HIV/AIDS Analysis

Human immunodeficiency virus infection / Acquired immune deficiency

Group Members

Rahulkumar Nasit

Bhavesh Mishra

Avikumar Patel

Contents

Introduction

There are huge amounts of data generated in health care sectors, which can be used to make learn or find the meaningful pattern in the data. A crucial stage of data analysis is exploratory data analysis (EDA). A machine learning (ML) model can also be highly beneficial for the early diagnosis and treatment of persons with cardiovascular illness or who are at high cardiovascular risk.

The virus known as HIV (human immunodeficiency virus) targets the immune system of the body. AIDS (acquired immunodeficiency syndrome) can develop from HIV if it is not treated and increasing the risk of cancer and life healthcare-threatening infections. There isn't a cure that works right now. People who contract HIV are permanently infected. But HIV can be managed with the right medical attention.

In the majority of cases, HIV is a sexually transmitted infection. HIV can, however, also be passed from a mother to her child through breastfeeding, pregnancy, or childbirth. Sharing injectable supplies, like as needles, can also result in non-sexual transmission.

As, currently there is no cure for HIV infection, which makes it a huge global public health concern. Though HIV infection has become a manageable chronic health condition, growing access to comprehensive HIV prevention, diagnosis, treatment, and care—including for opportunistic infections has allowed people living with HIV to lead long and healthy lives. Both the individual and society suffer when HIV is discovered too late. The improvement of early HIV diagnosis must be a top focus in healthcare.

There is not so much analysis found on the internet regarding HIV/AIDS. That is why our main objective in this project is, we will thoroughly examine the global data on HIV/AIDS and create some insightful visualizations. Additionally, we'll develop a machine learning model that can forecast how HIV patients' treatments will develop. By using machine learning's classification technique, we will create the prediction model to check the progression in HIV patients that give us the best result.

Related work

This section's major goal is to highlight and contrast the results obtained by one group with those of another. Since this is to ensure that our present project results are comparable to or in line with theirs, the fields of work should be similar.

The other group members have worked on the following things in python files for the HIV/ADIS datasets

- Using hiv-death-rates-by-age to examine the age groups with the highest mortality rates.
- Examine life expectancy table
- Examine new cases of hiv infections

Although they only used three datasets in this, the dataset itself comprises all 31 datasets. These datasets have been used to determine which age groups have the highest death rates, life expectancy trends over time, the best countries for life expectancy, and the nations with the fewest new infections. We decided to work with only a few datasets initially, but as we moved forward, we continued to use all of the datasets. We finished our work with the 31st dataset. Additionally, we've cleaned up and created some insightful data about the HIV/AIDS epidemic using Tableau.

For the model prediction, they have used the logistic regression and KNN model to predict the progression of the HIV patients with given dataset. Both the KNN model and the logistic regression gave them an accuracy of 82%. However, their recall and precision ratings were low. Therefore, we have experimented with many categorization models, including Random Forest, Decision Tree, Support Vector Model, and Naive Bayes Model. Additionally, we created confusion matrices for each machine learning models.

Methods

We will briefly describe the procedures followed, exploratory data analysis, preprocessing methods, machine learning models, and the assessment metrics we plan to utilize for our dataset in this section.

HIV/AIDS datasets:

We have used tableau for the cleaning operations such as filtering, renaming, splitting or remove the fields as it easier to combine and analyze the data and preparing visualizations. For example some the datasets contain the negative value in year columns we have to removed or ignore those values.

HIV patients' Progression Predict datasets:

Importing Libraries, Packages, and Dataset

We imported the necessary libraries and packages that are required to run a particular area of code or all of the code sections in the Python notebook. Then, for later processing, we imported our dataset.

Checking for Outliers

Outliers are data points that are far from the other observations or data points. Although the data may have been entered wrongly, a random variation may have allowed us to discover anything of interest. Using this information, we examined outliers in our columns to comprehend the distribution of data points.

Data Preprocessing

The process of preprocessing data is used to make the data more useable and comprehensible. Preprocessing the data is always a good idea because the raw data is usually imperfect and may lead to errors later.

We have checked for the null values in columns of the dataset that we have used for prediction, and we found that there in no missing values in the columns.

Machine Learning Models

A machine learning model is a file that has been taught to recognize specific patterns. An approach that enables a model to analyze and learn from the data is used to train it.
We used 5 Machine Learning Models for our dataset. The algorithms are

- Logistic Regression
- Random Forest Model
- Decision Tree Model
- Support Vector Machine Model
- KNN model
- Naïve Bayes Model

Evaluation Metrics

Evaluation metrics are used to describe a model's capacity. These measures are employed to assess the model's effectiveness. These measures must be used since they provide a clearer picture of how each model is operating at its highest level and what needs to be changed to raise the scores later.

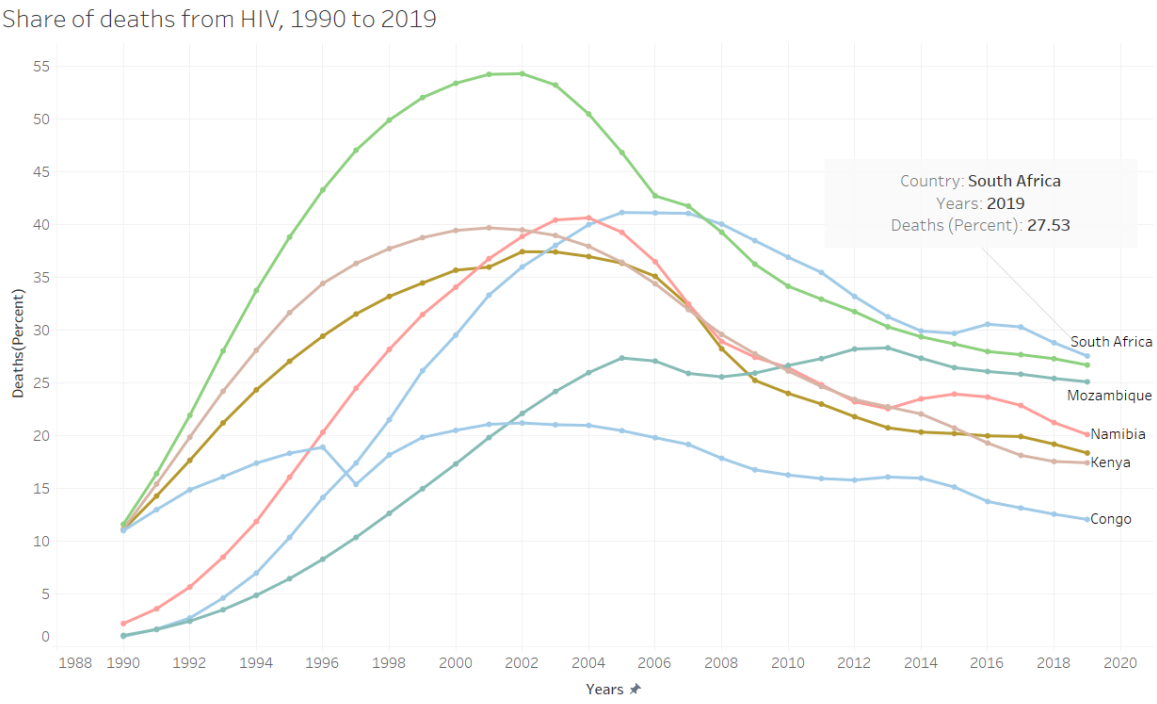
The metrics we have used

- Classification Report
- F1 Score
- Precision
- Recall
- Accuracy
- Confusion Matrix

Results

HIV/AIDS datasets:

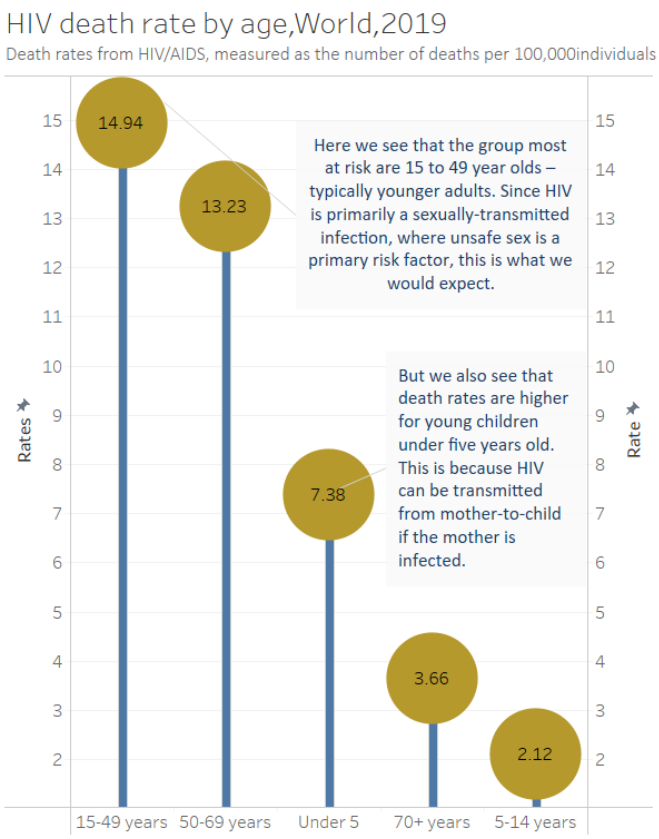
Here, we have mentioned some of our key results.

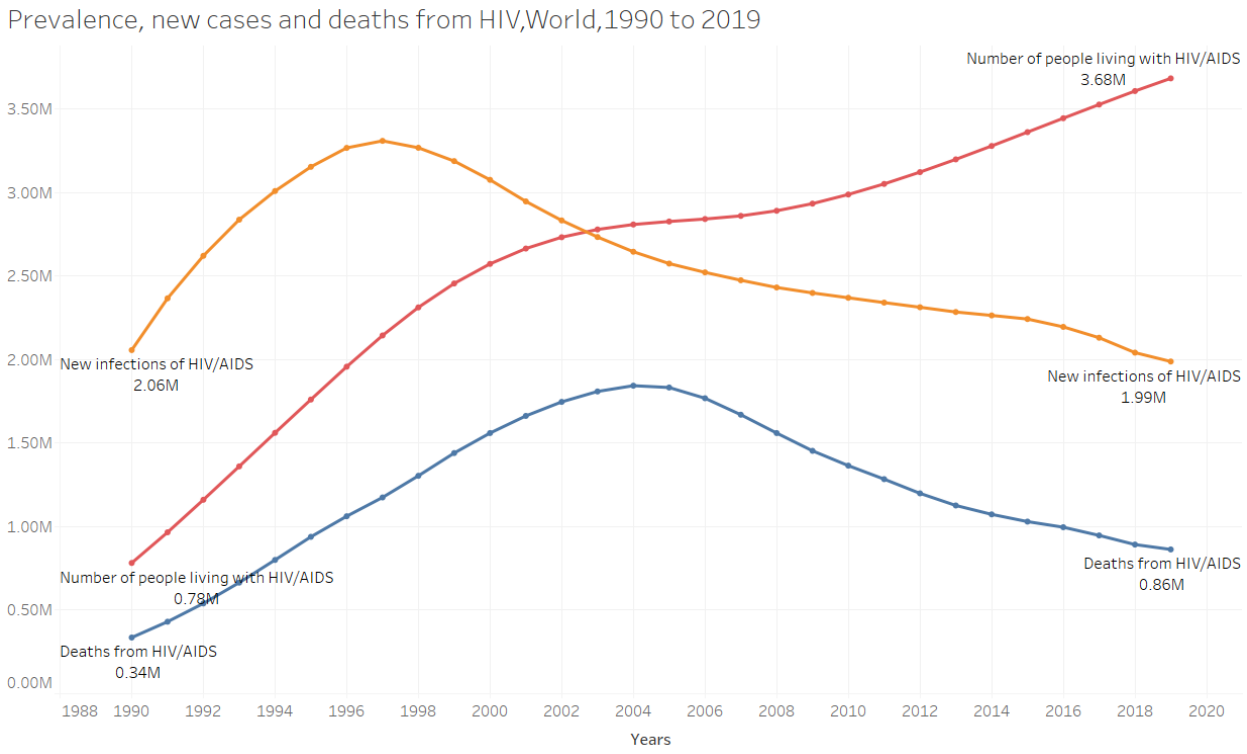


HIV/AIDS contributed to 1.5% of deaths worldwide in 2019.

However, a number of countries, especially in Southern Sub-Saharan Africa, have a rather high percentage. In South Africa, 28% of all fatalities in 2019 were caused by HIV/AIDS. A number of countries with sizable market shares included Botswana (27%), Mozambique (25%), Namibia (20%), Zambia (18%), Kenya (17%), and Congo (12%).

Children under the age of five and younger people have the highest mortality rates.



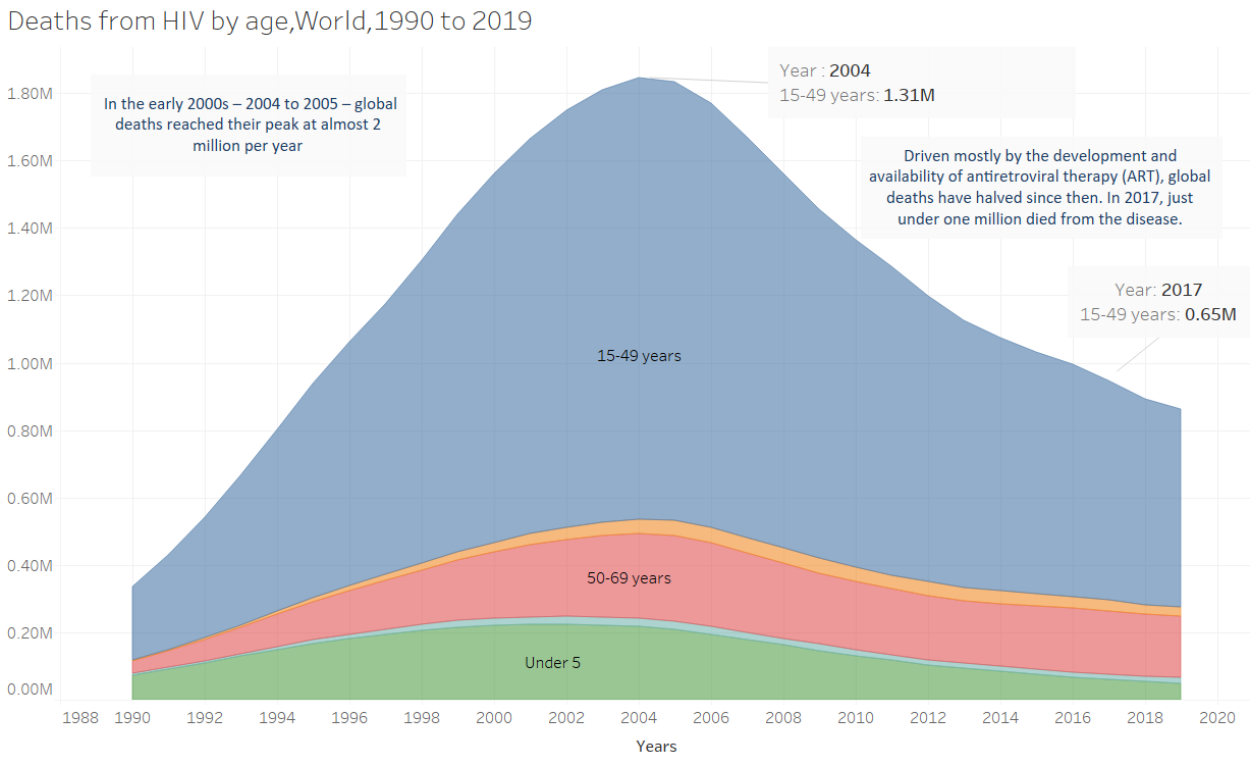


The 1990s saw a sharp rise in the number of AIDS- and HIV-related fatalities.

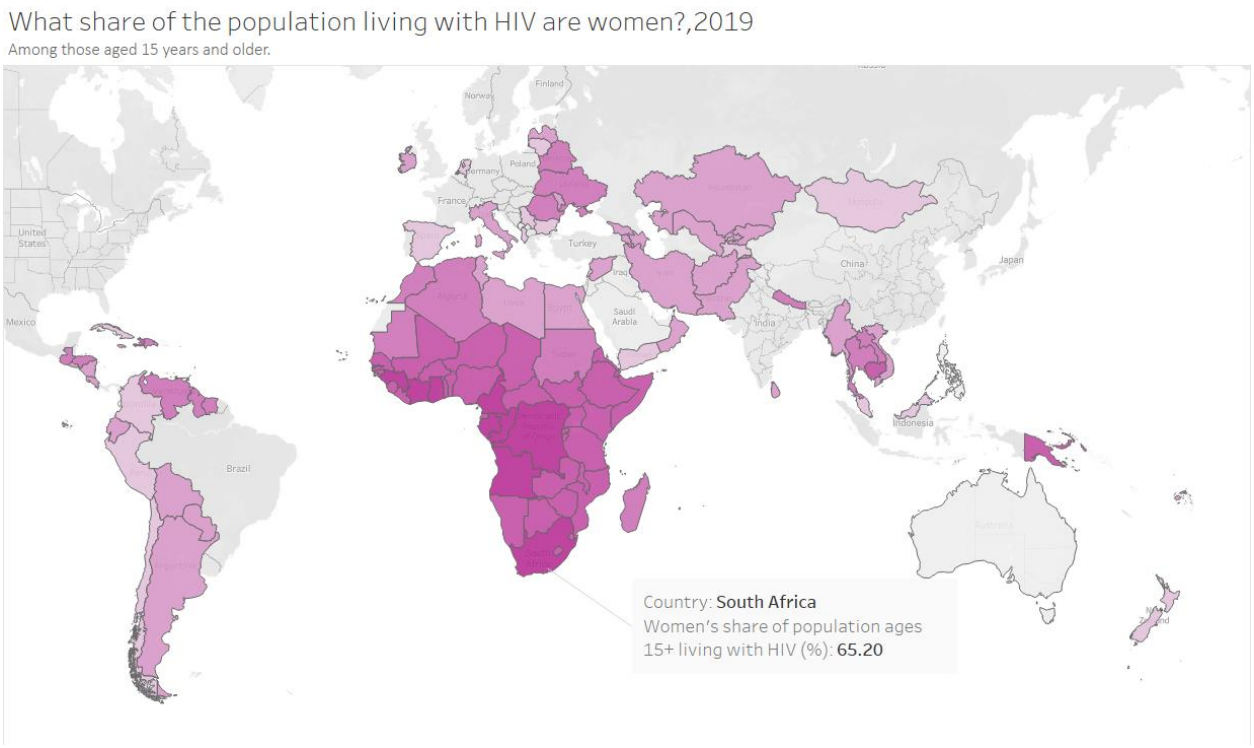
Every year between 1996 and 2001, more than 3 million people contracted HIV. After that, there has been a decrease in the number of new infections, which reached below 2 million in 2019.

Throughout the 1990s, there was an increase in AIDS-related deaths, which peaked in 2004 and 2005, when about 2 million individuals perished. The graph also demonstrates the ongoing rise in the number of HIV-positive people.

Within a decade, the number of people dying from HIV/AIDS worldwide was cut in half.



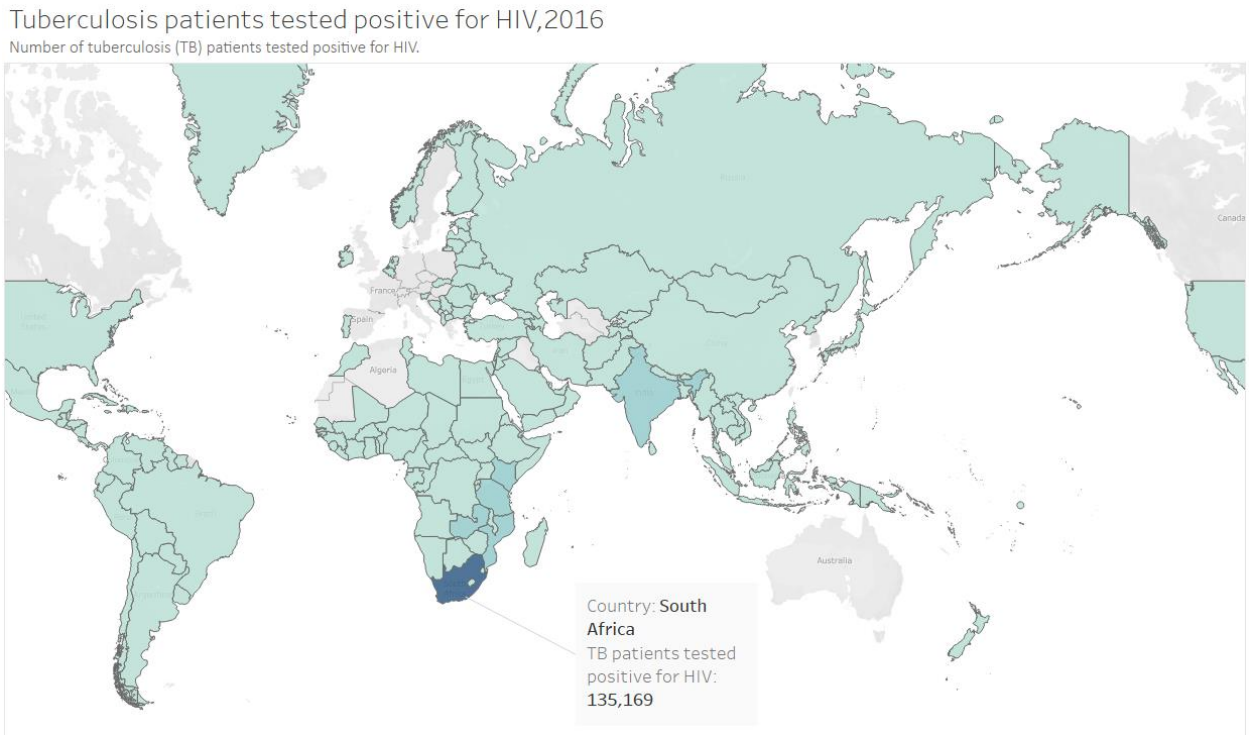
Is HIV/AIDS more common in men or women?

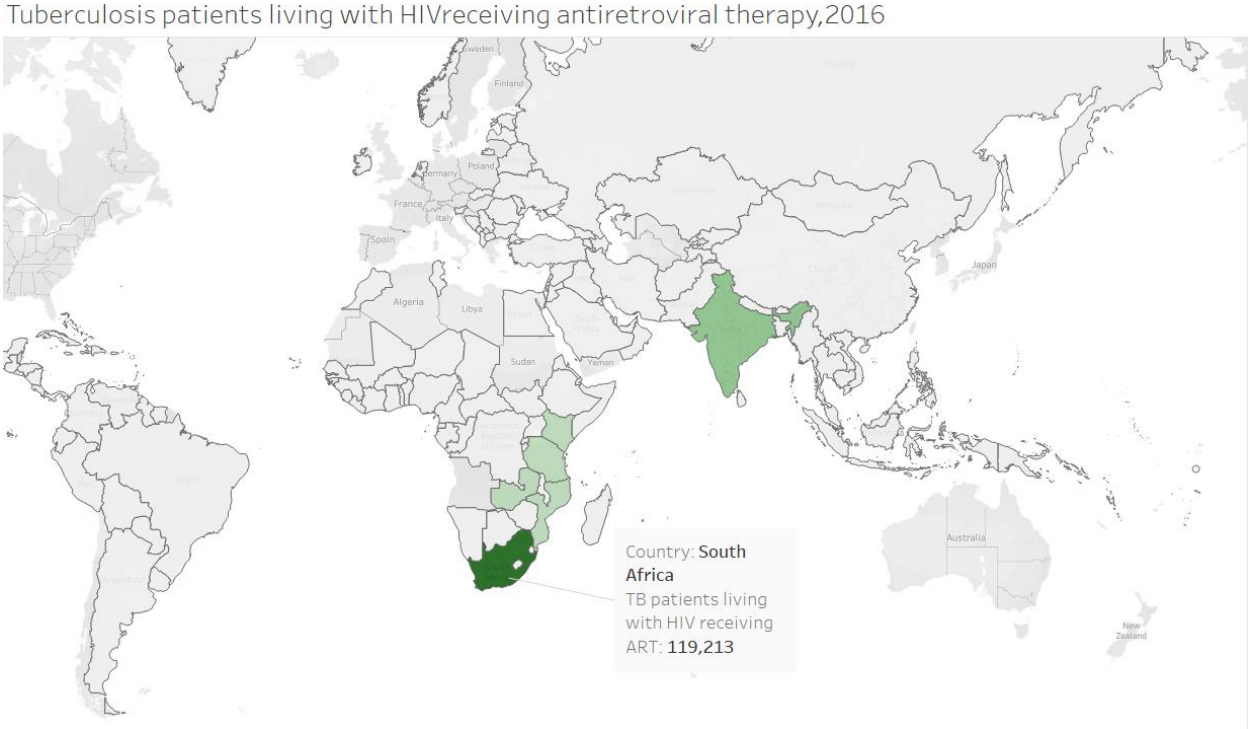


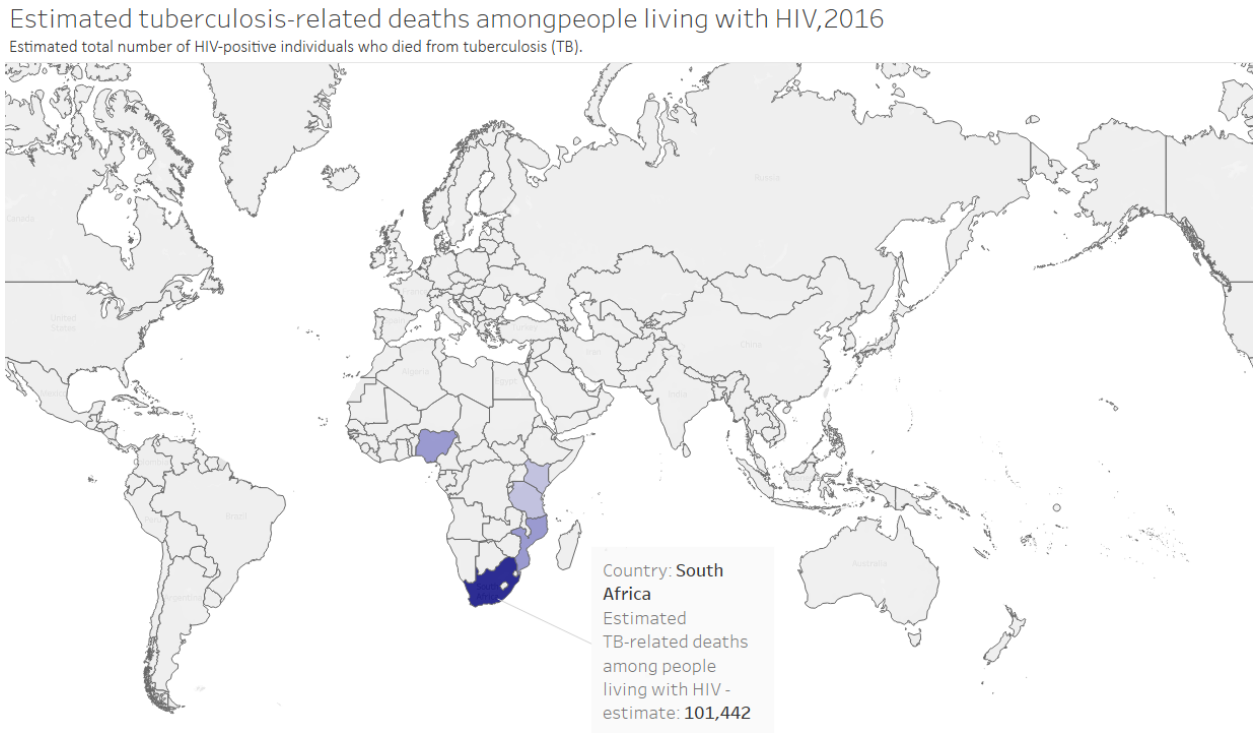
Men and women have different rates of AIDS-related death and HIV prevalence, respectively. The percentage of women who are HIV-positive is depicted in the graph.

As we can see, while HIV prevalence is higher in men in most other locations, it tends to be higher in women in South Africa. In opposition, more men than women often pass away from AIDS each year.

HIV-related tuberculosis among people

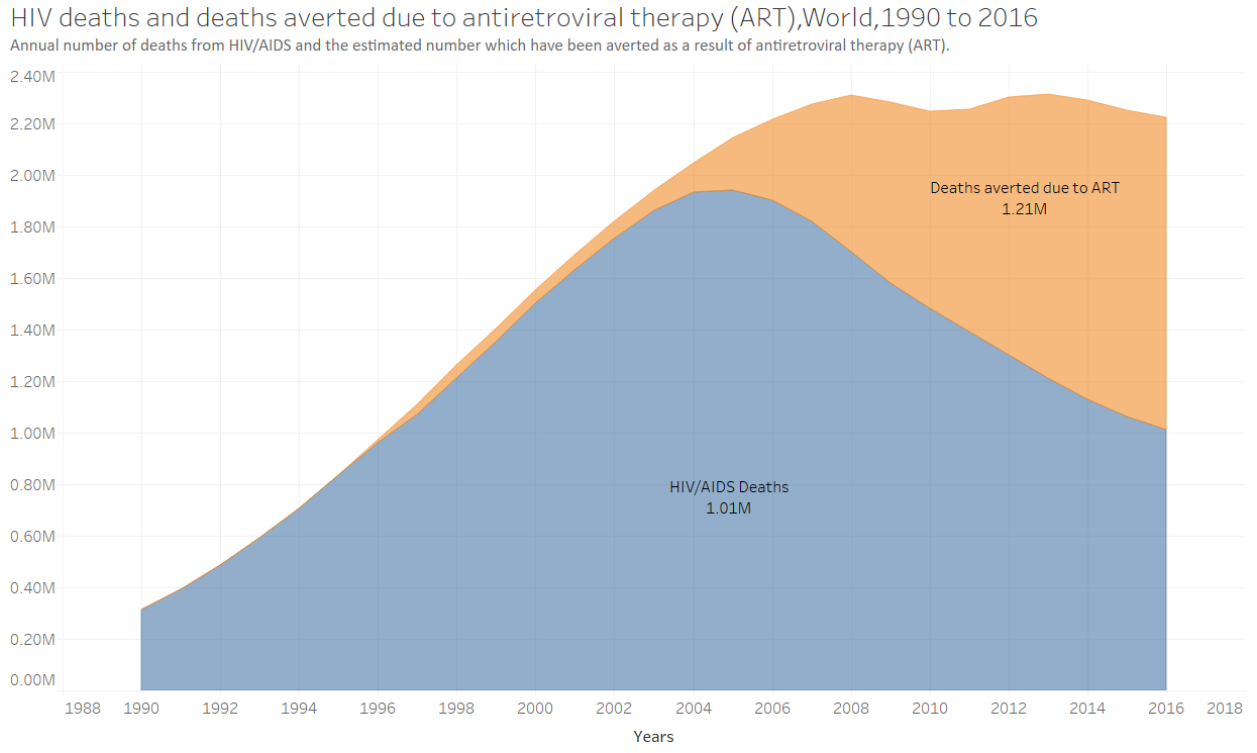






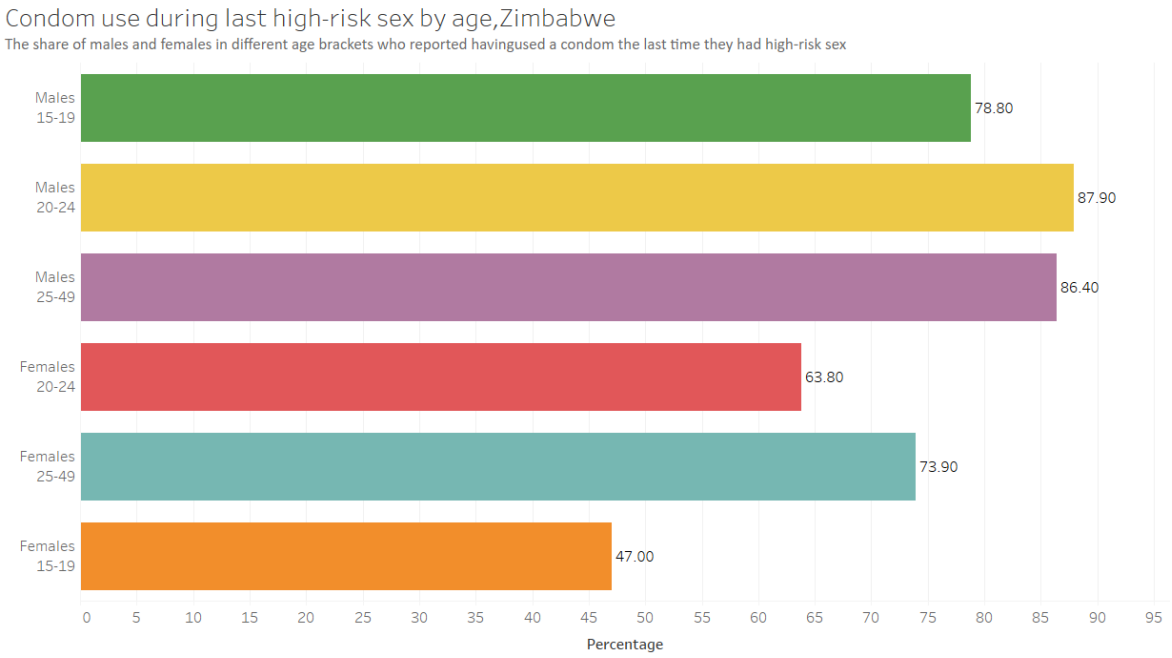
How can HIV/AIDS be prevented?

A few decades ago, the possibility of having HIV for longer than ten years considered incredibly unlikely. Longer lifespans are now a realistic expectation for those with HIV/AIDS because to antiretroviral therapy (ART).



Since ART is introduced in late 1980s, very large number of lives have been saved. The area graph shows the number of HIV related deaths that happened each year as well as the number of deaths revert due to ART.

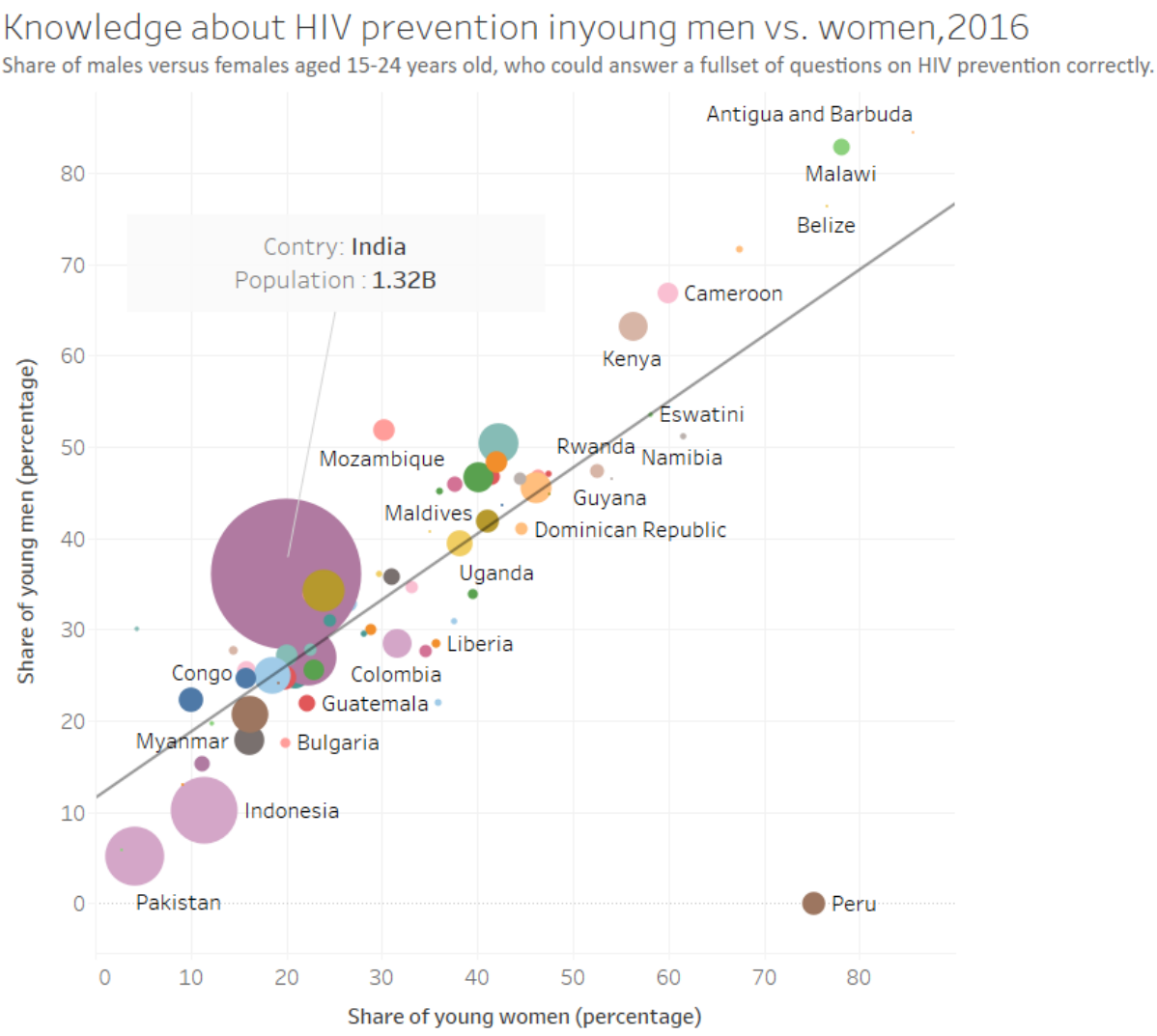
Share of people practicing safe sex



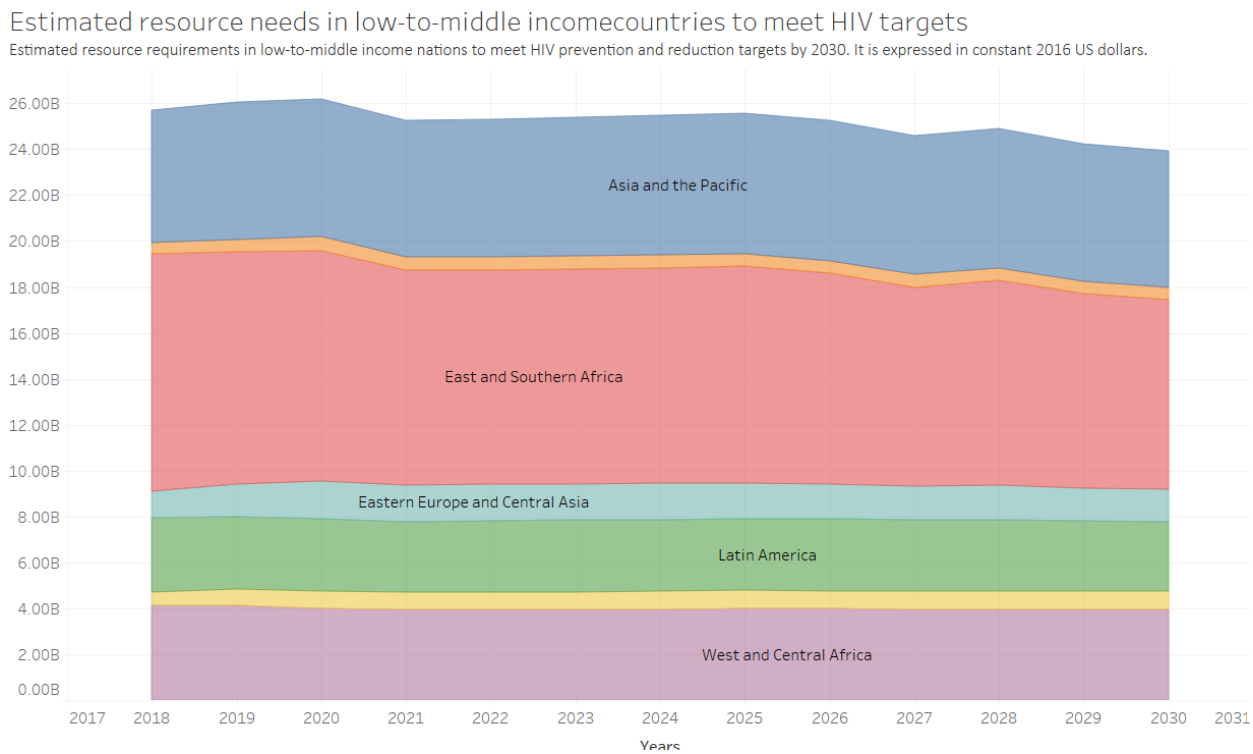
HIV is primarily spread through sexual activity.

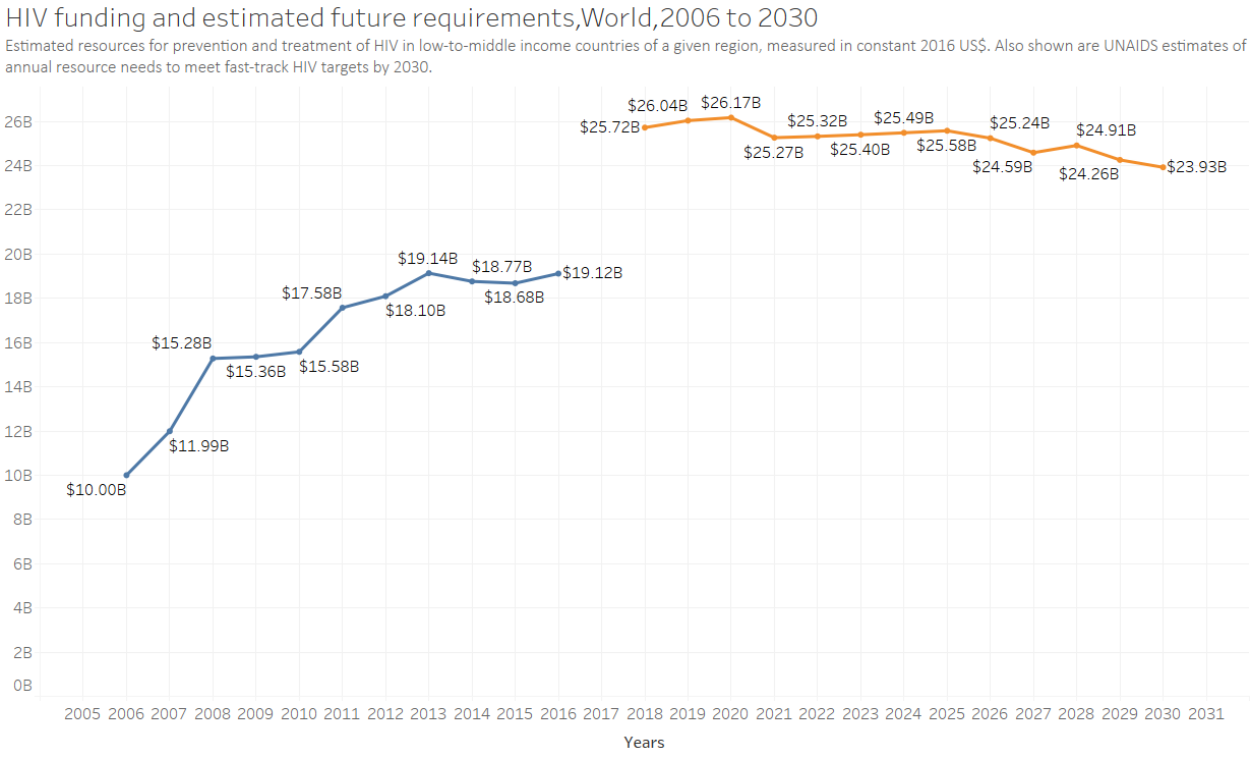
Condoms can be used to prevent sexual transmission In situations of "high-risk sex," which is defined as sexual activity with a partner who is not a spouse or roommate, condom use is more prevalent, as seen in the charts above.

Education on HIV/AIDS



financing for the fight against HIV/AIDS





HIV patients' Progression Predict datasets:

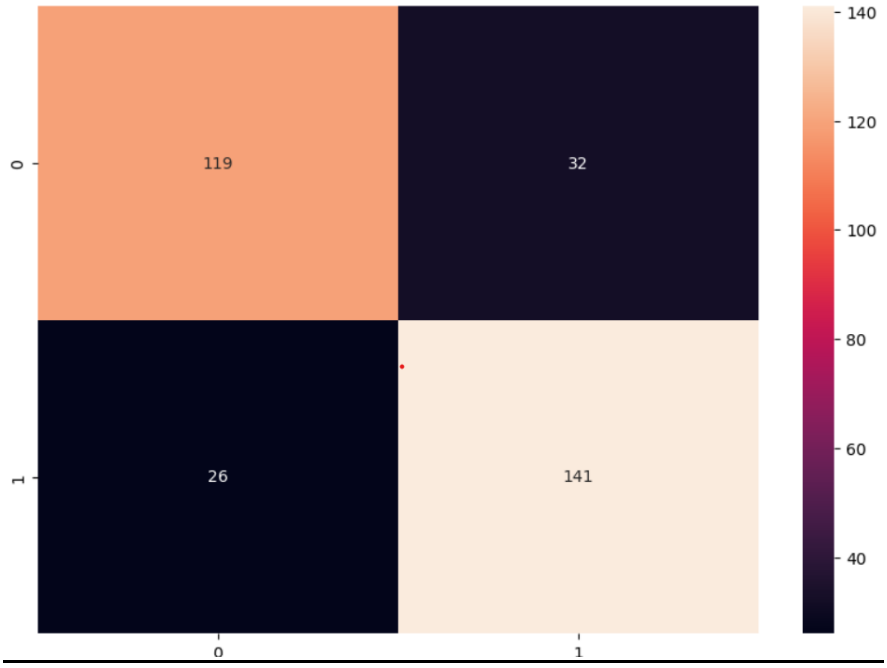
Following are the models that we have used for the progression prediction with it accuracy
Logistic Regression – 75%
Random Forest – 82%
Decision Tree – 80%
Support Vector Machine – 60%
KNN – 69%
Naïve Bayes – 74%

Performance of the best two models

Random Forest Model

	precision	recall	f1-score
0	0.82	0.79	0.80
1	0.82	0.84	0.83
accuracy			0.82

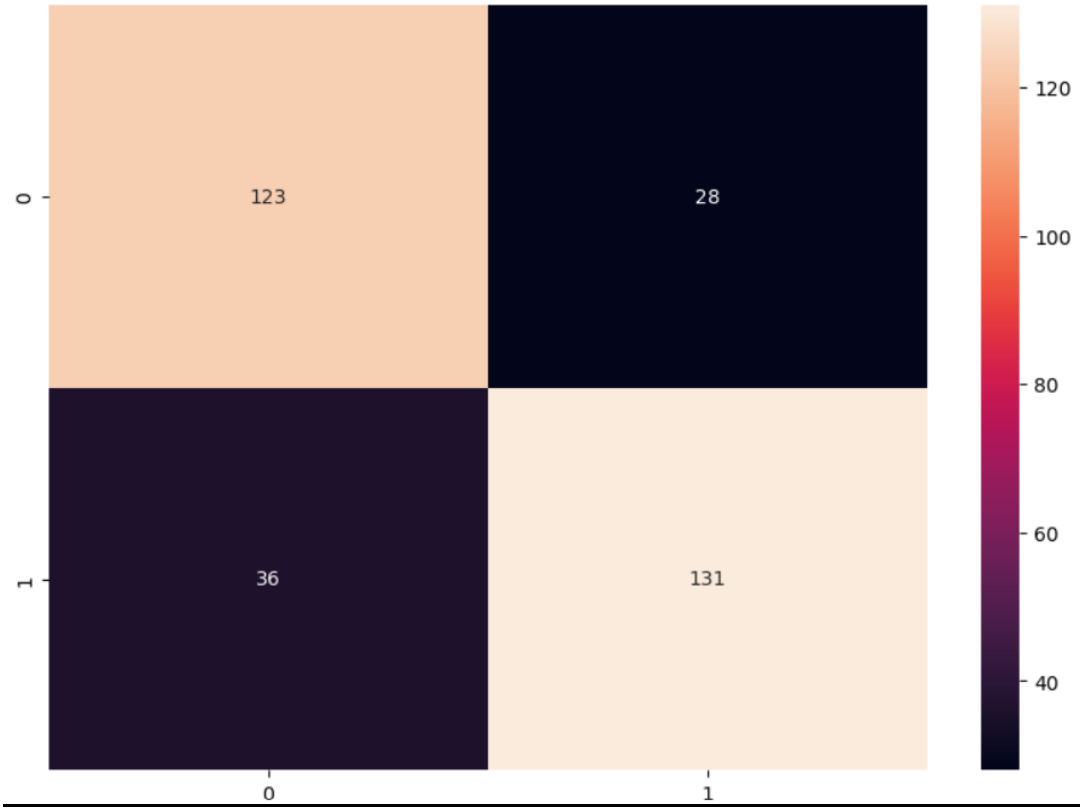
Confusion matrix



Decision Tree Model

	precision	recall	f1-score
0	0.77	0.81	0.79
1	0.82	0.78	0.80
accuracy			0.80

Confusion matrix



Discussion

We have finished compiling the comprehensive analysis of the global HIV/AIDS data. We were unaware of any other methods of HIV protection besides avoiding sex until we worked on this project, which taught us about the ART therapy used to lessen the HIV virus. We can learn about

HIV/AIDS through the initiative learning more than just these things. As a result, we can state with pride that we succeeded in achieving our goals.

Regarding related work, we would like to note that our work is unquestionably consistent with that; in fact, we accomplished considerably more. The biggest problem we ran into was trying to create an HIV prediction model with the main datasets because there were some restrictions. So, in order to solve this problem, we must use a different dataset. After further research, we have ultimately settled on one and created a machine learning algorithm-based model to predict HIV patient progression.

Conclusion

By doing the analysis of the HIV dataset, we have concluded the following things:

- With over one million deaths every year, HIV/AIDS is the leading cause of death in various countries.
- About one million deaths per year are found in some countries due to HIV/ADIS
- Sub-Saharan Africa has the highest mortality rates.
- Younger People and children experience the highest rates of death when HIV is transmitted from a mother.
- Over the past ten years, the number of deaths globally has reduced by 50%, showing that things are getting better. Life expectancy in Sub-Saharan Africa was significantly impacted by HIV.
- The prevention of AIDS-related deaths has relied heavily on antiretroviral therapy (ART).
- If the world wants to reach its 2030 targets, funding for HIV treatment and prevention needs to grow.

Contributions

Name	**Contribution**
Rahulkumar Vitthalbhai Nasit	Project work Work on the following set of datasets with data exploration and visualizations • HIV Deaths • People Living with HIV • HIV Prevention Work on the following prediction model for HIV patient progression • Logistic Regression

	• Random Forest • Decision Tree Final Report Related work | Results
Bhavesh Trigunanand Mishra	Project work Work on the following set of datasets with data exploration and visualizations • Share of population of HIV • TB Patient with HIV Work on the following prediction model for HIV patient progression • Support Vector Machine Model • KNN Model • Naïve Bayes Model Final Report Methods | Discussion
Avikumar Patel	Project work Work on the following set of datasets with data exploration and visualizations • HIV Expenditure – Current status and Future Requirements • Life Expectancy • Work on power point presentation Final Report Introduction | Conclusion

References

Datasets
https://www.kaggle.com/datasets/programmerrdai/hiv-aids?resource=download&select=life-expectancy.csv
https://github.com/jerofad/HIV-1_Progression-Prediction/tree/master/data

Machine learning
https://learn.microsoft.com/en-us/windows/ai/windows-ml/what-is-a-machine-learning-model
https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/
https://www.kaggle.com/code/mohammedsakibrahman/hiv-progression-prediction-newbie-approach

HIV/AIDS
https://en.wikipedia.org/wiki/HIV/AIDS
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7260108/
https://github.com/jerofad/HIV-1_Progression-Prediction
https://ourworldindata.org/

Front page image
https://news.emory.edu/stories/2012/07/international_aids_meeting_emory/campus.html

Appendices

- **HIV-Report.docx:** Contains the final report
- **HIV.twb:** Includes all findings with visualization for the HIV/AIDS
- **Healthcare.ipynb:** Includes all of the project's source code, including import, preprocessing, exploratory data analysis, machine learning algorithms, and assessment metrics.
- **Github link:** Contains all the above-mentioned files
  https://github.com/rn34/Healthcare