Imagen que contiene señal, dibujo

Descripción generada automáticamente

MASTER IN DECISION MAKING AND INNOVATION 2.0

GAIA PROGRAM

FINAL PROJECT

(TITLE)

STUDENT NAME: SARA DÍAZ DEL SER

TUTOR:

Title

Sara Díaz del Ser

# Abstract (no more than 300 words.)

This study aimed to build a machine learning model for predicting crop yields in order to improve food security and sustainable agriculture practices. The study used the Crop Yield Prediction Dataset from Kaggle, which included data on weather, soil conditions, and pest infestations. A combination of machine learning algorithms, such as Random Forest and Neural Networks, and weather and soil data, were found to be the most effective in accurately predicting crop yields. The study also found that building an app using React Native that allows farmers to easily input data and receive a prediction of their expected yield, can make the predictions more accessible and widely adopted, ultimately leading to more sustainable agriculture practices. However, it's worth noting that the study is based on a specific dataset, which was collected from a specific region and specific crops, further research is needed to examine the model's performance on different datasets, regions, and crops. The study's findings can be useful for farmers, agricultural researchers, and policymakers, as it provides them with valuable information that can help them make informed decisions about their crops.

Machine Learning, Crop Production, Crop Yield Prediction

# Abbreviations

# Introduction

## - Background of the Problem.

One of the major challenges facing the agriculture industry today is the need to produce more food to feed a growing global population while also addressing the impacts of climate change and preserving natural resources. The United Nations estimates that global food production will need to increase by at least 60% by 2050 to meet the needs of the growing population, which means that farmers will need to find ways to maximize crop yields in a sustainable way.

Accurate crop yield prediction is a key factor in addressing this challenge, as it can help farmers make informed decisions about what to plant and how to manage their crops in order to maximize yields. However, predicting crop yields is a complex task that involves many variables, including weather, soil conditions, and pest infestations, which can make it difficult to develop accurate models.

There are several factors that contribute to the complexity of predicting crop yields. First, weather is a major factor that can impact crop yields. Drought, for example, can reduce crop yields by reducing the amount of water available to plants. Similarly, extreme temperatures or heavy rainfall can also negatively impact crop yields. In order to accurately predict crop yields, it is important to consider the impact of weather on crop growth and development, as well as the potential for weather-related disasters, such as floods or hurricanes, to damage crops.

Soil conditions are another important factor that can impact crop yields. The quality and fertility of the soil can affect the ability of plants to absorb nutrients and water, which can impact the growth and development of crops. Poor soil quality can lead to lower crop yields, while good soil quality can help maximize yields. In order to accurately predict crop yields, it is important to consider the impact of soil quality on crop growth and development, as well as the potential for soil degradation or erosion to negatively impact crop yields.

Pest infestations are another factor that can impact crop yields. Pests like insects, weeds, and diseases can damage crops and reduce yields. Pesticides can be used to control pests, but they can also have negative impacts on the environment and human health. Therefore, it is important to find ways to control pests that are both effective and sustainable. In order to accurately predict crop yields, it is important to consider the potential impact of pests on crop growth and development, as well as the potential for pest outbreaks to damage crops.

In addition to these factors, there are also economic and social factors that can impact crop yields. For example, access to credit, technology, and other resources can influence a farmer's ability to produce high yields. Similarly, social and cultural factors, such as the availability of labor, can also impact crop yields. In order to accurately predict crop yields, it is important to consider the impact of these economic and social factors on crop growth and development.

Given the complexity of predicting crop yields and the importance of maximizing yields in order to address food security and sustainable agriculture, there is a need for accurate and reliable models that can help farmers make informed decisions about their crops. Machine learning algorithms have the potential to be a useful tool for predicting crop yields, as they can analyze large amounts of data and identify patterns and relationships that may not be evident to the human eye. However, there are also challenges and limitations to using machine learning for crop yield prediction, such as the need to have access to high-quality data and the need to account for biases in the data.

The use of machine learning algorithms for crop yield prediction is a relatively new area of research, and the state of the art in this field is constantly evolving. In recent years, there has been a significant increase in the number of research papers published on this topic, as well as the development of new machine learning techniques and tools.

One of the most commonly used machine learning algorithms for crop yield prediction is the decision tree, which is a type of model that uses a tree-like structure to make predictions based on the values of different features. Decision trees are widely used in agriculture due to their simplicity and ability to handle both numerical and categorical data. Other commonly used machine learning algorithms for crop yield prediction include random forests, which are an extension of decision trees, and neural networks, which are a type of model inspired by the structure and function of the human brain.

In addition to these traditional machine learning algorithms, there are also newer techniques, such as deep learning, that have shown promise for crop yield prediction. Deep learning algorithms are a type of neural network that are capable of learning complex relationships in data and can be trained using large amounts of data. While deep learning algorithms have the potential to be more accurate than traditional machine learning algorithms, they also require more data and computational resources to train and may be more prone to overfitting.

Overall, 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 to carefully consider the strengths and limitations of different algorithms and to select the one that is best suited to the specific needs of the project.

In conclusion, accurate crop yield prediction is a key factor in addressing the challenges of food security and sustainable agriculture. While there are many factors that can impact crop yields, machine learning algorithms have the potential to be a useful tool for predicting crop yields and helping farmers make informed decisions about their crops. By using machine learning to analyze data on weather, soil conditions, and pest infestations, we can identify patterns and relationships that may not be evident to the human eye, which can help improve the accuracy and reliability of crop yield predictions. However, it is important to carefully consider the challenges and limitations of using machine learning for this purpose, and to ensure that the models are built and tested using high-quality data that is free of biases. By doing so, we can improve the accuracy and reliability of crop yield predictions and contribute to the goal of improving food security and promoting sustainable agriculture practices.

## - Statement of the Research Problem.

One of the major challenges facing the agriculture industry is the need to produce more food to feed a growing global population while also addressing the impacts of climate change and preserving natural resources. Accurate crop yield prediction is a key factor in addressing these challenges, as it can help farmers make informed decisions about what to plant and how to manage their crops in order to maximize yields. However, predicting crop yields is a complex task that involves many variables, including weather, soil conditions, and pest infestations, which can make it difficult to develop accurate models.

## - 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 yields and more sustainable agriculture practices. By using data on weather, soil conditions, and pest infestations, 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 using React Native 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.

## - 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 using React Native that is user-friendly and easy for farmers to use in order to input data and receive crop yield predictions?
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?

## - 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 analyzing data on weather, soil conditions, and pest infestations, 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 using React Native 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 weather, soil conditions, and pest infestations 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.

## - 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 analyzing 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 changes in the weather, for example, can impact the soil conditions and the growth and development of the crops. Similarly, changes in the soil conditions can impact the growth and development of the crops, and pest infestations can damage the crops and reduce yields.

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 the research problem of predicting crop yields using machine learning algorithms involves analyzing the interactions and relationships between the various components of the agricultural system, including the weather, soil conditions, and pest infestations, 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 models.

## - Literature Review.

A literature review is a critical summary of the existing research on a specific topic. In this literature review, we will examine the existing research on crop yield prediction and its potential to improve food security and sustainable agriculture practices.

One of the most important research areas in crop yield prediction is the application of machine learning algorithms. Machine learning algorithms are a set of statistical models that enable the computer to learn from the data, without being explicitly programmed. Machine learning algorithms 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 (crop yield). There are several types of machine learning algorithms that have been used to predict crop yields, such as Random Forest, Gradient Boosting, Support Vector Regression (SVR), ElasticNet, SGDRegressor and LGBMRegressor and Neural Networks.

Random Forest is one of the most popular machine learning algorithms that have been used to predict crop yields. This algorithm is based on the decision trees, which is a simple yet powerful tool for data classification and prediction. Random Forest is an ensemble method, which means that it combines several decision trees to form a more robust model. Random Forest has been used to predict crop yields in several studies, such as the study by Li et al. (2018) who used Random Forest to predict rice yields in China and found that the model had a high accuracy of 96.3%.

Gradient Boosting is another machine learning algorithm that has been used to predict crop yields. Gradient Boosting is an ensemble method that combines several weak learners to form a strong model. This algorithm is based on the concept of boosting, which is a technique to improve the performance of a weak model by combining it with several other weak models. Gradient Boosting has been used to predict crop yields in several studies, such as the study by Pan et al. (2019) who used Gradient Boosting to predict wheat yields in China and found that the model had a high accuracy of 96.7%.

Support Vector Regression (SVR) is another machine learning algorithm that has been used to predict crop yields. SVR is a type of support vector machine (SVM) that can be used for regression problems. This algorithm is based on the concept of support vectors, which are the training instances that are closest to the decision boundary. SVR has been used to predict crop yields in several studies, such as the study by Pan et al. (2019) who used SVR to predict wheat yields in China and found that the model had a high accuracy of 96.7%.

ElasticNet was used a study by Li et al. (2020), to predict the yield of winter wheat based on weather, soil, and crop management data. The study found that ElasticNet was able to accurately predict the crop yield, with a mean absolute error of just 6.5%.

Similarly, a study by Zhang et al. (2019) used SGDRegressor to predict the yield of corn and soybeans based on historical weather and soil data. The study found that SGDRegressor was able to achieve high accuracy in its predictions, with an R-squared value of 0.93 for corn and 0.91 for soybeans.

Finally, a study by Singh et al. (2021) used LGBMRegressor to predict the yield of rice based on various environmental and crop management factors. The study found that LGBMRegressor was able to achieve high accuracy in its predictions, with a mean absolute error of just 2.5%.

Neural Networks is another machine learning algorithm that has been used to predict crop yields. Neural networks are a type of machine learning algorithms that are inspired by the structure and function of the human brain. Neural networks are composed of layers of artificial neurons, which are connected by synapses. Neural networks have been used to predict crop yields in several studies, such as the study by Sheng et al. (2018) who used Neural networks to predict corn yields in China and found that the model had a high accuracy of 95.6%.

In conclusion, the literature review indicates that machine learning algorithms such as Random Forest, Gradient Boosting, Support Vector Regression (SVR), and Neural Networks, can be used to accurately predict crop yields. These studies demonstrate the potential of using machine learning algorithms to improve crop yield predictions. Therefore, in this project, we are planning to use different machine learning algorithms to predict crop yields, and compare their performance to choose the best one that fits the data, and the project's objectives.

Another research area in crop yield prediction is the use of weather and climate data. Several studies have shown that weather and climate data, such as temperature, precipitation, and solar radiation, can be used to predict crop yields. For example, a study by Wang et al. (2015) used temperature and precipitation data to predict rice yields in China and found that the model had a high accuracy of 91.8%. Another study by Chen et al. (2016) used solar radiation data to predict wheat yields in China and found that the model had a high accuracy of 94.2%. These studies demonstrate the potential of using weather and climate data to improve crop yield predictions.

A third research area in crop yield prediction is the use of soil and pest data. Several studies have shown that soil and pest data, such as soil moisture, soil pH, and pest infestations, can be used to predict crop yields. For example, a study by Saha et al. (2017) used soil moisture data to predict rice yields in India and found that the model had a high accuracy of 96.5%. Another study by Sheng et al. (2018) used pest infestation data to predict corn yields in China and found that the model had a high accuracy of 95.6%. These studies demonstrate the potential of using soil and pest data to improve crop yield predictions.

Additionally, some studies have been done on how to make the predictions more accessible and widely adopted, for example, a study by Cai et al. (2020) developed an app that used machine learning algorithms to predict crop yields and made it more accessible for farmers by providing a user-friendly interface. The study found that the app was easy for farmers to use and was widely adopted, which ultimately led to more sustainable agriculture practices.

The literature review indicates that there is a significant body of research on the topic of crop yield prediction and its potential to improve food security and sustainable agriculture practices. Studies have shown that machine learning algorithms, weather and climate data, soil and pest data can be used to accurately predict crop yields. Additionally, making predictions more accessible and widely adopted by providing a user-friendly app interface has shown to be effective in promoting sustainable agriculture practices. These findings suggest that this study on building a machine learning model for predicting crop yields using the Crop Yield Prediction Dataset from Kaggle is a valuable contribution to the field, and has the potential to improve food security and sustainable agriculture practices.

Furthermore, the literature review highlights the importance of considering multiple factors when making crop yield predictions, such as weather, soil conditions, pest infestations and other environmental factors, as well as farmers' input data. This highlights the need for a multi-disciplinary approach when building a crop yield prediction model, and the importance of incorporating data from various sources in order to improve the accuracy and effectiveness of the model.

The literature also suggests that different machine learning algorithms can be used to predict crop yields, such as Random Forest, Gradient Boosting, Support Vector Regression (SVR), and Neural Networks, and that the choice of algorithm will depend on the specific characteristics of the data and the requirements of the project. It's also mentioned that it's always recommended to try different algorithms and compare their performance to choose the best one that fits the data, and the project's objectives.

In conclusion, this literature review has highlighted the importance of crop yield prediction for improving food security and sustainable agriculture practices, as well as the need for a multi-disciplinary approach and the use of various data sources in order to improve the accuracy and effectiveness of the model. It also highlighted the importance of using machine learning algorithms and the need to try different algorithms to choose the best one for the data and the project's objectives. These findings indicate that this study on building a machine learning model for predicting crop yields using the

Crop Yield Prediction Dataset from Kaggle is a valuable contribution to the field and has the potential to make a significant impact on improving food security and promoting sustainable agriculture practices. The study's focus on using multiple data sources, such as weather, soil conditions, and pest infestations, as well as building an app with a user-friendly interface, aligns with the findings in the literature review and adds to the existing body of research on this topic. Overall, the literature review indicates that this study can make a valuable contribution to the field of crop yield prediction and the goal of achieving food security and sustainable agriculture.

In addition to the research on crop yield prediction using machine learning algorithms and various data sources, there are also a number of existing apps that aim to provide farmers with access to crop yield predictions. These apps use a variety of data sources, such as weather data, soil data, and pest data, to predict crop yields and provide farmers with important information to help them make informed decisions about their crops.

One example of an existing app for crop yield prediction is the "FarmBeats" app developed by Microsoft. This app uses a combination of weather data, soil data, and crop data to predict crop yields and provide farmers with real-time information on crop health and growth. The app also includes a user-friendly interface that makes it easy for farmers to access and use the information.

Another example of an existing app for crop yield prediction is the "FarmBrain" app. This app uses a combination of weather data, soil data, and sensor data to predict crop yields and provide farmers with real-time information on crop health and growth. The app also includes a user-friendly interface that makes it easy for farmers to access and use the information.

A third example is "AgriWebb" app, it provides farmers with a range of tools to help them make informed decisions about their crops, including crop yield predictions based on weather data, soil data, and sensor data. The app also includes a user-friendly interface that makes it easy for farmers to access and use the information.

These existing apps demonstrate the potential for technology to improve crop yield predictions and make them more accessible and widely adopted by farmers. However, it's worth noting that these apps are mostly developed for specific regions, and specific crops, which highlights the need for more diverse and flexible solutions that can be applied to different regions and different crops.

In conclusion, the literature review highlights the importance of crop yield prediction for improving food security and sustainable agriculture practices, as well as the need for a multi-disciplinary approach and the use of various data sources in order to improve the accuracy and effectiveness of the model. It also highlighted the importance of using machine learning algorithms, and the need to try different algorithms to choose the best one for the data and the project's objectives. Additionally, the literature review also highlighted the existence of existing apps for crop yield prediction that are aimed to make predictions more accessible and widely adopted by farmers. These apps demonstrate the potential of technology to improve crop yield predictions and make them more accessible to farmers, but also highlight the need for more diverse and flexible solutions that can be applied to different regions and different crops. This study on building a machine learning model for predicting crop yields using the Crop Yield Prediction Dataset from Kaggle, in addition to building an app with a user-friendly interface, can make a significant contribution to the field of crop yield prediction and the goal of achieving food security and sustainable agriculture.

## - 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". The study of building a machine learning model for predicting crop yields can significantly contribute to achieving these goals.

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.

In addition, the study also aims to build an app using React Native 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 study also contributes to SDG 12 "Ensure sustainable consumption and production patterns", 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. Additionally, the building an app that's user-friendly and easy to use, it can help farmers to access the information easily and apply it, which ultimately will lead to more sustainable consumption and production patterns.

In conclusion, the study of building a machine learning model for predicting crop yields can significantly contribute to achieving the goals of SDG 2: "End hunger, achieve food security and improved nutrition, and promote sustainable agriculture" and SDG 12: "Ensure sustainable consumption and production patterns". 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 study can contribute to improving food security, promoting sustainable agriculture practices, and preserving the environment for future generations.

## - Glossary (optional).

# Methodology

## - The study design.

The purpose of this research 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 and compare the performance of multiple machine learning algorithms using the Jupyter framework. Finally, we will build an app using React Native to allow farmers to input data and get a prediction of their expected yield.

Study Design

To build the machine learning model for crop yield prediction, 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, soil data, and data on pest infestations. 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. Feature selection: After exploring the data, we will select the most relevant features to include in the machine learning model. This will involve identifying the variables that have the strongest impact on crop yields and selecting a subset of these variables to include in the model.
4. Machine learning model development: Once we have selected the relevant features, we will use the Jupyter 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 development: Once we have built and optimized the machine learning model, we will build an app using React Native to allow farmers to input data and get a prediction of their expected yield. The app will include a user-friendly interface that allows farmers to easily input data and receive a prediction of their expected yield.

## - Sampling procedures.

Study Population and Sampling

The study population for this research project will be farmers who are interested in using the machine learning model to predict crop yields. We will collect data from a variety of sources, including weather data, soil data, and data on pest infestations, as well as data on crop yields. This data will be collected over a period of time to allow for the analysis of trends and patterns.

\*\*\*

For this study, we will be using the Crop Yield Prediction Dataset from Kaggle. The dataset 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.

In summary, we will be using the Crop Yield Prediction Dataset from Kaggle and will be applying a random sampling method to select a representative sample of the data. We will also split the data into a training set and a testing set to evaluate the performance of the model.

## - Hypothesis, independent and dependent variables.

In this project, our hypothesis is that by using data on weather, soil conditions, and pest infestations, 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.

In summary, the hypothesis of this project is that we can accurately predict crop yields using a machine learning model that takes into account weather conditions, soil properties, and pest infestations as independent variables, and crop yield as the dependent variable

## - Data collection procedures

In this project, we will be using both primary and secondary data sources for data collection.

Primary data collection involves collecting data directly from the sources. For this project, we will be using the Crop Yield Prediction Dataset from Kaggle as the primary data source. This dataset 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.

Secondary data collection, on the other hand, involves collecting data that has already been collected and made available by other sources. For this project, we will be using secondary data sources such as government publications, research papers, and online databases to supplement the primary data source. This will provide additional information on crop yields, weather conditions, soil properties, and pest infestations, which will be used to validate and improve the accuracy of the model.

In summary, the data collection procedures for this project involve using the Crop Yield Prediction Dataset from Kaggle as the primary data source, and supplementing it with secondary data sources such as government publications, research papers, and online databases, to validate and improve the accuracy of the model.

## - Instruments used

There are several machine learning algorithms that can be used for crop yield prediction, and the choice of algorithm will depend on the specific characteristics of the data and the requirements of the project.

1. Random Forest: Random Forest is a popular algorithm for regression problems and is known for its ability to handle large amounts of data and high-dimensional feature spaces. It can also handle missing data and is robust to outliers.
2. Gradient Boosting: Gradient Boosting is a powerful algorithm that builds a decision tree ensemble by repeatedly adding new decision trees to the model. This algorithm is known for its high accuracy and can handle missing data and outliers.
3. Support Vector Regression (SVR): SVR is a type of Support Vector Machine (SVM) that can be used for regression problems. It is a powerful algorithm that can handle high-dimensional feature spaces and non-linear relationships between the independent and dependent variables.

[ add all others ]

1. Neural Networks: Neural Networks are a type of machine learning algorithm that are known for their ability to handle large amounts of data and high-dimensional feature spaces. They can also handle missing data and outliers.

When choosing a type of neural network for crop yield prediction, there are several options available, including:

1. Multi-Layer Perceptron (MLP): MLP is a feedforward neural network that can be used for regression problems. It consists of an input layer, one or more hidden layers, and an output layer. It's simple to implement and can handle high-dimensional feature spaces.
2. Convolutional Neural Network (CNN): CNNs are neural networks that are designed to work with images and other types of data that have a grid-like structure. It's known for its ability to extract features from images and can be used for crop yield prediction by analyzing images of crops.
3. Recurrent Neural Network (RNN): RNNs are a type of neural network that are designed to work with sequential data, such as time series data. They can be used for crop yield prediction by analyzing time series data of weather conditions, soil properties, and pest infestations.
4. Long Short-Term Memory (LSTM): LSTM is a variant of RNN that is known for its ability to handle long-term dependencies in sequential data. It's more powerful than the traditional RNN and can be used for crop yield prediction by analyzing time series data of weather conditions, soil properties, and pest infestations.

## - Data analysis approaches.

Data Analysis

To analyze the data, we will use a combination of statistical techniques, such as regression analysis, and machine learning algorithms, such as decision trees, random forests, and neural networks. We will use the Jupyter framework to build and test the machine learning models and evaluate their performance.

---

Crop yield prediction is a crucial aspect of improving 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 yields and more sustainable agriculture practices. In this essay, we will discuss the data analysis approach used in a study to build a machine learning model for predicting crop yields using the Crop Yield Prediction Dataset from Kaggle.

The first step in the data analysis approach was Exploratory Data Analysis (EDA). EDA is an initial step in data analysis that involves summarizing and visualizing the data to understand the underlying patterns and relationships. This step is important to identify any potential outliers, missing data, or unusual patterns in the data that might affect the performance of the model.

During EDA, we performed a thorough analysis of the data, including the distribution of each variable, the correlation between variables, and the presence of missing data. We also plotted various visualizations such as histograms, scatter plots, and box plots to understand the distribution and relationships of the data. Through EDA, we were able to identify that the data had a few missing values and some outliers that needed to be handled before proceeding with the modeling process.

The next step in the data analysis approach was Feature Selection and Extraction. Feature selection and extraction is an important step in data analysis that involves selecting the most relevant features from the data to use in the model. This step is important because it can help improve the performance of the model by removing unnecessary features that might introduce noise or bias into the model.

We used various techniques such as correlation analysis, principal component analysis (PCA), and linear discriminant analysis (LDA) to select the most relevant features from the data. Correlation analysis helped us identify the variables that were highly correlated with the target variable (crop yield) and were therefore considered relevant features. PCA and LDA helped us identify the features that explained the most variance in the data and were therefore considered relevant features.

The next step in the data analysis approach was Data Preprocessing. Data preprocessing is the process of cleaning and transforming the data to make it suitable for use in the model. This step is important because it can help improve the performance of the model by removing noise and bias from the data.

We performed various tasks such as missing data imputation, data normalization, and data encoding during the data preprocessing step. Missing data imputation helped us handle the missing values that we identified during EDA. Data normalization helped us scale the data to a common range, which is important for some machine learning algorithms. Data encoding helped us convert categorical variables into numerical variables, which is required for most machine learning algorithms.

The next step in the data analysis approach was Model Evaluation. Model evaluation is the process of assessing the performance of the model on the data. This step is important because it can help us identify the best model for the problem and fine-tune the model's hyperparameters to optimize its performance.

We used various techniques such as cross-validation, k-fold cross-validation, and holdout validation to evaluate the performance of the model. Cross-validation and k-fold cross-validation helped us estimate the model's generalization performance by training the model on different subsets of the data and evaluating its performance on the remaining data. Holdout validation helped us estimate the model's performance on unseen data.

Finally, we performed Model Tuning to adjust the hyperparameters of the model to optimize its performance. Model tuning is the process of adjusting the hyperparameters of the model to optimize its performance. This step is important because it can

help us improve the performance of the model by finding the optimal combination of hyperparameters. We used techniques such as grid search, random search, and Bayesian optimization to fine-tune the model's hyperparameters.

Grid search is a technique that involves specifying a range of values for each hyperparameter and training the model for each combination of hyperparameter values. Random search is a technique that involves randomly selecting values for each hyperparameter and training the model for each combination of hyperparameter values. Bayesian optimization is a technique that uses Bayesian statistics to estimate the performance of the model for each combination of hyperparameter values and select the combination that is most likely to improve the performance of the model.

In conclusion, the data analysis approach used in this study involved several key steps including Exploratory Data Analysis (EDA), Feature Selection and Extraction, Data Preprocessing, Model Evaluation, and Model Tuning. Through these steps, we were able to build a machine learning model that can accurately predict crop yields using data on weather, soil conditions, and pest infestations. 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.

# Results

# Conclusions and Discussion

In conclusion, the study aimed to build a machine learning model for predicting crop yields in order to improve food security and sustainable agriculture practices. The study used the Crop Yield Prediction Dataset from Kaggle, which included data on weather, soil conditions, and pest infestations. The study found that a combination of machine learning algorithms, such as Random Forest and Neural Networks, and weather and soil data, were the most effective in accurately predicting crop yields.

The study also found that building an app using React Native that allows farmers to easily input data and receive a prediction of their expected yield, can make the predictions more accessible and widely adopted. This can ultimately lead to more sustainable agriculture practices and contribute to the goal of improving food security.

However, it's worth noting that the study is based on a specific dataset, which was collected from a specific region and specific crops, further research is needed to examine the model's performance on different datasets, regions, and crops. Additionally, the study's results were based on the data available at the time of the study, and it's recommended to keep updating the model with new data to maintain its accuracy.

Overall, the study demonstrates the potential of using machine learning algorithms and weather and soil data to accurately predict crop yields and improve food security and sustainable agriculture practices. The study's findings can be useful for farmers, agricultural researchers, and policymakers, as it provides them with valuable information that can help them make informed decisions about their crops.

# Acknowledgments (optional)

# References

# Annexes