UNIVERSITY OF NAIROBI

DEPARTMENT OF COMPUTER SCIENCE

Predictive Analytics for Retention in Care and Antiretroviral Therapy Adherence Using Supervised Learning: A Case Study of County Health Facilities in Kenya.

GEOFFREY SAGWE OMBUI

P52/37791/2020

SUPERVISOR/PROJECT MENTOR

PROF. ANDREW MWAURA KAHONGE

Research Project Report Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Computational Intelligence, Department of Computer Science, University of Nairobi.

August 2023

DECLARATION

This project is my original work and has not been presented in any other institution for the purpose of academic award. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the research itself.

Signature:..... **Date:**01/sep/2023

Name: Geoffrey Sagwe Ombui **Registration Number:** P52/37791/20

Supervisor's Approval

This thesis report has been submitted in partial fulfilment of the requirements for the Degree of Master of Science in Computational Intelligence of the University of Nairobi with my approval as the University Supervisor.

Signed Date:1/sep/2023

Prof. Andrew Mwaura Kahonge:.....

DEDICATION

I dedicate this research to God Almighty for his grace and good health throughout the research. A special profound gratitude to my loving parents Ombui, and Anne Kemunto, and my siblings Dennis, Lydia and Brian for their words of encouragement and continuous support through the research process and friends who supported me through the journey.

ACKNOWLEDGEMENT

I acknowledge the support of my supervisor, Professor. Andrew Kahonge for his unwavering support and immense contribution and guidance. I would also like to acknowledge the support and help from CEMA team headed by Prof Thumbi Mwangi for allowing me to conduct my research and proving funding and resources. I would also like to thank the experts and panelists including, Dr. Mwendwa Kinyari, Dr. Evans Miriti and Mrs. Selina who were involved in the validation for this research project: Without their passionate participation and input, the validation could not have been successfully conducted.

LIST OF ABBREVIATIONS

PLWH	Persons living with HIV
ART	Antiretroviral Therapy
CD4	Cluster of Differential 4 Cells
LTFU	Lost to Follow Up
HIV	Human Immunodeficiency Virus
UNAIDS	United Nations Programme on HIV/AIDS
CEMA	Centre for Epidemiological Modelling and Analysis
CHIDH	Center for Health Informatics and Digital Health
CDC	Center for Disease Control and Prevention
WHO	World Health Organization
VL	Viral Load
EMR	Electronic Medical Records
ANN	Artificial Neural Networks
RF	Random forests
OLS	Ordinary Least Squares
NB	Naive Bayes
CAI	Case Adherence Index
GEE	Generalized Estimating Equations
TCE	Treatment Change Episode
LMICs	Low-and Middle-Income Countries
MOH	Ministry of Health
IQR	Inter-Quartile Ranges

DEFINITION OF TERMS

Retention in care – Remaining connected in care and starting diagnosis of HIV infection till lifelong services.

Viral Suppression – Suppressing or reducing the function and replication of a virus.

Lost To Follow Up (LTFU) – Dropping out from clinical care program and failing to attend a clinic within a specified interval after a previous visit.

Machine Learning – Is a field of Artificial Intelligence that allows computers to act without being explicitly programmed.

Antiretroviral Therapy (ART) – The treatment that suppresses or stops a retrovirus by preventing the virus from multiplying, which reduces the amount of HIV in the body.

Linkage to Care – Accessing medical care tests within 30 days (1 Month) of HIV diagnosis and completion of a first medical clinic visit.

Cohort Data – A kind of behavioral analytics that breaks the data in a dataset into related groups before analysis.

Comorbidities – The condition of having two or more diseases at the same time. More than one disease or condition present in the same person at the same time.

Data Visualization – Creating compelling graphical representation to make it easier to identify real-time trends, outliers and new insights through use of graphics and visual analytics.

Data quality – To measure objective elements of data such as completeness, accuracy, and consistency.

ABSTRACT

Developing countries in sub-Saharan Africa are scaling up ART programmes to reduce HIV transmission for patients infected by the diseases (UNAIDS, 2014). In healthcare organizations, a great problem is faced by healthcare providers to know the ART adherence and status of HIV/AIDS patients. In this research, a predictive model using supervised learning is developed to let clinicians and healthcare providers know the ART adherence of PLHIV using features of the patients' treatment history. The research explores the use of different machine learning methods to be able to detect records of patients defaulting and switching ART treatment. The methodology used was CRISP-DM data mining process. The dataset collected from the Ministry of Health sampling unit illustrates measurable baseline and clinical variables such as body weight, ART regimen, patients enrolled in care, Z-Score and phenotype. Data preprocessing and transformation was done to ensure the dataset collected was clean. Predictive model was designed in the process of data collection, dataset preprocessing like missing data, outlier data, feature selection and feature transformation and normalization of the dataset. Data was split into train and test set i.e., 80% training and 20% for test set and model designing, training and evaluation was performed using Anaconda.

The baseline results from the benchmark and performance evaluation showed that ensemble random forest algorithm performed the best with training accuracy of 81% and AUC of 79.3% compared to other binary algorithms and classification error rate of 0.333. The machine learning model that performed poorly was Naïve Bayes with an accuracy score of 20.0%. The researcher retrospectively followed 21551 records of patients who were seeking care at comprehensive health care units and county health facilities.

Keywords: Supervised Learning, Demographic, HIV, Retention in care, Viral Suppression, Data Mining, CRISP-DM methodology, Lost to Follow Up (LTFU), Area Under Curve, ART regimen, Ministry of Health (MOH).

Table of Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE: INTRODUCTION

1.1 Background of the Study

HIV epidemic remains a challenge globally with the international community improving access to HIV care and prevention as a life-saving antiretroviral treatment. For countries to provide effective prevention and long-term HIV care with ART, it requires effective and integrated patient monitoring system at the health facility level. According to (Granich RM, 2009) observational analyses and mathematical models suggests that ART treatment could reduce the epidemic spread hence improve economic productivity. In diverse settings using temporal trends will improve effectiveness of ART programs in middle-income countries (Grimsrud A, 2014). In Kenya a significant improvement in reduction in mortality rates and viral suppression reduced deaths rates in 2019 with estimate models demonstrating scale up of ART treatment-initiated scale. Overall life expectancy among patients with the survival between 5 and 10 years less than among uninfected people with deaths occur due to ART interruptions (Slaymaker E, & Hosegood V, 2014). ART initiation is key approach in reaching a worldwide target of 90% for patients to know their status, after receiving ART regimen to suppress viral replication (UNAIDS, 2017). Patients taking ART and retained in care are able to prevent the viral level eliminating effectively transmitting HIV to others. A key pillar of public health plans is retention in care to eliminate HIV transmission (CDC, 2018). As machine learning models and datasets proliferates, healthcare services are harnessing this AI technology to predict and prevent adverse outcomes of HIV.

When health outcomes and clinical data are monitored, machine learning algorithms can predict clinical health outcomes (Jacob O & Sindri, 2019). Under CDC program, Palladium has tested machine learning algorithms to assess the clinical and biomarker data of specific demographic to predict the risk of treatment failures on a set of de-identified patient records to identify consistencies in the predictions (Palladium, 2019). The machine learning models compares attributes and clinical variables of persons who exhibits treatment failure with viral suppression. Supporting Patient education, psychosocial support, adherence counseling, and patient-centered models of care are critical interventions to maximize retention. Strategies are needed to accurately and easily target the subset of individuals at highest risk for defaulting on treatment in limited resources setting with tailored interventions, thus enhancing programmatic effectiveness and efficiency. To explore predictive models for ART regimen could discover patterns and identify key factors on patient clinical historical data (H. E. Gendelman, 2019).

Machine learning technique can harness, enhance HIV prevention and increase the prediction capability by processing large amounts of data.

The method can be implemented to identify patterns of HIV risk behaviors and optimize HIV treatment modalities for targeted interventions from a number of novel data sources. In this study, we examined antiretroviral therapy initiation and retention for ART treatment-eligible patients in health facilities sampling unit. We aim to scale up interventions and improve timely linkage to treatment for persons starting ART and those enrolled for ART using machine learning algorithms for routinely collected agglegated data phenotypic and clinical data in sub-county health facilities in Kenya programmes and identify predictors of the key clinical outcomes.

1.2 Problem Statement

The problem of combating HIV lies in keeping virally suppressed patients retained in care for their lifetime in developing effective treatment (Avishek Kumar, 2020). ART uptake and viral suppression treatment highlight serious problem of patients with fewer health care visits and poor access to care hence avoiding such discussions with healthcare providers (Mayer KH, 2011). Clinicians when recommending ART to patients may delay the treatment and the low self-efficacy where patients are hesitant to take ARTs when prescribed by their healthcare providers assuming patients have started it (Kremer H, 2006). Poor linkage to care and treatment where clients who are referred, do not get to the facilities is a major bottleneck to treatment in Kenya (Wachira J, 2014). The community and individual side effects of patients defaulting ART include earlier mortality, viral resistance and increased high health care costs.

Manually ordering of ART and ARV regimen in ART Centers, by itself has a problem and is more prone errors considering the side effects of aforementioned parameters could greatly cause side effects. Clinicians often fail to recognize the immediate needs of patients disengaging with HIV treatment (Renju J, et al, 2017). Interventions using peer navigation, intensive management, and multi-faceted outreach programs is effective for timely delivery of results. To monitor ART treatment programs, population surveys to measure prevalence through linkage to care and viral suppression will help in evaluating treatment programs (WHO, 2012) and to identify patients not in adherence to ART treatment in achieving undetectable viral load (Gardner EM, 2011). The research aim to identify persons living with HIV/AIDS to explore the feasibility of recruiting PLHIV not on ART and compare them with respect to clinical characteristics and to explore patients who have taken ART and individuals retained in care at 12 months and switching of ARTs after virological failure from HIV care using machine learning models.

1.3 Overall Objective

Develop a machine learning model for retention in care and ART adherence for community-based testing treatment program in County HIV clinics.

1.4 Specific Objectives

1. To evaluate the performance of machine learning model for easy mapping coverage of treatment cohorts.
2. To monitor and assess treatment outcomes of patients switching from first to second line ART to achieve viral suppression.
3. To review experimental results and architectures applicable for effective monitoring of patients' cohort data.

1.5 Research Questions

1. How to compare and estimate predictive outcomes of first-line and second line ART?

2. How to quantify retention in care to establish baseline data for targeted mapping outreach for population level-subgroups?
3. How to identify patients enrolled in care and starting on ART using predictive models?

1.6 Significance of the Study

We present a cohort data for persons living with HIV/AIDS recruited at various county health facilities. The study explores persons' experiences and treatment at different phases of HIV care, access to ART, retention in care to understand the context of ART adherence and switching of ART after virological failure.

The research provides an ART adherence, retention in care and switching of first line and second line therapy that often characterize the lives of patients in each study setting. High-quality care and clinical outcomes means patients seek first care in starting ART quickly, retained in care and achieves viral suppression. We derive the prediction model by measuring quality of ART treatment using patient ART outcomes to the quality of clinical care and treatment for healthy facilities. The research constructs patient-level indicator to standardize ART treatment outcomes including, viral suppression, immune recovery presentation of retention with different groups, and monitoring for ART adherence.

1.7 Scope of the study

The research focuses on antiretroviral treatment in County health facilities HIV Treatment Cohorts. The targeted stakeholders include; HIV Patients, Health care providers and clinicians. The main study aims at developing a real-time and effective machine learning model for easy mapping coverage of treatment cohorts to identify patients defaulting ART adherence and to offer a viable intervention using predictive analytics techniques to make the ART adherence and switching of therapies more objective and reliable, thus enabling a faster prevention, to achieve viral suppression.

CHAPTER TWO: LITERATURE REVIEW

2.0 Introduction

The literature reviews use of predictive analytics, other machine learning models technologies for retention and viral suppression. Related works and findings by other authors were reviewed and a research gap identified. The study evaluated the literature and models used to predict the viral suppression and ART adherence and highlight the research shortcomings. In order to better analyze the application of predictive analytics for retention, an analysis of these is pertinent.

2.1 Retention and Viral Suppression.

ART treatment decreases morbidity among HIV positive individuals (Bendavid E, & Holmes CB, 2012). Switching of therapy and treatment failure occurs in the 6 and 7 months after initiating ART(Keiser O, & Tweya H, 2009). Lack of adherence increase viral transmission and development of resistance. Leveraging of knowledge base and clinical outcomes of patients will prevent treatment failure and clinical decision making (Revell AD, 2013).

Strict antiretroviral therapy adherence of the ART regimen by the patients is key to sustaining viral suppression thus improving quality of life reducing the drug resistance and defaulting ART treatment leads to patient losing future treatment and developing drug resistance. Significant number of patients not initiating ART leads to a gap of 90% ART coverage (UNAIDS, 2019). Potential ART users perceive treatment as essential to new treatment strategy. ART naïve patients in UK found willing to start ART at a percentage of 43% - 47% (Rodger AJ, 2014). If patients miss their appointment they are considered loss to follow-up and not retained in care for treatment.

2.1.1 Feedback loop to surveillance

To optimize HIV surveillance, Martin Holt (1995) describes monitoring not only measures the epidemic but also shaping how it is constituted. Surveillance and monitoring identifies problems for intervention to provide a mechanism for the success and interventions.

Looping process demarcates subject-positions such as high-risk and low-risk in determining patients likely to be poor in health and reaching out to them for retention. Patients identified as out of care are part of HIV monitoring and surveillance in data to care feedback loop.

Data to care programs provides safety to medication adherence as part of broader developments in the governance of HIV (CDC, 2017).

In the research paradigm, new parameters are created through daily medication adherence by targeting patients for additional support services (Guta, *et al.*, 2016).

Patients unable to sustain viral suppression owing to physiological factors related to ability to access services (Kiweewa, *et al.*, 2019). Persons unable to comply with the terms of public health interventions are positioned in the era of being subjected to increased scrutiny and outreach by partnering entities (Dombrowski, *et al.*, 2017).

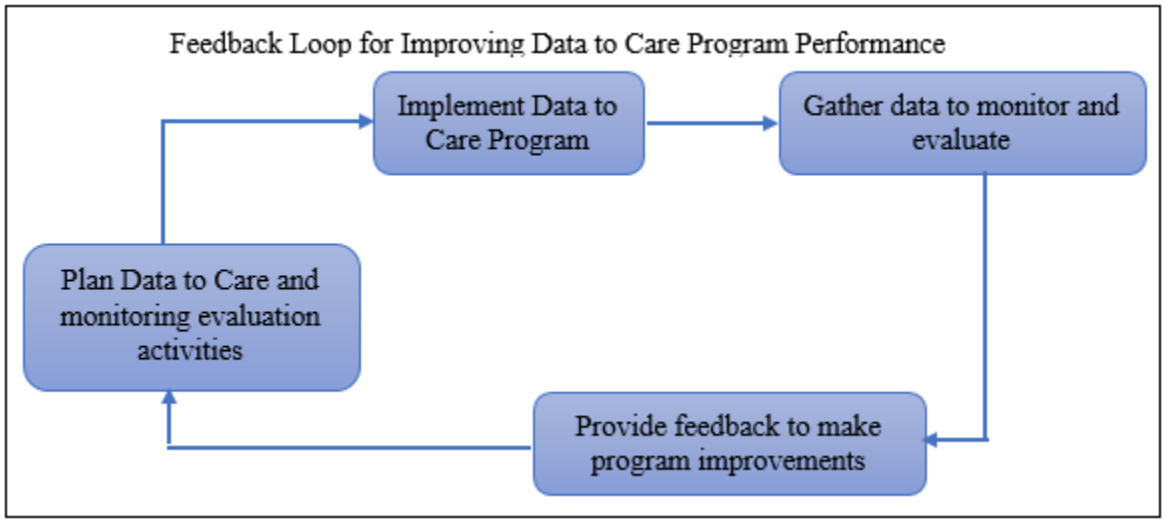


Figure 1: Toolkit visualization of the HIV Data to care Feedback Loop

2.1.2 Predictive Analytics

Predictive analytics applies various quantitative methods on data to make real-time predictions based on future performance of current and historical data. Predictive models use historical and analytical data to forecast future outcome based on criterion variables (Hair, 2007). Predictive analytics goals include, actionable information and predicting future events including large amounts of data for volume, speed and value.

2.2 Machine learning algorithms

Prediction of clinical outcomes guide decision making and may be limited by virological data not easily available (Wang D, & Larder B, 2009). Models that avoid genotype data use complex classifiers such as neural networks (NN) and random forests (RF) classifier as the backbone of prediction not easily interpretable by healthcare providers (Revell AD, 2013).

2.2.1 Decision Trees

Represents a sequence of rules for classification in the form of explicit rules (Brachman, R. J, 1996). Extracting rules from each path in the root to a leaf node makes a decision-tree readable creating a path to each leaf, transformed as IF-THEN rules.

$$I(S_1, \ldots S_m) = -\sum_{i=1}^{m} \log_2(pi)$$

$$p_i = \frac{s_i}{S}$$

where *p* is probability of arbitrary sample of *ci*. Entropy *E*(*Ai*), is partitioned by attribute *Ai*:

$$E(A_i) = \sum_{j=1}^{q} \frac{s_{1j} + s_{2j} + \cdots + s_{mj}}{S} I(s_{1j}, s_{2j}, \cdots, s_{mj})$$

$$I(s_{1j}, s_{2j}, \cdots, s_{mj}) = -\sum_{i=1}^{m} p_{ij} \log_2^{(p_{ij})}$$

2.2.2 Logistic Regression

It is a regression problem approach for Ordinary Least Squares. It predicts dichotomous outcome and estimates the models to predict the odds of its occurrence.

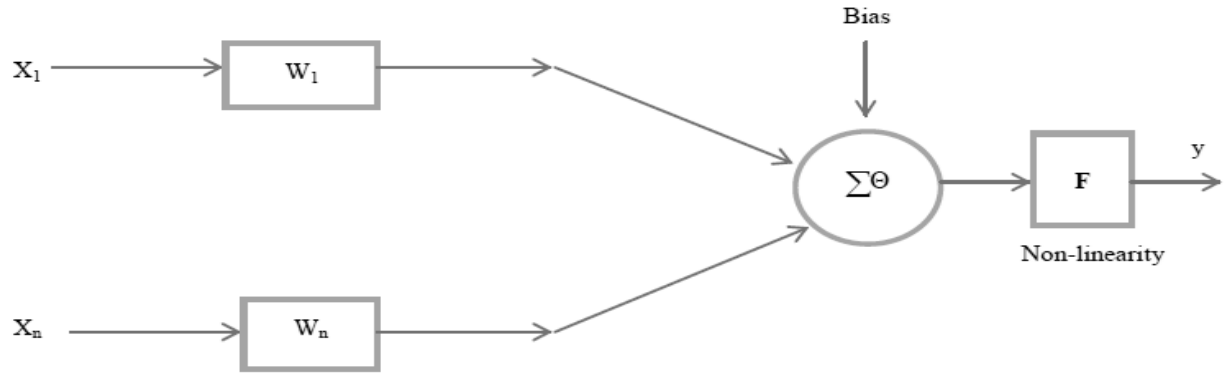


Figure 2: neuron model

2.2.3 Naive Bayes Classifiers

The model is a classification of probability theory to find the likely outcomes. Bayesian calculates the probability *P* (*A B, C*). Conditional probability of *A* and *B* is given by the probability

$$P(A|B,C) = \frac{P(A,B|C)}{P(B|C)}$$

P (*A B, C*)

2.2.4 Artificial Neural Network for classification and regression

Artificial Neural network is an algorithm for reasoning based on human brain. Neural networks models generalizes patterns from training set for prediction and classification task applied also in time-series prediction and undirected data mining (Lu, H., Setiono, R, 1996).

It is applied in both prediction and classification to classify unseen patterns and high tolerance to noise, thus used as predictive data models.

$$y_i = F\left\{\sum_{i=1}^{n} w_i x_i\right\}$$

Index i represent neuron and the popular nonlinear neurons are sigmoid functions. Neural networks test and train by adjusting the weights for selected attributes.

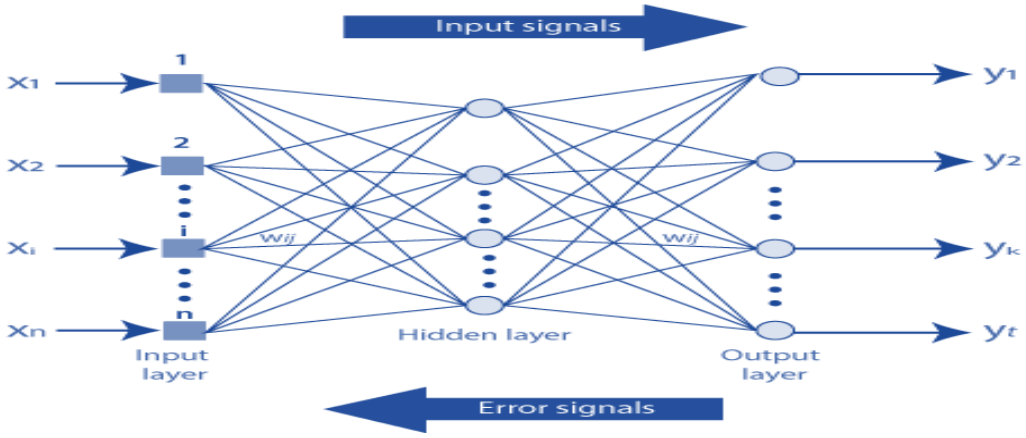


Figure 3: Artificial Neural Networks (ANNs)

2.2.5 Random Forest Algorithm

It utilizes tree-structured classifiers, X represents the input data, n represents distributed random vectors Ali, Khan (2012). The classifier use random sampling to train each decision tree, a process called bagging. Bagging concept use decision trees to develop independent sample of striped down of the dataset (Liaw & Wiener, 2017). The input variables used are patients retained in care, patients starting ART, treatment history and therapy switch at 12 months.

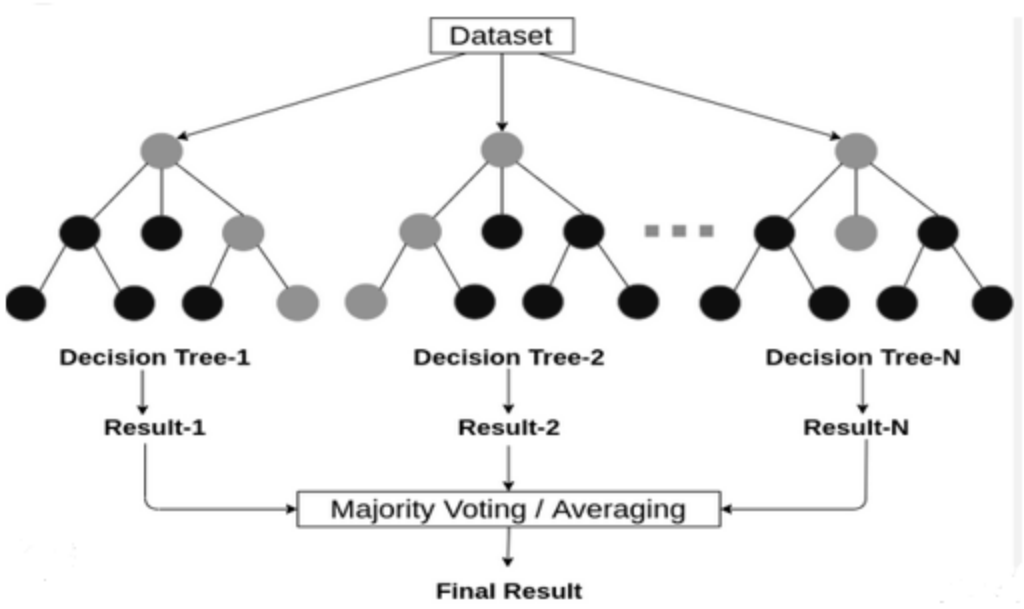


Figure 4: Random Forest Algorithm

2.2.6 K-Nearest Neighbor

KNN is a supervised learning algorithm using training set for regression and classification problem. Neighbors use distance measure which is the Manhattan, Euclidean, or Minkowski to determine K value of neighbors which considers the case of their distance. The K smallest distances obtained and selected as the neighbors are ranked.

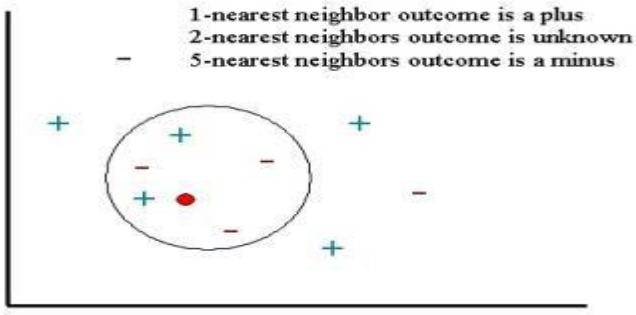


Figure 5: K-Nearest Neighbor classification diagram

The red dot is the point of interest, the distance is determined by the Euclidean distance, k is number of features and x is instance class.

$$D(x_i, x_j) = \sqrt{\sum_{k}^{n=1}((xik - xjk))2}$$

KNN model evaluates k value dimensional vectors for the training sample Shen, et al., (2016).

2.3 Data Mining

Data mining is used by organizations to detect data for insights relevant to their business requirements and needs. The main task in data mining is knowledge discovered from new patterns from the raw data to integrate combined data from multiple sources for decision making. The preprocessed data is stored in data warehouse and selected for analysis purpose. From analyzed and transformed dataset new pattern can be discovered and evaluated. Knowledge representation come in the form of visualization for end users.

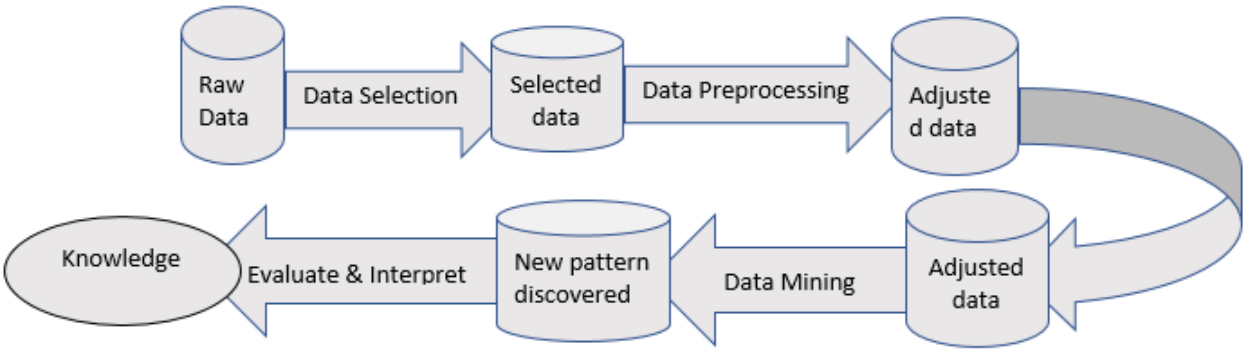


Figure 6: Knowledge discovering process

2.4 Other Authors Related Work and Findings

As prevalence of HIV affects more patients, there remains a gap toward adherence and retention to warrant careful treatment assessment (Grabowski MK, 2017). In a clinical setting, retention in care and ART treatment represents the health efforts to control and identify when patients missing clinical visits for medications on vulnerable periods where viral suppression target health systems investments (Ware NC, 2013). According to (R.S Butt, & I Ahmad, 2019) ART influenced by patients with nonlinear control models and backstepping controls was aligned to improve the efficiency of infected persons using T cells in Pakistan.

Sub-Saharan African countries a research done to distinguish adherence and nonadherence of patients' experiencing failure and results show that 72.9% of patients were adhered to ART (T. Heestermans, 2016). A study conducted in Zambia researchers proposed allocation of budget to sustain long-term survival and collaborative research done in Malawi, South Africa, and Zambia researchers implemented stochastic simulation model to monitor the effect viral load to address the mortality rates increasing in these countries.

In South Coast of Africa, patient's data was used to monitor and mitigate risks and reduce the problems using SMS reminders and mobile health apps to notify the HIV+ mothers for prevention(J. Davey, 2012), using this approach reported cases reduced and children were saved. Bootstrapping was used to correlate different parameters and confidence was considered in the interpretation of clinical proofs. ART drug through blood plasma technique was given to patients and evaluated with data to improve patients health in USA(R. Baraldi, K. Cross,2014). Data mining technique with ART therapy shows the performance of 80.5% prediction rate, ART predictive model achieved 66% results using hospital data in Ethiopia. A research done by (T. D Chala, 2019), presents knowledge Discovery in Databases (KDD) model in patients in details. A study done using Zambian population-based HIV Impact suggests that self-reporting ART use and viral suppression 90% of people thus potentially overestimating suppression (Barradas DT, 2017). Researchers examined viral suppression and retention in 4 provinces in Zambia with 1,200,000 adults, a sampling-based approach was used to select facilities with probability proportional to facility size. Predictive model estimated viral suppression and overall retention in Zambia with a precision to assess site-to-site variation (Geng EH, 2010).

2.4.1 Baseline Medication adherence.

Adherence is assessed and self-reported using case adherence index (CAI) measuring participants' on medications, classified as poor adherence or good adherence (Mannheimer SB, 2006). A category is reflected to patients' not on medication and three-category variable predicts viral suppression.

2.4.2 Statistical and Predictor selection.

According to (Hanley JA, 2003) adjusting the correlation by intervention examines the relationship of selected predictors with study outcomes viral suppression.

GEE use an exchangeable correlation matrix structure and binomial distribution for binary outcomes with estimation of parameters. Descriptive statistics generates variables, using median, standard deviations and confidence for frequencies and percentages, and quantitative variables for categorical variables.

2.5 Research Gap

According to literature review ART perceptions and intention using cross-sectional designs suggest lack of investing on ART perceptions in predicting ART behaviors in the implementation of new ART strategy (Yang et al, 2021).

Predictive models using virological resistance genotype data is not easily available. Genotype data using complex classifiers such as neural networks (ANN) are not easily interpretable by healthcare providers and where computing facilities may not be available (Revell AD, 2013). Random forest and Decision trees are popular with clinicians due to its ease of application. The research assesses the performance of supervised learning classification-based methods to predict retention in care in HIV patients switching therapy.

This research aims at incorporating ensemble methods and adaptive boosting to predict an unseen sample to correlate with an output variable to evaluate predictive performance of the model.

2.6 Conceptual Design

Conceptual design identify the variables required to explain a phenomenon in the research investigation mapping the actions required to realize the research objectives (Regoniel, 2016).

In this section, the conceptual framework for the research is presented for retention and viral suppression for HIV treatment cohorts. Various definitions of performance and scalability is then discussed and tested.

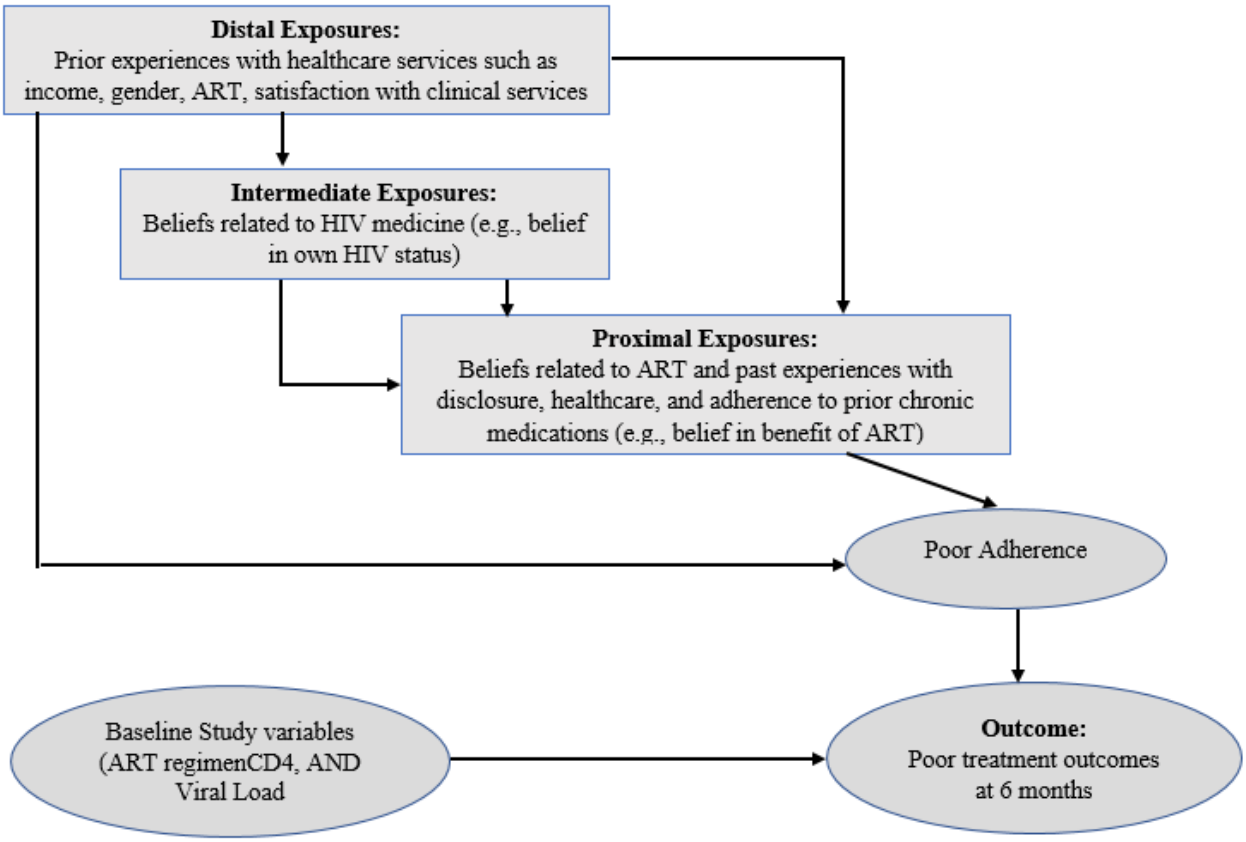


Figure 7: Hierarchical framework.

2.6.1 Operationalization of variables

Independent Variables

Factors influencing adherence to ART and care drugs are categorized as follows; Enabling Resources such as health worker attitudes, financial support, social factors, access to ART, Poor documentation, defaulter tracing, Occupation status, transfer in status, tracing status, location of ART management, time on ART , Predisposing Characteristics i.e., lack of acceptance and knowledge, Custom beliefs, stigma and fears, Age at initiation, gender, marital status, occupation type, religion. Need Factors; Initial weight, WHO stage at ART initiation, baseline CD4, weight at follow-up, side effects, Tuberculosis, ART continuation decision.

2.6.2 Intervening Factors

Mediating variables are hypothetical variables which can't be observed in an experiment but used to explain causal links in other variables. They do not cause the outcome but may modify it. In this study they are broadly categorized into; Health behavioral factors (adherence patterns, service appointment and missed appointments/defaulting ART drugs). Environmental factors (regional, national and district policies, societal norms, national ART guidelines).

2.6.3 Dependent Variables

These are the factors being affected by the independent variables, adherence of PLHIVs to ART and therefore determining their morbidity (improved immunity), staying alive longer or death.

2.7 Conceptual Model

Dataset used was obtained from facility sampling unit and was cleaned for training using data mining technique and machine learning algorithms. Dataset attributes were extracted, including the phenotype, ART regimen, patients enrolled in care, ART status, first-line and second-line ART to transform attribute construction, aggregation, and normalization.

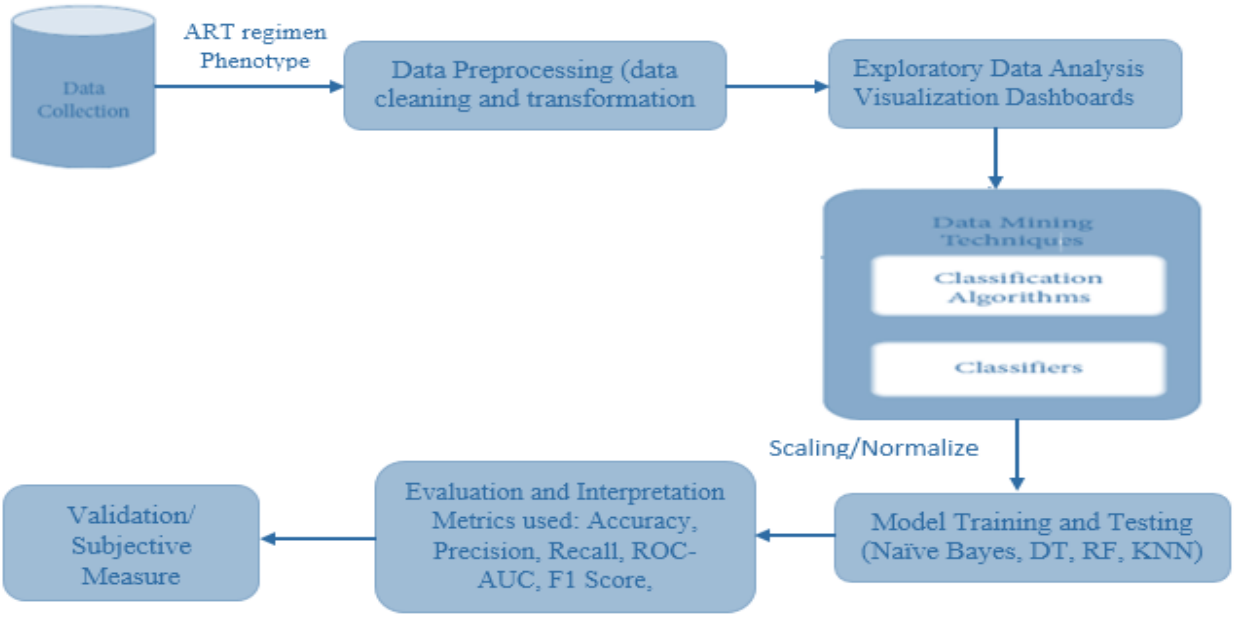


Figure 8: Conceptual Model of research work.

CHAPTER THREE:

RESEARCH METHODOLOGY

3.0 Introduction

The chapter analyzes the data from the Ministry of Health (MOH) by extracting features and establishing the predictive model for retention and viral suppression for cohort datasets. The section covered the research design paradigm used, the analysis methods and tools used to achieve the objectives.

3.1 Research Design

Research design analyzes data and measurement to obtain answers to the research questions (Wanjugu, 2015). The study used quantitative research design with emphasis in the objective measurements, statistical and numerical analysis of data, and aggregation of cohort dataset which is in numeric form. The entire process of this study and methodology was guided by (CRISP-DM). Research design served to outline the work done and described how the results was accomplished to identified research objectives. The following phases were taken to achieve the intended results.

1. Evaluating the performance of machine learning model for easy mapping coverage of treatment cohorts.

2. Monitor and assess treatment outcomes of patients switching from first to second line ART to achieve viral suppression.

3. Reviewing experimental results and architectures applicable for effective monitoring of patients' cohort data.

3.2 Overview of CRISP-DM

The researcher adopted CRISP-DM data mining model in the software development lifecycle. We chose this model due to its high level of flexibility and ability to perform regular iterations. CRISP-DM is intuitive for industry-specific-applications and easy to read documentation.

The discussion of the proposed methodology and research objectives is broken down into the following tasks.

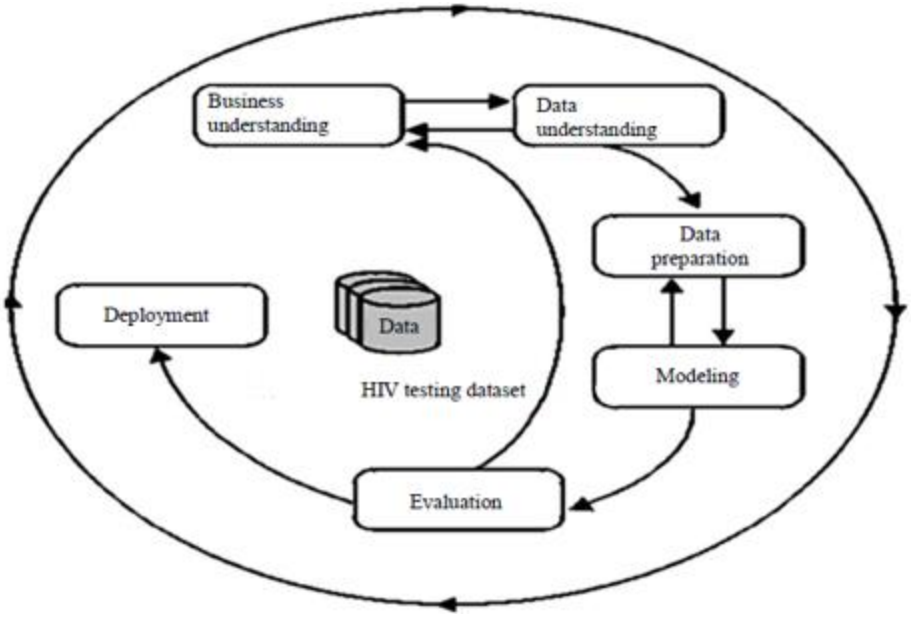


Figure 9. The CRISP-DM process model.

3.2.1 Business Understanding

In this phase all detailed business requirements of artifact to be developed were gathered. Understanding of the business was to identify which problems a business had. This was critically important for patients to understand about their ART regimen progress and make informed decisions improving availability for provider-initiated testing and counseling in health facilities.

3.2.2 Data Understanding

The overall goal was to ascertain, assemble and scrutinize the data sets that helped the researcher achieve the project goals. Attributes were identified for data mining purpose including subsets to form hypotheses for hidden information. Data mining including data cleaning, attribute selection and transformation were performed and the selection of a data set and analysis of its attributes and structure. The data was cleaned and prepared for modeling. We did data analysis just to get familiar with the data and features of the dataset.

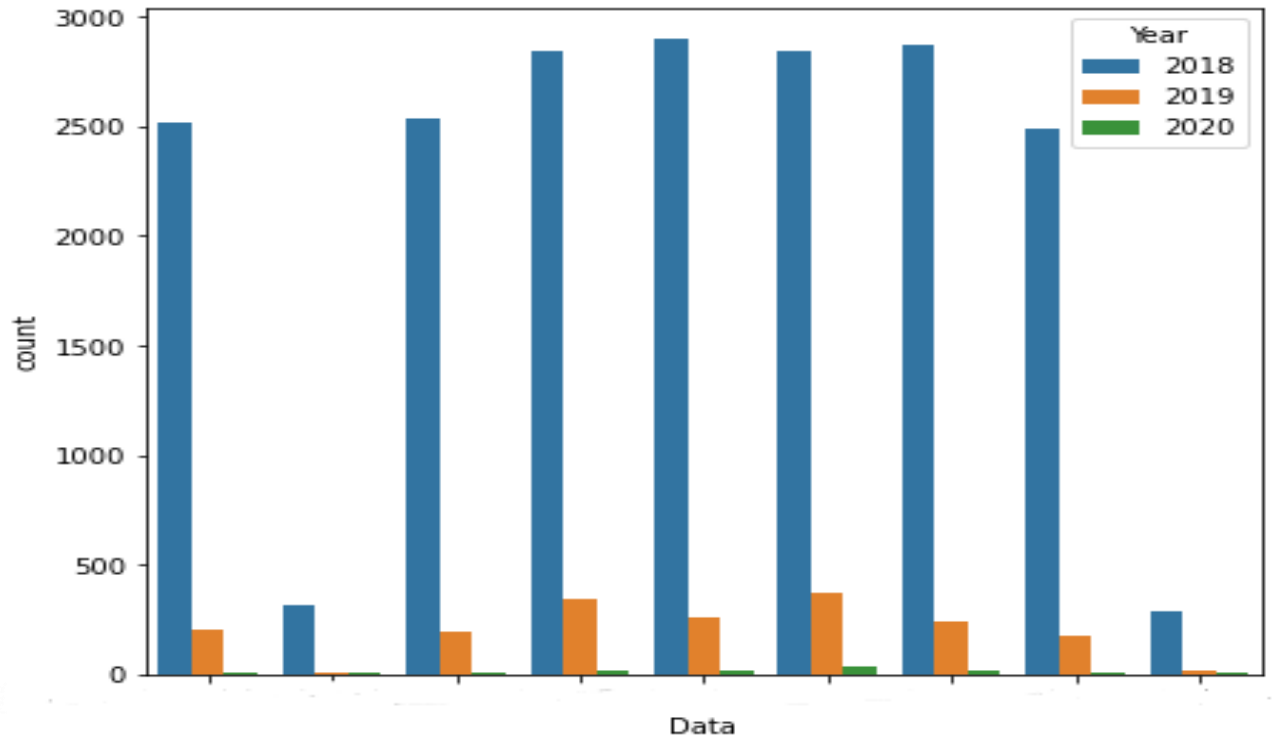


Figure 10: Study Sample of ART cohort data from 2018 to 2020

In figure 10 we can note that we had more patients hospitalized and retained for ART treatment in 2018 as compared to 2019 and 2020.

```
In [119]: dfCountySum.plot(kind="bar", figsize=(15,6))
Out[119]: <AxesSubplot:xlabel='County'>
```

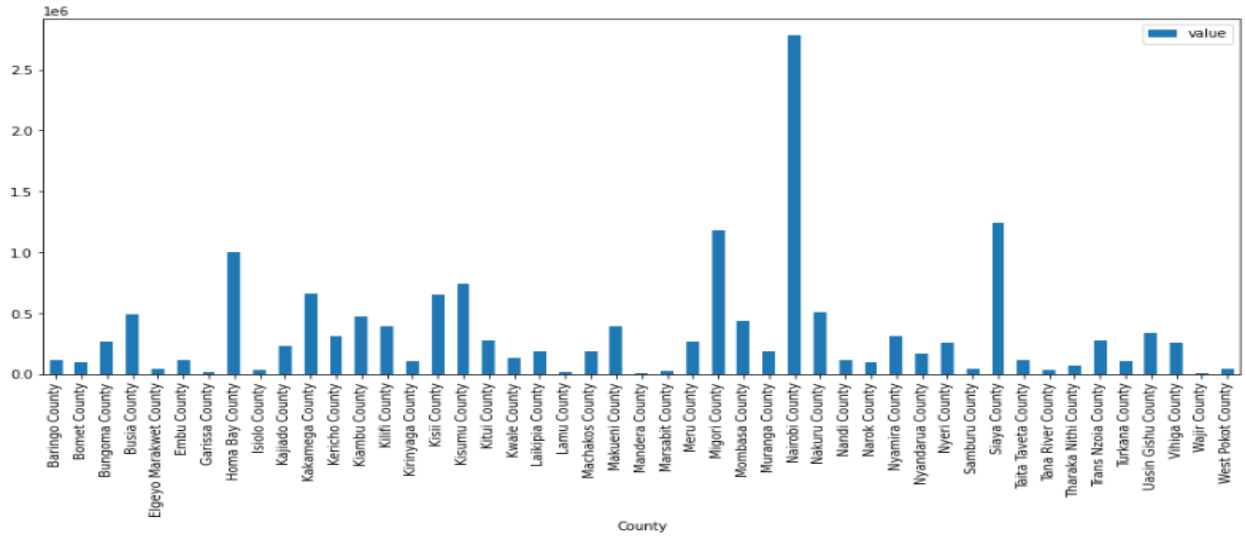


Figure 11: Plot showing patients enrolled in County health facilities

In figure 11, Nairobi County was the highest of patients seeking and enrolled for ART treatment followed by South Nyanza Counties including Siaya County, Migori County, Homa Bay County, Kisumu County and Kisii County across the Country.

Both counties show increase in ART use with equitable access to ART across socio-economic and geographic status. The evidence of disparities for ART in HIV clinics, and lingering inequities indicating some populations may require increased adherence support due to eligibility criteria.

3.2.3 Data Preparation

The data preparation tasks included data cleaning, integration and transformation, we performed the below steps: The tasks included were:

1. Data cleaning: The data collected was classified into HIV cohort 2018, 2019 and 2020 clinical variables for antiretroviral treatment. A python function was used to create the labels for the dataset.
2. The dataset containing missing columns was dropped and filled with missing data (NaN) values with median for continuous variable and 0/1 for discrete variables. The remaining final data had 21551 entries which were sufficient for training.
3. Data selection: The categorical columns were converted to one-hot encoding, label encoding based on the categories available.
4. Integrate and format data: data was reformatted by converting string values that stored numbers to numeric values to perform mathematical operations.

3.2.4 Modeling Phase

In the Modeling phase, a modeling technique was selected and applied to the dataset. The created model was then assessed. This phase not only evaluated the developed model but also the development process. This phase has the following tasks:

1. Selecting modelling techniques: The researcher chose the supervised learning classification models such as Naïve Bayes, Decision Trees, Random Forest Classifier and K-Nearest Neighbors algorithm to execute the tasks.
2. Generating test design: For the modeling approach, the researcher splits the data into testing, training and validation sets.

3.2.5. Evaluation

In this phase, we checked different machine learning models performance. This criteria was used based on performance metrics. This included: Confusion Matrix, F1 Score, ROC-Area Under Curve (ROC-AUC), Precision, Recall, Accuracy, Sensitivity and Specificity. The evaluation phase has the following tasks:

1. Evaluating the results and checking the impact for data mining goals according to business success criteria, reviewing and identifying failures and possible alternative actions for unexpected paths.

3.2.6. Deployment

The final phase had the following tasks.

1. Monitoring and maintenance: The researcher developed monitoring plan to avoid the post-project phase and operational phase of a model, the researcher documented the project summary of a final presentation of data mining results.
2. Research conducted a review of project retrospective about what went well and how to improve in the future.

3.3 Study design and data collection

The research evaluated cases of patients who were retained in care for ART at various county health facilities in the year of 2018 to 2020. Significant dataset was collected from ART Center, facility sampling unit MOH; that was, 21551 health facilities records were collected.

Required parameters like patients enrolled in care, patients currently on ART, those starting ART, patient weight, patients Max Z-Scores and outlier weight and patient ART regimen are collected from the repository. Table 1 below summarizes the cohort dataset.

	Cohort Dataset (2018, 2019, 2020)
Data ID	9
Data	9
County	47
Sub-County	301
County Assembly Ward	1207
Orgunit ID	3241
Orgunit Name	3241

Table 1: Features of cohort dataset 2018, 2019 and 2020

3.3.1 Data source and study population

This is a longitudinal data for patients enrolled and retained in care for ART from January 2018 to December 2020 of patient cases recorded at County health facilities spread throughout the Country. The examined cases include duration from the time patients were enrolled in HIV care, enrollment to eligibility for ART and time from eligibility to initiation of ART.

The following predictors including clinical information and antiretroviral treatment records and other HIV key indicators were used. ART therapy at 12 months, anonymized patient who had accessed HIV care and, during the period 2018-2020, who had recorded/started their ART clinic visit.

```
In [113]: x=range(1,37)
          plt.rcParams['figure.figsize']=(15,6)
          plot(x,df4["value"],lineLabel=r"Month vs ART_Treatment History", title="Baringo 2018-2020",  xLabel="month",yLabel="value")
          plt.show()
          plt.rcParams["figure.figsize"] = plt.rcParamsDefault["figure.figsize"]
```

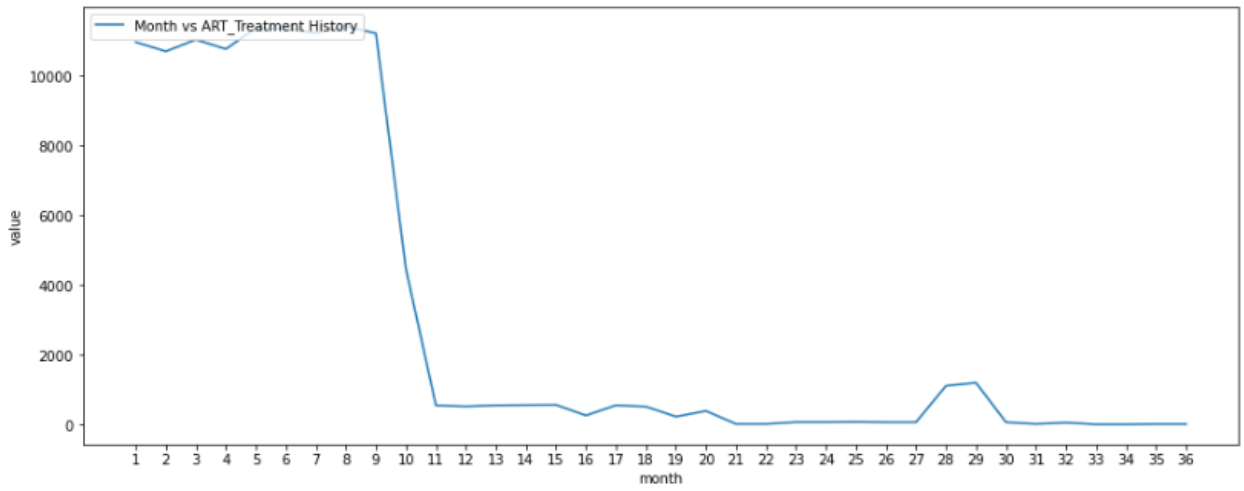


Figure 12: Graph showing antiretroviral treatment records.

3.4 Data Analysis

Exploratory data analysis from the time patient were enrolled to ART start and effects on survival were used to identify patients currently starting on ART for targeted mapping outreach. For qualitative and descriptive statistics, percentages and number were used to provide categorical variables. The study presented exploratory analysis to provide the intuition for the dataset in selecting and interpreting qualitative features to have global view of the data. The research analyzed the frequency of features and correlation between the different key features of ART treatment. This was presented using a heatmap as shown below.

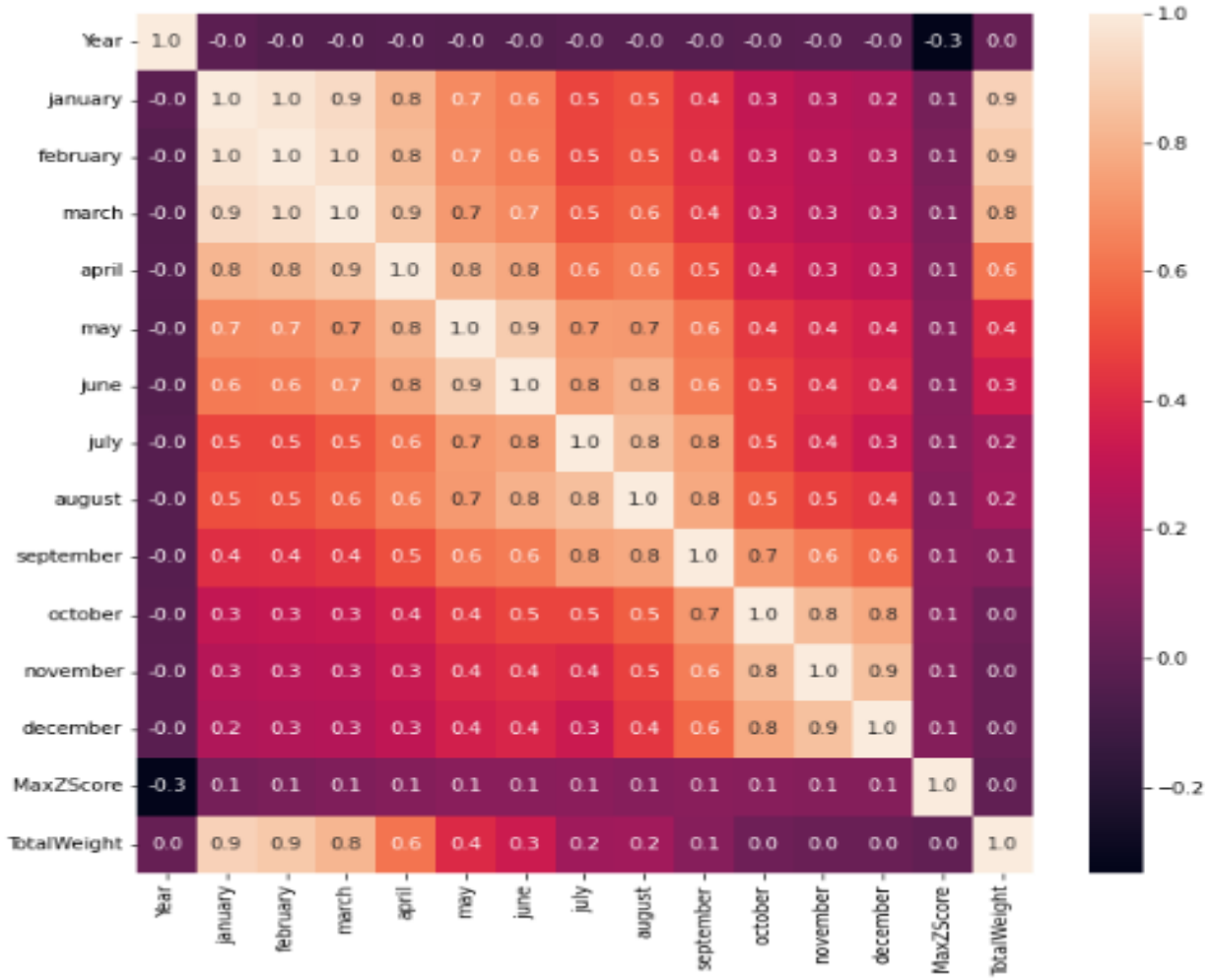


Figure 13: Features correlation heatmap

We used correlation heatmap in order to get the correlation coefficient against the label i.e. YEAR, we can deduce that most of the features have very low correlation, -1 and 1. 0 implied weaker correlation, a value of 0 implied no correlation, 1 implies positive correlation and a value closer to -1 implied stronger negative correlation. The largest correlation visible was YEAR. We created a histogram of frequency count of hospitalized_arv. Hospitalized_arv is a type of patients who are hospitalized under antiretroviral therapy).

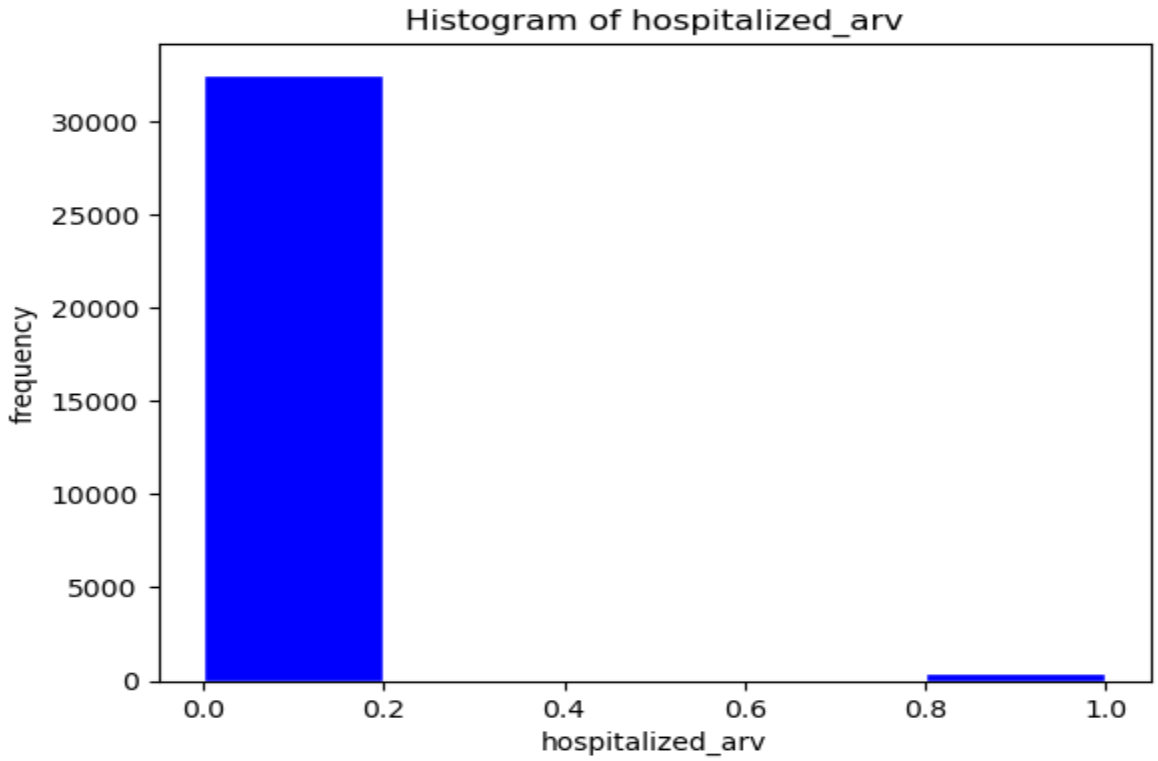


Figure 14: Data Visualization using Histogram

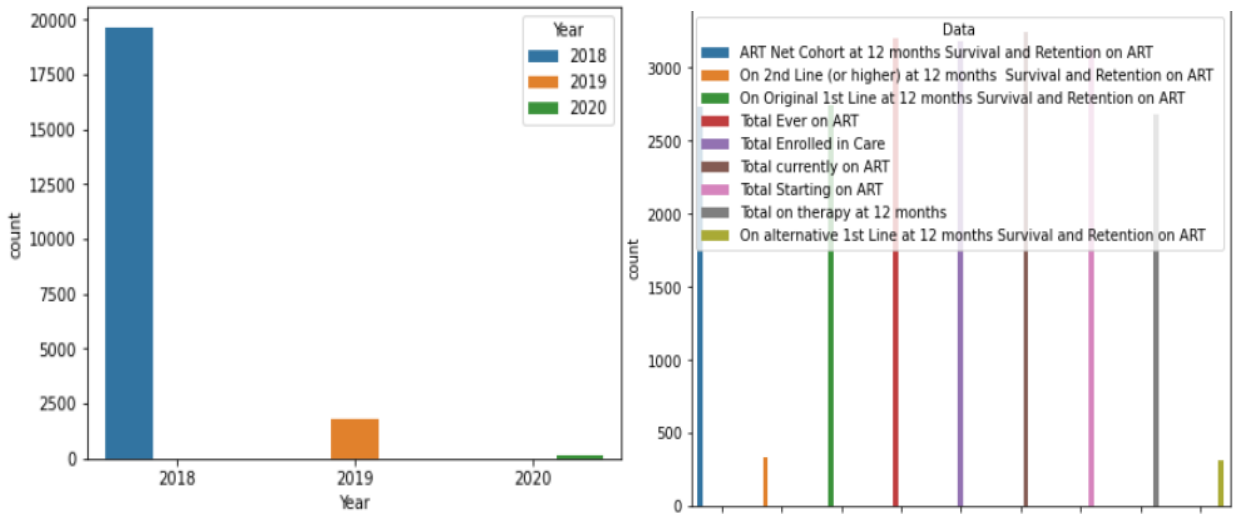


Figure 15: Data Analysis

ART enrollment increased significantly in both Counties, consistent with ART program scale-up. Proportion of patients on ART was higher in Nairobi County than other counties in chort years of 2018, 2019 and 2020. Nairobi County had the highest number of patients taking ART regimen to other counties, and the proportion of patients on ART nearly doubled over this period.

3.4.1 Power BI Client Integration Dashboard

We performed Exploratory Data Analysis (EDA) and created entities in PowerBI's visuals interactive dashboard for analytics report using Power BI client integration to review and validate reports allowing clinicians to gain insights from the data for visualization.

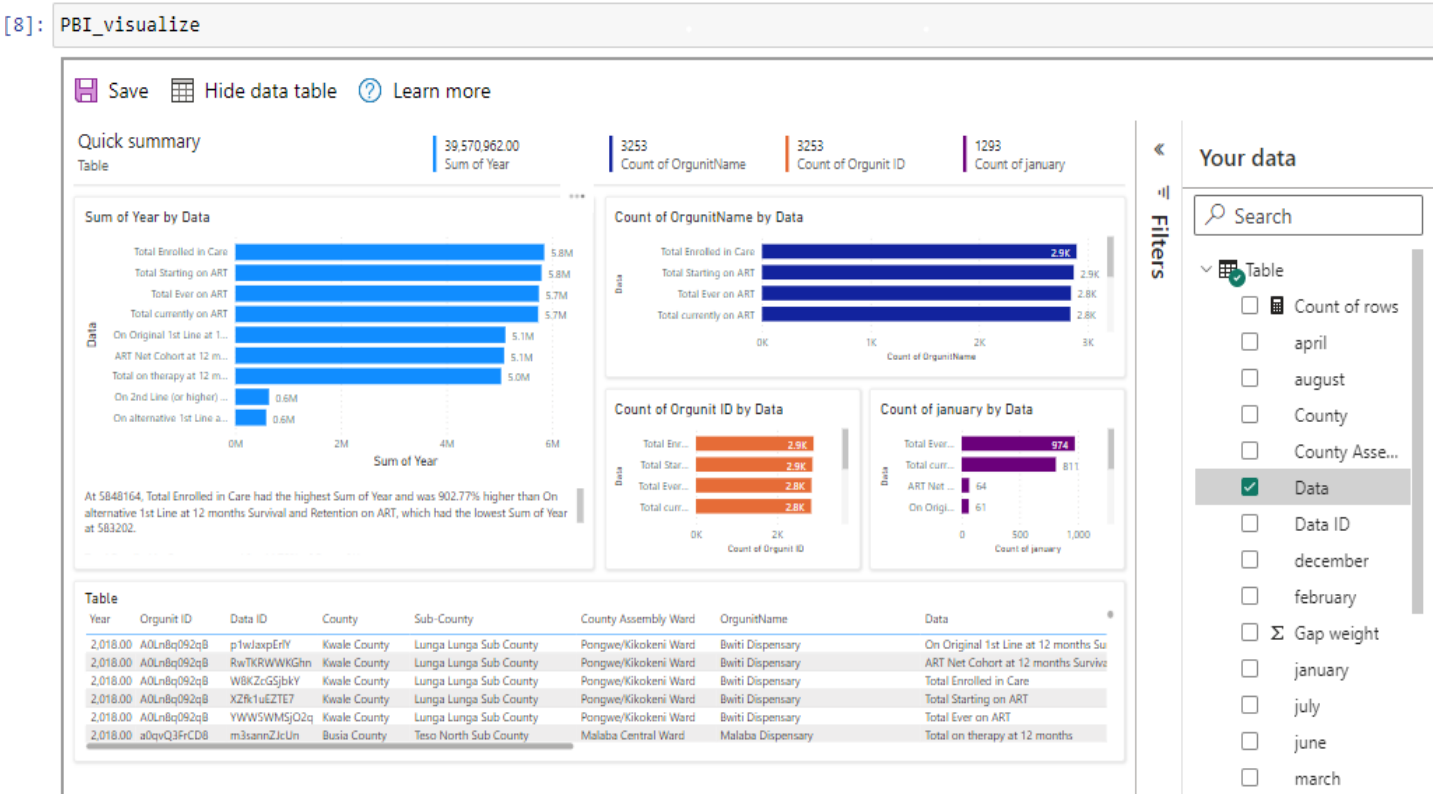


Figure 16: Power BI client integration report

Patients hospitalized and enrolled for ART care had the highest sum of the year and was 90.2% higher than on alternative 1st line at 12 months survival and retention on ART which had the lowest sum of the year at 58.32%.

3.4.2 Data Cleaning

We cleaned the datasets collected, removing duplicates, white spaces and highlighting errors, filling in missing data which improved the quality of the dataset. Columns were dropped that had large number of missing entries.

We filled the missing data with the **median** values for the continuous variable and 0/1 for discrete variables and replaced missing values with "NaN", while performing mathematical imputations.

	Year	Orgunit ID	Data ID	County	Sub-County	County Assembly Ward	OrgunitName	Data	january	february	...	august	september	october	november
0	2018	jqrGlRD2MbV	RwTKRWWKGhn	Mandera County	Mandera East Sub County	Neboi Ward	Mandera County Refferal Hospital	ART Net Cohort at 12 months Survival and Reten...	4.0	NaN	...	NaN	NaN	NaN	NaN
1	2018	MmojLQkJWtM	pHCaA5yineM	Mandera County	Banissa Sub County	Banissa Ward	Banisa sub county Hospital	On 2nd Line (or higher) at 12 months Survival...	NaN	NaN	...	NaN	NaN	NaN	NaN
2	2018	jqrGlRD2MbV	p1wJaxpErlY	Mandera County	Mandera East Sub County	Neboi Ward	Mandera County Refferal Hospital	On Original 1st Line at 12 months Survival and...	4.0	NaN	...	NaN	NaN	NaN	NaN
3	2018	MmojLQkJWtM	YWW5WMSjO2q	Mandera County	Banissa Sub County	Banissa Ward	Banisa sub county Hospital	Total Ever on ART	13.0	NaN	...	12.0	11.0	13.0	12.0
4	2018	cpmUjkUfMqG	YWW5WMSjO2q	Mandera County	Mandera South Sub County	Elwak North Ward	Elwak District Hospital	Total Ever on ART	46.0	NaN	...	NaN	NaN	NaN	NaN

```
df[2:13].fillna(0, inplace=True)
```

```
####filling NaN values with median for continous variables and 0/1 for discrete variables.

df['Year'] = df['Year'].fillna(df['Year'].median())
df['january'] = df['january'].fillna(df['january'].median())
df['february'] = df['february'].fillna(df['february'].median())
df['march'] = df['march'].fillna(df['march'].median())
df['april'] = df['april'].fillna(1)
df['may'] = df['may'].fillna(df['may'].median())
df['june'] = df['june'].fillna(df['june'].median())
df['july'] = df['july'].fillna(1)
df['august'] = df['august'].fillna(df['august'].median())
df['september'] = df['september'].fillna(0)
df['october'] = df['october'].fillna(0)
df['november'] = df['november'].fillna(1)
df['december'] = df['december'].fillna(1)
df['County'] = df['County'].fillna(0)
df['Data'] = df['Data'].fillna(0)
```

Figure 17: Filling missing values

3.4.3 Data Transformation

The dataset was transformed into a format that ML algorithm can understand, features were normalized using Scikit learn's preprocessing class called StandardScaler. The encoded categorical data added dummy variables for the categorical data to make the data more readable and understandable. We normalized the data using the StandardScaler () class by Scikit learn library between a scale of 0 and 1. This is to ensure that outliers do not skew the model.

Out[584]:

	Year	County	january	february	march	april	may	june	july	august	...	october	november	december	MaxZScore
0	0	23	93	785	748	1236	1	2782	4340	4352	...	6527	7930	0	0
1	0	23	93	785	748	1236	1	2782	4340	4352	...	6527	7930	0	0
2	0	23	93	785	748	1236	1	2782	4340	4352	...	6527	7930	4937	0
3	0	23	262	785	748	1236	1	2782	4340	4352	...	6527	7930	7975	110
4	0	23	425	785	716	1183	1	2768	4182	4217	...	6463	7930	7975	0

Figure 18: Data Transformation

3.4.4 Feature Engineering

The categorical variables were dummified using label encoder and one-hot encoding. We investigated/evaluated cardinality and number of labels within the categorical variable. We dropped categorical columns with high cardinality since they pose serious problems in the machine learning model such as space consumption and curse of dimensionality.

```
Out[59]: [('Data ID', 9),
          ('Data', 9),
          ('County', 47),
          ('Sub-County', 301),
          ('County Assembly Ward', 1207),
          ('Orgunit ID', 3241),
          ('OrgunitName', 3241)]

In [60]: d.items()

Out[60]: dict_items([('Orgunit ID', 3241), ('Data ID', 9), ('County', 47), ('Sub-County', 301), ('County Assembly Ward', 1207), ('Orguni
         tName', 3241), ('Data', 9)])
```

```
# Columns that will be dropped from the dataset
high_cardinality_cols = list(set(object_cols)-set(low_cardinality_cols))

print('Categorical columns that will be one-hot encoded:', low_cardinality_cols)
print('\nCategorical columns that will be dropped from the dataset:', high_cardinality_cols)
```

```
Categorical columns that will be one-hot encoded: ['Data ID', 'County', 'Data']

Categorical columns that will be dropped from the dataset: ['OrgunitName', 'County Assembly Ward', 'Sub-County', 'Orgunit ID']
```

Figure 19: Investigate cardinality and create an instance One-hot-encoder

3.4.5 Binary Classification Problem

Predictive analytics for retention in care is a classification problem where models classifies clinical features and predicts which class the features fall. In this case, the quality criteria for splitting the classification techniques is to formalize partitions in a tree using entropy.

Other heuristic is to use Gini uncertainty(Gini impurity) to maximize the criterion and interpret the maximization of the number of clinical outcomes of the same class that are in the same subtree.

We create a tree with the sample (Gini uncertainty and information gain).

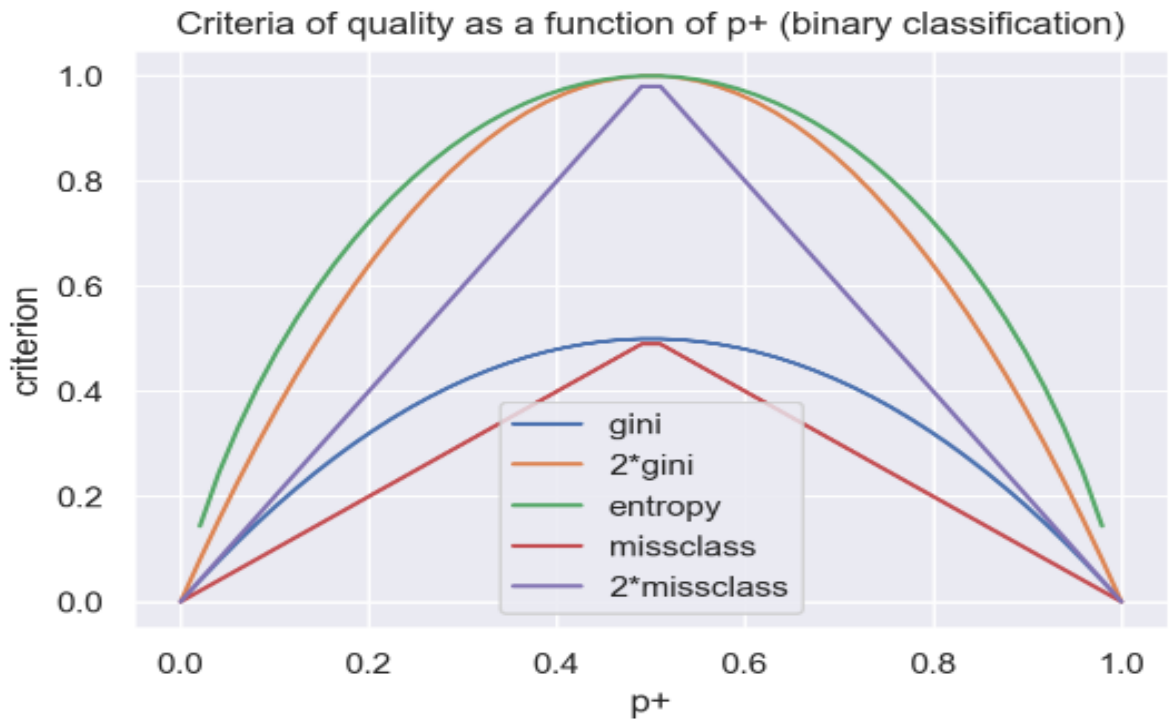


Figure 20: Binary classification using decision tree

3.5 Ethical Considerations

The research applied ethical guidelines for developing predictive retention and viral suppression system among HIV cohorts. For data handling – the researcher ensured protection of privacy avoiding collecting identifiable data. The research used publicly available cohort dataset after obtaining consent and rights from the ministry of health. The research disclosed the use of the data as per the scope of research and was granted access to develop the model, ensuring that the data given only contained aliases that could not be linked to any particular patient, and the data was only used for the purpose of this research.

CHAPTER FOUR:

RESEARCH FINDINGS AND DISCUSSIONS

4.1 Introduction

The objective of this research was to develop a model for prediction of ART adherence status for community-based testing treatment program in County HIV clinics. We aim to identify predictors of aggregated data using the following key outcomes; patients enrolled in care, those currently on ART, patients starting on ART and switching therapy.

4.2 Evaluation Results and Findings

Identifying patients currently enrolled in care and starting on ART represent a classification algorithms over a hyper-parameter grid used to develop the models.

The performance metric was evaluated using accuracy for unseen test set to the total number of input samples, sensitivity which classifies positive outcomes, precision represent the proportion of positive outcomes predicted in the unseen test set and specificity is the proportion of negative outcomes in the unseen test set (Maskew M, & Sharpey-Schafer K, 2010). Area under the curve (AUC) and (ROC) evaluates classification performance of the model. F1_score performance measure since it takes both recall and precision into consideration.

4.2.1 Model building

Total of 21551 predictor features were investigated including features scaling relating to cases/patient currently on ART, patients ever on ART, patients enrolled in care), hospitalized under ART, those starting on ART and first line and second line ART as well as patient testing (e.g. MaxZScore and Total Weight) among others. The model was evaluated for prevalence of each outcome % of original and alternative 1st line and 2nd line at 12 months survival and retention on ART and sampled on a 80/20, that is (80%) training and (20%) test set for visits classified as enrolled in care and on ART Net cohort at 12 months survival and negatives for patients currently on ART and starting on ART. Classifier algorithm was trained by predictor features and specified target outcomes to produce optimal configuration.

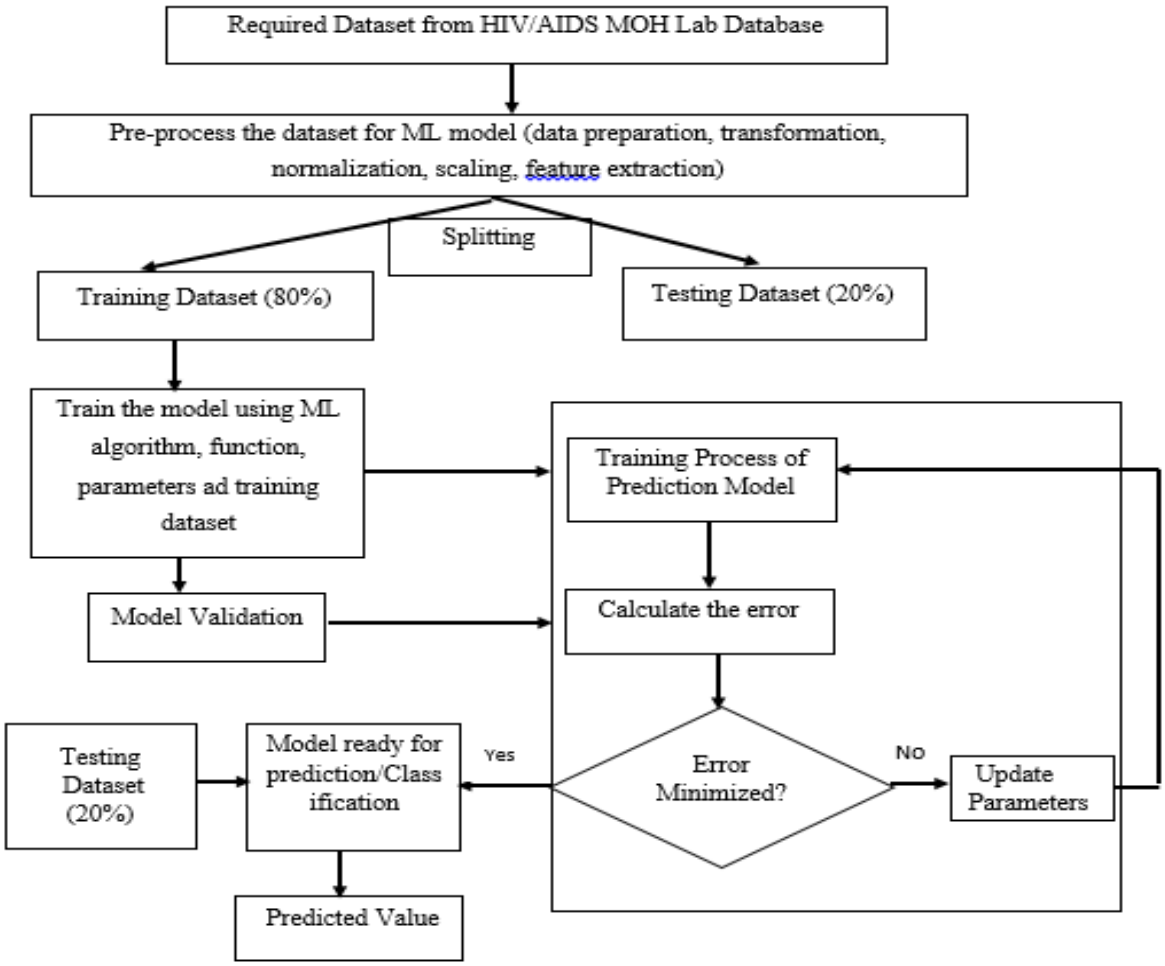


Figure 21: Steps involved in building predictive Analytics model

4.2.2 Classification Accuracy

Random forest performed better than other binary classifier models because it has bagging approach for randomly composed decision trees whose results are aggregated for the cohort datasets and the random nature limits overfitting making them attractive modelling tools for complex hyperspaces with non-linear separations in classes.

Model	Accuracy_Score (%)
Naive Bayes (NB)	20%
Decision Trees (DT)	68%
Random Forest Classfier (RF)	81%
K Nearest Neighbor (KNN)	71%

Table 2: Accuracy score for cohort datasets

4.2.3 Confusion Matrix

True Positives occurred when prediction of observation belonging to a certain class was collected predicted for correctly classified predictions. We predicted True Negatives and False Positives observations.

```
Confusion matrix

 [[ 36   0   0 ...   0   0   0]
 [  0  81   0 ...   0   0   0]
 [  0   0 249 ...   0   0   0]
 ...
 [  0   0   0 ...  64   0   0]
 [  0   0   0 ...   0   8   0]
 [  0   0   0 ...   0   0  25]]

True Positives(TP) =  36

True Negatives(TN) =  81

False Positives(FP) =  0

False Negatives(FN) =  0
```

Figure 22: Confusion matrix

		Predicted Value	
		Retention on ART	Virally Suppressed
Actual Values	Retention on ART	36 (TP)	0 (FN)
	On viral suppression	0 (FP)	81 (TN)

Table 3: Confusion matrix of predictive model on cohort testing dataset

4.2.4 F1-Score

F1 score is the weighted average of the precision and recall where its best value is at 1.0 and worst value is at 0.0. F1-score is lower than accuracy measures as they embed precision and recall into their performance. We used ensemble random forest and other binary classifier method for predictive factors of retention on ART, ensemble random forest delivered better results of 79%.

Model	F1-Score (%)
Decision Trees (DT)	68%
Naïve Bayes (NB)	20%
Random Forest (RF)	79%
K Nearest Neighbor (KNN)	42%

Table 4: F1-Score for classification model performance

4.2.5 Classification metrices

The research used classification report to evaluate the model performance. Precision is correctly predicted positive outcomes belonging to a particular ratio of true positives (TP). A model's sensitivity is correctly predicted positive ratio observations and cannot be influenced by the uneven class distribution (C. Goutte, 2005).

The classification algorithms was trained with unbalanced 80:20 sample of 15085 patients cases of retention on ART, the algorithms classified 6466 of the test set with accuracy score of 55%. Total cases of patients enrolled on ART correctly identified performing a higher recall of 100% score all positives.

Classification Report	Score (%)
Precision (Specificity)	100%
Recall (Sensitivity)	100%
True Positive Rate	95%
False Positive Rate	0%

Table 5: Classification report for precision and recall

4.2.6 Cross Validation, Selection and Model Training

Classification performance of machine learning algorithms was tested for classifiers such as trees, ensemble trees, Euclidean distance, and Gaussian distribution. Cross-validation was performed to identify patterns and correlation in the data for workflow in the deployment.

This allowed models developed to account ART treatment outcomes occurring before 12 months of prediction and tested on treatment. The baseline performance metric was evaluated using accuracy and positive predictive value.

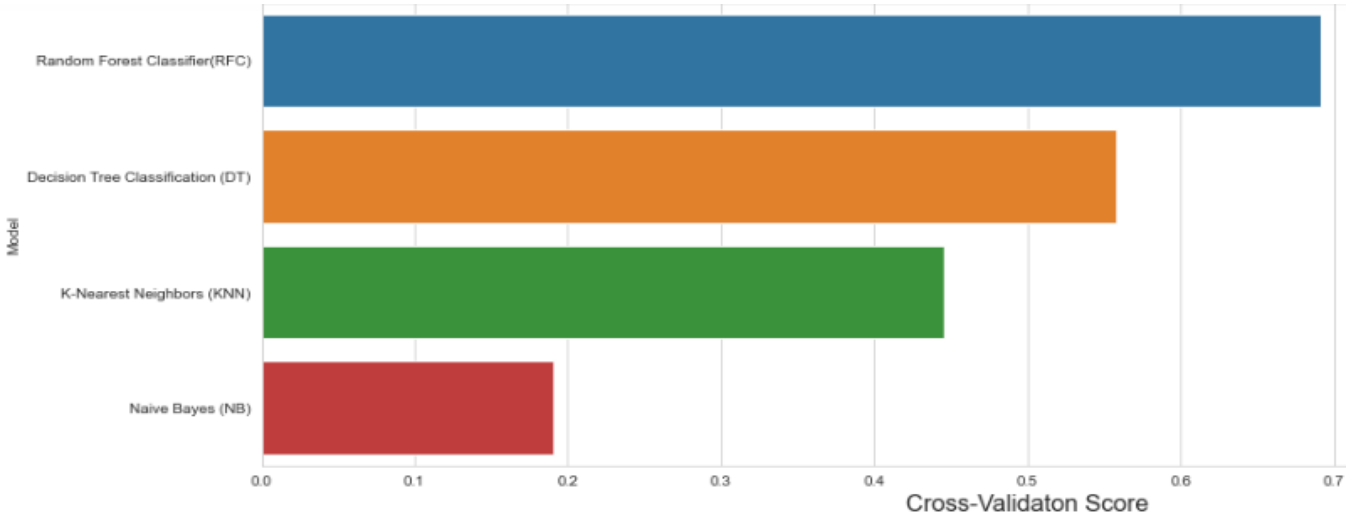
```
predict = pd.DataFrame(data = models, columns=['Model', 'True Positive', 'False Positive', 'True Negative',
                                               'False Negative', 'Accuracy(training)', 'Accuracy(test)',
                                               'Cross-Validation'])
predict
```

	Model	True Positive	False Positive	True Negative	False Negative	Accuracy(training)	Accuracy(test)	Cross-Validation
0	K-Nearest Neighbors (KNN)	8	3	53	1	1.000000	0.433310	0.445940
1	Naive Bayes (NB)	0	1	87	0	0.192053	0.180005	0.190603
2	Decision Tree Classification (DT)	9	0	88	0	0.610035	0.555788	0.558411
3	Random Forest Classifier(RFC)	13	0	88	0	0.770940	0.689863	0.690893

Model	Training (Accuracy %)	Test (Accuracy %)	Cross–Validation (%)
K Nearest Neighbor (KNN)	100%	43%	45%
Naïve Bayes (NB)	20%	18%	19%
Decision Tree (DT)	61%	56%	56%
Random Forest (RF)	77%	69%	69%

Table 6: Accuracy score of training/test set and cross-validation

For baseline performance of a classifier we assessed error rate in the formation of the classifier, dividing the dataset into 10 parts to represent same proportions as in the full dataset.



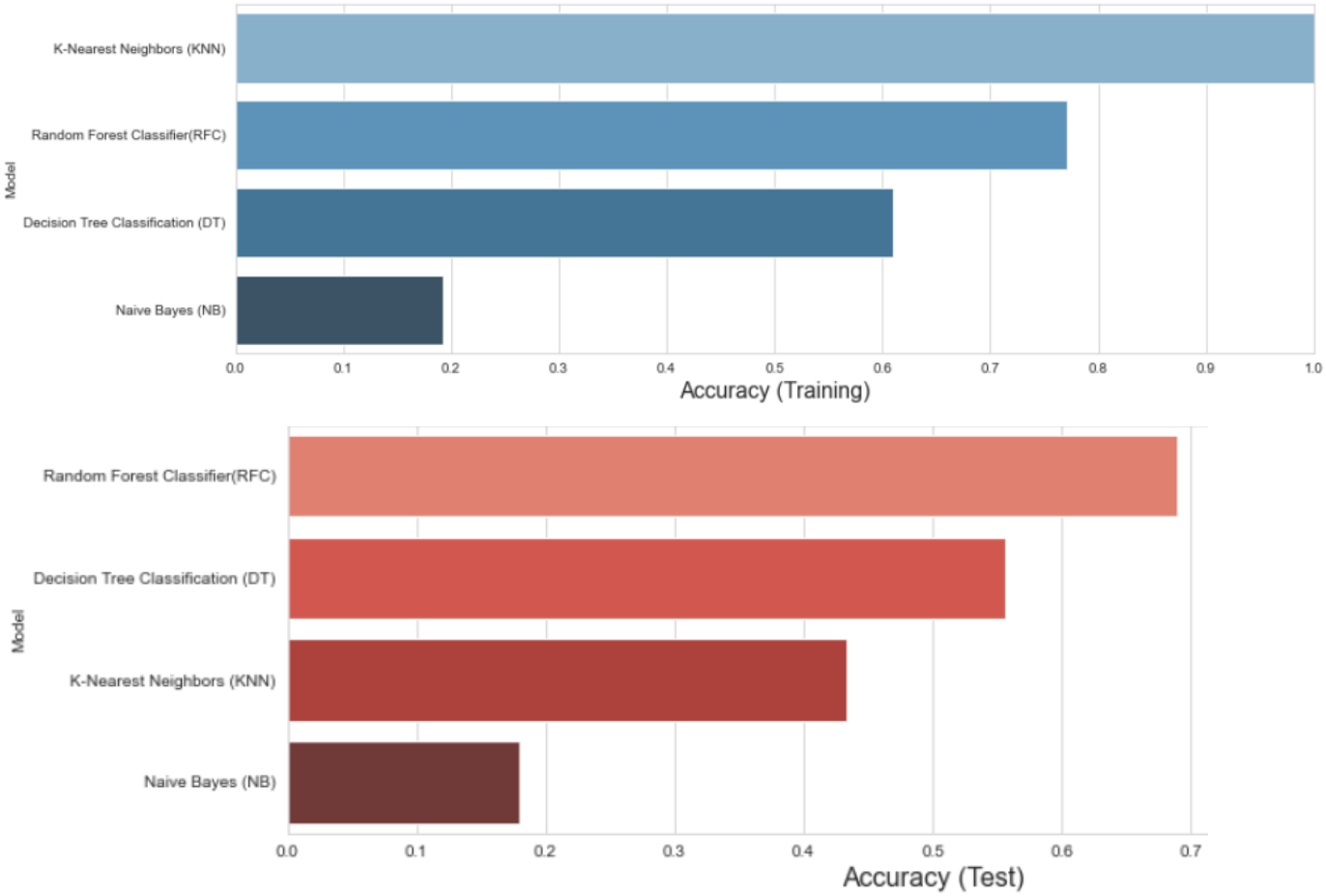


Figure 23: Visualizing Models Performance

4.2.7 The ROC - AUC Curve Analysis

ROC Curve focus on true positive rate and false positive rate at various threshold levels against the false positive rate. If we lower the threshold levels, more items will be classified as positive, increasing both true positives and false positives. The curve (AUC) evaluates the predictive classification performance of the model, 0.5 indicated no predictive power and 1.0 indicated perfect predictive power.

```
#create ROC curve
plt.plot(fpr,tpr,label="AUC="+str(auc))
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
plt.legend(loc=4)
plt.show()
```

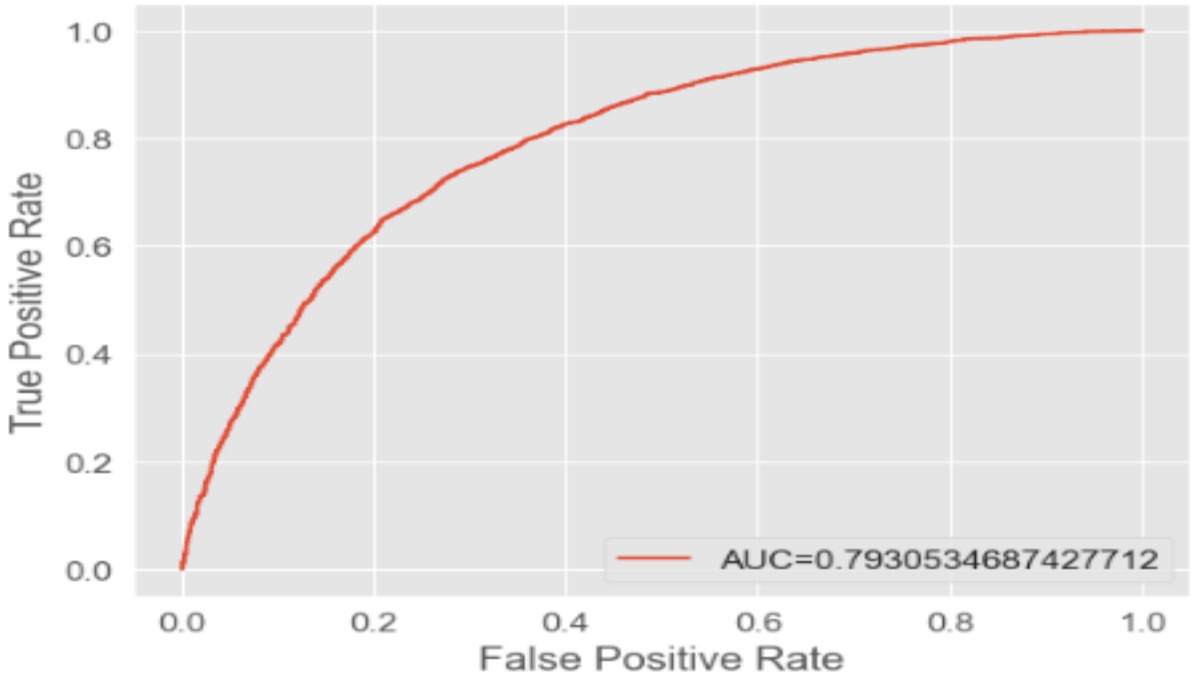


Figure 24: AUC graph for TPR vs FPR

AUC Curve represents a measure of separatability, the higher the AUC, the better the model is at distinguishing between patients on ART adherence and those switching and defaulting ART. The bigger the area covered, the better the machine learning models is at distinguishing the given classes and ideal value for AUC is 1.

```
# Calculates 10 coordinates of the ROC Curve
#tpr, fpr = get_n_roc_coordinates(y_test, resolution = 10)
# Plots the ROC curve
plot_roc_curve(tpr, fpr)
```

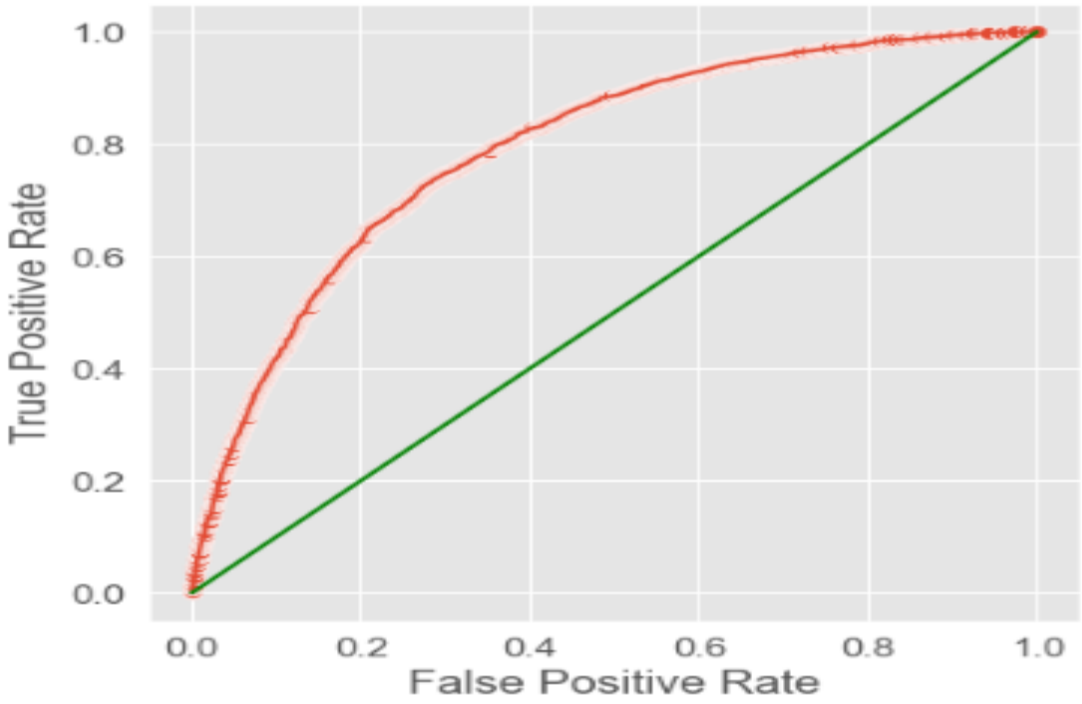


Figure 25: ROC curve to choose a threshold level.

4.2.8 Observations

We compared our model to baseline class categories which corresponds to 47 Counties. Class0 represents patients on retention and ART care, class1 label represents persons virally suppressed. We predicted the probabilities with the highest class probability. The threshold level was 1 and observed the probability of whether a person is on ART care or virally suppressed.

.

```
Out[150]: Text(0, 0.5, 'Frequency')
```

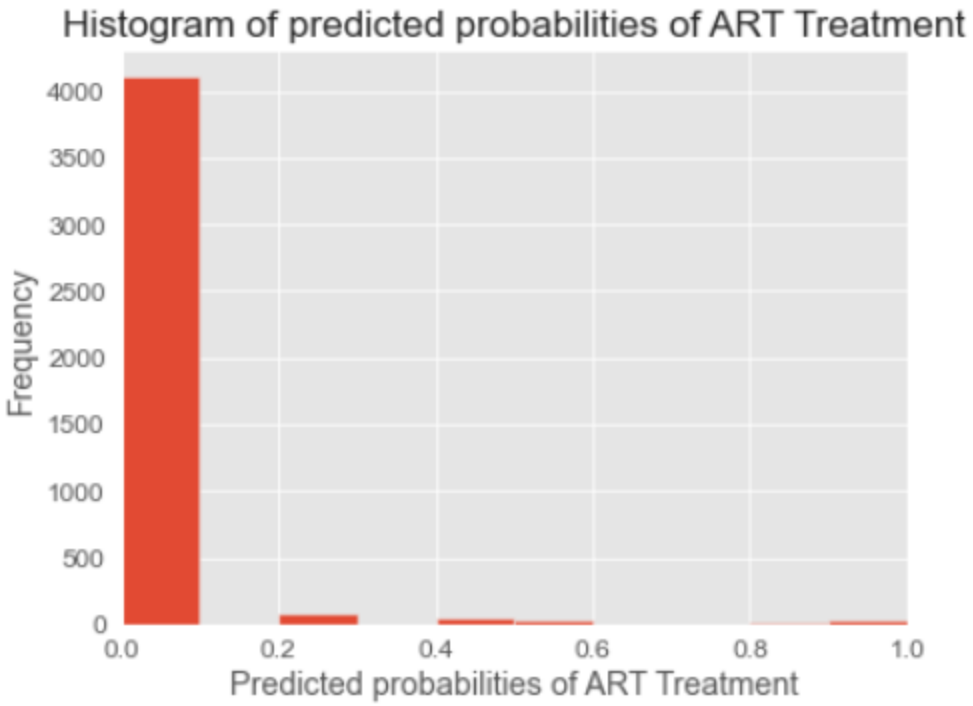


Figure 26: Histogram of Predicted probabilities

From the above histogram there are 4000 observations with probability between 0.0 and 0.1 where patients are likely to go for ART Treatment which is highly positive skewed relative to the number of observations with probability > 0.5. Majority of observations predict that patients are likely to start and enroll for ART Treatment.

4.3 Elbow Method in Supervised Learning

Optimal value of K reduces effects of the noise on the classification making classes less distinct to the count of the nearest neighbors. Elbow methods selects the optimal number of KNN clustering and different k values are created in the loop then tracked for the error_rate for each of these models with a list. Optimal value of K is 15 and the graph of the error rate increase after 15. The error_rate value tends to get consistent. From the plot shown below, the smallest error we got was 0.33 at k = 15

```
plt.figure(figsize=(10,6))
plt.plot(range(1,30),error_rate,color='blue', linestyle='dashed', marker='o',
         markerfacecolor='red', markersize=10)
plt.title('Error Rate vs. K Value')
plt.xlabel('K')
plt.ylabel('Error Rate')
```

```
Text(0, 0.5, 'Error Rate')
```

Figure 27: Error rate vs K-value

4.4 Clinical Characteristics of Participants

For three-year period, 3130 patients were hospitalized and had started ART, 2739 patients had taken first line therapy at 12 months on survival and retention. 329 switched to second line therapy at 12 months due to failure of the first line ART on survival and retention. Regarding medication adherence and total therapy at 12 months patients cases recorded had good adherence while on ART Net cohort at 12 months retention.

After at least 12 months, 329 cases of patients showed high survival and retention for patients who switched therapy to alternative first-line at 12 months were 313 and likely to have viral suppression compared to delayed patients on original first line ART at 12 months. Inter-quartile ranges (IQR) were computed between the alternative first-line therapy and start of second-line therapy after assumption tests where participants delayed to switch the therapy.

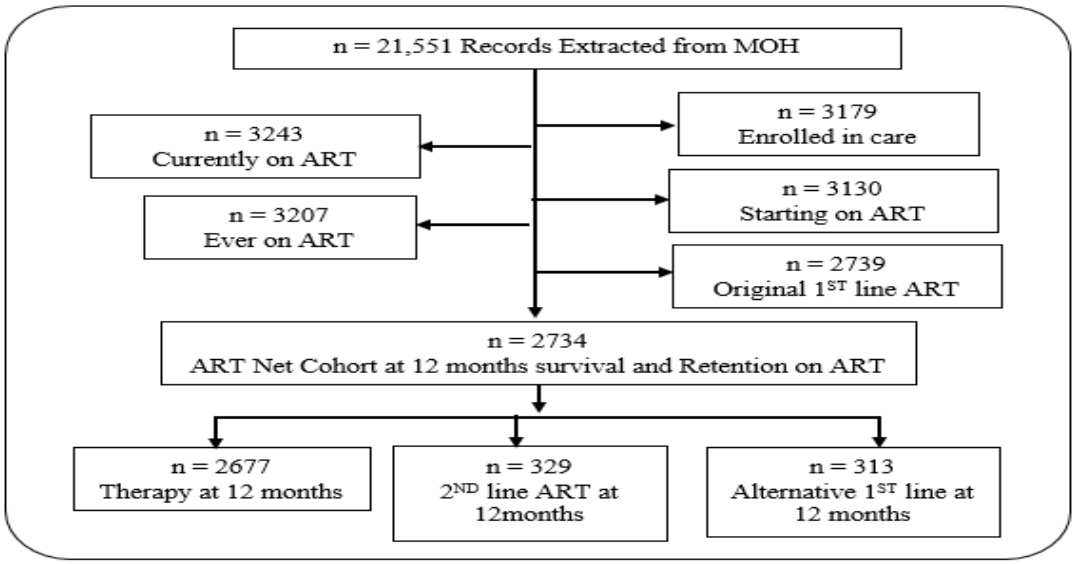


Figure 28: Flowchart of clinical outcome and therapy switch

4.5 Discussion

The research demonstrates how ML algorithms can identify patients on ART within County health clinics. Healthcare providers have difficulty predicting patients on ART treatment, risking missing appointments and determining persons most likely to benefit from resource intensive environment. ART predictive algorithm is developed to predict the treatment and adherence status from the MOH laboratory database sampling unit. This data comprised of PLWH who are on ART treatment, which constitutes total patients enrolled for ART, those currently on ART, patients starting ART, Ever on ART, original first line and second line on ART. We examined data mining techniques to build a model that predicts retention on ART and viral suppression.

The ensemble, trees, Gaussian distribution, Euclidean distance identify predictive factors of ART treatment outcome. Ensemble random forest had a higher performance than the other three algorithms with f1 score of 79% and 77% when all variables were included. The baseline results indicated the testing option of ensemble method (random forest) performing best with accuracy score of 81%, K Nearest neighbor was the second with a classification accuracy of 71%, the decision tree induction method followed (68%). Naïve Bayes classifier performed least accuracy of 20%. Random forest, k nearest neighbor algorithm and decision tree predictive models were able to predict patients retained in care for ART treatment. When a high certainty of patients are retained in care the algorithm calibrates to a specificity of 100% and optimal trade-off between sensitivity and specificity of ART for viral suppression is required.

Naïve Bayes performed poorly on cohort datasets because it introduces bias and independence assumptions on how weights are estimated given a data sample. The model was therefore not suitable for predictive modelling of retention and viral suppression. Random forest gave an exemplary performance on ART cohort data because of random subspace method and bagging preventing overfitting.

4.5.1 Model Verdict

The model training accuracy and testing accuracy during model training phase was performed to benchmark binary classifier and its performance in classifying patients on ART care and virally suppressed from cohort data. The average training accuracy and test accuracy provided the model's performance metric to rank the models. Both training and test accuracy score show that random forest outshines other algorithms. The results demonstrates that random forest performed well on the cohort data.

Model	Avg-train accuracy (%)	Avg-test accuracy (%)	Rank
Decision Tree (DT)	61%	56%	2
Naïve Bayes (NB)	20%	18%	4
Random Forest (RF)	77%	69%	1
K Nearest Neighbor (KNN)	100%	43%	3

Table 7: Models Average Performance

4.6 Contribution to Research

We developed models to predict retention and ART adherence for community-based testing treatment program for patients who are linked to ART treatment in limited resource settings. Based on ART regimen taken by patients there is a chance to know the level of treatment and adherence status for patients who are linked to ART treatment. Patients are advised to visit the healthy facilities within a specified period of time for the purpose of care, follow-up and treatment, if they missed their appointment because of lack of commitments etc.,

Prediction models which is developed using supervised learning, binary classification algorithms improve delivery of care and baseline results at lower cost in terms of time and resources. The model helps clinicians and general practitioners to estimate the ART status and adherence of patients based on feature values when they are unable to know the status levels for patients who are on ART treatment. The model is designed to support routine outreach by health department informing everyday treatment decisions.

4.7 Limitations and Challenges

For the collected data, it was not possible to trace missing data and availability of quality data was a challenge obtained from secondary sources that's the ministry of health (MOH) database sampling unit and contained missing values, unknown values and duplicates. Scope of the study was limited to the use of secondary data to develop the model and getting the data from local hospitals was very challenging task. Virological and clinical factors were limited to the attributes available, but the focus on the impact of ART adherence, including reasons for regimen change or adherence to ART regimens limits further interpretation of results.

CHAPTER FIVE:

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

Retention and viral suppression is imperative for patients and health facilities. We demonstrate how supervised learning models can identify patients on ART for continuous care and improve retention. We retrospectively follow 21551 data sample records within county health facilities located in 47 Counties for conducting the experiment.

We use binary classification algorithms to evaluate the performance allow prioritization of resources to patients likely to benefit and data-driven modelling approach was used to find the relationships between the model state features. Design for the model included the conceptual framework for developed model and evaluation of the predictive model was done by testing the prediction outputs generated from the trained model.

Research investigated patient on ART treatment to identify patients enrolled in care, starting ART, those currently on ART, Original and alternative 1st Line at 12 months Survival and Retention on ART to develop the model and help clinicians formulate medical policies for better planning. ART coverage was high for participants who diagnosed after the treatment strategy. The future implementation interventions may consider to modify the ART-related perceptions for patients diagnoses before the implementation of alternative ART strategy while expanding the accessibility of ART service.

Emphasis is placed on prediction and tracing patients on ART and those retained in care including searching non/participating facility records for patients not on ART. The study has shown the use of supervised learning, classification algorithms in prediction of ART adherence and treatment status can produce predictions with high accuracy and low errors, illustrating the ability of supervised learning to be used in ART centers and comprehensive care units to perform prediction of ART for patients who are retained in care and switching therapy.

5.2 Recommendations for future work

The following recommendations are forwarded:

I. Integrating ART adherence and treatment prediction model to end users to make nurses and general practitioners work easy.

II. More research would be done on how to estimate PLHIV CD4 counts by using other treatment features like virological, clinical and immunological treatment features.

III. Other type of algorithms such as neural networks and recurrent neural network can develop the model leading to the determination of the optimal neural network for efficient distribution.

IV. Redefining or including predictors and exploring interactions/extensions for behavioral data or clinical measures in Nation-wide settings.

More study to estimate the patient's ART adherence could provide a basis for computing the adherence percentage which could be used as a baseline data to the developed model and interlink existing systems to provide seamless integration of the systems and eliminate the need to have to manually transfer data, and ultimately eliminate the errors attributable to human intervention.

REFERENCES

UNAIDS (2014), Three Ones key principles and coordination of National Responses to HIV/AIDS: *guiding principles for national authorities and their partners.* Geneva.

Granich RM, Gilks CF, Dye C,(2009); Williams BG. Universal voluntary HIV testing with immediate antiretroviral therapy as a strategy for elimination of HIV transmission: a mathematical model.

Grimsrud A, Balkan S, Casas EC, et al (2014); Outcomes of antiretroviral therapy over a 10- year period of expansion: a multicohort analysis of African and Asian HIV programs.

Slaymaker E, Hosegood V, (2014) Scale and distribution of excess deaths among HIV positive adults by diagnosis, care and treatment history in African population based cohorts 2007 – 2011.

UNAIDS, 2017. 90-90-90: *An ambitious treatment target to help end the AIDS epidemic.* http://www.unaids.org/sites/default/files/media_asset /90-90-90_en.pdf.

Center for Disease Control (2018). *Understanding the HIV care continuum* https://www.cdc.gov/hiv/pdf/library/factsheets/cdc-hiv-care-continuum.pdf

Jacob Odhiambo, Sindri (2019); Cutting-edge technologies, Predictive Analytics for Antiretroviral therapy.

Parradium, (2019); HMIS – II Project; clinical and biomarker information to predict the risk of treatment failure.

H E. Gendelman (2019)The Promise of Long-Acting Antiretroviral Therapies: *From Need to Manufacture*

Avishek Kumar, Christina Sung (2020). A Machine Learning System for Retaining Patients in HIV Care.

Mayer KH. 2011; Introduction: linkage, engagement, and retention in HIV care: essential for optimal individual-and community-level outcomes in the era of highly active antiretroviral therapy.

MOH, (2020); *Guidelines on Comprehensive HIV Service Delivery and use of antiretroviral drugs for treating and preventing HIV in Kenya.*

Kremer H, Ironson, (2006); Why people with HIV share or don't share with their physicians whether they are taking their medications as prescribed.

Wachira J, Naanyu V, Koech B, Akinyi J, (2014) Health facility barriers to HIV linkage and retention in western Kenya.

Renju J, et al; (2017). Side effects are central effects; a multi-country qualitative study to understand the challenges of retention in HIV care and treatment programmes in sub–Saharan Africa.

World Health Organization, (2012) Framework for metrics to support effective treatment as prevention. Geneva, Switzerland: *https://www.who.int/iris/handle/10665/75387*

Gardner EM, McLees MP, Steiner JF, (2011) The spectrum of engagement in HIV care and its relevance to test-and-treat strategies for prevention of HIV infection.

Bendavid E, Holmes CB, Bhattacharya J, Miller G (2012) HIV development assistance and adult mortality in Africa. JAMA;307(19):2060–7.

Keiser O, Tweya H, Boulle A, Braitstein P.(2009) *Switching to second-line antiretroviral therapy in resource-limited settings:* comparison of programmes with and without viral load monitoring. AIDS (London, England).

Revell AD, Alvarez-Uria G, Wang D, Lane HC, et al (2013) Potential impact of a free online HIV treatment response prediction system for reducing virological failures and drug costs after antiretroviral therapy failure in a resource-limited setting. Biomed Res Int. 2013.

The Joint United Nations Programme on HIV and AIDS, (2019); 90-90-90 an ambitious treatment target to help end the AIDS epidemic. *https://www.unaids.org/sites/default/files/media_asset/90-90-90/en.pdf*

Rodger AJ, Philips A, Speakman A, (2014); Attitudes of people in the UK with HIV who are antiretroviral (ART) Naïve to starting ART at high CD4 counts for potential health benefit or to prevent HIV transmission.

Centers for Disease Control and Prevention, CDC (2017). "Data to Care program guidance: Using HIV surveillance data to support the HIV care continuum

A. Guta, S.J. Murray, and M. Gagnon, (2016) HIV, viral suppression and new technologies of surveillance and control," *Body & Society*, doi: *https://doi.org/10.1177/1357034X15624510*

F. Kiweewa, A. Esber, E. Musingye, D. Reed, T.A. HIV (2019), virologic failure and its predictors among HIV-infected adults on antiretroviral therapy in the African Cohort Study," *PLoS ONE*, doi: *https://doi.org/10.1371/journal.pone.0211344*

J.C. Dombrowski, J. Bove, J.C. Roscoe, J. Harvill (2017) Out of care' HIV case investigations: A collaborative analysis across 6 states in the northwest US*,"* JAIDS Journal of Acquired Immune Deficiency Syndromes doi: *https://doi.org/10.1097/QAI.0000000000001237.*

Wang D, Larder B, Revell A, De Wolf F, et al. (2009); A comparison of three computational modelling methods for the prediction of virological response to combination HIV therapy.

Revell AD, Alvarez-Uria G, Montaner JS, Lane HC, et al. (2013); Potential impact of a free online HIV treatment response prediction system for reducing virological failures and drug costs after antiretroviral therapy failure in a resource-limited setting.

Brachman, R. J. and Anand, T. (1996) *The Process of Knowledge Discovery in Databases.*

Lu, H., Setiono, R. and Liu, H. (1996) Effective Data Mining Using Neural Networks. *IEEE Transactions on Knowledge and Data Engineering.*

Ali, Khan and Maqsood (2012); Random Forests and Decision Trees. *International Journal of Computer Science Issues.*

Liaw & Wiener, (2017) Modification of the random forest algorithm to avoid statistical dependence problems when classifying remote sensing imagery.

Revell, et al., (2011) Clinical Evaluation of the potential Utility of Computational Modelling as an HIV Treatment Selection Tool by Physicians with Considerable HIV experience.

Shen, et al., (2016); Shortlist selection with residual-aware distance estimator for k-nearest neighbor. *Locality constrained representation-based KNN classification.*

Grabowski MK, Serwadda DM, Gray RH, Nakigozi G, Kigozi G, Kagaayi J, et al (2017) HIV prevention efforts and incidence of HIV in Uganda. N Engl J Med. 2017. *https://doi.org/10.1056/NEJMoa1702150 PMID:*

Ware NC, Wyatt MA, Geng EH, Kaaya SF, Agbaji OO, Muyindike WR, et al (2013). Toward an understanding of disengagement from HIV treatment and care in sub-Saharan Africa: a qualitative study. PLoS Med. 2013; *https://doi.org/10.1371/journal.pmed.*

R.S Butt, & I Ahmad, (2019); Integral backstepping and synergetic control for tracking of infected cells during early antiretroviral therapy.

T. Heestermans, J. L Browne, (2016); Determinants of adherence to antiretroviral therapy among HIV-positive adults in Sub-Saharan Africa; A systematic review.

J. Davey, & A. Nguimfack, S. Hares, W. Ponce, (2012); Evaluating SMS reminders in improving ART and PMTCT adherence in Mozambique: Challenges in achieving scale.

R. Baraldi, K. Cross, C. McChesney et al, (2014); Uncertainty quantification for a model of HIV-1 patient response to antiretroviral therapy interruptions , *" in proceedings of the 2014 American Control Conference*

T. D Chala, (2019): Data mining technology enabled antiretroviral therapy (ART) for HIV positive patients in Gondar university hospital, Ethiopia, Bioinformation.

Barradas DT, Gupta S, Moyo C, Sachathep K, Nkumbula T, et al (2017). *Findings from the 2016 Zambia Population-based HIV Impact Assessment (ZAMPHIA):* HIV prevalence, incidence and progress towards the 90-90-90 goals. Abstract TUAC0301.

Geng EH, Bangsberg DR, Bwana MB, Yiannoutsos CT, et al (2010). Understanding reasons for and outcomes of patients lost to follow-up in antiretroviral therapy programs in Africa through a sampling-based approach. *https://doi.org/10.1097/QAI.0b013e3181b843f0.*

Mannheimer SB, Mukherjee R, Hirschhorn LR, et al (2006); The CASE adherence index. *A novel method for measuring adherence to antiretroviral therapy*. AIDS Care.

Hanley JA, Negassa A, Edwardes MD, (2003). Statistical analysis of correlated data using generalized estimating equations: An orientation. Am J Epidemiol.

Maskew M, & Sharpey-Schafer K, (2010). Machine learning to predict retention and viral suppression in South African HIV treatment cohorts.

Tinglong Yang, Xueying Yang, Linghua Li, (2021). HIV diagnosis period influences ART initiation: *findings from a prospective cohort study in China*

Revell, P. Khabo, (2013) Computational models as predictors of HIV treatment outcomes for the Phidisa cohort in South Africa

Regoniel, 2016. https://1715373997.rsc.cdn77.org/wp-content/uploads/2018/12/TOCThesis.pdf

Maskew M, & Sharpey-Schafer K, (2010). predictive modeling for retention and viral suppression in South African HIV treatment cohorts; *https://www.nature.com/articles/s41598-022-16062-0*

Cyril Goutte & Eric Gaussier, (2005), probabilistic Interpretation of Precision, Recall and F -score, with Implication for Evaluation

APPENDICES

Patients Retained for ART care: A data dictionary describing the features of the dataset.

1. List of Top Counties with Highest ART Coverage

-->Check frequency Distribution of values in workclass variable

County	ART Coverage and Adherence
Nairobi County	1607
Homa Bay County	1380
Kakamega County	1367
Bungoma County	1216
Siaya County	1029
Kisii County	969
Kisumu County	969
Nakuru County	775

1. Orgunit Name/Health Facilities with highest number of patients retained for ARTs

Orgunit Name	Patients Retained for ART care
Sex Workers Outreach Program (Kibra)	25
Swop Korogocho	22
SWOP Clinic Donholm	22
Hoymas VCT (Nairobi)	21
Balambala Sub-District Hospital	17

2. HIV care Cascade and Assessment of ART Adherence and Eligibility

Data__ART Status	Total ART Outcome
Total Currently on ART	3243
Total Ever on ART	3207
Total Enrolled in Care	3179
Total Starting on ART	3130
On Original 1st Line at 12 months Survival and Retention on ART	2739
ART Net Cohort at 12 months Survival and Retention on ART	2734
Total on therapy at 12 months	2677
On 2nd Line (or higher) at 12 months Survival and Retention on ART	329
On alternative 1st Line at 12 months Survival and Retention on ART	313