Industrial Internship Report on

Prediction of Agriculture Crop Production

Prepared by

Vikash Baabhu

Indian Institute of Information Technology, Tiruchirappalli

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 prediction of agriculture crop production.

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 Introduction 3

2 Problem Statement 4

3 Methodology 5

3.1 Data Collection 5

3.2 Data Preprocessing 6

3.3 Feature Selection 7

3.4 Data Visualization 8

3.5 Splitting the Dataset 10

3.6 Model Selection 11

3.7 Model Training 12

3.8 Prediction 13

4 Results and Analysis 14

5 Challenges and Hurdles 16

6 My learnings 16

7 Future work scope 17

1. Introduction

- The Prediction of Agriculture Crop Production project utilizes data analytics and machine learning to forecast crop yields and improve farming efficiency.
- Predicting crop production using machine learning is a valuable application in agriculture.
- By analyzing historical and real-time data, such as crop production, cost of cultivation and other relevant factors, we develop predictive models that offer valuable insights to future crop yields.
- This project aims to empower farmers with accurate predictions, personalized recommendations, and a user-friendly interface, enabling them to make informed decisions about crop selection, resource allocation, and risk management.

2. Problem Statement

- The agricultural sector lacks accurate predictive models and timely information, hindering farmers' ability to optimize crop production, make informed decisions, and adopt sustainable farming practices.

- This leads to lower yields, increased costs, and potential losses due to unpredictable weather, diseases, and pests.

- A comprehensive solution is needed that utilizes data analytics and machine learning to provide accurate predictions and personalized recommendations to farmers, empowering them to enhance productivity and reduce risks.

3. Methodology

3.1 Data Collection

Gather historical data on crop yields, crop production, cost of cultivation and other relevant factors. This data will be used to train and evaluate the machine learning model.

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

1. Take the data and create dataframe

```
[ ]  df1 = pd.read_csv("/content/datafile (1).csv")
     df2 = pd.read_csv("/content/datafile (2).csv")
     df3 = pd.read_csv("/content/datafile (3).csv")
```

3.2 Data Preprocessing

Clean the collected data by removing any outliers, handling missing values, and normalizing or scaling the data as necessary. Ensure that all variables are in a suitable format for analysis.

```
[ ] df3.head()
```

	Crop	Variety	Season/ duration in days	Recommended Zone	Unnamed: 4
0	Paddy	Chinsurah Rice (IET 19140)	Medium	Andhra Pradesh, Tamil Nadu, Gujarat, Orissa, a...	NaN
1	Paddy	(CNI 383-5-11)	NaN	NaN	NaN
2	Paddy	IGKVR-1 (IET 19569)	Mid-early	Chhattisgarh, Madhya Pradesh and Orissa under ...	NaN
3	Paddy	IGKVR-2 (IET 19795)	Medium	Chhattisgarh, Bihar and Orissa under both irr...	NaN
4	Paddy	CR Dhan 401 (REETA)	145-150	Orissa, West Bengal, Tamil Nadu and Andhra Pra...	NaN

```
[ ] df3.drop('Unnamed: 4',axis=1,inplace=True)
    df3.dropna(inplace=True)
```

```
[ ] df3
```

	Crop	Variety	Season/ duration in days	Recommended Zone
0	Paddy	Chinsurah Rice (IET 19140)	Medium	Andhra Pradesh, Tamil Nadu, Gujarat, Orissa, a...
2	Paddy	IGKVR-1 (IET 19569)	Mid-early	Chhattisgarh, Madhya Pradesh and Orissa under ...
3	Paddy	IGKVR-2 (IET 19795)	Medium	Chhattisgarh, Bihar and Orissa under both irr...
4	Paddy	CR Dhan 401 (REETA)	145-150	Orissa, West Bengal, Tamil Nadu and Andhra Pra...

3.3 Feature Selection

Identify the most important features that contribute to crop production. This step involves analyzing the correlations between different variables and selecting those that have the highest impact.

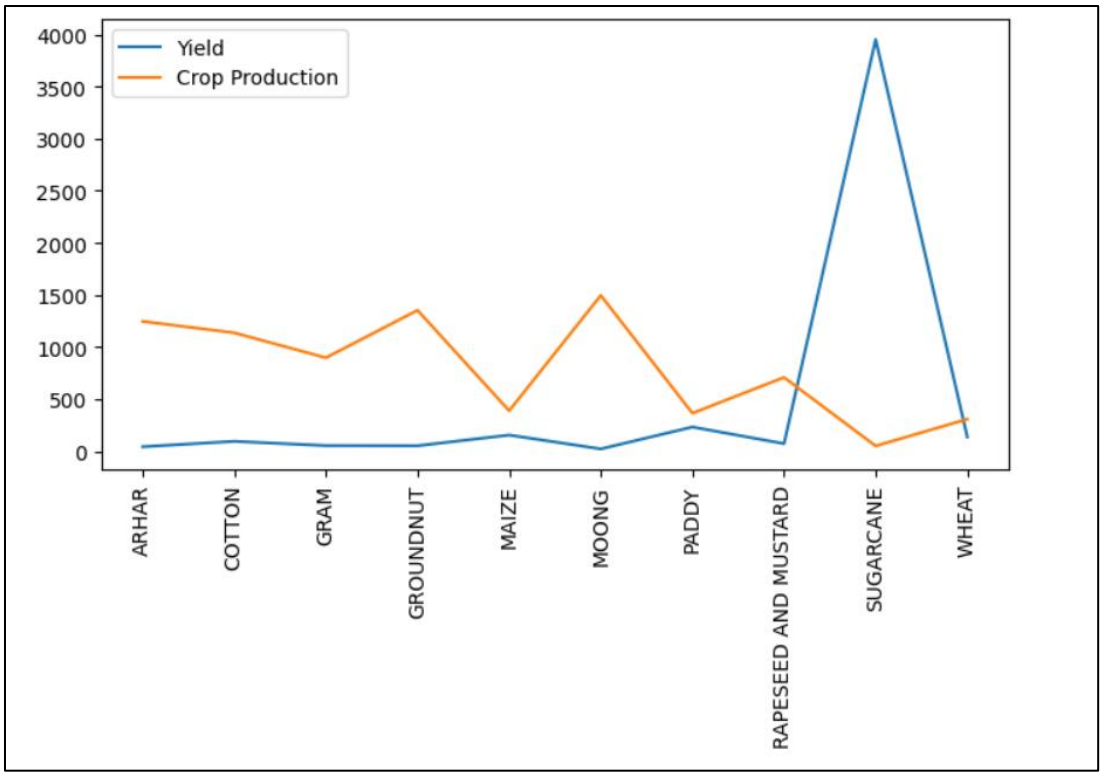
```
[ ]  x = df1.iloc[:,0:5].values

[ ]  y = df1.iloc[:,5].values
     y

     array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0], dtype=uint8)
```

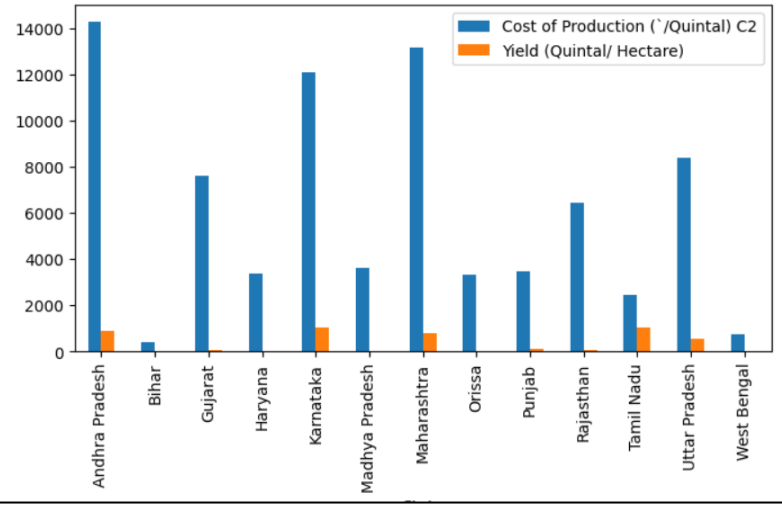
3.4 Data Visualization

Understanding and communicating insights from data. It allows you to represent complex information visually, making it easier to interpret and identify patterns, trends, and relationships within the data.



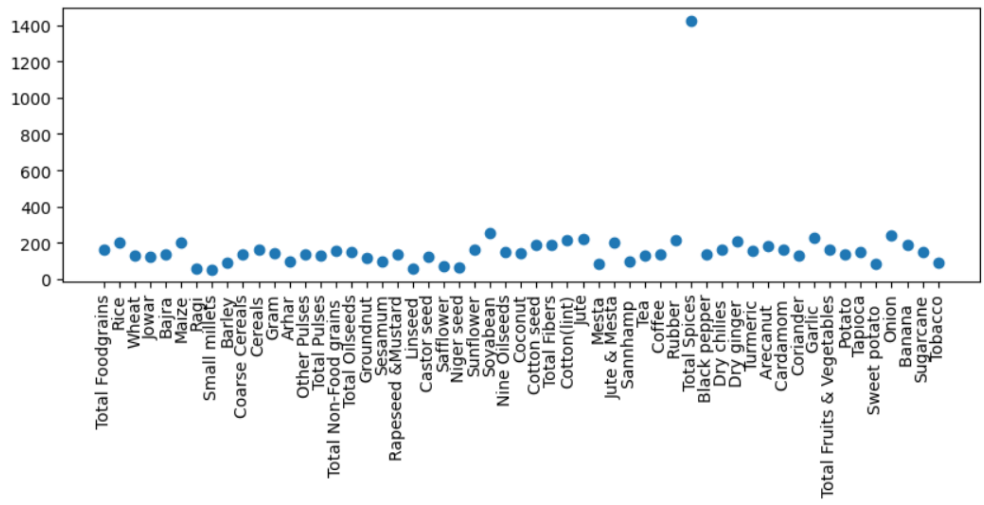
```
state_yield = df1.groupby(['State'])
index = list(state_yield.indices.keys())
state_yield.sum()[['Cost of Production (`/Quintal) C2', 'Yield (Quintal/ Hectare) ']].plot(kind='bar',figsize=(8,4))
```

```
<ipython-input-14-c3d619ff185f>:5: FutureWarning: The default value of numeric_only in DataFrameGroupBy.sum is deprecated.
  state_yield.sum()[['Cost of Production (`/Quintal) C2', 'Yield (Quintal/ Hectare) ']].plot(kind='bar',figsize=(8,4))
<Axes: xlabel='State'>
```




```
[ ] # Visualizing the production of various crops in the year 2006-07

    plt.subplots(figsize=(10,3))
    plt.scatter(x='Crop',y='Production 2006-07',data = df2)
    plt.xticks(rotation=90)
    plt.show()
```



3.5 Splitting the dataset

Divide the dataset into training and testing sets. The training set will be used to train the machine learning model, while the testing set will be used to evaluate its performance.

```
[ ]  from sklearn.model_selection import train_test_split
     x_train,x_test,y_train,y_test = train_test_split(x,y,random_state = 0)
```

```
[ ]  print(x.shape)
     print(x_train.shape)
     print(x_test.shape)
```

```
     (49, 5)
     (36, 5)
     (13, 5)
```

```
[ ]  print(y.shape)
     print(y_train.shape)
     print(y_test.shape)
```

```
     (49,)
     (36,)
     (13,)
```

3.6 Model selection

Choose an appropriate machine learning algorithm for your prediction task. Some common algorithms used for crop production prediction include linear regression, decision trees, random forests.

```
[ ]  from sklearn.linear_model import LinearRegression
     model = LinearRegression()
```

3.7 Model training

Train the selected model using the training dataset. The model will learn the patterns and relationships between the input features and the target variable (crop yield).

```
[ ]  model.fit(x_train,y_train)
```

```
▾ LinearRegression
LinearRegression()
```

3.8 Prediction

Once the model is trained, use it to make predictions on new, unseen data. You can input the type of crop, crop production and other relevant variables to forecast crop yields for future seasons.

```
[ ]  y_pred = model.predict(x_test)
     y_pred

     array([ -76.13208609,   84.05677448,  158.7158742 ,  726.61270001,
              37.41996066, -249.96481962, -127.60574732, -150.29125906,
              87.74800299,   45.94284648,  -54.91120403,  -58.1689958 ,
             282.23546753])
```

4. Results and Analysis

Here I predicted the model using Linear Regression, and Random Forest Regressor and achieved highest accuracy of 92 % using Random Forest Regressor.

```
[35] from sklearn.metrics import mean_squared_error , mean_absolute_error , r2_score

     Model_perf = pd.DataFrame(columns=['Model_Name','MSE','R2_Score'])

     LR_mse = mean_squared_error(y_test,y_pred)
     LR_R2 = r2_score(y_test,y_pred)

     new_row = {'Model_Name':'Linear Regression','MSE':LR_mse , 'R2_Score': LR_R2}
     Model_perf = Model_perf.append(new_row,ignore_index = True)
     Model_perf

<ipython-input-35-330479dbb3b6>:9: FutureWarning: The frame.append method is deprecated and
  Model_perf = Model_perf.append(new_row,ignore_index = True)
```

	Model_Name	MSE	R2_Score
0	Linear Regression	21422.851797	0.698323

```
[36] from sklearn.ensemble import RandomForestRegressor
     rf = RandomForestRegressor()
     rf.fit(x_train,y_train)

     y_pred = rf.predict(x_test)
     mse = mean_squared_error(y_test,y_pred)
     r2 = r2_score(y_test,y_pred)

     new_row = {'Model_Name':'Random Forest Regressor','MSE':mse , 'R2_Score': r2}
     Model_perf = Model_perf.append(new_row,ignore_index = True)
     Model_perf
```

```
<ipython-input-36-a210aa37c4df>:10: FutureWarning: The frame.append method is deprecated
  Model_perf = Model_perf.append(new_row,ignore_index = True)
```

	Model_Name	MSE	R2_Score
0	Linear Regression	21422.851797	0.698323
1	Random Forest Regressor	5296.585662	0.925413

GitHub Link :-

Vikashbaabhu/Agriculture-Crop-Production (github.com)

5. Challenges and Hurdles

- The first challenge that I faced was I couldn't merge the dataset, that is I don't how to group the given data.
- Second, I don't know how to convert the object datatype (String) to integer for performing the linear regression.
- I faced many other hurdles also but it was easy for me to rectify, but these two are the major hurdles that I faced.

6. My Learnings

- First, I referred many websites like GeeksforGeeks and YouTube and learned how to merge the dataset and how to group the data.
- Next, I referred some eBooks and many websites like stack overflow for converting the object datatype (String) to integer.
- These were the new skills that I learned during this internship.

7. Future work scope

- ✓ Extend the project's reach to encompass a wider range of geographical regions, considering the diverse climate conditions, soil types, and crop preferences.
- ✓ Explore the integration of emerging technologies such as remote sensing, drones, and precision agriculture tools.
- ✓ Incorporate disease and pest management modules into the predictive models.
- ✓ Expand the decision support system to include optimization modules for resource allocation.
- ✓ Implement a feedback loop mechanism that collects user feedback and incorporates it into model improvement processes.
- ✓ Develop educational materials and training programs to empower farmers with the necessary knowledge and skills to effectively utilize the predictive models and make data-driven decisions.