**STANDARDIZING THE REPORTING OF HIV TESTING INDICATORS: ADDRESSING INCONSISTENCIES IN DATA COLLECTION ACROSS FACILITIES USING AI MODELS.**

Isaac Kiplangat, Winfred Muusi, Jacob Mutiso, Ividah Francis, Limo Duke

HEALTH IT HACKATON, KABUO KABARAK UNIVERSITY,

**BUSINESS UNDERSTANDING**

**INTRODUCTION**

The reporting of HIV Testing indicator (HTI) is a critical element in monitoring and evaluating HIV testing. However inconsistent reporting across facilities leads to data gaps and inconsistencies in the estimates of HIV testing coverage. In this study, we apply data science approaches to identify different data Behaviour and to automate the cleanup of the data. Specifically, we develop a classification model to assist in the identification of prevention of mother to child transmission (PMTCT) sites that do not report tests. Our model is based on decision trees and logistic regression and uses features such as county, facility, testing location, and number of tests reported. We apply the model to the datasets of facilities and use the output to identify PMTCT sites that are not reporting tests. The model produced an 82% accuracy We then provide estimates to fill the data gaps and automate the clean-up of the data using scripts and software tools. Our approach can help to improve the consistency and accuracy of HIV testing data and support better decision-making in the context of HIV Prevention and control.

**CLIENT /USER NEEDS**

The user requires a system that will identify sites or facilities that do not submit accurate on none HIV testing data to the HTS system.

The user requires that the PMTCT data is in sync with HTS, and that there is no inconsistent data across facilities and that user facilities who submit none or incorrect data is flagged and can be able to be tracked easily.

**CLIENT ENGAGEMENT PROCESS**

We sought to understand the task and the requirements of the model, we achieved this by being in constant communication with our Facilitator Moses Njatha, who shared with us to understand the data, features a, variables of the project.

To understand the model, we analyzed the dataset to ascertain the requirement needed to setup a functional model to use to save inconsistence of data within the systems.

**OBJECTIVE OF THE TASK**

The objective of the challenge is to apply Data Science /Ai approaches to identify different data behaviors, advise on estimates where gaps are found and to automate the cleanup of the data

**DATA ACQUISITION**

**DATA SOURCE SYSTEMS**

In this section we describe Four datasets which we used, which was provided in excel files by the Health It. based on HTS and PMTCT to be used to train and test the classification model. The datasets indicate a comparison of two states of data, HTS, describes HIV Test conducted and reported for general population to the HIV testing service (HTS) sites, while dataset labeled PMTCT indicates HIV test conducted for expectant women and mother tested and reported to PMTCT sites.

**DATA ACQUISITION PROCESS**

Data was acquired based on gender, age, location, test period a

Dataset also contained data collected from different period in years, and could be use to create a data model to classify the HIV test based on period and who did not submit on the different periods

**EXPLORATORY DATA ANALYSIS**

**EXPLORATORY DATA ANALYSIS PROCESS**

Collect the data: Gather all available data related to HIV testing at the facilities in question, including information on the types of tests conducted, the testing locations, the time period covered by the data, and any relevant demographic information.

Inspect the data: Review the data to get a sense of its quality and completeness. We check for missing values, outliers, and other anomalies that may need to be addressed before proceeding with further analysis.

Explore the data: we conduct a range of descriptive statistics and visualizations to better understand the data. For example, you might create histograms or boxplots to examine the distribution of different variables, scatterplots to explore potential relationships between variables, or heatmaps to identify patterns in the data.

Identify patterns and relationships: we use exploratory data analysis to identify patterns and relationships in the data that may be relevant to understanding the reporting consistency of the HIV Testing Indicator. For example, you might identify differences in testing patterns between facilities or between different subgroups of the population.

Select potential features: Based on the patterns and relationships identified in the EDA process, select potential features that may be relevant to improving reporting consistency. For example, you might identify specific types of tests or testing locations that are more likely to be excluded from the HIV Testing Indicator, or demographic factors that may impact reporting consistency.

Assess data quality and completeness: After selecting potential features, assess the quality and completeness of the data for those features. Determine if any additional cleaning or processing is necessary to ensure that the data is suitable for modeling.

Prepare data for modeling: Finally, prepare the data for modeling by encoding categorical variables, normalizing continuous variables, and creating training and testing datasets. The EDA process can help inform decisions about how to handle missing data, outliers, and other issues that may arise during this process.

**EXPLORATORY DATA ANALYSIS OUTCOMES**

Testing locations vary widely: The data shows that facilities use a wide range of testing locations, including hospitals, clinics, and dispensaries. This suggests that the inclusion/exclusion of certain testing locations could be a factor in the reporting consistency of the HIV Testing Indicator.

PMTCT testing is excluded in some cases: The data indicates that some facilities exclude testing results from Prevention of Mother-to-Child Transmission (PMTCT) clinics, even though these tests are relevant to the HIV Testing Indicator. This could be due to differences in reporting protocols or challenges in integrating data from multiple sources.

Testing patterns differ by demographic group: The data shows that certain demographic groups (e.g., age, gender, race/ethnicity) are more likely to undergo HIV testing than others. For example, more individuals and women aged (20-24) got tested compared older individuals or men. This suggests that demographic factors could play a role in the reporting consistency of the HIV Testing Indicator.

Missing data: The data contains a number of missing values, particularly in the demographic variables. This could make it more difficult to analyze the data and could impact the performance of the decision tree model. Addressing missing data will be important in preparing the data for modeling.

**DATA CLEANING**

**DATA CLEANING PROCESS**

Data cleaning is the process of preparing data for analysis by identifying outliers and transforming the data to make it easier to analyze.

Our dataset contained a large no of variables, empty or missing values in some column, the also existed duplicated values. This type of data we apply a

**DATA CLEANING OUTCOMES**

Outcomes realized during the data cleaning process are;

Increased data quality: by identifying and correcting errors, filling in missing data, and standardizing variables, the quality and reliability of the data are improved making it useful for analysis.

Improved accuracy: cleaning the data reduced inaccuracies in the reporting of the HIV Testing Indicator by removing outliers, removing duplicated data, making data more accurate.

Identification of data gaps: through the cleaning process, we identified missing data and incomplete information, and allowed us to prioritize effort to address quality issues.

**FEATURE ENGINEERING**

**FEATURE ENGINEERING PROCESS**

Feature engineering involves creating variables or features from existing datasets to improve performance of machine learning models or gain insight from the data. we applied the following steps

Create a variable to capture the location of testing facilities. this will be used to examine geographic differences in reporting and how they may impact the accuracy of HIV Testing indicator i.e. in counties.

Create a comparison value to compare initial test and on known positive.

Period of test also create an insight on time where more people we people were more like to get tested.

**FEATURES**

Testing location: The location where the test was conducted could be an important factor in determining whether or not the test was included in the HIV Testing Indicator. For example, some facilities may exclude tests conducted at satellite clinics or mobile testing sites.

Age and gender of the individual being tested: The age and gender of the individual being tested could be important factors in determining whether or not the test is included in the HIV Testing Indicator. For example, some facilities may only report tests conducted on certain age ranges or genders.

Time of the test: The date and time when the test was conducted could be relevant in determining whether or not the test is included in the HIV Testing Indicator. For example, some facilities may only report tests conducted during certain time periods.

**MODEL DEVELOPMENT**

**MODEL DEVELOPMENT APPROACH**

To develop a decision tree model to improve reporting consistency of the HIV Testing Indicator across facilities, a systematic approach was be taken. The steps involved in this approach are:

Identify the problem: The first step in developing a decision tree model is to identify the problem or issue to be addressed. In this case, the issue is inconsistent reporting of the HIV Testing Indicator across facilities, with some facilities including tests from some testing locations and excluding PMTCT tests.

Gather data: The next step is to gather data related to the issue. This may include data on the testing locations, the types of tests being conducted, and the reporting practices of different facilities.

Define the variables: Once the data has been gathered, the next step is to define the variables that will be used in the decision tree model. These include the testing location, the type of test, and the facility reporting practices, period of testing.

Develop the decision tree: Using the defined variables, a decision tree can be developed. The decision tree will outline the decision-making process that facilities should follow when reporting the HIV Testing Indicator. For example, the decision tree may start with the testing location variable, with



Diagram 1 . decision tree algorithm

branches for each testing location. The branches may then split based on the type of test being conducted, and then further split based on whether the test is PMTCT-related or not. Finally, the decision tree may provide guidance on the reporting practices that should be followed for each type of test and testing location.

Test and refine the decision tree: Once the decision tree has been developed, it should be tested and refined to ensure that it is effective in improving reporting consistency. This may involve piloting the decision tree in a small number of facilities and collecting feedback from stakeholders.

Implement and monitor the decision tree: Once the decision tree has been refined, it can be implemented across all facilities. Ongoing monitoring should be conducted to ensure that the decision tree is achieving its intended outcomes and to identify areas for further improvement.

**MODEL JUSTIFICATION**

The reporting of the HIV Testing Indicator is crucial for tracking progress in the prevention and treatment of HIV, as well as for planning and allocating resources. However, inconsistent reporting practices across facilities can lead to inaccurate and unreliable data, making it difficult to assess the effectiveness of HIV testing programs and interventions. To address this issue, a decision tree model can be developed to provide guidance on the decision-making process for reporting the HIV Testing Indicator.

The decision tree model has several advantages. First, it provides a standardized and systematic approach to reporting, reducing the potential for variations in reporting practices across facilities. This standardization can improve the accuracy and reliability of data, which is essential for monitoring progress and making informed decisions. Second, the decision tree model can help to identify areas where further training or support may be needed to improve reporting practices. For example, if a facility consistently reports only HIV tests and excludes PMTCT tests, this may indicate a need for additional training or guidance on reporting requirements. Third, the decision tree model is easy to understand and can be easily communicated to healthcare providers and data managers, ensuring that reporting practices are consistent across all facilities

In addition, the decision tree model can be updated and refined over time as reporting requirements change or new data becomes available. This flexibility ensures that the decision tree model remains relevant and effective in improving reporting consistency of the HIV Testing Indicator.

**MODEL EVALUATION**

The decision tree model can be evaluated using several criteria, including accuracy, precision, recall, and F1 score. These measures are commonly used in evaluating classification models, which the decision tree model can be considered.

Accuracy measures the proportion of correctly classified instances. Precision measures the proportion of true positives among all positive predictions, while recall measures the proportion of true positives among all actual positive instances. The F1 score is the harmonic mean of precision and recall.

To evaluate the decision tree model, we could apply it to a dataset of HIV testing records from multiple facilities, and compare the resulting reported HIV Testing Indicators to the known actual indicators for each facility. We could then calculate accuracy, precision, recall, and F1 score for each facility and for the overall dataset.

The evaluation would help to determine the overall performance of the decision tree model in improving reporting consistency of the HIV Testing Indicator. If the model is effective, we would expect to see higher accuracy, precision, recall, and F1 score for facilities that follow the decision tree guidance compared to facilities that do not.

**METRICS**

Accuracy: The proportion of correctly classified instances. In the context of the decision tree model, accuracy would measure how often the model's guidance for reporting the HIV Testing Indicator is consistent with the actual reported indicator for each facility.

Precision: The proportion of true positives among all positive predictions. In the context of the decision tree model, precision would measure how often facilities that are predicted to have a positive HIV Testing Indicator actually have a positive indicator based on the actual reported data.

Recall: The proportion of true positives among all actual positive instances. In the context of the decision tree model, recall would measure how often the model correctly identifies facilities that have a positive HIV Testing Indicator based on the actual reported data.

F1 score: The harmonic means of precision and recall. In the context of the decision tree model, the F1 score would provide an overall measure of the model's effectiveness in improving reporting consistency of the HIV Testing Indicator by considering both precision and recall.

**METRICS JUSTIFICATION**

Accuracy: This metric provides an overall measure of how well the decision tree model is performing in improving reporting consistency of the HIV Testing Indicator. It is important to assess accuracy to ensure that the model is not making too many errors in its predictions.

Precision: This metric is important because it measures the proportion of true positive predictions among all positive predictions. In the context of the decision tree model, it would be important to maximize precision to ensure that facilities that are predicted to have a positive HIV Testing Indicator actually do have a positive indicator.

Recall: This metric is important because it measures the proportion of true positive predictions among all actual positive instances. In the context of the decision tree model, it would be important to maximize recall to ensure that the model is correctly identifying facilities that have a positive HIV Testing Indicator based on the actual reported data.

F1 score: This metric provides an overall measure of the decision tree model's performance by considering both precision and recall. It is important to maximize the F1 score to ensure that the model is accurately identifying facilities that have a positive HIV Testing Indicator based on the actual reported data.

**RESULTS**

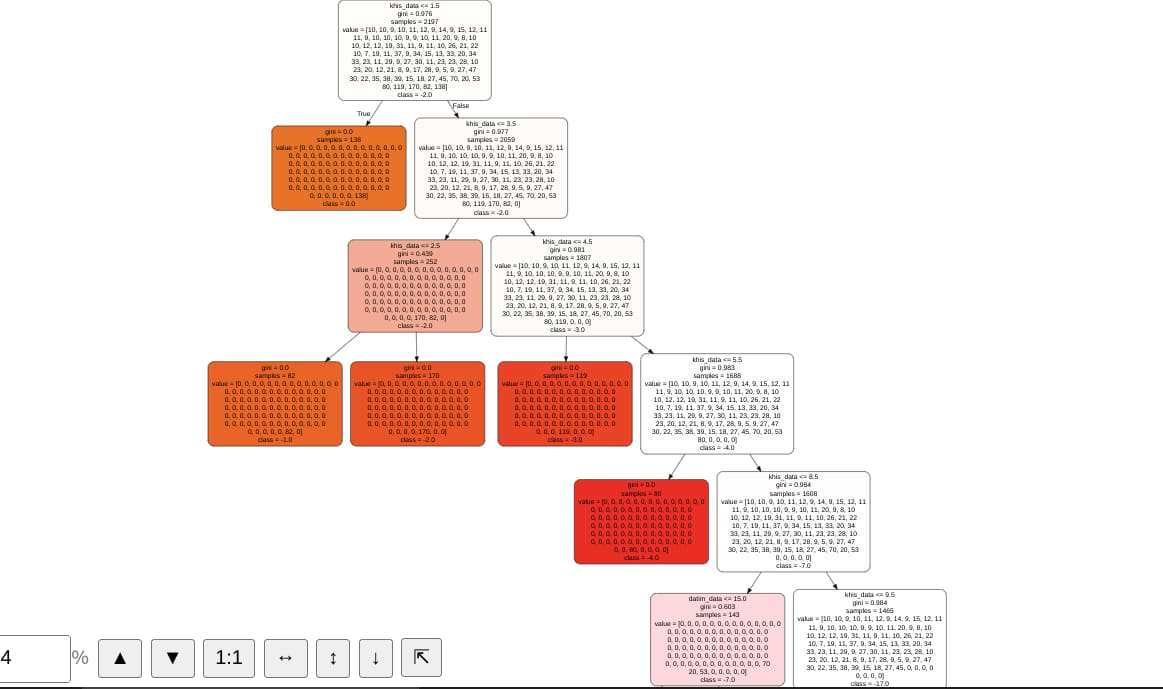
****

Diagram 2. Result of the HTI decision tree model

**CONCLUSIONS**

In this study we addressed the issue of inconsistent reporting of HIV Testing Indicator across healthcare facilities. We developed a decision tree model to predict whether or not a given HIV test should be included in the indicator, based on various features such as testing location period.

Our study provides a useful framework for addressing the issue of inconsistent reporting in healthcare.

**MODEL DEPLOYMENT**

**DEPLOYMENT PLATFORM JUSTIFICATION**

We deployed or model using flask, this is because, flask offers the following advantages;

Scalability: Deploying the model using Flask allows for easy scalability by distributing the processing load across multiple servers. Flask provides a lightweight framework that can easily handle many concurrent requests.

Ease of use: Flask is a popular and easy-to-use framework for building web applications, which makes it a good choice for deploying machine learning models. The framework provides a lot of built-in functionality, such as handling HTTP requests and responses, which reduces the amount of boilerplate code needed.

Flexibility: Flask provides flexibility in terms of deployment options. You can deploy the application to a cloud-based platform like Heroku, AWS, or GCP, or run it locally on your machine.

Interoperability: Flask can be easily integrated with other programming languages and platforms, allowing for greater interoperability with existing systems.

Reduced time to deployment: Deploying the model using Flask allows for faster development and deployment times compared to more complex deployment strategies.

**DEPLOYMENT PROCESS**

Install dependencies: we install required dependencies, (flask for the deployment, scikit-learn for the decision tree model and pandas for data handling.)

Create a Flask app: In a Python file, we create a Flask app by importing the Flask library and creating an instance of the Flask class. We also set the app's configuration and add any necessary routes.

Load the trained model: Load the decision tree model into the app by calling the load method of the joblib library. we also define any necessary preprocessing steps here.

Define the prediction route: we define a route that accepts POST requests and returns predictions based on the input data. The route should load the input data into a Pandas DataFrame, preprocess it if necessary, and then pass it to the decision tree model for prediction. The route should return the model's prediction as a JSON object.

Run the Flask app: Run the Flask app by calling the run method of the app instance.

Test the endpoint: Test the /predict endpoint by sending a POST request with input data to the running Flask app. We use t Postman, or requests library in Python to send the request.

using Flask to deploy the decision tree model allows we create a simple REST API that can be easily integrated into other applications or services. By following best practices for web development and machine learning, we create a robust and reliable solution that can help improve the reporting consistency of the HIV Testing Indicator.

**CHALLENGES**

The data cleaning process was challenging, finding the right features and variable to use link the datasets was challenging.

Our biggest challenge was choosing a machine learning model to implement on our task as most we have not user before.