Industrial Internship Report on

”Prediction of Agriculture Crop Production in India”

Prepared by

Parth Sharma

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 4 weeks’ time. My project focused on the **Prediction of Agricultural Crop Production in India** using machine learning models. I created a model by using Agricultural yield datasets, combining information on cultivation costs and crop varieties. Using this dataset, I trained and evaluated both a Linear Regression model and a Random Forest Regressor to accurately predict crop yield. 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

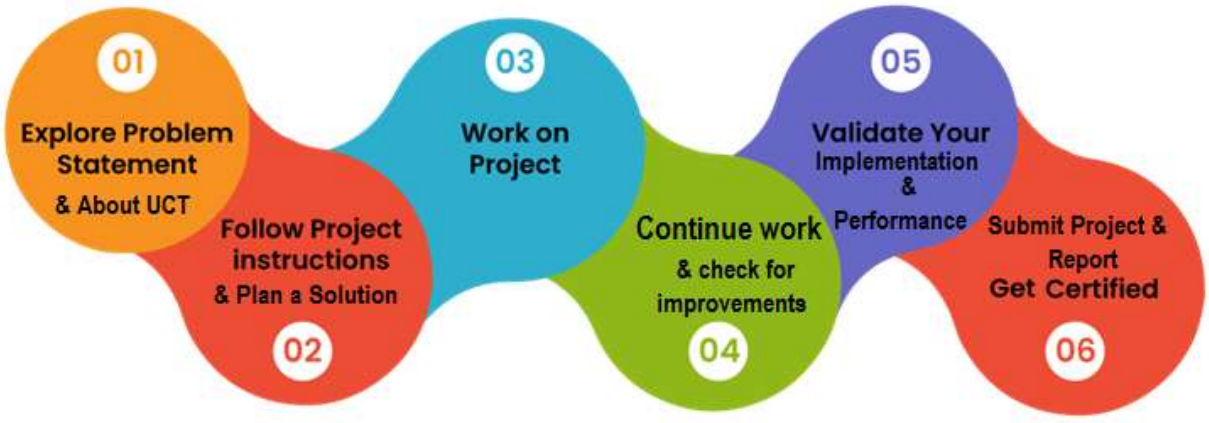
The internship was 4 weeks of learning and working on project. During the first week, I have learned the basics of data science and machine learning, I have studied the basics and the introduction. In second week, i have learned mathematical concepts of machine learning and at the same time started working on my project implementing these concepts. In the next 3 and 4 weeks, I have started working on the models building, testing and report making.

An internship is incredibly important for career development because it bridges the gap between theoretical knowledge and practical application. It allows you to take what you've learned in the classroom and apply it to real-world projects, solving actual business problems. This hands-on experience not only solidifies your skills but also makes your resume more appealing to future employers, as it demonstrates that you're ready to contribute from day one. Additionally, an internship is a perfect opportunity to start building your professional network, which can lead to future job opportunities, mentorship, and invaluable career advice.

My project was a data science and machine learning initiative focused on **predicting agricultural crop yield in India**. The problem statement was to create an accurate and comprehensive model that could forecast yield by integrating data from multiple sources.

Opportunity given by USC/UCT.

The program was planned for 4 weeks. The 4 weeks included learning about data science and machine learning. At the same time, a project was to decided to work on. Including this, weekly quizzes to test ourselves on the new concepts were conducted. A weekly report including all the learnings, challenges and progress made in the project was to be submitted.



This internship was an invaluable experience that transformed my understanding of data science and machine learning from theoretical concepts to practical skills. I learned to apply foundational principles and mathematical concepts by directly implementing them in Python code. This hands-on process allowed me to make continuous progress on my project as I acquired new knowledge. Beyond the technical skills, I gained crucial experience in teamwork and collaboration, and I'm very thankful for the guidance of my mentors and peers throughout this journey.

Thanks to everyone, who have helped me.

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)

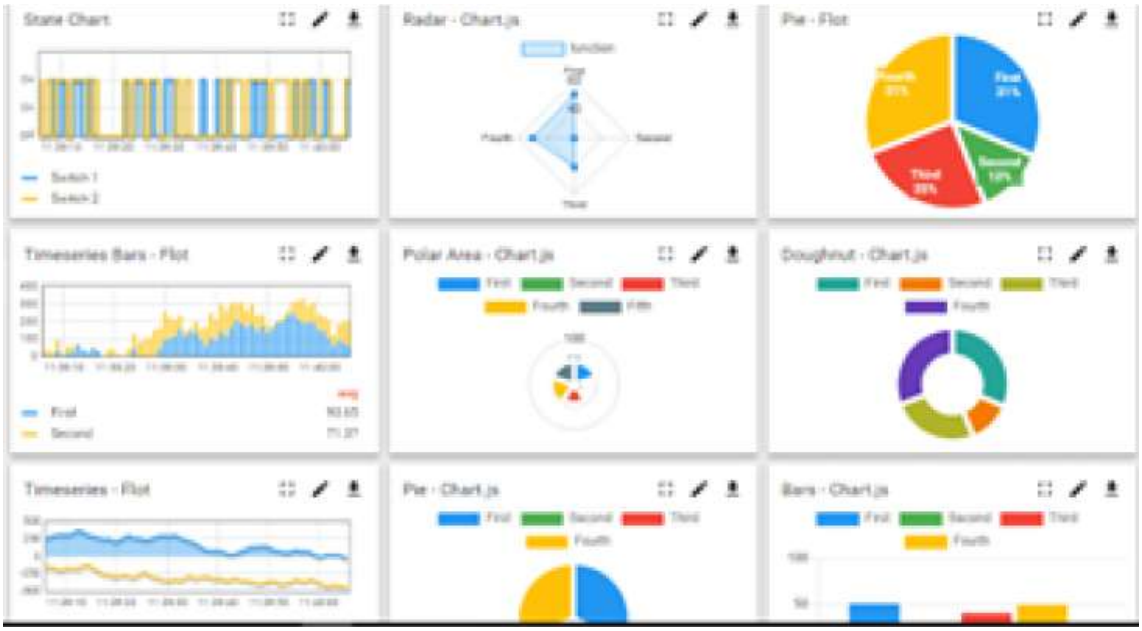
UCT Insight is an IOT platform designed for quick deployment of IOT applications on the same time providing valuable "insight" for your process/business. It has been built in Java for backend and ReactJS for Front end. It has support for MySQL and various NoSql Databases.

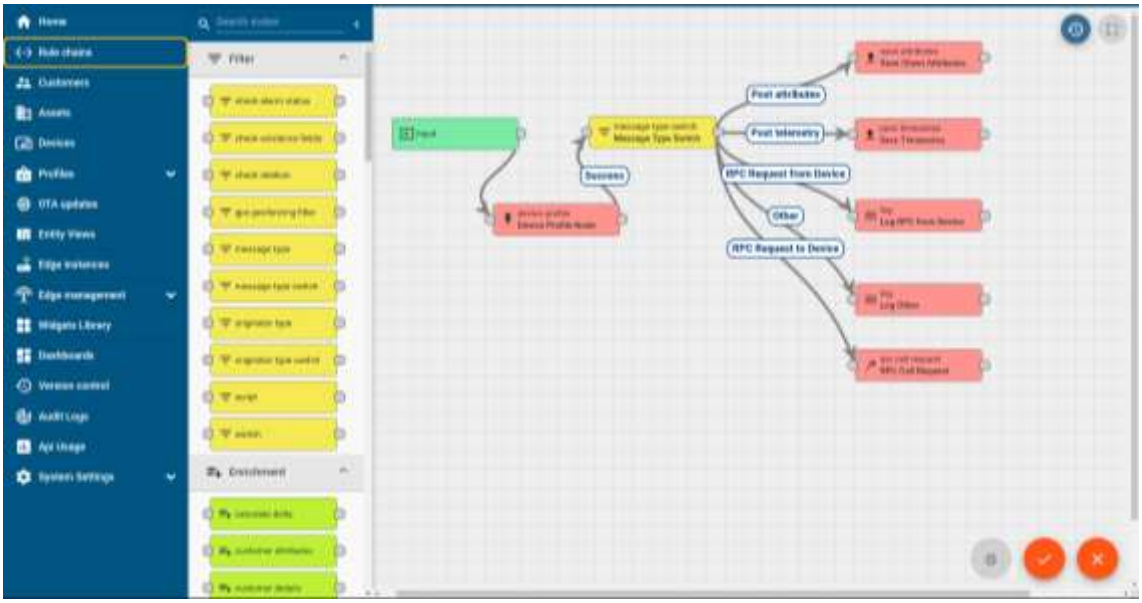
- It enables device connectivity via industry standard IoT protocols - MQTT, CoAP, HTTP, Modbus TCP, OPC UA

- 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


Post attributes
Post telemetry
RPC Request from Device
Other
RPC Request to Device
Success

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.


Demo Org
Demo
CNC_S7_81
CNC_S7_82
CNC_S7_84
Utilization
Utilization
Utilization

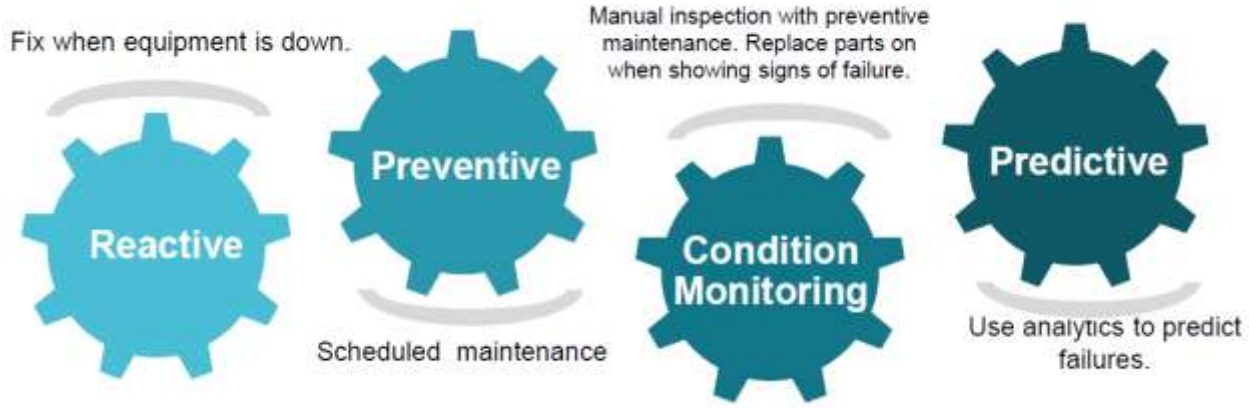

3D View
12:30
Search...
Stopped
Workplace Type
194724_27462-100-A3
Availability
Performance
Quality
OEE
Monitor
42.23%
400.00v
99.21%
99.21%
0.00%
0.00v
24.53%
24.53%
0.0%
200.00v
62.35%
62.35%

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

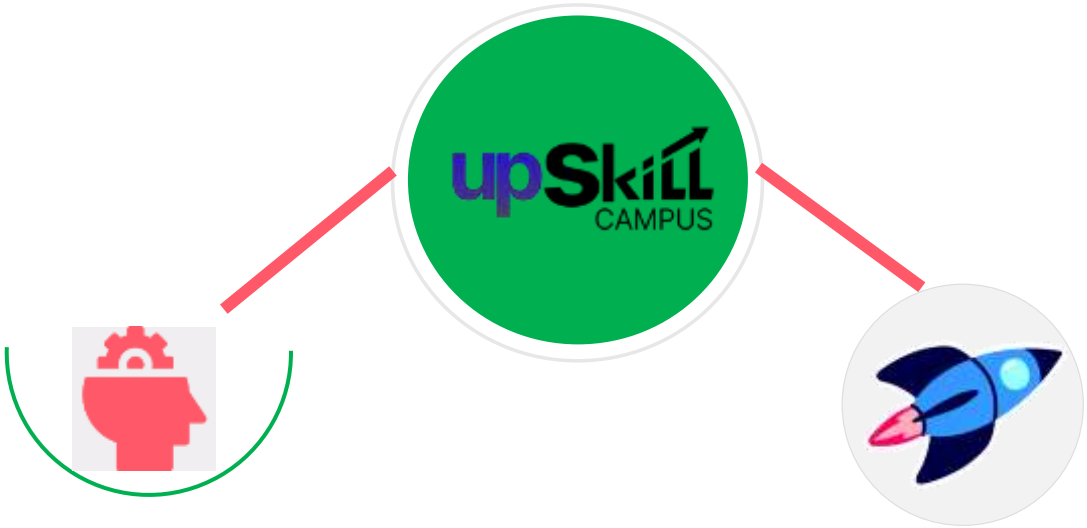
UCT is providing Industrial Machine health monitoring and Predictive maintenance solution leveraging Embedded system, Industrial IoT and Machine Learning Technologies by finding Remaining useful life time of various Machines used in production process.



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.

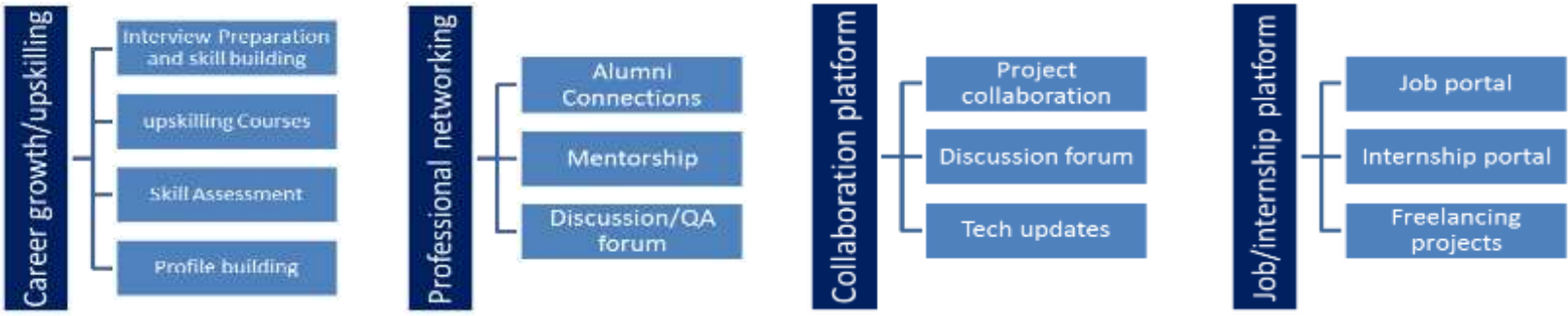
USC is a career development platform that delivers **personalized executive coaching** in a more affordable, scalable and measurable way.



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] Elbasi, E., Zaki, C., Topcu, A. E., Abdelbaki, W., Zreikat, A. I., Shdefat, A., Saker, L., & Cina, E. **Crop Prediction Model Using Machine Learning Algorithms**.

[2] Yamparala, R., Shaik, H. S., Guntaka, N. S., & Marr, P. (2022). **Crop Yield Prediction using Random Forest Algorithm**.

[3] Babu, G. M., Reddy, K. A., Vyshnavi, N., Harika, J., & Suhani, S. M. (Year). **Crop yield prediction using Random Forest Algorithm**.

2.6 Glossary

Terms	Acronym
Random Forest	A type of ensemble learning model that builds multiple decisions trees and averages their predictions to provide a more accurate and stable forecast for continuous values.
R^2 Score	A metric that represents the proportion of the variance in the target variable.
Mean Squared Error	MSE: A metric that measures the average squared difference between the predicted and actual values. A lower MSE means a more accurate model.
Feature	A measurable property or characteristic of a phenomenon being observed. In this project, features include State, Season, Cost etc.
Target	The variable that machine learning model is being trained to predict. In this project, the target is Yield.

3 Problem Statement

Crop yield production can be a very helpful thing for agricultural sector. Crop yield depends on many factors such as type of crop, state, season, variety, cost of cultivation or production etc. Using these factors, we can predict crop yield and provide help to the agriculture sector. The prediction will help in resource planning, policy-making and market forecasting. It will help the farmers to make better decisions and it will also help the government to ensure food security in our country. The challenge is building a predictive model that can accurately predict yield by using diverse factors. The project for focuses on building this model, and it testing it, for results and further improvements.

4 Existing and Proposed solution

Existing Solutions

Many existing solutions for crop yield prediction often rely on simple, isolated models that use a single dataset. A common approach is to use a basic linear regression model on a limited set of features, such as the cost of cultivation. The main limitations of these solutions are:

- **Limited Accuracy:** Simple models cannot capture complex, non-linear relationships between multiple features.
- **Incomplete Data:** Using a single dataset provides an incomplete picture. Factors like crop variety and growing season are often overlooked, leading to less accurate predictions.

Proposed Solution

My proposed solution is to build a more comprehensive and accurate predictive model by **merging multiple datasets**. By combining data on cultivation costs with information on crop variety and season, the model can learn from a richer feature set. The project then evaluates two different machine learning models, **Linear Regression** and **Random Forest Regressor**, to determine which provides the best performance.

Value Addition

The key value addition of this approach is the creation of a **holistic and robust model**. Instead of making predictions based on limited information, the model integrates diverse factors, leading to a more accurate and reliable forecast. The comparison between the two models also provides a data-driven recommendation for the best predictive tool for this specific problem.

4.1 Code submission (Github link)

4.2 Report submission (Github link) : first make placeholder, copy the link.

5 Proposed Design/ Model

The project's design is like this - loading the raw datasets and then standardizing the 'Crop' column to ensure consistent naming. The two datasets were then merged using a left join on this standardized column. Any missing values that resulted from the merge were handled by filling them with 'Not Available'. Following this, all categorical features (like crop and state names) were one-hot encoded, a technique that converts them into a numerical format suitable for machine learning models. The preprocessed data was then split into training and testing sets. Two models, a Linear Regression and a Random Forest Regressor, were trained on the data. Their performance was evaluated using metrics like the R2 score and Mean Squared Error (MSE), and the results were used to make a practical prediction and complete the project.

5.1 High Level Diagram (if applicable)

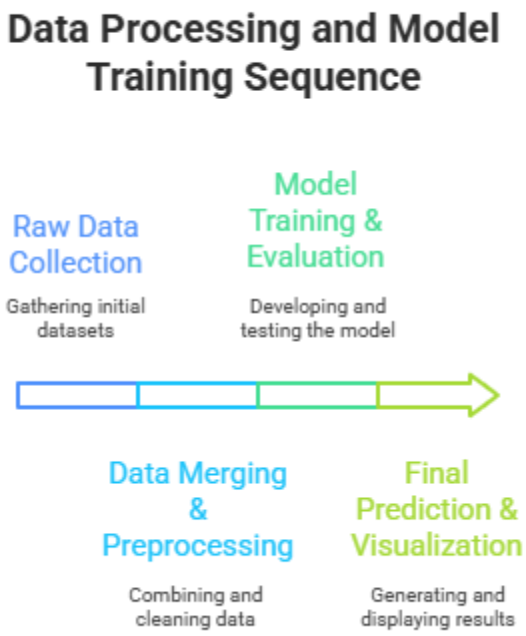
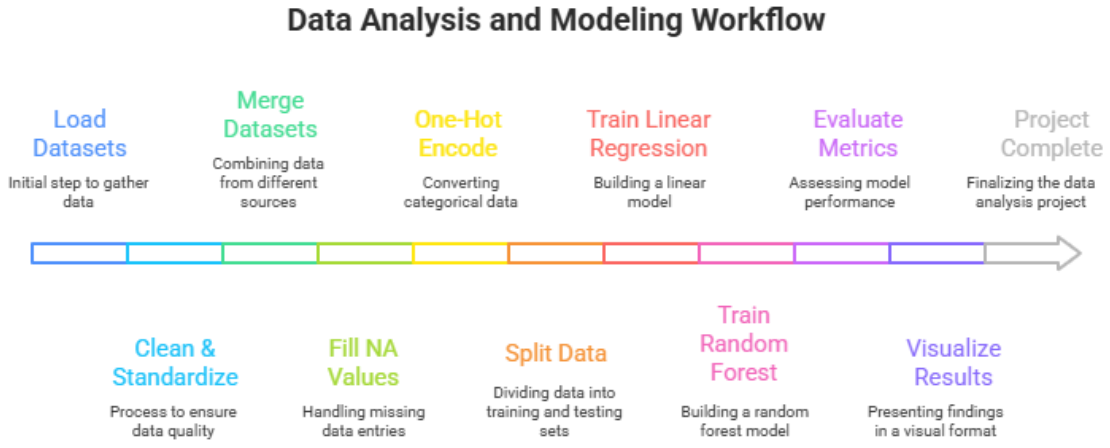


Figure 1: HIGH LEVEL DIAGRAM OF THE SYSTEM

5.2 Low Level Diagram (if applicable)



6 Performance Test

The key performance constraint for this project is **accuracy**. Since the goal is to provide a reliable prediction, the model's accuracy is paramount. Another important constraint is **computational speed**, as a model that takes too long to train or predict is not practical for real-world applications.

6.1 Test Plan/ Test Cases

To test the model's performance, the dataset was split into a training set (80%) and a testing set (20%). The models were trained on the training data and then evaluated on the unseen testing data. The key test cases were:

- Does the model provide a good fit for the data? (Measured by R2)
- Is the model's prediction error low? (Measured by MSE)
- Is the model's prediction accurate on a single, randomly selected data point?

6.2 Test Procedure

The testing procedure was executed to assess the performance of the trained machine learning models on new, unseen data. First, the models were used to generate predictions on the dedicated test set. These predicted values were then compared against the actual values from the same test set. The models' effectiveness was quantitatively measured using two key metrics: the R2 score, which indicates how well the model's predictions align with the actual data, and the Mean Squared Error (MSE), which calculates the average squared difference between the predicted and actual values. This rigorous evaluation provided an objective measure of the models' accuracy and reliability.

6.3 Performance Outcome

The models were evaluated on the test set, and the following outcomes were observed:

- **Linear Regression:**
    - R2 Score: 0.9446
    - Mean Squared Error (MSE): 1501.3239
- **Random Forest Regressor:**

- R2 Score: 0.9988
- Mean Squared Error (MSE): 31.5877

The test results clearly show that the **Random Forest Regressor significantly outperformed the Linear Regression model**. Its high R2 score (0.9988) indicates that it fits the data exceptionally well, and its low MSE (31.5877) confirms its superior accuracy in predicting crop yield. The computational time for both models was well within acceptable limits for a dataset of this size.

7 My learnings

The project was an incredible learning experience. I have learned not only theoretical but practical knowledge about data science and machine learning. During the first week, I have learned the basics of data science and machine learning, I have studied the basics and the introduction. Over the next couple of weeks, I have learned the theoretical aspects of machine learning, new mathematical concepts related to machine learning. Also, I have learned how to program these concepts in python. At the same time, I am learning these concepts, I have started working on my project and made continuous progress as I learn more and more concepts. Except from the technical aspect, I have learned how to work with other people and I want to thank my mentors and peers for helping me during this project.

8 Future work scope

While this project successfully built a strong predictive model, there is always room for future improvements and expansions, including:

- **Incorporating More Datasets:** Integrating additional data, such as weather patterns (e.g., rainfall, temperature) and soil quality, would provide even more features and likely improve the model's accuracy.
- **Exploring More Models:** We could experiment with other advanced models like Gradient Boosting (e.g., XGBoost, LightGBM) to see if they can capture more complex relationships and achieve even higher accuracy.
- **Hyperparameter Tuning:** The performance of the Random Forest model can be further optimized by tuning its hyperparameters (e.g., number of estimators, max depth) to achieve even better results.
- **Developing a User Interface:** The project could be extended by building a simple web application where a user can input various factors and get an instant crop yield prediction. This would make the model more accessible and practical for end-users.