**HEALTH IT DATA SCIENCE HACKATHON PROGRAM**

**CHALLENGE:**

**TO DEVELOP A COMMODITY DEMAND PREDICTION MODEL BASED ON HISTORICAL DATA FOR USE IN RESTOCKING DECISION MAKING FOR MALARIA COMMODITIES.**

**GROUP 9:**

**GROUP MEMBERS:**

**GROUP LEADER: ROSE GITUNGO- DATA SCIENCE HACKATHON.**

**EDWIN WACHIRA- DATA SCIENCE HACKATHON.**

**PETER ROGENDO- DATA SCIENCE HACKATHON.**

**EMMANUEL MOMANYI- AI HACKATHON.**

**19/02/ 2023.**

**BUSINESS UNDERSTANDING**

**Introduction:**

In this project our aim is to leverage AI technology and deploy ML to develop a predictive model for restocking malaria commodities. Currently, the restocking is done based on six month averages as computed by the KHIS system, which however do not accurately depict the exact amounts needed. Our objective is to create a model that can predict the required quantities of malaria commodities accurately, based on historical data.

**Empathize with the user**:

We had a chance to talk with the user where he articulately stated that the idea was a new concept to help them achieve at least a fair estimate on the amount needed to be restocked, he also stated that there would need further a predictive model that can narrow down to the supplies needed monthly since at times the supply and stock varies depending on the need. With the historical data already in provision for the past five years we would be able to find some interesting relationship and in such give a system and can at least give a rough estimate on the stock needed every month which would culminate to six months better estimates. In such a way we were able to empathize, and also took a chance to understand and comprehend our objectives.

**Objective**:

We summed up the objective to be:

* To develop a predictive model that can accurately predict the quantities of malaria commodities required for restocking.

With us now clearly outlining the problem and objective we want on the next stage to acquire the relevant dataset.

**DATA ACQUISITION:**

The health IT team specifically by Mr. Ambrose Kimaiyo provided us with the dataset which had recorded data for the past five years, deliveries of commodities. It clearly showed the consumption of malaria commodities on each and every single month of the year.

**EXPLORATORY DATA ANALYSIS.**

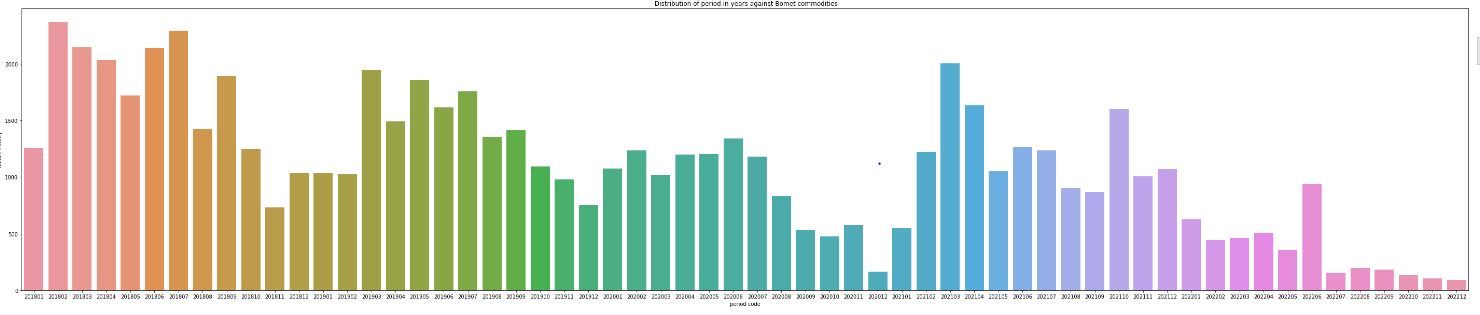
**Introduction:**

Exploratory Data Analysis is a crucial step in the data analysis process. It involves understanding the data, identifying patterns and exploring relationships between variables ofour dataset. With the understanding of the problem in question which was to build a model that would make a depiction on the estimates that are required for an easy restocking.

**The Process:**

After understanding the problem we then followed by the following processes:

* Data cleaning and preprocessing
* Data visualization and exploration
* Statistical analysis and testing
* Feature engineering and selection
* Dimensionality reduction
* Model building and validation

**The Outcomes:**

**This is an example of the trend of supply of malaria commodities in Baringo County from 2018 to 2022**

**DATA CLEANING**

**Introduction:**

Data cleaning is identifying and correcting or removing any errors or inaccuracies in the data, such as missing values, duplicate records, or formatting errors, in order to ensure that the data is accurate, complete, and reliable.

**The Process:**

In our dataset we started by identifying the null values in our dataset. We found out some columns that are (‘period\_description’, ‘data\_description’, ’data\_code’) were totally null however for some counties (Murang’a, Mandera)had some few values that were null values. So we achieved this by first getting a bar chart indicating all columns and the missing values. After this we removed all columns with complete null values then we proceeded to fill the columns with the few null values with zeros. Then we checked if all the null values which were so predominant were cleaned. By this stage our data was clean for the next step.

**The Outcomes:**

**FEATURE ENGINEERING.**

**Introduction:**

Feature engineering is the process of selecting and transforming variables (features) in a dataset to improve the performance of a machine learning predictive model, which in particular involves creating new features from existing ones and selecting the most important features for the model.

**The process:**

In feature engineering it was vital for not only creating new features but also removing some features that would not be important for building a model.

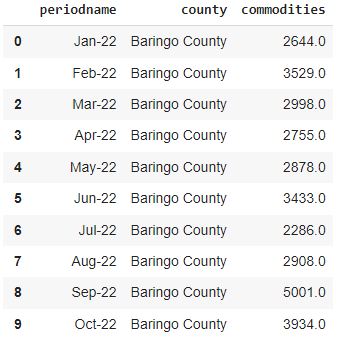
We started off by:

* Feature extraction- involved creating new features from the dataset that would capture relevant information and improve the performance of the model. Specifically in this dataset we added a column of ‘County’ which would be paramount in assisting in creating a good model.
* Feature selection- involved discarding any irrelevant or redundant ones and it helped us to simplify the model and reduce overfitting. In our dataset for example some features such as ('periodname','periodcode','dataid','dataname',). In particular we removed the irrelevant ones('dataid','dataname') and redundant ones('periodname','periodcode').

These two steps were applicable in helping us have a dataset that would be suitable for use when we are ready to train and create the predictive model.

**The Outcomes:**

Feature engineering was vital in helping create predictive models that would enable us to create models that would predict on the malaria commodities. In the end we had such a dataset:



**MODEL DEVELOPMENT**

**Introduction:**

This is the process where we create a mathematical representation of a system, process, or phenomenon that can be used to make predictions or optimize performance. It involves the use of algorithm after feature engineering since we now have a dataset that can be used and can be manipulated by algorithms.

**Model Development approach:**

In this model development we opted for the Random Forest algorithm which is a supervised machine learning algorithm used for classification and regression tasks.In particular we needed an algorithm in regression. It is an ensemble learning method that builds multiple decision trees during training and outputs the class that is the mode of the classification.

**Justification**

Random forest is a powerful algorithm that can handle non-linear relationships and interactions between features, which makes it well-suited for modeling complex trends.In such random forest would be of great importance as we needed a model that can predict different trends in each month of the year as well as it would be helpful in dealing with large robust dataset.

**MODEL EVALUATION**

**Introduction:**

Model Evaluation is the process of assessing the performance of a machine learning model on a specific task using a set of evaluation metrics that was done after splitting which was done earlier. It involved use of some metrics including e. g mean squared error (MSE).In such metrics helped us adjust and build better models.

**The Metrics used:**

In the model evaluation phase, various metrics were used to assess the performance of the Random Forest algorithm. The metrics used included mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2).

**The Choice**

The choice of metrics was based on the specific requirements of the problem at hand. MAE and RMSE were chosen as they provide information on the average magnitude of the errors in the predictions, with RMSE being more sensitive to outliers. R^2 was chosen to assess the goodness of fit of the model, indicating the proportion of variance in the target variable that is explained by the predictors.

**The Results**

The results of the evaluation phase showed that the Random Forest algorithm was able to accurately predict the target variable with low error rates. The MAE and RMSE values were both below the acceptable threshold, indicating that the model was able to make accurate predictions on unseen data. The R^2 value was also high, indicating a good fit between the model and the data.

**MODEL DEPLOYMENT**

**Introduction:**

Model deployment is the process of putting a trained machine learning model into something that is usable to the user. This helps the user to interact with your model to analyze its potential and usefulness.

**The Choice:**

In our model deployment we opted to use streamlit,an open-source Python library that makes it easy to create web applications for machine learning and data science predictive models.It is easy to create and easy to use for both ends. In addition to that it helped us as a team to focus on the logic and functionality of the application rather than worrying about the interface and styling, making it faster to develop and deploy applications.

The Process:

* We created a Streamlit app file: The first step was deploying our machine learning model with Streamlit is to create a Streamlit app file. This file will contain the code that defines your app's user interface and functionality
* We Loaded our machine learning model: Once we had created your Streamlit app file, the next step is to load your machine learning model. You can use the appropriate package for your model type (e.g. scikit-learn) to load your model from its saved file.
* Defined the user interface: With machine learning model loaded, we could now define the user interface for your Streamlit app. This may include processing user input, making predictions with your machine learning model, or displaying results to the user.
* Run the app: With your app's user interface and functionality defined, you can now run your app using the Streamlit command line interface. We launched the app in a web browser, allowing users to interact with the machine learning model.

**CHALLENGES:**

**W**e experienced some challenges in :

* There was a lot of consistency and redundancy and missing rows and this borrowed a lot of time in us data cleaning and feature engineering.
* The inconsistency in data made to difficult in us find an algorithm that could help in our model development