Industrial Internship Report on

”Prediction of Agricultural Crop Production in India”

Prepared by

Deepashree. K. A

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 is about predicting the agricultural crop production in India which is achieved through data analysis techniques and machine learning algorithms. 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

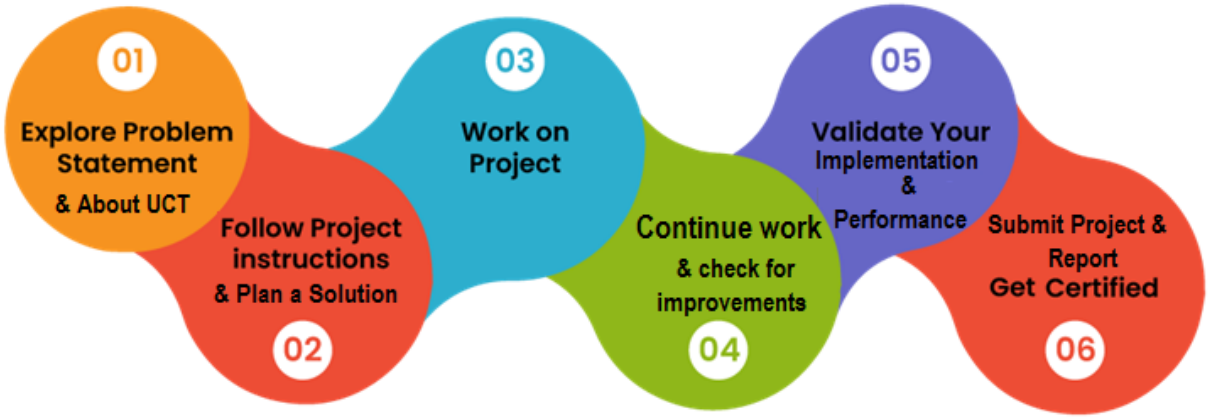
During the 1st week, I was introduced to the project and gained an understanding of the significance of predicting agricultural production. I familiarized myself with the project objectives and expectations. During the 2nd week and 3rd week, Data Collection and cleaning and preprocessing of the data to handle missing values and outliers was completed. During the 4th week combining the various datasets and eliminating null values was accomplished. During the 5th and 6th week Implemented and compared machine learning algorithms for predicting agricultural production as well as Conducted model training and evaluated the model's performance using appropriate metrics.

Relevant internships play a crucial role in career development for several reasons: Hands-on Experience: Internship provides practical, real-world experience that can't be gained solely through coursework or textbooks. They allow you to apply theoretical knowledge in a professional setting, bridging the gap between academia and the workplace. Skill Development: Internships offer the opportunity to develop and refine specific skills relevant to your chosen field. These skills can include technical skills, soft skills (communication, teamwork, problem-solving), and industry-specific knowledge. Career Insight: Internships provide insight into the day-to-day realities of a job or industry. This can help you make informed decisions about your career path and whether it aligns with your interests and values.Job Opportunities: In some cases, internships can lead directly to job offers. Many employers prefer to hire candidates who have interned with them because they are already familiar with the company culture and operations. Networking: Internships provide a chance to build a professional network.

Brief about Your project/problem statement: The problem is to develop a predictive model for agricultural production that utilizes data-driven approaches such as data science and machine learning. The aim is to accurately forecast crop production outcomes based on various input variables.

Opportunity given by USC/UCT.

How Program was planned



Your Learnings and overall experience:

Throughout this machine learning project, I gained valuable knowledge and insight. I learnt the importance of data quality during the data collection and cleaning phases. Selecting the right machine learning model deepened my understanding of model selection. Facing challenges during the model fitting and testing and training phase improved my problem solving skills. Working with Python and ML libraries increased my technical proficiency, and I see opportunities for further improvement and expansion in future work, such as incorporating more data sources and deploying models. Looking ahead, there is ample room for improvement and expansion of the project. Future work could involve incorporating additional data sources, exploring advanced model architectures, or deploying the model in a production environment.

A very sincere thanks to all (my teammate and upskill team), who have helped me directly or indirectly for the completion of this project.

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.

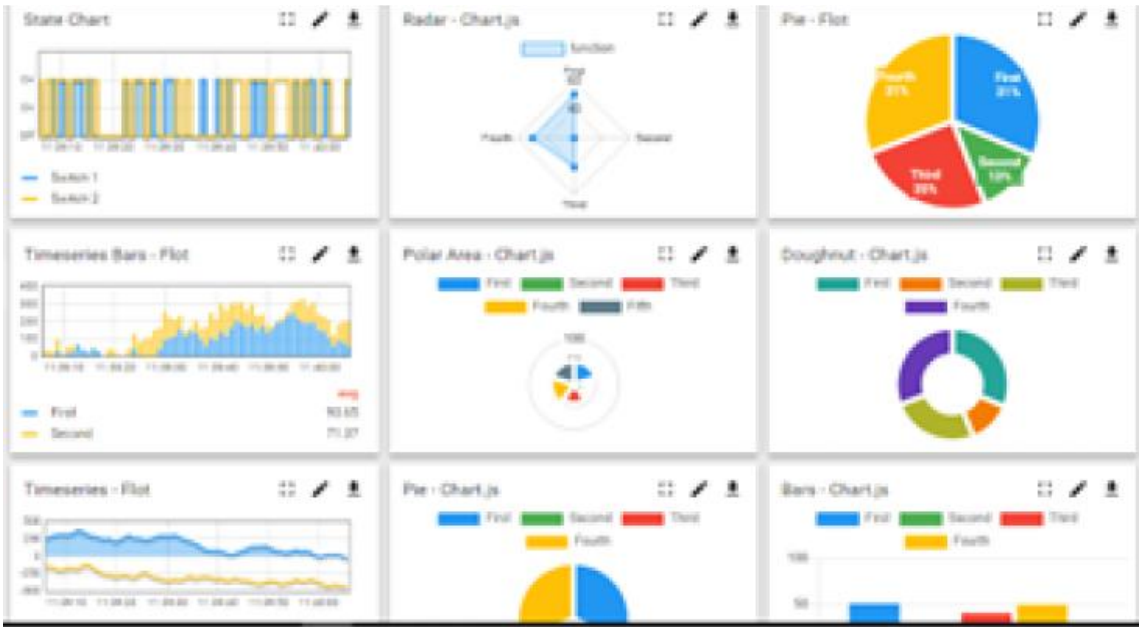
i. UCT IoT Platform (UCT Insight)

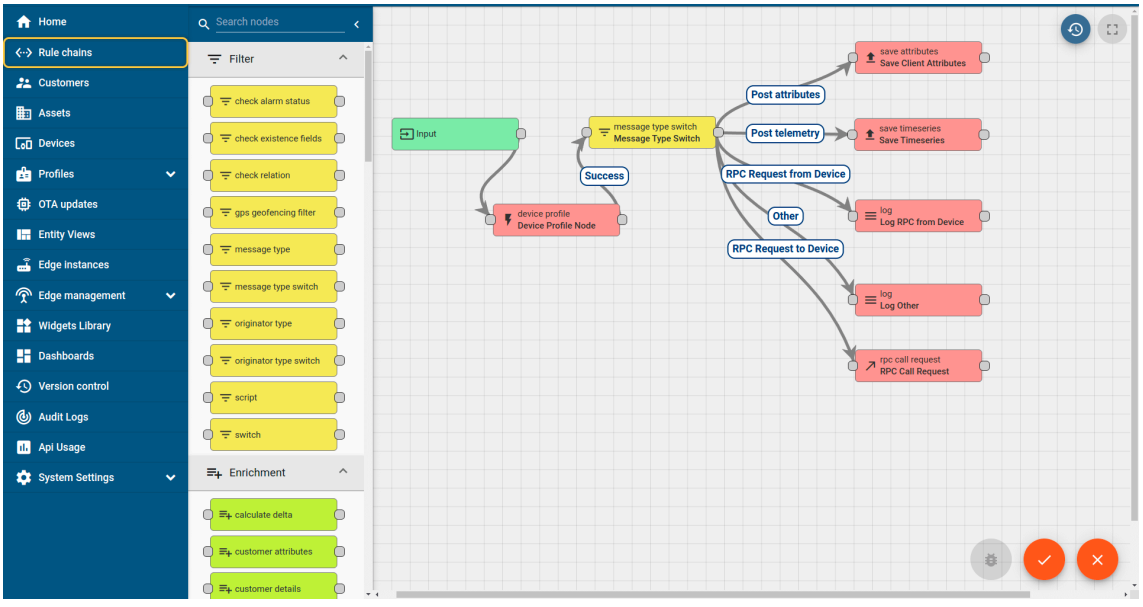
UCT Insight is an IOT platform designed for quick deployment of IOT applications on the same time providing valuable "insight" for your process/business. It has been built in Java for backend and ReactJS for Front end. It has support for MySQL and various NoSql Databases.

- 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.

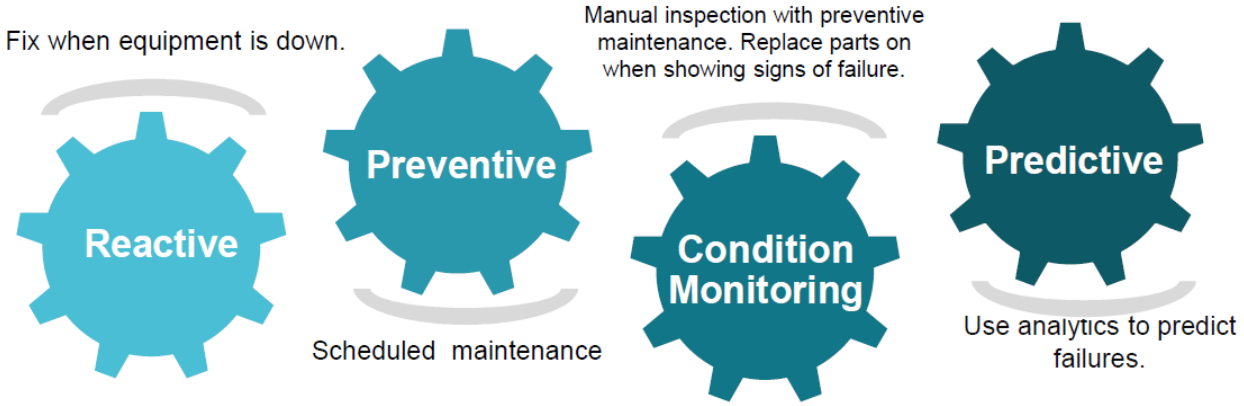
Machine	Operator	Work Order ID	Job ID	Job Performance	Job Progress		Output		Rejection	Time (mins)				Job Status	End Customer
					Start Time	End Time	Planned	Actual		Setup	Pred	Downtime	Idle		
CNC_S7_81	Operator 1	WO0405200001	4168	58%	10:30 AM		55	41	0	80	215	0	45	In Progress	i
CNC_S7_81	Operator 1	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 teschnology and providing solution in Agritech, Smart cities, Industrial Monitoring, Smart Street Light, Smart Water/ Gas/ Electricity metering solutions etc.

iv. Predictive Maintenance

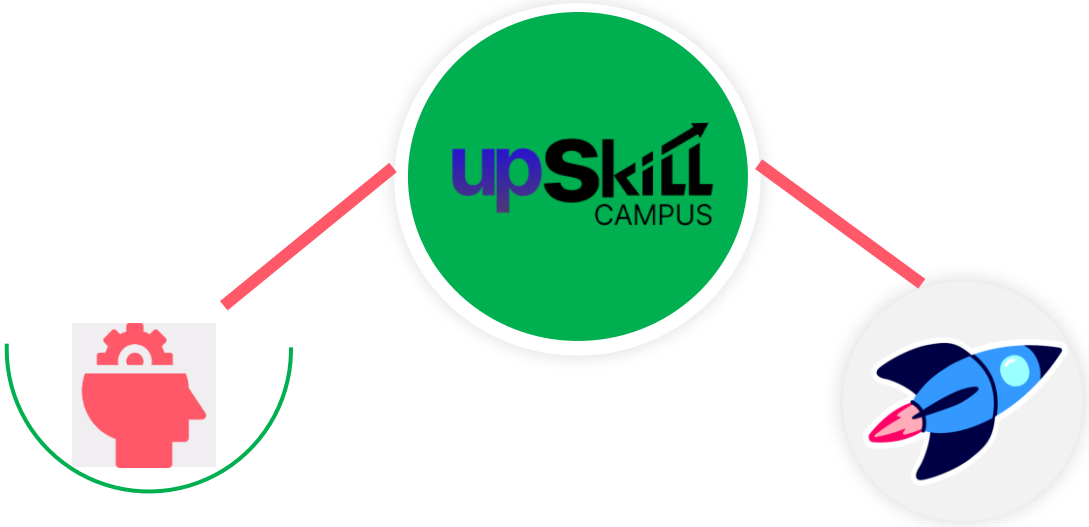
UCT is providing Industrial Machine health monitoring and Predictive maintenance solution leveraging Embedded system, Industrial IoT and Machine Learning Technologies by finding Remaining useful life time of various Machines used in production process.



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.

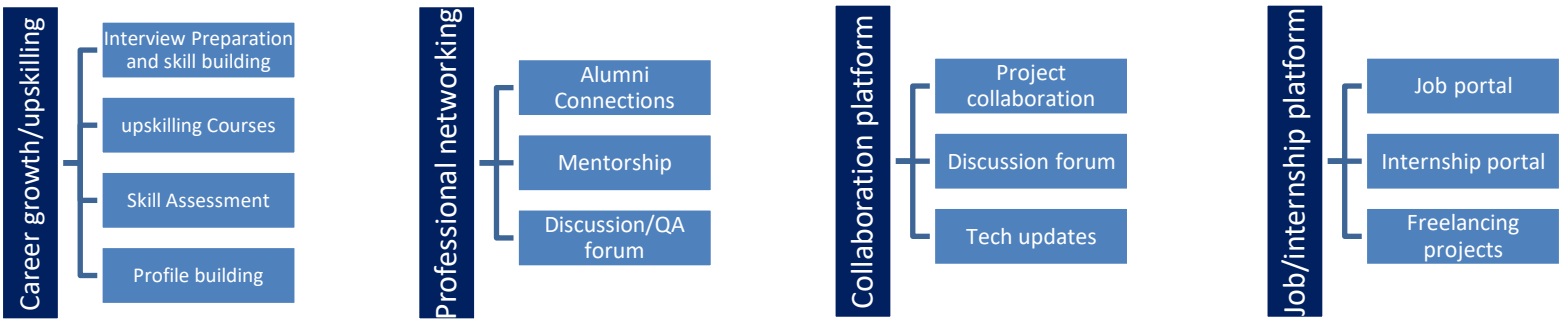
USC is a career development platform that delivers **personalized executive coaching** in a more affordable, scalable and measurable way.



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://github.com/ (for project submission).

[2] Google search and Chrome (information from various websites).

2.6 Glossary

Terms	Acronym
RMSE	Root Mean Squared Error
GUI	Graphical User Interface
Joblib	Joblib is a set of tools to provide lightweight pipelining in Python

3 Problem Statement

The aim of this project is to predict agricultural crop production in India for crops by analysing various factors such as the types of crop to be produced, region where produced, cost of production, etc provided in the data set. Most of the times farmers may not know cost of crop production for the particular crops in particular state. Using this prediction model, the prediction should be able to help educate farmers in making better decisions on the cost of production, growing their crops, using their resources more efficiently, boosting their agricultural productivity and make Farming more sustainable and efficient.

4 Existing and Proposed solution

Existing solutions and applications include:

Disease Detection: Computer vision and image recognition algorithms can identify plant diseases and pests based on images captured in the field. Early detection allows for timely intervention.

Weed Control: Machine learning-powered robots or drones equipped with cameras and sensors can distinguish between crops and weeds, enabling targeted weed control.

Soil Health Analysis: ML models can assess soil quality and nutrient levels, offering recommendations for optimal fertilization and soil management.

Irrigation Management: Smart irrigation systems use data from soil moisture sensors and weather forecasts to automate irrigation, conserving water and improving crop health.

There are various machine learning algorithms that can be used to implement the predictive model, some of which include

1. Linear regression: Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. The goal of the algorithm is to find the best linear equation that can predict the value of the dependent variable based on the independent variables. The equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable.
2. Random Forest: Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.
3. SVM (Support Vector Machine): Support Vector Machine (SVM) is a powerful machine learning algorithm used for linear or nonlinear classification, regression, and even outlier detection tasks. SVMs can be used for a variety of tasks, such as text classification, image classification, spam detection, handwriting identification, gene expression analysis, face detection, and anomaly detection. SVMs are adaptable and efficient in a variety of applications because they can manage high-dimensional data and nonlinear relationships.

4.1 Code submission (Github link):

https://github.com/Ka-Deepashree/upskillcampus

4.2 Report submission (Github link) :

https://github.com/Ka-Deepashree/upskillcampus

5 Proposed Design/ Model

5.1 High Level Diagram (if applicable)

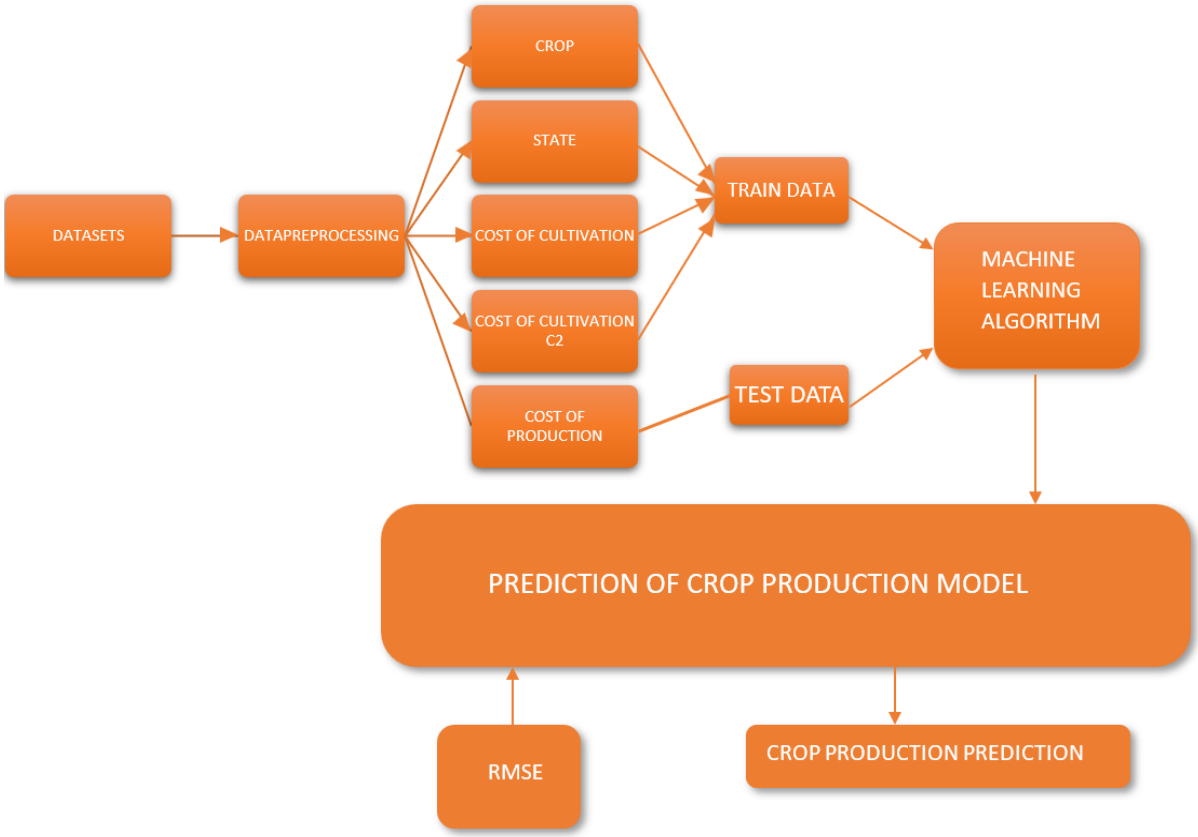


Figure 1: HIGH LEVEL DIAGRAM OF THE SYSTEM

5.2 Low Level Diagram (if applicable)

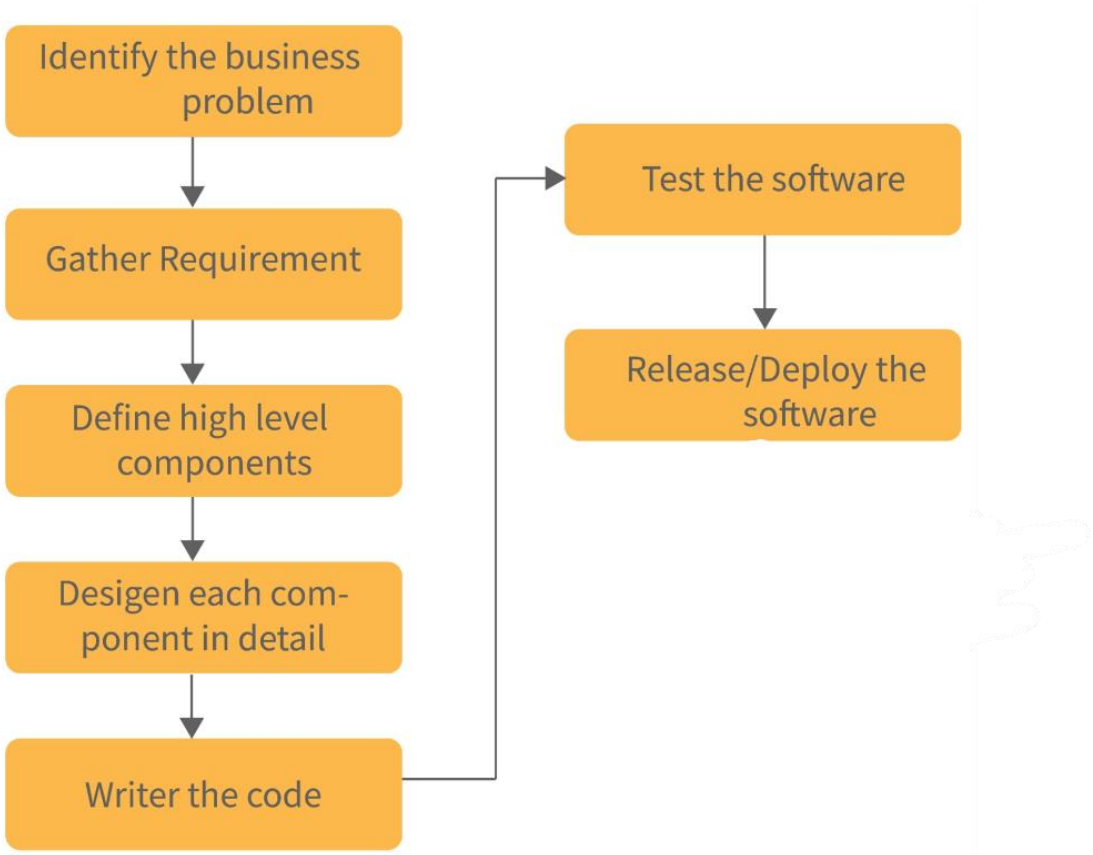


Figure 2: LOW LEVEL DIAGRAM OF THE SYSTEM

6 Performance Test

6.1 Test Plan/ Test Cases

TEST CASE 1: Ensuring Accurate Data preprocessing (removing or replacing null values, eliminating duplicate values, etc)

TEST CASE 2: Concatenating or merging the various datasets

TEST CASE 3: Label Encoding the values of string (or object) column i.e., mapping the unique values of a columns to unique numbers to avoid errors during model fitting or evaluation.

TEST CASE 4: Selecting the machine learning algorithm and model (comparing the ml models based on RMSE values)

TEST CASE 5: GUI

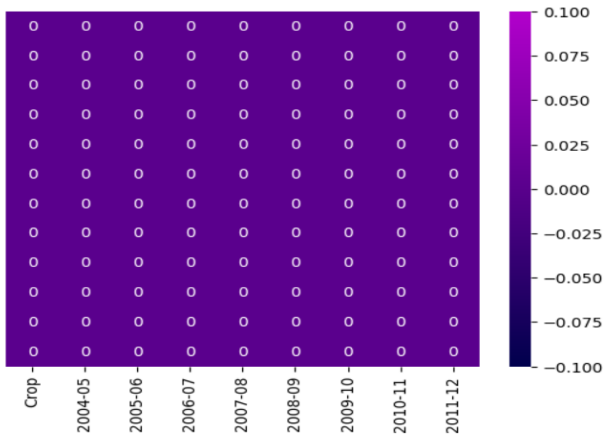
6.2 Test Procedure

LOADING THE DATAFILES AND DATA CLEANING AND PREPROCESSING

```
dataf_0=pd.read_csv("/content/datafile_0.csv")
dataf_1=pd.read_csv("/content/datafile_1.csv")
dataf_2=pd.read_csv("/content/datafile_2.csv")
dataf_3=pd.read_csv("/content/datafile_3.csv")
dataf_4=pd.read_csv("/content/datafile_4.csv")
```

```
sns.heatmap(dataf_0.isnull(), yticklabels=False, annot=True, cmap=custom_cmap)

# Show the plot
plt.show()
```




```
[10] dataf_0=dataf_0.dropna(how='all')
```

MERGING THE DATA:

```
datafile_list= sorted(glob('datafile_*.csv'))
datafile_list
```

```
[36] merged_data=pd.concat(pd.read_csv(datafile).assign(sourcefilename=datafile)for datafile in datafile_list)
     merged_data
```

```
[37] merged_data.to_csv('merged.csv')
```

```
[38] mdata=pd.read_csv("merged.csv")
     mdata
```

	Unnamed: 0	Crop	2004-05	2005-06	2006-07	2007-08	2008-09	2009-10	2010-11	2011-12	...	3-2005	3-2006	3-2007	3-2008	3-2009	3-2010	3-2011	3-2012
0	0	Rice	100.0	101.0	99.0	105.0	112.0	121.0	117.0	110.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	1	Wheat	100.0	101.0	112.0	115.0	117.0	127.0	120.0	108.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	2	Coarse Cereals	100.0	107.0	110.0	115.0	113.0	123.0	122.0	136.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	3	Pulses	100.0	108.0	134.0	124.0	124.0	146.0	137.0	129.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	4	Vegetables	100.0	109.0	103.0	118.0	113.0	124.0	128.0	115.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
619	424	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	778.0	669.0	807.0	1015.0	1031.0	541.0	1051.0	1080.0
620	425	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	1054.0	985.0	1951.0	1943.0	2424.0	2516.0	2476.0	3625.0
621	426	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	1577.0	1383.0	1404.0	1454.0	1550.0	1541.0	1563.0	1759.0
622	427	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	1222.0	1188.0	1318.0	1319.0	1280.0	1160.0	1328.0	1320.0
623	428	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	2443.0	2132.0	2589.0	2729.0	3319.0	3531.0	3700.0	3455.0

0s completed at 12:39 PM

LABEL ENCODING CROP NAME AND STATE NAME AND RETRIEVING UNIQUE NUMBER FOR EACH

```
[50]
     # Create a mapping of unique crop names to unique numbers
     crop_name_mapping = {crop: idx for idx, crop in enumerate(mdata['Crop'].unique())}

     # Replace crop names with unique numbers
     mdata['Crop'] = mdata['Crop'].map(crop_name_mapping)
```

```
[51]
     # Create a mapping of unique state names to unique numbers
     state_name_mapping = {state: idx for idx, state in enumerate(mdata['State'].unique())}

     # Replace state names with unique numbers
     mdata['State'] = mdata['State'].map(state_name_mapping)
```

```
[52] print(mdata['Crop'])
     print(mdata['State'])
```

```
0       0
1       1
2       2
3       3
4       4
       ..
619    49
620    49
621    49
622    49
623    49
Name: Crop, Length: 624, dtype: int64
0      0
```

Since user will not have knowledge of unique number for each unique crop name and state name (encoded to avoid error ValueError during prediction), reverse mapping of crop name and state name provides the user the number assigned to the entered crop and state

```
#reverse mapping dictionary 'crop_name_mapping'
# Enter the crop name you want to convert to the unique number
crop_name = input('Enter the crop')  # Replace with the crop name you want to look up

# Use the reverse mapping dictionary to get the unique number from the crop name
unique_number = crop_name_mapping.get(crop_name, -1)  # Default to -1 if not found

if unique_number != -1:
    print("Unique Number for Crop Name:", unique_number)
else:
    print("Crop Name not found ")
```

```
Enter the cropRice
Unique Number for Crop Name: 0
```

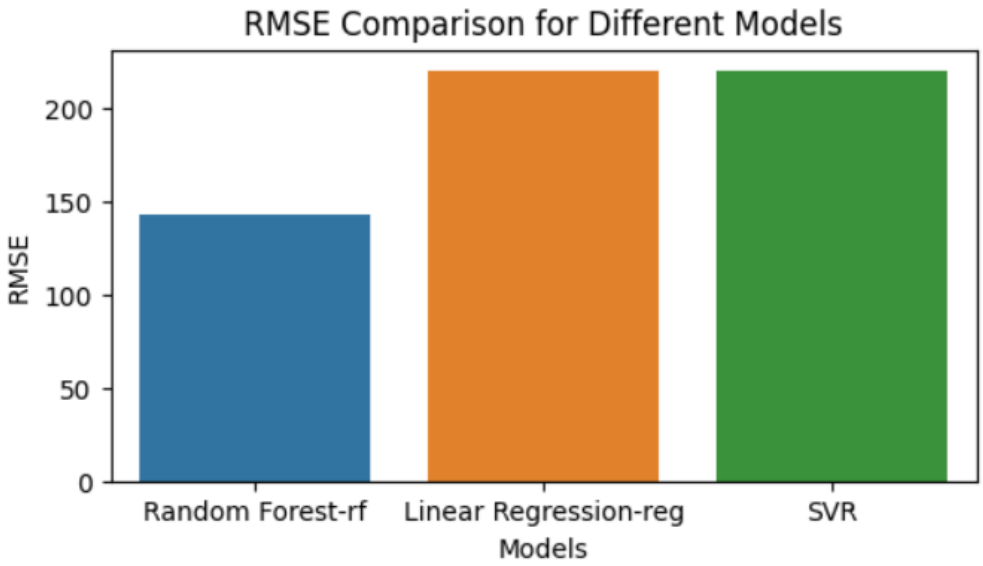
SPLITTING DATA INTO TRAINING SET AND TESTING SET

```
[ ] X = mdata[['Crop','State','Cost of Cultivation (`/Hectare) A2+FL', 'Cost of Cultivation (`/Hectare) C2',]]
    Y = mdata["Cost of Production (`/Quintal) C2"]
```

```
[ ] print("Shape of X:", X.shape)
    print("Shape of Y:", Y.shape)
```

```
[ ] X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=42)
```

6.3 Performance Outcome



USING RANDOM FOREST REGRESSOR

```
[155] # Use the model to make predictions on new data
      a=input("Enter the crop")
      b = input("Enter the State\n")
      d = input("Enter the cultivation cost per hectare\n")
      e = input("Enter the cultivation cost per hectare in c2\n")
      new_data = pd.DataFrame({'Crop': [a],'State': [b], 'Cost of Cultivation (`/Hectare) A2+FL': [d], 'Cost of Cultivation (`/Hectare) C2': [e]})
      areaYield = rf.predict(new_data)
      print("The Cost of production is ", areaYield)
```

```
Enter the crop12
Enter the State
1
Enter the cultivation cost per hectare
10593.15
Enter the cultivation cost per hectare in c2
16528.68
The Cost of production is  [2169.9346]
```

Predicted cost of production = 2169.9346

Actual cost of production=2172.46

SAVE MODEL USING JOBLIB

```
[244] import joblib
```

```
[245] joblib.dump(rf,'model_crop')

      ['model_crop']
```

```
[246] model=joblib.load('model_crop')
```

```
[247] model.predict(new_data)

      array([2169.9346])
```

7 My learnings

Throughout this machine learning project, I gained valuable knowledge and insight. I learnt the importance of data quality during the data collection and cleaning phases. Selecting the right machine learning model deepened my understanding of model selection. Facing challenges during the model fitting and testing and training phase improved my problem solving skills. Working with Python and ML libraries increased my technical proficiency, and I see opportunities for further improvement and expansion in future work, such as incorporating more data sources and deploying models. Looking ahead, there is ample room for improvement and expansion of the project. Future work could involve incorporating additional data sources, exploring advanced model architectures, or deploying the model in a production environment.

8 Future work scope

One idea that couldn't be included during this project is creating GUI (Graphical User Interface) which provides visuals that explain the results obtained from algorithms. Predicting optimal planting times, soil health and using more data sets to narrow the accuracy of prediction could also have been implemented.

Additionally we can use image processing to detect the type of crop, or weeds, plant diseases, the solution to eradicate the diseases or pests, weeds using safe and biological toxic free medicines, suggest the type of fertilizers best for the crops. These techniques can further improve the production of crop and result in greater yield.