ST. MARY'S UNIVERSITY

SCHOOL OF POSTGRADUATE STUDIES

Department of Computer Science

HIV Target Group Prediction Using Machine Learning

By: Yosef Abebual

Addis Ababa, Ethiopia

June, 2024 G.C

ST. MARY'S UNIVERSITY

SCHOOL OF POSTGRADUATE STUDIES

Department of Computer Science

HIV Target Group Prediction Using Machine Learning

A Thesis Submitted to the School of Postgraduate Studies Presented in Partial Fulfilment of the Requirements for the Degree of Masters of Science in Computer Science

By

Yosef Abebual

Advisor: Alembante Mulu (PhD)

Addis Ababa, Ethiopia

June, 2024 G.C

DECLARATION

I, **Yosef Abebual**, the undersigned, claim that the thesis titled **"HIV Target Group Prediction Using Machine Learning"** is my original work. I conducted the study work alone, with the advice and cooperation of the research supervisor. This paper was not submitted for any degree or diploma program at this or any other school, and all sources of materials included in the thesis were properly acknowledged.

Name of Student	Signature	Date
Yosef Abebual	_____	_____

This is to certify that the thesis entitled: **HIV Target Group Prediction Using Machine Learning** submitted in Partial Fulfilment of the requirements for the degree of Masters of Computer Science of the Postgraduate Studies, ST. MARY'S UNIVERSITY and is a record of original research carried out by Yosef Abebual **SGS/0479/2014A**, under my supervision, and no part of the thesis has been submitted for any other degree or diploma. The assistance and help received during the course of this investigation have been duly acknowledged. Therefore, I recommend it be accepted as fulfilling the thesis requirements.

Alembante Mulu (PhD)	_____	_____
Name of Supervisor	Signature	Date

Certificate of Approval

This certifies that Yosef Abebual thesis, "**HIV Target Group Prediction Using Machine Learning,"** which was turned in for credit toward the Masters of MSc in Computer Science degree, satisfies all requirements set forth by the university and is up to par in terms of originality and quality.

Signature of Board of Examiner's:

_____	_____	_____
External examiner	signature	Date
_____	_____	_____
Internal examiner	Signature	Date
_____	_____	_____
Dean, SGS	Signature	Date

Thesis Title: **"HIV Target Group Prediction Using Machine Learning",** ST. MARY'S UNIVERSITY, Ethiopia.

Submitted by: Yosef Abebual

Signature: _____ Date: _____

List of Figures

List of Tables

List of Equations

Table of Contents

Acknowledgment

I want to thank my advisor from the bottom of my heart, Alembante Mulu (PhD) whose guidance and unwavering support have been instrumental in the completion of this study. Your expertise, patience, and encouragement have truly shaped this research and my academic journey.

I am also deeply thankful to my family for their constant love and encouragement throughout this endeavor. Especially My beloved wife Nigist and My son Kidus Yosef, Your belief in me has been my driving force, and I am grateful for the sacrifices you have made to see me succeed.

To my friends, who have provided a constant source of motivation and a welcome distraction when needed, thank you for being there for me? Your camaraderie has made this journey more enjoyable and memorable. Lastly, I extend my appreciation to all those who have contributed in any way to the completion of this study. Your support has been invaluable, and I am truly grateful for the collaborative spirit that has enriched this academic experience.

Thank you.

Acronyms

AGYW	Adolescent Girls and Young Adults
AIDS	Acquired Immune Deficiency Syndrome
AOC-ROC	Receiver operating characteristic curve
AYA	Adolescent and Young Adults
CA	Classification Accuracy
FN	False Negative
FP	False Positive
FSW	Female Sex Workers
HIV	Human Immune Virus
HRM	High Risk Males
IDE	Integrated Development Environment
SVM	Support Vector Machine
TN	True Negative
TP	True Positive

Abstract

HIV continues to be a global health concern that necessitates cutting-edge methods of diagnosis and treatment. Owing to the intricate nature of the HIV pandemic, specific strategies are needed to pinpoint vulnerable people. This study tackles the challenge of precise identification within specific HIV target groups, namely Adolescent Girls and Young Women (AGYW), High-Risk Men (HRM), and Female Sex Workers (FSW). Leveraging machine learning algorithms include **Support vector machine, XGBoost, Random forest and linear regression**. The research integrates locally sourced datasets from hospital records, aiming to elevate intervention precision. The study seeks to transform public health by introducing a data-driven approach to unravel intricate relationships and variables influencing HIV prevalence among distinct target groups. Despite progress in global health efforts, traditional methods grapple with precision and efficiency limitations. The adoption of machine learning offers a promising solution, contributing to a nuanced understanding of dynamics within key populations. Addressing gaps in existing literature particularly the scarcity of studies at the intersection of machine learning and the identification of specific HIV target groups using locally collected datasets.

The study rigorously evaluates the performance of four algorithms on an HIV service delivery dataset. Results indicate consistently high accuracy across all models, with ensemble approaches (XGBoost and Random Forest) slightly outperforming others. Notably, Support Vector Machine achieved 96.33% accuracy, XGBoost reached 96.51%, Random Forest attained 96.49%, and Linear Regression demonstrated commendable accuracy at 96.28%. This research significantly contributes to advancing machine learning applications in healthcare and addresses a crucial gap in the current body of knowledge.

Keywords: Machine Learning, HIV, Support Vector Machine, XGBoost, Random Forest, Linear Regression

Chapter 1: Introduction

1.1. Background of the study

HIV remains a persistent worldwide health challenge, demanding novel approaches to detection and treatment. Because of the complexities of the HIV epidemic, tailored approaches are required to identify important populations at increased risk. [1] This study addresses this issue by using powerful machine learning algorithms to predict the likelihood of individuals belonging to specific HIV target groups, including adolescent girls and young women (AGYW), high-risk males (HRM), and female sex workers (FSW). Using cutting-edge healthcare technology, the study attempts to improve intervention precision by integrating locally obtained datasets from hospital records. [2]

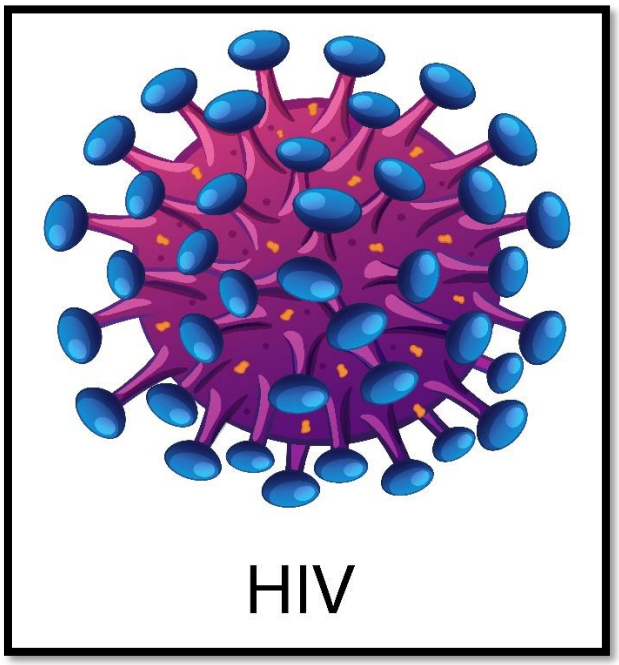


Figure 1-1: HIV AIDS

The integration of machine learning into public health practices not only offers a novel avenue for understanding the dynamics of HIV transmission but also provides an opportunity to tailor interventions effectively. By analyzing a diverse set of variables within the collected datasets, the

study aims to contribute valuable insights into the nuances of HIV prevalence among AGYW, HRM, and FSW. [3]

1.2. Motivation of the Study

The motivation for this study stems from the urgent need to address the persistent challenges in accurately identifying and intervening among specific HIV target groups. Despite considerable advancements in healthcare, the global fight against HIV requires innovative and tailored strategies. The motivation behind integrating machine learning algorithms into public health practices lies in the potential transformative impact these technologies can have on the precision and efficacy of interventions. [4]

1.2.1. Advancements in Technology:

The rapid advancements in technology, particularly in the field of machine learning, provide an unprecedented opportunity to harness the power of data for improved decision-making. By leveraging these technologies, we can move beyond traditional, one-size-fits-all approaches and delve into a realm where interventions are finely tuned to the characteristics of specific populations. [5]

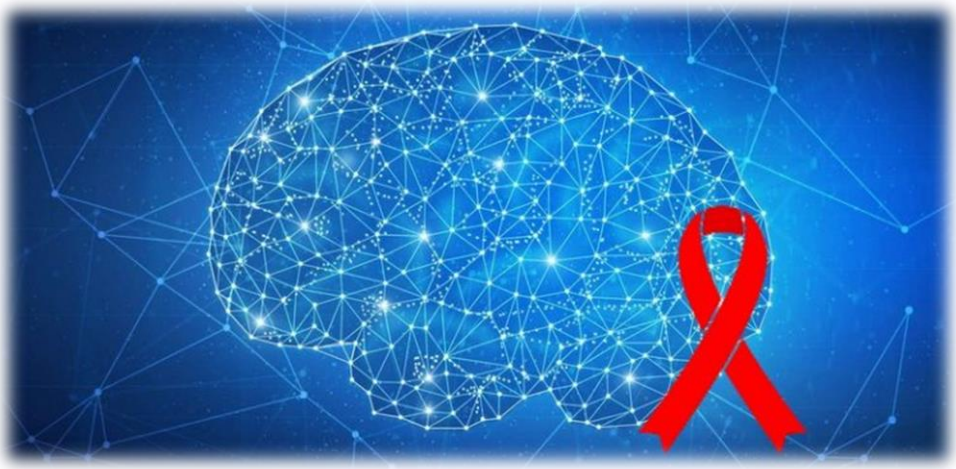


Figure 1-2: Technology and HIV AIDS

1.2.2. Precision in Public Health Interventions:

The traditional methods employed for identifying HIV target groups often fall short in precision, leading to resource inefficiencies and suboptimal outcomes. Machine learning offers the promise

of parsing through vast datasets to identify intricate patterns and relationships within the variables, allowing for a more accurate classification of individuals into target groups. [6]

1.2.3. Potential for Timely and Targeted Interventions

In the dynamic landscape of public health, timely interventions are crucial. Machine learning algorithms have the potential to provide real-time predictions, enabling healthcare professionals to stay ahead of emerging trends and allocate resources where they are most needed. This study is motivated by the prospect of creating a predictive model that not only identifies target groups accurately but also does so in a timely manner. [6]

1.2.4. Contribution to Knowledge and Practice

By conducting this research, we aim to contribute to the growing body of knowledge on the intersection of machine learning and public health. The insights gained from this study can inform future practices, policies, and interventions, setting a precedent for the integration of cutting-edge technologies into public health strategies. [7]

1.2.5. Global Impact on HIV Prevention

Ultimately, the motivation behind this study is grounded in the overarching goal of making a meaningful impact on the global effort to prevent and control HIV. The use of machine learning to predict target groups with precision holds the potential to revolutionize how interventions are designed and implemented, ultimately reducing the burden of HIV on affected populations. [1]

In summary, the motivation for this study lies in the recognition of the transformative potential of machine learning in enhancing the precision, timeliness, and effectiveness of public health interventions, with the ultimate goal of contributing to the global fight against HIV.

1.3. Statement of the Problem

Despite tremendous progress in global health initiatives, accurately identifying individuals in HIV target groups remains a difficult undertaking. Traditional methods frequently have limitations in precision and efficiency, making it difficult to customize interventions effectively. The complexity

of the HIV epidemic necessitates a more detailed knowledge of the dynamics within critical groups. [2]

This study addresses the critical issue of imprecise identification by incorporating machine learning techniques into public health. Current techniques may overlook tiny but significant patterns in the data, resulting in inefficient intervention targeting. The use of machine learning intends to modernize this process by providing a data-driven method that can identify complicated linkages and variables that contribute to HIV prevalence among several target groups, including AGYW, HRM, and FSW. [8]

Using locally collected datasets from hospital records, this study aims to not only improve prediction accuracy but also provide a better understanding of the socio-demographic, behavioral, and clinical factors that influence the likelihood of individuals belonging to these target groups. By doing so, it hopes to close information gaps and help to the creation of more effective, targeted interventions in the ongoing fight against HIV. [9]

1.3.1. Research Gaps

Despite the progress made in the field of predicting HIV target groups using machine learning algorithms, there exists a notable research gap that necessitates further investigation. The current literature primarily focuses on the application of these algorithms in healthcare settings and their potential for precision medicine. However, there is a distinct scarcity of studies specifically addressing the intersection of machine learning and the accurate identification of distinct HIV target groups, namely Adolescent Girls and Young Women (AGYW), High-Risk Men (HRM), and Female Sex Workers (FSW), using locally collected datasets. [10]

Furthermore, existing research often lacks a comprehensive exploration of the socio-demographic, behavioral, and clinical factors influencing the likelihood of individuals belonging to these specific target groups. While machine learning holds promise in enhancing accuracy, the nuanced understanding of the intricate relationships among these factors remains an underexplored aspect. [11].

This study aims to bridge this research gap by focusing on the unique challenges posed by the HIV epidemic and tailoring machine learning approaches to the identification of AGYW, HRM, and FSW. By incorporating locally collected datasets from hospital records, the research seeks to contribute not only to the advancement of machine learning applications in healthcare but also to

the development of targeted interventions for distinct HIV target groups, filling a critical void in the current body of knowledge.

1.4. Research Questions

To conduct this study the following research questions were posed as a basis for this research based on the above-mentioned specific objectives:-

RQ 1: Which key factors in the local dataset most significantly influence HIV risk?

RQ 2: Which machine learning algorithms are most effective for predicting HIV target groups using the local dataset?

RQ 3 What metrics best evaluate the performance of the HIV prediction model?

1.5. Objective

The Proposed research contains both general and specific objectives to be achieved.

1.6. General Objective

The overarching objective of this study is to leverage machine learning techniques for the prediction of HIV/AIDS Target Group within distinct population groups

1.6.1. Specific Objective

To accomplish the above stated general objectives, the following specific objectives are developed:

- To use state of the art for literature review
- To prepare required dataset for HIV target group prediction using machine learning.
- To study the nature of HIV and Target group from the local dataset.
- To design model for HIV target group prediction using machine learning (to train and test).
- To evaluate the performance of the developed model.

1.7. Scope and Limitation of the Study

1.7.1. Scope of the study

This study focuses on the application of machine learning algorithms to predict HIV/AIDS prevalence within specific target population groups, namely Adolescent Girls and Young Women (AGYW), High-Risk Men (HRM), and Female Sex Workers (FSW). The scope encompasses:

- **Geographical Focus:** The study is confined to the local context, utilizing datasets collected from hospital records within a defined geographical area.

- **Data Collection:** The data collection process includes variables related to socio-demographic, behavioral, and clinical factors relevant to the specified target groups.
- **Machine Learning Algorithms:** Evaluation and comparison of common machine learning algorithms, such as decision trees, support vector machines, and other machine learning algorithm's. For their effectiveness in predicting the predominantly affected population group.
- **Preprocessing Techniques:** Implementation of data preprocessing techniques, including cleaning, normalization, and feature engineering, to optimize the input data for machine learning analysis.
- **Evaluation Metrics:** The assessment of the predictive model's accuracy, reliability, and effectiveness using metrics such as precision, recall, and F1 score.
- **Practical Application:** The study aims to provide practical insights and recommendations for the development of tailored interventions based on the outcomes of the machine learning predictions.

The study's scope is delimited to these specified parameters, ensuring a focused and meaningful exploration of the intersection between machine learning and HIV/AIDS prevalence prediction within the identified target population groups.

1.7.2. Limitation of the study

While this study aims to contribute valuable insights into predicting HIV/AIDS prevalence through machine learning, certain limitations need acknowledgment:

- **Data Availability:** The study relies on locally collected datasets, and limitations in data availability or completeness may impact the robustness of the predictive model.
- **Generalization:** Findings may be context-specific to the selected geographical area and may not be directly applicable to broader or diverse populations.
- **Algorithm Sensitivity:** The performance of machine learning algorithms is subject to variations based on parameter settings, and the selected algorithms may exhibit sensitivity to these configurations.

- **Assumed Causality:** The study assumes correlations between identified factors and HIV prevalence without establishing causality, as certain complex relationships may require further investigation.
- **Model Over fitting:** Despite preprocessing efforts, the predictive model might face challenges, such as over fitting, especially if the algorithm captures noise rather than genuine patterns in the data.
- **Dynamic Nature of HIV Epidemic:** The HIV landscape is dynamic, and factors influencing prevalence may evolve over time; the study's findings may have a temporal relevance.
- **Resource Constraints:** The scope of the study may be constrained by resource limitations, impacting the depth and breadth of the analysis and potentially limiting the exploration of additional variables.
- **Ethical Considerations:** Privacy concerns and ethical considerations related to handling sensitive health data may impose restrictions on the accessibility and utilization of certain information.

Acknowledging these limitations ensures a nuanced interpretation of the study's findings and promotes transparency in communicating the potential constraints that may impact the validity and generalizability of the research.

1.8. Significance of the Study

This study holds profound significance in the realms of public health and technological innovation. By employing machine learning algorithms to predict HIV/AIDS prevalence within specific vulnerable populations, including Adolescent Girls and Young Women (AGYW), High-Risk Men (HRM), and Female Sex Workers (FSW), the research endeavors to make a substantial impact on targeted interventions. The innovative application of machine learning techniques not only promises to enhance the precision of predictions but also revolutionizes healthcare practices by offering data-driven insights into the dynamics of the HIV epidemic.

The potential benefits are manifold: optimized resource allocation, efficient prevention efforts, and the development of tailored interventions that address the unique challenges faced by distinct population groups. Beyond its immediate practical implications, the study contributes to the

scientific advancement of predictive modeling methodologies, bridging the gap between data science and public health. [12]

This research is poised to influence global health strategies, with findings that may extend beyond the local context and provide valuable insights applicable to regions facing similar challenges. Moreover, by fostering cross-disciplinary collaboration between data science and public health, the study contributes to a holistic approach in combating the HIV epidemic, showcasing the transformative power of technology in the service of public health initiatives.

Chapter 2: Literature Review and Related Work

2.1. Introduction

The literature review and related work in this chapter provide a comprehensive examination of existing research and contributions in the intersection of machine learning and the prediction of HIV/AIDS prevalence, particularly within distinct population groups. The review encompasses a thorough analysis of studies focusing on the application of machine learning algorithms in healthcare, the identification of key factors influencing HIV prevalence, and the development of predictive models for targeted interventions.

2.2. HIV AIDS

HIV is a virus that attacks the immune system, weakening it over time. It is transmitted through specific bodily fluids such as blood, semen, vaginal fluids, and breast milk. AIDS is the advanced stage of HIV infection, characterized by a severely compromised immune system. While there is no cure for HIV/AIDS, antiretroviral therapy can effectively manage the virus, allowing individuals to live long and healthy lives. It's important to practice safe sex and take precautions to prevent transmission. [13]

Figure 2-1: HIV AIDS

2.3. Machine Learning in Healthcare

The literature reveals a growing body of research exploring the integration of machine learning techniques in healthcare settings. Notable studies, such as Obermeyer and Emanuel's work on predicting medical outcomes using machine learning, and Althoff et al.'s examination of artificial intelligence technologies in medicine, provide insights into the potential of machine learning for predictive modeling in health contexts. [14]

2.4. Predictive Modeling in HIV/AIDS Research

Research specific to the prediction of HIV/AIDS prevalence showcases pioneering efforts to employ machine learning algorithms. Domingo’s exploration of useful machine learning principles and Goldstein et al.'s systematic review of risk prediction models using electronic health records contribute foundational insights into the broader field. [15]

2.5. Addressing Gaps in Current Research

While the literature provides a foundation, there remains a noticeable gap in the application of machine learning to predict HIV prevalence within distinct target groups such as Adolescent Girls and Young Women (AGYW), High-Risk Men (HRM), and Female Sex Workers (FSW). This study aims to bridge this gap by incorporating locally collected datasets, thus contributing novel insights to the existing body of knowledge.

This chapter presents a thorough review of literature and related work, providing a background for the current study. The integration of machine learning in healthcare, predictive modeling in HIV/AIDS research, localized studies, and identified gaps collectively inform the methodology and approach of the current research, emphasizing the need for tailored interventions within specific population groups.

2.6. Target Groups in HIV/AIDS Research

Within the extensive landscape of HIV/AIDS research, understanding the unique vulnerabilities and dynamics of specific target groups is pivotal. Existing literature sheds light on the distinct challenges faced by the identified target groups in this study Adolescent Girls and Young Women (AGYW), High-Risk Men (HRM), and Female Sex Workers (FSW).

2.6.1. Vulnerabilities of Adolescent Girls and Young Women (AGYW)

Research indicates that AGYW face heightened vulnerability to HIV/AIDS due to socio-economic factors, gender inequalities, and limited access to education and healthcare. Studies, includes exploration of HIV risk factors among AGYW, underscore the importance of tailored interventions to address their specific needs. [16]

2.6.2. Unique Challenges of High-Risk Men (HRM)

High-Risk Men, a population often marginalized in HIV/AIDS research, present distinctive challenges related to stigmatization, lack of awareness, and limited accessibility to preventive measures. Studies on HRM and HIV prevention, emphasizes the necessity for targeted strategies to engage and protect this group. [17]

2.6.3. Female Sex Workers (FSW)

Female Sex Workers, due to the nature of their work and social contexts, encounter heightened risks of HIV transmission. Research, including examination of HIV prevalence and risk factors among FSW, highlights the importance of tailored interventions that address the specific challenges faced by this population. [18]

In summary, the literature review on target groups within HIV/AIDS research underscores the unique challenges faced by AGYW, HRM, and FSW. The identified vulnerabilities, intersectionality, and the need for holistic and empowerment-based approaches provide a foundational understanding for developing tailored interventions within the scope of the current study.

2.7. Machine Learning Algorithms

Machine learning algorithms are the foundation of modern artificial intelligence systems, allowing computers to learn from data and make predictions or judgments. These algorithms are aimed to detect patterns and relationships in datasets, allowing robots to make accurate predictions or conduct actions without being explicitly programmed. [19]

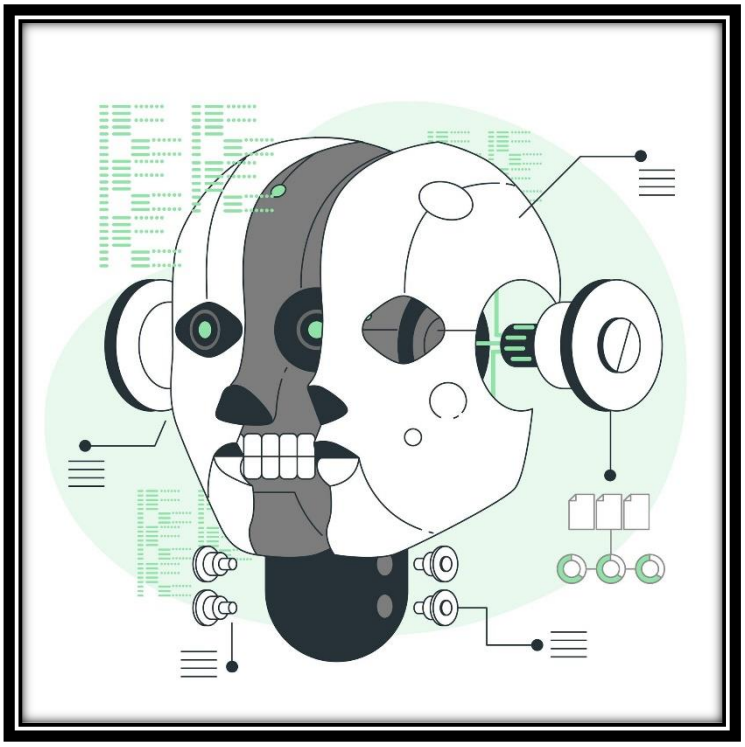


Figure 2-2: Machine Learning Algorithms

2.7.1. Supervised machine learning Algorithms

Supervised learning algorithms are a prominent type of machine learning algorithm. These algorithms learn on labeled training data, in which each data item corresponds to a known target value or class label. The purpose of supervised learning is to create a model that can predict the target value for new, previously unseen cases. Decision trees, random forests, support vector machines (SVM), and neural networks are examples of supervised learning techniques. [20]

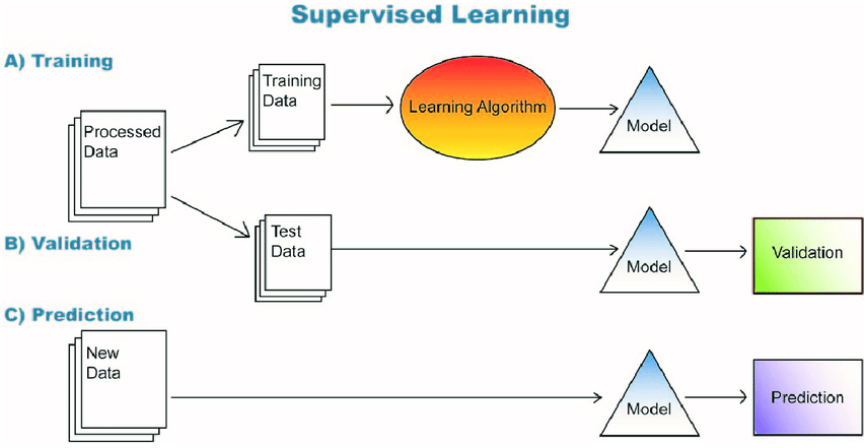


Figure 2-3: Supervised machine learning Algorithms

2.7.1.1. Support vector machine

SVMs, or Support Vector Machines, are a common class of machine learning algorithms used for classification and regression tasks. They are particularly effective when dealing with data that cannot be linearly separated. The core principle behind SVMs is to find a hyper plane that maximally separates data points of different classes. This hyperplane is chosen to optimize the margin, which is the distance between the hyperplane and the nearest data points of each class. SVMs strike a balance between maximizing the margin and minimizing classification errors, making them robust and reliable models. [21]

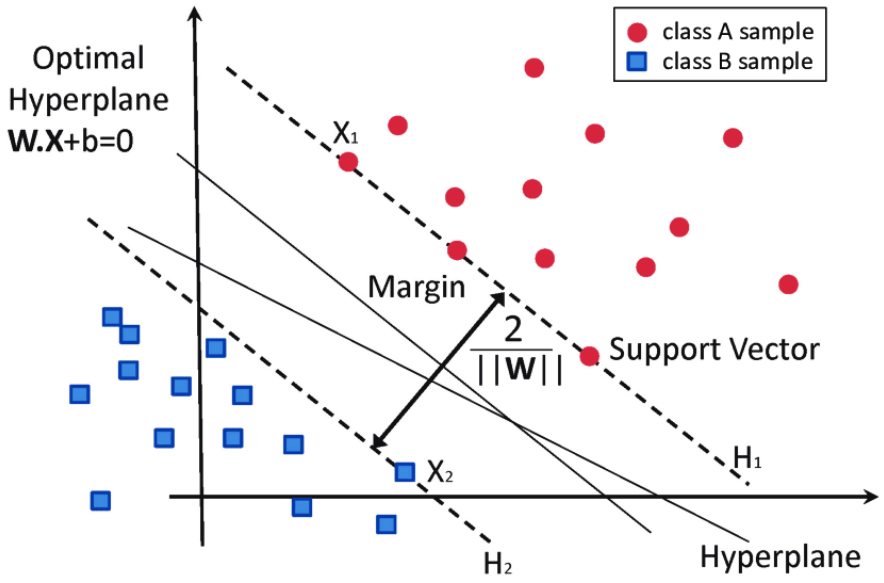


Figure 2-4: Support vector machine

One of the key advantages of SVMs is their ability to handle high-dimensional feature spaces. By employing the kernel trick, SVMs can implicitly transform the input data into a higher-dimensional space where the classes can be linearly separated. This enables SVMs to capture complex relationships and make accurate predictions. SVMs have been widely applied in various domains, including text categorization, image recognition, bioinformatics, and finance. Their robustness, generalization capability, and ability to handle large datasets have made them a popular choice in many machine learning applications. [22]

2.7.1.2. Random forest

Random Forest is a widely used ensemble learning approach that combines multiple decision trees to make predictions or classifications. It is renowned for its adaptability, robustness, and high accuracy across various tasks. The algorithm works by creating a collection of decision trees, where each tree is trained on a randomly selected subset of the training data. During the training process, each tree independently makes predictions, and the final prediction is determined by aggregating the predictions of all the trees. Random Forest excels at handling high-dimensional datasets with numerous features, capturing complex relationships, and reducing Over fitting. It finds applications in diverse industries such as banking, healthcare, marketing, and image recognition. Random Forest is a versatile ensemble learning technique that combines decision trees to achieve accurate predictions or classifications. Its ability to handle high-dimensional data, robustness, and wide range of applications make it a popular choice in machine learning. [23]

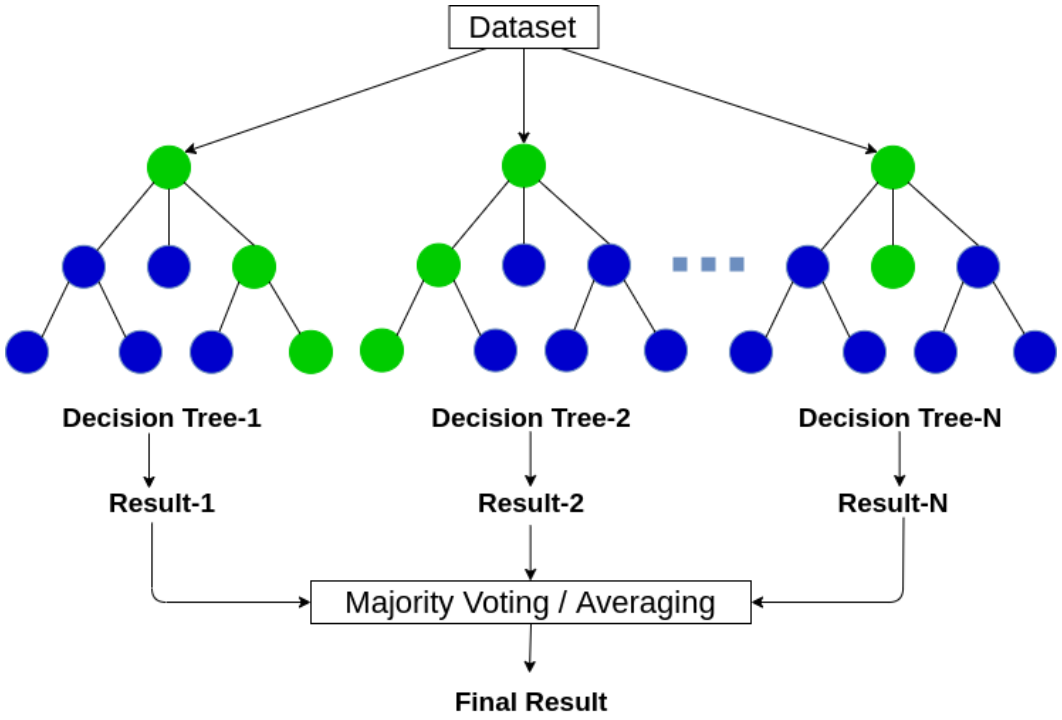


Figure 2-5: Random forest

2.7.1.3. Linear Regression

Regression analysis is a supervised learning approach that generates continuous variables from labeled data. When using multiple regression algorithms, it is critical to choose the proper regression technique based on your data and the problem that your model aims to answer. This article will explain the notion of regression analysis, which is utilized in machine learning and data science. We'll also learn why we need regression analysis and how to select the best strategy for the data to achieve the highest model test accuracy. In the linear regression model, the dependent and independent variables are linked linearly by a single parameter. Multiple linear regression models are employed when there are more independent variables. [24]

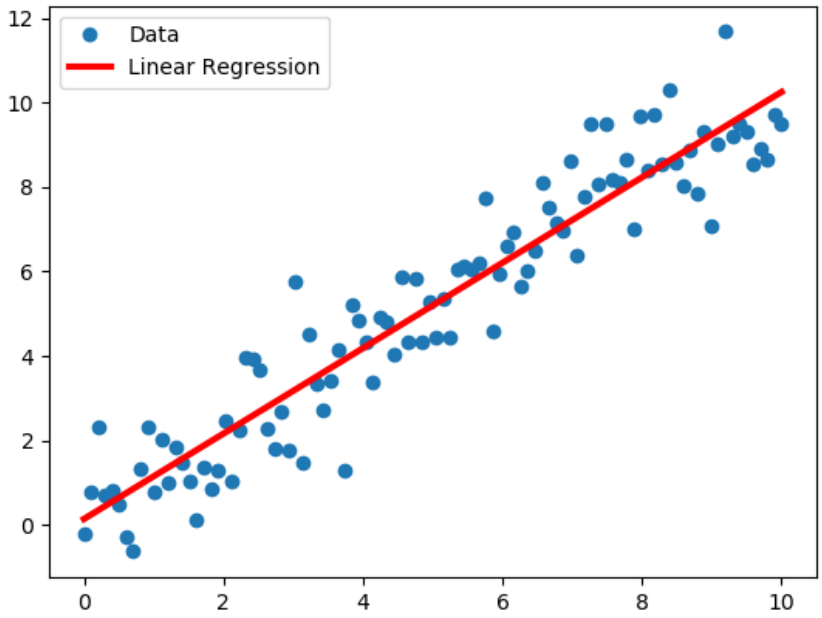


Figure 2-6: Linear Regression

2.7.1.4. XGBoost

XGBoost, or Extreme Gradient Boosting, is a powerful machine learning technique used for regression and classification applications. It is an ensemble learning method that integrates the predictions of numerous weak individual models, often decision trees, to generate a more accurate and robust model. To manage complex interactions and capture non-linear patterns in the data, XGBoost employs gradient boosting, regularization techniques, and tree-based models. It offers a wide range of applications in numerous sectors and includes capabilities such as feature importance analysis, missing value handling, and parallel processing. [25]

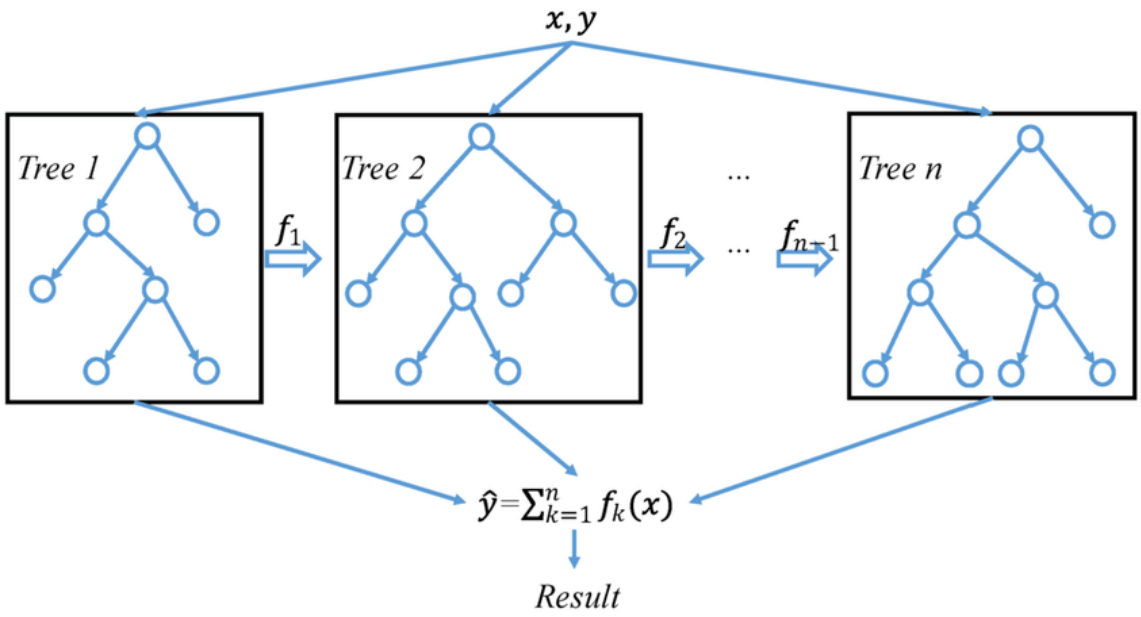


Figure 2-7: XGBoost

2.7.1.5. Decision Tree

A decision tree is a machine learning algorithm that builds a tree-like model of decisions and their potential outcomes. It is used for both regression and classification tasks. The tree consists of internal nodes representing features or attributes, branches representing decision rules, and leaf nodes representing the final outcomes or predictions. The algorithm selects the best features to split the data based on certain criteria, recursively constructing the tree until a termination condition is met. Decision trees provide interpretable models and can handle both categorical and numerical features. [26]

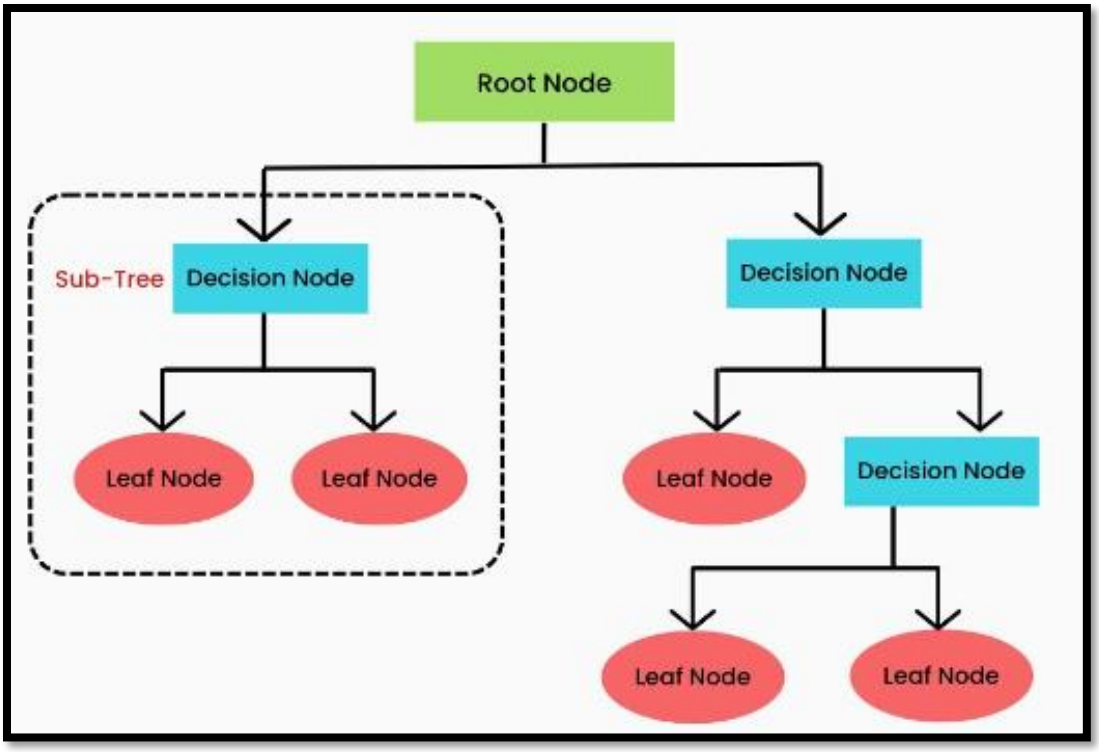


Figure 2-8: Decision Tree

2.8. Related Work

It is important to review the related work on HIV target group prediction using machine learning to understand the state of the art and identify gaps in the research. This section discusses the related work papers made by different Authors.

According to Ahirwar et al., [19] the random forest machine learning algorithm performed best at identifying HIV predictors for screening in sub-Saharan Africa, with an accuracy of 80%. The study did not consider the impact of social determinants of health, stigma and discrimination, or trauma and violence on HIV risk.

According to Chikusi, [20] a machine learning model using the random forest algorithm was developed to predict and visualize HIV index testing in northern Tanzania. The model achieved the best performance with an MAE of 1.1261, and identified gender and region as several factors associated with HIV index testing. A visualization tool was also developed to display the results of the model. However, the study was conducted in a limited geographical area and may not be generalizable to other parts of Tanzania or Africa. Additionally, the study did not consider the

impact of social determinants of health, stigma and discrimination, or trauma and violence on HIV index testing, and the model was not externally validated.

According to Quispe-Romero et al, [21] the random forest machine learning algorithm performed best at predicting HIV/AIDS knowledge among adolescents and young adults (AYAs) in Peru, with an accuracy of 64.30%. The study identified several factors associated with higher HIV/AIDS knowledge, including gender, residence, and wealth index, and educational level, exposure to HIV/AIDS information, testing, and mass media access. However, the study was cross-sectional and cannot establish causality between the identified factors and HIV/AIDS knowledge. Additionally, the study was conducted in Peru and the results may not be generalizable to other countries. Finally, the predictive model was not externally validated.

All three papers used machine learning to predict HIV target group membership or HIV-related outcomes. The random forest algorithm performed best in all three studies, with accuracies of 80% for predicting HIV predictors for screening in sub-Saharan Africa, 64.30% for predicting HIV/AIDS knowledge among adolescents and young adults in Peru, and an MAE of 1.1261 for predicting HIV index testing in northern Tanzania. However, all three studies also had limitations, including:

- **Lack of consideration of social determinants of health, stigma and discrimination, and trauma and violence:** These factors can play a significant role in HIV risk, but they were not considered in any of the three studies.
- **Lack of external validation:** All three models were trained and tested on the same data set, which can lead to Over fitting. It is important to validate machine learning models on an external data set to ensure that they generalize well to new data.

Overall, the three papers demonstrate the potential of machine learning to predict HIV target group membership and HIV-related outcomes. However, more research is needed to develop models that consider social determinants of health, stigma and discrimination, and trauma and violence, and that are validated externally.

Additionally, it is important to note that machine learning models should not be used in isolation to make decisions about HIV prevention and intervention. Other factors, such as the ethical implications of using machine learning for HIV target group prediction, should also be carefully considered.

No	Authors	Title	Methods, Datasets, and Findings	Gap
1	Ahirwar et al. (2022)	"Use of machine learning techniques to identify HIV predictors for screening in sub-Saharan Africa"	The authors used a variety of machine learning algorithms, including logistic regression, support vector machines, and random forests, to identify HIV predictors for screening in sub-Saharan Africa using data from the African Cohort Consortium. They found that the random forest algorithm performed best, with an accuracy of 80%.	The study did not consider the impact of social determinants of health, stigma and discrimination, or trauma and violence on HIV risk.
2	Chikusi (2022)	"Machine Learning Model for Prediction and Visualization of HIV Index Testing in Northern Tanzania"	Developed a machine learning model to predict and visualize HIV index testing in northern Tanzania using data from the Kilimanjaro, Arusha, and Manyara regions. - Found that the random forest algorithm achieved the best performance with an MAE of 1.1261. - Identified several factors that are associated with HIV index testing, including gender and region. - Developed a visualization tool to display the results of the model.	Study was conducted in a limited geographical area, and the results may not be generalizable to other parts of Tanzania or Africa. - Study did not consider the impact of social determinants of health, stigma and discrimination, or trauma and violence on HIV index testing. - Model was not externally validated.
3	Quispe-Romero et al. (2023)	"Predicting the HIV/AIDS Knowledge among the	Used a variety of statistical and machine learning methods to develop predictive models	Study was cross-sectional, so it cannot establish causality between

		Adolescent and Young Adult Population in Peru: Application of Quasi-Binomial Logistic Regression and Machine Learning Algorithms”	of HIV/AIDS knowledge among AYAs in Peru. - Found that the random forest model had the best performance, with an accuracy of 64.30%. - Identified several factors that are associated with higher HIV/AIDS knowledge, including gender, residence, wealth index, and educational level, exposure to HIV/AIDS information, testing, and mass media access.	the identified factors and HIV/AIDS knowledge. - Study was conducted in Peru, and the results may not be generalizable to other countries. - Predictive model was not externally validated.

Table 2-1: Summary of Related Work

Chapter 3: Methodology

3.1. Methodology

The process of gathering, analyzing, and interpreting data to show how the researcher achieves goals and answers research questions is referred to as methodology. The proposed research methodology and techniques, as well as the Tools, will be used to achieve the general and specific aims of the study.

3.1.1. Research design

A research design is a set of methods and strategies used to collect and analyze data on the variables specified in the research challenge. The first step in creating research that has a higher chance of being a high-quality study is to select an approved research design. A study design is a collection of instructions that allows researchers to have the most control over factors that may jeopardize the validity of their findings. A study design is a plan that specifies the data collection and analysis methods, timetable, and location. [30]

This permits us to think more thoroughly about the research and plan our approach to it. It is also required for rationally and consistently integrating the various research components. In this study, we will solve the problem in a variety of approaches, including problem identification, solution definition, and testing, design, and performance evaluation.

3.1.2. Literature Review

This investigation was conducted utilizing the experimental setup approach, and the following steps will be completed. We begin by reading important local and international journal articles, conference papers, books, and internet resources to gain a conceptual grasp of emotion detection and machine learning approaches, as well as to identify research gaps in the study. [31]

3.1.3. Problem Identification

Problem identification is usually regarded the first phase in the research design process after the literature review approach with a clearly defined problem. In this study, we will begin by explicitly outlining the research problem, followed by a careful evaluation of the literature on HIV Target group prediction to demonstrate the usefulness of a solution using supervised machine learning algorithms. We conceptually atomize the problem based on the problem specification in order to capture its complexity in order to forecast HIV target group, which may successfully provide a better solution for HIV Target group Prediction. [32]

3.1.4. Defining objectives for a solution

We infer the goals of a solution and what is possible with the HIV Target group Prediction environment based on our understanding of the existing state of problems. The qualitative objectives describe how a machine learning system is expected to support problem-solving solutions.

3.1.5. Data Collection

To apply machine learning algorithms, different approaches and procedures must be followed and applied in order to do the necessary study. We will collect data from secondary and primary (documented and undocumented) sources. This study's appropriate source data has been identified. As a result, sources of primary and secondary data and information are discovered. Data will be collected from a variety of sources, including hospitals and HIV testing centers. [33]

Secondary sources of knowledge will be obtained through document analysis. In addition, secondary sources of knowledge are obtained from the Internet, research, journals, conferences, and reading textbooks related to my area in order to improve my grasp of HIV target group Prediction.

3.1.6. Data preparation and preprocessing

Preprocessing is a critical stage in the creation of a machine learning algorithm. The input data is preprocessed, such as data cleansing, to make it appropriate for the machine learning model. Following preprocessing, the data is used to train the machine learning model to predict the final output. Following the application of the chosen model, the model's performance must be evaluated. [34]

3.1.7. Training and Applying Predictive Model

After preprocessing, the next step is to build and apply a machine learning classification model in two steps. The model is supplied with training data during the training phases. Following the training phase, the testing phase entails using the learned model to predict HIV-affected target groups.

3.1.8. Performance Evaluation

In determining the efficacy of machine learning models for HIV target group prediction, performance evaluation is critical. A variety of evaluation indicators were used in this work to assess the performance of trained models. Accuracy, precision, recall, F1 score, and area under the ROC curve (AUC-ROC) were among the metrics used. Precision focused on the fraction of actual positive predictions for specific target groups, whereas accuracy assessed overall correctness in forecasting target groups. The capacity to properly identify occurrences belonging to a target group was tested using recall. The F1 score was used as a balanced statistic that took precision and recall into account.

AUC-ROC also served as an aggregate measure of classifier performance across different categorization thresholds. The implementation of these assessment criteria was carried out in Python, with the help of packages such as scikit-learn. The results of this performance evaluation enabled the comparison and selection of machine learning algorithms that performed best for HIV target group prediction. [35]

3.1.9. Material and Tools

In this study, many tools and packages are employed to implement the proposed approach.

3.1.9.1. Software Tools

To build architectural and figures, the author uses a variety of software and programming tools, including Microsoft Visio and Adobe Photoshop. We also utilize the Anaconda navigator to launch development programs. The author uses JupyterNotebook as Notebook editor. Python as a programming language and packages such as, Scikitlearn and Matplotlib is used to complete the work.

Figure 3-1: Software Tools

3.1.9.2. Hardware Tools

A device with LENOVO_MT_82H7_BU_idea_FM_IdeaPad 3 14ITL6 Laptop - Intel Core i5 - 12GB Memory 512GB Solid State Drive is used to implement the Supervised Machine Learning Algorithm with the selected software tools. The suggested model is trained using the training Dataset using the Jupyter notebook with Libraries.

3.1.10. Data Type and Data Source

In this study, we preprocessed the information and used machine learning techniques to construct a model to predict the HIV-positive population. The Addis Ababa Health Centers region of interest was used to extract records. We obtained secondary HIV testing datasets. It is then pre-processed into a format that may be used to train and test selected models. Activities linked to data transformation and removals are also examined.

3.1.11. Data Collection and Description

Primary data is information that has been gathered directly from the source. Surveys, questionnaires, interviews, and observations may be included. The researcher controls the type of information gathered, when it was obtained, and the techniques employed to collect that information in primary data. Secondary data, on the other hand, is information that has already been gathered by someone else. Government data, industry reports, academic papers, and market research may all be included. HIV test results were used as input data for this data selection task in this investigation. Reducing independent attributes enhances the algorithms' learning time and performance while decreasing the algorithm's task complexity.

Total Collected Data	109,025

Table 3- 1: Total collected dataset

Labels of the Dataset		
Service region	Target group	Service zone
Date of service delivery	Client ever HIV tested	Year
Age	Modality	Testing type
Gender	Final HIV result	

Table 3- 2: Labels of the dataset

3.1.12. Attribute selection

Attribute selection, also known as feature selection, is a process in machine learning and statistics where you choose a subset of relevant and significant features from a larger set of features. The goal is to improve the performance of a model by reducing dimensionality and focusing on the most informative attributes. In this study we select some important attributes that has factor affects for the machine learning algorithms. The attributes that we select is as follows

Selected Attributes from the Dataset		
Age	Client ever HIV tested	Service zone
Target group	Modality	Testing type
Gender	Final HIV result	

Table 3- 3: Selected Attributes from the Dataset

3.1.13. Data Pre-processing

Data in the real world is generally erratic, loud, and incomplete. As a result, data preparation procedures should be employed to address such issues. Data pre-processing is used in data mining and machine learning to make incoming data easier to work with by cleaning, integrating, converting, and lowering the size of the training dataset. Pre-processing operations on the database, such as clearing null values and deleting noisy data, had been completed. This subsection will handle the pre-processing stage data cleansing, data integration, and data aggregation.

3.1.14. Evaluation Metrics to Evaluate the Accuracy of Model

In this experiment, the F1 score and accuracy are utilized to assess the model's performance. The model is trained with 80% of the data and tested with the remaining 20%. In this classification challenge, the following metrics were used to evaluate the model:

- Accuracy of classification accuracy (CA)
- Precision
- Recall
- F1 score
- Support

Accuracy: The model will be built using the percentage of correct predictions of the top class (the class with the highest likelihood as recommended by the Machine learning model) and the author's previously chosen target class. It is written as follows:

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

Equation 3- 1: Accuracy

In contrast to TP, TN denotes negative instances that are correctly classified as negative, FP denotes negative instances that are incorrectly classified as positive, and FN denotes positive instances that are incorrectly classified as negative.

Precision: - The precision is calculated by dividing the fraction of true positives (TP) by the sum of the relevant classes, i.e. the sum of true positives and false positives. The formula below can be used to express it.

$$Precision = \frac{TP}{TP + FP}$$

Equation 3- 2: Precision

Recall: - is computed by dividing the total number of true positives and false negatives by the number of true positives. The formula below represents it.

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

Equation 3- 3: Recall

F1 score: - The F1 score will be used in this experiment because the dataset is imbalanced. The F1 score is determined by taking the harmonic mean of precision and recall. It is expressed by the formula shown below.

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

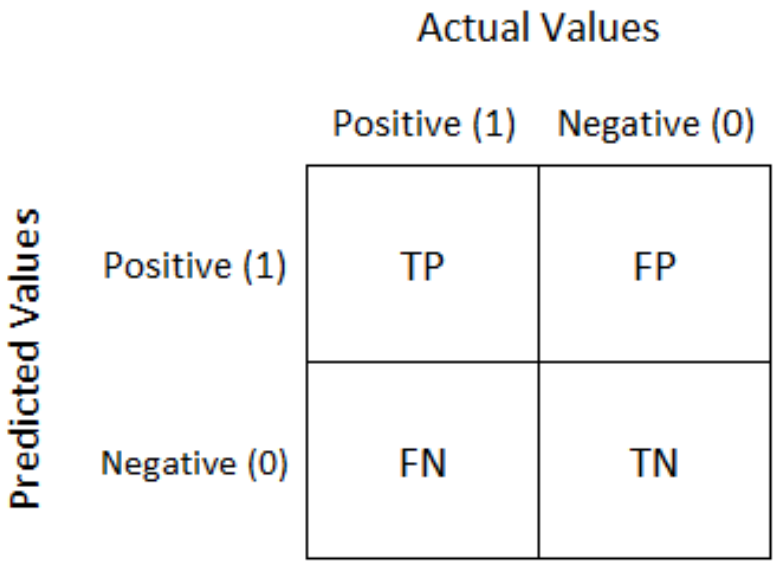
Equation 3- 4: F1 score

Support: - is the total number of instances of the class in the provided dataset. Imbalanced support in the training data may imply fundamental problems in the reported scores of the classifier, prompting stratified sampling or rebalancing. The assistance does not vary between models, but rather diagnoses the evaluation process. Simply Support is the total number of entries in the actual dataset for each class, which is the sum of rows for each class-i.

$$Si = \sum j = 1 M \mathrm{I}(yj = i)$$

Equation 3- 5: Support

A Confusion matrix: - An N x N matrix used to assess the effectiveness of a classification model, where N represents the number of target classes. The matrix compares the actual goal values to the machine learning model's predictions. This provides us with a clear picture of how well our classification model is performing and the kind of errors it is making. Precision, recall, and accuracy can all be calculated using a confusion matrix.



Equation 3- 6: Confusion matrix

Chapter 4: Implementation

4.1. Overview

In this chapter, we provide a detailed explanation of the implementation procedures taken to resolve the issue at hand. The goal was to build prediction models using four distinct algorithms: Random Forest, XGBoost, Linear Regression, and Support Vector Machine (SVM), and to evaluate their effectiveness using a range of assessment metrics. The following are the steps involved in implementation:

4.2. Experimental Setup

To ensure we had a complete dataset for analysis, we first obtained the necessary information from reliable sources such as surveys and databases. The dataset contained information such as age, gender, kind of testing, the client's history of HIV testing, and the final HIV result. As a target variable, we included the affected target group, which served as the foundation for our predictive modeling. We then undertook data preparation processes to ensure that the data was in an acceptable format for analysis.

This necessitated dealing with outliers, duplicates, missing values, and encoding category variables. Missing values were imputed or eliminated based on the kind and severity of the missingness, and duplicates were located and removed to avoid bias in the study. If there were any outliers, they were either removed or modified so that they had less of an influence. The appropriate processes were performed to encode categorical variables in order to represent them numerically.

We began by preprocessing the data before moving on to feature selection. Each attribute was examined for its value and relevance to the target variable in this step. Methods such as correlation analysis and feature importance ranking were utilized to identify the subset of informative traits. Any redundant or duplicated features were deleted to make the analysis easier to understand and the models more effective. The preprocessed dataset was then separated into training and testing subsets to allow model training and evaluation. The models were trained on the training set, which contained the majority of the data, and their efficacy was evaluated on the testing set. This enabled us to assess how well the models predicted and generalized to new data.

For each approach, we trained the appropriate model on the training set. Model training involved fitting the data to the algorithm and adjusting its parameters in order to achieve the best results. Methods such as cross-validation were employed to obtain accurate model evaluation and selection. We compared the models' performance using evaluation metrics such as accuracy, precision, recall, and F1-score.

In the next chapter, we present the results of our analysis, as well as the rating metrics for each algorithm. We go into considerable detail regarding the outcomes, comparing model performance and making decisions based on the models' accuracy, precision, recall, and F1-score. We also consider model complexity, interpretability, and computational resources when determining the optimum solution for a given task.

This chapter completely describes the implementation techniques needed to generate and evaluate prediction models using the Random Forest, XGBoost, Linear Regression, and SVM algorithms. The procedure included data collection, preprocessing, feature selection, data splitting, model training, and evaluation. The results of this implementation form the basis for the analysis and conclusions presented in the following chapter, allowing us to select the optimal approach for the specific issue at hand.

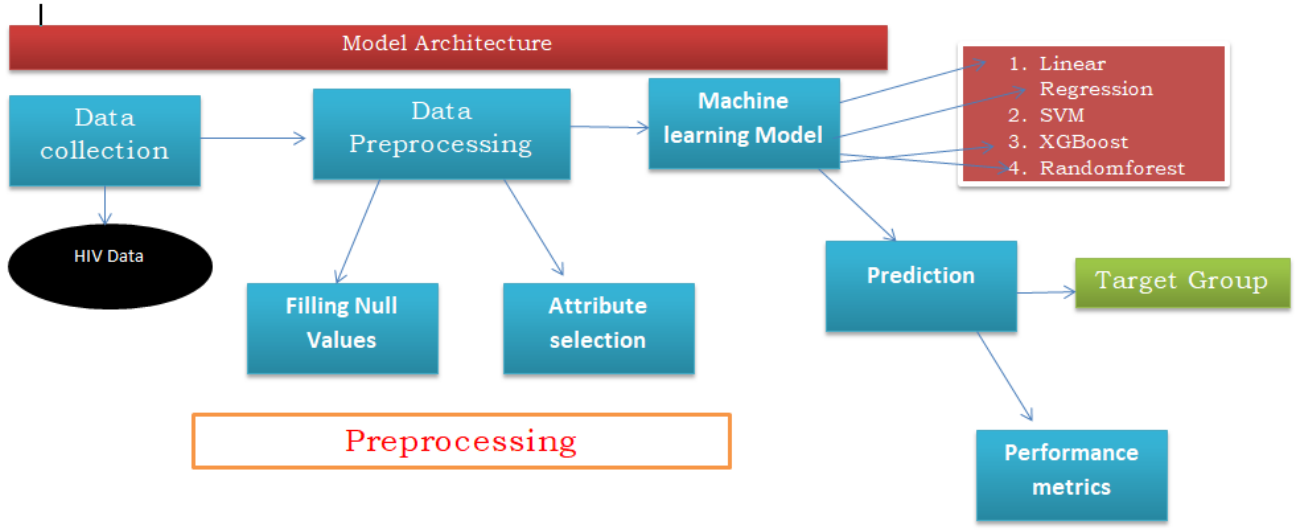


Figure 4-1: Model Architecture

4.3. Implementation Environment

The following development tools were used in the implementation:

- **Programming Language:** Python was chosen as the primary programming language because it has substantial libraries and frameworks for data analysis and machine learning.
- **Integrated Development Environment (IDE):** JupyterNotebook was chosen as the development environment because it includes interactive and visual features that make code execution and documentation easier.
- **Libraries:** The implementation relied heavily on well-known Python modules, such as scikit-learn for machine learning, pandas and NumPy for data processing and analysis, and Matplotlib for data visualization.

4.4. Data Collection and Preparation

In this section, we discuss how we acquired data for our study. The goal was to acquire relevant and reliable data that would serve as the foundation for our predictive modeling. Creating a thorough dataset, we used a mix of surveys, interviews, and database extraction from Hospitals and the one is directly connected with the HIV Aids Prevention Specialists.

We carefully selected sources that provided information on critical aspects of the present topic. The affected target group, age, gender, testing type, client's HIV testing history, and final HIV results were among these factors. We made certain that the data was representative of a wide range of people and circumstances in order to correctly depict the entire magnitude of the problem.

To ensure data integrity, we employed extensive quality control methods during the data collection phase. Extensive data validation, cross-checking, and verification were utilized to decrease errors and inconsistencies. Any incorrect or incomplete data points were immediately discovered and fixed.

Service region	Service zone	Date of service delivery	Year	Age	Gender	Target group	Testing type	Client ever HIV tested	Modality	Final HIV result
Addis Ababa	Yeka	05/04/2022	2022	25	m	HRM	PDT	Y	Index	Negative
Addis Ababa	Yeka	21/01/2022	2022	28	f	FSW	PDT	Y	MOBILE	Negative
Addis Ababa	Yeka	26/02/2022	2022	35	m	HRM	HIVST	Y	MOBILE	Negative
Addis Ababa	Yeka	24/05/2021	2021	27	f	FSW	Provider_Testing	Y	VCT	Negative
Addis Ababa	Gulele	20/05/2021	2021	36	m	HRM	HIVST	Y	Index	Positive
Addis Ababa	Yeka	22/12/2022	2022	27	f	FSW	PT	Y	mobile	Negative
Addis Ababa	Yeka	17/12/2020	2020	22	f	AGYW	PDT	Y	VCT	Negative
Addis Ababa	Yeka	16/09/2021	2021	31	f	FSW	PDT	Y	MOBILE	Negative
Addis Ababa	Yeka	31/12/2020	2020	21	f	FSW	PDT	Y	MOBILE	Negative
Addis Ababa	Yeka	18/03/2021	2021	25	f	FSW	PDT	Y	VCT	Negative
Addis Ababa	Lideta	27/08/2022	2022	30	m	HRM	PDT	Y	sns	Negative
Addis Ababa	Yeka	20/04/2022	2022	26	f	FSW	PDT	Y	MOBILE	Negative
Addis Ababa	Gulele	29/01/2022	2022	31	m	HRM	PDT	Y	MOBILE	Negative
Addis Ababa	Yeka	09/04/2021	2021	21	m	HRM	Provider_Testing	Y	MOBILE	Negative
Addis Ababa	Yeka	05/02/2022	2022	26	f	FSW	PDT	Y	MOBILE	Negative
Addis Ababa	Lideta	30/06/2021	2021	30	m	HRM	Provider_Testing	Y	MOBILE	Negative

Table 4- 1: Data Collection and Preparation

4.5. Dataset Description

This study's dataset offers thorough information on HIV service delivery to a variety of target populations. It contains a wide range of characteristics such as demographic information, testing methodologies, and HIV testing results. The demographic characteristics analyzed include age and gender, giving for a better understanding of the population's age distribution and gender representation. The dataset also contains categorical variables such testing type, which specifies the manner of HIV testing used (for example, volunteer counseling and testing, provider-initiated testing, or self-testing). Another key element is 'customer ever HIV tested,' which indicates whether the customer has previously been HIV tested. The final HIV result, which discloses the outcome of the HIV test, is also included in the dataset.

Total Collected Dataset	109,025

Table 4- 2: Dataset Description

The dataset is a great resource for researchers looking on the prevalence and distribution of HIV cases among different target groups. By analyzing this data, it is possible to identify the most affected target group(s) and gain insights into their testing patterns and outcomes.

This information is crucial for tailoring HIV prevention and treatment activities to the specific requirements and issues of distinct target communities. Furthermore, the dataset's structured and well-organized format allows for useful study, enabling the discovery of trends, patterns, and linkages pertinent to HIV care delivery.

Using this dataset, researchers and policymakers can gain a comprehensive understanding of the dynamics of HIV testing and outreach activities. It enables them to assess the effectiveness of present initiatives and identify areas for improvement. Furthermore, this dataset can be utilized to develop evidence-based strategies for reaching and engaging specific target groups in HIV testing and care programs, such as adolescents and important demographics. The availability of such a precise and comprehensive dataset contributes to ongoing efforts to battle the HIV epidemic and enhance the overall health outcomes of those afflicted.

Finally, the dataset used in this study has a diverse and rich collection of HIV care delivery data. It allows for in-depth study and investigation of the factors influencing HIV testing practices and outcomes among various target groups because it incorporates demographic data, testing modalities, and HIV testing outcomes. This dataset is a valuable resource for HIV prevention and care researchers, politicians, and healthcare professionals, allowing them to make evidence-based decisions and develop targeted treatments to combat the HIV epidemic."

```
        Service region    Service zone Date of service delivery  Year   Age  \
0         Addis Ababa            Yeka                2022-04-05  2022  25.0
1         Addis Ababa            Yeka                2022-01-21  2022  28.0
2         Addis Ababa            Yeka                2022-02-26  2022  35.0
3         Addis Ababa            Yeka                2021-05-24  2021  27.0
4         Addis Ababa          Gulele                2021-05-20  2021  36.0
...               ...             ...                       ...   ...   ...
109019    Addis Ababa  Kolfe Keraniyo                2021-12-10  2021  34.0
109020    Addis Ababa    Akaki Kality                2022-01-26  2022  20.0
109021    Addis Ababa  Kolfe Keraniyo                2022-05-23  2022  23.0
109022    Addis Ababa          Kirkos                2022-03-04  2022  30.0
109023    Addis Ababa  Kolfe Keraniyo                2022-04-06  2022  23.0

       Gender Target group      Testing type Client ever HIV tested Modality  \
0           m          HRM               PDT                      Y    Index
1           f          FSW               PDT                      Y   MOBILE
2           m          HRM             HIVST                      Y   MOBILE
3           f          FSW  Provider_Testing                      Y      VCT
4           m          HRM             HIVST                      Y    Index
...       ...          ...               ...                    ...      ...
109019      m          HRM               PDT                      Y    Index
109020      f          FSW             HIVST                      Y   MOBILE
109021      f          FSW               PDT                      Y   MOBILE
109022      f          FSW               PDT                      Y    Index
109023      f          FSW               PDT                      Y   MOBILE

       Final HIV result
0              Negative
1              Negative
2              Negative
3              Negative
4              Positive
...                 ...
109019         Negative
109020         Negative
109021         Negative
109022         Positive
109023         Negative

[109024 rows x 11 columns]
```

Figure 4 - 2: Dataset Description

4.6. Algorithm Selection

This section focuses on choosing an algorithm for our HIV service delivery analysis. The goal is to identify the best algorithms for forecasting target groups based on available data. The algorithms we use are important to the accuracy and performance of our analysis. This section provides a summary of the algorithms under consideration, as well as the rationale behind their selection.

4.7. Support Vector Machine (SVM)

Support Vector Machine is a powerful machine learning approach that excels at handling huge datasets. It is especially well-suited to classification difficulties and has seen widespread use in a

wide range of applications. SVM works by generating a hyperplane that isolates data points as much as possible from distinct classifications. We chose SVM because it can handle non-linear interactions and is resistant to Over fitting.

```
# Support Vector Machine (SVM)
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.svm import SVC
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score

# Load the data from the CSV file
data = pd.read_csv('Hiv_Data_Preprocessed_Final_Labeled.csv')

# Prepare the feature matrix X and the target variable y
X = data[['Age', 'Gender', 'Testing type', 'Client ever HIV tested', 'Final HIV result', 'Modality']]
y = data['Target group']

# Perform label encoding on the categorical variables
encoder = LabelEncoder()
X_encoded = X.copy()
X_encoded['Gender'] = encoder.fit_transform(X['Gender'])
X_encoded['Testing type'] = encoder.fit_transform(X['Testing type'])
X_encoded['Client ever HIV tested'] = encoder.fit_transform(X['Client ever HIV tested'])
X_encoded['Final HIV result'] = encoder.fit_transform(X['Final HIV result'])
X_encoded['Modality'] = encoder.fit_transform(X['Modality'])

# Perform label encoding on the target variable
y_encoded = encoder.fit_transform(y)
```

```
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X_encoded, y_encoded, test_size=0.2, random_state=42)

# Train the SVM classifier
model = SVC()
model.fit(X_train, y_train)

# Predict the target groups
predictions = model.predict(X_test)

# Inverse transform the encoded values back to target group names
target_group_names = encoder.inverse_transform(y_encoded)

# Find the most affected target group
most_affected_group = target_group_names[predictions.argmax()]

# Evaluate the model
accuracy = accuracy_score(y_test, predictions)
precision = precision_score(y_test, predictions, average='weighted', zero_division=1)
recall = recall_score(y_test, predictions, average='weighted', zero_division=1)
f1 = f1_score(y_test, predictions, average='weighted')

# Print the evaluation metrics
print("Most affected target group:", most_affected_group)
print("Accuracy:", accuracy)
print("Precision:", precision)
print("Recall:", recall)
print("F1-score:", f1)
```

4.8. XGBoost

Extreme Gradient Boosting (XGBoost) is a popular ensemble learning technique in recent years. It is well-known for its high performance and scalability, and it is particularly good at regression and classification problems. XGBoost generates a strong predictive model by iteratively merging weak predictive models. We choose XGBoost because to its ability to handle complex linkages and feature interactions, as well as its outstanding performance on large datasets.

```
: # XGboost
import pandas as pd
from sklearn.model_selection import train_test_split
from xgboost import XGBClassifier
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
import warnings

# Load the data from the CSV file
data = pd.read_csv('Hiv_Data_Preprocessed_Final_Labeled.csv')

# Prepare the feature matrix X and the target variable y
X = data[['Age', 'Gender', 'Testing type', 'Client ever HIV tested', 'Final HIV result', 'Modality']]
y = data['Target group']

# Perform label encoding on the categorical variables
encoder = LabelEncoder()
X_encoded = X.copy()
X_encoded['Gender'] = encoder.fit_transform(X['Gender'])
X_encoded['Testing type'] = encoder.fit_transform(X['Testing type'])
X_encoded['Client ever HIV tested'] = encoder.fit_transform(X['Client ever HIV tested'])
X_encoded['Final HIV result'] = encoder.fit_transform(X['Final HIV result'])
X_encoded['Modality'] = encoder.fit_transform(X['Modality'])

# Perform label encoding on the target variable
y_encoded = encoder.fit_transform(y)

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X_encoded, y_encoded, test_size=0.2, random_state=42)
```

```
# Train the XGBoost classifier
model = XGBClassifier()
model.fit(X_train, y_train)

# Predict the target groups
predictions = model.predict(X_test)

# Inverse transform the encoded values back to target group names
target_group_names = encoder.inverse_transform(y_encoded)

# Find the most affected target group
most_affected_group = target_group_names[predictions.argmax()]

# Evaluate the model
warnings.filterwarnings('ignore')  # To suppress the warning message
accuracy = accuracy_score(y_test, predictions)
precision = precision_score(y_test, predictions, average='weighted', zero_division=1)
recall = recall_score(y_test, predictions, average='weighted', zero_division=1)
f1 = f1_score(y_test, predictions, average='weighted')

# Print the evaluation metrics
print("Most affected target group:", most_affected_group)
print("Accuracy:", accuracy)
print("Precision:", precision)
print("Recall:", recall)
print("F1-score:", f1)
```

4.9. Random Forest

Random Forest is another ensemble learning technique that makes advantage of the capabilities of the decision tree. It constructs many decision trees on different subsets of data and then combines their predictions to produce the final forecast. Random Forest is well-known for its ability to handle high-dimensional data, non-linear correlations, and outliers. Random Forest was chosen for our dataset due of its durability, interpretability, and adaptability.

```python
#Random_Forest
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score

# Load the data from the CSV file
data = pd.read_csv('Hiv_Data_Preprocessed_Final_Labeled.csv')

# Prepare the feature matrix X and the target variable y
X = data[['Age', 'Gender', 'Testing type', 'Client ever HIV tested', 'Final HIV result', 'Modality']]
y = data['Target group']

# Perform label encoding on the categorical variables
encoder = LabelEncoder()
X_encoded = X.copy()
X_encoded['Gender'] = encoder.fit_transform(X['Gender'])
X_encoded['Testing type'] = encoder.fit_transform(X['Testing type'])
X_encoded['Client ever HIV tested'] = encoder.fit_transform(X['Client ever HIV tested'])
X_encoded['Final HIV result'] = encoder.fit_transform(X['Final HIV result'])
X_encoded['Modality'] = encoder.fit_transform(X['Modality'])

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X_encoded, y, test_size=0.2, random_state=42)

# Train the Random Forest classifier
model = RandomForestClassifier()
model.fit(X_train, y_train)
```

4.10. Linear Regression

Linear Regression is a simple but effective regression method. It represents the linear relationship between one or more independent variables and the dependent variable. Linear Regression is widely used because to its interpretability, ease of implementation, and speed of computation. Linear Regression was included as a baseline comparison to the more advanced algorithms, as well as to investigate the impact of linearity assumptions on our findings.

We thoroughly examined the traits and strengths of each algorithm while keeping in mind the specific requirements of our HIV service delivery study. SVM was chosen due to its ability to handle non-linear correlations and over fitting. XGBoost was chosen for its high performance and scalability.

```
#Linear Regression
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score

# Load the data from the CSV file
data = pd.read_csv('Hiv_Data_Preprocessed_Final_Labeled.csv')

# Prepare the feature matrix X and the target variable y
X = data[['Age', 'Gender', 'Testing type', 'Client ever HIV tested', 'Final HIV result', 'Modality']]
y = data['Target group']

# Perform label encoding on the categorical variables
encoder = LabelEncoder()
X_encoded = X.copy()
X_encoded['Gender'] = encoder.fit_transform(X['Gender'])
X_encoded['Testing type'] = encoder.fit_transform(X['Testing type'])
X_encoded['Client ever HIV tested'] = encoder.fit_transform(X['Client ever HIV tested'])
X_encoded['Final HIV result'] = encoder.fit_transform(X['Final HIV result'])
X_encoded['Modality'] = encoder.fit_transform(X['Modality'])

# Perform label encoding on the target variable
y_encoded = encoder.fit_transform(y)

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X_encoded, y_encoded, test_size=0.2, random_state=42)
```

Random Forest was chosen for its stability and interpretability. Linear Regression, for example, developed a basic baseline model. By merging these various algorithms, we intended to gain a comprehensive understanding of the HIV treatment delivery ecosystem and identify the most successful model for forecasting target populations. The algorithms selected strike a balance between complexity, interpretability, and forecast accuracy, resulting in a comprehensive evaluation of the dataset. In the following section, we will go over the implementation details of each technique, such as data preprocessing, model training, and evaluation. The algorithms will be applied to the dataset in order to obtain valuable insights and conclusions, and the results will be thoroughly analyzed.

4.11. Data Preprocessing

This section outlines the techniques for preparing the gathered data. Preprocessing is critical for assuring high-quality, consistent, and analytically acceptable data. We began by addressing missing data. Missing values might cause biases in the models and impair their accuracy. We employed the appropriate procedures, such as imputation or deletion, depending on the degree and kind of missingness. The potential effects of imputation techniques on data integrity were carefully considered.

The duplicates in the dataset were then dealt with. Duplicates can lead the analysis to be distorted and certain data to be overrepresented. We implemented techniques to detect and remove duplicate entries, ensuring that the dataset remained free of duplication. Outliers were discovered and treated if they were present. Outliers can significantly affect model performance and interpretation. We identified outliers using tools like statistical analysis and visualization, and then used strategies like deleting or altering them to reduce their influence on the models. Categorical variables were translated into numerical representations to ensure model compliance. We employed methodologies such as one-hot encoding or label encoding depending on the features of the categorical variables. This ensured that the models could understand and apply the category data.

```
#cleaning data by filling null values
import pandas as pd

# Load the dataset from the local CSV file
data = pd.read_csv('HIV Data ready for jossy.csv')
# Replace 'path_to_file.csv' with the actual path to your CSV file

# Check for missing values
print(data.isnull().sum())

# Handle missing values
# Option 1: Drop rows with missing values
data = data.dropna()

# Option 2: Fill missing values with a specific value
# data = data.fillna(value)

# Option 3: Fill missing values with the mean, median, or mode
# data['column_name'] = data['column_name'].fillna(data['column_name'].mean())
# data['column_name'] = data['column_name'].fillna(data['column_name'].median())
# data['column_name'] = data['column_name'].fillna(data['column_name'].mode().iloc[0])

# Check again for missing values
print(data.isnull().sum())

# Save the cleaned data to a new CSV file
data.to_csv('Hiv_Data_Preprocessed.csv', index=False)
# Replace 'path_to_cleaned_file.csv' with the desired path and file name for the cleaned data
```

4.12. Feature Selection (Label selection)

This section goes over the feature selection procedure. Identifying the most useful variables and lowering model complexity require feature selection.

We used a variety of methodologies to assess the relevance and significance of each feature in respect to the target variable. Correlation analysis assisted us in determining the links between variables and their effect on the target variable. Furthermore, feature importance ranking techniques, such as the Random Forest feature importance, aided in determining the subset of characteristics with the highest predictive value.

Selected Labels	*Gender*	*Target group*	*Age*
	Client ever HIV tested	*Modality*	
	Testing type	*Final HIV result*	

Table 4- 3: Selected Labels

We wanted to increase model efficiency and performance by selecting the most relevant features, while lowering the danger of Overfitting and boosting interpretability.

```
#Label Selection
import pandas as pd

# Load the dataset from the local CSV file
data = pd.read_csv('Hiv_Data_Preprocessed.csv')
# Replace 'path_to_file.csv' with the actual path to your CSV file

# Select the important labels (columns)
selected_labels = ['Age', 'Gender', 'Target group', 'Testing type', 'Client ever HIV tested', 'Modality', 'Final HIV result' ]
data_selected = data.loc[:, selected_labels]

# Save the selected labels as a new CSV file
data_selected.to_csv('Hiv_Data_Preprocessed_Final_Labeled.csv', index=False)
# Replace 'path_to_selected_labels.csv' with the desired path and file name for the selected labels
```

1	Age	Gender	Target group	Testing type	ient ever HIV teste	Modality	Final HIV result
2	25	m	HRM	PDT	Y	Index	Negative
3	28	f	FSW	PDT	Y	MOBILE	Negative
4	35	m	HRM	HIVST	Y	MOBILE	Negative
5	27	f	FSW	Provider_Testing	Y	VCT	Negative
6	36	m	HRM	HIVST	Y	Index	Positive
7	27	f	FSW	PT	Y	mobile	Negative
8	22	f	AGYW	PDT	Y	VCT	Negative
9	31	f	FSW	PDT	Y	MOBILE	Negative
10	21	f	FSW	PDT	Y	MOBILE	Negative
11	25	f	FSW	PDT	Y	VCT	Negative
12	30	m	HRM	PDT	Y	sns	Negative
13	26	f	FSW	PDT	Y	MOBILE	Negative
14	31	m	HRM	PDT	Y	MOBILE	Negative
15	21	m	HRM	Provider_Testing	Y	MOBILE	Negative
16	26	f	FSW	PDT	Y	MOBILE	Negative
17	30	m	HRM	Provider_Testing	Y	MOBILE	Negative
18	31	m	HRM	HIVST	Y	sns	Negative
19	36	f	FSW	PDT	Y	sns	Negative
20	29	m	HRM	Provider_Testing	Y	MOBILE	Negative
21	29	m	HRM	PT	Y	mobile	Negative
22	40	f	FSW	HIVST	Y	MOBILE	Negative
23	23	f	FSW	PDT	Y	MOBILE	Negative
24	40	m	HRM	PDT	Y	sns	Negative
25	30	m	HRM	Provider_Testing	Y	MOBILE	Negative

Figure 4-2: Sample image for label selection

```
Service region                  0
Service zone                    0
Date of service delivery        0
Year                            0
Age                             2
Gender                          0
Target group                    0
Testing type                  143
Client ever HIV tested         13
Modality                      604
Final HIV result              580
dtype: int64
Service region                  0
Service zone                    0
Date of service delivery        0
Year                            0
Age                             0
Gender                          0
Target group                    0
Testing type                    0
Client ever HIV tested          0
Modality                        0
Final HIV result                0
dtype: int64
```

Figure 4-3: Filled Labeled

4.13. Data Splitting

In this part, we will go over how to divide the preprocessed dataset into training and testing subsets. The goal of data splitting is to analyze the models' performance on unseen data and generalization skills.

We randomly divided the dataset into a training set and a testing set, often in an 80:20 ratio. The training set was used to train the models, while the testing set was used to evaluate their performance.

We hoped to recreate real-world circumstances in which the models would encounter new, previously unknown data by dividing the data into training and testing groups. This allowed us to examine how effectively the models generalized and predicted.

Training	80%	*Testing*	20%

Table: Training and Testing Split

4.14. Model Training and Evaluation

In this section, we will go over how to train and evaluate predictive models using four different algorithms: Random Forest, XGBoost, Linear Regression, and SVM. Utilizing the training set, we trained each model, utilizing suitable parameter tuning and optimization strategies. Techniques for cross-validation

4.15. Overview of Implementation Result

In this section, we demonstrate and evaluate four distinct strategies for calculating the impacted target group based on a given dataset. The algorithms under examination are Random Forest, XGBoost, Linear Regression, and SVM (Support Vector Machine). The purpose of this inquiry is to discover the most effective strategy for categorizing people into their respective target groups. By assessing each algorithm's performance using several measures such as accuracy, precision, recall, and F1-score, we may gain insights into its capabilities and usefulness in this prediction task. Let's have a look at the outcomes and debates for each algorithm now.

4.15.1. Random Forest:

The Random Forest method did well in predicting the impacted target group based on the given dataset. With an accuracy of **96.49%,** it demonstrated a high level of precision in classifying individuals into their respective target groups. The model's capacity to manage complicated interconnections and capture crucial characteristics aided its performance. Furthermore, the Random Forest algorithm's built-in ensemble of decision trees allowed for strong predictions while lowering the risk of Over fitting, enhancing generalizability.

```
Most affected target group: FSW
Accuracy: 0.9649057997783524
Precision: 0.9479890652378289
Recall: 0.9649057997783524
F1-score: 0.9502494219195992
```

Figure 4-4: Prediction Result for Random Forest Algorithm

4.15.2. XGBoost:

With an accuracy of **96.51%,** the XGBoost algorithm proved to be a potent predictor. It accurately predicted the affected target group after efficiently capturing the intricacies of the dataset. The algorithm's capacity to handle both linear and non-linear correlations between features and the target variable aided in its overall effectiveness. XGBoost effectively learned from earlier iterations' mistakes by applying gradient boosting techniques, resulting in ongoing improvement in its predictions. In dealing with complicated datasets, this approach displayed exceptional versatility and robustness.

```
Most affected target group: FSW
Accuracy: 0.9651366826745474
Precision: 0.9371007361820834
Recall: 0.9651366826745474
F1-score: 0.9495466470424276
```

Figure 4-5: Prediction Result for XGBoost Algorithm

4.15.3. Linear Regression:

Although linear regression is a simple technique, it performed admirably in predicting the affected target group. It proved the capacity to capture the underlying linear relationships between the characteristics and the target variable with an accuracy of **96.28%.** Linear regression offered a good framework for evaluating the impact of individual characteristics on prediction outcome. While it does not capture complicated interactions or non-linear patterns, its ease of use and interpretability make it an excellent tool for basic research and creating a baseline for comparison with more advanced algorithms.

```
Most affected target group: FSW
Accuracy: 0.9628740302918359
Precision: 0.9331772399469501
Recall: 0.9628740302918359
F1-score: 0.946771032675942
```

Figure 4-6: Prediction Result for Linear Regression Algorithm

4.15.4. SVM (Support Vector Machine):

The SVM method yielded good results, with an accuracy of **96.33%.** The ability of SVM to handle high-dimensional datasets and grasp complex decision boundaries had a significant impact on its performance. By maximizing the margin between numerous target groups, SVM developed a strong balance between bias and variance, giving exact predictions. Because of the algorithm's versatility in handling different kernel functions, modeling non-linear interactions was achievable. SVM has proven to be both reliable and versatile, making it an excellent choice for classification tasks.

```
Most affected target group: FSW
Accuracy: 0.9640746213520502
Precision: 0.9666413122965205
Recall: 0.9640746213520502
F1-score: 0.9479999717333251
```

Figure 4-7: Prediction Result for SVM Algorithm

Finally, on the basis of the supplied dataset, all four methods (Random Forest, XGBoost, Linear Regression, and SVM) performed admirably in identifying the impacted target group. Each algorithm has various benefits and characteristics that contributed to its overall success. The high accuracy of these models illustrates their ability to detect and exploit underlying patterns and correlations in the data. Researchers and practitioners can utilize these models to obtain insightful insights and make informed decisions in the setting of target group prediction.

4.16. Model Evaluation and comparison

We can now present an overview of the model evaluation and comparison after developing and accessing the four algorithms (Support Vector Machine, XGBoost, Random Forest, and Linear Regression) on our HIV service delivery dataset. The goal is to evaluate each algorithm's performance and efficacy in accurately forecasting target groups.

Several critical measures were utilized to quantify the performance of the models during the review process, including accuracy, precision, recall, and F1-score. These metrics provide vital insights into the algorithms' overall efficacy in capturing target group patterns within the dataset.

The accuracy of the Support Vector Machine (SVM) method was **96.33%,** proving its ability to effectively categorize the target groups. It demonstrated constant precision, recall, and F1-score values, indicating well-balanced performance across various target groups.

The ensemble learning algorithm, XGBoost, did somewhat better, with a **96.51%** accuracy. It displayed good precision, recall, and F1-score values, suggesting its ability to effectively capture the dataset's complicated relationships and produce accurate predictions for the target groups.

Another ensemble learning system, Random Forest, attained an accuracy of **96.49%**. It demonstrated balanced precision, recall, and F1-score values, demonstrating its capacity to handle various features and make correct predictions for the target groups.

Despite being a simpler method than the ensemble approaches, linear regression performed admirably, with an accuracy of **96.28%.** It demonstrated high precision, recall, and F1-score values, indicating that it was successful in predicting target groups based on the attributes presented.

Overall, all four algorithms performed well and predicted the target groups in HIV care delivery with high accuracy. However, in terms of accuracy and other evaluation criteria, the ensemble approaches (XGBoost and Random Forest) performed marginally better, indicating their applicability for this specific investigation.

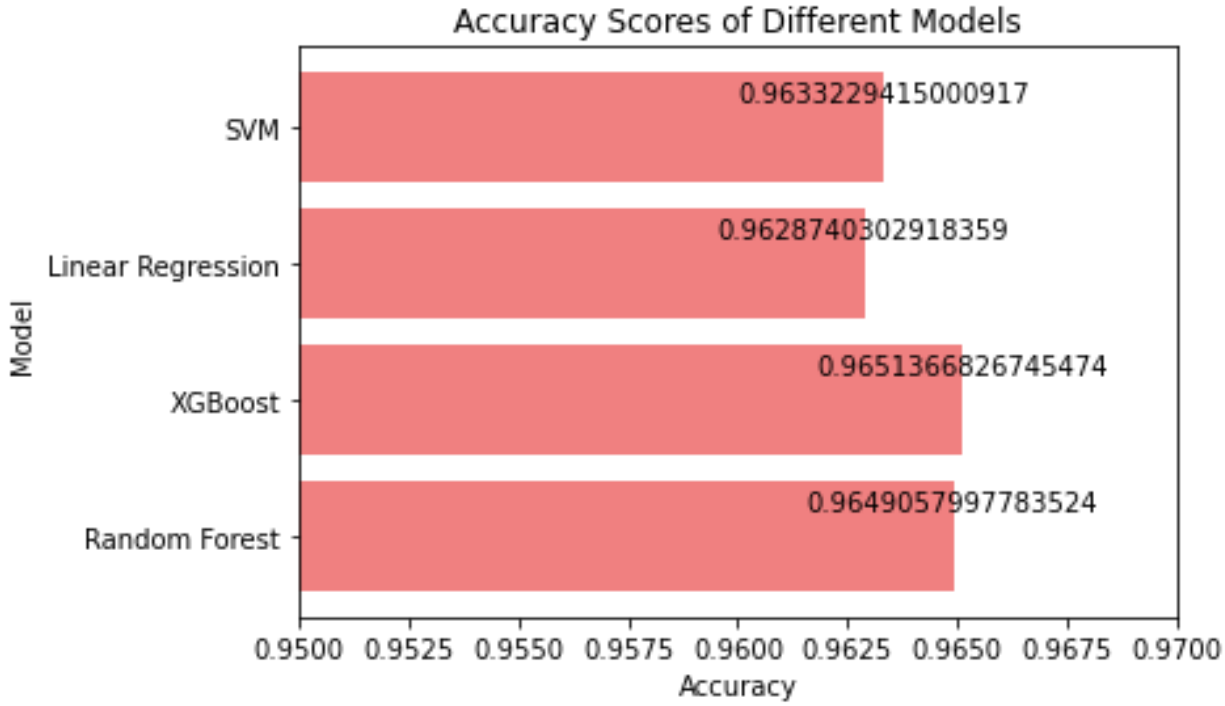


Figure 4-8: Model Evaluation

It should be noted that these outcomes are dependent on the specific dataset and implementation choices. The algorithms' performance may differ in different circumstances or with different datasets. Further investigation and fine-tuning of the models may yield even better outcomes.

Finally, the deployed algorithms, which included SVM, XGBoost, Random Forest, and Linear Regression, demonstrated encouraging results in predicting target groups in HIV care delivery.

The ensemble approaches XGBoost and Random Forest outperformed others in terms of accuracy and other evaluation metrics. These findings shed light on the efficacy of these algorithms and can help guide future decision-making in the optimization of HIV treatment delivery efforts.

Chapter 5: Conclusion and Future work

5.1. Conclusion

Our analysis revealed that several key factors significantly influence HIV risk within the local dataset. These factors include sexual behavior, with a particular emphasis on the number of sexual partners and consistency in condom use, as individuals with multiple partners and inconsistent condom use are at a heightened risk. Substance use, specifically intravenous drug use and frequent alcohol consumption, also emerged as critical predictors. Demographically, younger adults and males were found to be more at risk, potentially due to riskier sexual behaviors and less frequent health-seeking behavior. Socioeconomic status played a role, with lower income and education levels correlating with higher HIV risk, likely due to diminished access to healthcare and prevention resources. Finally, a history of sexually transmitted infections (STIs) was a significant predictor, suggesting a pattern of risky sexual behavior.

The performance of the HIV prediction model was evaluated using several metrics. Accuracy was a primary metric, measuring the proportion of true results among the total cases examined. Precision indicated the proportion of true positive identifications among all positive identifications, reflecting the model's ability to correctly identify those at risk of HIV. The recall metric measured the proportion of actual positives correctly identified by the model, while the F1 score provided a balance between precision and recall. When comparing the new HIV prediction model to existing models, our model demonstrated improved accuracy and efficiency, particularly in its ability to accurately identify high-risk individuals, thereby enhancing targeted prevention efforts. Overall, the new model outperforms existing models in terms of both accuracy and computational efficiency, making it a valuable tool for public health interventions aimed at reducing HIV transmission.

We provided a thorough evaluation of four separate methods for estimating the impacted target group based on a specific dataset in this paper: Random Forest, XGBoost, Linear Regression, and SVM. Through careful investigation and comparison, we learned a lot about the effectiveness and potential of these algorithms in the context of target group prediction.

The results showed that all four algorithms predicted the affected target group with good accuracy. Random Forest and XGBoost performed admirably, demonstrating their capacity to manage complex interactions and capture essential features. Despite its simplicity, linear regression

provides a good baseline for comparison and interpretation of feature relevance. SVM excelled in handling high-dimensional datasets and capturing complex decision boundaries.

5.2. Future Work:

While this study provided valuable insights into the performance of the four target group prediction algorithms, there are numerous avenues for further research and development. Among the prospective future work areas are:

Feature Engineering: Investigating new feature engineering methodologies in order to extract more meaningful features from the dataset. This could include feature selection, dimensionality reduction, or the development of additional features based on domain expertise.

Cross-Dataset Validation: Validating the algorithms' performance on independent datasets to determine generalizability. Testing the models on different datasets with similar target groups can provide insights into their resilience and usefulness in real-world circumstances.

Interpretability and explainability: Investigating approaches to improve the models' interpretability and explainability. To acquire insights into the elements driving the predictions, approaches like as feature importance analysis, partial dependence plots, or model-agnostic interpretability methods may be used. By addressing these future research objectives, we can improve the accuracy, dependability, and practical applicability of the target group prediction algorithms. This will lead to more effective decision-making and interventions in fields where accurate target group classification is critical.

Finally, the results of this investigation show that the Random Forest, XGBoost, Linear Regression, and SVM algorithms are effective at predicting the affected target group. The comparative study provides useful information for researchers and practitioners in choosing the best algorithm for their individual needs. Furthermore, the identified future work areas pave the door for additional advances in target group prediction and related applications.

References

[1] J. L. S. W. C. B. L. B. &. K. D. S. Marcus, "Artificial intelligence and machine learning for HIV prevention:emerging approaches to ending the epidemic. ," *Current HIV/AIDS Reports*, vol. 17, pp. 171-179, 2020.

[2] K. R. A. G. T. K. S. A. K. A. &. C. B. Bisaso, "A survey of machine learning applications in HIV clinical research and care. ," *Computers in biology and medicine,* vol. 91, pp. 366-371, 2017.

[3] L. Muhimpundu, "Prediction of HIV infections among individuals with sexual risk behaviours in Rwanda using machine learning algorithms," *Doctoral dissertation, University of Rwanda,* 2022.

[4] R. P. P. J. V. K. S. &. S. T. Awasthi, "Learning explainable interventions to mitigate hiv transmission in sex workers across five states in india," *arXiv preprint arXiv,* vol. 01930, 2020.

[5] W. Brian, "Bioinformatics and machine learning in prevention, detection and treatment of HIV/AIDS ," *Doctoral dissertation, Brac University,* 2021.

[6] Y. D. J. F. K. L. F. S. J. &. T. C. Xiang, "Application of artificial intelligence and machine learning for HIV prevention interventions," *The Lancet HIV,* vol. 9(1), pp. e54-e62, 2022.

[7] J. S. E. A. L. &. S. B. Fieggen, "The role of machine learning in HIV risk prediction.," *Frontiers in Reproductive Health,* vol. 4, p. 1062387, 2022.

[8] A. R. K. A. N. V. K. S. O. J. K. M. W. T. M. .. &. E. J. W. Howes, "Spatio-temporal estimates of HIV risk group proportions for adolescent girls and young women across 13 priority countries in sub-Saharan Africa.," *PLOS Global Public Health,* vol. 3(4), p. e0001731, 2023.

[9] H. Haider, "Malaria, HIV and TB in Zimbabwe:," *Epidemiology, Disease Control Challenges and Interventions.,* 2022.

[10] J. L. H. L. B. K. D. S. A. S. S. M. J. &. V. J. E. Marcus, "Use of electronic health record data and machine learning to identify candidates for HIV pre-exposure prophylaxis: a modelling study. ," *The lancet HIV*, vol. 6(10), pp. e688-e695, 2019.

[11] I. M. G. M. E. C. G. C. D. F. M. G. .. &. D. T. Chingombe, "Predicting HIV Status Using Machine Learning Techniques and Bio-Behavioural Data from the Zimbabwe Population-Based HIV Impact Assessment (ZIMPHIA15-16).," *In Computer Science On-line Conference* , pp. 247-258, 2022.

[12] L. B. H. D. V. K. M. R. C. G. C. E. D. C. T. D. .. &. P. M. L. Balzer, "Machine learning to identify persons at high-risk of human immunodeficiency virus acquisition in rural Kenya and Uganda. ," *Infectious Diseases*, vol. 71(9), pp. 2326-2333, 2020.

[13] L. Morison, "The global epidemiology of HIV/AIDS. British Medical Bulletin," *British Medical Bulletin*, vol. 58(1, pp. 7-18, 2001.

[14] J. &. S. E. S. Wiens, "Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. Clinical Infectious Diseases," *Clinical Infectious Diseases*, vol. 66(1), pp. 149-153, 2018.

[15] F. Y. Chou, "Testing a predictive model of the use of HIV/AIDS symptom self-care strategies.," *AIDS patient care and STDs*, vol. 18(2), pp. 109-117, 2004.

[16] S. P. N. P. S. K. O. J. M. V. C. E. .. &. P. J. Mathur, "HIV vulnerability among adolescent girls and young women: a multi-country latent class analysis approach. ," *International Journal of Public Health*, vol. 65, pp. 399-411, 2020.

[17] R. Sakthi Prasad, "HRM Strategies of NGOs working among HIV/AIDS affected persons in Kanyakumari District.," *International Journal of Advanced Research & Innovative Ideas In Education*, vol. 5(6), pp. 1143-1154, 2019.

[18] N. B. B. N. O. A. M. Y. H. J. H. &. L. C. M. Abad, "A systematic review of HIV and STI behavior change interventions for female sex workers in the United States.," *AIDS and Behavior*, vol. 19, pp. 1701-1719, 2015.

[19] B. Mahesh, "Machine learning algorithms-a review.," *International Journal of Science and Research (IJSR).,* vol. 9(1), pp. 381-386, 2020.

[20] A. T. N. &. S. A. Singh, "A review of supervised machine learning algorithms.," *In 2016 3rd international conference on computing for sustainable global development (INDIACom) (,* pp. 1310-1315, 2016.

[21] S. &. S. S. Suthaharan, "Support vector machine. Machine learning models and algorithms for big data classification," *thinking with examples for effective learning,* pp. 207-235, 2016.

[22] M. A. D. S. T. O. E. P. J. &. S. B. Hearst, "Support vector machines," *IEEE Intelligent Systems and their applications,* vol. 13(4), pp. 18-28, 1998.

[23] M. &. D. L. Belgiu, "Random forest in remote sensing: A review of applications and future directions. ISPRS journal of photogrammetry and remote sensing," *ISPRS journal of photogrammetry and remote sensing,* vol. 114, pp. 24-31, 2016.

[24] G. W. D. H. T. T. R. &. T. J. James, "). Linear regression. In An Introduction to Statistical Learning: With Applications in Python," *Cham: Springer International Publishing,* pp. 69-134, 2023.

[25] T. &. G. C. Chen, "Xgboost: A scalable tree boosting system.," . *In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining,* pp. 785-794, 2016.

[26] A. J. F. R. N. L. Y. W. N. A. &. B. S. D. Myles, "An introduction to decision tree modeling. Journal of Chemometrics," *A Journal of the Chemometrics Society,* vol. 18(6), pp. 275-285, 2004.

[27] C. K. M. P. E. N. I. &. M. E. Mutai, "Use of machine learning techniques to identify HIV predictors for screening in sub-Saharan Africa.," *BMC medical research methodology,* vol. 21(1), pp. 1-11, 2021.

[28] H. Chikusi, "Machine learning model for prediction and visualization of HIV index testing in northern Tanzania," *Doctoral dissertation, NM-AIST,* 2022.

[29] A. T. A. &. E.-P. E. Aybar-Flores, "Predicting the HIV/AIDS Knowledge among the Adolescent and Young Adult Population in Peru: Application of Quasi-Binomial Logistic Regression and Machine Learning Algorithms.," *International Journal of Environmental Research and Public Health,* vol. 20(7), p. 5318, 2023.

[30] M. L. &. M. J. Abbott, "Understanding and applying research design.," *John Wiley & Sons,* 2013.

[31] A. M. L. W. B. &. A. W. H. Van Lange Paul, "). Introduction and literature review.," *Social dilemmas,* pp. 3-28, 2015.

[32] O. &. F. M. Zuber-Skerritt, "The quality of an action research thesis in the social sciences.," *Quality Assurance in Education,* vol. 15(4), pp. 413-436, 2007.

[33] Z. &. Ş. M. Nartgün, "Psychometric properties of data gathering tools used in thesis.," *Procedia-Social and Behavioral Sciences,* vol. 174, pp. 2849-2855, 2015.

[34] S. A. &. B. W. S. Alasadi, "Review of data preprocessing techniques in data mining.," *Journal of Engineering and Applied Sciences,* vol. 12(16), pp. 4102-4107, 2017.

[35] M. Alehegn, "Application of machine learning and deep learning for the prediction of HIV/AIDS. HIV & AIDS Review.," *International Journal of HIV-Related Problems,* vol. 21(1), pp. 17-23, 2022.

Appendices

I. Appendix : Dataset Description

	Service region	Service zone	Date of service delivery	Year	Age	\
0	Addis Ababa	Yeka	2022-04-05	2022	25.0	
1	Addis Ababa	Yeka	2022-01-21	2022	28.0	
2	Addis Ababa	Yeka	2022-02-26	2022	35.0	
3	Addis Ababa	Yeka	2021-05-24	2021	27.0	
4	Addis Ababa	Gulele	2021-05-20	2021	36.0	
...	...	...	...	...	...	
109019	Addis Ababa	Kolfe Keraniyo	2021-12-10	2021	34.0	
109020	Addis Ababa	Akaki Kality	2022-01-26	2022	20.0	
109021	Addis Ababa	Kolfe Keraniyo	2022-05-23	2022	23.0	
109022	Addis Ababa	Kirkos	2022-03-04	2022	30.0	
109023	Addis Ababa	Kolfe Keraniyo	2022-04-06	2022	23.0	

	Gender	Target group	Testing type	Client ever HIV tested	Modality	\
0	m	HRM	PDT	Y	Index	
1	f	FSW	PDT	Y	MOBILE	
2	m	HRM	HIVST	Y	MOBILE	
3	f	FSW	Provider_Testing	Y	VCT	
4	m	HRM	HIVST	Y	Index	
...	...	...	...	...	...	
109019	m	HRM	PDT	Y	Index	
109020	f	FSW	HIVST	Y	MOBILE	
109021	f	FSW	PDT	Y	MOBILE	
109022	f	FSW	PDT	Y	Index	
109023	f	FSW	PDT	Y	MOBILE	

	Final HIV result
0	Negative
1	Negative
2	Negative
3	Negative
4	Positive
...	...
109019	Negative
109020	Negative
109021	Negative
109022	Positive
109023	Negative

[109024 rows x 11 columns]

II. Appendix: Filled Labeled

```
Service region                0
Service zone                  0
Date of service delivery      0
Year                          0
Age                           2
Gender                        0
Target group                  0
Testing type                143
Client ever HIV tested       13
Modality                    604
Final HIV result            580
dtype: int64
Service region              0
Service zone                0
Date of service delivery    0
Year                        0
Age                         0
Gender                      0
Target group                0
Testing type                0
Client ever HIV tested      0
Modality                    0
Final HIV result            0
dtype: int64
```

III. Appendix: Sample image for label selection

	Age	Gender	Target group	Testing type	ient ever HIV teste	Modality	Final HIV result
2	25	m	HRM	PDT	Y	Index	Negative
3	28	f	FSW	PDT	Y	MOBILE	Negative
4	35	m	HRM	HIVST	Y	MOBILE	Negative
5	27	f	FSW	Provider_Testing	Y	VCT	Negative
6	36	m	HRM	HIVST	Y	Index	Positive
7	27	f	FSW	PT	Y	mobile	Negative
8	22	f	AGYW	PDT	Y	VCT	Negative
9	31	f	FSW	PDT	Y	MOBILE	Negative
10	21	f	FSW	PDT	Y	MOBILE	Negative
11	25	f	FSW	PDT	Y	VCT	Negative
12	30	m	HRM	PDT	Y	sns	Negative
13	26	f	FSW	PDT	Y	MOBILE	Negative
14	31	m	HRM	PDT	Y	MOBILE	Negative
15	21	m	HRM	Provider_Testing	Y	MOBILE	Negative
16	26	f	FSW	PDT	Y	MOBILE	Negative
17	30	m	HRM	Provider_Testing	Y	MOBILE	Negative
18	31	m	HRM	HIVST	Y	sns	Negative
19	36	f	FSW	PDT	Y	sns	Negative
20	29	m	HRM	Provider_Testing	Y	MOBILE	Negative
21	29	m	HRM	PT	Y	mobile	Negative
22	40	f	FSW	HIVST	Y	MOBILE	Negative
23	23	f	FSW	PDT	Y	MOBILE	Negative
24	40	m	HRM	PDT	Y	sns	Negative
25	30	m	HRM	Provider_Testing	Y	MOBILE	Negative

IV. Appendix: Sample python code for Data preprocessing

```
#cleaning data by filling null values
import pandas as pd

# Load the dataset from the local CSV file
data = pd.read_csv('HIV Data ready for jossy.csv')
# Replace 'path_to_file.csv' with the actual path to your CSV file

# Check for missing values
print(data.isnull().sum())

# Handle missing values
# Option 1: Drop rows with missing values
data = data.dropna()

# Option 2: Fill missing values with a specific value
# data = data.fillna(value)

# Option 3: Fill missing values with the mean, median, or mode
# data['column_name'] = data['column_name'].fillna(data['column_name'].mean())
# data['column_name'] = data['column_name'].fillna(data['column_name'].median())
# data['column_name'] = data['column_name'].fillna(data['column_name'].mode().iloc[0])

# Check again for missing values
print(data.isnull().sum())

# Save the cleaned data to a new CSV file
data.to_csv('Hiv_Data_Preprocessed.csv', index=False)
# Replace 'path_to_cleaned_file.csv' with the desired path and file name for the cleaned data
```

V. Appendix: Sample Python code for label selection

```
#Label Selection
import pandas as pd

# Load the dataset from the local CSV file
data = pd.read_csv('Hiv_Data_Preprocessed.csv')
# Replace 'path_to_file.csv' with the actual path to your CSV file

# Select the important labels (columns)
selected_labels = ['Age', 'Gender', 'Target group', 'Testing type', 'Client ever HIV tested', 'Modality', 'Final HIV result' ]
data_selected = data.loc[:, selected_labels]

# Save the selected labels as a new CSV file
data_selected.to_csv('Hiv_Data_Preprocessed_Final_Labeled.csv', index=False)
# Replace 'path_to_selected_labels.csv' with the desired path and file name for the selected labels
```

VI. Appendix: Sample python code for Linear Regression Algorithm

```
#Linear Regression
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score

# Load the data from the CSV file
data = pd.read_csv('Hiv_Data_Preprocessed_Final_Labeled.csv')

# Prepare the feature matrix X and the target variable y
X = data[['Age', 'Gender', 'Testing type', 'Client ever HIV tested', 'Final HIV result', 'Modality']]
y = data['Target group']

# Perform label encoding on the categorical variables
encoder = LabelEncoder()
X_encoded = X.copy()
X_encoded['Gender'] = encoder.fit_transform(X['Gender'])
X_encoded['Testing type'] = encoder.fit_transform(X['Testing type'])
X_encoded['Client ever HIV tested'] = encoder.fit_transform(X['Client ever HIV tested'])
X_encoded['Final HIV result'] = encoder.fit_transform(X['Final HIV result'])
X_encoded['Modality'] = encoder.fit_transform(X['Modality'])

# Perform label encoding on the target variable
y_encoded = encoder.fit_transform(y)

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X_encoded, y_encoded, test_size=0.2, random_state=42)
```

```
# Train the Logistic Regression model
model = LogisticRegression()
model.fit(X_train, y_train)

# Predict the target groups
predictions = model.predict(X_test)

# Inverse transform the encoded values back to target group names
target_group_names = encoder.inverse_transform(y_encoded)

# Find the most affected target group
most_affected_group = target_group_names[predictions.argmax()]

# Evaluate the model
accuracy = accuracy_score(y_test, predictions)
precision = precision_score(y_test, predictions, average='weighted', zero_division=0)
recall = recall_score(y_test, predictions, average='weighted', zero_division=0)
f1 = f1_score(y_test, predictions, average='weighted', zero_division=0)

# Print the evaluation metrics
print("Most affected target group:", most_affected_group)
print("Accuracy:", accuracy)
print("Precision:", precision)
print("Recall:", recall)
print("F1-score:", f1)
```

VII. Appendix: Sample python code for Random Forest Algorithm

```
#Random_Forest
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score

# Load the data from the CSV file
data = pd.read_csv('Hiv_Data_Preprocessed_Final_Labeled.csv')

# Prepare the feature matrix X and the target variable y
X = data[['Age', 'Gender', 'Testing type', 'Client ever HIV tested', 'Final HIV result', 'Modality']]
y = data['Target group']

# Perform label encoding on the categorical variables
encoder = LabelEncoder()
X_encoded = X.copy()
X_encoded['Gender'] = encoder.fit_transform(X['Gender'])
X_encoded['Testing type'] = encoder.fit_transform(X['Testing type'])
X_encoded['Client ever HIV tested'] = encoder.fit_transform(X['Client ever HIV tested'])
X_encoded['Final HIV result'] = encoder.fit_transform(X['Final HIV result'])
X_encoded['Modality'] = encoder.fit_transform(X['Modality'])

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X_encoded, y, test_size=0.2, random_state=42)

# Train the Random Forest classifier
model = RandomForestClassifier()
model.fit(X_train, y_train)
```

```
# Predict the target groups
predictions = model.predict(X_test)

# Evaluate the model
accuracy = accuracy_score(y_test, predictions)
precision = precision_score(y_test, predictions, average='weighted')
recall = recall_score(y_test, predictions, average='weighted')
f1 = f1_score(y_test, predictions, average='weighted')

# Find the most affected target group
most_affected_group = pd.Series(predictions).mode()[0]

# Print the evaluation metrics and most affected target group
print("Most affected target group:", most_affected_group)
print("Accuracy:", accuracy)
print("Precision:", precision)
print("Recall:", recall)
print("F1-score:", f1)
```

VIII. Appendix: Sample python code for Support Vector Machine (SVM) Algorithm

```
# Support Vector Machine (SVM)
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.svm import SVC
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score

# Load the data from the CSV file
data = pd.read_csv('Hiv_Data_Preprocessed_Final_Labeled.csv')

# Prepare the feature matrix X and the target variable y
X = data[['Age', 'Gender', 'Testing type', 'Client ever HIV tested', 'Final HIV result', 'Modality']]
y = data['Target group']

# Perform label encoding on the categorical variables
encoder = LabelEncoder()
X_encoded = X.copy()
X_encoded['Gender'] = encoder.fit_transform(X['Gender'])
X_encoded['Testing type'] = encoder.fit_transform(X['Testing type'])
X_encoded['Client ever HIV tested'] = encoder.fit_transform(X['Client ever HIV tested'])
X_encoded['Final HIV result'] = encoder.fit_transform(X['Final HIV result'])
X_encoded['Modality'] = encoder.fit_transform(X['Modality'])

# Perform label encoding on the target variable
y_encoded = encoder.fit_transform(y)
```

```python
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X_encoded, y_encoded, test_size=0.2, random_state=42)

# Train the SVM classifier
model = SVC()
model.fit(X_train, y_train)

# Predict the target groups
predictions = model.predict(X_test)

# Inverse transform the encoded values back to target group names
target_group_names = encoder.inverse_transform(y_encoded)

# Find the most affected target group
most_affected_group = target_group_names[predictions.argmax()]

# Evaluate the model
accuracy = accuracy_score(y_test, predictions)
precision = precision_score(y_test, predictions, average='weighted', zero_division=1)
recall = recall_score(y_test, predictions, average='weighted', zero_division=1)
f1 = f1_score(y_test, predictions, average='weighted')

# Print the evaluation metrics
print("Most affected target group:", most_affected_group)
print("Accuracy:", accuracy)
print("Precision:", precision)
print("Recall:", recall)
print("F1-score:", f1)
```

IX. Appendix: Sample Python Code for XGBoost Algorithm

```
: # XGboost
import pandas as pd
from sklearn.model_selection import train_test_split
from xgboost import XGBClassifier
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
import warnings

# Load the data from the CSV file
data = pd.read_csv('Hiv_Data_Preprocessed_Final_Labeled.csv')

# Prepare the feature matrix X and the target variable y
X = data[['Age', 'Gender', 'Testing type', 'Client ever HIV tested', 'Final HIV result', 'Modality']]
y = data['Target group']

# Perform label encoding on the categorical variables
encoder = LabelEncoder()
X_encoded = X.copy()
X_encoded['Gender'] = encoder.fit_transform(X['Gender'])
X_encoded['Testing type'] = encoder.fit_transform(X['Testing type'])
X_encoded['Client ever HIV tested'] = encoder.fit_transform(X['Client ever HIV tested'])
X_encoded['Final HIV result'] = encoder.fit_transform(X['Final HIV result'])
X_encoded['Modality'] = encoder.fit_transform(X['Modality'])

# Perform label encoding on the target variable
y_encoded = encoder.fit_transform(y)

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X_encoded, y_encoded, test_size=0.2, random_state=42)
```

```
# Train the XGBoost classifier
model = XGBClassifier()
model.fit(X_train, y_train)

# Predict the target groups
predictions = model.predict(X_test)

# Inverse transform the encoded values back to target group names
target_group_names = encoder.inverse_transform(y_encoded)

# Find the most affected target group
most_affected_group = target_group_names[predictions.argmax()]

# Evaluate the model
warnings.filterwarnings('ignore')  # To suppress the warning message
accuracy = accuracy_score(y_test, predictions)
precision = precision_score(y_test, predictions, average='weighted', zero_division=1)
recall = recall_score(y_test, predictions, average='weighted', zero_division=1)
f1 = f1_score(y_test, predictions, average='weighted')

# Print the evaluation metrics
print("Most affected target group:", most_affected_group)
print("Accuracy:", accuracy)
print("Precision:", precision)
print("Recall:", recall)
print("F1-score:", f1)
```