DOCUMENTATION GROUP 2 HACKATHON

GROUP TWO

THALMA THANDIE

BRIAN KIMUTAI

BRANDSON KIPKIRUI

MUSHINDI RACHEL

EPHRAIM BLESSING MWEREZA

BUSINESS UNDERSTANDING

The client user/ their needs clearly described

NACSCOP is currently transitioning its data collection tools, and as a result, there will be two different sets of tools used for data collection and reporting. It is important to have an unbiased approach in selecting indicators that takes into account relevant facility characteristics. The client; the NASCOP personnel need to make a non-bias decision based on a machine learning model that advises on the indicator to use based on historical data values, facility characteristics, changes in the indicator etc. in making a choice of what indicator to use.

The client engagement process

We had online meetups with the client facilitated by the Kabarak bootcamp's patrons. The client was able to share information on the current order of operations and highlight the real-time challenges experienced with working with the current system.

The objective of the task

To develop a machine learning model that advises on the indicator to use based on historical data values, facility characteristics, changes in the indicator etc. In making a choice on what indicator to use

Data Acquisition

Source Systems

The data used was obtained from the normal running of the resident system in the organization. It was shared to us electronically in excel format.

Data Acquisition process

Raw data was obtained from the resident system and shared to us electronically in excel format

Exploratory Data Analysis

The exploratory data analysis process

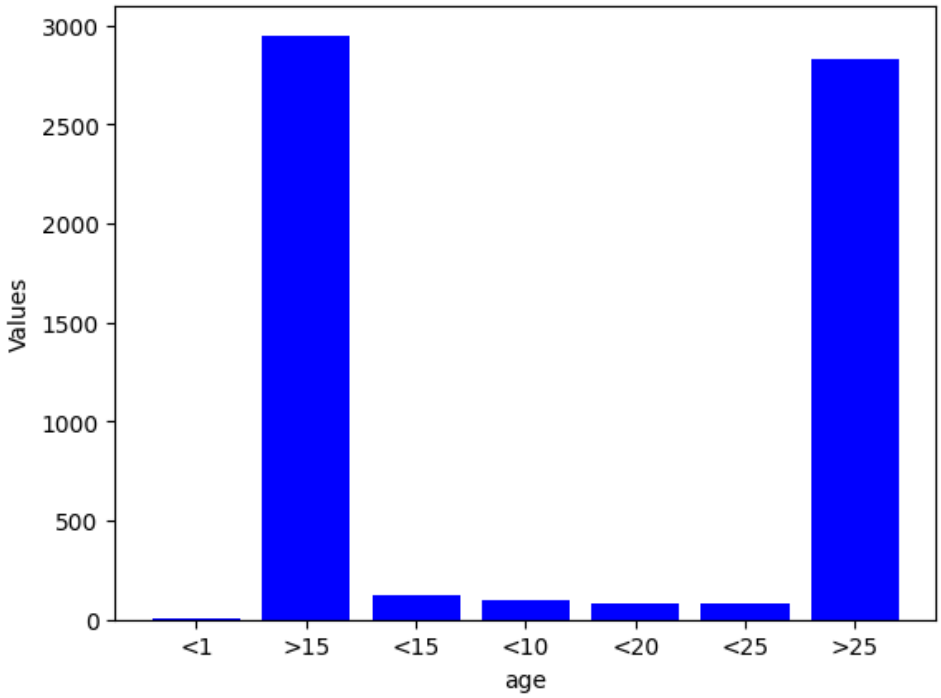
We used the python libraries tool for visualization, these were matplotlib and seaborn.

Under the libraries, we used bar plots to plot the graphs.

Different bar graphs have shown the relationships among variables in the data frame.

1. **Bar Chart of values against age.**

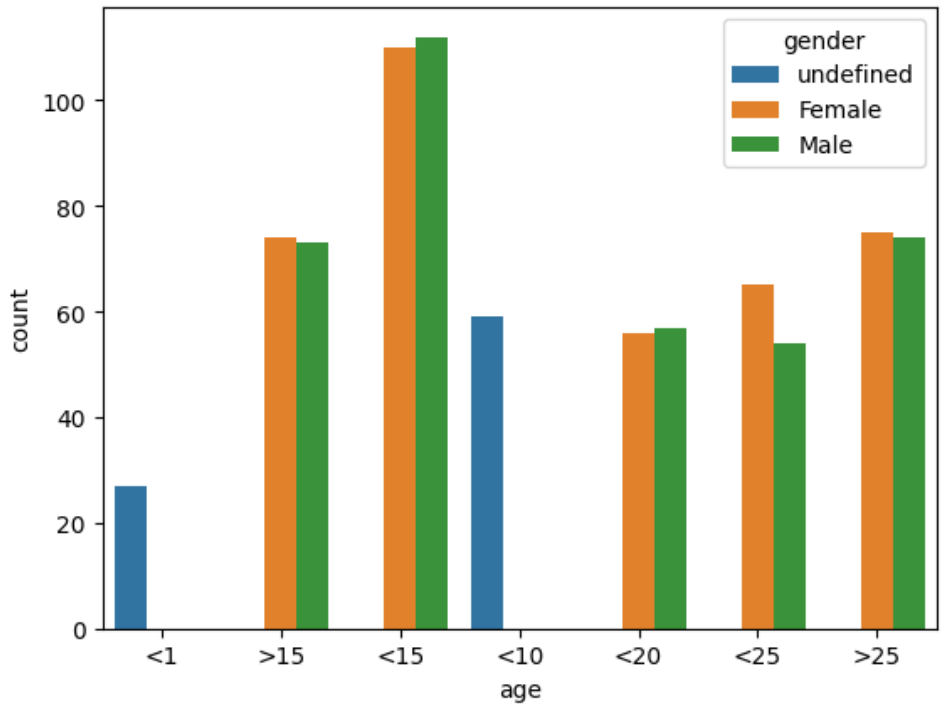
We can be able to visualize that age >1 have a smaller number of people(values) receiving ARVs but there are high number of people receiving ARVs at age >15 and <25.



2. **Bar Chart Showing a relationship between age and gender.**

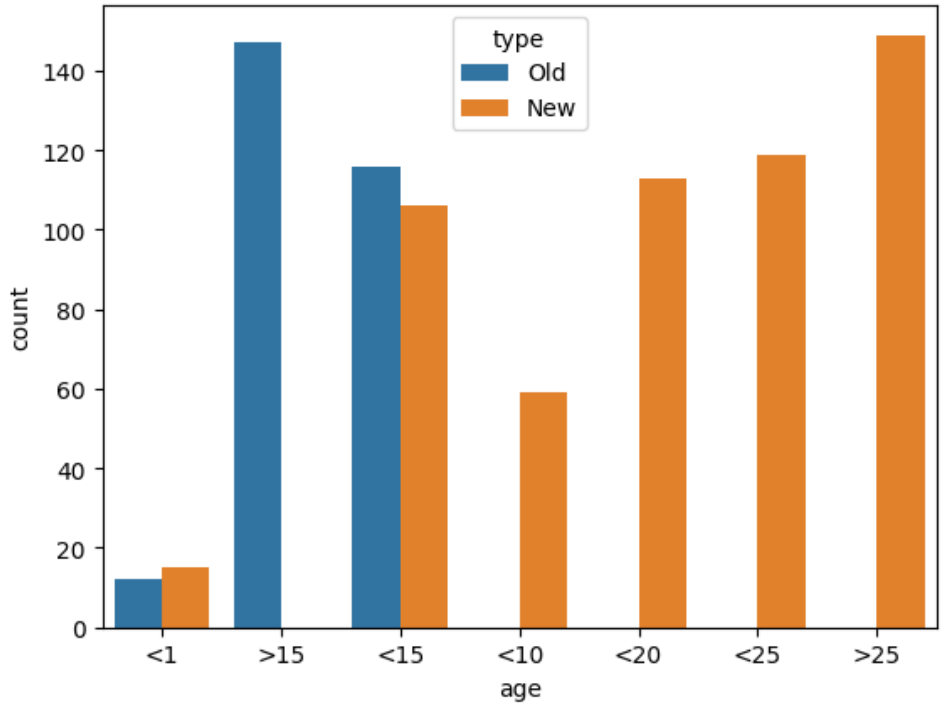
The bar graph shows that male who are less than 15 years receiving ARVs are slightly fewer compared to females.

Males who are greater 15 years receive more ARVs compared to female.



3. Bar Chart showing relationship between age and the tool used.

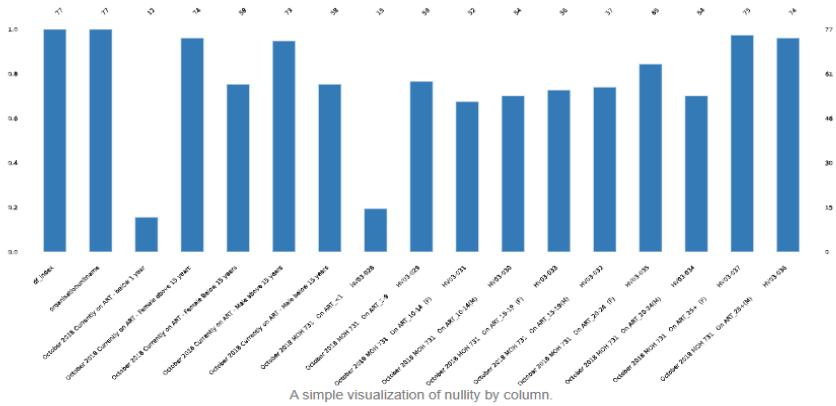
This plot shows that the new tool is specific to the age and is majorly used by people with 25 years and above, while the old tool does not break down the ages to further divisions and is mostly used by people above the age of 15.



Data Cleaning

Data cleaning process

The initial data frame contained rows that had totally null values distributed across the dataset. By use of the pandas dropna() function were able to delete the null rows in the dataset which were 771 rows dropped. As shown:



During the transition from the old tool to the new tool, there were facilities that opted to use either the old or the new tool or both tools. There were 771 facilities that used neither of the tools which translated to the totally null rows. The null rows needed to be dropped for they were insignificant when trying to predict trend of the data entry per tool with different ages.

We identified the age, type of tool as relevant features that could be useful in the classification model.

Data cleaning outcome

The data cleaning process resulted in reduced dataset size due to the deletion of rows with totally null values, the new data set size was 77 rows with 16 columns. This should result to improved model performance in that the classification model can be developed with a more accurate and reliable data

Removing the first row and setting it as the new column names enabled us to have new column names on the data frames. Dropping the null rows resulted in a more consistent dataset. From pivoting we had a transformed data frame with three columns time_and _type, variable and value making the dataset easier for data analysis

Feature Engineering

The feature engineering process used

The initial dataset did not have preset column names which would make it complicated to analyze the necessary features for the solution model.

We set the first row as the new column names for the dataset as the first step in feature engineering.

We used the pandas melt function to unpivot the data from wide format to a long format, increasing the number of rows while reducing the numbers of columns. As a result, we had three new columns; time_and_ type, variable and value. We carried out pivoting to reshape the data for easier analysis.

We created new features, new columns - age, gender, and type as part of feature engineering. These columns are likely to be important features in any machine learning model that we build, as they provide additional information about each observation that can be used to make predictions. These columns are highly correlated, and observations made of these features could be used for prediction.

We did one hot encoding of the variable column to transform the categorical data to numerical data, a format that can be used by the model.

Model Development

The model development approach

We trained the models from the training set.

Here are the steps used to train each model:

a. **Decision Trees:** we trained the decision tree model on the training data using an appropriate criterion and hyperparameter tuning. We evaluated the model's performance on the testing set using the accuracy metric.

   The model had an accuracy of: 0.15291008279042861

   Mean squared error of: 93851.79091627456

   Mean absolute error: 105.5769257150567

b. **Random forest regression**: we trained the random forest model on the training data using appropriate number of trees and hyperparameter tuning. We evaluated the model's performance on the testing set using the accuracy metric.

   Mean Absolute Error: 106.1593037836019
   Mean Squared Error: 93306.85784402715
   Accuracy score: 0.15782855378121263

c. **Linear regression:**

Mean Absolute Error: 132.2239048545531
Mean Squared Error: 106855.64421092803
Accuracy score: 0.03553956803274749

d. **Gradient boosting**

Mean Absolute Error: 105.5905908325453
Mean Squared Error: 93879.27146382436
Accuracy score: 0.1526620481549511

The justification of the chosen model

The algorithm selected for use is the decision trees because it gives the highest accuracy among other algorithms used.

This is also because decisions to be reached majorly followed the if-then format, making this model to be relevant.

Model evaluation

Metrics used

Justification for the choice of metric

Given the fact that the model developed is to be used to advise on the indicator to be used based on historical data and facility characteristics. The probability that the model would advise correctly with limited errors should be high to limit chances on making decisions on particular indicators.

Results presentation and Discussion

Model Deployment

The choice of model deployment platform was a web application: stream lit. This is because it's open-source and easier to implement and would take limited time to implement given the short duration of the entire project.

The process of deployment

Install Streamlit if you have not yet installed it on your computer. We used ‘pip’, python package manager, by running the command ‘pip install streamlit’ in the terminal.

We then wrote the code for the streamlit app. This was done typically by importing necessary libraries, defining functions for data processing and then creating a user interface using streamlit’s widgets.

We then tested our app code locally on our computer by running the ‘streamlit run app.py’ (where ‘app.py’ is the name of the app file). This launches a web server which we accessed in our web browser

Challenges

Many rows with missing values