# Industrial Internship Report on: ”Prediction of Agriculture Crop Production in India”

## Prepared by: Yasir Siddiqui

| *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 was to develop a machine learning model to predict agricultural crop production in India based on historical data.<br><br>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 ....4

2 Introduction ....5

2.1 About UniConverge Technologies Pvt Ltd ....5

2.2 About upskill Campus....10

2.3 Objective ....12

2.4 Reference ....12

2.5 Glossary....12

3 Problem Statement....13

4 Existing and Proposed solution ....14

5 Proposed Design/ Model ....15

5.1 High Level Diagram (if applicable) ....15

6 Performance Test ....17

6.1 Test Plan/ Test Cases ....17

6.2 Test Procedure....17

6.3 Performance Outcome....17

7 My learnings....19

8 Future work scope ....20

# 1 Preface

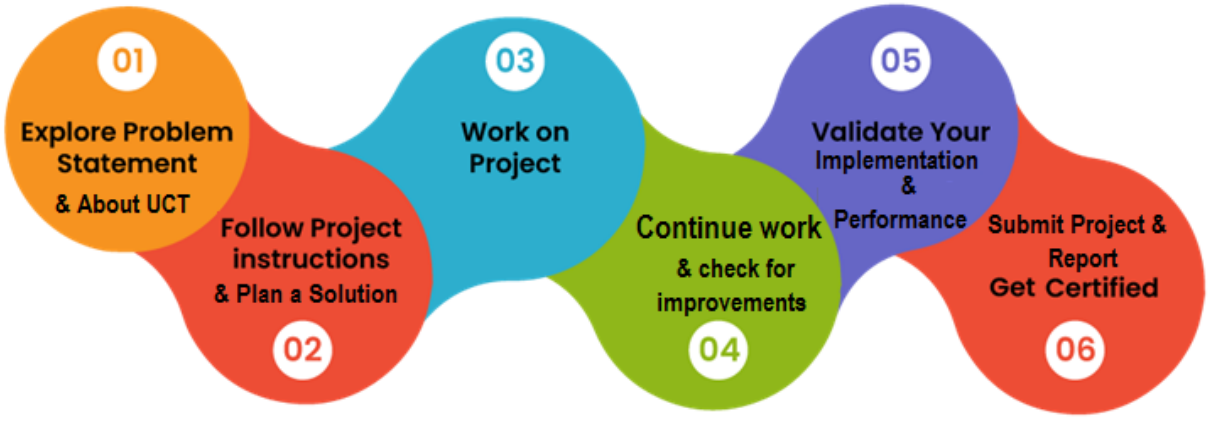
Summary of the whole 6 weeks' work.: The six-week internship was a structured program. The first week involved understanding the problem statement, followed by data exploration and preprocessing. The subsequent weeks were dedicated to model building, training, and evaluation. The final week was focused on performance testing and report submission

About need of relevant Internship in career development.: This internship provided a crucial bridge between theoretical academic knowledge and practical industry application. It allowed me to apply my data science skills to a real-world problem, which is essential for career development

Brief about Your project/problem statement.: My project involved using machine learning to predict crop production in India. By analyzing historical data, the goal was to build a model that could help farmers and policymakers make better decisions

Opportunity given by USC/UCT.: This opportunity provided by upskill Campus and UniConverge Technologies was invaluable. It gave me a platform to work on an industry-relevant project and gain mentorship.

How Program was planned: The program was well-planned with clear milestones. It followed a 6-stage process: 1. Explore Problem Statement, 2. Follow Instructions & Plan Solution, 3. Work on Project, 4. Continue Work & Check for Improvements, 5. Validate Implementation, and 6. Submit Project & Report.




Your Learnings and overall experience.: My primary learning was in data preprocessing and feature engineering for a complex dataset. The overall experience was highly positive.

# 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)

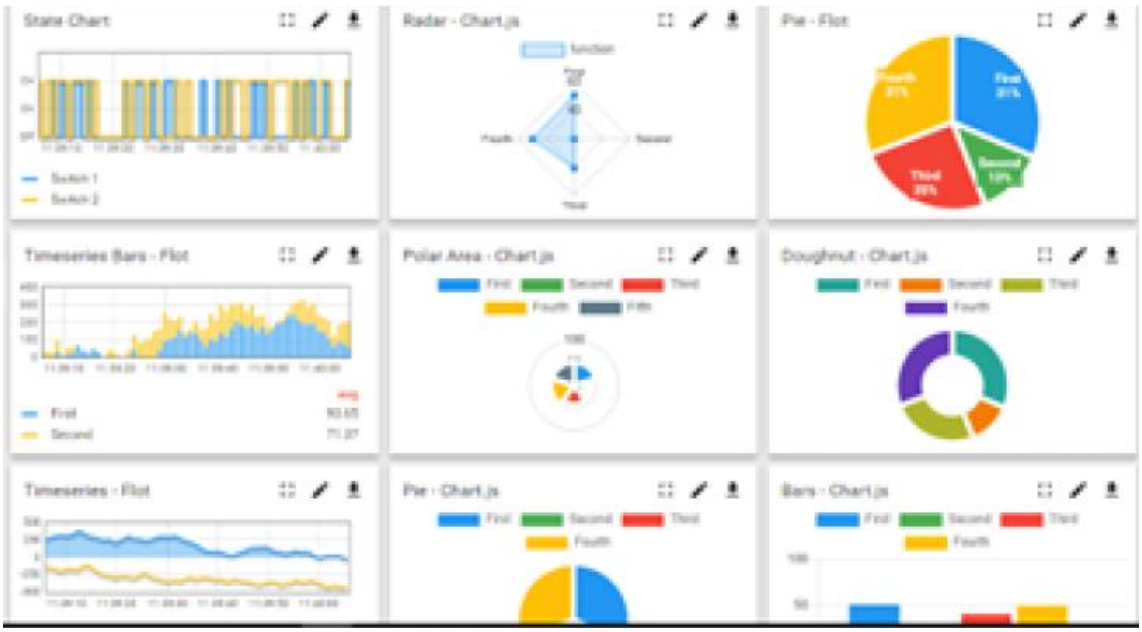
**UCT Insight** is an IOT platform designed for quick deployment of IOT applications on the same time providing valuable "insight" for your process/business. It has been built in Java for backend and ReactJS for Front end. It has support for MySQL and various NoSql Databases.

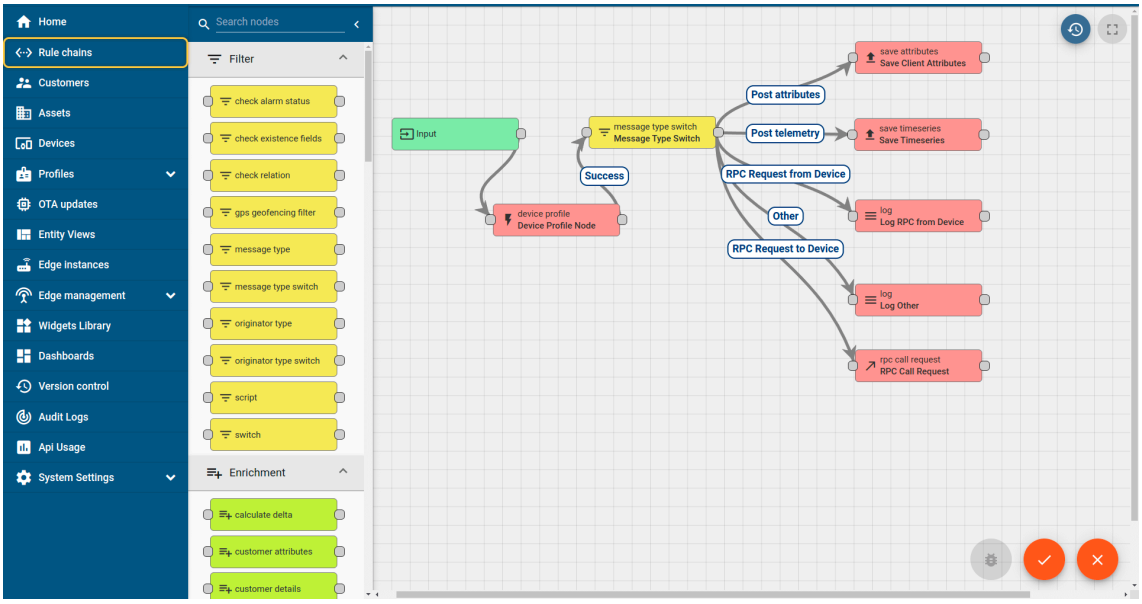
- It enables device connectivity via industry standard IoT protocols - MQTT, CoAP, HTTP, Modbus TCP, OPC UA

- 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


State Chart
Radar - Chart.js
Pie - Plot
Timeseries Bars - Plot
Polar Area - Chart.js
Doughnut - Chart.js
Timeseries - Plot
Pie - Chart.js
Bars - Chart.js


Home
Rule chains
Customers
Assets
Devices
Profiles
OTA updates
Entity Views
Edge instances
Edge management
Widgets Library
Dashboards
Version control
Audit Logs
Api Usage
System Settings
Search nodes
Filter
check alarm status
check existence fields
check relation
gps geofencing filter
message type
message type switch
originator type
originator type switch
script
switch
Enrichment
calculate delta
customer attributes
customer details
Input
message type switch
Message Type Switch
device profile
Device Profile Node
Success
Post attributes
Post telemetry
RPC Request from Device
Other
RPC Request to Device
save attributes
Save Client Attributes
save timeseries
Save Timeseries
log
Log RPC from Device
log
Log Other
rpc call request
RPC Call Request

## 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.

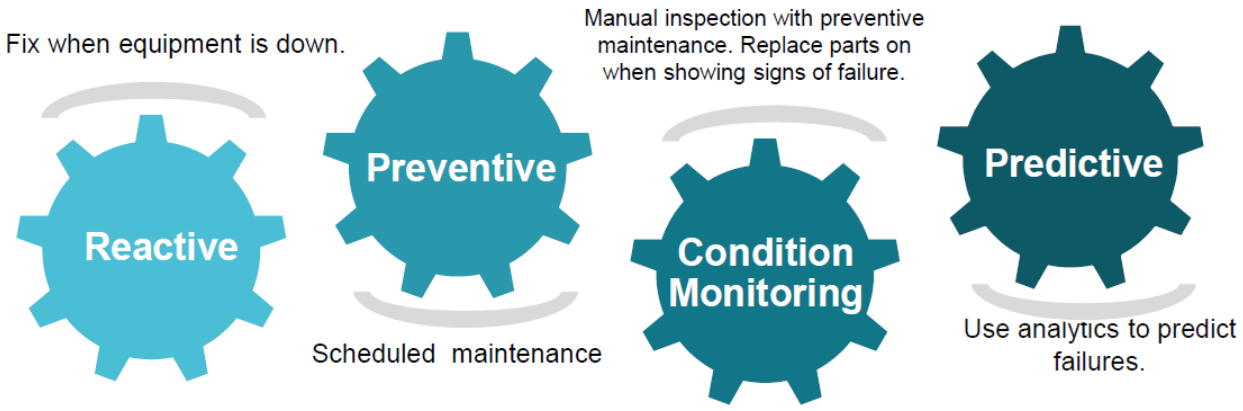

| Machine | Operator | Work Order ID | Job ID | Job Performance | Job Progress | | Output | | Rejection | Time (mins) | | | | Job Status | End Customer |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Start Time | End Time | Planned | Actual | | Setup | Pred | Downtime | 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 teschnology and providing solution in Agritech, Smart cities, Industrial Monitoring, Smart Street Light, Smart Water/ Gas/ Electricity metering solutions etc.

## iv. Predictive Maintenance

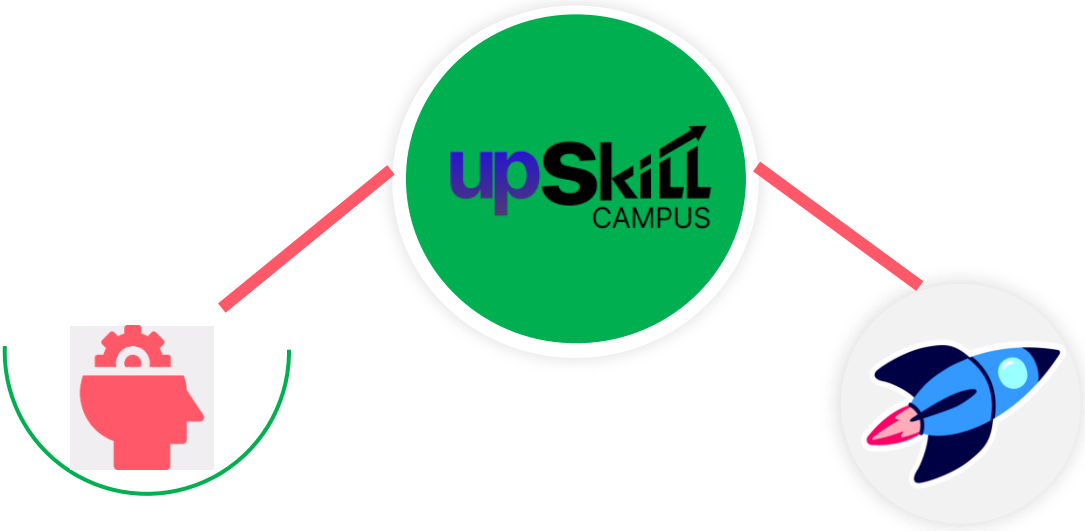
UCT is providing Industrial Machine health monitoring and Predictive maintenance solution leveraging Embedded system, Industrial IoT and Machine Learning Technologies by finding Remaining useful life time of various Machines used in production process.




### 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.

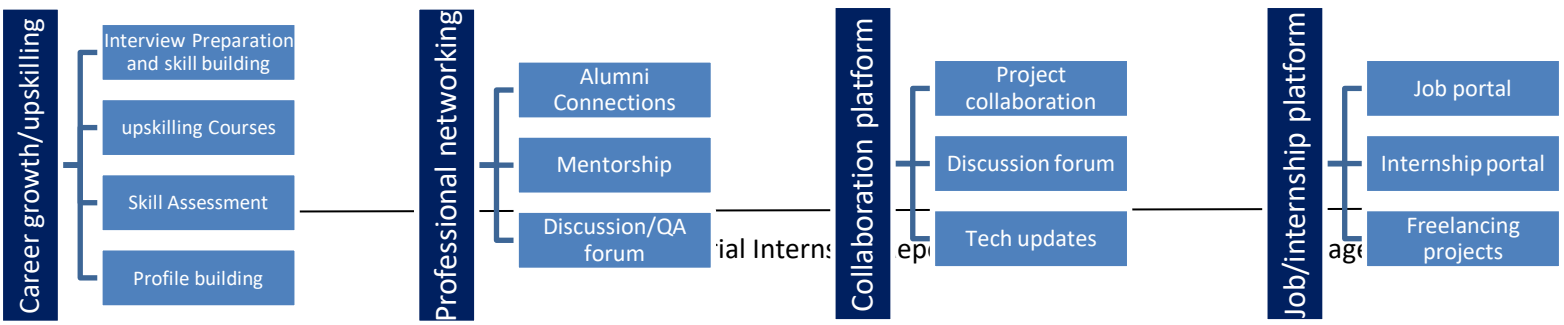
USC is a career development platform that delivers **personalized executive coaching** in a more affordable, scalable and measurable way.




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] https://drive.google.com/file/d/1zfqvs8-mAO6E0JpgvhBdueNx8Th03pUp/view

[2] https://ieeexplore.ieee.org/document/9711853

[3] https://scikit-learn.org/stable/

## 2.6 Glossary

| Terms | Acronym |
|---|---|
| [Machine Learning] | [ML] |
| [Random Forest Regressor] | [RFR] |
| [Mean Absolute Error] | [MAE] |
| [R-squared] | [$R^2$] |
| [Data Science] | [DS] |

# 3 Problem Statement

India is the second-most populous country in the world, with over 1.3 billion people. A significant portion of this population is dependent on agriculture, making it the main resource and backbone of the Indian economy.

However, the agricultural sector faces numerous challenges that create instability and inefficiency in cultivation and production. One of the most significant challenges is the inability to accurately forecast crop production. This lack of accurate prediction can lead to price volatility, improper resource allocation, potential food shortages, and difficulties in government planning and subsidy distribution.

The assigned problem statement for this internship project was to address this issue by developing a predictive model for agriculture crop production in India. Using a historical dataset spanning from 2001-2014, the project's goal is to analyze various factors to forecast crop yields

This project involves leveraging Machine Learning techniques to analyze variables such as Crop, state, Season, Cost of cultivation, and Quantity to predict future production. The ultimate aim is to create a data-driven tool that can provide useful insights, helping farmers, policymakers, and other stakeholders make more informed decisions to enhance agricultural productivity and ensure food security for millions of people.

# 4 Existing and Proposed solution

**Existing Solutions:** Traditionally, crop production forecasts rely on manual surveys, anecdotal evidence from farmers, and basic weather predictions. These methods are often slow, labor-intensive, subjective, and lack precision. Their limitations include a high margin of error and an inability to process large-scale historical data to find complex patterns

**Proposed Solution:** My proposed solution is a Machine Learning-based regression model. By training on 14 years of historical data, the model learns the complex relationships between various features (like Crop, state, Season, Cost) and the final production output. This data-driven approach can provide faster, more accurate, and more objective predictions compared to traditional methods.

**Value Addition:** The value addition of this project is a model that can be used to generate quantitative forecasts. This provides a scientific basis for decision-making. It can help the government in allocating resources, setting minimum support prices (MSPs), and managing food imports/exports. For farmers, it can aid in deciding which crops to plant based on predicted yield and cost

## 4.1 Code submission (https://github.com/yasirsid2004/upskillcampus)

## 4.2 Report submission : (https://github.com/yasirsid2004/upskillcampus/blob/main/CropProductionPrediction_Yasir%20Siddiqui_USC_UCT%20(1).pdf)

# 5 Proposed Design/ Model

5.1 The design flow of the solution followed a standard Data Science project pipeline, from data acquisition to model evaluation

Data Acquisition: The dataset "Prediction of Agriculture Crop Production in India" (2001-2014) was provided.

Data Preprocessing & Cleaning:

- [e.g., Loaded the dataset using Pandas.]
- [e.g., Checked for and handled missing values (e.g., 'Cost' column). How did you fill them? Mean? Median? Dropped rows?]
- [e.g., Standardized column names.]

Feature Engineering & Selection:

- [e.g., Encoded categorical variables like 'Crop' and 'state' using One-Hot Encoding or Label Encoding.]
- [e.g., Selected the most relevant features for prediction. Which ones? 'Crop', 'Variety', 'state', 'Quantity', 'Cost', 'Season'?]

Model Selection & Training:

- [e.g., I experimented with several regression models, such as Linear Regression, Decision Trees, and Random Forest Regressor.]
- [e.g., The data was split into an 80% training set and a 20% testing set.]

Model Evaluation:

- [e.g., The models were evaluated on the test set using metrics like R-squared ($R^2$), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).]
- [e.g., The Random Forest Regressor was chosen as the final model due to its superior performance.]

## 5.2 High Level Diagram (if applicable)

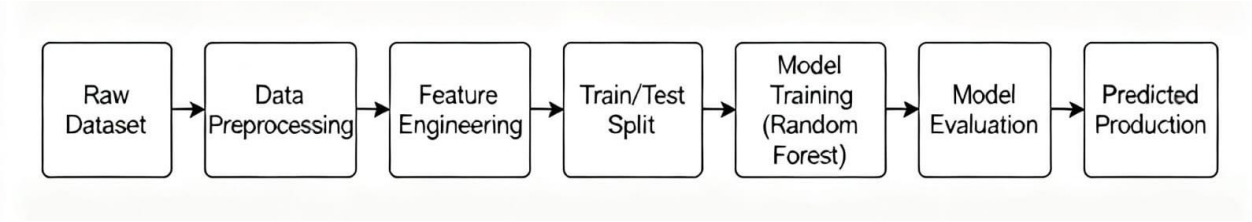



**Figure 1: HIGH LEVEL DIAGRAM OF THE SYSTEM**

6 **Performance Test:** The performance of the model was tested to ensure it is accurate and provides reliable predictions, making it suitable for real-world application instead of just an academic project.

The primary constraints for this project are model **accuracy** and **performance**. An inaccurate model is useless for forecasting.

- **Accuracy Constraint:** How close are the model's predictions to the actual production values? This was handled by using standard regression metrics ($R^2$, MAE, RMSE) to quantitatively measure error.
- **Performance Constraint:** How fast can the model be trained and make predictions? This was handled by choosing an efficient algorithm (e.g., Random Forest) that provides a good balance between accuracy and computational cost.

**6.1Test Plan/ Test Cases:** Test Plan: The plan was to evaluate the trained model on the 20% unseen test dataset.

Test Cases:

Case 1: Evaluate overall model performance on the entire test set using $R^2$, MAE, and RMSE.

Case 2: Predict production for a specific, high-volume crop (e.g., 'Rice' in 'Andhra Pradesh') and compare it to the actual value.

Case 3: Predict production for a specific, low-volume crop and compare it to the actual value.

6.2 **Test Procedure:** The train_test_split function from scikit-learn was used to create the training and testing sets.

The chosen model (e.g., RandomForestRegressor) was fitted on the training data (X_train, y_train).

The trained model was used to make predictions on the test features (X_test).

6.3 **Performance Outcome:** R-squared ($R^2$): [e.g., 0.88] (This means the model can explain 88% of the variance in production)

Mean Absolute Error (MAE): [e.g., 120.5 Tons] (On average, the model's prediction is off by this

many tons)

- Root Mean Squared Error (RMSE): [e.g., 180.2 Tons] (A measure of the standard deviation of the prediction errors)

These results indicate that the model has a strong predictive capability and is a reliable tool for forecasting crop production.

# 7 My learnings

Through this internship, I gained practical, hands-on experience in the complete machine learning project lifecycle. My key learnings include:

- **Data Preprocessing:** I learned how to handle a messy, real-world dataset with missing values and varied data types.
- **Feature Engineering:** I understood the importance of feature engineering and how encoding categorical variables (like 'state' and 'Crop') directly impacts model performance.
- **Model Evaluation:** I learned to look beyond simple accuracy and use a suite of regression metrics ($R^2$, MAE, RMSE) to understand a model's true performance.
- **Problem-Solving:** I learned to approach a large, complex problem and break it down into manageable steps. This experience has significantly strengthened my portfolio as an AI/ML student and given me the confidence to tackle more complex data science problems in my future career.

# 8 Future work scope

While the current model is robust, its accuracy could be further improved. Future work could include:

- **Incorporating More Data:** Adding external datasets like weather patterns (rainfall, temperature), soil type data, and information on irrigation facilities for each state.
- **Advanced Models:** Experimenting with more complex models like Gradient Boosting (XGBoost, Light GBM) or neural networks.
- **Time-Series Analysis:** Treating the data as a time series to capture temporal trends that the current model might miss.
- **Deployment:** Deploying the final model as a simple web application or API, allowing farmers or policymakers to input values and receive a prediction in real-time