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.)

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 was also developed, which will make 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

# Introduction

## - Background of the Problem.

The agriculture industry faces a daunting task in the coming years: producing enough food to meet the demands of over 8 billion human beings. 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 [[reference](https://www.un.org/en/chronicle/article/feeding-world-sustainably)]. This of course, while also tackling climate change's effects on the industry and preserving the planet’s natural resources.

[add graph: predicted global food consumption]

Exploring sustainable techniques for maximizing crop production is crucial in order to overcome this challenge. Successfully predicting crop yield is an important first step. 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 in many ways, mostly because of a number of 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 are a set of statistical models that enable the computer to learn from the data, without being explicitly programmed. 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, but these models have limitations, such as a need for high-quality data and the devastating possibility of bias.

Even as machine learning techniques and tools are constantly have seem incredible improvement in recent years, the subject of crop yield prediction using machine learning is still relatively young and continually changing.

Among the most commonly used machine learning algorithms in agriculture are decision trees, which use a tree-like structure to make predictions based on different features. They are widely used due to their simplicity and ability to handle both numerical and categorical data. Additionally, random forests, an extension of decision trees, and neural networks, a model inspired by the structure and function of the human brain, are also commonly used in crop yield prediction.

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 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 different algorithms and select the best approach for each specific need.

The ability to accurately predict crop yields is crucial for tackling the challenges of food security and sustainable agriculture. Machine learning algorithms have emerged as a potentially valuable tool for predicting crop yields and aiding farmers in making informed decisions about their crops.

These algorithms are capable of analysing data about temperature, rainfall, location and pesticide use, to detect hidden patterns and relationships that may not be visible to the human eye. This can enhance the precision and dependability of crop yield predictions.

Nevertheless, it is vital to thoroughly consider the challenges and limitations associated with machine learning for this purpose and ensure that the models are constructed and validated using high-quality, unbiased data. By doing so, we can enhance the accuracy and reliability of crop yield predictions, ultimately advancing the goal of promoting sustainable agricultural practices and improving food security.

## - 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.

## - 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.

## - 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?

## - 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.

## - 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 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 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 models.

## - Literature Review.

The agricultural sector faces various challenges, including the need to produce more food to feed a growing population while simultaneously 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. Machine learning (ML) algorithms have emerged as a potentially valuable tool for predicting crop yields and aiding farmers in making informed decisions about their crops. These algorithms are capable of analysing data on temperature, rainfall, location and pesticide use to detect hidden patterns and relationships that may not be visible to the human eye. In this literature review, we aim to explore the use of machine learning techniques for crop yield prediction and examine the current state of research in this field. We will also examine the challenges and limitations associated with using machine learning for crop yield prediction and identify potential areas for future research.

[ add industry trends ]

[ delete this: ?]

Most current 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, compare their performance, and select the one that best fits the data and the project's objectives.

However, there are also several challenges and limitations associated with machine learning algorithms that must be considered.

One of the primary challenges is the need for high-quality, unbiased data. Machine learning algorithms rely heavily on large datasets to make accurate predictions, and if the data used to train the model is incomplete or biased, it can lead to inaccurate predictions. Additionally, the quality of the data used can affect the reliability of the predictions. For instance, if the data used is outdated or does not account for recent changes in weather patterns or pest infestations, the model may not produce accurate results.

Another challenge associated with machine learning for crop yield prediction is the potential for overfitting. Machine learning algorithms can learn complex relationships within the data, but if the model becomes too complex, it may start to fit the noise in the data rather than the underlying patterns. This can lead to inaccurate predictions when the model is applied to new data.

Also, machine learning algorithms require significant computational resources to train, which can be a challenge for farmers or researchers with limited access to high-performance computing. This can limit the scalability of the models and make it difficult to apply them on a large scale.

Another challenge to mention is explainability, or­—more accurately—lack thereof. The interpretation of the results produced by machine learning models can be challenging. Unlike traditional statistical models, machine learning algorithms do not always produce a clear explanation for their predictions, which can make it difficult to identify the factors that are driving the predictions. This can limit the usefulness of the models in terms of informing decision-making.

To conclude this literature review, we can confirm that the development of machine learning models for crop yield prediction has the potential to significantly contribute to the improvement of food security and sustainable agriculture practices. By accurately predicting crop yields, farmers can make informed decisions about what to plant and how to manage their crops, leading to higher production and more sustainable agriculture practices.

Through this literature review, we have identified several machine learning algorithms commonly used in crop yield prediction, including decision trees, random forests, neural networks, and deep learning. We have also discussed the importance of high-quality, unbiased data in constructing and validating these models, as well as the challenges and limitations associated with using machine learning for crop yield prediction.

Moving forward, future research in this field should focus on improving the accuracy and reliability of these models, as well as making them more accessible to farmers. As well as exploring the use of additional data sources, such as satellite imagery and drone technology, to further enhance the precision and dependability of crop yield predictions. Additionally, there is a need for more research on the integration of machine learning models with precision agriculture techniques, such as variable rate application of fertilizers and pesticides, to optimize crop management practices and reduce environmental impact.

Overall, the potential benefits of machine learning for crop yield prediction are significant, and continued research in this field has the potential to contribute to the goal of promoting sustainable agricultural practices and improving food security.

## - 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.

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 "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, 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.

# Methodology

## - The 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 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, 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 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 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 development: Once we have built and optimized the machine learning model, we will build an app using the React Native framework 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.

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.

## - 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.

## - Data collection procedures

This project will make use of the Crop Yield Prediction Dataset from Kaggle, which is a secondary data source. This dataset contains information on various factors that can affect crop yields, such as weather conditions, soil properties, and pest infestations.

## - Instruments used

### Frameworks

* Jupyter Notebook is an open-source framework that allows users to create and share documents that contain live code, 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.
* React Native. An open-source framework developed by Facebook that allows developers to build mobile applications using the same fundamental principles as React, a popular JavaScript library for building user interfaces. It allows developers to build mobile apps for both iOS and Android platforms using a single codebase, instead of having to develop separate codebases for each platform. Additionally, I 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:

* Pandas 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 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: Seaborn 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 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 (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 is a popular open-source gradient boosting library that provides an implementation of gradient boosting algorithms.
* Plotly is a data visualization library that provides interactive visualization tools. It was used to create interactive visualizations during the Data Analysis phase.

### 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.

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

1. Random Forest Regression: 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): 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: 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.
4. Adaboost Regression: 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: 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: 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. XGBRegressor: An implementation of gradient boosting regression using the XGBoost library. It is optimized for speed and scalability, and provides a number of advanced features such as parallel computing and automatic tuning of hyperparameters.
8. LGBMRegressor: 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.
9. 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.

## - Data analysis approaches.

To analyze the data, we used a combination of statistical techniques and data science approaches, and machine learning techniques, which involved two steps:

1. Data exploration and visualization. A Exploratory Data Analysis (EDA) was performed, which involved summarizing and visualizing the data in order to attempt to understand the underlying patterns and relationships.
2. Feature selection. Involved selecting the most relevant features from the data to use in the model, to help improve the performance of the model by removing unnecessary features that might introduce noise or bias.

# Results

* EDA graphs
* Model comparison graphs
* Model tuning graphs
* App interface

# Conclusions and Discussion

Due to the large-scale origin of the data, the application seems more oriented toward big producers, instead of individual farmers.

# Acknowledgments (optional)

# References

# Annexes