

2022-08

# A prediction tool for dolutegravir associated hyperglycaemia among HIV patients in Uganda

Favor, Wisdom

NM-AIST

---

<https://dspace.nm-aist.ac.tz/handle/20.500.12479/1612>

*Provided with love from The Nelson Mandela African Institution of Science and Technology*

**A PREDICTION TOOL FOR DOLUTEGRAVIR ASSOCIATED  
HYPERGLYCAEMIA AMONG HIV PATIENTS IN UGANDA**

**Wisdom Ceaser Favor**

**A Project Report Submitted 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**

**Arusha, Tanzania**

**August, 2022**

## ABSTRACT

The initiation of Dolutegravir based antiretroviral therapy has provided a potent treatment option for persons living with human immunodeficiency Virus (PLHIV). However, clinical research has shown overwhelming evidence that use Dolutegravir (DTG) results into consequential hyperglycaemia. The incidence and prevalence rates of Dolutegravir associated hyperglycaemia among PLHIV are unknown. Therefore, identification of patients susceptible to dolutegravir associated hyperglycaemia is critical to lessening morbidity and mortality associated with uncontrolled high blood glucose level among patients on antiretroviral therapy (ART) and care. The current procedures practiced in screening for hyperglycemia among PLHIV being switched to DTG and DTG Associated Hyperglycemia related literature were appraised. Various machine learning classification algorithms were employed to come up with the most appropriate model. The purpose of the study was to develop an efficient DTG associated hyperglycaemia Screening tool for treatment experienced PLHIV being switched to DTG in Uganda. The study found Extreme Gradient Boost Classification model as the best with performance and evaluation metrics as follows: model accuracy is 0.99, probability to classify positives is 0.87, precision probability for predicting positives is 0.67, Area under the receiver operating Characteristic curve is 0.90, Area under a precision, recall curve as 0.86, F1 score is 0.76, and Cohen Kappa score as 0.72 inter alia. Therefore, the study recommends the adoption and use of the DTG associated hyperglycaemia screening tool while switching HIV treatment experienced patients to a DTG based regimen. This is because the world over is using Machine learning tools to support Medicare. The researcher also suggests stratifying the data by gender and build DTG associated prediction models for women and men separately because men are not affected in any by variables of pregnancy and post-menopausal phase like women. The research prosers lessening of bureaucratic tendencies and high charges levied on research protocols and data acquisition as a way of promoting these innovative studies among others.

## DECLARATION

I, Wisdom Ceaser Favor do declare to the Senate of the Nelson Mandela Africa Institution of Science and Technology that this Project report is my own original work and that it has neither been submitted nor being concurrently submitted for degree award in any other institution.

Wisdom Ceaser Favor



---

Name and Signature of Candidate

Date

The Above declaration is confirmed by:

Prof. Kisangiri Michael

---

Name and Signature of Supervisor 1

Date

Dr. Ramadhan Sinda

---

Name and Signature of Supervisor 2

Date

## **COPYRIGHT**

This Project report is copyright material protected under the Berne Convention, the Copyright Act of 1999 and other international and national enactments, in that behalf, on intellectual property. It must not be reproduced by any means, in full or in part, except for short extracts in fair dealing; for researcher private study, critical scholarly review or discourse with an acknowledgment, without a written permission of the Deputy Vice Chancellor for Academic, Research and Innovation, on behalf of both the author and the Nelson Mandela African Institution of Science and Technology.

## CERTIFICATION

The undersigned certify that have read and hereby recommend for acceptance by the Senate of the Nelson Mandela African Institution of Science and Technology, the Project Report titled “*A Prediction Tool for Dolutegravir Associated Hyperglycaemia Among HIV Patients in Uganda*” in Partial Fulfillment of the Requirements for the Award of the Degree of Master of Science in Embedded and Mobile Systems of the Nelson Mandela African Institution of Science and Technology.

Prof. Kisangiri Michael

---

Name and Signature of Supervisor 1

Date

Dr. Ramadhan Sinda

---

Name and Signature of Supervisor 2

Date

## ACKNOWLEDGEMENTS

At this moment, I am grateful to the Lord God Almighty for blessings, good health and availing me an opportunity to join Master's program at Nelson Mandela African Institution of Science and Technology (NM-AIST). On the same note gave me resolve, strength, and courage in my entire learning experience in pursuit of this achievement.

I am also immensely grateful to my sponsor Centre of Excellence for ICT in East Africa for the scholarship and financial as support extended in pursuit of my studies at NM-AIST.

I as well express my inmost appreciation to the research advisors: Dr. Balaba Martin and Dr. Agnes Kiraga of Infectious Disease Institute Uganda, and my supervisors: Prof. Kisangiri Michael and Dr. Sindi Ramadhani of NM-AIST who provided top notch guidance and expertise whenever challenging situations in this study arose. Though I did this Project work, they offered guidance and direction whenever direction seemed unclear.

In an exceptional way I express my gratitude to Mr. and Mrs. Robert Mutyaba for their inspirational support right from the start in pursuit of this program of Master of Science in Embedded and Mobile Systems. Following the same singular honor, I do acknowledge the ever-present support and unwildering counsel of Dr. Opeken Chris.

The unmeasurable contributions of NM-AIST community, colleagues, classmates, and lecturers frolicked a substantial assistance in the pursuit of my studies at NMA-IST. Also I do recognize extraordinary support of the HIV support organisations (Infectious Disease institute, The Aids Support Organisation and Mildmay) which supported me with data provision to make this Project success.

I finally, in a very special way express my selfsame unfathomable gratitude to Mum and Dad, Mr. & Mrs. Bameka Emmanuel, my two Children Khisa Monica and Emmanuel Blessed Favor, my dear friend Miss Olivia Namusoke for their unwavering support and incessant motivation all through the study and research journey of this Project. You all were pivotal to the writing and completion of this Project report.

Thank you all.

## **DEDICATION**

This Project report is dedicated to my Late Mother Mrs. Jane Bameka for her unenduring social and moral support in this journey, without which success here would have not been possible.



## TABE OF CONTENTS

ABSTRACT .....	i
DECLARATION .....	ii
COPYRIGHT .....	iii
CERTIFICATION.....	iv
ACKNOWLEDGEMENTS .....	v
DEDICATION .....	vi
TABE OF CONTENTS .....	vii
LIST OF TABLES .....	xii
LIST OF FIGURES.....	xiii
LIST OF APPENDCES .....	xvi
LIST OF ABBREVIATIONS AND SYMBOLS.....	xvii
CHAPTER ONE .....	1
INTRODUCTION.....	1
1.1 Background of the Problem.....	1
1.2 Statement of the Problem .....	2
1.3 Rationale of the Project .....	2
1.4 Project Objectives.....	3
1.4.1 Main Objective .....	3
1.4.2 Specific Objectives.....	3
1.5 Project Questions.....	3
1.6 Significance of the Project.....	3
1.7 Delineation of the Project.....	5
CHAPTER TWO.....	6
LITERATURE REVIEW .....	6
2.1 Project Framework .....	6
2.2 Human Immunodeficiency Virus Antiretroviral Therapy in Africa.....	6
2.3 Human Immunodeficiency Virus Antiretroviral Therapy in Uganda .....	7

2.4	Use of Data in Human Immunodeficiency Virus Therapy in Uganda .....	8
2.5	Machine Learning in Disease Diagnosis .....	8
2.5.1	Benefits of Using Learning in Disease .....	8
2.5.2	Challenges of Using Machine Learning in Disease Screening .....	9
2.6	Related Machine Learning Prediction models of Communicable Diseases.....	9
2.6.1	Cardiovascular Disease .....	10
2.6.2	Cancer.....	10
2.6.3	Diabetes.....	11
CHAPTER THREE.....		13
MATERIALS AND METHODS .....		13
3.1	Project Jurisdiction .....	13
3.2	Infectious Disease Institute.....	13
3.2.1	Vision .....	13
3.2.2	Mission .....	13
3.2.3	Core Values .....	13
3.3	Mildmay Uganda .....	14
3.3.1	Vision .....	14
3.3.2	Mission .....	14
3.3.3	Core Values.....	14
3.4	The Aids Support Organization.....	15
3.4.1	Vision .....	15
3.4.2	Mission .....	15
3.4.3	Core Values .....	15
3.5	The Study Population .....	17
3.6	Data Collection Methods.....	18
3.7	Used Sampling Technique.....	18
3.8	Data Preprocessing .....	21
3.8.1	Numpy.....	23

3.8.2	Pandas.....	24
3.8.3	Missing no .....	24
3.8.4	Warnings .....	24
3.8.5	Category_Encoders .....	24
3.8.6	Pickle.....	24
3.8.7	Imputena.....	24
3.8.8	Smote.....	24
3.9	Data Modelling.....	36
3.10	Prediction Tool Development Life Cycle.....	39
3.10.1	Software Development.....	39
3.11	Development and Implementation of the Application Tool.....	45
3.11.1	System Development Languages and Technologies.....	45
3.12	Other Requirements.....	47
3.13	Assumptions and Dependencies .....	47
3.14	Summary.....	48
CHAPTER FOUR .....		49
RESULTS AND DISCUSSION .....		49
4.1	Introduction .....	49
4.2	Demographics of the Collected Data.....	49
4.2.1	Gender Composition .....	51
4.2.2	Age Distribution.....	51
4.2.3	Body Mass Index of the Population .....	52
4.2.4	Number of Pregnant Women.....	52
4.3	Human Immunodeficiency Virus Therapy analysis .....	53
4.3.1	Numbers Previous Regimen Initiation .....	53
4.3.2	Number of Patients Categorized by Dolutegravir Regimen.....	54
4.4	Infections Analysis .....	55
4.4.1	Tuberculosis-Status .....	55

4.4.2	Box and Whisker Plot of Viral Load.....	56
4.4.3	Box and Whisker Plot for the CD4 Count.....	56
4.4.4	Systolic Pressure and Diastolic Pressure.....	57
4.4.5	First Regimen Duration.....	58
4.5	Model Development Process, Results and Prospects for the Developed Systems.....	59
4.6	Model Evaluations and Validation.....	59
4.6.1	Performance Evaluations.....	59
4.6.2	Accuracy Evaluations.....	61
4.6.3	Other Classification Metrics.....	62
4.6.4	Model Classification.....	63
4.7	Summary.....	68
4.8	Selection of Predictor Variables.....	68
4.9	Optimal Feature Selection.....	69
4.10	Dolutegravir Associate Hyperglycaemia Prediction Tool.....	70
4.10.1	User Registration.....	70
4.10.2	Users.....	71
4.10.3	Validation.....	72
4.10.4	Validation of XG-Boost Model.....	72
4.10.5	Assessment of the Validation of XG-Boost Learning.....	72
4.10.6	Class Prediction Error for XG-Boost.....	73
4.10.7	Threshold Report for CG-Boost.....	74
4.10.8	XG-Boost Confusion Matrix.....	75
4.11	Unit Testing.....	76
4.12	Integration Testing.....	78
4.13	System Testing.....	78
4.14	User Acceptance Testing.....	79
4.15	Discussion.....	80
CHAPTER FIVE.....		82

CONCLUSION AND RECOMENDATIONS .....82

5.1 Conclusion.....82

5.2 Recommendations .....83

REFERENCES .....85

APPENDICES.....90

POSTER PRESENTATION .....95

## LIST OF TABLES

Table 1: Functional Requirements.....	41
Table 2: Non-Functional Requirements.....	41
Table 3: System Testing Results .....	78
Table 4: User Acceptance Results.....	80

## LIST OF FIGURES

Figure 1:	Project mapping framework of answers to project research questions .....	6
Figure 2:	A map of Uganda Showing Districts and where data was acquired .....	16
Figure 3:	The above map kampala shows headquarters of IDI, TASO and Mildmay .....	17
Figure 4:	Chart showing Data contribution by the three Organizations .....	20
Figure 5:	Dolutegravir Associated Hyperglycaemia Model Building process .....	23
Figure 6:	Visualization of the data frame variables.....	26
Figure 7:	Visualization of statistical descriptive analysis of the variables in the data frame ..	27
Figure 8:	Visualization of the data frame completeness and comprehensiveness.....	29
Figure 9:	Visualization of the categorical variable of the dataframe .....	31
Figure 10:	Visualization of the data frame out look after fitting it with categorical encoded data .....	33
Figure 11:	Visualization of Dataframe variable completeness and comprehensiveness after imputation .....	35
Figure 12:	Classification imbalance within the data .....	36
Figure 13:	Box plot comparison of models' performance accuracy.....	39
Figure 14:	Development Life Cycle of the DTG Associated Hyperglycaemia Prediction Tool	40
Figure 15:	Conceptual Frame Work for the Application.....	42
Figure 16:	Conceptual Framework for Visualization of learning patterns of the XG-Boost model .....	43
Figure 17:	Application User Contextual Diagram.....	43
Figure 18:	Use Case Diagram for the prediction Showing User .....	44
Figure 19:	Use Case Diagram Showing Administrator .....	44
Figure 20:	Application login Interface .....	46
Figure 21:	Application contact us Interface .....	47
Figure 22:	Shows data completeness .....	50
Figure 23:	Gender distribution .....	51

Figure 24:	Patients Age Distribution .....	52
Figure 25:	Body Mass Index Distribution among the HIV Patients.....	52
Figure 26:	Number of Pregnant HIV Patients .....	53
Figure 27:	First (Baseline) Regimen Treatment Analysis of Patients .....	54
Figure 28:	Current Regimen Treatment of Patients.....	54
Figure 29:	Number of TB-Infected HIV Patients .....	55
Figure 30:	The HIV patients ViralLoad Distribution .....	56
Figure 31:	The HIV Patients CD4 count Distribution.....	57
Figure 32:	The Distribution of Systolic Pressure and Diastolic Pressure of HIV Patients .....	58
Figure 33:	Period each HIV patient took on the First Regimen .....	58
Figure 34:	Creation function of the eight algorithms for different models (XG-Boost Classifier, Random Forest Classifier, SVC, Gaussian-NB, Decision Tree Classifier, KN-Neighbors Classifier, Linear Discriminant Analysis, Logistic Regression) .....	60
Figure 35:	Comparison accuracy of the Seven Models .....	60
Figure 36:	The XG-Boost Classification Report .....	63
Figure 37:	Random Forest Classification Report .....	63
Figure 38:	The ROC Curve for XG-Boost .....	65
Figure 39:	The ROC Curve for Random Forest Classifier.....	66
Figure 40:	Precision Recall Curve for XG-Boost Model .....	67
Figure 41:	Precision Recall Curve for Random Forest Model .....	67
Figure 42:	The feature importance bar graph of XG-Boost Model.....	69
Figure 43:	Graph of optimal number of features.....	70
Figure 44:	User Registration Interface .....	70
Figure 45:	User Management Interface.....	71
Figure 46:	The Dictionary Interface .....	71
Figure 47:	Prediction Interface Page .....	72
Figure 48:	Graph of XG-boost Learning Curve .....	73
Figure 49:	Validation graph of Class prediction Error for XG-Boost model.....	74



Figure 50: Validation threshold Report for XG-Boost .....75

Figure 51: Confusion Matrix of the XG-Boost Model .....75

Figure 52: Calibration Graph for the XG-Boost model.....76

Figure 53: Negative Prediction Results .....77

Figure 54: Positive Prediction Results.....77

## LIST OF APPENDICES

Appendix 1:	TASO Research Ethics Committee Approval Letter .....	90
Appendix 2:	TASO Administrative Clearance for Data Acquisition.....	91
Appendix 3:	Mildmay Administrative Clearance for Data Acquisition .....	91
Appendix 4:	Data Abstraction tool .....	93
Appendix 5:	Screening tool.....	94

## LIST OF ABBREVIATIONS AND SYMBOLS

AIDS	Acquired Immunodeficiency Syndrome
ART	Antiretroviral Therapy
BMI	Body Mass Index
CAD	Coronary Artery Disease
CSV	Comma Separated Values
DNN	Deep Neural Network
DTG	Dolutegravir
FBS	Fasting Bloods Sugar
IDI	Infectious Diseases Institute
INSTIs	Integrase Strand Transfer Inhibitors
MICE	Multivariate Imputation by Chained Equations
ML	Machine Learning
NRTIs	Nucleoside/ Nucleotide Reverse Transcriptase Inhibitors
Open-MRS	Open Medical Records System
PLHIV	People Living with Human Immunodeficiency Virus
RBS	Random Blood Sugar
ROC Curve	Receive Operating Characteristic Curve
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
TASO	The Aids Support Organization
TASO-REC	The Aids Support Organization Research Ethics Committee
UNAIDS	United Nations program on HIV and AIDS
USA	United States of America
WHO	World Health Organisation



# CHAPTER ONE

## INTRODUCTION

### 1.1 Background of the Problem

Human Immunodeficiency Virus (HIV) is among the largest pandemics in the world Since 1994. HIV causes Acquired Immunodeficiency Syndrome (AIDS). Prior to 1980s, research approximates about 100 000 to 300 000 people were living with HIV in Uganda (Rachel, 2021). In June 1982, a group of cases among gay men in Southern California suggested that the cause of the immune deficiency was sexually transmitted and in 1987 zidovudine was the first treatment for HIV. Consequently, treatments to reduce mother to child transmission were made by Scientists (Healthline Editorial Team, 2020). In the 1990's, the prevalence of AIDS cases had reached 307 000 and the real numbers believed to be one million in United States of America (USA) and worldwide around 8-10 million people were alleged to be living with HIV (Chin, 1991). The United Nations program on HIV and AIDS (UNAIDS) has been advocating for accelerated, comprehensive and coordinated global action on the HIV/AIDS pandemic. Thus, leading to reduced New HIV infections by 40% since the peak in 1998. In 2019, around 1.5 million (1.0 million–2.0 million) people were newly infected with HIV, compared to 2.8 million (2.0 million–3.7 million) people in 1998 (UNAIDS, 2021). Since the start of the AIDS epidemic in 1980, there has been research to show Uganda lowering its HIV/AIDS prevalence rate from 18% to 6.5% between 2014 and 2016 (Jay & Marta, 2018). In 2017 Uganda HIV and AIDS report showed the incidence rate of HIV is still down to 83 500 in 2015 from 95 000 in 2014 and 160 000 in 2010 (Ondoa, 2016).

However, several studies from 2005 to 2017 showed that HIV patients were becoming multi drug resistant (Kaleebu & Aceng, 2018). This has led to the increase in numbers of more new HIV infections and bring a burden to HIV health services delivery. The HIV/AIDS drug resistance led to the research into Integrase strand transfer inhibitors (INSTIs) and Nucleoside/nucleotide reverse transcriptase inhibitors (NRTIs) to improve antiretroviral therapy (Bhavik *et al.*, 2014). The INSTIs drugs include bicitgravir, dolutegravir (DTG), elvitegravir, raltegravir which are being adopted as the first line of HIV treatment today (Fabrice & Cotelte, 2015). In the due course of implementation of DTG INSTI based regimen as the first line of treatment, research has revealed a challenge of comorbidities among HIV patients (hyperglycaemia) mainly non-communicable diseases (Charlson, 2019). These non-communicable diseases are challenging to manage among HIV patients, therefore, the need to have early diagnosis, prediction and detection (Kansiime *et al.*, 2019). This is also proven by the challenges the Ugandan medical fraternity is facing in managing

HIV patients with non-communicable diseases (hyperglycaemia, diabetes and hypertension).

Despite the challenges above, the INSTI (bictegravir, dolutegravir, elvitegravir, raltegravir) drug regimens achieve HIV viral load suppression, reduce multidrug resistance and lower infection rate. Thus, world health Organisation advised countries to adopt Dolutegravir (DTG) base regimen (DTG in combination with other drugs) as the first line treatment, of which many sub-Saharan governments accepted and Uganda was no exception. The Uganda Ministry of Health (MoH) gave screening guidelines to be used, before switching an HIV patient to DTG. The screening guidelines include Random blood sugar (RBS) test, Liver function test (minimum ALT and AST) and Hepatitis B test. Hence any HIV patient who passes these tests (doesn't have diabetes, hepatitis B and liver related illnesses) is started on a DTG based ART regimen. The random blood sugar test is aimed at not switching hyperglycaemic HIV patients to DTG. This does not cater for those HIV patients who would turn hyperglycaemic if switched to DTG.

## **1.2 Statement of the Problem**

There is a countrywide switching of HIV patients from their previous ART regimens to a DTG based regimen as the first line of treatment, despite all the proof which research has shown that DTG based regimen cause Hyperglycaemia in some patients (Lirri, 2018). The screening guidelines given by MoH before switching treatment experienced HIV patients to a DTG based regimen identifies patients who have already developed hepatitis B, Hyperglycaemia, liver diseases (abnormalities) and hyperglycaemia. Nevertheless, they do not cater for screening of the HIV patients' possibilities of development of hyperglycaemia if initiated on DTG. Yet the common challenge, manifesting among some HIV patients switched to a DTG based regimen is developing Hyperglycaemia (Abebe *et al.*, 2014). Therefore, this study aims at collecting, processing and modelling HIV patients' data using machine learning algorithms and develop a screening tool to predict development of hyperglycaemia by an HIV patient if switched to DTG based regimen.

## **1.3 Rationale of the Project**

This being an era of evidence-based HIV care delivery in an information age where HIV ART and care services are data driven, the development and use of this tool purposes to provide a fundamental step in using data to guide transitioning of only HIV patients, who are less likely to develop hyperglycaemia to DTG based ART Regimen. This will also open up the treatment of clients with drugs which will improve their health and livelihood. Furthermore, this tool will help ART therapy and HIV care providers to give care and treatment with informed decisions bearing in mind the health outcomes of HIV patient in context.

## **1.4 Project Objectives**

### **1.4.1 Main Objective**

To appraise and develop a DTG associated hyperglycaemia prediction Tool for treatment experienced HIV patients.

### **1.4.2 Specific Objectives**

- (i) To identify and collect data for the risk factors for developing DTG associated hyperglycemia among HIV treatment experienced patients.
- (ii) To determine the most important features in the prediction DTG associated hyperglycaemia among HIV treatment experienced patient.
- (iii) To build and test the most appropriate DTG associated hyperglycaemia prediction model using machine learning classification methods.
- (iv) To validate the DTG associated hyperglycaemia prediction model and tool.

## **1.5 Project Questions**

- (i) How to appraise and develop a DTG associated hyperglycemia prediction tool for treatment experienced HIV patients?
- (ii) What are the risk factors for developing DTG associated hypoglycaemia among HIV treatment experienced patients?
- (iii) What are the most important features in the prediction of DTG associated hyperglycaemia among HIV experienced patients in Uganda?
- (iv) What is the finest machine learning classification algorithm for building the best DTG associated hyperglycaemia prediction model?
- (v) How to validate a DTG associated hyperglycaemia prediction tool?

## **1.6 Significance of the Project**

Research in HIV therapy ART made its breakthrough in 1990's with patients having relief and improvement in livelihood across the globe. Also, there has been increasing research breakthroughs in therapeutic treatment with the most recent being the Integrase strand transfer inhibitors.

However, all the above are still under research and development to find a permanent cure. The research breakthroughs now have brought hope to HIV patients amidst the challenge of sickness. Nonetheless these Integrase strand transfer inhibitors (DTG) have some serious side effects and at times trigger other chronic illness (hyperglycaemia) (Lo *et al.*, 2019). Consequently, the development of a DTG associated hyperglycaemia prediction tool will enable switching of only HIV patients who are less likely to develop hyperglycaemia to DTG based regimen ART. This will lessen the burden of HIV care and treatment on health service providers, thus giving health workers opportunity to maintain or improve the health of HIV patients.

In the recent past world health organization advice to countries across the globe to adopt DTG based regimen as the first line of treatment for HIV patients. Thus, necessitates the development and implementation of a DTG associated hyperglycaemia prediction tool critical for screening of treatment experienced HIV patients in Uganda before being switched to DTG. This tool can be used as a bench mark for development of tools for other countries because development of DTG associated hyperglycemia is a global challenge among people living with HIV initiated on DTG, despite the difference in the genetic makeup of people in countries across the globe.

Research in HIV care and treatment is still important in the ever-staggering HIV incidence rates, though screening for other possibly triggered illnesses due to treatment using a machine learning models can important for the HIV ART and care. The prediction tool will also help improve the screening for possibility of development of Hyperglycaemia among treatment experience HIV patients switched to DTG, because it will continue learning from the predictions it makes.

The prediction tool will be used by clinicians and doctors to make their work efficient and help improve quality of health of HIV patients in care by reducing the threat of increasing difficult to manage HIV patients with DTG associated hyperglycaemia. In essence, this study will include the following contributions:

- (i) This study enlightens HIV ART therapy practitioners (Doctors, Nurses and Clinicians) on the use of machine learning algorithmic models, in the development of medical screening tools, to help in diagnosis and prediction of DTG associated hyperglycemia among HIV patients being switched to DTG in Uganda. This helps doctors and health practitioners devise means of mitigating the hyperglycemia challenge early before it happens.
- (ii) This study moreover presents a prediction tool to be used by doctors, clinician's researchers plus the whole scientific fraternity to screen and find HIV patients who would develop



hyperglycaemia when switched to DTG based regimen. This will help clinician and doctors to switch to DTG only HIV patients that won't develop hyperglycaemia.

## **1.7 Delineation of the Project**

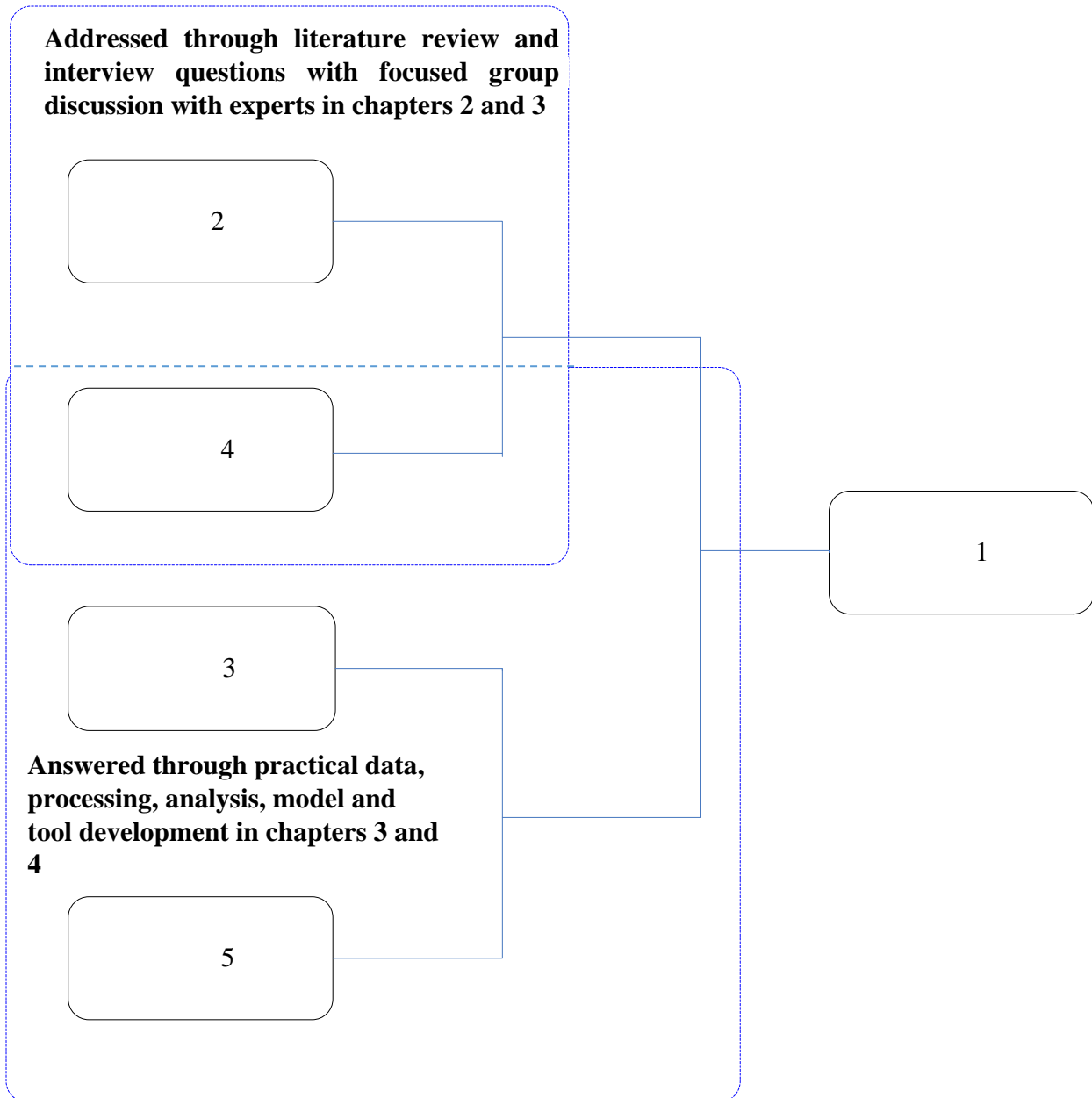
The study considered and used secondary data variables of HIV patients who were previously on other first line regimens for at least six months and currently on DTG, never switched to second line regimen, and had no diabetes. This is because the tool is to be used to predict development of hyperglycaemia of treatment experienced HIV patient is switched to DTG. Treatment naïve patients initiated on DTG at baseline have been excluded because there is less data for such criteria to support building a model as this will bring model bias or even leave the model “data hungry”. Also, the HIV patients who have never been switched or initiated onto DTG were excluded. The study is to develop a prediction tool for screening treatment experienced patients for a possibility to develop hyperglycaemia if they are switched on DTG. However, it does not determine and compare the probabilities of developing hyperglycaemia if you are switched on other INSTI regimens (bictegravir, elvitegravir, raltegravir) to give a clinician opportunity to select the best regimen for the HIV patient.

## CHAPTER TWO

### LITERATURE REVIEW

#### 2.1 Project Framework

The present Project work was guided by the mapping frame as shown in Fig. 1 during addressing the research questions of the Project.



**Figure 1: Project mapping framework of answers to project research questions**

#### 2.2 Human Immunodeficiency Virus Antiretroviral Therapy in Africa

There is still a surging prevalence of HIV in Sub-Saharan Africa with no effective cure for HIV at present though can be suppressed by a combination of medicines called antiretroviral (ARV)

therapy consisting of three or more ARV drugs (Regimen) (Sidibe, 2018). The registered decrease in HIV-related deaths in African Region largely as a result of steady scale-up of antiretroviral therapy (ART). In 2018 it was estimated that around 16.3 million people in Sub-African had access to treatment which catered for at least 64% of the total estimated number of people living with HIV who have access to antiretroviral therapy globally. Furthermore, 52% of PLHIV on treatment had a suppressed viral load that keeps a strong immune system working to prevent illness. Furthermore, viral suppression reduces chances of HIV transmission through sex, needle sharing and from mother to child (during pregnancy, at birth and breastfeeding) (WHO Regional Office Africa, 2021). In the recent past, world Health Organization had given guidelines for sub-Saharan countries to use Dolutegravir (DTG) based regimen as the first line of treatment for HIV patients (Chan, 2016).

However, worldwide there is a rise in the prevalence of non-communicable diseases among HIV patients (Lelani *et al.*, 2019). Unfortunately, there is no clear data to show the burden of non-communicable diseases in sub-Saharan Africa (Gouda *et al.*, 2019). Despite efforts to suppress HIV, there is rise in non-communicable diseases among people living with HIV in sub-Saharan Africa (Levitt *et al.*, 2011). According to research done in year (2017- 2018) there is increase in the prevalence of diabetes among people living with HIV compared to those without (Gilmaria *et al.*, 2020).

### **2.3 Human Immunodeficiency Virus Antiretroviral Therapy in Uganda**

Uganda started with support and care of People living with HIV in the 1980's. During that time, there was limited knowledge about HIV ART therapy. As research progressed in HIV treatment, the government of Uganda and WHO started programs of ART therapy access to all HIV patients in Uganda. The WHO guide lines adopted by Ugandan government policy of test and treat also increased easy access to ART. In Uganda today about 1.5 million people are living with HIV, and 1.2 million people living with HIV are on ART therapy which represents coverage of greater than 90% of the people living with HIV being on ART (UNAIDS, 2020). Uganda has also adopted the WHO guidelines and is also transitioning treatment experienced HIV patients on to DTG. In March 2018, Ugandan Government recommended initiating treatment naïve and experienced patients onto Dolutegravir based regimens. But according to Antonella *et al.* (2016), the treatment experienced HIV patients developed hyperglycaemia. The spring study of treatment naïve HIV patients also revealed development hyperglycaemia by HIV patients (Antonio & José, 2015). According to Ddamulira *et al.* (2020), the prevalence of non-communicable diseases is increasing among people living with HIV in Uganda at alarming rate. Thus, studies on patients that were

initiated by infectious disease institute Makerere (Kampala, Uganda) showed that there were treatment naïve and treatment experienced HIV patients who developed hyperglycaemia after being initiated on DTG base regimens (Lamorde *et al.*, 2020).

## **2.4 Use of Data in Human Immunodeficiency Virus Therapy in Uganda**

There is use of data to guide HIV ART therapy implementation by various non-governmental health organizations in Uganda. A lot of statistical data has been largely used under monitoring and evaluation to guide, improve and successfully provide HIV ART among people living with HIV in Uganda. This has been achieved through evidence-based health service provision or implementation policies put in place by funders /Donors and government alike. Consequently, there is statistical data about HIV incidence rate, prevalence, ART coverage and the demographics among people living with HIV to show impact made by the interventions in place (Opito *et al.*, 2020). These have been all geared towards achieving the ambitious treatment target to help end the AIDS epidemic called UNAIDS goals of 90-90-90 (UNAIDS, 2017). Furthermore, statistical information is used to guide health workers in the provision of HIV ART to people living with HIV thus helping them attain a healthy and long life (Mutabazi *et al.*, 2014).

However, there has been no adoption and use of data with machine learning algorithms to aid screening of people living with HIV before initiating them on ART therapy, which would greatly contribute to achieving the UNAIDS 90-90-90 aims of putting an end to HIV epidemic and also enable patients live a quality health life.

## **2.5 Machine Learning in Disease Diagnosis**

Machine learning has been used for development of prediction tools of other diseases in the health domain to its benefit though they posed some challenges.

### **2.5.1 Benefits of Using Learning in Disease**

The use of machine learning therefore, can considerably advance diagnosis, treatment and prognosis of diseases in patients (Schaefer *et al.*, 2020). Hence, machine learning prediction models would help isolate those HIV patients that would develop hyperglycemia from those that won't if switched to DTG ART. On that account if a machine learning prediction tool is adopted during HIV ART administration, it provides timely, accurate, and relevant information to clinicians and doctors decision making of who is to be initiated on DTG and who isn't. Leading to a proliferation quality of health and long life for people living with HIV.

Machine learning plays a critical role in interpretation of chronic diseases (Battineni *et al.*, 2020), and implementation of a predictive analytic system to predict the likeliness of having asthma and extent of diabetes in an individual (Ritesh, 2015) demonstrates that how critical machine learning models are in the prediction chronic illnesses to help to aid in deliberations of accurate informed decision.

The DNN deep learning model of infectious diseases helps to look into the future and plan for such disease interventions with medical supplies, logistical support and medical personnel in a more effectively and precisely manner (Sangwon *et al.*, 2018). Fatma *et al.* (2020) developed a model for detection of hypertension (cardiovascular disease) with high specificity, sensitivity, error, accuracy and performance which predicts the likely of developing hypertension thus supporting hypertension management and Medicare provision. Hence the need to have need a DTG associated hyperglycaemia prediction tool to support switching to DTG in HIV ART therapy

### **2.5.2 Challenges of Using Machine Learning in Disease Screening**

The timely translation of ML research into clinically endorsed and properly structured controlled systems which help everybody is a challenge. Moreover, there are challenges of required further work to identify: (a) algorithmic bias and unfairness while developing mitigations to address disease prediction and diagnosis, (b) reducing brittleness at the same time improving generalizability, and (c) to advance procedures for enriched interpretable machine learning predictions and achievement to those goals will cause transformational roles for patients (Kelly *et al.*, 2019). This issue points out the need to have tailored ML prediction models to answer a specific challenge the necessity to build a DTG associated hyperglycaemia prediction tool.

There is a challenge of ML approaches to convey results in an easy form for the health care professionals to comprehend and interpret the outcome buoyancy. That should a cause for machine learning to a substitute for clinicians and doctors because patients always want the human touch of an empathetic relationship with a medical professional providing care (Deshmukh, 2020). At the backdrop of such challenges this research aims to develop a DTG associated prediction tool whose results are easy to understand and interpret by the health experts (Doctors, nurses and clinicians).

### **2.6 Related Machine Learning Prediction models of Communicable Diseases**

There is a lot of research carried to develop machine learning models for communicable diseases like Cardiovascular disease, Cancer and Diabetes.

### **2.6.1 Cardiovascular Disease**

Cardiovascular disease is the frenzy of blood vessels and heart. It is among the top causes of demises globally. Machine learning Support vector machine (SVM) models are auspicious for predicting coronary artery disease (CAD) and stroke risk, even though there is further research needed to compare human expertise and ML models (Krittanawong *et al.*, 2020), which indicate that machine learning can serve exclusively as a predictor of CAD instead of a diagnostic tool. It is against such evidence that this study is undertaken to develop a DTG associated hyper glycaemia prediction tool.

Machine learning prediction of coronary artery disease (CAD) advances in three steps: (a) model exploration; (b) model refinement; (c) and monitoring and maintenance which improves on efficiency and accuracy (Akella & Akella, 2021). Random forest model can be used to attain great precision established on feature and pattern selection in cardiovascular events. This research offers proof and guidance why it is critical for the researcher to investigate the best classification model to be used to develop the DTG associated hyperglycemia.

Although there are limitations that are experienced can be overcome by using large-scale real-world datasets to build a model and prediction tool (Balakrishnan *et al.*, 2021). This research has chosen use of HIV treatment experienced participants' large dataset, to address such challenges right at development through model threshold discrimination, cross validation and calibration.

### **2.6.2 Cancer**

Cancer is a disease that is genetic and also caused by adverse events, environment and in the recent past has claimed a number of lives. It can be predicted the following ways: (a) the prediction of cancer susceptibility, (b) the prediction of cancer relapse, (c) the prediction of cancer survival. The ML concepts were defined and their use in cancer predictions. Studies that have been proposed focusing on the development of predictive models using supervised ML methods and classification algorithms targeting to predict valid disease outcomes using multidimensional heterogeneous data (Kouroua *et al.*, 2015). In the prediction events this disease Multi-classification model was used developed and used. This is similar to this research, on the choice of bi-classification model to be built and used of the prediction tool.

Machine learning algorithms can be applied on data from a cancer database to predict outcomes. Improved predicted results have the potential to support clinicians or doctors take precise decisions concerning treatment and or contribute in the development of prospect social and care needs (Gupta *et al.*, 2014). This can help greatly health workers in regions where screening equipment for cancer

are rare and lives can be saved. Those benefits even make this study essential given that Uganda is a low developed country with similar challenges or even worse in HIV ART therapy, care and support.

The use of machine learning algorithms in the classification of women having and those not having breast cancer showed great precision and accuracy compared to sophisticated-model-based approaches. As well efficiency and accuracy in prediction methods are vital for personalized medicine because they facilitate categorization of prevention strategies and personalized medical management (Ming *et al.*, 2019). Also, this study main goal is to develop a DTG hyperglycaemia prediction tool to be used to identify HIV treatment experienced would turn hyperglycemic if switched to DTG patients. That acts as prevention for DTG associated hyperglycaemia as only patients are transitioned to DTG.

### **2.6.3 Diabetes**

Diabetes mellitus is a chronic non-communicable disease that causes many complications which have claimed many lives of people and research has also sighted that machine learning will play a critical role in prediction of the diseases (Woldaregay *et al.*, 2019). Research also has shown that having asthma and extent of diabetes in an individual can be predicted by machine learning predictive analytic system (Razvan *et al.*, 2013). Also, research in early detection of diabetes has shown that there is a possibility of building a system using logistic regression models to predict future occurrence of diabetes in patients and provide necessary preventive measures (Maniruzzaman *et al.*, 2018). That is why this study is critical in the adoption necessary preventive measures for DTG associated hyperglycaemia.

In the prediction of Diabetes Mellitus, fasting glucose is an important index among all other indices that are required. Using three different classification algorithms like random forest, decision tree, and neural network; random forest was better than others in prediction accuracy. The best performance result of pima Indians in prediction of diabetes is 0.7721 and random forest result is 0.8084 for luzhou dataset, which indicated machine learning being better at predicting diabetes mellitus (Zou *et al.*, 2018). Therefore, machine learning prediction tools are dependable and more efficient in the prediction of hyperglycaemia events.

Research has shown that machine learning techniques are critical in predictive analytics over big data of diabetic patients in health care systems, which is helpful to health care practitioners in making informed decisions about patients' health and treatment. The performance and precision of these decisions depend on the applied Machine Learning algorithms for the early prediction of

diabetes. It was further concluded that Machine Learning random forest Algorithms are more precise in the early prediction of diabetes (Sarwar *et al.*, 2018).

Therefore, a lot of research has been made to come up with models that can predict development of diabetes in patients using Machine learning models, but there is no research and tool developed to predict DTG associated hyperglycaemia in HIV patients. This is because it's an emerging challenge that came because of adopting DTG based regimens as a first line of treatment among people living with HIV in Uganda and perhaps sub-Saharan Africa since 2017. Thus, my interest in developing a DTG associated hyperglycaemia prediction tool to help in screening treatment experienced HIV patients before being switched to DTG.



## **CHAPTER THREE**

### **MATERIALS AND METHODS**

#### **3.1 Project Jurisdiction**

The study covered three leading HIV support, care and Therapy organisations (Infectious Disease Institute, Mildmay and The AIDS Support Organisation) in Uganda. These randomly chosen organisations have their branches in various regions of the country which included Eastern, western, Southern, Northern. However, they all have their headquarters in Kampala.

#### **3.2 Infectious Disease Institute**

Infectious Disease Institute (IDI) was founded by the Academic Alliance for AIDS Care and Prevention in Africa and its headquarters are found at Makerere University, IDI-Mckinnell knowledge Centre. The Organisation performs training of health care workers, treating HIV patients and conducting cutting-edge research through the Academic Alliance. They aim at providing excellent care for People Living with HIV (PLHIV) in Uganda and Africa, maintaining strategic emphasis on HIV prevention, and conducting research relevant to improving the outcome of the HIV epidemic. The IDI has 322 124 people living with HIV on ART therapy. Their vision, mission and core values of IDI are:

##### **3.2.1 Vision**

A healthy Africa free from the burden of infectious diseases

##### **3.2.2 Mission**

To strengthen health systems in Africa, with a strong emphasis on infectious diseases, through research and capacity development.

##### **3.2.3 Core Values**

The core Values are:

- (i) Caring
- (ii) Integrity
- (iii) Excellence

- (iv) Innovation
- (v) Teamwork
- (vi) Accountability

### **3.3 Mildmay Uganda**

Mildmay is an Organisation established in 1998 as a Centre of Excellence for provision of comprehensive HIV&AIDS prevention, care, treatment and training with its headquarters located along Entebbe road, Naziba hill, Lweza, Kampala. The organization's diversification and evolvement into comprehensive healthcare, family- centered approach to prevention care, and treatment. That made the organisation to be recognized by World Health Organisation as the organisation with the best practice in HIV prevention and care. Mildmay has its programmes concentrated in 16 districts within Uganda and still expanding to other areas in Uganda. About 259 789 people living with HIV are under care at Mildmay Uganda. The Vison, Mission ad Values of Mildmay are:

#### **3.3.1 Vision**

To transform communities for sustainable health

#### **3.3.2 Mission**

To empower communities for health and sustainable livelihoods by providing quality healthcare, developing human resources for health and generating evidence to influence health policy.

#### **3.3.3 Core Values**

- (i) Integrity
- (ii) Customer Centricity
- (iii) Innovation
- (iv) People Development
- (v) Open Communication

### **3.4 The Aids Support Organization**

The Aids Support Organization (TASO) is an HIV and AIDS therapy organization, established in 1986/87 as a support group for people living with HIV and AIDS at Mulago National Referral Hospital kampala. To-date its headquarters are at Mulago National referral hospital. It was established with a fundamental objective of creating critical based community-based support to respond to the AIDS epidemic through providing counseling services to HIV infected persons and their families. The TASO has support centers in all the five regions (East, West, North, South and Central) of the country with 435 631 HIV patients on ART therapy under its care and support. The Vision Mission and core values are:

#### **3.4.1 Vision**

To see World without HIV/AIDS.

#### **3.4.2 Mission**

To contribute to a process of preventing HIV infection, restoring hope and improving the quality of life of persons, families and communities affected by HIV infection and disease.

#### **3.4.3 Core Values**

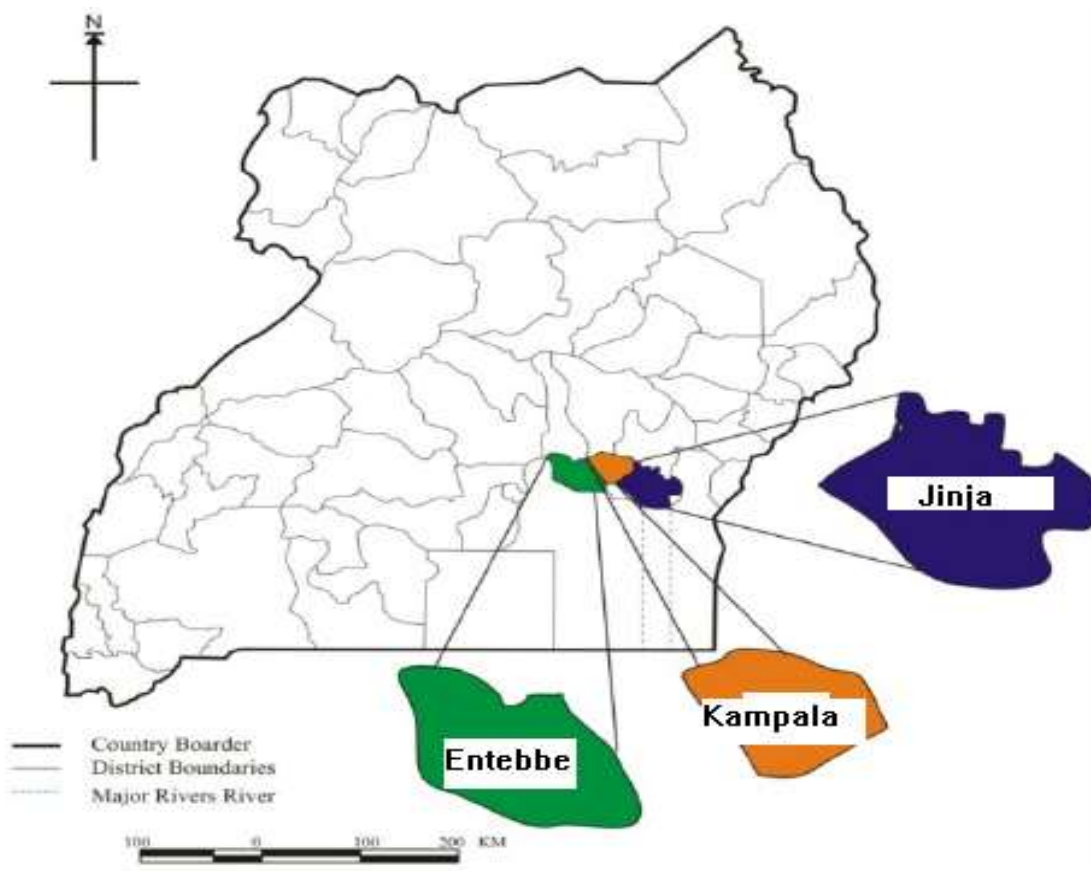
- (i) Obligated to people infected and affected by HIV and AIDS
- (ii) Integrity
- (iii) Family Spirit
- (iv) Equal rights, equal opportunities and shared responsibility
- (v) Human dignity

The research data from these organisations was collected urban and peri-urban. This is because DTG is a recently world health organisation approved regimen to be adopted as the first line of HIV ART Therapy, it had not been widely used except these pilot districts of Kampala, Jinja and Entebbe.

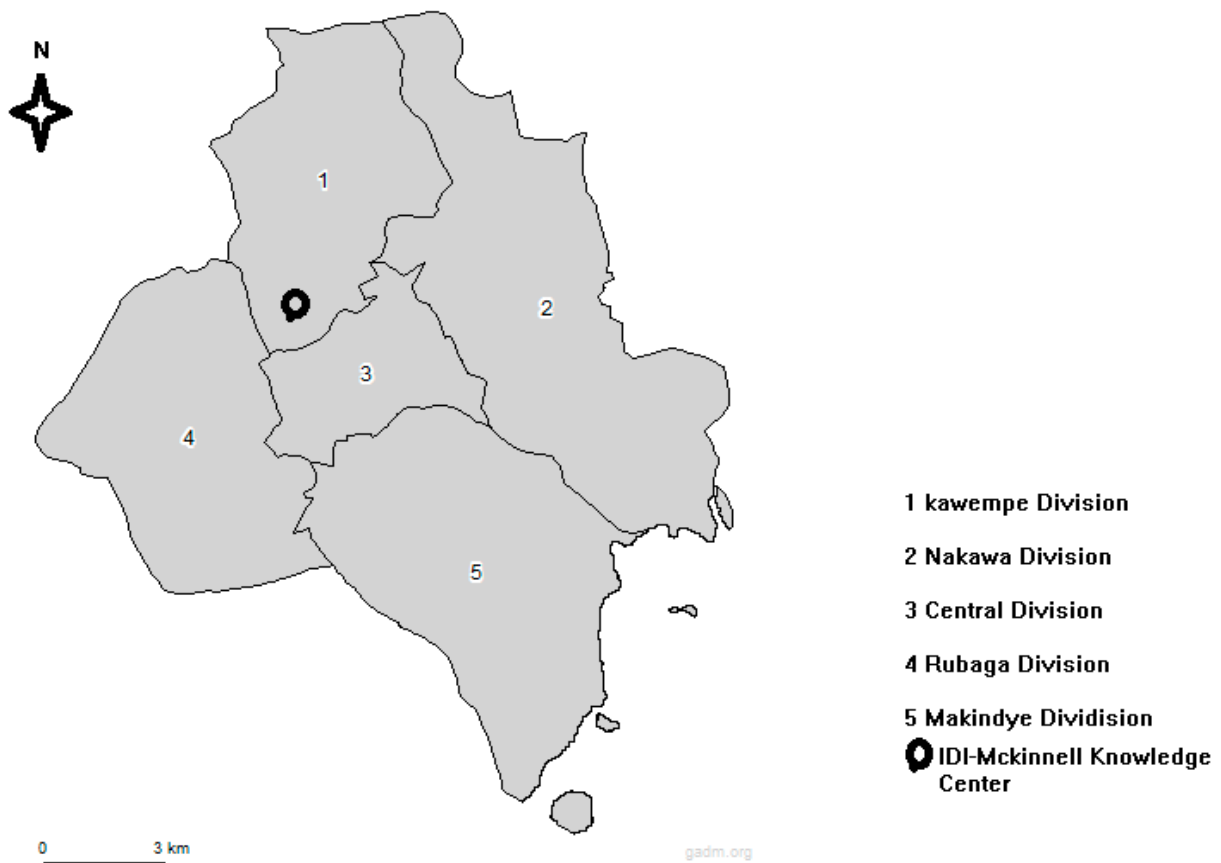
The study predominantly constituted of acquired secondary data and of treatment experience HIV patients already switched on DTG. Though to identify the risk factors of DTG associated hyperglycaemia and assessing its challenges secondary and primary data was used. Secondary data comprised of published articles, information from websites, conference proceedings, HIV reports,

electronic databases, and electronic sources *inter alia*. While primary data comprised of Responses from interview questionnaires and focus group discussions.

The maps in Fig. 2 and 3 show districts where data was acquired from and the locations of the organisations headquarters.



**Figure 2: A map of Uganda Showing Districts and where data was acquired**



**Figure 3: The above map kampala shows headquarters of IDI, TASO and Mildmay**

### **3.5 The Study Population**

The population of interest selected for the present Project was categorized as follows:

- (i) **Participants:** According to the research objectives, the HIV treatment experienced patients already switched on DTG would provide enough data to assess the DTG associated hyperglycemia and develop the screening tool.
- (ii) The Project focused on the leading HIV ART therapy, care and support organisations: because these have large datasets critical for the development of the DTG Associated Hyperglycaemia prediction tool and they are direct beneficiaries of this kind of innovation in their line of work.
- (iii) The study also considered time duration on DTG of at least three months to capture the minimum duration pharmacokinetics process to take place in the body.
- (iv) Diabetic before switching to DTG and those on the second line of treatment HIV patients were excluded; diabetic patients before switching to DTG meant that their hyperglycemic conditions were due to other causes other than DTG and those patients on the second line hard underlying causes which caused their switching to the second line of treatment. Yet

this study is only interest in only patients switching to a DTG based regimen as their first line of treatment.

- (v) Geographical choice of Entebbe, Kampala and Jinja areas was because they comprised of towns and peri-urban areas which are representative of places that are worst hit by HIV epidemic and they were among the first locations to switch HIV patients DTG as their first line of ART therapy.

### **3.6 Data Collection Methods**

The researcher used interviews questions and focused group discussions with HIV ART therapy experts to enhance depth of research quality. Interviews were held with HIV ART therapy experts (Doctors, Clinicians and Nurses) having wealth of experience in management of HIV and related illness; to seek opinion on the risk factors for development of DTG associated hyperglycaemia, seek policy trends for the screening procedure to switch treatment experienced HIV patients to DTG. Precisely and discuss issues ranging from:

- (i) The trending screening procedure for treatment experienced HIV patients being switched to DTG.
- (ii) The risk factors of DTG associated hyperglycaemia.
- (iii) Importance of DTG in ART therapy in the HIV support care and treatment.
- (iv) Contribution of Information technology in switching HIV patients to DTG.
- (v) The existing Government policy on switching HIV patients to DTG.
- (vi) Their Views on using prediction tools to screen HIV patients before switching to DTG.

The interviews used because of experts' handful schedules and limited time to respond to questionnaires.

### **3.7 Used Sampling Technique**

Basing on the research objectives, the study adopted non-probability which necessitated choosing treatment experienced HIV participants who met the confounding criteria not at random. This come with the advantages of: (a) The research being certain of collecting a dataset descriptive of participants of interest, (b) Easy computation of sampling error, and (c) The outcome can be used for prediction on the same population.

Using the experts and literature review information an abstraction data collection tool was designed to be used in the collection of secondary data for the treatment experienced HIV patients on DTG who met the confounding criteria.

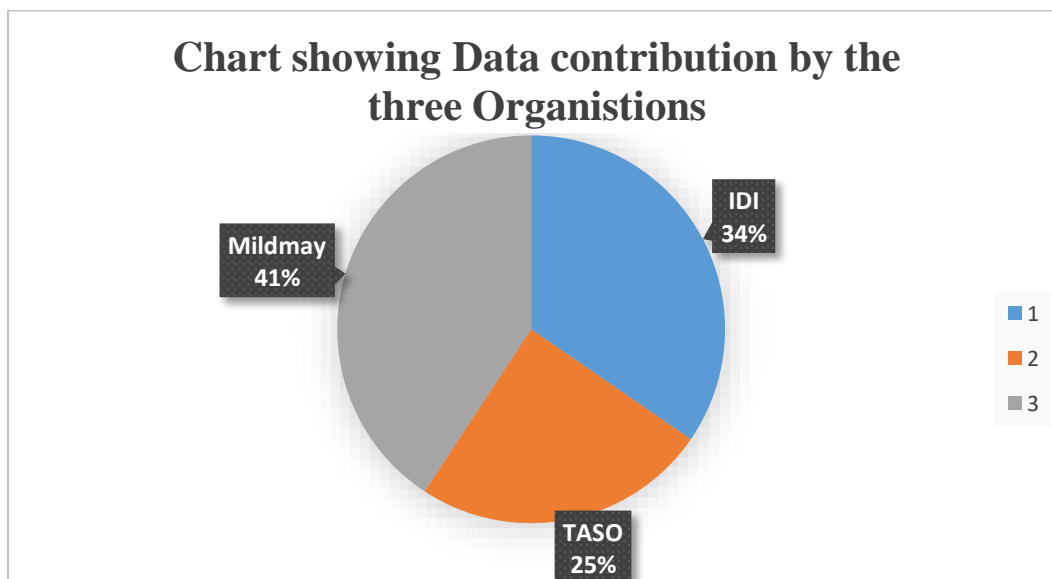
The data abstraction tool herein as Appendix 4, contained the risk factors for DTG associated hyperglycaemia which was secondary data for chosen participants in the study. Authorisation to access secondary data for the study participants in this research was sought from TASO Research Ethical Committee (TASO-REC), herein as Appendix 1. This was followed by seeking and acquiring administrative clearances from the chosen HIV ART therapy, care, and support organisations (TASO, Mildmay and IDI) to acquire/collect the data of participants of interest herein as Appendix 2 and 3. Hyperglycaemia is a condition that occurs when the Random Blood Sugar (RBS) of an individual is 200 mg/dl and above (11.11 mmol) or when the Fasting Blood Sugar 165 mg/dl and above (9.17 mmol). The DTG associated hyperglycaemia is hyperglycaemia contracted as a result of using DTG as a regimen for HIV ART therapy.

The study participants' descriptive variables identified to correspond to the risk factors for DTG associated hyperglycaemia were listed to form the secondary data collection abstraction tool herein as Appendix 4. The abstraction tool had the following data variables:

- (i) Family medical history of hypertension
- (ii) Family medical history of diabetes mellitus
- (iii) History of alcohol and Substance Abuse
- (iv) Smoking
- (v) Age
- (vi) Gender
- (vii) Baseline or Previous Regimen
- (viii) Most recent Height before switching to DTG
- (ix) Most recent Weight before switching to DTG
- (x) Duration taken on Baseline/Previous Regimen
- (xi) Most recent Diastolic Pressure before switching to DTG

- (xii) Most recent Systolic Pressure before switching to DTG
- (xiii) Pregnancy status at the time of switching to DTG
- (xiv) Most recent Viral Load before Switching to DTG
- (xv) Most recent CD4 Count before Switching to DTG
- (xvi) TB Status before Switching to DTG
- (xvii) Most recent FBS/RBS while on DTG
- (xviii) Most recent Insulin levels before switching to DTG
- (xix) Most recent Skin thickness before switching to DTG

The researcher was able to collect a total of 16 6066 records of data; where IDI provided 57 471 records, TASO 40 950 records, and Mildmay 67 645 records as illustrated in the pie chart Fig. 4



**Figure 4: Chart showing Data contribution by the three Organizations**

During data acquisition the researchers enjoyed some of these benefits of secondary data collection:

- (i) The data was readily available from a variety of sources
- (ii) Due to its availability the process became less time consuming
- (iii) The data acquisition process allowed the researcher to generate new insights from data.
- (iv) It facilitated longitudinal analysis which was a research objective.



- (v) The data collection process was easily executed because it never required any specialized training.
- (vi) The researchers acquired a significant amount of secondary data with a wide variety of sources among others.

Though as well the following challenges were faced:

- (i) There were bureaucratic tendencies most especially during research protocol ethical approval.
- (ii) The study lugged behind time schedule during administrative clearances, because no organisation availed the researcher administrative clearance for data acquisition without research ethical approval letter.
- (iii) The data was not specific to research needs there for had a lot of “*noise*” of unwanted information which for data processing
- (iv) The data collected had inadequacy tendencies in characteristics of quality for the study like:
  - (a) Completeness: as some key variables like systolic and diastolic pressure lack more than three quarters of their values.
  - (b) Comprehensiveness: were some of the values for data variables like skin thickness, Insulin levels, family history of hypertension, family history of diabetes, alcohol or substance abuse and smoking were un available.
  - (c) Inaccuracy tendencies: there instances with in some removed data records were men were breast feeding, pregnant among others.

Participants confidentiality was ensured by withholding patients’ names from the data provide to the researchers.

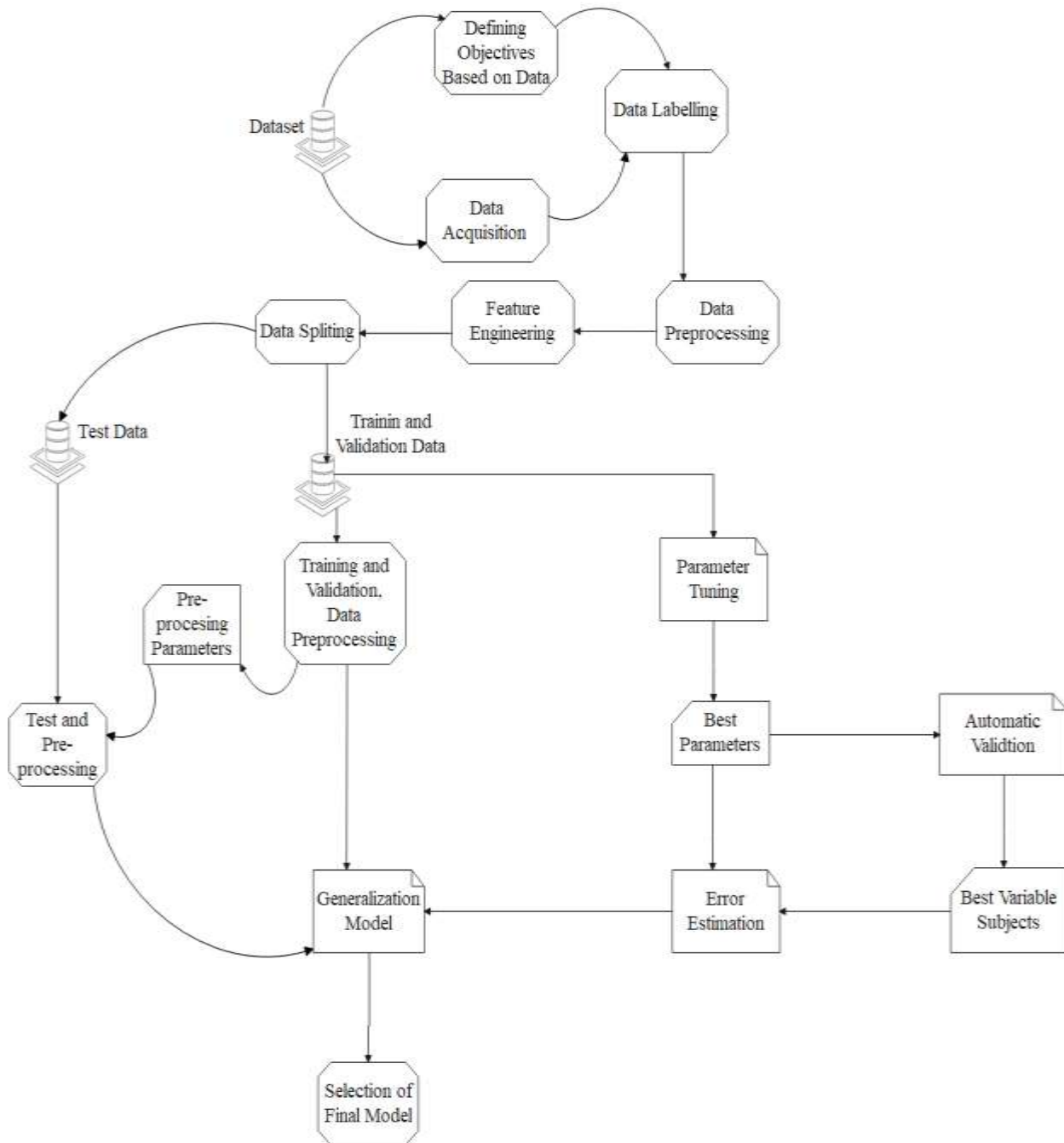
### **3.8 Data Preprocessing**

To further ensure data security and confidentiality of the acquired data, the researcher stored it on an external hard disk with encryptions and passwords which was accessible to authorised persons with the required passwords. During processing and modelling the researchers also assigned the participants with arbitrary unique identifiers and dropped the columns with participants unique

identifiers as provided from the organisations. This was aimed at further anonymising the participants data to ensure privacy. Fields like Age, duration taken on previous/Baseline Regimen and Body Mass Index (BMI) of population of interest were derived from date of birth, subtraction baseline regimen initiation date from DTG regimen initiation date, weight and height respectively. The data was then labelled, organised by the researcher in excel sheet and saved in a comma-separated values (CSV) format. This was done to make the data file ready for processing and manipulation in python pandas dataframe format using anaconda environment. The choice of the anaconda environment was influenced by the following benefits:

- (i) Anaconda is an open-source free capable of analysing large datasets with ease more than ordinary statistical tools ordinary.
- (ii) This environment has many data science and analysis packages to tap into which are fundament for this research.
- (iii) It eases model development, management and deployment.

We installed anaconda development environment and made sure it had the following data processing, analysis and manipulation libraries (Numpy, Pandas, Missingno, Warnings, Category\_Encoders, Pickle, Imputena, Smote) and data manipulation, visualisation and modelling libraries (matplotlib.pyplot, seaborn, sklearn.model\_selection) were import into the anaconda development environment. Below is the diagrammatic representation of the DTG associated hyperglycaemia data processing model building process in Fig. 5.



**Figure 5: Dolutegravir Associated Hyperglycaemia Model Building process**

After setting up the anaconda development environment with all the necessary libraries, we started and opened Jupiter notebook and created the development page. From the environment page we imported:

### 3.8.1 Numpy

That was used to work on the imported dataframe structure as array numbers in manipulation, working on data in domain of linear algebra, Fourier transform, and matrices.

### **3.8.2 Pandas**

It was used as a fast, powerful, flexible and easy data analysis and manipulation tool, built basing on Python programming language.

### **3.8.3 Missing no**

It was used to provide the ability to comprehend the distribution of missing values by use of revealing visualizations then check the correlation of columns missing values with the target column.

### **3.8.4 Warnings**

This library was used when writing scripts to temporarily suppress warning when you use deprecated functions.

### **3.8.5 Category\_Encoders**

It is a library under Scikit-learn-style transformers used to encode categorical datasets into numerical datasets.

### **3.8.6 Pickle**

This library module was important for serialization and de-serialization python object structures created during data processing and manipulation for model lifting.

### **3.8.7 Imputena**

This library was used when assigning values to missing values in variables of the data-frame by inference from the value processes to which it contributes.

### **3.8.8 Smote**

This is a library that was used to solve data classification imbalance problem which could make our model biased towards the majority class because it does not have enough data to learn about the minority. This is always a problem with disease prediction models as will be witnessed in the following steps ahead.

Then we imported the excel CSV file into jupyter notebook opened as a panda dataframe file with that line of script (`data = pd.read_csv('nudata_bmi$age_not_normalized_2.csv')`). We took a quick view of the first four records of the dataframe to visualise how the data looks like with the help of

this script (`data.head (4)`) as shown in the Fig. 6. The researcher used statistical descriptive analysis to understand the data distribution, total number, mean, standard deviation away from the mean, the quartiles, minimum and maximum values for the various data variables in the dataframe using (`data.describe()`) processing method. Figure 7 shows the described data.

selineRegimen	ViralLoad	TBStatus	Pregnant	CD4Count	CurrentRegimen	FirstRegimenDuration	bminorm	RBS	SystolicPressure	DiastolicPressure
TDF-3TC-NVP	8250.0	No signs or symptoms of TB	NOT APPLICABLE	NaN	TDF/3TC/DTG	4.0	0.1	9.92	NaN	NaN
TDF-3TC-NVP	8250.0	No signs or symptoms of TB	NOT APPLICABLE	NaN	TDF/3TC/DTG	4.0	0.1	11.70	NaN	NaN
TDF-3TC-NVP	8250.0	No signs or symptoms of TB	NOT APPLICABLE	NaN	TDF/3TC/DTG	4.0	0.1	11.48	NaN	NaN
TDF-3TC-NVP	8250.0	No signs or symptoms of TB	NOT APPLICABLE	NaN	TDF/3TC/DTG	4.0	0.1	12.58	NaN	NaN

**Figure 6: Visualization of the data frame variables**

	Age	ViralLoad	CD4Count	FirstRegimenDuration	bminorm	SystolicPressure	DiastolicPressure	Label
count	9077.000000	9.035000e+03	4540.000000	9072.000000	8977.000000	1807.000000	1807.000000	9077.000000
mean	44.704638	2.278218e+03	503.881872	8.644010	0.167822	122.006087	81.368013	0.177261
std	12.457151	3.838783e+04	290.569466	4.197451	0.383076	25.910645	23.282371	0.381911
min	4.000000	0.000000e+00	0.000000	-0.530000	-0.400000	9.000000	1.000000	0.000000
25%	38.000000	1.000000e+00	314.000000	5.770000	-0.230000	108.000000	71.000000	0.000000
50%	46.000000	1.000000e+00	475.500000	8.900000	0.200000	120.000000	80.000000	0.000000
75%	53.000000	5.000000e+01	661.000000	11.500000	0.480000	132.000000	89.000000	0.000000
max	92.000000	2.820000e+06	3536.000000	46.720000	1.100000	720.000000	855.000000	1.000000

Figure 7: Visualization of statistical descriptive analysis of the variables in the data frame

The dataframe was also processed with this command (`mno.martrix (data)`) to visualize the completeness and comprehensiveness of the variables of the dataframe. The visualization revealed diastolic pressure, Systolic pressure and CD4 count having significant random missingness shown in Fig. 8.



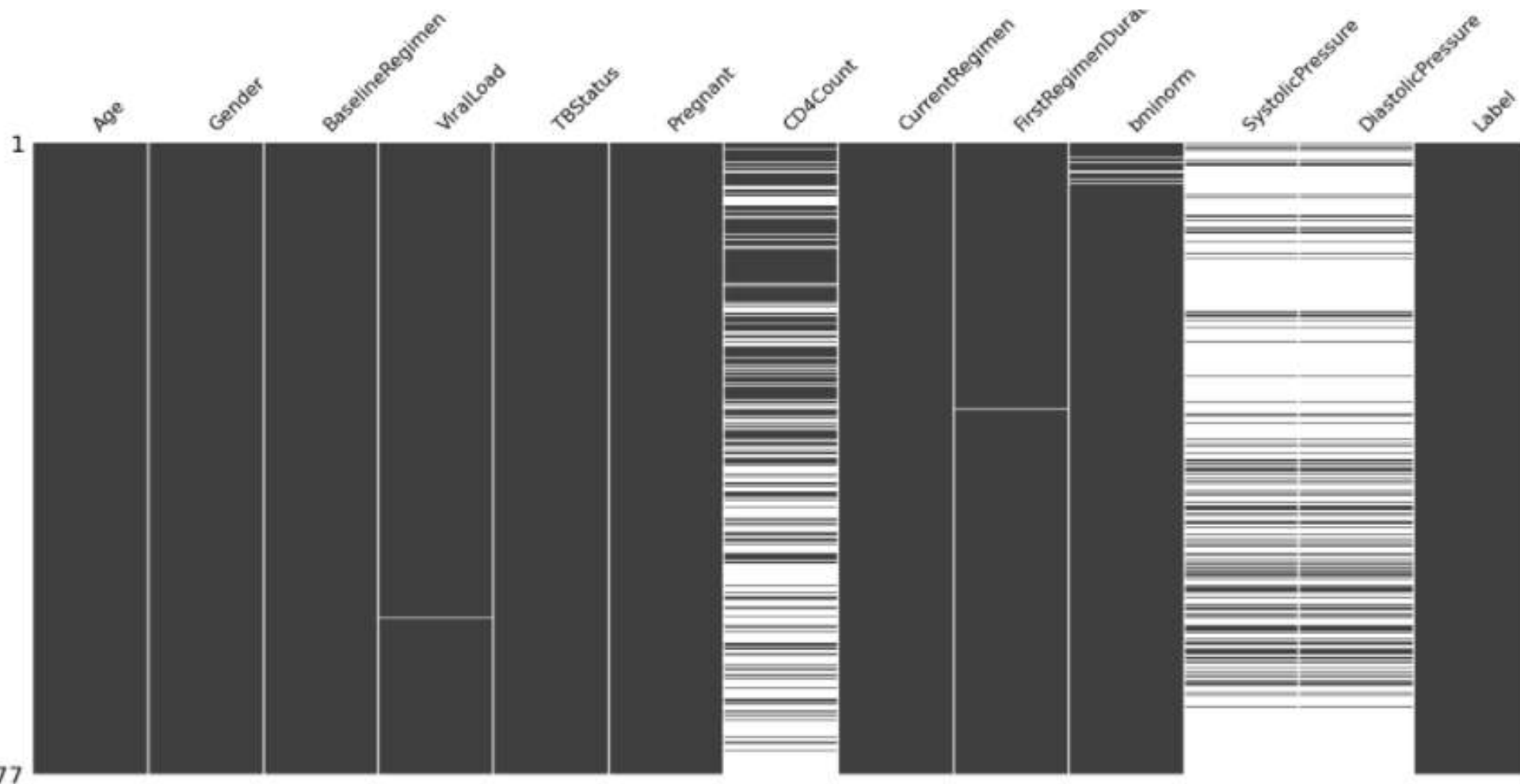


Figure 8: Visualization of the data frame completeness and comprehensiveness

This led to further processing of this data, to know the exact number of null values by columns or variables using this python command(`data.isnull().sum()`) giving the results shown below.

(i)	Age	0
(ii)	Gender	0
(iii)	BaselineRegimen	2
(iv)	ViralLoad	42
(v)	TBStatus	0
(vi)	Pregnant	0
(vii)	CD4Count	4537
(viii)	CurrentRegimen	0
(ix)	FirstRegimenDuration	5
(x)	bminorm	100
(xi)	SystolicPressure	7270
(xii)	DiastolicPressure	7270
(xiii)	Label	0
(xiv)	dtype: int64	

The outcome showed that, the research study is going into this modelling with data quality issues of completeness and comprehensiveness in the dataframe, as it was showed in the assessment that more than three quarters never had systolic and diastolic pressure data, half of the patients had no records of CD4-Count, 100 of the patients didn't have body mass index (BMI<sub>norm</sub>) data, 42 patients their viral-load records were null, 2 never had current regimen records, and 5 their baseline records were null.

The data visualisation made immediately after the importation of the data frame revealed that there were categorical variables in the dataset so the researchers had a brief detailed visualisation of them with this command (`data.select_dtypes(include = 'object')`) which gave results in Fig. 9.

	<b>Gender</b>	<b>BaselineRegimen</b>	<b>TBStatus</b>	<b>Pregnant</b>	<b>CurrentRegimen</b>
<b>0</b>	F	TDF-3TC-NVP	No signs or symptoms of TB	NOT APPLICABLE	TDF/3TC/DTG
<b>1</b>	F	TDF-3TC-NVP	No signs or symptoms of TB	NOT APPLICABLE	TDF/3TC/DTG
<b>2</b>	F	TDF-3TC-NVP	No signs or symptoms of TB	NOT APPLICABLE	TDF/3TC/DTG
<b>3</b>	F	TDF-3TC-NVP	No signs or symptoms of TB	NOT APPLICABLE	TDF/3TC/DTG
<b>4</b>	F	TDF-3TC-NVP	No signs or symptoms of TB	NOT APPLICABLE	TDF/3TC/DTG
...	...	...	...	...	...
<b>9072</b>	F	TDF-3TC-EFV	No signs or symptoms of TB	No	TDF/3TC/DTG
<b>9073</b>	F	TDF-3TC-EFV	No signs or symptoms of TB	No	TDF/3TC/DTG
<b>9074</b>	F	TDF-3TC-EFV	No signs or symptoms of TB	No	TDF/3TC/DTG
<b>9075</b>	F	TDF-3TC-EFV	No signs or symptoms of TB	No	TDF/3TC/DTG
<b>9076</b>	F	TDF-3TC-EFV	No signs or symptoms of TB	No	TDF/3TC/DTG

9077 rows × 5 columns

Figure 9: Visualization of the categorical variable of the dataframe

The above visualisation led to our knowledge and understanding of the number of different categorical values recorded under each of these different variables as displayed above using the following python scripts (`data.Gender.unique()`, `list(data.BaselineRegimen.unique())`, `data.TBStatus.unique()`, `data.Pregnant.unique()`, `data.CurrentRegimen.unique()`) and each had the following results as shown below.

Gender had these values:

```
array(['F', 'M'], dtype=object)
```

Baseline Regimen:

```
array(['TDF-3TC-NVP', 'others', 'AZT-3TC-NVP', 'TDF-3TC-EFV',  
      'ABC-3TC-EFV', nan, 'AZT-3TC-EFV', 'D4T-3TC-EFV', 'D4T-3TC-NVP',  
      'ABC-3TC-NVP', 'TDF-FTC-EFV', 'TDF-FTC-NVP', 'AZT-3TC-ALLV'],  
      dtype=object)
```

TB-Status:

```
array(['No signs or symptoms of TB', 'Currently on TB treatment',  
      'TB Treatment Completed', 'Suspect TB - referred or sputum sent',  
      'TB Diagnosed - TB LAM', 'TB Diagnosed - Gene Xpert'], dtype=object)
```

Pregnant:

```
array(['NOT APPLICABLE', 'No', 'No ', 'Not Applicable', 'Breast feeding',  
      'Yes'], dtype=object)
```

CurrentRegimen:

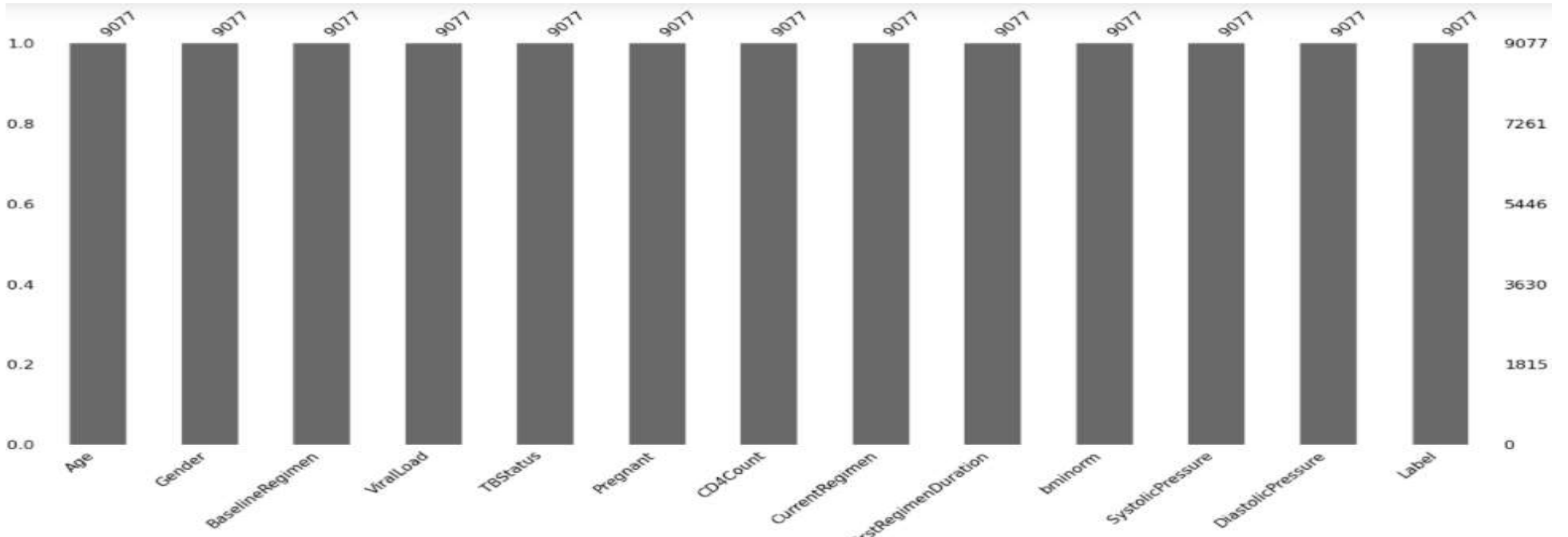
```
array(['TDF/3TC/DTG', 'ABC/3TC/DTG', 'AZT/3TC/DTG'], dtype=object)
```

The results displayed for each variable uncovered that the pregnant variable had some categorical values of ('No', 'No ') and (NOT APPLICABLE, Not Applicable) being recognised by python as different values yet they are the same. So we decided to solve that challenge by replacing one with the other using this script (`data.Pregnant.replace({'No ': 'No'}, inplace=True)`). Knowing that machine learning algorithms don't carry out computations with such categorical values the researcher used an ordinal encoding method to encode all the values for the above categorical variables. Then fitted and transformed the encoded categorical data values for the variables above into the dataframe. A visualisation method, was applied on the dataframe which gave the output Fig. 10.

	Age	Gender	BaselineRegimen	ViralLoad	TBStatus	Pregnant	CD4Count	CurrentRegimen	FirstRegimenDuration	bmin
0	13	0	0.0	8250.0	0	-1.0	NaN	0	4.0	
1	13	0	0.0	8250.0	0	-1.0	NaN	0	4.0	
2	13	0	0.0	8250.0	0	-1.0	NaN	0	4.0	
3	13	0	0.0	8250.0	0	-1.0	NaN	0	4.0	
4	13	0	0.0	8250.0	0	-1.0	NaN	0	4.0	
...	...	...	...	...	...	...	...	...	...	...
9072	21	0	3.0	0.0	0	0.0	NaN	0	0.0	
9073	46	0	3.0	1.0	0	0.0	NaN	0	4.4	
9074	27	0	3.0	1.0	0	0.0	NaN	0	0.0	
9075	35	0	3.0	1.0	0	0.0	NaN	0	0.9	
9076	27	0	3.0	0.0	0	0.0	NaN	0	0.1	
9077 rows x 13 columns										

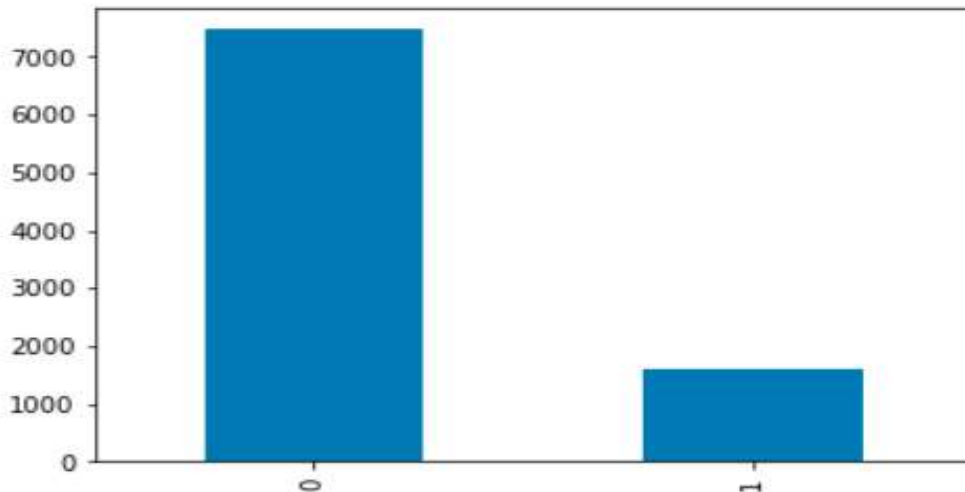
Figure 10: Visualization of the data frame out look after fitting it with categorical encoded data

At this point, still there was a challenge of missing or null values. This was solved using multivariate imputation by chained equations (MICE) algorithm. It works by specifying a multivariate imputation model on a variable-by-variable basis using a set of given densities, one for each incomplete variable. The MICE Starts from an initial imputation, and allures imputations by repetitions over the conditional densities. Number of iterations (like 10–20) may be enough (Brownlee , Iterative Imputation for Missing Values in Machine Learning, 2020). The methods used in imputing of null values was `(impt = imp.mice(dft,3))`, then later the researcher rendered the result into another dataframe using this python command `(dt_imp = pd.DataFrame(impt[0]))`, after visualisation of the dataframe in Fig. 11 was made to ascertain for completeness and comprehensive of variables.



**Figure 11: Visualization of Dataframe variable completeness and comprehensiveness after imputation**

The above graphic visualisation of the dataframe reveals that there are no anymore missing values. Thus, the dataframe was ready to be taken through further processing to check for a classification imbalance check which is a common phenomenon with data in the health sector especially for the sick (patients) being much less than those who are not normal. Which is a common occurrence when creating screening prediction models for diseases. A python script was run (`dt_imp.label.value_counts().plot(kind = 'bar')`) which revealed a significant classification imbalance challenge were the number of HIV patients with hyperglycaemia was slightly less than a quarter of those without hyperglycaemia as shown in Fig. 12.



**Figure 12: Classification imbalance within the data**

This imbalance challenge will be addressed in the steps that were followed while building models.

### 3.9 Data Modelling

At this moment we imported (from `sklearn.model_selection import train_test_split`) library to help us split the data at various stages of model development and testing. The dataset was split into two datasets (Train and Test datasets) but being mindful of making sure that they have equivalent label outcomes of zeros and ones. This was done using (`train, test = train_test_split(dt_imp, test_size = 0.25, shuffle = True, random_state = 42)`) method. A method (`train.shape, test.shape`) was also used to show the size of the two datasets as `((6807, 13), (2270, 13))`. The test dataset was kept for further use in performance evaluations of the built model and train dataset was further split into y-train the label column and x-train all other columns except the label by the method:

```
y = train.Label
x = train.drop(columns='Label')
```

But previously after using MICE algorithm the data values of X variables got a lot of infinite decimal places that could affect the model if the data is used to develop the model in its format,



therefore we defined a function below to make finite decimal places for all the values for variables of X.

```
def clean_dataset(df):  
    assert isinstance(df, pd.DataFrame), "df needs to be a pd.DataFrame"  
    df.dropna(inplace=True)  
    indices_to_keep = ~df.isin([np.nan, np.inf, -np.inf]).any(1)  
    return df[indices_to_keep].astype(np.float64)  
x = clean_dataset(x)  
np.all(np.isfinite(x))
```

To avoid biasing the model, data was again split the into Y (test and train) and X (test and train) and stratified by Y outcome variable to make sure that the datasets for model training and testing during development get equivalent classes of outcome values of ones for positive and zeros for negative.

Since we had earlier recognized a classification imbalance challenge in the data, we imported an imblearn python library with algorithm called Synthetic Minority Oversampling Technique (SMOTE), (from imblearn.over\_sampling import SMOTE) to help us in fixing the classification imbalance problem as shown in the data. The SMOTE works by oversampling the minority class duplicating data points in the minority class, although these data points don't add any new information to the model, instead new data points can be synthesized from the existing data points (Brownlee, SMOTE for Imbalanced Classification, 2021). The SMOTE library (smote=SMOTE (random\_state=42,sampling\_strategy='minority',n\_jobs=-1,k\_neighbors=6)) was called to make synthetic data points of the minority class as a way of alleviating the imbalance during modelling. A SMOTE method was applied to the dataset that is going to be used during model training by this script (X\_sm,y\_sm = smote.fit\_resample(X\_train, Y\_train)) and there after the researcher checked to see whether a balanced classification has been achieved using the python scripts:

```
print('After OverSampling, the shape of train_X: {}'.format(X_sm.shape))  
print('After OverSampling, the shape of train_y: {} \n'.format(y_sm.shape))  
print("After OverSampling, counts of label '1': {}".format(sum(y_sm==1)))  
print("After OverSampling, counts of label '0': {}".format(sum(y_sm==0)))
```

These scripts gave the output below showing that the issue classification imbalance is sorted.

After OverSampling, the shape of train\_X: (8928, 12)

After OverSampling, the shape of train\_y: (8928,)

After OverSampling, counts of label '1': 4464

After OverSampling, counts of label '0': 4464

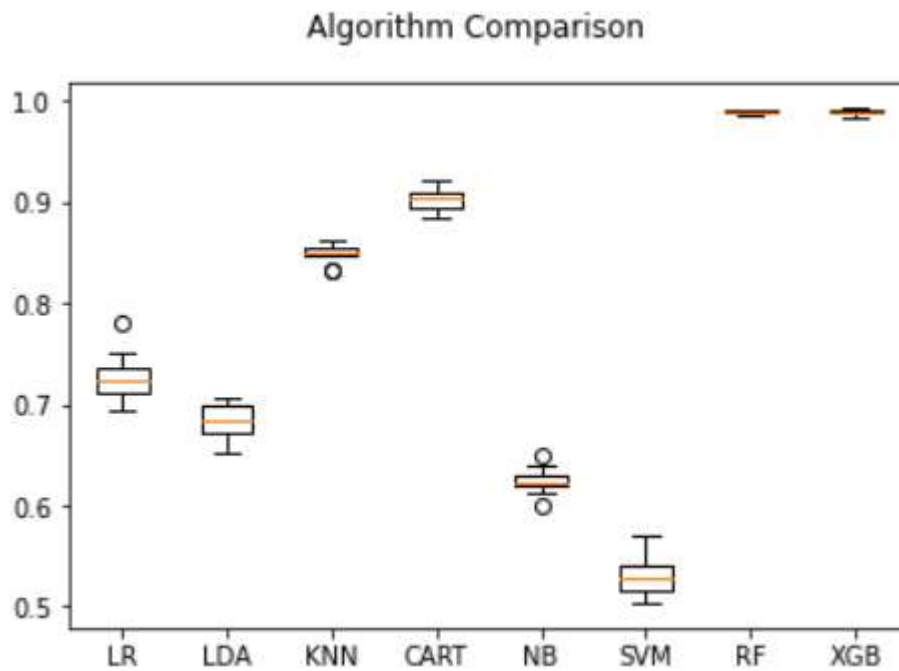
But at this point the researcher knew that the data which was created is at the bottom of the dataset so there was need to reshuffle the dataset again. To do the dataset shuffling, an algorithm to shuffle data was called from SKlearn.util python library and the (`data = shuffle(X_sm)`) method was used to shuffle the dataset. At this point the dataset was ready to be used for modelling but it had been recombined for the processing which took place as above, so we had to re-split the dataset into Y for outcome variable and X for the predictor variables as shown below:

```
y_ = data['label']  
x_ = data.drop(columns=['label'])
```

From the SKlearn library various modelling algorithms were called, which included Logistic Regression (LR), Decision Tree Classifier (CART), KNeighbors Classifier (KNN), Linear Discriminant Analysis (LDA), Gaussian Naïve Bayes (NB), Random Forest Classifier (RF), extreme Gradient-Boost Classifier (XG-Bost) and Support Vector Machine (SVM). These models were prepared and appended to list of result output, giving the output score for each model accuracy as shown below:

```
('LR',          0.6655323320615922, 0.022924093923434648)  
(LDA',         0.6956482895509095, 0.009397749131901653)  
(KNN',         0.8588820403755124, 0.01181342611770981)  
(CART',        0.9017091439646737, 0.005322954436443641)  
(NB',          0.611635491123665, 0.013094741151642107)  
(SVM',         0.5267342073759893, 0.026598382411637332)  
(RF',          0.9884425707121837, 0.0030008891703574058)  
(XG-Boost',    0.9885222954808809, 0.0035661880444705983)
```

The models were subjected to the same dataset 10 times and the results for models accuracy variations were plotted on a graph using the box plot comparison algorithm, to compare the performance and accuracy score of each model. The variation or steadiness of accuracy of each model when subjected to predictions of the same data over ten iterations are as shown in Fig. 13



**Figure 13: Box plot comparison of models' performance accuracy**

Finally, two best models were chosen to be assessed: - Random Forest Classifier and Extreme Gradient Boost were further tested with other metrics for accuracy and performance using the classification report. The aim was to select the best model to be implemented in the DTG Associate hyperglycaemia prediction tool. Then researcher started on assessing the feature importance of the best model in order to have the maximum number of features that should be included in the prediction tool to avoid model redundancy where the model has predictor variables which don't contribute to the outcome (prediction). Further assessment for Validation score, threshold discriminant analysis and model calibrations were also investigated. This helped the researchers to have only features that affect the accuracy and performance of the model, to be included on the DTG associated hyperglycaemia prediction tool development. The results of these evaluations and performances will be discussed in detail in chapter four.

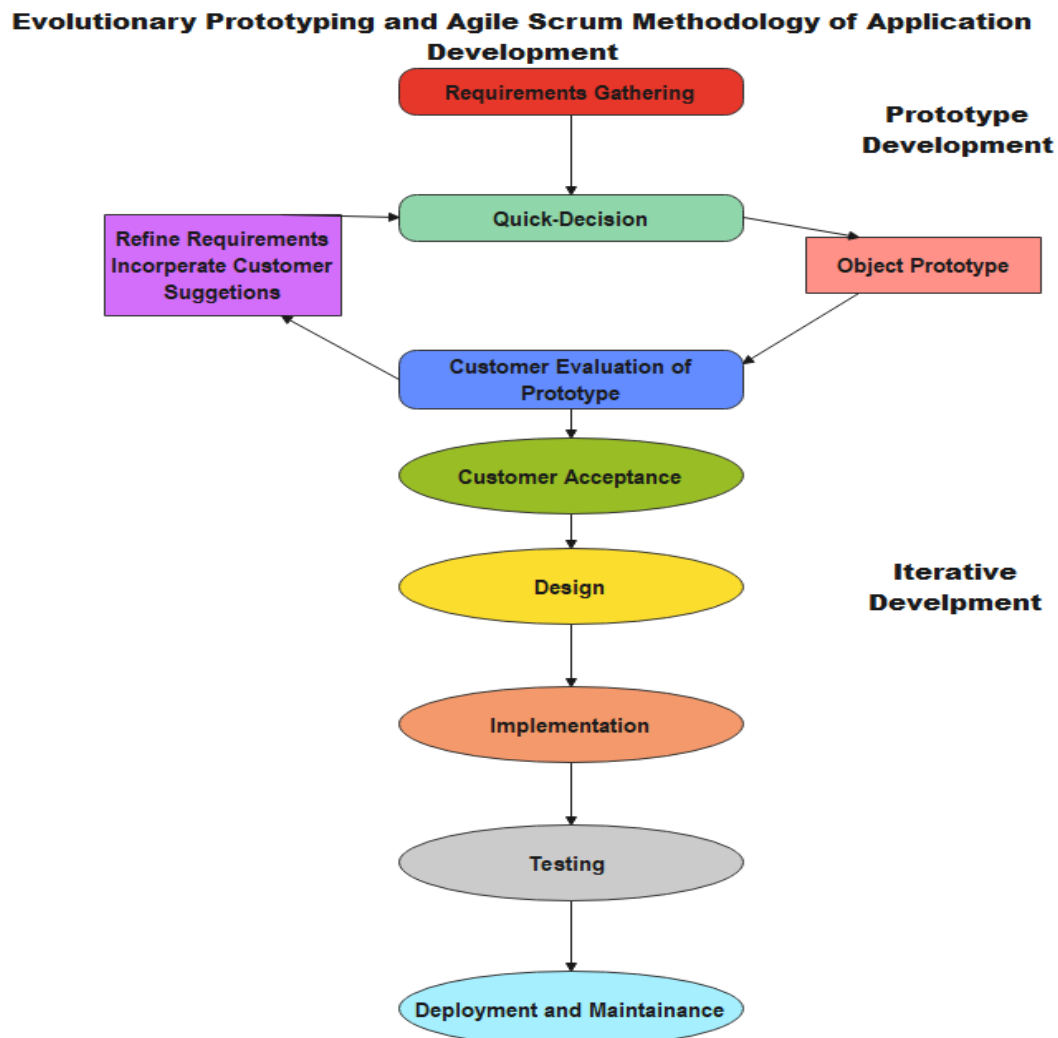
### 3.10 Prediction Tool Development Life Cycle

The prediction tool is developed using various stages of software development life cycle as illustrated in sections and the chosen software development methodology exhibited in Fig. 14.

#### 3.10.1 Software Development

The DTG Associated hyperglycaemia prediction tool was developed using Evolutionary prototyping and agile scrum software development methodologies. Evolutionary methodology was used most specially to help with the requirements specification process because the application tool was new, and its requirements were not properly spelt out neither by the developers nor the users

of the system. That rigorous user interaction with the system during development made application testing and validation inherent within the application development process right from requirements specifications through all phases up to application development and deployment. It also eased user training during tool testing and deployment, because the user would have already interacted with system during systems development, so they would be familiar with the application user interface. Agile scrum methodology also is used because of its flexibility through iteration of software development thus yielding improved quality, customer satisfaction, employee satisfaction, and organizational synergy during application development of which they were key aspects for the success of the development of a prediction tool.



**Figure 14: Development Life Cycle of the DTG Associated Hyperglycaemia Prediction Tool**

**(i) Requirements Analysis**

The user requirements for the DTG associated hyperglycaemia prediction tool were collected, analyzed, and categorized into two that is; functional and non-functional requirements.

## (ii) Functional Requirements

According to Gabriela (2017), functional requirements describe what a system should do. These requirements include: The acceptable inputs by the system, outputs of the system, Information stored by the systems or shared with other systems, computational performances done by the system, the timing and synchronization of the above. The DTG associated Hyperglycaemia prediction tool functional requirements are in the Table 1.

**Table 1: Functional Requirements**

<b>Requirement</b>	<b>Description</b>
Registration (Clinicians or Doctor)	The application shall perform registration and login for access of doctors or clinicians who will use it to predict for likelihood of development hyperglycaemia if a patient is switched to DTG
Display of computational results	The interface shall be used to display the results of the prediction computations
Manage user Accounts	The administrator shall Add, delete and manage users.

## (iii) Non-Functional Requirements

Non-functional requirements are restraints which need to be followed (design and implementation) while developing a system (Shahid & Tasneem, 2017). They define the performance of the DTG associated hyperglycaemia application tool as stated in the Table 2.

**Table 2: Non-Functional Requirements**

<b>Requirement</b>	<b>Description</b>
Performance	The application shall process the DTG associated hyperglycaemia prediction and show results in the shortest time possible, and processing of the login credentials shall be done very fast.
Security	The application show allows user-access by username and encrypted password authentication.
Usability	The Application shall be easily accessed and user friendly to authentic users.
Robustness	The application will be in position to recover from failure due to connection challenges.
Availability	It will be available as and when needed.
Language	The available interaction language is English.

#### (iv) Architectural Design

The DTG associated hyperglycaemia prediction tool was designed to be used by clinicians/ doctors/ nurses, or any other health work involved in the screening of ART experienced patients to be switched to DTG, and an administrator to perform application administration tasks. The health worker will acquire feature inputs from the HIV patient (Age, Gender, Body mass Index, Baseline Regimen, Duration on Previous Regimen, Current Regimen, Pregnancy, Viral-load, CD4 Count, Systolic pressure, Diastolic pressure) and feed them into the DTG Associated hyperglycaemia prediction tool. The application will use the built-in model to learn feature patterns of hyperglycaemia negative and positive HIV patients, and use them to make future predictions about the would be hyperglycaemia status of the HIV patient whose inputs are fed in the application. The DTG associated hyperglycaemia prediction tool will use the developed Extreme gradient boost model to make the predictions. The application model will be doing all that by a method called classification, where the HIV patient will belong to either of the two classes of hyperglycemia (positive “1”, negative “0”). This application tool will even become more accurate in its predictions as it will continue learning from the new other predictions that it will be making. The administrator of the DTG associated hyperglycaemia prediction application tool will be able to add, remove and manage application users. The outcome/output of the application will be displayed to the user on a new results page. The AT this page the user will click on the predict page button to go to the next predict page. The prediction model and the application will be uploaded in centralized place. The application will be able to support multiple users at the same time.

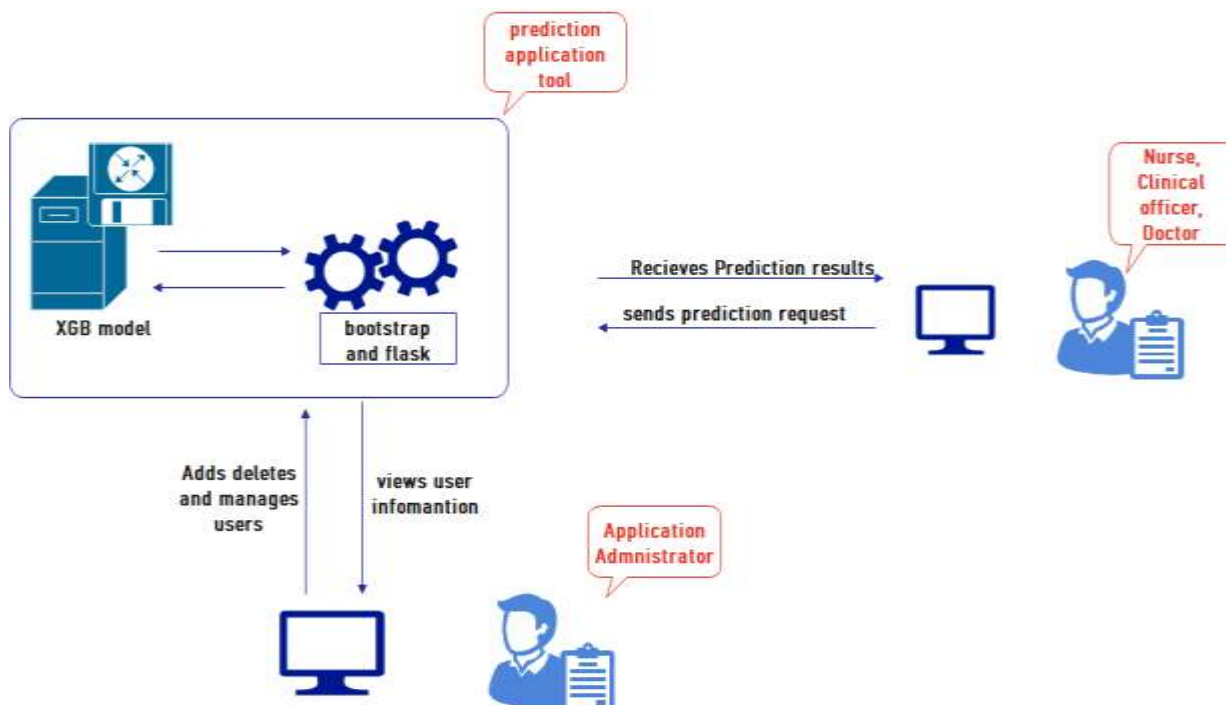
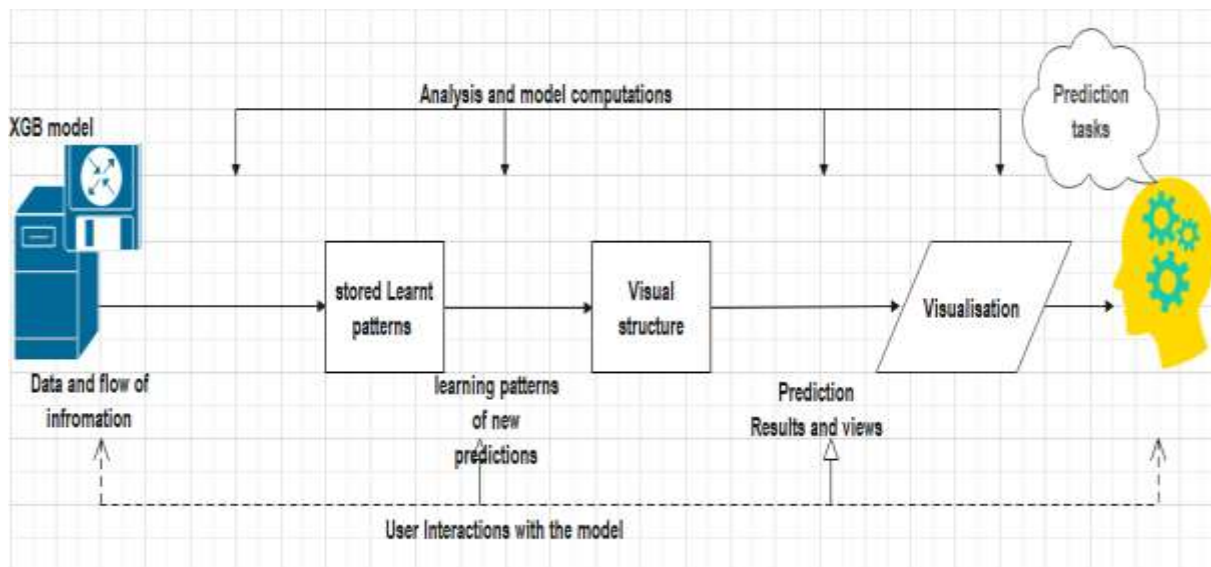


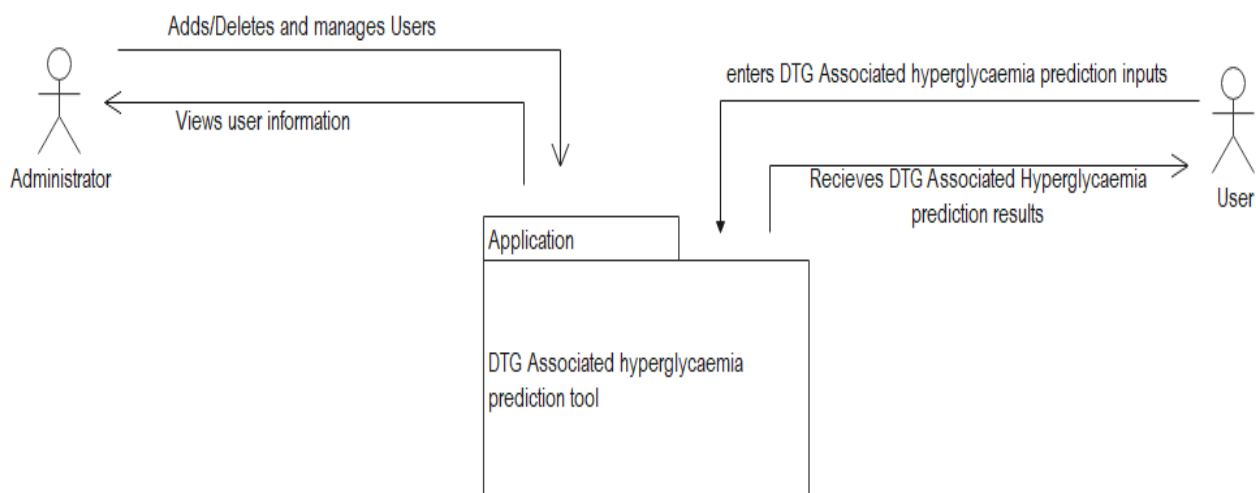
Figure 15: Conceptual Frame Work for the Application



**Figure 16: Conceptual Framework for Visualization of learning patterns of the XG-Boost model**

**(v) Context Diagram**

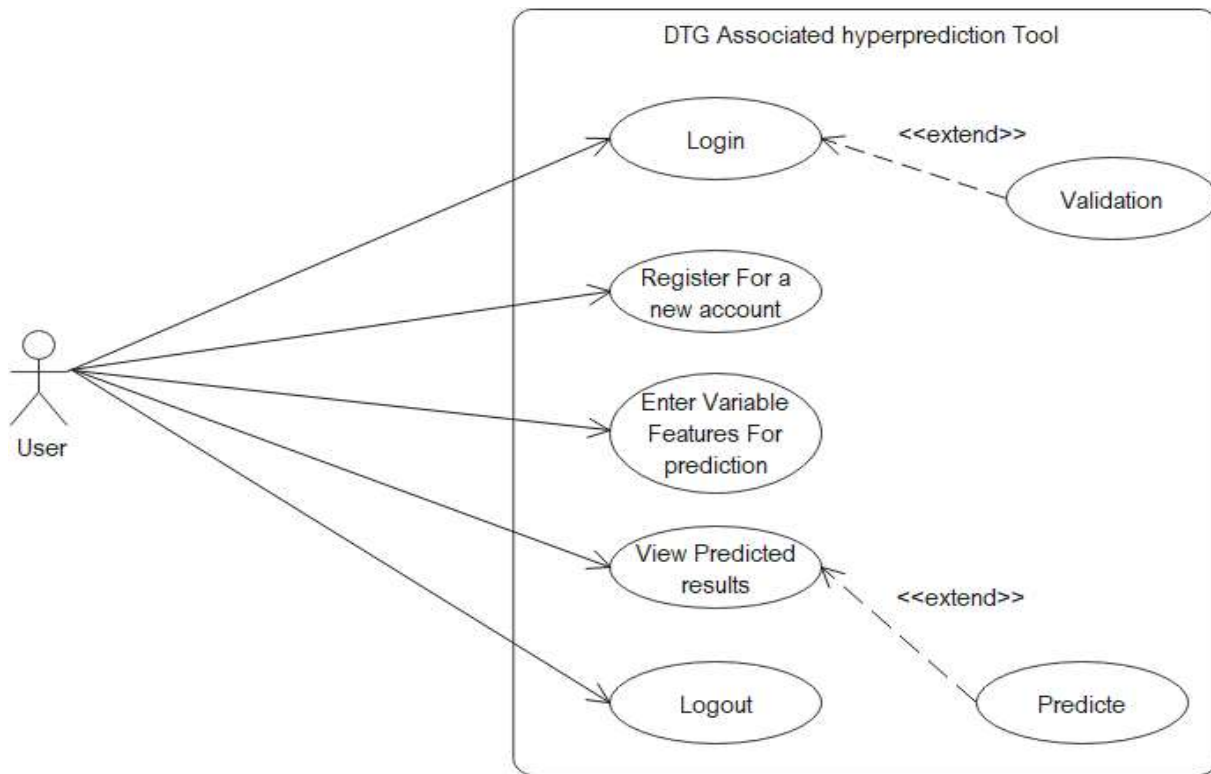
Contextual diagram in Fig. 16 defines the relationship between the system and users or other related systems; exhibiting system process interactions with existing entities. It generalizes the functionalities of the application as a whole in relation to external entities.



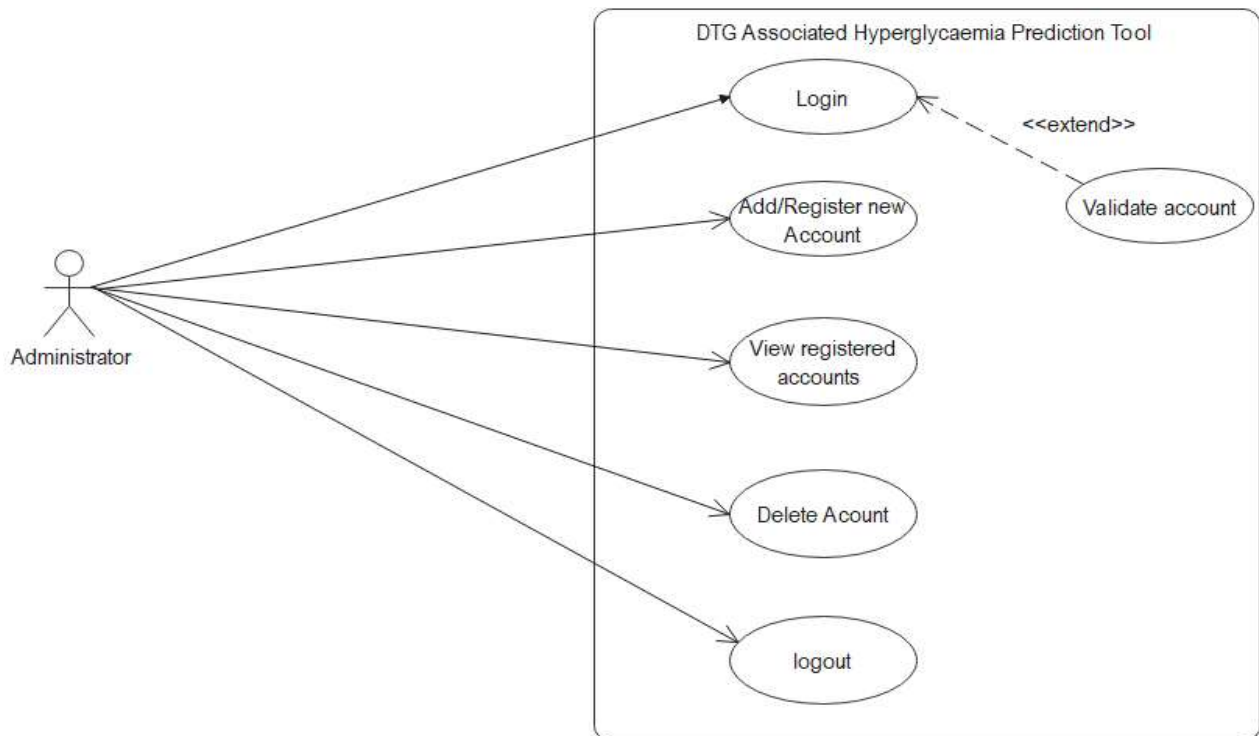
**Figure 17: Application User Contextual Diagram**

**(vi) Use Case Diagram**

Use Case Diagram in Fig. 17 and 18 shows the user and system interactions. It's a detailed description of user processes or actions within the system. Use cases are actions within a Use Case Diagram.



**Figure 18: Use Case Diagram for the prediction Showing User**



**Figure 19: Use Case Diagram Showing Administrator**

Actors (External Entities and user cases) of the tool are health workers (doctors, nurses or clinicians) and System Administrator. The Use cases are: Login, register for a new account, give inputs for the predictions, submit and receive predictions. The administrator will register users, delete accounts, view user information, maintain and manage the application.



### 3.11 Development and Implementation of the Application Tool

The DTG associated hyperglycaemia prediction tool was developed using python, Extensible Markup Language (XML) and Hyper Text Markup Language (HTML). These development languages were used in the Atom development tool.

#### 3.11.1 System Development Languages and Technologies

These are scripting and software; with client-side languages executed by the browsers at the clients' computer. But the process of executing the source code happens at the application host computer. This is because the inputs are captured and transferred from the end user computer to the application computer where the prediction model is hosted along with the application.

##### (i) Extensible Markup Language

The Extensible Markup Language (XML) is basically for formatting, displaying data, transporting, data sharing and data availability. It is a composition of inherent and user defined features (tags) like the header tags (such as h1, h2, h3...), paragraph tags (<p> My paragraph </p>), image tags (<img src="" alt="" > </img>), anchor tags (a href="" target=""> Home </a>) and <footnote> for footnotes respectively like shown below using CDATA tags.

```
<xml>
  <title>Your HTML title</title>
  <htmlData><![CDATA[<html>
    <head>
      <script/>
    </head>
    <body>
      Your HTML's body
    </body>
  </html>
  ]>
</htmlData>
<footnote></footnote>
</xml>
```

The CDATA, means Character Data; it's used for distinguishing related purpose within XML and Standard Generalized Markup Language (SGML). The XML was used in front end implementation, particularly in designing of front-end interfaces that facilitate user interactions with the application. It offers a simple structuring of application pages, that are greatly improved by other technologies like Cascading Style Sheets (CSS) and Python scripts which were also used basically for interface connections and design. The user friendliness, flexibility and easier interaction with other languages was for the choice of XML.

## (ii) Cascading Style Sheet

Cascading Style Sheet (CSS) is a cornerstone language and technology used to describe the presentation of information written in markup languages. The XHTML provides basic tools necessary for information structuring in an application. But CSS, gives the ability to present information in a consistent similar format each time it's viewed. This presentation includes styling, background images, colors, headings, sections, headers, footers and many others, which were the basic building blocks of application tool interface design.

## (iii) Python Scripting Language

Python scripting languages was used for developing the XGB-classifier model and the backend programming of the application interface. Python was the application development language for interface connection and Session management. It was also used to capture and transferring inputs for predictions to the prediction model. As well user authentication and validation at login into the system are in python language. The development environment used for creating the back-end software is called flask. Flask is a micro web framework written in Python. It is classified as a micro-framework because does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. It validates the username and password login credentials using the MySQL database. Below are some screen shots of the developed login (Fig. 20) and contact us (Fig. 21) application interfaces.

---

Please Fill Your Information Here Below

User Name

Password

Login

**Figure 20: Application login Interface**



Contact Us:

---

Name

Email

Subject

Message

---

**Figure 21: Application contact us Interface**

### 3.12 Other Requirements

- (i) Web browser; Google Chrome, Mozilla Firefox, Torch and Internet Explorer.
- (ii) Operating system; Windows 7,8,10, Mac OS and Unix-based operating (suse-linux, Mandrake-linux, Mint-linux, Ubuun).
- (iii) RAM 4 GB
- (iv) 256 GB Harddisk
- (v) i3 core

### 3.13 Assumptions and Dependencies

- (i) It is assumed that Users (Nurses, Doctors, clinicians have Access to computers when screening patients switching) and Administrators have access to a smartphone, personal computers.

- (ii) Inputs used for predictions are readily available to users before switching HIV patients to DTG.
- (iii) The function of the application tool depends on the operational performance of XGB-classifier model and flask Python scripting language.
- (iv) Availability of the system depends on installation of the following operational requirements (Flask, numpy, Flask-Admin, Flask-Bootstrap, Flask-WTF, WTFForms, pandas, flask-sqlalchemy, werkzeug, Flask-Login, gunicorn).

### **3.14 Summary**

The methodology chapter presented the data sources and techniques employed in the carrying out this study. It described the persons of interest, data collection, analysis, and data modeling algorithms used. It also explicitly stated the functional and non-functional requirements required for the development and deployment of the application tool which is critical in prediction of the likelihood of development of hyperglycaemia if a treatment experienced HIV patient is switched to DTG. Moreover, it defined the architectural design of the application implemented which illuminated clearly the application development discussing technologies and reasons as to why they were preferred. Thus, contributing factor to maintaining the quality of health conditions of patients and a long life.

## CHAPTER FOUR

### RESULTS AND DISCUSSION

#### 4.1 Introduction

Having described the materials and methods used in chapter three by this study, the researcher now unveils results with deliberation, explanations and descriptions for the collected, processed, analysed, and modeled data for the research and development of the DTG associate hyperglycaemia prediction tool.

#### 4.2 Demographics of the Collected Data

From the total of 166 066 collected data; 62 358 (38%) were men and 103 361 (62%) were women. When data processing was executed 9077 records is what the researchers remained with, which is (6%) of total collected data. This was because of the occurrences systematic missingness of data within various variables of the dataset which may cause biasness within the model. Further processing revealed many null values especially with in variables of; systolic and diastolic pressure, CD4 Count, Body Mass Index, Viral load, first line regimen, and Baseline regimen as shown by the visualization in Fig. 22.

(i)	Age	0
(ii)	Gender	0
(iii)	BaselineRegimen	2
(iv)	ViralLoad	42
(v)	TBStatus	0
(vi)	Pregnant	0
(vii)	CD4Count	4537
(viii)	CurrentRegimen	0
(ix)	FirstRegimenDuration	5
(x)	bminorm	100
(xi)	SystolicPressure	7270
(xii)	DiastolicPressure	7270

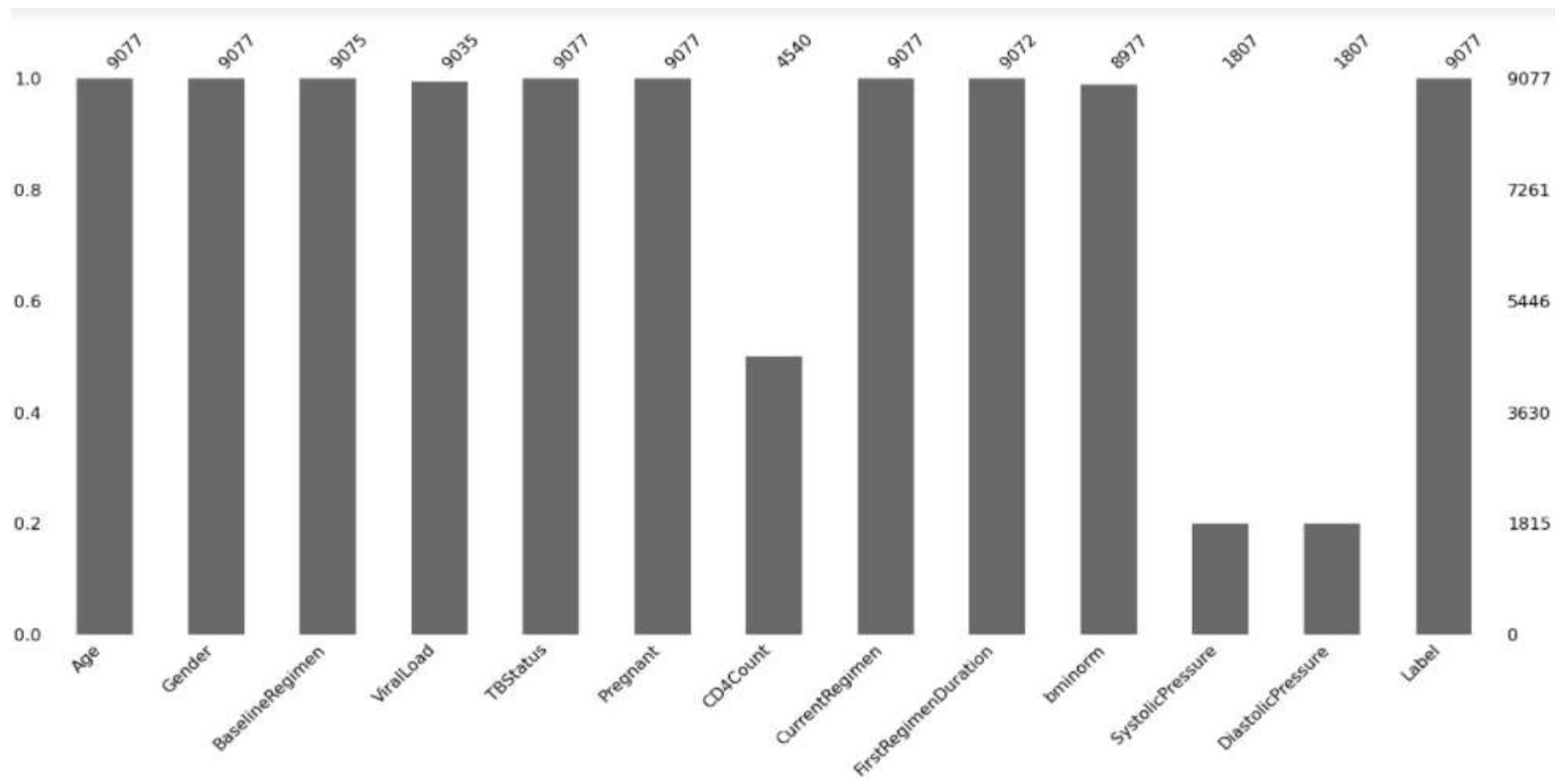
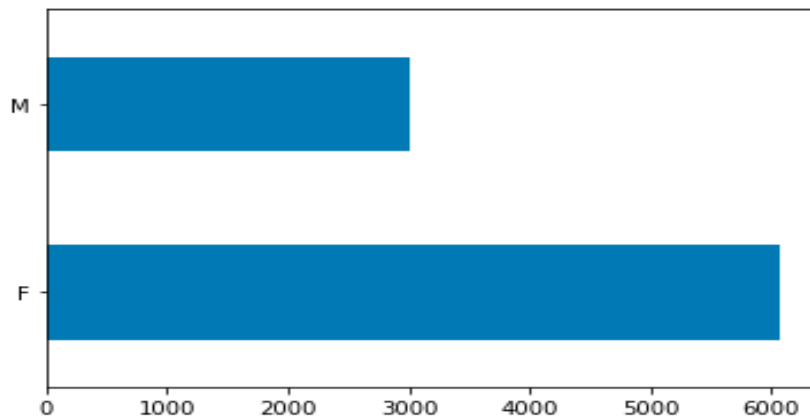


Figure 22: Shows data completeness

## 4.2.1 Gender Composition

```
F    6062
M    3015
Name: Gender, dtype: int64

<matplotlib.axes._subplots.AxesSubplot at 0x1c7
```

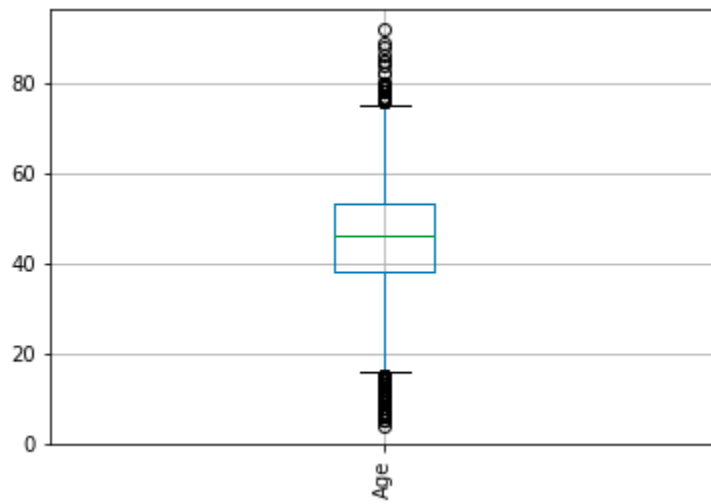


**Figure 23: Gender distribution**

Using the filtered and processed data a big imbalance of gender distribution revealed 6062 (2/3) were women and 3015 (1/3) were men, which was a true representation of the population that is seeking treatment and support for HIV. It has been stated over the years that women seek HIV testing, counselling and treatment more than men, hence their reflection even in our data distribution. This implies that there is more data for women for the model to learn from than men. As illustrated in the visualization in the Fig. 23.

## 4.2.2 Age Distribution

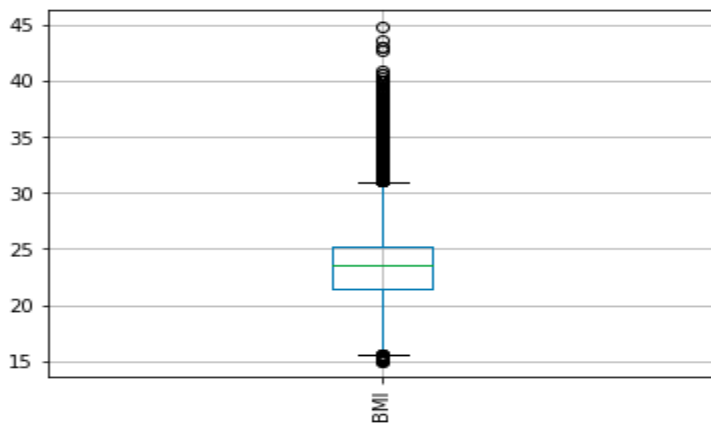
The researchers also made age analysis using a box and whisker plot; the outcome in Fig. 24 shows that  $\frac{1}{4}$  of the HIV patients are 0 – 39 years of age,  $\frac{1}{2}$  of the HIV patients are less or equal to 48 years,  $\frac{3}{4}$  of the HIV patients are  $\leq 57$  years of age, and  $\geq 58$  years are in the last quarter of the age distribution. This implied that the model will synthetically learn about the participants age, because the median is in the middle of the box and the Whiskers are equal distance from ends of the box. The variations in the ages of participants follows a normal distribution from 18 – 77 years of Age, as shown above.



**Figure 24: Patients Age Distribution**

#### 4.2.3 Body Mass Index of the Population

The BMI box and whisker plot in Fig. 25 also shows the four quarters into which the participants are distributed; with the first quarter of participants' BMI ranging from 14 – 20, second quarter of participants' BMI ranges from 21 – 23, the third quarter of participants' BMI ranges from 24 – 25, while the last quarter of participants' BMI ranges from 25 - 31. Analysis from the box and whisker revealed that the distribution of the BMI is skewed to the right, because the median is way above the middle of the box though the whiskers are at equal distance from the box. This indicates a more variation of BMI in the second quarter which means that the model will be able to synthetically learn from the first second and fourth quarters of this variable than its third quarter.



**Figure 25: Body Mass Index Distribution among the HIV Patients**

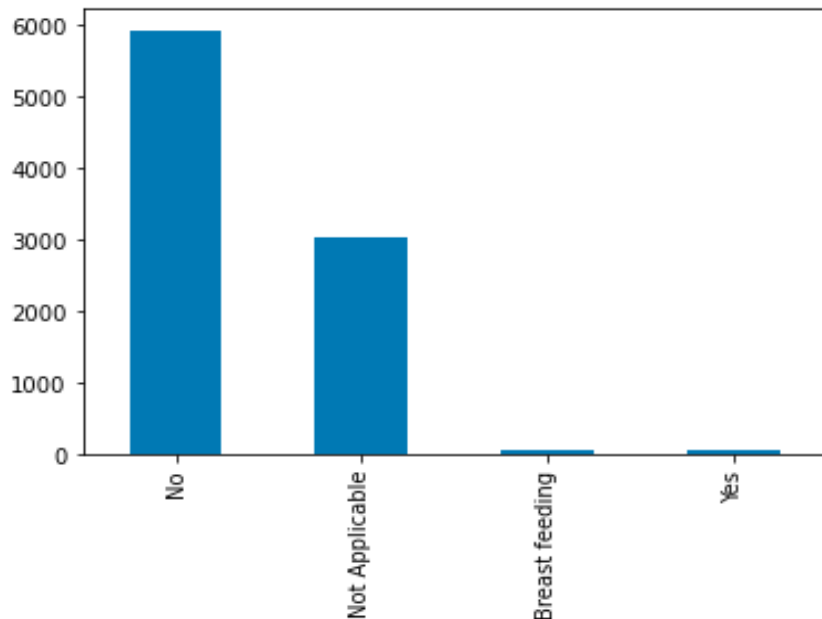
#### 4.2.4 Number of Pregnant Women

The data of the pregnancy variable also revealed that from the number of 6062 women; 5938 (93%) are not pregnant, 3015 were not applicable (implying men), only 68 (4%) of women were pregnant, and 56 (3%) are nursing mothers. The data of women who were not pregnant is present more



learning patterns to the model than the men, nursing mothers and pregnant women all combine. The illustration of this variable distribution is as shown in Fig. 26.

```
No          5938
Not Applicable 3015|
Breast feeding  56
Yes           68
Name: Pregnant, dtype: int64
AxesSubplot(0.125,0.125;0.775x0.755)
```



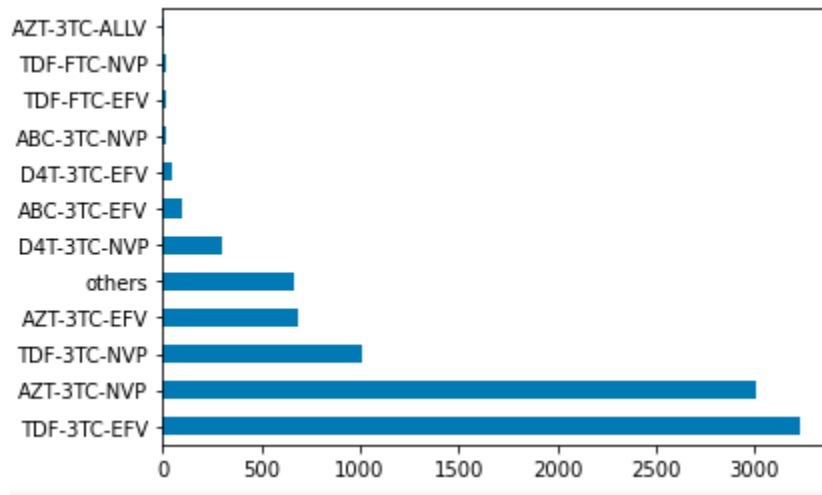
**Figure 26: Number of Pregnant HIV Patients**

### 4.3 Human Immunodeficiency Virus Therapy analysis

Here the researcher will assess how the model will learn from the treatment information in use for the prediction.

#### 4.3.1 Numbers Previous Regimen Initiation

The previous regimen data revealed 12 different types which were taken by the participants before switching to a DTG based regimen. From the close analysis it was found that; 3230 had taken TDF-3TC-EFV, 3001 had used AZT-3TC-NVP, 1008 had used TDF-3TC-NVP, 682 had taken AZT-3TC-EFV, 666 had taken others, 297 had taken D4T-3TC-NVP, 100 had taken ABC-3TC-EFV, 44 had taken D4T-3TC-EFV, 17 had taken ABC-3TC-NVP, 14 had taken TDF-FTC-EFV, 14 had taken TDF-FTC-NVP, and 2 had taken AZT-3TC-ALLV. It was noted and observed that two previous regimens TDF-3TC-EFV and AZT-3TC-NVP were taken by 3/4 of the participants in this research. These two regimens will present the model with more learning patterns than all the other remaining 10 regimens combined, while creating a model which is data hungry for learning patterns of the regimens as shown in Fig. 27.

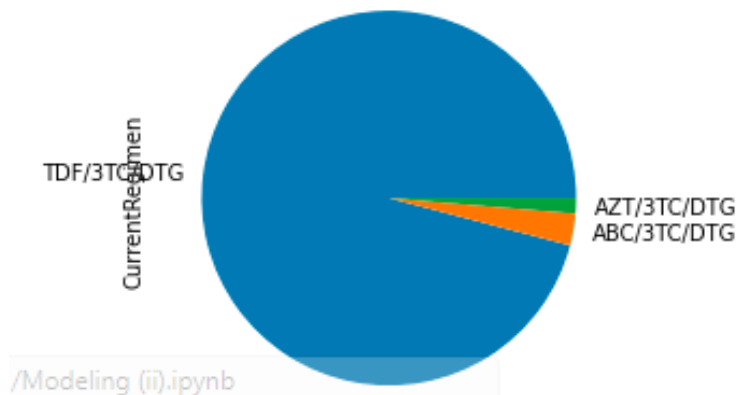


**Figure 27: First (Baseline) Regimen Treatment Analysis of Patients**

### 4.3.2 Number of Patients Categorized by Dolutegravir Regimen

Similarly, the researcher assessed the distribution and categorization of participants by the type of DTG Regimen they took and it was found out that: 8708 HIV participants were initiated on TDF/3TC/DTG regimen representing more than three quarters of the total number of participants in this study. The 250 participants were switched to ABC/3TC/DTG and lastly 119 switched to AZT/3TC/DTG. This analysis revealed that most participants were switched to TDF/3TC/DTG regimen which implies that this regimen will present the model with more learning patterns than the other two DTG regimens as revealed in the Fig. 28.

```
TDF/3TC/DTG      8708
ABC/3TC/DTG      250
AZT/3TC/DTG      119
Name: CurrentRegimen, dtype: int64
AxesSubplot(0.260833,0.125;0.503333x0.755)
```



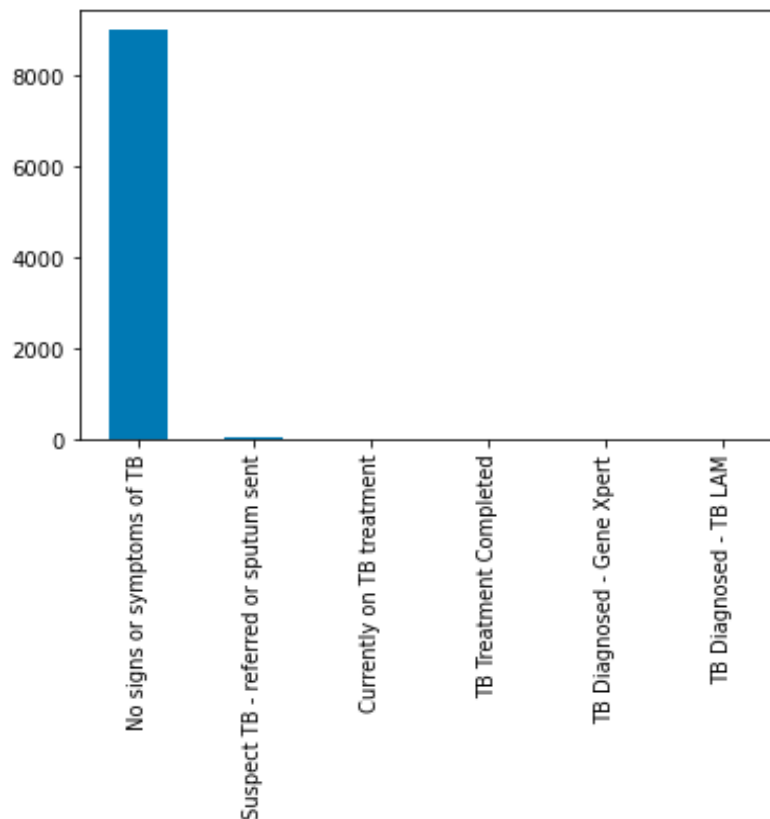
**Figure 28: Current Regimen Treatment of Patients**

## 4.4 Infections Analysis

In this part patients' health data was analysed which include blood pressure, Viral load, CD4 Count, TB- Status, and this help us to under to understand whether the data was synthetically distributed or not. This would in a long to determine the model's performance.

### 4.4.1 Tuberculosis-Status

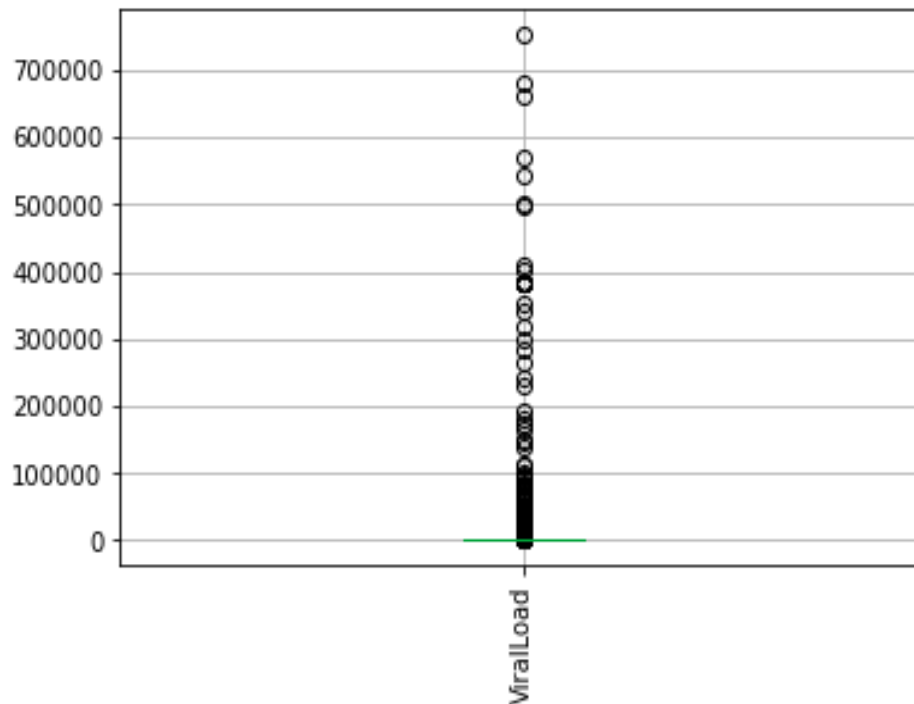
For TB infection, the data showed that; 9006 HIV patients had never been infected with TB representing 99% of the total number of participants in the study being TB negative, 41 (0.5%) of participants were suspected to have TB, 14 (0.25%) of participants are currently on TB treatment, 13 (0.2%) of participants had completed the TB treatment, 2 (0.03%) of participants were diagnosed with multi-drug resistance TB, and 1 (0.02%) was tested with positive TB. The analysis gives the total number of 17 participants positive with TB, participants cured of TB as 13. The model will have more learning patterns of TB negative participants while almost learn nothing from cured and positive participants. The Fig. 29 presents the TB-Status visualization.



**Figure 29:** Number of TB-Infected HIV Patients

#### 4.4.2 Box and Whisker Plot of Viral Load

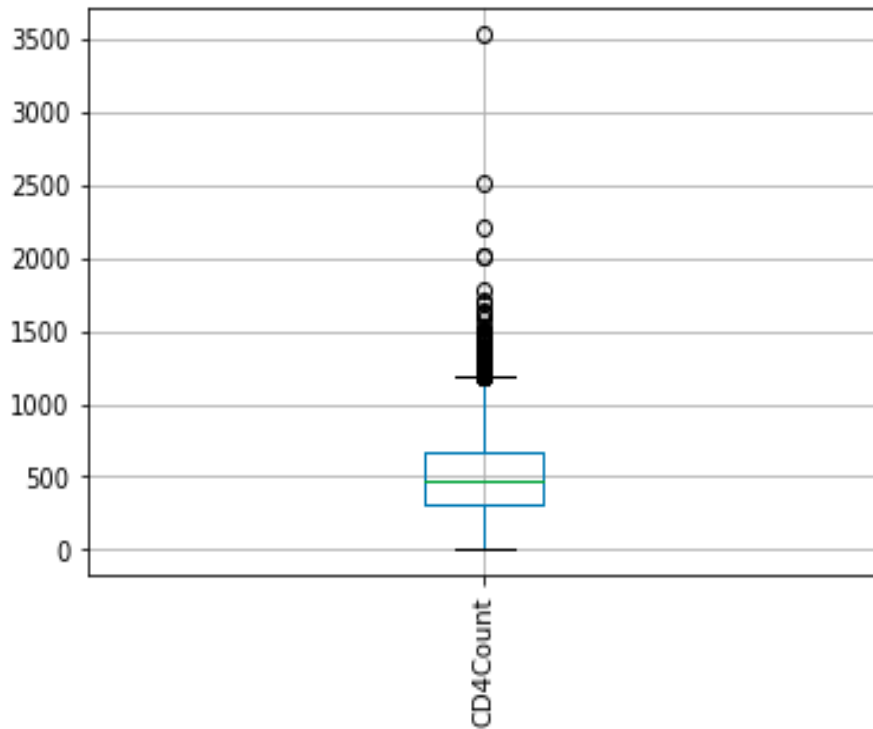
The participants viral load box and whisker plot shows that most of the viral-load is suppressed with outliers in the upper whisker. This implies that, the model will learn more about patterns with suppressed viral-load than those with higher viral-load demonstrated in the Fig. 30.



**Figure 30: The HIV patients ViralLoad Distribution**

#### 4.4.3 Box and Whisker Plot for the CD4 Count

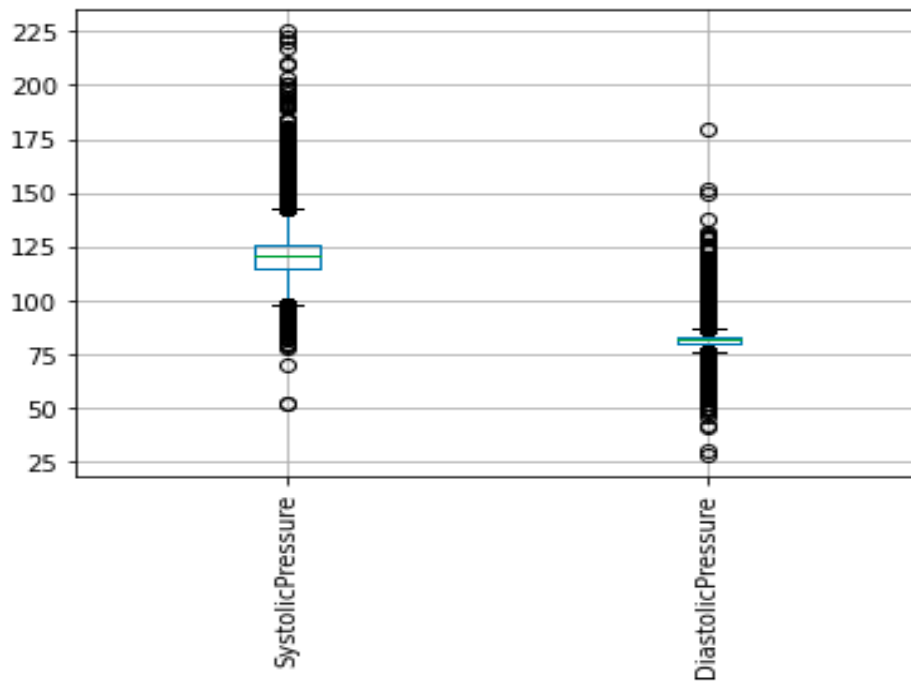
The CD4 box and whisker above reveals  $\frac{1}{2}$  of the participants under study have minimum (450) CD4 and below, while the remaining half lay within normal CD4 count range. Moreover, from the analysis of this graph, we observed that the data is skewed to the right (positively skewed). Further, it reveals variations of CD4 count in normal range (450 – 1200) with in the second half. Indicating positive skewness of the CD4-variable which implying that the model will synthetically learn from participants with minimum CD4 below than those within range of normal CD4 as illustrated in Fig. 31.



**Figure 31: The HIV Patients CD4 count Distribution**

#### **4.4.4 Systolic Pressure and Diastolic Pressure**

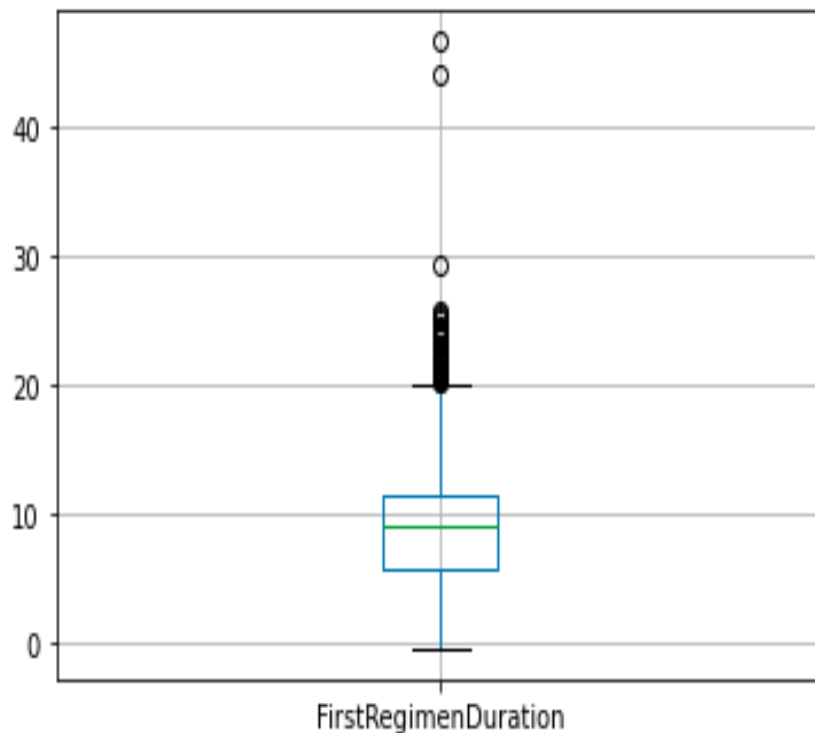
The systolic pressure of the HIV patients is normally distributed from 96 - 148 and skewed to the left (negatively skewed). This means that the model will synthetically learn more about participants with systolic pressure in the third and fourth quartile than those with systolic pressure in the first and second quartile. The study shows lower normally distributed varying diastolic pressure as compared to systolic pressure which is the normal expected phenomena. For the diastolic pressure normal distribution varies from 75 – 87 and skewed to the left (negatively skewed). Implying that the model will learn synthetically more from participants with diastolic pressure in the fourth and third quarters than those with diastolic pressure within the first and second quarters. This data in Fig. 32 showed that model will learn more from HIV patients with normal blood pressure than any other.



**Figure 32: The Distribution of Systolic Pressure and Diastolic Pressure of HIV Patients**

#### 4.4.5 First Regimen Duration

The box and whisker plot in Fig. 33 for the first regimen duration is positively skewed. But with a close analysis reveals that the model synthetically will learn more from participants with first regimen duration in the third quarter than participants with first regimen duration in any of the other three quarters.



**Figure 33: Period each HIV patient took on the First Regimen**

## **4.5 Model Development Process, Results and Prospects for the Developed Systems**

The aim here is to understand and account for research results. This will entail the model development process, what were the assessment metrics of the model concerning performance evaluation, accuracy and validation of the models. Binary classification algorithms were chosen for the model development process because our prediction involved identification of two classes of being hyperglycemic or not hyperglycemic (1 or 0). The eight chosen binary classifiers included:

- (i) Ensemble classifiers (XG-boost, Random Forest) these improve machine learning results by combining several models.
- (ii) Decision tree classifier (Decision Tree Classifier) create the classification model by building a decision tree creates the classification model by building a decision tree.
- (iii) Gradient boost classifiers (KN-Neighbors Classifier, Linear Discriminant Analysis, Logistic Regression, SVC) combine many weak learning models together to create a strong predictive model.
- (iv) Gaussian Naive Bayes supports continuous valued features and models each as conforming to a Gaussian (normal) distribution.

The above algorithms a representation of the binary classifiers according to classification method and were used to assess which algorithm and method produces the most efficient model for the prediction.

## **4.6 Model Evaluations and Validation**

During this phase the researchers validated and evaluated all the models for accuracy to have the model selected for application prediction development.

### **4.6.1 Performance Evaluations**

When the researcher ran model creation function of the eight algorithms for different models (XG-Boost Classifier, Random Forest Classifier, SVC, Gaussian-NB, Decision Tree Classifier, KN-Neighbors Classifier, Linear Discriminant Analysis, Logistic Regression), it produced results which indicated that Random Forest Classifier and XG-Boost Classifiers had the highest degree of accuracy as show Fig. 34.

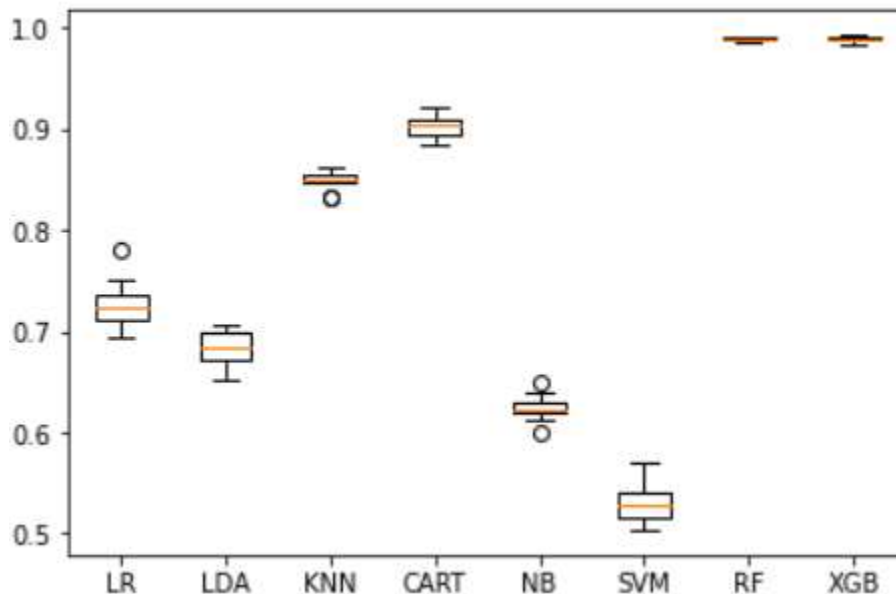
```

('LR', 0.6770543277822008, 0.03305775709051789)
('LDA', 0.6949211881611664, 0.016850844040097842)
('KNN', 0.8553654074523841, 0.015775757690777905)
('CART', 0.8940641369278955, 0.007089790093361343)
('NB', 0.6122583150275729, 0.0210735981744737)
('SVM', 0.5290422857790226, 0.02095709708902302)
('RF', 0.9881160399056702, 0.0021961512865160703)
('XGB', 0.9891230537507869, 0.002245744820137256)

```

**Figure 34: Creation function of the eight algorithms for different models (XG-Boost Classifier, Random Forest Classifier, SVC, Gaussian-NB, Decision Tree Classifier, KN-Neighbors Classifier, Linear Discriminant Analysis, Logistic Regression)**

But this was not enough assessment for accuracy, so the researcher continued to compare the models' performance accuracy using the boxplot algorithm comparison. This helps to analyze the model accuracy more in terms of skewness and Variations in performance accuracy due to the number of predictions runs each model is subjected to with the same predictor variables' values. The model comparison algorithms plot in Fig. 35.



**Figure 35: Comparison accuracy of the Seven Models**

The visualized diagram of Fig. 35 shows ten iterations of data prediction by each machine learning algorithm. Skewedness was observed in the learning patterns each algorithm and very large variations in performance accuracy for (Support Vector Machine, Gaussian-NB, Decision Tree Classifier, K-Neighbors Classifier, Linear Discriminant Analysis and Logistic Regression) with the exception of (Random Forest Classifier and XG-Boost classifier). This implied that those six models' performance accuracy can vary so high to the number of iterations a model can be subjected to the same variable data-inputs except for the two algorithms (XG-Boost classifier and Random Forest Classifier). Yet in the real-world repetition of the same variable inputs is a real phenomenon. Therefore, Random Forest Classifier and XG-Boost classifier were chosen as the



models with the best performance accurate because they have no skewness in their learning pattern and their accuracy is not affected by iterations of same variable inputs as shown in Fig. 35.

#### 4.6.2 Accuracy Evaluations

During this Part, the researcher is looking at the models' accuracy of predicting correctly each class. To be precise: The classification of participants with hyperglycaemia as truly positive, and the classification participants without hyperglycemia as true negatives as hyperglycaemia. In this phase a confusion matrix method was employed to evaluate each of the two models with the following metrics:

##### (i) Precision

Is the measure of ability of a model to correctly identify positive cases as true positive (positive predictive value) and Negative cases as truly negative (Negative predictive value) from all the predicted positive cases. Below the formula

$$\text{Negative predictive value} = \frac{\text{True negative}}{\text{True negative} + \text{False negative}}$$

$$\text{positive predictive value} = \frac{\text{True Positive}}{\text{True Positive} + \text{False positive}}$$

##### (ii) Recall

Is the proportion of correctly identified positives from all the positively predicted cases. Or the proportion correctly identified negatives from all the negatively predicted case. This is very important most especially at a moment when falsely identifying positive cases as negative can be very costly in terms of lives.

$$\text{(Specificity)Proportion of correctly identified negatives} = \frac{\text{True negative}}{\text{True negative} + \text{False Positive}}$$

$$\text{(Sensitivity)Proportion of correctly identified positives} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

##### (iii) Accuracy

Measures of ability to correctly classify. It's mostly a useful metric when all classifications have same importance.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{False Positive} + \text{True Negative} + \text{False Negative}}$$

**(iv) F1-Score**

It is the harmonic mean of Recall and Precision which gives a finer measure of the miss-classified cases as compared to the Accuracy Metric.

$$\text{F1 - Score} = 2X \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}$$

**(v) Macro Average**

Is the method that can be used when you want to know how the application performs overall across the sets of data. Though it is not advisable to come up with any specific decision with this average.

**(vi) Weighted Average**

Accounts for class imbalance by computing the average of binary metrics in which each class's score is weighted by its presence in the true data sample.

**4.6.3 Other Classification Metrics**

(i) Null Error Rate

(ii) This is how often you would be wrong if you always predicted the majority class. This can be a useful baseline metric to compare your classifier against. However, the best classifier for a particular application will sometimes have a higher error rate than the null error rate.

(iii) Misclassification Rate (aka Error Rate): Overall, how often is it wrong?

(iv) Cohen's Kappa: It is a measure of how well the model performs in comparison to how it could perform by chance.

(v) Receiver operator Curve (ROC Curve): This is a commonly used graph that summarizes the performance of a classifier over all possible thresholds. It is generated by plotting the True Positive Rate (y-axis) against the False Positive Rate (x-axis) as you vary the threshold by assigning observations to a given class.

(vi) Precision Recall Curve: is the curve which shows the relationship between precision (= positive predictive value) and recall (= sensitivity) for every possible cut-off.

(vii) But since our research project has singled out ensemble bi-classification models (XG-Boost Classifier and Random Forest Classifier) as the best two from which to select one for the DTG associated hyperglycaemia prediction tool development: The models will be evaluated using only the following metrics: Precision, Recall, Accuracy, F1-Score, ROC curve, Precision recall curve and Cohen’s Kappa which look at the accuracy performance abilities of sensitivity, specificity, positive predictive value, Negative predictive value and ability of the model to predict not by chance, which are critical for our DTG associated hyperglycaemia prediction application.

#### 4.6.4 Model Classification

For this accuracy evaluation testing with the classification, the researcher used test data-set that was set aside during the first dataset split. The data-test split had 1362 records in total with 186 positive and 1176 negative participants. Table 36 presents the classification reports with the results of the accuracy evaluations.

	precision	recall	f1-score	support
0.0	0.98	0.93	0.96	1176
1.0	0.67	0.89	0.77	186
accuracy			0.93	1362
macro avg	0.83	0.91	0.86	1362
weighted avg	0.94	0.93	0.93	1362

**Figure 36: The XG-Boost Classification Report**

	precision	recall	f1-score	support
0.0	0.97	0.92	0.94	1184
1.0	0.60	0.83	0.70	178
accuracy			0.91	1362
macro avg	0.79	0.87	0.82	1362
weighted avg	0.92	0.91	0.91	1362

**Figure 37: Random Forest Classification Report**

##### (i) Precision Metric

The results of both classification reports under precision metrics showed the positive predictive values (PPV) for XG-Boost and Random Forest Classifier models are 0.67 and 0.60 respectively. This meant that XG-Boost classifier model has higher probability of classifying HIV patients as true hyperglycaemia positive from all the classified hyperglycaemia positive participants, than Random Forest classifier model. The classification report also revealed that Negative predictive

values for the models are 0.98 and 0.97 for XG-Boost classifier and Rand Forest Classifier Respective. Also, this meant that XG-Boost classifier has a higher probability of classifying true negative hyperglycaemia HIV patients as negatives from the hyperglycaemia negative participants than the Random Classifier.

### **(ii) Recall Metric**

On a close look at the Recall metrics, the XG-Boost classifier and Random Forest Classifier observations show their probabilities of specificity as 0.93 and 0.92 respectively. Thus XG-Boost classifier model exhibited a higher specificity than the Random Forest classifier model. This implied that the XG-Boost classifier can correctly classify true negative hyperglycaemia HIV Patients as negatives from the hyperglycaemia negative HIV patients better than the Random Forest Classifier model.

Still under recall, it was observed that the Sensitivity of the two models' probabilities are 0.89 and 0.83 for XG-Boost Classifier model and Random Forest Classifier model respectively. Hence XG-Boost classifier model had higher sensitivity than the Random Forest Classifier model. This pointed out still the fact that, the XG-Boost model is better than Random forest model at classifying true hyperglycaemia positive HIV patients as positive from the group of hyperglycaemia positive, participants switched to DTG. Both models' sensitivity metric scores exceeded the one required for a health application which is 0.8. And this is the most important metric for our research as we are more interested in rightly classifying HIV patients who would become Hyperglycaemia positive if switched to DTG.

### **(iii) F1-Score Metric**

The F1-Score of the negative class of each model is 0.96 and 0.94 for XG-Boost classifier and Random forest classifier respectively. That indicated a higher harmonic mean of Recall and Precision of the negative class for XG-Boost classifier than the Random forest classifier. Implying on Average XG-Boost classifier is better than Random Forest Classifier at correctly classifying true negative hyperglycaemia HIV patients as negative if switch to DTG.

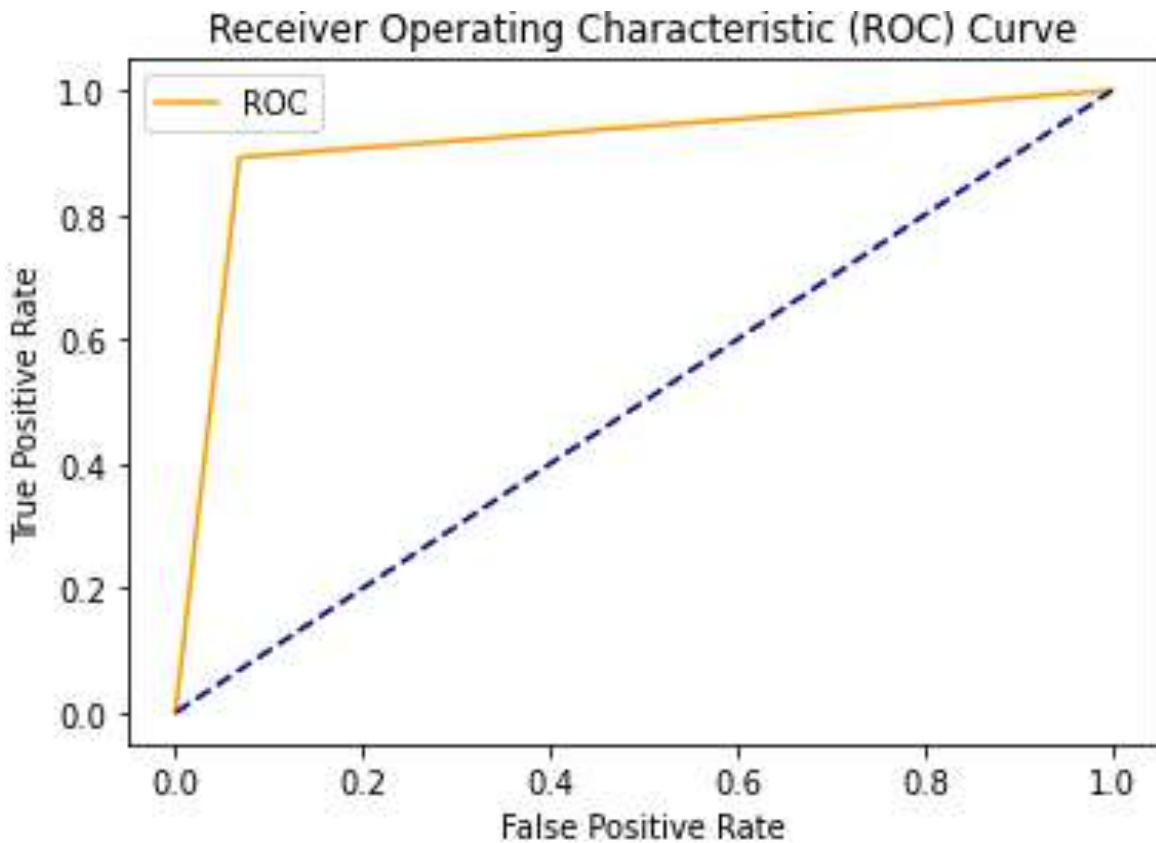
As well from the classification report the F1-Score of the positive class of each model is as follows 0.77 and 0.70 for XG-Boost classifier and Random forest classifier respectively. Hence indicating a higher Recall and Precision harmonic mean of the class for XG-Boost classifier than the Random forest classifier. That meant XG-Boost classifier is better than Random Forest Classifier at correctly classifying true positive hyperglycaemia HIV patients as positive if switch to DTG.

**(iv) Accuracy Metric**

The accuracy of each module is 0.93 and 0.91 for XG-Boost classifier and Random Forest Classifier respectively. This shows that the XG-Boost classifier model has a higher overall classification accuracy than the Random Forest Model.

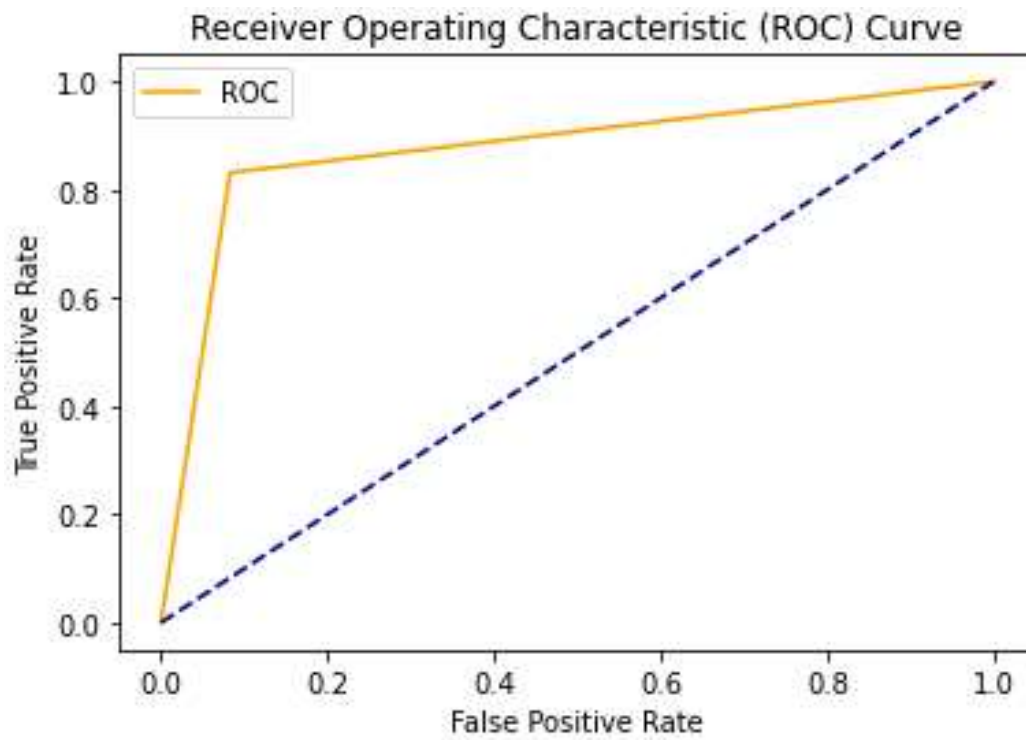
**(v) Receive Operator Characteristics Curves**

For studies designed to classify or predict the diseased from the un-diseased it's important to assess the model performance accuracy for the benefit of the patient care and health care system. The ROC curves are useful tools in the assessment of the performance of a diagnostic test over the range of possible values of a predictor variable. The area under a ROC curve gives a measure of discrimination and allows researchers to relate the performance of two or more diagnostic tests.



**Figure 38: The ROC Curve for XG-Boost**

The area under a curve for the XG-Boost classification model is 0.91 which is way above the diagonal 45-degree blue line. This indicates that the model has an excellent discriminatory ability to classify hyperglycemia positive and Negative HIV patients. The 0.91 suggests a chance for a XG-Boost classification model to appropriately distinguish a hyperglycaemia negative HIV patient from a positive one.



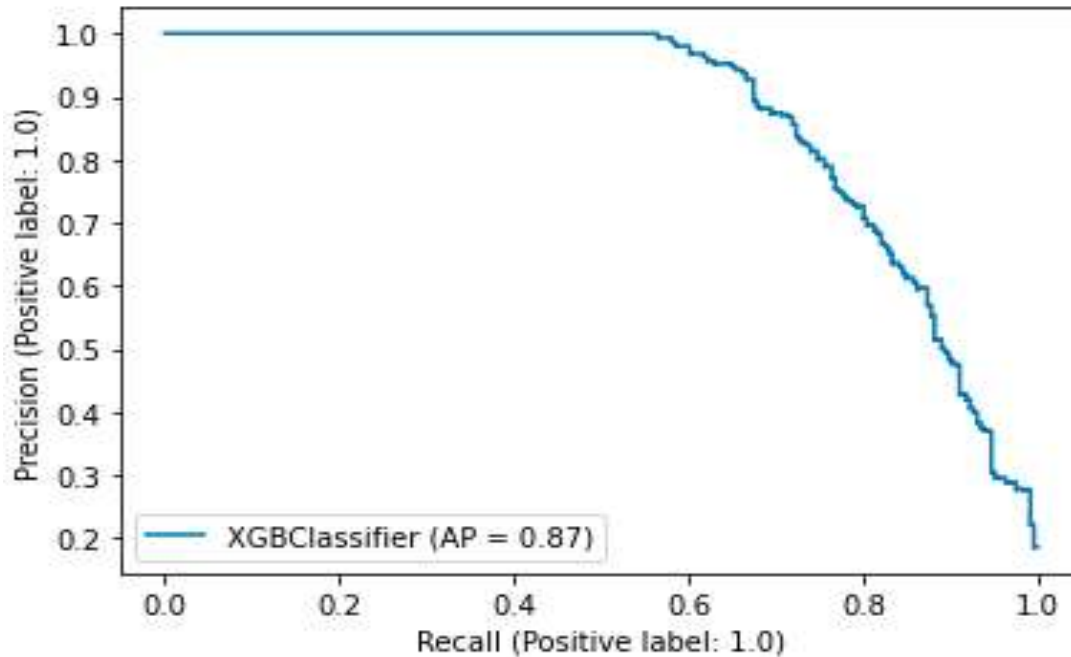
**Figure 39: The ROC Curve for Random Forest Classifier**

The random forest classifier has area under curve as 0.87. This is also above the 45-degree diagonal blue line implying the Random forest has classification ability to appropriately distinguish the positive hyperglycaemia positive HIV patients from negatives ones. The 0.87 area under a curve suggests that the Random forest model has a probability 87% to distinguish a positive hyperglycaemia HIV patient from a negative. In conclusion, based on the two AUC curve results of XG-Boost and Random Forest ensemble Classifiers of 0.91 and 0.87, respectively, analysis from the ROC-AUC curves shows that the XG-Boost classification model is having 4% more chances of distinguishing hyperglycaemia positive HIV patients from negative ones than the Random Forest classification model. This implies the probability to classify a HIV patient as hyperglycaemia positive by XG-Boost classification model is higher than the Random Forest Classification Model by 0.04. Thus being able to identify the highest number HIV patients who developed hyperglycemia when switched to DTG as hyperglycaemia positive than Random Forest classifier.

**(vi) Precision Recall Curves**

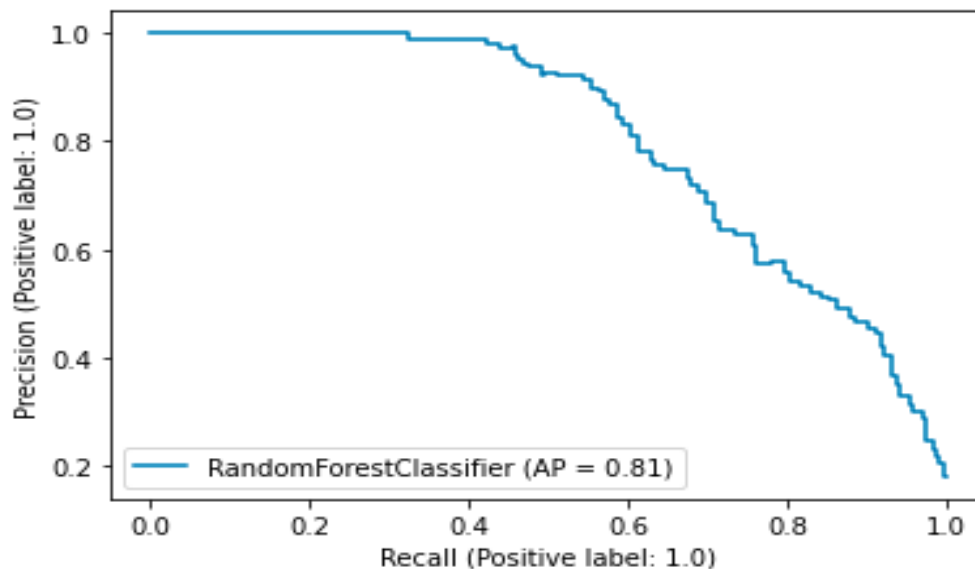
The precision-recall plot is able to show the performance difference between balanced and imbalanced cases. It is also useful to reveal the performance of high-ranking instances and research dataset had a big imbalance in classes. Also, the Precision Recall curve eliminate the false positive, false native element in the performance measure output. It uses the precision (positive predictive value) on the vertical axis and Recall on the horizontal axis as its parameters which are used to

calculate the area under the curve using the trapezoidal rule. That is why the Precision Recall curve is used to show the difference.



**Figure 40: Precision Recall Curve for XG-Boost Model**

The 0.87 area under XG-Boost curve (Fig. 40) shows that 87% of the XG-Boost model predicted positives are actually true positives. This implies the model has high ability to correctly identify HIV patients who turn hyperglycemic true positives (sensitivity) if switched to DTG. And for this research this is a metric with the highest priority. The model developed is intended to predict correctly those patients who turn hyperglycemic if switched to DTG.



**Figure 41: Precision Recall Curve for Random Forest Model**

Looking at the 0.81 area under Random forest curve as shown Fig. 41 means that 81% of Random forest model predicted positives are actually true positives. This exhibits that this model has high ability to correctly identify HIV patients who turn hyperglycaemia positive as true positives (Sensitivity) if switched to DTG. But from a comparison point of view the area under a curve of the XG-Boost model and Random Forest model reveals that XG-Boost model 6% more sensitivity than Random forest model. Thus the XG-boost model has better accuracy to predict positives than Random forest model.

#### **(vii) Cohen's Kappa Scores**

It is critical a performance measure of a classification model as compared to how well it can perform "NOT" just by chance.

#### **(viii) XG-Boost Model Cohen's Kappa Score**

The Cohen kappa score for XG-Boost model is 0.72. This shows that the XG-Boost model performance not by chance is 72%. This implies that the model ability to predict new out-comes using patterns from previously learned classifications is 72%.

#### **(ix) Random Forest Model Cohen's Kappa Score**

The Cohen kappa score for Random Forest model is 0.64. This also meant that the Random Forest model performance not by chance is 64%. Thus, model ability to predict new outcomes using patterns from previously learned outcomes is 64%.

In conclusion XG-Boost has 8% more "NOT" by chance abilities to predict the new outcome using patterns from previously learned outcomes than the Random forest classification model.

### **4.7 Summary**

The metrics used to assess accuracy and performance of the two models, have all shown that XG-boost model has a better performance and accuracy as compared Random Forest model. Therefore, based on that the XG-boost model was chosen for the DTG associated hyperglycaemia prediction tool development.

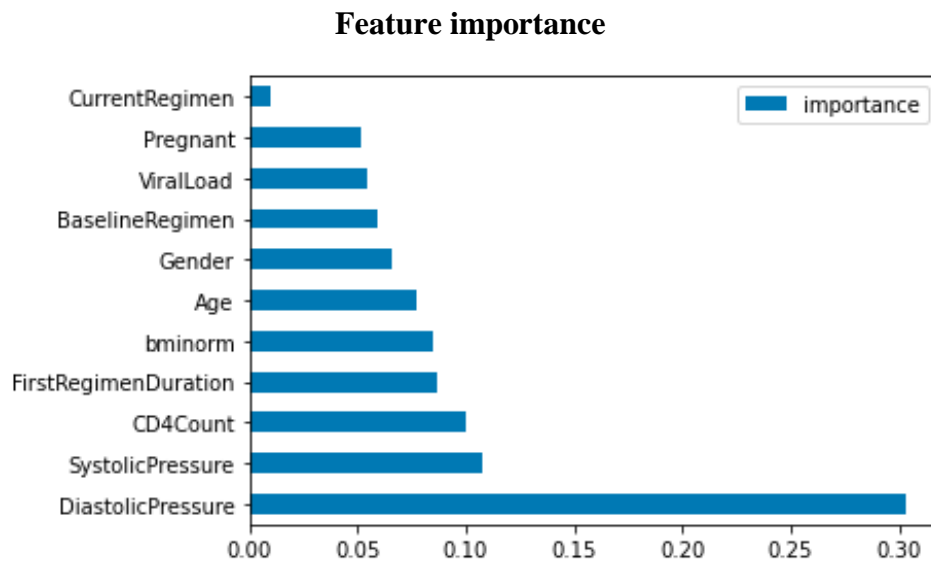
### **4.8 Selection of Predictor Variables**

Feature Importance is calculated by a measure each decision tree attribute split point improves the performance measure and, it is weighted by the particular node's number of observations. There after a simple harmonic mean of the feature importance is taken across all the decision trees within



the model. Feature importance scores are critical in a prediction model provision of insight into the data, model, and a foundation for dimensionality reduction and feature selection which improve the efficiency and effectiveness of a model prediction. After running the feature importance method, the output in Fig. 42 of horizontal bar chart shows in descending order of importance each feature from the one with highest importance score (Diastolic pressure) to the one with the least importance (Current Regimen). And from the feature importance chart we have seen that TB-Status effect on the prediction outcome of this model can be negligible (has no effect).

But to decide whether to exclude it from the predictor variables to be selected it's important to use the optimal feature selection graph as shown in Fig. 43.

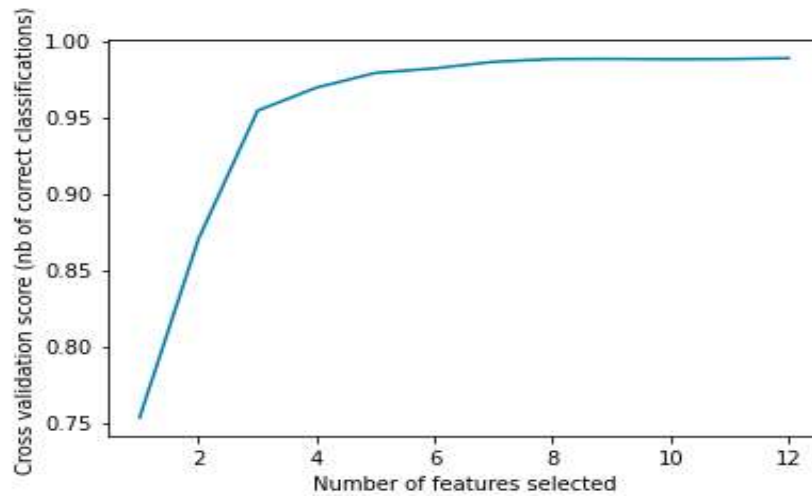


**Figure 42: The feature importance bar graph of XG-Boost Model**

#### 4.9 Optimal Feature Selection

After assessment of the feature importance with the optimal feature selection graph above, result indicates 12 as the optimal number of features (variables) that can be used to develop the DTG associated hyperglycaemia prediction tool. The Fig. 43 helps to evaluate and know the relevant and optimal number of columns (predictors) for the model which helps reduce application redundancy. This will help remove less important features that do not contribute much to our target variable in order to achieve better accuracy for our model. In addition, cross validation score shows that all the 12 variables must be included in the model predictor variables despite the feature importance graph showing that TB-status had no impact on the target variable.

Optimal number of features : 12



**Figure 43: Graph of optimal number of features**

#### 4.10 Dolutegravir Associate Hyperglycaemia Prediction Tool

A responsive DTG Associated Hyperglycaemia prediction application was developed to be accessed via a web browser on computers and laptops. The application and model reside on a local computer (desktop) that is accessed by a local user but also can be accessed remotely by a user on another computer. The responsive DTG Associated application functionalities are: login, user registration, prediction page, variable dictionary, Results page.

##### 4.10.1 User Registration

The application gives the administrator privileges to register users of the application. For this study the administrator can register someone as a user like shown in the Fig. 44. Registration is a done on easy-to-use graphical web interface as the python bootstrap library validates the form as user information is inserted. If the record is incorrect, the user information is not accepted by the system. The User details include: Names, Email address, Password, Username.

The screenshot shows a web interface for user registration. At the top, there are navigation links: Admin, Home, User, and Add User. Below the navigation is a heading: "Please Fill Your Information Here Below". The form contains three input fields: Email, Username, and Password. A blue "Sign Up" button is located at the bottom left of the form.

**Figure 44: User Registration Interface**

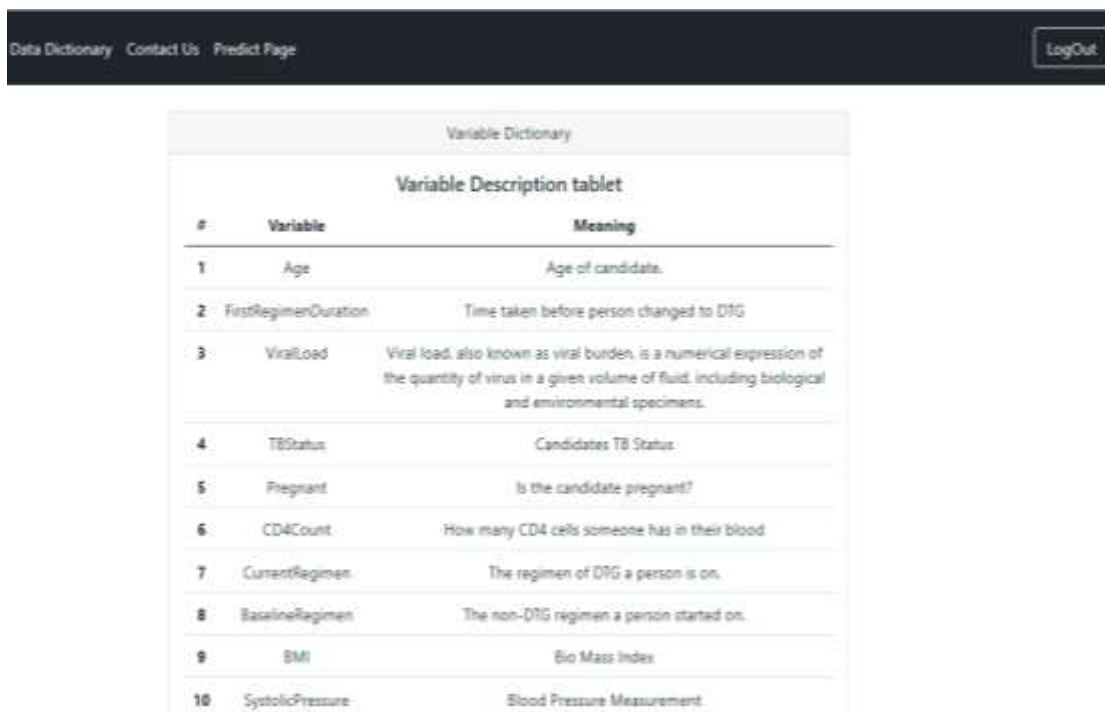
The administrator still manages the user accounts which allows the administrator to delete or edit the user accounts of the application as required from time to time as shown in Fig. 45.



**Figure 45: User Management Interface**

#### 4.10.2 Users

The users of this application have around three interfaces with features that can be readily accessed when logged-in. The variable dictionary page Fig. 46 which describes the meaning of all the variable being used in the prediction. The prediction page Fig. 47 where the predictor inputs are submitted and the prediction Results pages Fig. 48 and 49 where the output of the predictions (likely to develop hyperglycemia OR Not likely to develop hyperglycemia) are displayed. Lastly the contact us page (Fig. 47) which is used to contact the systems administrator just in-case the user experienced any challenge.



**Figure 46: The Dictionary Interface**

Please Fill Your Information Here Below:

Age(Years)- (Enter Values between 1 and 100)

Gender:- (Please Choose from the dropdown list below)

FirstRegimenDuration(Years)- (Enter Values between 1-35)

ViralLoad-(Copies)- (Enter values between 0 and 1 Million)

TBStatus:- (Choose from the dropdown below)

**Figure 47: Prediction Interface Page**

### 4.10.3 Validation

Validation is the commendation by examination and delivery of unbiased proof specific requirements of a certain proposed use are met. This guarantees user requirements satisfaction. Validation is not only testing but requirements must be clearly stated and exhibited (Kamalrudin & Sidek, 2015).

This DTG Associated Hyperglycaemia prediction tool has several validation stages which were selected and they included: Prediction Model validation, unit testing, integration testing, system testing, acceptance testing and regression testing.

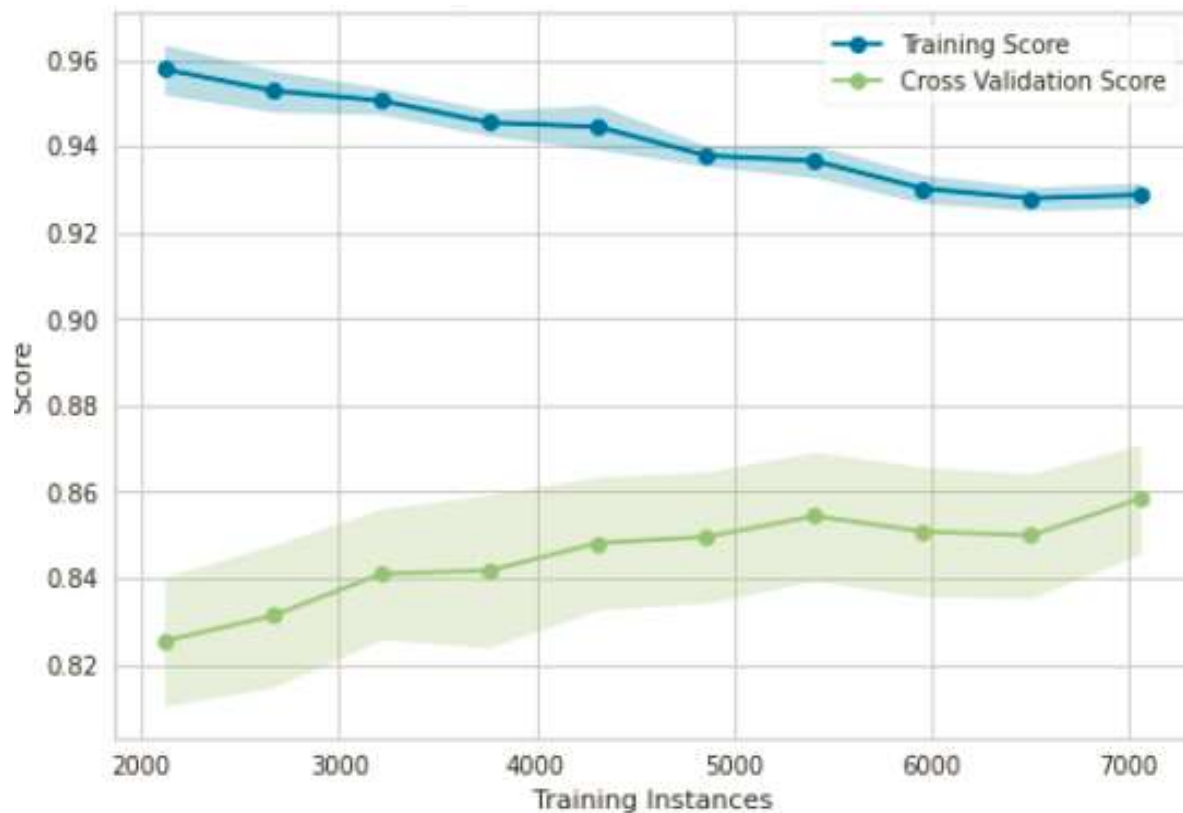
### 4.10.4 Validation of XG-Boost Model

Validation of the XG-boost model was done using synthetic model validation methods ideal situations for model deployment. The model will be tested for training cross validation score, Classification prediction error, discrimination threshold, and calibration evaluations.

### 4.10.5 Assessment of the Validation of XG-Boost Learning

The learning curve for the XG-boost (Fig. 48) shows that both the validation score and the training score are converging to a value that is very high with the increase in size of the training instances. This indicates that the health workers and patients will benefit more from the XG-Boost model

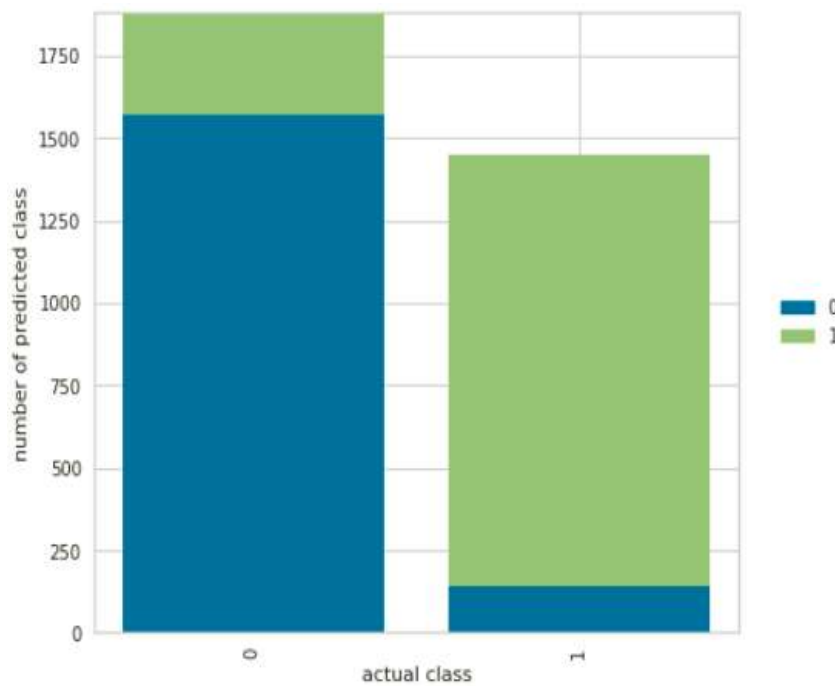
with more prediction data. This is because it will increase preciseness and accuracy when the mode is used more and more over a long time as it will from its predictions.



**Figure 48: Graph of XG-boost Learning Curve**

#### 4.10.6 Class Prediction Error for XG-Boost

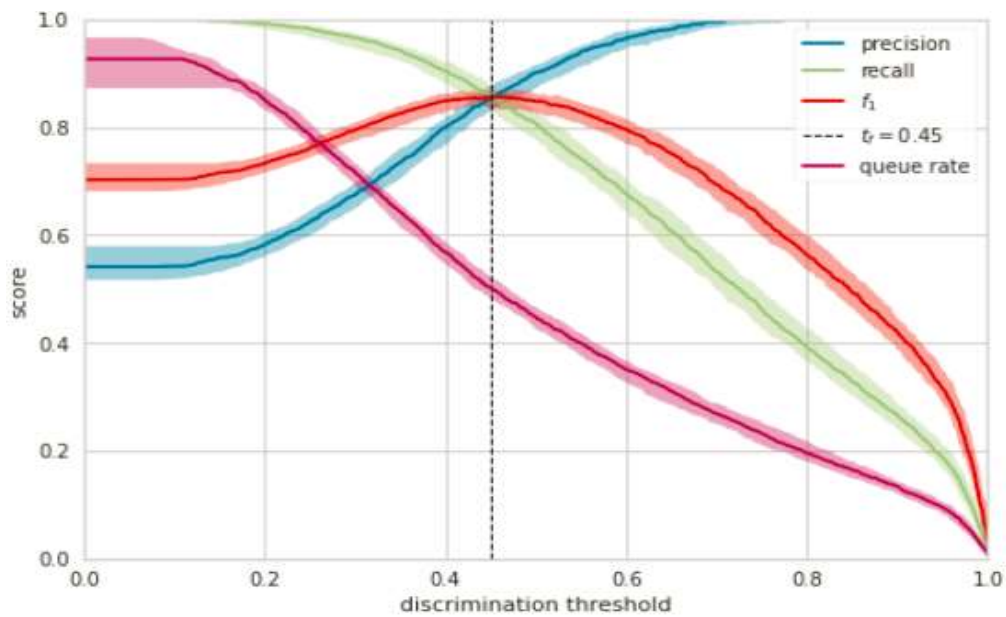
Figure 49 shows the two predicted classes of HIV patients the Hyperglycemia positive (1) and the hyperglycemia negative (0) from the number of HIV patients that were provided. The (0) hyperglycaemia negative classification with a blue color, has a small number of miss classified hyperglycaemia positive HIV patients as negative but with a green color (Type II Error). Also, the (1) hyperglycaemia positive classification with a green color has also a small number of miss classified hyperglycaemia negative HIV patients classified as positive but with a blue color (Type I error). This shows the classification error of the XG-boost model for the two classes. The bar chart shows that our XG-Boost model has a lower Type II error in class (0) relative Type I error in class (1). This is also an objective for this model because it always important for any developed health application to minimize error type II. Because this type of error undermines the very purpose for the application development. That miss classification causing error type II ends up classing would be positive as negatives thus causing the very problem it's trying to solve hence defeating its pupose.



**Figure 49: Validation graph of Class prediction Error for XG-Boost model**

#### 4.10.7 Threshold Report for CG-Boost

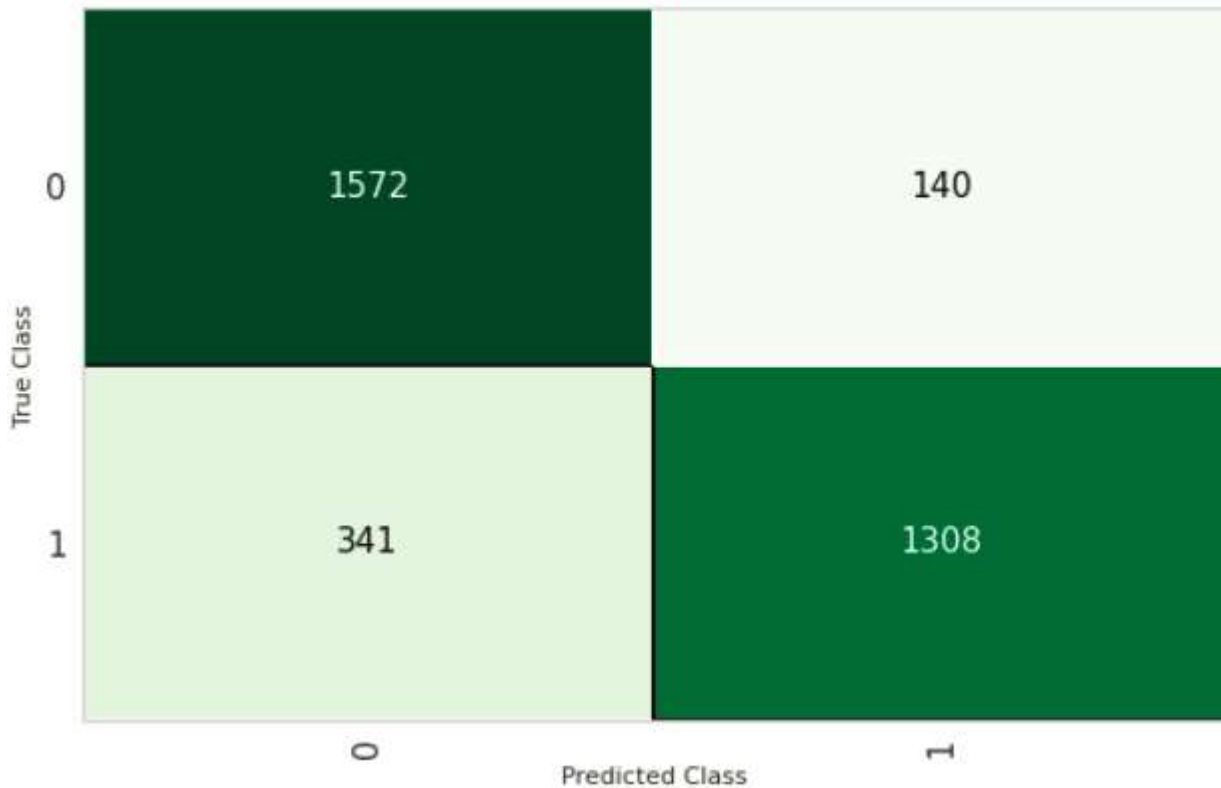
A discrimination threshold is the probability that a positive class is chosen over a negative class. The threshold discrimination at times is not the optimal threshold. The tuning of discrimination threshold in classifiers is made with the consideration of metrics of F1-Score, Precision, and Recall. Therefore, the XG-boost Classifier harmonic mean of precision and recall (F1-score) were tuned to make it behave optimally for DTG associated prediction tool. Figure 50 with graphs of Recall, Precision, F1-Score, Queue rate shows a threshold report of the XG-boost model adjusted to increase its sensitivity to false positives. At threshold discrimination the graphs show the score/probability (0.84) of the precision of the XG-boost is equal to probability of the recall as well the probability of the F1-Score. This means that all the classified hyperglycaemia positives HIV patients by the XG-boost model are true positive. This means also that the XG-boost model will only predict an HIV to be hyperglycaemia positive only if the probability of the being positive is 0.84 or above. This is a characteristic which is very important for our model because the cost of an HIV missing out on their right medication since miss-classification can be at times life threatening. From the observation of two graphs Recall and Precision, the researcher noted an inverse relationship with respect to discrimination threshold. The queue rate trailing the Precision, Recall and F1-Score suggests the cost of predicting at discrimination threshold as 0.5 delay probability. It even goes on decreasing with increasing precision and continues trailing all the other curves (Recall and F1-Score).



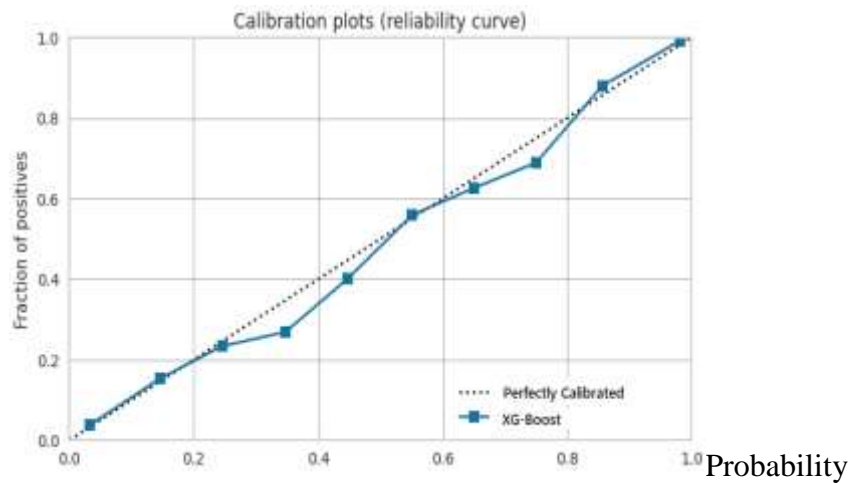
**Figure 50: Validation threshold Report for XG-Boost**

#### 4.10.8 XG-Boost Confusion Matrix

The Fig. 51 of confusion matrix (truth table) shows the true known classes of hyperglycaemia positive (1) and negative (0) participants versus the predicted classes of hyperglycaemia positive and negative participant. This helps show number of false negatives and false positives. But as well you can see the number of true negatives and true positive.



**Figure 51: Confusion Matrix of the XG-Boost Model**



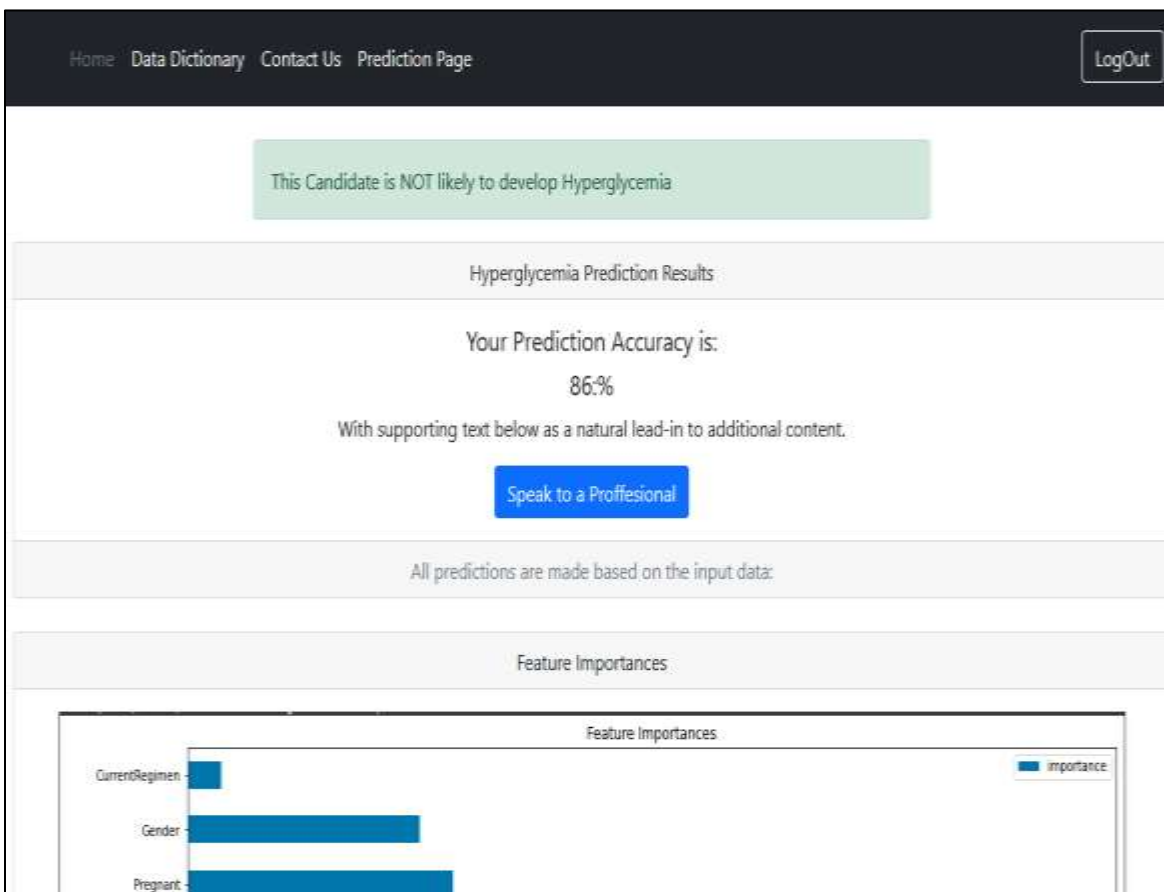
**Figure 52: Calibration Graph for the XG-Boost model**

Model calibration is the process where a model is trained to apply a post-processing operation, that enhances its probability estimation. Hence, a sample of participant where estimated to be positive with a 0.85 probability, the expectation of them being true positives is 85%. But the reality is not always true. Since we are developing a critical tool for screening DTG associated hyperglycaemia among HIV patients being switched to DTG, then the actual true positive is important. Figure 52 shows the calibration plot with the fraction of positives on the Y-axis against probability on x-axis. A reliability graph of XG-boost model (blue-line graph) was plotted to mimic the perfect estimation graph (black dotted line). The results show a relatively great calibration of the XG-Boost model. This implied a reliable probability prediction of positives by the XG-Boost model. This was because the model never showed pessimistic nor optimistic tendencies in its graph.

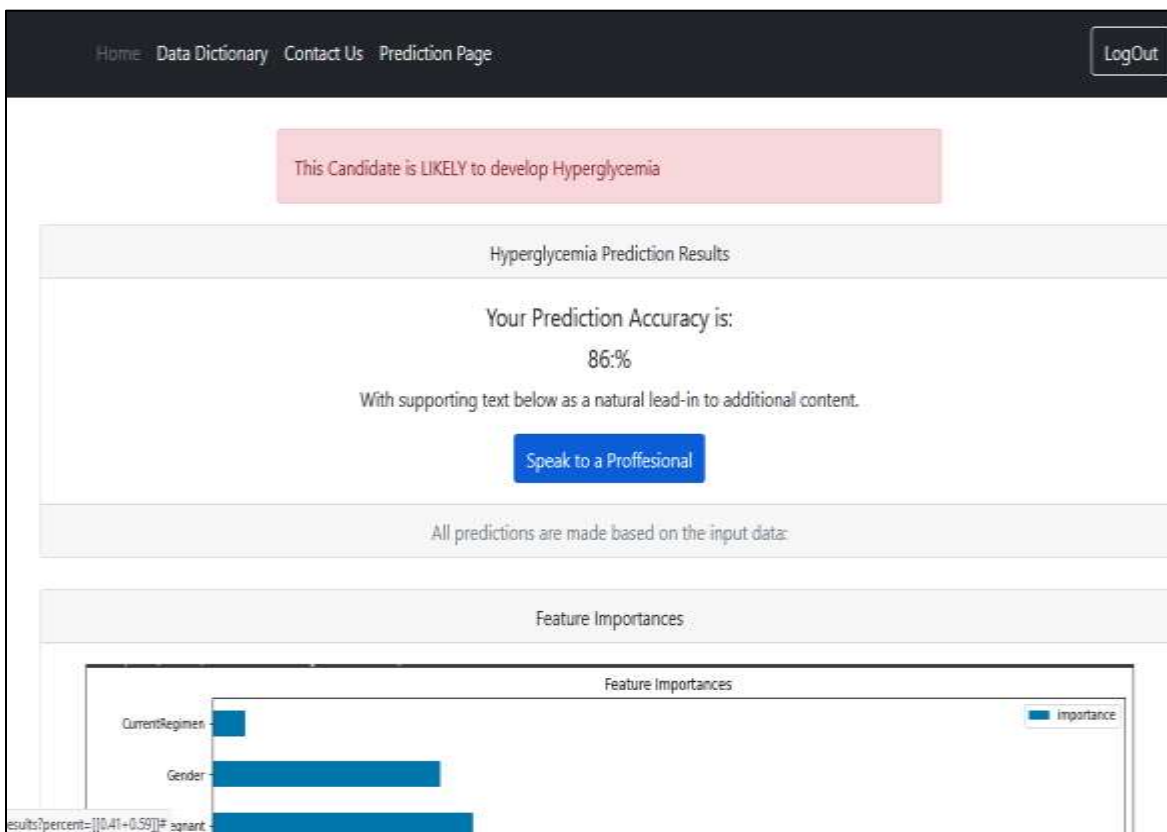
#### 4.11 Unit Testing

Unit testing is the verification of functional behavior of the smallest module of the system (Dybå & Dingsøy, 2008). The DTG associated hyperglycaemia prediction tool, units that were tested are: Administrative user management module, Login and user account module, Logout module, Admin Contact module and the Prediction page. The user can login into the prediction page with their username and password (Fig. 46) the first welcome page is data dictionary (Fig. 53) which describes the variable to be entered into prediction page (Fig. 54). For the user to access the prediction page selects the prediction tab on the top left corner of the top black bar. After all the required variables are entered, the user makes a prediction using the prediction button at the bottom thus displaying the prediction results pages in Fig. 53 and 54.





**Figure 53: Negative Prediction Results**



**Figure 54: Positive Prediction Results**

## 4.12 Integration Testing

Integration is meant to verify whether the smallest modules validated in the unit testing stage, function together appropriately and confirm to requirements specification in the low-level design (Nidhra, 2012). The functional units (login, admin management, user home module, user variable dictionary, user prediction page, admin contact module, Prediction results module) were verified and integrated to ascertain whether they properly work together. For instance, the login page was designed to authenticate users' access to specified account, view of information and Tabs for access to other unit modules of predict page, contact page, and Variable Dictionary.

## 4.13 System Testing

System testing is when the integration is tested to verify that the tool developed meets the specific end user and business requirements. The output of the integrated and validated functional modules is a system. Functional features that are visible to the end user make up a system (Nidhra, 2012). The DTG associated hyperglycaemia prediction tool passed all integration tests carried out while running on a work station (desktop). The variables submitted to the model from DTG Associated hyperglycaemia prediction tool were successfully interpreted, used for the prediction and even its patterns learnt from by the model. The summary of the system testing results are in the Table 3.

**Table 3: System Testing Results**

<b>Requirement</b>	<b>Description</b>	<b>Test Score</b>
Registration (Users)	Prediction tool shall allow administrator to register users and allow users to login for access, and the same applies to all new users	Pass
Manage user Accounts	The administrator shall Add, delete and edit user accounts	Pass
Contact us	Users shall contact Application administrator using that module when they encounter a challenge	Pass
View Prediction Results	It shall display prediction information to the user	Pass
Variable Data Dictionary	This page will have descriptive information of the various predictor variables.	Pass
Predict Page	This will allow the user predictor variables values input and prediction process.	Pass

#### **4.14 User Acceptance Testing**

Acceptance testing is anticipation and meeting of users' or customers' expectations in systems development and deployment. Acceptance testing objective is to affirm whether the software is functioning appropriately and confirms to the business process requirements. Privileged users interact with the functional unit of the application to assess correctness, speed, ease of use and responsiveness which are the fundamental issues of terms of reference for this testing (Nidhra, 2012). Questionnaires were given out to the application privileged users to find out their views about the developed application. A total of 18 privileged users were trained on how to use the DTG Associated hyperglycaemia prediction tool and were given 2 days to interact with the tool. Then later were given questionnaires to assess their tool experience. The user reactions were computed on five-point mean score of a Likert scale (5 = Strongly Agree, 4 = Agree, 3 = Neutral, 2 = Disagree and 1 = Strongly Disagree) as shown in Table 4.

The mean score of the study shows that each validated feature was above 3.5 and many of the sampled privileged users accepted the tool. They expressed readiness to use the tool to screen HIV patients for possibility of development of DTG associated hyperglycemia if switched to DTG. Their recommendation was to use the tool to predict out comes of using the DTG based regimen and informed decision of ART transition basing on facts.

**Table 4: User Acceptance Results**

Validation Feature	Respondents					Mean Score
	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	
The DTG Associated hyperglycaemia prediction Tool is easy to use	0	0	2	5	11	4.50
The DTG Associated hyperglycaemia prediction Tool shall impress many health workers to use it for prediction.	0	0	1	6	12	4.83
The DTG Associated hyperglycaemia prediction Tool will help health workers to save many HIV patients from developing hyperglycaemia	0	0	2	4	12	4.55
The DTG Associated hyperglycaemia prediction Tool will greatly help in the screening of HIV patients who would develop hyperglycaemia before switching to DTG	0	0	1	3	14	4.72
I will recommend this tool to my team/ colleagues	0	0	0	1	17	4.94

#### 4.15 Discussion

The study revealed that the health workers used a manual Ministry of Health tool herein as Appendix 5 to screen HIV patients that must not be switched to DTG based regimen. The tool only looked out for the only patients who had already developed hyperglycaemia and Diabetes, (Mwebesa & Musinguzi, 2020).

The above ministry of health manual screening tool only looks out for patients that had already developed Hyperglycaemia and diabetes. It does not look out for development of hyperglycaemia if a patient is switched to DTG based regimen. This is where our DTG associated hyperglycaemia prediction tool becomes of critical importance in solving what would be challenging and life-threatening situations, of switching an HIV patient to DTG and develops hyperglycaemia. The above mentioned MOH screening tool only looks at age, BMI, History of hypertension as the only risk factors for hyperglycaemia. However, our tool has considered a diverse number of risk factors: Age, Gender, Baseline or Previous Regimen, Height, Weight, Duration taken on, Baseline/Previous Regimen, Diastolic Pressure, Systolic Pressure, Pregnancy, Viral Load, CD4 Count. The feature importance for the predictor variables shows all the above risk factors playing a significant role in contribution to the development of DTG associated hyperglycaemia as shown in the feature importance bar scores. There is also lack of knowledge and computerized tool to aid the screening diagnosis of HIV patients for the possibility of developing DTG associated hyperglycaemia hence this study being fundamental in this aspect and can also be used as a genesis to embracing computer aided health screening tools.

## CHAPTER FIVE

### CONCLUSION AND RECOMENDATIONS

#### 5.1 Conclusion

In Chapter four the researcher described analysis and results entailed in this study and developed application. Further the application functionality and interfaces where extensively discussed as well.

The chapter firmly presents contribution and conclusion of the study as well as recommendations for further research and investigations. The increasing number of treatment experienced HIV patients who are turning hyperglycemic, is a growing concern for the health workers and medical researchers. Though the ministry of health has tried to put in place a screening for diabetic and Hyperglycemia positive patients not be switched to DTG, still the underlying challenge of those who turn hyperglycemic after switching to DTG is still glaring. Despite other underlying challenges in HIV therapy and treatment researcher and health workers are strugglingly handling participants case by case as they emerge.

Research also showed that many of the HIV hyperglycemic patients turn positive after switching to DTG. Though health workers (Doctors, nurses, clinicians) are desperate to use any other different method to support solving the challenge of HIV patients turning hyperglycemic when switched to DTG. The developed prediction tool will help prevent the issue of HIV patients turning hyperglycaemia after being switched to DTG. Thus, help reduce the burden on health services provision due to reduced number of HIV patients turning hyperglycemic. The health domain is searching for ways of providing health services that suit individuals given that people have different genomic compositions.

The use of the prediction tool will help doctors switch to initiate on DTG, only participants who won't develop hyperglycaemia and thus maintain participants' health and long life. Thus, the developed DTG Associated hyperglycemia prediction tool addresses the following challenges that include:

- (i) Screening participants who would develop hyperglycemia if switched to DTG.
- (ii) Help doctors initiated only HIV treatment experienced patients who won't develop hyperglycemia onto DTG.

- (iii) Support doctors in decreasing the number of HIV patients who would develop hyperglycaemia if switched to DTG.
- (iv) Support the doctors to initiate treatment experience HIV patients to a therapy that would give them longevity of life.

This study will not just deliver a developed solution of a DTG associated hyperglycemia prediction tool to screen HIV patients who turn hyperglycemia positive if initiated on DTG, but also lessen the burden of comorbidities (hyperglycemia and HIV) exert on the health services provided to patients. This tool also will be proof of a pay of a data driven health service implementation policy that was advocated more than 10 years ago by WHO.

## **5.2 Recommendations**

This study was confounded to a non-probabilistic random sampling methodology of selecting participants who do not have diabetic, or hyperglycemic and they have been switched to a DTG based regimen for at least the past six months. Hence given the confounding factors and the procedural methodology used in the model development and confirming to health standards of accuracy for a health screening tool, the study advises the health sector to adopt the use of this screening tool while switching participants to a DTG based regimen.

The user (nurses, clinician, Doctors) and health-informaticians, should use this tool as foundation for contribution in the development of more prediction tool for health services implementation. There has been initiation and continuous improvement to data driven and evidence-based health services delivery. But for this to be realized, the health service implementers should invest more in health informaticians to help them realize collection of good and quality data relevant for the development of such vital tools. There should be revised means of getting and giving data without charging high research fees in-order to encourage innovation in health services which improve quality of health of patients.

The feminine human subjects have different hormonal changes when it comes pregnancy than men. Therefore, for further research should be conducted with stratified data according to gender and build separate models for women and men, thus use the pregnancy and post-menopausal as only variables for women. This improves the accuracy of the models due to the hormonal differences in both genders and they affect both of them differently with age and gender category. Furthermore, to make these models more comprehensive and improve accuracy, the left-out variables like history of hypertension, smoking, alcohol consumption, history of diabetes, Insulin levels, skin thickness should be included in each of the models for further study. And as well a separate study for a

prediction tool to screen for possibility of development of DTG associated hyperglycemia if treatment Naïve HIV patients are initiated onto a DTG based regimen.



## REFERENCES

- Abebe, M., Kinde, S., Belay, G., Gebreegziabxier, A., Challa, F., Gebeyehu, T., Nigussie, P., & Tegbaru, B. (2014). Antiretroviral treatment associated hyperglycemia and dyslipidemia among HIV infected patients at Burayu Health Center, Addis Ababa, Ethiopia: a cross-sectional comparative study. *BMC Research Notes*, 7(1), 1-8.
- Abdeldjouad, F. Z., Brahami, M., & Matta, N. (2020, June). *A Hybrid Approach for Heart Disease Diagnosis and Prediction using Machine Learning Techniques*. In *International Conference on Smart Homes and Health Telematics* (pp. 299-306). Springer, Cham. <https://scholar.google.com>
- Akella, A., & Akella, S. (2021). Machine learning algorithms for predicting coronary artery disease: efforts toward an open source solution. *Future Science*, 7(6), FSO698.
- Bernardino, J. I., & Antela, A. (2015). Efficacy of dolutegravir in treatment-naïve patients. The SPRING-1, SPRING-2, SINGLE and FLAMINGO trials. *Enfermedades Infecciosas Y Microbiología Clínica*, 33, 14-19.
- Balakrishnan, M., Christopher, A. A., Ramprakash, P., & Logeswari, A. (2021). Prediction of Cardiovascular Disease using Machine Learning. *Journal of Physics: Conference Series*, 1767(1), 012013.
- Bailly, F., & Cotelle, P. (2015). The preclinical discovery and development of dolutegravir for the treatment of HIV. *Expert Opinion on Drug Discovery*, 10(11), 1243-1253.
- Battineni, G., Sagaro, G. G., Chinatalapudi, N., & Amenta, F. (2020). Applications of machine learning predictive models in the chronic disease diagnosis. *Journal of Personalized Medicine*, 10(2), 21.
- Brownlee, J. (2020). *Iterative Imputation for Missing Values in Machine Learning*. <https://machinelearningmastery.com/iterative-imputation-for-missing-values-in-machine-learning/>
- Brownlee, J. (2021). *SMOTE for Imbalanced Classification*. <https://scholar.google.com>
- Bunescu, R., Struble, N., Marling, C., Shubrook, J., & Schwartz, F. (2013). *Blood Glucose Level Prediction using Physiological Models and Support Vector Regression*. In *2013 12<sup>th</sup>*

*International Conference on Machine Learning and Applications*. [https:// scholar. google. com](https://scholar.google.com)

Chae, S., Kwon, S., & Lee, D. (2018). Predicting infectious disease using deep learning and big data. *International Journal Of Environmental Research And Public Health*, 15(8), 1596.

Chan, M. (2016). *Consolidated Guidelines on the use of antiretroviral Drugs for Treating and Preventing HIV Infection*. Geneva, Switzerland: WHO. [https:// scholar. google. com](https://scholar.google.com)

Charlson, F., Sorsdahl, K., & Ahmadzada, S. (2019). Burden of non-communicable diseases in sub-Saharan Africa, 1990–2017: Results from the Global Burden of Disease Study 2017. *The Lancet*, 7, 1375-1387.

Castagna, A., Maggiolo, F., Penco, G., Wright, D., Mills, A., Grossberg, R., Molina, J. M., Chas, J., Durant, J., Moreno, S., & Doroana, M. (2014). Dolutegravir in antiretroviral-experienced patients with raltegravir-and/or elvitegravir-resistant HIV-1: 24-week results of the phase III VIKING-3 study. *The Journal of Infectious Diseases*, 210(3), 354-362.

Chin, J. (1991). Global estimates of HIV infections and AIDS cases: 1991. *AIDS*, 5, 57-62.

Coetzee, L., Bogler, L., De Neve, J. W., Bärnighausen, T., Geldsetzer, P., & Vollmer, S. (2019). HIV, antiretroviral therapy and non-communicable diseases in Sub-Saharan Africa: Empirical evidence from 44 countries over the period 2000 to 2016. *Journal of the International AIDS Society*, 22(7), e25364.

Da Cunha, G. H., Franco, K. B., Galvão, M. T. G., Lima, M. A. C., Fontenele, M. S. M., Siqueira, L. R., Ramalho, A. K. L., & Fehine, F. V. (2020). Diabetes mellitus in people living with HIV/AIDS: prevalence and associated risk factors. *AIDS Care*, 32(5), 600-607.

Ddamulira, C., Nsereko, N., Musoke, M., & Kiyingi, F. P. (2020). Community Based Non Communicable Disease Services as a Predictor of Improved Quality of Life of People Living with HIV in Uganda: A Randomized Controlled Trial. *Journal of Environmental Science and Public Health*, 4(4), 304-317.

Deshmukh, S. K. (2020). Machine Learning for Healthcare: Emerging Challenges and Opportunities in Disease Diagnosis. *Journal of Cellular Signaling*, 1(3), 76-78.

Gouda, H. N., Charlson, F., Sorsdahl, K., Ahmadzada, S., Ferrari, A. J., Erskine, H., Leung, J., Santamauro, D., Lund, C., Aminde, L. N., & Mayosi, B. M. (2019). Burden of non-

- communicable diseases in sub-Saharan Africa, 1990–2017: Results from the Global Burden of Disease Study 2017. *The Lancet Global Health*, 7(10), e1375-e1387.
- Gupta, S., Tran, T., Luo, W., Phung, D., Kennedy, R. L., Broad, A., Campbell, D., Kipp, D., Singh, M., Khasraw, M., & Matheson, L. (2014). Machine-learning prediction of cancer survival: A retrospective study using electronic administrative records and a cancer registry. *BMJ Open*, 4(3), e004007.
- Hamoodi, S. A. (2014). Website Development Life Cycle. *International Journal of Computer and Information Technology*, 03(05), 22-79.
- Healthline Editorial Team. (2020). *Healthline*. <https://www.healthline.com/health/hiv-aids/medications-list>.
- Kaleebu, P., & Aceng, J. R. (2018). *National HIV Drug Resistance Prevention, Monitoring and Surveillance Activities, National Status Report*. Kampala: The Uganda Minister of Health. <https://scholar.google.com>
- Kansiime, S., Mwesigire, D., & Mugerwa, H. (2019). Prevalence of non-communicable diseases among HIV positive patients on antiretroviral therapy at joint clinical research centre, Lubowa, Uganda. *Plos One*, 2019, 1-11.
- Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1), 1-9.
- Krittanawong, C., Virk, H. U. H., Bangalore, S., Wang, Z., Johnson, K. W., Pinotti, R., Zhang, H., Kaplin, S., Narasimhan, B., Kitai, T., & Baber, U. (2020). Machine learning prediction in cardiovascular diseases: A meta-analysis. *Scientific Reports*, 10(1), 1-11.
- Kouroua, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadisab, D. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 2015, 8-17.
- Lamorde, M., Atwiine, M., Owarwo, N. C., Ddungu, A., Laker, E. O., Mubiru, F., Kiragga, A., Lwanga, I. B., & Castelnuovo, B. (2020). Dolutegravir-associated hyperglycaemia in patients with HIV. *The Lancet HIV*, 7(7), e461-e462.
- Levitt, N. S., Steyn, K., Dave, J., & Bradshaw, D. (2011). Chronic noncommunicable diseases and HIV-AIDS on a collision course: Relevance for health care delivery, particularly in low-

resource settings: Insights from South Africa. *The American Journal of Clinical Nutrition*, 94(6), 1690S-1696S.

Lirri, E. (2018). *Uganda Switches to HIV Super Drug Dolutegravir*. <https://scholar.google.com/>

Lo, J., Oyee, J., Crawford, M., Grove, R., DeMasi, R., Curtis, L., Fettiplace, A., Vannappagari, V., Payvandi, N., Aboud, M., & Van Wyk, J. (2019). *Dolutegravir and Insulin Resistance*. In *Conference on Retroviruses and Opportunistic Infections*. <https://scholar.google.com>

Maniruzzaman, M., Rahman, M., Al-MehediHasan, M., Suri, H. S., Abedin, M., El-Baz, A., & Suri, J. S. (2018). Accurate diabetes risk stratification using machine learning: Role of missing value and outliers. *Journal of Medical Systems*, 42(5), 1-17.

Ming, C., Viassolo, V., Probst-Hensch, N., Chappuis, P. O., Dinov, I. D., & Katapodi, M. C. (2019). Machine learning techniques for personalized breast cancer risk prediction: Comparison with the BCRAT and BOADICEA models. *Breast Cancer Research*, 21(1), 1-11.

Mutabazi-Mwesigire, D., Seeley, J., Martin, F., & Katamba, A. (2014). Perceptions of quality of life among Ugandan patients living with HIV: A qualitative study. *BMC Public Health*, 14(1), 1-10.

Mwebesa, H., & Musinguzi, J. (2020). *Consolidated Guidelines for the Prevention and Treatment of Hiv and Aids in Uganda*. Kampala, Uganda: Ministry of Health. <https://www.health.go.ug>

Ondoa, C. J. (2016). *The Uganda HIV and AIDS Country Progress Report July 2015 - June 2016*. Kampala: Uganda Aids Commission. <https://scholar.google.com/>

Opito, R., Mpagi, J., Bwayo, D., Okello, F., Mugisha, K., & Napyo, A. (2020). Treatment outcome of the implementation of HIV test and treat policy at the AIDs Support Organization (TASO) Tororo clinic, Eastern Uganda: A retrospective cohort study. *Plos One*, 15(9), e0239087.

Rachel, N. (2021). *The History of HIV and AIDS in the United States*. <https://www.healthline.com/health/hiv-aids/history>.

Ritesh, J. (2015). *Predictive Modeling For Chronic Conditions*. Boca Raton: Florida Atlantic University. <https://scholar.google.com>

- Sarwar, M. A., Kamal, N., Hamid, W., & Shah, A. M. (2018). *Prediction of Diabetes Using Machine Learning Algorithms in Healthcare: International Conference on Automation and Computing*. Newcastle Upon Tyne. <https://www.goiole.com>
- Schaefer, J., Lehne, M., Schepers, J., Prasser, F., & Thun, S. (2020). The use of machine learning in rare diseases: A scoping review. *Orphanet Journal of Rare Diseases*, 15(1), 1-10.
- Sidibe, M. (2018). UNAIDS DATA. Geneva: UNAIDS. <https://www.google.com>
- Shah, B. M., Schafer, J. J., & DeSimone Jr, J. A. (2014). Dolutegravir: A new integrase strand transfer inhibitor for the treatment of HIV. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, 34(5), 506-520.
- UNAIDS. (2017). *An Ambitious Treatment Target to Help End the AIDS Epidemic*. Geneva, Avenue Appia, Switzerland. <https://www.google.com>
- UNAIDS. (2020). *Uganda*. <https://www.unaids.org/>
- UNAIDS. (2021). *UNAIDS-Factsheet\_en.pdf*. <https://www.unaids.org/shttps://www.unaids.org>
- Vithalani, J., & Herreros-Villanueva, M. (2018). HIV Epidemiology in Uganda: survey based on age, gender, number of sexual partners and frequency of testing. *African Health Sciences*, 18(3), 523-530.
- WHO. (2021). *Regional Office Africa*. <https://www.afro.who.int/>
- Woldaregay, A. Z., Årsand, E., Walderhaug, S., Albers, D., Mamykina, L., Botsis, T., & Hartvigsen, G. (2019). Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes. *Artificial Intelligence in Medicine*, 98, 109-134.
- Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. *Frontiers in Genetics*, 9, 515.



# Appendix 2: TASO Administrative Clearance for Data Acquisition



## The AIDS Support Organisation (TASO) Uganda Ltd.

TASO Headquarters  
Mulago Hospital Complex  
P.O. Box 10443, Kampala-Uganda  
Tel: +256 414 532 580/1  
Fax: +256 414 541 288  
Email: [mail@tasouganda.org](mailto:mail@tasouganda.org)  
Website: [www.tasouganda.org](http://www.tasouganda.org)

12<sup>th</sup> April 2021

Our Ref: TASO REC/ADMC04/2021-JG-REC-009

**Favor Ceaser Wisdom,**  
Nelson Mandela African Institute of Science & Technology  
[favorwisdom2014@gmail.com](mailto:favorwisdom2014@gmail.com)

Dear Caesar,

**REF: ADMINISTRATIVE CLEARANCE FOR PROTOCOL: "A PREDICTION TOOL FOR DTG ASSOCIATED HYPERGYCEMIA AMONG HIV PATIENTS IN UGANDA"**

Thank you for submitting a request for administrative clearance for the above referenced study.

After a favourable administrative review, TASO (U) Ltd has no objection to grant you permission to implement the study at the TASO Mulago and TASO Entebbe clinics (sites) as enclosed in your letter dated 26<sup>th</sup> March 2021.

You are reminded to comply with the provisions of the protocol approved by TASO Research Ethics Committee on 26<sup>th</sup> March 2021, and to follow the guidelines of Uganda National Council for Science and Technology, July 2014 and the COVID19 guidelines, July 2020 in carrying out this research project.



**Amendments:** All proposed changes to the study (including personnel, procedures, or documents) must be approved by the REC in advance through the amendment process.

**Adverse Events/Unanticipated Problems:** You must inform the REC of all unanticipated problems and adverse events that occur during your research study.

It is a requirement by TASO that you submit the end of study report upon completion of the study.

For further correspondence with us, our contact person is Dr. Levicatus Mugenyi, Research Manager, [lmugenyi005@gmail.com](mailto:lmugenyi005@gmail.com) +256 701099537 or Dr. Kagimu David, Vice Chairperson, TASO REC, [kagimud@tasouganda.org](mailto:kagimud@tasouganda.org) +256 752774157

Sincerely,

**Dr. Etukoit Bernard Michael,**  
**EXECUTIVE DIRECTOR.**

CC: Chairperson, TASO RESEARCH ETHICS COMMITTEE (REC)  
Center Program Managers: Mulago, Entebbe  
Center Program Managers: Entebbe

**TASO COLLEGE OF HEALTH SCIENCES (TASOCS)**  
Kampala CR (Garden Road)  
John Mukasa  
P.O. Box 3445, Kampala  
Tel: +256 414 541 417  
Fax: +256 414 541 288  
Email: [tasocs@tasouganda.org](mailto:tasocs@tasouganda.org)

**SERVICES CENTER**  
**TASO ENTebbe**  
Plot 12-13, Independence Avenue  
P.O. Box 124, Entebbe  
Tel: +256 414 532 577/578/579/580/581  
Fax: +256 414 541 288  
Email: [entebbe@tasouganda.org](mailto:entebbe@tasouganda.org)

**TASO MULAGO**  
Plot 4, Mulago Hospital Road  
P.O. Box 347, Mulago  
Tel: +256 414 532 582/583/584/585  
Fax: +256 414 541 288  
Email: [mulago@tasouganda.org](mailto:mulago@tasouganda.org)

**TASO JUBA**  
Juba National Hospital  
P.O. Box 517, Juba  
Tel: +256 414 532 586/587/588/589/590  
Fax: +256 414 541 288  
Email: [juba@tasouganda.org](mailto:juba@tasouganda.org)

**TASO MUKAMA**  
Mukama Hospital  
P.O. Box 1075, Mukama  
Tel: +256 414 532 591/592/593/594/595  
Fax: +256 414 541 288  
Email: [mukama@tasouganda.org](mailto:mukama@tasouganda.org)

**TASO MADINDI**  
Madindi Hospital  
P.O. Box 111, Madindi  
Tel: +256 414 532 596/597/598/599/600  
Fax: +256 414 541 288  
Email: [madindi@tasouganda.org](mailto:madindi@tasouganda.org)

**TASO MWALE**  
Mwala Hospital  
P.O. Box 2200, Mwala  
Tel: +256 414 532 601/602/603/604/605  
Fax: +256 414 541 288  
Email: [mwala@tasouganda.org](mailto:mwala@tasouganda.org)

**TASO MBARARA**  
Plot 25, East of Hospital Road  
P.O. Box 1016, Mbarara  
Tel: +256 414 532 606/607/608/609/610  
Fax: +256 414 541 288  
Email: [mbarara@tasouganda.org](mailto:mbarara@tasouganda.org)

**TASO WALUGU**  
Walugulu Hospital  
P.O. Box 14460, Walugulu  
Tel: +256 414 532 611/612/613/614/615  
Fax: +256 414 541 288  
Email: [walugulu@tasouganda.org](mailto:walugulu@tasouganda.org)

**TASO KUMUNDA**  
Kumunda Hospital  
P.O. Box 280, Kumunda  
Tel: +256 414 532 616/617/618/619/620  
Fax: +256 414 541 288  
Email: [kumunda@tasouganda.org](mailto:kumunda@tasouganda.org)

**TASO BODDIT**  
Boddit Hospital  
P.O. Box 422, Boddit  
Tel: +256 414 532 621/622/623/624/625  
Fax: +256 414 541 288  
Email: [boddit@tasouganda.org](mailto:boddit@tasouganda.org)

**TASO TOSORO**  
Plot 20, Cox Road  
P.O. Box 1774, Tosoro  
Tel: +256 414 532 626/627/628/629/630  
Fax: +256 414 541 288  
Email: [tosoro@tasouganda.org](mailto:tosoro@tasouganda.org)

**RESEARCH PROGRAMS**  
**GRANTS MANAGEMENT UNIT /**  
**GLOBAL FUND**  
House 19, Plot 19  
Wanda Road  
P.O. Box 10443, Kampala  
Tel: +256 414 532 592/593/594/595/596  
Fax: +256 414 541 288  
Email: [grants@tasouganda.org](mailto:grants@tasouganda.org)

**TASO BAHAMUKA PROJECT**  
Plot 12, Independence Avenue  
P.O. Box 124, Entebbe  
Tel: +256 414 532 577/578/579/580/581  
Fax: +256 414 541 288  
Email: [bahamuka@tasouganda.org](mailto:bahamuka@tasouganda.org)

**TORORO LABORATORY UNIT**  
P.O. Box 111, Tororo  
Tel: +256 414 532 631/632/633/634/635  
Fax: +256 414 541 288  
Email: [tororo@tasouganda.org](mailto:tororo@tasouganda.org)

### Appendix 3: Mildmay Administrative Clearance for Data Acquisition



26 April 2021

**Favor Ceaser Wisdom**

Nelson Mandela African Institution of Science and Technology  
Principal Investigator

Dear Favor,

**Administrative clearance for a Research Protocol Titled: "A Prediction Tool for DTG Associated Hyperglycaemia Among HIV Patients in Uganda."**

Thank you for submitting your protocol to Mildmay Uganda for administrative clearance to conduct research at Mildmay Uganda's supported District Health facilities.

Following approval of your study by the Nelson Mandela African Institution of Science and Technology on 20<sup>th</sup> January 2021 and The Aids Support Organisation (TASO) Uganda on 26<sup>th</sup> March 2021 which expires on 25<sup>th</sup> March 2022, you are hereby authorised to access data/research participants at Mildmay Uganda.

Please note that you are required to share your findings with Mildmay Uganda.

This clearance is valid for the same period as that provided by the Research Ethics Committee which approved the study.

Yours Sincerely,

A handwritten signature in black ink, appearing to read "Mary Oditi".

Mary Oditi

**Director Research and Strategic Information**

**Mildmay Uganda**  
PO Box 24985  
Kampala  
Uganda  
tel: +256 312 210 200  
fax: +256 312 210 205  
[www.mildmay.org/uganda](http://www.mildmay.org/uganda)  
NGO NO. S.5914 / 9191



**DATA ABSTRACTION TOOL**

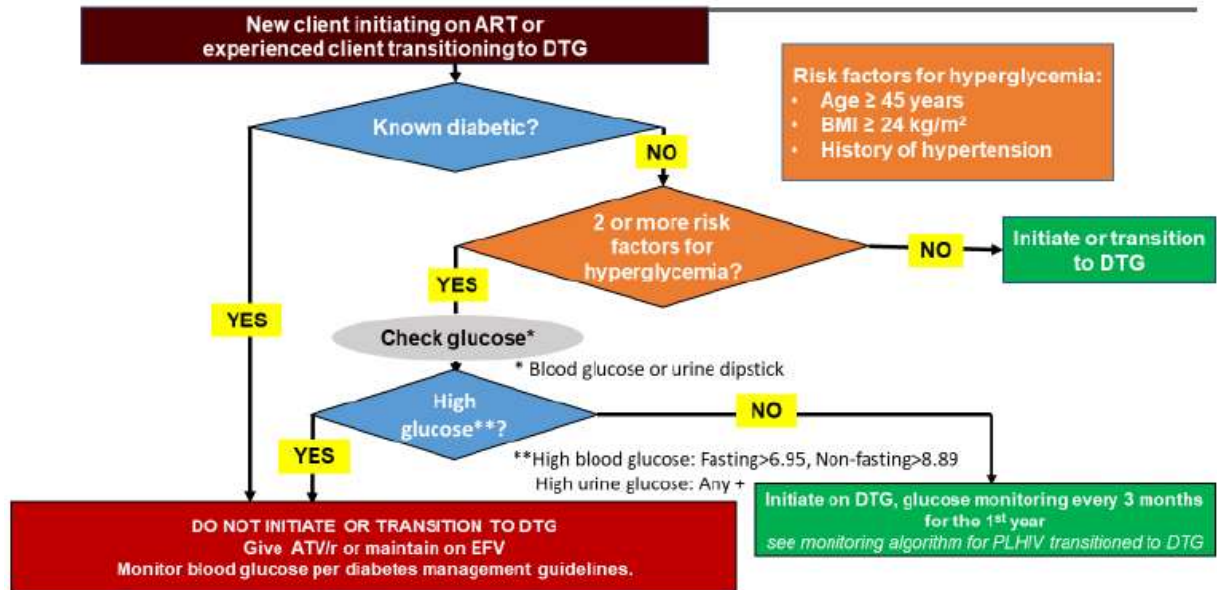
**A PREDICTION TOOL FOR DTG ASSOCIATED HYPERGYCEMIA AMONG HIV PATIENTS IN UGANDA**

- ✓ glucose levels
- ✓ Data of birth
- ✓ Insulin levels
- ✓ Blood pressure
- ✓ height
- ✓ Random Blood glucose
- ✓ Random Blood Sugar / Fasting blood Sugar
- ✓ latest CD4 count at the transition to DTG
- ✓ latest Viral Load count at DTG Transit
- ✓ Alcohol
- ✓ Smoking
- ✓ Family medical histrol (cancer, Hypertension, Diabetes)
- ✓ Previous Regimen
- ✓ Date initiated on the first Regimen
- ✓ Date initiated on DTG
- ✓ pregnant
- ✓ duration of pregnancy
- ✓ Weight at initiation on DTG
- ✓ Current weight while on DTG
- ✓ DTG regiment initiated on
- ✓ TB status
- ✓ Cholesterol.



## Appendix 5: Screening tool

### Screening Algorithm for PLHIV initiating or transitioning to DTG



## POSTER PRESENTATION