Table of Contents

[Abstract 1](#_Toc174389759)

[Acknowledgements 1](#_Toc174389760)

[Chapter 1: Introduction 1](#_Toc174389761)

[I. Overview 1](#_Toc174389762)

[II. Research Questions 2](#_Toc174389763)

[Chapter 2: Background 2](#_Toc174389764)

[I. Dataset 2](#_Toc174389765)

[II. Literature Review 2](#_Toc174389766)

[III. Algorithms 7](#_Toc174389767)

[a. SVM 7](#_Toc174389768)

[b. Logistic Regression 9](#_Toc174389769)

[c. Decision Tree 9](#_Toc174389770)

[d. XGBoost 10](#_Toc174389771)

[Chapter 3: Methodology 12](#_Toc174389772)

[I. Tools and Techniques 12](#_Toc174389773)

[II. EDA and Visualization 12](#_Toc174389774)

[Chapter 4: Results and Conclusion 13](#_Toc174389775)

[I. Critical Discussion 13](#_Toc174389776)

[II. Conclusion 13](#_Toc174389777)

[III. Future Work 13](#_Toc174389778)

[Chapter 5: Legal, Ethical and Professional Issues 13](#_Toc174389779)

[References 13](#_Toc174389780)

[Appendices 13](#_Toc174389781)

# Abstract

# Acknowledgements

# Chapter 1: Introduction

## Overview

The main aim of this project is to develop a reliable machine learning model that can predict AIDS outcomes and compare multiple machine learning algorithms using this application. The project will use the Streamlit framework so that it can create an interactive web application which can be accessible easily.

AIDS is a disease that stands for Acquired ImmunoDeficiency Syndrome, it is widely known as a life-threatening condition caused by a virus known as Human Immunodeficiency Virus (HIV) (*HIV and AIDS*, no date). Currently there is no cure known cure for HIV, but early detection plays a major part in extended the life of the person affected with AIDS. This is where ML can help in the early diagnosis, or at least help the medical professionals to speed up the process (He *et al.*, 2022).

Initially, an extensive AIDS dataset will be utilized to develop a machine learning model. Suitable algorithms, such as logistic regression, random forests, and neural networks could be considered for this classification task. After the model has been trained and validated to ensure it predicts accurately, the focus will shift to developing the web interface using Streamlit. Streamlit allows for easy integration of interactive widgets, such as sliders, buttons, and text inputs, which facilitates efficient collection of user inputs. These inputs then act as parameters for the machine learning model. In the Python script that powers the website, functions will be incorporated to process these inputs through the model and display the predictions on the web page. Streamlit also supports embedding data visualizations, which will be used to present charts and graphs that can help users understand the model’s predictions.

Upon the completion of this project, it is expected that the website will function as a tool that aids in the prediction of AIDS outcomes, leveraging the power of machine learning. This will not only demonstrate the practical application of data science in solving significant health issues but also provide a user-friendly platform for users to interact with the model. The integration of interactive elements and data visualizations will further enhance the understanding and accessibility of the predictive analysis, making it a valuable resource for users and researchers alike.

## Research Questions

* How do various machine learning algorithms compare in terms of accuracy and efficiency in the detection of AIDS?
* Which features have the highest impact on predicting the infection status?
* How do patient outcomes differ between various demographic groups (e.g., age, gender, weight)?

# Chapter 2: Background

## Dataset

The dataset has 4 files each with varying sizes of patient records. One important thing to note is the data does not contain patient identifying information like name, email, etc., Other personal information like age and gender are present. The dataset contains 23 features related to patient along with the target variable. Some of the important types of features available are patient demographic data, treatment type, health indicators, CD4 cell count and other features. The target variable “infected” indicates if the patient is infected with AIDS or not.

## Literature Review

In this research, the popular HIV and AIDS trending topic in Twitter is selected for the sentiment analysis, thus using supervised machine learning to classify the Twitter sentiments. The method of research is best presented in the form of a number of steps, which correspond to the evaluation of the analysis. The first procedure included data collection where the experimental base was equipped with a Twitter API via the help of the Tweepy API, collecting a sufficient amount of data set of 83,495 tweets. Tweets were filtered using keywords like ‘AIDS’ and ‘HIV’, the data set comprised of attributes like username, geographical location, and the content of the tweets. These set of data was useful in informing various other set of analysis which have been elaborated as follows.

After the data was collected, emphasis was in the pre-processing of the text. Cleaning of collected data was also carried out to achieve a suitable state in order to analyzed. Some of the pre-processing techniques that were performed include; deleting the unneeded features, stripping the punctuations, numbers and non-essential words such as ‘stop words’, words with a length of three characters or below, URLs, words that have been repeated and HTML encodings. These steps were essential as it enabled the conversion of the raw data into a well framed format that could be easily handled with the help of a machine learning algorithm. The text pre-processing made it possible to eliminate noise from the dataset hence improving on the quality of the data used in the analysis.

The study then moved to comparative sentiment analysis; it used the TextBlob and Vader Python libraries. The following libraries were used to perform the sentiment analysis on the pre-processed tweets and the sentiment results were depicted with the help of Python word cloud library. The results indicated most frequently used positive and negative words related to HIV and AIDS that helped to assess public attitudes to the subject. The sentiment analysis was fine-tuned using the vectorization techniques that uses the Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), and N-Grams. These methods assigned values to each word in the training data which is also important for sentiment classification.

The analytical method that props up the study was primarily based on supervised machine learning algorithms. Three classification algorithms were utilized: such as logistic regression, support vector machine (SVM), and multinomial Naive Bayes. These algorithms were then applied on the pre-processed and vectorized data as an effort to look for the class values by means of a defined target function. These algorithms’ performances were measured based on the metrics such as accuracy, precision and recall which were defined with the help of confusion matrix. Of all the three models, the best performing was the support vector machine especially when it was used in combination with the n-gram model. The highest accuracy was attained by the SVM which was 99% in the sentiment classification, this justifies the usefulness of the developed models in sentiment analysis of social media data.

In the study, the effect of the size of the vocabulary used in the sentiment analysis also formed part of the focus. Vocabulary sizes were set differently for the Bag of Words, TF-IDF and N-Grams models to let the researcher compare the impact of using larger set of various and specific words on the analysis of the sample. In the study, it was revealed that the best results were obtained from the support vector machine model which was fine-tuned with n-grams feature representation of the text which gives the F1 score of 0: 999 which is very close to accuracy and recall which are the main parameters of a good classifier. This result applies to the feature selection and model tuning as important determinants of the performance in sentiments analysis.

All in all, the study gives a detailed analysis of the sentiment regarding HIV/AIDS on the Twitter platform through the application of standard multivariate text analysis methods. The results show that it is possible to achieve higher values of overall accuracy of a sentiment classifier when the support vector machine is used, fine-tuned with proper features. This choice of methodology, together with the focus on low error rates and high efficiency, enriches the methodological base of sentiment analysis and contributes to the qualitative development of practical research in topics related to public health, based on the analysis of big data from social networks. The findings so underscore the effectiveness of machine learning when it comes to gauging the disaster on mood on pertinent heath issues and responding to public health campaigns/ policies.

This research focuses on the use of the ML models for the HIV disease and patient progression and patient outcomes, to enhance trial matching for clinical trials. So, with the use of some of the most popular methods such as Random Forest, SVM, K-NN, NB, XGBoost, and Logistic Regression the above objective was tested for its efficiency. The empirical results reveal that for the dataset used in the study Random Forest and XGBoost algorithms have higher accuracy equal to 87. 85% and 89. The gender differences in the two cellular subpopulations associated with HIV disease progression were 0. 02%, 0. 02%, respectively. These rigorous models also assert the positive outlook in correctly categorizing individuals living with HIV into the ideal outcome categories, thus improving the efficiency of clinical trials’ procedures as well as the care given to the patient.

The methods used in this study are laid down in the following ways; Model selection is the first step. Due to the fact that it is rather difficult to forecast HIV related outcomes, the paper examines several of the most effective algorithms for classification. From mid-level models, Random Forest, an ensemble learning method, has been used because of its robustness and the capability in dealing with big data and no overfitting issue. This algorithm employs many decision trees in coming up with predictions which makes it strong for the study of the algorithm. Two more important and effective classification algorithms used were Support Vector Machine (SVM) due to the fact that it is applicable to linear as well as non-linear data. SVM has also high metrics of discrete effectiveness in handing of high-dimensional data environment. K-Nearest Neighbors (K-NN) is one of the simplest classification algorithms that categorizes data points according to the majority class of their nearest neighborhood; this makes it possible to compare its prediction of HIV progression. The Naïve Bayes (NB) classifier was added to the list due to its ability to be a probabilistic classifier with a focus on the Bayes’ theorem and simple yet effective in classification. The Gradient Boosting Methods were more used because of the chance of creating successive decision trees which brings enhancement in the predictive model. Last but not the least, an analysis was done to check for the feasibility of using Logistic Regression, a linear classifier, for the prediction of HIV progression and the outcomes of the patients depending on the features corresponding to the problem domain.

The data splitting process used in this study represented a way of achieving the validity of the predictive models as well as their transportability across samples. The data set was split into the training, validation and the test with 80%, 10% and 10% respectively. This distribution facilitated a way that the model development and the model evaluation were balanced appropriately to provide accurate predictions with regard to HIV progression and patient outcomes. The key to be predicted using the ML models was the censoring indicator (cid) which was a binary attribute representing if the data of the patient corresponds to censoring or failure with regard to HIV progression. The ability to predict cid accurately can greatly enhance the assessment of the effectiveness of treatment, the determination of costs involved in treatment and more importantly, the care of patients and hence this feature will be of primary interest to the study.

Training of the model was done using a set of efficient tools and languages with Python as the key language owing to the availability of related libraries. The source management was well supported by the scikit-learn library containing all-toolsets necessary for the model selection, training, and also integration of different ML algorithms into the process. XGBoost model that has a high gradient-boosting efficiency in this study improved the accuracy of predictions. A cross-validation process was employed with the help of GridSearchCV so that all of the models were optimized to provide the best possible predictive outcomes. Feature scaling was performed with the help of StandardScaler tool, which is intended to provide fair and unbiased training of models.

In order to validate the performance of the developed ML models, several performance metrics were used which includes accuracy, precision, recall, F1-score and ROC AUC. All these metrics were useful in establishing a strong foundation for the assessment of the models’ ability to provide predictions. In Random Forest and XGBoost, the analytical findings of the study showed higher precision, recall, F1-score, and ROC AUC. Random Forest demonstrated highest precision almost to the extent of 0. The last model with a score close to 851 was XGBoost with nearly 0. 846. The same is true for recall: these models demonstrated a higher value compared to other ones, with Random Forest equal to ~0. 624 and XGBoost around 0 as is depicted in the graph below. 653. Another compelling evaluation metric: balanced F1-score which confirmed the efficiency of XGBoost and Random Forest with distinction: the F1-score of XGBoost was close to 0. 74 and Random Forest about 0. 72. XGBoost had the best ROC AUC score of approximately 0., Kappa value of 0.81, proving that the criterion has high discriminating ability.

Therefore, the study establishes the feasibility of using the superior ML models, Random Forest and XGBoost in HIV progression and patient outcome prediction. Some of these models have been proven useful in classifying persons into relevant categories in patient-trial matching hence improving trial’s efficiency and patient’s care. They have identified the fact that an additional improvement to these models can be achieved by performing hyperparameter optimization, feature selection and incorporating other data sources to enhance the precision, accuracy and recall of the models. The authors of the study have emphasized that as ML advances on, the study results are anticipated to help to the current endeavors of applying superior analysis in combating HIV as well as other complicated health conditions.

The research is based on a case-study of utilization of thirteen different ML algorithms for the purpose of estimating the effect of ART on HIV positive populations in 75 countries during 11 years. The main research question is to identify the optimal set of ML algorithms that provide accurate prognosis of the results of ART on clients who have HIV, expressed in the percentage of people who are alive to those who died of the disease. Matching the type of study, the work uses the supervised and unsupervised ML algorithms of KNN, Logistic Regression, SVM, XGBoost, Decision Tree, Random Forest, and K-Means.

Information for this research was obtained from the following agencies: World Health Organization (WHO), the World Bank. The datasets were obtained from the WHO, and included indicators of ART coverage, total HIV related deaths, and total HIV positive population. Information on the total population of every nation was also provided by the World Bank. Hence, for a period between the year 2011 and 2021, the dataset has information of 75 countries in 825 rows, where each row contains the stated indicator for a particular country of its given year.

Data cleaning was performed before the actual integration wherein instances with missing values were dropped and the two datasets were merged using the ‘Location’ variable. The final dataset had ten major variables, which are ART coverage, HIV related death rate and the total population of the country among others. This dataset was then used as base to train and/or test various instances of ML.

The following supervised ML methods were used in the study to analyze the changes in the population with HIV, due to ART; Another one was made up of a new target variable, which has been named ‘Index,’ defined as the share of population alive this year. If the ‘Alive (%)’ was less than 96 %, the record was coded as ‘Small’, otherwise ‘Big’ As the target variable, other such features which were used to train the ML models were ART coverage and number of people with HIV.

To assess the efficiency of each supervised ML technique, intrinsics of accurate classification algorithm: accuracy, occurred precision, taken recall, and F1-score were used. The data set was divided as training and test data with 66. 7% to 33. 3%, respectively. The models were trained and tested and the tests results were also compared. Thus, two algorithms of supervised learning – XGBoost and Random Forest were identified as most efficient in terms of accuracy. Random Forest, for instance, had the best result, having an accuracy of 90% which placed it higher than all the other models of determining the impact of ART in the population with HIV.

Apart from supervised learning the study also used K-Means a type of unsupervised learning method to categorize similar countries on their ART coverage, HIV related deaths etc. The clustering analysis resulted in two distinct clusters: the first group was composed of 15 mostly populous countries while the second group was having 60 non-populous countries. The coherencies of the clusters were further explored and characterized, by correlation coefficients of the Pearson type in reference to the impact of ART to HIV-related mortalities. The first aggregate with a higher population had a stronger negative relationship between ART coverage and deaths suggesting that increased coverage of ART had a superior impact to the mortality in these countries.

In the conclusion, the study finds out that the ML algorithms; Random Forest and XG boost yields high accuracy in the prediction of outcomes of ART on populations with HIV. Another advantage of the clustering analysis was the possibility to identify the differences in effectiveness of ART between certain countries. The results imply that ART has a higher level of efficiency in decreasing the HIV—related mortality rates in large population countries because the negative coefficient in the first cluster is higher.

The study also recognizes that more studies could be carried out to investigate other variables including the GDP of each country in other to determine the effects in HIV prevalence, mortality rates and ART coverage. Moreover, improving the models’ features besides the current GLM could also improve the predictive capabilities of the models. Future studies could therefore increase the understanding of the effectiveness of ART by broadening the criteria of investigation to include younger populations as well as people who have not yet begun treatment.

The study incorporates a new feature of machine learning alongside blood biomarkers that make a shift in HIV diagnosis to better diagnosis in the early stage. Conventional diagnostic such as ELISA has comparably low sensitivity and specificity and PCR method takes a relatively longer time than antibody-based tests. There is an alternative in the form of machine learning which… is designed to discover complex patterns in large amounts of data. The present work pursued the goal of using ML algorithms to diagnose HIV infection or detect biomarkers in the bloodstream that may be indicative of early-stage HIV infection, which can be a much faster and more accurate method than the current ones. This approach extending into routine practice does not only enhance the early identification but can also be likely to have great effects on enhancing public health status by lowering chances of passing on HIV.

The research used a descriptive research approach, and data collected from clinical databases were used to assess the possibility of using ML in the diagnosis of HIV. This study used a database of deidentified blood tests, participants’ basic demographics, and clinical characteristics of HIV-confirmed participants. Some of the measures quantified in the study were viral load, CD4 cell counts and inflammation markers such as; C reactive protein (CRP) and interleukin 6 (IL-6). Thus, these biomarkers were chosen based on the fact that they have already been shown to be connected to HIV progression and immunity. In this study, missing values and outliers were dealt with before the data was load for analysis in order to have a consistent data set for analysis. Other feature selection methods such as Principal Component Analysis (PCA) where applied to make feature selection and reduce the resultant dimensions without losing important information.

Using biomarker data, the study conducted classification experiment that tried different ML algorithms to predict HIV status including: SVM, RF, and Neural Networks. The data set was then split in a training and validation set in the ratio of 8:2. Okonkwo et al. assessed the accuracy and hold-out validation of the models using the area under the receiver operating characteristic curve (AUC-ROC), sensitivity and specificity. In details, accuracy of the models is in the following order; SVM, with the accuracy of 92. 5% thus proving that it has the efficiency and ability in dealing with high dimensionality and decision space. In the second position came Random Forest with a level of accuracy of 90. 3% while giving evidence of the ability to capture intricate patterns though enhanced on ensemble learning. The Neural Network model with slightly lower accuracy of 88. 7% and was able to learn nonlinear pattern recognition especially when hyperparameters were fine tuned.

Another suggestion of the study was the selection of the features, moreover it explained how much each biomarker contributed. The most important feature turned out to be CD4 cell count, which accords with the medical understanding of CD4 as the indicator of immune status and pathogen progress in HIV positive patient. Viral load was another critical biomarker; this is because viral load was defined as the amount of virus in the blood and would help in assessing viral replication and guide treatment. Other inflammatory-specific markers such as CRP and IL-6 were also significant suggesting increased immune response to HIV. Such singular biomarkers contribute to the discovery of a lengthy biomarker panel for the classification of HIV, encouraging further, more specific, therapeutic approaches.

For further proof and to make sure that the model is well portable, the developed model was tested on an independent dataset that contains different patient samples. This matched well with the results obtained at the first training and validation of the model where the model showed acceptable reliability at various clinical situations. Out of the four classifications techniques, SVM specifically has high accuracy on the independent dataset with a higher AUC-ROC indicating that the model can easily classify and distinguish people living with HIV and those who are HIV negative. This validation shows the effectiveness of the work of the developed model if applied to a wide range of patients that one can meet in real-life scenarios.

The results of the study could be useful in determination of the HIV abbreviation and other health concerns. On the basis of the results obtained based on the use of scientific approaches such as ML and blood biomarkers, the study offers an efficient and new way of HIV screening. Considering the high accuracy rates which have been committed for certain models as SVM, this method may be incorporated into practice to increase quick and definite diagnosis of different diseases. Nevertheless, the study also shows limitations; for example, the existence of bigger and more diversified data sets for validation, or the feasibility problems of the model in health-care organizations.

Further studies should emphasize the collection of new data that would be used for further training of ML algorithms and additionally include more longitudinal data to account for temporal dynamics of HIV progression. Furthermore, extending the biomarkers list and the opportunities of the multi-modal data analysis and integration may improve diagnostic performance of the model. Other for-the-record investigations such as usability trials and pilot projects in care delivery organizations are required to transfer this science from theory to application. Engaging public health organizations will be important for ensure that the model can be easily infused into relevant healthcare systems, and to support the broad application of the research in combatting HIV.

## Algorithms

### SVM

Support Vector Machine or SVM is a type of learning algorithm or model which comes under both classification as well as regression methods. Its main aim is to find the right hyperplane that divides various classes present in a dataset. This hyperplane is chosen in such a way that it’s the maximum distance possible from all the support vectors of different classes. It is true that a larger margin makes the overall generalization when the model reconstructs with new data.

SVM work in a way that input data is transformed in to an n-dimensional space where each of them is features. The next step of the given model is to look for the hyperplane, which would bring the best classification of the data in classes. In two dimensions the hyperplane is a line in three dimensions it is a plane that is just like any other plane produced from the plane of the scale. This hyperplane lies in the plane that is perpendicular to the normal vector and bisects the distance between the two farthest points of the set of support vectors, so unlike other classifiers, SVM is relatively intolerant of outliers.

When data are not separable in a linear fashion, then SVM introduces a ‘soft’ margin system which can accept moderate misclassification in order to balance the margin and the volume of mistakes. Furthermore, use of the kernel trick means that relationships that are nonlinear can be dealt with by mapping the samples into a higher dimensional space where they are linear. Linear kernels, polynomial kernels, radial basis function (RBF) kernels, and sigmoid kernels are among the widest spread ones and the choice of the kernel depends on the data set and the problem being solved.

SVM can also be used for regression problems or Support Vector Regression (SVR) in which the aims are to find the best function that models input attributes and target value with acceptable level of loss.

SVM is suitable for datasets, which have many features and can work when the number of features is bigger than the number of instances. Although it can be very powerful, it can be more computationally expensive compared to the other approaches and therefore, getting the proper kernel as well as its hyperparameters is very crucial. However, these are some of the challenges, which do not make SVM less viable as a tool, more so in classification problems, where boundaries between the classes are well defined.

### Logistic Regression

Logistic regression is also one of the most popular methods of binary classification, aimed at approximation of the instance’s place in the space of two classes. Unlike linear regression which make a continuous quantity prediction, logistic regression predictions the probability of an event with the use of a sigmoid function. The characteristic feature of this function is its ability to bring any real value to the range between 0 and 1, thus making it possible to estimate probabilities. The model generates a measure of how likely it is for an input to be in the positive class and this measure is then compared to a baseline most of which is set at 0. 5. If the probability is more than this value the instance is classified as positive and otherwise as negative.

This model is similar to linear regression but instead of summing the input features, logistic regression sums the weighted by their coefficients feature values and a bias term. However, rather than outputting this linear combination it passes the summed values through the logistic function so that it creates a probability. The weights used in the neural networks are also determined from the training data by a process referred to as back propagation which seeks to minimize a loss function and in the case of binary classification, the binary cross entropy loss. This loss function quantifies the difference between the predicted probabilities and actual class labels and the weights are adjusted using gradient of the loss function so as to minimize the loss.

Sometimes, logistic regression introduces L1 (Lasso) or L2 (Ridge) regularization, which reduces the weights’ magnitude, and, thus, the complexity of the model and its susceptibility to overfitting. In this case, logistic regression performs well when extracting the features and their relationship with the results contain a linear nature. At the same time, it might fail with non-linear relations unless feature engineering becomes the center of attention. However, logistic regression is fast and fairly useful by its practicality of being applied in a straightforward manner.

### Decision Tree

Decision trees are categorized as a supervised machine learning algorithms works both on classification and regression. They work according to a principle of carving the data in feature space so as to give a tree-of-decisions. It is made up of nodes which contain decision making and edges that show the results, until the final nodes or leaves that offer the decisions.

The tree construction starts with the selection of the feature that provides ‘maximum information gain’ – this measures the goodness of the split of data. Splitting methods used are Gini index, Information gain which uses the concept of entropy, and variance reduction as splitting criterion for regression problems. One of the measures that fits in classification is the Gini Impurity which predicts how often a randomly chosen feature would be misclassified if the its label were chosen at random. This means that we are looking for a state where Gini is as low as possible, and if it reaches a certain level of low, then the node is said to be ‘pure’ as it contains many samples of the same class. However, in regression the aim is to minimize or reduce variance within in subsets.

They are also very easy to explain, since the steps taken by the decision process can be easily depicted visually. For example, each path from the root node to a leaf node is a decision rule making it easy to see how some predictions are made.

Nonetheless, decision trees have their shortcoming as follows: They have the drawback of overfitting especially where the tree is deep; an effect arises, which causes poor performance on the new data. Some of the ways through which overfitting can be prevented include, pruning, and some rules such as setting maximum depth of decision trees and minimum observation levels in splits. Also, decision trees are highly prone to overfitting, which makes them quite unstable – that is, some slight changes in the data will lead to the construction of very different trees. This instability is well explained and can be managed by ensembles such as the Random Forest or the Gradient Boosting where many trees are constructed to act as a single tree but have a higher stability and accuracy.

### XGBoost

XGBoost, pronounced as ‘cross GB’, stands for Extreme Gradient Boosting, an improved version of the GB with superior characteristics related to speed, performance and scalability. For binary and multiclass classification models, as well as for regression models, it is widely used. Apply of this algorithm in different machine learning challenges, as well as in real-life problems, has put it amongst the most commonly used ones.

In other words, XGBoost is a more efficient form of boosting, which is a type of ensemble learning technique that involves using several ‘weak’ models, the common one being decision trees. In each, new model is developed mainly for eliminating the flaws of the earlier models and achieve a better result for predictive function. XGBoost works on this by partitioning the hyperplane space into subspaces to make the training faster, and the model better.

The XGBoost algorithm then involves a base model and in the case of XG the base model is a decision tree which is trained to the dataset. It then determines the residuals or the discrepancy for the values actually observed and the values that have been predicted by the model. In subsequent iterations the new models in fact are trained on these residuals and progressively improved on. The recommended number of iterations is used until the residual drops below this limit or until a certain number of iterations is achieved.

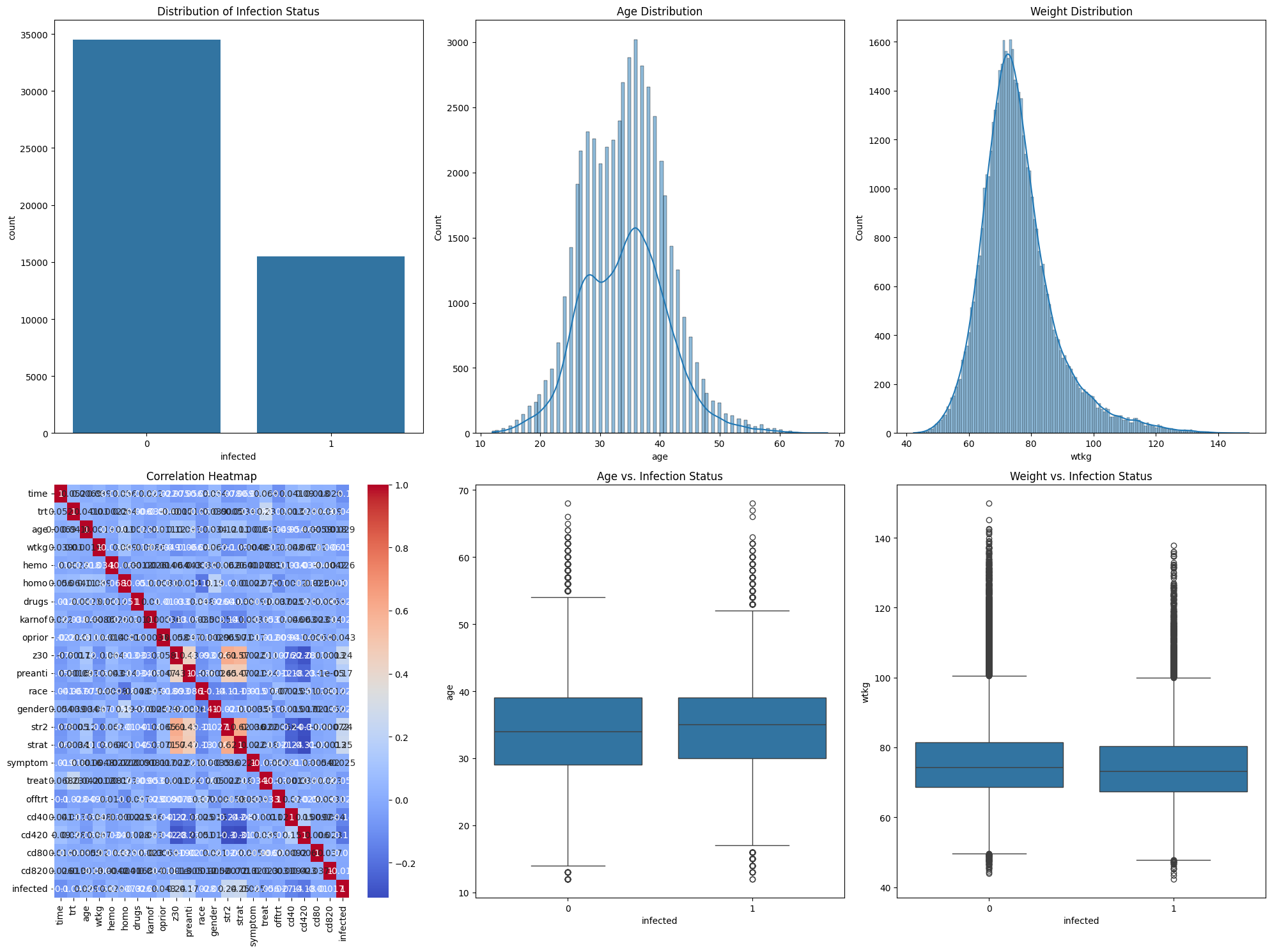
Another advantage of XGBoost is that it is equipped with regularization to avoid overfitting, which is another extensive issue in the field of machine learning: the model performs well on the training data but poorly on the other data. One may use conventional L1 and/or L2 norms that XGBoost, implements using the Alpha parameter and Gamma parameter. L1 regularization helps in removing the features that are not much relevant in the model known as selection type of regularization L2 also known as ridge regression does not make much difference between large coefficients and small coefficients but does penalize a large value of coefficient leading to a smoother model. These above regularization techniques can aid to avoid problem of overfitting so that the XGBoost models can perform better in future unseen data.

Another feature of XGBoost is that optimization is based on second derivatives. Unlike the gradient boosting algorithm in general, which is based on first order derivatives, XGBoost utilizes, in fact, first and second-order derivatives (the Hessian). This makes it easier for the algorithm to estimate the curvature of the loss function, which in turns help to produce better estimate the of model parameters in order to increase the efficiency of the model.

It likewise has several engineering optimizations to boot that helps the model to increase the efficiency and scalability. Another interesting aspect is that it has been designed with the provision of handling missing values and spars matrices, thus finding its application in a number of actual datasets. Also, XGBoost has the ability of processing data in blocks, which makes the algorithm to benefit from multi-core processors and distributed systems and this makes the algorithm to take little time for training especially when working with large data sets.

# Chapter 3: Methodology

## EDA and Visualization



The first plot in the grid is a count plot of the target variable ‘infected’. The plot above shows the number of infected people and the number of people not infected. From the visualization above it can be seen that a greater percentage of the sample population is not infected; this could mean that the sample dataset is skewed. The nature of this distribution is important to understand for further analysis for it might affect the outcome of any predictive models applied to the data.

The second plot in concerned with the dispersion of the age variable. This distribution is plotted as a histogram also adding a kernel density estimate (KDE) to allow to observe smoothness of the data distribution. It seems that it is nearly normally distributed with a slight positive skewness with regards to the age distribution. This means that more than half of the people in the dataset are within a given age bracket which would create a background to justify if age influences the infection status.

The third plot also shows the distribution of weight variable (wtkg). The same as in the case of age distribution, the use of histogram with KDE is applied. The distribution of the weight also looks pretty much normal with part of the distribution showing certain spread, which indicates variability among the subjects included in the study with regard to weight. A closer examination of this distribution contributes to the knowledge of the part played by weight whenever there is infection.

The fourth plot is the correlation matrix in the form of heat map representing the correlation between all the numerical fields of the data set. The heatmap depicts a matrix of positive or negative correlation values and the strength of the correlation with the help of colors for graphical representation although annotations exhibit the actual correlation coefficients. Most of the time, this kind of visualization is useful for tension-seeking, which can be crucial in the process of feature selection, or in general understanding the data shape.

These are two graphs with the Group being the infected status, while the Variables being, age and wtkg with fifth plot being a box plot for a creation of variable age and the sixth plot being a box plot for the creation of variable wtkg. This conjecture arises from a simple boxplot of age versus infection status, and plotting the data may reveal slight differences in the exact age distribution between the two groups, though it is entirely expected that infected people would be slightly older, on average. The boxplot of weight against infection status gives a similar information about how distribution of weight is affected by infection status, but the differences appear less dramatic.

# Chapter 4: Results and Conclusion

## Critical Discussion

Building the AIDS prediction model and the Streamlit app for the model was quite a creative and stages process. With an array of data loading and understanding as the initial steps, followed by the feature engineering, model training, and finally web application development. Below is a breakdown of the strategy I used and the performance I managed to record at the end of it.

First, I began by importing the AIDS classification dataset into a pandas DataFrame. This dataset includes many attributes that could suggest a person’s propensity to be infected with AIDS. The variables used in the analysis consisted of numerical variables such as age, weight, as well as hemoglobin level, and categorical variables such as gender and race/ethnicity. From the above information and the help of pandas’., head(), . info(), and . By running the |describe()| functions, I was about to understand the structure of the data, the missing values and the basic statistic. This was important in the first instance because it dictated further actions, I was to undertake in data preprocessing and feature selection.

With more information of the data, I went to the next step, which is visualizing it. First, with help of matplotlib and seaborn libraries I created a number of plots that may emphasize some significant properties of the given data set. These were followed by a probability density function of the distribution of the infection status, age and weight distribution and a heatmap to see how features were dependent on each other. The boxplots of the age and weight with the status of infection were especially useful as they outlined the possible trends and relations which are useful for making predictions. These enabled me distinguish which of those features would be more likely perhaps provide a probability of an AIDS infection.

Then, I enrolled the data of stylometric features into model training. Eliminating the dependent variable, the features (X) and the target variable (y) was the infection status. I then divided the data into training dataset from the ones to be used for testing to minimize dependency on the data set used to train the model. This step plays a crucial role in anticipation of over fitting when a model is capable of learning from the training data and fail to generalize well from new unseen data.

In order to enhance the nature of models, I used feature scaling from the StandardScaler of scikit-learn. This step normalized the features thus making each one of them contribute in equal proportions in predicting the model. Also, for feature selection, I used Select K Best which scores features according to their importance and only takes the best K features after ranking the features tool. This does not only decrease the data dimensionality but also improves the model’s performance since unnecessary features will have been removed. In this work, the features with the 10 highest ANOVA F-values were chosen for this project.

Fortunately, the chosen features were proven to be significant to construct a reliable predictive model. With these features, I decided to train a Random Forest Classifier, this being a powerful ensemble learning method whose major strength is accuracy and handle with big data sets. Based on the accuracies achieved and precision, recall and F1-Score calculated below, I determined that the selected features were optimal for training the model. The model had a total accuracy of 70%, and had a 73% accuracy for non-infected and a 52% for the infected cases. The same was evident from the confusion matrix where the model had a better performance in identifying non infected patients but poor performance in identifying infected patients whereas evident from lower recall and F1-score of infected class.

After training and evaluating the model, it was saved via joblib with a view to using it in the Streamlit application. It was necessary to make the application interactive and convenient to use to allow users inputting a set of feature values and getting in a result the output of the probability of this person having the AIDS virus.

When creating the Streamlit application, the very first step was to load the trained model and the used dataset to make sure that all the features are consistent and the same preprocessing was applied as during the training stage. To normalize the inputs to a form, I translated the names of the raw features to understandable labels with descriptions of about 1-3 words long. This mapping also involved replacing technical names of some columns such as ‘wtkg’ and ‘cd4’ as more easily understood ones such as ‘Weight (kg)’ and ‘CD4 Count (immune system health).’’

It was preceded by the application interface which consisted of six columns and input fields of features to be learned. This layout was chosen to enable any user easily find any information within the application with minimal difficulty. These could be typed directly, or default values which could be estimated from the data set being considered, presented by the means of the columns. Such a setup was possible and provided both settings for serious users who would want to input serious data as well as the normal user who would have the default settings.

After that users input the input data, then they click the “Predict” button to make the prediction process to be running. To the used application, the same preprocessing steps that were used while training the models were used and therefore the input data was normalized and the top selected features were used. From this, the model produced a prediction whereby the model hinted at the likelihood of the particular person being infected with AIDS or not.

The prediction result was comparatively easy to understand and would be labelled with either ‘Infected’ or ‘Not infected’ The application also provided the possibility of the two results thus giving the users an indication of the model’s confidence level. This was especially helpful in a number of cases where the marginal likelihood of the event outcome might be perhaps 51/49 and such decision making is made relatively easier.

In addition, it showcased the features used in the model so that the user has credence in the reality of the prediction. This information was relayed in a manner that would permit the user to actually understand the numerous layers of processing that occur prior to reaching the ultimate outcome; there was much discussion about machine learning, but here it was translated down to basic, palatable usage that people could comprehend.

Analyzing the results presented in confusion matrix and classification report, the conclusion might be made that the performance of the model is quite decent, however, there are to be improvements on the infected class. The reduced recall rate and F1-score for infected instances indicates that the model for classifying infected persons may require the optimization or the introduction of more parameters to register the class ‘impact. This is something that should be taken into account in the future work because in the medical setting it is highly significant to identify all the infected people.

## Conclusion

To sum up, the presented work required a number of steps related to the data exploration and visualization, data preprocessing, model training, and building real-life applications. The end product is a Streamlit application that lends itself naturally to predicting the likelihood of AIDS infection based on the proposed set of features. Although the proposed model has relatively low false positive and high true negative values, some improvements for enhancing the model’s sensitivity and specificity will be required in the future research in relation to infected cases. I found this project as both enjoyable and challenging as it enabled me learn more about the application of machine learning in healthcare system.

Unfortunately, this AIDS prediction project is not for the public use because of high risk factors and its requirement of medical supervision. While the model’s results may be helpful, they are no indication of professional medical prognosis and could result in self-diagnosis or unnecessary stress. The current limitation of the application in terms of efficiency, specifically, lower detection rate of infected people is justified by the necessity to involve the experts and validate the results. Protecting the users and the veracity of the medical content is important, and since this platform does not involve a direct interaction with healthcare practitioners, there is a possibility of wrong interpretation or usage of the results. Thus, this project has to be limited to clinical and the like settings where physicians can testify to preparing requisite background for its application.

## Future Work

Subsequent researches involved in this AIDS predicting project would therefore aim at improving the predicting model especially as regards the infected cases. This could mean that to accommodate the new features required to capture more detail of patient’s data, like lifestyle or genetic characteristics, or any other relevant detail, then it would require multiple models. Trying to employ more enhanced methodologies as a part of the ML approaches, such as deep learning or ensemble models, can also enhance the accuracy of the predictions. However, it suggests the necessity for developing extensive validation with greater datasets and cross ethnicity population samples. As such, integrating the application in a secure, healthcare regulated fully-supervised by medical professionals’ environment would be the only way to safely implement the application in a real word scenario. Finally, an easily navigable interface with accompaniments on the forecast might enable the user take suitable medical advice or action.

# Chapter 5: Ethical Issues

The ethical aspect of this AIDS prediction project is rather large: the question is whether it is moral to collect, store, and process people’s information and use mathematical models to make predictions based on this data. There is always the question of privacy when dealing with people’s health information and the laws that govern such information such as the HIPAA must be complied with. In addition, the model must have predictive capability because anything that is predicted wrong may cause some harm, include mere complacency or fear that is not warranted. There is also the disinhibitory use if individuals, or people who are not experts, provide interpretations of the predictions to make fatalistic or irresponsible decisions. In order to counter these issues, the project must apply ethical standards: The operation with the data has to be explained thoroughly and the user must be informed about the shortcomings of the model, that means additional disclaimers are needed, moreover, to protect the user’s health the tool has to be only available with the permission of a doctor.

# References

# Appendices