# Profitability and Yield prediction on agricultural crops of India

## 1. INTRODUCTION

Across the globe India is the second largest country having a population of more than 1.3 Billion. Many people are dependent on agriculture but the sector lacks efficiency and technology especially in our country. By bridging the gap between traditional agriculture and data science, effective crop cultivation can be achieved.

In developing countries, farming is considered as the major source of revenue for many people. In modern years, agricultural growth is engaged by several innovations, environments, techniques and civilizations. In addition, the utilization of information technology may change the condition of decision making and thus farmers may yield the best way. For the decision making process, data mining techniques related to agriculture are used. Data Mining is the process of analyzing, extracting and predicting the meaningful information from huge data to extract some pattern. This process is used by companies to turn the raw data of their customer to useful information. The process of Data Mining includes first selection of data followed by pre- processing of data and then transforming the data to get patterns which can then be used to predict useful insights. Preprocessing includes finding outliers and detecting missing values whereas transformation finds the correlation between objects. Applying the data mining techniques on historical climate and crop production data several predictions can be made on the basis of knowledge gathered which in turn can help in increasing crop productivity.

## 2. MOTIVATION

Agriculture is the most important sector that influences the economy of India. It contributes to 18% of India‘s Gross Domestic Product (GDP) and gives employment to 50% of the population of India. People of India have been practicing Agriculture for years but the results are never satisfying due to various factors that affect the crop yield [3]. To fulfill the needs of around 1.2 billion people, it is very important to have a good yield of crops. Due to factors like soil type, precipitation, seed quality, lack of technical facilities etc. the crop yield is directly influenced. We focus on implementing crop yield prediction systems by using Machine learning techniques by doing analysis on agriculture dataset. For evaluating performance Accuracy is used as one of the factors. The classifiers are further compared with the values of Precision, Recall and F1 Score. Lesser the value of error, more accurate the algorithm will work. The result is based on comparison among the classifiers.

### 2.1 Scope

The scope of this project is to investigate a dataset of crop records for the agricultural sector using machine learning techniques. Identifying crop predictions by farmers is more difficult. We try to reduce this risk factor behind selection of the crop.

### 2.2. Objectives

**Objectives**

1. Data Preprocessing
2. Data Visualization
3. Using various algorithms and comparing the accuracy

## 3. RELATED WORK

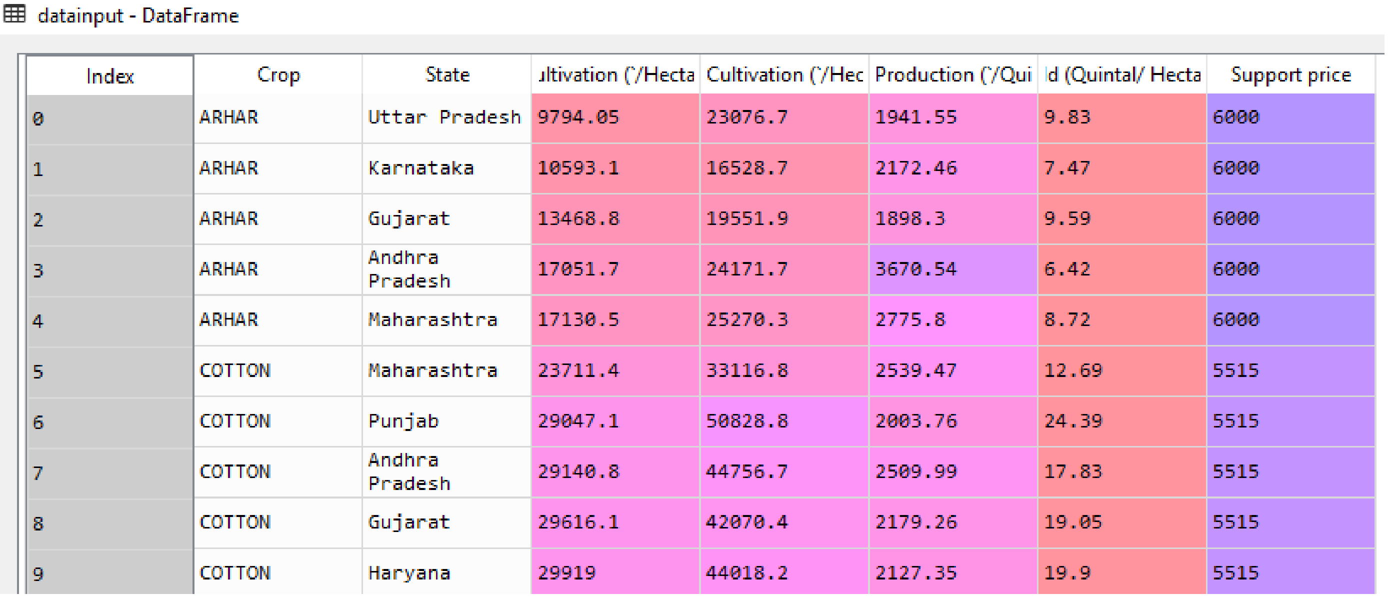
In agriculture, Machine Learning is considered as a novel field, as a variety of work has been done with the help of machine learning in the field of agriculture. There are different philosophies made and evaluated by the researchers all through the world in the field of agriculture and related sciences.

CH. Vishnu VardhanChowdary, Dr.K.Venkataramana [5], developed id3 algorithm for getting improved and great quality of crop yield of Tomato and is executed in Php platform and datasets are used as csv. Temperature, area, humidity and the production of tomato crop are the different parameters used in this study.R. Sujatha and P. Isakki [6], utilizes data mining techniques for prediction. This model worked on different parameters such as crop name, land area, soil type, pH value, seed type, water and also foreseen the boom and diseases of plants and in this way empowered to choose the descent crop dependent on climatic data and required parameters.N. Gandhi, L. J. Armstrong, O. Petkar and A. K. Tripathy [7], proposed the SVM for crop yield prediction of rice. In this method, the dataset used consists of different parameters such as place, temperature, precipitation and manufacturing. On this dataset, the implemented classifier is sequential minimal optimization.

## 4. DATASET

We have considered 2 datasets. One finds out the profit and classifies it if there is profit or loss. The second dataset predicts the production.

### 4.1 Dataset 1.



datainput - DataFrame

| Index | Crop | State | ıltivation (`/Hecta | Cultivation (`/Hec | Production (`/Qui | ld (Quintal/ Hecta | Support price |
|---|---|---|---|---|---|---|---|
| 0 | ARHAR | Uttar Pradesh | 9794.05 | 23076.7 | 1941.55 | 9.83 | 6000 |
| 1 | ARHAR | Karnataka | 10593.1 | 16528.7 | 2172.46 | 7.47 | 6000 |
| 2 | ARHAR | Gujarat | 13468.8 | 19551.9 | 1898.3 | 9.59 | 6000 |
| 3 | ARHAR | Andhra Pradesh | 17051.7 | 24171.7 | 3670.54 | 6.42 | 6000 |
| 4 | ARHAR | Maharashtra | 17130.5 | 25270.3 | 2775.8 | 8.72 | 6000 |
| 5 | COTTON | Maharashtra | 23711.4 | 33116.8 | 2539.47 | 12.69 | 5515 |
| 6 | COTTON | Punjab | 29047.1 | 50828.8 | 2003.76 | 24.39 | 5515 |
| 7 | COTTON | Andhra Pradesh | 29140.8 | 44756.7 | 2509.99 | 17.83 | 5515 |
| 8 | COTTON | Gujarat | 29616.1 | 42070.4 | 2179.26 | 19.05 | 5515 |
| 9 | COTTON | Haryana | 29919 | 44018.2 | 2127.35 | 19.9 | 5515 |

We combined data from different sources.
The data contains columns:

- Crops
- State
- Cost of Cultivation (`/Hectare) A2+FL
- Cost of Cultivation (`/Hectare) C2
- Cost of Production (`/Quintal) C2

- Yield produced.

The profit for each row was calculated using the formula

C1 -> Cost of cultivation(`/Hectare) A2+FL
C2 -> Cost of Cultivation (`/Hectare) C2
Cp -> Cost of Production (`/Quintal)
***Profit = (Yield *Support Price) - (C1 + C2 + (Yield*Cp))***

The govt. fixes support prices[2] per Quintal for various commodities, for example various Kharif and Rabi crops.
If the yield produced will result in profit based on support prices declared by the government, class 1 was allotted; else it was classified as class 0.

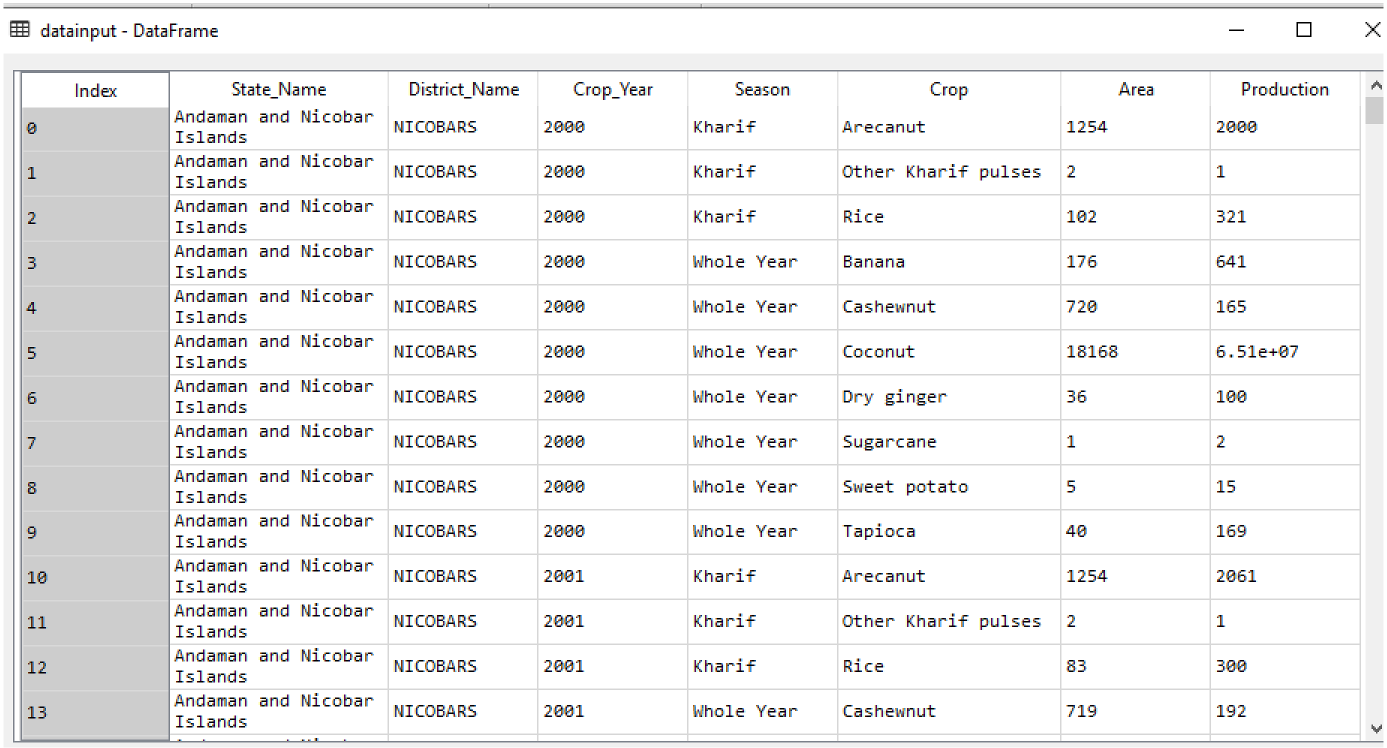
**Advantages**

- This dataset is compiled by using data from an official government site which proves its authenticity.
- Farmers can directly find out if the crop they are about to sow will result in profit after cultivation

**Disadvantage**

- Does not have many instances

## 4.2 Dataset 2



| Index | State_Name | District_Name | Crop_Year | Season | Crop | Area | Production |
|---|---|---|---|---|---|---|---|
| 0 | Andaman and Nicobar Islands | NICOBARS | 2000 | Kharif | Arecanut | 1254 | 2000 |
| 1 | Andaman and Nicobar Islands | NICOBARS | 2000 | Kharif | Other Kharif pulses | 2 | 1 |
| 2 | Andaman and Nicobar Islands | NICOBARS | 2000 | Kharif | Rice | 102 | 321 |
| 3 | Andaman and Nicobar Islands | NICOBARS | 2000 | Whole Year | Banana | 176 | 641 |
| 4 | Andaman and Nicobar Islands | NICOBARS | 2000 | Whole Year | Cashewnut | 720 | 165 |
| 5 | Andaman and Nicobar Islands | NICOBARS | 2000 | Whole Year | Coconut | 18168 | 6.51e+07 |
| 6 | Andaman and Nicobar Islands | NICOBARS | 2000 | Whole Year | Dry ginger | 36 | 100 |
| 7 | Andaman and Nicobar Islands | NICOBARS | 2000 | Whole Year | Sugarcane | 1 | 2 |
| 8 | Andaman and Nicobar Islands | NICOBARS | 2000 | Whole Year | Sweet potato | 5 | 15 |
| 9 | Andaman and Nicobar Islands | NICOBARS | 2000 | Whole Year | Tapioca | 40 | 169 |
| 10 | Andaman and Nicobar Islands | NICOBARS | 2001 | Kharif | Arecanut | 1254 | 2061 |
| 11 | Andaman and Nicobar Islands | NICOBARS | 2001 | Kharif | Other Kharif pulses | 2 | 1 |
| 12 | Andaman and Nicobar Islands | NICOBARS | 2001 | Kharif | Rice | 83 | 300 |
| 13 | Andaman and Nicobar Islands | NICOBARS | 2001 | Whole Year | Cashewnut | 719 | 192 |

In the second dataset we have the following columns:

- State_Name
- District_Name
- Crop_Year
- Season
- Crop
- Area
- Production.

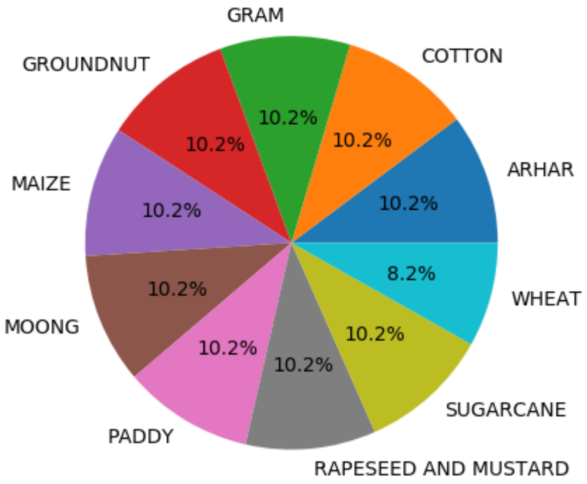
We will be predicting the production of the crops using regressors.

**Advantages:**

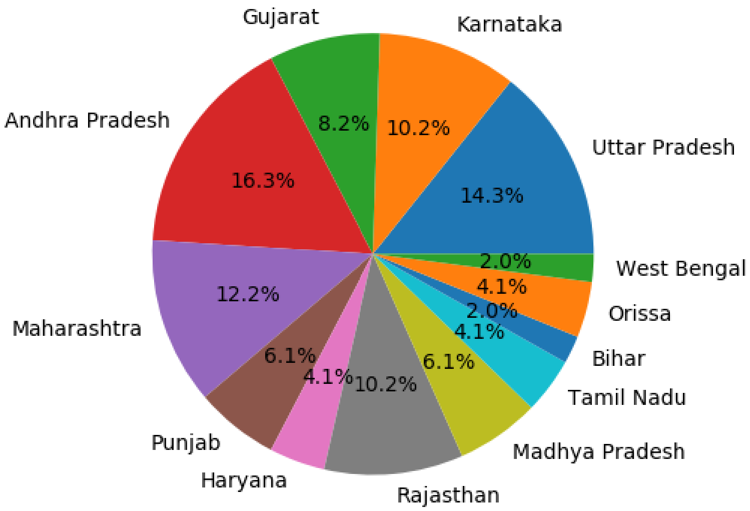
- Huge dataset about 2 lakh entries
- Takes season as well as crop year into consideration

**Disadvantages**:

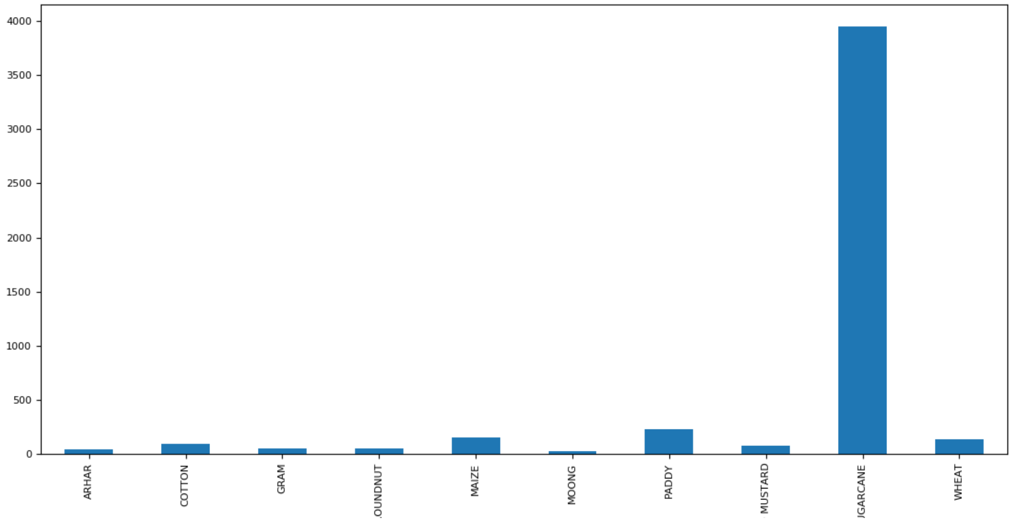
- Many missing values
- Rainfall and temperature are not considered.
- Numerous Categories: The categorical variables have many values for each attribute.



*Crop-wise distribution in percentage*



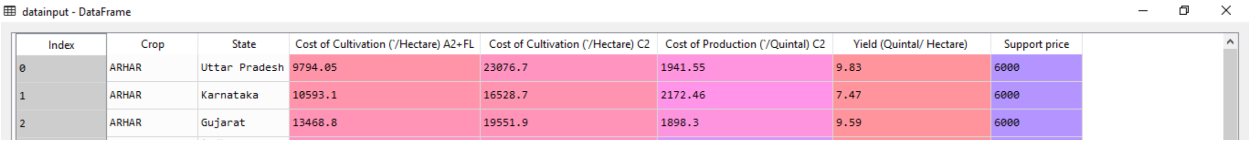
*State wise distribution in dataset*



*Total Yield crop wise*

## 5. DATA-PREPROCESSING

After adding the support price column and profit in our dataset and labelling them as 0 and 1, preprocessing techniques were applied such as missing values. The crops and state columns were encoded using labels and the one hot encoder was applied so as to avoid ranking.



datainput - DataFrame

| Index | Crop | State | Cost of Cultivation (`/Hectare) A2+FL | Cost of Cultivation (`/Hectare) C2 | Cost of Production (`/Quintal) C2 | Yield (Quintal/ Hectare) | Support price |
|---|---|---|---|---|---|---|---|
| 0 | ARHAR | Uttar Pradesh | 9794.05 | 23076.7 | 1941.55 | 9.83 | 6000 |
| 1 | ARHAR | Karnataka | 10593.1 | 16528.7 | 2172.46 | 7.47 | 6000 |
| 2 | ARHAR | Gujarat | 13468.8 | 19551.9 | 1898.3 | 9.59 | 6000 |

*A snapshot of Dataframe*

Dataset 2 contains many missing values. During the preprocessing step these rows are dropped since the number of instances is very large.

# 6. PROPOSED SOLUTION

In the system, we propose tests of many algorithms and by studying the classification report we compare the algorithms and choose the best one. It has to find accuracy of the training dataset, accuracy of the testing dataset, specification, False Positive rate, precision and recall by comparing algorithms using python code.

The following Involvement steps are :

1. Define a problem
2. Preparing data
3. Evaluating algorithms
4. Improving results
5. Predicting results

We will be applying classification algorithms on dataset 1 and regression for prediction of Production on dataset 2.

Algorithms applied:

1) Classification:
    a) Decision Tree
    b) Logistic Regression
    c) K nearest neighbour
    d) Random forest Classifier
2) Clustering
3) Regression:
    a) Decision Tree
    b) Random forest

## 7. PERFORMANCE EVALUATION

Let us first briefly understand some of the performance evaluation metrics:

### 7.1. General Definitions

1. True Positive (TP) depicts the number of instances where the system detects for a condition when it is really present.

2. True Negative (TN) depicts the number of instances where the system does not detect a condition when it is absent.

Observations to the total predicted positive observations. Low false positive rate means high precision. In this research the precision 0.788 is obtained which is pretty good.

### 7.2. Recall

Positive observed values proportion is correctly predicted. (Actual defaulter‘s model will correctly predict the proportion)

Recall = TP / (TP + FN)

Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class -yes.

### 7.3. F1 Score

F1 score is the process of finding the calculated weighted average of Precision and Recall. The score is considered for both false positives and false negatives. Intuitively it is not easy to understand accuracy, but F1 is usually more useful than accuracy, especially if uneven class distribution is considered. Accuracy is the best way, if false positives and false negatives have similar cost. To better look at the precision and recall, the cost of false positives and false negatives should be very different.

General Formula:

F- Measure = 2TP / (2TP + FP + FN)

F1-Score Formula:

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

### 7.4 Precision

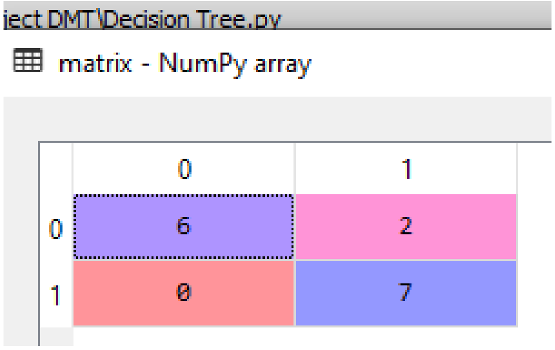
When the model predicts positive, how often is it correct?

Precision = TP/TP+FP

Precision helps when the costs of false positives are high. So let's assume the problem involves the detection of skin cancer. If we have a model that has very low precision, then many patients will be told that they have melanoma, and that will include some misdiagnoses. Lots of extra tests and stress are at stake. When false positives are too high, those who monitor the results will learn to ignore them after being bombarded with false alarms.

## 1. Decision Tree:

1.1 Confusion matrix :



ject DMT\Decision Tree.py

matrix - NumPy array

|   | 0 | 1 |
|---|---|---|
| 0 | 6 | 2 |
| 1 | 0 | 7 |

1.2 Classification report :

| Class | Precision | Recall | F1 Score | Support (num of examples) |
|---|---|---|---|---|
| 0 | 0.78 | 1.00 | 0.88 | 7 |
| 1 | 1.00 | 0.75 | 0.86 | 8 |
| Accuracy | 0.87 |  |  | 15 |

1.3 R2 score (dataset 2) :

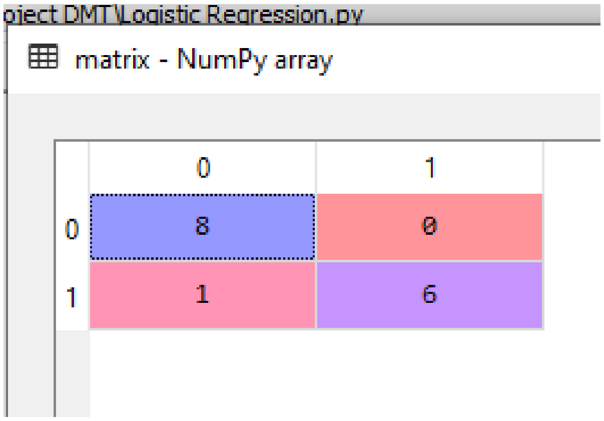
The R2 score comes out to be 0.84 using the Decision tree regressor.

The mean absolute error comes out to be 167163.3086041714

-----------------------------------------------------------------------------------------------

## 2. Logistic regression

2.1 Confusion Matrix:



oject DMT\Logistic Regression.py

matrix - NumPy array

|   | 0 | 1 |
|---|---|---|
| 0 | 8 | 0 |
| 1 | 1 | 6 |

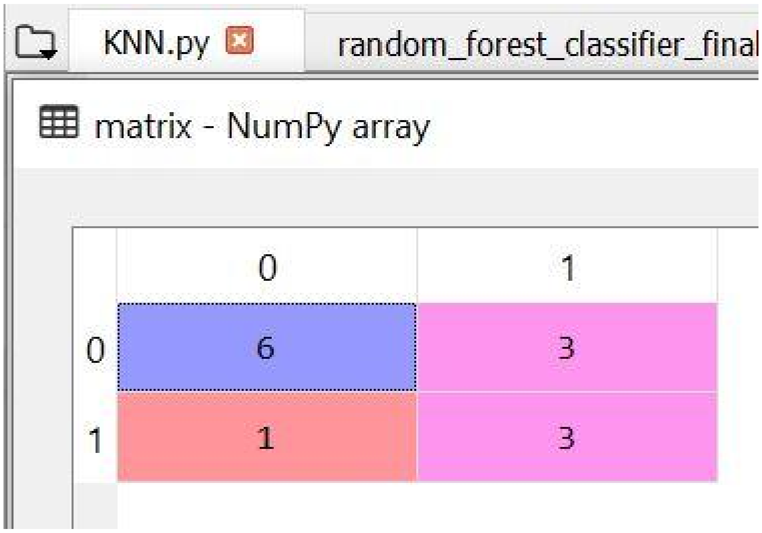
2.2 Classification report :

| Class | Precision | Recall | F1 Score | Support (num of examples) |
|---|---|---|---|---|
| 0 | 1.00 | 0.86 | 0.92 | 7 |
| 1 | 0.89 | 1.00 | 0.94 | 8 |
| Accuracy | 0.93 | | | 15 |

-----------------------------------------------------------------------------------------------

## 3. K nearest neighbor classifier

3.1 Confusion Matrix :



|   | 0 | 1 |
|---|---|---|
| 0 | 6 | 3 |
| 1 | 1 | 3 |

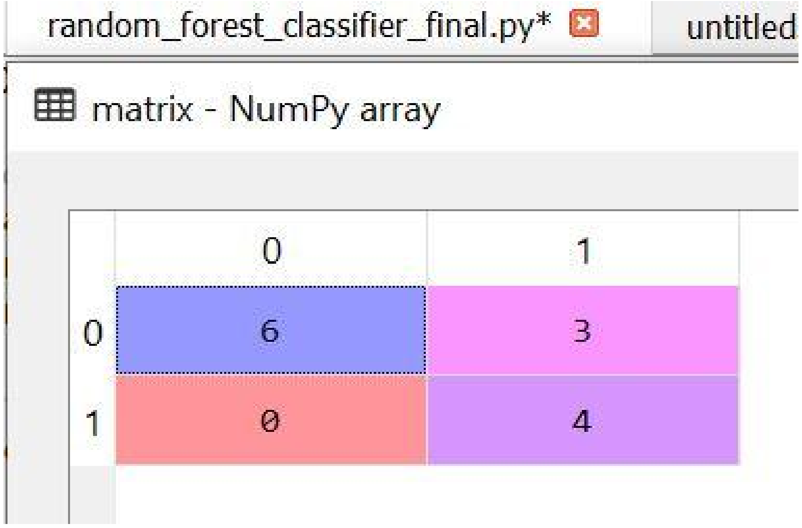
3.2 Classification report :

| Class | Precision | Recall | F1 Score | Support (num of examples) |
|---|---|---|---|---|
| 0 | 0.50 | 0.75 | 0.60 | 4 |
| 1 | 0.86 | 0.67 | 0.75 | 9 |
| Accuracy | 0.69 |  |  | 13 |

---

## 4. Random Forest Classifier

4.1 Confusion Matrix :



|   | 0 | 1 |
|---|---|---|
| 0 | 6 | 3 |
| 1 | 0 | 4 |

4.2 Classification report :

| Class | Precision | Recall | F1 Score | Support (num of examples) |
|---|---|---|---|---|
| 0 | 0.57 | 1.00 | 0.73 | 4 |
| 1 | 1.00 | 0.67 | 0.80 | 9 |
| Accuracy | 0.77 | | | 13 |

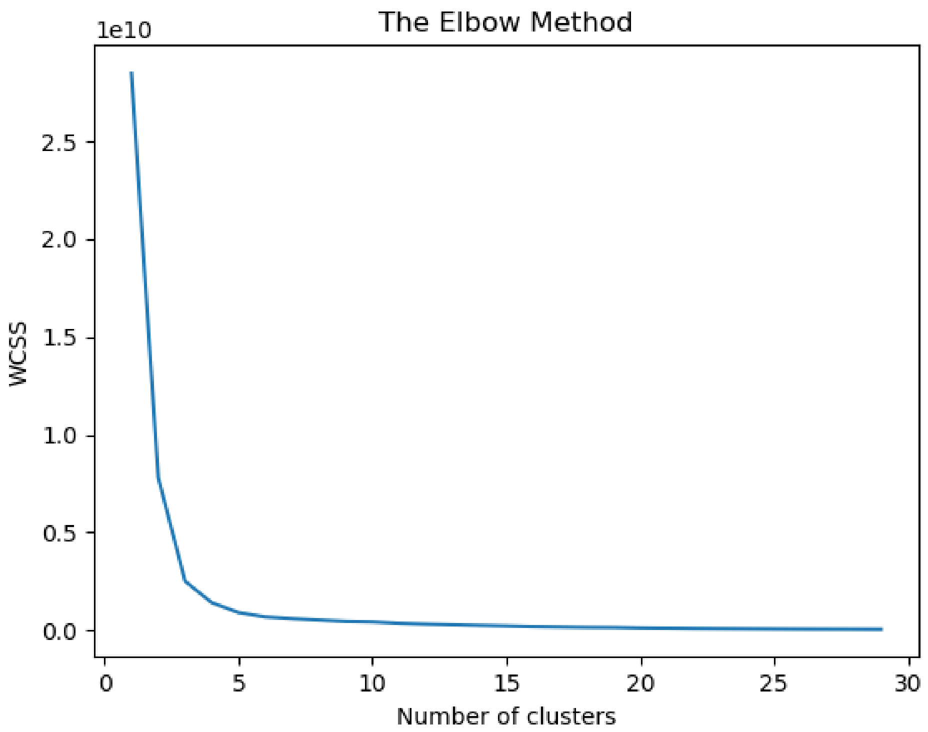
4.3 Random forest regressor on dataset 2

The mean absolute error comes out to be 155503.99436675265  
The R2 score is 0.91

---

## 8. Clustering

After applying clustering we plotted the elbow graph to check how many clusters gave optimal results. An ideal way to figure out the right number of clusters would be to calculate the Within-Cluster-Sum-of-Squares (**WCSS**). **WCSS** is the sum of squares of the distances of each data point in all clusters to their respective centroids. The idea is to minimise the sum.



This graph shows that 3 clusters are best suited for the dataset.

---------------------------------------------------------------------------------------------------

## 9. RESULTS AND DISCUSSION

| Algorithm | Precision | | Recall | | F1 Score | | Accuracy |
|---|---|---|---|---|---|---|---|
| | Class 0 | Class 1 | Class 0 | Class 1 | Class 0 | Class 1 | |
| Logistic Regression | 1.0 | 0.89 | 0.86 | 1.0 | 0.92 | 0.94 | 0.93 |
| Decision Tree | 0.78 | 1.0 | 1.0 | 0.75 | 0.88 | 0.86 | 0.87 |
| Random forest | 0.86 | 0.75 | 0.75 | 0.86 | 0.80 | 0.80 | 0.80 |
| K nearest | 0.50 | 0.86 | 0.75 | 0.67 | 0.60 | 0.75 | 0.69 |

| Algorithm | R2 score | Mean absolute error |
|---|---|---|
| Decision Tree | 0.84 | 167163.3086041714 |
| Random Forest | 0.91 | 155503.99436675265 |

For classification algorithms, Logistic regression performed the best for predicting the profit on a given crop, state, costs of cultivation (C1,C2), cost of production (Cp), and support prices provided by the government for the year 2020-21.

As we saw the second dataset did not perform so well. Additional columns like rainfall and temperature need to be added to improve the accuracy of the models.

## REFERENCES

[1] https://data.gov.in/

[2] https://farmer.gov.in/mspstatements.aspx

[3] N. P. Sastra and D. M. Wiharta, ―Environmental monitoring as an IoT application in building smart campus of UniversitasUdayana,‖ in Proc. Int. Conf. Smart Green Technol. Elect. Inf. Syst. (ICSGTEIS), Oct. 2016, pp. 85−88.

[4] M. Suganya., Dayana R and Revathi.R, Crop Yield Prediction Using Supervised Learning Techniques, International Journal of Computer Engineering and Technology, 11(2), 2020, pp. 9-20

[5] CH. Vishnu Vardhanchowdary, Dr.K.Venkataramana, Tomato Crop Yield Prediction using ID3, March 2018,IJIRT Volume 4 Issue 10 pp,663-62.

[6] R. Sujatha and P. Isakki, A study on crop yield forecasting using classification techniques 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16), Kovilpatti, 2016, pp. 1-4.

[7] N. Gandhi, L. J. Armstrong, O. Petkar and A. K. Tripathy, Rice crop yield prediction in India using support vector machines 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), KhonKaen, 2016, pp. 1-5.