
# MAKERERE UNIVERSITY

## COLLEGE OF COMPUTING AND INFORMATION SCIENCES

### SCHOOL OF COMPUTING AND INFORMATICS TECHNOLOGY

### DEPARTMENT OF COMPUTER SCIENCE

## END OF SEMESTER PROJECT- MCS7103 MACHINE LEARNING

## TITLE: CROP YIELD ANALYSIS AND PREDICTION

**Prepared by:**

**GLADYS OBOL ACAA**

**2025/HD05/26332U**

**2500726332**

# 1.0 Introduction

Agriculture remains the backbone of many economies [1]. In Uganda the sector accounts for the largest share of employment (47%) and close to 80% of all households in Uganda are involved in agriculture [2].

Farm productivity is often constrained by unpredictable weather conditions, pest infestations, soil degradation, inefficient use of resources, and lower crop yields which directly affect food security and farmer livelihoods. In 2023, maize production in Uganda is projected to reach 3,192,047 metric tons (MT), which corresponds to a 8% increase over 2022 production levels.

Traditional methods of predicting crop yields like maize rely on historical averages and subjective judgments, which are often inaccurate and fail to capture the complexity of modern agricultural ecosystems.

Using Machine Learning models (ML),farmers and agricultural planners can make informed decisions about crop selection, input optimization, and resource allocation ultimately improving yield and sustainability.

# 1.1 Problem Statement

Farmers lack accurate and timely predictions of crop yields due to limited access to intelligent analytical tools. Existing methods are unable to integrate large datasets involving climate, soil quality, and management practices, resulting in yield uncertainty and economic loss.

Therefore, there is a need for a machine learning based system that can analyze farm data and predict yield outcomes to guide farmers in making evidence-based decisions.

# 1.2 Objectives

## 1.2.1 Main Objective

To design and develop a Machine Learning–based model for analyzing farm data and predicting crop yields e.g maize with high accuracy.

## 1.2.2 Specific Objectives

1. To collect and preprocess data on soil, weather, and crop characteristics.

2. To train and compare multiple machine learning algorithms for yield prediction (e.g., Logistic Regression, Random Forest,).
3. To visualize the relationship between different farm parameters and yield outcomes.

# 2. Literature Review

Recent advances in machine learning (ML) have significantly improved the predictive accuracy of crop yields, a critical factor for agricultural planning and food security. A variety of agricultural datasets encompassing genomic, phenomic, climatic, and remote sensing data have been utilized to develop ML models for yield prediction.

In Zimbabwe, an applied study combined historical climate data and satellite imagery to train ML models including Random Forest (RF), Support Vector Machines (SVM), and Neural Networks. Results showed RF achieving the highest accuracy (87.5%) in maize yield prediction.[4]

Hybrid approaches combining crop growth simulation models with ML algorithms such as LightGBM, XGBoost, and RF have proven effective in the US Corn Belt. These models utilize crop modeling features alongside weather and soil data to improve prediction accuracy by 7-20%, emphasizing the need for multidimensional datasets that reflect environmental and phenological factors.[9]

To fill these gaps, the research objective is to develop a machine-learning based approach to estimate and predict crop yield at a fine scale but over a large area of interest.

To do this, supervised learning algorithms are deployed to calibrate and validate model performance.

Specifically, this report addresses the following questions.

- How does model performance vary across different regression and machine-learning approaches?
- Which period of the time series earth observation data are most important to estimate crop yield?
- How useful are ancillary spatial datasets, such as soil properties, rainfall, and temperature variables?
- What is the prediction accuracy?

# 3. Methodology

## Data

Data was collected from Kaggle Agricultural Datasets.

The datasets used in this study include rainfall data, temperature data, crop yield, soil pH and soil nutrients data. It provides insights into various aspects of agriculture, specifically focusing on farm-level details, crop types, soil pH, and associated metrics such as yield, temperature,humidity,and rainfall.

## Climate data

Estimates of weather parameters proved to be important predictors of crop yield. The variables include temperature and humidity.

## Soil properties

Crop yield may be evidently affected by soil properties such as fertility level and water holding capacity.

Soil pH and soil nutrients properties were selected in the analysis.

## Tools used

- Python, jupyter Notebook for programming
- Scikit Learn,Random Forest, Pandas, Numpy, Matplotlib, Plotly, seaborn for data handling.

## Model Development

I tested a variety of prediction approaches, including regression-based predictions and alternative machine learning algorithms, i.e. support vector machine (SVM), regression, and random forests.

- Regression-based methods can capture the spatial variation of the yield and derive the general relationship between yield and biophysical variable measurements.
- Random Forest – The ensemble nature of Random Forest, coupled with techniques like bagging and feature randomness, makes it robust to overfitting, especially when dealing with noisy or high-dimensional datasets.
- Support Vector Machine – can be used for both linear and non-linear classification tasks, making them a flexible choice for different types of data.

# 4. Results

## EXPLORATORY DATA ANALYSIS

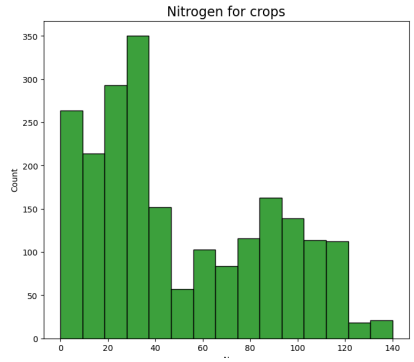
For:

- The soil nutrient properties of Nitrogen, Phosphorus and Potassium.
- The soil pH
- Climate data of temperature, rainfall and humidity.

### 1]Nitrogen

Nitrogen is an essential nutrient for plant growth, development and reproduction.

```
plt.figure(figsize=(8,7))
sns.histplot(x='N',data=crop_df,color='g');
plt.title("Nitrogen for crops",{'fontsize':16});
```

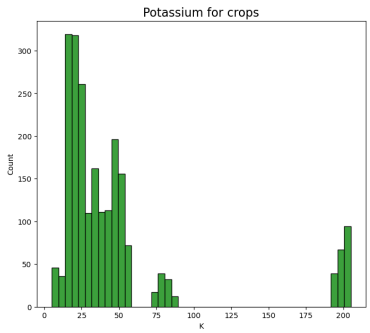



### 2]Potassium

Importance of Potassium in Plants:

- The rate of respiration by plants is largely the determining factor for proper uptake and transport of potassium by plants.
- Potassium also facilitates protein and starch synthesis in plants.
- It activates enzymes responsible for specific functions.

```
plt.figure(figsize=(8,7))
sns.histplot(x='K',data=crop_df,color='g');
plt.title("Potassium for crops",{'fontsize':16});
```

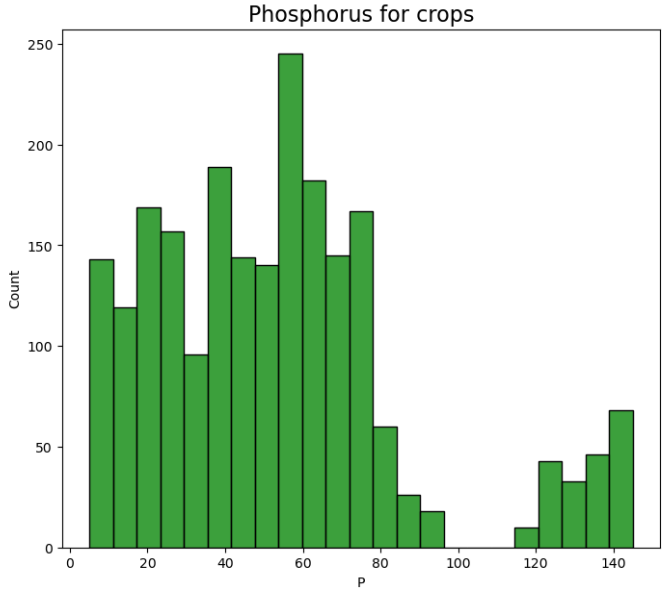


## 3]Phosphorus

Importance of Phosphorus to Plants :

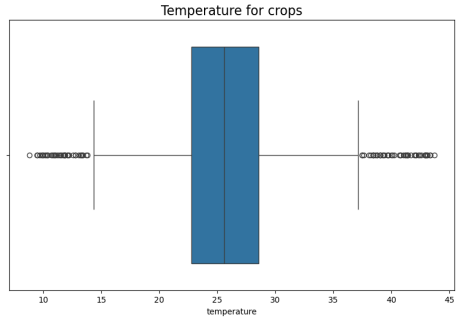
- Phosphorus is important for cell division and development of new tissues.
- Adding phosphorus to plants helps for root growth.
- It is also recommended for early growth of plants.

```
plt.figure(figsize=(8,7))
sns.histplot(x='P',data=crop_df,color='g');
plt.title("Phosphorus for crops",{'fontsize':16});
```



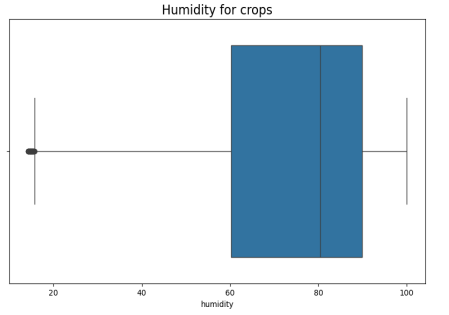

## 4]Temperature

```
plt.figure(figsize=(10,6))
sns.boxplot(x=crop_df.temperature);
plt.title("Temperature for crops",{'fontsize':16});
```




## 5]Humidity

```
plt.figure(figsize=(10,6))
sns.boxplot(x=crop_df.humidity);
plt.title("Humidity for crops",{'fontsize':16});
```

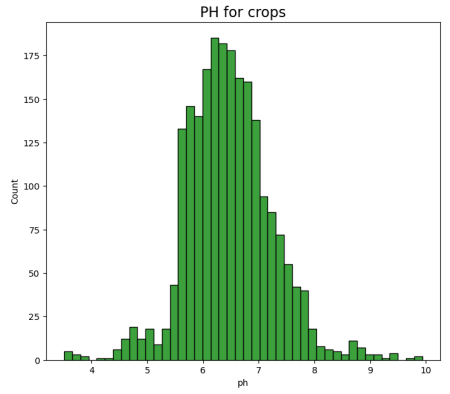



## 6] PH

pH stands for ‘potential of hydrogen’ and refers to the amount of hydrogen found in the soil.

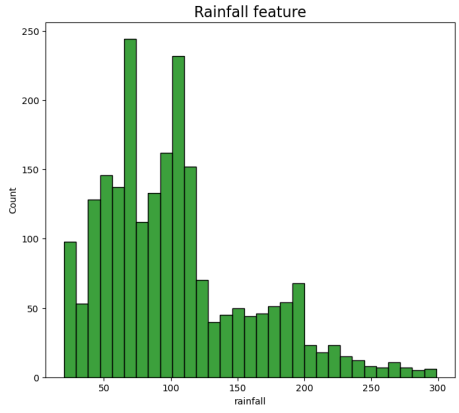
pH can affect a plant's ability to absorb vital nutrients from the soil. If pH is too acidic or alkaline, this can stunt or retard root growth and consequently, restrict water and nutrient uptake.

```
plt.figure(figsize=(8,7))
sns.histplot(x='ph',data=crop_df,color='g');
plt.title("PH for crops",{'fontsize':16});
```




## 7]Rainfall

```
plt.figure(figsize=(8,7))
sns.histplot(x='rainfall',data=crop_df,color='g');
plt.title("Rainfall feature",{'fontsize':16});
```




# Training the data for modeling

**DATA PREPROCESSING**

```
X = crop_df.drop(['label','no_label'],axis=1)
y = crop_df.no_label
```

```
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.2)
```

```
X_train.head()  #feature scaling
```

| | N | P | K | temperature | humidity | ph | rainfall |
|---|---|---|---|---|---|---|---|
| **1503** | 8 | 120 | 201 | 21.186674 | 91.134357 | 6.321152 | 122.233323 |
| **838** | 31 | 58 | 15 | 28.318869 | 60.194614 | 6.167855 | 45.365213 |
| **1425** | 117 | 25 | 53 | 29.118585 | 92.125430 | 6.413927 | 24.520202 |
| **101** | 61 | 44 | 17 | 26.100184 | 71.574769 | 6.931757 | 102.266244 |
| **1188** | 22 | 18 | 31 | 30.764552 | 47.937915 | 5.956027 | 90.385035 |

```
scalar = StandardScaler()
X_train = scalar.fit_transform(X_train)
X_test = scalar.transform(X_test)
```

**BUILDING MODEL**

```
models = {
    LogisticRegression(max_iter=500):'Logistic Regression',
    RandomForestClassifier():'Random Forest',
    SVC():'Support Vector Machine'
}
for m in models.keys():
    m.fit(X_train,y_train)
for model,name in models.items():
     print(f"Accuracy Score for {name} is : ",model.score(X_test,y_test)*100,"%")
```

```
Accuracy Score for Logistic Regression is :  97.95454545454545 %
Accuracy Score for Random Forest is :  99.54545454545455 %
Accuracy Score for Support Vector Machine is :  97.95454545454545 %
```

```
rf = RandomForestClassifier()
rf.fit(X_train,y_train)
```

**Best Model**: Random Forest with `99.54545454545455`% accuracy

**Challenges**
The model's performance could be influenced by overfitting to the small test dataset, the potential overfitting could limit the model's generalizability to unseen data.

# 5. Conclusion

The literature shows a clear trend: multi-source datasets (weather + soil + field labels) yield the best ML performance for crop yield prediction.

The project

- Enhanced data-driven decision-making in agriculture.
- Improved yield forecasting, reducing uncertainty and waste.

# 6. References


1. MAAIF Ministry of Agriculture, Animal Industry and Fisheries. https://www.agriculture.go.ug
2. Agricultural Survey 2019 Report. Uganda Bureau of Statistics (UBOS). Kampala, Uganda. https://www.ubos.org/wp-content/uploads/publications/04_2022AAS2019_Report.pdf
3. https://pmc.ncbi.nlm.nih.gov/articles/PMC10853138/
4. https://ieomsociety.org/proceedings/2023vietnam/46.pdf
5. https://www.mdpi.com/2072-4292/17/10/1717
6. https://www.nature.com/articles/s41598-020-80820-1#Sec2