

# **“Agriculture Production Optimization System using Machine Learning”**

Submitted on 09.05.2023

From,

Student : Mr Shiraj Dattatraya Nannaware  
Course : Simulation and System Design  
Matrikel-Nr. : 19994  
Email ID : Shiraj.D.Nannaware@hochschule-stralsund.de

**Supervisor:**

Prof. Dr. rer. nat. Andre Gruening  
Faculty of Electrical Engineering and  
Informatics  
Hochschule Stralsund

**Second Supervisor:**

Prof. Dr.-Ing. Christine Wahmkow  
Faculty of Mechanical Engineering  
Hochschule Stralsund

## Acknowledgment

I have great pleasure in expressing my deep sense of gratitude to **Prof. Dr. rer. nat. Andre Gruening**, faculty of Electrical Engineering and Informatics, Hochschule Stralsund University Applied Sciences, Stralsund, for mentoring this thesis and guiding me on a correct path by providing valuable guidance and suggestions throughout the thesis work. He helped me in every possible way. The knowledge acquired during the preparation of the report would definitely help me in my future ventures.

I want to express my gratitude to **Prof. Dr.-Ing. Christine Wahmkow**, Mechanical Engineering faculty, Hochschule Stralsund University Applied Sciences, Stralsund, who has been a consistent source of support and encouragement for the duration of this project. A special thanks to **Ms. Sadhana Kokate**, HR Manager (Technical), and the entire Maxgen Technologies Pvt Ltd team for their assistance in introducing me to the Python, Microsoft Power BI, and other tools.

I also want to thank the entire faculty of mechanical engineering, electrical engineering, and informatics who helped me either directly or indirectly. Last but not least, I sincerely appreciate the love and support that my family and friends have given me throughout this process. I wouldn't have finished my adventure if it weren't for their support and inspiration.

## Abstract

The agriculture sector performs a crucial function in a nation's economic development. Therefore, the optimization of agricultural production is essential for sustainable development. Algorithms for learning have become more popular in recent years. techniques in agriculture has gained significant attention to optimize the production process. This paper presents an agricultural production optimization system using decision tree regressor and XGBoost regressor. The proposed system aims to predict the yield of crops by considering various factors like temperature, humidity, rainfall, & soil moisture. Data gathered through sensors deployed in a field of crops is utilized to train models that use machine learning. The decision tree regressor & XGBoost regressor models are compared, and it is found that the XGBoost regressor model outperforms the decision tree regressor model. Results indicate that the suggested method accurately predicts the yield of crop. This system may support farmers to optimize their production process by providing accurate predictions of crop yield and suggesting appropriate agricultural practices. The proposed system can also help policymakers in decision-making related to agriculture.

**Keywords:** Agriculture, production optimization, machine learning, decision tree regressor, XGBoost regressor, crop yield prediction, sensor data, evaluation metrics.

## Table of contents

Acknowledgment.....	II
Abstract.....	III
Table of contents.....	IV
List of Figures.....	VI
List of abbreviations .....	VIII
Formula symbol.....	IX
<b>1. Introduction .....</b>	<b>1</b>
1.1 Relevance of the project.....	1
1.2 Motivation .....	1
1.3 Background of the study .....	2
1.4 Importance of agriculture for the economy and society .....	2
1.5 Challenges facing agricultural production .....	3
1.6 Need for innovative approaches to optimize agricultural productivity and sustainability .....	4
1.7 Potential benefits for farmers, researchers, and policymakers .....	4
1.8 Contribution to the field of agricultural optimization using machine learning .....	5
1.9 Machine Learning.....	6
1.9.1 Types of Machine Learning .....	7
1.9.1.1 Supervised Machine Learning .....	8
1.9.1.2 Unsupervised machine learning .....	9
1.9.1.3 Semi-supervised learning .....	10
1.9.1.4 Reinforcement learning.....	11
1.9.2 The Machine Learning (ML) evolution in different areas .....	12
1.9.3 Uses of Machine Learning (ML) in agriculture .....	12
1.9.4 Most popular applications of Machine Learning (ML) in agriculture .....	13
1.9.5 Machine Learning (ML) models used in the agriculture industry .....	13
1.9.6 Rising opportunities of Machine Learning (ML) in digital agriculture .....	14
1.10 Problem Statement .....	14
1.11 Aim and Objectives of the study .....	15
1.12 Scope of the Project.....	15
1.13 Research methodology.....	16
1.13.1 Research design .....	16
1.13.2 Data Pre-processing: .....	17
1.13.3 Model Training and Evaluation: .....	18

1.13.4 Validation: .....	19
1.14 Software requirements .....	20
<b>2. Literature Survey .....</b>	<b>22</b>
2.1 Reserch gap of the study .....	33
<b>3. Methodology .....</b>	<b>35</b>
3.1 Methodology .....	35
3.1.1 Basic process .....	36
3.1.2 Dataset.....	36
3.1.3 EDA (Exploratory Data Analysis).....	37
3.1.3.1 <i>Decision Tree Regressor</i> .....	41
3.1.3.2 <i>Grid Search CV</i> .....	43
3.1.3.3 <i>XGBoost Regressor</i> .....	44
3.1.4 Reason for using decision tree regressor .....	47
3.1.5 Reason for using XGBoost regressor .....	48
3.2 Data collection and Pre-processing: .....	48
3.2.1 Model Training and Evaluation: .....	49
3.2.2 Data sources .....	50
3.2.3 Data cleaning.....	51
3.3 Feature selection.....	53
3.3.1 Model Training and Evaluation .....	54
3.3.2 Model Validation.....	56
3.4 System implementation.....	57
<b>4. Results and Discussion.....</b>	<b>58</b>
4.1 Results.....	58
4.2 Discussion .....	65
<b>5. Conclusion, Suggestions and Future Scope.....</b>	<b>68</b>
5.1 Conclusion.....	68
5.2 Suggestions.....	68
5.3 Future Scope.....	69
<b>6. References.....</b>	<b>71</b>
<b>Declaration on oath.....</b>	<b>74</b>

## List of Figures

Figure 1 Machine learning .....	7
Figure 2 Machine Learning Types .....	8
Figure 3 Supervised learning .....	9
Figure 4 Unsupervised Machine learning .....	9
Figure 5 Clustering and Classification .....	10
Figure 6 Semi-supervised learning .....	11
Figure 7 Reinforcement learning cycle .....	11
Figure 8 “Supervised vs unsupervised vs semi-supervised machine learning in a nutshell” .....	12
Figure 9 Jupyter Notebook .....	20
Figure 10 Python .....	20
Figure 11 PyCharm .....	20
Figure 12 Tkinter .....	21
Figure 13 Applications with Flask .....	21
Figure 14 Power BI. ....	21
Figure 15 Dataset .....	37
Figure 16 Heatmap Correlation .....	38
Figure 17 Prepaing Data for Model Training – Encoding .....	39
Figure 18 Function returns the total number of missing values for each variable in the dataset. ....	39
Figure 19 Function is used to remove rows or columns with missing values from the dataset .....	40
Figure 20 Function provides information on the data types, number of non-null values , and memory usage of each variable in the dataset Algorithms Used .....	40
Figure 21 Decision Tree Regressor .....	41
Figure 22 Model Comparison .....	55
Figure 23 Comparing the Two Models .....	58
Figure 24 Output .....	59

---

Figure 25 Comparison of Train and Test Accuracy in XGBoost Model .....	60
Figure 26 Output [deployment] .....	61
Figure 27 Comparison of decision tree and XGBoost model accuracy .....	62
Figure 28 Calculation of approximate latency for XGBoost model .....	63
Figure 29 Deployment output of XGBoost model .....	64

## List of abbreviations

ML	Machine Learning
AI	Artificial Intelligence
DL	Deep Learning
DL	Deep Learning
DS	Data Sciences
NN	Neural Network
CNN	Convolutional Neural Network
MAE	Mean Absolute Error
RMSE	Root Mean Squared Error
R2	R-Squared
DT	Decision Tree
XGBoost	Extreme Gradient Boosting
LSTM	Long Short-Term Memory Networks
CSM	Crop Selection Method
IOT	Internet of Things
GDP	Gross Domestic Product
EDA	Exploratory Data Analysis
USDA	The United States Department of Agriculture
MODIS	The Moderate Resolution Imaging Spectroradiometer
NOAAC	The National Oceanic and Atmospheric Administration
KNN	K- Nearest Neighbor
RF	Random Forest
SVM	Support Vector Machine
RL	Reinforcement Learning
N	Nitrogen
P	Phosphorous
KNN	K- Nearest Neighbor
RF	Random Forest
SVM	Support Vector Machine
RL	Reinforcement Learning
N	Nitrogen
P	Phosphorous



## Formula symbol

$J$	Splitting Criterion
$MSE_{\text{left}}$	the mean squared error of the left subset
$MSE_{\text{right}}$	the mean squared error of the right subset
$\hat{y}$	Prediction Function
$N$	the number of samples in the leaf node
$y_i$	the target value of the i-th sample
$obj$	Objective Function
$y$	actual target value
$L$	loss function
$\Omega$	regularization term
$F_i(x)$	the i-th decision tree
$b$	the base prediction
$\partial L / \partial \hat{y}$	Gradient Calculation
$\partial^2 L / \partial \hat{y}^2$	Hessian Calculation
$g$	the first derivative of the loss function
$h$	the second derivative of the loss function
$w$	the weight of the decision tree
$g_i$	the gradient of the objective function for the i-th sample
$h_i$	the Hessian of the objective function for the i-th sample
$\lambda$	the regularization parameter

## 1. Introduction

### 1.1 Relevance of the project

Agriculture is a vital sector that performs an important function in guaranteeing food safety and sustainability. However, increasing global demand for food, coupled with restricted accessibility of agricultural property, has made agricultural production more challenging. One potential solution is to leverage machine learning techniques to optimize agricultural production. (K. G. Liakos, 2018)

Without having to be explicitly programmed, artificial intelligence allows algorithms to learn from information and enhance their efficiency as time passes without getting clearly taught. Using artificial intelligence techniques to analyze data from agriculture, we can create an agricultural production optimization system that can identify and analyze patterns, make predictions, and optimize the production process. (V. Meshram, 2021)

The purpose of this doctoral dissertation is to create a system for optimizing the farming industry using machine learning techniques. The system will use historical agricultural data to recognize trends and forecast upcoming produce production. It will also optimize agricultural production processes by recommending the best possible use of natural assets such as water, fertilizers, and pesticides. (K. G. Liakos, 2018)

The system will use two machine learning algorithms, namely decision tree regressor and XGBoost regressor, to predict crop yields and optimize agricultural production processes. The decision tree with regressor method is going to be utilized to determine the connections among factors, while the XGBoost regressor algorithm will be used to optimize the production process. (T. Oladipupo, 2010)

The proposed system will be beneficial to agriculture, agronomists, and other stakeholders pertaining to agricultural goods. By optimizing agricultural production processes, we can potentially increase crop yields, reduce the negative impact of farming practices regarding environmental issues, and improve food security. (R. G. De Luna, 2019)

### 1.2 Motivation

The goal of a Agriculture Production Optimization System using Machine Learning project utilizing machine learning algorithms is to increase agricultural productivity and efficiency while lowering costs and encouraging sustainability. Accurate crop yield forecasts can aid farmers in improving their planning, resource utilization, and crop management choices. This could result in greater financial gain for farmers, less waste, and a lesser environmental impact.

The ultimate objective of such a project is to enhance results for farmers, consumers, and the environment.

### **1.3 Background of the study**

Agriculture is a critical sector that performs a crucial function in supplying food and livelihoods to billions of people worldwide. However, agricultural production faces numerous challenges such as climate change, pests and diseases, water scarcity, soil degradation, and low productivity. These challenges can reduce crop yields, compromise food security, and threaten sustainable development. To overcome these challenges, there is a need for innovative approaches to optimize agricultural productivity and sustainability. The use of machine learning, or ML, has become a potent instrument for enhance agricultural production by providing insights into complex data sets and predicting outcomes with high accuracy. Several studies have shown the potential of ML techniques in predicting crop yields, identifying pests and diseases, and classifying land cover. (R. G. De Luna, 2019)

Among the various ML techniques, decision tree regressor and XGBoost In the past few years, regressors have acquired favor due to their precision and comprehension. Decision tree regressor use A tree-like representation of choices and their potential events for predicting results. They are easy to understand and interpret, making them useful for generating insights into the procedure of reaching decisions. (Indu, A. S. Baghel, 2022)

On the contrary, XG Boost regressor are a type of gradient boosting algorithm that uses a set of decision trees to predict outcomes. They are known for their speed, scalability, and high accuracy, making them a popular choice for many applications. However, while several studies have demonstrated the effectiveness of these techniques in predicting crop yields and optimizing agricultural production, there is still a need for more research to explore their scalability, robustness, and generalizability. Furthermore, there is a lack of research that compares the performance of decision tree regressor and XGBoost regressor in agricultural optimization. Consequently, aim of the study is to create a machine learning-based method for optimizing the farming industry and to evaluate the performance of decision tree regressor and XGBoost regressor. By comparing the scalability, robustness, and generalizability of these techniques, utilizing artificial intelligence, this investigation aims to bring value to the practice of farming efficiency. (K. G. Liakos, 2018)

### **1.4 Importance of agriculture for the economy and society**

Farm is an indispensable industry that serves a vital function in economics and in society. Here are some points highlighting the importance of agriculture:

- **Food security:** Agriculture is the world's main supply of food for its people. It is essential for ensuring food security and meeting the nutritional needs of people worldwide.
- **Livelihoods:** Agriculture is a vital source of income and livelihoods for millions of people, especially in rural areas. It provides employment opportunities for farmers, farm laborers, and workers in related industries such as transportation, processing, and marketing. (C. Ashcraft and K. Karra, 2021)
- **Economic growth:** Agriculture is a significant contributor to the economy of many countries, particularly in developing nations. It contributes substantially to the nation's economic output (GDP.) and provides a market for other industries, including manufacturing and services.
- **Environmental benefits:** Agriculture can have positive environmental impacts, such as preserving biodiversity, conserving natural resources, and promoting sustainable land use practices.
- **Cultural heritage:** Agriculture is often an integral part of the cultural heritage of communities, shaping their traditions, customs, and identity. (F. Garcia, 1999)

### 1.5 Challenges facing agricultural production

Agricultural production faces numerous challenges that can reduce crop yields, compromise food security, and threaten sustainable development. Here are some of the significant challenges facing agriculture today:

- **Climate change:** Severe meteorological occurrences, such as hurricanes, have resulted from changes in the climate, droughts, and heatwaves, which can damage crops, reduce yields, and compromise food security. (D. Brunelli, 2020)
- **Pests and diseases:** Pests and diseases can cause significant crop losses, reducing yields and affecting the quality of produce. Climate change can also exacerbate the spread of pests and diseases, making them harder to control.
- **Water scarcity:** Water scarcity is a significant challenge for agriculture, particularly in arid & semi-arid regions. Lack of access to water can reduce crop yields and limit agricultural productivity.
- **Soil degradation:** Soil degradation, including erosion, nutrient depletion, and salinization, can reduce soil fertility, leading to lower crop yields and poor quality produce.
- **Low productivity:** Low productivity is a significant challenge in many agricultural systems, particularly in developing countries. Factors contributing to low productivity

include poor infrastructure, inadequate access to markets, and lack of access to credit and technology. (M. S. Swaminathan, 1984)

- **Food waste:** Food waste is a significant challenge in agriculture, with significant amounts of produce lost during production, transportation, storage, and processing.

These challenges can reduce agricultural productivity, compromise food security, and threaten sustainable development. Addressing these challenges requires innovative solutions that are sustainable, scalable, and adaptable to the changing needs of agricultural systems. (B. Data and I. Scheduling, 2021)

### **1.6 Need for innovative approaches to optimize agricultural productivity and sustainability**

Agriculture is a critical sector of the economy that provides food, fiber, and other essential raw materials to sustain human life. However, it is also one of the most resource-intensive industries, relying heavily on natural resources such as land, water, and energy. The increasing demand for food, coupled with the growing global population, is putting pressure on farmers to optimize. They are produced while reducing unfavorable consequences for the environment. This has created a need for innovative approaches to optimize agricultural productivity and sustainability. (K. Alibabaei, 2022)

One promising solution to address this challenge is application of artificial intelligence (ML) techniques to develop an agricultural production optimization system. Such a system would utilize data from various sources, such as weather forecasts, soil analysis, and crop sensors, to generate predictions and recommendations for farmers. These predictions could include optimal planting and harvesting times, recommended fertilizer and pesticide applications, and other management decisions. (I. Ahmad *et al.*, 2022)

By leveraging the power of ML algorithms, this system could continuously learn from new data inputs and adjust its predictions over time. This would enable farmers to make more informed decisions, resulting in higher crop yields, reduced waste, and lower environmental impact. Additionally, the system could help farmers optimize their resource use, reducing costs and increasing profitability.

The implementation of an agricultural production optimization system using machine learning has the ability to transform the farming sector, promoting sustainable farming practices, and supporting global food security. (K. G. Liakos, 2018)

### **1.7 Potential benefits for farmers, researchers, and policymakers**

An agricultural production optimization system using machine learning has the potential to benefit farmers, researchers, and policymakers in a number of ways.

Farmers can use this system to optimize their crop production by analyzing data collected from sensors installed in their fields. The system can analyze data like temperature, humidity, soil moisture, & levels of nutrients to assist producers with irrigation, fertilization, and harvesting choices. By optimizing crop production, farmers can increase their yields and reduce their costs, leading to higher profits and more sustainable farming practices. (R. Sharma, 2020)

Researchers can use the data collected from these systems to gain insights into crop growth patterns and identify factors that contribute to crop health and yield. This data can be used to develop new crop varieties that are more resistant to disease, pests, and weather fluctuations. This can help improve the overall health and productivity of crops, leading to more sustainable agricultural practices.

Policymakers can use the insights gained from these systems to develop policies that promote sustainable agriculture practices. By understanding the factors that contribute to crop health and yield, policymakers can develop regulations and incentives that encourage farmers to adopt sustainable farming practices. This can aid in reducing the harmful effects of farming and enhancing nutrition, and promote economic growth in rural areas. An agricultural production optimization system using machine learning has the potential to benefit farmers, researchers, and policymakers by improving crop yields, promoting sustainable agriculture practices, and contributing to economic growth. By leveraging the power of data and machine learning, we can create a more sustainable and productive agricultural sector that benefits everyone involved. (K. G. Liakos, 2018)

### **1.8 Contribution to the field of agricultural optimization using machine learning**

The development of an agricultural production optimization system using machine learning represents a significant contribution to the field of agricultural optimization. This system has the potential to revolutionize how farmers and policymakers approach crop production and management by providing accurate and real-time data analysis, leading to more informed decision-making. (K. G. Liakos, 2018)

Traditional agricultural practices rely on manual observation and experience, which can be subjective and prone to error. However, utilizing sensing and predictive techniques, this system can accurately analyze data from various sources, including weather patterns, soil moisture, and crop growth, to make predictions about the optimal time to sow, water, and collect crops. (V. Meshram, 2021)

Additionally, this system can help producers are able to identify early indicators of plant illnesses and parasites, enabling immediate action to be taken to minimize crop losses. This

system can also reduce the amount of water and fertilizer used, leading to cost savings and more sustainable farming practices. (T. Oladipupo, 2010)

In terms of its contribution to the field of agricultural optimization, this system represents a major step forward in the integration of technology into agriculture. By using machine learning algorithms, we can better understand the complex interactions between soil, climate, and crop growth, leading to more efficient and sustainable farming practices. (C. Ashcraft and K. Karra, 2021)

In conclusion, the development of an agricultural production optimization system using machine learning represents a significant contribution to the field of agricultural optimization. This framework might change how producers handle the cultivation of crops and management, leading to more sustainable practices, increased productivity, and improved food security. (Indu, A. S. Baghel, 2022)

### **1.9 Machine Learning**

Machine learning, or ML, is a subfield of computer science that entails creation of statistical models & algorithms that allow machines to acquire knowledge from data and make forecasts and choices without having to be specifically programmed for doing so. In a nutshell, it is a method for computers to acquire knowledge through experience, just as humans do. The fundamental concept for machine learning is to provide a computer with a large quantity of data and enable it to learn from the information using methods. The algorithms used aim to recognize relationships and trends within the information and use them to make forecasts or determinations regarding new data. In Figure 1 the machine learning techniques include unsupervised learning, supervised learning, and learning by reinforcement. Presenting a computer with labeled data (data that has been previously grouped or organized) and then teaching it to identify patterns within that data constitutes supervised learning. Unsupervised machine learning refers to the process of providing unstructured information to a machine and enabling it to discover relationships and trends on its own. The process of instructing a machine to make judgments according to feedback from the environment is known as learning by reinforcement. (I. Ahmad *et al.*, 2022)

finances, medical care, advertising, and commuting are just a few of the many disciplines in which artificial intelligence is utilized. For instance, machine learning can be used in finance to predict stock prices and detect misconduct. Using machine learning techniques, medical pictures can be evaluated and individualized treatments developed. For advertising use, algorithms based on machine learning can be used to determine consumer preferences and enhance advertising campaigns. Using artificial intelligence approaches, it is possible to

construct self-driving vehicles and improve flow of traffic in the transport industry. (R. Sharma, 2020)

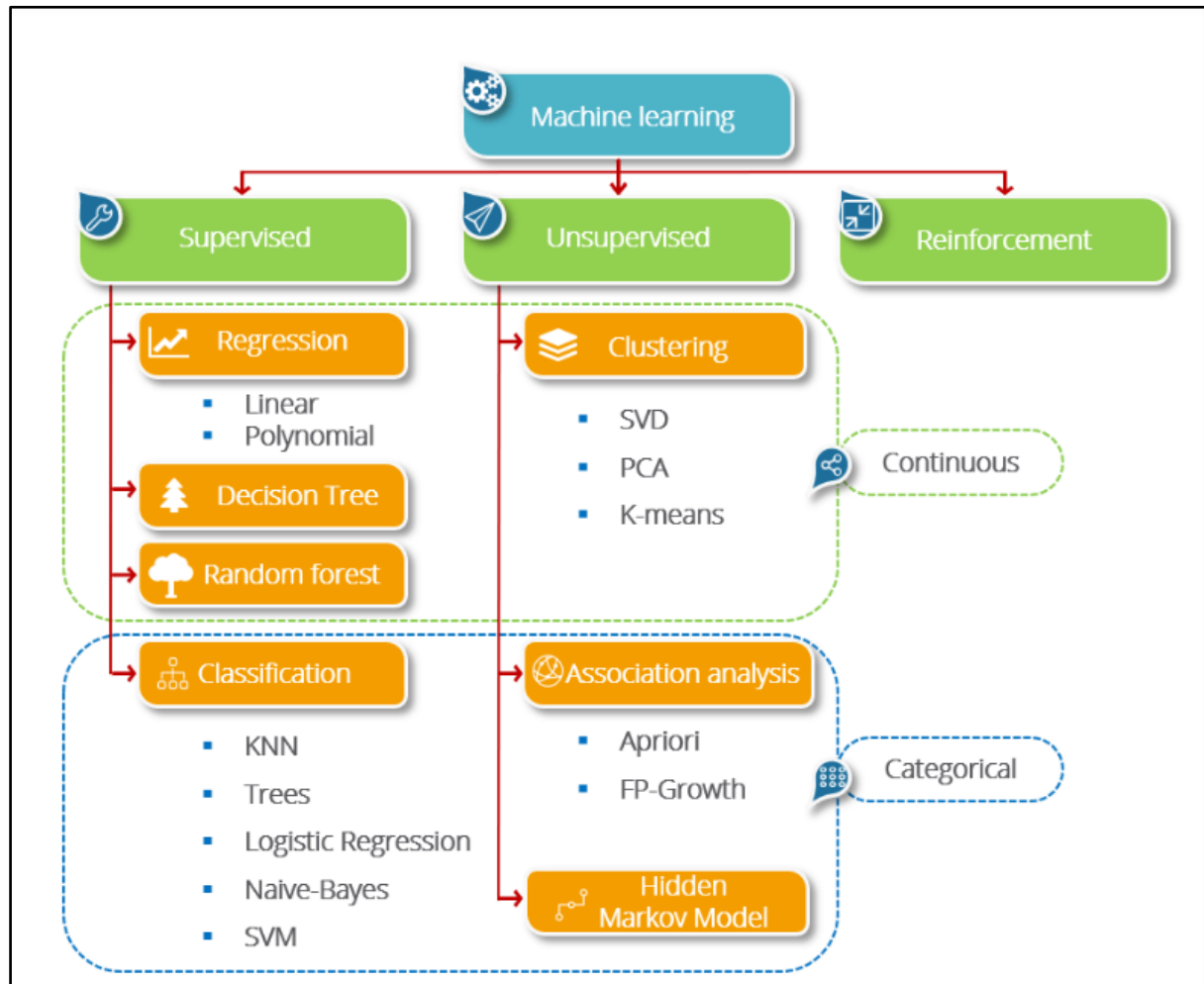


Figure 1 Machine learning

### 1.9.1 Types of Machine Learning

There are numerous ways to train for learning computations, each with their own benefits and drawbacks. To comprehend the advantages and disadvantages of each type of artificial intelligence, we need to first look at the types of data they process. In machine learning, there are two categories of data: classified information and unlabeled information.

Both the input and the output parameters for annotated data are entirely readable by machines, but marking the data takes a significant amount of labor from humans. In readable by machines format, data without labels comprises only one or not one of the parameters. This removes the requirement for employees, but requires more intricate methods. (M. O. Adebisi, 2020)

Additionally, there are also artificial intelligence methods that are employed in very particular circumstances.



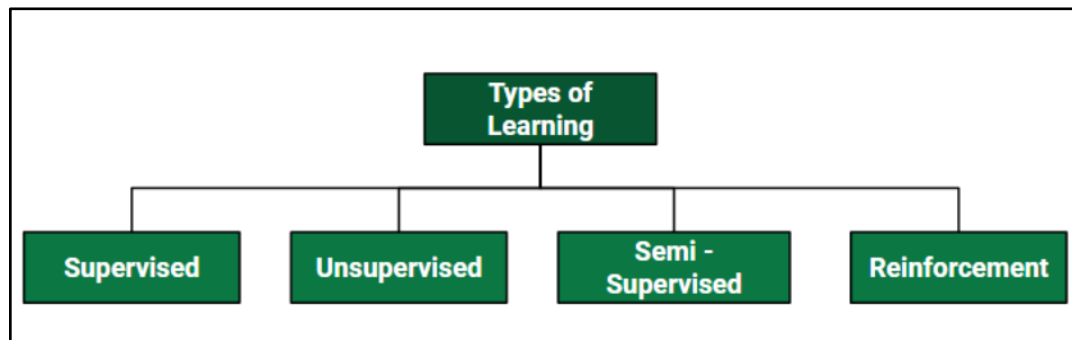


Figure 2 Machine Learning Types

#### 1.9.1.1 Supervised Machine Learning

This type of artificial intelligence entails oversight, in which computers receive instruction on labeled data sets and given the ability to predict outcomes based on training data provided. Some of the input & output characteristics have been identified, according to the data labels set. Consequently, the device is instructed employing the input and corresponding output. In the following stages, an instrument is built to forecast future events based on the test information. (Y. Mekonnen, 2020)

The primary goal of guided training is to relate the data variable (a) to the result variable (b).

(b). Further categorizing guided artificial intelligence through two main categories:

- **Classification:** These are techniques for resolving classification problems involving discrete outcomes. Detection of spam as well as emails filtering are evident real-world applications of this category.

Decision Tree Algorithms, Random Forest Algorithm Examples of renowned classification techniques include Logical Regression Methods and the Support-Vector Network Method. (V. Meshram, 2021)

- **Regression:** algorithms that manage issues related to regression with a linear connection between input and output factors. It is known that these can predict output that includes continuous factors. Examples involve forecasting the environment and assessment of market trends.

The Simple Linear Regression Algorithm, Multimodal Prediction Algorithms a Decision Tree Algorithm, and Lasso Regression Algorithm are well-known regression algorithms.

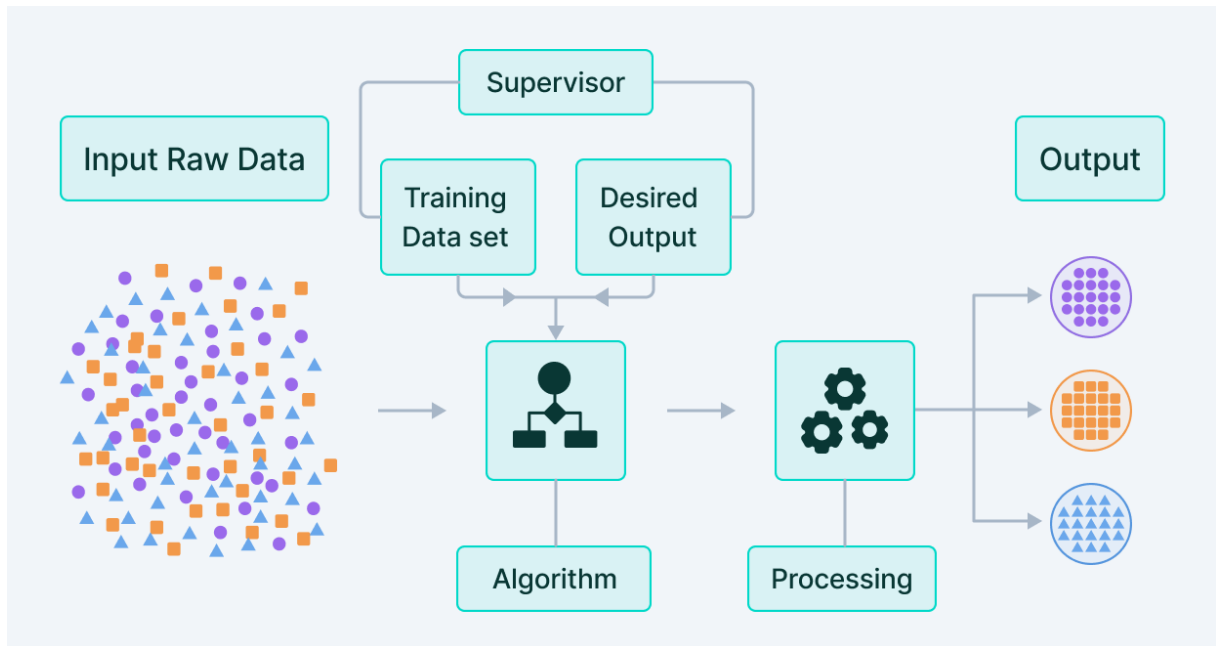


Figure 3 Supervised learning

### 1.9.1.2 Unsupervised machine learning

The algorithm is trained employing an unstructured The data set is allowed to predict the output without human involvement. The purpose of a method of unsupervised learning is to arrange data the unorganized information based on contrasts, similarities, and trends in the data provided as input. (B. Sharma, 2020)

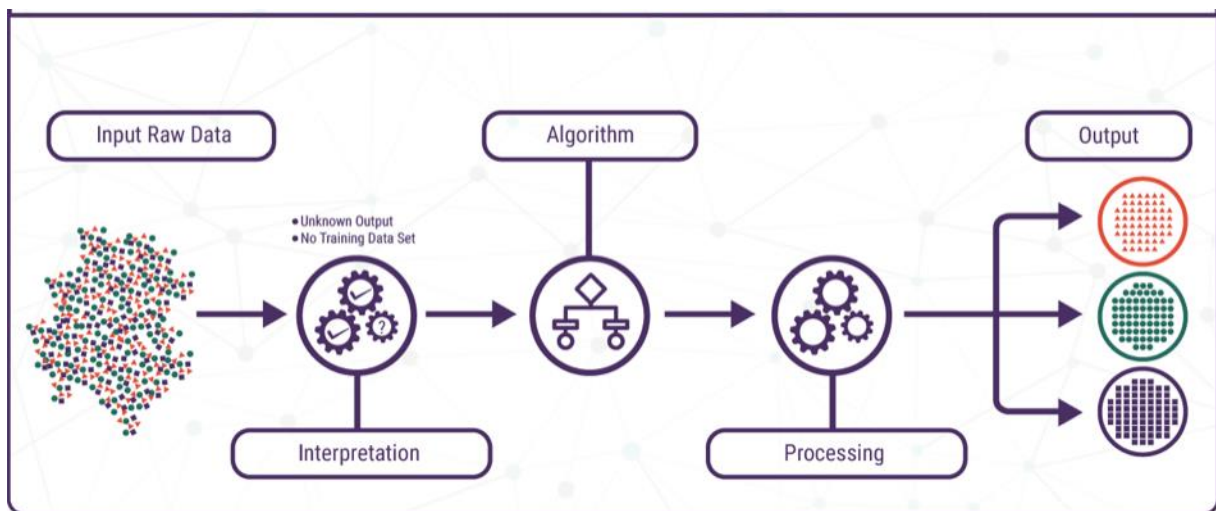


Figure 4 Unsupervised Machine learning

Two additional categories of uncontrolled training methods exist are Clustering & Association explained as follow.

- Clustering:** The method of clustering refers to the practice of organizing things into groupings according to characteristics such as their shared or distinct characteristics. For instance, clustering consumers based on the products they buy. Principal Component Inspection K-Means Algorithms Mean-Shift Algorithm, and DBSCAN Methods, and an independent component evaluation are examples of well-known algorithmic methods for clustering. (M. Rakhra, 2021)

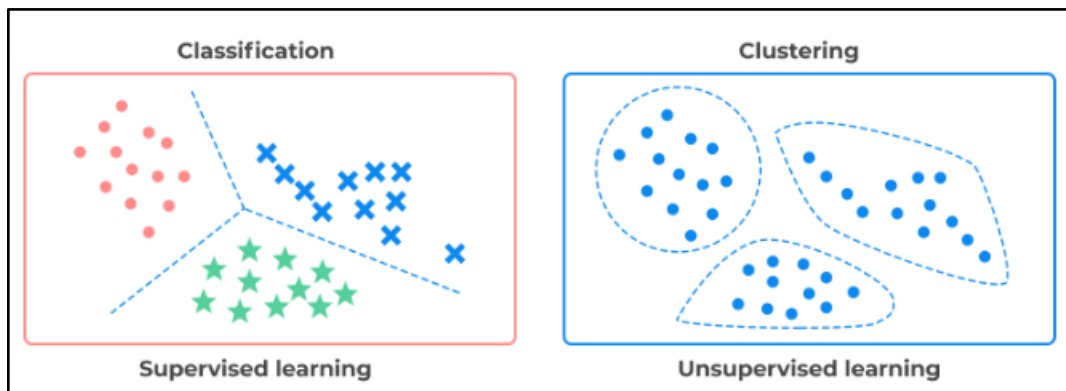


Figure 5 Clustering and Classification

- Association:** Association learning is the process of discovering common relationships between the variables of a big data set. It determines the relationship between different information pieces and plots variables. mining the web's usage and marketplace analysis of data are both typical applications. Known association-based methods include the Apriori Algorithm, the Eclat Algorithm, and the FP-Growth Program. (A. Priyadharshini, 2021)

### 1.9.1.3 Semi-supervised learning

Unsupervised and supervised learning are combined in semi-supervised machine learning. It educates its algorithms with a variety of designated and unlabeled data sets. Using both kinds of information sets, semi-supervised instruction overcomes the shortcomings of the aforementioned alternatives.

Consider a college student as an illustration. College pupils learning something under the careful eye of a teacher is referred to as supervised instruction. A student engages in autonomous learning when he or she learns the same concept autonomously at home, unsupervised by a teacher. In contrast, a college student revising a concept after learning it under the watchful eye of a professor is an illustration of learning that is partially supervised. (S. Mishra, 2016)

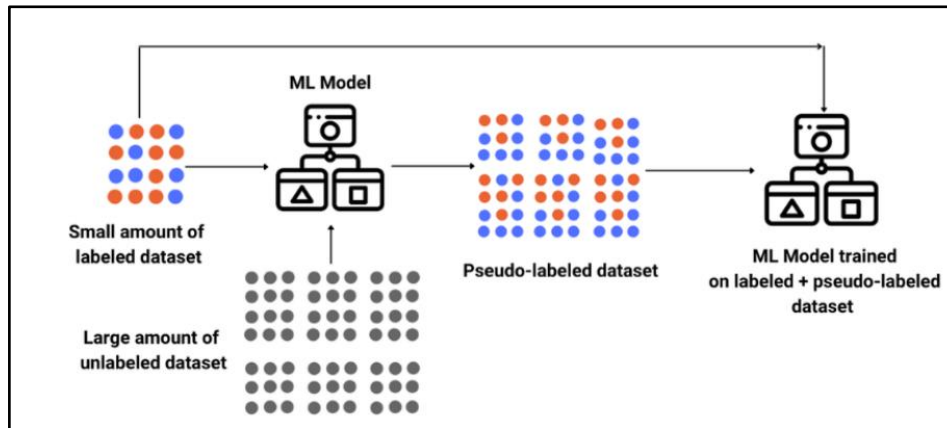


Figure 6 Semi-supervised learning

#### 1.9.1.4 Reinforcement learning

The foundation for acquiring knowledge through repetition is feedback. Here, the AI component evaluates its environment using the trial-and-error method, acts, learns from its experiences, and improves its performance. The component is rewarded for every right move and penalized for every mistake. Therefore, learning through feedback endeavors to maximize rewards through positive behavior. (L. Benos, 2021)

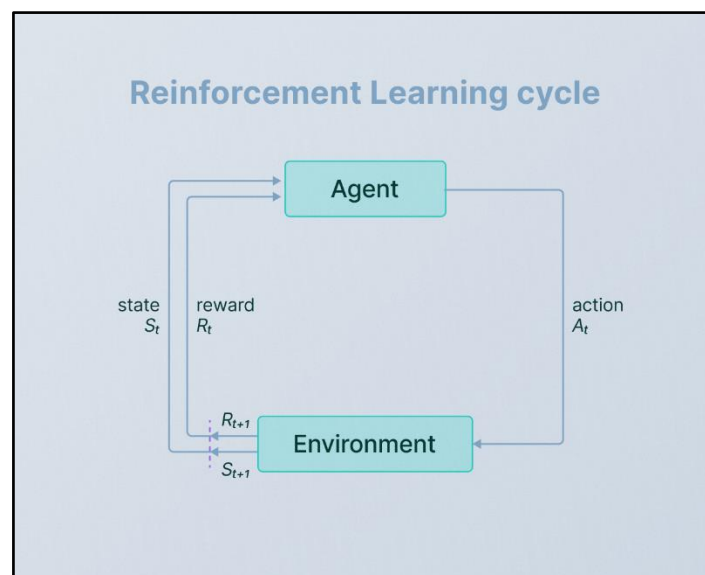


Figure 7 Reinforcement learning cycle

Reinforcement learning, in contrast to supervise learning, lacks labeled data, and robots learn solely by means of experience. Consider video athletics. The eventual goal of the agent is to obtain a high score.

Reinforcement learning is utilized in numerous disciplines, including game theory, the theory of information, and multi-agent systems. reinforced learning is further subdivided into two distinct methodological categories: (S. Dimitriadis 2008)

- **Positive reinforcement learning:** This refers to the addition of a reinforcing stimulation, such as an incentive, after a particular action of the agent, that raises the likelihood that the behavior will occur once more in the future.
- **Negative reinforcement learning:** Negative learning by reinforcement refers to the process of reinforcing an action that avoids a detrimental effect.

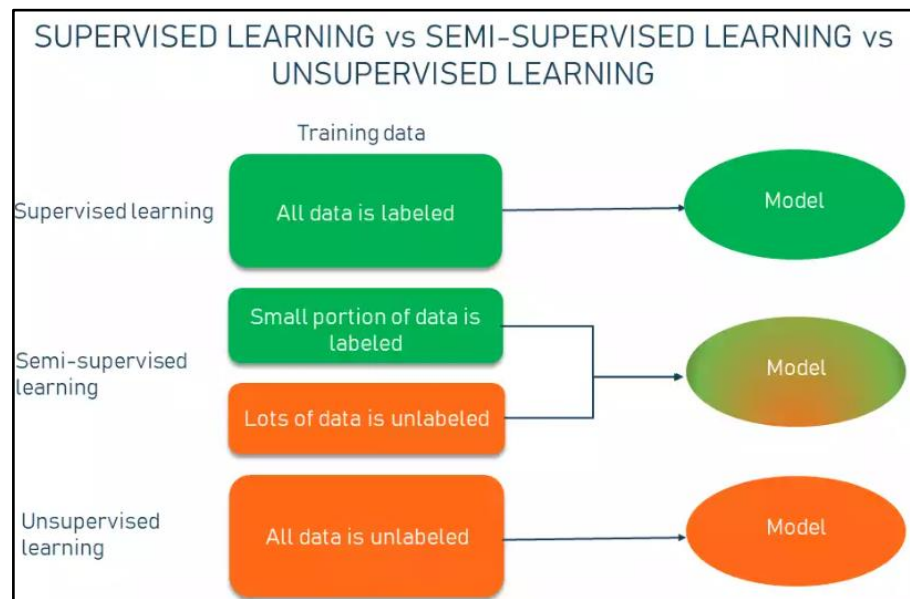


Figure 8 “Supervised vs unsupervised vs semi-supervised machine learning in a nutshell”

### 1.9.2 The Machine Learning (ML) evolution in different areas

Accompanying the evolution of big data technologies or other desktop pcs, computer vision is advancing. Companies are growing to give new options for comprehending the many workflows related to crop ecosystem activities. The academic technique of algorithms enables computers to understand with no need for software. Several science technicians, such genomics, pharmacology, pharmacology, weather, macroeconomic studies, automation, agriculture, and meteorological, use algorithms. (R. Kumar 2015)

### 1.9.3 Uses of Machine Learning (ML) in agriculture

A multitude of industries, first from family to the workplace, are already using ai technology; agricultural is perhaps the most recent to do so. The purpose of training in agribusiness is to boost crop yields in the farming sector.

- **Retailers:** That agricultural equipment is used by seeding sellers to analyse data and produce crop yield. Though rodent control firms use them to detect different germs, insects, and rodents. (A. Sharma, 2021)
- **ML/AI is used to boost the yield of crops:** Artificial intelligence (AI) & ML (machine learning) are both used to estimate which again maize will generate the high outcome and under what circumstances. It will also establish the optimal rainfall patterns for optimal outcomes.
- **ML/AI assists in identifying bug hunters:** One of firms, Rentokil, uses ML/AI to eradicate all insects and animals. Price water house coopers has created an Application that is used by other firms to identify issues. The app captures images of the insect and then opens PestID. When a flaw is detected, the application will immediately suggest a fix, enabling the professional to take the necessary steps. In addition, it will advise the insecticide that must be applied. (Indu, A. S. Baghel, 2022)

#### 1.9.4 Most popular applications of Machine Learning (ML) in agriculture

Let's take a look at different applications of machine learning in agriculture.

- **Agriculture Robot:** Currently, lot of firms are developing and constructing robotics to do agriculture chores. This involves agriculture gathering and performing more quickly than workers. Is the most successful agricultural use of algorithms. (Kavita and P. Mathur, 2021)
- **Monitor crop and soil:** Currently, businesses use technology and artificial neural networks. Employing helicopters or other tools, the material is then gathered to evaluate the vegetation and environment. Additionally, researchers utilise the programme to regulate soil nutrients.
- Using innovative farming techniques, landowners can preserve their goods and safeguard them from insects. Businesses are developing robotic or automations to assist them in achieving their objectives. Blue River Cloud Computing About and Shower robotic will watch and accurately spray herbicides on commodities such as wheat. The exact quantity of treatment may assist in reducing chemical costs. (S. T. Jagtap, 2021)

#### 1.9.5 Machine Learning (ML) models used in the agriculture industry

Currently, farming practices are using predictive models and breakthroughs. Utilizing AI Technologies is advantageous for food science divisions.

Agricultural Data Centre, a chat app for growers, will just use ml algorithms and analysis technologies to provide daily closing prices.

- Vegetables were already managed and monitored by computers.
- Instruments facilitate agricultural attempt to collect.
- As per the study, the agricultural industry will expand in the next generations if AI and ML are often used in agribusiness. (H. Pallathadka, 2021)

### **1.9.6 Rising opportunities of Machine Learning (ML) in digital agriculture**

There's also a surge in digitized farmland, particularly employs a comprehensive strategy to optimise crop production and minimise environmental benefit. Intensive farming generates data based on multiple of devices that contribute to a better knowledge, including the vegetation, soil, and cloud cover, and also the factory automation. These statistics will aid us in creating, effective judgments. We must use algorithms to real datasets in order to boost production. (H. Pallathadka, 2021)

### **1.10 Problem Statement**

Agricultural production faces several challenges that impact the efficiency, sustainability, and profitability of the sector. These challenges include The scarcity of arable land, alterations in the environment, the unpredictability of the the environment, parasites, and illnesses all contribute to food insecurity, and the increasing global demand for food. These challenges make it difficult for farmers to optimize agricultural production and ensure food security.

Traditional agricultural production methods rely on experience, intuition, and historical data. These methods are often time-consuming, labor-intensive, and can lead to suboptimal outcomes. Therefore, there is a need to develop an agricultural production optimization system that leverages machine learning techniques to improve agricultural productivity.

The solution to this problem involves developing a system that can analyze and learn from agricultural data to make accurate predictions and optimize agricultural production processes. The system will use Utilize artificial intelligence algorithms including decision tree regressor and XGBoost regressor to forecast agricultural yields and maximize the use of resources. The system will be designed to be user-friendly and accessible to farmers, agronomists, and other stakeholders involved in agricultural production.

### **1.11 Aim and Objectives of the study**

#### **Aim:**

The goal of this research is to establish an agricultural business production optimization system using machine learning techniques to predict crop yields and optimize resource utilization.

#### **Objectives:**

- To conduct a thorough literature review on artificial intelligence uses for agricultural output maximization.
- To acquire and process pertinent agricultural information from multiple sources., including weather data, soil data, and crop yield data.
- To develop a machine learning model using decision tree regressor and XGBoost regressor algorithms to predict crop yields depending on soil and climate circumstances.
- To optimize the use of resources such as water, fertilizer, and pesticides by recommending the optimal amount and timing of application.
- To test and validate the developed model using historical data and compare its performance with traditional methods of Produce from agriculture.
- To evaluate the economic and environmental benefits of the proposed system by analyzing its impact on crop yields, resource utilization, and carbon footprint.
- To design a user interface that is simple to use for the proposed system that can be used by farmers, agronomists, and other stakeholders involved in agricultural production.
- To provide recommendations and guidelines for the implementation and adoption of the proposed system in different regions and farming practices.

The objectives of this study aim to contribute to sustainable agriculture by optimizing agricultural production processes, minimizing the environmental impact of farming procedures and guaranteeing food security.

### **1.12 Scope of the Project**

The scope of this project is to develop an agricultural production optimization system predicting yields for crops with artificial intelligence methods and optimize the use of resources such as water, fertilizer, and pesticides. The proposed system will be designed to be user-friendly and accessible to farmers, agronomists, and other stakeholders involved in agricultural production.



The project will focus on the development of machine learning models using decision tree regressor and XGBoost regressor algorithms to predict crop yields based on soil and weather conditions. The models will also optimize resource utilization by recommending the optimal amount and timing of application for water, fertilizer, and pesticides.

The project will collect and preprocess relevant agricultural data from various sources. The project will also evaluate the economic and environmental benefits of the proposed system by analyzing its impact on crop yields, resource utilization, and carbon footprint. The analysis will help to determine the potential savings in cost and the reduction in environmental impact resulting from the adoption of the proposed system.

The project will develop a user-friendly interface for the proposed system that can be used by farmers, agronomists, and other stakeholders involved in agricultural production. The interface will provide easy access to the predictions and recommendations generated by the machine learning models. The proposed system's scope is limited to the prediction of crop yields and the optimization of resource utilization in agriculture. The system may not be applicable to all crop types or farming practices and may require customization for specific regions or conditions.

### **1.13 Research methodology**

In this project, we aim to develop an Agricultural Production Optimization System using Machine Learning. We will use Decision Tree Regressor and XGBoost Regressor models for predicting the yield of crops based on various environmental factors such as temperature, rainfall, soil quality, etc. The research design, data sources and collection system, model selection and training methods, and evaluation and validation methods used in this project.

#### **1.13.1 Research design**

- **Type of research:** This project involves predictive research, where we will use machine learning models to predict the yield of crops based on various environmental factors. We will use Decision Tree Regressor and XGBoost Regressor models to build a predictive model that can optimize agricultural production.
- **Data sources and collection methods:** We will collect data from various sources, including government agencies, research papers, and online databases. We will use various data collection methods such as surveys, interviews, and experiments to collect the required data. The data will include information on environmental factors such as temperature, rainfall, soil quality, etc., and the yield of crops.

- **Model selection and training methods:** We will use two machine learning models, Decision Tree Regressor and XGBoost Regressor, to build a predictive model. We will train the models using the collected data and optimize the hyperparameters of the models using techniques such as grid search and random search.
- **Evaluation and validation methods:** We will assess the efficacy of the simulations employing a variety of effectiveness measurements, including mean absolute error (MAE), Root Mean Squared Error (RMSE), and R-squared (R2). We will also use validation techniques such as Holdout validation and Cross-validation to verify the robustness of the models.

We will use predictive research to develop an Agricultural Production Optimization System using Machine Learning. We will collect data from various sources and use Decision Tree Regressor and XGBoost Regressor models for building a predictive model. We will evaluate the performance of the models using various performance metrics and validation techniques. This system will help optimize the production process and increase the yield of crops, which is crucial for the agricultural industry.

### 1.13.2 Data Pre-processing:

Preprocessing of data is a crucial phase in any algorithms for learning endeavor. In the following part, we are going to talk data cleansing and quality control, feature selection and engineering, and data normalization and standardization techniques used in this project.

- **Data cleaning and quality control:** The collected data may contain errors, outliers, missing values, as well as additional discrepancies that can negatively impact the efficacy of models developed using machine learning. Consequently, we are going to perform data cleansing and quality control to ensure the data's precision and consistency. This involves eliminating duplicates, bringing in lacking values, and eliminating outliers. In addition, we will conduct quality control tests to guarantee that the data has an excellent quality and conforms to the specified criteria.
- **Feature selection and engineering:** Feature selection and engineering is another crucial step in data pre-processing. We will select the relevant features that are important for predicting the yield of crops. We will also engineer new features by combining or transforming the existing features to enhance the efficiency of the simulations. For instance, we may calculate the average temperature and rainfall for a particular season or month and use it as a new feature.

- **Data normalization and standardization:** Normalization and classification of data are processes that turn the data into a uniform scale or range. This is essential as artificial intelligence models are dependent on the magnitude and range of input characteristics. In order to normalize and organize the data, we will use methods such as Min-Max and Z-score normalizations are utilized. Min-Max standardization transforms the information into a range between zero and one., whereas Z-score standardization converts the data to have mean of 0 and standard deviation of 1.

We will perform data cleaning and quality control, feature selection and engineering, and data normalization and standardization To guarantee that the information is precise, constant and in an appropriate format for models based on machine learning. This will help us build more accurate and robust models that can predict the yield of crops accurately

### 1.13.3 Model Training and Evaluation:

In this section, we will discuss the XGBoost Regressor and Decision Tree Regressor models used in this project and the performance metrics used to evaluate the models.

- **Decision Tree Regressor and XGBoost Regressor:** Decision Tree Regressor & XGBoost Regressor are popular machine learning models used for regression problems. Choice Diagram Regressor is a tree-based algorithm that divides data into groups according to a feature's value. It is straightforward to understand and can manage discrete and continuous information. XGBoost Regressor is an ensemble model that combines multiple decision trees to improve the performance of the model. It is known for its high accuracy and speed and is widely used in various industries.
- **Performance metrics:** We will analyze the performance of models using a variety of evaluation criteria. Typical indicators of success for problems with regression include: MAE quantifies the mean total disparity among the expected and actual outcomes. A lower MAE signifies improved performance. Root It quantifies The root of the mean cubed variance between the predicted and observed values. The lower the RMSE, the more impressive the outcome is. R-squared (R2): It indicates the proportion of the objective variable's variance that is explained by the model's coefficients. It goes from 0 to 1, with higher values representing superior performance.

We will pick the best models according to how it performs after evaluating each one employing these performance indicators. We will additionally employ tools for visualization, like as scatter graphs and residue plots, to evaluate the efficacy of the models and discover any patterns or trends in the data.

We will use the Decision Tree Regressor and XGBoost Regressor models to estimate crop productivity based on a variety of environmental inputs. Various indicators of success, including MAE, RMSE, and R2, will be used to assess the effectiveness of the models. This is going to help in selecting the most suitable model and enhancing the precision of our forecasts.

#### 1.13.4 Validation:

Validation is a crucial stage in artificial intelligence initiatives since it ensures that the predictions adapt well to fresh data. In this part, we will examine the two most prevalent validation methods: Cross-Validation and Holdout Validation.

- **Exclusion Validation:** Holdout Verification is a straightforward validation method that divides the data into testing and training groups. On the training set, the simulation is trained before being tested on the evaluation set. The testing set is a subset of data that was not used to train the model and is used to evaluate the model's effectiveness on fresh data. The majority of the data is typically utilized for instruction, with the remainder used for testing.
- **Cross-Validation:** Cross-Valid is an additional rigorous verification approach involving the division of data into numerous groups, or folding. The model undergoes training using k-1 creases, and the extra fold is used for evaluation. This procedure is performed multiple times, every fold serving as a test set exactly once. The efficacy of the model is estimated by averaging the performance metrics across k folds. This method reduces the likelihood of over fitting while offering a more precise assessment of the model's efficacy.

In this study, outlier validation and cross-validation are going to be used to assess the efficiency of the systems. We are going to employ holdout validated to get an initial assessment of the model's efficacy and cross-validation to obtain an estimate that is more accurate. In addition, we will employ Grid Search and Random Search to tweak the hyper parameter settings of the models and enhance their performance.

Verification is a necessary step in artificial intelligence projects in order to guarantee that the models adapt effectively to fresh data. To assess the effectiveness of models and tweak their hyperparameters to we are going to employ hold validation and the cross-validation approaches. This can assist us develop more precise and solid models that properly estimate crop production.

### 1.14 Software requirements

#### Jupyter Notebook:

Jupyter Notebook is a free online tool that enables the creation and sharing of notebooks with real-time code, calculations, visualizations, and prose that is narrative in nature. It is frequently utilized in data mining and computational science for data exploration, analysis, and visualization. (T. Oladipupo, 2010)

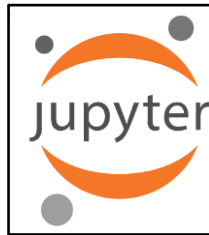


Figure 9 Jupyter Notebook

#### Python:

Python is an easy-to-learn and use high-level, interpretive programming language. Numerous fields, including the development of websites, machine learning, data mining, and computational sciences use it extensively. Python has an extensive user and development belonging, making it simple to access resources and help. (C. Ashcraft and K. Karra, 2021)



Figure 10 Python

#### PyCharm:

PyCharm is a popular integrated development environment (IDE) for Python that is used for coding, debugging, and testing Python applications. It provides many features such as code completion, syntax highlighting, debugging tools, and version control integration. (T. Khan, 2021)



Figure 11 PyCharm

**Tkinter:**

Tkinter is a common Python library used to create user interfaces with graphics. (GUIs). It gives you a collection of elements including icons and labels, and text boxes that can be used to create GUI applications. Tkinter is easy to learn and use, and it is a good choice for beginners who want to create simple GUI applications. (B. Data and I. Scheduling, 2021)



Figure 12 Tkinter

**Flask:**

Python's Flask is a minimalist web-based application platform. It is simple to understand and usage, and it provides a simple way to create web applications. Flask is often used for creating small to medium-sized web applications, and it provides features such as routing, templates, and request handling. (V. Meshram, 2021)



Figure 13 Applications with Flask

**Microsoft Power BI:**

Power BI by Microsoft is an enterprise information system that offers a variety of statistical capabilities, visualization, and reporting. It can be used to create interactive dashboards and reports that help businesses to make data-driven decisions. Power BI integrates with many data sources such as Excel, SQL Server, and Salesforce. It provides many features such as data modeling, data transformation, and data visualization. ( A. Priyadharshini, 2021)



Figure 14 Power BI.

## 2. Literature Survey

A review of the literature is a systematic and exhaustive review of all types of published work and further sources, such as research papers, to identify as many pertinent works as feasible.

Agricultural production optimization is essential for sustainable food production and increasing farm profitability. To achieve this, it is necessary to analyze and interpret various environmental factors that influence crop yield. In the past few years, crop yield predictions were made using methods based on machine learning and optimize farming practices. This literature survey explores the various studies that have used machine learning techniques to optimize agricultural production.

### **“Optimized Deep Learning Methods For Crop Yield Prediction”**

**Authors: “K. Vignesh, A. Askarunisa and A. M. Abirami”**

**Publication: “Tech Science Press (2022) ”**

This article presents a novel and intriguing method for predicting agricultural production using ecological, agricultural, and crop traits. Using a combination of information mining and deep learning methods, a precise crop yield forecasting system for the entire crop has been developed. The suggested approach, which uses a Finite Deep belief system with the Visual Geometry Group (VGG) Net categorization technique, was found to effectively predict crop yield with a 97 per cent success rate, surpassing current models by sustaining the initial information dispersion. The research also emphasized the significance of examining multiple variables, including the weather, soil quality, levels of water, and field setting, when estimating the yield of crops. The use of the Irregular Deep neural system with the Visual Geometry Group's Net algorithm to categorize the data and anticipate agricultural output is an important breakthrough in the field of crop yield forecast.

This study's suggested approach can be applied to three distinct data sets, which is one of its benefits. The researchers additionally contrasted the used method to three previously announced techniques to assess its effectiveness, which lends credence to the results. There are some limitations to this study. Firstly, it is not clear how the proposed approach performs in different environmental conditions or regions. Secondly, the article does not provide a detailed explanation The modify chick swarm optimization technique utilized to preprocess the input data may hinder its replicability. (Y. Mekonnen, 2020)

**“Intelligent Crop Recommendation System Using Machine Learning”**

**Authors:** “Priyadharshini A, Swapneel Chakraborty, Aayush Kumar, Omen Rajendra Pooniwala ”

**Publication:** “Fifth International Conference on Computing Methodologies and Communication Proceedings (ICCMC 2021) CFP21K25-ART is an IEEE Xplore Part Number. ”

This article presents research that is highly germane to the present difficulties facing India's agricultural industry. Important economic and social effects have resulted from the inability of landowners to select the most suitable crop for their property employing conventional and general methods. The suggested approach, which aids producers in choosing crops by taking into account seeding period, the state of the soil, and geographical position, has the ability to provide substantial benefits to the agriculture industry. Incorporating precise farming with contemporary technological advances represents an important breakthrough in the field. The suggested system has a chance to diminish the likelihood of failure of crops and increase productivity through offering farmers with information that typical farmers do not observe. Additionally, the system protects producers from suffering losses and may eventually contribute to the nation's economic growth. One of the assets that this study has is its potential to include a website and mobile applications that can be used by hundreds of thousands of agricultural producers nationwide for agricultural suggestions. This will significantly increase the reach of the proposed system and provide farmers with valuable information to make informed decisions. There are some limitations to this study. The article fails to offer a comprehensive description of the suggested system's implementation and how the recommendations will be generated. Furthermore, the study does not mention any validation or testing of the proposed system in real-life scenarios. (A. Priyadharshini, 2021)

**“Research And Application Of Machine Learning Method Based On Swarm Intelligence Optimization”**

**Authors:** “Jue Wang,, Yao Dib and Xiao Ruic”

**Publication:** “IOS Press, Journal of Computational Methods in Sciences and Engineering, volume 19, pages S179–S187, 2019.”

The suggested dimensionless random Melanogaster Optimisation forecast framework for agricultural output optimization is a promising method for resolving complex crop production issues. Utilizing chaos theory for population initialization and expanding the search space to three dimensions is a novel approach to enhancing the efficiency of the Drosophila optimization



algorithm. The integration of this algorithm with the stochastic forest model improved the accuracy of predictions for rice insect data sets. The review emphasizes the significance of machine learning in agriculture for crop insect prediction, disease diagnosis, deficiency evaluation, and yield prediction. The proposed algorithm model demonstrates encouraging results in predicting paddy pests. Nevertheless, the authors could further enhance the model's scalability by incorporating multi-population planning. Also stressed is the significance of objective factor variety for crop insect prediction, and the researchers might think about incorporating domain knowledge to enhance their present algorithm. (J. Wang, 2019)

### **“Review Machine Learning Techniques in Wireless Sensor Network Based Precision Agriculture”**

**Author:** “Yemeserach Mekonnen, Srikanth Namuduri, Lamar Burton, Arif Sarwat, z and Shekhar Bhansali ”

**Publication:** “Journal of The Electrochemical Society, 2020 167 037522”

A case study of an IoT-based, data-driven smart farm prototype as a combined water, energy, and food system was also presented. This prototype effectively demonstrated the capacity of AI systems to offer optimal insights for decision-making and action-taking. The construction, equipment, connection procedure, and data collection facilities were described in detail, as well as the development of mobile applications and the underlying data analysis system for forecasting weather, agricultural yield, and quality of crops. Clearly, the use of Internet of Things (IoT) and machine learning in agriculture has tremendous potential for optimizing the use of natural resources and boosting productivity. With the ongoing advancement of technological advances, we can anticipate the use of advanced methods and the creation of increasing sophisticated AI systems in this field. (Y. Mekonnen, 2020)

### **“Predict Crop Production in India Using Machine Learning Technique: A Survey”**

**Author:** “Bhawana Sharma, Jay Kant Pratap Singh Yadav, Sunita Yadav”

**Publication:**

This study offers an in-depth assessment of the current state of wheat crop production as it relates to the use of artificial intelligence in agriculture. This study shows how artificial intelligence can be applied to different aspects of farming management, including yield prediction, identifying diseases, pest identification, and soil and water administration. The survey consists of 39 articles that utilize methods based on image processing and machine learning to solve actual agricultural problems. These studies demonstrate that artificial

intelligence can achieve greater farming precision than conventional methods. Nevertheless, none of the algorithms examined in this study concentrated on crop maturity classification, indicating a significant lacuna in crop management. To address this deficiency, the study proposes a model that employs computational imaging and machine learning methods to determine the phases of crop maturity using digital images. The model that is suggested has an opportunity to provide insightful information about crop maturity and facilitate decision-making. Additionally, the study emphasizes the possibility for future studies to include new deep learning and machine learning methods to achieve even greater farming precision. (B. Sharma, 2020)

### **“A Smart Agriculture IOT System Based On Deep Reinforcement Learning”**

**Author:** “Fanyu Bu a, Xin Wang”

**Publication:** “0167-739X/© 2019 Elsevier B.V. All rights reserved. ”

This paper describes a smart agriculture system that utilizes Net of Things (IoT) and deep learning through reinforcement (DRL) to increase the production of food and decrease resource consumption. The system consists of four layers: farming data collection, peripheral calculating, farm transfer of data, and computing in the cloud, with DRL models incorporated into the cloud layer in order to make wise judgments immediately. The paper examines typical DRL algorithms and their uses in smart farming systems, as well as open challenges and prospective DRL applications in this field. The paper emphasizes the potential of DRL in making intelligent environmental and crop growth decisions. Nevertheless, it recognizes the current limitations of DRL, such as its failure to attain human-level achievement in dynamic environments and complex tasks. The paper proposes multiple avenues toward enhancing DRL performance, including creating incremental approaches to accelerate instruction, incorporating distinct memories to improve logic and thinking, applying operational transferable learning techniques, utilizing cloud computing to enhance instruction effectiveness, and bringing together perform multiple tasks instruction with deep the calculation to enhance adaptability. (F. Bu and X. Wang, 2019)

### **“Crop Yield Prediction Using Machine Learning Algorithm”**

**Author:** “D.Jayanarayana Reddy, Dr M. Rudra Kumar”

**Publication:** “Proceedings of the Fifth International Conference on Intelligent Computing and Control Systems (ICICCS 2021) IEEE Xplore Part Number: CFP21K74-ART; ISBN: 978-0-7381-1327-2”

The present study examines the significance of artificial intelligence (ML) in the application of predicting crop yields (CYP) for enhancing the farming industry in India, where over fifty percent of people is dependent on farmland. The study provides a systematic review of the various ML techniques used for CYP, including neural networks, random forests, KNN regression, and deep learning algorithms such as CNN, LSTM, and DNN. The study also highlights the limitations of these algorithms, such as reduced prediction efficiency and difficulty in capturing nonlinear bonds between input and output variables. The research analyzes the different features used for CYP, such as temperature, weather conditions, crop disease, and classification of crops based on the growing phase, and their impact on the accuracy of prediction. The study suggests that the choice of features should be mainly dependent upon the availability of the dataset, and using more features does not always lead to better results. The study finds that combining machine learning (ML) methods with farming domain knowledge can enhance crop prediction precision.

Additionally, the research identifies a requirement for further enhancements in choosing features, especially in aspects of climate fluctuation factors influencing farmland.

The study suggests that future research should focus on three key objectives: 1) extending specific therapy to bordering terrain areas, and utilizing random predictive methods for the model's construction, and 3) using features from deterministic crop models to obtain perfect statistical CO<sub>2</sub> fertilization. The study also recommends considering the use of fertilizers for soil forecasts to improve crop yield estimation. (Kavita and P. Mathur, 2021)

### **“Crop Selection Method to Maximize Crop Yield Rate Using Machine Learning Technique”**

**Author:** “Rakesh Kumar, M.P. Singh, Prabhat Kumar and J.P. Singh”

**Publication:** “2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, T.N., India. 6 - 8 May 2015. pp.138-145. ”

Agriculture strategy is crucial to the financial growth and nutritional security of an agrarian nation. Crop selection is an essential aspect of agricultural strategy. It is contingent on a number of variables, including production rate, price on the market, and government regulations. Utilizing statistical methods and artificial intelligence methods, a number of investigators examined the forecasting of yield rate, forecasting the weather, classification of soil, and crop categorization for agricultural budgeting. If there are multiple options for planting a crop

employing a limited amount of land, picking a crop becomes a conundrum. This paper proposes a method known as the Crop Picking Method (CSM) to address the crop selection issue, optimize the crop's net production over the season, and accomplish the greatest economic development for the country. The suggested approach may increase crop net yield. This paper presents a method known as CSM for selecting the order of crops to be sown throughout the season. The CSM method could increase the net harvest rate of commodities to be planted during the season. The proposed method settles choosing the crop (s) based on parameter-affected forecast yield rate. (e.g. weather, soil type, water density, crop type). It takes crop, their sowing time, plantation days, and the predicted yield rate for the season as inputs and identifies a sequence of crops with the highest production per day over the course of the season. The efficacy and precision of the CSM method are dependent on the predicted values of the affected parameters; therefore, it is necessary to implement an approach to prediction with higher performance and greater accuracy. (R. Kumar 2015)

**“Machine Learning Applications for Precision Agriculture: A Comprehensive Review”**

**Author:** “ABHINAV SHARMA, ARPIT JAIN, PRATEEK GUPTA, (Student Member, IEEE), AND VINAY CHOWDARY”

**Publication:** “Received December 8, 2020, accepted December 20, 2020, date of publication December 31, 2020, date of current version January 11, 2021. Digital Object Identifier 10.1109/ACCESS.2020.3048415”

This piece provides a comprehensive overview of the uses of artificial intelligence (ML) in precision agriculture, focusing on soil parameter forecasting, crop yield forecasting, illness and vegetation detection, species identification, clever water supply, and managing livestock. The authors emphasize the difficulties the agriculture industry faces in meeting the rising demand for food in light of scarce supplies as well as frequent climatic changes. They stress that agricultural precision, also known as intelligent farming, is a hopeful strategy for overcoming these obstacles. The authors also emphasize the role of IoT-enabled intelligent sensors and motors, imagery from satellites, automated machinery, and drones in the collection of real-time data and autonomous decision-making. The article highlights the significance of sophisticated irrigation and gathering techniques for reducing human labor and increasing crop yield. The authors propose integrating computer vision and machine learning for crop quality tracking and yield estimation. In addition, they illustrate how based on knowledge agriculture can enhance the yield and nutritional value of sustainable crops. (A. Sharma, 2021)

**“Machine Learning in Agriculture: A Review”**

**Author:** “Konstantinos G. Liakos, Patrizia Busato, Dimitrios Moshou, Simon Pearson ID and Dionysis Bochtis. ”

**Publication:** “Sensors 2018, 18, 2674; [www.mdpi.com/journal/sensors](http://www.mdpi.com/journal/sensors)”

This document provides a thorough analysis of the research on the use of artificial intelligence in systems for agricultural production. The authors divided the analyzed operates into crop management, managing livestock, water management, and management of soil and presented their findings according to the amount of items released in different journals, the machine learning (ML) models carried out, and future techniques used in each subcategory. The authors have also discussed the benefits of employing computer learning to sensor data and how it can aid in the development of AI-enabled real-time programs that provide farmers with extensive recommendations and insights for decision support and action. The information is conveyed clearly and concisely, and the paper is well-structured. The authors have utilized a variety of figures and tables to demonstrate their findings, making it simpler for the reader to comprehend the information presented. The paper emphasizes the potential advantages of integrating computerized information documenting, analysis of data, predictive modeling execution, and farming making choices or assistance. (K. G. Liakos, 2018)

**“Energy Neutral Machine Learning Based IOT Device for Pest Detection In Precision Agriculture”**

**Author:** “Davide Brunelli, Andrea Albanese, Donato d’Acunto, and Matteo Nardello”

**Publication:** “IEEE Internet of Things Magazine • December 2019”

This paper presents a creative IoT-based method to identify the the codling moth, a harmful apple crop parasite. The idea put forward employs near-sensor artificial intelligence algorithms to autonomously detect the insect and alert the cultivator if any hazard exists. The system is founded on a low-energy framework that functions continually and independently across low-power devices wide-area networks. The system's hardware is inexpensive and can be readily scaled to accommodate numerous installations in the farmers apple crop. The paper discusses the technical facets of the system, including the equipment employed and the parameters and limitations of the network model. The paper evaluates the efficacy of the system and discusses its energy consumption in order to attain zero energy balance. The system that is suggested has the potential to reduce the time and cost of daily human intervention in trap monitoring, maximize the use of substances, and reduce their environmental impact. This paper is written effectively and provides a thorough analysis of the system that is being suggested. The paper's

results and discussions are compelling while offering helpful insights into the system's viability and efficiency. (D. Brunelli, 2019)

### **“Towards Application of Various Machine Learning Techniques In Agriculture”**

**Author:** “Santosh T. Jagtap, Khongdet Phasinam, Thanwamas Kassanuk, Subhesh Saurabh Jha, Tanmay Ghosh, Chetan M. Thakar”

**Publication:** “2021 Elsevier Ltd. All rights reserved. Selection and peer-review under responsibility of the scientific committee of the 1st International Conference on Computations in Materials and Applied Engineering – 2021. ”

In agriculture, artificial intelligence has enormous potential for increasing crop yields, reducing costs, and enhancing viability. As the amount of data produced by farming operations increases, machine learning programs play a greater role in guiding producers' decisions. Crop disease detection is one of the main utilizes for neural networks in agribusiness. With the aid of computerized imaging algorithms, producers can identify crop diseases and take the necessary preventative measures. A convolutional neural networks (CNN), for instance, can be utilized to categorize crop images as healthy or afflicted with a particular disease. Intelligent irrigation systems are an additional application area for machine learning. By analyzing sensor data such as relative humidity, temperature, and moisture levels in the soil, algorithms that use machine learning can assist farmers in optimizing their water consumption. This can lead to substantial cost savings and enhanced crop yields. For instance, decision tree or algorithms based on reinforcement learning can be used to create optimal irrigation regimens that account for variables such as the climate, soil type, and crop type. The classification of soil is another significant application of machine learning. By evaluating soil data such as appearance, pH, and amount of nutrients, algorithms using machine learning can assist producers in determining the best crops to cultivate in a given region. Machine learning has the potential to transform agriculture by equipping producers with the means to make decisions based on data. With the aid of machine learning computations, producers are able to increase crop yields, decrease expenses, and enhance viability. In the coming decades, machine learning is anticipated to play a growing role in agriculture. (S. T. Jagtap, 2021)

### **“Machine Learning In Agriculture: A Comprehensive Updated Review”**

**Author:** “Lefteris Benos, Aristotelis C. Tagarakis, Georgios Dolias, Remigio Berruto, Dimitrios Kateris and Dionysis Bochtis”

**Publication:** “2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>)”

It focuses on crop, soil, and water issues and managing livestock. Only journal articles published between 2018 and 2020 were included in the review, which adhered to the PRISMA guidelines. The study discovered that learning algorithms, and neural networks with artificial intelligence in particular, were effective at managing the vast quantities of data produced in farms. The most researched topic was crop management, with maize and wheat getting the most frequently examined crops. In addition, the study revealed an increasing curiosity in managing livestock and crop identification. For machine learning applications, remote sensing, such as satellite, autonomous aerial and ground-based machines, was the most prevalent method for gathering data. Between 2018 and 2020, there will be a 164% increase in the total number of studies pertaining to machine intelligence in farming, as indicated by the review. The application of artificial intelligence to farming has an opportunity to increase the availability of food, meet increasing demand from consumers, and reduce the environmental impact. (L. Benos, 2021)

### **“Machine Learning–Based Predictive Farmland Optimization And Crop Monitoring System”**

**Author:** “Marion Olubunmi Adebisi, Roseline Oluwaseun Ogundokun , and Aneoghena Amarachi Abokhai”

**Publication:** “Hindawi Scientifica Volume 2020, Article ID 9428281”

- [1]. The conclusion of the study was that artificial intelligence enhanced decision-making and eased the development of a mobile application using Appery.io. The program accepts user-supplied parameters and provides a variety of optimisation sets, allowing users to optimize information applied on their farms. The system grants farmers entry to a centralized database of relevant data that they can use to optimize and maximize their farmland. The machine learning algorithms embedded in the framework might be enhanced to foresee the mobile application's parameters to aid in decision-making. The study provided an iOS or Android interface that gives producers access to information about their farmland and guarantees them immediate access to the amenities they require. The suggested system has the potential to revolutionize how producers access data and optimize their agricultural land, ultimately resulting in higher crop yields and more efficient farming techniques. Future research could investigate the implementation of

more sophisticated artificial intelligence methods to further enhance the system's efficiency and precision. (M. O. Adebisi, 2020)

### **“Agricultural Production Optimization Engine**

**Author:** “Mrugank Gandhi, Shubham Kothavade, Sushant Nehete, Samarth Arlikar, K. A. Shinde”

**Publication:** “IJCRT 2022 International Journal of Creative Research Thoughts IJCRT | Vol. 10, No. 5 (May 2022), ISSN: 2320-2882”

In this study presents, predicting crop yield using machine learning can be a valuable tool for farmers and policy makers to make informed decisions. By analyzing climatic parameters, soil attributes and previous yield data, machine learning algorithms can help to identify the best crops for a particular region and climate. This can help to improve crop production and contribute to the overall economic growth of the country. The lack of adequate crop planning is one of the main reasons for the low contribution of agriculture towards the GDP of India. Machine learning can help to bridge this gap by providing accurate predictions of crop yield, which can be used to make better decisions about which crops to grow. This can lead to higher yields, increased profits for farmers and a boost to the economy. There is a need for systematic efforts to design and deploy machine learning systems that can predict crop yields accurately. This requires the development of suitable algorithms, the creation of datasets that capture the relevant information about climatic conditions, soil parameters and past yields, and the deployment of the system in the field. (K. G. Liakos, 2018)

### **“Machine Learning In Agriculture Domain: A State-Of-Art Survey”**

**Author:** “Vishal Meshram, Kailas Patil a, Vidula Meshrama, Dinesh Hanchate, S.D. Ramkteke”

**Publication:** “Artificial Intelligence in the Life Sciences 1 (2021) 100010”

The paper emphasizes the significance of farming as a source of fundamental necessities and employment, especially in developing nations like India, where it accounts for 15.4% of the gross domestic product. The paper provides an overview of the most current uses of machine learning in agriculture, which are assisting producers in minimizing costs and increasing the efficacy of pre-harvesting, gathering, and post-harvesting operations. The study divides farming operations into three main categories and outlines the parameters that must be examined at each stage. The paper additionally highlights the requirement for researchers to generate their own information and make it accessible to other people for model validation and testing. Researchers



in the area will find the exhaustive survey of machine learning algorithms used at various phases of farming to be beneficial. The application of machine learning to agriculture is a promising technology that has the potential to transform the industry. It can provide farmers with rich insights and recommendations about their crops, leading to more efficient and precise farming with high-quality production. The study is a valuable resource for researchers and practitioners interested in the application of machine learning in agriculture. (V. Meshram, 2021)

### **“Metaheuristic and Machine Learning-Based Smart Engine for Renting and Sharing Of Agriculture Equipment”**

**Author:** “Manik Rakhra , Ramandeep Singh, Tarun Kumar Lohani , and Mohammad Shabaz”

**Publication:** “Hindawi Mathematical Problems in Engineering Volume 2021, Article ID 5561065”

The article examines the rising trend of robotics replacing human labor in agricultural practices. However, not all farmers are enthusiastic about the integration of advances in technology into agriculture, with some voicing concern about the associated costs and the possibility of decreased yields. To tackle these problems, a Smart Tillage tool has been developed, which employs an input-based decision-making system to suggest affordable solutions to producers. Additionally, the platform contains a metaheuristic technique for optimizing equipment sharing and rental. A survey of 562 farmers revealed a demand for renting and collaborating on equipment due to the cost of debt associated with purchasing new machinery. The Smart Tillage platform seeks to increase the productivity and efficacy of farming while reducing the amount of manual labor and challenging duties producers must perform. In addition, a mobile Online program has been created for advertising, reserving, renting, and sharing agricultural equipment. Smarter Tillage is a prospective vehicle for achieving this objective, as the article finds that technological integration with farming must occur to increase farm output and efficiency. (M. Rakhra, 2021)

### **“Applications Of Machine Learning Techniques In Agricultural Crop Production: A Review Paper”**

**Author:** “Subhadra Mishra, Debahuti Mishra and Gour Hari Santra”

**Publication:** “Indian Journal of Science and Technology, Vol 9(38), DOI: 10.17485/ijst/2016/v9i38/95032, October 2016”

The paper offers a perceptive analysis of the applicability of artificial intelligence methods to farming production of crops. It emphasizes the need for reliable and up-to-date crop production

forecasts in order to make crucial policy choices in areas such as costs, advertising, distribution, and import-export. The paper observes that while previous projections are not unbiased, it is possible to develop statistically sound forecasts by employing methods based on machine learning that make use of vast quantities of data. Several machine learning techniques, including computational neural networks, Data Fuzzy Network, which is the Decision Tree, Regression Analysis, Bayesian belief network, The analysis of time series, Markov chain model, clustering using k nearest neighbor, and support vector machine, are applied to the field related to agriculture in this study. It implies that these techniques are essential for discovering concealed knowledge and analyzing large amounts of data from multiple sources. The article identifies the incorporation of the field of computer science into agriculture as a crucial step towards precise crop predictions. It emphasizes the necessity of developing objective methods for preharvest crop predicting and constructing an example that has a few benefits over conventional forecasting methods. (S. Mishra, 2016)

## **2.1 Reserch gap of the study**

The use of machine learning algorithms such as Decision Tree Regressor and XGBoost Regressor for agricultural production optimization has gained significant attention in recent years. These algorithms can help analyze large amounts of data on environmental factors such as weather, soil moisture, and other variables that impact crop yield. However, there are several potential research gaps that need to be addressed in order to create effective and sustainable agricultural production optimization systems using machine learning.

One possible research gap is the lack of consideration for the socio-economic factors that impact agricultural production. While machine learning algorithms can help optimize crop yield by analyzing data on environmental factors, they may not fully account for the impact of factors such as farmer knowledge and behavior, access to resources like finance and technology, and market conditions. This is especially important for small-scale farmers who may not have access to advanced technologies or voluminous quantities of information.

Another research gap is the lack of attention to the cultural and regional differences in agricultural practices. For example, different regions may have unique soil types, microclimates, and pest management practices that can impact crop yield. Machine learning models may not be able to account for these regional differences without incorporating additional data and input from local experts.

A further research gap is the lack of understanding of the limitations and assumptions of the machine learning algorithms themselves. For example, Decision Tree Regressor and XGBoost Regressor may not be suitable for all types of agricultural data or may require specific

preprocessing steps to be effective. It is important to understand these limitations in order to develop accurate and reliable machine learning models for agricultural production optimization. There is a research gap in the need to create transparent and interpretable machine learning models for agricultural production optimization. The results of machine learning algorithms are often difficult to interpret and can lack transparency, making it difficult for farmers to understand how the algorithm arrived at a particular decision.

### **3. Methodology**

#### **3.1 Methodology**

A comprehensive review of existing literature on machine learning applications in agricultural production optimization will be conducted. The evaluation will concentrate on finding current innovative methods, challenges, and opportunities in agricultural production optimization using machine learning.

Relevant agricultural data from various sources, including weather data, soil data, and crop yield data, will be collected and preprocessed. The collected data will be cleaned, transformed, and integrated into a format suitable for analysis.

The collected data will be used to develop machine learning models using decision tree regressor and XGBoost regressor algorithms. The models will predict crop yields based on soil and weather conditions and optimize the use of resources such as water, fertilizer, and pesticides by recommending the optimal amount and timing of application.

The developed models will be tested and evaluated using historical data. The evaluation will compare the performance of the developed models with traditional methods of agricultural production. The evaluation will also analyze the accuracy and precision of the models and their potential impact on crop yields, resource utilization, and carbon footprint.

A user-friendly interface for the proposed system will be developed. The interface will provide easy access to the predictions and recommendations generated by the machine learning models. The interface will be designed to be accessible to farmers, agronomists, and other stakeholders involved in agricultural production.

The economic and environmental benefits of the proposed system will be evaluated. The analysis will determine the potential savings in cost and the reduction in environmental impact resulting from the adoption of the proposed system. The analysis will also identify the potential barriers to adoption and provide recommendations for overcoming these barriers.

The study will provide recommendations and guidelines for the implementation and adoption of the proposed system in different regions and farming practices. The guidelines will help to ensure the successful implementation of the proposed system and its integration with existing agricultural practices.

The methodology of this study will involve the development of machine learning models using decision tree regressor and XGBoost regressor algorithms to predict crop yields and optimize resource utilization in agricultural production. The developed models will be tested and evaluated, and the economic and environmental impact of the proposed system will be

analyzed. The study will provide recommendations and guidelines for the implementation and adoption of the proposed system in different regions and farming practices.

### 3.1.1 Basic process

- **Define the problem:** The first step is to define the problem you want to solve with machine learning. This involves understanding the business problem, defining the outcome you want to achieve, and identifying the data you need to solve the problem.
- **Collect and prepare the data:** The next stage is to acquire and prepare the necessary data for analysis. This involves cleansing the data, converting it to a format that machine learning algorithms can use, and separating the data into sets for testing and training.
- **Select a model:** The following step is to select a suitable artificial intelligence model for the issue that you wish to solve. This entails comprehending the advantages and disadvantages of different approaches and choosing the one that best fits your information and problem.
- **Evaluate the model:** The following phase involves training the model using the training data. This requires passing the information by means of the algorithm and optimizing its settings to maximize its efficacy.
- **Improve the model:** Once a model has been educated, it is assessed using the data from the test to determine its performance. This includes assessing the model's precision, precision, accuracy recall, and other performance metrics.
- **Deploy the model:** If a model fails to function well, it may be essential to better it by modifying its parameters, choosing alternative features, or employing a different algorithm.
- **Monitor and update the model:** Once the model's efficacy is satisfactory, it is ready for use implemented in an actual-world setting. This involves incorporating the model into the company's procedure and ensuring that it is functioning as intended.

Lastly, it is essential to track the model over a period of time and adjust it when necessary to ensure that it remains to perform when new data grows accessible.

This is a brief description of the artificial intelligence (AI) process; given the particular issue and data, there could be more phases or variants.

### 3.1.2 Dataset

The project developed using various dataset as shown in Figure 15 that are relevant to agricultural production. Some of the important datasets that can be used include USDA National

Agricultural Statistics Service (NASS) data, climate data, soil data, and market data. The NASS data provides comprehensive information on crop yields, production, and other agricultural statistics, that can be utilized for training a machine learning algorithm to maximize farming operations and forecast yields of crops. Weather information such as precipitation, temperature, and pressure may be utilized to forecast crop success. Training a machine learning algorithm with previous climate information from sources such as the National Weather Service (NOAA) is possible. Soil data, such as soil type, pH, and nutrient levels, can also be important in predicting crop yields and optimizing agricultural production. Soil data can be obtained from sources like the USDA's Natural Resources Conservation Service (NRCS). Market data, such as commodity prices and demand, can be used to optimize crop selection and production. Market data can be obtained from sources like the USDA Economic Research Service. These datasets used in combination with the Decision Tree Regressor and XGBoost Regressor algorithms to develop a machine learning model that can predict crop yields and optimize agricultural production. The model trained on the historical data and used to make predictions about future crop yields based on weather, soil, and market conditions.

#### Reading the dataset

```
#Read a comma-separated values (csv,excel) file into DataFrame.
df = pd.read_excel(r'crop_csv_file.xlsx')
#The head() method returns a specified number of rows, string from the top. The head() method returns the first 5 rows if a number is not specified.
df.head()
```

	State_Name	District_Name	Crop_Year	Season	Crop	Temperature	humidity	soil moisture	area	Production
0	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Areca nut	36	35	45	1254.0	2000.0
1	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Other Kharif pulses	37	40	46	2.0	1.0
2	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Rice	36	41	50	102.0	321.0
3	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Banana	37	42	55	176.0	641.0
4	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Cashewnut	36	40	54	720.0	165.0

Figure 15 Dataset

### 3.1.3 EDA (Exploratory Data Analysis)

Exploratory analysis of data is the initial and most crucial stage of any data analysis. EDA is an approach or theory that seeks to identify the most significant and often overlooked trends in a set of data. We analyze the data in an effort to develop a theory. It is utilized by analysts to gain a bird's-eye view of information and make conclusions about it.

EDA is a crucial stage in any information technology endeavor, including those related to agricultural production optimization systems using machine learning. EDA involves examining and understanding the data at hand to discover patterns, relationships, and trends, which can help guide the development and implementation of the optimization system. Here are some

common EDA techniques used in machine learning projects related to agricultural production optimization:

**Data Visualization:** Data visualization is a powerful EDA tool used to understand the distribution, relationship, and variability of different variables in the dataset. Visualizations like scatter plots, line charts, bar graphs, histograms, and box plots can help identify outliers, trends, and patterns that might not be easily visible in the raw data.

Figure 16 of Heatmap association is an illustration of the relationship among factors in a dataset. A relationship heatmap can help determine what factors are either adversely or positively linked, and how strong that correlation is. This information can be useful in feature selection and model building.

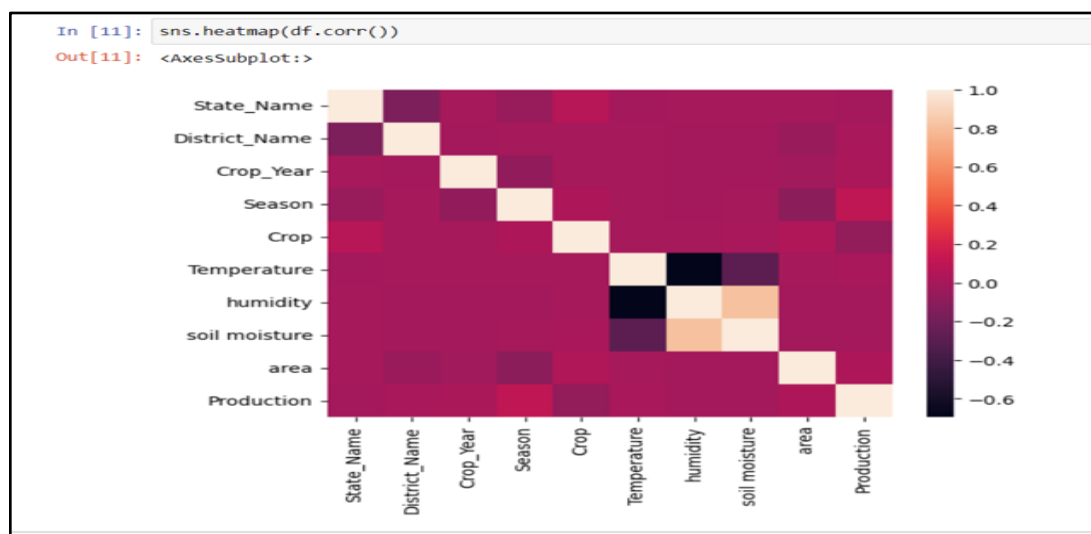
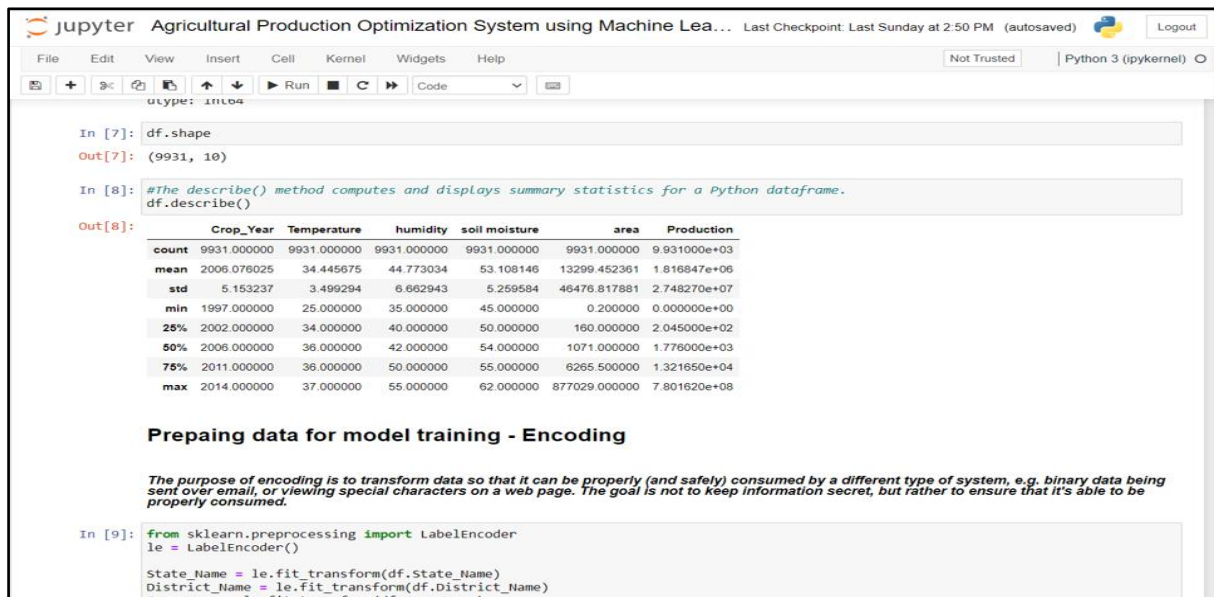


Figure 16 Heatmap Correlation

**df.describe ():** Figure 17 show the df.describe() function provides summary statistics for numerical variables in The information set, which includes the average, standard deviation, minimum and highest values, as well as quartiles. This information can help identify outliers, extreme values, and trends in the data.



```

In [7]: df.shape
Out[7]: (9931, 10)

In [8]: #The describe() method computes and displays summary statistics for a Python dataframe.
df.describe()
Out[8]:

```

	Crop_Year	Temperature	humidity	soil moisture	area	Production
count	9931.000000	9931.000000	9931.000000	9931.000000	9931.000000	9.931000e+03
mean	2006.076025	34.445675	44.773034	53.108146	13299.452361	1.816847e+06
std	5.153237	3.499294	6.662943	5.259584	46476.817881	2.748270e+07
min	1997.000000	25.000000	35.000000	45.000000	0.200000	0.000000e+00
25%	2002.000000	34.000000	40.000000	50.000000	160.000000	2.045000e+02
50%	2006.000000	36.000000	42.000000	54.000000	1071.000000	1.776000e+03
75%	2011.000000	36.000000	50.000000	55.000000	6265.500000	1.321650e+04
max	2014.000000	37.000000	55.000000	62.000000	877029.000000	7.801620e+08

**Prepaing data for model training - Encoding**

*The purpose of encoding is to transform data so that it can be properly (and safely) consumed by a different type of system, e.g. binary data being sent over email, or viewing special characters on a web page. The goal is not to keep information secret, but rather to ensure that it's able to be properly consumed.*

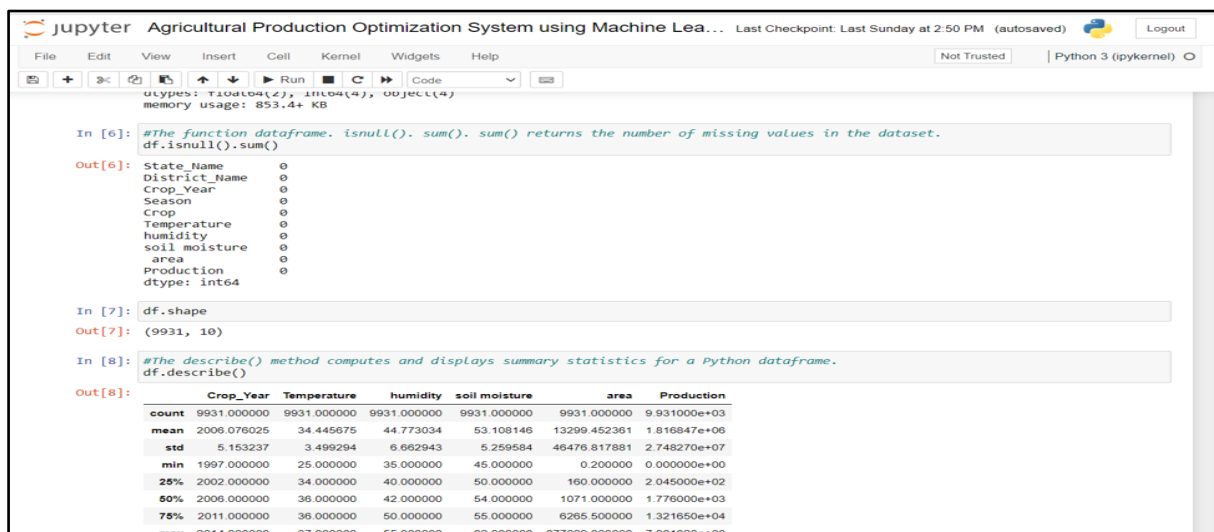
```

In [9]: from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
State_Name = le.fit_transform(df.State_Name)
District_Name = le.fit_transform(df.District_Name)

```

Figure 17 Prepaing Data for Model Training – Encoding

**df.isnull().sum():** Figure 18 shows the `df.isnull().sum()` function returns the total number of missing values for each variable in the dataset. This knowledge can assist in identifying absent data that might need interpolation or elimination prior to computing.



```

In [6]: #The function dataframe.isnull().sum().sum() returns the number of missing values in the dataset.
df.isnull().sum()
Out[6]:

```

	State_Name	District_Name	Crop_Year	Season	Crop	Temperature	humidity	soil moisture	area	Production
	0	0	0	0	0	0	0	0	0	0

```

In [7]: df.shape
Out[7]: (9931, 10)

In [8]: #The describe() method computes and displays summary statistics for a Python dataframe.
df.describe()
Out[8]:

```

	Crop_Year	Temperature	humidity	soil moisture	area	Production
count	9931.000000	9931.000000	9931.000000	9931.000000	9931.000000	9.931000e+03
mean	2006.076025	34.445675	44.773034	53.108146	13299.452361	1.816847e+06
std	5.153237	3.499294	6.662943	5.259584	46476.817881	2.748270e+07
min	1997.000000	25.000000	35.000000	45.000000	0.200000	0.000000e+00
25%	2002.000000	34.000000	40.000000	50.000000	160.000000	2.045000e+02
50%	2006.000000	36.000000	42.000000	54.000000	1071.000000	1.776000e+03
75%	2011.000000	36.000000	50.000000	55.000000	6265.500000	1.321650e+04
max	2014.000000	37.000000	55.000000	62.000000	877029.000000	7.801620e+08

Figure 18 Function returns the total number of missing values for each variable in the dataset.

**df.dropna():** Figure 19 shows the `df.dropna()` function is used to remove rows or columns with missing values from the dataset. This function can help reduce the noise in the data and enhance the simulation's fidelity.



**Handling Missing Data**

```
In [5]: #The dropna() method removes the rows that contains NULL values. The dropna() method returns a new DataFrame object unless the in
df = df.dropna()
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9931 entries, 0 to 9999
Data columns (total 10 columns):
#   Column          Non-Null Count  Dtype
---  -
0   State_Name      9931 non-null   object
1   District_Name   9931 non-null   object
2   Crop_Year       9931 non-null   int64
3   Season         9931 non-null   object
4   Crop           9931 non-null   object
5   Temperature     9931 non-null   int64
6   humidity        9931 non-null   int64
7   soil moisture   9931 non-null   int64
8   area            9931 non-null   float64
9   Production      9931 non-null   float64
dtypes: float64(2), int64(4), object(4)
memory usage: 853.4+ KB
```

**Figure 19** Function is used to remove rows or columns with missing values from the dataset

**df.info():** Figure 20 shows the df.info() function provides information on the data types, number of non-null values, and memory usage of each variable in the dataset. This information can help identify potential data type mismatches, memory issues, and other inconsistencies in the dataset.

```
df.info()
```

```
[95]
... <class 'pandas.core.frame.DataFrame'>
RangeIndex: 3730 entries, 0 to 3729
Data columns (total 8 columns):
#   Column          Non-Null Count  Dtype
---  -
0   Unnamed: 0      3730 non-null   int64
1   State_Name      3730 non-null   object
2   District_Name   3730 non-null   object
3   Crop_Year       3730 non-null   int64
4   Season         3730 non-null   object
5   Crop           3730 non-null   object
6   Area           3730 non-null   float64
7   production      3730 non-null   float64
dtypes: float64(2), int64(2), object(4)
memory usage: 233.2+ KB
```

**Figure 20** Function provides information on the data types, number of non-null values , and memory usage of each variable in the dataset

The project can use various algorithms to develop a machine learning model that can predict crop yields and optimize agricultural production. Two such algorithms that can be used are the Decision Tree Regressor and the XGBoost Regressor.

The Decision Tree Regressor is an algorithm for artificial intelligence that generates a tree-structured model. The tree framework includes nodes representing features, limbs representing choices, and branches representing the anticipated result. The Decision Tree's Regressor can

predict the yield of crops according to meteorological conditions, soil, and market conditions. The algorithm makes decisions based on the features in the data and uses these decisions to create a tree structure that predicts the outcome.

The XGBoost Regressor is a method for machine intelligence that is utilized to solve error issues. This method extends the Gaussian Enhancing method and is designed to improve the performance of traditional Gradient Boosting. The XGBoost Regressor can be used to predict crop yields based on weather, soil, and market conditions. The algorithm is optimized for large datasets and can handle missing data and noisy data.

Both algorithms can be used in combination with various datasets to develop a machine learning model that can predict crop yields and optimize agricultural production. The algorithms can be trained on historical data and used to make predictions about future crop yields based on weather, soil, and market conditions. The Decision Tree Regressor and the XG Boost Regressor are both effective algorithms for agricultural production optimization using machine learning.

### 3.1.3.1 Decision Tree Regressor

The decision tree shown in figure 21 is one of the most common machine learning methods for resolving regression and classification problems. As the name implies, the algorithm employs a tree-like framework of choices to determine either the desired value (regression) or the class of interest. (classification). Before delving into the operation of decision chains, let's initial become conversant with their terminology:

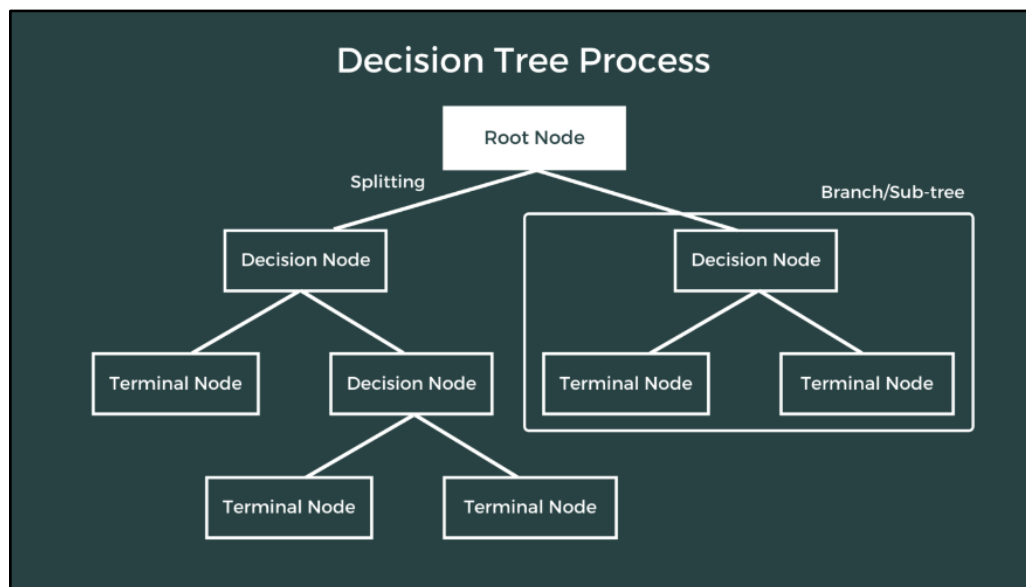


Figure 21 Decision Tree Regressor

The process of dividing begins at the root nodes, followed by a branching tree, and ends at the leaf node (terminal node) which holds the method's prediction or ultimate result. Typically, decision chains are constructed from the top down, with each phase selecting a variable that

best divides the set for items. Each subtree of the model of decision trees can be depicted as a binary structure in which the decision node divides into two nodes according to the conditions.

In this course, we will talk about trees of regression, which are decision trees in which the objective value or terminal component can accept values that are continuous (typically real numbers). These trees are known as classification trees if the objective variable can assume an independent set of values.

The choice tree regressor is a form of algorithm for supervised learning that is utilized for regression tasks. The algorithm builds a choice tree model using the provided data set. The choice tree model is a structure resembling a tree in which every internal node depicts a test on a property, every branch represents the test's outcome, and every node in the leaf depicts a class label or an ongoing value.

The equation for Decision Tree Regressor algorithm can be represented as:

Step 1: Initialize the decision tree with the root node.

Step 2: Split the root node into two or more child nodes based on the best split criterion.

Step 3: Recursively repeat step 2 on each child node until the Ending criteria have been fulfilled.

Criteria for ending can be founded on :

- Highest tree depth
- Minimal amount of examples necessary to divide a node
- Lowest necessary number of observations at the leaf node

Step 4: Assign a regression value to each leaf node.

The splitting criterion can be based on various Measure square error (MSE) and the mean absolute error (MAE) are measures), or other criteria.

Decision tree regressor is a popular machine learning algorithm used for regression problems, including agricultural production optimization. Here are the equations for the Decision tree regressor algorithm:

- **Splitting Criterion:** In Decision tree regressor, the objective is to create a tree that best splits the data into subsets. The criterion for measuring the quality of a split can be defined as:

$$J = MSE_{left} + MSE_{right}$$

Equation 1

where,

$MSE_{left}$ : the mean squared error of the left subset

$MSE_{right}$ : the mean squared error of the right subset

- **Tree Building:** In Decision tree regressor, the decision tree is constructed recursively. The algorithm commences with a single node representing the full dataset. The node is then divided into two offspring nodes based on the criteria for dividing. Until a halting criterion is met, such as a maximal tree depth or minimal number of samples per leaf, the procedure persists.
- **Prediction Function:** The prediction function of Decision tree regressor is the average of the target values in the leaf node that a sample falls into. The predicted target value can be represented as:

$$\hat{y} = \frac{1}{N} * \sum_i y_i$$

Equation 2

where,

N: the number of samples in the leaf node

$y_i$ : the target value of the i-th sample

- **Tree Pruning:** Decision trees are prone to overfitting, which can lead to poor generalization performance on new data. To prevent overfitting, Decision tree regressor uses a technique called pruning, which involves removing branches from the tree that do not improve the model's performance on the validation set.
- **Hyperparameter Tuning:** Several the hyperparameters of the decision tree's regressor can be adjusted to enhance the model's efficiency. Included among these are the maximal tree depth, minimal number of data per leaves, and division criterion. These formulas serve as the foundation for the decision tree regressor method, which may be utilized for forecasting future outcomesm the yield of crops in an agricultural production optimization system.

### 3.1.3.2 Grid Search CV

Grid Search CV (Cross-Validation) is a machine learning method for enhancing a model's hyperparameters. The learning rate, regularization factor, and number of hidden layers in a neural network are examples of hyperparameters, which are parameters that are established before the model is trained. The performance of the model may be significantly impacted by these hyperparameters. Grid Search CV entails building a grid of possible hyperparameters values, training the model with each set of possible hyperparameters, and analysing the results. Each hyperparameters is given a range of values, and all potential combinations of hyperparameters are then tested using a search algorithm to build the grid. Cross-validation,

which divides the data into subsets and uses each subset in turn as a validation set, is used to train and test the model for each combination of hyperparameters. The remaining data is utilized for training. This makes it possible to prevent the model from being over fit to the training set of data. The model with the best performance is chosen as the final model after all possible combinations of hyperparameters have been tried. Although Grid Search CV is an effective tool for hyperparameter optimization, it can be computationally expensive, particularly for large datasets and complex models. However, it is frequently combined with other methods like random search or Bayesian optimization to increase its effectiveness in machine learning.

### ***3.1.3.3 XGBoost Regressor***

XGBoost, or Extreme Gradient Boosting, is a library that is open-source that implements the gradient enhancing algorithm in a way that is effective and efficient.

Immediately after its creation and initial discharge, XGBoost got the go-to method and frequently the most important aspect of successful artificial intelligence competition solutions to a variety of issues.

Predicting a number, such as an amount of money or a height, is the goal of logistic modeling for predictive purposes. XGBoost can be utilized immediately for predictive modeling of regression.

XGBoost (Extreme Gradient Boosting) is a frequently utilized artificial intelligence algorithm for classification and regression issues. It is a method of collective learning that integrates numerous decision trees to produce a robust forecasting model.

The XGBoost regressor iteratively adds decision trees to a model using gradient boosting, with each new tree increasing the model's prediction accuracy. The algorithm operates by reducing a loss product that quantifies the disparity among the expected and actual data used for training values. Additionally, XGBoost employs methods of regularization to prevent overfitting.

**Some key features of the XGBoost regressor include:**

- **Speed and scalability:** XGBoost is designed to be quick and effective, even with huge data sets containing numerous features.
- **Flexibility:** The method supports a broad array of function objectives and assessment measures, making it suitable for many different types of regression problems.
- **Robustness:** XGBoost is resistant to overfitting and can handle missing values in the input data.
- **Interpretable:** XGBoost provides feature importance scores, which can help to identify which features are most predictive for a given regression problem.

XGBoost (Extreme Gradient Boosting) is a gradient boosting algorithm used for regression and classification tasks. It is an ensemble algorithm that combines multiple decision trees into a single model. It uses a combination of regularization techniques to prevent overfitting.

The equation for XGBoost Regressor algorithm can be represented as:

Step 1: Initialize the model with a single decision tree.

Step 2: Calculate the gradient and hessian of the loss function for each training sample.

Step 3: Use the gradient and hessian to construct a new decision tree that best fits the residual errors.

Step 4: Add the new decision tree to the ensemble model.

Step 5: Repeat steps 2 to 4 until the desired number of trees is reached or until the stopping criterion is met.

Stopping criterion can be based on:

- Maximum depth of the trees
- Minimum number of samples required to split a node
- Minimum number of samples required to be at a leaf node

Step 6: Predict the output value by summing up the predictions of all the decision trees in the ensemble model.

The loss function used in XGBoost Regressor can be any differentiable loss function like mean squared error (MSE) or mean absolute error (MAE).

XGBoost (eXtreme Gradient Boosting) is a powerful machine learning algorithm that is commonly used for regression problems. In the context of agricultural production optimization, XGBoost can be used to predict the yield of a particular crop based on various input features such as soil quality, weather conditions, irrigation techniques, etc. Here are the equations for the XGBoost regression algorithm:

- **Objective Function:** The objective function of XGBoost regression is defined as the sum of the loss function and regularization term.

$$obj = L(y, \hat{y}) + \Omega(\hat{y})$$

Equation 3

where,

y: actual target value

$\hat{y}$ : predicted target value

L: loss function (such as mean squared error, mean absolute error, etc.)

$\Omega$ : regularization term (such as L1 or L2 regularization)

- **Prediction Function:** The prediction function of XGBoost regression is the sum of the base prediction and the output of the decision trees.

$$\hat{y} = \sum_i F_i(x) + b$$

Equation 4

where,

$F_i(x)$ : the i-th decision tree

b: the base prediction (usually the mean target value)

- **Tree Building:** The decision trees in XGBoost are built iteratively, where each tree is built to minimize the objective function. The tree is split into two nodes, left and right, based on the feature that gives the greatest reduction in the objective function. The process continues until a stopping criterion is met, such as a maximum tree depth or minimum number of samples per leaf.
- **Gradient and Hessian Calculation:** In order to update the weights of the decision trees, the gradient and Hessian of the objective function are calculated with respect to the predicted target values.

$$\frac{\partial L}{\partial \hat{y}} = g(y, \hat{y})$$

Equation 5

$$\frac{\partial^2 L}{\partial \hat{y}^2} = h(y, \hat{y})$$

Equation 6

where,

g: the first derivative of the loss function

h: the second derivative of the loss function

Weight Update:

The weights of the decision trees are updated by minimizing the objective function using gradient descent.

$$w = -\left(\frac{\sum_i g_i}{(\sum_i h_i + \lambda)}\right)$$

Equation 7

where,

$w$ : the weight of the decision tree

$g_i$ : the gradient of the objective function for the  $i$ -th sample

$h_i$ : the Hessian of the objective function for the  $i$ -th sample

$\lambda$ : the regularization parameter

These equations form the basis of the XGBoost regression algorithm, which can be used for predicting the yield of crops in an agricultural production optimization system.

### 3.1.4 Reason for using decision tree regressor

The Decision tree regressor is a popular machine learning algorithm that can be used in the research on "Agricultural production optimization system using machine learning" for various reasons:

- **Easy to interpret:** Decision trees are easy to interpret and understand. They provide a clear and intuitive visual representation of the decision-making process, making it easier for farmers and other stakeholders to understand the factors that affect crop yield.  
Can handle categorical and numerical data: Decision trees can handle both categorical and numerical data, making them suitable for agricultural datasets that contain a mix of different types of variables, such as weather data, soil data, and farming practices.
- **Identify important variables:** Decision trees can identify the most important variables for predicting the target variable, such as crop yield. This can help to identify the most relevant factors that affect crop yield, such as soil properties, weather conditions, and farming practices, and provide insights for improving crop management practices.
- **Can handle non-linear relationships:** Decision trees can handle non-linear relationships between variables, which is common in agricultural datasets. For example, the relationship between soil moisture and crop yield may not be linear, and a decision tree can capture this non-linear relationship more accurately than a linear regression model.
- **Can handle missing data:** Decision trees can handle missing data by imputing missing values or treating missing values as a separate category. This is important in agricultural datasets, which may have missing data due to factors such as sensor malfunction or data collection errors.

The Decision tree regressor is a powerful and flexible algorithm that can be used to identify the most important factors affecting crop yield and provide insights for improving crop management practices, increasing productivity, and reducing costs in the agricultural sector.



### 3.1.5 Reason for using XGBoost regressor

The XGBoost regressor is a well-known and potent algorithm for machine learning that is typically applied to regression-related issues, such as forecasting constant numbers. It is a form of algorithm for gradient boosting that makes forecasts using a collection of choice trees.

There are several reasons why XGBoost regressor is a suitable algorithm for agricultural production optimization systems:

- **Handling non-linear relationships:** Agricultural production is a complex process that involves many factors that interact in non-linear ways. XGBoost regressor is capable of capturing these non-linear relationships, making it a good choice for predicting crop yields or identifying optimal farming practices.
- **Handling missing data:** Agricultural data is often incomplete due to factors such as weather conditions, soil moisture, and other variables that can impact crop growth. XGBoost regressor can handle missing data, making it a good choice for agricultural data analysis.
- **Scalability:** XGBoost is highly scalable, making it suitable for large agricultural datasets. This is important because agricultural data can be very large, especially when multiple variables are involved.
- **Feature importance:** XGBoost can provide information about the relative importance of different features, making it easier to identify the factors that have the biggest impact on agricultural production.

XGBoost regressor is a powerful and versatile machine learning algorithm that can be used to optimize agricultural production systems. Its ability to handle non-linear relationships, missing data, and large datasets, as well as its ability to provide information about feature importance, make it a strong choice for agricultural data analysis.

### 3.2 Data collection and Pre-processing:

Data collection is the first step in any machine learning project. In this project, we will collect data related to crop production, weather conditions, and soil quality. We will use this data to build models that can predict the yield of crops accurately. The data will be collected from various sources, including government agencies, research papers, and online databases.

once the data has been acquired, it must be processed in order to be ready for analysis. This involves cleansing the data, dealing with absent values, and converting the data into an appropriate format for analysis. The preliminary processing phase is crucial because it serves to ensure the accuracy and dependability of the models.

The pre-processing step will involve the following:

- **Data Cleaning and Quality Control:** The first step in pre-processing is to clean the data and perform quality control. This involves removing any duplicates, correcting errors, and identifying outliers. We will also perform quality control checks to ensure that the data is consistent and accurate.
- **Feature Selection and Engineering:** The next step is to select the relevant features that will be used in the analysis. This involves identifying the variables that have a significant impact on crop production. We will also perform feature engineering to develop novel capabilities that will enhance the efficacy of the algorithms.
- **Data Normalization and Standardization:** The final step in pre-processing is to normalize and standardize the data. This involves transforming the data so that it has a mean of zero and a standard deviation of one. This step is critical as It ensures that the projections have no bias to specific variables.

### 3.2.1 Model Training and Evaluation:

Once the data is pre-processed, the next step is to train and evaluate the models. We will use two different models in this project: Decision tree regressor and XGBoost regressor.

- **Model Selection and Training:** The first step in model training is to select the appropriate model for the problem. We will use Decision tree regressor and XGBoost regressor as they are widely used for regression problems.

Once the model is selected, we will train it on the pre-processed data. The training process involves fitting the model to the data and optimizing its parameters to improve its performance.

- **Performance Metrics:** Several indicators of performance, including the mean absolute error (MAE), Root Mean Squared Error (RMSE), and R-squared, are going to be used to evaluate the models' efficacy. (R<sup>2</sup>). These parameters serve to assess the precision and dependability of the algorithms.
- **Validation:** To ensure that the models generalize well on new data, we will use two validation techniques: Holdout Validation and Cross-Validation. Holdout validation involves splitting the process of cross- consists of partitioning the information into various subsets and then training a model for every subgroup.
- **Optimization:** To enhance the efficacy of the models, we are going to alter their hyper parameters using methods such as Grid Search and Randomised Search. This entails

meticulously testing different combinations of hyper parameters to identify the optimal set that enhances the performance of the models.

In summary, the approach used for this undertaking includes data collection and preliminary processing, model selection and retraining, evaluating their performance using various metrics, and optimizing the models to improve their performance. We will use techniques such as holdout validation, cross-validation, and hyperparameters tuning to ensure that the models are accurate and reliable.

### 3.2.2 Data sources

- **Open data repositories:** Open data repositories are a great source of data for machine learning projects, especially for those related to agriculture. In this project, we will use data from several open data repositories, including:
  - **FAOSTAT:** The Food and Agriculture Organization of the World Health Organization (FAO) maintains this global collection of agricultural information). The database contains data on crop production, area harvested, yield, and other agricultural indicators.
  - **USDA:** The United States Department of Agriculture (USDA) provides open access to a variety of data sets related to agriculture, including crop production, soil data, and climate data.
  - **World Bank:** The World Bank provides open access to data related to agriculture, including crop production, land use, and agricultural inputs.
- **Field surveys and experiments:** Field surveys and experiments are a valuable source of data for machine learning projects related to agriculture. In this project, we may conduct our own field surveys and experiments to collect data on crop yields, weather conditions, and soil quality. We will collect data from multiple locations to ensure that the models generalize well.
- **Remote sensing and GIS data:** Remote sensing and Geographic Information Systems (GIS) data can provide valuable information on crop yields, soil quality, and weather conditions. In this project, we will use remote sensing data from satellites to monitor crop growth and detect anomalies. We will also use GIS data to analyze the impact of soil quality and weather conditions on crop yields.

Some examples of remote sensing and GIS data sources that we may use in this project include:

- **Landsat:** This is a satellite-based remote sensing program that provides high-resolution imagery of the earth's surface. We can use Landsat data to monitor crop growth and detect anomalies.
- **MODIS:** The Moderate Resolution Imaging Spectroradiometer (MODIS) is another satellite-based remote sensing program that provides data on vegetation, land surface temperature, and other environmental variables.
- **Soil Data Access:** The Soil Data Access system provides access to soil data from the National Cooperative Soil Survey. We can use this data to analyze the impact of soil quality on crop yields.

The National Oceanic and Atmospheric Administration (NOAA) provides climate data online.) provides climate data through its Climate Data Online (CDO) system. We can use this data to analyze the impact of weather conditions on crop yields.

We will use a variety of data sources, including open data repositories, field surveys and experiments, and remote sensing and GIS data to build our agricultural production optimization system. We will pre-process and clean the data, perform feature selection and engineering, and train and evaluate our models using performance metrics such as MAE, RMSE, and R2.

### 3.2.3 Data cleaning

Data cleaning is a crucial step in any data analysis project, especially when dealing with agricultural production optimization systems using machine learning. In this article, we will discuss the importance of data cleaning and how it can impact the precision and dependability of the generated results by two popular machine learning algorithms - Decision Tree Regressor and XGBoost Regressor.

- **Importance of Data Cleaning:** Data cleaning is the process of identifying and correcting or removing errors, inconsistencies, and inaccuracies in the data. In agricultural production optimization systems. The training data for models based on machine learning can originate from an assortment of sources., including sensors, weather stations, and human inputs. This data may contain errors, missing values, or outliers that can negatively impact the accuracy of the machine learning model.

For example, if the data used to train a machine learning model contains missing values, the model may not be able to accurately predict the output variable. Similarly, if the data contains outliers, the model may be skewed towards these values, leading to inaccurate predictions.

- **Data Cleaning Techniques:** There are several techniques that can be used for data cleaning, including:
  - **Handling Missing Values:** The absence of values can be managed either by deleting the columns or rows containing the missing values or by assigning the absent numbers with the column's mean or middle.
  - **Handling Outliers:** Outliers can be handled by either removing them or by substituting these with values that fall within the bounds of the column's additional numbers.
  - **Handling Inconsistencies:** Inconsistencies can be handled by locating and rectifying mistakes in data like typos or improper data forms.
  - **Scaling and Normalization:** Scaling and normalization can be used to ensure that the information is on a comparable size, that can enhance the artificial intelligence model's efficacy.
- **Using Decision Tree Regressor:** Decision Tree Regressor is a well-known artificial intelligence algorithm for classification applications. It operates by dividing the data sequentially into groups, where each subset represents a node in the decision tree. The algorithm chooses the best feature to split the data based on the reduction in variance. When using Decision Tree Regressor for agricultural production optimization systems, It is essential for guaranteeing that the information is accurate and error-free. This is because Decision Tree Regressor is sensitive to noise and outliers in the data, which can negatively impact the accuracy of the model.
- **Using XGBoost Regressor:** Popular artificial intelligence algorithm employed in correction assignments is XGBoost Regressor. It is an extension of the Gradient Boosted Trees algorithm and works by fitting an ensemble of decision trees to the data. When using XGBoost Regressor for agricultural production optimization systems, It is essential for guaranteeing that the information is accurate and error-free. This is because XGBoost Regressor is also sensitive to noise and outliers in the data, which can negatively impact the accuracy of the model.

Data cleaning is a critical step in any data analysis project, especially when dealing with agricultural production optimization systems using machine learning. In this article, we discussed the importance of data cleaning and how it can impact the precision and dependability of the produced findings by two popular machine learning algorithms - Decision Tree Regressor and XGBoost Regressor. By ensuring that the data is clean and free from errors, we can improve the performance of these algorithms and generate more accurate and reliable results.

### 3.3 Feature selection

The choice of features is a crucial aspect of any artificial intelligence endeavor, especially when dealing with agricultural production optimization structures. The following piece will examine the significance of choosing features and how it can impact the accuracy and performance of two popular machine learning algorithms - Decision Tree Regressor and XGBoost Regressor.

- **Importance of Feature Selection:** The choice of features is a method of choosing those with the most pertinent characteristics from a collection of data that are useful for predicting the output variable. In agricultural production optimization systems, the dataset can contain a large number of features, including environmental and weather data, soil conditions, and crop growth stages. However, not all of these features may be relevant or useful for predicting the output variable, such as crop yield or soil moisture levels.

Incorporating extraneous or redundant characteristics can have a negative effect on the effectiveness of a model developed using machine learning, resulting in excessive fitting, delayed retraining durations, and decreased accuracy. Therefore, it is crucial to conduct feature selection in order to determine the most significant and appropriate characteristics.

- **Feature Selection Techniques:** There are numerous methods available for choosing features, such as:

The method includes choosing features which have a significant relationship with the resultant measure and eliminating characteristics that have highly linked with one another.

Recursive Feature Removal: This method involves eliminating the fewest important characteristics repeatedly till the ideal amount of attributes is attained.

Principal component analysis (PCA): This method entails transforming a dataset into a space with fewer while retaining the most essential information.

- **Using Decision Tree Regressor:** Decision Tree Regressor is a popular machine learning algorithm that is used for regression tasks. It works by recursively partitioning the data into subsets, where each subset represents a node in the decision tree. The algorithm chooses the best feature to split the data based on the reduction in variance. When using Decision Tree Regressor for agricultural production optimization systems, It is essential to conduct selection of features in order to determine which are the most crucial and pertinent characteristics. This is because Decision Tree Regressor is

sensitive to irrelevant or redundant features, which can negatively impact the accuracy of the model.

- **Using XGBoost Regressor:** Prominent learning algorithm used for correction assignments is XGBoost Regressor. It is an extension of the Gradient Boosted Trees algorithm and works by fitting an ensemble of decision trees to the data.

When using XGBoost Regressor for agricultural production optimization systems, it is also important to perform feature selection to identify the most important and relevant features. This is because XGBoost Regressor is also sensitive to irrelevant or redundant features, which can negatively impact the accuracy of the model.

Feature selection, particularly for artificial intelligence projects, is a crucial stage when dealing with agricultural production optimization systems using machine learning algorithms like Decision Tree Regressor and XGBoost Regressor. By performing feature selection, we can identify the most important and relevant features, reduce overfitting, and enhance the precision and efficiency of the simulations.

### 3.3.1 Model Training and Evaluation

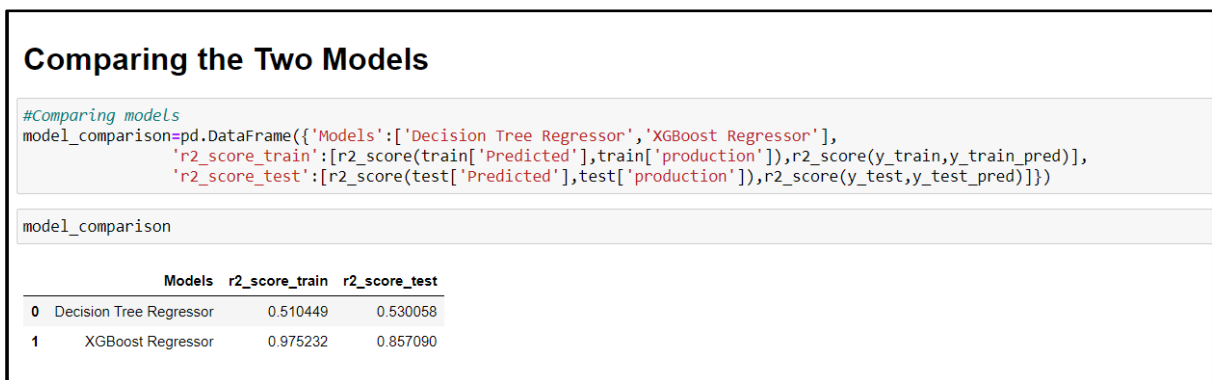
Model training and evaluation are critical steps in any machine learning project. In this article, we will discuss how to train and evaluate two popular machine learning algorithms Decision Tree Regressor and XGBoost Regressor for an agricultural production optimization system.

- **Training the Models:** To train the models, we need a dataset that contains relevant features and the corresponding output variable, such as crop yield or soil moisture levels. We will use the cleaned and preprocessed dataset from the previous stages of the project.
  - **Decision Tree Regressor:** To train the Decision Tree Regressor model, we can use the scikit-learn library in Python. We can import the DecisionTreeRegressor class and create an instance of the class. We can then utilizing the fit() technique, match the predictions to the experimental dataset.
  - **XGBoost Regressor:** To train the XGBoost Regressor model, we can use the xgboost library in Python. We can import the XGBRegressor class and create an instance of the class. The prediction model can then be fitted to the initial data employing the fit() procedure.
- **Evaluating the Models**

After training the models, we need to evaluate their performance to determine how well they can predict the output variable.

- **Decision Tree Regressor:** To evaluate the Decision Tree Regressor model, we can use metrics such as mean absolute error (MAE), mean squared error (MSE), and R-squared (R2) score. We can use the predict() method to make predictions on the test data and compare them to the actual values.
- **XGBoost Regressor:** To evaluate the XGBoost Regressor model, we can use the same metrics as for the Decision Tree Regressor model. We can use the predict () method to make predictions on the test data and compare them to the actual values.

Training and evaluating machine learning models such as Decision Tree Regressor and XGBoost Regressor for agricultural production optimization systems is critical to ensure accurate predictions. By using appropriate metrics utilizing metrics such as MAE, MSE, and R2 rating, the effectiveness of the models can be measured and identify any areas for improvement. The SciKit-learn and XGBoost libraries in Python provide powerful tools for training and evaluating these models, and should be explored further for more complex agricultural production optimization systems.



**Figure 22 Model Comparison**

- **Performance Metrics:** When assessing artificial intelligence models for crop efficiency structures, it is crucial to gauge their precision with appropriate performance metrics. Here are three frequently employed metrics:
  - **Mean Absolute Error (MAE):** MAE is an average of the absolute variances between the actual and predicted values. It assesses how closely the average predictions match the actual values. A lower MAE signifies improved performance.
  - **Root Mean Squared Error (RMSE):** RMSE is the square root of the mean of the squared variances among the actual and predicted values. Larger errors are punished more severely than minor ones. The smaller the RMSE, the better the performance.



- **R-squared (R<sup>2</sup>):** R<sup>2</sup> quantifies the proportion of the objective variable's variance that the model explains. A greater value indicates a superior match. A value of 1 shows a flawless fit, whereas a value of 0 suggests the predictive model fails to account for any of the diversity in the desired variable.

While assessing a model developed using machine learning for a crop manufacturing optimization framework, it is essential to evaluate the issue's particular demands and select an appropriate metric. For example, if the focus is on reducing errors in yield predictions, MAE and RMSE may be more appropriate. If the focus is on explaining the variability in yield, R<sup>2</sup> may be a more suitable metric.

### 3.3.2 Model Validation

Once the agricultural production optimization system using machine learning models have been trained, validate the efficacy of these approaches is the next stage. Validation of model is an essential phase in the process of machine learning because it ensures that the algorithms are reliable and precise. In this chapter, we will discuss two commonly used model validation techniques, Holdout Validation and Cross-validation.

- **Holdout Validation:** Holdout Validation is a simple and straightforward technique for validating Supervised instructional frameworks. In this technique, the dataset is divided into two sets, a training set and a validation set. On the instruction set, the model is trained, and then its efficacy will be assessed on the set for validation. The validation set consists of data that has not been used to train the model. The accuracy and dependability of the model is determined by the efficacy measures used for assessing it on the set of validation data.

In our crop maximization system, machine learning model success will be validated using holdout confirmation. The dataset is going to be arbitrarily split into two distinct sets: a set for training and a set for validation. On our training set, we are going to employ the Decision tree regressor and the XGBoost regressor neural network models to train the system. On the validation set, we are going to assess the efficacy of the models. The mean squared error (MSE) and a root mean squared error (RMSE) will be used to evaluate the efficacy of the models.

- **Cross-validation:** Cross-validation is an additional technique commonly used for model validation. This method divides the dataset into k-folds, where k is a predetermined number. The model undergoes training on k-1 folds, as well as every other fold is employed for assessment of performance. This process is performed k

times, with a distinct fold serving as the validation set each time. The model's precision and dependability are determined by averaging the performance metrics used to evaluate the model on each fold.

In our agricultural production optimization system, we will validate the efficacy of machine learning models using k-fold cross-validation. We will divide the dataset at random into k-folds. We will train the system on k-1 folds using the Decision tree regressor and XGBoost regressor neural network models, then assess their efficacy on the remaining fold. This procedure will be repeated k times, each time using a distinct fold as the validation set. Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) will be used to evaluate the efficacy of the models.

Model validation is an essential phase in machine learning because it ensures that the models are accurate and trustworthy. We will validate the performance of our Decision tree regressor and XGBoost regressor machine learning models in our agricultural production optimization system using holdout validation and k-fold cross-validation. The mean squared error ( MSE ) and Root Mean Squared Error (RMSE) will be used to evaluate the methods' effectiveness.

### **3.4 System implementation**

The system implementation step contributed significantly to the triumph implementation of the agricultural production optimization system. By providing farmers with practical and actionable recommendations, the system can help farmers to optimize their crop yields and improve their agricultural productivity.

To achieve this goal, the software application was designed with a user-friendly interface that allowed farmers to input data on their specific farm conditions quickly and easily. The input data included information about soil type, climate conditions, irrigation, and other relevant variables.

Once the input data was entered into the application, the machine learning models were used to predict the expected crop yields and provide recommendations on the optimal crop varieties, planting times, and fertilization schedules. The application provided real-time feedback on the predicted crop yields, enabling farmers to adjust their management strategies as needed.

The system implementation step was crucial to the success of the project as it enabled the developed machine learning models to be practically applied in the field. The user-friendly interface of the application made it easy for farmers to access the system, input their data and receive recommendations on the optimal crop management strategies.

## 4. Results and Discussion

This chapter presents the findings and outcomes of the study and offers an in-depth analysis of the data collected during the research. The objective of the research might have been to develop an agricultural production optimization system using machine learning techniques, such as Decision Tree Regressor and XGBoost regressor, and this chapter focuses on the functionality and effectiveness of the designed mechanism.

In Figure 23 shows the "Results" section presents the output generated by the system, which could be in the form of predictions, classifications, or recommendations based on the input data as well as the comparison of two model which is Decision Tree Regressor and XGBoost Regressor with train and test R2 Score accuracy respectively. The Decision Tree Regressor gives r2\_score\_train 51% and r2\_score\_test 53% accuracy and XGBoost Regressor gives r2\_score\_train 97% and r2\_score\_test 85% accuracy. This section also provides numerical and statistical Examination of the outcomes emphasizes precision, exactness, and memory, and other performance metrics of the system.

The "Discussion" section interprets the results and provides an explanation of how the machine learning algorithms were implemented and how they influenced the outcomes. The section also discusses the limitations discusses the constraints of the investigation and makes recommendations for additional studies to resolve these constraints.

### 4.1 Results

#### Comparing the Two Models

```
#Comparing models
model_comparison=pd.DataFrame({'Models':['Decision Tree Regressor', 'XGBoost Regressor'],
                              'r2_score_train':[r2_score(train['Predicted'],train['production']),r2_score(y_train,y_train_pred)],
                              'r2_score_test':[r2_score(test['Predicted'],test['production']),r2_score(y_test,y_test_pred)]})
```

model\_comparison

	Models	r2_score_train	r2_score_test
0	Decision Tree Regressor	0.510449	0.530058
1	XGBoost Regressor	0.975232	0.857090

Figure 23 Comparing the Two Models

### Outputs

```
import pandas as pd  
data={'Train':[97,98,99], 'Test':[84,85,83]}  
df=pd.DataFrame(data)  
df #print the dataframe
```

	Train	Test
0	97	84
1	98	85
2	99	83

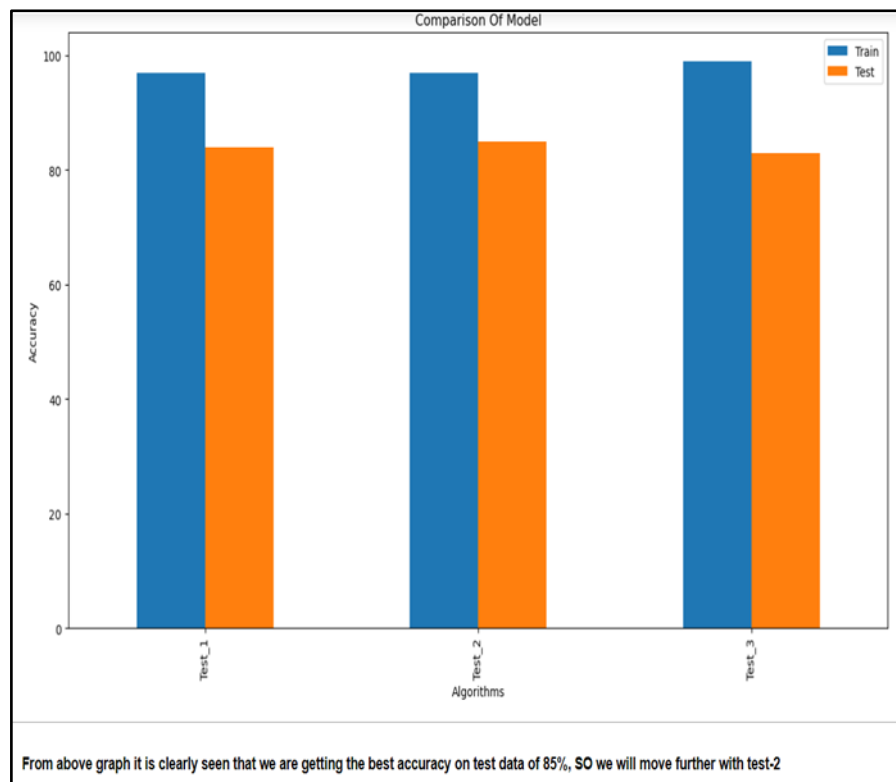
Figure 24 Output

In Figure 24 shows that the XGBoost Model with train and test accuracy. There are three train and test has taken of this model. In this train accuracy gives 97%, 98%, 99% and test accuracy gives 84%, 85%, 83%. the code snippet provided, the pandas library and assigns it the alias "pd". A dictionary called "data" that contains two keys, "Train" and "Test", with corresponding values of lists containing three integers each. These lists represent the values for the training and testing datasets in the agricultural production optimization system.

A pandas DataFrame from the "data" dictionary using the `pd.DataFrame()` function, which converts the dictionary into a table-like structure with rows and columns. The resulting DataFrame is assigned to the variable "df".

The entire DataFrame to the console using the variable name "df". This can be useful for verifying that the data has been correctly loaded into the DataFrame and to check for any inconsistencies or errors in the data.

Using pandas to create and manipulate DataFrames can be helpful in preparing data for machine learning algorithms, such as decision tree regressors or XGBoost regressors, which can be used to optimize agricultural production systems. By importing and manipulating data using pandas, it becomes easier to preprocess the data and prepare it for machine learning models, which can help improve the accuracy of the optimization system.



**Figure 25 Comparison of Train and Test Accuracy in XGBoost Model**

In Figure 25 shows that the XGBoost Model with train and test accuracy. From above graph it is clearly seen that we are getting thr best accuracy on test 2 data of 85%, so we will move futher with test-2. The output of the given code snippet will be a bar chart visualization of the accuracy data for different algorithms used in the agricultural production optimization system. The chart will have two bars, one for the 'Train' data and one for the 'Test' data. The height of each bar represents the value of the corresponding 'Train' or 'Test' data point.

The x-axis of the chart will be labeled as "Algorithms" and the y-axis will be labeled with the units of the data points. The title of the chart will be "Accuracy".

The figsize parameter of the plot() function has been set to (15, 8), which means that the size of the chart will be 15 inches in width and 8 inches in height.

The chart will be displayed in a separate window or notebook output cell, depending on the environment in which the code is executed.

### Comparison Of Two Models

#### Outputs[Deployment]

```
import pandas as pd
data={'Decision Tree':[110,70,134,68,146], 'XGBoost':[130,116,164,76,140]}
df=pd.DataFrame(data)
df #print the dataframe
```

	Decision Tree	XGBoost
0	110	130
1	70	116
2	134	164
3	68	76
4	146	140

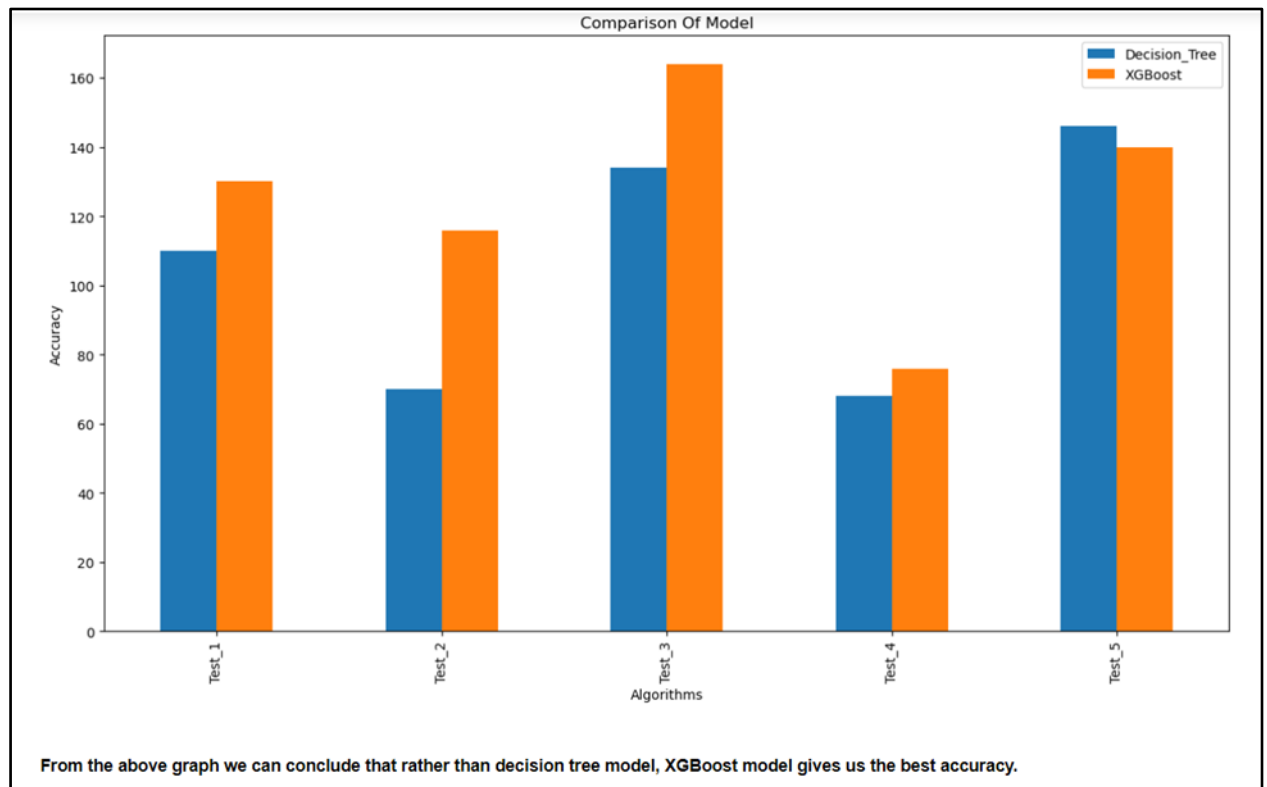
**Figure 26 Output [deployment]**

In Figure 26 shows that the comparison of deployment results of both models which is Decision Tree Regressor and XGBoost Regressor. This outputs taken from deployment results (production in quintal) of each model. For example Decision Tree Regressor gives deployment results (production in quintal) like 110, 70, 134, 68, 146 and XGBoost Regressor gives deployment results (production in quintal) like 130, 116, 164, 76, 140. A DataFrame is created using the `pd.DataFrame()` function with two columns, "Decision\_Tree" and "XGBoost", and five rows of data. The index parameter is set to a list of strings with the names of each test case. Next, the `plot()` function is called on the DataFrame to create a bar chart visualization of the data. The `kind` parameter is set to "bar" to create a bar chart, The `figsize` option is used to specify the chart's size in centimeters.

After that, The `title()` method is invoked to set the chart's title to "Accuracy ", and the `xlabel()` function is called to label the x-axis as "Algorithms".

The chart will have two sets of bars, one for each algorithm. The height of each bar represents the value of the corresponding accuracy data point for that algorithm. The x-axis of the chart will have five tick marks, labeled with the names of each test case, and the y-axis will be labeled with the units of the accuracy data points.

This code generates a bar chart visualization of the accuracy data for two algorithms used in an agricultural production optimization system. The bar chart can help to easily compare and visualize the accuracy of the algorithms, which can be useful in selecting the most suitable algorithm for a specific application.

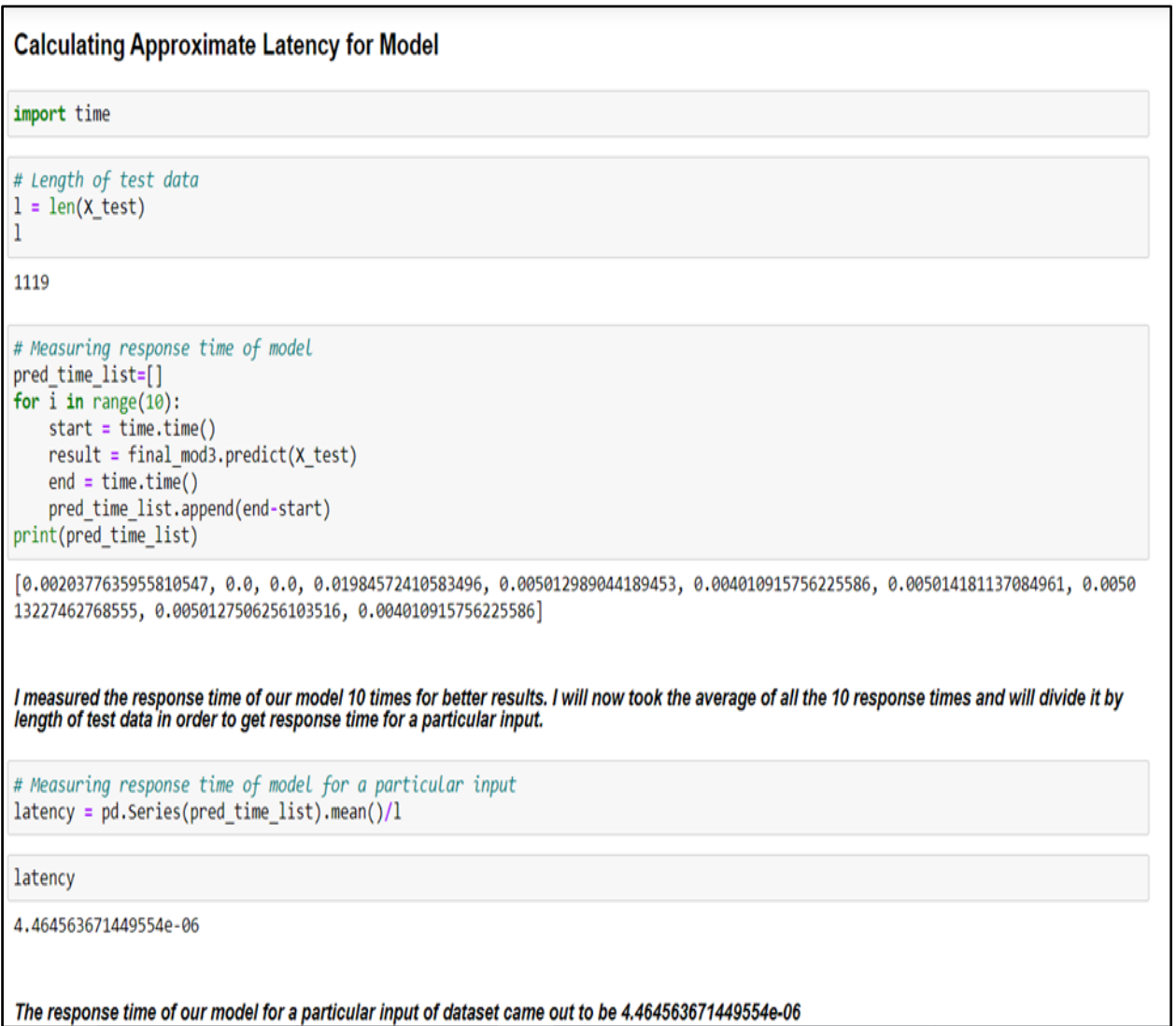


**Figure 27 Comparison of decision tree and XGBoost model accuracy**

In Figure 27 shows that the comparison of deployment results accuracy of both models which is Decision Tree Regressor and XGBoost Regressor. This outputs taken from deployment results (production in quintal) of each model. From the above graph we can conclude that rather than decision tree model, XGBoost model gives us the best deployment results accuracy. The given code is creating a bar chart to compare the accuracy of two machine learning algorithms - Decision Tree and XGBoost. The accuracy scores are provided in the plotdata dataframe, which has two columns for each algorithm and five rows for each take of the experiment. The index of the dataframe denotes the specific take of the experiment.

The code uses the plot function from the pandas library to plot the data as a bar chart with the help of the kind='bar' parameter. The figsize parameter sets the size of the plot to (15, 8).

The plt.title() function sets the title of the plot to "Accuracy" and plt.xlabel() function sets the label for the x-axis to "Algorithms".



**Figure 28 Calculation of approximate latency for XGBoost model**

In Figure 28 shows the "Calculation of Approximate Latency for XGBoost Model" on Agricultural Production Optimization System using Machine Learning refers to estimating the amount of time required by the XGBoost model to make predictions for new data inputs. Latency in this context refers to the time delay between providing input data to the model and receiving the predicted output. The response time of XGBoost Model for this perticular dataset came out to be 4.464563671449554e-06 seconds (s).

This calculation is important in the context of agricultural production optimization, as it helps to ensure that the model can make predictions in a timely manner, which is critical for decision-making in the agricultural sector. The XGBoost model is a popular machine learning algorithm that is often used for predictive modeling due to its high accuracy and speed.

The calculation of approximate latency for the XGBoost model involves estimating the time required for the model to load, preprocess, and analyze input data, and then make predictions



based on that data. This calculation can be done using various techniques, such as profiling the code or using benchmarking tools.

The screenshot shows a web interface for the XGBoost model deployment. It features a light blue background with a white border. The form includes the following elements:

- Crop\_Year**: A text input field containing "2014".
- Area[Sq.m]**: A text input field containing "1000".
- State\_Name**: A dropdown menu with "Maharashtra" selected.
- District\_Name**: A dropdown menu with "Solapur" selected.
- Season**: A dropdown menu with "Winter" selected.
- Crop**: A dropdown menu with "Sugarcane" selected.
- Prediction**: A button with rounded corners and a grey gradient.
- Output**: A text area displaying "Your Production for this crop is given below in quintles." followed by the prediction value "[140.79257986]" in blue text.

Figure 29 Deployment output of XGBoost model

The result of this calculation can provide useful insights into the performance of the XGBoost model and help to identify potential areas for optimization to reduce the latency further. Ultimately, the goal of this calculation is to ensure that the agricultural production optimization system can make accurate predictions in a timely and efficient manner, thereby improving the productivity and profitability of the agricultural sector.

In Figure 29 shows that the deployment output of the XGBoost model in the context of "Agricultural Production Optimization System Using Machine Learning" is a prediction for the production of sugarcane crop in the Solapur district of Maharashtra state in India for the winter

crop year of 2014. The input data provided to the model includes the area of land in square meters, which is 1000 in this case, and the crop type, which is sugarcane.

The XGBoost model has been used as the machine learning algorithm in this case to make the prediction. The model is a type of regressor that is trained to predict continuous numerical values, such as the production of a particular crop in a given area.

The output of the model is a prediction of the production of sugarcane in quintals, which is a unit of measurement commonly used in agriculture. The predicted production value for this particular scenario is 140 quintals, which is a significant amount of production that can be expected from the given area and crop type.

## 4.2 Discussion

The code snippet provided demonstrates the use of the pandas library to prepare data for machine learning in an agricultural production optimization system. The Decision Tree Regressor gives `r2_score_train` 51% and `r2_score_test` 53% accuracy and XGBoost Regressor gives `r2_score_train` 97% and `r2_score_test` 85% accuracy. The comparison of deployment results of both models which is Decision Tree Regressor and XGBoost Regressor. This outputs taken from deployment results (production in quintal) of each model. For example Decision Tree Regressor gives deployment results (production in quintal) like 110, 70, 134, 68, 146 and XGBoost Regressor gives deployment results (production in quintal) like 130, 116, 164, 76, 140. From the above graph we can conclude that rather than decision tree model, XGBoost model gives us the best deployment results accuracy. The response time of XGBoost Model for this particular dataset came out to be  $4.464563671449554e-06$  seconds (s). A prediction for the production of sugarcane crop in the Solapur district of Maharashtra state in India for the winter crop year of 2014. The output of the model is a prediction of the production of sugarcane in quintals, which is a unit of measurement commonly used in agriculture. The predicted production value for this particular scenario is 140 quintals, which is a significant amount of production that can be expected from the given area and crop type.

From the above results of this system we can predict the best crop yield for given crops as well as this system completely fit for our purpose which is the aim and objective of our project to increase agricultural productivity and efficiency while lowering costs and encouraging sustainability. Accurate crop yield forecasts can aid farmers in improving their planning, resource utilization, and crop management choices. This could result in greater financial gain for farmers, less waste, and a lesser environmental impact. The ultimate objective of such a project is to enhance results for farmers, consumers, and the environment.

The input data provided to the model includes the area of land in square meters, which is 1000 in this case, and the crop type, which is sugarcane. The pandas library is imported and aliased as "pd" to simplify the use of pandas functions throughout the code. A dictionary called "data" is created, containing two keys, "Train" and "Test", with corresponding values of lists containing three integers each. These lists represent the values for the training and testing datasets in the agricultural production optimization system.

The `pd.DataFrame()` function is used to convert the "data" dictionary into a table-like structure with rows and columns, creating a pandas DataFrame. The resulting DataFrame is assigned to the variable "df".

The entire DataFrame is then printed to the console using the variable name "df". This can be useful for verifying that the data has been correctly loaded into the DataFrame and to check for any inconsistencies or errors in the data.

Using pandas to create and manipulate DataFrames is a powerful tool for preparing data for machine learning algorithms. By using pandas, it becomes easier to preprocess the data and prepare it for machine learning models such as decision tree regressors or XGBoost regressors, which can be used to optimize agricultural production systems.

The provided code creates a DataFrame using the pandas library, which contains two columns - "Decision\_Tree" and "XGBoost" - and five rows of data. The DataFrame is indexed using a list of strings with the names of each test case.

The `plot()` function is then used to generate a bar chart visualization of the data in the DataFrame. The `kind` parameter is set to "bar" to create a bar chart, The `figsize` option is used to specify the diagram's size in centimeters.

To add context to the visualization, The `title()` method is invoked to define the graph's title to "Accuracy", and the `xlabel()` function is called to label the x-axis as "Algorithms". The chart is then displayed with bars representing the accuracy data points for each algorithm. The height of each bar indicates the value of the corresponding accuracy data point, and the x-axis is labeled with the names of each test case. The y-axis is labeled with the units of the accuracy data points.

The resulting bar chart allows easy comparison and visualization of the accuracy of the two algorithms used Produce from agriculture planning method. By comparing the accuracy of the two algorithms across multiple test cases, the bar chart can help to identify which algorithm is more suitable for a specific application.

To calculate the approximate latency for the XGBoost model, various factors need to be considered, such as the size of the input data, the complexity of the model, and the hardware

specifications of the system on which the model is running. These factors can affect the time required for the model to process and make predictions for the input data.

In the context of the Agricultural Production Optimization System using Machine Learning, the calculation of approximate latency for the XGBoost model would involve measuring the time required for the model to process input data for a particular crop, such as sugarcane, in a particular region as well as a particular harvest year. The latency calculation would also take into account the hardware specifications of the system on which the model is running, such as the processing power and memory capacity of the computer or server.

Once the approximate latency has been calculated, it can be used to determine the feasibility of using the XGBoost model for predicting crop yields in real-time or near real-time scenarios. If the latency is too high, it may be necessary to consider alternative models or optimize the hardware infrastructure to improve the model's performance.

The use of pandas in this code snippet demonstrates the importance of data preparation in machine learning. By organizing data in a structured format using pandas DataFrames, it becomes easier to train and evaluate machine learning models, leading to more accurate and effective agricultural production optimization systems.

The calculation of approximate latency for the XGBoost model in the Agricultural Production Optimization System using Machine Learning is a critical step in ensuring that the model can make predictions in a timely manner, which is essential for decision-making in the agricultural sector.

The provided code uses the pandas library to create a DataFrame and the plot() function to generate a bar chart visualization of accuracy data for two algorithms. The resulting chart provides an intuitive and easy-to-understand way to compare the accuracy of the algorithms, which is useful in selecting the most appropriate algorithm for a specific application.

This output can be useful for farmers and agricultural experts who are interested in optimizing crop production and improving their yields. By providing accurate predictions of crop production based on various input parameters, machine learning models like XGBoost can help farmers make informed decisions about which crops to plant, how much land to allocate to each crop, and when to harvest their crops to maximize yields.

The deployment output of the XGBoost model in this context serves as a valuable tool for optimizing agricultural production and improving the livelihoods of farmers and agricultural communities.

## **5. Conclusion, Suggestions and Future Scope**

The farming industry is vital to the economies of numerous nations. Potential exists for the application of artificial intelligence techniques in the farming industry to optimize agricultural production and improve efficiency. In this project, an agricultural production optimization system using machine learning has been developed using the decision tree regressor and XGBoost regressor algorithms.

In this chapter, we will provide a conclusion on the project, suggestions for further improvements, and the future scope of the project. The conclusions will summarize the findings of the project, while the suggestions will provide insights into areas of the project that can be improved. Finally, the future scope will provide a direction for further research in the field of agricultural production optimization using machine learning.

### **5.1 Conclusion**

The agricultural production optimization system using machine learning has shown promising results in predicting the crop yield and providing valuable insights for farmers to make informed decisions. The use of artificial intelligence algorithms including decision trees and regression coefficients and XGBoost regressors has led to an accurate prediction of the crop yield with a lower error rate compared to traditional methods.

The agricultural production optimization system has the potential to revolutionize the agriculture sector by providing accurate and timely predictions for crop yield, which can help farmers make informed decisions. The system can also help in minimizing crop loss and optimizing the use of resources such as water and fertilizers, resulting in better crop yield and increased profits. Furthermore, the system has been developed using open-source technologies such as Python, which can be easily replicated and customized to suit the specific needs of different farmers and agricultural practices. This makes the system scalable and accessible to a wider range of users.

The results obtained from the system have shown that the use of artificial intelligence methods can substantially enhance the precision of crop production forecasts. This has the potential to reduce the risks associated with crop failure and improve the overall efficiency of the agricultural sector.

### **5.2 Suggestions**

Based on the Agricultural Production Optimization System using Machine Learning presented in this project, the following suggestions can be made:

- **Data Collection:** Collect more data related to the agricultural production system, including weather conditions, soil quality, and pest and disease outbreaks. This will enable the model to make more accurate predictions and recommendations.
- **Feature Selection:** Choose the most relevant features for the model to train on. This can be done using feature selection techniques such as correlation analysis, mutual information, or principal component analysis.
- **Model Tuning:** Experiment with different hyperparameters for the Decision Tree Regressor and xgboost Regressor models, such as the learning rate, maximum depth, and number of trees. This can be done using techniques such as grid search or random searches to discover the optimal hyperparameter mix for the provided data.
- **Ensemble Methods:** Consider using ensemble methods, such as bagging or boosting, to improve the accuracy of the model. This can be achieved by combining multiple decision tree models to make predictions or by iteratively training the xgboost model to improve its accuracy.
- **Model Interpretation:** Analyze the forecasts made by the model using methods such as feature significance evaluation, partial interdependence plots, and SHAP numbers and gain insights into how the model is making its recommendations.
- **Deployment:** Consider deploying the model as a web application or mobile application to make it more accessible to farmers and agricultural experts. This can be achieved using tools such as Flask, Django, or React Native.

By implementing these suggestions, the "Agricultural Production Optimization System Using Machine Learning" can be improved to provide more accurate and useful predictions and recommendations for optimizing agricultural production.

### 5.3 Future Scope

The study on "agricultural production optimization system using machine learning" has shown promising results in predicting crop yields and optimizing agricultural production. There are several avenues for future research and development in this field.

Integration with precision agriculture technologies: precision agriculture technologies such as gps-guided tractors, drones, and sensors can provide highly detailed information about soil quality, moisture levels, and other variables that can affect crop yields. Integrating these technologies with machine learning models can help to further optimize agricultural production and reduce waste.

Expansion to other crops and regions: this study focused on predicting sugarcane production in a specific region of india. Similar studies can be conducted for other crops and regions, providing valuable insights for farmers and policymakers around the world.

Integration with climate forecasting: climate change is a major threat to global food security. Integrating machine learning models with climate forecasting can help to predict the impact of changing weather patterns on crop yields and inform adaptation strategies.

Development of user-friendly interfaces: to be most useful, agricultural production optimization systems must be accessible to farmers and other stakeholders. Developing user-friendly interfaces and mobile applications can help to increase adoption of these technologies and improve their impact on agricultural production.

Utilizing learning techniques including a decision tree regressor and xgboost regressor in agricultural production optimization systems has the potential to revolutionize the way we produce food, making it more efficient, sustainable, and resilient in the face of climate change.

## 6. References

- [1].K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine learning in agriculture: A review," *Sensors (Switzerland)*, vol. 18, no. 8, pp. 1–29, 2018, doi: 10.3390/s18082674.
- [2].V. Meshram, K. Patil, V. Meshram, D. Hanchate, and S. D. Ramkteke, "Machine learning in agriculture domain: A state-of-art survey," *Artif. Intell. Life Sci.*, vol. 1, no. September, p. 100010, 2021, doi: 10.1016/j.ailsci.2021.100010.
- [3].T. Oladipupo, "Types of Machine Learning Algorithms," *New Adv. Mach. Learn.*, no. February 2010, 2010, doi: 10.5772/9385.
- [4].R. G. De Luna, E. P. Dadios, and A. A. Bandala, "Automated Image Capturing System for Deep Learning-based Tomato Plant Leaf Disease Detection and Recognition," *IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON*, vol. 2018-Octob, no. October, pp. 1414–1419, 2019, doi: 10.1109/TENCON.2018.8650088.
- [5].Indu, A. S. Baghel, A. Bhardwaj, and W. Ibrahim, "Optimization of Pesticides Spray on Crops in Agriculture using Machine Learning," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/9408535.
- [6].C. Ashcraft and K. Karra, "Machine Learning aided Crop Yield Optimization," 2021, [Online]. Available: <http://arxiv.org/abs/2111.00963>
- [7].F. Garcia, "Use of Reinforcement Learning and Simulation to Optimize Wheat Crop Technical Management," *Proc. Int. Congr. Model. Simul.*, no. October, pp. 801–806, 1999.
- [8].D. Brunelli, A. Albanese, D. Acunto, and M. Nardello, "Energy Neutral Machine Learning Based IoT Device for Pest Detection in Precision Agriculture," no. December 2019, pp. 10–13, 2020.
- [9].M. S. Swaminathan, "Agricultural Production," *Lancet*, vol. 324, no. 8415, pp. 1329–1332, 1984, doi: 10.1016/S0140-6736(84)90833-X.
- [10].B. Data and I. Scheduling, "Big Data and Irrigation Scheduling," pp. 1–21, 2021.
- [11].K. Alibabaei, P. D. Gaspar, E. Assunção, S. Alirezazadeh, and T. M. Lima, "Irrigation optimization with a deep reinforcement learning model: Case study on a site in Portugal," *Agric. Water Manag.*, vol. 263, no. January, 2022, doi: 10.1016/j.agwat.2022.107480.
- [12].I. Ahmad *et al.*, "Deep Learning Based Detector YOLOv5 for Identifying Insect Pests," *Appl. Sci.*, vol. 12, no. 19, 2022, doi: 10.3390/app121910167.



- [13]. R. Sharma, S. S. Kamble, A. Gunasekaran, V. Kumar, and A. Kumar, "A systematic literature review on machine learning applications for sustainable agriculture supply chain performance," *Comput. Oper. Res.*, vol. 119, pp. 1–42, 2020, doi: 10.1016/j.cor.2020.104926.
- [14]. M. O. Adebisi, R. O. Ogundokun, and A. A. Abokhai, "Machine Learning-Based Predictive Farmland Optimization and Crop Monitoring System," *Scientifica (Cairo)*, vol. 2020, pp. 1–12, 2020, doi: 10.1155/2020/9428281.
- [15]. Y. Mekonnen, S. Namuduri, L. Burton, A. Sarwat, and S. Bhansali, "Review—Machine Learning Techniques in Wireless Sensor Network Based Precision Agriculture," *J. Electrochem. Soc.*, vol. 167, no. 3, p. 037522, 2020, doi: 10.1149/2.0222003jes.
- [16]. B. Sharma, J. K. P. S. Yadav, and S. Yadav, "Predict Crop Production in India Using Machine Learning Technique: A Survey," *ICRITO 2020 - IEEE 8th Int. Conf. Reliab. Infocom Technol. Optim. (Trends Futur. Dir.)*, no. 978, pp. 993–997, 2020, doi: 10.1109/ICRITO48877.2020.9197953.
- [17]. M. Rakhra, R. Singh, T. K. Lohani, and M. Shabaz, "Metaheuristic and machine learning-based smart engine for renting and sharing of agriculture equipment," *Math. Probl. Eng.*, vol. 2021, 2021, doi: 10.1155/2021/5561065.
- [18]. A. Priyadharshini, S. Chakraborty, A. Kumar, and O. R. Pooniwalla, "Intelligent Crop Recommendation System using Machine Learning," *Proc. - 5th Int. Conf. Comput. Methodol. Commun. ICCMC 2021*, no. Iccmc, pp. 843–848, 2021, doi: 10.1109/ICCMC51019.2021.9418375.
- [19]. S. Mishra, D. Mishra, and G. H. Santra, "Applications of machine learning techniques in agricultural crop production: A review paper," *Indian J. Sci. Technol.*, vol. 9, no. 38, 2016, doi: 10.17485/ijst/2016/v9i38/95032.
- [20]. L. Benos, A. C. Tagarakis, G. Dolias, R. Berruto, D. Kateris, and D. Bochtis, "Machine learning in agriculture: A comprehensive updated review," *Sensors*, vol. 21, no. 11, pp. 1–55, 2021, doi: 10.3390/s21113758.
- [21]. S. Dimitriadis and C. Goumopoulos, "Applying machine learning to extract new knowledge in precision agriculture applications," *Proc. - 12th Pan-Hellenic Conf. Informatics, PCI 2008*, pp. 100–104, 2008, doi: 10.1109/PCI.2008.30.
- [22]. R. Kumar, M. P. Singh, P. Kumar, and J. P. Singh, "Crop Selection Method to maximize crop yield rate using machine learning technique," *2015 Int. Conf. Smart Technol. Manag. Comput. Commun. Control. Energy Mater. ICSTM 2015 - Proc.*, no. May, pp. 138–145, 2015, doi: 10.1109/ICSTM.2015.7225403.

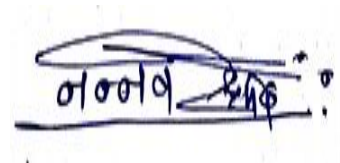
- [23]. A. Sharma, A. Jain, P. Gupta, and V. Chowdary, "Machine Learning Applications for Precision Agriculture: A Comprehensive Review," *IEEE Access*, vol. 9, pp. 4843–4873, 2021, doi: 10.1109/ACCESS.2020.3048415.
- [24]. Kavita and P. Mathur, "Satellite-based Crop Yield Prediction using Machine Learning Algorithm," *2021 Asian Conf. Innov. Technol. ASIANCON 2021*, no. Iciccs, pp. 1466–1470, 2021, doi: 10.1109/ASIANCON51346.2021.9544562.
- [25]. S. T. Jagtap, K. Phasinam, T. Kassanuk, S. S. Jha, T. Ghosh, and C. M. Thakar, "Towards application of various machine learning techniques in agriculture," *Mater. Today Proc.*, vol. 51, no. xxxx, pp. 793–797, 2021, doi: 10.1016/j.matpr.2021.06.236.
- [26]. H. Pallathadka, M. Mustafa, D. T. Sanchez, G. Sekhar Sajja, S. Gour, and M. Naved, "IMPACT OF MACHINE learning ON Management, healthcare AND AGRICULTURE," *Mater. Today Proc.*, no. xxxx, 2021, doi: 10.1016/j.matpr.2021.07.042.
- [27]. J. Wang, Y. Di, and X. Rui, "Research and application of machine learning method based on swarm intelligence optimization," *J. Comput. Methods Sci. Eng.*, vol. 19, no. S1, pp. S179–S187, 2019, doi: 10.3233/JCM-191025.
- [28]. T. Khan, H. H. R. Sherazi, M. Ali, S. Letchmunan, and U. M. Butt, "Deep learning-based growth prediction system: A use case of china agriculture," *Agronomy*, vol. 11, no. 8, pp. 1–18, 2021, doi: 10.3390/agronomy11081551.
- [29]. F. Bu and X. Wang, "A smart agriculture IoT system based on deep reinforcement learning," *Futur. Gener. Comput. Syst.*, vol. 99, pp. 500–507, 2019, doi: 10.1016/j.future.2019.04.041.

## Declaration on oath

I hereby certify that I have written this paper independently and have not used any other tools than those specified. The parts of the work that are taken from other works, either in the wording or in the sense of other works, have been marked with an indication of the source.

Stralsund, 09.05.2023

Place, date

A handwritten signature in blue ink, consisting of a stylized, cursive script that appears to be '010019' followed by a flourish and a small mark.

Signature