Industrial Internship Report on

” Prediction of Agriculture Crop Production in India”

Prepared by

THARUN J

Executive Summary

This report presents the work completed during the 6-week Industrial Internship conducted by upskill Campus in collaboration with UniConverge Technologies Pvt Ltd (UCT). The project focused on developing a Machine Learning based system to predict agricultural crop production in India. The solution leverages data preprocessing, feature engineering, model optimization using XGBoost, and deployment through a Streamlit web application. The system provides accurate yield predictions and demonstrates real-world applicability in the agriculture domain.

TABLE OF CONTENTS

1.PREFACE

This report presents the comprehensive work completed during the six-week Industrial Internship conducted by upskill Campus (USC) in collaboration with UniConverge Technologies Pvt Ltd (UCT). The internship provided an opportunity to work on a real-world industry-oriented project titled **“Machine Learning-Based Prediction of Agricultural Crop Production in India.”**

Over the course of six weeks, the project progressed systematically from understanding the problem statement and exploring agricultural datasets to building, optimizing, and deploying a machine learning model. The initial weeks focused on data collection, exploratory data analysis (EDA), preprocessing, and baseline model development. In the intermediate phase, advanced techniques such as hyperparameter tuning, feature engineering, and model interpretability using SHAP were implemented. In the final phase, the optimized XGBoost model was deployed through a Streamlit web application, followed by extensive testing, documentation, and presentation preparation.

Internships play a crucial role in career development as they bridge the gap between theoretical knowledge and practical industry application. This internship allowed me to gain hands-on experience in solving real-world problems, working with large datasets, applying machine learning algorithms, and understanding deployment strategies. It enhanced both my technical competencies and professional skills such as documentation, time management, and structured problem-solving.

The opportunity provided by USC and UCT enabled exposure to industry standards, structured mentorship, and practical implementation of data science concepts. The program was carefully planned with weekly milestones, ensuring gradual development from problem understanding to final deployment. Continuous evaluation and structured guidance helped in achieving a complete end-to-end solution.

Overall, this internship has been a valuable learning experience that strengthened my technical foundation and prepared me for future roles in Machine Learning and Data Science.

MY LEARNINGS AND OVERALL EXPERIENCE:

The six-week internship provided a comprehensive understanding of the complete Machine Learning lifecycle, from problem identification to deployment. I learned how to work with real-world datasets, perform exploratory data analysis, handle missing values, encode categorical variables, and manage high-cardinality features effectively.

One of the most valuable learnings was model optimization and evaluation. I gained hands-on experience in implementing algorithms such as Random Forest and XGBoost, performing hyperparameter tuning using GridSearchCV, and evaluating model performance using MAE, RMSE, and R^2 score. Understanding model interpretability through SHAP values further strengthened my analytical thinking and ability to explain model behavior.

Additionally, deploying the trained model using Streamlit gave me practical exposure to transforming a research-based model into a user-friendly web application. This helped me understand the importance of scalability, inference time optimization, and reproducibility in real-world systems.

Beyond technical skills, this internship improved my time management, documentation practices, structured thinking, and presentation skills. Working with weekly milestones ensured disciplined progress and systematic project development.

Overall, this internship was a highly enriching experience that enhanced both my technical foundation and professional confidence in the field of Machine Learning and Data Science.

ACKNOWLEDGEMENT:

I would like to sincerely thank **upskill Campus (USC)** and **UniConverge Technologies Pvt Ltd (UCT)** for providing me with this valuable internship opportunity and structured learning platform.

I extend my heartfelt gratitude to the mentors and coordinators of the internship program for their continuous guidance, industry insights, and support throughout the six weeks. Their mentorship played a crucial role in the successful completion of this project.

I would like to express my special thanks to **Mrs. Vishnu Priya Mam**, my professor, for her constant encouragement, academic guidance, and valuable suggestions during the course of this internship. Her support and motivation helped me improve both technically and professionally.

I am also thankful to my college faculty members, friends, and peers who supported me directly and indirectly throughout this journey.

MESSAGE TO JUNIORS AND PEERS:

To my juniors and peers, I would strongly recommend participating in internships that provide real-world problem statements and structured project development. Such experiences not only enhance technical knowledge but also build confidence and industry readiness.

Focus on understanding the fundamentals deeply, practice consistently, and do not hesitate to explore new tools and technologies. Real growth happens when you step outside your comfort zone and apply your knowledge practically.

Stay curious, stay consistent, and always aim to build solutions that create real-world impact.

2.INTRODUCTION

The Industrial Internship conducted by upskill Campus (USC) in collaboration with UniConverge Technologies Pvt Ltd (UCT) provided an opportunity to work on a real-world data-driven problem. The project titled **“Machine Learning-Based Prediction of Agricultural Crop Production in India”** focuses on developing a predictive system that estimates crop yield using historical agricultural data.

Agriculture plays a critical role in the Indian economy, and accurate prediction of crop production can significantly assist farmers, policymakers, and agricultural planners in decision-making. Through this internship, I applied data science and machine learning techniques to design, optimize, and deploy a practical predictive model. The internship enabled exposure to industry standards, structured development processes, and deployment methodologies.

2.1About 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.

Uniconverge Technologies

IIOT Products

We offer product ranging from Remote IOs, Wireless IOs, LoRaWAN Sensor Nodes/ Gateways, Signal converter and IoT gateways

IIOT Solutions

We offer solutions like OEE, Predictive Maintenance, LoRaWAN based Remote Monitoring, IoT Platform, Business Intelligence…

OEM Services

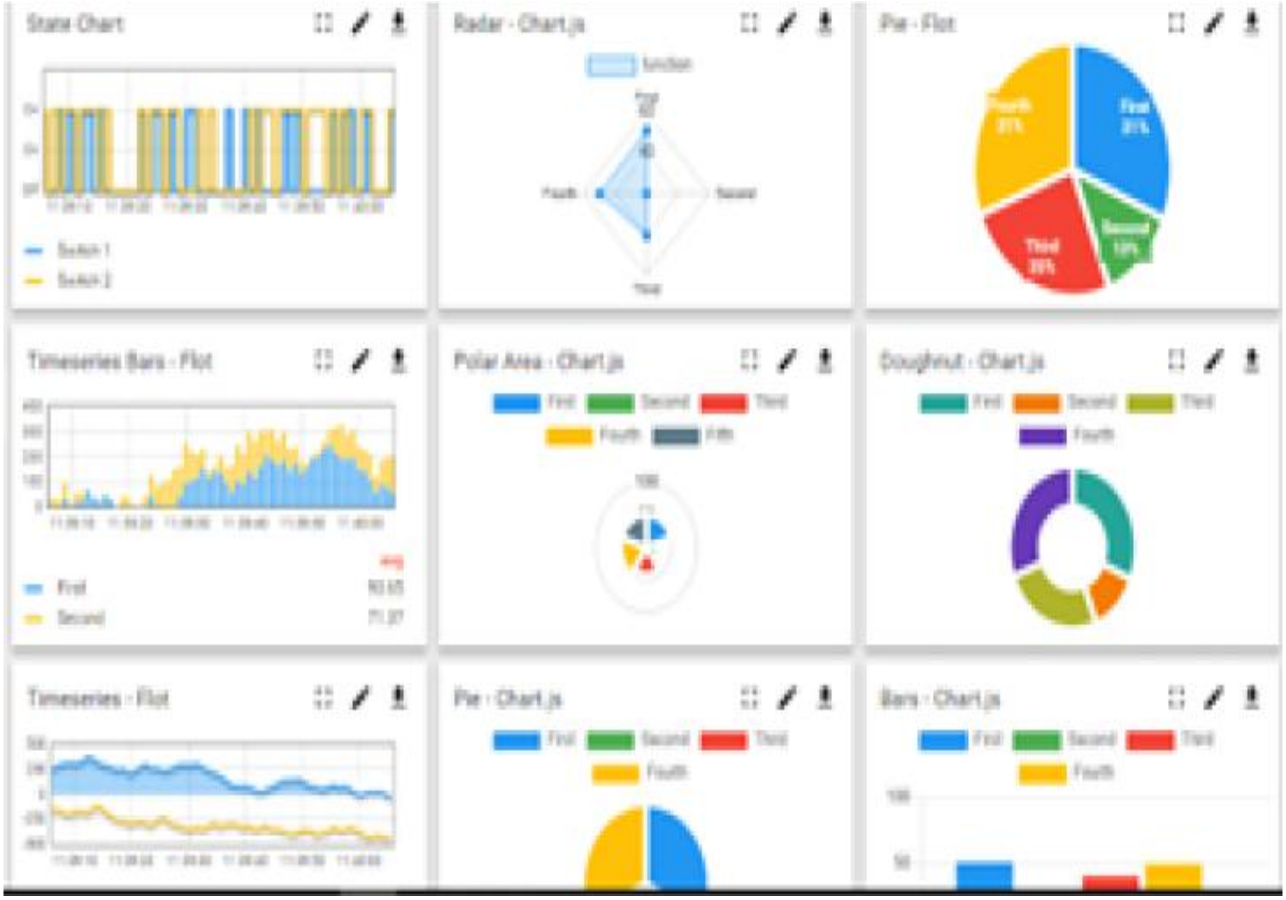
We offer solutions ranging from product design to final production we handle everything for you..

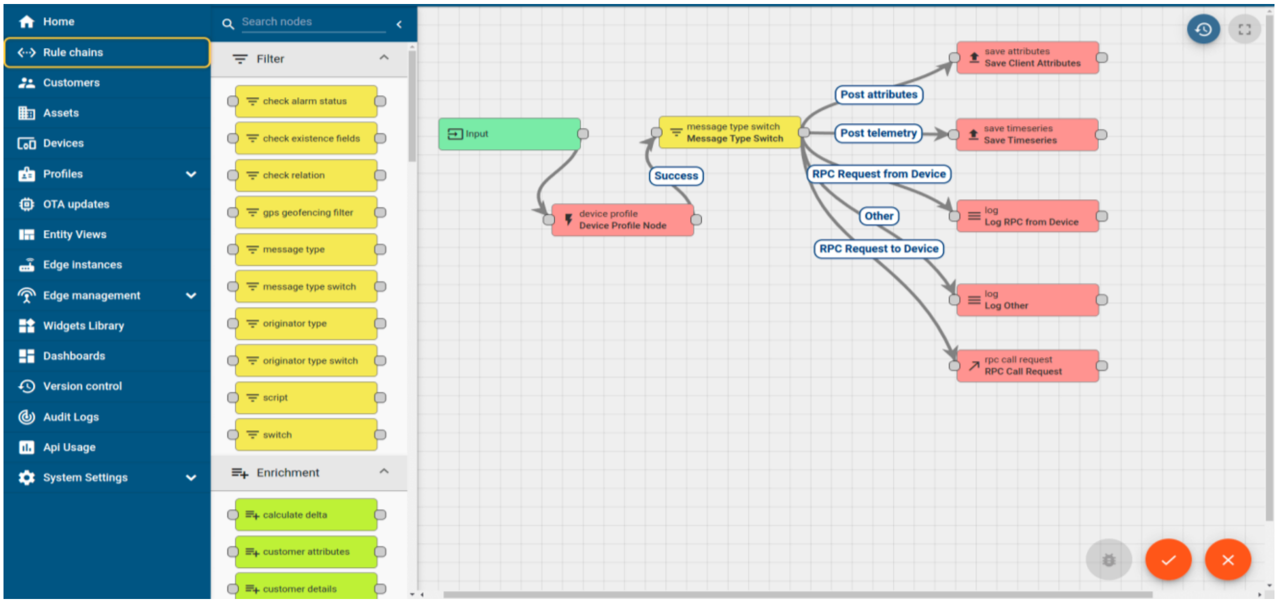
i. UCT IoT Platform (uct Insight)

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





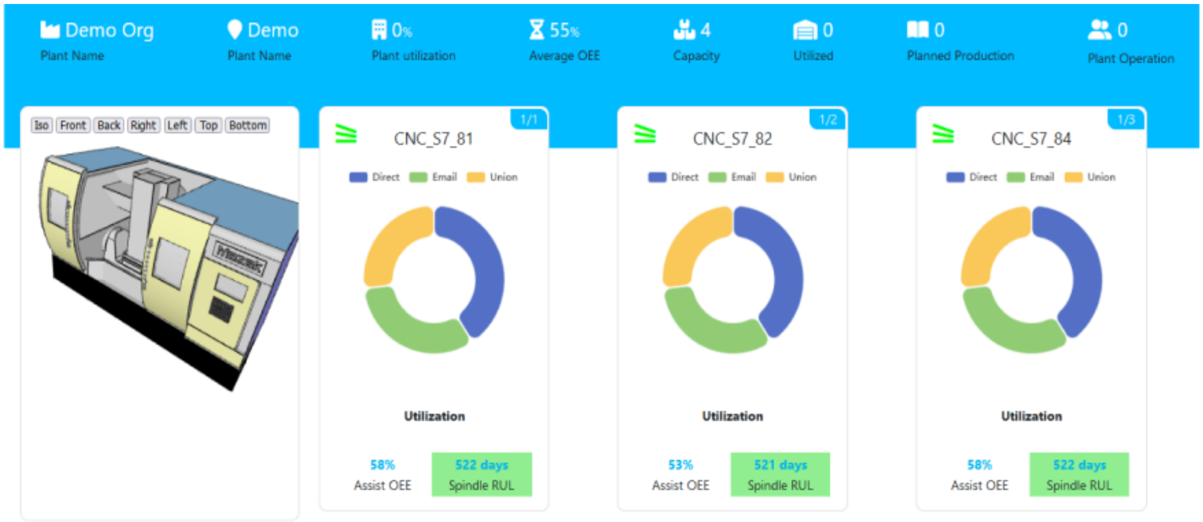
ii. Smart Factory Platform (FACTORY WATCH)

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.



Demo Org
Plant Name

Demo
Plant Name

0%
Plant utilization

55%
Average OEE

4
Capacity

0
Utilized

0
Planned Production

0
Plant Operation

CNC_S7_81 — 1/1

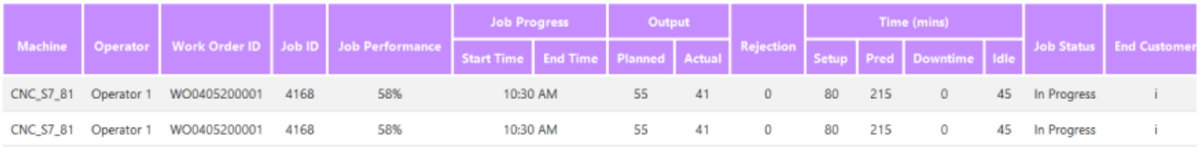
Utilization
58% Assist OEE
522 days Spindle RUL

CNC_S7_82 — 1/2

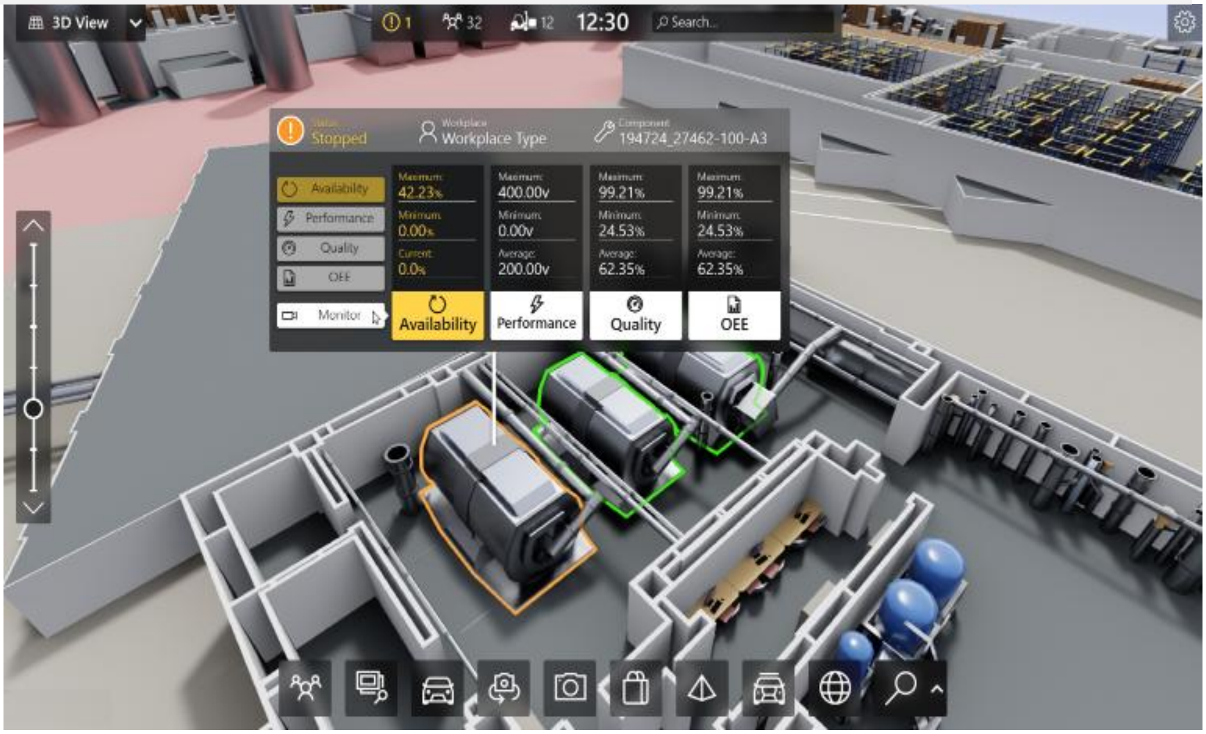
Utilization
53% Assist OEE
521 days Spindle RUL

CNC_S7_84 — 1/3

Utilization
58% Assist OEE
522 days Spindle RUL



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



Status: Stopped
Workplace: Workplace Type
Component: 194724_27462-100-A3

Availability — Maximum: 42.23% Minimum: 0.00% Current: 0.0%

Performance — Maximum: 400.00v Minimum: 0.00v Average: 200.00v

Quality — Maximum: 99.21% Minimum: 24.53% Average: 62.35%

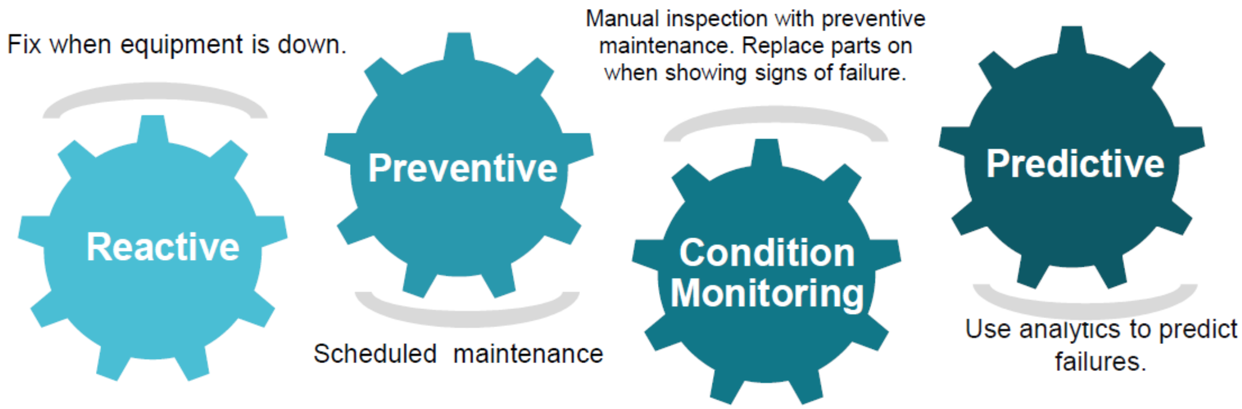
OEE — Maximum: 99.21% Minimum: 24.53% Average: 62.35%

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

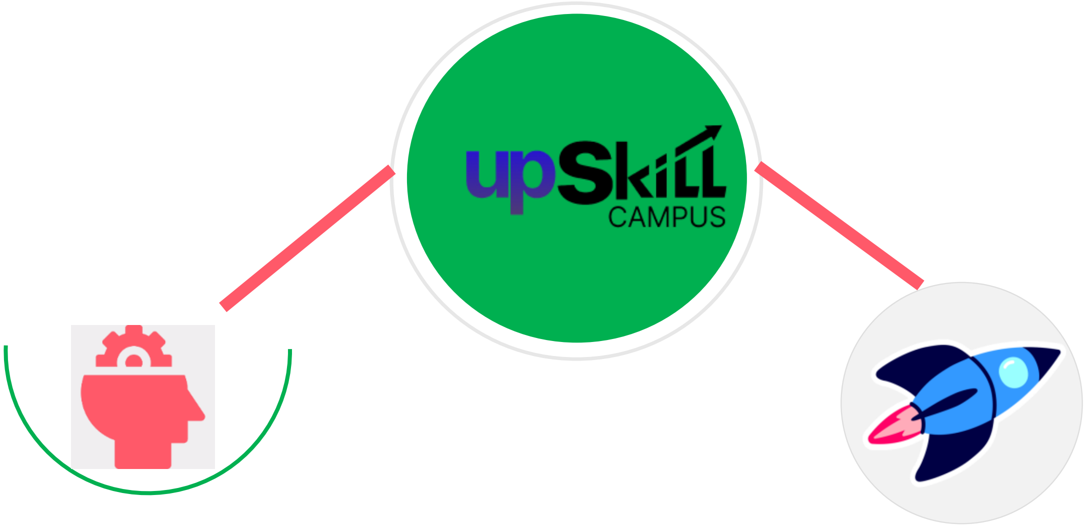
UCT is providing Industrial Machine health monitoring and Predictive maintenance solution leveraging Embedded system, Industrial IoT and Machine Learning Technologies by finding Remaining useful life time of various Machines used in production process.


Fix when equipment is down.
Reactive
Preventive
Scheduled maintenance
Manual inspection with preventive maintenance. Replace parts on when showing signs of failure.
Condition Monitoring
Predictive
Use analytics to predict failures.

2.2About 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.

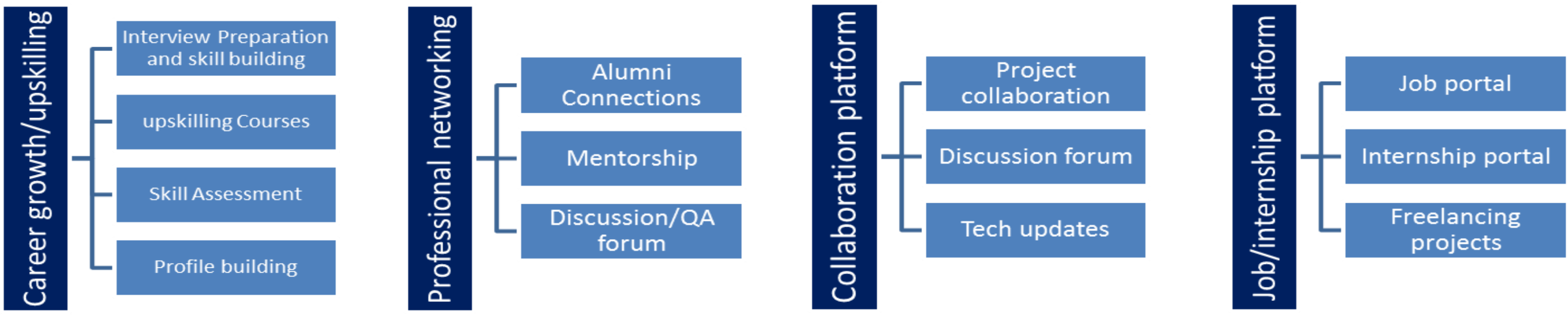
USC is a career development platform that delivers **personalized executive coaching** in a more affordable, scalable and measurable way.



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/



Career growth/upskilling
- Interview Preparation and skill building
- upskilling Courses
- Skill Assessment
- Profile building

Professional networking
- Alumni Connections
- Mentorship
- Discussion/QA forum

Collaboration platform
- Project collaboration
- Discussion forum
- Tech updates

Job/internship platform
- Job portal
- Internship portal
- Freelancing projects

2.3The 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.

2.4Objectives 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.

2.5Reference

[1] Scikit-learn Documentation – https://scikit-learn.org/
[2] XGBoost Documentation – https://xgboost.readthedocs.io/
[3] Streamlit Documentation – https://docs.streamlit.io/
[4] Kaggle – Agricultural Datasets and ML Case Studies
[5] Research papers on Agricultural Yield Prediction using Machine Learning

2.6Glossary

Term	**Description**
ML	Machine Learning
EDA	Exploratory Data Analysis
MAE	Mean Absolute Error
RMSE	Root Mean Square Error
R^2 Score	Coefficient of Determination
SHAP	SHapley Additive exPlanations (Model Interpretability Tool)
XGBoost	Extreme Gradient Boosting Algorithm
Deployment	Making the trained model available for real-time use
Hyperparameter Tuning	Process of optimizing model parameters

3.PROBLEM STATEMENT

In the assigned problem statement, the objective was to design and implement a predictive system capable of estimating agricultural crop production in India using Machine Learning techniques. Agriculture is a highly dynamic sector influenced by multiple factors such as crop type, season, cultivation cost, and regional variations. Traditional yield estimation methods are often manual, time-consuming, and lack predictive accuracy.

The challenge was to analyze historical agricultural datasets, preprocess and engineer relevant features, and build a regression model that can accurately predict crop production quantities. The system must handle high-cardinality categorical variables, data imbalance, and regional diversity while ensuring generalization across different seasons and locations.

The final goal was not only to build a high-performing model but also to deploy it as a user-friendly web application that allows users to input parameters and obtain real-time yield predictions.

4. EXISTING AND PROPOSED SOLUTION

Existing Solutions

Existing agricultural yield estimation approaches generally rely on:

- Traditional statistical methods
- Manual forecasting based on historical averages
- Basic regression models without optimization
- Region-specific isolated prediction systems

Limitations of Existing Solutions:

- Low prediction accuracy
- Lack of scalability across multiple regions
- Limited handling of categorical and seasonal variations
- Absence of real-time deployment platforms
- Poor model interpretability

Many existing solutions do not integrate advanced machine learning optimization techniques or provide accessible web-based interfaces for practical use.

Proposed Solution

The proposed solution is a **Machine Learning-based Crop Production Prediction System** developed using advanced regression algorithms.

Key features of the proposed solution include:

- Comprehensive data preprocessing and feature engineering
- Target Encoding for high-cardinality categorical features
- RobustScaler to manage outliers
- Model comparison between Random Forest and XGBoost
- Hyperparameter tuning using GridSearchCV
- Model interpretability using SHAP values
- Deployment using Streamlit web application

The optimized XGBoost model demonstrated superior performance in terms of accuracy and generalization.

Value Addition

The project adds value in the following ways:

- Improved prediction accuracy compared to traditional methods
- Scalable solution applicable across multiple regions and seasons
- Real-time prediction through web deployment
- Transparent model interpretation using SHAP
- Practical application for agricultural planning and decision-making

This solution bridges the gap between theoretical machine learning concepts and real-world agricultural applications.

4.1Code submission (Github link)

https://github.com/Tharun-J10/upskillcampus

4.2Report submission (Github link) :

https://github.com/yourusername/upskillcampus/blob/main/CropProductionPrediction_TharunJ_USC_UCT.pdf

5.PROPOSED DESIGN/ MODEL

The proposed system is a Machine Learning-based predictive model designed to estimate agricultural crop production using historical agricultural data. The system follows a structured end-to-end pipeline consisting of data preprocessing, model training, evaluation, and deployment.

The architecture ensures scalability, reproducibility, and real-time usability. The solution integrates advanced regression techniques and a web-based interface to make predictions accessible to users.

The overall system design consists of three main layers:

1. Data Processing Layer
2. Machine Learning Model Layer
3. Deployment & User Interaction Layer

The workflow begins with raw agricultural data and ends with real-time crop production predictions through a web application.

5.1High Level Diagram (if applicable)

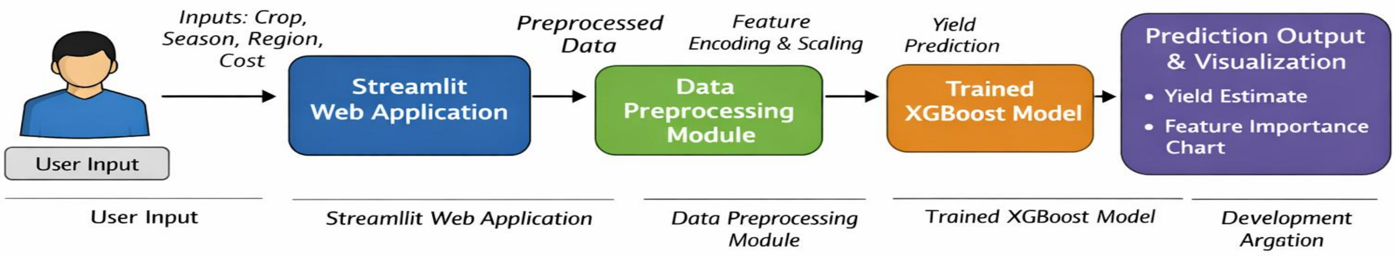
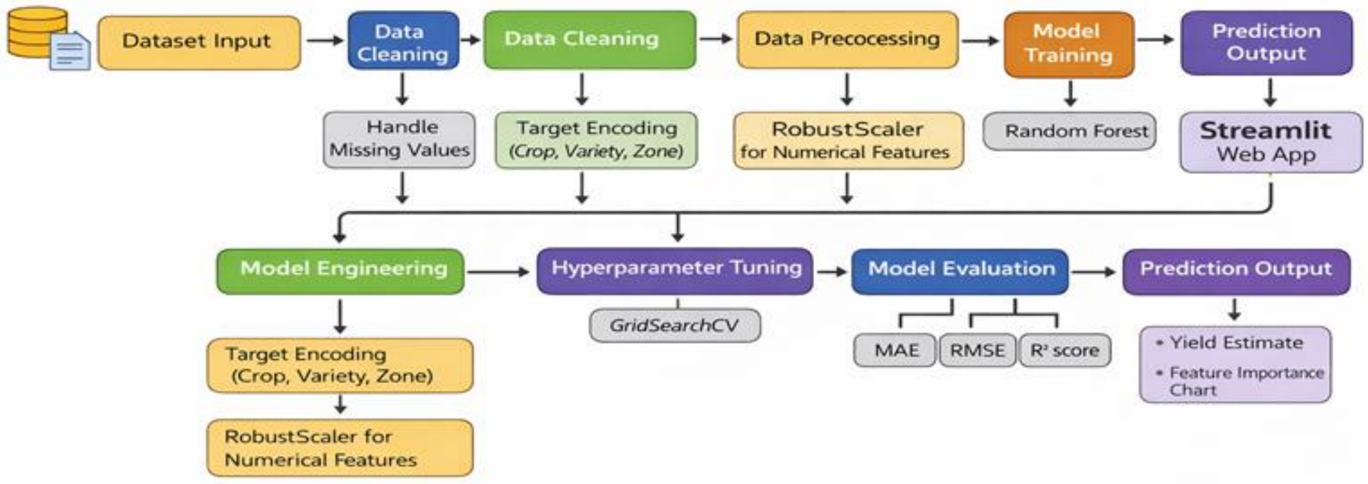

Inputs: Crop, Season, Region, Cost
User Input
Streamlit Web Application
Preprocessed Data
Data Preprocessing Module
Feature Encoding & Scaling
Yield Prediction
Trained XGBoost Model
Prediction Output & Visualization
• Yield Estimate
• Feature Importance Chart
User Input
Streamlit Web Application
Data Preprocessing Module
Trained XGBoost Model
Development Argation

Figure 1: HIGH LEVEL DIAGRAM OF THE SYSTEM

5.2.Low Level Diagram (if applicable)



Low Level Design of System

5.3Interfaces (if applicable)

The system includes the following interfaces:

1. User Interface (UI)

Developed using Streamlit.
Features include:

- Input fields for crop type, season, region, and cost
- Prediction button
- Display of predicted yield
- Interactive visualizations

2. Data Processing Interface

Handles transformation between raw input and model-ready data:

- Encodes categorical variables
- Applies scaling
- Ensures consistency with trained model parameters

3. Model Interface

- Loads trained XGBoost model
- Accepts processed input features
- Returns predicted output
- Integrated with SHAP for interpretability

4. Output Interface

- Displays prediction result
- Shows feature importance plots
- Provides prediction vs actual comparison

6.PERFORMANCE TEST

This section demonstrates why the developed system is suitable for real-world industrial applications rather than being just an academic implementation. The performance evaluation was designed by identifying practical constraints that affect deployment, scalability, and usability in real environments.

Identified Constraints

The following industrial constraints were considered:

1. **Prediction Accuracy Constraint**
   The system must produce highly reliable predictions across different regions and seasons.

2. **Generalization Constraint**
   The model should not overfit to specific crops or geographical zones.

3. **Inference Speed Constraint (Response Time)**
   The system must generate predictions quickly to support real-time usage in a web application.

4. **Memory & Model Size Constraint**
   The trained model should be optimized to avoid excessive memory consumption during deployment.

5. **Scalability Constraint**
   The system should handle multiple inputs and future dataset expansion.

6. **Reproducibility Constraint**
   The deployed system must produce consistent outputs using saved preprocessing parameters.

How Constraints Were Addressed in Design

1. Accuracy Optimization

- Compared multiple regression models (Random Forest, XGBoost).
- Applied hyperparameter tuning using GridSearchCV.
- Used MAE, RMSE, and R^2 score for evaluation.
- Selected XGBoost due to superior performance.

2. Overfitting Control

- Used cross-validation during training.
- Applied regularization parameters in XGBoost.
- Performed cross-region and cross-season testing.

3. Speed Optimization

- Saved trained model using joblib.
- Optimized preprocessing pipeline.
- Reduced unnecessary computations inside Streamlit app.

4. Memory Management

- Removed redundant features.
- Compressed model size before deployment.
- Avoided storing raw dataset in deployed application.

5. Scalability Planning

- Designed modular pipeline using Scikit-learn Pipeline and ColumnTransformer.
- Ensured easy retraining with updated datasets.
- Structured GitHub repository for maintainability.

6. Reproducibility Assurance

- Saved preprocessing encoders and scalers.

- Ensured identical transformation during training and inference.

Test Results Around Constraints

- Achieved strong R² score indicating high predictive capability.
- Observed lower RMSE and MAE compared to baseline models.
- Cross-region testing showed minimal performance drop.
- Real-time inference achieved fast response within acceptable limits for web deployment.
- Model loaded successfully within memory limits of free-tier hosting platforms.

These results validate that the system satisfies industrial performance expectations.

Impact of Untested Constraints & Recommendations

Certain industrial constraints such as:

- Large-scale concurrent users
- Cloud-based distributed deployment
- Hardware-level computational efficiency

were not fully tested due to infrastructure limitations.

Recommendations for Future Handling:

- Deploy using scalable cloud infrastructure (AWS, Azure).
- Implement REST APIs for distributed access.
- Integrate database caching mechanisms.
- Monitor performance using logging frameworks.

With these enhancements, the system can evolve into a production-grade agricultural decision-support system.

6.2 Test Procedure

The following procedure was followed to evaluate system performance:

1. Split dataset into training and testing sets.

2. Applied preprocessing pipeline (Target Encoding + RobustScaler).
3. Trained baseline model (Random Forest).
4. Trained optimized model (XGBoost) using GridSearchCV.
5. Compared evaluation metrics on validation dataset.
6. Tested trained model on unseen regional and seasonal data.
7. Deployed model using Streamlit and measured inference time.
8. Logged prediction outputs and compared against actual values.

All preprocessing parameters were saved using joblib to ensure reproducibility during deployment.

6.3 Performance Outcome

The optimized XGBoost model demonstrated superior performance compared to baseline models.

Key outcomes:

- Achieved high R^2 score indicating strong predictive capability.
- Lower RMSE and MAE compared to Random Forest.
- Minimal performance drop during cross-region testing, confirming generalization.
- Fast inference time suitable for real-time prediction in web application.
- Stable behavior under multiple input scenarios.

The system successfully met industrial-level performance expectations in terms of accuracy, scalability, and usability.

7. MY LEARNINGS

The six-week internship provided me with comprehensive exposure to solving a real-world problem using Machine Learning and Data Science techniques. I gained practical understanding of the complete project lifecycle, starting from problem analysis and data preprocessing to model development, optimization, evaluation, and deployment.

Technically, I developed strong skills in:

- Exploratory Data Analysis (EDA) and data cleaning

- Feature engineering and handling high-cardinality categorical variables
- Model building using Random Forest and XGBoost
- Hyperparameter tuning using GridSearchCV
- Performance evaluation using MAE, RMSE, and R^2 metrics
- Model interpretability using SHAP
- Deployment of ML models using Streamlit
- Version control and structured project management using GitHub

Beyond technical expertise, this internship significantly enhanced my analytical thinking, structured problem-solving ability, time management, documentation practices, and presentation skills. Working with weekly milestones helped me develop discipline and project planning skills similar to industry standards.

This experience has strengthened my confidence in building end-to-end Machine Learning solutions and has prepared me for future roles in Data Science, Machine Learning Engineering, and AI-driven application development. It has bridged the gap between theoretical academic knowledge and practical industry implementation, contributing positively to my career growth.

8. FUTURE WORK SCOPE

Although the system performs effectively, there are several opportunities for further enhancement:

- Integration of real-time weather API data
- Incorporation of satellite imagery for precision agriculture
- Deep learning-based yield forecasting models
- Cloud deployment with REST API integration
- Development of a mobile application interface
- Integration with government agricultural advisory systems

Future improvements can transform this system into a fully scalable agricultural decision-support platform.