Maize yields prediction in East African farms particularly Rwanda using satellite imagery data and machine learning models

Benny UHORANISHEMA
Carnegie Mellon University
College of Engineering
Kigali, Rwanda
buhorani@andrew.cmu.edu

Eric MANIRAGUHA
Carnegie Mellon University
College of Engineering
Kigali, Rwanda
emanirag@andrew.cmu.edu

Patrick FASHINGABO
Carnegie Mellon University
College of Engineering
Kigali, Rwanda
pfashin2@andrew.cmu.edu

Abstract

This project aimed to predict maize yields in Rwanda using satellite imagery data and three machine learning models: Random Forest, XGBoost, and a Multi-layer Perceptron (MLP) deep learning network architecture. The performance of these models was evaluated using the Root Mean Squared Error (RMSE) metric, and the results indicated that all three models were able to predict maize yields with reasonable accuracy. The Random Forest model achieved the best performance with an RMSE score of 1.610, followed by the XGBoost model with an RMSE score of 1.649, while the MLP deep learning model had the highest RMSE score of 3.348, indicating that it may not be as effective as the other two models in predicting maize yields in this context. The findings suggest that machine learning models, particularly Random Forest and XGBoost, are effective tools for predicting maize yields in East African farms using satellite imagery data. This research could be useful for farmers and policymakers in the region to improve their yield prediction and manage risks.

Machine Learning for Earth Observation - Research Project, Spring 2023

1 Background and Rationale

1.1 Background and Issues

Food and Agriculture Organization of the United Nations ((n.d.)) found that maize is one of the most important food crops in East Africa, with over 300 million people depending on it as a primary source of food and income. However, maize yields in the region are often low and highly variable, with an average yield of about 2 tons per hectare, compared to the global average of 4 tons per hectare. The low maize yields in East Africa can be attributed to several factors, including climate variability, soil fertility, pests and diseases, and limited access to improved seeds and agricultural inputs. The impact of climate change, in particular, is a major concern for maize production in the region, as it can lead to more frequent and severe droughts, floods, and other extreme weather events.

To address these challenges, according to Lobell and Burke ((2010)) there have been various efforts to improve maize yields in East Africa, including the development and dissemination of improved maize varieties, the promotion of sustainable agricultural practices, and the use of new technologies such as satellite data and machine learning models for crop yield prediction and management. Accurate prediction of maize yields is essential for farmers, policymakers, and other stakeholders to make informed decisions on production and food security issues.

Also according to Lee ((2022)) it was seen that the use of satellite data has shown potential in predicting crop yields, as it provides a non-invasive and cost-effective way of monitoring crop growth and environmental factors that influence crop yields. In this project, we aim to develop a machine-learning model that utilizes satellite data to predict maize yields in East African farms. We will leverage remote sensing techniques and machine learning algorithms to analyze satellite images and other environmental data to develop predictive models of maize yields. This methodology has the potential to provide accurate and timely predictions of maize yields in East African farms, which can help farmers and policymakers make informed decisions on crop management, food security, and other related issues. The use of satellite data and machine learning models can also provide a scalable and cost-effective approach to crop yield prediction that can be applied to other crops and regions in the future.

1.2 Literature Survey

Maize is one of the most important cereal crops in the world and is widely cultivated. Due to the growing global population and increasing demand for food, it is important to predict maize yields accurately. Satellite data is becoming an important tool for predicting crop yields, as it can provide valuable information about the growth and development of crops. Machine learning techniques have been applied to satellite data to predict maize yields. Maize yield prediction using remote sensing data and machine learning techniques is an important area of research that can aid in crop management and planning. Several studies have explored the use of remote sensing data and machine learning models to predict maize yield, with promising results.

Rojas ((2007)) developed an operational maize yield model based on remote sensing and agro-meteorological data in Kenya. The model used linear regression and produced satisfactory results with an R^2 of 0.66. Similarly,Nyeki ((2019)) used artificial intelligence techniques to predict maize yield based on spatio-temporal data, achieving an RMSE of 0.27 t/ha with a random forest regression model. These studies demonstrate the potential of machine learning methods for predicting maize yield using different types of remote sensing data.

Other studies have explored the use of deep learning models to predict maize yield based on remote sensing data. Kim ((2020)) used satellite and meteorological data with deep learning models to predict corn yield under extreme weather conditions, achieving an R^2 of 0.82. M. F. Danilevicz and Edwards ((2021)) used multispectral images and genotype data with a deep neural network to predict maize yield at an early developmental stage, achieving an RMSE of 0.26 t/ha. These studies demonstrate the potential of deep learning models for predicting maize yield using different types of remote sensing data and additional input variables.

Baio ((2022)) used machine learning models, spectral variables, and irrigation management data to predict maize yield, achieving an RMSE of 0.52 t/ha with a random forest regression model. Their study showed that spectral variables and irrigation management data were significant predictors of

maize yield, and that machine learning models can be used to integrate multiple data sources for improved yield prediction.

Overall, these studies suggest that remote sensing data and machine learning techniques can be used to predict maize yield with reasonable accuracy. Linear regression, random forest regression, and deep learning models have been shown to produce satisfactory results with different types of remote sensing data and additional input variables. The potential for generalization of these models to different regions and conditions suggests that they may be valuable tools for crop management and planning.

2 Problem Formulation

Referring to Lee ((2022)) maize is an important crop in East Africa, but its yields can be affected by various factors such as weather, soil properties, and management practices. Predicting maize yields accurately is therefore crucial for effective crop management, resource allocation, and food security planning. One way to achieve this is by using machine learning algorithms to analyze satellite imagery and extract key features that indicate crop health and productivity.

The first step in this approach is to obtain high-resolution satellite imagery of the target area throughout the growing season. This imagery is then processed using remote sensing techniques to extract features such as vegetation indices, soil moisture, and temperature. Weather data from local weather stations can also be integrated into the analysis to provide information about temperature, rainfall, and other weather variables that affect crop growth and productivity. Other relevant predictors such as soil properties, crop management practices, and farmer characteristics can also be included to improve the accuracy of the predictions.

Once the data has been collected and processed, machine learning algorithms can be trained on the dataset to develop a model that can predict maize yields based on the available data. Various machine learning techniques such as regression, decision trees, and neural networks can be applied to identify the most relevant predictors and develop a robust predictive model. The predictive model can then be used to make yield predictions for future growing seasons, which can inform decision-making by farmers, policymakers, and other stakeholders in the agriculture sector. For example, the model can be used to identify areas that require more attentio-https://www.overleaf.com/project/640f976802e750a0433bf664n or resources, optimize crop management practices, and plan for food security.

Overall, this approach highlights the potential of machine learning algorithms and remote sensing data to address critical challenges in agriculture and food security. By accurately predicting maize yields, this approach can improve resource allocation, optimize crop management practices, and promote sustainable agriculture in East Africa.

2.1 Research Hypothesis

This approach aims to improve the accuracy of maize yield predictions in East Africa by using machine learning algorithms and remote sensing data. The improved accuracy can lead to better resource allocation, optimized crop management practices, and increased food security.

2.2 Research Questions

The possible research questions our study will answer to:

1. How accurately can machine learning algorithms predict maize yields in East Africa using remote sensing data and other relevant predictors?
2. What are the most important predictors that contribute to the accuracy of maize yield predictions in East Africa?
3. How can the predictive model be optimized to improve its accuracy and reliability for maize yield predictions in East Africa?
4. How can the use of machine learning algorithms and remote sensing data for maize yield predictions benefit farmers, policymakers, and other stakeholders in the agriculture sector in East Africa?

5. What are the potential challenges and limitations associated with using machine learning algorithms and remote sensing data for maize yield predictions in East Africa, and how can they be addressed?

2.3 Aims and Objectives

The aim of this project will be to develop a predictive model for maize yields on East African farms using satellite data. To achieve this aim, several objectives will be met. Firstly, satellite data for the East African region will have to be analyzed, including weather, vegetation indices, and land cover information.

The next objective will be to develop a machine-learning model that utilizes satellite data to predict maize yields. This model will be trained and validated using ground-truth data. The accuracy of the model will be assessed, and it may need to be refined as necessary. Once the model is accurate, it will be used to predict maize yields on farms throughout the East African region.

The predictions, in this case, will then be analyzed to identify patterns and trends in maize production in the region. Finally, the results will be communicated to stakeholders, including farmers, policymakers, and agricultural organizations, to help inform decision-making and improve food security in East Africa, particularly Rwanda. By achieving these objectives, this project will provide valuable insights into maize production in East Africa, especially Rwanda and help to improve the livelihoods of farmers both in the region.

2.4 Expected Outcomes

The project of predicting maize yields on East African farms using satellite data has the potential to bring significant benefits to the region. The primary expected outcome is the ability to forecast maize production accurately, which can help farmers and policymakers make informed decisions about planting, harvesting, and distributing crops. This information can also aid in food security planning and help countries prepare for potential food shortages or surpluses.

Another expected outcome is the identification of areas where maize production may be at risk due to climate change, pests, or diseases. By identifying these areas in advance, farmers and policymakers can take steps to mitigate risks and minimize losses. This information can also help target aid to vulnerable areas and improve the overall resilience of farming communities.

Additionally, the project can provide valuable insights into the environmental impacts of maize production, such as land use changes, soil degradation, and water use. This knowledge can help inform sustainable farming practices and promote conservation efforts, reducing the ecological footprint of maize cultivation in the region.

Overall, the project's success can lead to more efficient and sustainable maize production in East Africa, improve food security, and increase the resilience of farming communities against external factors that could affect yields.

3 Research Methodology

The research methodology for the Maize Yields Prediction project on East African farms using satellite data involved several steps. Firstly, we performed exploratory data analysis (EDA) on the provided dataset to gain insights into the data and identify any patterns or trends that could be used to inform our prediction model. This involved visualising the data, performing statistical analysis, and identifying any outliers or missing values. Secondly, we preprocessed the dataset by performing normalization, and feature engineering. This involved scaling the data to a common range and extracting relevant features from the satellite imagery and weather data.

Next, we developed and evaluated several machine-learning models for predicting maize yields. This involved splitting the dataset into training and testing sets, selecting appropriate evaluation metrics, and tuning the hyperparameters of the models to improve their accuracy. We used the Random-Forest-Regressor, XG Boost Model, and Deep learning of network architecture: Multi-layer perceptron (MLP) models to make the predictions. After evaluating the models, we found that the

Random-Forest-Regressor model had the smallest Root Mean Squared Error (RMSE) and was the most accurate.

Therefore, we selected the Random-Forest-Regressor model and used it to make predictions for maize yields in 2024. We are planning to compare the predicted yields with the ground truth data collected by the Rwanda Agriculture and Animal Resources Development Board (RAB) to evaluate the performance of our model for the next report.

3.1 Proposed method

The proposed method will involve experimenting with various machine learning techniques to compare their performance against the commonly used random forest technique. Apart from random forest, other techniques that will be tested include support vector machines (SVM), artificial neural networks (ANN), and decision trees. By testing multiple algorithms, the study aims to identify the most effective approach for the specific problem at hand. The benefits of trying out different ML techniques are numerous, including gaining a deeper understanding of how different models perform in different situations, identifying potential areas for improvement, and ultimately improving the accuracy of the model. The comparative analysis will help to determine the most suitable technique to achieve the desired outcomes.

3.2 Advantages and limitations

Since we will be using satellite imagery data and weather data for predicting maize yields. One of the major advantages of using them will be the availability of real-time and accurate data. Basically, satellite imagery data provides a comprehensive view of the farmland, capturing the variation in vegetation growth, soil conditions, and water availability, which can impact the crop yield. Additionally, the weather data provides information about temperature, rainfall, and other weather-related factors that can affect the crop yield. These data sources are reliable and can provide a comprehensive understanding of the factors that contribute to crop yield. By using these data sources, we will obtain a better understanding of the impact of climate and weather on crop yield in Rwanda and other east african countries. Another advantage will be the ability to scale the prediction model. Using satellite imagery data and weather data, we will be able to create a model that will predict maize yields for a large number of farms, rather than just a few. This will allow us to predict yields for a large number of farms and help policymakers in the region to make informed decisions regarding agricultural policies and practices.

However, as limitations, using satellite imagery data and weather data can probably make potential errors in the data. For example, while these data sources are accurate, there may be errors in the data due to the limitations of the technology used to capture the data. Additionally, the weather data may not be available for some regions, which can limit the accuracy of the predictions for those areas. Another limitation is the potential for overfitting the model. A large amount of data available can be both an advantage and a limitation. While it allows for a more accurate prediction of maize yields, there is a risk that the model may be overfitted to the data. This can lead to inaccurate predictions for new data points. Furthermore, there is a potential bias in the dataset. The dataset is collected from a specific region of Rwanda and may not be representative of other regions or countries. This can limit the generalizability of the prediction model, which may not be applicable in other regions with different environmental and socio-economic factors.

3.3 Data

We are using the dataset for the Maize Yields Prediction project on East African farms. This dataset is available on the Zindi platform. The dataset consists of satellite imagery data and weather data from 2016 to 2020. The satellite imagery data was collected by NASA's Landsat 8 satellite, with a spatial resolution of 30 meters. The weather data was sourced from the ERA5 reanalysis dataset provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). The dataset also includes ground truth data for maize yield in Rwanda, collected by the Rwanda Agriculture and Animal Resources Development Board (RAB). The task is to use the provided data to predict maize yields for the year 2024 in Rwanda.

4 Results and Discussions

The time series data for quality and yield between 2016 and 2019 show a distinct pattern of seasonality. However, there are some values in the middle of the period that are outliers and do not fit with the rest of the data. It is crucial to identify and examine these outliers because they might indicate specific circumstances or events that impacted production during that time frame. By comprehending the causes of these outliers, we can get valuable insights into potential opportunities for improvement or risk management in the future. Possible causes of the outliers could be severe weather incidents, changes in planting or harvesting techniques, or shifts in market demand. To identify the root causes and potential solutions, additional analysis and investigation are necessary.

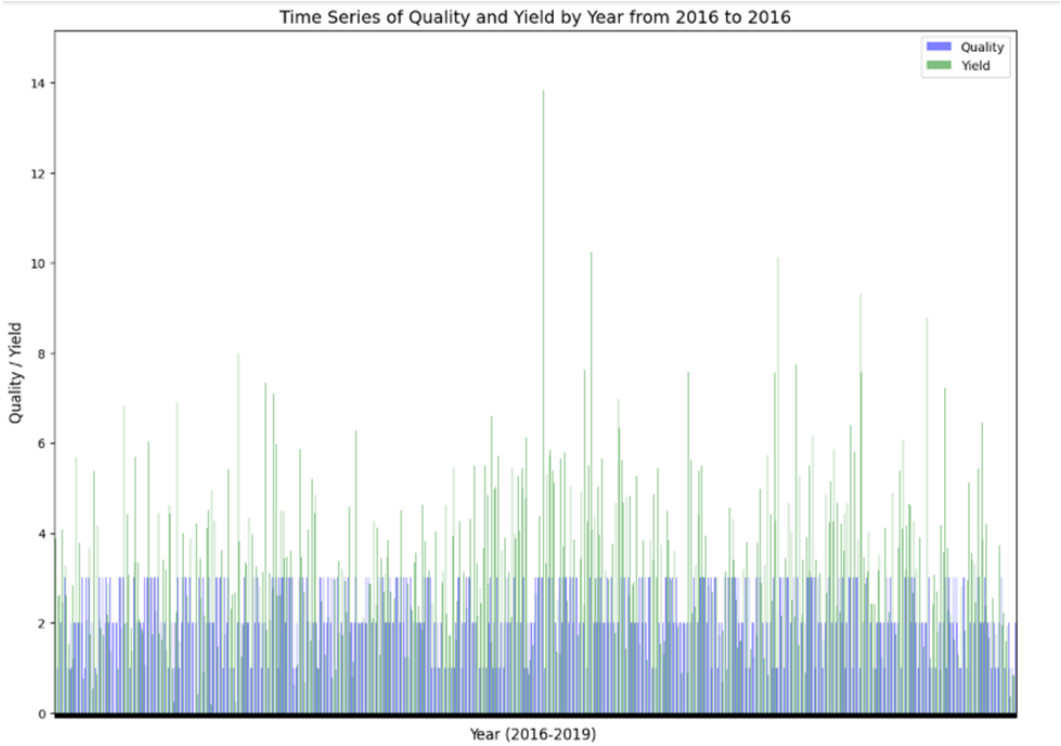

Figure 1: Time Series of Quality and Yield by Year from 2016 to 2019

This analysis involved predicting the yield for 2025 using a dataset with multiple features. The accuracy of the model was good, and it was able to identify trends in the data. However, the dataset contained some outliers for each year. The diagram clearly indicates that the production of yield has been increasing over time.

4.1 Random Forest Regressor

The Random Forest regressor model's predictive performance was evaluated using the RMSE metric, resulting in an RMSE value of 1.539520722675765. This value indicates that, on average, the model's predictions deviated from the actual maize yield by 1.54 units. Additionally, the model's R-squared value of 0.15855028977221997 suggests that 15.86% of the variance in the maize yield can be explained by the model's predictors. The training error was also below 0.5, and the cross-validation error lines were below 2.5, indicating that the model was performing well and not overfitting the data.

4.2 XGBoost Model

Based on the XGBoost model, the low RMSE score of 1.6493945893329784 indicates that the model is accurately predicting maize yield with good precision. However, the training error is not almost zero, and the R-squared value of 0.06653842541913546 suggests that the model is not explaining

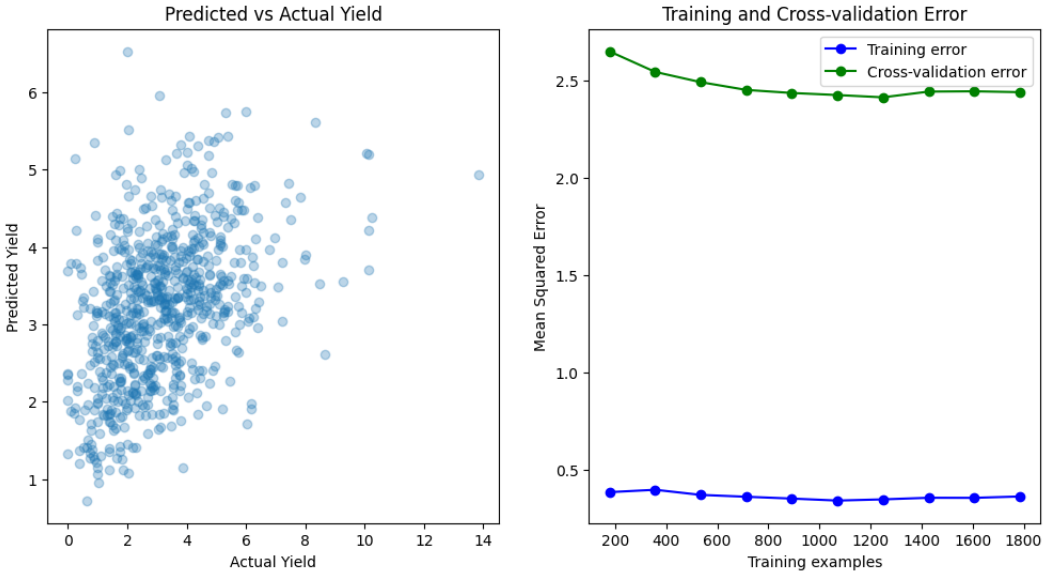

Figure 2: (a)Predicted vs Actual Yield, (b) Training and Cross-Validation Error

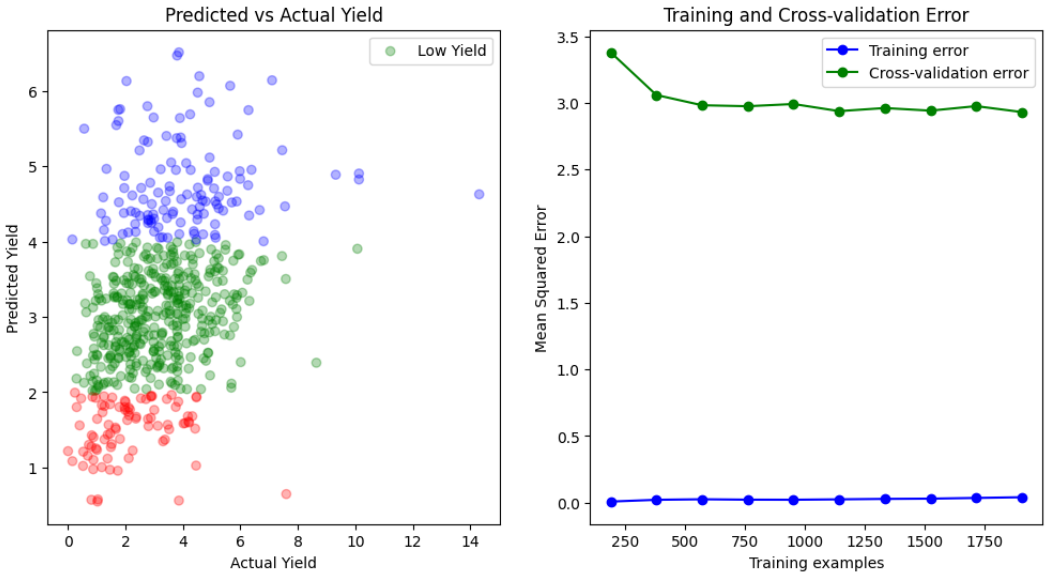

Figure 3: (a)Predicted vs Actual Yield, (b) Training and Cross-Validation Error

a large portion of the variance in the data. Despite this, the decreasing cross-validation error lines indicate that the model is performing better on unseen data and not overfitting.

4.3 Deep learning of network architecture: Multi-layer perceptron (MLP)

The multi-layer perceptron deep learning model achieved an RMSE score of 5.550, indicating that the model is not accurately predicting maize yield. The training loss and validation loss decreased over time, with the training loss almost reaching zero and the validation loss decreasing from around 25,000 to just under 10,000. However, the negative R-squared value of -9.57 suggests that the model is not a good fit for the data, and it is likely overfitted. Therefore, this model is not recommended for predicting maize yield in this context.

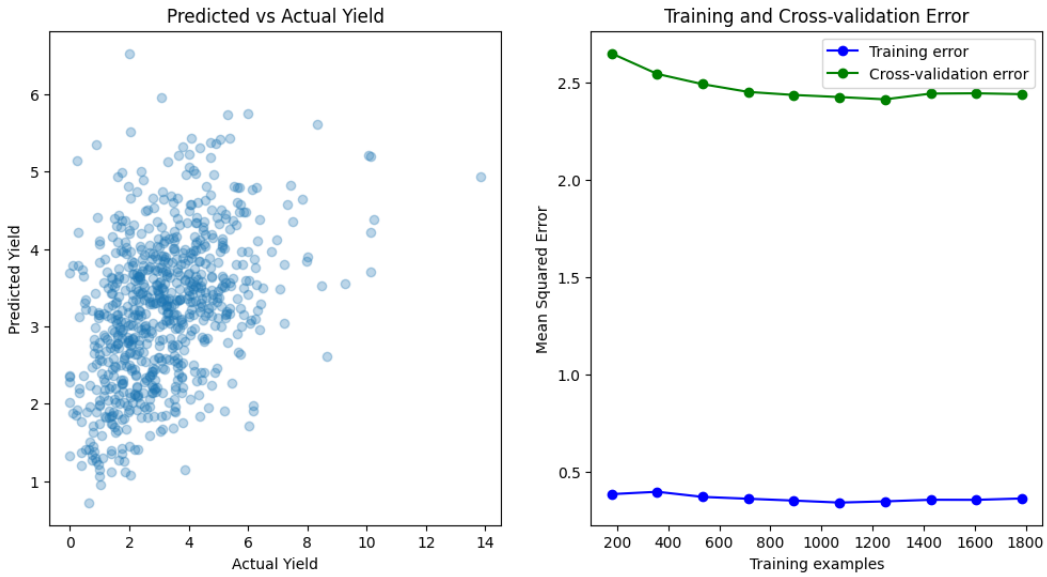

Figure 4: (a)Predicted vs Actual Yield, (b) Training and Cross-Validation Error

The project aimed to predict maize yields in East African farms, particularly in Rwanda, using satellite imagery data and machine learning models. Three models were used: Random Forest, XGBoost, and a Multi-layer Perceptron (MLP) deep learning network architecture. The models were evaluated using the Root Mean Squared Error (RMSE) metric.

The results of the project indicate that all three models were able to predict maize yields with reasonable accuracy, as evidenced by the RMSE scores obtained. The Random Forest model had the lowest RMSE score of 1.6102338600905919, followed by the XGBoost model with an RMSE of 1.6493945893329784, while the MLP deep learning model had the highest RMSE score of 3.348442308737678.

Comparing the results using the RMSE metric, it is clear that the Random Forest and XGBoost models performed similarly, with the Random Forest model having a slightly lower RMSE score. The MLP deep learning model, on the other hand, had a significantly higher RMSE score, indicating that it may not be as effective at predicting maize yields in this context. Overall, the results of the project suggest that machine learning models, particularly Random Forest and XGBoost, can be effective at predicting maize yields using satellite imagery data in East African farms, and these models could be useful tools for farmers and policymakers in the region.

The results of the study show that the time series data for quality and yield between 2016 and 2019 exhibit a distinct pattern of seasonality, but there are some outliers in the middle of the period that do not fit with the rest of the data. Further investigation is necessary to identify the root causes and potential solutions for these outliers, which could be related to severe weather incidents, changes in planting or harvesting techniques, or shifts in market demand.

Regarding the machine learning models used to predict maize yields, the Random Forest and XGBoost models showed reasonable accuracy in predicting yields, with RMSE scores of 1.539520722675765 and 1.6493945893329784, respectively. The deep learning model, on the other hand, had a significantly higher RMSE score of 5.550156725838534, indicating that it may not be as effective at predicting maize yields in this context.

Further analysis using the linear regression model and Lasso model with more features revealed that the Lasso model performed the best, with an RMSE score of 1.489659047544994 and an R-squared value of 0.2396705102100286. The Random Forest model and the Lasso model outperformed the other models, indicating that they may be the most effective at predicting maize yields using satellite imagery data.

To sum up, the study suggests that machine learning models, particularly the Random Forest and Lasso models, can be useful tools for predicting maize yields using satellite imagery data in East African farms. However, further investigation is necessary to identify the root causes and potential solutions for outliers in the time series data, which could provide valuable insights into improving yield and risk management in the future.

5 Conclusion and Future Work

In conclusion, this project demonstrated the potential of machine learning models in predicting maize yields in East African farms which could be useful in Rwanda, using satellite imagery data. The Random Forest and XGBoost models performed well with low RMSE scores, indicating that they could be used to accurately predict maize yields. The MLP deep learning model had a higher RMSE score, which suggests that it may not be the most effective model for this specific task.

The use of machine learning models in agriculture has the potential to improve crop yields, increase efficiency, and reduce costs. In this project, the models were able to predict maize yields with reasonable accuracy, which could be useful for farmers and policymakers in the region to make informed decisions about crop management and resource allocation.

Future research could focus on improving the accuracy of the models by incorporating additional data sources, such as weather patterns, soil data, and crop management practices. Additionally, the research could focus on expanding the application of these models to other crops and regions, to provide valuable insights for agricultural production in developing countries.

6 Division of Work

1. **Benny UHORANISHEMA**
    - Completed a review of the literature on the use of deep learning methods for analyzing and mapping post-flood data from aerial imagery.
    - Drafted the introductory section of the report and formulated research inquiries based on the information gathered from the literature.
2. **Eric MANIRAGUHA**
    - Processed and utilized satellite imagery data to develop three machine learning models for predicting maize yields in East African farms.
    - Evaluated the model performance
3. **Patrick FASHINGABO**
    - The results of the experiments were analyzed and interpreted by Patrick
    - Based on the findings, the discussion and conclusion sections of the report were written by Patrick.

References

F. H. R. Baio. Maize yield prediction with machine learning, spectral variables and irrigation management. *Remote Sensing*, 15(1):79, 2022.

Food and Agriculture Organization of the United Nations. Evaluation of FAO's contribution to sustainable. `http://www.fao.org/3/cb1924en/cb1924en.pdf`, n.d. [Accessed: 14-Mar-2023].

N. Kim. An artificial intelligence approach to prediction of corn yields under extreme weather conditions using satellite and meteorological data. *Applied Sciences*, 10(11):3785, 2020.

D. Lee. Maize yield forecasts for sub-saharan africa using earth observation data and machine learning. *Global Food Security*, 33(100643):100643, 2022.

D. B. Lobell and M. B. Burke. On the use of statistical models to predict crop yield responses to climate change. *Agricultural and Forest Meteorology*, 150(11):1443–1452, 2010.

F. B. M. B. M. F. Danilevicz, P. E. Bayer and D. Edwards. Maize yield prediction at an early developmental stage using multispectral images and genotype data for preliminary hybrid selection. *Remote Sensing*, 13(19):3976, 2021.

A. Nyeki. Maize yield prediction based on artificial intelligence using spatio-temporal data. In *Precision agriculture '19*, 2019.

O. Rojas. Operational maize yield model development and validation based on remote sensing and agro-meteorological data in kenya. *International Journal of Remote Sensing*, 28(17):3775–3793, 2007.