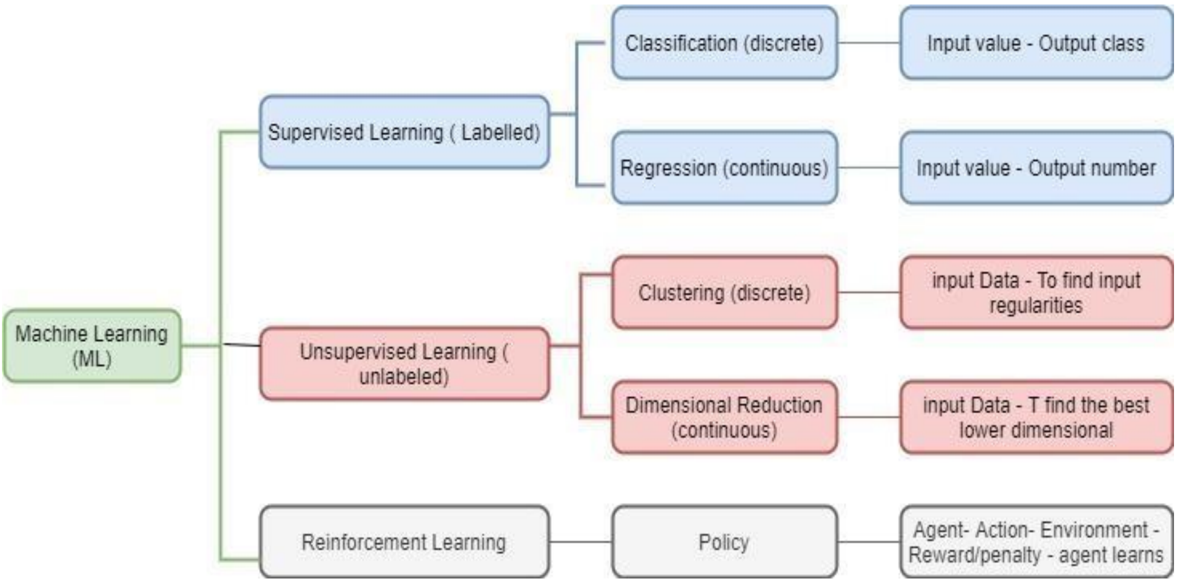
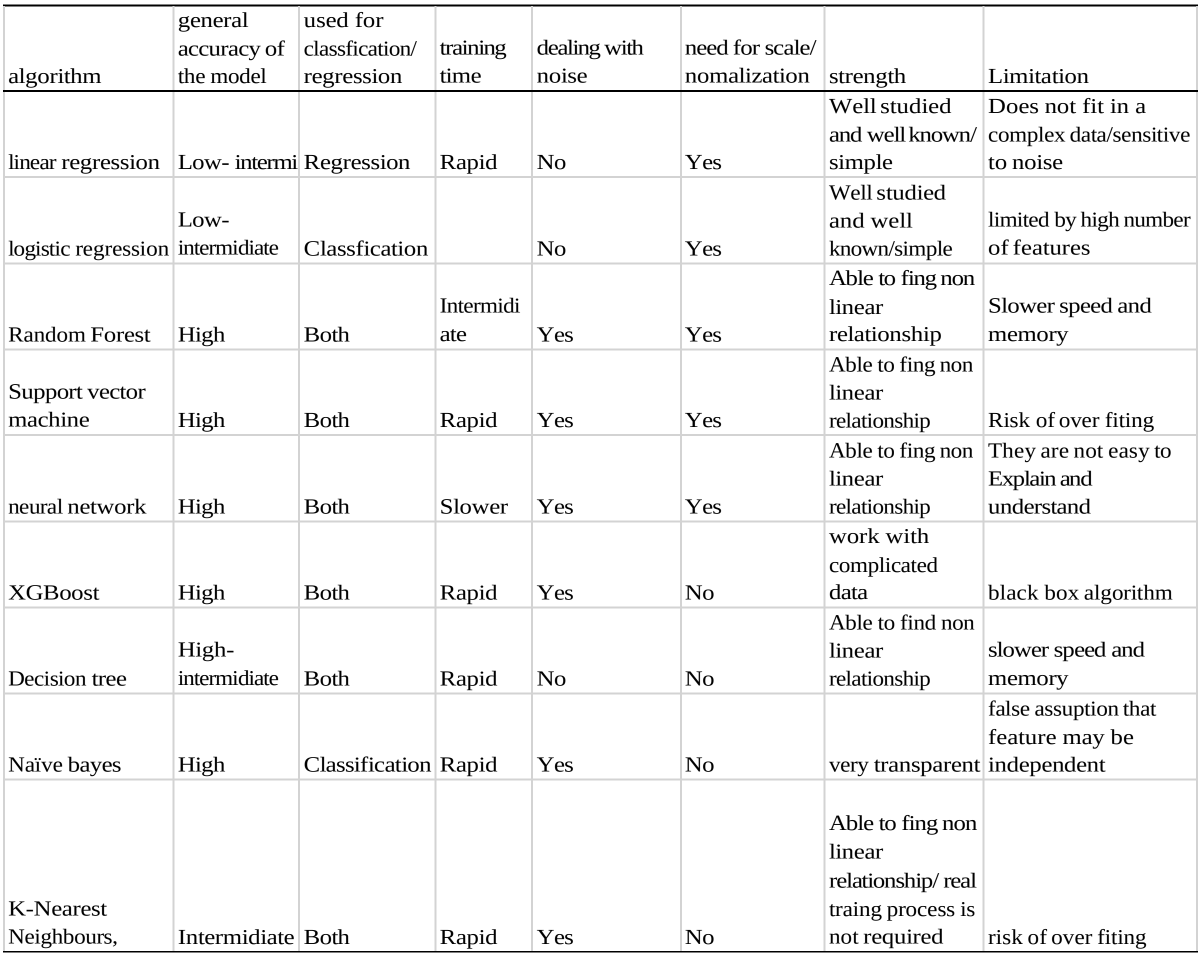

Machine Learning (ML)
Supervised Learning (Labelled)
Classification (discrete)
Input value - Output class
Regression (continuous)
Input value - Output number
Unsupervised Learning (unlabeled)
Clustering (discrete)
input Data - To find input regularities
Dimensional Reduction (continuous)
input Data - T find the best lower dimensional
Reinforcement Learning
Policy
Agent- Action- Environment - Reward/penalty - agent learns

Figure 7: Framework for machine learning (Rashidi et al., 2019)

In machine learning, there is no a single algorithm that excels at solving every challenge. Instead, there are other things to think about when deciding which strategy is ideal for a particular circumstance. Choosing needs having a basic awareness of the types, weaknesses, and strengths of machine learning algorithms.

In general, picking a machine learning method depends on a number of variables, such as the size, the structure, the kind, and the learning strategy (supervised vs unsupervised), as well as the necessity for accuracy and speed in data processing. The following algorithms fall under supervised learning: support vector machines, neural networks, linear regression, logistic regression, random forests, XGBoost, decision tree, Naïve Bayes, and K-Nearest Neighbors.

Figure 8 describes the summary of the algorithm used in supervised ML learning, showing each algorithms' application, strength, and limitation.



algorithm	general accuracy of the model	used for classfication/ regression	training time	dealing with noise	need for scale/ nomalization	strength	Limitation
linear regression	Low- intermi	Regression	Rapid	No	Yes	Well studied and well known/ simple	Does not fit in a complex data/sensitive to noise
logistic regression	Low-intermidiate	Classfication		No	Yes	Well studied and well known/simple	limited by high number of features
Random Forest	High	Both	Intermidiate	Yes	Yes	Able to fing non linear relationship	Slower speed and memory
Support vector machine	High	Both	Rapid	Yes	Yes	Able to fing non linear relationship	Risk of over fiting
neural network	High	Both	Slower	Yes	Yes	Able to fing non linear relationship	They are not easy to Explain and understand
XGBoost	High	Both	Rapid	Yes	No	work with complicated data	black box algorithm
Decision tree	High-intermidiate	Both	Rapid	No	No	Able to find non linear relationship	slower speed and memory
Naïve bayes	High	Classification	Rapid	Yes	No	very transparent	false assuption that feature may be independent
K-Nearest Neighbours,	Intermidiate	Both	Rapid	Yes	No	Able to fing non linear relationship/ real traing process is not required	risk of over fiting

Figure 8: Comparison of commonly used supervised algorithms

To discover which regression method will most accurately forecast the quantity of HIV Index testing, three algorithms were chosen and their performance was compared. Based in literature, XGBoost, Random Forest (RF), and Artificial Neural Network were these algorithms (ANN). The effectiveness and usefulness of the aforementioned algorithms were factors in their selection. For instance, they can both handle classification and regression issues, establish a nonlinear relationship, and deal with complex data. To choose the top-performing ML algorithm, the study took into account all three. Therefore, the following is an explanation of these algorithms:

(i) XGBoost

The ensemble algorithm XGBoost, which is based on gradient boosting, has been described as an effective and trustworthy machine learning method for tackling problems (Santhanam *et al.*, 2017). It is the finest open-source library in terms of performance, speed, and parameter setup (Bent & Mart, 2020). Both classification and regression predictive XGBoost addresses modeling issues. It is the top algorithm for the Kaggle tournament.

(ii) Random Forest

A network of decision trees is used in the ensemble learning method known as random forest. Breiman proposed it in 2001. It works for both regression and classification (Lin *et al.,* 2017; Zhang *et al.,* 2018). Multiple randomized decision trees are combined in the random forest technique. It is used to solve bigger-scale issues. For instance, random sampling accelerates the overfitting problem's degradation (Rashidi *et al.,* 2019). The dataset for the ensemble decision tree is trained using a randomly generated dataset. The output of each decision tree will be decided. The random forest algorithm's creation is depicted in the Fig. 9.

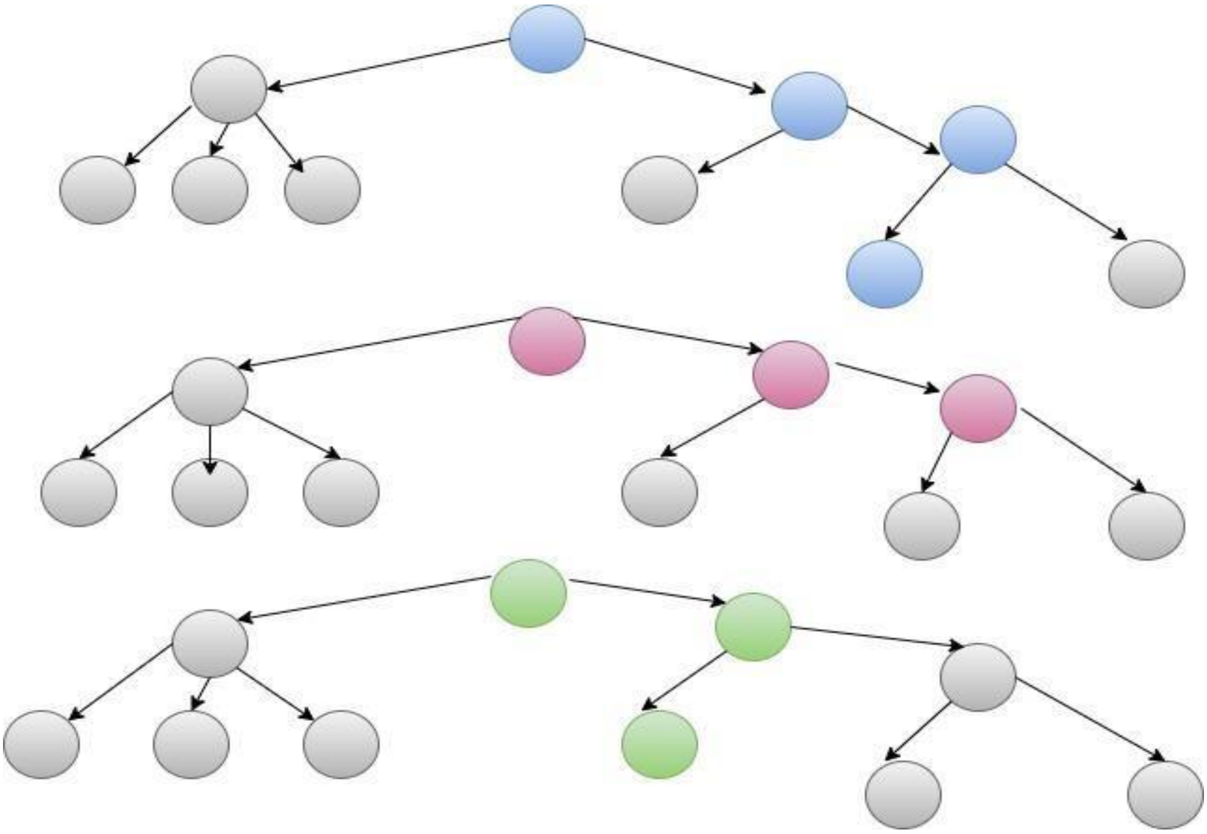


Figure 9: Random Forest algorithm diagram (Lin et al., 2017)

(iii) Artificial Neural Network

An artificial neural network, also known as a neural network, is a collection of nodes called neurons that are connected to one another (Vakili & Rezaei, 2020). It functions similarly, to how the human brain does, with a network of interconnected neurons that can communicate with one another. Iteratively, the network is asked to find a solution to a problem. More connections are made, which increases success and decreases failure.

It employed nonlinear combination to offer the capability of nonlinear relationship for neural network modeling. It can be applied to situations involving classification and regression. A random gradient (SGD) and backpropagation technique are used to train the neural network (Kaczmarek, 2019). Figure 10 depicts the structure of ANN.

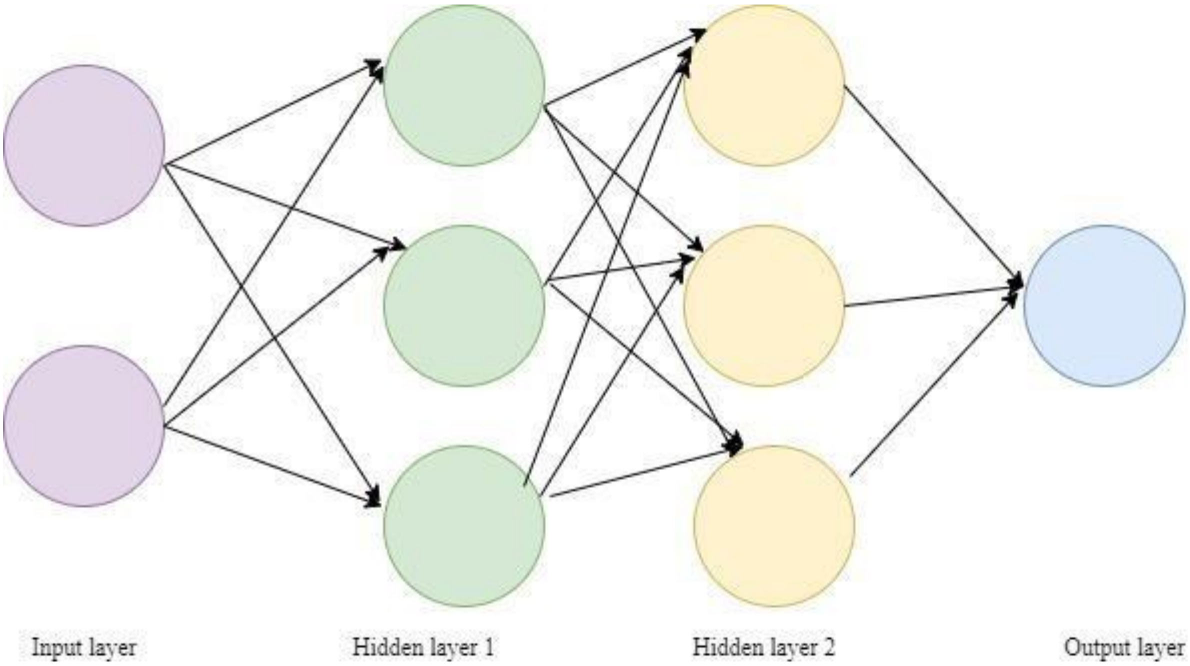


Figure 10: Artificial neural network structure

(a) Strength

- An artificial neural network is very effective in dealing with nonlinear relationships and in complex problem.
- The artificial neural network is very flexible in handling data structure with any type of variable relationship.
- Network performance is assured by only providing more training data for the network.

(b) Limitation

- It is very hard to explain and understand this model due to its complexity.
- To be keen is necessary for adjusting the super parameters in the process of training.

(c) Artificial neural network with the tensor flow

Tensor flow is an open-source library that is popular and was released in 2015 by the google brain team. It is used on python and applied in machine learning and deep learning (Dertat, 2017). It is scalable; the computation can be done across the machine and works with the large dataset; support faster debugging and model building; works efficiently with multi-dimension arrays. It has a large community and provides a tensor board to visualize the model.

3.3.6 Model Development Procedures

Major tasks involved in model creation include data acquisition, preprocessing, feature selection, and model selection. The description of the experimental techniques is shown in Fig. 11. The model developed using the three algorithms: Random Forest, XGBoost, and Neural Network. The preprocessed dataset was divided into 80% training and 20% for testing

In Random Forest Repressor, model compilation used the following n-estimators = 100, criterion =MAE (Mean Absolute Error) and n-jobs = -1. The model fit was performed, and the result was recorded. In addition, feature importance was performed to identify which feature has a high contribution to predicting the result.

In XGBRegressor, Model compilation used n-estimators = 100, max-depth 8, learning rate 0.1, sun sample 0.5 and criterion= MAE, metric MSE. In addition, feature importance was performed.

In Neural Network Repressor with the tensor flow, model compilation used were as follows: Dense 1, optimizer Adam, metric Mean Absolute Error (MAE), and epochs 100. Epoch is defined as one full cycle through a training dataset. This means that One cycle takes 100. Figure 10 illustrates the framework to follow.

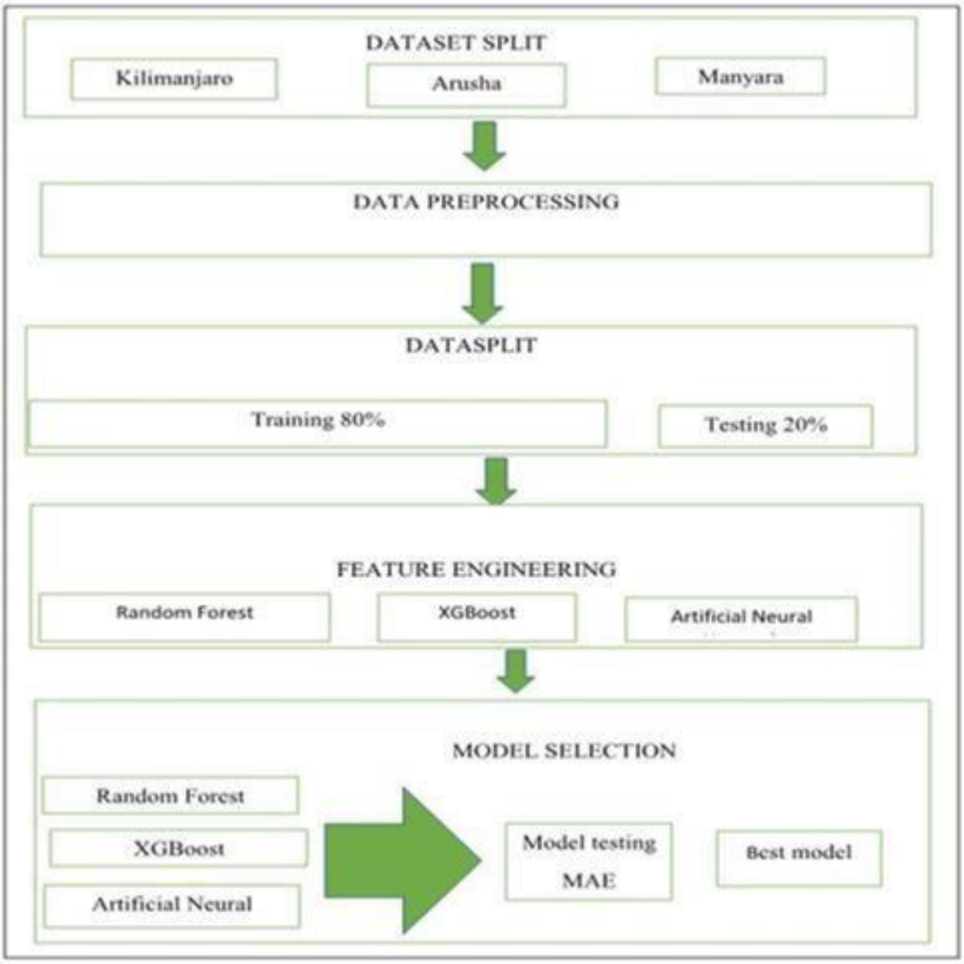


Figure 11: Machine-learning framework

3.3.7 Model Evaluation Metrics

Model evaluation involves selecting the model that best represents the data and assessing how well the model will perform with hypothetical data. Regression models can be evaluated using a wide range of indicators (Brownlee, 2021). The most popular metrics, according to the literature review, are MSE, RMSE and MAE. The mean or average of the squared discrepancies between the actual output and the expected goal values in a dataset is used to calculate the MSE. The mean squared error has been extended to include RMSE. The average of the absolute error is used to determine the MAE score. The following are mathematical representations of these measures by Wang and Lu (2018).

$\mathrm{MSE} = 1\backslash n\sum_{i=1}(y_i - y\hat{}_i)$

$\mathrm{RMSE}= \sqrt{}\ \mathbf{1}\backslash\boldsymbol{n}\sum_{\boldsymbol{i}=\mathbf{1}}(\boldsymbol{y_i} - \boldsymbol{y}\hat{}_{\boldsymbol{i}})^2$

MAE $= 1\backslash n \sum_{i=1}^{n} |y_i\text{-}y\text{^}_i|$

Where,

n = total number of datasets

∑ = summation of

$\mathbf{y_i}$ = real output

$\mathbf{y\text{^}_i}$ = predicted output

Therefore, the model selection was based on the value of the metrics obtained. The small the value, the desired model.

3.3.8 Model Deployment

(i) Prototype Development

We followed the evolutionary prototyping approach as shown in Fig. 12. We selected this approach due to its flexibility to interact with the system used to receive feedback from users, review, and update the developed prototype (Nyandowe & Zakariyah, 2014). As a result, the prototype was improved considering users' feedback, and the process continues until the user gets satisfied, then the final product is developed, tested, and maintained. Prototype stages are explained as follows:

(ii) Initial Requirement Gathering

The first step of evolutionary prototyping is requirement gathering. This is the process of identifying systems' users and their roles as well as knowing what to achieve through the system**.** Requirements of the system can be divided into two categories: Functional requirements and non-functional requirements. Performance is referred to as functional criteria for a system**,** while non-functional requirements focus on criteria or attributes that the system must comply.

(iii) Design

The second phase is system design; this refers to defining components, interfaces, and data to satisfy the required system**.** Two phases employed in this stage are conceptual design and logical design. The conceptual design shows entities and their relationship, validating the specified requirement. On the other hand, physical design involves the logical schema, which describes the data in details.

(iv) Prototyping

This phase involved building the prototype considering a conceptual model referencing the specified requirements.

(v) User evaluation

The end-users were involved in each design stage to evaluate the system and satisfy the final product aligned with their expectations. This stage was conducted in each iteration. The cycle continues until the final product is achieved as shown in Fig. 12.

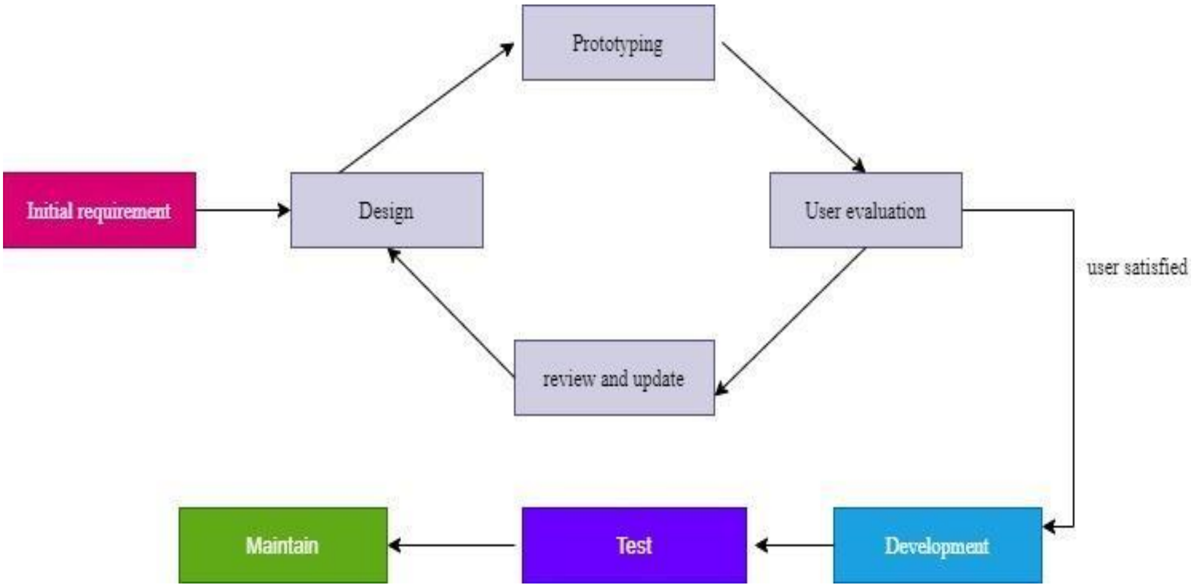


Figure 12: Evolutionary prototype diagram

(vi) Structural Design

The prototype structure was constructed from the user's requirement. Through the internet, the users are linked to the server. The users are connected to the server through the internet. Through a prototype interface, the user inputs contacts with HIV. Then transferred via the internet to the server. The hosting platform used for model deployment was Heroku. Then the model predicts the result. Figure 13 shows the presentation of the system, and Table 4 shows the user and their roles in the system.

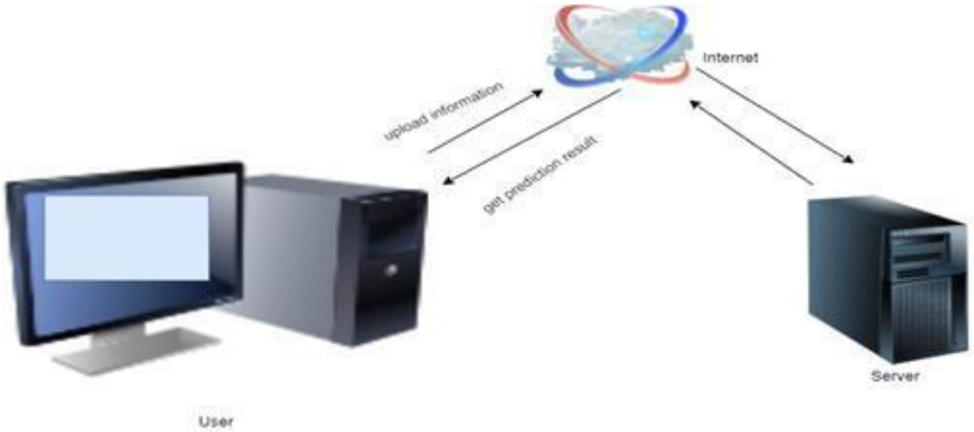


Figure 13: Architecture design diagram

Table 4: User roles and function to the system

USER	ROLES
Administrator	Add /delete user, inter HIV index information, view the prediction, and visualize the data.
A health worker at the facility	Inter HIV index information, view prediction and see a visualization of the data.

3.3.9 Prototype evaluation

(i) Technical Evaluation

Technical evaluation assessment was done considering the criteria of the non-function requirements. These include accessibility, Availability, simplicity, scalability, and consistency. Evaluators rated the system on either Low, Average, or High.

(ii) Evaluation by User

The final user evaluated the system on the System's capability to predict the number of HIV Index testing, the capability of the system to visualize the report, and the clarity of the result predicted.

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Results

4.1.1 Contributing Factors to HIV Index Testing

The findings from focus group discussion (FGD) showed that, client's gender (sex) female is willing to list their partners by 70%; age between (15-45) by 60%, people with no position by 70%; HIV awareness by 60 %, marital status (married) by 60%, education level by 50, and other were routine to go to the hospital and readiness of the client to list their partners. Table 5 shows the result in detail. Then the common factors from group discussion and the dataset were put together in order to get feature importance for predicting our model.

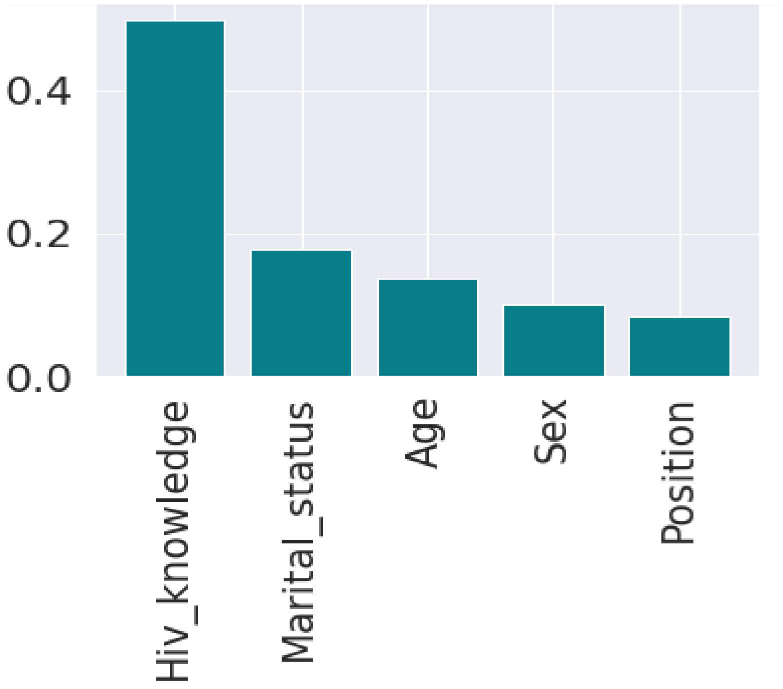
Table 5: Factors contributing to HIV index testing

	Demographic Characteristics	Respondent	Percentage (%)
Gender	Female	4	30
	Male	7	70
Age	Children (0-14)	2	20
	Youth (15-45)	5	50
	Adult (46 and above)	3	30
Education	Primary	5	50
	Secondary	3	30
	Collage	2	20
Position	Yes	3	30
	No	7	70
HIV awareness	Yes	3	30
	No	7	70
Marital Status	Married	6	60
	Not married	4	40
Go-to-hospital	Everyday	5	50
	Medium	3	30
	Rarely	2	20
Readyness- to-list partners	Yes	7	70
	No	3	30

4.1.2 Extracting features importance

The term "feature importance" describes a method of scoring each input characteristic for a certain model. Its score indicates each feature’s importance. The more points a feature receives, the more significant an impact it will have on the model used to forecast a given variable. Random Forest Importance was the method utilized in this study to determine the feature importance.

Experiment outcomes from feature extraction showed that, HIV awareness has a strong coefficient of (0.5), which is followed by marital status (0.17), age (0.15), sex (0.13), and position (an individual's influence in the community and their line of work) (0.1). Figure 14 displays the outcomes of the feature engineering, and Table 6 lists the features that were chosen for model development along with their values.



.

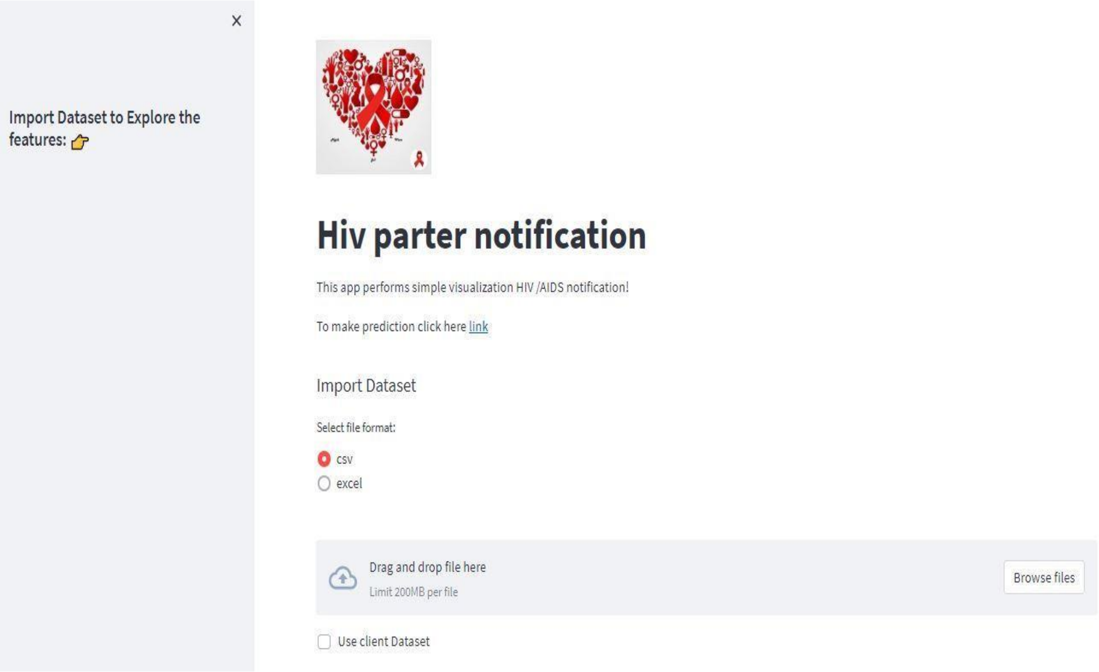
Figure 14: Feature extraction diagram

Table 6: Features chosen for model development

Variable	Description	Value
Age	Client's age	number
Sex	Client's gender (Male/ Female).	Male 1, Female 0
Position/occupation	Client with social influence/occupation, status (No/Yes)	NO/YES 1/0
Marital Status	Married/ not married (divorce, widow, widower, and never married)	NO/YES 0/1
HIV knowledge	Client with HIV knowledge and none (yes/ no)	NO/YES 1/0

4.1.3 Data Visualization

The section illustrates data insight from a variety of perspectives. Figure 15 displays the partner notification dashboard for HIV (Kingbo *et al*., 2020). The amount of HIV indices per ClientID is shown in Fig. 16. Figure 17 shows the number of HIV indexes broken down by site and status. Figure 18 shows the HIV index in relation to HIV status and the nature of the association. The number of HIV index cases by location and age distributions are shown in Fig. 19 and Fig. 20.



×

Import Dataset to Explore the features: 👉

Hiv parter notification

This app performs simple visualization HIV /AIDS notification!

To make prediction click here link

Import Dataset

Select file format:

● csv

○ excel

Drag and drop file here

Limit 200MB per file

Browse files

□ Use client Dataset

Figure 15: Dashboard for HIV partner notification app

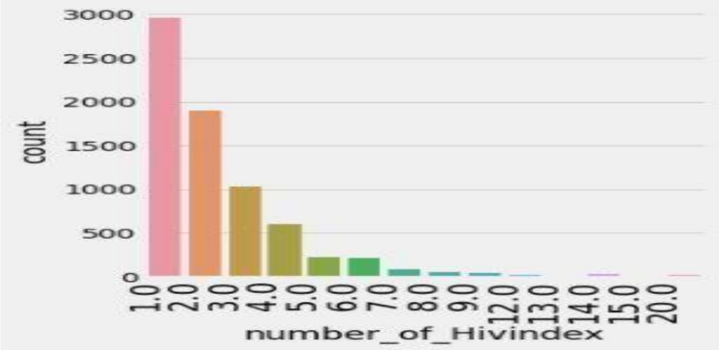


Figure 16: The total number of HIV-positive contacts per client_id

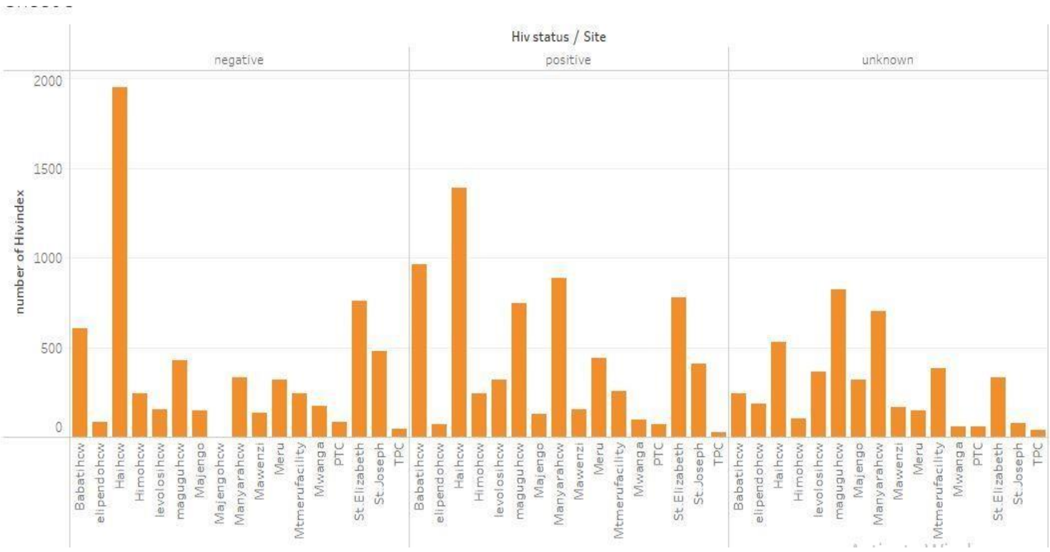


Figure 17: HIV index number by place and status

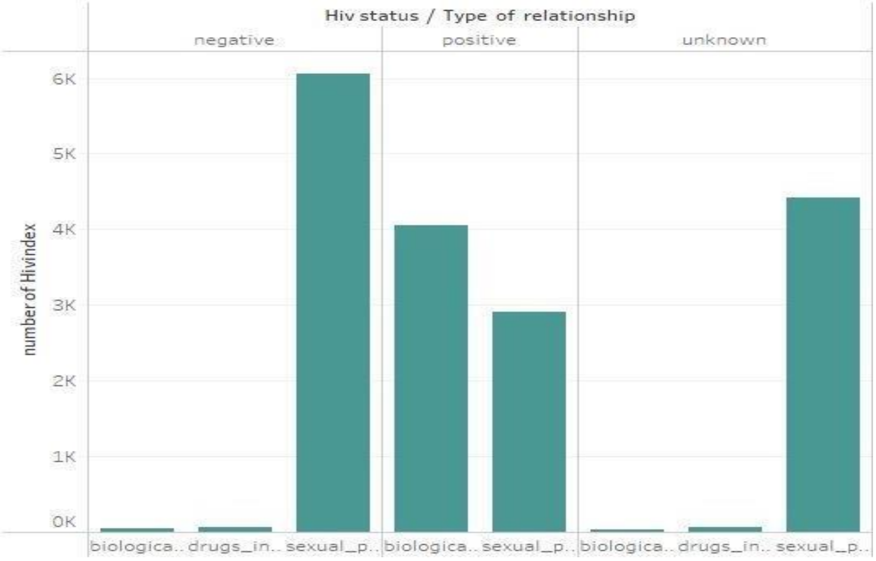


Figure 18: HIV index count in relation to HIV status and type of relationship

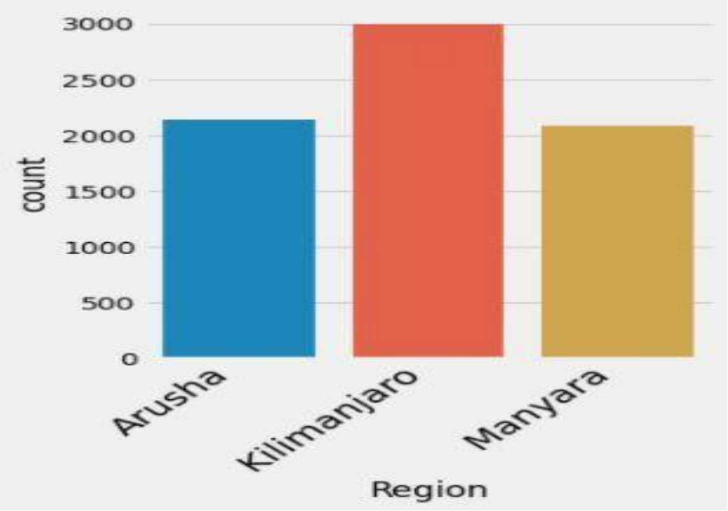
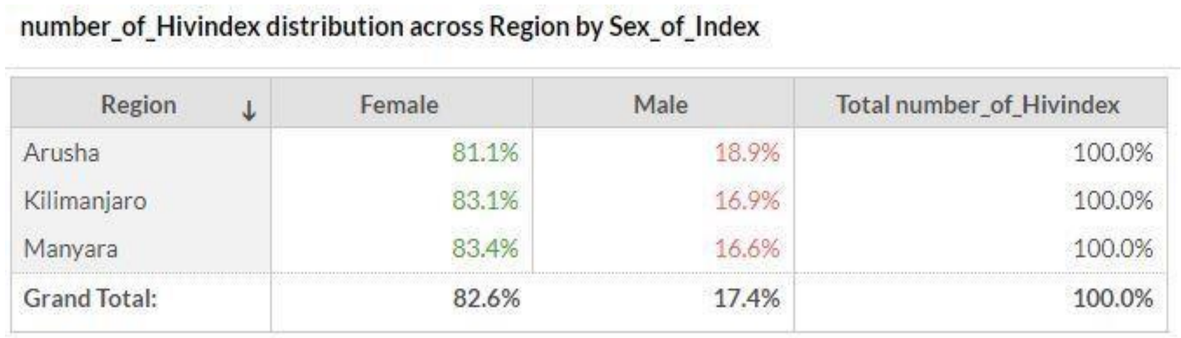


Figure 19: Diagram displaying each region's overall HIV index numbers



number_of_Hivindex distribution across Region by Sex_of_Index

Region ↓	Female	Male	Total number_of_Hivindex
Arusha	81.1%	18.9%	100.0%
Kilimanjaro	83.1%	16.9%	100.0%
Manyara	83.4%	16.6%	100.0%
Grand Total:	82.6%	17.4%	100.0%

Figure 20: Regional gender distribution

4.1.4 Development and Evaluation of Model

The results of the model creation utilizing the three algorithms are presented in Table 7. Random forest outperformed the other two by having the lower MAE value. The desired model increases as the value decreases.

Table 7: Result obtain during model development

Serial number(S/N)	Model name	Metric (MAE)
1	Random Forest	1.1261
2	XGBoost	1.2340
3	ANN	1.1268

A random forest was the algorithm with the best performance. Its Mean Absolute Error (MAE) was the lowest, at 1.1261. After the model was improved, using the most effective GridSearchCV parameters, the outcome remained unchanged. The model was then saved and made ready for use.

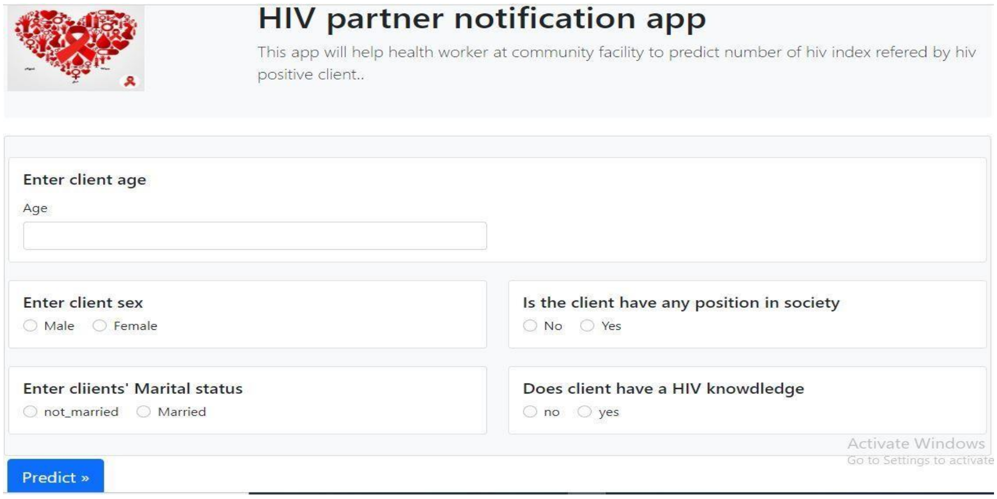
4.1.5 Model deployment

(i) User's Requirements

The following were suggested as functional and non-functional needs for the proposed system as of the focus group discussion, as shown in Table 8. The interface to capture user input for model building is shown in Fig. 21 and the result page for the predicted number of HIV index in Fig. 22, respectively.

Table 8: Proposed system requirements

Functional Requirements	Non-functional requirements
The system shall predict the number of HIV index.	The system shall allow upgrading when needed. Interoperability.
The system shall provide a visualization report.	The system shall be easy to use (simplicity).
The predicted result shall be clear and understandable.	A system shall be capable of responding to user inquiries immediately.
	A system shall be able to provide support, change and restructure over time (Maintainability).



HIV partner notification app

This app will help health worker at community facility to predict number of hiv index refered by hiv positive client..

Enter client age

Age

Enter client sex

○ Male ○ Female

Is the client have any position in society

○ No ○ Yes

Enter cliients' Marital status

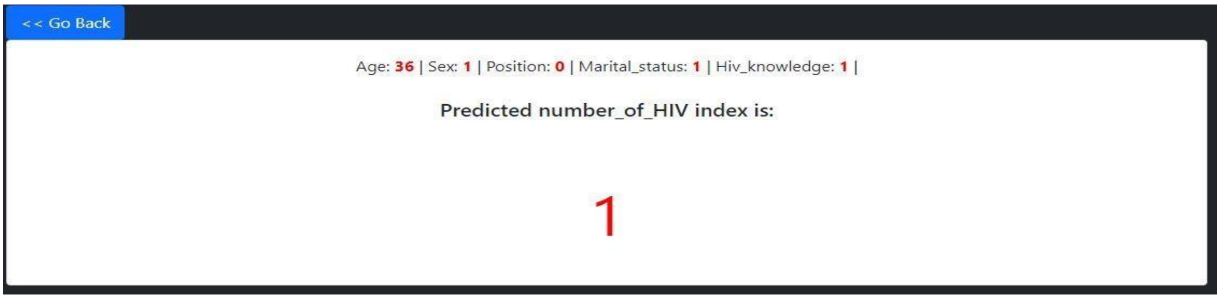
○ not_married ○ Married

Does client have a HIV knowdledge

○ no ○ yes

Predict »

Figure 21: User input interface



Age: 36 | Sex: 1 | Position: 0 | Marital_status: 1 | Hiv_knowledge: 1 |

Predicted number_of_HIV index is:

1

Figure 22: Prediction result interface

(ii) Prototype Evaluation

Beta testing results showed that 93% of the respondent on the simplicity of the application, accessibility and Performance scored 88%, Performance and prediction got 80% clarity of the result 78%, Interoperability and visualization 75% as shown in Fig. 23.

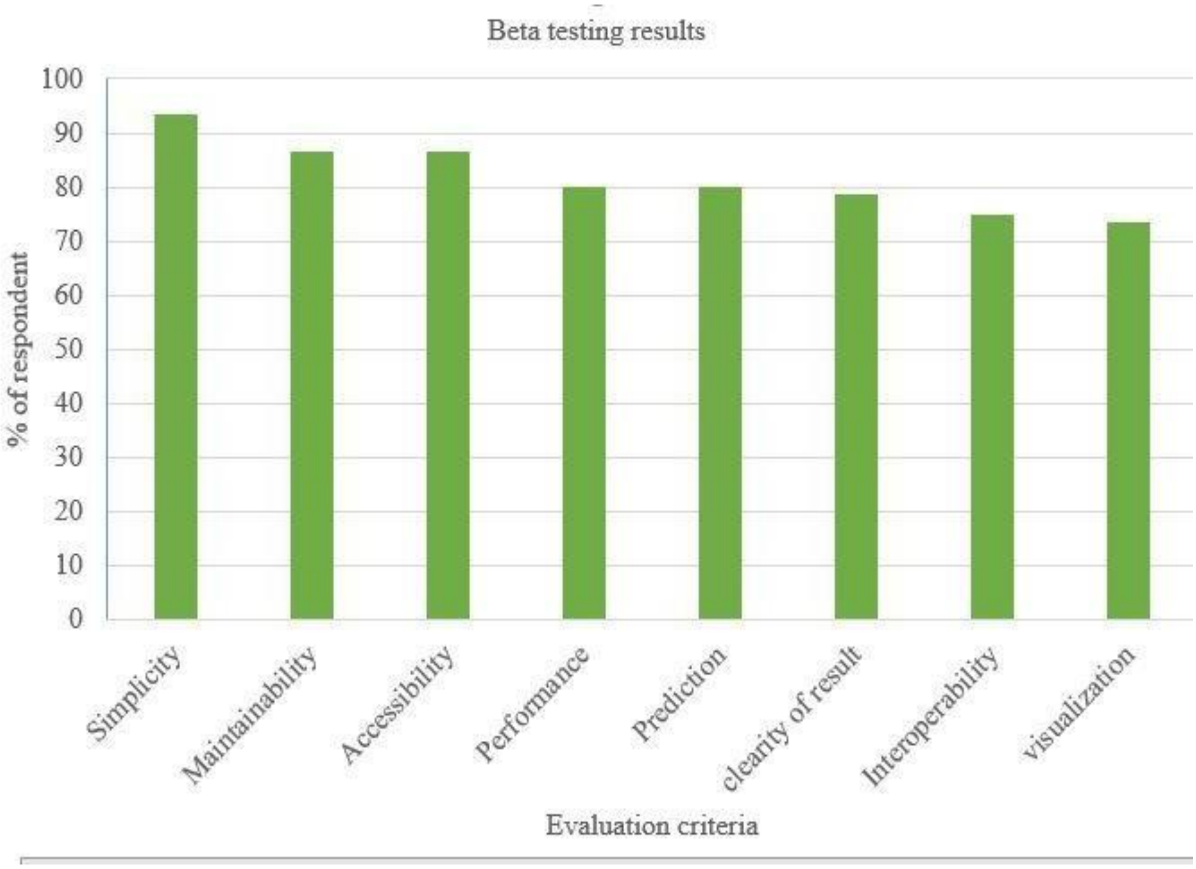


Figure 23: Prototype evaluation diagram

4.2 Discussion

The process of understanding the domain knowledge was done thoroughly. The significant number of respondents specified clients' understanding of HIV /AIDS (HIV knowledge) is the main contributing factor HIV index testing. Similar findings were reported by Gitige *et al*. (2021). Partners mentioned HIV knowledge in the context of awareness of elicitation method, awareness of HIV status partner, and fear-free from rejection. Another factor mentioned by the responded was age, the most affected group was youth from (16-39), followed by Marital status, Married people were ready to list their sexual partners so as to be tested, and inked into care and treatment. Sex Female gender is leading in eliciting their sexual partner, drug injecting and their biological children under 15 years. position/occupation and the level of education were listed as less contributing factor. Similarly, reviewed work (Henley *et al*., 2013) listed the same factor.

Additionally, the findings from the feature engineering showed that HIV knowledge was the main contributing factor for HIV index testing. The followed factors were marital status, age, sex and the last was position. The technique used in this study was the random forest technique. The selection of the mentioned technique was its ability to deal with relationship of non-linearity variables. However, the viewed literature work used Statistical tools like STATA and Epi Info, and Microsoft excel (Gitige *et al.*, 2021; Mwango *et al.,* 2020), for analyzing the feature importance. This study used machine learning techniques to avoid ignoring of the nonlinear features (Gupta, 2020).

In this study, a prototype was created based on the final user requirements and established into the user interface for easier interaction. The user wanted the interface that can capture user input of HIV-positive client information one after another and the predicted result for that specific client. Additionally, the user wanted an interface for uploading data to perform statistical analysis and visualize different reports about that data. Usability is vital in any web application Anjum *et al.* (2018).

This study comes out with a prototype that serves the two requirements. The statistical analysis does not need statistical tools like Stata and SPSS. The User just uploads file and select by one click the function he/she wants to perform. Moreover, technical evaluations were carried out on the criteria such as simplicity, maintainability, interoperability and accessibility. The prototype of this study is developed in such a way that technical experts can do maintenance for the future and it is accessible and can operate in various operating systems.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

The goal of the project was to create a machine-learning model for estimating the prevalence of HIV index in northern Tanzania. The dataset used were from the three regions of Arusha, Kilimanjaro and Manyara. The dataset had 11 features. Feature extraction was performed to find a feature with high contribution to the target. The result showed that HIV knowledge is the main factor for HIV index testing. The experiment was carried out using three algorithms, where by random forest was the best in terms of performance by having smallest value of MAE.

The model developed was deployed into prototype for the purpose of interpreting results from machine learning model. Moreover, the interface for doing statistical analysis and visualizing different report was incorporated into the prototype developed.

These outcomes indicate that the developed machine-learning model and the prototype were able of making prediction number of HIV index elicited by the HIV positive client who will be at risk of contracting the HIV/AIDS disease. Hence, health workers and other stakeholders for early intervention can take further steps.

Additionally, the visualization report created into prototype will assist government, NGO'S related to HIV/AIDS and policy makers to find better strategies to end up the HIV/AIDS as heath threat to society by 2030.

5.2 Recommendations

Tanzania is one of the sub-Saharan countries with a large rate of people living with HIV. Therefore, the study recommends using the developed prototype (client partner notification) in decision support and medical care guidelines to reach the target 95 95 95 by 2025 set by WHO., meaning that 95% of PLHIV know their status, 95% are enrolled to health care and treatment; and 95% their health has been improved.

The study found that the main factor contributing to HIV index elicitation is the knowledge of HIV/AIDS. Therefore, we recommend HIV awareness should be continuous excise by all

HIV/AIDS stakeholders. Using various media to reach all people with regardless of their residence, education level, occupation, gender and age.

The prototype that was created is a web application. Therefore, access to it requires an internet connection. We advise other researchers to take into account creating mobile applications that can operate offline, allow users to access, and interact with the system at anytime and anywhere without the need for an internet connection because of the poor and inconsistent internet accessibility, particularly in rural areas.

Three northern regions with low rates of HIV prevalence were taken into consideration in the study. We advise other studies to conduct research in other regions, particularly those with high rates of HIV prevalence, in order to produce more comprehensive information that will aid in model development and lead to the development of a good solution for finding HIV index testing.

Although a good prototype (a HIV notification app) was created, the model was built based on collected data. This was brought on by incomplete health care data, inadequate knowledge of social, economic, and behavioral variables related to health. Information such as place of residence (rural/unban), economic situation (poor/middle/reach), level of education, and religion could affect the outcome. Therefore, the study recommends that the health care system, especially the unit dealing with HIV/AIDS to use the automated system and review the data to be collected for both hospitals and stakeholders to facilitate quality data collection.

REFERENCES

Anjum, N., Sarker, M., & Ahmed, S. (2018). Evaluation of Web Usability Requirement Model and Web Application Interface Components. *International Journal of Advanced Research in Computer Science and Software Engineering*, *8*(12), 1–8.

Avert. (2020). *Tanzania 90-90-90 progress*. *August*, 2020.

Bain, L. E., Nkoke, C., & Noubiap, J. J. N. (2017). UNAIDS 90-90-90 targets to end the AIDS epidemic by 2020 are not realistic: Comment on "Can the UNAIDS 90-90-90 target be achieved? A systematic analysis of national HIV treatment cascades". *BMJ Global Health*, *2*(2), e000227. https://doi.org/10.1136/bmjgh-2016-000227

Balzer, L. B., Havlir, D. V., Kamya, M. R., Chamie, G., Charlebois, E. D., Clark, T. D., Koss, C. A., Kwarisiima, D., Ayieko, J., Sang, N., Kabami, J., Atukunda, M., Jain, V., Camlin, C. S., Cohen, C. R., & Bukusi, E. A. (2019). *Machine Learning to Identify Persons at High-Risk of Human Immunodeficiency Virus Acquisition in Rural Kenya and Uganda*. *Xx Xxxx*, 1–8. https://doi.org/10.1093/cid/ciz1096

Bent, C., & Mart, G. (2020). *A Comparative Analysis of XGBoost A Comparative Analysis of XGBoost*. *November 2019*.

Brownlee, J. (2021). Regression Metrics for Machine Learning. In *Machine Learning Mastery*. https://machinelearningmastery.com/regression-metrics-for-machine-learning/

COP. (2019). *Tanzania_COP19-Strategic-directional-Summary_public*.

Dertat, A. (2017, August 8). Applied Deep Learning—Part 1: Artificial Neural Networks. *Towards Data Science*. Applied Deep Learning - Part 1: Artificial Neural Networks

Gitige, C. G., Kwesigabo, G. P., Panga, O. D., Samizi, F. G., Abade, A. M., Mbelele, P. M., & Kishimba, R. S. (2021). Factors associated with Partners Elicitation during HIV Index client's testing in Dar es Salaam Region, Tanzania. *Journal of Interventional Epidemiology and Public Health*, *4*(3). https://doi.org/10.37432/jieph.2021.4.3.41

Gupta, A. (2020). Feature Selection Techniques in Machine Learning. In *Https://Www.Analyticsvidhya.Com/Blog/2020/10/Feature-Selection-Techniques-in-Machine-Learning/*.

Gwajima, D., & Wright, D. (2021). *Strategicfocus*. *March*, 2021.

Henley, C., Forgwei, G., Welty, T., Golden, M., Adimora, A., Shields, R., & Muffih, P. T. (2013). Scale-up and case-finding effectiveness of an HIV partner services program in Cameroon: An innovative HIV prevention intervention for developing countries. *Sexually Transmitted Diseases*, *40*(12), 909–914.

Jeren, D., & Abebe, W. (2017). *Hiv Testing Services Hiv Self-Testing and Partner*. *December*, 7.

Kaczmarek, J. S. D. W. A. (2019). Multiple regression models and Artificial Neural Network (ANN) as prediction tools of changes in overall quality during the storage of spreadable processed Gouda cheese. *European Food Research and Technology*, *245*(11), 2539–2547. https://doi.org/10.1007/s00217-019-03369-y

Khan, H., & Khattak, M. (2006). Hiv / Aids. In *The Professional Medical Journal* (Vol. 13, Issue 04, pp. 627–631). https://doi.org/10.29309/tpmj/2006.13.04.4940

Leo, J., Luhanga, E., & Michael, K. (2019). Machine learning model for imbalanced cholera dataset in Tanzania. *The Scientific World Journal*, 9(1), 1–14.

Lin, W., Wu, Z., Lin, L., Wen, A., & Li, J. I. N. (2017). *An Ensemble Random Forest Algorithm for Insurance Big Data Analysis*. *5*.

Mahachi, N., Muchedzi, A., Tafuma, T. A., Mawora, P., Kariuki, L., werq Semo, B., Bateganya, M. H., Nyagura, T., Ncube, G., Merrigan, M. B., Chabikuli, O. N., & Mpofu, M. (2019). Sustained high HIV case-finding through index testing and partner notification services: Experiences from three provinces in Zimbabwe. *Journal of the International AIDS Society*, *22*(S3), 23–30. https://doi.org/10.1002/jia2.25321

Mwango, L. K., Stafford, K. A., Blanco, N. C., Lavoie, M. C., Mujansi, M., Nyirongo, N., Tembo, K., Sakala, H., Chipukuma, J., Phiri, B., Nzangwa, C., Mwandila, S., Nkwemu, K. C., Saadani, A., Mwila, A., Herce, M. E., & Claassen, C. W. (2020). Index and targeted community-based testing to optimize HIV case finding and ART linkage among men in Zambia. *Journal of the International AIDS Society*, *23*(S2), 51–61. https://doi.org/10.1002/jia2.25520

National Bureau of Statistics. (2018). Tanzania HIV Impact Survey (THIS) 2016-2017. *Tanzania HIV Impact Survey (THIS) 2016-2017, December 2017*, 2016–2017.

Nyandowe, I. T., & Zakariyah, S. S. (2014). *User Testing and Feedback*. *October*, 1–2. https://doi.org/10.13140/2.1.1473.4080

Prabhakaran, S. (2014). Machine Learning Methods for HIV / AIDS Diagnostics and Therapy Planning. *Ph.D Thesis*.

Raposo, L. M., Rosa, P. T. C., & Nobre, F. F. (2020). *Random Forest Algorithm for Prediction of HIV Drug Resistance Random Forest Algorithm for Prediction of HIV Drug Resistance*. *March*. https://doi.org/10.1007/978-3-030-38021-2

Rashidi, H. H., Tran, N. K., Abb, H., Betts, E. V., Howell, L. P., & Green, R. (2019). *Artificial Intelligence and Machine Learning in Pathology: The Present Landscape of Supervised Methods*. *6*. https://doi.org/10.1177/2374289519873088

Santhanam, R., Uzir, N., Raman, S., & Banerjee, S. (2017). Experimenting XGBoost Algorithm for Prediction and Classification of Different Ramraj S, Nishant Uzir, Sunil R and Shatadeep Banerjee Experimenting XGBoost Algorithm for Prediction and Classification of Different Datasets. *International Journal of Control Theory and Applications*, *9*(March), 651–662.

SAP Insights. (2019). *What Is Machine Learning? | Definition, Types, and Examples | SAP Insights*. https://insights.sap.com/what-is-machine-learning/

Singh, Y., Narsai, N., & Mars, M. (2013). Applying machine learning to predict patient-specific current CD 4 cell count in order to determine the progression of human immunodeficiency virus (HIV) infection. *African Journal of Biotechnology*, *12*(23).

Smyrnov, P., Sereda, Y., Lytvyn, A., & Denisiuk, O. (2016). *Improving HIV case-finding with machine learning ML algorithm has performed better or equally well in comparison with rule based algorithm on making decision who should receive additional recruitment coupons due to higher probability of undiagnosed HIV ca*. 1223.

Streamlit. (2022). *Streamlit: The fastest way to build and share data apps*.

Tawasanet. (2019). *TAWASANET: Where We Work*. http://www.tawasanet.or.tz/

UNAIDS. (2020). Data 2020. *Programme on HIV/AIDS*, 1–248.

United Nations Joint Programme on HIV/AIDS (UNAIDS). (2014). To help end the AIDS epidemic. *United Nations*, 40.

Vakili, M., & Rezaei, M. (2020). *Performance Analysis and Comparison of Machine and Deep Learning Performance Analysis and Comparison of Machine and Deep Learning Algorithms for IoT Data Classification*. *January*, 0–13.

Wang, W., & Lu, Y. (2018). Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model. *IOP Conference Series: Materials Science and Engineering*, *324*(1). https://doi.org/10.1088/1757-899X/324/1/012049

Zhang, Q., Chai, Y., Li, X., Young, S. D., & Zhou, J. (2018). Using internet search data to predict new HIV diagnoses in China: A modelling study. *BMJ Open*, *8*(10). https://doi.org/10.1136/bmjopen-2017-018335

APPENDICES

Appendix 1: Codes for Model Development

```
# -*- coding: utf-8 -*-

"""Randomforest.ipynb

Automatically generated by Colaboratory.

Original file is located at

    https://colab.research.google.com/drive/1UKNNUp25ZHAqlsrJ_tBPUDCYEWv2oeEX

"""

"""Import libraries"""

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

from google.colab import files

uploaded = files.upload()

import io

client_data = pd.read_csv(io.BytesIO(uploaded['client_info.csv']))

# Dataset is now stored in a Pandas Dataframe

client_data

data = client_data.drop(columns=['Date','Residence', 'Contact_no', 'CTC_NO'])

data

#checking the data

data.isnull().sum()
```

```
data_clean = data.apply(lambda x: x.fillna(x.value_counts().index[0]))

data_clean

from sklearn.preprocessing import LabelEncoder

# Turn all categories into numbers using label encoder

le = LabelEncoder()

data_clean['Sex'] = le.fit_transform(data_clean['Sex'])

data_clean['Position'] = le.fit_transform(data_clean['Position'])

data_clean['Marital_status'] = le.fit_transform(data_clean['Marital_status'])

data_clean['Hiv_knowledge'] = le.fit_transform(data_clean['Hiv_knowledge'])

data_clean.head()

# Create X & y values

X = data_clean.drop(['client_id', 'number_of_Hivindex'], axis=1)

y = data_clean['number_of_Hivindex']

# View features

X.head()

print(X.shape)

#import Library for Random forest to work

from sklearn.ensemble import RandomForestRegressor

from sklearn.metrics import mean_absolute_error

from sklearn.metrics import mean_squared_error

# Create training and test sets

from sklearn.model_selection import train_test_split
```

```
X_train, X_test, y_train, y_test = train_test_split(X,
                                                    y,
                                                    test_size=0.2,
                                                    random_state=42)
model1 = RandomForestRegressor(n_estimators=100,
                               criterion="mae",
                               n_jobs=-1)
#fitting the model
model1.fit(X_train, y_train)
y_pred = model1.predict(X_test)
print (mean_absolute_error(y_pred,y_test))
print (mean_squared_error(y_pred, y_test))
from sklearn.model_selection import RandomizedSearchCV
from sklearn.utils import shuffle
rf = RandomForestRegressor()
param_grid = {}
grid_model = GridSearchCV(model1, param_grid)
grid_model.fit(X_train, y_train)
grid_model.fit(X_test,y_test)
print(grid_model.best_params_)
y_pred = model1.predict(X_test)
print (mean_absolute_error(y_pred,y_test))
print (mean_squared_error(y_pred, y_test))

import pickle

file_name = "Bestmodel.pkl"

# save

pickle.dump(model1, open(file_name, "wb"))

# load

model1_loaded = pickle.load(open(file_name, "rb"))

predictions = model1.predict(X_test)

predictions
```

Appendix 2: Codes for Model Deployment

```
from flask import Flask, render_template, request

import pickle

import numpy as np

import pandas as pd

# Initialise the Flask app

app = Flask(__name__)

# Use pickle to load in the pre-trained model

filename = "Bestmodel.pkl"

model = pickle.load(open(filename, "rb"))

# Set up the main route

@app.route('/', methods=["GET", "POST"])

def main():

    if request.method == "POST":

    # Extract the input from the form

    Age = request.form.get("Age")

    Sex = request.form.get("Sex")

    Position = request.form.get("Position")

    Marital_status = request.form.get("Marital_status")

    Hiv_knowledge = request.form.get("Hiv_knowledge")

    Ag = int(Age)

    Sx = int(Sex)

    Psition = int(Position)
    Marital_sttus = int(Marital_status)

    Hiv_knowledg = int(Hiv_knowledge)

    # Get the model's prediction

    # Given that the prediction is stored in an array we simply extract by indexing

    arr = np.array([[Ag,Sx,Psition,Marital_sttus,Hiv_knowledg]])

    # arr = np.array([[Age,Sex,Position,Marital_status,Hiv_knowledge]])

     data = pd.DataFrame(data=arr,

                columns = ['Sex', 'Age', 'Position', 'Marital_status', 'Hiv_knowledge'] )

    #pred = xg_model.predict(arr)

    #return render_template('result.html', data=pred)

    prediction = model.predict(data)[0]

     return render_template("result.html",

                          original_input={'Age':Ag,

                                  'Sex':Sx,

                                  'Position':Psition,

                                  'Marital_status':Marital_sttus,

                                  'Hiv_knowledge':Hiv_knowledg},

                          result=int(round(prediction,0))

                          )

  # If the request method is GET

  return render_template("index.html")
```

Appendix 3: Guiding Questions for Focus Group Discussion

Question 1: Do the client aware of HIV/AIDS?

Question 2: How often people go to hospital for HIV testing?

Questions 3: Which factors contribute to Spread HIV/AIDS in Tanzania?

Question 4: which age group does HIV mostly affect?

Question 5: Which group age is responsive to HIV index testing?

Question 6: What is responsiveness regarding to HIV issues between women and men?

Question 7: How is the HIV positive client ready to mention his/her chain?

Question 8: Which Marital status are mostly affected with HIV/AIDS?

Question 9: Can level of Education being an influence for HIV index testing?

Question 10: Is position of a person in a society be a factor to hinder HIV index testing?

Appendix 4: Prototype Evaluation Questionnaires

(a) Technical expert evaluation

Evaluation Criteria	High	Average	Low
Simplicity			
Performance			
Maintainability			
Accessibility			
Interoperability			

(b) End users' evaluation

Evaluation Criteria	High	Average	Low
The capability of the prototype to predict HIV Index testing.			
The capability of the system to visualize report.			
The capability of the system to predict the clear and understandable result.			

Appendix 5: Codes for Prediction Result Page

```
<!DOCTYPE html>
<html lang="en">
<head>
    <meta charset="UTF-8">
    <meta http-equiv="X-UA-Compatible" content="IE=edge">
    <meta name="viewport" content="width=device-width, initial-scale=1.0">
    <!-- Bootstrap CSS -->
    <link href="https://cdn.jsdelivr.net/npm/bootstrap@5.0.2/dist/css/bootstrap.min.css"
rel="stylesheet" integrity="sha384-
EVSTQN3/azprG1Anm3QDgpJLIm9Nao0Yz1ztcQTwFspd3yD65VohhpuuCOmLASjC"
crossorigin="anonymous">
   <title>Result</title>
</head>
<body class="bg-dark">
<div class="container ">
<br>
    <a href="/" class="btn btn-primary"> << Go Back </a>
    <br>
<div class="card raise text-center">
  <!-- <div class="card-header">
 </div> -->
  <div class="card-body">
    <!-- <h5 class="card-title">Special title treatment</h5>
    <p class="card-text">With supporting text below as a natural lead-in to additional
content.</p>
    <a href="#" class="btn btn-primary">Go somewhere</a> --> {% if result %}
```

Appendix 6: Poster Presentation

MACHINE LEARNING MODEL FOR PREDICTION AND VISUALIZATION OF HIV INDEX IN NORTHEN TANZANIA

1. Happyness S Chikusi (Student) 2. Prof. Shubi Kaijage & Dr. Judith Leo (Supervisors)

Background

Infection with the human immunodeficiency virus and acquired immunodeficiency syndrome (HIV/AIDS) continue to pose a threat to Tanzanian society. Various tactics have been used to improve the number of persons who are aware of their HIV status. HIV index testing stands out among these methods as the most effective way to count the number of HIV contacts who may be at risk of catching HIV from HIV-positive individuals. The current HIV index testing, however, is manual, which presents a number of difficulties, including inaccuracies, is time-consuming, and is expensive to operate. tstudy presents the findings of the machine-learning model.

Research Objective

To develop machine learning model for prediction of HIV index testing in northern Tanzania

Method

Evolutionary prototyping was used in this study. This was selected because, suitable for interactive refine at each step. Saves time and used in AI and Machine learning

Result

The study come out with Machine learning model that deployed into prototype, which is simple to use, accurate and faster.

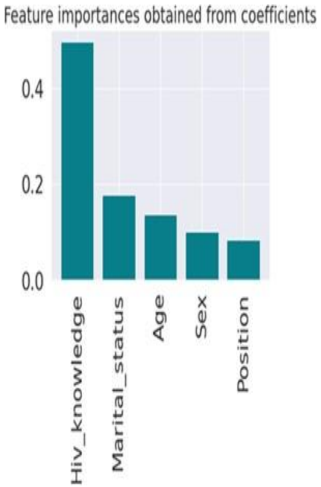


Figure1: Feature extraction for HIV index

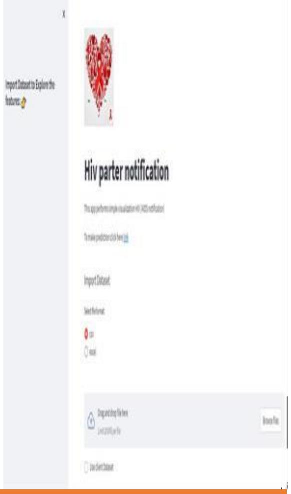
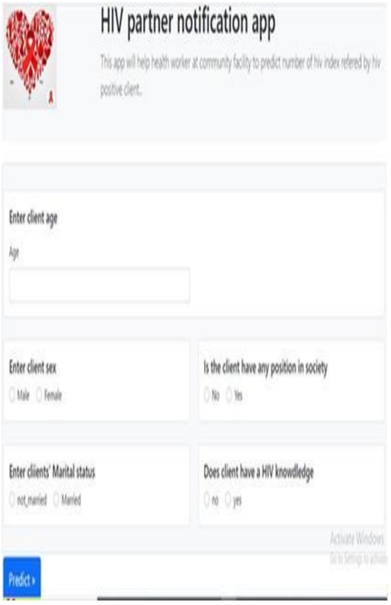


Figure 2: Dashboard


HIV partner notification app

This app will help health worker at community facility to predict number of hiv index refered by hiv positive client.

Enter client age

Age

Enter client sex

○ Male ○ Female

Is the client have any position in society

○ No ○ Yes

Enter clients' Marital status

○ not_married ○ Married

Does client have a HIV knowledge

○ no ○ yes

Figure 3: User input interface

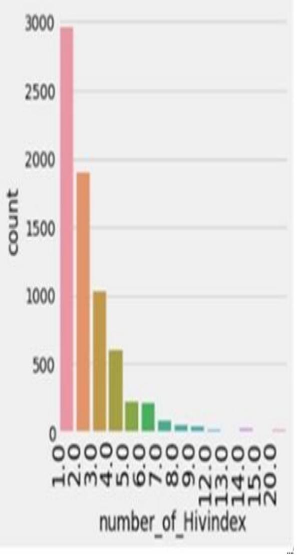


Figure 4: HIV + contacts per client_id

By conclusion, this work helped us realize the importance of machine learning in predicting and visualizing HIV index tests in general. The created model can help decision-makers build a viable intervention to stop the spread of HIV and AIDS in our communities.