**A MACHINE LEARNING MODEL FOR PREDICTING CROP YIELDS IN SUB-SAHARAN AFRICA**

1. Emmanuel Chesire
2. Stanley Njoroge
3. Dedan Okware
4. Lennox Mwita

**INTRODUCTION**

Crop production in Sub-Saharan Africa is a critical component of the region's economy, providing food and income for millions of people. However, the region's agricultural productivity remains low, with yields far below those achieved in other regions of the world. This is due to a range of factors, including poor soil quality, inadequate access to inputs such as fertilizers and improved seeds, and the effects of climate change. Efforts to improve crop production in the region have focused on increasing access to inputs, improving soil fertility, and developing crop varieties that are better adapted to local conditions. The development and adoption of new technologies, such as precision agriculture and digital farming, are also offering new opportunities for increasing crop yields and improving livelihoods in the region.

## **BUSINESS UNDERSTANDING/**

### PROBLEM

The low productivity of crop production in Sub-Saharan Africa remains a persistent challenge, with yields far below those achieved in other regions of the world. This poses a significant threat to food security and the livelihoods of millions of farmers in the region. Factors such as poor soil quality, inadequate access to inputs, poor timely planting, and the effects of climate change are contributing to this challenge.

THE NEED TO DEVELOP A CROP YIELDS PREDICTION MODEL

The aim of developing a machine learning model to predict crop yields in Sub-Saharan Africa is to provide farmers with valuable insights into the upcoming growing season. By analyzing data on weather patterns, soil quality, irrigation systems, and farming techniques, the model can provide accurate predictions of crop yields. Farmers can then use this information to make informed decisions about when to plant and harvest crops, reducing waste and increasing income. This will improve food security and help to address the persistent challenge of low crop productivity in the region. The development and implementation of this technology will offer new opportunities for agricultural innovation and improved livelihoods for farmers in Sub-Saharan Africa.

Overall, the development and implementation of a machine learning model to predict crop yields in Sub-Saharan Africa is a critical step towards improving the lives of farmers and addressing the challenges of food security in the region.

CLIENT ENGAGEMENT PROCESS

The user engagement process is defined by the following steps which will ensure a user-centered crop yield prediction model is developed.

1. Defining target audience – This includes identifying the specific group of people i.e farmers who will be using the crop yield prediction model.
2. Understanding user needs – After defining the target audience, we need to understand their needs and preferences.
3. Provide excellent services – The crop yield prediction model should be able to perform its task correctly.
4. Continuously improve the crop yield prediction model – improvements should be done on the model in order to meet changing user needs.
5. Measure user engagement – Measuring the engagement helps to determine whether the

system was effective or not.

### OBJECTIVES

1. To develop a crop yield prediction model for farmers in the Sub-African Saharah region.
2. To deploy a crop prediction model into a web interface to enhance user experience.
3. To develop a crop yield prediction model using machine learning which can be able to predict crop yields quickly and efficiently.

## **DATA ACQUISITION**

## SOURCES OF DATA

Having understood the business logic we used several websites to obtain our data and datasets. They included:

1. <https://climateknowledgeportal.worldbank.org/download-data> -in this particular site, we were able to get information about rainfalls and temperature. This is key information when matters of crop production are raised.
2. <https://www.fao.org/faostat/en/#data/QC> - in this site, we were able to get information about crops, pesticides, and crop yields.

The data corrected varied depending on different countries within the Africa Sub-Saharan.

### DATA ACQUISITION PROCESS

After identifying our data sources, we were able to come up with a process which would enable us to get as much information from the data as possible. We used the Extraction Transformation and Loading (ETL) tool, this is a process used in data acquisition to collect and move data from different sources, transform the data to make it compatible with the destination system, and load the transformed data into the target system. The first step was to go through the different websites shown above and understand the connection between them.We were able to obtain and separate useful information which was necessary for training our model from other information. We then recorded the data in four csv files which was an appropriate data format for our model. The files included:

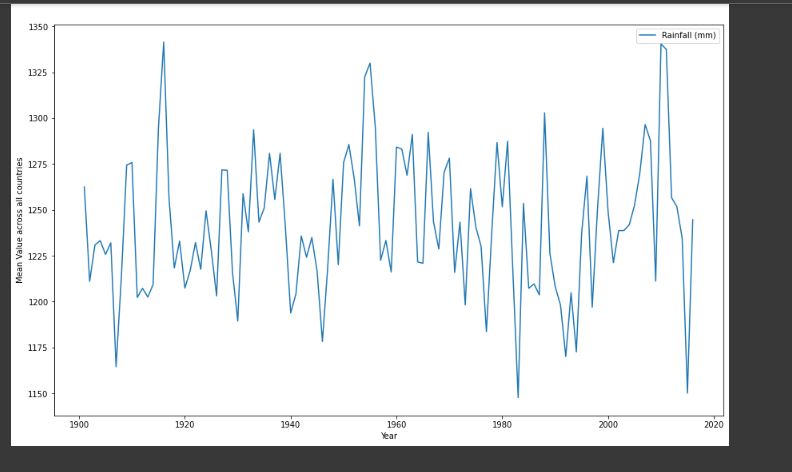
1. A Rainfall file
2. A yields file
3. A crops file
4. A Pesticides file

The final step was loading the data into our system for use in training the model.

## **EXPLORATORY DATA ANALYSIS**

EDA is the process of analyzing and summarizing the main characteristics of a dataset. Its purpose is to explore the topic patterns of the dataset.

The EDA techniques used included exploratory visualizations, to analyze the trends of the weather patterns , temperatures, utilized pesticides, and the corresponding yields obtained over the years.

****

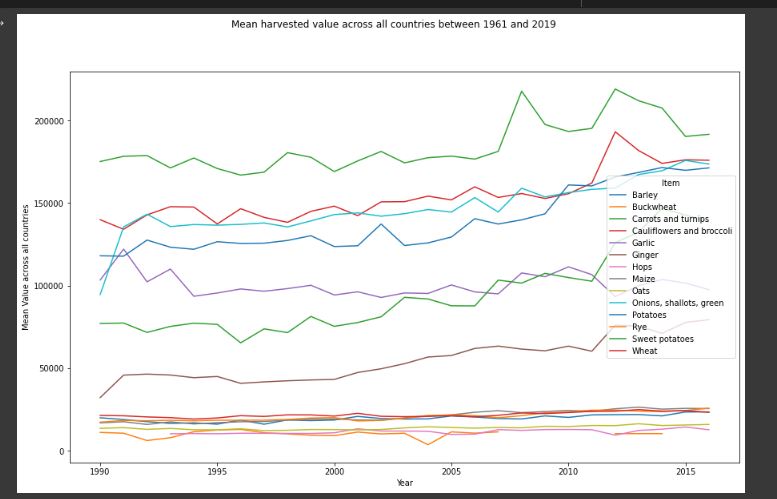
### EXPLORATORY DATA ANALYSIS PROCESS

### Data cleaning and preprocessing – We cleaned the data to remove empty or incomplete data from the dataset and some of the inconsistencies.

### Identifying patterns – This involves identifying related patterns and analyzing some of the relationship between the different files of data we had.

### Feature selection – This involved selecting important relationships that was used to train the crop prediction model.

### Visualizations – This includes use of histograms and box plots to examine the distribution of rainfall, pesticides, temperature and the corresponding crop yields.

****

## **DATA CLEANING**

We categorized the data into four different csv files and performed cleaning operations on each file separately.

### DATA CLEANING PROCESS

* We first went through the different datasets from the websites and identified the relationship between them
* Dropped empty lines
* Dropped null values
* Dropped data above the required range
* Categorized the data into different into four different csv files
* Identified duplicated patterns and removed unnecessary characters and formatted the data

### DATA CLEANING OUTCOMES

1. We were able to acquire a dataset that contained a complete relationship between rainfall, temperatures, pesticides and crop yields.
2. Improved data quality: we removed errors, inconsistencies, and irrelevant data which made it more reliable for the next step.

## **FEATURE ENGINEERING**

Feature engineering is the process of selecting and transforming raw data into features that can be used to train a machine learning model.

### FEATURE ENGINEERING PROCESS

After doing the data cleaning process and data transformation, we were able to remain with a better dataset as our training set.

These steps were followed for the feature engineering process:

### FEATURES USED

Year

This is the year in which the data used was recorded. This was an helpful independent variable in preicting the yiels

Rainfall

Amount of rainfall in a particular year. Rain is one of the influencing factors of crop production, increase in rain means high yields

Pestisides

Pesticides used per country arranged according to year.

Country

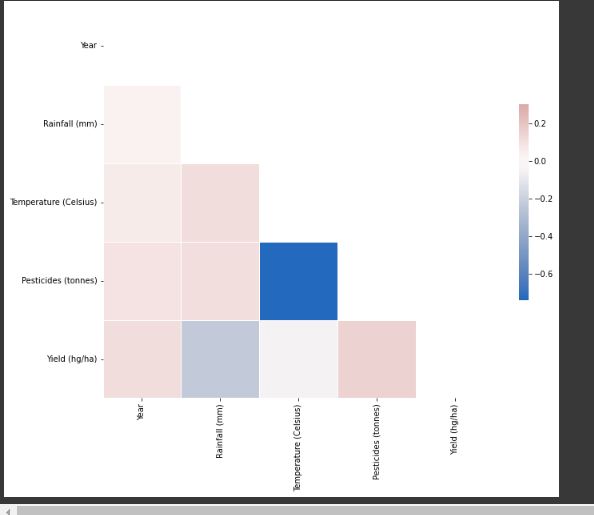
This is the subsahran african countries where data was collected from.

Yield

Dependent variable being predicted in the dataset

Items

Types of crops planted in that particular area and their yields recorded.



## **MODEL DEVELOPMENT**

The model development approach chosen is an unsupervised learning approach. It involves training a model on an unlabeled dataset, where the input data is not accompanied by any target variable. The goal of the model is to discover patterns in the data, without any specific guidance as to what to look for.

### JUSTIFICATION FOR MODEL USED

The model was used because it is useful when there are no labels or targets for the data, or when data exploration is required. In our model it is supposed to traverse through the given dataset and come up with an accurate prediction of the crop yield. It is supposed to learn patterns so as to be able to perform adequately.

## **MODEL EVALUATION**

Machine learning models are evaluated by use of metrics. A metric is a measure of something that can be used to track and compare performance. Metrics can be used to measure and compare performance and progress. They provide an objective way to measure and compare progress in order to make informed decisions and identify areas for improvement.

### METRICS USED

The following types of metrics were applied:

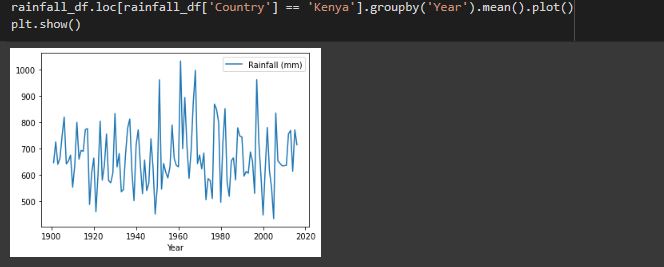
1. Precision

Precision metric is a measure of accuracy and consistency in a model. This metric was used to assess how accurately our model was able to predict the true value of a target variable. Precision measured how many of the predictions made by the model were correct. The higher the precision, the more accurate the predictions of the model are.

2. F1-score

The F1-score is a metric that combines precision and recall into a single score. This metric was used to assess the overall performance of a model. The F1-score takes into account both precision and recall to give an overall measure of how well a model is performing. The higher the F1-score, the better the performance of the model.

RESULTS FROM DIFFERENT METRICS

****

### JUSTIFICATION FOR METRIC USED

The precision, recall, and F1-score metrics are useful for assessing the performance of a model. Precision measures how accurately the model is able to predict the true value of the target variable, while recall measures how many of the true values the model is able to identify. The F1-score combines these two metrics into a single score, allowing for a more comprehensive assessment of the model's performance. These metrics were useful for evaluating the model and helped to identify areas for improvement.

## **MODEL DEPLOYMENT**

Model deployment in machine learning refers to the process of integrating a trained machine learning model into a production environment, where it can be used to make predictions or perform other tasks in real-time. Model deployment is the final step in the machine learning pipeline, and it involves making the model available to end-users or other systems that can make use of its outputs.

### DEPLOYMENT METHOD USED

We used Streamlit for our model deployment because it is an open-source Python library that allows data scientists and developers to quickly create and deploy interactive web applications for machine learning and data science projects.

### PROCESS OF MODEL DEPLOYMENT

After training and saving our model, we used the following process to deploy our model:

1. Installed Streamlit: we Installed Streamlit on our local machine using pip.
2. Defined our Streamlit app: we created a Python script that defined our Streamlit app. This script imported our model, loaded any required data, defined the UI elements, and defined the functionality of the app.
3. Tested the app locally: we tested our app locally to make sure it was working correctly. We did this by running the Streamlit app using the command “streamlit run app.py” in our terminal.
4. Deployed the app: After testing our app locally and we were satisfied with the results, we deployed it to a web server.

## **CHALLENGES**

* The main challenge faced is cleaning the data. The data was hug.
* Inadequate time: There was limited time for us to complete the project.
* A Lot of time was invested in searching for the appropriate datasets.