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

SVM is an example of supervised machine learning training model which is used for classification as well as regression analysis. From the concept of SVM, the goal is to identify the right hyperplane to different the classes in a given data set. The major emphasis is done on the greatest distance between classes, which is the actual shortest distance of the data points of one class from the nearest data points of the other class, called the support vectors. It is also important due to an important concept related to training models – the fact that a higher margin almost always results in improved generalization when the model needs to work with new data.

The process starts with associating the given input data with points in a given n-dimensional space where n is equal to the number of features in a given data set. In this space, value of features for each data point is plotted. The purpose is to identify a line or a surface that splits the set data and forms on side of one class and the other side of another class of the data. In Euclidean two-space, the hyperplane is a line while in three-space it is a plane with the same dimension as the space in which it is contained. In, for instance, the third dimension, the hyperplane is even more complicated than in the second dimensional space but the idea is equivalent.

It is woe of the hyperplane is that which has maximum margin, this means the distance of the hyperplane to the closest point of the different classes. This margin is determined by the points called the support vectors and these are the data points which are nearest to the hyperplane. These support vectors are however very important in the determination of the position and orientation of the hyperplane. The rest of the features do not alter the hyperplane and thus, for the same reason, SVM is less sensitive to outliers.

If the data is separable that is the classes can be separated by a line in two-dimensional space or a plane in three dimensions, SVM determines the line or plane that separates the classes with the maximum margin of separation. However, there are occasions whereby the data is not cleanly separable by the straight leaning line. In such cases SVM employs a kind of technique known as the ‘soft margin’ approach. This approach permits some amounts of misclassifications or errors where data can be on the margin or on the wrong side of the hyperplane. The balance between the margin and classification errors which is optimizing the margin as well as minimizing the errors is regulated by parameters of how much should the model focus on the large margin if it means that one or two errors have to be made.

When the data is not separable in the best feature space, then SVM uses what is termed as the kernel trick. Kernel trick is one of the mathematical functions that maps the original feature space to a higher dimension space where the data could be separable by a hyperplane. It enables SVM to model non-linearity between the features and the target variable without actually mapping the data to a higher dimension. Some of the widely employed Kernels are; Linear kernel, Polynomial kernel, Radial basis function kernel, and Sigmoid kernels. Which kernel to select is not a principal concern of the user, but depends on the characteristics of the data set and the problem under solution.

Once the hyperplane is decided then in future any new data point can be labeled by simply seeing in which side of the hyperplane they lie. The separation by hyperplane is clearly defined and hence the model provides a clear distinction between the classes. Apart from the classification problems, SVM can also be applied in regression problems and these are called the support vector regression or SVR. Therefore, in SVR the aim is to look for such a function that brings out the dependency between the target feature and the input features with the least possible error.

Another advantage of SVM is its versatility in working with large numbers of features and thus is effectively useful when working with data having many features. Further, the most important advantage of SVM is its ability to work in cases where the number of features is larger than the number of samples. However, SVM can be computationally expensive and the required selection of a kernel and the appropriate hyperparameters can time-consuming and must be optimized to yield the best results.

More importantly, SVM is one of the most powerful machine learning algorithms that have impressive performance specially in classification where the number of features accompanied by a clear margin between classes. Due to its capability to build complicated decision boundaries by involving kernel functions, it might be very effective for solving many non-linear problems, although it depends highly on support vectors for increasing the reliability and speed in many uses.

### Logistic Regression

As for logistic regression, it is a commonly chosen method for classifiers of binary data where it is aimed to define the likelihood of an instance to be on one of the two categories. While linear regression is used to predict the quantity, logistic regression is used to predict the category, it models the probability of an event using sigmoid function.

This logistic function is able to map any real valued number into the interval [0. 1] and hence can be used for probability estimation. Depending on the result of the logistic function, it is possible to determine the probability of the current input to belong to the positive class. This probability is then compared to a threshold that is normally 0. 5, to arrive at a decision of classification. For the probability greater than the positive probability threshold, the model SHALL predict the member of the positive class; otherwise, the member of the negative class.

The input features are summed over the features comparable to the approach of linear regression using a weight sum of the feature values and a bias. However as opposed to it logistic regression does not output this linear combination straight away but it follows it by the logistic function converting it into a probability. The weights in the model, therefore, are obtained from the training data through a process known as back propagation which seeks to minimize a Loss function; in this case, the Binary Cross Entropy Loss which compares the probabilities predicted with the actual class labels.

The training procedure and the model consist of tuning the weights so that the loss function is minimized; this is often solved by gradient-based methods. Gradient descent continues to update the weights in the direction that lowers the loss which in turn makes the training data optimized to fit the model. Often adopted techniques ‘L1 (Lasso) or L2 (Ridge) regularization’ are used here to control and reduce maximum weights to limit overfitting of the model to go for a simpler solution that has more potential of generalizing over future instances.

The strength of logistic regression is especially manifested in the cases when the probability of occurrence of a certain event depends on the features that define this event in a linear way. It is also understandable as the weights show in which direction and at what value each feature affects the probability of the given class. However, logistic regression may not be very effective in catering for non-linear relationships between the independent variables and the dependent variable unless a feature engineering or transformation is done.

This is how non-geometric methods of Machine learning work; logistic regression is a particularly effective method of binary classification if the data set is linearly separable or if we want a probability value as our result instead of a class label. It is much used in many practical applications due to its efficiency and relatively simple to put into practice.

### Decision Tree

Decision tree is one of the most popular algorithms of supervised learning which can be applied to classification as well as regression analysis. The idea of the decision tree is based on the attribute test in each sub-sample or subset to construct a tree-like model of decisions. The elements used in decision tree include the nodes and edges and we have internal nodes, decision on a feature, edges, a result of the decision of the internal node, leaf node, final prediction or outcome.

Decision trees are constructed by partitioning data into the best feature to split in the root node. The idea is to find the feature which has maximum information gain and this measure indicates a feature by which classes are good separated or by which entropy of data set is minimized. There are numerous methods of splitting that can be employed: for example, Gini impurity, information gain, based on entropy, variance reduction in the case of regression.

Gini impurity is mostly used in classification problems. It is used to estimate how often, when a random element of the dataset is selected and randomly labeled according to the distribution of labels in the subset, the unlabeled element is misclassified. The Gini Impurity is the sum of probabilities of a node being in class i and a node being in class j, where class i is distinct from class j A lower Gini impurity in the node means that most of the samples are of one class and hence is termed as a ‘pure node’. Information gain on the other hand has an assessment of entropy which is a measure of disorder within the data set. The method is to choose the feature that gave the largest reduction of the entropy for the split.

After cross tabulation on the feature with the highest values, the data is split according the different values of this feature. This process is repeated for all the subsets thus bringing branches of the tree until any of the stopping conditions is reached. Stopping criteria could be the maximum depth of the tree or a minimum number of samples have to be present in a node or node has to be pure, all the samples in the node belong to the same class.

When for the task, it is a regression, then the decision tree provides a continuous output instead of the class label. The tree is constructed in the same way but the splitting is done according to a measure which aims at minimizing the variance of the subsets rather than certain classification measures such as Gini’s index or entropy.

Another strength of decision trees is that they are understandable. The structure of the tree can be illustrated in a clear manner so that the procedure of decision-making can be easily understood. Every path from the root to a leaf corresponds to a rule or decision path which is easy to see exactly how the predictions are arrived at from the input features.

Decision trees can also be non-parametric, that means that it does not presuppose any distribution of data and, therefore, can unearth intricate relations between features. It can be applied for nominal and ordinal data and the most important thing is that they are versatile.

However, there are some disadvantages that can be mentioned in connection with decision trees. They are however, prone to over fitting especially when the tree is made deeper and starts to capture noises within the data and not the actual trend. This can lead to poor generalization to unseen new data Hence, having a small validation set is a good indication that the model is not generalizing well to new unseen data. To avoid overfitting parameters there are approaches like pruning where least important branches are shaved off or constraints where, for example, maximal depth or minimum samples per leaf or node are set.

Decision trees are also: ‘unstable’ – small changes to the data result in large changes to the structure of the tree. This issue can be managed by the usage of assembling methods, for instance, Random Forest or Gradient Boosting that are built from multiple decision trees.

In conclusion, decision trees are a versatile productivity in machine learning with a booth and accurate performance and, therefore quite effective when applied with combination of additional elements to combat with their disadvantages.

### XGBoost

XGBoost is an optimized version of Gradient Boost is called Extreme Gradient Boosting used for solving binary as well as multiclass classifications and regression problems. It is characterized by high speed, great performance, and high scalability and that is why it is popular in competitions in the field of machine learning and practical use. First, XGBoost is an extension of the boosting technique, which is a type of ensemble learning, the process of incrementally fitting several ‘weak’ prediction functions in an attempt to improve the function’s accuracy.

The general principle of boosting is the successive training of a series of ‘weak’ classifiers where each subsequent classifier seeks to correct the mistakes made by the previous one. XGBoost is an advanced optimization of gradient boosting where the parameter-space is divided into subsets which enables faster training and improves the prediction quality of the model.

Within the construct of XGboost, gradient boosting is chief, and it built by reducing the loss function with the help of gradients and its calculation by gradient descent. The algorithm begins with a first model that is traditionally the decision tree that is trained with the training set. The actual or real target values are also estimated and then the residuals or the differences between the predicted and actual target values are computed. In each of the subsequent iterations, the data is re-trained with the residuals which implies that each of the new models is learning to rectify the mistakes of the now precedent model. This process continues for the given n number of iterations or until the residuals are of an acceptable magnanimity.

Another of the major improvements in XGBoost is the utilization of the different methods of regularization to reduce over learning. The concept of regularization is an attempt to minimize the risk by introducing a term for model complexity, so the algorithm will have to find the best way how to fit the training data as well as how to generalize to unseen data. XGBoost introduces two types of regularization: L1 (Lasso) which helps in making the model sparse and second is L2 (Ridge) in which large coefficients values are penalized. This regularization aids in achieving levels of smoother and less over-fitting of the model to the data as a way of predicting new data.

Another particularly significant aspect of XGBoost design is second-order derivatives of the loss function to improve the model updating process. Though, basic gradient boosting provided by ML is function of only the derivative of loss function while, XGBoost uses up to the second order derivatives known as ‘Hessian’. It also makes it easier for the algorithm to approximate the curvature of the bowl in which the loss functions are placed, and hence improve the delta or change in the model parameters.

XGBoost also contains several engineering optimizations that have been performed to improve the efficiency and scalability of the algorithm. These are; The employed algorithm is a learner that can deal with missing values and sparse matrices this makes it suitable for real datasets since most of them contain a number of missing values. Moreover, XGBoost works with the blocks of the data, which enables a more successful usage of such resources as multi-core processors or distributed computing systems. This ability of parallelization helps reduce the training time, and especially in large datasets.

The algorithm also provides for tree boosting using exact greedy, approximate as well as the histogram-based techniques where each of them has different aspects of its applicability namely speed and accuracy. Exact greedy boosting involves the construction of each tree when all possible split is considered with a guarantee of selecting the best. While the Approximate boosting is faster due to the use of heuristics to obtain the approximate promising splits. Histogram based boosting divides the continuous variables into intervals to increase efficiency of the decision tree construction in addition to possibility of reducing the number of splits to be considered.

Feature importance is another important aspect that is available in XGBoost. The last possibility is that after training, the algorithm can explain which features were the most important in terms of the model’s choice. This information can then be used to decipher other patterns in the data set and can also be employed in selection of features in other models that will be constructed.

Besides, XGBoost is much flexible in other aspects; the algorithm has a number of hyperparameters that can be adjusted for achieving better performance for a specific dataset and a certain task. These are the hyperparameters such as learning rate, maximum depth of the trees, the minimum weight of a child node, subsampling ratio among others. These parameters can be tweaked by a practitioner so that the best performance is optimized out of the model.

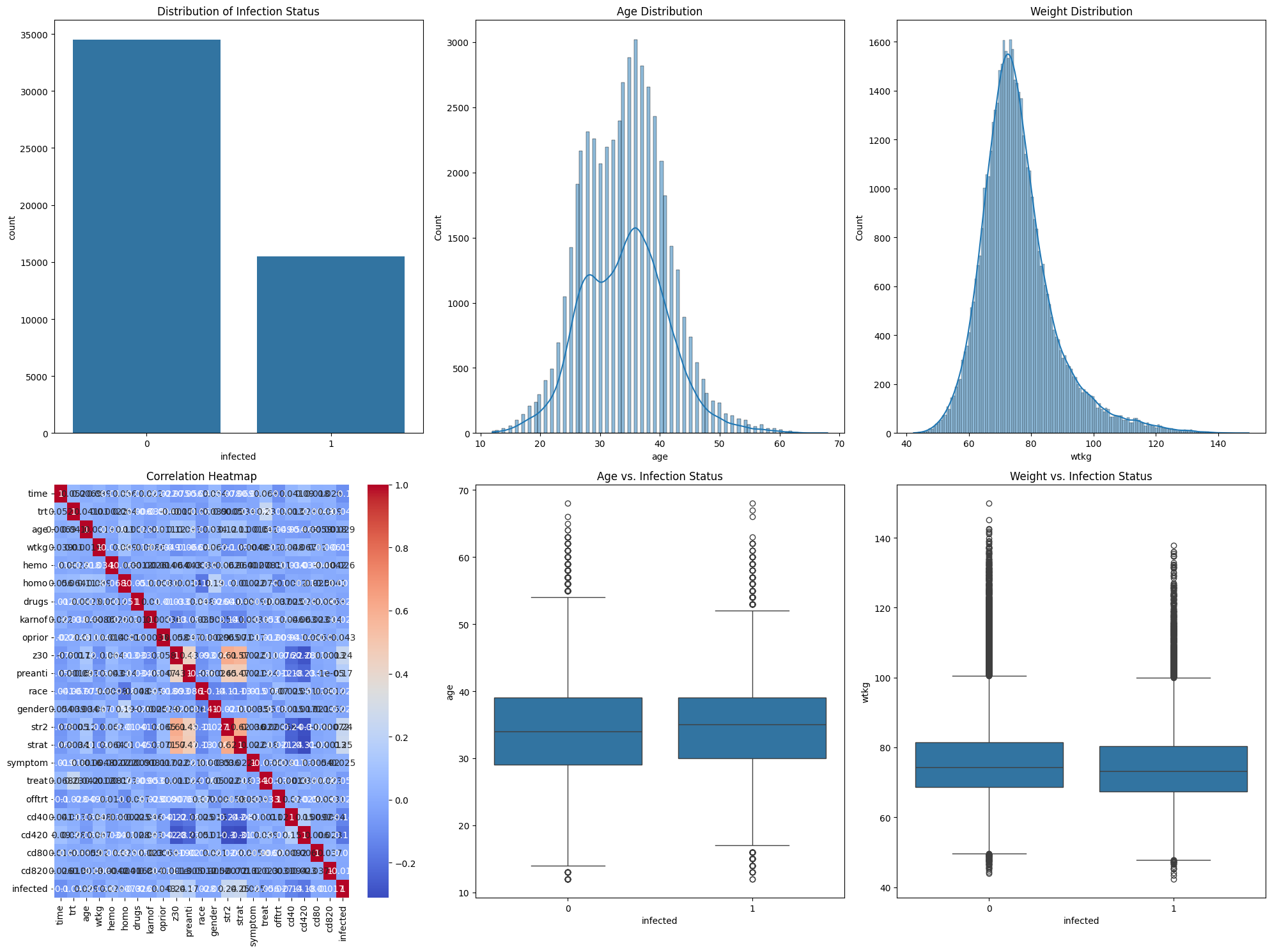
There are a number of advantages of XGBoost to be able to work effectively with different types of data on tabular structure and structure time series data in addition, it is able to work with a certain type of text and image data after preliminary data preparation. However, this versatility, alongside with high robustness and speed, makes XGBoost one of the most popular tools in the machine learning field at the present time.

In conclusion, XGBoost may be viewed as a state of the art and efficient library of gradient boosting algorithm for both classification and regression. Being capable of using regularization, second order optimization and other high end engineering it is an optimal tool when it comes to developing predictors. Furthermore, the efficiency in term of scalability and advanced ability to process big data and high dimensionality combined with flexibility and versatility and tuning capability put XGBoost as one of the most important algorithms in the machine learning world.

# Chapter 3: Methodology

## Tools and Techniques

## EDA and Visualization



# Chapter 4: Results and Conclusion

## Critical Discussion

## Conclusion

## Future Work

# Chapter 5: Legal, Ethical and Professional Issues

# References

# Appendices