1. Introduction

1.1 Background

Agriculture is a critical sector for many African countries, serving as a key source of income and sustenance for numerous farming households. However, ensuring food accessibility and profitability remains a significant challenge, given the various socio-economic and environmental factors that influence agricultural productivity and sustainability, primarily climate according to McKinsey and Company[1]. In this context, the Agricultural Survey of African Farm Households dataset, obtained from Kaggle[2], provides a valuable resource for understanding the dynamics of farming practices, income levels, and food security across the African continent.

1.2 Objective

The primary objective of this analysis is to develop predictive models that can accurately assess the factors contributing to food accessibility and profitability for farming households in Africa. By leveraging machine learning techniques, we aim to identify the crucial determinants affecting food accessibility and profitability, thereby providing insights that can inform policymakers, non-governmental organisations, and other stakeholders in the agricultural sector. Through this analysis, we seek to contribute to the advancement of sustainable agricultural practices and the improvement of food security in the region.

1.3 Dataset Overview

The dataset utilized in this analysis, the Agricultural Survey of African Farm Households[3], contains comprehensive information collected from farming households across eleven African countries. It includes various socio-economic, demographic, and agricultural indicators, offering a holistic view of the challenges and opportunities faced by these communities. The data set, obtained from a survey, specifies farming systems characteristics that can help inform about the importance of each system for a country's agricultural production and its ability to cope with short- and long-term climate changes or extreme weather events.

---

[1] "Effects of climate change on agriculture in Africa | McKinsey." 18 May. 2020, https://www.mckinsey.com/capabilities/sustainability/our-insights/how-will-african-farmers-adjust-to-changing-patterns-of-precipitation. Accessed 9 Nov. 2023.
[2] "Kaggle: Your Machine Learning and Data Science Community." https://www.kaggle.com/. Accessed 9 Nov. 2023.
[3] "Agricultural Survey of African Farm Households - Kaggle." https://www.kaggle.com/datasets/crawford/agricultural-survey-of-african-farm-households. Accessed 9 Nov. 2023.

2. Data Exploration and Preparation

2.1 Overview

The Agricultural Survey of African Farm Households dataset comprises 9597 rows that represent the households and 1753 columns with details about the households across the following African countries:
1. Burkina Faso
2. Cameroon

3. Ghana
4. Niger
5 Senegal
6. Egypt
7. Ethiopia
8. Kenya
9. South Africa
10. Zambia
11. Zimbabwe

After cleaning and preprocessing the data, we obtained valuable insights that lay the foundation for our subsequent analysis.

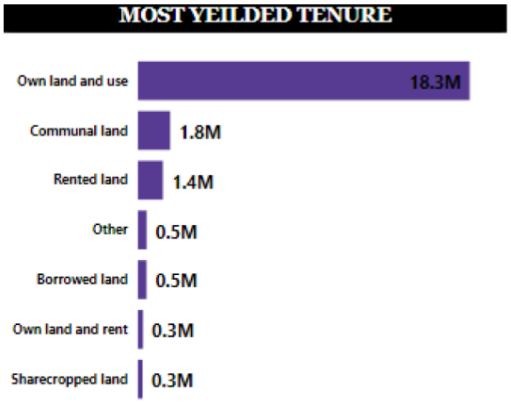
2.2 Key Findings

2.2.1 Total Yield and Transportation Costs

Our analysis reveals that the total agricultural yield recorded across the dataset amounts to an impressive 29 million units, indicating the substantial production output within the surveyed farming households. In parallel, the total transportation cost associated with the agricultural activities in the dataset is estimated to be 351 million, underscoring the significance of transportation expenditure in the agricultural supply chain.

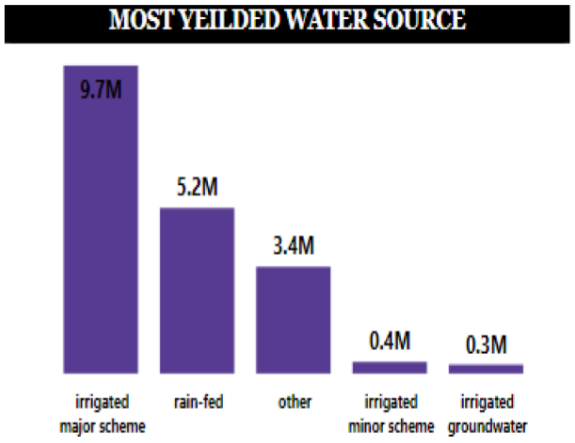
2.2.2 Tenure and Water Sources

Among the surveyed farming households, the predominant form of land tenure is "own land" as opposed to other options such as "community land," "borrowed land," or "rented land." This suggests a significant preference for land ownership and management, and its correlation with yield, within the surveyed communities.



There are different tenure used in this analysis, but the tenure that yielded the most is the '**Own land and Own use**' that gave 18.3M yield in season 1.

In terms of water sources utilized for agricultural activities, our analysis highlights that the majority of farming households rely on "irrigated major schemes" which yields the most compared to other options like "rain-fed" or "irrigated minor schemes." This finding underscores the critical role of large-scale irrigation systems in supporting agricultural productivity and yield stability.



There are different water sources used for planting

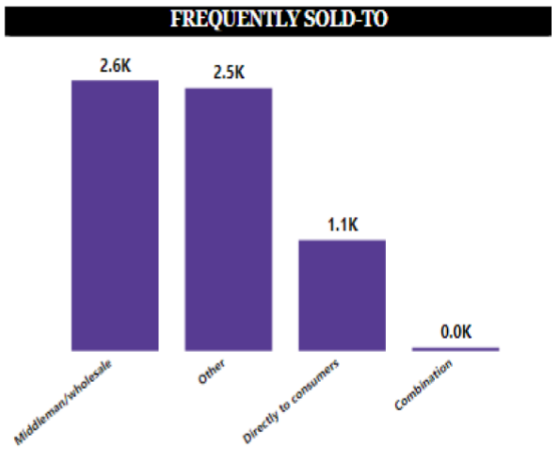
From the visual, irrigated major scheme(public) source was the water source that yielded the most with 9.7M.

2.2.3 Transportation Modes and Sales Channels

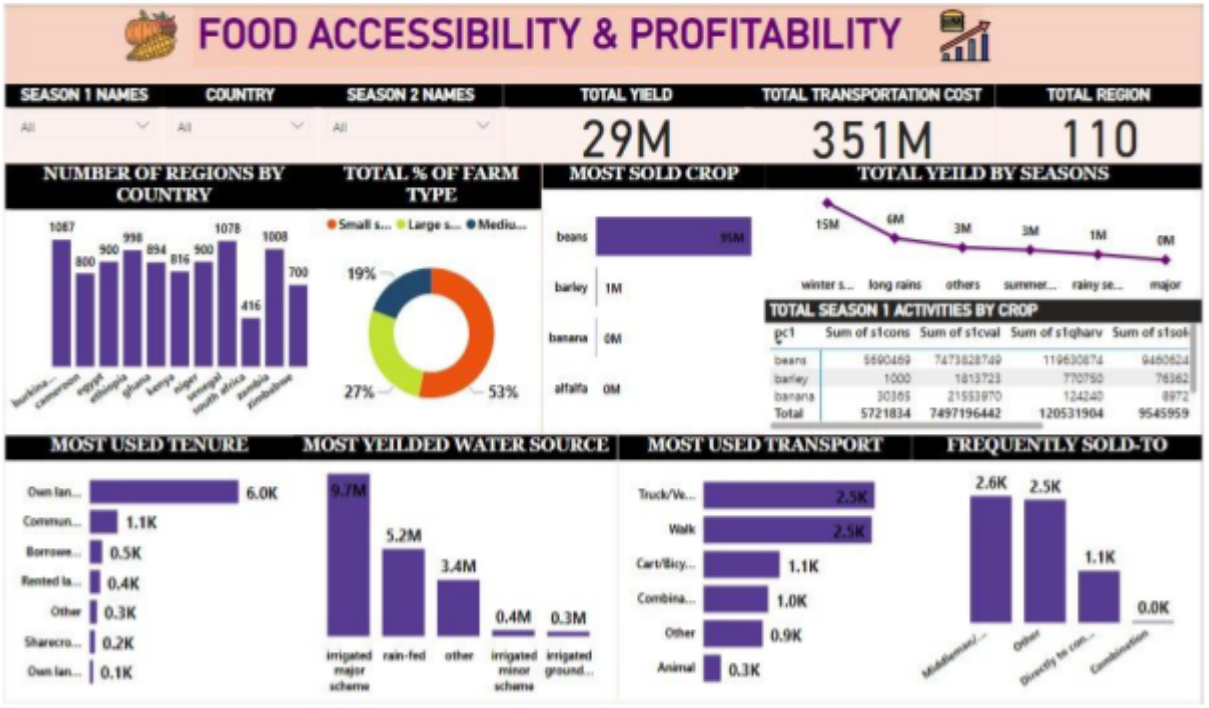
Furthermore, our examination of transportation preferences within the dataset indicates that "trucks" are the most commonly utilized mode of transport for agricultural goods, indicating

the importance of efficient and reliable transportation infrastructure for facilitating the movement of agricultural produce.

In terms of sales channels, our findings reveal that a significant portion of agricultural products are sold to "middlemen," suggesting the prevalence of intermediary market structures within the agricultural value chain. This highlights the need for a deeper understanding of the dynamics between farmers and intermediaries to ensure fair and transparent pricing mechanisms.



From the visual, the output gotten from the harvested crops of the farm were mostly sold to the wholesalers(middleman) and then others, then sold directly to consumers.



Data Visualization & Exploration of Agricultural Survey of African Farm Households Dataset

2.3 Data Preprocessing

The data, as regards the survey, was observed to be in sections. Hence, we identified the sections relating to our topic and created individual data frames for the selected sections to allow for easy cleaning of each section.

The sections include:

1. Section 3- Tenure issues

2. Section 4 – Details on Farming Activities

3. Section 7 – Climate Adaptation

Cleaning Section 3 involved selecting essential columns, rectifying anomalies within the 'Country' ('adm0'), 'Region' ('adm1'), and 'season name' columns, and addressing outliers in the coded categorical columns. Similarly, Section 4 underwent a similar cleaning process, involving the removal of unrelated columns pertaining to seasons 3 and crops 2-6. Additionally, outliers in the coded categorical columns were corrected.

The climate adaptation columns were binary coded. The cleaning process for this section entailed rectifying data that deviated from the binary format and converting unrecoverable data into NA values. In all three sections, columns with over 80% null values were dropped, as they lacked substantial information for analysis.

Following individual section cleaning, the cleaned data frames were concatenated for further processing. The date columns in the dataset were in poor condition. In an attempt to salvage relevant information, a parser was utilized to extract the months from a significant number of rows.

Subsequently, the data was duplicated and labeled into two sets: one for modeling and another for analysis. The model data was further refined by removing columns related to season 2, streamlining the model for analysis of a single season. Column names were standardized by applying a replacement dictionary for clarity and readability. Additionally, categorical columns in the analysis data were restructured, converting the existing codes into their corresponding string values using a replacement dictionary tailored for each column.

These preprocessing steps enabled us to derive meaningful insights that accurately reflect the dynamics of food accessibility and profitability within African farming communities and ensured the dataset was prepared for in-depth analysis and modelling endeavors.

3. Model Training

3.1 Model Selection

In our analysis, we employed two machine learning algorithms, namely Random Forest Regressor and LightGBM Regressor model. These algorithms were selected based on their demonstrated efficacy in handling complex, multi-dimensional datasets and their ability to capture non-linear relationships within the data.

3.2 Feature Selection

Furthermore, to enhance the robustness and interpretability of our models, we utilized Recursive Feature Elimination (RFE), a technique that systematically selects the most relevant features for model training. The reason for utilizing RFE for feature selction was simple:
1. Features names were clearly defined

2. Data is numerical

3. The are a lot of features in the data

By iteratively removing less significant features, RFE allowed us to streamline the feature space (from 1753) and focus on the most influential variables (54 columns and 6823 rows) affecting food accessibility and profitability.

3.3 Model Training Process

We split the preprocessed dataset into training and testing sets to ensure the evaluation of model performance on unseen data. The Random Forest Regressor leveraged an ensemble

learning technique to capture complex relationships within the data, while the LightGBM Regressor utilized gradient boosting to handle the large-scale datasets efficiently.

3.4 Evaluation Metrics

Following the model training phase, we evaluated the performance of the trained models using the Root Mean Square Error (RMSE) metric. RSME metric is used to evaluate the accuracy of a regression model. The lower the RMSE, the better the model is at predicting the target variable. The formula to calculate RMSE is as follows:

RMSE=n∑i=1n(Pi−Oi)2

where Pi is the predicted value for the ith observation in the dataset, Oi is the observed value for the ith observation in the dataset, and n is the sample size 1.

For example, suppose we have a regression model that uses “hours studied” to predict “exam score” of students on a particular college entrance exam. We can calculate the RMSE of this model to assess how well it fits the dataset. If the RMSE is low, it means that the model is able to predict the exam score accurately based on the number of hours studied. Conversely, if the RMSE is high, it means that the model is not able to predict the exam score accurately based on the number of hours studied 1.

The RMSE values obtained for the Random Forest Regressor and the LightGBM Regressor were 0.616921 and 0.363188, respectively.

4. Results

In our quest to decode the enigmatic patterns of food accessibility and profitability in African farming communities, based on the comprehensive model training and evaluation process, we successfully developed the LightGBM Regressor powered predictive model that accurately predicts food accessibility and profitability in African farming households. The significantly lower RMSE value obtained for the LightGBM Regressor underscores its effectiveness in capturing the intricate dynamics of agricultural production and market behavior, thus offering valuable insights for policymakers, stakeholders, and organizations seeking to implement targeted interventions and strategies for sustainable agricultural development and improved food security across the African continent.