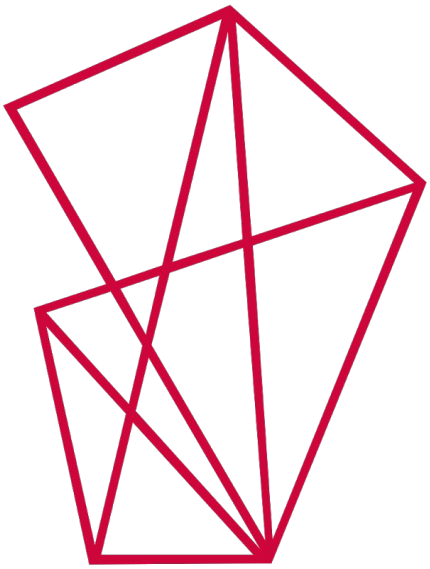


GAIA
by FUE

MASTER IN DECISION MAKING AND INNOVATION 2.0

GAIA PROGRAM

FINAL PROJECT

TITLE: APPLICATION OF MACHINE LEARNING ALGORITHMS TO PERFORM CROP YIELD PREDICTION.

STUDENT NAME: SARA DÍAZ DEL SER

TUTOR: ADOLFO MELENDEZ

Application of Machine Learning Algorithms to Perform Crop Yield Prediction

Sara Díaz del Ser

Abstract

The aim of this research was to develop a machine learning model to forecast crop yields with the objective of enhancing food security and encouraging sustainable agriculture practices. The Kaggle Crop Yield Prediction Dataset was utilized, containing data on pest infestations, soil conditions, and weather. The most effective method of forecasting crop yields was discovered to be a combination of weather and pesticide data with machine learning algorithms like Random Forest and Gradient Boost. An app mockup was also designed, with the aim of making it easier for farmers to input data and obtain their anticipated yield prediction, thus making the forecasts more accessible and widely accepted. It should be noted, however, that this research was based on a particular dataset obtained from a generalized global data. As a result, further research is required to assess the performance of the model on specific datasets, regions, and crops. This study's results are valuable to farmers, agricultural researchers, and policymakers because it provides them with crucial information to help them make informed decisions about their crops.

Machine Learning, Crop Production, Crop Yield Prediction, Agricultural AI, Regression Algorithms

Abbreviations

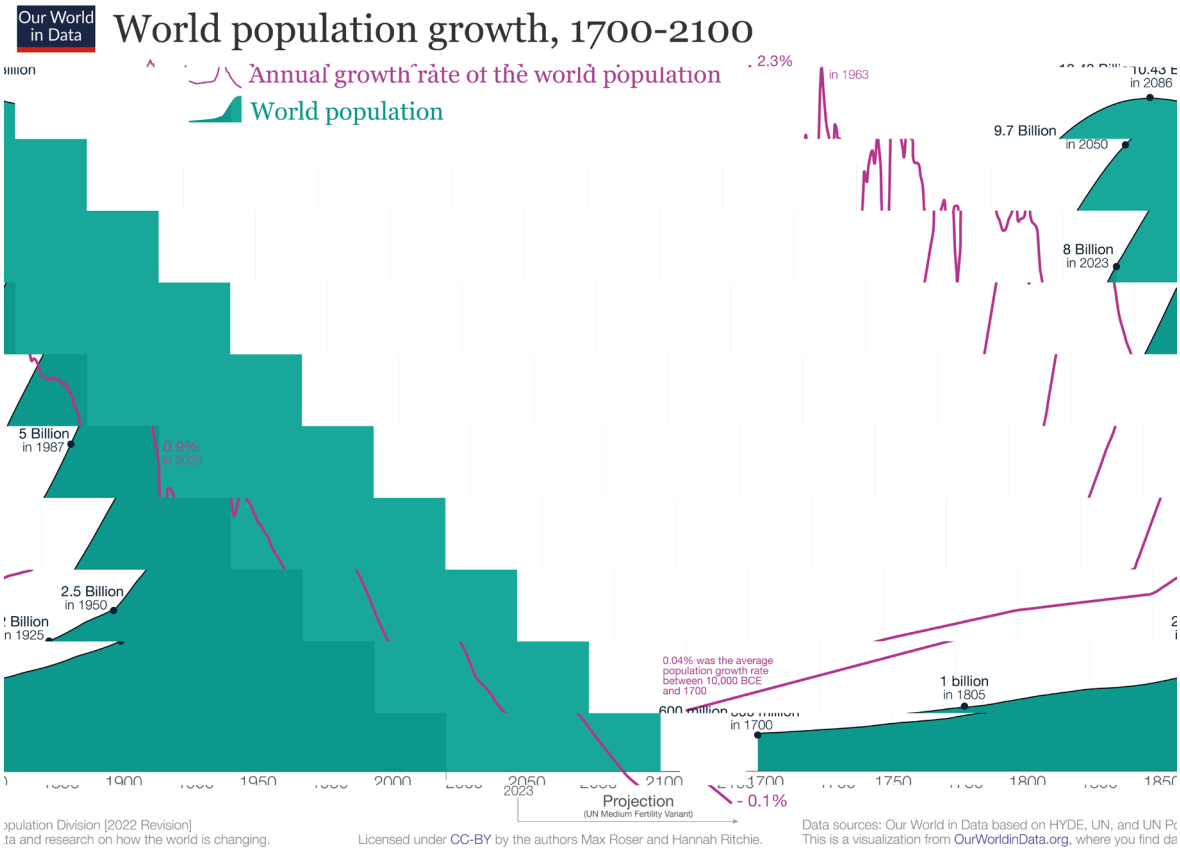
Adjusted R-Squared	Adj. R^2
Artificial Intelligence	AI
Exploratory Data Analysis	EDA
Machine Learning	ML
Mean Absolute Error	MAE
Mean Absolute Error Percentage	MAE%
R-Squared	R^2
Root Mean Squared Error	RMSE
Sustainable Development Goal	SDG

1. Introduction

1.1. Background of the Problem

Issues in the Agricultural System

The agriculture industry faces a daunting task in the coming years: producing enough food to meet the demands of over 8 billion human beings[47]. According to the United Nations, global food production must rise by a minimum of 60% by 2050 to meet the population's rapidly increasing needs[60][63]. This of course, while also tackling climate change's effects on the industry and preserving the planet's natural resources.



The current agricultural system is being confronted with significant sustainability challenges that stem from various interrelated factors [60][63].

One key issue is its heavy reliance on finite resources, including water, fossil fuels, and soil. Continuous extraction and usage of these resources have led to their depletion and have resulted in adverse environmental consequences. For instance, excessive water extraction has caused aquifer depletion and water scarcity in many regions. The extensive use of fossil fuels in farming operations contributes to greenhouse gas emissions, exacerbating climate change and its associated impacts.

Chemical inputs, such as fertilizers and pesticides, also pose significant challenges to the sustainability of the agricultural system. While these inputs are often used to enhance crop productivity and protect against pests and diseases, their excessive and improper application can lead to soil degradation, water pollution, and loss of biodiversity. Improper management of chemical inputs can result in nutrient imbalances in the soil, decreased soil fertility, contamination of water sources, and harm to beneficial organisms.

The agricultural system's vulnerability to climate change further compounds its sustainability challenges. Climate change manifests in the form of erratic weather patterns, increased frequency and intensity of droughts, floods, and extreme temperatures. These climate-related factors disrupt traditional agricultural practices, impacting crop growth and productivity. Changes in rainfall patterns can result in water stress for crops, while extreme temperatures can cause heat stress, affecting plant development and overall yields. The unpredictability of weather patterns makes it challenging for farmers to plan their planting and harvesting schedules, exacerbating the risks associated with agricultural production.

Another critical aspect of the current agricultural system is the concentration of production in large-scale commercial operations. This concentration leads to economic and social disparities, posing challenges to small-scale farmers and rural communities. Large-scale operations often benefit from economies of scale, access to capital, and advanced technologies, giving them a competitive advantage over smaller farmers. This can lead to the displacement of small-scale farmers, loss of livelihoods, and exacerbation of economic inequalities within rural areas.

In addition to environmental and economic concerns, the current agricultural system also poses health risks. Agricultural workers, who come into direct contact with chemicals during farming activities, face occupational hazards such as pesticide exposure, which can have long-term health implications. Furthermore, the production of processed foods associated with the industrialized agricultural system often involves the use of additives, preservatives, and unhealthy ingredients. The consumption of these foods has been linked to various diet-related health issues, including obesity, cardiovascular diseases, and other chronic conditions.

Addressing these sustainability challenges requires a transformation of the agricultural system towards more sustainable and resilient practices. This includes promoting sustainable resource management, such as efficient water usage, reduced reliance on fossil fuels, and improved soil conservation techniques. Encouraging organic farming practices and integrated pest management can help minimize the use of harmful chemicals while maintaining productivity. Diversifying agricultural systems, supporting small-scale farmers, and promoting equitable access to resources and markets are essential for reducing economic and social disparities within the sector. Additionally, promoting healthier and more sustainable food choices can contribute to improved public health outcomes[43][44][52].

Artificial Intelligence as a Potential Solution

The application of Artificial Intelligence (AI) and Machine Learning (ML) in agriculture provides opportunities to address these challenges and improve productivity.

AI and ML techniques enable computers to analyze vast amounts of data and learn patterns [4] and relationships that can be utilized to enhance agricultural practices. By leveraging AI algorithms, farmers can make data-driven decisions and optimize their resource utilization. This is particularly relevant in the context of the sustainability challenges discussed earlier[39].

Some direct applications of AI techniques in agriculture include: crop yield prediction, disease and pest detection, resource management, land optimization, and precision agriculture, among others [13][20].

- Crop Yield Prediction: machine learning models can analyze historical weather data, soil conditions, and other environmental factors to forecast real-time crop

performance. This information helps farmers plan their planting and harvesting schedules more effectively, optimizing resource allocation and minimizing waste.

- Disease and Pest Detection: by analyzing images of crops, machine learning algorithms can identify visual cues associated with diseases or pest infestation. Early detection enables farmers to take timely action, such as targeted pesticide application or the implementation of preventive measures, reducing the reliance on harmful chemicals and minimizing crop losses[29][51].
- Resource Management: for example, through predictive modeling and data analysis, machine learning algorithms can forecast water demand and identify areas within a field that require varying levels of irrigation. This optimization minimizes water usage while ensuring that crops receive adequate hydration[23].
- Land Optimization: by analyzing sensor data and providing insights into soil health, nutrient levels, and other key indicators. This information helps farmers make informed decisions about land management practices, such as crop rotation, soil conservation techniques, and optimal planting densities[37].
- Precision Agriculture: by collecting and analyzing data from sensors and other sources, farmers can fine-tune their fertilizer application rates, irrigation schedules, and other interventions. This targeted approach minimizes resource wastage, reduces environmental impacts, and enhances overall productivity[53][64].

Crop Yield Prediction

Using AI techniques to sustainably maximize crop production is an excellent first step to overcoming the challenges posed to the current system. Crop yield prediction enables farmers to make well-informed decisions regarding crop management and species selection to optimize production.

However, forecasting crop yield is a complicated task that involves several variables, such as temperature, rainfall, and pest control. As with most weather-influenced events, developing precise models that output reliable information on forecasted crop yield is a challenging task.

Predicting agricultural yields is complicated due to various factors. One essential factor is weather, which can significantly impact crop yields. For instance, drought can lead to a reduction in crop yields by depriving plants of water, and other extreme conditions such as high temperatures or heavy rainfall can also have a negative impact. Additionally, it is important to consider weather-related disasters, such as floods or hurricanes, which destroy crops and massively hinder production.

Pest infestations also pose a threat to crop yields. Insects, weeds, and diseases can damage crops, leading to reduced production. Pesticides, while effective in controlling pests, can have a negative impact on both the environment and human health. This translates into a growing need for sustainable pest control methods and reduced pesticide's use.

There are many other factors that affect crop yield. A farmer's ability to produce high yields can be impacted by their access to capital, technology, and other resources. Similar to how economic considerations can influence crop yields, social and cultural aspects like labour

availability do too. Thus, assessing how these economic and social factors affect agricultural growth and development is necessary in order to anticipate crop yields.

Machine learning algorithms can and have been widely used in crop yield prediction because they can learn the underlying patterns and relationships between the input variables and the output variable (in this case, crop yield). There are several types of machine learning algorithms that can be used to estimate agricultural yields (svm, decision trees, random forest, etc.), but these models have limitations, such as a need for high-quality data and the devastating possibility of bias.

In addition to traditional machine learning algorithms, newer techniques, such as deep learning, have shown promise in crop yield prediction. Deep learning algorithms can learn complex relationships in data and can be trained using large amounts of data. However, they require more data and computational resources to train, and may be more prone to overfitting (generalize badly) than traditional machine learning algorithms.

The state of the art in the machine learning field for crop yield prediction is constantly evolving, and there are many different approaches and techniques that can be used to build accurate and reliable models. It is important must carefully consider the strengths and limitations of each of the different algorithms and select the best approach for each specific need[5][21][27][55].

1.2. Statement of the Research Problem

The agriculture industry is confronted with the great challenge of producing sufficient food to meet the demands of a growing population, while also addressing climate change and preserving natural resources. Precise crop yield prediction is vital in overcoming these challenges, as it enables farmers to make informed decisions about crop management and selection to optimize yields. However, forecasting crop yield is a complicated task that involves numerous variables, including weather conditions, pest infestations and multiple socio-economic variables, all of which poses a challenge in developing accurate models.

1.3. Purpose of the Study

The purpose of this study is to build a machine learning model for predicting crop yields in order to improve food security and sustainable agriculture practices. Accurate crop yield prediction can help farmers make informed decisions about what to plant and how to manage their crops, which can lead to higher production and more sustainable agriculture practices. By using data on geographical location, temperature, rainfall, and pesticide use, we aim to develop a model that can accurately predict crop yields and help farmers make informed decisions about their crops.

In addition, we also aim to build an app that will allow farmers to easily input their data and receive a prediction of their expected yield. This app will provide a user-friendly interface and be easy for farmers to use, which will make our model more accessible. By providing farmers with access to accurate crop yield predictions, we hope to contribute to the goal of improving food security and promoting sustainable agriculture practices.

1.4. Research questions

1. What are the most important factors that impact crop yields and how can these factors be incorporated into a machine learning model for predicting crop yields?

2. How accurate are different machine learning algorithms at predicting crop yields and which algorithms are the most reliable?

3. How can we ensure that the data used to train and test machine learning models for crop yield prediction is high-quality and free of biases?

4. How can we design an app that is user-friendly and easy for farmers to use?

5. How can we evaluate the accuracy and reliability of the machine learning models and app developed in this project and what measures can be taken to improve their performance?

6. How can the machine learning models and app developed in this project be used to improve food security and promote sustainable agriculture practices?

7. What are the limitations of the machine learning models and app developed in this project and how can these limitations be addressed in future research?

1.5. Research objectives

The research objectives of this project can be classified as prediction and influence, and there are three main ones:

1. To build a machine learning model for predicting crop yields that is accurate and reliable. This will involve collecting and analysing data on temperature, rainfall, location and pesticide use, and comparing the performance of different machine learning algorithms in order to identify the one that is most accurate and reliable.

2. To build an app that allows farmers to input data and receive a prediction of their expected crop yield. The app should be user-friendly and easy for farmers to use, and should provide a convenient and accessible way for farmers to access crop yield predictions.

3. To understand the factors that impact the accuracy and reliability of machine learning algorithms in predicting crop yields. This will involve examining the impact of temperature, rainfall, location and pesticide use on the accuracy of the models, and identifying ways to account for these factors in order to improve the accuracy and reliability of the predictions.

Overall, the research objectives of this project are to build a machine learning model and app that can help farmers make informed decisions about their crops and improve food security and sustainable agriculture practices. By providing accurate and reliable predictions of crop yields, we hope to contribute to the goal of improving food security and promoting sustainable agriculture practices.

1.6. Theoretical Framework

One of the theories that can be used to understand the research problem of predicting crop yields using machine learning algorithms is the systems theory. Systems theory is a framework that seeks to understand complex systems by analysing the interactions and relationships between their components. In the context of this research problem, systems theory can be used to understand the factors that impact crop yields and the ways in which these factors interact and influence each other.

According to systems theory, a system is composed of a set of components that interact with each other and with the environment in order to achieve a specific goal. In the context

of agriculture, the components of the system might include the weather, soil conditions, pest infestations, and the crops themselves. The goal of the system is to produce high crop yields in a sustainable way.

Systems theory suggests that the components of a system are interdependent and that changes in one component can affect the other components. In the context of agriculture, this means that, for example, changes in the weather, such as increased rainfall or extreme temperatures, can directly impact the soil conditions and subsequently affect the growth and development of the crops.

One of the key concepts in systems theory is the feedback loop, which refers to the way in which changes in a system can produce feedback that impacts the system itself. In the context of agriculture, feedback loops might include the way in which changes in the weather impact the soil conditions, which in turn impact the growth and development of the crops, which can then impact the overall crop yield.

Using systems theory to understand this research problem, involves analysing the interactions and relationships between the various components of the agricultural system, including weather conditions such as temperature, rainfall, geographical conditions and pesticide use, and the ways in which these components influence each other and impact crop yields. It also involves considering the feedback loops that exist within the system and the ways in which changes in one component can impact the other components. By doing so, we can develop a deeper understanding of the complexity of predicting crop yields and identify the factors that are most important for improving the accuracy and reliability of the crop prediction models.

1.7. Literature Review

Agricultural technologies, especially those driven by AI, are revolutionizing the way we approach food production. For this literature review, we performed a thorough exploration of the rising trends in agricultural technology[13][15][20][28][30][31][62][67].

- Precision agriculture. Precision agriculture is a trend that utilizes technologies such as sensors, drones, and AI to optimize farming practices. By collecting data and analyzing it with AI algorithms, farmers can make informed decisions about crop management, leading to maximized yields and reduced resource waste. However, as many have pointed out, the implementation of precision agriculture requires significant investment in technology and training, which can be a barrier for small-scale farmers. Additionally, reliance on technology and data may create a dependency that could pose challenges in the event of technical failures or data inaccuracies. [2]
- Vertical Farming. Vertical farming is another important trend that relies on AI for monitoring and optimization. In vertical farming, crops are grown in controlled environments, often indoors, in vertical stacks. This approach presents advantages such as efficient land and resource use, reduced dependence on traditional agricultural land, and the ability to grow crops in urban areas. It also allows for year-round production and reduced exposure to pests and diseases. However, vertical farming requires substantial initial investment in infrastructure and energy for lighting and climate control. Scaling up vertical farming operations may also face regulatory challenges and consumer acceptance of produce grown in indoor environments.

Either way, vertical farming is predicted to see increased adoption as a sustainable solution for food production.

- Smart Irrigation. Smart irrigation systems are transforming water management in agriculture. These systems employ AI and sensor technologies to optimize water usage by delivering the right amount of water to crops based on their needs. By reducing water waste and improving efficiency, smart irrigation contributes to sustainable agriculture. Smart irrigation systems contribute to water conservation and improved crop water management, resulting in resource savings and increased water-use efficiency. By providing crops with the right amount of water at the right time, farmers can minimize water waste and mitigate the risks of drought. Nevertheless, the cost of implementing smart irrigation systems can be a barrier for farmers with limited financial resources. It's been said that the reliance on technology and accurate data for irrigation decisions may also pose challenges in regions with limited connectivity or unreliable data networks [23].
- Crop Monitoring and Disease Detection. Crop monitoring and disease detection are critical for ensuring crop health and minimizing yield losses. AI-based image recognition and data analytics are employed to monitor crops, detect diseases, and provide timely interventions. These techniques offer the advantage of early identification and targeted interventions, reducing yield losses and the use of chemical treatments. By using AI-based image recognition and data analytics, farmers can efficiently monitor large areas and detect diseases with precision. However, the effectiveness of disease detection algorithms relies on accurate training data, and the availability of such data may be limited for certain crop-disease combinations. Integration of AI technologies into existing farming practices may also require training and adjustment periods, but researchers are confident that the role of AI in crop monitoring and disease detection will only grow, leading to improved resilience and higher productivity[29].
- Autonomous Farming. Autonomous farming is a trend that combines AI, robotics, and IoT devices to automate various farming tasks. From planting and harvesting to monitoring and data collection, autonomous systems can revolutionize farming operations. Autonomous farming provides benefits such as increased efficiency, reduced labor costs, and improved accuracy in farming operations. By automating tasks like planting, harvesting, and monitoring, farmers can save time and focus on higher-level decision-making. However, the adoption of autonomous systems requires significant capital investment and technical expertise. Lastly, concerns over job displacement and the potential impact on rural communities have been raised, especially as the levels of automation and autonomy grow rapidly.
- Genetic Modification. Gene editing technologies like CRISPR offer precise modification of crop DNA, allowing for enhancements in traits such as yield, resilience, and nutritional content. This trend in agricultural biotechnology enables the development of crops with improved characteristics, contributing to increased food production and resilience. Gene editing technologies hold promise for developing crops with enhanced traits, such as higher yields and improved resilience. However, the use of gene editing raises ethical and regulatory concerns. Public acceptance, transparency, and proper regulation are crucial to ensure the responsible and safe use of gene editing in agriculture, but researchers are

confident that the ongoing advancement of gene editing techniques will continue to drive further progress in crop enhancement.

- Food Cultivation. Cultivated meat and seafood production is gaining traction as an innovative solution to traditional farming practices. By growing animal cells in a lab setting, this approach reduces the need for raising livestock, resulting in more sustainable and resource-efficient food production. This type of alternative production offers potential benefits such as reduced environmental impact, improved animal welfare, and more efficient resource utilization. It has the potential to address the many challenges and issues associated with traditional livestock farming. However, the scalability and cost-effectiveness of cultivated meat and seafood production are still being developed. The acceptance and perception of cultured products by consumers and regulatory bodies also need to be addressed, especially as the cultured meat and seafood industry is predicted to grow significantly in the future, driven by advances in biotechnology.
- Biological Pest Control. Biological pest control is an alternative approach to traditional pesticide use that many are hailing as the safer, less technologically-driven option. By utilizing natural predators, parasites, and pathogens, farmers can control pests in a more environmentally friendly manner. Biological pest control methods provide an environmentally friendly alternative to synthetic pesticides, reducing chemical use and potential harm to ecosystems and human health. These methods can be effective in managing pests and diseases sustainably. However, the success of biological pest control relies on understanding ecological dynamics, and it may not be applicable in all farming contexts or for all pests. Proper implementation and monitoring are necessary to ensure the desired results.
- Blockchain. Blockchain technology is finding its place in the agricultural supply chain. By providing transparency and traceability, blockchain ensures the authenticity and quality of products. This technology holds the potential to enhance transparency, traceability, and trust in the agricultural supply chain and improve food safety. However, the adoption of blockchain in agriculture requires collaboration and standardization across the industry. As blockchain adoption in supply chain management expands, we can expect increased traceability and improved consumer confidence in agricultural products.
- Predictive Modeling. Data analytics and predictive modeling are transforming agriculture by leveraging AI and machine learning to analyze vast amounts of data collected from various sources. By utilizing historical and real-time data, farmers can make data-driven decisions and predictions, optimizing their farming practices and improving productivity. However, the successful implementation of data analytics requires access to reliable and accurate data, as well as the necessary infrastructure and technical skills. Data privacy and security concerns must also be addressed to ensure the responsible use of data in agriculture. Once these concerns are address, the consensus is that predictive modeling will further empower farmers with actionable insights for efficient resource management and decision-making.

1.8. Justification

Sustainable Development Goal (SDG) 2 aims to "End hunger, achieve food security and improved nutrition, and promote sustainable agriculture", while SDG 12 aims to "Ensure

sustainable consumption and production patterns". This project can significantly contribute to achieving these goals[59].

One of the main benefits of this study is that it can improve food security by providing farmers with access to accurate crop yield predictions. Accurate crop yield predictions can help farmers make informed decisions about what to plant and how to manage their crops, which can lead to higher yields and more sustainable agriculture practices. By providing farmers with access to this information, they can better plan and manage their crops, which can lead to increased food production and improved food security.

Another benefit of this study is that it can promote sustainable agriculture practices. Sustainable agriculture is an approach to farming that aims to meet the needs of present and future generations by preserving the natural resources that support agriculture. Accurate crop yield predictions can help farmers make decisions about what crops to plant, how to manage their land and water resources, and how to control pests and diseases. By using these predictions, farmers can reduce the use of chemical pesticides and fertilizers, which can help to preserve the environment and promote sustainable agriculture practices.

Additionally, study also aims to build an app that will allow farmers to easily input data and receive a prediction of their expected yield. This app will provide a user-friendly interface and be easy for farmers to use, which will make it more accessible and widely adopted. By providing farmers with access to accurate crop yield predictions, we hope to contribute to the goal of improving food security and promoting sustainable agriculture practices.

The project also contributes to SDG 12 by providing farmers with accurate crop yield predictions, they can make better decisions about what crops to plant and how to manage them. This can lead to higher yields and more sustainable agriculture practices, which can ultimately lead to reducing the use of chemical pesticides and fertilizers, and preserving the environment.

In conclusion, by providing farmers with access to accurate crop yield predictions, they can better plan and manage their crops, which can lead to increased food production and improved food security, and promoting sustainable agriculture practices. The app will make the predictions more accessible and widely adopted, thus, leading to more sustainable consumption and production patterns. Ultimately, this project can contribute to improving food security, promoting sustainable agriculture practices, and preserving the environment for future generations.

2. Methodology

Hypothesis, Independent and Dependent Variables

In this project, our hypothesis is that by using data on temperature, rainfall, location and pesticide use, we can develop a machine learning model that can accurately predict crop yields.

The independent variables in this project would be the factors that can affect crop yields such as weather conditions, soil properties, and pest infestations. These are the variables that we will use as input to train the machine learning model.

The dependent variable in this project is the crop yield. This is the variable that we want to predict based on the independent variables. The model will use the information on weather, soil, and pests to make predictions about the crop yield.

Study Design

The purpose of this project is to build a machine learning model for predicting crop yields in order to improve food security and sustainable agriculture practices. To achieve this goal, we will use the *Crop Yield Prediction Dataset* provided by Rishi Patel [14] and compare the performance of multiple machine learning algorithms using the Jupyter Notebook (Python) framework. Finally, we will build an app to allow farmers to input data and get a prediction of their expected yield.

To build the machine learning model for crop yield prediction [9], we will follow the following steps:

1. Data collection and preparation: The first step in building the machine learning model will be to collect and prepare the data for analysis. The Crop Yield Prediction Dataset includes data on a variety of factors that may impact crop yields, including weather data, and data on pesticide use. We will use this data to train and test our machine learning models.

2. Data exploration and visualization: Before building the machine learning model, we will explore and visualize the data to better understand the relationships between the different variables and identify any patterns or trends. This will involve using tools like scatter plots and histograms to visualize the data and identify any potential outliers or anomalies.

3. Data pre-processing and feature selection: After exploring the data, we will perform the necessary pre-processing and select the most relevant features to include in the machine learning model. This will involve normalization, addressing class imbalance and feature selection.

4. Model development: Once we have selected the relevant features, we will use the Jupyter Notebook framework to build and test multiple machine learning models using different algorithms. This will involve splitting the data into training and testing sets, building the models using the training data, and evaluating their performance using the testing data. We will compare the performance of different algorithms, such as decision trees, random forests, and neural networks, to identify the model that performs the best.

5. Model evaluation and improvement: After identifying the best-performing machine learning model, we will evaluate its performance in more detail and make any necessary improvements. This may involve adjusting the model's hyperparameters, adding or removing features, or using more data to train the model.

6. App prototype development: Once we have built and optimized the machine learning model, we will build an app prototype using the React Native framework to allow farmers to input data and get a prediction of their expected yield. The initial interface will be designed using Canva then later implemented in a React Native framework. The prototype will include a user-friendly interface that allows farmers to easily input data and receive a prediction of their expected yield.

Frameworks and Libraries

Frameworks

- Jupyter Notebook[41] is an open-source framework that allows users to create and share documents that contain live code cells, equations, visualizations, and narrative text. It supports several programming languages, including Python, the language used in this project. This framework was used for the Data Analysis, as well as the model comparison and optimization, as it allows for interactive data analysis.
- Canva is a graphic design tool that provides a range of templates, design elements, and customization options, making it accessible for designing intuitive and visually appealing interfaces.
- React Native[45] is an open-source framework developed by Facebook that enables developers to build native mobile applications for iOS and Android platforms using JavaScript and React principles. It works well for integrating small machine learning models, which is why it was chosen for the development of our app.

Libraries

The following Python libraries were used in this project for data manipulation, visualization, numerical computations, machine learning, and interactive visualizations:

- Pandas[35] is a data manipulation library that provides tools for data cleaning, exploration, and analysis. It was used to read the data and to prepare it for the machine learning models.
- Matplotlib[32] is a data visualization library that provides a range of plotting functions and tools. It was used to create and visualize graphs, in order to explore and analyze the data.
- Seaborn[50] is another data visualization library that provides more advanced visualization techniques and styles It was also used to create and visualize graphs, in order to explore and analyze the data.
- Numpy[34] is a fundamental library for scientific computing in Python. It provides support for multi-dimensional arrays and matrices, as well as a range of mathematical functions. It was used to perform certain numerical computations and linear algebra operations.

- Scikit-learn[49] (sklearn) is a machine learning library that provides a range of algorithms for classification, regression, clustering, and dimensionality reduction. It also provides tools for data preprocessing, model selection, and evaluation. It was used to build and train the different machine learning models, and compare them.
- Xgboost[11] is a popular open-source gradient boosting library that provides an implementation of gradient boosting algorithms.
- Plotly[40] is a data visualization library that provides interactive visualization tools. It was used to create interactive visualizations during the Data Analysis phase.

2.1. Data Collection and Preparation

The quality and relevance of the dataset used for analysis play a critical role in determining the performance of the machine learning model. In this section, we delve into the crucial initial step of collecting and preparing the data necessary for building an accurate crop yield prediction model.

Sampling Procedures

For this study, we will be using the Crop Yield Prediction Dataset from Kaggle [14], which is a widely recognized platform for data science competitions and datasets. The dataset can be considered a secondary data source and contains information on various factors that can affect crop yields, such as weather conditions, soil properties, and pest infestations. The data is collected from various locations around the world and covers a wide range of crops.

In terms of sampling procedures, we will be using a random sampling method to select a representative sample of the data from the dataset. This will ensure that the sample is representative of the population and that the results of the study can be generalized to the larger population. We will also ensure that the sample is diverse and includes data from different locations, crops, and weather conditions to increase the robustness of the model.

We will also split the data into a training set and a testing set to evaluate the performance of the model. The training set will be used to train the model, while the testing set will be used to evaluate the model's performance and accuracy. This will help us to ensure that the model is able to accurately predict crop yields on new, unseen data.

2.2. Data Exploration and Visualization

In order to build an effective machine learning model, it is essential to carry out a comprehensive exploration and visualization of the data, which is outlined in this section. By utilizing tools such as scatter plots and histograms, we aim to gain valuable insights, identify outliers, and detect anomalies within the dataset[38].

An Exploratory Data Analysis (EDA)[65] will be performed, which consists of summarizing and visualizing the data in order to attempt to understand the underlying patterns and relationships. This process consists of four main steps:

Step 1: Loading the Dataset and Initial Exploration

The first few rows of the dataset are displayed to get a glimpse of the data structure and to verify that it is loaded correctly.

	Area	Item	Year	hg/ha_yield	average_rain_fall_mm_per_year	pesticides_tonnes	avg_temp
0	Albania	Maize	1990	36613	1485.0	121.0	16.37
1	Albania	Potatoes	1990	66667	1485.0	121.0	16.37
2	Albania	Rice, paddy	1990	23333	1485.0	121.0	16.37
3	Albania	Sorghum	1990	12500	1485.0	121.0	16.37
4	Albania	Soybeans	1990	7000	1485.0	121.0	16.37

Figure 1. First 5 rows of dataset displayed as a Pandas' DataFrame.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 28242 entries, 0 to 28241
Data columns (total 7 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Area                   28242 non-null  object
 1   Crop                   28242 non-null  object
 2   Year                   28242 non-null  int64
 3   Crop Yield (hg/ha)     28242 non-null  int64
 4   Rainfall (mm/year)     28242 non-null  float64
 5   Pesticides (tonnes)    28242 non-null  float64
 6   Temperature (Celsius)  28242 non-null  float64
dtypes: float64(3), int64(2), object(2)
memory usage: 1.7+ MB
```

Figure 2. Content breakdown of the dataset.

As can be seen (Figures 1 and 2), dataset with 28,242 entries and 7 columns. Here is a breakdown of the information provided about each column:

- Area: This column contains categorical data (object type) representing the country or area where the crop yield data is recorded. It provides a geographical context to analyze and compare crop yields across different regions.
- Crop: This column also contains categorical data (object type) indicating different types of crops being measured. It allows for the analysis and prediction of specific crops individually.
- Year: This column represents the year in which the crop yield data is recorded as an integer (int64 type). It provides a temporal dimension, allowing for the analysis of yield trends over time.
- Crop Yield (hg/ha): This column contains numerical data as an integer (int64 type) representing the yield measured in hectograms per hectare. It serves as the target variable for regression modeling.
- Rainfall (mm/year): This column contains numerical data as a floating-point number (float64 type) representing the annual rainfall measured in millimeters. It can provide significant value on the impact of weather conditions on crop yield.
- Pesticides (tonnes): This column contains numerical data as a floating-point number (float64 type) representing the amount of pesticides used measured in tonnes. It can provide insight on the impact of plant health on crop yield.

- Temperature (Celsius): This column contains numerical data as a floating-point number (float64 type) representing the temperature measured in Celsius. It can also provide significant value on the impact of weather conditions on crop yield.

Step 2: Summary Statistics

The summary statistics table generated provide an overview of the dataset's characteristics. The dataset contains 28,242 observations for each variable: Year, Crop Yield (hg/ha), Rainfall (mm/year), Pesticides (tonnes), and Temperature (Celsius).

	count	mean	std	min	25%	50%	75%	max
Year	28242.0	2001.544296	7.051905	1990.00	1995.0000	2001.00	2008.00	2013.00
Crop Yield (hg/ha)	28242.0	77053.332094	84956.612897	50.00	19919.2500	38295.00	104676.75	501412.00
Rainfall (mm/year)	28242.0	1149.055980	709.812150	51.00	593.0000	1083.00	1668.00	3240.00
Pesticides (tonnes)	28242.0	37076.909344	59958.784665	0.04	1702.0000	17529.44	48687.88	367778.00
Temperature (Celsius)	28242.0	20.542627	6.312051	1.30	16.7025	21.51	26.00	30.65

Figure 3. Summary Statistics table for the dataset. Each column represents a feature in the dataset and each row corresponds to one of the following statistics metrics: count, mean, standard deviation, minimum, 25th percentile, median, 75th percentile, maximum values.

- Year. The dataset spans from 1990 to 2013, with a total count of 28,242 observations. The mean year is around 2001.5, indicating a relatively balanced distribution of data over the years. The standard deviation of 7.05 suggests a moderate spread of data around the mean.
- Crop Yield (hg/ha): The crop yield variable has a mean of approximately 77,053 hg/ha. The standard deviation of 84,956.61 indicates a substantial variation in crop yield across the dataset. The minimum and maximum values represent the lowest and highest crop yields observed, with 50 and 501,412 hg/ha, respectively. The dataset's quartiles reveal that 25% of the observations have a crop yield below 19,919.25 hg/ha, while 75% have a yield below 104,676.75 hg/ha.
- Rainfall (mm/year): The average annual rainfall in the dataset is approximately 1,149.06 mm. The standard deviation of 709.81 suggests a moderate variation in rainfall levels. The minimum and maximum values represent the lowest and highest observed rainfall, with 51 and 3,240 mm/year, respectively. The quartiles show that 25% of the observations have rainfall levels below 593 mm/year, while 75% have levels below 1,668 mm/year.
- Pesticides (tonnes): The mean pesticide usage is around 37,076.91 tonnes. The standard deviation of 59,958.78 indicates a substantial variation in pesticide usage across the dataset. The minimum and maximum values represent the lowest and highest levels of pesticide usage, with 0.04 and 367,778 tonnes, respectively. The quartiles reveal that 25% of the observations have pesticide usage below 1,702 tonnes, while 75% have usage below 48,687.88 tonnes.
- Temperature (Celsius): The average temperature in the dataset is approximately 20.54 degrees Celsius. The standard deviation of 6.31 indicates a moderate variation in temperature levels. The minimum and maximum values represent the lowest and highest observed temperatures, with 1.3 and 30.65 degrees Celsius, respectively. The quartiles show that 25% of the observations have temperatures

below 16.70 degrees Celsius, while 75% have temperatures below 26.00 degrees Celsius.

Step 3: Data Cleaning

If required, this step involves converting data types or handling missing values. In our case, the were no missing values (*nan*) and so no data cleaning was required.

```
# Shows NA or NAN
dataset.isna().any()
```

✓ 0.0s

```
Area                            False
Crop                            False
Year                            False
Crop Yield (hg/ha)              False
Rainfall (mm/year)              False
Pesticides (tonnes)             False
Temperature (Celsius)           False
dtype: bool
```

Figure 4. Missing values (nan) in the dataset.

Step 4: Visualization of the Data

Various plots and graphs are generated, to help identify patterns, trends, and relationships within the data.

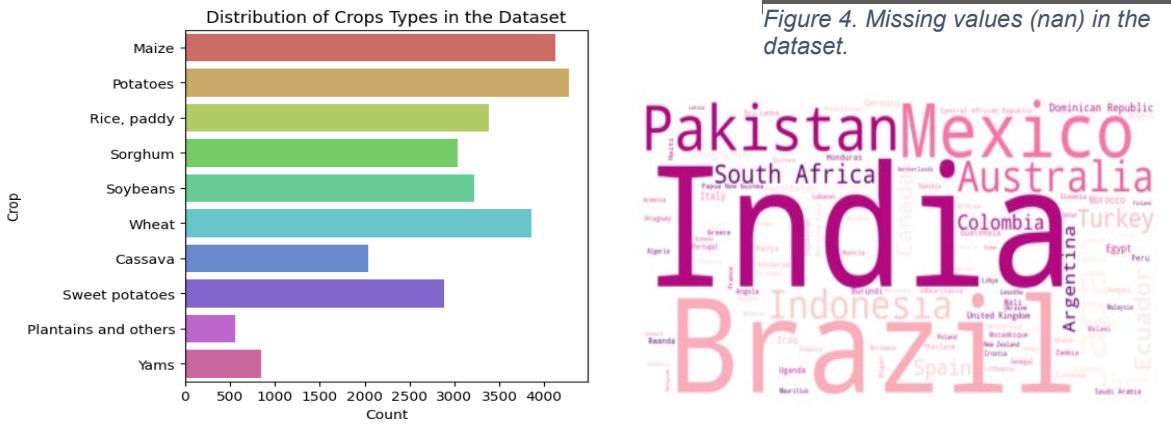


Figure 5. Distribution of categorical features Area and Crop in the dataset. (A) Count plot illustrating the distribution of crop types in the dataset. Each bar represents a specific crop, and the height of the bar corresponds to the count or frequency of that crop in the dataset. (B) Word cloud visualization based on the frequency of different areas in the dataset. The size of each area name in the word cloud corresponds to its frequency, with larger text indicating a higher count.

The distribution of the categorical feature "Crops" in the dataset is displayed in Figure 5A. By examining the countplot, we can observe the distribution of crop types and identify some imbalances or biases in the dataset. Potatoes have the highest count with 4,276 occurrences, followed closely by Maize with 4,121 occurrences. Wheat is the third most frequent crop with 3,857 occurrences, while Rice, paddy ranks fourth with 3,388 occurrences. Soybeans and Sorghum also have considerable counts, appearing 3,223 and 3,039 times, respectively. Sweet potatoes have a count of 2,890, while Cassava appears 2,045 times in the dataset. Yams and Plantains and others have lower counts, with 847 and 556 occurrences, respectively.

The varying frequencies of different crop types suggest that certain crops are more prevalent in the dataset, while others are relatively less represented. This imbalance in the distribution of crop types may introduce bias in the model's predictions. Crops with higher frequencies, such as Potatoes, Maize, and Wheat, may have more reliable and abundant data, allowing the model to learn and generalize better for these crops. On the other hand, crops with lower frequencies, such as Yams or Plantains and others, may have fewer observations, potentially leading to limited representation and less accurate predictions for these specific crop types.

To mitigate the impact of the imbalanced distribution of crop types, several approaches can be employed. In this case, we will employ advanced modeling techniques that are more robust to imbalanced data, such as ensemble methods (i.e. Random Forest Regressor and Gradient Boosting Regressor), which can help address the potential challenges posed by the uneven distribution of crop types.

The distribution of the categorical feature "Area" in the dataset is displayed in a Wordcloud in Figure 5B and in a Choropleth Map in Figure 6. Both representations provide insights into the prominence of various areas in the dataset, which offers valuable information for understanding agricultural patterns and potential impacts on crop yield.

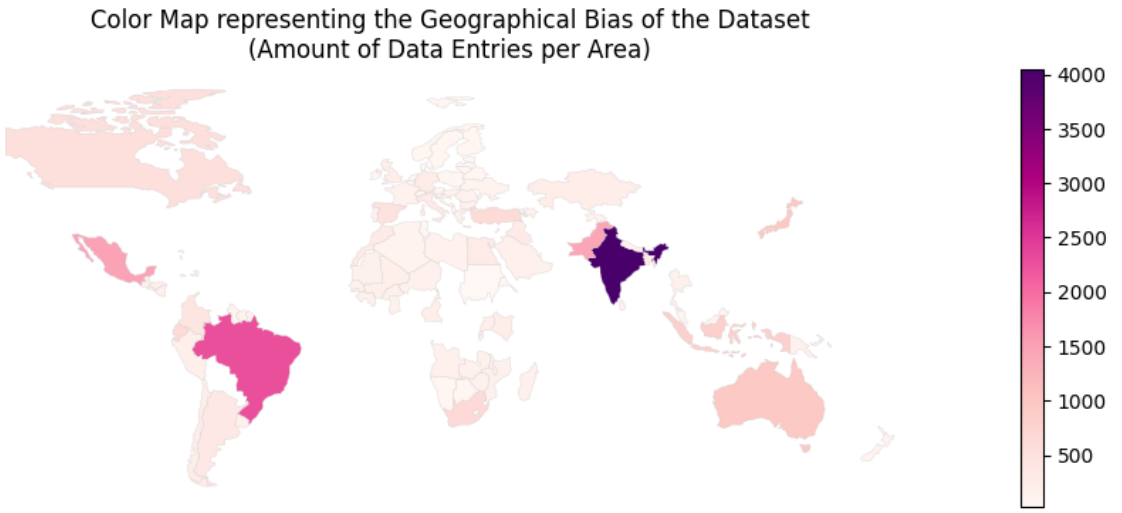


Figure 6. Representation of geographical bias in the dataset via a choropleth map. The map shows the distribution of data entries per area or country. The color intensity represents the amount of data entries, with darker shades indicating a higher number of entries. The legend provides information about the range of data counts and the corresponding color scale.

By examining the map in Figure 6, we can observe the concentration of data entries in certain regions. Some areas, such as India (4,048), Brazil (2,277), Mexico (1,472), Pakistan (1,449), Japan, and Australia, have a significant number of data entries, as represented by the darker shades. These areas likely contribute to the overall patterns and insights derived from the dataset. In contrast, there are areas with relatively fewer data entries, indicated by lighter shades. The distribution of data entries across different areas can introduce geographic bias in the dataset, as certain regions may be overrepresented or underrepresented compared to others.

The presence of bias can be attributed to several factors. For example, it could be reflective of the actual agricultural landscape, where certain countries have a larger share of global crop production. India and Brazil, in particular, are known for their extensive agricultural activities and diverse crop cultivation. The bias towards these countries may be a true representation of their significant contributions to overall crop production. Additionally, there might be economic factors and market dynamics contributing to this bias.

To mitigate the impact of this imbalanced distribution of areas, we will be using the same data-balancing approaches as with the crop types feature.

The distribution of numerical variables is best represented by histograms and KDE plots (See Figures 7 and 8).

The histograms provide valuable insights into the distribution of each numerical variable. Understanding the shape, skewness, and presence of outliers helps determine the appropriate regression models and transformation techniques.

To facilitate comparison and interpretation, the numerical variables have been normalized using z-score normalization. This normalization technique transforms the variables to have a mean of 0 and a standard deviation of 1, allowing for a standardized comparison across variables.

The KDE plots also differentiate the distributions based on different crop types. This provides an additional layer of information and allows for the examination of potential differences or patterns in the distribution of variables across crop types.

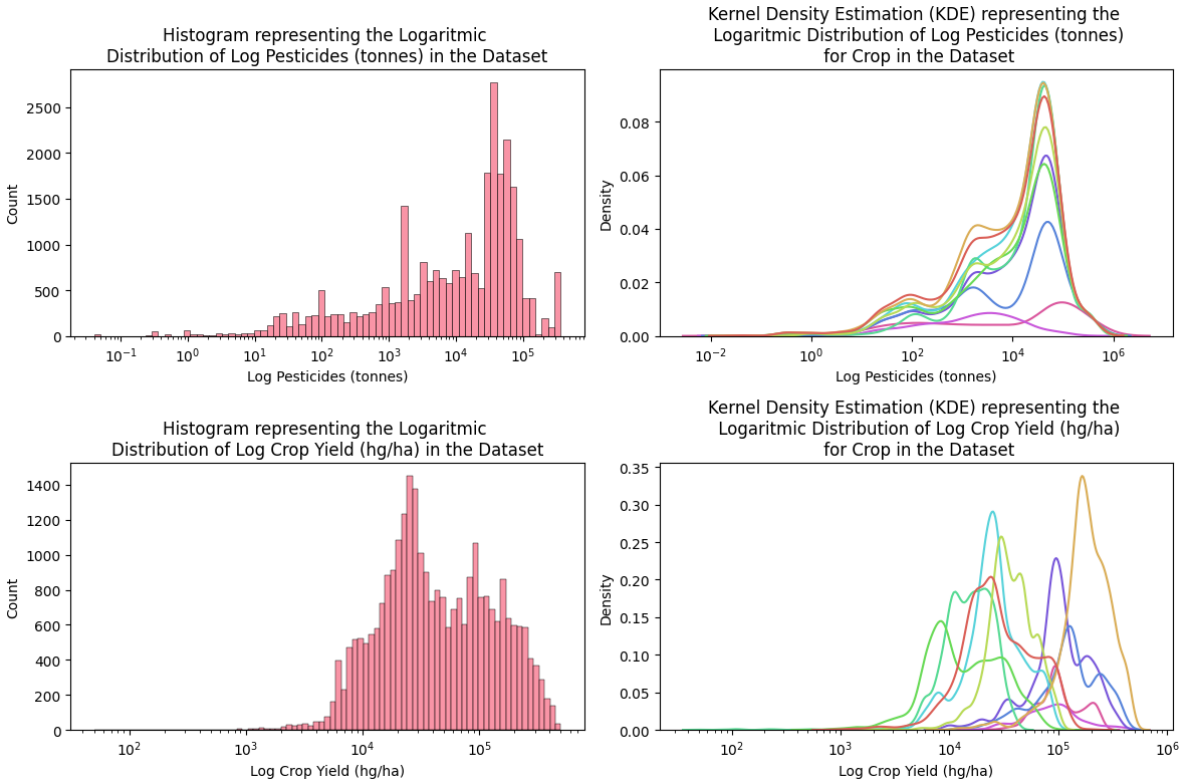


Figure 7. Distribution of features Crop Yield and Pesticides in a logarithmic scale. Represented by histogram (left) and Kernel Density Estimation or KDE (right).

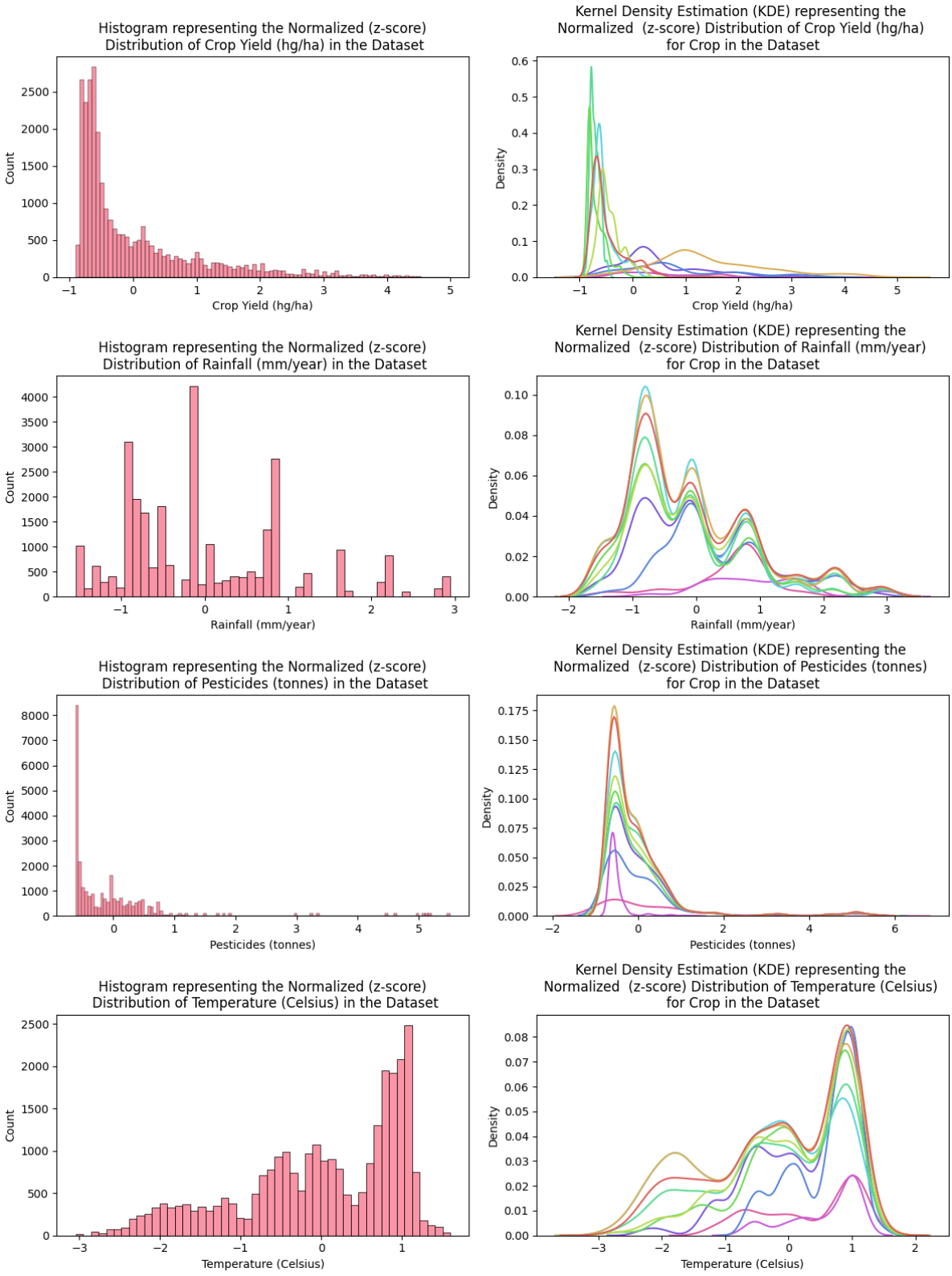


Figure 8. Distribution of normalized (z-score) numerical features in the dataset represented by Histograms (LEFT) and kernel density estimation (KDE) plots (RIGHT). The histograms provide a visual representation of the distribution of each normalized variable. The KDE plots offer a smoothed estimation of the underlying distribution, as well as differentiating the distributions based on different crop types, represented by different colors.

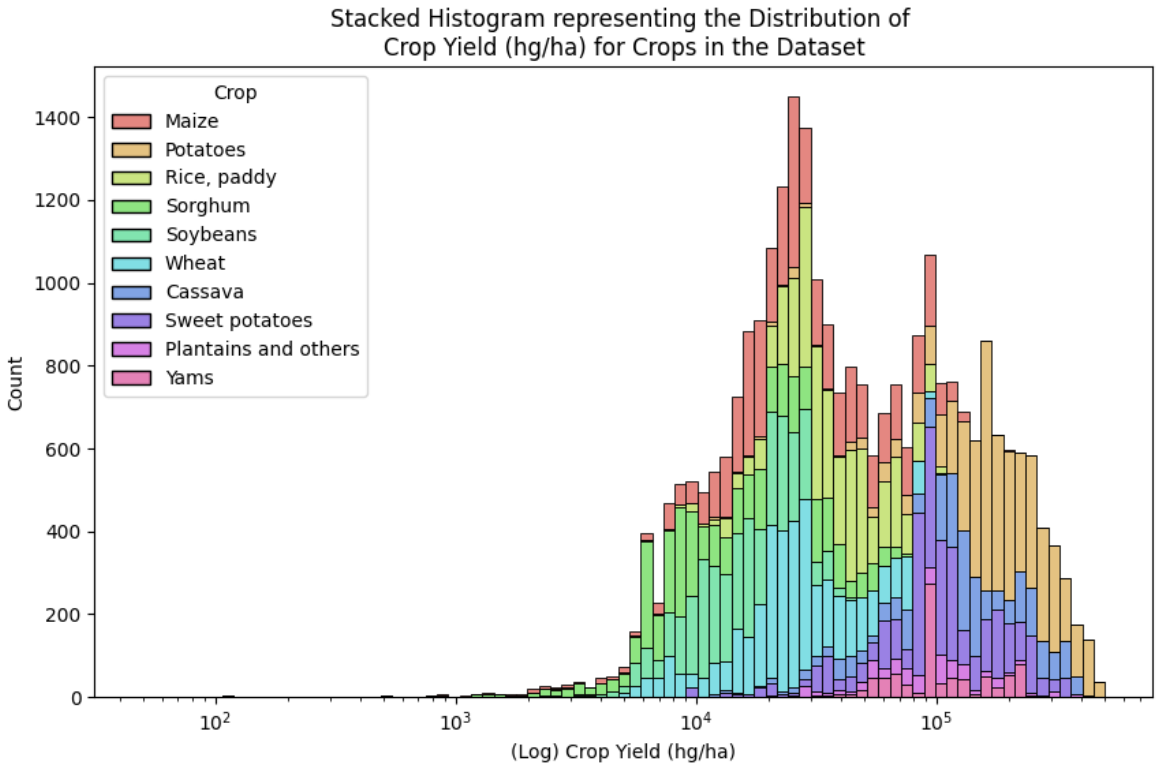


Figure 9. Stacked histogram representing the distribution of "Crop Yield (hg/ha)" for different crop types in the dataset. The histogram is plotted on a logarithmic scale, since the original data spans several orders of magnitude. The stacked format allows for visualizing the contribution of each crop type to the overall distribution.

Figure 9 provides a visual representation of the distribution of crop yields for different crops in the dataset. It highlights variations in yield among different crop types. For example, Potatoes show higher yield, while Maize appears to be the most frequent.

Representing the variables for different crops throughout time can give valuable insight on historical growth rates and time trends.

As can be seen in Figures 10 and 11, Crop Yield (hg/ha) demonstrates significant variations across crops. Maize and Cassava exhibit the highest growth rates, with increases of 70.16% and 53.44% respectively. This suggests that these crops have experienced substantial improvements in yield over the years. On the other hand, Sorghum and Soybeans show relatively lower growth rates, indicating more modest advancements in yield for these crops.

When examining Rainfall (mm/year) growth, several crops demonstrate negative growth rates, implying a decrease in rainfall levels over time. Maize and Potatoes exhibit the most substantial decline, suggesting potential challenges in these crops due to decreased water availability. Conversely, Sweet Potatoes and Yams experience positive growth rates, indicating an increase in rainfall levels, which might be advantageous for their cultivation.

The growth in Pesticides (tonnes) usage demonstrates varying patterns across crops. Yams exhibit the highest growth rate of 272.58%, indicating a substantial increase in pesticide application. Other crops such as Sorghum, Rice, and Sweet Potatoes also show significant growth in pesticide usage. This information highlights the potential dependence of these

crops on pesticides for pest management and the need for sustainable agricultural practices.

Temperature (Celsius) growth provides insights into the impact of changing climatic conditions on crop cultivation. Notably, Yams exhibit the highest growth rate of 3.32%, suggesting a notable rise in temperature over time. Sweet Potatoes and Cassava also experience positive growth rates, indicating a similar trend. These findings emphasize the need to consider temperature changes and their potential implications for crop selection and farming practices.

	Crop Yield (hg/ha)	Rainfall (mm/year)	Pesticides (tonnes)	Temperature (Celsius)
Cassava	53.435269	-0.144771	126.230013	2.024532
Maize	70.162578	-2.881368	73.088973	-0.516946
Plantains and others	2.631679	-1.218515	121.097860	1.168221
Potatoes	35.365157	-3.049699	67.811490	-0.617003
Rice, paddy	37.119535	-1.727324	79.840531	1.139067
Sorghum	23.381925	-3.442016	92.133888	-0.506401
Soybeans	16.308424	-0.833926	75.193956	0.906708
Sweet potatoes	32.688431	2.668179	83.722964	2.760178
Wheat	31.486740	-2.188853	69.348251	-0.726807
Yams	17.704218	2.666552	272.578943	3.316241

Figure 10. Average growth percentage (%) per Crop type through the years for all numerical features in the dataset.

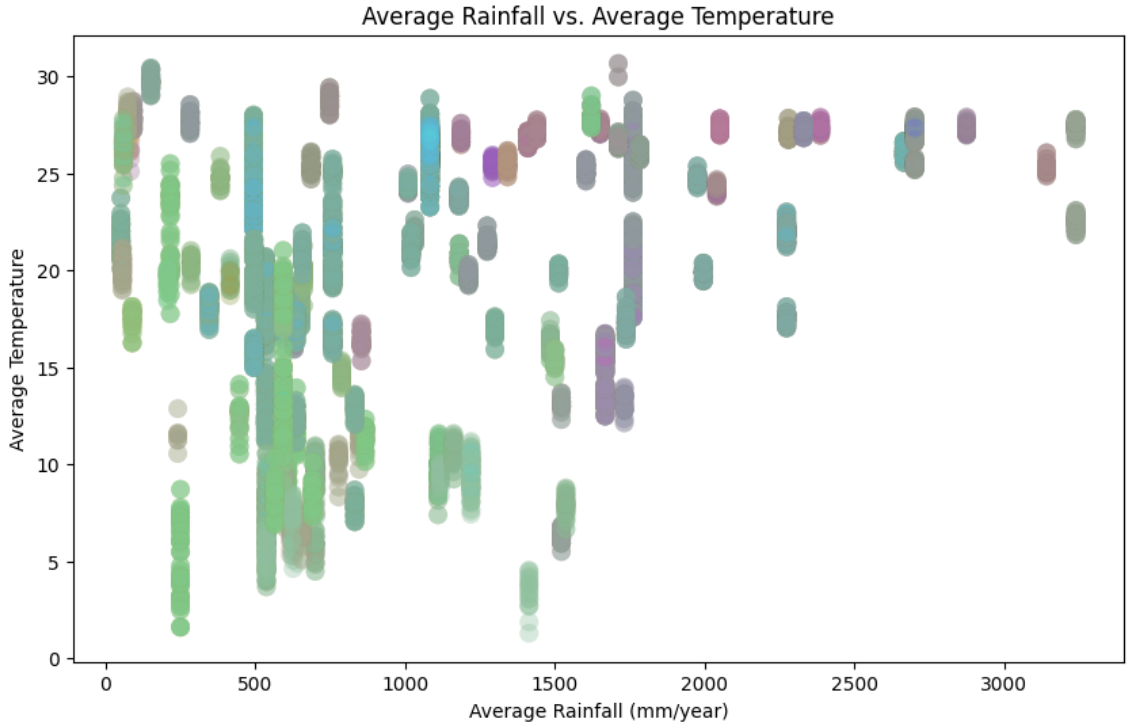


Figure 11. Scatter plot visualizing the relationship between two variables, 'Rainfall (mm/year)' and 'Temperature (Celsius)', in the dataset. Each data point represents the average rainfall and temperature for a specific crop.

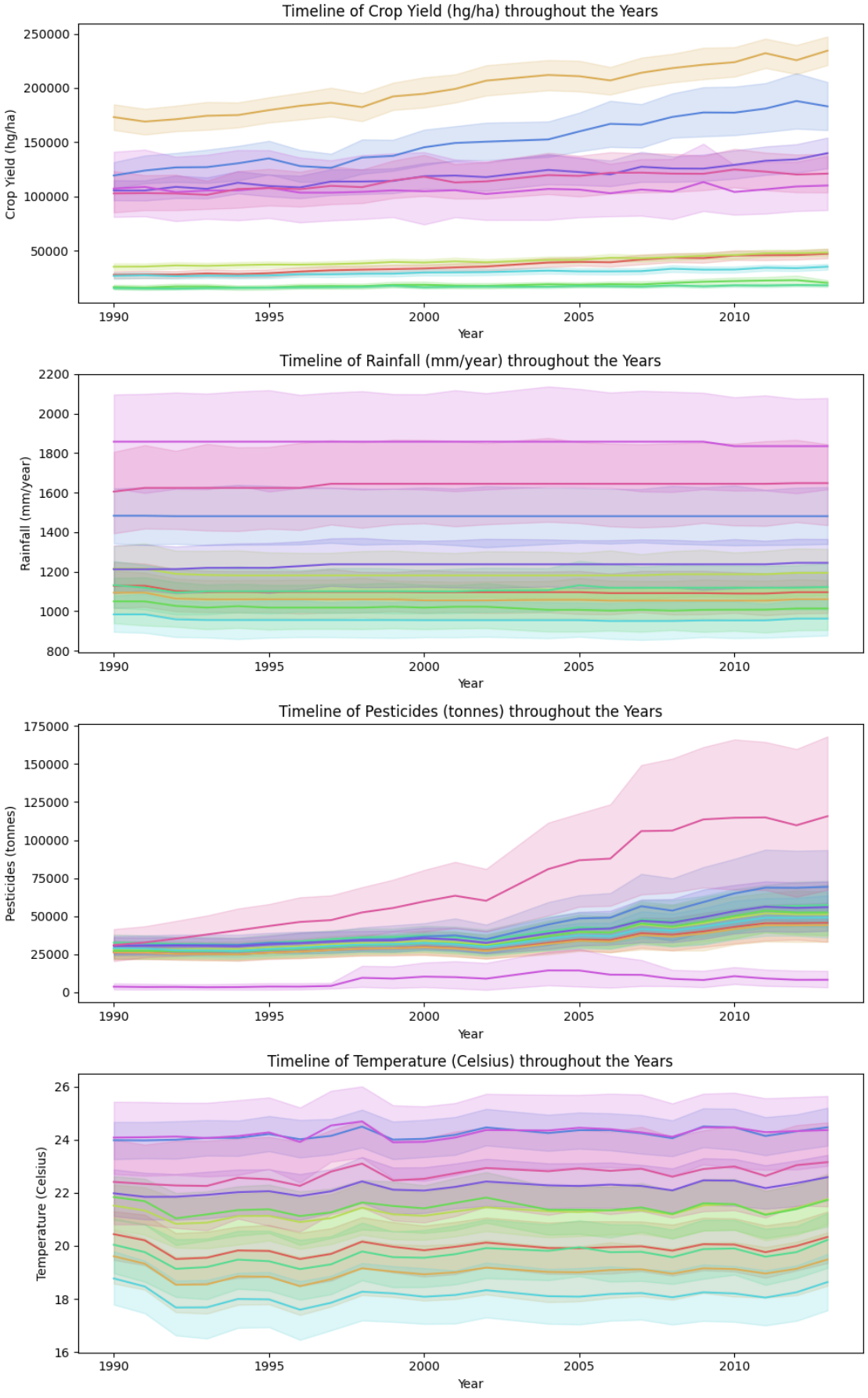


Figure 12. Line plots representing the change in various variables over the years. Each subplot displays the timeline of a specific variable for different crop types.

The representation of Average Rainfall versus Average Precipitations (Figure 12) can help identify any potential patterns or trends between these two variables across different crops.

Understanding the relationship between rainfall and temperature is relevant for regression modeling because these variables can significantly influence crop yield. As seen in the scatterplot, we can gain insights into the nature of the relationship between these variables and identify a potential correlation between them. This will be further examined during feature selection with a correlation matrix.

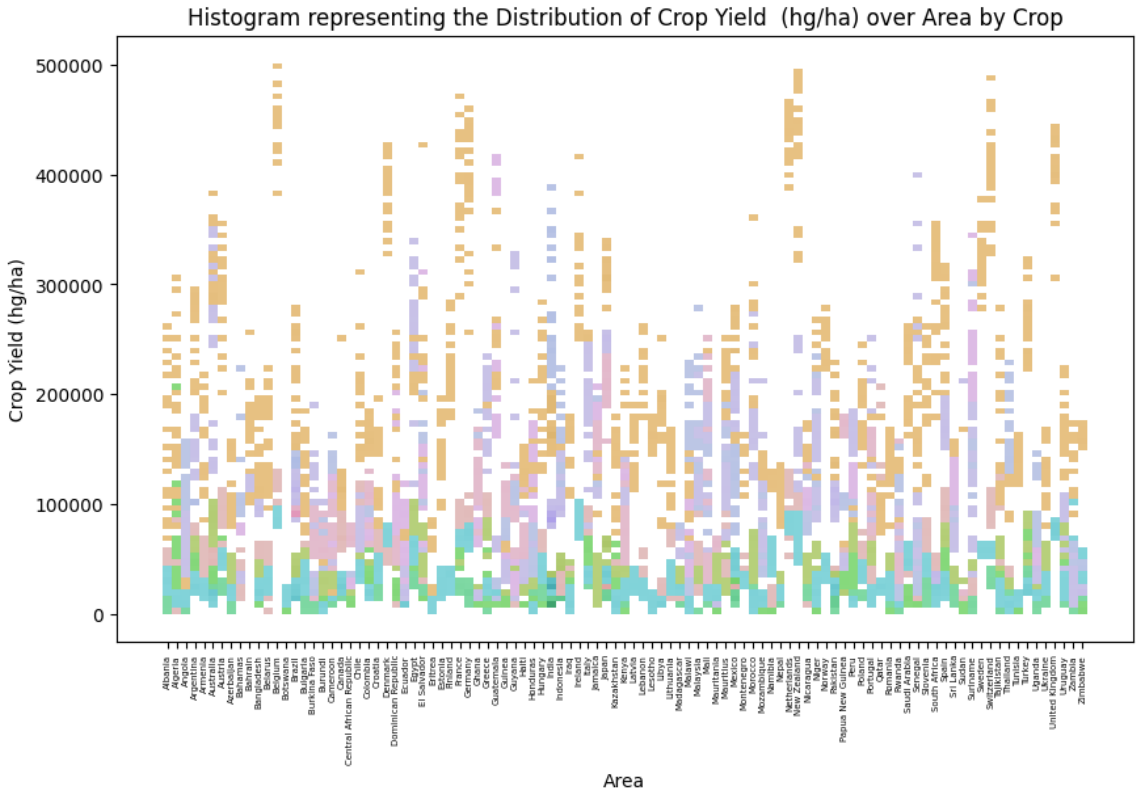


Figure 13. Histogram plot visualizing the distribution of 'Crop Yield (hg/ha)' across different areas in the dataset. Each bar represents the frequency or count of crop yield values for a specific area, with the color encoding representing different crop types. The x-axis denotes the areas, while the y-axis represents the crop yield in hectograms per hectare (hg/ha).

By examining the distribution of 'Crop Yield (hg/ha)' across different areas (Figure 13), we can identify the certain tendencies, such high-yield patterns for Potatoes, and low-yield patterns for Wheat, Sorghum and Soybeans, as well as visualize the spread or dispersion of crop yield values for each area. Additionally, this plot allows for the identification of any outliers or unusual patterns in crop yield distribution, which in this case weren't significant.

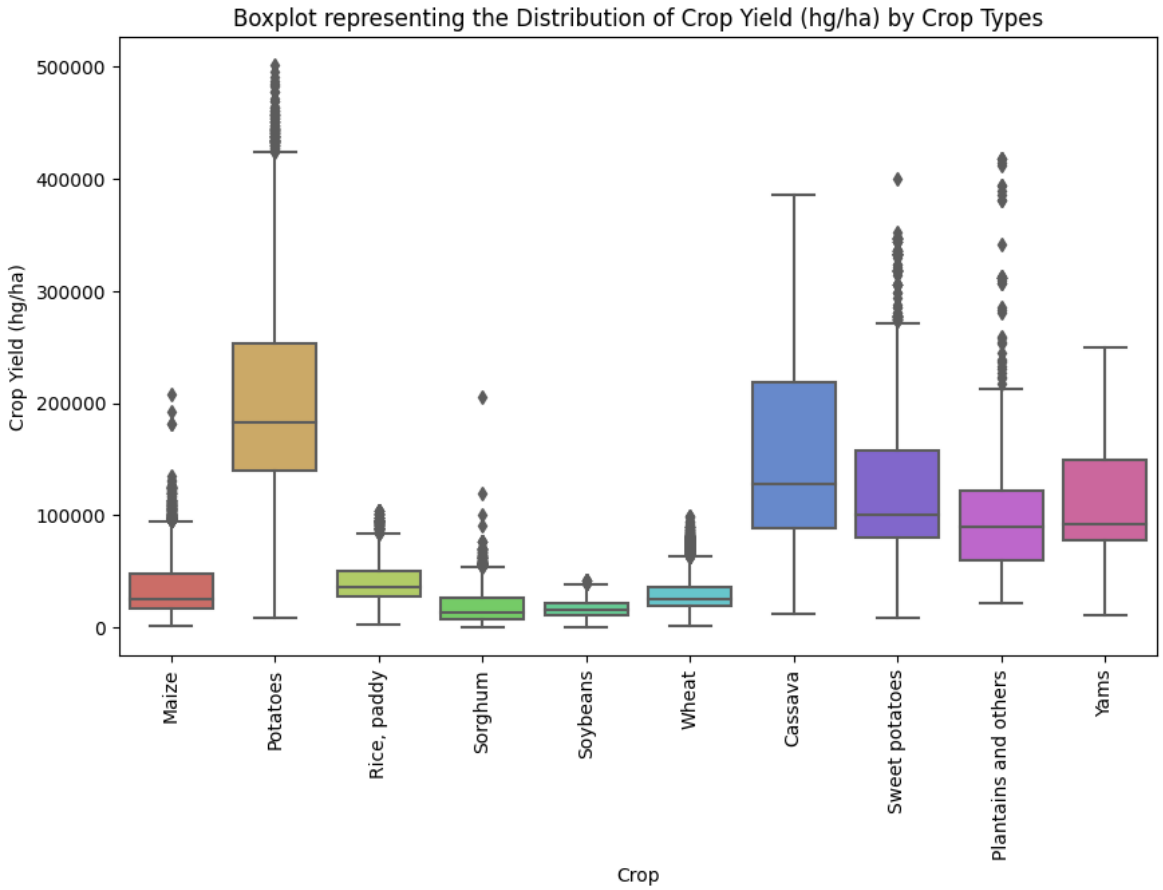


Figure 14. Box plot visualization of the distribution of 'Crop Yield (hg/ha)' for different crop types in the dataset. Each box represents the interquartile range (IQR), which spans from the 25th percentile (lower quartile) to the 75th percentile (upper quartile) of the crop yield values. The line inside the box represents the median, and the whiskers extend to the minimum and maximum values within 1.5 times the IQR.

The box plot (Figure 14) reveals interesting patterns and differences in the distribution of crop yields among the different crop types. Specifically, cassava and potatoes exhibit large boxes, indicating a relatively wide spread of crop yield values. The median line positioned about one-fourth from the bottom suggests that the majority of crop yield values for these crops are concentrated towards the lower end. This indicates that cassava and potatoes might have a higher likelihood of lower crop yields compared to the other crops.

On the other hand, sweet potatoes, plantains and others, and yams are represented by medium-sized boxes, suggesting a moderate spread of crop yield values. The almost-squared shape of these boxes indicates a relatively balanced distribution of values, with the median line positioned closer to the center. This suggests that these crops have a more consistent range of crop yields, with a comparable likelihood of achieving both higher and lower yields.

In contrast, the remaining crops exhibit small boxes, indicating a narrower spread of crop yield values. This suggests that these crops have a more concentrated range of yields, with relatively little variation. The smaller boxes imply that these crops may have a higher likelihood of achieving more consistent or predictable crop yields compared to cassava, potatoes, sweet potatoes, plantains and others, and yams.

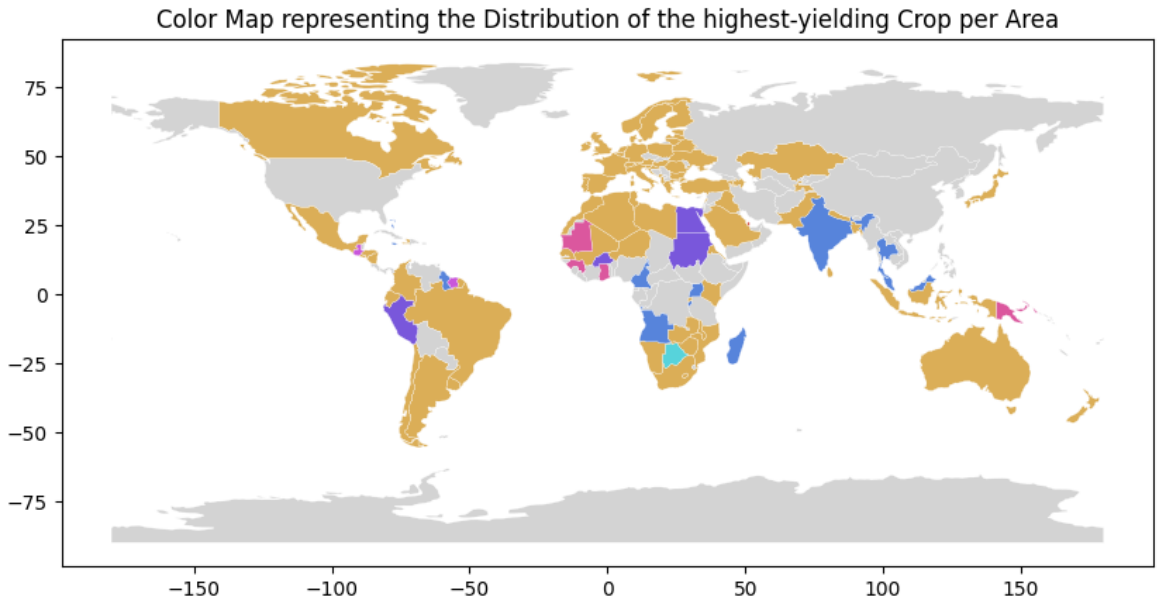


Figure 15. World map visualization representing the distribution of the highest-yielding crop per area. Each area is color-coded based on the corresponding crop.

2.3. Data Preprocessing

The process of developing an effective machine learning model relies in the pre-processing and feature selection stage. This section outlines the process of preparing the data for analysis, specifically addressing categorical features, and then selecting the most relevant features for inclusion in the model.

Feature Selection

One of the main methods for feature analysis is the correlation matrix[33][38]. This matrix (represented as a heatmap in Figure 13) measures the strength and direction of the direct relationship between pairs of numerical variables: Crop Yield (hg/ha), Rainfall (mm/year), Pesticides (tonnes), Temperature (Celsius), and Year.

Looking at the correlation values, it can be observed that the correlation between Crop Yield and Rainfall is very close to zero (0.001). This indicates a weak or negligible linear relationship between these variables. Similarly, Crop Yield has a weak positive correlation (0.065) with Pesticides and a weak negative correlation (-0.115) with Temperature. The correlation between Crop Yield and Year is also weak (0.092), indicating a slight positive relationship.

On the other hand, Rainfall and Temperature show a moderate positive correlation (0.313), suggesting that higher temperatures are associated with higher rainfall. There is also a moderate positive correlation (0.181) between Rainfall and Pesticides, implying that regions with higher rainfall tend to use more pesticides.

The correlation matrix provides valuable insights into the relationships between the variables in the dataset. We will use this data to consider performing feature selection based on the strength of correlation with the target variable (Crop Yield) and the degree of correlation between the input variables themselves.

In terms of the correlation with Crop Yield, we can see that Rainfall has a very weak correlation (close to zero) with Crop Yield. This suggests that Rainfall may not be a strong predictor of Crop Yield and could potentially be excluded from the feature set.

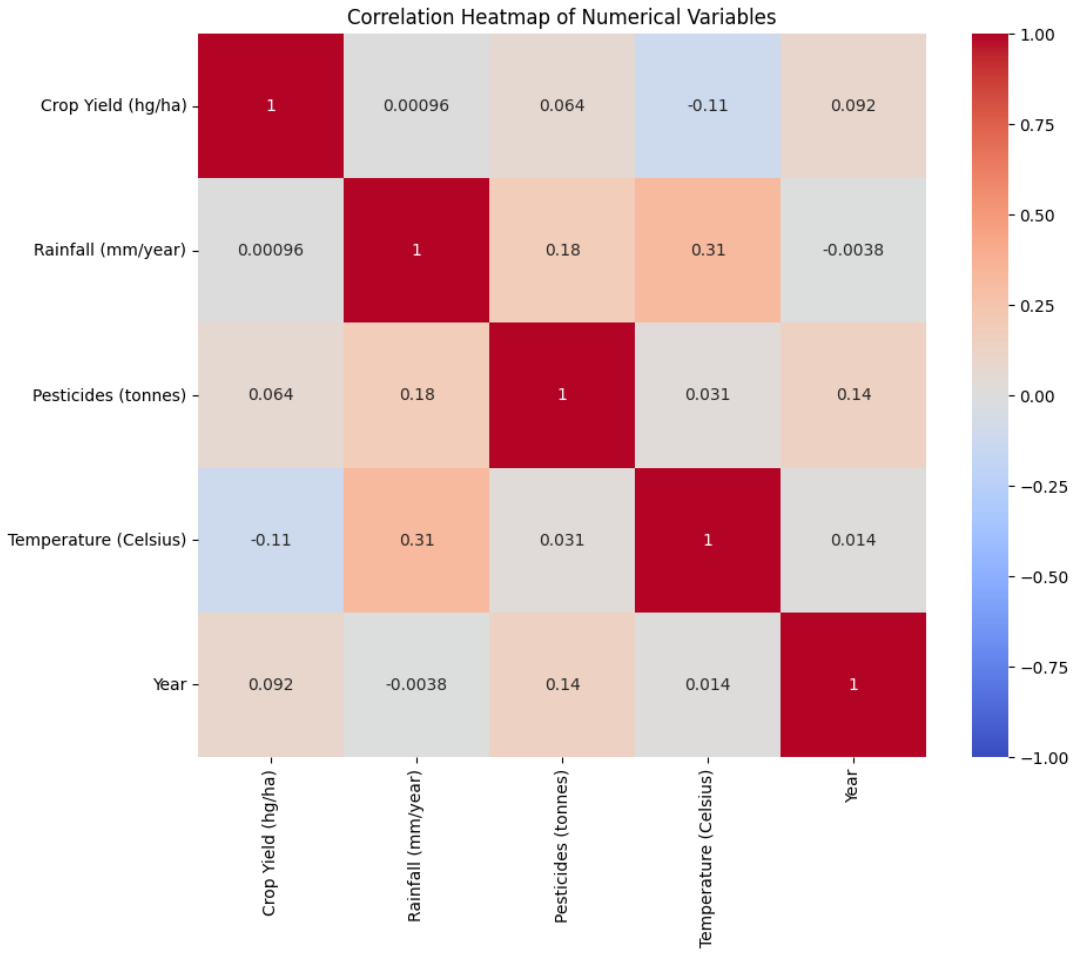


Figure 16. Correlation Matrix. Represents the direct correlation between numerical features in the dataset. The colors represent correlation values ranging from -1 (blue) to 1 (red). A correlation of +1 indicates a perfect positive correlation, 0 indicates no correlation, and -1 indicates a perfect negative correlation

Pesticides and Temperature exhibit weak correlations with Crop Yield. While these correlations are not very strong, they still indicate some degree of relationship between these variables and Crop Yield. Therefore, we will retain Pesticides and Temperature as input features for the regression model.

When considering the correlations between the input variables themselves, we observe a moderate positive correlation between Rainfall and Temperature. This suggests that these two features may provide similar information and including both of them in the model might introduce multicollinearity. Thus, we will select one of the correlated variables (Temperature) and exclude the other (Rainfall) to avoid redundancy and potential issues with model interpretation.

One-Hot-Encoding Categorical Variables

One-hot encoding is a technique used to represent categorical data in a numerical format so that it can be used in machine learning algorithms, which only take numerical input[33][38].

The process of one-hot encoding involves converting each unique category in a categorical feature into a new binary feature in a new dataset.

Since, our dataset has multiple categorical features, we will be using the One-hot-encoding technique to correctly format these features.

```
# Handle categorical variables
# OneHotEncoder
encoder = OneHotEncoder(handle_unknown='ignore')

# Fit and transform the categorical data in Area and Item
encoded_data = encoder.fit_transform(dataset[["Crop", 'Area']])

# Convert the encoded data back into a Pandas DataFrame
encoded_df = pd.DataFrame(encoded_data.toarray(),
columns=encoder.get_feature_names_out(["Crop", 'Area']))
df = pd.concat([dataset, encoded_df], axis=1)
```

Splitting the Dataset

Before inputing the data into a machine learning algorithm, the dataset has to be divided into features and labels, representing the input variables and the corresponding output variable, respectively[33][38].

```
# Create training (80%) and test (20%) sets
features_train, features_test , labels_train , labels_test = train_test_split(features,
labels, test_size =0.2, random_state=42)
```

The test_size parameter is set to 0.2, indicating that 20% of the data will be used for testing, while the remaining 80% will be used for training. The random_state parameter is set to 42 to ensure reproducibility, meaning that the same random split will be generated each time the code is run.

This train-test split is crucial for building and evaluating machine learning models. The training set is used to train the model on a subset of the data, allowing it to learn patterns and relationships between the features and labels. The test set, which is unseen during training, is then used to evaluate the model's performance and assess its generalization ability on new, unseen data. This separation helps to prevent overfitting, where the model memorizes the training data but fails to generalize well to new data.

2.4. Model Development

In order to provide accurate predictions, the selection and training of a good machine learning model becomes essential. This section outlines the process of building, comparing, training and testing multiple machine learning models using different algorithms, facilitated by the Jupyter Notebook framework.

Machine Learning Algorithms

There are several machine learning algorithms that can be used for regression prediction, and the choice of algorithm often depends on the specific characteristics of the data and the requirements of the project[33][38].

For our project, we compared the performance of different machine learning algorithms, which are explained as follows.

1. Random Forest Regression[8][16][61][66]: An ensemble algorithm that uses multiple decision trees to make a prediction. Random forest combines the predictions of multiple trees to produce a more robust and accurate result, and it is less prone to overfitting compared to decision trees.

2. Support Vector Regression (SVR) [10][54]: A non-linear algorithm that uses a technique called kernel trick to transform the input data into a higher-dimensional space. SVR is suitable for problems with complex relationships between the independent and dependent variables, but it can be slow to train and may not scale well to large datasets.

3. Gradient Boosting Regression [11][16]: An ensemble algorithm that builds a series of weak learners and combines their predictions to produce a final result. Gradient boosting can handle both linear and non-linear relationships, and it is often considered one of the best regression algorithms in terms of accuracy. However, it can be slow to train and may overfit the data if the number of trees is too large.

   XGBRegressor is an implementation of gradient boosting regression using the XGBoost library[42]. It is optimized for speed and scalability, and provides a number of advanced features such as parallel computing and automatic tuning of hyperparameters [7] [48].

4. Adaboost Regression [16][24]: An ensemble algorithm that builds a series of weak learners and gives more weight to instances that are misclassified by the previous weak learners. Adaboost can handle both linear and non-linear relationships, and it is fast to train. However, it can be prone to overfitting if the number of weak learners is too large.

5. ElasticNet [68]: A linear regression algorithm that combines the penalties of L1 and L2 regularization, which help to prevent overfitting and ensure stability. ElasticNet can handle highly correlated features and sparse data, but it may not perform well on non-linear problems.

6. SGDRegressor [17][46]: A linear regression algorithm that uses the Stochastic Gradient Descent optimization method to find the optimal coefficients. It is most commonly suitable for large-scale problems and online learning, but it can be sensitive to the choice of hyperparameters.

7. LGBMRegressor [19][36]: An implementation of gradient boosting regression using the LightGBM library. It is designed for fast training and is suitable for large-scale problems and also provides a number of advanced features such as parallel computing and handling of missing values.

Evaluation Metrics

To evaluate the results of each model, both during the model comparison and during the training and testing of the selected model, we will use the following standardized metrics[22][56][57][58]:

- Mean Absolute Error (MAE). Measures the average magnitude of errors in a set of predictions.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right|$$

A lower MAE indicates that the model's predictions are closer to the actual values.

- Root Mean Squared Error (RMSE). [reference] Measures the average magnitude of errors in a set of predictions that considers the direction of the errors.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$$

A lower RMSE value indicates that the model's predictions are closer to the actual values.

- R-Squared (R^2). Measures the proportion of variance in the dependent variable that is explained by the independent variables.

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

An R^2 value of 1 indicates a perfect fit between the observed data and the model's predictions, while an R^2 value of 0 indicates that the model doesn't explain any of the variation in the data. Values between 0 and 1 indicate the proportion of the variation in the data that is explained by the model.

- Adjusted R-Squared (Adj. R^2). Measures the proportion of variance in the dependent variable that is explained by the independent variables, but takes into account the number of independent variables in the model.

  The formula for adjusted R-squared is:

$$Adjusted\ R^2 = 1 - (1 - R^2) * \frac{n-1}{n-k-1}$$

  where:
  - R^2 is the regular R-squared (coefficient of determination) value.
  - n is the number of observations (sample size).
  - k is the number of independent variables (model's degrees of freedom).

  The adjusted R^2 value gives a better indication of the model's performance than R^2, especially in models with many predictors.

Cross Validation

Cross-validation is a resampling method that uses different samples of the data to test and train a predictive model on different iterations. It allows to estimate how accurately a model performs while protecting it against overfitting [1][18][25][26].

KFold and RepeatedKFold are two cross-validation techniques commonly used in machine learning to evaluate the performance of a model on unseen data.

The KFold method splits the dataset into k folds, where k is specified by the user, and trains the model on k-1 folds and tests it on the remaining one. This process is repeated k times with each fold being used as the test set once. The RepeatedKFold method is similar to KFold, but it allows repeating the k-fold cross-validation process multiple times, each with different randomization of the data. This is useful when the dataset is small, and there is a high variance in the results obtained from a single run of KFold.

Since we have a medium-sized dataset with a significant variability in the target variable, we will use like RepeatedKfold.

2.5. Model Evaluation and Improvement

Once the best-performing model has been identified, we need to evaluate its performance and implement the necessary changes to maximize its accuracy and efficiency.

This section involves the process of evaluating and improving the selected model. Once the best-performing model has been identified, a comprehensive evaluation must be conducted, examining various performance metrics (R-Squared, MAE, RMSE, Adjusted R-Squared, and MAE%).

Following the evaluation, the next step involves adjusting the model's hyperparameters through techniques such as GridSearch, which systematically explores various combinations of hyperparameter values to optimize performance. By fine-tuning these hyperparameters, the model can be further tailored to achieve higher accuracy and reliability[18][33][38].

2.4. App Prototype Development

In order to provide farmers with a usable solution, we development of an app prototype. This section outlines the process of designing a user-friendly app interface using Canva and then creating an app prototype using the React Native framework.

To begin with, after successfully constructing and optimizing the machine learning model, we focus on creating a visually appealing interface that enhances the overall user experience. The initial design of the interface will be done using Canvas, allowing for meticulous attention to detail and optimal visual representation of the data. Following this design phase, the prototype will be implemented in the React Native framework.

The development phase of the prototype involves translating the visual design into functional components and screens that can be rendered within the app. This is a complicated process that consists of multiple steps, the first four being the following:

- Component Structure: Involves breaking down the Canva-designed mockup into individual components that align with the app's screen hierarchy.
- Styling and Layout: Consists of creating and appling the styles and layout defined in the Canva mockup using React Native's styling system.
- Navigation and Routing: Involves implementing navigation and routing between screens to ensure a seamless user experience
- Data Binding and Interactivity: Consists of connecting the user interface components to the underlying data and logic of the app.

3. Results

3.1. Models Comparison

For model selection, we experimented with various regression models to determine the most suitable one for our task. We evaluated the performance of the following models: Support Vector Machines (SVM), Random Forest, AdaBoost, ElasticNet, XGBoost, Stochastic Gradient Descent Regression (SGDR), and LightGBM (LGBM).

By testing multiple models, we aimed to identify the model that provides the best performance in terms of accuracy, predictive power, and generalization to unseen data. Each model has its own strengths and weaknesses, and comparing their results allows us to make an informed decision about which model is most appropriate for our specific regression task.

The detailed results of these model performance are presented in Appendix A, while a more concise result summary can be seen in the following table:

Model	R-squared	MAE	RMSE	Adj. R-squared	MAE%
SVM	-0.195	56983.75	93115.23	6752.16	11.78%
RandomForest	0.987	3660.70	9715.70	74.50	0.76%
AdaBoost	0.642	40639.97	50960.26	2023.09	8.40%
ElasticNet	0.755	29498.73	42136.01	1383.43	6.10%
XGradBoost	0.986	4737.67	10050.96	79.66	0.98%
SGDR	0.750	28734.01	42624.71	1415.68	5.94%
LGBMR	0.945	11978.47	19949.99	310.90	2.48%

Table 1. Performance metrics for the seven regression models being compared for crop yield prediction. Each row displays the R-squared value (R^2), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Adjusted R-squared (Adj. R^2), and MAE (%) for the respective model.

The comparison of these models is crucial in order to select the one that exhibits the highest performance metrics, such as R-squared, mean squared error (MSE), or root mean squared error (RMSE).

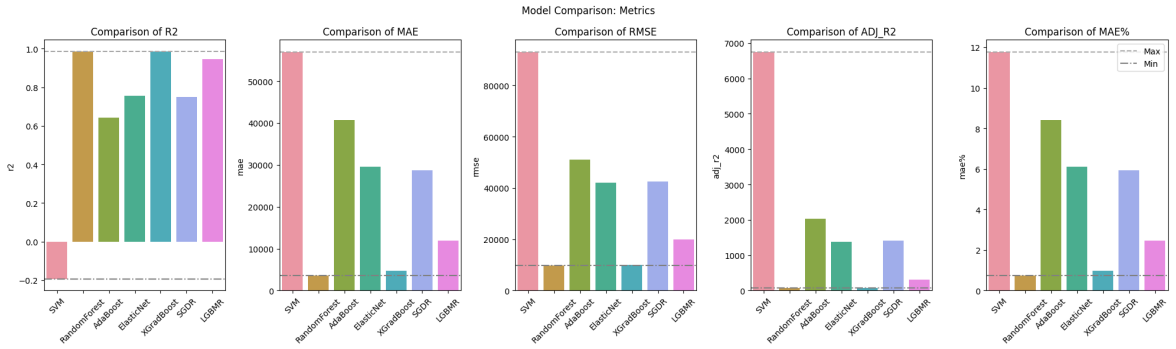


Figure 17. Each of the five bar plots are representing a different metric: R-squared (R^2), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Adjusted R-squared (Adj. R^2), and MAE% (MAE percentage). Each bar plot displays the performance of various models for the corresponding metric. Horizontal lines indicate the maximum and minimum scores achieved across all models.

As we can see in Figure 17, the RandomForest and XGradBoost models consistently demonstrate strong performance across multiple metrics. They have high R-squared

values, low MAE and RMSE values, and low MAE% values, indicating accurate predictions and good overall fit. ElasticNet also shows competitive performance in terms of MAE and RMSE, but the Adjusted R-Squared value is too high, implying overfitting of the model. On the other hand, the SVM and AdaBoost models perform relatively poorly in terms of these metrics.

Based on these results, we can confidently select either XGradBoost or RandomForest as our model our choice. In this case, we've selected XGradBoost, since this algorithm can later be more easily exported.

3.2. Model Training and Hyperparameter Tuning

After conducting a thorough comparison of different models, we have selected the XGBoost Regressor as our choice for performing crop yield prediction. To optimize the performance of the model, we employed grid search and a pipeline approach to identify the best hyperparameters.

The parameter grid used for grid search included various options for 'n_estimators', 'max_depth', 'learning_rate', 'min_child_weight', 'gamma', 'reg_alpha', and 'reg_lambda', as displayed in Table 2.

rgs__n_estimators	100, 200, 300
rgs__max_depth	3.00, 6.0,12.0
rgs__learning_rate	0.05, 0.1, 0.2
rgs__min_child_weight	1.00, 10.0
rgs__gamma	0.00, 0.1, 0.5
rgs__reg_alpha	0.00, 0.1, 0.5
rgs__reg_lambda	0.00, 0.1, 0.5

Table 2. Hyperparameters options for the performed GridSearch.

By exploring different combinations of these hyperparameters, we aimed to find the optimal configuration for our XGBoost model.

To evaluate the performance of the selected model, we generated a train-validation graph (Figure 18). As we can see in the graph, the training score remained consistently high, close to 1, while the validation score surpassed 0.98. This indicates that the model generalizes well to unseen data, showing promising predictive capabilities.

Furthermore, we obtained several performance metrics to assess the accuracy and reliability of the model. The best score achieved by the model was 0.9857, which is a strong indication of its predictive power. The corresponding set of best parameters identified through grid search consisted of 'gamma' = 0.5, 'learning_rate' = 0.2, 'max_depth' = 12, 'min_child_weight' = 1, 'n_estimators' = 300, 'reg_alpha' = 0.5, and 'reg_lambda' = 0.5. These parameter values were found to optimize the model's performance based on the given dataset and prediction task.

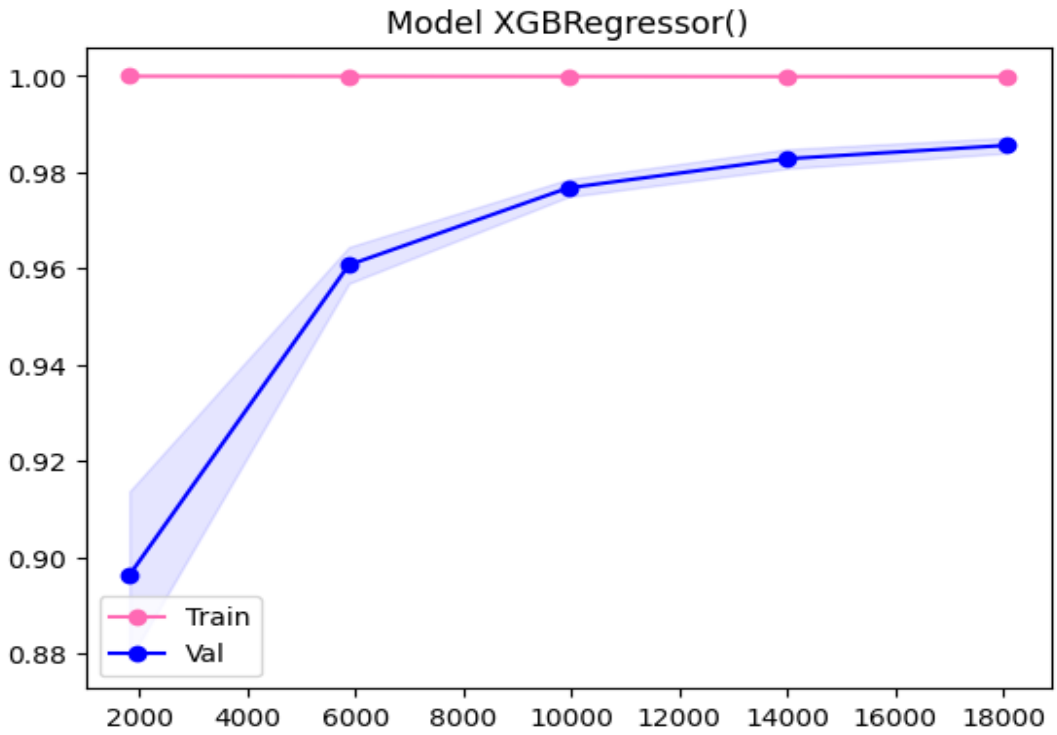


Figure 18. Learning curve showing the performance of the model during training (pink) and validation (blue) stages. The shaded areas represent the standard deviation of the scores, while the solid lines represent the average scores. The model exhibits high performance on both.

Model	R^2	MAE	RMSE	Adj. R^2	MAE%
XGBoost	0.988286	3661.640	9217.860	67.160	0.75692

Table 3. Performance metrics for the XGBoost model we selected for crop yield prediction.

Crop	R^2	MAE	MAE%
Cassava	0.985485	4798.500	1.285
Maize	0.970338	2404.510	1.917
Plantains and others	0.924819	8117.964	2.044
Potatoes	0.965109	8693.648	1.841
Rice, paddy	0.962729	2102.584	2.117
Sorghum	0.943424	1872.057	2.482
Soybeans	0.908057	1195.224	2.886
Sweet potatoes	0.976809	4859.903	1.435
Wheat	0.951262	2046.192	2.126
Yams	0.986989	3473.811	1.627

Table 4. Performance metrics for trained XGradBoost Model, for individual crops: Cassava, Maize, Plantains and others, Potatoes, Rice (paddy), Sorghum, Soybeans, Sweet potatoes, Wheat, and Yams. Each row in the table provides the R-squared value (R^2), Mean Absolute Error (MAE) and MAE (%) for the respective crop.

Analyzing the evaluation metrics for the trained model, we can see it shows a strong performance in predicting crop yield. A R-squared value (R^2) of 0.988 suggests that the model can explain approximately 98.8% of the variance in the crop yield data, and a MAE (%) of 0.76%, which suggests a low relative error. Also, a Adjusted R-squared (Adjusted R^2) value of 67.160, indicates the proportion of variance is explained after adjusting for the number of predictors.

3.3. App Prototype

Model Exporting

In order to use the trained model in a React Native application, the model must be exported into a JSON format. This is a multi-step process that consists of the following:

Firstly, the trained model needs to be serialized into a format that can be easily exported. The Xgboost library provides functionality to convert a trained model into a binary format using the pickle module in Python.

Next, the serialized model file needs to be converted into a JSON format. The Xgboost library in Python provides specific functions for which allows exporting the model's structure and details as a JSON file. This file can then be accessed within the React Native application by importing it as a resource or making it available through an API endpoint.

```
# Serialize and save the model using joblib
joblib.dump(model, 'xgb_regressor.joblib')

# Convert the trained model to a JSON format
model.get_booster().dump_model('xgb_trained_model.json')
```

In the React Native application, the JSON file can be loaded and parsed to retrieve the necessary information for performing predictions. This typically involves extracting the hyperparameters, reconstructing the tree structure, and implementing the prediction logic using the model data.

```
import xgboost from 'xgboost';
const modelFilePath = './assets/xgb_trained_model.json';

// Load the JSON model file
const jsonModel = require(modelFilePath);
// Extract the serialized data
const serializedModel = jsonModel.model;
const hyperparameters = jsonModel.hyperparameters;
const featureNames = jsonModel.feature_names;

// Create an XGBoost model object from the serialized data
const model = xgboost.Booster({ model_file: serializedModel });
```

Mockup Design

The initial interface design process involves creating a visual representation of the app prototype's graphic interface using Canva.

Using Canva's design features, we created a mockups that outline the structure and layout of the app prototype's interface. This mockup considers elements such as user inputs, result visualization, navigation menus, buttons, and other interactive components.

As can be seen in Figure 19, the app comprises five screens:

1. Logo Screen: The app's logo, featuring the name "Crop Yield Predictor," captures attention while establishing the brand identity. It sets the tone for a reliable and efficient farming tool.
2. Home Screen: Upon launching the app, users are greeted with a visually appealing main menu screen. Here, users can access three features: the *Crop Yield Prediction*, their past prediction *History*, and a *Q&A* section. Intuitive icons and clear labels enhance ease of navigation.
3. Crop Selection Screen: When selecting *Crop Yield Prediction*, users are presented with a comprehensive list featuring a range of ten options.
4. Crop Conditions Screen: Once the crops of interest are selected, this screen allows for users to input their data.
5. Predicted Yield Screen: After inputting the required conditions, users are presented with a detailed summary of the predictions made in the results screen. The highest-yielding crop is marked with a crown icon for easier identification.

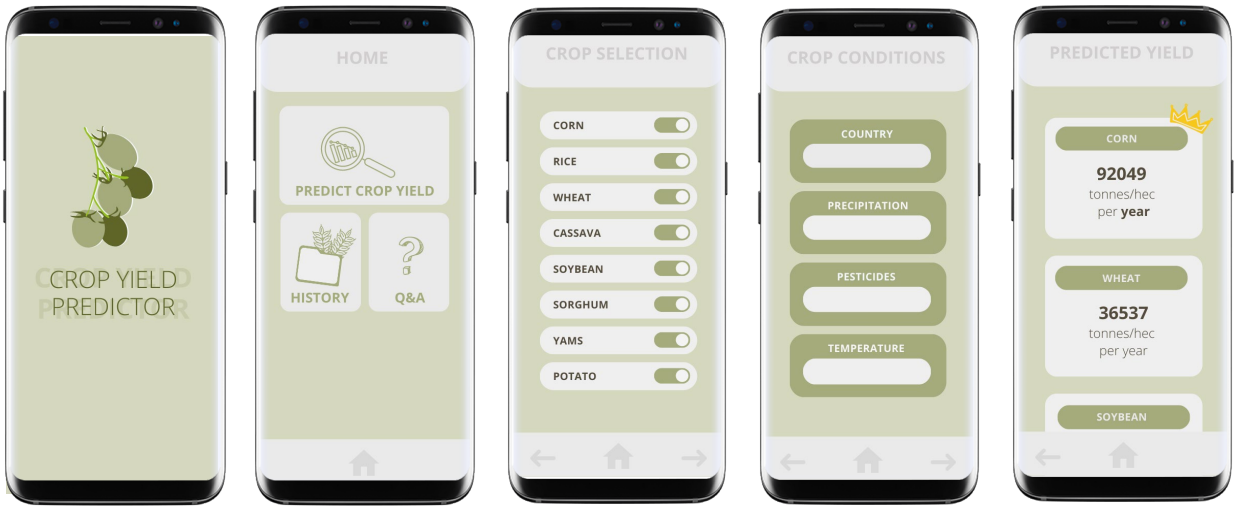


Figure 19. App Mockup showcasing the user interface design of the Crop Yield Predictor app. The mockup consists of 5 screens: Logo, Home, Crop Selection, Crop Conditions, Predicted Yield (results).

Prototype Development

Due to time constraints, the completion of the app mockup was prioritized over the subsequent step of implementing it in the React Native framework for prototype deployment. As a result, the development phase of the app prototype remains unfinished within the given timeframe.

Out of the steps outlined in Section [reference], the completion of the app development in the React Native framework was limited to the first two stages:

- Component Structure: The Canva-designed mockup was successfully broken down into individual components that correspond to the app's screen hierarchy.

- Styling and Layout: The styles and layout defined in the Canva mockup were created and applied using React Native's styling system.

However, the subsequent steps, including navigation and routing, as well as data binding and interactivity, were not completed. These stages involve implementing seamless navigation between screens and connecting user interface components to the underlying data and logic of the app for interactivity.

Although these remaining steps are pending, we aim to prioritize their completion in future iterations of the app development process.

4. Discussion and Conclusions

4.1. Discussion

The project successfully accomplished its goals of finding and analyzing a dataset of crop yields, training a regression model to predict crop performance and designing a prototype application that provides valuable information and predictions for agricultural decision-making. The prototype development and deployment remained unfinished due to time constraints and the complexity of the task.

In regard to the model itself, it is important to note that the nature of the data used in the project primarily represents crops on a large scale. This means that the predictions generated by the model may be less applicable and useful for small-scale farmers who operate with different conditions, resources, and management practices.

Large-scale agricultural operations often have standardized practices, access to advanced technologies, and greater resources, which can result in more consistent and predictable outcomes. Therefore, the predictions derived from the model are better suited to assisting larger agricultural enterprises in optimizing their production processes and making informed decisions regarding resource allocation, crop selection, and harvest planning.

On the other hand, small-scale farmers often face unique challenges, such as limited resources, diverse cropping systems, and variations in environmental conditions. These factors can significantly influence crop performance and make it challenging to generalize predictions based solely on large-scale data. Therefore, while the predictions may still provide some valuable insights to small-scale farmers, they should be interpreted with caution and considered alongside local knowledge and specific contextual factors.

To improve the applicability of the application for small-scale farmers, future work could involve collecting and incorporating more localized and diversified data sources, considering factors such as soil quality, microclimate variations, and specific management practices commonly used by small-scale agricultural operations. By expanding the scope of data collection and analysis, the application could provide more tailored and accurate predictions for a broader range of agricultural contexts, including small-scale farming.

4.2. Future Improvements

Due to time constraints and other limitations, the project is currently very rudimentary, leaving ample room for improvement.

Most notably, the development and deployment of the app prototype, including the implementation of the Canva-designed mockup in the React Native framework, remain a work in progress. The completion of this phase of the project is the main priority for future efforts in improving the project.

Additionally, after evaluating the reach and usability of the current project, we have compiled the following list of improvements opportunities to enhance the project and it's impact.

- Geographic focus: It would be beneficial to concentrate on a specific geographical area and make predictions that are more tailored to that region, considering factors such as soil type, climate, and local agricultural practices. This would result in more precise predictions.
- Exploration of seasonal averages: Instead of relying solely on annual averages, it would be valuable to explore seasonal averages in data analysis. Crop yields can be affected by seasonal factors such as temperature variations, precipitation patterns, and daylight duration. By examining seasonal averages, the application could capture the impact of these factors on crop growth and incorporate them into the prediction model, allowing for more accurate and detailed predictions throughout different seasons.
- Integration of weather forecasts: Integrating real-time weather conditions into the application would be a significant improvement. By accessing weather forecast data, the application could incorporate the most recent information such as temperature, rainfall, and humidity into the prediction model. This dynamic approach would enable the provision of more accurate and up-to-date predictions based on current weather conditions, thereby improving the reliability of crop performance predictions.
- Analysis of pesticide effects: In addition to considering the amount of pesticides used, it would be valuable to investigate how different types of pesticides affect crop yields. By incorporating information about specific types of pesticides used, the application could analyze their impacts on crop health and performance, providing valuable insights to optimize pesticide selection and usage and improve crop productivity.
- Expansion to more crops or different species: As a future development, it would be beneficial to expand the application to include more types of crops or even different species within the same crop. This would enhance its utility and applicability by providing information on performance and yield variations in different crops or genetic varieties, enabling more informed decision-making and personalized recommendations for specific agricultural scenarios.

4.3. Conclusions

The project has made significant progress in achieving its goals, including finding and analyzing a dataset of crop yields, training a regression model to predict crop performance, and designing a prototype for an application that would provide valuable information and predictions for agricultural decision-making. The development and deployment phase of the project is still in-progress, but this is an expected outcome given the complexity of the task.

In general, there is a lot of room left for improvement, and we hope to continue working on this project, incorporating new features and making changes in the future.

All in all, the project not only successfully achieved most of its goals but also demonstrated the immense potential that lies within the application of artificial intelligence in the field of agriculture.

5. References

[1] 3.1. Cross-validation: evaluating estimator performance. (2023, May 23). Scikit-learn. https://scikit-learn.org/stable/modules/cross_validation.html

[2] AenVerde, R. (2023). La agricultura de precisión y la inteligencia artificial, principales tendencias en FoodTech. A En Verde - Medio De Información De Agricultura. https://www.aenverde.es/la-agricultura-de-precision-y-la-inteligencia-artificial-principales-tendencias-en-foodtech/

[3] Agriculture And Food Systems Unsustainable - Our World. (2023, May 23). https://ourworld.unu.edu/en/agriculture-and-food-systems-unsustainable

[4] AI For Everyone. (2023, Feb 1). Coursera. https://www.coursera.org/learn/ai-for-everyone?=

[5] Ansarifar, J., Wang, L., & Archontoulis, S. V. (2021a). An interaction regression model for crop yield prediction. Scientific Reports, 11(1). https://doi.org/10.1038/s41598-021-97221-7

[6] Ansarifar, J., Wang, L., & Archontoulis, S. V. (2021b). An interaction regression model for crop yield prediction. Scientific Reports, 11(1). https://doi.org/10.1038/s41598-021-97221-7

[7] Bottou, L. (2016, June 15). Optimization Methods for Large-Scale Machine Learning. arXiv.org. https://arxiv.org/abs/1606.04838

[8] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/a:1010933404324

[9] Brownlee, J. (2016). Machine Learning Mastery With Python: Understand Your Data, Create Accurate Models, and Work Projects End-to-End. Machine Learning Mastery.

[10] Chang, C., & Lin, C. (2011). LIBSVM. ACM Transactions on Intelligent Systems and Technology, 2(3), 1–27. https://doi.org/10.1145/1961189.1961199

[11] Chen, T., & Guestrin, C. (2016a). XGBoost. https://doi.org/10.1145/2939672.2939785

[12] Chen, T., & Guestrin, C. (2016b). XGBoost. https://doi.org/10.1145/2939672.2939785

[13] Columbus, L. (2021, February 17). 10 Ways AI Has The Potential To Improve Agriculture In 2021. Forbes. https://www.forbes.com/sites/louiscolumbus/2021/02/17/10-ways-ai-has-the-potential-to-improve-agriculture-in-2021/

[14] Crop Yield Prediction Dataset. (2021, December 1). Kaggle. https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset

[15] Fernández, F. R. (2020). Inteligencia Artificial y Agricultura: nuevos retos en el sector agrario. Revista Campo Jurídico, 8(2), 123–139. https://doi.org/10.37497/revcampojur.v8i2.662

[16] Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences, 55(1), 119–139. https://doi.org/10.1006/jcss.1997.1504

[17] Friedman, J. H. (2002). Stochastic gradient boosting. Computational Statistics & Data Analysis, 38(4), 367–378. https://doi.org/10.1016/s0167-9473(01)00065-2

[18] Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. “O’Reilly Media, Inc.”

[19] Giba, B. (2021). Elastic Net Regression Explained, Step by Step. Machine Learning Compass. https://machinelearningcompass.com/machine_learning_models/elastic_net_regression/

[20] Gonzalez, W. (2023, February 2). How AI Is Cropping Up In The Agriculture Industry. Forbes. https://www.forbes.com/sites/forbesbusinesscouncil/2023/02/02/how-ai-is-cropping-up-in-the-agriculture-industry/

[21] Gonzalez-Sanchez, A., Frausto-Solís, J., & Ojeda-Bustamante, W. (2014). Attribute Selection Impact on Linear and Nonlinear Regression Models for Crop Yield Prediction. The Scientific World Journal, 2014, 1–10. https://doi.org/10.1155/2014/509429

[22] Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). The Elements of Statistical Learning. In Springer series in statistics. Springer Science+Business Media. https://doi.org/10.1007/978-0-387-84858-7

[23] How AI can help water go further in farming. (2022, May 20). World Economic Forum. https://www.weforum.org/agenda/2021/01/ai-agriculture-water-irrigation-farming/

[24] insidelearningmachines. (2023, May 9). Understanding the Adaboost Regression Algorithm - Inside Learning Machines. Inside Learning Machines. https://insidelearningmachines.com/adaboost_regression_algorithm/

[25] Introduction to Statistical Learning. (2023, May 23). https://trevorhastie.github.io/ISLR/

[26] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer Science & Business Media.

[27] Lecerf, R., Ceglar, A., López-Lozano, R., Van Der Velde, M., & Baruth, B. (2019). Assessing the information in crop model and meteorological indicators to forecast crop yield over Europe. Agricultural Systems, 168, 191–202. https://doi.org/10.1016/j.agsy.2018.03.002

[28] Lenniy, D. (2023, March 23). The Future of AgriTech: Trends and Innovations in Agriculture to Watch in 2023. Intellias. https://intellias.com/innovations-in-agriculture/

[29] Liu, J., & Wang, X. (2021). Plant diseases and pests detection based on deep learning: a review. Plant Methods, 17(1). https://doi.org/10.1186/s13007-021-00722-9

[30] Mansour, K. (2022, December 23). Infographic – Top trends in agricultural biotechnology > Early Metrics. Early Metrics. https://earlymetrics.com/infographic-agricultural-biotechnology-trends/

[31] MassChallenge. (2023, May 31). Agriculture Innovation: 10 Tech Trends to Watch in 2023 - MassChallenge. https://masschallenge.org/articles/agriculture-innovation/

[32] Matplotlib — Visualization with Python. (2023, May 23). https://matplotlib.org/

[33] Nantasenamat, C. (2021, December 15). How to Build a Machine Learning Model - Towards Data Science. Medium. https://towardsdatascience.com/how-to-build-a-machine-learning-model-439ab8fb3fb1

[34] NumPy. (2023, May 23). https://numpy.org/

[35] pandas - Python Data Analysis Library. (2023, May 23). https://pandas.pydata.org/

[36] Pattnaik, S. (2021, December 24). A LightGBM Autoregressor — Using Sktime - Towards Data Science. Medium. https://towardsdatascience.com/a-lightgbm-autoregressor-using-sktime-6402726e0e7b

[37] Paz, D. B., Henderson, K., & Loreau, M. (2020). Agricultural land use and the sustainability of social-ecological systems. Ecological Modelling, 437, 109312. https://doi.org/10.1016/j.ecolmodel.2020.109312

[38] PedregosaFabian, VaroquauxGaël, GramfortAlexandre, MichelVincent, ThirionBertrand, GriselOlivier, BlondelMathieu, PrettenhoferPeter, WeissRon, DubourgVincent, VanderplasJake, PassosAlexandre, CournapeauDavid, BrucherMatthieu, PerrotMatthieu, & DuchesnayÉdouard. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. https://doi.org/10.5555/1953048.2078195

[39] Pictet. (2023, March 14). Agritech: Can cutting edge innovations solve the global food crisis? The Pictet Group. https://www.pictet.com/tw/en/insights/agritech-cutting-edge-innovations-food-crisis

[40] Plotly. (2023, May 23). https://plotly.com/python/

[41] Project Jupyter. (2023, May 23). Home. https://jupyter.org/

[42] Python Package Introduction — xgboost 1.7.6 documentation. (2023, May 23). https://xgboost.readthedocs.io/en/stable/python/python_intro.html

[43] Rahman, M. (2012). Review of Factors Affecting Agricultural Ecosystem. https://www.semanticscholar.org/paper/Review-of-Factors-Affecting-Agricultural-Ecosystem-Rahman/dfbaff68de28c836bfa4e48926a97893f134fc64

[44] Raza, A., Razzaq, A., Mehmood, S. S., Zou, X., Zhang, X., Yan, L., & Xu, J. (2019). Impact of Climate Change on Crops Adaptation and Strategies to Tackle Its Outcome: A Review. Plants, 8(2), 34. https://doi.org/10.3390/plants8020034

[45] React Native · Learn once, write anywhere. (2023, May 30). https://reactnative.dev/

[46] Regression Example with SGDRegressor in Python. (2020, September 15). https://www.datatechnotes.com/2020/09/regression-example-with-sgdregressor-in-python.html

[47] Roser, M. (2013, May 9). Future Population Growth. Our World in Data. https://ourworldindata.org/future-population-growth

[48] Ruder, S. (2016, September 15). An overview of gradient descent optimization algorithms. arXiv.org. https://arxiv.org/abs/1609.04747

[49] scikit-learn: machine learning in Python — scikit-learn 1.3.0 documentation. (2023, May 23). https://scikit-learn.org/stable/index.html

[50] seaborn: statistical data visualization — seaborn 0.12.2 documentation. (2023, May 23). https://seaborn.pydata.org/

[51] Selvaraj, M. G., Vergara, A., Ruiz, H. H., Safari, N., Elayabalan, S., Ocimati, W., & Blomme, G. (2019). AI-powered banana diseases and pest detection. Plant Methods, 15(1). https://doi.org/10.1186/s13007-019-0475-z

[52] Shahmohamadloo, R. S., Febria, C. M., Fraser, E. D. G., & Sibley, P. K. (2021). The sustainable agriculture imperative: A perspective on the need for an agrosystem approach to meet the United Nations Sustainable Development Goals by 2030. Integrated Environmental Assessment and Management, 18(5), 1199–1205. https://doi.org/10.1002/ieam.4558

[53] Sharma, V., Tripathi, A., & Mittal, H. (2022). Technological revolutions in smart farming: Current trends, challenges & future directions. Computers and Electronics in Agriculture, 201, 107217. https://doi.org/10.1016/j.compag.2022.107217

[54] Sharp, T. (2023, April 3). An Introduction to Support Vector Regression (SVR) - Towards Data Science. Medium. https://towardsdatascience.com/an-introduction-to-support-vector-regression-svr-a3ebc1672c2

[55] Shastry, A. (2017). Prediction of Crop Yield Using Regression Techniques. https://www.semanticscholar.org/paper/Prediction-of-Crop-Yield-Using-Regression-Shastry-Ha/5209efb2db7aea2c54ee1cc81d21aee51818b49f

[56] sklearn.metrics.mean_absolute_error. (2023, May 23). Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_error.html

[57] SmritiS. (2023). What is Mean Squared Error, Mean Absolute Error, Root Mean Squared Error and R Squared? Studytonight. https://www.studytonight.com/post/what-is-mean-squared-error-mean-absolute-error-root-mean-squared-error-and-r-squared

[58] Stephanie. (2020, December 28). Absolute Error & Mean Absolute Error (MAE) - Statistics How To. Statistics How To. https://www.statisticshowto.com/absolute-error/

[59] THE 17 GOALS | Sustainable Development. (2023, May 14). https://sdgs.un.org/goals

[60] The State of Food and Agriculture 2022. (2022). In FAO eBooks. https://doi.org/10.4060/cb9479en

[61] Thorn, J. (2021, December 15). Random Forest Explained - Towards Data Science. Medium. https://towardsdatascience.com/random-forest-explained-7eae084f3ebe

[62] Top Trends in Agritech Software Development. (2023, May 14). NCube. https://ncube.com/blog/top-trends-in-agritech-software-development

[63] United Nations. (2023, May 23). Feeding the World Sustainably | United Nations. https://www.un.org/en/chronicle/article/feeding-world-sustainably

[64] Veum, K. S., Sudduth, K. A., Kremer, R. J., & Kitchen, N. R. (2017). Sensor data fusion for soil health assessment. Geoderma, 305, 53–61. https://doi.org/10.1016/j.geoderma.2017.05.031

[65] What is Exploratory Data Analysis? | IBM. (2023, Mar 11). https://www.ibm.com/topics/exploratory-data-analysis

[66] Yiu, T. (2021, December 10). Understanding Random Forest - Towards Data Science. Medium. https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[67] Young, S. (2020). The Future of Farming: Artificial Intelligence and Agriculture. Harvard International Review. https://hir.harvard.edu/the-future-of-farming-artificial-intelligence-and-agriculture/

[68] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B-statistical Methodology, 67(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

[69] OpenAI. (2023, June 10). ChatGPT 3. https://openai.com

Assisted in identifying grammar and spelling errors, offering suggestions for improving sentence structure, and aiding in rephrasing certain complex passages. I am aware of the limitations of AI-based tools and, therefore, exercised caution while incorporating their suggestions into the final content.

[70] Diaz del Ser, S. (2023). GitHub. https://github.com/saradiazdelser/application-of-machine-learning-algorithms-to-perform-crop-yield-prediction

Appendix A: Model Comparison Details

In this appendix, you will find the detailed results of the each model's performance, consisting of the Train-Val learning curve and the broken-down performance metrics for each individual crop.

Random Forest

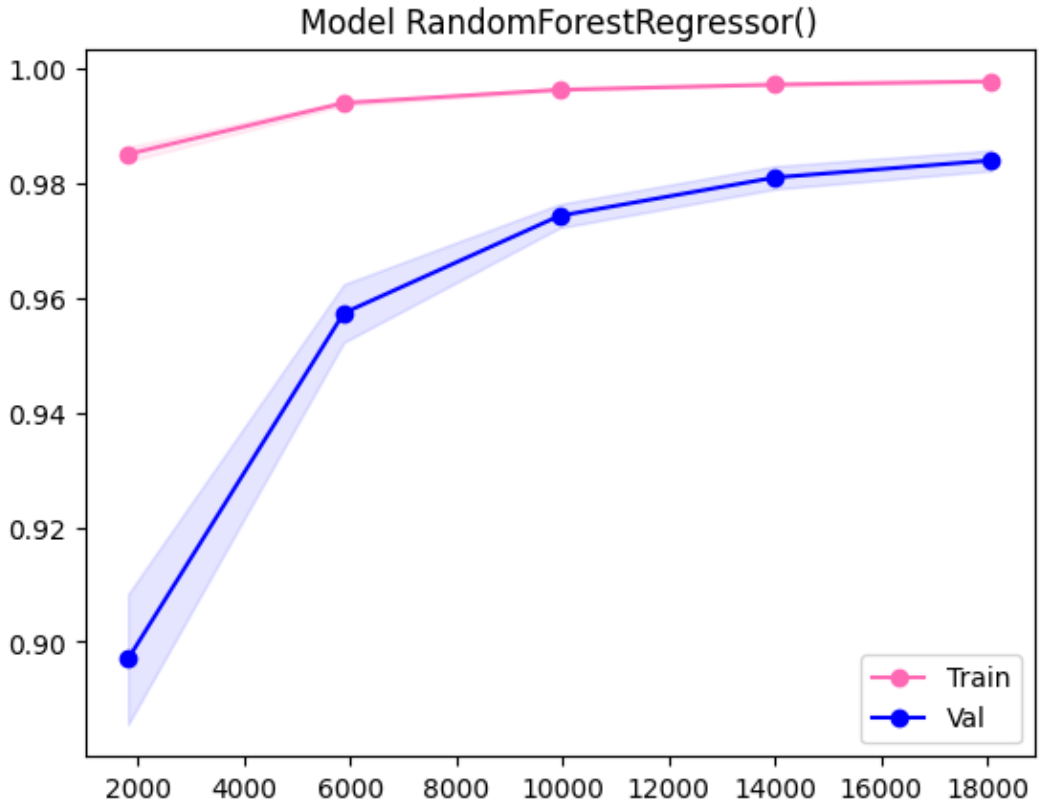


Figure 20. Learning curve showing the performance of the RandomForestRegressor() model during training (pink) and validation (blue) stages. The shaded areas represent the standard deviation of the scores, while the solid lines represent the average scores.

Crop	R-squared	MAE	MAE%
Cassava	0.98	5223.98	1.40%
Maize	0.97	2291.53	1.83%
Plantains and others	0.92	8784.33	2.21%
Potatoes	0.96	8869.78	1.88%
Rice, paddy	0.96	2000.53	2.01%
Sorghum	0.93	1792.31	2.38%
Soybeans	0.94	969.23	2.34%
Sweet potatoes	0.97	5210.53	1.54%
Wheat	0.96	1724.45	1.79%
Yams	0.98	3393.64	1.59%

Table 5. Performance metrics for the RandomForestRegressor() model, for individual crops: Cassava, Maize, Plantains and others, Potatoes, Rice (paddy), Sorghum, Soybeans, Sweet potatoes, Wheat, and Yams. Each

row in the table provides the R-squared value (R^2), Mean Absolute Error (MAE) and MAE (%) for the respective crop.

These results provide an overview of the performance of different crops in terms of regression metrics. Cassava and Yams stand out as the top-performing crops with high R-squared values and low MAE% values. Maize, Sweet potatoes, and Potatoes also show strong performance. Plantains and others, Sorghum, Rice, paddy, Soybeans, and Wheat have slightly lower but still reasonable performance.

SVR

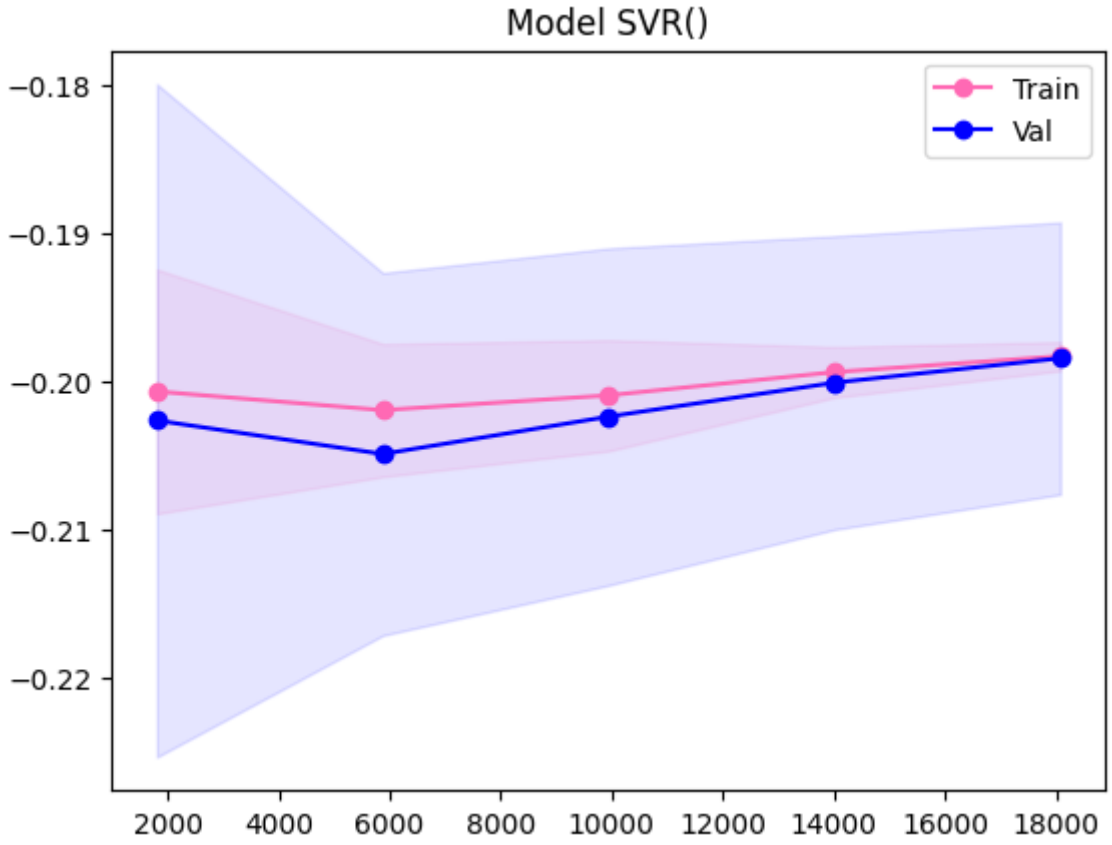


Figure 21. Learning curve showing the performance of the SVR() model during training (pink) and validation (blue) stages. The shaded areas represent the standard deviation of the scores, while the solid lines represent the average scores.

Crop	R-squared	MAE	MAE%
Cassava	-1.45	106867.49	28.62%
Maize	-0.01	21743.61	17.33%
Plantains and others	-0.84	69548.58	17.51%
Potatoes	-2.99	162430.62	34.40%
Rice, paddy	-0.02	15232.00	15.33%
Sorghum	-1.39	21604.62	28.64%
Soybeans	-8.23	21494.59	51.91%
Sweet potatoes	-1.33	82027.93	24.23%
Wheat	-0.24	17567.72	18.26%

Crop	R-squared	MAE	MAE%
Yams	-1.88	77142.42	36.13%

Table 6. Performance metrics for the SVR() model, for individual crops: Cassava, Maize, Plantains and others, Potatoes, Rice (paddy), Sorghum, Soybeans, Sweet potatoes, Wheat, and Yams. Each row in the table provides the R-squared value (R^2), Mean Absolute Error (MAE) and MAE (%) for the respective crop.

These results provide an overview of the performance of different crops in terms of regression metrics. They suggest that the SVR() model is not performing well. The negative R-squared values indicate poor fits, and the high MAE and MAE% values highlight significant prediction errors. Further analysis and model improvements are necessary to enhance the accuracy of yield predictions for these crops.

AdaBoost

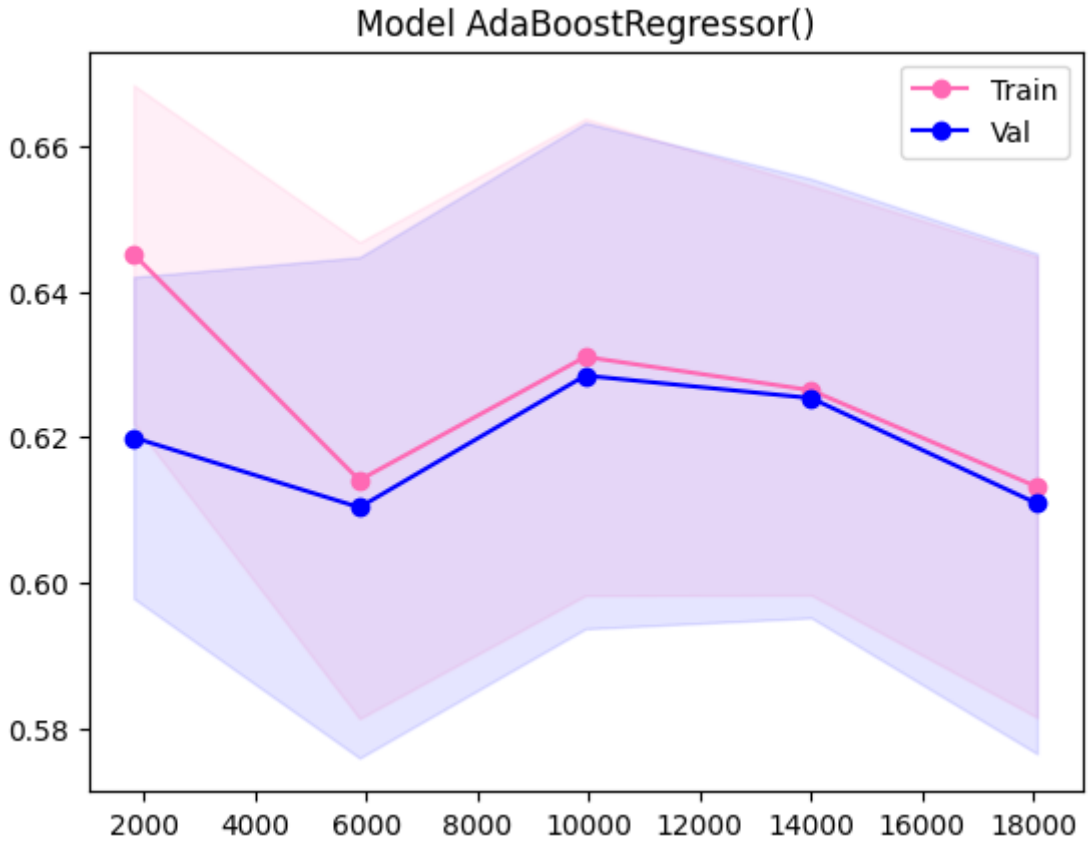


Figure 22. Learning curve showing the performance of the AdaBoostRegressor() model during training (pink) and validation (blue) stages. The shaded areas represent the standard deviation of the scores, while the solid lines represent the average scores.

Crop	R-squared	MAE	MAE%
Cassava	0.69	40138.14	10.75%
Maize	-0.67	31373.62	25.01%
Plantains and others	-0.31	52844.22	13.31%
Potatoes	0.39	57967.74	12.28%
Rice, paddy	-0.59	22345.85	22.49%
Sorghum	-5.72	37438.57	49.64%

Crop	R-squared	MAE	MAE%
Soybeans	-28.96	40223.99	97.14%
Sweet potatoes	-0.04	57312.93	16.93%
Wheat	-2.27	30656.76	31.86%
Yams	-1.10	60819.66	28.48%

Table 7. Performance metrics for the AdaBoostRegressor() model, for individual crops: Cassava, Maize, Plantains and others, Potatoes, Rice (paddy), Sorghum, Soybeans, Sweet potatoes, Wheat, and Yams. Each row in the table provides the R-squared value (R^2), Mean Absolute Error (MAE) and MAE (%) for the respective crop.

These results provide an overview of the performance of different crops in terms of regression metrics. Cassava shows a relatively good fit and lower prediction errors, while Soybeans and Sorghum exhibit higher errors and poorer model performance.

Elastic Net

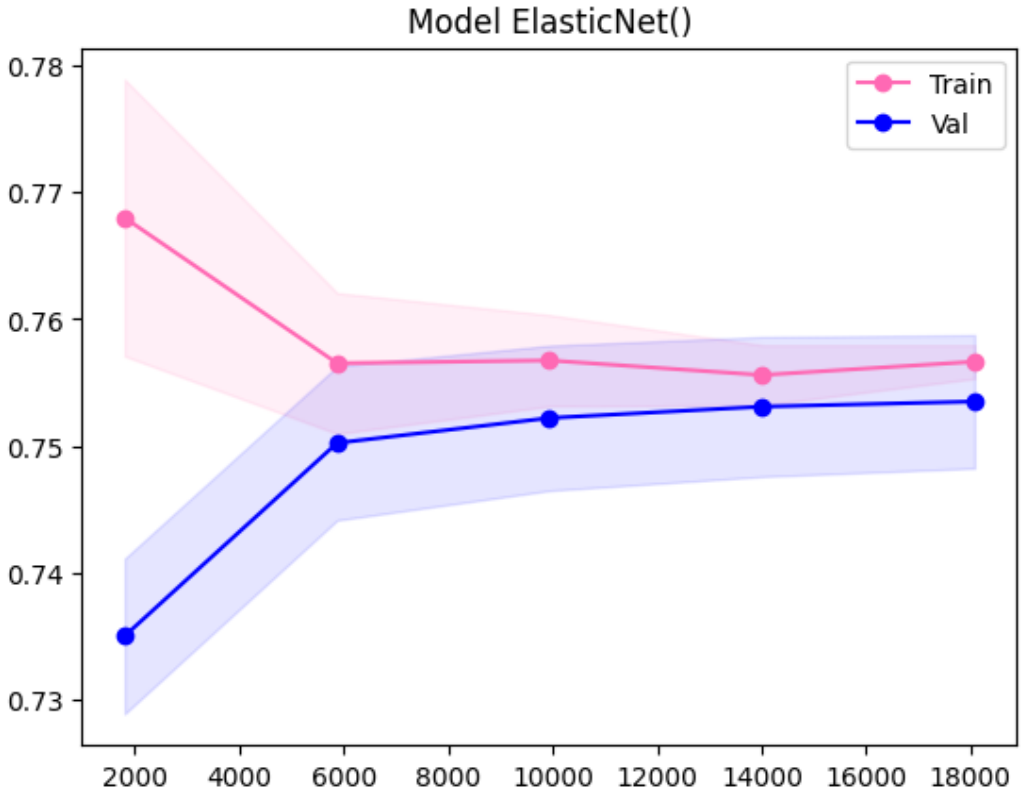


Figure 23. Learning curve showing the performance of the ElasticNet() model during training (pink) and validation (blue) stages. The shaded areas represent the standard deviation of the scores, while the solid lines represent the average scores.

Crop	R-squared	MAE	MAE%
Cassava	0.29	60626.70	16.24%
Maize	-0.03	20625.98	16.44%
Plantains and others	0.34	40340.70	10.16%
Potatoes	0.56	49413.80	10.47%
Rice, paddy	-0.08	15506.72	15.61%

Crop	R-squared	MAE	MAE%
Sorghum	-1.09	17268.22	22.90%
Soybeans	-11.64	20384.09	49.23%
Sweet potatoes	0.42	38919.12	11.50%
Wheat	-1.31	20572.88	21.38%
Yams	0.62	28941.00	13.55%

Table 8. Performance metrics for the ElasticNet() model, for individual crops: Cassava, Maize, Plantains and others, Potatoes, Rice (paddy), Sorghum, Soybeans, Sweet potatoes, Wheat, and Yams. Each row in the table provides the R-squared value (R^2), Mean Absolute Error (MAE) and MAE (%) for the respective crop.

These results provide an overview of the performance of different crops in terms of regression metrics. Yams have the highest R-squared value of 0.62, indicating a relatively good fit for predicting yam crop yield. On the other hand, crops like sorghum, soybeans, wheat, and rice have negative R-squared values, suggesting that the models are not able to effectively capture the variance in their yields. Additionally, soybeans have the highest MAE% of 49.23%, indicating a relatively large prediction error compared to the mean yield. Plantains and others have the lowest MAE% of 10.16%.

XGBoost

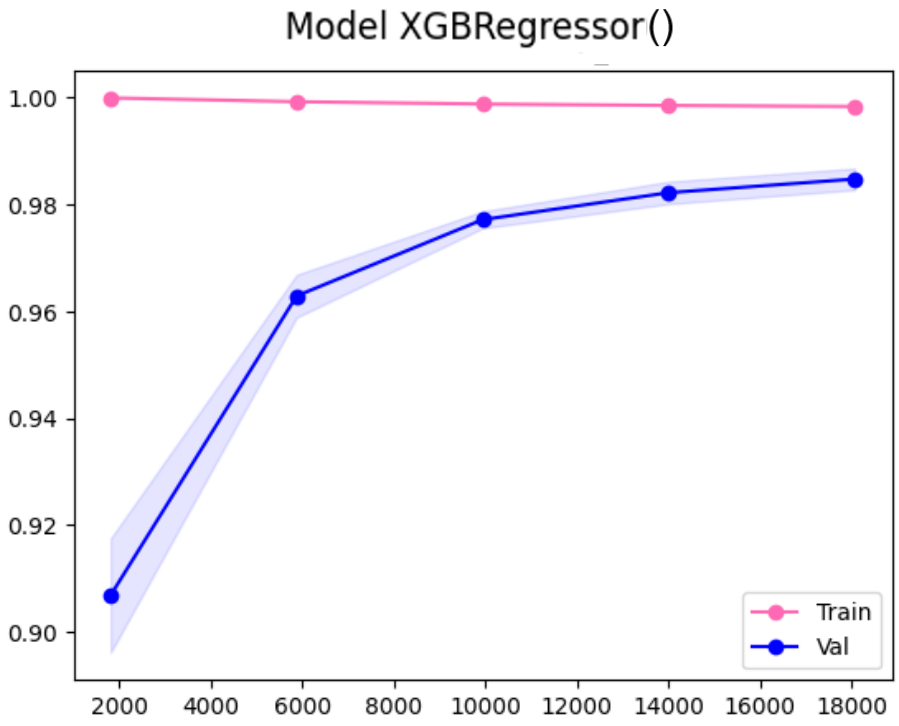


Figure 24. Learning curve showing the performance of the XGBRegressor() model during training (pink) and validation (blue) stages. The shaded areas represent the standard deviation of the scores, while the solid lines represent the average scores.

Crop	R-squared	MAE	MAE%
Cassava	0.98	6396.96	1.71%
Maize	0.96	3136.46	2.50%
Plantains and others	0.91	10929.70	2.75%
Potatoes	0.96	10311.71	2.18%

Crop	R-squared	MAE	MAE%
Rice, paddy	0.95	2938.25	2.96%
Sorghum	0.91	2706.50	3.59%
Soybeans	0.88	1861.61	4.50%
Sweet potatoes	0.96	6537.89	1.93%
Wheat	0.95	2584.70	2.69%
Yams	0.98	4740.12	2.22%

Table 9. Performance metrics for the XGBRegressor() model, for individual crops: Cassava, Maize, Plantains and others, Potatoes, Rice (paddy), Sorghum, Soybeans, Sweet potatoes, Wheat, and Yams. Each row in the table provides the R-squared value (R^2), Mean Absolute Error (MAE) and MAE (%) for the respective crop.

These results provide an overview of the performance of different crops in terms of regression metrics. Cassava and Yams exhibit high R-squared values of 0.98, indicating that the models explain a substantial amount of variance in their yields. On the other hand, Soybeans have a relatively lower R-squared value of 0.88, suggesting that the model may not capture as much variability in the crop yield. Additionally, Soybeans have the lowest MAE of 1861.61, followed by Wheat with 2584.70. Higher MAE values are observed for crops like Plantains and others and Yams, indicating larger prediction errors.

SGD

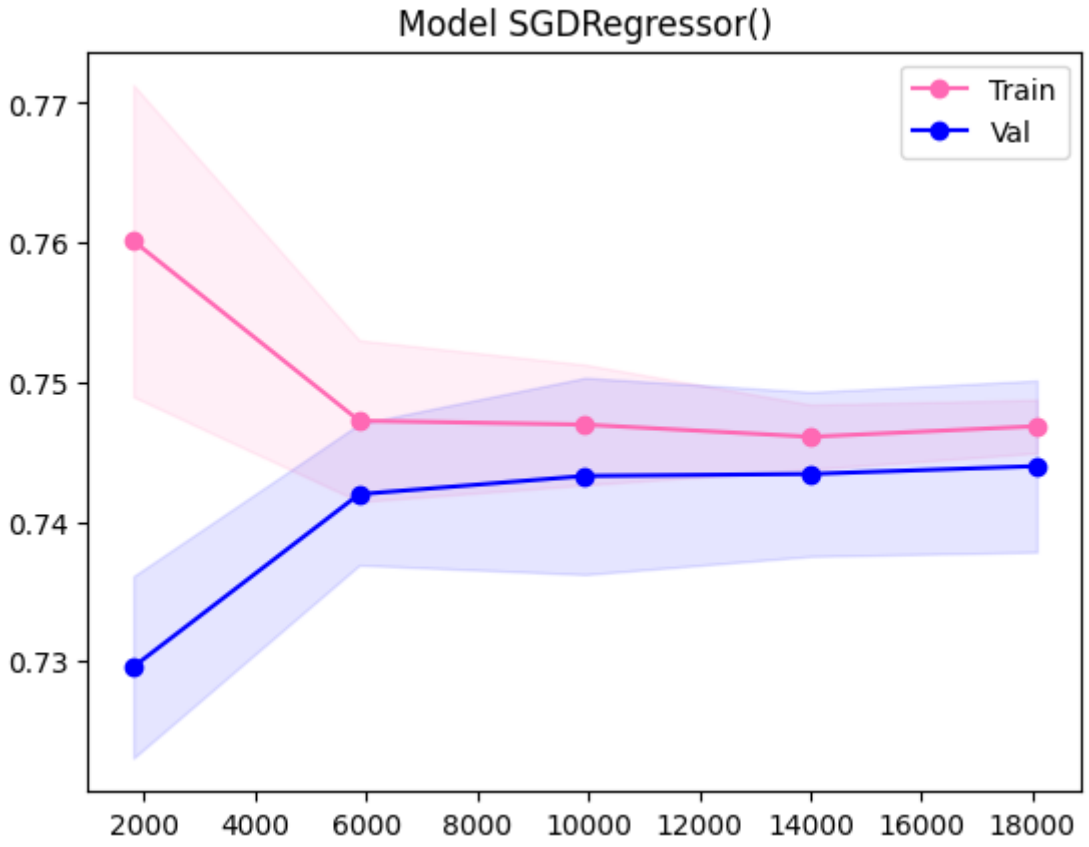


Figure 25. Learning curve showing the performance of the SGDRegressor() model during training (pink) and validation (blue) stages. The shaded areas represent the standard deviation of the scores, while the solid lines represent the average scores.

Crop	R-squared	MAE	MAE%
Cassava	0.22	59075.06	15.82%
Maize	0.04	19362.39	15.43%
Plantains and others	0.36	44042.39	11.09%
Potatoes	0.53	49710.78	10.53%
Rice, paddy	0.08	14402.89	14.50%
Sorghum	-0.75	16069.20	21.31%
Soybeans	-9.96	18604.54	44.93%
Sweet potatoes	0.37	40484.53	11.96%
Wheat	-0.90	18503.58	19.23%
Yams	0.62	28669.87	13.43%

Table 10. Performance metrics for the SGDRegressor() model, for individual crops: Cassava, Maize, Plantains and others, Potatoes, Rice (paddy), Sorghum, Soybeans, Sweet potatoes, Wheat, and Yams. Each row in the table provides the R-squared value (R^2), Mean Absolute Error (MAE) and MAE (%) for the respective crop.

These results provide an overview of the performance of different crops in terms of regression metrics. Potatoes, Plantains and others, and Yams have relatively higher R-squared values and lower MAE% values, indicating better predictive performance. On the other hand, Sorghum, Soybeans, and Wheat show poor performance with low or negative R-squared values and higher MAE% values, indicating significant prediction errors.

LGB

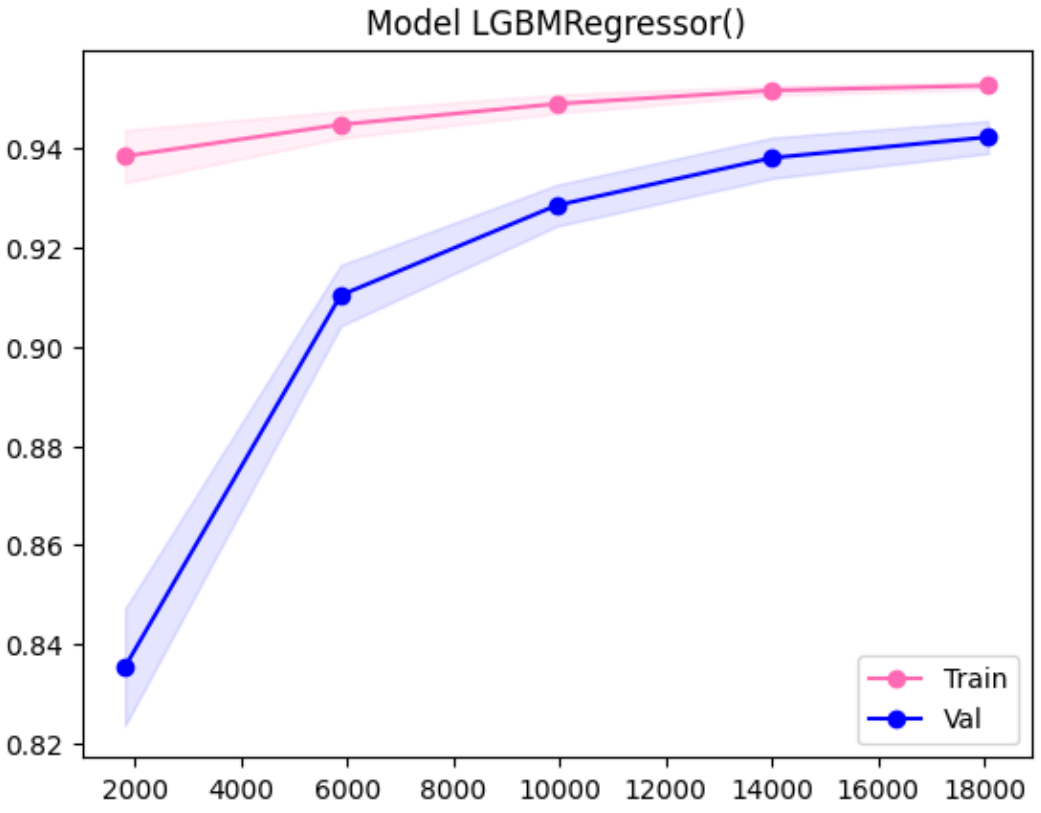


Figure 26. Learning curve showing the performance of the LGBMRegressor() model during training (pink) and validation (blue) stages. The shaded areas represent the standard deviation of the scores, while the solid lines represent the average scores.

Crop	R-squared	MAE	MAE%
Cassava	0.92	17181.56	4.60%
Maize	0.73	9449.19	7.53%
Plantains and others	0.69	25849.32	6.51%
Potatoes	0.87	24009.92	5.09%
Rice, paddy	0.71	7550.55	7.60%
Sorghum	0.73	6156.96	8.16%
Soybeans	0.23	4923.69	11.89%
Sweet potatoes	0.86	16522.38	4.88%
Wheat	0.74	6863.68	7.13%
Yams	0.87	13840.96	6.48%

Table 11. Performance metrics for the LGBMRegressor() model, for individual crops: Cassava, Maize, Plantains and others, Potatoes, Rice (paddy), Sorghum, Soybeans, Sweet potatoes, Wheat, and Yams. Each row in the table provides the R-squared value (R^2), Mean Absolute Error (MAE) and MAE (%) for the respective crop.

Cassava, Potatoes, and Yams consistently demonstrate strong performance with high R-squared values and relatively low MAE% values. On the other hand, Soybeans shows the weakest performance with the lowest R-squared value and highest MAE% value. The rest (Maize, Rice, and Wheat) perform moderately well with reasonably high R-squared values and moderate MAE% values.

All Models

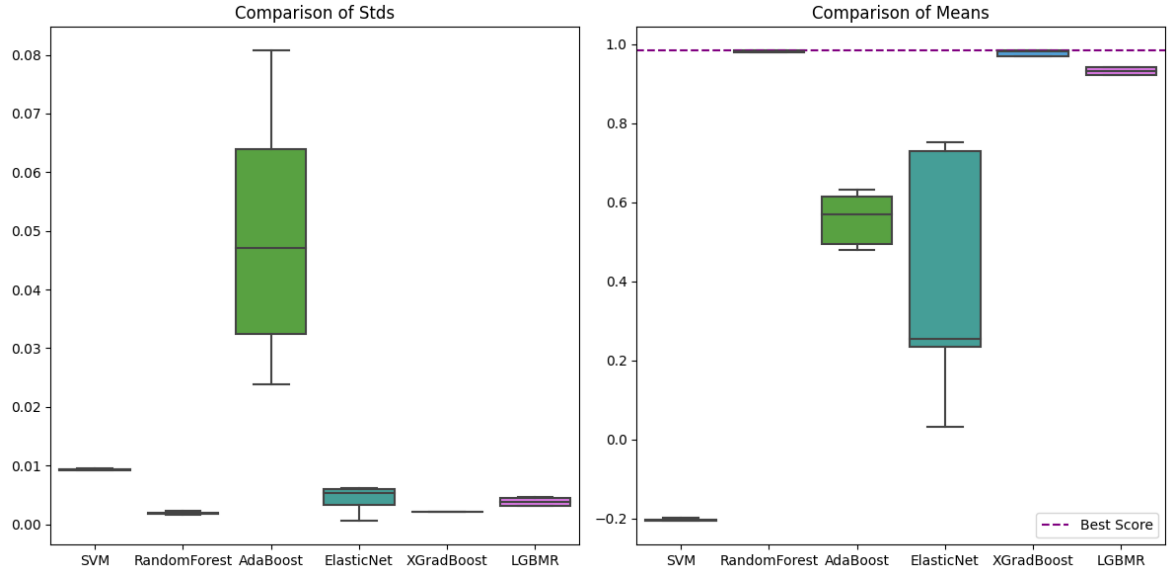


Figure 27. Boxplot comparison of standard deviations and means for different models. On the LEFT is displayed the standard deviations of the model, while on the RIGHT the boxplot compares the means of the models. Each box represents the distribution of that metric for a specific model. A horizontal purple line represents the best score achieved across all models.

Appendix B: Code Repository

In this appendix, you will find the link to the code repository used for the implementation of the data analysis algorithms. The code repository contains the Python scripts used to analyse, visualize and pre-process the data, perform model comparison and selection, perform model training and testing, as well as hyperparameter tuning.

Code Repository Link: https://github.com/saradiazdelser/application-of-machine-learning-algorithms-to-perform-crop-yield-prediction

Please note that the code repository may be updated in the future, and any major revisions will be documented in the repository's README file.