Industrial Internship Report on

”Prediction of Agriculture Crop Production in India: Crop Prediction”

Prepared by

[Adeline Christabel]

Executive Summary
This report provides details of the Industrial Internship provided by upskill Campus and The IoT Academy in collaboration with Industrial Partner UniConverge Technologies Pvt Ltd (UCT). This internship was focused on a project/problem statement provided by UCT. We had to finish the project including the report in 6 weeks’ time. My project, titled **"Prediction of Agriculture Crop Production in India,"** involved developing a machine learning model to predict the most suitable crop to grow based on critical agricultural attributes such as nitrogen, phosphorus, potassium levels, temperature, humidity, pH, and rainfall. The project aimed to address the real-world challenge of optimizing crop selection for improved productivity and resource utilization in agriculture. This internship gave me a very good opportunity to get exposure to Industrial problems and design/implement solution for that. It was an overall great experience to have this internship.

TABLE OF CONTENTS

1 Preface

The six-week industrial internship provided by Upskill Campus (USC) and The IoT Academy, in collaboration with Uniconverge Technologies Pvt. Ltd. (UCT), has been an invaluable learning journey. This internship was focused on solving a real-world problem posed by UCT, requiring not only technical implementation but also comprehensive reporting within a tight six-week timeline. The experience gave me an opportunity to immerse myself in industrial challenges, bridging the gap between academic knowledge and practical applications.

Internships like these are critical for career development, as they provide hands-on exposure to real-world problems and industrial workflows. They help in applying theoretical knowledge to practical scenarios, improving technical skills, and fostering problem-solving abilities. Furthermore, internships offer a platform to work collaboratively, manage time effectively, and develop professional communication skills, all of which are vital for career growth.

My project, **"Prediction of Agriculture Crop Production in India,"** aimed to address one of the most significant challenges in modern agriculture: optimizing crop selection. I developed a machine learning model that predicts the most suitable crop to grow based on attributes like nitrogen, phosphorus, potassium, temperature, humidity, pH, and rainfall. The project involved preprocessing a dataset, evaluating multiple machine learning models (e.g., Random Forest, Gradient Boosting, and Support Vector Machines), and deploying the best-performing model using Streamlit for real-time predictions.

Opportunity Provided by USC and UCT:
The internship offered by USC and UCT was a well-structured program that enabled participants to solve industrial-grade problems while enhancing their technical and professional capabilities. The opportunity to work on a problem statement from UCT provided a real-world context to our work, making the learning process both engaging and impactful.
The program was meticulously planned, with clear milestones set for each week. The entire process was divided into six clear and well-defined stages.

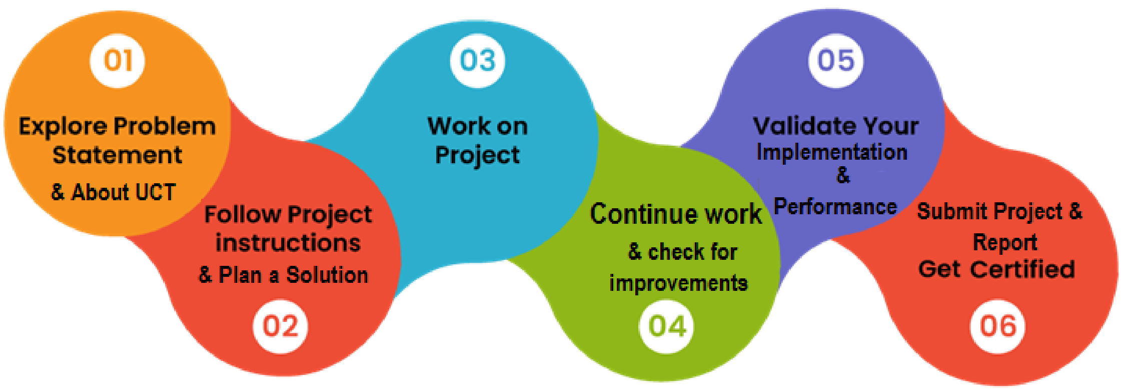
Week 1: Explore Problem Statement & About UCT
The program began with a thorough exploration of the problem statement provided by Uniconverge Technologies Pvt. Ltd. (UCT). This stage involved understanding the core objectives, challenges, and industrial relevance of the project.

Week 2: Follow Project Instructions & Plan a Solution
In this phase, we followed the project guidelines and crafted a structured plan to address the problem. This involved brainstorming potential approaches, identifying resources, and defining milestones for the next stages of the project.

Week 3: Work on Project
The third stage focused on the implementation of the project. This included activities like data preprocessing, feature engineering, applying various machine learning models, and evaluating their performance based on metrics like R^2 score.

Week 4: Continue Work & Check for Improvements
After initial implementation, efforts were made to refine the solution. This stage involved iterating on the models, tuning hyperparameters, and ensuring the system's scalability and reliability.

Week 5: Validate Your Implementation & Performance
Validation of the model's implementation and performance was a critical step. Testing was performed using real-world scenarios, ensuring the model predicted accurate results under varying conditions. Additionally, the deployment process was tested through a Streamlit application.



Week 6: Submit Project & Report, Get Certified
Finally, the project, along with detailed documentation and the report, was submitted. This marked the culmination of the six-week program, leading to certification for successful completion.

My Learnings and Overall Experience:
Throughout this internship, I learned the importance of:

- End-to-end project development, from data preprocessing to application deployment.
- Evaluating and selecting the best machine learning models based on performance metrics.
- Practical testing of models using real-world scenarios to ensure reliability.
- Effective time management and collaborative problem-solving.

The overall experience was immensely rewarding, as it not only deepened my technical expertise but also improved my professional skills, such as documentation, communication, and presenting solutions to industrial problems.

Acknowledgments:
I would like to express my heartfelt gratitude to all those who contributed to my internship journey:

- **Upskill Campus (USC) and The IoT Academy** for organizing this excellent program.

- **UniConverge Technologies Pvt. Ltd. (UCT)** for providing a challenging and meaningful problem statement.
- My mentors and guides at UCT and USC, who provided invaluable feedback and support.
- My peers, who made this journey enjoyable and collaborative.

2 Introduction

2.1 About UniConverge Technologies Pvt Ltd

A company established in 2013 and working in Digital Transformation domain and providing Industrial solutions with prime focus on sustainability and RoI.

For developing its products and solutions it is leveraging various **Cutting Edge Technologies e.g. Internet of Things (IoT), Cyber Security, Cloud computing (AWS, Azure), Machine Learning, Communication Technologies (4G/5G/LoRaWAN), Java Full Stack, Python, Front end** etc.

i. UCT IoT Platform (uct Insight)

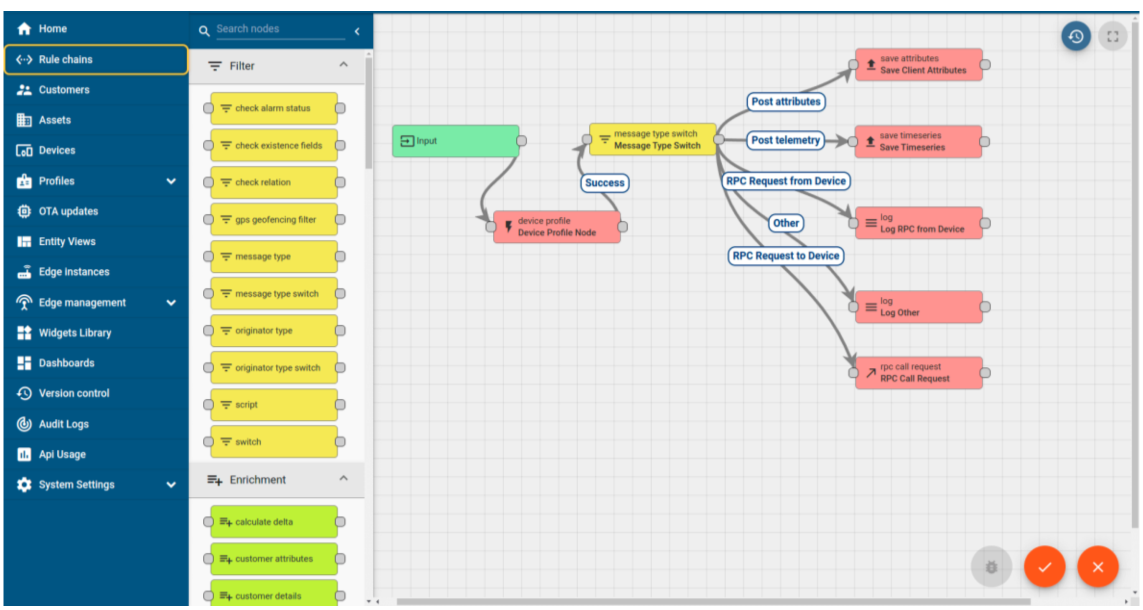
UCT Insight is an IOT platform designed for quick deployment of IOT applications on the same time providing valuable "insight" for your process/business. It has been built in Java for backend and ReactJS for Front end. It has support for MySQL and various NoSql Databases.

- It enables device connectivity via industry standard IoT protocols - MQTT, CoAP, HTTP, Modbus TCP, OPC UA

- It supports both cloud and on-premises deployments.

It has features to

- Build Your own dashboard
- Analytics and Reporting
- Alert and Notification
- Integration with third party application (Power BI, SAP, ERP)
- Rule Engine



ii. Smart Factory Platform (FACTORY WATCH)

Factory watch is a platform for smart factory needs.
It provides Users/ Factory

- with a scalable solution for their Production and asset monitoring
- OEE and predictive maintenance solution scaling up to digital twin for your assets.
- to unleased the true potential of the data that their machines are generating and helps to identify the KPIs and also improve them.
- A modular architecture that allows users to choose the service that they what to start and then can scale to more complex solutions as per their demands.

Its unique SaaS model helps users to save time, cost and money.

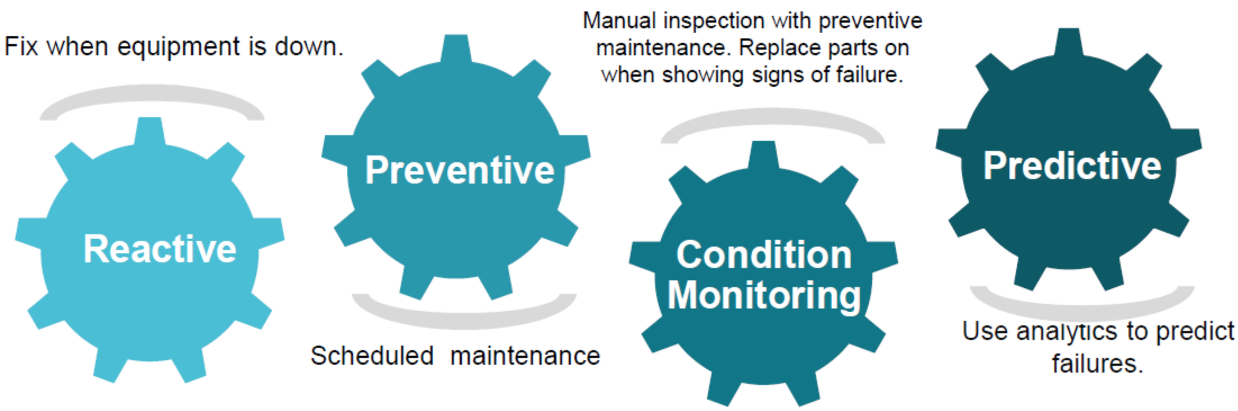
Machine	Operator	Work Order ID	Job ID	Job Performance	Job Progress		Output		Rejection	Time (mins)				Job Status	End Customer
					Start Time	End Time	Planned	Actual		Setup	Pred	Downtime	Idle		
CNC_S7_81	Operator 1	WO0405200001	4168	58%	10:30 AM		55	41	0	80	215	0	45	In Progress	i
CNC_S7_81	Operator 1	WO0405200001	4168	58%	10:30 AM		55	41	0	80	215	0	45	In Progress	i

iii. LoRaWAN based Solution

UCT is one of the early adopters of LoRAWAN technology and providing solution in Agritech, Smart cities, Industrial Monitoring, Smart Street Light, Smart Water/ Gas/ Electricity metering solutions etc.

iv. Predictive Maintenance

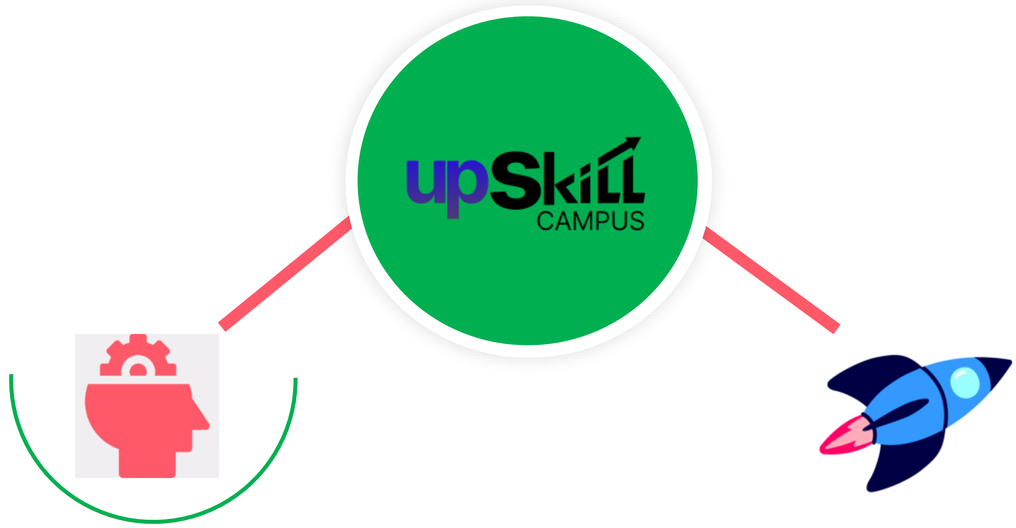
UCT is providing Industrial Machine health monitoring and Predictive maintenance solution leveraging Embedded system, Industrial IoT and Machine Learning Technologies by finding Remaining useful life time of various Machines used in production process.



2.2 About upskill Campus (USC)

upskill Campus along with The IoT Academy and in association with Uniconverge technologies has facilitated the smooth execution of the complete internship process.

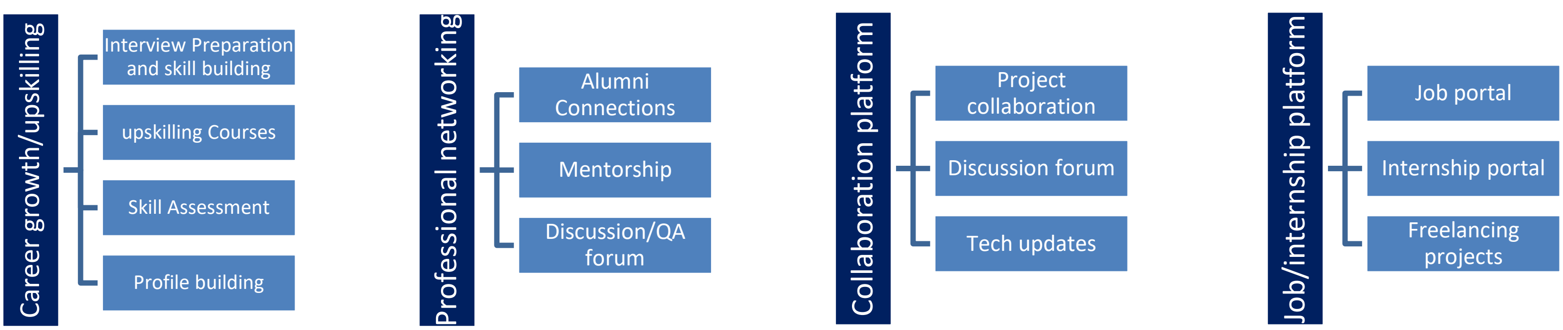
USC is a career development platform that delivers personalized executive coaching in a more affordable, scalable and measurable way.



Seeing need of upskilling in self paced manner along-with additional support services e.g. Internship, projects, interaction with Industry experts, Career growth Services

upSkill Campus aiming to upskill 1 million learners in next 5 year

https://www.upskillcampus.com/



2.3 The IoT Academy

The IoT academy is EdTech Division of UCT that is running long executive certification programs in collaboration with EICT Academy, IITK, IITR and IITG in multiple domains.

2.4 Objectives of this Internship program

The objective for this internship program was to

- get practical experience of working in the industry.
- to solve real world problems.

- ☛ to have improved job prospects.
- ☛ to have Improved understanding of our field and its applications.
- ☛ to have Personal growth like better communication and problem solving.

2.5 Reference

[1] Dataset Source: Kaggle

[2] Tools and Libraries Documentation: https://www.geeksforgeeks.org/best-python-libraries-for-machine-learning

[3] E-books provided by Upskill Campus (USC):

- *Introducing Data Science and Machine Learning*
- *Machine Learning*

2.6 Glossary

Terms	Acronym
Nitrogen	N
Phosphorus	P
Potassium	K
Machine Learning	**ML**
Mean Squared Error	MSE
R-Squared Score	R^2 Score
Support Vector Machine	SVM
Internet of Things	IoT

3 Problem Statement

Agriculture plays a vital role in the Indian economy, contributing significantly to employment and food production. However, farmers often face challenges in selecting the right crops to grow, as it depends on numerous factors such as climatic conditions, soil health, and nutrient levels. Inaccurate decisions can lead to reduced yields, economic losses, and unsustainable farming practices.
To address this issue, this project focuses on developing a predictive model for **"Prediction of Agriculture Crop Production in India".** The model aims to assist farmers and agricultural stakeholders in identifying the most suitable crop to cultivate under specific conditions.

The model predicts the crop to be grown based on the following attributes:

1. **Temperature** – The prevailing environmental temperature.
2. **Humidity** – The amount of moisture present in the air.
3. **pH** – The acidity or alkalinity of the soil.
4. **Rainfall** – The amount of precipitation received.
5. **Nitrogen (N)** – The nitrogen content in the soil.
6. **Phosphorus (P)** – The phosphorus level in the soil.
7. **Potassium (K)** – The potassium content in the soil.

Using machine learning techniques, the dataset containing these attributes is analysed, pre-processed, and trained to predict the crop most likely to yield optimum results under given conditions.
This solution provides a data-driven approach to assist farmers in decision-making, promoting sustainable agriculture, improving crop yields, and minimizing risks associated with inappropriate crop selection.

4 Existing and Proposed solution

Various solutions have been developed to predict suitable crops for agriculture based on environmental and soil parameters. These include:

1. **Traditional Decision Support Systems (DSS):**
   - These systems rely on predefined rules and historical data.
   - **Limitations:**
     - Lack adaptability to dynamic environmental conditions.
     - Limited ability to learn from new datasets, reducing accuracy.
2. **Single Algorithm Models:**
   - Many existing implementations use a single machine learning algorithm, such as Linear Regression or Decision Trees.
   - **Limitations:**
     - Single models often fail to capture complex relationships in the data.
     - Performance is inconsistent across different datasets.
3. **Small Datasets in Use:**

- Previous approaches often relied on limited datasets, which reduce the generalizability and accuracy of predictions.
- **Limitations:**
    - Insufficient instances result in models that lack robustness.
    - Predictions may not be reliable for diverse scenarios.

➢ Proposed solution:

Overview:

The proposed solution uses a comprehensive machine learning-based approach to predict the most suitable crop to cultivate based on conditions like temperature, humidity, pH, rainfall, and nutrient levels (N, P, K). The solution compares the performance of multiple machine learning algorithms to identify the best-performing model based on its R^2 score.

Implementation Details:

1. **Algorithms Used:**
    - Linear Regression
    - Random Forest
    - Decision Tree
    - Gradient Boosting
    - Support Vector Machine (SVM)
    - K-Nearest Neighbors (KNN)
2. **Model Selection:**
    - The dataset is passed through all the algorithms.
    - The performance of each model is evaluated using the R^2 score metric.
    - The model with the highest R^2 score is selected as the best-performing model.
3. **Dataset Considerations:**
    - The dataset contains 2200 instances, which is essential for achieving better performance and generalizability.
    - Smaller datasets are avoided as they lead to inaccurate predictions and poor model training.
4. **Value Addition:**
    - Comparative Analysis: Unlike existing solutions, the proposed model evaluates multiple algorithms to ensure optimal performance.
    - Enhanced Dataset Usage: The inclusion of a reasonably large dataset (2200 instances) improves the model's ability to make accurate predictions across diverse conditions.
    - Performance-Oriented Selection: By selecting the model with the highest R^2 score, the proposed solution ensures that the predictions are reliable and robust.

Advantages of the Proposed Solution:

1. Higher accuracy and reliability due to the comparative evaluation of multiple machine learning models.
2. Ability to generalize across diverse environmental and soil conditions due to the larger dataset.
3. A systematic and data-driven approach to assist farmers in making informed decisions.

4.1 Code submission (Github link):

Link of Repository contains code,.csv file, report document, demo video:
https://github.com/AdelineChristabel/upskillcampus

GitHub link of code-
https://github.com/AdelineChristabel/upskillcampus/blob/main/CropPrediction.py

4.2 Report submission (Github link):

Report Link:

5 Proposed Design/ Model

The proposed design follows a systematic workflow to ensure a robust and accurate crop prediction model. The design consists of several stages, each contributing to building and refining the solution. Here's the detailed design flow:

1. Problem Definition:

- Objective: To predict the most suitable crop for cultivation based on environmental and soil attributes such as temperature, humidity, pH, rainfall, and nutrient levels (Nitrogen, Phosphorus, and Potassium).

2. Data Collection and Preprocessing:

➢ Steps:

- Dataset Collection: A dataset containing approximately 2200 instances is used, with attributes such as temperature, humidity, pH, rainfall, and soil nutrients.
- Handling Missing Values: Use imputation techniques to replace missing or null values.
- Normalization/Scaling: Apply techniques like Min-Max Scaling or Standardization to normalize the numerical features for algorithms that are sensitive to scale (e.g., KNN, SVM).
- Feature Selection: Identify and retain the most relevant features by analyzing correlations and importance scores.
- Data Splitting: Divide the dataset into training and testing sets, typically with an 80%-20% split.

3. Exploratory Data Analysis (EDA):

- Purpose: Understand the dataset's structure and identify trends, correlations, and outliers.

➢ Techniques Used:

- Descriptive Statistics: Summarize the central tendencies and variability of features.
- Visualization: Create correlation heatmaps, scatter plots, and histograms to analyze relationships between features.
- Outlier Detection: Use boxplots or statistical methods to identify and handle outliers.

4. Model Development:

- Algorithm Selection: Train and test multiple machine learning models to find the best fit for the problem.
- Models considered include:

1. Linear Regression
2. Random Forest
3. Decision Tree
4. Gradient Boosting
5. Support Vector Machine (SVM)
6. K-Nearest Neighbours (KNN)

➢ Steps:

- Train each model using the training dataset.
- Evaluate each model's performance using the R^2 score and other metrics.
- Select the model with the highest R^2 score for final predictions.

5. Model Optimization:

- Techniques:
- Cross-Validation: Apply k-fold cross-validation to validate model performance and reduce overfitting.

6. Prediction and Final Output:

- Process:
- Input new data points (temperature, humidity, pH, rainfall, N, P, K).
- Use the optimized model to predict the most suitable crop.
- Generate output as the predicted crop for the given conditions.

7. Model Evaluation:

- Metrics Used:
- R^2 Score: Measures the proportion of variance explained by the model.
- Mean Absolute Error (MAE): Quantifies prediction error.
- Root Mean Square Error (RMSE): Measures the standard deviation of residuals.
- Validation:
- Test the model on unseen data to assess generalization.
- Conduct performance comparison between different algorithms.

8. Application Testing with Streamlit:

After implementing the machine learning model, a Streamlit app was developed and tested to provide a user-friendly interface for making predictions. The app allows users to input environmental and soil parameters and receive real-time crop predictions.

- Steps:

1. **User Interface Design:**
   - Developed an intuitive input form for users to provide the following parameters:
     - Temperature
     - Humidity
     - pH
     - Rainfall
     - Nitrogen (N)
     - Phosphorus (P)
     - Potassium (K)
2. **Backend Integration:**
   - The trained and optimized machine learning model was integrated into the app using the Streamlit framework.
3. **Predictions** are generated in real-time based on user inputs.
4. **App Features:**
   - Input Validation: Ensures that users provide valid numerical values for each parameter.
   - Real-Time Predictions: Displays the most suitable crop based on the entered data.
5. **Testing and Evaluation:**
   - Tested the app with various sample inputs to ensure accurate predictions.

5.1 High Level Diagram

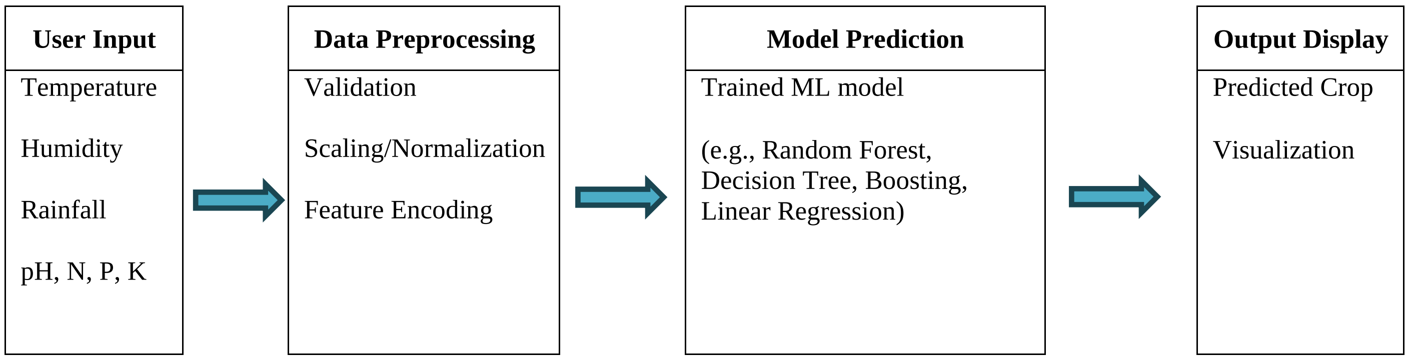
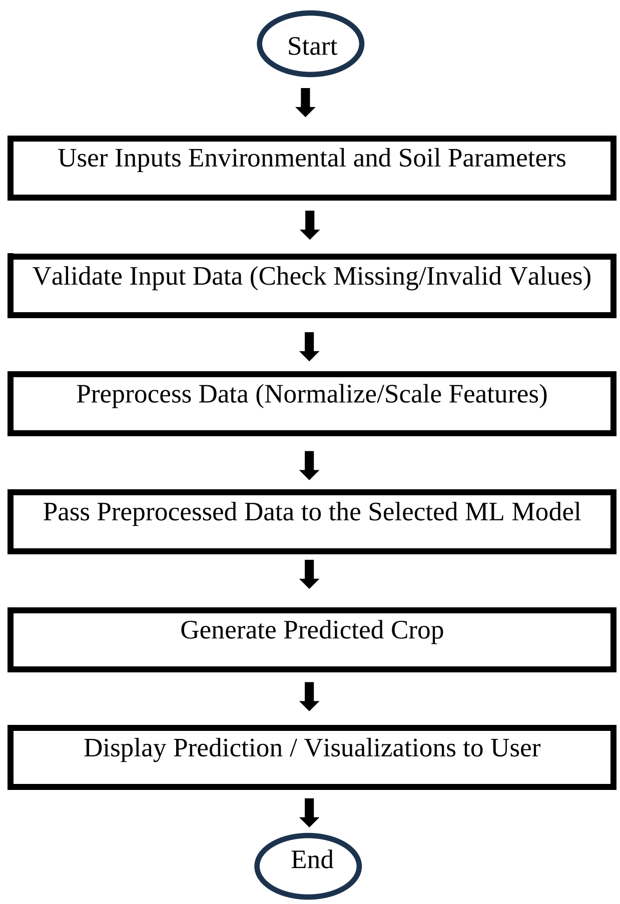


Figure: HIGH LEVEL DIAGRAM OF THE SYSTEM

5.2 Interfaces

Flow chart:



6 Performance Test

1. Identified Constraints and Their Impact

- **Memory Usage:**
- Constraint: Large datasets and complex models like Random Forest or Gradient Boosting require significant memory.
- Impact: Limited memory could lead to slower processing or system crashes.
- **Processing Speed (MIPS):**
- Constraint: Real-time predictions necessitate quick data processing and model execution.
- Impact: A slower processing speed could make the system unsuitable for large-scale or time-sensitive applications.
- **Dataset Size:**
- Constraint: The dataset used contains approximately 1800 - 2200 instances, which is relatively small for building a highly generalized machine learning model.
- Impact: Small datasets might reduce the model's accuracy and its ability to handle diverse inputs.
- **Accuracy:**
- Constraint: Achieving high accuracy is crucial for reliable crop predictions.
- Impact: Low accuracy could reduce trust and usability for end-users (e.g., farmers, agricultural planners).

2. Handling Identified Constraints in the Design

- **Memory Optimization:**
- Used optimized libraries (e.g., scikit-learn) with efficient memory management for training and prediction.
- **Processing Speed:**
- Selected algorithms with efficient runtime complexity.
- Preprocessed and scaled data before feeding it into the model, ensuring faster execution.
- **Accuracy Improvement:**
- Tested multiple machine learning algorithms (e.g., Random Forest, Gradient Boosting, SVM) and selected the one with the highest R^2 score.

3. Industrial Application Potential

- **Real-Time Predictions:** The system’s ability to provide quick and accurate predictions makes it ideal for deployment in agricultural advisory services.
- **Scalable Design:** With larger datasets and memory optimizations, the system can be scaled to serve entire regions or countries.
- **Practical Usability:** The Streamlit-based interface ensures accessibility for farmers and agricultural consultants with minimal technical expertise.

6.1 Test Plan/ Test Cases

This section describes the test plan and the specific test cases executed in your Streamlit app for crop prediction. The testing involved uploading a CSV file, entering custom attribute values, and verifying predictions against the dataset.

➢ **Objective:**
To validate the accuracy, functionality, and usability of the crop prediction system by testing predictions for custom inputs and verifying results against the dataset.

- **Scope:**
- Manual testing of attribute values entered by the user.
- Verification of predictions for selected rows from the uploaded dataset.
- Ensuring the system provides accurate and relevant results.

➢ **Testing Tools Used:**
- Application: Streamlit
- Validation Source: Uploaded CSV dataset

➢ **Testing Scenarios:**
1. Manual Input Testing
2. Row Selection and Prediction Verification

6.1.1 Test Cases

1. Manual Input Testing

Test Case	Description	Input	Expected Output	Actual Output	Status
1	Predict crop for given attribute values	N: 2.567, P: -3.880, K: 1.560, Temp: 23.0, Humidity: 1.66, pH: 5.0, Rainfall: 2.46	Recommended Crop: Rice	Recommended Crop: Rice	Pass
2	Predict crop for another set of attributes	N: -0.356, P: -0.6996, K: -0.035, Temp: 24.0, Humidity: 2.348, pH: 4.6, Rainfall: 2.108	Recommended Crop: Pomegranate	Recommended Crop: Pomegranate	Pass

2. Row Selection and Prediction Verification

Test Case	Description	Steps	Expected Output	Actual Output	Status
3	Predict crop using a selected row from dataset	1. Upload CSV file. 2. Select a row (row number). 3. Load row's values into attribute fields and submit prediction.	Matches dataset entry's crop value	Matches dataset entry's crop	Pass
4	Validate result against dataset	Select multiple rows and verify predictions against the corresponding dataset values.	Correct predictions for all rows	Correct for all rows	Pass

6.2 Test Procedure

Execution Steps:

1. Uploading the Dataset:
   - The user uploads a CSV file containing attributes such as nitrogen (N), phosphorus (P), potassium (K), temperature, humidity, pH, and rainfall.
2. Manual Input Testing:
   - Enter custom attribute values into the input fields of the Streamlit app.
   - Submit the values and verify the predicted crop against expected outcomes based on domain knowledge.
3. Row Selection Testing:
   - After uploading the dataset, select specific rows.
   - Verify that the app correctly loads the row's values into the respective fields.
   - Submit the prediction and compare the result with the corresponding entry in the dataset.

6.3 Performance Outcome

- The system accurately predicted the recommended crops for both custom attribute inputs and selected rows from the dataset.
- The ability to load row values from the dataset enhanced testing flexibility and helped validate predictions against known data.
- The manual input feature and row selection feature were user-friendly and performed as expected.

7 My learning

Working on the project "Prediction of Agriculture Crop Production in India" has been a transformative experience that has significantly enhanced my technical and professional skills. This project provided me with a deep understanding of the end-to-end data pipeline, including data preprocessing, feature scaling, and normalization, which are essential for ensuring high-quality inputs for machine learning models. I gained hands-on experience with multiple machine learning algorithms such as Linear Regression, Random Forest, Decision Tree, Gradient Boosting, Support Vector Machine, and K-Nearest Neighbors. By evaluating these models using metrics like R^2 score, I was able to select the best-performing algorithm and gain a comprehensive understanding of their strengths and limitations. Additionally, integrating the model into a user-friendly web application using Streamlit taught me how to deploy machine learning solutions effectively for real-world use cases.

The project also emphasized the importance of performance testing, where I explored metrics like response time, memory usage, and scalability to ensure the system's robustness. By handling large datasets and testing the application, I developed a practical understanding of optimizing performance. I also learned how to manage my time effectively by handling multiple phases of the project, from research and implementation to testing and deployment, within a structured timeline.

Moreover, the project helped me enhance my communication skills by documenting technical findings in a clear and structured manner. These learnings have equipped me with industry-ready skills, bridging the gap between academic knowledge and professional requirements. Understanding how to address real-world challenges in agriculture has also deepened my appreciation for the practical impact of technology and its role in solving domain-specific problems. Overall, this project has been a valuable step in my career growth, providing me with the confidence and skills needed to contribute effectively to future projects and professional environments.

8 Future work scope

While the project "Prediction of Agriculture Crop Production in India" has achieved its primary goals, there is significant potential for future enhancements that could further improve its performance and usability. Due to time constraints, several ideas could not be implemented but can serve as a foundation for future work.

One of the key areas for improvement is obtaining larger and more diverse datasets. Expanding the dataset to include additional attributes such as soil type, crop variety, and regional weather patterns would enhance the model's accuracy and generalization capabilities. Furthermore, exploring advanced machine learning models, such as deep learning architectures or ensemble techniques, could provide more robust and precise predictions.

Making the application more user-friendly is another area of focus. Enhancing the user interface with features such as multi-language support, real-time data visualization, and personalized crop recommendations based on user preferences could make the system more accessible and effective for a broader audience. Additionally, integrating IoT devices, such as sensors for real-time soil pH, moisture, and temperature data, would enable dynamic and real-time predictions, aligning with the needs of modern precision agriculture.

Incorporating sustainability metrics and resource optimization tools could further expand the application's utility. For instance, the model could recommend not only the best crop to grow but also optimal water usage, fertilizer requirements, and sustainable farming practices. Future work could also involve deploying the application on mobile platforms to reach farmers and agricultural stakeholders in remote areas, ensuring that the benefits of this system are widely accessible.

Overall, these enhancements would not only make the model more accurate and versatile but also increase its value as a practical tool for the agricultural industry. By addressing these areas in the future, the system can evolve into a comprehensive decision-support tool for sustainable and efficient farming.